Technical Note

VEROGEN

Establishing robust thresholds and filters for the ForenSeq MainstAY Kit

Out-of-the-box settings enable reliable implementation of NGS.

Highlights:

- **Configurable analysis methods.** Default thresholds based on extensive testing.
- **Multi-level filtering and sorting capabilities.** Streamlined workflow built for high volume laboratories.
- **Project and sample roll up summaries.** Low touch analysis and data assessments.

Introduction

Human DNA profiling using PCR at polymorphic short tandem repeat (STR) loci followed by capillary electrophoresis (CE) size separation and length-based allele typing has been the standard in the forensic community for over 20 years.

Next generation sequencing (NGS) addresses a number of challenges inherent in current methods such as the need to limit the STRs that can be simultaneously detected to <30 loci. This forces the forensic analysis of additional Y-STRs and X-STRs, if desired or required, in separate workflows. Furthermore, NGS enables forensic laboratories to process degraded or PCR inhibited DNA samples, commonly encountered in criminal casework, missing person cases, or mass disaster investigations, while providing the genetic discrimination power of sequenced-based typing where alleles of the same length are identified not only by the number of STR repeats but by the actual nucleotide-by-nucleotide STR sequence itself.

The ForenSeq MainstAY kit is optimized to support high quality reference samples and low quality casework samples. It enables the simultaneous typing of established 27 autosomal STRs (auSTRs), 25 Y-STRs, and Amelogenin in a single workflow. When coupled with the MainstAY Analysis Module in the Universal Analysis Software (UAS), this workflow provides a cost-effective option for forensic laboratories considering NGS as part of their operational workflows.

To ensure reproducible data analysis, the default MainstAY Analysis Method in the software has robust settings and thresholds based on developmental validation by Verogen scientists. To support the varying needs of forensic laboratories, the software also allows users to create their own analysis methods with custom thresholds. This technical note outlines the methods used to establish the default settings and guidelines used within the software as well as additional functionality available in the UAS to enable streamlined high-volume data analysis.

Low-touch features for high-volume STR analysis and reporting

The UAS is a forensically designed software that enables the analysis of data generated by ForenSeq MainstAY DNA libraries that were sequenced using the MiSeq FGx Sequencing System and the MiSeq FGx Reagent Micro Kit. It supports run setup, sample management and includes extensive data review, analysis, and report generation tools for human identification with STRs.

When sequencing is complete, the MiSeq FGx Sequencing System automatically transfers raw base calls to UAS as BCL files. UAS converts the base calls into sequence reads in FASTQ file format. The reads are then demultiplexed and assigned to the appropriate sample based on the index adapter sequences specified in the sample sheet. The use of unique dual indexes (UDIs) improves demultiplexing efficiency and optimizes data recovery. Unassigned reads are trimmed and the FASTQ files are aligned using the ForenSeq MainstAY manifest to generate raw alignments. Unaligned reads, which include primer dimers, primer adapters, chimeric reads, off-target reads, and allele drop-outs, are filtered to ensure high quality data for downstream interpretation and reporting. Alleles are aggregated and the number of reads associated with the alleles is counted and parsed according to the STR motif. This information is used to determine if the sample is potentially a mixture, call genotypes, and assign the appropriate QC flag to help with data visualization.

For research, forensic, or paternity use only. Not for use in diagnostic procedures.

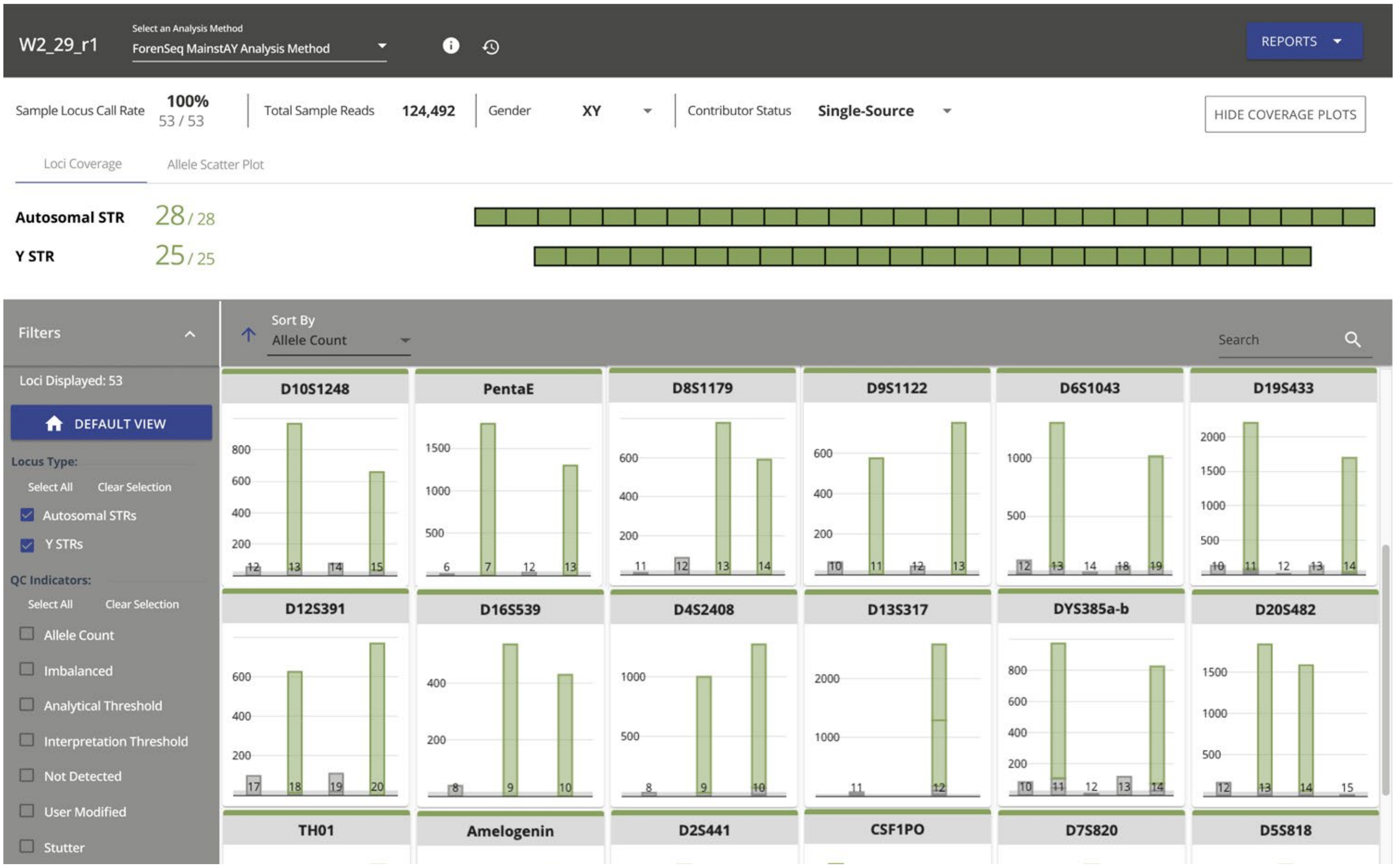
The MainstAY Analysis Software is specifically designed to support the high operational efficiency needs of forensic analysts who are simultaneously analyzing 96 samples generated using the ForenSeq MainstAY Kit on a MiSeq FGx Reagent Micro Kit. Once a sequencing run is analyzed, selecting a sample from a project allows in-depth assessment of the markers with high level sample data such as call rate, total reads, gender, and contributor status, as well as flexible data visualization options like loci-based heatmaps or allele-based scatter plots (Figure 1).

The sample details page also includes sorting options and filters based on a variety of QC indicators to facilitate easy data organization. The sample details page can be maximally viewed by hiding the project sidebar and the filtering options. Detailed histogram views enable simultaneous review of multiple markers, and locus details provide additional information, such as sequence context and user modifiable options to type or untype an allele. Data can be quickly reviewed using the sample roll up summaries, filter and sort capabilities, and histogram view of all STRs. Table 1 summarizes the complete list of filtering and sorting options.

To supplement the guided exploration of onscreen results, the software offers three types of reports for

Table 1: Filters and sort options.

Feature	Description
Metadata filters	Locus Type
QC indicator filters	Allele Count, Imbalanced, Analytical Threshold, Interpretation Threshold, Not Detected, User Modified, Stutter, No QC Indicators
Sorting	Allele Count, Amplicon Size, Chromosome, Intralocus Balance, Intensity, STR Name, STR Type



W2_29_r1

Select an Analysis Method: ForenSeq MainstAY Analysis Method

Sample Locus Call Rate 100% 53 / 53 | Total Sample Reads 124,492 | Gender XY | Contributor Status Single-Source

Autosomal STR 28 / 28

Y STR 25 / 25

Loci Displayed: 53

Locus Type:

☑ Autosomal STRs

☑ Y STRs

QC Indicators:

□ Allele Count

□ Imbalanced

□ Analytical Threshold

□ Interpretation Threshold

□ Not Detected

□ User Modified

□ Stutter

Sort By: Allele Count

Figure 1: Universal Analysis Software allows visualization of STR read intensity and repeat length for each locus. The display can be filtered or sorted to modify the view.

packaging STR data:

- **Project**. Results for all samples in a project compiled in one report.
- **Sample**. Locus with genotypes for each STR type along with a summary of thresholds, QC indicators, and coverage information.
- **CODIS**. Variants from selected samples complied for upload to CODIS.

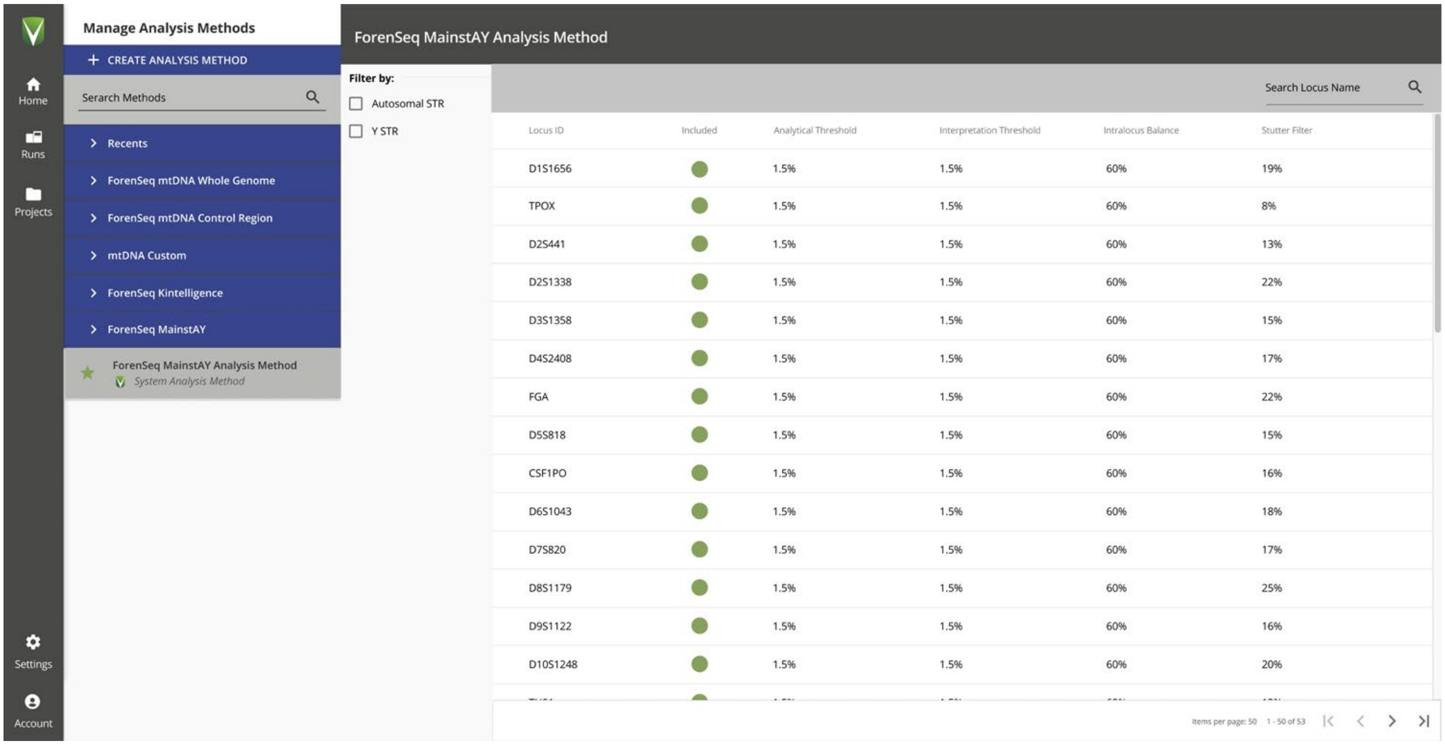
Thresholds and guidelines for confident STR calls and reporting

The UAS provides a default analysis method—the MainstAY Analysis Method—that calls and evaluates STRs typed using settings recommended by Verogen (Table 2). Forensic laboratories can implement this method or use it as a template, configuring the settings based on their own assessments of the ForenSeq MainstAY Kit as shown in Figure 2. Configurable settings include analytical threshold (AT), interpretation threshold (IT), intralocus balance, (ILB) and which loci to analyze. The UAS also includes guidance for the sample read count, a metric that users can access from the sample representation bar chart presented on the user interface (Figure 2). Additionally, users can create multiple analysis methods and reanalyze samples using the different methods.

Analytical and interpretation thresholds

The analytical threshold default setting for the ForenSeq MainstAY Analysis Method was established by analyzing one sequencing run of 96 negative amplification control (NTC) libraries generated using the ForenSeq MainstAY Library Prep Kit on the MiSeq FGx Sequencing System with the MiSeq FGx Reagent Micro Kit. AT and IT values were determined for a locus by multiplying the analysis parameter percentage value by the sum of read counts at that locus. In cases of low coverage, a minimum read number of 650 reads was used for the locus in determination of the threshold values. The average (0.0%) plus three times the standard deviation (0.5%) was used to set the analytical threshold of 1.5%. The interpretation threshold was set equal to the analytical threshold. Table 2 summarizes the final thresholds implemented in the default ForenSeq MainstAY Analysis Method.

To verify the analytical threshold setting of 1.5%, libraries were prepared from extracted gDNA from 305 individuals of African American, Caucasian, or admixed American ancestry1 using the ForenSeq MainstAY Library Prep Kit. The libraries were sequenced at a plexity of 96 samples with the MiSeq FGx Reagent Micro Kit on the MiSeq FGx. DYS389II, DYS481, and DYS612 demonstrated higher noise



ForenSeq MainstAY Analysis Method

Locus ID	Included	Analytical Threshold	Interpretation Threshold	Intralocus Balance	Stutter Filter
D1S1656	●	1.5%	1.5%	60%	19%
TPOX	●	1.5%	1.5%	60%	8%
D2S441	●	1.5%	1.5%	60%	13%
D2S1338	●	1.5%	1.5%	60%	22%
D3S1358	●	1.5%	1.5%	60%	15%
D4S2408	●	1.5%	1.5%	60%	17%
FGA	●	1.5%	1.5%	60%	22%
D5S818	●	1.5%	1.5%	60%	15%
CSF1PO	●	1.5%	1.5%	60%	16%
D6S1043	●	1.5%	1.5%	60%	18%
D7S820	●	1.5%	1.5%	60%	17%
D8S1179	●	1.5%	1.5%	60%	25%
D9S1122	●	1.5%	1.5%	60%	16%
D10S1248	●	1.5%	1.5%	60%	20%

Figure 2: Default analysis method for the ForenSeq MainstAY workflow that displays the out-of-the-box AT, IT, ILB, and Stutter Filters. Laboratories can use the analysis method as a template and customize filters based on their internal validations.

relative to the other loci in the multiplex, and higher ATs of 4.5%, 2%, and 2%, respectively, were set for these loci.[1]

Sample read count guideline for quality control

The sample representation chart in the software displays the numbers of reads (intensity) for each sample in a run as well as the percent coefficient of variance (%CV) for the samples across the run (Figure 3). A guideline is shown in light blue for the number of reads recommended per sample as a measure of quality control. If a sample has fewer reads than the recommended guideline, it could potentially have low or no coverage of some loci. This guideline is set using analyses of libraries generated from male DNA. Libraries generated with female DNA only have signal for the autosomal STRs and therefore require

Table 2: Default thresholds and guidelines for the ForenSeq MainstAY Analysis Method.

Setting	Description	Default Value	Config.
Analytical threshold (AT)	The value that a read count must reach for the software to type an allele.	>1.5%[1,2]	Yes
Interpretation threshold (IT)	Single auSTR allele read counts greater than IT are called homozygous. Counts below IT but above AT are called ambiguous.	>1.5%[1,2]	Yes
Intralocus balance	The balance of read counts between typed alleles at a heterozygous locus.	60%	Yes
Sample representation	The number of reads per sample for a run recommended to achieve optimal coverage of all loci.	15,000 reads[3]	No

1. In cases of low coverage, a minimum of 650 reads is used for the calculation of the analytical threshold thus setting the minimum number of reads for a call at 11 reads.
2. DYS389II has an analytical threshold of >4.5%, and DYS481 and DYS612 have an analytical threshold of >2%.
3. The sample representation value is provided as a guideline only and is not a setting in the software.

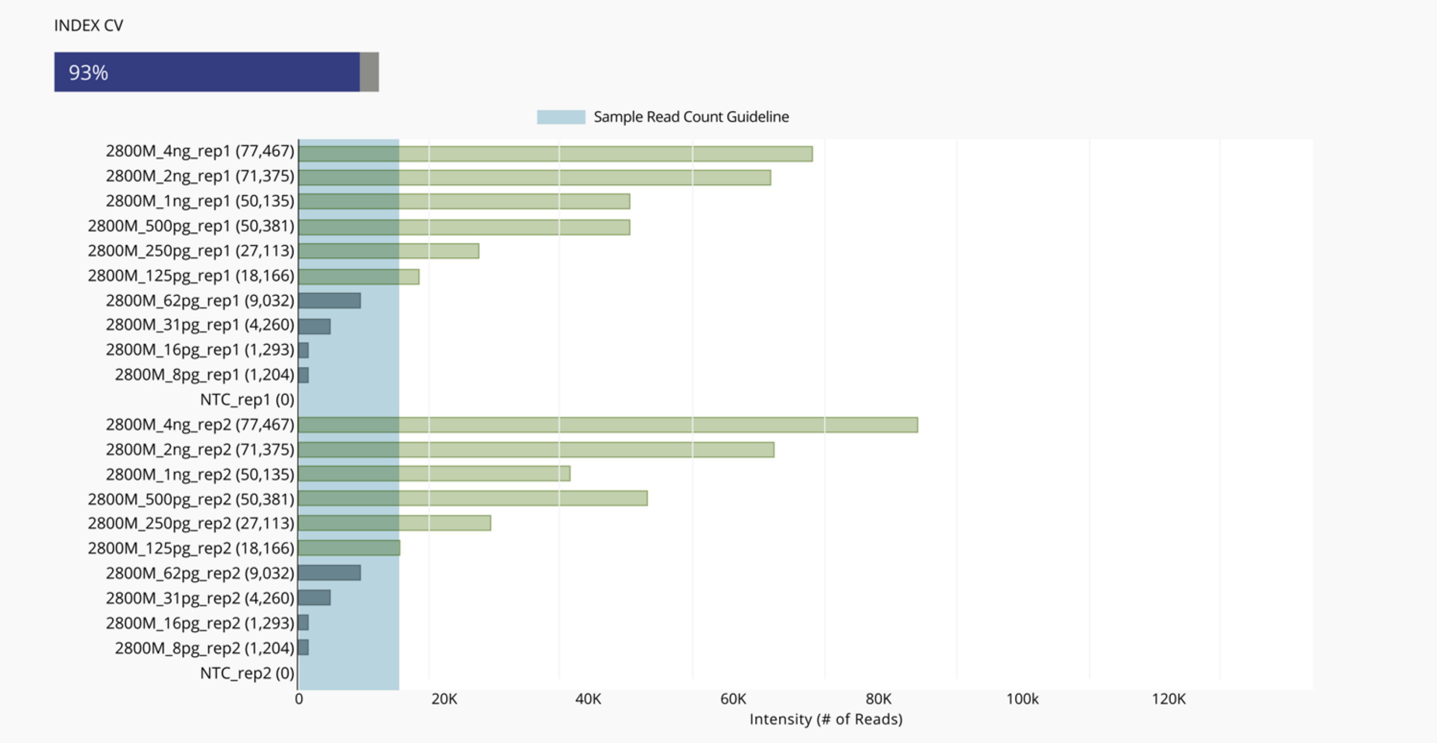


Figure 3: Sample representation for a ForenSeq MainstAY Analysis Method displays the read counts for each sample relative to the sample read count guideline, indicated with the blue shading (96-sample plexity; not all samples are shown).

lower total sample reads (~10,000 reads per sample) than libraries generated with male DNA for optimal coverage of all loci.

To determine the read count guideline for the ForenSeq MainstAY samples, data from two sensitivity experiments were analyzed with the UAS v2.3 software. The sensitivity studies were performed using libraries generated using the ForenSeq MainstAY Library Prep kit with three different male DNAs: 2800M DNA2 (40 libraries), 9948 DNA3 (33 libraries), and NA24385 DNA4 (40 libraries, positive amplification control DNA). The libraries for these experiments were sequenced at 96-sample plexity using the MiSeq FGx Reagent Micro Kit on two different MiSeq FGx instruments.

The thirty-three 9948 DNA libraries were prepared in triplicate at 4,000, 2,000, 1,000, 500, 250, 125, 63, 31, 16, 8, and 4 pg of gDNA. The minimum, maximum, and average total read counts were 1,534, 56,886, and 22,970 respectively. The 18 samples with total reads greater than or equal to 15,000 returned full profiles in the software with no loci below the analytical threshold. Of these 18 libraries, at most one locus had coverage lower than 100 reads.[3]

The 40 DNA libraries, each consisting of 2800M and NA24385 DNA, were prepared in quadruplicate at inputs 4,000, 2,000, 1,000, 500, 250, 125, 63, 31, 16, and 8 pg of gDNA. The minimum, maximum, and average aligned read counts were 834, 147,095, and 42,797 respectively. The 53 samples with total reads greater than or equal to 15,000 returned full profiles in the software with no loci below the analytical threshold. Of these 53 libraires, at most six loci had coverage lower than 100 reads.[2,5]

To verify the read count guideline, two runs of libraries were prepared using the ForenSeq MainstAY Library Prep Kit. Libraries were prepared either from 160 pg extracted gDNA extracted from 94 individuals of African American, Caucasian, or admixed American ancestry1 with two negative amplification controls or with 1 ng gDNA extracted from 95 Coriell cell-line DNA samples5 from individuals of African, Asian, or Caucasian ancestry with one negative amplification control. The libraries were then sequenced utilizing MiSeq FGx Reagent Micro Kits on the MiSeq FGx. Of the 94 160-pg samples, only one sample returned a partial profile with one locus (DYS389II) with reads below the analytical threshold.[4,5]

This sample was one of three samples having total reads of <15,000. The other two returned full profiles. Of the 95 1-ng Coriell samples, one sample returned a partial profile with one locus (DYS19) with reads below the analytical threshold. This sample had 28,106 reads, while the 12 samples with reads below 15,000 all returned full profiles. Based on these results, a read count guideline of 15,000 total reads for each sample is recommended to provide confidence that samples yield full profiles. However, samples with total number of reads lower than the guideline might still generate sufficient data for analysis.

Stutter filtering

Stutter is defined as polymerase slippage while copying repetitive sequences during PCR for library preparation or clustering on the MiSeq FGx flow cell to generate library molecules with less than or greater than the repeat length of the original allele. Stutter is computed as sequence stutter, where the UAS checks the repeat units in the STR sequence. This computation can enable the separation of a minor contributor allele from stutter when the sequences differ. A sequence is considered stutter, with offset n repeat units if:

- The sequences differ in length by k times the reference repeat unit length.
- The number of repeat units in the potential stutter is identical to an adjacent unit of length of the reference repeat unit.
- The are no (other) mismatches between the sequences.

In the case of tri-, tetra-, and penta-nucleotide repeats, the reference repeat unit lengths are 3, 4, and 5 respectively. The software assesses differences in the sequences for k = -2, -1, and +1 repeat units, as described in Table 3.

Table 3: Positions used by the UAS for assessing difference in sequences.

Stutter Repeat	Positions[1]
Tri-nucleotide repeats	n-6, n-3, n+3
Tetra-nucleotide repeats	n-8, n-4, n+4
Penta-nucleotide repeats	n-10, n-5, n+5

1. n is the length in the base pairs of a given allele

The stutter filter setting depends on the differences in repeat units. The stutter filter setting is applied for any given locus at the user defined value for the k=-1 position. The stutter filter setting at the k=+1 and k=-2 positions is the square of the user defined value for the k=-1 position. The expected stutter intensity is the product of the stutter filter setting and the called allele intensity at a repeat length. For example, a stutter filter setting of 10% (0.1) is squared (0.1 x 0.1 = 0.01 or 1%) before multiplying by the intensity of the called parent allele for either the k=+1 or k=-2 position. A stutter quality control indicator for elevated stutter is displayed when both of these conditions exist:

- Uncalled read intensity, with sequence of a potential stutter of a called allele.
- Uncalled read intensity is greater than the maximum expected stutter % of the potential parent allele.

The stutter thresholds were set using the default analytical threshold with all of the 1-ng libraries described above and an additional set of six sequencing runs of 32 NA24385 1-ng positive amplification control DNA libraries. The percent stutter for the k=-1, k=-2, and k=+1 stutter products relative to the major allele were calculated. The mean, standard deviation, and maximum for the percent stutter for each stutter type and for each locus was calculated and the distribution of stutter analyzed. Several different calculation methods were used to set conservative stutter filters in order to filter the stutter for these samples. The mean plus three standard deviations of the k=-1 stutter products was used to set the stutter filters for nineteen of the loci (k=-1 mean + 3SD, Table 4). Seven loci with higher k=-1 stutter were set using the maximum k=-1 stutter (k=-1 max, Table 4).

The MainstAY library prep kit was optimized to minimize PCR bias against size which results in more k=+1 stutter than seen with traditional CE STR kits. To conservatively

Table 4. Default settings for stutter filtering in the default ForenSeq MainstAY Analysis Method

Autosomal STRs			Y-STRs		
Locus	**Stutter threshold**	**Method**	**Locus**	**Stutter threshold**	**Method**
CSF1PO	0.16	k=+1 max	**DYS19**	0.15	k=+1 max
D1S1656	0.19	k=-1 mean+3SD	**DYS385a-b**	0.32	k=-1 max
D2S441	0.13	k=+1 max	**DYF387S1**	0.21	k=-1 mean+3SD
D2S1338	0.22	k=-1 max	**DYS389I**	0.15	k=-1 mean+3SD
D3S1358	0.15	k=+1 max	**DYS389II**	0.19	k=-1 max
D4S2408	0.17	k=+1 max	**DYS390**	0.13	k=-1 mean+3SD
D5S818	0.15	k=+1 max	**DYS391**	0.15	k=-1 mean+3SD
D6S1043	0.18	k=+1 max	**DYS392**	0.17	k=-1 mean+3SD
D7S820	0.17	k=+1 max	**DYS437**	0.13	k=+1 max
D8S1179	0.25	k=-1 mean+3SD	**DYS393**	0.15	k=+1 max
D9S1122	0.16	k=-1 mean+3SD	**DYS438**	0.1	k=+1 max
D10S1248	0.2	k=-1 max	**DYS439**	0.11	k=+1 max
D12S391	0.29	k=-1 mean+3SD	**DYS448**	0.03	k=-1 mean+3SD
D13S317	0.15	k=+1 max	**DYS460**	0.16	k=+1 max
D16S539	0.16	k=-1 mean+3SD	**DYS481**	0.47	k=-1 mean+3SD
D17S1301	0.21	k=-1 max	**DYS505**	0.16	k=-2 max
D18S51	0.18	k=-1 mean+3SD	**DYS522**	0.2	k=+1 max
D19S433	0.19	k=-1 max	**DYS533**	0.15	k=+1 max
D20S482	0.15	k=-1 mean+3SD	**DYS549**	0.15	k=+1 max
D21S11	0.16	k=+1 max	**DYS570**	0.18	k=-1 mean+3SD
D22S1045	0.23	k=+1 max	**DYS576**	0.17	k=+1 max
FGA	0.22	k=+1 max	**DYS612**	0.38	k=-1 mean+3SD
PentaD	0.05	k=-1 mean+3SD	**DYS635**	0.15	k=+1 max
PentaE	0.12	k=-1 max	**DYS643**	0.11	k=+1 max
TH01	0.12	k=-1 mean+3SD	**Y-GATA-H4**	0.16	k=+1 max
TPOX	0.08	k=+1 max			
vWA	0.17	k=-1 mean+3SD			

filter stutter products for loci with k=+1 stutter, default stutter filters were set using the square root of the maximum k=+1 stutter (k=+1 max) for the twenty five loci (indicated in Table 4). One locus (DYS505) has higher k=-2 stutter, therefore the stutter filter for this locus was set using the square root of the maximum k=-2 stutter (k=-2 max, Table 4).

Contributor status and gender estimation

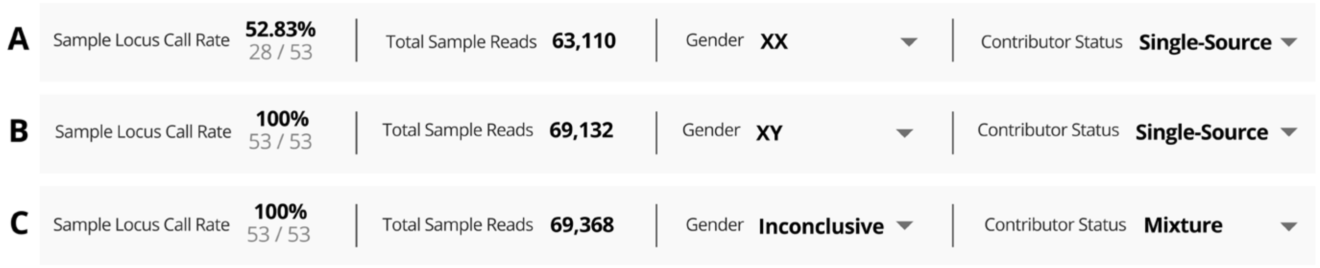
The contributor status and biological sex for each sample is calculated after analyzing the STRs and applying the thresholds and stutter filters (Figure 4). The software first determines the contributor status as contributor status can impact the biological sex determination.

The UAS determines the contributor status of the sample as shown in Table 5 after STR analysis. Negative amplification controls are always designated as inconclusive in the software. Autosomal loci, DYS385a-b, and DYF387S1 Y-STR loci have a maximum of two possible typed alleles, and the remaining Y-STR have a maximum of one possible typed allele. If three or more loci in a sample have more than the possible number of alleles typed for those loci, the sample is called a mixture. A sample is marked single-source if there are no loci with a number of typed alleles exceeding the possible number of alleles. If there are one to two loci with more than the possible number of alleles typed for those loci, the sample is marked inconclusive.

The biological sex of the sample is based on detection of Y-STR loci, coverage of the autosomal STRs, and contributor status (Table 6). Negative amplification controls are always designated as inconclusive in the software as are samples that have a contributor status of mixture. Samples are designated XX when there are no Y-STRs typed and at least 50% of the autosomal STRs have a typed allele. Samples are designated XY if at least 5 Y-STR loci are typed. If a sample is not XX or XY by these criteria, it is marked inconclusive.

Conclusion

The ForenSeq MainstAY kit and the MiSeq FGx Sequencing System enable simultaneous PCR amplification and sequencing of autosomal and Y-STRs in a single reaction with one workflow, for maximum information potential and operational efficiency. When coupled with the UAS, this workflow demonstrates robust, reliable, reproducible, and semi-automated allele calling that meets established forensic validation guidelines. The workflow maintains backward compatibility of allele calling with existing law enforcement STR databases. Quality indicators and project roll up views in the UAS



A
Sample Locus Call Rate 52.83% 28 / 53
Total Sample Reads 63,110
Gender XX
Contributor Status Single-Source

B
Sample Locus Call Rate 100% 53 / 53
Total Sample Reads 69,132
Gender XY
Contributor Status Single-Source

C
Sample Locus Call Rate 100% 53 / 53
Total Sample Reads 69,368
Gender Inconclusive
Contributor Status Mixture

Figure 4: Screenshots of the sample overview section on the results page for a biologically female sample (A), biologically male (B), and a mixture of male and female samples (C).

Table 5. Contributor Status determination in the MainstAY Analysis Module

Single-source	Mixture	Inconclusive
For a sample to receive a contributor status of single-source, all of the following conditions must apply: • The sample is not a negative control. • Of the loci in the sample, no loci have more alleles called than is possible for that locus (e.g., 2 alleles possible for autosomal STR loci).	For a sample to receive a contributor status of mixture, all of the following conditions must apply: • The sample is not a negative control. • Of the loci in the sample, at least 3 loci have more alleles called than is possible for that locus (e.g., 2 alleles possible for autosomal STR loci).	Negative control samples are always assigned a contributor status of inconclusive. A sample is also assigned a contributor status of inconclusive if it meets this condition: • Of the loci in the sample, one or two loci have more alleles called than is possible for that locus (e.g., 2 alleles possible for autosomal STR loci).

Table 6. Biological Sex Determination in the MainstAY Analysis Module

XX	XY	Inconclusive
For a sample to receive the designation of XX, all of the following conditions must apply: • The sample is not a negative control. • Of the autosomal STR loci in the sample, at least 50% of the loci have a signal above the analytical threshold. • Of the Y-STR loci in the sample, no Y-STR loci have signal greater than the analytical threshold.	For a sample to receive the designation of XY, all of the following conditions must apply: • The sample is not a negative control. • Of the Y-STR loci in the sample, at least 5 loci have a signal above the analytical threshold.	Negative control samples are always assigned as inconclusive. A sample is also assigned a inconclusive if any of these conditions apply: • The contributor status of the sample is designated a mixture (i.e., >= 3 loci with more alleles than expected for that locus). • Of the Y-STR loci in the sample, 1–4 Y-STR loci have an allele above the analytical threshold. • Of the autosomal STR loci in the sample, fewer than 50% of the loci have a signal above the analytical threshold and no Y-STR loci are typed.

enable quick evaluations of information content in each run to verify semi-automated allele calls relative to default thresholds and filters.

Individual laboratories can implement rigorously tested default analysis parameters or modify them by enacting internal policies, giving them a powerful tool for analyzing NGS data on their own terms. Importantly, the UAS combines these extensive data management and analysis capabilities in an audit-controlled, intuitive user interface that is designed to simplify NGS STR analysis. Because of the seamless integration of end-to-end sequencing and analysis, laboratories that are new to NGS and those planning on scaling their operations can implement the ForenSeq MainstAY workflow without compromising quality or ease-of-use.

References

1. Innogenomics Inc, New Orleans, Louisiana, USA
2. Promega Corp, Madison, Wisconsin, USA
3. MCLAB, South San Francisco, California, USA
4. Verogen, San Diego, California, USA
5. Coriell Institute for Medical Research, Camden, New Jersey, USA

Product documentation is available for download at **www.verogen.com/support**

For research, forensic, or paternity use only. Not for use in diagnostic procedures.