Technical Note

VEROGEN

Analysis Settings in ForenSeq Universal Analysis Software v2.0

Verogen analysis methods offer fast, flexible, and accurate variant calling of mitochondrial DNA data.

Summary: Providing three analysis methods with default settings, one method for each assay type, Verogen software streamlines analysis of sequencing data. Apply one of these rigorously tested methods for confidence in high-quality results or create a copy and set up a custom analysis method that tunes your output to precise specifications.

Introduction

ForenSeq™ Universal Analysis Software (UAS) v2.0 offers a dynamic and intuitive solution for analyzing mitochondrial DNA (mtDNA) data, integrating seamlessly with the MiSeq FGx® Sequencing System to support a diverse range of mtDNA assays. Verogen next-generation sequencing (NGS) technology sequences and analyzes the mitochondrial genome (mtGenome) in fewer than two days, enabling applications that require improved data quality and resolution for challenging samples with fast turnaround times. To simplify and accelerate this complex analysis, the software includes three default analysis methods. Each method targets a distinct workflow, from control region and whole genome interrogation with ForenSeq mtDNA library prep to custom analysis of libraries prepared with third-party or home-brew assays.[1]

Although the default analysis methods provide efficient and reliable variant calling, some laboratories have unique requirements for forensic casework or DNA databasing. To ensure flexibility in these cases, Verogen designed ForenSeq UAS v2.0 with configurable settings so laboratories can develop and set thresholds to suit a variety of analysis needs. Without the obstacle of analysis methods that are misaligned with the desired specifications, laboratories are free to take full advantage of this user-friendly analysis solution. This technical note describes how Verogen determined the optimum default thresholds for mtDNA analysis and documents the supporting data.

Table 1: Default settings for Verogen analysis methods

Setting	Verogen mtDNA Control Region Analysis Method	Verogen mtDNA Whole Genome Analysis Method	Verogen mtDNA Custom Analysis Method
Analytical Threshold	10%	6%	10%
Interpretation Threshold	10%	6%	10%
Minimum Quality Score	Q30	Q30	Q30
Minimum Read Count	64 reads	45 reads	64 reads
Library Type*	Primer-directed sequencing	Primer-directed sequencing	Non-directed sequencing

* Customizable when using the Verogen mtDNA Custom Analysis Method as a template.

Flexible analysis methods to tune output

Advance your casework with a proven Verogen analysis method or set up a custom analysis method to expand options and unlock a broader range of applications. An analysis method is a collection of settings that inform allele or variant calls from NGS data. For each of the three assay types included in ForenSeq UAS v2.0—ForenSeq mtDNA Control Region, ForenSeq mtDNA Whole Genome, and mtDNA Custom—the software provides a corresponding default analysis method. These default methods also serve as a straightforward template for creating a custom analysis method: copy any of the default methods and with a few clicks modify the settings to suit your requirements. Customizable settings include minimum quality score (Q-score), minimum read count, analytical threshold (AT), and interpretation threshold (IT) (Table 1). Provided the custom analysis method is based on the Verogen mtDNA Custom Analysis Method, the library type setting is also customizable. After analyzing data with the default method, you can apply a custom method to calibrate results and easily switch between analysis methods to review results with a variety of settings applied. The software preserves data from all analyses.

Summary of thresholds

- Minimum Q-score is the quality threshold required to call a base. Q-scores are an established metric for measuring base calling accuracy and assessing the quality of sequencing data.[2]
- Minimum read count defines the lower limit of the base-specific number of reads the software uses to assign a base call for a position in the mitochondrial genome (mtGenome). To be eligible for base calling, a position must reach the minimum read count.
- AT is the value that a read count must meet to contribute to a call. If a read count falls below this threshold, the call is not visible in the software and not reported as part of the call.
- IT is the value that a base must meet to contribute to a call (Figure 1). A base with a total number of reads greater than or equal to the IT contributes to the call for the mtDNA coordinate. When multiple bases exceed this threshold, the International Union of Pure and Applied Chemistry (IUPAC) code for the bases is reported.

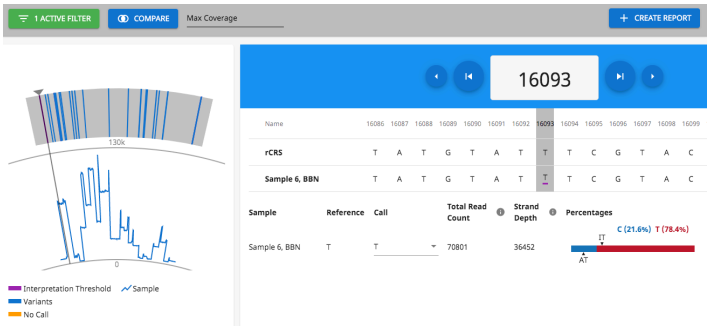

Figure 1: The software indicates when the total number of reads for a base do not meet the interpretation threshold. In this control region example, the number of T reads falls below the specified IT value.

Table 2: Default amplicons for non-directed sequencing

Setting	MTL F1 – MTL R1	MTL F2 – MTL R2
Forward Primer Start Coordinate	9397	15,195
Reverse Primer Start Coordinate	1892	9796
Forward Primer Length	20	20
Reverse Primer Length	20	20

Summary of library types

The MiSeq FGx System can generate custom mtDNA data from two library types: primer-directed sequencing and non-directed sequencing. When creating an analysis method based on the Verogen mtDNA Custom Analysis Method, you can choose the library type and add or remove amplicons and their corresponding settings for start coordinates and primer lengths. By default, the non-directed sequencing library type includes the amplicons MTL F1 – MTL R1 and MTL F2 – MTL R2. Table 2 presents the default settings included with each of these amplicons. All settings are adjustable, including the amplicon names.

- With primer-directed sequencing, one strand is always sequenced as Read 1 and the opposite strand is always sequenced as Read 2. Examples of primer-directed sequencing are the Verogen ForenSeq mtDNA kits, which use a PCR-based approach to create sequencing targets (amplicons).
- With non-directed sequencing, either strand can be sequenced as Read 1 or Read 2. An example of non-directed sequencing is the Illumina Nextera XT DNA Library Prep Kit, which uses a transposase-based approach to prepare sequencing libraries. Non-directed sequencing uses a library prep protocol that targets the whole mtGenome.[3]

Determining thresholds for accurate calls

Minimum Q-score

For all analysis methods, Verogen selected a default minimum Q-score of Q30, which is considered the benchmark for NGS data quality. Verogen NGS technology leverages the Illumina method of quality score assignment, which is based on the Phred algorithm. When the algorithm assigns Q30 to a base, the probability of an incorrect base call is only 1 in 1000 for an accuracy rate of 99.9%. When sequencing quality reaches Q30, virtually all reads are perfect. Lowering the base call accuracy to even 99% (Q20) elevates the probability of an incorrect base call to 1 in 100.[2]

Minimum read count

To determine the default minimum read count for the Verogen mtDNA Control Region Analysis Method, Verogen assessed background signal on two MiSeq FGx Systems using MiSeq FGx Reagent Micro Kits. ForenSeq mtDNA Control Region libraries and 48 water-only negative amplification controls were evaluated across two runs. Data analyzed in ForenSeq UAS v2.0 with the Verogen mtDNA Control Region Analysis Method indicated a mean of 34 reads per position in the negative amplification controls, with a standard deviation of 30 reads across the 1157 positions evaluated in the control region. Calculating one standard deviation above the mean read number provided the default minimum read count of 64 reads per base at a position.

To determine the default minimum read count for the Verogen mtDNA Whole Genome Analysis Method, Verogen applied the same methods used for control region analysis. ForenSeq mtDNA Whole Genome libraries and 16 water-only negative amplification controls were evaluated over two runs on the MiSeq FGx System using MiSeq FGx Reagent Kits. Data analyzed in ForenSeq UAS v2.1 indicated a mean of one read per amplicon in the negative amplification controls with a standard deviation of 14 reads across the 245 amplicons that cover the entire mtGenome. Calculating three standard deviations above the mean read number provided the default minimum read count value of 45 reads per base at a position.

When the default minimum read count is applied, the software does not report a call or coverage for a position with fewer than 63 reads for control region or 44 reads for whole genome for a specific base (A, G, C, or T). The software does report a call when at least 64 reads for control region or 45 reads for whole genome are present at a position for at least one base *and* the call complies with other parameters. A call is supported when it meets or exceeds the AT, minimum Q-score, and minimum read count. For example, when 100 T reads are present at a position that meet the AT and Q-score, the software calls a base. Conversely, when a position has 50 T reads and 50 A reads, the software does not call a base and does not report coverage for control region, regardless of whether the minimum AT and Q-score are met. Similarly, when a position has 30 T reads and 30 A reads, the software does not call a base and does not report coverage for whole genome, regardless of whether the minimum AT and Q-score are met.

Analytical and interpretation thresholds

To determine the default AT for the Verogen mtDNA Control Region Analysis Method, background signal was assessed across four operators using four MiSeq FGx Systems with MiSeq FGx Reagent Micro Kits. Libraries were generated with the ForenSeq mtDNA Control Region Kit and sequenced over 10 runs. Each run sequenced 2–18 control DNA libraries (HL60) for a total of 63 positive amplification controls at 100 pg each. After sequencing, data were analyzed in ForenSeq UAS v2.0 using a minimum read count of 64 and AT and IT values of 0%. For the Verogen mtDNA Whole Genome Analysis Method, Verogen conducted a similar assessment across six operators using five MiSeq FGx Systems with MiSeq FGx Reagent Kits. Libraries were generated with the ForenSeq mtDNA Whole Genome Kit and sequenced over six runs. Each run sequenced 2–16 HL60 libraries for a total of 90 positive amplification controls at 100 pg each. After sequencing, data were analyzed in ForenSeq UAS v2.1 using a minimum read count of 45 and AT and IT values of 0%.

Verogen collated unexpected variants in the HL60 libraries for each analysis method to confirm the optimum percentage of reads for the default AT and IT. Table 3 summarizes the percentage of unexpected variants across the runs with the AT and IT. For control region, each maximum data point occurred within the hypervariable I (HVI) and hypervariable II (HVII) C-stretches or the AC repeat at positions 523–524. Excluding these locations, the maximum percentage of unexpected variants was 2.1%. This default AT, which rounds the maximum unobserved variant percentage to the nearest whole integer, supports detection of the known HL60 haplotype from the positive amplification controls.

When an AT of 3.7% (0.7% +3SD) is set in the Verogen mtDNA Control Region Analysis Method, the software calls 3.5% of unexpected variants (147 of 4157) in the C-stretch and AC repeat. Consider lowering the AT if detection of less than 10% heteroplasmy or minor contributors in mixtures is desired. For example, with careful interpretation of the HVI and HVII C-stretches and AC repeat at positions 523–524, a 3.7% AT can help detection and interpretation of heteroplasmy or minor contributors at 5%.

Similarly, for the Verogen mtDNA Whole Genome Analysis Method, applying an AT of 3% (0.7% +3SD, rounded up) results in the software calling 0.6% (91 of 16,203) of unexpected variants in the C-stretch, AC repeat, and other homopolymeric regions (for example, poly A tract at 12,418–12,425). Consider lowering the AT if detection of less than 6% heteroplasmy or minor contributors in mixtures is desired. For example, with careful interpretation of the HVI and HVII C-stretches, AC repeat at positions 523–524, and homopolymeric regions, a 3% AT can help detection and interpretation of heteroplasmy or minor contributors at 5%.

Sample read count guideline for quality control

The software provides the sample read count guideline on a sample representation bar chart, which displays the number of reads (intensity) and read distribution for each sample in a run (Figure 2). Although not customizable, the read count guideline complements analysis method settings by providing a useful metric for reviewing run and sample quality. To determine the sample read count guidelines, Verogen conducted sensitivity studies and evaluated known samples using data generated from two runs sequencing ForenSeq mtDNA Control Region libraries with the MiSeq FGx Reagent Micro Kit and four runs sequencing ForenSeq mtDNA Whole Genome libraries with the MiSeq FGx Reagent Kit.

Data from the two runs of ForenSeq mtDNA Control Region libraries were analyzed in ForenSeq UAS v2.0 with a minimum read count of 64, an AT of 0%, and an IT of 0%. Forty-seven known samples were sequenced from a dilution series of positive amplification controls in quadruplicate—HL60 at inputs of 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, and 10,000 pg—and a water-only negative amplification control in triplicate. An additional 48 libraries were sequenced from 46 high-quality gDNA samples, one 2800M sample, and one HL60 sample at 100 pg.[4,5]

For the 44 HL60 samples in the dilution series, total reads per sample ranged from 25,272 to 77,132. Verogen set the sample read count guideline by rounding the mean of 51,202 reads per sample to 50,000. All positions achieved complete coverage and all variants were accurately called, even for samples with read counts below the guideline, such as a 1 pg gDNA sample with only 25,272 total reads. Of the known samples in this study, 10 achieved complete coverage of all positions and total reads that exceeded the guideline. Thirty-eight samples had total

Table 3: Unexpected variant range with AT and IT for control region and whole genome

Metric	Verogen mtDNA Control Region Analysis Method	Verogen mtDNA Whole Genome Analysis Method
Number of positive amplification controls	63	90
Unexpected variant range	0.1–9.7% (4157 bases of 72,954)	0.1%–5.3% (16,203 bases of 1,491,210)
50th percentile for unexpected variants (average)	0.7% (1% standard deviation)	0.7% (0.5% standard deviation)
95th percentile for unexpected variants	2.2%	1.6%
Maximum percentage of unexpected variants	3.9–9.7%	1.9%–5.3%
AT setting	10%	6%
IT setting	10%	6%

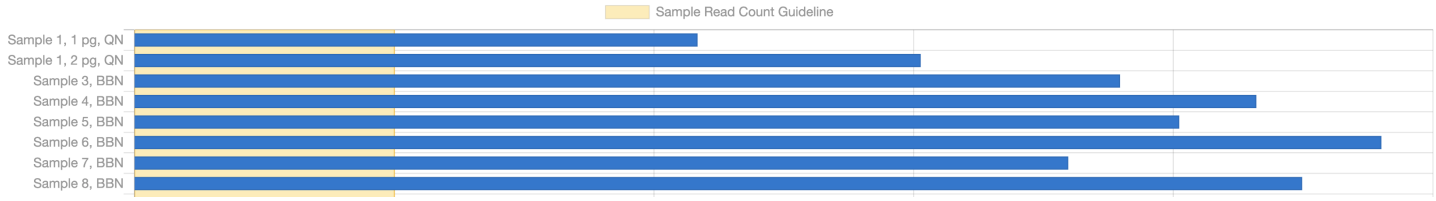

Figure 2: Sample representation for a Verogen mtDNA Control Region Analysis Method compares the read counts for each sample to the sample read count guideline, which is indicated with yellow shading.

reads less than the read count guideline: 26 of these 38 had complete coverage, demonstrating utility for a broad range of sample results. Of the remaining 12 samples, 1–5 positions in the HVII C-stretch were not called in seven samples and approximately 50 positions were not called in five samples. The sample read count guideline is not a threshold, so complete coverage and 100% variant calls in the control region is possible for samples with fewer reads.

Data from the four runs of ForenSeq mtDNA Whole Genome libraries were analyzed in ForenSeq UAS v2.1 with a minimum read count of 45, an AT of 0%, and an IT of 0%. Forty-eight known samples were sequenced from a dilution series of positive amplification controls—HL60 at inputs of 2, 5, 10, 20, 30, 50, 100 pg—and a water-only negative amplification control in duplicate. An additional 95 libraries were sequenced from 93 high-quality gDNA samples, one 2800M sample, and one HL60 sample at 100 pg.[4,5]

For the 42 HL60 samples in the dilution series, total reads per sample ranged from 173,231 to 693,427. Verogen set the sample read count guideline by rounding the mean of 433,329 reads per sample to 400,000. For samples with at least 5 pg gDNA input, over 99% of the mtGenome was covered and detection of all variants was achieved for total reads per sample > 302,175. Of the 95 known samples, 68 achieved complete coverage of all positions and 90 had total reads that exceeded the guideline. Three samples had ≥ 387,127 total reads, while 22 samples displayed > 99% coverage of the mtGenome and three samples displayed > 98% coverage. All hypervariable regions (HVI, HVII, and HVIII) were called correctly in all known samples. Samples with 2 pg gDNA input, where reads per sample were 173,231–277,615, showed ≥ 92% coverage of the mtGenome with all expected variants detected. For samples with < 20 pg gDNA input, pairing the optional second purification with the manual quantification method can increase reads per sample.[6]

Conclusion

After sequencing on the MiSeq FGx System, ForenSeq UAS v2.0 quickly and accurately demulitplexes mtDNA data and calls variants for reliable analysis. Analysis is completed in under one hour with results visualized on an intuitive user interface rich with dynamic, interactive features and easy-to-use tools for guided exploration of sequencing data and meticulous reporting. The default analysis methods provide high-quality results, both for high-level metrics and variant calling performance, and the quality of results is similar across methods. Settings that comprise the default analysis methods are configurable, enhancing flexibility and accessibility for optimum control of your data. Laboratories can take advantage of this user-friendly, forensics-focused software while tailoring it to meet specific needs. Simply copy one of the built-in analysis methods and edit the default settings. Leveraging the capability to create analysis methods grants laboratories instant access to a powerful tool for analyzing NGS data on their own terms.

References

1. Jäger, Anne C., Michelle L. Alvarez, Carey P. Davis, Ernesto Guzmán, Yonmee Han, Lisa Way, Paulina Walichiewicz, et al., "Developmental Validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories," *Forensic Science International: Genetics* 28 (May 2017): 52–70. https://doi.org/10.1016/j.fsigen.2017.01.011.
2. Illumina, *Quality Scores for Next-Generation Sequencing*, October 2011, pub. no. 770-2011-030.
3. Illumina, *Human mtDNA Genome*, February 2016, document # 15037958.
4. Coriell Institute, Camden, NJ
5. Promega, Madison, WI
6. Verogen, *ForenSeq mtDNA Whole Genome Kit Reference Guide*, August 2020, document # VD2020006