Scientific Working Group on DNA Analysis Methods

Addendum to "SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories" to Address Next Generation Sequencing

Contents

Background.....1

Core Elements 2

1. Interpretation.....3

2. Mixture Interpretation.....8

3. Comparison of References & Statistical Weight...12

Glossary.....13

References.....15

Supplementary Tables.....19

Supplementary Examples..... 21

Addendum to "SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories" to Address Next Generation Sequencing

Background

This addendum provides guidelines for the interpretation of DNA typing results from short tandem repeat (STR) data developed via next generation sequencing (NGS). At this time, these guidelines are intended for use by laboratories that will employ binary approaches to interpret sequence-based STR data. Sequence-based STR data are backwards compatible with fragment length-based STR typing. That is, sequence-based STR alleles can be converted to fragment length-based STR alleles. Throughout this document, the word "allele" is used to represent both forms of data, unless prefaced with "sequence-based" or "length-based" to specify one or the other. For demonstration purposes, all examples use an arbitrary analytical threshold of 50 reads and a stochastic threshold of 200 reads. When the interpretation guidance for sequence-based data differs from the guidance for length-based NGS data, this is clarified in the document. This addendum was approved by the SWGDAM membership on April 23, 2019.

The following system has been used to supplement the parent document:

1. If the same numerical heading is present in the addendum and in the parent document, then the text in the addendum is intended to replace the text in the parent document for NGS data interpretation.
2. If the text associated with a numerical heading in the parent document requires additional information specific to NGS data interpretation, the numerical heading is present in the addendum, followed by "In addition to the information in the parent document…", followed by the additional text.
3. Subheadings non-duplicative of the parent document are used in the addendum to allow for inclusion of distinct, NGS-specific information.
4. If the text in the parent document does not apply to NGS, then that numerical designation is included in the addendum along with the text "this section does not apply to NGS data interpretation."
5. If a numerical heading or subheading from the parent document is not addressed in the addendum, then the parent document text should be applied to NGS data interpretation. If capillary electrophoresis (CE)-specific terms appear in the parent document in sections not addressed in this addendum, then the NGS-relevant term should be inferred, as follows:

   Peak = Signal

   Peak Height Ratio (PHR) = Allele Count Ratio (ACR)

   RFU = Read Count

   Refer to the glossary at the end of this document for NGS-specific definitions.

Core Elements

The Core Elements of the parent document remain unchanged for NGS data interpretation with the exception of Core Element III, which does not apply as internal size standards and allelic ladders are not utilized in the NGS process. The remainder of the Core Elements are applicable to NGS data

interpretation as the same requirements, expectations and statistical approaches apply to STR data developed by both CE and NGS.

III. The laboratory shall establish criteria to address locus and allele designation from NGS data.

Section 1. Interpretation

Introduction:

With NGS technologies for human identification, STR typing results are derived through application of analytical software during and after sequencing of DNA libraries, and the demultiplexing of indexed samples. For each sample, software detects the sequence of nucleotides in a DNA strand and translates that information into digital sequence data (reads). The resulting reads are differentiated by locus and compiled, then sequences and associated read counts are reported by the software with descriptors. These descriptors may include length-based allele designation, allele sequence (in nucleotides), signal intensity (measured as read count), and sequence-based allele nomenclature [for more information on nomenclature, see references 1, 2]. Allele designations are assigned by the software based on length and/or sequence information.

To ensure the accuracy of the computer-generated allele assignments, the DNA analyst must verify that established quality criteria for the sequencing run and sample data have been met and that the correct genotyping results were obtained for a known positive control. If the laboratory is interpreting sequence-based data, the correct sequence-based genotype for the positive control must be verified. If the laboratory is interpreting length-based data, the correct length-based genotype must be verified. Additionally, if a sample is amplified using multiple kits and/or platforms that contain redundant loci, the DNA analyst must address the concordance of the genotyping results for the regions sequenced in common.

The results of the analysis controls [i.e., reagent blank(s), positive control(s), negative control(s)] are evaluated. If the reagent blank(s), positive control(s), and negative control(s) yield results that are within their prescribed specifications, the DNA analyst interprets the DNA typing results from each sample.

1.1 Analytical Threshold

The analytical threshold should be validated based on internally derived empirical data. An analytical threshold defines the minimum read count at and above which detected signal can be reliably

distinguished from background noise. Non-reproducible noise may be detected above the analytical threshold. However, usage of an exceedingly high analytical threshold to minimize this sporadic noise signal increases the risk of allelic data loss.

1.1.1 Analytical thresholds may be based on fixed read count values and/or read count percentages (e.g., allele read count divided by total locus read count, and referenced in this document as percentage-based threshold). Analytical thresholds may vary by locus as well.

1.2 Sequencing Run Evaluation: The laboratory must develop criteria to evaluate the quality of the run and the run data. Run quality may be measured by results from positive controls and sequencing standards, for example, or by other defined run metrics/parameters. If run quality metrics are used for this purpose, they should be defined during validation. Metrics used for this purpose may include such parameters as phasing, loading density, cluster density, total reads per sample, total reads per run, forward/reverse strand balance, Q-scores, etc.

1.3.1 The laboratory must establish criteria for evaluation of controls, including but not limited to: reagent blank, positive and negative controls. As applicable, additional controls may be useful for troubleshooting and quality control purposes.

1.4 Locus Designation: The laboratory must have criteria to address locus designations and locus assignment for alleles. A positive control may be used to verify correct locus designations, provided the analytical software has been previously validated for this purpose.

1.4.1 Locus designations must include the range of reported sequence defined according to positions on an identified and relevant human genome reference sequence (e.g., hg 19, GRCh38, etc.).

1.5 Allele Designation: The laboratory must establish criteria for designating alleles according to length-based and/or sequence-based data. The laboratory may designate alleles as numerical values or as sequences in accordance with the guidance of the International Society of Forensic Genetics (ISFG) [1, 2]. SWGDAM will stay abreast of developments regarding standardization of STR sequence nomenclature and provide additional guidance when appropriate. Sequence-based allele designations

must consider the sequence range reported, as ranges may vary based on NGS assay and/or analytical software.

1.5.1 Length-based allele designation is based conceptually on the number of repeat units contained within the core repeat; in practice, it should be based on the fragment size to maximize concordance with CE methods. Sequence-based allele designation is operationally derived from the DNA sequence of the allele per the defined range. This may include the repeat region sequence data only or the repeat region plus additional flanking region sequence data.

1.5.1.1 Analytical software for NGS should have sufficient developmental validation demonstrating concordance in length-based calls between NGS and CE or other platforms.

1.5.1.2 If sequence data will be used for interpretation, the laboratory must establish criteria for reporting alleles of the same length that differ by sequence (isoalleles).

1.5.1.2.1 For backward compatibility to existing length-based STR databases, laboratories should search the length-based allele.

1.5.1.3 If sequence data will be used for interpretation, the laboratory should have a policy for addressing instances where the NGS-developed length-based and sequence-based alleles are discrepant (for example, due to flanking region insertions or deletions that may be captured in the sequence-based allele designations of the analytical software). The discrepancy should be documented and where possible a reason given for the non-concordance.

1.5.2 The laboratory must establish guidelines for the designation of alleles containing an incomplete repeat motif or any reported sequence resulting in other than a whole number allele.

1.5.3 This section does not apply to NGS data interpretation.

1.5.4 Similar to microvariant and off-ladder alleles in CE analysis, laboratories should be aware that previously unobserved sequences and/or unusual motifs may be encountered. When such sequences are recognized, data quality criteria including but not limited to read depth, quality

scores (Q-scores), and forward/reverse strand balance should be considered. Additionally, it may be useful to perform a literature search to determine if the sequence or motif has been reported previously. Resequencing may be useful if the sequence or motif is not found in the literature. Laboratories should also consider if the unusual sequence has been properly designated by the software (length-based and sequence-based designations).

1.6 Non-allelic signal

Because forensic DNA typing characterizes STR loci using PCR and sequencing technologies, some data that result from this analytical process may not represent actual alleles that originate from the sample. It is therefore necessary, before the STR typing results can be used for comparison purposes, to identify any potential non-allelic signal. Non-allelic signals may be PCR products (e.g., stutter and non-specific amplification product) or analytical artifacts as determined by validation studies.

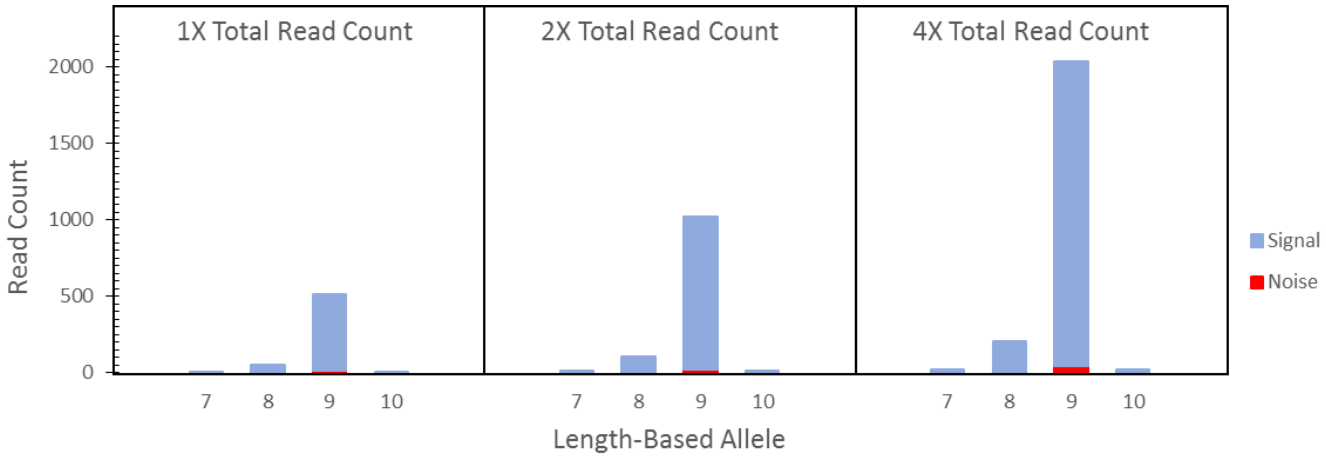
1.6.1 The laboratory must establish criteria based on empirical data (obtained internally and/or externally), and specific to the NGS assay and detection systems used, to address the interpretation of non-allelic signal. These guidelines should address identification of non-allelic signal and the uniform application, across all loci of a DNA profile, of the criteria to identify non-allelic signal.

1.6.1.1 In general, the empirical criteria are based on qualitative (sequence) and/or quantitative (read count) characteristics of signals. As an example, noise signal may be distinguished from allele signal based on the sequence of the artifact and/or its reproducibility. Stutter signals may be characterized based on amplicon length, sequence and read count relative to a parent allele.

1.6.1.2 The laboratory must establish expectations for the proportionality of noise (e.g., amplification background, sequencing errors, sequencing background) relative to signal, and data interpretation should be based on the limits established by validation. With higher read count, increased total reads are detected for the true allele as well as noise. As read count signal may vary across loci, and as noise may be proportional to signal, locus specific analytical thresholds may be required.

The following table and figure demonstrate an example of noise in direct proportion to signal, for a single allele in a single-source sample with increasing total read counts. The designation "Signal" represents allele signal and stutter signal, or the read count of the most common sequence (or consensus) for each length-based allele. The designation "Noise" is the cumulative number of all other sequences for each length-based allele. In this example, noise is consistently 2% of signal, regardless of signal intensity (read count). "ND" indicates not detected.

	1X Total Read Count		2X Total Read Count		4X Total Read Count	
Length-Based Allele or Stutter	Signal	Noise	Signal	Noise	Signal	Noise
7	5	ND	10	ND	20	ND
8	50	1	100	2	200	4
9	500	10	1000	20	2000	40
10	5	ND	10	ND	20	ND



1.7.1 The guidance from the parent document applies. However, laboratories may apply locus specific stochastic threshold values. For examples of how fixed read count thresholds and percentage based thresholds may result in different allelic dropout designations, see Supplementary Examples E1 through E3.

1.7.1.1 NGS-specific measures to enhance sensitivity may also include kit-based procedural modifications (e.g., changes to sample normalization steps, DNA sample concentration, etc.), sequencing step modifications (e.g., reduction in the number of samples pooled, concentration of the sequence pool, etc.), and reduction of analytical/stochastic thresholds.

1.9 In addition to the information in the parent document, alleles of the same length may be differentiated by sequence level information (isoalleles), and these sequence-based alleles may be useful for contributor number estimations.

Section 2: Mixture Interpretation Overview and Strategies

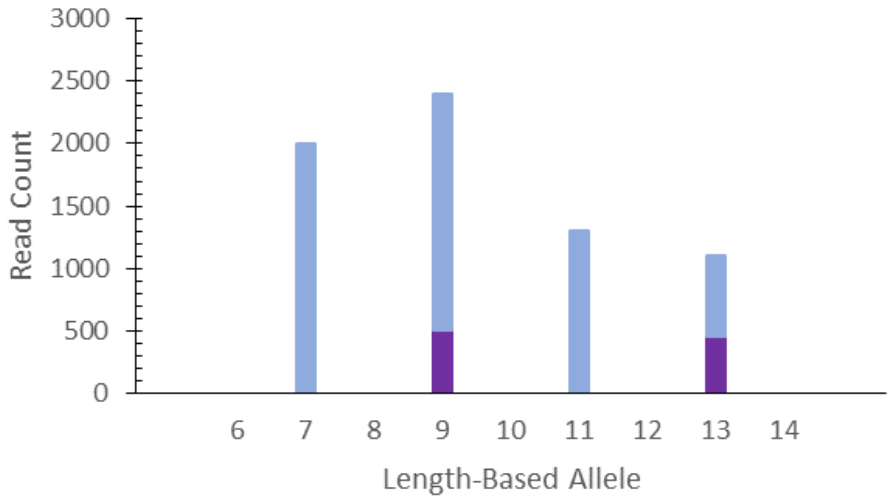
Introduction:

Generally the principles that apply to STR mixture interpretation strategies for CE data described in the parent document also apply to sequence-based data. When interpretation of the sequence data is limited to the length-based allele, then the same principles apply. This section addresses differences a laboratory may encounter when evaluating mixture profiles at the sequence level. In these situations, more alleles may be resolved and in some cases stutter may be more readily distinguished from alleles of minor contributors. Sequence specific examples that correlate to CE examples are included.

2.1 In addition to the text in the parent document, if assumptions are made that rely on sequence information, such as number of contributors being based on isoalleles, then the criteria shall be supported by the data and shall be defined and documented. In the following example table and figure, this mixture appears as a minimum two person mixture by length, but a minimum of three contributors is indicated by sequence. *To highlight this difference, and for example purposes only, stutter alleles are not being considered and are therefore not represented in the figure.*

Length-Based Allele	Sequence	Read Count
7	[TCTA]7	2000
9	[TCTA]9	1900
9	[TCTA][TCTG][TCTA]7	500
11	[TCTA]11	1300

13	[TCTA]13	650
13	[TCTA][TCTG][TCTA]11	450

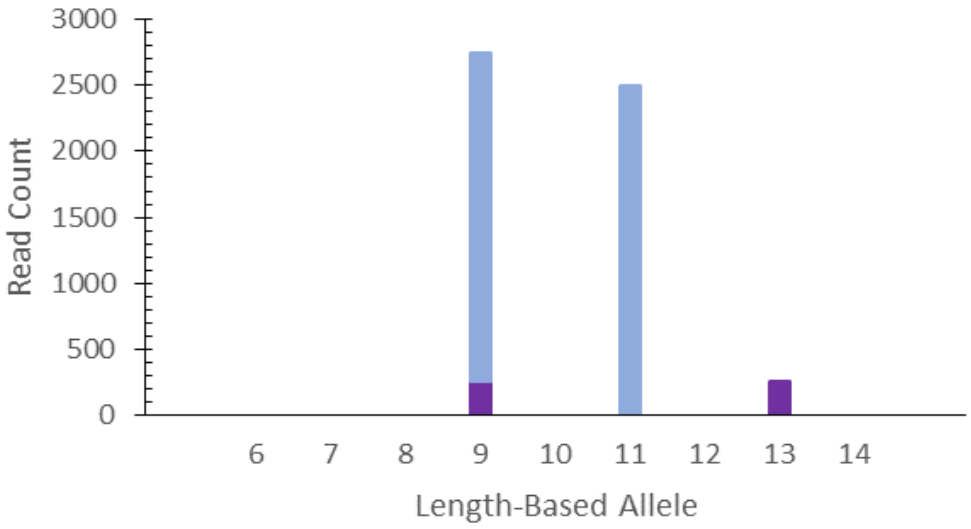


2.2 In addition to the text in the parent document, for sequence-based NGS data, pair-wise comparison of all potential genotypic combinations should consider sequence data and the ACR of all sequence combinations, including isoalleles.

2.3 In addition to the text in the parent document, sequence-based interpretation of mixtures must be based on sequence-based mixture studies, including known contributors with alleles that overlap by length but are differentiated by sequence (isoalleles).

2.4.1 In addition to the text in the parent document, for sequence-based NGS data, the sequence data may differentiate length-based shared alleles (isoalleles), and this may allow determination of major and minor sequence alleles which are the same length. In the following example table and figure, two different sequences are observed for the length-based allele 9. Based on the read counts and ACR expectations, one isoallele can be attributed to the major contributor and the other to the minor contributor. *For purposes of this example, stutter alleles are not considered, and are therefore not represented in the table or figure.*

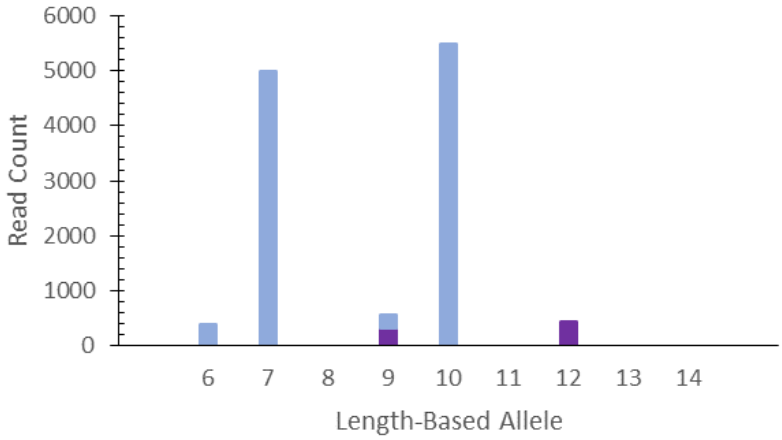
Length-Based Allele	Sequence	Read Count
9	[TCTA]9	2500
9	[TCTA][TCTG][TCTA]7	250
11	[TCTA]11	2500
13	[TCTA]13	250



2.6.3.1 At some loci, sequence information may allow differentiation of stutter and same-length minor contributor alleles. In order to use this information, it is important to characterize sequence-based stutter expectations during validation. In the following example table and figure, two different sequences are observed for the length-based allele 9. Based on the sequences, read counts, ACR and stutter expectations, one isoallele is consistent with stutter of the major contributor and the other isoallele is attributable to the minor contributor.

Length-Based Allele	Sequence	Read Count
6	[TCTA]6	380
7	[TCTA]7	5000
9	[TCTA]9	300

9	[TCTA][TCTG][TCTA]7	250
10	[TCTA][TCTG][TCTA]8	5500
12	[TCTA]12	425



The stutter threshold at this locus is 10%. The profile is assumed to be a two-person mixture.

Major contributor sequence-based genotype determined as:

Length-Based Allele	Sequence	Read Count	ACR
7	[TCTA]7	5000	91%
10	[TCTA][TCTG][TCTA]8	5500	

Minor contributor sequence-based genotype determined as:

Length-Based Allele	Sequence	Read Count	ACR
9	[TCTA]9	300	71%
12	[TCTA]12	425	

Stutter alleles that do not need to be incorporated into statistic:

Length-Based Allele	Sequence	Read Count	Stutter

6	[TCTA]6	380	7.6%
9	[TCTA][TCTG][TCTA]7	250	4.5%

Section 3. Comparison of References and Statistical Weight of Probative Inclusions

Introduction:

Generally, the principles that apply to CE based comparisons of reference profiles to evidence profiles also apply to profiles derived from sequence data. When comparisons of profiles derived from sequence data is limited to the length-based allele designations, then the same principles apply as CE based comparisons. This section addresses differences a laboratory may encounter when comparing profiles at the sequence level.

3.2.4.1 Reported Range: If comparisons are conducted using sequence data, then sequence-based allele frequency data shall be used for statistical analysis. The sequence-based allele frequency data utilized should consist of the same sequence range (see section 1.4.1) as the allele being compared. It is possible to truncate a sequence-based allele frequency data set if it represents a larger range than the laboratory's desired reporting range. However, if the sequence-based allele frequency data set is of a smaller range than the laboratory's desired reporting range, the laboratory's reported range must be truncated to match the frequency data (see Supplementary Tables S1-S3 for more information).

3.4.2.2.1 For statistical calculations of sequence-based single source genotypes, isoalleles are considered heterozygote genotypes and the formula is 2pq.

3.4.2.2.1.1 Laboratories performing length-based analysis should have policies for interpretation and statistical analysis when isoalleles are encountered at a locus. This is recommended to avoid concealing potentially exculpatory information.

3.4.2.3.1 Currently, guidance does not exist regarding theta values for sequence-based data; therefore, the existing NRC II guidance should be followed (NRC II 4.4a, where typically $\theta = 0.01$ for most U.S. groups or 0.03 for some isolated populations).

Glossary

Allele Count Ratio: the relative ratio, or intralocus balance, of two alleles at a given locus. This is commonly expressed as a percentage, and is generally calculated for a given locus by dividing the count of the allele with the lower signal value by the count of the allele with the higher signal value. Allele Count Ratios may also be referred to as allele coverage ratios or read count ratios. In all cases, the ratios are analogous to CE-based Peak Height Ratios.

Cluster Density: the density of clonal clusters on a sequencing flow cell. Optimal cluster density maximizes sequencing performance in terms of data quality and total sequence data output. The term applies to those NGS chemistries that employ glass flow cell sequencing technology.

Demultiplexing: the bioinformatic sorting of samples that have been simultaneously analyzed on a sequencer using the indices (also known as barcodes) associated with individual samples.

Flanking Region: DNA sequence located between PCR primer binding sites and the core repeat region of the short tandem repeat (STR) amplicon. The extent of this region varies according to PCR primer placement and is therefore expected to vary by kit. The amount of flanking region reported may further vary by bioinformatic program design, data quality, allele length, etc.

Flanking Region Polymorphisms: nucleotide variants in regions of short tandem repeat (STR) amplicons that are 5′ or 3′ adjacent to the core repeat region.

Forward/Reverse Balance (or strand balance): a measure of the distribution of forward and reverse reads aligned at each nucleotide position. A relatively even distribution of reads from both strands provides a measure of support for the nucleotide call. While strand imbalance or bias can, under certain circumstances, indicate reduced support for the affected nucleotide calls, in some assays, and in particular genomic regions, only one strand is routinely sequenced. As forward/reverse balance can be used as a quality metric, expectations for strand balance should be established during validation.

Index: a molecular barcode, typically consisting of DNA sequence(s) covalently bound to genetic material from a sequencing library that provides sample identity and allows for multiple samples to be analyzed simultaneously.

Isoallele: alleles that are identical by length-based analysis, but different by sequence-based analysis. Also known as isometric alleles.

Library (or sequencing library): a work product consisting of genetic material prepared for analysis on a next generation sequencing instrument.

Loading Density: the percentage of wells successfully loaded across the physical surface of a sequencing chip. Higher values reflect greater coverage or loading of the chip. The term applies to those NGS chemistries that employ sequencing chips with wells.

Next Generation Sequencing or NGS: (also known as massively parallel sequencing, deep sequencing and high throughput sequencing) is a term used to describe modern sequencing technologies other than Sanger sequencing.

Phasing: the rate at which single molecules within a sequencing cluster become out of sync with each other during the sequencing process. Individual strands may be a base (or more) ahead of the majority of the cluster (pre-phasing), or they may lag behind the majority of the cluster (phasing). Together, pre-phasing and phasing offer a measure of the performance of the chemistry/sequencing run, with higher values indicating lower signal to noise ratios (i.e., more noise from phasing). The term applies to sequencing by synthesis chemistries.

Quality score (or Q-score): a metric that is used to indicate whether a base has been called correctly. Specifically, it is the probability that a given base has been miscalled. Mathematically, it is defined as -10log10(e), where e is the estimated probability of the base call being incorrect. Higher Q scores indicate a lower probability of base-calling error, while lower Q scores indicate a higher probability of error.

Reads: the raw sequence data produced by a sequencing instrument, generated as a result of detection of the sequence of nucleotides in a DNA strand and the translation of that information into digital sequence data.

Read count: the absolute number of reads of a given sequence. Read count (X) is a measurement of signal and, as such, is analogous to relative fluorescent units (RFUs) in capillary electrophoresis based analysis.

References

[1] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmao, D.R. Hares, J.A. Irwin, J.L. King, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C.V. Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54-63.

[2] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson, "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, Forensic Sci. Int. Genet. 34 (2018) 162-169.

Additional Resources

The following articles, though not directly referenced in this document, provide relevant background information that may be helpful to laboratories performing next generation sequencing.

Assay Information

A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J.K. Pond, J. Varlaro, K.M. Stephens, C.L. Holt, Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories, Forensic Sci. Int. Genet. 28 (2017) 52–70.

R.S. Just, L.I. Moreno, J.B. Smerick, J.A. Irwin, Performance and Concordance of the ForenSeq System for Autosomal and Y Chromosome Short Tandem Repeat Sequencing of Reference-type Specimens, Forensic Sci. Int. Genet. 28 (2017) 1-9.

P. Fattorini, C. Previdere, I. Carboni, G. Marrubini, S. Sorcaburu-Cigliero, P. Grignani, B. Bertoglio, P. Vatta, U. Ricci, Performance of the ForenSeq™ DNA Signature Prep Kit on Highly Degraded Samples, Electrophoresis 00 (2017) 1-12.

L.I. Moreno, M.B. Galusha, R.S. Just, A closer look at Verogen's Forenseq™ DNA Signature Prep kit autosomal and Y-STR data for streamlined analysis of routine reference samples, Electrophoresis 39 (2018) 2685-2693.

V. Sharma, H.Y. Chow, D. Siegel, E. Wurmbach, Qualitative and Quantitative Assessment of Illumina's Forensic STR and SNP kits on MiSeq FGx, PLoS ONE 12 (2017) e0187932.

C. Hollard, L. Ausset, Y. Chantrel, S. Jullien, M. Clot, M. Faivre, E. Suzanne, L. Pène, F. Laurent, Automation and developmental validation of the ForenSeq™ DNA Signature Preparation kit for high-throughput analysis in forensic laboratories, Forensic Sci. Int. Genet. 40 (2019) 37-45.

K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F.J. Laros, P. de Knijff, Massively parallel sequencing of short tandem repeats—Population data and mixture analysis results for the PowerSeq system, Forensic Sci. Int. Genet. 24 (2016) 86-96.

E.A. Montano, J.M. Bush, A.M. Garver, M.M. Larijani, S.M. Wiechman, C.H. Baker, M.R. Wilson, R.A. Guerrieri, E.A. Benzinger, D.N. Gehres, M.L. Dickens, Optimization of the Promega PowerSeq™ Auto/Y system for efficient integration within a forensic DNA laboratory, Forensic Sci. Int. Genet. 32 (2018) 26–32.

X. Zeng, J. King, S. Hermanson, J. Patel, D. Storts, B. Budowle, An evaluation of the PowerSeq™ Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing, Forensic Sci. Int. Genet. 19 (2015) 172-179.

P. Müller, A. Alonso, P.A. Barrio, B. Berger, M. Bodner, P. Martin, W. Parson, Systematic evaluation of the early access Applied Biosystems Precision ID Globalfiler Mixture ID and Globalfiler NGS STR panels for the Ion S5 system. Forensic Sci. Int. Genet. 36 (2019) 95-103.

Z. Wang, D. Zhou, H. Wang, Z. Jia, J. Liu, X. Qian, C. Li, Y. Hou, Massively parallel sequencing of 32 forensic markers using the Precision ID GlobalFiler™ NGS STR Panel and the Ion PGM™ System, Forensic Sci. Int. Genet. 31 (2017) 126-134.

Bioinformatics

B. Young, J. King, B. Budowle, L. Armogida, A technique for setting analytical thresholds in massively parallel sequencing-based forensic DNA analysis, PLoS ONE 12(5):e0178005.

J.L. King, F.R Wendt, J. Sun, B. Budowle, STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems, Forensic Sci. Int. Genet. 29 (2017) 21-28.

J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. Laros, FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognize and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27-40.

S.L. Friis, A. Buchard, E. Rockenbauer, C. Borsting, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci. Int. Genet. 21 (2016) 68-75.

NGS Data quality metrics

N. Aziz, Q. Zhao, L. Bry, D. Driscoll, B. Funke, J. Gibson, W. Grody, M. Hegde, G. Hoeltge, D. Leonard, J. Merker, R. Nagarajan, L. Palicki, R. Robetorye, I. Schrijver, K. Weck, K. Voelkerding, College of American Pathologists' laboratory standards for next-generation sequencing clinical tests, Arch Pathol Lab Med 139 (2015) 481-493.

A.S. Gargis, L. Kalman, M.W. Berry, D.P. Bick, D.P. Dimmock, T. Hambuch,… I.M. Lubin, Assuring the quality of next-generation sequencing in clinical laboratory practice, Nature Biotechnology 30 (2012) 1033-1036.

G. Pont-Kingdon, F. Gedge, W. Wooderchak-Donahue, I. Schrijver, K. Weck, J. Kant, D. Oglesbee, P. Bayrak-Toydemir, E. Lyon, Design and Analytical Validation of Clinical DNA Sequencing Assays. Arch Pathol Lab Med, 136 (2012) 41‑46.

H. Rehm, S. Bale, P. Bayrak-Toydemir, J. Berg, K. Brown, J. Deignan, M. Friez, B. Funke, M. Hegde, E. Lyon; for the Working Group of the American College of Medical Genetics and Genomics Laboratory

Quality Assurance Committee, ACMG clinical laboratory standards for next-generation sequencing, Genet Med 15 (2013) 733-747.

Population Data

K. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. Population Data for 27 Autosomal STR Loci, Forensic Sci. Int. Genet. (2018) 37:106-115.

N. Novroski, J. King, J. Cihlar, L. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Sci. Int. Genet. 25 (2016) 214-226.

L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D. Syndercombe Court, Concordance of the ForenSeq system and characterization of sequence-specific autosomal STR alleles across two major population groups, Forensic Sci. Int. Genet. 334 (2018) 57-61.

STR Sequence Information

K. Gettings, L. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, P.M. Vallone, STRSeq: A Catalogue of Sequence Diversity at Human Identification Short Tandem Repeat Loci, Forensic Sci. Int. Genet. 31 (2017) 111-117.

M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97-102.

A.O. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int. Genet. 26 (2017) 58-65.

Supplementary Tables S1-S3

Sequence-based population allele frequency data used for interpretation should consist of the same sequence range as the allele(s) being compared. The following example describes how a sequence-based allele frequency data set based on a larger range than the laboratory's desired reporting range could be truncated for use.

In the following example, the first table represents a published sequence-based allele frequency data set for one locus and one population. The published data contains 14 bases of 5′ flanking region and 36 bases of 3′ flanking region. Polymorphisms which differentiate these sequences from the Genome Reference Consortium Human genome build 38 (GRCh38) are highlighted in yellow.

Supplementary Table S1

		Full Sequence		
Allele	Frequency	5' Flank	Repeat Region	3' Flank
5	0.0005	CAGACAGACAGGTG	GATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC
8	0.0212	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
9	0.1134	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC
9	0.0483	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
9	0.0005	CAGACAGACAGGTG	GGTAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
9	0.0005	CAGAGAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
10	0.0772	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC
10	0.0304	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
10	0.0005	CAGACAGACAGGTG	GATAGATAGATAGATAGATACAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
11	0.2698	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
11	0.0174	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC
11	0.0043	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGACAC
12	0.2539	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
12	0.0024	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC
12	0.0005	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGACAC
13	0.1371	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
14	0.0212	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
14	0.0005	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC
15	0.0005	CAGACAGACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC
	1			

A laboratory would like to use this data set to generate statistics; however, the laboratory is only including 8 bases of 5′ flanking region and 20 bases of 3′ flanking region in their analysis. Therefore, they must first truncate the published data set to match their reported range, as follows (hyphens indicate bases removed from the previous table, and are included for demonstration purposes only):

Supplementary Table S2

		Full Sequence		
Allele	Frequency	5' Flank	Repeat Region	3' Flank
5	0.0005	------GACAGGTG	GATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
8	0.0212	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
9	0.1134	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
9	0.0483	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
9	0.0005	------GACAGGTG	GGTAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
9	0.0005	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
10	0.0772	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
10	0.0304	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
10	0.0005	------GACAGGTG	GATAGATAGATAGATAGACAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
11	0.2698	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
11	0.0174	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
11	0.0043	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
12	0.2539	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
12	0.0024	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
12	0.0005	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
13	0.1371	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
14	0.0212	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
14	0.0005	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
15	0.0005	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
	1			

And then collapse/combine the frequencies associated with sequences which are no longer differentiated, e.g. the second and fourth "9" sequences, the first and third "11" sequences, and the first and third "12" sequences. This yields the following frequency data set that the laboratory will use for statistical analysis:

Supplementary Table S3

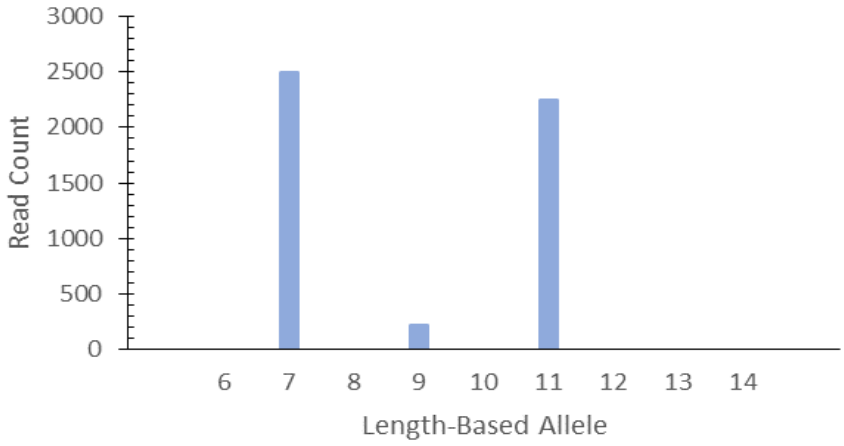
		Full Sequence		
Allele	Frequency	5' Flank	Repeat Region	3' Flank
5	0.0005	------GACAGGTG	GATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
8	0.0212	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
9	0.1134	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
9	0.0488	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
9	0.0005	------GACAGGTG	GGTAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
10	0.0772	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
10	0.0304	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
10	0.0005	------GACAGGTG	GATAGATAGATAGATAGACAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
11	0.2741	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
11	0.0174	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
12	0.2544	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
12	0.0024	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
13	0.1371	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
14	0.0212	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
14	0.0005	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAACCAGA----------
15	0.0005	------GACAGGTG	GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA	TCATTGAAAGACAAAACAGA----------
	1			

Supplementary Examples E1-E3

Laboratories should be aware of the differences between percentage-based and fixed read count analytical and stochastic thresholds*. For example, percentage-based and fixed read count stochastic thresholds may yield different allelic dropout designations in mixtures with high-level major and low-level minor contributors since the percentage-based stochastic threshold will primarily depend on the read count of the major contributor(s). The following two scenarios demonstrate how a 10% difference in the major contributor read counts results in different outcomes for equivalent minor allele read count under a percentage-based threshold, while the minor allele is treated independently and consistently under a fixed read count value threshold.

Supplementary Example E1.

Length-Based Allele	Sequence	Read Count
7	[TCTA]7	2500
9	[TCTA]9	230
11	[TCTA]11	2250



* The percentage-based stochastic threshold discussed in these examples does not necessarily correspond to the percentage-based "interpretation threshold" implemented in some commercially available data analysis packages. Laboratories employing an "interpretation threshold" should fully understand the parameter and how it relates to analytical and stochastic thresholds.

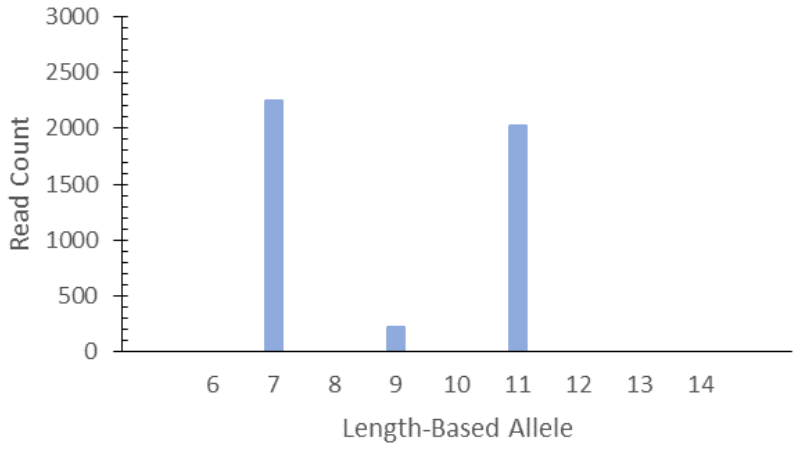
Based on the overall profile, this is determined to be a two-person mixture.

Scenario 1: The laboratory uses a fixed read count value analytical threshold of 50X and stochastic threshold of 200X. **The minor allele is above the stochastic threshold** and the possibility of dropout need not be considered.

Scenario 2: The laboratory uses a percentage-based analytical threshold of 1%, with a minimum value of 50X, and a percentage-based stochastic threshold of 5%, with a minimum value of 200X. The locus read count results in an analytical threshold of 50X (based on 1% of the total locus read count) and a stochastic threshold of 249X (based on 5% of the total read count). **Due to the high read count of the major contributor, the minor allele is below the stochastic threshold** and the possibility of dropout must be considered.

Supplementary Example E2.

Length-Based Allele	Sequence	Read Count
7	[TCTA]7	2250
9	[TCTA]9	230
11	[TCTA]11	2025



Based on the overall profile, this is determined to be a two-person mixture.

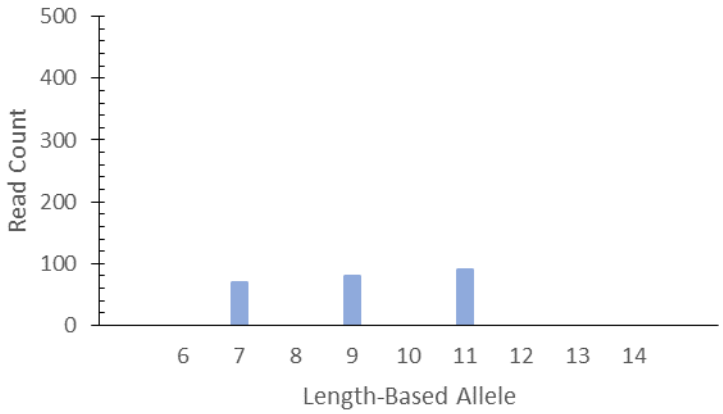
Scenario 1: The laboratory uses a fixed read count value analytical threshold of 50X and stochastic threshold of 200X. **The minor allele is above the stochastic threshold** and the possibility of dropout need not be considered.

Scenario 2: The laboratory uses a percentage-based analytical threshold of 1%, with a minimum value of 50X, and a percentage-based stochastic threshold of 5%, with a minimum value of 200X. The locus read count results in an analytical threshold of 50X (45X based on 1% of the total locus read count; minimum value of 50X) and a stochastic threshold of 225X (based on 5% of the total read count). **The minor allele is above the stochastic threshold** and the possibility of dropout need not be considered, despite the minor contributor read count being the same as in Example E1.

Supplementary Example E3.

As shown in the following example, in the case of a low-level mixture, the percentage-based stochastic threshold that is validated alongside a minimum read count threshold will likely perform similarly to a fixed read count value stochastic threshold. This is because the validated minimum value will typically be higher than the percentage-based value in low level mixtures. However, consideration of all possible scenarios is beyond the scope of this document and it is the laboratory's responsibility to fully consider and provide criteria for the implementation of a percentage-based threshold.

Length-Based Allele	Sequence	Read Count
7	[TCTA]7	70
9	[TCTA]9	80
11	[TCTA]11	90



Based on the overall profile, this is determined to be a two-person mixture.

Scenario 1: The laboratory uses a fixed read count value analytical threshold of 50X and stochastic threshold of 200X. All three alleles are below the stochastic threshold; therefore, the possibility of dropout must be considered.

Scenario 2: The laboratory uses a percentage-based analytical threshold of 1%, with a minimum value of 50X, and a percentage-based stochastic threshold of 5%, with a minimum value of 200X. Because the minimum value stochastic threshold is higher than the percentage-based value (70+80+90=240, then 240*0.05=12X), the possibility of dropout must be considered.