APPLIED GENETICS
National Institute of Standards and Technology
U.S. Department of Commerce

THE USE OF STATISITCAL ANALYSIS TO ASSESS NOISE AND ZYGOSITY IN TARGETED SEQUENCING OF FORENSIC STR MARKERS

Poster # 58 at:
29th Int Symposium on Human Identification (ISHI)
Phoenix, Arizona (September 24-27, 2018)

Email: sarah.riman@nist.gov
Phone: (301) 975-4162
Poster available for download from STRBase:
http://strbase.nist.gov/NISTpub.htm#Presentations

Sarah Riman[1], Hari Iyer[2], Lisa Borsuk[1], Peter M. Vallone[1]
[1] U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA
[2] U.S. National Institute of Standards and Technology, Statistical Design, Analysis, and Modeling Group, Gaithersburg, MD 20899-8314, USA

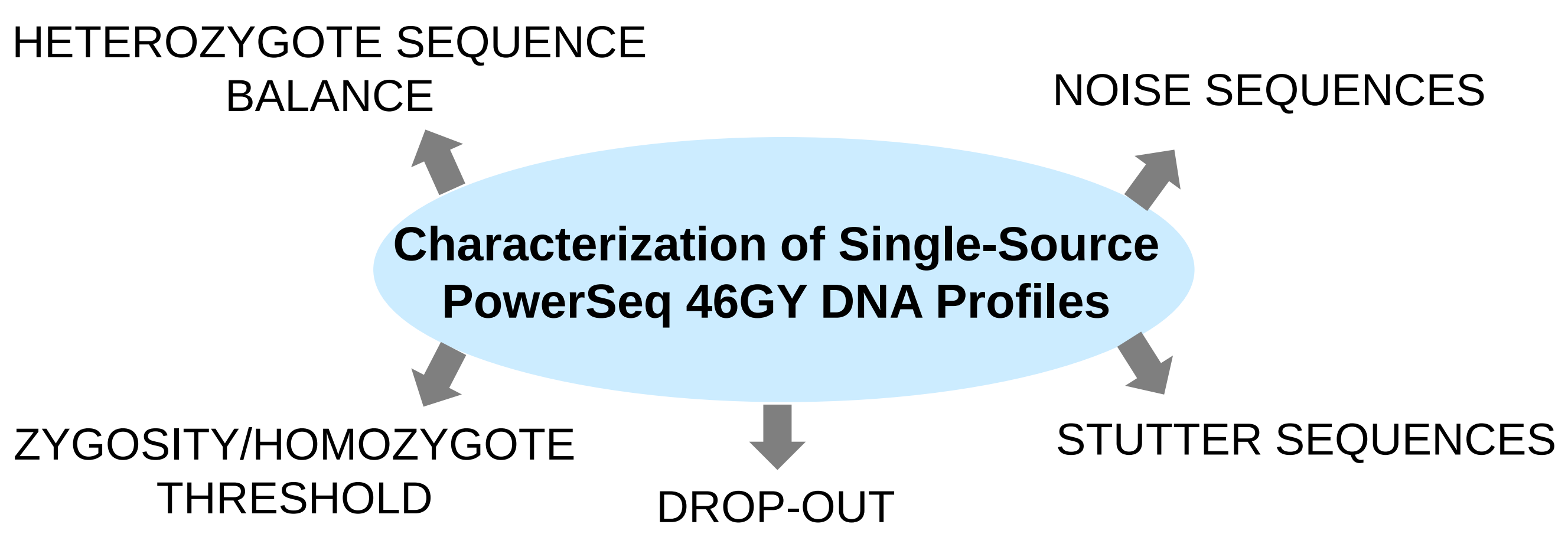
Funding: NIST Special Programs Office: Forensic DNA, FBI Biometrics Center of Excellence: Forensic DNA Typing as a Biometric tool.

Disclaimer: Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce or the Department of Justice. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.
All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

Abstract

The sequencing of STR markers provides additional information due to the underlying sequence variation that is typically masked by traditional fragment-based genotyping. The interpretation of STR profiles generated by targeted sequencing methods are susceptible to the same factors as profiles generated using capillary electrophoresis. These factors include signal noise, stutter artifacts, heterozygote imbalance, and allelic drop-out/in. Our goal is to characterize and understand how these behave in targeted sequence datasets. Here, we developed a framework using statistical tools to systematically interpret and understand the characteristics of single-source DNA profiles generated by targeted sequencing. Data were generated from sensitivity studies using known single-source samples amplified with the PowerSeq 46GY System Prototype with varying DNA target masses ranging from 15 pg to 500 pg. The STR loci were sequenced on the Illumina MiSeq platform and raw FASTQ data files were analyzed in STRait Razor [1] without applying any thresholds (i.e. at a coverage ≥ 1). Boxplots were then used to visualize and compare the distribution of the true allelic, back stutter, and noise sequences. Histograms were developed to visualize the distribution of the heterozygote sequence balance. Laboratories evaluating thresholds to interpret single-source NGS profiles should evaluate the tradeoff between true positives (true allelic sequences) and false positives (non-specific sequences) as well as the risks of misidentifying heterozygous and homozygous genotypes. This can be accomplished by constructing the Receiver Operating Characteristic (ROC) plots. All the data were analyzed globally (all DNA quantities combined), as well as investigated per DNA quantity and per locus. Analyses presented can be applied to sequence data generated by similar targeted sequence multiplexes and/or NGS platforms.

1 Aim of the Study



Characterization of Single-Source PowerSeq 46GY DNA Profiles

HETEROZYGOTE SEQUENCE BALANCE

NOISE SEQUENCES

ZYGOSITY/HOMOZYGOTE THRESHOLD

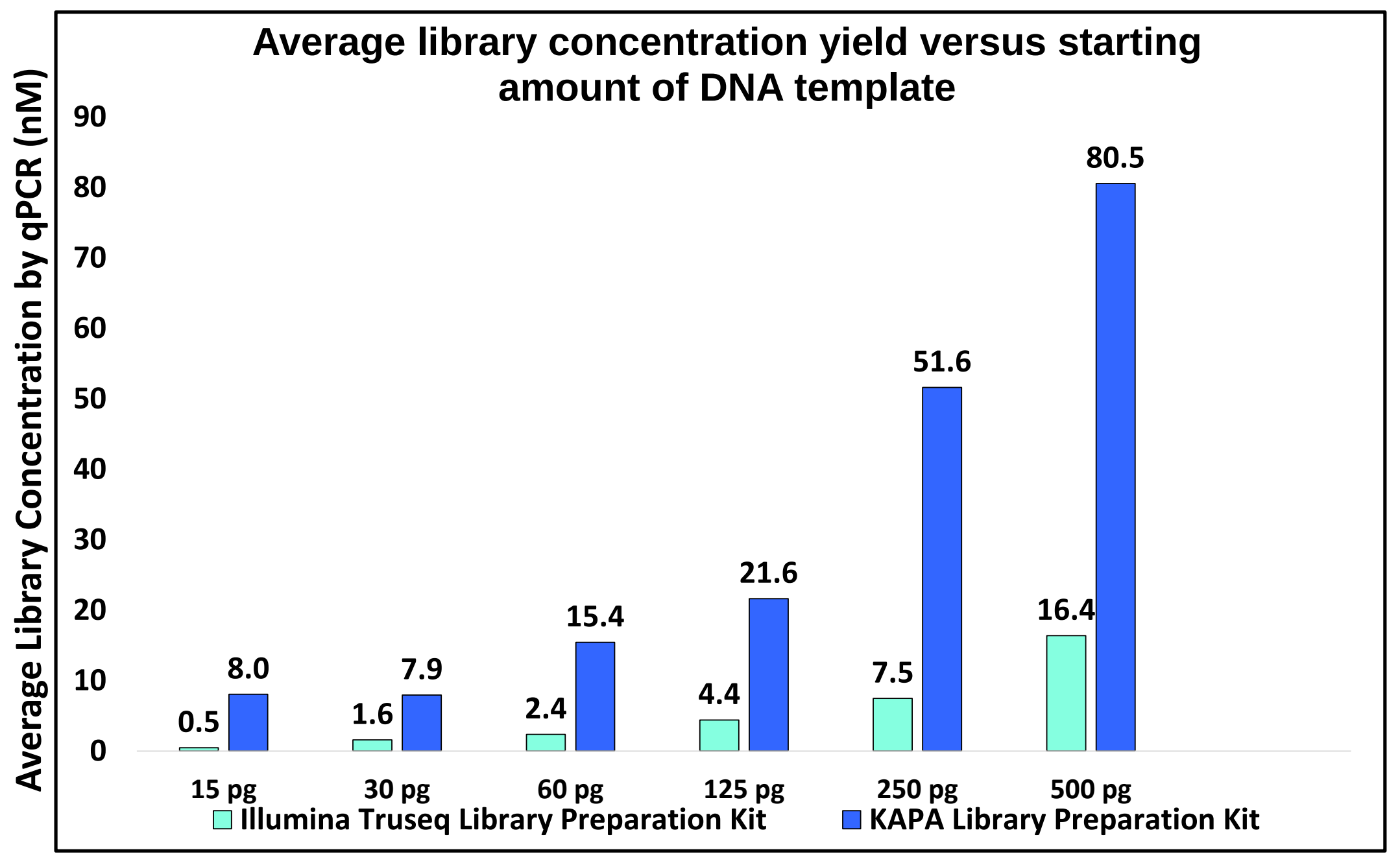
DROP-OUT

STUTTER SEQUENCES

2 Materials and Methods

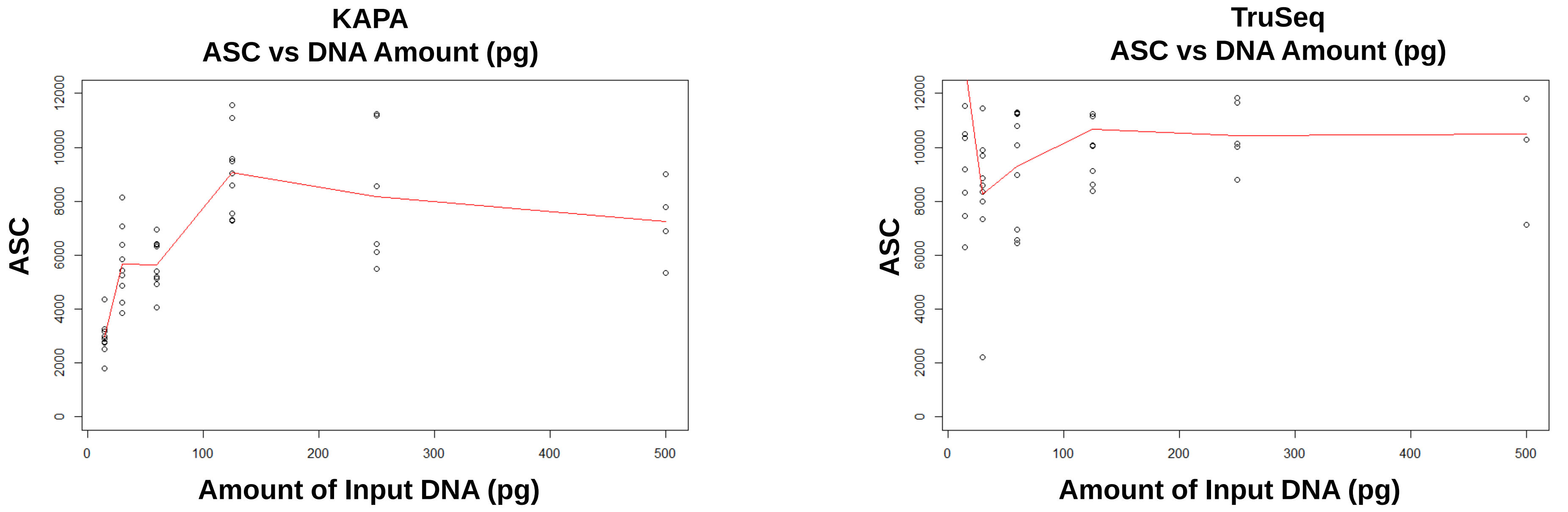
- STR loci were amplified from three unique DNA extracts in triplicate using varying amounts of DNA template (15, 30, 60, 125, 250, and 500 pg), with the PowerSeq 46GY System Prototype (Promega). In this study, the analysis focused solely on the 22 autosomal STR loci, AMEL, and DYS391).
- Purified PCR products were subjected to DNA library preparation using two kits according to the each manufacturers recommended protocol: TruSeq DNA PCR-free (Illumina) or KAPA Hyper Prep (KAPA Biosystems).
- The libraries were normalized, pooled, and sequenced on an Illumina MiSeq instrument with the v3 chemistry.
- STR sequence regions were extracted from the FASTQ files using STRait Razor [1].
- A minimum depth of <u>**coverage of 1X**</u> was used for the data analysis.
- Analysis of the data was performed in *R*.

3 Impact of Library Construction Kits on Library Yield



- The yield of PCR products successfully converted to adapter-ligated libraries was observed to be higher with the KAPA kit as compared to the TruSeq kit.
- ***Note:** the higher yield observed with the KAPA kit did not exhibit any significant difference on correct allele calling, zygosity, and heterozygote sequence balance when compared to the TruSeq kit.*

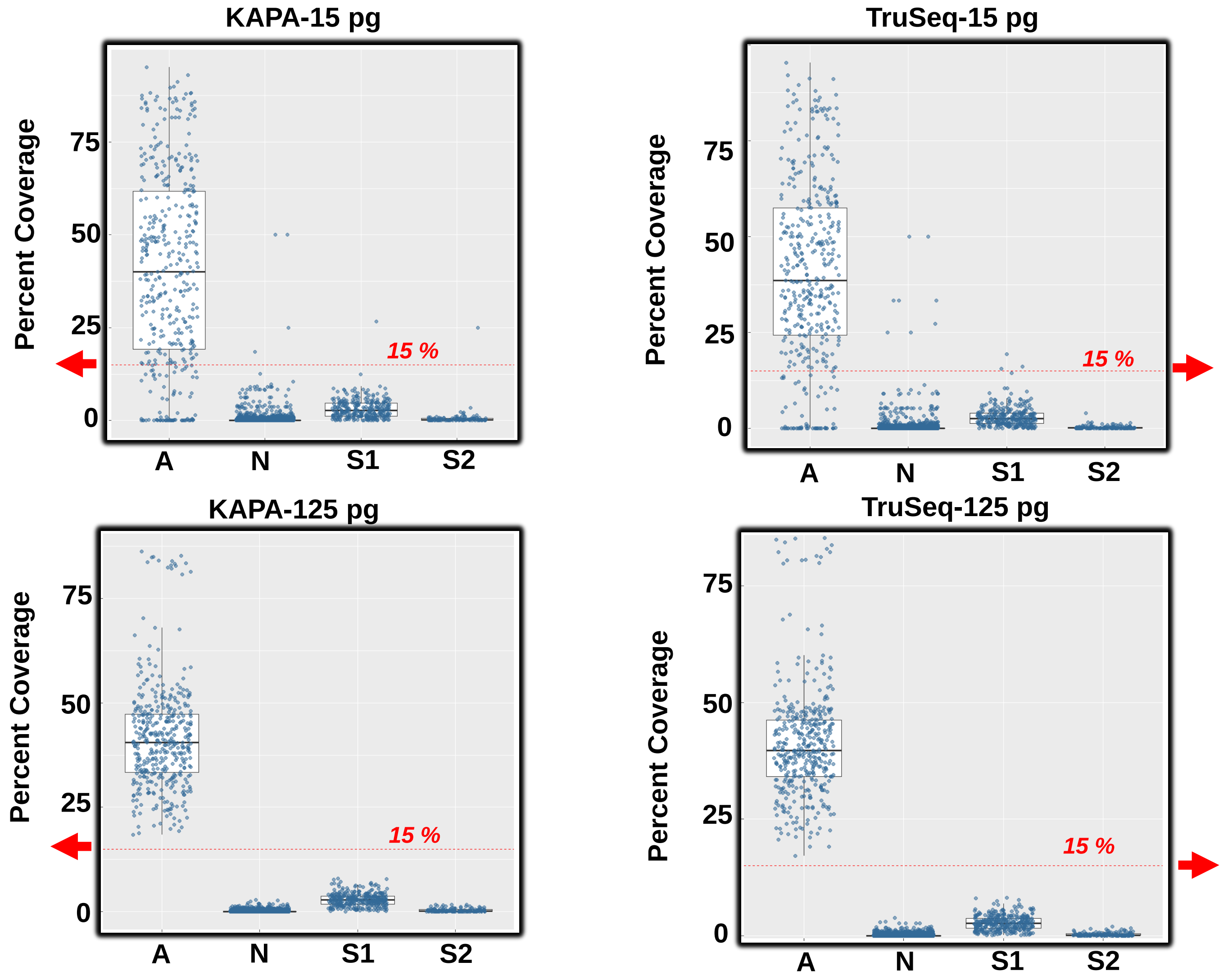
4 Average Sequence Coverage (ASC) of Allelic Sequences versus DNA Template Amount



$$ASC_{Profile} = \frac{\sum \text{coverage of allelic sequences}}{\text{total number of loci}}$$

- ***The average sequence coverage of known alleles did not show a strong correlation (R < 0.4) to the amount of input DNA present at the start of the amplification for either of the library kits.***
- This observation can be attributed to library normalization, a procedural step performed prior to sequencing to ensure that libraries are added in equimolar amounts to the sequencing pool.

5 Impact of DNA Template Amount on the Distribution of Known Allele, Stutter, and Noise Sequences



KAPA-15 pg: At a percent coverage of 15%:
- ***339 sequences are called***
  - *333 can be attributed to **A***
  - *4 can be attributed to **N***
  - *1 can be attributed to **S1***
  - *1 can be attributed to **S2***

TruSeq-15 pg: At a percent coverage of 15%:
- ***361 sequences are called***
  - *350 can be attributed to **A***
  - *8 can be attributed to **N***
  - *3 can be attributed to **S1***
  - *0 can be attributed to **S2***

KAPA-125 pg: At a percent coverage of 15%:
- ***411 sequences are called***
  - *411 can be attributed to **A***
  - *0 can be attributed to **N***
  - *0 can be attributed to **S1***
  - *0 can be attributed to **S2***

TruSeq-125 pg: At a percent coverage of 15%:
- ***411 sequences are called***
  - *411 can be attributed to **A***
  - *0 can be attributed to **N***
  - *0 can be attributed to **S1***
  - *0 can be attributed to **S2***

A = Known allele sequences

N = Noise sequences

S1 = Back stutter of the longest uninterrupted stretch of the basic repeat motifs within an allelic sequence [2,3]

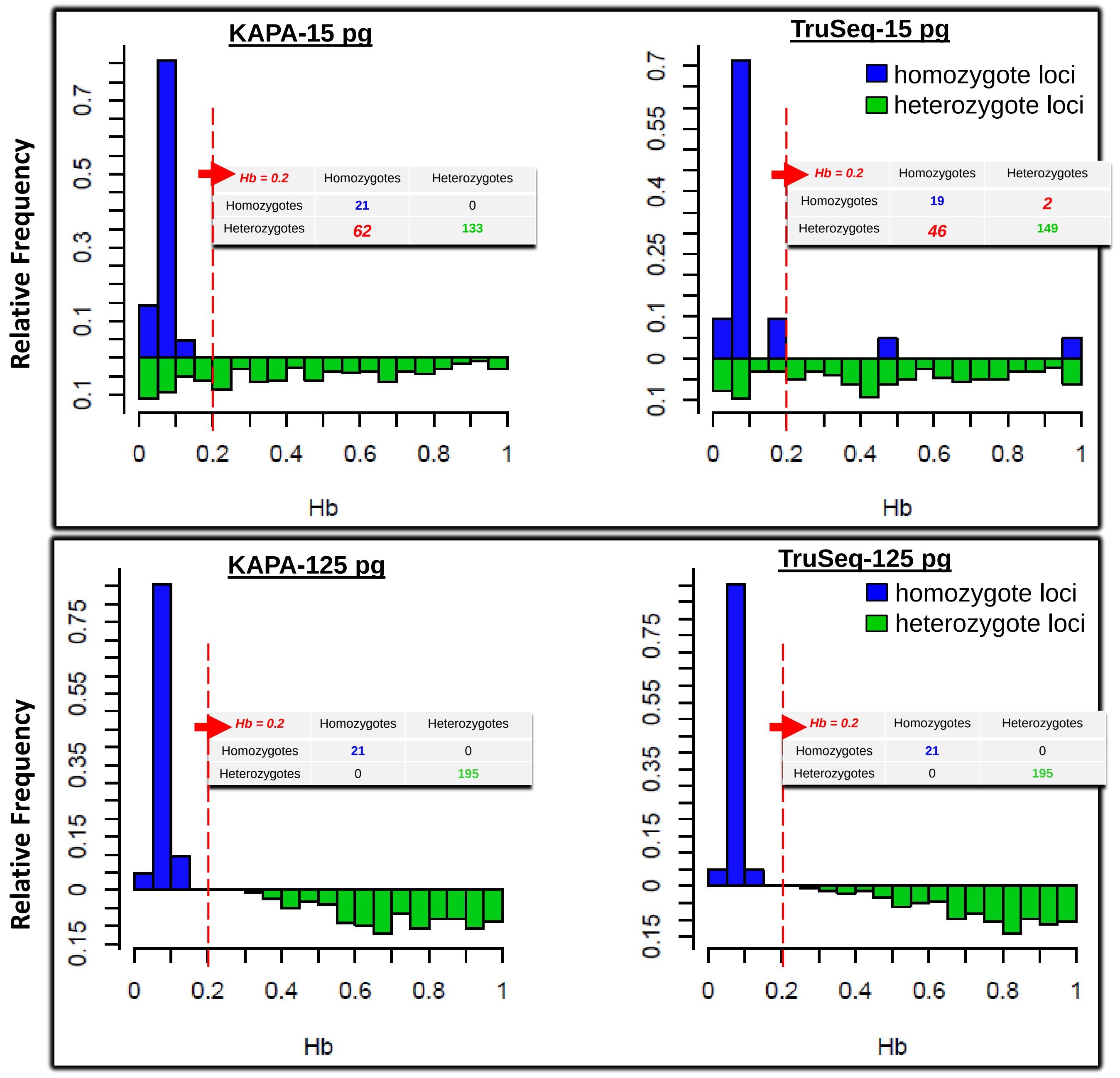
S2 = Back stutter sequences not attributed to S1

$$\textbf{Percent Coverage} = \frac{\text{Coverage of any sequence from a given locus}}{\text{Total coverage of the locus}} \times 100$$

- ***As expected, improved discrimination between known alleles (A) and the remainder of the sequences (N, S1, and S2) is observed as the amount of DNA template increases.***
- A threshold used to interpret single-source profiles should evaluate the tradeoff between the known allelic, stutter, and noise sequences. This can be accomplished through the construction of Receiver Operating Characteristic (ROC) curves. ROC curves plot the true positive rate vs. false positive rate as a function of percent coverage.
- ***Note:** In the example above, a value of 15 % is <u>ONLY</u> used for illustrative purposes and not as a recommended threshold. Each lab should perform sensitivity experiments and establish a threshold for interpretational purposes.*

6 Inferring Zygosity using Heterozygote Sequence Balance

In this dataset, there are 21 known homozygotes and 195 heterozygotes



KAPA-15 pg

Hb = 0.2	Homozygotes	Heterozygotes
Homozygotes	21	0
Heterozygotes	62	133

TruSeq-15 pg

Hb = 0.2	Homozygotes	Heterozygotes
Homozygotes	19	2
Heterozygotes	46	149

KAPA-125 pg

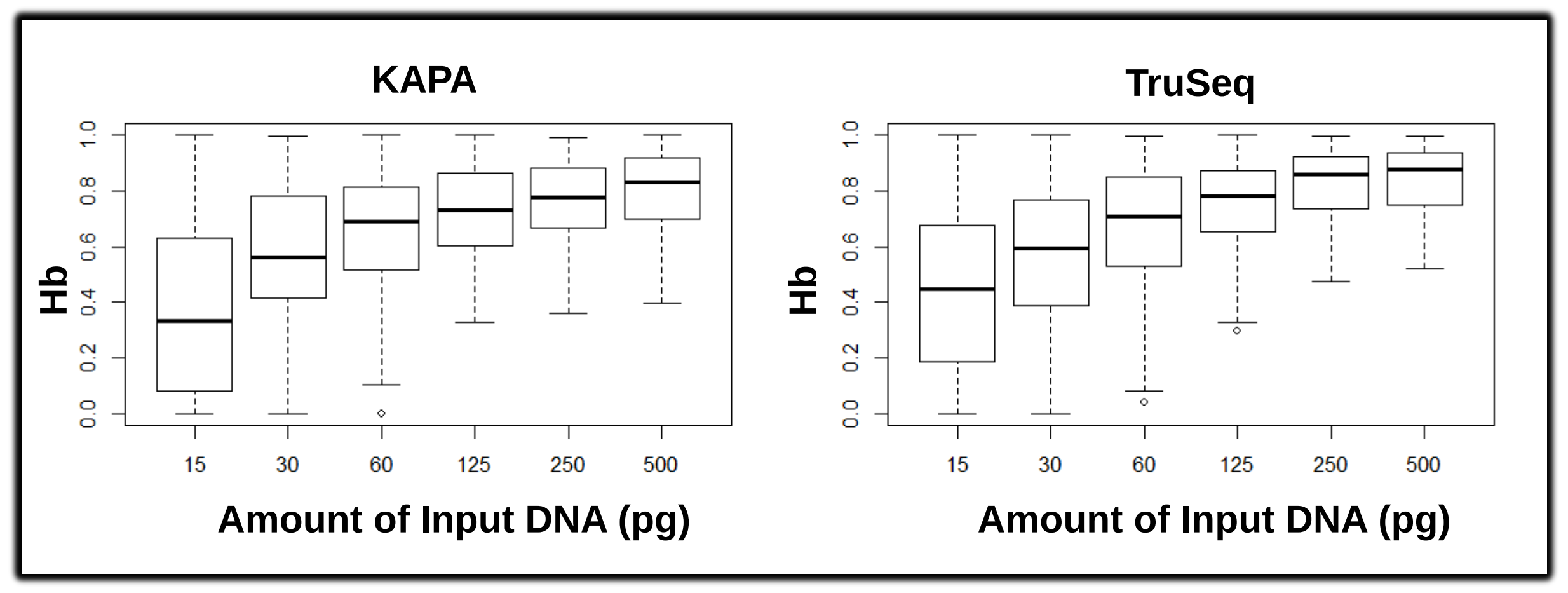
Hb = 0.2	Homozygotes	Heterozygotes
Homozygotes	21	0
Heterozygotes	0	195

TruSeq-125 pg

Hb = 0.2	Homozygotes	Heterozygotes
Homozygotes	21	0
Heterozygotes	0	195

- Heterozygote balance (Hb) was calculated using the following definition:
  Hb = low coverage allele/high coverage allele
- Histograms were developed by using the calculated Hb and the known zygosity at each locus across the different DNA amounts and library kits.
- A zygosity threshold for single-source profiles should evaluate the risks of misidentifying heterozygous and homozygous genotypes. This can be accomplished through the construction of ROC plots.
- ***Note:** In the example above, an Hb = 0.2 is <u>ONLY</u> used for illustrative purposes and not as a recommended threshold. Each lab should perform sensitivity experiments and establish a threshold for interpretational purposes.*

7 Hb versus DNA Input



- ***The variation of the heterozygote balance decreased as the amount of DNA increased.***
- ***The heterozygote balance is equally variable for both library kits as a function of DNA amount.***

<u>Acknowledgement:</u> The authors express gratitude to Doug Storts and Spencer Hermanson at Promega for providing PowerSeq Auto reagents and technical guidance.

<u>Reference:</u> [1] 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 science international. Genetics 29 (2017) 21-28.
[2] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic acids research 24(14) (1996) 2807-12.
[3] C. Brookes, J.A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic science international. Genetics 6(1) (2012) 58-63.