MRHSRS-24-12049: AccuGenomics Internal Standards Enhance NGS Pathogen Surveillance

Tom Morrison[1] & Anthony C. Fries[2]

[1]AccuGenomics, Inc. Wilmington, NC. [2]US Air Force School of Aerospace Medicine Public Health, Wright-Patterson AFB, OH

Abstract

COVID-19 NGS surveillance laid the groundwork for how sequencing data can be incorporated into operations to assess countermeasures and evaluate evolving threats. However, it takes months to develop a robust NGS test with limited assurance that performance is consistent. AccuGenomics internal standards (IS) accelerate new NGS development by providing instant feedback on data yields and accuracy. Operationally, the controls ensure results and facilitate investigations. AccuGenomics manufactures and supports the integration of internal standards to improve quality control and standardization of routine NGS manufacturing tests. Standardized Nucleic Acid Quantification for sequencing (SNAQ^{TM}-SEQ) internal standards are mixtures of synthetic nucleic acid controls that when added to each sample biochemically covary in yield and sequence detection in targeted or non-targeted NGS assays. Additionally, AccuGenomics supports the integration of internal standards into manufacturing bioinformatic pipelines. This poster provides examples of using internal standards to measure test capture efficiency and duplication rate; limit controls as a sensitivity QC for adventitious agents; and ddPCR like accuracy for abundance measurements of manufacturing host/vector and therapeutic vector.

Introduction

Next Generation Sequencing (NGS) has demonstrated great potential as a single method to detect many biomarkers with high sensitivity and specificity. Unfortunately, NGS is highly complex and relies on analytical surrogates that indirectly demonstrate testing proficiency. AccuGenomics makes internal standards (IS) for NGS testing in a similar approach as analytic mass spectrophotometry uses IS to ensure testing accuracy. The IS are designed to biochemically and bioinformatically covary with biomarkers of interest to directly ensure testing sensitivity and analytic accuracy for every sample tested. The IS are manufactured to have $< 10^{-8}$ base error rate and IS mixtures are prepared using 2 to 3 different abundance methods under quality control systems to create reference quality materials. AccuGenomics tunes the IS mixture for the testing platform and supports integration into bioinformatic pipelines. This poster discusses uses of the IS mixtures relevant to biomanufacturing environment.

Methods

Lyophilized SARS-CoV-2 ssRNA IS and a double blinded set of 73 SARS-CoV-2 were shipped to three outbreak testing laboratories. IS was added to each sample prior to using an amplicon tiling sequencing approach (ARTICv4.1), and the resulting FASTQ were processed through VSOFT docker container to create IS QC and viral strain reports.

1. SNAQ™-SEQ SARS-CoV-2 Internal standards

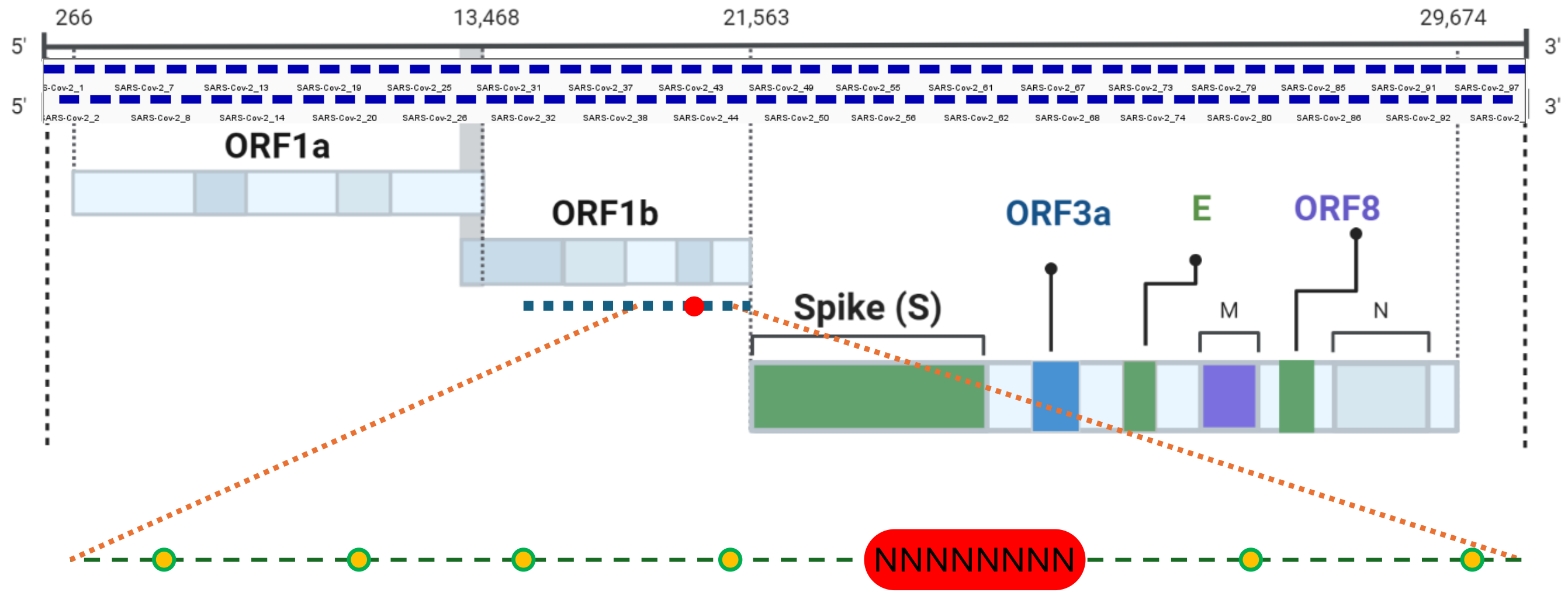

Figure 1. SARS-CoV-2 genome. The SARS-CoV-2 coding regions regions (middle boxes) are sequenced by the ARTICv4.1 protocol by enrichment of a series of overlapping amplicons (top blue bars). SNAQ-SEQ complexity capture controls (bottom) indicates unique bases (circles) and degenerate bases (red box) that span a conserved viral genomic region.

- SNAQ™ SARS-CoV-2 internal standards consist of tiled ssRNA of 3-4 ARTIC v4.1 amplicons per tile of Wuhan reference sequence with unique base changes every 80 positions to allow bioinformatic identification
- Two regions contain degenerate bases used to provide an estimate of library complexity capture by comparing input vs. unique control reads

3. Complexity Capture Quality Control

	Lab1	Lab2	Lab3
POOL1	11.4 ± 1.6%	3.2 ± 0.9%	3.6 ± 0.9%
POOL2	17.1 ± 1.7%	7.2 ± 1.9%	6.1 ± 1.2%

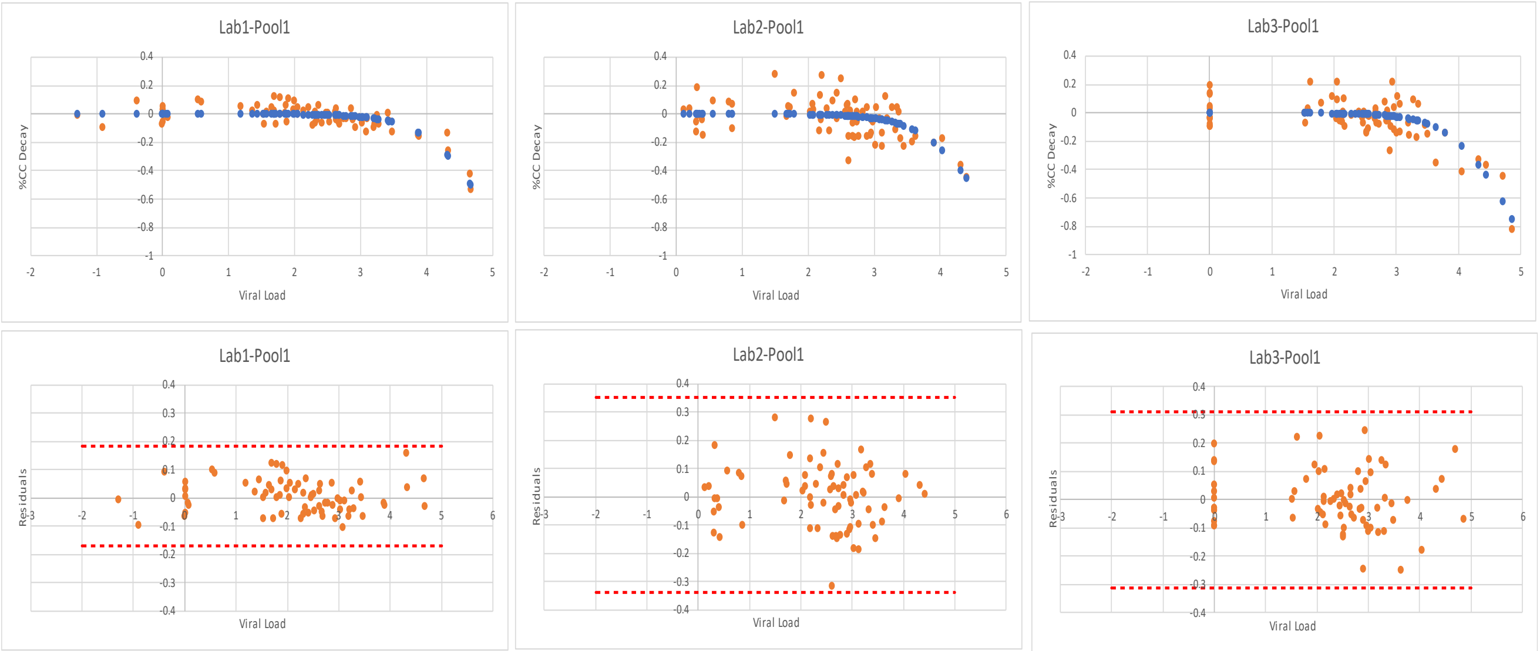

Figure 3. SNAQ™ Complexity Capture QC Profile for each Study Laboratory. The loss of unique Complexity Control Count (output / input) may be modeled as a function of viral load. The top table indicates the nominal complexity capture for the indicated labs (column heading) and indicated ARTIC multiplex PCR pool (rows). Flow cell sample load normalization leads to competition between NT and IS reads. With increasing viral load (x-axis), the Unique CC count drops. The %CC response was normalized using the lab's average %CC capture in low viral load samples (Pool1 %CC table in previous slide). The log10 CC change with log10 viral load (y-axis, top row plots) was modeled as Frac_CC = (IS x A) / (viral_load + IS x A), where A is lab & pool specific fudge factor influenced mostly by the read duplication rate. The residuals of the model (y-axis, lower plot rows) were found to be normally distributed by the Jarque-Bera test, from which 99.9% confidence intervals (red dashes) were calculated using Excel norm.inv function. For brevity, Pool 2 data not shown. CC QC indicates a significant per sample drop in test sensitivity if %CC deviates more than 1.5, 2, 2-fold from model for Lab1,2,3, respectively. Of note, SNAQ™ was designed to QC <10,000 genomic input samples because these samples are more challenging to obtain good sequence.

How to more accurately detect NGS library preparation testing errors?

- SNAQ™ complexity capture (CC) acts as a full process control to measure how well each sample's viral genome was captured as sequencing reads
- Table indicates each lab's nominal CC rate
- A CC model was created from each lab's CC vs viral Load response; the residuals to this model indicate if a sample CC is nominal
- SNAQ™ could detect a >two-fold drop in complexity capture with high specificity
- SNAQ™ CC indicated each sample met nominal testing performance

2. Standardized Viral Abundance

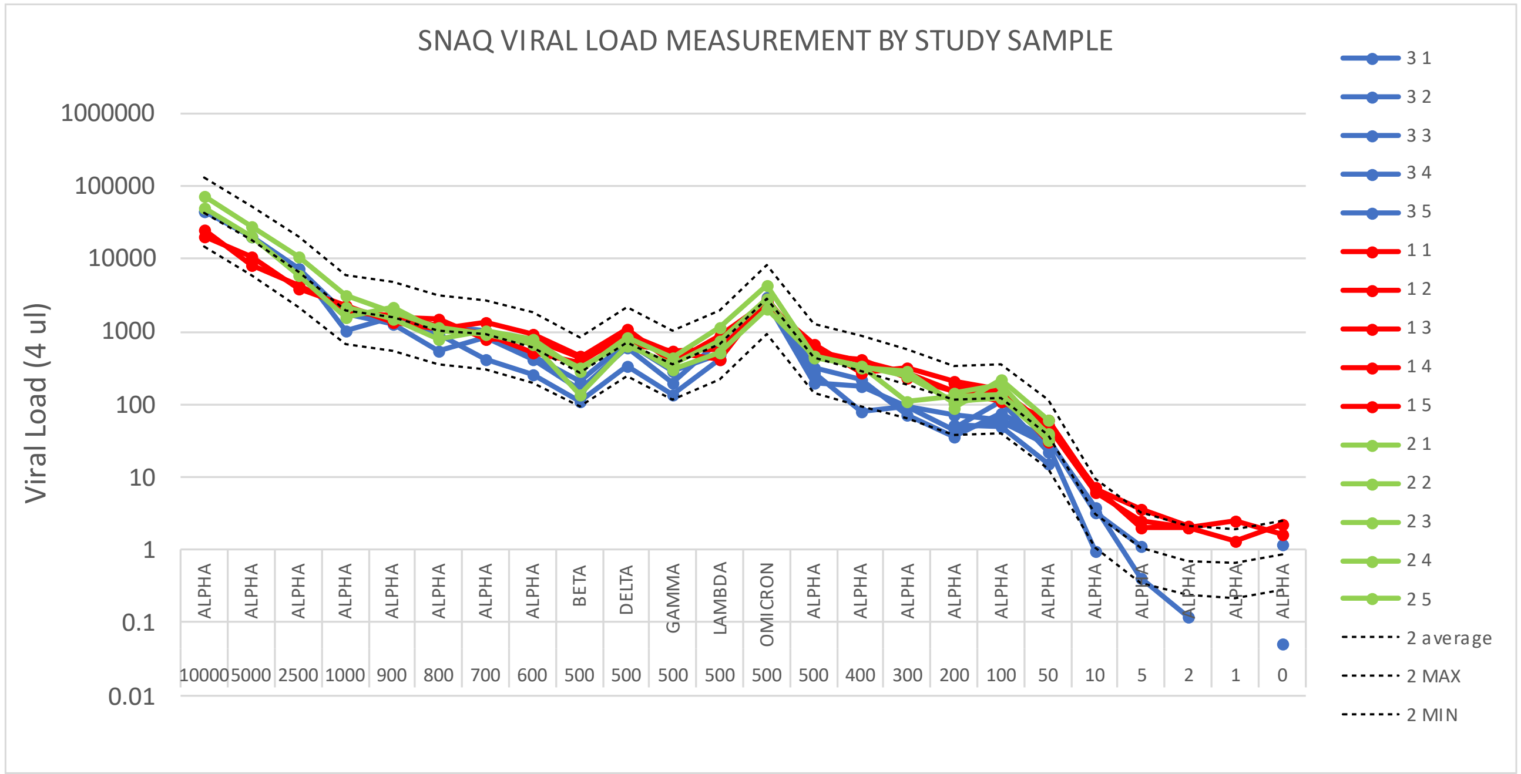

Figure 2. SNAQ™ Based Viral Load Measurement of Study Samples. Study FASTQ were analyzed using SNAQ™-VSOFT v1.2beta to estimate viral load (y-axis) of each study sample (x-axis). X-axis indicates COVID-19 genome and approximate genomic input; legend indicates lab and replicate number. The mean (central dashed line) and three-fold difference from mean (outer dashed lines) indicate that 163 of 165 samples with greater than 100 genomic abundance were within 3-fold of mean.

- Poor test results can arise from lower-than-expected genomic input or errors in NGS testing procedure
- SNAQ™ abundance directly measures genomic input in the NGS results by indicating sequence failure association with low sample input
- Results demonstrate SNAQ™ abundance measurements varied less than three-fold for samples ranging from $>10^4$ to 10 copy viral genomes sequenced in three different laboratories using different sequencing instruments and altered ARTIC NGS protocols
- SNAQ™ was able to standardize viral abundance measurements

4. SNAQ™ VAF LOD Estimation

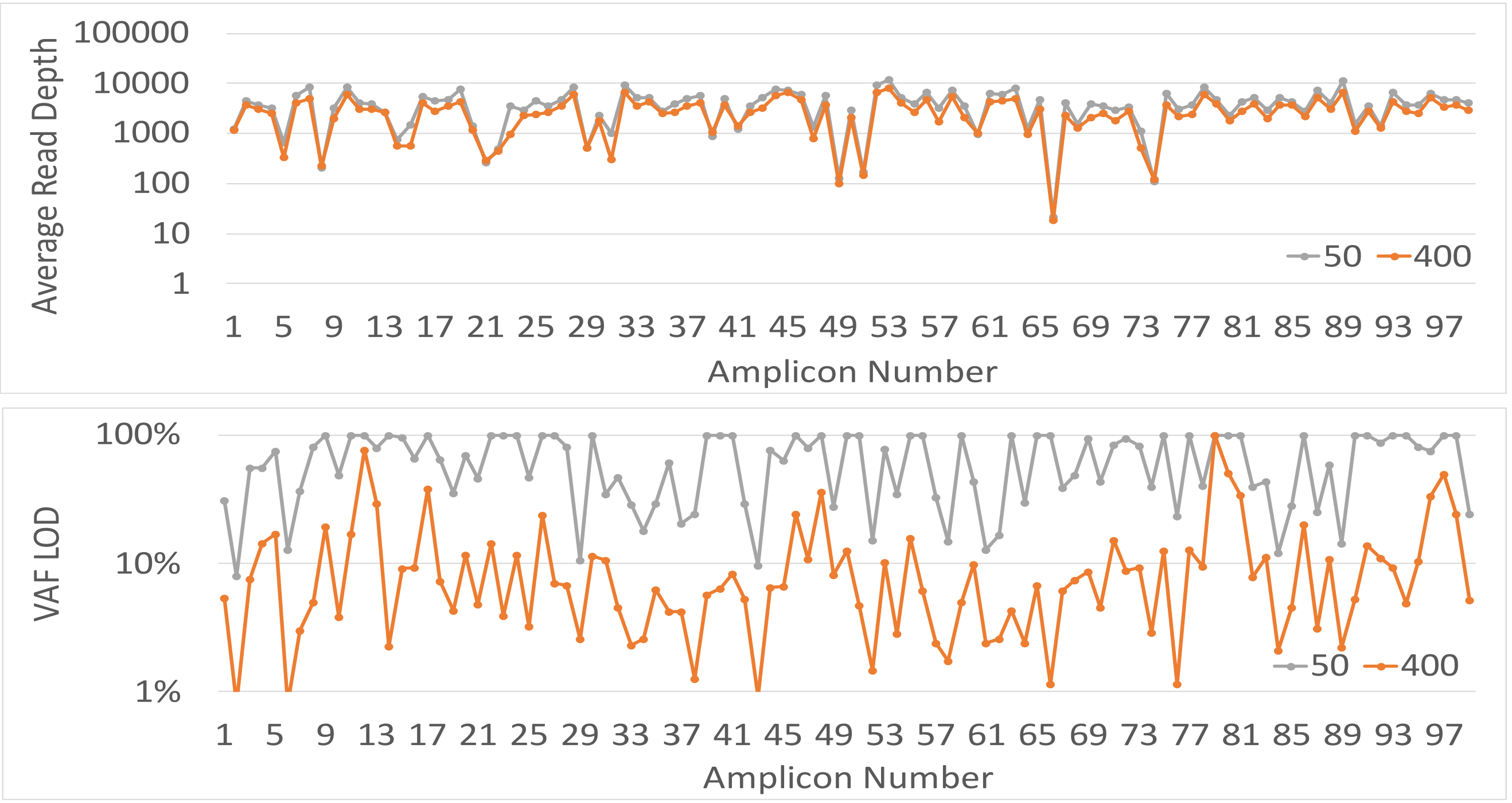

Figure 4. Read Depth versus Complexity Capture as Indication of Testing Sensitivity. Average read depth (top plot) and SNAQ™ VAF LOD (bottom plot) per amplicon were calculated using snaq-vsoft v1.1 software for two samples at indicated 50 and 400 Alpha COVID-19 genomic input (legend) by lab 1. Artic COVID-19 NGS testing uses 98 tiled amplicons to which each read pair was mapped, base count extracted, and average read depth calculated. Read depth for read pairs mapping to two adjacent amplicons were split between adjacent amplicons in proportion to each amplicons mapped reads. VAF was calculated by first using the read depth ratio for the matching NT and IS amplicons to calculate NT yield, complexity capture was used to estimate how many of the NT reads were unique, and then a Poisson calculation was used to indicate the 95% confidence of detecting a given VAF based on the number of unique captured NT amplicon templates.

- Average read depth by amplicon for COVID-19 Alpha NIIMBL study samples represents current practice for sequencing sensitivity
- Results indicate similar read depth for the two samples
- Expectation would be very similar variant calling sensitivity, this expectation would be incorrect
- SNAQ™ coverage estimates how many unique templates were captured in each amplicon
- Amplicon coverage allows estimation of VAF LOD on a per sample per amplicon basis
- Plot depicts 95% confidence of detecting variant at or above the indicated VAF
- LOD model predicted >95% of variant detected in all low viral load samples
- SNAQ™ coverage provides a per sample per region variant sensitivity measurement

Summary

IS-calculated interlaboratory viral load replicates were within three-fold of each other. An IS complexity capture profile for each laboratory indicated that, although there were inter laboratory sensitivity differences, each sample was nominally assessed. The IS analysis gave a per-sample indication of variant allele fraction sensitivity.

Use of AccuGenomics IS in pathogen surveillance monitoring facilitates:

1. Intrahost minor variant analysis for transmission studies,
2. A shift away from digital PCR to NGS for more accurate strain characterization of environmental samples,
3. Standardized viral load measurement in samples.

AccuGenomics has IS for COVID-19 (ARTICv4.1 & Midnight-1200 protocols), Influenza A and B strains, and accompanying software containers for analysis, and has demonstrated ability to rapidly create similar controls for novel pathogens.