Introduction

Spike-ins are RNA transcripts of known sequences. It was first used to calibrate variations in RNA hybridization assays. Now, it is also adapted to be used in sequencing. In the Piantadosi Lab, spike-ins are used as quality control. During the preparation step, each sample is assigned and added a specific spike-in. When the sequencing results come out, if the highest-read spike-in matches the assigned spike-in, the sample is sequenced correctly.

However, a cutoff for the number of spike-in reads has not been set. The Piantadosi Lab has only been checking whether the assigned spike-in is detected and marking it as confirmed on the Sequentially Sequenced SARS Samples Sheet (SSSSS) on Microsoft Teams, despite there may be other spike-ins detected with higher reads.

Therefore, the goal of this project is to find a cutoff for the number of spike-in reads. Along the way, a few questions come up and are also investigated:

- Are there any spike-in aliquots with low concentrations?
- What is the upper cutoff for the read difference between spike-ins with the highest read and the second-highest read?
- Is there any cross-contamination of spike-ins in a batch?

Methods

A total of 1606 Covid (n = 1295), Dengue (n = 173), and BWH (n = 138) samples were used. Spike-in count information was downloaded from DNAnexus, and other information (e.g., assigned spike-in, total reads, etc.) for each sample was obtained from SSSSS on Microsoft Teams.

Standardization was performed for calculating both cutoffs. To determine the cutoff for the number of spike-in reads, read per million (RPM) was used. RPM = highest spike-in reads / total reads * 1000000. To determine the cutoff for the read difference, the second-highest-to-the-highest spike-in-read ratio (SHR) was used. SHR ranges from 0 – 1. SHR = second highest spike-in reads / highest spike-in reads.

The spike-in confirmation status was re-determined for each sample. Spike-in is confirmed if and only if the highest-read spike-in matches the assigned spike-in; otherwise, it is not confirmed.

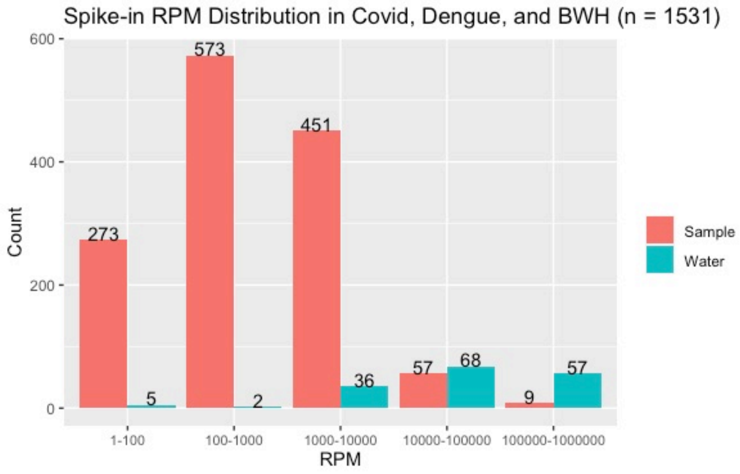
All analyses were done in R.

Results

RPM cutoff

Comparing the RPM values between samples and water (Figure 1), most water have higher RPM values within the 10,000 – 1,000,000 range. Thus, analysis on determining the RPM cutoff will only be performed on samples. Water will be excluded.

Figure 1.



Looking at samples with confirmed spike-ins, shown in Figure 2 and Table 1, most of the samples (74.78%) have RPM values within the 100 – 10,000 range. Interestingly, ~20% of the samples have RPM values within the 1 – 100 range. But according to Table 2, the 5th percentile corresponds to 12.41 RPM. Hence, the lower cutoff for RPM value should be within the 12.41 – 100 range.

Figure 2.

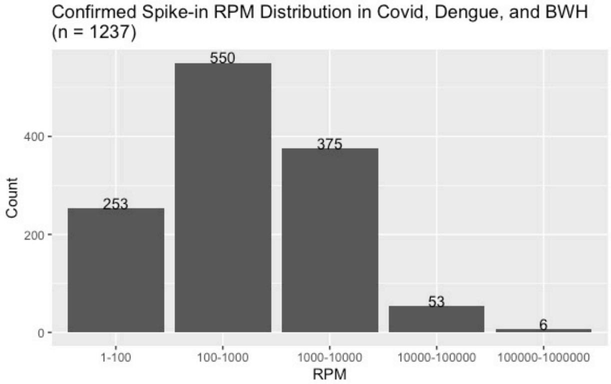


Table 1.

RPM Range	**Percentage**
1-100	20.45%
100-1000	44.46%
1000-10000	30.32%
10000-100000	4.28%
100000-1000000	0.49%

Table 2.

Percentile	**RPM**
20	94.28
15	61.85
10	32.56
5	12.41
3	7.24

Spike-in with low concentrations

Samples with RPM values lower than 100 have spike-in aliquots with low concentration added. According to Table 3, the aliquots of ERCC-00023, ERCC-00077, ERCC-00054, and ERCC-00067 may have low concentrations. Only percentages greater than 50% are shown.

Table 3.

Assigned Spike-ins	Low RPM Count	Total Count	Percentage
ERCC-00023	24	27	88.89%
ERCC-00077	16	20	80.00%
ERCC-00054	18	23	78.26%
ERCC-00067	12	24	50.00%

<u>Ratio cutoff</u>

Only samples with two or more spike-ins detected are analyzed. As shown in Figure 3 and Table 4, most of the samples (83.55%) have SHR values within the 0 – 0.1 range, while the rest (16.45%) lie within the 0.1 – 1 range. According to Table 5, the 85th and 90th percentiles, corresponding to a 0.16 and 0.43 SHR value respectively, seem most appropriate. Thus, the SHR cutoff should be within the 0.16 – 0.43 range.

Figure 3.

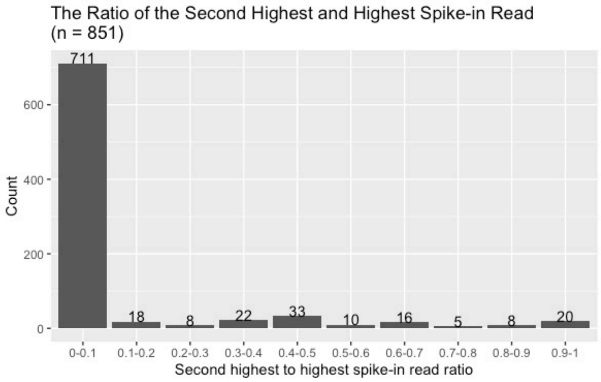


Table 4.

SHR Range	Count	Percentage
0 - 0.1	711	83.55%
≥ 0.1	140	16.45%

Table 5.

Percentile	SHR
80	0.05
85	0.16
90	0.43
95	0.63
97	0.84

<u>Spike-in with possible contamination</u>

Samples with SHR values greater than 0.43 are further analyzed. According to Table 6, the batches sequenced on December 1, 2022, February 13, 2023, February 2, 2023, and December 8, 2022, might have been contaminated while preparing the library. Only percentages greater than 50% are shown.

Table 6.

Sequenced Date	Virus	Count	Total Count	Percentage
2022-12-01	Dengue	15	18	83.33%
2023-02-13	Dengue	12	15	80.00%
2023-02-02	Dengue	13	21	61.90%

2022-12-08	Dengue	11	18	61.11%

Conclusion

The lower cutoff for spike-in RPM is 12.41 - 100, while 12.41 (5th percentile) is the lowest value. This means, that for any samples with RPM lower than 12.41, their assigned spike-in aliquots should be checked to confirm if they have the appropriate concentration.

The upper cutoff for spike-in SHR is 0.16 - 0.43, while 0.43 (90th percentile) is the highest value. This indicates that, for any samples with SHR greater than 0.43, their assigned spike-in aliquots should be checked to confirm if there is potential contamination.

For future direction, having more data could narrow the range of the cutoffs.

Supplement: R Function

Confirm_spikein3.0.1 is an R function developed to calculate whether the expected spike-ins are confirmed and other spike-ins detected and their reads. The calculated results are exported as an Excel file, which can then be copied to SSSSS on Microsoft Teams.