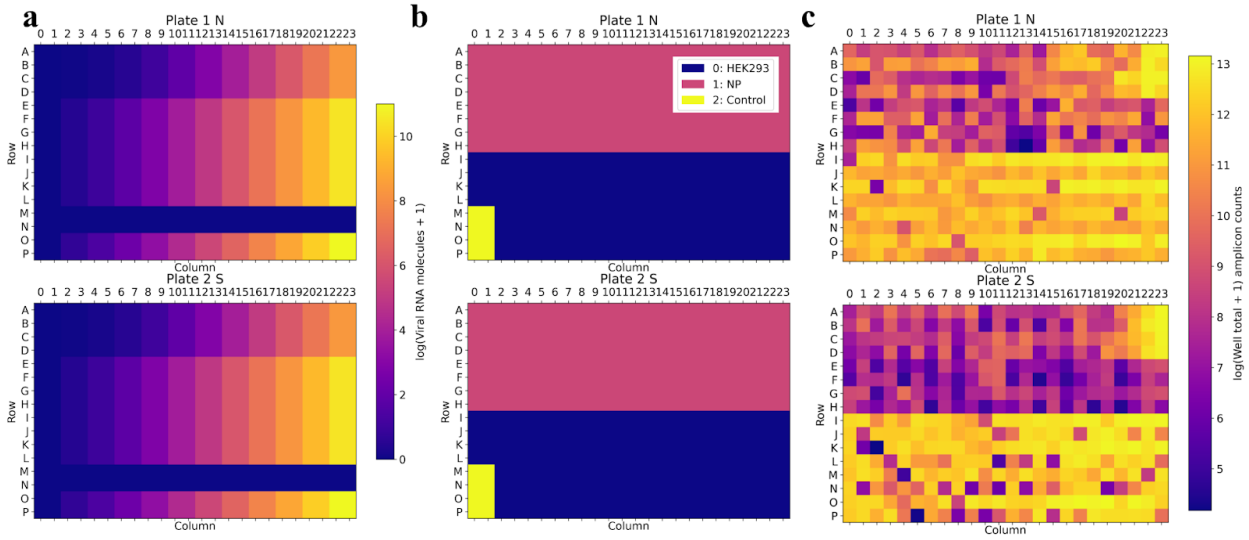
Supplementary Material

Fast and accurate diagnostics from highly multiplexed sequencing assays

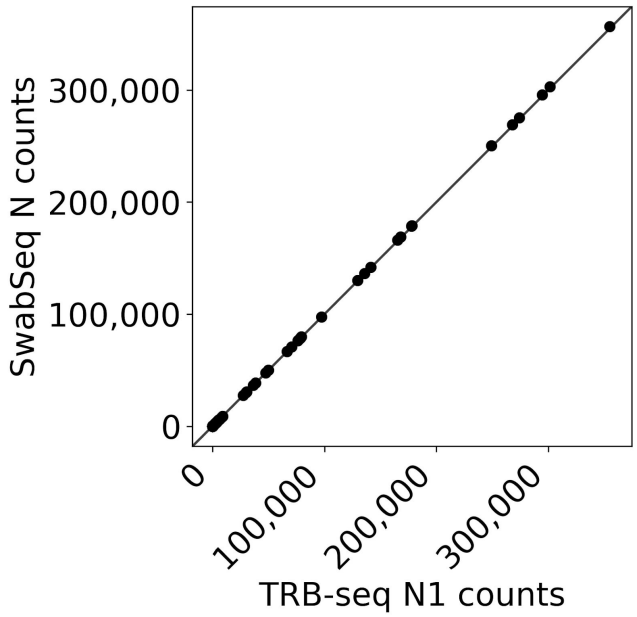
A. Sina Booeshaghi[1], Nathan B. Lubock[2], Aaron R. Cooper[2], Scott W. Simpkins[2], Joshua S. Bloom[2,3], Jase Gehring[4], Laura Luebbert[5], Sri Kosuri[2], & Lior Pachter[5,6,*]

1. Department of Mechanical Engineering, California Institute of Technology, Pasadena, CA
2. Octant Inc., Emeryville, CA
3. Department of Human Genetics, University of California, Los Angeles, Los Angeles, United States
4. Department of Genome Sciences, University of Washington, Seattle, WA
5. Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA
6. Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA

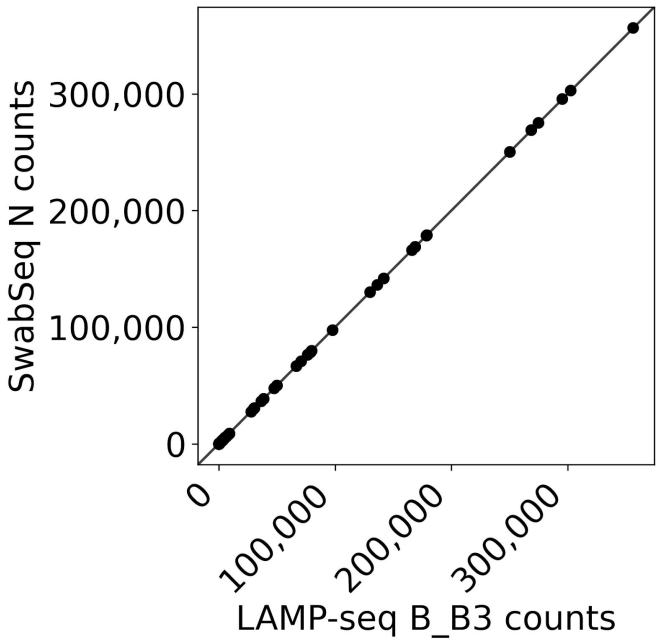
*Address correspondence to lpachter@caltech.edu



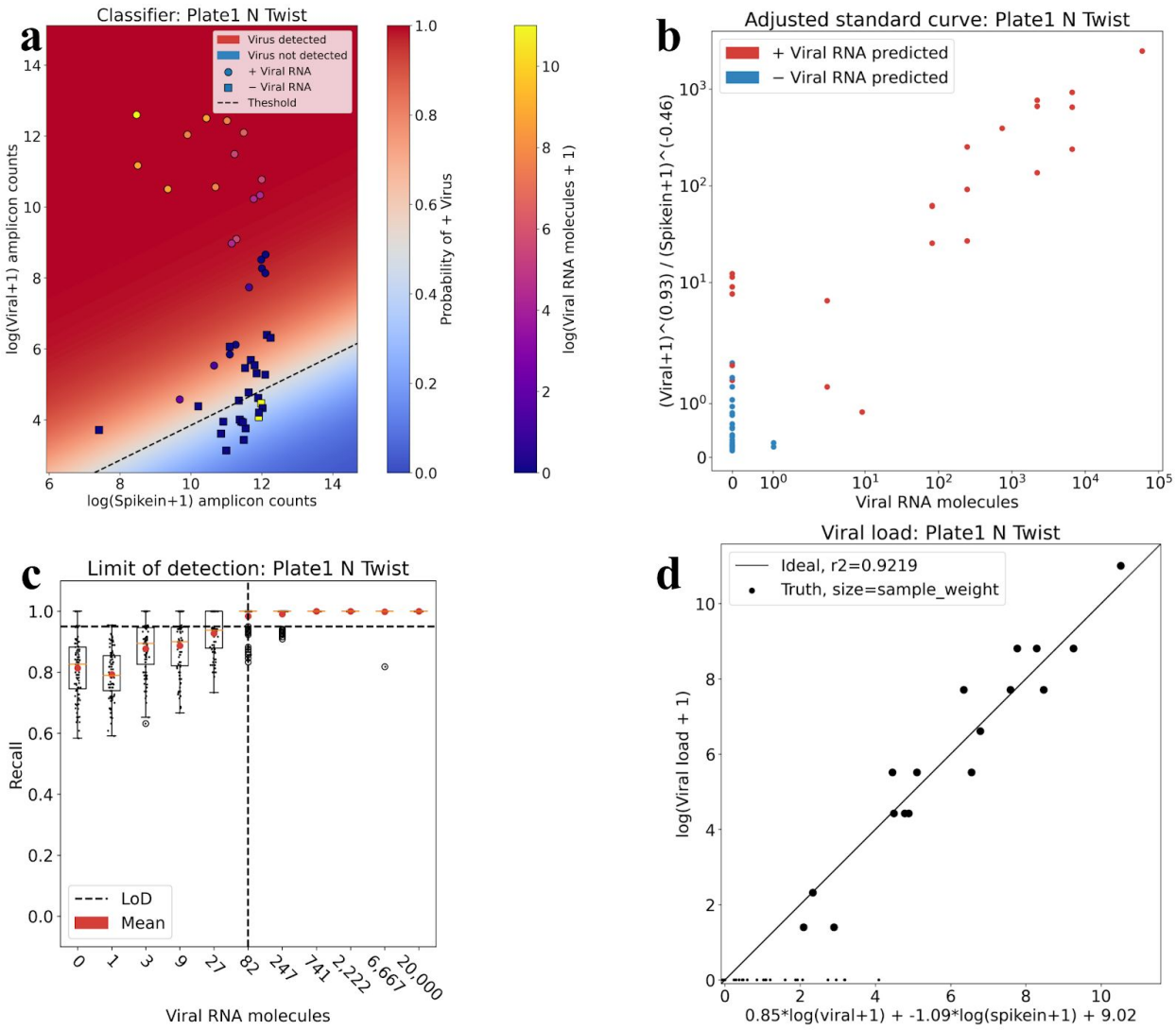
Supplementary Figure 1: SwabSeq experimental design. The SwabSeq experiment consisted of two 384-well plates, each with a titration series of viral RNA from two companies, Twist and ATCC, for a total of 768 uniquely barcoded samples (a). HEK293 lysate, nasopharyngeal (NP) lysate, and controls were included in all the wells of each plate (b). The first plate was used to test primers bound to the SARS-CoV-2 N gene and the second plate was used to test primers bound to the SARS-CoV-2 S gene. Samples were barcoded using the i5 and i7 Illumina indices, the former indexing the well and the latter the plate. Total amplicon counts per well were greater for HEK293 than for NP lysate (c). The code to reproduce this figure is here: code.



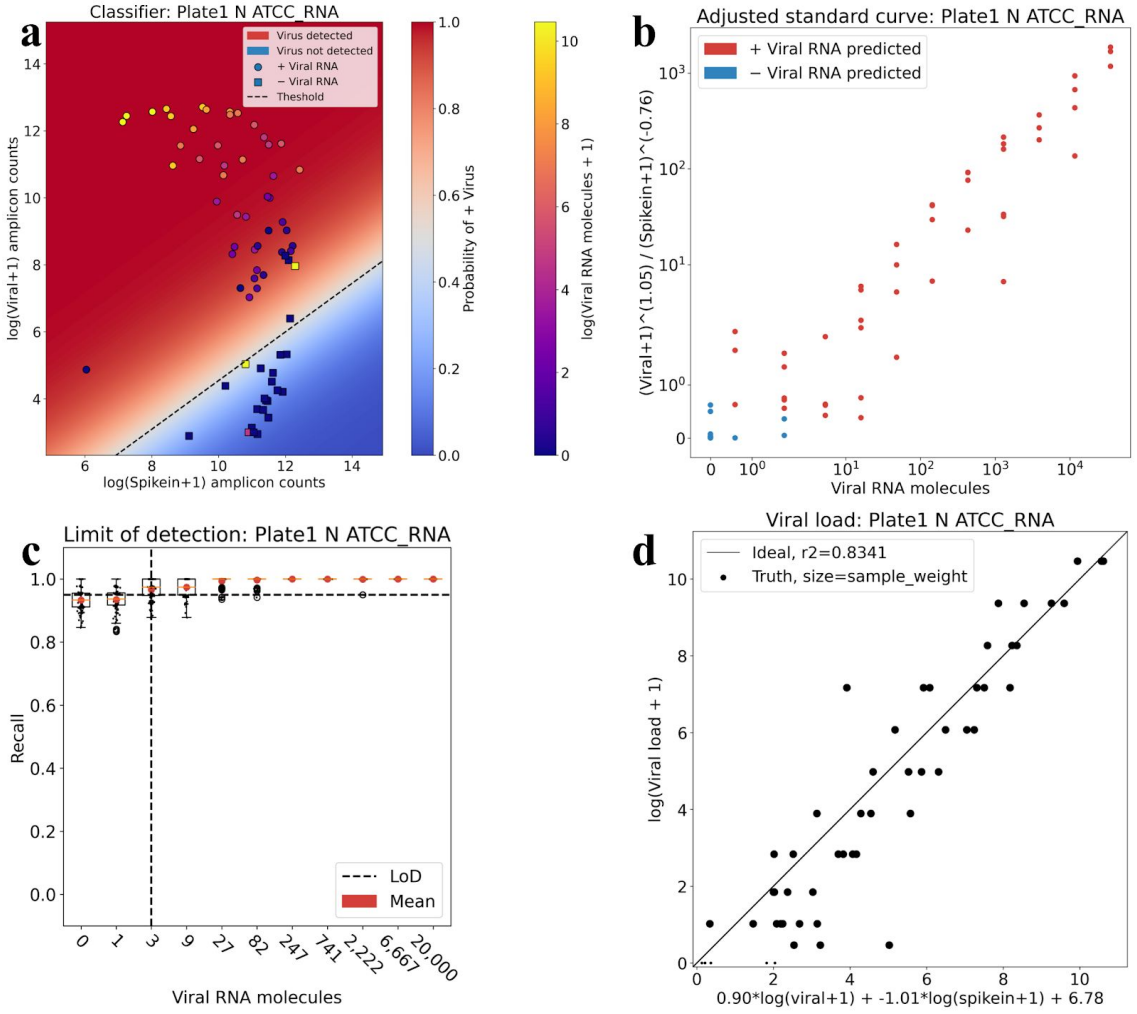
Supplementary Figure 2: Validation of the TRB-seq workflow with the SwabSeq workflow where each point is a sample and the counts for each sample corresponding to the N gene for SwabSeq and the N1 gene for TRB-seq are plotted. The code to reproduce this figure is here: code.



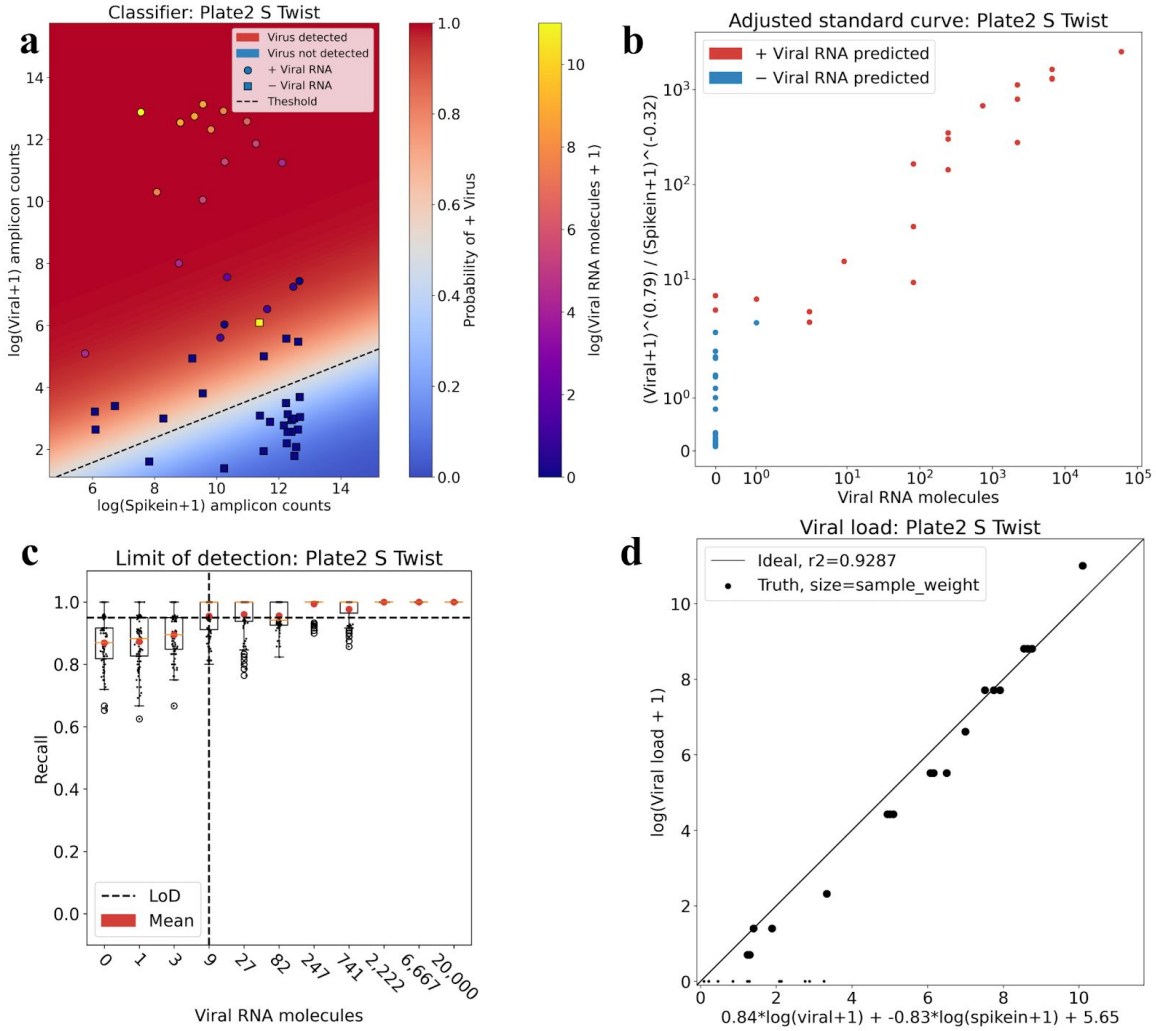
Supplementary Figure 3: Validation of the LAMP-seq workflow with the SwabSeq workflow, where each point is a sample. The counts for each sample corresponding to the N gene for SwabSeq and the B_B3 gene target for LAMP-seq are plotted. The code to reproduce this figure is here: code.



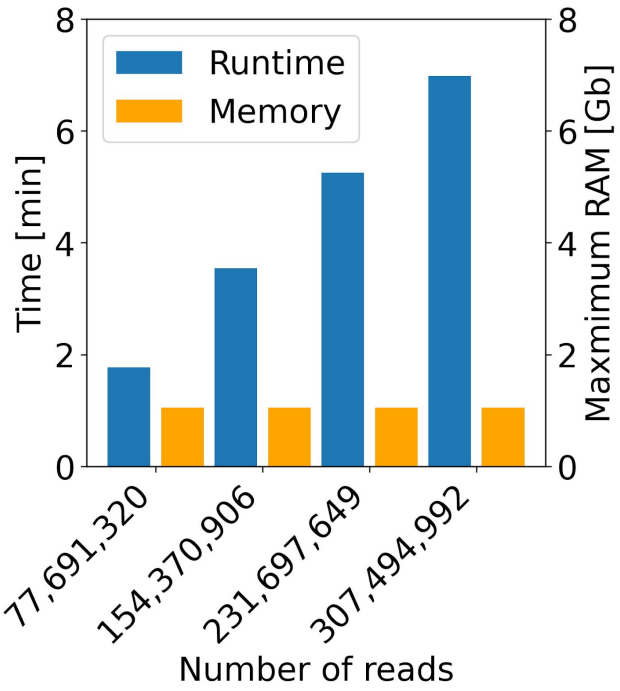
Supplementary Figure 4: Sample classification, viral load prediction and limit of detection. a) Positive and negative samples from the Plate 1 N Twist experiment can be effectively separated using logistic regression. Points correspond to samples and are colored by the known amount of viral RNA per sample. The probability of each sample of having a non-zero amount of viral RNA is given by the logistic function and is painted as orthogonal to the logistic regression boundary. The shape of the point indicates whether it was predicted to be positive for viral RNA (circle) or negative (square). **b)** The standard curve measuring spike-in and virus vs the known amount of viral RNA per sample with optimal exponential coefficients determined by logistic regression; samples are colored by their predicted classification. **c)** The limit of detection as estimated from 99 rounds of split/test and logistic regression to classify samples with non-zero amount of viral RNA. The limit of detection is defined as the number of RNA molecules for which the recall is greater than 19/20 (=0.95) **d)** The viral load per sample can be predicted with (a weighted) linear regression using the log counts from each gene. Each point is a sample, with perfect predictions lying on the diagonal line. Size of points represents their weight, with points weighted so that each titer is represented with equal weight. The code to reproduce each figure is here: code (a) and code (b), code (c), code (d).



Supplementary Figure 5: Sample classification, viral load prediction and limit of detection. a) Positive and negative samples from the Plate 1 N ATCC RNA experiment can be effectively separated using logistic regression. Points correspond to samples and are colored by the known amount of viral RNA per sample. The probability of each sample of having a non-zero amount of viral RNA is given by the logistic function and is painted as orthogonal to the logistic regression boundary. The shape of the point indicates whether it was predicted to be positive for viral RNA (circle) or negative (square). **b)** The standard curve measuring spike-in and virus vs the known amount of viral RNA per sample with optimal exponential coefficients determined by logistic regression; samples are colored by their predicted classification. **c)** The limit of detection as estimated from 99 rounds of split/test and logistic regression to classify samples with non-zero amount of viral RNA. The limit of detection is defined as the number of RNA molecules for which the recall is greater than 19/20 (=0.95) **d)** The viral load per sample can be predicted with (a weighted) linear regression using the log counts from each gene. Each point is a sample, with perfect predictions lying on the diagonal line. Size of points represents their weight, with points weighted so that each titer is represented with equal weight. The code to reproduce each figure is here: code (a) and code (b), code (c), code (d).



Supplementary Figure 6: Sample classification, viral load prediction and limit of detection. a) Positive and negative samples from the Plate 2 S Twist experiment can be effectively separated using logistic regression. Points correspond to samples and are colored by the known amount of viral RNA per sample. The probability of each sample of having a non-zero amount of viral RNA is given by the logistic function and is painted as orthogonal to the logistic regression boundary. The shape of the point indicates whether it was predicted to be positive for viral RNA (circle) or negative (square). **b)** The standard curve measuring spike-in and virus vs the known amount of viral RNA per sample with optimal exponential coefficients determined by logistic regression; samples are colored by their predicted classification. **c)** The limit of detection as estimated from 99 rounds of split/test and logistic regression to classify samples with non-zero amount of viral RNA. The limit of detection is defined as the number of RNA molecules for which the recall is greater than 19/20 (=0.95) **d)** The viral load per sample can be predicted with (a weighted) linear regression using the log counts from each gene. Each point is a sample, with perfect predictions lying on the diagonal line. Size of points represents their weight, with points weighted so that each titer is represented with equal weight. The code to reproduce each figure is here: code (a) and code (b), code (c), code (d).



Supplementary Figure 7: Runtime and memory footprint of kallisto | bustools shown for a total of 1 lane (77,691,320 reads), 2 lanes (154,370,906 reads), 3 lanes (231,697,649 reads), and 4 lanes (307,494,992 reads) of Illumina NextSeq 550 data. The code to reproduce this figure is here: code.

Lane	Index 1	Read 1	Read 2
1	Lane 1 Index 1	Lane 1 Read 1	Lane 1 Read 2
2	Lane 2 Index 1	Lane 2 Read 1	Lane 2 Read 2
3	Lane 3 Index 1	Lane 3 Read 1	Lane 3 Read 2
4	Lane 4 Index 1	Lane 4 Read 1	Lane 4 Read 2

Supplementary Table 1: Links to all of the SwabSeq FASTQ files analyzed.

Software	Version
Anndata	0.7.1
bustools	0.40.0 (branch: covid)
awk (GNU awk)	4.1.4
grep (GNU grep)	3.1
kallisto	0.46.2 (branch: covid)
kb_python	0.24.4 (branch: count-kite)
Matplotlib	3.0.3
Numpy	1.18.1
Pandas	0.25.3
Scipy	1.4.1
sed (GNU sed)	4.4
sklearn	0.22.1
starcode	v1.3 17-07-2018
tar (GNU tar)	1.29
time (GNU time)	1.7

Supplementary Table 2: Software used.