A novel likelihood-based model to estimate SARS-CoV-2 viral titer from next-generation sequencing (NGS) data

Heather L Wells[1]; Joseph E Barrows[1]; Mara Couto-Rodriguez[1], Xavier O Jirau Serrano[1], Marilyne Debieu[1], Karen Wessel[2], Colleen B Jonsson[3], Bradley A Connor[4,5,6], Christopher E Mason[1,7-10], Dorottya Nagy-Szakal[1,11], Niamh B O'Hara[1,11]

1. Biotia Inc., New York, NY, USA
2. Zots Klimas, Germany
3. The University of Tennessee, Health Science Center
4. Weill Cornell Medicine, New York, NY, USA
5. The New York Center for Travel and Tropical Medicine, New York, NY, USA
6. Geosentinel, New York, NY, USA
7. Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University, New York, NY, USA
8. The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
9. The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA
10. The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA
11. SUNY Downstate Health Sciences University, The Department Cell Biology/College of Medicine, New York, NY, USA

Contact us!
Heather L Wells
wells@biotia.io

Niamh B O'Hara, PhD
ohara@biotia.io
biotia.io

ABSTRACT

Objectives The quantitative level of a pathogen present in a host is a major driver of infectious disease (ID) state and outcome. However, the majority of ID diagnostics are qualitative. Next-generation sequencing (NGS) is an emerging ID diagnostics and research tool that can be used to provide insights such as tracking transmission and identifying novel strains.

Methods We built a novel likelihood-based computational method that uses NGS data generated by hybrid capture (Twist Bioscience) to quantify viral titer. We used de-identified clinical specimens tested for SARS-CoV-2 using qRT-PCR and SARS-CoV-2 hybrid capture (Twist Bioscience). A subset of samples were also tested using the ARTIC platform. Given the proportion of the genome covered at varying depths for a single sample as input data, our model estimated the Ct of that sample as the value that produces the maximum likelihood of generating the observed hybrid capture NGS genome coverage data.

Findings The model fit on 119 training samples produced a good fit to the 28 testing samples, with a coefficient of correlation (r^2) of 0.8. The accuracy of the model was high (mean absolute percent error of ~10.5%), meaning our model is able to predict the Ct value of each sample within a margin of ±10.5% on average. Because of the nature of the commonly used ARTIC protocol, we found that all quantitative signals in this data were lost during PCR amplification and the model is not applicable for quantification of samples captured this way. The ability to model quantification is a major advantage of the hybrid capture protocol over ARTIC and other PCR-amplification protocols.

Conclusion To our knowledge, this is the first model to incorporate sequence data mapped across the genome of a pathogen to quantify the level of that pathogen in a clinical specimen. This has implications in ID diagnostics, research, and metagenomics.

STUDY DESIGN

Sample collection and extraction De-identified nasopharyngeal (NP) specimens were collected and processed under the IRB numbered Pro00042824 (Advarra). RNA from NP specimens were isolated and purified by manual extraction using Direct-zol DNA/RNA MiniPrep kit (250µl input volume, Zymo Research, Irvine, CA).

qPCR RT-qPCR was performed using four different SARS-CoV-2 assays, CDC SARS-CoV-2 Assay (N1, N2 target), Roche Cobas SARS-CoV-2 Assay (Orf1ab and E gene), GenArraytion COVID-19 Real-Time Assay and BGI RT-PCR kit.

SARS-CoV-2 NGS Assay Extracted and purified RNA samples were converted to cDNA TruSeq compatible libraries using Twist Library preparation kit, and libraries were pooled in 8-plex hybridization reactions for enrichment with the SARS-CoV-2 research panel.

ARTIC SARS-CoV-2 NGS Assay A subset of RNA samples were processed with an amplicon based approach using the NEBNext® ARTIC SARS-CoV-2 FS Library Prep Kit (Illumina®) workflow.

Sequencing All enriched library pools were spiked with 1% PhiX and sequenced on an Illumina NextSeq 550 platform using a NextSeq 500/550 High Output kit (Illumina, San Diego, CA) set to 150bp single-end reads.

COVID-DX The COVID-DX Pipeline included removal of low-quality reads, alignment to SARS-CoV-2 and off-target human and microbial genomes, extraction of mapped reads, modeling of coverage using a sliding window to determine presence/absence of the SARS-CoV-2 virus, genetic variant calling, viral clade estimation, and phylogenetic tree generation. COVID-DX combined Cromwell, WDL, Docker, and GATK Best Practices on the Microsoft Azure cloud.

ALGORITHM

Model fitting We fit deterministic sigmoidal functions to percent genome coverage at varying levels of depth using formula **[Equation 1]** where a_j is the exponential rate of increase at depth j, b_j is the inflection point of the function at depth j, $y_{i,j}$ is the coverage of sample i at depth j, x_i is the Ct value of sample i, and the carrying capacity of the sigmoidal function is one. The sigmoidal function is subtracted from one to model Ct value, which has an inverse logarithmic relationship with viral titer. We also implemented a Gaussian-shape to the standard deviation of the error profile to this function, such that variation would be highest at the inflection point and lowest at the tails of the sigmoidal function, using formula **[Equation 2]**, where b_j is the inflection point of the sigmoidal function at depth j, x^* is the Ct value, s_j is the standard deviation of the Gaussian curve at depth j, and s_j is the standard deviation of the normally-distributed error profile around the value x^* at depth j. The deterministic sigmoidal function was first fit to the data using nonlinear least-squares and the "port" algorithm in R. To minimize the effect of outliers, a random sampling consensus (RANSAC) algorithm was implemented; further, only coverage values in the range (0.0001, 0.9999) were used such that the algorithm would fit to the points in the middle of the sigmoidal curve and not at the asymptotes. Finally, a normally distributed error profile with Gaussian-shaped standard deviation was fit using maximum likelihood estimation (MLE). The likelihood equation **[Equation 3]** was maximized using the "Brent" method in R with the value of s constrained with boundaries (1,10) to avoid incorrectly inflated error profiles and flattened sigmoidal curves. Separate functions were fit to coverage values at varying levels of read depth j to adequately capture differentiation at both high and low Ct values.

$$y_{i,j} = 1 - \frac{1}{1+e^{-a_j*(x_i-b_j)}} \quad [1]$$

$$\sigma_j = e^{-\frac{1}{2}\left(\frac{x^*-b_j}{s_j}\right)^2} \quad [2]$$

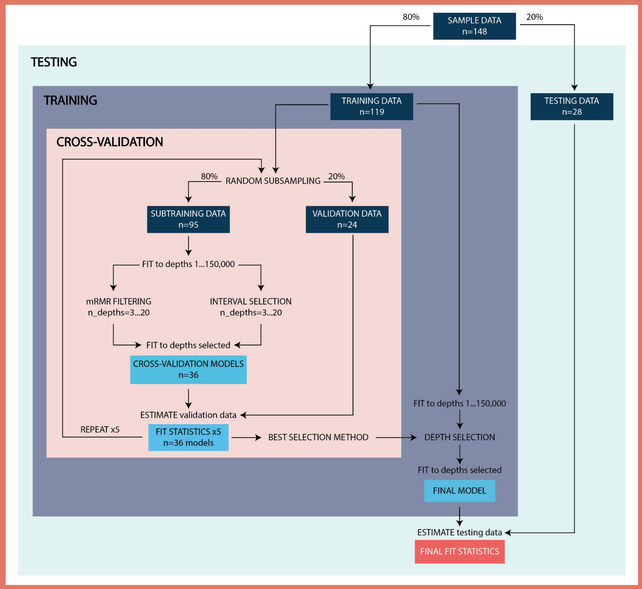
$$y_{i,j} \sim Normal\left(1 - \frac{1}{1+e^{-a_j*(x_i-b_j)}},\ e^{-\frac{1}{2}\left(\frac{x_i-b_j}{s_j}\right)^2}\right) \quad [3]$$

$$p_j^* = \begin{cases} Prob(y_{i,j} \le y^*), & Ct^* \le b \\ Prob(y_{i,j} \ge y^*), & Ct^* > b \end{cases} \quad [4]$$

$$NLL^* = -\log \prod_{j=1\ldots n} p_j^* \quad [5]$$

$$2 * \log\left(\frac{L(Ct_{GI})}{L(Ct_{est})}\right) \le 3.84 \quad [6]$$

Estimation of Ct values To compile all of the coverage models at different levels of read depth into a single estimator of Ct, we again used a likelihood-based method. Given the parameters estimated for the sigmoidal function and error profile during model fitting, we calculated the likelihood that a particular Ct value would generate the set of observed coverage values for a sample at the levels of read depth chosen for the models. For each depth model, this was calculated by estimating the probability p_j^* of generating the observed value of coverage at depth j, $y_{i,j}$, for any value Ct^* according to the distribution $y_{i,j} \sim$ Normal(y^*, $s(x^*)$), with the standard deviation of the normally distributed error calculated at x^*, **[Equation 4]**. Finally, the negative log-likelihood (NLL^*) was calculated by finding the negative log-product of all p_j^* for all depth levels $\{j_1 \ldots j_n\}$, **[Equation 5]**. To find the value Ct_{est} with the highest likelihood of generating the observed data Y_j, this function was minimized using the *optimize* function in R. Again, only values in the range (0.0001, 0.9999) were included to minimize the effect of asymptotic values on the estimate. Confidence intervals were estimated by finding the Ct values to the left and the right of Ct_{est} such that a likelihood ratio test (LRT) compared to Ct_{est} was significant with $p \le 0.05$, such that **[Equation 6]**. This value was estimated using the *uniroot* function in R.



IMPLEMENTATION

Model selection Our training set consisted of 80% randomly selected samples from our full data set (n=119). In order to perform the cross-validation and depth selection, we first fit our model (Equations 1-3) to every depth value between 1 and 150,000 one at a time to five subtraining sets selected using random subsampling (a further 80% split to the training data, n=95). This resulted in r^2 values for each depth from 1 to 150,000 for each of the five cross-validations. Of these, we performed our two selection methods: mRMR filtering and logarithmic interval selection. Each selection method was performed to select between 3 and 20 depth values to use in the final model for testing. This resulted in the comparison of a total of 36 models (Figure 1) five separate times. The selected depths for each of the 36 models were those that were included in the likelihood estimator of Ct values (Equations 4-6) used on each of the five validation sets. To evaluate each model, we calculated the r^2 value and the MAPE of the Ct estimates of the validation sets (the remaining 20% in each cross) compared to the laboratory-measured Ct values. Models were ranked according to r^2 and MAPE values and the model with the best rank sum was chosen as the best model. Our final model was the one that used the logarithmic interval selection method to select 19 depths, which resulted in an average r^2 value of 0.625 and MAPE of 11.88% across the five validation sets. The final model included depth set [2, 4, 8, 15, 18, 41, 87, 119, 275, 364, 665, 1962, 2189, 4788, 10193, 13292, 31542, 81041, 82391].

Training The entire training set (n=119) was fit to these depths using Equations 1-3.

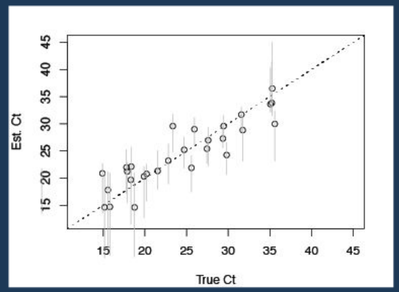
Testing The random sample of 20% of our sequenced samples that were initially withheld as a final independent testing set (n=29) was then tested on this model using Equations 4-6. Samples with laboratory-measured Ct values less than the minimum Ct value or greater than the maximum Ct value within the training set (<13.19 and >39.40) were excluded (n=1). Our final model resulted in a r^2 value of 0.80 and a MAPE of 10.5%.

CONCLUSIONS

In this study we present a novel method to estimate viral titer of SARS-CoV-2 using NGS data incorporating coverage across the viral genome. We built a robust model using a likelihood approach, and we were able to calculate 95% confidence intervals for our Ct value estimates, which we found capture 100% of our laboratory test sample Cts. To our knowledge, this is the first quantitative model for SARS-CoV-2 NGS data. This new approach allows for simultaneous detection of genetic variants and viral quantification in one single assay.

There are only a handful of studies that have used NGS data to quantify pathogens. These studies have focused on quantifying human immunodeficiency virus (HIV) and human papillomavirus (HPV), and been limited to correlating the number of viral sequence reads to viral titer (PCR cycle threshold or Ct value). This approach can only approximate viral titer with wide confidence intervals and is undermined when read duplication during PCR amplification results in uneven genome coverage.

Given the proportion of the genome covered at varying depths for a single sample as input data, our model estimates the Ct of that sample as the value which produces the maximum likelihood of generating the observed genome coverage data. We propose that such an algorithm in combination with genome variant identification will have important implications for cohorting clinical patients, monitoring disease and response to therapy, and supporting critical studies of vaccine and therapeutic efficacy against numerous SARS-CoV-2 variants.



Above: Results of the final model fit to the testing data. Each point is shown with lines representing 95% confidence intervals. The dotted line is the 1:1 line. The laboratory-measured Ct ("true" Ct) is shown on the x-axis and the Ct value estimated by the model is shown on the y-axis. The final model resulted in an R2 value of 0.80 and a MAPE of 10.5%.

LIMITATIONS AND FUTURE WORK

Next step will be to increase the number of samples we are using for training and testing our model. Future studies are needed to collect clinical metadata in relation to the viral titer that enable monitoring of disease outcome, therapeutic and vaccine responses.

REFERENCES

Nagy-Szakal et al. Microbiology Spectrum, 2021. Targeted Hybridization Capture of SARS-CoV-2 and Metagenomics Enables Genetic Variant Discovery and Nasal Microbiome Insights.

ACKNOWLEDGEMENTS

The COVID-DX software pipeline work used CromwellOnAzure, the Microsoft Genomics supported implementation of the Broad Institute's Cromwell workflow engine on Azure. The quantification model work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant (ACI-1548562). Specifically, it used the Bridges system, which is supported by NSF award (ACI-1445606), at the Pittsburgh Supercomputing Center (PSC). This work used resources of the COVID-19 HPC Consortium. We thank Nicholas Nystrom, Paola Buitrago, Julian Uran, David O'Neal and collaborators for their assistance with PSC resources access, use, and software installation support. We also thank Zaineb Bello and Caitlin Otto for their laboratory operation support; the Abesse Team, Pierre Davidoff, Shay David and Cory Mason for IT support, and John Papciak for the conference preparation.

DISCLAIMER: HLW, JEB, MCR XJS, MD, BAC, CEM, DNS and NBO are employees or consultants at Biotia Inc.