SARS-CoV-2 NGS Assay Powered by Biotia COVID-DX Software

Dorottya Nagy-Szakal[1,2], Mara Couto-Rodriguez[1], Joseph Barrows[1], Heather Wells[1], Marilyne Debieu[1], Courteny Hager[1], Kristin Butcher[3], Siyuan Chen[3], Robert J. Boorstein[4], Christopher E. Mason[1,5], Niamh B. O'Hara[1,2]

[1]Biotia Inc., New York, NY, USA
[2]SUNY Downstate Health Sciences University, The Department Cell Biology/College of Medicine, New York, NY, USA
[3]Twist Bioscience, South San Francisco, CA, USA
[4]Lenco Diagnostic Laboratories, Inc., New York, NY, USA
[5]Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University, New York, NY, USA

Contact us!
Dorottya Nagy-Szakal, MD, PhD
nagy-szakal@biotia.io
+1.832.316.4949
biotia.io

ABSTRACT

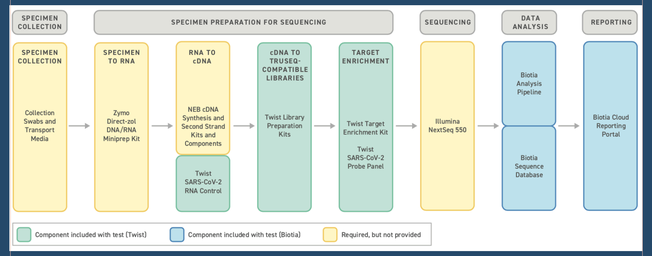
Background: COVID-19 has quickly spread throughout the world, causing an international public health emergency with an alarming global shortage of COVID-19 diagnostic tests. We developed and clinically validated a next-generation sequencing (NGS)-based target enrichment assay (Twist Bioscience) with the COVID-DX Software (Biotia) for the detection, characterization, and surveillance of the SARS-CoV-2 viral genome.

Methods: The SARS-CoV-2 NGS assay consists of components including library preparation, target enrichment, sequencing, and a COVID-DX Software analysis tool. Library preparation starts with extracted RNA from nasopharyngeal (NP) swabs followed by cDNA synthesis and conversion to Illumina TruSeq-compatible libraries using the Twist Library Preparation Kit via Enzymatic Fragmentation and Unique Dual Indices (UDI). The library is then enriched for SARS-CoV-2 sequences using a panel of dsDNA biotin-labeled probes, specifically designed to target the SARS-CoV-2 genome and sequenced on an Illumina NextSeq 550 platform. The COVID-DX Software analyzes sequence results and provides a clinically oriented report, including the presence/absence of SARS-CoV-2 for diagnostic use. An additional research use only report describes the assay performance, coverage across the viral genome, genetic variants, and phylogenetic analysis. Additionally, we generated Nextera Flex DNA sequencing libraries and utilized the BIOTIA-DX Pipeline to assess the microbiome composition, antimicrobial resistance (AMR) profiles, and virulence factors of subset samples.

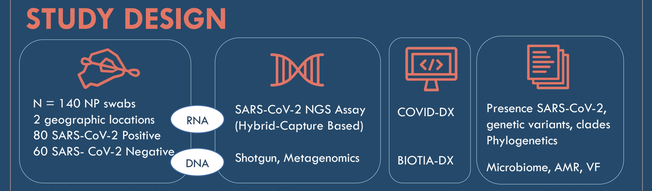
Results: The SARS-CoV-2 NGS Assay was validated on 60 positive and 60 negative clinical samples. To measure the sensitivity and specificity of the assay, the positive and negative percent agreement (PPA, NPA) was defined in comparison to an orthogonal EUA RT-PCR assay (PPA [95% CI]: 95.2% [90 %-100%] and NPA [98.3% CI]: 100% [95.2%-100%]). Data reported using our assay defined the limit of detection to be 800 copies/ml using heat-inactivated SARS-CoV-2 viral genome in clinical matrices. *In-silico* analysis provided >99.9% coverage across the SARS-CoV-2 viral genome and no cross-reactivity with evolutionarily similar respiratory pathogens. We identified new mutations, including 26 in the spike protein. Metagenomic analysis revealed 8 taxa significantly increased in COVID-19-positive patient samples.

Conclusion: The SARS-CoV-2 NGS Assay powered by the COVID-DX Software can be used to detect the SARS-CoV-2 virus and provide additional insight into genetic variants to track transmission, stratify risk, predict outcome and therapeutic response, and control the spread of infectious disease.

WORKFLOW ASSAY + SOFTWARE



STUDY DESIGN



SARS-CoV-2 NGS ASSAY: The SARS-CoV-2 NGS Assay is a highly sensitive nucleic acid hybridization capture-based assay used for the detection, characterization, and environmental monitoring of the SARS-CoV-2 virus. It utilizes Twist Bioscience's unique ability to rapidly develop virus-specific panels by DNA synthesis, and Biotia's comprehensive software and reporting capabilities (COVID-DX (v1.0)).

COVID-DX: The COVID-DX Pipeline includes 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 with a background of 3,365 global samples (GISAID). COVID-DX combines Cromwell, WDL, Docker, and GATK Best Practices on the Microsoft Azure cloud.

BIOTIA-DX: The BIOTIA-DX is a metagenomic analysis tool that identifies and estimates the abundance of organisms present in environmental or clinical samples using a lightweight data structure based on k-mers.

VALIDATION AND PERFORMANCE

To assess inclusivity *in silico*, we performed BLASTN alignment using 994 probes against 3.4 million viral nucleotide sequences from NCBI Virus (including ~50,000 SARS-CoV-2 genome nucleotide sequences from GISAID). Our analysis identified 50,987 sequences in the GISAID database with high identity matches (150 sequences with 100% mean identity and 50,816 sequences with ≧80% mean percent identity to our probes). False positives occurred only in closely conserved viruses such as bat coronavirus, infectious bronchitis virus, and transmissible gastroenteritis virus strains. Additional alignment to 26 clinically significant microbial genomes and the human genome showed no homology (≧80%) between the cohort genomes and the probes, enabling the use of this combined genome to filter off-target reads.

Discovery sample set
30+ and 30- NP swabs

Validation sample set
30+ and 30- NP swabs

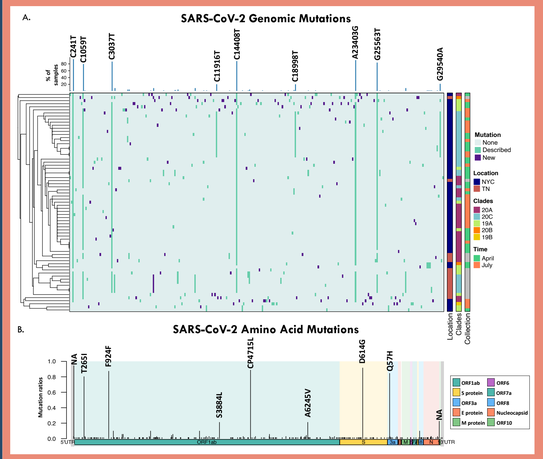
Analytical sensitivity
LoD 800 copies/ml

Inclusivity
Cross-reactivity

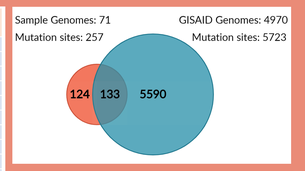
Clinical evaluation

The LoD (analytical sensitivity) was determined to be 800 copies/ml.
The positive and negative percent agreement was calculated in relation to the EUA RT-PCR comparator method with the combined discovery and independent validation set (n=120; PPA [95% CI]: 95.2% [90 %-100%] and NPA [98.3% CI]: 100% [95.2%-100%]).

SARS-CoV-2 GENETIC VARIANTS



Gene	Protein	New Mutations Compared to GISAID
Orf1ab	NSP2	13
Orf1ab	NSP3	25
Orf1ab	NSP4	2
Orf1ab	NSP5	5
Orf1ab	NSP6	1
Orf1ab	NSP9/Replicase	4
Orf1ab	NSP10/RNA synthase	1
Orf1ab	NSP12/RNAdepRNApol	11
Orf1ab	NSP13/Helicase	6
Orf1ab	NSP14/NSP11	6
Orf1ab	NSP15/NSP11	1
Orf1ab	NSP15/EndoRNAse	2
Orf1ab	NSP16/NSP13	5
S gene	Spike (RBD only)	26 (7)
Orf3a	Orf3a Protein	1
E/M Linking	NA: E/M Linking Region	1
M gene	Membrane/Matrix Protein	2
Orf6	Orf6 Protein	3
Orf8	Orf8 Protein	1
N gene	Nucleocapsid	8



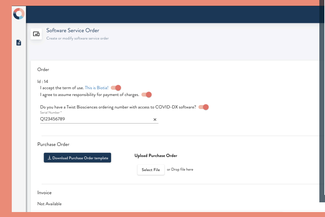
124 new mutations identified, including 26 in the spike protein

71 SARS-CoV-2 positive samples were collected from 2 geographic locations (NY, TN) and processed by the SARS-CoV-2 NGS Assay, including COVID-DX Software

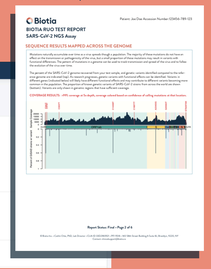
- 733 mutations were detected at 257 different mutation sites
- 124 new mutations were identified, the majority located at Orf1ab and 26 at the spike protein
- Samples collected from 2 geographic locations share similar genetic mutations

Nucleotide Mutation	AA Mutation	Geographic Location	Frequency in Samples	Frequency in GISAID	Gene	Protein Name	Synonymity	AA Alt Properties
Most Frequent Mutations in All Samples								
A23403G	D614G	All	65	11	S gene	Spike	non-synonymous	radical
C241T	extogenic	All	67	15	5'UTR	NA	NA	NA
C14408T	P4715L	All	63	12	ORF1ab	NSP12/RNApol	non-synonymous	radical
C3037T	F924F	All	62	12	ORF1ab	NSP3	synonymous	NA
G25563T	Q57H	All	60	12	ORF3a	ORF3a Protein	non-synonymous	conservative
C1059T	T265I	All	57	7	ORF1ab	NSP2	non-synonymous	conservative
G29540A	extrogenic	All	16	5	N/Orf10 Linking Region	NA	NA	NA
C11916T	S3884L	All	15	2	ORF1ab	NSP7/Replicase	non-synonymous	radical
C18998T	A6245V	All	15	1	ORF1ab	NSP14/NSP11	non-synonymous	conservative
Unique Mutations at Each Geographic Location								
G29540A	extrogenic	NY	16	5	N/Orf10 Linking Region	NA	NA	NA
C11916T	S3884L	NY	15	2	ORF1ab	NSP7/Replicase	non-synonymous	radical
C18998T	A6245V	NY	15	1	ORF1ab	NSP14/NSP11	non-synonymous	conservative
C27964T	S24L	NY	7	3	ORF8	ORF8 Protein	non-synonymous	radical
C4113T	A1283V	NY	4	2	ORF1ab	NSP3	non-synonymous	conservative
C3411T	A1049V	TN	7	1	ORF1ab	NSP3	non-synonymous	conservative
T6394C	D2043D	TN	7	0	ORF1ab	NSP3	synonymous	NA
C10319T	L3352F	TN	7	4	ORF1ab	NSP5	non-synonymous	conservative
A4197G	E1311G	TN	4	1	ORF1ab	NSP3	non-synonymous	radical
G8179A	R2638R	TN	4	3	ORF1ab	NSP3	synonymous	NA
C13924T	Y5220Y	TN	4	1	ORF1ab	NSP12/RNApol	synonymous	NA

User Portal to order tests, upload files for analysis, download report



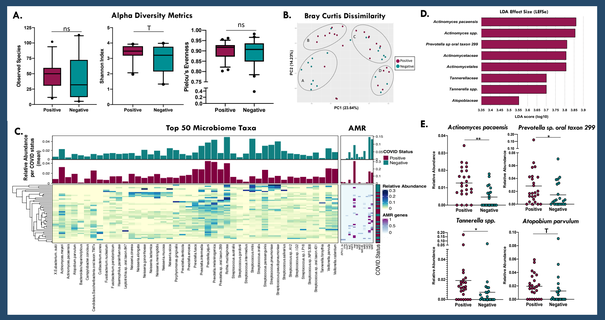
COVID-DX Report



SEQUENCING METRICS

- For the SARS-CoV-2 NGS Assay, an average of 15.1M and 1.5M reads were obtained from positive and negative samples, respectively. HS metrics showed an average of 43% on target reads, ranging from 0.005% to 99.5%. Additionally, our target enrichment approach yielded an average fold enrichment of 46791x, ranging from 5.9x to 108,602x. Subsampling to 500,000 reads per sample did not significantly change fold enrichment (mean: 108598x) or on-target reads (mean: 40.3%).
- As for NP metagenomics, an average of 10M and 7.3M reads were obtained from positive and negative samples, respectively. After removal of reads mapping to human DNA, an average of 1.7M (CVP) and 1.9 M (CVN) microbial reads were used for further analysis.

METAGENOMICS



A total of 106 metagenomic libraries were prepared with Illumina's Nextera Flex Kit, sequenced, and subsequently processed with the BIOTIA-DX Software. After human sequences removal, only 21 COVID-19 negative (CVN) samples and 25 COVID-19 positive (CVP) samples were further analyzed for their microbiome composition, AMR profiles, and virulence factors.

Alpha Diversity metrics showed no significant difference in richness or evenness between CVP and CVN samples (Panel A). Bray Curtis Dissimilarity Index (Panel B) exhibited 4 clusters, of which cluster A and cluster B are largely dominated by CVN and CVP samples, respectively. Overall, the microbiome profiles are in concordance with previously reported respiratory microorganisms, such as *Streptococcus, Veillonella*, *Rothia*, and *Prevotella species* (Panel C). Linear Discriminant Analysis Effect Size (LEfSe) revealed 8 bacterial taxa that are significantly increased in CVP samples (Panel D), of which 3/4 species-level taxa were further confirmed to be significant when comparing their relative abundances using a Mann-Whitney U- test (Panel E). Additionally, a total of 13 AMR genes were detected among all samples (CVP = 19, CVN = 12) with *mel*, which confers macrolide resistance, being significantly overrepresented in CVP samples.

CONCLUSIONS

The SARS-CoV-2 NGS Assay powered by the COVID-DX Software can be used to detect the SARS-CoV-2 viral RNA in clinical settings and provide additional insight into genetic variants to track transmission, stratify risk, and predict outcome and therapeutic response.

Our approach allowed the identification of 124 new genetic mutations, including 26 in the spike protein, in addition to mapping genetic mutations in two geographic locations with different mutation patterns.

The metagenomic study revealed alteration of the respiratory tract microbiome and AMR profile in COVID-19 positive patient samples that needs to be further analyzed in correlation with extensive clinical metadata.

NGS-based genetic epidemiology and infectious disease diagnostics are valuable assets to fight against the current COVID-19 pandemic and control the spread of infectious diseases.

FUTURE WORK

We are confirming the novel variants using a combination of bioinformatic tools and additional sequencing. Further work includes building a maximum likelihood model to estimate viral titer, additional optimization to achieve faster turnaround time with built-in automation, increase specificity and sensitivity, and extension to saliva and other specimens. We are expanding our probe panel to detect other respiratory pathogens for use in clinics and research.

Biotia focuses on enabling personalized, data-driven pathogen discovery from the individual to the community level by providing precision infectious disease diagnostics, NGS-guided patient treatment, global databases, and predictive analysis, with the ultimate goals of preserving the efficacy of vaccines and antimicrobials and moving toward the eradication of death due to infectious diseases.

ACKNOWLEDGEMENTS

We thank Agnes Berki, PhD, for her work on the SARS-CoV-2 genetic variants and the University of Tennessee (Colleen B Johnson, PhD, Shruti Bansal, Mariah K Taylor) for providing clinical samples and running orthogonal qPCR testing.