Title: SARS-CoV-2 Whole Genome Sequencing Data Collection and Analysis Guidelines

Study Number: N/A

Parent Compound Number(s): PF-07302048

Alternative Compound Identifiers: N/A

Author

Qi Yang, PhD
Associate Research Fellow, Viral Vaccines

Zhenghui Li, PhD
Senior Scientist, Bacterial Vaccines & Technology

Approvers

Kena Swanson, PhD
Director, Viral Vaccines

Warren Kalina, PhD
Senior Director, Clinical & Diagnostic Assay Development

Paul Liberator, PhD
Senior Director, Bacterial Vaccines & Technology

(b) (6)
(b) (6)

Pfizer Vaccine Research and Development
401 N. Middletown Rd.
Pearl River, NY

TABLE OF CONTENTS

LIST OF TABLES 2
LIST OF FIGURES 2
1. BACKGROUND 3
1.1. Targeted NGS for SARS-CoV-2....................3
1.2. Objective 3
2. MATERIALS AND METHODS....................3
2.1. Nucleic Acid Extraction 3
2.2. Ion Torrent....................4
2.2.1. Ion Torrent Library Preparation and Sequencing 4
2.2.2. Ion Torrent Bioinformatics Workflow....................4
2.3. Illumina 5
2.3.1. Illumina Library Preparation and Sequencing....................5
2.3.2. Illumina Bioinformatics Workflow 5
2.4. Lineage Classification 6
2.5. Rules for Lineage Assignment 6
3. COMPARISON OF ION TORRENT AND ILLUMINA PLATFORMS....................6
4. WORKFLOW AND DATA ENTRY AND STORAGE....................9
4.1. Workflow and LIMS Upload Process 9
4.2. Data Repository....................9
5. REFERENCES 10

LIST OF TABLES

Table 1. SARS-CoV-2 Lineage Assignments of PCR Positive NP Swabs from Ion Torrent and Illumina NGS Platforms 7

LIST OF FIGURES

Figure 1. Workflow for Capturing SARS-CoV-2 Sample Sequence Data 9

1. BACKGROUND

1.1. Targeted NGS for SARS-CoV-2

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus that causes coronavirus disease 2019 (COVID-19), the illness responsible for the COVID-19 pandemic. SARS-CoV-2 is a positive-sense single-stranded RNA virus approximately 30,000 bases in length. Given the opportunity, RNA viruses have a propensity to evolve - hundreds of thousands of distinct SARS-CoV-2 viral lineages have been identified using next generation sequencing (NGS) platforms. Some of these are now classified as either variants of concern (VOC) or variants of interest (VOI) based on disease characteristics, epidemiology, and other factors. NGS provides an effective, unbiased way to identify new coronavirus variants. This information is valuable to understand viral evolution, geographic distribution, and transmission. The sequence data may be used to inform public health decisions, mitigate viral spread, and/or formulate recommendations for next generation vaccines.

The viral content in clinical specimens that are PCR positive in SARS-CoV-2 molecular diagnostic assays can be highly variable. In addition, the composition of nucleic acid extracted from PCR positive clinical specimens is complex, including contributions from the human host and microbes in addition to SARS-CoV-2. For these reasons, a shotgun metagenomics approach for determination of SARS-CoV-2 genome sequence from clinical specimens is not effective. Amplicon-based approaches that enrich for SARS-CoV-2 content prior to NGS library construction are a necessary alternative. Following random primed cDNA synthesis of total RNA in the nucleic acid extracted from a clinical specimen, selective SARS-CoV-2 enrichment is achieved by multiplex PCR using oligonucleotides to generate amplicons that are tiled across the viral genome. A library of these PCR amplicons is constructed for each sample and then sequenced using NGS platforms.

1.2. Objective

The objective of this document is to describe the process used to determine the viral sequence and lineage in SARS-CoV-2 RT-PCR (Cepheid) positive specimens from clinical study C4591001.[1] Determination of SARS-CoV-2 lineage is an exploratory study objective.

Capturing SARS-CoV-2 sequence information from a SARS-CoV-2 PCR positive specimen can be achieved using different sequencing platforms, two of which are manufactured by Ion Torrent and Illumina.

2. MATERIALS AND METHODS

2.1. Nucleic Acid Extraction

Midturbinate swabs that are positive in the SARS-CoV-2 RT-PCR (Cepheid) assay at either the N or E gene target are advanced for viral sequence. Nucleic acid extraction is performed using the MagMAX™ Viral/Pathogen Ultra Nucleic Acid Isolation Kit processed on a KingFisher Flex or KingFisher Presto.

2.2. Ion Torrent

2.2.1. Ion Torrent Library Preparation and Sequencing

The SARS-CoV-2 viral genome sequencing is performed manually using the Ion AmpliSeq™ technology and the Ion GeneStudio™S5 plus system.

The Ion AmpliSeq™ SARS-CoV-2 Research Panel consists of two primer pools that target 237 PCR amplicons specific to SARS-CoV-2 and 5 human expression controls. Oligonucleotide primers based upon available SARS-CoV-2 nucleotide sequences direct the amplification of the viral genome with amplicon lengths of 125-275 bp. The panel provides greater than 99% coverage of the SARS-CoV-2 genome (~30 kb), achieving detection limits as low as 20 viral copies.

Briefly, SARS-CoV-2 viral RNA content in the nucleic acid purified from swabs is quantified using TaqMan™ 2019-nCoV Assay Kit v1, the TaqMan™ 2019-nCoV Control Kit v1, and the TaqPath™ 1-Step RT-qPCR Master Mix, CG to determine the optimal number of target amplification cycles. cDNA is synthesized with the SuperScript VILO cDNA synthesis Kit. Libraries are prepared using the Ion AmpliSeq™ library kit plus[2] and Ion AmpliSeq™ SARS-CoV-2 research assay panel according to the manufacturer's instructions.[3] The prepared library undergoes template preparation with the Ion Chef according to the manufacturer's instructions.[4] The enriched templates are then loaded onto an Ion 530 chip for semiconductor sequencing on the Ion GeneStudio™ S5 plus sequencer according to the manufacturer's instructions.[4]

2.2.2. Ion Torrent Bioinformatics Workflow

Raw sequencing reads generated by the Ion Torrent sequencer are quality and adaptor trimmed by Ion Torrent Suite and the resulting reads are then mapped to the complete genome of the SARS-CoV-2 Wuhan-Hu-1 isolate (GenBank accession number MN908947.3) using TMAP 5.14.0.

Variant calling is carried out with the Torrent Variant Caller using the BAM file from the mapping of the cleaned sequence reads to the reference sequence of SARS-CoV-2. Samples not achieving mean sequence coverage of (b) (4) (b) (4) [redacted] (b) (4) uniformity do not receive a lineage assignment and are designated as indeterminant (IND). Samples designated as IND are submitted for NGS using the Illumina platform (Section 2.3).

All identified SARS-CoV-2 nucleotide variants are annotated using the (b) (4) software.[5] For each of the variants identified, the output consists of: 1) the nucleotide of the reference at each position and the alternative sequence, 2) the codon of the reference and the alternative codon if the nucleotide variant results in a nonsynonymous substitution, and 3) the nucleotide translation and information about the mutation (synonymous, missense plus deletions).

For each sample, a whole-genome sequence is generated in a FASTA file using IRMA (Iterative Refinement Meta-Assemble) and iterative optimization of read gathering and assembly.

2.3. Illumina

2.3.1. Illumina Library Preparation and Sequencing

Briefly, nucleic acid purified from swabs is digested with DNase using the Invitrogen TURBO DNA-free™ Kit (AM1907) followed by RNA purification using the Qiagen RNeasy MinElute Cleanup Kit (74204). Initially performed manually, these steps (DNase digestion and subsequent RNA purification) have since been programmed as additional steps at the end of the primary RNA extraction using the KingFisher Presto (Section 2.1). Synthesis of cDNA is performed using random sequence primers according to the AmpliSeq for Illumina On-Demand, Custom and Community Panels Reference Guide.[6]

The cDNA is used as template to specifically enrich for SARS-CoV-2 content by PCR. The AmpliSeq for Illumina SARS-CoV-2 panel of PCR primers is a 2-pool design, containing a total of 247 amplicons/primer pairs (Pool 1: 125 amplicons, Pool 2: 122 amplicons). These include 242 SARS-CoV-2 viral-specific targets and 5 human gene expression controls. PCR products range in size from 125-275 bps in length and cover > 99% of the viral genome and all potential lineages of the virus. Oligonucleotide primers directing the amplification of the viral amplicons are based upon available SARS-CoV-2 nucleotide sequences. Universal Next Generation Sequencing Adaptors are ligated to the ends of the SARS-CoV-2 AmpliSeq amplicons. The amplicon libraries are purified with magnetic beads and loaded to a flow cell for sequence determination using the Illumina NextSeq instrument according to the manufacturer's instructions.

2.3.2. Illumina Bioinformatics Workflow

A FASTQ file for each sample containing sequence data from the clusters that pass filter is imported into CLC Genomic Workbench version 20.0.3. Any read with header information that has a "failed" flag due to a poor-quality score is removed during the file import process by CLC Genomics Workbench. Primary reads are then aligned to the SARS-CoV-2 Wuhan-Hu-1 reference genome (Genbank accession MN908947.3) using the "Map Reads to Reference" function. Consensus SARS-CoV-2 sequence is generated by using the "Extract Consensus Sequence" function (b) (4) The coverage of the Spike gene is verified by using the "QC for Targeted Sequencing" function. Only isolates that have (b) (4) coverage at each nucleotide position across the entire Spike gene are advanced for lineage assignment. Single nucleotide variants (SNV) are called using the "Low Frequency Variant Detection" function (b) (4) (b) (4) Samples that do not meet these acceptance criteria are considered indeterminant (IND) and not advanced for lineage assignment. Samples designated as IND are submitted for NGS using the alternative Ion Torrent platform (Section 2.2).

Nucleotide sequence coding for the Spike protein is extracted and translated into amino acid sequence followed by alignment to the Spike amino acid sequence of the Wuhan-Hu-1 strain (Genbank accession number MN908947.3) using the CLC Genomics Workbench. Non-synonymous substitutions relative to the Wuhan-Hu-1 Spike sequence are recorded.

2.4. Lineage Classification

SARS-CoV-2 lineage assignment for data from both Ion Torrent and Illumina NGS platforms is based on Pangolin (b) (4) software,[7] which runs a multinomial logistic regression model trained against lineage assignments based on isolate data from GISAID, a global science initiative established in 2008 that provides open-access to genomics data of influenza virus and SARS-CoV-2.[8]

Viral lineage is determined for samples sequenced using either the Ion Torrent or Illumina NGS platforms, and in some cases using both platforms.

2.5. Rules for Lineage Assignment

SARS-CoV-2 lineage designations are determined from NGS data using either the Ion Torrent or Illumina sequencing platforms. If acceptance criteria described in Sections 2.2.2 and Section 2.3.2 are met, a SARS-CoV-2 lineage is assigned and entered into the database. If acceptance criteria are not met using the initial sequencing platform the result is considered indeterminant (IND), and the sample is routed for assay using the alternative platform. If the NGS determined using the alternative platform meets acceptance criteria, a SARS-CoV-2 lineage is assigned and entered into the database. If the NGS result is considered IND using both platforms (ie, data do not meet acceptance criteria for either platform) and the Cepheid RT-PCR Ct value for that sample is ≤ 34 for either the N or E gene target, the sample is entered into the database as IND.

A sample is designated QNS (quantity not sufficient) if,

1. NGS quality does not meet the respective acceptance criteria (ie, sample is considered IND) for both the Ion Torrent (Section 2.2.2) and Illumina (Section 2.3.2) platforms, and

2. The sample has a Cepheid RT-PCR Ct value greater than 34 for both the N and E gene target.

3. COMPARISON OF ION TORRENT AND ILLUMINA PLATFORMS

An initial group of (b) (4) Cepheid RT-PCR positive swab specimens from clinical study C4591001 were processed as described above, using each of the two NGS platforms to evaluate their suitability in determination of SARS-CoV-2 NGS and viral lineage. Selected RT-PCR negative swabs were also included as negative controls. Sample barcodes, Ct values for the N and E gene amplicons from the Cepheid PCR assay, and SARS-CoV-2 lineage assignments determined from nucleotide sequence collected using the Ion Torrent and Illumina NGS platforms are listed in Table 1. The Illumina NGS did not meet acceptance criteria for (b) (4) of the samples (22.5%). Several of these had very high Cepheid PCR Ct values (ie, low SARS-CoV-2 RNA content). Lineage assignments could not be made from Ion Torrent NGS for (b) (4) of the samples (14.1%); (b) (4 of these (b) (4) were also designated as IND from Illumina NGS.

Concordance between the two platforms was high; lineage assignments were the same in (b) (4) of the (b) (4) (83.1%) samples. In most instances ((b) (4)) the lack of concordance was in samples with limited viral RNA and where one of the two platforms did not achieve sufficient sequence coverage and received an IND designation. This analysis supported implementation of a rule where samples initially designated as IND on one platform would be flexed to the alternative NGS platform for evaluation provided that sufficient sample volume remained.

Two samples in the group of (b) (4) were assigned different SARS-CoV-2 lineages from data collected using the respective NGS platforms. The viral lineage determined for sample RR03351 was P.2 for Ion Torrent and B.1.1.28 for Illumina. These two lineages reside very close to one another on the SARS-CoV-2 phylogenomic tree, differing from one another by a small number of nucleotide changes. In the second case, the viral lineage determined for sample BM9F0G was B.1.351 for Ion Torrent and B.1 for Illumina.

The following rules for lineage assignment will be followed in these types of instances. (b) (4) (b) (4). Samples RR03351 and BM9F0G (b) (4) (b) (4) (b) (4) (b) (4) led to a B.1.351 lineage assignment for sample BM9F0G and P.2 for sample RR03351.

Table 1. SARS-CoV-2 Lineage Assignments of PCR Positive NP Swabs from Ion Torrent and Illumina NGS Platforms

Sample Barcodes	Cepheid E (Ct)[a]	Cepheid N (Ct)*	Ion Torrent Lineage	Illumina Lineage
(b) (4)				

Table 1. SARS-CoV-2 Lineage Assignments of PCR Positive NP Swabs from Ion Torrent and Illumina NGS Platforms

Sample Barcodes	Cepheid E (Ct)[a]	Cepheid N (Ct)*	Ion Torrent Lineage	Illumina Lineage

(b) (4)

Table 1. SARS-CoV-2 Lineage Assignments of PCR Positive NP Swabs from Ion Torrent and Illumina NGS Platforms

Sample Barcodes	Cepheid E (Ct)[a]	Cepheid N (Ct)*	Ion Torrent Lineage	Illumina Lineage
(b) (4)				

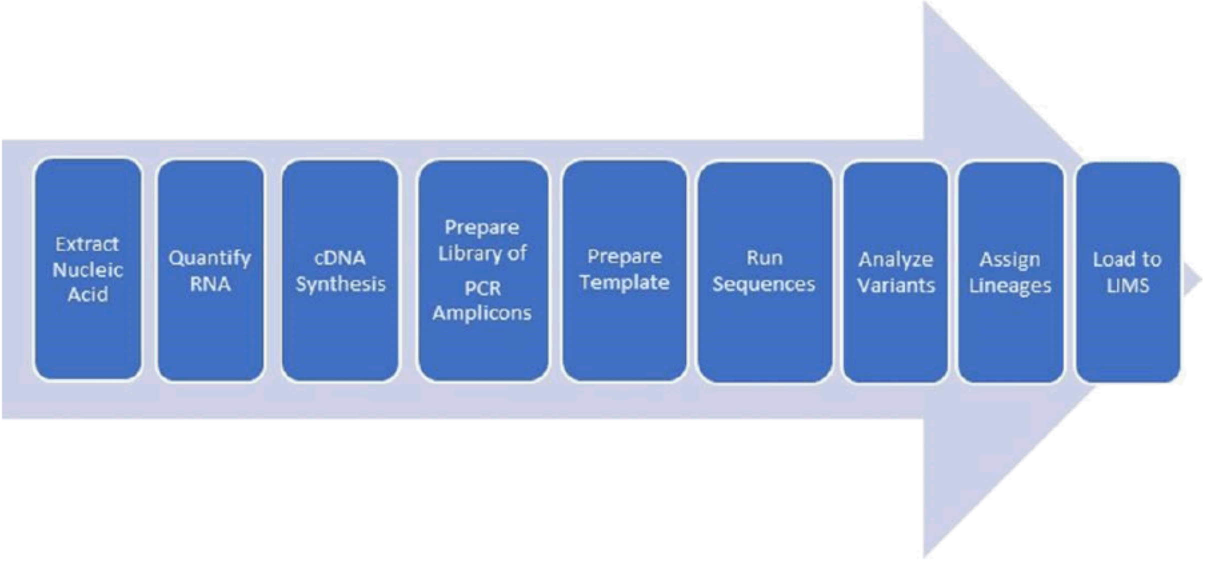
4. WORKFLOW AND DATA ENTRY AND STORAGE

4.1. Workflow and LIMS Upload Process

Figure 1 describes the general workflow for sample processing, data analysis, lineage assignment, and LIMS upload of the lineage assignment and associated barcode identifier.

The verified lineage and barcode data are tabulated in an Excel file. The barcode and lineage classification are documented in an electronic notebook and confirmed by a witness. The lineage is then manually entered into LIMS. Cross verification of LIMS entries is performed by a second colleague prior to data transfer to the clinical database.

Figure 1. Workflow for Capturing SARS-CoV-2 Sample Sequence Data



4.2. Data Repository

The primary data is deployed to the secure(b) (4) servers for long-term storage. (b) (4) (b) (4) (b) (4)

5. REFERENCES

[1] VR-VR-10080. Report on Method Validation of a Cepheid Xpert® Xpress PCR Assay to Detect SARS-CoV-2.

[2] ThermoFisher. Ion AmpliSeq™ Library Kit Plus USER GUIDE. Publication MAN0017003 version C.0.

[3] Ion AmpliSeq™ SARS-CoV2 Research Panel, Instructions – for use on an Ion GeneStudio™ S5 Series System. Publication MAN0019277, version B.0.

[4] ThermoFisher. Ion 510™ & Ion 520™ & Ion 530™ Kit – Chef USER GUIDE. Publication MAN0016854, version F.0.

[5] (b) (4)

[6] Illumina. AmpliSeq for Illumina On-Demand, Custom and Community Panels Reference Guide Document 1000000036408, v09 May 2020. Accessed from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/ampliseq-for-illumina/ampliseq-for-illumina-custom-and-community-panels-reference-guide-1000000036408-09.pdf

[7] Pangolin, https://github.com/cov-lineages/pangolin

[8] https://www.gisaid.org/

Document Approval Record

Document Name:	VR-VTN-10436
Document Title:	SARS-CoV-2 Whole Genome Sequencing Data Collection and Analysis Guidelines

Signed By:	Date(GMT)	Signing Capacity
Yang, Qi	03-Jun-2021 13:54:29	Author Approval
Liberator, Paul	03-Jun-2021 15:02:28	Final Approval
Li, Zhenghui	03-Jun-2021 15:06:44	Author Approval
Swanson, Kena	03-Jun-2021 21:13:34	Final Approval
(b) (6)	03-Jun-2021 23:50:48	(b) (6)
Kalina, Warren	04-Jun-2021 15:42:11	Manager Approval

090177e19732889a\Approved\Approved On: 04-Jun-2021 15:42 (GMT)