Submitting Next Generation Sequencing Data to the Division of Antiviral Products

Guidance for Industry

Technical Specifications Document

For questions regarding this technical specifications document, contact CDER at cder-edata@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)

July 2019
Technical Specifications

Revision History

Date	Version	Summary of Revisions
July 2019	1.0	Initial Version

Table of Contents

Contains Nonbinding Recommendations

Submitting Next Generation Sequencing Data to the Division of Antiviral Products

Guidance for Industry
Technical Specifications Document[1]

This guidance represents the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA office responsible for this guidance as listed on the title page.

1.0 Introduction

The purpose of this technical specifications document is to provide the current thinking of FDA's Division of Antiviral Products (the Division) in regard to the submission of next generation nucleotide sequence analysis procedures and data in support of resistance assessments for the development of antiviral drug products.

The Division performs independent analyses of all resistance data associated with antiviral drug products being developed to ensure that the emergence of resistance is carefully characterized and explained in the label of newly approved antiviral drug products. Providing accurate resistance information is imperative for protecting public health to prevent the emergence of novel resistant and cross-resistant viral variants that have the potential to infect others and cause major outbreaks of disease that cannot be controlled by approved drug products. In addition, the resistance information provides important guidance for health care professionals who oversee the use of antiviral drug products and is included in the drug product information approved by the Division. Moreover, the Division can request data for nucleotide sequence analysis of host-targeted genes for polymorphism analysis to determine if different population-based alleles have an effect on efficacy.

Next generation sequencing (NGS) is an emerging technology that sponsors frequently employ when performing sequence-based resistance analysis. Because this technology generates the nucleotide sequence for all RNAs or DNAs in a clinical sample, NGS adds complexity to the resistance analysis process. In contrast to Sanger nucleotide sequence analysis, which provides an average sequence of the virus population, NGS provides nucleotide sequence information for

[1] This technical specifications document has been prepared by the Division of Antiviral Products in the Center for Drug Evaluation and Research at the Food and Drug Administration. You may submit comments on this guidance at any time. Submit comments to Docket No. FDA-2017-D-6821 (available at https://www.regulations.gov/docket?D=FDA-2017-D-6821) (see the instructions for submitting comments in the docket).

individual viruses within a viral population, frequently providing millions or billions of short sequences per sample. The complexity of the data makes it challenging for reviewers to analyze and validate the sequence information particularly because there are currently no standardized bioinformatics analysis approaches for analyzing these large datasets.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

2.0 Acceptable NGS Platforms

The Division will accept nucleotide sequencing data generated from most standard NGS platforms provided the sponsor submits the appropriate details for the sequencing platform, the protocols used for sample preparation, the raw NGS data in fastq format, and the methods used to analyze the data. We recommend the sponsor communicate with the Division early in the process and provide these details before submitting the sequencing data. In addition, we recommend the sponsor submit a mock NGS dataset before any formal submissions to ensure that the appropriate data formats and processes are acceptable. Sponsors should consider the information in the following sections when preparing NGS submissions.

3.0 Summary of Information to Submit to the Division

3.1 NGS Protocol and Data Analysis Methods

The sponsor should submit a detailed protocol that describes sample processing and NGS analysis procedures. See sections 4.0, NGS Protocol, and 5.0, Describing NGS Data Analysis Methods and Reporting Results, for specific methods to be described.

3.2 Raw NGS Data in Fastq Format

The sponsor should provide all of the raw NGS data from each sequence run in the fastq format. An assembled read mapping can be submitted in .fas, .ace, .sam, or .bam formats, but this is optional. The sponsor should provide the appropriate reference sequences (with concurrence from the division) and accession numbers used for any reference mappings. For reference mapping to a baseline sample or gene of interest, the sponsor should provide the baseline or reference consensus sequence and state how this sequence was derived. For de novo assemblies, the sponsor should provide all contigs greater than 200 nucleotides in the fastq format.

3.3 Frequency Tables

The sponsor should provide a frequency table reporting all amino acid substitutions that differ from baseline at frequencies greater than or equal to 1 percent. See section 7.0, Frequency Table Example, for an example of a frequency table.

3.4 Summary Tables

The sponsor should consider the following:

a. Provide the data for each study in the analysis-ready virology resistance data format used for population-based sequencing showing the baseline and failure sequences and using blank cells when amino acids are identical to the reference sequence. We recommend that the sponsor consult with the Division to establish the sensitivity level for the resistance dataset (table populated with amino acid substitutions present at a specific frequency, i.e., 15 percent) or provide justification for the sensitivity level selected.

b. Provide a table presenting a high-level summary of the predominant substitutions for each study or study subgroup (i.e., by genotype, by dose).

3.5 Other Analysis-Ready Datasets (as applicable)

We recommend that the sponsor consult with the Division before submitting other analysis-ready datasets related to the resistance analysis of NGS data.

4.0 NGS Protocol

4.1 General Protocol Design Elements

Sponsors should include the following general protocol design elements in NGS protocols:

a. A description of the subjects, study time points, and sample matrices to be analyzed.

b. A description of the NGS platform to be used including all associated performance characteristics.

c. Target gene region name(s) and size(s) to be analyzed.

d. General analysis strategy (e.g., identify changes relative to a prototypical strain, compare sequences from different time points in the same subject).

e. The coverage level to be attempted. We recommend a target for coverage of greater than 5,000 reads. However, we recognize that some samples with lower viral loads may not produce assemblies at this level of coverage. Sponsors should identify samples that did not reach this level of coverage.

f. A description of the approach used to identify, filter, or process sequencing errors.

4.2 Sample Preparation

The key to reliable sequencing results is linked to template preparation, so that the sample being sequenced is representative of the population being analyzed. The following information should be described:

a. Methods for extracting nucleic acids from samples.

b. Methods for purifying viral sequences from contaminating background nucleic acids.

c. Methods for concentrating viral nucleic acids. Include the estimated target copy number input for reverse transcription polymerase chain reaction (RT-PCR) (viral RNA) or PCR (viral DNA) reactions for each sample.

d. Methods for denaturing secondary structure.

e. Methods for generating double stranded DNA (dsDNA). Include a description of the primers.

f. Methods for purifying dsDNA for sequencing.

g. Methods for NGS library preparation.

h. Methods for adding barcodes for multiplexing (when applicable).

NOTE: Many NGS protocols that have been published in the scientific literature have used a specific (amplicon) or nonspecific (random primers) PCR amplification stage to increase the concentration of dsDNA for sequencing. However, while this approach will amplify the predominant genotypes, it is not likely to detect the minor variants that may be important for the rise of resistance mutations over time. Therefore, any protocol that uses a PCR amplification step before NGS should provide evidence that amplifications are representative of the target population and minor variants would still be present in the NGS data. We recommend employing approaches that correct for PCR resampling bias and (RT-)PCR and sequencing error, such as complementary DNA barcoding.

5.0 Describing NGS Data Analysis Methods and Reporting Results

Submissions of sequence data must include a thorough description of the analysis pipeline used to analyze the sequencing dataset and the raw sequence information so that the Division can conduct an independent analysis of the data. The information in the following sections should be provided in these reports.

5.1 Summary Statistics for Each Sequence Run

The report should include the total number of reads sequenced per sequence run, the quality scores of the reads, the average length of the reads, a description of any error assessments

performed to reduce the effect of platform-specific errors, and information regarding what trimming parameters were used to clean up potential sequence errors or remove reads that failed to meet any quality or length thresholds.

5.2 A Description of How Sequence Barcodes Were Processed

The report should include a description of the methods used to add barcode sequences to the sequences of interest, the sequences of the barcode tags used to sequester reads after the sequence run, the program used to sort into bins the reads by barcode, a description of how barcode sequences were trimmed from the sequences, and a description of any precautions taken to eliminate potential cross-barcode contamination.

5.3 Contig and Mapping Reports

Two approaches can be used for NGS analyses of data to support the development of antiviral drug products: (1) mapping of short reads to a references sequence or (2) de novo assembly of short reads to assemble contigs.

5.3.1 Mapping sequence reads to a reference sequence and calling variants

Sponsors should consider the following for inclusion in the report on mapping sequence reads to a reference sequence and calling variants:

a. Identify the reference sequence(s) used for mapping and provide the nucleotide and amino acid sequences and an accession number.

b. Identify the program and describe the algorithm used to conduct the mapping of reads to the reference sequence and provide a list of parameter settings used along with a rationale for each. The report should provide specific information about the algorithm including, but not limited to, percentage of mismatches tolerated, number of insertions and deletions (indels) tolerated, and identity and similarity cutoffs.

c. Describe the algorithm employed to call variants and provide the parameters used. In addition, describe how single nucleotide polymorphisms (SNPs) and indels were distinguished from sequence errors.

5.3.2 Reporting results from the read mapping and variant calling

Sponsors should consider the following for inclusion in the report in results from the read mapping and variant calling:

a. Provide a summary of the coverage achieved for each read mapping and include a coverage graph.

b. Provide the number of sequences that were present in each read mapping.

c. Provide a frequency table for all read mappings showing predominant SNPs and indels.

d. Provide a combined frequency table of the NGS results from each clinical trial in amino acid format that includes the following: the amino acids at each position, the frequency for variants at each position, and the coverage at each position where variation occurred (see example in section 7.0, Frequency Table Example).

e. Provide a summary table showing predominant changes across the population of interest.

5.3.3 De novo assembly of contigs and calling variants (when applicable)

Sponsors should consider the following for inclusion in the report on de novo assembly of contigs and calling variants:

a. Provide contig statistics showing the number of contigs generated from the reads.

b. Identify the program and describe the algorithm used to conduct the de novo assembly of short reads into contigs and provide a list of parameter settings used along with a rationale for each. The report should provide specific information about the algorithm including, but not limited to, percentage of mismatches tolerated, number of indels tolerated, and identity and similarity cutoffs.

c. Describe the algorithm employed to compare contigs and call variants, and provide the parameters used. In addition, describe how SNPs and indels were distinguished from sequence errors.

d. Describe how contigs were annotated by identifying the program that was used to identify similarity to known sequences, identify the database that was queried, provide the search criteria that were used, define the cutoffs used for percent identity, and provide details on how contaminants or other viruses were identified and reported.

5.3.4 Reporting results from the de novo assembly and variant calling

Sponsors should consider the following for the report on de novo assembly and variant calling:

a. Provide a summary of the coverage for each contig.

b. Provide the number of sequences in each contig and the number of reads that did not form contigs.

c. Describe contaminants or other viruses that were detected during annotation and report the percent representation of each contaminant.

d. Provide a frequency table for all major contigs showing predominant SNPs and indels.

e. Provide a combined frequency table of the NGS results from each clinical trial in amino acid format that includes the following: the amino acids at each position, the frequency for variants at each position, and the coverage at each position where variation occurred (see example in section 7.0, Frequency Table Example).

f. Provide a summary table showing predominant changes across the population of interest.

g. We recommend performing de novo assembly of all reads that do not map to a reference sequence and annotating the contigs to assess for potential cross contamination of samples.

6.0 NGS File Types and Submission Procedures

6.1 Files Submitted Via Portable Hard Drive

The raw NGS data in the fastq format and frequency tables should be sent to the Division on a secured, portable hard drive following the guidelines outlined in the technical specifications document *Transmitting Electronic Submissions Using eCTD Specifications* (April 2019).[2]

6.1.1 Contents of the portable hard drive

Note that only the raw NGS data, the frequency table(s), and a table of contents should be submitted on the hard drive. Additional files, such as those with a .exe extension, may result in rejection of the submission.

If the hard drive is password protected (not required or recommended at this time), the sponsor should consult with the Division ahead of time to ensure that the password is provided to the appropriate personnel in the Center for Drug Evaluation and Research (CDER) document room.

6.1.2 Labeling fastq sequence files

All NGS data should be provided in the fastq format, which includes nucleotide sequences and quality information. This should include NGS sequence runs for each subject at various time points, including baseline and a time point close to when treatment failure was observed. The sponsor should provide the fastq files labeled by unique subject identifier and time point. For example, for study ABC123, subject 0001, and three samples (baseline, week 16, and follow up week 4) the files should be labeled as follows:

a. ABC123-0001.baseline.fastq
b. ABC123-0001.WEEK16.fastq
c. ABC123-0001.WEEK4FU.fastq

[2] Available at https://www.fda.gov/media/76812/download. See also the Electronic Regulatory Submission and Review web page at https://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/default.htm.

If a sample is resequenced, distinguish the first and second sequence runs as follows:

a. ABC123-0001.baseline-1.fastq
b. ABC123-0001.baseline-2.fastq

6.2 Files submitted via the CDER electronic document gateway

The NGS protocol, analysis-ready resistance datasets, and any other supporting information should be submitted via the CDER electronic document gateway.

7.0 Frequency Table Example

STUDYID	USUBJID	NGSPL	VISIT	AAPOS	AAREF	AASUB	TCOV	VCOV	AAFREQ
ABC123-999	0123	Illumina	BL	81	R	K	4317	156	0.036
ABC123-999	0123	Illumina	BL	98	K	R	2841	99	0.035
ABC123-999	0123	Illumina	Day 2	98	K	R	9487	366	0.039
ABC123-999	0123	Illumina	Day 3	98	K	R	9474	378	0.040
ABC123-999	0123	Illumina	BL	120	R	Q	4310	200	0.046
ABC123-999	0123	Illumina	Day 2	120	R	Q	12722	470	0.037
ABC123-999	0123	Illumina	Day 3	120	R	Q	12466	489	0.039
ABC123-999	0123	Illumina	BL	147	I	V	3456	742	0.215
ABC123-999	0123	Illumina	Day 2	147	I	V	13456	2709	0.201
ABC123-999	0123	Illumina	Day 3	147	I	V	13297	1934	0.145
ABC123-999	0123	Illumina	BL	150	A	V	3107	43	0.014
ABC123-999	0123	Illumina	Day 2	150	A	T	13116	167	0.013
ABC123-999	0123	Illumina	BL	154	K	R	2987	124	0.042
ABC123-999	0123	Illumina	Day 2	154	K	R	13434	1350	0.101
ABC123-999	0123	Illumina	Day 3	154	K	R	13077	1206	0.092
ABC123-999	0123	Illumina	Day 3	155	R	K	12459	9837	0.781
ABC123-999	0123	Illumina	Day 3	156	P	S	13385	172	0.013
ABC123-999	0123	Illumina	BL	186	V	I	6155	129	0.021
ABC123-999	0123	Illumina	Day 2	186	V	I	17698	269	0.015
ABC123-999	0123	Illumina	Day 3	186	V	I	16474	460	0.028
ABC123-999	0123	Illumina	BL	206	K	H	9698	165	0.017
ABC123-999	0123	Illumina	Day 2	206	K	R	24601	292	0.012
ABC123-999	0123	Illumina	Day 3	210	S	N	23001	255	0.011
ABC123-999	0123	Illumina	Day 3	254	H	R	25145	290	0.012

STUDYID = study protocol number; **USUBJID** = unique subject ID; **NGSPL** = next generation sequencing platform used for sequencing; **VISIT** = study visit that the sample was collected from; **AAPOS** = amino acid position in the target gene; **AAREF** = amino acid present at this position in the reference sequence; **AASUB** = amino acid substitution detected by sequencing; **TCOV** = total coverage at the nucleotide site; **VCOV** = total coverage at the nucleotide position of the variant; **AAFREQ** = frequency of the substitution detected; **BL** = baseline.