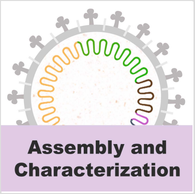


3 ▾

Oct 19, 2021

Cecret Workflow for SARS-CoV-2 Assembly and Lineage Classification V.3

1

Erin L Young[1], Technical Outreach and Assistance for States Team[2]

[1]Utah Public Health Laboratory; [2]Centers for Disease Control and Prevention

protocol .

TOAST_public
Tech. support email: **toast@cdc.gov**

ycm6

The opinions expressed here do not necessarily reflect the opinions of the Centers for Disease Control and Prevention or the institutions with which the authors are affiliated. The protocol content here is under development and is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

This protocol provides instructions to install and run the Cecret workflow as part of the StaPH-B Toolkit. Cecret produces SARS-CoV-2 consensus sequence assemblies from either single- or paired-end Illumina reads in fastq (or fastq.gz) format and assigns lineage classifications using Pangolin and Nextclade. This document applies to all whole-genome sequencing runs on the Illumina platform and downstream bioinformatics for public health laboratories.

For technical assistance, please contact: **TOAST@cdc.gov**

https://github.com/StaPH-B/staphb_toolkit

Erin L Young, Technical Outreach and Assistance for States Team 2021. Cecret Workflow for SARS-CoV-2 Assembly and Lineage Classification . **protocols.io**
https://protocols.io/view/cecret-workflow-for-sars-cov-2-assembly-and-lineag-by72pzqe
Technical Outreach and Assistance for States Team

(ILLUMINA MENU) Protocols for SARS-CoV-2 Library Prep with ARTIC Primers, Bioinformatic Analysis, and Database Submission

SARS-CoV-2, Genomics, Pangolin, StaPH-B

protocol ,

Oct 19, 2021

Oct 19, 2021

Oct 19, 2021 ycm6

54234

Part of collection
(ILLUMINA MENU) Protocols for SARS-CoV-2 Library Prep with ARTIC Primers, Bioinformatic Analysis, and Database Submission

:

The opinions expressed here do not necessarily reflect the opinions of the Centers for Disease Control and Prevention or the institutions with which the authors are affiliated. The protocol content here is under development and is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Software Dependencies

Load software dependencies

1 Cecret and the StaPH-B Toolkit require the following dependencies:

1. Singularity or Docker,
2. Python 3.6 or later,
3. Java version 8 or later.

Additional instructions are provided here: https://staph-b.github.io/staphb_toolkit/install/

Ensure all dependencies are in your system's PATH environment variable.

Within certain high-performance computing environments, these software can be loaded using GNU module commands similar to:

Loading dependencies for staphb_toolkit

```
module load Python/3.9.1
module load java/jdk1.8.0_221
module load nextflow/20.04.1
module load singularity/3.5.3
```

Installing StaPH-B Toolkit

2 The Cecret assembly workflow can be installed as part of the StaPH-B Toolkit using the following commands:

staphb_tookit Install

```
git clone https://github.com/StaPH-B/staphb_toolkit.git
cd staphb_toolkit/packaging/
python3 setup.py install --user
cd ../
export PATH=$PATH:$(pwd)
```

Running Cecret Workflow

3 Run the Cecret workflow to perform SARS-CoV-2 genome sequence assembly and lineage classification

The input directory for Cecret should contain a set of single or paired-end (default) fastq.gz

The input directory for Cecret should contain a set of single or paired end (default) fastq.gz or fastq reads from amplicon prepared libraries. By default, Cecret is configured to use the ARTIC V3 primer set but can be customized to use other primer sets.

In the examples below, the *--profile* argument is set to use Singularity containers, but Cecret works with Docker containers as well (*--profile* docker).

Detailed descriptions of parameters are provided here: https://github.com/UPHL-BioNGS/Cecret

Cecret Input and Output File Paths

```
#Input Sequencing Reads File Path:
/Full_Path_to_Fastq_File_Directory/INPUT/SRR11953697

#Output Directory:
/Full_Path_to_Cecret_Output_Directory/SRR11953697_cecret_output
```

Cecret Workflow Command

```
mkdir -p cachedir
SINGULARITY_CACHEDIR=/PATH/for/cache
staphb-wf cecret /Full_Path_to_Fastq_File_Directory/INPUT/SRR11953697 --output /Full_Path_to_Cecret_Output_Directory/SRR11953697_cecret_output --profile singularity
```

Default cache directory is ~/.singularity/cache. -resume is available if you would like to retry an interrupted workflow.

Getting/Using an optional configuration file

```
staphb-wf cecret --get_config
staphb-wf cecret /Full_Path_to_Fastq_File_Directory/INPUT/SRR11953697
--output
/Full_Path_to_Cecret_Output_Directory/SRR11953697_cecret_output --
profile singularity -c
/Full_Path_to_Config_File_Directory/date_cecret.config
```

Output Files

4 Information about output files

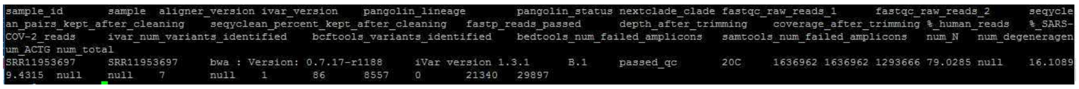
Complete output files structure

```
cecret_run_results.txt              # information about the sequencing run
that's compatible with legacy workflows
covid_samples.csv                   # only if supplied initially
cecret
|-aligned
| |-pretrimmed.sorted.bam
|-bamsnap
| |-sample
|   |-ivar
|     |-variant.png                 # png of variants identified via ivar
|   |-bcftools
|     |-variant.png                 # png of variants identified via bcftools
|-bedtools_multicov
| |-sample.multicov.txt             # depth per amplicon
|-consensus
| |-sample.consensus.fa             # the likely reason you are running this
workflow
|-fastp
| |-sample_clean_PE1.fastq          # clean file: only if
params.cleaner=fastp
| |-sample_clean_PE2.fastq          # clean file: only if
params.cleaner=fastp
|-fastqc
| |-sample.fastqc.html
| |-sample.fastqc.zip
|-filter                            # optional: turns aligned bams into fastq files
| |-sample_filtered_R1.fastq
| |-sample_filtered_R2.fastq
| |-sample_filtered_unpaired.fastq
|-iqtree                          # optional: relatedness parameter must be set
to true
| |-iqtree.treefile
|-ivar_trim
| |-sample.primertrim.bam           # aligned reads after primer trimming.
trimmer parameter must be set to 'ivar'
|-ivar_variants
| |-sample.variants.tsv             # list of variants identified via ivar and
corresponding scores
|-kraken2
| |-sample_kraken2_report.txt        # kraken2 report of the percentage of
reads matching virus and human sequences
|-logs
| |-process_logs                    # for troubleshooting puroses
|-mafft                            # optional: relatedness parameter must be set
to true
| |-mafft_aligned.fasta             # multiple sequence alignement
generated via mafft
|-nextclade                         # identfication of nextclade clades and
variants identified
| |-sample_nextclade_report.csv       # actually a ";" deliminated file
|-pangolin
| |-lineage_report.csv              # identification of pangolin lineages
|-samtools_ampliconstats
| |-sample_ampliconstats.txt          # stats for the amplicons used
|-samtools_coverage
| |-aligned
| | |-sample.cov.hist               # histogram of coverage for aligned reads
| | |-sample.cov.txt                # tabular information of coverage for
aligned reads
| |-trimmed
|   |-sample.cov.trim.hist           # histogram of coverage for aligned
reads after primer trimming
|   |-sample.cov.trim.txt            # tabular information of coverage for
aligned reads after primer trimming
|-samtools_depth
| |-aligned
| | |-sample.depth.txt              # read depth for each read position
| |-trimmed
|   |-sample.depth.txt              # read depth for each position
|-samtools_flagstat
| |-aligned
| | |-sample.flagstat.txt           # samtools stats for aligned reads
| |-trimmed
|   |-sample.flagstat.trim.txt       # samtools stats for trimmed reads
|-samtools_plot_ampliconstats
| |-sample.*.png                     # images corresponding to amiplicon
performance
|-samtools_stats
| |-aligned
| | |-sample.stats.txt               # samtools stats for aligned reads
| |-trimmed
|   |-sample.stats.trim.txt          # samtools stats for trimmed reads
|-seqyclean
| |-sample_clean_PE1.fastq           # clean file
| |-sample_clean_PE2.fastq           # clean file
|-snp-dists                          # optional: relatedness parameter must be set
to true
| |-snp-dists                        # file containing a table of the number of snps
that differ between any two samples
|-submission_files                   # optional: is only created if
covid_samples.txt exists
| |-sample.genbank.fa                # fasta file with formatting and header
including metadata for genbank
| |-sample.gisaid.fa                 # fasta file with header for gisaid
| |-sample.R1.fastq.gz               # renamed raw fastq.gz file
| |-sample.R2.fastq.gz               # renamed raw fastq.gz file
|-summary
| |-sample.summary.txt               # individual results
|-summary.csv                        # tab-delimited summary of results from
the workflow
reads
| |-sample_S1_L001_R1_001.fastq.gz    # initial file
| |-sample_S1_L001_R2_001.fastq.gz    # inital file
work                                 # nextflows work directory. Likely fairly large.
vadr
  |-vadr*                            # vadr output
```

Cecret Consensus Sequence and Summary Report Paths

```
#Consensus Sequence Path:
/Full_Path_to_Cecret_Output_Directory/SRR11953697_cecret_output/conse

#Summary Report Path:
/Full_Path_to_Cecret_Output_Directory/SRR11953697_cecret_output/sumn
```



```
sample_id       sample  aligner_version ivar_version    pangolin_lineage        pangolin_status nextclade_clade fastqc_raw_reads_1      fastqc_raw_reads_2      seqycle
an_pairs_kept_after_cleaning    seqyclean_percent_kept_after_cleaning   fastp_reads_passed      depth_after_trimming    coverage_after_trimming %_human_reads   %_SARS-
COV-2_reads     ivar_num_variants_identified    bcftools_variants_identified    bedtools_num_failed_amplicons   samtools_num_failed_amplicons   num_N   num_degeneragen
um_ACTG num_total
SRR11953697     SRR11953697     bwa : Version: 0.7.17-r1188     iVar version 1.3.1      B.1     passed_qc       20C     1636962 1636962 1293666 79.0285 null    16.1089
9.4315  null    null    7       null    1       86      8557    0       21340   29897
```

Example Summary.txt output file

Additional documentation for the Cecret workflow is available here:
https://github.com/UPHL-BioNGS/Cecret

And further details about the StaPH-B Toolkit are available here:
https://staph-b.github.io/staphb_toolkit

Before submitting the resulting SARS-CoV-2 consensus sequence assemblies to public repositories, such as NCBI GenBank or GISAID, refer to the following documentation describing submission criteria and minimum quality control thresholds:

GenBank Submission Criteria: About GenBank Submission (nih.gov)

GISAID Submission Criteria: **Gisaid inclusion criteria.pdf**

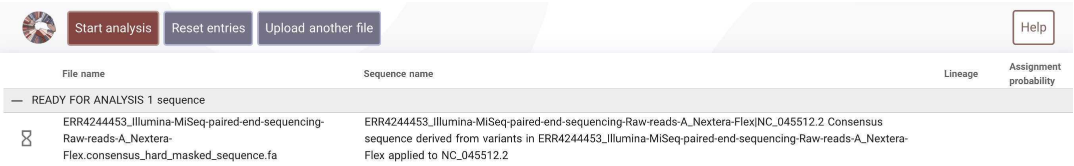
Alternative Lineage Assignment

5 The SARS-CoV-2 consensus sequence assembly generated by the Cecret workflow can also be uploaded to other lineage assignment software.

5.1 Upload the consensus sequence for each sample to the **Pangolin COVID-19 Lineage Assigner** at:

https://pangolin.cog-uk.io/

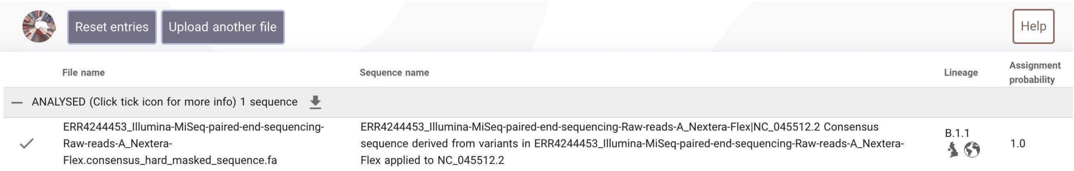
Click the 'Start analysis' button:



	File name	Sequence name	Lineage	Assignment probability
— READY FOR ANALYSIS 1 sequence				
⌛	ERR4244453_Illumina-MiSeq-paired-end-sequencing-Raw-reads-A_Nextera-Flex.consensus_hard_masked_sequence.fa	ERR4244453_Illumina-MiSeq-paired-end-sequencing-Raw-reads-A_Nextera-Flex|NC_045512.2 Consensus sequence derived from variants in ERR4244453_Illumina-MiSeq-paired-end-sequencing-Raw-reads-A_Nextera-Flex applied to NC_045512.2		

Pangolin COVID-19 Lineage Assigner example

The Pangolin COVID-19 Lineage Assigner returns the lineage classification and assignment probability:

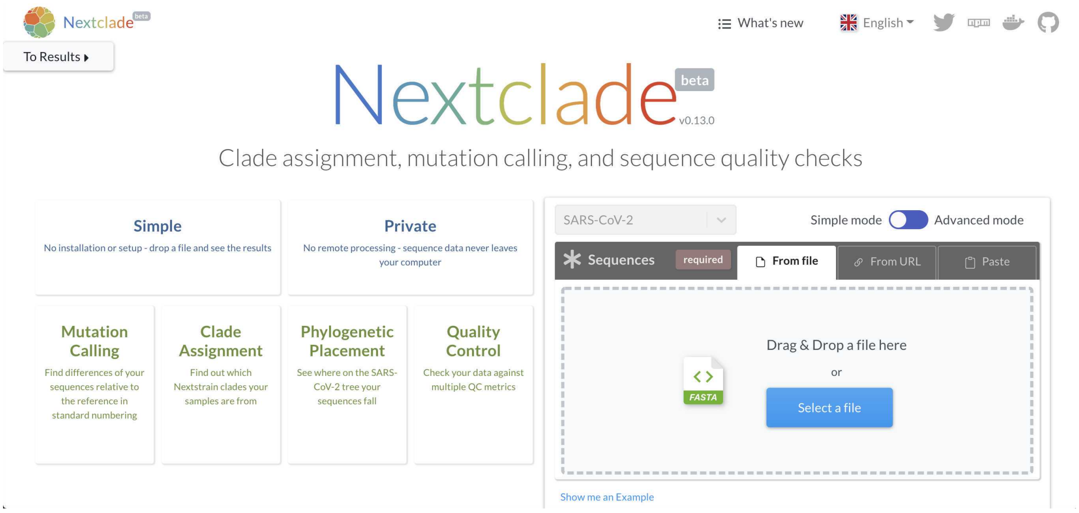


	File name	Sequence name	Lineage	Assignment probability
— ANALYSED (Click tick icon for more info) 1 sequence				
✓	ERR4244453_Illumina-MiSeq-paired-end-sequencing-Raw-reads-A_Nextera-Flex.consensus_hard_masked_sequence.fa	ERR4244453_Illumina-MiSeq-paired-end-sequencing-Raw-reads-A_Nextera-Flex|NC_045512.2 Consensus sequence derived from variants in ERR4244453_Illumina-MiSeq-paired-end-sequencing-Raw-reads-A_Nextera-Flex applied to NC_045512.2	B.1.1	1.0

Pangolin COVID-19 Lineage Assigner output

5.2 Or upload the consensus sequence for each sample to the **Nextclade** clade assignment web portal at:

https://clades.nextstrain.org/



Clade assignment, mutation calling, and sequence quality checks

Simple: No installation or setup - drop a file and see the results

Private: No remote processing - sequence data never leaves your computer

Mutation Calling: Find differences of your sequences relative to the reference in standard numbering

Clade Assignment: Find out which Nextstrain clades your samples are from

Phylogenetic Placement: See where on the SARS-CoV-2 tree your sequences fall

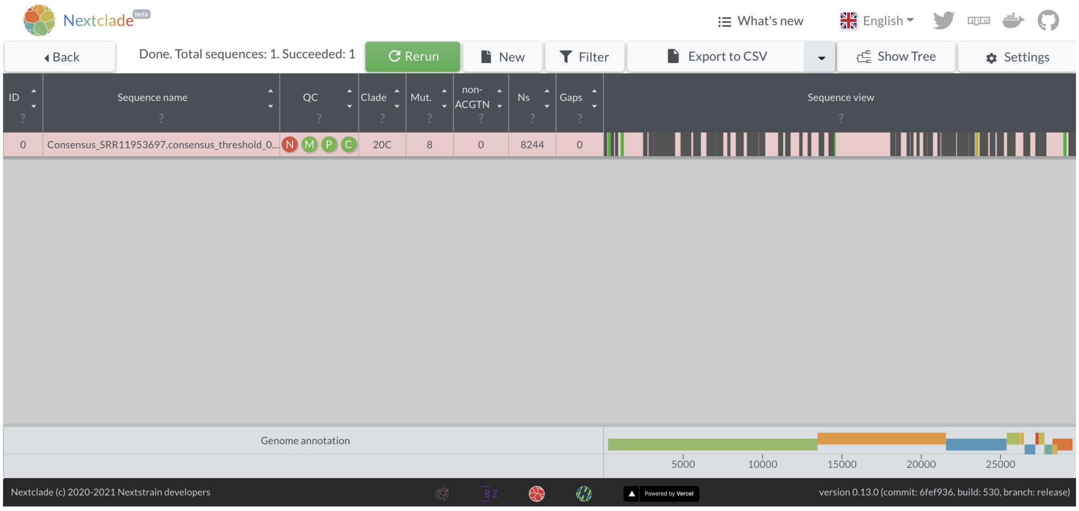
Quality Control: Check your data against multiple QC metrics

Drag & Drop a file here

or

NextClade assignment web portal

The Nextclade server provides clade classification as well as QC metrics and a list of amino acid substitutions. A summary output file can be downloaded with the 'Export to CSV' button.



Done. Total sequences: 1. Succeeded: 1

ID	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Gaps
0	Consensus_SRR11953697.consensus_threshold_0...	N M P C	20C	8	0	8244	0

Nextclade (c) 2020-2021 Nextstrain developers

version 0.13.0 (commit: 6fef936, build: 530, branch: release)

Nextclade clade assignment output