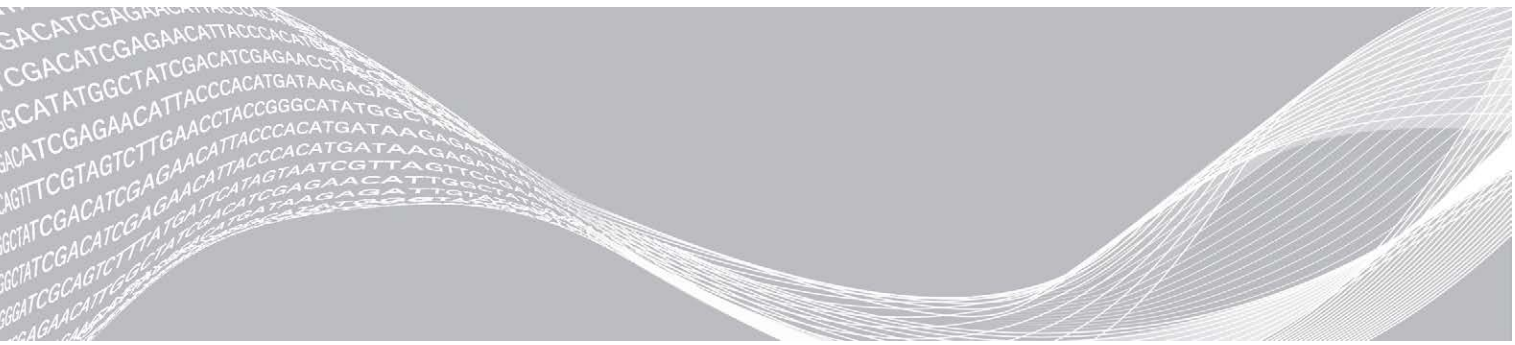
illumina®

Illumina DRAGEN COVID Pipeline

Software Guide



Document # 1000000158680 v01

ILLUMINA PROPRIETARY

October 2021

For Research Use Only. Not for use in diagnostic procedures.

Revision History

Document	Date	Description of Change
Document # 1000000158680 v01	October 2021	• Added new commands to the Running the Illumina DRAGEN COVID Pipeline section. • Added command for uninstalling prior Illumina DRAGEN COVID Pipeline version in the Install the Illumina DRAGEN COVID Pipeline section. • Updated content throughout for v1.0.1 release and compatibility with new ARTIC v4 primers. • Updated output folders and files in Analysis Methods and Output Structure sections. • Updated the detected sequence variants depth criteria to "greater than or equal to 10" in Variant Calling and Consensus Sequence Generation for Research Use Only section.
Document # 1000000158680 v00	April 2021	Initial release.

Table of Contents

Overview

The Illumina DRAGEN COVID Pipeline analyzes sequencing reads of RNA libraries prepared using any ARTIC v3 and v4 gene panel assays. The Illumina DRAGEN COVID Pipeline uses the Illumina DRAGEN Bio-IT Platform to perform analysis to determine the presence of SARS-CoV-2 and generates results in BAM, VCF, and FASTA formats.

The Illumina DRAGEN COVID Pipeline requires a sample sheet. For information on creating a sample sheet, see *BCL Convert Software Guide (document # 100000094725)*.

In addition to SARS-CoV-2 analysis, the Illumina DRAGEN COVID Pipeline performs small variant calling for samples that meet the minimum threshold of SARS-CoV-2 virus targets detected. The minimum threshold is configurable with a default of 90 targets. See the options in *Running the Illumina DRAGEN COVID Pipeline* on page 3. The Illumina DRAGEN COVID Pipeline generates a consensus sequence in FASTA format. Variant calls and consensus sequences are generated for informational purposes only and not for patient reporting.

Installation Requirements

Illumina DRAGEN COVID Pipeline contains the following minimum operating requirements.

The Illumina DRAGEN COVID Pipeline is compatible with a DRAGEN Server v2 and v3.

By default, the software includes the following items:

- Linux CentOS 7.3 operating system, or later.

The following additional software is required before installing Illumina DRAGEN COVID Pipeline.

- Docker version 18.09, or later.

Storage Requirements

The DRAGEN Server provides NVMe SSD located in `/staging` directory to use as the software output directory.

If using the DRAGEN Server v2, store sequencing run data in a network-attached folder to make sure the required disk space is available on the NVMe SSD drives for analysis output. Network-attached storage is required for long-term storage for both DRAGEN Server v2 and v3.

Analysis output is automatically written to the `/staging/dragen-covid-pipeline_analysis_<timestamp>` to make sure the DRAGEN Server processes read and write data on the NVMe SSD. You can modify this location using the command-line.

Before beginning analysis, develop a strategy to copy data from the DRAGEN Server to a network-attached storage. Delete output data on the DRAGEN Server as soon as possible.

The following are the run and analysis output sizes for each sequencing system per 36 bp. Output folder size can vary based on the number of positive samples. The following table are recommended storage requirements.

Sequencing System	Run Folder Output (GB)	Analysis Output (GB)
NovaSeq 6000 SP flow cell	20	60
NovaSeq 6000 S4 flow cell	225–240	860
NextSeq 500/550 and 550Dx HO flow cell	12	30

Install the Illumina DRAGEN COVID Pipeline

Use the instructions in this section to install the Illumina DRAGEN COVID Pipeline.

Illumina recommends running Docker as a non-root user by adding the user to the docker group. It is possible to run the Illumina DRAGEN COVID Pipeline as root but not recommended. For more information, see the Docker website.

The Illumina DRAGEN COVID Pipeline installation script uninstalls any existing DRAGEN software on the server. If you would like to use a different DRAGEN pipeline, you will need to uninstall the Illumina DRAGEN COVID Pipeline, and download a DRAGEN software installation package from the DRAGEN support page.

1 Contact your local Illumina Field Application Scientist to obtain the Illumina DRAGEN COVID Pipeline installer package.

2 Install Docker 18.09 or later using the install instructions for CentOS provided in the Docker documentation.

3 Install the DRAGEN Server license using the instructions provided in the *Illumina DRAGEN Server Site Prep & Installation Guide*.

4 Download the Illumina DRAGEN COVID Pipeline installation script provided in the email from Illumina. The link expires after 1 week.

5 Store the install script in the `/staging` directory.

6 To run the software in a screen virtual terminal, enter the following command: `screen`.

```
screen -S <name>
```

In addition to other uses, the command avoids terminating the execution due to terminal inactivity.

7 To update the run script permissions, enter the following command:

```
chmod +x /staging/install_dragen-covid-pipeline-RUO-v1.0.1.run
```

8 To uninstall previous versions of the DRAGEN COVIDSeq Test Pipeline or the Illumina DRAGEN COVID Pipeline, enter one of the following commands:
- DRAGEN COVIDSeq Test Pipeline: `/staging/uninstall_covidseq-1.1.0.sh`
- Illumina DRAGEN COVID Pipeline: `/staging/uninstall_dragen-covid-pipeline-RUO-1.0.0.sh`

9 To run the installation script, enter the following command:

```
/staging/install_dragen-covid-pipeline-RUO-1.0.1.run
```

The script removes any previously installed DRAGEN software. Depending on your previous version of DRAGEN, you might need to power cycle your server after install.

Running the System Check

Make sure that the system is functioning properly by running the `check_dragen-covid-pipeline-1.0.1.sh` script. The self-test script checks the following functions:

- If all required services are running.
- If the proper Docker image is installed.
- If the Illumina DRAGEN COVID Pipeline successfully runs on a test data set.

The self-test runs for approximately five minutes. If the self-test prints a failure message, contact Illumina Technical Support and provide the `/staging/check_dragen-covid-pipeline_<timestamp>.tgz` output file.

To detach from the screen process at anytime, enter `crtl-a d`.

Running the Illumina DRAGEN COVID Pipeline

The Illumina DRAGEN COVID Pipeline is started by selecting the shell script using the command line, and then running the software with Docker. Analysis outputs are located in the `/staging/dragen-covid-pipeline_analysis_<timestamp>` directory.

This location ensures that the server is on an optimized NVMe SSD.

Do not move files or press CRTL+C when the app is running. Moving files during the analysis can cause the analysis to fail or provide incorrect results. Pressing CRTL+C stops the analysis and might cause an error. If an error does occur, restart the server.

1 If detached from the screen process, enter the following command to reattach to screen:

```
screen -r name
```

2 To run the Illumina DRAGEN COVID Pipeline, enter the following command-line argument:

```
dragen-covid-pipeline.sh --runFolder <FULL_PATH_TO_RUN_FOLDER>
```

3 [Optional] Enter any of the other following available commands:

Option	Description
`--ampliconThreshold`	Specifies the minimum number of SARS-CoV-2 targets detected for the sample to have a result of SARS-CoV2-Detected. The default value is 90.
`--analysisFolder`	Full path to the alternative analysis folder. For high performance, the folder must be on an NVMe SSD partition. Make sure to use a different folder than the test data folder `/staging/illumina/covid`. If the Illumina DRAGEN COVID Pipeline is uninstalled, the test data folder is deleted.
`--fastMode`	Turns off alignment, variant calling, and consensus sequence FASTA generation to improve speed.
`--fastqFolder`	Full path to the FastQ folder. Use of this command requires the specification of a sample sheet using the `--sampleSheet` command. Do not use this command if you are using the `--runFolder` command.
`--help`	Displays a help screen, and then exits.
`--lane`	To exclude BCL Convert on a single lane, set the option to the value 1, 2, 3, or 4.
`--primers`	Specifies primers version number. Default value is v3. Set this value to v4 if you are using ARTIC v4 primers.
`--runFolder`	Full path to the BCL folder. Do not use this command if you are using the `--fastqFolder` command.
`--sampleSheet`	Full path to the sample sheet. If your sample sheet is not located in the root directory of your run folder or is not named SampleSheet.csv, the `--sampleSheet` command is required.
`--version`	Displays the version of the software, and then exits.

Process Lane Subsets or Multiple Flow Cells

To analyze a subset of lanes, create a copy of the sample sheet, and then remove all samples that are not in the lanes to process. Specify this new sample sheet on the command line.

To process a single lane, use the `--lane` option with the value set to 1, 2, 3, or 4. The option restricts demultiplexing BCL to FASTQ to the specified lane and increases overall performance. Only include samples from the specified lane in your sample sheet.

To analyze multiple flow cells, perform multiple, serial executions of the software. Only initiate a new analysis after the previous is completed. Running multiple executions of the software concurrently on the same server can cause the analysis to fail or produce incorrect results.

Each flow cell includes a separate run folder.

Analysis Methods

The Illumina DRAGEN COVID Pipeline performs analysis using the following steps. Each of the first seven steps creates a subfolder in Logs_intermediates subfolder under the analysis folder.

1 Validates the sample sheet fields.
This step generates the `SampleSheetValidation` subfolder.

2 Converts BCL data in the run folder to FASTQ sample data. All samples from the run are available as FASTQ files compressed in a gzip. If the `--lane` option is used, only samples in the specified lane are available as FASTQ files.
This step generates the `FastqGeneration` subfolder.

3 For each sample, Illumina DRAGEN COVID Pipeline determines the presence of SARS-CoV-2 and an internal (human) control. The read coverage per target is compared to a fixed target threshold to determine covered targets. The number of covered targets is then used to detect SARS-CoV-2 (≥ virusThreshold) and the internal control (≥ humanThreshold).
The step generates the `VirusDetection` subfolder.

4 For each sample with a result of SARS-CoV-2 Detected and at least 90 SARS-CoV-2 targets detected, Illumina DRAGEN COVID Pipeline aligns FASTQ files to the SARS-CoV-2 reference genome. To change the number of targets detected from the default of 90, use the `--ampliconThreshold` option at run time.
This step generates the `MapAlign` subfolder.

5 For each sample with a result of SARS-CoV-2 Detected and at least 90 SARS-CoV-2 targets detected, Illumina DRAGEN COVID Pipeline performs variant calling to determine any variants present in the sample with respect to the SARS-CoV-2 reference genome. To change the number of targets detected from the default of 90, use the `--ampliconThreshold` option at run time. This step produces VCF files containing detected variants for each processed sample. See *Variant Calling and Consensus Sequence Generation for Research Use Only* on page 7 for more information.
This step generates the `VariantCalling` subfolder.

6 For each sample with a result of SARS-CoV-2 Detected and at least 90 SARS-CoV-2 targets detected, Illumina DRAGEN COVID Pipeline generates a consensus genome in FASTA format using variant calls and coverage metrics as input. To change the number of targets detected from the default of 90, use the `--ampliconThreshold` option at run time. See *Variant Calling and Consensus Sequence Generation for Research Use Only* on page 7 for more information.
This step generates the `ConsensusFasta` subfolder.

7 For each sample, Illumina DRAGEN COVID Pipeline generates a viral only BAM file.
This step generates the `NonViralReads` subfolder.

8 The Illumina DRAGEN COVID Pipeline generates a combined coverage report (coverage.csv file) that determines if the sample had greater than or equal to 30x coverage.

For information about downstream analysis using third-party software, see *DRAGEN COVID Lineage Tools* on page 7.

Detection Algorithm

The Illumina DRAGEN COVID Pipeline uses a kmer based algorithm to detect SARS-CoV-2 and any external controls. The algorithm uses a kmer reference database to match kmers from the sequencing read to kmers from the SARS-CoV-2 reference genome (NC_0455). To create the kmer reference list, the SARS-CoV-2 reference genome is split in 32 bp kmers, and then any kmers that contain cross-reactivity are removed. To measure cross-reactivity, the kmer reference list uses the NCBI database of 100k human and animal pathogens. Bat and pangolin viruses are not included because of the similarity to the SARS-CoV-2 genome.

The kmer algorithm is performed using the following process:

1. The sequencing read is split into 32 bp kmers.
2. The kmers are matched to the kmer reference list.
3. Each of the reference kmers is labeled with a corresponding amplicon from either SARS-CoV-2 or external control.
4. If an amplicon contains at least 150 matched read and reference kmers, the amplicon is detected.
5. If the following number of amplicons are present, the algorithm detects SARS-CoV-2 or the external control.
   - SARS-CoV-2 is detected if there are at least 5 SARS-CoV-2 amplicons.
   - External control is detected if there are at least 3 external control amplicons.

Output Structure

The Illumina DRAGEN COVID Pipeline outputs results in the following folder structure. Key output files are shown below.

- Logs_Intermediates
  - SampleSheetValidation
  - ConsensusFasta
    - <sampleID>
      - <sampleID>.consensus_metrics.csv
  - FastqGeneration
    - Reports
      - Demultiplex_Stats.csv
      - Adapter_Metrics.csv
  - MapAlign
  - NonViralReads
    - <sampleID>
      - <sampleID>_virus_only.bam
  - VariantCalling
  - VirusDetection
- Sample_Analysis
  - coverage.csv
  - <sampleID1>
    - <sampleID1>.fasta
    - <sampleID1>.fasta.md5sum
    - <sampleID1>_tumor.bam
    - <sampleID1>_tumor.bam.bai
    - <sampleID1>_tumor.bam.md5sum
    - <sampleID1>.hard-filtered.vcf.gz
    - <sampleID1>.hard-filtered.vcf.gz.md5sum
    - <sampleID1>.hard-filtered.vcf.gz.tbi
    - <sampleID1>.consensus_filtered_variants.vcf.gz
    - <sampleID1>.consensus_filtered_variants.vcf.gz.md5sum
    - <sampleID1>.consensus_filtered_variants.vcf.gz.tbi
  - <sampleID2>
- Errors.tsv

AnalysisLog.txt

run.log

task_duration.csv

Variant Calling and Consensus Sequence Generation for Research Use Only

Illumina DRAGEN COVID Pipeline performs variant and consensus sequence generation for each sample with a result of SARS-CoV-2 Detected and at least 90 SARS-CoV-2 virus targets detected. To change the number of targets detected from the default of 90, use the `--ampliconThreshold` option at run time. Variant calls and consensus sequences are for information purposes only and should not be used for patient reporting.

Variant calling and consensus sequence generation is not performed for invalid samples.

The variant calling output file is generated in VCF 4.2 file format and located in Sample_Analysis/<Sample ID>/<Sample ID>.hard-filtered.vcf.gz.

The consensus filtered variant calling output file is located in Sample_Analysis/<Sample ID>/<Sample ID>.consensus_filtered_variants.vcf.gz.

To generate a consensus sequence in FASTA format, detected sequence variants that meet the following criteria are applied to the SARS-CoV-2 reference sequence (NCBI Accession NC_045512.2).

- All DRAGEN quality filters pass.
- Allele frequency is greater than or equal to 0.5.
- Depth is greater than or equal to 10.

Regions of sequence with coverage below 10 are masked as low-confidence. Hard-masking is applied, and all bases in low-confidence regions are converted to "N". A soft-masked sequence is also provided and indicates all low-confidence regions with lower case characters.

The hard-masked consensus FASTA is available in Sample_Analysis/<Sample ID>/<Sample ID>.fasta.

Uninstall Illumina DRAGEN COVID Pipeline

The Illumina DRAGEN COVID Pipeline includes an uninstall script located in the `/usr/local/bin` called `uninstall_dragen-covid-pipeline-1.0.1.sh`.

The uninstall script removes the following assets:

- All scripts (dragen-covid-pipeline.sh, check_dragen-covid-pipeline-1.0.1.sh, uninstall_dragen-covid-pipeline-1.0.1.sh).
- The Illumina DRAGEN COVID Pipeline Docker image.
- Data stored in /staging/illumina/dragen-covid-pipeline.

The script does not uninstall Docker.

To uninstall the Illumina DRAGEN COVID Pipeline, enter the following command as root.

```
/usr/local/bin/uninstall_dragen-covid-pipeline-1.0.1.sh
```

DRAGEN COVID Lineage Tools

You can use the DRAGEN COVID Lineage Tools software package for complimentary downstream analysis for Illumina DRAGEN COVID Pipeline.

DRAGEN COVID Lineage Tools packages the third-party software applications Pangolin and NextClade, which provide lineage and clade determination of viral RNA sequences found in FASTA files. The DRAGEN COVID Lineage Tools software package is not supported by Illumina Technical Support.

The software is packaged as a Docker image and deployed as a TAR file.

Load and run DRAGEN COVID Lineage Tools, as follows.

1 Download the TAR file from the Illumina DRAGEN COVID Pipeline support page.

2 To extract the image, load the TAR file into Docker by entering the following command:

```
docker load < dragen-covid-lineage-tools-1.0.1.tar
```

3 To see all the images on your server, enter the following command:

```
docker images
```

The response displays all docker images on the system, which should include the image dragen-covid-lineage-tools:1.0.1.

4 To run the image as a container, enter the following command:

```
docker run -t -v {PATH_TO_FASTA_FOLDER}:/opt/illumina/fasta-folder:ro -v
   {PATH_TO_ANALYSIS_FOLDER}:/opt/illumina/analysis-folder:rw dragen-
   covid-lineage-tools:1.0.1 --update-nextclade --update-pangolin --max-
   ambig 0.6
```

For descriptions of the options used, refer to the following table:

Option	Description
`{PATH_TO_ANALYSIS_FOLDER}`	Set to an absolute folder path on your server. The analysis output is saved to the folder location.
`{PATH_TO_FASTA_FOLDER}`	Set to an absolute folder path on your server that contains your FASTA files. FASTA files can be nested inside the root folder. For example, the output from the Illumina DRAGEN COVID Pipeline produces a Sample_Analysis folder, which contains {Sample_ID} subfolders. Inside of the subfolders are the sample FASTA files. Point to the root Sample_Analysis folder and the software recursively finds all the nested FASTA files.
`--max-ambig`	Specifies the maximum proportion of Ns allowed for Pangolin to attempt assignment. The default value is 0.5.
`--update-nextclade`	To update the NextClade knowledge base on the software, include the `--update-nextclade` flag. Make sure the server is connected to the internet for the update to occur.
`--update-pangolin`	To update the Pangolin knowledge base on the software, include the `--update-pangolin` flag. Make sure the server is connected to the internet for the update to occur.

Outputs

The DRAGEN COVID Lineage Tools software generates the following outputs:

Output	Path	Description
Combined Clade Report	./analysis-folder/combined_clade_report.csv	A combined CSV file containing all sample clade reports.
Combined FASTA	./analysis-folder/combined_fasta.fasta	A concatenated file of all processed FASTA files.
Combined Lineage Report	./analysis-folder/combined_lineage_report.csv	A combined CSV file containing all sample lineage reports.
Log File	./analysis-folder/covid-lineage-tools-<timestamp>.log	Details the functions performed by the software and any errors encountered.
Per Sample Clade Reports	./analysis-folder/{sample-id}/{sample-id}.clade_report.csv ./analysis-folder/{sample-id}/{sample-id}.clade_report.tsv ./analysis-folder/{sample-id}/{sample-id}.clade_report.json	Clade reports produced by NextClade for each sample. The report is named after the sample ID found in the FASTA file name. NextClade generates the same report in following file formats: • csv • tsv • json For example, the file sample1.fasta would have the following reports: • sample1.clade_report.csv • sample1.clade_report.tsv • sample1.clade_report.json
Per Sample Lineage Report	./analysis-folder/{sample-id}/{sample-id}.lineage_report.csv	A lineage report produced by Pangolin for each sample. The report is named after the sample ID found in the FASTA file name. For example, the file sample1.fasta would have a report with the name sample1.lineage_report.csv.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Website: www.illumina.com
Email: techsupport@illumina.com

Illumina Technical Support Telephone Numbers

Region	Toll Free	International
Australia	+61 1800 775 688	
Austria	+43 800 006249	+43 1 9286540
Belgium	+32 800 77 160	+32 3 400 29 73
Canada	+1 800 809 4566	
China		+86 400 066 5835
Denmark	+45 80 82 01 83	+45 89 87 11 56
Finland	+358 800 918 363	+358 9 7479 0110
France	+33 8 05 10 21 93	+33 1 70 77 04 46
Germany	+49 800 101 4940	+49 89 3803 5677
Hong Kong, China	+852 800 960 230	
India	+91 8006500375	
Indonesia		0078036510048
Ireland	+353 1800 936608	+353 1 695 0506
Italy	+39 800 985513	+39 236003759
Japan	+81 0800 111 5011	
Malaysia	+60 1800 80 6789	
Netherlands	+31 800 022 2493	+31 20 713 2960
New Zealand	+64 800 451 650	
Norway	+47 800 16 836	+47 21 93 96 93
Philippines	+63 180016510798	
Singapore	1 800 5792 745	
South Korea	+82 80 234 5300	
Spain	+34 800 300 143	+34 911 899 417
Sweden	+46 2 00883979	+46 8 50619671
Switzerland	+41 800 200 442	+41 56 580 00 00
Taiwan, China	+886 8 06651752	
Thailand	+66 1800 011 304	
United Kingdom	+44 800 012 6019	+44 20 7305 7197
United States	+1 800 809 4566	+1 858 202 4566
Vietnam	+84 1206 5263	

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download from support.illumina.com.

Document # 1000000158680 v01

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

illumina®