# xGen™ SARS-CoV-2 Amplicon Panel dockerized data analysis guidelines

This guide is for customers using the xGen SARS-CoV-2 Amplicon Panel who are comfortable with command line tools and want a ready-to-use variant calling analysis workflow that runs on local Unix machines. A pre-packaged Docker image can be downloaded from the Docker hub and run using the accompanying pipeline from Github to obtain the consensus sequence and variant (clade and subclade) designation from paired-end FASTQ files (Figure 1). The expected outputs and file descriptions are listed in the final section below. All the necessary tools and reference files are pre-installed and configured in a Docker image posted in a Docker hub repository.

## System Requirements

The dockerized analysis workflow is compatible with any operating system (OS) as long as Docker (Linux) or Docker Desktop is installed (Mac OS & Windows) and at least 8 GB memory (≥32 GB RAM recommended) is available to run the recommended pipeline.

1. Linux server (Ubuntu or similar)
2. Mac OS 10.14.6/Mojave and above
3. Windows 10 and above (contact Application support for guidelines)

## Steps to run analysis:

**Note:** Although the platform name still says Swift it is compatible with the xGen SARS-CoV-2 Amplicon Panel

1. Obtain the Swift Sarscov2 analysis Docker image from Docker Hub using the following command: (Docker Hub login may be required) docker pull swiftbiosci/sarscov2analysis:latest
2. Obtain pipeline (*run_swift_sarscov2_docker.sh*) and respective masterfile (Sarscov2 [v1 or v2], Sgene) to run the above image from Github using the following command:
   Git clone https://github.com/swiftbiosciences/sarscov2analysis_docker.git
3. Copy the above pipeline to a folder along with FASTQ files.

Run the analysis with offline mode (default -v) or online mode (-u) as (from the linux terminal) run_swift_sarscov2_docker.sh -v masterfile.txt


**Figure 1. Analysis overview.**

With the rate at which SARS-CoV-2 is spreading and evolving the pangolin version in the docker pipeline is no longer up to date. However, the consensus FASTA generated by the pipeline can be easily used with the live, current version of Pangolin for lineage calling found here.

## Understanding pipeline results and output

Below is a summary of pipeline example results and outputs. The three "run-level" (all samples summary) reports for quick results are:

a. The excel file summarizes panel performance for metrics, including Aligned (%), PCT_10X, Mean Coverage, Coverage Uniformity (%), Dimers (%), and Human_Aligned (%).

b. Nextclade clade (lineage) identification report.

c. Pangolin variant designation report.

Additional "per sample" result files and their respective folders/directories (*dirname/*) are as follows:

a. Fastqc (.html): *fastqc/*

b. Alignment files (.bam & .bai): *bam/*

c. Picard target PCR metrics (.txt): *metrics/*

d. Per-amplicon coverage plots (.pdf): *plots/*

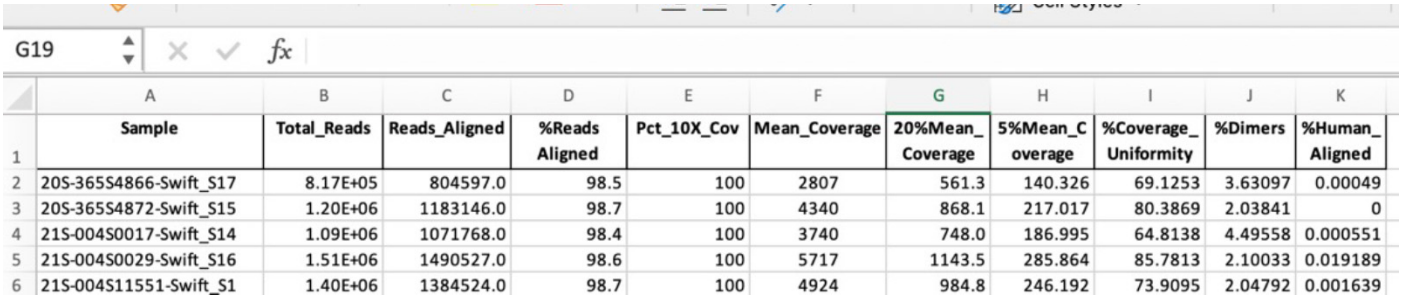
e. GATK HaplotypeCaller variant call files (*gatkHC.vcf): *vcf/*

f. Bcftools consensus fasta (.fa): *consensus/*

g. Nextclade and Pangolin QC (.csv): *nextclade/*, *pangolin/*

h. Standard output from each processing step is saved with respective tool's name-appended log file which can be found in the *logs/* folder.

## Sample Report Output from the pipeline



G19

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Sample | Total_Reads | Reads_Aligned | %Reads Aligned | Pct_10X_Cov | Mean_Coverage | 20%Mean_ Coverage | 5%Mean_C overage | %Coverage_ Uniformity | %Dimers | %Human_ Aligned |
| 2 | 20S-365S4866-Swift_S17 | 8.17E+05 | 804597.0 | 98.5 | 100 | 2807 | 561.3 | 140.326 | 69.1253 | 3.63097 | 0.00049 |
| 3 | 20S-365S4872-Swift_S15 | 1.20E+06 | 1183146.0 | 98.7 | 100 | 4340 | 868.1 | 217.017 | 80.3869 | 2.03841 | 0 |
| 4 | 21S-004S0017-Swift_S14 | 1.09E+06 | 1071768.0 | 98.4 | 100 | 3740 | 748.0 | 186.995 | 64.8138 | 4.49558 | 0.000551 |
| 5 | 21S-004S0029-Swift_S16 | 1.51E+06 | 1490527.0 | 98.6 | 100 | 5717 | 1143.5 | 285.864 | 85.7813 | 2.10033 | 0.019189 |
| 6 | 21S-004S11551-Swift_S1 | 1.40E+06 | 1384524.0 | 98.7 | 100 | 4924 | 984.8 | 246.192 | 73.9095 | 2.04792 | 0.001639 |

## Sample "Nextclade report" (truncated)

| seqNan | substitutions | clade | lMutat | aminoacidCh | minoacidCha | insertions | otalInsertion | deletions | totalGaps | missing | totalMissing |
|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | C241T,T1187G,A1715G,A1810G | 20A | 25 | ORF1a:S308 | 15 | 1140:A,1633 | 8 | 511-525,686-695,2 | 32 | 1-26,27803- | 81 |
| S2 | C241T,C882T,T1187G,C1912T,C | 20A | 15 | ORF1a:A206 | 8 | 6700:T,1060 | 5 | 686-695,11082 | 10 | 1-26,29858- | 76 |
| S3 | A187G,C241T,C1059T,T1187G,G | 20C | 29 | ORF1a:T265 | 18 | 5677:T,6700 | 8 | ########## | 3 | 1-26,29861- | 76 |
| S4 | C241T,C1059T,C2395T,T2597C,C | 20C | 31 | ORF1a:T265 | 16 | 6700:T,7251 | 7 | 5343,9268-9274,99 | 23 | 1-26,29860- | 76 |

## Sample "Pangolin Lineage report"

| Sample | lineage | robabili | EARN_ | status | note | taxon | LineageName | Most_common_cou | Date_Range | Numberof_t | Days_sinceL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | B.1.404 | 1 | ##### | passed_qc | | NC_045512.2 | B.1.404 | USA Mexico | June-16 Dece | 81 | 72 |
| S2 | B.1 | 1 | ##### | passed_qc | | NC_045512.2 | B.1 | USA UK Spain | January-28 Ja | 19204 | 57 |
| S3 | B.1.2 | 1 | ##### | passed_qc | | NC_045512.2 | B.1.2 | USA UK Australia | March-09 Jar | 3872 | 57 |
| S4 | B.1.429 | 1 | ##### | passed_qc | | NC_045512.2 | B.1.429 | USA New_Zealand | October-16 D | 136 | 62 |

# References:


Andrews S. FastQC. Babraham Bioinformatics. Accessed January 7,2022. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Burrow-Wheeler Aligner. Accessed January 7, 2022. https://github.com/lh3/bwa

Picard. Broad Institute. Accessed January 7, 2022. https://broadinstitute.github.io/picard/

Genome Analysis Toolkit. Broad Institute. Access January 7, 2022. https://gatk.broadinstitute.org/hc/en-us

Bcftools. Samtools. Accessed January 7, 2022. https://github.com/samtools/bcftools

Nextclade. Nextstrain. Accessed January 7, 2022. https://clades.nextstrain.org/

Cov-lineages/pangolin. Accessed January 7, 2022. https://github.com/cov-lineages/pangolin

# xGen™ SARS-CoV-2 Amplicon Panel Dockerized Data Analysis Guidelines

Technical support: applicationsupport@idtdna.com