PROTOCOL

Computational Protocol for Assembly and Analysis of SARS-nCoV-2 Genomes

Mukta Poojary[1,2], Anantharaman Shantaraman[1], Bani Jolly[1,2] and Vinod Scaria[1,2]

[1]CSIR Institute of Genomics and Integrative Biology, Mathura Road, Delhi
[2]Academy of Scientific and Innovative Research (AcSIR)

[*]Corresponding email: MP mukta.poojary@igib.in; AS anantharaman3898@gmail.com; BJ bani.jolly@igib.in; VS vinods@igib.in

ABSTRACT

SARS-CoV-2, the pathogen responsible for the ongoing Coronavirus Disease 2019 pandemic is a novel human-infecting strain of Betacoronavirus. The outbreak that initially emerged in Wuhan, China, rapidly spread to several countries at an alarming rate leading to severe global socio-economic disruption and thus overloading the healthcare systems. Owing to the high rate of infection of the virus, as well as the absence of vaccines or antivirals, there is a lack of robust mechanisms to control the outbreak and contain its transmission. Rapid advancement and plummeting costs of high throughput sequencing technologies has enabled sequencing of the virus in several affected individuals globally. Deciphering the viral genome has the potential to help understand the epidemiology of the disease as well as aid in the development of robust diagnostics, novel treatments and prevention strategies. Towards this effort, we have compiled a comprehensive protocol for analysis and interpretation of the sequencing data of SARS-CoV-2 using easy-to-use open source utilities. In this protocol, we have incorporated strategies to assemble the genome of SARS-CoV-2 using two approaches: reference-guided and *de novo*. Strategies to understand the diversity of the local strain as compared to other global strains have also been described in this protocol.

KEYWORDS: Next generation sequencing SARS-CoV-2, Whole Genome Sequencing SARS-CoV-2, Genome assembly protocol, *de novo* assembly SARS-CoV-2

Citation: Poojary M Et al (2020) Computational Protocol for Assembly and Analysis of SARS-nCoV-2 Genomes. Research Reports 4: e1-e14 doi:10.9777/rr.2020.10001

Introduction

The SARS-related coronaviruses (SARS-CoV) are enveloped, positive-sense, single-stranded RNA viruses. The SARS-CoV viruses have one of the largest genomes among RNA viruses, with a genome of about 30,000 base pairs. The RNA genome has a 5′ methylated cap and a 3′ polyadenylated tail. The SARS-CoV-2 is unique among known coronavirus strains in the presence of the 5′ cap and a polybasic cleavage site, a feature that is commonly known to cause an increase in pathogenicity and transmissibility. The 3′ tail of the SARS-CoV viruses allows them to be directly translated by the host cell's ribosome on viral entry. Like other coronaviruses, the SARS-CoV genome has two large overlapping open reading frames (ORFs) both of which produce polyproteins [1].

SARS-CoV-2 is a strain of SARS-related Coronavirus that is responsible for the ongoing COVID-19 (Coronavirus Disease 2019) pandemic. The disease was first identified during an outbreak in Wuhan, China, and has continued to spread at an alarming rate globally causing major social and economic disruptions in a large number of nations and thus causing significant strain to the healthcare of these countries [2,3]. COVID-19 testing currently involves real time Reverse Transcription Polymerase Chain Reaction (RT-PCR) assays as well as surveillance using serology [4,5].

Sequencing of the viral genome can potentially help understand the molecular underpinnings of pathology as well as provide an opportunity to understand the epidemiology of viruses. The genome can also provide a way to type strains and understand unique characteristics including mutation rates and evolution of the pathogen [6,7]. It is also increasingly evident that the genome of the strains could also provide insights into the global diversity of the virus.

Next Generation sequencing (NGS) provides a scalable high-throughput approach to understand the genomes of organisms. NGS is now widely used in clinical and research settings to understand pathogens and their genetic underpinnings. This protocol has been put together to aid bioinformaticians involved in genome analysis and interpretation of next generation sequence data for the SARS-CoV-2 genome. Two approaches for genome assembly, reference based as well as de novo approaches are discussed here. All software and datasets used are available in the public domain and emphasis has been placed to use only tools and resources that are freely available.

Pre-requisites

Knowledge and Skills

Basic knowledge of the Next Generation Sequencing file formats and working experience in the command line on Linux operating systems is expected from the user.

Computational Hardware

A workstation or server with approximately 8GB RAM with a Linux based operating system should be sufficient to run all these steps with ease.

The following steps are under the assumption that the user has a Debian based Linux distribution as the Operating System and has superuser privileges. The commands have been validated on Ubuntu (18.04 LTS) Linux Distribution. The protocol is written with the assumption that the sequence reads are based on Illumina sequencing chemistry (California, USA) and generated as per standard protocols.

Software and Datasets

Before proceeding with the analysis, download and install all the required tools and its dependencies.

The following software is required to be installed before performing the analysis.

Table 1

Table 1. List of requisite tools used for different steps in the analysis pipeline.

Analysis step	Tool
Quality check	FastQC
Trimming and Adapter removal	Trimmomatic
Pre and post processing	samtools
Alignment	HISAT2
Evaluation of species diversity	Kraken and Krona
Consensus genome and variant calling	BEDTools, Seqtk, BCFtools and VarScan
De novo Genome assembly	MEGAHIT and SPAdes
Assembly evaluation and statistics	QUAST
Phylogenetic Analysis	Clustal omega, MAFFT and MEGA

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Trimmomatic[8], samtools [9], HISAT2 [10], BEDTools[11], BCFtools (https://github.com/samtools/bcftools), Seqtk[12], VarScan[13,14], Kraken [15], Krona [16],MEGAHIT [17], SPAdes[18], QUAST[19], MAFFT[20], MEGA[21].

The following datasets are required for the analysis from Sequence Read Archive (SRA) - SRP251618.

This dataset is a single end viral RNA sequencing data from University of Washington from 14 patients infected with SARS-CoV-2.

The following section provides details on how to systematically install all the prerequisites and download the requisite datasets. Figure 1 provides an overview of the different steps involved in the processing of sequence files.

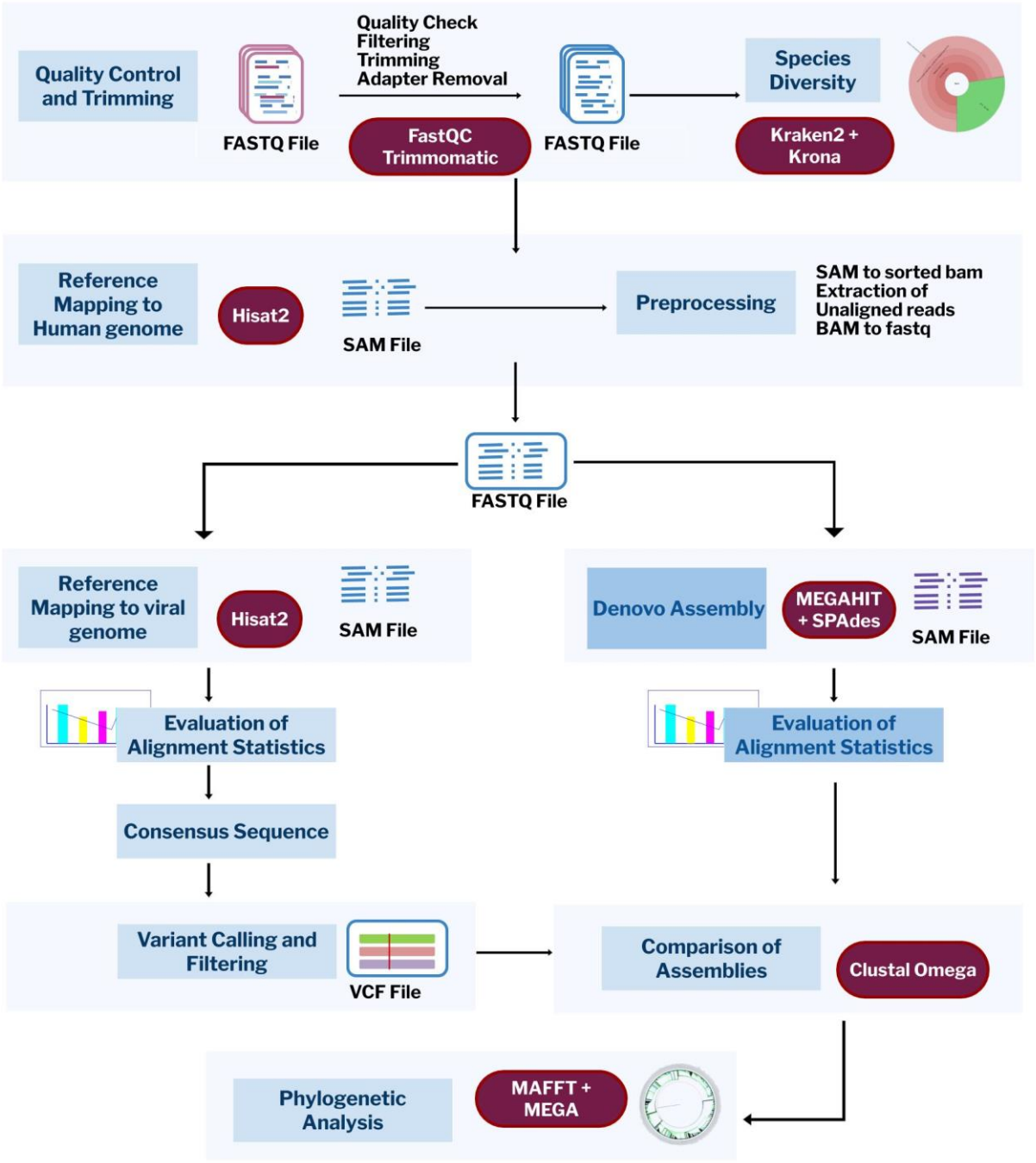

Figure 1. The different steps described in this protocol and the software used in each of the analysis steps.

Downloading and Installation of requisite software tools

FastQC

FastQC is a simple, easy-to-use, fast tool used for quality control analysis of high throughput sequencing datasets. It provides a graphical summary of several parameters required to assess the quality of sequencing data such as overall ready quality, per tile quality, sequence length, adapter contamination and over- representated sequences. To download and install the tool, use the following command

```
sudo apt-get install -y fastqc
```

Trimmomatic

Trimmomatic is an easy-to-use tool which performs read filtering and trimming along with removal of any adaptor contaminated sequences. To download and install the tool, use the following command

```
sudo apt-get install -y trimmomatic
```

SAMtools

SAMtools is a set of utilities for post-processing different sequence alignments such as the SAM, BAM and CRAM formats. To download and install the utility, use the following command

```
sudo apt-get install -y samtools
```

HISAT2

HISAT2 is a fast and efficient assembly tool for Next Generation sequence data. To download the software, use the following command

```
sudo apt-get install -y hisat2
```

If the command gives an error saying "Following repository not found" update your repositories using command,

```
sudo apt-get update
```

BEDTools

BEDTools suite provides a multitude of tools to perform various genomic analysis such as intersect, merge, shuffle etc. involving multiple file formats widely used in genomic studies such as BAM, BED, VCF, GTF/GFF. To download the suite, use the following command

```
sudo apt-get install -y bedtools
```

BCFtools

BCFtools is a set of utilities which help in manipulating the Variant call format (VCF) and its binary format (BCF). To download the utility, use the following command

```
sudo apt-get install -y bcftools
```

Seqtk

Seqtk is an ultrafast utility to parse and process sequences in FASTA and FASTQ formats. To download the utility, use the following command

```
sudo apt-get install -y seqtk
```

VarScan

VarScan is a robust mutation caller for targeted as well as high throughput whole genome or exome scale data generated from different sequencing platforms. To download the software, use the following command

```
git clone https://github.com/dkoboldt/
varscan.git
```

MEGAHIT

MEGAHIT is an ultra-fast denovo assembler which makes use of succinct de Bruijn graph (SdBG) to achieve low memory assembly. To download the assembler, use the command

```
wget
https://github.com/voutcn/megahit/rele
ases/download/v1.2.9/MEGAHIT-1.2.9-
Linux-x86_64-static.tar.gz
```

This command should download the file MEGAHIT-1.2.9-Linux-x86_64-static.tar.gz. Proceed to unzip the file using command

```
tar zvxf MEGAHIT-1.2.9-Linux-x86_64-static.tar.gz
```

The executable can be found in the bin folder and run by the command ./megahit

Kraken

Kraken is an efficient tool to assign taxonomic labels to metagenomic data. The software can be downloaded from the GitHub using the command

```
wget https://github.com/DerrickWood/kraken2/archive/master.zip
```

Proceed to unzip the downloaded file using command, followed by installation

```
unzip master.zip
```

Change directory to kraken2-master

```
cd kraken2-master/
```

Install kraken

```
sudo ./install_kraken2.sh /usr/bin
```

Downloading minikraken database

The minikraken database is a pre-built 8 GB database built using complete bacterial, archaeal, and viral genomes in RefSeq (as of March, 2020) and available for download at the Kraken website.

```
wget ftp://ftp.ccb.jhu.edu/pub/data/kraken2_dbs/minikraken_8GB_202003.tgz
tar -xvf minikraken_8GB_202003.tgz
export KRAKEN2_DB_PATH="<path/to/folder/containing/minikraken/database>"
```

Krona

Krona is a tool which enables interactive visualization of metagenomic classification from Kraken datasets and also displays the read abundances.

You may need to install GitHub using the following command, if not installed already.

```
sudo apt-get install git-core git-gui git-doc
```

Krona can be downloaded using the following GitHub link

```
git clone https://github.com/marbl/Krona.git
```

Change directory to KronaTools

```
cd Krona/KronaTools/
```

Proceed for installation using the following command

```
sudo ./install.pl
mkdir taxonomy
./updateTaxonomy.sh
./updateAccessions.sh
```

SPAdes

SPAdes (St. Petersburg genome assembler) provides genome assemblies using data from different sequencing platforms such as Illumina, Ion torrent and PacBio.

```
wget http://cab.spbu.ru/files/release3.14.0/SPAdes-3.14.0-Linux.tar.gz
tar -xzf SPAdes-3.14.0-Linux.tar.gz
```

The scripts can be found in this folder: SPAdes-3.14.0-Linux/bin/

QUAST

QUAST (QUality ASsessment Tool) toolkit generates metrics which describe the quality of the assembly. To download and install the tool, use the following commands

```
git clone https://github.com/ablab/quast.git
cd quast
```

```
python setup.py install_full
```

MAFFT

MAFFT (Multiple Alignment using Fast Fourier Transform) is a multiple sequence alignment tool. To download and install the tool, use the following command

```
sudo apt-get install mafft
```

MEGAX

MEGAX (Molecular Evolutionary Genetics Analysis version 10) is a tool that can analyze molecular evolution and generate phylogenetic trees. To download and install MEGAX Command Line Interface, use the following commands

```
wget
https://www.megasoftware.net/do_force_
download/megacc_10.1.1_amd64_beta.tar.
gz
tar -zxvf
megacc_10.1.1_amd64_beta.tar.gz
```

The executable script for MEGAX can be found in the current folder.

Download Reference Dataset for Phylogenetic Analysis

The Global Initiative on Sharing All Influenza Data (GISAID) gives public access to the most complete repository of sequencing data for SARS-CoV-2[22]. We have downloaded the dataset of 2,124 (as of 27 March, 2020) complete sequences of coronavirus. This can be done by registering for an account on GISAID. Account activation can take upto a few hours, after which the dataset can be downloaded by logging into the GISAID database and then navigating to the *Browse* option in the EpiCoV™ database.

Download Reference Genome

Before proceeding with the analysis, we also need to download the reference genomes for both Human as well as SARS-CoV2.

The latest version of the human genome can be downloaded from

```
https://www.gencodegenes.org/human/
```

The latest version of the SARS-CoV-2 genome can be downloaded from NCBI. We have downloaded the genome with the accession number NC_045512.2

Indexing the Reference genome

Both the reference genomes need to be indexed with the HISAT2 tool.

Input: Downloaded reference genome FASTA files

```
hisat2-build
<genome/GRCh38.p13.genome.fa>
<genome/GRCh38.p13.genome>
hisat2-build <genome/NC_045512.2.fna>
<genome/NC_045512.2>
```

Downloading the test dataset

The FASTQ files of viral RNA from patients affected with SARS-CoV-2 can be downloaded from SRA using the following link. The data used in the demonstration is RNA sequencing data of 14 patients infected with SARS-CoV-2 sequenced by University of Washington.

```
https://trace.ncbi.nlm.nih.gov/Traces/
sra/?study=SRP251618
```

Quality Assessment and Trimming

The quality assessment of the RNA seq data can be analyzed using FastQC tool.

Input: FASTQ file

```
fastqc <fastq file>
```

If the quality of the data is satisfactory, you may proceed to the next step. If the quality of the data is poor, Trimmomatic may be used to trim the adaptors and bad quality reads.

Input: raw FASTQ file

The following command may be used for single end reads,

```
TrimmomaticSE <input.fastq>
<output.fastq>
ILLUMINACLIP:</usr/share/trimmomatic/T
ruSeq3-SE.fa>:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15
MINLEN:36
```

```
java -jar trimmomatic- <version> .jar
SE <input.fastq> <output.fastq>
ILLUMINACLIP:TruSeq3-SE:2:30:10
LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

For paired end reads, use the following command.

```
TrimmomaticPE <input_fwd.fastq>
<input_rev.fastq>
<output_fwd_paired.fastq>
<output_fwd_unpaired.fastq>
<output_rev_paired.fastq>
<output_rev_unpaired.fastq>
ILLUMINACLIP:/usr/share/trimmomatic/Tr
uSeq3-SE.fa:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15
MINLEN:36
```

```
java -jar trimmomatic- <version> .jar
PE <input_fwd.fastq> <input_rev.fastq>
<output_fwd_paired.fastq>
<output_fwd_unpaired.fastq>
<output_rev_paired.fastq>
<output_rev_unpaired.fastq>
ILLUMINACLIP:TruSeq3-SE:2:30:10
LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

The different parameters used in the command are explained below

SE/ PE: Single end or Paired End sequencing data
ILLUMINACLIP: Remove adaptor and Illumina-related sequences
SLIDINGWINDOW: Number of basepairs to be read in a single window: Minimum quality
LEADING: Number of basepairs to be trimmed from the start of the read
TRAILING: Number of basepairs to be removed from the end of the read
MINLEN: Minimum length of read after trimming

Assembly and Analysis

Analysis of species diversity present in the RNA sample

The analysis of the species diversity in the RNA sample can be done using metagenomic approach. Kraken tool will be used for estimating the species diversity in the data and Krona tool will be used for visualization of the diversity.

Input: trimmed FASTQ file

For unpaired reads,

```
kraken2 --db minikraken_8GB_20200312 -
-threads 16 --report <file.kreport>
<trimmed.fastq> > <trimmed.kraken>
```

For paired reads,

```
kraken2 --db minikraken_8GB_20200312 -
-threads 16 --report <file.kreport> --
paired <file1.fastq> <file2.fastq> >
<trimmed.kraken>
```

Visualization by Krona

Input: Kraken output

```
ktImportTaxonomy -s 3 -t 4 -o
<visualization_output>.html
<trimmed.kraken>
```

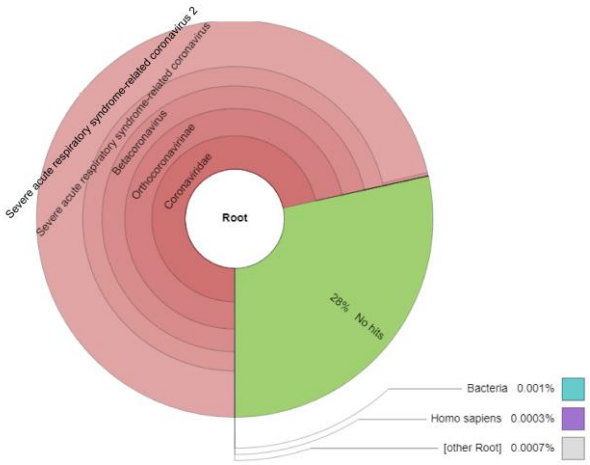

Figure 2. Pictorial representation of abundance of SARS-CoV-2 reads by Krona

Reference based assembly against the human reference genome

After verifying that a considerable amount of the reads belong to the SARS-CoV-2, we can perform reference based assembly against the human reference genome. This is done in order to discard reads that specifically belong to the human host.

Reference based assembly is only possible, if we are sure of the organism and also the reference genome sequence of that organism's genome is known. It is computationally less intensive and more accurate.

RNA sequencing, in this case, was done with the help of a certain human bodily fluid sample suspected to have a considerable viral load. Hence, there would be a considerable contamination of human RNA which needs to be eliminated. This can be achieved by aligning all the reads against the human reference genome and extracting all reads not aligning to the human reference genome.

Input: Indexed reference genome and trimmed FASTQ file

Reference based assembly can be done with the tool HISAT2 using the following command.

```
hisat2 -x <Humangenomeprefix> -U
<unpaired_input.fastq> -S <human.sam>
-p 16 --dta-cufflinks --summary-file
<humanUnpaired.log>
```

Pre-processing and extracting unaligned reads from the BAM file

In order, to extract the unaligned reads, the sam file needs to be sorted and converted into a bam file. This can be done with the help of samtools.

Input: SAM output after aligning to human reference genome

```
samtools sort -@ 8 <human.sam> -o
<human.bam> -O BAM
```

The unaligned reads can be extracted using samtools via three steps.

Input: BAM file generated from previous step

Extracting unmapped reads whose mate is mapped

```
samtools view -u -f 4 -F264<human.bam>
> temp1.bam
```

Extracting mapped reads whose mate is unmapped

```
samtools view -u -f 8 -F 260
<human.bam> > temp2.bam
```

Extracting reads whose mate and pair are both unmapped

```
samtools view -u -f 12 -F 256<
human.bam> > temp3.bam
```

Merging the three temporary bam files

```
samtools merge -u - temp[123].bam |
samtools sort -n -o <unmapped.bam>
```

Converting the BAM file to FASTQ file

The bam file of the unaligned reads against the human reference genome needs to be converted

to the fastq file inorder to be realigned to the viral reference genome and de novo assembly of the viral genome.

Input: Merged BAM file generated from previous step

```
bamToFastq -i <unmapped.bam> -fq
<unmapped.fq>
```

Alignment to the viral reference genome

The FASTQ file containing unaligned reads can be used to be aligned to the viral reference genome.

Input: Indexed SARS-CoV2 reference genome

```
hisat2 -x
<NC_045512.2_viral_reference_genome> -
U <unmapped.fq> -S <covid.sam> -p 16 -
-dta-cufflinks --summary-file
<unmapped.log>
```

Evaluation of Alignment statistics to the viral reference genome

The alignment statistics can be evaluated with the help of samtools using the following command

```
samtools flagstat <covid.sam>
```

The output of flagstat provides three important results, the Total Number of Reads present before alignment, Total Number of aligned Reads and Alignment Percentage. The alignment percentage against the viral genome should be atleast above 80%.

Table 2. Summarised Alignment statistics against the SARS-CoV-2 genome.

Sample ID	Number of mapped reads	Total number of reads	Alignment Percentage
WA3-UW1	128595	131949	97.46%

Generating a consensus sequence

A consensus genome is the sequence representation of the reads aligned to the genome.

The consensus file can be generated with the help of samtools and bcftools utilities using the following command.

```
samtools sort -@ 8 <covid.sam> -o
<covid.bam> -O BAM

samtools faidx
<covid_reference_genome>

samtools mpileup -uf
<covid_reference_genome> <covid.bam> |
bcftools call -c | vcfutils.pl vcf2fq
> <consensus.fq>

seqtk seq -aQ64 -q20 -n N
<consensus.fq> > <consensus.fasta>
```

The variants can be called with the help of samtools and bcftools utilities using the following command.

```
samtools mpileup -uf
<covid_reference_genome> <covid.bam> |
bcftools call -cv -Ob > <variant.bcf>
```

The variants file can be converted to a readable VCF format from the binary BCF format with the help of bcftools utilities using the following command.

```
bcftools view <variant.bcf> >
<variant.vcf>
```

The variants can also be called with the help of VarScan tool using the following command,

```
samtools mpileup -f
<covid_reference_genome> <covid.bam> >
<covid.pileup>

java -jar VarScan.v2.4.4.jar
mpileup2cns <covid.pileup> --output-
vcf 1 --variants><covid.vcf>
```

The minimum read depth to call a variant is at least 10 reads which should support the presence of the variant.

De Novo assembly of the viral reference genome

De Novo assembly of the viral reference genome can be done using both the following two tools, and the assemblies thus produced can be compared.

1. MEGAHIT
2. SPAdes

Input: FASTQ file of reads which were not aligned correctly to the human reference genome.

De Novo assembly with the help of Megahit tool can be performed using the following command.

```
./megahit -r <unmapped.fq> -o
<denovo_covid_genome>
```

After the assembly with MEGAHIT is completed, the log file has all the details of the assembly results. The assembled contigs are saved in final.contigs.fa file in the output folder named while executing the command. This file will be used to understand and evaluate the assembly quality and statistics.

The contig assembled should have a length approximately equal to the known genome size of SARS-CoV-2 which is ~30,000bps.

De Novo assembly with the help of SPAdes tool can be performed using the following command.

```
python spades.py -t <8> -s
<unmapped.fq> -o
<spades_output_folder>
```

After the assembly with SPAdes is completed, the log file has all the details of the assembly results. The assembled contigs are saved in contigs.fa file in the output folder named while executing the command. This file will be used to understand and evaluate the assembly quality and statistics.

Evaluation of the de novo assembly statistics

The contigs formed from the *de novo* assembly can be evaluated on the basis of following parameters

1. Number of large contigs
2. Length of largest contig
3. N50 score : length of largest contig which covers at least 50% of the assembly

The quality of the assembly can be evaluated with the help of Quast toolkit. The quast tool calculates several metrics which are used to evaluate the quality of the genome assembly generated with the help of MEGAHIT and SPAdes. The following command generates a comprehensive report of the assembly metrics.

Input: contig files generated by MEGAHIT and SPAdes.

```
python quast.py <contigs.fasta> -r
<NC_045512.2.fna covid reference
genome> --single <trimmed.fastq> -o
<quast_output>
```

The report.txt file in the output folder defined while executing the command gives a summary of the assembly metrics along with the above mentioned parameters.

Evaluating the similarity between the assembled genome and the reference genome

The consistency between both the reference and the assembled genomes can be evaluated by Multiple sequence alignment. Multiple sequence alignment can be done online with the help of Clustal omega tool[23]. The sequences of the contigs generated by the two assemblers and the reference sequence can be submitted in the form of FASTA format to the tool.

The result will be displayed in the form of the following representation.

```
NC_045512.2     CAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCT   184
spades          CTGGGGGCAAATTGTGCAATTTGCGGCCAATGTTTGTAATCAGTTCCTTGTCTGATTAGT   715
megahit         CTGGGGGCAAATTGTGCAATTTGCGGCCAATGTTTGTAATCAGTTCCTTGTCTGATTAGT   720
                * *       * *    ****   * *        *  *  ***  ** ****   *  *

NC_045512.2     GCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTC   244
spades          TCCTGGTCCCCAAAATTTCCTTGGGTTTGTTCTGGACCACGTCTGCCGAAAGCTTGTGTT   775
megahit         TCCTGGTCCCCAAAATTTCCTTGGGTTTGTTCTGGACCACGTCTGCCGAAAGCTTGTGTT   780
                 *  * * *  *   **** *  *  ***     ** **      *    ** **  **

NC_045512.2     CGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACAC   304
spades          ACATTGTATGCTTTAGT---GGCAGTACGTTTTTGCCGAGGCTTCTTAGAAGCCTCAGCA   832
megahit         ACATTGTATGCTTTAGT---GGCAGTACGTTTTTGCCGAGGCTTCTTAGAAGCCTCAGCA   837
                    ***   *   **    *   *   *  **  **  ** ***   **     **
```

Figure 3. Sample output of Clustal omega showing the similar / dissimilar nucleotide sequence positions between the two *de novo* assemblies and the reference viral genome.

Summarising the Assembly for Submission and Sharing

Visualizing the evolution and diversity of different SARS-CoV-2 strains can help increase our epidemiological understanding of the pathogen. Real-time interactive data visualization can be done using Nextstrain[24]. Nextstrain is a collection of open-source tools for phylodynamics analysis and visualization. To use Nextstrain for visualizing our data, we need the following files:

1. A single FASTA file having a collection of pathogen sequences that are to be analyzed.
2. The reference sequence in GenBank (.gb) format downloaded from NCBI using accession number NC_045512.2.
3. A tab-delimited metadata file that describes the sequences given in the FASTA file.

Required fields for the metadata file: Strain, Virus, Date

Additional fields (if using published data): Accession, Authors, URL, Title, Journal, Paper_URL.

To infer ancestral traits, additional information fields such as region, country and city can be added to the metadata.

Nextstrain requires the information added against the fields to be in a particular format.

A sample spreadsheet has been prepared to show the formats that are accepted by Nextstrain. The information filled against all fields for a strain in the Sample Metadata spreadsheet is taken from the metadata section of the GenBank record for that strain.

A Sample Metadata Spreadsheet is linked here as Supplementary Data 1

Phylogenetic analysis

A phylogenetic tree is a branching diagram that depicts the evolutionary ties between a group of organisms based on their genetic relatedness. For this analysis, we use the GISAID dataset of 2,124 complete sequences of coronavirus. Multiple sequence alignment for these sequences can be done using MAFFT.

Input: A single FASTA file containing 2,124 sequences of coronavirus from GISAID

```
mafft --thread 4
<gisaid_cov2020_sequences.fasta> >
<gisaid_cov2020_sequences_MSA.fasta>
```

The phylogenetic tree can then be generated from the multiple sequence alignment using MEGAX. We calculate the tree using the neighbor-joining method with 100 bootstrap replicates and maximum composite likelihood model. MEGAX requires an Analysis Options File (.mao) which specifies the analysis to be performed and the options that will be set during the analysis. The .mao file can be generated using the MEGA-Prototyper interface provided by MEGA. The .mao file with our required options looks like this:

```
[ MEGAinfo ]
ver                                 = 10200107-x86_64
[ DataSettings ]
datatype                            = snNucleotide
containsCodingNuc                   = True
MissingBaseSymbol                   = ?
IdenticalBaseSymbol                 = .
GapSymbol                           = -
Labelled Sites                      = All Sites
Labels to Include                   =
[ ProcessTypes ]
ppInfer                             = true
ppNJ                                = true
[ AnalysisSettings ]
Analysis                            = Phylogeny Reconstruction
Scope                               = All Selected Taxa
Statistical Method                  = Neighbor-joining
Phylogeny Test                      = ====================
Test of Phylogeny                   = Bootstrap method
No. of Bootstrap Replications       = 100
Substitution Model                  = ====================
Substitutions Type                  = Nucleotide
Model/Method                        = Maximum Composite Likelihood
Fixed Transition/Transversion Ratio = Not Applicable
Substitutions to Include            = d: Transitions + Transversions
Rates and Patterns                  = ====================
Rates among Sites                   = Uniform Rates
Gamma Parameter                     = Not Applicable
Pattern among Lineages              = Same (Homogeneous)
Data Subset to Use                  = ====================
Gaps/Missing Data Treatment         = Pairwise deletion
Site Coverage Cutoff (%)            = Not Applicable
Select Codon Positions              = 1st, 2nd, 3rd, Non-Coding
System Resource Usage               = ====================
Number of Threads                   = 4
Genetic Code Table                  = Not Applicable
Genetic Code                        = Not Applicable
```

Figure 4. Screenshot summary of the MEGA Analysis Options File (infer_NJ_nucleotide.mao) used for the phylogenetic analysis

Input: Multiple Sequence Alignment file in FASTA format produced by MAFFT

```
megacc -a <infer_NJ_nucleotide.mao> -d
<gisaid_cov2020_sequences_MSA.fasta> -
o <gisaid_cov2020_tree>
```

The command will generate a phylogenetic tree in Newick format (.nwk). The generated tree can be visualized using FigTree software[25]. FigTree can be

downloaded from https://github.com/rambaut/figtree/releases.

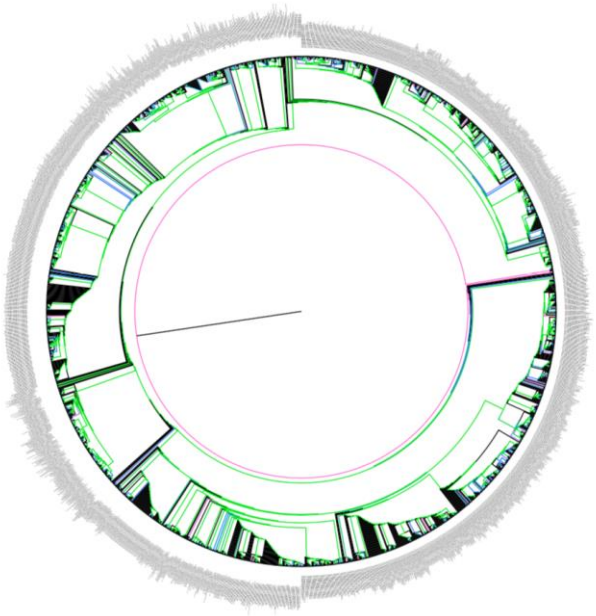

Figure 5. Phylogenetic tree of SARS-CoV-2 based on neighbor-joining method using GISAID dataset of 2,124 complete nucleotide sequences

In summary, we provide a comprehensive protocol for reference based as well as de novo assembly of SARS-nCoV-2 genomes. The protocol is based on free/open source software, which would allow the analysis to be replicated across the globe.

Code Availability

The configuration file for creating a custom Conda environment containing all the tools and their dependencies required for running this pipeline is provided as a GitHub repository (https://github.com/banijolly/vslab-ncov2019-genome).

Acknowledgements

Authors acknowledge help, support and assistance from Disha Sharma and Abhinav Jain in preparing the protocol and Bharath Ram, Prakrithi P, Ankit Pathak and Mohit Mangla for evaluating the protocol. This work was funded by the Council of Scientific and Industrial Research (CSIR India) through grants to Vinod Scaria, CSIR-IGIB. MP acknowledges a BINC Fellowship from the Department of Biotechnology. BJ acknowledges a GATE Fellowship from the Council of Scientific and Industrial Research. The funders had no role in the preparation of the manuscript or the decision to publish.

Conflict of interest statement

The author declares no competing or conflict of interests. The funders had no role in study design, writing of the manuscript and decision to publish.

Authors' contributions

VS conceptualised the protocol and sections. MP, BJ designed the protocol. MP, BJ, AS validated the steps. All authors contributed to writing the manuscript.

REFERENCES

van Boheemen, S. *et al.* Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio* **3**, (2012).

Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* vol. 395 689–697 (2020).

Emery, S. L. *et al.* Real-Time Reverse Transcription–Polymerase Chain Reaction Assay for SARS-associated Coronavirus. *Emerging Infectious Diseases* vol. 10 311–316 (2004).

Chan, J. F.-W. *et al.* Improved molecular diagnosis of COVID-19 by the novel, highly sensitive and specific COVID-19-RdRp/Hel real-time reverse transcription-polymerase chain reaction assay validated and with clinical specimens. *J. Clin. Microbiol.* (2020) doi:10.1128/JCM.00310-20.

Wang, C. *et al.* The establishment of reference sequence for SARS-CoV-2 and variation analysis. *Journal of Medical Virology* (2020) doi:10.1002/jmv.25762.

Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review* (2020) doi:10.1093/nsr/nwaa036.

Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* vol. 25 2078–2079 (2009).

Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–34 (2014).

lh. lh3/seqtk. *GitHub* https://github.com/lh3/seqtk.

Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* vol. 22 568–576 (2012).

Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* vol. 25 2283–2285 (2009).

Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* vol. 12 (2011).

Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).

Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

FigTree. http://tree.bio.ed.ac.uk/software/figtree/.