SARS-CoV-2 genomic landscape

Alt-text Figure 1 - The SARS-CoV-2 virus with its proteins

Schematic illustration of SARS-CoV-2: a purple sphere covered with green spikes. Spike, Nucleocapsid, Membrane, Envelope and ssRNA are indicated inside the sphere.

Alt-text Figure 2 - Various enzymes and the spike protein of SARS-CoV-2

Illustrative image of the SARS-CoV-2 genome and its proteins: a bar with rectangles representing SARS-CoV-2 genes and protein structures of protease, endoribonuclease and spike.

Overview of genomic, sub-genomic and anti-genomic sequences

Alt-text Figure 3 - SARS-CoV-2 replication cycle

SARS-CoV-2 replication cycle: a viral particle binds to the cell membrane and is internalised. Viral genomic and sub-genomic RNA is transcribed, and proteins are synthesised. A new viral particle is assembled and it is released from the cells to infect new host cells.

How is SARS-CoV-2 sequencing done?

Alt-text Figure 4 - Methods for SARS-CoV-2 genome sequencing

This is a figure about methods used for SARS-CoV-2 genome sequencing. DNA strands are represented by thick black lines. The figure is made up of three separate diagrams labelled A, B, and C. Diagram A is a workflow detailing Illumina’s Nextera DNA Flex Enrichment protocol. The first step of this workflow shows a strand of DNA, the First-strand, which forms a map from which cDNA, (complementary DNA) strands will be synthesised. The second step of this workflow shows that the complementary strands are now the maps for which another set of complementary DNA strands are synthesised. This is called Second strand DNA synthesis. The third step in the workflow is called Bead-linked tagmentation. Strands of DNA are shown with small grey extensions representing molecular tags. The fourth step is called Indexing PCR, and the small grey tags are now represented in red. These are labelled: Sequencing library. The fifth step shows Eppendorf tubes containing a small volume of liquid being added together - these are Pooled samples. The sixth step shows a magnet with a probe being used to match the red-tagged DNA fragments. The seventh step shows an Eppendorf tube which is labelled: Enriched Library. The eighth step is called QC and sequence and shows that the Enriched Library can be used in an Illumina sequencing machine. Diagram B is a workflow for the ARTIC protocol. The first step of this workflow shows a strand of DNA, the First-strand, which forms a map from which cDNA strands will be synthesised. The second step of this workflow is called Multiplex PCR (2 pools), Untailed Primers. Two Eppendorf tubes are shown being added together, and are labelled: Combine pools and QC. The third step is labelled Barcode addition / NGS library preparation. Strands of DNA are shown with small red extensions and are labelled: sequencing library. The fourth step is labelled: Normalise, QC and sequence, and shows that the sequencing library can be used in Oxford Nanopore and Illumina machines. Diagram C is a workflow for the Tailed Amplicon Method. The first step of this workflow shows a strand of DNA, the First-strand, which forms a map from which cDNA strands will be synthesised. The second step of this workflow is called Multiplex PCR (2 or 4 pools), Tailed primers. Four Eppendorf tubes with small volumes of liquid are shown being combined. The third step is called indexing PCR, and strands of DNA with small red extensions are shown. These are labelled: Sequencing Library. The fourth step is called Normalise, QC and sequence, and shows that the sequencing library can be used in Illumina machines.

Amplicon-based sequencing

Alt-text Figure 5 - Flowchart of a sequencing protocol from viral RNA isolation to sequencing analyses

A flow diagram showing the processes following RNA extraction and how many days these take. The first flow is ordered as follows: RNA extraction, RT-PCR, PCR Amplification, Library Preparation, Illumina Sequencing, and Sequence Analysis. These processes take three days. The second flow is RNA extraction, qPCR detection, and Report. These processes take two days.

Alt-text Figure 6 - The ARTIC sequencing protocol

The ARTIC sequencing protocol. The protocol is as follows: A single strand of RNA is completed with cDNA synthesis. This undergoes multiplex PCR with untailed primers. Pools are combined and quality controlled. Barcodes are added or NGS libraries prepared. These form a sequencing library. This library is normalised, quality controlled and sequenced in either Oxford Nanopore or Illumina machinery.

Alt-text Figure 7 - The Midnight protocol

This is a diagram of the Midnight protocol. It shows the steps and the time taken for each. The process begins with RT-PCR (265 minutes), and then pooling (5 minutes). A 1200bp amplicon is rapidly barcoded (15 minutes), and then these are subjected to sample pooling, SPRI clean, quantification, and rapid adapter addition (all taking 35 minutes). Finally, these are loaded into the sequencer (10 minutes).

Building a workflow on Galaxy

Alt-text Figure 8 - A screenshot of the steps used to create a new workflow on Galaxy

A screenshot from the Galaxy website depicts the following instructions: Hands-on: creating a new workflow. 1) Click on Workflow in the top panel of the Galaxy page. On the top right you will see 2 buttons: Create and Import. To create a new workflow click on Create. Enter a Name and Annotation for your workflow and click Save. The Workflow Editor will open with a new, empty workflow loaded. Below the information is the image of a blank (empty) New Workflow with a menu bar Tools on the left and on the right fields with Name, Version and Annotation to be completed.

Alt-text Figure 9 - The annotation naming of a galaxy workflow

A screenshot showing the window to create a New Workflow on Galaxy, where the following information is complete on respective fields: Name: flipping the burger workflow; Annotation: an example workflow for demonstrating data flipping. Below, the buttons Create and Cancel.

Alt-text Figure 10 - The blank canvas of a galaxy workflow

A screenshot showing a blank canvas on Galaxy entitled: flipping the burger workflow.

Alt-text Figure 11 - Adding the input dataset tool

A screenshot showing a box in the middle of the canvas with the following information: load burger data > output (input).

Alt-text Figure 12 - Adding the reverse file tool and link to the input dataset

A screenshot of the Galaxy canvas showing a second box connected to the box described in Figure 11. The second box contains the information: flip burger > outfile (input).

Alt-text Figure 13 - The file upload menu

A screenshot of a Galaxy window entitled: Workflow: flipping the burger workflow. The Load Burguer Data field contains the information: 5: tac on data 1. There is an Upload (veritical arrow) button on the right-hand side.

Alt-text Figure 14 - The file upload menu on the paste input section

A screenshot of the Drag-and-Drop window. Below are buttons to search for files on the computer: Choose local files; Choose remote files; Paste/Fetch data, followed by the buttons Start; Pause; Reset; Cancel.

Alt-text Figure 15 - The run workflow window after the upload set

A screenshot showing the uploaded file and its specifications. Name: burger.txt. Size: 39 b. Type: txt. Genome: unspecified. Content: Bread, Onions, Cheese.

Alt-text Figure 16 - The workflow complete screen

A screenshot of a green box with the information: Executed tac and successfully added 1 job to the queue. The tool uses this input: 6: burger.txt. It produces this output: 10: tac on data 6.

Garbage in/Garbage out: importance of generating good data and cleaning up

Alt-text Figure 1 - Building a model without accounting for missing values

Graph of annual income versus education showing a weak positive correlation.

Alt-text Figure 2 - Building a model removing missing values

Graph of annual income versus education showing a strong positive correlation.

Data cleaning and quality control

Alt-text Figure 3 - Two sample graphs from a FastQC output

A screenshot of two illustrative graphs of a FastQC output showing good quality data. Details in the main text.

Alt-text Figure 4 - The average quality of the aligned reads

Screenshot of an illustrative graph showing a QualiMap BAMQC output indicating a quality score of ~ 60. Details in the main text.

FastQC and MultiQC tools

Alt-text Figure 5 - The top of a typical MultiQC report

An illustrative screenshot of MultiQC report. It shows general statistics of the analysed data such as %Assigned, M Assigned, % Aligned, M Aligned, % BP Trimmed, % Dups, % CG. Details in the main text.

Assess quality with FastQC

Alt-text Figure 6 - An example of a plot generated from FastQC analysis showing the Per Base Sequence Quality across a sample

A screenshot of a quality graph from FastQC. It shows the quality scores across all databases from sequencing. Details in the main text.

Assembly of SARS-CoV-2 genome and sequence alignment

Alt-text Figure 7 - Two types of reads assembly methods

Schematic illustration describing two different methods for genome assembly. Resequencing aligns reads to a reference genome and identifies variants. *De novo* assembly constructs a genome sequence from overlaps between reads. Details in the main text.

Alt-text Figure 8 - Identification of a point mutation

Schematic illustration showing a reference genome and a query read being compared and highlighting a point mutation T>C. Details in the main text.

Assembly tutorial

Alt-text Figure 9 - Galaxy homepage

Screenshot of Galaxy’s homepage. A red arrow points to “Login or Register”. Details in the main text.

Alt-text Figure 10 - Rename history
Screenshot of Galaxy’s homepage. A red arrow points to “Rename history”. Details in the main text.

Alt-text Figure 11 - Upload data
Screenshot of Galaxy’s homepage. A red arrow points to “Upload Data”. Details in the main text.

Alt-text Figure 12 - Upload window
Screenshot of Galaxy upload window. A red arrow points to “Paste/Fetch data”. Details in the main text.

Alt-text Figure 13 - Start upload
Screenshot of Galaxy upload window with a list of files in it. A red arrow points to “Start”. Details in the main text.

Alt-text Figure 14 - Uploaded files
Screenshot of Galaxy upload window. The window is green to indicate it is ready. A red arrow points to “Close”. Details in the main text.

Alt-text Figure 15 - Processing files
Screenshot of Galaxy homepage. A red arrow points to the uploaded data highlighted in yellow in the “History” menu. Details in the main text.

Alt-text Figure 16 - Files processed
Screenshot of Galaxy homepage. A red arrow points to the uploaded data highlighted in green in the “History” menu. Details in the main text.

Alt-text Figure 17 - Enabling file selection
Screenshot of Galaxy homepage. A red arrow points to the “operations on multiple datasets” tick-box in the “History” menu. Details in the main text.

Alt-text Figure 18 - Selecting files
Screenshot of Galaxy homepage. A red arrow points to the tick-boxes in front of the datasets in the “History” menu. Details in the main text.

Alt-text Figure 19 - Building list of dataset pairs
Screenshot of Galaxy homepage. One red arrow points to the “For all selected” in the “History” menu and another red arrow points to “Build List of Dataset Pairs”. Details in the main text.

Alt-text Figure 20 - Creating a collection
Screenshot of Galaxy “Create a collection of paired datasets” window. One red arrow points to a field containing “_1”, and a second arrow points to a field containing “_2”. At the bottom of the window, a red arrow points to “Name” and another red arrow points to “Create collection”. Details in the main text.

Alt-text Figure 21 - Disabling operation in multiple datasets
Screenshot of Galaxy homepage. A red arrow points to the “operations on multiple datasets” tick-box in the “History” menu. Details in the main text.

Alt-text Figure 22 - Choosing alignment tool
Screenshot of Galaxy “Edit dataset attributes” window. A red arrow points to the search field in the “Tools” menu. The field contains the query “bwa-mem”. Another red arrow points to “Map with BWA” in the “Tools” menu. Details in the main text.

Alt-text Figure 23 - Setting up the aligment
Screenshot of Galaxy analyses window. A red arrow points to the field “Will you select a reference genome from your history” with the information “Use a genome from history” selected in the field. A second red arrow points to the field “Single or pair-end reads” with the information “Paired collection” selected in the field. A third red arrow points to the field “Select analysis mode” with the information “1_Simple Illumina mode” selected in the field. Details in the main text.

Alt-text Figure 24 - How to view the results
Screenshot of Galaxy’s output window. A red arrow points to “View data” in the “History” menu. Details in the main text.

Alt-text Figure 25 - Snippet of the BAM file - part 1
Screenshot of a BAM file output. It has the columns QNAME, FLAG, RNAME, POS, MAPQ, CIGAR, MRNM and MPOS. Details in the text.

Alt-text Figure 26 - Snippet of the BAM file - part 2
Screenshot of a BAM file output. It has the columns ISIZE and SEQ. Details in the text.