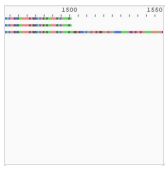


May 21, 2021

Protocol for variant calling in SARS-Cov-2 enabling long indel detection

Juanjo Bermúdez[1]

[1]Contignant Technologies SL

dx.doi.org/10.17504/protocols.io.btrunm6w

Coronavirus Method Development Community Contignant

Juanjo Bermúdez

ABSTRACT

Protocol for SARS-Cov-2 variant calling departing from raw reads in a run. First, we will de-novo assemble the genome using s-aligner, then we will use standard software for variant calling: BWA, samtools, and freebayes. The advantage of this protocol is that it also finds long indels if they exist. Other protocols based on reference mapping will often miss these.

DOI

dx.doi.org/10.17504/protocols.io.btrunm6w

PROTOCOL CITATION

Juanjo Bermúdez 2021. Protocol for variant calling in SARS-Cov-2 enabling long indel detection. **protocols.io** https://dx.doi.org/10.17504/protocols.io.btrunm6w

CREATED

Mar 29, 2021

LAST MODIFIED

May 21, 2021

PROTOCOL INTEGER ID

48660

1 Assemble the genome

Protocol for assembling SARS-Cov-2 runs with s-aligner
by Juanjo Bermúdez

PREVIEW RUN

1.1

Index the reads in the run. You can use as input a FASTA file (preferred) or a FASTQ. The file can also be compressed with gzip, having a .gz extension. The script will uncompress the file and pass it to FASTA format if it's a FASTQ file.

./index.sh your_run_id /mnt/c/your_run.fastq.gz

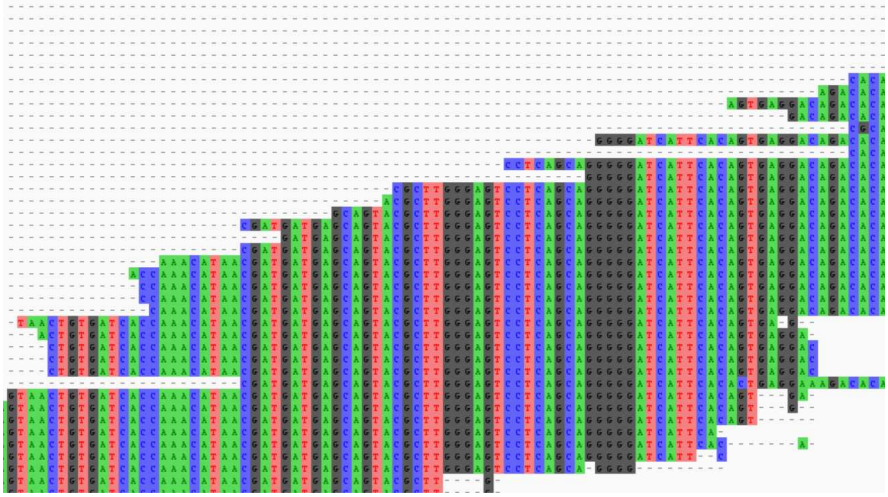
1.2

Assemble

./assemble.sh your-run-id -output alignments 500 > results/your-run-id-500.fa

Assemble up to processing 500 reads

results/your-*run*-id-500.fa should have the largest contig bigger than 1.200bp and at least one of the files containing aligned reads to form a contig should look like that:



Step 1.2 includes a Step case.

<1.200bp

>1200bp

step case

<1.200bp

The largest contig in your-*run*-id-500.fa is shorter than 1200bp

1.3

Citation: Juanjo BermÃÂºdez (05/21/2021). Protocol for variant calling in SARS-Cov-2 enabling long indel detection.
https://dx.doi.org/10.17504/protocols.io.btrunm6w

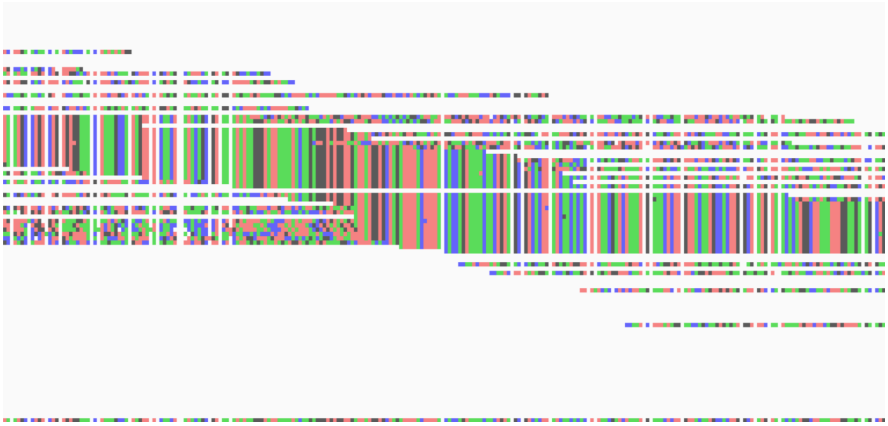
Map your reads to a reference genome to see what's going wrong.

./map.sh your-run-id sequences/your-run-id/sra_data.part-71.fa reference.fa 4000 > results/map-your-run-id.fa

Step 1.3 includes a Step case.
Data not trimmed
Too many chimeras
No overlapping
Runs are short
Contamination

step case

Data not trimmed



If the data is not trimmed and has adaptors at the extremes of the runs it will look like the image above.

1.4 Trim your data and start again.

2 Call variants.sh to find variants. The script will call BWA, samtools, varscan and freebayes to generate a VCF file with the resulting variants.

./scripts/variants.sh RUN-ID ./results/RUN-ID-xxxx.fa PATH_TO_SARS_COV_2_REFERENCE_FASTA

Citation: Juanjo BermÃÂºdez (05/21/2021). Protocol for variant calling in SARS-Cov-2 enabling long indel detection.
https://dx.doi.org/10.17504/protocols.io.btrunm6w