 doi:10.20944/preprints202005.0343.v1

Community Research Amid COVID-19 Pandemic: Genomics Analysis of SARS-CoV-2 over Public GALAXY Server

Ambarish Kumar [1,*], Bjoern Gruening [2] and Ali Haider Bangash [3]

[1] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India;

[2] University of Freiburg, Germany

[3] Shifa College of Medicine, STMU, Pakistan

Abstract

Citizen Science has come up to perform analytics over the SARS-CoV-2 genome. Public GALAXY servers provide an automated platform for genomics analysis. Study includes design of GALAXY workflows for RNASEQ assembly and annotation as well as genomic variant discovery and perform analysis across four samples of SARS-CoV-2 infected humans obtained from the local population of Wuhan, China. It provides information about transcriptomics and genomic variants across the SARS-CoV-2 genome. Study can be extended to perform evolutionary and comparative study across each species of coronaviruses. Augmented and integrated study with cheminformatics and immunoinformatics will be a way forward for drug discovery and vaccine development.

Keywords: GALAXY, SARS-CoV-2, Assembly, Annotation, Genomic Variant Discovery

Introduction

SARS-CoV-2 outbreak has brought the community researchers all together. Publicly accessible computational platforms like public GALAXY servers are providing web-accessible resources for community research. Available tools in the GALAXY tools shed enable users to perform integrated analysis of SARS-CoV-2 genome - Genomics, Cheminformatics, Immunoinformatics, Genomic Variant Discovery, Phylogenomics etc. Present study is an analysis of the SARS-CoV-2 genome - assembly, annotation and genomic variants discovery. It will provide an insight into the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2020 doi:10.20944/preprints202005.0343.v1

SARS-CoV-2 genome and will help to understand the functioning and adaptation of SARS-CoV-2 virus inside its host. Further, impact of the SARS-CoV-2 virus is very specific to the population. An automated and reproducible pipeline for integrative study of Genomics, Cheminformatics and immunoinformatics of SARS-CoV-2 infection will pave the path for population specific vaccine development. GALAXY can be developed as a dedicated workbench for all possible analysis of viruses belonging to the coronaviruses family.

Materials and Methods

- Fasta format reference genome sequence of SARS-CoV-2 and its gff3 format annotation - NCBI Reference Sequence: NC_045512.2.

- RNASEQ Data - Paired-end illumina RNASEQ reads obtained from SARS-CoV-2 infected Wuhan, China population. NCBI SRA accessions are - SRR10903401, SRR10903402, SRR11092064 and SRR11092057. None of the samples has biological or technical replicates.

- Computational platform - Public GALAXY/Europe server - https://usegalaxy.eu/

There are three separate workflows for assembly and annotation of SARS-CoV-2 genome. Shared and published GALAXY workflows for assembly and annotation are as follows.

Workflow	Shared and published GALAXY workflow
Assembly using TOPHAT2 and annotation (alternate)	https://usegalaxy.eu/u/ambarishk/w/covid-19-assembly-using-tophat2-and-annotation-alternate
Assembly using Unicycler and annotation.	https://usegalaxy.eu/u/ambarishk/w/covid-19-unicycler-assembly-and-annotation
Assembly using StringTie and annotation	https://usegalaxy.eu/u/ambarishk/w/covid-19-stringtie-assembly-and-annotation
Assembly using TOPHAT2 and annotation.	https://usegalaxy.eu/u/ambarishk/w/covid-19-assembly-using-tophat-and-annotation

Table-01: Shared and published GALAXY workflows for assembly and annotation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2020 doi:10.20944/preprints202005.0343.v1

There is an alternate workflow for assembly and annotation using TOPHAT2 which has been used into the present study. It does not has data pre-processing steps to remove host reads. Sample may not has sufficient number of reads of relevant virus due to which TOPHAT2 run may fail to produce splice junction or splice junction index.

Pre-processing of NGS data

Steps to pre-process RNASEQ illumina reads are common in all GALAXY workflows - assembly and annotation as well as genomic variant calling.

Included steps and GALAXY tools for data preprocessing are as follows.

1. Input dataset - uploaded text file containing list of NCBI SRA accession for each sample.
2. Faster Download and Extract Reads in FASTQ - download and extract fastq reads from NCBI SRA.
3. Fastp - quality score correction and adapter trimming.
4. MultiQC - Generate reports for fastp.
5. BWA-MEM - map host reads to human genome.
6. Filter SAM or BAM, output SAM or BAM - filter out unmapped reads to host genome (human).
7. MergeSamFiles - merge aligned SAM file of each sample.
8. Samtools fastx - separate out paired-end illumina reads of each sample.

Alignment

There are two aligners used in assembly and annotation workflows.

1. TOPHAT2 - aligner used in GALAXY workflow for assembly using TOPHAT2 and annotation.
2. HISAT2 - aligner used in GALAXY workflow for assembly using StringTie and annotation.

Assembly

Used assembler for each assembly and annotation workflows are as follows.

1. Unicycler - assembler used in GALAXY workflow for assembly using Unicycler and annotation.
2. Cufflink - assembler used in GALAXY workflow for assembly using TOPHAT2 and annotation.
3. StringTie - assembler used in GALAXY workflow for assembly using StringTie and annotation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2020 doi:10.20944/preprints202005.0343.v1

Abundance estimation

Abundance estimation of expressed transcripts and genes are produced by Cufflink and StringTie in their respective workflows for assembly and annotation.

1. Cufflink produces transcript expression and gene expression.
2. StringTie produces coverage, gene count and transcript count.

Annotation

All steps for annotation are common in all three GALAXY workflows for assembly and annotation.

1. Transdecoder - annotates coding DNA sequences and coded proteins.
2. Glimmer3 - Knowledge based gene prediction. It predicts open reading frames in the SARS-CoV-2 genome. Glimmer ICM builder servers as a trained model for running Glimmer3.
3. Jackhmmer - Per domain and per sequence hits of SARS-CoV-2 proteins in human proteomes.
4. Antismash - genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. It runs genome-wide pfam analysis.
5. BLASTP - coded proteins are matched against the known SARS-CoV-2 proteins as BLAST hits.
6. Reciprocal BLAST hits - orthologs of SARS-CoV-2 proteins to human proteomes and vice-versa.

Genomic variant discovery

For genomic variant calling there are two separate GALAXY workflows. Shared and published GALAXY workflows for genomic variant calling are as follows.

Workflow	Shared and published GALAXY workflow
Genomic variant calling using GATK4	https://usegalaxy.eu/u/ambarishk/w/gatk4
Genomic variant calling using VARSCAN	https://usegalaxy.eu/u/ambarishk/w/varscan

Table-02: Shared and published GALAXY workflows for genomic variant calling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2020 doi:10.20944/preprints202005.0343.v1

Data pre-processing in each of the workflows for genomic variant calling is the same as mentioned in the above section. Apart from data pre-processing, included GALAXY tools and steps of genomic variant discovery in respective the workflow for GATK4 and VARSCAN are as follows.

1. Genomic variant discovery using GATK4

1.1 Bowtie2 - alignment of RNASEQ reads with SARS-CoV-2 genome.
1.2 AddOrReplaceReadGroups - add or replace sequencing information of the read group.
1.3 SortSam - sort SAM file as per genomic coordinates.
1.4 MarkDuplicates - locate duplicate aligned reads.
1.5 GATK4 Mutect2 - call somatic mutations - SNPs and INDELs.
1.6 VcfAllelicPrimitives - split allelic primitives (gaps or mismatches) into multiple VCF lines.
1.7 SnpEff eff - annotate variants.
1.8 SnpSift Extract Fields - convert VCF into tabular file.
1.8 Concatenate datasets - tail-to-head concatenation of genomic variants across all samples.
1.9 Unique - find unique occurrences of each variant record.

2. Genomic variant discovery using VARSCAN

2.1 Bowtie2 - alignment of RNASEQ reads with SARS-CoV-2 genome.
2.2 Samtools sort - sort SAM format aligned files.
2.3 samtools mpileup - generate mpileup file.
2.4 VarScan mpileup - call SNPs and INDELs from mpileup file.
2.5 VcfAllelicPrimitives - split allelic primitives (gaps or mismatches) into multiple VCF lines.
2.6 SnpEff eff - annotate variants.
2.7 SnpSift Extract Fields - convert VCF into tabular file.
2.8 Concatenate datasets - tail-to-head concatenation of tabular files.

Results

Output of each step of the workflows can be obtained from shared and published GALAXY workflows.

Shared GALAXY history of each workflow for assembly and annotation are as follows.

History name	Shared and published GALAXY history
Assembly using TOPHAT2 and	https://usegalaxy.eu/u/ambarishk/h/covid-19-tophat2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2020 doi:10.20944/preprints202005.0343.v1

annotation.(alternate)	
Assembly using Unicycler and annotation.	https://usegalaxy.eu/u/ambarishk/h/covid-19-unicycler
Assembly using StringTie and annotation	https://usegalaxy.eu/u/ambarishk/h/covid-19-stringtie

Table - 03: Shared GALAXY history of each workflow for assembly and annotation

Shared GALAXY history of each workflow for genomic variant calling are as follows.

History name	Shared and published GALAXY history
Genomic variant discovery using GATK4	https://usegalaxy.eu/u/ambarishk/h/covid-19-gatk
Genomic variant discovery using VARSCAN	https://usegalaxy.eu/u/ambarishk/h/covid-19-varscan

Table- 04: Shared GALAXY history of each workflow for genomic variant calling

Conclusions

Alignment statistics obtained from TOPHAT2 shows the metagenomic nature of the data. A clear visualisation of intermediate results of SARS-CoV-2 genomics analysis can be obtained from the GALAXY analysis run steps. Literature search about ortholog hits between SARS-CoV-2 proteins and human proteomes will support the findings and reveal more about adaptation of SARS-CoV-2 virus into humans as a host.

Future work

Study design will help to perform analysis of SARS-CoV-2 infected human samples across various geolocations. Also, It will help to study and develop a curated repository of analysis of viruses belonging to the coronaviruses family. Integrated cheminformatics and immunoinformatics analytics will be a way forward towards drug discovery and vaccine development.

Acknowledgements

GALAXY Europe is acknowledged for providing computational infrastructure and support while implementing the GALAXY tools. This work is carried out during virtual biohackathon covid-19-bh20 - https://github.com/virtual-biohackathons/covid-19-bh20. Nevertheless delicious food at my home and my kid is acknowledged for keeping me entertained and revived during COVID-19 lockdown.

References

[1] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update**, *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W537–W544, doi:10.1093/nar/gky379

[2] **antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline**
Kai Blin, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H Medema, & Tilmann Weber. *Nucleic Acids Research* (2019) doi: 10.1093/nar/gkz310.