Comprehensive Analysis and Annotation of the SARS-CoV-2 Genome in Vero E6 Cells

Prepared by: Simranjit Kaur Kang, Supervised by: Professor Dr. Sarath Janga
Course: FA23: GENOMIC DATA ANAL & PRECIS MED
Class ID: 35088
For script and raw data please refer GitHub : **https://github.com/SKVirk27/SARS-CoV-2-Genome-Assembly-and-Annotation.**

Abstract

This comprehensive research explores the genomic analysis of SARS-CoV-2 in the Vero E6 cell line, derived from African green monkey kidney cells. The study meticulously conducts genome assembly and annotation, focusing on essential viral proteins. A significant achievement is the development of a bioinformatics pipeline, adaptable for both specific genomic and broader metagenomic analyses. This innovative approach provides a powerful tool for future research in genomic and metagenomic studies. The findings significantly enhance our understanding of SARS-CoV-2's behavior in cell culture environments, offering vital insights for the development of diagnostics and vaccines. This project, by integrating advanced genomic techniques and bioinformatics tools, marks a notable advancement in virology and genomic research fields.

Introduction

The advent of the COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has underscored the urgency of genomic research in infectious diseases. Originating in Wuhan, China in late 2019, the virus has rapidly spread globally, demanding swift scientific action (University of California - San Diego, 2021) . This study delves into the genomic structure of SARS-CoV-2, using the Vero E6 cell line to unpack its complex genomic features. By leveraging advanced bioinformatics tools for genome assembly and annotation, we aim to shed light on viral proteins critical to the virus's lifecycle, thereby contributing to the ongoing efforts to curb the pandemic. Our innovative bioinformatics pipeline, capable of detailed genomic and metagenomic analyses, stands as a testament to the progress in virology research, offering new perspectives for diagnostics and vaccine development in the face of this unprecedented health challenge.

Methodology Overview

Our genomic exploration of SARS-CoV-2, leveraging next-generation sequencing, commenced with a rigorous FASTQC quality assessment, safeguarding data fidelity. The SPAdes toolkit orchestrated precise genome assembly, augmented by Bandage for intuitive genomic visualization. Alignments, crafted with Bowtie2 precision, paved the way for variant discovery via Samtools and BCFtools. SnpEff was paramount in annotating these variants, offering functional insights, and predicting phenotypic consequences, thereby enriching the tapestry of our viral genetic understanding. This methodological amalgamation fortified the foundation of our genomic discoveries.

Dataset Analysis:

Employing Illumina MiniSeq's paired-end sequencing, we explored into datasets SRR11528306 and SRR11528307, each a genomic compendium of over 177M and 186.5M bases. The datasets, reflective of the viral genome with an average read length of 131 bp and a GC content nearing 38%, were pivotal for elucidating viral replication and mutation dynamics within Vero E6 cells. Published on April 15, 2020, their analysis has been indispensable in tracing the evolutionary trajectory of the virus, a cornerstone for public health intervention strategies.

Initial Data Processing and Quality Control:

The fidelity of our sequencing data underwent stringent scrutiny with FASTQC, Andrews (2010) describes FastQC as a quality control tool for high throughput sequence data . Establishing a quality baseline critical for trustworthy assembly.

Genome Assembly:

Employing SPAdes, According to Prjibelski et al. (n.d.), the SPAdes tool is used for genome assembly. We navigated the complexities of metagenomic assembly, stitching high-quality reads into contiguous genomic sequences.

Comparative Analysis:

Prior to alignment, a comparative genomic analysis was conducted using MUMmer, aligning our assembled contigs with reference genomes. This crucial step allowed us to discern structural variations and highlight genomic rearrangements that may play pivotal roles in the virus's evolution and pathogenicity.

Visualization and Annotation:

Post-assembly, the Bandage tool provided a visual assessment of the genomic structure , (Wick et al., 2015) pinpointing regions for targeted analysis. Subsequent automated annotation of genes within the contigs was carried out by PROKKA (Seemann, 2014), generating a detailed genomic annotation necessary for understanding viral function. Buels et al. (2016) provide JBrowse as a web platform for genome analysis. We visualized our gene annotation and location in JW Browser to get better understanding.

Alignment and Variant Calling:

Post-comparative analysis, the accurate alignment of our sequences to a reference genome was executed with Bowtie2 (Langmead, 2010). This foundational step paved the way for the critical phase of variant detection and calling through Samtools and BCFtools, uncovering mutations that hold the key to the virus's evolutionary dynamics (Li, Handsaker, Danecek, & the samtools team, n.d.). For in-depth functional annotation and impact assessment of the identified variants, SnpEff was utilized, Cingolani et al. (2012) discuss how SnpEff annotates and predicts the effects of single nucleotide polymorphisms. Providing a comprehensive annotation that encompasses both predicted phenotypic impacts and potential changes to protein function.

Database Construction for SnpEff Annotation:

To facilitate detailed annotation of SARS-CoV-2 variants, we constructed a specialized SnpEff database. This was achieved by downloading the nucleotide, protein sequences, CDS FASTA files, and corresponding GFF3 annotation files from the NCBI database, using the RefSeq accession number NC_045512. The integration of these files into SnpEff allowed for an enriched analysis, linking genomic variations to potential functional changes in the viral genome — a crucial step for interpreting the impact of mutations observed during our study.

Each step in this methodology is designed to build upon the previous, ensuring a robust and comprehensive analysis of the SARS-CoV-2 genome. This structured approach underlines the reliability of our results and the potential impact of our findings on the broader scientific community's understanding of this novel coronavirus.

Results

Quality Assessment and Sequencing Data:

The sequencing data quality assessment for SRR11528306 and SRR11528307, as conducted with FASTQC, confirmed high-quality metrics, with a sequence length range of 2-151 bp and an approximate GC content of 38%. The quality metrics underscored most high-quality sequence scores, indicating the reliability of the sequencing data for further genomic analysis.

1a

1b

Figure 1a, b : Quality Assessment Visualization - This image validates the sequencing data quality, with high-quality zones indicative of the dataset's robustness for genome assembly.

Genomic Assembly and Visualization of SARS-CoV-2:

Our meticulous genomic assembly, detailed in Figure 2, has culminated in the identification of significant contigs indicative of a comprehensive representation of the SARS-CoV-2 genome. Specifically, the SPAdes assembly output for sample SRR11528307 yielded a largest contig of 29,555 base pairs with an extraordinary coverage depth of 2693.87X, and for sample SRR11528306, a contig of 29,849 base pairs with a coverage of 2531.28X was observed. Such depth and breadth of coverage are instrumental in ensuring a high-quality assembly, reducing the potential for errors. Complementing this, the Bandage tool provided a visual confirmation of the contigs' integrity, emphasizing the near-complete nature of the viral genomic segments crucial for further investigative analysis (refer to Figure 2)

2a 2b

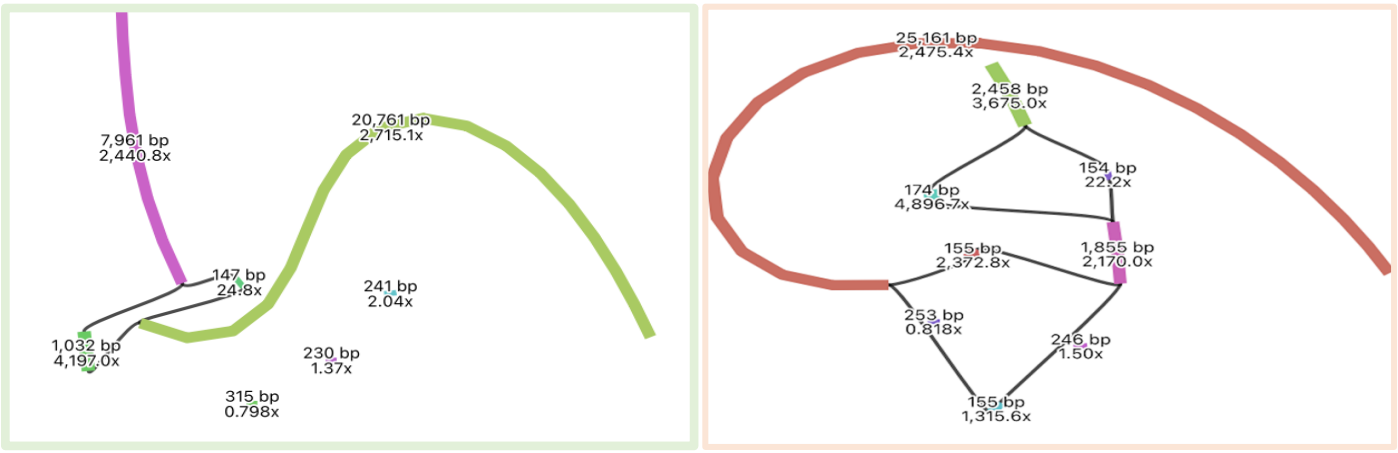

Figure 2: Bandage Assembly Map - This figure illustrates the contigs resulting from genome assembly, with coverage depth highlighting the confidence in our genomic reconstruction.

MUMmer Alignment Results:

Our comparative genomic analysis using MUMmer showed a high degree of sequence conservation between the assembled contigs and the reference genome, highlighting the genomic stability of SARS-CoV-2.

SRR ID	Start	End	Query Start	Query End	Alignment Length	% Identity	Reference Length	Query Length	Coverage R	Coverage Q
SRR11528307	149	29703	29555	1	29555	99.99	29903	29555	98.84	100
SRR11528306	253	25490	25238	1	25238	99.98	29903	25239	84.4	100

Table 1. The MUMmer alignment analysis presented above demonstrates a remarkable sequence identity of approximately 99.98% between our SARS-CoV-2 genome assemblies and the reference sequence. This high level of genomic fidelity underscores the conservation of key viral regions and the accuracy of our assembly process. The table also shows the extensive coverage, indicating the depth and comprehensiveness of our sequencing efforts. This data is pivotal for future studies on viral evolution and potential therapeutic targets.

Prokka Gene Annotation Insights:

Annotation with Prokka identified key proteins crucial for viral replication and highlighted several hypothetical proteins, offering potential targets for further investigation into viral pathogenesis and treatment strategies.

Locus_tag	Location	Gene	Product
SRR11528306_prokka_00001	1260	N	Nucleoprotein
SRR11528306_prokka_00002	366		hypothetical protein
SRR11528306_prokka_00003	366	7a	Protein 7a
SRR11528306_prokka_00004	669	M	Membrane protein
SRR11528306_prokka_00005	456	3a	Protein 3a
SRR11528306_prokka_00006	3849	S	Spike glycoprotein
SRR11528306_prokka_00007	7788	rep	Replicase polyprotein 1ab
SRR11528306_prokka_00008	13218	1a	Replicase polyprotein 1a
SRR11528307_prokka_00001	1260	N	Nucleoprotein
SRR11528307_prokka_00002	366		hypothetical protein
SRR11528307_prokka_00003	366	7a	Protein 7a
SRR11528307_prokka_00004	186		hypothetical protein
SRR11528307_prokka_00005	669	M	Membrane protein
SRR11528307_prokka_00006	828	3a	Protein 3a
SRR11528307_prokka_00007	3822	S	Spike glycoprotein
SRR11528307_prokka_00008	7788	rep	Replicase polyprotein 1ab
SRR11528307_prokka_00009	13218	1a	Replicase polyprotein 1a

Table 2: The table presents a summary of key genes and their associated proteins identified in the SARS-CoV-2 genome from samples SRR11528306 and SRR11528307, including both well-characterized and hypothetical proteins, highlighting areas for further functional research.

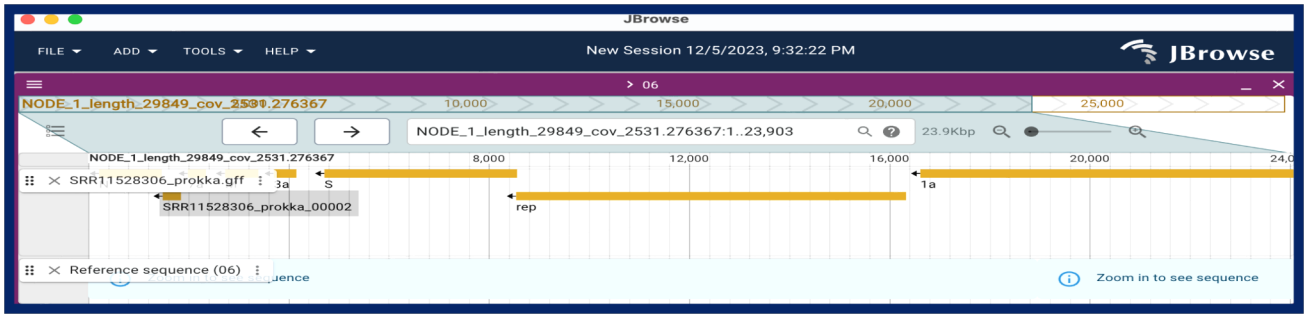

Figure 3: JBrowse Visualization of Prokka Annotations. This interactive display allows for an in-depth examination of the genomic annotations provided by Prokka. It showcases the locations and predicts functions of essential viral proteins, as well as the placement of hypothetical proteins, offering a map for future research into the virus's life cycle and potential therapeutic targets.

Sequence Alignment and Variant Calling:

Subsequent alignment with Bowtie2 and variant calling with Samtools and BCFtools demonstrated a near-perfect match with the reference genome, underscoring the accuracy of our assembly. Filtered VCF files, annotated with SnpEff, revealed mutations that could significantly impact viral transmission and pathogenicity.

Sample_ID	Gene Name	Effect	Impact	Allele	Position	Ref	Alt	Amino Acid Change
SRR11528307	ORF1ab	missense	MODERATE	G	2832	A	G	Lys856Arg
SRR11528307	ORF1ab	missense	MODERATE	T	11083	G	T	Leu3606Phe
SRR11528307	N	synonymous	LOW	C	28688	T	C	Leu139Leu
SRR11528307	S	downstream	MODIFIER	T	29716	A	T	*4332
SRR11528306	ORF1ab	missense_variant	MODERATE	T	884	C	T	Arg207Cys
SRR11528306	ORF1ab	missense_variant	MODERATE	A	1397	G	A	Val198Ile
SRR11528306	ORF1ab	synonymous_variant	LOW	T	3040	C	T	Tyr107Tyr
SRR11528306	ORF1ab	missense_variant	MODERATE	T	8653	G	T	Met2796Ile
SRR11528306	ORF1ab	missense_variant	MODERATE	T	11083	G	T	Leu37Phe
SRR11528306	ORF3a	synonymous_variant	LOW	T	25413	C	T	Ile7Ile
SRR11528306	N	synonymous_variant	LOW	C	28688	T	C	Leu139Leu

Table 3: It presents the mutational profile of the SARS-CoV-2 genome as derived from the samples SRR11528306 and SRR11528307. Highlighted are the missense mutations with a moderate impact on ORF1ab, which may influence the virus's replication efficiency and protein functionality. The synonymous mutations in the N gene and the ORF3a gene, while not altering protein sequences, could potentially affect the virus's evolutionary dynamics and interactions with host cell mechanisms. The downstream modifier in the S gene points to changes that could impact viral propagation and immune response evasion.

Discussion

This research has elucidated key aspects of the SARS-CoV-2 genome through SPAdes assembly, identifying the largest contigs which are instrumental for understanding the virus's structure. The notable mutations such as Lys856Arg and Leu3606Phe in the ORF1ab gene, revealed through SnpEff annotation, suggest potential modifications in protein function, which may impact viral replication and pathogenicity. The synonymous mutations and downstream modifiers highlight the virus's adaptability and potential immune evasion mechanisms. Future efforts will aim to explore the roles of hypothetical proteins and their interactions within the host, which are vital for innovative therapeutic approaches. Our findings lay the groundwork for detailed functional genomic studies, crucial for ongoing COVID-19 research. The goal is to extend this work to include a broader range of isolates, enhancing our understanding of SARS-CoV-2's diversity and informing public health strategies. This research underpins the global initiative to develop precise interventions against this pandemic, with the aspiration of informing vaccine design and antiviral drug development.

References

1. University of California - San Diego. (2021, March 18). Novel coronavirus circulated undetected months before first COVID-19 cases in Wuhan, China. UC San Diego Health. https://health.ucsd.edu/news/press-releases/2021-03-18-novel-coronavirus-circulated-undetected-months-before-first-covid-19-cases-in-wuhan-china/.
2. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (n.d.). Using SPAdes De Novo Assembler. PMID: 32559359. https://doi.org/10.1002/cpbi.102/
4. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153/
5. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6(2), 80–92. https://doi.org/10.4161/fly.19695
6. Buels, R., Yao, E., Diesh, C. M., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biology, 17, 66. https://doi.org/10.1186/s13059-016-0924-1
7. Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. Bioinformatics, 31(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383
8. Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Current Protocols in Bioinformatics, Chapter 11, Unit 11.7. https://doi.org/10.1002/0471250953.bi1107s32
9. Li, H., Handsaker, B., Danecek, P., & the samtools team. (n.d.). bcftools. Retrieved from https://samtools.github.io/bcftools/bcftools.html