Comparison of SARS-CoV-2 variants with INSaFLU and galaxyproject

Nazli Sila Kara¹, Meltem Kutnu², Yasemin Utkueri³, Funda Yilmaz⁴, Elif Bozlak^{5,6}, and Evrim Fer⁷

1 Department of Medical Biochemistry, Faculty of Medicine, Istinye University, Istanbul, Turkey **2** Department of Biological Sciences, Middle East Technical University, Ankara, Turkey **3** Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey **4** Radboud University, Donders Centre for Cognition, Nijmegen, 6525 HR, The Netherlands **5** Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, Vienna, 1210, Austria **6** Vienna Graduate School of Population Genetics, Vienna, Austria **7** Department of Genetics, University of Arizona, Tucson, USA

BioHackathon series:
COVID-19 Biohackathon
Virtual conference 2020

Submitted: 11 May 2020

License
Authors retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Published by BioHackrXiv.org

Introduction

Covid-19 Virtual BioHackathon 2020 hosts many ongoing projects on workflows about either for the development of the existing ones or combining two or more of them for better functioning. Some of these workflows are Galaxy Project, INSaFLU and nf-core. In these pipelines, the inputs are genome sequence reads mostly coming from Illumina but also Oxford Nanopore, which are analyzed through these workflows by branching whether it is single end or paired end, as output, the annotated variant lists including single nucleotide polymorphisms (SNPs) and small insertion deletion (indel) are provided. We compared published variants by Galaxy project with produced variants by INSaFLU workflow in the course of our study.

- galaxyproject/SARS-CoV-2

Galaxy, a scientific community and open-source platform, provides workflows, datasets and histories enabling accessibility, reproducibility, and transparency in computational biology, mostly for genomic research [@nekrutenko2020no]. In Galaxy Projects, there are many workflows including genomic sequence assembly, variant calling. In this study, we used galaxyproject/SARS-CoV-2 workflow proposed by Galaxy community and this workflow will be named as "Galaxy workflow" in the latter part of this report. Steps of Galaxy workflow for variant analysis include mapping with BWA (for paired end data) or bowtie2 (for single end data), filtering steps, variant calling with lofreq and annotation with SnpEff, to choose a variant calling algorithm for the workflow, they compared FreeBayes, Mutec2 and lofreq, and based on the results, they used lofreq in the workflow.

- INSaFLU

INSaFLU has been originally developed as a web-based platform for influenza-based bioinformatics analysis and authors suggest that it can be also used for rapid assessment of variants for the novel coronavirus SARS-CoV-2 [@borges2018insaflu]. In order to determine variant positions, it uses bwa for mapping, samtools and FreeBayes algorithms for variant calling and SnpEff for variant annotation. The INSaFLU interface provides many features which are gene and whole-genome consensus sequences, phylogenetic trees, as well as gene, protein and genome alignments. Besides, it shares annotated variant files.

Taking into account the impact of the differences in analysis steps of different workflows (Baichoo et al., 2018), our project for Covid-19 Virtual BioHackathon 2020 aimed to compare variants determined by the INSaFLU workflow for the same samples that have been analyzed with the Galaxy workflow between 20.03.2020 and 02.04.2020. For this purpose, the Illumina

reads were retrieved from the SRA database, then all samples were uploaded to the INSaFLU interface and workflow was run with updated reference genome NC045512.2. The resulting variants were used to detect the shared and unique variants in two different workflows. In the final step, some general properties of the variants were analyzed.

Results

Variant Detection by INSaFLU and Comparison with Galaxy

In the first place, we wanted to have a variant list for the INSaFLU workflow to compare with published variant files from Galaxy. To run the INSaFLU workflow, we uploaded fastq files obtained from SRA by SRA toolkit. Although we tried to use all of the samples that were already analyzed in the Galaxy workflow, it was not possible to upload some of the samples because of technical limitations. Due to the lack of time, limited storage, computer pace and internet connection capacities, we used the web interface of INSaFLU workflow (Borges, Pinheiro, Pechirra, Guiomar, & Gomes, 2018) which allows the upload of files smaller than 300 Mb. Because of the file size limitation, we could not try all samples with INSaFLU, while the large size samples have been analyzed in Galaxy workflow. Despite the file size limitation, we were able to upload 50 samples (Supplementary Table 1) to INSaFLU and an INSaFLU was run with these samples. The updated reference genome for COVID-19 (NC045512.2) was used in INSaFLU analyses. The workflow was run with these samples and the reference genome. In the end of the run, we obtained 2 result files, which were validated variants and minor intra-host SNVs (variants below 50% frequency) in .tsv format. The second technical limitation that we faced in INSaFLU was losing some of the samples because of a coverage error. In the lost samples, the percentages of locus size covered by at least 1-fold and by at least 10-fold were less than 100%. The coverage of these samples did not allow the detection of variants (Borges et al., 2018). Despite these technical difficulties, we had a sufficient number of samples (see figure 1) and variants (see figure 2) at the end of the INSaFLU run and we were able to compare the variants from the remaining samples with the Galaxy workflow output.

Figure 1: Number of samples used in the workflows. Samples that are used in the INSaFLU workflow were chosen from published variant files at Galaxy.

Comparison of the samples from both pipelines gave results for 24 samples (see figure 2). Therefore, we continued with these 24 samples to compare found variants. We detected 597 “shared” variations for these 24 samples from the algorithms besides the unique ones, which were 321 in INSaFLU and 2256 in Galaxy (see figure 2). The uniquely found variants in the Galaxy workflow is almost 7 times more than the uniquely found variants by INSaFLU. The reason for that could be the higher sensitivity of the lofreq (Wilm et al., 2012) algorithm which is designed for microbial sequences used in Galaxy. Besides, the different number of total variants detected in different workflows could also be explained by different mapping and filtering parameters used in these workflows. When we looked at the distribution of the shared and unique variants among 24 samples (see figure 2), surprisingly we saw that many of the unique variants in the Galaxy workflow are coming from a few samples (SRR11059942, SRR11059944, SRR11059945), whereas INSaFLU did not detect as many. These samples and their sequence data could be analyzed in detail to find out the reason for such a high amount of unique variants in further studies. Besides, for some samples such as SRR11059940, SRR11059943 and SRR11241255, INSaFLU detected more unique variants than Galaxy. In all samples except SRR11454612, INSaFLU and Galaxy detected shared variants. Variants subjected to Figure 2b were given in the supplementary as four different .tsv files (galaxy_shared.tsv, galaxy_unique.tsv, insaflu_shared.tsv, insaflu_unique.tsv) by adding all of the samples analyzed in Galaxy and INSaFLU in which the same variants were detected. In the insaflu_unique.tsv and insaflu_shared.tsv files, the variants were given as they were annotated in the INSaFLU results, while the galaxy_unique.tsv and galaxy_shared.tsv files include the annotations from the Galaxy results. These files include more than 24 samples because we aimed to take all of the samples in the database which carry the target variants.

Figure 2: a) Venn diagram showing the number of shared and unique variants identified by INSaFLU and Galaxy workflows. b) Bar plot showing the number of shared and unique variants detected by Galaxy and INSaFLU workflows per sample. Red shows shared variants, blue shows INSaFLU-specific variants and green shows Galaxy-specific variants. Number of variants for each condition is given as a function of logarithm in base 2.

Properties and frequencies of variants detected in the samples

By using files that we created in the previous step, we compared detected variants coming from Galaxy and INSaFLU to see different mutation types and abundances of shared variants in the all sequenced samples.

First, we compared the variation types (synonymous and non-synonymous) of all variants. The Number of different variation types in 4 different files that we created is given in Figure 3. Although the same variants are defined in different workflows, they were annotated differently in their results. This gives rise to a difference in the number of variant classes for shared variants. Overall, the amount of non-synonymous mutations are greater than the synonymous ones. Among the non-synonymous variants, 275 missense variants were found in the replicase polyprotein 1ab, and 67 missense variants were found in the spike glycoprotein. Several missense variations were also found in ORF's 7a and 7b, ORF8, ORF3, ORF10, ORF3a, M (membrane protein coding sequence) and in the nucleoprotein sequence.

The protein coding sequences of two proteins, replicase polyprotein 1ab and Spike, have the highest number of non-synonymous variants. When we have a closer look at these proteins, we see that replicase polyprotein is involved in viral RNA replication and transcription (Wu et al., 2020), whereas Spike glycoprotein is a mediator for the attachment of virus to human ACE2 receptor (Zhou et al., 2020). Variations on the coding sequences of these proteins can be further investigated.

Figure 3: Bar plot showing the number of synonymous and non-synonymous variants identified by Galaxy and INSaFLU. Galaxy variants are green, INSaFLU variants are blue. Numbers are given as a function of logarithm in base 2

When we analyzed the reference and alternative allelic states of the shared variants (597), we found that almost half of the (282) variants were multiallelic, whereas 270 variants were biallelic among samples and workflows. This again shows how different workflows create differences in the results. We also found different reference alleles coming from different workflows for the same variants in 45 variants. The possible reason for that is using the updated reference (NC045512.2) for the INSaFLU analysis. When we look at the abundances of shared biallelic variants (detected by both algorithms) among all samples, we see that while some variants are unique for individual samples (Supplementary Table 2), some are more frequent among all samples (see figure 4). We conclude that higher frequency variants which are detected by both workflows in same properties and in many samples are potential segregating sites in the population.

Figure 4: Distribution of biallelic variants seen in more than 5 samples. X-axis shows the variants names in a combination of chromosome and position and Y-axis shows the number of samples which carry the variants.

Conclusion

During the week of Covid-19 Virtual BioHackathon 2020, we compared two different workflows; GalaxyProject and INSaFLU to evaluate their performances in terms of variant calling. We wanted to see whether INSaFLU could identify the same variants with GalaxyProject or not. As a result, we could detect shared variants as well as many unique variants in both workflows. We then analyzed mutation types and their abundances among samples.

Two main issues we have encountered from INSaFLU during our analysis were the limited file upload size and sample coverage. Some of the samples with an exceeding file size could not be uploaded to INSaFLU. Moreover, coverage of numerous assembled reads were found to be low after running the program. Therefore, we were compelled to continue with samples that have higher coverage.

We next identified shared and unique variants for both Galaxy and INSaFLU. We could observe 597 shared variants, with almost 50% of them occurring within the coding sequence of replicase polyprotein 1ab. While categorizing shared variants as either non-synonymous or synonymous, we observed that annotations of some variants were given differently (see figure 3). Additionally, we looked at the allelic properties of the variants, and observed that almost half of them were multiallelic. Distinct reference alleles which come from Galaxy and INSaFLU workflows were present for the same variants. The reason for this could be further investigated. Some of the shared variants were found to be more frequent, hinting to potentially segregating alleles.

Based on our results, we conclude that using multiple workflows can be advantageous to eliminate some of wrong variants as well as identifying true candidates that could be missed (Hwang et al., 2019). The shared and unique variants that we shared in the subject of the study would be considered in this aspect for further studies.

Future Work

The comparison of variants from different workflows can be enriched by adding another workflow to further validate the variant results. The nf-core/viralrecon workflow (Ewels et al., 2020) is one of the strong candidates for such aim, which performs low frequency variant analysis and SNP annotation. The resulting variants from our study could also be compared with the already published variants from different studies such as NextStrain (Hadfield et al., 2018) to see their accuracy. In addition, mutation patterns and types of mutations in shared files could be investigated.

As a result of possible effects of non-synonymous mutations on protein structures and functions, the non-synonymous variants would be interesting to analyze in further studies. These variants could be analyzed in detail to see their possible phenotypic effects on the virus pathogenicity or virus-host interaction. Moreover, it could be highly beneficial to model the effects of non-synonymous mutations at protein structure level. Conformational changes within protein structures resulting from such variants could be illustrated via protein modeling and possible functional changes could be further examined. Modeling of proteins which take part in host infection and host-virus interaction could be useful, since computer aided drug design techniques enable modeling and synthesis of antiviral drugs that could have inhibitory effects on such proteins (Klimenko, 2017). These drugs could be utilized to block host-virus interaction or viral replication pathways and further contribute to decrease the levels of viral infection on the host.

GitHub Repository

The repository contains the supplementary material [<https://github.com/rsgturkeygroup3/BioHackathon>]

Acknowledgements

We would like to express our very great appreciation to the organization team of Covid-19 Virtual BioHackathon 2020 for enabling us to take part in the help for COVID-19 pandemic. Also special thanks to RSG Turkey representatives for their assistance.

References

- Baichoo, S., Souilmi, Y., Panji, S., Botha, G., Meintjes, A., Hazelhurst, S., Bendou, H., et al. (2018). Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support african genomics. *BMC bioinformatics*, 19(1), 457.
- Borges, V., Pinheiro, M., Pechirra, P., Guiomar, R., & Gomes, P., J. (2018). INSaFLU: An automated open web-based bioinformatics suite “from-reads” for influenza whole-genome-sequencing-based surveillance. *Genome Medicine*, 10(1), 46.
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3), 276–278.
- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., et al. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34(23), 4121–4123.

Hwang, K.-B., Lee, I.-H., Li, H., Won, D.-G., Hernandez-Ferrer, C., Negron, J. A., & Kong, S. W. (2019). Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Scientific reports*, 9(1), 1–10.

Klimenko, K. (2017). *Computer-aided drug design of broad-spectrum antiviral compounds* (PhD thesis). Universite de Strasbourg.

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., et al. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, 40(22), 11189–11201.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., et al. (2020). A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798), 265–269.

Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., & Cheng, F. (2020). Network-based drug repurposing for novel coronavirus 2019-nCoV/sars-cov-2. *Cell Discovery*, 6(1), 1–18.