CHALMERS

Department of Biology and Biological Engineering
Applied Bioinformatics, BBT045
2021-03-08

Reproducing: "Using Next Generation Sequencing to Study the Genetic Diversity of Candidate Live Attenuated Zika Vaccines"

Collins, Natalie D Shan, Chao Nunes, Bruno T D Widen, Steven G Shi, Pei-Yong Barrett, Alan D T Sarathy, Vanessa V

Supervisors:
Aleksej Zelezniak
Filip Buric
Sandra Viknander

Authors: **Email:**
Styrbjörn Käll skall@student.chalmers.se
Baptiste Faussurier bapfau@student.chalmers.se

1 Introduction

Zika Virus (ZIKV) is a single-strand positive RNA virus from the family Flaviridiae. It is transmitted via mosquito (Aedes Aegypti) in different areas around the world including Asia, Africa, and more recently Central America and is associated with several diseases [1]. There is currently no treatment against ZIKV and global solutions are based on prevention spread of carrier mosquitos, similar to malaria and dengue fever [2]. Nevertheless, for the treatment of ZIKV the development of a vaccine remains an active area of research.

One group of vaccines used to treat ZIKV is the so called Live-Attenuated-Vaccines (LAVs), which are attenuated variants of the virus. This means that the virus remains viable and is able to infect host cells and trigger the same immune response as the wild types in order to yield a stronger protection. In order to amplify LAVs, the virus has to infect a cell which results in mutations in the viral RNA. If such mutations would cause the virus to regain it's virulence the vaccine is not suitable [3].

In this study, we have chosen to reproduce *"Using Next Generation Sequencing to Study the Genetic Diversity of Candidate Live Attenuated Vaccines"* by Collins et al. [4]. They studied three LAVs (and WT) engineered from the wild type ZIKV strain FSS13025 by deletions of 10, 20, or 30 nucleotides in the 3'UTR region of the genome. The authors wanted to not only see whether the deletions were preserved but also to analyse how the diversity of the viral population changed across passages in Vero cells by using a work pipeline described in an earlier study by the same authors [5].

In this reproduction we have chosen to limit ourselves to only focus on two ZIKV strains: the WTZIKV and the 10-3'UTR-ZIKV (LAV with deletion of 10 nucleotides in the 3'UTR region).

2 Materials and Methods

2.1 Viruses and NGS sequencing

The original study used WTZIKV FSS13025 and 10D-3'UTRZIKV FSS13025 transfected in Vero cells. They were incubated for 9 days, harvested and clarified by centrifugation to generate passage 0 (P0) ZIKV clones. P0 ZIKVs were then passaged into new Vero cultures and harvested to generate P1, then serially through P5. RNA was extracted from culture supernatants, and cDNA libraries were constructed using random hexamers with the TruSeq RNA v2 kit (Illumina) and sequenced on Illumina HiSeq1500 to generate paired end reads for each strain and passage.

2.2 Data Acquisition and QC

The data used for this reproduction was acquired from ArrayExpress, id E-MTAB-8905. The raw data appeared to already have primers removed due to reads having equal read length as the length specified in the Illumina protocol (50 bp). The raw data files were trimmed using Trimmomatic v0.39 by the same settings as stated in the original study. In addition to this the software fastQC was used to visualize the quality of reads before and after trimming.

2.3 *de novo* assembly and alignment

The original study uses *de novo* assembled genome onto which the reads are aligned. The assembly was done with ABySS v2.2.5 and was also attempted here. However, this assembly was not used since it failed to build a continuous genome from the P0 reads due to unknown reasons. Therefore, this step could not be reproduced and subsequently the following step of disassembling the *de novo* sequence and aligning it also failed.

Instead the alignment was done using a reference genome acquired from GenBank, corresponding to the WT FSS13025 strain, id MH158236.1. This was used both for the WTZIKV and the 10D-3'UTR with the knowledge that it would differ from the original study. The alignment was performed using Bowtie2 v2.4.2, a newer version than the one used in the original study, but using the same parameters. The alignment was compressed using SAMtools' v1.11 function "view" and sorted by coordinate using Picard-tools' v2.18.7 function "SortSAM", updated versions from the original study. As a last step before the downstream processing, PCR duplicates were removed using Picard-tools' function "MarkDuplicates".

2.4 Variant detection and Shannon entropy

Variant detection was performed using Vphaser2 v2.0, the same version as in the original study. The list of SNVs and the coverage across the reference genome were stored as .txt files for further

analysis.

Shannon entropy was calculated using Rstudio v3.6 and the package deepSNV v3.10. The sorted .bam files from "MarkDuplicates" were loaded and converted to count matrices using the function "bam2R". These in turn were used to calculate the relative frequencies with the function "RF" which was used to calculate the Shannon entropy from a custom R function provided in the original study. The Shannon entropies were stored as .txt files for further analysis. Custom R-scripts were generated to visualize the results inside Rstudio v3.6 and can be found on GitHub, or through https://github.com/StyrbjornKall/team_project. A full analysis pipeline comparing this reproduction to the original study is seen in Figure 1.

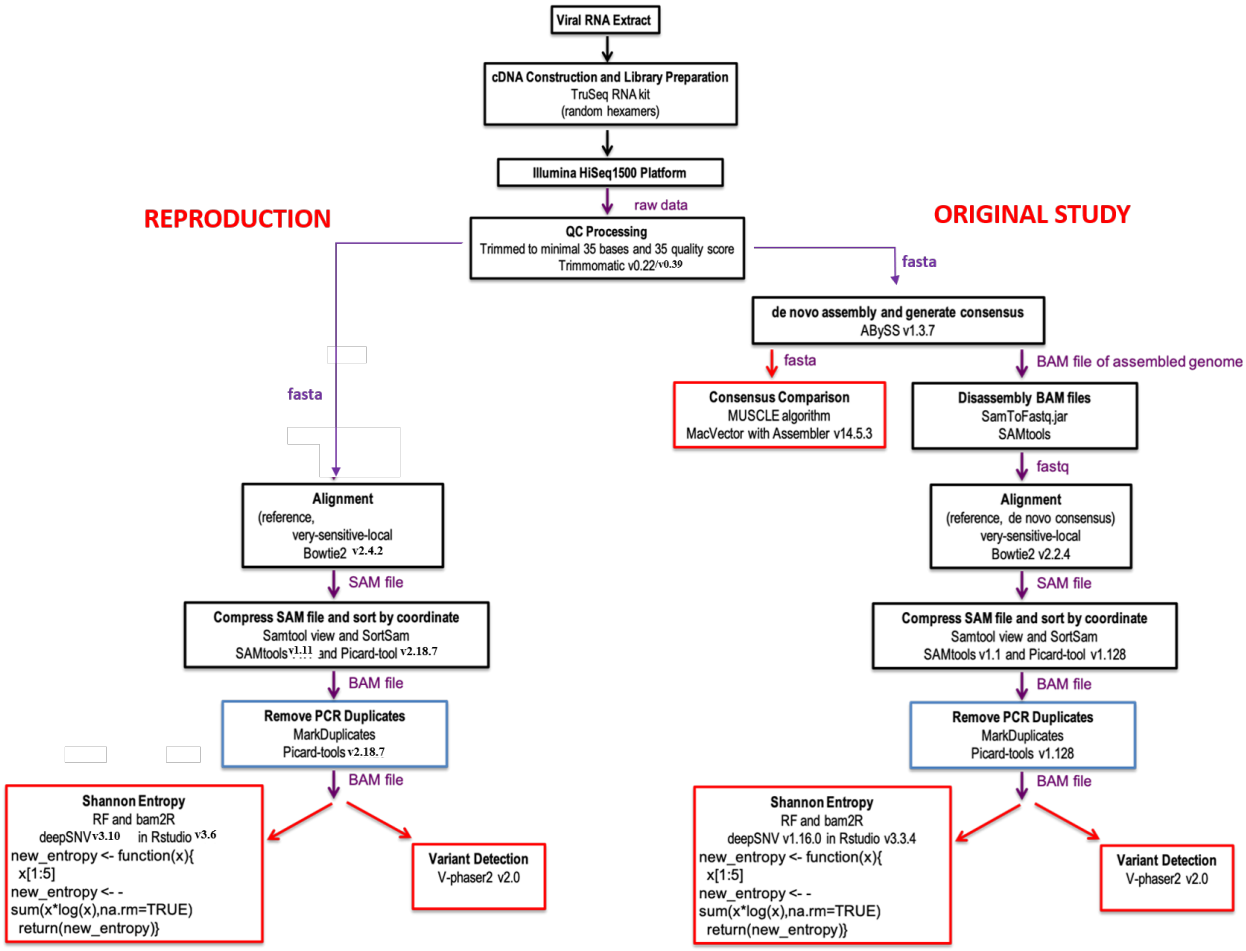
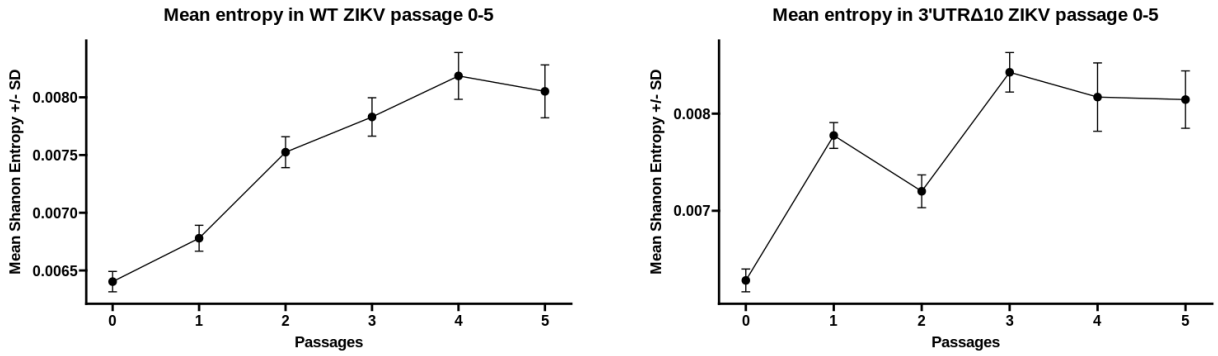


Figure 1: The project pipelines used in this reproduction to the left and the pipeline used in the original study to the right, adapted from original study's supplementary material [5]. The reproduction does not rely on *de novo* assembly for alignment.

3 Results

3.1 Shannon entropy

From the calculated Shannon entropies the mean and standard error was calculated for WTZIKV and 10D-3'UTR mutant, see Figure 2.



(a) WT

(b) 10D-3'UTR

Figure 2: The mean Shannon entropy (+/ standard error of the mean) and passage P0 through P5 for the two strains used in this reproduction. The results show similar trends as the previously reported study.

These results are similar to what was reported in the original study with the WT strain having a gradual increase in mean Shannon entropy across all passages and the 10D-3'UTR mutant having a dip in entropy in passage P2. The main difference is that in the reproduction, entropies are doubled from what was previously reported and entropy in 10D-3'UTR P3 is a lot higher. The conclusion however, remains the same, that with sequential transplantation into new host cells the genome is changing and the diversity increases. This indicates that the sampled virus population is adapting to its host.

Furthermore, the original study reports that entropy is increasing with mean coverage and has a significant Spearman correlation. In Figure 3 the mean entropy is related to mean coverage across the genome for both WT and mutant.

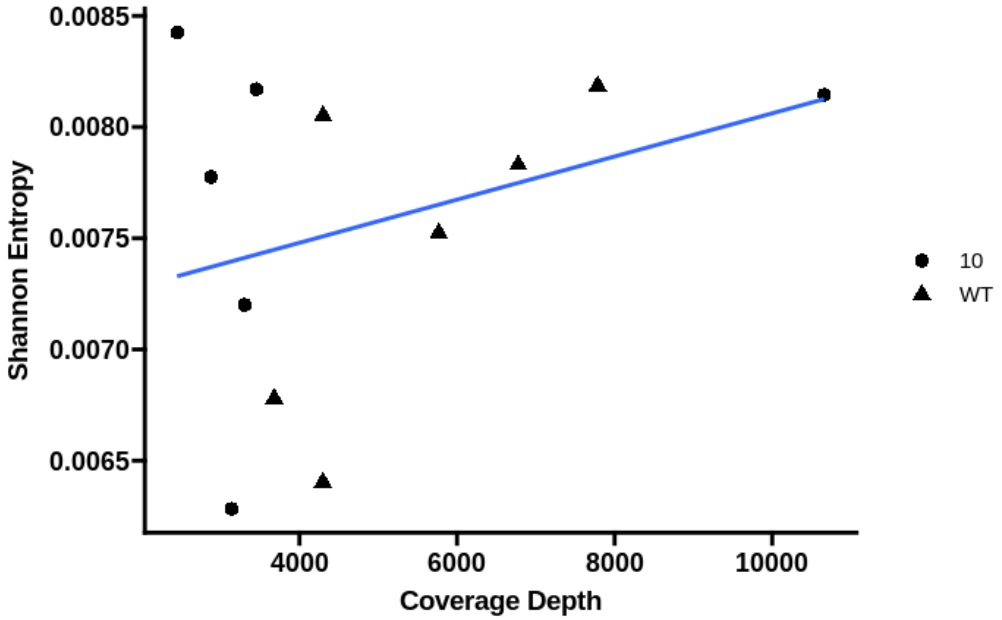
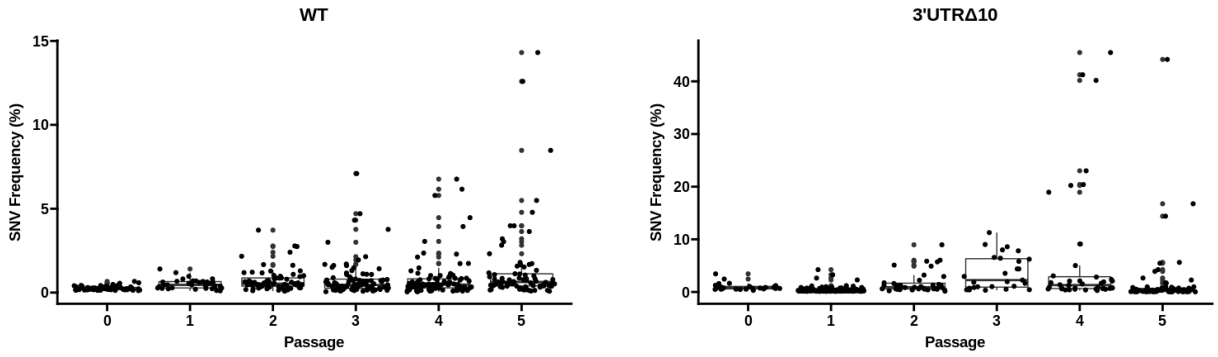


Figure 3: The mean Shannon entropy related to tge mean coverage depth for the WT and 10D-3'UTR mutant strain. Similar trend and positions can be seen in the original study but the p-value for the Spearman correlation is in this case 0.573 and therefore not significant.

The p-value for the two-sided Spearman correlation is in this case 0.573 and therefore not significant. This is different from the original study, reporting a p-value < 0.0001. The difference could be an effect of this study only using two of the four strains. It means that in this case there is no correlation between mean Shannon entropy and mean coverage, indicating that increasing the amount of sequenced RNA will not affect the calculated diversity in the population.

3.2 Variant detection

From the SNVs detected using Vphaser2 v2.0, Figure 4 was created, showing the frequencies for each detected SNV across passages for the WT and 10D-3'UTR mutant.



(a) WT **(b)** 10D-3'UTR

Figure 4: The mean frequency (+/ standard error of the mean) of the detected SNVs in each passage for the WT and 10D-3'UTR mutant. From the figure it is apparent that frequencies increase across passages for both samples.

This shows similar results as the original study showing that more variants are present in later passages. Furthermore, the total number of SNVs were plotted against the mean coverage in Figure 5 showing a non-significant correlation between coverage and the total number of SNVs detected, p-value 0.0749. This is not the same result as in the original study where a p-value of 0.0007 was reported.

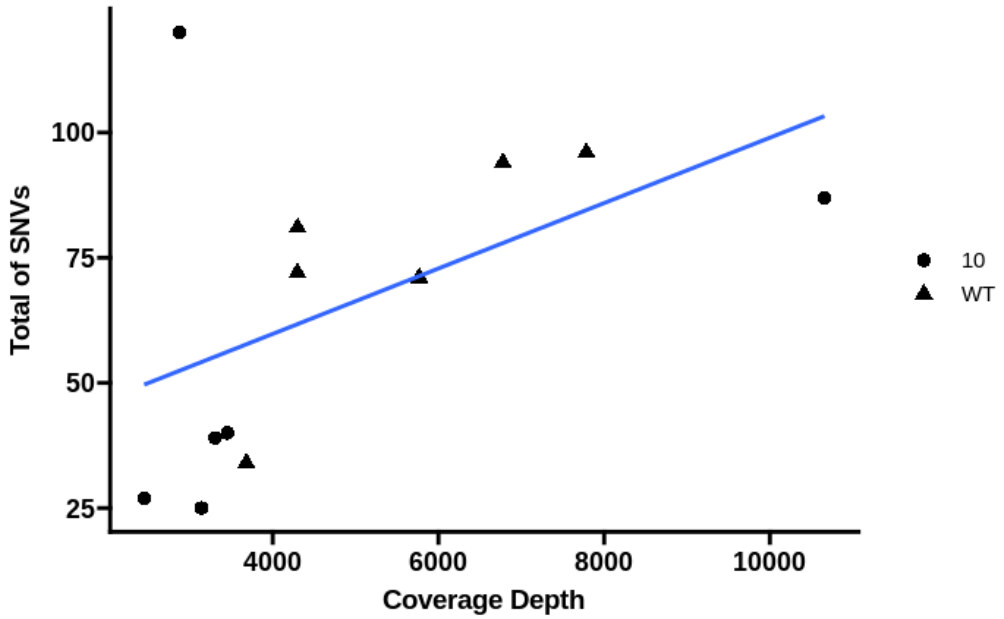


Figure 5: The total number of SNVs present at each passage in both WT and 10D-3'UTR related to the mean coverage. The calculated two-sided Spearman correlation has a p-value of 0.0749

As a final variant analysis, the positions of SNVs present with minimum 5 % frequency were mapped against the reference genome in Figure 6. The genome has been labelled by its genes.

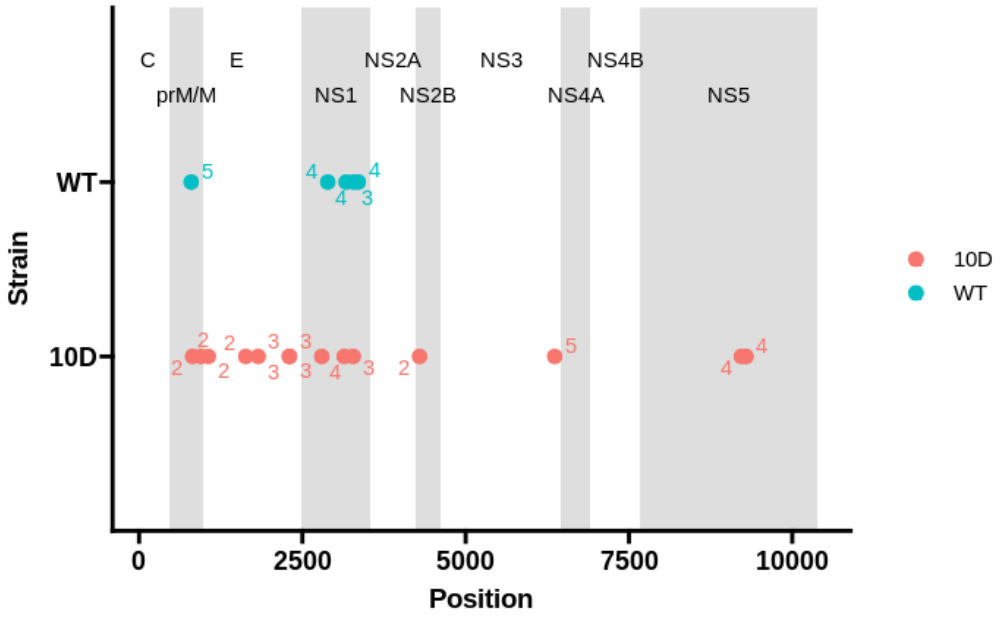


Figure 6: SNVs with frequency greater than 5 % in WT and 10D-3'UTR mutant mapped onto the reference genome with labelled genes.

From Figure 6 it is apparent that the SNVs are present in prM/M, E and NS1 genes with the 10D-3'UTR mutant also having SNVs in the NS2B, NS3 and NS5 gene. This resembles the results reported in the original study apart from the mutations found in NS2B and NS3. This suggests that the proteins encoded by prM/M and NS1 are important for host adaptation in both viruses and E having importance to the 10 deletion mutant, thus also agreeing with what was previously reported. However, this also suggests that NS5 may have importance to the 10D strain.

4 Discussion

This reproduction show similar results as the original study but differ in some areas. The provided pipeline was followed apart from the steps using ABySS, see Figure 1. The divergence can be explained by both inadequate instructions from the study, not including all raw data files as well as the time-span and knowledge we currently possess. The major differences can be seen in the calculated Shannon entropies and the found SNVs. The similarities are that the basic trends for the most part remain the same.

The most important factor that limits this study's reproduction is the missing labels of the raw data files. The study examines three candidate LAVs but only the 10 deletion mutant has the passage labels. Therefore, it is impossible to reproduce the results from the other two mutants.

The second most important factor arises from the inadequate methods and instructions. The original study references instructions from an earlier study by the same authors which, although it remains very similar, does not analyse the exact same thing. This immediately introduces some degree of uncertainty to the reproducibility. Furthermore, the supplementary material of the referenced study leaves questions unanswered, as clear descriptions of the input, parameter settings and output of softwares are absent. For the most part this is not an issue, however for the *de novo* assembly by ABySS the reproduction fails. Instead of aligning to the generated *de novo* sequences a reference genome from GenBank had to be used for further analysis. This could have been solved by easily uploading the generated *de novo* sequences along with the raw data.

Moreover, the use of the *de novo* sequences remain incomplete in the instructions. It is unclear whether a *de novo* sequence is produced for each passage and is used for alignment or if the sequence is only produced from the reads from the first passage and then later used as reference genome for all sequential passages.

In this reproduction we have tried to clarify things and be clear with where deviations from the original study has been made. In hindsight the structuring of the data files could have been restructured. The directories generated are overstructured, meaning there are a lot of subdirectories to different outputs from different softwares. It makes it easy to follow the data generation but also makes it difficult to reproduce since it makes it much harder to make a script that automates the whole project's pipeline. The reason for using this structure arose due to there being a lot of different raw data files belonging to many different groups (mutants and passages) which at the beginning seemed like good data structures. With more labour and time this would have been fixed and automated by a script that, given a list of raw files, could run the whole pipeline from start to finish.

5 Conclusion

In this reproduction we confirm that genetic diversity increases over passages for ZIKV in Vero cells. This is seen by both the increasing Shannon entropy and increasing numbers and frequencies of SNVs. The reproduction neither finds a relation between coverage and entropy or coverge and SNVs. The differences can be assumed to arise from not using all raw data files and not being able to reproduce the *de novo* assembly.

References

1. World Health Organization. Zika virus. 2018. Available from: https://www.who.int/news-room/fact-sheets/detail/zika-virus [Accessed on: 2021 Mar 5]
2. World Health Organization. Vector-borne diseases. 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases [Accessed on: 2021 Mar 5]
3. Yadav DK, Yadav N, and Khurana SMP. Chapter 26 - Vaccines: Present Status and Applications. *Animal Biotechnology*. Ed. by Verma AS and Singh A. San Diego: Academic Press, 2014 :491–508. DOI: https://doi.org/10.1016/B978-0-12-416002-6.00026-2. Available from: https://www.sciencedirect.com/science/article/pii/B9780124160026000262
4. Collins ND, Shan C, Nunes BT, Widen SG, Shi PY, Barrett AD, and Sarathy VV. Using next generation sequencing to study the genetic diversity of candidate live attenuated zika vaccines. Vaccines 2020; 8:161
5. Collins ND, Widen SG, Li L, Swetnam DM, Shi PY, Tesh RB, and Sarathy VV. Inter-and intra-lineage genetic diversity of wild-type Zika viruses reveals both common and distinctive nucleotide variants and clusters of genomic diversity. Emerging microbes & infections 2019; 8:1126–38