Xinhui Cai Thesis

Computational identification of within-host diversity of SARS-CoV-2 and its benefits for mRNA vaccine design

Xinhui Cai

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney

Dec 28th, 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Xinhui Cai, declare that this thesis is submitted in fulfillment of the requirements for the award of Master of Analytics, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 28/12/2022

Acknowledgment

Through out the time of this research I have received a great amount of help and support from many people, without them this research would be impossible. The first person I'd like to thank is my principle supervisor, prof.Jinyan Li. During my master studies, prof.Li helps me form the idea of identifying co-existing strains from the early SARS-CoV-2 sample and provides me valuable feedback and guidance at each stage of my research and helping me with the scholarship which reduces my financial burden considerably. I would also like to thank the other members in the team for their support during my research, especially Tian Lan and Pengyao Ping. Tian made a great contribution in obtaining the protein docking information and Pengyao develops an important part of the double-model error correction method. Dr.Xuan Zhang developed the InsEC which is a source of inspiration for the strain identification algorithm I designed, she also has given me some useful advice from time to time. My parents support me firmly during my research which I appreciate very much. The sequencing read information of SARS-CoV-2 reference genome sequence provided by Dr.Si Haorui from Prof.Zhengli Shi's lab is very important to my research and I appreciate it very much.

I would like to express my gratitude to my co-supervisor Prof.Brian Oliver for providing very helpful insights for both of my papers and playing an important role in forming some of the ideas presented in the papers. His knowledge in virology and immune system has been very helpful.

The Australian Research Council has provided financial support to this research through the scholarship I received. I would like to thank the organization for its recognition and generosity.

I also want to thank the panel of my candidate assessments, Prof.Wei Liu, Prof.Longbing Cao, Prof.Jinyan Li and Prof.Brian Oliver, who have shared their valuable opinions on each stage of my research.

Finally, for people special to me and support me throughout my study and life, I am very grateful and happy to have them.

Xinhui Cai

December 2022 in Sydney

Abstract

Since the emergence of COVID-19 in Wuhan in late 2019, the world has been largely affected. Despite extensive research about the virus itself (SARS-CoV-2) causing the pandemic, some questions regarding its emergence are still unaccounted for. This paper contains an investigation and analysis of the sequencing reads and the SARS-CoV-2 reference genome sequence assembled from those reads. By retracing the steps of assembling the reference sequence, a probability that multiple strains of SARS-CoV-2 co-existed inside the patient's body is found. The assembly tool, MEGAHIT, applies an assembly process that tends to ignore multiple routes to form contigs. Therefore, we design a workflow that could rectify this potential issue and identify multiple strains from the COVID-19 patient sample. It involves error correction, extracting relevant reads, strain identification, phylogenetic study, and protein structure analysis. The results indicate that more than one strain of SARS-CoV-2 could be produced from the sample that was used to produce the reference sequence. Their binding affinity and phylogenetic relationships with the published SARS-CoV-2 reference genome, SARS-CoV, and some other variants of SARS-CoV-2 are also revealed. The discovered strains show some possible structural differences that affect the protein binding affinity between the spike protein and human ACE2. Consequently, this workflow for identifying within-host diversity highlights the existence of co-existing strains with distinct nucleotide sequences, emphasizing the importance of considering these variations when designing mRNA vaccines.

Keywords: SARS-CoV-2 strains; error correction; Illumina short reads; phylogenetic study;mRNA vaccine design

Contents