Bioinformatics

Chu Duc Thang

May 2021

1 Introduction

This is a summary of findings/knowledge/questions about the summary COVID-19 research in University of Alberta. Please take a note that this document is a personal reflection, not a official record so most of writings and explanation is for self-understanding. If you read this document and find out any mistakes, please contact me at chuducthang77@gmail.com so we can discuss.

2 Lesson of basic biology

- The most basics - DNA (including 4 nucleotide A, T, C, G to create different combinations) → RNA (including 4 different nucleotide A, U, C, G). Note that: A of DNA will link with U of RNA → for every 3 nucleotide in RNA forms amino acid → protein. Note that: in virus, the RNA encodes the information, not like DNA in human.
- Structural part of coronavirus: M (membrane), N, S (spike protein), E (envelop)

3 Summary of Coronavirus Evolution: An overview

- Coronavirus (Cov): enveloped, single strand, non-segmented positive-sense RNA viruses (approximately 30kb)
- Phylogentic Analysis: Multiple sequence alignment (MSA) - MAFFT, Protein structural modelling
- Virology
  - 16 non-structural protein and 4 structural protein (nucleocapsid - N, envelop - E, membrane - M, spike - S)
  - M: binds N, viral assembly and budding
  - E: viral shape, dictate how the virus release and the environment they leave

  - S: binding cell receptor, human-human transmission.

- Nucleotide sequence similarity
  - Previous research: Very similar to genome of bat
  - MSA: high similarity supports the intra-sequence similarity
  - Sequence similarity vs sequence identity
  - 3 groups of coronavirus: Alpha (Human coronavirus HCov), Beta (Hcov, SAR, MER), and Delta (mostly animals)

- Type of SAR-COV-2
  - Previous research: 3 types including A (USA/Australia vs Wuhan), B (Wuhan) vs C (Europe/East Asia)
  - RaXML: 3 clades including A (America vs Europe/South Asia), B (East Asia), C (China)

- Envelop protein
  - Help the membrane protein
  - 3 domains: N-terminus, transmembrane domain (TMD), C-terminus
  - Previous research: 4 mostly important mutation including T55S, V56F, E69, deletion of G
  - MSA: highly identical to corona viruses
  - Universal to accept amino acid
  - Highest number of mutation occurs in C-terminus domain
  - Most mutations of envelop protein occured in clade a (USA)

- Membrane protein
  - 3 parts: N-terminal domain, 3 transmembrane domain, carboxy-terminal dormain
  - Previous research: No mutation observed so far in membrane protein
  - MSA: agreement with the previous result

- Spike protein
  - Binding of host cell receptors, human-to-human transmission rate
  - S1 domain (receptor binding - finding ACE2 in human) and S2 domain (membrane fusion)
  - Most important mutation - immunogenicity, viral tropism, pathogenesis
  - G476S and V483A
  - MSA: 614 positions mutated throughout the time, from 5 to 300^th mutated between March and July. From 939 to 1978, no mutations

  - Period: < 100 in Feb, increase to $>$ after April, but reduces to 1971 in May and June, before dropping to 180 positions in July
  - Most difference: N-terminal domain and receptor binding domain. Difference from other proteins, the similarity of SARS-COVID-2 only similar to 1 type of bat coronavirus, even difference from SARS-Cov
  - 1000 fold stronger and stable than other coronavirus, N-terminal domain is well conserved, C-terminal more variations,

- Intermediate host: Potential candidates are pangolin (100% similarity of envelop protein, 98% membrane protein, food habit), turtle and snake
- Hibernating: Using z-score to show the structural similarity with SARS-COV-2. Highest z-score is Myl-CoV, which is hibernating for 4 months in the winter
- Drug and Vaccine
  - Vaccine: Messenger RNA (Moderna/BioNTech) vs evoking cellular immunity (T-cell responses)
  - Drug: Remdesivir

4 Questions:

- There are 4 structural proteins in the virus, but why don't we focus on N (nucleocapsid)?
- Besides MSA, what is the method the research used to confirm the previous research finding?
- What does antigenic, antigenic determinant?
- What does cell mean? What is the difference between b-cell and c-cell?
- What is clade S, L, V, G, GV, GH, GR?
- Asking about the envelop tree? Is CA/2020-03-28 similar to WA/2020-03-19 (for envelop protein)? Is the length in phylogenetic analysis important?
- What is the difference between migration graph and evolution graph?
- Tutorial again on how to use the MSA