Sequencing the SARS-CoV-2 (COVID-19) Virus

Basic Principles of Genome Sequencing

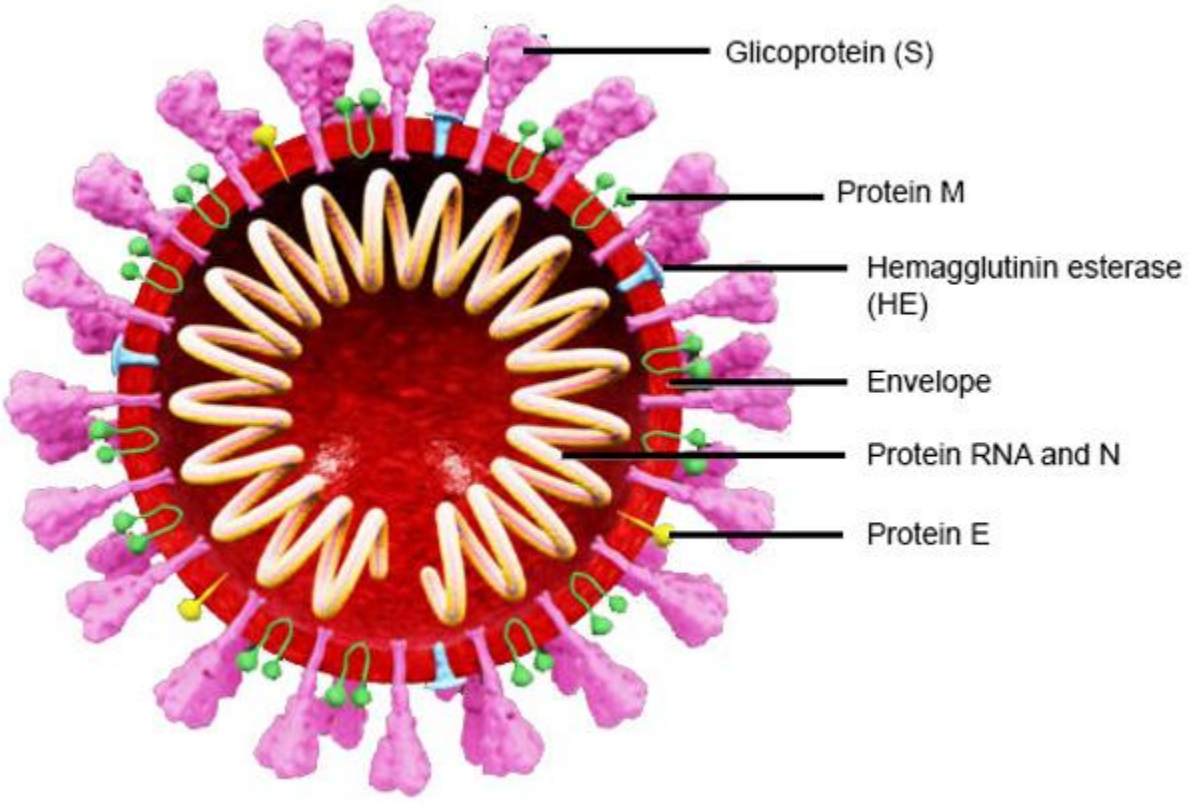
Steps for all projects

1. Identify the genotype/individual that you want to sequence
    - Should represent a source of greatest interest or utility
2. Determine the methodology you will use
    - Library preparation protocol
        - This generates the material that will be sequenced
    - Nucleic acid sequencing protocol
        - Collects the data that will be assembled and annotated
    - Sequence assembly
        - Describes genome organization
    - Annotation
        - Names the genes that are expressed
3. Compare to other species
    - How does your species fit taxonomically with other species

Let's use sequencing of the SARS-CoV-2 (COVID-19) virus as an example of a sequencing project!!!

Genomics of the SARS-CoV-2 (COVID-19) Virus

General Structure of the SARS-CoV-2 Virus



General Background of SARS-CoV-2 Respiratory Disease

Infection agent

- Coronavirus

Coronavirus Taxonomy

- **Realm**: Riboviria
    - **Encode**
        - RNA-dependent RNA polymerase (RdRp) or
        - RNA-dependent DNA polymerase (RdDp)
            - **Reverse Transcriptase**
    - **Examples**
        - HIV, Ebola virus, Flu virus, Coronoaviruses, Measles, Mumps
- **Order**: Nidovirales
    - **Infects**: Vertebrates and invertebrates
    - **Encode**
        - Positive-single strand RNA genome
        - Viral envelope
- **Family**: Coronaviridae
    - **Infects**: Mammals, birds, amphibians
    - **Encode**
        - Positive-single strand RNA genome
        - Viral envelope
        - Spikes
- **Subfamily**: Orthocoronavirinae
    - **Infects**: Mammals, birds, amphibians
    - **Encode**
        - Positive-single strand RNA genome
        - Viral envelope
        - Spikes

Coronavirus History

- 1931: First discovered in chickens in ***North Dakota***
- 1940s: Discovered in mice and pigs
- 1960s: Human coronavirus detected

Most Recent Human Coronaviruses

- **2003: SARS-CoV**
  - ***S***evere ***A***cute ***R***espiratory ***S***yndrome-**Co**rona**V**irus
    - Deaths: 774
    - Fatality Rate: 9.2%
- **2012, 2015, 2018: MERS-CoV**
  - ***M***iddle ***E***ast ***R***espiratory ***S***yndrome-**Co**rona**V**irus
    - Deaths: 858
    - Fatality Rate: 37%
- **2019-???: SARS-Cov-2**
  - ***S***evere ***A***cute ***R***espiratory ***S***yndrome-**Co**rona**V**irus-**2**

	August 12,2020	**September 10, 2021**
Cases	**20,388,405**	**223,296,909**
Deaths	**743,599**	**4,608,047**
Fatality rate	**3.7%**	**2.1%**

Coronavirus Life Cycle

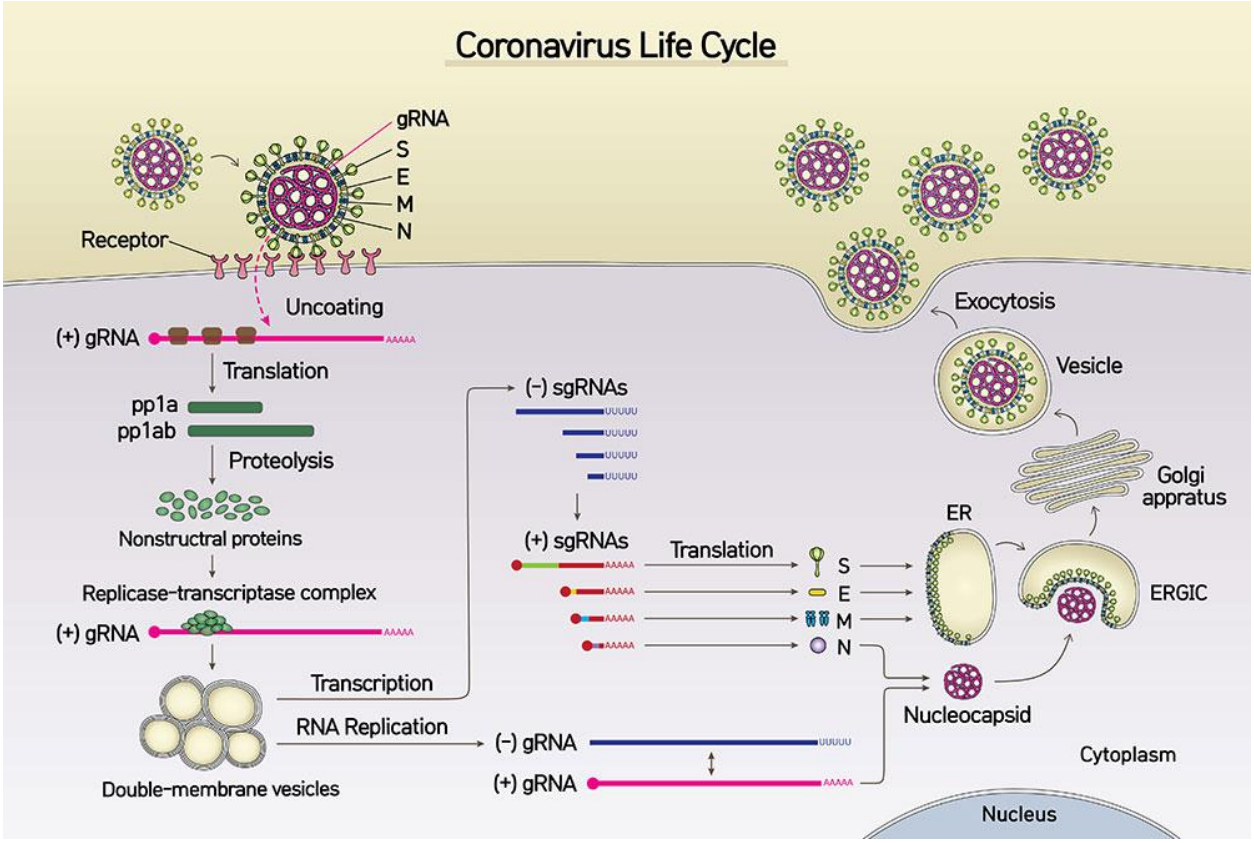
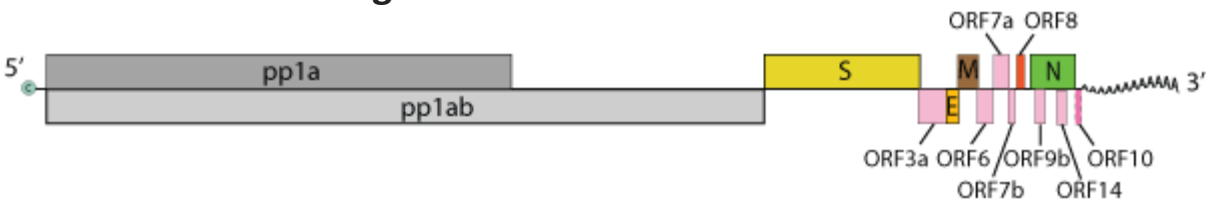


Figure 1 The life cycle of SARS-CoV-2. When the spike protein of SARS-CoV-2 **binds to the receptor** of the host cell, the virus **enters the cell**, and then the **envelope is peeled off**, which let genomic RNA be present in the cytoplasm. The **ORF1a and ORF1b RNAs** are **made by genomic RNA**, and then **translated** into **pp1a** and **pp1ab proteins**, respectively. Protein **pp1a and ppa1b** are **cleaved by protease** to **make** a total of **16 nonstructural proteins**. **Some nonstructural proteins form a replication/transcription complex (RNA-dependent RNA polymerase, RdRp)**, which use the **(+) strand genomic RNA as a template**. The **(+) strand genomic RNA produced through the replication** process becomes the **genome of the new virus particle**. **Subgenomic RNAs** produced through the transcription are **translated into structural proteins** (**S**: *spike protein*, **E**: *envelope protein*, **M**: *membrane protein*, and **N**: *nucleocapsid protein*) which **form a viral particle**. **Spike, envelope and membrane proteins enter the endoplasmic reticulum**, and the **nucleocapsid protein** is **combined** with the **(+) strand genomic RNA** to become a **nucleoprotein complex**. They **merge** into the **complete virus particle in the endoplasmic reticulum-Golgi apparatus compartment**, and are **excreted** to **extracellular region** through the **Golgi apparatus** and the **vesicle**.

Coronavirus Gene Organization



Coronavirus Gene Expression

From: Sawicki et al. 2007. A contemporary view of coronavirus transcription. J. of Virology 81:20-29.

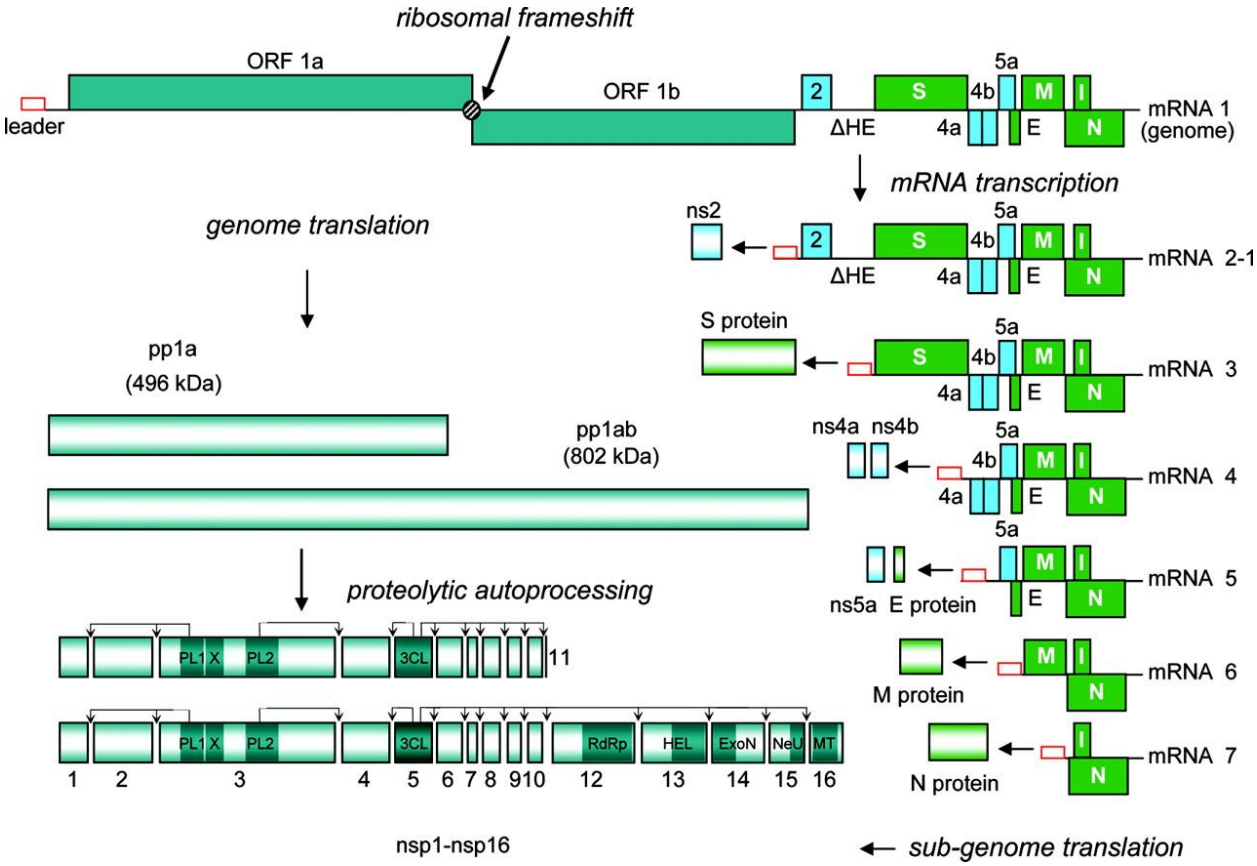
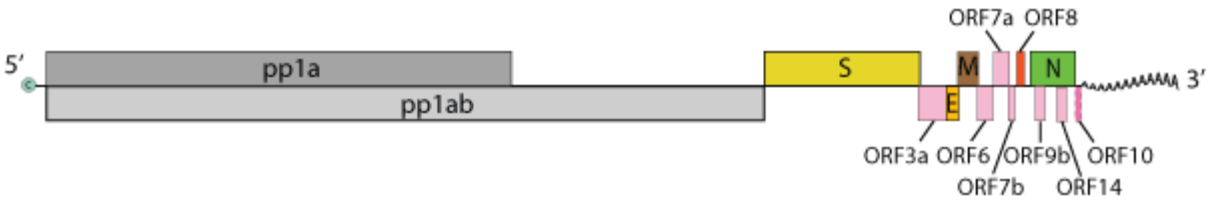


FIG. 1. Organization and expression of the MHV-A59 genome. The structural relationships of the MHV-A59 genome- and subgenome-length mRNAs are shown. The virus ORFs are depicted in teal (nsp1-nsp16 genes), blue (ns2, ns4a, ns4b, and ns5a genes), and green (S, M, E, N, and I structural protein genes). The ORFs are defined by the genomic sequence of MHV-A59 as published by Coley et al. (20). The open red box represents the common 59-leader sequence, and the barred circle represents the programmed (1) frameshifting element. The translation products of the genome and subgenome-length mRNAs are depicted, and the autoproteolytic processing of the ORF1a and ORF1a/ORF1bpolyproteins into proteins nsp1 to nsp16 is shown. A number of confirmed and putative functional domains in the nsp proteins are also indicated. NeU, uridylate-specific endoribonuclease; PL1, papain-like protease 1; PL2, papain-like protease 2.

SARS-CoV-2 Gene expression



ORF 1ab Genes

- Encodes mRNA transcribed into a polyprotein
  - Polyprotein
    - Almost exclusively found in viruses
    - A single translated protein processed into multiple proteins
    - Each protein can have a function in pathogenicity
    - Coronavirus polyprotein
      - ***Largest polyprotein of any RNA virus***
- ORF 1a and 1b
  - Translated from mRNA in the envelope of the original particle
    - ORF 1a: ~500 kDa polyprotein
      - nsp1 – nsp11
      - nsp = Non-structural protein
    - ORF 1b: ~800 kDa polyprotein
      - nsp12 – nsp16
    - **nsp12 = RNA dependent RNA polymerase; replicase**

Sequencing the SARS-CoV-2 Virus

Notes summarized article published in *Feb, 2020*:
Wu et al. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265-269.

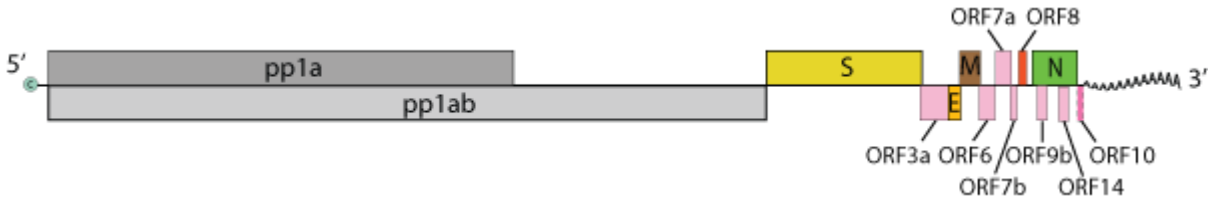
Patient

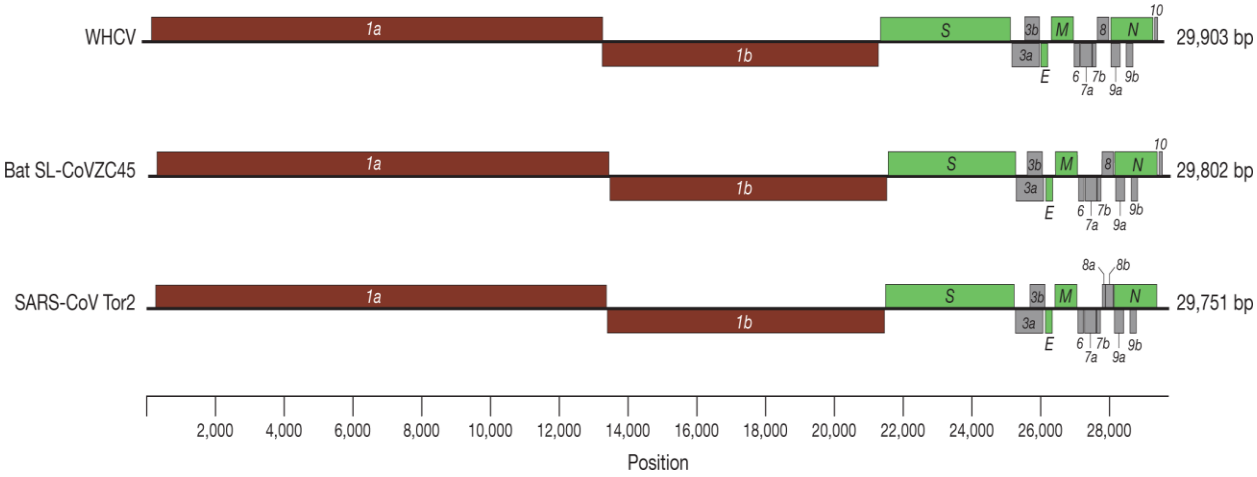
- 41-year-old man
    - Worked in Wuhan market
    - Symptoms
        - Fever, chest tightness, unproductive cough, pain, and weakness for 1 week on presentation
        - Mild hypoxaemia (low blood oxygen level)
        - Air-space shadowing and patchy consolidation in both lungs
- Negative for:
    - Influenza virus, *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* and Human Adenovirus
        - Screened by:
            - Commercial pathogen antigen-detection kits
            - Confirmed by PCR tests

Sequencing approach

- All work performed in biosafety level 3 laboratory
- Common requirements in a BSL-3 laboratory include: (https://consteril.com/biosafety-levels-difference/)
    - Standard personal protective equipment must be worn, and respirators might be required
    - Solid-front wraparound gowns, scrub suits or coveralls are often required
    - All work with microbes must be performed within an appropriate BSC
    - Access hands-free sink and eyewash are available near the exit
    - Sustained directional airflow to draw air into the laboratory from clean areas towards potentially contaminated areas (Exhaust air cannot be re-circulated)
    - A self closing set of locking doors with access away from general building corridors

- Collected bronchoalveolar lavage fluid (BALF)
- Performed deep meta-transcriptomic sequencing
    - Total RNA was extracted from 200 µl of BALF
    - Meta-transcriptomic library constructed
    - Pair-end (150-bp reads) sequencing
        - Illumina MiniSeq
            - ***56,565,928 sequence reads***
- *deNovo* genome assemby
    - Longest assembly
        - ***30,474 nucleotides (nt)***
- Confirmed by reverse-transcription PCR (RT–PCR)
    - 5′/3′ rapid amplification of cDNA ends (RACE)
        - ***Whole genome sequence = 29,903 nt***
- Strain designated as WH-Human 1 coronavirus (**WHCV**)
    - Also been referred to as '2019-nCoV'
    - Closely related to a bat SARS-like coronavirus (CoV)
        - China strain: bat **SL-CoVZC45**
        - 89.1% nucleotide identity
- Viral load in the BALF sample
    - 3.95×10^8 copies per ml





WHCV genome organization

- Determined by sequence alignment to
  - Human coronavirus: **SARS-CoV Tor2**
  - Bat coronavirus: **Bat SL-CoVZC45**

Supplementary Table 5. Predicted gene functions of WHCV ORFs. In 5' to 3' order.

ORF name	Size (nt)	Proposed function
ORF 1a	21,291 for both	**Encoded nonstructural proteins** (*nsp1* to *nsp11*): essential for viral replication, viral assembly, immune response modulation, etc
ORF 1b		**Encoded nonstructural proteins** (*nsp12* to *nsp16*): essential for viral replication
S	3,822	**Spike protein**: binding to cell receptor and mediate virus-cell fusion
ORF 3a	828	Accessory protein
ORF 3b		Accessory protein
E	669	**Envelope protein**: virus assembly and morphogenesis
M		**Membrane protein**: virus assembly
ORF 6		Accessory protein
ORF 7a		Accessory protein
ORF 7b		Accessory protein
ORF 8	366	Accessory protein
N	1,260	**Nucleocapsid protein**: forms complexes with genomic RNA, interact with M protein for viral assembly
ORF 9a		Accessory protein
ORF 9b		Accessory protein
ORF 10		Accessory protein

Phylogeny: Discovering the ancestry of SARS-CoV-2

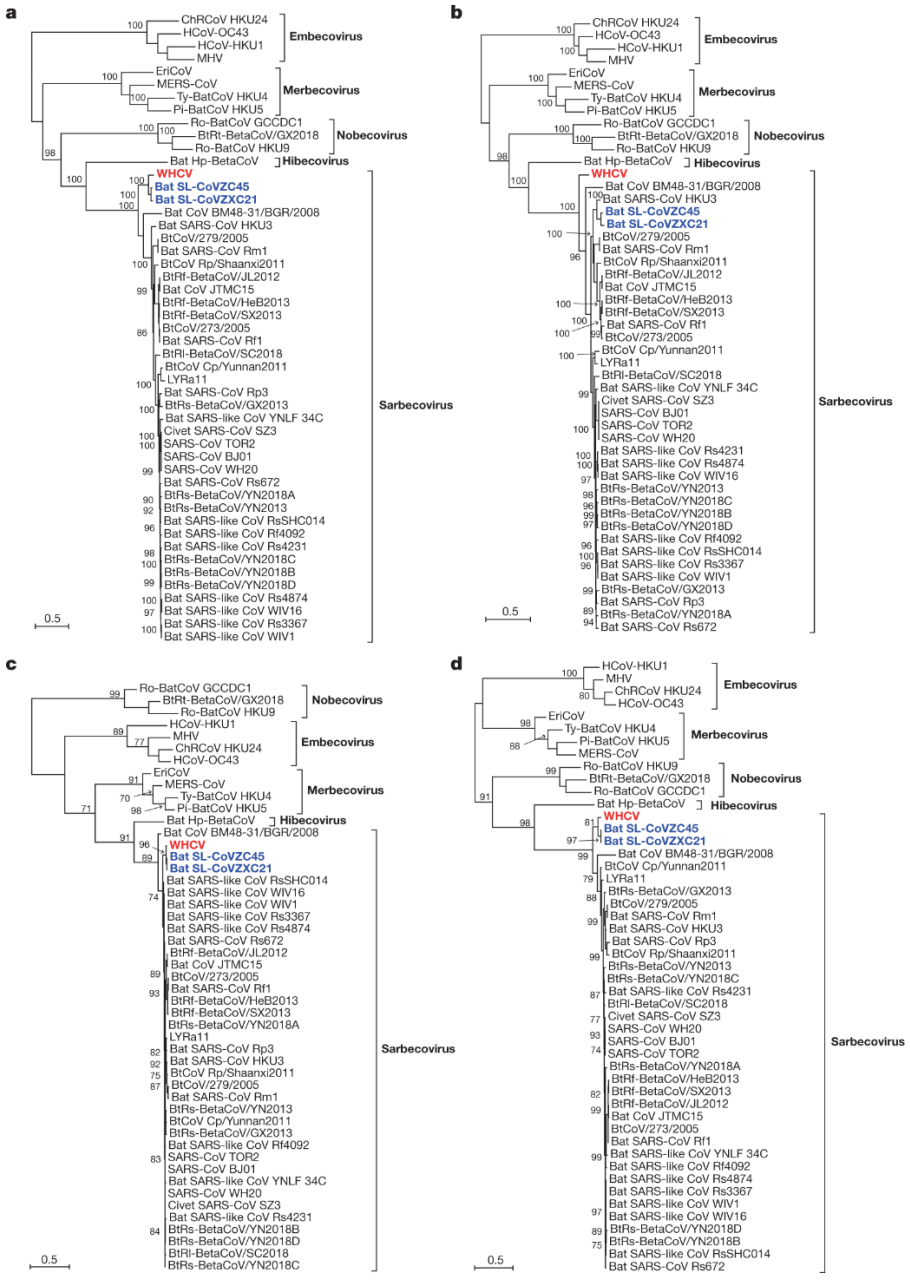


Fig. 2: Maximum likelihood phylogenetic trees of nucleotide sequences of the *ORF1a*, *ORF1b*, *E* and *M* genes of WHCV and related coronaviruses. **a**, Phylogenetic trees of ***ORF1a***. **b**, Phylogenetic trees of ***ORF1b***. **c**, Phylogenetic trees of ***E***. **d**, Phylogenetic trees of ***M***. EriCoV, Erinaceus coronavirus. Numbers (>70) above or below the branches indicate percentage bootstrap values for the associated nodes. The trees were mid-point rooted for clarity only. The scale bar represents the number of substitutions per site.

Covid Variants

https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html

- How discovered
  - Genome sequencing of infected patients
- How classified
  - **Variant of Interest** (Eta, Iota, Kappa, Unnamed)
    - "A variant with specific genetic markers that have been associated with changes to receptor binding, reduced neutralization by antibodies generated against previous infection or vaccination, reduced efficacy of treatments, potential diagnostic impact, or predicted increase in transmissibility or disease severity."
  - **Variant of Concern** (Alpha, Beta, Gamma, Delta)
    - "A variant for which there is evidence of an increase in transmissibility, more severe disease (e.g., increased hospitalizations or deaths), significant reduction in neutralization by antibodies generated during previous infection or vaccination, reduced effectiveness of treatments or vaccines, or diagnostic detection failures."
  - **Variant of High Consequence** (None)
    - "A variant of high consequence has clear evidence that prevention measures or medical countermeasures (MCMs) have significantly reduced effectiveness relative to previously circulating variants."
- How defined
  - Amino acid variation in the spike protein
  - Amino acid variation in receptor binding domain

Variants of Concern in United States

Data from: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html

	Alpha	Beta	Gamma	Delta
PANGO definition	B.1.1.7	B.1.351	P.1	B.1.617.2
Country first detected	United Kingdom	South Africa	Japan & Brazil	India
# spike protein mutations	10	10	11	15
RBD mutations	N501Y	K417N, E484K, N501Y	K417T, E484K, N50Y	K417N, L452R, T478K
Comments	**Increased** transmission	**Increased** transmission **Significant** reduction in antibody effectiveness **Reduced** effectiveness of convalescent plasma	**Significant** reduction in antibody effectiveness **Reduced** effectiveness of convalescent plasma	**Increased** transmission **Potential** reduction in antibody effectiveness **Potential r**educed effectiveness of convalescent plasma

RBD = Receptor Binding Domain