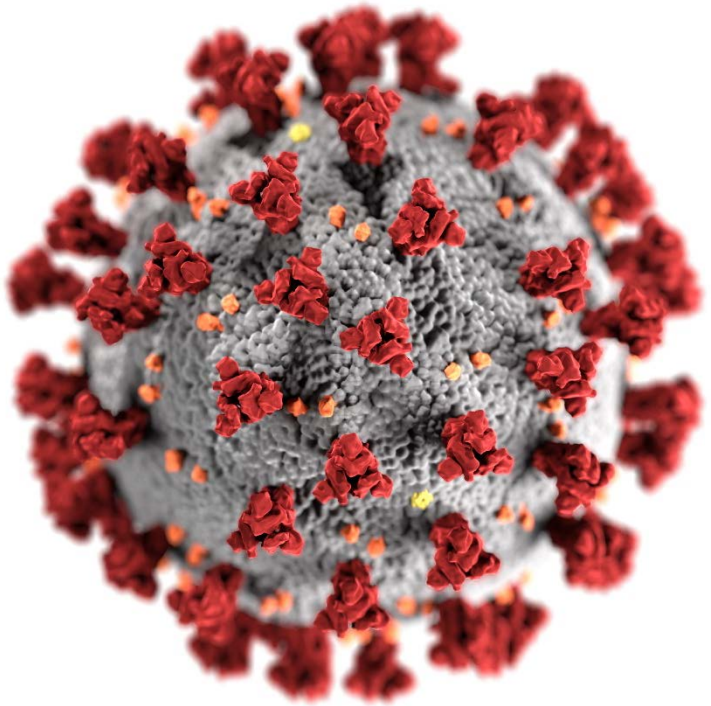
The SARS-CoV-2 genome

COVID-19 Genomic Epidemiology Toolkit: Module 1.2

Shatavia S. Morrison, PhD
Bioinformatics Unit Lead
Centers for Disease Control and Prevention



cdc.gov/coronavirus

Toolkit map

Part 1: Introduction

1.1 What is genomic epidemiology?

1.2 The SARS-CoV-2 genome

1.3 How to read phylogenetic trees

Part 2: Case Studies

2.1 SARS-CoV-2 sequencing in Arizona

2.2 Healthcare cluster transmission

2.3 Community Transmission

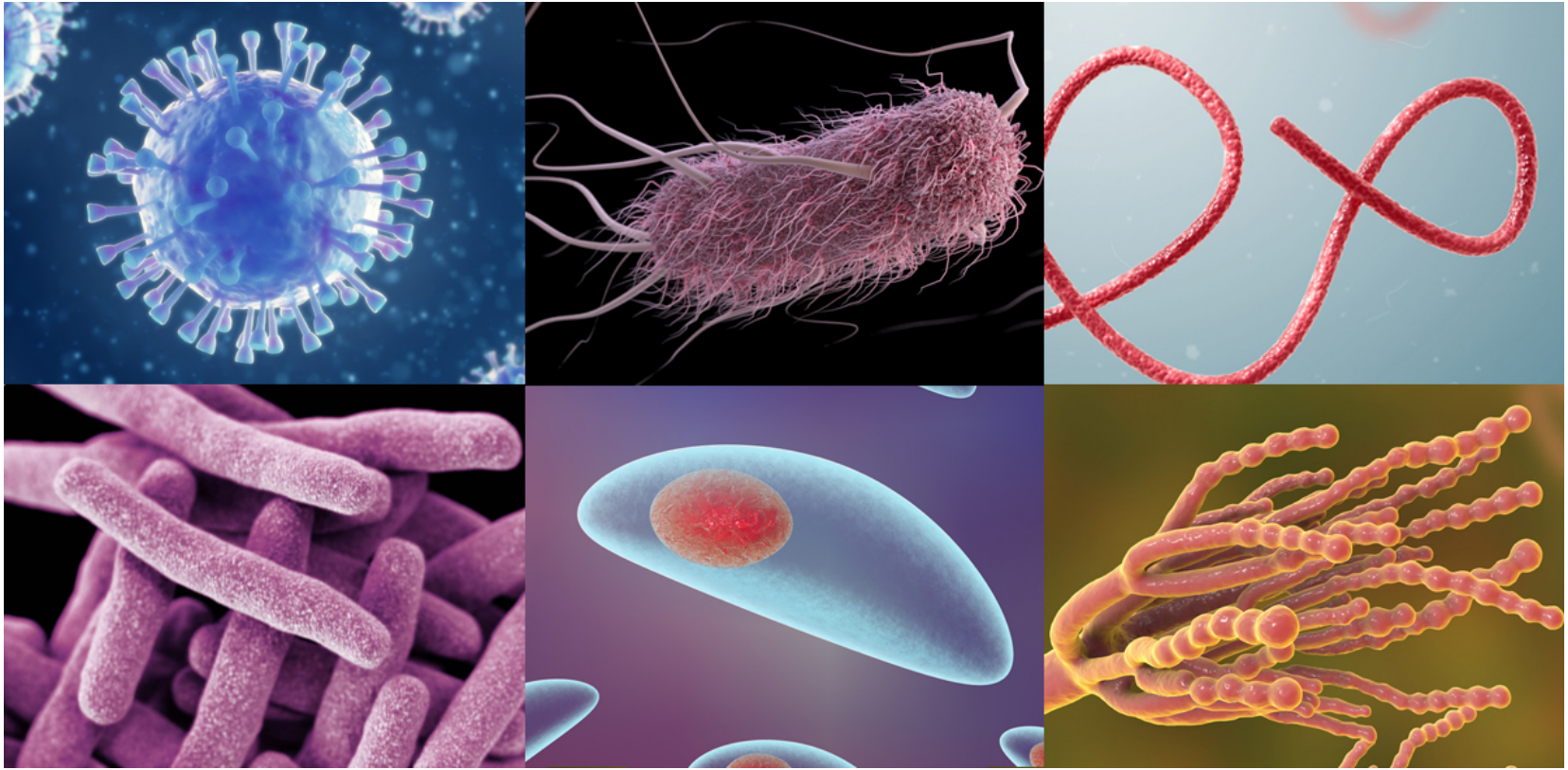
Part 3: Implementation

3.1 Getting started with Nextstrain

3.2 Getting started with MicrobeTrace

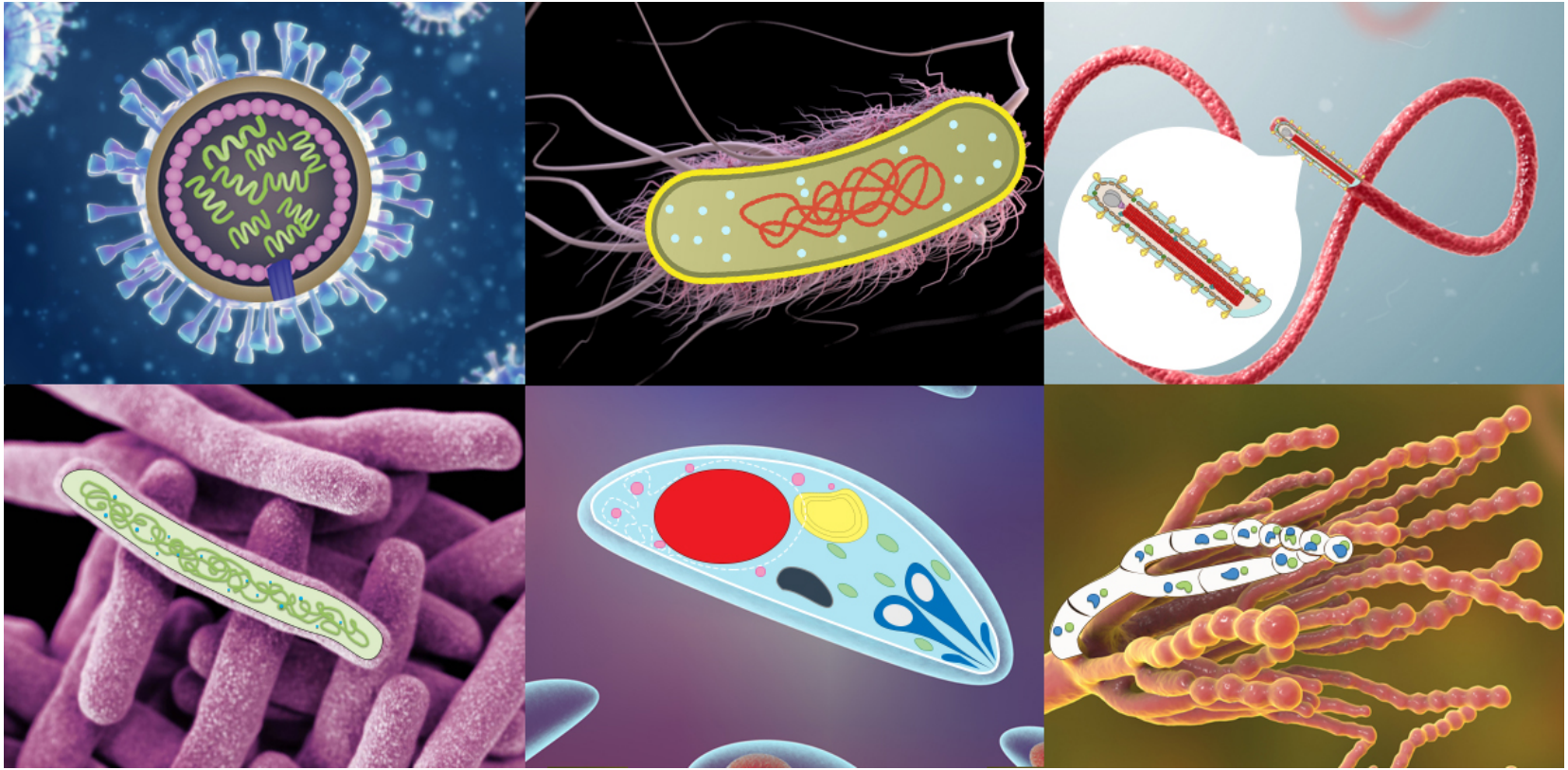
3.3 Linking epidemiologic data

Microbial pathogens are diverse



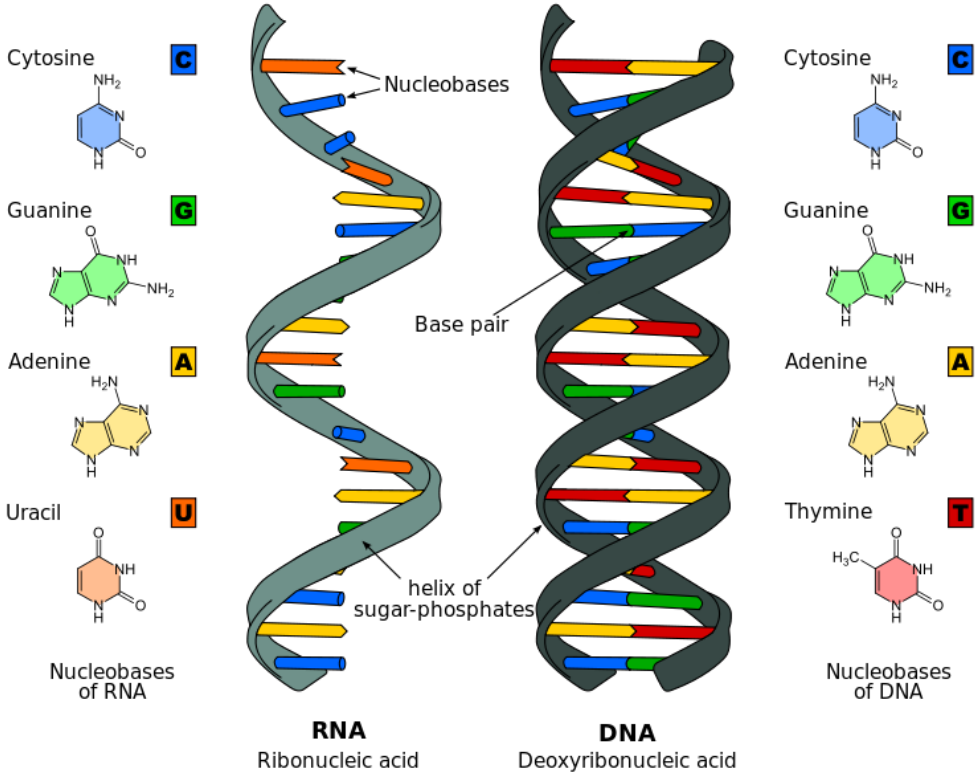
Images: Virus (Getty Images), *E. coli* (PHIL- CDC), Ebola (Getty Images), *Mycobacterium tuberculosis* (PHIL - CDC), *Toxoplasma gondii* (Getty), Fungi *Penicillium* (Getty)

(Almost) Every microbial pathogen has a genome



Structure illustrations: Virus (Getty Images), *E. coli* (Getty Images), Ebola (CDC), *Mycobacterium tuberculosis* (CDC), *Toxoplasma gondii* (CDC), Fungi *Penicillium* (CDC)

Nucleotides are the building blocks of genomes



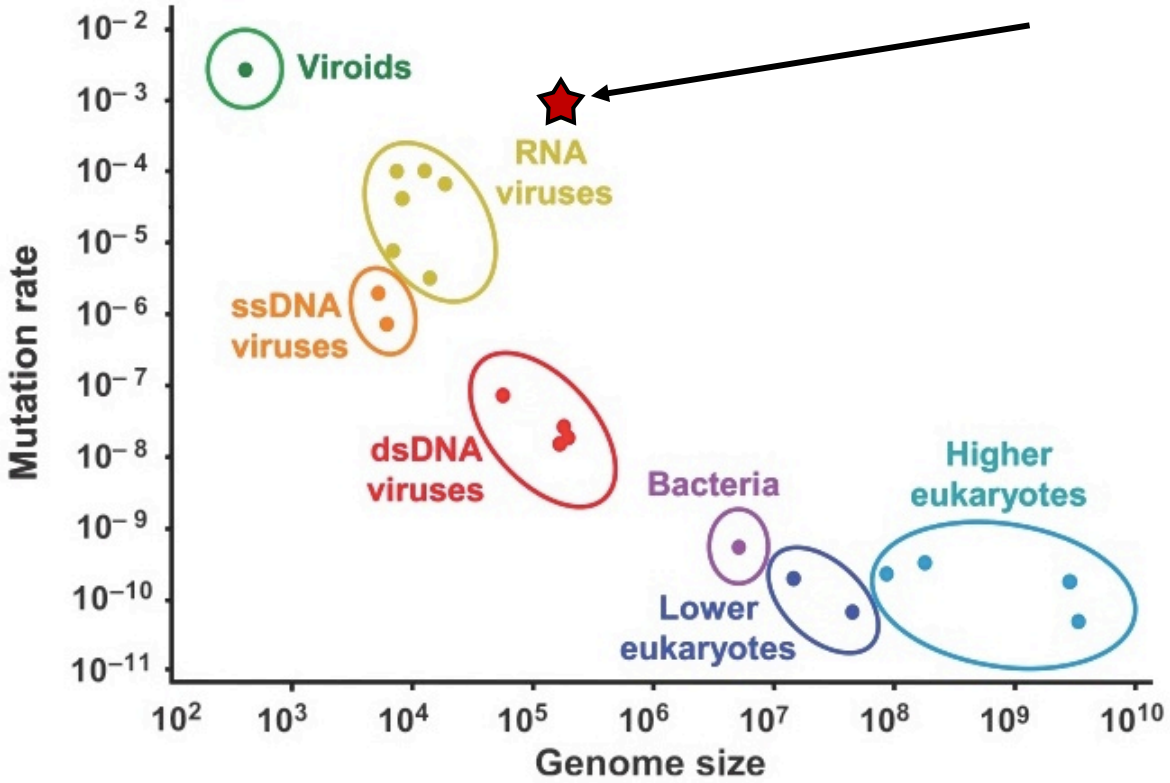

Studying the entire genome

Variations in genome size

SARS-CoV-2

Nucleotides: ~30,000

Substitution rate: ~10^{-4} -10^{-3}



Adapted from Gago, S et al. (2009) Extremely High Mutation Rate of a Hammerhead Viroid | Science (sciencemag.org)

Viruses

- Compact genomes
  - 10,000s nucleotides
- Variable structure, composition
- Either RNA or DNA genomes
- Often highly variable
  - Particularly true of ssRNA viruses

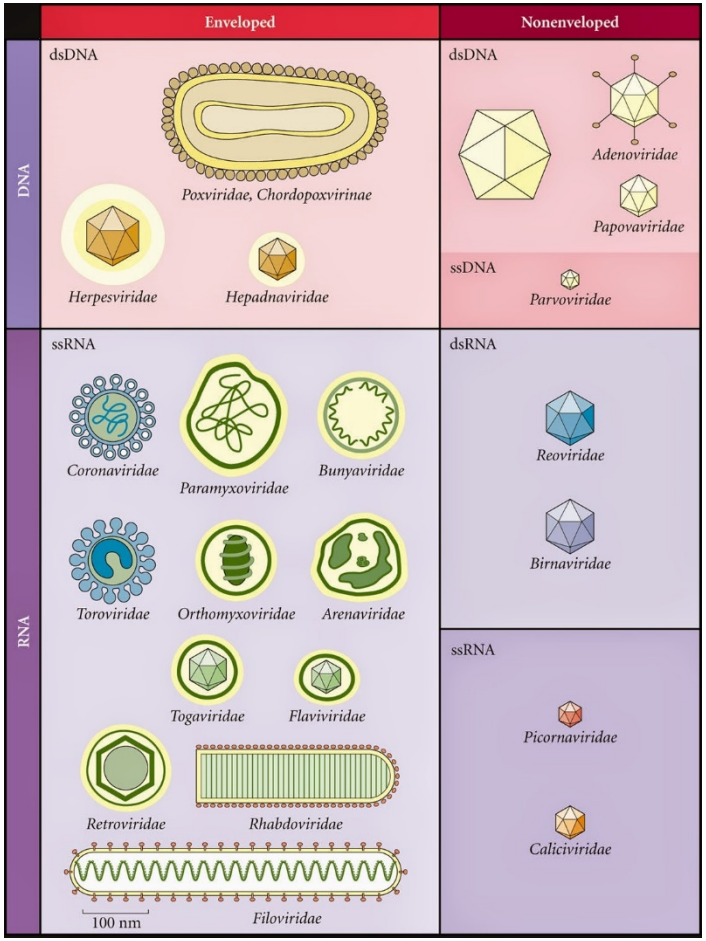
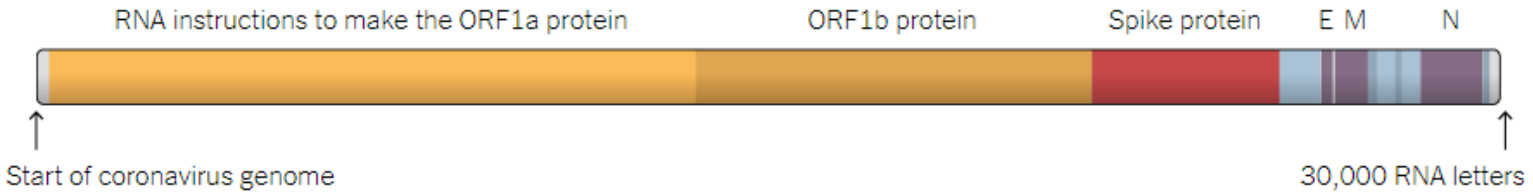
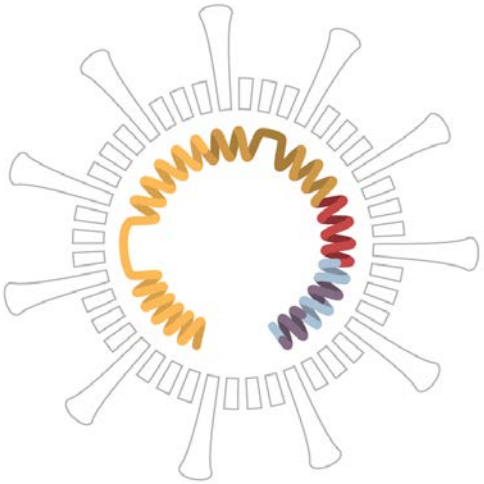


Image from Pathogen Profile Dictionary https://ppdictionary.com/viruses.htm

The SARS-CoV-2 genome



- RNA virus (single-stranded, positive-sense)
- Linear genome = ~30,000 nucleotides
- 11 coding-regions (genes)
- 12 potential gene products
  - e.g., Spike protein



Naqvi et al. (2020) Insight into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approaches, PMID:32544429

Images from *The New York Times* "How Coronavirus Mutates and Spreads" www.nytimes.com/interactive/2020/04/30/science/coronavirus-mutations.html

Fingerprinting and phylogenetics

- Mutations in the genome produce a fingerprint that can be used to infer ancestral relationships (phylogeny), the topic of Module 1.3

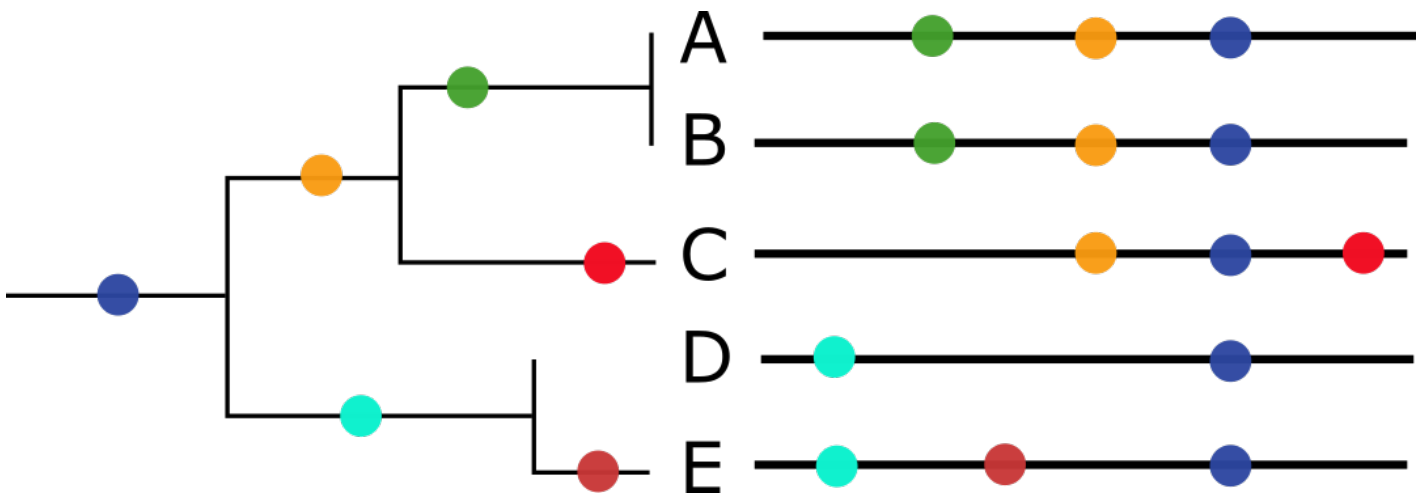
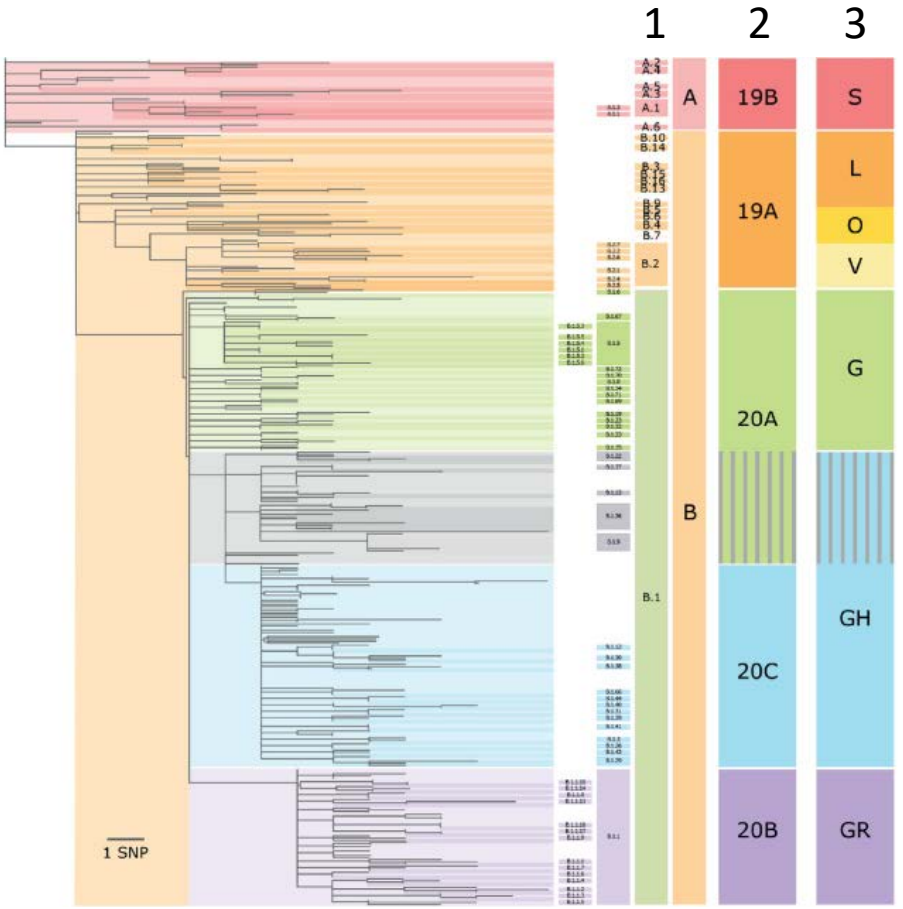


Image from Trevor Bedford Group: https://docs.nextstrain.org

SARS-CoV-2 clades:

Clade naming conventions

1. Pangolin Lineages
   - cov-lineages.org
2. Clades by Nextstrain ****
   - nextstrain.org
3. Clades by GISAID
   - gisaid.org



Adapted from Alm et al. 2020

Rationale for sequencing of SARS-CoV-2

- Monitor trends at the national level
  - Monitor emergence of important new strains
  - Monitor trends after interventions such as vaccination

- Better understand epidemiology at the local level
  - Investigate transmission in healthcare settings
  - Investigate clusters in other settings
  - Reveal important, unsuspected clusters
  - Provide evidence for or against suspected transmission

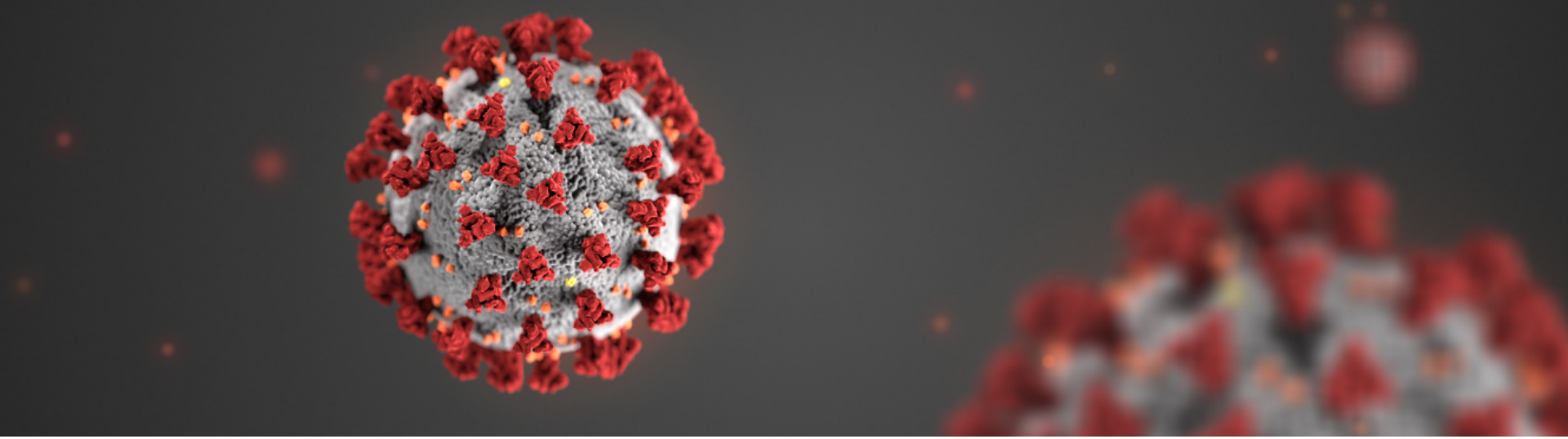
Summary

- SARS-CoV-2 contains a linear RNA genome of ~ 30,000 nucleotides
- Whole genome sequencing can be used to identify genetic mutations in the SARS-CoV-2 genome
- Genome fingerprinting and phylogenetics can be used to:
  - Separate circulating SARS-CoV-2 into ‘clades’ or ‘lineages’ with standard nomenclature
  - Identify potential outbreak clades or source attribution

Learn more

- Other introduction modules
  - What is genomic epidemiology? – Module 1.1
  - How to read a phylogenetic tree – Module 1.3

- COVID-19 Genomic Epidemiology Toolkit
  - Find further reading
  - Subscribe to receive updates on new modules as they are released
  - go.usa.gov/xAbMw



For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.