Bio 208FS: Computing on the Genome

Paul Magwene

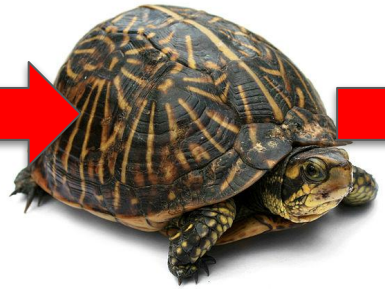
Goals for today

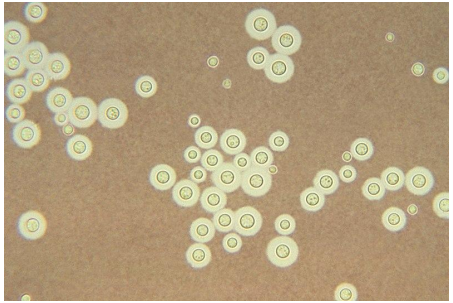
- Introductions
- Course policies
- Course overview
- First assignment

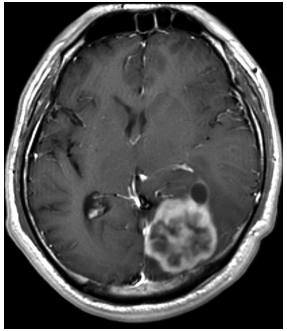
Who am I?

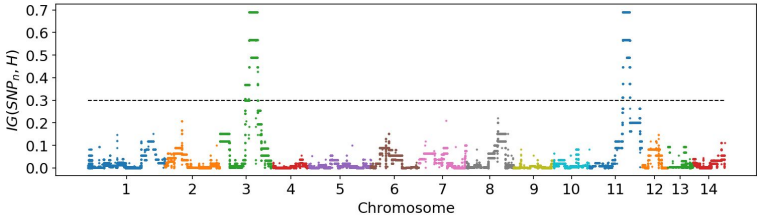
- Associate Professor in the Biology Department
- Secondary appointment in Molecular Genetics & Microbiology
- Former Director of the Computational Biology & Bioinformatics grad program at Duke (2015-2020)
- Joined Duke in fall of 2004
- Prior to coming to Duke my academic journey included:
  - Undergrad at Harvard
  - Grad school at the University of Chicago
  - Postdocs at Yale and UPenn

My scientific journey: Dinosaurs to Genomes









Who are you?

- Where are you from?
- What are your academic interests/likely major?
- What got you interested in the Genetics & Genomics Focus cluster?
- Do you have any experience with computer programming languages? If so, which ones?
- Do you had the opportunity to participate in scientific research?
- Anything else you'd like to share!

Course policies and grading

- Policies
  - Duke Compact
  - Academic integrity
  - Missed class time
- Grading
  - Short assignments
  - Longer projects
  - Late submissions

See the course wiki:
https://github.com/bio208fs-class/bio208fs-lecture/wiki/Expectations-and-Policies

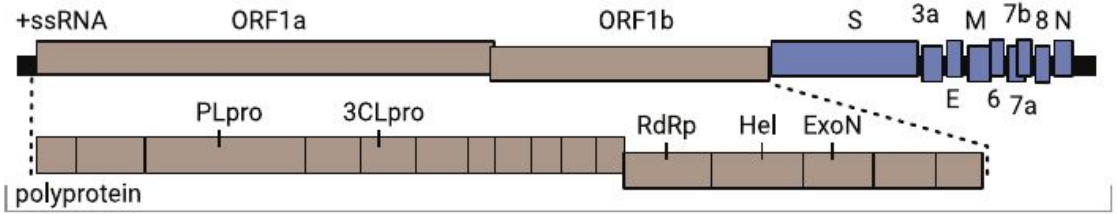
Computing on the Genome: Learning objectives

- Genome biology
  - structure, function, variation, evolution
- Genome technology
  - high throughput sequencing
  - sequence data as a proxy for numerous aspects of molecular function
- Genome bioinformatics
  - Assembly
  - Gene calling and annotation
  - Sequence alignment
  - Building phylogenetic trees
- Working with "big data": algorithmic and quantitative approaches
  - Data curation, filtering
  - Data exploration and visualization
  - Clustering and dimensional reduction techniques
- An applied introduction to computer programming

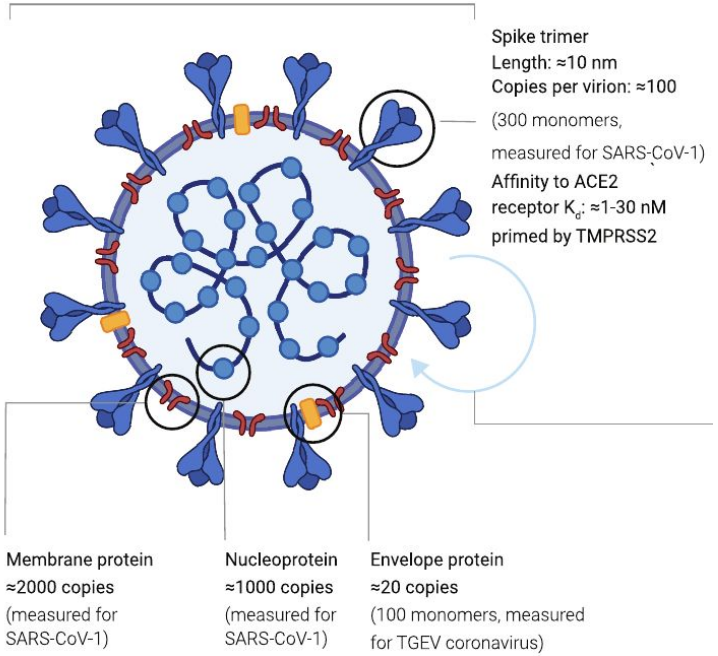
~5% of the COVID-19 (SARS-CoV-2) Reference Genome

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA

From sequence to structure

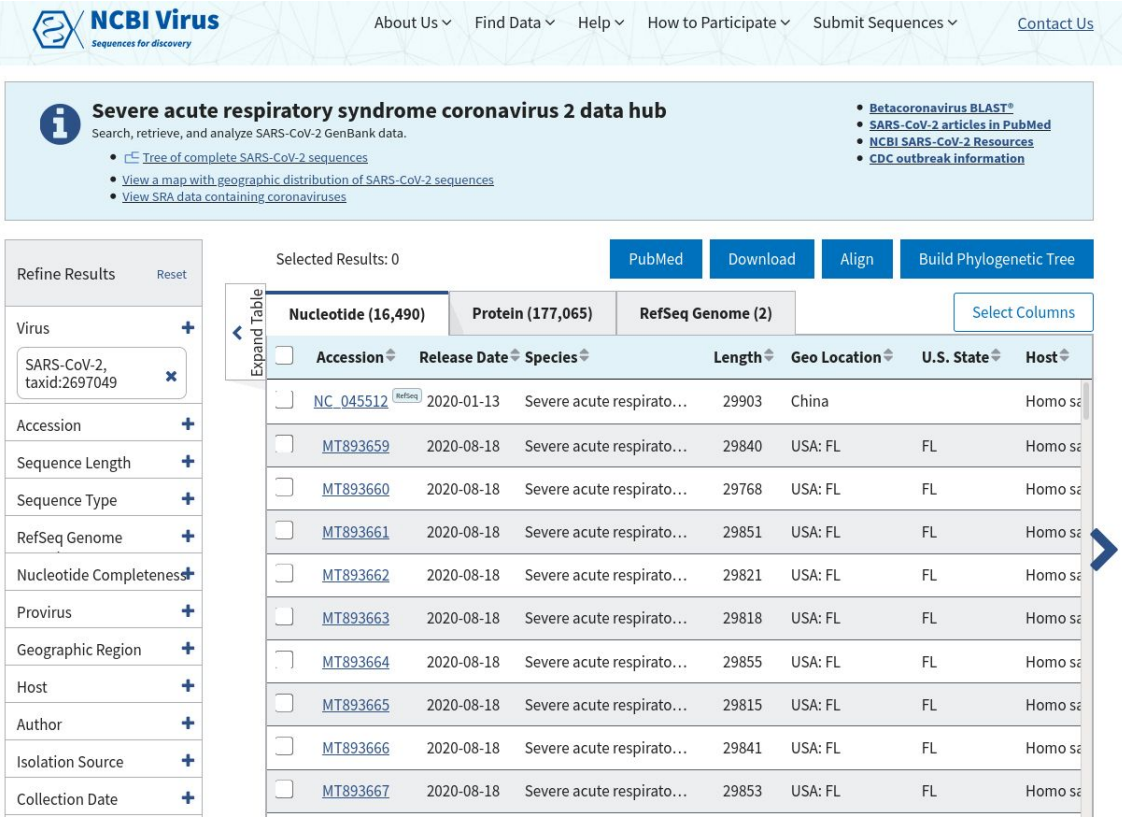


Length: ≈30kb; β-coronavirus with 10-14 ORFs (24-27 proteins)



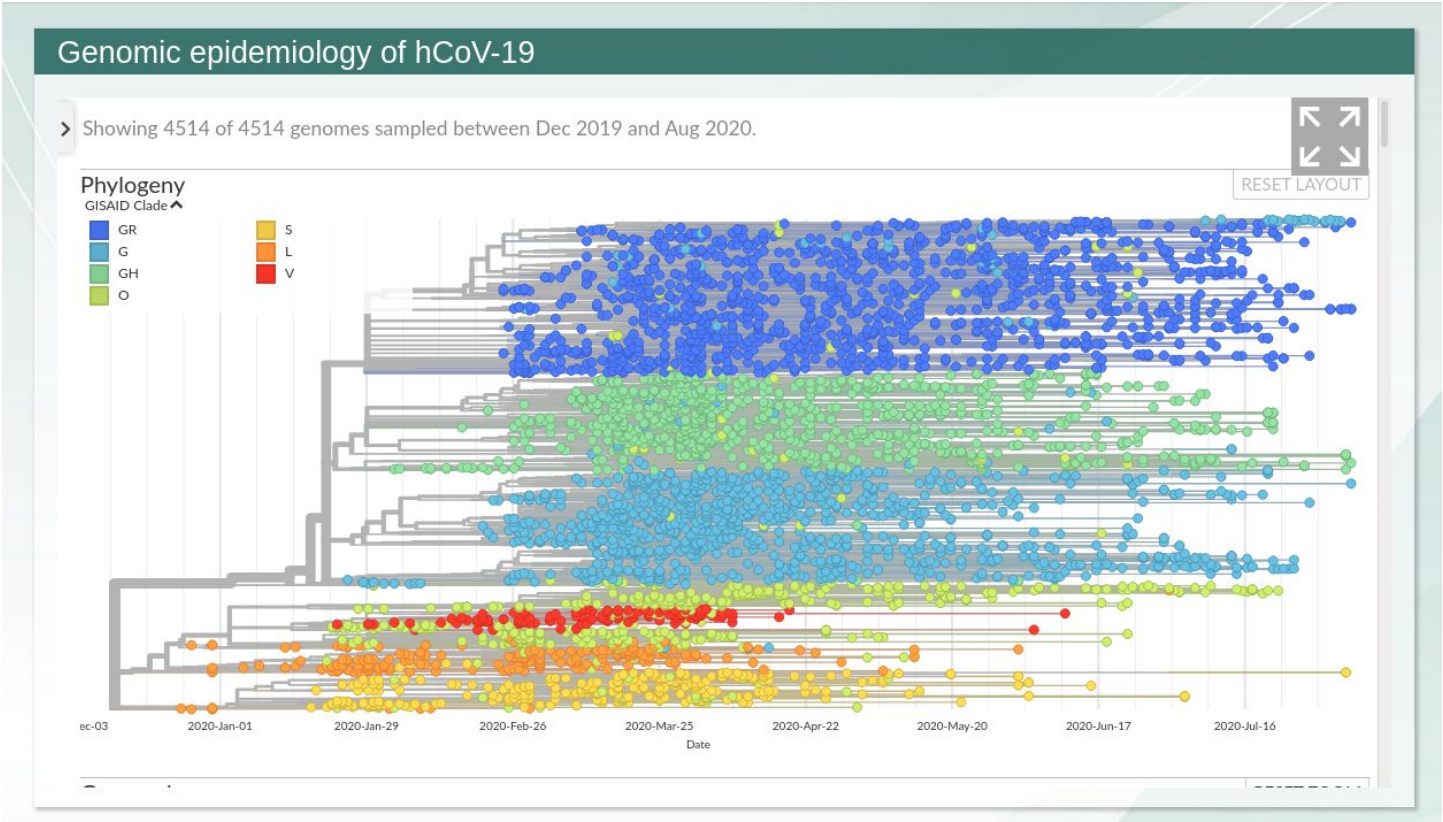
Images from [Bar-On et al. 2020, eLife]

A Vast Database of Information about Genome Variation in COVID-19



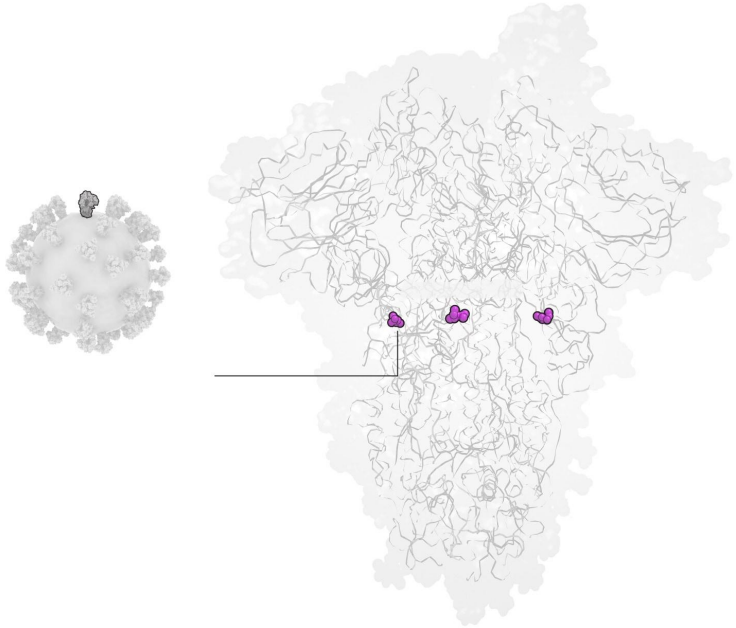
See the [NCBI SARS-Cov-2 Resources page]

Genetic Epidemiology of SARS: Evolution in Action



From GISAID website

Genomic signatures of selection in COVID-19



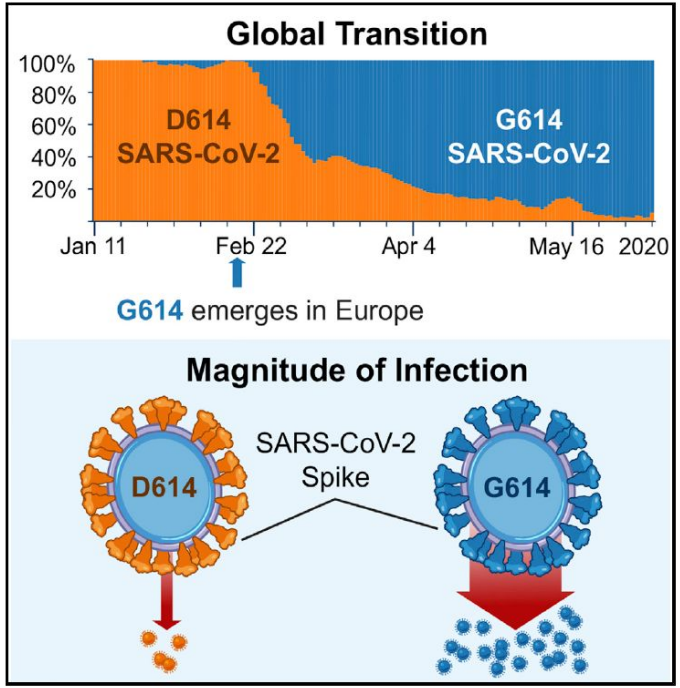


Image from the Washington Post

Image from Korber et al. 2020, Cell

First assignment: Install the Anaconda Python Data Science Toolkit

- Download at: https://www.anaconda.com/products/individual