

# Bio 208FS: Computing on the Genome

Paul Magwene

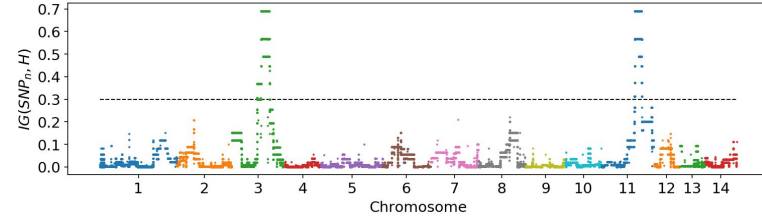
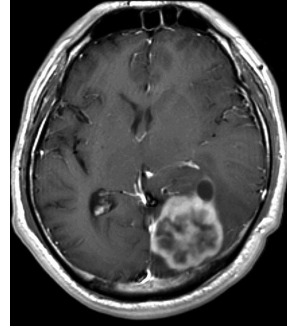
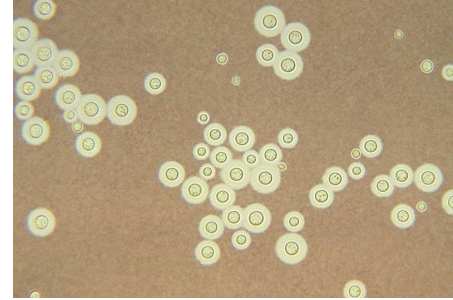
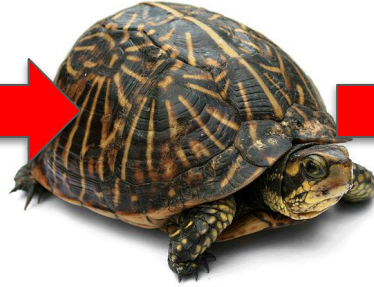
# Goals for today

- 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

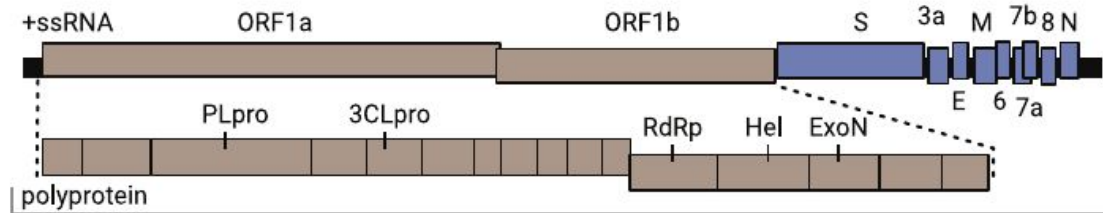
- Genome biology
  - structure, function, variation, evolution
- Genome technology
  - high throughput sequencing
  - sequence data as a proxy for numerous aspects of molecular function
- Genome bioinformatics
  - Assembly
  - Gene calling and annotation
  - Sequence alignment
  - Building phylogenetic 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

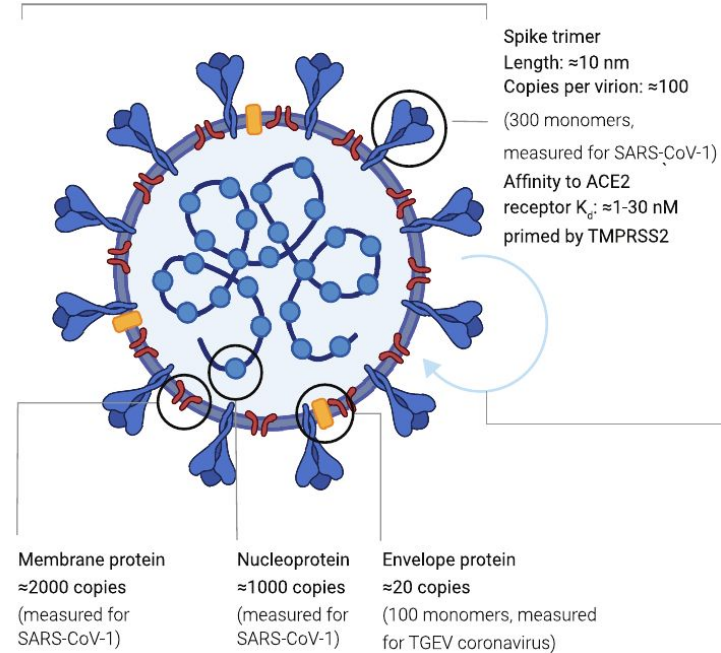
>NC\_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome  
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA  
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC  
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG  
TTGCAGCCGATCATCAGCACATCTAGGTTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC  
CCTGGTTTCAACGAGAAAACACACGTCCAACCTCAGTTTGCCTGTTTTACAGGTTTCGCGACGTGCTCGTAC  
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG  
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTTCATCAAACGTTCCGAT  
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTACGTACGGTC  
GTAGTGGTGAGACACTTGGTGTCCTTGTCCTCATGTGGGCGAAATACCAAGTGGCTTACCGCAAGGTTCT  
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA  
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACCTGGAACACTAAACATAGCAGTGGTG  
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGGCATACTCGCTATGTCGATAACAACCTTCTGTGG  
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG  
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG  
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA  
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA  
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTTCGATCTGTCTATCCAGTTGCGTCAC  
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAACCTTCATGGCA




# From sequence to structure




Length:  $\approx 30\text{kb}$ ;  $\beta$ -coronavirus with 10-14 ORFs (24-27 proteins)



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

**NCBI Virus**  
Sequences for discovery

About Us ▾ Find Data ▾ Help ▾ How to Participate ▾ Submit Sequences ▾ [Contact Us](#)

**Severe acute respiratory syndrome coronavirus 2 data hub**  
Search, retrieve, and analyze SARS-CoV-2 GenBank data.

- [Tree of complete SARS-CoV-2 sequences](#)
- [View a map with geographic distribution of SARS-CoV-2 sequences](#)
- [View SRA data containing coronaviruses](#)

- [Betacoronavirus BLAST®](#)
- [SARS-CoV-2 articles in PubMed](#)
- [NCBI SARS-CoV-2 Resources](#)
- [CDC outbreak information](#)

Refine Results [Reset](#)

Virus [+](#)

SARS-CoV-2, taxid:2697049 [×](#)

Accession [+](#)

Sequence Length [+](#)

Sequence Type [+](#)

RefSeq Genome [+](#)

Nucleotide Completeness [+](#)

Provirus [+](#)

Geographic Region [+](#)

Host [+](#)

Author [+](#)

Isolation Source [+](#)

Collection Date [+](#)

Selected Results: 0

PubMedDownloadAlignBuild Phylogenetic Tree

Expand Table

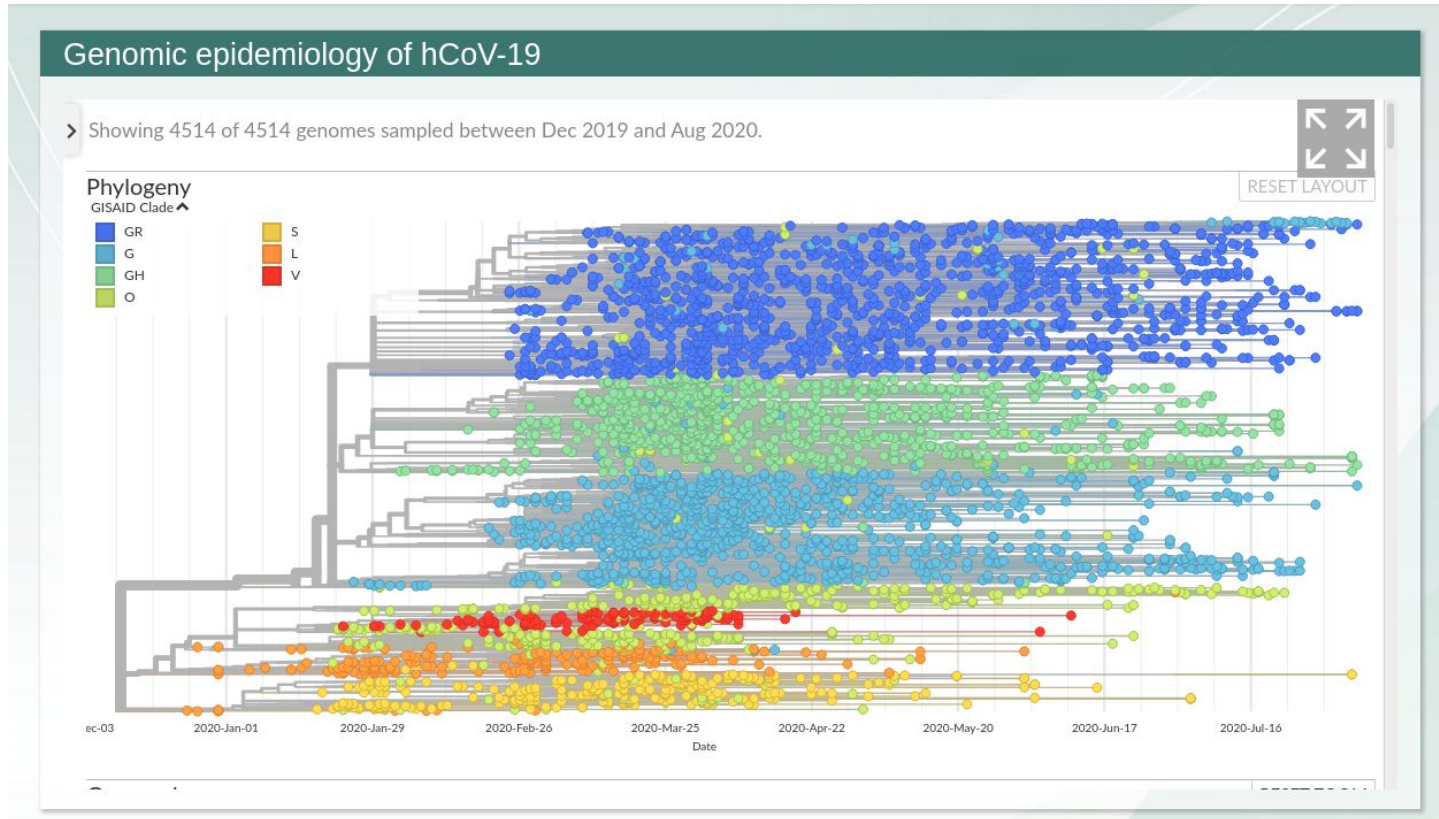
Nucleotide (16,490)Protein (177,065)RefSeq Genome (2)

Select Columns

<input type="checkbox"/>	Accession ▾	Release Date ▾	Species ▾	Length ▾	Geo Location ▾	U.S. State ▾	Host ▾
<input type="checkbox"/>	<a href="#">NC_045512</a> <small>RefSeq</small>	2020-01-13	Severe acute respirato...	29903	China		Homo se
<input type="checkbox"/>	<a href="#">MT893659</a>	2020-08-18	Severe acute respirato...	29840	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893660</a>	2020-08-18	Severe acute respirato...	29768	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893661</a>	2020-08-18	Severe acute respirato...	29851	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893662</a>	2020-08-18	Severe acute respirato...	29821	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893663</a>	2020-08-18	Severe acute respirato...	29818	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893664</a>	2020-08-18	Severe acute respirato...	29855	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893665</a>	2020-08-18	Severe acute respirato...	29815	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893666</a>	2020-08-18	Severe acute respirato...	29841	USA: FL	FL	Homo se
<input type="checkbox"/>	<a href="#">MT893667</a>	2020-08-18	Severe acute respirato...	29853	USA: FL	FL	Homo se

See the [NCBI SARS-Cov-2 Resources page](#)

# Genetic Epidemiology of SARS: Evolution in Action



From [GISAID website](https://gisaid.org/)

# Genomic signatures of selection in COVID-19

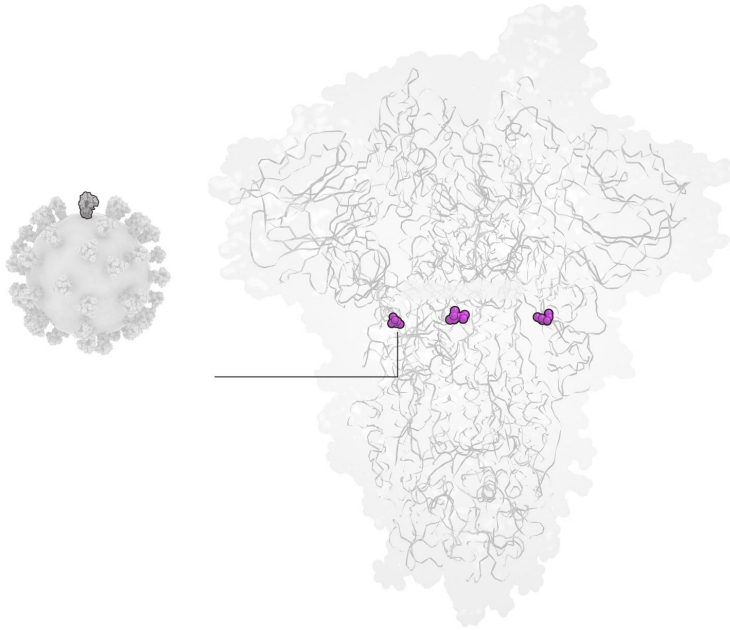


Image from the [Washington Post](#)

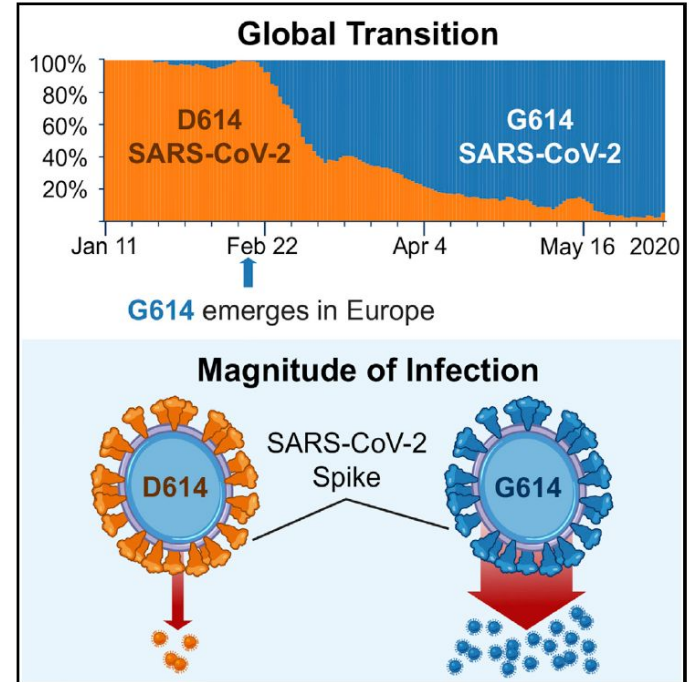


Image from [Korber et al. 2020, Cell](#)

# First assignment: Install the Anaconda Python Data Science Toolkit

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

