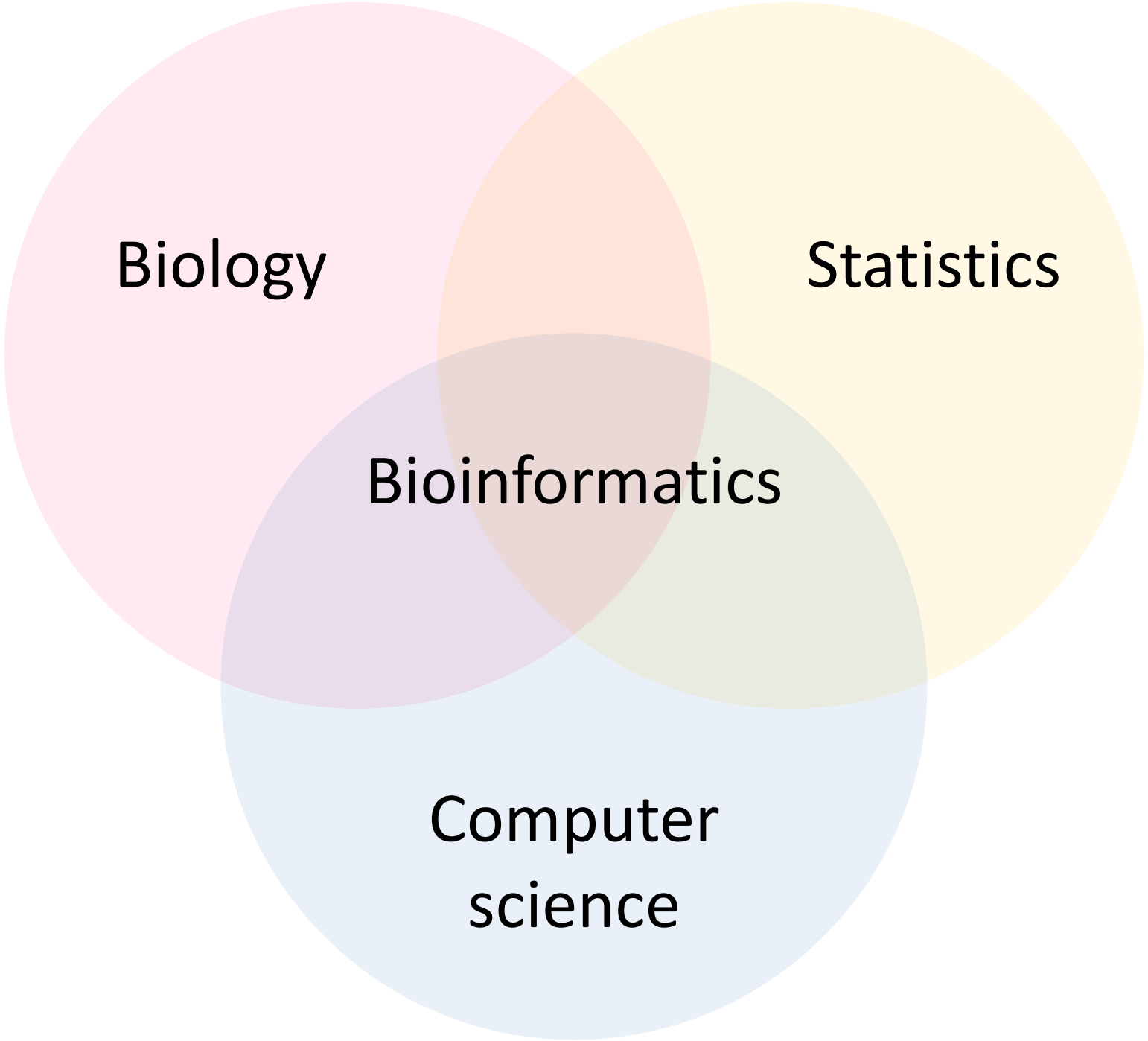
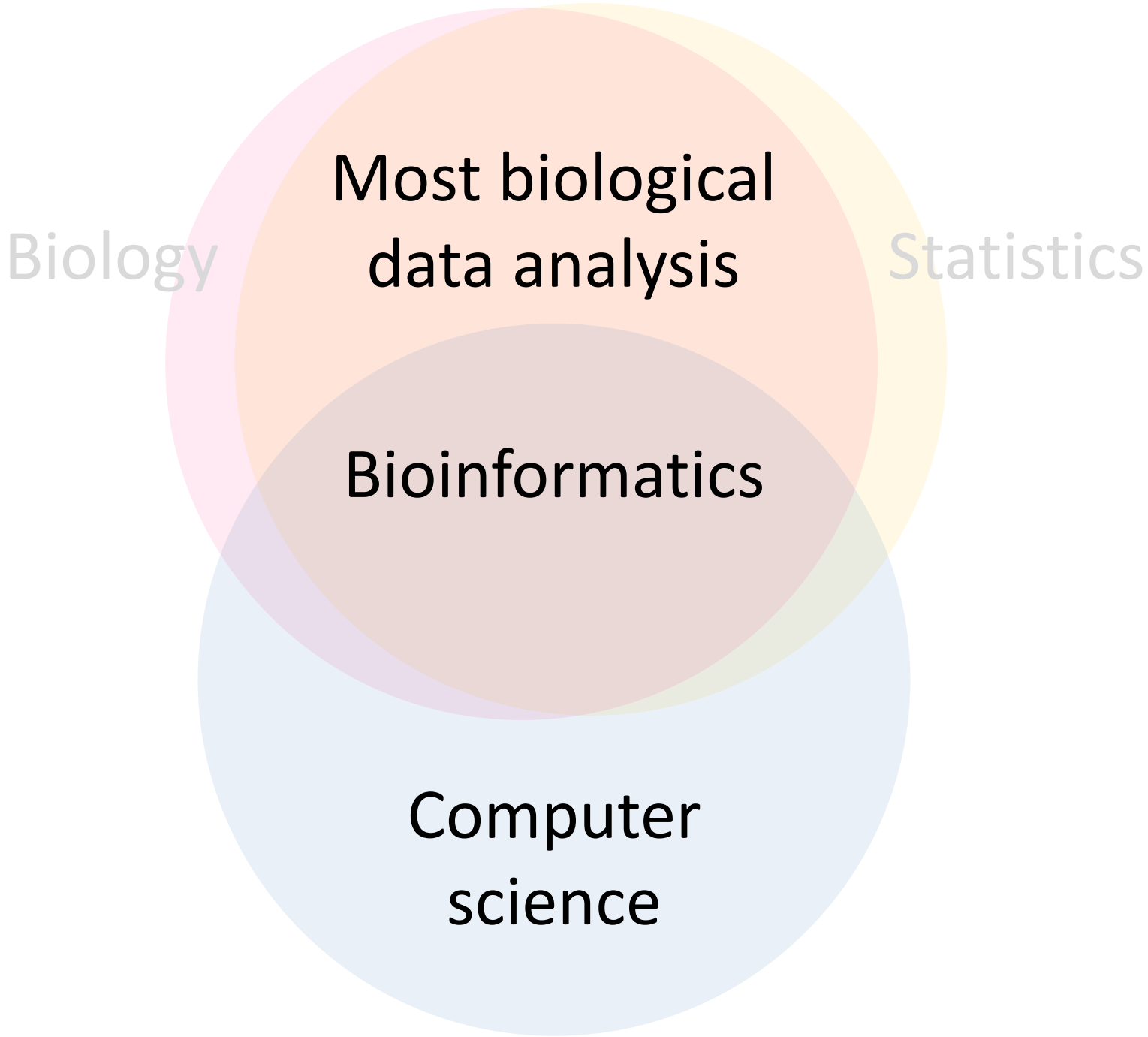
Bioinformatics for RPIP

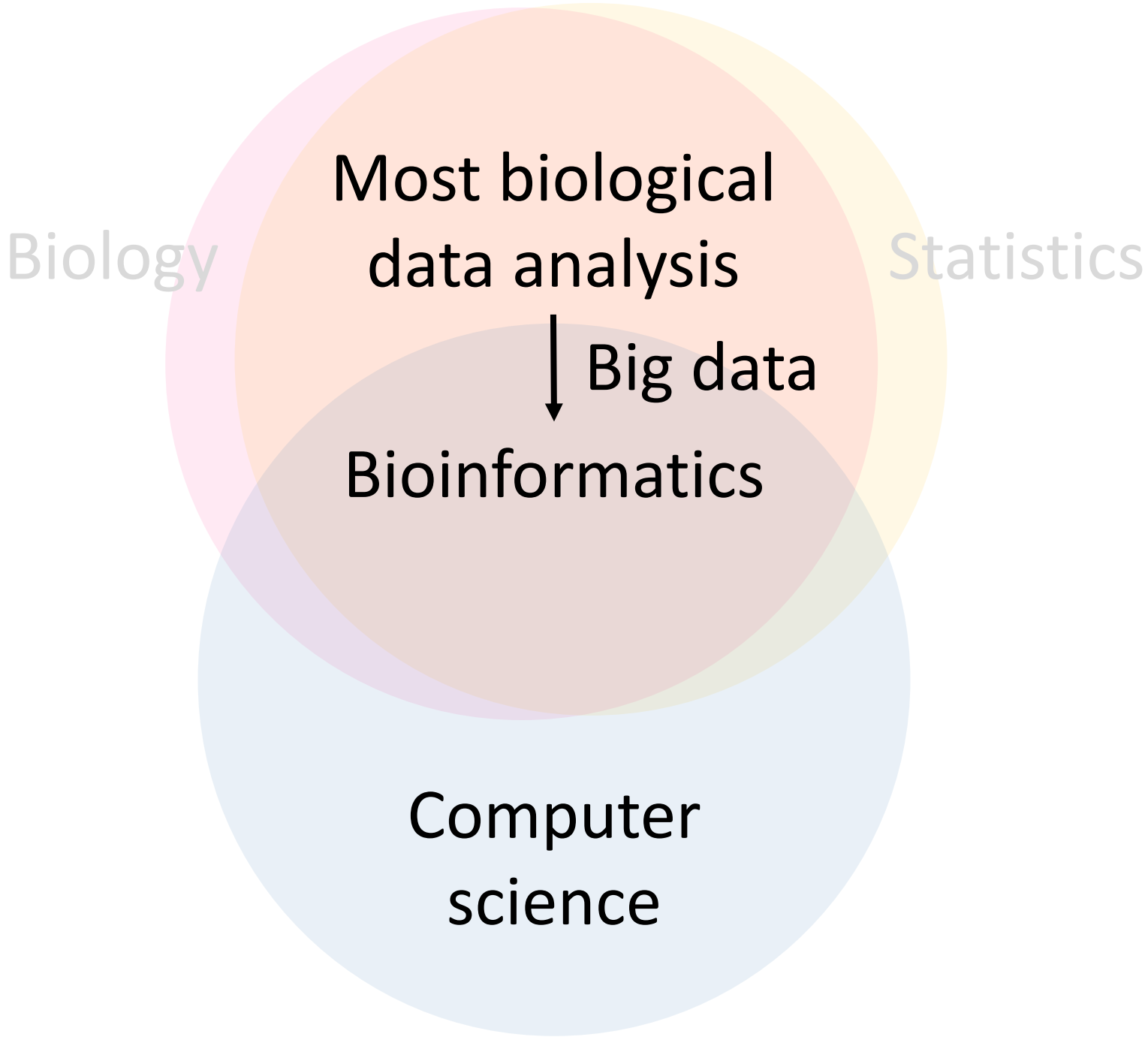
Kim Dill-McFarland

2021.05.18

kadm@uw.edu @kdillmcfarland


Biology
Statistics
Bioinformatics
Computer
science


Biology
Most biological data analysis
Statistics
Bioinformatics
Computer science


Biology
Statistics
Most biological data analysis
Big data
Bioinformatics
Computer science

Big data are defined by

- Volume
- Variety
- Velocity
- Variability

- 100s of samples
- 1000s of sequences per sample
- 100s of new samples per week
  + a need for fast results
- Every sample (person) is unique

Goal

Build a bioinformatic pipeline to efficiently and reproducibly analyze Respiratory Pathogen ID/AMR Panel sequences

Specific aims

- Generate COVID19 consensus genomes for upload to public database
- Track variants of interest and variants of concern
- Track co-infections
- Create phylogenetic trees to
  - Track local variants
  - Identify local outbreaks
- Create reproducible, informative reports

Building a pipeline: Consider the data

Respiratory Pathogen ID/AMR Panel (RPIP)

- Targeted sequencing of
  - All genes in SARS-CoV-2
  - All genes in influenza A/B
  - Identifying genes for 180 bacteria, 50 fungi, 40 viruses
  - Identifying genes for 1200 antimicrobial resistance markers

- Sequences from MiniSeq
  - 7 million (mid)
  - 20 million (rapid)
  - 22 million (high)

Raw data

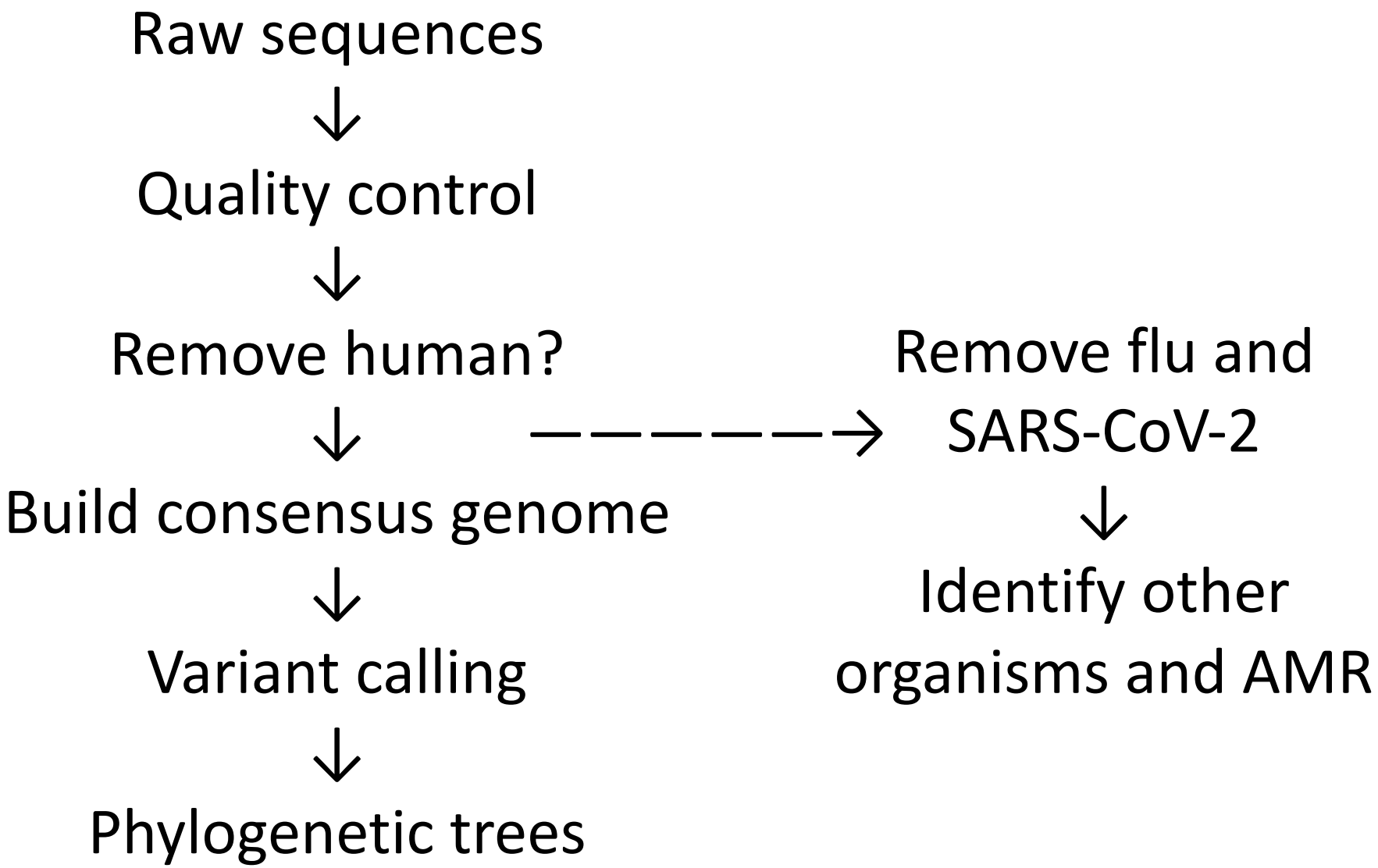
.fastq

@SRR001666.1	@Unique ID
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCAC	Sequence
+SRR001666.1	+Unique ID
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC	Quality scores

.fasta

@SRR001666.1	@Unique ID
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCAC	Sequence

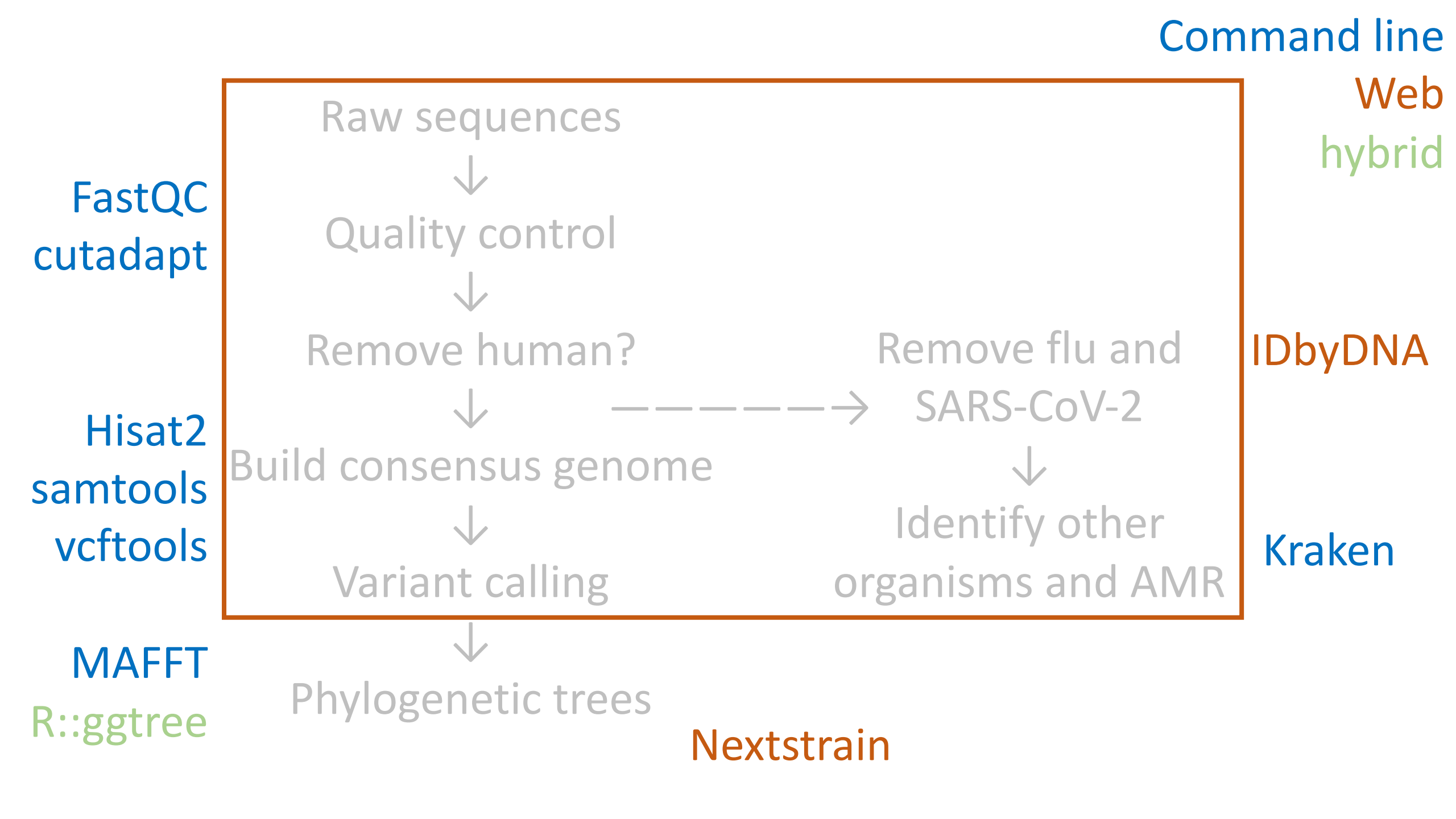
Building a pipeline: Outline a workflow


Raw sequences
↓
Quality control
↓
Remove human?
↓
Build consensus genome
↓
Variant calling
↓
Phylogenetic trees
Remove flu and
SARS-CoV-2
↓
Identify other
organisms and AMR

Building a pipeline: Research software

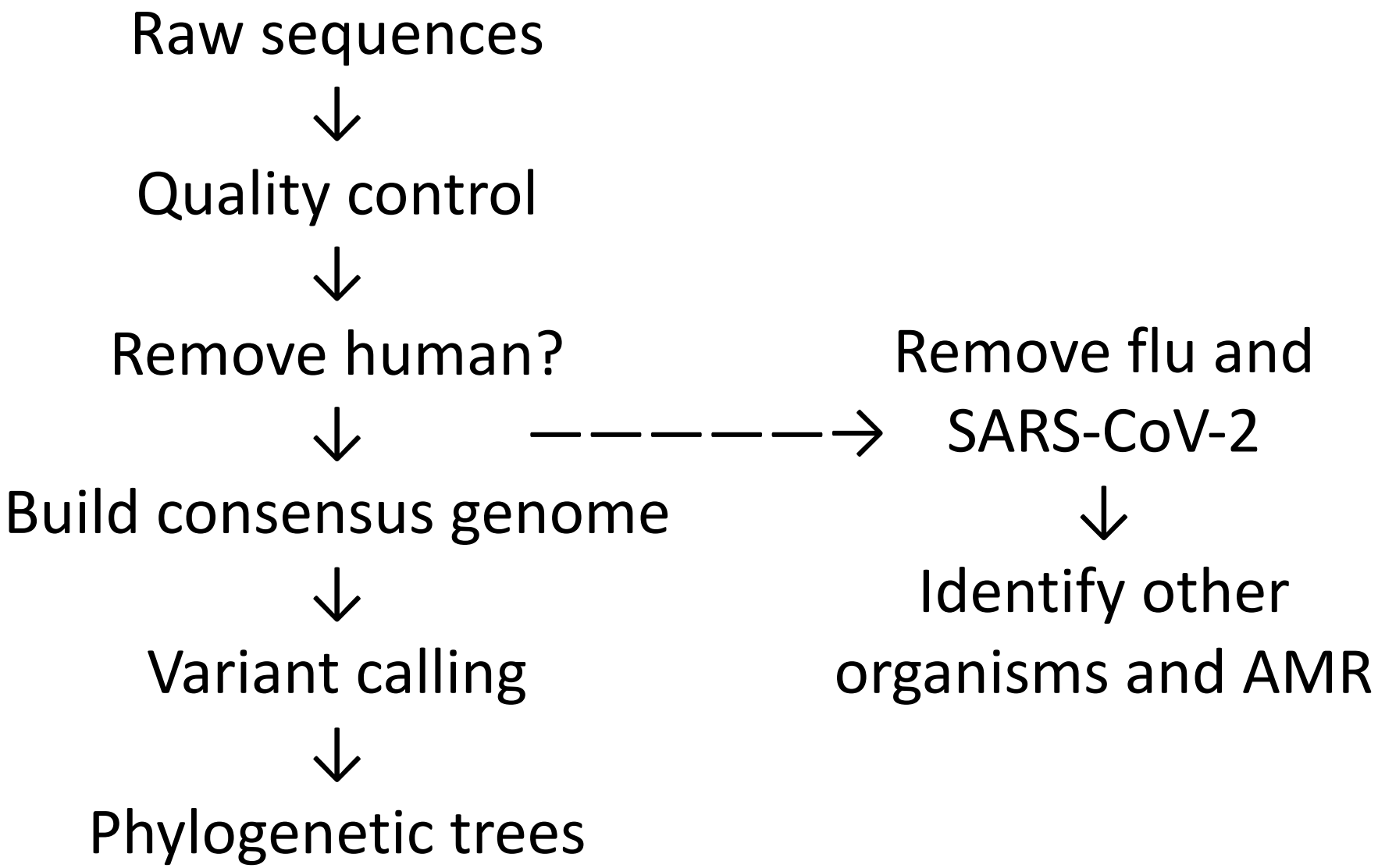
Software considerations

0. Usefulness
   - Does it do what you need it to do?
1. Documentation
   - Manuals, tutorials, GitHub
   - Open-source or black box
2. Cost
   - Ownership, annual membership, cost-per-use
3. Platform
   - Local (Windows, Mac, Linux), cloud, web
   - Command line, graphical user interface (GUI), hybrid

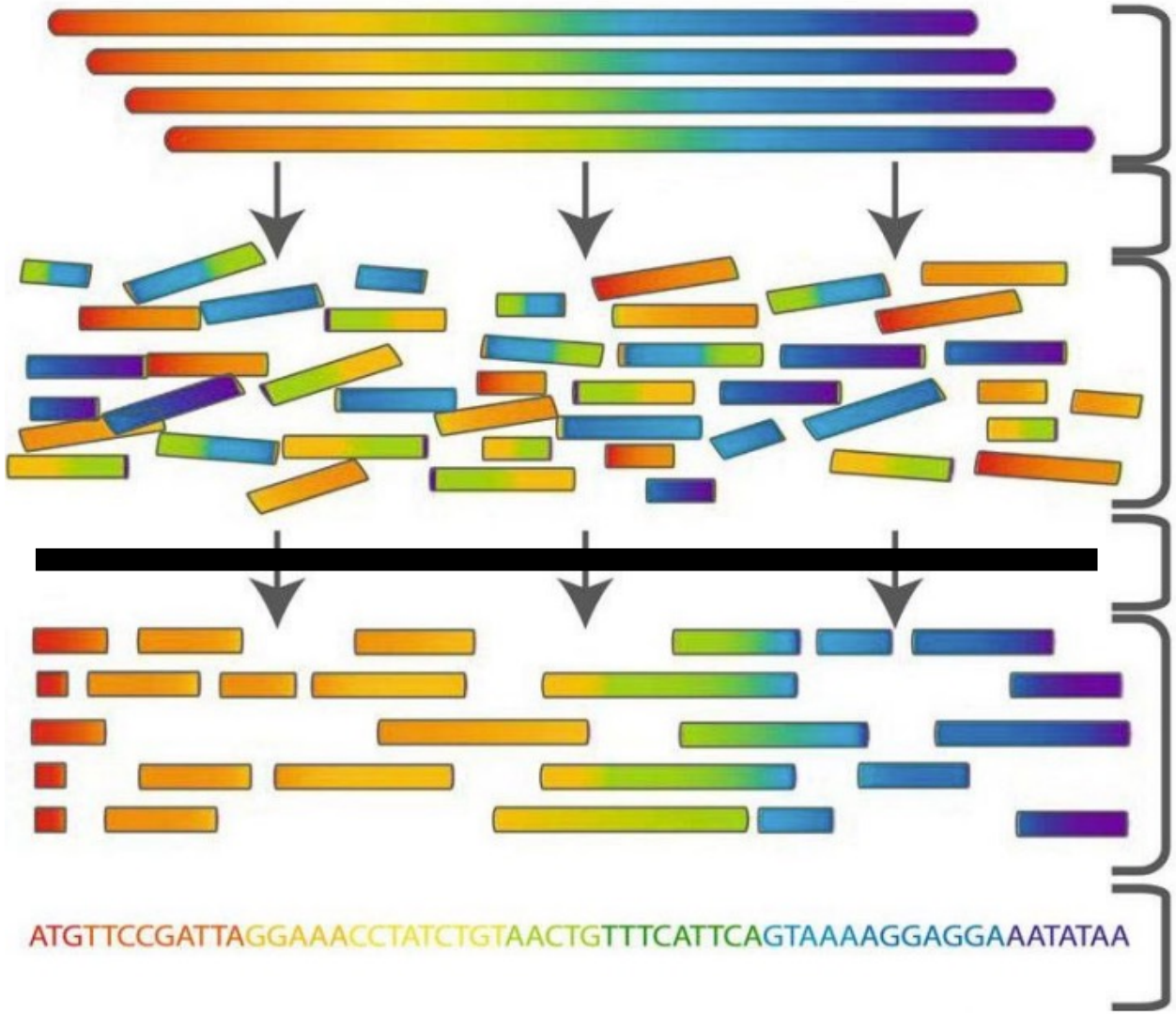

Command line
Web
hybrid
FastQC
cutadapt
Raw sequences
Quality control
Remove human?
Hisat2
samtools
vcftools
Build consensus genome
Variant calling
Remove flu and
SARS-CoV-2
IDbyDNA
Identify other
organisms and AMR
Kraken
MAFFT
R::ggtree
Phylogenetic trees
Nextstrain

Specific aims

- Generate COVID19 consensus genomes for upload to public database
- Track variants of interest and variants of concern
  - How often does it update the global database?
- Track co-infections
- Create phylogenetic trees to
  - Track local variants
  - Identify local outbreaks
- Create reproducible, informative reports
  - How will software updates impact results?


Raw sequences
↓
Quality control
↓
Remove human?
↓
Build consensus genome
↓
Variant calling
↓
Phylogenetic trees
Remove flu and
SARS-CoV-2
↓
Identify other
organisms and AMR

Building a consensus genome



DNA in sample

Shear / amplify

Sequenced reads

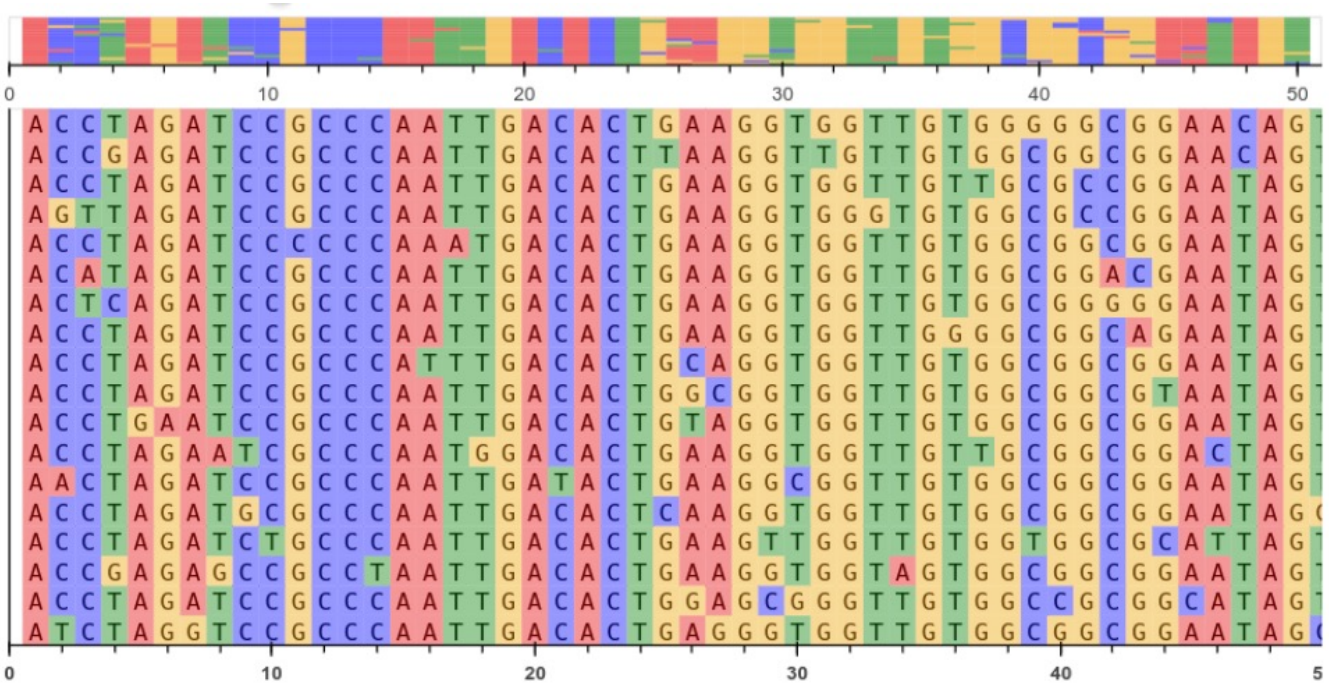
Reference genome

Reconstruct ~~*de novo*~~

Consensus

Torsten Seemann

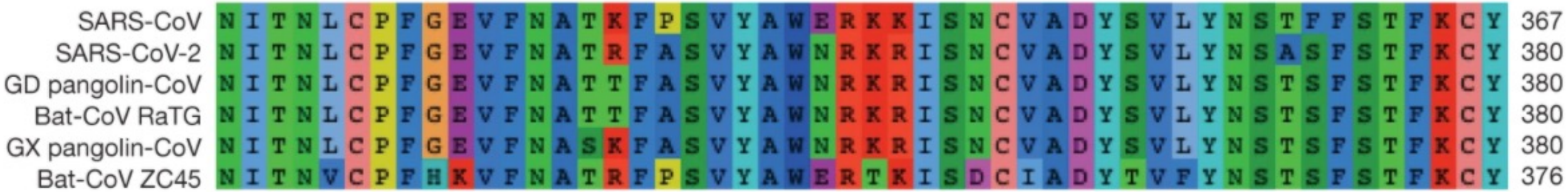
Errors in consensus



- Real or sequencing error? → PhiX control DNA
- If it's real…
  - IUPAC codes for uncertainty → R = A/G, N = ACTG, etc
  - Mixed population

Variant calling

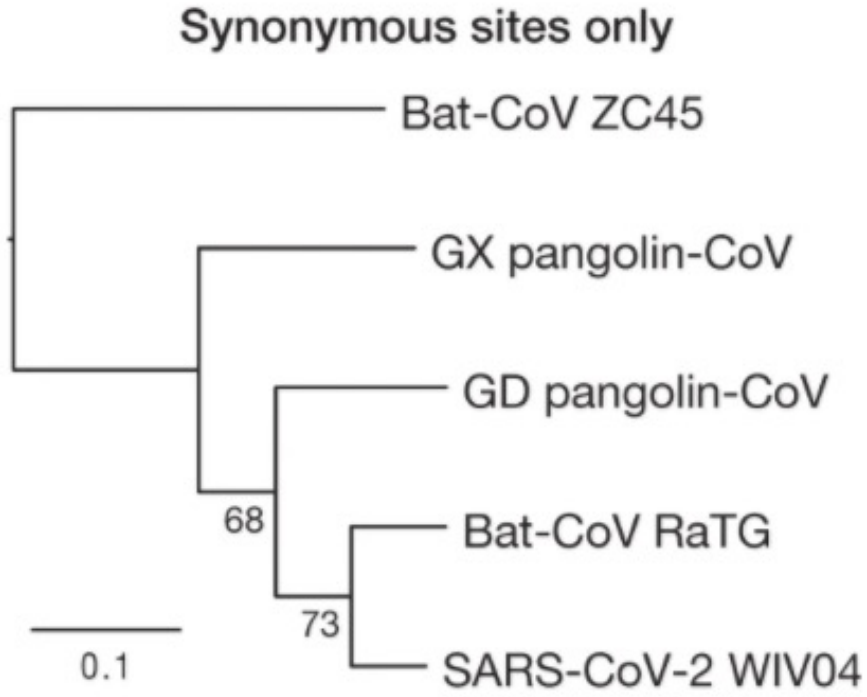
- Compare to database with variants of interest/concern
- Usually translated alignment



Lam et al 2020 doi: 10.1038/s41586-020-2169-0

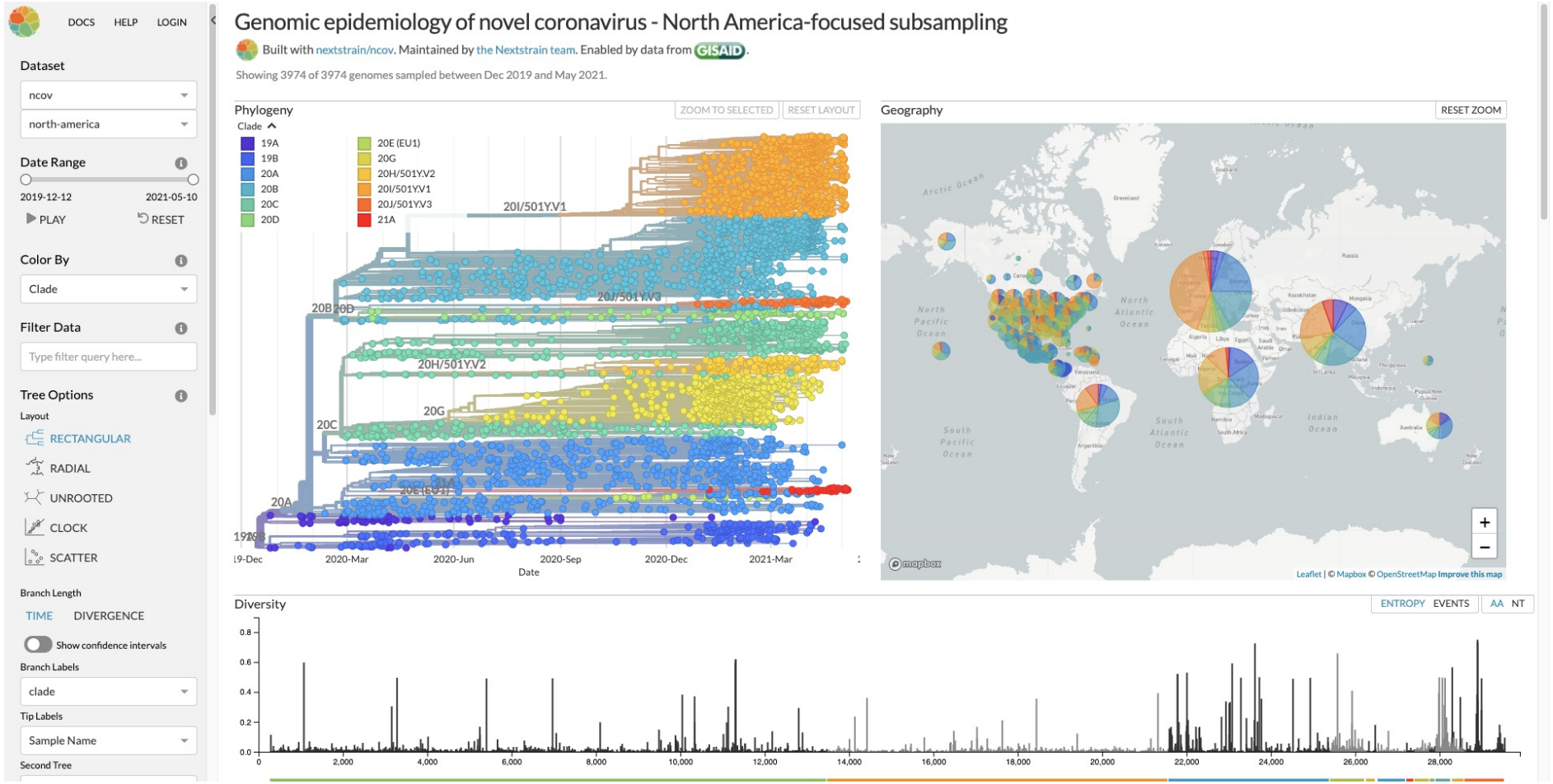
Phylogenetic tree

- Based on translated alignment
- Groups more similar “species”



Lam et al 2020 doi: 10.1038/s41586-020-2169-0

NextStrain



https://nextstrain.org/ncov/north-america

Building a pipeline: Putting it all together

Pipeline wrappers

- Custom executable script
  - Short input in command line like `my_pipeline.sh data.fastq`
  - Download data + databases and run locally
- Web-based workflows (Terra Bio, Shiny apps)
  - Run executable script in the cloud
  - Data + databases can also be stored in the cloud
- Hybrid with IDbyDNA
  - Most steps on web
  - Download consensus sequence
  - Trees run locally, web, or cloud

Next steps

- Decide on desired pipeline format (platform, wrapper, specific software, etc)
- Build and test pipeline on pilot data

Addtl resources

- NextStrain https://nextstrain.org/
- Shiny apps https://shiny.rstudio.com/gallery/
- Terra Bio https://terra.bio/
  - Example workflow usage https://www.youtube.com/watch?v=HObb_J9fPc0&t=604s