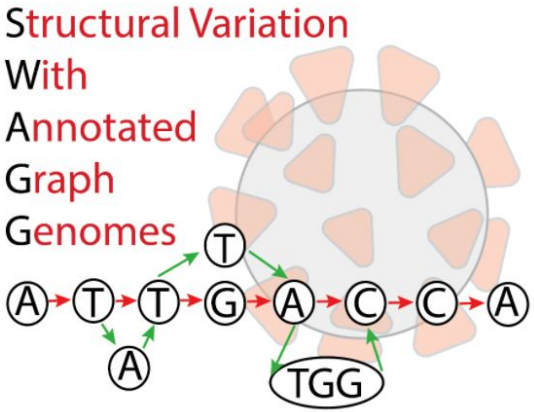
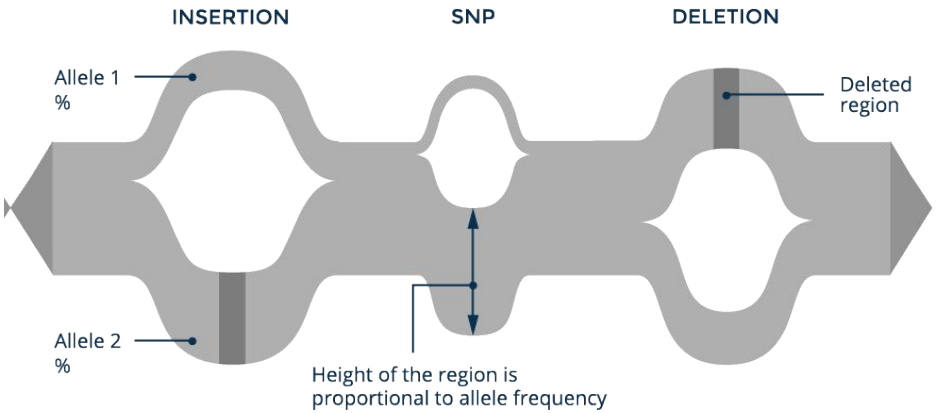
Structural Variation with Annotated Graph Genomes (SWAGG)



Ahmed A., Alejandro G., Daniel C., Dreycey A., Eric D., Fawaz D., Glenn H., **Michael J.**, Sagayamary S., Zeng Q.

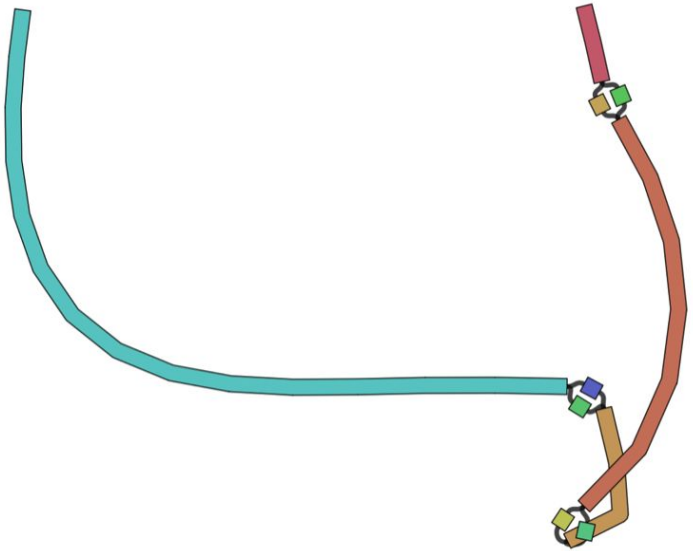
OVERALL GOAL

Genome Graphs

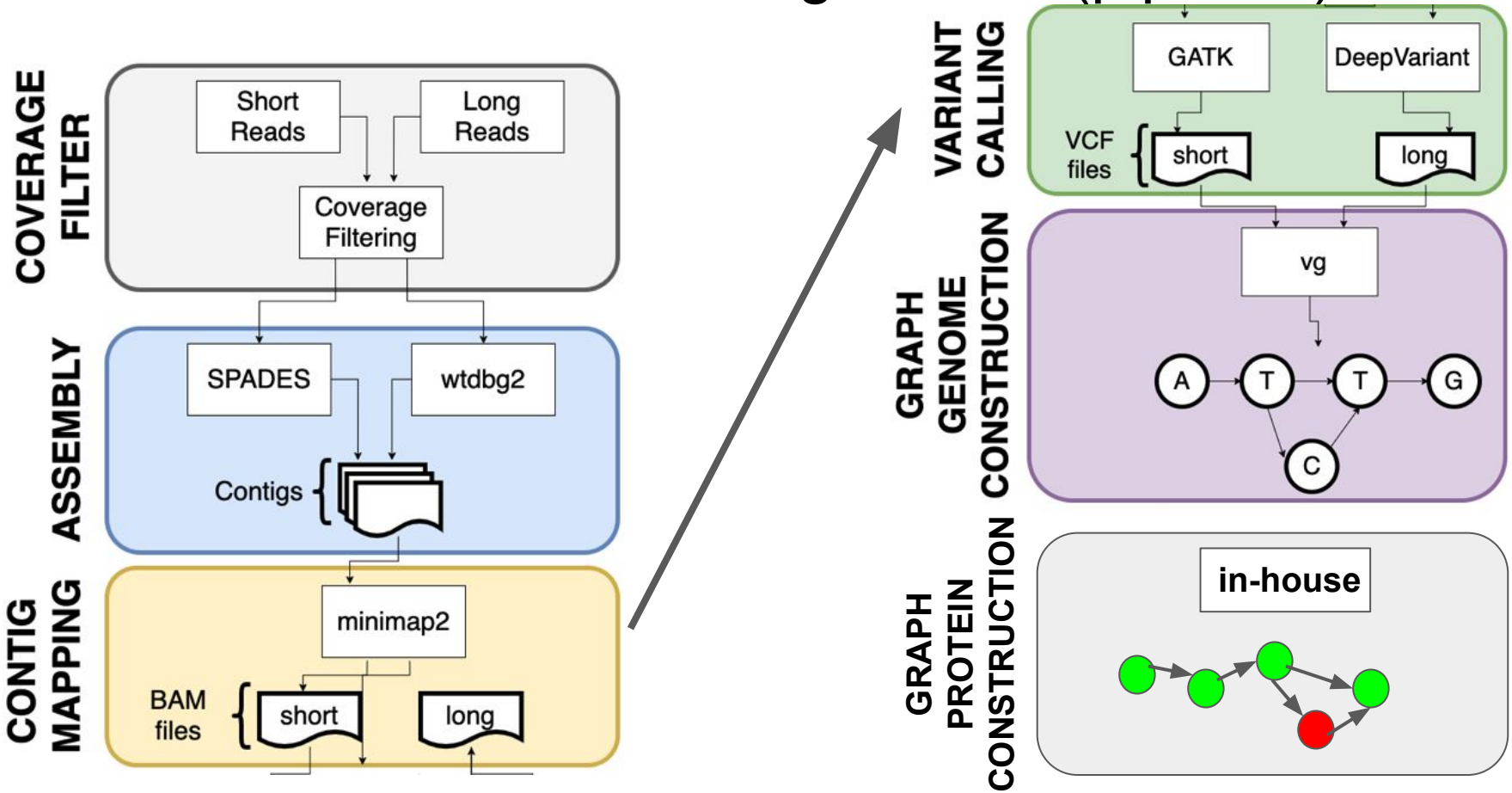


https://www.sevenbridges.com/graf/

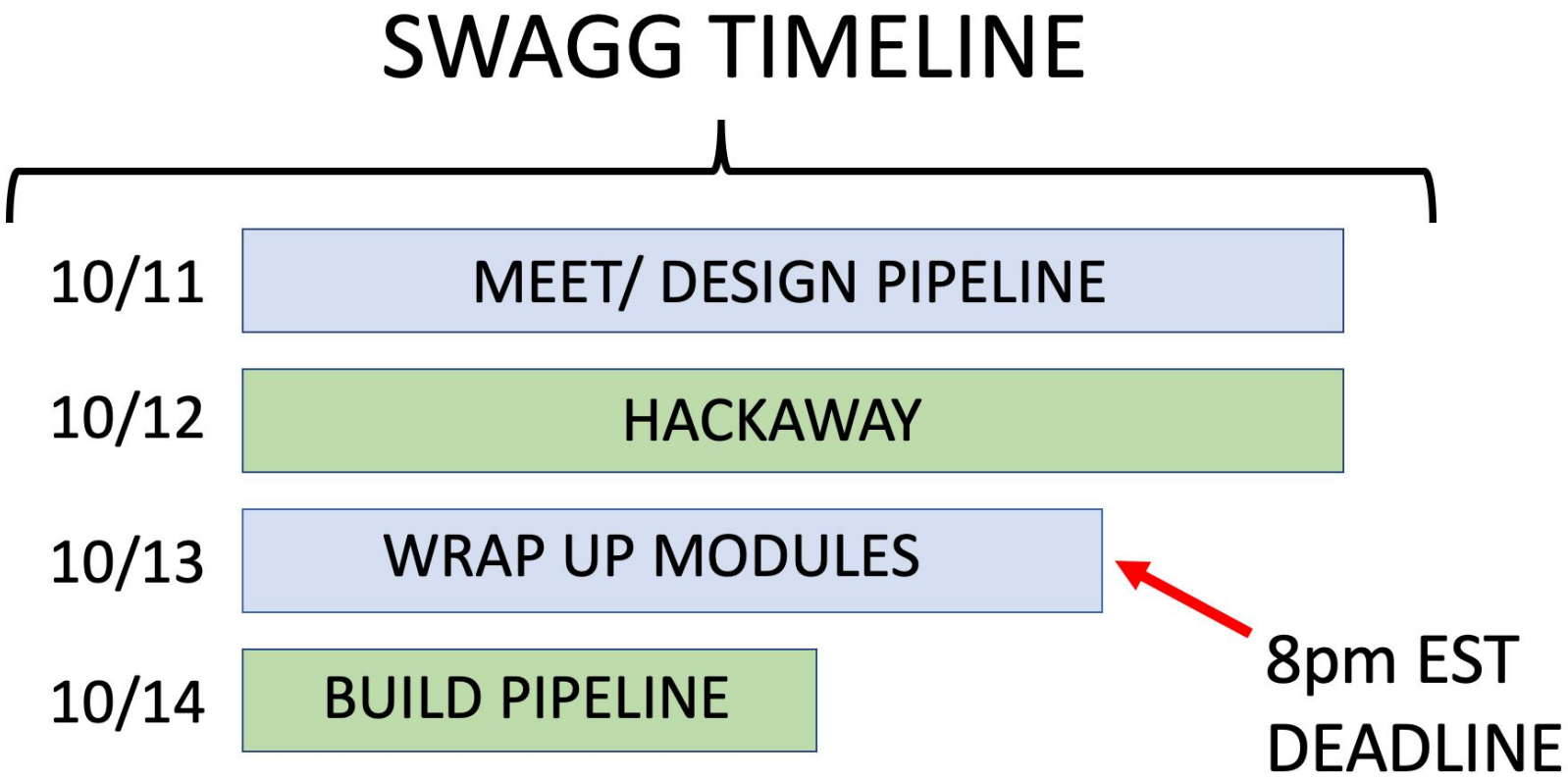
Protein Graphs



OVERALL GOAL - How to get there (pipeline)



Project Management - 3.5 day timeline



Project Management - Distributed Modules

Read Simulator ==> Dreycey

INPUT: Fasta genome files OUTPUT: Simulated NanoPore, Illumina, and PacBio fasta files

Coverage Filter ==> Dreycey

TOOLS: Minimap2 INPUT: Reads and Reference Genome OUTPUT: SAM file, BAM file, and Co

Assembly (short reads) ==> Alej and Fawaz and Dreycey

TOOL(s): Spades URL: https://github.com/ablab/spades INPUT: Fasta short read files - we ca
contigs

Assembly (long reads) ==> Qian

TOOL(s): wtdbg2 - michael posted long read files
URL: https://github.com/ruanjue/wtdbg2 INPUT: Fasta long read files
OUTPUT: Fasta contigs

Mapping (short and long reads) ==> Dreycey

TOOL(s): MiniMap2 URL: https://github.com/lh3/minimap2
INPUT: Reference file and assembled contigs
OUTPUT: BAM, SAM, Coverage

Variant Calling (short reads) ==> Alej and Fawaz and Dreycey

TOOL(s): GATK
URL: https://github.com/broadinstitute/gatk
INPUT: Reference file and BAM file (short read BAM, from contigs from Spades)
OUTPUT: VCF

Structural Variant Calling (short reads) ==> Daniel

TOOL(s): GRIDSS URL: https://github.com/PapenfussLab/gridss INPUT: Reference
from Spades) OUTPUT: VCF, targeted breakpoint assemblies

Variant Calling (long reads) ==> Ahmed

TOOL(s): DeepVariant URL: https://github.com/google/deepvariant
INPUT: Reference file and BAM file (long read BAM, from contigs from wtdgb2) OU

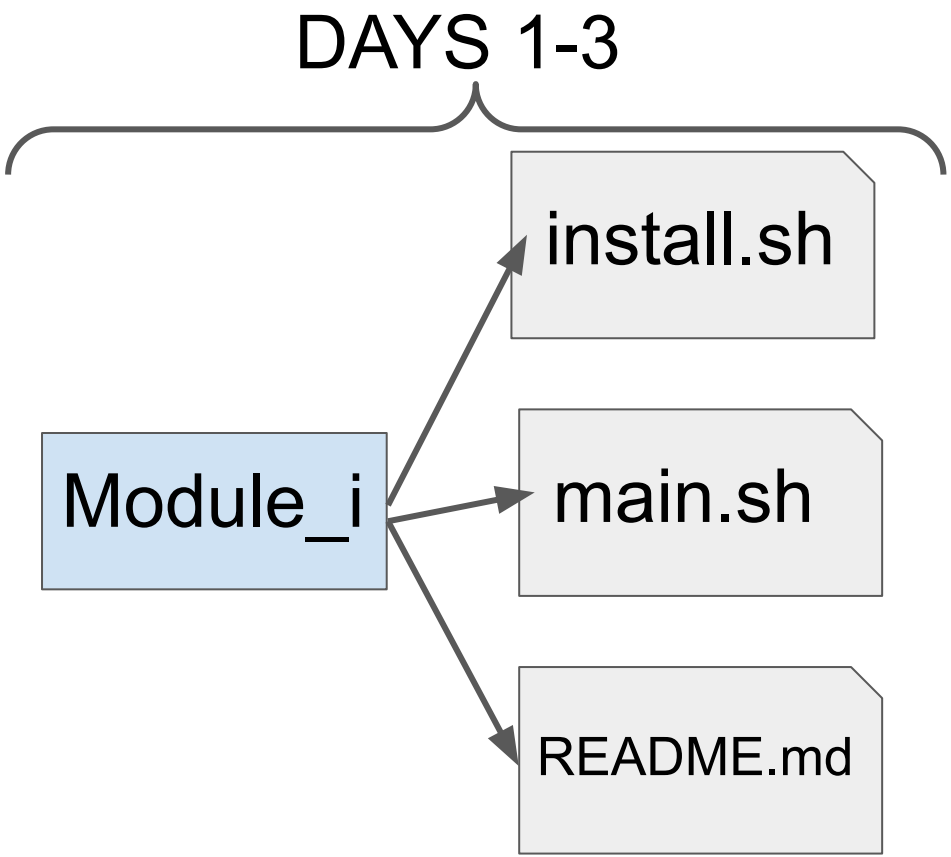
Graph Proteome Construction ==> Fawaz

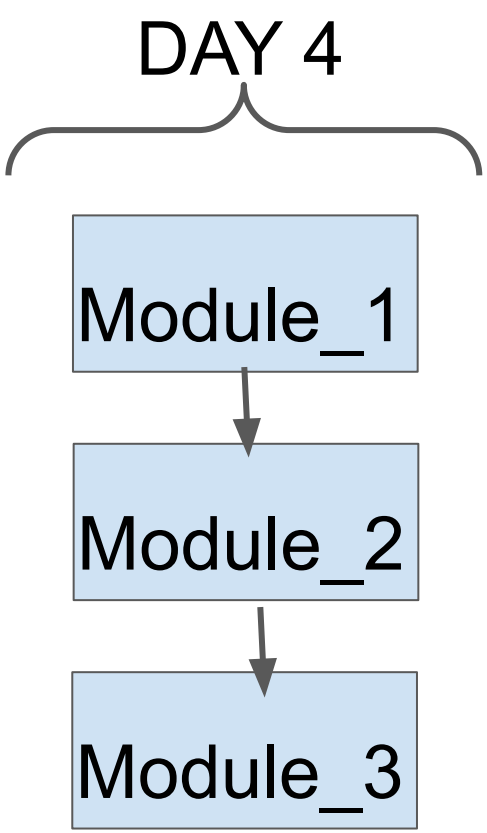
TOOL(s): in-house
URL:
INPUT:F asta consensus genome file and a corresponding VCF
OUTPUT: graph proteins

Graph Genome Construction ==> Sagayamary, Glen

TOOL(s): vg
URL: https://github.com/vgteam/vg
INPUT: Fasta consensus genome file and a corresponding VCF
OUTPUT: graph genome

Project Management - Distributed Modules





Project Management - Results

SWAGG PROGRESS

10/12

- ✓ Read Simulator
- Coverage Filter
- Assembly S
- Assembly L
- Variant S
- Variant L
- Read Mapping
- Protein Graph
- ✓ Genome Graph

10/13-AM

- ✓ Read Simulator
- ✓ Coverage Filter
- ✓ Assembly S
- Assembly L
- Variant S
- ✓ Variant L
- Read Mapping
- ✓ Protein Graph
- ✓ Genome Graph

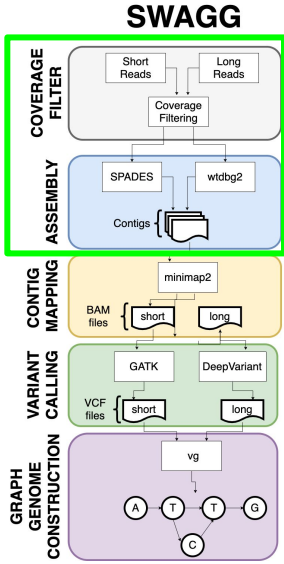
10/13-PM

- ✓ Read Simulator
- ✓ Coverage Filter
- ✓ Assembly S
- ✓ Assembly L
- ✓ Variant S
- ✓ Variant L
- ✓ Read Mapping
- ✓ Protein Graph
- ✓ Genome Graph

10/14

- ✓ Read Simulator
- ✓ Coverage Filter
- ✓ Assembly S
- ✓ Assembly L
- ✓ Variant S
- ✓ Variant L
- ✓ Read Mapping
- ✓ Protein Graph
- ✓ Genome Graph

To assemble, or not?



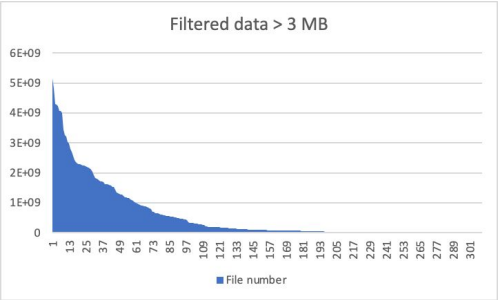
Coverage filter step: Short/long-reads (Sapoval, et al. *bioRxiv* 2020) mapped to SARS-CoV-2 Wuhan-Hu-1

- minimap2
  - handles paired as single, but higher mapping efficiency compared to HISAT2 and Bowtie2

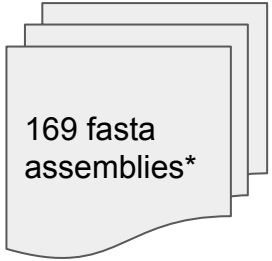
Many patient samples had low SARS-CoV-2 coverage (bottom figure)

Passed files are assembled with:

- SPAdes (right figure), or
- Wtdbg2
  - Accepts FASTA



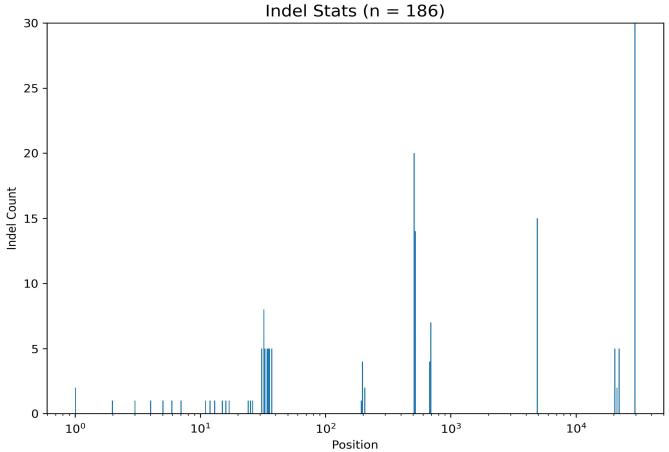
Covid Pangenome Graph



abPOA

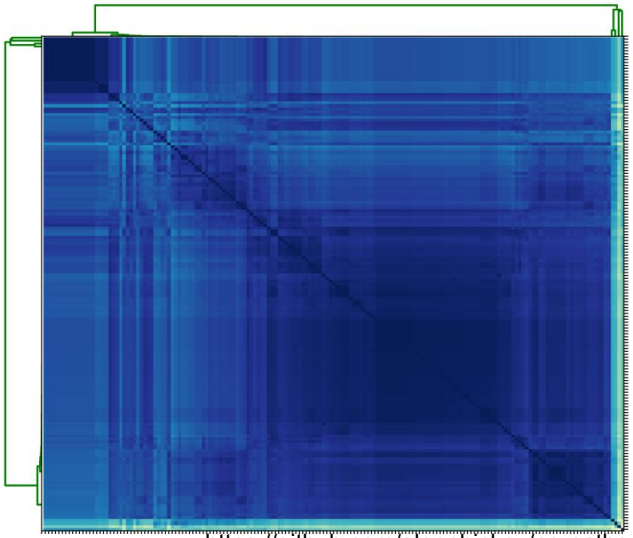
Pangenome Graph

Back to VCF if needed. Whole pipeline can be optionally run on VCF in first place



SV** hotspots

Efficient reference-free clustering

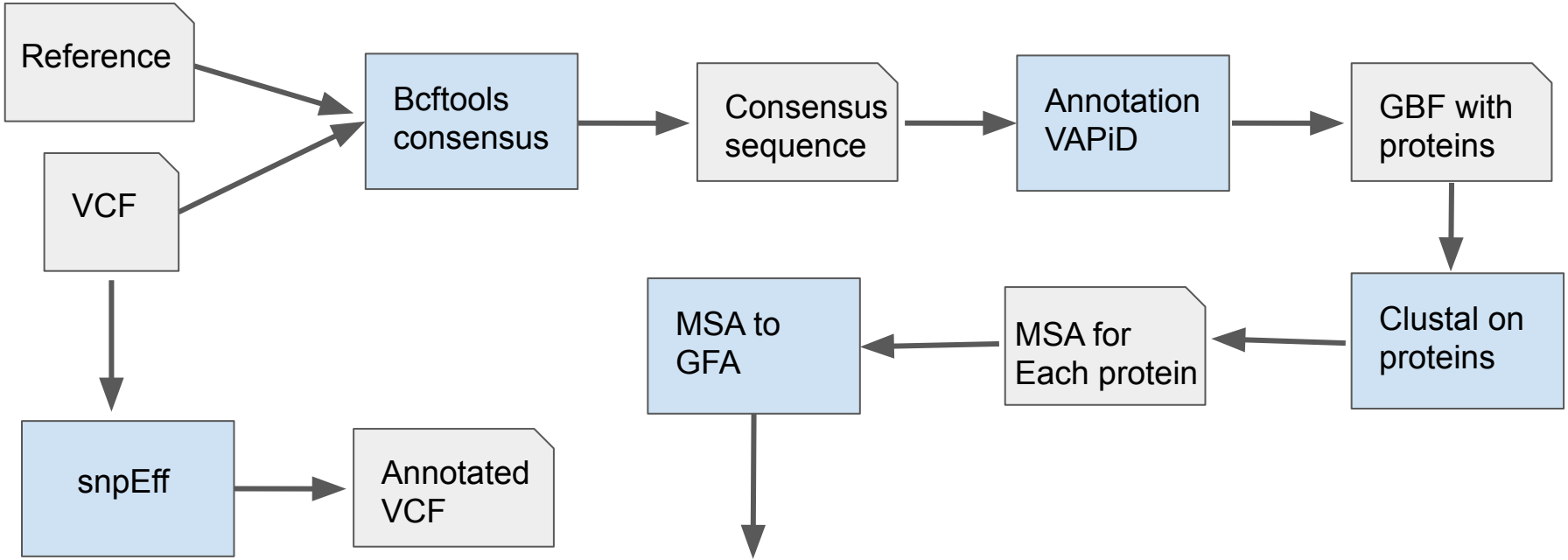


*From https://github.com/hpobio-lab/viral-analysis
**No SV input data yet -- ran on small indels

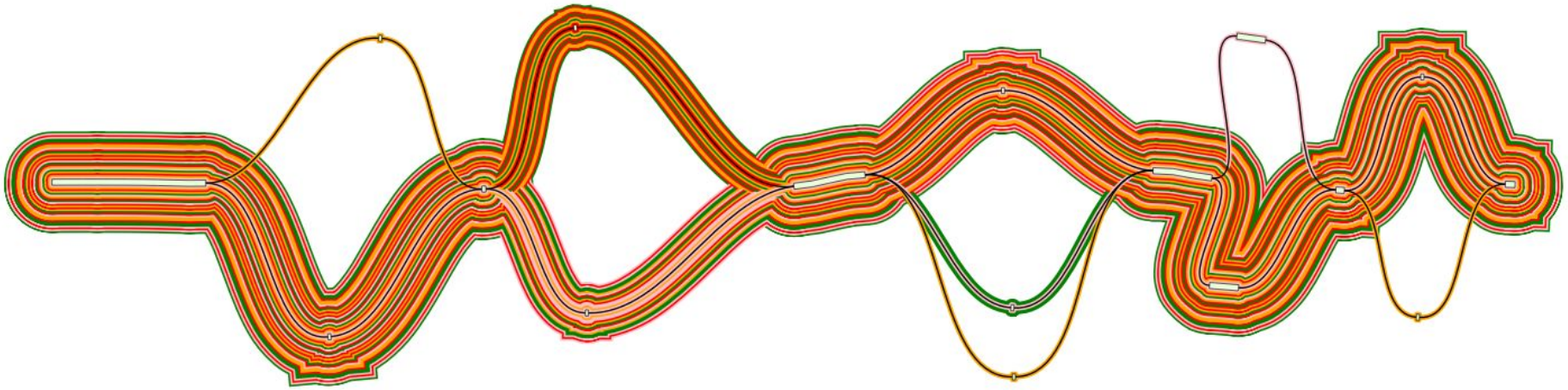
https://github.com/glennhickey/pg-pathcomp

Protein Graphs

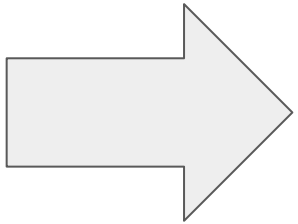
Did the variants introduce new amino acids or stop a stop codon? Can we see that in a graph?

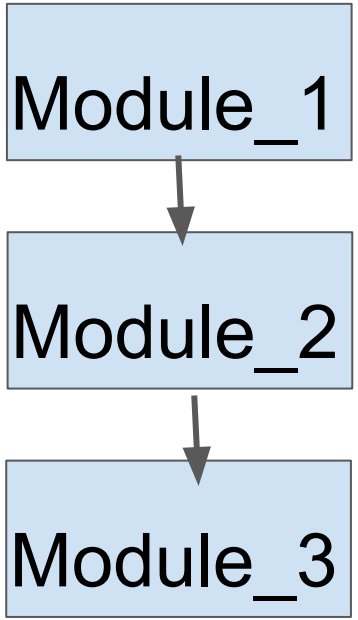


Protein graph with paths representing the original sample. This graph here is Nucleocapsid Phosphoprotein generated from 26 samples



Next Steps - construct the pipeline





El Fin