#team-foo-bar19

This team's effort was focused on the annotation prediction similarity and comparative analysis of all SARS-Cov2 (COVID19) genomes publically available. The intersection of host specific (Human, Bat and Pangolin) genomes as well as a similarity of predicted gene clusters were both processed using open source tools deployable on DataBiology work unit instances. The main goal was to explore the possibility of variant calling on the COVID109 genomes available and exploit potential variants with severe consequences for the virus. The single nucleotide variant frequency and the metrics of said variants will form the main part of the results.

Outline

- Question of regions of interest
- Methodology and input sequences
- Results and potential route
- Future work and requirements

Questions and region of interest

- Genes were clustered to identify conserved regions within and between different host groups.
- Conserved regions were used to find potential hotspots for variant analysis
- Variant calling against human SARS-Cov2 consensus reference
- Comparison between species

Input data

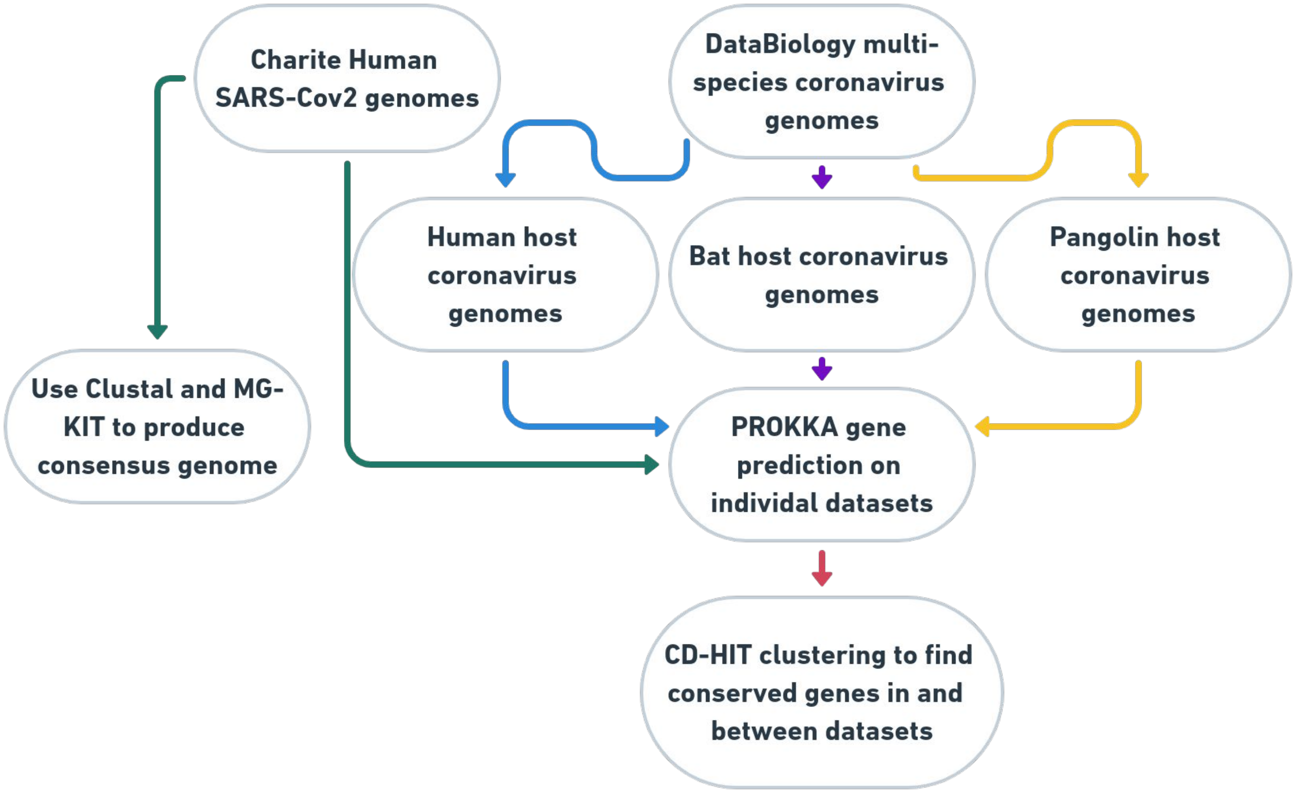
- DataBiology ~ 3800 genomes (multi species, SARS, MERS, COVID19=SARS-Cov2)
  - Human
  - Bat
  - Pangolin
  - Camelids (excl.)
  - Ruminants (excl.)
  - Carnivores (excl.)
- Charite 117 german human SARS-Cov2 genomes
- SRR11508492 (velvet assembly)

Tools

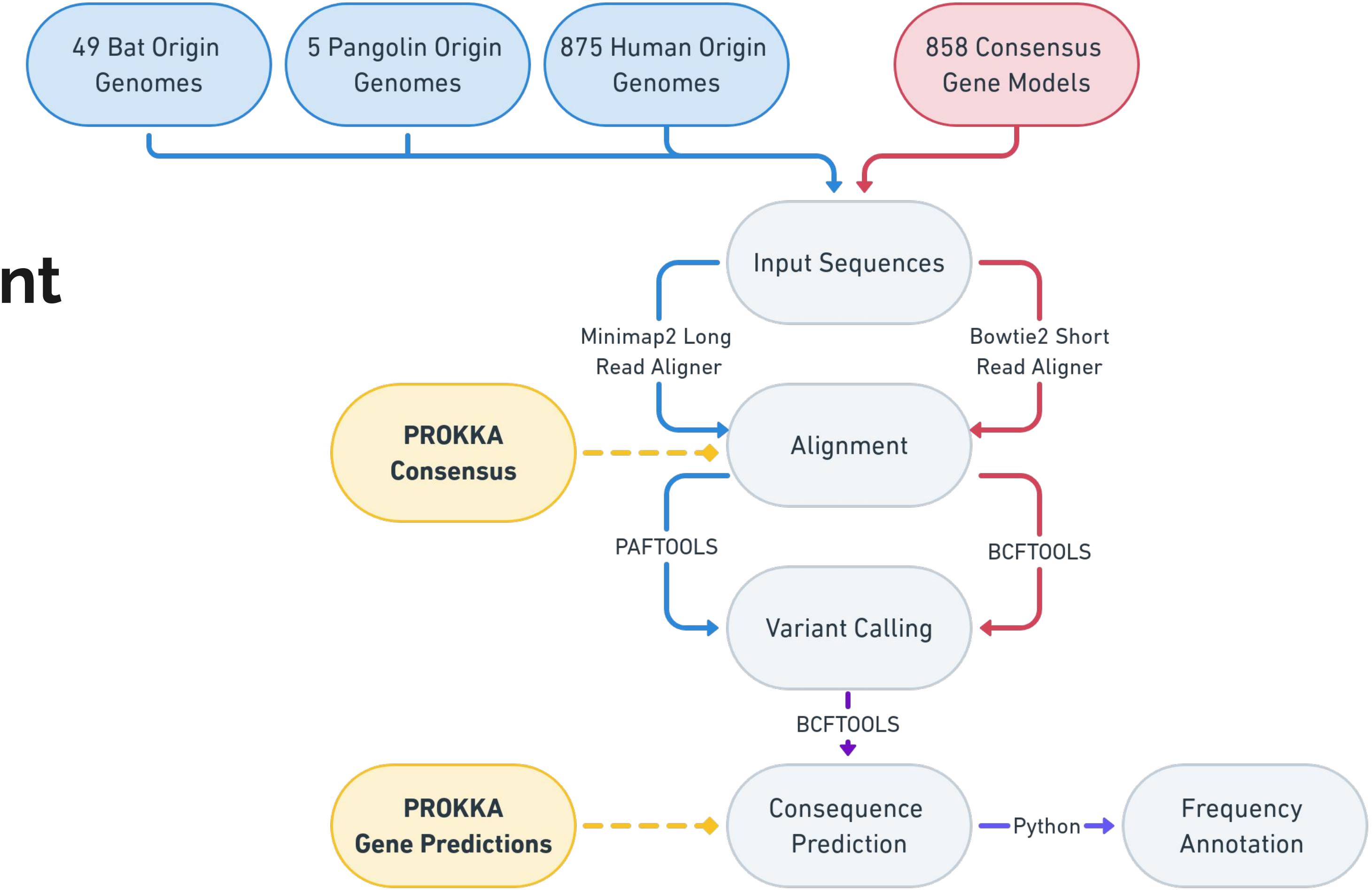
All the included tools are open source and ready to compile on databiology platform:

- **PROKKA** (Nick)
  - Prodigal
  - ncbi-blast+
- **CD-Hit** (Nick)
- **MG-Kit and Clustal** (Nick)
- **Variant calling pipeline** (David)
- **Metrics vs Variant effects** (David, Nick, Mazdak, Barbara)
- **R and RStudio** (Mazdak, Barbara)
- **Python** (David, Nick)
- **Bash scripting** (David, Nick, Mazdak, Barbara)

Gene Clustering & Human Consensus Genome


Charite Human SARS-Cov2 genomes
DataBiology multi-species coronavirus genomes
Human host coronavirus genomes
Bat host coronavirus genomes
Pangolin host coronavirus genomes
Use Clustal and MG-KIT to produce consensus genome
PROKKA gene prediction on individal datasets
CD-HIT clustering to find conserved genes in and between datasets

Genome Alignment



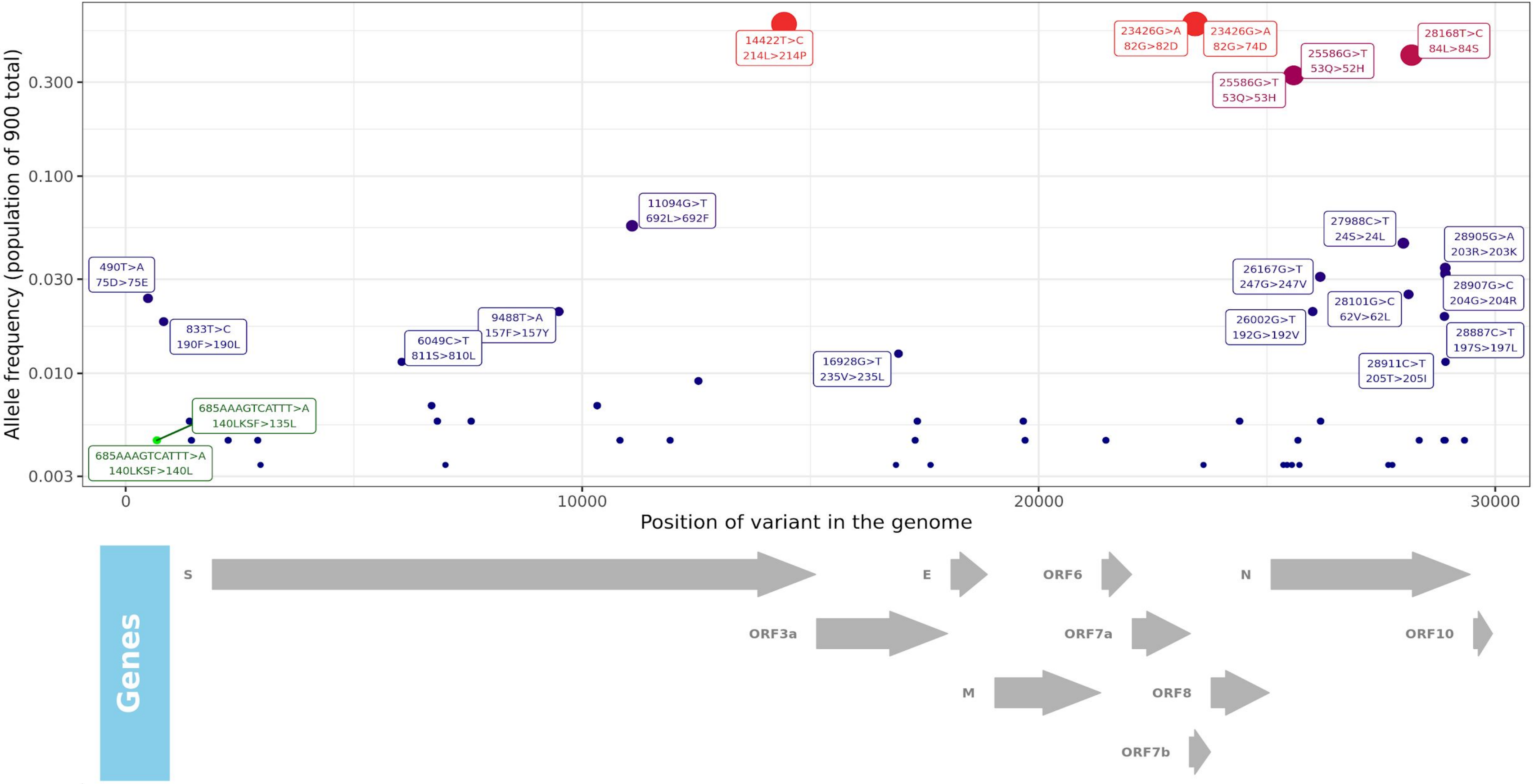
Results

Human SARS-Cov2 genomes (875) and Bat SARS-Cov2 (~49) from 2019
vs
Charite (~117) reference SARS-Cov2 population variant calling results:

- Removing AF < 0.01 and > 0.98
- Removing missense and frame shifts with high allele counts (VAC > 80%)
- Removing synonymous and variant without prediction

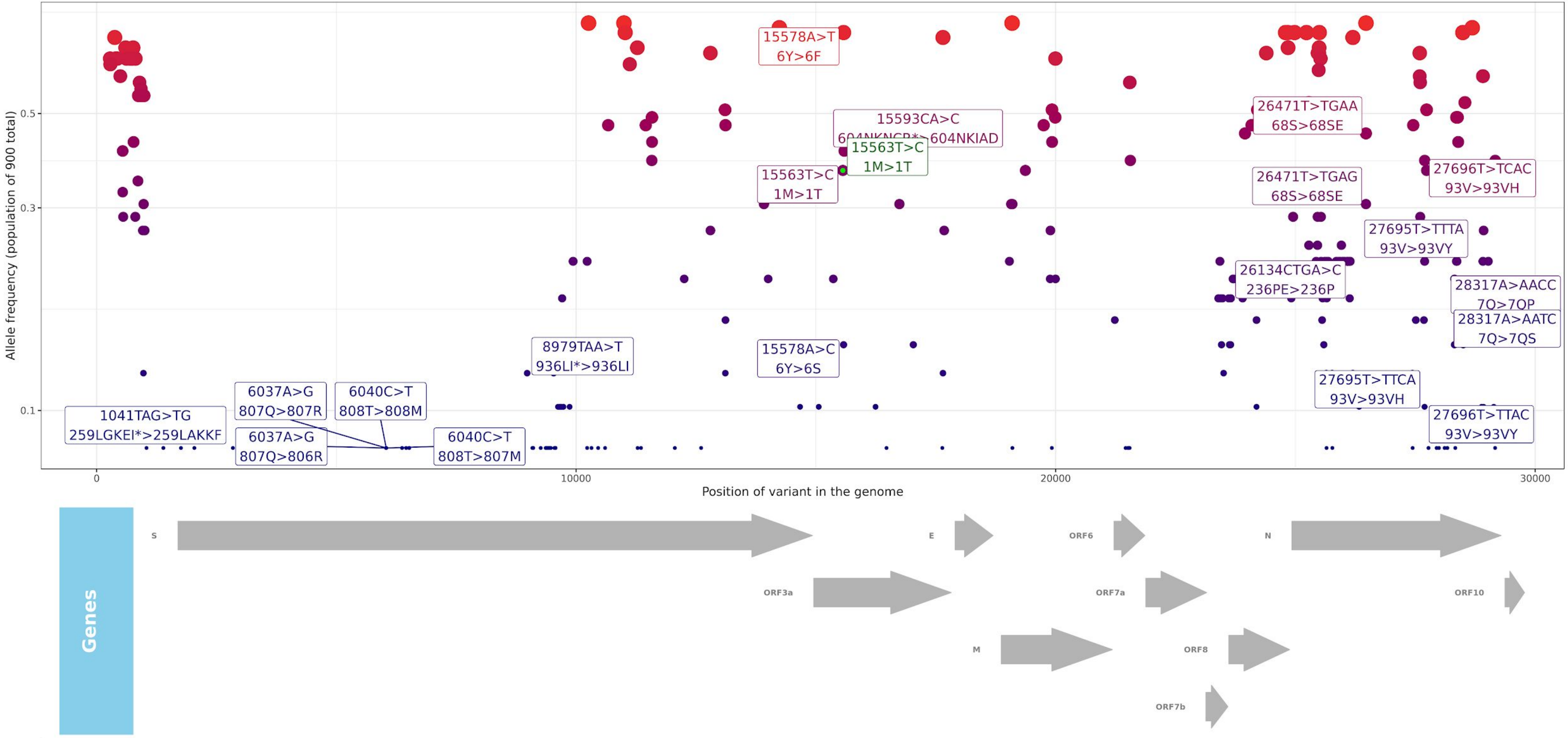
Allele frequency distribution of human SARS-Cov2 variants

Charite 117 population from Germany used as refrence



Allele frequency distribution of bat SARS-Cov2 samples

Charite 117 population from Germany used as refrence



Outline

- Question of regions of interest
- Methodology and input sequences
- Results and potential route
- Future work and requirements
  - SARS-Cov2 pangeome or graph based *de novo* assembly
  - Overlay of the variant effect prediction with PDB structure change
  - Extend the work towards potential vaccine epitope or drug target plans