Evolutionary inference from genomes

BIOLM0030 Genome Biology and Genomics, Session 5

Tom Williams

Session overview

Statistical comparison of genomes to make evolutionary inferences

Variant calling and phylogenetics

Case study: origin of SARS-CoV-2 variants and identification of causative mutations.

1. Variant calling

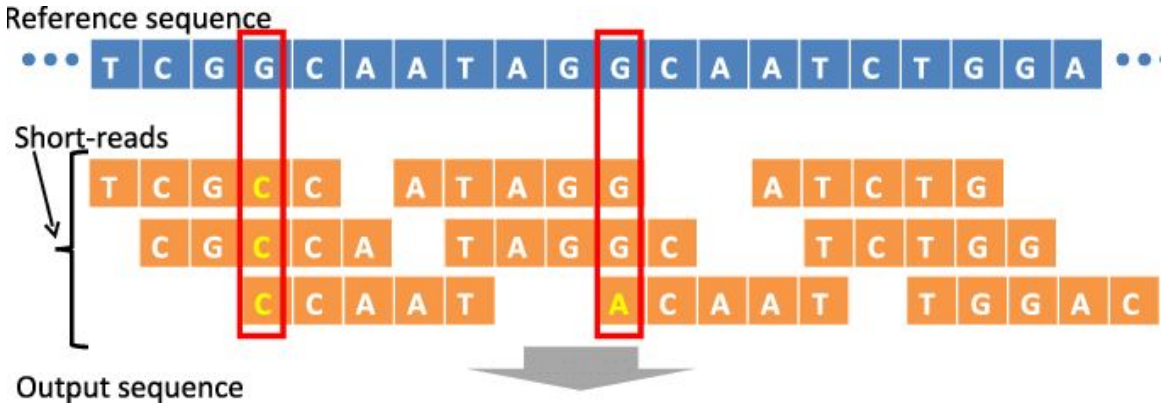
How does a sampled individual differ from a reference genome?

Variant calling (identification) using short reads

Map short reads to a reference genome and identify the differences.

Does not require *de novo* assembly of the new individual, so very useful for organisms with large genomes (or when coverage is low).

Answers the question: how does this individual differ from the reference?



Waidyasooriya et al.

Applications of read mapping/variant calling

Evolutionary: identify variations in a population

Biomedical: identify changes associated with disease (e.g., cancer)

Epidemiological: identify differences among variants of a virus (e.g., SARS-CoV-2)

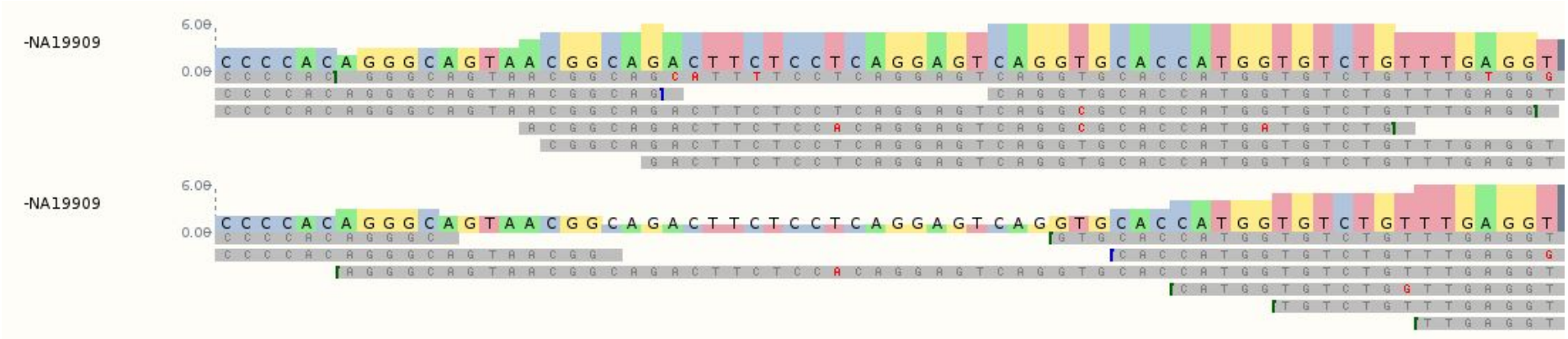
Given a patient sample:

- Which mutations does it contain relative to the Covid-19 reference genome?
- (Phylogenetics) Where does the strain sit in the global diversity of SARS-CoV-2?

Variant calling pipeline: 1. Read mapping

For a set of short reads, where do they align to the reference genome?

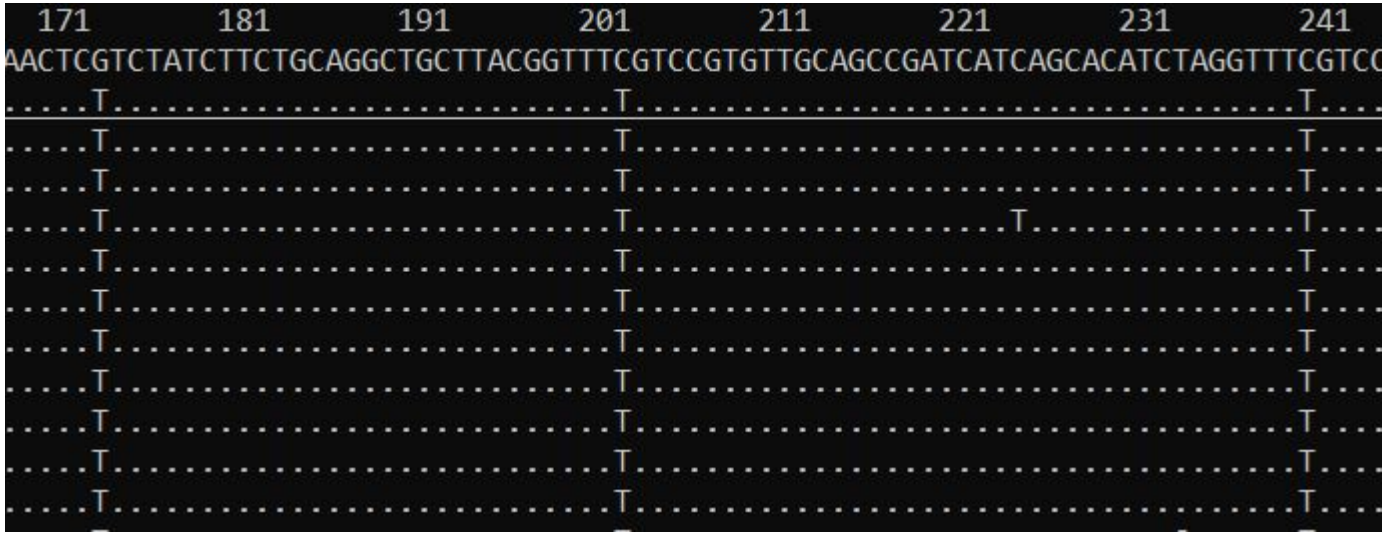
Burrows-Wheeler Transform: algorithm for efficient mapping of short reads to a large genome, with gaps and mismatches. Implemented in BWA and Bowtie(2).



EBI/EMBL training website; Li et al. (2009)

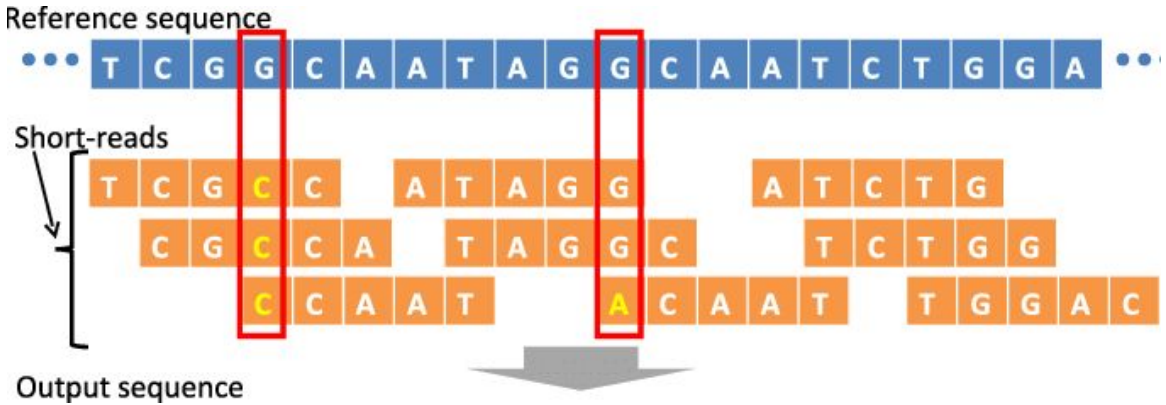
Variant calling: the SAM (Sequence Alignment/Map) format

A format that stores the alignment of reads against a reference.



Variant calling pipeline: 2. Assessing the probability of a variant

Reads contain errors. So we need to compute the probability of a variant given the observed data (reads).



Variant calling pipeline: 2. Assessing the probability of a variant

Reads contain errors. So we need to compute the probability of a variant given the observed data (reads): **P(variant|data)**.

The Phred quality score is a direct meausre of error probability:

Q (Phred) = -10*$\log_{10}$*P*, P = per-base error probability.

If P = 0.001 (1 in 1000 error), Q = 30; P = 0.0001 (1 in 10000), Q = 40

Li et al. (2011)

Variant calling pipeline: 2. Assessing the probability of a variant

Reads contain errors. So we need to compute the probability of a variant given the observed data (reads): **P(variant|data)**.

The Phred quality score is a direct meausre of error probability:

Q (Phred) = -10*$\log_{10}$*P*, P = per-base error probability.

If P = 0.001 (1 in 1000 error), Q = 30; P = 0.0001 (1 in 10000), Q = 40

Therefore, **P (variant | data)** depends on number of reads and the Phred scores.

Variant calling pipeline: 2. Assessing the probability of a variant

The error rate associated with P(variant|data) can also be expressed as a Phred Quality score ($-10\log_{10}P$), with the same interpretation (QUAL = 20: 1% error rate).

Variant calling pipeline 3: The VCF format

A standard-ish format across all variant-calling tools.

Provides genomic position, nature of differences, quality score, and sometimes other information.

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  variant2_sorted.bam
gi|1798174254|ref|NC_045512.2|  174 .   G   T   170 .   DP=21;VDB=2.7996e-05;SGB=-0.692067;MQ0F=0;AC=1;AN=1;DP4=0,0,20,0;MQ=60  GT:PL   1:200,0
gi|1798174254|ref|NC_045512.2|  203 .   C   T   173 .   DP=26;VDB=0.000740843;SGB=-0.692914;MQ0F=0;AC=1;AN=1;DP4=0,0,25,0;MQ=60 GT:PL   1:203,0
gi|1798174254|ref|NC_045512.2|  241 .   C   T   174 .   DP=35;VDB=0.0375731;SGB=-0.693127;MQ0F=0;AC=1;AN=1;DP4=0,0,33,0;MQ=60   GT:PL   1:204,0
gi|1798174254|ref|NC_045512.2|  1059    .   C   T   225 .   DP=69;VDB=0.405778;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,46,18;MQ=60    GT:PL   1:255,0
gi|1798174254|ref|NC_045512.2|  2692    .   A   T   225 .   DP=115;VDB=0.256327;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,46,49;MQ=60   GT:PL   1:255,0
gi|1798174254|ref|NC_045512.2|  3037    .   C   T   228 .   DP=106;VDB=0.932596;SGB=-0.693147;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;AC=1;AN=1;DP4=1,0,46,45;MQ=60 GT:PL   1:255,0
gi|1798174254|ref|NC_045512.2|  5230    .   G   T   225 .   DP=95;VDB=0.639395;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,52,32;MQ=60    GT:PL   1:255,0
gi|1798174254|ref|NC_045512.2|  10323   .   A   G   225 .   DP=130;VDB=0.198947;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,54,54;MQ=60   GT:PL   1:255,0
gi|1798174254|ref|NC_045512.2|  11282   .   AGTTTGTCTGGTTT  AGTTT   99  .   INDEL;IDV=1;IMF=0.00862069;DP=116;VDB=0.997554;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=11,20,48,37;MQ=60  GT:PL   1:255,129
gi|1798174254|ref|NC_045512.2|  11287   .   GTCTGGTTTT  G   228 .   INDEL;IDV=102;IMF=0.902655;DP=113;VDB=0.986883;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=8,13,50,42;MQ=60   GT:PL   1:255,0
gi|1798174254|ref|NC_045512.2|  11288   .   TCTGGTTTTA  T   190 .   INDEL;IDV=1;IMF=0.00952381;DP=105;VDB=0.688774;SGB=-0.693147;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=7,14,47,37;MQ=60   GT:PL   1:255,38
gi|1798174254|ref|NC_045512.2|  11296   .   T   G   26.4242 .   DP=5;VDB=0.02;SGB=-0.453602;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=60  GT:PL   1:56,0
```

Variant calling pipeline 4: Predicting the effect of variants

Many variants have no impact on phenotype.

To predict the phenotypic effect, if any, of a variant, we need information about the genomic context: is it within a gene? Will it change the protein product?

Variant calling pipeline 4: Predicting the effect of variants

Many variants have no impact on phenotype.

To predict the phenotypic effect, if any, of a variant, we need information about the genomic context: is it within a gene? Will it change the protein product?

We need to compare the coordinates of the identified variants with our genome annotation. **snpEff** is a tool for doing this.

Variant calling pipeline 4: Predicting the effect of variants

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	1.471%
disruptive_inframe_deletion	9	2.206%
downstream_gene_variant	165	40.441%
intergenic_region	3	0.735%
missense_variant	32	7.843%
stop_gained	4	0.98%
synonymous_variant	11	2.696%
upstream_gene_variant	178	43.627%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	165	40.842%
EXON	58	14.356%
INTERGENIC	3	0.743%
UPSTREAM	178	44.059%

Variant calling pipeline 4: Predicting the effect of variants

Number of effects by type and region

Type

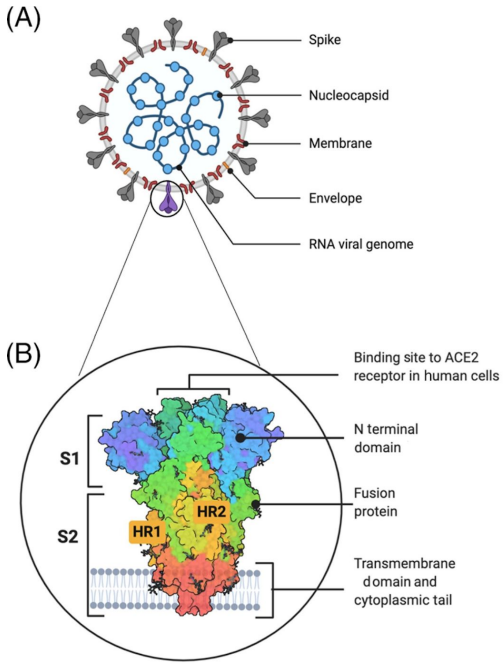
Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	1.471%
disruptive_inframe_deletion	9	2.206%
downstream_gene_variant	165	40.441%
intergenic_region	3	0.735%
missense_variant	32	7.843%
stop_gained	4	0.98%
synonymous_variant	11	2.696%
upstream_gene_variant	178	43.627%

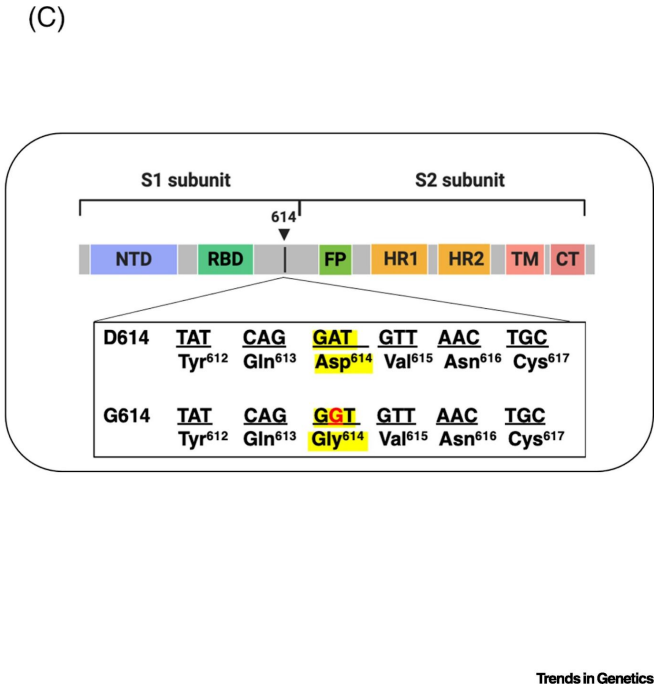
Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	165	40.842%
EXON	58	14.356%
INTERGENIC	3	0.743%
UPSTREAM	178	44.059%

Which kind(s) of variants would you predict to most likely affect phenotype?

Missense mutations: amino acid substitutions





SARS-CoV-2 enters human cells via the ACE2 receptor.

Amino acid substitutions in the spike protein (surface protein) can alter how that interaction happens.

Zhang et al. (2020)

2. Phylogenetics

How is this variant related to others?

What is phylogenetics?

The study of the evolutionary relationships among evolving entities (species, genomes, genes, individuals).

"The time will come, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of nature"

- Darwin (1857)

"One of the grand biological ideas is to be able to work out the complete detailed quantitative phylogenetic tree --- the history of the origins of all living species, back to the very beginning."

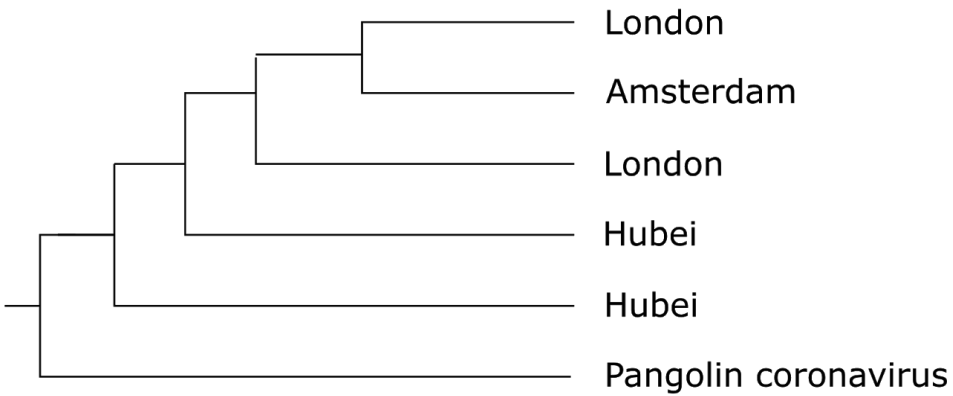
- Dayhoff and Eck (1972)

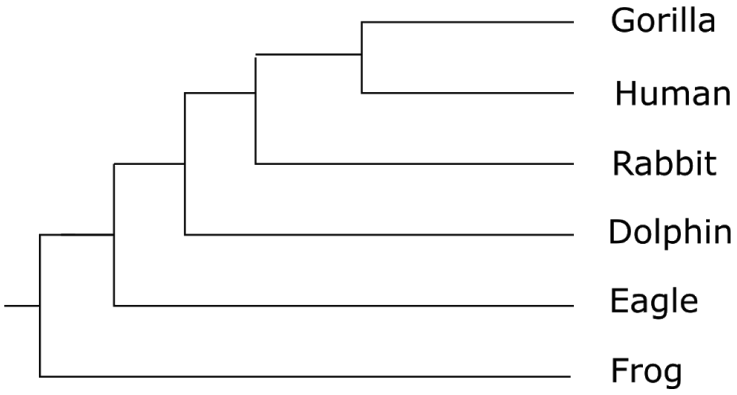
(Various practical applications)

Phylogenetics has very direct applications in tracing the COVID-19 pandemic

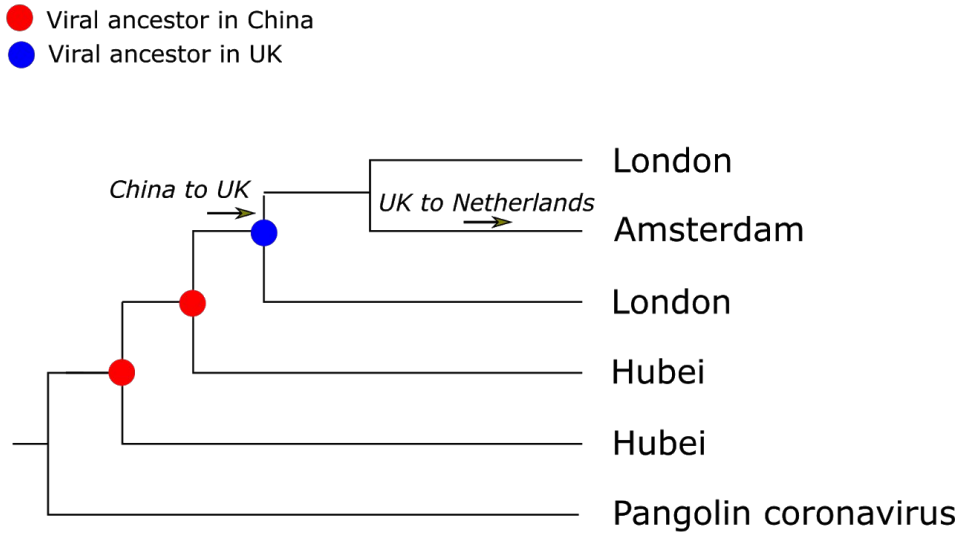
- A wonderful subject, but here we'll focus on a very pressing, real world application
- The COG-UK project uses phylogenetics to study the epidemiology of COVID-19:
  - Where do new strains originate?
  - How do they spread between and within countries?

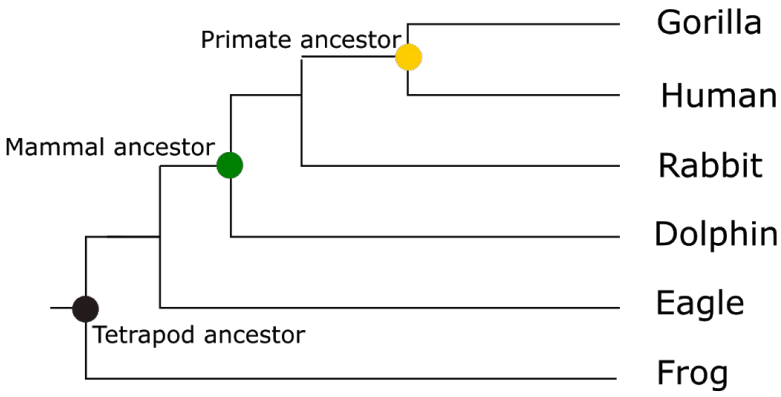
Tracing the spread of a virus is like tracing the ancestry of cellular lifeforms





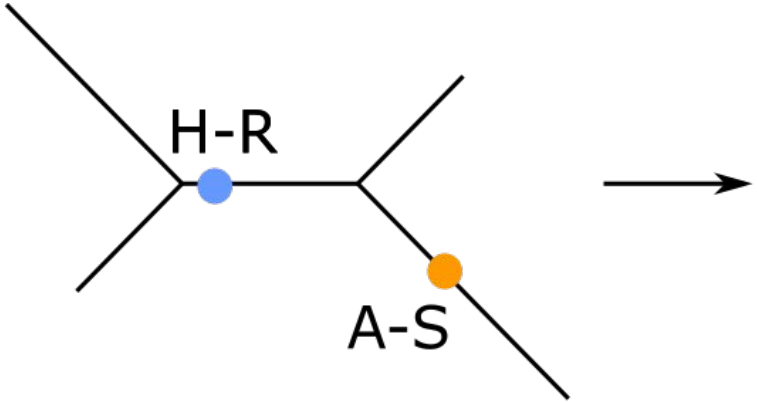
Tracing the spread of a virus is like tracing the ancestry of cellular lifeforms

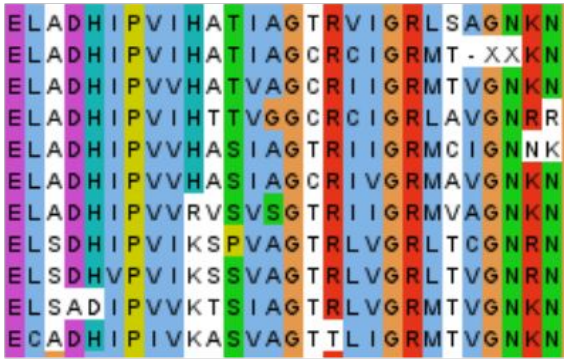




NB: Interpreting phylogenies is about ancestral and derived states, common ancestors; not about “progress”.

Phylogeny inference makes use of substitution models





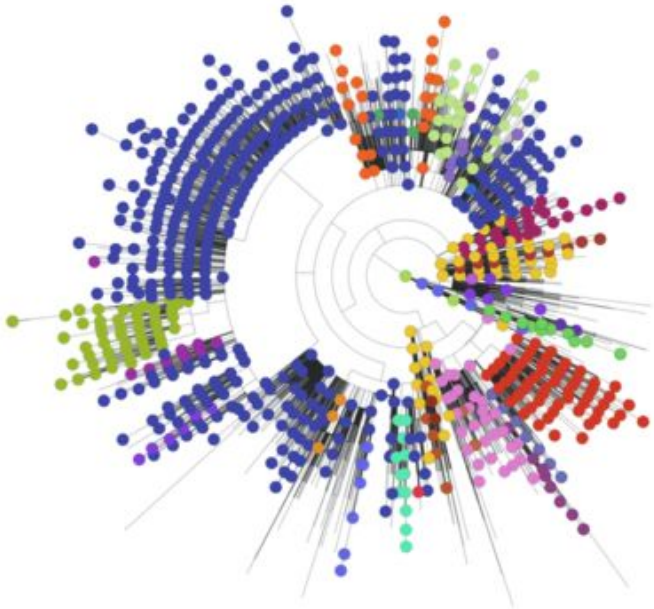
Substitution model

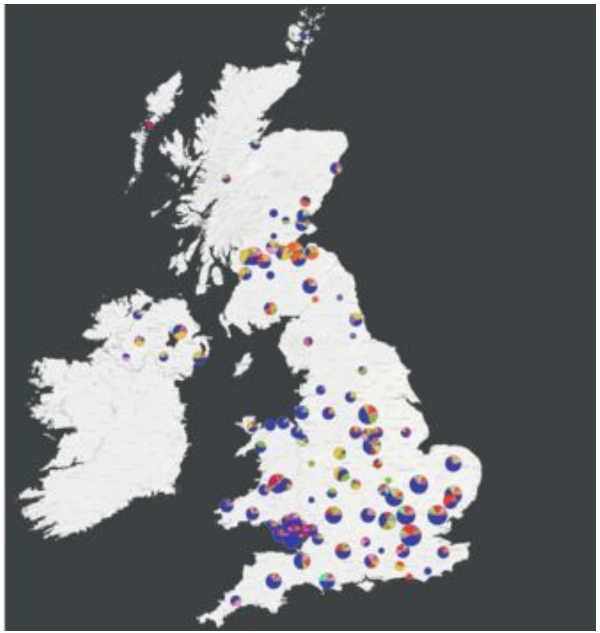
(sequences evolve along tree, experiencing substitutions)

Alignment

Prob(data) ~ G; RaXML, IQ-Tree, PhyML, MrBayes, RevBayes, PhyloBayes

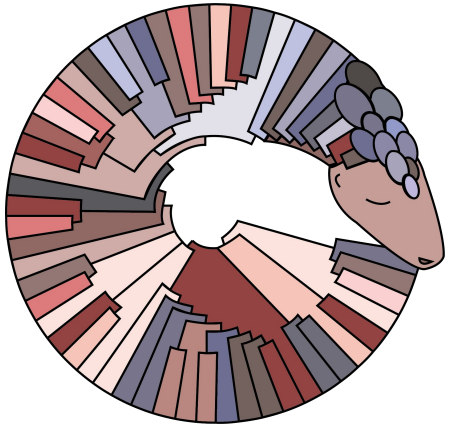
COG-UK/NextStrain maintain an alignment and phylogeny of all sequenced COVID19 genomes





MicroReact database: live demo

Session 5 practical



Identify and track COVID19 variants!

Part 1: Use variant calling to characterise the mutations in two new variant strains. Predict the impact of any detected changes.

Part 2: Infer the geographic origin of the two variants using Pangolin/MicroReact.