What is computational biology?

David Welch

School of Computer Science, Centre for Computational Evolution, and Te Pūnaha Matatini

What is biology?

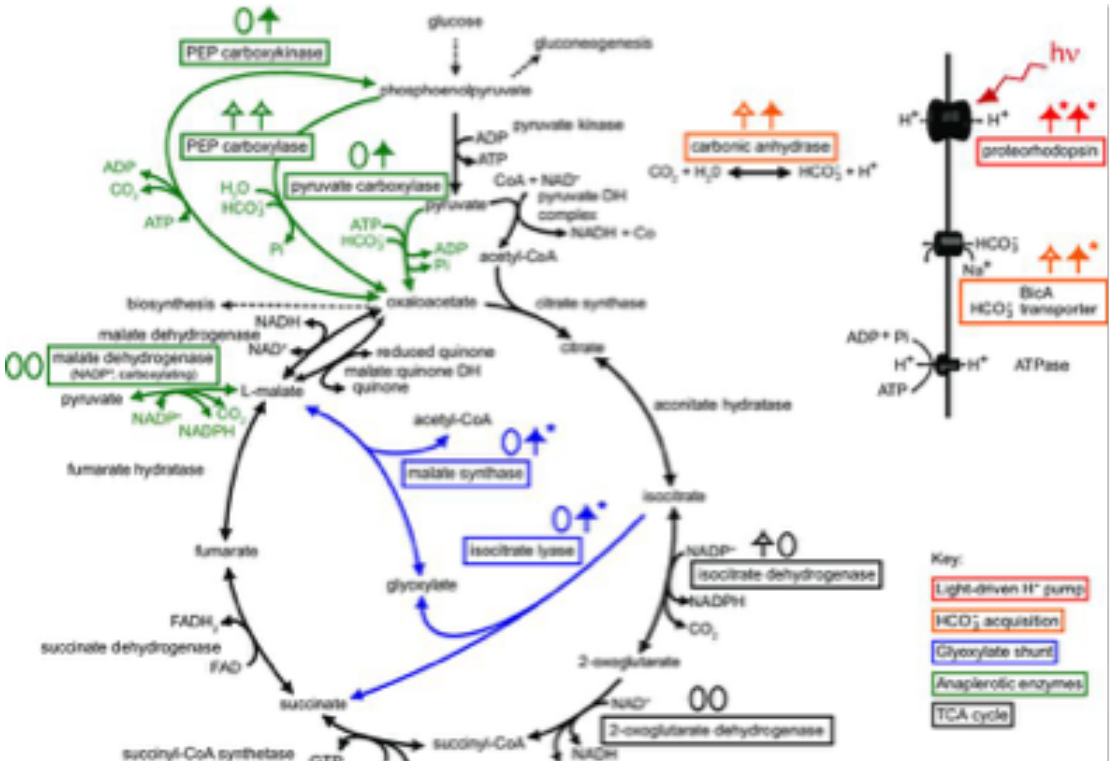
Biology is the study of life. This is a broad target!

- **individual organisms**
- **populations of organisms**
- **evolution of populations and species**
- **ecological systems** (interactions between diverse populations)
- **organism subsystems** (e.g. organs)
- **cell biology**
- **genetics**
- **metabolism**
- **ethology** (behaviour)
- **the origin of life** (how is life different from non-life, how did life emerge?)
- And more….!

Biological systems differ from eg physical ones

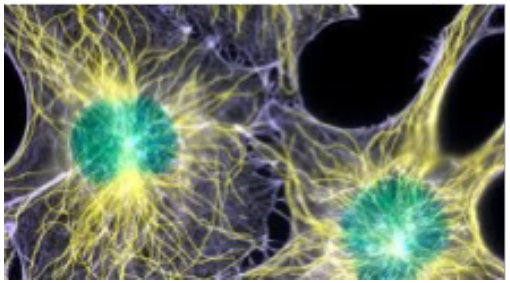
Biological systems are **complex**

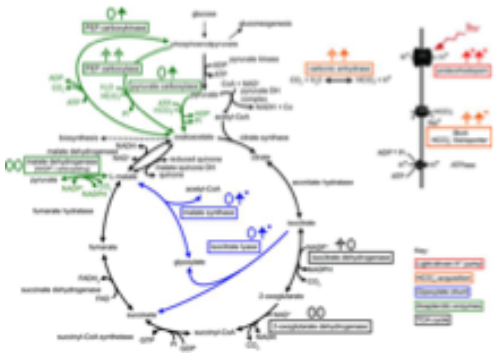
- lots of interacting parts
- parts are different from each other (**heterogeneity**)
- the parts interact **non-linearly**

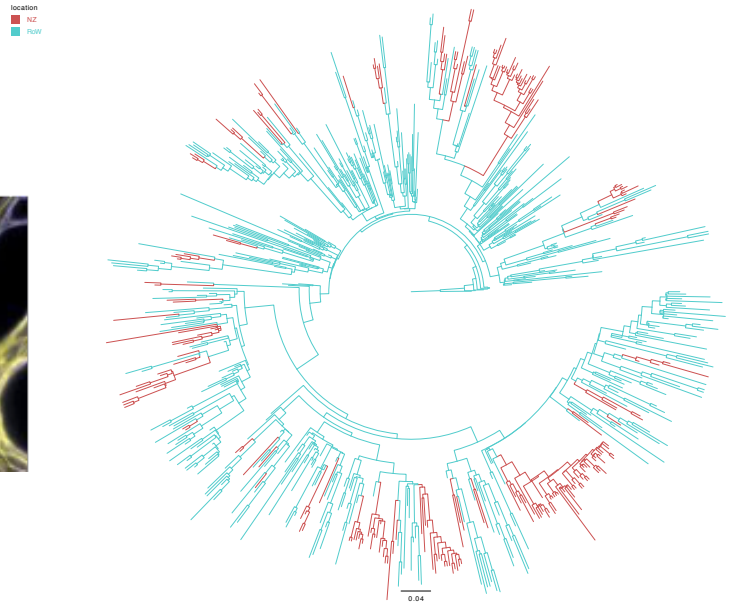


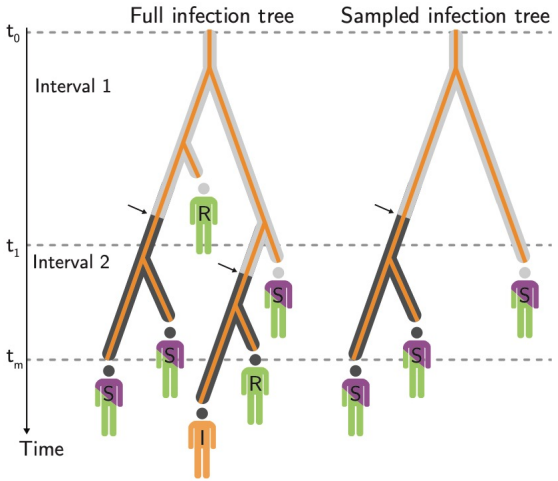
How can computers help?

- Modelling
- Simulation
- Statistical inference
- Data processing
- Visualisation
- Data collection
- Automation
- …









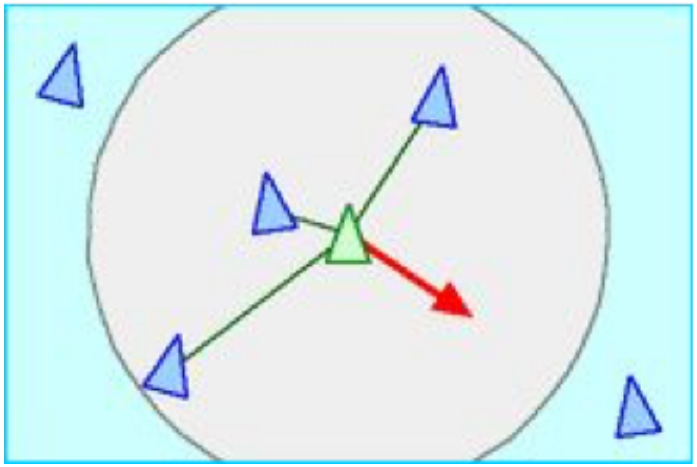
3 examples

- Boids: modelling, simulation, emergence
- Alignment and structural prediction: algorithms, gamification, heuristics
- Phylodynamics of COVID-19: inference, big data, visualisation

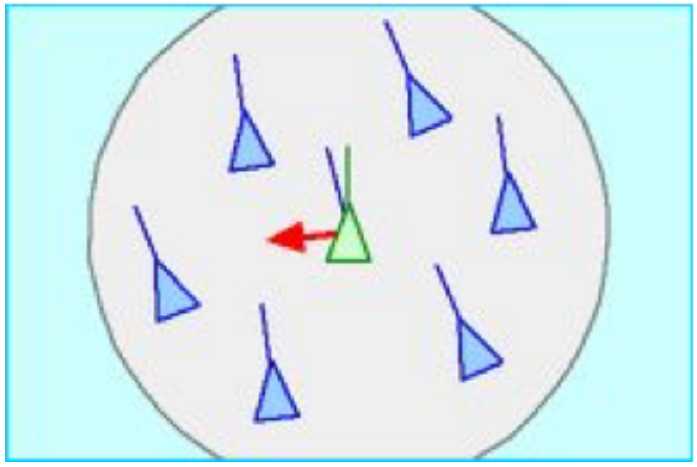
Boids: a simple model of flocking

Each bird is following a few simple rules and the complex pattern simply emerges from them

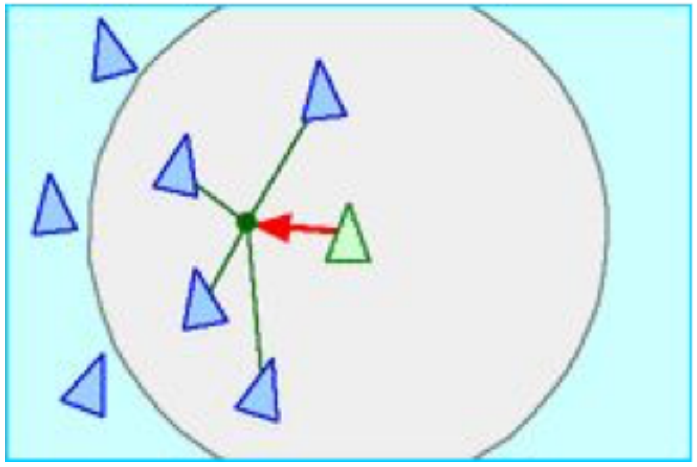
Separation



Alignment

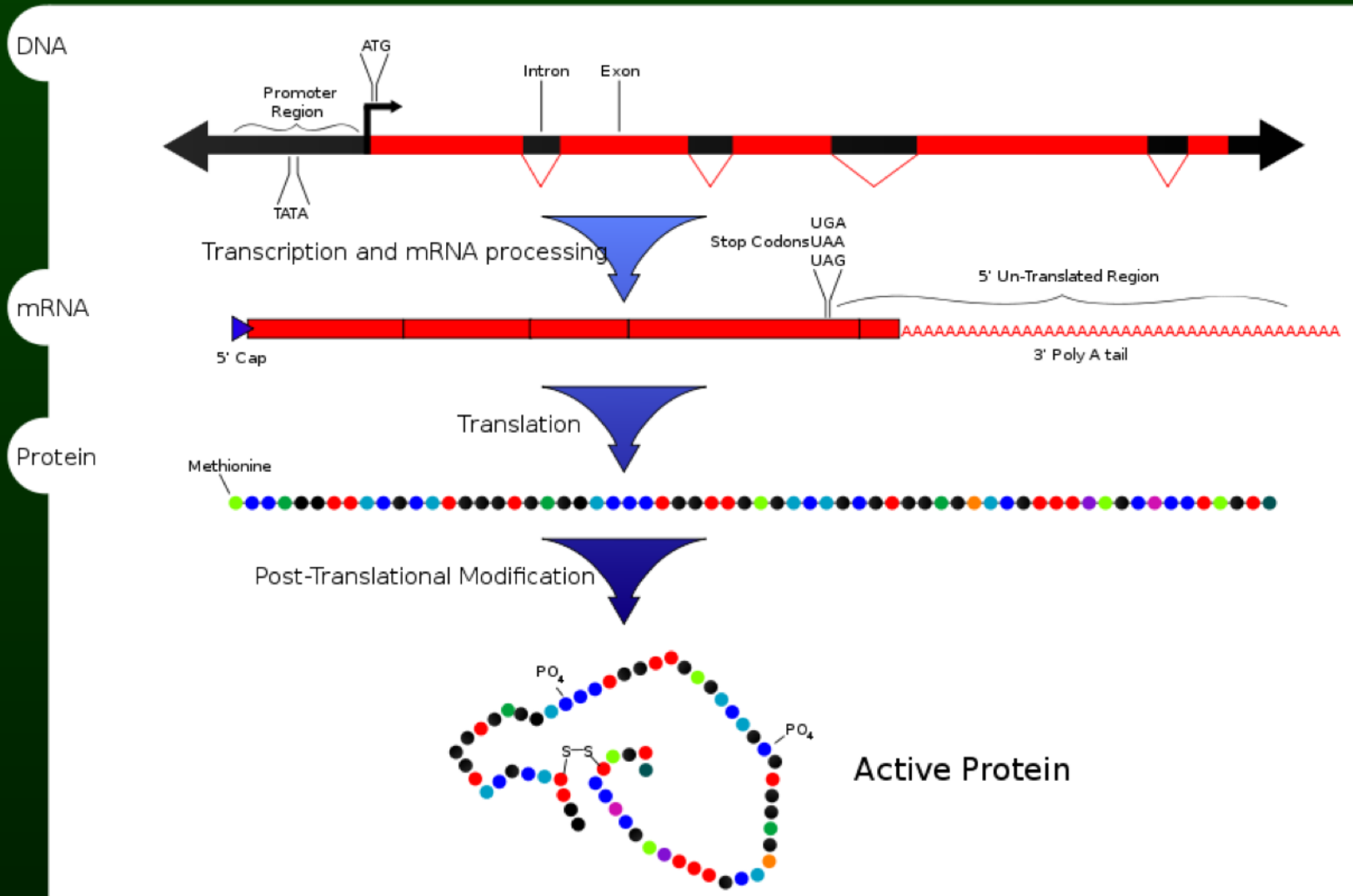


Cohesion

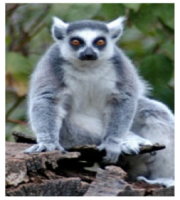


Images taken from Craig Reynold's website. http://www.red3d.com/cwr/boids/

Central Dogma of Molecular Biology : Eukaryotic Mode



Alignments of Human Beta-Globin to Other Globins



```
Human beta-globin                    VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                                       LTPEE   VT LWGKVNV  VGGEALGRLLVVYPWTQRFFESFGDLS PDA MGN
Ring-tailed lemur beta-globin        TFLTPEENGHVTSLWGKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDA MGN

PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
PKVKAHGKKVL AFS GL HLDNLKGTFA LSELHC  LHVDPENF LLGNVLV VLAHHFG  F P  QAA QKVV GVANALAHKYH
PKVKAHGKKVLSAFSEGLHHLDNLKGTFAQLSELHCVALHVDPENFKLLGNVLVIVLAHHFGNDFSPQTQAAFQKVVIGVANALAHKYH
```

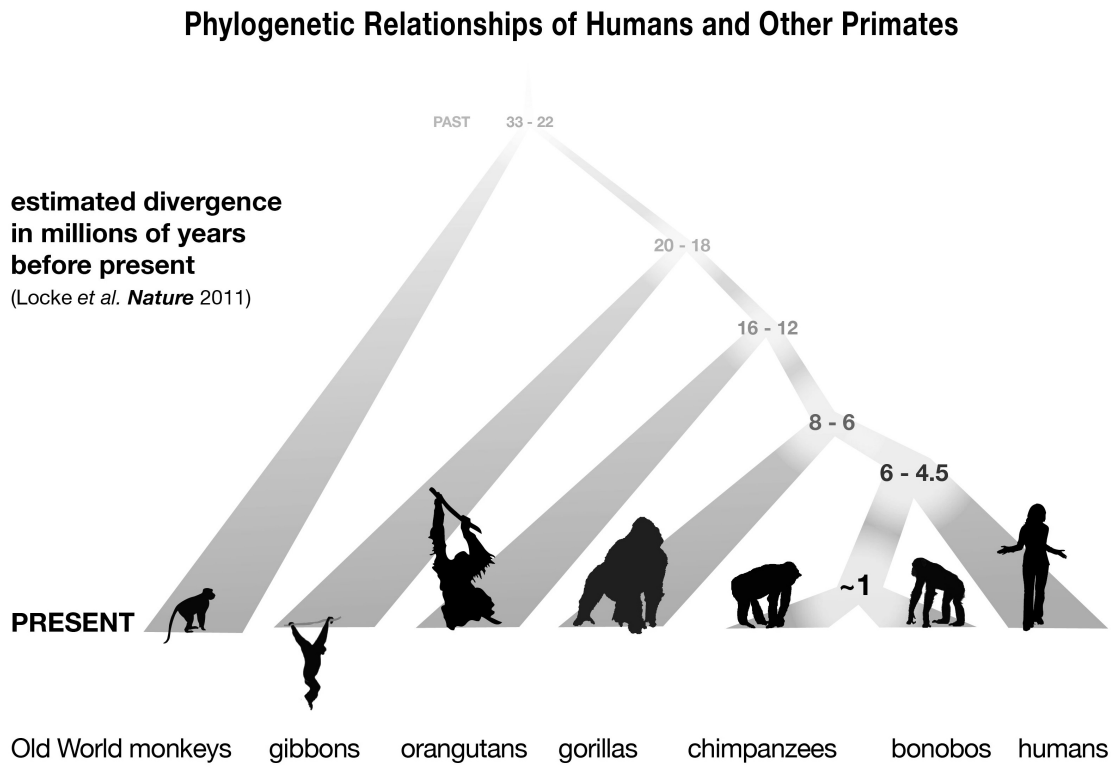
```
Human beta-globin                    VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
                                     V  T  E SA   LWGK N DE G  AL R L VYPWTQR F  FG LS P A MGNP
Goldfish beta-globin                 VEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNP

KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-KEFTPPVQAAYQKVVAGVANALAHKYH
KV AHG  V G          DN K T A LS  H  KLHVDP NFRLL        A  FG   F   VQ A QK    V  AL   YH
KVAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFGPSGFNADVQEAWQKFLSVVVSALCRQYH
```

```
Human beta-globin                  VHLTPEEKSAVTALW----GKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKA
                                     L       V   W    G  N   VG E     L          F  F   S        P V
Bloodworm globin IV                MGLSAAQRQVVASTWKDIAGSDNGAGVGKECFTKFLSAHHDIAAVF-GFSGAS-------DPGVAD

HGKKVLGAFSDGLAHL-DNLKGTFATLSELHCDK----LHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
 G KVL    D   HL D  K             K     H   E F  LG  L     H  G   T     A     A    AL
LGAKVLAQIGVAVSHLGDEGKMVAEMKAVGVRHKGYGYKHIKAEYFEPLGASLLSAMEHRIGGKMTAAAKDAWAAAYADISGALISGLQ
```

```
Human beta-globin                   VHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
                                    V  T      V       K N             L   P     F         P     NPK
Soybean leghemoglobin               VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPT----NPK

VKAHGKKVLGAFSDGLAHLDNLKGTFA--TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
   H  K       D    L       A   L   H  K   DP  F       L         G          A         A A
LTGHAEKLFALVRDSAGQLKASGTVVADAALGSVHAQKAVTDPQ-FVVVKEALLKTIKAAVGDKWSDELSREWEVAYDELAAAIKKA--
```

How to align multiple sequences?

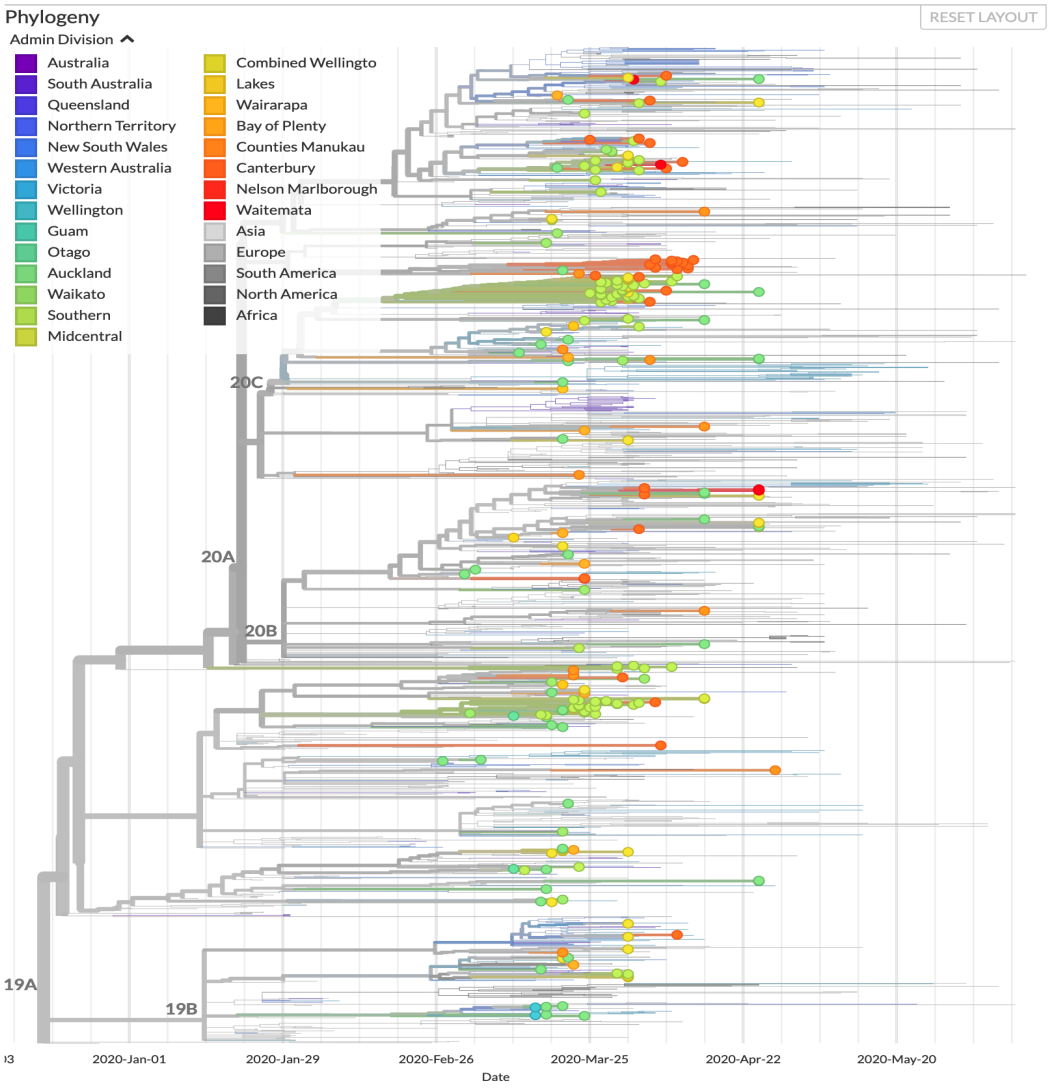
A phylogeny shows relationships between sampled species, populations, or individuals

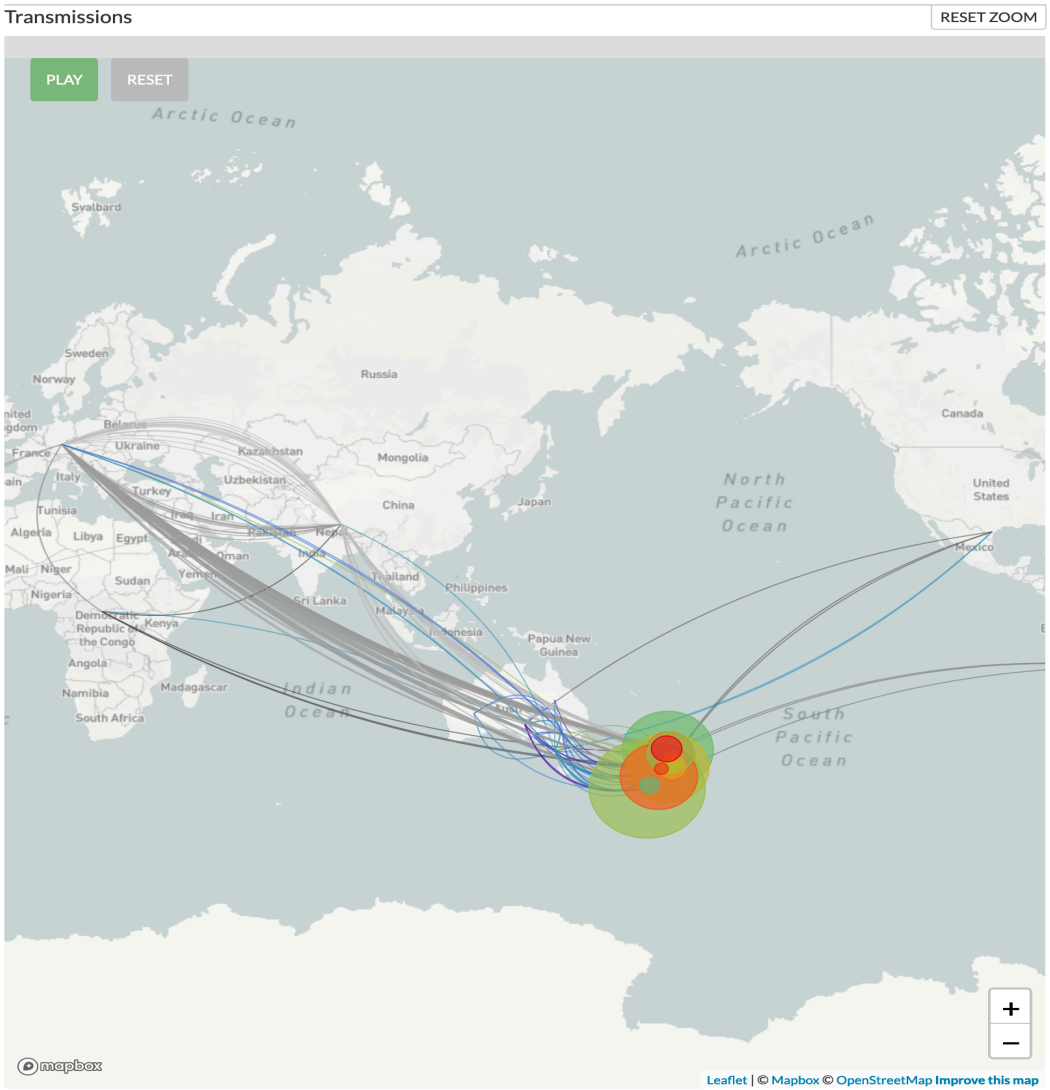


Phylodynamics of COVID-19

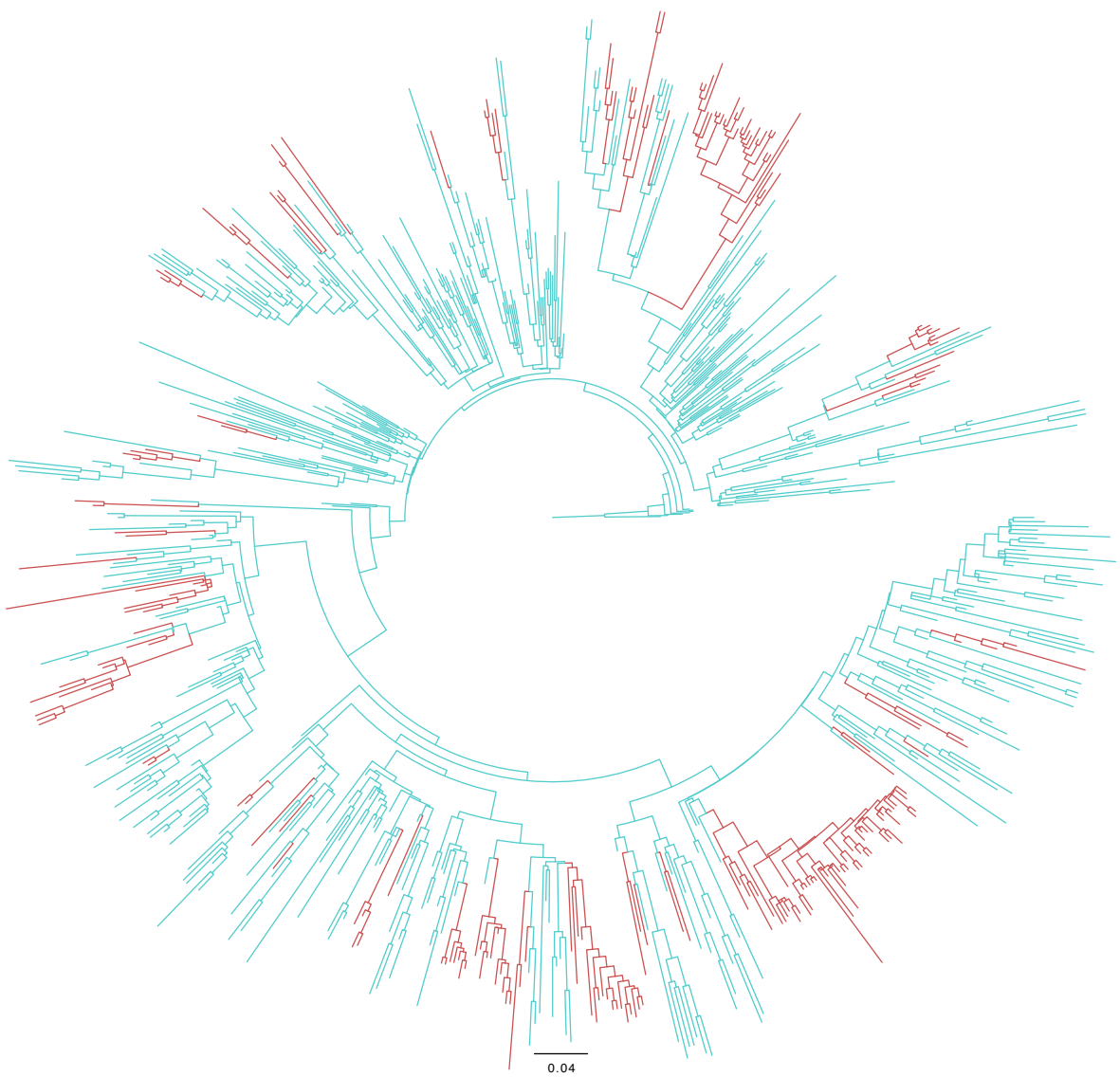
- Phylogenetics builds trees from genomes by modelling mutation processes, typically ignores population altogether.
- building a phylogenetic tree can tell us a lot about a virus: suggests local transmission chains, places sequences geographically, etc
- Phylodynamics accounts for the dynamics in the population of interest, incorporates information like
  - number of cases
  - Location and movement
  - Sampling dates

Nextstrain.org provides an excellent global overview

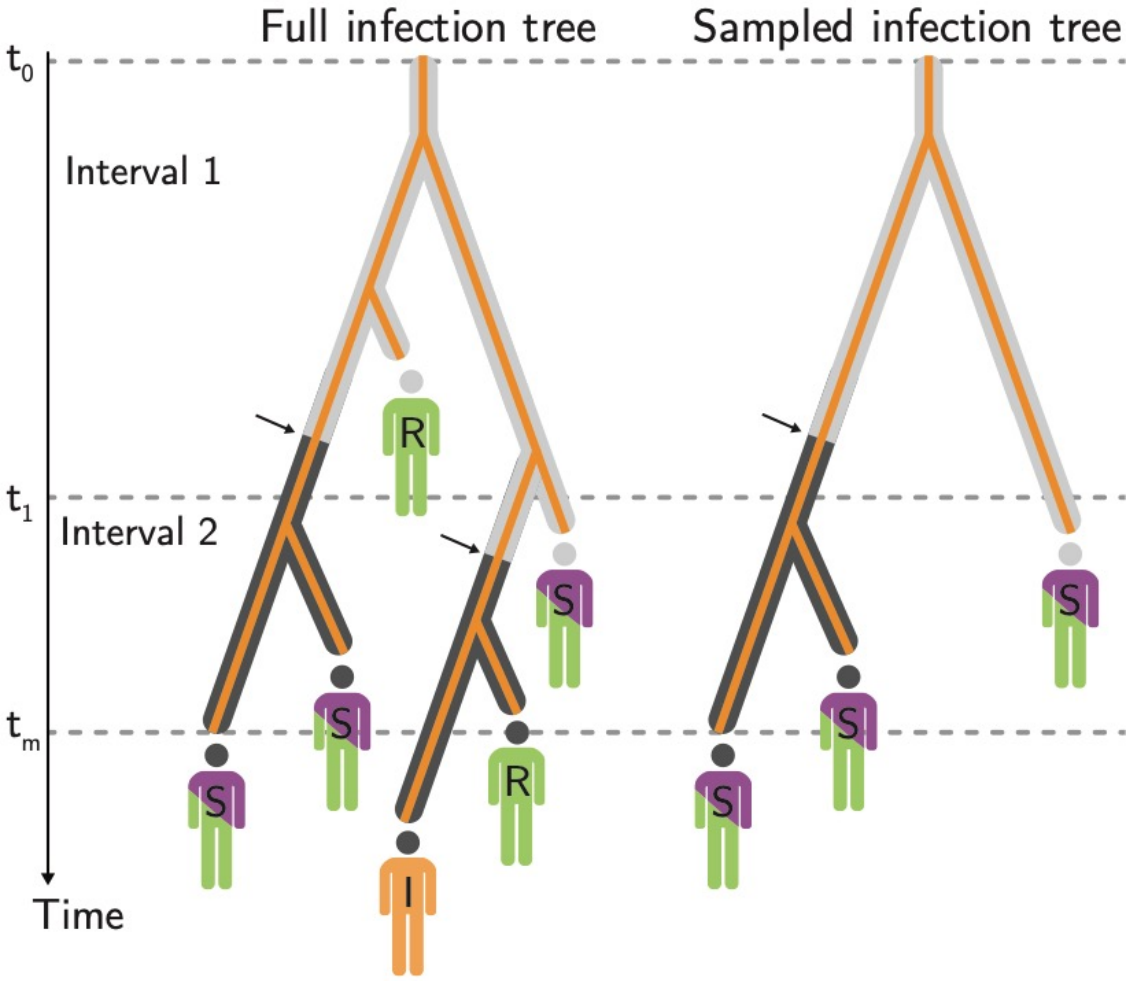




220 NZ samples, 500 for rest of world, time increases towards leaves



A basic phylodynamic model



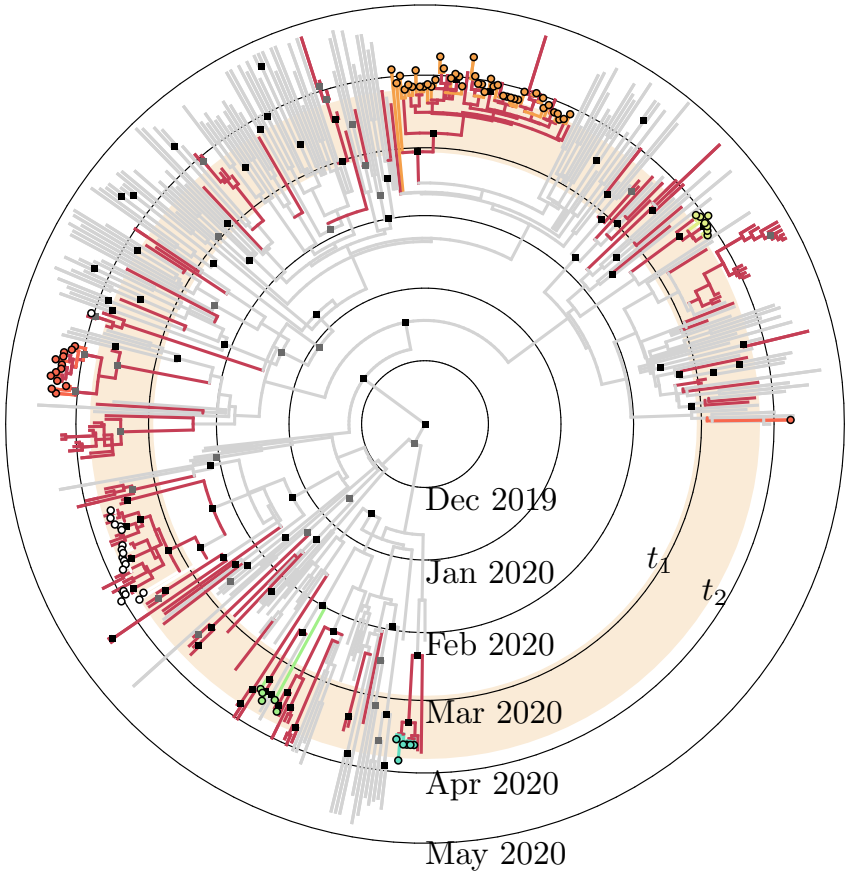
Two locations: World and NZ

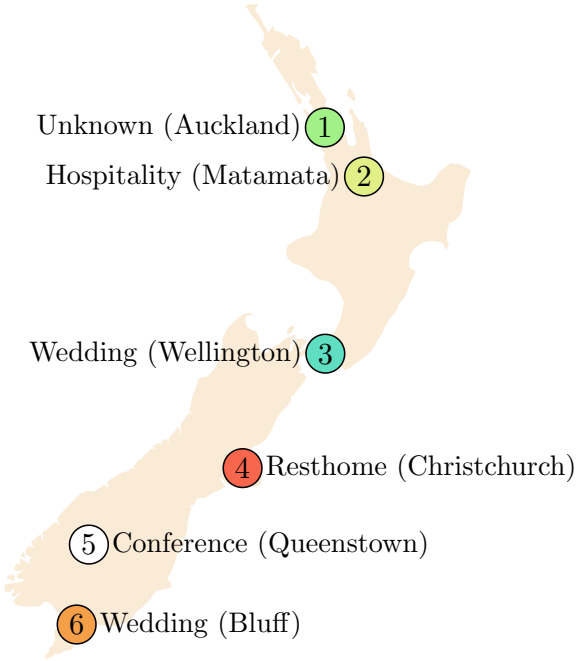
Parameters

Migration rates
Transmission rates (“birth”)
Removal rates (“death”)
Sampling rates

Parameters can differ in different intervals

Major outbreaks in New Zealand



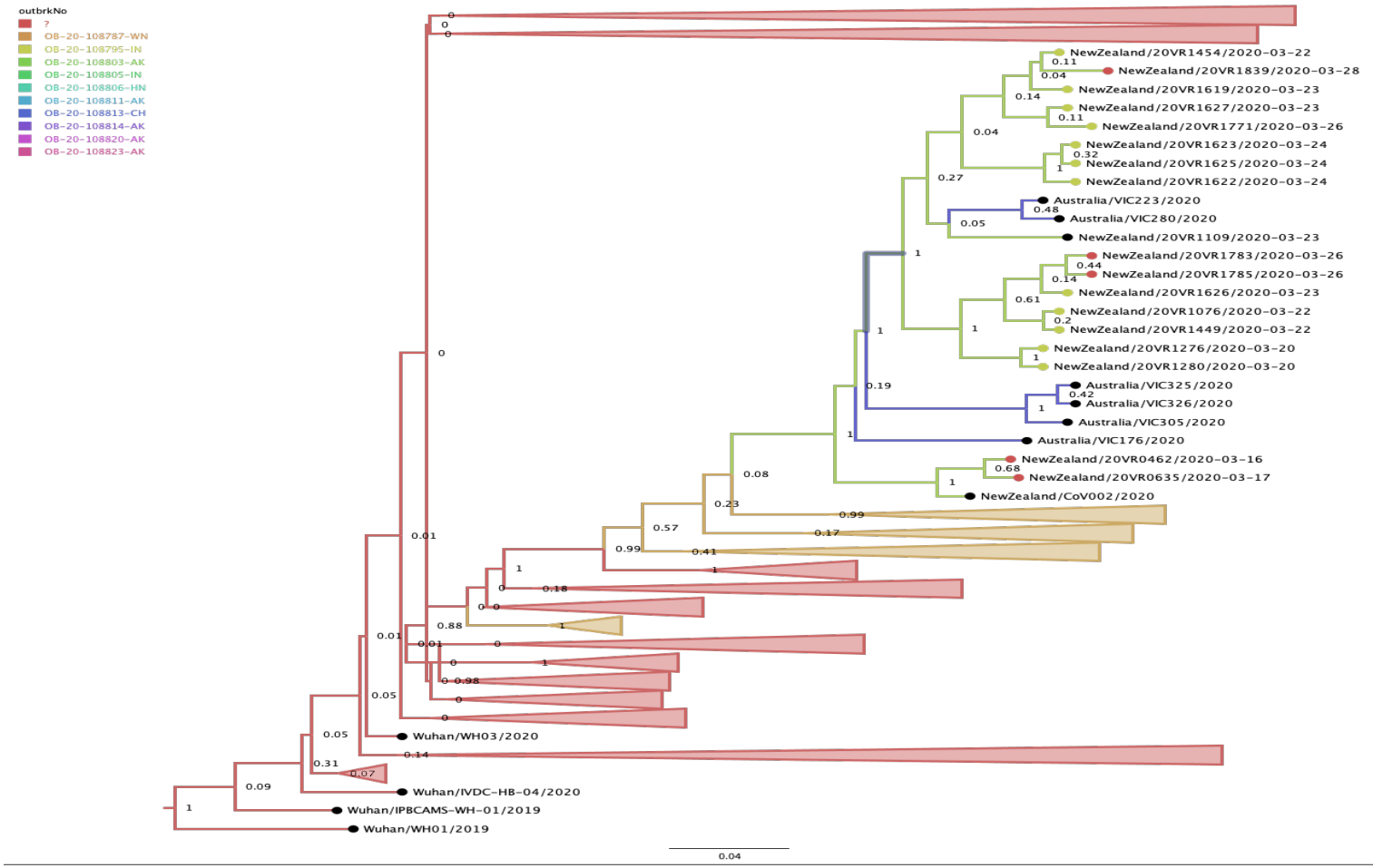


$\mathcal{RW}$ deme
$\mathcal{IS}$ deme

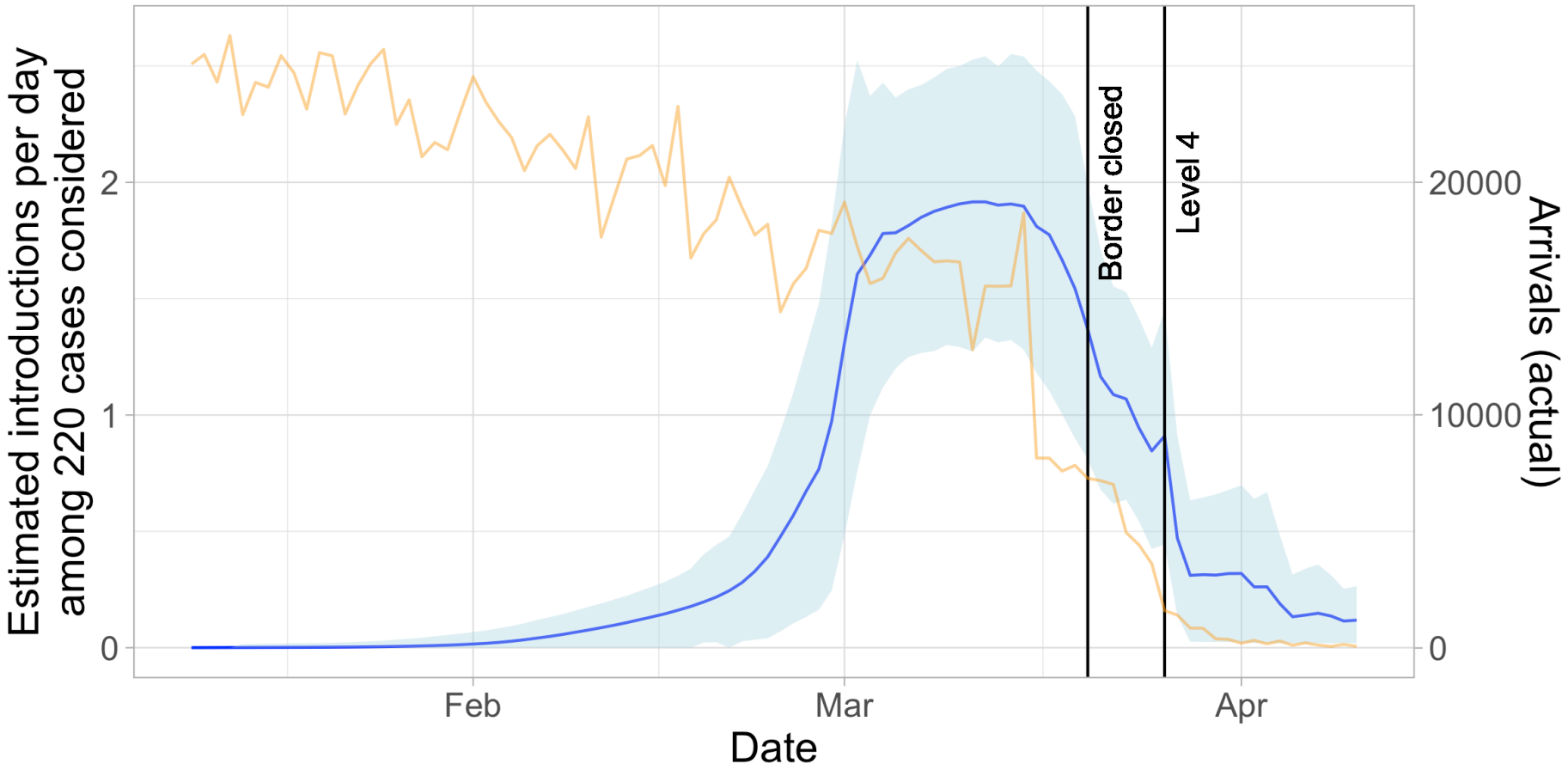
t_1: First reported case in $\mathcal{IS}$
t_2: Mobility decrease in $\mathcal{IS}$

Clade support:
$p \geq 0.9$
$0.5 \leq p < 0.9$

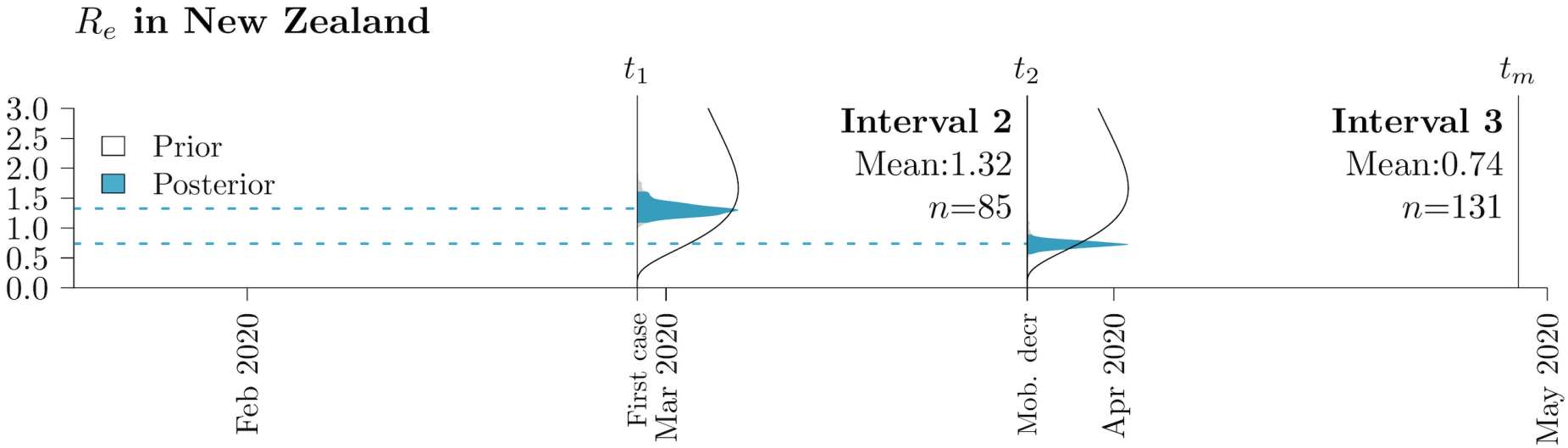
Hereford conference – TMRCA 25 Feb-2 Mar



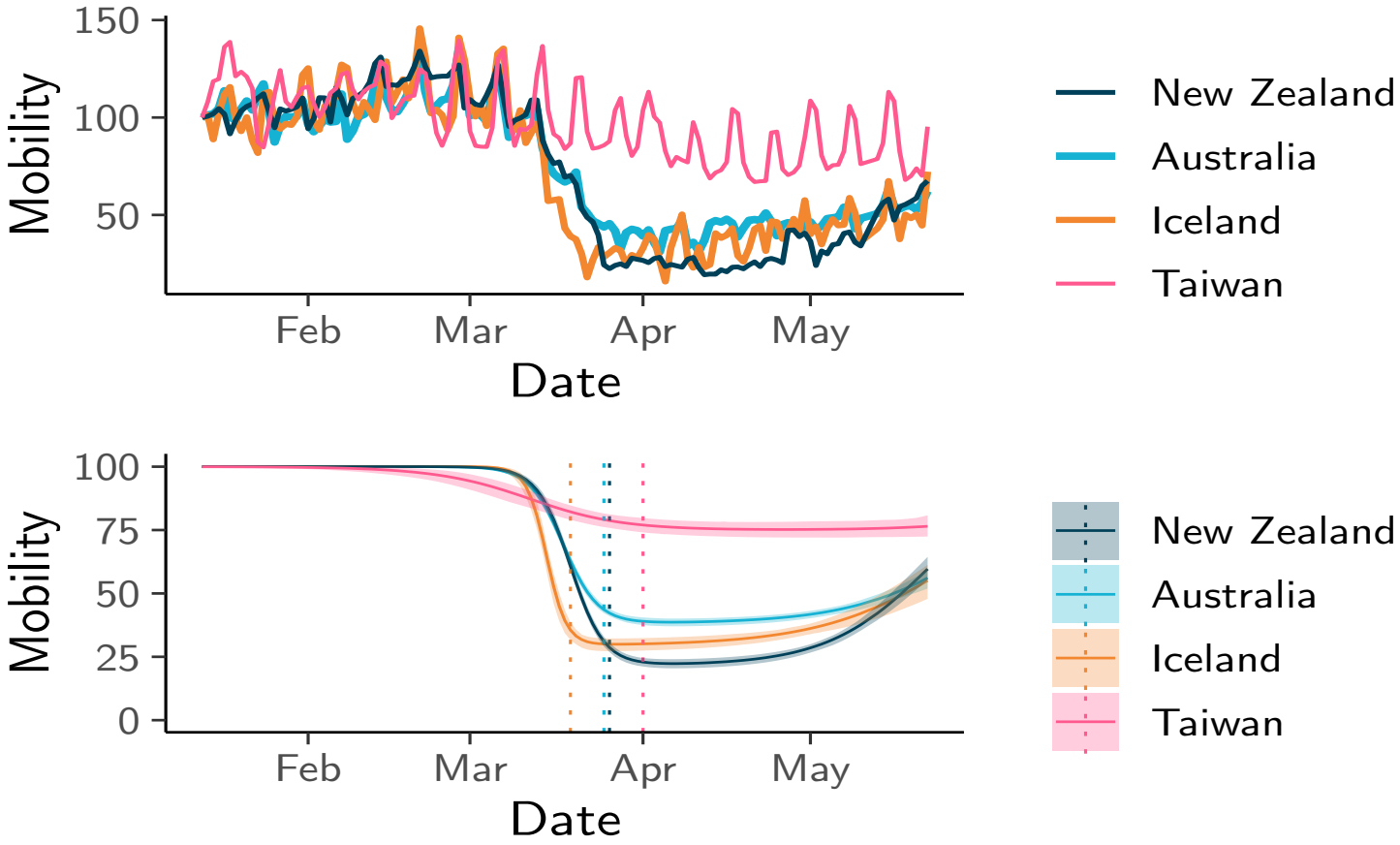
When were cases introduced to New Zealand?



Estimating the reproduction number



Using mobility data to define intervals



Challenges in computational biology

- More data being generated than we can handle...
- …and data is noisy, often highly correlated
- High levels of cross disciplinary work required
- Models are easy to build and simulate but inference is hard
- Finding balance of simplicity and complexity is very hard
- How can we combine multiple sources of data?

Links

- Murmuration: https://www.youtube.com/watch?v=DmO4Ellgmd0
- Boids: https://www.youtube.com/watch?v=nbbd5uby0sY
- Protein folding animation: https://www.youtube.com/watch?v=jVOyaT56LEU
- Foldit: https://fold.it/portal/info/about
- Multiple alignment game: https://phylo.cs.mcgill.ca/play.php
- Nextstrain COVID: https://nextstrain.org/ncov/global