

What is computational biology?

David Welch

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

What is biology?

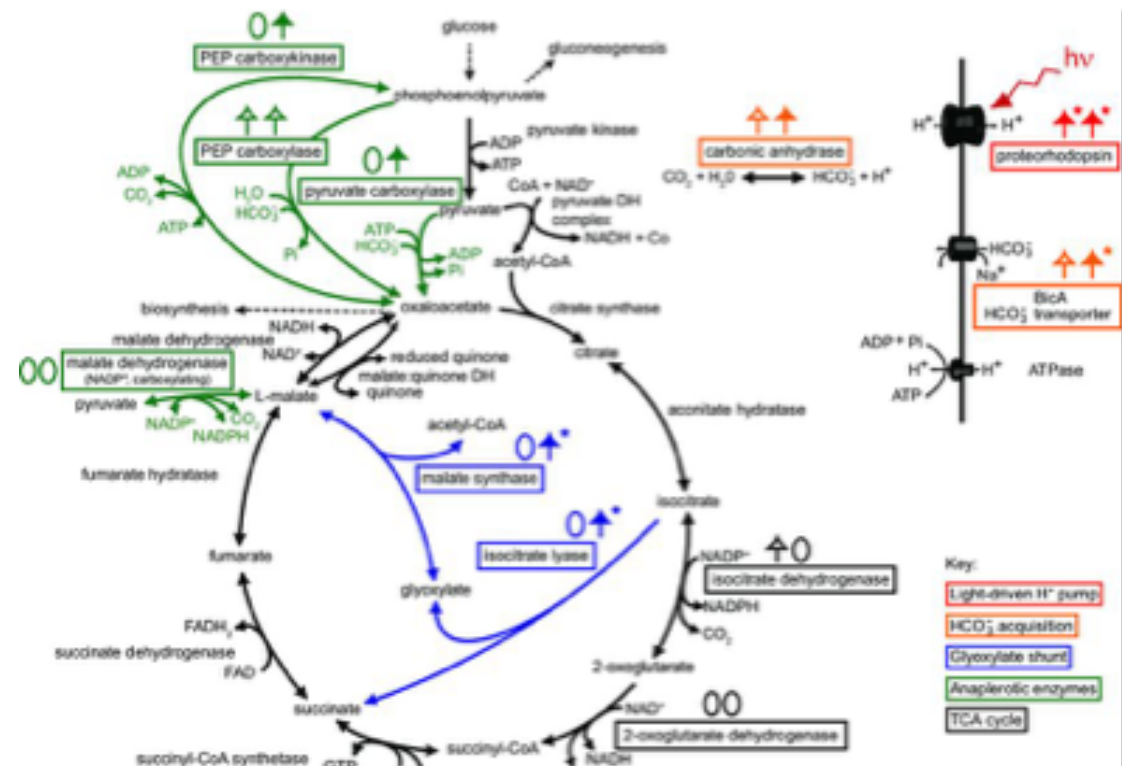
Biology is the study of life. This is a broad target!

- **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

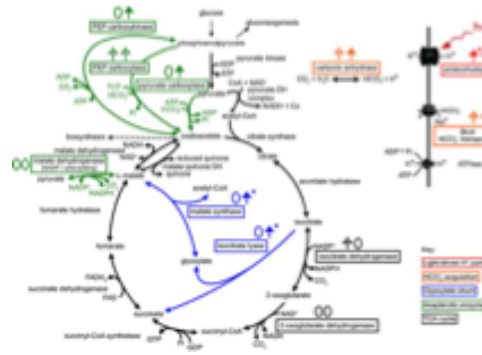
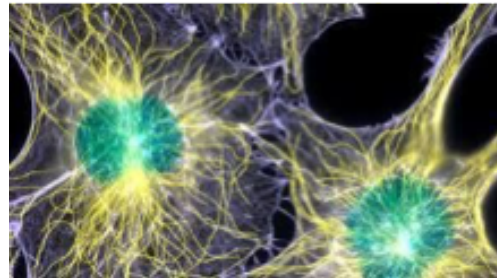
Biological systems are **complex**

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

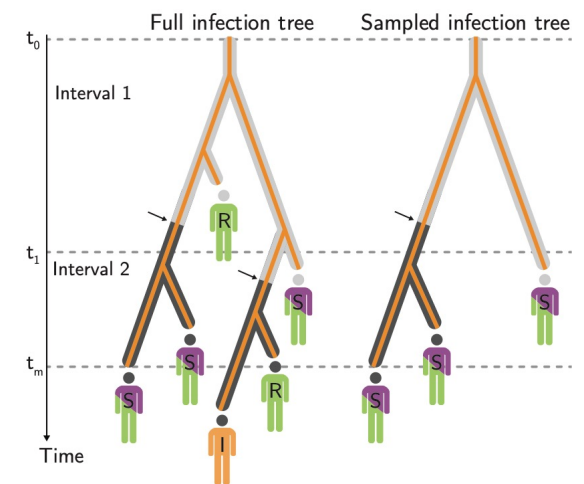
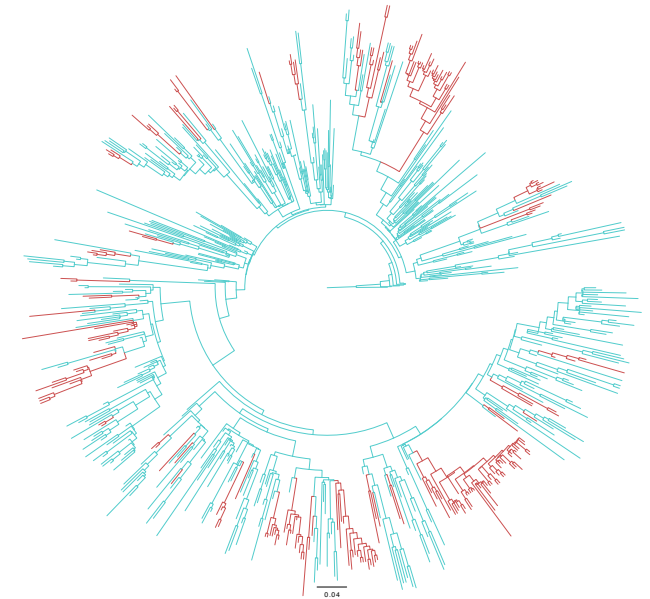


How can computers help?

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



location
NZ
FOW



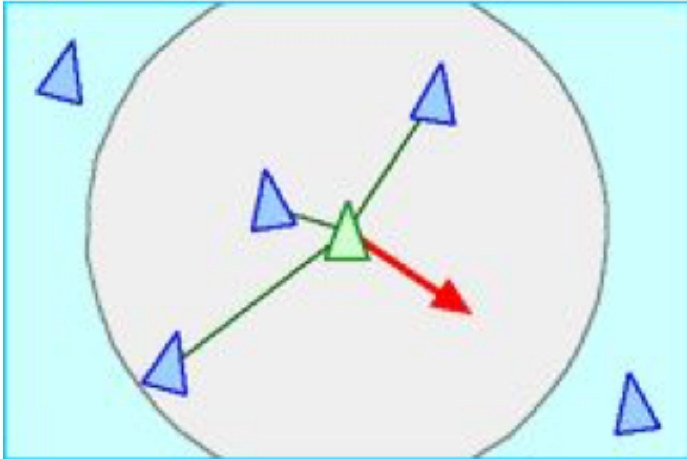
3 examples

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

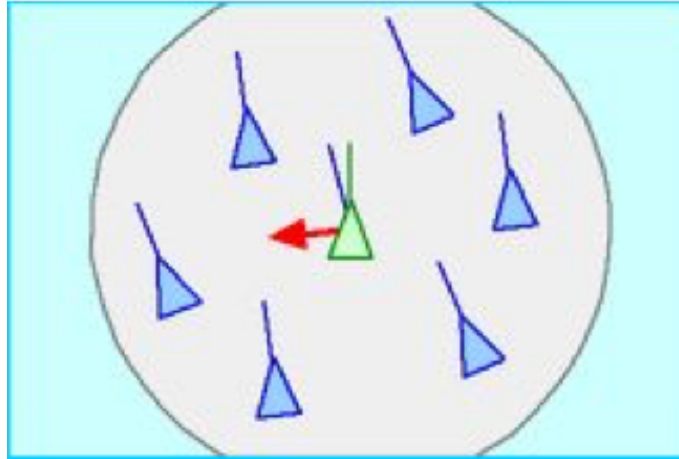
Boids: a simple model of flocking

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

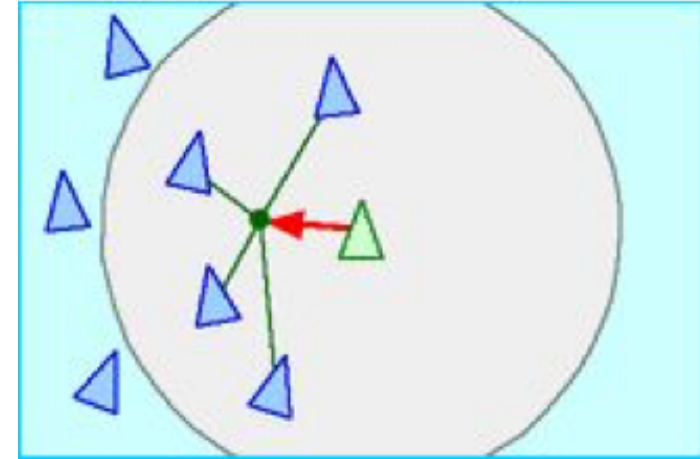
Separation



Alignment



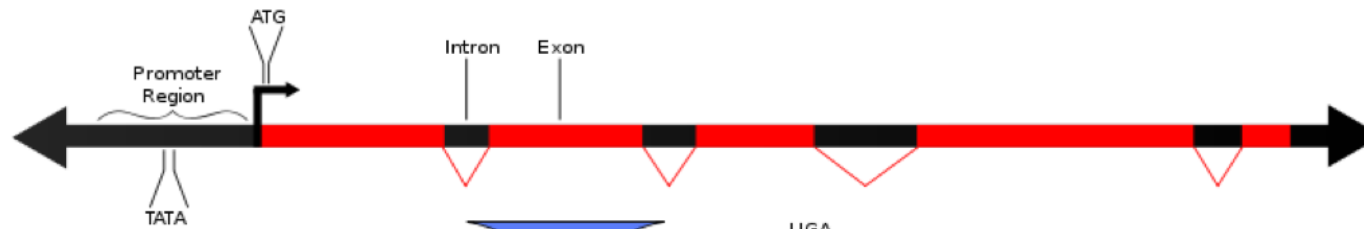
Cohesion



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

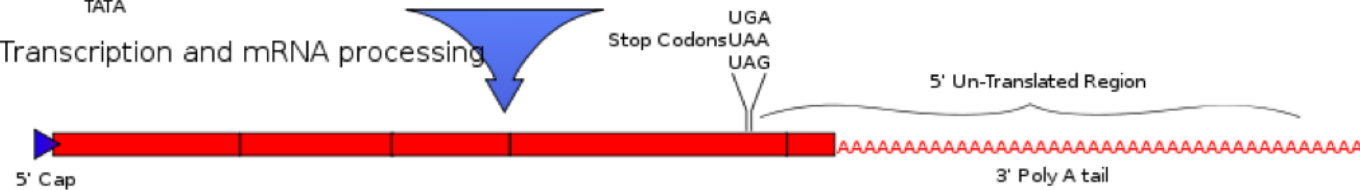
Central Dogma of Molecular Biology : Eukaryotic Mode

DNA



Transcription and mRNA processing

mRNA

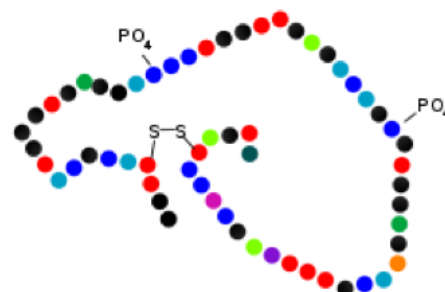


Translation

Protein

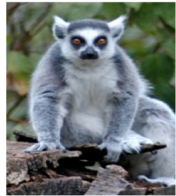


Post-Translational Modification



Active Protein

Alignments of Human Beta-Globin to Other Globins



Human beta-globin	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
Ring-tailed lemur beta-globin	LTPEE VT LWGKVVN VGGEALGRLLVVYPWTQRFFESFGDLS PDA MGN
	TFLTPEENGHVTS LWGKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDA MGN
	PKVKAHGKKVLGAFSDGLAHL DNLKGTFA TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAA YQKVVAGVANALAHKYH
	PKVKAHGKKVL AFS GL HLDNLKGTFA LSELHC LHVDPENF LLGNVLV VLAHHFG F P QAA QKVV GVANALAHKYH
	PKVKAHGKKVLSAFSEGLHHL DNLKGTFAQLSELHCVLHVDPENFKLLGNVLVIVLAHHFGNDFSPQTQAAFQKVVIGVANALAHKYH



Human beta-globin	VHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
Goldfish beta-globin	V T E SA LWGK N DE G AL R L VYPWTQR F FG LS P A MGNP
	VEWTD AERSAI IGLWGK LNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNP
	KVKAHGKKVLGAFSDGLAHL DNLKGTFA TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-KEFTPPVQAA YQKVVAGVANALAHKYH
	KV AHG V G DN K T A LS H KLHVDP NFRLL A FG F VQ A QK V AL YH
	KVAAHGRTVMGGLERA IKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFGPSGFNADVQEA WQKFLSVVVSALCRQYH



Human beta-globin	VHLTPEEKSAVTALW----GKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKA
Bloodworm globin IV	L V W G N VG E L F F S P V
	MGLSAAQRQVVASTWKDIAGSDNGAGVGKECF TKFLSAHHDIAAVF-GFSGAS-----DPGVAD
	HGKKVLGAFSDGLAHL-DNLKGTFA TLSELHCDK----LHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAA YQKVVAGVANALAHKYH
	G KVL D HL D K K H E F LG L H G T A A AL
	LGAKVLAQIGVAVSHLGDEGKMVAEMKAVGVRHKGYGYKHIKAEYFEPLGASLLSAMEHRIGGKMTAAAKDAWAAAYADISGALISGLQ



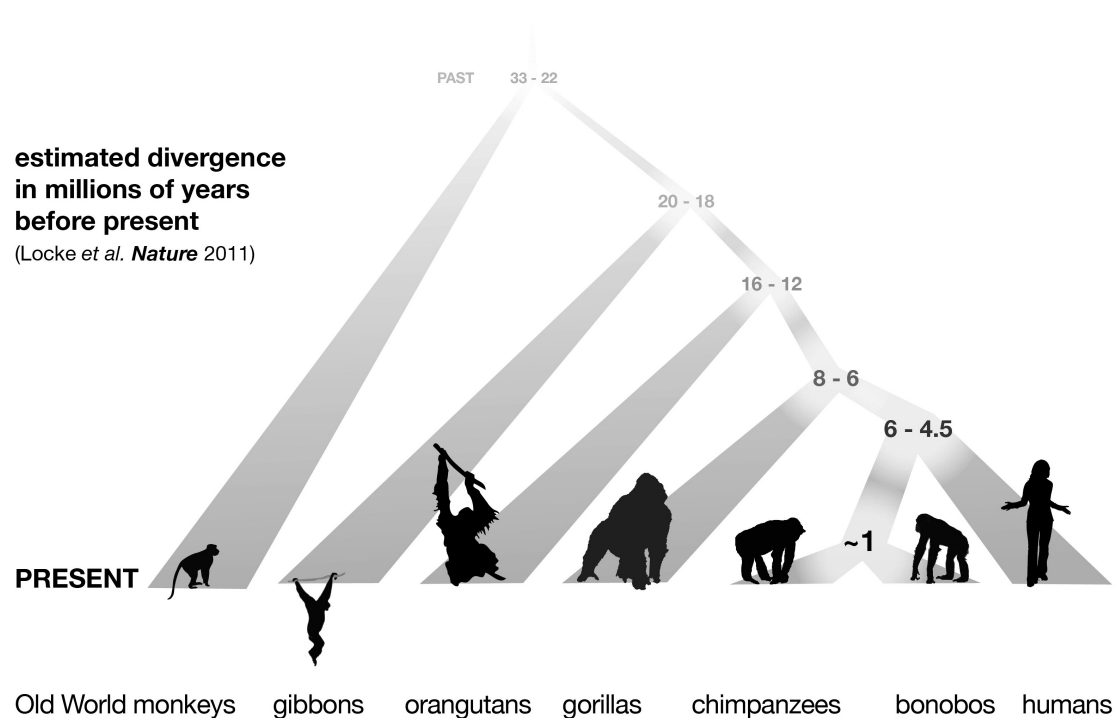
Human beta-globin	VHLTPEEKSAVTALWG--KVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Soybean leghemoglobin	V T V K N L P F P NPK
	VAFTEKQDALVSSSF EAFKANIPQYSVVFYTSILEKAPAAKD LFSFLANGVDPT----NPK
	VKAHGKKVLGAFSDGLAHL DNLKGTFA--TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAA YQKVVAGVANALAHKYH
	H K D L A L H K DP F L G A A A
	LTGHA EKL FALVRDSAGQLKASGTVVADAALG SVHAQKAVTDPQ-FVVVKEALLKTIKAAVGDKWSELSREWEVAYDELAAAIKKA--

How to align multiple sequences?

C_aminophilum	AGCT.YCGCA	TGRAGCAGTG	TGAAAA....	ACTCCGGT	GGTACAGGAT
C_colinum	AGTA..GGCA	TCTACAAGTT	GGAAAA....	ACTGAGGT	GGTATAGGAG
C_lentocellum	GGTATTCGCT	TGATTATNAT	AGTAAA....	GATTTATC	GCCATAGGAT
C_botulinum_D	TTTA.TGGCA	TCATACATAA	AATAATCAAA	GGAGCAATCC	GCTTTGAGAT
C_novyi_A	TTTA.CGGCA	T....CGTAG	AATAATCAAA	GGAGCAATCC	GCTTTGAGAT
C_gasigenes	AGTT.TCGCA	TGAAACA...	GC.AATTAAA	GGAGAAATCC	GCTATAAGAT
C_aurantibutyricum	A.NT.TCGCA	TGGAGCA...	AC.AATCAAA	GGAGCAAT.C	ACTATAAGAT
C_sp_C_quinii	AGTT.T.GCA	TGGGACA...	GC.AATTAAA	GGAGCAATCC	GCTATGAGAT
C_perfringens	AAGA.TGGCA	T.CATCA...	TTCAACC AAA	GGAGCAATCC	GCTATGAGAT
C_cadaveris	TTTT.CTGCA	TGGGAAA...	GTC.ATGAAA	GGAGCAATCC	GCTGTAAGAT
C_cellulovorans	ATTC.TCGCA	TGAGAGA...	.TGATACAAA	GGAGCAATCC	GCTATAAGAT
C_K21	TTGR.TCGCA	TGATCKAAAC	ATCAAAGGAT	..TTTTCTTTGGAAAATTCC	ACTTTGAGAT
C_estertheticum	TTGA.TCGCA	TGATCTTAAC	ATCAAAGGAA	..TTT..TTCGG..AATTTTC	ACTTTGAGAT
C_botulinum_A	AGAA.TCGCA	TGATTTTTCTT	ATCAAAGATT	..T.....ATT..	GCTTTGAGAT
C_sporogenes	AGAA.TCGCA	TGATTTTTCTT	ATCAAAGATT	..T.....ATT..	GCTTTGAGAT
C_argentinense	AAGG.TCGCA	TGACTTTTAT	ACCAAAGGAG	..T.....AATCC	GCTATGAGAT
C_subterminale	AAGG.TCGCA	TGACTTTTAT	ACCAAAGGAG	..T.....AATCC	GCTATGAGAT
C_tetanomorphum	TTTT.CCGCA	TGAAAAACTA	ATCAAAGGAG	..T.....AAT.C	GCTTTGAGAT
C_pasteurianum	AGTT.TCACA	TGGAGCTTTA	ATTAAAGGAG	..T.....AATCC	GCTTTGAGAT
C_collagenovorans	TTGA.TCGCA	TGGTCGAAAT	ATTAAAGGAG	..T.....AATCC	GCTTACAGAT
C_histolyticum	TTTA.ATGCA	TGTTAGAAAG	ATTAAAGGAG	CAATCC	GCTTTGAGAT
C_tyrobutyricum	AGTT.TCACA	TGGAATTTGG	ATGAAAGGAG	..T.....AATTC	GCTTTGAGAT
C_tetani	GGTT.TCGCA	TGAAACTTTA	ACCAAAGGAG	..T.....AATCT	GCTTTGAGAT
C_barkeri	GACA.TCGCA	TGGTGTT...	.TTAATGAAA	ACTCCGGT	GCCATGAGAT
C_thermocellum	GGCA.TCGTC	CTGTTAT...	.CAAAGGAGA	AATCCGGT	...ATGAGAT
Pep_prevotii	AGTC.TCGCA	TGGNGTTATC	ATCAAAGA..	TTTATC	GGTGTAAGAT
C_innocuum	ACGGAGCGCA	TGCTCTGTAT	ATTAAAGCGC	CCTTCAAGGCGTGAAC....	ATGGAT
S_ruminantium	AGTTTCCGCA	TGGGAGCTTG	ATTAAAGATG	GCCTCTACTTGTAAGCTATC	GCTTTGCGAT

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

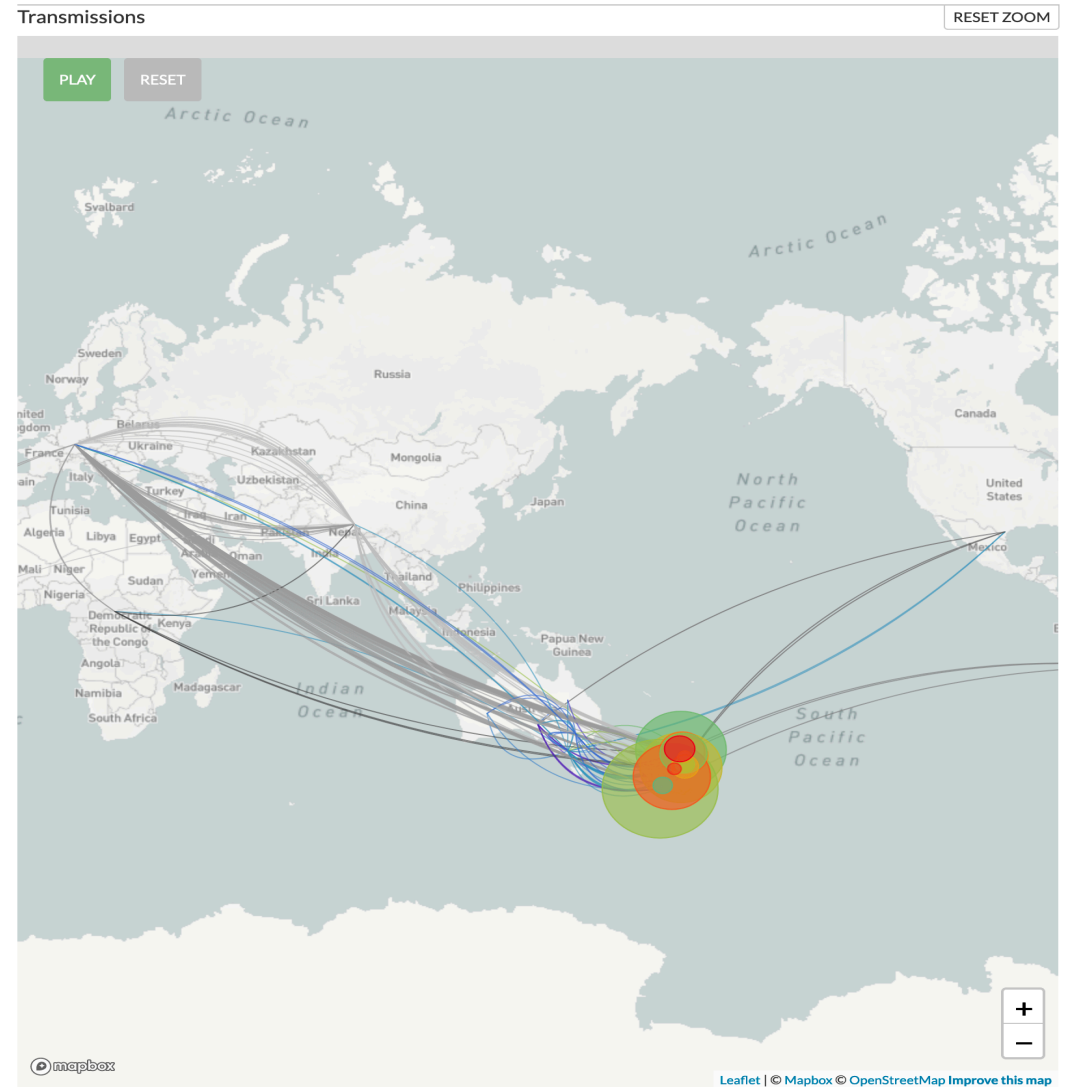
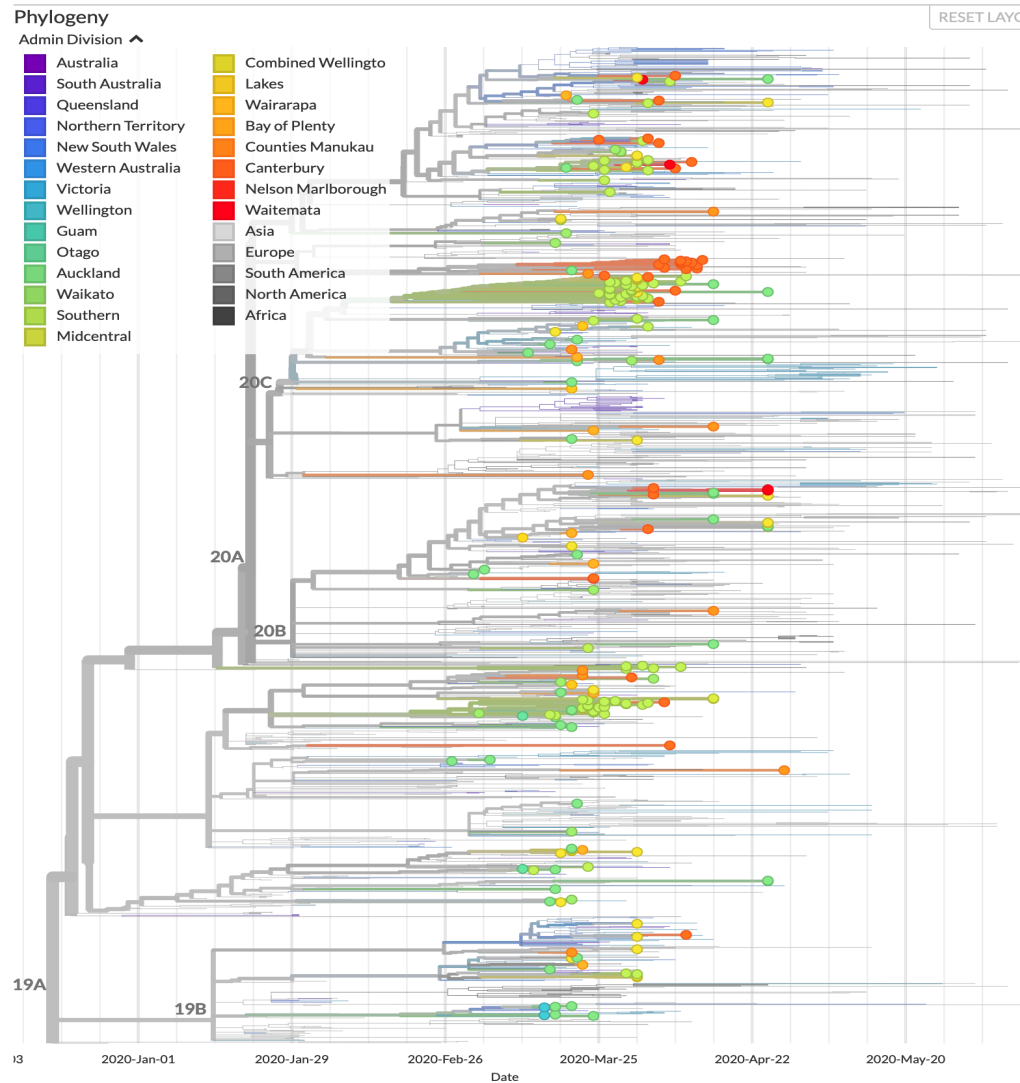
Phylogenetic Relationships of Humans and Other Primates



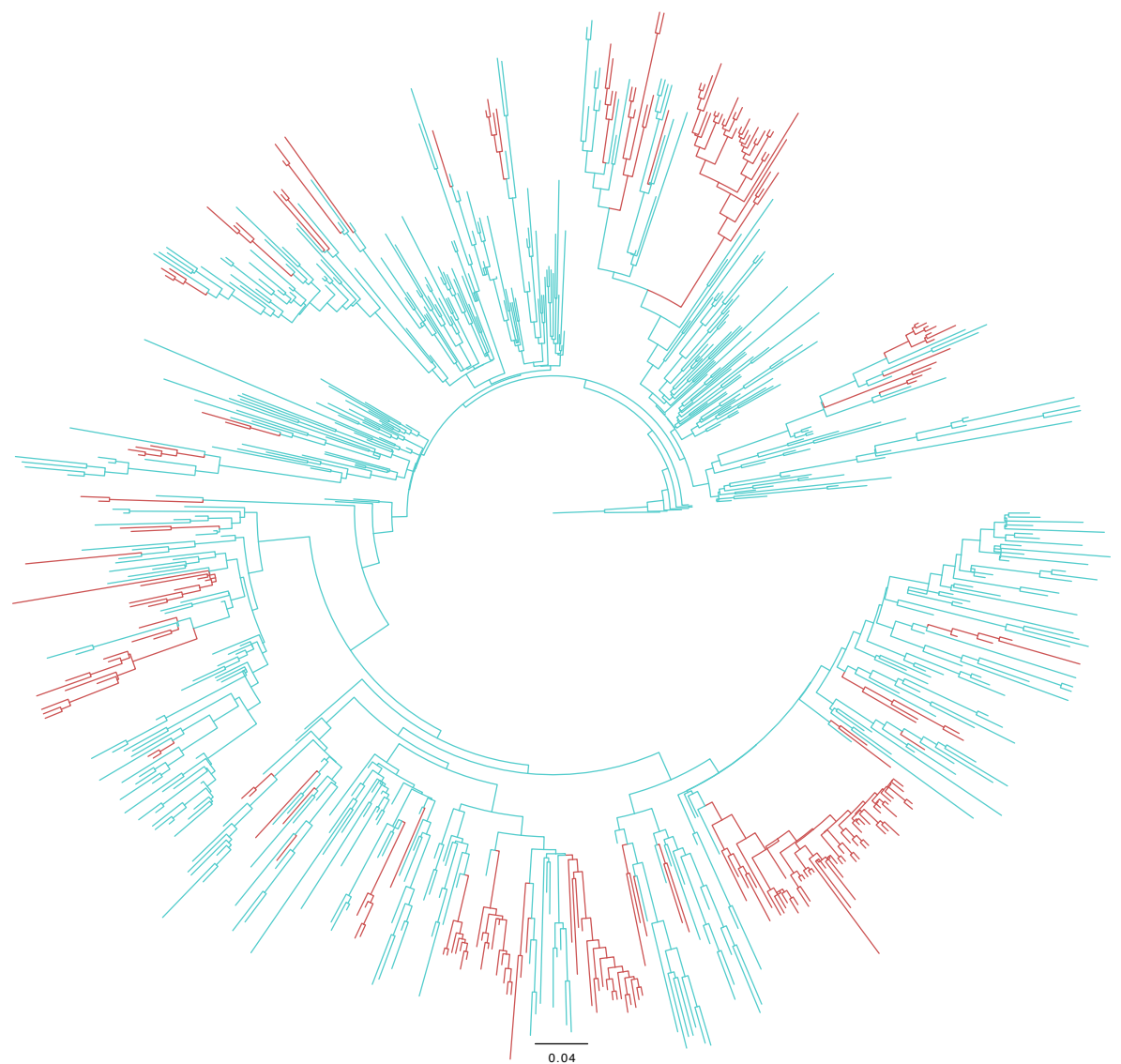
Phylodynamics of COVID-19

- 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

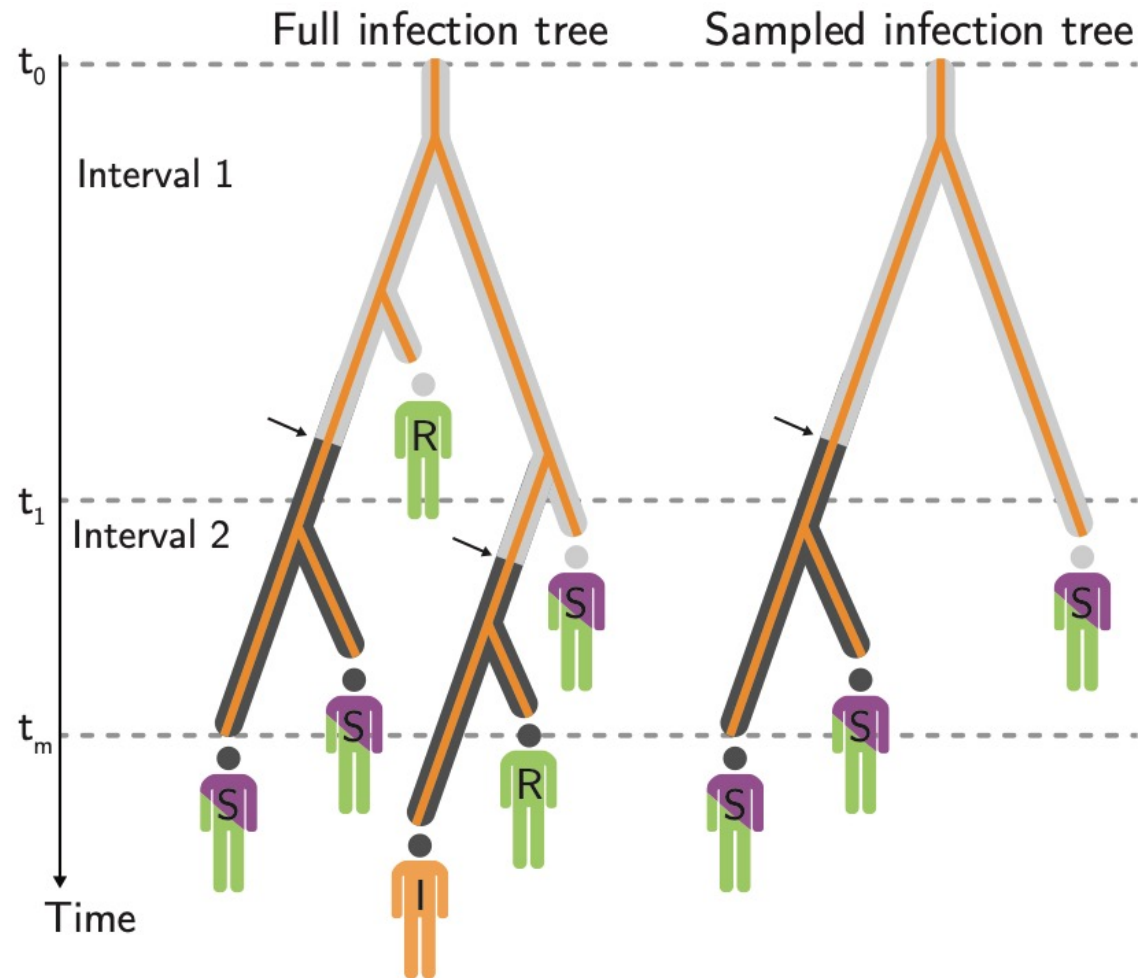
Nextstrain.org provides an excellent global overview



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



A basic phylodynamic model



Two locations: World and NZ

Parameters

Migration rates

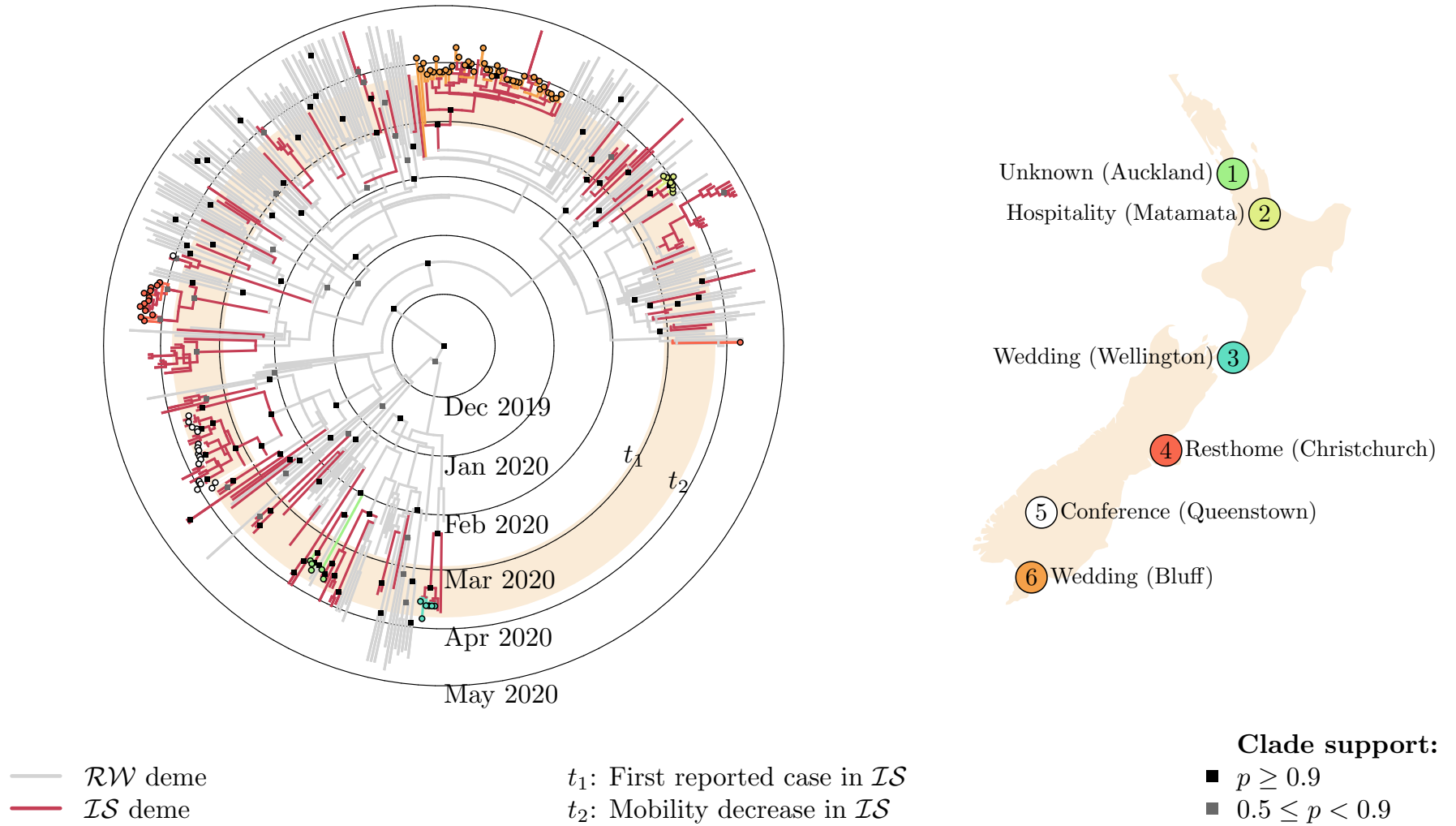
Transmission rates ("birth")

Removal rates ("death")

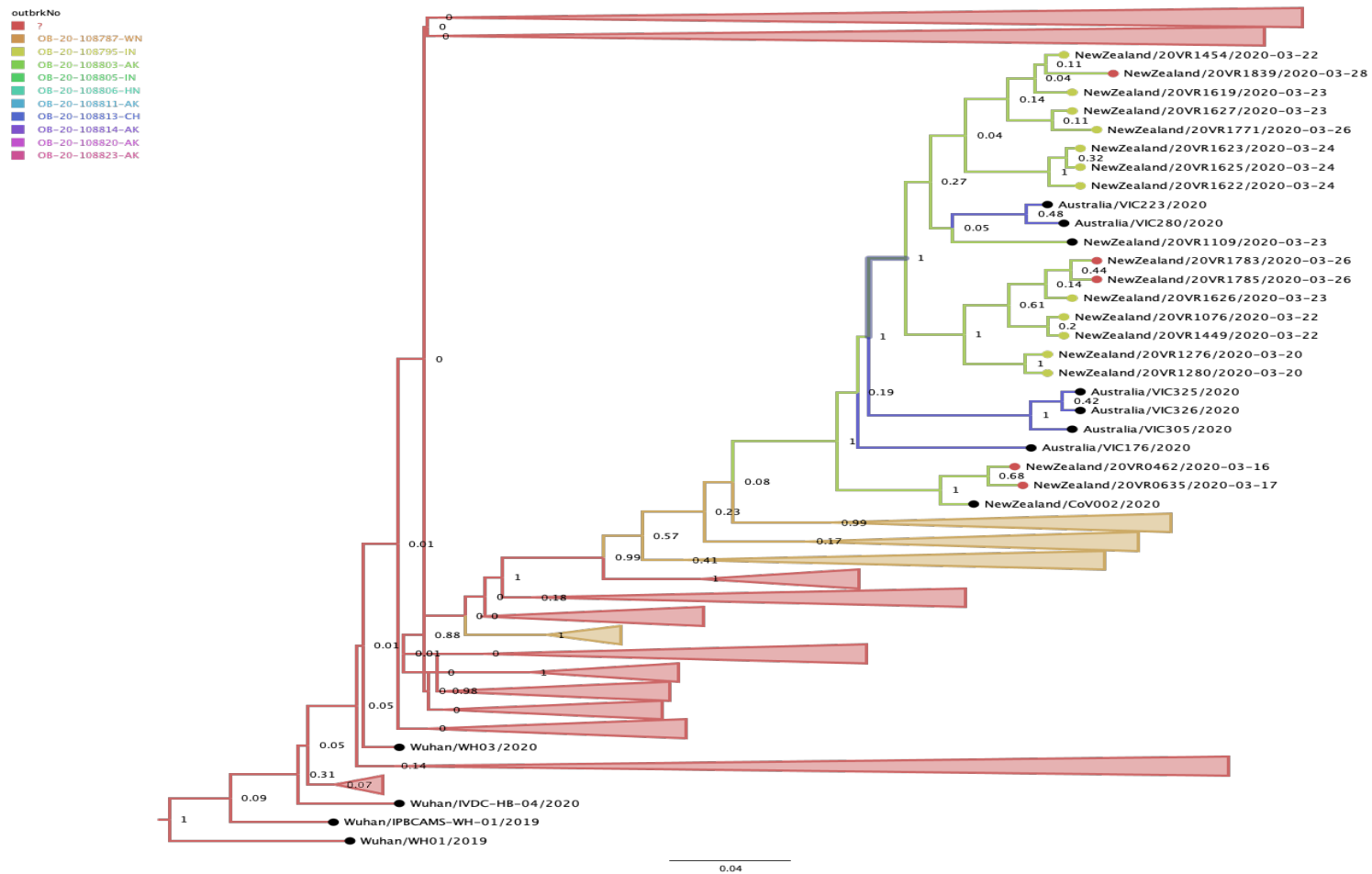
Sampling rates

Parameters can differ in different intervals

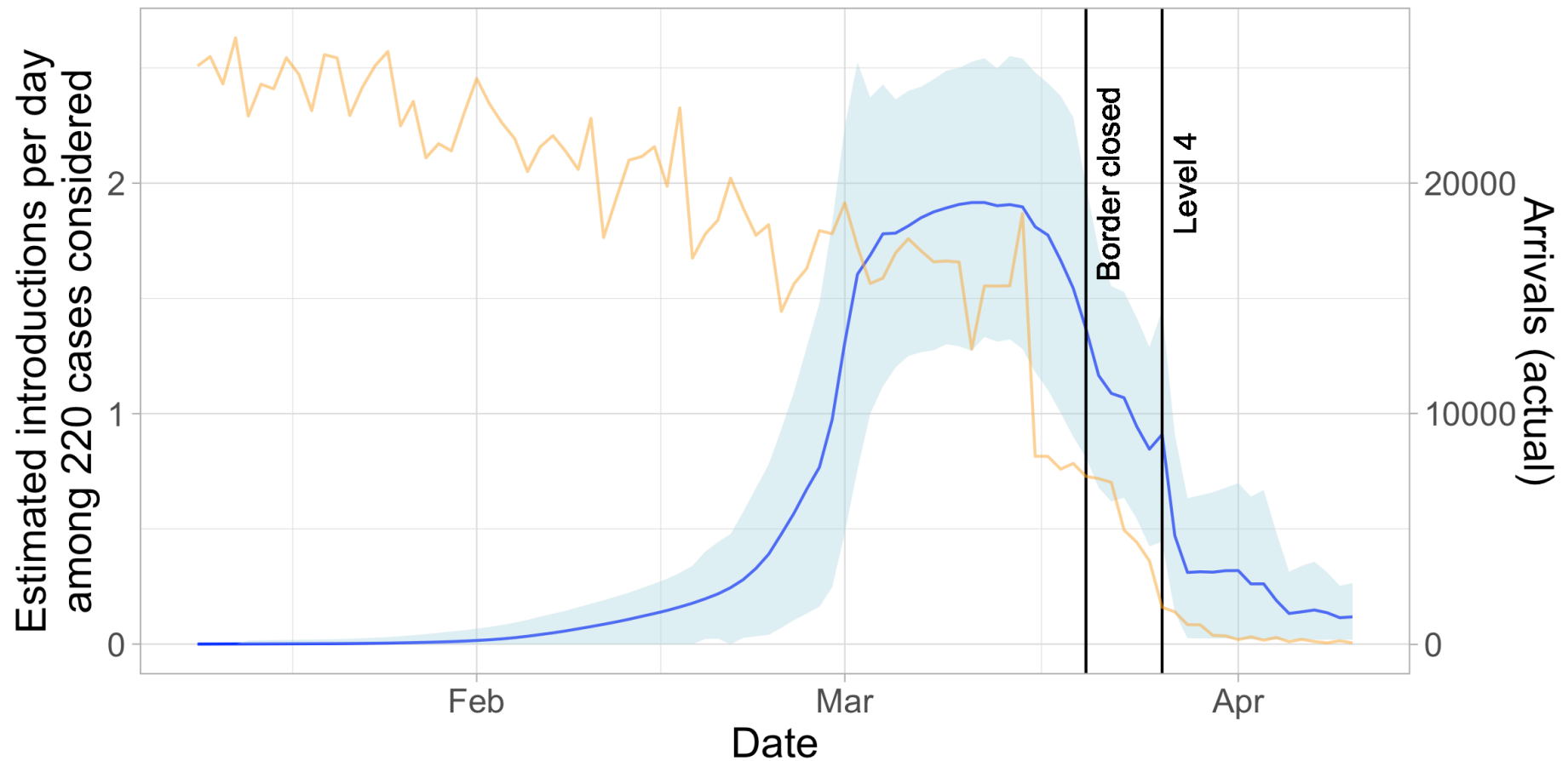
Major outbreaks in New Zealand



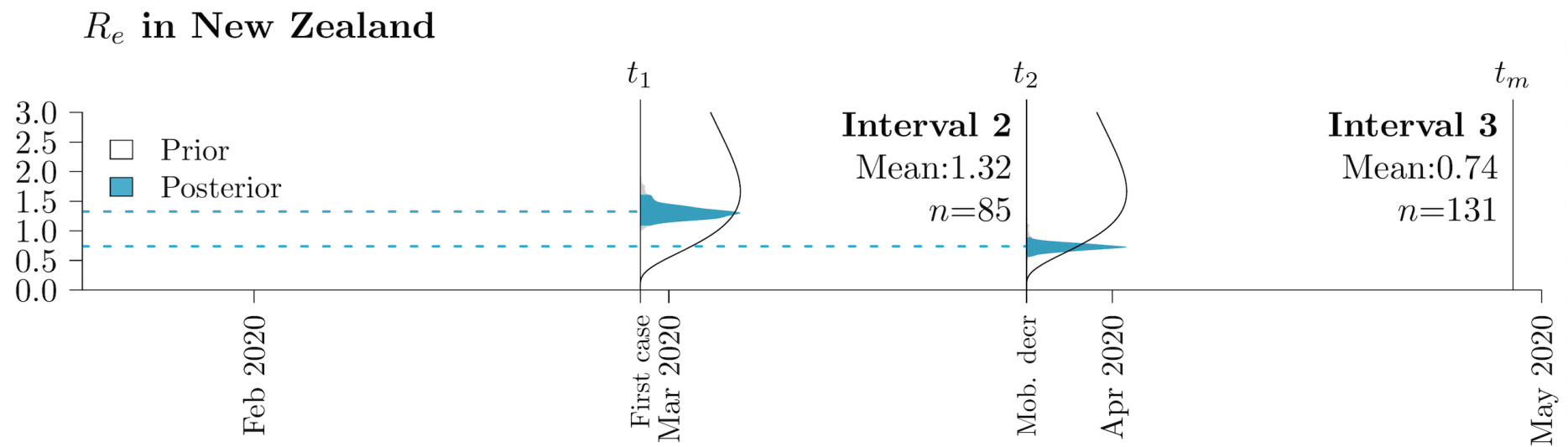
Hereford conference – TMRCA 25 Feb-2 Mar



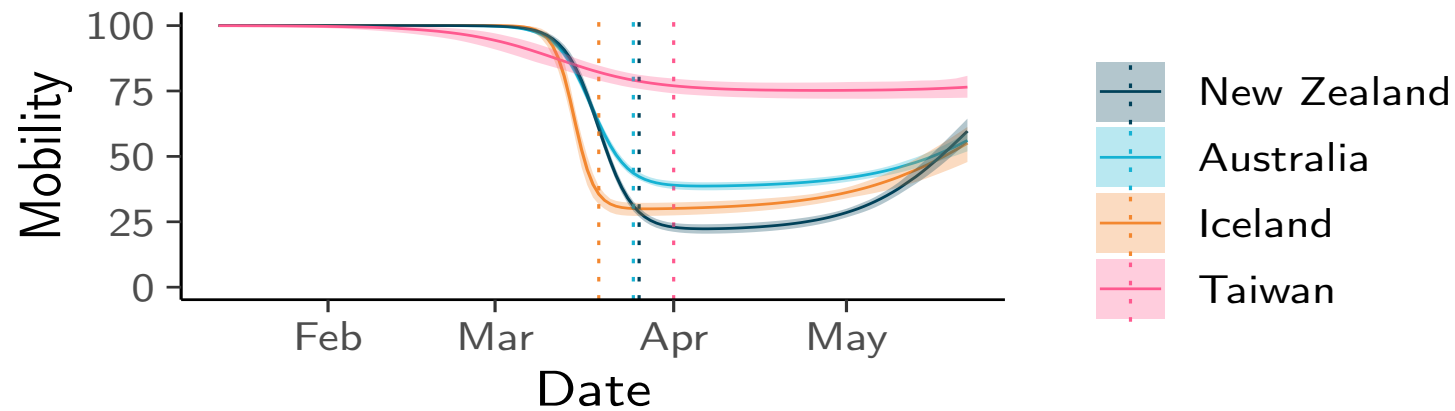
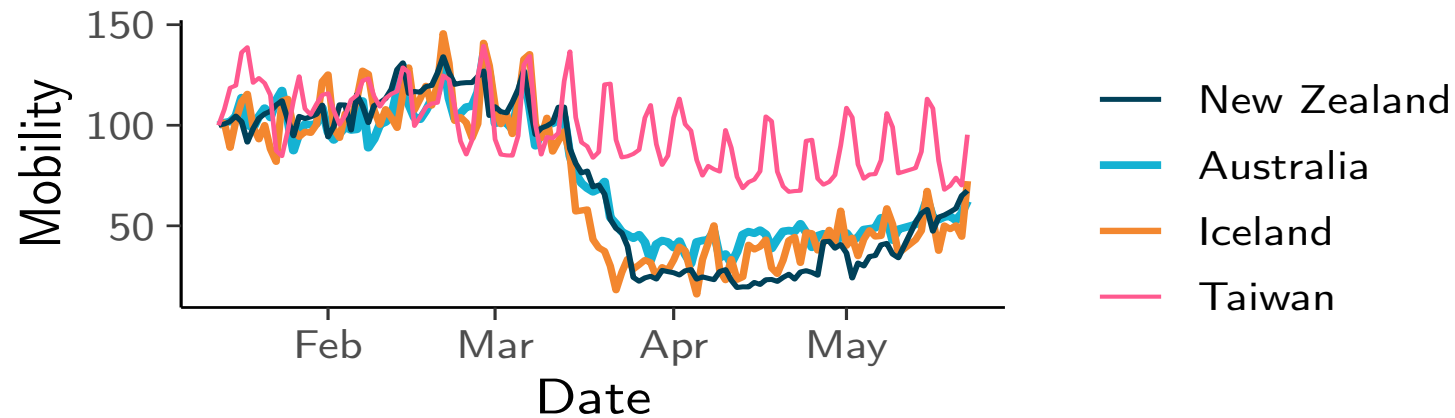
When were cases introduced to New Zealand?



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>