Themistoklis Konstantinopoulos, s210232

1: Introduction to monkeypox virus: genome and biology

Monkeypox virus (MPV) is an orthopoxvirus with an approximately 197,000-bp long genome. MPV is a discrete species exhibiting differences from other orthopoxviruses pathogenic to humans. It probably evolved from an ancestor cowpox-like virus. Recombination between species in the orthopoxviruses family is observed, as in the case of the Vaccinia virus. It replicates in the cytoplasm of cells (Shchelkunov et al., 2002).

Inverted terminal repetitions (ITR), short tandem repeats (STP), and hairpins are found at the end of the monkeypox genome. These ITRs are 6379 bp long. The complete genome consists of 190 open reading frames, each containing more than 60 amino acid residues. The virus contains all essential genes for orthopoxviruses and they are located in the central region of the genome, which is the most conserved. These genes are involved in host range, cell proliferation, growth rate, and immune evasion. The genome encodes only one full-size Kelch protein (PV-GRI G3L) in comparison to other Orthopoxvirus species which encode more. Three of the ORFs are coding the proteins F13L which is necessary to form the extracellular virus, K4L, dispensable for replication in tissues, and a lysophospholipase homologous protein (Shchelkunov et al., 2002).

Orthopoxviruses are large and complex. The shape of the capsule is brick-like with sizes ranging from 220 to 450nm and 140 to 260nm in length and width respectively. Their core consists of core fibrils and nucleoprotein complexes. The 'palisade layer' engulfs the membrane. The whole structure is enveloped inside an outer envelope and a membrane which is made of surface tubules. The empty spaces inside the virion are called lateral bodies (figure 1) (Sklenovská, 2020). The virus may or may not be inside an outer lipoprotein envelope depending on the exocytosis strategy of the virus. The mature virus contains at least 80 proteins (Resch et al., 2007).

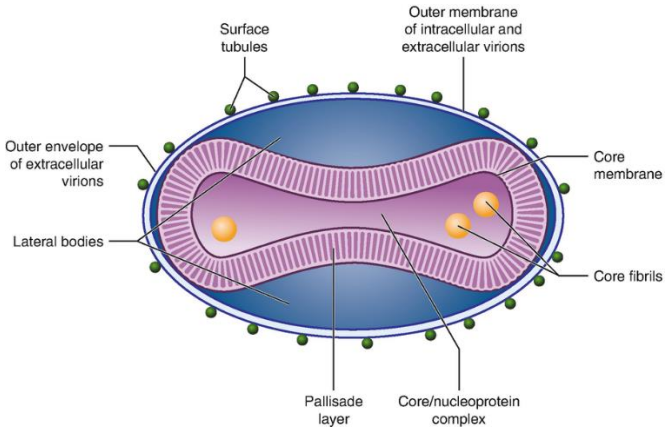

Figure 1. Schematic representation of orthopoxviruses morphology (Sklenovská, 2020)

As mentioned before, the virus replicates in the cytoplasm of the hosts' cells. After entering the cell, the virus undergoes two uncoating events. Then, the DNA is replicated, and recombination takes place. Early mRNA that is released after the first uncoating is modifying the host's membrane, growth factors, and host responses. Intermediate mRNA, which is released after the replication, translates into late transcription factors. These factors induce the production of enzymes responsible for the transcription system and late viral proteins. In the end, morphogenesis takes

place, and new viruses are released. Nuclear factors and Golgi organelle also contribute to this process (figure 2) (Resch et al., 2007).

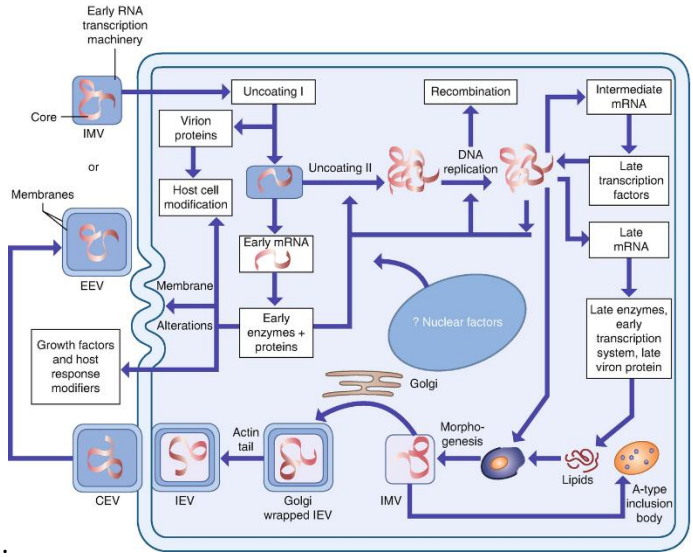

Figure 2. Life cycle of orthopoxviruses (Sklenovská, 2020)

MPV has a wide range of mammal hosts (Reynolds et al., 2010). Thus, it can maintain a reservoir in wild animals while sporadically causing human disease. Additionally, since human vaccination plans don't include wild animals, the virus escaped the pox eradication plan (Giulio & Eckburg, 2004). In monkeys, around 70% of the hosts are asymptotic carriers (Magnus et al., 1959).

The virus can be transmitted from human to human and can undergo zoonosis (i.e., be transmitted from animals to humans). The virus particles can be found in respiration droplets and body fluids and thus can be transmitted by the touch of infected fluids by soft tissue (like eyes, inner nose, or mouth). The virus incubation period is 10 to 14 days, and the disease lasts from two to four weeks. Symptoms, in the beginning, include muscle pain, swelled lymph nodes, fever, headache, and then, a rash appears. Itching of the rash areas can lead to an eruption. Lesions can be found all over the body and they are particularly abundant and large on the palms and the soles (CDC, 2022).

2: Collection of data sets

The NCBI (GenBank) virus database was used to find and collect the samples of interest. The search followed the criteria: species: Monkeypox virus, complete genomes and viruses found in all hosts. Twenty-five samples with a collection date before May 2022 were selected (table 1). Additionally, six samples from the recent outbreak were retrieved (collection date May 2022) (table 2). The complete genome of a Vaccinia virus strain was retrieved to be used as an outgroup (table 3). In two cases where the collection month and year were known but the exact date was not specified, the date was set as the 15th of the respective month (samples MK783032.1 and ON563414.2).

Table 1. Samples including complete genome sequences of Monkeypox Virus that were collected before the 2022 outbreak.

Accession	*Geo_Location*	*Host*	*Collection_Date*
JX878407.1	Democratic Republic of the Congo	*Homo sapiens*	09-10-06
JX878411.1	Democratic Republic of the Congo	*Homo sapiens*	30-11-06
JX878413.1	Democratic Republic of the Congo	*Homo sapiens*	14-12-06
JX878415.1	Democratic Republic of the Congo	*Homo sapiens*	26-12-06
JX878417.1	Democratic Republic of the Congo	*Homo sapiens*	14-12-06

JX878418.1	Democratic Republic of the Congo	*Homo sapiens*	02-01-07
JX878421.1	Democratic Republic of the Congo	*Homo sapiens*	22-03-07
JX878422.1	Democratic Republic of the Congo	*Homo sapiens*	20-03-07
JX878424.1	Democratic Republic of the Congo	*Homo sapiens*	25-03-07
JX878426.1	Democratic Republic of the Congo	*Homo sapiens*	27-05-07
JX878427.1	Democratic Republic of the Congo	*Homo sapiens*	25-05-07
JX878429.1	Democratic Republic of the Congo	*Homo sapiens*	04-09-07
MK783028.1	Nigeria: Rivers State	*Homo sapiens*	09-11-17
MK783029.1	Nigeria: Rivers State	*Homo sapiens*	06-12-17
MK783030.1	Nigeria: Rivers State	*Homo sapiens*	30-11-17
MK783031.1	Nigeria: Rivers State	*Homo sapiens*	09-11-17
MK783032.1	Nigeria: Rivers State	*Homo sapiens*	15-11-17
MN648051.1	Israel	*Homo sapiens*	04-10-18
MN346690.1	Cote d'Ivoire	*Pan troglodytes verus*	13-03-17
MN346693.1	Cote d'Ivoire	*Pan troglodytes verus*	08-04-17
MN346694.1	Cote d'Ivoire	*Pan troglodytes verus*	28-03-17
MN346696.1	Cote d'Ivoire	*Pan troglodytes verus*	03-04-17
MN346698.1	Cote d'Ivoire	*Pan troglodytes verus*	24-01-17
MN346700.1	Cote d'Ivoire	*Pan troglodytes verus*	31-01-17
MN346702.1	Cote d'Ivoire	*Pan troglodytes verus*	11-05-18

Table 2. Monkeypox Virus sequences from 2022 outbreak retrieved from GenBank.

Accession	*Geo_Location*	*Host*	*Collection_Date*
ON563414.2	USA: MA	*Homo sapiens*	15-05-22
ON568298.1	Germany	*Homo sapiens*	19-05-22
ON585031.1	Portugal	*Homo sapiens*	15-05-22
ON585033.1	Portugal	*Homo sapiens*	15-05-22
ON585035.1	Portugal	*Homo sapiens*	15-05-22
ON585038.1	Portugal	*Homo sapiens*	15-05-22

Table 3. Information regarding the outgroup selected for the current study.

Accession	*Species*	*Country*	*Host*	*Collection_Date*
OK422495.1	Vaccinia virus	India	*Bubalus bubalis*	

The sequences were downloaded as fasta files. The selection fields for naming the sequences were accession number and collection date. The merge of the three datasets was made in Linux environment by using the following command:

```
cat file1.fasta file2.fasta file3.fasta > file4.fasta
```

3. Model Selection

File1.fasta (table 1 samples) and file4.fasta (all samples) were aligned by using the mafft software (Katoh & Standley, 2013) with the following commands in Linux environment:

For file1:

```
mafft --auto monkeypox_dataset1.fasta > old.mafft_aligned.fasta
```

For file4:

```
mafft --auto monkeypox_dataset3.fasta > new.mafft_aligned.fasta
```

There was an attempt to align the sequences by using clustalw multi-sequence alignment tool (Thomson et al., 1994) but the computational time required sets this method not applicable. Nevertheless, the command that was used was:

```
Clustalw -infile=monkeypox_dataset3.fasta -
output=clustalw_aligned.fasta -align -type=DNA
```

The software jmodeltest (Posada, 2008) was used to calculate maximum likelihood and Akaike information criterion values for both datasets. Likelihood scores for 56 different models were calculated. A fixed BIONJ-JC base tree and seven substitution sets were used.

AIC, deltaAIC, and weight values were calculated as in equations 1-3.

$$AIC = -2lnL + 2K \quad \text{eq. 1}$$

$$\Delta AIC_i = AIC_i - \min(AIC) \quad \text{eq.2}$$

$$weight_i = \frac{\exp(-0.5\Delta AIC_i)}{\sum_i \exp(-0.5\Delta AIC_i)} \quad \text{eq.3}$$

Where K, the number of parameters.

Dataset 1 (old samples)

The following table summarizes the best five models, based on AIC values. The best model is TPM1uf+G with a weight value of 0.3763. The next best models are TVM+G,TIM1+G, and TPM1uf+I+G with weight values of 0.1586, 0.1388, and 0.1382 respectively.

Table 4. Likelihood and AIC values for best fitted models in old sequence dataset.

Model Name	*-ln(L)*	*p*	*AIC*	*deltaAIC*	*weight*	*Cumulative weight*
TPM1uf+G	285031.0197	53	570168.0395	0.0	0.3763	0.3763
TVM+G	285029.8838	55	570169.7677	1.7282	0.1586	0.5349
TIM1+G	285031.0173	54	570170.0346	1.9951	0.1388	0.6737
TPM1uf+I+G	285031.0213	54	570170.0426	2.0031	0.1382	0.8119
TVM+I+G	285029.8858	56	570171.7716	3.7321	0.0582	0.8702

Dataset 2 (old and recent samples)

The following table summarizes the best five models, as they were calculated based on their AIC values. The most probable model is the TPM1uf+G, with a weight value of 0.3683. The next best models are TIM1+G, TPM1uf+I+G and TVM+G with weight values of 0.1767, 0.1344, and 0.1266 respectively.

Table 5. Likelihood and AIC values for best fitted models in complete monkeypox sequence dataset.

Model name	*-ln(L)*	*p*	*AIC*	*deltaAIC*	*weight*	*Cumulative weight*
TPM1uf+G	315379.8534	67	630893.7068	0.0	0.3683	0.368
TIM1+G	315379.588	68	630895.176	1.4692	0.1767	0.545
TPM1uf+I+G	315379.8617	68	630895.7235	2.0167	0.1344	0.679
TVM+G	315378.9213	69	630895.8425	2.1357	0.1266	0.806
TIM1+I+G	315379.6015	69	630897.2029	3.4961	0.0641	0.87

The weight of a model shows how probable this model is to be selected as the best fitting model in each trial. For example, a weight of 0.37 indicates that 37 out of 100 times a specific model is the best fitting model.

The selected model for both cases is TPM1uf+G (even though phylogenetic tree reconstruction and analysis for all models with weight>0.10 would be appropriate due to the small number of samples and low value of best model weight). The TPMuf assumes that the base frequencies are unequal, and the six substitution rates are coupled as rAC=rGT, rAT=eCG, and rAG=rCT (Arbiza et al., 2011).

There are 53 and 67 parameters for the old and new dataset respectively, of which 12 are highly important (table 6).

Table 6. Most important prior parameters, their values and meaning.

Parameter name	Value data 1	Value data 2	description
fA	0.3359	0.3359	Adenine frequency
fC	0.165	0.1653	Cytosine frequency
fG	0.1648	0.1650	Guanine frequency
fT	0.3342	0.3359	Thymine frequency
rAC	1	1	Rate of mutation A→ C
rAG	3.40451	3.5295	Rate of mutation A→G
rAT	0.39398	0.3489	Rate of mutation A→T
rCG	0.39398	0.3489	Rate of mutation C→G
rCT	3.40451	3.5295	Rate of mutation C→T
rGT	1	1	Rate of mutation G→T
rates	gamma	gamma	Shape of rate distribution
Shape	0.0220	0.2720	Site evolution rate of heterogeneity

4: Bayesian tree of dataset 2

Bayesian reconstruction calculates the probability of the hypothesis given the data, also known as the posterior probability. This relationship can be expressed as:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \qquad \text{eq.4}$$

Where P(H|D) is the posterior probability of the hypothesis given the data, P(D|H) is the probability of the data given the hypothesis (i.e., likelihood), P(H) is the prior probability of the hypothesis, and P(D) the marginal probability of the data.

Due to difficulties in calculating these probabilities analytically or numerically, we are using Markov Chain Monte Carlo algorithms. These algorithms are using the whole parameter space (i.e., all possible parameter values) and they calculate the likelihood for the data based on these values by the computation of the product of the prior probability times the likelihood for each parameter value. Each step changes a parameter value base on the proposal distribution, which is the defined distribution for the specific parameter. If the newly calculated product of likelihood times prior probability is higher than the last step's value, then the parameter value change is accepted. Otherwise, it can be accepted by calculating the probability of these new parameter values.

In case the parameter space has many 'hills' (i.e., areas with higher and lower probability values in parameter space) the algorithm can use several chains (Metropolis Coupled MCMC MC3 analysis). One of these chains is called cold and is performing MCMC analysis finding local maximum probabilities, whilst the other chains (called heated) can travel through the parameter space 'freely'. In each step heated and cold chains may swap positions based on the proportional ratio between the product of the prior times the likelihood.

Parameters that are calculated as described above include among others the tree shape, branch lengths, nucleotide frequencies, substitution rates, and proposed tree rearrangements. The results include the maximum a posteriori (most often tree topology in MCMCMC output) and therefore we can calculate the posterior probability of a group as the frequency of the clade in sampled trees.

The aligned fasta files were converted to nexus by using seqconverter with the following command in Linux environment:

```
Seqconverter -I auto -O nexus new.mafft_aligned.fasta >
alignmaf_3.nexus

Seqconverter -I auto -O nexus old.mafft_aligned.fasta >
alignmaf_1.nexus
```

Subsequently, the MrBayes software was opened by the command:

```
mb
```

and the second dataset was loaded by using the command:

```
execute alignmaf_3.nexus
```

The model that was selected was GTR, which allows six independent substitution rates (Lecocq et al., 2013). Based on the model selection in section 4 TPM1uf+G the parameter values were set according to table 6. The following commands were used:

```
lset nst=6 rates=gamma ngammacat=4

prset shapepr=fixed(0.2720) statefreqpr=dirichlet(0.3359, 0.1653, 0.1650,
0.3359)  revmatpr = dirichlet(1.0000, 3.5295, 0.3489, 0.3489,
3.5295,1.0000)

outgroup OK422495.1
```

Finally, the model running duration was set to 5000000 generations but was interrupted around 1800000 because it reached convergence. Three chains (two warm, one cold) were used. The diagnostic frequency was defined for every 5000 generations. The sd value was 0.008. The following command initiated the run:

```
mcmc ngen=5000000 samplefreq=1000 nchains=3 diagnfreq=5000
```

After the analysis finished, an *end* was added at the end of the two .t files and the following commands were used to get a summary of the results:

```
sumt contype=halfcompat showtreeprobs=no relburnin=yes burninfrac=0.25
sump relburnin=yes burninfrac=0.25
```

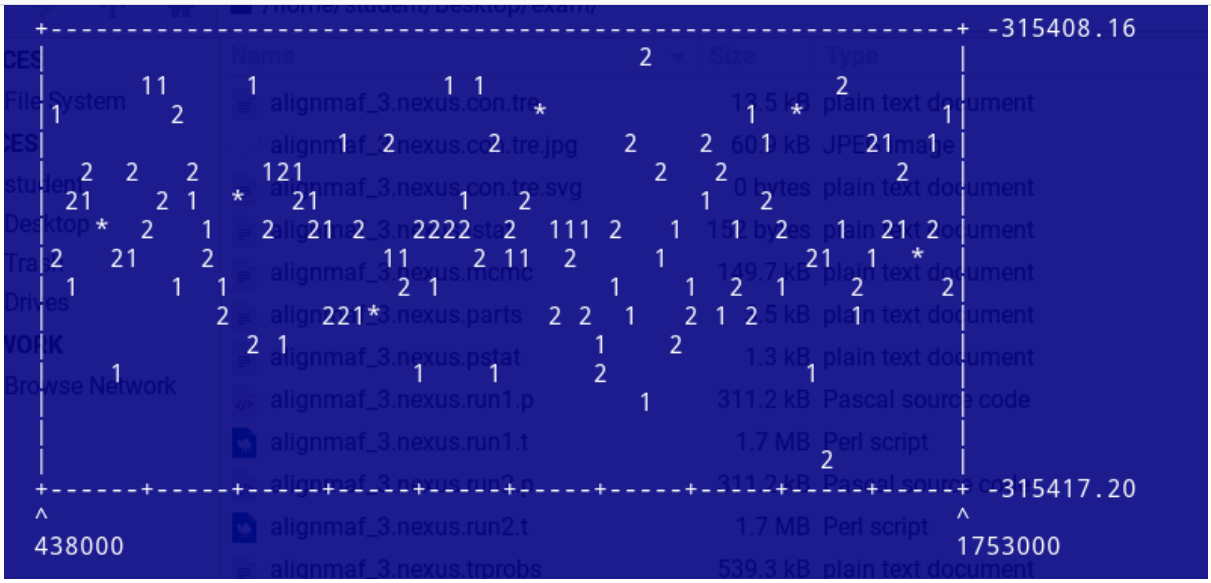

Figure 3. ‚Rough plots' of number of generations (x-axis) and log probability (y axis) of the observed data. There is no trend.

The estimated parameters and information regarding their convergence are summarized in table 7. Overall, the model has reached its converged state and the results of the GTR model do somehow confirm the selection of the TPM1uf model (figures 4-6).

Table 7. Summary statistics for important parameters of the selected model. HPD: highest probability density, ESS: effective sample size, PSRF: Potential Scale Reduction Factor (close to 1 indicates converged results). TL: tree length, r(N<->N): nucleotide substitution rate, pi(N): nucleotide frequency. All values rounded to three decimal places.

			95% HPD Interval				
Parameter	Mean	Variance	Lower	Upper	Median	avg ESS	PSRF

TL	0.036	0.000	0.035	0.037	0.036	1207.950	1.000
r(A<->C)	0.098	0.000	0.091	0.105	0.098	1001.330	1.000
r(A<->G)	0.366	0.000	0.354	0.379	0.366	781.170	1.000
r(A<->T)	0.036	0.000	0.032	0.039	0.036	947.830	1.001
r(C<->G)	0.036	0.000	0.028	0.043	0.036	941.390	1.000
r(C<->T)	0.357	0.000	0.345	0.371	0.357	699.650	1.000
r(G<->T)	0.106	0.000	0.098	0.114	0.106	949.710	1.000
pi(A)	0.336	0.000	0.334	0.338	0.336	299.450	1.001
pi(C)	0.165	0.000	0.164	0.167	0.165	501.380	1.000
pi(G)	0.165	0.000	0.163	0.166	0.165	323.140	1.001
pi(T)	0.334	0.000	0.332	0.336	0.334	386.200	1.000

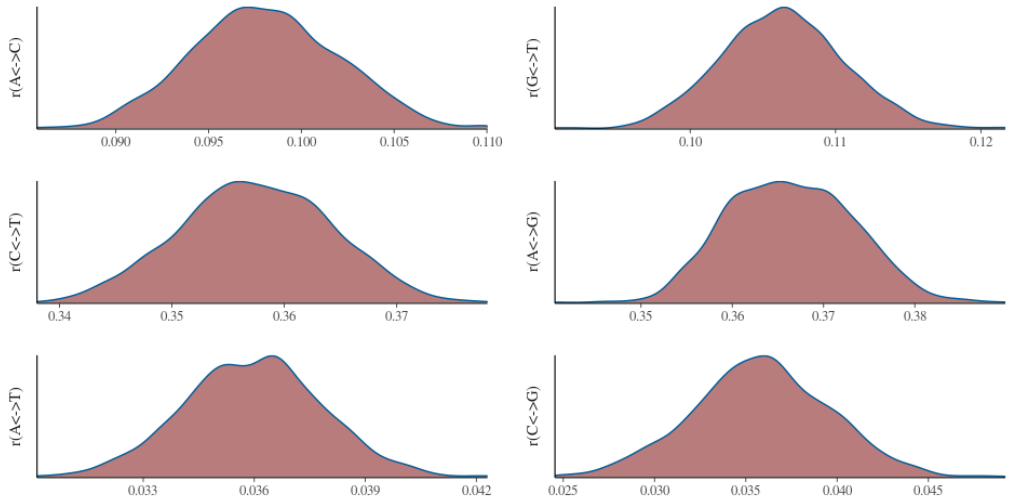

Figure 4. posterior probability of transition mutation rates AG, CT, and transversions AT, CG, GT, AC. As expected, transitions occur in higher frequencies than transversions.

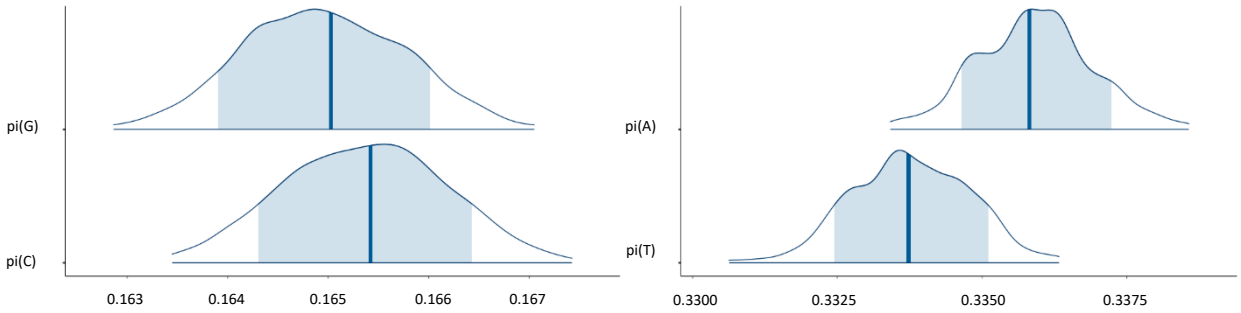

Figure 5. posterior probability of nucleotide frequencies

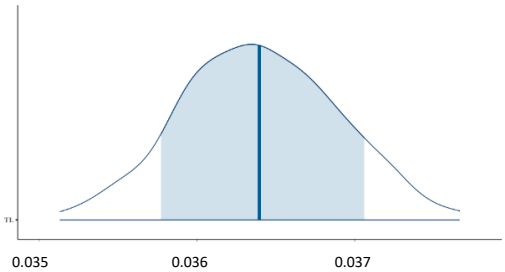

Figure 6. Posterior distribution of Tree length These plots were generated in R by using the package bayesplot (Gabry & Mahr, 2022) (APENDIX 1 for code)

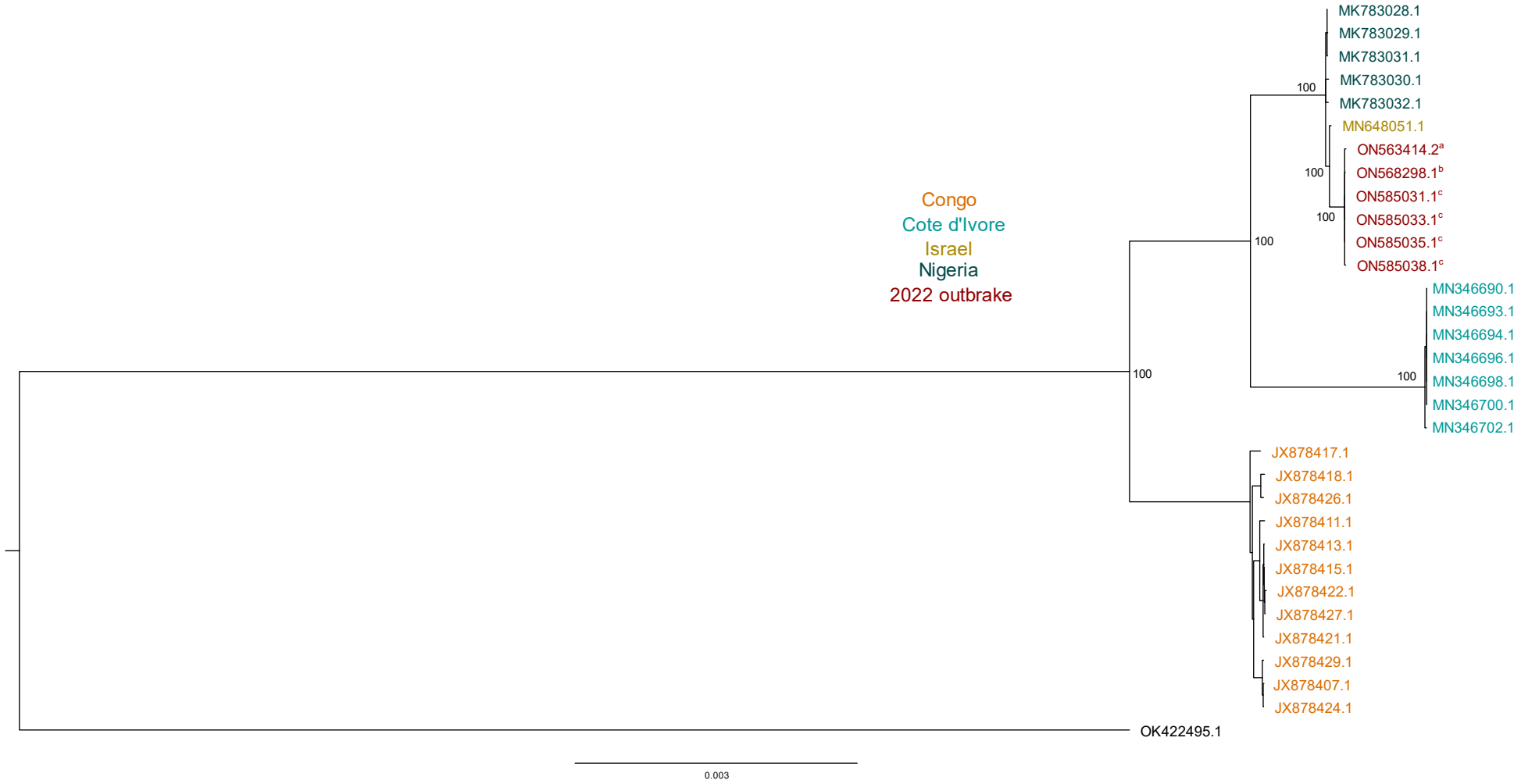

Figure 7. Phylogenetic tree based on MCMCMC analysis by using model 1 parameters as described in text. Location of 2022 samples: a: USA, b: Germany, c: Portugal. Node labels indicate the posterior probability for each clade.

The Bayesian inference analysis revealed two main clades. The first contains all samples from Congo, whilst the second consists of the rest of the samples. Therefore, it is showed that the recent outbreak comes from the western African clade, and it seems related to the strain that caused the case in Israel in 2018. Pangolin samples from Cote d'Ivoire are located in the western clade, but there is no relation between them and the recent samples. It seems like a strain from Nigeria can travel overseas.

The tree was examined in Rstudio by using the packages ape and apetree (Paradis & Schliep, 2019) (APPENDIX 1). The branch (edge) that links the recent outbreak with the Israel case was identified and the branch length was extracted. Based on this analysis, the branch length is 2.507813e-04 substitutions per site. The Monkeypox DNA is around 200000 bases long and is a double-stranded molecule, therefore the total nucleotide substitutions are approximately 100 for the whole genome.

5: Bayesian clock model tree for dataset 1

Beauti and the Beast software was used to calculate the rate of the molecular clock in dataset 1 (Bouckaert et al., 2019). At first, the nexus files generated in step 3 were loaded to beauti, which was used to set up the model. The tip dates were extracted from the dataset. One partition was used for the whole dataset. A gamma site model was selected, with a GTR substitution model and the values for substitution rates were set according to table 6.

A strict and a relaxed log-normal clock was tested, with the default value of '1.0' and the option 'estimate' on. A birth-death skyline serial model was selected, although there was a trial to run a contemporary version with dimension value for rho (ρ) changed, but with no luck as a bug was stopping the analysis. The priors were set according to table 8. Priors represent our initial beliefs about the true value of the parameters. The generated file was saved as clockanalysis.xlm in the same folder as the rest of the exam files. The duration of the run was set to 50000000 generations.

The interfaced version of BEAST was opened by using the command:

```
Beast -options
```

Table 8. Priors' setup for BEAST analysis

Parameter	Distribution shape	Value
Infectious period	Gamma	0-inf
Reproductive number	Log-normal	0-inf
BDSKY origin*	Log-normal	1,1.25
Sampling proportion	Beta	0-1
Substitution rates	Gamma	Limits around values of table 6
Reproduction number	Gamma	Initial value 2.1 (WHO, 2020)

The molecular clock quantitative measures the number of substitutions, or mutations, which accumulate in the gene sequences over time. However, a molecular clock can only provide information regarding the proportional duration of two time periods, and it is not able to estimate the exact duration (in days) of these periods. Therefore, in studies where evolutionary history is examined and there is the possibility of sampling sequences over a specific timescale, the dates of these samples can be used to calibrate these durations into days or years.

The software Tracer (Rambaut et al., 2018) was used to check if the model is converged.

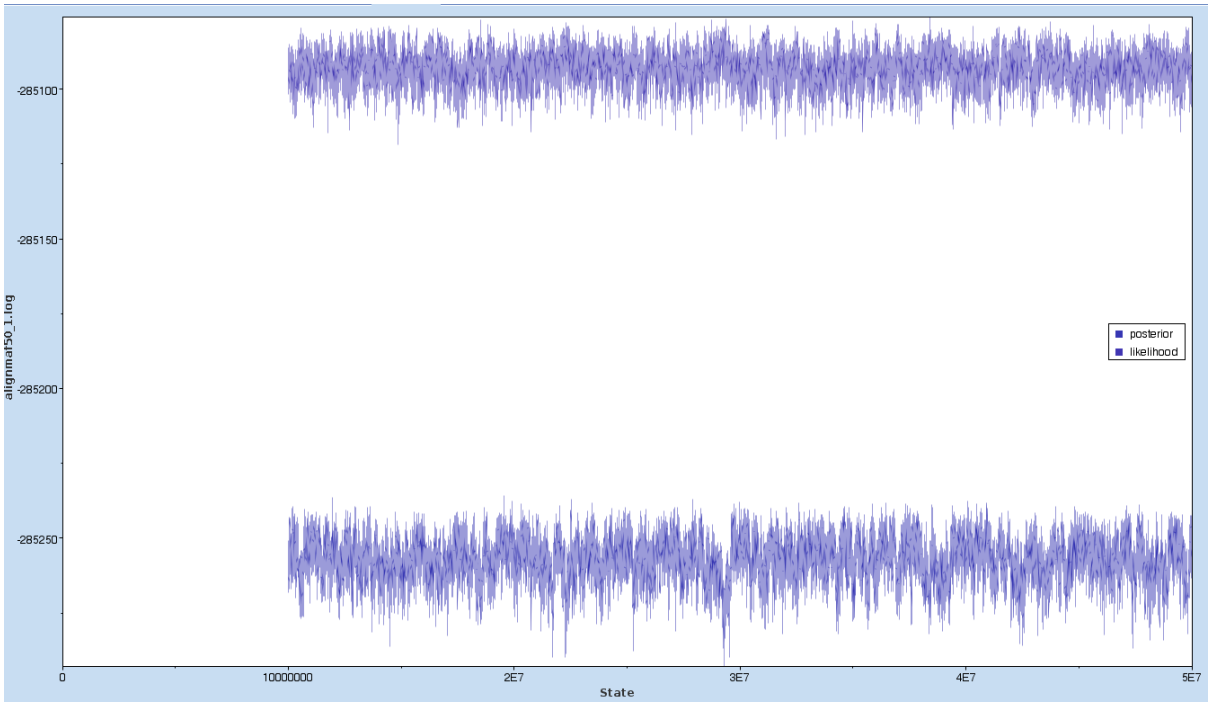

Figure 8. likelihood (top) and posterior likelihood (bottom) of the analyzed tree.

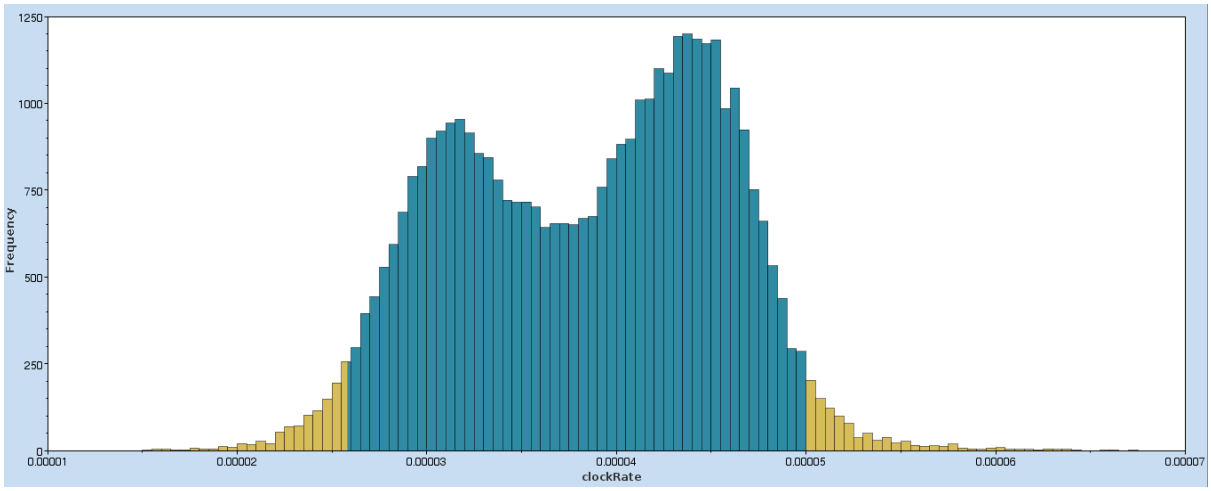

Figure 9. Posterior Distribution plot for clock rate.

Table 9. Summary statistics for the BEAST analysis.

Statistic	Mean	ESS	95% HPD interval
posterior	-2.85E+05	457	[-285270.4487, -285243.734]

likelihood	-2.85E+05	807	[-285103.0771, -285082.9325]
prior	-164.089	247	[-178.8838, -151.0574]
treeLikelihood	-2.85E+05	807	[-285103.0771, -285082.9325]
TreeHeight	7.51E+01	227	[55.5736, 101.713]
FreqA	3.36E-01	1033	[0.3339, 0.3381]
FreqC	1.65E-01	972	[0.1634, 0.1666]
FreqG	1.65E-01	904	[0.1632, 0.1665]
FreqT	3.34E-01	936	[0.3322, 0.3363]
rateAC	9.52E-01	440	[0.7605, 1.1683]
rateAG	3.43E+00	207	[3.0019, 3.8735]
rateAT	4.12E-01	471	[0.3368, 0.492]
rateCG	3.49E-01	998	[0.2504, 0.4576]
rateGT	1.08E+00	368	[0.854, 1.311]
rateCT	3.44E+00	229	[3.0023, 3.8861]
clock rate	3.84E-05	234	[2.5803E-5, 4.9994E-5]

Based on the BEAST analysis the clock rate is equal to 3.84e-5 substitutions per site per year. For the whole genome of the virus (200000bp), this is translated to around 15 substitutions per genome per year (both DNA strands).

The clock rate is lower than RNA viruses (like the COVID19 (mini project 3)), thus vaccines should provide longer immunity periods. However, the calculated rate is higher than similar viruses (e.g., the variola virus has a clock rate of around 9.0e-6 (Firth et al., 2010)). Such high mutation rates can be observed in other double-stranded DNA viruses like HPV (Firth et al., 2010).

The nature of the genetic material highly affects the clock rate. For example, RNA viruses show the highest mutation rates, whilst DNA viruses' mutation rate is lower. Additionally, the replication mechanisms and the location inside the cell the virus replicates affect the substitution frequency. Host post-processing mechanisms (as in Monkeypox) can also result in modified genomic sequences. Of course, random events, like replication errors can lead to genetic diversity. At the population level, vaccination can prevent transmission, and thus reduce the clock rates, as the virus has a limited number of susceptible hosts. Behavioural mechanisms (such as isolation, or hygiene-related actions) can act similarly, by not allowing the transmission of the virus.

6: Comparison of observed branch length below recent outbreak with estimated clock rate

The basal relative of the recent outbreak in Dataset 2 is the sample from Israel (figure 7). This sample was collected on the 4th of October 2018, i.e., 3.61 years before the date of the recent samples.

Based on the clock rate in section 5, we would expect around 54 substitutions in the whole genome, but instead, around 100 were observed. Therefore, it seems like some beneficial mutations increased the virus' asymptotic hosts and reproduction rates. There are also reports of microevolutionary events in the monkeypox virus genome and gene loss (Isidro et al., 2022).

To further investigate the nature of these mutations we could identify the locations of all differences between the 'old' and new sequences and then test for positive, negative, or neutral selection by using tools like codeml (Yang, 2007). Next, we could identify which genes are affected and try to predict how these changes will affect the biology and life cycle of the virus.

7: Conclusion

The monkeypox virus is a DNA virus with a genome of around 200,000 DNA nucleotides. Monkeypox virus has a much lower genetic variation (15 -7.5 per DNA strand- mutations per year) than RNA viruses such as SARS-CoV-2 but shows a high variation in this study. Young people have not received the vaccine for smallpox, and they are susceptible to disease. Alongside with the asymptotic hosts, the disease can travel and create epidemic events in areas where the immunity against orthopoxviruses is reduced (Beer & Rao, 2019).

Our analysis indicated the West Clade (Nigeria) as the origin of the disease. The limited number of samples that fulfilled the selection criteria indicated that pangolins in Cote d'Ivoire had nothing to do with the recent outbreak. Additionally, recent cases relate to other incidents around the globe some years ago. More samples from the western Africa clade, including non-Human hosts, can provide further insight regarding the exact event that drives this epidemic wave and what drove this evolutionary event. Some strains lose genes while staying among human populations. The recent outbreak and the accessibility in whole viral genome sequencing can provide further insight about the evolutionary mechanisms acting on the virus. The long gap in sample dates may cause some problems in the clock rate estimation. The clock rates show two picks close to each other. This could be indicating a shift between the oldest samples and the most recent ones, or it could be due to the conserved area in the genome, that faces less mutation events. Nevertheless, to increase the model accuracy, we could post-process the data files and use partitioning in them.

All files related to this exam can be accessed here: www.github.com/themiskon/molevol

References

Arbiza, L., Patricio, M., Dopazo, H., & Posada, D. (2011). Genome-wide heterogeneity of nucleotide substitution model fit. Genome biology and evolution, 3, 896-908.

Beer, E. M., & Rao, V. B. (2019). A systematic review of the epidemiology of human monkeypox outbreaks and implications for outbreak strategy. PLoS neglected tropical diseases, 13(10), e0007791.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology, 15(4), e1006650.

CDC. (2022). Center for Disease Control and Prevention. Monkeypox. Available online at: Monkeypox | Poxvirus | CDC. (Last accessed on 29.05.2022)

Di Giulio, D. B., & Eckburg, P. B. (2004). Human monkeypox: an emerging zoonosis. The Lancet infectious diseases, 4(1), 15-25.

Gabry, J., Mahr, T. (2022). "bayesplot: Plotting for Bayesian Models." R package version 1.9.0, https://mc-stan.org/bayesplot/.

Huelsenbeck, J. P., Ronquist, F., & Teslenko, M. (2015). Command reference for mrbayes ver. 3.2. 5.

Isidro, J., Borges, V., Pinto, M., Ferreira, R., Sobral, D., Nunes, A., Santos, J. D., Mixão, V., Santos, D., Duarte, S., Vieira, L., Borrego, M. J., Núncio, S., Pelerito, A., Cordeiro, R., Gomes J. P. (2022). Multi-country outbreak of Monkeypox virus: genetic divergence and first signs of microevolution. Genome Reports. Available at: https://virological.org/t/multi-country-outbreak-of-monkeypox-virus-genetic-divergence-and-first-signs-of-microevolution/806. (Last accessed 29.05.2022)

Katoh, Standley 2013 (Molecular Biology and Evolution 30:772-780). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. (Outlines version 7)

Lecocq, T., Vereecken, N. J., Michez, D., Dellicour, S., Lhomme, P., Valterova, I., ... & Rasmont, P. (2013). Patterns of genetic and reproductive traits differentiation in mainland vs. Corsican populations of bumblebees. PLoS One, 8(6), e65642.

Magnus, P. V., Andersen, E. K., Petersen, K. B., & Birch-Andersen, A. (1959). A pox-like disease in cynomolgus monkeys. Acta Pathologica Microbiologica Scandinavica, 46(2), 156-176.

Paradis E., Schliep K. (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R." Bioinformatics, 35, 526-528.

Posada, D. 2008. jModelTest: Phylogenetic Model Averaging. Molecular Biology and Evolution 25: 1253-1256.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard M. A. (2018) Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. Systematic Biology. syy032. doi:10.1093/sysbio/syy032

Resch, W., Hixson, K. K., Moore, R. J., Lipton, M. S., & Moss, B. (2007). Protein composition of the vaccinia virus mature virion. Virology, 358(1), 233-247.

Reynolds, M. G., Carroll, D. S., Olson, V. A., Hughes, C., Galley, J., Likos, A., ... & Damon, I. K. (2010). A silent enzootic of an orthopoxvirus in Ghana, West Africa: evidence for multi-species involvement in the absence of widespread human disease. The American journal of tropical medicine and hygiene, 82(4), 746.

Shchelkunov, S. N., Totmenin, A. V., Safronov, P. F., Mikheev, M. V., Gutorov, V. V., Ryazankina, O. I., ... & Moss, B. (2002). Analysis of the monkeypox virus genome. Virology, 297(2), 172-194.

Sklenovská, N. (2020). Monkeypox Virus. In: Malik, Y.S., Singh, R.K., Dhama, K. (eds) Animal-Origin Viral Zoonoses. Livestock Diseases and Management. Springer, Singapore. https://doi.org/10.1007/978-981-15-2651-0_2

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), 4673-4680.

World Health Organization. (2020). Bulletin of the World Health Organization.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24(8), 1586-1591.

Appendix 1. R code for creation of graphs and identification of branch lengths

```
library(ape)
library(apTreeshape)
library(adephylo)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(broom)
library(zoom)
#set working directory
setwd("~/Desktop/exam")
#read tree file
Monkeytree<- read.nexus("alignmaf_3.nexus.con.tre")
plot(Monkeytree, show.tip.label=TRUE)
add.scale.bar(length=0.005)
edgelabels(cex=0.8, frame = 'none')
zm()
#find branch number and extract the value from variable Monkeytree.edge.length
########################          Posterior Plots          ############################
library(magrittr)
library(tidyverse)
library(bayesplot)

setwd("~/Desktop/exam")

df = read_tsv("alignmaf_3.nexus.run1.p", skip=1)
burnin = df$Gen %>%
  max() %>%
  multiply_by(0.25) %>%
  floor()


df2 = df %>%
  filter(Gen > burnin)

mcmc_trace(df2, pars=c("pi(A)",'pi(T)'))
mcmc_trace(df2,pars = 'pi(T)')
mcmc_trace(df2,pars='pi(G)')
mcmc_trace(df2,pars='pi(C)')
color_scheme_set("mix-blue-red")
mcmc_trace(df2, pars = c("pi(A)", "pi(T)","pi(G)","pi(C)"),
         facet_args = list(ncol = 1, strip.position = "left"))

#plots of frequency posterior probability
mcmc_areas(
  df,
  pars = c("pi(A)", "pi(T)"),
  prob = 0.8, # 80% intervals
  prob_outer = 0.99, # 99%
  point_est = "mean"
)

mcmc_areas(
  df,
  pars = c("pi(G)","pi(C)"),
  prob = 0.8, # 80% intervals
  prob_outer = 0.99, # 99%
  point_est = "mean"
```

```
)

mcmc_dens(
  df2,
  pars = c("r(A<->C)",'r(G<->T)','r(C<->T)',"r(A<->G)",'r(A<->T)','r(C<->G)'),
  facet_args = list(ncol = 2, strip.position = "left"),
  prob = 0.8, # 80% intervals
  prob_outer = 0.99, # 99%
  point_est = "mean"
)

mcmc_areas(
  df2,
  pars = c("TL"),
  prob = 0.8, # 80% intervals
  prob_outer = 0.99, # 99%
  point_est = "mean"
)
```