Progress

TECE meeting 2016-08-18

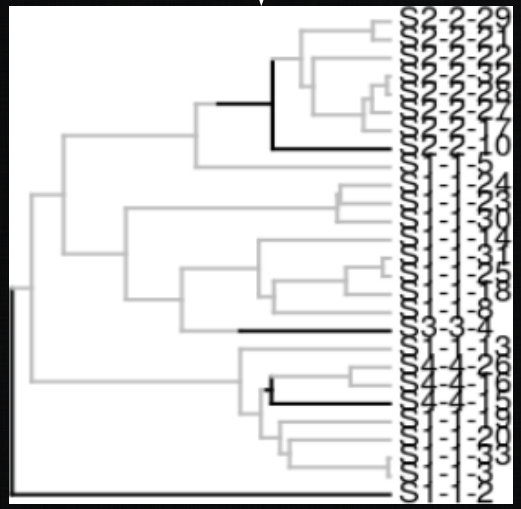
www.github.com/richelbilderbeek/Science

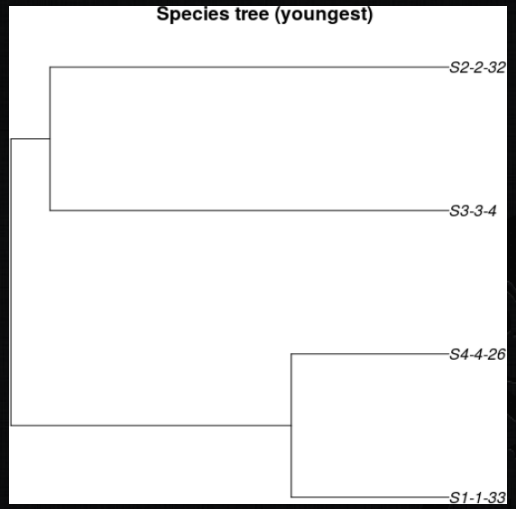
Research question

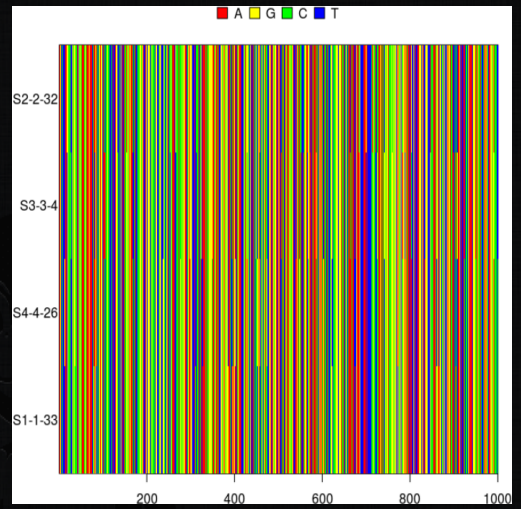
If speciation in nature takes time, what is the error made in inferring a phylogeny?

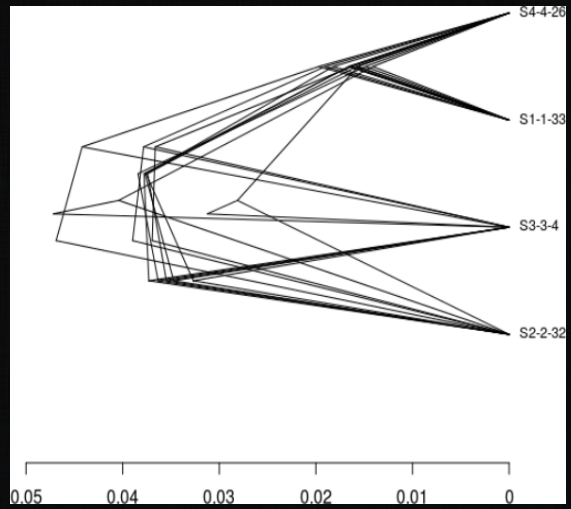
- Under which conditions does it hurt?
- Are these conditions relevant and/or realistic?
- Tool used: BEAST2, assumes speciation is instantaneous

Approach







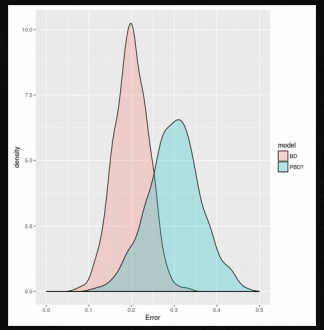


Simulate an incipient species tree

Sample species trees

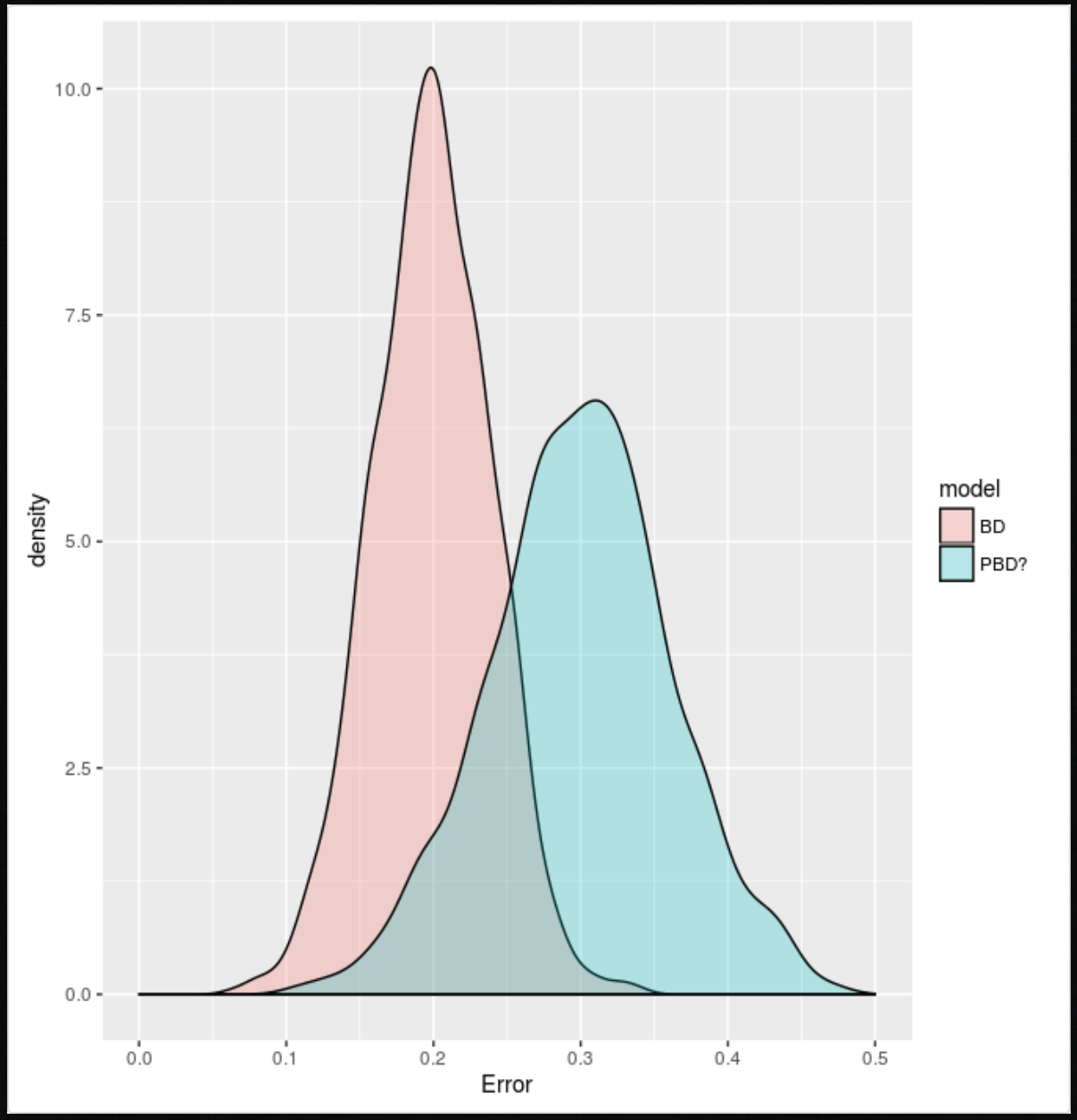
Simulate DNA alignments

Infer species tree posterior



Measure difference/error

Goal



Work in progress

- Under which parameters can I simulate?
- Measure the nLTT statistics for those parameters
- Finish article

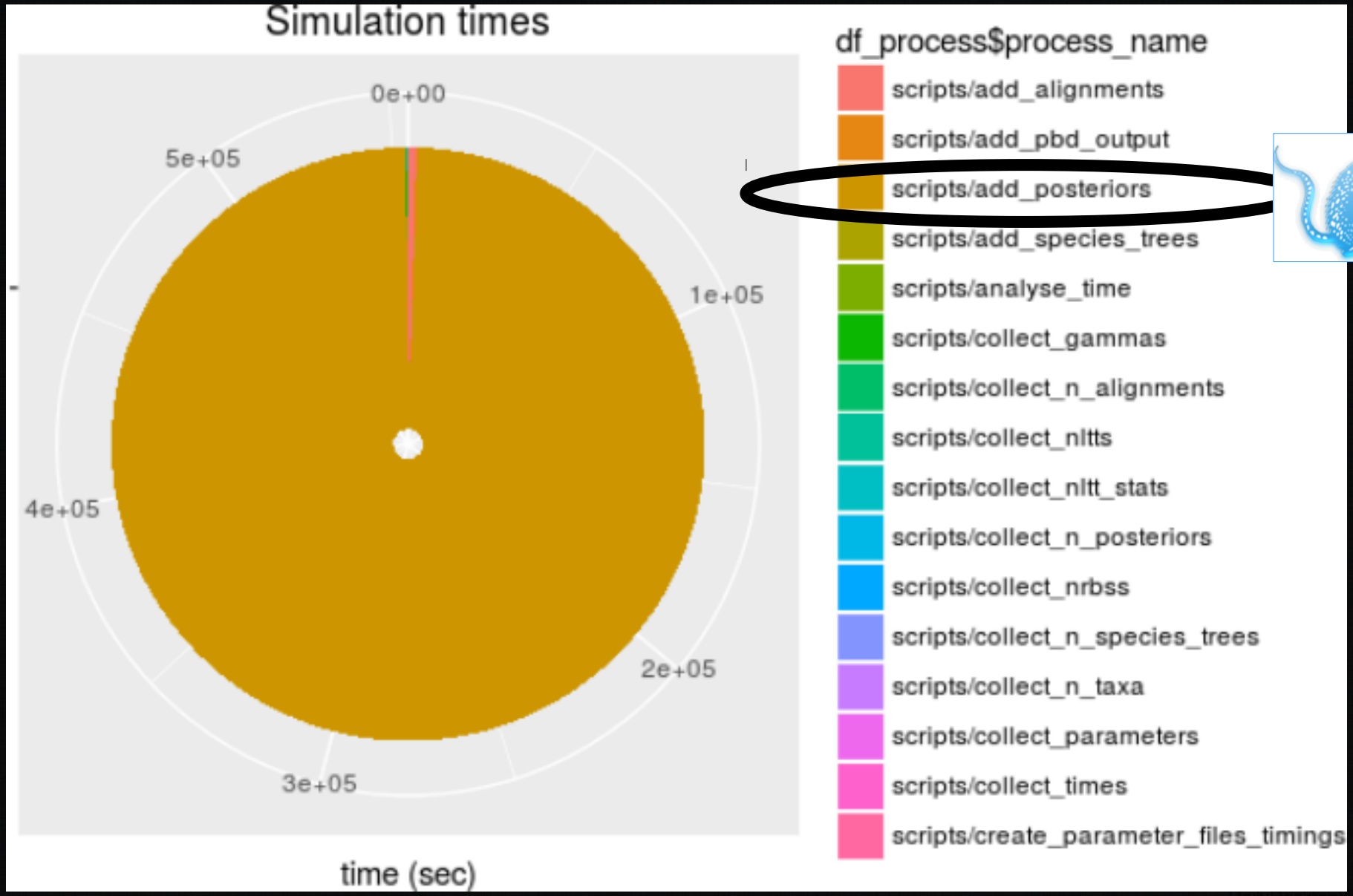
Under which parameters can I simulate?

Under which parameters can I simulate?

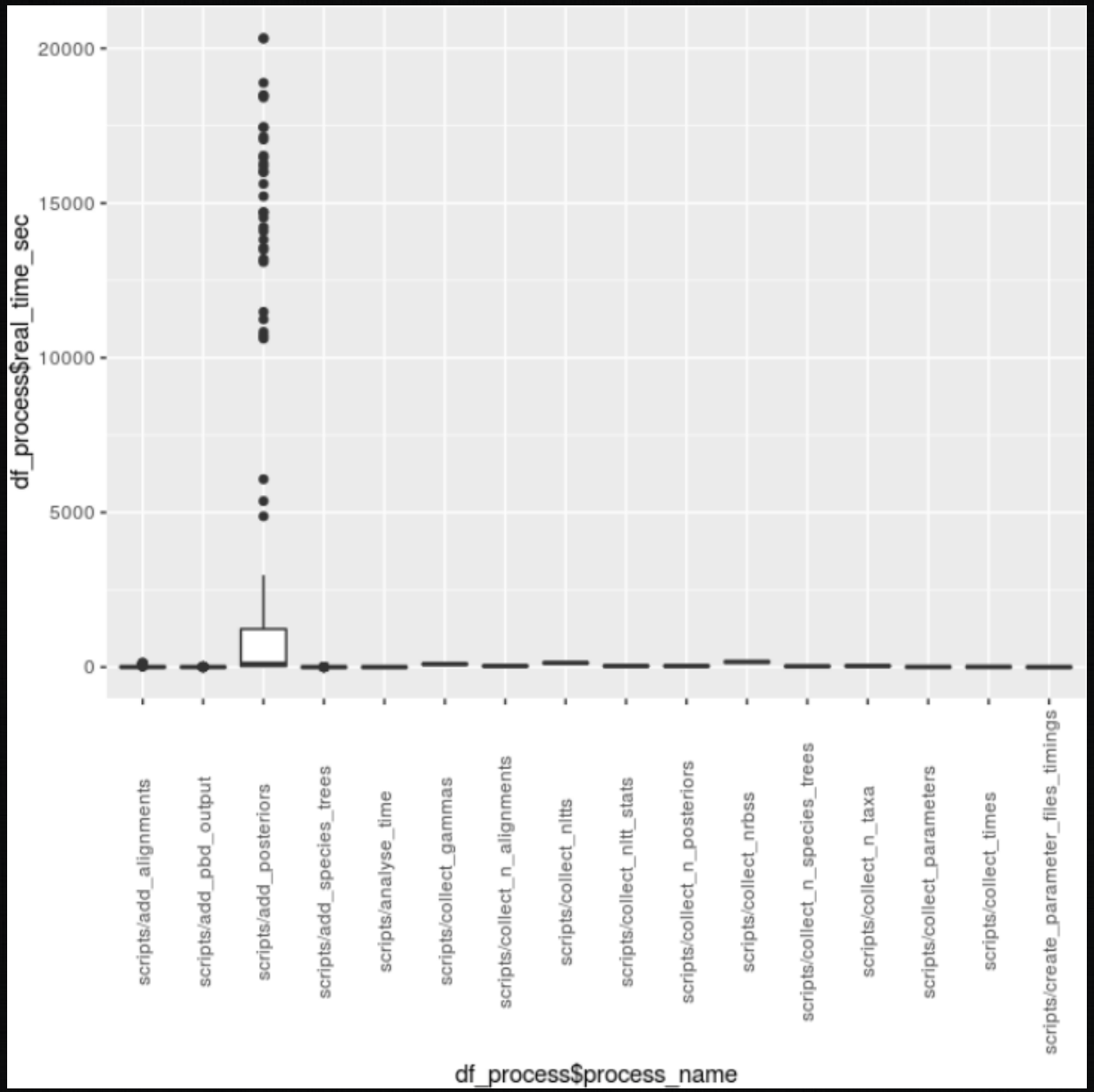
Premature optimization is
the root of all evil
(or at least most of it)
in programming

Donald Knuth

Under which parameters can I simulate?

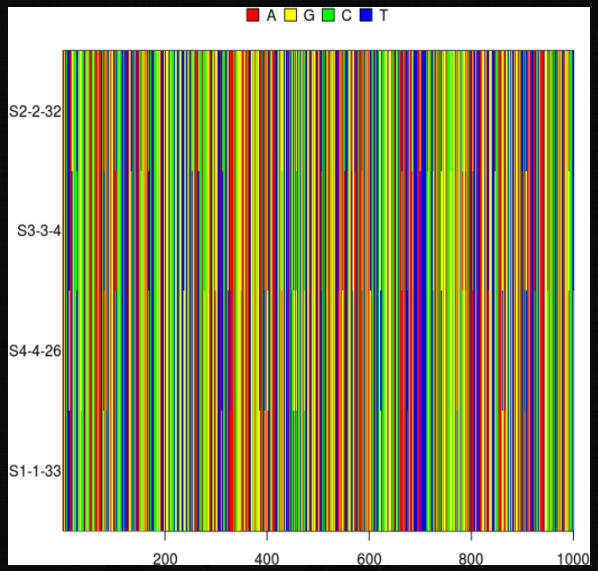


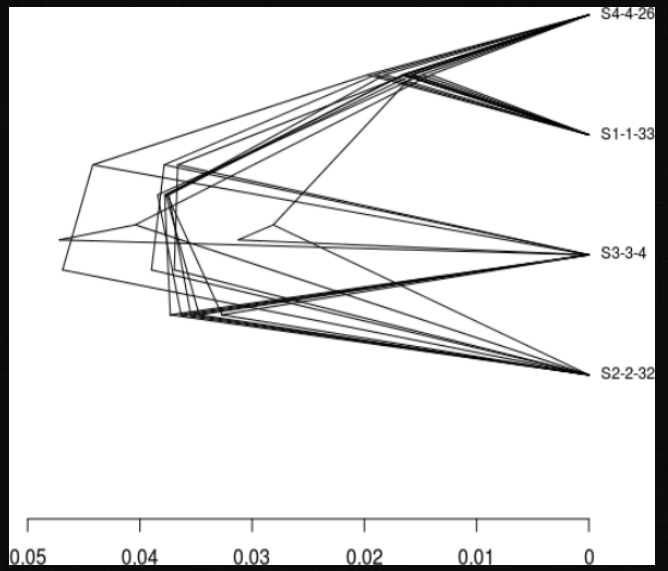
Under which parameters can I simulate?



add_posteriors

- Runtime speed determined by
  - Alignment length?
  - Number of taxa?

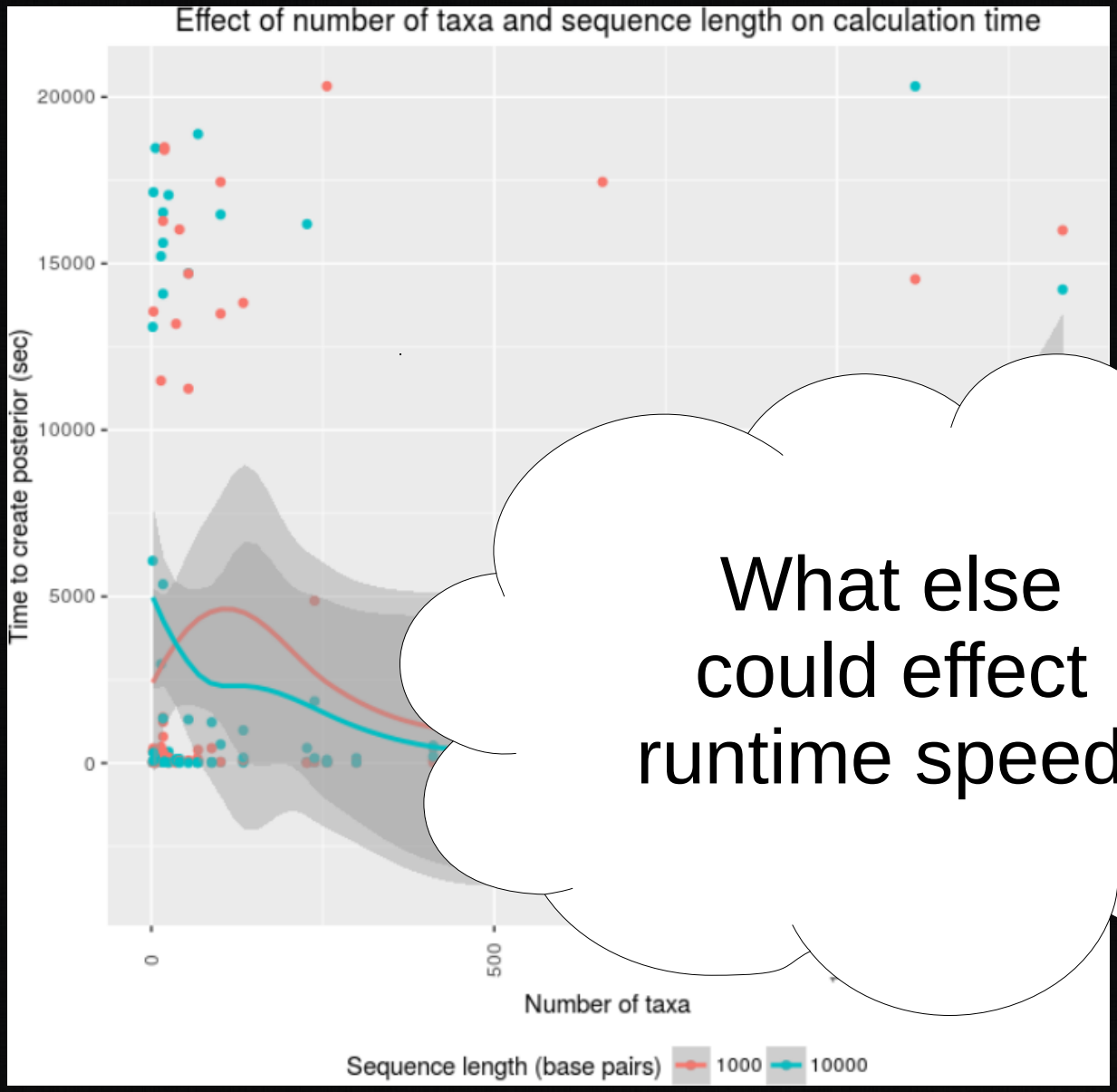




Simulate DNA alignments

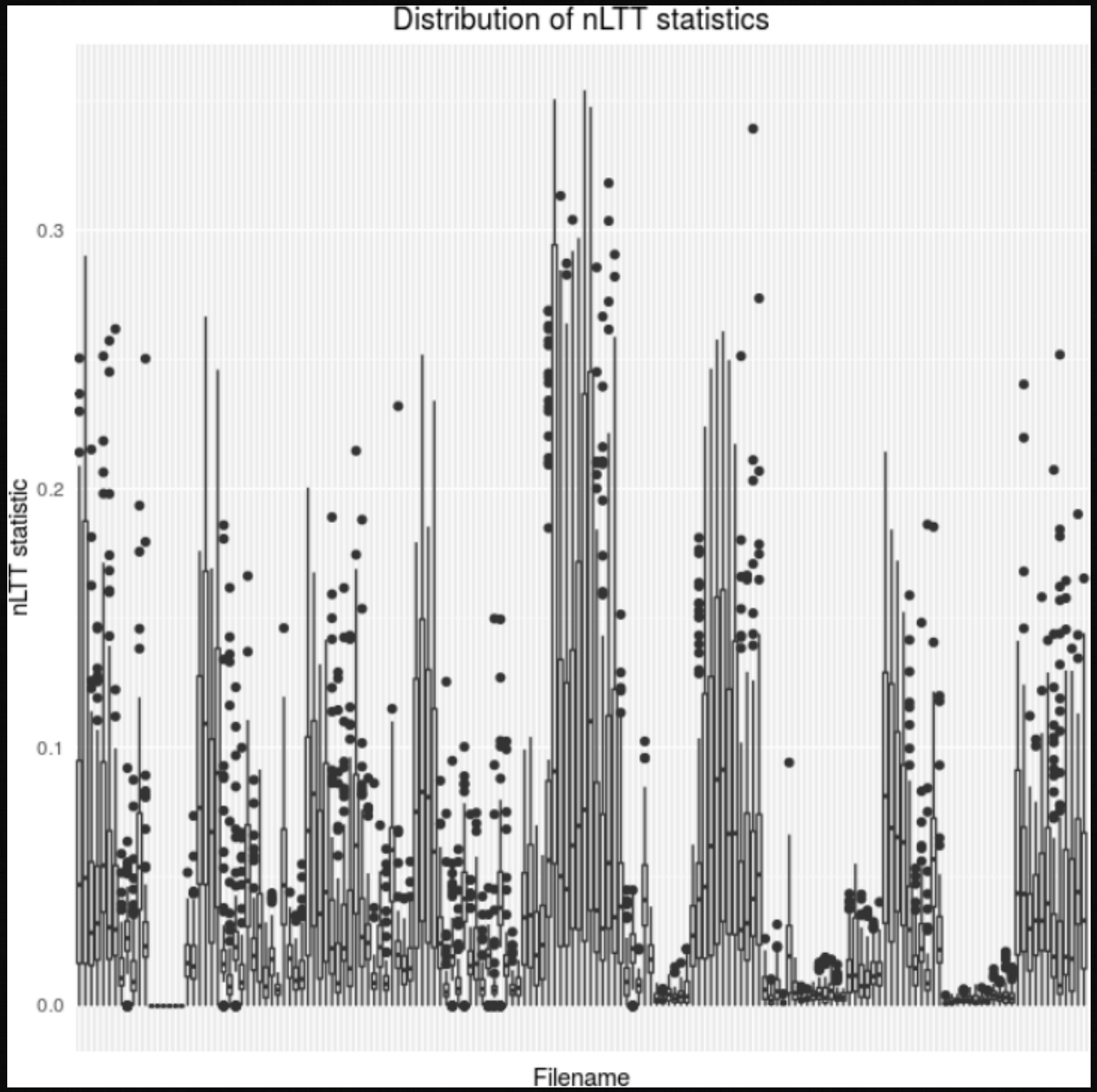
Infer species tree posterior

Under which parameters can I simulate?

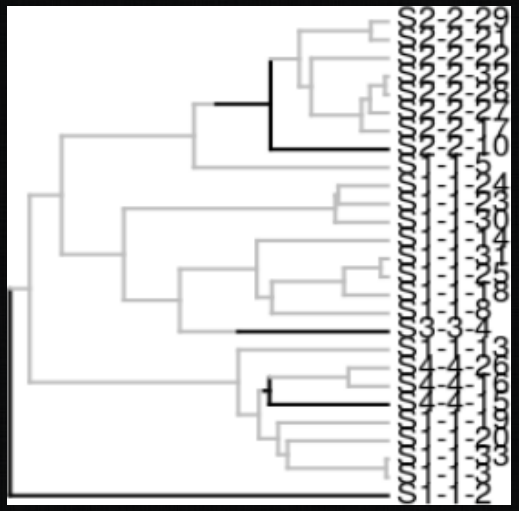


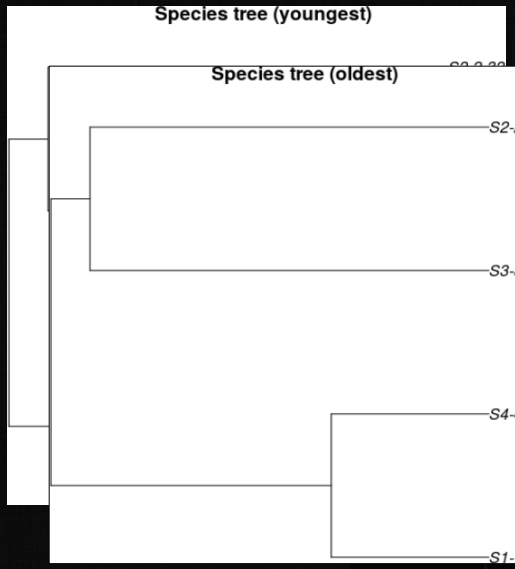
What else could effect runtime speed?

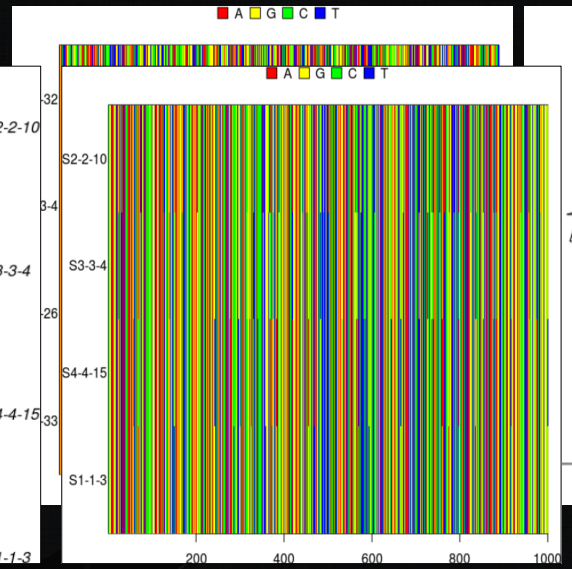
Measure the nLTT statistics for those parameters

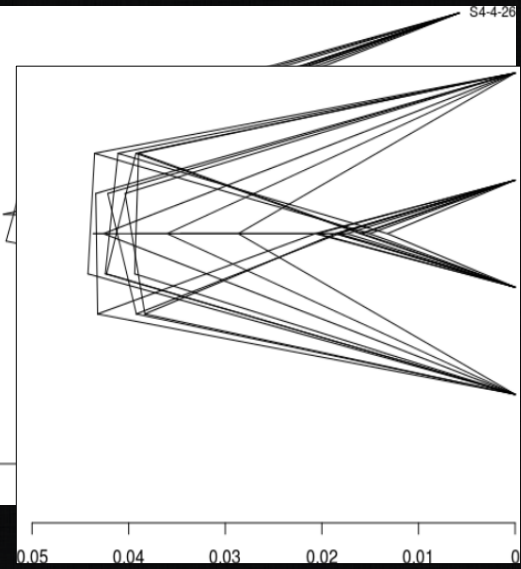


Measure the nLTT statistics for those parameters



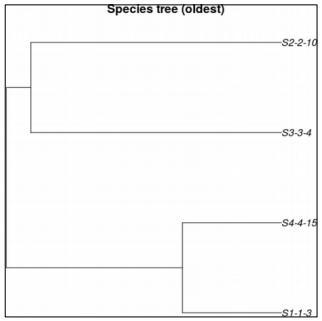




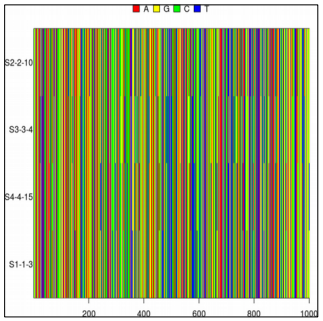


Goal

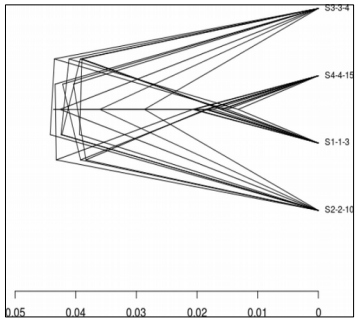
There was no/a difference
in nLTT statistics distributions between



sampled
species trees



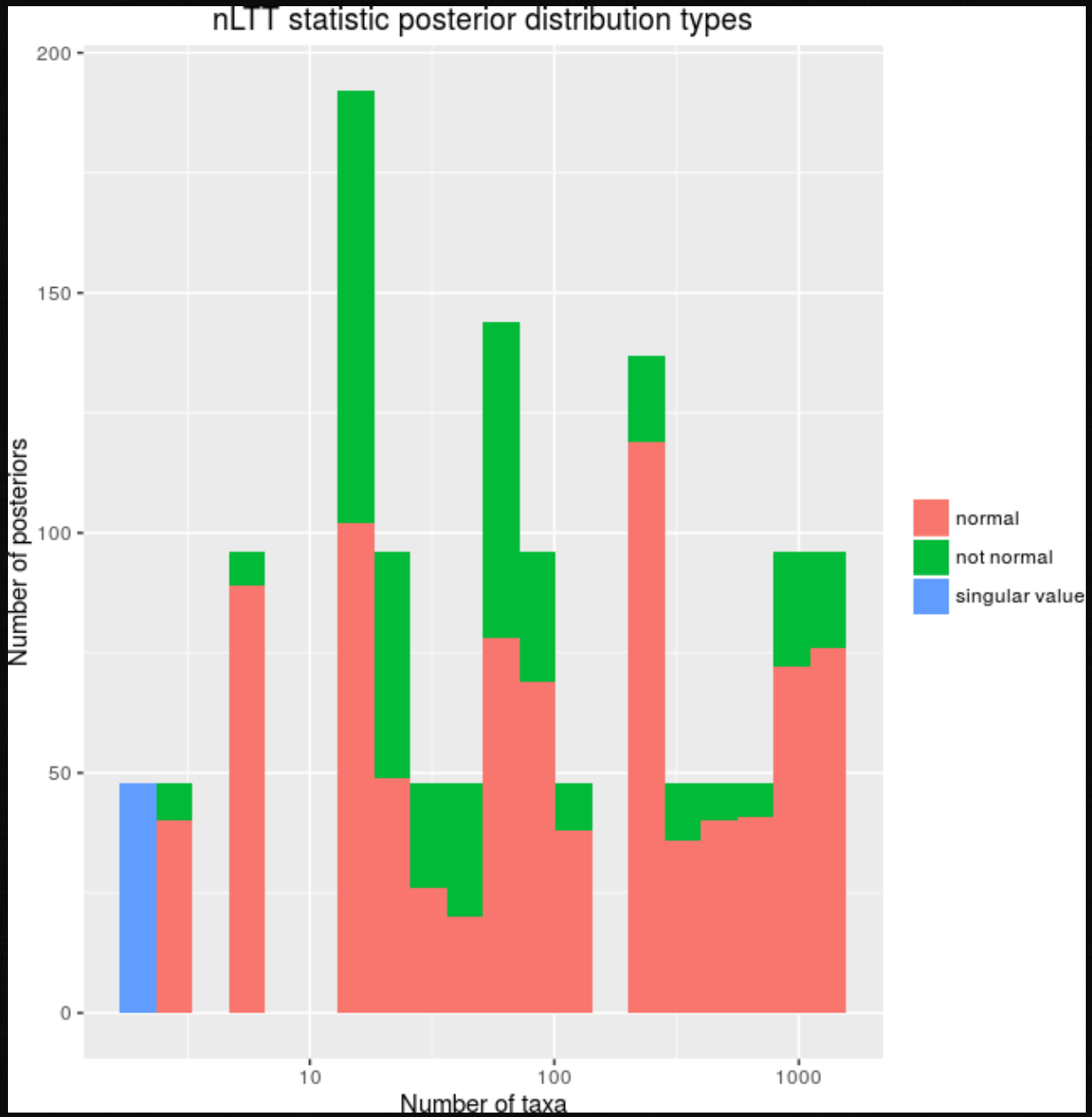
simulated
alignments

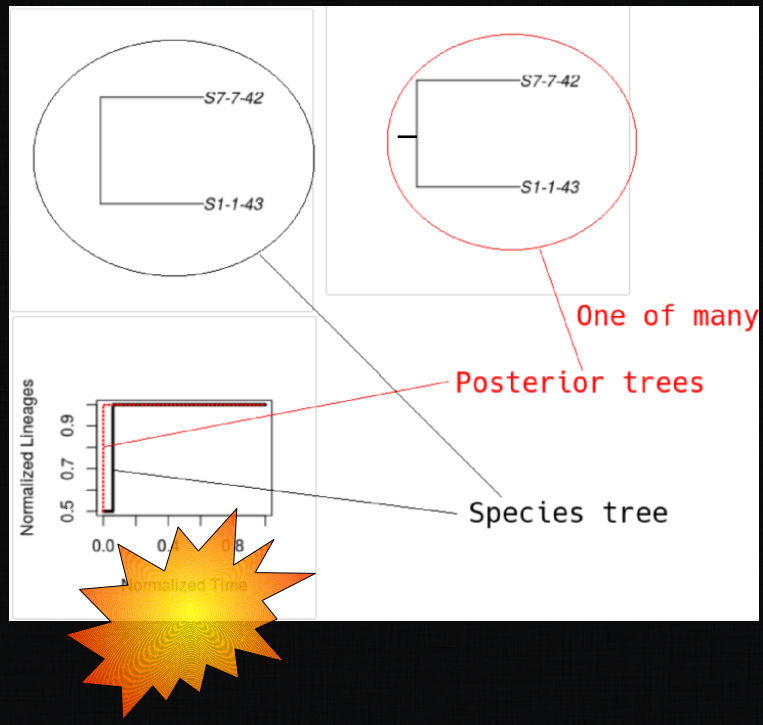


BEAST2
runs

so we reduced/kept
the number of replicates

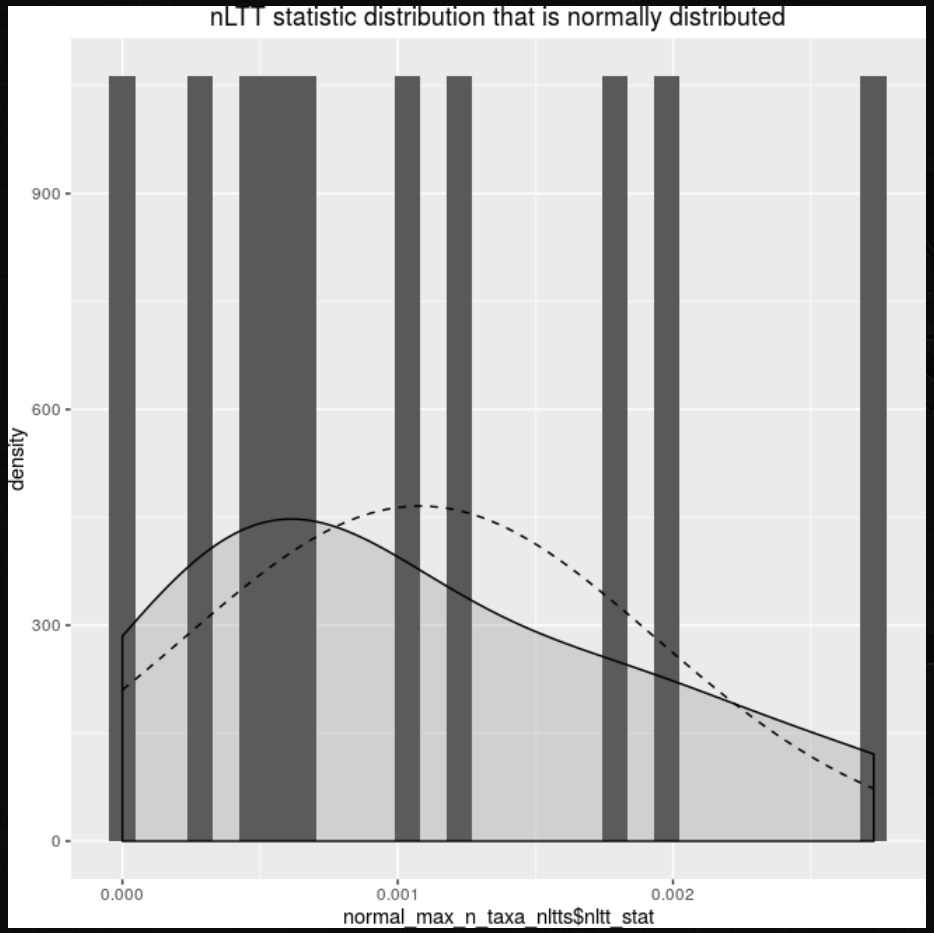
nLTT statistic distribution

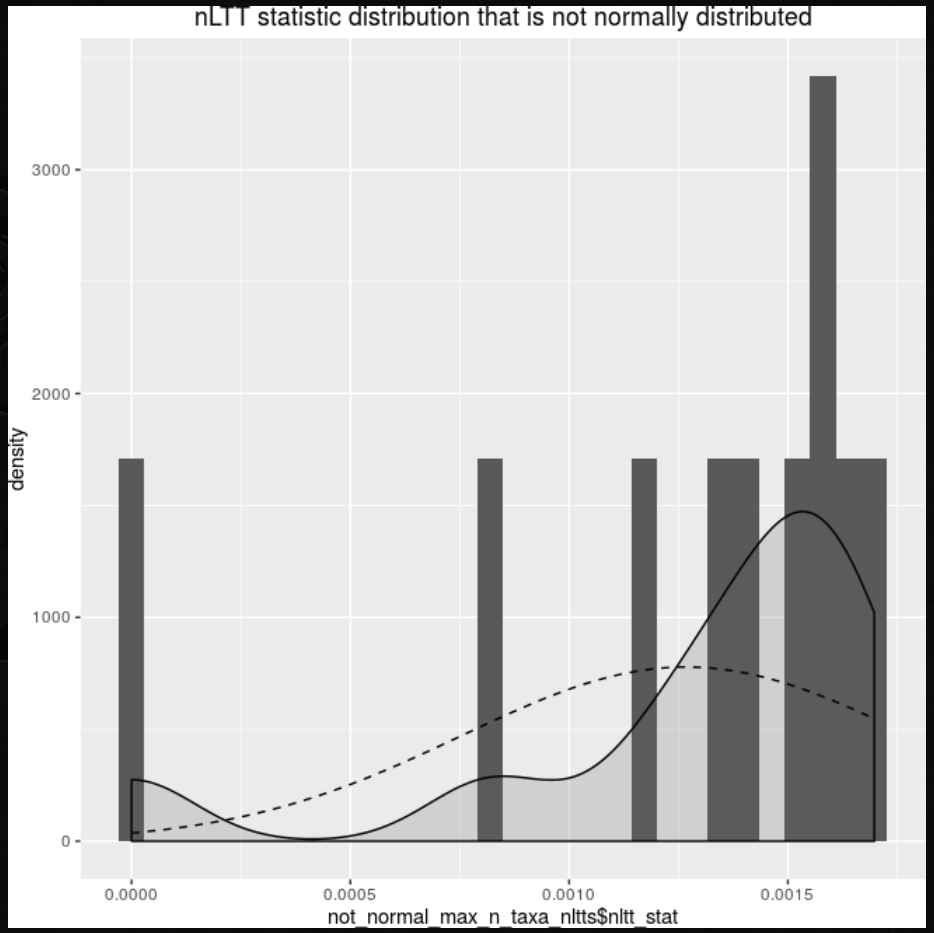




idn	n
FALSE	394
TRUE	895
NA	48

nLTT statistic distributions

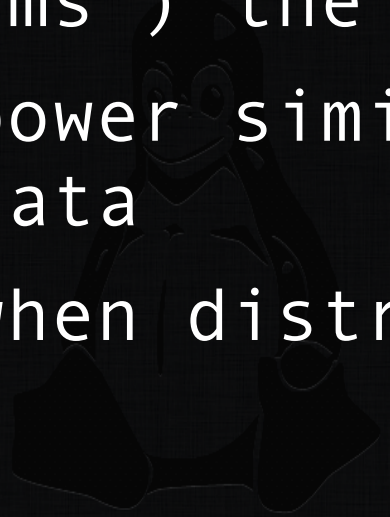




Use a test that does not assume normality

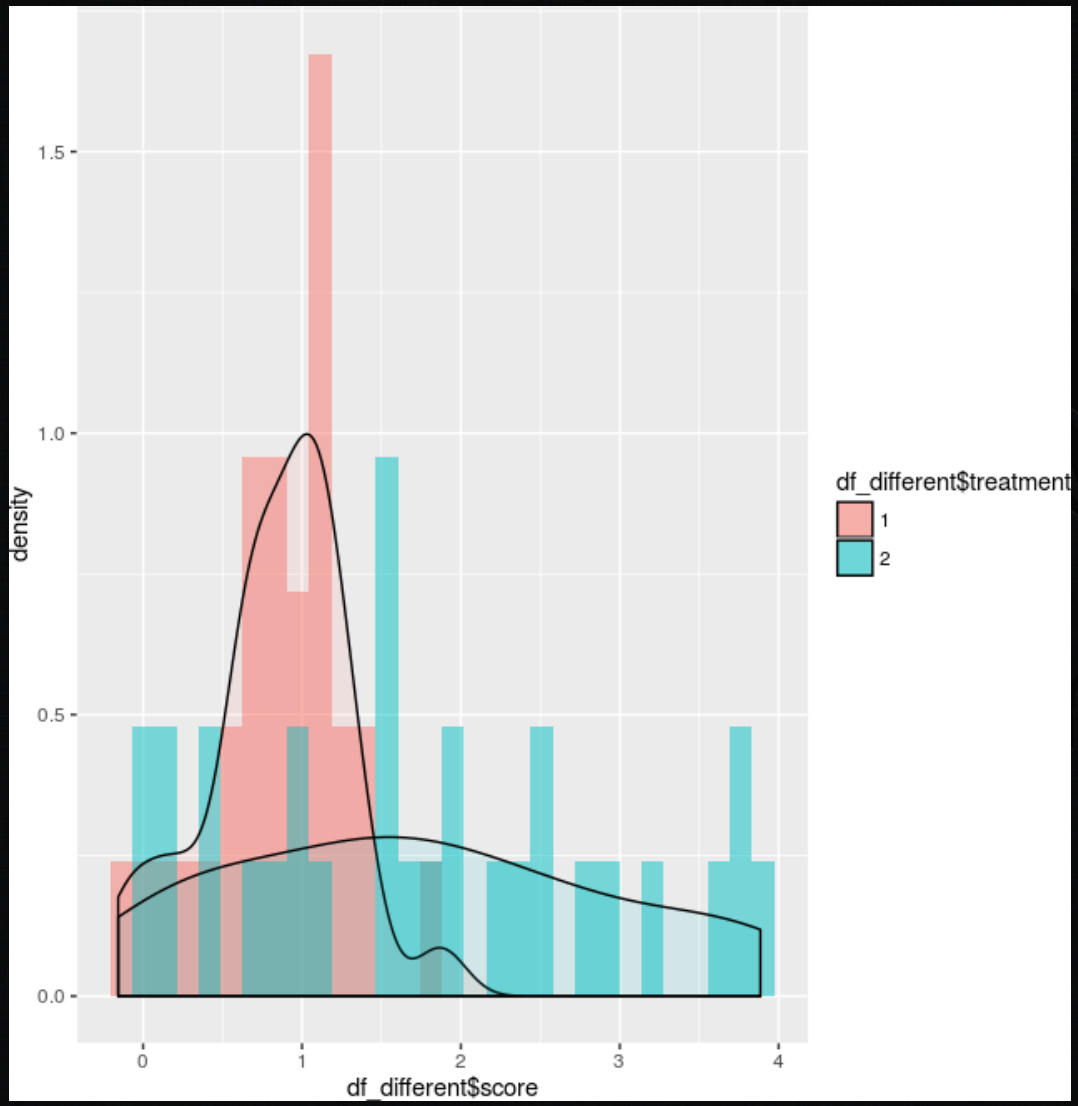
NEW: Two-sample trimmed t [1]

- Recommended by [2]
  - Removes (‘trims’) the extreme values
  - Statistical power similar for normally distributed data
  - More robust when distribution is long-tailed



[1] Yuen, Karen K. "The two-sample trimmed t for unequal population variances." Biometrika 61.1 (1974): 165-170.
[2] Field, A. Miles. "J. & Field, Z.(2012)." Discovering statistics using R.

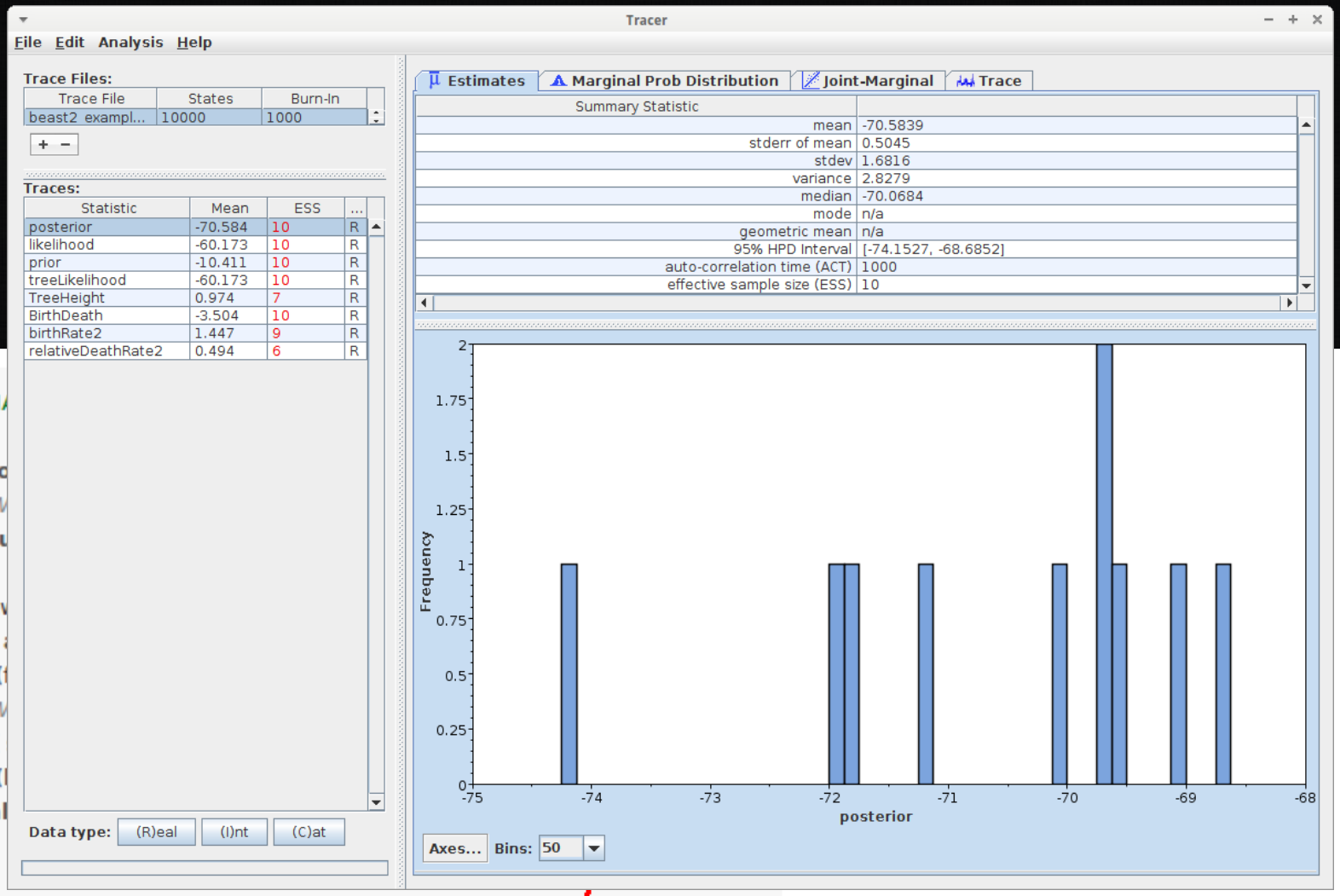
Example



1: rnorm(n = 30, mean = 1, sd = 0.5)
2: runif(n = 30, min = 0, max = 4)

data_set	test	p_value
same	yuen	1.0000000
different	yuen	0.0081425
same	yuenbt	1.0000000
different	yuenbt	0.0095000
same	pb2gen	0.9849750
different	pb2gen	0.0033389

NEW: measure ESS




```
esses <- rep(NA

for (i in seq_alo
  # Estimates W
  ewob <- as.nu

  n <- length(ev
  first_index <-
  testit::assert(f
  # Estimates W
  ewb <- ewob[
  testit::assert(l
  esses[i] <- cal
}
esses
```

```
## [1]  2.668464 10.000000 10.000000 10.000000 10.000000  6.657254 10.000000
## [8]  8.905181  6.217762
```

`RBeast::calc_ess`

Collaborations

https://github.com/richelbilderbeek/nLTT

Thijs Janzen

https://github.com/beast-dev/RBeast

Marc Suchard

https://github.com/olli0601/rBeast

Oliver Ratmann

Work in progress

- Under which parameters can I simulate (1) within 10 days (2) for an ESS > 200?
- Draw conclusions from the nLTT statistics
- Finish article