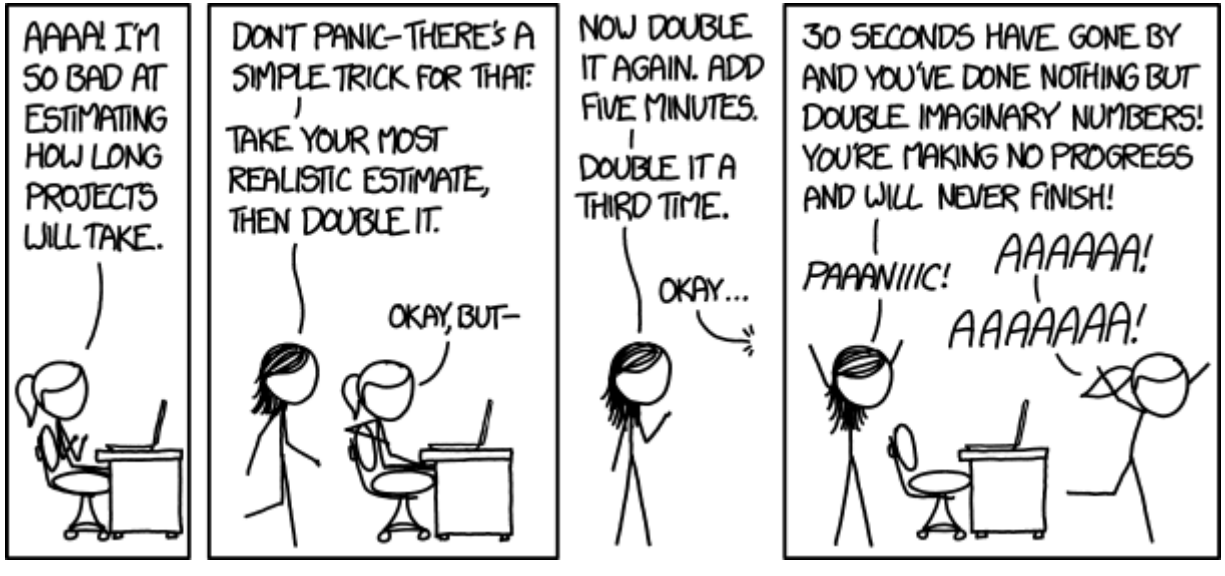
MIMM4750G

Molecular clocks



Please don't forget to submit your proposal abstracts today!

What is a molecular clock?

- The molecular clock is a hypothesis that evolution at a molecular level (*e.g.*, proteins) occurs at a constant rate.
- Based on linear relationship between the time that species diverged from their common ancestor, and the number of amino acid differences in homologous proteins (*right*).
- The molecular clock does *not* "tick".

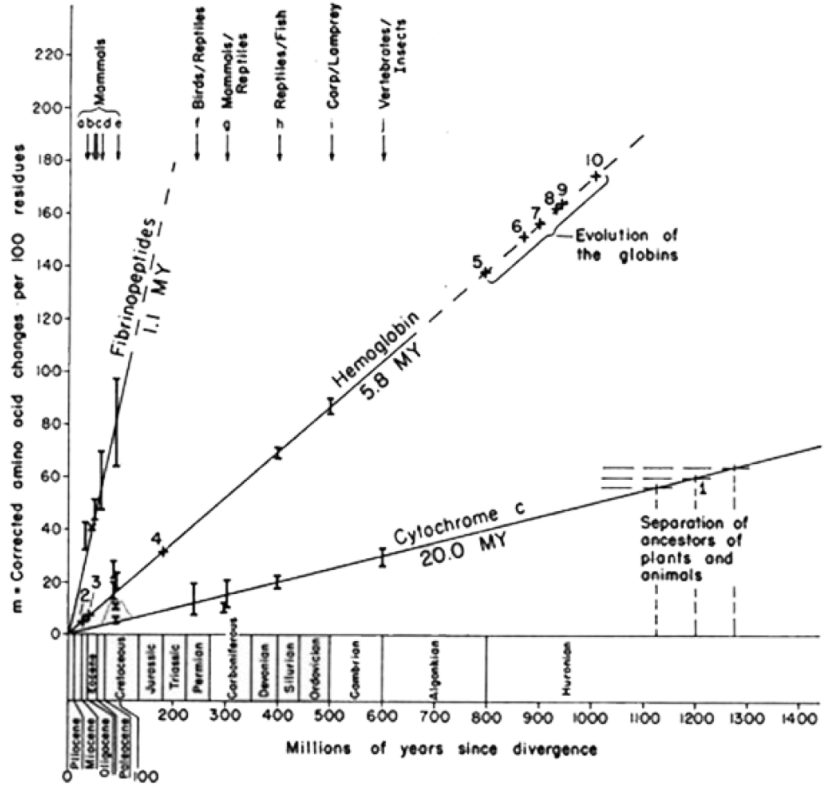


Figure from RE Dickerson (1971) J Mol Evol 1: 26.

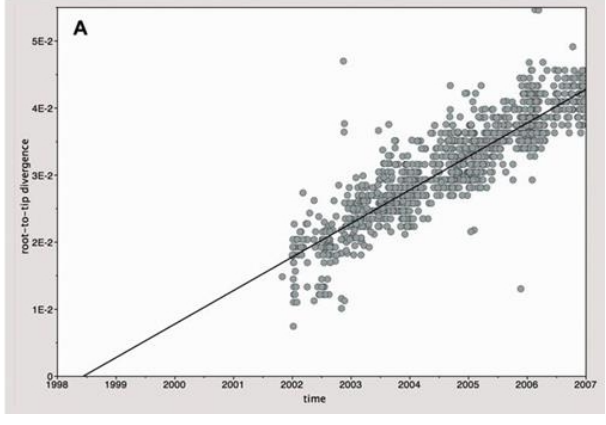
Neutral theory

- The neutral theory: "the great majority of evolutionary changes at the molecular level [...] are caused not by Darwinian selection but by random drift"
- Without selection, nucleotide substitutions should be a random process at a constant rate.
- Today, this model is denoted the "strict" clock and is seldom supported.

Motoo Kimura

Testing for a clock

- If a clock exists, then sequences should become more distant from the common ancestor as a *linear trend* over time.

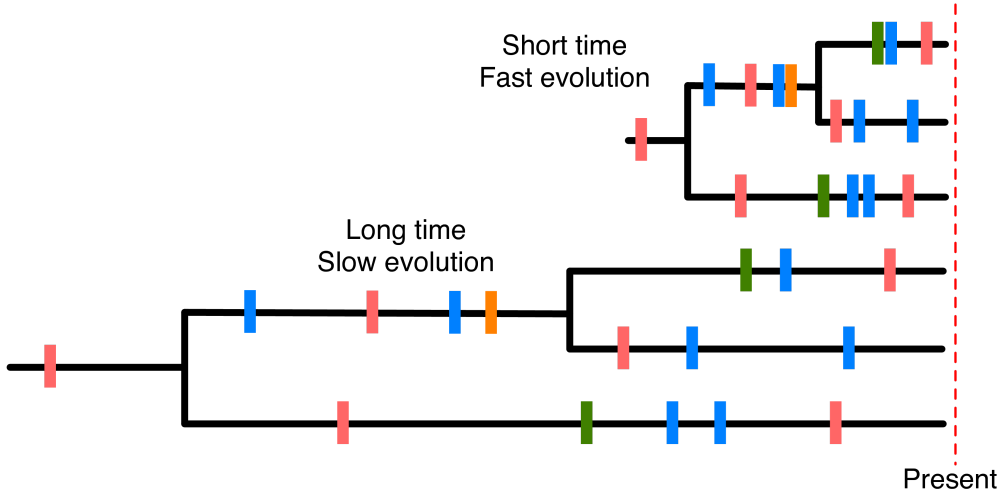


- (above) Plot of total branch length between root and tip, against sample collection date of tip, for influenza A virus (Rambaut *et al.* 2015 Virus Evolution).

Clocks and trees

- When we reconstruct a tree by fitting a model of evolution, the lengths of branches are measured in units of *expected numbers of substitutions*.
- This weird unit exists because most models are scaled to $\mu \times t$ (the rate of evolution multiplied by time).
- We say that the rate and time are *confounded* - it is impossible to estimate one parameter without knowing the true value of the other.

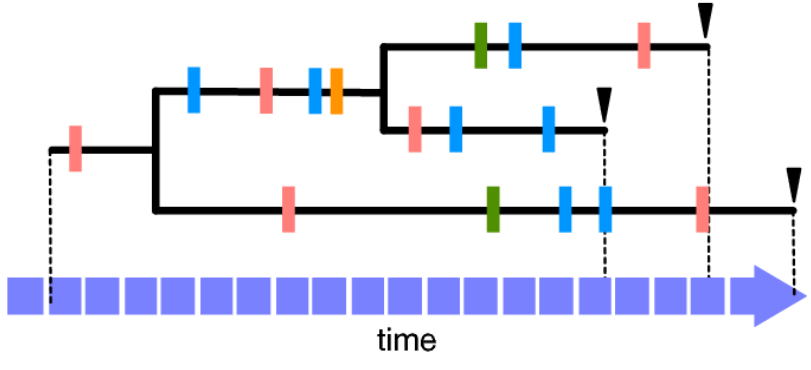
Confound it!



- A tree can be stretched back in time ($\uparrow t$) and explain the data with *exactly the same likelihood* if we decrease the rate ($\downarrow \mu$).
- Conversely, if we compress the tree forward in time ($\downarrow ts$), we obtain the same likelihood by speeding up the rate ($\uparrow \mu$).

Estimating the clock

- For infectious diseases, measurable evolution can occur on a time scale of weeks.
- This means that a branch of the tree can "grow" in the time between sampling different infections.



- We can use sampling times to "pin down" the tree and prevent the free-scaling problem.

Rooting the tree

- If we are directly measuring *time*, then we must impose a direction of time's flow on the tree.
- This means that we must locate the tree's *root* (the earliest point in time).
- Rooting trees is difficult! By definition, it is the point that is the most distant in time — it cannot be directly observed.

Outgroup rooting

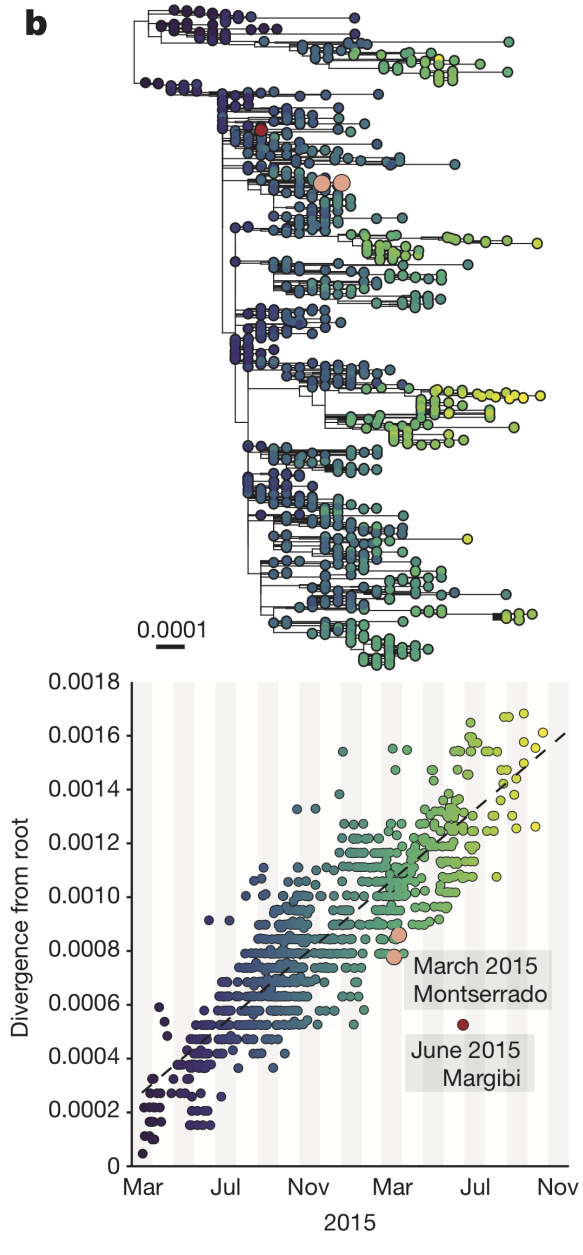
- An *outgroup* is a species or infection that is not closely related to the sample set.
- We place the root at the point where the branch leading to the outgroup intersects the tree.
- If the outgroup is too *close*, then the root is too influenced by sampling bias (our choice of outgroup).
- If the outgroup is too *distant*, then where it intersects the tree is subject to random variation.

Choosing an outgroup

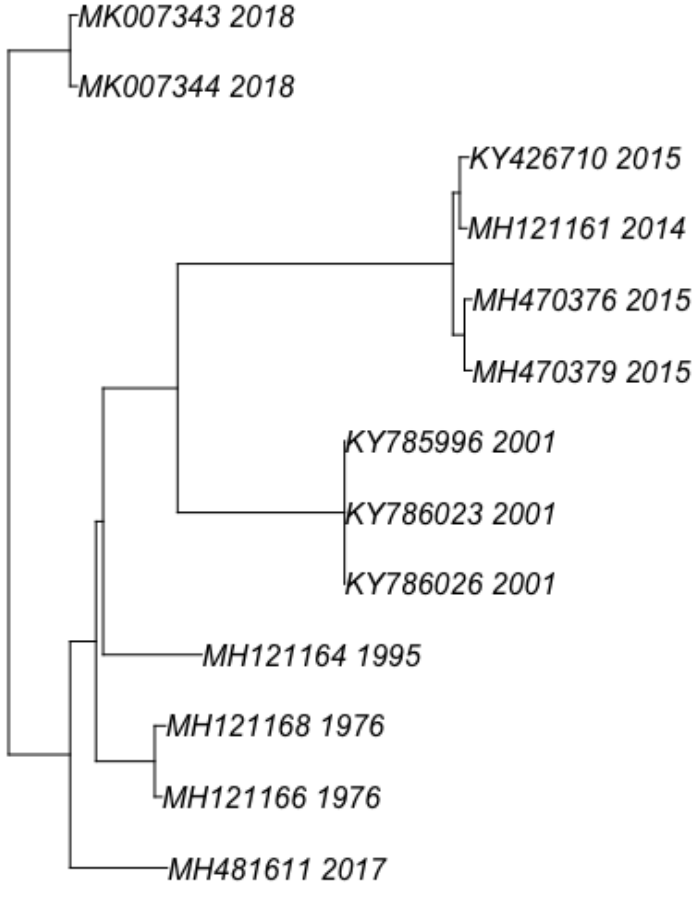
- There is no objective framework to select an "optimal" outgroup.
- We often reach an informal consensus about which outgroup to use, without any other rationale.
- *e.g.*, HIV-1 subtype D is an outgroup for HIV-1.
- If an early isolate of the pathogen is available, it is sometimes used as an outgroup.
- *e.g.*, West Nile virus isolate B956 (isolated in 1940).
- **Use different outgroups to see how robust your conclusions are to choice of outgroup!**

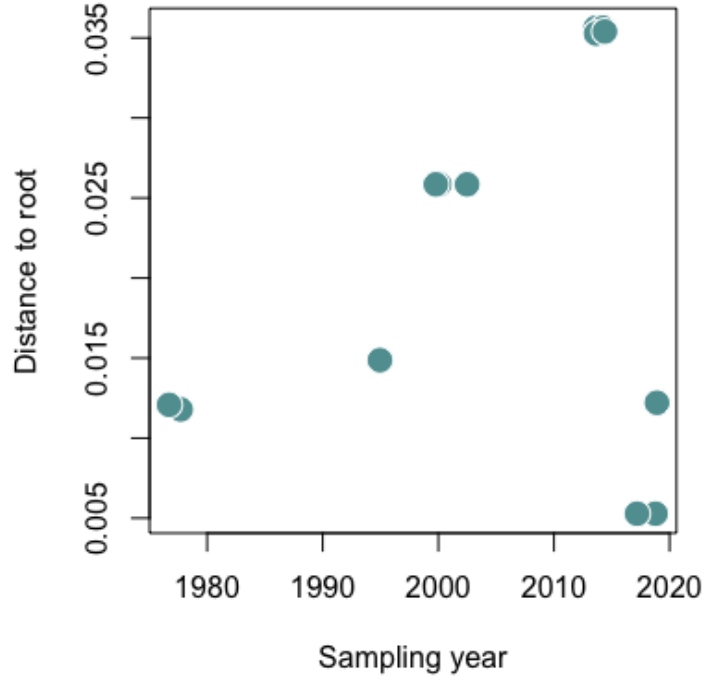
Root-to-tip regression

- We can simultaneously estimate the clock (rate of evolution) *and* locate the root.
- Try different placements of the root and look at the plot that results.
- The "best" root should give the cleanest (positive) linear trend.
- (*right*) Phylogeny and root-to-tip plot for Ebolavirus isolates from 2013-2016 epidemic in West Africa.

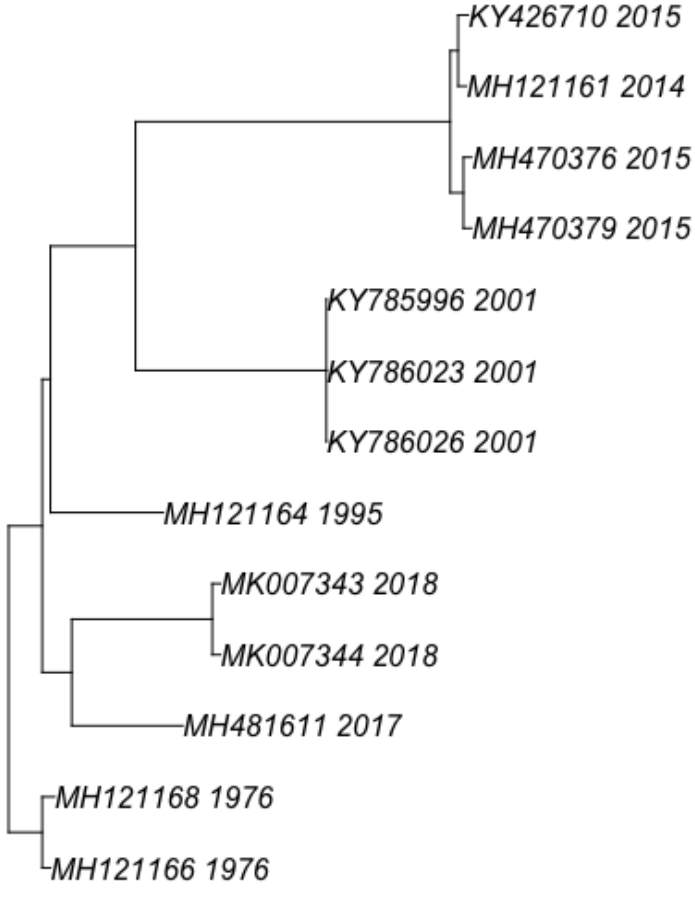


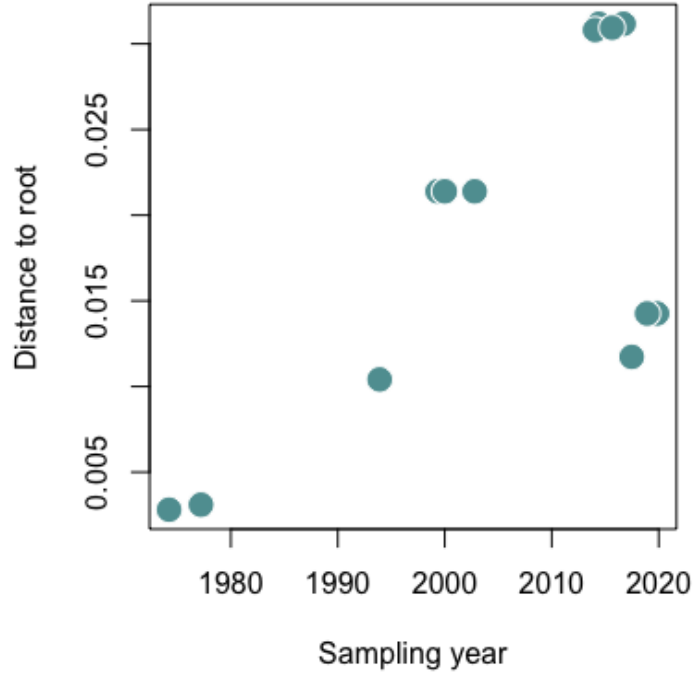
"Bad" root


MK007343 2018
MK007344 2018
KY426710 2015
MH121161 2014
MH470376 2015
MH470379 2015
KY785996 2001
KY786023 2001
KY786026 2001
MH121164 1995
MH121168 1976
MH121166 1976
MH481611 2017


Distance to root
0.005
0.015
0.025
0.035
1980
1990
2000
2010
2020
Sampling year

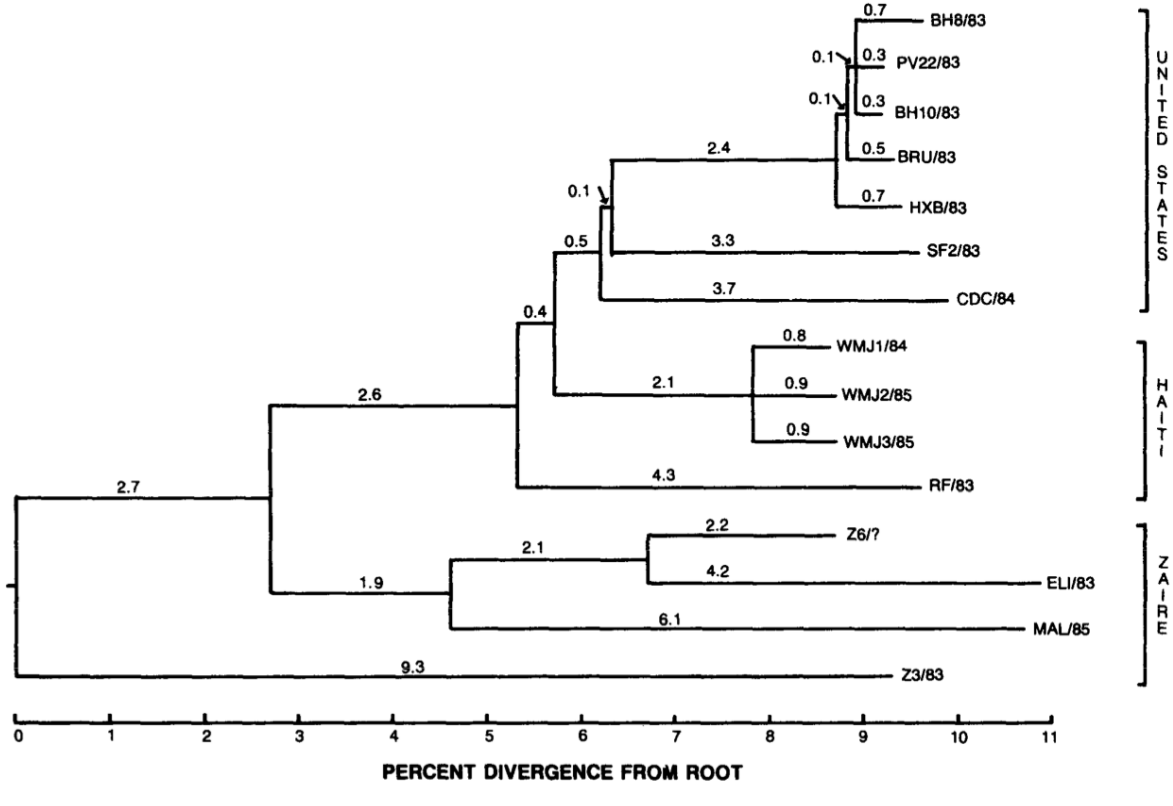
"Good" root


KY426710 2015
MH121161 2014
MH470376 2015
MH470379 2015
KY785996 2001
KY786023 2001
KY786026 2001
MH121164 1995
MK007343 2018
MK007344 2018
MH481611 2017
MH121168 1976
MH121166 1976


Distance to root
0.005
0.015
0.025
1980
1990
2000
2010
2020
Sampling year

Origin of HIV-1

- This concept was first used to estimate the origin date of HIV-1.



Tree from Li, Tanimura and Sharp (1988) Mol Biol Evol 5: 313.

Origin of HIV-1

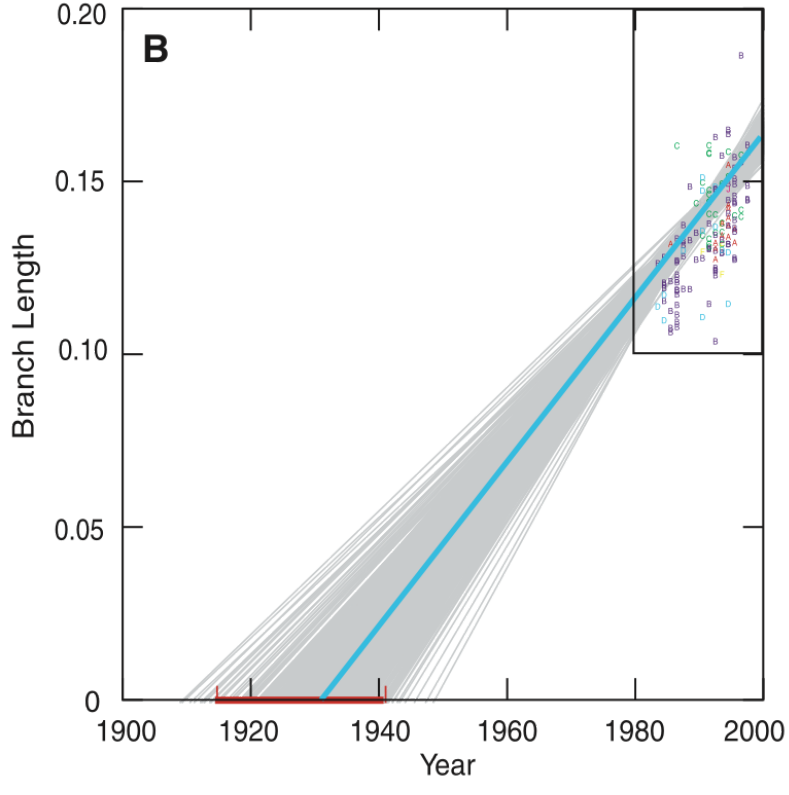


Figure from B Korber *et al.* (2000) Science 288.