Introduction

Richèl J.C. Bilderbeek[1]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

March 3, 2020

1

INTRODUCTION

Once upon a time, there was the first living organism, the First Universal Common Ancestor (FUCA). We do not know when it lived. We can depict the evolutionary history of FUCA as seen in the left panel of figure 1.1.

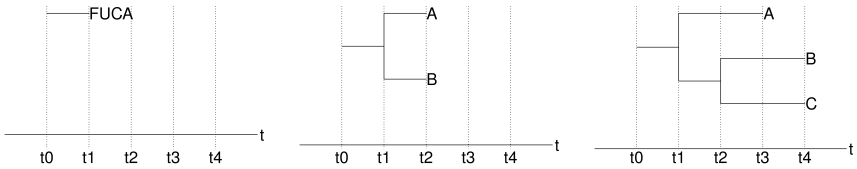

Figure 1.1 | Left: Evolutionary history of the First Universal Common Ancestor (FUCA). Middle: Evolutionary history of the two decendants of FUCA. Right: Evolutionary history of the three decendants of FUCA, of which one went extinct. Assuming B and C will give rise to all contemporary biodiversity, the Last Universal Common Ancestor (LUCA) must have come into existence at timepoint t1. In all cases, time goes from past (left) towards the present (right).

One unknown day, FUCA speciated, resulting in two species. This event doubled the biodiversity on Earth. The two species, which we will call species A and B are sister species. We do not know what caused the speciation. We can depict the evolutionary history of FUCA as seen in the middle panel of figure 1.1.

Both species A had their unknown histories: they speciated themselves, and they and/or their descendants went extinct. Extinction is a common event. Let's assume A and/or its clade went exctint and that species B created a sister species C. Species B and C will give rise to all contemporary biodiversity. This ancestor of species B and C is called the Last Universal Common Ancestor and lived around [then]. We can depict the evolutionary history of FUCA as seen in the right panel of figure 1.1. **[RJCB: Make this story a bit more coherent]**

The biodiversity derived from LUCA is important to us humans (apart from that is has created us) for many reasons. Biodiversity is found so important, that, for example, the European Union has an explicit Biodiversity Strategy, which aims to halt the loss of biodiversity and ecosystem services (`https://ec.europa.eu/environment/nature/biodiversity/strategy/index_en.htm` **[RJCB: Move reference to references]**). Ecosystem services are features of biological systems that are positive for human well-being, for example food, carbon sequestration, waste decomposition and pest control. A review paper Cardinale *et al.* 2012 shows that biodiversity unsually improves ecosystem services.

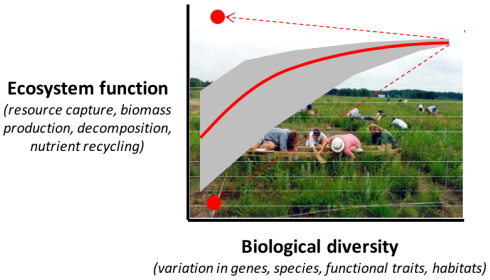

Figure 1.2 | A diversity-function relationship found to be typical from hundreds of studies. The red line represents an average, where the grey polygon represents a 95% confidence interval. The red dots show the lower and upper limit for monocultures. From Cardinale *et al.* 2012

Speciation is the process that increases biological diversity. This process is studied from multiple angles, among others, we can study the mechanism ('what causes a speciation event?') or we can study the patterns of many of such events ('is speciation rate constant through time?')

The mechanism of a speciation event has many facets. For more than half a century ago, it was hypothesized that speciation is caused by geographical isolation (e.g. Mayr, 1942) or due to ecological factors (e.g. Lack, 1947).

Instead of looking at the mechanism behind each speciation event, we can also look at patterns of speciation events through evolutionary time, asking question such as 'How often do speciation and extinction events take place?' 'Are speciation and extinctions rates constant or do they change?', 'What causes a change in speciation or extinction rate?' or 'Is there an upper limit on the number of species?'.

There are two methods to research speciation patterns back in evolutionary time: the use of fossils or using molecular phylogenies.

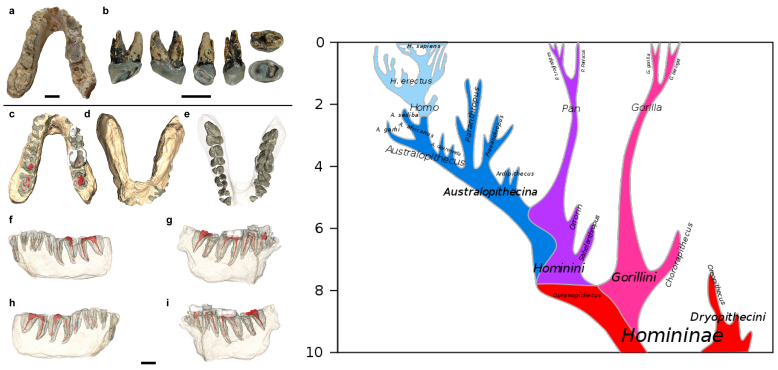

Figure 1.3 | Left: El Graeco fossil, from Fuss *et al.* 2017. Right: Evolution of the Homininae, based on Stringer 2012

Using fossils is a classic way to look back in evolutionary time. Fossils show a glimpse of the biodiversity in the past. We can deduce the age of fossils, by the rock layers they are found in. Using fossils has its limitations. First, it is mostly species with hard body parts

that are suitable to fossilize. Of such species, an organisms is still only rarely preserved, of which only a fraction under ideal circumstances. Of these fossils, only a fraction is discovered. One example of a famous fossil is 'El Graeco', which may be the oldest known hominin Fuss *et al.* 2017, where homonins are the tribe (taxanomic group) we Homo sapiens share with the Panini.

Using molecular phylogenies is the modern way to look back in evolutionary time. It is the use of heritable molecules (e.g. DNA, RNA, or protein) of contemporary species to infer phylogenies. The field of phylogenetics is the research discipline that intends to infer the most accurate phylogenies possible, regarding topology, speciation and extinction times, optionally adding morphological data and/or fossil data. Phylogenetics is applied in many settings, among others, species classification, forensics, conservation ecology and epidemiology Lam *et al.* 2010.

One example of the importance of an accurate phylogenetic tree is demonstrated in Bush *et al.* 1999. This study investigated which loci of the H3 hemagglutinin surface protein are under selection, by constrasting asynonymous and synonymous mutation rates along the branches of a phylogeny. In a preliminary analysis by the authors, they noted that most selection rates were either below or above the statistical threshold depending on the phylogeny. This study contributed to the selection of recommended composition of influenza virus vaccines.

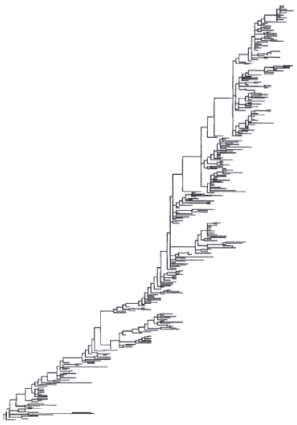

Figure 1.4 | Phylogeny of the human influenza virus type A subtype H3, from Bush *et al.* 1999

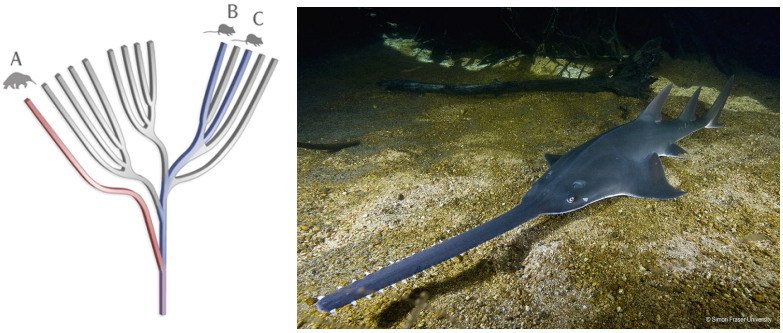

Figure 1.5 | Left: The ED (evolutionary distinctiveness) of species A is higher than that of species B or C, as more evolutionary history will be lost when that species goes extinct. Right: The Largetooth Sawfish (Pristis pristis) is at number 1 of the EDGE (ED = 'Evolutionary Distinctiveness', GE = Globally Endangered status) list, with an EDGE Score of 7.38 and an ED of 99.298.

Another example of the importance of an accurate phylogenetic tree comes from conservation biology, in which phylogenies are used to calculate an EDGE ('Evolutionarily Distinct and Globally Endangered') score. Species with a high EDGE score are prioritized in conservation. To calculate an EDGE score, one needs a metric of evolutionary distinctiveness ('ED') and globally 'endangeredness' ('GE'). The GE score is a conservational status, ranging from from zero ('Least Concern') to four ('Critically Endangered'). The ED embodies the amount of evolutionary history lost when a species would go extinct, which can be calculated from a (hopefully accurate) phylogeny.

Figure 1.6 | PHYLIP logo

Phylogenetics has taken a huge flight, due to the massively increased computational power and techniques. A first milestone in this file is Felsensteyn's work in 1980, creating (and still maintaining) PHYLIP, the first software package for classical phylogenetic analysis. Another milestone is the Metropolis-Hasting algorithm, which allowed Bayesian phylogenetics to thrive, resulting in contemporay tools such as BEAST, BEAST2, MrBayes and RevBayes.

Figure 1.7 | BEAST2 logo (left) and example output (right)

A clear example of the power of modern phylogenetics, is the Tree Of Life: it uses 3,083 genomes of 2,596 amino-acid positions to create one big phylogeny of all (sequenced) life on Earth, which took 3,840 computational hours on a modern supercomputer Hug *et al.* 2016.

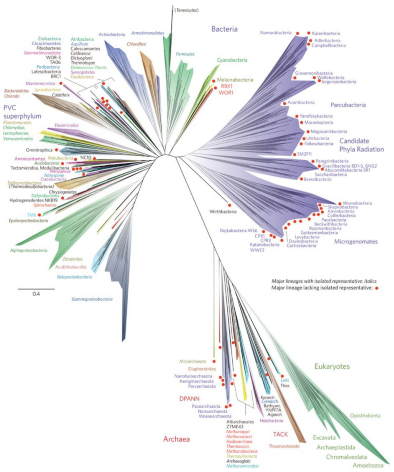

Figure 1.8 | Tree of Life, from Hug *et al.* 2016

To create such a tree from protein sequences, one has to specify an evolutionary model. This evolutionary model embodies our set of assumptions, such as the evolution of a protein sequence (also called the site model), the rate(s) at which this happens (the clock model) and the rate(s) at which a branching/speciation event takes place (the tree model). For example, the amino acids of the Tree Of Life are assumed to change/mutate according to the LG model Le & Gascuel 2008, which is a model that uses the average transition rates found in nature.

There are many evolutionary models to choose from, and selecting which one to use is hard, due to the many sets of assumptions to choose from. In general, modellers are looking for that set of assumptions that is as simple as possible, but not simpler. And even then, sometimes an overly simplistic model is picked regardless, due to computation constraints.

By using a model comparison, one has a rational way to select an evolutionary model that is as simple as possible, but not simpler. A model comparison algorithm selects the evolutionary model that is most likely to have generated the data, without being overly complex. The idea is that the best evolutionary model should result in the most accurate phylogenetic trees.

Because model comparison is hard, there have been multiple studies that investigate the effect of picking the wrong evolutionary models.

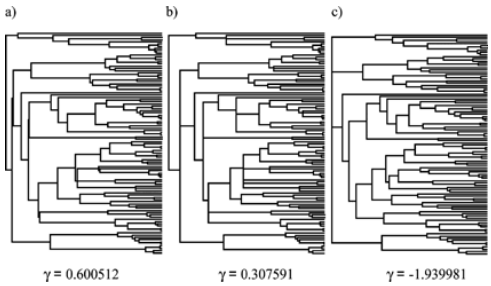

Figure 1.9 | Figure from Revell *et al.* 2005. At the left was the true tree. In the middle the inferred tree, that used the generative model At the left the inferred tree when using a too simple inference model

One example that demonstrates the effect of using a too simple inference model comes from Revell at al Revell *et al.* 2005. They simulated many phylogenies and respective DNA sequences using different DNA substitution models. After this, they inferred phylogenies from the simulated alignments with either the correct or a simpler DNA substitution model. Ideally, the inferred phylogenies match the phylogenies the alignments are based upon. They found that when the DNA model is the correct one, inference of the phylogenies is satisfactory. However, when using an overly simplistic DNA model, the inferred trees show a slowdown in their speciation rates, also when the original trees were simulated with a constant speciation rate. This study shows that a decreasing speciation rate may be attributed to an overly simplistic DNA model, instead of an interesting biological process.

A more recent example that demonstrates the effect of using a too simple inference model is about assuming a wrong clock model. A clock model embodies our assumptions regarding the mutation rates in the history of different taxa. The simplest clock model, called the strict clock model, assumes these mutation rates are equal across all taxa. Using a wrong clock model has a profound impact on the inferred phylogenetic trees, unless we can specify the timing of some early speciation events Duchêne *et al.* 2014.

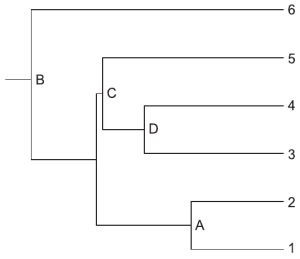

Figure 1.10 | Phylogeny with speciation events labelled A to D, where B is the earliest speciation event. Figure from Duchêne *et al.* 2014.

The tree model is the most important piece of an evolutionary model, with regard to speciation. The assumptions of a tree model is called the tree prior, where 'prior' refers to the knowledge known before creating a phylogeny. The tree prior specifies how likely processes that determine the shape of a tree occur. These two processes are the formation

of a new branch and the termination of an existing branches. In the context of speciation, we call these two events a speciation and an extinction event respectively.

There are two standard tree models, called the Yule (also: pure-birth(and (standard) birth-death model. The most basic speciation model is the Yule model [Yule, 19..] which assumes that speciation is constant and there is no extinction. [Research on fossils with Yule model would be fun]. The Yule model predict that the number of extant species grows exponentially through time.

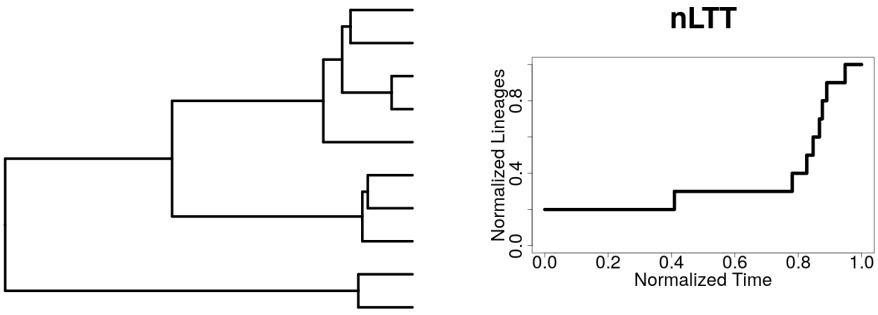

Figure 1.11 | Left: An example Yule tree Right: A lineages-through-time plot of the example Yule tree. In all cases, time goes from past (left) towards the present (right).

The Birth-Death model [Nee et al., 1994] is an extension of the Yule that allows for a constant extinction rate. If the speciation rate exceeds the extinction rate, also the BD model predicts that the number of extant species grows exponentially through time. If the extinction rate exceeds the speciation rate, the number of lineages is expected to decline exponentially. The latter is biologically irrelevant.

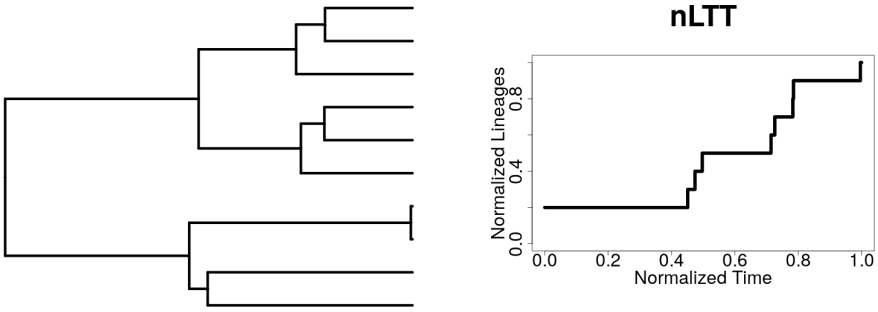

Figure 1.12 | Left: An example Birth-Death tree Right: A lineages-through-time plot of the example Birth-Death tree. In all cases, time goes from past (left) towards the present (right).

It is clear that an exponential growth in the expected number of lineages is biologically nonsense. To state the obvious: a finite area (Earth) results in a finite number of species.

Applying the BD model to molecular data already shows that it does not always hold, see figure [below]

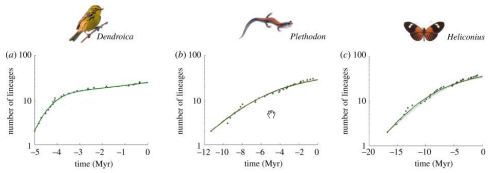

Figure 1.13 | An LTT plot for bird/lizards that shows a slowdown in speciation rate, adapted from Etienne *et al.* 2012. Because the number of lineages on the y-axis are plotted on a logarithmic scale, exponential growth would show as a straight line.

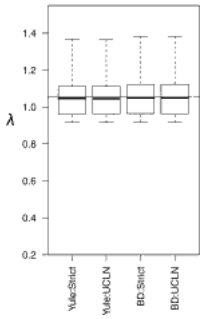

Figure 1.14 | Estimation of the speciation rate (lambda) on inferred trees using 4 evolutionary models. The original trees had 100 taxa and were simulated with a strict clock model and BD tree model, with a speciation rate of 1.104. Adapted from Sarver *et al.* 2019.

A recent study that investigates the effect of picking a wrong standard tree prior, comes from Sarver et al, 2019 Sarver *et al.* 2019. In this study, they first simulate trees using either a Yule or a birth-death tree model, after which they simulate an alignment from that phylogeny using two different standard clock models. From these alignments, they inferred the original trees using all of the four different clock and tree prior combinations. They show that, regardless which priors are used, the estimated speciation and diversification rates from the inferred trees are similar to those of the original tree.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard, novel tree model. I will described the two new biological tree models that have been investigated, as well as the re-usable framework to do so.

The first novel and non-standard tree model is the protracted birth-death model (PBD) Etienne & Rosindell 2012. Where the standard BD models assume that a speciation event creates two new species instantly, the PBD model assumes that one of these two species is an incipient species. An incipient species is a new species that is not yet recognized as such, although complete reproductive isolation is already present. A biological example is from Fennessy *et al.* 2013 in which some new giraffe species have been discovered by sequencing part of their DNA. Although these new species have been 'discovered' recently, the had been no gene flow between species for already two million years.

Using the BD model in species that are slow to speciate, will cause an underestimation of the number of lineages in the present (as in the giraffes), in effect possibly giving the illusion that speciation slows down, where in reality it does not.

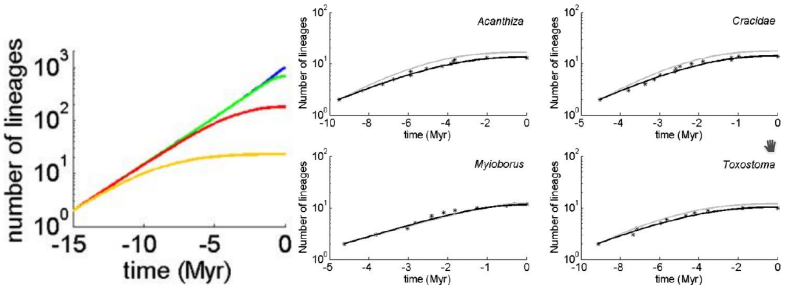

Figure 1.15 | Left: example lineage-through-time plots, for different speciation completion rates: yellow = 0.01, red = 0.1, green = 1.0, blue = 10. Note the slowdown in the accumulation of new lineages when speciation completion rate is lowered. Right: number of species through time plots for four bird phylogenies, (after Phillimore & Price 2008) Both figures are adapted from Etienne & Rosindell 2012

The second novel and non-standard tree model is the multiple-birth death (MBD) model [Laudanno et al., 2020]. Where the standard BD models assume that a speciation event occurs in one species only at a time, the MBD models allows for speciation events to occur in multiple species at the same time. The biological idea behind this model, is that when a habitat (lake or mountain range) gets split into two, this may trigger speciation events in multiple species of both communities at the same time. This mechanism is posed as an explanation for high biodiversity in lake Tangyanika, where the water level rises and falls with ice ages, splitting up and merging the lake again and again, triggering co-occurring speciation events each change.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard tree model, using the phylogenetic software called BEAST2 Bouckaert *et al.* 2019, an abbreviation of 'Bayesian Evolutionary Analysis by Sampling Trees'.

We chose to use BEAST2 Bouckaert *et al.* 2019 over other phylogenetic software, because BEAST2 is popular, beginner-friendly, flexible, has a package manager and a modular well-designed software architecture. The beginner-friendliness comes from the BEAST2 program called BEAUti, in which the user can set up his/her evolutionary model from a graphical user interface. There are many (in the order of dozens to hundreds) options to set up an evolutionary inference model. These choices are categorized in a site model, clock model and a tree prior.

A site model embodies the way the -in our case of DNA- nucleotides change over time. One can specify the proportion of nucleotides that changes, or let it be estimated. And how dissimilar different transition rates may be between different nucleotides. Most essential is the nucleotide substitution model, which entails the relation between the twelve transition rates from any of the four nucleotides to any of the other three nucleotides. The simplest model (called JC69) assumes all are equal, where the most complex model (called GTR) assumes that all may differ. BEAST2 has four site models, yet there is a package that

contains 18 more standard models.

To give an idea about the scale of BEAST2, I will zoom in on the proportion invariants, which can be set to a certain value, or be estimated. If the value is set to a certain value, BEAST2 assumes this as the truth. If the value is to be estimated by BEAST2, then one must additionally specify an initial value and a distribution how likely the different values are. By default, BEAST2 assumes a uniform distribution from 0 to 1, in which all values are equally likely, but there are ten other distributions that can be picked as well. So, for one simple value, there is already a plethora of options (and there more I will not discuss!). Within BEAST2, this liberty is the rule, instead of the exception, rendering it very flexible.

The clock model embodies how the mutation rates vary between different species. The simplest clock model, called the strict clock, assumes that mutation rates are identical in all species. Two, called relaxed-clock, models assume that mutation rates between branches are indepent, yet all rates are from one same distribution. The last standard clock model assumes that all species have a same mutation rate at the same time, yet these mutation rates may vary through time.

The tree prior specifies how a tree is built up, or, in our context, how speciation takes place in time, at the macro-evolutionary level. In our context, these are the Yule and Birth-Death model, which I already described earlier.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard tree model, using practices from Open Science.

Open Science has many facets.

* Open eductional resources * Open access * Open data * Open methodology * Open tools

* Reproducible research: no HARKing

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard tree model. It does so, by using the same experimental setup, called 'pirouette', which is described in chapter 3. This framework is built up a foundation of R packages called 'babette', which is described in chapter 2.

Figure 1.16 | Environment that follows an unknown speciation model.

In the end, we want to know how well we can infer a phylogeny from molecular data found in the field. That field, outside, which follows an unknown speciation model. Let us just hope our inference is robust to whatever novel model we throw at it.

Ik heb even naar de intro en synthese gekeken zoals je me die op 10-2 toestuurde. Met de punten die we gisteren bespraken (o.a. meer detail over de ingredienten van BEAST2),

denk ik dat het een aardige inleiding wordt. Zou je in je inleiding ook kunnen ingaan op je methode van werken? Zowel Open Science en preregistration en het (samen)werken via GitHub?

Het zou ook goed zijn als je aangeeft wat precies de hordes zijn die je over moet bij het maken van packages als babette en pirouette. Ik sprak laatst iemand die dacht dat je een theoretische studie heel snel kan doen, alleen maar even de computer zijn gang laten gaan. Het is goed voor de lezer (zowel je vrienden/familie als de leescommissie) om inzicht te krijgen in het proces van "even de computer zijn gang laten gaan". Een soort "making of" dus. Als het beter past in een box of in de synthese, dan mag dat wat mij betreft ook.

REFERENCES

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.H., Xie, D., Zhang, C., Stadler, T. & Drummond, A.J. (2019) Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, **15**, 1–28.

Bush, R.M., Fitch, W.M., Bender, C.A. & Cox, N.J. (1999) Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular biology and evolution*, **16**, 1457–1465.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A. *et al.* (2012) Biodiversity loss and its impact on humanity. *Nature*, **486**, 59–67.

Duchêne, S., Lanfear, R. & Ho, S.Y. (2014) The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular phylogenetics and evolution*, **78**, 277–289.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1300–1309.

Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, **61**, 204.

Fennessy, J., Bock, F., Tutchings, A., Brenneman, R. & Janke, A. (2013) Mitochondrial dna analyses show that z ambia's s outh l uangwa v alley giraffe (g iraffa camelopardalis thornicrofti) are genetically isolated. *African Journal of Ecology*, **51**, 635–640.

Fuss, J., Spassov, N., Begun, D.R. & Böhme, M. (2017) Potential hominin affinities of graecopithecus from the late miocene of europe. *PloS one*, **12**, e0177127.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K. *et al.* (2016) A new view of the tree of life. *Nature microbiology*, **1**, 16048.

Lam, T.T.Y., Hon, C.C. & Tang, J.W. (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical laboratory sciences*, **47**, 5–49.

Le, S.Q. & Gascuel, O. (2008) An improved general amino acid replacement matrix. *Molecular biology and evolution*, **25**, 1307–1320.

Phillimore, A.B. & Price, T.D. (2008) Density-dependent cladogenesis in birds. *PLoS biology*, **6**.

Revell, L.J., Harmon, L.J. & Glor, R.E. (2005) Under-parameterized model of sequence evolution leads to bias in the estimation of diversification rates from molecular phylogenies. *Systematic Biology*, **54**, 973–983.

Sarver, B.A., Pennell, M.W., Brown, J.W., Keeble, S., Hardwick, K.M., Sullivan, J. & Harmon, L.J. (2019) The choice of tree prior and molecular clock does not substantially affect phylogenetic inferences of diversification rates. *PeerJ*, **7**, e6334.

Stringer, C. (2012) What makes a modern human. *Nature*, **485**, 33–35.

1.1. PHOTO ATTRIBUTION

Figure 1.3 is made by Dbachmann and taken from `https://en.wikipedia.org/wiki/File:Hominini_lineage.svg`.

The phylogeny of figure 1.5 (left side) is by Aglondon, from `https://commons.wikimedia.org/wiki/File:Edge_tree.png`. The Largetooth Sawfish of figure 1.5 (right side) is taken from `fromhttp://www.edgeofexistence.org/species/largetooth-sawfish`

The image of figure 1.16 is from `https://commons.wikimedia.org/wiki/File:The_Earth_seen_from_Apollo_17.jpg`.