Testing hypotheses in macroevolution

CrossMark

Lindell Bromham

Centre for Macroevolution and Macroecology, Research School of Biology, Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

ABSTRACT

Experimental manipulation of microevolution (changes in frequency of heritable traits in populations) has shed much light on evolutionary processes. But many evolutionary processes occur on scales that are not amenable to experimental manipulation. Indeed, one of the reasons that macroevolution (changes in biodiversity over time, space and lineages) has sometimes been a controversial topic is that processes underlying the generation of biological diversity generally operate at scales that are not open to direct observation or manipulation. Macroevolutionary hypotheses can be tested by using them to generate predictions then asking whether observations from the biological world match those predictions. Each study that identifies significant correlations between evolutionary events, processes or outcomes can generate new predictions that can be further tested with different datasets, allowing a cumulative process that may narrow down on plausible explanations, or lead to rejection of other explanations as inconsistent or unsupported. A similar approach can be taken even for unique events, for example by comparing patterns in different regions, lineages, or time periods. I will illustrate the promise and pitfalls of these approaches using a range of examples, and discuss the problems of inferring causality from significant evolutionary associations.

© 2015 Elsevier Ltd. All rights reserved.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Science*

---

1. Experiments in deep time

"The fact that we expect our theories to have exceptions makes it hard to test them… It makes me envious of my colleagues in molecular biology. They can usually settle their problems by experiment: I seem to live with mine. Of course, my problems are more interesting." John Maynard Smith (1990) Taking a Chance on Evolution, New York Review of Books, June 14 1990.

I once had a postprandial argument with a fruitfly geneticist. He said that the work of comparative evolutionary biologists such as myself was all well and good, but it was not real science, because the gold standard of science was the manipulative experiment. If you don't set up a replicated experiment where you apply a treatment to some but not all samples then observe any difference between treatment and controls, then you aren't really doing science,

because any other approach does not allow you to make causal statements. While this "manipulationist" attitude toward establishing causality is less popular with many philosophers of science, it is still promulgated by some experimental scientists, as my conversation with the *Drosophila* geneticist demonstrates (Woodward, 2013). Setting aside the problem of making causal statements from the results of experiments, which is trickier than it first appears, is his statement about comparative evolutionary biology fair?

Experiments in evolution have a long history (de Varigny, 1892). For example, in the 1880's William Dallinger showed that it was possible to increase the thermal tolerance of microbes over many generations by selecting for slight increases each generation, such that the organisms sampled from the end of a multigenerational selection experiment could tolerate high temperatures that would have killed all individuals from the first generation. Experiments on evolutionary change such as this one have built an important body of work in population genetics and behavioural ecology (Buckling, Maclean, Brockhurst, & Colegrave, 2009).

E-mail address: Lindell.Bromham@anu.edu.au.

But there are many evolutionary phenomena that we might wish to study that are not open to experimental manipulation. Here it is helpful to make a distinction between microevolution and macroevolution. Definitions may vary, but I will consider that microevolution describes changes in the relative frequency of heritable traits in a population over generations. On the assumption that the genetic constitution of the population changed over time, as the individuals with a genetic capacity for greater thermal tolerance out-reproduced those less able to cope with the higher temperatures, Dallinger's increase in thermal tolerance in microbes is a classic case of microevolution. In contrast, the study of macroevolution focuses on changes in biodiversity over time, space and lineages, describing and explaining changes in the representation of lineages in the biota. Macroevolution has sometimes been considered a controversial topic, because it was considered by some to represent a challenge to the Darwinian hypothesis that the large scale differences in species found in different times and places are the result of the accumulation of many small genetic changes in populations. When I was an undergraduate we were discouraged from using the word "macroevolution" as it was thought to imply that there were non-Darwinian mechanisms shaping diversity that did not originate in microevolutionary processes (for background to this debate see Sterelny, 2007; Turner, 2011). Some researchers feel that observed patterns of biodiversity in space and time cannot be fully explained in terms of microevolutionary processes and must therefore call upon special macroevolutionary phenomena (e.g. Butterfield, 2007; Carroll, 2000; Erwin, 2000; Gould, 2002). However, these days macroevolution is an entirely respectable way to describe evolutionary studies that focus on patterns of representation of lineages, rather than on changes on genetic variants within single species, irrespective of the ultimate cause of those patterns (e.g. Goldberg, Roy, Lande, & Jablonski, 2005; Levinton, 2001; Nee, 2006; Purvis, 1996). So, in common with many biologists, when I use the term macroevolution I am describing observations about the distribution of biodiversity across large spatial, temporal and biological scales, regardless of the mechanisms that created the patterns of interest.

While the biological patterns under study differ between macroevolution and microevolution, most biologists work under the assumption that these patterns are all generated by the same basic processes: that is, that macroevolutionary patterns are generated by microevolutionary mechanisms acting over long periods of time. In practice, the timescales involved in macroevolutionary change put it well beyond the reach of direct observation or experimentation. Changes in biodiversity over such long timescales are generally not open to manipulation. Even the longest running evolutionary experiments, encompassing tens of thousands of generations of bacteria growing in a laboratory, illustrate phenomena of population divergence but do not generate biodiversity to the degree normally considered under macroevolutionary phenomena (Barrick et al., 2009). The practical upshot of this is that macroevolution is not studied as it happens, but after the fact, by observing the results of naturally acting processes, not by manipulating them directly.

As someone interested in macroevolution, I never perform classic, manipulative experiments. Yet in common with experimental biologists, my aim is to uncover causal relationships, by using hypotheses to generate predictions which are then compared to observation, through careful attention to experimental design and statistical analysis of my data. This view of a scientific test as a comparison of hypothesis predictions to observations can be applied to experimentation, observations, modelling and comparisons, such that we can test a scientific idea by sitting in a bird hide recording behaviour under different naturally occurring circumstances, or by comparing fossils from different time periods, or by using a computer program to simulate the diversification of species under different models of speciation.

Here, I am using "prediction" in the informal sense of using what you know to make an informed guess about something you don't know. In the case of macroevolution, we are not usually in the business of making forward predictions about future events that are yet to happen. We might find it useful in some cases to make future prediction—for example which kinds of species are most likely to go extinct (e.g. Cardillo, Mace, Gittleman, & Purvis, 2006)—but it will rarely be the path to usefully discriminating macroevolutionary hypotheses. Instead, the word "prediction" is often used to refer to the use of prior knowledge to identify the most likely outcome, a process that is agnostic with respect to the timing of the outcome. In other words, a prediction is a statement about what is likely to happen if a particular set of condition is realized, given a particular hypothesis (Cleland, 2002), and it can just as well apply to what did happen in the past when a particular set of conditions occurred as it does to future events brought about by experimental manipulation.

There are two points that I would like to emphasize in this paper. One is that comparative studies, or "natural experiments", do not need to be seen as poor cousins to classical manipulative experiments, nor as imperfect attempts to mimic laboratory experiments. In fact, laboratory, field and comparative tests often employ similar design and analytical frameworks, and generally have the same basic goal of seeking relevant observations that allow discrimination of alternative causal hypotheses (Jeffares, 2008; Morgan, 2013; Okasha, 2011). Indeed, comparative tests or field observations can share not only the strengths of manipulative experiments (e.g. replication, ability to isolate variables of interest) but also some of the weaknesses (e.g. lack of transparency of causal mechanisms, difficulty in controlling covarying factors).

The second main point I would like to illustrate is that, in common with experimental studies, most comparative macroevolutionary studies do not provide a definitive test of a hypothesis. The heroic stories told of science usually focus on world-shaking discoveries (hence the obsession with Nobel laureates) or elegantly decisive "killer tests" that put competing ideas permanently to rest. While there is an entirely understandable tendency to focus on the rather more captivating examples where exciting new discoveries provide definitive answers to big questions, much of the progress is actually made by a more pedestrian accumulation of corroborating evidence from a range of investigations, weighed against case studies where a particular explanation can be rejected as less satisfying than an alternative (Currie, 2014; Stanford, 2011). Each study conducted adds grains to the balance in which hypotheses are weighed. This process of circling round a hypothesis by gathering evidence for and against a particular explanation using a collection of independent tests is not so different from classical experimental science, where the results of well-designed and precisely-executed manipulative experiments are often not as clear cut as might have been hoped.

Take the example of two high profile studies that aimed to test the hypothesis that caloric restriction increases lifespan (Sinclair, 2005). Two independent long-term manipulative experiments were carried out over several decades to discover whether restricting calorie consumption resulted in extended lifespan in primates, conducted by the National Institute for Aging (NIA) and Wisconsin National Primate Research Centre (WNPRC). Although each experiment was designed to be a definitive test of the hypothesis, they came to opposite conclusions: while both reported health benefits of calorie restriction, the WNPRC found that calorie restriction resulted in a 30% increase in lifespan (Colman, et al., 2009), but the NIA found no significant increase in lifespan in calorie restricted monkeys (Mattison, et al., 2012). In addition to

differences in the monkeys tested and the diets used, each research team considered that the others' experimental design had failed to isolate the salient causal mechanism, through inappropriate contrast between experimental (restricted calorie) and control monkeys. The NIA team considered that WNPRC effect may have been driven by allowing control monkeys *ad libitum* access to a highly processed diet, such that the control monkeys shorter life-span could be attributed to the negative health effects of a junk food diet rather than by higher calories *per se*. The WNPRC team consider that by not allowing the control monkeys unrestricted amounts of food, the NIA experiment effectively had two calorie restricted lines rather than a treatment and a control (Austad, 2012). Following these two large experiments we now know more about diet and longevity in captive monkeys: monkeys allowed to eat as much monkey chow as they like tend to have shorter lives than monkeys whose intake is restricted. These find-ings may well be telling us something important about the effect of diet on lifespan, but the experiments have not, as hoped, provided a killer test of the caloric restriction hypothesis.

In this paper, I am not aiming to examine in detail the theoretical or philosophical underpinnings of the nature of scientific testing. Instead, I will present a view from the trenches, using examples from my own research and teaching in the field of macroevolution to highlight a few of the different ways that we conduct tests when investigating evolutionary phenomena that play out over millions of years. The examples I will discuss may seem like an odd collec-tion of tales, but they have been chosen to illustrate a few of the approaches that are used in testing hypotheses without recourse to experimental manipulation. Parasitic plants will demonstrate how it is possible to construct something that looks like a controlled experiment on past processes. Long-lived rockfish will provide an example of the way multiple studies can be combined to reduce the influence of confounding factors and thus weigh up alternative causal explanations. Salt tolerant plants will illustrate the chal-lenges of inferring processes of macroevolution from patterns observed in the present day. Dinosaur extinctions will allow us to ask whether we can do science on single historical events that are not replicated. And the Cambrian explosion will provide a platform for asking whether there are some unique evolutionary events that are so singular that there is no way of drawing general conclusions about patterns and processes of macroevolution.

2. Parasitic plants: comparative studies are designed like experiments

Most plants are autotrophs: they generate their own energy for growth and metabolism by capturing energy from sunlight. But some plants get their energy from other plants by tapping into their host's body. There are a wide variety of parasitic plants that connect to their host plants in different ways: for example mistletoes that grow on the branches of trees; sandalwood trees that connect to their host through their roots; or the giant *Rafflesia* flowers that grow entirely hidden in their host's tissues until they flower.

The reason I became interested in parasitic plants is that it had been noted that some parasitic plant lineages seem to have pecu-liarly fast rates of molecular evolution (Lemaire, Huysmans, Smets, & Merckx, 2011; Nickrent & Starr, 1994). This is interesting for two reasons. On a practical level, faster rates of molecular evolution in parasites makes it more challenging to work out their evolutionary history using analysis of DNA sequences: just as the highly modified morphology of parasitic plants can make phylogenetic placement difficult, so their highly modified gene sequences can make them difficult to place within a molecular phylogeny (Bellot & Renner, 2014). But my interest in parasitic plants was in what they might tell us about the influences on the rate at which

genomes evolve, which is something that I have been interested in for a while (Bromham, 2009).

Various theoretical models have suggested that parasites might benefit from raised mutation rates, as it might increase the rate at which random changes to the genome generate novel ways around their host's immune system (e.g. Haraguchi & Sasaki, 1996). These theoretical predictions received some support from experiments that showed that bacteria subject to constantly changing environ-ments, or to persistent parasite attack, could evolve increased mutation rates (e.g. Chao & Cox, 1983; Pal, Maciá, Oliver, Schachar, & Buckling, 2007). But the costs of increased mutation might limit the long-term viability of the strategy: for every useful mutation, many more deleterious mutations would be generated, potentially reducing the average fecundity of "mutator" lineages when viewed over many generations. Any offspring with a reduced mutation rate would be likely to leave more descendants than its mutator cousins, pushing the average mutation rate back down again.

Given these theoretical models, what would we expect to see in the real world? Could parasitic mutators maintain their advantage, or would they lose their competitive edge under the weight of deleterious mutations? You can make observations of some kinds of bacteria growing under changed circumstances, such as bacteria in hospitals where antibiotics are used (Bjorkholm, et al., 2001), or do experiments on microbial parasites of laboratory animals (Giraud, et al., 2001). But how can we tell if the increased mutation rate seen in some laboratory experiments is a general feature of the natural world? Should the tens of thousands of species of non-microbial parasites in the world, from tapeworms to cuckoos, also benefit from higher rates of mutation? The influence of parasitic lifestyle on rates of molecular evolution is not an experiment you can do in the lab, because you can't turn a free-living species into a parasitic one then let its genome evolve for millions of years to see how it changes over time. But actually, this experiment has already been run many times, unsupervised, in the natural world. This is the basis of the comparative method.

Parasitic plants provide a nice illustration of the comparative method for investigating the influence of parasitism on rates of molecular evolution because we can identify multiple independent origins of parasitism in the flowering plant phylogeny, where a free-living photosynthetic plant lineage has given rise to a parasitic lineage. This is like an experiment that has been run again and again using different kinds of plants. What happens to mutation rates when you turn a lineage from the laurel family into a parasite? Look at the climbing, vine-like dodder (*Cassytha*). What about when you turn a lineage from the forget-me-not family into a parasite to produce the clumping, succulent "sand food" plants such as *Lennoa*? Parasitic plants are highly variable, but if you have enough of these independently derived parasitic lineages you can start to ask whether, despite the great differences between them, there are also any commonalities. Each of these parasitic lineages is like a single experimental treatment line. On its own, it doesn't tell us anything about general patterns, because the mutation rates in any specific parasitic lineage could have been affected by many different factors including chance. Just as in a lab experiment, the key to making general statements is repetition. If the same pattern is seen again and again, in many different backgrounds and cir-cumstances, then you start to believe that it can't just be due to chance. In other words, if you have enough independent trials, you can do a statistical test on the results.

To test the hypothesis that parasitic plants have faster rates of molecular evolution, we compared DNA sequences from 12 sepa-rate parasitic plant lineages, each of which represents an inde-pendent evolutionary origin of parasitism, to their non-parasitic relatives. For each pair, consisting of a parasitic plant and a non-parasitic relative, we asked which lineage has acquired more

changes in their gene sequences since they last shared a common ancestor. If being a parasitic plant has no effect on rate of molecular evolution, then we would expect the parasitic lineage to have accumulated more changes, by chance, in roughly half of the examples. So we are looking for significant departures from the random expectation. Fig. 1 displays the results for genes from the three different genomes. You can see that in most cases the parasite has a faster rate of molecular evolution than its non-parasitic relative. This provides strong support to the hypothesis that parasitic plants have faster rates of molecular evolution than their non-parasitic relatives, because we would not expect this pattern to occur by chance in the absence of a causal link (whether direct or indirect) between parasitism and molecular evolution.

But, as noted by our *Drosophila* geneticist friend, this proves the correlation, but does not reveal the cause. There are actually many reasons why parasitic plants might have faster rates of molecular evolution. Parasitic plant species vary in many ways, and yet may also be similar in other ways that are less obvious. Many parasitic plants are likely to have smaller population sizes than their free-living relatives, and small population size generally results in faster rates of molecular evolution (Charlesworth, 2009; Woolfit, 2009). Parasites also often experience relaxed selection as they jettison the traits needed for independent life. A fully parasitic plant does not need to maintain the equipment for producing its own energy from sunlight so can let the genes responsible for leaf structures, photosynthetic enzymes, and energy producing pathways decay, resulting in a rapid rate of change in these genes (dePamphilis, Young, & Wolfe, 1997). Parasitic plants are typically smaller in size than their free-living relatives, and shorter plants tend to have faster rates of molecular evolution (Bromham, Hua, Lanfear, et al., 2015; Lanfear, et al., 2013). So it may not be parasitism *per se* that is driving this pattern, but some other factor that comes with the whole parasitic package, such as generation time, height, population size, or some other factor we haven't even thought of.

Of course, this problem of confounding factors is not unique to comparative tests. Recall that the dispute over experimental evidence for caloric restriction increasing lifespan in monkeys also turned, at least partly, on confounding factors. Both experiments were designed to test the influence of caloric intake on longevity, but caloric intake could not be easily separated from other aspects of diet that might have an influence on health and longevity, such that the identified effect on longevity might have been at least partly due to the negative influence of living in a cage and eating lots of monkey chow rather than being strictly a function of number of calories consumed. Nonetheless, confounding factors are also a problem for comparative studies in macroevolution. But, just as experimental scientists can react to possible confounding factors by redesigning their experiments to better isolate salient factors, so too can comparative biologists. The next example illustrates one way of using multiple tests to reject some possible causal explanations, thereby narrowing in on plausible explanations for an observed correlation.

3. Long-lived rockfish: using correlations to close in on causal factors

One of the most consistent patterns in genomic evolution is that smaller animals and plants have faster rates of molecular evolution than their larger relatives (Bromham, 2009). There have been many hypotheses put forward to explain this correlation between body size and rate of molecular evolution. One is that smaller animals and plants have more rapid turnover of generations, so they copy their genomes more per unit time, and thus have more chances for replication errors to alter the DNA sequence of the genome (Bromham, Rambaut, & Harvey, 1996). Another is that smaller bodied animals have higher mass-specific metabolic rates, so their tissues produce more DNA-damaging metabolites such as free oxygen radicals, which might generate more mutations (Martin & Palumbi, 1993). An alternative explanation is that because large-bodied plants and animals tend to have longer lives, there is selective pressure to decrease the mutation rate in order to reduce the chances of a life-shortening mutation ruining their chances of successful reproduction (Galtier, Blier, & Nabholz, 2009). Similarly, large animals tend to have fewer offspring, investing more energy into each one, so the loss of a single offspring due to the deleterious effects of mutation represents a much larger loss of fitness, driving selection pressure to reduce the mutation rate (Welch, Bininda-Emonds, & Bromham, 2008). These are all plausible explanations, and they are not mutually exclusive, so two or more of these mechanisms may act in concert to produce the observed outcomes.

To test which of these hypotheses has significant explanatory power, we need to make predictions, then test those predictions against observation. But the problem is that all of these hypotheses make similar predictions, because the causal factors are all correlated: smaller animals often also have shorter generations, more offspring, shorter lifespans and higher metabolic rates. One way to untangle the different effects is to use multivariate statistics to ask what effect each variable has over and beyond its covariation with the other variables. Such tests have revealed that metabolic rate generally does not provide significant explanatory power beyond its covariation with size, generation time, longevity and fecundity (Bromham, et al., 1996; Lanfear, Thomas, Welch, & Bromham, 2007; Welch, et al., 2008). But this multivariate approach has been less helpful for separating out the effects of generation time, size, longevity and fecundity, all of which are significantly correlated with rates. Either these factors all influence rates, or they are so tightly correlated that it's not possible to separate out the effects of the different variables. How can we distinguish these two possibilities?

One approach is to try to find an example where the covariation between traits is broken, which would allow us to separate out the effects of each variable. Rockfish (*Sebastes*) provide an ideal test of the effects of size, longevity and fecundity on rates of molecular evolution. Rockfish are famously long-lived, with some species capable of living for a century or two (Love, Yoklavich, & Thorsteinson, 2002). What's more, they have no apparent senescence: size increases throughout life, and fecundity increases with size, with no apparent decline in offspring fitness with age (de
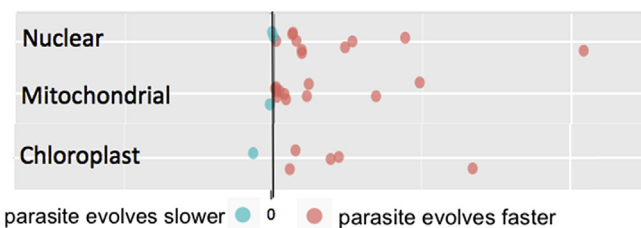

Fig. 1. Rate of molecular evolution in parasitic plants compared to their non-parasitic relatives, estimated from sequences from the three cellular genomes of angiosperms (nuclear, mitochondrial and chloroplast). Each dot represents a comparison between a parasitic plant and its non-parasitic relative. Since each parasite included represents an independent evolutionary origin of parasitism in flowering plants, each comparison represents a statistically independent observation of the relationship between rates of molecular evolution and parasitism. The dot is red if the parasite has the faster rate, blue if the non-parasite has the faster rate. If there were no association between parasitism and rate of molecular evolution, then we would expect roughly equal numbers of red and blue dots for each genome. However, not only are most comparisons positive (parasite has faster rate), but the few that are negative (parasite has slower rate) are close to zero.

Bruin, Gosden, Finch, & Leaman, 2004). So like mammals, reptiles and birds, larger-bodied rockfish species tend to have longer generation times and longer lifespans than their smaller relatives. But unlike mammals, bigger, longer-lived rockfish tend to have more offspring.

Using much the same approach as for the parasitic plants, we compared DNA sequences from closely-related pairs of rockfish species, and asked whether the one with the longer lifespan had accumulated fewer changes in these genes since their last shared common ancestor). Just like in mammals, the longer-lived rockfish species had slower rates of molecular evolution. But this observation is more than just a useful confirmation that the link between longevity and rates of molecular evolution applies to some fish as well as mammals and birds, because it allows us to rule out some of the possible causes (Fig. 2). The longevity effect in rockfish can't be explained as an influence of fecundity, because in these fish fecundity increases with size, so if fecundity was driving the rate variation then we should see the bigger fish having faster rates. Moreover, it doesn't really look like a replication frequency effect either, because longer-lived, more highly fecund fish are likely to have copied their genome more times per unit time. And it doesn't seem to fit with metabolic rate as a driving force either, because fish that live in deeper, colder waters don't have slower rates than those that live in the warmer surface waters. So while the correlation between size and rates is the same in rockfish as in mammals, we can actually rule out three of the four possible explanations for the pattern proposed for mammals (copy frequency, fecundity, metabolic rate).

Does this mean we have proved the fourth explanation (selection for lower rates in longer lived species) correct? Of course not, because there may be alternative hypotheses we haven't even considered that fit the data as well or better. More generally, we can't guarantee that the causal links between longevity and rate of molecular evolution in fish are the same as those in mammals and birds. But we have made progress. We have shown that in rockfish, we can't easily assign the body size effect to a DNA replication frequency effect, but we could explain this pattern in terms of selection for increased DNA repair in longer lived species. So the

findings of this study add some support to the plausibility of the selection-for-longevity model. Further studies may add more support, or they may not. This incremental approach, step by step along explanatory paths, is a common *modus operandi* in evolutionary biology. Generally, no one study provides a definitive answer, but each moves us closer to understanding the major forces shaping macroevolution.

The first two examples illustrated how comparative studies can resemble experiments by considering whether particular effects are repeated in different lineages when a defined change occurs: for example, when a plant lineage becomes parasitic, it usually also has an increase in rate of molecular evolution. Then we considered how we can weigh different causal explanations for the same correlational pattern by using statistical analysis to untangle the effects of different variables, or by contrasting cases where the variables differ in key ways. These comparative methods rely on making comparisons between lineages with different traits, using repeated patterns in present day species to infer likely processes. Is it possible to add a temporal dimension to comparative tests, so that instead of simply saying that two traits tend to be found together, we can determine whether one trait actually causes the evolution of another trait? Clearly we cannot directly witness the acquisition of a new trait and follow its long-term consequences over millions of years. So we must try to infer the occurrence of trait changes and their macroevolutionary effects from present day patterns in biodiversity. One way to do this is to use a phylogeny (evolutionary tree) to infer the repeated origins of a particular trait, and ask whether the fate of those lineages with the trait is different from their relatives without it. We can also use a phylogeny to try to establish temporal patterns of trait acquisition, in order to support or reject different hypothesised causal connections between traits. One complex suite of traits that has evolved multiple times is salt tolerance in plants.

4. Salt tolerant plants: phylogenies can reveal macroevolutionary processes

Salt is toxic to plants and imposes physiological drought by making it difficult to draw water from the soil. Some plant lineages have special adaptations that allow them to reduce the harmful effects of environmental salt, and to mitigate the water stress it imposes, such as compartmentalising or excreting salt, and improving water efficiency (Flowers, Galal, & Bromham, 2010). But salt tolerant plants are relatively rare: less than one quarter of one percent of flowering plants are known to be salt tolerant (Bromham, 2015). Since there is a large amount of salt affected land (up to 10% of the land surface), it would seem that species that could tolerate salinity and exploit these areas would be at an advantage.

The relative paucity of salt tolerant lineages has been interpreted as a sign that it is difficult to evolve. One possible explanation is that salt tolerance is a complex trait requiring many different genes and changes to many physiological and anatomical traits. Furthermore, it has proved surprisingly difficult to breed salt tolerant crops: decades of intense research have thusfar not produced any commercially-viable salt tolerant varieties of major crops (Ashraf & Foolad, 2012; Cheeseman, 2014). Yet salt tolerance has arisen in many different lineages: one third of plant orders contain some salt tolerant species (Flowers, et al., 2010). This suggests that, despite the genetic complexity and physiological costs of salt tolerance, it has evolved in a wide variety of different kinds of plants. So if there are lots of ecological opportunities for salt tolerant plants, and if salinity tolerance can evolve in a wide variety of plant types, then why are there so few halophytes (salt tolerant plants)?

	Mammals		**Rockfish**	
	Large-bodied species	Small-bodied species	Large-bodied species	Small-bodied species
Copy error effect predicts:	⬇	⬆	⬆	⬇
Selection for longevity predicts:	⬇	⬆	⬇	⬆
Observation:	⬇	⬆	⬇	⬆

Fig. 2. Conducting the same comparative test in both mammals and rockfish helps to distinguish two explanations for differences in rate of molecular evolution between species. In mammals, the observed pattern (lower mitochondrial mutation rates in larger-bodied species) is compatible both with a copy error effect (large animals copy their germline less often per unit time) and with a selective explanations (large animals must reduce their mutation rates to be a viable life history strategy). But in rockfish, the same observation discriminates between the two explanations, because larger-bodied rockfish produce more eggs throughout their lifetime, so should have more opportunity to accumulate copy errors. Note that we cannot directly extrapolate the rockfish result to discriminate between causal explanations in mammals, nor can we say we have proven selection for longevity in rockfish, as there may be other plausible explanations not tested here.

Given the importance of developing salt tolerant crops, there is a huge body of experimental science conducted on a range of different species. But these experiments can't address this macro-evolutionary question: why does salt tolerance evolve often yet there are so few salt tolerant species? We explored what the distribution of species on phylogenies could tell us about macroevolutionary patterns of salt tolerance (Bennett, Flowers, & Bromham, 2013; Moray, Hua, & Bromham, 2015). We were expecting to find that salt tolerant species clustered into specialised clades, groups of related species that all descend from a common origin, all adapted to exploiting saline environments. But we were surprised to find that, instead of occurring in groups of similarly adapted species, as might be expected if salt tolerance is a rare and complex adaptation, salt tolerant species were scattered throughout the phylogeny, each one corresponding to a relatively recent origin, and with few or no close salt tolerant relatives. Instead of being difficult to evolve, salt tolerance seems to evolve very frequently in a wide range of lineages. For example, the 200 known salt tolerant grasses arise from over seventy independent origins of salt tolerance. This suggests that, far from being difficult to achieve, salt tolerance is a lot easier to evolve than we might have supposed given the rarity of halophytes. This raises some important questions. If salt tolerance is so easy to evolve, why aren't more species exploiting saline environments? Why is salt tolerance scattered across the twigs of the evolutionary tree, rather than defining large branches? And are these evolutionary patterns in any way connected to the frustrating lack of success in breeding commercially-viable salt tolerant crop species?

One possible explanation for the surprisingly large number of independent origins of salt tolerance is that some lineages may be better equipped to evolve salt tolerance than others. There could be "enabling traits" that, while they don't actually confer resistance to salt, make it easier for a plant to develop salt tolerance (Bromham, 2015). For example, it has long been recognized that there is a non-random association between salt tolerance and C_4 photosynthesis, which is a modification of the standard C_3 photosynthetic pathway that allows more efficient water use (Sage, 2004). But this correlation between photosynthesis pathway and salt tolerance could have different causal explanations: salt tolerant plants could be more likely to develop C_4 (to conserve water) or C_4 plants could be more likely to be able to adapt to saline environments because they are less effected by reduced water availability, or the two traits might be indirectly connected through some other variable (for example, C_4 plants are more common in open habitats which are more likely to be salt affected: Edwards & Donoghue, 2013). Is it possible to isolate not just the correlation between two traits, but the likely causal direction of that link?

When you plot origins of salt tolerance on a phylogeny, the origins of salt tolerance occur within C_4 lineages far more often than you would expect by chance (Bromham & Bennett, 2014). This would appear to be a clear case of directionality allowing a correlational statement to be turned into a causal statement. It looks like an experiment where you apply a treatment (evolve C_4 photosynthesis) and get a response (evolve salt tolerance) that occurs significantly more often in the treatment (C_4) lines than the control (C_3) lines. Actually, we can't be quite so confident of this conclusion, because we can only reconstruct the most recent origins of salt tolerance, those that that have led to living salt tolerant species, with any confidence. But if we are right in our hypothesis that salt tolerance evolves very often, but doesn't persist, then we have to assume that it has also evolved often in the past, even if those past origins left no present day descendants. So, while the pattern of correlation on the phylogeny suggests that C_4 precedes salt tolerance, we can't rule out the possibility that C_4 photosynthesis is more likely to arise in salt tolerant lineages (Fig. 3).
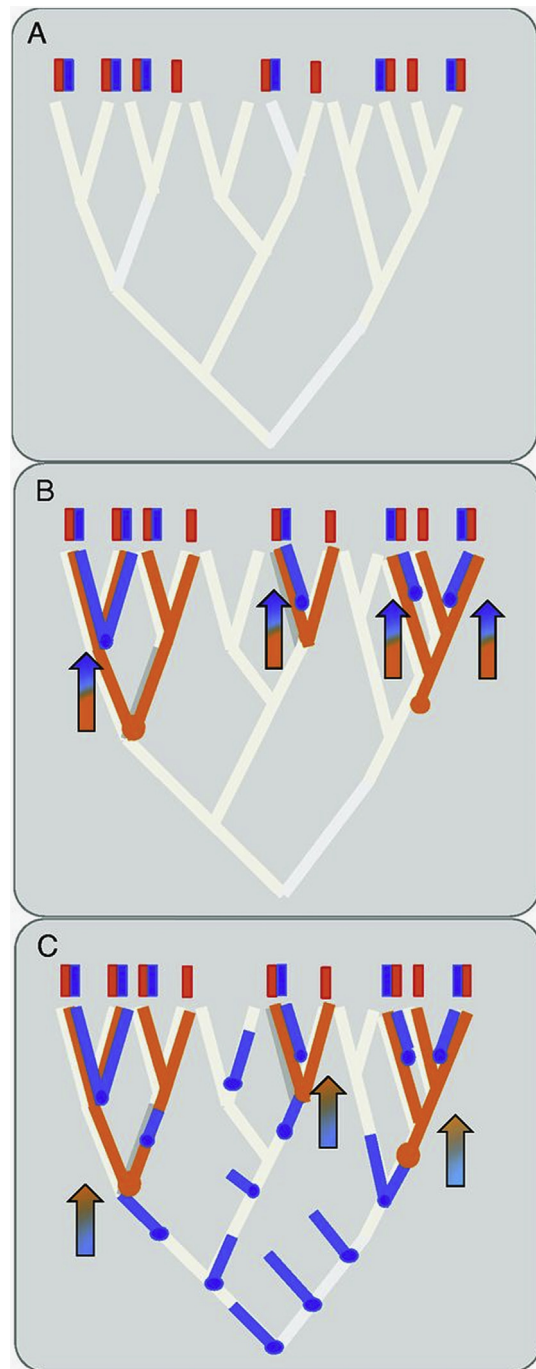

Fig. 3. Phylogenetic tests of evolutionary links between traits. You can detect significant correlations between traits if you plot them on a phylogeny and find they co-occur more often than you would expect by chance (A). If you could determine a consistent order of trait acquisition, you could turn this correlation into a causal statement. In this case, there are fewer, deeper origins of orange inferred, and most more recent origins of blue occur within orange lineages (B). One interpretation is that orange always precedes blue, therefore giving a direction to any inferred causal connection. But, given that blue has many recent origins, it appears to be relatively labile, being gained and lost often. Unless this is due to a recent environmental change, we can assume that it has also been gained often in the past, then lost again (by reversal or extinction), in which case it is possible that orange always arose in a blue lineage (C). Figure reproduced with permission from (Bromham, 2015).

Correlation between salt tolerance and "enabling traits" such as C_4 might explain why salt tolerance is easier to evolve than might be expected. But why aren't there more halophytes? And why are most of the salt tolerant lineages fairly young, with each new evolutionary origin giving rise to only a few descendants, instead of defining large groups of salt tolerant specialists? One possible explanation is that salt tolerance is easy to evolve but hard to maintain, because the stresses imposed by saline conditions and the investment in tolerance mechanisms limit growth rates to an extent that makes salt tolerant species uncompetitive in the long run (Munns & Gilliham, 2015). Under this hypothesis, new salt tolerant species that can exploit saline habitats are constantly arising, but few salt-tolerant lineages persist or diversity. Because it concerns long term costs and benefits, this is a macroevolutionary hypothesis that cannot be settled through conventional manipulative experiments. Experiments will be very helpful for quantifying the physiological costs of salt on plant growth, or comparing the different mechanisms employed by plant species, but they wont allow us to evaluate how these costs play out over evolutionary time in diverse ecosystems.

I want to use our attempts to explain the patterns of evolution of salt tolerance to illustrate another common approach to evaluating macroevolutionary hypotheses, which is to compare observed patterns to the patterns expectation under different macroevolutionary models. The first question we wanted to ask was: just how surprising is it that salt tolerance is scattered across the twigs of the phylogeny, rather than defining large branches of salt tolerant species? Could this pattern have arisen simply by chance even if salt tolerance has no particular effect on tempo and mode of macroevolution? What kinds of macroevolutionary process could create a pattern of a trait on a phylogeny such that the inferred origins of the trait are nearly all on the "twigs of the tree of life" and rarely on the deeper branches (Schwander & Crespi, 2008)? To work out just how surprising a particular evolutionary pattern is, we need to develop a null model that tells us what range of outcomes we should expect to occur just by chance—in other words, we need a formal way of evaluating just how surprising the observed pattern really is. Stochastic models of clade growth have a long history in evolutionary biology (Nee, May, & Harvey, 1994; Raup, Gould, Schopf, & Simberloff, 1973), and have been instrumental in testing claims about the significance of observed macroevolutionary patterns (e.g. Mooers, Gascuel, Stadler, Li, & Steel, 2012; Pybus, Rambaut, Holmes, & Harvey, 2002; Rabosky, 2009).

To generate the expected distribution of traits on phylogenies under different macroevolutionary scenarios, we used computer programs to simulate the evolution of lineages. In these simulations, each lineage has a defined probability, at any given point in time, of going extinct, or of speciating to produce two daughter lineages, each of which also has a chance of speciation or extinction in each subsequent time step. These simulated lineages can also gain or lose a hypothetical trait with a defined probability. Importantly, we can also specify that gaining a particular trait influences the probabilities of speciation or extinction (Maddison, Midford, & Otto, 2007). So we can evolve hundreds of alternative evolutionary histories, each representing one possible outcome of evolution under a particular set of conditions. Then we can ask whether any of these simulations produces patterns similar to those we observe in the real world.

When we applied this null model approach to the distribution of salt tolerant species across the grass family phylogeny, we found that the observed pattern of salt tolerance was much more "tippy" than expected (Bromham, Hua, & Cardillo, 2015; Moray, et al., 2015). This suggests that something is causing salt tolerant lineages to be younger and less species-rich that we would expect if salt tolerance had no effect on lineage persistence or diversification. To compare different possible macroevolutionary explanations of this pattern, we evolved thousands of simulated phylogenies under different evolutionary models, involving different combinations of speciation and extinction rates and trait transition rates. We found that we could reject all alternative models as inconsistent with the observed phylogenetic patterns, except for models with a high rate of gain of salt tolerance, and an even higher rate of loss. In other words, this twiggy pattern of salt tolerance is just what we would expect to see if salt tolerance evolves often but then is lost again almost immediately. In this way, there are many salt tolerant lineages, but they are mostly very young. We can reject all the alternative models we tested because the simulations suggest they cannot produce patterns like the observed pattern (Bromham, Hua, & Cardillo, 2015). Interestingly, one of the models that we tested and rejected as being inconsistent with the data is the "dead man walking" scenario, where frequent gain of salt tolerance leads to higher extinction rates. This high gain-high extinction model is invoked to explain the observation that most asexual lineages are fairly young: asexuality arises often but limits the ability of the lineage to persist and diversify (Maynard Smith, 1978; Williams, 1975).

Finding a model consistent with the observed pattern, and rejecting all other tested models, doesn't prove that we have identified the correct explanation, but it does tell us that the pattern we observe could result from the macroevolutionary dynamics of a trait with frequent gain and rapid loss. For all we know, the models tested may be unrealistic—for example, the speciation and extinction values might be too different to those in the real world. Null model tests are only as good as the null model tested. But, like the previous examples, they provide useful way of asking whether the hypothesized process could have produced the observed pattern.

All of the previous examples relied upon being able to mimic experimental design by looking at repeated instances of the evolution of a particular trait—parasitism in plants, longevity in rockfish, salt tolerance in plants. We could then use statistical tests to ask whether there was a significant correlation between evolutionary outcomes: for example parasitism results in raised rates of molecular evolution more often than expected by chance, C_4 plant lineages are more likely to evolve salt tolerance. We used multiple tests across different groups to rule out some alternative hypotheses as inconsistent with observations, which allowed us to narrow in on more plausible causal explanations: for example, rockfish show that the correlation between longevity and mutation rates is general across many vertebrate groups, but it isn't likely to be the result of consistent patterns of copy frequency or metabolic rate. But there are many evolutionary phenomena that we wish to study that do not occur again and again in different lineages. We might want to study the impact of a particular event that occurred in a specific place and time. One of the best examples is the final extinction of the dinosaurs at the end of the Cretaceous period.

5. Dinosaur extinctions: testing multiple predicted effects of a single event

In popular parlance, the term dinosaur means a large, lumbering reptile, now extinct. Actually, there are lots of large extinct reptiles that are not dinosaurs (such as sail-backed lizards, pterosaurs, plesiosaurs) and lots of small fast-moving species of dinosaurs that are not extinct (birds). But I am going to use "dinosaur" in the common sense of the word, to refer to all non-avian members of the Dinosauria (which defines a taxonomic group containing all the descendants of the last common ancestor of the triceratops and the sparrow). There were many hundreds of different dinosaur species, which arose and went extinct throughout the Mesozoic ("Age of

Reptiles", 252 to 66 million years ago). But the disappearance of the last non-avian dinosaur lineages, along with the plesiosaurs (aquatic reptiles) and pterosaurs (flying reptiles), at the boundary between the Cretaceous and Palaeogene strata, 66 million years ago, has puzzled biologists for over a century. This boundary is known variously as the K—Pg (Cretaceous—Paleogene) or KT (Cretaceous—Tertiary) boundary.

The discovery of raised iridium levels at this boundary led to the hypothesis that a massive extra-terrestrial object hit the earth at the end of the Cretaceous period, wiping out the dinosaurs in a stroke of cosmic bad luck (Alvarez, Alvarez, Asaro, & Michel, 1980; Schulte, et al., 2010). This hypothesis has become so widely accepted that it is now commonly cited as a fact, particularly in public discourse on dinosaur extinctions (Miller, 2014). The impact extinction hypothesis has been used as an exemplar of successful hypothesis testing in the historical sciences (e.g. UnderstandingScience.org, Cleland, 2002). However, the impact extinction hypothesis is not the only possible explanation for dinosaur extinctions. For example, some scientists think that massive volcanism was the trigger of dinosaur extinctions, or rapid climate change, or that they were naturally in decline anyway (Archibald, et al., 2010; Courtillot & Fluteau, 2010).

How can we test the veracity of an explanation of a single past evolutionary event? The impact extinction hypothesis proposes that a discrete event—the impact of an extra-terrestrial bolide (such as an asteroid or comet) that happened on a particular day at a particular time about 66 million years ago—changed the outcome of evolutionary history by wiping out some large, previously successful clades, changing the world forever and allowing other clades to flourish in their places (including our own mammalian lineage). We can't go back in time and witness the event happening, nor can we rerun the event. All we can do is piece together the story from the fragments left in the present day.

The end-Cretaceous impact was a one-off event, so we cannot look for the kind of correlations we considered in the examples above, where we asked whether the same effect was observed when the same evolutionary event was repeated again and again. But we can ask whether similar events produce similar effects. The Chicxulub crater on the Yucatan peninsula, which suggests the impact of an extra-terrestrial object around 10 km wide, is currently considered the most likely source of the iridium layer. This is the biggest known impact crater from the Phanerozoic, the last half a billion years of earth's history that marks the diversification of the animal and plant kingdoms. But there were other bolide impacts during the Mesozoic. For example, the Kara crater in Russia is estimated to be half as big as Chicxulub and has been dated to around 70 Million years ago. This suggests that the dinosaurs lived through a large impact without any extraordinary extinctions only 5 million years before the end of the Cretaceous. Mesozoic craters of similar size to Kara are reported from South Africa and Canada. There are also a number of large craters dated to the Cenozoic ("Age of Mammals", 66 myr ago until the present day). A "meteor shower" marked by several large impacts around 36 million years ago includes the Popigai crater in Siberia, two thirds as big as Chicxulub yet with no associated increase in extinctions. So while the Chicxulub impact itself is a non-repeatable event, we might ask why the proposed biotic effects of that bolide impact are not matched by other large impacts (White & Saunders, 2005).

Just as correlation does not prove causality, in this case lack of similar effects across different events doesn't disprove causal connection between the Chicxulub impact and the dinosaur extinctions. Chicxulub is the biggest known Phanerozoic impact crater, so maybe it was greater than some threshold size over which bolide impacts cause global destruction. Maybe there was something particularly nasty about the Chicxulub impact, for example a low entry angle could have kicked up more dust and caused a more severe "impact winter", or the particular point of impact could have been, unluckily, particularly devastating for global biodiversity. Maybe it really was a one-off and comparisons with other impacts are uninformative. Even if it was a one-off, we can still test the hypothesis in a non-statistically replicated way, by using the hypothesis to make predictions then ask if these predictions are met by observations.

For example, the "killing mechanisms" of the impact are said to include a deadly thermal pulse as hot ejecta from the impact sent a burst of radiant heat to the surface of the earth, and acid rain caused by the ejection of oxides of sulphur and nitrogen into the atmosphere (Schulte et al., 2010). The predicted effect is that any unsheltered animals and plants would be killed at or soon after the moment of impact (Morgan, Artemieva, & Goldin, 2013). What observations support this prediction? No large-bodied terrestrial animal species are known to make across the boundary that marks the end of the Cretaceous period. Not only do the dinosaurs and pterosaurs bow out at the K—Pg, so do the larger crocodiles, and many bird and mammals species also fail to persist from the late Cretaceous to the early Tertiary (Robertson, McKenna, Toon, Hope, & Lillegraven, 2004). A spike in fern spores at many locations has been taken as evidence that much of the vegetation was burnt away by the impact, leaving space for ferns to colonize (Vajda, Raine, & Hollis, 2001). What observations speak against this prediction? Animals that should have been sheltered from the blast, such as aquatic reptiles and burrowing dinosaurs, suffer extinction, as do animals that might have survived as buried eggs, such as some pterosaurs (Varricchio, Martin, & Katsura, 2007). Lineages that seem to be particularly vulnerable to acid rain, such as amphibians and fish, show no evidence of mass extinction at the K—Pg boundary (Friedman & Sallan, 2012; MacLeod, et al., 1997). Modelled effects of ocean acidification have been considered insufficient to cause widespread marine extinctions (Tyrrell, Merico, & McKay, 2015). There is currently little evidence to support a mass burning of vegetation (Belcher, Collinson, Sweet, Hildebrand, & Scott, 2003).

We can apply the same kind of prediction-observation tests to other hypothesized effects of the impact, such as an "impact winter" generated by a layer of ejected material in the atmosphere that might have blocked sunlight for years after the event. The predicted effect would be that species dependent on sunlight for energy (e.g. photosynthesizing plants) or warmth (e.g. "cold blooded" reptiles) would not survive. Just as the predictions of the immediate effects of the impact, there are both observations that speak for and against the impact winter hypothesis. While there is turnover of green plant taxa in some areas, there is no global plant mass extinction (Spicer & Collinson, 2014). Polar dinosaur species that must have been able to survive long, dark winters also went extinct, yet many reptiles dependent on environmental heat survived the K—Pg boundary, including snakes, lizards and small crocodiles (Buffetaut, 2004). While some marine plankton suffered mass extinction (planktonic foraminifera), others do not seem to have been badly affected (e.g. benthic foraminifera, dinoflagellates: Alegret, Thomas, & Lohmann, 2012).

So while the hypothesis that dinosaur extinctions were caused by a massive extra-terrestrial impact seems plausible, and some patterns of extinction support the hypothesis, many of the predictions of the hypothesis are curiously unsupported by observations concerning the distribution of extinctions in time, space and across lineages. What are we to make of these counter-observations? Unlike the comparative examples above, we cannot apply a formal statistical procedure to weigh up these observations for and against hypotheses, because the observations do not represent statistically independent instances of treatment and

response. Instead, we must weigh up the evidence informally: how many supporting observations do we need to make a hypothesis convincing? Is a single observation counter to prediction enough to reject a hypothesis? If not, how many observations that don't fit predictions do we need before we consider a hypothesis untenable as a causal explanation? Clearly, these are matters of opinion and subjective judgement.

One way to get around the subjective weighing of supporting and refuting observations is if we could identify a "killer test" that could provide a definitive way of confirming or denying the hypothesis. Is there a killer test for the impact extinction hypothesis? How about if we found reliable evidence on non-avian dinosaurs from after the K–Pg boundary? This would tell us that some dinosaurs survived the short and medium-term effects of the Chicxulub impact, but it wouldn't rule out that the impact had wiped out all the other dinosaurs, even if not all dinosaur lineages succumbed immediately (Fastovsky & Sheehan, 2005).

The opposite pattern looks more promising as a killer test: if dinosaurs were already all extinct by the time the bolide hit, then it can't have caused their mass extinction. But here we run into problems of temporal resolution. The impact happened on a particular day at a particular time. But dinosaur fossils are rare in time and space: most species are known from few fossils, and we can assume that there are many other dinosaur species yet undiscovered. Furthermore, relatively few places on earth contain dinosaur fossils. In particular, while late Cretaceous dinosaur remains have been found in many places in the world, dinosaur fossils from very last stages of the Cretaceous, just before the impact, are primarily known from the Midwest of north America, so we do not have a clear picture of whether dinosaurs in the rest of the world had already declined or disappeared by the time the bolide hit (Archibald, 2014; Brusatte, Butler, Prieto-Marquez, & Norell, 2012). Even the famous Hell Creek formation, which has an apparently continuous record of vertebrate fossils from the latest Cretaceous through the boundary to the early Paleogene, has been interpreted by different researchers as either supporting a pre-impact decline of dinosaur diversity, or as evidence of an instantaneous mass extinction at the K–Pg boundary (Archibald, 2014; Lyson et al., 2011; van Loon, 2012). If it could be proved beyond doubt that all non-avian dinosaurs had already disappeared before the impact, then that would be a killer test that would rule out the impact extinction hypothesis. But in reality we may never be able to obtain the necessary evidence, if we don't know for sure whether lack of dinosaur fossils just before the boundary is due to their disappearance from the biosphere or their failure to leave fossils. So even an apparent killer test can be reduced to weighing up plausibility due to the unavoidable limits on resolution of the fossil record.

All of the above examples relied on the idea of replication: if we are to establish causal connections, then we need to make a connection between cause and effect over multiple examples. This may rely on comparing repeated instances of evolution of a given trait, or looking for a consistent, statistically-significant link between particular traits, or seeking similar effects of an environmental perturbation on many different lineages, or a detecting repeated pattern at many different locations or time periods. The dinosaur impact extinction hypothesis illustrates how even for a hypothesis concerning a single, unrepeated event, we can still devise multiple tests and weigh up the evidence over many locations and lineages, and compare to similar events at other times and places. But what if there was a key evolutionary event that was so singular, so entirely dependent on the particular state of organisms and environment at that particular point in time, that it has no parallels in any other event? Claims of this kind have been made for the diversification of animal phyla in the early Cambrian period (541 − 485 million years ago).

6. Cambrian explosion: even unique events are open to investigation

The origin of the animal phyla has been described as the greatest unsolved mystery in evolutionary biology (Schopf, 2001). The animal kingdom is divided into at least 30 phyla, and some biologists consider that these represent fundamental "body plans" for animals. Examples of phyla with characteristic body plans include Arthropoda (six legs, jointed exoskeleton), Echinodermata (radial symmetry, water vascular system) and Chordata (head at one end, tail at the other, nerve chord running down the back). The first undisputed fossils of nearly of all of the readily-preservable animal phyla appear during the Cambrian period (many phyla of small soft-bodied animals have no fossil record: Valentine, 2004). Some have interpreted the near-simultaneous appearance of members of the modern animal phyla, perhaps within a period as short as ten million years, as the sign of a remarkable evolutionary event, the like of which has not happened before or since (Marshall, 2006). There are no other instances in the fossil record of this degree of morphological change in such a short period. The most extraordinary evolutionary radiations don't even come close. Consider the Hawaiian honeycreepers or Darwin's finches on the Galapagos: in both cases a large number of different species have evolved in less than 10 million years, with variation in size, shape and habits, which allow them to exploit different niches. Yet these remarkably rapid radiations produced slightly different types of birds, not entirely different body plans.

There is a huge body of scientific work on the Cambrian explosion, and the more work that is done the clearer our view is of this remarkable period of evolution (Budd, 2013). There have been a great many hypotheses for the explosion of animal forms in the Cambrian fossil record (mostly not mutually exclusive) and it would be impossible to review them here. Instead, I would like to briefly consider the implications of attempting to explain why there is a burst of diversity in Cambrian (why not before?) and why there are no equivalent degrees of evolutionary change in any subsequent period (why never again?). Some do this by proposing a particular series of events or set of conditions that only occurred at that time and not before or since (such as an enabling rise in oxygen levels or the melting of "snowball earth": e.g. Knoll & Carroll, 1999). Other explanations focus on ecological changes triggering a rapid cascade of events that led to the establishment of familiar trophic structures (e.g. Butterfield, 2011). Some propose that genetic or developmental capacity for evolutionary change was at its highest in the Cambrian, prompting an explosive diversification of forms, but then became locked in to canalized networks, preventing the further exploration of body plan space (e.g. Peterson, Dietrich, & McPeek, 2009). Other explanations have rested on the evolutionary origin of a "key innovation" which then allows the animal lineage to evolve a wider range of forms (for example features of development: Giribet, 2002). An alternative view of the Cambrian explosion is that the differences between phyla are the result of long period of accumulation of differences that began when the phyletic lineages diverged in the deep past but somehow left no unambiguous traces of their early diversification in the Precambrian fossil record. Under this hypothesis, the explosion of phyla is in the early Cambrian is a burst of recognizable forms rather than marking the origins of the lineages (Fortey, Jackson, & Strugnell, 2003).

All of these hypotheses suggest non-repeating events due to the historicity of evolution (Williams, 1992). We don't see phyla forming in other periods (or the present day) either because the conditions that triggered the evolution of the phyla were never repeated, or because the formation of phyla relied on a specific evolutionary change that happened at a particular point in time, or because phylum-level differences represent half a billion years of

accumulated differences so cannot form in more recent time periods. If hypothesis testing requires repeatability and predictability, then are these explanations of a unique event inaccessible to testing? At one end of the spectrum we can consider uniformitarian hypotheses that assume that evolutionary change occurs by the same basic mechanisms in all periods, albeit faster or slower at various times. In that case, we can observe the same evolutionary mechanisms operating today, making the assumption that the process of divergence we can observe between *Drosophila* populations in the laboratory, or parasitic plant lineages in the field, is the same process that has led, over vast time periods, to the diversification of the phyla (Coyne, 2006). At the other end are non-uniformitarian hypotheses that suggest that conditions or opportunities at a particular point in the past were so entirely unlike conditions in the present day, or recent time periods, that we cannot expect to be able to extrapolate from observations in any other time, place or lineage (Carroll, 2000; Erwin, 2011).

I am optimistic enough to think that there is always a toe-hold we can get on even the most distant and singular evolutionary events. Such events may be viewed through a glass darkly, such that we can rarely expect to find an unambiguous trace of past processes. But, in common with much of evolutionary biology, we can circle around hypotheses asking what we would expect to see if they were true. For example, if the invention of a particular genetic architecture triggered the Cambrian explosion, but then became canalised disallowing future change of the same magnitude, then even if we can't go back and watch that process unfold half a billion years ago, we can consider the fate of lineages without that architecture, and we can investigate the ability of that genetic architecture to change in subsequent periods. The hypothesis that expansion of the homeobox genes led to body plan diversification in the bilaterian animals has been supported by an increasingly sophisticated understanding of the role that these genes play in body patterning during development (Holland, 2015), however metazoan lineages without this rich variety of homeobox genes are not noticeably less diverse than those with, and many inferred expansions of homeobox genes in the post-Cambrian period are not associated with body plan evolution (Bromham, 2011). These observations suggest that although investigation of the role of homeobox genes will continue to illuminate the processes of animal evolution, it seems unlikely to provide killer tests of hypotheses concerning the Cambrian explosion.

Currie (2015a) is similarly optimistic about making explanatory progress in the "historical sciences" by drawing together on many different sources of evidence. Even single events have multiple, contingent outcomes, so the coherency between lines of evidence builds a picture of the plausibility of explanations. This approach from coherency need not be seen as something different from classical hypothesis testing, given that these contingencies can be expressed as predictions: "if this hypothesised cause of a certain observation is true, then I should also expect to see…". Acknowledging the challenges of untangling complex webs of causation in deep evolutionary history does not necessitate being pessimistic about our ability to make progress.

7. Predictions and tests: the bright future of "historical" science

"So long as we feel sure that in existing nature we have a key for interpreting the mysteries of the past, we need never despair: whereas, had the causes acting in the remoter ages differed in either kind or degree from those now operating, our science must forever have continued one of mere conjecture and ingenious speculation"

Lyell C (1861) A manual of elementary geology: or the ancient changes of the Earth and its inhabitants as illustrated by geological monuments. 6th Ed. New York: D. Appleton and Company.

Any readers who have made it this far may be wondering what conclusions can possibly be drawn from this varied collection of research projects. I hope that these examples have illustrated a few of the approaches used for hypothesis testing in macroevolution. The parasitic plants illustrated how repeated evolution of a particular state allows us to construct a replicated, statistically sound "experiment" to test evolutionary links between species traits (in this case, parasitism and rate of molecular evolution). The rockfish showed us that conducting similar tests on many different groups allows us not only to identify general patterns, but also to break down some of the co-varying relationships, allowing us to reject some explanations and get closer to a plausible causal explanation for correlations between traits (in this case, longevity and mutation rate). The salt tolerant plants showed us how distribution of species on a phylogeny can be used to generate and test hypotheses about macroevolutionary processes, such as identifying traits that evolve often but do not persist, by comparison to the expected pattern of traits under various alternative models of the evolutionary process. We also saw how phylogenies can potentially reveal causal connections by highlighting order of acquisition of traits, although care must be taken to consider the effect of past changes that cannot be inferred from current patterns (such as origins of salt tolerance that do not lead to present-day salt tolerant species).

These three examples all relied on the evolutionary phenomenon we are interested in being repeated. But what about investigating unique events? The impact extinction hypothesis for dinosaur extinctions showed how even if an event is unique, we can generate multiple predictions and compare to other similar events, as long as we can make the assumption that the processes occurring in one event are shared. For example, the survival of amphibians and fish across the K–Pg boundary speaks against the hypothesis that impact-induced acid rain was a driver of mass extinction, on the assumption that frogs of old were as sensitive to acid rain as frogs of today. The Cambrian explosion example illustrated how the plausibility of hypotheses for unique events can be weighed by considering what additional observations we would expect to make if they were true. The implicit assumption of shared process leading to repeatability of outcomes, whether justified or not, is what unites experimental and comparative evolutionary biology. The fruitfly geneticist's manipulative experiments provide definitive proof that particular microevolutionary processes can occur, whereas my comparative macroevolutionary tests are aimed at working out what actually did happen in the past and why.

Evolutionary biology is commonly described by philosophers as a "historical science". This label might risk giving the inaccurate impression that evolutionary biologists are primarily interested in reconstructing the series of historical events that have shaped life on earth. Some are, but many aren't. There are several broad senses in which history imposes on research into evolution. First, some (but by no means all) evolutionary research is aimed at reconstruction of past events, such as "What happened at the K–Pg boundary?". Second, many evolutionary processes play out over timescales that do not permit direct observation or manipulation, therefore can only be studied through the traces of the past left in present day observations, for example using phylogenies of extant plant species to ask "Why are salt tolerant species rare?". Third, evolution is at all times contingent on past states, thus any evolutionary account must take historicity into account: the question

"Why is there a burst of animal diversity in the Cambrian fossil record?" can only be answered by considering the state of animal lineages in the preceding period.

Examinations of the explanatory power of the "historical science" of evolutionary biology have often focused on iconic case studies, such as the dinosaur extinctions and Cambrian explosion, which is unsurprising given that they are the most intriguing and perplexing historical events and as such present particular challenges to scientific investigation (Cleland, 2013; Turner, 2014). But there is an alternative, complementary approach, which I have previously referred to as the small picture approach to the big picture (Bromham, 2011): sometimes an effective way to test ideas about macroevolutionary processes is to ask whether the predicted effects are observed in particular lineages, places or time periods. Such small picture tests will hardly ever provide a killer test or a smoking gun. But there are good reasons to pay some attention to less exciting case studies, such as those presented in this paper. These case studies are often chosen, or at least can be selected *post hoc*, because features of the case provide particular leverage on a question. Just as paleontologists reconstructing past organisms or events may be "methodological omnivores" (Currie, 2015b), combining a range of approaches and techniques to build a picture of organisms or environments we cannot directly witness, so evolutionary biologists may obtain a wide diet of case studies and types of data to reconstruct past events and mechanisms. Some explanations of historical events might be discriminated by finding a "smoking gun", an evidential trace that could not have existed in the absence of a particular event (Cleland, 2002). But progress in explaining evolutionary past and process does not rely upon isolating such uniquely decisive pieces of evidence: it can be built by assembling many observations, none of which is in itself decisive, into a coherent picture (Currie, 2015a).

Assembling a set of small picture tests might also help to overcome data limitations of core foci of investigation. There may be an absolute limit to the degree to which dinosaur remains in the latest Cretaceous can discriminate the impact hypothesis from alternative hypotheses, but there is a great deal of other corollary evidence we could draw on, such as the extinction rates in other less sexy lineages, or the biotic influences of other impact craters. A small picture approach allows the investigator to choose the case study that maximizes available data. Although it is commonly assumed that both the quantity and quality of evidence of evolutionary events declines with time elapsed, this is not always strictly the case. Evolutionary processes and events can be difficult to investigate at the "shallow end" when the recency of evolutionary events has left a paucity of traces (e.g. recent diversification may have low signal), as well as at the "deep end" where the passage of time has erased the traces of the past (e.g. ancient diversification may have high noise). For example, estimating molecular dates of divergence may be challenging both for recent events, where there may be high amounts of polymorphism and few substitutions, and for deep events, where multiple hits can erase phylogenetic signal (Ho et al., 2011). Researchers overcome these limitations by choosing the data and the methods that suit the event or process under investigation: analyse slowly evolving genes under a phylogenetic model for deep splits and fast evolving loci under a coalescent framework for shallow events (Bromham, 2008). Similarly, while we might expect fewer informative fossils for very deep events, there is not a strict decay in palaeontological evidence with time, as the quantity and quality of material depends on the discovery of fossil beds with appropriate preservational environments for the taxa in question. One of the reasons that the Cambrian explosion is so fascinating is that the exceptional preservation in a number of localities provides an astounding level of detail into diversity and morphology through the preservation of soft tissues of animals that lived over

half a billion years ago. We can select case studies, taxa, time periods or regions that give us the clearest window on the processes we are interested in, choosing those that maximize the signal or reduce the noise.

Both experimental and comparative scientists would like to conduct "killer tests" that unambiguously discriminate between hypotheses, and the same amount of care needs to be invested in the comparison of treatment (what happens if…) and the control (what happens in the absence of…). Determining an appropriate comparison between treatment and control is as much an issue for comparative tests as for manipulative experiments. For a comparative experiment, the "control" takes the form of the expected distribution in the absence of a causal relationship or key conditions. In the three comparative studies illustrated here, we ask whether the distribution of traits in extant species is compatible with a model where no link between the traits exist. In the parasitic plants and rockfish, our null expectation is that if there is no link between rates of molecular evolution and the trait of interest (parasitism, longevity), then when we compare pairs of lineages with and without the trait, there should be a random distribution of rate differences between them. If we see a non-random association between traits and rates, then we conclude that there is some link between the two. In the salt tolerance example, we generated a null distribution by asking how would we expect these species to be distributed on the phylogeny if there was no link between salt tolerance and diversification rate or trait transition rate.

Just as in manipulative experiments, the experimental design of a comparative study may fail to exactly discriminate causal mechanisms, and so the results of comparative tests must likewise be interpreted with caution. In the parasitic plants example, we can say with confidence that parasitic plants tend to have higher rates of molecular evolution, but we cannot make a clear statement about the causal mechanism, because there are multiple traits of interest that might vary consistently between parasitic plants and their non-parasitic relatives (such as parasitic behaviour, height, population size). In the rockfish example, we attempted to use confounding traits to our advantage by contrasting the findings to the situation in mammals which have different trait relationships. In the salt tolerant plants, we can say that the distribution is significantly different than we would expect under a defined model where evolving salt tolerance has no influence on a lineages diversification rate or trait transition rate, and that it is more compatible with what we would expect to see if salt tolerance lineages have a higher rate of extinction or trait loss.

8. Conclusions

Hypothesis testing in macroevolution presents both challenges and opportunities. Many discussions of research in "historical science" have focussed on the limitations of doing science on events or processes that cannot be directly observed. But research in macroevolutionary biology has a number of great advantages not available to manipulative experiments in microevolution, such as vast timescales and replication over many different kinds of lineages or environments. Experimental and comparative studies share many of the same challenges, such as untangling the factors that might covary with the "treatment" lines and detecting unintended consequences of "control" lines. Good experimental design and a sceptical eye for the results are essential in both. Most importantly, while all researchers might dream of conducting the killer test that comprehensively puts debate to rest, both experimental and comparative approaches to questions in evolutionary biology typically require researchers to circle around hypotheses, weighing evidence for and against, considering the plausibility of alternative explanations, inching toward understanding one test at a time.

Acknowledgements

References

Alegret, L., Thomas, E., & Lohmann, K. C. (2012). End-cretaceous marine mass extinction not caused by productivity collapse. *Proceedings of the National Academy of Sciences United States of America, 109*, 728-732.

Alvarez, L. W., Alvarez, W., Asaro, F., & Michel, H. V. (1980). Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science, 208*, 1095-1108.

Archibald, J. D. (2014). What the dinosaur record says about extinction scenarios. *Geological Society of America Special Papers, 505*, 213-224.

Archibald, J. D., Clemens, W. A., Padian, K., Rowe, T., Macleod, N., Barrett, P. M., … Sahni, A. (2010). Cretaceous extinctions: Multiple causes. *Science, 328*, 973.

Ashraf, M., & Foolad, M. R. (2012). Crop breeding for salt tolerance in the era of molecular markers and marker-assisted selection. *Plant Breeding, 132*, 1-20.

Austad, S. N. (2012). Ageing: Mixed results for dieting monkeys. *Nature, 489*, 210-211.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., … Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature, 461*, 1243-1247.

Belcher, C. M., Collinson, M. E., Sweet, A. R., Hildebrand, A. R., & Scott, A. C. (2003). Fireball passes and nothing burns—The role of thermal radiation in the Cretaceous-Tertiary event: Evidence from the charcoal record of North America. *Geology, 31*, 1061-1064.

Bellot, S., & Renner, S. S. (2014). Exploring new dating approaches for parasites: The worldwide Apodanthaceae (Cucurbitales) as an example. *Molecular Phylogenetics and Evolution, 80*, 1-10.

Bennett, T. H., Flowers, T. J., & Bromham, L. (2013). Repeated evolution of salt-tolerance in grasses. *Biology Letters, 9*, 20130029.

Bjorkholm, B., Sjolund, M., Falk, P. G., Berg, O. G., Engstrand, L., & Andersson, D. I. (2001). Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proceedings of the national Academy of Sciences United States of America, 98*, 14607-14612.

Bromham, L. (2008). *Reading the Story in DNA: A beginner's guide to molecular evolution*. Oxford University Press.

Bromham, L. (2009). Why do species vary in their rate of molecular evolution? *Biology Letters, 5*, 401-404.

Bromham, L. (2011). The small picture approach to the big picture: Using DNA sequences to investigate the diversification of animal body plans. In K. Sterelny, & B. Calcott (Eds.), *The major transitions in evolution revisited*. MIT press.

Bromham, L. (2015). Macroevolutionary patterns of salt tolerance in angiosperms. *Annals of Botany, 115*, 333-341.

Bromham, L., & Bennett, T. H. (2014). Salt tolerance evolves more frequently in C4 grass lineages. *Journal of Evolutionary Biology, 27*, 653-659.

Bromham, L., Hua, X., & Cardillo, M. (2015). Detecting macroevolutionary self-destruction from phylogenetic patterns. *Systematic Biology*. in press.

Bromham, L., Hua, X., Lanfear, R., & Cowman, P. (2015). Exploring the relationships between mutation rates, life history, genome size, environment and species richness in flowering plants. *American Naturalist, 185*, 507.

Bromham, L., Rambaut, A., & Harvey, P. H. (1996). Determinants of rate variation in mammalian DNA sequence evolution. *Journal of Molecular Evolution, 43*, 610-621.

de Bruin, J.-P., Gosden, R. G., Finch, C. E., & Leaman, B. M. (2004). Ovarian aging in two species of long-lived rockfish, *Sebastes aleutianus* and *S. alutus*. *Biology of Reproduction, 71*, 1036-1042.

Brusatte, S. L., Butler, R. J., Prieto-Marquez, A., & Norell, M. A. (2012). Dinosaur morphological diversity and the end-Cretaceous extinction. *Nature Communications, 3*, 804.

Buckling, A., Macleacn, R. C., Brockhurst, M. A., & Colegrave, N. (2009). The beagle in a bottle. *Nature, 457*, 824-829.

Budd, G. E. (2013). At the origin of animals: The revolutionary cambrian fossil record. *Current Genomics, 14*, 344-354.

Buffetaut, E. (2004). Polar dinosaurs and the question of dinosaur extinction: A brief review. *Palaeogeography, Palaeoclimatology, Palaeoecology, 214*, 225-231.

Butterfield, N. J. (2007). Macroevolution and macroecology through deep time. *Palaeontology, 50*, 41-55.

Butterfield, N. J. (2011). Animals and the invention of the Phanerozoic Earth system. *Trends in Ecology & Evolution, 26*(2), 81-87.

Cardillo, M., Mace, G. M., Gittleman, J. L., & Purvis, A. (2006). Latent extinction risk and the future battlegrounds of mammal conservation. *Proceedings of the national Academy of Sciences United States of America, 103*, 4157-4161.

Carroll, R. C. (2000). Towards a new evolutionary synthesis. *Trends in Ecology & Evolution, 15*, 27-32.

Chao, L., & Cox, E. C. (1983). Competition between high and low mutating strains of *Escherichia coli*. *Evolution, 37*, 125-134.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics, 10*, 195-205.

Cheeseman, J. M. (2014). The evolution of halophytes, glycophytes and crops, and its implications for food security under saline conditions. *New Phytologist, 206*, 557-570.

Cleland, C. E. (2002). Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science, 69*, 447-451.

Cleland, C. E. (2013). Common cause explanation and the search for a smoking gun. *Geological Society of America Special Papers, 502*, 1-9.

Colman, R. J., Anderson, R. M., Johnson, S. C., Kastman, E. K., Kosmatka, K. J., Beasley, T. M., … Weindruch, R. (2009). Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science, 325*, 201-204.

Courtillot, V., & Fluteau, F. (2010). Cretaceous extinctions: The volcanic hypothesis. *Science, 328*, 973-974.

Coyne, J. A. (2006). Comment on "Gene regulatory networks and the evolution of animal body plans". *Science, 313*, 761.

Currie, A. (2014). Narratives, mechanisms and progress in historical science. *Synthese, 191*, 1163-1183.

Currie, A. (2015a). Hot blooded gluttons: Dependency, coherency & method in the historical sciences. *British Journal for the Philosophy of Science* (in press)

Currie, A. (2015b). Marsupial lions and methodological omnivory: Function, success and reconstruction in paleobiology. *Biology & Philosophy*, 1-23.

Edwards, E. J., & Donoghue, M. J. (2013). Is it easy to move and easy to evolve? Evolutionary accessibility and adaptation. *Journal of Experimental Botany, 64*, 4047-4052.

Erwin, D. H. (2000). Macroevolution is more than repeated rounds of microevolution. *Evolution & Development, 2*, 78-84.

Erwin, D. H. (2011). Evolutionary uniformitarianism. *Developmental Biology, 357*, 27-34.

Fastovsky, D. E., & Sheehan, P. M. (2005). Comment and reply: The extinction of the dinosaurs in North America. *GSA Today, 18*, 4-10.

Flowers, T. J., Galal, H. K., & Bromham, L. (2010). Evolution of halophytes: Multiple origins of salt tolerance in land plants. *Functional Plant Biology, 37*, 604-612.

Fortey, R. A., Jackson, J., & Strugnell, J. (2003). Phylogenetic "fuses" and evolutionary explosions: Conflicting evidence and critical tests. In P. Donoghue, & M. Smith (Eds.), *Telling the evolutionary time: Molecular clocks and the fossil record* (p. 42). Boca Raton: CRC Press.

Friedman, M., & Sallan, L. C. (2012). Five hundred million years of extinction and recovery: A phanerozoic survey of large-scale diversity patterns in fishes. *Palaeontology, 55*, 707-742.

Galtier, N., Blier, P. U., & Nabholz, B. (2009). Inverse relationship between longevity and evolutionary rate of mitochondrial proteins in mammals and birds. *Mitochondrion, 9*, 51-57.

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., & Taddei, F. (2001). Costs and benefits of high mutation rates: Adaptive evolution of bacteria in the mouse gut. *Science, 291*, 2606-2608.

Giribet, G. (2002). Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion? *Molecular Phylogenetics and Evolution, 24*(3), 345-357.

Goldberg, E. E., Roy, K., Lande, R., & Jablonski, D. (2005). Diversity, endemism, and age distributions in macroevolutionary sources and sinks. *The American Naturalist, 165*, 623-633.

Gould, S. J. (2002). *The structure of evolutionary theory*. Harvard University Press.

Haraguchi, Y., & Sasaki, A. (1996). Host-parasite arms-race in mutation modifications — Indefinate escalation despite a heavy load? *Journal of Theoretical Biology, 183*, 121-137.

Ho, S. Y., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., & Cooper, A. (2011). Time-dependent rates of molecular evolution. *Molecular Ecology, 20*, 3087-3101.

Holland, P. (2015). Did homeobox gene duplications contribute to the Cambrian explosion? *Zoological Letters, 1*(1), 1.

Jeffares, B. (2008). Testing times: Regularities in the historical sciences. *Studies in History and Philosophy of Science Part C, 39*, 469-475.

Knoll, A. H., & Carroll, S. B. (1999). Early animal evolution: Emerging views from comparative biology. *Science, 284*, 2129-2137.

Lanfear, R., Ho, S. Y. W., Davies, T. J., Moles, A. T., Aarssen, L., Swenson, N. G., … Allen, A. P. (2013). Taller plants have lower rates of molecular evolution: The rate of mitosis hypothesis. *Nature Communications, 4*, 1879.

Lanfear, R., Thomas, J. A., Welch, J. J., & Bromham, L. (2007). Metabolic rate does not calibrate the molecular clock. *Proceedings of the national Academy of Sciences United States of America, 104*, 15388-15393.

Lemaire, B., Huysmans, S., Smets, E., & Merckx, V. (2011). Rate accelerations in nuclear 18S rDNA of mycoheterotrophic and parasitic angiosperms. *Journal Plant Research, 124*, 561-576.

Levinton, J. S. (2001). *Genetics, palaeontology and macroevolution* (2nd ed.). Cambridge: Cambridge University Press.

van Loon, A. (2012). Time and space in stratigraphy: The '3-m gap' story. *Proceedings of the Geologists' Association, 123*, 394-396.

Love, M. S., Yoklavich, M., & Thorsteinson, L. K. (2002). *The rockfishes of the northeast Pacific*. Univ of California Press.

Lyson, T. R., Bercovici, A., Chester, S. G., Sargis, E. J., Pearson, D., & Joyce, W. G. (2011). Dinosaur extinction: Closing the '3 m gap'. *Biology Letters, 7*, 925-928.

MacLeod, N., Rawson, P., Forey, P., Banner, F., Boudagher-Fadel, M., Bown, P., … Evans, S. (1997). The Cretaceous-Tertiary biotic transition. *Journal of the Geological Society, 154*, 265-292.

Maddison, W. P., Midford, P. E., & Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology, 56*, 701-710.

Marshall, C. R. (2006). Explaining the Cambrian "explosion" of animals. *Annual Review of Earth and Planetary Sciences, 34*, 355-384.

Martin, A. P., & Palumbi, S. R. (1993). Body size, metabolic rate, generation time and the molecular clock. *Proceedings of the National Academy of Sciences United States of America, 90*, 4087-4091.

Mattison, J. A., Roth, G. S., Beasley, T. M., Tilmont, E. M., Handy, A. M., Herbert, R. L., … Bryant, M. (2012). Impact of caloric restriction on health and survival in rhesus monkeys from the NIA study. *Nature, 489*, 318-321.

Maynard Smith, J. (1978). *The evolution of sex*. Cambridge University Press.

Maynard Smith, J. (1990). *Taking a chance on evolution*. New York Review of Books. June 14 1990.

Miller, S. (2014). The public impact of impacts: How the media play in the mass extinction debates. *Geological Society of America Special Papers, 505*, 505-524.

Mooers, A., Gascuel, O., Stadler, T., Li, H., & Steel, M. (2012). Branch lengths on birth-death trees and the expected loss of phylogenetic diversity. *Systematic Biology, 61*, 195-203.

Moray, C., Hua, X., & Bromham, L. (2015). Salt tolerance is evolutionarily labile in a diverse set of angiosperm families. *Journal of Evolutionary Biology, 15*, 90.

Morgan, M. S. (2013). Nature's experiments and natural experiments in the social sciences. *Philosophy of the Social Sciences, 43*, 341-357.

Morgan, J., Artemieva, N., & Goldin, T. (2013). Revisiting wildfires at the K-Pg boundary. *Journal of Geophysical Research: Biogeosciences, 118*, 1508-1520.

Munns, R., & Gilliham, M. (2015). Salinity tolerance of crops — What is the cost? *New Phytologist*. http://dx.doi.org/10.1111/nph.13519.

Nee, S. (2006). Birth-death models in macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, 1-17.

Nee, S., May, R. M., & Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences, 344*, 305-311.

Nickrent, D. L., & Starr, E. M. (1994). High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants. *Journal of Molecular Evolution., 39*, 62.

Okasha, S. (2011). Experiment, observation and the confirmation of laws. *Analysis, 71*, 222-232.

Pal, C., Maciá, M. D., Oliver, A., Schachar, I., & Buckling, A. (2007). Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature, 450*, 1079-1081.

dePamphilis, C. W., Young, N. D., & Wolfe, A. D. (1997). Evolution of plastid gene rps2 in a lineage of hemiparasitic and holoparasitic plants: Many losses of photosynthesis and complex patterns of ratevariation. *Proceedings of the national Academy of Sciences of the United States of America, 94*, 7367-7372.

Peterson, K. J., Dietrich, M. R., & McPeek, M. A. (2009). MicroRNAs and metazoan macroevolution: Insights into canalization, complexity, and the Cambrian explosion. *BioEssays, 31*, 736-747.

Purvis, A. (1996). Using interspecies phylogenies to test macroevolutionary hypotheses. In P. H. Harvey, A. J. Brown, J. Maynard Smith, & S. Nee (Eds.), *New uses for new phylogenies* (pp. 153-168). Oxford: Oxford University Press.

Pybus, O., Rambaut, A., Holmes, E., & Harvey, P. (2002). New inferences from tree shape: Numbers of missing taxa and population growth rates. *Systematic Biology, 51*, 881-888.

Rabosky, D. L. (2009). Ecological limits and diversification rate: Alternative paradigms to explain the variation in species richness among clades and regions. *Ecology Letters, 12*, 735-743.

Raup, D. M., Gould, S. J., Schopf, T. J., & Simberloff, D. S. (1973). Stochastic models of phylogeny and the evolution of diversity. *The Journal of Geology*, 525-542.

Robertson, D. S., McKenna, M. C., Toon, O. B., Hope, S., & Lillegraven, J. A. (2004). Survival in the first hours of the Cenozoic. *Geological Society of America Bulletin, 116*, 760-768.

Sage, R. F. (2004). The evolution of C4 photosynthesis. *New Phytologist, 161*, 341-370.

Schopf, J. W. (2001). Solution to Darwins dilemma: Discovery of the missing precambrian record of life. In J. Chela-Flores, T. Owen, & F. Raulin (Eds.), *First steps in the origin of life in the universe*. Netherlands: Springer (pp. 23—23)

Schulte, P., Alegret, L., Arenillas, I., Arz, J. A., Barton, P. J., Bown, P. R., … Willumsen, P. S. (2010). The Chicxulub Asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science, 327*, 1214-1218.

Schwander, T., & Crespi, B. J. (2008). Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. *Molecular Ecology, 18*, 28-42.

Sinclair, D. A. (2005). Toward a unified theory of caloric restriction and longevity regulation. *Mechanisms of Ageing and Development, 126*, 987-1002.

Spicer, R. A., & Collinson, M. E. (2014). Plants and floral change at the Cretaceous-Paleogene boundary: Three decades on. *Geological Society of America Special Papers, 505*.

Stanford, P. K. (2011). Damn the consequences: Projective evidence and the heterogeneity of scientific confirmation. *Philosophy of Science, 78*, 887-899.

Sterelny, K. (2007). *Dawkins vs Gould: Survival of the fittest*. Icon Books.

Turner, D. (2011). *Paleontology: A philosophical introduction*. Cambridge University Press.

Turner, D. D. (2014). Philosophical issues in recent paleontology. *Philosophy Compass, 9*, 494-505.

Tyrrell, T., Merico, A., & Armstrong McKay, D. I. (2015). Severity of ocean acidification following the end-Cretaceous asteroid impact. *Proceedings of the National Academy of Sciences United States of America, 112*, 6556-6561.

Vajda, V., Raine, J. I., & Hollis, C. J. (2001). Indication of global deforestation at the Cretaceous-Tertiary boundary by New Zealand fern spike. *Science, 294*, 1700-1702.

Valentine, J. W. (2004). *On the origin of phyla*. Chicago: University of Chicago Press.

de Varigny, H. (1892). *Experimental evolution*. London: Macmillan and Co.

Varricchio, D. J., Martin, A. J., & Katsura, Y. (2007). First trace and body fossil evidence of a burrowing, denning dinosaur. *Proceedings of the Royal Society B — Biological Sciences, 274*, 1361-1368.

Welch, J. J., Bininda-Emonds, O. R. P., & Bromham, L. (2008). Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evolutionary Biology, 8*, 53.

White, R. V., & Saunders, A. D. (2005). Volcanism, impact and mass extinctions: Incredible or credible coincidences? *Lithos, 79*, 299-316.

Williams, G. C. (1975). *Sex and evolution*. Princeton Univesrit Pres.

Williams, G. C. (1992). *Natural selection: Domains, levels and challenges*. Oxford: Oxford University Press.

Woodward, J. (2013). Causation and manipulability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/win2013/entries/causation-mani/%3E

Woolfit, M. (2009). Effective population size and the rate and pattern of nucleotide substitutions. *Biology Letters, 5*, 417-420.