Zootaxa 3435: 40–67 (2012)
www.mapress.com/zootaxa/

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

Article

The limits of understanding in biological systematics

KIRK FITZHUGH

Research & Collections Branch, Natural History Museum of Los Angeles County, 900 Exposition Boulevard, Los Angeles, California 90007, USA. E-mail: kfitzhug@nhm.org

Abstract

Ernst Mayr's (1961, *Science* 131: 1501–1506) distinction between proximate and ultimate causation in biology is examined with regard to the acquisition of understanding in biological systematics. Rather than a two-part distinction, understanding in systematics is characterized by relations between three explanatory components: descriptive (observation statements)—proximate (ontogenetic hypotheses)—ultimate (e.g. specific and phylogenetic hypotheses). Initial inferential actions in each component involve reasoning to explanatory hypotheses via abductive inference, providing preliminary understanding. Testing hypotheses, to critically assess understanding, is varied. Descriptive- and proximate-level hypotheses are routinely tested, but ultimate hypotheses present inherent difficulties that impose severe limits, contrary to what is usually claimed in the systematics literature. The problem is compounded by imprecise considerations of 'evidence' and 'support.' For instance, in most cases, the 'evidence' offering 'support' for phylogenetic hypotheses, as cladograms, is nothing more than the abductive evidence (premises) used to infer those hypotheses, i.e. character data and associated phylogenetic-based theories. By definition, such evidence only offers initial, trivial understanding, whereas the pertinent evidence sought in the sciences is test evidence, which cannot be supplanted by character data. The pursuit of ultimate understanding by way of spurious procedures such as contrived testing, Bremer support, and resampling methods are discussed with regard to phylogenetic hypotheses.

Key words: biological systematics, causal explanation, cladistics, falsification, inference, testing

> *Every hypothesis should be put to the test by forcing it to make verifiable predictions. A hypothesis on which no verifiable predictions can be based should never be accepted, except with some mark attached to it to show that it is regarded as a mere convenient vehicle of thought—a mere matter of form.*
> Peirce (1935: 5.599)

> *Systematics is on a dangerous path towards irrelevancy to the remainder of biology because meaningful dialogue or assessment is no longer attempted, and is essentially impossible.*
> Mooi and Gill (2010: 27)

Introduction

The subject of causation is probably the most fundamental consideration in all the sciences, as it is the continual desire of scientists to acquire understanding of the phenomena we encounter (Hempel 1965; Rescher 1970; Popper 1983, 1992; Salmon 1984a; Van Fraassen 1990; Strahler 1992; Mahner & Bunge 1997; Hausman 1998; de Regt *et al*. 2009). Such understanding entails the interplay between our activities of explanation and prediction in conjunction with available hypotheses and theories. For instance, de Regt and Dieks (2005: 150, emphasis original) define the 'criterion for understanding phenomena' as, "*A phenomenon P can be understood if a theory T of P exists that is intelligible (and meets the usual logical, methodological and empirical requirements)*." Regarding biology specifically, Leonelli (2009: 197, emphasis original) characterizes understanding as "*the cognitive achievement realizable by scientists through their ability to coordinate theoretical and embodied knowledge that apply to a specific phenomenon*." A notable exegesis on causation and understanding in biology was provided by

Accepted by M. R. de Carvalho: 13 Jun. 2012; published: 23 Aug. 2012

Mayr [1961: 1503, see also 1982, 1993, 1994; a similar perspective was independently developed by Tinbergen (1963) with respect to ethology], in which he distinguished proximate and ultimate causation:

> ...proximate causes govern the responses of the individual (and his organs) to immediate factors of the environment while ultimate causes are responsible for the evolution of the particular DNA code of information with which every individual of every species is endowed.

Beatty (1994: 334) summarized Mayr's distinction as follows:

> The proximate causes of an organism's traits occur within the lifetime of the organism. They involve the expression of the information contained in the organism's genetic material, as mediated by the environment. The ultimate causes occur prior to the lifetime of the organism, within the evolutionary history of the organism's species.

In his recent treatment of the subject, Ariew (2003) provided some crucial clarifications. In rightly removing Mayr's emphasis on DNA and information, Ariew (2003: 555) characterized proximate causation more generally as "the causal capacities of structural elements" that are part of the life history of an organism. But in contradistinction to ultimate causation, Ariew suggested that what is really being referred to is 'evolutionary explanation,' which subsumes not only natural selection, which was the only cause referred to by Mayr (1961), but other causes as well, such as mutation, recombination, and genetic drift. And while evolutionary explanations would serve to address causal questions regarding the properties of organisms, Ariew (2003: 558, 560) saw these as "statistical population-level" explanations: "Evolutionary explanations range over statistical attributes of a population..." The intent was to counter the argument that ultimate explanations can be reduced to a series of individual-level proximate explanations.

The purpose of the present paper will be to examine Mayr's (1961) proximate-ultimate distinction from the perspective of the acquisition of causal understanding in biological systematics. Systematics addresses causal questions that span proximate and ultimate explanatory realms, with the latter consisting of several pertinent classes of explanations. While there is a broad range of causes that can be regarded as evolutionary, I will not follow Ariew (2003) in grouping all ultimate causes under that term. Although ultimate explanations entail a variety of hypotheses, some of which are presented below, two of the most prominent in systematics are specific (i.e. species taxa; Fitzhugh 2005b, 2009) and phylogenetic (cladograms, or supraspecific taxa; Fitzhugh 2008b). The issue to be addressed in this paper is the extent to which ultimate explanations in biological systematics not only lead to initial causal understanding but also result in further, critical assessments of that understanding, per the goal of scientific inquiry. In other words, in what capacity do such hypotheses provide understanding, and how pervasive is empirical support for or against that understanding as a consequence of testing?[1] What I will point out is that ultimate explanations in biological systematics are typically marginal vehicles for understanding. They are 'explanation sketches' as characterized by Hempel (1965: 423–424). These sketches, either as species 'descriptions' or graphic representations of phylogenetic hypotheses referred to as cladograms, are rarely ever filled out as full explanations amenable to the acts of testing so often claimed in the systematics literature (e.g. Wiley 1975; Gaffney 1979; Eldredge & Cracraft 1980; Rieppel 1988; Faith & Cranston 1992; Kluge 1997a, 1997b, 1999, 2001; Grandcolas *et al*. 1997; Siddall & Kluge 1997; Wenzel 1997; Schuh 2000; de Queiroz & Poe 2001, 2003; Farris *et al*. 2001; Faith & Trueman 2001; Faith 2004, 2006; Wheeler 2004, 2010; Franz 2005; Helfenbein & DeSalle 2005; Egan 2006; Grant & Kluge 2008; Schuh & Brower 2009; Faith *et al*. 2011; Wiley & Lieberman

[1] To be clear from the start, my reference to hypothesis testing is only in reference to *explanatory* hypotheses, not *statistical* hypotheses. If the principle goal of scientific inquiry is to extend causal understanding, then there is the expectation that a historical science like systematics should strive for that goal. The pursuit of causal understanding begins with one's reactions to observations in the form of why-questions regarding observed states of affairs that are unexpected or surprising. We infer explanatory hypotheses that suggest possible past causal conditions, serving as answers to those questions. As such, cladograms *qua* topologies are not explanatory hypotheses. Rather, they are diagrams implying a variety of explanatory hypotheses regarding past causal events. Cladograms therefore are not statistical constructs. While there are a host of methods that are commonly implemented under the guise of 'testing' cladograms, e.g. bootstrap, jackknife, permutation and likelihood ratio tests, none of these are addressing the actual explanatory hypotheses to which hypothesis testing would be directed.

2011). Rather than promoting increased ultimate (evolutionary) understanding through the testing of specific and phylogenetic hypotheses, the tendency is for investigators to revert to the pursuit of enhancements of descriptive aspects (observation statements) regarding organisms or some proximate explanations. Two complicating issues, which have been addressed elsewhere, are the following: (1) the lack of emphasis on the causal questions that prompt the inferences that serve as answers (Fitzhugh 2006a–c), and (2) the tendency to confuse those inferences with the testing of resultant hypotheses (Fitzhugh 2006a, 2008a, 2010a; e.g. Schmidt 2009). The consequence has been methodological mischaracterizations of those inferences and hypothesis testing under such headings as 'parsimony,' 'maximum likelihood,' and 'Bayesianism' (e.g. Siddall & Kluge 1997; Faith & Trueman 2001; Haber 2005; Schmidt 2009), among others.

Biological understanding: descriptive, proximate, ultimate

Mayr (1961) characterized biology as two separate fields, functional and evolutionary, seeking proximate and ultimate causes, respectively. But Mayr (1961: 1501) also recognized a third field, "purely descriptive structural biology." In terms of either proximate or ultimate causal hypotheses, these would follow from one's description(s) of the perceived effects in need of explanation, i.e. the properties of organisms. What is interesting is that while Mayr downplayed description in lieu of proximate and ultimate explanations, the inferences, and subsequent communication of our observation statements do in fact serve the purpose of explanation as well. Observation statements are not theory neutral constructs (Hanson 1958; Popper 1992; Godfrey-Smith 2003). We receive sense data as a consequence of interactions with objects. To those data we apply any variety of theories to infer concepts and statements that accord us some degree of understanding of those perceptions, providing bases for subsequent actions. For instance, my sense perceptions of a group of objects might lead me to conclude that "This is a glass of water." The conclusion is inferred by that class of non-deductive reasoning known as abduction, wherein effects are conjoined with one or more theories to infer a tentative cause (Peirce 1878, 1931, 1932, 1933a, 1933b, 1934, 1935, 1958a, 1958b; Hanson 1958; Achinstein 1970; Fann 1970; Reilly 1970; Curd 1980; Nickles 1980; Thagard 1988; Josephson & Josephson 1994; Hacking 2001; Magnani 2001; Psillos 2002, 2007; Godfrey-Smith 2003; Norton 2003; Walton 2004; Aliseda 2006; Fitzhugh 2005a, 2005b, 2006a, 2006b, 2008a–c, 2009, 2010a; Schurz 2008). Abductive inference can be schematized as:

[1]

• auxiliary theory(ies)
• theory(ies) relevant to the effects perceived
• perceived effects
══════════════════
• explanatory hypothesis, *H*.

While uttering "This is a glass of water" is an observation statement, per my current understanding of theories such as glass and water, the statement also serves the purpose of explaining why my perceptions are the case—it is the existence of the objects referred to as glass and water that are the causes of my sense perceptions (Schurz 2008).

In considering the nature of understanding in biological systematics, it would be neglectful not to include descriptions as a class of causal understanding fundamental to proximate and ultimate understanding. If it is the case that the goal of science is to engage in a continual process of acquiring causal understanding by way of initial abductive inferences and subsequent testing of hypotheses and theories, then observation statements are the fundamental starting points that lead to the pursuit of proximate and ultimate understanding (cf. Mayr 1982). Like observation statements, which provide descriptive understanding, proximate and ultimate understanding also originate as products of abductive reasoning, differing only in the respective sets of theories employed in each. Extensive treatments of the nature of abductive inference in systematics can be found in Fitzhugh (2005a, 2006a, 2006b, 2008a, 2008b, 2009, 2010a). Examples of specific and phylogenetic inferences will be provided later in relation to testing such hypotheses.

With explanatory hypotheses providing initial understanding of sets of objects and/or events, serving as answers to specifiable causal questions, subsequent testing would serve to assess and potentially expand or revise that understanding through confirming evidential support, or hypothesis revision or replacement as results of disconfirming evidence. In its simplest form, testing of explanatory hypotheses would first involve deductive inferences to predictions of potential test evidence in the form of consequences that should be encountered if the cause-effect relations stated by a theory as well as initial conditions presented in the hypothesis are the case (Schurz 2008; Fitzhugh 2010a):

[2]

- auxiliary theory(ies)
- theory(ies) relevant to the effects perceived
- specific causal conditions presented in explanatory hypothesis, *H* (*ex* **[1]**)
- proposed conditions needed to carry out test

- perceived effects (originally prompting *H*; cf. **[1]**)
- *predicted test evidence, i.e. effects related as closely as possible with the specific causal conditions of the hypothesis*.

Ideally, this potential test evidence should consist of effects with the lowest probability of occurrence if the causal conditions stated in the hypothesis did not transpire (Peirce 1958a; Mayo 1996; Achinstein 2001; Fitzhugh 2010a)[2]—what Cleland (2001, 2002, 2011a, 2011b) referred to as ‘smoking gun’ evidence. And by the very nature of the deduction from the specific causal events stated in the hypothesis, test evidence would be a class of effects independent of that upon which the hypothesis was originally inferred (Popper 1992; see also Tucker 2011). Pursuant to deriving predictions, testing (inductive *sensu stricto*) would be performed to determine whether or not observed test consequences support the hypothesis:

[3]

- auxiliary theory(ies)
- theory(ies) relevant to the effects perceived
- actual test conditions
- actual confirming/disconfirming evidence (observations of *predicted test evidence* in **[2]**/alternate observations)

- *H* is confirmed/disconfirmed.

As will be noted later, the systematics literature too often fails to provide cogent distinctions between our observations of organisms we wish to explain by way of past evolutionary events (*qua* abduction—**[1]**) and the potential or actual evidence needed (via either de- or induction—**[2]**, **[3]**) to engage in the empirical evaluations of those hypotheses. Failure to recognize the abductive nature of systematics inferences that serve as answers to either implicit or explicit causal questions, and the proper mechanics of testing, has led to erroneous efforts that conflate hypothesis inference with testing as well as other spurious notions of assessing hypothesis support (e.g. Schmidt 2009).

With this interplay between the inferences of hypotheses and their being tested, Mayr’s three classes of understanding can be brought to the context of systematics and summarized as follows:

- **Descriptive understanding.** Critical assessment over time of answers to questions regarding properties (effects) instantiated by organisms; e.g. “What accounts for my sense perceptions of this organism in my visual field as opposed to some other object?” The question entails observations for the purpose of adding to one’s repertoire of properties that characterize those objects, with observation statements serving as perceptual hypotheses answering the questions. There is the subsequent process of testing those hypotheses by way of investigations into component parts that make up the more inclusive properties of the objects.
- **Proximate causal understanding.** Critical assessment over time of answers to questions regarding properties (effects) instantiated by organisms at a moment in their life history as opposed to some other time.

[2] Reference here to low probability of test evidence is not equivalent to the ‘improbable evidence’ of character data in relation to alternative hypotheses as suggested by Faith (2004, 2006), Faith and Cranston (1991, 1992), Faith and Trueman (1998, 2001), and Faith et al. (2011). As will be described later, no amount of character data used to infer cladograms can serve the added purpose of then testing those hypotheses.

Answers are by way of proximate causes, i.e. explanations of manifestations of characters by way of intrinsic and extrinsic causes occurring during the lifetime of an organism (Beatty 1994; Ariew 2003); e.g. ontogenetic hypothesis: “Why does this individual have property *Y* at this stage of its life history as opposed to property *X* that occurs at another stage?” An ontogenetic hypothesis regarding manifestation of *Y* is provided. Continued evaluation of this proximate understanding would be by the consequences of testing the hypothesis that served as the initial answer.

- **Ultimate causal understanding.** Critical assessment over time of answers to questions regarding properties (effects) instantiated by groups of organisms. Answers are by way of intrinsic and extrinsic causal processes occurring among multi-generational groups of individuals, resulting in differential expressions of properties; e.g. phylogenetic hypothesis: “Why do individuals to which species hypotheses *b-us* and *c-us* apply have character *Y* as opposed to character *X* as observed among members of *a-us*, *x-us*, etc.?” A phylogenetic hypothesis regarding character *Y* origin and fixation in an ancestral population and subsequent population splitting (‘speciation’) can be provided. The continued enhancement of this ultimate understanding would be by the consequences of testing the various causal components in the hypothesis that served as the initial answer.

Initial understanding in biological systematics

Segregating understanding into three distinct classes when speaking of biological systematics provides a way to indicate the internested relations that exist in systematics research. In other words, there tends to be a three-tier system of inquiry: (1) descriptive understanding, in the form of observation statements, is a consequence of our desire to attain understanding of sense data by way of the existence of objects with particular properties, and that level of understanding naturally leads to (2) sets of questions answerable by proximate causes, given that we routinely observe individuals at different points in their life history, and (3) with observations among groups of organisms there are additional questions to which a variety of ultimate causes can be applied.

Hennig (1966: fig. 6) stressed the importance of recognizing the nuances that exist among various biological systematics hypotheses (Fig. 1). He noted that systematists deal with at least seven classes of hypotheses within a spectrum of understanding transcending descriptive, proximate, and ultimate contexts. These hypotheses are distributed accordingly (Fig. 2):

descriptive—individual (semaphoront)
proximate —ontogenetic
ultimate —tokogenetic, cyclomorphic, sexual dimorphic, polymorphic, specific (species), and phylogenetic.

As indicated in the previous section, all of these hypotheses are products of abductive reasoning, as responses to implicit or explicit questions regarding one’s perceptions of individuals (Fitzhugh 2005a, 2005b, 2006a–c, 2008a–c, 2009; Table 1).

Continued understanding in biological systematics

The relations between the classes of understanding presented above tend to be operationally internested from the perspective that our opportunities to test hypotheses that are (1) descriptive, (2) proximate, and (3) ultimate, respectively, tend to be increasingly more restricted and thus less frequent (Fitzhugh 2010a). Compare, for instance, the ability to (1′) test the descriptive hypothesis that I observe a lizard with three toes (I–III, as opposed to II–IV) in contrast to four toes (I–IV), to (2′) the proximate hypothesis explaining by way of ontogeny the presence of toes I–III in this adult, to (3′) an ultimate (phylogenetic) hypothesis explaining the presence of toes I–III in contrast to I–IV among individuals to which several different species hypotheses apply. Assessing the observation statement would minimally require (1″) observations of the expected skeletal components that comprise I–III (as opposed to II–IV), while testing the ontogenetic hypothesis would require (2″) the commitment of time and resources to observe causal relations among limbs and toes within individuals over some period of time during their life history. Testing the phylogenetic hypothesis would be the most challenging, as it would require (3″) access to test evidence not only regarding the specific causal events associated with origin and fixation of the three-toe condition in an ancestral population, but also evidence of the cause(s) that led to subsequent splitting(s) of the population(s), colloquially referred to as 'speciation.'

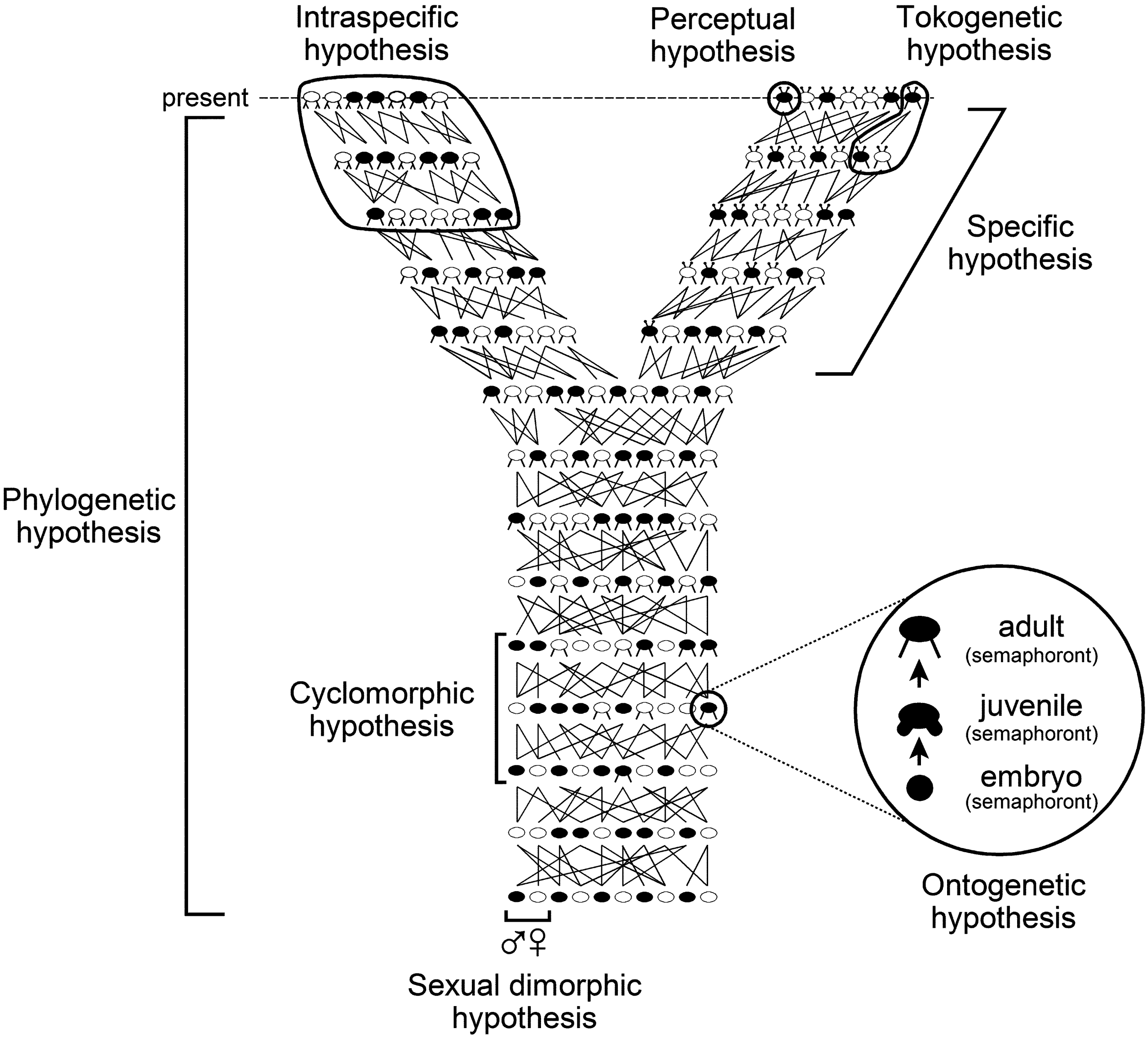


FIGURE 1. Hypotheses commonly encountered in biological systematics. Modified from Fitzhugh (2008b: fig. 1) based on Hennig's (1966) figure 6.

Assessing our causal understanding in biological systematics, whether descriptive, proximate, or ultimate, is determined by the extent to which hypothesis testing, *sensu* **[3]**, is accomplished. While the act of inferring an explanatory hypothesis provides initial understanding, in that the hypothesis serves as an answer to at least one specifiable question, it is a hallmark of the sciences that engaging in critically evaluating such hypotheses allows us to not only gauge the current status of understanding but to alter or revise that understanding over time. But as noted earlier, the ability to test biological systematics hypotheses becomes progressively more difficult as one proceeds from descriptive, to proximate, to ultimate explanations. A prominent limiting factor for the testing of higher-level ultimate hypotheses is the span of time between those effects that prompt inferences of hypotheses and the causal events themselves. The greater the span of time from a hypothesized cause(s) and observed effects the more likely will be the eradication of relevant evidence required to test those hypotheses (Cleland 2011a). The consequence of the inherent difficulty with testing ultimate explanations is that two of the most prominent classes of evolutionary hypotheses, specific and phylogenetic (cf. Fig. 1), are almost never legitimately tested, such that there is no actual enhancement of understanding via hypothesis revision or replacement on the basis of empirical assessments of causal relations. Rather, once ultimate hypotheses have been inferred, the tendency among systematists is to direct focus back to the inferences of additional or revised descriptive and proximate hypotheses, if at all, in a purported attempt to refine ultimate understanding *qua* testing (see below). For instance, Hennig (1966: 122, emphasis added) mistakenly considered this maneuver as a process of testing phylogenetic hypotheses:

> Thus the question of whether kinship relations based on a single character or a single presumed transformation series of characters correspond to the actual phylogenetic relationships of the species *is tested by means of other series of characters: by trying to bring the relationships indicated by the several series of characters into congruence. In the final analysis this is again the method of "checking, correcting, and rechecking"....*

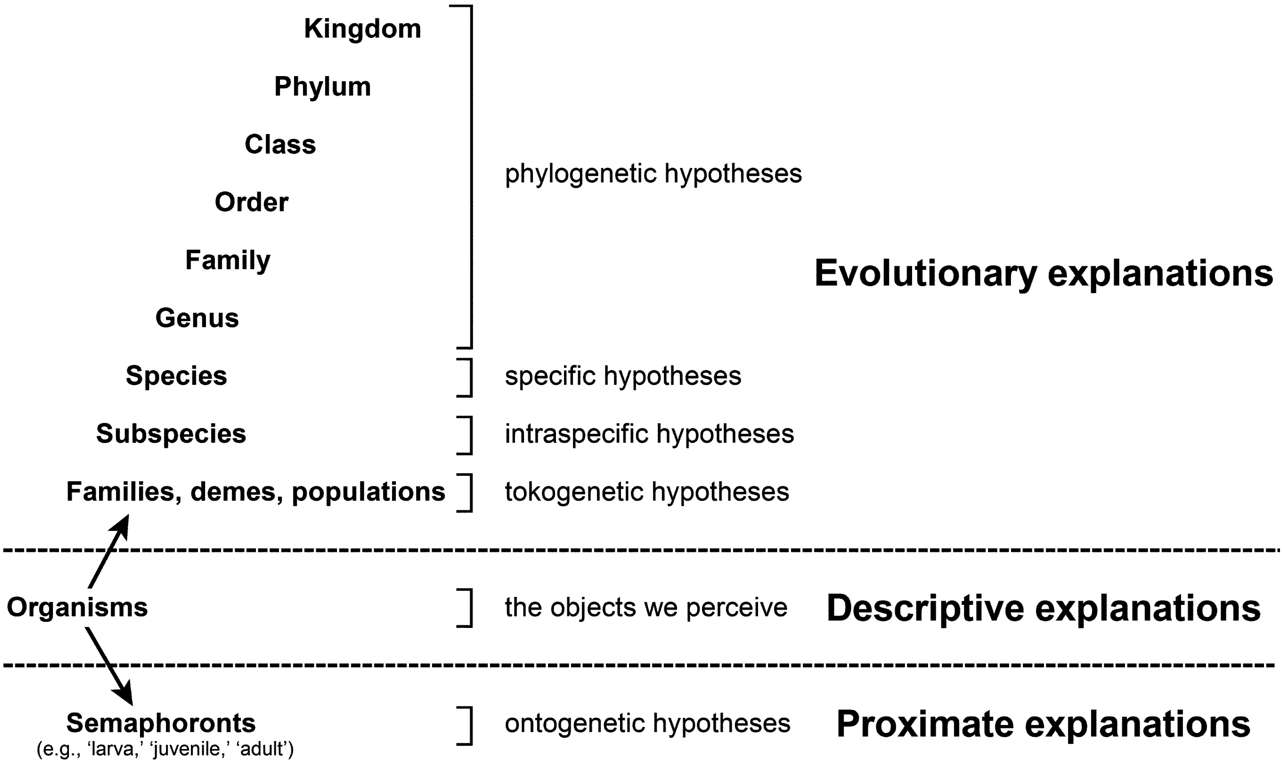


FIGURE 2. Relations between systematics hypotheses (cf. Fig. 1) utilized in descriptive, proximate, and ultimate explanatory contexts.

Likewise, the more characters involved in this process the better (Hennig 1966: 132, emphasis added): "For phylogenetic systematics this means that the *reliability of its results* increases with the number of individual characters that can be fitted into transformation series." Rindal and Brower (2011: 331; see also Brower 2006, 2010) go so far as to make the claim that character congruence on cladograms is the only means of phylogenetic assessment. Unfortunately, in the nearly 40 years since Hennig's statements, the emphasis on congruence or acquiring more character data under the guise of testing has grown in prominence, most notably through the peculiar attempts to ally this activity with Karl Popper's notions of corroboration or falsification (e.g. Wiley 1975; Gaffney 1979; Eldredge & Cracraft 1980; Rieppel 1988; Kluge 1997a, 1997b, 1999, 2001; Siddall & Kluge 1997; Farris *et al.* 2001; Faith & Trueman 2001; Faith 2004; de Queiroz & Poe 2001; Lee & Camens 2009; Wiens 2009; Faith *et al.* 2011). As will be outlined in the next section, the result has been the development of a cycle of at best minimizing, and at worst impeding causal understanding as a consequence of conflating the inferences of evolutionary hypotheses with their being tested.

TABLE 1. Comparisons of perceptual, ontogenetic, tokogenetic, 'intraspecific' specific, and phylogenetic hypotheses (modified from Fitzhugh 2008b). See Figure 1 for graphic representations of each hypothesis.

Causal questions:	Relations:	Represented by:
'Why do I have these sense perceptions?'	*Perceptual hypothesis* – An individual exists.	Observation statement
	Ontogenetic hypothesis –	Semaphoront names, e.g. 'embryo,' 'larva,' 'adult'
'Why does this individual have character *v*(1) at time t_2 in contrast to character *v*(0) at t_1?'	This individual has character *x*(1) at time t_2 because it is part of the ontogenetic trajectory.	
	Tokogenetic hypothesis –	Families, demes, populations, communities, etc.
'Why are these individuals observed at this location in contrast to some other location?'	Individuals are at this location because they are products of past tokogenetic events among other individuals.	
'Why does this individual, or individuals, have character *w*(1) in contrast to *w*(1) ?'	This individual, or individuals, has character *w*(1) because the genetic capacity to exhibit the character was passed on from their parent(s).	
	'*Intraspecific*' *hypothesis* –	Polymorphism
'Why do individuals to which species hypothesis *X-us x-us* refers have either character *x*(0) or *x*(1) in contrast to only character *x*(0) observed among other individuals?'	The reproductively isolated population is polymorphic because character *x*(1) originated in the population, such that observed individuals with *x*(0) and *x*(1) are products of past tokogenetic events among individuals with those characters.	
	Species hypothesis –	Species names
'Why do these individuals have character *y*(1) in contrast to character *y*(0) observed among individuals to which other species hypotheses have been applied?'	Individuals have character *y*(1) because it originated among individuals with character *y*(0), and *y*(1) eventually became fixed throughout the population, such that individuals observed in the present are products of past tokogenetic events involving individuals with that character.	
	Phylogenetic hypothesis –	Supraspecific names
(i) 'Why do these individuals, to which species hypotheses *A-us a-us* and *A-us b-us* refer, have character *z*(1) in contrast to character *z*(0)?' (applicable to gonochoristic or cross-fertilizing hermaphroditic organisms)	(i) Individuals have character *z*(1) because this character originated within a population with character *z*(0), and *z*(1) eventually became fixed throughout the population, and there was subsequent splitting of that population into two or more populations.	
(ii) 'Why do these individuals have character *z*(1) in contrast to character *z*(0)?' (applicable to obligate asexual, parthenogenetic, and self-fertilizing organisms)	(ii) Individuals have character *z*(1) because this character originated by some unspecified mechanism(s) during tokogeny, i.e. asexual, parthenogenetic, or self-fertilizing, from an earlier individual with *z*(0).	

The process of acquiring causal understanding in biological systematics

The previous two sections outlined what is required to proceed from sense data to three general classes of understanding via inferences of explanatory hypotheses. The last section ended with the observation that some authors in biological systematics have developed a protocol that falsely claims to move ultimate understanding forward by actions that are not valid test procedures. The nature of this confusion will be addressed in part in this section. The subsequent section will examine common approaches to evaluating phylogenetic hypotheses, either in the context of testing or stipulating evidential support.

Inferences of ultimate hypotheses. Consider the following common tactic. A systematist examines specimens, amassing a variety of 'morphological,' histological, reproductive, behavioral observations and/or nucleotide sequences. Select properties of observed individuals are described, leading (usually implicitly) to inferences of specific hypotheses, colloquially referred to as 'species descriptions' (*contra* Table 1; Nogueira *et al.* 2010; Fitzhugh 2008b, 2009, 2010b). More inclusive phylogenetic hypotheses, in the form of cladograms, are often included, with some 'clades' given formal (supraspecific) names. While typically unstated, the goals of these actions are to determine explanatory accounts of differentially shared characters among observed individuals, as answers to implicit causal questions (Table 1). Given the regularity with which such ultimate hypotheses as specific and phylogenetic are inferred, to what extent has initial understanding of organismal properties been achieved? While we can readily identify descriptive understanding of the objects of interest, in terms of the characters instantiated by observed individuals, what is garnered in terms of ultimate understanding with specific and phylogenetic hypotheses is usually nothing more than vague answers to questions regarding shared characters (Fitzhugh 2009). Species-level hypotheses are rarely referred to as providing explanatory accounts of particular characters. And when such hypotheses are specified as such, they are only in the imprecise sense that characters originated and became fixed in an ancestral population. Explicit details regarding the causal factors involved in character fixation are not presented.

Cladograms are equally vague causal accounts, at best implying that particular characters originated by some unspecified mechanism(s) and were subsequently fixed among members of an ancestral population by some unspecified mechanism(s), followed by at least one population splitting event by some unspecified mechanism(s) (Fig. 1). The meager explanatory standings of specific and phylogenetic hypotheses are consequences of the fact that neither type of inference is made on the basis of detailed theories regarding character origin, fixation, and/or population splitting (Fitzhugh 2006a, 2009). Overall, while we can identify answers to causal questions that provide initial descriptive, proximate, and ultimate understanding (Table 1, Figs. 1–2), it is the latter that is the least detailed in terms of conveying causal structure.

Testing and support issues. From the basic outline of systematics practice just presented, ranging from observation statements to abductive inferences of specific and phylogenetic hypotheses, there is the subsequent matter of judging the explanatory merits of these hypotheses as matters of both assessing and extending causal understanding. Especially with the advent of cladistics, the emphasis on testing or garnering evidential support has been almost exclusively directed toward phylogenetic as opposed to intraspecific or specific hypotheses, so it will be with this former class of ultimate causation that hypothesis evaluation will be examined here.

The subject of evidential support for phylogenetic hypotheses has received substantial attention, whether under the heading of hypothesis testing or techniques purported to measure support, e.g. Bremer support or various resampling protocols. But support in relation to cladograms has two quite different connotations.[3] In one sense it relates to abductively inferred hypotheses (**[1]**) and in another to subsequent testing by induction **([3])**. More generally, and regardless of mode of inference, support refers to relations between premises and conclusion(s) (Longino 1979; Salmon 1984b; Achinstein 2001). This means a distinction can be made between the support for *initial* understanding that is the product of abduction, as opposed to support for *assessing*, *expanding*, or *revising* that understanding as consequences of testing via induction (Hanson 1958; Norton 2003).

[3] There is a third connotation, sometimes applied by advocates who regard cladograms as ahistorical, non-causal (and thus non-explanatory) diagrams. For instance, Brower (2011: 447; see also Brower 2006, 2010; Turjak & Trontelj 2012) refers to cladogram support as "a measure of the relative quantity of evidence favoring a hypothesis of relationships [*sic*], not a measure of whether or not the hypothesis corresponds to the actual pattern of historical cladogenesis of the taxon in question…." Since such a view has no relation to the pursuit of causal understanding that is the goal of scientific inquiry, there is no need to give it consideration in the context of hypothesis support addressed in this paper.

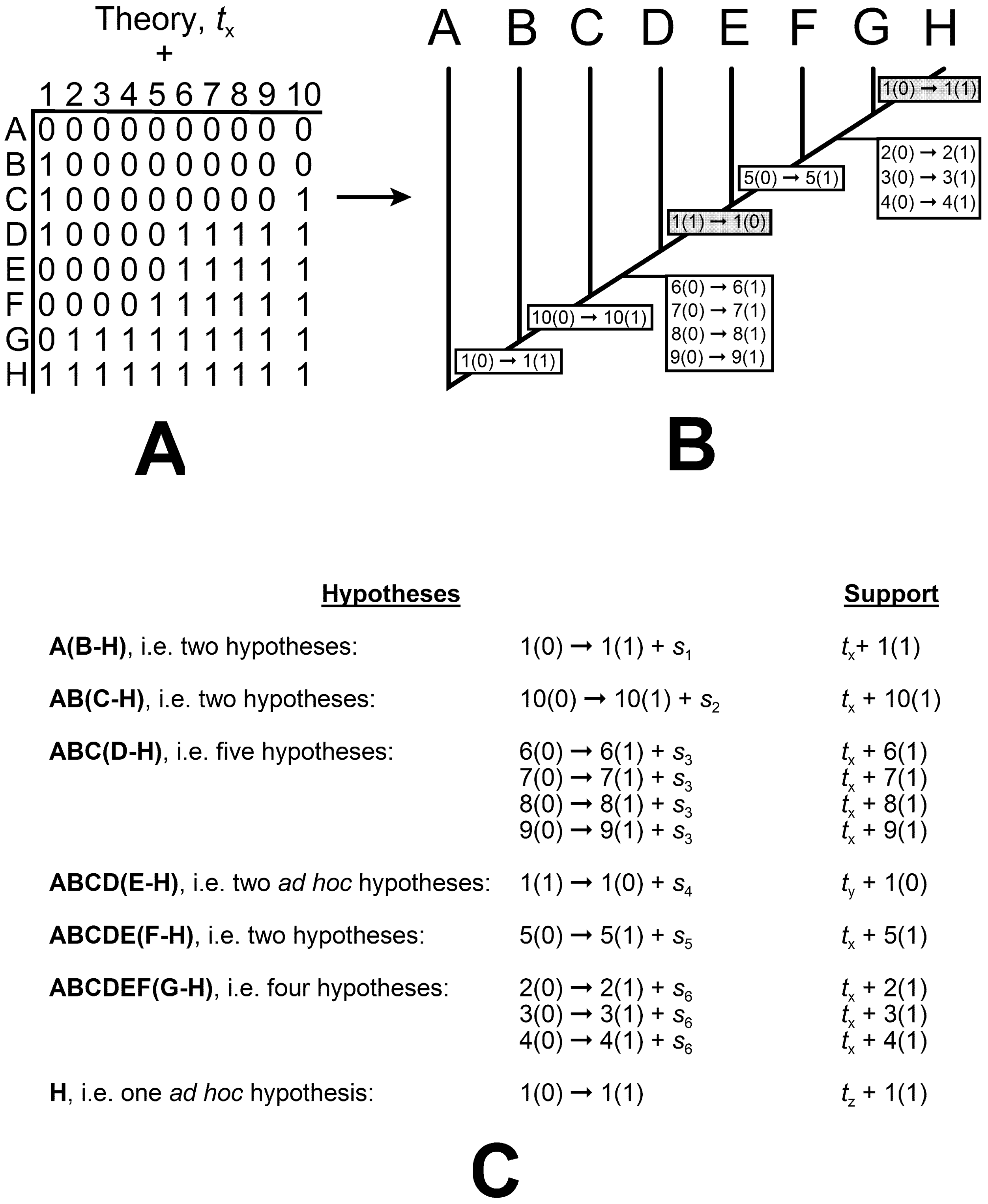
Referring to the schematic outline of abduction in **[1]**, the understanding afforded is simply a matter of conjoining as completely as possible observed effects in need of explanation with some theory(ies). Regardless of the causal depth stipulated by the theory, the initial understanding provided in the conclusion is no more than a tentative causal accounting. Coupled with the fact that such inferences should apply a given theory as fully as possible to effects, thus maximizing explanations of effects as instances of the conjoined theory, considerations of support for cladograms are largely trivial (cf. Schurz 2008). Indeed, applying a theory as fully as possible to observed effects has the default consequence of maximizing support, i.e. explaining those effects as completely as possible as instances of the theory. As the initial understanding conveyed by cladograms is quite meager, tallying support for the hypotheses they imply can be performed directly. Consider the example in Fig. 3. The premises comprise at a minimum relevant theories (and attendant background knowledge) and observed effects (Fig. 3A). In this instance a generic 'descent with modification/common ancestry' theory (actually several theories) would be one entailing novel character origin and fixation in ancestral populations, and subsequent population splittings (a fuller explication is given below in **[10]** in **Increasing causal understanding—testing as intended**):

[4] If character *x*(0) exists among individuals of a reproductively isolated, gonochoristic or cross-fertilizing hermaphroditic population and character *x*(1) originates by mechanisms *a, b, c… n*, and becomes fixed within the population by mechanisms *d, e, f… n* (=ancestral species hypothesis), followed by event(s) *g, h, i… n*, wherein the population is divided into two or more reproductively isolated populations, then individuals to which descendant species hypotheses refer would exhibit *x*(1).

While the theory in **[4]** is one of strict common cause, two common causes are referenced: proximate character origin/fixation and distal population splitting. Both classes of events are required given the diagrammatic representations offered by cladograms. Invoking **[4]** follows from the causal questions one explicitly or implicitly asks regarding shared characters among individuals to which two or more specific-level hypotheses apply (cf. Table 1; Fitzhugh 2006a, 2006b, 2008b, 2008c, 2010a). Indeed, these questions are codified in data matrices by way of the inclusion of outgroups (taxon 'A' in this example; Fitzhugh 2006c), and if the intent is to explain the fidelity of one's observations, then conjoining the theory in **[4]** with observations of shared characters will offer explanatory accounts that maintain that fidelity to the greatest extent possible (Fitzhugh 2006a). There is the alternative view subsumed under the phrases 'maximum likelihood' (Felsenstein 1981, 2004; Swofford *et al.* 1996; Huelsenbeck & Crandall 1997; Haber 2011) and 'Bayesian' (Huelsenbeck *et al.* 2001; Huelsenbeck & Ronquist 2001; Archibald *et al*. 2003; Ronquist *et al*. 2009) that asserts that common ancestry should be considered in conjunction with stochastic rates of character change and 'branch lengths.' The failure of likelihood and Bayesian approaches in the context of phylogenetic inference is that neither considers the relations between the causal questions represented in a data matrix and the abduction of explanatory hypotheses. The concept of likelihood is superfluous to abduction since inferred hypotheses automatically accord the highest probability on the character data being explained, as consequences of the conjunctions of theory(ies) with observed effects (cf. **[1]**; see example in **Increasing ultimate causal understanding—testing as intended**). The attendant argument from statistical consistency that has been used to promote 'maximum likelihood' methods is meaningless for abductive reasoning. As is the case with statistics, which pertains to induction *sensu stricto*, consistency is only relevant to hypothesis testing, under the view that such testing will 'in the long run' eventually lead to 'true' hypotheses. Consistency has no relevance for abduction, and by extension phylogenetic inference (Peirce 1901, 1932: 2.777, emphasis added):

> [Abduction] is the only kind of reasoning which supplies new ideas, the only kind which is, in this sense, synthetic. Induction is justified as a method which must in the long run lead up to the truth, and that, by gradual modification of the actual conclusion. *There is no such warrant for* [*abduction*]. The hypothesis which it problematically concludes is frequently utterly wrong itself, and even the method need not ever lead to the truth; for it may be that the features of the phenomena which it aims to explain have no rational explanation at all. Its only justification is that its method is the only way in which there can be any hope of attaining a rational explanation.

The Bayesian perspective is misdirected because the 'evidence' of interest in phylogenetic inference is not test evidence. The protracted emphasis on 'optimality criteria' in phylogenetic inference, especially parsimony versus likelihood, has needlessly detracted from the more salient issue of stipulating the appropriate theory relative to observations in need of being explained (Fitzhugh 2006a).



	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	0	0	0	1
D	1	0	0	0	0	1	1	1	1	1
E	0	0	0	0	0	1	1	1	1	1
F	0	0	0	0	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1
H	1	1	1	1	1	1	1	1	1	1

Hypotheses		Support
A(B-H), i.e. two hypotheses:	1(0) → 1(1) + s_1	t_x+ 1(1)
AB(C-H), i.e. two hypotheses:	10(0) → 10(1) + s_2	t_x + 10(1)
ABC(D-H), i.e. five hypotheses:	6(0) → 6(1) + s_3 7(0) → 7(1) + s_3 8(0) → 8(1) + s_3 9(0) → 9(1) + s_3	t_x + 6(1) t_x + 7(1) t_x + 8(1) t_x + 9(1)
ABCD(E-H), i.e. two *ad hoc* hypotheses:	1(1) → 1(0) + s_4	t_y + 1(0)
ABCDE(F-H), i.e. two hypotheses:	5(0) → 5(1) + s_5	t_x + 5(1)
ABCDEF(G-H), i.e. four hypotheses:	2(0) → 2(1) + s_6 3(0) → 3(1) + s_6 4(0) → 4(1) + s_6	t_x + 2(1) t_x + 3(1) t_x + 4(1)
H, i.e. one *ad hoc* hypothesis:	1(0) → 1(1)	t_z + 1(1)

FIGURE 3. Relations between the (A) premises and (B) conclusion of an abductive inference to phylogenetic hypotheses, and (C) the direct indication of the abductive support for those hypotheses. Note that such support refers not to 'branch support,' but rather the actual support for the two classes of hypotheses implied by cladograms, i.e. character origin/fixation [$X(0)$ → $X(1)$] and subsequent population splitting events (s_n). See text for discussion.

Applying the theory in **[4]** to effects leads to a minimum of two classes of hypotheses being implied by a cladogram: character origin/fixation and population splitting events. Distinguishing these hypotheses is often overlooked in lieu of only indicating character changes on branches (Fig. 3B), which is consistent with emphasis on support being characterized as 'branch' or 'group support' and the view that quantification (e.g. via indirect methods like resampling or Bremer support, see below) of such 'support' has some sort of relevance, largely under the misconception that the effects being explained by theory, i.e. character data, are relevant to assessing support in terms of the pursuit of understanding. But branch/group support is not equivalent to actually indicating support in terms of relations between premises and conclusion(s), and any assessment of understanding comes not from the premises used to infer hypotheses, but rather as consequences of testing *sensu* **[3]**. Regardless, the matter of documenting initial understanding provided by cladograms is uncomplicated and trivial. It is uncomplicated because associating particular premises to individual hypotheses implied by a cladogram is simple to document. It is trivial for the fact that the act of applying the theory as fully as possible to effects will result in the simplest (i.e. most parsimonious *and* with greatest likelihood) thus best supported hypothesis(es), albeit the explanatory scope of a cladogram is quite meager given the theory applied. But in such an instance, to say a hypothesis is best supported is only to acknowledge that effects are explained as fully as possible as instances of the applied theory(ies). Rather than branch or group support, the actual support for the hypotheses in Fig. 3B are summarized in Fig. 3C. The cladogram in Fig. 3B implies 18 hypotheses—12 regarding character origin/fixation and six for population splitting events. Among these hypotheses, three are *ad hoc*, as their inference goes beyond the stipulated theory in **[4]**. Support for all hypotheses is represented by the conjunctions of theory, indicated in the premises or *ad hoc*, and the particular characters being explained.

Acknowledging that character data plus theory abductively lead to explanatory hypotheses (Fig. 3A–B), it is necessary to recognize that referring to support for those hypotheses is not conveyed by branching diagrams. In other words, support is not in terms of relations between premises and cladograms, but rather premises and individual hypotheses of character origin/fixation and population splitting events that are *implied by* cladograms. For instance, it would be incorrect to say that 'group' (DEFGH) is better supported than (EFGH) because the branch subtending (DEFGH) shows more (non-*ad hoc*) character changes than does the branch for (EFGH) (Fig. 3B). There are instead five equally supported hypotheses implied by (DEFGH) and two implied by (EFGH) (Fig. 3C). Support is not considered among groups or clades but rather among the various hypotheses implied by those groups/clades.

While abductive support for the hypotheses implied by a cladogram are maximized by the extent to which *ad hoc* hypotheses are avoided, as a matter of the conjunction of theory and characters, such support is unremarkable in that it is necessitated by the premises. Note as well that this support is not relevant to assessing the causal conditions implied by the cladogram or stated in the hypotheses. Actual evaluative support for those causal conditions would have to come from testing, which requires evidence well beyond character data (cf. **Increasing causal understanding—testing as intended**). These considerations have implications for indirect measures of support garnered by resampling or Bremer support methods, discussed later.

Considerations of support in terms of test evidence has special significance for cladograms, given the emphasis on testing that has been associated with the development of phylogenetic systematics (cf. **Increasing causal understanding—testing as intended**). Identifying support for hypotheses presented in cladograms require evidence of the form shown in the premises in **[3]**, especially 'actual confirming/disconfirming evidence,' which allows for either supporting a hypothesis or suggesting alternatives. A cursory comparison of abductive and inductive inferences in **[1]** and **[3]**, respectively, indicates that the class of evidence used to support the initial inferences of hypotheses asserting particular causal events is not the same as the class of evidence required as tests of those hypotheses (Fitzhugh 2006a, 2008a, 2010a).

Testing á la Popper. A perspective begun in the 1970's, continuing to the present, is that evolutionary hypotheses in the form of cladograms are routinely tested as a consequence of the introduction of new characters. It was remarked in the previous section that this was a view held by Hennig (1966), but it was the attempted association of Karl Popper's (e.g. 1959) writings on testing that placed the greatest emphasis on equating character data with test evidence (e.g. Wiley 1975; Gaffney 1979; Eldredge & Cracraft 1980; Wiley 1981; Rieppel 1988; Faith & Cranston 1992; Kluge 1997a, b, 1999, 2001; Grandcolas *et al.* 1997; Siddall & Kluge 1997; Wenzel 1997; Schuh 2000; de Queiroz & Poe 2001, 2003; Farris *et al.* 2001; Faith & Trueman 2001; Faith 2004, 2006; Schuh & Brower 2009; Jenner 2003; de Queiroz 2004; Franz 2005; Helfenbein & DeSalle 2005; Wägele 2005; Egan 2006;

Grant & Kluge 2008; Wheeler 2004, 2010; Brower 2011; Faith *et al.* 2011; Rindal & Brower 2011; Hovenkamp 2012; cf. Fitzhugh 2006a, Vogt 2008 for critical overviews; see also Sober & Steel 2002, and Sober 2008 for similar, yet non-Popperian perspectives). The purported process of Popperian testing in systematics is as follows. With the introduction of new characters, 'predicted' or otherwise, and their integration into an existing data matrix, a new round of cladogram inference is performed. If the topology/topologies of old and new cladograms are the same, it is claimed that this is an instance of corroboration (*sensu* Popper), whereas differences between topologies would mean the earlier cladogram(s) has been falsified. For example, given the observations 0011 and 0111 among individuals to which respective species hypotheses *a-us*, *b-us*, *c-us*, and *d-us* apply, the inferred cladogram would be (*a-us* (*b-us* (*c-us*, *d-us*))). Additional characters, 0101 and 0101, are subsequently observed and a new cladogram inferred, (*a-us* (*c-us* (*b-us*, *d-us*))). The standard position is that (*a-us* (*b-us* (*c-us*, *d-us*))) has been falsified in lieu of (*a-us* (*c-us* (*b-us*, *d-us*))).

There are fundamental problems with this approach. The relation to Popper, much less testing in any hypothetico-deductive sense (e.g. **[2]**, **[3]**), is entirely illusory. While only infrequently speaking of the testing of hypotheses as opposed to theories, Popper (1962: 241; 1966: 260–269, 362–364; 1983: 192–193, 349–352; 1992: 132–134; 1994: 124, 133) held the already articulated view (e.g. Peirce 1932: 2.776; 1958a: 7.182, 7.206; Haack *in* Haack & Kolenda 1977: 69; Brent 1998: 117; see also Fitzhugh 2006a, 2010a) that test evidence must be consequences not only aligned as closely as possible with the causal conditions stated in the hypothesis, but those consequences must be of a variety different from and independent of the effects the hypothesis explains (Cleland 2001, 2002, 2011b). Character data cannot serve as test evidence of hypotheses intended to causally account for those effects. This confusion between evidence used to (abductively—**[1]**) infer a hypothesis and evidence (deductively or inductively—**[2]**, **[3]**) inferred for the purpose of subsequently testing (by way of induction) that hypothesis is what Lipton (1991, 2004, 2005; see also Maher 1988; Mayo 1996; Achinstein 2001; Fitzhugh 2010a; Cleland 2011b) characterized as accommodation versus prediction. A phylogenetic hypothesis is inferred to accommodate character data. In turn those data offer no opportunity to critically assess the causal conditions inferred to explain those data. Only relying on accommodated data makes it too easy to claim support for a hypothesis, with no inherent risk of refutation. It is not the case that (*a-us* (*b-us* (*c-us*, *d-us*))) has been falsified/disconfirmed relative to (*a-us* (*c-us* (*b-us*, *d-us*))) by the inclusion of new observations of characters. Indeed, no test has been performed. Observations of new characters cannot be validly deduced or predicted from an existing topology, such that those characters are consequences relevant to the assessment of the causal conditions asserted by the hypothesis (Sober 1988; Fitzhugh 2005a, 2006a, 2008a, 2010a; Vogt 2008). Since testing pertains to assessing causal claims, the relevant test observations would have to be effects that are as closely related as possible to the nuances of the conditions expounded in each hypothesis. Cladogram (*a-us* (*b-us* (*c-us*, *d-us*))), as a causal account, only has explanatory relevance to 0011 and 0111. The causal conditions presented in the hypothesis, albeit extremely vague, do not lend themselves to making predictions of other characters. Second, the inferences that led to (*a-us* (*b-us* (*c-us*, *d-us*))) and (*a-us* (*c-us* (*b-us*, *d-us*))) are both abductive, and as such, the hypotheses have no evaluative capacity relative to one another. At best one can say that (*a-us* (*b-us* (*c-us*, *d-us*))) has been *replaced by* (*a-us* (*c-us* (*b-us*, *d-us*))) for the fact that the explanations of new observations 0101 and 0101 have relevance to the explanations of old observations 0011 and 0111 (see next section). Proceeding from the first inference to the next has resulted in no net positive or negative change in understanding. Rather, the two hypotheses are equivalent in that both only provide initial, equally vague answers to different sets of causal questions.

Testing via disjunct hypotheses. Related to the misconception that character data alone serve as tests of phylogenetic hypotheses, there is the popular tendency to compare hypotheses inferred from different sets of data, most commonly 'morphology' versus nucleotide sequences (e.g. Asher *et al.* 2003; Asher *et al.* 2008; Chen *et al.* 2003; von Dohlen *et al*. 2006; Crespo *et al*. 2007; Springer *et al*. 2007; Bourlat *et al*. 2008; Dunn *et al*. 2008; Prasad *et al*. 2008; Bailey *et al*. 2010; Regier *et al*. 2010; Meredith *et al.* 2011; Philippe 2011; Rota-Stabelli *et al.* 2011; Vila *et al*. 2011; Crawford *et al*. 2012; see also examples discussed in Mooi & Gill 2010). The reasoning here is that congruence among topologies inferred from 'independent' data sets provides a measure of support or corroboration [*sic*] (e.g. Lienau & DeSalle 2010) for the overall 'phylogeny' of a group of organisms. There is, however, a two-fold problem with this approach. First, comparing cladograms inferred from sets of data that are explanatorily relevant to one another violates one of the basic tenets of rational reasoning—the requirement of total evidence (RTE). The RTE stipulates that if evidence has relevance, positive or negative, to the support for a particular conclusion, then that evidence must be taken into consideration as part of the premises used to infer that conclusion (Carnap 1950; Barker 1957; Hempel 1962, 1965, 1966, 2001; Salmon 1967, 1984a, 1984b, 1989, 1998; Sober 1975; Fetzer 1993; Fetzer & Almeder 1993; Lecointre & Deleporte 2005; Fitzhugh, 2006a, 2006b). While regarded as necessary to engage in rational non-deductive reasoning (the requirement is automatically satisfied in deduction), the RTE has largely been either misconstrued, pro or con (Eernisse & Kluge 1993; Kluge & Wolf 1993; Nixon & Carpenter 1996; Kluge 1989, 1998, 2004; Rieppel 2003a; Lecointre & Deleporte 2005), or ignored (Bull *et al*. 1993; de Queiroz 1993; Miyamoto & Fitch 1995; Levasseur & Lapointe 2001) in the systematics literature. Oddly, the RTE is often equated with Popper's notion of corroboration (Nixon & Carpenter 1996; Kluge 2004; Lienau & DeSalle 2010; cf. Fitzhugh 2006b) or verificationism (Bucknam *et al*. 2006), when in fact the principle transcends *all* inferential practices. Simply applying the RTE is not tantamount to testing. Regardless, systematists routinely speak of topological similarities and differences between cladograms inferred from different sets of data. Yet just such a maneuver that violates the RTE also indicates the evidential relevance of those data to one another for the sake of causally accounting for their occurrences. Ignoring the RTE in such instances provides no basis for critically assessing understanding from the perspective of testing. Rather it is the explanatory worthiness of cladograms, meager as it is, that is scarified in the name of a version (in name only) of testing that is just as distorted and vacuous as equating 'total evidence analyses' with testing.

The most apparent consequence of denying the RTE in biological systematics research has been the view that disparate cladograms can be evaluated against one another, with congruence between topologies offering empirical support for a 'phylogeny.' Note however that in the context of cladograms, the term phylogeny refers to the sum total of causal events explaining relevant properties of organisms (cf. Table 1, Fig. 1). Drawing comparisons of branch arrangements *between* cladograms inferred from different data sets is, by definition, an exercise divorced from phylogeny. Such comparisons are without epistemic merit. Consider the following sets of data and inferred explanatory hypotheses:

[5] (a) • auxiliary theory(ies)
• phylogenetic theory(ies) *A*, *B*, *C*,...*n*
• 'morphology' data set *X*

• (*a-us* (*b-us* (*c-us*, *d-us*)))

(b) • auxiliary theory(ies)
• phylogenetic theory(ies) *A*, *B*, *C*,...*n*
• nucleotide data set *Y*

• (*a-us* (*b-us* (*c-us*, *d-us*))).

Both sets of inferences lead to explanatory hypotheses that provide some degree of initial causal understanding of the respective sets of observations. Does the 'congruence' between these topologies provide evidential support beyond what is initially offered by the separate sets of premises? In other words, do mere similarities in branch arrangements offer assessments of our ultimate causal understanding of observations? No. The two hypotheses, as explanatory constructs, have no relevant meaning to one another. Each provides vague causal accountings, per the theories applied in the inferences, to their respective sets of data. What are relevant, per the RTE, are the observations in need of being explained. And that relevance obviates separate inferences. That relevance is all the more apparent from the fact that internodal branches, nodes, and terminal branches of disparate cladograms represent the same specifiable classes of past causal events (Fitzhugh 2009: fig. 19), i.e. novel character origin/fixation followed by population splitting.

It may be asserted that as the inferences in **[5]** are products of 'independent' data sets, congruence of results offers positive support (e.g. Rota-Stabelli *et al*. 2010). For instance Chen *et al*. (2003: 264, emphasis original) claim,

> The congruence of inferences separately drawn from *independent* data is considered as strong indicator of reliability. If we keep in mind the fact that molecular homoplasy may have different effects on tree reconstruction from one gene to another, obtaining the same clade from separate analysis of several genes despite this fact renders the clade even more reliable. In other words, obtaining the same tree or even some common clades means that there is a common structure in these data sets that must come from common evolutionary history.

While such a notion of independence of evidence might appear consistent with what is actually required for testing (e.g. Popper 1992; Cleland 2001, 2002, 2011a; Fitzhugh 2006a, 2010a), the similarity is simply a consequence of the misuse of terms. Claiming, for instance, that ‘morphological’ characters are ‘independent’ of nucleotides might make sense if one is segregating these features according to particular criteria for the sake of establishing a classificatory arrangement of observations, e.g. ‘cellular’ as opposed to ‘molecular.’ But independence *qua* classes does not automatically translate into independence of test evidence. An equally serious consequence is that for results from one abductive inference to provide a basis for judging the veracity of results from a separate inference requires some extra-evidential criterion for weighing one hypothesis against another. A typical determining factor is, as alluded to in the quote from Chen *et al*. (2003) the *a priori* perceived problem of homoplasy. If one class of data is deemed to have an inordinate amount of homoplasy this could lead to ‘incorrect’ results. Partitioning data and comparing the separate cladogram topologies is claimed to allow one to judge reliability of the ‘overall phylogeny.’ What makes this an extra-evidential criterion is that one must impose hypotheses of homoplasy prior to even inferring such hypotheses from the data at hand. But such a criterion is erroneous because homoplasy is a class of *ad hoc* hypothesis that is the product of the abductive inference of phylogenetic hypotheses (Fitzhugh 2006a, 2006b). Asserting homoplasy prior to such inferences is nonsensical as it does nothing more than reduce one to concluding that what they perceive as the same characters among a group of organisms are not the same. This is not a matter of homoplasy, but rather the fact that one either cannot trust their own basic cognitive abilities or that they have already explained their observations. In the case of the latter, the ‘same’ characters should immediately be regarded as different. It is only subsequent to this step that one would then engage in phylogenetic inference. If such an extra-evidential criterion did exist, and it does not, it would have to come into consideration prior to making the inferences, again as a simple matter of evidential relevance. Comparing phylogenetic hypotheses for partitioned sets of relevant data is both irrational and counter to subsequently assessing scientific understanding (Fitzhugh 2006a, 2006b, 2008c).

Related to the situation outlined in **[5]**, there is the alternate approach of taking partitioned data sets and applying different theories in the inferences of phylogenetic hypotheses. For example,

[6] (a)	• auxiliary theory(ies)	**(b)**	• auxiliary theory(ies)
	• phylogenetic theory(ies) *A, B, C,... n*		• phylogenetic theory(ies) *J, K, L,... n*
	• ‘morphology’ data set *X*		• nucleotide data set *Y*
	• (*a-us* (*b-us* (*c-us*, *d-us*)))		• (*a-us* (*b-us* (*c-us*, *d-us*))).

Usually poorly articulated, and questionable in their justification (cf. Fitzhugh 2006a, 2006b), the common phylogenetic ‘theories’ include what are referred to as parsimony, maximum likelihood, and Bayesian. As with the previous example, it is customary to conclude that congruence between topologies offers some measure of support. The inherent problem that precludes such a conclusion, beyond violating the RTE and the specious argument from independence, is that one must assume that cladograms represent something beyond the causal conditions they imply per the theories used in their inference. Otherwise, to draw comparisons between cladograms that connote different causal parameters is a meaningless exercise. As the only scientifically viable way to interpret cladograms is that they are sets of vague explanatory hypotheses (Fig. 3B), comparisons of cladograms/hypotheses inferred from different theories is not a surrogate for testing and can provide no empirical critique of any degree of ultimate understanding.

Finally, it might be argued that inferences of disjunct phylogenetic hypotheses using partitioned data sets are consistent with William Whewell’s (1847) ‘consilience of inductions.’ The commonly referred to characterization of this doctrine comes from volume two of Whewell’s (1847: 469, emphasis original) *Philosophy of the Inductive Sciences*, aphorism XIV: “*The Consilience of Inductions* takes place when an Induction, obtained from one class of facts, coincides with an Induction, obtained from another different class. This Consilience is a test of the truth of the Theory in which it occurs.” But as Laudan (1981: 165, emphasis original) noted in his analysis of Whewell’s writings on the subject, the principle is implemented under circumstances not always related to testing:

(1) When an hypothesis is capable of explaining two (or more) known classes of facts (or laws);
(2) When an hypothesis can successfully *predict* “cases of a *kind different* from those which were contemplated in the formation of our hypothesis;”

(3) When an hypothesis can successfully predict or explain the occurrence of phenomena which, on the basis of our background knowledge, we would not have expected to occur.

Laudan (1981: 166) pointed out that (1) is just a consequence of compiling relevant data for abductive inferences, much along the lines of what is stipulated by the RTE. It is an action of "…*formal unification* or *simplification* of our theories and hypotheses. By reducing two classes of phenomena—which had hitherto required separate and (seemingly) independent hypotheses or theories for their explanation—to one general hypothesis or theory," and "achieves a reduction in the theoretical baggage required to 'carry' the known phenomena." The circumstances surrounding (2) and (3) are more crucial in that these are actions leading to increased empirical content and understanding as consequences of testing per **[2]** and **[3]** (cf. **Increasing causal understanding—testing, as intended**, below). Whewell's consilience of inductions cannot justify the inferences of disparate phylogenetic hypotheses from partitioned sets of data, much less purported empirical comparisons of those hypotheses as matters of testing. Circumstance (1) simply brings together relevant empirical content for the initial purpose of inferring an explanatory account. And just as was noted in the previous section, **Testing á la Popper**, (1) is not a surrogate for testing.

Resampling methods. Attempts to garner support [*sic*] for phylogenetic hypotheses have also come from the adoption of procedures such as the bootstrap (Felsenstein 1985, 2004; Efron 1979; Efron & Tibshirani 1993; Efron *et al*. 1996; Holmes 2003; Soltis & Soltis 2003), jackknife (Farris *et al*. 1996; Miller 2003), and permutation tests (Faith & Cranston 1991; cf. Egan 2006 for a review of all of these approaches). These methods were originally developed to test statistical hypotheses through a process of random resampling, with or without replacement depending on the method, from among members of an original sample distribution to determine confidence intervals on a population parameter. Applications of these methods to phylogenetic hypotheses occur by randomly sampling characters from an original data matrix to create contrived data matrices of the original dimensions, from which new cladograms are inferred. The frequencies of groups or clades occurring among the cladograms are compared to groups/clades present in the original cladogram(s). The idea is that the more frequent the occurrences of contrived clades identical (in topology only) to those in the original cladogram(s) being 'tested,' the greater the support accorded those clades.

There are several problems associated with attempting to claim support for phylogenetic hypotheses by the use of resampling methods (cf. discussion in Fitzhugh 2006a). The first is that these methods are intended for testing statistical, not explanatory hypotheses (but see Farris 2002, and Goloboff *et al*. 2003, for views that resampling does not require statistical assumptions). Statistical hypotheses characterize the properties of a class or population, while explanatory hypotheses offer past causal conditions or events that account for specific, present effects. The distinction with regard to testing is the nature of the test evidence. Statistical hypotheses rely on test evidence that is the same class of effects from which hypotheses are inferred. Effects that serve as test evidence for explanatory hypotheses are independent of the class of effects from which hypotheses are inferred, such that that evidence consists of effects that are related as narrowly as possible to the hypothesized causal events/conditions, thus having the lowest probability of occurrence if hypothesized causal conditions did not occur (cf. **[2]**, **[3]**; see also **Descriptive, proximate, and ultimate understanding**; **Testing à la Popper**; **Increasing causal understanding—testing as intended**). The application of resampling methods to evaluate the initial, abductive understanding afforded by phylogenetic hypotheses fails for much the same reason that the addition of new character data cannot serve as test evidence (see **Testing à la Popper**)—character data of organisms are not relevant test evidence to judge the veracity of the hypotheses of causal conditions inferred to explain those data. A second notable problem is related to what was outlined earlier (**Testing and support issues**) regarding the distinction between abductive support for hypotheses versus support by way of testing. What is at issue when speaking of support for a cladogram is not topological groups, but rather the separate hypotheses of character origin/fixation and population splitting events implied by cladograms (Fig. 3). As one cannot establish that a 'clade' or 'group' inferred from a resampling procedure refers to the empirically identical phylogenetic hypotheses (cladogram) being evaluated, any comparisons are between nothing more than branching diagrams, not clades-as-composite-hypotheses. As was noted under **Testing and support issues**, abductive support for hypotheses is immediately constrained by the conjunctions of theory and effects (cf. **[1]**). And while that support can be directly tabulated (Fig. 3), it is not the requisite support needed to evaluate the causal claims among those hypotheses.

Bremer 'support'. A fourth type of technique to determine hypothesis support is the Bremer support analysis or decay index (Bremer 1988, 1994; Davis 1995)[4]. As with resampling methods discussed earlier, the support to which Bremer refers is individual clades in a cladogram, determined by how many extra steps are required on a cladogram to eliminate each clade. The idea is that the larger the number of steps required to alter topologies the better supported is the overall cladogram. Fitzhugh (2006a: 100, emphasis added) pointed out that Bremer support offers no semblance of support for phylogenetic hypotheses:

> The examination of cladograms of greater length for the eventual ‘collapse’ of what is interpreted as the ‘same clade’ cannot have any empirical meaning regarding the actual evidential support that allows for that clade in the original hypothesis being ‘evaluated.’ Hypotheses of greater length stand on their own as independent causal accounts that are derived from premises that differ from those used to infer a ‘minimum-length’ cladogram. The difference in premises would have to refer to the interpretations of at least some shared similarities as not being the same characters, which would imply a different set of causal questions from what were originally asked. *The overall consequence is that the exercise of comparing these hypotheses with the hypothesis in question, much less noting the number of steps required to collapse clades of that hypothesis, cannot provide the indication of support that proponents have suggested*.

As with the other purported test or evaluative procedures outlined so far in this section, Bremer support is nothing but an exercise in ‘branch manipulation.’ It has no relation to the observations that resulted in a hypothesis, much less assessing the underlying causal events it implies (*contra* Grant & Kluge 2007).

Bremer support has also been used as an argument to combine ‘morphological’ and nucleotide sequence data. For instance, in their inferences of phylogenetic hypotheses among placental mammals from partitioned and combined data, Lee and Camens (2009: 2244) noted that “when morphology is combined with extensive molecular data, morphology increases branch (Bremer) support for every clade in the preferred [*sic*] tree.” This is a matter of incorrectly conflating the requirement of total evidence (RTE) with testing. Bremer support has no relation to, and provides no epistemic basis for the RTE. As discussed earlier, the RTE is not an *ex post facto* criterion. On the matter of evidential support for clades, Bremer-support values are detached from reality. As clades are diagrammatic representations capable of nothing more than implying hypothesized causal events (Fig. 3), no amount of character data can be brought to bear on the subject of support for or against such hypotheses that account for those data. Pertinent support, and thus further causal understanding, is garnered as a consequence of proper testing.

Increasing ultimate causal understanding—testing as intended. The relation between observed characters of organisms and specific/phylogenetic hypotheses is one of effects being explained by ultimate causes. This is a direct consequence of the overarching goal of scientific inquiry, i.e. to pursue causal understanding by way of explanations of observed effects and predictions of future phenomena (Hempel 1965: 139). It is this relation that establishes what would be required to test specific and phylogenetic hypotheses. A contrived example presented by Fitzhugh (2010a) can be used to schematically outline this process. To begin, consider as background knowledge that there are known groups of organisms to which specific hypotheses *a-us*, *b-us*, etc., have been applied in the past. New individuals with unanticipated or surprising characters are subsequently observed (Fig. 4A). These new observations lead to three causal questions[5]:

[7] Q_1: Why do some of these individuals have a white spot in contrast to completely black?
Q_2: Why do some of these individuals have antennae in contrast to a smooth dorsum?
Q_3: Why do individuals to which specific hypotheses *x-us* and *y-us* apply have ventral appendages?

Answers to questions Q_1 and Q_2 are provided by way of the abductive inferences of specific hypotheses *x-us* and *y-us*, respectively. For instance, the inference providing an answer to Q_2 has the form (cf. Fitzhugh 2009):

[4] The assessment in this section also applies to the derivative ‘ratio of explanatory power’ (REP) of Grant and Kluge (2007, 2008). Contrary to claims by these authors, hypothesis support in the context of phylogenetic inference—given that it is abductive—is nothing more than the relation between premises and conclusion(s) (cf. **[1]**). Making a distinction between support and optimality as suggested by Grant and Kluge is gratuitous.

[5] Regardless of observations being ‘morphological’ or nucleotide sequences, the causal questions would be identical in form (cf. Fitzhugh 2006a–c).

[8] **Species Theory:** If character *x*(1) originates by mechanisms *a, b, c…n*, among gonochoristic or cross-fertilizing hermaphroditic individuals of a reproductively isolated population with character *x*(0), and *x*(1) subsequently becomes fixed throughout the population during tokogeny by mechanisms *d, e, f… n*, then individuals observed in the present will exhibit character *x*(1).

Observations (effects): Individuals have a dorsal margin with antennae in contrast to a smooth dorsal margin as seen among individuals to which other specific hypotheses (*a-us*, *b-us*, etc.) refer.

Causal Conditions (specific hypothesis *y-us*): The antennate dorsal margin condition originated by unspecified mechanisms within a reproductively isolated population with smooth dorsal margins and eventually became fixed throughout the population during tokogeny by additional unspecified mechanisms.

Answering Q_3 is also by way of abduction, but in this instance to a phylogenetic hypothesis (cf. Fitzhugh 2009):

[9][6] **Phylogenetic Theory:** If character *x*(0) exists among individuals of a reproductively isolated, gonochoristic or cross-fertilizing hermaphroditic population and character *x*(1) originates by mechanisms *a, b, c… n*, and becomes fixed within the population by mechanisms *d, e, f… n* (=ancestral species hypothesis), followed by event(s) *g, h, i… n*, wherein the population is divided into two or more reproductively isolated populations, then individuals to which descendant species hypotheses refer would exhibit *x*(1).

Observations (effects): Individuals to which specific hypotheses *x-us* and *y-us* refer have ventrolateral margins with appendages in contrast to smooth as seen among individuals to which other species hypotheses (*a-us*, *b-us*, etc.) refer.

Causal Conditions (phylogenetic hypothesis *X-us*): Ventrolateral margin appendages originated by some unspecified mechanism(s) within a reproductively isolated population with smooth ventrolateral margins, and the appendage condition became fixed in the population by some unspecified mechanism(s) (= ancestral species hypothesis), followed by an unspecified event(s) that resulted in two or more reproductively isolated populations.

Although separately inferred as responses to questions Q_1–Q_3, the answers are graphically represented by cladogram (*a-us bus* (*x-us y-us*)). Notice the stark contrast between this portrayal of phylogenetic inference and Sober's (2002: 157) characterization: "The first problem is… one infers a tree;… one uses an inferred tree to solve a further problem [of ancestral character transformation]."

[6] This inference requires comment, as it departs from standard approaches, i.e. 'parsimony,' 'maximum likelihood,' 'Bayesian.' While there is some resemblance to what is commonly referred to as 'parsimony analysis,' the premises are determined by question Q_3, not parsimony. Parsimony is, however, a relevant factor in linking the question to an inference that maintains as much as possible the empirical content in the question (Sober 1975). As abductive inference using a common cause theory will lead to 'most parsimonious' conclusions, the hypothesis is by definition also of maximum likelihood in the sense of (abductive) support accorded the hypothesis by the premises (Sober 1988; Fitzhugh 2006a, 2006b). What is referred to as 'maximum likelihood analysis' (ML) (*sensu* Felsenstein 1981, 2004; Swofford *et al.* 1996; Huelsenbeck & Crandall 1997), however, is problematic in that it implements a theory ('model') that is at odds with phylogenetic-based causal questions. ML concerns itself with branch lengths, thus the explanatory scope cannot be phylogenetic, but rather specific or intraspecific (Fig. 1; cf. Fitzhugh 2006a, 2006b). 'Bayesian analysis' (e.g. Huelsenbeck *et al*. 2001; Huelsenbeck & Ronquist 2001; Archibald *et al*. 2003; Ronquist *et al*. 2009) is erroneous for the fact that Bayesianism addresses determinations of hypothesis acceptance/belief on the basis of test evidence *subsequent to* hypothesis inference. 'Bayesian analysis' erroneously estimates phylogenetic hypotheses using posterior probabilities derived from the character data explained by those hypotheses. To make such 'inferences,' one must rely upon empirically empty cladograms and a theory that is inconsistent with relevant causal questions (as in ML), and treat character data as test evidence [*sic*].

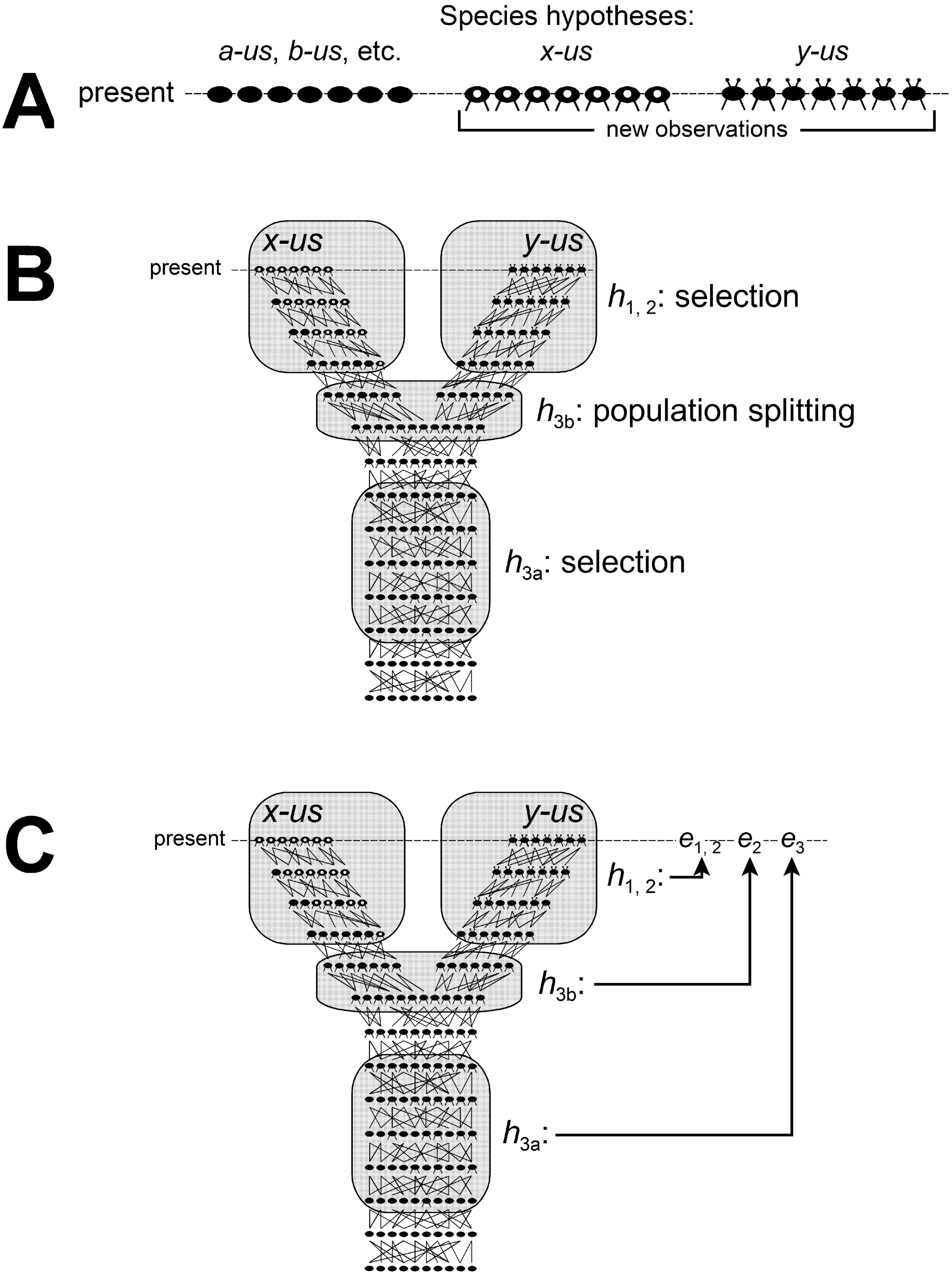


FIGURE 4. Relations between observations (A) leading to specific and phylogenetic hypotheses (B), and the evidence required to test those hypotheses (C). Modified from Fitzhugh (2010a: fig. 2).

With regard to testing (*a-us bus* (*x-us y-us*)), we have to acknowledge what hypotheses are involved. If (*a-us bus* (*x-us y-us*)) conveys the products of the above inferences, then a minimum of four hypotheses are candidates for testing (Figure 4B). Specific hypotheses *x-us* and *y-us* provide respective explanations of characters by way of origin and subsequent fixation in ancestral populations (Fig. 4B, $h_{1,\ 2}$: selection). The phylogenetic hypothesis actually entails at least two hypotheses, origin and subsequent fixation of appendages (Fig. 4B, h_{3a}: selection) and subsequent population splitting (Fig. 4B, h_{3b}: population splitting; cf. Fig. 3). The test evidence required for $h_{1–3}$ would have to be respective effects that could be associated as narrowly as possible with each of the causal events in the hypotheses (Fig. 4C, e_1–e_3; cf. **[2]**, **[3]**). But this would first necessitate filling out the causal conditions with sufficient specifics such that useful predictions of relevant test evidence can be specified. For instance, hypotheses $h_{3a–b}$, represented by (*a-us bus* (*x-us y-us*)) (Fig. 4C), might be expanded to state that ventrolateral appendages originated within an ancestral population, and that feature conferred a selective advantage in competition for food resources, leading to fixation of the character in the population. This modest increase in detail might allow for predicting consequences from these conditions that could serve as potential test evidence, e.g. presence of food remains indicating a shift in diet associated with individuals with appendages, correlated with increasing frequency of remains of individuals with appendages in a particular region. In similar fashion, evidence of a population splitting event would require stipulating causal specifics, e.g. vicariance via tectonic events, from which effects as narrowly associated as possible with such a class of events might be predicted. Schematically, predictions of potential test evidence for $h_{3a–b}$ would have the form (cf. **[2]**),

[10] **Phylogenetic Theory:** If character *x*(0) exists among individuals of a reproductively isolated, gonochoristic or cross-fertilizing hermaphroditic population and character *x*(1) originates by mechanisms *a, b, c… n*, and becomes fixed within the population by mechanisms *d, e, f… n* (=ancestral species hypothesis), followed by event(s) *g, h, i… n*, wherein the population is divided into two or more reproductively isolated populations, then individuals to which descendant species hypotheses refer would exhibit *x*(1).

Causal Conditions (phylogenetic hypothesis *X-us*): Ventrolateral margin appendages originated by events X_1, X_2, X_3,... *n* within a reproductively isolated population with smooth ventrolateral margins, and the appendage condition became fixed in the population via events Y_1, Y_2, Y_3,... *n* (= ancestral species hypothesis), followed by events Z_1, Z_2, Z_3,... *n* that resulted in two or more reproductively isolated populations.

Original observations (effects): Individuals to which specific hypotheses *x-us* and *y-us* refer have ventrolateral margins with appendages in contrast to smooth as seen among individuals to which other species hypotheses (*a-us*, *b-us*, etc.) refer.

Predicted test consequences: Effects $X_{1'}$, $X_{2'}$, $X_{3'}$,... *n* and $Y_{1'}$, $Y_{2'}$, $Y_{3'}$,... *n* should be observed, indicating the causal events of character origin and fixation of appendages, respectively, among individuals of an ancestral population (cf. Fig. 4C: h_{3a}), and effect(s) $Z_{1'}$, $Z_{2'}$, $Z_{3'}$,... *n* should be observed, indicating occurrences of causal events resulting in splittings of populations into separate, reproductively isolated groups (cf. Fig. 4C, h_{3b}).

Notice that potential test consequences from the selection hypothesis are independent of the effects that prompted inference of the hypothesis. While both classes of evidence are inferred to be effects of a common causal event, evidence of selective advantages for the presence of appendages lie beyond the mere presence of those appendages. This is the independence of evidence referred to by Popper (e.g. 1992: 132–133) regarding effects being explained by a hypothesis and effects serving as test evidence for that hypothesis. Within the mechanics of testing specific or phylogenetic hypotheses, simply segregating suites of characters into different classes is not tantamount to this type of independence.

Actually carrying out the testing of hypotheses h_{3a} and h_{3b} (Fig. 4C) would have the form presented in **[3]**:

[11] **Auxiliary theory(ies):** Stated as relevant and necessary to the test.

Phylogenetic Theory: If character *x*(0) exists among individuals of a reproductively isolated, gonochoristic or cross-fertilizing hermaphroditic population and character *x*(1) originates by mechanisms *a, b, c… n*, and becomes fixed within the population by mechanisms *d, e, f… n* (=ancestral species hypothesis), followed by event(s) *g, h, i… n*, wherein the population is divided into two or more reproductively isolated populations, then individuals to which descendant species hypotheses refer would exhibit *x*(1).

Actual test conditions: Descriptions of actions taken to enable potential observations of test results.

Test results: Effects $X_{1'}$, $X_{2'}$, $X_{3'}$,... *n* and $Y_{1'}$, $Y_{2'}$, $Y_{3'}$,... *n* are observed, indicating the causal events of character origin and fixation of appendages, respectively, among individuals of an ancestral population (cf. Fig. 4C: h_{3a}), and effect(s) $Z_{1'}$, $Z_{2'}$, $Z_{3'}$,... *n* are observed, indicating occurrences of causal events resulting in splittings of populations into separate, reproductively isolated groups (cf. Fig. 4C, h_{3b}).

Conclusions: Hypotheses h_{3a} and h_{3b} are confirmed.

There is, however, the consequence of realizing the actual limitations to testing specific and phylogenetic hypotheses—the time that has elapsed between the hypothesized causes and observations of effects in the present can severely limit or preclude the existence of test evidence. While testing is open to being potentially accomplished, it might not be feasible given inherent constraints. In the absence of both filling out (*a-us bus* (*x-us y-us*)) to the point of providing potential test evidence (cf. **[2]**, **[10]**) and actually engaging in testing (cf. **[3]**, **[11]**), ultimate causal understanding provided by specific and phylogenetic hypotheses remains rudimentary in that it is only the simple conjunctions of theories and effects-to-be-explained (cf. **[1]**, **[8]**, **[9]**).

Conclusions

Scientific understanding occurs by way of explanation through the fitting of observations into some broader theoretical framework, not only by offering initial information about possible causes but also through the ability to anticipate and investigate consequences related to those causes as matters of critical evaluation. Understanding is also context dependent in that it is a state of mind contingent on what individuals regard as being sufficient for meeting their standard of understanding. What provides an adequate explanation for one individual might be unsatisfactory to another. As a result, some yardstick by which to judge the adequacy of understanding is required. Surely any modicum of consensus on the adequacy of hypotheses in the sciences should come from the extent to which results from empirical testing are manifested. Therein lay two problems for biological systematics. Ultimate hypotheses, especially specific and phylogenetic, are often devoid of causal details, such that the state of understanding, beyond the initial explanatory notions presented under the rubric of ‘taxa’ or cladograms are neither pursued nor enhanced. Even if these hypotheses are filled out to the point that valid test predictions can be stipulated, it is likely that actual testing will be impractical in nearly all instances, as noted earlier. Instead, critical assessments of hypotheses are stalled, such that there is the tendency to orient back to original character observations, such as interesting correlations among features, or pursuing investigations of the finer structural components of characters (descriptive) or their ontogenetic development (proximate). In other words, the general reaction is to move backward in an epistemic sense to consider the enhancement of descriptive and proximate understanding, rather than actually pursuing ultimate understanding in terms of critical hypothesis assessment. Taken at face value, there is nothing wrong with such a maneuver, but we do need to be cognizant that enhancing descriptive or proximate understanding does nothing to promote continued ultimate understanding as conceived as testing in the sciences. Several of the classes of systematics hypotheses fail to provide substantive growth in causal understanding for the fact that our tendency is to maneuver away from testing those hypotheses.

It is with the advent of nucleotide sequencing that the magnitude of this problem has increased. For instance, in what way does an ultimate hypothesis lead to the pursuit of causal understanding of nucleotides in a particular sequence? While we can readily associate epistemic consequences of causal questions regarding observations of ‘morphological’ characters and the cladograms that serve as vague answers (cf. **[7]**–**[11]**), what are the merits to asking questions of the form, "Why do I observe a T at position 482 as opposed to A in this string of nucleotides"? Indeed, such questions are necessarily implied, just as they are for any other classes of characters, in a data matrix (Fitzhugh 2006c). In other words, what understanding is attained by inferring cladograms to answer such questions if the accumulation of sequence data cannot promote a process of reexamining and testing descriptive, proximate or ultimate understanding? Are such questions at the level of individual nucleotides even appropriate or relevant? The routine pattern in 'molecular' systematics is to (a) sequence nucleotides, (b) infer cladograms (or groups of mutually exclusive cladograms from partitioned data sets and/or contradictory theories), (c) publish cladograms, (d) proceed to another project and repeat steps (a)–(c), or (d) perform more sequencing and repeat steps (a)–(c). This approach is exemplified by research programs determining metazoan phylogenetic relationships, e.g. Giribet *et al*. (2000); Halanych (2004); Philippe and Telford (2006); Dunn *et al*. (2008); Philippe *et al*. (2009); Schierwater *et al*. (2009); Philippe *et al*. (2011). With regard to the phylogenetic hypotheses inferred, there is nothing in this pattern of activity that enhances evolutionary understanding—neither in terms of the vague explanatory nature of the cladograms produced, nor by the fact that proper testing is infeasible. There is the clear indication that causal understanding of sequence data is far less important than simply deriving branching diagrams from which one might refer to taxa (= explanatory hypotheses) that have been previously characterized using morphology. This is not a problem limited to considerations of sequence data. Analogous instances of steps (a)–(d) also can be found among groups of organisms with both extensive neontological (morphology, sequence data) and paleontological data, e.g. cetacean phylogeny (cf. review by Uhen 2010 for neontological and paleontological morphological studies; Gatesy 1998, Montgelard *et al*. 2007, O'Leary & Gatesy 2008, Spaulding *et al*. 2009, Xiong *et al*. 2009 regarding sequence data). To that end, the view of O'Leary and Gatesy (2008: 400; see also Spaulding *et al*. 2009: 1, 12) exemplifies this mischaracterization of pursuing understanding, by equating adherence to the requirement of total evidence with testing (cf. **Testing á la Popper**, **Testing via disjunct hypotheses**):

> Continued synthesis of molecular and morphological data from extant and extinct taxa remains the strongest test of phylogenetic hypotheses and the best summary of the common signal in the diverse data available for phylogenetics….

Mooi and Gill (2010: 27) echo the problem just described:

> Solving character conflict is at the crux of systematics. Conflicting hypotheses of relationship can be addressed through: (1) a declaration of one to be true based on our own authority, (2) a re-examination of characters supporting each to discover, understand and potentially resolve conflicts, (3) the introduction of an additional source of data (either from other character complexes or with different or additional taxa) to produce yet another tree, (4) the presentation of the data in a manner where conflict is obscured and avoids scrutiny.

The issue of 'character conflict' is, however, a contrived problem, readily solved by correctly applying the requirement of total evidence in the act of abductively inferring hypotheses. More substantial is the fact that the maneuvers outlined by Mooi and Gill are symptomatic of the erroneous view that ultimate causal understanding in systematics can be achieved solely by manipulations of characters, under the headings of testing, Popperian corroboration, support, etc.

In light of the above consequences, we need to acknowledge that biological systematics is a venue for increasing ultimate understanding that is inherently limited. The greatest strength of systematics has been as a vehicle for prompting one to revert back to considerations of descriptive and proximate understanding, rather than actually pushing forward with critically evaluating ultimate understanding through the process of testing. This conclusion is not to impugn the importance of systematics, but rather to point out that the perceived productivity of systematics research programs, especially those that are specific and phylogenetic, are far more constrained than usually assumed. Recognizing these inherent limitations would aid in better streamlining systematics research to maximize productivity in the sense of actually shifting causal understanding to more descriptive and proximate causal levels.

My intent in outlining standard approaches in systematics in terms of Mayr's (1961) proximate and ultimate causes in biology is to highlight the too often unrealized boundaries actually imposed on the field. Ignoring those limitations has resulted in the development of methodological and research pursuits that are at odds with the goal of attaining causal understanding as part of scientific inquiry. At its best, systematics enhances descriptive understanding, and within limits the pursuit of proximate causal understanding. Where it has been especially remiss is in elevating the importance of specific and phylogenetic hypotheses beyond what they usually are—initial, very vague explanation sketches—as well as claiming increases in evolutionary understanding where none exists.

Acknowledgements

Sincere thanks to Lorenzo Prendini, Robert Mesibov, John A. Long, George D.F. Wilson, and Rasmus Winther for useful comments on earlier drafts.

References

Achinstein, P. (1970) Inference to scientific laws. *In*: Stuewer, R.H. (Ed.), *Volume V: Historical and Philosophical Perspectives of Science*. Minnesota Studies in the Philosophy of Science. University of Minnesota Press, Minneapolis, pp. 87–111.

Achinstein, P. (2001) *The Book of Evidence*. Oxford University Press, New York, 290 pp.

Aliseda, A. (2006) *Abductive Reasoning: Logical Investigations into Discovery and Explanation*. Springer, Dordrecht, 225 pp.

Archibald, J.K., Mort, M.E. & Crawford, D.J. (2003) Bayesian inference of phylogeny: a non-technical primer. *Taxon*, 52, 187–191.

Ariew, A. (2003) Ernst Mayr's 'ultimate/proximate' distinction reconsidered and reconstructed. *Biology & Philosophy*, 18, 553–565.

Asher, R.J., Geisler, J.H. & Sánchez-Villagra, M.R. (2008) Morphology, paleontology, and placental mammal phylogeny. *Systematic Biology*, 57, 311–317.

Asher, R.J., Novacek, M.J. & Geisler, J.G. (2003) Relationships of endemic African mammals and their fossil relatives based on morphological and molecular evidence. *Journal of Mammalian Evolution*, 10, 131–194.

Bailey, A.L., Brewer, M.S., Hendrixson, B.E. & Bond, J.E. (2010) Phylogeny and classification of the trapdoor spider genus *Myrmekiaphila*: an integrative approach to evaluating taxonomic hypotheses. *PLoS ONE*, 5, e12744. doi:10.1371/journal.pone.0012744.

Barker, S.F. (1957) *Induction and Hypothesis*. Cornell University Press, New York, 203 pp.

Beatty, J. (1994) The proximate/ultimate distinction in the multiple careers of Ernst Mayr. *Biology & Philosophy*, 9, 333–356.

Bourlat, S.J., Nielsen, C., Economou, A.D. & Telford, M.J. (2008) Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. *Molecular Phylogenetics and Evolution*, 49, 23–31.

Bremer, K. (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, 42, 795–803.

Bremer, K. (1994) Branch support and tree stability. *Cladistics*, 10, 295–304.

Brent, J. (1998) *Charles Sanders Peirce: A Life*. Indiana University Press, Bloomington, 412 pp.

Brower, A.V.Z. (2006) The how and why of branch support and partitioned branch support, with a new index to assess partition incongruence. *Cladistics*, 22, 378–386.

Brower, A.V.Z. (2010) Stability, replication, pseudoreplication and support. Cladistics, 26, 112–113.

Brower, A.V.Z. (2011) Repeatability and reality. *Cladistics*, 27, 447–448.

Bucknam, J., Boucher, Y. & Bapteste, E. (2006) Refuting phylogenetic relationships. *Biology Direct*, 1, 26.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L. & Waddell, P.J. (1993) Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42, 384–397.

Carnap, R. (1950) *Logical Foundations of Probability*. University of Chicago Press, Chicago, 607 pp.

Chen, W.-J., Bonillo, C. & Lecointre, G. (2003) Repeatablility of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, 26, 262–288.

Cleland, C.E. (2001) Historical science, experimental science, and the scientific method. *Geology*, 29, 987–990.

Cleland, C.E. (2002) Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science*, 69, 474–496.

Cleland, C.E. (2011a) Philosophical issues in natural history and historiography. *In*: Tucker, A. (Ed.), *A Companion to the Philosophy of History and Historiography*. Wiley-Blackwell, Malden, pp. 44–62.

Cleland, C.E. (2011b) Prediction and explanation in historical natural science. *British Journal for the Philosophy of Science*, 62, 551–582.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K. & Glenn, T.C. (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, doi: 10.1098/rsbl.2012.0331.

Crespo, A., Lumbsch, H.T., Mattsson, J.-E., Blanco, O., Divakar, P.K., Articus, K., Wiklund, E., Bawingan, P.A. & Wedin, M.

(2007) Testing morphology-based hypotheses of phylogenetic relationships in Parmeliaceae (Ascomycota) using three ribosomal markers and the nuclear *RPB1* gene. *Molecular Phylogenetics and Evolution*, 44, 812–824.

Curd, M.V. (1980) The logic of discovery: an analysis of three approaches. *In*: Nickles, T. (Ed.), *Scientific Discovery, Logic and Rationality*. D. Reidel Publishing Company, Dordrecht, pp. 201–219.

Davis, J.I. (1995) A phylogenetic structure for the monocotyledons, as inferred from chloroplast DNA restriction site variation, and a comparison of measures of clade support. *Systematic Botany*, 20, 503–527.

de Queiroz, A. (1993) For consensus (sometimes). *Systematic Biology*, 42, 368–372.

de Queiroz, K. (2004) The measurement of test severity, significance tests for resolution, and a unified philosophy of phylogenetic inference. *Zoologica Scripta*, 33, 463–473.

de Queiroz, K. & Poe, S. (2001) Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Systematic Biology*, 50, 305–321.

de Queiroz, K. & Poe, S. (2003) Failed refutations: further comments on parsimony and likelihood methods and their relationship to Popper's degree of corroboration. *Systematic Biology*, 52, 322–330.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sørensen, M.V., Haddock, S.H.D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q. & Giribet, G. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452, 745–750.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.

Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York, 436 pp.

Efron, B., Halloran, E. & Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93, 7085–7090.

Egan, M.G. (2006) Support versus corroboration. *Journal of Biomedical Informatics*, 39, 72–85.

Eldredge, N. & Cracraft, J. (1980) *Phylogenetic Patterns and the Evolutionary Process: Method and Theory in Comparative Biology*. Columbia University Press, New York, 349 pp.

Eernisse, D. & Kluge, A.G. (1993) Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molecular Biology and Evolution*, 10, 1170–1195.

Faith, D.P. (2004) From species to supertrees: Popperian corroboration and some current controversies in systematics. *Australian Systematic Botany*, 17, 1–16.

Faith, D.P. (2006) Science and philosophy for molecular systematics: which is the cart and which is the horse? *Molecular Phylogenetics and Evolution*, 38, 553–557.

Faith, D.P. & Cranston, P.S. (1991) Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. *Cladistics*, 7, 1–28.

Faith, D.P. & Cranston, P.S. (1992) Probability, parsimony, and Popper. *Systematic Biology*, 41, 252–257.

Faith, D.P. & Trueman, J.W.H. (1998) When the topology-dependent permutation test (T–PTP) for a null hypothesis of non-monophyly returns significant support for monophyly, should that be equated with (a) rejecting a null hypothesis, (b) rejecting a null hypothesis of 'no structure', (c) failing to falsify a hypothesis of monophyly, or (d) none of the above? *Systematic Biology*, 45, 577–584.

Faith, D.P. & Trueman, J.W.H. (2001) Towards an inclusive philosophy for phylogenetic inference. *Systematic Biology*, 50, 331–350.

Faith, D.P., Köhler, F., Puslednik, L. & Ballard, J.W.O. (2011) Phylogenies with corroboration assessment. *Zootaxa*, 2946, 52–56.

Farris, J. (2002) RASA attributes highly significant structure to randomized data. *Cladistics*, 18, 334–353.

Fann, K.T. (1970) *Peirce's Theory of Abduction*. Martinus Nijhoff, The Hague, 62 pp.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D. & Kluge, A.G. (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, 12, 99–124.

Farris, J.S., Kluge, A.G. & Carpenter, J.M. (2001) Popper and likelihood versus "Popper*." *Systematic Biology*, 50, 438–444.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17, 368–376.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39, 783–791.

Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, 664 pp.

Fetzer, J.H. (1993) *Philosophy of Science*. Paragon House, New York, 197 pp.

Fetzer, J.H. & Almeder, R.F. (1993) *Glossary of Epistemology/Philosophy of Science*. Paragon House, New York, 149 pp.

Fitzhugh, K. (2005a) Les bases philosophiques de l'inférence phylogénétique: une vue d'ensemble. *Biosystema*, 24, 83–105.

Fitzhugh, K. (2005b) The inferential basis of species hypotheses: the solution to defining the term 'species.' *Marine Ecology*, 26, 155–165.

Fitzhugh, K. (2006a) The abduction of phylogenetic hypotheses. *Zootaxa*, 1145, 1–110.

Fitzhugh, K. (2006b) The 'requirement of total evidence' and its role in phylogenetic systematics. *Biology & Philosophy*, 21, 309–351.

Fitzhugh, K. (2006c) The philosophical basis of character coding for the inference of phylogenetic hypotheses. *Zoologica Scripta*, 35, 261–286.

Fitzhugh, K. (2008a) Fact, theory, test and evolution. *Zoologica Scripta*, 37, 109–113.

Fitzhugh, K. (2008b) Abductive inference: implications for ‘Linnean’ and ‘phylogenetic’ approaches for representing biological systematization. *Evolutionary Biology*, 35, 52–82.

Fitzhugh, K. (2008c) Clarifying the role of character loss in phylogenetic inference. *Zoologica Scripta*, 37, 561–569.

Fitzhugh, K. (2009) Species as explanatory hypotheses: refinements and implications. *Acta Biotheoretica*, 57, 201–248.

Fitzhugh, K. (2010a) Evidence for evolution versus evidence for intelligent design: parallel confusions. *Evolutionary Biology*, 37, 68–92.

Fitzhugh, K. (2010b) Revised systematics of *Fabricia oregonica* Banse, 1956 (Polychaeta: Sabellidae: Fabriciinae): an example of the need for a uninomial nomenclatural system. *Zootaxa*, 2647, 35–50.

Franz, N.M. (2005) Outline of an explanatory account of cladistic practice. *Biology & Philosophy*, 20, 489–515.

Gaffney, E.S. (1979) An introduction to the logic of phylogeny reconstruction. *In*: Cracraft, J. & Eldredge, N. (Eds), *Phylogenetic Analysis and Paleontology*. Columbia University Press, New York, pp. 79–111.

Gatesy, J. (1998) Molecular evidence for the phylogenetic affinities of Cetacea. *In*: Thewissen, J.G.M. (Ed.), *The Emergence of Whales*. Plenum, New York, pp. 63–112.

Giribet, G., Distel, D.L., Polz, M., Sterrer, W. & Wheeler, W.C. (2000) Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. *Systematic Biology*, 49, 539–562.

Godfrey-Smith, P. (2003) *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press, Chicago, 272 pp.

Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J. & Szumik, C.A. (2003) Improvements to resampling measures of group support. *Cladistics*, 19, 324–332.

Grandcolas, P., Deleporte, P. & Desutter-Grandcolas, L. (1997) Testing evolutionary processes with phylogenetic patterns: test power and test limitations. *In*: Grandcolas, P. (Ed.), *The Origin of Biodiversity in Insects: Phylogenetic Tests of Evolutionary Scenarios. Mémoires du Muséum National d=Histoire Naturelle*, 173, 53–71.

Grant, T. & Kluge, A.G. (2007) Ratio of explanatory power (REP): a new measure of group support. *Molecular Phylogenetics and Evolution*, 44, 483–487.

Grant, T. & Kluge, A.G. (2008) Clade support measures and their adequacy. *Cladistics*, 24, 1051–1064.

Haack, S. & Kolenda, K. (1977) Two fallibilists in search of the truth. *Proceedings of the Aristotelian Society, Supplement*, 51, 63–104.

Haber, M.H. (2005) On probability and systematics: possibility, probability, and phylogenetic inference. *Systematic Biology*, 54, 831–841.

Haber, M. (2011). Phylogenetic inference. *In*: Tucker, A. (Ed.), *A Companion to the Philosophy of History and Historiography*. Wiley-Blackwell, Malden, pp. 231–242.

Hacking, I. (2001) *An Introduction to Probability and Inductive Logic*. Cambridge University Press, New York, 302 pp.

Halanych, K.M. (2004) The new view of animal phylogeny. *Annual Review of Ecology and Evolutionary Systematics*, 35, 229–256.

Hanson, N.R. (1958) *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, New York, 241 pp.

Hausman, D.M. (1998) *Causal Asymmetries*. Cambridge University Press, New York, 300 pp.

Helfenbein, G.K. & DeSalle, R. (2005) Falsifications and corroborations: Karl Popper’s influence on systematics. *Molecular Phylogenetics and Evolution*, 35, 271–280.

Hempel, C.G. (1962) Deductive nomological vs. statistical explanation. *In*: Feigl, H. & Maxwell, G. (Eds), *Minnesota Studies in the Philosophy of Science, Volume. 3*. University of Minnesota Press, Minneapolis, pp. 98–169.

Hempel, C.G. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, New York, 505 pp.

Hempel, C.G. (1966) Recent problems of induction. *In*: Colodny, R.G. (Ed.), *Mind and Cosmos*. University of Pittsburgh Press, Pittsburgh, pp. 112–134.

Hempel, C.G. (2001) *The Philosophy of Carl G. Hempel: Studies in Science, Explanation, and Rationality. In*: Fetzer, J.H. (Ed.). Oxford University Press, New York, 423 pp.

Hennig, W. (1966) *Phylogenetic Systematics*. University of Illinois Press, Urbana, 263 pp.

Holmes, S. (2003) Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18, 241–255.

Hovenkamp, P. (2012) Syncretism and corroboration. *Cladistics*, 28, 115–116.

Huelsenbeck, J.P. & Crandall, K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28, 437–466.

Huelsenbeck, J.P. & Ronquist, F. (2001) MrBayes: bayesian inference of phylogeny. *Bioinformatics*, 17, 754–755.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. & Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294, 2310–2314.

Jenner, R.A. (2003) Unleashing the force of cladistics? Metazoan phylogenetics and hypothesis testing. *Integrative and Comparative Biology*, 43, 207–218.

Josephson, J.R. & Josephson, S.G. (Eds) (1994) *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press, New York, 306 pp.

Kluge, A.G. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology*, 38, 7–25.

Kluge, A.G. (1997a) Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. *Zoologica Scripta*, 26, 349–360.

Kluge, A.G. (1997b) Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics*, 13, 81–96.

Kluge, A.G. (1998) Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics*, 14, 151–158.

Kluge, A.G. (1999) The science of phylogenetic systematics: explanation, prediction, and test. *Cladistics*, 15, 429–436.

Kluge, A.G. (2001) Philosophical conjectures and their refutation. *Systematic Biology*, 50, 322–330.

Kluge, A.G. (2004) On total evidence: for the record. *Cladistics*, 20, 205–207.

Kluge, A.G. & Wolf, A.J. (1993) Cladistics: what=s in a word? *Cladistics*, 9, 183–199.

Leonelli, S. (2009) Understanding in biology: the impure nature of biological knowledge. *In*: de Regt, H., Leonelli, S. & Eigner, K. (Eds), *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press, Pittsburgh, pp. 189–209.

Lecointre, G. & Deleporte, P. (2005) Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, 34, 101–117.

Lee, M.S.Y. & Camens, A.B. (2009) Strong morphological support for the molecular evolutionary tree of placental mammals. *Journal of Evolutionary Biology*, 22, 2243–2257.

Levasseur, C. & Lapointe, F.-J. (2001) War and peace in phylogenetics: a rejoinder to total evidence and consensus. *Systematic Biology*, 50, 881–891.

Lienau, E.K. & DeSalle, R. (2010) Is the microbial tree of life verificationist? *Cladistics*, 26, 195–201.

Lipton, P. (2004) *Inference to the Best Explanation*. Routledge, New York, 219 pp.

Lipton, P. (2005) Testing hypotheses: prediction and prejudice. *Science*, 307, 219–221.

Longhorn, S.J., Pohl, H.W. & Vogler, A.P. (2010) Ribosomal protein genes of holometabolan insects reject the Halteria, instead revealing a close affinity of Strepsiptera with Coleoptera. *Molecular Phylogenetics and Evolution*, 55, 846–859.

Longino, H.E. (1979) Evidence and hypothesis: an analysis of evidential relations. *Philosophy of Science*, 46, 35–56.

Magnani, L. (2001) *Abduction, Reason, and Science: Processes of Discovery and Explanation*. Kluwer Academic, New York, 205 pp.

Maher, P. (1988) Prediction, accommodation, and the logic of discovery. *In*: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pp. 273–285.

Mahner, M. & Bunge, M. (1997) *Foundations of Biophilosophy*. Springer, New York, 423 pp.

Mayo, D.G. (1996) *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, Chicago, 493 pp.

Mayr, E. (1961) Cause and effect in biology. *Science*, 131, 1501–1506.

Mayr, E. (1982) *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Harvard University Press, Cambridge, 974 pp.

Mayr, E. (1993) Proximate and ultimate causation. *Biology & Philosophy*, 8, 95–98.

Mayr, E. (1994) Response to John Beatty. *Biology & Philosophy*, 9, 359–371.

Meredith, R.W., Janecka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L.L., Stadler, T., Rabosky, D.L., Honeycutt, R.L., Flynn, J.J., Ingram, C.M., Steiner, C., Williams, T.L., Robinson, T.J., Burk-Herrick, A., Westerman, M., Ayoub, N.A., Springer, M.S. & Murphy, W.J. (2011) Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science*, DOI*:* 10.1126/science.1211028.

Miller, J.A. (2003) Assessing progress in systematics with continuous jackknife function analysis. *Systematic Biology*, 52, 55–65.

Miyamoto, M.M. & Fitch, W.M. (1995) Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, 44, 64–76.

Montgelard, C., Douzery, E.J.P. & Michaux, J. (2007) Classification and molecular phylogeny. *In*: Miller, D.L. (Ed.), *Reproductive Biology and Phylogeny of Cetacea*. Science Publishers, Enfield, New Hampshire, pp. 95–125.

Mooi, R.D. & Gill, A.C. (2010) Phylogenies without synapomorphies—a crisis in fish systematics: time to show some character. *Zootaxa*, 2450, 26–40.

Nickles, T. (1980) Introductory essay: scientific discovery and the future of philosophy of science. *In*: Nickles, T. (Ed.), *Scientific Discovery, Logic and Rationality*. D. Reidel Publishing Company, Dordrecht, pp. 1–59.

Nixon, K.C. & Carpenter, J.M. (1996) On simultaneous analysis. *Cladistics*, 12, 221–241.

Nogueira, J.M.D.M., Fitzhugh, K. & Rossi, M.C.S. (2010) A new genus and new species of fan worms (Polychaeta: Sabellidae) from Atlantic and Pacific Oceans—the formal treatment of taxon names as explanatory hypotheses. *Zootaxa*, 2603, 1–52.

Norton, J.D. (2003) A material theory of induction. *Philosophy of Science*, 70, 647–670.

O'Leary, M.A. & Gatesy, J. (2008) Impact of increased character sampling on the phylogeny of Cetartiodactyla (Mammalia): combined analysis including fossils. *Cladistics*, 24, 397–442.

Peirce, C.S. (1878) Illustrations of the logic of science. Sixth paper.—Deduction, induction, and hypothesis. *Popular Science Monthly*, 13, 470–482.

Peirce, C.S. (1901) Reasoning. *In*: Baldwin, J.M. (Ed.), *Dictionary of Philosophy and Psychology, Including many of the Principal Conceptions of Ethics, Logic, Aesthetics, Philosophy of Religion, Mental Pathology, Anthropology, Biology, Neurology, Physiology, Physical Science, and Education and giving a Terminology in English, French, German, and Italian. Vol. II*. The Macmillan Company, New York, pp. 426–428.

Peirce, C.S. (1931) *Collected Papers of Charles Sanders Peirce, Volume 1, Principles of Philosophy. In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 393 pp.

Peirce, C.S. (1932) *Collected Papers of Charles Sanders Peirce, Volume 2, Elements of Logic*. *In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 535 pp.

Peirce, C.S. (1933a) *Collected Papers of Charles Sanders Peirce, Volume 3, Exact Logic*. *In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 433 pp.

Peirce, C.S. (1933b) *Collected Papers of Charles Sanders Peirce, Volume 4, The Simplest Mathematics*. *In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 601 pp.

Peirce, C.S. (1934) *Collected Papers of Charles Sanders Peirce, Volume 5, Pragmatism and Pragmaticism*. *In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 455 pp.

Peirce, C.S. (1935) *Collected Papers of Charles Sanders Peirce, Volume 6, Scientific Metaphysics*. *In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 462 pp.

Peirce, C.S. (1958a) *Collected Papers of Charles Sanders Peirce, Volume 7, Science and Philosophy. In*: Hartshorne, C., Weiss, P. & Burks, A. (Eds). Harvard University Press, Cambridge, 415 pp.

Peirce, C.S. (1958b). *Collected Papers of Charles Sanders Peirce, Volume 8, Correspondence and Bibliography. In*: Burks, A. (Ed.). Harvard University Press, Cambridge, 352 pp.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G. & Baurain, D. (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology*, 9, e1000602. doi: 10.1371/journal.pbio.1000602.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D.J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. & Manuel, M. (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, 19, 706–712.

Philippe, H. & Telford, M. (2006) Large-scale sequencing and the new animal phylogeny. *Trends in Ecology and Evolution*, 21, 614–620.

Popper, K.R. (1959) *The Logic of Scientific Discovery.* Basic Books, Inc., New York, 480 pp.

Popper, K.R. (1962) *Conjectures and Refutations: The Growth of Scientific Knowledge*. Basic Books, Publishers, New York, 412 pp.

Popper, K. (1966) *The Open Society and Its Enemies. Volume II. The High Tide of Prophecy: Hegel, Marx, and the Aftermath*. Princeton University Press, Princeton, 420 pp.

Popper, K.R. (1983) *Objective Knowledge: An Evolutionary Approach*. Oxford University Press, New York, 395 pp.

Popper, K.R. (1992) *Realism and the Aim of Science.* Routledge, New York, 420 pp.

Popper, K. (1994) *The Poverty of Historicism*. Routledge, New York, 166 pp.

Prasad, A.B., Allard, M.W., NISC Comparative Sequencing Program & Green, E.D. (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Molecular Biology and Evolution*, 25, 1795–1808.

Psillos, S. (2002) Simply the best: a case for abduction. *In*: Kakas, A.C. & Sadri, F. (Eds), *Computational Logic: Logic Programming and Beyond*. Springer, New York, pp. 605–625.

Psillos, S. (2007) *Philosophy of Science A–Z.* University Press, Edinburgh, 280 pp.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W. & Cunningham, C.W. (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, 463, 1079–1083.

de Regt, H.W. & Dieks, D. (2005) A contextual approach to scientific understanding. *Synthese*, 144, 137–170.

de Regt, H.W., Leonelli, S. & Eigner, K. (2009) Focusing on scientific understanding. *In*: de Regt, H., Leonelli, S. & Eigner, K. (Eds), *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press, Pittsburgh, pp. 1–17.

Reilly, F.E. (1970) *Charles Peirce=s Theory of Scientific Method.* Fordham University Press, New York, 200 pp.

Rescher, N. (1970) *Scientific Explanation*. The Free Press, New York, 242 pp.

Rieppel, O. (1988) *Fundamentals of Comparative Biology.* Birkhäuser Verlag, Boston, 202 pp.

Rieppel, O. (2003) Semaphoronts, cladograms and the roots of total evidence. *Biological Journal of the Linnean Society,* 80, 167–186.

Rindal, E. & Brower, A.V.Z. (2011) Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data. *Cladistics*, 27, 131?334.

Ronquist, F., van der Mark, P. & Huelsenbeck, J.P. (2009) Bayesian phylogenetic analysis using MrBayes. *In*: Lemey, P., Salemi, M. & Vandamme, A.-M. Eds), *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, pp. 210?266.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H. & Telford, M.J. (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B*, 278, 298–306.

Salmon, W.C. (1967) *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh, 157 pp.

Salmon, W.C. (1984a) *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 305 pp.

Salmon, W.C. (1984b) *Logic.* Prentice-Hall, Inc., Englewood Cliffs, 180 pp.

Salmon, W.C. (1989) Four decades of scientific explanation. *In*: Kitcher, P. & Salmon, W.C. (Eds.), *Scientific Explanation. Minnesota Studies in the Philosophy of Science, Volume XIII*. University of Minnesota Press, Minneapolis, pp. 3–219.

Salmon, W.C. (1998) *Causality and Explanation*. Oxford University Press, New York, 434 pp.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.J., Hadrys, H., Dellaporta, S.L., Kolokotronis, S.-O. & DeSalle, R. (2009) Concatenated analysis sheds light on early metazoan evolution and fuels a modern ''urmetazoon'' hypothesis. *PLoS Biology*, 7, e1000020. doi: 10.1371/journal.pbio.1000020.

Schmidt, H.A. (2009) Testing tree topologies. *In*: Lemey, P., Salemi, M. & Vandamme, A.-M. Eds), *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, pp. 381?404.

Schuh, R.T. (2000) *Biological Systematics: Principles and Applications*. Cornell University Press, Ithaca, 236 pp.

Schuh, R.T. & Brower, A.V.Z. (2009) *Biological Systematics: Principles and Applications*. Second edition. Cornell University Press, Ithaca, 311 pp.

Schurz, G. (2008) Patterns of abduction. *Synthese*, 164, 201–234.

Siddall, M.E. & Kluge, A.G. (1997) Probabilism and phylogenetic inference. *Cladistics*, 13, 313–336.

Sober, E. (1975) *Simplicity*. Oxford University Press, New York, 189 pp.

Sober, E. (1988) *Reconstructing the Past: Parsimony, Evolution, and Inference*. MIT Press, Cambridge, 265 pp.

Sober, E. (2002) Reconstructing the character states of ancestors: a likelihood perspective on cladistic parsimony. *Monist*, 85, 156–176.

Sober, E. (2008) *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press, New York, 392 pp.

Sober, E. & Steel, M. (2002) Testing the hypothesis of common ancestry. *Journal of Theoretical Biology*, 218, 395–408.

Soltis, P.S. & Soltis, D.E. (2003) Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 18, 256–267.

Spaulding, M., O'Leary, M.A. & Gatesy, J. (2009) Relationships of Cetacea (Artiodactyla) among mammals: increased taxon sampling alters interpretations of key fossils and character evolution. *PLoS ONE*, 4, e7062. doi: 10.1371/journal.pone.0007062.

Springer, M.S., Burk-Herrick, A., Meredith, R., Eizirik, E., Teeling, E., O'Brien, S.J. & Murphy, W.J. (2007) The adequacy of morphology for reconstructing the early history of placental mammals. *Systematic Biology*, 56, 673–84.

Strahler, A.N. (1992) *Understanding Science: An Introduction to Concepts and Issues*. Prometheus Books, Buffalo, 409 pp.

Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. (1996) Phylogenetic inference. *In*: Hillis, D.M., Moritz, C. & Mable, B.K. (Eds), *Molecular Systematics*. Sinauer Associates, Sunderland, pp. 407–514.

Thagard, P. (1988) *Computational Philosophy of Science*. The MIT Press, Cambridge, 240 pp.

Tinbergen, N. (1963) On aims and methods in ethology. *Zeitschrift fur Tierpsychologie*, 20, 410–433.

Tucker, A. (2011) Historical science, over- and underdetermined: A study of Darwin's inference of origins. *British Journal for the Philosophy of Science*, 62, 805–829.

Turjak, M. & Trontelj, P. (2012) A method for measuring support for synapomorphy using character state distributions on phylogenetic trees. *Cladistics*, doi: 10.1111/j.1096-0031.2012.00403.x

Uhen, M.D. (2010) The origin(s) of whales. *Annual Review of Earth and Planetary Sciences*, 38, 189–219.

Van Fraassen, B.C. (1990) *The Scientific Image*. Clarendon Press, New York, 235 pp.

Vila, R., Bell, C.D., Macniven, R., Goldman-Huertas, B., Ree, R.H., Marshall, C.R., Bálint, Z., Johnson, K., Benyamini, D. & Pierce, N.E. (2011) Phylogeny and palaeoecology of *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the New World. *Proceedings of the Royal Society B*, doi: 10.1098/rspb.2010.2213.

Vogt, L. (2008) The unfalsifiability of cladograms and its consequences. *Cladistics*, 24, 62–73.

Von Dohlen, C.D., Rowe, C.A. & Heie, O.E. (2006) A test of morphological hypotheses for tribal and subtribal relationships of Aphidinae (Insecta: Hemiptera: Aphididae) using DNA sequences. *Molecular Phylogenetics and Evolution*, 38, 316–329.

Wägele, J.-W. (2005) *Foundations of Phylogenetic Systematics*. Verlag Dr. Friedrich Pfeil, München, 365 pp.

Walton, D. (2004) *Abductive Reasoning*. The University of Alabama Press, Tuscaloosa, 303 pp.

Wenzel, J.W. (1997) When is a phylogenetic test good enough? *In*: Grandcolas, P. (Ed.), *The Origin of Biodiversity in Insects: Phylogenetic Tests of Evolutionary Scenarios*. *Mémoires du Muséum National d=Histoire Naturelle*, 173, 31–45

Wheeler, Q.D. (2004) Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society B*, 359, 571–583.

Wheeler, Q.D. (2010) Do we need to describe, name, and classify all species? *In*: Williams, D.M. & Knapp, S. (Eds), *Beyond Cladistics: The Branching of a Paradigm*. University of California Press, Berkeley, pp. 67–75.

Wiens, J.J. (2009) Paleontology, genomics, and combined-data phylogenetics: can molecular data improve phylogeny estimation for fossil taxa? *Systematic Biology*, 58, 87–99.

Wiley, E.O. (1975) Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. *Systematic Zoology*, 24, 233–243.

Wiley, E.O. (1981) *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons, New York, 439 pp.

Wiley, E.O. & Lieberman, B.S. (2011) *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. Wiley-Blackwell, Hoboken, New Jersey, 406 pp.

Xiong, Y., Brandley, M.C., Xu, S., Zhou, K. & Yang, G. (2009) Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. *BMC Evolutionary Biology*, 9: 20 doi: 10.1186/1471-2148-9-20.