The Future of Systematics: Tree-Thinking Without the Tree
(Forthcoming in Philosophy of Science Dec 2012)

Joel D. Velasco
California Institute of Technology

Abstract: Phylogenetic trees are meant to represent the genealogical history of life and apparently derive their justification from the existence of the tree of life and the fact that evolutionary processes are tree-like. However, there are a number of problems for these assumptions. Here it is argued that once we understand the important role that phylogenetic trees play as models which contain idealizations, we can accept these criticisms and deny the reality of the tree while justifying the continued use of trees in phylogenetic theory and preserving nearly all of what defenders of trees have called "the importance of tree-thinking."

1. Introduction to Phylogenetic Trees

Evolutionary biology is a historical science. Phylogenetic trees represent the history of lineages and lineage splits through time. Constructing trees is the starting point for nearly every study in evolutionary biology today. Trees provide the historical information that is used as the essential background framing for explanations, reconstructing events in the past, and understanding trends through time. It is not surprising that a common paraphrase of Dobzhansky's famous dictum about evolution is that "Nothing makes sense except in the light of phylogeny."

The ability to properly read and understand trees is what Robert O'Hara (1988) memorably termed "tree-thinking." As he later put it, "Just as beginning students in geography need to be taught how to read maps, so beginning students in biology should be taught how to read trees and to understand what trees communicate" (O'Hara 1997). This is especially important given our common tendencies toward group (class/kind) thinking and ladder thinking as well as the ubiquity of misleading diagrams of evolutionary history (see Baum and Smith 2012).

It was Darwin in *The Origin* (1859) who first used trees to emphasize the importance and explanatory power of common ancestry. For example, the phenomenon of homology was well-known before *The Origin*, but it was unclear why, for example, bats, birds, crocs, frogs, humans, and turtles would have the same (homologous) bone structure in their limbs - a humerus attached to a radius and ulna, attached to carpals, metacarpals, and phalanges, etc. - despite the obviously different functions that these limbs performed. We can see how Darwin explained homology by examining figure 1.

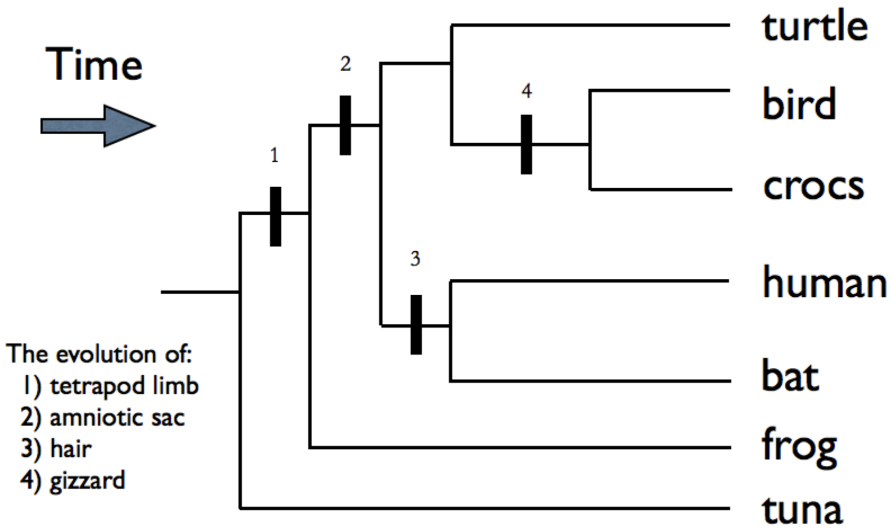

Figure 1: A phylogenetic tree of some chordate groups with the origin of a few key traits marked.

The tree represents the passage of time from left to right with the tips of the tree representing extant groups. The tetrapods: amphibians, reptiles, mammals, and birds all have four limbs with a common limb structure, whereas all these except the amphibians have a water-tight egg. Darwin's explanation of why these groups have the common tetrapod limb structure is that it evolved once in the evolutionary past in one particular group and has been transmitted down the braches to all the descendants of that group. Similarly, the water-tight egg evolved one time (on the amniote branch after it split from the amphibian branch) and was passed down.

We can already see from this example the key features of the information that trees convey. We assume that traits are passed downstream on a branch and we assume that they are not passed on to any non-downstream branches. Lineage separation represents isolation.

Moving from the fact that homologous traits could possibly be explained by common ancestry to the truth of the nested hierarchy and the branching structure of the tree is not trivial. Perhaps other structures might also explain homologous traits. On the special creation model, the tetrapod limb is simply a good design structure. It is not that hard to see why a creator might reuse a design multiple times if it is simply a good design. On a ladder or great chain of being picture, we can imagine that there is some kind of hierarchy of traits, the lower creatures having the basic ones while the more advanced creatures have more and more of the advanced traits until finally, humans are the only earthly species with the highest traits (like intelligence).

The easiest way to see the problem with the ladder is to think about multiple traits. On the ladder picture, a single trait can divide species into two groups – those that have the trait in question and those that don't (hair divides mammals from everything else). We could think of a linear ordering of all of these groups with those that don't have the trait as more ancestral than those that do. But when we look at a second trait (say feathers), it will group different organisms into the 'higher' category (birds vs. non-birds). So who is higher on the evolutionary ladder, mammals or birds? The obvious answer is that there are at least two ladders. And if we look at traits only found in turtles (like the shell) it is obvious that neither birds nor mammals came from turtles, rather, there is a third line leading to turtles and so on. This reasoning leads to a branching explanation of trait distribution, not a ladder explanation.

Of course naturalists were long aware of the distribution of traits and Lamarck's ladder can't be used to explain homology. Notice in Figure 1 that all of the groups that have a water-tight egg also have four limbs. One group is nested inside the other. On the other hand, there are no organisms with gizzards that also have hair. Those groups are completely disjoint. If every trait evolved only once, we could map the traits onto a tree and form a perfect group-within-group structure that would be the phylogenetic tree of life. It would directly give us our trait-based classification of groups in a nested hierarchy. Linnaeus already had the hierarchical part of the tree structure, but he had no explanation for this. The branching of genealogy explains the hierarchy of classification and nothing else does. This is why Darwin considered the evidence from hierarchical classification to be the strongest evidence for his theory. Nesting is logically *consistent* with special creation, but it is *explained* by a tree-like common ancestry.

This centrality of tree-like ancestry manifests itself in the ubiquitous use of phylogenetic trees across different problem areas. Treatments of many problems in systematics require the use of phylogenetic trees. Studying adaptation, biogeography, or comparative biology all require the use of trees. More general framework questions such as to what extent there is a molecular clock can only be answered once a phylogeny is in place. Testing trends, such as whether mammals have increased in size over time, requires the tree. Basically, for any questions that can only be answered in the light of knowing the history of the group, the phylogenetic tree is the essential background information needed to get started answering such a question. More generally, historical information is required for systematic studies. What I will call the *realist* defense of trees says why this is so: trees are essential precisely because they represent the real history.

2. Worries for the Tree of Life

The centrality of phylogenetic trees extends to the construction of the universal tree of life. But there are serious problems with that tree. Many biologists have claimed that there is no such thing and that tree-thinking can often constrain us and blind us to the truth. For detailed arguments, see the special issues on the tree of life in *Biology and Philosophy* (2010) and *Biology Direct* (2011). For brief overviews, see O'Malley et al. (2010) and Velasco (forthcoming). Here, I will mention only some of the most serious issues, which have an important feature in common.

A natural way to think about individual phylogenetic trees is that they are subtrees of the big, universal tree of life, which represents how all species are connected (Velasco 2010). Some worries are specific to the universal tree of life. For example, it is not clear that there is a single, universal common ancestor at the root of the tree. But here, it must be pointed out that this is a special problem for the *universal* tree and for deep time questions such as the origin of cells. Even if in the end, we must give up the belief in the universal tree, this has no direct consequences for the existence of trees at smaller scales. If we are looking at a tree of Darwin's finches, then as long as they do in fact have a uniquely shared origin, there is no parallel problem with there being a single rooted tree in this domain which serves exactly the same explanatory purposes and plays exactly the same inferential role as before. There is no need to overgeneralize to all life in order to justify the use of particular trees.

Perhaps the most dramatic problem with the tree is endosymbiosis. In endosymbiosis, one organism comes to live inside another and eventually becomes an obligate symbiote. Over time, they are so tightly interconnected, that it is no longer appropriate to think of the situation as one organism living inside another, but rather, as one integrated organism. Different understandings of endosymbiosis will be more or less strict about what counts as a new entity, but even on the most conservative understandings, the origins of mitochondria and chloroplasts will count as two separate lineages merging into one and the origin of the eukaryotic cell itself might be such an event (O'Malley 2010). Endosymbiosis could hardly be more important to the history of life on this planet. But endosymbiosis involves the combination and complete transformation of two lineages - sometimes as distantly related as is possible on Earth - and fuses them into a completely new lineage of a new kind of organism.

As mentioned earlier, the tree typically represents genealogical connections between species and thus separations of lineages represent the genetic isolation of different species. But we know that organisms of different species can and do hybridize. Mallet (2005) surveys a variety of studies of hybrids and concludes that at least 25% of plant and 10% of animal species form hybrids with other species in nature.

The most common way for different species to share genetic information is via lateral gene transfer where genetic information is physically transferred from one organism to another pre-existing organism. Lateral gene transfer (LGT) had been conclusively shown in the 1940s to happen in the lab and it was known to be responsible for the spread of antibiotic resistance in the 1960s (Sapp 2009). But just how widespread in nature LGT was, was not known. It is now known to be pervasive. Traits that are transferred from one lineage to another via LGT show that there is something wrong with our tree explanation of the distribution of many homologous traits. It isn't that there is some other explanation to explain their nested group-within-group structure. Rather, for a great many traits, they are not distributed in a group-within-group hierarchy in the first place. LGT undermines central systematic concepts like *phylogeny, lineage,* and *species* and O'Malley and Boucher (2005) consider LGT to be the major cause of an ongoing paradigm shift in microbiology beyond just its role in systematics.

Endosymbiosis between lineages, hybridization between species, and lateral gene transfer between organisms all show that there are possible routes by which changes on one lineage can have cross-stream, rather than just downstream, effects on other lineages. Crucially, this is inconsistent with a branching tree pattern. If we take seriously the realist idea that the tree of life represents the genealogical pattern of evolutionary history, then the "tree" will not be a phylogenetic tree but, rather, a complicated web or network of some sort. Even very simplified drawings will look something like figure 2.

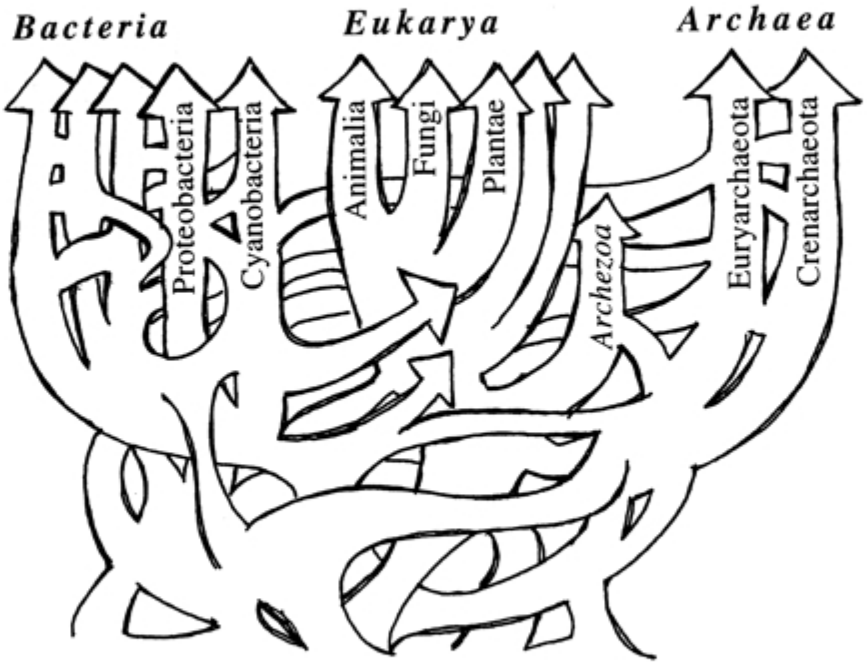

Figure 2: A web of life which represents the history of lineages when we take into account various lateral transfers of genetic information as well as vertical reproduction. Taken from Doolittle (1999).

The empirical evidence is clear; there are a great many non-tree-like processes which produce non-tree-like genealogical patterns in nature. The realist is forced to become a realist about networks rather than trees.

3. Modeling and Idealization

These attacks on the reality of the tree structure in nature are not easily dismissed. Defenders of trees will often attempt to minimize just how bad things are by arguing that endosymbiosis and hybridization are rare, or that lateral gene transfer happens only in prokaryotes (only in the vast majority of life on earth?!). Common responses from those who attack the use of trees is that these phenomena are much more widespread than commonly appreciated. While I agree that the extent of such phenomena is widely under-appreciated, such empirical questions can hardly matter to the realist who is trying to insist that evolutionary history is accurately represented by a single tree. It is not. Whether hybridization is rare or common, it is clearly not nonexistent and from the realist point of view, that is all that matters.

However, debates about the extent of these phenomena make perfect sense if we think about trees as models. Here the relevant question is not whether the history is a tree, but how good is a model that depicts history as a tree (Franklin-Hall 2010). We can call the *modeling* defense of trees the view that trees are useful because they are very good models. Models contain idealizations. Roughly, idealizations deliberately simplify or alter something complicated in order to better understand it or to better understand something else. Philosophers of science have pointed to a number of different ways that idealizations are used in scientific practice. Here I will follow Weisberg (2007) in distinguishing three kinds of idealizations. While Weisberg's particular taxonomy of idealizations is not essential to my project, it does provide a useful expository strategy.

First, there are examples of *minimalist idealizations*. Here, the model contains only the core causal factors that are relevant in the situation at hand. Obviously, there are many aspects of genealogy that a tree does not represent. If we have a tree representing the great ape species, we know that traits are actually inherited by individual organisms, not by species. But for at least some uses of trees, these details are simply ignored as irrelevant to the problem at hand. We can treat character traits as being transmitted by lineages along branches of the tree without any problems. This kind of idealization has been called *abstraction* (cf. Cartwright 1989) or *Aristotelian idealization* (Frigg and Hartmann 2006). It is easy to see how abstracting away from details that don't matter could help us better understand and explain some phenomenon like the particular distribution of traits.

While sometimes we can abstract away from irrelevant details, other times, we idealize away details that do make a difference. However, even in some of those cases we are justified in thinking that the effects are small for the purposes at hand and that, therefore, it is permissible to ignore them. For example, tree models assume that once a lineage splits from another, changes such as mutations on one lineage don't affect isolated lineages. Hybridization and lateral gene transfer show this assumption to be false. Sometimes we simply ignore these effects as real, but small. Inferences that use trees will assume no such effects, but as long as you are aware of these possibilities, you will reason that you are just making a fallible inference when you infer that, for example, a trait has evolved independently in two different lineages.

The systematics of the past is dominated by minimalist idealization. Those who built the first trees such as Darwin and Haeckel were well aware of hybridization. But often, they were attempting to capture broad-scale features of the history such as the connections between the different classes of vertebrates or the different orders of birds. The messy details of speciation were simply irrelevant for these purposes. In contrast, systematists of the late 20th century built trees at all scales and here the details of hybridization do matter. Yet often, these details were simply ignored.

We now know that simply ignoring lateral effects will not do. Here, the empirical studies on the extent of lateral gene transfer and hybridization really matter. In these cases, there is also a particular kind of problem that arises from initially just ignoring these lateral effects. If you don't bother looking for lateral events, you will build a tree that is capable of fitting the data and may not notice that anything has gone awry. Standard phylogenetic methods simply assume that the data is generated from a tree process and then find the best tree to fit the data. This is no way to detect whether a tree is the right kind of structure in the first place (Sober 2008). One simply cannot detect lateral events unless you specifically look for them. And there is no way of knowing ahead of time if you are working with a group where lateral events play an important causal and explanatory role.

Now that we know that lateral events can significantly affect our understanding of the history of particular groups, systematists today often attempt to directly estimate the incidence and effects of lateral events such as gene transfer or hybridization, as well as the effects from vertical transmission.

In these cases, a second type of idealization is often used. This is *Galilean idealization*, the introduction and later removal of deliberate distortions. The name is due to examples such as where Galileo assumes that the Earth is flat over a short distance or that a rolling ball is perfectly spherical (McMullin 1985). Sometimes the removal of the distortion will make a non-negligible difference, but this difference can then be estimated and we can still see that the original idealization was an essential part of the overall inference process.

Phylogenetic networks that contain lateral branches as well as horizontal branches are sometimes constructed by first building the tree that best fits the data. Then we look for anomalies in the data, by, for example, adding lateral branches one at a time and checking to see if the measure of fit (say a Parsimony or Likelihood score) is improved by a sufficient amount. If so, add the branch and then look for more. While this is an improvement over ignoring possible lateral effects, this procedure is analogous to doing a multiple regression by examining one variable at a time and there are well known methodological problems with all such greedy methods (Velasco and Sober 2010). Some kind of "find the best network all at once" approach is needed.

Most of the work that I described above still operates under the realist assumption that we are trying to find and represent the one true history of the group. This history is complicated, but in principle, it is what we want to know if we are to make further inferences. Much of the best work in systematics today in fact does quite a good job of inferring very complicated historical structures and patterns. But after reflecting on the fact that trees and networks are simply models of the history, new possibilities are opened up. It is here that I think the most exciting and ultimately the most fruitful work in systematics will be done. Here, we acknowledge that we are building models to use for inferences and that these models will contain idealizations which are not entirely accurate. But we accept that there is no need to show that these idealizations are harmless. In fact, they can be beneficial to our inferences and explanations.

The third kind of idealization is *multiple models idealization* which involves "the practice of building multiple related but incompatible models, each of which makes distinct claims about the nature and causal structure giving rise to a phenomenon" (Weisberg 2007, 645). Here, we do not expect a single, best model to be the final product. Rather, each model involves tradeoffs of better representing some aspects of reality at the cost of a worse representation of others. An obvious parallel is the case of maps. Some maps of New York City focus on the layout of the streets and ignore topographical features, others carefully note changes in altitude but would not help you if you were lost trying to find the library (Toulmin 1953, Kitcher 2001). This kind of tradeoff is clear in the decision to represent a group by a tree or a network. We can represent the history of lineages quite well with trees, but this will ignore connections such as those due to lateral gene transfer. But complicated networks that attempt to represent all such connections distort or even destroy the history of the lineages and cellular reproduction events. In general, we can imagine that different models of the phylogeny of the same groups idealize different aspects of this genealogy in different ways. For specific purposes, one model might be better than another. But in general, there is no reason to expect that a single model will fare better for answering a number of different questions about the groups in question just as there is no best map. For certain purposes such as phylogenomic inference (inferring the function of a gene from phylogenetic information), a phylogenetic tree that depicts three separate domains of life and some of the major clades within them really is a valuable tool in understanding diversity (Brown and Sjölander 2006). But not representing the endosymbiotic events which resulted in the origins of the plastids, mitochondria, and possibly all of the eukaryotes, would be tremendously misleading if wish to understand some of the most important innovations in the history of life on Earth.

In a review of plastid evolution, Archibald (2009) utilizes a phylogenetic tree as an inference tool to discover exactly where and when endosymbiotic events occurred in the past. This single figure contains both a network and a special tree contained within this network.

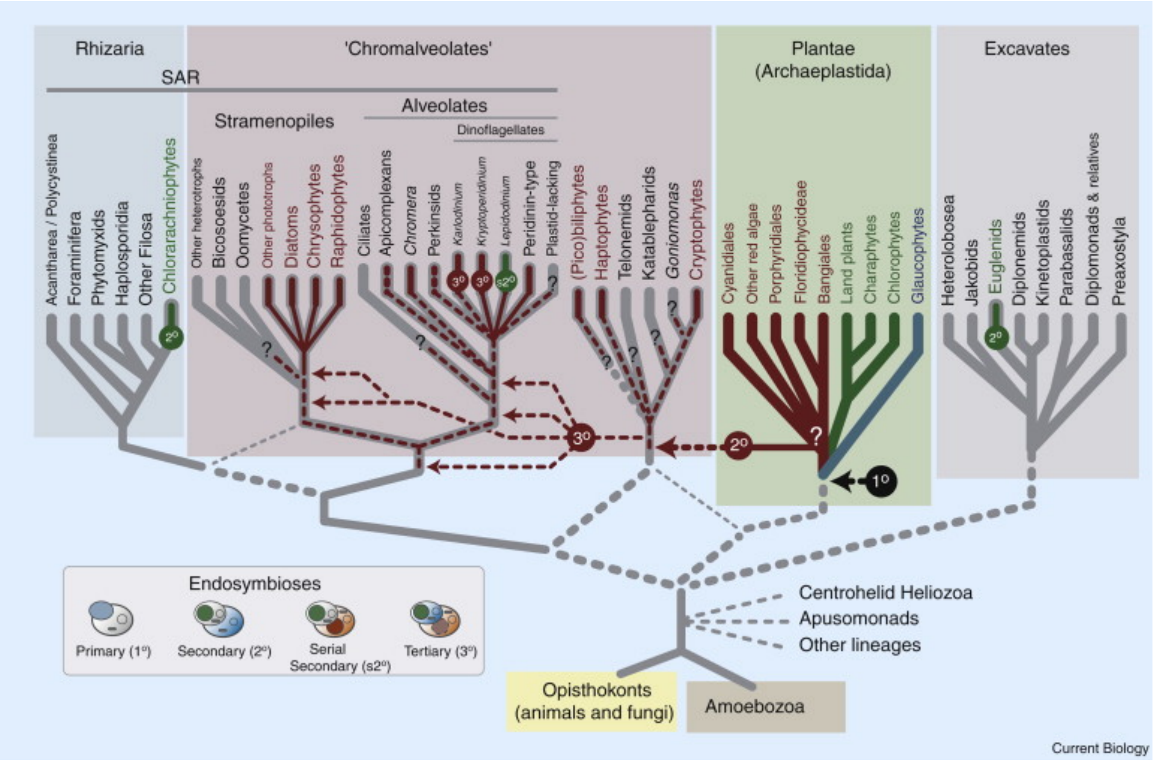

Figure 3: Hypothesis for the origin and spread of photosynthesis in eukaryotes (due to Archibald 2009).

In order to reconstruct something like figure 3, first, note that the gain of an internal symbiote and transition to a new kind of organism is simply <u>*one kind*</u> of change which gets passed along to descendants along the branch. Now use the current distribution of the symbiotes, together with information from other traits, to reconstruct an incomplete tree of the host cells that can itself be used to reconstruct the history of the past acquisition of symbiotes.

Archibald's reasoning exemplifies tree-thinking without the tree. We know that the history of these organisms has been heavily influenced by lateral transfer events such as repeated endosymbiosis followed by the massive sharing of genetic information between symbiote and host. But in order to understand this history, we must recognize that bits of the history such as the vertical descent of the host cells form a tree.

This also exemplifies multi-model inference because in order to understand one aspect of the history, we need to downplay or distort the history of other events. It would be a mistake to look at arbitrary gene trees to build the host tree in the first place since many of these will not share the host's phylogeny. To represent the history of such genes in the same model as the vertical inheritance of others would obscure the unique role of the host cell. Here the tree model and the network model idealize different things and there is a necessary tradeoff in the representations. Which model is to be preferred depends on what you are trying to do.

On the smaller scale, it is obvious that forcing the history of Darwin's finches or human populations into a tree structure ignores information about migration and introgression between populations, which is essential to understanding their history. For example, it has been argued that there has been so much admixture in human history, that the concept of race makes no biological sense (Templeton 1998; but see Andreasen 2004). On the other hand, trees of human populations are quite useful for understanding major ancestral migratory events (Soares et al. 2009). Obviously there are migrations between human populations. But whether the amount of migration is "sufficient" to be best represented as collapsing the distinction between different lineages depends on what aspects of the lineages we are focusing on. For different inference purposes, different representations of genealogical history focusing on different aspects of that history are appropriate. Looking at trees of populations at the subspecific level such as those of humans is common in the growing field of phylogeography (Avise 2000). This is a clear example of the superiority of the modeling defense of trees since if there were genetic isolation between these lineages they would represent different species.

One might think that we could get everything we want by simply "representing everything" in one giant model. This is false. Not only because it would be computationally intractable, which of course it would be, but also because models that focus on the right things are more explanatory than models which have no focus. An important part of the future of systematics will be recognizing how to take advantage of the fact that different models of the same phenomena can be independently fruitful (and fruitfully combined as well). Learning new and interesting things about past history will always be important. But learning when it is okay to ignore things we know can be equally important.

4. Conclusion

We have seen that phylogenetic trees are ubiquitous in biology. The justification for the use of trees has traditionally been that evolutionary processes are in fact tree-like. This justification is faulty. Attempting to interpret phylogenetic trees in a literal way leads to the view that these trees entail many falsities about evolutionary history. Attacks on the universal tree of life thus appear to be justified. The goal of this paper is to argue that these attacks are not in conflict with the continued and justified use of trees and tree-thinking in biology. The use of phylogenetic trees can often be completely justified even if they are not entirely accurate representations of the world. Instead, these trees are models which contain idealizations. These models are used to better understand the world. Sometimes, for some purposes, a tree model is inappropriate. But often, trees are entirely appropriate and perhaps even the best models we have.

Modeling and idealizations are widespread throughout the sciences. There is no particular reason to think that systematics should be any different. Evolutionary history is complicated. It is a sign of the advancement of the science of systematics that we not only take advantage of standard modeling practices from other disciplines, but that we understand that this is what it is that we are doing. It is true that belief in the existence of the tree of life as the big, universal, grand unifying, scale-free representation of all of the history of life should probably go away (if indeed biologists ever did believe there was such a tree). Whether we can still talk about the tree of life as some modified version of this idea is, I think, an open question. Whether the problems with the universal tree extend to smaller trees as well will depend on the particular details of the case in question. But whatever the outcome of these debates, phylogenetic trees and the importance of tree-thinking are here to stay. The future of tree-thinking is bright as long as we can recognize the importance of tree-thinking without the tree.

Acknowledgements: Thanks to audiences at the 2010 PSA meeting, the 2011 ISHPSSB meeting, Caltech, Ohio State, Gustavus Adolphus, and Texas Tech who heard parts of versions of this paper and to Matt Barker, Kenny Easwaran, Luke Glynn, Matt Haber, Matt Slater, and Elliott Sober who also provided comments on this paper.

References:

Andreasen, Robin O. 2004. “The Cladistic Race Concept: A Defense.” *Biology and Philosophy* 19:425-442.

Archibald, John M. 2009. “The Puzzle of Plastid Evolution.” *Current Biology* 19 (2): R81‑R88. http://dx.doi.org/10.1016/j.cub.2008.11.067

Avise, John C. 2000. *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, MA.

Baum, David A., and Stacy D. Smith. 2012. *Tree Thinking: An Introduction to Phylogenetic Biology*. Roberts and Company Publishers.

Brown, Duncan, and Kimmen Sjölander. 2006. “Functional Classification Using Phylogenomic Inference.” *PLoS Computational Biology* 2 (6): e77. PMID 16846248

Cartwright, Nancy. 1989. *Nature's Capacities and their Measurement*. Oxford: Oxford University Press.

Darwin, Charles. 1859. *The Origin of Species*. John Murray, London.

Doolittle, W. Ford. 1999. “Phylogenetic Classification and the Universal Tree.” *Science* 284 (5423): 2124-2128.

Franklin-Hall, Laura R. 2010. “Trashing Life's Tree.” *Biology and Philosophy* 25 (4): 689-709.

Frigg, Roman, and Stephan Hartmann. 2012. “Models in Science.” *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), ed. Edward N. Zalta. URL = <http://plato.stanford.edu/archives/spr2012/entries/models-science/>.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. Oxford University Press, Oxford

McMullin, Ernan. 1985. “Galilean Idealization.” *Studies in History and Philosophy of Science* 16:247-73.

Mallet, James. 2005. “Hybridization as an Invasion of the Genome.” *Trends in Ecology and*
Evolution 20:229–37.

O’Hara, Robert. 1988. “Homage to Clio, or, Toward an Historical Philosophy for Evolutionary Biology.” *Systematic Zoology* 37 (2): 142-155.

O’Hara, Robert. 1997. “Population Thinking and Tree Thinking in Systematics.” *Zoologica Scripta* 26 (4): 326-329.

O'Malley, Maureen A., William Martin, and John Dupré. 2010. “The Tree of Life: Introduction to an Evolutionary Debate.” *Biology and Philosophy* 25:441-453.

O'Malley, Maureen A. 2010. “The First Eukaryote Cell: An Unfinished History of Contestation.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 41 (3): 212–224.

O’Malley, Maureen A., and Yan Boucher. 2005. “Paradigm Change in Evolutionary Microbiology.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:183–208

Sapp. J. 2009. *The New Foundations of Evolution: On the Tree of Life*. Oxford University Press, Oxford.

Soares, Pedro, Luca Ermini, Noel Thomson, Maru Mormina, Teresa Rito, Arne Rohl, Antonio Salas, Stephen Oppenheimer, Vincent Macaulay, and Martin B. Richards. 2009. "Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock." *American Journal of Human Genetics* 84:740–759.

Sober, Elliott. 2008. *Evidence and Evolution*. Cambridge University Press.

Templeton, Alan R. 1998. "Human Races: A Genetic and Evolutionary Perspective." *American Anthropologist* 100 (3):632-650.

Toulmin, Stephen. 1953. The Philosophy of Science: An Introduction. Hutchinson's University Library, London.

Velasco, Joel D. 2010. "Species, Genes, and the Tree of Life." *British Journal for the Philosophy of Science* 61:599-619.

Velasco, Joel D. forthcoming. "The Tree of Life." In *The Cambridge Encyclopedia of Darwin and Evolutionary Thought*, ed. Michael Ruse. Cambridge University Press.

Velasco, Joel D., and Sober, Elliott. 2010. "Testing for Treeness: Lateral Gene Transfer, Phylogenetic Inference, and Model Selection." *Biology and Philosophy* 25:675-687.

Weisberg, Michael. 2007. "Three Kinds of Idealization." *Journal of Philosophy* 104 (12):639-659.