# Species, Genes, and the Tree of Life

**Joel Velasco**

**Stanford University**

**Abstract:** A common view is that species occupy a unique position on the Tree of Life. Evaluating this claim this requires an understanding of what the Tree of Life represents. The Tree represents history, but there are a least three biological levels that are often said to have genealogies: species, organisms, and genes. Here I focus on defending the plausibility of a gene-based account of the Tree. This leads to an account of species that are determined by gene genealogies. On this view, an exclusive group is a group of organisms that forms a clade for a higher proportion of the genome than any conflicting clade. Taxa occupy a unique position in what can be called the 'primary concordance tree'. But each gene has its own historical 'tree of life'. I conclude by arguing that both organismal pedigrees with their corresponding Tree as well as gene genealogies and their trees are objectively real and play important, but different, roles in biological practice.



## 1 Introduction

Biological systematics is often said to be in the business of recovering the Tree of Life. The Tree is supposed to be the great record of the evolutionary history of all life and it represents how all life on earth is descended from a common ancestor. One understanding of the Tree is that it is a record of the history of species and

speciations (Wiley [1981]; Cracraft and Donoghue [2004]; Dawkins [2004]; Hodkinson and Parnell [2006]). However, there are numerous different species concepts in the literature (25 are listed in Wilkins [2009]), and different species concepts lead to different histories of speciation and so on this understanding of the Tree, to multiple, incompatible, Trees of Life. Therefore, it might appear that there can be no single, objective Tree (LaPorte [2005]). However, the species problem has no immediate bearing on the objectivity of the Tree as the Tree of Life is metaphysically prior to, and independent of, any particular species concept that we choose (Velasco [2008]). For example, many systematists doubt the existence of species in Bacteria (Mayr [2001]; Franklin [2007]) yet there is no doubt that bacteria belong on the Tree of Life.

If the Tree of Life does not depend on our classification system, then what does it depend on? The answer is that it depends on evolutionary history – in particular, genealogical history. But this does not settle the meaning, for there are multiple ways of understanding genealogy. We have seen one way – the genealogy of species. However, there are at least two other levels of biological organization that have genealogical histories – organisms and genes.

One way to understand how there could be an objective genealogy for bacteria even if they do not form species is to recognize that bacteria are organisms with particular genealogies at that level of organization. As organisms, they are genealogically connected to all other life and to each other and it might be these organismal connections that the Tree represents. We could think of the entire Tree in this way. This is what Doolittle and Bapteste ([2007]) call 'the tree

of cells'. On some views of species, this conclusion contradicts the view that the Tree of Life shows how species are connected. However, a better option is to hold on to the view that the Tree of Life represents the history of species, and require that we use a species concept where the genealogy of a species is determined by the genealogical histories of the organisms that make up that species. This is the strategy used by defenders of some versions of the Phylogenetic Species Concept. On this view, species are units of phylogeny, and so must be groups of organisms united by a shared history (Mishler and Donoghue [1982]; Baum and Donoghue [1995]). With the right definition of species, the Tree of Life can represent the history of species and of organisms at the same time. In this way, we have a basis for phylogenetic classification. In such a system, only clades can be taxa. Since which groups are clades depends on the Tree, our classification depends on the Tree, not the other way around.

It is important to note that the question of what the Tree of Life is matters for all aspects of systematics and not just for classification. A correct phylogenetic tree is just a subtree of the full Tree of Life and so any area that depends on phylogenies—which is a huge portion of modern biology—depends on the Tree even if we decide that taxa need not be clades or that species need not be phylogenetic units.

## 2 The organism centered view of the Tree

It is possible to spell out a genealogical species concept according to which the genealogy of a species is completely determined by the genealogy of the

organisms in the species. Here each taxon will have a unique place on the Tree of Life (and therefore a unique genealogy) as long as each taxon forms an exclusive group of organisms. This means, roughly, that everything in the group is more closely related to the rest of the group than to anything outside it (de Queiroz and Donoghue [1990]; Velasco [2008], [2009]).

Such a species concept would ignore all other considerations such as the character states of particular organisms. This 'reductive'[1] understanding of species genealogy gives us a picture of what taxa are, sets the metaphysical stage that allows us to have a concept of the genealogy of taxa, and gives us a picture of one possible meaning of 'the Tree of Life' (Velasco [2009]).

Although this view leaves us with a questionable metaphor since the Tree of Life is no longer strictly a tree in the mathematical sense of each node having exactly one parent, it does solve many potential problems of the species-genealogy view of the Tree. For example, many traditional species concepts do not seem to apply across the whole of the Tree. In addition, many traditionally recognized species seem to have multiple histories or inconsistent phylogenetic connections because these groups are paraphyletic or even polyphyletic. Both problems are apparently solved by thinking at the organismal level.

But in fact, organisms are not the panacea that this view makes them seem to be. Though of broader scope than 'species', it is not obvious that the concept of 'organism' applies everywhere that we could do phylogenetic studies. For example, a famous phylogenetic study of varying strains of HIV was used by the CDC to prove that a Florida dentist was passing the virus to his patients (Palca

[1992]; Smith and Waterman [1992]). However, if organisms must be able to reproduce relatively independently of other organisms, then viruses will not be organisms and there is no organismal history here that permits one to make sense of the phylogeny. While it is by no means clear what sufficient and necessary conditions there might be for being an organism, many of the proposed definitions will leave viruses out (Wilson [2005]).

More general concerns about what organisms are threaten to undermine their usefulness for phylogenetics. For example, we could ask whether or not a human organism includes its gut flora (the trillions of microorganisms that live in our digestive tracts). Symbiotic relationships such as these are widespread through all branches of life (Wilson [2005]). If organisms are defined in terms of some type of functional or causal dependency, then surely gut flora are part of the human organism. We could not live without them. But if this is the case, an organism does not have a single genealogy, but rather, has many parts, which may have different genealogies.

The biological situation might then be parallel to the history of an ordinary physical object such as a car. It is clear that different parts of a car such as its engine, tires, or steering column might have different histories and in some contexts, asking about 'the' history of the car just seems to be asking about the history of its various parts.[2] An obvious way to attempt to avoid this in the biological case is by defining organisms (or at least 'organisms' for the purpose of our definition of taxa) in such a way that the gut flora is made up of trillions of different organisms, each with its own genealogy, while the human organism,

which does not include the gut flora, has its own separate, but unique, genealogy. The natural way to do this is to identify an organism (or an organism's genealogy) with a single genome type.[3] An intestinal bacterium has a separate genome from its host's genome, so they are separate organisms.[4]

Despite the fact that organisms are clearly more than just their genomes, thinking about the role that genealogy plays in taxonomy leads to the view that for many purposes, the genealogy of genomes is in some sense more fundamental for phylogeny than the history of 'whole organisms' in the more conventional sense of the term. Nevertheless, looking solely at an organism's genome does not solve the multiple origins problem – genomes themselves often do not have unique genealogies. Just as species are composed of organisms that have different genealogical histories, organisms themselves have parts (different genes) that have different genealogical histories.

Just as organisms have genealogies, genes also have genealogies. Any token copy of a gene in some individual has the structure that it does because it inherited it through replication from some 'parent' gene. Going back through many replication events, we eventually get to genes that were located in a different individual (usually a parent). If we examine two token copies of a gene (whether in two different individuals or in the same individual), they share a common ancestor in the past which is called the point of coalescence. The study of gene genealogies within a population (or set of taxa) is called coalescent theory and has become an important part of modern population genetics (Halliburton

[2004]; Hein et al. [2005]). Exactly how the genealogy of genes relates to taxonomy is the subject of the remainder of this paper.

## 3 Gene genealogies

I have just cited a few problems with the organismal conception of the Tree of Life. I do think that whole organism genealogical studies are important for many projects in modern biology and there are important problems that require us to think about phylogenies in this way. I will come back to these points at the end of the paper. Now I will present a third way of understanding genealogy; one based on genetic histories. To understand how taxonomy might be affected by the genealogy of genes, we have to first understand exactly what genealogy means at this level of biological organization.

It is easy to see that different genes in an organism can have different histories. In purely asexual organisms, during reproduction, the entire genome is passed from parent to offspring and so all genes will have concordant (identical) histories. However, in biparental organisms, the genome is made up of a set of genes, a subset of which came from each parent. In the case of lateral gene transfer, a subset of genes is passed from one organism to another without this involving an act of reproduction. Like sexual reproduction, this also causes different genes in the same genome to have discordant (non-identical) histories.

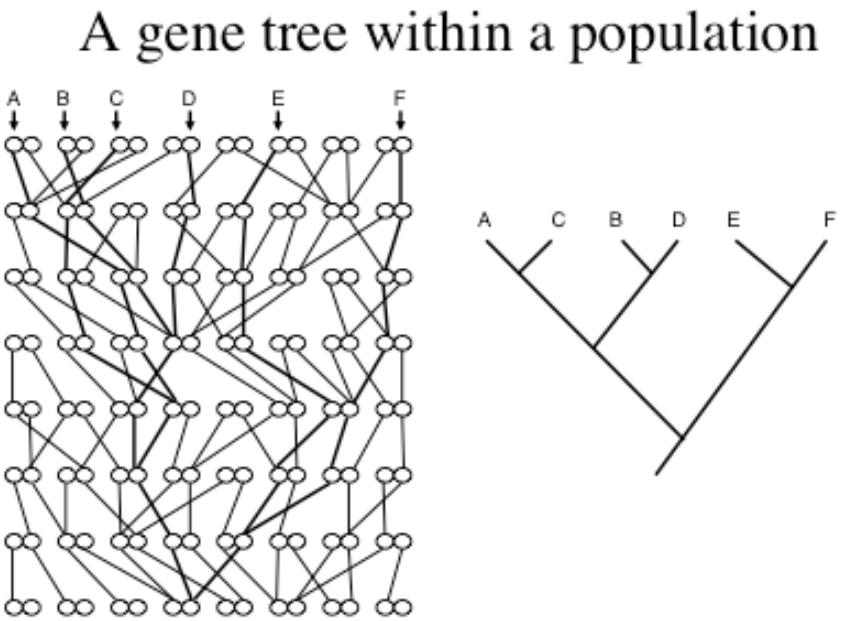
If homologous genes (genes that are 'the same gene' because they share common ancestry – for example, they are cytochrome c genes) are sampled from two organisms, we can ask, 'How far back in time do these genic lineages

coalesce?' The more recently they coalesce, the more closely related the genes are. We say that gene copy A is more closely related to copy B than it is to copy C if A and B coalesce with each other more recently than either does with C. This is completely analogous to the case of common ancestry of uniparental organisms in which there is no lateral transfer. If genes here are, by definition, non-recombining genetic elements, each gene has a unique genealogical history and gene genealogies form strictly divergent branching trees.

Figure 1 represents the history of a number of copies of a homologous gene present in different individuals in a population. The organisms are diploids, meaning that they have two copies of each homologous chromosome represented by two connected circles. The vertical lines are lines of gene descent from parent to offspring. Some organisms pass on multiple copies of the gene while others pass on none.

The lineages in bold represent a single token gene (a copy of a single allele) and all of its descendents as they spread through the population. If we sample a number of copies of this gene at the tips (A-F) they will form a tree based on their times of coalescence with the other tips. In this case, token copy A is more closely related to token copy C than to token copy B since A and C coalesce one generation more recently. Since each gene has only one parent, gene genealogies are purely diverging trees. If we extended this population back through time, eventually we would get to a point where we could trace the history of all 16 copies of the gene at the present time back to their most recent common ancestor (MRCA).

**Figure 1.** A population of organisms with the history of a single gene (a copy of a single allele) highlighted. The diagram is inspired by those in Maddison (1995).




Within a single population of organisms, different genes can form different trees. For some genes, my sister and I have copies that coalesce in one of our parents. For other genes, they coalesce very deeply in the past. So the full network of organismal genealogies does not determine any of the gene genealogies. Not every genetic lineage goes through the most recent common ancestor of two organisms. If phylogenetic trees are to be useful for tracing the histories of traits and traits are at least partially determined by gene histories, then knowing only the organismal histories (and only a fraction of the organismal history when we are basing our taxa only on MRCAs) will not suffice to fully understand the history of heritable traits. Knowing only about recency of

common ancestry at the organismal level provides only an incomplete genealogical history of the traits those organisms have.

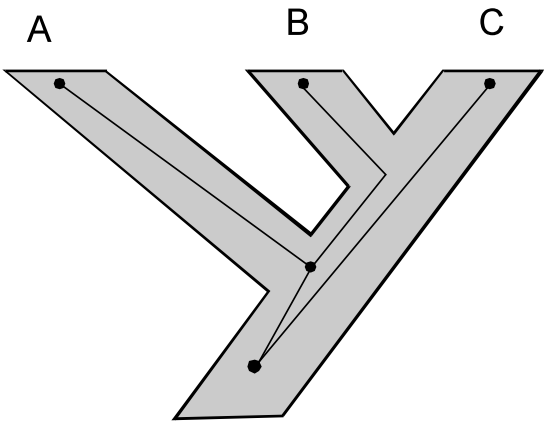
While it is obvious that different genes have different trees *within a population*, it is less obvious is that different genes indicate different relationships *between populations*. Of course, in cases of horizontal gene transfer, this is clear. If a bacterium transfers a gene to a human, that gene will suggest that the human is closer to a particular group of bacteria than to chimpanzees. Any non-transferred genes would indicate otherwise. Disagreement among gene trees could also be the result of species hybridization or even symbiogenesis (the merging of two organisms). But as modern phylogenetic studies have made clear, disagreement among gene trees is common, and is to be expected as the result of ordinary branching speciation events.

In many cases, two copies of a gene in a single population will coalesce earlier in time than when that population split from another. But this means that a copy of a gene is often more closely related to a copy from another population than it is to some copies in its own population. And 'population' is just a placeholder here – this kind of non-exclusivity extends above the level of traditional species: often gene copies in one species are closer to copies in another species than to other copies in its own species. If the time to coalescence extends past two speciation events, then we can have gene discordance at the level of species histories. In other words, gene tree topologies can differ from the topology of the corresponding 'species tree'. For example, if we sampled a gene from a human, a chimp, and a gorilla, the expectation is that the genes from the

human and the chimp would coalesce more recently with each other than either does with the one from a gorilla. But because some gene coalescent times are very different than others, some genes will exhibit different histories than this – some genes will indicate that humans and gorillas form a clade and others will indicate that chimps and gorillas do. This is possible even if we have defined human, chimp, and gorilla in a way that guarantees that the organisms in each form genealogically exclusive groups.

Figure 2 provides an example of how gene trees can indicate relationships that differ from those given by species trees.

**Figure 2.** A, B, and C represent separate taxa as determined by recency of common ancestry of the organisms in them. The thick branches contain lineages and the skinny black lines highlights a single gene's history where the token copies from A and B coalesce more recently than either does with C.




In this case, the 'fat branches' represent organismal histories while the single dots represent token copies of genes and the skinny black lines are the genic lineages. In this case, while the organisms at the tips in C are more closely related to those

in B than to those in A, this particular gene has a history in which A and B from a clade that does not include C.

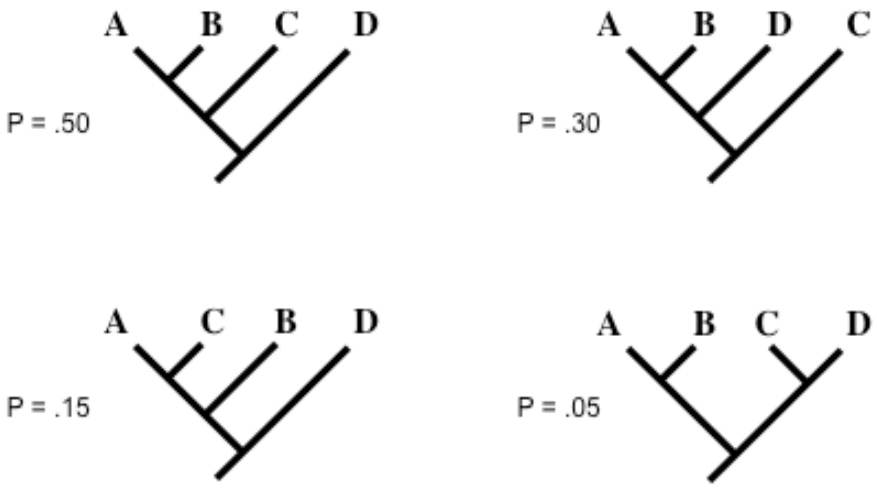
Since the 1980's, this phenomenon has become widely known and studied and is typically referred to as the relationship between gene trees and species trees (Tateno et al. [1982]; Hudson [1983]; Tajima [1983]; Wilson et al. [1985]). Since gene sequences are the primary evidence used to infer phylogenies of taxa, at the very minimum this presents an important epistemological issue in attempting to infer phylogenies. If we simply sampled one copy of a homologous gene from a number of taxa, even if we were able to correctly infer the phylogeny of the gene, the gene need not have the same phylogeny as the taxa we wish to study. The problem is not restricted to genes per se. Imagine the over-simplistic case where a phenotypic character is completely controlled by one gene. This means that the character distribution represents misleading data, but not because of homoplasy as is typically assumed, but because the branching history of the character really is just different from the branching history of the taxa.

While this is a very important issue, it could be merely an epistemological difficulty. We can acknowledge that genes have genealogies just like organisms, populations, species, and taxa do, and that the genealogical relations at these varying levels are just different. When building species trees, if we could ignore gene histories and directly infer the history of species, we would do that. But there are other possible responses – such as that in some important metaphysical sense, the genealogy of organisms or species (or both) <u>just is</u> (or maybe is determined by) the genealogy of genes. As Maddison ([1995]) puts it, 'one

possible interpretation of a species phylogeny is that it depicts the lines by which genetic information was passed on *and nothing more*' (Maddison [1995], p. 285). Maddison interprets Baum ([1992]) and Baum and Shaw ([1995]) as holding this position, though it is not clear that they do.

On this view, a species phylogeny is composed directly of gene trees and the genealogical relationships among these species do not form a single branching pattern, but rather are composed of a number of branching patterns. A species tree is thus a 'statistical distribution' of gene trees which might be conveyed by a statement like 'half of the genes in species A-D show a D(C(B,A)) topology but three other topologies are represented by 30, 15, and 5 percent of the genes' (Figure 3).

**Figure 3.** The phylogeny of species A, B, C, and D represented as a statistical distribution. In this case, only 4 of the 15 possible topologies are represented.



In the end Maddison rejects this position because it 'views populations as just bags of genes that happen to have been entangled in their history' (Maddison [1995], p. 286), though he is a bit more favorable to the position in (Maddison [1997]).

While this is a rather stark view of the nature of populations, species, or taxa generally, Maddison does point out the benefits that this view has for thinking about phylogeny. For inferences that rely on phylogenies, we are concerned with the actual, historical distribution of traits (like genes) and facts about what could have happened but did not happen are irrelevant. What Maddison fails to realize is that phylogenies can have the required properties without this entailing the stark metaphysical picture that says species are just collections of genes. We can still define taxa such as species as groups of organisms where the organisms are genealogically related; we just need to understand phylogeny in a way that properly takes these gene genealogies into account.

## 4 Exclusivity as recentness of genetic coalescence

In order for a species, or any taxon, to have a unique genealogy, it must be composed of a genealogically exclusive group of organisms. Again, an exclusive group is roughly, a group of organisms that are more closely related to each other than to any organisms outside the group. A natural way to define how closely related two organisms are is in terms of their recentness of common ancestry. But

we know that not all of our genes come from our most recent common ancestors. So perhaps we need to have a broader understanding of genealogical relatedness that takes this into account. For example, on the MRCA view, siblings are more closely related than first cousins are because they share a parent rather than just a grandparent as their MRCA. Notice that only one common ancestor determines this relationship. On this view, two half-siblings are just as closely related as two full siblings are. An alternate way to think about the situation is that full siblings are more closely related than half-siblings because they share *two* parents, not just one. An explanation of this fact that you might find in an introductory genetics text is that half-siblings are expected to share one quarter of their genes whereas siblings are expected to share one half.[5] With more distant relationships, it becomes more obvious that we may be more concerned with trying to combine the history of a number of our ancestors rather than simply being concerned with a single ancestor in common who happens to be the most recent of all of a large number of distinct common ancestors that we might share.

Genetic relatedness provides a way to measure relatedness that can come apart from recency of organismic common ancestry. Since different genes have different histories, we have to think of the organism's genealogy as a combination of gene histories. This takes into account the idea that not all of an organism's genes are passed down from a single ancestor.

One way of doing this is to alter our understanding of what an exclusive group of organisms is. A precise definition of exclusivity in terms of common ancestry has been developed by David Baum in collaboration with Elliott Sober

and Joel Velasco (Baum [2009]; Velasco [2009]). However, we could also define exclusivity in terms of gene histories and historically, it was first made precise in this way. Baum and Shaw ([1995]) define an exclusive group of organisms as one whose genes coalesce more recently within the group than between any member of the group and any organism outside the group. To hold on to the view that species have a unique place on the Tree of Life, they then use this definition in what they call the 'Genealogical Species Concept' (GSC), which holds that a species is an exclusive group with no exclusive groups inside it.

This definition does seem to implicitly define exclusivity as what we might call '100% exclusivity' where every gene must have a within-group coalescence. This definition has a number of interesting consequences. One is that since gene trees are strictly diverging, we can be sure that exclusive groups really can be treated as a phylogenetic unit with a unique position in any phylogeny. The phylogenies of a number of exclusive groups will form a strict hierarchy and thus can be represented as a diverging tree. Another interesting consequence is that it removes the possibility of gene trees disagreeing with 'species trees'. If you do have a gene tree that disagrees with what you thought the species tree was, either you have inferred the gene tree incorrectly, you are wrong about the species tree, or you are mistaken about which groups are species.

These 100% exclusive groups have a number of very useful properties from the standpoint of phylogenetics precisely because their definition entails that they will form a precise nested hierarchy. But we have not truly gotten rid of the discordance among gene trees – rather, all we have done is shifted the border

between interspecies relationships (the diverging relationships which are traditionally in the realm of systematics) and intraspecies relationships (the reticulate, network-like relationships which are traditionally outside the realm of systematics.) Rather than recognizing that the border between strictly diverging and even somewhat reticulate does not lie where we have traditionally located species, Baum and Shaw have simply defined species as being precisely at this border. But as we will see, despite its theoretical virtues, this view is simply too radical in its consequences concerning which groups we should call species.

## 5 From 100% to less

Requiring of an exclusive group that *all* of its genes coalesce more recently with each other than with any outside genes is a very stringent requirement. While it is true that if two populations are isolated long enough, drift alone will eventually lead to each being exclusive with respect to the other, achieving this 'reciprocal exclusivity' can take a very long time. Any instance of horizontal gene transfer instantly collapses the source and recipient into the same exclusive group, which is a taxonomic disaster. But even with ad hoc adjustments to our definition so we are only interested in the vertical history of genes, 100% exclusivity is still far too strong. Some polymorphisms which survive through lineage splits may be under balancing or frequency dependent selection, meaning that there will be pressure to keep at least two different types of alleles in the population. As long as the polymorphism persists, that gene will not show an exclusivity pattern concordant with other genes in the population.

These polymorphisms often exhibit coalescent times well past conventionally defined borders for species. For example, the major histocompatibility complex (MHC) is a large gene family that plays an important role in the immune system and can be found in most vertebrates. These genes contain the most extreme examples known of genetic polymorphisms in humans with the HLA-A, HLA-B, and HLA-DRB1 genes having roughly 250, 500, and 300 known alleles respectively. There are a number of possible explanations for how such extreme diversity could exist; for example, it may partially be the result of frequency dependent selection driven by pathogens. Since pathogens will evolve to attack the most common type of defense, genes in the minority will have an advantage. Clearly, the precise explanation does not matter here – what matters is that many of these polymorphisms are ancient, having existed for millions of years.

Figueroa et al. ([1988]) and Lawler et al. ([1988]) show that many of these polymorphisms predate the split of chimpanzees and humans, meaning that some human copies will be closer to chimpanzee copies than to other human copies. This means that if we enforce 100% exclusivity requirements, humans and chimps will have to be the same species. Ayala and Escalante ([1996]) suggest that some could be far older – perhaps as old as 30 million years as some human alleles appear to be more closely related to those of orangutans and even some old world monkeys. If this were correct, it is likely that the smallest exclusive group that contains all humans would be the entire clade Catarrhini consisting of Old World monkeys and the Apes (currently classified as approximately 100 extant species).

This group would itself either be a species, or more likely, it may have some smaller exclusive groups within it so groups such as humans would be metaspecies (not part of any basal exclusive group and so part of no species). While this view does place species on firm ontological ground, it lacks much of the important theoretical and practical virtues in recognizing that there are genuine phylogenetic histories of far smaller groups (like humans).

This account is simply too radical to be an account of *species*. Just how conservative a species concept has to be will depend on other virtues that it has and how it compares to other concepts. I would suggest that it is not obvious that *Homo sapiens* (as currently understood) exactly delimits a species – for example, perhaps *Homo neanderthalensis* should be lumped together with it. However, a reasonable criterion would surely recognize that humans and chimpanzees are different species. If we did shift the meaning of species so that there could not be any discordant gene histories between species, then we would simply need to invent another taxonomic rank (perhaps subspecies) which would recognize groups such as humans and chimpanzees and would end up functioning in almost precisely the way that 'species' functions now.

An obvious fix is to suggest that 100% exclusivity is not required for a group to form a taxon. Shaw ([2001]) drops the strict requirement and says, 'Exclusivity, where members of the daughter population are more closely related to each other than to organisms outside the population, applies once the <u>majority</u> of gene copies each find their closest genealogical relative within that population' (Shaw [2001], p.881) (emphasis added). Dawkins ([2004]) also suggests that in

the future, we will think of phylogenies as recording what is true for a majority of genes.

Baum ([2007]) points out that 50% is an arbitrary cutoff that has no real biological importance and instead suggests that the best concept of exclusivity makes use of what he calls 'concordance factors' – the proportion of the genome for which a given clade is true. An exclusive group is simply any clade that has a higher concordance factor than any contradictory clade. This view suggests that a divergent phylogeny arises over time and that there may be no precise moment at which reticulating network histories suddenly diverge. The 'primary concordance tree', which represents all of the exclusive groups, is thus not the 'full' Tree of Life which respects all aspects of genealogy; rather, it is something like the best single tree that gets the most right (with respect to gene genealogies) if you have to pick just one tree.

The primary concordance tree is useful for taxonomy precisely because it is a single tree and so we can use it as a basis for phylogenetic classification. There is an obvious correspondence between a tree and a hierarchy of groups within groups, which is an important goal, if not a requirement, of a taxonomy. This tree serves the purposes of classification, but the real, 'full' genealogical history is a very complicated collection of gene trees. Many of these will overlap and this overlap is what is represented on the primary concordance tree, but genes that have histories that differ from most other genes have real histories too that should not be dismissed or suppressed as some type of 'error'. Rather, each gene has a real history. There is a difficult practical question of how to present as

much relevant information as possible in a precise but compressed way, but there are no serious ontological difficulties with this picture of phylogenetic history as the history of genes.[6]

It should be noted that concordance factors were developed to describe the meaning of phylogeny. They do not, by themselves, immediately lead to any specific definition of species. On a phylogenetic view of taxonomy, all taxa must be clades. The gene-based concordance view says that by 'clades' here, we mean clades on the primary concordance tree, which are exclusive groups. A phylogenetic view of species says that since species are taxa, they must be clades as well. But which clades are species? That is *ranking* question which remains to be answered. A variety of answers are possible – for example, Baum (2009) argues that there is no single, objective ranking criterion, but rather, we should use a set of 'semi-subjective' criteria. But importantly, however ranking is determined, as long as species as necessarily exclusive groups, then any tree of species will be consistent with the primary concordance tree.

We have now moved to a view in which phylogenies represent statistical distributions of gene genealogies just as Maddison described. But rather than adopt the picture of taxa as nothing but 'bags of genes', we have traveled through the very natural route of thinking of taxa (and therefore, species) as groups of organisms that form genealogical units. What has changed is simply how we measure how closely organisms are related to each other. The idea that the genealogy of an organism is tied to the genealogy of its genes is a consequence of understanding what kind of properties genealogies are supposed to have.

Genealogies are supposed to represent the actual historical paths that allowed various bits of information from past organisms to causally influence the traits of organisms today. A precise understanding would have to restrict the causal relation in question using some sort of heritability conditions, but this need not be dealt with here. While genes do not capture all of the relevant heritable information, genes are obviously a large part of our genealogical history and ignoring discordant gene histories, as the whole organism based view of the tree does, leaves us with a concept of phylogeny cannot represent an important aspect of evolutionary history.

## 6 Criticisms of the genealogical species concept

While this gene-based version of a phylogenetic species concept has received relatively little attention in the literature, there have been some criticisms, which I will examine here. In their book *Speciation,* Coyne and Orr defend the popular Biological Species Concept (BSC), which they define in the most common way, as 'Species are groups of interbreeding natural populations that are reproductively isolated from other such groups' (Coyne and Orr [2004], p. 30). Coyne and Orr argue that something akin to the 'relaxed' version of the Genealogical Species Concept (GSC) above is the best phylogenetic concept, which in turn they seem to take to be the most serious alternative to the BSC. They then argue that the BSC should be favored over the GSC (Coyne and Orr [2004], pp. 467-70).

Coyne and Orr favor the BSC over this relaxed version of the GSC for three reasons. First, they say that the GSC results in many groups not being

members of any species at all.[7] Their main concern seems to be the way in which metaspecies are formed: 'At the moment when an isolated population becomes monophyletic [to be a criticism of the GSC, this should read 'exclusive' which is different from monophyletic], every individual in every other population instantly loses its status as belonging to any species. It seems odd that, without any change in its own genetic composition, a group can lose species status based on what happens in a remote population' (Coyne and Orr [2004], p. 467).

First, we should clarify exactly the kind of case Coyne and Orr have in mind. It is actually impossible for a group that counts as a species to lose this status on the GSC view (unless it becomes a higher taxon). What they have in mind is that a new species forms from within a larger species. Then, according to Coyne and Orr, this leaves the remainder of the old species not inside any species.

This criticism is problematic on several fronts. In the first place, the entire problem relies on the existence of a precise moment of speciation within the lifetime of individual organisms that is in no way required by the GSC. In fact, relying on tracking species through time already borders on question begging against an exclusivity view of classification that is explicitly synchronic. But if we are concerned about species through time, there are some obvious ways of trying to extend time-limited groups to time-extended groups (Baum [1998]) so it might be possible to proceed in examining this criticism.

Not only do Coyne and Orr incorrectly assume that the GSC attempts to track species through time, they also assume that once a group within a species becomes monophyletic [exclusive] it must be a new species. This relies on a

particular criterion of ranking – namely, that species are by definition, basal exclusive groups (as in Baum and Shaw [1995]). This criterion may make sense on the 100% exclusivity idea (in which case their example of *Drosophilia simulans i*s not an example), but is obviously inappropriate for the plurality concept. For example, two siblings might share 50% of their genes and so will form an exclusive group, but of course they are not a new species. On the relaxed version of the GSC, ranking is made using other criteria and we are not required to have metaspecies at all – every organism can be a member of one and only one species and so there are no metaspecies (Baum [2009]).

A factor in this first criticism was that a population could change what species it is in without any genetic or phenotypic changes to the organisms in a population. This idea is repeated in their second criticism, that 'little of biological import occurs at the completion of genealogical speciation. What significance, for example, can one impute to the moment at which the proportion of loci showing exclusivity rises from 50% to 50.1%?' As argued before, 50% is an arbitrary benchmark, which is why Baum ([2007]) and ([2009]) uses a plurality concept where the clade simply has a higher concordance factor than any other contradictory clade. Here, if we track changes through time, we will say that new exclusive groups arise when there is a change in which clade represents the most common pattern. At what precise point is there enough divergence to represent a change from one taxon to two? It is impossible to say. But this is a strength of the view, not a problem. As Baum says, 'I would propose that the search for a particular CF [concordance factor] threshold that denotes the boundary between

reticulation and divergence is doomed. The acquisition of a divergent structure accrues gradually as a result of gene lineage extinction in reproductively isolated populations or demes (see, for example, Avise and Ball [1990]; Avise and Wollenberg [1997]; Maddison [1997])' (Baum [2007], pp. 425-6).

The third criticism that Coyne and Orr make represents a real difference in the goals of the systematist who favors the GSC and one who favors the BSC and it is hard to see how any response can be adequate without question begging. But I will attempt a brief defense here. The criticism centers around the idea that genealogical speciation will often be transitory since allopatric populations may become exclusive without being intrinsically isolated from other species and this is no guarantee that they will stay exclusive when reproductive barriers are removed. The evolution of intrinsic barriers to reproduction is no guarantee that there will always be such barriers, but it is surely correct that external barriers such as geographical separation are much more fluid and apt to change over evolutionary time. This type of comparison between the BSC and versions of phylogenetic concepts can be found elsewhere such as in Avise and Ball ([1990]), Avise and Wollenberg ([1997]), and Avise ([2000]).

If we delimit taxa based only on actual reproductive history and ignore the difference between temporary and permanent barriers as versions of the PSC do, then our *current* taxa might not *stay* taxa. The focus of Coyne and Orr ([2004]) throughout is to describe the importance of gene flow and reproductive isolation in population genetics. For them, species are the fundamental units of evolution. An opposing goal represented by Nelson ([1979]), Donoghue ([1985]), and

virtually all defenders of phylogenetic species concepts is to determine what species concept fits best in systematic theory. In systematic theory, our primary interest is in recovering and representing evolutionary history. In systematics, species play the role of the fundamental units of phylogeny. We can describe patterns of distribution of traits across taxa without any reference to units of evolution or possible patterns in the future.

The BSC delimits species by their potential for sharing descendants. Versions of the PSC group organisms by their shared ancestry. Which is better? Obviously, the best concept will depend on what we use the species concept for (Baum and Donoghue [1995]). Hybrid views which allow non-genealogical groups to be species, but then demand that these same species be the fundamental units in a phylogenetic taxonomy are unacceptable. It may be a consistent position to use 'species' to refer to special groups such as reproductively compatible groups and then simply invent a new term for fundamental taxa which are the basic units in a phylogenetic classification, but this is not the typical strategy of defenders of the BSC. Rather, they agree that species are taxonomic units and that higher taxa are groups of species delimited by their history, and they argue that the BSC is the proper way to delimit such groups since species, unlike other taxa, are special and have an additional role to play. But nothing could play both roles. If species are to be units of phylogeny, they must be genealogical units which are united by their past.

Taxa that are united by their unique, shared past need not stay that way. But there is no particular reason that they must stay that way in order to serve the

central purposes of systematics – recovering and representing evolutionary history. It is certainly of biological interest to determine which of these groups are likely to stay exclusive, but like allowing paraphyletic groups as taxa, attempting to build these forward-looking interests into the definition of taxa necessarily disrupts the central goals of systematics (de Queiroz [1988]).

## 7 The Tree of Life?

It is obvious that the study of gene genealogies is an important part of biology. Gene genealogies also provide an important source of evidence for organism pedigrees. Certainly, organisms have real genealogies that we might have a practical reason to care about, but it might be thought that, biologically, all of the interesting genealogical action takes place at the level of genes. I do not think that this is right. While concordance factors are valuable tools in understanding exactly what phylogenies represent, organisms really do have a history of their own that is valuable for certain biological purposes. There are many processes and forces of evolution that act on whole organisms and not on genes directly. Careful study of such processes may require that we understand the specific patterns of mating between organisms independently of knowing anything about which genes are in which organisms. For example, monophyletic groups of organisms are valuable when tracing biogeographical patterns across time and space because geographic range is a heritable trait from parent to offspring. Here, it is appropriate to think of lineages of organisms on their own terms rather than being defined by how closely related their genes are. And of course, there is far

more to heredity than genes. Studying phenomena such as epigenetic changes requires examining organismal pedigrees independent of any gene genealogies. Essentially, the whole field of developmental systems theory depends on looking at the history of whole organisms as opposed to just their genes (Oyama [2000]).

Where does this leave us with respect to the Tree of Life? On the organism-based view, the Tree of Life connects all organisms via parent-offspring reproductive relationships. On the gene-centric picture of species, all taxa, including species, occupy a particular place on what can be called the 'primary concordance tree' (Baum [2007]). This tree is something like the single phylogenetic tree that gets the most right if we have to pick just one tree. But each gene has a unique tree and there are multiple Trees of Life, each of which is objectively correct. This seems to lead back to the Maddisonian picture of phylogeny with a particular view of taxonomy added on. To get a single diachronic object, we could take all of these gene trees and combine them in a single representation that we might call the 'Web of Life' or the 'Net of Life' (Kunin et al. [2005]) or the 'Synthesis of Life' (Bapteste et al. [2004]).

In light of the disagreement between various understandings of the Tree and how it relates to classification, phylogeny, and genealogy, what happens to our apparent starting place that species and other taxa should occupy a particular place on the Tree of Life? A view of exclusivity such as (Velasco [2009]) keeps to this tradition, though as lateral gene transfer and other causes of gene tree discordance increase, these taxa will be driven farther apart from the taxa described by concepts based on genetic coalescence. These gene-based views

hold that taxa have a unique place on the primary concordance tree. Which tree is important to biology? Clearly, they both are.

Genes have real genealogies that are biologically important to study. Organisms have genealogical histories that are important to study. Once we specify each of these two levels, it becomes easier to see that there is no independent third level 'the species level' that has a genealogy that is important to study. If we insist on a view of species that is not simply reducible to either of these lower levels, and then we insist that species have genealogies as defined by speciation events, we can construct a 'phylogeny' of species, but it would not function in biological explanations in the way that we want it to. Species would have ancestors and thus genealogies, but these genealogical histories would not track heritable traits over time, since these follow the gene histories and are constrained by organism pedigrees. In cases where this species genealogy conflicts with these lower level histories, it is unclear why we should accept that the concept of genealogy applies at the species level at all.

Of course, we can sensibly talk about relationships between species. The history of a species, like any taxon, is just is the history of the organisms in it. This is the 'reductionist' part of the picture I would defend. On the other hand, organisms have histories independent of the histories of their genes, that is why there are two levels rather than just one. If species are exclusive groups of organisms, then it makes perfect sense to talk about the genealogy of a species. Exclusive groups of organisms could be defined either in terms of organism pedigrees or in terms of gene genealogies. Each is biologically important for

certain types of explanations. Either gives us a sensible way to talk about species as units of phylogeny.

## Endnotes

1. I say 'reductive' because the genealogy of a species supervenes on the lower level genealogies of organisms. I put 'reductive' in quotes because on some understandings of reduction, to reduce species genealogy implies that species do not have genealogies. I do not want to assume this eliminative form of reductionism.

2. This analogy comes from a discussion by Elliott Sober at the 'Questioning the Tree of Life' session at the 2008 Philosophy of Science Association meeting in Pittsburgh, PA.

3. I say 'type' here so that multiple cells in my body, which arguably have different token genomes (or definitely do in the case of mutations) can still be a part of me. Worries about monozygotic twins and the like might require an extra clause.

4. There are disputes about what counts as a single genome, but on every view, gut bacteria have a different genome. On my view, mitochondrial DNA is part of the human genome since it is inside the cell. Thus mitochondria (now) are not separate organisms. On other views, they are not part of the human genome and

so by my definition, they are separate organisms. I take this to be an unacceptable conclusion so I am relying on defining the genome in a way that rules that out.

5. 'Sharing one half of their genes' is a common way of expressing a particular relationship that is very difficult to describe. With siblings, it is equivalent to claiming that the expectation is that one half of their genes will coalesce in one or the other of their parents whereas with half-siblings, one quarter of their genes are expected to coalesce in a parent. But this type of translation is difficult to generalize. For example, we say that siblings whose parents are themselves cousins share 9/16 of their genes (of the 50% of their homologous genes that come from different parents, 1/8 of them will be shared between the parents since they are cousins so 1/2+1/8*1/2 = 9/16). In this case, 9/16ths of their genes are expected to coalesce within the last three generations, not just within the parents.

6. I say there are no 'serious' difficulties, but there certainly are difficulties. Details of the view can be found in (Baum [2007] and [2009]), but they leave many technical questions open. For example, to calculate a concordance factor, we need to know the 'percentage of the genome that forms a clade'. But how is this calculated? Does each gene count equally? Or each base pair so that longer genes have heavier weights? What about genes that have homologues in some, but not all of the taxa under study? There are multiple ways of making this view precise; the only question is which of the ways is best. But I would suggest that

for our purposes, the picture of taxa and genealogy is clear enough without these details.

7. Every organism is part of some exclusive group. But the smallest exclusive group it is in might have a smaller exclusive group inside it that doesn't contain the organism in question. The 'leftover' organisms that aren't part of any basal exclusive group form a metaspecies (by the definition given in Baum and Shaw [1995]) – other papers such as (de Queiroz and Donoghue [1988]) use an epistemic definition which would not fit the example here).

**Acknowledgements**

I would like to thank an anonymous referee from *BJPS* as well as Matt Barker, David Baum, Jamie Brett, Marc Ereshefsky, Laura Franklin-Hall, Ehud Lamm, Brent Mishler, and Elliott Sober for valuable discussion and comments on earlier versions of this paper. I would also like to thank audiences from the Bay Area Biosystematists, the Philosophical Pizza Munch at the California Academy of Sciences, the Humanities Fellows at Stanford University, and the University of Minnesota for providing feedback on talks based on this paper.

## References


Avise, J. C. and Ball, R. M. [1990]: 'Principles of genealogical concordance in species concepts and biological taxonomy', *Oxford Surveys in Evolutionary Biology*, **7**, pp. 45–67.

Avise, J. C. and Wollenberg, K. [1997]: 'Phylogenetics and the origin of species', *Proceedings of the National Academy of Sciences USA*, **94**, pp. 7748-55.

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

Ayala, F. J. and Escalante, A. E. [1996]: 'The evolution of human populations: A molecular perspective', *Molecular Phylogenetics and Evolution* **5**(1), pp. 188–201.

Bapteste, E., Boucher, Y., Leigh, J. and Doolittle, W. F. [2004]: 'Phylogenetic reconstruction and lateral gene transfer', *Trends in Microbiology*, **12**, pp. 406-11.

Baum, D. A. [1992]: 'Phylogenetic species concepts', *Trends in Ecology & Evolution*, **7**(1), pp. 1-2.

Baum, D. A. [1998]: 'Individuality and the existence of species through time', *Systematic Biology*, 47(4), pp. 641-53.

Baum, D. A. [2007]: 'Concordance trees, concordance factors, and the exploration of reticulate genealogy', *Taxon*, **56**(2), pp. 417–26.

Baum, D. A. [2009]: 'Species as ranked taxa', *Systematic Biology*, **58**, pp. 74-86.

Baum, D. A. and Donoghue, M. J. [1995]: 'Choosing among alternative 'phylogenetic' species concepts', *Systematic Botany*, **20**(4), pp. 560-73.

Baum, D. A. and Shaw, K. L. [1995]: 'Genealogical perspectives on the species problem', in P. C. Hoch and A. G. Stephenson (*eds*), 1995, *Experimental and Molecular Approaches to Plant Biosystematics*, St. Louis, MO: Missouri Botanical Garden, pp. 289-303.

Coyne, J. A. and Orr, H. A. [2004]: *Speciation*, Sunderland, MA: Sinauer Associates.

Cracraft, J. and Donoghue, M. J. [2004]: 'Introduction', in J. Cracraft and M. J. Donoghue (*eds*), *Assembling the Tree of Life*, Oxford: Oxford University Press, USA, pp. 1-4.

Dawkins, R. [2004]: *The Ancestor's Tale*, New York: Houghton Mifflin Company.

de Queiroz, K. [1988]: 'Systematics and the darwinian revolution', *Philosophy of Science*, **55**(2), pp. 238-59.

de Queiroz, K. and Donoghue, M. [1988]: 'Phylogenetic systematics and the species problem', *Cladistics*, **4**, pp. 317-38.

Donoghue, M. J. [1985]: 'A critique of the biological species concept and recommendations for a phylogenetic alternative', *The Bryologist*, **88**(3), pp. 172-81.

Doolittle, W. F. and Bapteste, E. [2007]: 'Pattern pluralism and the Tree of Life hypothesis', *Proceedings of the National Academy of Sciences*, **104**(7), pp. 2043-7.

Figueroa, F., Günther, E. and Klein, J. [1988]: 'MHC polymorphism predating speciation', *Nature [London]*, **335**, pp. 265–7.

Franklin, L. R. [2007]: 'Bacteria, sex, and systematics', *Philosophy of Science*, **74**, pp. 69-94.

Halliburton, R. [2004]: *Introduction to population genetics*, Upper Saddle River, NJ: Pearson/Prentice Hall.

Hein, J., Schierup, M. and Wiuf, C. [2005]: *Gene genealogies, variation and evolution : A primer in coalescent theory*, New York: Oxford University Press.

Hodkinson, T. R. and Parnell, J. A. N. [2006]: 'Introduction to the systematics of species rich groups', in T. R. Hodkinson and J. A. N. Parnell (*eds*), 2006,

*Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa*, Boca Raton, FL: CRC, pp. 1-20.

Hudson, R. R. [1983]: 'Testing the constant-rate neutral model with protein sequence data', *Evolution*, **37**, pp. 203-17.

Kingman, J. F. C. [1982]: 'The coalescent', *Stochastic Processes and their Applications*, **13**(3), pp. 235-48.

Kunin, K., Goldovsky, L., Darzentas, N. and Ouzounis, C. A. [2005]: 'The net of life: Reconstructing the microbial phylogenetic network', *Genome Research*, **15**, pp. 954-9.

LaPorte, J. [2005]: 'Is there a single objective, evolutionary tree of life?', *The Journal of Philosophy*, **102**(7), pp. 357-74.

Lawler, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P. and Parham, P. [1988]: 'HLA-A and B polymorphisms predate the divergence of humans and chimpanzees', *Nature*, **335**, pp. 268-71.

Maddison, W. P. [1995]: 'Phylogenetic histories within and among species', in P. C. Hoch and A. G. Stephenson (*eds*), 1995, *Experimental and Molecular Approaches to Plant Biosystematics*, St. Louis, MO: Missouri Botanical Garden, pp. 273-87.

Maddison, W. P. [1997]: 'Gene trees in species trees', *Systematic Biology*, **46**(3), pp. 523-36.

Mayr, E. [2001]: *What Evolution Is*, New York: Basic.

Mishler, B. D. and Donoghue, M. J. [1982]: 'Species concepts: A case for pluralism', *Systematic Zoology*, **31**, pp. 491-503.

Nelson, G. [1979]: 'Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's *Familles des Plantes* [1763-1764]', *Systematic Zoology*, **28**, pp. 1-21.

Oyama, S. [2000]: *The Ontogeny of Information*: *Developmental Systems and Evolution*, Second edition, Durham, N.C.: Duke University Press.

Palca, J. [1992]: 'CDC Closes the Case of The Florida Dentist, *Science*, **256**, pp. 1130-1.

Shaw, K. L. [2001]: 'The genealogical view of speciation', *Journal of Evolutionary Biology*, **14**, pp. 880-2.

Smith, T. F. and Waterman, M. S. [1992]: 'The Continuing Case of the Florida Dentist', *Science*, **256**, pp. 1155-6.

Tajima, F. [1983]: 'Evolutionary relationship of DNA sequences in finite populations', *Genetics*, **105**, pp. 437–60.

Tateno, Y., Nei, M. and Tajima, F. [1982]: 'Accuracy of estimated phylogenetic trees from molecular data I. Distantly related species', *Journal of Molecular Evolution*, **18**, pp. 387-404.

Velasco, J. D. [2008]: 'Species Concepts Should Not Conflict with Evolutionary History, but often do', *Studies in the History and Philosophy of Biological and Biomedical Sciences*, **39**, pp. 407-14.

Velasco, J. D. [2009]: 'When monophyly is not enough: Exclusivity as the key to defining a phylogenetic species concept', *Biology & Philosophy*, **24**, pp. 473-86.

Wiley, E. [1981]: *Phylogenetics: The theory and practice of phylogenetic systematics*, New York: Wiley-Interscience.

Wilkins, J. S. [2009]: *Species: a history of the idea*, Berkeley: University of California Press.

Wilson, A. C., Cann, R. L., Carr, S. M., George, M., Jr., Gyllensten, U. B., Helm-Bychowski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E.M., Sage, R. D. and

Stoneking, M. [1985]: 'Mitochondrial DNA and two perspectives on evolutionary genetics', *Biological Journal of the Linnean Society*, **26**(4), pp. 375-400.

Wilson, R. A. [2005]: Genes and the Agents of Life: The Individual in the Fragile Sciences: Biology, New York, NY: Cambridge University Press.