**Marcel!**

**Thank you for your time and attention on this – I know it’s not a easy read. I’ve addressed all of your straightforward points, but there are a few where I could really use some more feedback – mostly, I need some help where I can’t figure out how to bridge different nomenclatures, or where I’m not sure which part of a concept needs to be emphasized. I’ve marked these with a ‘\*\*’.**

**The major changes are as follows :**

**[1] I scrapped the conclusion and wrote a new one.**

**[2] I collected the discussion of methods that don’t really work for this application (phylogenetic signal and cospeciation tests by permutation) into a new section called Background, which I place just after the Introduction. To this, I added a new subsection introducing previous work on the graph theory approach to tree topology and explaining how I build on it.**

**The organization is still not so great, but this chapter is going to be taken apart into two or three papers for publication. I would have saved myself a lot of trouble if I’d written it that way to being with. I can’t tell you why I thought this would be easier.**

Dear Russell,

As promised I concentrated on the unpublished parts of your thesis and just checked that I was happy with the published parts as making up an acceptable body of work for a PhD thesis. Overall, I’m more than satisfied that you have enough work: I would have actually said to not include everything in your thesis to make it more focused but I’m okay with what you have Jonathan have decided on this.

The main things that need to be done is to make the thesis Abstract complete. This gets published and needs to cover the entire contents of the thesis not just the unpublished parts. It needs to be written as a publishable abstract. (More details on this and lots of other things below.) For Chapter 1 it needs some rewriting. The Introduction fails to reference existing methods on which you build and there is quite a lot of introductory material in the “Results and Discussion” section. The Results and Discussion section is generally quite poorly organized and I would recommend separating these sections, so that there is a clear discussion of results that you have presented. The conclusion also is mainly a philosophical reflection rather than summing up what you found. I therefore think Chapter 1 needs another round before it is an acceptable part of any thesis or before it could be published in an appropriate peer-reviewed journal. It was also poorly proof read and perhaps not even spell checked, so more care is needed. Chapter 10 seems tidy enough overall, and gives a clear description of the new methods developed.

Cheers, Marcel.

I’ve marked more important points with a \*.

\* The abstract just reports on the unpublished parts of the thesis. It should give an overall abstract that covers the published chapters not just the details of the unpublished ones. The aim is to say what the thesis as a whole does and currently it does not do this. Given that it gets published online on dissertation abstracts it is good to write a reasonable abstract. It may provide incentive for somebody to look up your thesis or to check if you published papers on the subjects covered in it. You need to provide a full abstract not just parts of the thesis. A reader would be entirely mystified by "and has been published previously."

**\*\* Added complete descriptions for each chapter, which results in an nine page abstract (eek!). Is that really useful?**

**My rationale for leaving out the complete abstracts for chapters that are already in press was that, usually, the only reason someone ends up looking at a dissertation is as the result of a detailed literature search. Any search thorough enough to lead the researcher to my dissertation will surely turn up the published work first, right? It shouldn’t be too mysterious if the abstract includes the complete citation, I hope.**

**I don’t mind either way, I just don’t really understand what is expected. I tried digging up random dissertations, and most of them appear to be either a bunch of chapter abstracts pasted together, or they varied from a few sentences to several pages. I really have no idea what the abstract should look like, and everyone I ask just tells me “it should be an abstract for the dissertation.” :-/**

Throughout the Abstract words like “explores” are used. These could be replaced with stronger, more precise words such as tests, analyzes and say what was analyzed. Otherwise you reduce everything to being an exploration of patterns, and its clear that some of your chapter go well beyond this.

**Clarified.**

Abstract page XXX: The subtitle “Prediction of ecological function in the microbiome using machine learning on the graph spectra of coevolving subnetworks”. Learning on the graph spectra of … does not have a clear meaning. Reword to clarify.

Abstract page xxx. Delete this “The first chapter represents the largest single component

of this dissertation.” It doesn’t add anything to the content.

**Done.**

Abstract page xxx. The paragraph beginning “We propose…” is very wooly and roundabout in what it actually says. For instance, “forms the basis of feature space into which labeled and unlabeled interactions can be projected.” The next sentence said you used clustering methods on this space. Why not simply state what you analyzed and why? An abstract is meant to be a concise statement of what you did, how, what you found and usually some take-home message. Keeping it to this would be appropriate given that it is published in online dissertation abstracts.

**This chapter only describes the development of a method. There are no reliable or trustworthy scientific results at this point, and so I want to be careful not to represent it as such. The fish dataset was used to insure that the method I developed would be relevant to real data, and as an example of how the analysis would work. I do not have a large enough dataset of training data from the literature to really claim to have analyzed the fish dataset with this method; I’ve merely shown how it *could* be analyzed.**

Page xxx-xxxi. The sentence spanning the page does not say why you did this. What did training aim to do and was it to see if it worked, or why use this data and then a bacteria and archaea and fish dataset? The logic is far from clear and needs to be more carefully explained.

Page xxxi. In the following sentence change "develops" to "develop" in "Chapters five and six present results from a microbial survey of the Indian Ocean, and develops a citizen science approach for performing oceanographic research using small sailboats."

**Fixed.**

Page xxxi. In Part 3 the wording here is terrible "that develop the necessary tools for exploring the gene regulatory networks of halophilic archaea." I could explore it in many ways including searching the internet, so necessary tools doesn't tell me anything useful.

**Clarified.**

Page xxxi. Correct the grammar and clarify "has been published as preprint." Anything can be put in a preprint server and so it doesn't mean very much. I suspect you mean something different.

**That is exactly what I mean. It is a PeerJ Preprint. My main collaborator stopped working on haloarchaea, and my own research went in a different direction. I probably won’t pursue peer review.**

Page xxxi. Give the necessary context so that readers know what "peak finders" refers to.

**Adding the full abstract for Chapter 7 should take care of this.**

Page xxxii top. "freely licensed dependencies" is unclear.

**I agonized over this phrasing a lot, actually. Originally, I had written “non-proprietary,” but that didn’t sit right. I also considered “OSI-approved license,” but I don’t perfectly agree with OSI either. Which licenses are or aren’t “free” is a topic of considerable controversy within the open source / free software / software libre movement (there isn’t even a consensus about what to call the movement). I chose this language very carefully. There isn’t really a way to be more clear without taking what amounts to a political stand, and I didn’t think this was the right place to do that. Anyone who was motivated to learn my opinion about what “freely licensed” means can look at the licenses of the dependencies (they’re in the build manifest) and decide for themselves if they agree with me. I hope that’s OK.**

Acknowledgments were incomplete in what you sent us.

**Not to worry. I have an extensive list of people who need to be thanked. I decided to leave it out for this draft because I’m still double-checking it, and I didn’t want leave the impression that anyone was left out on purpose. It will be the final version.**

In the Introduction references are included but outside of sentences. They normally appear before the period as a part of a sentence. The references to "a crescent moon," "circumnavigated the globe," and "the jetwash of" are playful but would be hard for anybody who is a nonnative English speaker to extract the meaning. I would suggest to change the language to plain English that can be understood easily be an international audience, this is meant to be science after all.

**Heh. OK, OK.**

**Can I keep “circumnavigated the globe…?” It really is true.**

Introduction page 1. This contradicts itself "Things have cooled somewhat, though the pace of discovery continues to accelerate." Then you go from plural subjects to "While it revolutionized," so "it" doesn't make sense.

**Clarified. I want to convey that excitement has cooled a bit, but the pace of discovery has not.**

Page 2. In the following, what you claim about physiology could equally well be said about cell biology: “Physiology paints a picture of what behaviors are possible. Cell biology tells a story about what resources are required for survival.”

**I agree that the two subjects overlap. Cell biology and physiology are literally the same thing when the organism happens to be unicellular. It’s not perfectly obvious where to draw the line between the two, or if it even makes sense to make an explicit distinction. I’m pointing at where I think is the center of mass is in terms of human interest in those topics. It’s sort of like physics and chemistry – there is a difference between them, but the difference is a human convention, not a natural distinction. As a physicist (a lapsed one, anyway), I know I could work out the transition rates among the isomers of glucose, but if someone asked me, I would probably beg them to ask a chemist instead. That doesn’t mean that nature draws any meaningful distinction between physics and chemistry. So, I’m not saying that physiology can’t tell me what resources are required for survival – I’m saying that if I wanted to know, I would probably start by pestering a cell biologist.**

**This is just my own personal perspective. It seemed like as good a spot as any to express it.**

Page 2. The interpretation of phylogenomics is strange here since it considering organisms relative to other organisms that reveals things about evolution.

**Interpreting of phylogenomics *is* strange! :-)**

**The terms means a lot of things to different people, though I think it’s reasonable to consider them as the same activity through the lens of different questions. Wikipedia says that phylogenomics concerns itself with four general areas of research :**

**1) Prediction of gene function**

**2) Establishment and clarification of evolutionary relationships**

**3) Gene family evolution**

**4) Prediction and retracing** [**lateral gene transfer**](https://en.wikipedia.org/wiki/Lateral_gene_transfer)

**Placing phylogenetics in the context of genomics (or visa versa, if you prefer) admits a very broad range of questions. It’s not surprising (or a bad thing) that different people gravitate toward different applications. I think I might be being a little obtuse, but the idea I’m trying to get at is that while a collection of phylogenomic data might answer many different kinds of questions, the question one might want to answer might not be the question answered most readily. There is a sort of epistemological grain to it, and it’s easier to follow the grain.**

**Or, to look at it another way, you can ignore the grain, but then you will have to generate a lot of data that answers questions you weren’t asking.**

Page 2, sixth line from the bottom “its value” is unclear what it is referring to. The subject of the previous two paragraphs is “sound reasoning” and sound reasoning “as a target one may reason towards” is strange.

**Clarified.**

Page 3. Put the parts about 9 being published together so that the short second paragraph can be deleted and repetition cut down. Most of the rest of it beyond saying you describe author contributions is nonsense.

**Deleted.**

Bottom of page 3. This needs correcting “culmination of largest effort”. The largest effort.

**Fixed.**

Page 3-4. It would be clearer if you referred to chapters by numbers rather than first, second etc. repeatedly.

**Fixed.**

**Sorry about that – originally, the chapters were numbered hierarchically (1.1, 1.2…), but a few people found this annoying.**

Page 7. This sentence doesn’t make much sense “The interpenetrating dispersal of many organisms is the process that drives the assembly of ecosystems.” Dispersal is an increase in dispersion. I think you mean mixing or something similar.

**Right. An increase in dispersion is an increase in the spacial extent over which something is distributed. Two things that disperse into overlapping spaces don’t necessary mix, and I thought it would be going a step too far to call it mixing. But if you think it’s reasonable to call it mixing, though, I’ll go with that.**

Page 7. This is dissatisfying: “very little is understood about their natural histories. Microscopy, culture-based and molecular genetic methods have revealed a great deal about the physiology of microorganisms, but they have revealed only a patchy, distorted view of microbial natural history.” The purpose should be to put your work in a broader context yet you don’t cite any previous studies and say what we do know is patchy and distorted. A more careful context would be more appropriate.

**Good point. I often take it for granted that everyone worries about culturability and the Great Plate Count Anomaly and bias in culture collections and things like that, but I shouldn’t treat it as received wisdom. I’ve added some references to the classic Staley & Konopka review that coined the GPCA term in 1985, and a more recent review from 2013 that discusses the progress towards addressing the problem, particularly work from the Giovannoni lab on culturing difficult organisms.**

Last full paragraph on page 7. What about using chips to identify bacteria, doesn’t that make it far easier and more precise?

**Unfortunately, no. :-(**

**There are several chip-based approaches (PhyloChip, for example), but they don’t have a significant advantage over sequencing in terms of precision or ease of use (and certainly not cost), and have the unfortunate characteristic of introducing database bias at the detection step. They are mostly useful in applications where the community has already been very thoroughly characterized (which is to say, not very many of them).**

Page 8. Whitaker or Whitacker. Both are used on this page.

**Fixed.**

Correct extremeophiles. Extremophiles I presume.

**Fixed.**

Page 9. It is usual to give a page number for a quotation within a book.

**Fixed.**

Page 11. This sentence seems backwards: To date, substantial prior knowledge of the ecology, physiology and genetics has been required for all species of bacteria and archaea for which a dispersal-structured biogeography has been demonstrated.

**It does sound backwards, doesn’t it? It is what I mean to say, though.**

**Ecology, because one has to know where to find the organism in order to locate it and collect specimens it in several geographically isolated places. Physiology, because one has to understand at least enough to grow the organism in pure culture, which is usually necessary before one can amplify genes and sequence them (there are other ways to go about this, but nobody has used them for studying dispersal yet). Genetics, because one has to design primers targeting a collection of genes. By the time you can build the high resolution trees you’d need to say anything about biogeography, you’ve already had to learn quite a lot about the organism.**

**That’s fine if you started out with a pretty good idea that the organism would be interesting. This is how Whitaker and Papke approached their systems. People studying dispersal of infectious diseases through populations have a similar point of view. It’s not such a great state of affairs if the thing that would clue you into the possibility that an organism might be interesting is its biogeography.**

Page 16. In the Title, using learning on could be expressed in a better way. Perhaps, applying learning to or using learning to analyze/understand etc.

Page 16. “preparation and DNA sequencing was performed” should be were performed. Shortly thereafter this sentence needs attention “Software was written, analysis was performed by Russell Y. Neches with advice from Matthew D. McGee.”

**Fixed.**

Page 17, Abstract. This is unclear “Such phylogenies resemble “tangled trees".” Also abstracts usually end with a take-home message, and one might be added.

\* Introduction to Chapter 1. Given that you build on an existing method, it would be helpful to readers to introduce these methods. This serves both to orient readers to where your method fits into the realm of possible choices for the problem you are trying to solve, and would orient readers to the kind of methods that you intend to use.

Top of page 18. “in observational protocol” should be something like “in the observational protocol” to be grammatical.

**Fixed.**

Page 18. Fix “While is well established.”

**Fixed.**

Page 18. “or link these observations” link to what, each other I presume but should be clearer.

**Clarified.**

Page 19. Is this statement really true: coevolution is always the outcome of an ecological process that has a significant impact on the life history of an organism? Can it not be fitness without changing life history? I wonder about this especially given the disclaimer at the end of the paragraph.

**Fixed.**

**You’re right. Life history already has a specific meaning that I didn’t intend to evoke. Fitness would be a better way of thinking about it.**

Top of page 20. Fix “identification the mechanism of an interaction is…”.

**Fixed.**

Page 20. The purpose for stating this is unclear: “(more subdivisions are possible with more data).”

**Yeah, that is kind of a puddle of mush. I’ve been meaning to ask you about what would be the best terminology to use.**

**What I mean is that if you have data describing more kinds of interactions, then you can take a concept like “mutualism” and break it down into more specific kinds of mutualism (frugivory and pollination, in this case). With more data, it is likely that you could split those up into even more specific kinds of interactions. But… “kind of interaction” is not very illustrative of the concept, and I’m not sure how to describe a hierarchy of such things without being obtuse.**

Top of page 21. The connection of your system (bottom of page 20) to the following could be clearer, especially given the use of the literature data: “The habitat we have chosen to illustrate this method is a group…”.

Bottom of page 21. This needs fixing “These rectangular blocks are where the any between the two joined graphs must exist.”

**Fixed.**

Page 22. Correct “of each edge connected connected to that node.”

**Fixed.**

Page 24. Fix “from Hafner et al..”

**\*\* Should I add another citation? Or is something else wrong?**

Page 24. This sentence is incomplete. “In this way, a feature space constructed over the Laplacian spectra of networks of interaction species with known ecologies (parasitic verses mutualistic interactions, for example).” Presumably the feature space is constructed, generated or similar.

**Fixed.**

\* Page 24. The part about training a classifier is a clearer explanation than is used elsewhere. I think it would be good to add this to the chapter Introduction and Abstract, as well as the thesis Abstract.

Figure 1.3 doesn’t illustrate anything very clearly to me. Perhaps explain more fully what you are trying to show.

**The differences between the panels due the permutations are difficult to see – if you compare the locations of the dots in the off-diagonal blocks, you’ll notice that they are arranged differently in each panel. I tried a couple of different ways to make them pop out more obviously, but it’s still difficult to see with real data (as I’ve done here). This figure will probably be part of the supplement when this goes for publication.**

**I will add a schematic diagram to make this more obvious.**

Figure 1.5 a seems clear enough and useful, but part b is somewhat mysterious. Why is the top left section all purple for instance?

**This was another not-quite useful avenue I pursued, and will probably be relegated to the supplement when this goes for publication, or perhaps expanded into a little theory paper by itself. It’s one of several potentially interesting side-tracks that I bypassed on my way to Just Making The Damned Thing Work.**

**Before I tried the machine learning approach, the plan was to try to classify the type of interaction by Euclidean distance to the most similar interaction from literature. To see how well this was going to work, I computed the pairwise distances between the spectral density profiles of all the interactions in the training database, and threw some hierarchical clustering at it. I learned that while the interactions cluster weakly by type – mustualisms mostly with other mutualisms, nulls mostly with other nulls, and so on – the associations are messy. This is how I finally realized that this would have to be a machine learning problem; evidently, the space is non-Euclidean, and I didn’t have a reasonable theoretical framework to put it into. Machine learning is designed exactly for this kind of situation.**

**I included it here for completeness, though it doesn’t add much to the argument as the chapter stands right now.**

Bottom of page 29. Fix “feeding began as soon as it deemed safe.”

**Fixed.**

Page 31. “stool was removed”. Presumably a stool sample?

**As much of the stool I could get (usually all it). Fish poops aren’t very big. Originally, this said “stool samples were collected,” but it was pointed out that this might be misleading.**

Section 1.4.5 the title could be more grammatical. Presumably a phylogeny, or phylogenies.

**Fixed (changed it to “Building the OTU tree”).**

Top of page 35. Pearson’s I presume.

**I think that might be something else. I’ve heard it called “Pearson’s correlation coefficient,” “Pearson’s r,” “PCC,” “Pearson’s product-moment correlation coefficient,” or “bivariate correlation.”**

Middle page 35. Correct “where the tips of to Independently generated random trees.”

**Fixed.**

Bottom of page 35. Presumably the sequence of events could also be expressed as a likelihood, so explaining the first part in terms of likelihood is not a useful comparison. Also the “enlarged” question on the top of page 36 seems like a more localized version of the first question, so I would link the two.

**Clarified.**

\* The “Results and Discussion” initial section 1.5 before 1.5.1 begins seems more like introductory material or methods. I suggest moving it to an earlier section and giving a clear separate Results section and separate Discussion section. Currently they seem like a crutch for poor organization. Likewise, most of Section 1.5.1 seems like introduction or methods. Material from the literature is presented without relating it in any clear way to your own findings, which seems more like Introduction. Generally this section seems somewhat rambling and unfocused. The first paragraph on page 40 is the first part that reads like results. In section 1.5.2 it is only in the last paragraph that we get any results, and the material before this is introductory (before this the reference to Figure 1.1.3 is used to illustrate how things can be shown but nothing else).

**\*\* These two sections were originally part of the introduction, but I moved them to *Discussion & Results* after receiving feedback that they distracted from setting up the conceptual purpose of the research (I tend to agree). They don’t really fit in D&R either.**

**I have moved them to a new section called *Background*, which I have placed immediately after the introduction. I have also added a new subsection to *Background* introducing the previous work on using graph theory to examine tree topology, and explaining how I build on it.**

Page 36. The two parts of this sentence are disconnected and the intent is not clear: “Unfortunately, trait-based models do not account for interactions among traits, or a way to account for traits that have their own evolutionary model (i.e., traits that are themselves organisms).”

**\*\* This is another concept that I’ve struggled to articulate. Tests for phylogenetic signal are aimed at “conventional” traits, like presence or absence of spots, or the ability to synthesize one’s own tryptophan. These kinds of traits don’t exist independently of the organism that has them, and because of that, they don’t have an ecological or evolutionary history outside the context of their host organism. Mathematically, they could be treated as orthogonal. There are counterexamples, as always; a tail color trait, for example, could be said to “interact” with the trait of presence of absence of a tail. But for the most part, traits do not “interact” outside the context of their host’s phenotype.**

**Microbes are organisms unto themselves, but in the context of their relationship with a host organism, you can treat them as traits of the host. Conceptually, this works just fine as long as you only have one species of microbe. Once you have more than one microbe, the traits violate the assumption of orthogonality found in models of phylogenetic signal. Microbe A and Microbe B have a relationship with one another which can be expressed in their relative positions in the tree of life.**

**This is a problem because all tests for phylogenetic signal are correlation tests of some kind, and the non-independence of supposedly independent variables plays havoc on interpretation. This is not necessarily a fatal problem, but I decided I didn’t want to go down the path of inventing a new measure of phylogenetic signal (I have some half-baked attempts that I’d prefer never saw the light of day).**

Legend to Figure 1.9. Most of this is discussion and should be in the main text rather than the figure legend.

**Done.**

Page 38. Correct “Fist of all.” Then the following sentence needs fixing: “which is itself inferred from an probabilistic model (an approximate maximum likelihood model, in this case), which is based on nucelotide transitions inferred from from an alignment.” Repeated use of “which” clauses so the meaning is lost, spelling of “nucleotide”, and “from from”.

**Fixed.**

Figure 1.1 there is no scale to compare panels and in the legend “Pierson”.

**Fixed spelling.**

**The scales are patristic distances, which vary wildly among different cases. This figure is meant to illustrate the the *distributions* of patristic distances only, so that the reader can examine whether or not the distributions obey the assumptions of the Pearson’s correlation test (they don’t). I modeled the presentation on a couple of different textbook treatments of Anscombe's quartet, illustrating a similar point.**

\* Section 1.6 is largely a philosophical reflection on the findings rather than a summary or logical summation of what can be concluded from your analyses.

**Scrapped and re-written.**

Page 257 second from last line. Change “on microplate reader” to “on a microplate reader”.

**Fixed.**

Page 258. Does “laboratory grade water” have a clear meaning? Filtered, tap, etc?

**Clarification added.**

**Usually, it just means “pure, according to your own lab’s standards for working with DNA.” Different labs have different requirements depending on available equipment, funding and… frankly… superstition.**