Editor type comments

Two experts and I have reviewed the preprint entitled “Modelling coinfections to detect within-host interactions from genotype combination prevalences” and all have come to similar conclusions. First, all of us appreciate the topic and think that the work itself has value.

Thanks!

There were several noted areas where additional analyses would improve the impact, although most of these were minor. The primary area that the authors should focus on to improve the manuscript is on the presentation. All of the reviewers believed that the much of the work was under-presented (not enough detail), ambiguously described, and in several areas confusing. It is important to note that none of the reviewers identified all or even most of the areas where the presentation needs alteration, instead noting some examples of the types of presentation that caused confusion.

Given the unanimity, there must indeed have been a strong presentation problem. We were trying to keep the manuscript short but given this feedback, we now provide more details about the method and the biological background.

Globally, we found this an important piece of work and hope to see an improved version in the near future. Below you will find my personal review for what it is worth. Dustin

Thanks.

Reviewer #1

There are several important academic areas in the study of infectious diseases that have not been well explored, often because of technical difficulty. The topic of the current work, “Modelling coinfections to detect within-host interactions from genotype combination prevalences,” is certainly one of the. Both theory and several important examples have demonstrated that co-infection dynamics can have outsized impacts on epidemiology. Unfortunately, for the outcome of pathogen interactions are difficult to decipher from correlational data. The current work aims to identify the impacts of pathogen interactions on transmission dynamics using epidemiological data. This is important and would be a major advance.

Thanks for the accurate summary.

The primary issue with the current manuscript is in the presentation, which I found to be ambiguous in some places and confusing in others. A potential issue is that you have assumed that I am smarter than I am, but I should point out that those like me will be the primary audience of papers that introduce analytical approaches for empirical data. I would suggest more hand holding in the presentation. Below I try to provide some examples to help guide the resubmission. At the end, I also point out some areas where additional analyses could be useful.

As indicated above, we wanted to be as concise as possible, which probably led to the confusion.

First, the method section must come before the results and discussion as one must read the methods to make sense of the results For example, parameter set #3 is referred to in the results without a description of what that is and why it may be important. Additionally, there are definitions that are in the methods that are needed to understand the results (ie “target runs”, “competition intensity”).

We now use a more classical format with the methods first.

Second, in many descriptions there is an absence of explicit purpose which seems to affect the structure of the manuscript. That is, it is not always clear what is trying to be accomplished with each section and several sections that is necessary. This is true for the overall paper as well. By the end it is somewhat that the purpose is to introduce and validate a method to assess transmission dynamics from epidemiological data given potential coinfections (this is my inference anyway), but it is difficult to infer this from the abstract or the introduction.

We have tried to clarify the purpose of the study, which is to validate a new method and compare it to an existing one.

Another potential point or subpoint is that there is information that cannot be used by other methods so you have less precise results. Anyway, being more explicit throughout will help structure the manuscript and help the reader.

About this point, we apologize for the lack of clarity. All methods are allowed to use the same data (prevalence of each type of coinfection in the population). This data is "digested" into summary statistics, which may appear as generating more information but the raw data is still the same.

One of the major issues seems to be a lack of precision in the descriptions, again throughout. Take the first paragraph of the Discussion as an example ("This is due to the fact that when sharing a host, parasites can interact in various ways [42]. The goal of this study was to determine to what extent the prevalence of parasite combinations can inform us on such interactions."). It is hard to get a solid footing on what the interactions are, what is meant by parasite combinations, and what information you are looking for.

We now detail different types of within-host interactions described between parasites (e.g. competition for resources, immune mediated apparent competition, competition for public goods). However, it is important to stress that we only have access to the average of all the interactions taking place between parasite genotypes in a host.

The expression "parasite combination" referred to the set of genotypes present in a host. We have gone through the manuscript to homogenize the wording and now refer to the prevalences of the combinations of genotypes.

Finally, the term information was perhaps too vague. Our goal is to detect statistically significant within-host interactions.

Similarly, better descriptions of the figures in the text, especially with regard to what the reader is supposed to learn from them, is essential. For example, Fig 3 is intense but it is never explicitly stated what the take home message of the figure is.

Figure 3 was indeed intense since it tried to summarise the performance of 3 existing methods on simulated datasets with known values.

To simplify the reading, we now only report the best of the three methods (which happens to be the most original one). This allows us to explain the results in more details.

Furthermore, we now compare a setting with only a single type of hosts (as assumed implicitely by these classical models) to show that they perform much better in such settings.

I am also not clear on what the numbers in the circles of Fig 2A are or what is meant by "different prevalences" mean.

Circles were supposed to indicate a prevalence measure that can be used as a summary statistics. These prevalences can refer to that of specific combinations of genotypes (in the middle), host ranks / number coinfectinge genotypes (on the left) or total genoytpe prevalence (on the top).

The methods are overall pretty clear, but more hand-holding would be helpful. For example, the summary statistics used seem important – spending greater time describing them and what we learn from the different types would be useful. In addition, as the ABC is the centerpiece, a fuller description is probably warranted.

Summary statistics are indeed an important aspect of any ABC approach and we now present them more thoroughly. Regarding the ABC, this is more delicate because we do not think this is the best paper to get into the technical details. However, we did try to improve the explanation of the essence of the method.

Some comments on the science itself.

The ABC is used to infer parameters of the model, but what if you assume the incorrect underlying model, is it substantially worse than the other heuristic methods where there are fewer assumptions. I would like to see what occurs if you assumptions about the epidemiology are incorrect.

This is indeed a limitation of the ABC approach we used because we do not compare between models.

To address this concern, we used our learning dataset (which assumes two host types) to analyse target datasets generated using models that assume a single host type. We show that this does not affect our ability to infer the competition parameter.

We also mention in the discussion other ways in which we could explore this further. One possibility would be to use different ABC regression methods (e.g. random forests) to compare between models (i.e. our model with 2 host types to a model with a single host type). Another possibility would be to use a much more detailed individual-based model to simulate the data, while keeping our simple model with two host types to perform the analysis. Unfortunately, both of these seem beyond the scope of this study.

Can you explain how "we assumed that interactions between HPV types take place through the recovery rates." affects either model or biological inference? I think you may mean how coinfection affects clearance rates, but I still cannot quite figure out how this impacts the interpretation.

In our model, within-host interactions are not described explicitly. Instead, we have a within-host black box and these interactions only affect epidemiological parameters. Since we neglect virulence, an infected host is only defined by two parameters: the rate at which it transmits its viruses and the rate at which it clears it. In theory, interactions between coinfecting genotypes could affect both their transmission rates or their clearance rates. Here, we assume this is the latter for biological reasons detailed in the article. It is not obvious for us to predict whether assuming an interaction through the transmission rates would change the results or not... The main reason to assume that it should not is that in the end we are working with equilibrium prevalences to decreasing transmission rate or increasing clearance rate should even out.

It is not clear what happens to the clearance rates if there are 3 strains present. That is, is recovery (1+k)^2 or is it still 1+k.

We now clarified it. The clearance rate is unchanged: it remains 1+k if a low competitor is coinfecting a host with one or more better competitors.

I don't understand why a in (2) denotes assortment between host types rather than within host types

Sorry, this is an English mistake: we did mean within host types.

What is the upper case delta in the updated master equation (w/ two host classes)?

Well spotted: \Delta should not be there because it is the matrix refering to host deaths (which are assumed to be 0 here).

Some other things to consider

Why the paper is centered on HPV is not clear to me. It applies to many potential pathogens and you did not actually use any data from HPV.

A first reason is that there is currently a debate as to whether different HPV types interact when coinfecting a host.

A second reason is that this is one of the only biological examples we know of, where this method could be applied. Indeed, most datasets do not have the sampling power to perform the inference. We do know this data exist but basically to analyse it we need to first show we have the power to infer something out of it.

The sentence describing 6D is confusing.

We divided the sentence into several sentences to clarify our interpretation.

A definition of what a significant association between parasites is should be explicitly stated in the results section for each test. Currently it is a bit confusing "Depending on whether k is greater or lower than 1, we expect host classes containing genotypes from the second group to be under- or over-represented respectively"

We apologize for the mistake because 0 should be the threshold value instead of 1.

Contrarily to other methods, we allow for interactions between genotypes to be facilitating or competing (this is given by the sign of k). Depending the scenario we are in, we expect some genotypes to be favoured by coinfections (and to be more prevalent than if there were no interactions), whereas in the other ot will be the opposite.

- why is there not a positive correlation between interaction intensity and the probability that the test is significant for the combination network?

In general, the combination network was rarely significant (at most 30% of the 100 test values). We cannot speculate as to why the effect goes in the unexpected direction. However, this is the first time this method has been tested with a dataset simulated from an epidemiological model so it validity to detect interactions can be questioned.

We have removed the results on the networks because in addition to not performing well they were very slow to run.

Reviewed by anonymous reviewer, 2018-03-12 19:49

This paper presents a model along with methods to infer the strength and direction of within-host interactions of coinfecting parasites from epidemiological data. The manuscript generates simulated data sets using an explicit epidemiological model, of the SIS type that allows multiple parasite genotypes and cotransmission (unlike previous models). The text implements previously published tests on data generated from the SIS model and then proceeds to implement an ABC regression to

infer interactions between parasite genotypes. The key results are that 1) using an mechanistic epidemiological model improves the inference of previous heuristic methods, but that their ability to infer parasite interactions are limited in the presence of host heterogeneity and 2) that including information on the parasite genotype combinations in addition to rank (i.e. genotypes per host) and host heterogeneity (i.e. superspreaders or not) greatly improves inference of within host interactions from epidemiological data. I note that my expertise in epidemiological theory is very limited and I cannot assess the technical merit of the models. Therefore, I deliver this review from the perspective of an (very interested) experimentalist.

Thanks for the accurate summary.

This paper tackles a very interesting subject matter. As someone that studies viral coinfections as a main focus, it is really exciting to see that viral interactions can, in theory, be inferred from epidemiological data. Moreover, it is very gratifying to see that there are researchers trying to do just that. In certain sections, the manuscript does a very good job of explaining the methods and results, but these are sometimes not in the right place within the manuscript (see Major comments below). The paper takes a very even-handed and judicious approach, by giving proper credit/inclusion to prior work and opting to model data that can be realistically obtained, respectively. The paper represents another leap forward in rebooting epidemiological modeling to include the complexities of viral coinfection.

Thanks

This manuscript could be greatly improved by guiding the reader more through the results and their biological interpretation, given the selected Intro/Results/Discussion/Methods format. In particular, the results section should include brief, basic methodological explanations and state the biological interpretation of results to guide readers. The most important example of this is the nature of the viral interaction being modeled (see Major Comment #2). The paper has a number of findings that are of great relevance to clinicians and experimentalists, but they are not mentioned in the discussion. I think this undersells the impact of this work beyond epidemiological modeling. Additionally, there are some issues with the figures that warrant attention.

As pointed out in our reply to reviewer #1, we have rewritten the manuscript to improve clarity.

Overall, I find the work to be highly interesting and potentially impactful, but the text hard to follow in some places. Providing more clarity will allow the paper to highlight and communicate these exciting results. This paper makes seamless connections between basic viral biology and its epidemiological consequences, tying them together with realistic models, thus illustrating the fundamental utility of biological modeling. More generally, I think this paper continues to push modeling towards embracing the full range of interactions among coinfecting viruses, which empirical studies are increasingly uncovering. In doing so, it provides crucial tools for experimentalists to focus their efforts and collaborate with theoreticians to better understand coinfections and their epidemiological consequences. With some revisions, this paper will be an outstanding contribution to the epidemiological literature and to the emerging coinfection (viral and parasite) research community.

Thanks.

Major comments: 1. This paper follows an Introduction/Results format, which usually is used to appeal to a broader readership by guiding readers through the results including the most essential background on methodological procedures. However, a lot of this basic information is not present in the results, but appears in the methods. These essential bits of information should be included in the results, even if they make the Results text longer, because they serve to guide the reader that is outside the subfield that is unlikely to peruse the methods in detail.

There are many examples of brief, elegant explanations that in the methods text that would have greatly aided the understanding of the results. A few examples of these explanations in the methods that should be in the results:
-Line 323 First sentence
-Lines 334-337
-A summary of lines 352-370, especially the three sets of summary statistics and the rationale for selecting them.
-Line 303-306 (see #2 below)
-Lines 341-346, especially the statement "...simulating many datasets, for which by definition the underlying parameters are known, and comparing them to the target dataset the parameters of which we want to estimate."

This was also raised by the Editor. It turns out nowadays a majority of journals seem to prefer to have the model at the end (even one like *PLoS Computational Biology*!) but we agree that here having it first helps the reading.

The interactions between the genotypes that are a focus of this study were initially expressed as "...we assume that any interaction between HR and LR types takes place through the recovery rate." Upon first reading it was not clear to me what was meant by this statement. I believe it corresponds to the following example given in the methods: "e.g. how the presence of genotype A affect the clearance rate of genotype B."

It is very, very important that this crucial interaction (the main focus of the study!) be crystal clear to the reader, upon first mention - more so with such a good, brief example explanation provided already in the methods.

This is a good point and we now give a real example to clarify this interaction.

There are some inconsistencies in how terms or parameters are mentioned in the text versus the figures. Rectifying these inconsistencies can greatly improve the readability of the paper. See comments below on Figures S1, 4, 5, and 6.

We apologize for these and corrected the figures.

The discussion seemed very centered around the modeling, but contained little information for epidemiologists or experimentalists. This despite the fact that the model results have practical implications that could advance research; for instance, the need for sensitive testing that distinguishes multiple HPV types.

This was due to our initial idea to focus more on the methods.

We now better discuss the implications to detect such interactions in the field, especially in systems other than HPV.

The use of grayscale in several graphs limited contrast and hindered the interpretations offered in the text. I suggest different shapes or colors as alternatives (see below for specific comments).

We now added colors.

Minor Comments: Line 49 - "specific functional response" is vague and leaves the reader guessing. Response of the host to the parasite? Response of the parasite to coinfection? I gather from the references cited that the text is attempting to refer to biological interactions broadly speaking, but not sure.

We now expand on this a bit more for clarity.

Line 56 - Suggestion: "This is consistent with a key result of the study, which identifies the ‘number of lifetime sex..."

Thanks

Line 95 - “have a competitive advantage (or disadvantage) when competing with non-oncogenic types”

Since this is a hypothesis, it is very confusing for some readers to put in parentheses the opposite prediction within the same statement.

The reason for this is that we investigate both possibilities in our model (since *k* can be positive or negative).

Since this is indeed misleading, we now refer more generally to interactions.

Line 97 - “interaction between HR and LR types takes place through the recovery rate.”

A little explanation of the biological implication of the interaction occurring through the recovery rate would greatly aid readers not steeped in epidemiology. I presume that this assumption means that the interaction between HR and LR occurs via the immune system as the host will have been infected and recovered before infection with the next type. It is important to have this point be clear because it goes to one of the key aspects of the paper, the interaction between the viruses.

Our problem here is that such interactions have not been demonstrated yet (partly because of the methodological issues we raise in this study). So although we do think there could be immune-mediated competition, this could be interpreted as very provocative from HPV experts.

That being said, we do know that there is evidence for cross-immunity between different HPV types (not necessarily HR or LR though).

There is also evidence of interaction at the cellular level between different HPV types.

We added a paragraph that explain this issue in details and summarises the current evidence available supporting the existence for within-host interactions.

Line 110 - “These have been tested by generating distributions but without any epidemiological model.”

A brief description about how these distributions were generated would guide the reader. For example - Were the distributions generated using a statistical distribution? This would aid readers not familiar with the literature.

In general, these earlier methods follow a similar approach than ours. They assumed that the observed prevalence data was following a specific distribution in absence of interaction. They then tested their ability to detect significant differences compared to this expected “neutral” distribution.

We tried to clarify this presentation.

Figure 3 - The grayscale is very difficult to distinguish.

We now use ggplot colors.

Line 117 - This paragraph would benefit from a wrap up sentence that explains the statistical result in light of what is being investigated. What does "most combinations lead to significant tests" mean? My understanding is that with a 1k sample size (and more so with a 5k sample size) this test can detect interactions even when their intensity is low.

We did not want to be too harsh but in general the probability to detect an interaction seems independent of the intensity of the interaction, which is not ideal, and seems strongly dependent on the dataset size, which is even more problematic.

As indicated above, we removed the results based on the network approach.

Line 118 - "the positive association between interaction intensity and test significance". Again, what this statistical finding means should be outlined. I would presume that if interaction intensity is high, it would be more easily detected by any test. So ideally one would want a test that could deliver significant results even at a low interaction intensity, correct?

We clarified the text because in the initial version we did not explain sufficiently well that all these tests were doing was testing whether the observed distribution of prevalence of each genotype combination was statistically different or not from a "neutral" expectation. In other words, the only result these tests can return is being significant or not.

Line 122 - Brief explanation of connectance may be beneficial to the reader. Most readers will understand ChiSq significance pretty intuitively (as a departure of expected values, which suggests an interaction between the parasite types), but connectance in this context may be less familiar to many readers.

Upon reading the methods there is an elegant and brief explanation of connectance that should be included in this section of the text: "that is the proportion of observed edges relative to the number of edges. Here, individuals are connected if they share the same 336 parasite (parasite network) or the same combination of parasites (combination network)." Similarly, the association screening approach explanation from the methods can be also be placed in this section.

We agree and have now moved the methods earlier in the manuscript.

Line 122-135 - Combination network (earlier called coinfection combination network) and parasite network are dropped in this paragraph with no definition whatsoever. My understanding until reading this statement was that the data being discussed were parasite coinfection combinations.

Hopefully having the methods first has improved this.

Figure S1 - The text mentions this figure depicts the "correlation between interaction intensity and the prevalence of each host combination", but the figure states it has the "Correlation between interaction intensity and combination, rank or genotype prevalence". Presumably, Class# corresponds to host combination, Rank# corresponds to rank, and Tot# corresponds to genotype prevalence. These labels should be clearly indicated in the figure legend and ideally would correspond to the names in the text, i.e. Prev# for genotype prevalence and Combination# for host combination.

This interpretation of the figure is correct (we indeed added the ranks and total prevalences). This is now clarified in the caption.

Line 141 - This sentence should be reworded to avoid the use of prediction twice: "The fraction of predictions that match our expectation is generally close" OR "The fraction of correct predictions is generally close" (Also sentence needs a period)

Our apologies for the hasty writing.

Line 144 - contact structure is sufficient to blur the effect of within-host interactions OR is sufficient to blur the ability of this test to detect within host interactions? I thought the contact structure and within-host interactions are set (or at least constant) in these simulated data sets, no?

Here we meant contact structure between the two host types (i.e. via the assortativity parameter) and not a true network contact structure.

Line 148 - k should be defined in the text at first appearance. I see k is defined in Figure 5, but this text is referencing Figure 4, where k is not defined either.

We now have the methods first but also remind the reader what k is.

Line 149-152 and Figure 4 - The grayscale that indicates the different ranks cannot be distinguished (even on a high resolution screen), making it difficult to see the clear pattern described in the text. This is a really important figure in the paper so the ranks should be more distinct. Different symbols would work well if non-color figures are desired.

We now have colors.

Line 154 - Suggestion: "... same prevalence in single infection (see rank 1 data points)."

Thanks!

Line 157 - At the end of the paragraph, the text states the goal is to infer parameter k, but at the beginning of the paragraph there are two graphs with known k parameters. I did not understand this discrepancy. Is it that the first model doesn't have HR and LR genotypes? Some clarification will aid the reader. Upon reading the methods, the statement on line 343 greatly clarifies this paragraph: "It consists in simulating many datasets, for which by definition the underlying parameters are known, and comparing them to the target dataset the parameters of which we want to estimate." Including this statement in the results would broaden accessibility of the text to readers outside the field. This is a general comment throughout the manuscript that I outline in Major Comments.

We apologize for the lack of clarity and have now moved the methods up front.

Line 162 - TYPO: "We assessed the performance of.." (no 's' after performance)

Thank your for the careful reading.

Line 165 - This paragraph, together with the figure is well written and the results are very clear.

Thanks!

Line 175 - It is not clear to me what is meant by "runs". Readers may benefit from a little more explanation regarding how analyses differ.

We rephrased and now refer to it as a single inference.

Line 198 - If "proportion of errors" in the text is the same as "error probability" in Figure 6D, as I believe it is, the text and figure should match each other for increased clarity.

We now always refer to the probability also in the text.

Line 228 - These seem like important results, that (contrary to the statement) are being reported in the discussion and should be included in the results. Moreover, they are important results that justify the focus of the paper on k, which increases the accuracy of inference more than many other parameters. Minimally they should be included in a supplement.

We added a supplementary figure to show our ability to infer the other parameters of the model for the run shown in Figure 5.

Line 303 - Again, this statement "interactions between HPV types take place through the recovery rates" is not very clear in biological terms to me. Now reading further, if this statement: "how the presence of genotype A affect the clearance rate of genotype B" is what is meant, this should be stated in the methods and the main text. This is a crucial explanation of the biological process that is being modeled and is not clear in the text.

Again, we apologize for the lack of clarity originating from having the methods at the end.

Reviewed by Erick Gagne, 2018-03-22 06:20

This manuscript sets out to test a potential important improvement on SIS models that incorporate co-infection information. This is a significant contribution as it is known that co-infections of multiple pathogens, or multiple genotypes of the same pathogens, can influence transmission, susceptibility, and virulence. This can be particularly important for pathogens such as HPV, where different genotypes can have a wide variety of disease outcomes, and this is a good choice of focus for the manuscript.

The models being developed seem appropriate (see some comments below) and to be an improvement on existing approaches. I think these results are worthy of publication. I particularly found the testing of existing methods followed by the implementation and testing of Approximate Bayesian Computation (ABC) an appropriate and informative approach that illustrates the current shortcomings and the benefit of the ABC modeling well. However, a shortcoming of my review is that, although I am a disease ecologist with experience in many of the concepts of the manuscript, my work relies heavily upon molecular techniques and not this type of modeling. I have experience implementing similar models (ABC with genetic data) but not developing them.

Thank you for this accurate summary.

A major area of revision that I recommend prior to submission for publication is the focus of the introduction. The manuscript touches on many relevant points and outlines some clear arguments but these could be better linked and structured. The problem and focus of the manuscript should be introduced earlier. This will allow the reader to follow why the specific arguments and discussion of existing modeling approaches are being outlined. In addition to the general structure, the introduction needs to be revised for clarity. I recognize this is a pre-print and the main objective is to receive feedback on the study so this is not a major concern for now. The opening sentence "With the advent of next generation sequencing, an increasing number of infections turn out to be coinfections by multiple genotypes" makes it sound like NGS is resulting in an increased number of coinfection, when it means NGS is allowing the discovery of most infections being coinfections.

We have tried to clarify the introduction with regards to the state-of-the-art and improve the link with existing data.

In other places, it is unclear if multiple infections/coinfections is referring to multiple pathogens or multiple genotypes of the same pathogen. For example, the second paragraph in which they define

multiple infections as multiple genotypes but then provide an example of HIV and malaria (i.e., multiple pathogens).

This is an excellent point, which we needed to clarify. The main answer is that the type of focus (same species or different species) should affect the underlying epidemiological model. For instance, the one we used here is appropriate for various genotypes of HPV but to focus on HIV-malaria coinfections, it would have to be completely redone.

The summary sentence of the introduction (Lines 34-36) needs to be refined for clarity. For example, it is not clear from the rest of the introduction what is meant by "but it is unclear whether these interactions are sufficiently strong to be detected at the population level." Do the authors mean it is unclear if these interactions can be diagnostically detected or if these interactions have biological effects at the population level.

We believe that these two answers are linked. Currently, there is a methological problem that makes it difficult to detect these interactions even if they exist (this is what we aim at addressing with our new method). Once we have demonstrated our power to infer existing interactions, we will be able to analyse real datasets to see if interactions occur. This is why we focused on HPV because the existence of interactions is strongly debated.

The sections discussing current modeling approaches would benefit from being reworked to have better conclusions of each section- as written, the pitfalls and benefits of each method are not clear. For example, the "parasite combinations" section concludes abruptly (Line 79-80). I recommend explaining why the lack of an explicit epidemiological model is problematic and what important information these current models lack. The issues I highlight above relate to a general issue with the clarity of the writing throughout (with the exception that the methods section is well written).

We partly addressed this by moving the methods before the results. We also only present one of the methods in the main text for clarity.

If the manuscript is being submitted to a journal in which the methods come after the results (as is the current structure), the results need a bit more information so the reader can follow.

We agree with this and made the change.

A minor concern with the model that I have is in regards to not considering vaccinations. The fact that natural immunity is low but vaccine immunity is high makes it is unclear why immunization was not included in the model. This would seem to be an important component-specifically as vaccination against specific genotypes could presumably increase frequency of other genotypes.

About HPV, we did not want to include vaccination for several reasons. First, the existing datasets that could already be analysed have been collected prior to vaccinations (they were usually the control arm of the vaccine). Second, expect for a few countries (Australia, the USA), vaccine coverage is low (France is below 20%). In fact, countries with limited resources still do not have access to the vaccines.

Introducing vaccination in the model could already be done by assuming that our two host types correspond to vaccinated and unvaccinated hosts (instead of superspreaders and normalspreaders). This would only require a minimal adjustment of the model. A more detailed possibility would be to extend the model to 4 host types but this would be challenging using our ODE-based model and would likely require an individual based approach.

My major take away is that the analyses are well done and represent a worthwhile contribution to the literature. A careful revision of the text will greatly strengthen the work and is recommended before publication.

Thanks!

A few minor comments:

Line 28: “virus loads also seem to be differ” Should be ‘seem to differ’ or ‘seem to be different’

Done.

Line 47-49: For submission to most journals, further explanation is needed as to why a negative binomial distribution is thought to indicate host population structure or a specific functional response. In addition, “a specific functional response” is vague and needs further explanation.

We now try to describe this in a bit more details but there are years of parasitology research on the topic, which makes it difficult.

Line 109-110: Revise for clarity “First we use existing methods developed to detect significant associations between parasites from coinfection data.”

We now used a less complicated formulation.

Line 228-229: “We do not report it here but the accuracy of the inference varied widely across parameters.” This is interesting and worth reporting in the results.

As indicated to reviewer #2, we now report the distribution for the other parameters in Figure 5.