Many thanks for accepting to recommend our article and for the helpful comments. As detailed below, the main change concerns the discussion, which was largely rewritten, and several shortened and rearranged passages. Our answers are in green italic font, quotes from the manuscript within quotation marks in green normal font. Line numbers refer to the revised manuscript, except if stated differently.

Reviews

Reviewed by anonymous reviewer, 17 Feb 2023 14:01

This is a very interesting and relevant review about MTBC evolution. The authors clearly state the assumptions of population genetics models and which ones are met by the known biology and evolution of MTBC. Nevertheless, I think the flow and structure of the article can be improved.

For example, linkage and linked selection is mentioned in the introduction, but in the section on positive selection it is not clear which consequences this has for the inference of selection.

The passage on linked selection was removed from the introduction. The challenge with linked selection is that, while crucial from a theoretical point of view, it is virtually absent in the empirical literature we discuss. Potentially relevant mechanisms of linked selection (Muller's ratchet & background selection) are now considered in the discussion, l. 616-659, where we make an argument for background selection as a potential key mechanism in the MTBC.

Regarding the consequences of linked selection on the inference of selection, we briefly touch upon this issue and refer to a review (l. 483f):

"Identifying signatures of positive selection in linked genomes is challenging since most tests rely on the comparison of haplotypes within genomes (Shapiro 2009)."

I also find it odd to start with recombination instead of mutation since the latter is the basic process generating diversity and all other processes act on this diversity. Of course, the rationale of the authors can be different, but then that must be clear in the flow of the text.

We agree that this order was unusual. Our idea was to start with recombination because extreme clonality and lack of HGT is the hallmark of monomorphic bacteria and a premise of the review. We now swapped the two sections because it indeed feels more logical and fits with the order in which the processes are mentioned in the introduction.

It is also unclear why genetic drift and purifying selection are addressed in the same section (and positive selection in a different one) although drift interferes with selection independent of the direction of selection. Also, this results in dN/dS being introduced twice (lines 368 and 616).

Our original intention was indeed to have separate sections on drift and selection, which would be cleaner conceptually. However, from the empirical standpoint this proved difficult because genetic drift and purifying selection are usually discussed together, while papers interested in detecting *positive selection rarely care about purifying selection. A sentence was added in the introduction to genetic drift in order to justify this procedure (l. 292-295):*

“Increased genetic drift thus implies reduced purifying selection, and the same genomic evidence (see below) underlies claims as to the relative importance of the two processes. For this reason genetic drift and purifying selection are treated together, while a separate section is dedicated to positive selection.“

The second passage on dN/dS was substantially shortened to avoid repetition. We now mainly point out the difference between genome-wide dN/dS based on pairwise comparisons, which underlies the argument for increased drift / relaxed purifying selection, and the more sophisticated models used to detect positive selection.

The manuscript is also a bit lengthy and I would suggest to shorten sections that are not directly relevant for MTBC or for clonal evolution, e.g., the section on DCT starting on line 112.

We went through the manuscript and shortened/removed different sections (line numbers refer to the first draft):

- *introduction: two paragraphs on linked selection were removed (l. 51-70)*
- *details on DCT were removed (l. 112-116)*
- *basics of genetic drift were shortened (l. 340-358)*
- *basics on dN/dS as a signature of positive selection were shortened (l. 616-631)*
- *simulations: the box was removed, parts were recycled in the discussion*

Due to the extended discussion, the revised manuscript ended up being almost as long as the original. We hope that it feels less lengthy as unnecessary details were replaced with a more interesting, broadened discussion.

The translation of the per-site mutation rate into the per-year mutation rate is very simplistic (line 220). Does this assume that all mutations are neutral and how are dynamics in the population included? The authors correctly distinguish between the mutation rate and the molecular clock rate earlier in the manuscript (line 191), so they should also correctly distinguish them throughout the manuscript.

In this context, the translation from generations to years serves to make the simple point that the MTBC evolves slowly if one considers its long generation time. This passage was rewritten without reference to a rate per year, which is not necessary in this context (l. 107-110):

“The bacteria of the MTBC have long generation times ranging from 18 h in nutrient rich medium to potentially much longer time-spans in vivo (Colangeli et al., 2020). Scaled to clock time, mutation rates are thus indeed low in the MTBC compared to other bacteria, at least in the laboratory (Gibson et al., 2018).”

It is unclear which conclusions the authors want to state in the paragraph starting at line 439. They search for an explanation for the high dN/dS and invoke selection at synonymous sites. But then it is shown that there is evidence for positive selection at synonymous sites, which would results in low dN/dS and not high ones.

The sequence of arguments in this section was confusing. We now clarify that we do not agree with the argument for positive selection on synonymous sites, which rests on the assumption that intergenic sites evolve neutrally. Here the rewritten passage (l.381-388):

“ Under the assumption that intergenic sites are free from selective pressures, Wang & Chen conclude that synonymous sites are more diverse than expected by chance and therefore evolve under diversifying, that is, positive selection. Alternatively, and in line with the initial hypothesis of purifying selection at synonymous sites, higher synonymous than intergenic diversity is also expected when intergenic sites are even more constrained than synonymous sites. Intergenic regions in bacteria are packed with regulatory motives and can hardly be assumed to evolve neutrally (Molina and Van Nimwegen, 2008; Rocha, 2018). ”

The discussion section contains a highly relevant appeal for including proper simulations in data analysis. Nevertheless an actual discussion is missing. I suggest to add a discussion on the consequences of clonal evolution on MTBC genome evolution, for example on their genomic architecture, on the efficiency of selection, or on linkage and the distribution of fitness effects (what about epistasis?). Such a conclusion is really expected by the reader since the authors state in the introduction "In this review, we present the main hypotheses about what drives the evolution of the MTBC, and how they have been arrived at." (line 91) So, what are these main hypotheses?

We fully agree on this criticism, which is also pointed out by reviewer 2 (“a lack of a synthetic view on the main hypotheses”). The discussion was largely rewritten and divided into two parts (l.583-587):

“ In the following, we discuss a unifying scenario, the evolutionary optimum hypothesis, to connect the different threads laid bare above and to make a case for background selection as a key process in monomorphic bacterial pathogens. This speculative exercise is followed by a discussion of simulations as a key tool to transition to a more quantitative understanding of evolutionary dynamics under extreme clonality.”

In the first part, the two main hypotheses in the MTBC literature, strong genetic drift versus strong purifying selection, are discussed in terms of linked selection, that is, Muller’s ratchet and background selection, and we present an argument against Muller’s ratchet and for effective purifying selection.

Further comments:
line 239: It is unclear what the authors want to say with the sentence starting at line 239. How it the AT bias reflected by stress-induced mutagenesis and how does this relate to the GC rich genome?

This sentence was removed, it contained unnecessary detail.

line 605: I would suggest to remove "human" from the sentence to focus on MTBC migration instead. An explanation might be added that this is driven by human migration.

“Human” was removed from the sentence.

line 691: It would be easier to follow, if the purpose of the simulation was mentioned, before the simulation details are described.

As stated above, the discussion section was restructured: the simulation part is now contained in an outlook section after the discussion. The simulation is now introduced as follows (l. 675-680):

“To conclude this review, we present an exemplary simulation that captures some realistic aspects of the within-host population dynamics of a clonal pathogen (script and detailed description on https://doi.org/10.5281/zenodo.8042695). Such simulations could be used to better understand the patterns of genetic variation expected in an infected individual, and the bias introduced through punctual sampling of a structured population and culturing (Morales-Arce et al., 2021).”

Fig 4c,d: What are the solid and dashed boxes?

Solid and dashed boxes show results for simulations for two different selection coefficients: s = 0 (solid), s=9.5e-4 (dashed). This information was added to the figure legend.

Line 710: Where can Fig. B1 be found?

The correct reference is Figure 4c; this was changed. Also on lines 699 and 704 figure legends were corrected as they were from a previous version of the manuscript.

Review by anonymous reviewer, 17 Apr 2023 06:28

In their review article, Stritt and Gagneux provide a thorough and well-written analysis of the literature on the evolution of the Mycobacterium tuberculosis complex (MTBC), a monomorphic bacteria. This bacterial species complex exhibits a large genome size, high GC content and low level of genetic diversity, those genomic features occur in the context of a slow growth rate and clonaliy (i.e. lack of HGT). The authors highlight the challenges presented by the extreme clonality of the MTBC and other monomorphic bacterial populations, to investigate the evolution of their genetic characteristics. They provide a comprehensive overview of the literature investigating the basic evolutionary processes (i.e. recombination, mutation rates, genetic drift and selection) occurring in MTBC populations. The paper is actually divided into four sections corresponding to each of these processes and how they have been evidenced and sometimes quantified in MTBC. In each part, the authors raise several questions. The authors review experimental and empirical studies of the MTBC complex that have attempted to answer these questions. They provide clear and sometimes detailed explanations of the models/hypotheses that have been put forwrad to explain the evolution of these monomorphic bacteria, and then discuss the limitations of these models. Doing so they underline that some studies are may be “too” narrative, without clear hypothesis testing. The authors highlight a key area for future investigations: simulation studies and propose an existing tool to conduct such studies. The paper is somehow unusual in its form even for a review paper (with a discussion proposing a protocols an an analytical tool for going forward), but it is not a problem as it is the purpose of journals/platform such as PCI to have papers that do not follow “preformatted” guidelines. Overall, Stritt and Gagneux have produced a well-written and comprehensive review of the literature on the evolution of monomorphic bacteria, with a focus on the MTBC. Altogether, it provides a valuable resource for researchers studying the evolution of monomorphic bacterial populations and underlines weaknesses in the analytical tools used .
I have nevertheless, a few suggestions to improve the manuscript and may be make it useful for a larger audience.

Two general comments.

First, for a review paper, there is in my opinion a lack of synthetic view on the main hypotheses found in the literature on how these genomes evolve, the predictions derived from these hypotheses (and how or whether these predictions have been thoroughly tested). Overall I gathered that: lack of HGT but intra-chromosomal recombination has been demonstrated (inferred from experimental evidence and population genomic investigation on MTBC); rates of evolution have been measured and are rather slow but actually very variable according studies and many studies could suffer from methodological studies; strong genetic drift is often referred to but difficult to measure (with Dn/ds quantification or estimation of Ne); positive selection (on resistance genes) is recurrently evidenced. Can the authors come up with a table summarizing studies that have put forward clear hypotheses and predictions or attempted to quantitatively estimate each factor (may be for each section)?

For instance:

-lack of HGT: list of studies with experimental evidence, list of studies with empirical evidence (pop genomic studies), which dataset they have worked on and what they have concluded.

-low rate of evolution: a table summarizing studies that have measured this rate with which method (experimental work, Beast estimates) and on which dataset. (I guess figure 3 does the job but it should come earlier in the paragraph)

-strong genetic drift: prediction (low Ne, overabundant nonsynonymous polymorphism), list of studies estimating dn/ds to estimate genetic drift in MTBC, measuring Ne using Bayesian skylineplots (again dataset used, methods and their conclusions…)

And etc.. for positive selection. May be the literature is too all over the place to make such a table… It's a suggestion to shorten the paper : a lot of what is specified in the text could then be summarized in the table and the text would be more about the methodological issues of the approaches presented.

The problem we see with such a summary table is that there are indeed very few studies that "have put forward clear hypotheses and predictions or attempted to quantitatively estimate each factor". As we tried to convey in the original manuscript, only few studies in the large MTBC literature focus on basic evolutionary questions.

We now try to make up for the missing overview in the revised discussion (l. 579ff), where we revisit the drift versus selection debate in the context of the evolutionary optimum hypothesis and linked selection (Muller's ratchet, background selection). We also discuss the hypothesis that low mutation rates constrain the evolution of the MTBC, which is at odds with the fast evolution of resistance within patients or rapid adaptation in experimental evolution studies.

Another general comment I have is that, the article is focused primarily on the MTBC, and for now it is difficult to see what is applicable to other monomorphic bacterial populations or other microorganisms. It would have been helpful to provide more discussion on how this review may be generalized to other systems. Many of the characteristics of MTBC populations (low GC contents, strong genetic drift, lack of HGT) reminded me of intracellular bacterial symbionts. Can the authors in the discussion widen their scope and argue how what we learn from MTBC population genetics is informative for other model species (which study also suffers from narratives that are not always put to the test)? Could they specify which of their recommendation (e.g. question the use of Bayesian skypline plots for estimating Ne) applies to "all"/ "most" bacterial population studies.

The rewritten discussion now offers a broader perspective: we discuss bacterial endosymbionts and another monomorphic bacterial pathogen, M. leprae, and how their peculiar genome compositions have been interpreted. More specifically, the lack of diversity in the >1000 pseudogenes of M. leprae is presented as evidence against Muller's ratchet in monomorphic bacterial pathogens.

The approach of this review is to use the MTBC as a model for monomorphic bacterial pathogens because it is the most diverse and the most extensively studied among them. The MTBC can thus be used to probe deeper than it is possible in other highly clonal organisms.

Our recommendations that modeling assumptions should be treated more critically, and that ideally simulations should be used to test intuitions and methods, hold for any empirical study making claims about mutation, recombination, genetic drift, and natural selection. We think that they are *particularly relevant in the field of bacterial pathogens, where evolution is primarily of interest in the context of antibiotic resistance and evolutionary concepts are often used superficially.*

Specific comments

Throughout the manuscript, it is not always easy to follow what are generalities (for population genomics study of bacteria, populations genomics in general) or specific to the study system. Make sure that it is clear when the study you are citing is focused on MTBC.

We could not quite figure out which passages this comment is referring to. Paragraphs usually contain the necessary context ("in the MTBC", "in M. canettii", "in bacteria" etc).

Figue 2 is a nice summary of some genetic characteristics of MTBC, but it is difficult to understand how and on which data they have been built from, please give some sort of mat and meth, at least in a sup mat, with a link to a dataset.

The legend now refers to Zwyer et al. (2021), from where the SNP alignment underlying pairwise genetic distances was taken (Fig. 2a), and to Bobay & Ochman 2018, from where the data underlying the plots b to d was taken. This information was also added in the "Data and code availability" section. All data sets and the plotting code are now available on Zenodo (https://doi.org/10.5281/zenodo.8042695).

The paper is long, and sometimes gives too many details on very general matters that are interesting, well written, but can be a bit confusing as it distracts from the main messages. May be go through the paper and see how you can shorten non-essential paragraphs. For instance line 51 to 64 is a general paragraph on bacterial phylogenies which is not very useful in the rest of the paper (I don't think the authors come back to how "linked selection" has affected inferences on evolutionary processes explaining MTBC evolution).

We fully agree. As stated above, we shortened/removed different sections, including the one mentioned here on bacterial phylogenies. The paper is still long, but hopefully more focussed.

Less important but still , line 129: the authors mention that lack of HGT could be an adaptation to parasitism but do not further review a study that have investigated this hypothesis in MTBC..it's a bit "off track" in this paragraph.

This passage was rewritten (l.238-245):

"Rather than a mere side effect, as implied in the lack of opportunity hypothesis, absence of HGT could be an evolutionary strategy with a genetic basis. The predominance of clonality in a wide range of pathogenic organisms could indicate that clonality is adaptive by preventing the breakup of favorable allele combinations (Tibayrenc and Ayala, 2017). Further investigation into the genetic and environmental determinants of extreme clonality would be worthwhile, and the M. canettii-MTBC system provides a great opportunity to elucidate the poorly understood evolutionary transition to extreme clonality characteristic of many obligate pathogens."

We here wish to point out that extreme clonality can and should also be studied from an evolutionary perspective, which is largely missing in the MTBC literature. There the default explanation is "lack of opportunity", although this has not been studied systematically.

The definition of genetic drift from line 339 is a bit long (delete last sentence of the §), shorten the intro on genetic drift. I believe this review is intended to readers who are already familiar with population genetics concepts.

This section was shortened. The audience we hope to address includes epidemiologists, infection biologists and microbiologists who might lack a background in population genetics. But of course a proper introduction to this complex concept would require even more space. The references in the paragraph hopefully provide an entry point for the intrigued.

The § from line 629, on dN/dS measure and its caveats is also very general, it should either be deleted or come earlier, when dN/ds are mentioned in the context of estimating genetic drift.

This paragraph was shortened: we now focus on explaining the difference between dN/dS in the context of purifying versus positive selection, while other generalities were removed.

Very specific comments:
P12: Define genome erosion: genome size diminution or low coding density?

We think that the first sentence of the paragraph provides the necessary context (l. 352-354):

"Strong genetic drift leaves other signs than an excess of nonsynonymous mutations, including pseudogenization, proliferation of selfish genetic elements, or an increased proportion of transversions."

Line 432: you mean "positive selection at synonymous sites"?

We indeed mean purifying selection: a high genome-wide dN/dS might result when there are fewer synonymous mutations than expected under neutrality, that is, when purifying selection removes polymorphisms at synonymous sites. As also pointed out by reviewer 1, this section was confusing as we did not make clear that we disagree with Wang & Chen, who propose that there is positive selection on synonymous sites. This should now be clearer in the rewritten version (l. 381-388).

Line 719 "sputum": I guess you meant some?

"Sputum" was correct, it is a main material from which MTBC cells are isolated. This passage was rewritten without referring to sputum.