Dear authors,

First of all, I would like to apologise for the delay in the handling process of your preprint. Unfortunately we have not been able to receive one additional review we were expecting. We have now decided to move forward without it, but will let you know if we receive it in the coming days.

You will see, however, that two reviewers have been able to examine your preprint and their **evaluation is generally positive**. I agree with their assessment, and would like to congratulate you on this work. There are a small number of issues that require your attention before the preprint can be recommended; please see the comments attached.

Most of the feedback relates to improving the clarity and presentation of your results. A few comments have to do with providing phylogenetic support to some assumptions (same history of orthogroups selected for phylogenomics) and claims (evolutionary history of specific genes). If it's possible to provide these, they could strengthen some of the focal points in your analyses. Of particular importance are the comments concerning this article's Supplementary tables. Please make sure that these are accessible as part of your revision.

I would like to add that it would be helpful if you add to Fig. 1's legend the evolution model, in this case selected by MFP, used for the displayed phylogeny. Note that mixture models are generally useful to analyse long alignments, and these are not considered by the default Model Finder usage in IQTree. I recommend that you consider at least one such mixture model, perhaps based on the empirical model selected by MFP if appropriate, for your main phylogenomic tree. Additionally, please note that the cladogram in Fig. S5 has been incompletely printed.

Thank you very much for submitting your preprint for evaluation at PCI Microbiology.

With kind regards,

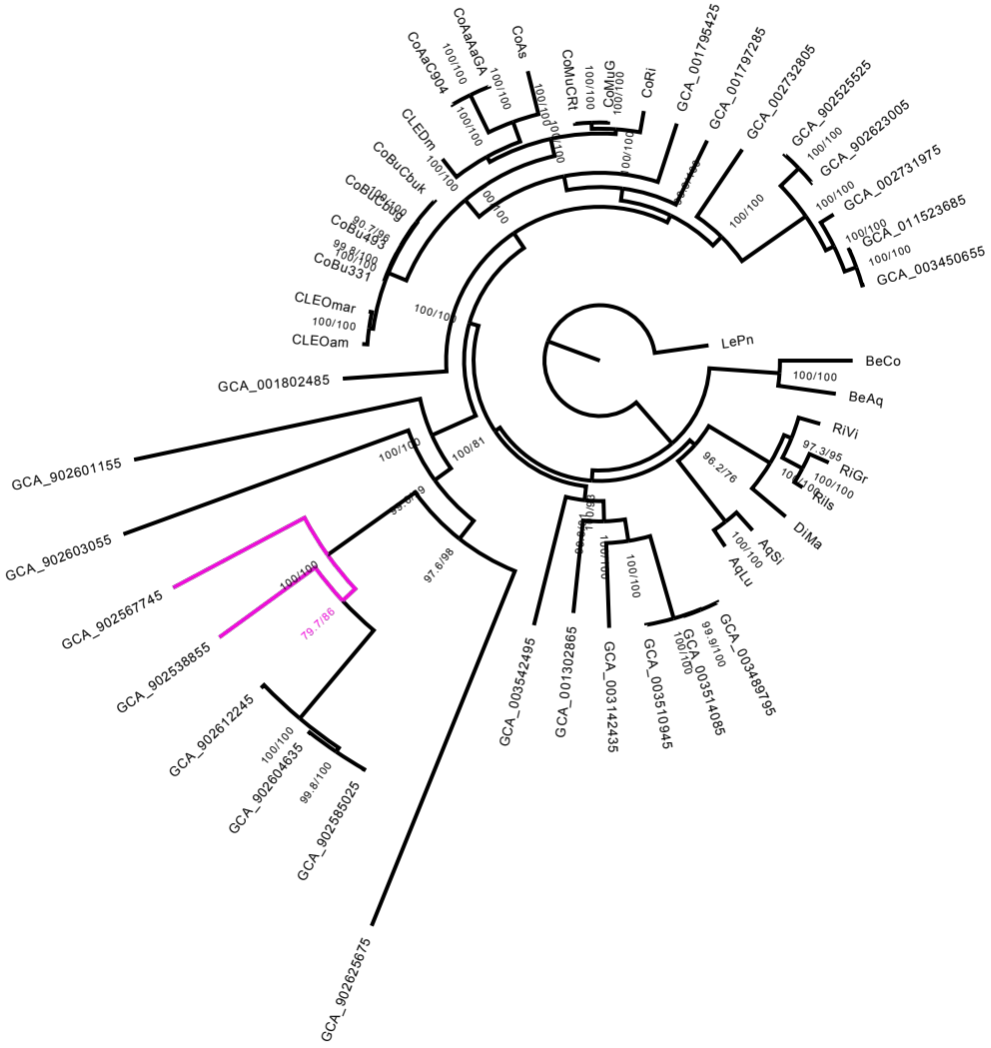
Daniel Tamarit

Dear Dr. Tamarit,

We appreciate the reviewer's constructive comments and believe the manuscript's overall quality has increased. We like to mention that we took out the molecular analysis results. Our decision was based on reviewers' comments and the fact that this part seemed out of the general arguments developed in the manuscript and it does not affect the main conclusions. In general, we tried to address most of the reviewer's concerns and suggestions. Our answers can be found below, We updated the manuscript version on BioArxiv, highlighting in blue any modification we introduced on the new version. We acknowledge the editor's comment to improve the species tree. Therefore, we ran IQTREE with ModelFinder and allowed different mixture models (-m MF -madd LgaM,LG4X,C10,C20,C30,C40,C50,C60). In addition, we ran a LG+C20 ML tree (tree below). The topology between the LG+R6 and LG+C20 was almost identical except for one bifurcation, inside a Coxiella MAGs cluster (purple branch), better resolved in the LG+C20. However, according to the Bayesian Information Criterion LG+R6 outperformed the tested mixture models. Also, this inner node does not affect the other analysis performed in work. Hence, we prefer to keep the originally submitted tree. Both three are available at https://itol.embl.de/shared/dsantosgarcia In addition, we added the model to the figure legend.

On behalf of all authors,

Diego



Review by **Sophie Abby**, 04 Dec 2022 18:30

In this article, Santos-Garcia and colleagues explore the evolutionary history of *Coxiella burnetti* and related symbionts using comparative genomics and phylogenomics, and by taking advantage of two newly sequenced genomes from *Coxiella*-like endosymbionts (Coxiella-LE) isolated from ticks. In particular, they investigate the evolution several specific traits that may explain the pathogenic VS endosymbiotic lifestyle of the *Coxiella burnetti* and *Coxiella*-LE (mainly vitamins B synthesis, T4SS and pH resistance). Overall, the article is clearly written and should be of good interest for the microbiologists community. The general context is well exposed, making appealing the read of the results section. The analyses are well-suited, and the claims in general well-supported, even though some details could be added (see my comments below). My main comment is that I believe the paper could greatly benefit from more explicit illustrations on evolutionary scenarios, for example by adding sketch species trees on main figures (as done for Fig. 5) where evolutionary history of key traits are discussed (Fig 3, 4, 6), and maybe even a conclusion figure on a tree summing up the different results and main evolutionary hypotheses.

Major comments

- Despite a very thorough supplementary material accessible at BiorXiv and at Figshare, I could not find the Sup Tables. Maybe have I overlooked it. Otherwise, could these tables also be made available via e.g. Figshare?

Our mistake. Supplementary Tables are now available in FigShare. Sorry for any inconvenience.

- For the phylogenomics analysis, it is unclear whereas individual gene trees of concatenated markers for the species tree construction were individually examined. Could the authors clarify whether/how the consistency of their evolutionary history was evaluated?

COPs used for phylogenomics were not manually evaluated. However, we think the COPs selected ensure a robust reconstructions. Indeed, we used those COPs selected by OrthoFinder. The STAG algorithm paper (https://www.biorxiv.org/content/10.1101/267914v1), although not peer-reviewed, seems to support the robustness of the method and COPs selected by Orthofinder to reconstruct species trees. According to the commands we used, the right section from Orthofinder's manual is: https://github.com/davidemms/OrthoFinder#multiple-sequence-alignment-species-tree-method--m-msa.

- When observing Fig. S2, I do not understand from where comes the support for the claim of the sentence on L. 265-267: "COG classification of the COPs showed that basic cellular processes, such as translation and transcription (J) or replication, recombination, and repair (L), along with other functions such as co-enzyme transport and metabolism (H), are over-represented in reduced Coxiella-LE genomes (Fig S2)."

Also, "over-represented" compared to which genome? Could the authors clarify this? Would a representation of the %age of COPs in COG categories rather than showing absolute numbers help here as the genomes have different sizes?

Maybe we used the wrong word. Indeed, those COG categories (most belonging to the informational machinery, transcription, translation, and some other essential cell processes) are not over-represented, just retained during the genome reduction process since they are generally required by the symbiont. We have tried to clarify that section and added the % of COPS to Figure S2. Now the sections reads as: "COG classification of the COPs showed that basic cellular processes, such as translation and transcription (J), replication, recombination, and repair (L), or post-translational modifications and chaperonines (O) are retained in reduced *Coxiella*-LE genomes compared to *C. burnetii* (Fig S2). Additionally, co-enzyme transport and metabolism (H) is also retained in reduced *Coxiella*-LE genomes (Fig S2), as already reported in other facultative symbionts suffering a genomic shrinkage and evolving towards a more obligatory status (Manzano-Marin2012)."

- In sentence on line 272-273: "The number of COPs involved in replication, recombination, and repair (COG category L, Fig S2) reflects genome size and is negatively correlated with both synonymous (dS) and non-synonymous (dN) substitution rates (Fig S4), especially the latter"

it is difficult to link dS and dN values on Fig S4 with genome size, as it is not directly displayed on the figure. Could this (genome size and number of COP in cat. L) be reminded e.g. next to the color legend for genomes? If a correlation has been computed, could the statistical details and values be reported here? And does it still hold when accounting for phylogenetic inertia? Also, from the plots in Fig S4, it seems that some of the discussed increases or decreases in values when comparing

genomes are not significant (e.g for the comparison on lines 279-280, same or overlapping statistical groups?). If double-checked, could the authors specify it in the main text?

We didn't perform a proper correlation analysis, we just highlighted a trend. We apologize for the wording used that meant some statistical support. Besides, almost all highly reduced *Coxiella*-LEs are associated with *Ambvlyomma* ticks and medium size with *Rhipicephalus* ticks. This issue makes us suspect that, indeed, it could be possible that phylogeny is a cofounder in those analysis. All that said, we decided to discard the molecular analysis (see our comment below).

- I am not really sure what it the overall conclusion drawn from the dN, dS analysis. Could the authors elaborate on e.g. what was hypothesized/expected versus what was obtained here?

The hypothesis, proposed by Gottlieb *et al*. 2015 (doi:10.1093/gbe/evv108), was that the loss of the mismatch repair genes *mutSL* could be one of the main factors driving genome difference among *Coxiella*-LE. The loss of *mutSL* in some *Coxiella*-LE lineages might have increased mutation rates, then, allowing a higher rate of genome erosion and smaller genomes.

Our data is in line with Gottlieb *et al*. 2015 proposal that smaller genomes without *mutSL* are evolving faster than larger genomes with *mutSL*. However, we think that dN and dS analysis is no longer relevant to the way our work is presented. This is somewhat reinforced by the reviewer's concerns. Therefore, we have decided to delete the molecular analysis section hoping that our decision will not impact the quality of the presented work.

- Could the evolutionary scenarios be corroborated by the analysis of gene phylogenies? For instance, in paragraph on the proposal for ancestral presence of the different vitamins B and co-factors (L. 285-292), it would have been important to have checked whether the gene trees follow the species tree. If verified, then the ancestral presence would clearly be supported (with subsequent losses in the case of patchy gene distributions as proposed). Goes the same for the discussion of the evolution and distribution of the other traits. The case made for the evolution of the dot/Icm T4SS is clearer, as the gene order was analysed along the phylogeny, in order to help support the claim for ancestrality and gene patchiness (Fig. 5).

We performed single-gene ML trees for the different OCPs containing the Dot/Icm T4SS and the vitamins and co-factors discussed in the main text. Despite some topological differences, especially the *Coxiella* MAGs position. Most trees support the ancestral statutes of the discussed features with the highest support for the *Coxiella burnetii*/*Coxiella*-LE clade. All trees are now available at https://itol.embl.de/shared/dsantosgarcia

- In addition, it could help interpret evolutionarily Figure 3, Figure 4 and Figure 6 on vit. B metabolic potential, presence of T4SS and pH regulation mechanisms, if a sketch species tree was added to remind the relationships between genomes, for example below the species names (or on the left for T6SS).

We added a cladogram representing Coxiellacea species phylogenetic relationships to Figures 2, 3, 4, 5, and 6, and Figures S3 and S8 to homogenize figures.

- Upon first mention of the Sha/Mrp operon history, it is unclear from genomic contexts represented on Fig 5 that the operon has been acquired by transfer (L. 325). What are the pieces of evidence? Is it not present outside of the represented species' genomes set (e.g. from other "outgroup"

genomes)? Referring much earlier to data from Fig S8-S15 could help the reader understand. Or maybe add a mention like “see later” to make the reader understand that further details would be provided later on in the text.

We agree on the reviewers point and we modified the text accordingly. Now, the paragraph describing the HGT origin of the Sha/Mrp operon is fused with its first mention when describing the genomic context of the PAI: “Among the predicted HGT (Table S7), the presence of a Sodium Hydrogen/Multiple resistance and pH (Sha/Mrp) antiporter. The Sha/Mrp antiporter is located upstream from the Dot/Icm SS and is composed by six genes (shaABCDEFG) organized as an operon (Fig 5). This operon might have been acquired from a *Coxiella* relative, such as *Coxiella* sp. GCA_001802485, but its origin is probably from *Legionella* (Fig ??-S15). Indeed, only *B. cookevillensis* encodes another Sha operon (Fig ??), but one which is unrelated to that one of the *C. burnetii* lineage (Fig ??-S15), supporting different HGT events.”

We slightly modified the old paragraph mentioning the HGT origin of the Sha/Mrp to keep the reading flow: “Already mentioned above, the Na /H Sha/Mrp antiporter, was acquired laterally by the MRCA of the *C. burnetii* lineage (Fig ??-S15). Besides their role as cation antiporters, Sha/Mrp antiporters have other functions than pH homeostasis (Ito2017), including virulence and host-colonization (Kosono2005).”

- Paragraph 343-351 on expression profiles in the different morphotypes: even if somehow linked to pathogenicity evolution, I do not see well the rationale of putting this paragraph at this point of the text. Also, I do not see any associated Materials and Methods. Could the authors clarify where the data come from and how were they analysed?

We agree that the paragraph doesn’t fit correctly in the section. However, some of these proteins are later on discussed and we prefer to keep the paragraph. The screened proteins list was obtained from *Coleman et al. 2007*. Since the analysis is just to check for the presence/absence of those proteins in the OCPs, we considered is somehow implicit and does not require and specific M&M sentence. However, we added the following sentence to M&M: “Obtained COPs table was queried to retrieve specific subsets of COPs and to check for the presence/absence of COPs in different Coxiellaceae.”

We rephrased the subsection title and paragraph 343-351 to avoid confusions. We feel that adding it in first term will avoid breaking the reading flow of the Icm/Dot and the PAI results. Now the subsection title is: “Evolution of Coxiellaceae Virulence: Phase-Specific proteins and the Dot/Icm System”.

The discussed paragraph is now: “It is known that *C. burnetii* encodes several proteins which are over- or under-expressed in the different morphotypes (SCV or LCV) and may play important roles in pathogenicity (Coleman2007). Those proteins are defined, according to their expression profiles in the morphotypes, as LCV^{Hi} /SCV^{Lo} and SCV^{Hi}/LCV^{Lo}. Among these phase-specific proteins, the small cell variant protein A (ScvA) and histone-like Hq1 (HcbA) are thought to be involved in nucleoid condensation in SCVs (Coleman2007). Therefore, we assessed the presence of LCV^{Hi}/SCV^{Lo} and SCV^{Hi}/LCV^{Lo} proteins among the different Coxiellaceae genomes (Table S9).“

We expect that those changes increased the readability of the subsection.

Minor comments

- L. 165-166: I have doubts on the order things were performed to build the concatenate of COPs to build the species tree? Alignments of individual genes should be treated separately (alignment AND position filtering) before being concatenated. It is unclear what was done, or maybe even suggested that Gblocks was used after the concatenation. Could you please clarify?

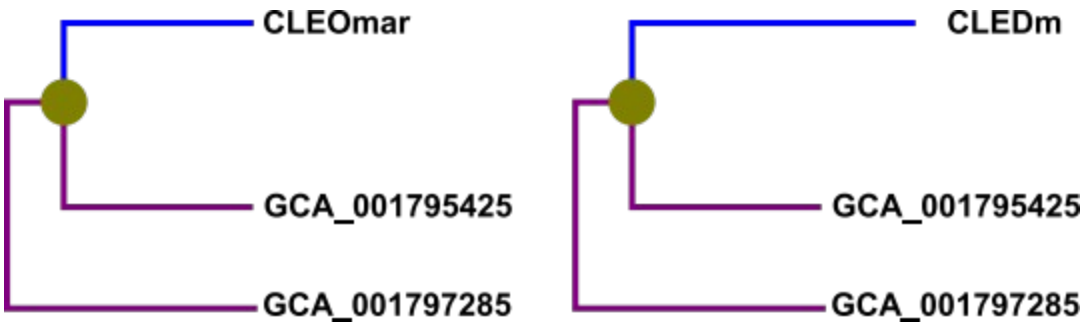
OrthoFinder computes a species tree based on a concatenated alignment of individually aligned COPs. Since OrthoFinder uses IQTREE with default parameters, it does not calculate node support. Hence, we used OrthoFinder's results to obtain a final species tree with support values. For that, we used the same concatenated alignment but pruning positions with a gap in more than 50% of the sequences (Gblocks with -b=5). Finally we used IQTREE again to obtain the Maximum Likelihood tree with node support. We tried to clarify the paragraph and now it reads as follows: "A first species tree was obtained with OrthoFinder. In brief, 348 individually aligned COPs were selected by OrthoFinder to build a concatenated alignment. Then, the species tree of the Coxiellaceae dataset was inferred using the STAG algorithm (doi:10.1101/267914). To obtain node support values, a second species tree was computed as follows: (i) OrthoFinder concatenated alignment (107812 positions) was pruned with Gblocks v0.91b (half-gaps, 21499 selected positions in 143 blocks) (Castresana2000); (ii) IQ-TREE v2.0.3 was used to infer the Maximum Likelihood phylogenomic tree using the best suggested evolutionary model (-m MFP) and ultrafast bootstrap (-bb 1000) and SH-like approximate likelihood ratio test (-alrt 1000) (Nguyen2015; Kalyaanamoorthy2017)".

We agree with the reviewer on the convenience of pruning each alignment separately. This is done by default by OrthoFinder on each alignment before their concatenation. However, for being a bit more astringent, we pruned the concatenated alignment with Gblocks. We are aware of the introduction of erroneous conserved blocks due to the concatenation but we think their effect was buffered by the large amount of positions used. Indeed, the topology of both trees (OrthoFinder automatic method and our pruned concatenation) was the same. Hence, we expect the reviewer will agree with us that the quality of the inferred tree is enough for the presented work.

- L. 190-191. I am not sure to understand, could the authors please give more details on the "triplet approach"?

We used the triplet approach previously in *Sodalis* symbionts (Santos-Garcia *et al*. 2017, doi: 10.1093/gbe/evx202). It was a way to standardize divergence times among without having a fossil record.

For this approach, an unrooted sub-tree from the species tree is used. In the sub-tree, two basal species to the clade of interest are fixed while the third species is the one we want to compute *dS* and *dN* parameters. In that way, we can compare branches leading to the species of interest (blue branche) since their distance to the common ancestor (green dot), the node between the two fixed species (purple branches), is always the same. We hope the concept is more clear now.



- In Figure 2 presenting the species phylogeny, it is sometimes difficult to see to which branch do the support values correspond. It is probably not always easy, but could the values be placed closer to the corresponding branches?

We tried to alleviate the visualization problem by deleting all support values above 99 and placing the remaining values as close as possible to the node they belong to. We also added the following sentence to the Figure 1 legend: “SH-aLRT/ultrafast bootstrap support values numbers are displayed at each node if they are below 99.”

- Line 312, maybe specify it is a Type IV secretion system (and use T4SS instead of simply SS)?

We changed the nomenclature of the Dot/Icm SS to Dot/Icm T4SS over all the text and supplementary material according to reviewer’s suggestion.

- Line 317, I’d say it is not only its presence in most genomes, but also its distribution along the phylogeny of these species’ genomes that do suggest its ancestrality.

We changed the sentence according to reviewer’s suggestion: “Nonetheless, the presence of the Dot/Icm T4SS in most Coxiellaceae genomes along all their phylogeny indicates its ancestral state”.

- The beginning of the section on pH resistance is somehow redundant with the paragraph on lines 68-73 in introduction.

We deleted the following sentence to reduce the redundancy: “Generally, acidophilic bacteria present both specific and shared, with neutrophilic ones, mechanisms to deal with low pH (Baker-Austin and Dopson, 2007; Krulwich et al., 2011; Lund et al., 2014)”. We kept the rest of the introduction since explains why we compared the pH mechanism among Coxiellaceae.

- In discussion on sha operons origins (L. 536-542), could the respective environments inhabited by the potential donors help decipher which lineage is more likely the donor?

Coxiella MAGs were assemble from environmental samples. For example, the *Coxiella* sp. GCA_001802485 MAG (https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_001802485.1/) was obtained from any biological material from groundwater retained by a 1.2um filter (Anantharaman *et al*. 2016. doi: 10.1038/ncomms13219). With the mentioned pore size, the filter might have retained both prokaryotic and eukaryotic cells (e.g. amoebae), hence it is not clear the exact environment of *Coxiella* MAGs (aquatic, sediments, host-associated, etc). Closest *Legionella* species are mostly associated with aquatic environments (*L. nagasakiensis, L. israeliensis, L. yabuuchiae, L. gresilensis, L. jamestowniensis*) and some of them reported to infect humans (*L. nagasakiensis, L. tucsonensis, L. jamestowniensis, L. clemsonensis*). Hence, it seems both *Coxiella* and *Legionella* are able to inhabit aquatic and host-associated environments, making difficult to use

this information to decipher the origin of the Sha/Mrp antiporter. Further *Coxiella* genus taxonomic sampling might help to establish the exact origin of this antiporter.

Typos

L. 193: typo “LRT”

Corrected

L. 213: missing word “copies”?

Corrected

Fig. S4 titles: “Synonymous”

Corrected

Fig. 5 legend: slashes instead of back-slashes?

Corrected

L. 386: “cations”?

Corrected

L. 397 (and elsewhere): “amoebae”?

Corrected

L. 428: “we found traces” “that” missing?

Sentence rephrased to avoid redundancy: “As also found by Brenner2021, we found traces of acid-resistance systems (sfcA, gadC, and cfa) in CLEDm and some *Coxiella*-LE from *Rhipicephalus* tick species.”

L. 549: two “.”

Corrected

L.564: replace “then” by “thus”?

Corrected

Review by anonymous reviewer, 06 Dec 2022 03:01

This paper by Santa-Garcia et al. describes the sequencing and analysis of two new Coxiellaceae genomes, from symbionts of ticks. This clade of bacteria is of interest for understanding the evolution along the parasitism-mutualism spectrum, as it contains members having both types of relationships with their hosts. The authors identify several clade-specific genes which suggests that lifestyle transitions are reflected by distinct changes in the genome. Their results support the idea that the ancestral Coxiellaceae bacterium contained pre-adaptations that not only facilitated parasitism, but also enabled a transition to obligate mutualism in some hosts.

Overall, this paper is well written and easily understood. Although this manuscript is fundamentally a rather simple and straightforward genome description paper, the authors’ evolutionary analyses add some more depth and understanding to the existing story of Coxiellaceae evolution.

Major comments/questions:

1. The sequenced genomes were highly fragmented, which inhibited some of the synteny analyses. I suggest using long-read sequencing approaches to gain a better assembly quality, if feasible.

We agree with reviewer's comment that closed, or almost closed, genomes should be desirable. The manuscript describes only two new genomes. *Coxiella*-LEs CLEDm and CLEOmar. While the genome of CLEDm is closed, it is true that CLEOmar is highly fragmented. However, the original sequenced material is no longer available, precluding us to use long-reads sequencing to improve CLEOmar genome contiguity.

2. More justification is required for using protein isoelectric points to determine adaptation to acidic conditions. Previous papers using this method should be cited and similar analyses for known acidophilic bacteria should be used for comparison/validation. Additionally, I wonder why proteome-wide analysis was done, instead of focusing on membrane proteins or the proteins mostly expressed within acidic compartments (LCV genes?).

Unfortunately, we are not aware of any paper that empirically probes the correlation between pI and the pH range of an enzyme. However, following the reviewer's suggestion, we included some acidophilic bacteria in our analysis (Table S12). Indeed, GadB, and some GadA, enzymes' from acidophilic bacteria range from 4.96 to 5.18 pI and have a predicted charge between -10 and -4.9 at 5.5 pH (lysosome). In the case of *C. burnetii* PanD, it presents a 4.79 pI and a -2.3 negative charge at 5.5 pH. *C. burnetii* PanD predictions are closer to GadB from acidophilic bacteria than to PanD from other *Coxiellaceae* species (except some MAGs). We interpret this result as a suggestion that PanD might work in acidic conditions.

To outline this uncertainty, we toned down the corresponding paragraph in Discussion. Now it can be read as: "This enzyme presents the lowest pI (4.7) in C. 17burnetii (CBU_0422) when compared to almost all other PanD from the Coxiellaceae. Indeed, its pI and predicted charge at 5.5 pH (the pH of the lysosome) are closer to those of GadB from some acidophile bacteria. Therefore, we propose that PanD might have been co-opted to work as part of the AR2 system by decarboxylating L-glutamate under acidic conditions (Kelkar2013). However, the ability of PanD from C. burnetii to perform in acidic environments should be empirically validated."

Answering the reviewers request on why we used proteome-wide analysis, it was suggested that acidophilic bacteria might present more basic proteomes as a way of buffer the leaking of H^+ from the acidic environment. Hence, we tested if *C. burnetii*, the only known acidophile among Coxiellaceae, presented a more basic proteome. That was the reason for the proteome-wide analysis.

3. The authors infer from their phylogenomic tree and the presence/absence of respective genes that the Dot/Icm system and vitamin biosynthesis pathways were present in the Coxiellaceae ancestor and lost in certain descendant lineages (page 10). This can be explicitly tested by building phylogenies with the genes making up these systems, and determining if these are congruent with the phylogenomic tree.

We performed single-gene ML trees for the different OCPs containing the Dot/Icm T4SS and the vitamins and co-factors discussed in the main text. Despite some topological differences, especially the *Coxiella* MAGs position. Most trees support the ancestral statutes of the discussed features with

the highest support for the *Coxiella burnetii*/*Coxiella*-LE clade. All trees are now available at https://itol.embl.de/shared/dsantosgarcia

4. Is there co-speciation/co-phylogeny between ticks and their symbionts?

Co-speciation events between ticks and nutritional mutualistic symbionts have been widely reported in the literature. For example, congruence between *Amblyomma* (Binetrui *et al.* 2020 – doi: 10.1111/mec.15373) or *Rhipicephalus* (Duron *et al.* 2017 – doi:10.1111/mec.14094) ticks and *Coxiella*-LEs. However, testing for phylogenetic congruence between hosts and symbionts requires wider genera sampling, which is out of the scope of this work.

5. I cannot find the supplementary table file for this manuscript, so I cannot comment on those results. Please make sure they are available with the paper, if they are not already.

Our mistake. Supplementary Tables are now available in FigShare. Sorry for any inconvenience.

Minor points:

1. Figure 1 and Figure 6, the green/red and blue/purple are not colorblind friendly.

Figures 1 and 6 have been modified following the reviewer suggestion. The new selected colors have been tested using a color blindness accessibility simulator (https://www.color-blindness.com/coblis-color-blindness-simulator/).

2. Line 343; this mentions gene expression. It is not clear if any transcriptomics done in this study, or is this data from another publication?

There is no transcriptomics data in our analysis. The screened protein list was obtained *from Coleman et al. 2007.* We rephrased the paragraph to avoid confusions: "*C. burnetii* encodes several proteins which are over- or under-expressed in the different morphotypes (SCV or LCV) and may play important roles in pathogenicity (Coleman2007). Those proteins are defined, according to their expression profiles in the morphotypes, as LCV^{Hi} /SCV^{Lo} and SCV^{Hi}/LCV^{Lo}. Among these phase-specific proteins, the small cell variant protein A (ScvA) and histone-like Hq1 (HcbA) are thought to be involved in nucleoid condensation in SCVs (Coleman2007). Therefore, we assessed the presence of LCV^{Hi}/SCV^{Lo} and SCV^{Hi}/LCV^{Lo} proteins among the different Coxiellaceae genomes (Table S9)."

3. Figure S5; the tree on the left side of the figure is cut off.

Corrected