nature portfolio

Peer Review File

Genomic evidence of two-staged transmission of the early seventh cholera pandemic

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript reports the analysis of the completely sequenced genomes of 22 Vibrio cholerae serogroup O1 isolates collected between 1961 and 1979 from several countries. The main aim of this article was to investigate in detail the epidemiology of wave-1 7PET, which was analysed using appropriate methodologies at the SNP typing level. Additionally, this study comprehensively analysed the sequences of the genomes sequenced in this work regarding their virulence, antibiotic resistance, and other factors. Although the study claimed to have identified two-step transmission by the 7PET wave-1 line as a major finding, there is a lack of strong evidence to support this claim.

Major Comments:

According to the landmark study of three separate cholera waves reported by Mutreja et al., (PMID: 21866102), wave 1 isolates spanned a period of approximately 16 years (1973-1993, n = approximately 40 isolates). However, the current exploratory analysis selected isolates spanning the period 1961 to 1979 that were obtained exclusively from NCBI. What is the rationale for choosing this shorter period instead of up to 1993 as reported in Mutreja et al.'s work? The addition of all of these isolates from Mutreja et al.'s work could serve as baseline data for additional genomes sequenced in this work. This is an important sample selection gap that introduces a sample selection bias. Authors could consider including the entire wave-1 collection of Mutreja et al., to the current study collection and analysis. This would be the best possible approach to obtain a detailed representation.

Other comments:

1. The representation of Figure 1 could be better. Also, this can be a supplementary figure as this is more of a descriptive statistic. Certain gene categories display non-synonymous mutations as 100%. This needs to be checked.
2. Have these 22 isolates been sequenced before? I noticed that one of the isolates listed as being sequenced in the current study is already available from NCBI under GenBank: CP013305.1 (CRC1106). This requires clarification.
3. On what basis were these 29 genomes classified as 7 clusters? Looking at Figure 2, it appears that the clusters were assigned based on the geographic location from which these isolates originated. Moreover, no substantial explanations were provided. When assigning a cluster, it is necessary to explain this decision. For example, assigning clusters 1 and 4 to a single isolate is not usually practised. Please clarify.
4. Lines 79-80. Does these six isolates represent the years between 1961 and 1979? Mention the year of the isolates from which the sequences made.
5. Lines 91-92. The excluded set of genes is from the 21 functional categories?
6. It would be interesting to analyse the signal transduction mechanisms with reference to the two-compartment systems (TCS), which are known to regulate QS and biofilm formation. These analyses can include classical vs wave 1, wave 2 and wave 3 El Tor sequences to identify the evolutionary changes if any.
7. Lines 103-112. Description about C1 is missing. How the clusters (1-7) chronologically made? Is there any SNPs variation amongst these isolates?
8. Generally, V. cholerae, lack the typical enoyl-acyl carrier protein encoding gene (acp). A novel enoyl-ACP reductase isoform, FabV, in V. cholerae has been reported (PMID: 18032386). The haemolysin co-regulated protein encoding gene (hcp) is not reported in classical biotype of V. cholerae O1. The origin and presence of these genes in 7th pandemic isolates of V. cholerae is thus important to discuss.
9. V. cholerae regulate c-di-GMP levels and biofilm formation in response to distinct sensory pathways. VieSAB nearly silent in V. cholerae of the El Tor biotype, while in classical biotype, the VieA system is the primary sensory input. The role of Vie system is pre-7th pandemic V. cholerae has not been reported. The above aspects can be included in the analysis.
10. Lines 134 to 144: SNPs identified as unique to the early seventh pandemic, but not unique to all seventh pandemic isolates, due to lack of clarity regarding reciprocity or parallelism. Analysis

had to be made on the entire Wave 1 collection from Mutreja et al.'s study to conclusively determine whether they were unique to Wave 1 and not Wave 2 and Wave 3.
11. Lines 142-144. It would be interesting to analyse the classical biotype sequences to confirm the overlap characteristics in pre-7th pandemic V. cholerae O1.
12. Lines 147 to 151: Are these 150 genomes downloaded from NCBI part of wave 1? How was this confirmed? This has not been mentioned clearly.
13. Labelling of the phylogenetic tree should be done in detail, including the year and location of isolation, to infer any spatiotemporal clustering of isolates from which assigned clusters could be seen.
14. Lines 155 to 157: Isolate ID M646 and M714 have been typed as wave-2, however, assigned as cluster-C7.2 along with the wave-1 isolates as shown in Figure 2. How can this be explained?
15. Lines 161 to 164: As noted above, assigning a new cluster as C8 and C9 for a single isolate (as shown in Figure 3) is not typically practised. Additionally, no explanation was given for the polytomies seen in Figure 3 in the first part of Stage 2 of the transmission, which was highlighted. There is also no clear evidence to label it as a two-step transmission, as no spatiotemporal signals were evident.
16. Line 169. It is worthy to analyse the Pre-7th Pandemic and early 7th pandemic isolates for variation in the tcpA and ctxB genes. Many of the recent 7th pandemic V. cholerae O1 El Tor had classical and other new CT genotypes.
17. Has any QRDR mutation testing been done at the chromosomal level? There is no information about this.
18. What was the source of the study isolates? This has not been mentioned anywhere in the manuscript.
19. The use of pre-pandemic strain C5 is unclear, instead of the M66 strain.
20. Detailed descriptive SNP mapping results describing the position and genes could be supplementary data, as it dilutes the main content of the manuscript (Lines: 117 to 132)
21. Line 181. It is IncC or IncA/C?
22. Lines 211 to 213: how was the deduplication factor ruled out?
23. Line 216: Why Sis were screened only in 25 genomes?
24. Lines 357-359. Specify whether these mutations continued in the three waves.
25. Line 363. Specify the DNA repair genes.
26. The results and discussion parts are confined to the analysis of pre-and early 7th pandemic isolates. Most discussions contain repetitions of study results and are long. This needs to be taken into account and made more relevant. It will be more informative if the authors extend the analysis to 6th pandemic classical isolates, as some of the sequences are available in the public domain.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts

Reviewer #3 (Remarks to the Author):

Review of NCOMMS-24-11779 - Genomic evidence of two-staged transmission of the early seventh cholera pandemic.

This study explores genomic evolutionary aspects of the first wave of the seventh cholera pandemic. The study used a combination of Illumina and Nanopore sequencing on 22 previously unpublished genomes from 1961-1979, adding resolution to epidemiological and genomic investigations of that wave. The main findings are that there were different stages to this wave, as well as the identification of several mutations which could have played an important part in

altering the virulence.

These findings build upon and further develop what was previously demonstrated in papers such as Mutreja et al., 2011 and Hu et al., 2016, and the current paper brings an incremental increase in understanding compared to these landmark papers. Methodically, the paper does not do anything novel, but the manuscript is well written and technically sound, and the conclusions are all backed up by the data.

I have a handful of concerns and suggestions. One is that phylogenies were created using all identified SNPs, the majority of which are non-synonymous. These sites will be subject to selective pressures and homoplasies and do not carry an unbiased phylogenetic signal. Although it is not uncommon to use both synonymous and non-synonymous SNPs in phylogenetic analyses, these particular data have a very high ratio of non-synonymous SNPs, suggesting strong deviations from many of the assumptions underlying phylogeny. Perhaps the authors could also create a phylogeny using only synonymous and intergenic SNP sites and compare the phylogenetic congruence between the two.

Somewhat related to the previous point, it was surprising to me that the authors only used the maximum likelihood phylogeny and not any dating approaches like BEAST or similar since the emergence of the different stages are such a central finding in the paper. From visual inspection of the trees it looks like there is a decent temporal signal, but again this would probably require looking only at synonymous/intergenic sites. I don't think the authors NEED to add this, but I include it here as a suggestion that could add some details to their findings.

It is tricky to navigate the nomenclature in this manuscript. Focus is on the seventh pandemic, waves 1 and 2 (of 3). Wave 1 can be further divided into two "Stages" (a finding of this paper). (However, from Fig2 it seems that some wave 2 isolates are also defined as Stage 2 even though the stages only pertained to wave 1?). Furthermore, the 22 isolates can also be divided into seven clusters (C1-C7), with some of these having subclusters (C7.1-C7.2). To add to the confusion, there is both isolate C5 and cluster C5. I think the paper would benefit from some clearer definitions around these terms. On a related note, some abbreviations (such as GS) are introduced without explanation.

The discussion is good, but there are two places where I think the authors go to far in their speculation:

Line 300-302: "The transmission from Africa to European countries was probably due to Portuguese troops travelling frequently between South/West Africa and Portugal in the early 1970s" – While reference 28 does indeed mention this as a possible source for introduction to Portugal, it seems like a stretch to say that this was the most probable route of introduction to Europe in general. With the amount of traffic between South/West Africa and Europe by the 1970s there must be many competing hypotheses that have not been examined.

Line 343-346: "The low recombination rate in the seventh pandemic was likely due to the transmission amongst human hosts without any environmental survival stress or opportunity to recombine with isolates from another source or local gene pool" – Although reference 33 is cited, I could not find any such explicit statement in that study. Are the authors saying the seventh pandemic is exclusively human-to-human transmission without any role for environmental sources?.

As a final note, I am missing some information about the completeness of the genome assemblies. Were they all closed? The methods section mentions 28 and then later 29 completed genomes from this study, but the study only has data from 22 isolates on NCBI SRA so some of these must be sequenced previously? Also, the figure legend from Fig5 makes it seem as if at least four isolates were not closed. I think genome completeness should be added to supplementary table 1.

Reviewer #1 (Remarks to the Author):

This manuscript reports the analysis of the completely sequenced genomes of 22 Vibrio cholerae serogroup O1 isolates collected between 1961 and 1979 from several countries. The main aim of this article was to investigate in detail the epidemiology of wave-1 7PET, which was analysed using appropriate methodologies at the SNP typing level. Additionally, this study comprehensively analysed the sequences of the genomes sequenced in this work regarding their virulence, antibiotic resistance, and other factors. Although the study claimed to have identified two-step transmission by the 7PET wave-1 line as a major finding, there is a lack of strong evidence to support this claim.

Major Comments:

According to the landmark study of three separate cholera waves reported by Mutreja et al., (PMID: 21866102), wave 1 isolates spanned a period of approximately 16 years (1973-1993, n = approximately 40 isolates). However, the current exploratory analysis selected isolates spanning the period 1961 to 1979 that were obtained exclusively from NCBI. What is the rationale for choosing this shorter period instead of up to 1993 as reported in Mutreja et al.'s work? The addition of all of these isolates from Mutreja et al.'s work could serve as baseline data for additional genomes sequenced in this work. This is an important sample selection gap that introduces a sample selection bias. Authors could consider including the entire wave-1 collection of Mutreja et al., to the current study collection and analysis. This would be the best possible approach to obtain a detailed representation.
Response:

We have now explained why we studied isolates from 1961 to 1979. This was based on epidemiology that there was a break (the lull) from cholera outbreaks in the 1980s. Mutreja et al.'s work defines wave 1 with year up to 1993 including the South America outbreak in 1991. Epidemiologically the first intercontinental spread to Africa in the 1970s was a major event and was less than 10 years from the start of the pandemic. We believe the time period was better defined as early 7th pandemic period. We revised the introductory paragraph on the 7th pandemic and more specifically defined the pandemic period studied with the revised text is as follows in lines 64-66:

However, the transmission of the seventh pandemic in its earlier years from 1961 to 1979 (before entering the low cholera period in the 1980s) remained less well studied due to the lack of genomic data.

Unfortunately, the Mutreja et al. study used 54 bp paired-end Illumina sequencing. Although most advanced at the time, genomes assembled using such short length reads no longer pass current quality filtering standard. Ten isolates falling into the same period (1961-1979) but only five isolates passed our MGT quality control. We didn't include them in our study to maintain consistency in sequence quality.

It should be noted that by no means we imply any invalid conclusions from such short reads sequence data. The three waves are confirmed divisions of the 7th pandemic major events across continents. Also as shown in our previous study using multilevel genome typing (MGT), we can type 7th pandemic strains into the 3 waves using MGT.

Other comments:

1. The representation of Figure 1 could be better. Also, this can be a supplementary figure as this is more of a descriptive statistic. Certain gene categories display non-synonymous mutations as 100%. This needs to be checked.
Response:

We have moved Figure 1 to Supplementary Figure S1 as suggested. Additionally, we have now updated the figure and have included genes in multiple categories.

We manually checked some genes. They were correct calls. Since the coverage was high for both Illumina (85-fold average) and nanopore sequencing (160-fold average), and the genomes were assembled using a hybrid of data from both sequencing methods we believe the calls are more likely to be correct.

2. Have these 22 isolates been sequenced before? I noticed that one of the isolates listed as being sequenced in the current study is already available from NCBI under GenBank: CP013305.1 (CRC1106). This requires clarification.
Response:

Our apologies. It is our oversight we didn't state that CRC1106 was the same strain as M806 which was our lab number. CRC1106 was sequenced by Hu et al using PacBio. We compared the two genomes and found there were differences (70 SNPs and 3 structural differences). The differences could be sequencing errors or genuine minor differences between stocks of the same strain. To maintain consistency, we used the genome sequence we sequenced.

In lines 453-459 (Methods):

A total of 22 V. cholerae isolates were sequenced in this study (Table 1). These isolates were collected by other laboratories from different countries, with source laboratories listed in Table 1. These strains have used in our previous studies [15] and were historical archived strains. Note that one isolates (M806) was the same strain as CRC1106 sequenced by Hu et al. using PacBio [8]. Comparison of the two genome sequences showed 70 base differences and three genomic structure differences (Supplementary Table S8). Therefore, we used our sequence to represent the strain for consistency.

3. On what basis were these 29 genomes classified as 7 clusters? Looking at Figure 2, it appears that the clusters were assigned based on the geographic location from which these

isolates originated. Moreover, no substantial explanations were provided. When assigning a cluster, it is necessary to explain this decision. For example, assigning clusters 1 and 4 to a single isolate is not usually practised. Please clarify.
Response:

We have now removed cluster designations in the complete genome tree and introduced them in the tree with all genomes and explained the reason why we call them clusters.

4. Lines 79-80. Does these six isolates represent the years between 1961 and 1979? Mention the year of the isolates from which the sequences made.
Response:

We have now rewritten the sentence to clarify this as below in lines 80-83.

Six publicly available complete genomes from Asia including five early seventh pandemic isolates (1961-1978) and one pre-seventh pandemic strain (C5) from 1957 were retrieved from the GenBank and included in this study.

5. Lines 91-92. The excluded set of genes is from the 21 functional categories?
Response:

There were 35 genes (with 43 SNPs) that were allocated to more than 1 category. We have now included these genes and allocated them into each of their multiple categories and results are now as Supplementary Figure S1.

6. It would be interesting to analyse the signal transduction mechanisms with reference to the two-compartment systems (TCS), which are known to regulate QS and biofilm formation. These analyses can include classical vs wave 1, wave 2 and wave 3 El Tor sequences to identify the evolutionary changes if any.
Response:

Thank you for the suggestion. We have performed the analysis on all the genes with mutations in the signal transduction category and found an additional 4 TCS genes with elevated mutation rates. We have discussed this in the discussion. We also included 6th pandemic isolates but there were small number of isolates and all were isolated well after the pandemic. We didn't describe nor draw any conclusions on the 6th pandemic isolates. Overall the TCS genes with elevated mutations were present and remain elevated in all three waves. The revised text is as follows in lines 364-375:

Since luxO and hapR had elevated mutation rate, we examined all 43 genes in the signal transduction category with mutations in the early seventh pandemic isolates for elevated

mutation rate and found an additional four genes (arcA, dctB, dctR and VC0694) with a persistent elevated mutation rate in the seventh pandemic across different waves (Supplementary Fig S5). ArcA is a global regulator of the ArcB/A two component system (TCS) and known to regulate virulence gene expression and biofilm formation [36, 37]. DctB (VC1925) and DctR (VC1926) pair as a TCS that sense C4-dicarboxylates [38, 39], but little is known of regulatory targets or functions. VC0694 encodes an uncharacterised TCS histidine kinase with VC0693 encoding its cognate response regulator and plays a role in intestinal colonization in infant mice [40] and biofilm formation [41]. It seems that mutations in a subset of the TCSs may have played an adaptive role in the development of the seventh pandemic, which has not been recognised previously.

7. Lines 103-112. Description about C1 is missing. How the clusters (1-7) chronologically made? Is there any SNPs variation amongst these isolates?
Response:

Description of clusters have now moved to the draft genome section. This part of text is now deleted.

8. Generally, *V. cholerae*, lack the typical enoyl-acyl carrier protein encoding gene (acp). A novel enoyl-ACP reductase isoform, FabV, in V. cholerae has been reported (PMID: 18032386). The haemolysin co-regulated protein encoding gene (hcp) is not reported in classical biotype of V. cholerae O1. The origin and presence of these genes in 7th pandemic isolates of V. cholerae is thus important to discuss.
Response:

Our study addresses how the 7th pandemic developed from pre-7th pandemic strain, therefore we focus on changes from that evolutionary point. *hcp* has no mutational changes in the early pandemic isolates and thus plays no adaptive role for its pandemic leap.

In terms of gene content difference between the 6th and 7th pandemic clones, these genes are interesting in terms of evolution of pandemic clones. However, many studies have examined this. Again our focus was on difference between early stages of the 7th pandemic and its difference to its pre-pandemic most recent ancestor. Therefore we refrained from expanding the genome difference comparison to the other strains well outside the framework of this study.

9. V. cholerae regulate c-di-GMP levels and biofilm formation in response to distinct sensory pathways. VieSAB nearly silent in V. cholerae of the El Tor biotype, while in classical biotype, the VieA system is the primary sensory input. The role of Vie system is pre-7th pandemic V. cholerae has not been reported. The above aspects can be included in the analysis.
Response:

We didn't identify any SNPs in vieA gene in the early 7th pandemic isolates. There was no SNP difference in this gene between the pre-7th pandemic strain C5 and the early 7th pandemic isolates. Thus there were no mutational changes in this gene that played a role in turning the 7th pandemic precursor to the 7th pandemic clone.

10. Lines 134 to 144: SNPs identified as unique to the early seventh pandemic, but not unique to all seventh pandemic isolates, due to lack of clarity regarding reciprocity or parallelism. Analysis had to be made on the entire Wave 1 collection from Mutreja et al.'s study to conclusively determine whether they were unique to Wave 1 and not Wave 2 and Wave 3.

Response:

We apologise for the lack of clarity in this. All our statements on the unique SNPs to the to early 7th pandemic was relative to the pre-7th pandemic, not to wave 2 or wave 3 or the entire 7th pandemic set of strains. We have now expanded the analysis to all 7th pandemic isolates through mapping. Three SNPs remain 7th pandemic specific and one (on *tcpF*) was unique to all except E9120 (an N at the site) and six other early isolates in cluster CL1 and thus the SNP arose after CL1 divergence. Therefore, we have excluded the *tcpF* SNP as it is not specific to the entire 7th pandemic. Please note that we used *acfB* (VC0840) as the gene symbol for *pctA_1* in this revision. The revised text is as follows.

Results section lines 133-138:

We interrogated 7574 seventh pandemic isolates from all three waves on these SNPs. The SNP on csrD reversed back to the allele on the C5 strain in three seventh pandemic isolates (ERR576981 (wave 1), ERR4175611 (wave 2), ERR9716121 (wave 3)) and thus was seventh pandemic specific. The SNPs on acfB and luxO were found to be present in all the seventh pandemic isolates examined. Thus, we can ascertain that the three SNPs observed were unique to the seventh pandemic with the exceptions described above.

Discussion section lines 325-330:

We confirmed these mutations were present in all or nearly all wave 2 and wave 3 isolates and thus maintained in the seventh pandemic clone. Our findings suggest that the accumulations of nonsynonymous mutations in the three genes that play key roles in adaptation and virulence may have enabled the pandemicity of the seventh pandemic clone in the initial development in Indonesia after it diverged from its precursor.

11. Lines 142-144. It would be interesting to analyse the classical biotype sequences to confirm the overlap characteristics in pre-7th pandemic V. cholerae O1.

Response:

We presume reviewer's reference to pre-7th means early 7th pandemic. This is an interesting suggestion and interesting line of enquiry on the evolution of the pandemic clones. However the limited number of 6th pandemic isolates/genomes available would not allow us to draw any reliable conclusions. Therefore we didn't perform the analysis on the 6th pandemic isolates for comparison.

12. Lines 147 to 151: Are these 150 genomes downloaded from NCBI part of wave 1? How was this confirmed? This has not been mentioned clearly.
Response:

Not all belonged to wave 1. We used MGT (Cheney et al. mSystems 2021) to type them into MGT2 STs which were designed to type isolates into the 3 waves. The results were in supplementary Table S9. We clarified this in lines 167-171 as below:

The seventh pandemic was divided into waves [9] and the three waves can be distinguished using multilevel genome typing (MGT) [23]. By MGT typing, 173 isolates/genomes were identified as MGT2 ST1 which belongs to wave 1. Three isolates/genomes, M646 and M714 and one Illumina sequenced genome (ERR025383) were typed as MGT2 ST2, which belongs to wave 2 was derived from a precursor in C7.2.

13. Labelling of the phylogenetic tree should be done in detail, including the year and location of isolation, to infer any spatiotemporal clustering of isolates from which assigned clusters could be seen.
Response:

We redid the figure as Fig 2 with all data labelled on the tree. Year of isolation is shown using a heatmap with reds for 1960s and blues for 1970s. Continent / Subcontinent of origin for each isolate is displayed with a colour bar and Country of origin is displayed when that country was mentioned specifically in the text. Only complete genomes are named for clarity. We have also marked keys events including date of divergence on the tree.

14. Lines 155 to 157: Isolate ID M646 and M714 have been typed as wave-2, however, assigned as cluster-C7.2 along with the wave-1 isolates as shown in Figure 2. How can this be explained?
Response:

The pandemic clone was divided into 3 waves, in fact there were demarcation of 3 clades. Wave 2 is a clade within wave 1 and evolved from wave 1 and wave 3 is a clade within wave 2 and evolved from wave 2 (Mutreja et al. Figure 1). These waves were not due to strains evolved in parallel. The later one was derived from an early one. In this study, we discovered the origin of wave 2 lineage. CL7.2 gave rise to wave 2, while all other CL7.2 and wave 1 offsprings on the tree died out.

Therefore what we described is consistent with the wave demarcation of Mutreja et al.

15. Lines 161 to 164: As noted above, assigning a new cluster as C8 and C9 for a single isolate (as shown in Figure 3) is not typically practised. Additionally, no explanation was given for the polytomies seen in Figure 3 in the first part of Stage 2 of the transmission, which was highlighted. There is also no clear evidence to label it as a two-step transmission, as no spatiotemporal signals were evident.

Response:

We have removed the cluster labelling for singletons and referred to the branch by strain name.

The polytomies are due to the isolates evolving at the same time so it is a star phylogeny at that node point. Star phylogeny is due to a population expansion event. The start of the pandemic is a massive population expansion so it isn't surprising we see a star phylogeny. we have added explanation to the text as follows.

We have reworked on figure 2 and have marked the stages with events. We marked isolates by years using colour heat map and also marked them by continents. Hope the spatiotemporal signal of two stage transmission is now more clearly presented.

In results in lines 146-148:

It should be noted that along the phylogenetic tree and within the clusters, there were many isolates showing a star phylogeny. Such branching patterns are typical of rapid population expansion of a pandemic organism.

16. Line 169. It is worthy to analyse the Pre-7th Pandemic and early 7th pandemic isolates for variation in the tcpA and ctxB genes. Many of the recent 7th pandemic V. cholerae O1 El Tor had classical and other new CT genotypes.

Response:

There were no changes in both genes. We've added this sentence in line 180-182:

All ctxB genes were typed to ctxB genotype 3. All the tcpA genes of the complete genomes from this study and the strain C5 were identical.

17. Has any QRDR mutation testing been done at the chromosomal level? There is no information about this.

Response:

We checked QRDR mutations using AMRFinderPlus (version: 3.12.8), none of the QRDR mutations were identified in this study.

We added this sentence in lines 192-193:

None of the other isolates from the early pandemic period carried any plasmids and no isolates carried AMR genes or mutations.

In lines 506-507 (Methods):

We used AMRFinderPlus version 3.12.8 with database version 2024-05-02.2 to identify point mutations in the assemblies [66].

18. What was the source of the study isolates? This has not been mentioned anywhere in the manuscript.
Response:

We have now added to the methods and have also moved supplementary table S1 to a main table. The revised text is as below in lines 453-456:

A total of 22 V. cholerae isolates were sequenced in this study (Table 1). These isolates were collected by other laboratories from different countries, with source laboratories listed in Table 1. These strains have used in our previous studies [15] and were historical archived strains.

19. The use of pre-pandemic strain C5 is unclear, instead of the M66 strain.
Response:

From Hu et al. (PNAS, cited ref [8]), there were six stages leading to the 7th pandemic. M66 isolated in 1937 was in stage 5 while C5 isolated in 1957 was in stage 6 which was the closest to the 7th pandemic lineage. Therefore we chose strain C5 instead of strain M66. We added this sentence in lines 83-85:

C5 was the closest pre-seventh pandemic isolate [8] and was used as an outgroup and for comparison to identify seventh pandemic specific changes.

20. Detailed descriptive SNP mapping results describing the position and genes could be supplementary data, as it dilutes the main content of the manuscript (Lines: 117 to 132)
Response:

We moved it to supplementary figure (Supplementary Fig S2) as suggested.

21. Line 181. It is IncC or IncA/C?
Response:
Our typo, it's IncA/C. We changed it to IncA/C in line 190.

22. Lines 211 to 213: how was the deduplication factor ruled out?
Response:

We didn't observe any duplicate genes in C5 that were deduplicated in the early 7th pandemic isolates. We added this to the text as follows in lines 220-221.

No duplicated C5 genes were deduplicated to single copy in the early seventh pandemic isolates.

23. Line 216: Why Sis were screened only in 25 genomes?
Response:

The chromosomes 2 of four strains in this study were not closed, so they can't be used to analyse for the short repeats on superintegrons.

24. Lines 357-359. Specify whether these mutations continued in the three waves.
Response:

Yes, these elevated mutation rates continued in the three waves. We have added to the text as follows in line 360-362:

The elevated mutation rate in luxO and hapR was maintained throughout the three seventh pandemic waves (Supplementary Fig S56).

25. Line 363. Specify the DNA repair genes.
Response:

We have specified the DNA repair genes. The revised text is as follows in lines 379-382.

Mutators that carry mutations in DNA mismatch repair genes have increased mutation rates and may facilitate adaptation [43]. However, our complete genomes did not contain any mutator mutations with the mismatch repair genes examined (mutS, mutH, mutL and uvrD).

26. The results and discussion parts are confined to the analysis of pre-and early 7th pandemic isolates. Most discussions contain repetitions of study results and are long. This needs to be taken into account and made more relevant. It will be more informative if the authors extend the analysis to 6th pandemic classical isolates, as some of the sequences are

available in the public domain.
Response:

We have tried to reduce summaries of the results in discussion. We have refrained from discussion of the 6th pandemic as our focus was on the early 7th pandemic development. Further there were limited number of genomes available from the 6th pandemic as stated.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts

Reviewer #3 (Remarks to the Author):

Review of NCOMMS-24-11779 - Genomic evidence of two-staged transmission of the early seventh cholera pandemic.

This study explores genomic evolutionary aspects of the first wave of the seventh cholera pandemic. The study used a combination of Illumina and Nanopore sequencing on 22 previously unpublished genomes from 1961-1979, adding resolution to epidemiological and genomic investigations of that wave. The main findings are that there were different stages to this wave, as well as the identification of several mutations which could have played an important part in altering the virulence.

These findings build upon and further develop what was previously demonstrated in papers such as Mutreja et al., 2011 and Hu et al., 2016, and the current paper brings an incremental increase in understanding compared to these landmark papers. Methodically, the paper does not do anything novel, but the manuscript is well written and technically sound, and the conclusions are all backed up by the data.

I have a handful of concerns and suggestions. One is that phylogenies were created using all identified SNPs, the majority of which are non-synonymous. These sites will be subject to selective pressures and homoplasies and do not carry an unbiased phylogenetic signal. Although it is not uncommon to use both synonymous and non-synonymous SNPs in phylogenetic analyses, these particular data have a very high ratio of non-synonymous SNPs, suggesting strong deviations from many of the assumptions underlying phylogeny. Perhaps the authors could also create a phylogeny using only synonymous and intergenic SNP sites and compare the phylogenetic congruence between the two.

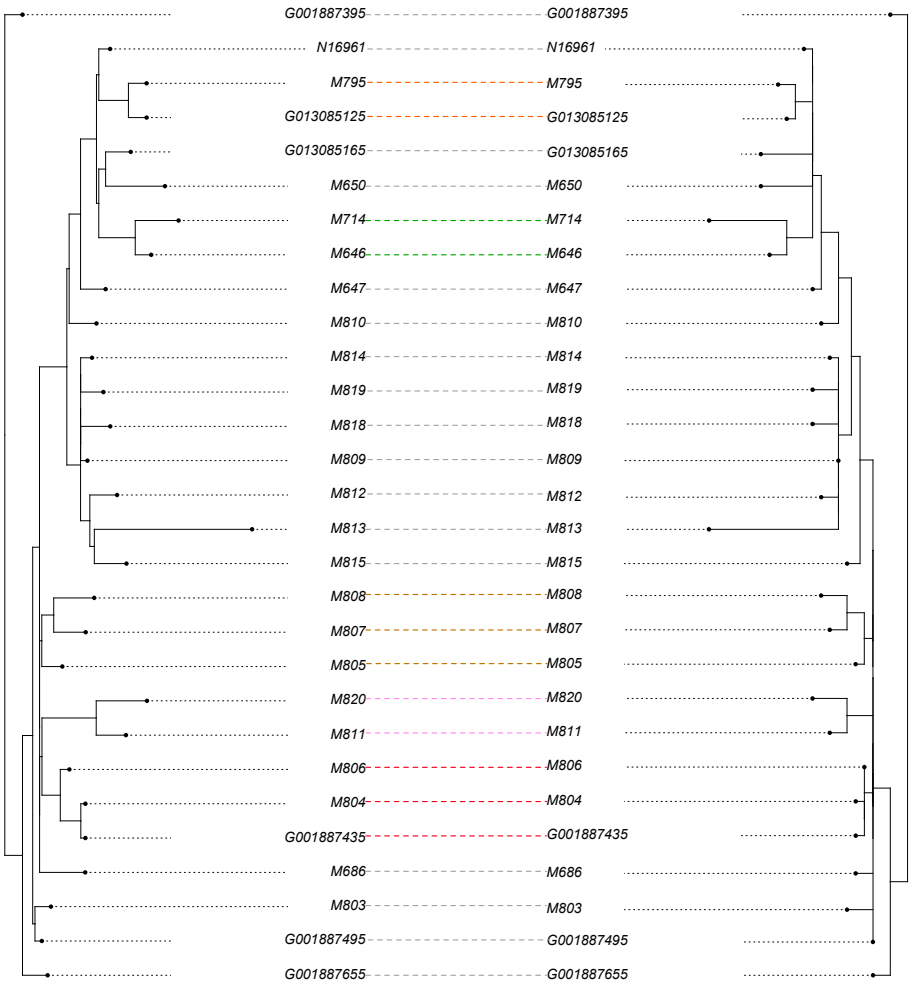
Response:

We have revised various parts of the MS and the conclusion to make it clearer that our study offered some major advances on the evolution of the 7th pandemic cholera. We highlight key findings here.

1) Identifying the two stages of initial pandemic spread provides novel insight into how the 7th pandemic started. It had always been assumed that the pandemic spread as a single event from Indonesia in 1961 with one homogenous strain. Our analysis of more than one isolates from Indonesia showed that wasn't the case. There was already diversity developed in the 7th pandemic clone before it spread and multiple variants of the pandemic spread simultaneously to other parts of Asia.
2) No studies up to date pinpointed any genetic events that made the 7th pandemic clone a pandemic organism. We found that the mutations in three key regulatory genes that may be the critical changes as they were present in all stage 1 isolates. Hu et al 2016 did not identify these key pandemic-enabling mutations. We believe our findings were fundamentally important to understand how the 7th pandemic evolved right at the start.

3) We have identified the precursors of the wave 2.
4) We found elevated mutation rates in *luxO/luxR* regulatory genes in the early 7th pandemic. We also found 4 additional two component system regulatory genes with elevated mutation rates in the entire 7th pandemic.

We have constructed a tree using sSNPs to avoid homoplasy as shown below. The left is the tree with all SNPs, right with only synonymous SNPs. The coloured lines indicate the common subtrees. The figure was generated using the phytools R package, then further sorted manually so that the nodes aligned, and adjoining lines were coloured when a similar subtree structure was seen. The two trees are very similar.



Somewhat related to the previous point, it was surprising to me that the authors only used the maximum likelihood phylogeny and not any dating approaches like BEAST or similar since the emergence of the different stages are such a central finding in the paper. From visual inspection of the trees it looks like there is a decent temporal signal, but again this would probably require looking only at synonymous/intergenic sites. I don't think the authors NEED to add this, but I include it here as a suggestion that could add some details to their findings.

Response:

We appreciate the Bayesian approach (BEAST) to construct a time measured phylogeny that allows computation of rates and evolutionary timeframe. We have attempted for both the complete genome dataset and the total dataset including both complete and Illumina draft genomes. Unfortunately for the complete genome dataset, TempEst analysis found that there wasn't enough temporal signal to perform BEAST analysis.

For the Illumina draft genomes and the complete genomes together, there was good temporal signal based on TempEst analysis. However, there were 8 genomes with very long terminal branch lengths which led to violation of all the evolutionary models tested in BEAST analysis. One of the 8 genomes was a strain from Indonesia isolated in 1961 which diverged the earliest, so we attempted BEAST analysis by 1) removing the 8 genomes; 2) keeping the Indonesian one but removing the other seven genomes. In both cases, forcing strain C5 (the prepandemic strain) as the outgroup, the date estimate for the MCRA of the 7th pandemic was 1940s. In particular, in the latter, the ESS for clock rate and overall priors were very low. Therefore, we did not include the results in the manuscript.

It is tricky to navigate the nomenclature in this manuscript. Focus is on the seventh pandemic, waves 1 and 2 (of 3). Wave 1 can be further divided into two "Stages" (a finding of this paper). (However, from Fig2 it seems that some wave 2 isolates are also defined as Stage 2 even though the stages only pertained to wave 1?). Furthermore, the 22 isolates can also be divided into seven clusters (C1-C7), with some of these having subclusters (C7.1-C7.2). To add to the confusion, there is both isolate C5 and cluster C5. I think the paper would benefit from some clearer definitions around these terms. On a related note, some abbreviations (such as GS) are introduced without explanation.
Response:

To deal with clarity on the divisions by wave, stage and clusters, we have renamed clusters as CL which also avoids confusion with strain C5. We have now defined better the early 7th pandemic from 1961 to 1979 which is within wave 1. The two stages were used to divide the early 7th pandemic period and hopefully this helps to reduce confusion. We have now explained why we named some clades as clusters, which were clades with certain spatial or temporal epidemiological significance and for the convenience of discussion.

We added the abbreviation GS in line 246 for genomic structures.

The discussion is good, but there are two places where I think the authors go to far in their speculation:

Line 300-302: "The transmission from Africa to European countries was probably due to Portuguese troops travelling frequently between South/West Africa and Portugal in the early 1970s" – While reference 28 does indeed mention this as a possible source for introduction to

Portugal, it seems like a stretch to say that this was the most probable route of introduction to Europe in general. With the amount of traffic between South/West Africa and Europe by the 1970s there must be many competing hypotheses that have not been examined.
Response:

We have revised the discussion as suggested. The revised text is as follows in lines 308-309.

However, there were many other possible transmission routes through trade and human travel between Africa and Europe.

Line 343-346: "The low recombination rate in the seventh pandemic was likely due to the transmission amongst human hosts without any environmental survival stress or opportunity to recombine with isolates from another source or local gene pool" – Although reference 33 is cited, I could not find any such explicit statement in that study. Are the authors saying the seventh pandemic is exclusively human-to-human transmission without any role for environmental sources?.
Response:

We have rewritten the explanation as the sentence is a too brief summary of the explanation put forward by Reference 33. In the discussion of Ref 33 (the last 2 paragraphs before conclusion section), two hypotheses were put forward, 1) they spend less time in the environment during pandemic and outbreak periods so less opportunity for recombination; 2) when they were in the environment, they stay in the environment in a viable but nonculturable form (VBNC).

In lines 343-347 we changed the sentence to:

As hypothesised by Hu et al. [32], the low recombination rate in the seventh pandemic clone may be due to that they spend less time in the environment during pandemic and outbreak periods, and they may also enter a viable but nonculturable form with less opportunity for recombination when they are in the environment.

As a final note, I am missing some information about the completeness of the genome assemblies. Were they all closed? The methods section mentions 28 and then later 29 completed genomes from this study, but the study only has data from 22 isolates on NCBI SRA so some of these must be sequenced previously? Also, the figure legend from Fig5 makes it seem as if at least four isolates were not closed. I think genome completeness should be added to supplementary table 1.

Response:
Apologies for the confusing numbers.

There were 28 genomes included in this study, the 29th genome was the reference genome N16961. We've corrected this in lines 481-483 to make it clearer.

The SNPs of publicly available genomes and the complete genomes in this study were called by SaRTree pipeline using proportion threshold 20 and all recombinant SNPs were removed.

We added one column of genome completeness in Table 1 and marked all the closed chromosomes and unclosed chromosomes.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have made efforts in addressing the reviewers' comments. While the overall improvement is evident, there are areas that still require clarification. Some key issues remain unaddressed, and the justifications provided are inadequate. Following are the major comments that need to be addressed.

1. Inclusion of Isolates from Mutreja's Study: The 5 out of 10 isolates that passed the quality filter from Mutreja's study should be included in the analysis. If included already, it needs to be mentioned clearly.
2. Gene Categorization (Please refer previous Q.1): The number of genes in each category is missing, making it difficult to draw any conclusions from Supplementary Figure 1. Inclusion of a supplementary table categorizing the genes would enhance clarity.
3. SNP Differences (Q.2): The difference of 70 SNPs in isolate CRC1106/M806 is significant and cannot be attributed to sequencing errors or different stock cultures, as this exceeds the expected SNP accumulation rate in V. cholerae (3.3 SNPs/year). This requires proper justification or omission of the SNP differences.
4. Identification of Isolates (Q.10): The analysis of 7574 7PET isolates is only briefly mentioned in lines 133-134. The revised manuscript lacks details on how these isolates were identified as 7PET/wave:1-3 among the total global collection (NCBI). This should be clearly outlined in the methods and results sections.
5. Justification of 150 Isolates (Q.12): The provided justification for the selection of 150 isolates does not support the data. Further clarification is needed.
6. Phylogenetic Tree (Q.13): The tree contains too many variables. Consider removing the bootstrap values and clearly mark the study isolates. It would be useful for the readers if the complete genomes are marked at the branch tips and country of origin uniformly across the tree. Mark any missing data as unknown.
7. Figure 3: This figure could be improved by highlighting genes related to AMR/MGEs while removing other gene names. This will make better readability and emphasis the key information.
8. Figure 5: The legend needs more detailed explanation. Currently, it lacks the study isolate IDs. Indicate if representative isolates from each cluster/2 stages were chosen. This could also be provided as a supplementary figure.
9. Non-Inclusion of BEAST Analysis: The omission of BEAST analysis should be mentioned as a limitation of the study.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts

Reviewer #1 (Remarks to the Author):

The authors have made efforts in addressing the reviewers' comments. While the overall improvement is evident, there are areas that still require clarification. Some key issues remain unaddressed, and the justifications provided are inadequate. Following are the major comments that need to be addressed.

1. Inclusion of Isolates from Mutreja's Study: The 5 out of 10 isolates that passed the quality filter from Mutreja's study should be included in the analysis. If included already, it needs to be mentioned clearly.
Response:

We apologise for the unclear description of these 5 isolates. Among these 5 isolates, three of them have been included in the analysis (ERR025383, ERR025384, ERR025386), we have added the remaining two (ERR018188 and ERR018191), both from India isolated in 1979 on the phylogenetic tree (Figure 2) in this revision for completeness. It is our oversight that these two were filtered out due to their genome size of ERR018191 (3892417 bp) and ERR018188 (3899905 bp) falling just under the cutoff of 3.9 MB in our original (slightly more stringent) filters on genome size but passed in our current MGT typing filter. Adding the 2 genomes didn't affect the conclusions of this study.

2. Gene Categorization (Please refer previous Q.1): The number of genes in each category is missing, making it difficult to draw any conclusions from Supplementary Figure 1. Inclusion of a supplementary table categorizing the genes would enhance clarity.
Response:
A supplementary table of the genes in each category has been added as supplementary Table S2.

3. SNP Differences (Q.2): The difference of 70 SNPs in isolate CRC1106/M806 is significant and cannot be attributed to sequencing errors or different stock cultures, as this exceeds the expected SNP accumulation rate in V. cholerae (3.3 SNPs/year). This requires proper justification or omission of the SNP differences.
Response:

We apologise for not thoroughly investigating these differences. The differences identified were done through MUMmer4 alignment of the two assembled genomes. we have now used progressiveMauve to perform genome alignment. The three genomc structural differences were also identified by Mauve and are consistent. However for SNPs, there were differences. Excluding indels, the common SNPs from the two programs were 32 SNPs. We examined the location of the SNPs (on genes or intergenic regions) and whether they were clustered SNPs. 10 SNPs were on 16S ribosomal RNA genes, 8 SNPs on 23S ribosomal RNA genes, one SNP on an IS200/IS605-like element, 5 SNPs on or 3' proximal to a DUF1289 domain-containing protein gene, and 4 SNPs on a pseudogene. We further found the latter two gene/pseudogene contained repetitive regions and thus these SNPs were most likely assembly errors. For 16S and 23S genes, it is known that there is inter-operon variation as there are up to 9 copies of 16S-23S operons in *V. cholerae* genome. We

further performed Illumina reads mapping and found only 2 consecutive SNPs at M806 genome position 2861457-2861458 were called and none of the other SNPs was identified by reads mapping. Therefore, the previously identified SNP differences were largely non-genuine. We cannot ascertain further whether the two SNPs, also identified by reads mapping, were genuine SNP difference as they are consecutive base changes. The bases of M806 are the same as N16961 while the ones of CRC1106 were different. Therefore if these two SNPs were genuine, it could be a single event and would fall within expectations of the mutation rates and that the SNP differences were due to different cultures. Overall we can conclude that these genomes were very similar and most likely represent the same strain.

4. Identification of Isolates (Q.10): The analysis of 7574 7PET isolates is only briefly mentioned in lines 133-134. The revised manuscript lacks details on how these isolates were identified as 7PET/wave:1-3 among the total global collection (NCBI). This should be clearly outlined in the methods and results sections.
Response:

The MGT method of assignment of the 7PET isolates to different waves was established in our previous study when we developed the MGT typing system (Cheney et al. 2021 cited in the manuscript). The difficulty and the need to assign an isolate to a wave by phylogenetic analysis was one of the reasons we developed MGT. We would like to refer reviewer to that publication for algorithm details. Essentially, in that study, we identified specific gene/alleles to demarcate the 3 waves and incorporated the loci into MGT2 level to differentiate them using MGT2 sequence types (STs). MGT2 ST1, ST2 and ST3 correspond to wave 1, wave 2, wave 3 respectively. We clarified the correspondence of MGT STs to waves in the methods in lines 514-516:

MGT2 ST1, ST2 and ST3 correspond to the seventh pandemic wave 1, wave 2 and wave 3 respectively [23].

5. Justification of 150 Isolates (Q.12): The provided justification for the selection of 150 isolates does not support the data. Further clarification is needed.
Response:

The 150 isolates (now 152 with the addition of 2 extra isolates) were filtered by year (1961-1979) and genome quality (MGT typing filters). We have added in the results to specify their year range and have now added in the methods to specify both filters.

In results (lines 140-142):
The newly sequenced complete genomes were compared with 152 Illumina sequenced genomes from the early seventh pandemic period (1961-1979) by phylogenetic analysis (Fig 2, Supplementary Fig S3).

In methods (lines 466-468):
Additionally, raw sequence data of 152 V. cholerae isolated between 1961 and 1979 from African and Asian countries with genome sequence quality that passed MGT typing filters were downloaded from NCBI for comparison (Supplementary Table S10).

6. Phylogenetic Tree (Q.13): The tree contains too many variables. Consider removing the bootstrap values and clearly mark the study isolates. It would be useful for the readers if the complete genomes are marked at the branch tips and country of origin uniformly across the tree. Mark any missing data as unknown.
Response:

We have removed the bootstrap values from the tree (Figure 2) and marked the complete genomes sequenced in this study. Since the isolates on the tree were from 39 countries, we only highlighted the countries where the isolate was discussed in the manuscript as featured isolate/genomes. For the other countries, we marked them all with the same grey colour and referred readers to supplementary figure for details. There are no isolates with unknown country, so our previous presentation can lead to misunderstanding although we explained this in the figure legend. This is now amended and clearly stated in the colour legend.

7. Figure 3: This figure could be improved by highlighting genes related to AMR/MGEs while removing other gene names. This will make better readability and emphasis the key information.
Response:

We have now highlighted antibiotic resistance, antiseptic resistance and metal resistance genes by different colours. We prefer keeping the other genes on the plasmid as other readers may have an interest on the other genes carried by the plasmids.

8. Figure 5: The legend needs more detailed explanation. Currently, it lacks the study isolate IDs. Indicate if representative isolates from each cluster/2 stages were chosen. This could also be provided as a supplementary figure.
Response:

The representative isolates were chosen by the different genomic structures, not related to the clusters. We have also moved figure 5 to supplementary as Supplementary figure S5 as suggested.

Additional details were also added to the figure legend:

Four representative genome structures (GS1, GS2, GS3 and GS4 represented by N16961, E1162, M647 and M803 respectively) were aligned by progressiveMauve. The distribution of these structures amongst the 29 complete genomes in the study are shown in Figure 1.

9. Non-Inclusion of BEAST Analysis: The omission of BEAST analysis should be mentioned as a limitation of the study.
Response:

Omission of BEAST analysis was a limitation of the data not a limitation of the study. As detailed in response to reviewer 3 comments in the last revision, we did perform BEAST analysis on both complete genome dataset and the total dataset and we

described the outcomes of attempted analysis in the response. From our understanding, response to reviewers will be published with the paper if accepted by this journal, readers will be able to read the outcomes of the BEAST analysis and limitations associated with the datasets on BEAST analysis in our previous response. However, we appreciate that readers may not read the responses and thus we added the following to the manuscript in the methods (in lines 498-509), although this does create a bit oddity in style that results like description in methods.

Note that we attempted BEAST analysis on the complete genome dataset and the total dataset. Unfortunately for the complete genome dataset, TempEst (v1.5.3) analysis found that there was not enough temporal signal to perform BEAST analysis. For the total dataset of Illumina draft genomes and the complete genomes together, there was good temporal signal based on TempEst analysis. However, there were eight genomes with very long terminal branch lengths which led to violation of all the evolutionary models tested in BEAST analysis. One of the eight genomes was a strain from Indonesia isolated in 1961 which diverged the earliest, so we attempted BEAST analysis by 1) removing the eight genomes; 2) keeping the Indonesian one but removing the other seven genomes. In both cases, forcing strain C5 (the prepandemic strain) as the outgroup, the date estimate for the MCRA of the seventh pandemic was 1940s. In particular, in the latter, the effective sample size (ESS) for clock rate and overall priors were very low.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The explanations for the comments by the authors are adequate, and the manuscript has been revised effectively. However, in Figure 3, the gene aadB (also known as ant(2'')-Ia) is marked twice. If this specifies the presence of two gene copies, it should be highlighted distinctly. If there is only one copy, remove the duplicate label. Additionally, ensure that aadB is marked in red like the other AMR encoding genes (if two copies). This correction is needs to be addressed, but it does not require any further review and has to be made before final submission.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #1 (Remarks to the Author):

The explanations for the comments by the authors are adequate, and the manuscript has been revised effectively. However, in Figure 3, the gene aadB (also known as ant(2'')-Ia) is marked twice. If this specifies the presence of two gene copies, it should be highlighted distinctly. If there is only one copy, remove the duplicate label. Additionally, ensure that aadB is marked in red like the other AMR encoding genes (if two copies). This correction is needs to be addressed, but it does not require any further review and has to be made before final submission.

Response:

Thank you for the comments and picked up the error. *aadB* was a duplicated label and we have removed it from Figure 3.