nature portfolio

Peer Review File

Hotspots of genetic change in Yersinia pestis

Corresponding Author: Dr Yujun Cui

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Title: Unraveling evolution's puzzle: Hotspots of genetic change in Yersinia pestis

Authors: Yarong Wu, Youquan Xin, Xiaoyan Yang, Kai Song, Qingwen Zhang, Haihong Zhao, Cunxiang Li, Yong Jin, Yan Guo, Yafang Tan, Yajun Song, Huaiyu Tian, Zhizhen Qi, Ruifu Yang, Yujun Cui

MS Summary:
In this study, the authors examine 3,318 genomes belonging to the species Yersinia pestis and ask whether rates of mutation are even throughout the genome. After arranging the genomes into a phylogeny, they measure rates of variation and find that a cluster of independent, small genomic regions (40 – 4436 bp) exhibit increased variation relative to the genomic average. They find that these genomic 'hot regions' are sometimes found throughout the phylogeny, but other times are only observed in certain clades. Further, they find an association between hot regions and a particular gene class, namely regulatory genes, which were determined using information in the STRING database. They discuss the opposing explanations for these rates of divergence – increased homologous recombination rates (i.e. more horizontal gene transfer in certain genomic regions); increased positive selection for change at hot regions; and increased localised mutation rates (driven by either local nucleotide context or a broader mechanism). They find little evidence for increased homologous recombination and mutable nucleotide motifs, and while they find evidence for diversifying selection and positive epistasis on some of the hot regions, they argue that some of the observed mutations are neutral or deleterious (either synonymous or experimentally shown to not affect an expected phenotype), and so cannot be attributed to selection. They therefore propose a new model, described as a "localised SOS-response", to explain their results. This model suggests that certain bacterial genomic regions transiently increase their mutation rates to increase localised adaptability under stress.

The manuscript was well presented, results were clearly explained and flowed logically, and the proposed model is thought-provoking. However, I do have strong reservations about the theory. The authors acknowledge that their model diverges significantly from previously established mutational/evolutionary theory. As presented, I do not believe that the evidence in this paper is strong enough to support their argument. Biased mutation rates can be difficult to directly demonstrate, most often they are implicated by erasing other possibilities, including selection. I do not feel there is enough evidence to discount selection in this work (see minor comments 2 and 3) or to reconcile this new model with established theory on the evolution of mutation rates.

Major comments:

There is very little evidence from any analysed taxa that selection interacts with localised genomic mutation rates, instead: "selection operates on the genome-wide deleterious mutation rate" (Lynch, 2016). One potential exception to this is the recent evidence presented by Monroe et al. (2022), who reported that gene bodies have evolved lower mutation rates in A. thaliana. However, even this study attempts to match the recognised framework for how selection interacts with mutation rate: "This intuitive model fits established theory showing that adaptive mutation bias could evolve despite drift when the length of sequence affected (L segment) is large" (Monroe, 2022). The authors here go further by saying that 1. Selection acts on very small segments of the genome (as small as 40bp) and that 2. Selection drives mutation rates up (or drives rates down unevenly) to build these genomic hot regions as an evolutionary risk management strategy. This argument requires further evidence:
A. If the authors argue that these regions are under positive selection for increased variation, they must perform further

modelling that reconciles or challenges Lynch's models. Such a model should demonstrate how hot regions could be maintained by selection over generational time despite genetic drift and the increased mutational load in these regions. (The authors argue that not all mutations in these regions will be adaptive and so the load may cause them to degrade as hotbeds of evolvability over time).
B. If the authors argue to repeal selection as the driver but still argue that the hot regions are the result of mutation bias, they should at least present evidence for a viable mechanism that could drive this effect, as they find no evidence of known drivers of local hypermutability.

Minor comments:

1. Did the authors observe a change in mutation spectrum as well as rate in the hot regions? It is highly unlikely that rates of all mutation types would raise simultaneously if indeed a mechanism is causing local mutation rates to spike. Identifying mutation spectrum would also provide a clue as to the mechanism driving the observation.

2. Line 277: "However, the majority of HR variations (93.11%) had low allele frequencies … This suggests that while some variations in HRs may be under positive selection, most are purged over time, contrary to expectation if positive selection was the sole force at play."

What about a combination of selection and genetic drift? The efficiency of selection scales with Ne, so a smaller population could 'purge' these alleles and prevent them reaching fixation. I don't believe this result is suggestive of mutation bias as is implicated by the study.

3. Line 285: "Within HR06, we identified 36 variations … Despite this, only the previously reported K43R mutation in rpsL demonstrated a resistance phenotype … This suggest that many of the variations in HR06 do not offer a selective advantage under antibiotic pressure."

I do not believe this discounts selection as a primary driver. This idea is explored by the authors in the following section –

Line 320: "These segments then become hotbeds of mutation, supplying a pool of genetic diversity for subsequent selection processes. As environmental pressures subside, these regions may revert to genomic stability."

What the authors are suggesting is that segments within the hot regions are subject to fluctuating selection and revert to genomic stability after an environmental stress has relaxed. Therefore, couldn't this explain why mutations in rpsL do not offer a resistance phenotype in many of the strains tested? These strains may have been collected during a period of relaxed stress and genomic stability, and so have reverted their phenotypes. In line with their arguments, this result therefore doesn't discount fluctuating selection as the primary driver of increased variation, even at standardised mutation rates.

4. The NLM model figure should not be presented in the results section without a mathematical model to formalise the argument. It is more appropriate to present this idea in the discussion section of the manuscript.

5. Figure 3: Why is the text of some of the hot regions (on the x axis) in red? Why are they not ordered numerically?

6. In general, I find the arguments for the model teleological, suggesting that these hot regions are maintained for future contingencies. If this is inaccurate – it should be directly addressed in the text. Line 329 compares the NLM hypothesis to the localized SOS response in bacteria, and claims, "that a mechanism may exist within bacteria to selectively accelerate mutation in targeted regions, thereby facilitating the rapid emergence of beneficial traits for survival and adaptation". But this is not what the SOS response in bacteria does – more error-prone DNA polymerases are recruited to repair DNA lesions, and increased mutagenesis is simply a secondary consequence. But there is experimental evidence to show that this does not actually increase evolvability (Torres-Barcelo, Kojadinovic et al 2015). There is a large amount of literature that debates the evolution of evolvability that I feel is missing – a paragraph should be added to the discussion to directly address this.

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)
The authors present a population genomic analysis of Yersinia pestis, that includes many new genomes that have not been sequenced before.

These new genomic data are a valuable contribution to the field. The authors also find that SNP-based genetic variation in Y. pestis is concentrated in a small number of genomic "hotspots".

The authors consider several hypotheses that could drive these patterns of variation based on potential evolutionary "forces" (selection, recombination, mutation) and conclude that selection and recombination are insufficient to explain the observed patterns of variation. They therefore conclude that local variation in mutation rate must be involved in driving these patterns of variation. Further, they present a hypothesis that they call "the Natural Localized Mutagenesis (NLM) model, which posits that localized hypermutation in targeted genomic regions of bacteria may be invoked as an evolutionary risk management strategy. This approach facilitates the rapid emergence of a wide array of mutations, fueling adaptive selection while mitigating the risks of global genomic instability".

My evaluation is that these data are valuable, but that the overall conclusion is highly questionable. It may or may not be true, but the evidence presented to support the final conclusion (the NLM model) is limited and depends on several questionable premises.

Before explaining my concerns, I want to state that my concerns are less related to specific technical aspects of the authors' study, and more related to two key research gaps in the field: 1) a lack of of realistic mathematical models (null models or otherwise) for bacterial genome evolution 2) challenges in drawing definitive conclusions about mutation rates from sequence variation in natural isolates (which is related to research gap 1).

My major concerns with this paper are as follows:

1) unrealistic null model of uniform genomic evolution in bacteria in the absence of selection. This is a research gap in the field due to recent findings. Recent work has shown that E. coli actually shows a wave pattern of local mutation rates in the chromosome, related to chromosome organization in the nucleoid ("The symmetrical wave pattern of base-pair substitution rates across the Escherichia coli chromosome has multiple causes"). I have also found the same wave pattern in synonymous mutations in the Lenski experiment ("Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with Escherichia coli"). Importantly, the wave pattern only occurs in one LTEE population which did not evolve any mutations in topA, fis, and dusB, which are genes that affect DNA topology and chromosome structure. This paper updates my previous work that the authors kindly cited, which found a uniform mutation rate in the LTEE. My new analysis shows that the apparently uniform mutation rate actually depended on both mutator allele as well as the mutations in topA, fis, and dusB.

In addition, several cancer studies have shown variation in genomic mutation rates that is largely explained by chromatin structure and epigenetic variation (see review: "The effects of chromatin organization on variation in mutation rates in the genome")

Based on the authors' interests, it would be incredibly valuable to conduct a mutation accumulation experiment to determine genomic mutation rates in Y. pestis, but this is probably substantial work for a separate paper.

2) Selection in the variation hotspots. The authors find clear evidence of positive selection (dN/dS > 1, multiple homoplasies on the tree (parallel evolution), under the assumption of negligible recombination) in several, but not all hotspots. In itself this is a striking finding in itself, and not a concern. However, positive and negative selection can bias nearby neutral variation in complicated ways that is hard to control for.

A key finding in this analysis is the low frequency of some allelic variants in the hotspots, which the authors interpret as variants which are being purged from the population (i.e. deleterious new mutations which are observed, but being purged by purifying selection). The problem with this interpretation is that those variants may have been beneficial on short timescales, but deleterious on longer timescales (see the paper: "Antagonistic pleiotropy conceals molecular adaptations in changing environments"). The assumption that these variants are new deleterious mutants is hard to justify rigorously, but the authors' implication of higher mutation rates in regions under positive selection depends on this key assumption. An alternative interpretation is that these regions are evolving under strong positive selection, perhaps diversifying selection as well, and some of the variants in these region were under positive selection at first (driving them to be observed), but are now under negative selection (being purged). This makes more sense to me that assuming that these are new deleterious mutations being purged-- if deleterious, then seems unlikely they would observed in large bacterial populations (deleterious mutations should be immediately outcompeted unless hitchhiking with highly beneficial mutations).

3) lack of recombination in Y. pestis. The authors' analysis of recombination depends on one method, ClonalML. However, inferring recombination rates in bacteria is an active area of research, and some methods are wildly discordant with others. A new PNAS paper "Evolution of homologous recombination rates across bacteria" reports three widely diverging recombination/mutation ratio estimates for Yersinia pseudotuberculosis (parent species of Y. pestis) in their Supplementary Table 4: 557, 4.25, and 0.3. The latter number comes from ClonalML, so it is possible that the authors' recombination rate is an underestimate. In addition, another paper in Nature Methods "Inferring bacterial recombination rates from large-scale sequencing datasets" specifically analyze Y. pestis data an find an r/m ratio = 3, suggesting that recombination is a better explanation than mutation for most SNPs. It would be very valuable for the authors to use all four methods (the three referenced in the PNAS paper, and the method in the Nature Methods paper) to see how robust their conclusions about Y. pestis recombination rates are, and clarify the discordance in the literature. Separately, I wonder the extent to which the homoplasies may or may not be affected by recombination, and it would great if the authors could examine this further. Importantly, their authors' analysis depends on the key assumption that recombination is negligible, which needs additional support for robustness.

4) the extent to which synonymous variation (theta_S) in bacterial genomes reflects mutation rate variation. This is a highly questionable assumption. My current opinion is that the Martincorena paper correctly concluded variable mutation rates in E. coli, but using fundamentally flawed methods. My original rebuttal paper essentially made the point that thetaS -> mutation rate variation is highly questionable. My new paper "Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with Escherichia coli" again found no correlation between theta_S and mutation rates in the LTEE, even in the population showing the wave pattern in mutation rates.

I agree with the authors that the finding of elevated theta_S in the hotspots is very interesting, but I do not think this evidence is sufficient to conclude that mutation rate variation is the cause. I do think this is possible-- but I think this requires direct measurements of mutation rate variation, either through MA experiments, or through experiments like in the following paper: "Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing".

5) To my knowledge, there is basically no evidence that the "evolutionary risk-management hypothesis" presented by Martincorena et al. or the authors here can explain mutation rate variation, beyond speculation given their thetaS findings. See a separate rebuttal paper of the Martincorena paper here, based on a different line of argumentation: "No Gene-Specific Optimization of Mutation Rate in Escherichia coli". I do think chromosome structure in the nucleoid could potentially serve as a potential mechanism, but the hypothesis that selection can optimize which genomic regions have higher mutation rates remains controversial, and therefore needs compelling evidence. People have recently discovered targeted error-prone polymerases (diversity-generating elements) that can increase the mutation rate at particular loci as well. However, the authors' hypothesis would be more compelling if they had particular mechanisms in mind, and ways to directly test their hypothesis.

As described the specific evolutionary risk-management hypothesis presented here, in many ways is untestable, since the authors speculate that the localized mutation rates are triggered by some unknown stress. If the stress is not known, then negative experimental results could be attributed to simply not knowing the correct stress to use.

I don't think the authors' hypothesis is necessarily wrong-- simply that it needs more evidence, as well as definitive experimental designs to test.

I think it is more important to generate definitive data on Y. pestis mutation rates before jumping to more complicated hypotheses and conclusions.

6) contribution of copy number variation to apparent mutation rate variation. It is also possible that the hotspots are prone to copy number variation, which results in apparent increases in mutation rate (say, selection drives higher gene copy number, and those copies get mutations in proportion to their copy number). This was the cause of the Cairns "directed mutation" controversy findings-- the lac reporter gene would duplicate to a higher copy number under selection, and those copies would then get mutations at a rate proportional to copy number. This is one possibility that the authors have not yet considered, and worth studying as well as including in their discussion.

Additional comments:

The impact of the Yersinia pestis genome data presented here is limited by the lack of high quality genome assemblies; the analysis depends on alignment to the CO92 genome, and SNPs and indels outside of the CO92 genome or the core genome shared by these strains is ignored.

If the authors are able to include long-read sequencing of representative isolates on their phylogeny, they would be able to generate high-quality complete genome assemblies for different branches of the tree (by rerunning assembly with both short-read and long-read data), and would be able to examine the genome dynamics of Y. pestis over 100 years in much higher detail. My opinion is that such a set of high quality complete genome assemblies could potentially serve as a foundational resource for the Yersinia pestis community.

My overall assessment is that this paper would be improved by including more rigorous analyses of recombination, potentially including long-read data to generate complete genome assemblies and doing a deeper analysis of those complete genomes, and by tempering the language to focus on variation hotspots, and reducing the specific claim that mutation rate variation is necessary to explain their data. While possible, this seems like one of several possibilities that could explain the variation hotspots in Y. pestis (including positive and diversifying selection, and perhaps cryptic recombination as well).

Rohan Maddamsetti

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors provide a thoughtful and thorough response to the review's comments. The additional analyses have strengthened this manuscript and the removal of the NLM model which did not have enough supporting data I think was the

right call. The authors' discussion is also improved, and the appropriate place for exploration of possible hypotheses of a mechanism that requires further investigation. There is value in this type of big data analysis, and it is a nice contribution that highlights our current knowledge gaps bacterial evolution. I think this revised version is a nice paper and I congratulate the authors on their work.

Reviewer #3

(Remarks to the Author)
The authors have carefully revised their work in response to the reviewers' comments. They present one of the largest datasets of Y. pestis genomes, and analyze patterns of variation that can shed light on important open questions in bacterial evolution and genomics.

I have the following minor comments that I hope can further improve their paper:

The first line of the abstract should be edited to reflect the revised focus of the paper. Here is one suggestion: “The relative contributions of mutation rate variation, selection, and recombination in shaping genomic variation in bacterial populations remains poorly understood”.

Also on line 30: “natural selection” -> “positive selection” since negative/purifying selection can account for purging of variation.

Line 84: “genetically monomorphic” needs to be qualified that the authors focus their analysis on the Y. pestis core genome, relative to one focal strain. Should be “… genetically monomorphic nature of the Y. pestis core genome”.

Some fundamental numbers need to be reported here. When considering the collection of 3,318 Y. pestis genomes, how many genes are in this core genome? How many genes in the pangenome? What is size of the biggest genome and number of genes, and size of the smallest genome, and its number of genes?

If these are not critical for main text, these details can go into the Methods. This can help the reader understand some basic facts about genome diversity in this collection of genomes, and get a sense for diversity outside the core genome analyzed here.

Figure 1: the figure itself looks very nice, but the legends and labels are way too small. Legend for panel B in particular is very hard to read because it is so small. Perhaps some postprocessing to ‘cut’ the legend, and ‘paste’ it larger next to the tree could help.

Line 124: It is important to provide a formula for VHR in the Methods since this is a key piece of the authors’ argument in this section. I understand the formula for P in the Methods section “HR mutation distribution in phylogroups”. This formula could be simplified by summing over n rather than m, since the remaining n-m sites contribute 0 to the sum. As I understand, there is a vector of phylogroup proportions per HR, and the variance is calculated over this vector. This could be made clearer. If my understanding is correct, then line 124 should state “we calculated the dispersion of variation across phylogroups per HR” (not “within” as stated in the main text).

The labels in Figure 2B are not very professional, with the under_score characters and abbreviations like GenomeLen and TotalMuts, and this figure is hard to understand. I think it will probably be better to put these numbers into a table, or put these numbers directly into the main text and cut the figure. Or at least, improve the labels and annotate the total numbers for each bar, so that we can understand how the proportions correspond to particular numbers.

Figure 4a: I have no idea what the blue bars and red line are supposed to mean, or why these numbers matter. Furthermore the line is confusing because there is no biological connection between the points connected to the lines. Each red point represents an independent number, and the apparent “spikes” in the distribution don’t seem to be that meaningful for this paper.

These data should be more easily understood using a table with four columns, one for GO Term, one for observed gene count, one for Percentage of Observed Gene Count in Background, and one for FDR. But the blue bars and red line seem to be hard to understand and irrelevant for the point in the text. Why not just have a simple table with two columns, one for GO term, and one for FDR-corrected p-value for enrichment, and then use bold text to highly the GO terms related to regulation? If the blue bars and red line are more meaningful that I understand, then please interpret further in main text. But my understanding is that these are not that important, and even if so, may be better visualized in a table, rather than in a confusing graph with two different x-axes and plots superposed one on another.

In the methods, “Variation hotspot” should be preferred to “mutation hotspot”.

Comment on copy number analysis in the rebuttal letter: copy number variations can be detected by examining read coverage in each of the genomes. Even if repetitive regions are excluded, this does not mean that copy number variations have been excluded in these genomes. From direct experience from sequencing evolved genomes in the lab, copy number variations can be very common, and arise faster than SNPs, although they can be very unstable.

For instance, the authors could examine sequencing read coverage in the HR regions, and ask how often the read coverage

is 2X or even higher than the average read coverage across the core genome. By assuming a negative binomial distribution of reads (poisson is theoretical Lander-Waterman model for sequencing coverage, but in reality the variance is inflated due to experimental factors in library prep, machine etc), one can conduct tests for substantial higher read coverage in the HR regions compared to the mean coverage over the genome. If the authors have the read alignments across the core genomes on hand, and have the read alignments for the HR regions, then they just have to calculate the depth of these regions, the variance in the depth across the core genome, use the mean and variance for the whole core genome to parameterize the negative binomial, and then look at the probability of a HR region with elevated coverage, drawing from the negative binomial representing the genome-wide coverage distribution. Assuming that reads are say, 300bp long, one can break up an HR region into 500bp tiles, run independent tests, and if large regions (like 2000-10,000bp) have elevated coverage, those separate tests can be multiplied (because independent) to generate a p-value, and then that p-value can be Bonferroni-corrected or FDR corrected to get statistical significance for copy number variation. This is something I have done to search for CNVs in genomes during experimental evolution, there may be better approaches, but in my hands this idea is practical and straightforward.

That said, I do not think this CNV analysis is necessary for further revision. An analysis of CNVs from gDNA extracted in the lab may largely reflect growth selection under lab conditions rather than the record of historical selection and mutation processes in natural populations.

Rohan Maddamsetti

Response letter regarding the decision on Nature Communications manuscript # NCOMMS-24-22406-T

Reviewer Comments

Reviewer #1 (Remarks to the Author):

Title: Unraveling evolution's puzzle: Hotspots of genetic change in *Yersinia pestis*

Authors: Yarong Wu, Youquan Xin, Xiaoyan Yang, Kai Song, Qingwen Zhang, Haihong Zhao, Cunxiang Li, Yong Jin, Yan Guo, Yafang Tan, Yajun Song, Huaiyu Tian, Zhizhen Qi, Ruifu Yang, Yujun Cui

MS Summary:

In this study, the authors examine 3,318 genomes belonging to the species *Yersinia pestis* and ask whether rates of mutation are even throughout the genome. After arranging the genomes into a phylogeny, they measure rates of variation and find that a cluster of independent, small genomic regions (40 – 4436 bp) exhibit increased variation relative to the genomic average. They find that these genomic 'hot regions' are sometimes found throughout the phylogeny, but other times are only observed in certain clades. Further, they find an association between hot regions and a particular gene class, namely regulatory genes, which were determined using information in the STRING database. They discuss the opposing explanations for these rates of divergence – increased homologous recombination rates (i.e. more horizontal gene transfer in certain genomic regions); increased positive selection for change at hot regions; and increased localised mutation rates (driven by either local nucleotide context or a broader mechanism). They find little evidence for increased homologous recombination and mutable nucleotide motifs, and while they find evidence for diversifying selection and positive epistasis on some of the hot regions, they argue that some of the observed mutations are neutral or deleterious (either synonymous or experimentally shown to not affect an expected phenotype), and so cannot be attributed to selection. They therefore propose a new model, described as a "localised SOS-response", to explain their results. This model suggests that certain bacterial genomic regions transiently increase their mutation rates to increase localised adaptability under stress.

The manuscript was well presented, results were clearly explained and flowed logically, and the proposed model is thought-provoking. However, I do have strong reservations about the theory. The authors acknowledge that their model diverges significantly from previously established mutational/evolutionary theory. As presented, I do not believe that the evidence in this paper is strong enough to support their argument. Biased mutation rates can be difficult to directly demonstrate, most often they are implicated by erasing other possibilities, including selection. I do not feel there is enough evidence

to discount selection in this work (see minor comments 2 and 3) or to reconcile this new model with established theory on the evolution of mutation rates.

Response: Thank you for your detailed feedback on our manuscript. We have re-assessed our data and considered the suggestions provided by both you and Reviewer #3. Recognizing that our current evidence may not robustly support the initially proposed model, and after careful deliberation and discussion, we have revised the Results and Discussion sections to more clearly highlight natural selection as the primary driver of the observed patterns in hot regions, while considering mutation rate bias as a potential alternative explanation for a minority of these regions with high θ_s. Given the insufficient evidence for the NLM model, we have decided to remove it from our results. Accordingly, we have updated the Abstract and Background sections to reflect these changes. We have carefully addressed each of your comments, as detailed below.

Major comments:

Q1: There is very little evidence from any analysed taxa that selection interacts with localised genomic mutation rates, instead: "selection operates on the genome-wide deleterious mutation rate" (Lynch, 2016). One potential exception to this is the recent evidence presented by Monroe *et al.* (2022), who reported that gene bodies have evolved lower mutation rates in *A. thaliana*. However, even this study attempts to match the recognised framework for how selection interacts with mutation rate: "This intuitive model fits established theory showing that adaptive mutation bias could evolve despite drift when the length of sequence affected (L segment) is large" (Monroe, 2022). The authors here go further by saying that 1. Selection acts on very small segments of the genome (as small as 40bp) and that 2. Selection drives mutation rates up (or drives rates down unevenly) to build these genomic hot regions as an evolutionary risk management strategy. This argument requires further evidence:

A. If the authors argue that these regions are under positive selection for increased variation, they must perform further modelling that reconciles or challenges Lynch's models. Such a model should demonstrate how hot regions could be maintained by selection over generational time despite genetic drift and the increased mutational load in these regions. (The authors argue that not all mutations in these regions will be adaptive and so the load may cause them to degrade as hotbeds of evolvability over time).

B. If the authors argue to repeal selection as the driver but still argue that the hot regions are the result of mutation bias, they should at least present evidence for a viable mechanism that could drive this effect, as they find no evidence of known drivers of local hypermutability.

Response: Thank you for your insightful comment. We recognize that while the influence of natural selection on genome-wide mutation rates is well-established, evidence supporting the fine-tuning of mutation rates within specific genomic regions

remains limited. Upon reassessing our data, we have clarified in the revised Results and Discussion sections that natural selection predominantly shapes the patterns observed in hot regions. We also consider that mutation rate bias may explain a minority of these regions with high θ_s. Please refer to lines 236-264 and 342-351 in the revised main text. Given the lack of robust mathematical models to formalize the NLM model and the insufficient evidence supporting it, we have decided to remove the NLM model from our analysis. Additionally, based on the suggestions from both you and Reviewer #3, we have proposed two possible explanations for the pronounced trend of variation purging within hot regions: transient Darwinian selection and the combined effects of selection and genetic drift (see lines 326-341). The Abstract and Background sections have been revised to ensure consistency with these updates (see lines 27-32 and 73).

Minor comments:

Q2: Did the authors observe a change in mutation spectrum as well as rate in the hot regions? It is highly unlikely that rates of all mutation types would raise simultaneously if indeed a mechanism is causing local mutation rates to spike. Identifying mutation spectrum would also provide a clue as to the mechanism driving the observation.

Response: Thank you for your valuable feedback. According to your suggestion, we analyzed the number and proportion of different types of transitions and transversions across the entire chromosome, mutational hot regions (HRs), and non-hot regions (non-HRs), as shown in Fig. R1 (corresponding to Supplementary Fig. 6b). Our analysis revealed significant differences in the number of the six mutation types in HRs compared to the entire chromosome, non-HRs, and random samples from non-HRs (1000 repeats, each with the same number of mutation sites as the HRs, then averaged) (Chi-squared test, P=9.94×10^{-6}, P=2.77×10^{-4}, and P=0.011, respectively). Similar results were obtained for the five HRs, which had significantly higher θ_s values compared to the entire chromosome, non-HRs, and random samples from non-HRs (P=0.002, P=5.64×10^{-4}, and P=0.0031, respectively). Additionally, we found that HRs with high θ_s showed a greater proportion of G:C->T:A (transversion) and a lower proportion of G:C->A:T (transition) compared to other HRs and non-HRs, based on the calculated standardized residuals. The relevant results have been incorporated into the Results section of our revised manuscript (see lines 241-246). Although the underlying mechanisms remain unclear, we hope our findings will provide some insights for further research into this phenomenon.

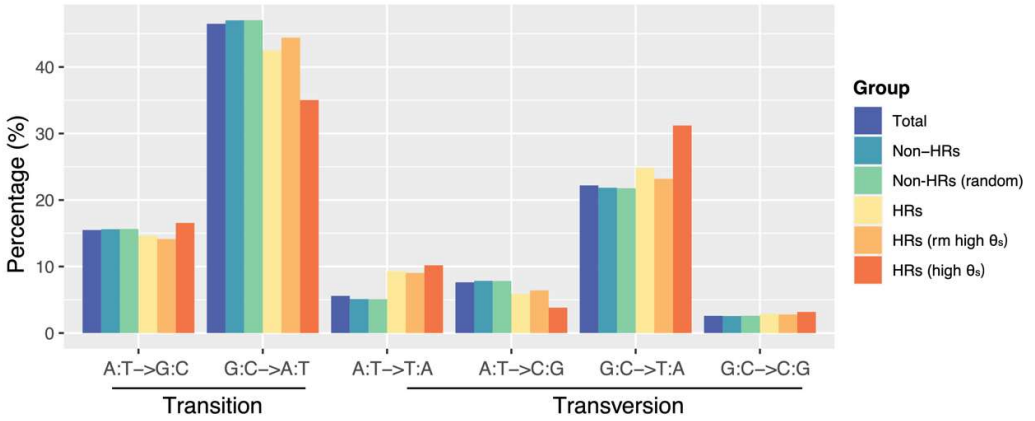

Fig. R1 Distribution of mutation spectrum in *Y. pestis*. The x-axis represents different types of transitions and transversions, while the y-axis indicates the proportion of each mutation type relative to the total number of mutations in each group. Different groups are represented by different colors. The "non-HR random" group refers to the mean value derived from 1000 random samples taken across the genome outside of mutational hot regions (i.e., non-HRs), with each sample containing the same number of SNP sites as found in the hot regions. The "HR (high θ_s)" group consists of the five HRs with elevated θ_s, while the "HR (rm high θ_s) " group includes all HRs except for these five. For each genomic site, SNPs with more than two alleles (i.e., non-biallelic SNPs) were counted multiple times based on changes relative to the reference sequence.

Q3: Line 277: "However, the majority of HR variations (93.11%) had low allele frequencies … This suggests that while some variations in HRs may be under positive selection, most are purged over time, contrary to expectation if positive selection was the sole force at play."

What about a combination of selection and genetic drift? The efficiency of selection scales with Ne, so a smaller population could 'purge' these alleles and prevent them reaching fixation. I don't believe this result is suggestive of mutation bias as is implicated by the study.

Response: Thanks for pointing this out. We acknowledge that a combination of selection and genetic drift may contribute to this phenomenon. *Y. pestis* has a low mutation rate, within a range of 3×10^{-10} to 1×10^{-7} substitutions per site per year, according to our previous study (2013, PNAS, Cui *et al.*, doi: 10.1073/pnas.1205750110) and subsequent studies incorporating ancient DNA (2019, Nature Communications, Spyrou *et al.*, doi: 10.1038/s41467-019-12154-0; 2022, Nature, Spyrou *et al.*, doi: 10.1038/s41586-022-04800-3). An estimation of the median (or mean) effective population size (N_e) of *Y. pestis*, based on the Bayesian Coalescent Skyline model (2018, Nature communications, Spyrou *et al.*, doi: 10.1038/s41467-018-04550-9), was 1×10^{3}-1×10^{4} (95% CI: 1×10^{2}-1×10^{5}), indicating a small N_e for *Y. pestis*. This has been supported by a comparison of N_e across 152 bacterial species and one archaeon (2018, BMC Evolutionary Biology, Bobay and Ochman, doi:

10.1186/s12862-018-1272-4), although the N_e value differed from Spyrou *et al.*'s estimation due to different evaluation methods. Thus, as the efficiency of selection decreases and the influence of genetic drift increases in the population with small N_e, advantageous mutations may be lost during the drift process, preventing them from reaching fixation. We have revised the Discussion section to include this possible explanation. Additionally, based on Reviewer #3's suggestion in Q2, we have also introduced an alternative hypothesis—transient Darwinian selection—to account for the low allele frequencies observed in many HR variations. Please refer to lines 326-341 in the revised main text.

Q4: Line 285: "Within HR06, we identified 36 variations … Despite this, only the previously reported K43R mutation in rpsL demonstrated a resistance phenotype … This suggest that many of the variations in HR06 do not offer a selective advantage under antibiotic pressure."

I do not believe this discounts selection as a primary driver. This idea is explored by the authors in the following section –

Line 320: "These segments then become hotbeds of mutation, supplying a pool of genetic diversity for subsequent selection processes. As environmental pressures subside, these regions may revert to genomic stability."

What the authors are suggesting is that segments within the hot regions are subject to fluctuating selection and revert to genomic stability after an environmental stress has relaxed. Therefore, couldn't this explain why mutations in rpsL do not offer a resistance phenotype in many of the strains tested? These strains may have been collected during a period of relaxed stress and genomic stability, and so have reverted their phenotypes. In line with their arguments, this result therefore doesn't discount fluctuating selection as the primary driver of increased variation, even at standardised mutation rates.

Response: Thank you for pointing this out. Indeed, the observation that "many of the variations in HR06 do not offer a selective advantage under antibiotic pressure" does not rule out the possibility of fluctuating selection as the primary driver of increased variation. This is likely an indication of transient Darwinian selection, resulting from changing environmental selection pressures or epistatic interactions with additional mutations at other sites that could revert the phenotype. Accordingly, we have revised the Results and Discussion section to suggest that selection is the main driver for the emergence of the HRs. Please see lines 249-270 and 326-341 in the revised main text.

Q5: The NLM model figure should not be presented in the results section without a mathematical model to formalise the argument. It is more appropriate to present this idea in the discussion section of the manuscript.

Response: Thank you for your suggestion. Following a re-evaluation of our data and careful consideration and discussion, we now present mutation rate bias as a possible

alternative explanation for the emergence of a minority of HRs (see lines 236-248 and 342-351 in the revised main text). We have also removed the NLM model from our analysis, including the associated figure and related content, due to the absence of a mathematical model to formalize it and the lack of robust supporting evidence.

Q6: Figure 3: Why is the text of some of the hot regions (on the x axis) in red? Why are they not ordered numerically?

Response: Thank you for pointing this out. The HRs labeled in red in Figure 3 indicate the presence of regulator-related genes within those regions (refer to Supplementary Table 6). The numbering of these HRs corresponds to their positions on the reference chromosome. However, Figure 3 is organized based on the clustering of the proportions of variant sites across different phylogroups for each HR by columns. As a result, the arrangement of the HRs follows the clustering results rather than numerical order. We have revised the figure legend in Figure 3, as detailed in lines 800 and 802-803 of the revised main text.

Q7: In general, I find the arguments for the model teleological, suggesting that these hot regions are maintained for future contingencies. If this is inaccurate – it should be directly addressed in the text. Line 329 compares the NLM hypothesis to the localized SOS response in bacteria, and claims, "that a mechanism may exist within bacteria to selectively accelerate mutation in targeted regions, thereby facilitating the rapid emergence of beneficial traits for survival and adaptation". But this is not what the SOS response in bacteria does – more error-prone DNA polymerases are recruited to repair DNA lesions, and increased mutagenesis is simply a secondary consequence. But there is experimental evidence to show that this does not actually increase evolvability (Torres-Barcelo, Kojadinovic et al 2015). There is a large amount of literature that debates the evolution of evolvability that I feel is missing – a paragraph should be added to the discussion to directly address this.

Response: Thank you for your valuable feedback. We understand your concerns regarding the robustness of our conclusions and the emphasis on the NLM model. After re-evaluating our data and careful consideration, we have removed the NLM model from our analysis. The revised Results and Discussion sections now present natural selection as the primary driver of the observed patterns in hot regions, while considering mutation rate bias as a possible explanation for the high θ_s in a few HRs compared to other genomic regions (see lines 236-264 and 342-351 in the revised main text). Additionally, as you correctly pointed out, our previous mention of the localized SOS response was not in line with the original definition of SOS (1967, PNAS, Witkin, doi: 10.1073/pnas.57.5.1275; 1975, (eds) Molecular Mechanisms for Repair of DNA, Basic Life Sciences, Radman, doi: 10.1007/978-1-4684-2895-7_48). To prevent confusion, we have removed this analogy to the SOS response in bacteria. Finally, as we have

reduced the focus on mutation rate bias, we did not include additional discussion on the debate regarding the evolution of evolvability in the Discussion section.

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.

Response: Thank you for participating in the co-review process. We have provided a point-by-point response to each comment.

Reviewer #3 (Remarks to the Author):

The authors present a population genomic analysis of *Yersinia pestis*, that includes many new genomes that have not been sequenced before.

These new genomic data are a valuable contribution to the field. The authors also find that SNP-based genetic variation in *Y. pestis* is concentrated in a small number of genomic "hotspots".

The authors consider several hypotheses that could drive these patterns of variation based on potential evolutionary "forces" (selection, recombination, mutation) and conclude that selection and recombination are insufficient to explain the observed patterns of variation. They therefore conclude that local variation in mutation rate must be involved in driving these patterns of variation. Further, they present a hypothesis that they call "the Natural Localized Mutagenesis (NLM) model, which posits that localized hypermutation in targeted genomic regions of bacteria may be invoked as an evolutionary risk management strategy. This approach facilitates the rapid emergence of a wide array of mutations, fueling adaptive selection while mitigating the risks of global genomic instability".

My evaluation is that these data are valuable, but that the overall conclusion is highly questionable. It may or may not be true, but the evidence presented to support the final conclusion (the NLM model) is limited and depends on several questionable premises.

Before explaining my concerns, I want to state that my concerns are less related to specific technical aspects of the authors' study, and more related to two key research gaps in the field: 1) a lack of of realistic mathematical models (null models or otherwise) for bacterial genome evolution 2) challenges in drawing definitive conclusions about mutation rates from sequence variation in natural isolates (which is related to research gap 1).

Response: We appreciate the valuable comments you have provided. We fully agree with the two key research gaps in bacterial studies you highlighted. And we understand you concerns regarding the NLM model. After re-evaluating our data and engaging in thorough discussion, we have revised the Results and Discussion sections (see lines

236-264 and 342-351 in the revised main text). We now attribute the observed patterns in hot regions primarily to natural selection, with mutation rate bias considered as a potential alternative explanation for a minority of HRs. Given the lack of a realistic mathematical model and insufficient supporting evidence, we have decided to exclude the NLM model from our analysis.

Nonetheless, with your profound insights and advice, we hope our research can provide data support and offer new perspectives to help advance the understanding of these gaps and encourage further in-depth exploration. Additionally, based on our current work, we have planned to conduct mutation accumulation experiments or maximum-depth sequencing-like experiments in *Y. pestis*. However, as you mentioned in comment Q1, considering the substantial work involved, we intend to publish these findings in a separate paper. Please find our detailed responses listed below.

My major concerns with this paper are as follows:

Q1: unrealistic null model of uniform genomic evolution in bacteria in the absence of selection. This is a research gap in the field due to recent findings. Recent work has shown that *E. coli* actually shows a wave pattern of local mutation rates in the chromosome, related to chromosome organization in the nucleoid ("The symmetrical wave pattern of base-pair substitution rates across the *Escherichia coli* chromosome has multiple causes"). I have also found the same wave pattern in synonymous mutations in the Lenski experiment ("Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with *Escherichia coli*"). Importantly, the wave pattern only occurs in one LTEE population which did not evolve any mutations in *topA*, *fis*, and *dusB*, which are genes that affect DNA topology and chromosome structure. This paper updates my previous work that the authors kindly cited, which found a uniform mutation rate in the LTEE. My new analysis shows that the apparently uniform mutation rate actually depended on both mutator allele as well as the mutations in *topA*, *fis*, and *dusB*.

In addition, several cancer studies have shown variation in genomic mutation rates that is largely explained by chromatin structure and epigenetic variation (see review: "The effects of chromatin organization on variation in mutation rates in the genome")

Based on the authors' interests, it would be incredibly valuable to conduct a mutation accumulation experiment to determine genomic mutation rates in *Y. pestis*, but this is probably substantial work for a separate paper.

Response: Thanks for your insightful suggestion. Indeed, chromatin structure and epigenetic variation can influence genomic mutation rates, as has also been reported in *Arabidopsis thaliana* (2022, Nature, Monroe *et al.*, doi: 10.1038/s41586-021-04269-6). The absence of a realistic null model makes it difficult to definitively determine genomic mutation rates in bacteria during evolution, particularly in natural isolates. We plan to conduct mutation accumulation experiments or maximum-depth sequencing-like experiments in *Y. pestis*, aiming to minimize the impact of natural selection, and

then integrate mutation rate analysis with chromatin structure and epigenetic variation. We hope this will contribute to the quantitative study of mutation rate changes in bacterial chromosomes. This work will be presented in a separate paper, and we look forward to further discussions with you in the future.

Q2: Selection in the variation hotspots. The authors find clear evidence of positive selection (dN/dS > 1, multiple homoplasies on the tree (parallel evolution), under the assumption of negligible recombination) in several, but not all hotspots. In itself this is a striking finding in itself, and not a concern. However, positive and negative selection can bias nearby neutral variation in complicated ways that is hard to control for.

A key finding in this analysis is the low frequency of some allelic variants in the hotspots, which the authors interpret as variants which are being purged from the population (i.e. deleterious new mutations which are observed, but being purged by purifying selection). The problem with this interpretation is that those variants may have been beneficial on short timescales, but deleterious on longer timescales (see the paper: "Antagonistic pleiotropy conceals molecular adaptations in changing environments"). The assumption that these variants are new deleterious mutants is hard to justify rigorously, but the authors' implication of higher mutation rates in regions under positive selection depends on this key assumption. An alternative interpretation is that these regions are evolving under strong positive selection, perhaps diversifying selection as well, and some of the variants in these region were under positive selection at first (driving them to be observed), but are now under negative selection (being purged). This makes more sense to me that assuming that these are new deleterious mutations being purged-- if deleterious, then seems unlikely they would observed in large bacterial populations (deleterious mutations should be immediately outcompeted unless hitchhiking with highly beneficial mutations).

Response: Thanks for your comment. We agree with your perspective that the low frequency of certain allelic variants in hotspots could result from an initial phase of positive selection followed by subsequent negative selection. This interpretation aligns with the concept of transient Darwinian selection, as demonstrated in population genomics studies on *Salmonella enterica* serovar Paratyphi A (2014, PNAS, Zhou *et al.*, doi: 10.1073/pnas.1411012111). We have incorporated this as a possible interpretation in the Discussion section (see lines 327-334 in the revised main text). Additionally, we present another hypothesis in the Discussion section (see lines 334-339), which considers the combined effects of selection and genetic drift, as noted by Reviewer #1 in Q3. However, due to the complexity of natural environmental changes, the lack of reliable mathematical models, and insufficient data, a quantitative analysis of this issue remains unfeasible.

Q3: lack of recombination in *Y. pestis*. The authors' analysis of recombination depends on one method, ClonalML. However, inferring recombination rates in bacteria is an

active area of research, and some methods are wildly discordant with others. A new PNAS paper "Evolution of homologous recombination rates across bacteria" reports three widely diverging recombination/mutation ratio estimates for *Yersinia pseudotuberculosis* (parent species of *Y. pestis*) in their Supplementary Table 4: 557, 4.25, and 0.3. The latter number comes from ClonalML, so it is possible that the authors' recombination rate is an underestimate. In addition, another paper in Nature Methods "Inferring bacterial recombination rates from large-scale sequencing datasets" specifically analyze *Y. pestis* data an find an r/m ratio = 3, suggesting that recombination is a better explanation than mutation for most SNPs. It would be very valuable for the authors to use all four methods (the three referenced in the PNAS paper, and the method in the Nature Methods paper) to see how robust their conclusions about *Y. pestis* recombination rates are, and clarify the discordance in the literature. Separately, I wonder the extent to which the homoplasies may or may not be affected by recombination, and it would great if the authors could examine this further. Importantly, their authors' analysis depends on the key assumption that recombination is negligible, which needs additional support for robustness.

Response: Thanks for pointing this out. *Y. pestis* is a recently emerged clone of *Y. pseudotuberculosis*, exhibiting low genetic diversity compared to its ancestor. Given that different methods and dataset sizes can impact analysis results, we have reanalyzed homologous recombination in both *Y. pestis* and *Y. pseudotuberculosis* under different dataset sizes using three different methods—ClonalFrameML (2015, PLOS Computational Biology, Didelot & Wilson), mcorr (2019, Nature Methods, Lin & Kussell), and another commonly used bacterial recombination software for large samples, Gubbins (2015, Nucleic Acids Research, Croucher *et al.*). Due to the high computational costs and significant disk space requirements of recABC (PNAS, 2024, Torrance *et al.*)—as noted in its README file, where a full run for 15 *Bacillus safensis* strains of 1.2-Mb alignments would take approximately 15 days and require around 140 GB of storage—we did not include this method in our analysis given the large scale of our dataset.

We acknowledge an error in our initial calculation of the recombination/mutation ratio (r/m) using ClonalFrameML, which led to the underestimation of the recombination rate in *Y. pestis*. The correct formula should have been r/m = (R/theta)/(1/delta)*nu, but we mistakenly used r/m = R/theta*(1/delta)*nu. Upon reanalysis, the r/m for *Y. pestis* ranged from 0.2 to 1.4 in ClonalFrameML, 0.2 to 1.5 in Gubbins, and 0.3 to 5.1 in mcorr (as detailed in Table R1 below).

Table R1. Inferred recombination parameters for *Y. pestis* and *Y. pseudotuberculosis* using different methods across different dataset sizes

Methods	Parameters	*Y. pestis*								*Y. pseudotuberculosis*	
		25S_run1#	**25S_run2#**	**64S_run1#**	**64S_run2#**	**186S_phylo1#**	**186S_phylo3#**	**1167S#**	**3318S**	**23S**	**559S**
ClonalFrameML	R/theta	0.0405318	0.0415261	0.0421348	0.0275513	0.0646501	0.0724579	0.119931	0.203946	0.264248	0.332961
	1/delta	0.00863485	0.0136034	0.0138607	0.0131459	0.0133575	0.0122634	0.0112573	0.0118904	0.00345528	0.00351633
	nu	0.09575	0.0893729	0.0813196	0.111126	0.080501	0.0868895	0.087517	0.0831898	0.0302724	0.0248206
	r/m	0.44944844	0.27282209	0.24720145	0.23289891	0.38962363	0.51338378	0.93237289	1.42688446	2.31512964	2.350261
	total_rec_cov#	0.00151511	0.00060927	0.00077454	0.00059843	0.00116723	0.00176855	0.00352217	0.00459713	0.71391057	0.87510376
mcorr	d_sample	5.9032E-05	4.87E-05	4.46E-05	4.40E-05	4.95E-05	4.42E-05	2.89E-05	2.93E-05	0.0132905	0.00562177
	theta_pool	0.0354488	0.0192678	0.0126326	0.0188637	0.0298331	0.0133325	0.0111916	0.00843078	0.0499017	0.0357244
	phi_pool	0.0105522	0.0517176	0.0249552	0.0252925	0.0144178	0.0298504	0.0276213	0.0429444	0.220919	0.135684
	r/m (gamma/mu)	0.297676 (0.229407-1.19047e+06)	2.68415 (0.0522536-1.67023e+19)	1.97547 (0.0484444-2.24226e+14)	1.3408 (0.197623-6.26948e+20)	0.483283 (0.12329-2.29022e+23)	2.23892 (0.0170158-1.38567e+10)	2.46804 (1.26916-10.6506)	5.09377 (1.5439-18.506)	4.42708 (2.6476-6.62872)	3.79807 (2.44646-5.50659)
	fbar	730.448	177.589	290.849	229.788	171.082	271.969	779.849	789.18	1562.22	881.386
	sample_rec_cov (c)	0.00144872	0.00222106	0.00296774	0.0018901	0.00105646	0.00284311	0.00244669	0.00336175	0.282744	0.162722
Gubbins	r/m*	0.4018143	0.20464881	0.18934316	0.16953317	0.33333333	0.46482476	0.89334392	1.48392415	0.64126271	0.97258105
	rho/theta*	0.01760939	0.01667509	0.01642091	0.01193401	0.02475979	0.02967511	0.06073839	0.1071723	0.02307059	0.0413309

	total_rec_cov#	0.00517966	0.00405941	0.00326728	0.00382957	0.00635339	0.00570341	0.00758374	0.01595118	0.64899704	0.98911077
	sample_rec_cov#	0.00190539	3.73E-05	9.53E-05	1.83E-05	8.19E-05	3.73E-05	4.60E-05	6.91E-05	0.18299045	0.30241898

* According to the Gubbins manual, the overall r/m ratio in Gubbins was calculated by dividing the sum of the number of SNPs inside recombination regions by the sum of the number of SNPs outside recombinations. Similarly, the overall R/theta was determined by dividing the sum of the number of recombination blocks by the sum of the number of SNPs outside recombination regions.

The total recombination coverage (total_rec_cov) is calculated as the ratio of the total length of reference-mapped recombination regions to the length of the non-repetitive core genome. The sample-related recombination coverage (sample_rec_cov) is the median of all individual ratios of the total length of recombination regions associated with each strain, including both internal branches and strain-specific events, to the length of the non-repetitive core genome.

However, the conclusion of our study—that recombination is rare in *Y. pestis*—remains unaffected; the error's primary impact was solely on the evaluation of the r/m ratio. It is important to note that besides the r/m ratio, which only considers genomic regions affected by recombination, the recombination coverage (i.e., the proportion of sites in a genome whose diversity originates from outside the sample) is another crucial factor that should be taken into account (2019, Nature Methods, Lin & Kussell).

We found that the recombination regions in *Y. pestis* are mainly scattered within individual strains, whereas *Y. pseudotuberculosis* exhibits more recombination events affecting internal branches, shared by multiple isolates through common descent (see Fig. R2 below). When we mapped the recombination regions identified in individual strains and internal branches to the reference chromosome, we found that the total recombination regions (merged overlapping regions) identified by ClonalFrameML accounted for 0.06%-0.5% of the non-repetitive core genome (present in 95% of strains) in different dataset sizes of *Y. pestis*, compared to 71.39%-87.51% in *Y. pseudotuberculosis*, with similar results also found in Gubbins (Table R1).

Additionally, since the outputs of Gubbins provides the strain names involved in each internal branch, we further evaluated the proportion of recombination regions in each strain. We found that only 816 out of 3318 *Y. pestis* strains (24.6%) had recombination regions identified by Gubbins, and the median recombination coverage in these 816 strains was 0.007% per stain. In contrast, all strains of *Y. pseudotuberculosis* had recombination regions being identified, with a median recombination coverage of 18.3%-30.2% per strain. The mcorr result also found similar result, that *Y. pestis* has a lower recombination coverage (range: 0.1%-0.3%) compared to its ancestor (range: 16.3%-28.3%) or other species (Supplementary Table 3 in Nature Methods), although our estimated recombination coverage in *Y. pestis* was lower than that reported in Nature Methods (3.3%). Taken together, these results implies that there are fewer recombination events contributing to the genetic diversity of the *Y. pestis* genome, with mutation being the predominant source of variation.

Furthermore, our investigation into the impact of recombination on mutational hot regions (HRs) in *Y. pestis* revealed minimal overlap. Only one HR (HR06) was found to partially overlap with a recombination region within the *tuf* gene (YPO0203), which encodes elongation factor Tu, identified by both ClonalFrameML and Gubbins in a single *Y. pestis* strain. However, the overlap accounts for just 5% of HR06's length, with only 4 out of 36 variations in HR06 falling within this segment. Excluding this region would not affect HR06's identification as a mutational hotspot. Additionally, a Gubbins-identified recombination region (1,228 bp) in one strain partially overlapped with HR026, covering 17.9% of HR026's length, with 1 out of 25 variations in HR026 falling within the shared region. Nine ClonalFrameML regions also overlapped with eight HRs, but these were short (2-17 bp) and strain-specific. These results were consistent with our findings of rare physically linked HR-related variations within the same genome (see Supplementary Fig. 7f and lines 218-224 in the main text).

Thus, despite a variable r/m ratio (0.17-5.09) across different datasets and tools, the recombination coverage was consistently low (≤1.6% overall, median ≤0.3% per sample), and recombination regions rarely overlapped with HRs. Collectively, these findings support that genetic diversity in *Y. pestis* is largely intrinsic to the sample, indicating that recombination plays a minor role in shaping HRs. We have supplemented these recombination reanalysis results using three methods under different datasets in both the Results and Methods sections and revised our original description of the r/m ratio, as detailed on lines 203-208, 213-217, and 537-563 of the revised main text. We have also added Table R1 and Figure R2 for additional information (see Supplementary Table 8 and Supplementary Fig. 7a-d).

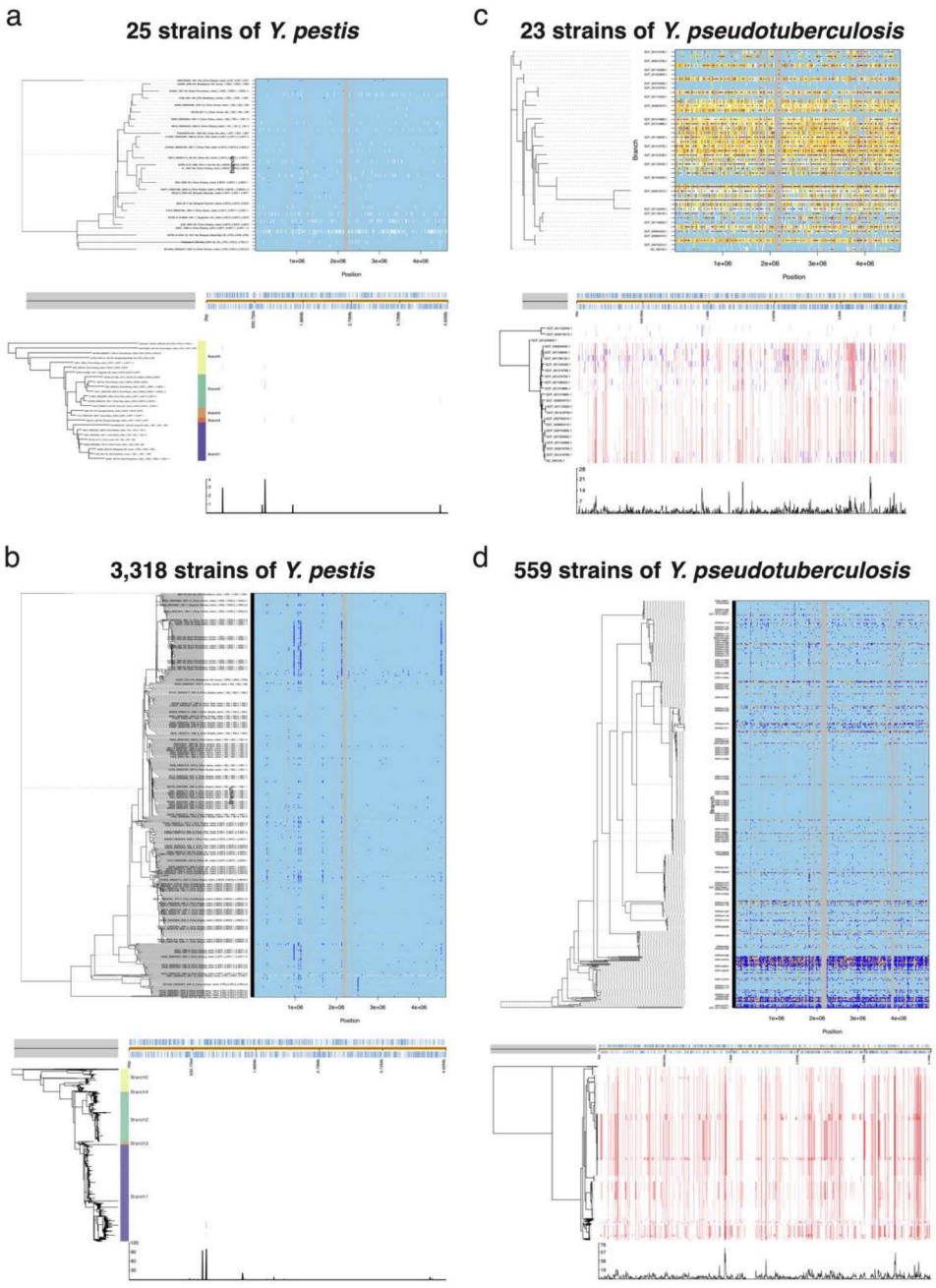

Fig. R2 Analysis of recombination in *Y. pestis* and *Y. pseudotuberculosis* using ClonalFrameML and Gubbins across different dataset sizes. a) 25 strains of *Y. pestis*; b) 3,318 strains of *Y. pestis*; c) 23 strains of *Y. pseudotuberculosis*; d) 559 strains of *Y. pseudotuberculosis*. Top panels show ClonalFrameML results, while bottom panels show Gubbins results. In ClonalFrameML, dark blue bars indicate recombination events, colored vertical bars represent branch substitutions, and grey areas are non-core regions. In Gubbins, red blocks indicate predicted recombinations on internal branches (shared among multiple strains due to common descent), and blue blocks represent recombinations unique to individual strains. The colored bars between the ML tree and heatmap represent the five major branches of *Y. pestis* (Branch 0-4).

As for the significant variation in r/m ratio estimates for *Y. pseudotuberculosis*, noted in Supplementary Table 4 (557, 4.25, and 0.3; PNAS, 2014), it can be attributed to differences in the dataset and methods used. The first (557) and last (0.3) values originate from ClonalFrame (the predecessor of ClonalFrameML), with the former based on a limited whole-genome core alignment of four complete genomes, and the latter from an analysis of seven genes across 43 MLST sequence types (STs). These discrepancies underscore the influence of sampling bias and genomic scope on the estimates. In the same publication's Supplementary Table 3, the r/m ratio comparison between recABC and ClonalFrameML, using core alignments of 23 representative *Y. pseudotuberculosis* genomes, yielded values of 4.25 and 2.16, respectively. Our reanalysis, employing non-repetitive core alignments of the same 23 strains and an expanded dataset of 559 strains with ClonalFrameML and mcorr, produced comparable results: 2.32 and 2.35 for ClonalFrameML, and 4.43 and 3.80 for mcorr (Table R1). However, Gubbins provided a lower r/m ratio for *Y. pseudotuberculosis* (0.64 and 0.97 across two datasets), showing less consistency with ClonalFrameML than observed in *Y. pestis*. Taken together, these discrepancies and results highlights the impact that software choice, dataset size, and alignment methods can have on recombination analysis outcomes. Your recommendation to utilize different software tools for recombination analysis to ensure the robustness of the findings is indeed valuable.

Finally, we examined the impact of recombination on homoplasies. We focused on the output files from ClonalFrameML and Gubbins, which provide specific site information for recombination regions to determine whether the homoplastic sites are located within these regions. The results showed that 7 homoplasies, belonging to 2 genes (YPO0348/*aspA* and YPO2210) and one intergenic region, were located in three ClonalFrameML-identified recombination regions, with lengths ranging from 2 to 17 bp (see Table R2 below). Six of these homoplasies were involved in hot regions HR07 and HR45 and were located in recombination regions present in individual strains. No homoplasies were found in the recombination regions identified by Gubbins.

Table R2. Homoplasies located within ClonalFrameML-identified recombination regions

Index	Ref_pos	Ref_allele	Mut_allele	Ref_gene	HR_ID	Rec_start	Rec_end	Rec_length	Rec_type
1	358874	C	A,G,T	YPO0348/*aspA*	HR07	358874	358875	2	individual
2	358875	A	G						
3	2484440	A	G	YPO2210	NA	2484440	2484441	2	internal node
4	4532763	T	A	intergenic region	HR45	4532761	4532777	17	individual
5	4532767	G	A						
6	4532769	C	T						
7	4532775	C	A,T						
8	4532777	A	G						

Q4: the extent to which synonymous variation (theta_S) in bacterial genomes reflects mutation rate variation. This is a highly questionable assumption. My current opinion is that the Martincorena paper correctly concluded variable mutation rates in *E. coli*, but using fundamentally flawed methods. My original rebuttal paper essentially made the point that thetaS -> mutation rate variation is highly questionable. My new paper "Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with *Escherichia coli*" again found no correlation between theta_S and mutation rates in the LTEE, even in the population showing the wave pattern in mutation rates.

I agree with the authors that the finding of elevated theta_S in the hotspots is very interesting, but I do not think this evidence is sufficient to conclude that mutation rate variation is the cause. I do think this is possible-- but I think this requires direct measurements of mutation rate variation, either through MA experiments, or through experiments like in the following paper: "Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing".

Response: We acknowledge that our current analysis lacks direct measurements of mutation rate variation and relies on the key assumption that recombination is negligible. From our reanalysis of recombination in both *Y. pestis* and its parent *Y. pseudotuberculosis* using different dataset sizes and methods (as discussed in our previous comment, Q3), we reached a more reliable conclusion. Although the r/m ratio of *Y. pestis* was indeed underestimated in our initial analysis, the overall recombination coverage in *Y. pestis* is low, whether mapping the recombination regions of all strains to the reference sequence or examining individual samples. Additionally, the mutational hot regions are minimally affected by recombination, and linkage events in these hot regions are rare. Therefore, we infer that recombination is negligible in *Y. pestis*.

Based on this, we assumed the core genome of *Y. pestis* has experienced the same effective population size (N_e). Given the neutral theory of molecular evolution ($\theta = 2N_e\mu$), and assuming N_e remains relatively constant, we use synonymous diversity (θ_s) as a measure of mutation rates. However, this approach may be controversial, as indicated by findings from the *E. coli* LTEE studies. We speculate that the disconnect between θ_s and μ observed in the LTEE studies may be due to factors such as variations in homologous recombination rates, which can lead to inconsistencies in the genetic characteristics of the populations.

Due to the lack of realistic mathematical model and direct supporting evidence, we have decided to exclude the NLM model from our analysis and moderated our initial assertion regarding mutation rate variation. We now consider mutation rate bias as one of the possible explanations for the emergence of hot regions in *Y. pestis* (see lines 236-264 and 342-351 in the revised main text). According to your suggestion, we plan to conduct mutation accumulation experiments or maximum-depth sequencing-like experiments in *Y. pestis* in future studies to determine genomic mutation rates directly.

Q5: To my knowledge, there is basically no evidence that the "evolutionary risk-management hypothesis" presented by Martincorena *et al.* or the authors here can explain mutation rate variation, beyond speculation given their thetaS findings. See a separate rebuttal paper of the Martincorena paper here, based on a different line of argumentation: "No Gene-Specific Optimization of Mutation Rate in *Escherichia coli*". I do think chromosome structure in the nucleoid could potentially serve as a potential mechanism, but the hypothesis that selection can optimize which genomic regions have higher mutation rates remains controversial, and therefore needs compelling evidence. People have recently discovered targeted error-prone polymerases (diversity-generating elements) that can increase the mutation rate at particular loci as well. However, the authors' hypothesis would be more compelling if they had particular mechanisms in mind, and ways to directly test their hypothesis.

As described the specific evolutionary risk-management hypothesis presented here, in many ways is untestable, since the authors speculate that the localized mutation rates are triggered by some unknown stress. If the stress is not known, then negative experimental results could be attributed to simply not knowing the correct stress to use.

I don't think the authors' hypothesis is necessarily wrong-- simply that it needs more evidence, as well as definitive experimental designs to test.

I think it is more important to generate definitive data on *Y. pestis* mutation rates before jumping to more complicated hypotheses and conclusions.

Response: We acknowledge that aside from elevated synonymous variation, our current evidence for localized mutation rates is insufficient. We have revised the manuscript to reduce the claim on mutation rate bias and present it as one possible explanation for a minority of HRs (see lines 236-264 and 342-351 in the revised main text). In future studies, we plan to validate this phenomenon by seeking potential evidence through mutation accumulation experiments and maximum-depth sequencing-like experiments. As for the underlying molecular mechanisms responsible for this phenomenon, there is not yet a clear understanding, and further investigation in the relevant fields may be needed once the phenomenon is confirmed.

Q6: contribution of copy number variation to apparent mutation rate variation. It is also possible that the hotspots are prone to copy number variation, which results in apparent increases in mutation rate (say, selection drives higher gene copy number, and those copies get mutations in proportion to their copy number). This was the cause of the Cairns "directed mutation" controversy findings-- the lac reporter gene would duplicate to a higher copy number under selection, and those copies would then get mutations at a rate proportional to copy number. This is one possibility that the authors have not yet considered, and worth studying as well as including in their discussion.

Response: We agree that gene copy number variation is a potential factor that influences mutation rates. However, the presence of repetitive sequences during the SNP and Indel calling processes may cause the mismapping of short reads, potentially

leading to inaccurate identification of genetic variants. To address this, we excluded mutations found within repetitive regions of the reference genome, including tandem and interspersed repeats. These were identified using Tandem Repeat Finder (minimum alignment score 50) and BLAST software (for nucleotide identities ≥95%), as detailed in the Methods section (lines 402-403). The excluded repetitive regions total approximately 326 kb, which constitutes 7% of the reference chromosome, with over half of this length (~170 kb) due to four prevalent types of insertion sequences (IS) of the *Y. pestis* genome. As a result, mutations within duplicated genes or genomic fragments of the reference genome were excluded from our study.

To address the potential impact of undetected duplications not preset in the reference, we extracted sequences from 45 HRs in the reference genome and compared them against 64 complete genome sequences from the NCBI GenBank database, as well as nearly 200 newly sequenced complete genomes from our ongoing research. Our analysis revealed that only HR06 contained a 413 bp segment (11.8% of the 3,499 bp region) with 88.9% nucleotide identity to another genomic region, indicating duplication. This segment is part of the *tuf* gene, which also exhibited a recombination signal (as detailed in Q3). Further BLAST analysis identified this gene as an out-paralog that predates the emergence of *Y. pestis*, due to the observation of a similar pattern in its ancestor, *Y. pseudotuberculosis*. No other HRs showed evidence of copy number variation. This indicates that the influence of gene or genomic fragment copy number variation on the identified HRs in our study is minimal.

Given the intriguing relationship between gene copy number variations and mutation rates, we plan to investigate these duplicated regions in future studies. We will employ complete genome sequences and integrate both short-read and long-read sequencing data to conduct a comprehensive analysis of the underlying dynamics.

Additional comments:

Q7: The impact of the *Yersinia pestis* genome data presented here is limited by the lack of high quality genome assemblies; the analysis depends on alignment to the CO92 genome, and SNPs and indels outside of the CO92 genome or the core genome shared by these strains is ignored.

Response: Yes, we selected CO92 (the most commonly used sequence for *Y. pestis*) as the reference genome and conducted mutational hot region analysis based on SNPs and indels identified from the core genome. Nonetheless, it is important to note that our core genome is a "soft-core genome", meaning it does not require presence in all strains. This approach reduces the loss of information due to the absence of certain regions in individual strains, which may result from poor sequencing quality or inherent gaps. In this study, the soft-core genome consists of genomic sequences present in at least 95% of the strains, totaling 4.15 Mb in length and representing 89.2% of the reference chromosome.

Furthermore, we analyzed gene gain and loss in 64 publicly available complete genome sequences of *Y. pestis* using Prokka and Roary software (-cd 95 -i 90), with additional verification through BLASTn. Our analysis revealed that *Y. pestis* undergoes few gene acquisition events during its evolution (see Fig. R3 below). Specifically, only 108 out of 4,000 pan-genes (2.7%) are absent in the CO92 chromosome, defined as having less than 90% nucleotide identity or less than 90% length coverage. This finding is consistent with our previous research and other studies which concluded that *Y. pestis* has a closed pangenome (2013, PNAS, Cui *et al.*, doi: 10.1073/pnas.1205750110; 2024, Peer Community Journal, Parmigiani *et al.*, 10.24072/pcjournal.415). Therefore, the impact of excluding SNPs and indels outside the CO92 genome or the core genome shared by 95% of strains on our results is minimal.

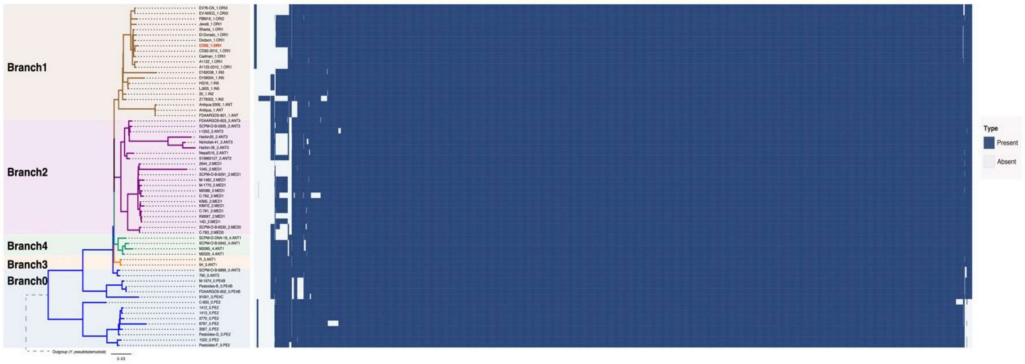

Fig. R3 Gene gain and loss analysis in 64 publicly available complete genome sequences of *Y. pestis*. On the left is the phylogenetic tree of *Y. pestis*, with the five major branches marked in different colors and the reference CO92 highlighted in red text. The heatmap on the right represents the distribution of the 4,000 pan-genes across the strains. The heatmap shows strains in rows and genes in columns, with dark blue marking gene presence and light gray indicating absence.

Q8: If the authors are able to include long-read sequencing of representative isolates on their phylogeny, they would be able to generate high-quality complete genome assemblies for different branches of the tree (by rerunning assembly with both short-read and long-read data), and would be able to examine the genome dynamics of *Y. pestis* over 100 years in much higher detail. My opinion is that such a set of high quality complete genome assemblies could potentially serve as a foundational resource for the *Yersinia pestis* community.

Response: Thank you for the suggestion. Regarding the generation of high-quality complete genome assemblies using long-read sequencing, we are currently undertaking this work. Given the higher cost of third-generation sequencing compared to second-generation sequencing, we have selected approximately 200 representative strains for long-read sequencing. Our preliminary results indicated that the complete sequences do not impact the current identification of core genome SNPs and indels, which aligns with our expectations as described in response to Q7. We'll present a detailed analysis of these complete sequences in a separate paper, including deciphering other important

genetic characteristics and the high frequency of genome rearrangements across the *Y. pestis* population. We hope that these new complete sequences will serve as a foundational resource for the *Y. pestis* community, as you have suggested.

Q9: My overall assessment is that this paper would be improved by including more rigorous analyses of recombination, potentially including long-read data to generate complete genome assemblies and doing a deeper analysis of those complete genomes, and by tempering the language to focus on variation hotspots, and reducing the specific claim that mutation rate variation is necessary to explain their data. While possible, this seems like one of several possibilities that could explain the variation hotspots in *Y. pestis* (including positive and diversifying selection, and perhaps cryptic recombination as well).

Response: Thank you for your valuable comments and suggestions. We have conducted a thorough re-analysis of recombination within the *Y. pestis* genome, as detailed in Q3. In light of the feedback from both you and Reviewer #1, we have re-evaluated our data and placed less emphasis on mutation rate variation. The Results and Discussion sections have been revised to more clearly highlight natural selection as the primary driver of the observed patterns in hot regions, while considering mutation rate bias as a potential alternative explanation for a minority of these regions (see lines 236-264 and 342-351 in the revised main text). In our forthcoming study, we plan to investigate different types of genomic variations in *Y. pestis* by using complete genome sequences, including frequent genome rearrangements, to provide a valuable dataset for the scientific community.

Response letter regarding the decision on Nature Communications manuscript # NCOMMS-24-22406A

Reviewer Comments

Reviewer #1 (Remarks to the Author):

The authors provide a thoughtful and thorough response to the review's comments. The additional analyses have strengthened this manuscript and the removal of the NLM model which did not have enough supporting data I think was the right call. The authors' discussion is also improved, and the appropriate place for exploration of possible hypotheses of a mechanism that requires further investigation. There is value in this type of big data analysis, and it is a nice contribution that highlights our current knowledge gaps bacterial evolution. I think this revised version is a nice paper and I congratulate the authors on their work.

Response: Thank you for recognizing our work.

Reviewer #3 (Remarks to the Author):

The authors have carefully revised their work in response to the reviewers' comments. They present one of the largest datasets of *Y. pestis* genomes, and analyze patterns of variation that can shed light on important open questions in bacterial evolution and genomics.

I have the following minor comments that I hope can further improve their paper:

Response: Thank you for your recognition and valuable feedback on our manuscript. Our detailed responses are provided below.

Q1: The first line of the abstract should be edited to reflect the revised focus of the paper. Here is one suggestion: "The relative contributions of mutation rate variation, selection, and recombination in shaping genomic variation in bacterial populations remains poorly understood".

Response: We have revised the abstract according to your suggestion (see Lines 20-21 in the revised main text).

Q2: Also on line 30: "natural selection" -> "positive selection" since negative/purifying selection can account for purging of variation.

Response: Done (see Line 30 in the revised main text).

Q3: Line 84: "genetically monomorphic" needs to be qualified that the authors focus their analysis on the *Y. pestis* core genome, relative to one focal strain. Should be "… genetically monomorphic nature of the *Y. pestis* core genome".

Response: Thank you for pointing this out. We have revised the phrasing in accordance with your suggestion (see Line 82 in the revised main text).

Q4: Some fundamental numbers need to be reported here. When considering the collection of 3,318 *Y. pestis* genomes, how many genes are in this core genome? How many genes in the pangenome? What is size of the biggest genome and number of genes, and size of the smallest genome, and its number of genes?

If these are not critical for main text, these details can go into the Methods. This can help the reader understand some basic facts about genome diversity in this collection of genomes, and get a sense for diversity outside the core genome analyzed here.

Response: We have added the range and median values for genome size and the number of annotated genes across the 3,318 *Y. pestis* genomes in our collection. Pan-genome analysis was conducted using both Panaroo and Roary, quantifying the total number of genes, as well as the core, soft-core, shell, and cloud genes. These details are now included in the Methods section (see Lines 391-403 in the revised main text).

Q5: Figure 1: the figure itself looks very nice, but the legends and labels are way too small. Legend for panel B in particular is very hard to read because it is so small. Perhaps some postprocessing to 'cut' the legend, and 'paste' it larger next to the tree could help.

Response: Thanks for your good suggestion. We have repositioned the legends and increased the font size of the labels to enhance clarity of Figure 1, particularly for panel B (see revised Figure 1).

Q6: Line 124: It is important to provide a formula for VHR in the Methods since this is a key piece of the authors' argument in this section. I understand the formula for P in the Methods section "HR mutation distribution in phylogroups". This formula could be simplified by summing over n rather than m, since the remaining n-m sites contribute 0 to the sum. As I understand, there is a vector of phylogroup proportions per HR, and the variance is calculated over this vector. This could be made clearer. If my understanding is correct, then line 124 should state "we calculated the dispersion of variation across phylogroups per HR" (not "within" as stated in the main text).

Response: We appreciate your valuable suggestion. The formula for V_{HR} has been added to the Methods section, where we clarify that V_{HR} represents the variance of phylogroup proportions (R) for phylogroups where $R > 0$ (i.e., those containing strains exhibiting variations within the HR). To avoid confusion with the P-value used in

statistical tests, we have replaced "*P*" with "*R*" in the formula. We have also simplified the formula for *R* by summing over *n*, as the use of *m* was redundant. The original text has been revised to read: "we quantified the dispersion of variation across phylogroups per HR by calculating…", as you correctly pointed out. See Lines 122, and 507-517 in the revised main text.

Q7: The labels in Figure 2B are not very professional, with the under_score characters and abbreviations like GenomeLen and TotalMuts, and this figure is hard to understand. I think it will probably be better to put these numbers into a table, or put these numbers directly into the main text and cut the figure. Or at least, improve the labels and annotate the total numbers for each bar, so that we can understand how the proportions correspond to particular numbers.

Response: Thank you for pointing this out. We have updated the labels in Figure 2b and included annotations for the total numbers of each bar, and the figure legend has been revised accordingly. Detailed numbers and percentages are now provided in the revised Supplementary Data 4, with corresponding references added to the main text. Please see Lines 108 and 831-835 in the revised manuscript.

Q8: Figure 4a: I have no idea what the blue bars and red line are supposed to mean, or why these numbers matter. Furthermore the line is confusing because there is no biological connection between the points connected to the lines. Each red point represents an independent number, and the apparent "spikes" in the distribution don't seem to be that meaningful for this paper.

These data should be more easily understood using a table with four columns, one for GO Term, one for observed gene count, one for Percentage of Observed Gene Count in Background, and one for FDR. But the blue bars and red line seem to be hard to understand and irrelevant for the point in the text. Why not just have a simple table with two columns, one for GO term, and one for FDR-corrected p-value for enrichment, and then use bold text to highly the GO terms related to regulation? If the blue bars and red line are more meaningful that I understand, then please interpret further in main text. But my understanding is that these are not that important, and even if so, may be better visualized in a table, rather than in a confusing graph with two different x-axes and plots superposed one on another.

Response: We acknowledge that the blue bars and red line are not central to the main findings of the paper. Our primary focus is on the significantly enriched GO terms (FDR < 0.05) identified for the 96 HR-related genes against the whole-genome background. The blue bars were originally used to represent the count of HR-related genes associated with each GO term (noting that each gene can correspond to multiple GO terms), while the red points indicated the ratio of HR-related genes relative to the whole-genome background for each GO term.

We agree that there is no biological connection between the red points, and the apparent "spikes" in the distribution are not meaningful in the context of this study. To avoid confusion, we have removed the red points and the connecting red line from Figure 4a. Additionally, we have revised the figure to highlight the enriched GO terms and their categorization into three main domains, with GO terms related to regulation presented in bold text.

While the count of HR-related genes in the corresponding GO terms is not the focus of our study, we believe it may provide useful context for readers to understand the distribution of genes across these terms. Therefore, we have retained this information but reduced its prominence in the updated Figure 4a. The figure legend has been revised accordingly (see Lines 850–854 in the revised manuscript).

Q9: In the methods, "Variation hotspot" should be preferred to "mutation hotspot".

Response: Done (see Line 463 in the revised main text).

Q10: Comment on copy number analysis in the rebuttal letter: copy number variations can be detected by examining read coverage in each of the genomes. Even if repetitive regions are excluded, this does not mean that copy number variations have been excluded in these genomes. From direct experience from sequencing evolved genomes in the lab, copy number variations can be very common, and arise faster than SNPs, although they can be very unstable.

For instance, the authors could examine sequencing read coverage in the HR regions, and ask how often the read coverage is 2X or even higher than the average read coverage across the core genome. By assuming a negative binomial distribution of reads (poisson is theoretical Lander-Waterman model for sequencing coverage, but in reality the variance is inflated due to experimental factors in library prep, machine etc), one can conduct tests for substantial higher read coverage in the HR regions compared to the mean coverage over the genome. If the authors have the read alignments across the core genomes on hand, and have the read alignments for the HR regions, then they just have to calculate the depth of these regions, the variance in the depth across the core genome, use the mean and variance for the whole core genome to parameterize the negative binomial, and then look at the probability of a HR region with elevated coverage, drawing from the negative binomial representing the genome-wide coverage distribution. Assuming that reads are say, 300bp long, one can break up an HR region into 500bp tiles, run independent tests, and if large regions (like 2000-10,000bp) have elevated coverage, those separate tests can be multiplied (because independent) to generate a p-value, and then that p-value can be Bonferroni-corrected or FDR corrected to get statistical significance for copy number variation. This is something I have done to search for CNVs in genomes during experimental evolution, there may be better approaches, but in my hands this idea is practical and straightforward.

That said, I do not think this CNV analysis is necessary for further revision. An analysis of CNVs from gDNA extracted in the lab may largely reflect growth selection under lab conditions rather than the record of historical selection and mutation processes in natural populations.

Response: Thank you very much for sharing your direct experience with copy number analysis in genomic studies. This insight will be invaluable for our future mutation accumulation experiments in *Y. pestis*.