

Brief antibiotic use drives human gut bacteria towards low-cost resistance

Corresponding Author: Dr Eitan Yaffe

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Summary

In this paper, the authors shotgun sequence 60 patients, each 16 sampling timepoints at a mean num reads of 18.8M/sample, before and after consumption of ciprofloxacin to identify evolutionary adaptations in response to the antibiotic. The major points of novelty for the paper are (1) significant sequencing effort, (2) introduction of novel computational approach for disentangling evolutionary changes from other sources of SNV changes including paralog and ortholog read stealing, and (3) a biological story showing repeated parallel evolution at the *gyrA* locus, involving both de novo and standing variation, as well as few reversions.

Positives

The paper explores the evolutionary response to a very commonly used antibiotic in a natural gut environment in a massive cohort of individuals. With this controlled dataset, the authors are able to pull out trends regarding the way that evolution proceeds in response to strong selective pressures. In particular, the authors are able to demonstrate that parallelism is a common feature of the gut microbiome, and that there are specific residues more prone to evolution. They also are able to show that reversions after antibiotics are rare. Overall, the systematic characterization of common trends in a diverse cohort receiving the same exact perturbation is a valuable contribution to the literature, as the rules and principles of evolution in the gut microbiome are still emergent.

In addition, this study takes a step forward into predicting when antibiotic resistance will occur for a plethora of species in the gut microbiome. These variables, such as species identity and initial abundance, are easy to measure, have clear causative interpretations, and effectively predict which species and resistance alleles will rebound. This sets the groundwork for a future of personalized medicine, where individual patient microbiomes' profiles can help inform clinical interventions.

Major Critiques

There are many axes on which the paper pushes novelty. I think that the data collection and biological findings stand for themselves. This review largely probes the novelty and robustness of PolyPlanner, the pipeline that the authors propose for detecting evolutionary changes from their temporal data, as this was a major point of innovation that the authors made a case for.

First, the authors put a large emphasis on distinguishing noise from true polymorphism. In line 142, the authors write: "To our knowledge, this approach is the first explicit modelling and classification of variant types and allows one to enrich for true polymorphic variants using longitudinal shotgun data."

We probed deeply into where the 'first' comes from, as there are in fact other approaches that do leverage temporal shotgun data that achieve similar results. In particular, Roodgar, Good et al. Genome Research explicitly models multiple time points (>15) to recover sweeping SNVs, whilst controlling for read stealing across putative orthologs and paralogs, and Garud, Good et al. 2019 PLoS Biology also does the same with 2 time points. The latter explicitly uses shotgun data, and while the former uses readcloud technology, Wolff et al. mBio later on use a similar approach on shotgun data representing >100 time

points. Some acknowledgement in the paper as to how explicitly this approach differs from the above ones would be very helpful in understanding the novelty of PolyPlanner.

To recapitulate the pipeline of PolyPlanner, after reading the supplement, this is what I have surmised, and embedded are some requests for clarification:

1. Identify populations.

A 'population' here is a MAG inferred using many time points. It seems that a population is a host x species combination. It would help if the authors could state that there are potentially numerous strains within each population to foreshadow what is yet to come.

2. Identify dynamic variants

Remove sequencing errors using the Quince et al. methodology

Then separate dynamic variants from paralog and ortholog variants. This is on the basis of the particular variant alt and ref having similar coverage as the local region. Adding a line in the results section saying "we apply a chisq test to..." would be very helpful in orienting the reader as to what the statistical machinery under the hood might be. This statistical test seems to me the major innovation of PolyPlanner as it's a new way of controlling for read stealing that does not exist in the literature.

3. Remove populations 1000+ dynamic variants from further analysis.

The value of 1000 doesn't seem to be justified or explained anywhere. I gather that this number is chosen to distinguish SNV changes that are blatantly due to strain fluctuations as opposed to evolution. One thing that is concerning is the fact that there seem to be a number of populations that do have 200-1000 dynamic SNVs in Figure S9 – these likely are experiencing strain fluctuations. I suppose point 4 below helps address this, but it does seem like the 1000 cutoff is arbitrary.

4. Apply Strain finder to the remaining populations.

Here, I gather that the definition of a strain is less important than identifying 'linkage groups', or, in other words, groups of SNVs inferred to move together on the same genomic background. Within each linkage group, a number of dynamic variants are found, as summarized in Figure S9. It is worth noting that in the supplement of Roodgar, Good et al., the veracity of strain finder is stress tested and it is made clear that many of the 'strains' inferred are noise. In fact, it's highly unlikely that there are 8 'strains' for a given species for a given sample – I have not seen this high of a number anywhere in the literature, but instead I take it that once again the authors here are just using Strain finder as a tool to recover linkage groups. Note, I put the word 'strain' in quotation marks because it seems irrelevant how divergent the strains are – the authors seem to really just care about linkage groups, once again.

Should we believe that the linkage groups that are recovered by strain finder are real, in Fig S9 there are several linkage groups that consist of very few dynamic SNVs – in fact, some consist of 1 dynamic SNV. Do the authors really believe that the 1 SNV that was inferred on the single strain was actually an evolutionary change? To judge if this is real, the authors could plot the actual SNV trajectory to verify that this is real and not some weird artifact, like in the SNV trajectory figures made in Roodgar, Good et al.

In addition, while the authors choose to focus on *gyrA* as a major variant contributing to resistance, the omission of strain dynamics with more than 1000 SNVs and the abundant signal of sweeping SNVs on large linkage groups (in Fig S9) does suggest that strain dynamics and/or recombination may be a significant driver of resistance as well. How important is it to quantify how these gut communities actually adapt, whether due to de novo *gyrA* mutation or to ecological responses (strain dynamics) and/or recombination (through standing genetic variation)?

To summarize:

As far as we can tell, the major novelty rests in how the authors recover dynamic variants. Others (eg. mentioned above), have dealt with read stealing, but PolyPlanner provides a statistical way of dealing with it that is novel. Now, whether controlling for read-stealing and donating is a Nature-worthy contribution is a question the editor must decide on. We do suggest that perhaps quantifying how much of a problem read stealing might be for inferring evolution could potentially push the case for PolyPlanner.

Minor Critiques

Highlight gyrase on plot Figure 2F

PetIn Fig 2C, the authors find a higher dN/dS ratio for smaller linkage group. Without really interpreting it, they seem to implicitly suggest that positive selection must be stronger within these 'younger' strain divergences. However, this pattern is not unique to this analysis, and has been shown in other papers (e.g. Garud Good et al 2019, Zhao and Lieberman et al. 2019); and is also consistent with other explanations, including increasing strength of purifying selection with strain divergence (e.g. due to HGT). The fact that these values are all $\ll 1$ and quite close to one another is also not super convincing.

While the number of sweeping variants are highlighted, it's not really clear how many independent sweep events occurred among populations. From Fig S9, this number could be as few as 185. It would be good to know as a better baseline for how parallel and convergent *gyrA* and other mutations actually are.

For the logistic regression analysis, it seems like you are actually regressing on a transformation of the same variable twice:

$\text{link-1}(\text{evolvability}_i) = \beta_1 * \text{base_abundance}_i + \beta_2 * (\text{base_abundance}_i / \text{treatment_abundance}_i) + \dots + \epsilon$ for population i

While base_abundance_i and $(\text{base_abundance}_i / \text{treatment_abundance}_i)$ may not be perfectly correlated with the addition of treatment abundance, it seems like there may be issues (see the increase in standard error for base abundance with the inclusion of 'decline'). It seems like, qualitatively, 'decline' is supposed to be measuring susceptibility to the antibiotic – perhaps a better (and more interpretable) variable would be whether the population contains a resistance variant, which the authors have already calculated for many species.

For Fig 1A, could the authors include labels for the cluster size for the broad categories of trajectories (e.g. Drop 1)? The color scheme is a bit hard to interpret for so many near 0, and it would be useful to know, e.g. how many total populations go extinct? How many drop? Etc.

Figure S6 seems actually quite important as this describes the main innovation – Make the circles bigger! It is hard to distinguish white from black.

In Fig S7, why does spurious variant density look so weird for 10 genuine mutations in high-coverage simulations?

Are the authors normalizing by genome length when measuring genome abundance? (Line 272 of Supplementary)

Where are the 20->80% thresholds coming from for dynamic variants?

Linkage group: I think a better cartoon is needed to describe the concept of a linkage group than what is shown in Figure 2. Moreover, I think further explanation is needed to describe the significance of small vs large linkage groups (and where they come from)

In Fig S7, the sensitivity for variants seems fairly low for coverage even from 50x to 100x. However, the median depth for all populations is only 50x. The main text seems to remedy this by only calling dynamic variants in MAGs with high abundance, but it would be useful to show the coverage and dynamic variant detected histogram. How many populations are lost with a 100x coverage cutoff? Does the graph for # variants detected vs coverage appear similar to S7? Are populations with 1-1000 dynamic variants (the rest of the analysis) correlated with lower coverage? How would this bias the rest of the results?

Referee #2

(Remarks to the Author)

This study describes evolution of resistance to ciprofloxacin in the microbiota of human volunteers during and after a 5-d treatment. It is fascinating to see sweeps of mutations in diverse microbial populations in humans. The results align well with previous knowledge about resistance to fluoroquinolone antibiotics including specific target mutations, and the dependency on population size and selective pressure. A main novel insight is the different evolvability of diverse taxa within the human microbiota in situ.

I have a few comments:

1. Clinical relevance

A single *gyrA* mutation as found in this study does not usually confer CLSI-classified resistance (this requires at least another mutation in *parC* or upregulation of efflux). The data reflect an evolution towards reduced susceptibility, not "persistent resistance" as stated in the title and in the text.

The mutations were detected only in abundant commensals but not in less abundant bacteria that could become pathogenic such as *E. coli* (line 262). Horizontal gene transfer of gyrase alleles is rare, limiting the spread of mutations from commensals to pathogens. Moreover, clinical data showing relevant resistance development during treatment of bacterial diarrhea seem to be rare (despite initial bacterial population sizes that are likely much higher than the estimates provided in Suppl. Note 1).

The clinical relevance of the evolutionary changes observed in this study may be limited.

2. Role of other mutations

Increased drug efflux can be a major component of resistance to fluoroquinolones, particularly at early steps of resistance evolution. However, efflux-associated mutations were rarely found in this study (only 4 variants in KO2004). Could this reflect a larger mutational space for upregulating efflux, which would make detecting the underlying mutations difficult compared to mutational hotspots such as the quinolone resistance-determining region on *GyrA*? This could also affect other *gyrA*-independent resistance mechanisms that are found in many Gram-positive and Gram-negative fluoroquinolone-resistant strains. Thus, the data might underestimate the level of resistance evolution.

3. Susceptibility to ciprofloxacin in the gut

Figs. 1A, S4, S5 show diverse responses of microbiota populations during and after ciprofloxacin treatment. This is great data and it would be very interesting to relate these results to reported data on fluoroquinolone susceptibility of the corresponding bacterial taxa under in-vitro conditions (e.g. PubMedID 34646011). This could yield insights into potential

differences between drug action in laboratory cultures vs. in the human gut (a large number of populations showed surprisingly weak responses to treatment).

In addition, the comparison of responses during treatment and the corresponding *gyrA* genotypes could provide actual data for the assumed impact of various *gyrA* genotypes on susceptibility of microbiota in the human gut.

Minor points:

Fig. 3c, title of y-axis should be rather "number of genomes"?

Lines 237-239, Fig. 3D: The calculation underlying the 10% probability of substitution from *gyrA* S83 towards resistance remains unclear. *gyrA* sweeps (including other mutations than S83) occurred in 63 populations (lines 207-208) out of 1866 analyzed populations (line 1866), ~60% of which had S83 prior to exposure? This would give a lower probability. Yet another calculation is given in line 267-271.

Referee #3

(Remarks to the Author)

Yaffe and colleagues present a very interesting study in 60 human subjects exposed to a short course of oral ciprofloxacin treatment. The main aim was to investigate the effect of the quinolone exposure on the gut microbiota – specifically on selection and maintenance of resistance traits.

Study participants were exposed to 2x500mg ciprofloxacin orally for 5 days and self-collected (a total of 16) stool samples 9 weeks prior and for each of the 2 consecutive days immediately prior to the start of ciprofloxacin, daily during antibiotic exposure (Days 0-4) and for the following four days (Days 5-8), and then on Days 10, 18, 28, and 77. Subjects had not taken any antibiotics for at least 6 months prior to the start of sampling.

Metagenome-assembled genomes were generated in 960 collected samples sequenced with high read per sample sequencing (18.8M). This was done by a novel tool named PolyPanner that compares individual time-point samples to a subject-specific metagenome generated from all samples from a given individual to identify fluctuations in single-nucleotide profiles (which then identifies polymorphic variants). 5655 populations spanning 924 species were analyzed in this study.

Key findings were:

- 1) Only few extinction events overall and return to baseline suggest a high resilience of the gut microbiota to such a treatment
- 2) Dynamic variants were mainly SNPs and found in 38% of populations analyzed. 1866 were further processed to identify putative selective sweeps evolving longitudinally (linkage-group based)
- 3) 7132 variants swept from pre-antibiotic to antibiotic "era" (frequency increase) in 513 populations (212 species)
- 4) *gyrA* mutations were concentrated on positions 83 and 87 (56 and 9 substitutions, respectively). These mutations are known to be involved in development of FQN resistance and *gyrA*:83 mutations occurred in 33 subjects in 38 species. Longitudinal analysis (comparison to baseline) suggested a transition to AA associated with resistance in approx. 10% of susceptible populations. Furthermore, there was evidence for pre-existing resistance.
- 5) Interestingly, *E. coli* - a key taxon – had an estimated 41% resistance rate at baseline without evidence for sweep in these populations during the study.
- 6) *gyrA* evolvability was not the only sweep driver and accounted for only 36% of overall evolvability. Evolvability was associated with FQN susceptibility as determined by abundance reduction (being high in Bacteroidetes and Actinobacteria and low in Firmicutes).
- 7) After removal of exposure there was only a low reversal of sweep variants in populations (frequency above 80% for most *gyrA* sweeps).

Overall, this is a very interesting pilot trial including the bioinformatics methods, which seems to be the main novelty (that I cannot judge to the code level). Assembling genomes from metagenomics data in such a rich environment is really challenging. PolyPanner seems to be a significant step forward in this matter.

The choice of ciprofloxacin is interesting as resistance "development" under (mono)treatment clinically occurs in several species (e.g., *S. aureus*) by mutations in the target genes (gyrase/topoisomerase).

Indeed, the authors demonstrate a sweep in *gyrA* mutations (exposure of ciprofloxacin resulted in a sweep of mutations in *gyrA* in ~10% of investigated microbial populations) in their work indicating that the use of FQN can lead to selection of such mutants.

However, the work has several limitations mainly being in the study design with no control group and additionally - for this journal - a phenotypic confirmation of key aspects such as MIC determination, confirmation of mutations in cultures populations, and assessments of fitness traits should be performed. Furthermore, the clinical relevance of the accumulation of those mutations remains unclear.

Specific strengths are the sophisticated novel bioinformatic analyses and the human trial, whereas specific weaknesses are the lack of a control group w/o antibiotic, the lack of phenotypic confirmation of resistant phenotype and associated predicted low fitness cost in key taxa. Although the discovery of *gyrA* mutant sweeps is interesting, the clinical consequence of those is unclear as invasive infections with most of those strains is rare and evidence for horizontal transfer was low.

Specifically:

250 Given the long-lasting effect demonstrated here; is there any information of whether or not these patients were exposed

to FQN before the 6 months of antibiotic-free interval they were supposed to have to be enrolled?

256 A control group would be useful to gather more evidence for this.

318 This should be confirmed by phenotypic fitness assays for some key taxa

344 This is a bit a stretch – FQN should not be a first line (empiric) treatment of UTI anyways and if they are, there is probably a clinical reason (e.g., *P. aeruginosa* where FQN are the only oral option). Assuming there is time to screen for abundance of gut-residing pathogens then there is also time to culture urine and initiate a targeted antibiotic treatment. This should therefore be rephrased, as antibiotic stewardship should already be good enough to prevent misuse of FQN for this indication.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

I thank the authors for doing a great job in addressing our comments! The work represents a really thorough investigation into the evolutionary dynamics of gut microbiomes when exposed to antibiotics.

I have a few minor comments that I feel if addressed would improve the paper a bit.

- I like the added note on the supplement about the difference between the coverage approach of Garud et al. and their more statistical approach for removing ortholog variants. However, I think I just don't have a good grasp on how much of a problem ortholog variants present. In their supplement, figure S10, they show that only 3.5% of the variants they detect get classified as "dynamic variants" or true variants. Given that they are assembling using MAGs, is having over 85% of the genomic variants detected in the data due to paralogs/orthologs a expected result? (It seems very high to me, but I do not have any data to compare it to).
- I do agree however that by using such a specific statistical test, they are putting a much higher requirement on what variants to consider as "true polymorphisms" (Or "dynamic variants") and therefore, the results they do find are a lot more reliable (in terms of me believing that the variants they found are due to true polymorphisms instead of potential ortholog or other variants).
- I still do not understand where the 1000 dynamic strain cutoff comes from (as a number) but the inclusion of 2C is nice to show that the number of genomes thrown out by this threshold is not a majority. I think explicitly citing a paper (either Lieberman/Zhao or Garud/Good and/or showing that the number of between host SNVs is $\sim 10^4$ would be useful for justifying this).

Git-Hub Comments:

- Github is generally well commented. I have a couple of small notes:
- It seems like some utility scripts are required to run the workflow before you even start using the PolyPlanner commands (PolyPlanner is a command-line tool). In particular, things like `utils/parse_bwa_sam.pl` or `utils/bin_summary.r`. I think it might be useful to include the expected version of perl or r required for those scripts?
- One step of their workflow is "paired reads are transformed into POP files". Is this a standard bioinformatic procedure (I have never done it and would not know how to immediately)? I wonder if including the specific command to perform that action would be useful for the pipeline to be easily accessed by others. Reading a little more into the github, I think it is the command `polypanner construct`. Might just be worth mentioning it directly in the workflow explanation.

Additional comments:

1. Fig S26 has an 'A'
2. What is x-coverage profile?
3. What do the different colors for the coverage plots in S10 and S26 mean?
4. Do you think that your method is underpowered to detect soft sweeps, or, is it underpowered to detect very soft sweeps or soft sweeps that do not meet the <0.2 to >0.8 criteria (e.g. a partial soft sweep)? Could it be that there are more soft sweeps in the data and currently we are just looking at extreme allele frequency changes that could be attributed to a hard or a soft sweep?

Referee #2

(Remarks to the Author)

The authors have addressed appropriately all my comments and questions. They might want to mention explicitly also the many species that survived rather well the treatment without detectable sweeps, in the paragraph discussing the survival of *gyrA* mutants (lines 429-437). These findings actually suggest that the antibiotic treatment was actually quite inefficient in the gut despite the rather high concentrations and the previously reported susceptibility of many of these species to fluoroquinolones.

Referee #3

(Remarks to the Author)

In the resubmitted version, Yaffe and colleagues have significantly revised the manuscript:

Key changes are:

- revised introduction for pointing out possible clinical consequences of sweeps
- a control group (4 subjects) was added with 16 sampling points "matching" the time points of the initial intervention group; the control group had a lower rate of sweep events; it might be worth adding more information on the sweep events in the control group - at least whether or not genes of interest (e.g., *gyrA* but also compared to the evolving populations without *gyrA* mutations) were involved
- rephrasing of resistance to evolution towards resistance (also now reflected in the title) and added novel analyses on populations that had selected sweeps in *gyrA*::83 (4 phyla with various substitutions) using PCR to analyze fitness effects. By doing this and addressing other points raised (or rephrasing / omitting) I feel that the comments have been adequately addressed. For an overall summary and detailed judgment I take the liberty to refer to my initial summary. I still believe that this is a very interesting and well done study that has gained strength by adding a (small) control group and performing additional fitness data. By downphrasing to "towards resistance" I also do not feel the need for extended phenotypic confirmations given the available literature.

(Remarks on code availability)

As stated in the initial review, my knowledge of programming languages is not sufficient to make a judgment on the code level.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

The authors have adequately addressed all of my comments.

(Remarks on code availability)

I reviewed the code briefly now, but in greater detail in my previous review. I believe the authors have taken steps to make the code accessible.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Review comments are provided verbatim *in italics* and our responses are [in blue](#). To facilitate the review process, we created a version of the manuscript with tracked changes. We also generated a clean version of the manuscript without these tracked changes.

Response to Reviewer #1

Summary

In this paper, the authors shotgun sequence 60 patients, each 16 sampling timepoints at a mean num reads of 18.8M/sample, before and after consumption of ciprofloxacin to identify evolutionary adaptations in response to the antibiotic. The major points of novelty for the paper are (1) significant sequencing effort, (2) introduction of novel computational approach for disentangling evolutionary changes from other sources of SNV changes including paralog and ortholog read stealing, and (3) a biological story showing repeated parallel evolution at the gyrA locus, involving both de novo and standing variation, as well as few reversions.

Positives

The paper explores the evolutionary response to a very commonly used antibiotic in a natural gut environment in a massive cohort of individuals. With this controlled dataset, the authors are able to pull out trends regarding the way that evolution proceeds in response to strong selective pressures. In particular, the authors are able to demonstrate that parallelism is a common feature of the gut microbiome, and that there are specific residues more prone to evolution. They also are able to show that reversions after antibiotics are rare. Overall, the systematic characterization of common trends in a diverse cohort receiving the same exact perturbation is a valuable contribution to the literature, as the rules and principles of evolution in the gut microbiome are still emergent.

In addition, this study takes a step forward into predicting when antibiotic resistance will occur for a plethora of species in the gut microbiome. These variables, such as species identity and initial abundance, are easy to measure, have clear causative interpretations, and effectively predict which species and resistance alleles will rebound. This sets the groundwork for a future of personalized medicine, where individual patient microbiomes' profiles can help inform clinical interventions.

[Thank you for these comments and for this constructive review. In the revised manuscript we have addressed to the best of our ability all concerns and questions. In brief, the manuscript now discusses how our approach compares with previous approaches and demonstrates the contribution of PolyPanner to the study of evolution in natural communities \(see detailed responses below\).](#)

[Also, as a natural extension of the work initially reported, we have added experimental data that expands on our finding regarding the low fitness cost of drug-resistance gyrA alleles \(Figure 6, lines 357-383\).](#)

[We also added experimental data and conducted new analysis in response to comments by the other reviewers:](#)

- [\(1\) We added 4 control subjects that did not take antibiotics \(reviewer #3\).](#)

- (2) We added an analysis that shows how the presence of resistance *gyrA* alleles prior to exposure is correlated with elevated abundance during antibiotics (reviewer #2).

Major Critiques

There are many axes on which the paper pushes novelty. I think that the data collection and biological findings stand for themselves. This review largely probes the novelty and robustness of PolyPlanner, the pipeline that the authors propose for detecting evolutionary changes from their temporal data, as this was a major point of innovation that the authors made a case for.

First, the authors put a large emphasis on distinguishing noise from true polymorphism. In line 142, the authors write: “To our knowledge, this approach is the first explicit modelling and classification of variant types and allows one to enrich for true polymorphic variants using longitudinal shotgun data.”

We probed deeply into where the ‘first’ comes from, as there are in fact other approaches that do leverage temporal shotgun data that achieve similar results. In particular, Roodgar, Good et al. Genome Research explicitly models multiple time points (>15) to recover sweeping SNVs, whilst controlling for read stealing across putative orthologs and paralogs, and Garud, Good et al. 2019 PloS Biology also does the same with 2 time points. The latter explicitly uses shotgun data, and while the former uses readcloud technology, Wolff et al. mBio later on use a similar approach on shotgun data representing >100 time points. Some acknowledgement in the paper as to how explicitly this approach differs from the above ones would be very helpful in understanding the novelty of PolyPlanner.

R1. We appreciate this comment. Accordingly, we have revised the Introduction to state more clearly how our approach compares to previous ones (lines 72-91). The two recent papers (Wolff et al. and Roodgar et al.) analyze temporal data but begin by identifying polymorphic sites on a per-sample basis using the method developed by Garud et al. (2019). In brief, this method evaluates samples individually, limiting its precision. In **Supplementary Note 1**, we explain how the use of longitudinal samples to identify polymorphic sites allows us to go beyond detecting sweeps and infer sites with more subtle temporal frequency changes.

To recapitulate the pipeline of PolyPlanner, after reading the supplement, this is what I have surmised, and embedded are some requests for clarification:

1. Identify populations.

A ‘population’ here is a MAG inferred using many time points. It seems that a population is a host x species combination. It would help if the authors could state that there are potentially numerous strains within each population to foreshadow what is yet to come.

R2. You are correct. We now state, “Each genome represented the population of a single bacterial species inhabiting the gut of a specific subject and can be seen as a consensus genome of one or more colonizing strains” (line 126).

2. Identify dynamic variants.

Remove sequencing errors using the Quince et al. methodology. Then separate dynamic variants from paralog and ortholog variants. This is on the basis of the particular variant alt and ref having similar coverage as the local region. Adding a line in the results section saying “we apply a chisq test to...” would be very helpful in orienting the reader as to what the statistical machinery under the hood might be. This statistical test seems to me the major innovation of PolyPlanner as it’s a new way of controlling for read stealing that does not exist in the literature.

R3. Thank you. We have revised the text to explain explicitly how PolyPlanner works using the Chi-Square test (line 155).

3. Remove populations 1000+ dynamic variants from further analysis.

The value of 1000 doesn’t seem to be justified or explained anywhere. I gather that this number is chosen to distinguish SNV changes that are blatantly due to strain fluctuations as opposed to evolution. One thing that is concerning is the fact that there seem to be a number of populations that do have 200-1000 dynamic SNVs in Figure S9 – these likely are experiencing strain fluctuations. I suppose point 4 below helps address this, but it does seem like the 1000 cutoff is arbitrary.

R4. As you correctly inferred, we omitted populations with >1000 variants since our focus was recent evolution. For clarity, we have revised the main text (line 197) and methods (line 274). We chose the value 1000 to ensure we include all populations that have a simple population structure (i.e., one strain colonization event), while allowing some populations with intermediate levels of variants to be included. The only result that could be affected by this choice is the convergent evolution analysis; there, our weighting scheme was designed to downplay variants in large populations so that analysis would be robust to this choice.

Per your suggestion, we added an analysis that compares evolutionary and ecological changes prior to dropping populations with >1000 variants (**Fig. 2B**).

4. Apply Strain finder to the remaining populations.

Here, I gather that the definition of a strain is less important than identifying ‘linkage groups’, or, in other words, groups of SNVs inferred to move together on the same genomic background. Within each linkage group, a number of dynamic variants are found, as summarized in Figure S9. It is worth noting that in the supplement of Roodgar, Good et al., the veracity of strain finder is stress tested and it is made clear that many of the ‘strains’ inferred are noise. In fact, it’s highly unlikely that there are 8 ‘strains’ for a given species for a given sample – I have not seen this high of a number anywhere in the literature, but instead I take it that once again the authors here are just using Strain finder as a tool to recover linkage groups. Note, I put the word ‘strain’ in quotation marks because it seems irrelevant how divergent the strains are – the authors seem to really just care about linkage groups, once again.

R5. Thank you for pointing this out. Yes, we infer strains for two purposes: (1) to weight dynamic variants appropriately in the convergent evolution analysis and (2) for visualization purposes. We have revised the text to state this explicitly (line 200). We also added a figure that shows that the convergent evolution analysis is robust to the choice of the weighting scheme (**Supp. Figure S18**, line 256).

Should we believe that the linkage groups that are recovered by strain finder are real, in Fig S9 there are several linkage groups that consist of very few dynamic SNVs – in fact, some consist of 1 dynamic SNV. Do the authors really believe that the 1 SNV that was inferred on the single strain was actually an evolutionary

change? To judge if this is real, the authors could plot the actual SNV trajectory to verify that this is real and not some weird artifact, like in the SNV trajectory figures made in Roodgar, Good et al.

R6. While some linkage groups might be a result of unknown artifacts, we believe that some *gyrA* variants are sweeping with a few or no passenger variants. We added plots as in Roodgar et al., including **Figure 2H** (one example of a *gyrA* sweep), **Supp. Figure S15** (more examples of *gyrA* sweeps), and **Supp. Figure S17** (examples of non-*gyrA* sweeps).

*In addition, while the authors choose to focus on *gyrA* as a major variant contributing to resistance, the omission of strain dynamics with more than 1000 SNVs and the abundant signal of sweeping SNVs on large linkage groups (in Fig S9) does suggest that strain dynamics and/or recombination may be a significant driver of resistance as well. How important is it to quantify how these gut communities actually adapt, whether due to de novo *gyrA* mutation or to ecological responses (strain dynamics) and/or recombination (through standing genetic variation)?*

R7. Following the reviewer's suggestion, we added a histogram of the number of populations stratified by the number of sweeping variants (**Fig. 2C**, line 195). This analysis supports the claim that recombination and ecological changes are contributing to resistance alongside evolutionary changes, but to a lesser degree.

To summarize:

As far as we can tell, the major novelty rests in how the authors recover dynamic variants. Others (e.g. mentioned above), have dealt with read stealing, but PolyPlanner provides a statistical way of dealing with it that is novel. Now, whether controlling for read-stealing and donating is a Nature-worthy contribution is a question the editor must decide on. We do suggest that perhaps quantifying how much of a problem read stealing might be for inferring evolution could potentially push the case for PolyPlanner.

R8. Indeed, the major technical novelty is in the way PolyPlanner deals with read-stealing (what we call spurious sites). In addition to providing a clear statistical test, the usage of longitudinal data allows PolyPlanner to identify polymorphic sites that have subtle temporal changes in their frequency.

To strengthen the case for PolyPlanner, we added a Figure that demonstrates the importance of filtering spurious sites (**Supp. Figure S10**, line 184). We also added an unsupervised clustering of dynamic variants that showcases the ability to uncover complex trends in frequency dynamics (**Figure 2B**, line 189). While the bulk of our analysis is focused on sweeping variants (14% of all dynamic variants), which carry the strongest signal, PolyPlanner enables the investigation of polymorphic sites with subtler changes in frequency.

The precision of PolyPlanner allowed us to reach novel biological findings. For example, we quantified for the first-time the probability of populations having sweeps of resistance mutations at *gyrA*:83. Our systematic approach was vital here, since this assessment required identifying populations for which sweeps could be identified if present or if not present, based on population coverage and the statistical tests of PolyPlanner. The precision of the tool was also central in the convergent evolution analysis, where vetted sweeps allowed us to pinpoint genes (other than *gyrA*) that were under selective pressure, such as transcription factors and efflux pumps.

Minor Critiques

Highlight gyrase on plot Figure 2F

R9. We revised the figure to highlight *gyrA* with a black circle.

In Fig 2C, the authors find a higher dN/dS ratio for smaller linkage group. Without really interpreting it, they seem to implicitly suggest that positive selection must be stronger within these ‘younger’ strain divergences. However, this pattern is not unique to this analysis, and has been shown in other papers (e.g. Garud Good et al 2019, Zhao and Lieberman et al. 2019); and is also consistent with other explanations, including increasing strength of purifying selection with strain divergence (e.g. due to HGT). The fact that these values are all $\ll 1$ and quite close to one another is also not super convincing.

R10. Indeed, this pattern is not unique to our analysis. We have revised figure 2C (Now **2E**, line 208) to stratify variants by their class (sweeping vs. non-sweeping), which shows that dN/dS reaches 1 for sweeping variants (both for *gyrA* variants and for non-*gyrA* variants). We now clearly state our interpretation and cite the relevant papers to refer readers to these key papers that expand on the population genetics theory of adaptive evolution (line 209).

*While the number of sweeping variants are highlighted, it’s not really clear how many independent sweep events occurred among populations. From Fig S9, this number could be as few as 185. It would be good to know as a better baseline for how parallel and convergent *gyrA* and other mutations actually are.*

R11. This is a helpful suggestion: we have added **Figure 2F** that shows the number of independent sweeps per genome. There were 629 independent sweep events (line 215).

For the logistic regression analysis, it seems like you are actually regressing on a transformation of the same variable twice:

$$\text{link-1}(\text{evolvability}_i) = \beta_1 * \text{base_abundance}_i + \beta_2 * (\text{base_abundance}_i / \text{treatment_abundance}_i) + \dots + \epsilon \text{ for population } i$$

While base_abundance_i and $(\text{base_abundance}_i / \text{treatment_abundance}_i)$ may not be perfectly correlated with the addition of treatment abundance, it seems like there may be issues (see the increase in standard error for base abundance with the inclusion of ‘decline’). It seems like, qualitatively, ‘decline’ is supposed to be measuring susceptibility to the antibiotic – perhaps a better (and more interpretable) variable would be whether the population contains a resistance variant, which the authors have already calculated for many species.

R12. We checked for co-linearity and found that the baseline and decline variables are correlated with a Pearson’s coefficient of 0.46. Given this moderate level of correlation, the models that include both variables are of interest, and we leave it for readers to choose their preferred model. We added a line in methods that alludes to this potential issue (line 366 in Methods).

Regarding the suggestion to use resistance at baseline as a predictor for evolution, we note that our analysis is focused on naïve populations that had serine at *gyrA*:83 at baseline. For these populations it would be impossible to use *gyrA*:83 at baseline as a predictor as it is serine by definition. Following the suggestion made by the reviewer, we made an attempt to broaden the scope to populations with any amino acid at *gyrA*:83. We added a matching Boolean predictor variable (“Resist”, true if *gyrA*:83 was

phenylalanine or leucine), and inferred 3 additional models: Resist, Resist+Base, Resist+Base+Phylum. It turned out that the Resist variable was not significant in any of these models and therefore we did not include the Resist variable in the revision.

As a side note, while reviewing the logistic regression analysis we discovered that we inadvertently omitted sweeping variants that had large LGs from this analysis. In the revision we corrected this issue and regenerated **Figure 4** and **Supp. Figures S23** (which had minor changes), and **Supp. Figures S24** (for which fewer models were found significant).

For Fig 1A, could the authors include labels for the cluster size for the broad categories of trajectories (e.g. Drop 1)? The color scheme is a bit hard to interpret for so many near 0, and it would be useful to know, e.g. how many total populations go extinct? How many drop? Etc.

R13. As suggested, we added the number of populations per cluster group to **Figure 1A**.

Figure S6 seems actually quite important as this describes the main innovation – Make the circles bigger! It is hard to distinguish white from black.

R14. Thanks for pointing this out. We have redone **Figure S6** and hope the new version is clear.

In Fig S7, why does spurious variant density look so weird for 10 genuine mutations in high-coverage simulations?

R15. The “bump” in the plot was due to an outlier genome with an unusual number of spurious variants. We switched from **mean** densities to **median** densities. In the process we found a minor plotting bug, which we have corrected while revising **Supp. Figs. S8** and **S9**.

Are the authors normalizing by genome length when measuring genome abundance? (Line 272 of Supplementary)

R16. Yes, we do normalize by genome length when measuring abundance and revised the Methods accordingly (line 292 in Methods). Thank you for catching this missing technical detail.

Where are the 20->80% thresholds coming from for dynamic variants?

R17. The thresholds were designed to capture the center of the dynamic range and mirror previous work by others (Garud et al. 2019). We have revised the text to explain this and cite this paper (line 191).

Linkage group: I think a better cartoon is needed to describe the concept of a linkage group than what is shown in Figure 2. Moreover, I think further explanation is needed to describe the significance of small vs large linkage groups (and where they come from).

R18. We have revised **Figure 2H** to present in more detail how linkage groups were defined. We also edited the text to better explain our interpretation of small linkage groups as an indicator of adaptive evolution within subjects (lines 205-210).

In Fig S7, the sensitivity for variants seems fairly low for coverage even from 50x to 100x. However, the median depth for all populations is only 50x. The main text seems to remedy this by only calling dynamic variants in MAGs with high abundance, but it would be useful to show the coverage and dynamic variant

detected histogram. How many populations are lost with a 100x coverage cutoff? Does the graph for # variants detected vs coverage appear similar to S7? Are populations with 1-1000 dynamic variants (the rest of the analysis) correlated with lower coverage? How would this bias the rest of the results?

R19. We now show x-coverage instead of abundance in the revised **Fig. 2A**. The transition to higher sensitivity around 100x coverage that emerged in simulated data is recapitulated in the real data.

Following your suggestion, we calculated the x-coverage and found that populations with 0, 1-1000, >1000 variants had a median coverage of 31x, 126X, 287x (respectively). This does not come as a surprise – as coverage increases, more polymorphic variants are expected to have sufficient statistical power to be classified as dynamic, even if we assume that the underlying density of polymorphic sites is the same.

However, we believe that the convergent evolution analysis is robust to the omission of populations with >1000 variants, as practically all variants in those populations are downweighted due to their large linkage groups. We believe other results, such as the slow sweep recovery and the modeling of *gyrA* evolvability, are also robust to this cutoff. We have added a sentence to the discussion to explain how population abundance and frequency dynamics affect the detection of polymorphic variants (line 423).

Referee #2 (Remarks to the Author):

This study describes evolution of resistance to ciprofloxacin in the microbiota of human volunteers during and after a 5-d treatment. It is fascinating to see sweeps of mutations in diverse microbial populations in humans. The results align well with previous knowledge about resistance to fluoroquinolone antibiotics including specific target mutations, and the dependency on population size and selective pressure. A main novel insight is the different evolvability of diverse taxa within the human microbiota in situ.

Thanks for these comments and this constructive and thoughtful review. In the revision we have attempted to address all concerns.

I have a few comments:

1. Clinical relevance

*A single *gyrA* mutation as found in this study does not usually confer CLSI-classified resistance (this requires at least another mutation in *parC* or upregulation of efflux). The data reflect an evolution towards reduced susceptibility, not “persistent resistance” as stated in the title and in the text.*

R20. We share this reservation and have adopted the term ‘towards’. But because the *gyrA* mutations we observed are indicative of high-level resistance, we have settled on “towards resistance”, recognizing that ‘towards’ is not the same as ‘arrival at’. Accordingly, we have revised the title to “Bacteria in the human gut evolve towards resistance with low fitness cost after a brief antibiotic exposure”.

*The mutations were detected only in abundant commensals but not in less abundant bacteria that could become pathogenic such as *E. coli* (line 262). Horizontal gene transfer of *gyrase* alleles is rare, limiting the spread of mutations from commensals to pathogens. Moreover, clinical data showing relevant resistance development during treatment of bacterial diarrhea seem to be rare (despite initial bacterial population sizes that are likely much higher than the estimates provided in Suppl. Note 1).*

The clinical relevance of the evolutionary changes observed in this study may be limited.

R21. We realize now that we could have explained the clinical relevance of this work with greater specificity and nuance. In our view, the clinical relevance of the work lies in demonstrating the capacity of the human gut in healthy individuals to support the evolution of AMR. The evolution we observe among commensals, which is frequent and involves all major gut phyla, demonstrates that pathogens that may be present in the gut during subclinical infections will be exposed to similar selective pressures and may respond similarly. While it is well established that resistant strains appear after fluoroquinolone treatment, genetic analysis has typically indicated that these resistant strains evolved elsewhere. Demonstrating that the human gut is a habitat that intrinsically supports the evolution of fluoroquinolone resistance is novel, and clinically important – strategies to combat strain colonization/transmission are quite different from strategies to combat strain evolution.

We have revised the Introduction to spell out the clinical relevance, as we see it (lines 30-50).

As a side note, there were only 7 *E. coli* populations with serine at gyrA:83, so the absence of gyrA sweeps in *E. coli* is within the range displayed by other commensals (as only ~10% of populations evolve on average). Regarding the transfer of resistance, we searched the literature and found some interesting data on the transfer of resistance-associated gyrA from commensals to pathogens, see Manoharan-Basil et al. 2022 (PMID 35369513), Ferrándiz et al. 2000 (PMID: 10722479) and Chen et al. 2020 (PMID 31740556).

2. Role of other mutations

Increased drug efflux can be a major component of resistance to fluoroquinolones, particularly at early steps of resistance evolution. However, efflux-associated mutations were rarely found in this study (only 4 variants in KO2004). Could this reflect a larger mutational space for upregulating efflux, which would make detecting the underlying mutations difficult compared to mutational hotspots such as the quinolone resistance-determining region on GyrA? This could also affect other gyrA-independent resistance mechanisms that are found in many Gram-positive and Gram-negative fluoroquinolone-resistant strains. Thus, the data might underestimate the level of resistance evolution.

R22. Thank you for raising this very valid point. Indeed, any case in which multiple mutations generate a similar phenotype can result in so called ‘soft’ sweeps. We added some text to the Discussion that clarifies this matter (line 425).

The new experimental data we generated by amplifying a region of gyrA (**Figure 6** in the revision) reveal this exact phenomenon, where we report on a soft sweep involving at least 7 sites in *Prevotella copri*.

3. Susceptibility to ciprofloxacin in the gut

Figs. 1A, S4, S5 show diverse responses of microbiota populations during and after ciprofloxacin treatment. This is great data and it would be very interesting to relate these results to reported data on fluoroquinolone susceptibility of the corresponding bacterial taxa under in-vitro conditions (e.g. PubMedID 34646011). This could yield insights into potential differences between drug action in laboratory cultures vs. in the human gut (a large number of populations showed surprisingly weak responses to treatment).

R23. We were also surprised by the relative resilience of many gut populations. Maier et al. 2021 (PMID 34646011) determined the MICs (minimal inhibitory concentration) of ciprofloxacin for 40 human gut commensals, and interestingly all but one were sensitive to ciprofloxacin (Extended Data Fig. 1 in Maier et al.). Their MICs of ciprofloxacin were relatively low (0.06-32mg/L, Extended Data Fig. 2 there), especially

in contrast to the high ciprofloxacin concentrations found in the human gut lumen (>800mg/L) (Pecquet et al., PMID 2211433). For example, the MIC of *Roseburia intestinalis* was 32 mg/L in line with a relative increase in abundance during antibiotics for the *Roseburia* genus (Supp. Figure S5). The discrepancy between MICs measured *in vitro* and responses *in vivo* can be explained by a multitude of factors such as temporal/spatial variations in drug concentrations, the formation of protective biofilms, presence of persister cells, and more. We added text on this issue to the Introduction (line 66) and dedicated a paragraph at the end of the Discussion that speculates on the factors that affect evolvability *in vivo* (lines 429-437).

In addition, the comparison of responses during treatment and the corresponding gyrA genotypes could provide actual data for the assumed impact of various gyrA genotypes on susceptibility of microbiota in the human gut.

R24. Thank you for this great suggestion! We added a plot of abundance trajectories that shows the dramatic effect of having a resistant allele at gyrA:83 at baseline. See **Fig. 3F** and **Supp. Fig. S22**.

Minor points:

Fig. 3c, title of y-axis should be rather “number of genomes”?

R25. You are correct, thanks for catching that.

Lines 237-239, Fig. 3D: The calculation underlying the 10% probability of substitution from gyrA S83 towards resistance remains unclear. gyrA sweeps (including other mutations than S83) occurred in 63 populations (lines 207-208) out of 1866 analyzed populations (line 1866), ≈60% of which had S83 prior to exposure? This would give a lower probability. Yet another calculation is given in line 267-271.

R26. We agree that the text was confusing. For the gyrA:83 substitution analysis we used 698 populations that (1) had a single *gyrA* gene that aligned to the *E. coli gyrA* at position #83, and (2) had sufficient coverage to detect sweeps (line 356 in Methods). At baseline there were 410 population with S83, out of which 42 had sweeps during antibiotics, leading to the estimate. We have revised the text to clearly state the number of populations used (line 266).

The original 1866 populations (those with one or more dynamic variants) were not suitable for this calculation, since we wanted to include only populations that had a resolved *gyrA* (n=576) and also include 122 populations that lacked dynamic variants but had sufficient coverage for us to detect sweeps.

The second parameter we estimated is *gyrA* evolvability, which equaled 0.11. This was defined as the probability of a sweep in *gyrA* in populations that had S83 and sufficient coverage. The reason we limited the analysis to S83 is because we wanted a conservative assessment of the contribution of non-*gyrA* mutations to resistance, as these populations had sweeps in non-*gyrA* genes despite having S83 that remained unchanged through the exposure. The evolvability calculation includes a few cases in which sweeps occurred in *gyrA* at positions other than 83, leading to an evolvability of 0.11 which is slightly higher than the probability of a substitution at S83 (0.102 or 10.2%).

While reviewing this analysis we encountered and fixed a minor bug in the assignment of amino acids to gyrA:83 which had a negligible effect on the results (e.g., from 9.6% to 10.2% for serine substitutions). We corrected all downstream figures accordingly.

Referee #3 (Remarks to the Author):

Yaffe and colleagues present a very interesting study in 60 human subjects exposed to a short course of oral ciprofloxacin treatment. The main aim was to investigate the effect of the quinolone exposure on the gut microbiota – specifically on selection and maintenance of resistance traits.

Study participants were exposed to 2x500mg ciprofloxacin orally for 5 days and self-collected (a total of 16) stool samples 9 weeks prior and for each of the 2 consecutive days immediately prior to the start of ciprofloxacin, daily during antibiotic exposure (Days 0-4) and for the following four days (Days 5-8), and then on Days 10, 18, 28, and 77. Subjects had not taken any antibiotics for at least 6 months prior to the start of sampling.

Metagenome-assembled genomes were generated in 960 collected samples sequenced with high read per sample sequencing (18.8M). This was done by a novel tool named PolyPanner that compares individual time-point samples to a subject-specific metagenome generated from all samples from a given individual to identify fluctuations in single-nucleotide profiles (which then identifies polymorphic variants). 5655 populations spanning 924 species were analyzed in this study.

Key findings were:

- 1) Only few extinction events overall and return to baseline suggest a high resilience of the gut microbiota to such a treatment*
- 2) Dynamic variants were mainly SNPs and found in 38% of populations analyzed. 1866 were further processed to identify putative selective sweeps evolving longitudinally (linkage-group based)*
- 3) 7132 variants swept from pre-antibiotic to antibiotic “era” (frequency increase) in 513 populations (212 species)*
- 4) gyrA mutations were concentrated on positions 83 and 87 (56 and 9 substitutions, respectively). These mutations are known to be involved in development of FQN resistance and gyrA:83 mutations occurred in 33 subjects in 38 species. Longitudinal analysis (comparison to baseline) suggested a transition to AA associated with resistance in approx. 10% of susceptible populations. Furthermore, there was evidence for pre-existing resistance.*
- 5) Interestingly, E. coli - a key taxon – had an estimated 41% resistance rate at baseline without evidence for sweep in these populations during the study.*
- 6) gyrA evolvability was not the only sweep driver and accounted for only 36% of overall evolvability. Evolvability was associated with FQN susceptibility as determined by abundance reduction (being high in Bacteroidetes and Actinobacteria and low in Firmicutes).*
- 7) After removal of exposure there was only a low reversal of sweep variants in populations (frequency above 80% for most gyrA sweeps).*

Overall, this is a very interesting pilot trial including the bioinformatics methods, which seems to be the main novelty (that I cannot judge to the code level). Assembling genomes from metagenomics data in such a rich environment is really challenging. PolyPanner seems to be a significant step forward in this matter. The choice of ciprofloxacin is interesting as resistance “development” under (mono)treatment clinically occurs in several species (e.g., S. aureus) by mutations in the target genes (gyrase/topoisomerase).

Indeed, the authors demonstrate a sweep in gyrA mutations (exposure of ciprofloxacin resulted in a sweep of mutations in gyrA in ~10% of investigated microbial populations) in their work indicating that the use of FQN can lead to selection of such mutants.

However, the work has several limitations mainly being in the study design with no control group and additionally - for this journal - a phenotypic confirmation of key aspects such as MIC determination, confirmation of mutations in cultures populations, and assessments of fitness traits should be performed. Furthermore, the clinical relevance of the accumulation of those mutations remains unclear.

R27. Thank you for a thoughtful and constructive review. We attempted to address all concerns raised. We added 4 control subjects that did not take antibiotics. We did not isolate strains and assess fitness in vitro, as the role of *gyrA* in resistance has already been established. However, we added experimental data that strengthen our claim that resistant *gyrA* alleles have little or no fitness cost in vivo. See detailed replies inline below.

Specific strengths are the sophisticated novel bioinformatic analyses and the human trial, whereas specific weaknesses are the lack of a control group w/o antibiotic, the lack of phenotypic confirmation of resistant phenotype and associated predicted low fitness cost in key taxa. Although the discovery of gyrA mutant sweeps is interesting, the clinical consequence of those is unclear as invasive infections with most of those strains is rare and evidence for horizontal transfer was low.

R28. In retrospect we realize we could have done a better job of explaining the clinical consequences of this study in the original manuscript. We have now revised the Introduction to clearly state why it is valuable to track the evolution of commensals in subjects taking antibiotics. In a nutshell, the capacity of the human gut to support evolution over relatively short periods of time is practically uncharted. Studies that compared pre- and post- antibiotic states either did not genotype strains or when they did so, they found that ciprofloxacin resistance emerges primarily through strain-turnover events, in which the resistant strains have evolved elsewhere. We have revised the Introduction to explain this subtle yet important point (line 35 and lines 59-71). Other antibiotics are also mostly uncharted.

One small note regards horizontal gene transfer (HGT). The approach presented in the paper did not assay HGT and therefore we cannot tell if HGT was frequent or infrequent in those subjects. Future work, using appropriate assays to monitor HGT will take place in subjects exposed to fluoroquinolones. We have revised the discussion to explain this limitation.

Specifically:

250 Given the long-lasting effect demonstrated here; is there any information of whether or not these patients were exposed to FQN before the 6 months of antibiotic-free interval they were supposed to have to be enrolled?

R29. Unfortunately, we do not have data on FQN exposure more than 6 months prior to enrollment.

256 A control group would be useful to gather more evidence for this.

R30. We have now added a control group comprised of 4 subjects that did not take antibiotics and were sampled with 16 samples in a similar fashion to the ciprofloxacin cohort (lines 12-15 in Methods). These subjects displayed reduced rates of sweep events, yet some evolutionary changes did occur during the 20 weeks of sampling (see **Figure 2G** in the revision, and line 216).

318 *This should be confirmed by phenotypic fitness assays for some key taxa*

R31. Isolating and assaying commensals is challenging and even if achieved, would not necessarily add significantly to our key findings. While in general genotype-phenotype associations should be confirmed *in vitro*, in this specific case, a body of existing literature has established the involvement of mutations at *gyrA*:83 (and nearby loci) in fluoroquinolone resistance. As a secondary issue, interpretation of MICs of poorly characterized commensal species can be problematic. What is novel in our study is the assessment of fitness *in vivo*, with the natural selective pressures of the human gut during and post antibiotic exposure.

To get a better grasp of the fitness cost of mutations in *gyrA* we have added new experimental data. We chose 8 representative species and amplified a short region that spans *gyrA*:83 in all of them using tailored primers (new **Figure 6**, lines 356-383 in Results, lines 393-449 in Methods). The two key findings from this new analysis were that (1) resistant alleles were present prior to the exposure in 2 out of the 8 cases, while the other 6 cases are consistent with *de novo* mutations; and (2), we found no measurable fitness cost for the resistance mutations.

344 *This is a bit a stretch – FQN should not be a first line (empiric) treatment of UTI anyways and if they are, there is probably a clinical reason (e.g., P. aeruginosa where FQN are the only oral option). Assuming there is time to screen for abundance of gut-residing pathogens then there is also time to culture urine and initiate a targeted antibiotic treatment. This should therefore be rephrased, as antibiotic stewardship should already be good enough to prevent misuse of FQN for this indication.*

R32. We agree and have removed the statement in question.

Our replies below are highlighted in blue.

Referee #1 (Remarks to the Author)

I thank the authors for doing a great job in addressing our comments! The work represents a thorough investigation into the evolutionary dynamics of gut microbiomes when exposed to antibiotics.

Thank you for taking the time to review our manuscript in detail. We greatly appreciate your thoughtful comments and constructive feedback, which have helped us enhance the clarity and completeness of the work.

I have a few minor comments that I feel if addressed would improve the paper a bit.

- I like the added note on the supplement about the difference between the coverage approach of Garud et al. and their more statistical approach for removing ortholog variants. However, I think I just don't have a good grasp on how much of a problem ortholog variants present. In their supplement, figure S10, they show that only 3.5% of the variants they detect get classified as “dynamic variants” or true variants. Given that they are assembling using MAGs, is having over 85% of the genomic variants detected in the data due to paralogs/orthologs an expected result? (It seems very high to me, but I do not have any data to compare it to).

This question highlights an important limitation of our approach. For each variant, we test two null hypotheses: one asserting it is a paralog and the other that it is an ortholog. If we fail to reject both hypotheses, the variant is classified as likely spurious. However, this does not rule out the possibility that the variant is in fact polymorphic. Variants with low coverage—whether due to insufficient sequencing depth or low allele frequency—are particularly prone to misclassification as spurious. In such cases, it becomes challenging to distinguish between dynamic variants, paralogs, and orthologs. We have clarified this limitation in Extended Data Figure 7 (previously Supp. Figure S10) with the following text:

“For each variant, we test two null hypotheses, i.e., that it is a paralog or an ortholog (both considered spurious variants). Note that variants with low coverage may be misclassified as spurious, even if they are true polymorphisms.”

- I do agree however that by using such a specific statistical test, they are putting a much higher requirement on what variants to consider as “true polymorphisms” (Or “dynamic variants”) and therefore, the results they do find are a lot more reliable (in

terms of me believing that the variants they found are due to true polymorphisms instead of potential ortholog or other variants).

- I still do not understand where the 1000 dynamic strain cutoff comes from (as a number) but the inclusion of 2C is nice to show that the number of genomes thrown out by this threshold is not a majority. I think explicitly citing a paper (either Lieberman/Zhao or Garud/Good) and/or showing that the number of between host SNVs is $\sim 10^4$ would be useful for justifying this.

The specific threshold is arbitrary, yet other thresholds between 100 and 10000 yield comparable results. To address this request, we added the following underlined comment in the main text:

“We omitted 267 populations with >1000 variants from downstream analysis, as our focus here was evolution *in vivo* (using empirical thresholds of inter-subject strain relatedness^{33,35}).

Where references 33 and 35 are Garud et al. and Zhao et al.

Git-Hub Comments:

- Github is generally well commented. I have a couple of small notes:
- It seems like some utility scripts are required to run the workflow before you even start using the PolyPlanner commands (PolyPlanner is a command-line tool). In particular, things like `utils/parse_bwa_sam.pl` or `utils/bin_summary.r`. I think it might be useful to include the expected version of perl or R required for those scripts?

We added the following lines to the README on github:

“There are several helper scripts written in Perl and R. These scripts were tested on Perl versions 5.30.3 and 5.30.0, and on R versions 4.3.1 and 4.4.1, but they are likely compatible with other comparable versions.”

- One step of their workflow is “paired reads are transformed into POP files”. Is this a standard bioinformatic procedure (I have never done it and would not know how to immediately)? I wonder if including the specific command to perform that action would be useful for the pipeline to be easily accessed by others. Reading a little more into the github, I think it is the command `polypanner construct`. Might just be worth mentioning it directly in the workflow explanation.

The POP format is specific to PolyPanner and not a standard format. To clarify, we added the following explanation on github:

“PolyPanner uses an intermediate binary representation of mapped reads, referred to as the POP format. Users do not interact directly with POP files; instead, PolyPanner

generates these files from mapped reads and outputs dynamic variants in a tab-delimited format.”

Additionally, we clarified the usage of the construct command on github:

“... paired reads are transformed into POP files using the construct command (see the syntax of the construct command [here](#)).”

Additional comments:

1. Fig S26 has an ‘A’

Thanks for pointing out this typo!

2. What is an x-coverage profile?

X-coverage is the number of reads that map to a specific position. An x-coverage profile is a vector of x-coverages over time. We added the underlined comment below to the main text for clarity.

“For each sample, reads were aligned to their respective subject assembly, generating single-nucleotide coverage profiles (temporal vectors) which represent fluctuations in read coverage of all assembled base pairs over the 16 time points.”

3. What do the different colors for the coverage plots in S10 and S26 mean?

Variants were colored in both figures according to the variant class, as specified in panel A in Supp. Figure S10 (now Extended Data Figure 7). We added the following text at the end of the legend of Supp. Figure S10:

“Variants in panels B-C colored according to their class, as specified in panel A.”

4. Do you think that your method is underpowered to detect soft sweeps, or, is it underpowered to detect very soft sweeps or soft sweeps that do not meet the <0.2 to >0.8 criteria (e.g. a partial soft sweep)?

Given sufficient sequencing depth, our method can detect any type of sweep, including soft sweeps. For example, as shown in Extended Data Figure 5 (previously Supplementary Figure S7), a sweep from 0 to 0.2 can be detected if the genome x-coverage exceeds $\sim 3000\times$ (see panel B). More generally, a soft sweep involving N loci that sweep together will fixate at a frequency of $1/N$. For instance, if a certain phenotype can be achieved through 100 different mutations (e.g., reducing transcription via mutations in a promoter region), and all these mutations are part of the standing variation, each mutation might fixate at a frequency of 0.01. In such cases, given our relatively poor sequencing depth ($\sim 100\times$), we would not detect any of these sweeps.

Could it be that there are more soft sweeps in the data and currently we are just looking at extreme allele frequency changes that could be attributed to a hard or a soft sweep?

Yes, this is likely. The amplification of *gyrA* (Figure 6) supports this hypothesis. The high resolution (~10000x) enabled us to detect low-frequency variants undergoing sweeps. For instance, in *Prevotella copri* (the last population in Figure 6), we observed seven variants that swept, some of which plateaued at frequencies well below 0.01.

Referee #2 (Remarks to the Author)

The authors have addressed appropriately all my comments and questions. They might want to mention explicitly also the many species that survived rather well the treatment without detectable sweeps, in the paragraph discussing the survival of *gyrA* mutants (lines 429-437). These findings actually suggest that the antibiotic treatment was actually quite inefficient in the gut despite the rather high concentrations and the previously reported susceptibility of many of these species to fluoroquinolones.

Thank you for taking the time to thoroughly review our manuscript. We greatly appreciate your thoughtful comments and constructive feedback, which have helped us enhance the clarity and depth of our work.

Regarding your observation, we added the following text to the end of that paragraph: "Intriguingly, most populations persisted in the gut without detectable sweeps in *gyrA* or other loci, suggesting that ciprofloxacin was relatively ineffective in the gut, despite its high concentrations and the previously reported *in vitro* susceptibility of most commensals (Maier et al. 2021)."

Referee #3 (Remarks to the Author)

In the resubmitted version, Yaffe and colleagues have significantly revised the manuscript:

Key changes are:

- revised introduction for pointing out possible clinical consequences of sweeps
- a control group (4 subjects) was added with 16 sampling points "matching" the time points of the initial intervention group; the control group had a lower rate of sweep events; it might be worth adding more information on the sweep events in the control group - at least whether or not genes of interest (e.g., *gyrA* but also compared to the evolving populations without *gyrA* mutations) were involved
- rephrasing of resistance to evolution towards resistance (also now reflected in the title) and added novel analyses on populations that had selected sweeps in *gyrA*::83 (4 phyla

with various substitutions) using PCR to analyze fitness effects.

By doing this and addressing other points raised (or rephrasing / omitting) I feel that the comments have been adequately addressed. For an overall summary and detailed judgment I take the liberty to refer to my initial summary. I still believe that this is a very interesting and well done study that has gained strength by adding a (small) control group and performing additional fitness data. By downphrasing to "towards resistance" I also do not feel the need for extended phenotypic confirmations given the available literature.

Thank you for your detailed review of our manuscript and for highlighting the key revisions we made. We greatly appreciate your thoughtful comments and constructive feedback, which have significantly contributed to improving the clarity and rigor of our study. Your insights, particularly regarding the addition of the control group and the nuanced framing of evolution towards resistance, have been invaluable in strengthening our work.