

Universal rules govern plasmid copy number

Corresponding Author: Dr Jeronimo Rodriguez-Beltran

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)

The manuscript by Ramiro-Martínez describes universal rules governing plasmid copy number. The authors map sequence coverage on plasmids and chromosomes and use it to quantify plasmid copy number (PCN). They analyse the PCN of plasmids in depth providing several very interesting results for plasmid biology. The observation of a scaling law between size and PCN is not really new (see below) and this should be acknowledged. Yet, the other results are very novel and some are very interesting. For example, the authors show that the distribution of copy number is approximately bi-modal, quite similar across species and they describe how it varies across families of plasmids. Particularly interesting is the section on dominance that tackles a topic that was in need of clarification: what happens in terms of replication when two plasmids with different PCN merge. They show a diversity of situations, within a general trend towards the plasmid having lower rather than higher PCN. Finally, they show that plasmid load DNA is conserved relative to chromosome size, which is interesting and unexpected (at least to me). I have three major comments and a few minor ones. The manuscript is clear and well-written.

Major comments:

- The scaling law is not really new. It was thought to exist for decades based on sparse data, but it did get demonstration recently in a study that is cited quite disparagingly. Shaw, Science Advances, 2021 Figure 3 shows exactly this scaling law. This should be acknowledged. All the more so, that the authors use Shaw et al data in figure S19 and find a perfect fit between the two!
- There is a pervasive lack of mention to variance in the data. As can be seen from figure S16, there is some variance around these trends. One thus needs to have some information on intervals of confidence. Not of the regression lines (those are given and it's fine), but about the data points. More precisely, if a plasmid has a given size what is the interval of confidence at 5% for the expected size? This matches a standard query of a researcher: given a plasmid what is the expectation of the size. This information is also important to evaluate the claims about constraints, since the figure does show a lot of variance, indicating some degree of flexibility. In my opinion Fig S16 should be in main text. Here, as elsewhere see below, the statistical analyses are sometimes inconsistent. The figure provides a linear regression but the statistics are about correlation. Both statistics should be given. And especially the R² to have an idea of the percent of variance that is explained by the model. The data in S16 suggests R² on the Gram+ should be about 0.5 which shows that there is a lot of unexplained variance which is an important information that is not apparent enough in the main text.
- Statistical analyses are for the most fine. But there are reporting issues. An example mentioned above concerns the figures showing regression lines but statistics corresponding to correlations. There are also statements about correlations that are grounded on coefficients of determination of regressions (line 282 for example). Sometimes tests are described too quickly and insufficiently (figure s12, statistical differences at which alpha level?). The reporting of the statistical analyses needs a thorough check. Whenever there is a linear regression there should be a supplementary table with R², estimates of the parameters, and P-values for all this.

Minor comments.

- The introduction should have more information on previous ideas about the association between PCN and plasmid size.
- "Higher copy number replicons were generally recessive to replicons showing lower copy numbers". I found this observation very interesting. From the point of view of fitness impact this makes a lot of sense. But from the point of view of the mechanisms this is intriguing because the high PCN should be dominant over the low PCN (because it will start replication until copy number is high enough). I think a more thorough discussion on this point would enrich the study. I also wonder if the oriVs of the high PCN in the fusions with low PCN show evidence of mutation accumulations that could justify the result. One could think that selection would favour mutations lowering PCN after co-integration.
- I found fig4DEF a bit difficult to follow and in lacking of more quantitative data on the frequency of each case.

- On the same point. While I understand that some cases are very rare and not much will come out of detailing them, the results on dominance could be detailed. Why is there dominance? (my comment above on mutations in oriV is an example).
- The figure 5C is a fit on medians, which vastly inflates the fit and gives a false impression the model is perfect. When it's probably not (as suggested by results in figure S15 and S16). These fits should account for the variation in the data not summary statistics.
- "Distribution of total DNA load per plasmid (x- axis) relative to chromosome size per genus (y-axis). The DNA load of each plasmid is calculated by multiplying the plasmid size by the copy number and then expressed as a proportion relative to the chromosome size" Given the importance of this measure one needs a more precise description in methods with mathematical formulas to remove ambiguity.
- " Our analysis revealed that, regardless of their host genus, size, or copy number, all plasmids accounted for approximately the same percentage of chromosomal DNA" This sentence needs nuance. The figure shows a variation over two orders of magnitude. Even if the averages are close the variance is large.
- " Moreover, intrinsic variability and PCN were positively and strongly correlated when we classified plasmids according to their replicon type, PTU, or plasmid cluster (Spearman's rank correlation $p < 10^{-2}$; Figure 3, Extended Data Fig. 9)." It's not completely clear to me why in the main text the coefficient of variation is different in terms of PCN but not in Extended Data Fig. 9.
- "These results demonstrate that most plasmids encode robust intrinsic replication control mechanisms that determine their characteristic PCN." Maybe rephrase? As such, this was well known.
- Several references lack page numbers or manuscript identification.
- Figure S12 lacks the colour code (mauve versus light orange) in the boxplots. I assume it's the same as in S11.
- Gram negative and gram positive are outdated taxonomical groups. Here, it's Proteobacteria and Firmicutes (or Pseudomonadota and Bacillota).

(Remarks on code availability)

The code in the git would deserve a manual for guidance and a toy example for a first experience.

Reviewer #2

(Remarks to the Author)

Universal rules govern plasmid copy number

Ramiro-Martínez et al. have produced a fantastic work of fundamental research towards a systematic understanding of plasmid copy number using publicly available data. In combination with Maddamsetti et al., these two articles were a pleasure to read. I did not have any major concerns with the article, with only some minor comments as below.

line 49: typo: "...and is mainly limited to..."

figure 1b: Could you clarify if "non-mobilisable" is based on the absence of any replicon machinery? Did you consider an oriT analysis?

fig 2b: Density plots are notoriously difficult to see raw numbers - could you add the raw numbers for each category?

line 271: "...only Escherichia and Salmonella being different..." - I'm not sure I can see in your figure what you are referring to in terms of Escherichia and Salmonella being different, can you edit the figure or describe more in text (numerically)?

line 329: "(keeping only circular genomes)" - could you describe how circularity was determined and if you mean circularity of all plasmids/chromosomes?

Line 335: Was there any QC of the raw reads before mapping?

Line 346: You state how many plasmids there were at the end of filtering, but I could not find how many plasmids there were initially. Could you add how many were removed because of $<30\times$ or <1 PCN?

Did you consider the presence of phage-plasmids, and how these might differ in PCN vs. plasmids? If not, please consider adding this as a future direction/limitation.

Line 355 (onwards): please add versions of tools (and DBs where relevant) including parameters used. The Github docs should also have which versions of tools the scripts have been tested.

378-380: "The relative percentage of plasmid DNA load was calculated as (plasmid size (bp) / chromosome size (bp) $\times 100$)" I just wanted to check this was correct, based on the previous sentence where plasmid DNA load is plasmid size $\times$ PCN. So should the relative percentage of plasmid DNA load be plasmid load (size $\times$ PCN) / chromosome size?

(Remarks on code availability)

The github repo is well-written and clear in its description of methods. My only minor comment would be to add versions of third-party tools used as a record of compatibility if tools are updated in future.

Reviewer #3

(Remarks to the Author)

The authors need to take into account the following major points:

#) Mathematical Dependence

Since the coefficient of variation (CV) is calculated by dividing the standard deviation by the mean, the mean and CV are mathematically dependent. Therefore, when interpreting the correlation between the mean and CV, it is important to carefully consider that this may not reflect a biological relationship, but rather result from mathematical dependence.

#) Non-Independence of Data

In analyses involving multiple plasmids, the assumption that data points are independent is made. However, this assumption could lead to misleading results. Organisms (and replicons such as plasmids) do not evolve independently; they share traits due to common ancestry. Thus, plasmids cannot be treated as independent data points. This issue is similar to the problem of pseudoreplication in experimental research, which can introduce significant bias (type I errors) in statistical analyses.

For example, carefully reconsider the following analysis in the manuscript:

Figure 3. Intrinsic PCN variability. (***: $p < 10^{-9}$, Wilcoxon rank sum exact test).

Here are quotes from a reference paper:

<https://pmc.ncbi.nlm.nih.gov/articles/PMC7612097/>

Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range

Importance of controlling for non-independence of genomes

#) Validating the results

It seems that the authors merely ran bioinformatics tools for plasmid analysis without carefully validating the results. For example, for plasmid sequences assigned as 'rep_type.s.' to 'IncP' by tools like 'mob_typer' or 'plasmidfinder', there are sequences where the 'predicted_mobility' is assigned as 'mobilizable' or 'non-mobilizable' instead of 'conjugative', and sequences where the 'Relaxase' is assigned to 'MOBQ' or 'MOBH' instead of 'MOBP'.

<https://github.com/PaulaRamiro/NpAUREO/blob/main/GeneralWorkflow.md>

#) DNA sequencing data

The authors stated, "we leveraged extensive DNA sequencing data to comprehensively analyze the copy number of thousands of diverse bacterial plasmids."

but if they are not certain about how these DNA sequencing data were obtained (e.g., bacterial growth conditions and DNA extraction methods used), the reliability of the results may be questionable.

The authors used raw reads and assemblies downloaded from the NCBI SRA database, but they should have considered (or at least discussed) this point.

Here are quotes from a reference paper:

<https://pmc.ncbi.nlm.nih.gov/articles/PMC6312765/>

Reconsidering plasmid maintenance factors for computational plasmid design

Plasmid size may be associated with copy number.

Plasmid copy number estimates can vary, depending on bacterial growth conditions and DNA extraction methods used [97,99]. Therefore, copy number data should be interpreted carefully.

Here are some minor points:

#) Supplementary Dataset 1

558670_0_supp_4562777_skztgk.xlsx

Please clarify the meaning of the column labels.

#) natural plasmids

The authors should clearly specify the method used to obtain only "natural plasmids" (i.e., to exclude artificial sequences) from the database.

#) phylogenetic boundaries

The authors should provide a clear definition of "phylogenetic boundaries" if they claim that such boundaries exist for organisms that have diverged from a common ancestor and thus form a continuum.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I thank the reviewers for having very seriously tackled all my comments. The answers and actions were appropriate. The only major comment that wasn't really tackled regards the oriV evolution issue, but since they are preparing one paper on the topic, I'll wait for that one.

(Remarks on code availability)

While the authors have added a toy example, I still think the Git should have more instructions for the use of the software, a FAQ, etc.

Reviewer #2

(Remarks to the Author)

All previous comments have been addressed adequately. This is a wonderful piece of work and I look forward to seeing it published.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

<https://github.com/PaulaRamiro/NpAUREO/blob/main/Example.md>

Please verify the usability and reproducibility of the code in different computer environments.

(Remarks on code availability)

git clone <https://github.com/PaulaRamiro/NpAUREO/>

<https://github.com/PaulaRamiro/NpAUREO/blob/main/Example.md>

Following the instructions in Example.md, the commands and output examples executed are as follows.

```
+ # bash #
wget https://ftp.ncbi.nlm.nih.gov/genomes/genbank/assembly_summary_genbank.txt
grep "Staphylococcus" assembly_summary_genbank.txt | grep "Complete Genome" > data_staphylococcus.txt

grep "GCA_001018645" data_staphylococcus.txt > staphylococcus_example.txt
```

```
esearch -db assembly -query SAMN03255442 | elink -target nucleotide -name assembly_nuccore_refseq | efetch -format
fasta > SAMN03255442.fasta
```

resulted in an empty file (file size 0).

Here is the output:

```
$esearch -db assembly -query SAMN03255442 | elink -target nucleotide -name assembly_nuccore_refseq | efetch -format
fasta
$esearch -db assembly -query SAMN03255442 | elink -target nucleotide -name assembly_nuccore_refseq
<ENTREZ_DIRECT>
<Db>nucleotide</Db>
<Count>0</Count>
<Step>3</Step>
<Elapsed>2</Elapsed>
</ENTREZ_DIRECT>
$esearch -db assembly -query SAMN03255442
<ENTREZ_DIRECT>
<Db>assembly</Db>
<WebEnv>MCID_67af67d482b9378dba0e079a</WebEnv>
<QueryKey>1</QueryKey>
<Count>2</Count>
```

```
<Step>1</Step>
<Elapsed>1</Elapsed>
</ENTREZ_DIRECT>
```

```
+ # bash #
```

```
esearch -db sra -query SAMN03255442 | efetch -format runinfo > SAMN03255442.numbers
```

```
grep "ILLUMINA" SAMN03255442.numbers | grep "PAIRED" > filtered_SAMN03255442.numbers
```

```
+ # bash #
```

```
fasterq-dump --split-3 SRR1955495
```

The above command output the following message.

```
2025-02-14T16:53:38 fasterq-dump.3.2.0 err: cannot find 'SRR1955495' ->
RC(rcVFS,rcQuery,rcExecuting,rcItem,rcNotFound)
fasterq-dump quit with error code 3
```

```
$du -sh *
18M NpAUREO
0B SAMN03255442
0B SAMN03255442.fasta
4.0K SAMN03255442.numbers
1.3G assembly_summary_genbank.txt
1.7M data_staphylococcus.txt
4.0K filtered_SAMN03255442.numbers
4.0K staphylococcus_example.txt
```

```
$uname -a
Darwin MacBook-Pro-14-inch-2021.local 21.2.0 Darwin Kernel Version 21.2.0: Sun Nov 28 20:28:41 PST 2021; root:xnu-
8019.61.5~1/RELEASE_ARM64_T6000 arm64
```

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/>

We would like to thank the editor and the reviewers for their helpful comments, which we think have greatly improved the quality of this manuscript. The revised version of the manuscript highlights the changes we have made following the reviewers' suggestions in green.

Here, we provide a point-by-point description of how we have addressed the reviewers' comments (in blue, and our responses in black):

Reviewer #1 (Remarks to the Author):

The manuscript by Ramiro-Martínez describes universal rules governing plasmid copy number. The authors map sequence coverage on plasmids and chromosomes and use it to quantify plasmid copy number (PCN). They analyse the PCN of plasmids in depth providing several very interesting results for plasmid biology. The observation of a scaling law between size and PCN is not really new (see below) and this should be acknowledged. Yet, the other results are very novel and some are very interesting. For example, the authors show that the distribution of copy number is approximately bi-modal, quite similar across species and they describe how it varies across families of plasmids. Particularly interesting is the section on dominance that tackles a topic that was in need of clarification: what happens in terms of replication when two plasmids with different PCN merge. They show a diversity of situations, within a general trend towards the plasmid having lower rather than higher PCN. Finally, they show that plasmid load DNA is conserved relative to chromosome size, which is interesting and unexpected (at least to me). I have three major comments and a few minor ones. The manuscript is clear and well-written.

We thank the reviewer for praising our work, providing insightful comments, and constructive criticism. Please find our point-by-point replies below.

Major comments:

The scaling law is not really new. It was thought to exist for decades based on sparse data, but it did get demonstration recently in a study that is cited quite disparagingly. Shaw, Science Advances, 2021 Figure 3 shows exactly this scaling law. This should be acknowledged. All the more so, that the authors use Shaw et al data in figure S19 and find a perfect fit between the two!

We agree that Shaw et al. present data in their excellent paper that aligns with the scaling law outlined in our research. Additionally, the concept of small plasmids existing in multiple copies while larger plasmids exhibiting lower PCN is not novel; it is supported by limited yet consistent evidence. However, while Shaw et al. was most likely the first work displaying the almost perfect linear relationship between PCN and size (in their Figure 3), it was not formally discussed (nor alluded to) in writing.

Moreover, there are key differences between Shaw et al. and our manuscript. First, our work addresses the association between plasmid size and PCN and builds on it to more formally demonstrate that the power law (i.e. the slope) is conserved across different bacterial species and even phyla (results section starting in new line 240, Figure 5). We

further demonstrate that plasmid DNA load is conserved relative to chromosomal size (results section starting in new line 288, Figure 5). Therefore, our study provides deeper detail into the association between PCN and size shown by Shaw et al.

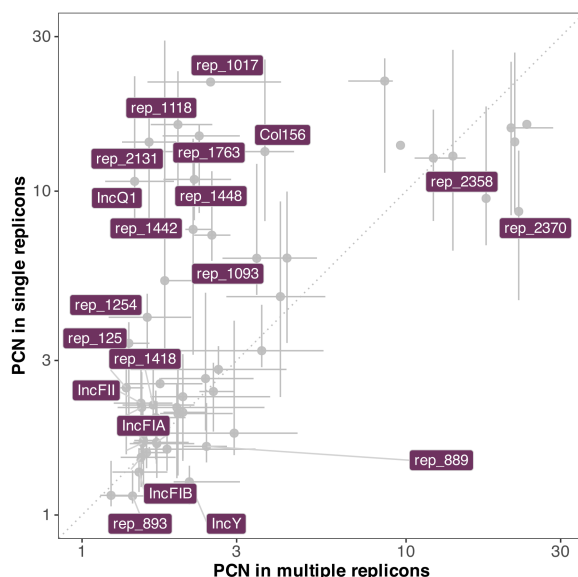
Second, the data of Shaw et al. is focused on Enterobacteriaceae. To provide a broader perspective, we expanded our analysis to other genera of Proteobacteria and included genomes belonging to (the distant phylum) Firmicutes, which further contributes to establishing the pervasiveness of the power law.

In summary, we believe that our manuscript discusses the scaling law more directly and delivers a more robust demonstration. Nonetheless, we agree that, in the previous version of the manuscript Shaw et al. was cited quite disparagingly. Therefore, we amended our introduction (see response to minor comment #1) and discussion sections to credit the work by Shaw et al. and two other studies showing compatible results:

Line 336: “This result agrees with previous observations with limited sampling(10,11) or concerning only Enterobacterial plasmids(9) and is further supported by a recent work identifying consistent scaling laws that relate plasmid size with copy number, protein-coding genes, and metabolic genes across ecological niches(45)”

There is a pervasive lack of mention to variance in the data. As can be seen from figure S16, there is some variance around these trends. One thus needs to have some information on intervals of confidence. Not of the regression lines (those are given and it's fine), but about the data points. More precisely, if a plasmid has a given size what is the interval of confidence at 5% for the expected size? This matches a standard query of a researcher: given a plasmid what is the expectation of the size. This information is also important to evaluate the claims about constraints, since the figure does show a lot of variance, indicating some degree of flexibility. In my opinion Fig S16 should be in main text.

We agree with the Reviewer that, as an extensive analysis of thousands of independent measures of PCN, our data shows significant variability. Although this variation was not explicitly shown in the previous version of some figures (i.e., Figure 4b, Figure 5a and 5c), the dispersion of the data was previously shown in related Supplementary Figures (Extended Data Fig. 11, 16, and 17).



However, we acknowledge that adding error bars and showing the actual variance of the data in the figures would improve the interpretation of our results, as it better conveys the idea that PCN is intrinsically noisy. Therefore, we have now added error bars depicting the standard deviation or the interquartile range in the appropriate figures (Figures 5a and 5c). Nevertheless, adding error bars to Figure 4b makes the interpretation of the data harder due to the many overlapping error bars and the large number of data points (see figure on the left).

Therefore, for this figure, we prefer to include variation in data as boxplots shown in Supplementary Figure 11, and now we explicitly refer the reader to that figure to ascertain variation as follows:

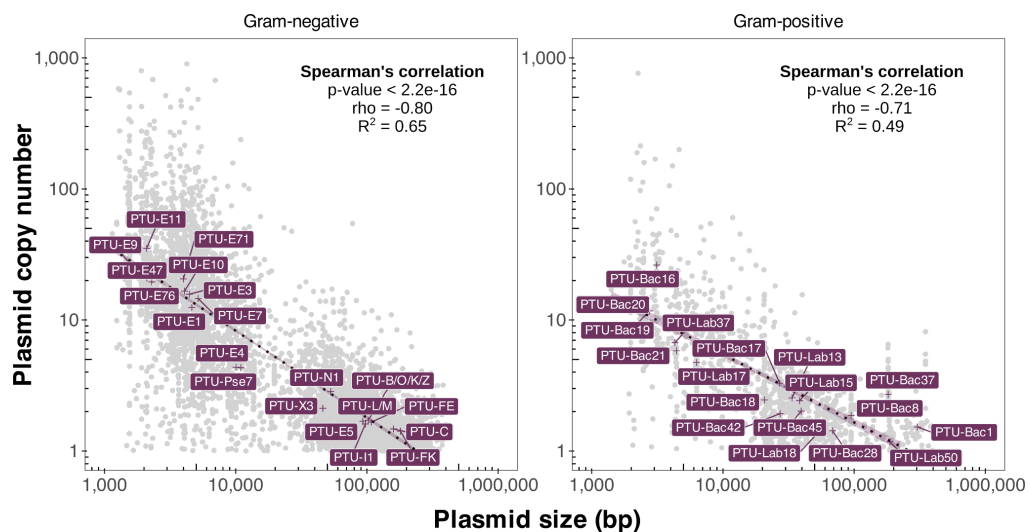
Lines 195 and 227 (Figure legend): “See Extended Data Fig. 11 for data represented as boxplots.”

Regarding Figure S16, we think that the main value of this figure is showing the data as it is (without aggregation). While this helps to visualise the individual variability of the data, we refrain from adding it to the main text because plotting all plasmids as separate data points might cause pseudoreplication (as many plasmids might share ancestry and be highly similar) and artificially improve the fit, as highlighted by Reviewer #3. Therefore, we decided to keep this figure as a Supplementary Figure.

Here, as elsewhere see below, the statistical analyses are sometimes inconsistent. The figure provides a linear regression but the statistics are about correlation. Both statistics should be given. And especially the R^2 to have an idea of the percent of variance that is explained by the model. The data in S16 suggests R^2 on the Gram+ should be about 0.5 which shows that there is a lot of unexplained variance which is an important information that is not apparent enough in the main text.

As stated below (see the response to the following point), we have improved the consistency of our statistical analyses and no longer use correlation and regression interchangeably. Moreover, as requested, we now included the R^2 values in Supplementary Figure S16 and the correlation statistics (see new figure below). As the Reviewer estimated, the R^2 for Gram+ is close to 0.5, indicating substantial unexplained variance. We now acknowledge this fact in the text as follows:

Line 250: “Although there was substantial unexplained variance, plasmid size and copy number were strongly and negatively associated (Extended Data Fig. 16).”



Statistical analyses are for the most fine. But there are reporting issues. An example mentioned above concerns the figures showing regression lines but statistics corresponding to correlations. There are also statements about correlations that are grounded on coefficients of determination of regressions (line 282 for example). Sometimes tests are described too quickly and insufficiently (figure s12, statistical differences at which alpha level?). The reporting of the statistical analyses needs a thorough check. Whenever there is a linear regression there should be a supplementary table with R^2 , estimates of the parameters, and P -values for all this.

We thank the reviewer for this comment and agree that there was some confusion regarding correlations and regressions in the previous version of the manuscript. We have carefully checked the manuscript and corrected affirmations about regression and correlation coefficients. We also removed linear fits that were unnecessary or not mentioned in the text (Extended data figure 19). We have also added regression parameters to Extended Data Figure S16, as suggested, and included the relevant regression parameters in Supplementary Dataset 7.

Regarding the alpha level, in the statistical analysis section (Methods) of the previous version of the paper, we stated (old line 416) that “The significance level was set at 0.05 for all statistical tests.” However, we acknowledge this may be hard to spot, and have added significance levels to all relevant figure legends (Extended data figure 12, Extended data figure 14, Extended data figure 4).

Minor comments.

The introduction should have more information on previous ideas about the association between PCN and plasmid size.

We now introduced a paragraph explaining (and referencing) previous ideas about the association between PCN and plasmid size. The section reads as follows:

Line 31: “Plasmids ensure their stability in microbial populations thanks to fine-tuned replication mechanisms that maintain a given number of plasmid copies per cell. Plasmid copy number (PCN) is thus a fundamental aspect of plasmid biology that governs plasmid lifestyles. Small plasmids typically lack active partition systems, so they are randomly distributed (segregated) to daughter cells. To avoid being stochastically lost during cell division, these plasmids rely on being present at a high PCN, which statistically guarantees their stable inheritance and persistence in the population (7). On the other hand, large plasmids are typically present at low PCN as they carry active segregation and partition systems that mechanistically ensure their persistence. Their low PCN likely reduces their metabolic load to the host, alleviating their fitness cost (8). Therefore, copy number and size are highly intertwined plasmid properties that have been shown to be negatively correlated (9–11).”

"Higher copy number replicons were generally recessive to replicons showing lower copy numbers". I found this observation very interesting. From the point of view of fitness impact this makes a lot of sense. But from the point of view of the mechanisms this is intriguing because the high PCN should be dominant over the low PCN (because it will start replication until copy number is high enough). I think a more thorough discussion on this point would enrich the study. I also wonder if the oriVs of the high PCN in the fusions with low PCN show evidence of mutation accumulations that could justify the result. One could think that selection would favour mutations lowering PCN after co-integration.

We agree with the reviewer that this point was scarcely discussed in the previous version of the manuscript, and now we have added a sentence that reads as follows:

Line 329: "In this regard, perhaps the most relevant result is that low PCN replicons are generally dominant to high PCN replicons. This suggests two non-mutually exclusive possibilities that await experimental validation: i) the replication machinery of the HCP (small) plasmid is inefficient for replicating a larger DNA molecule, and/or ii) selection favours larger plasmids that exist in low copy, as a mechanism to reduce fitness costs."

However, owing to the mechanism's speculative nature, we refrain from discussing it further (see also the response below). We also appreciate the reviewer's insightful suggestion regarding analysing oriV mutations in high PCN replicons. We are already analysing replicon dominance patterns in-depth for a forthcoming manuscript, and this idea is definitely worth exploring. However, we think that, at this point, it is beyond the scope of this article and thus have decided not to include it here.

- I found fig4DEF a bit difficult to follow and in lacking of more quantitative data on the frequency of each case.

We now provide a more detailed description of Figure 4DEF in the figure legend that we hope clarifies the figure. We have also included some minor modifications to the figure, making a clearer distinction between the expected and the observed data. We also changed labels to better reflect the expected outcomes of dominance (DOM_A or DOM_B):

Line 236 (figure legend): "Example of co-dominant replicons. Grey diamonds on panels d-f show predicted dominance outcomes: PCN additivity (Addit.), dominance of the higher (DOM_A) or lower (DOM_B) PCN replicon. Observed data (purple circles) depicts the observed median PCN for each case. Raw data for each replicon is available in Extended Data Fig. 12."

We have also included quantitative data on the frequency of each dominance case, which are reported in the text and Supplementary Figure 12 (see figure below).

The corresponding text now reads as:

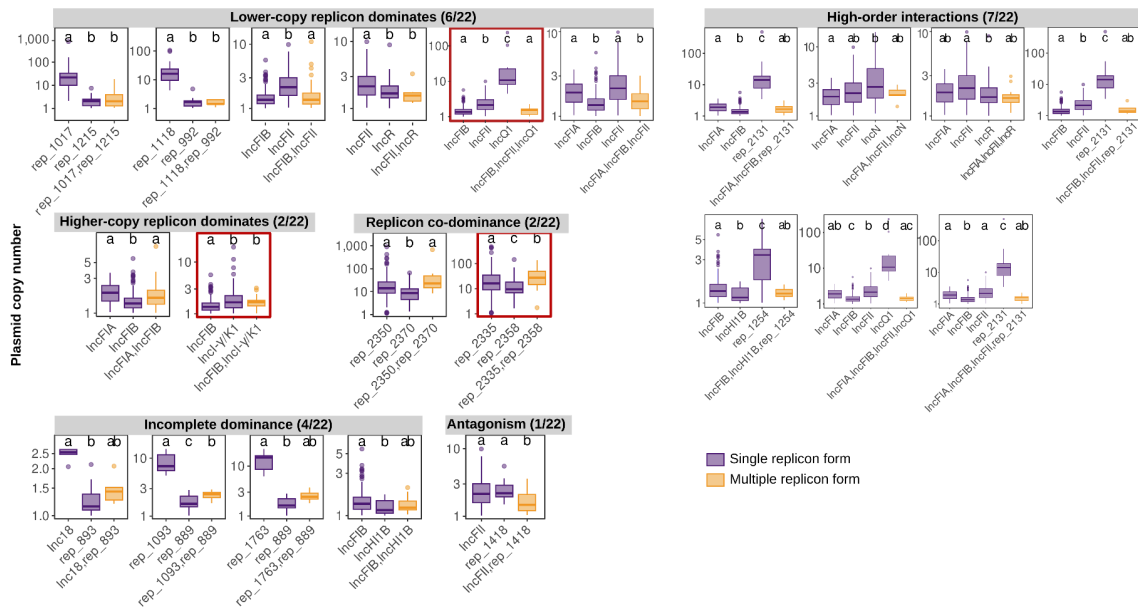
Line 203: "Higher copy number replicons were generally recessive to replicons showing lower copy numbers (6 of 22 cases; Figure 4d, Extended data Fig. 12)."

Line 208: "On the other hand, the higher copy replicon only dominated in plasmids containing multiple LCP-associated replicons, albeit PCN differences were

generally small (2 of 22 cases; Figure 4e, Extended Data Fig. 12, Extended Data Fig. 13).”

Line 211: “An interesting case of replicon dominance occurs when both replicons are co-dominant, resulting in an additive PCN (2 of 22 cases, Figure 4c,f, Extended Data Fig. 12, Extended Data Fig. 13).”

Line 217: “We also observed other interactions, such as incomplete dominance (4 of 22 cases), resulting in intermediate PCN, antagonism between replicons (1 of 22 cases), and high-order interactions occurring in plasmids showing more than two replicons (7 of 22 cases).”



On the same point. While I understand that some cases are very rare and not much will come out of detailing them, the results on dominance could be detailed. Why is there dominance? (my comment above on mutations in oriV is an example).

The reviewer raises an interesting question. Why do dominance relationships usually govern the PCN of multi-replicon plasmids? Multiple hypotheses could explain dominance patterns, including that one replicon is preferentially used depending on the host species, or even strain (we term this replicon preference), competition by the replication machinery between both replicons and general incompatibilities between the replicon and other general properties of the plasmid, such as the size.

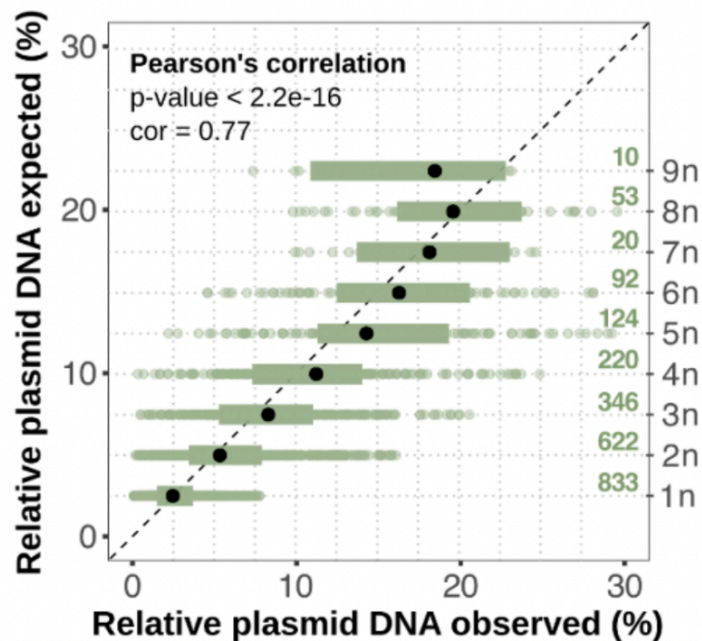
As mentioned earlier, we are working on a manuscript that uses bioinformatics and experimental methods to thoroughly address replicon dominance. This substantial undertaking warrants its dedicated manuscript, and we feel that including (some of) the findings here without experimental support would be too speculative.

Nevertheless, the manuscript delves into the possible mechanisms explaining replicon co-dominance (see Supplementary Figures 13 and 14). We included this information in the supplementary material for the sake of space. If the reviewer believes that adding these mechanistic explanations to the main text would improve the manuscript, we would be happy to do so in the next version.

- The figure 5C is a fit on medians, which vastly inflates the fit and gives a false impression the model is perfect. When it's probably not (as suggested by results in figure S15 and S16). These fits should account for the variation in the data not summary statistics.

Thank you for this observation. Initially, we chose to represent the data as single data points (median) for clarity. However, we agree that using summary statistics inflates the fit and might be confusing. To address this concern and to better represent data variability, we have changed this plot, which now displays individual data points and a measure of variation (interquartile range; see below for the new Figure 5C). We also included the number of samples in each category to improve statistical reporting, as per Nature Communications guidelines. The medians changed slightly as we revised all figures (following Reviewer's #2 comment) and noticed that the previous version of this figure inadvertently included linear plasmids.

c)



After calculating correlation statistics for all data points (instead of medians), we still observe a significant correlation, although far from perfect (Pearson product-moment correlation $r=0.77$, $p < 10^{-6}$). We believe that this change ensures that the variation in the data is clearly visualised and avoids the misleading impression of a perfect fit model. We have changed the text accordingly:

Line 310: "Remarkably, this expectation correlates well with the observed percentage of DNA content allocated to plasmids (Pearson product-moment correlation $r=0.77$, $p < 10^{-6}$; Figure 5c). Therefore, the proportion of plasmid DNA within any bacterial cell, indeed, seems to follow a discrete pattern."

- *"Distribution of total DNA load per plasmid (x- axis) relative to chromosome size per genus (y-axis). The DNA load of each plasmid is calculated by multiplying the plasmid size by the copy number and then expressed as a proportion relative to the chromosome size" Given the importance of this measure one needs a more precise description in methods with mathematical formulas to remove ambiguity.*

We appreciate the reviewer's feedback. We have now added formulas in the methods section to clarify this issue. It now reads as:

Line 441: "Plasmid DNA load (bp) was calculated as $\text{plasmid load (bp)} = \text{plasmid size (bp)} \times \text{PCN}$. The relative percentage of plasmid DNA load was calculated as $\frac{\text{plasmid load (bp)}}{\text{chromosome size (bp)}} \times 100$. Taken together, the relative percentage of plasmid DNA can also be expressed as $\frac{\text{plasmid size (bp)} \times \text{PCN}}{\text{chromosome size (bp)}} \times 100$."

- *" Our analysis revealed that, regardless of their host genus, size, or copy number, all plasmids accounted for approximately the same percentage of chromosomal DNA" This sentence needs nuance. The figure shows a variation over two orders of magnitude. Even if the averages are close the variance is large.*

The reviewer is right, and we've changed the sentence to add more nuance. The updated sentence now reads as:

Line 297: "Although there was substantial variation, our analyses revealed that, regardless of their host genus, size, or copy number, all plasmids tended to account for approximately the same percentage of chromosomal DNA."

- *" Moreover, intrinsic variability and PCN were positively and strongly correlated when we classified plasmids according to their replicon type, PTU, or plasmid cluster (Spearman's rank correlation $p < 10^{-2}$; Figure 3, Extended Data Fig. 9). " It's not completely clear to me why in the main text the coefficient of variation is different in terms of PCN but not in Extended Data Fig. 9.*

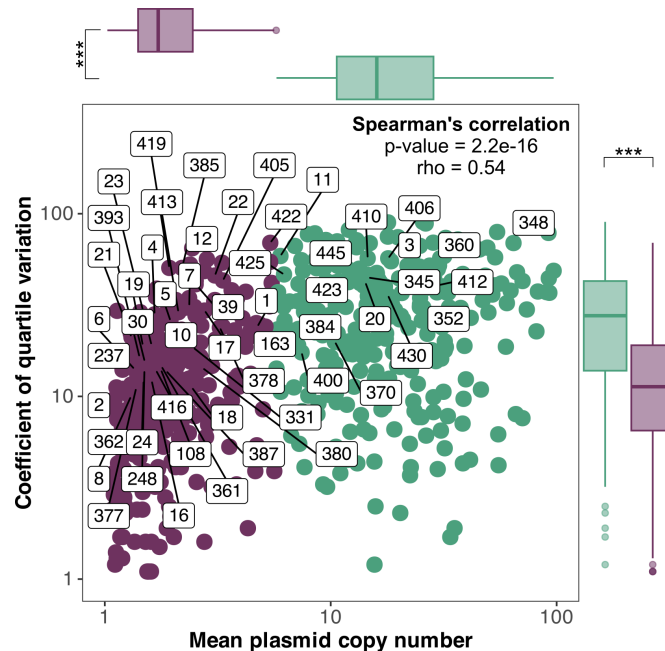
Thank you for highlighting this. Indeed, in the previous version of Figure 3, the CV for PTUs and replication types was statistically different between LCPs and HCPs. However, when grouping plasmids by cluster (old Extended Data Fig. 9), this difference was no longer significant, although the CV – PCN correlation was maintained.

Following Reviewer's #3 comment, we have changed the analyses concerning variation and PCN. In the new version of the manuscript, we no longer use CV to minimise autocorrelation (see response to Reviewer #3 on Mathematical Dependence). We now use a more robust measure of dispersion: the coefficient of quantile variation (CQV). Similar to the CV, the CQV measures dispersion and is used to make comparisons within and between data sets. The CQV is calculated as $(Q3 - Q1) / (Q3 + Q1)$, where Q1 and Q3 are respectively the first and third quartiles of each data set. Since it is based on quantile information, the CQV is less sensitive to outliers than the CV and is a more robust choice for data expressed as a ratio (<https://doi.org/10.1038/s41598-023-31711-8>).

Moreover, following Reviewer #3's comments, we have now complemented the statistical analysis of this figure to prevent type I errors. In addition to previous statistical analyses, we now perform a Bayesian multilevel model with the host of the plasmids added as a random effect. Statistically, this approach accounts for the potential shared ancestry among plasmid hosts, providing a more robust method to compare between groups.

After these changes, we found that the statistical assessment of our results remains qualitatively similar, with the only exception of the difference between HCPs and LCPs in variability (CQV) when grouped by cluster, which is now significant as estimated by Wilcoxon rank sum test and further supported by the Bayesian model.

This is now indicated in the new Extended Data Figure 9 legend as follows:



Line 76 (Extended data): “Asterisks indicate significant differences between HCP and LCP (***) for $p < 0.001$, effect size = 0.360; Wilcoxon rank sum exact test), even after accounting by host shared ancestry (as estimated by Bayesian multilevel models with host as a random effect; see Extended Dataset 6 for details)”.

- *"These results demonstrate that most plasmids encode robust intrinsic replication control mechanisms that determine their characteristic PCN." Maybe rephrase? As such, this was well known.*

We have rephrased the sentence to better highlight the novelty of our findings while avoiding redundancy with established knowledge. The revised sentence now reads as:

Line 324: “These results emphasise that intrinsic replication control mechanisms are crucial in determining each characteristic PCN, but also provide new insights into how these mechanisms differ among plasmid families.”

- Several references lack page numbers or manuscript identification.

We thank the reviewer for spotting these oversights. After thoroughly reviewing the references section, we updated references 7, 19, 44, and 51.

- Figure S12 lacks the colour code (mauve versus light orange) in the boxplots. I assume it's the same as in S11.

Thank you for noticing this error. We have added the appropriate legend to Supplementary Figure 12 to clarify the colour code. It now reads as:

Line 95: "Boxplots represent the plasmid copy number (y-axes) among different plasmid forms (x-axes): single replicons (purple) and multiple replicons (orange)."

- Gram negative and gram positive are outdated taxonomical groups. Here, it's Proteobacteria and Firmicutes (or Pseudomonadota and Bacillota).

The reviewer is right. We now refer to these taxons in both the methods (L67) and result section (L265) as Pseudomonadota and Bacillota and specify that in the context of our manuscript, we also refer to these groups as "Gram-negative" and "Gram-positive", given their historical use, that we believe still transmit better clarity to the reader.

Reviewer #1 (Remarks on code availability):

The code in the git would deserve a manual for guidance and a toy example for a first experience.

As requested, we have updated the "general workflow" and added a toy example to the repository. We agree that this might help users follow more easily.

Reviewer #2 (Remarks to the Author):

Universal rules govern plasmid copy number

Ramiro-Martínez et al. have produced a fantastic work of fundamental research towards a systematic understanding of plasmid copy number using publicly available data. In combination with Maddamsetti et al., these two articles were a pleasure to read. I did not have any major concerns with the article, with only some minor comments as below.

We thank the reviewer for our manuscript's enthusiastic assessment and comments. Please find our point-by-point replies below.

line 49: typo: "...and is mainly limited to..."

Thank you. This has now been corrected.

figure 1b: Could you clarify if "non-mobilisable" is based on the absence of any replicon machinery? Did you consider an oriT analysis?

To classify plasmid mobility, we used the *MOB_typer* function encoded in mob-suite (<https://github.com/phac-nml/mob-suite>). According to the authors of mob-suite, mobility prediction relies on identifying three key components: the relaxase, mate-pair formation, and oriT sequences. A plasmid is classified as "conjugative" if it contains both a relaxase and a mate-pair formation marker. Plasmids with either a relaxase or an oriT but lack the mate-pair formation marker are classified as "mobilizable." Those without both a relaxase and an oriT are categorised as "non-mobilizable".

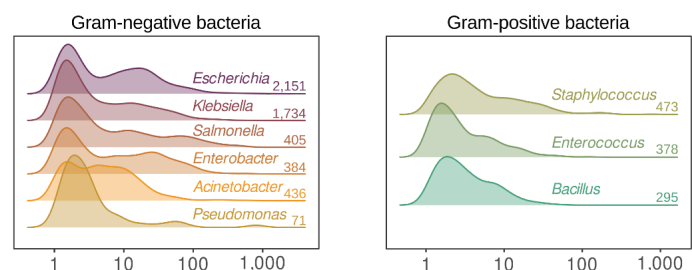
We honestly believe that additional analyses are unnecessary at this point, as plasmid mobility is not a key factor in our manuscript. We do understand, however, that MOB-typer might be overestimating the number of "non-mobilizable" plasmids, and now we include this information alongside an appropriate reference in the Methods and Discussion sections:

Line 359 (Discussion): "Fourth, our analyses rely on the accuracy of these bioinformatic tools for establishing meaningful plasmid classifications. While these methods are standard in the field, they could inadvertently introduce bias by mispredicting some plasmid properties (e.g., mobility) (53)."

Line 430 (Methods): "Finally, we also used MOB-typer (with the --multi flag) to predict plasmid mobility. We note, however, that this method likely overestimates the fraction of plasmids assigned to the "non-mobilizable" category (53, 61)."

Fig 2b: Density plots are notoriously difficult to see raw numbers - could you add the raw numbers for each category?

The number of samples was previously included in Figure 1b. However, we agree that it is important to provide clarity in Figure 2b as well. To address this, we have now added the number of samples for each genus to Figure 2b (see the figure on the right).

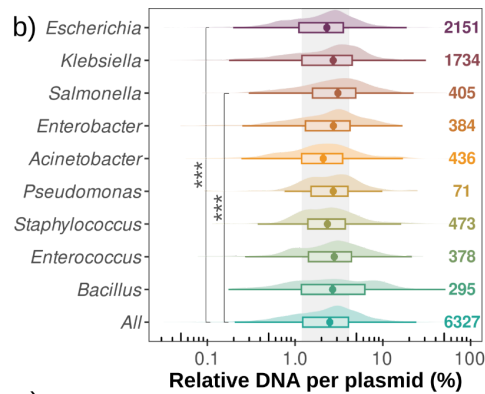


line 271: "...only Escherichia and Salmonella being different..." - I'm not sure I can see in your figure what you are referring to in terms of Escherichia and Salmonella being different, can you edit the figure or describe more in text (numerically)?

The text stated previously that "all plasmids accounted for approximately the same percentage of chromosomal DNA (median = 2.49 %, IQR: 1.22-4.06, Figure 5b, only *Escherichia* and *Salmonella* being different, albeit negligibly, from "All"; Kruskal-Wallis test followed by Dunn's test for pairwise multiple comparisons $p < 10^{-4}$).", although we acknowledge this was not clear enough. We have now added statistical information to the figure (as shown below).

Additionally, we edited the figure legend and main text to specify that this refers to statistical differences when comparing to "All" and include the effect size of the comparison (as requested by reviewer #3)

Line 279 and 300: “Only *Escherichia* and *Salmonella* significantly differ from “All”; Kruskal-Wallis test followed by Dunn's test for pairwise multiple comparisons $p < 10^{-4}$; effect size = 0.006.”



line 329: "(keeping only circular genomes)" - could you describe how circularity was determined and if you mean circularity of all plasmids/chromosomes?

Thank you. Indeed, in the previous version, the determination of topology was poorly described in the Methods section. Following this and other comments, we have now thoroughly revised and improved the description of our pipeline (see lines 383-408). In particular, plasmid topology (circular or linear) was checked using the NCBI metadata. We now specify this as follows:

Line 397: “Plasmids were identified using mob_suite (see Plasmid classification for details; n=8,660 plasmids), and their topology (circular or linear) was checked with a custom script that retrieves information from the NCBI database using its dedicated API (see Code Availability section). Plasmid contigs annotated as circular were kept for further analyses (n=8,091).”

Line 335: Was there any QC of the raw reads before mapping?

We did not apply any filtering or trimming steps to the raw reads because our primary goal was to retain as many samples as possible for downstream analyses. We found that aggressive QC measures often led to substantial losses in total read coverage and the number of viable samples that passed coverage thresholds. Instead, we indirectly accounted for read quality and coverage by setting minimum coverage requirements and calculating PCN later in the workflow.

Indeed, we removed 678 pair-end raw reads and their corresponding assemblies from the analyses because the alignments did not meet the criteria of CoverM. This is now indicated in the appropriate methods section as follows:

Line 396: “Some of the alignments did not meet the quality criteria of CoverM and were excluded from further analyses (n = 678)”

Furthermore, to ensure a more accurate and unbiased estimation of coverage for PCN calculation, we employed the trimmed mean option in CoverM. With this option, CoverM discards the 5% of positions with higher and lower coverage, respectively. Trimming helps to

eliminate potential outlier effects—such as sporadic alignments or exceptionally high coverage peaks that may stem from sequencing or mapping artefacts—thereby generating a more robust and representative coverage measure. This strategy ultimately mitigates biases in coverage estimation and leads to a more reliable determination of PCN across the dataset.

Line 346: You state how many plasmids there were at the end of filtering, but I could not find how many plasmids there were initially. Could you add how many were removed because of <30x or <1 PCN?

As stated above, we have completely rewritten this section. We now indicate how many samples were removed at each data processing step. Particularly, 736 plasmids were excluded because of coverage <30x and 1,000 because of PCN <1. We also excluded 28 extremely small plasmids (< 1kb). This section now reads as follows:

Line 403: “We removed plasmids belonging to assemblies with an absolute sequencing depth below 30x (n = 736) and plasmids showing a size < 1 kb (n = 28) or PCN <1 (n = 1,000).”

Did you consider the presence of phage-plasmids, and how these might differ in PCN vs. plasmids? If not, please consider adding this as a future direction/limitation.

Our main aim was to study PCN, so we did not specifically delve into phage-plasmids. Currently, tools for classifying elements as phage-plasmids are not as straightforward as tools for plasmid typing. Some groups of phage-plasmids have recently been defined, but these groups do not necessarily associate with a single specific replicon (doi: 10.1093/nar/gkab064). Thus, classifying phage-plasmids solely based on replicon typing does not seem appropriate. For this reason, we refrain from addressing phage-plasmids in this work.

Nonetheless, to address the Reviewer’s curiosity, we checked plasmids with replicon type IncY, as some P1-like phage-plasmids seem to associate with this replicon type (doi: 10.1093/nar/gkab064, Figure 4 therein). In our dataset, single replicon IncY (phage-)plasmids display a median PCN = 1.26 (Supplementary dataset 2) and vary between ~83 and ~114 kb in size (Supplementary dataset 1). Thus, these potential phage-plasmids display the size/PCN trade-off discussed, and their median relative DNA load converges towards 2.5% (IQR: 2.09 - 2.77). When these elements carry further replicons besides the IncY, their PCN tends to increase to ~3, thus displaying a sign of replicon dominance (Figure 4b, Extended Data Figure 11). Based on this anecdotal evidence, we have no reason to assume phage-plasmids may behave differently than “regular” plasmids.

Beyond phage-plasmids, it would be interesting to analyse how extrachromosomal elements in general (including phage-plasmids, linear plasmids, or even secondary chromosomes) behave regarding copy number. This would be, however, the topic for future work. We address this in the new discussion paragraph, which now reads as follows:

Line 366: “Finally, our analysis is restricted to the classical definition of plasmids (independently replicating circular DNA molecules). Yet, not all plasmids are circular, and many extrachromosomal genetic elements share properties with plasmids (e.g., phage-plasmids or secondary chromosomes)(55). In this sense, our study lays the foundation for future works addressing copy number variation in extrachromosomal genetic elements of non-model microorganisms. By revealing a traditionally neglected aspect of their biology, these studies will shed light on the complex interplay among different genetic elements and their bacterial hosts.”

Line 355 (onwards): please add versions of tools (and DBs where relevant) including parameters used. The Github docs should also have which versions of tools the scripts have been tested.

We thank the reviewer for spotting this oversight. We have updated both the main text and the GitHub repository to show versions of all tools and databases used in our analysis.

378-380: "The relative percentage of plasmid DNA load was calculated as (plasmid size (bp) / chromosome size (bp) x 100)" I just wanted to check this was correct, based on the previous sentence where plasmid DNA load is plasmid size x PCN. So should the relative percentage of plasmid DNA load be plasmid load (size x PCN) / chromosome size?

Yes, the reviewer is correct. We have amended this section to include formulas, as requested by Reviewer #1 and now it reads as follows:

Line 441: “Plasmid DNA load (bp) was calculated as $plasmid\ load\ (bp) = plasmid\ size\ (bp) \times PCN$. The relative percentage of plasmid DNA load was calculated as $\frac{plasmid\ load\ (bp)}{chromosome\ size\ (bp)} \times 100$. Taken together, the relative percentage of plasmid DNA can also be expressed as $\frac{plasmid\ size\ (bp) \times PCN}{chromosome\ size\ (bp)} \times 100$.”

Reviewer #2 (Remarks on code availability):

The github repo is well-written and clear in its description of methods. My only minor comment would be to add versions of third-party tools used as a record of compatibility if tools are updated in future.

We have now completed the information and updated the GitHub repository with the version number of the tools/databases that were missing.

Reviewer #3 (Remarks to the Author):

The authors need to take into account the following major points:

We thank the reviewer for their helpful comments. Please find our point-by-point replies below.

#) Mathematical Dependence

Since the coefficient of variation (CV) is calculated by dividing the standard deviation by the mean, the mean and CV are mathematically dependent. Therefore, when interpreting the correlation between the mean and CV, it is important to carefully consider that this may not reflect a biological relationship, but rather result from mathematical dependence.

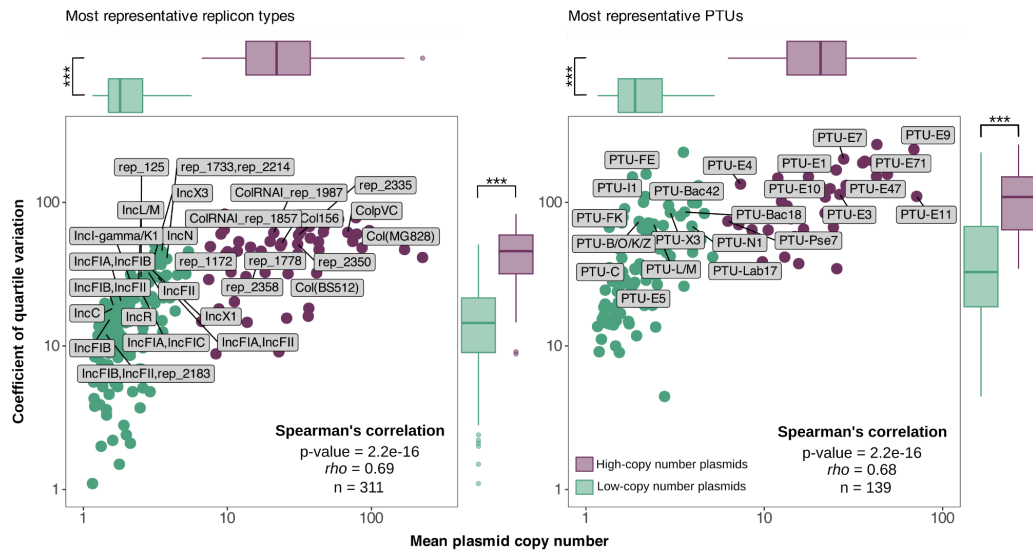
The purpose of the coefficient of variation (CV) is to normalise the standard deviation (SD) into a dimensionless measure (typically expressed as a percentage) that is appropriate for comparing variation when the difference between means is large. For example, a two-copy departure from the mean PCN entails a considerable difference for an IncF plasmid (~1 copy) but only a slight difference for a ColE1 plasmid (~20 copies). The CV allows us to compare these differences unbiasedly by normalising the variation.

However, we agree with the reviewer that CV and mean are mathematically dependent variables (the CV is calculated as the SD divided by the mean). To reduce mathematical dependence, we repeated the analyses using a more robust measure of dispersion: The coefficient of quantile variation (CQV). Similar to the CV, the CQV measures dispersion and is used to make comparisons within and between data sets. The CQV is less sensitive to outliers and is a more robust choice for data expressed as a ratio than the CV (<https://doi.org/10.1038/s41598-023-31711-8>). The CQV is calculated as $(Q3 - Q1) / (Q3 + Q1)$, where Q1 and Q3 are respectively the first and third quartiles of each data set. Since it is based on quantile information, the CQV and the mean PCN are no longer expected to be mathematically dependent.

Therefore, in the new version of the manuscript, instead of the CV, we exclusively use the CQV. This change does not lead to any major changes in our conclusions, and the new Figure 3 looks highly similar to the original Figure 3 (see below). We also changed the corresponding Extended data Figures (S8, S9 and S10), results, figure legends, and methods sections to reflect this change:

Line 172 (Results): “[...] HCPs showed significantly greater variability in their PCN than LCPs (measured as coefficient of quartile variation – CQV (31))”

Line 437 (Methods): “To measure PCN variation across our dataset, we calculated the quartile coefficients of dispersion (CQV). The CQV allows for robustly comparing the degree of variation from one plasmid group to another, even if the PCNs are drastically different (31). CQV was calculated with the R package [cvcqv](#)⁶⁴ v1.0.1.”



#) Non-Independence of Data

In analyses involving multiple plasmids, the assumption that data points are independent is made. However, this assumption could lead to misleading results. Organisms (and replicons such as plasmids) do not evolve independently; they share traits due to common ancestry. Thus, plasmids cannot be treated as independent data points. This issue is similar to the problem of pseudoreplication in experimental research, which can introduce significant bias (type I errors) in statistical analyses.

For example, carefully reconsider the following analysis in the manuscript:

Figure 3. Intrinsic PCN variability. (***: $p < 10^{-9}$, Wilcoxon rank sum exact test).

Here are quotes from a reference paper:

<https://pmc.ncbi.nlm.nih.gov/articles/PMC7612097/>

Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range

Importance of controlling for non-independence of genomes

The reviewer raises an interesting point. We agree that plasmids do not evolve independently, which can introduce biases such as pseudoreplication that inflate type I error rates. As suggested in the Reviewer's reference paper (and references therein), a possibility to account for phylogenetic pseudoreplication is to apply a mixed effect model that controls whether phylogenetic relationships (shared ancestry) confound the patterns across species. To do that, it would be necessary to establish a phylogenetic relationship between plasmids belonging to different groups and genera. However, constructing accurate plasmid phylogenies is challenging because of their high horizontal gene transfer and recombination

rates. Indeed, establishing a plasmid “core genome” is almost impossible because many plasmid groups do not share any genetic content.

Owing to the difficulty of determining plasmid phylogenies, in the previous manuscript version, we took the alternative approach of classifying plasmids into related groups that share a given property (e.g., PTUs, rep types, and clusters). Then, statistically, each plasmid group is accounted for only once, thereby reducing redundancy and minimising the influence of shared ancestry on our statistical results. This was the case for Figure 3, where each data point represented each plasmid group's mean PCN and CV (now CQV).

However, to ensure that shared ancestry does not significantly affect our statistical analyses and conclusions, we used an approach similar to that in the Reviewer's suggested paper. Briefly, we have performed Bayesian multilevel models, adding the hosts of each plasmid group (Replicon type, PTU, cluster) as random effects. Accounting for host-shared ancestry did not affect our conclusions, as we found the model substantially supported our previous findings in Figure 3, Extended Data Figure 8 and Extended Data Figure 9. As the Wilcoxon rank-sum exact test results and the Bayesian models largely agree, we have decided to keep both in the manuscript, as we believe it further reinforces our conclusions. We report the results of the Bayesian models in the new **Extended Data 6** and cite them in the appropriate figure legends and text.

Moreover, to ensure that statistical errors are minimised throughout the manuscript and in line with the Reviewer's comment on type I errors, we have included effect sizes in all pertinent analyses (Figure 2c, Extended Data Figure 8, Supplementary Figure 9, Figure 5b). Effect sizes quantify how substantial the observed effects are, offering context beyond just p-values and aiding in evaluating the practical significance of our results.

All in all, the analysis of effect sizes supported our interpretation of results. Indeed, Figure 2c demonstrated a moderate effect size (Cohen's $h = 0.1$), indicating that the observed differences are meaningful and that the associated p-values support our conclusions appropriately. In contrast, Figure 5b showed a very small effect size ($\eta^2 = 0.006$), suggesting that while the p-value is statistically significant, the actual difference is negligible and likely lacks biological importance, as stated in the manuscript. Furthermore, Figures 3a, 3b and Extended Data Figure 9 presented strong effect sizes ($h \geq 0.836$, $h \geq 0.512$, and $h \geq 0.360$, respectively) based on the Wilcoxon rank-sum test, highlighting substantial and significant differences that robustly reinforce our findings.

#) Validating the results

It seems that the authors merely ran bioinformatics tools for plasmid analysis without carefully validating the results. For example, for plasmid sequences assigned as 'rep_type.s.' to 'IncP' by tools like 'mob_typer' or 'plasmidfinder', there are sequences where the 'predicted_mobility' is assigned as 'mobilizable' or 'non-mobilizable' instead of 'conjugative', and sequences where the 'Relaxase' is assigned to 'MOBQ' or 'MOBH' instead of 'MOBP'.

<https://github.com/PaulaRamiro/NpAUREO/blob/main/GeneralWorkflow.md>

The reviewer makes a valid point. We acknowledge that bioinformatics tools like `mob_typer` and `plasmidfinder` may occasionally misclassify relaxase types or mobility categories, such as assigning some 'IncP' plasmids as 'mobilizable' or incorrectly

categorizing relaxases as 'MOBQ' or 'MOBH' instead of 'MOBP'. This is also related to Reviewer's #2 point about oriT detection. We now acknowledge in the discussion section that our analyses are limited by the tools we used and that it is impossible to avoid a certain degree of misclassification.

Line 357: "Third, our results derive only from a few bacterial taxa, primarily genera of clinical importance. To some extent, this is an unavoidable consequence of the lack of appropriate tools for plasmid classification (e.g., replicon type, PTU) beyond well-studied bacterial genera. Fourth, our analyses rely on the accuracy of these bioinformatic tools for establishing meaningful plasmid classifications. While these methods are standard in the field, they could inadvertently introduce bias by mispredicting some plasmid properties (e.g., mobility) (54). Although these factors probably account for some of the observed variability in PCN, they are unlikely to significantly impact our general conclusions, founded upon an analysis of thousands of diverse bacterial plasmids."

Due to the rare occurrence of errors, manually correcting them is impractical and unlikely to significantly affect results. Therefore, we chose automated classifications, confident that few misassignments won't change the study's conclusions.

#) DNA sequencing data

The authors stated, "we leveraged extensive DNA sequencing data to comprehensively analyze the copy number of thousands of diverse bacterial plasmids."

but if they are not certain about how these DNA sequencing data were obtained (e.g., bacterial growth conditions and DNA extraction methods used), the reliability of the results may be questionable.

The authors used raw reads and assemblies downloaded from the NCBI SRA database, but they should have considered (or at least discussed) this point.

Here are quotes from a reference paper:

<https://pmc.ncbi.nlm.nih.gov/articles/PMC6312765/>

Reconsidering plasmid maintenance factors for computational plasmid design

Plasmid size may be associated with copy number.

Plasmid copy number estimates can vary, depending on bacterial growth conditions and DNA extraction methods used [97,99]. Therefore, copy number data should be interpreted carefully.

Thank you. The reviewer raises a valid point. We have now added a new paragraph that enriches our discussion, where we cite the references provided. In a new discussion

paragraph, we address (technical) limitations when analysing PCN from sequencing data. It reads as:

Line 351: “Our study is not without limitations. First, PCN estimation might be subject to a certain degree of noise. PCN is an inherently plastic trait and may vary at different points of the host cell cycle or depending on growth conditions⁵⁰. We calculated PCN from deposited sequencing data and cannot exclude that some experimental factors (e.g., sequencing technology, DNA extraction protocol) may affect PCN determination (51-53).”

Here are some minor points:

#) Supplementary Dataset 1

558670_0_supp_4562777_skztgk.xlsx

Please clarify the meaning of the column labels.

Thank you for pointing this out. We have clarified the names of the columns and added a new sheet to the file explaining the meaning of each variable. Specifically, the description sheet “Dataset variables” reads as (Supplementary Dataset 1):

Variable	Description
Biosample	Accession number of the genomic assembly from which the plasmid data originates.
Contig	Specific sequence identifier for the plasmid or genomic contig.
Genus	Genus of the host organism of the plasmid.
Species	Species of the host organism of the plasmid.
Predicted_mobility	Classification of plasmid mobility (e.g., conjugative, mobilizable, non-mobilizable) as predicted by MOB_typer.
Size	Total size of the plasmid in base pairs
Relaxase	Type of relaxase gene (e.g., MOB _P , MOB _F), if present, or "No MOB" if absent, as predicted by MOB_typer.
PCN	Average number of plasmid copies per cell.
Percentage_DNA_load	Proportion of plasmid DNA (size × plasmid copy number) relative to the chromosome size (%).
Total_DNA_load	Total DNA contributed by the plasmid (size × plasmid copy number).
Replicon_type	Identified replicon types (e.g., IncFIA, IncFIB), extracted with MOB_typer.
Replicon_form	Indicates whether the plasmid has single or multiple replicons.
Chromosome_size	Total size of the bacterial chromosome in base pairs.
GC	GC content percentage of the plasmid.
Plasmids_per_cell	Total number of plasmid copies present in the cell.
Topology	Structure of the plasmid (e.g., circular).
Predicted_PTU	Predicted Plasmid Taxonomic Unit (PTU) classification, if applicable.

#) *natural plasmids*

The authors should clearly specify the method used to obtain only "natural plasmids" (i.e., to exclude artificial sequences) from the database.

The reviewer raises a valid point. However, distinguishing genuinely "natural" plasmids from synthetic ones can be difficult, if not impossible. In our original text, we used the term "natural" to refer to the plasmids in our database, but we recognise that some may contain synthetic elements.

To obtain an approximate assessment of the presence of synthetic plasmids within our collection, we conducted a BLAST search for 14 commonly used artificial plasmids (pUC19, pSEVA111, pSEVA141, pSEVA142, pSEVA131, pSEVA151, pBAD18, pBlueScript, pMAL, pJET1, pGEX, pGEM-T, pET28, pBR322) against the 6,327 plasmids present in our database. We classified synthetic plasmids as those in our database that contained at least 50% of one or more of the aforementioned artificial plasmids.

Following this criteria, we found only 11 synthetic plasmids (0.17%) in our database. Although this approximation may not be exhaustive and likely underrepresents the actual number of synthetic plasmids in our database, we feel it provides a basic estimate and suggests that synthetic plasmids are generally present at a low frequency. Nevertheless, we now acknowledge that our collection might include synthetic plasmids in the limitations paragraph of the discussion section:

Line 355 (Discussion): "Second, some plasmids in our database may be synthetic, and consequently, they might have been engineered to display an artificially high (or low) PCN."

We also revised the usage of the term "natural" throughout the manuscript and removed it where it referred to plasmids in our database (strikethrough text has been removed from the new version):

Line 16: "We found that the copy number of ~~natural~~ plasmids is highly variable, spanning nearly three orders of magnitude, and that it is intrinsically robust against changes in genomic context."

Line 52: "In this work, we developed a custom bioinformatic pipeline that leverages extensive DNA sequencing data from different studies to calculate PCN for 6,327 ~~naturally occurring~~, phylogenetically diverse plasmids."

Line 182: "In contrast to previous observations restricted to model laboratory plasmids (30, 31), these results indicate that, ~~for natural plasmids~~, the higher the PCN, the more relaxed the control of replication and segregation."

Line 72: “This approach gave rise to a dataset that comprises plasmids belonging to 139 PTUs, 238 distinct replicon types, and 2,200 sequence-based clusters, indicating that it captures a significant fraction of the extant ~~natural~~ plasmid diversity”

Line 184: “This result underscores the role of HCPs as plastic adaptive platforms ~~in natural settings.~~”

Line 326: “The stringency of PCN control is, however, different between plasmid lifestyles: ~~and naturally occurring~~ for instance, HCPs show more variation in PCN than LCPs”

#) *phylogenetic boundaries.*

The authors should provide a clear definition of "phylogenetic boundaries" if they claim that such boundaries exist for organisms that have diverged from a common ancestor and thus form a continuum.

We agree that, in the previous version of our manuscript, the term “phylogenetic boundary” was somewhat ambiguous. We intended to stress that our results hold across various bacterial species. To avoid confusion, we have now replaced the term “phylogenetic boundaries” with “across bacterial species” in the discussion and the abstract sections:

Line 20 (Abstract): “Finally, we uncover a universal scaling law that links copy number and plasmid size across bacterial species, indicating that pervasive constraints orchestrate the PCN-size trade-off”

Line 335 (Discussion): “Arguably, the most intriguing result of our work is that a PCN-size scaling law governs plasmid biology across bacterial species”

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I thank the reviewers for having very seriously tackled all my comments. The answers and actions were appropriate. The only major comment that wasn't really tackled regards the oriV evolution issue, but since they are preparing one paper on the topic, I'll wait for that one.

We thank the reviewer for their positive evaluation of our work.

Reviewer #1 (Remarks on code availability):

While the authors have added a toy example, I still think the Git should have more instructions for the use of the software, a FAQ, etc.

Thank you for your feedback. However, the objective of our work is not to present a new tool or software; rather, we utilise existing tools to derive our analyses. Our aim is not to introduce a tool for calculating copy number but to build upon the available software. Thus, we have provided a toy example to illustrate the use of our code as requested, and we encourage users to refer to the original documentation for further guidance.

Reviewer #2 (Remarks to the Author):

All previous comments have been addressed adequately. This is a wonderful piece of work and I look forward to seeing it published.

Thank you for your kind words and support. We truly appreciate your feedback.

Reviewer #3 (Remarks to the Author):

<https://github.com/PaulaRamiro/NpAUREO/blob/main/Example.md>

Please verify the usability and reproducibility of the code in different computer environments.

Reviewer #3 (Remarks on code availability):

git clone <https://github.com/PaulaRamiro/NpAUREO/>

<https://github.com/PaulaRamiro/NpAUREO/blob/main/Example.md>

Following the instructions in Example.md, the commands and output examples executed are as follows.

```
+ # bash #  
wget  
https://ftp.ncbi.nlm.nih.gov/genomes/genbank/assembly\_summary\_genbank.txt  
grep "Staphylococcus" assembly_summary_genbank.txt | grep "Complete  
Genome" > data_staphylococcus.txt
```

```
grep "GCA_001018645" data_staphylococcus.txt > staphylococcus_example.txt
```

```
esearch -db assembly -query SAMN03255442 | elink -target nucleotide -name  
assembly_nucline_refseq | efetch -format fasta > SAMN03255442.fasta
```

resulted in an empty file (file size 0).

Here is the output:

```
$esearch -db assembly -query SAMN03255442 | elink -target nucleotide -name  
assembly_nucline_refseq | efetch -format fasta  
$esearch -db assembly -query SAMN03255442 | elink -target nucleotide -name  
assembly_nucline_refseq  
<ENTREZ_DIRECT>  
<Db>nucleotide</Db>  
<Count>0</Count>  
<Step>3</Step>  
<Elapsed>2</Elapsed>  
</ENTREZ_DIRECT>  
$esearch -db assembly -query SAMN03255442  
<ENTREZ_DIRECT>  
<Db>assembly</Db>  
<WebEnv>MCID_67af67d482b9378dba0e079a</WebEnv>  
<QueryKey>1</QueryKey>  
<Count>2</Count>  
<Step>1</Step>  
<Elapsed>1</Elapsed>  
</ENTREZ_DIRECT>
```

```
+ # bash #
```

```
esearch -db sra -query SAMN03255442 | efetch -format runinfo >  
SAMN03255442.numbers
```

```
grep "ILLUMINA" SAMN03255442.numbers | grep "PAIRED" >  
filtered_SAMN03255442.numbers
```

```
+ # bash #
```

```
fasterq-dump --split-3 SRR1955495
```

The above command output the following message.

```
2025-02-14T16:53:38 fasterq-dump.3.2.0 err: cannot find 'SRR1955495' ->
RC(rcVFS,rcQuery,rcExecuting,rcItem,rcNotFound)
fasterq-dump quit with error code 3
```

```
$du -sh *
18M NpAUREO
0B SAMN03255442
0B SAMN03255442.fasta
4.0K SAMN03255442.numbers
1.3G assembly_summary_genbank.txt
1.7M data_staphylococcus.txt
4.0K filtered_SAMN03255442.numbers
4.0K staphylococcus_example.txt
```

```
$uname -a
Darwin MacBook-Pro-14-inch-2021.local 21.2.0 Darwin Kernel Version 21.2.0:
Sun      Nov      28      20:28:41      PST      2021;
root:xnu-8019.61.5~1/RELEASE_ARM64_T6000 arm64
```

Thank you for your feedback; it has significantly improved our manuscript. While our code was thoroughly tested on Linux, we acknowledge that it was not tested on macOS. We have now addressed this issue and incorporated the necessary commands to ensure compatibility and reproducibility on both Linux and macOS systems. Specifically, the problem occurred because some Linux commands do not work natively on macOS (e.g., `wget` versus `curl`). We now provide specific code when necessary for both Mac and Linux users to address these compatibility issues.