Is the distribution of plasmid lengths bimodal?

Ian Dewan[1 2] and Hildegard Uecker[3]

Research group Stochastic Evolutionary Dynamics, Department of Theoretical Biology, Max Planck Institute for Evolutionary Biology, Plön, Germany

[1]ORCiD: 0000-0001-8964-3800
[2]Corresponding author: dewan@evolbio.mpg.de
[3]ORCiD: 0000-0001-9435-2813

Abstract

The length of a plasmid is a key property which contributes to many aspects of plasmid biology. When distributions of plasmid lengths are shown in the literature, they are usually plotted with length on a logarithmic scale. However, a quantity and its logarithm have distinct distributions which may differ considerably in shape. Mistaking the distribution of log-lengths for the distribution of lengths can therefore lead to distorted conclusions about the distribution; in particular, the distribution of log-lengths may be bimodal when the distribution of lengths is only unimodal. This particular confusion has arisen in the literature where a bimodal log-length distribution is suggested to support a distinction between large and small classes of plasmids. While the length distribution is indeed bimodal within many bacterial families, it is not across the ensemble of all plasmids. We suggest that authors should be careful to show the plasmid length distribution, or to distinguish the two distributions, to avoid misleading inferences.

Keywords: plasmids, logarithmic scale, density estimation

Highlights:

- The distributions of lengths and log-lengths of plasmids are different, and have considerably different shapes.
- The typical practice of using a logarithmic scale for plasmid lengths leads to confusion between the two distributions.
- In particular, the distribution of log-lengths can be bimodal when the distribution of lengths is not.
- The length distribution within bacterial families is often bimodal, but across all plasmids, it is unimodal.
- Clearly distinguishing between the distributions will ensure that biological conclusions drawn from them are robust.

1 Introduction

Bacterial plasmids have a wide variety of sizes, from tiny cryptic plasmids less than a kilobase long, to multiple-megabase megaplasmids that rival the size of chromosomes (Ciok et al. 2016; Hall et al. 2021). The length is closely linked to other aspects of a plasmid's biology, since smaller plasmids contain fewer genes and may therefore be limited in their functions (plasmid-specific or related to the host phenotype), while very large plasmids may contain a wide variety of accessory genes.

It has generally been assumed in the literature that plasmids can be broadly divided into two groups on the basis of length, and that this division corresponds to a biologically meaningful distinction between two plasmid lifestyles. The shorter group, with a modal length less than 10 kbp (Hall et al. 2021), is said to comprise plasmids with high copy numbers which are mostly mobilizable or nontransmissible, while the longer group, with a modal length around 100 kbp, comprises low copy number plasmids which are more likely to be conjugative. Part of the evidence for the existence of two length-based groups is the distribution of plasmid lengths estimated using standard techniques (kernel density estimation or histograms): the resulting density estimates clearly have a bimodal shape (this is done for all plasmids together by Smillie et al. 2010; Garcillán-Barcia, Redondo-Salvo, and Cruz 2023 and for plasmids divided by host family by Hall et al. 2021; Rodríguez-Beltrán et al. 2021).

However, these density estimates are usually plotted with a logarithmic scale on the x-axis (plasmid length): when we plotted plasmid length data (pooled across bacterial families) on a linear scale (Dewan and Uecker 2023), the bimodal distribution disappeared (compare panels A and C of Figure 1). Choosing a log scale, which is a natural thing to do given that the lengths of plasmids extend over four orders of magnitude, has distorted the estimated distribution of plasmid lengths. In this letter, we explain the reason why this happens, and what implications, if any, it has for our understanding of plasmids.

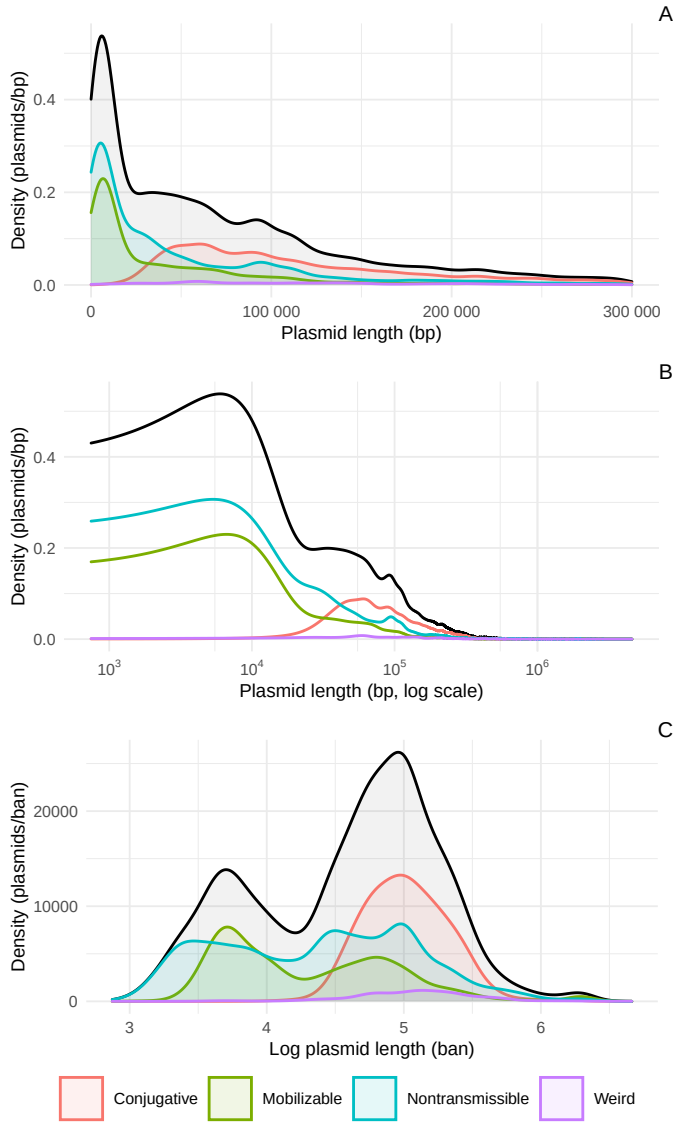

Figure 1: The distributions of plasmid lengths and log-lengths (in black: all plasmids; in colour: plasmids separated by mobility). Plasmid length data is taken from PLSDB (Schmartz et al. 2021, retrieved 1 March 2023), and the estimate made by kernel density estimation in R (R Core Team 2020). Panel A shows the estimated distribution of plasmid lengths (on a linear scale). Only plasmids of length less than 300 000 bp are shown for reasons of scale: 1844 plasmids in the database (5.34 % of the total) are longer than the limit; the longest plasmid in the database is 4 605 385 bp. Based on Figure 1 of Dewan and Uecker (2023). Panel B shows the same distribution as panel A, but with plasmid length on a log scale. Panel C shows the estimated distribution of plasmid log-lengths ("ban" are the units of the base-10 logarithm). To classify the plasmids by mobility, the annotations in PLSDB from MOB-typer (Robertson and Nash 2018), indicating putative relaxases or conjugative genes, have been used. Those plasmids with neither are nontransmissible, those with both conjugative, and those with only the relaxase mobilizable. The "weird" class includes those plasmids which had conjugative genes but no relaxase: these would be nontransmissible, but nonetheless have all the genes for a secretion system and mating pair formation. It is possible that these are misidentified conjugative plasmids, which have an unknown relaxase, or misidentified mobilizable or nontransmissible plasmids, which do not actually have conjugative genes.

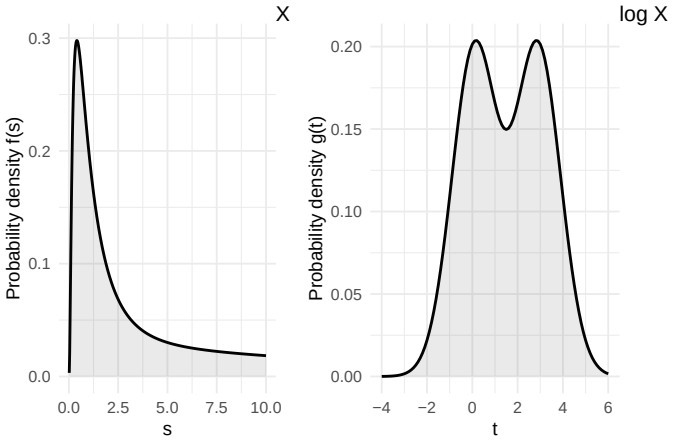

Figure 2: The probability density function of a distribution (left) and its the (natural) logarithm (right). The distribution of $\log X$ is an equal mixture of two normal distributions, with means 0.1 and 2.9 and standard deviation 1; the distribution of X is therefore a mixture of two log-normal distributions.

2 Analysis and Results

The distribution of log-lengths is not the distribution of lengths

What has gone wrong with figures showing the plasmid length distribution in the literature? They are in fact showing *not* the distribution of lengths but the distribution of log-lengths, and these two distributions may have quite different shapes; in general, the shape of a distribution changes, possibly drastically, when a transformation is applied to it. Intuitively, bins of equal sizes on the log scale correspond to bins of different sizes on the linear scale: a bin at small plasmid log-lengths contains a much narrower range of plasmid lengths than a bin at large plasmid log-lengths, and the latter might contain more sequences than the former for this reason alone. This might create a second mode even if the distribution is unimodal on a linear scale. This is illustrated in Figure 2 by a hypothetical example of a distribution of a random variable and the corresponding distribution of its logarithm. The distribution of the variable on the linear scale is unimodal, but the distribution of its logarithm is bimodal, precisely what happens with plasmid lengths.

We can show the change in the shape of the distribution formally by exam-

ining the effect of a log transformation on the density function of a distribution. Suppose we have a (positive) random variable X with a probability density function f. The defining characteristic of the probability density function is that it can be integrated over an interval to find the probability that X lies in that interval, or

$$P(a \leq X \leq b) = \int_a^b f(s)\, ds.$$

If we know f, we can find the probability density function of $\log X$:

$$\begin{aligned} P(a \leq \log X \leq b) &= P(e^a \leq X \leq e^b) \\ &= \int_{e^a}^{e^b} f(s)\, ds \\ &= \int_a^b \underbrace{f(e^t)e^t}_{=:g(t)}\, dt, \end{aligned}$$

so the probability density function of $\log X$ is $g(t) = f(e^t)e^t$. (Here we have used the natural logarithm, but the same argument applies to the "common" or base-10 logarithm usually used in log scales.) When we apply the change of variable $s = e^t$ in the above integral to pass from a variable s representing the value of X to a variable t representing the value of its logarithm, we f by a nonconstant—indeed increasing—function e^t, changing the shape of the function. If the probability density function f decreases too slowly (i.e. if it has a heavy tail), this creates a second mode.

The difference in the number of modes between Figure 1A and 1C is therefore simply a mathematical consequence of the log scale. Changes in shape of course also occur in distributions of subsets of plasmids as well, such as the distributions within mobility classes shown in Figure 1. Note, for example, that looking at the log-length distributions in Figure 1C would suggest that mobilizable plasmids have a bimodal distribution and nontransmissible plasmids are roughly equally abundant over a wide range of lengths; but the length distributions in Figure 1A show that both groups have a unimodal distribution with a long right tail.

To verify that these changes in distribution shape are indeed caused by the log transformation and not instead artifacts of the mechanism of density estimation, we have confirmed that the general shape of the length and log-length distributions is consistent under different estimation methods: different bandwidth selections for kernel density estimation, the empirical cumulative distribution function, and equal-area histograms. These do indeed all give similar results: see the supplement for details.

Within bacterial families, the plasmid length distribution is often bimodal, but the modes and the valley are incorrectly predicted by the log-length distribution

So far, we have considered the plasmid length distribution for all plasmids together, irrespective of their bacterial hosts. However, the size of plasmids differs between groups of bacteria, for example because bacteria with larger genome sizes in general have larger plasmids (Slater et al. 2008; Smillie et al. 2010). If the length distribution is really bimodal within groups, but the modes (and the antimode, or local minimum density, between them) have different locations in different groups because of a general difference in size, then the pooled distribution for all plasmids could be unimodal because modes and antimodes of different groups obscure each other. Looking at the length distribution by family, as shown in Figure 3, shows that some bacterial families indeed show a pronounced bimodal distribution on the linear scale, while for others there is no clear second mode. A caveat to the within family density estimates is that they are based on much fewer sequences than the estimate for all plasmids.

Nevertheless, even if both distributions have the same number of modes, looking at the log-length distribution can be misleading, since taking logarithms can change not only the number but also the location of modes (and of antimodes). This shift in the mode can, for example, be seen for conjugative plasmids in Figure 1. The modal log-length of a conjugative plasmid appears to be about 5 (corresponding to a length of 10^5 bp), while the modal length is smaller

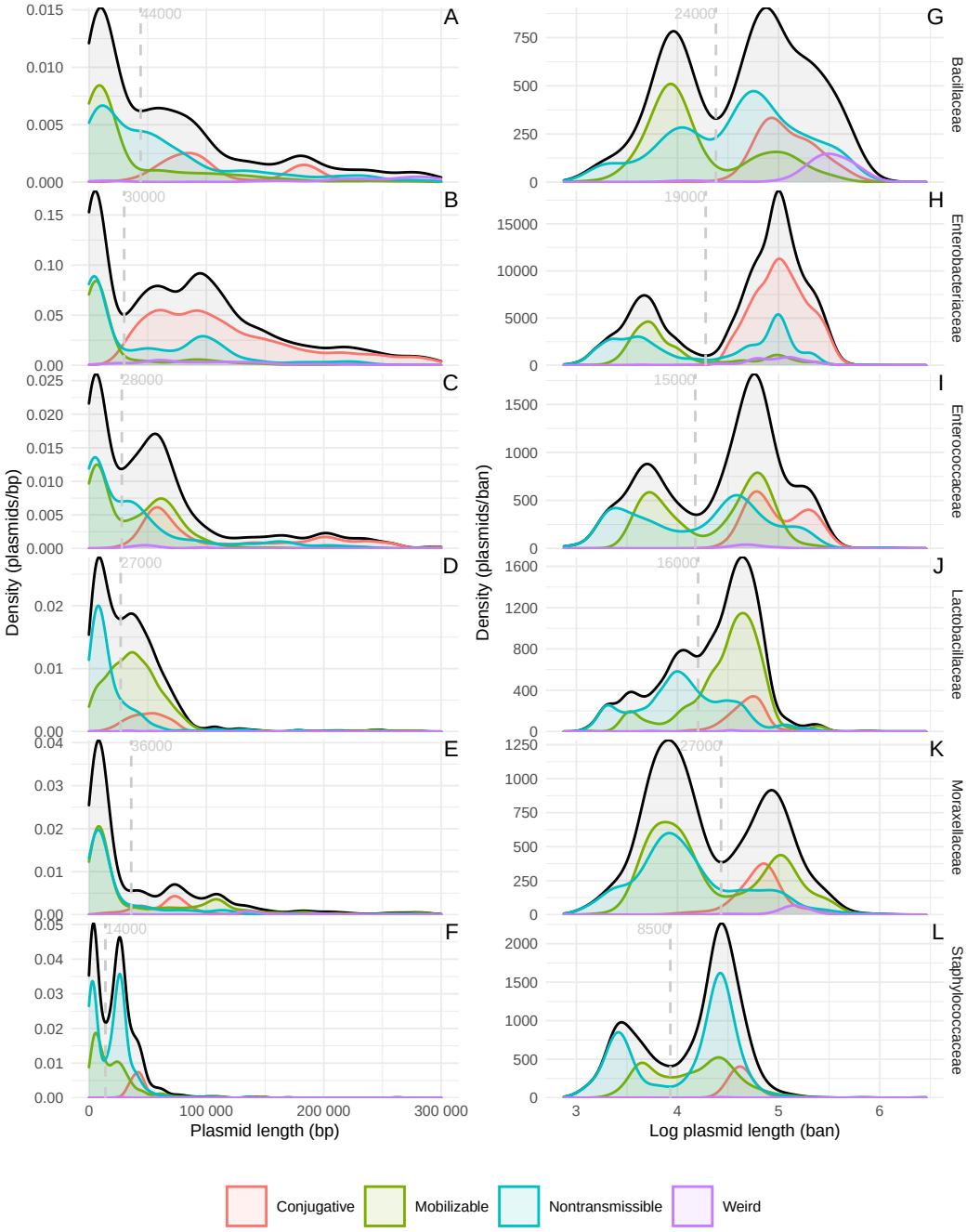

Figure 3: Distributions of plasmid length (left column) and log-length (right column) for plasmids from the six bacterial families with at least 1000 sequences in the database. Distributions for all plasmids shown in black, and for individual mobility classes in colour. Note that the length distributions are cut off at 300 000 bp for reasons of scale. The vertical lines show the location of the antimode of each distribution to the nearest kilobase. "Ban" are the units of the base-10 logarithm. For the data source and details on the figure construction see Figure 1.

(closer to 50 000 bp). Similarly, within bacterial families the antimode dividing small and large plasmids is not the same for the length and the log-length distributions (Figure 3). Thus even if there are distinct groups of "large" and "small" plasmids, trying to read their typical sizes or the dividing line between them off of an estimate of the distribution of log-lengths will lead to errors.

3 Discussion

We have shown that estimates of the distribution of plasmid lengths from available plasmid sequence data are not bimodal, at the global level at least, even though estimates of the distribution of log-lengths are, and that this kind of discrepancy is an ordinary feature of probability distributions without any biological meaning. Within bacterial families, such a bimodal length distribution often exists, but the modes and antimode are incorrectly predicted by the log-length distribution. It is important to note that this does not mean the distribution of log-lengths is wrong or not really bimodal. The log-lengths of plasmids do have a distribution, it is perfectly mathematically possible to estimate and plot it, and, apparently, that distribution really is bimodal (both for all plasmids and within most bacterial families). But this distribution is different from the distribution of lengths, and information about the distribution of lengths cannot be directly read off the graph of the distribution of log-lengths.

How did the confusion of the length and log-length distributions arise? We suspect that it is due to combining density estimation and a log scale automatically. It is common and reasonable, when plotting a graph with plasmid length on one axis, to use a log scale, because of the large range of plasmid lengths. However, "plotting a density on a log scale" can mean two very different things depending on the order that the density estimation and the log transformation are done. If the density estimation is done first, on the raw lengths, and then this estimate is plotted with a log-scaled axis, the result is like Figure 1B, which is a graph of the distribution of lengths (the numerical values of the plotted

functions are the densities of plasmid lengths), although an unconventional one. But if the log transformation is done first, and then the density estimated from the log-transformed data, one gets an estimate of the distribution of log-lengths. Since Figure 1B is almost never what is actually wanted (it is hard to interpret and puts excessive emphasis on the lower tail of the distribution), graphing software asked to estimate a density and use a log scale will almost always choose to do the log transformation first.

Irrespective of its distribution, plasmid length is an important factor in the biology of a plasmid, which is associated with other properties such as transmissibility. For example, it is fairly clear from Figure 1 that there is a lower bound on the size of a conjugative plasmid (presumably the size of the genes needed to code for the complete conjugation machinery), and below this bound there are only mobilizable and nontransmissible plasmids (see also Coluzzi et al. 2022). Small and large plasmids in *Klebsiella, Escherichia,* and *Salmonella,* as distinguished by the antimode of the log-length distribution, belong to different plasmid taxonomic units (Tal Dagan, private communication). Within those genera, the distribution of plasmid lengths on a linear scale is likely bimodal as well, and the shift in the antimode too small to affect general results (yet, this remains to be verified).

The division of plasmids into large and small is said to be associated with a corresponding division into low and high copy number plasmids. It would be interesting to see if the distribution of plasmid copy numbers is bimodal, and to what extent this is associated with length. Unfortunately, plasmid copy number is not as frequently measured as length, and is not included in databases of plasmid sequences, so it is difficult to perform a similar analysis of the copy number distribution.

To conclude, a clear understanding of the biology of plasmid length would be best served by presenting estimates of the distribution of plasmid lengths, and not of log-lengths, when discussing plasmid length. Since the fact that these distributions have different shapes is not immediately apparent, the log-length

distribution can cause confusion or be misleading.

Acknowledgments

We thank Álvaro San Millán, Dustin Hanke, and Tal Dagan for helpful discussions.

Code availability

The R programs used to generate the figures in the text and the supplement are available as supplementary material.

Funding information

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 400993799 (project 3 within the Research Training Group 2501 "Translational Evolutionary Research", `https://gepris.dfg.de/gepris/projekt/400993799`). I.D. is a member of the International Max Planck Research School for Evolutionary Biology and gratefully acknowledges the benefits provided by the program.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

Ciok, Anna et al. (2016). "Identification of miniature plasmids in psychrophilic Arctic bacteria of the genus *Variovorax*." In: *FEMS Microbiology Ecology* 92.4, fiw043. DOI: `10.1093/femsec/fiw043`.

Coluzzi, Charles et al. (2022). "Evolution of Plasmid Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids." In: *Molecular Biology and Evolution* 39.6, msac115. DOI: 10.1093/molbev/msac115.

Dewan, Ian and Hildegard Uecker (2023). "A mathematician's guide to plasmids: an introduction to plasmid biology for modellers." In: *Microbiology* 169.7, 001362. DOI: 10.1099/mic.0.001362.

Garcillán-Barcia, M. Pilar, Santiago Redondo-Salvo, and Fernando de la Cruz (2023). "Plasmid classifications." In: *Plasmid* 126, 102684. DOI: 10.1016/j.plasmid.2023.102684.

Hall, James P. J. et al. (2021). "What makes a megaplasmid?" In: *Philosophical Transactions of the Royal Society B* 377.1842, 20200472. DOI: 10.1098/rstb.2020.0472.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Robertson, James and John H. E. Nash (2018). "MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies." In: *Microbial Genomics* 4.8. DOI: 10.1099/mgen.0.000206.

Rodríguez-Beltrán, Jerónimo et al. (2021). "Beyond horizontal gene transfer: the role of plasmids in bacterial evolution." In: *Nature Reviews Microbiology* 19.6, pp. 347–359. DOI: 10.1038/s41579-020-00497-1.

Schmartz, Georges P. et al. (2021). "PLSDB: advancing a comprehensive database of bacterial plasmids." In: *Nucleic Acids Research* 50, pp. D273–D278. DOI: 10.1093/nar/gkab1111.

Slater, Frances R. et al. (2008). "Progress towards understanding the fate of plasmids in bacterial communities." In: *FEMS Microbiology Ecology* 66.1, pp. 3–13. DOI: 10.1111/j.1574-6941.2008.00505.x.

Smillie, Chris et al. (2010). "Mobility of Plasmids." In: *Microbiology and Molecular Biology Reviews* 74.3, pp. 434–452. DOI: 10.1128/mmbr.00020-10.