# Evolutionary trajectories of secondary replicons in multipartite genomes

**Authors:** Natalia O. Dranenko[1], Aleksandra D. Rodina[2], Yaroslav V. Demenchuk[2], Mikhail S. Gelfand[1,3*], Olga O. Bochkareva[4*#]

**Affiliations:**

[1] A.A. Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia

[2] School of Molecular and Theoretical Biology, Brentwood, CA, USA

[3] Skolkovo Institute of Science and Technology, Moscow, Russia

[4] Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria

[*] equal contribution [#] corresponding author, olga.bochkareva@univie.ac.at

ADR current affiliation: Immanuel Kant Baltic Federal University, Kaliningrad, Russia

YVD current affiliation: National Research University Higher School of Economics (HSE University), Moscow, Russia

OOB current affiliation: CUBE, Department of Microbiology, University of Vienna, Austria

**Key words:** bacteria, chromosome, chromid, evolution, pan-genome

**Data summary:** The datasets supporting the conclusions of this article are available at https://github.com/zaryanichka/MultipleReplicons

## Abstract

Most bacterial genomes have a single chromosome that may be supplemented by smaller, dispensable plasmids. However, approximately 10% of bacteria with completely sequenced genomes, mostly pathogens and plant symbionts, have more than one stable large replicon. Some secondary replicons are species-specific, carrying pathogenicity or symbiotic factors. Other replicons are common on at least the genus level, carry house-keeping genes, and may have a size of several million base pairs.

We analyzed the abundance and sizes of large secondary replicons in different groups of bacteria and identified two patterns in the evolution of multipartite genomes. In nine genera of four families, *Pseudoalteromonadaceae*, *Burkholderiaceae*, *Vibrionaceae*, and *Brucellaceae*, we observed a positive correlation between the sizes of the chromosome and the secondary replicon with the slope in the range of 0.6–1.2. This indicates that in these genera the replicons evolve in a coordinated manner, with comparable rates of gene gain/loss, hence supporting classification of such secondary replicons as ‘chromids’. The second, more common pattern, features gene gains and losses mainly occurring in the primary replicon, yielding a stable size of the secondary replicon. Such secondary replicons are usually present in only a low fraction of the genus’ species. Hence, such replicons behave as ‘megaplasmids’. A mixed situation was observed in symbiotic genera from the *Rhizobiaceae* family where the large secondary replicons are of stable size, but present in all species. These results may provide a general framework for understanding the evolution of genome complexity in prokaryotes.

## Significance

Large secondary replicons are observed in representatives of many taxonomic groups of bacteria. Traditionally, they are referred to as *second chromosomes*, *chromids*, or *megaplasmids*, with little consistency, in particular because their evolution remains understudied. Here we demonstrate that the sizes of secondary replicons follow two main evolutionary trends: replicons whose size scales linearly with the size of the main chromosome (the suggested term *chromids*) typically contain numerous essential genes (rRNA, tRNA, ribosomal proteins), while large secondary replicons of stable size (termed *megaplasmids*) contain fewer or none such genes.

## Introduction

Classical microbiology assumed bacterial genomes to be a single circular chromosome possibly supplemented by small nonessential circular plasmids. Since 1979, when linear plasmids in *Streptomyces* were discovered (Hayakawa et al. 1979), these assumptions started to change. Ten years later two chromosomes in the *Rhodobacter* genome were described (Suwanto and Kaplan 1989). This type of genome organization was named multipartite genomes, with about 10% of sequenced bacterial species being of that type (diCenzo and Finan 2017).

When a bacterial genome is formed by several large elements, the largest one, that holds the main fraction of house-keeping genes, is considered the primary chromosome, other parts being classified as chromids, megaplasmids, or plasmids based on their size and gene content (for details, see (diCenzo and Finan 2017)), but in the absence of universal definitions this classification is mainly based on arbitrary thresholds (Burton et al. 2013; Weiser et al. 2019; Cazares et al. 2020; Bottacini et al. 2015; diCenzo and Finan 2017). The term *chromid* was introduced by (Harrison et al. 2010) to describe large secondary replicons replicated by the plasmid-type mechanism but containing essential genes found in chromosomes in other species. Chromids were observed mostly in *Proteobacteria* but also in some distant genera, such as *Prevotella*, *Leptospira*, and *Deinococcus* (Harrison et al. 2010). The term *megaplasmid* is applied to plasmids larger than ~100 kb (Burton et al. 2013; Weiser et al. 2019; Cazares et al. 2020; Bottacini et al. 2015; diCenzo and Finan 2017). Genomes of some species may hold several secondary replicons of different types, e.g. *Burkholderia cepacia* that has a chromid and a huge essential megaplasmid (Bochkareva et al. 2018).

Evolutionary advantages of the multi-chromosomal genome organization are not clear. One hypothesis suggests that it has evolved to allow for gene accumulation and genome size expansion with two opposite directions of selection, one driving accumulation of new genes to increase the organism fitness, and the other reducing the genome to increase the replication rate. Multiple replicons may replicate simultaneously allowing for faster bacterial cell division even for large genomes (Guo et al. 2003), although there is no strong correlation between the number of replicons and the replication time (diCenzo and Finan 2017). The bacterial genome size varies from 112 kb for obligate symbionts (Bennett and Moran 2013) to ~15 Mb for some free-living organisms (Han et al. 2013). While multipartite genomes are larger on average, large single-chromosome genomes exist, and less than one third of genomes larger than 6Mb are multipartite (diCenzo and Finan 2017). Therefore, the genome size is not uniquely driven by the number of replicons. Another hypothesis suggests that secondary replicons are expected to be under weaker selective pressure and have more

hotspots of incorporation of new horizontally transferred genes (diCenzo et al. 2019). At that, secondary replicons may act as “test beds” that accumulate new genes and allow them to evolve more rapidly.

Here we study the relationships between the replicons’ sizes in multipartite genomes and interpret these observations in the context of evolution.

# Results

## Genome sizes

In the absence of a universally accepted definition of a multipartite genome, we define ‘multipartite genome’ as a genome with at least one secondary replicon longer than 500 Kbp, the threshold set at the first local minimum of the distribution of the secondary-replicon sizes (Supplementary Figure 1). We analyzed all representative bacterial genomes from RefSeq to select genera with at least one reference multipartite genome. Then we filtered out genera that comprised less than six genomes (see Methods). Thus, we analyzed genome configuration and size for 547 species from 36 genera.

The smallest genome (2.5 Mb) was of *Prevotella* sp. F0039 and the largest genome (12.2 Mb) was of *Streptomyces sp.* RLB1-9 (Supplementary Table 1). In agreement to previous observations, we did not observe any association between the genome size and taxonomy (Figure 1).

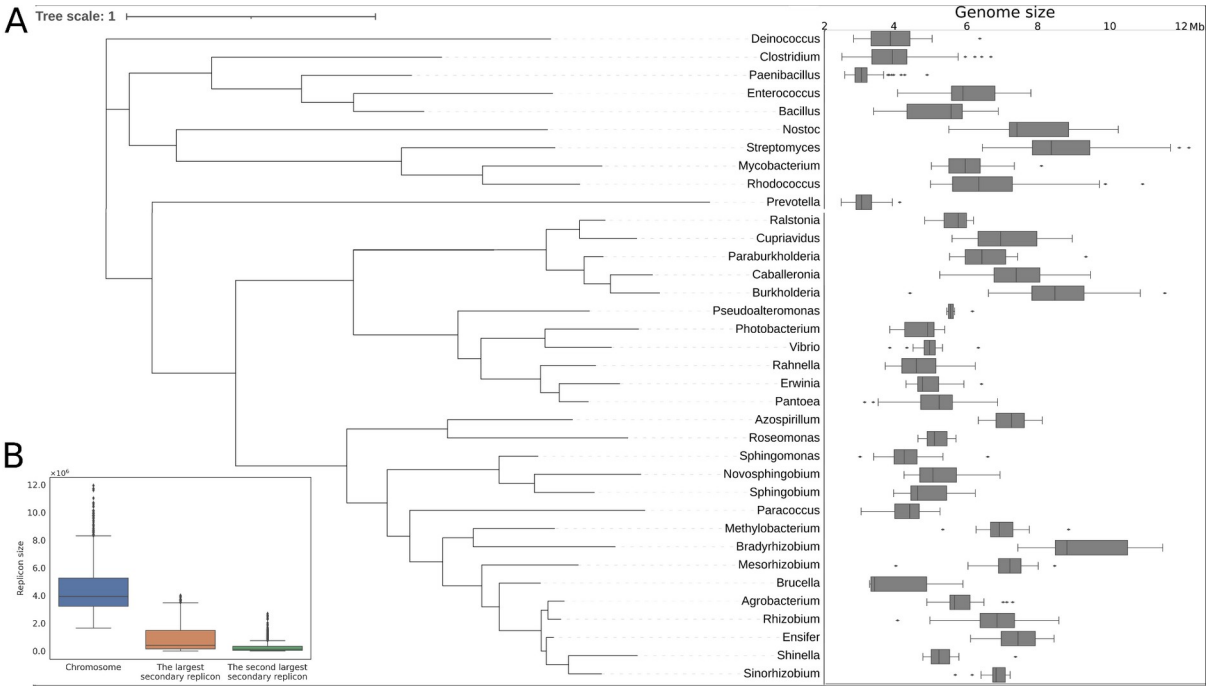



**Figure 1.** Replicon sizes in 36 genera with multipartite genomes. (A) Genome sizes of strains from genera that have members with multipartite genomes (see Methods). The cladogram topology follows the NCBI taxonomy (Schoch et al. 2020). (B) Replicon sizes; blue shows chromosomes; orange, the largest secondary replicon; green, the second-largest recondary replicon.

## Replicon sizes


3

We compare the sizes of chromosomes and secondary replicons in genera with multipartite genomes. The chromosomes in some genera reach 12 Mbp, and the secondary replicons may be of length 4 Mbp (Supplementary Figure 1,2). The intervals of sizes of the chromosome and the largest secondary replicon is very large, while the second-largest secondary replicons rarely reach the chromosome size.

## Relationships between replicon sizes

Visual comparison of sizes of the chromosomes and the largest secondary replicons revealed two trends: the *horizontal* trend formed by plasmids with a stable size and the *diagonal* trend reflecting the consistent growth or reduction of both replicons (Figure 2).

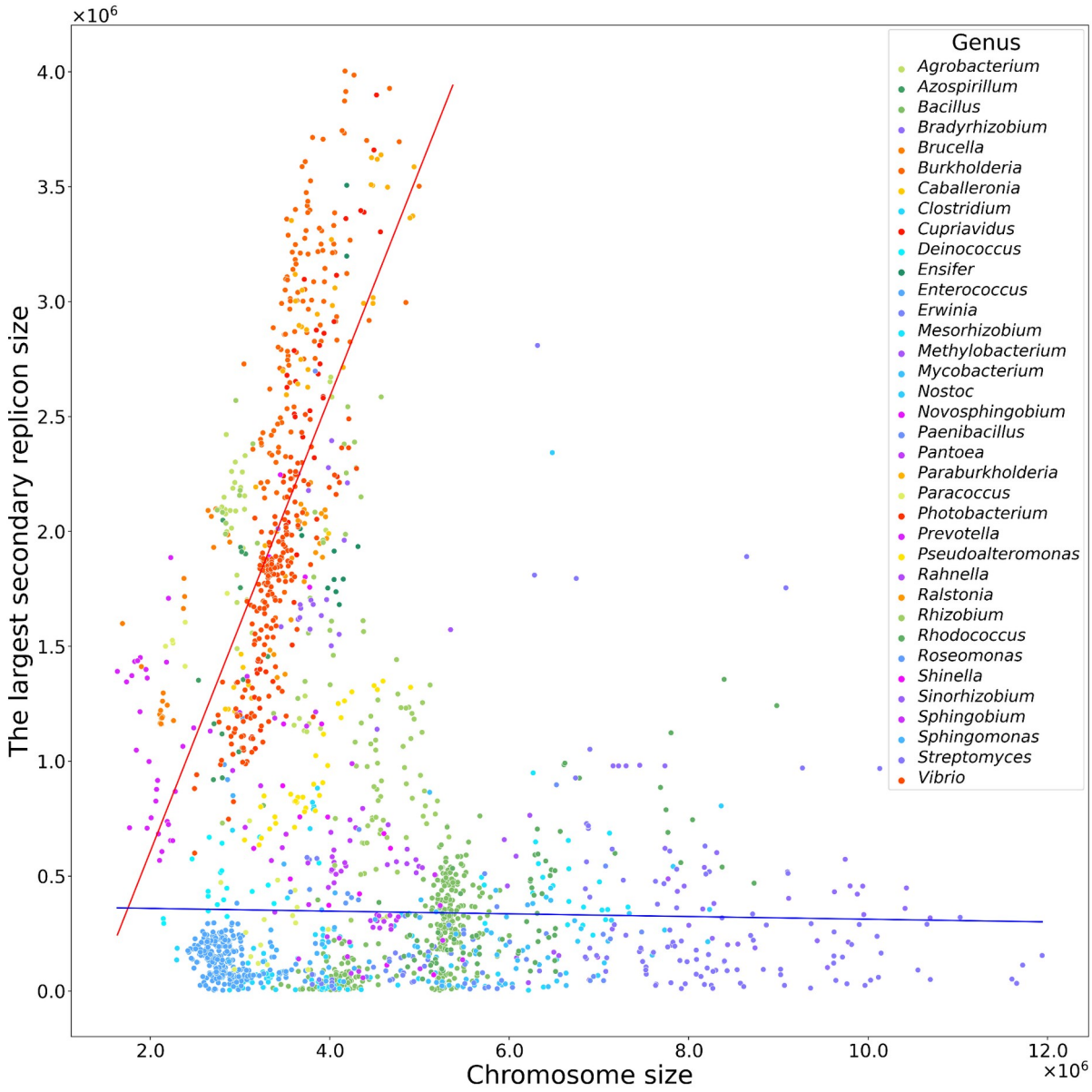


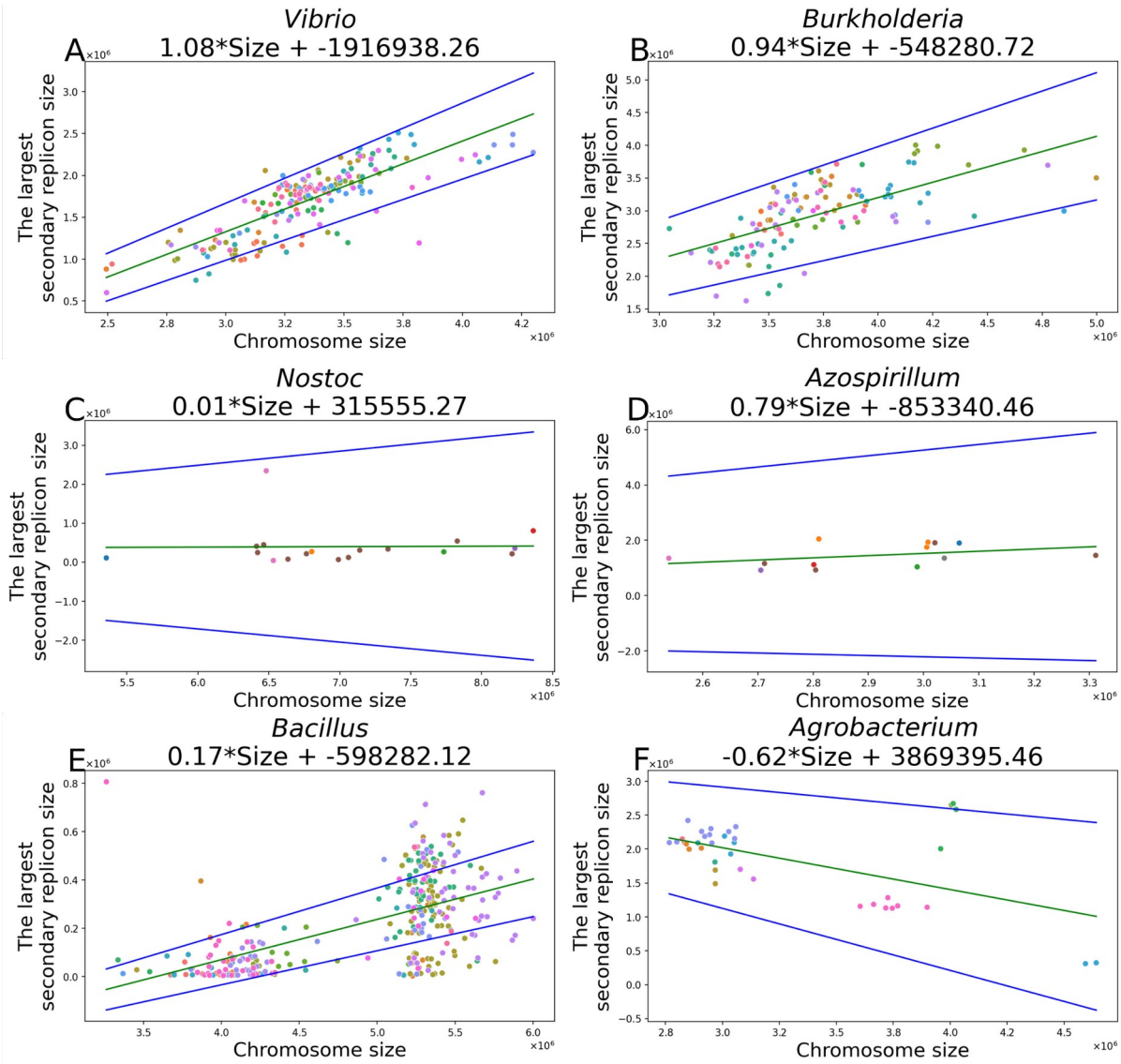

**Figure 2.** The chromosome size (horizontal axis) and the size of the largest secondary replicon (vertical axis) for strains with multipartite genomes from 36 genera. The dot color shows the bacterial genus (see the legend). Blue-purple colors show genera with significant positive correlation, yellow-red colors show genera with approximately constant size of the


4

secondary replicon, green colors show genera with non-significant positive correlation or non-homogeneous genera.

To confirm this hypothesis, we performed linear regression analysis for each genus separately (Supplementary Figure 3, Table 1). We observed a significant positive correlation between the sizes of chromosomes and the largest secondary replicons in the genera where the multipartite genome organization is common for all or at least most species (Figure 3A,B). In turn, in genera where species with multipartite genomes are rare, we observed no correlation (Figure 3C).

However, there were several exceptions. *Ensifer and Azospirillum,* genera with the multipartite genome organization being common for all species, demonstrated positive but non-significant correlation (Figure 3D). A more interesting case was that of heterogeneous genera with significant positive or negative correlation (*Paracoccus, Agrobacterium, Rhizobium, Bacillus, Rhodococcus*). Linear regression is not applicable in this case, as these genera are formed by diverse subgroups of strains yielding separate clusters in the scatterplots (Figure 3E,F). Indeed, it is well known that *Bacillus* spp. are highly diverse and should be reclassified into several genera (Gupta et al. 2020). These genera were excluded from further analysis.





5

**Figure 3.** Linear approximation of the relationships between sizes of the chromosome and the largest secondary replicon in: (A, B) *Vibrio* ($r$ = 0.65, slope coefficient 1.08 ± 0.12) and *Burkholderia* ($r$ = 0.8, slope coefficient 0.94 ± 0.2), genera with a strong positive correlation; (C) *Nostoc*, genus with an approximately constant size of the secondary replicon (slope coefficient 0.01 ± 0.34); (D) *Azospirillum*, genus with a not-significant correlation (linear regression has a high variance; $r$ = 0.4, slope coefficient 0.79 ± 1.14). (E,F) *Bacillus* and *Agrobacterium*, genera formed by several clusters of strains (the linear regression does not make sense). Green line reflects the resulting slope coefficient, blue lines reflect the 95% confidence interval for the slope.

**Table 1.** Abundance of multipartite genomes and the coefficients of linear regression between the sizes of the chromosome and the largest secondary replicon, per genus. Columns: Total — the number of species in the genus; multipartite — the number of species with strains having multipartite genomes (for species containing less than 5 strains the value is given in parentheses); slope and intercept — coefficients of the linear regression; slope_error and intercept_error — standard error of the estimated slope and intercept under the assumption of residual normality; Pearson correlation — the Pearson correlation coefficient; R2 — Coefficient of determination; p-value — the *p*-value for the null hypothesis that the slope is zero, using the Wald test with t-distribution of the test statistic; Range and Mode — the number of replicons, its range and mode value, in a genus; rRNAs, rProts, and tRNAs — the mean number of secondary replicons with the respective features, in a genus; Trend — strong positive correlation (PC), approximately constant size of the secondary replicon (HT), or non-significant positive correlation or non-homogeneous genera (NS).

| | Species | | Linear regression | | | | Statistics | | | Number of replicons | | Core elements on secondary replicons | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Genus | Total | Multi partite* | Slope | Slope error | Intercept, Mb | Intercept error, Mb | Pearson correlation | R2 | p-value | Range | Mode | rRNAs | rProts | tRNAs | Trend |
| *Cupriavidus* | 11 | 11 | 1,25 | 0,15 | -2,06 | 0,58 | 0,85 | 0,72 | 5,45E-09 | 2-8 | 3 | 1,0 | 0,2 | 1,5 | PC |
| *Photobacterium* | 6 | 6 | 1,12 | 0,13 | -2,31 | 0,44 | 0,90 | 0,81 | 1,56E-07 | 2-7 | 3 | 0,1 | 0,1 | 1,2 | PC |
| *Vibrio* | 63 | 63 | 1,08 | 0,06 | -1,92 | 0,19 | 0,81 | 0,65 | 3,42E-45 | 1-7 | 2 | 0,6 | 0,4 | 1,1 | PC |
| *Burkholderia* | 32 | 32 | 0,94 | 0,10 | -0,55 | 0,37 | 0,65 | 0,43 | 1,88E-16 | 1-8 | 2 | 1,6 | 1,2 | 1,7 | PC |
| *Brucella* | 14 | 14 | 0,94 | 0,15 | -0,65 | 0,39 | 0,80 | 0,64 | 3,10E-06 | 1-6 | 2 | 1,1 | 1,2 | 1,2 | PC |
| *Ensifer* | 5 | 5 | 0,86 | 1,00 | -1,37 | 4,03 | 0,28 | 0,08 | 4,08E-01 | 3-6 | 3 | 1,3 | 0,3 | 1,9 | NS |
| *Ralstonia* | 7 | 5(+2) | 0,83 | 0,20 | -1,16 | 0,75 | 0,66 | 0,44 | 5,48E-04 | 2-4 | 2 | 0,7 | 0,1 | 1,0 | PC |
| *Paraburkholderia* | 24 | 24 | 0,81 | 0,11 | -0,29 | 0,45 | 0,79 | 0,63 | 2,64E-08 | 2-8 | 3 | 1,2 | 1,0 | 1,9 | PC |
| *Azospirillum* | 8 | 8 | 0,79 | 0,57 | -0,85 | 1,65 | 0,39 | 0,15 | 1,89E-01 | 6-10 | 6 | 3,5 | 1,8 | 4,9 | NS |
| *Caballeronia* | 4 | 4 | 0,66 | 0,07 | -0,65 | 0,22 | 0,97 | 0,94 | 8,64E-05 | 4-8 | 5 | 1,1 | 0,4 | 2,5 | PC |
| *Pseudoalteromonas* | 26 | 20(+2) | 0,48 | 0,04 | -0,88 | 0,15 | 0,91 | 0,83 | 1,12E-13 | 1-5 | 2 | 0,3 | 0,8 | 0,4 | PC |
| *Rhodococcus* | 18 | 6(+1) | 0,21 | 0,02 | -0,87 | 0,14 | 0,72 | 0,52 | 3,46E-13 | 1-7 | 2 | 0,0 | 0,0 | 0,5 | NS |
| *Bacillus* | 52 | (+2) | 0,17 | 0,01 | -0,60 | 0,07 | 0,58 | 0,33 | 5,04E-30 | 1-16 | 1 | 0,0 | 0,1 | 0,3 | NS |
| *Deinococcus* | 17 | (+4) | 0,10 | 0,07 | 0,19 | 0,19 | 0,33 | 0,11 | 1,63E-01 | 1-11 | 4 | 0,3 | 0,1 | 1,0 | HT |
| *Mesorhizobium* | 11 | (+2) | 0,03 | 0,05 | 0,23 | 0,32 | 0,14 | 0,02 | 5,17E-01 | 1-5 | 1 | 0,0 | 0,0 | 0,3 | HT |
| *Clostridium* | 46 | 1(+1) | 0,01 | 0,03 | 0,12 | 0,10 | 0,03 | 0,00 | 7,65E-01 | 1-7 | 1 | 0,0 | 0,1 | 0,1 | HT |
| *Nostoc* | 7 | (+1) | 0,01 | 0,17 | 0,32 | 1,17 | 0,02 | 0,00 | 9,44E-01 | 1-16 | 6 | 0,0 | 0,2 | 0,7 | HT |
| *Mycobacterium* | 61 | (+1) | -0,02 | 0,04 | 0,27 | 0,23 | -0,07 | 0,00 | 6,63E-01 | 1-5 | 1 | 0,0 | 0,0 | 0,5 | HT |
| *Sphingomonas* | 30 | (+2) | -0,04 | 0,05 | 0,38 | 0,19 | -0,18 | 0,03 | 3,78E-01 | 1-6 | 1 | 0,0 | 0,0 | 0,2 | HT |
| *Enterococcus* | 17 | (+2) | -0,06 | 0,03 | 0,31 | 0,08 | -0,12 | 0,01 | 5,05E-02 | 1-13 | 2 | 0,0 | 0,0 | 0,1 | HT |
| *Roseomonas* | 5 | 0 | -0,07 | 0,09 | 0,75 | 0,38 | -0,32 | 0,10 | 4,39E-01 | 1-10 | 5 | 0,1 | 0,0 | 1,0 | HT |
| *Paenibacillus* | 52 | (+2) | -0,09 | 0,08 | 0,75 | 0,47 | -0,19 | 0,04 | 2,74E-01 | 1-7 | 1 | 0,2 | 0,3 | 0,2 | HT |
| *Erwinia* | 9 | (+1) | -0,10 | 0,07 | 0,62 | 0,31 | -0,30 | 0,09 | 1,62E-01 | 1-6 | 2 | 0,0 | 0,0 | 0,0 | HT |
| *Streptomyces* | 141 | 5(+1) | -0,11 | 0,03 | 1,27 | 0,28 | -0,31 | 0,09 | 1,00E-03 | 1-8 | 1 | 0,0 | 0,1 | 0,5 | HT |
| *Bradyrhizobium* | 23 | (+3) | -0,12 | 0,04 | 1,54 | 0,34 | -0,51 | 0,26 | 3,46E-03 | 1-5 | 1 | 0,0 | 0,0 | 0,5 | HT |
| *Sinorhizobium* | 5 | 3(+2) | -0,14 | 0,28 | 2,33 | 1,08 | -0,12 | 0,01 | 6,18E-01 | 2-7 | 4 | 0,1 | 0,4 | 1,8 | HT |
| *Methylobacterium* | 14 | (+2) | -0,22 | 0,18 | 1,78 | 1,15 | -0,34 | 0,12 | 2,58E-01 | 1-9 | 1 | 0,4 | 0,1 | 1,1 | HT |
| *Rahnella* | 4 | 3(+1) | -0,24 | 0,13 | 1,74 | 0,64 | -0,57 | 0,32 | 1,12E-01 | 2-4 | 3 | 0,0 | 0,0 | 0,0 | HT |
| *Pantoea* | 12 | 2(+3) | -0,27 | 0,09 | 1,57 | 0,39 | -0,46 | 0,21 | 6,44E-03 | 1-12 | 4 | 0,0 | 0,0 | 0,1 | HT |
| *Sphingobium* | 12 | 5(+1) | -0,37 | 0,10 | 2,13 | 0,39 | -0,62 | 0,38 | 1,04E-03 | 1-11 | 4 | 0,6 | 0,1 | 0,8 | HT |
| *Prevotella* | 19 | 6(+4) | -0,44 | 0,33 | 1,96 | 0,68 | -0,24 | 0,06 | 1,99E-01 | 1-4 | 2 | 1,0 | 1,0 | 1,0 | HT |
| *Rhizobium* | 24 | 18(+5) | -0,53 | 0,08 | 3,54 | 0,36 | -0,56 | 0,31 | 1,01E-08 | 2-9 | 5 | 0,2 | 1,0 | 0,9 | NS |
| *Agrobacterium* | 11 | 11 | -0,62 | 0,15 | 3,87 | 0,49 | -0,57 | 0,32 | 1,65E-04 | 2-10 | 4 | 0,9 | 0,8 | 1,2 | NS |
| *Novosphingobium* | 6 | (+2) | -0,62 | 0,56 | 3,38 | 2,09 | -0,36 | 0,13 | 3,01E-01 | 1-6 | 1 | 0,2 | 0,0 | 0,7 | HT |
| *Shinella* | 3 | 1(+1) | -0,68 | 0,25 | 3,57 | 1,06 | -0,77 | 0,60 | 4,19E-02 | 2-13 | 3 | 0,0 | 0,1 | 0,7 | HT |
| *Paracoccus* | 16 | (+4) | -0,88 | 0,15 | 3,35 | 0,48 | -0,79 | 0,63 | 1,20E-05 | 1-23 | 3 | 0,4 | 0,3 | 1,4 | NS |

This procedure allowed us to classify the genera into two groups:

- significant positive correlation was detected in *Pseudoalteromonas, Burkholderia, Caballeronia, Cupriavidus, Paraburkholderia, Ralstonia, Photobacterium, Vibrio, Brucella.*
- non-significant or no correlation was observed in *Bradyrhizobium, Clostridium, Deinococcus, Enterococcus, Erwinia, Mesorhizobium, Methylobacterium, Mycobacterium, Nostoc, Novosphingobium, Paenibacillus, Pantoea, Prevotella, Sinorhizobium, Rahnella, Roseomonas, Shinella, Sphingobium, Sphingomonas, Streptomyces.*

Separate linear regression analysis for the groups (Figure 2) yielded two trends. For the first group $S$=0.99×$C$–1373734 ($C$ is the chromosome size, S is the size of the largest secondary replicon), that is, $S \approx C + 1{,}4 \times 10^6$, and the correlation $R$=0.63. For the second group, $S$=–0.006×$C$+370674 with $R$=–0.03 and the 95% confidence interval for the correlation being (–0.0193, 0.0076); so this trend is essentially a horizontal one.

## Resulting trends

To get a general view, for each genus we constructed a Gaussian function with the mean equal to the slope coefficient and the standard deviation equal to the standard deviation of this coefficient (Figure 4A). Then we built a sum of these Gaussian functions (Figure 4B). The resulting distribution was bimodal, with the first peak at –0.05 and the second peak at ≈1.0. Hence, indeed, secondary replicons can be classified into two groups based on the relationships between their size and the chromosome size, with *megaplasmids* having no correlation, and *chromids* featuring a strong positive correlation.

Note that replicons classified as chromids have sizes in the range 0.5-4 Mb, while megaplasmids do not exceed 3.5Mb and many are shorter than 500 kb (Figure 4C). Thus, the selection of genera with at least one 500 kb secondary replicon indeed has been an appropriate initial filter to study the organization of multipartite genomes.


7

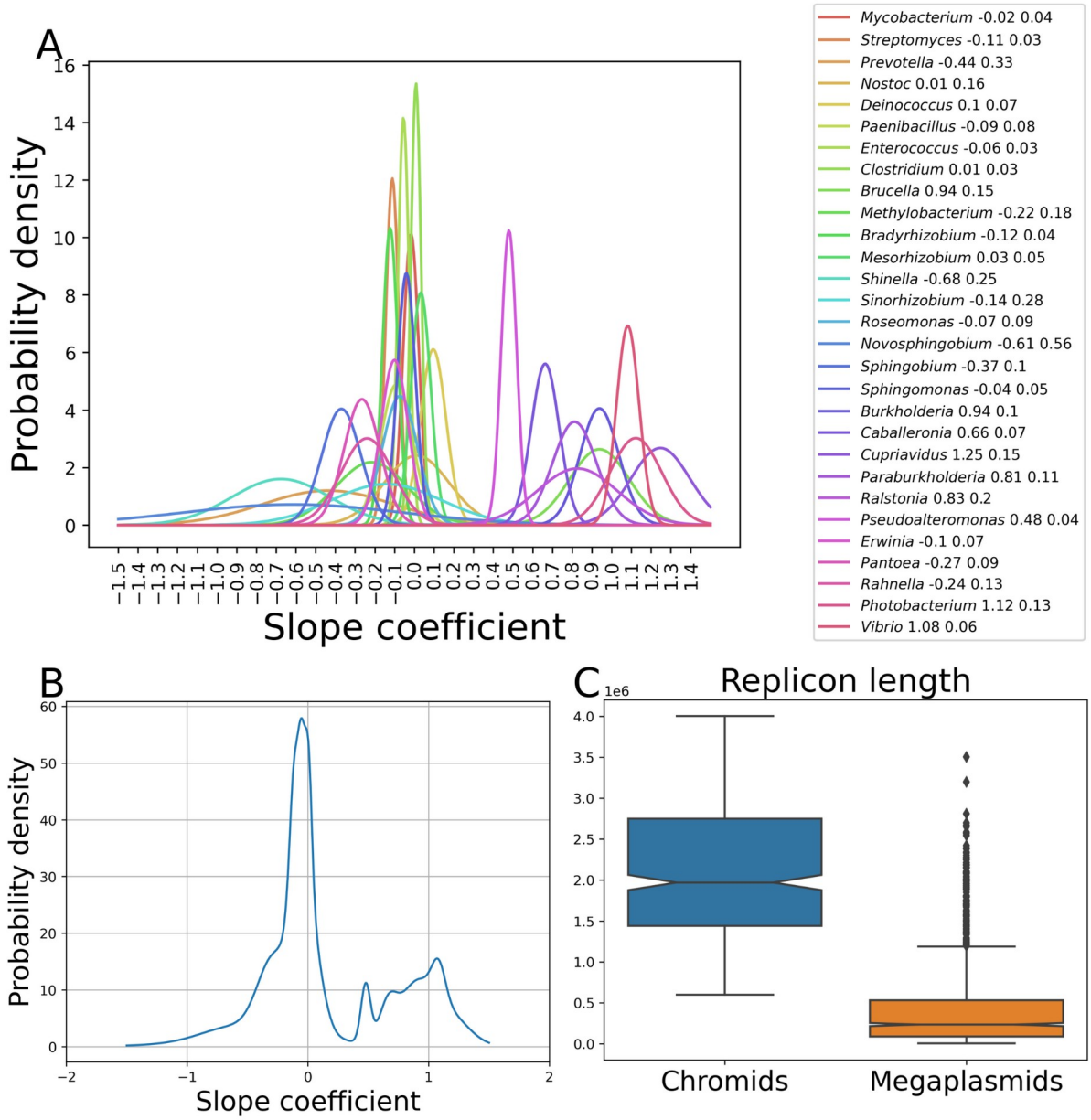



**Figure 4.** Summary of the regression analysis. (A) Gaussian functions derived from the slope coefficients (see the text), each colored line corresponds to a genus (listed in the legend; the numbers are the mean and the standard deviation). (B) Sum of the Gaussian functions. (C) Size distribution of the replicons classified as chromids and (mega)plasmids according to the regression analysis.

## Contribution of other replicons

If gene distribution between the components of a genome is shaped by selection forces, the total number of replicons in a genome may affect the relationships between their sizes. It may explain the absence of correlation in *Rhizobia* as they possess many large replicons in comparison to *Vibrio*, typically having only two large replicons.

In all genera with the largest secondary replicons characterized above as ‘megaplasmids’, including *Rhizobiaceae* family, no correlation between the chromosome size and the total


8

size of secondary replicons was observed (Supplementary Figure 4). Moreover, some genomes with chromids contain additional replicons of size comparable to that of chromids; their sizes do not demonstrate any consistent behavior (Supplementary Figure 5). In particular, in addition to a chromosome and a chromid, species of the *Burkholderia cepacia* complex (*Bcc*) possess a common megaplasmid (~1 Mb) involved in virulence and stress tolerance (Agnoli et al. 2012). No significant difference between trends in *Bcc* with those of other *Burkholderia* was observed, demonstrating that acquisition of megaplasmids had not affected the relationship between the chromosome and chromid sizes (Supplementary Figure 6).

## rRNA, tRNA, and ribosomal protein genes

We compared the occurrence of three categories of *core* genes, namely rRNA genes, tRNA genes, and genes encoding ribosomal proteins (r-protein genes) in chromids and megaplasmids. As the chromids are bigger than 0.5 Mb while the megaplasmids are usually smaller (Figure 5A), we avoid biases by dividing secondary replicons into three groups based on their size and type: chromids (487 replicons), megaplasmids longer than 500 kb (385 replicons), and megaplasmids shorter than 500 kb (1025 replicons).

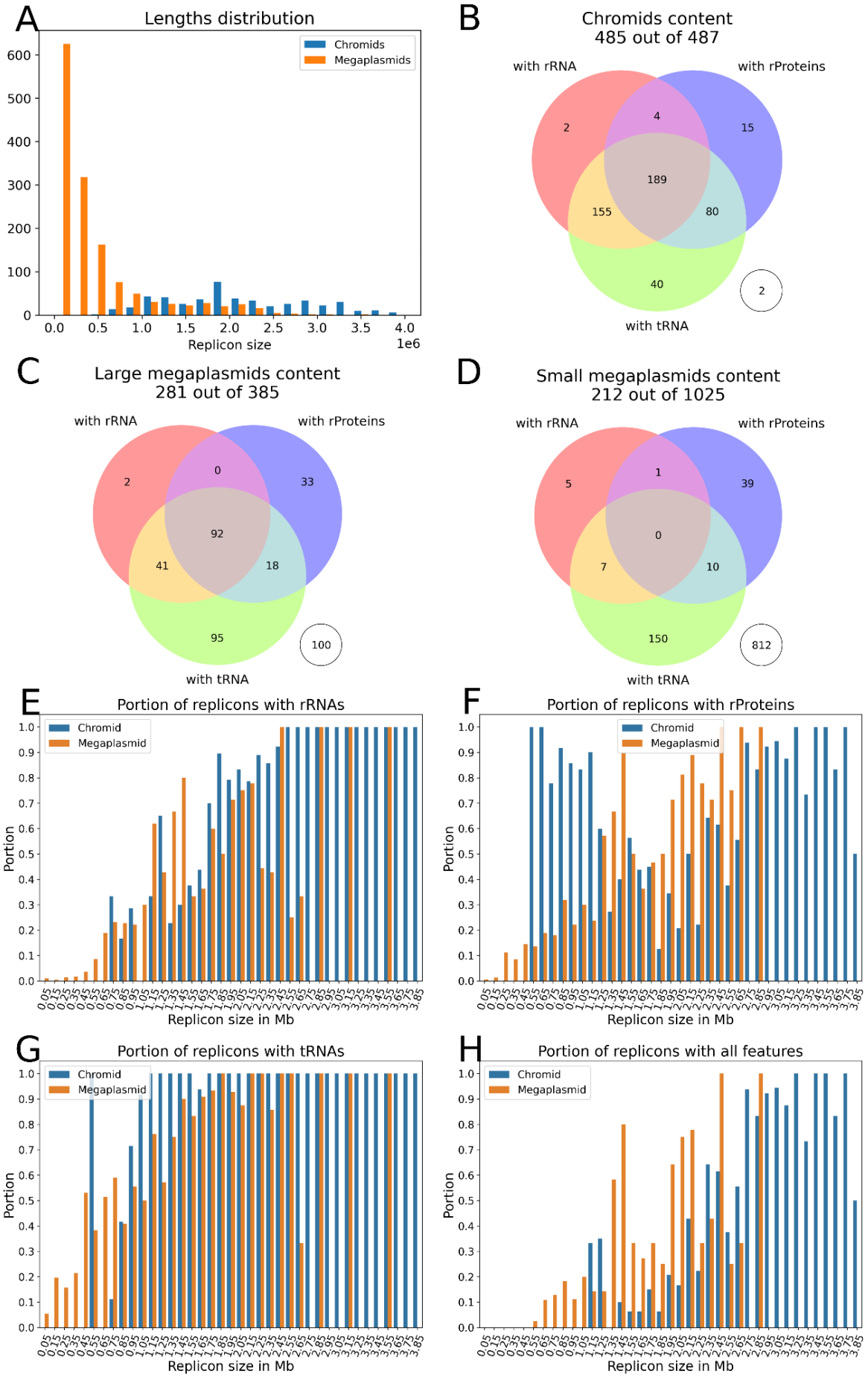
No such genes were found in only two chromids but more than one fourth of large megaplasmids (Figure 5B,C). In turn, ~40% chromids (mainly from *Brucella, Burkholderia, Paraburkholderia*) and only ~25% megaplasmids (mainly *Azospirillum* and *Prevotella*) featured at least one gene of each category. In small (mega)plasmids, we observed no replicons with at least one gene of each category, and only 10% of them possessed at least one of the core genes (Figure 5D).

Since the presence of the core genes could reflect the size of the secondary replicon rather than its type, we further calculated the fraction of replicons with core genes in cohorts defined by size (Figure 5E-H, Supplementary Figure 7). Indeed, the fraction of replicons with rRNA and tRNA genes is lower in smaller replicons. Surprisingly, although r-protein genes were rare in small megaplasmids, they were found in all size cohorts of chromids. In contrast to rRNA and tRNA genes, most r-protein genes are single-copy ones (Yutin et al. 2012) and their location in secondary replicons may explain the essentiality of the latter. To address this, we selected only those r-protein genes which were not present in chromosomes.

We observed such r-protein genes in the largest secondary replicons in *Brucella* (L33,S21), some *Vibrio* species (L20,L35), in of *Agrobacterium* species, and only several *Rhizobium* species (L36,L31,S1,S21,L28). In *Azospirillum*, the genus with an exceptionally high number of large replicons, the largest secondary replicon may contain (L33, L28) or (L13, S9) sets of r-protein genes.

The largest number of r-protein genes was observed in secondary replicons in the *Prevotella* genus. Note that genus *Prevotella* comprises species with a single ~3Mb chromosome and species with two replicons, ~2Mb and ~1Mb in size. Moreover, in two-replicon species, rRNA, tRNA, and r-protein genes are proportionately distributed between the larger and smaller replicons indicating formation of such multipartite genome organization via chromosome division or translocation of a large fragment from the chromosome to the plasmid via intragenomic recombination.


9


A
Lengths distribution
Chromids
Megaplasmids
Replicon size
1e6
B
Chromids content
485 out of 487
with rRNA
with rProteins
with tRNA
2
4
15
189
155
80
40
2
C
Large megaplasmids content
281 out of 385
with rRNA
with rProteins
with tRNA
2
0
33
92
41
18
95
100
D
Small megaplasmids content
212 out of 1025
with rRNA
with rProteins
with tRNA
5
1
39
0
7
10
150
812
E
Portion of replicons with rRNAs
Chromid
Megaplasmid
Portion
Replicon size in Mb
F
Portion of replicons with rProteins
Chromid
Megaplasmid
Portion
Replicon size in Mb
G
Portion of replicons with tRNAs
Chromid
Megaplasmid
Portion
Replicon size in Mb
H
Portion of replicons with all features
Chromid
Megaplasmid
Portion
Replicon size in Mb



10

**Figure 5. Occurrence of core features in secondary replicons.** (A) Length distribution of secondary replicons characterized as chromids and megaplasmids based on the regression analysis. (B–D) Venn diagrams showing the occurrence of rRNA, tRNA, and r-protein genes encoding ribosomal proteins in secondary replicons characterized as chromids (B) and large (C) and small (D) megaplasmids; numbers in empty circles show the number of replicons with none of these genes. (E–G) Fractions of chromids and megaplasmids, depending on length, with rRNA genes (E), r-protein genes (F), and tRNA genes (G), at least one core gene of any category (H).

# Discussion

The correlation between the chromosome and chromid sizes may reflect consistent growth or reduction of both replicons. In particular, in *Burkholderiaceae*, chromids show rapid accumulation of recently acquired genes associated with the adaptation to novel environments (diCenzo et al., 2019). Another example is an intracellular pathogen *B. mallei* that has relatively recently evolved from the extracellular pathogen *B. pseudomallei* (Bochkareva et al., 2018). This transition was accompanied by strong genome reduction, 12.5% in the chromosome and 31% in the chromid, that retained the proportion between the replicon sizes.

The second, more common trend demonstrates the absence of such correlation. This trend was mainly observed in genera with low fraction of multipartite genomes. Interestingly, it was also seen in megaplasmids of symbiotic genera from *Rhizobiaceae* (*Agrobacterium, Sinorhizobium, Rhizobium, Ensifer,* and *Shinella*), and *Azospirillaceae (Azospirillum)*. Note that large essential secondary replicons are common in *Rhizobiaceae* and *Azospirillum*, thus the absence of correlation could have been explained by the contribution of additional replicons. However, in these genera the correlation between the chromosome size and the total size of secondary replicons also was not observed. Moreover, acquisition of large megaplasmids by bacteria with chromids (e.g. *Bcc*) does not affect the chromosome-chromid size correlation.

The presented analysis may explain the origin and evolution of multipartite genomes. The size of a bacterial genome is associated with the lifestyle, with free-living bacteria having large genomes with open pangenomes, and intracellular pathogens having smaller genomes with smaller fractions of periphery (strain-specific) genes (Köstlbacher et al. 2021; Horesh et al. 2021). The correlation between the chromosome and the chromid sizes in *Pseudoalteromonas, Burkholderiaceae, Vibrionaceae,* and *Brucellaceae* both in genera and on the family level (Figure 6A-D) indicates that their chromids are involved in the gene gain/loss process, likely shaped by evolutionary selection under adaptation to different hosts and niches. In contrast, in *Rhizobiaceae* we observe no link between the chromosome and chromid sizes (Figure 6E)**.** *Rhizobiaceae* are plant symbionts, characterized by the ability to modify plant development, with genes essential for nodulation and nitrogen-fixation located in megaplasmids and in chromosomal genomic islands (Yang et al. 2020). The absence of correlation may indicate specific functionality and, as a result, high conservation of composition of *Rhizobiaceae* secondary replicons.


11

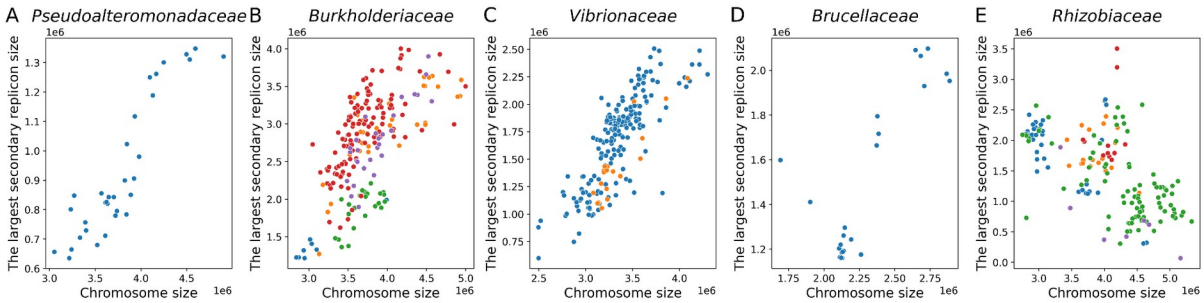



**Figure 6.** Relationships between the chromosome size and the largest secondary replicon size for five families with chromids: (A) *Pseudoalteromonadaceae* (*Pseudoalteromonas*), (B) *Burkholderiaceae* (*Caballeronia*, *Paraburkholderia*, *Ralstonia*, *Burkholderia*, *Cupriavidus*), (C) *Vibrionaceae* (*Vibrio*, *Photobacterium*), (D) *Brucellaceae* (*Brucella*), and (E) *Rhizobiaceae* (*Agrobacterium*, *Sinorhizobium*, *Rhizobium*, *Ensifer*, *Shinella*). Dot's color reflects different genera.

Current classifications of secondary replicons of bacteria are rather arbitrary. The suggested criteria rely on the replicon size threshold and the presence of core (universal) or essential genes (Harrison et al. 2010; Hall et al. 2022). However, the definition of essentiality depends on the research question (e.g. with or without competition between strains) (Gerdes et al. 2003), whereas the core fraction of pan-genome depends on the taxonomy level of the latter. According to the strictest definition, chromids carry at least one gene that is essential for cell viability. This definition relies on experimental evidence that a replicon carries an indispensable core gene (diCenzo and Finan 2017). Here, we have considered three categories of core genes, common for all bacteria, tRNA and rRNA genes which are often found in multiple copies in many replicons (Santos and Del-Bem 2022; Gifford, Dasgupta, and Barrick 2021) and genes encoding ribosomal protein, usually located in the chromosome (Yutin et al. 2012). Indeed, the core genes were more frequent in replicons classified as chromids, compared to megaplasmids.

Our results suggest an evolutionary approach to the definition of chromids and megaplasmids based on general, consistent trends in their sizes. One caveat is that the performed analysis might be affected by several factors such as errors in genome assemblies and taxonomical misclassification. Moreover, replicons in multipartite genomes may be affected by recombination between *rrna* repeats resulting in large inter-replicon translocations (Treangen et al. 2009; Bochkareva et al. 2018) or replicon fusion (Sozhamannan and Waldminghaus 2020; Bochkareva et al. 2018; Sozhamannan and Waldminghaus 2020; Mori and Kanaly 2022). Thus, in order to further classify secondary replicons, the history of the gene flow between replicons should be considered in addition to the content and architecture of extant genomes.

# Materials and Methods

## Genomes and taxonomy

We started with all representative strains available in the NCBI RefSeq database (O'Leary et al. 2016). If at least one representative strain had a secondary replicon of size exceeding 0.5 Mb, we added the genus to the analysis (Supplementary Table 1).

## Dataset filtering


12

Thus, we analyzed the genome composition for 627 species from 84 genera. We reduced the dataset by selecting single representatives for groups of closely related strains. For this we clustered the genomes with respect to their replicon length (the differences less than 5%). Then we filtered out genera with less than 6 representatives. Finally, the dataset comprised 547 species from 36 genera (Supplementary Table 2).

## Analysis of outliers

Genome assemblies of several strains were significantly different from members of the same taxonomic group.

Secondary replicons in *Prevotella copri* DSM 18205 strain FDAARGOS_1573 and *Prevotella* sp. E2-28 were smaller than 210 kb, while other members of this species have only one replicon so these strains were considered as not multipartite.

In *Burkholderia cenocepacia* strain 895, *Burkholderia cenocepacia* strain VC12802, *Burkholderia vietnamiensis* strain AU1233, *Cupriavidus metallidurans* strain Ni-2, *Cupriavidus* sp. KK10, and *Brucella pseudogrignonensis* strain K8 we observed fusion of chromosomes and secondary replicons (Supplementary Figure 8). Here, we used Sibelia (Minkin et al. 2013) software to align and compare the outliers and the closest complete genome of the same genera. To identify the pair for each of the outliers, we build a distance matrix using Mash (Ondov et al. 2016).

The assemblies of *Pseudoalteromonas spongiae* UST010723-006, *Pseudoalteromonas piratica* strain OCN003, and *Pseudoalteromonas spongiae* strain SAO4-4 are multipartite but the sizes differ dramatically from other members of this genus that may indicate recent recombination events or assembly errors.

All these genomes were excluded from the analysis.

## Statistical analysis

Linear regression analysis was performed with the scipy.stats.linregress Python module. To construct the Gaussian functions for each genus, we used the slope coefficient as the mean μ and the slope standard error as the standard error σ in the Gaussian function

$$\{1 / [\sigma \sqrt{(2\pi)}]\} \times \exp \{-½ [(x-\mu) / \sigma]^2\}.$$

## Calculation of abundance of multipartite genomes in genera

We considered a genome as multipartite if it had a secondary replicon larger than 0.5 Mb. If more than 75% of strains forming a species had multipartite genomes, the species was marked as multipartite (Supplementary Table 3). Then, for each genus, we calculated the fraction of species with multipartite genomes (Supplementary Table 4).

# Author statements

## Funding

The study was supported by RFBR via grant 20-54-14005 and Fonds zur Förderung der Wissenschaftlichen Forschung (FWF), Grant # I5127-B. The work of OOB is supported by

FWF, Grant # ESP 253-B. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Authors contribution

O.O.B. and M.S.G. conceived and designed the study. N.O.D., A.R., and Y. D. analyzed the data. N.O.D., O.O.B., and M.S.G. interpreted the results and wrote the graft. All authors read and approved the final version of the manuscript.

## Conflicts of interest

The authors declare that they have no competing interests.

## Ethical approval

Not applicable.

## Consent for publication

Not applicable.

## Acknowledgements

The project was initiated at the Summer School of Molecular and Theoretical Biology (SMTB-2021), supported by the Zimin Foundation.

# References

Agnoli, K., S. Schwager, S. Uehlinger, A. Vergunst, D. F. Viteri, D. T. Nguyen, P. A. Sokol, A. Carlier, and L. Eberl. 2012. “Exposing the Third Chromosome of Burkholderia Cepacia Complex Strains as a Virulence Plasmid.” *Molecular Microbiology* 83 (2). https://doi.org/10.1111/j.1365-2958.2011.07937.x.

Bennett, Gordon M., and Nancy A. Moran. 2013. “Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect.” *Genome Biology and Evolution* 5 (9): 1675.

Bochkareva, Olga O., Elena V. Moroz, Iakov I. Davydov, and Mikhail S. Gelfand. 2018. “Genome Rearrangements and Selection in Multi-Chromosome Bacteria Burkholderia Spp.” *BMC Genomics* 19 (1): 1–17.

Bottacini, Francesca, Mary O’connell Motherway, Eoghan Casey, Brian McDonnell, Jennifer Mahony, Marco Ventura, and Douwe van Sinderen. 2015. “Discovery of a Conjugative Megaplasmid in Bifidobacterium Breve.” *Applied and Environmental Microbiology* 81 (1): 166.

Burton, Jeremy P., Philip A. Wescombe, Jean M. Macklaim, Melissa H. C. Chai, Kyle MacDonald, John D. F. Hale, John Tagg, Gregor Reid, Gregory B. Gloor, and Peter A. Cadieux. 2013. “Persistence of the Oral Probiotic Streptococcus Salivarius M18 Is Dose Dependent and Megaplasmid Transfer Can Augment Their Bacteriocin Production and Adhesion Characteristics.” *PloS One* 8 (6). https://doi.org/10.1371/journal.pone.0065991.

Cazares, Adrian, Matthew P. Moore, James P. J. Hall, Laura L. Wright, Macauley Grimes, Jean-Guillaume Emond-Rhéault, Pisut Pongchaikul, et al. 2020. “A Megaplasmid Family Driving Dissemination of Multidrug Resistance in Pseudomonas.” *Nature*


14

*Communications* 11. https://doi.org/10.1038/s41467-020-15081-7.
diCenzo, G. C., A. Mengoni, and E. Perrin. 2019. “Chromids Aid Genome Expansion and Functional Diversification in the Family Burkholderiaceae.” *Molecular Biology and Evolution* 36 (3). https://doi.org/10.1093/molbev/msy248.
diCenzo, George C., and Turlough M. Finan. 2017. “The Divided Bacterial Genome: Structure, Function, and Evolution.” *Microbiology and Molecular Biology Reviews: MMBR* 81 (3). https://doi.org/10.1128/MMBR.00019-17.
Gerdes, S. Y., Scholle, J. W. Campbell, G. Balázsi, E. Ravasz, Daugherty, A. L. Somera, et al. 2003. “Experimental Determination and System Level Analysis of Essential Genes in Escherichia Coli MG1655.” *Journal of Bacteriology* 185 (19). https://doi.org/10.1128/JB.185.19.5673-5684.2003.
Gifford, I., A. Dasgupta, and J. E. Barrick. 2021. “Rates of Gene Conversions between Escherichia Coli Ribosomal Operons.” *G3* 11 (2). https://doi.org/10.1093/g3journal/jkaa002.
Guo, X., M. Flores, P. Mavingui, S. I. Fuentes, G. Hernández, G. Dávila, and R. Palacios. 2003. “Natural Genomic Design in Sinorhizobium Meliloti: Novel Genomic Architectures.” *Genome Research* 13 (8). https://doi.org/10.1101/gr.1260903.
Gupta, Radhey S., Sudip Patel, Navneet Saini, and Shu Chen. 2020. “Robust Demarcation of 17 Distinct Bacillus Species Clades, Proposed as Novel Bacillaceae Genera, by Phylogenomics and Comparative Genomic Analyses: Description of Robertmurraya Kyonggiensis Sp. Nov. and Proposal for an Emended Genus Bacillus Limiting It Only to the Members of the Subtilis and Cereus Clades of Species.” *International Journal of Systematic and Evolutionary Microbiology* 70 (11): 5753–98.
Hall, James P. J., João Botelho, Adrian Cazares, and David A. Baltrus. 2022. “What Makes a Megaplasmid?” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 377 (1842). https://doi.org/10.1098/rstb.2020.0472.
Han, Kui, Zhi-Feng Li, Ran Peng, Li-Ping Zhu, Tao Zhou, Lu-Guang Wang, Shu-Guang Li, et al. 2013. “Extraordinary Expansion of a Sorangium Cellulosum Genome from an Alkaline Milieu.” *Scientific Reports* 3. https://doi.org/10.1038/srep02101.
Harrison, P. W., R. P. Lower, N. K. Kim, and J. P. Young. 2010. “Introducing the Bacterial ‘Chromid’: Not a Chromosome, Not a Plasmid.” *Trends in Microbiology* 18 (4). https://doi.org/10.1016/j.tim.2009.12.010.
Hayakawa, Takaki, Teruo Tanaka, Kenji Sakaguchi, Noboru Otake, and Hiroshi Yonehara. 1979. “A LINEAR PLASMID-LIKE DNA IN STREPTOMYCES SP. PRODUCING LANKACIDIN GROUP ANTIBIOTICS.” *The Journal of General and Applied Microbiology* 25 (4): 255–60.
Horesh, Gal, Alyce Taylor-Brown, Stephanie McGimpsey, Florent Lassalle, Jukka Corander, Eva Heinz, and Nicholas R. Thomson. 2021. “Different Evolutionary Trends Form the Twilight Zone of the Bacterial Pan-Genome.” *Microbial Genomics* 7 (9): 000670.
Köstlbacher, Stephan, Astrid Collingro, Tamara Halter, Frederik Schulz, Sean P. Jungbluth, and Matthias Horn. 2021. “Pangenomics Reveals Alternative Environmental Lifestyles among Chlamydiae.” *Nature Communications* 12 (1): 1–15.
Minkin, Ilya, Anand Patel, Mikhail Kolmogorov, Nikolay Vyahhi, and Son Pham. 2013. “Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes.” *Algorithms in Bioinformatics*, 215–29.
Mori, J. F., and R. A. Kanaly. 2022. “Natural Chromosome-Chromid Fusion across rRNA Operons in a Burkholderiaceae Bacterium.” *Microbiology Spectrum* 10 (1). https://doi.org/10.1128/spectrum.02225-21.
O’Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids Research* 44 (D1). https://doi.org/10.1093/nar/gkv1189.
Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17 (1): 1–14.

Santos, F. B., and L. E. Del-Bem. 2022. “The Evolution of tRNA Copy Number and Repertoire in Cellular Life.” *Genes* 14 (1). https://doi.org/10.3390/genes14010027.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, et al. 2020. “NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools.” *Database: The Journal of Biological Databases and Curation* 2020 (January). https://doi.org/10.1093/database/baaa062.

Sozhamannan, S., and T. Waldminghaus. 2020. “Exception to the Exception Rule: Synthetic and Naturally Occurring Single Chromosome Vibrio Cholerae.” *Environmental Microbiology* 22 (10). https://doi.org/10.1111/1462-2920.15002.

Suwanto, A., and S. Kaplan. 1989. “Physical and Genetic Mapping of the Rhodobacter Sphaeroides 2.4.1 Genome: Presence of Two Unique Circular Chromosomes.” *Journal of Bacteriology* 171 (11). https://doi.org/10.1128/jb.171.11.5850-5859.1989.

Treangen, Todd J., Anne-Laure Abraham, Marie Touchon, and Eduardo P. C. Rocha. 2009. “Genesis, Effects and Fates of Repeats in Prokaryotic Genomes.” *FEMS Microbiology Reviews* 33 (3): 539–71.

Weiser, Rebecca, Angharad E. Green, Matthew J. Bull, Edward Cunningham-Oakes, Keith A. Jolley, Martin C. J. Maiden, Amanda J. Hall, et al. 2019. “Not All Pseudomonas Aeruginosa Are Equal: Strains from Industrial Sources Possess Uniquely Large Multireplicon Genomes.” *Microbial Genomics* 5 (7). https://doi.org/10.1099/mgen.0.000276.

Yang, L. L., Z. Jiang, Y. Li, E. T. Wang, and X. Y. Zhi. 2020. “Plasmids Related to the Symbiotic Nitrogen Fixation Are Not Only Cooperated Functionally but Also May Have Evolved over a Time Span in Family Rhizobiaceae.” *Genome Biology and Evolution* 12 (11). https://doi.org/10.1093/gbe/evaa152.

Yutin, Natalya, Pere Puigbò, Eugene V. Koonin, and Yuri I. Wolf. 2012. “Phylogenomics of Prokaryotic Ribosomal Proteins.” *PloS One* 7 (5). https://doi.org/10.1371/journal.pone.0036972.