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

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

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

OOB Present affiliation: CUBE, Department of Microbiology, University of Vienna, 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, to date, 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. Moreover, such secondary replicons are usually present in only a low fraction of the genus’ species. Hence, such replicons behave as ‘megaplasmids’. A mixed behavior was observed in symbiotic genera from the *Rhizobiaceae* family (*Rhizobium, Agrobacterium*, *Azospirillum, and Ensifer*) with

large, essential secondary replicons of stable size. These results may provide a general framework for understanding the evolution of genome complexity in prokaryotes.

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 *Rhodobacter* genome were described (Suwanto& Kaplan, 1989). This type of genome organization was named multipartite genomes (diCenzo&Finan, 2017), with about 10% of sequenced bacterial species being of that type (diCenzo&Finan, 2017).

When a bacterial genome is formed by several large elements, the largest one, that holds the main fraction of housekeeping 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&Finan, 2017)), but in the absence of universal definitions this classification is mainly based on arbitrary thresholds (Bottacini et al., 2015; Burton et al., 2013; Cazares et al., 2020; diCenzo&Finan, 2017; Weiser et al., 2019). 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 (Bottacini et al., 2015; Burton et al., 2013; Cazares et al., 2020; diCenzo&Finan, 2017; Weiser et al., 2019). Genomes of some species may hold several secondary replicons of different types, e.g. *Burkholderiacepacia* 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&Finan, 2017). The bacterial genome size varies from 112 kb for obligate symbionts (Bennett & Moran, 2013) to ~15 Mb for some free-living organisms (Han et al., 2013). While multipartite genomes are larger on average, single-chromosome genomes exist, and only one third of genomes larger than 6Mb are multipartite. Therefore, the genome size is not uniquely driven by multiparticity.

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 (diCenzo&Finan, 2017), we define 'multipartite genome' as a genome with at least one secondary replicon longer than 500 Kbp. 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*)**.

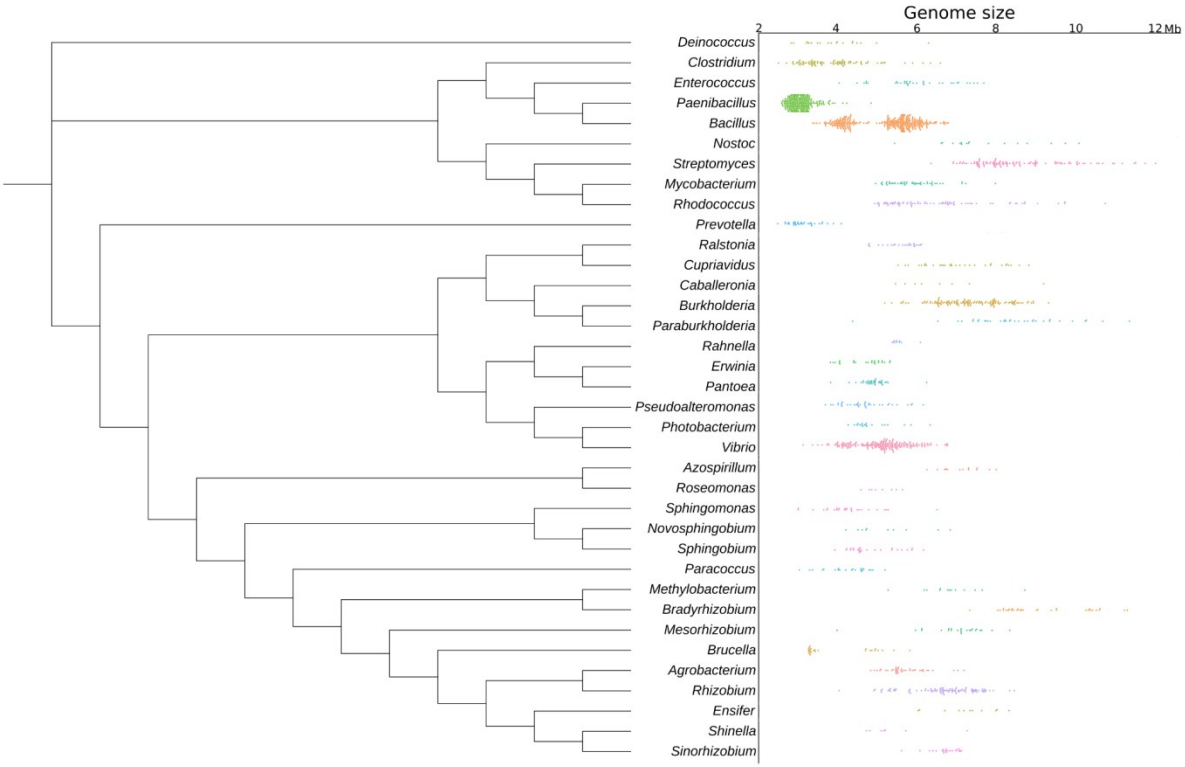


Figure 1. Genome sizes of strains from genera that have members with multipartite genomes (see Methods). The cladogram topology follows the NCBI taxonomy (Schoch et al., 2020).

Replicon sizes

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**(*Supplementaryfigure 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 of both replicons ***(Figure 2).***

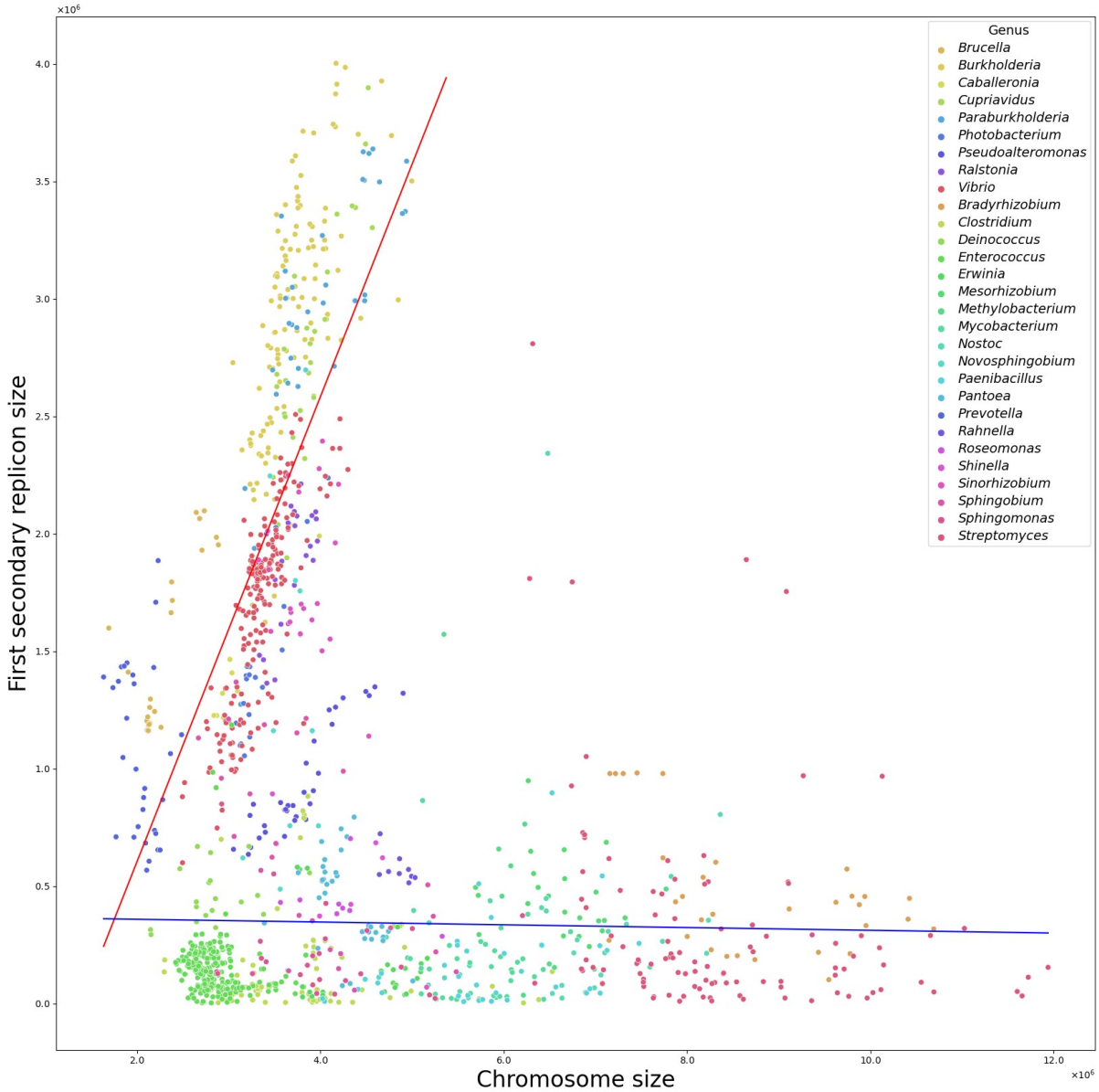


Figure 2. The chromosome size (horizontal axis) and the size of the largest secondary replicon (vertical axis) for strains with multipartite genomes from 36 genera. The dot color shows the bacterial genus (see the legend).

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.

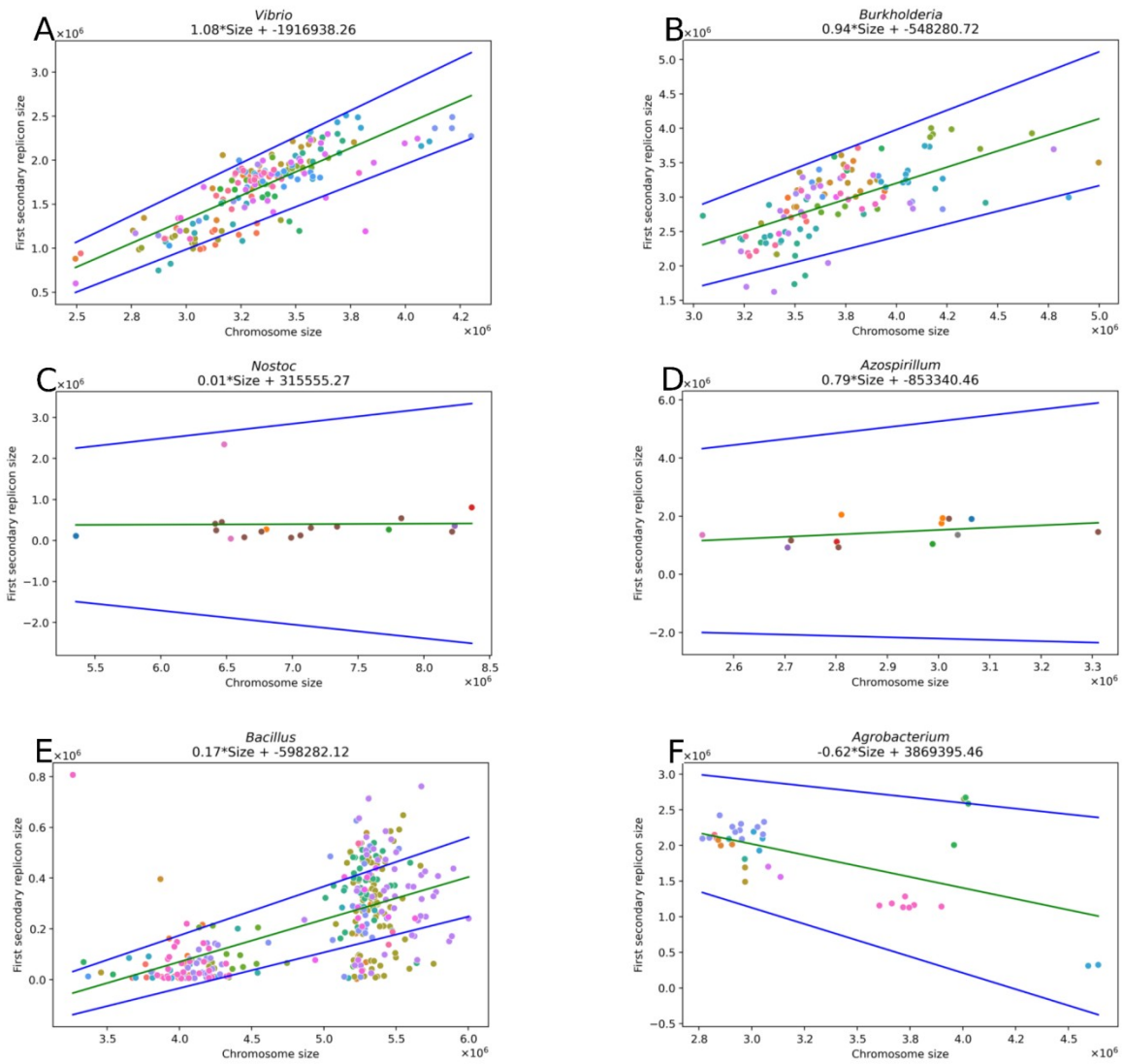


Figure 3. Linear approximation of the relationships between sizes of the chromosome and the largest secondary replicon in: (A, B) *Vibrio* (r = 0.65, slope coefficient 1.08 ± 0.12) and *Burkholderia* (r = 0.8, slope coefficient 0.94 ± 0.2), genera with a strong positive correlation; (C) *Nostoc*, genus 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. Color code: green — strong positive correlation; blue — approximately constant size of the secondary replicon; red — non-homogeneous genera (see the text); yellow — non-significant positive correlation.

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

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 ***(Supplementary figure 4, 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.

For nine genera where the correlation between the chromosome size and the first secondary replicon size was observed, we performed the same analysis for the second-largest secondary replicons ***(Supplementary figure 5)***; no correlation was observed.

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.

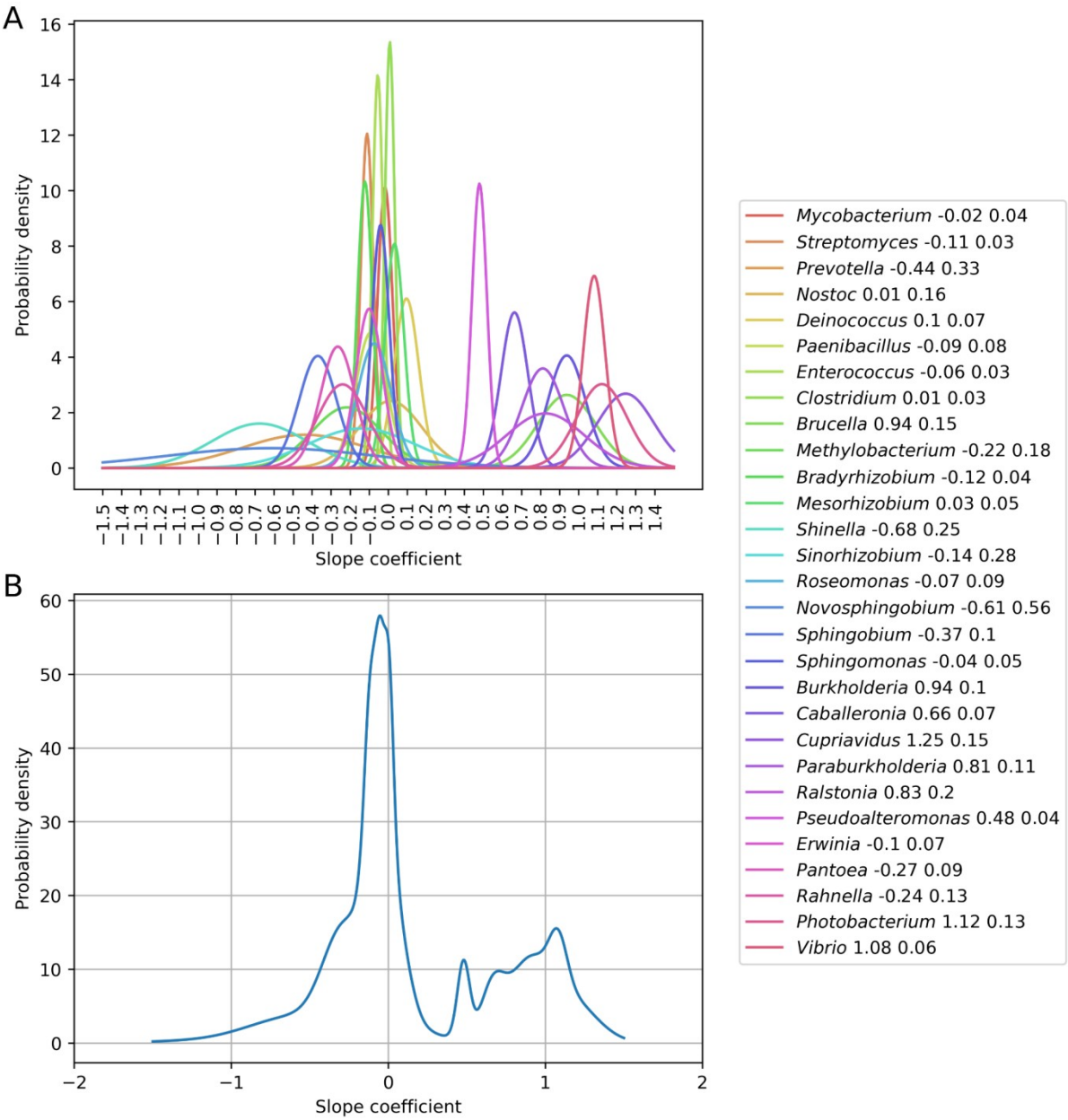


Figure 4. Summary of the regression analysis. (A) Gaussian functions derived from the slope coefficients (see the text), each colored line corresponds to a genus (listed in the legend; the numbers are the mean and the standard deviation). (B) Sum of the Gaussian functions.

Discussion

Positive correlation between the chromosome size and the secondary replicon sizes was observed in nine genera from four families: *Burkholderiaceae (Burkholderia, Paraburkholderia, Ralstonia, Caballeronia, Cupriavidus), Vibrionaceae (Vibrio, Photobacterium),Pseudoalteromonadaceae (Pseudoalteromonas),* and *Brucellaceae (Brucella).* It might be explained by acquisition and stabilization of a secondary replicon in the common ancestor of each family.

The second, more common trend demonstrates the absence of such correlation. This trend was mainly observed in genera with low fraction of multipartite genomes. It was also seen in megaplasmids of symbiotic genera from *Rhizobiaceae*(*Agrobacterium, Sinorhizobium, Rhizobium, Ensifer,* and*Shinella*), and *Azospirillaceae (Azospirillum)*.

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 symbionts having smaller genomes with smaller fractions of the periphery (strain-specific) genes (Horesh et al., 2021; Köstlbacher 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 5A-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 5E*).** *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.

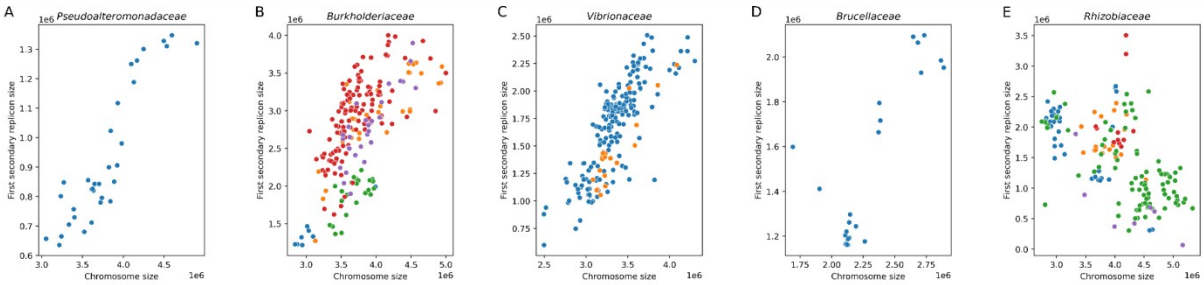


Figure 5. Relationships between the chromosome size and the largest secondary replicon size for five families with chromids: (A) *Pseudoalteromonadaceae* (*Pseudoalteromonas*), (B) *Burkholderiaceae* (*Caballeronia*, *Paraburkholderia*, *Ralstonia*, *Burkholderia*, *Cupriavidus*), (C) *Vibrionaceae* (*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 (Hall et al., 2022; Harrison et al., 2010). 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. Our results suggest an evolutionary definition based on the genome architecture and different roles of secondary replicons in multipartite genomes.

Meanwhile, 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 (Bochkareva et al., 2018; Treangen et al., 2009) or replicon fusion (Sozhamannan&Waldminghaus, 2020, Mori &Kanaly, 2022, Bochkareva et al., 2018). Thus,

the history of the gene flow between replicons should be considered in addition to the extant genome content and architecture in order to classify secondary replicons.

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

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 6)***. 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 zurFörderung der wissenschaftlichenForschung (FWF), Grant # I5127-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

Bennett, G. M., & Moran, N. A. (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, O. O., Moroz, E. V., Davydov, I. I., & Gelfand, M. S. (2018). Genome rearrangements and selection in multi-chromosome bacteria Burkholderia spp. *BMC Genomics*, *19*(1), 1–17.

Bottacini, F., Motherway, M. O., Casey, E., McDonnell, B., Mahony, J., Ventura, M., & van Sinderen, D. (2015). Discovery of a Conjugative Megaplasmid in Bifidobacterium breve. *Applied and Environmental Microbiology*, *81*(1), 166.

Burton, J. P., Wescombe, P. A., Macklaim, J. M., Chai, M. H. C., MacDonald, K., Hale, J. D. F., Tagg, J., Reid, G., Gloor, G. B., & Cadieux, P. A. (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, A., Moore, M. P., Hall, J. P. J., Wright, L. L., Grimes, M., Emond-Rhéault, J.-G., Pongchaikul, P., Santanirand, P., Levesque, R. C., Fothergill, J. L., & Winstanley, C. (2020). A megaplasmid family driving dissemination of multidrug resistance in Pseudomonas. *Nature Communications*, *11*. https://doi.org/10.1038/s41467-020-15081-7

diCenzo, G. C., & Finan, T. M. (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

diCenzo, G. C., Mengoni, A., & Perrin, E. (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

Gerdes, S. Y., Scholle, Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A. L., … Osterman, A. L. (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

Guo, X., Flores, M., Mavingui, P., Fuentes, S. I., Hernández, G., Dávila, G., & Palacios, R. (2003). Natural genomic design in Sinorhizobium meliloti: novel genomic architectures. *Genome Research*, *13*(8). https://doi.org/10.1101/gr.1260903

Gupta, R. S., Patel, S., Saini, N., & Chen, S. (2020). Robust demarcation of 17 distinct Bacillus species clades, proposed as novel Bacillaceae genera, by phylogenomics and comparative genomic analyses: description of Robertmurrayakyonggiensis 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), 5753–5798.

Hall, J. P. J., Botelho, J., Cazares, A., & Baltrus, D. A. (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, K., Li, Z.-F., Peng, R., Zhu, L.-P., Zhou, T., Wang, L.-G., Li, S.-G., Zhang, X.-B., Hu, W., Wu, Z.-H., Qin, N., & Li, Y.-Z. (2013). Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. *Scientific Reports*, *3*. https://doi.org/10.1038/srep02101

Harrison, P. W., Lower, R. P., Kim, N. K., & Young, J. P. (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, T., Tanaka, T., Sakaguchi, K., Otake, N., &Yonehara, H. (1979). A LINEAR PLASMID-LIKE DNA IN STREPTOMYCES SP. PRODUCING LANKACIDIN GROUP ANTIBIOTICS. *The Journal of General and Applied Microbiology*, *25*(4), 255–260.

Horesh, G., Taylor-Brown, A., McGimpsey, S., Lassalle, F., Corander, J., Heinz, E., & Thomson, N. R. (2021). Different evolutionary trends form the twilight zone of the bacterial pan-genome. *Microbial Genomics*, *7*(9), 000670.

Köstlbacher, S., Collingro, A., Halter, T., Schulz, F., Jungbluth, S. P., & Horn, M. (2021). Pangenomics reveals alternative environmental lifestyles among chlamydiae. *Nature Communications*, *12*(1), 1–15.

Minkin, I., Patel, A., Kolmogorov, M., Vyahhi, N., & Pham, S. (2013). Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes. *Algorithms in Bioinformatics*, 215–229.

Mori, J. F., & Kanaly, R. A. (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., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (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, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., &Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation*, *2020*. https://doi.org/10.1093/database/baaa062

Sozhamannan, S., & Waldminghaus, T. (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

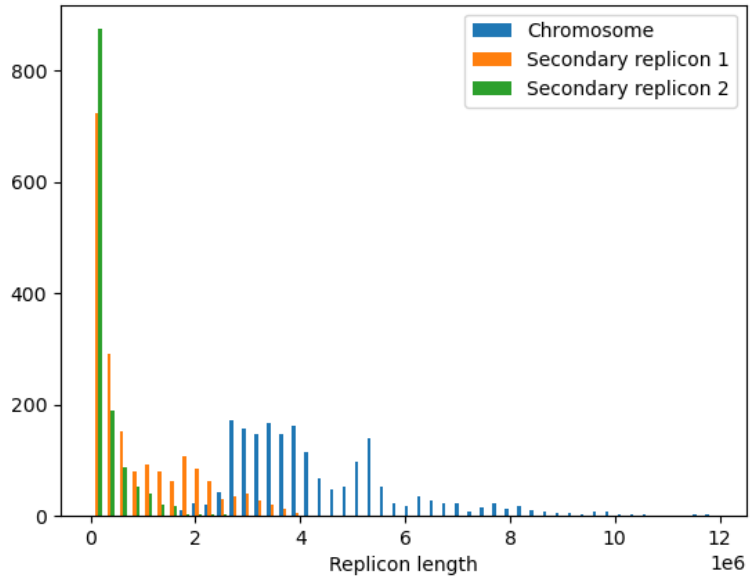
Suwanto, A., & Kaplan, S. (1989). Physical and genetic mapping of the Rhodobacter sphaeroides 2.4.1 genome: presence of two unique circular chromosomes. *Journal of Bacteriology*, *171*(11). https://doi.org/10.1128/jb.171.11.5850-5859.1989

Treangen, T. J., Abraham, A.-L., Touchon, M., & Rocha, E. P. C. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews*, *33*(3), 539–571.

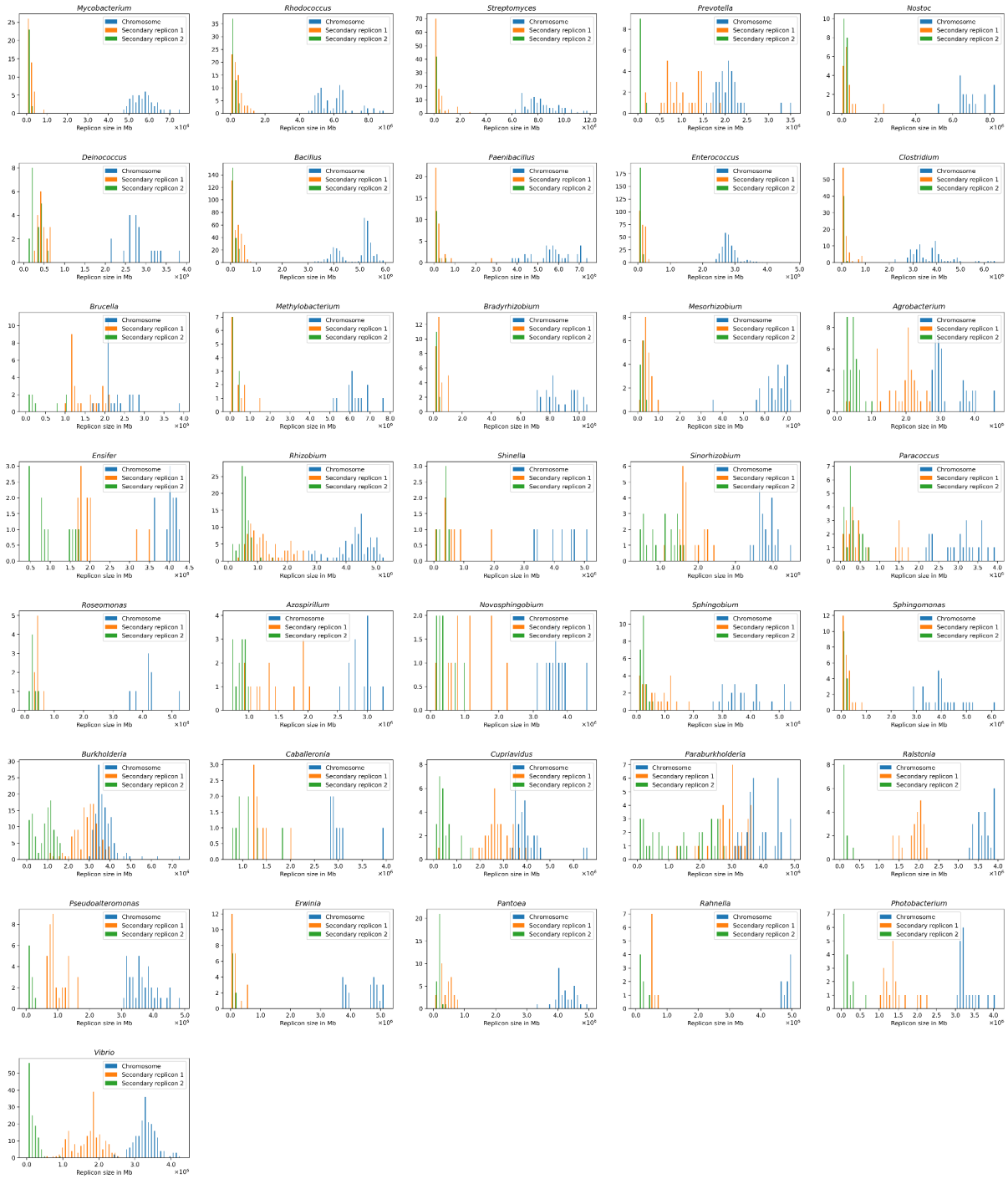
Weiser, R., Green, A. E., Bull, M. J., Cunningham-Oakes, E., Jolley, K. A., Maiden, M. C. J., Hall, A. J., Winstanley, C., Weightman, A. J., Donoghue, D., Amezquita, A., Connor, T. R., & Mahenthiralingam, E. (2019). Not all Pseudomonas aeruginosa are equal: strains from industrial sources possess uniquely large multireplicon genomes. *Microbial Genomics*, *5*(7). https://doi.org/10.1099/mgen.0.000276

Yang, L. L., Jiang, Z., Li, Y., Wang, E. T.,& Zhi, X. Y. (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

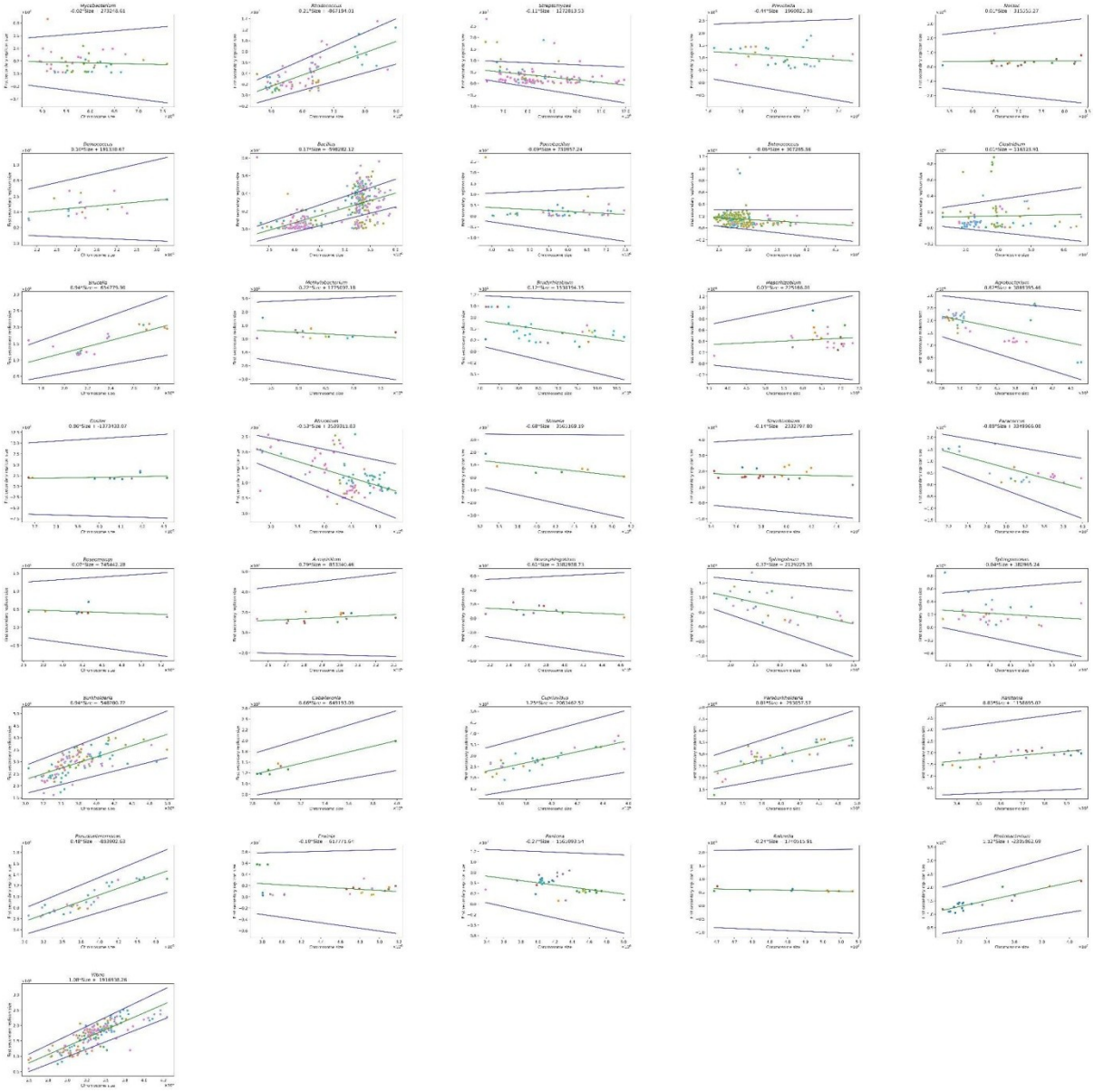
Supplementary information



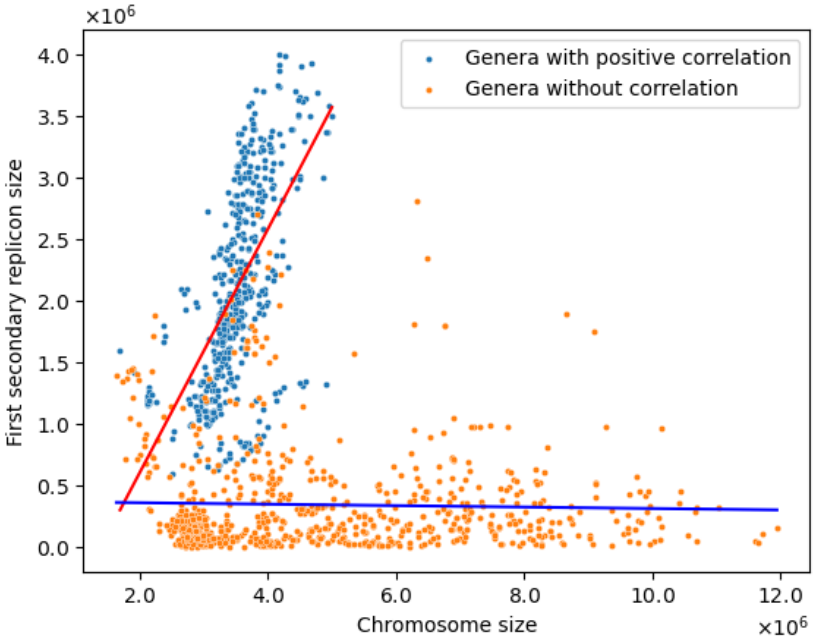
Supplementary figure 1. Replicon sizes in the 36 genera with multipartite genomes. Blue shows chromosomes; orange, the largest secondary replicon; green, the second-largest recondary replicon.



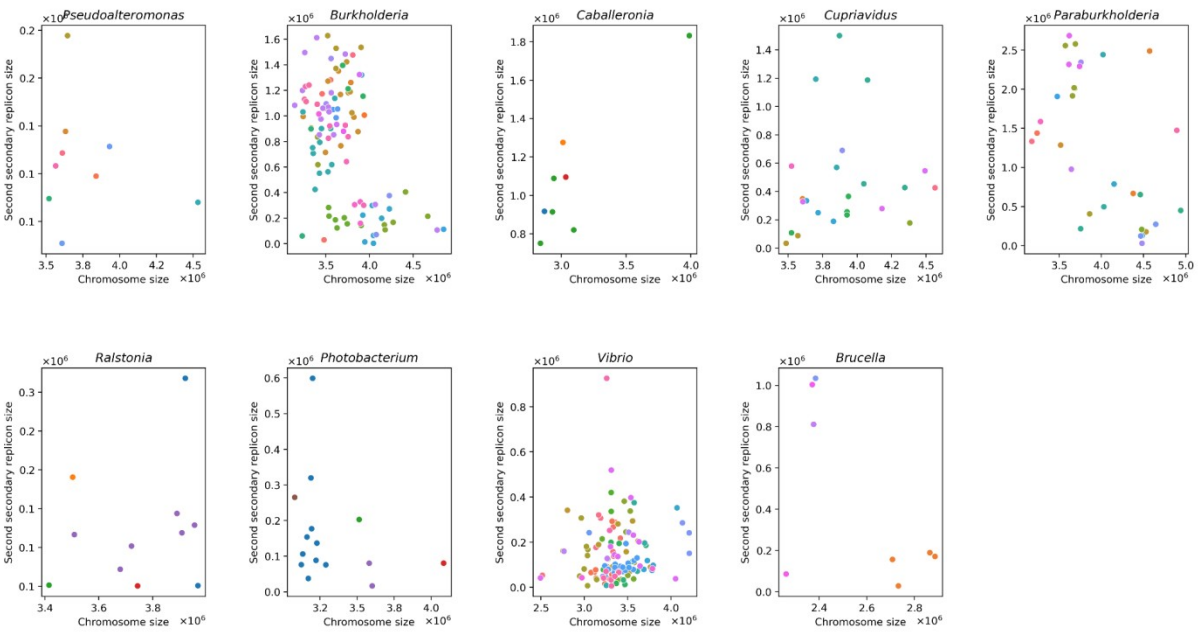
Supplementary figure 2. Replicon sizes in the 36 genera separately. Notation as in Supplementary figure 1.



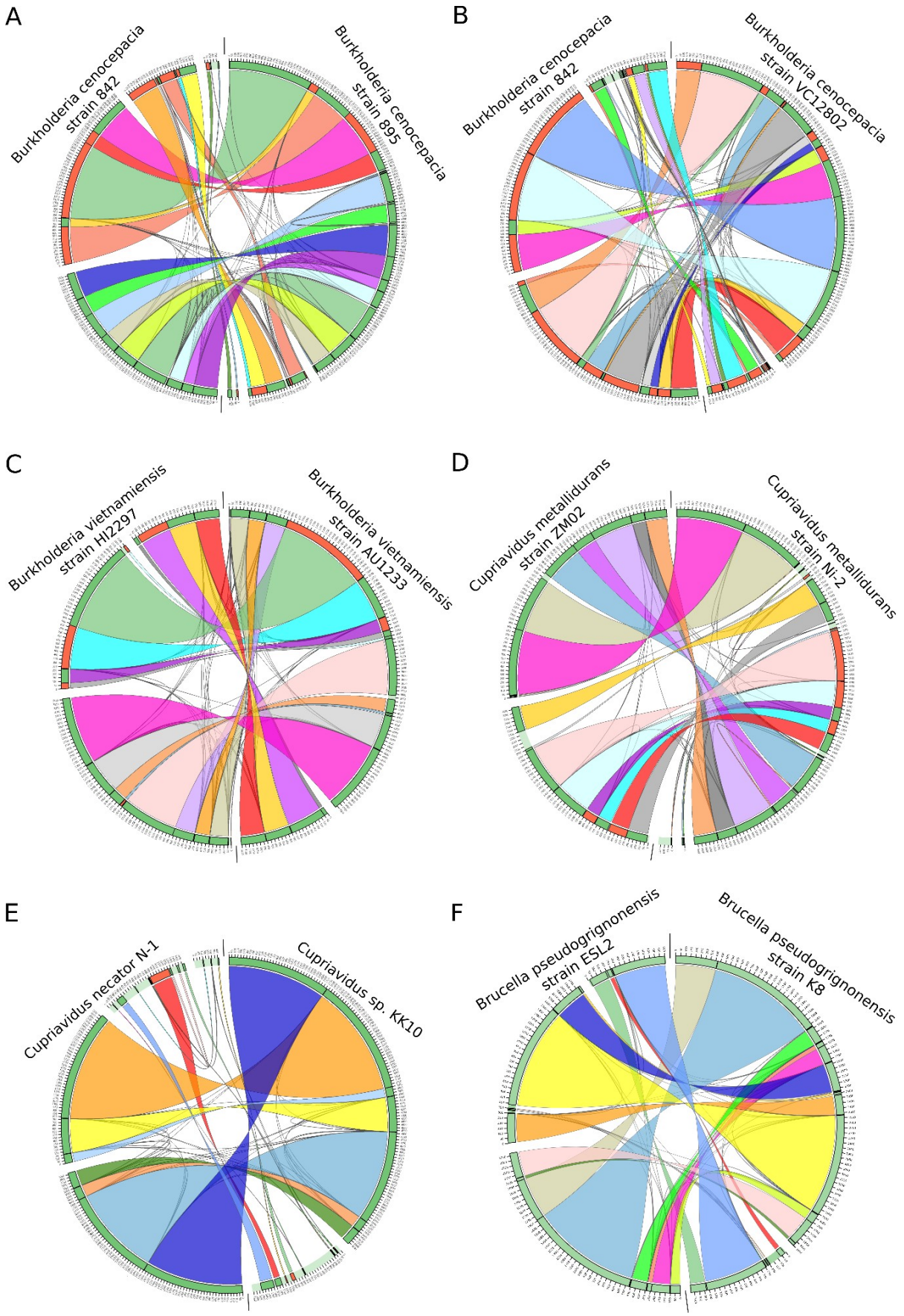
Supplementary figure 3. Linear approximation of the relationships between sizes of the chromosome and the largest secondary replicon in the genera with multipartite genomes. Different species in a genus are marked by different colors. Green line reflects the resulting slope coefficient, blue lines reflect the 95% confidence interval for the slope. The slope coefficients and intercepts are shown in the titles.



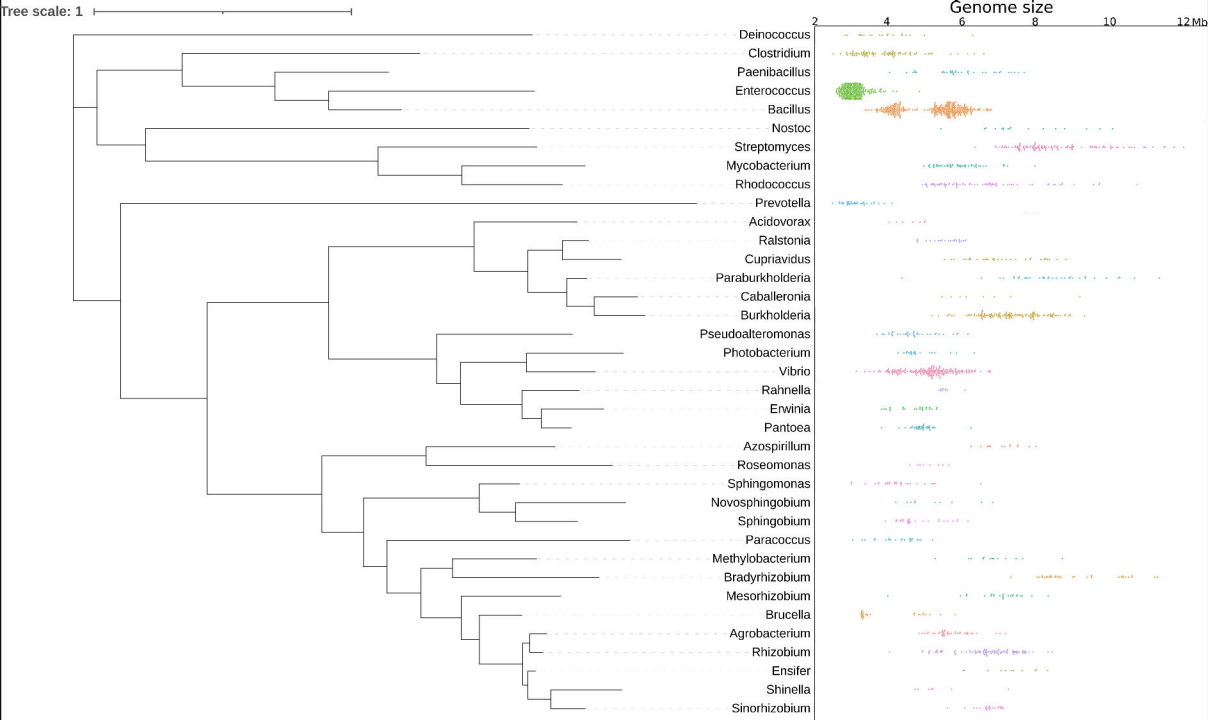
Supplementary figure 4. The chromosome size (horizontal axis) and the size of the largest secondary replicon (vertical axis) for strains from the genera with significant positive correlation (blue dots) and non-significant or no correlation (orange dots). Linear approximation of the relationships between sizes of the chromosome and the largest secondary replicon was shown separately in each group (red and blue lines correspondingly).

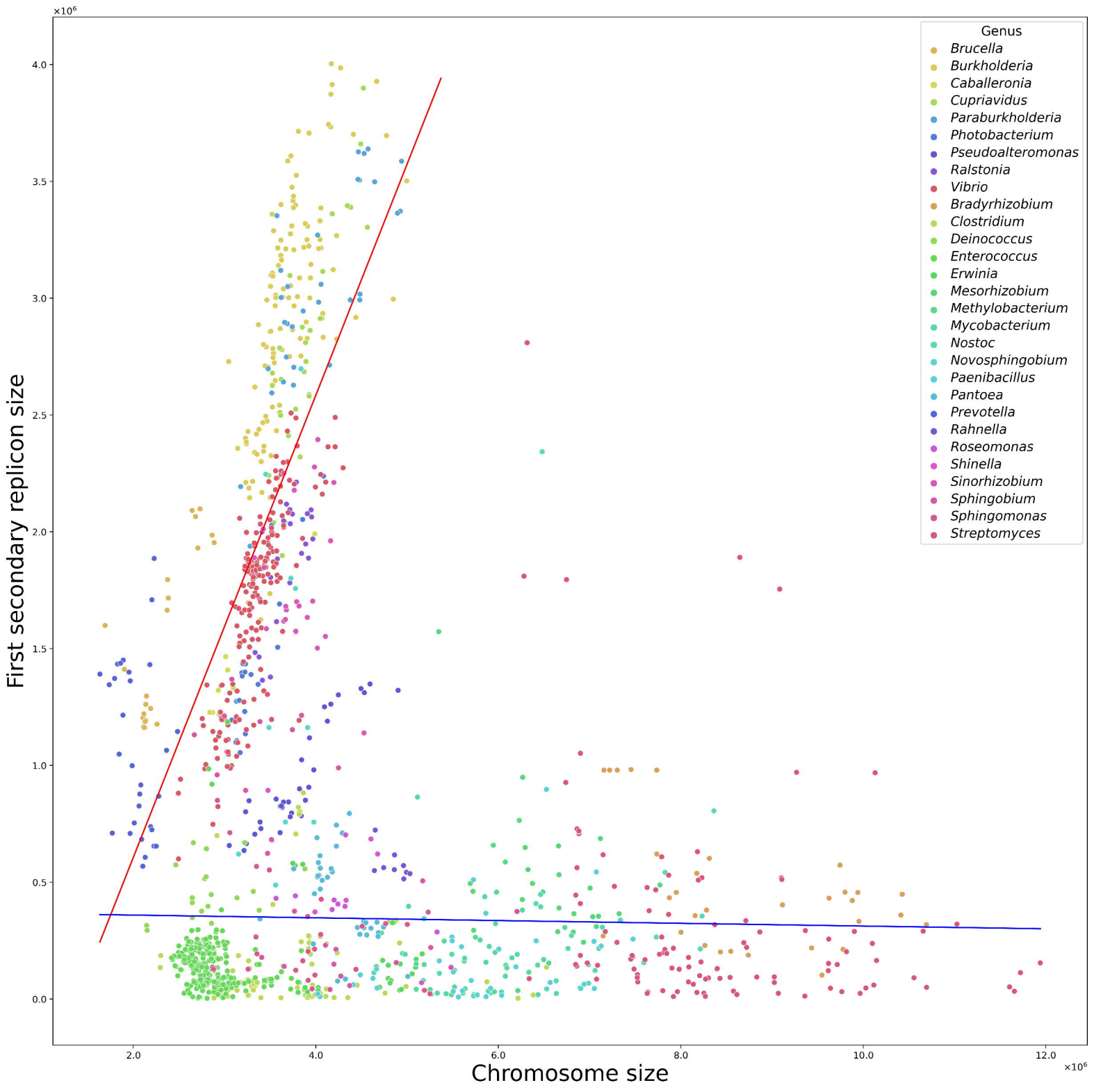


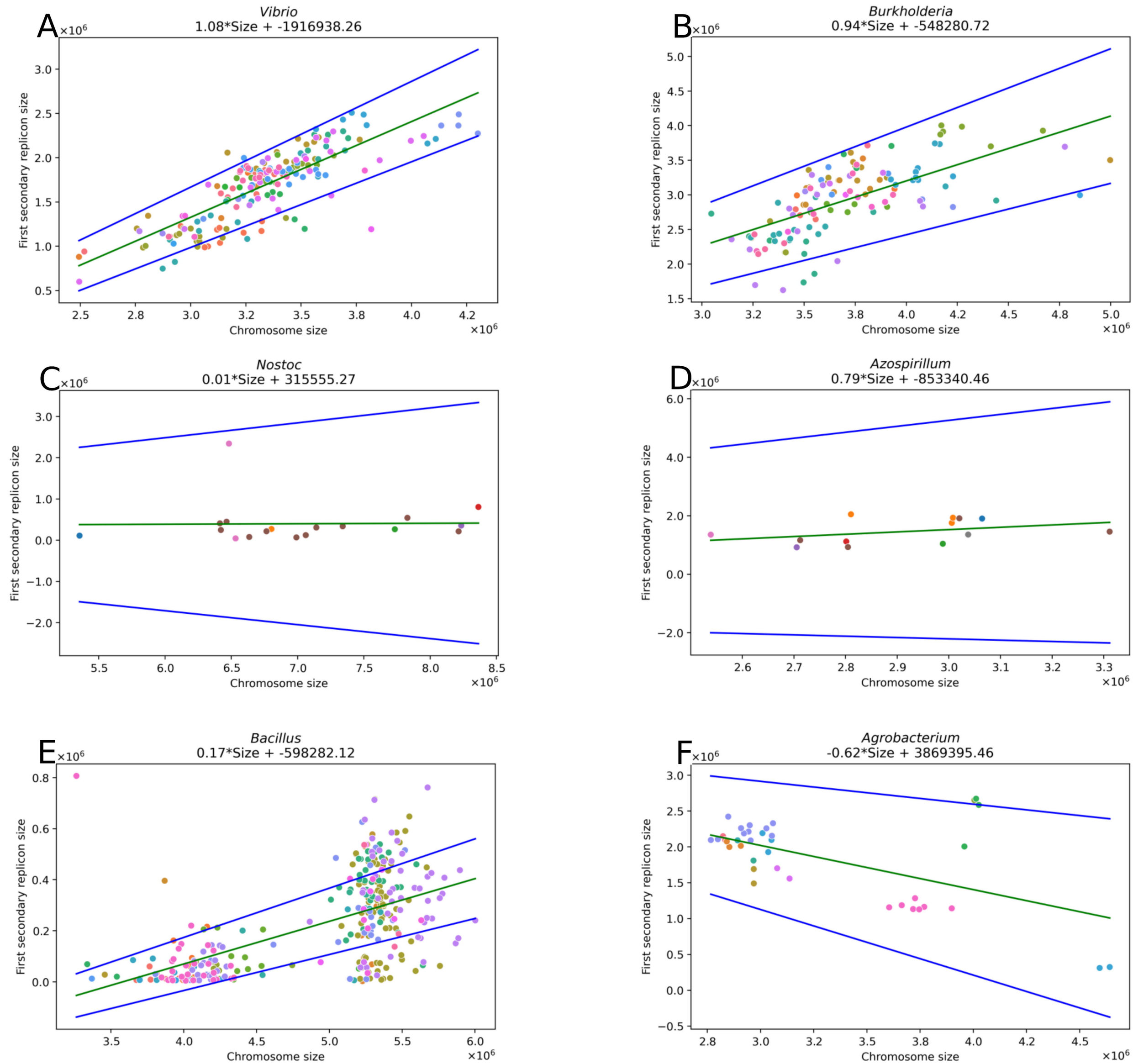
Supplementary figure5. Relationships between the chromosome and the second secondary replicon sizes for eight genera with the correlation between the chromosome and the largest secondary replicon was being observed. Different species in a genus are marked by colors.

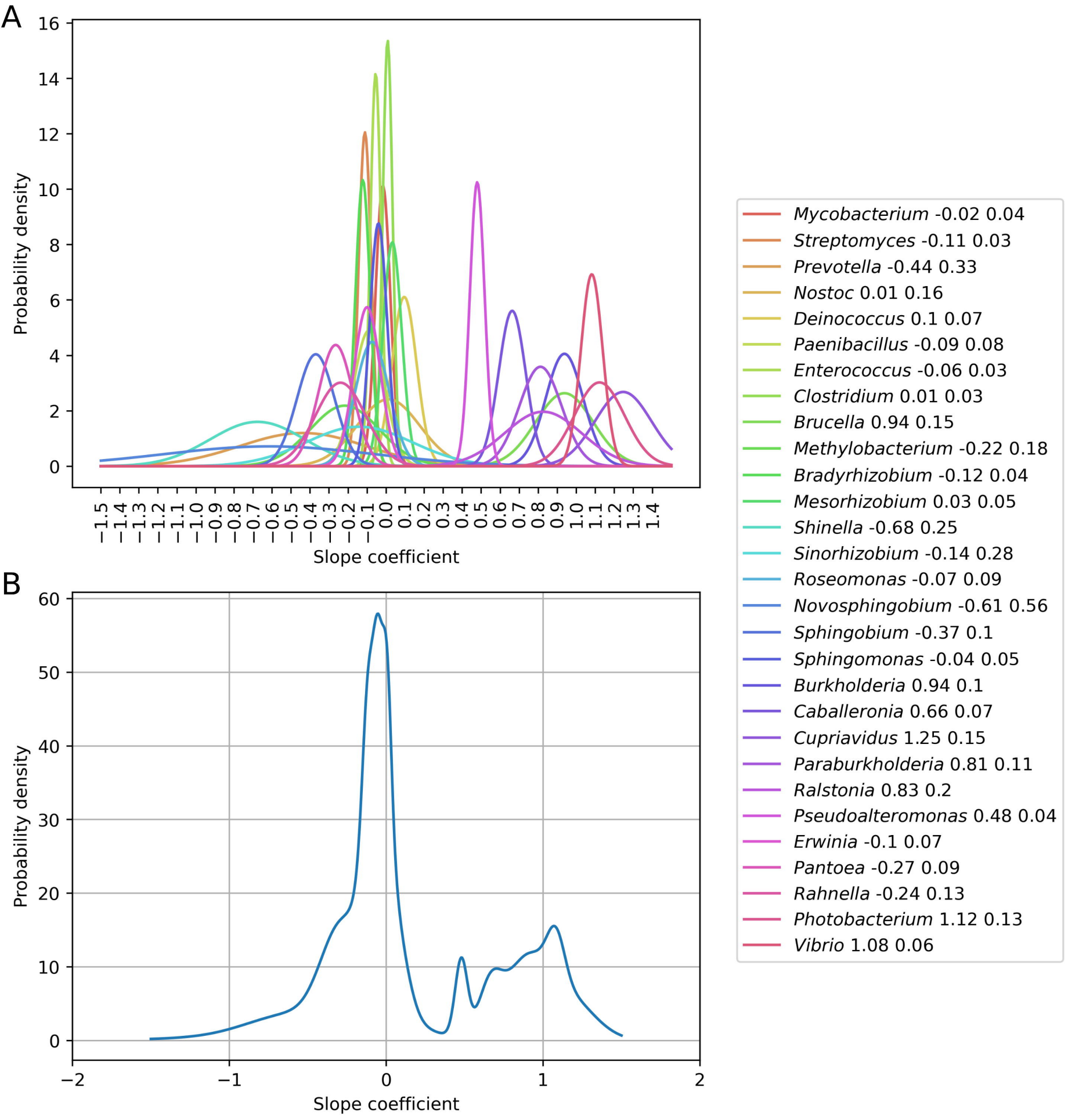

A
Burkholderia cenocepacia strain 842
Burkholderia cenocepacia strain 895
B
Burkholderia cenocepacia strain 842
Burkholderia cenocepacia strain VC12802
C
Burkholderia vietnamiensis strain HI2297
Burkholderia vietnamiensis strain AU1233
D
Cupriavidus metallidurans strain ZM02
Cupriavidus metallidurans strain Ni-2
E
Cupriavidus necator N-1
Cupriavidus sp. KK10
F
Brucella pseudogrignonensis strain ESL2
Brucella pseudogrignonensis strain K8

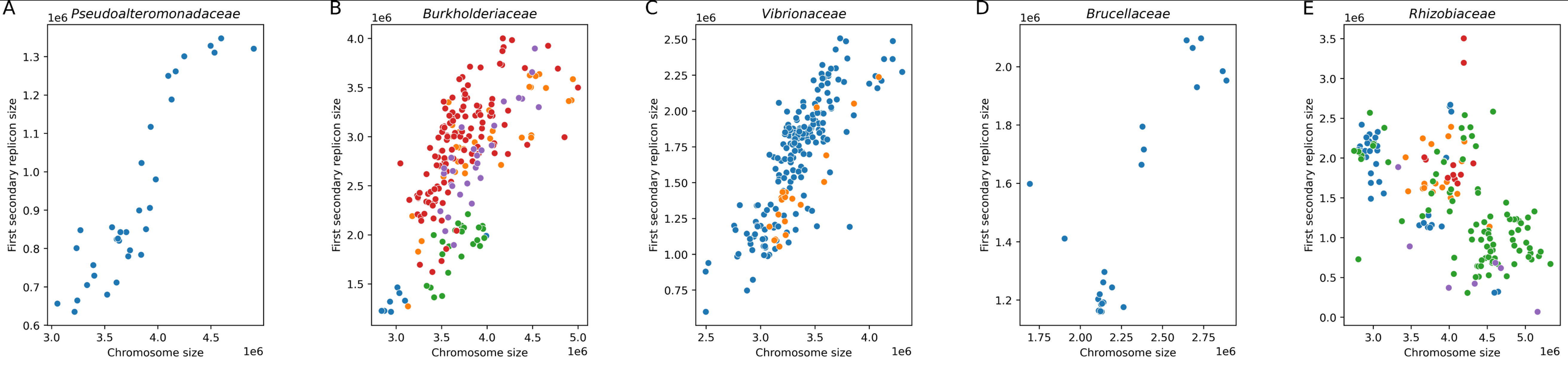
Supplementary figure 6. Pairwise comparisons of strains using circos diagrams (A) Burkholderia cenocepacia strain 895 and *Burkholderia cenocepacia* strain 842; (B) *Burkholderia cenocepacia* strain VC12802 and *Burkholderia cenocepacia* strain 842; (C) *Burkholderia vietnamiensis* strain AU1233 and *Burkholderia vietnamiensis* strain HI2297; (D) *Cupriavidus metallidurans* strain Ni-2 and *Cupriavidus metallidurans* strain ZM02; (E) *Cupriavidus* sp. KK10 and *Cupriavidus necator* N-1.


Tree scale: 1
Genome size
2
4
6
8
10
12 Mb
Deinococcus
Clostridium
Paenibacillus
Enterococcus
Bacillus
Nostoc
Streptomyces
Mycobacterium
Rhodococcus
Prevotella
Acidovorax
Ralstonia
Cupriavidus
Paraburkholderia
Caballeronia
Burkholderia
Pseudoalteromonas
Photobacterium
Vibrio
Rahnella
Erwinia
Pantoea
Azospirillum
Roseomonas
Sphingomonas
Novosphingobium
Sphingobium
Paracoccus
Methylobacterium
Bradyrhizobium
Mesorhizobium
Brucella
Agrobacterium
Rhizobium
Ensifer
Shinella
Sinorhizobium


×10^6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
First secondary replicon size
2.0
4.0
6.0
8.0
10.0
12.0
×10^6
Chromosome size
Genus
Brucella
Burkholderia
Caballeronia
Cupriavidus
Paraburkholderia
Photobacterium
Pseudoalteromonas
Ralstonia
Vibrio
Bradyrhizobium
Clostridium
Deinococcus
Enterococcus
Erwinia
Mesorhizobium
Methylobacterium
Mycobacterium
Nostoc
Novosphingobium
Paenibacillus
Pantoea
Prevotella
Rahnella
Roseomonas
Shinella
Sinorhizobium
Sphingobium
Sphingomonas
Streptomyces


A
Vibrio
1.08*Size + -1916938.26
First secondary replicon size
Chromosome size
×10^6
B
Burkholderia
0.94*Size + -548280.72
First secondary replicon size
Chromosome size
×10^6
C
Nostoc
0.01*Size + 315555.27
First secondary replicon size
Chromosome size
×10^6
D
Azospirillum
0.79*Size + -853340.46
First secondary replicon size
Chromosome size
×10^6
E
Bacillus
0.17*Size + -598282.12
First secondary replicon size
Chromosome size
×10^6
F
Agrobacterium
-0.62*Size + 3869395.46
First secondary replicon size
Chromosome size
×10^6


A
Probability density
Slope coefficient
B
Probability density
Slope coefficient
Mycobacterium -0.02 0.04
Streptomyces -0.11 0.03
Prevotella -0.44 0.33
Nostoc 0.01 0.16
Deinococcus 0.1 0.07
Paenibacillus -0.09 0.08
Enterococcus -0.06 0.03
Clostridium 0.01 0.03
Brucella 0.94 0.15
Methylobacterium -0.22 0.18
Bradyrhizobium -0.12 0.04
Mesorhizobium 0.03 0.05
Shinella -0.68 0.25
Sinorhizobium -0.14 0.28
Roseomonas -0.07 0.09
Novosphingobium -0.61 0.56
Sphingobium -0.37 0.1
Sphingomonas -0.04 0.05
Burkholderia 0.94 0.1
Caballeronia 0.66 0.07
Cupriavidus 1.25 0.15
Paraburkholderia 0.81 0.11
Ralstonia 0.83 0.2
Pseudoalteromonas 0.48 0.04
Erwinia -0.1 0.07
Pantoea -0.27 0.09
Rahnella -0.24 0.13
Photobacterium 1.12 0.13
Vibrio 1.08 0.06


A
Pseudoalteromonadaceae
B
Burkholderiaceae
C
Vibrionaceae
D
Brucellaceae
E
Rhizobiaceae
1e6
First secondary replicon size
Chromosome size
1e6