Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

The organization of bacterial genomes: Towards understanding the interplay between structure and function

Daan J. W. Brocken[1], Mariliis Tark-Dame[2] and Remus T. Dame[1,3]

Abstract

Genomes are arranged in a confined space in the cell, the nucleoid or nucleus. This arrangement is hierarchical and dynamic, and follows DNA/chromatin-based transactions or environmental conditions. Describing the interplay between local genome structure and gene activity is a long-standing quest in biology. Here, we focus on systematic studies correlating bacterial genome folding and function. Parallels on organizational similarities with eukaryotes are drawn. The biological relevance of hierarchical units in bacterial genome folding and the causal relationship between genome folding and its activity is unclear. We discuss recent quantitative approaches to tackle these questions. Moreover, we sketch a perspective of experiments necessary to iteratively and systematically build, test and improve structure–function models of bacterial chromatin.

Addresses

[1] Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, Leiden, The Netherlands

[2] Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

[3] Centre for Microbial Cell Biology, Leiden University, Leiden, The Netherlands

Corresponding author: Dame, Remus T. (rtdame@chem.leidenuniv.nl)

Email address: d.j.w.brocken@lic.leidenuniv.nl (D.J.W. Brocken), m.tark@uva.nl (M. Tark-Dame)

Current Opinion in Systems Biology 2018, **8**:137–143

This review comes from a themed issue on **Single cell and systems biology (2018)**

Edited by **Frank J. Bruggeman** and **Peter Swain**

For a complete overview see the Issue and the Editorial

Available online 27 February 2018

https://doi.org/10.1016/j.coisb.2018.02.007

Keywords

Bacterial genome organization, Nucleoid, Chromatin dynamics, Chromosome conformation capture, FROS, CRISPR/dCas9, High-resolution imaging.

Introduction

Genomes of all organisms, bacteria, archaea and eukaryotes, are arranged in the cell in a confined space, the nucleoid or nucleus. This arrangement is dynamic allowing for DNA transactions such as replication, transcription and repair to occur at appropriate times. A spectrum of mechanisms is involved in physically compacting and functionally organizing genomes in cells. Although at first sight the organization of the genomes of bacteria and eukaryotes may appear diverse, common principles are recognized [1]. The proteins providing structural and functional organization are generally not conserved at the protein sequence level. Nevertheless, several types of conserved structural features are evident [1]. Bacterial H-NS-family proteins, SMC proteins, and eukaryotic insulator proteins bridge DNA to form loops at different length scales [2–4]. In bacteria DNA decorated with architectural proteins is folded in looped structures [5], whereas in eukaryotes nucleosomal fibres, in which DNA is wrapped around histone proteins, are arranged into loops [6]. Although in eukaryotes much of genome regulation occurs at the level of nucleosomes (via histone tail modifications and nucleosome density) [7,8], at a coarse grained, structural level such molecular details are irrelevant. At a larger scale both in bacteria and eukaryotes, loops are arranged into structural domains, defined by genome activity [1,9]. An understanding of the interplay between structural and functional organization is emerging. The field is further advanced in eukaryotic organisms compared to bacteria, yet in both cases a lot of unanswered questions remain. Here, we discuss recent advancements in understanding bacterial genome organization, linking chromatin structure to function. Our focus is on systematic approaches aimed at determining characteristics of the dynamic organization of bacterial genomes, and lessons learned from similar studies in eukaryotic model systems.

State of the art

Most bacterial model organisms harbour a single circular chromosome. The bacterial chromosome has been primarily studied in *Bacillus subtilis*, *Escherichia coli* and *Caulobacter crescentus*, and unless otherwise indicated the information summarized here applies to these organisms. Bacteria have a cell cycle with a duration on the order of tens of minutes. As a consequence, genome folding and transcription are intimately coupled with genome replication. Current key question is to understand the structure–function relations within the bacterial chromosome, specifically the interplay between genome structure and gene activity.

The first systematic studies of bacterial chromosome structure aimed at defining the positioning of genomic loci within the cell. Two approaches based on fluorescence microscopy were used: *i*) fluorescent *in situ* hybridization (FISH) labelling of endogenous loci in fixed cells [10] and *ii*) fluorescent repressor-operator systems (FROS) involving binding of e.g. LacI-GFP to exogenous *lacO* operator sites integrated in the genome in living cells [11,12]. These studies have revealed that regions proximal to the initiation (*oriC*) and termination (*ter*) site of replication are not distributed randomly in the nucleoid but exhibit specific localization patterns throughout the cell cycle [12–14]. Visualizing the locations of up to about 100 defined genomic loci relative to *oriC* reveals a linear relationship between genomic and physical location, indicating a linear ordering [15–17]. The reproducible positioning of genomic loci at specific subcellular positions in individual cells and their linear organization appear as fundamental features of chromatin organization in bacteria.

In *E. coli*, *oriC* and *ter* are part of two distinct structural domains, the Ori and Ter macrodomains [18]. In addition, the *E. coli* genome contains two other structural domains flanking the Ter domain, called the Right and Left macrodomains, and two non-structured (NS) regions, flanking the Ori domain [19,20] (see Figure 1). The Ter domain stretches along the length of the cell to connect the two chromosomal arms, with an estimated packing density of only 1/10 compared to the rest of the genome [16]. Genome packing density may correlate with genome activity. The chromosome is organized as a dense nucleoid scaffold wherefrom large 'plectonemic' loops of negatively supercoiled DNA protrude. Such loops are probably formed by binding of a group of proteins called 'nucleoid-associated proteins' (NAPs) [21–23]. But these proteins only provide part of the answer.

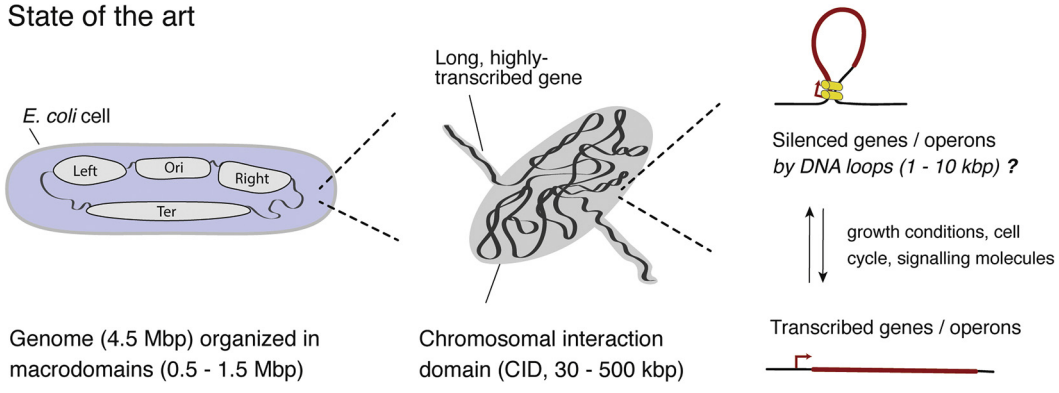
What is the relevance and evolutionary conservation of the different levels of organization? How are the different organizational units and their borders defined, and what are their dynamics upon DNA transactions? Below we discuss recent advances on quantitative approaches aiming to find answers to these fundamental questions.

Chromatin structure: organization in domains

Since the early 2000's chromosome conformation capture (3C), developed by Dekker and colleagues [24], and derivatives of the method, have promoted large advances in understanding genome folding and function [25–28]. These techniques yield maps of relative interaction frequency between different pairs of genomic sites averaged over a population of cells. The technique relies on chemical cross-linking, digestion, religation of digested fragments, followed by identification of the hybrid DNA molecules [24]. A large-scale variant, Hi-C [29], has been used to produce genome-wide contact matrices in several organisms. These matrices, structurally interpreted by modelling approaches, provide insight in global and local features of genome structures [28,30].

Among bacterial species, currently, Hi-C contact maps are available for *C. crescentus* [31–33], *B. subtilis* [34–36] and *Mycoplasma pneumoniae* [37]. Different genome features have been identified in these studies. In the *C. crescentus*, *B. subtilis*, *V. cholera and M. pneumoniae* genome-wide contact maps, several tens of chromosomal interaction domains (CIDs) have been identified [31,34,35,37,38] (see Figure 1). CIDs are highly self-interacting genomic regions. These regions vary in size from about 20 to 400 kbp, and they are analogous to so-called topologically associated domains (TADs) identified in contact maps of eukaryotic genomes [39,40]. An

Figure 1



State of the art. The bacterial genome is organised at different length scales. At the smallest length scales it has been hypothesised that genome folding is directly affected by environmental signals which are translated into a transcriptional response.

exception in terms of CID dimensions (15–33 kbp) is the genome of *M. pneumoniae*, which is only 800 kbp in size [37], 5x smaller than the genomes of other bacteria studied by Hi-C. For *E. coli*, a genome wide contact map was obtained using an other 3C-derivative [41]. Whereas regions of high internal interaction are identified, the relatively low resolution of these maps does not permit identification of CIDs. The regions of high interaction frequency might correspond to the macrodomains discussed above. Indications of the existence of domains on average ~10 kbp in size (microdomains) in *E. coli*, comparable in size to CIDs, come from EM imaging of isolated chromosomes, clustered gene activity [42] and *in vivo* recombination based assays in the closely related *Salmonella typhimurium* [43,44]. Finally, at a length scale between that of CIDs/microdomains and macrodomains, another organizational structure is proposed: the high-density chromosomal regions (HDRs), of which 10 per genome are estimated and sizing around 200 tot 250 kbp [35].

Considering the mechanisms that form CIDs and their boundaries, mounting evidence is pointing at long (>1 kbp), highly transcribed genes separating domains [31,32,34,35,38]. Inhibition of transcription with rifampicin almost completely eliminated these borders [31,34,36], and insertion of a highly expressed gene was sufficient to generate a new barrier [31]. Possibly, these highly transcribed genes cluster into transcription factories [45], imposing an organization of intervening sequences in looped domains. Borders not containing highly expressed genes often have low GC%, indicating the presence of horizontally acquired elements [31,35]. In that light it is interesting to note that low GC-content regions are targeted by the nucleoid-associated proteins H-NS and FIS [46–48]. Inhibition of gyrase reduces the sharpness and position of CID borders, which has been interpreted as supercoiling being important in establishing CIDs [31,37]. Thus, CIDs have been proposed to be connected by segments of decompacted chromatin, forming a higher-order "domains-on-a-string" organization. The biological relevance of this level of organisation is not clear.

CIDs and their borders are dynamic and correlate with changes in gene expression. It has been long known that bacteria change global gene expression patterns in response to environmental cues [49]. Indeed, contact maps acquired from bacterial cultures in different growth conditions, e.g. starved cells versus exponentially growing cells, exhibit clear alterations in CID boundary positioning [32]. Currently, there is not sufficient data to correlate the changes occurring in global genome folding (at CID level or higher) and changes in the expression of specific genes and operons in response to altered conditions.

Structure–function relations of the bacterial genome

Correlations, linking genome structure to gene activity, have been established by combining information from chromosome conformation capture, chromatin immunoprecipitation (ChIP) data of DNA-binding proteins and gene expression profiles. In eukaryotes insulator proteins are bound at domain boundaries and have been shown to be involved in boundary formation [6,39,50,51]. There is no direct evidence for the involvement of specific architectural proteins in CID boundary formation in bacteria, but for microdomains in *E. coli*, which might be the same as CIDs, the involvement of nucleoid-associated proteins FIS and H-NS has been suggested [52,53]. There is very limited information on structure–function relations, but more is known about the effect of DNA-binding of NAPs on gene expression. Genes bound by H-NS are generally expressed at low levels or silenced completely [47,48,54,55]. Such an effect on gene expression is not seen for binding of FIS, which primarily exerts effects on transcription indirectly by regulating the expression of other transcription factors [47,56]. High protein occupancy (including other DNA-binding proteins in addition to NAPs) along the genome is also associated with gene silencing [57].

Interplay between DNA-binding of NAPs and gene expression cannot provide us with an understanding of genome structure-driven regulation of gene activity (see Figure 1). To understand how local genome organization at the level of genes and operons affects functional biological outcome, e.g. state and level of gene expression, studies on single cells or pre-sorted small homogenous cell populations are needed. For instance, chromosome conformation capture studies of genome organization involve ensemble-averaging over the genomic conformation of all cells in a sample. This is a particularly important limitation for using this technique in bacteria with short cell cycles and/or which are hard to synchronize. The benefit of single-cell methods is that in contrast to yielding average characteristics of the population of cells analysed, they reveal intercellular variance within a population, allowing identification of differently behaving subpopulations of cells [58]. The solution to avoid ensemble-averaging in Hi-C is to use single cells as shown for eukaryotes [58,59].

3C techniques cannot be used to straightforwardly determine changes in genome conformation occurring at short time-scales, such as regulatory switches in response to environmental cues, due to limited time resolution. To quantify dynamic changes in chromatin structure at the scale of genes and operons (on the order of 1–10 kilobases), a different approach is required. Currently, high-resolution imaging of sets of *in vivo* fluorescently tagged loci encompassing the genomic

region of interest in living cells, e.g. using FROS, is best suited to answer these questions. Performing time-course experiments on these single cells and/or molecules allows for characterization of the true dynamics of genome organization and gene activity regulation. Positioning of loci relative to each other and cellular landmarks can be determined in real-time to reveal changes occurring upon varying growth conditions. Hensel and co-workers demonstrated the feasibility of such methods in bacteria: they investigated the formation of loops of 2.3 kb upon binding of the cI repressor using two FROS arrays flanking the operator elements that bind cI [60]. A positive correlation was established between loop formation due to repressor binding and gene activity (simultaneous positive autoregulation and silencing of a major lytic promoter). This approach can be extended to studies correlating local genome organization to gene activity.

Additionally, tracking of loci in living cells allows determination of diffusion constants (i.e. the space explored per time unit by a locus). These values vary as a function of growth phase and are subject to metabolic processes, ATP synthesis and temperature [61]. Moreover, these values differ dependent on subcellular localization and chromosomal coordinate [62–64]. Loci in the Ter macrodomain are least mobile; mobility increases along the chromosomal arms towards the Ori macrodomain [62]. It is not clear whether gene activity also correlates with its macrodomain-positioning, but there are indications that active genes in the Ori macrodomain are higher expressed than those in Ter. The mechanistic nature of these differences remains unclear, but cannot be simply attributed to a gene dosage effect [65]. Gene silencing does not correlate with macrodomain positioning, but with regions of high levels of nucleoid-associated protein binding [65,57,66], which might be more compactly organized compared to regions with active transcription. Systematic parallel studies of gene activity and global and local genome organization are needed to establish firm correlations.

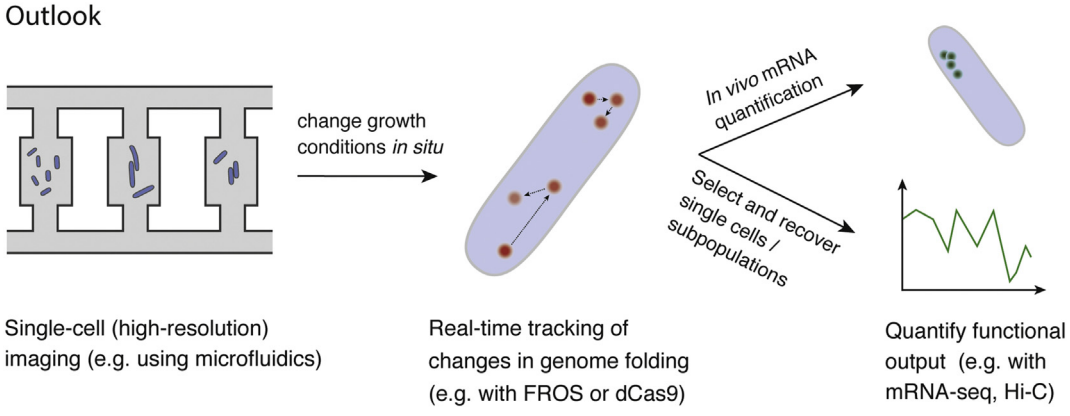
Outlook

FROS and Hi-C have advanced our knowledge of genome organization *in vivo*. Questions of structure–function relations can be addressed by either of these methods, but particularly powerful will be their mutual combination or combination with other genomic analyses (e.g. RNA-seq, ChIP-seq). We see three promising avenues for future research (see Figure 2):

1) application of the programmable and DNA-targeting platform CRISPR/dCas9, which is replacing FROS in many recent studies in eukaryotes [67–70], in bacteria. Advantage is that this approach obviates extensive time-consuming genome engineering,
2) establishing Hi-C for low number of cells permitting analyses of homogenous sub-populations extracted from intrinsically heterogeneous bacterial populations,
3) developing parallel imaging platforms for single bacterial cells under controlled variable conditions. This is essential for obtaining systematic data from time course studies. Clever microfluidic channel designs – some of which already utilized now – can be used to change growth conditions leading to a physiological response whilst cells are being imaged [71–73]. Application of microfluidics might be the key to recover (defined populations of) single cells, which can be processed for Hi-C or genomic analysis methods.

Instrumental to all these approaches is the verification of function i.e. quantification of the level of gene activity, which can be measured by (single-cell) mRNA-seq [74,75] or visualization of real-time kinetics of transcription *in vivo*. Transcription *in vivo* can be

Figure 2



Outlook. Development of new technologies and application of existing – but in bacteria unused – technologies is expected to establish structure–function relationships for bacterial genome organisation.

visualized e.g. by including RNA aptamers in mRNA transcripts targeted by MS2-GFP [76,77], the RNA-binding protein Pumilio [78,79] and the nuclease-deficient Cas9 (dCas9) from type II CRISPR/Cas systems [80], which can be (re-)programmed to specifically bind RNA [81].

It is remarkable that very few studies to quantitatively dissect structure–function relations on local scales have been published, whereas the techniques required are either already implemented and applied, or just need translation from eukaryotic to prokaryotic cells. We expect extensive data to become available in the next few years that can be used for iterative building, testing and improving of biological models of genome organization and dynamics.

Acknowledgments

Research on the topic of this review in the lab of R.T.D. is supported by grants from the Netherlands Organization for Scientific Research [VICI 016.160.613] and the Human Frontier Science Program (HFSP) [RGP0014/2014], M. T.-D. is supported by a grant of the Dutch Technology Foundation [12385] STW.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Luijsterburg MS, White MF, van Driel R, Dame RT: **The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes**. *Crit Rev Biochem Mol Biol* 2008, **43**: 393–418.

2. Phillips-Cremins JE, Corces VG: **Chromatin insulators: linking genome organization to cellular function**. *Mol Cell* 2013, **50**: 461–474.

3. van der Valk RA, Vreede J, Crémazy F, Dame RT: **Genomic looping: a key principle of chromatin organization**. *J Mol Microbiol Biotechnol* 2014, **24**:344–359.

4. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, *et al.*: **CTCF-mediated functional chromatin interactome in pluripotent cells**. *Nat Genet* 2011, **43**:630–638.

5. Song D, Loparo JJ: **Building bridges within the bacterial chromosome**. *Trends Genet* 2015, **31**:164–173.

6. Chetverina D, Fujioka M, Erokhin M, Georgiev P, Jaynes JB, Schedl P: **Boundaries of loop domains (insulators): determinants of chromosome form and function in multicellular eukaryotes**. *Bioessays* 2017, **39**:1600233.

7. Zentner GE, Henikoff S: **Regulation of nucleosome dynamics by histone modifications**. *Nat Struct Mol Biol* 2013, **20**: 259–266.

8. Lawrence M, Daujat S, Schneider R: **Lateral thinking: how histone modifications regulate gene expression**. *Trends Genet* 2016, **32**:42–56.

9. Cavalli G, Misteli T: **Functional implications of genome topology**. *Nat Struct Mol Biol* 2013, **20**:290–299.

10. Nath J, Johnson KL: **A review of fluorescence in situ hybridization (FISH): current status and future prospects**. *Biotech Histochem* 2000, **75**:54–78.

11. Robinett CC, Straight A, Li G, Willhelm C, Sudlow G, Murray A, Belmont AS: **In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition**. *J Cell Biol* 1996, **135**. 1685 LP-1700.

12. Webb CD, Teleman A, Gordon S, Straight A, Belmont A, Lin DCH, Grossman AD, Wright A, Losick R: **Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of *B. subtilis***. *Cell* 1997, **88**:667–674.

13. Gordon GS, Sitnikov D, Webb CD, Teleman A, Straight A, Losick R, Murray AW, Wright A: **Chromosome and low copy plasmid segregation in *E. coli*: visual evidence for distinct mechanisms**. *Cell* 1997, **90**:1113–1121.

14. Teleman AA, Graumann PL, Lin DC, Grossman AD, Losick R: **Chromosome arrangement within a bacterium**. *Curr Biol* 1998, **8**:1102–1109.

15. • Viollier PH, Thanbichler M, McGrath PT, West L, Meewan M, McAdams HH, Shapiro L: **Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication**. *Proc Natl Acad Sci USA* 2004, **101**:9257–9262.
Tracking the positions of 112 genomic loci in *C. crescentus* showed that genomic loci are not distributed randomly throughout the nucleoid; a linear correlation was found between chromosomal position and physical localization in the cell.

16. Wiggins PA, Cheveralls KC, Martin JS, Lintner R, Kondev J: **Strong intranucleoid interactions organize the *Escherichia coli* chromosome into a nucleoid filament**. *Proc Natl Acad Sci* 2010, **107**:4991–4995.

17. Wang X, Montero Llopis P, Rudner DZ: ***Bacillus subtilis* chromosome organization oscillates between two distinct patterns**. *Proc Natl Acad Sci USA* 2014, **111**:12877–12882.

18. Niki H, Yamaichi Y, Hiraga S: **Dynamic organization of chromosomal DNA in *Escherichia coli***. *Genes Dev* 2000, **14**: 212–223.

19. Valens M, Penaud S, Rossignol M, Cornet F, Boccard F: **Macrodomain organization of the *Escherichia coli* chromosome**. *EMBO J* 2004, **23**:4330–4341.

20. Boccard F, Esnault E, Valens M: **Spatial arrangement and macrodomain organization of bacterial chromosomes**. *Mol Microbiol* 2005, **57**:9–16.

21. Dame RT: **The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin**. *Mol Microbiol* 2005, **56**:858–870.

22. Dillon SC, Dorman CJ: **Bacterial nucleoid-associated proteins, nucleoid structure and gene expression**. *Nat Rev Microbiol* 2010, **8**:185–195.

23. Dame RT, Tark-Dame M: **Bacterial chromatin: converging views at different scales**. *Curr Opin Cell Biol* 2016, **40**:60–65.

24. Dekker J: **Capturing chromosome conformation**. *Science* 2002, **295**:1306–1311.

25. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)**. *Nat Genet* 2006, **38**:1348–1354.

26. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Singh Sandhu K, Singh U, *et al.*: **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions**. *Nat Genet* 2006, **38**:1341–1347.

27. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA: **The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules**. *Nat Struct Mol Biol* 2011, **18**:107–114.

28. Dekker J, Mirny L: **The 3D genome as moderator of chromosomal communication**. *Cell* 2016, **164**:1110–1121.

29. Lieberman-Aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, *et al.*: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**:289–294.

30. Dekker J, Marti-Renom MA, Mirny LA: **Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data**. *Nat Rev Genet* 2013, **14**:390–403.

31. • Le TBK, Imakaev MV, Mirny LA, Laub MT: **High-resolution mapping of the spatial organization of a bacterial chromosome**. *Science* 2013, **342**:731–734.
The first bacterial high-resolution chromatin conformation capture study, identifying chromosomal interaction domains (CIDs), analogous to topological associated domains (TADs in eukaryotes), and that CID boundaries are enriched in highly-expressed genes.

32. •• Le TB, Laub MT: **Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries**. *EMBO J* 2016, **35**:1582–1595.
Hi-C and fluorescent imaging are combined to further resolve the definition of CIDs and inter-CID regions. These authors define transcription rate and length as drivers of CID boundaries, and show that this level of genome organization is dynamic, changing as function of gene expression (e.g. induced by changed growth conditions).

33. Tran NT, Laub MT, Le TBK: **SMC progressively aligns chromosomal arms in Caulobacter crescentus but is antagonized by convergent transcription**. *Cell Rep* 2017, **20**:2057–2071.

34. Wang X, Le TBK, Lajoie BR, Dekker J, Laub MT, Rudner DZ: **Condensin promotes the juxtaposition of DNA flanking its loading site in Bacillus subtilis**. *Genes Dev* 2015, **29**:1661–1675.

35. Marbouty M, Le Gall A, Cattoni DI, Cournac A, Koh A, Fiche JB, Mozziconacci J, Murray H, Koszul R, Nollmann M: **Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging**. *Mol Cell* 2015, **59**:588–602.

36. Wang X, Brandão HB, Le TBK, Laub MT, Rudner DZ: **Bacillus subtilis SMC complexes juxtapose chromosome arms as they travel from origin to terminus**. *Science* 2017, **355**:524–527.

37. •• Trussart M, Yus E, Martinez S, Baù D, Tahara YO, Pengo T, Widjaja M, Kretschmer S, Swoger J, Djordjevic S, *et al.*: **Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae**. *Nat Commun* 2017, **8**:14665.
Hi-C contact maps from *M. pneumoniae* show that in this small-sized bacterial chromosome, highly self-interacting organizational structures of similar length are found, as in other bacteria. This supports the view that CIDs are a conserved feature in bacterial genomes.

38. Val M-E, Marbouty M, de Lemos Martins F, Kennedy SP, Kemble H, Bland MJ, Possoz C, Koszul R, Skovgaard O, Mazel D: **A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae***. *Sci Adv* 2016, **2**:e1501914.

39. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012, **485**:376–380.

40. Dekker J, Heard E: **Structural and functional diversity of topologically associating domains**. *FEBS Lett* 2015, **589**:2877–2884.

41. Marbouty M, Cournac A, Flot JF, Marie-Nelly H, Mozziconacci J, Koszul R: **Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms**. *Elife* 2014, **3**, e03318.

42. Postow L, Hardy CA, Arsuaga J, Cozzarelli NR: **Topological domain structure of the *Escherichia coli* chromosome**. *Genes Dev* 2004, **18**:1766–1779.

43. Higgins NP, Yang X, Fu Q, Roth JR: **Surveying a supercoil domain by using the gamma delta resolution system in *Salmonella typhimurium***. *J Bacteriol* 1996, **178**:2825–2835.

44. Deng S, Stein RA, Higgins NP: **Transcription-induced barriers to supercoil diffusion in the *Salmonella typhimurium* chromosome**. *Proc Natl Acad Sci USA* 2004, **101**:3398–3403.

45. Papantonis A, Cook PR: **Transcription factories: genome organization and gene regulation**. *Chem Rev* 2013, **113**:8683–8705.

46. Singh K, Milstein JN, Navarre WW: **Xenogeneic silencing and its impact on bacterial genomes**. *Annu Rev Microbiol* 2016, **70**:199–213.

47. Grainger DC, Hurd D, Goldberg MD, Busby SJW: **Association of nucleoid proteins with coding and non-coding segments of the *Escherichia coli* genome**. *Nucleic Acids Res* 2006, **34**:4642–4652.

48. Lucchini S, Rowley G, Goldberg MD, Hurd D, Harrison M, Hinton JCD: **H-NS mediates the silencing of laterally acquired genes in bacteria**. *PLoS Pathog* 2006, **2**:0746–0752.

49. Dorman CJ: **Flexible response: DNA supercoiling, transcription and bacterial adaptation to environmental stress**. *Trends Microbiol* 1996, **4**:214–216.

50. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome**. *Cell* 2012, **148**:458–472.

51. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell* 2014, **159**:1665–1680.

52. Noom MC, Navarre WW, Oshima T, Wuite GJL, Dame RT: **H-NS promotes looped domain formation in the bacterial chromosome**. *Curr Biol* 2007, **17**:913–914.

53. Hardy CD, Cozzarelli NR: **A genetic selection for supercoiling mutants of *Escherichia coli* reveals proteins implicated in chromosome structure**. *Mol Microbiol* 2005, **57**:1636–1652.

54. Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC: **Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella**. *Science* 2006, **313**:236. LP-238.

55. Oshima T, Ishikawa S, Kurokawa K, Aiba H, Ogasawara N: ***Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase**. *DNA Res* 2006, **13**:141–153.

56. Kahramanoglou C, Seshasayee ASN, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, Benes V, Fraser GM, Luscombe NM: **Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli***. *Nucleic Acids Res* 2011, **39**:2073–2091.

57. Vora T, Hottes AK, Tavazoie S: **Protein occupancy landscape of a bacterial genome**. *Mol Cell* 2009, **35**:247–253.

58. • Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P: **Single-cell Hi-C reveals cell-to-cell variability in chromosome structure**. *Nature* 2013, **502**:59–64.
First application of single-cell Hi-C in eukaryotic cells, producing contact maps that reveal cell-to-cell variance in genome organization.

59. Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, Mendelson-cohen N: **Cell-cycle dynamics of chromosomal organization at single-cell resolution**. *Nature* 2017, **547**:61–67.

60. •• Hensel Z, Weng X, Lagda AC, Xiao J: **Transcription-factor-mediated DNA looping probed by high-resolution, single-molecule imaging in live *E. coli* cells**. *PLoS Biol* 2013, **11**.
Using FROS insertions, a pair of genomic loci with a genomic distance of 2.3 kb (a distance relevant to regulation of gene expression) is tracked. These authors showed that repressor cI is able to bridge two operator elements, forming a loop required for simultaneous repression of a lytic promoter and positive autoregulation.

61. • Weber SC, Spakowitz AJ, Theriot JA: **Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci**. *Proc Natl Acad Sci USA* 2012, **109**:7338–7343.
E. coli and *S. cerevisiae* were imaged using FROS showing that fluctuations in the position of genomic loci are not just due to thermal fluctuations, but are affected by metabolic activity and ATP levels in the cell.

62. • Javer A, Long Z, Nugent E, Grisi M, Siriwatwetchakul K, Dorfman KD, Cicuta P, Cosentino Lagomarsino M: **Short-time**

movement of *E. coli* chromosomal loci depends on coordinate and subcellular localization. *Nat Commun* 2013, **4**:1–8.
The fluctuations of genomic loci as described in Weber et al. (2012) are shown to be also dependent on chromosomal and subcellular localization: the dynamic behaviour of loci differs depending on the distance to the origin of replication.

63. Javer A, Kuwada NJ, Long Z, Benza VG, Dorfman KD, Wiggins PA, Cicuta P, Lagomarsino MC: **Persistent super-diffusive motion of *Escherichia coli* chromosomal loci**. *Nat Commun* 2014, **5**:1–8.

64. • Espeli O, Mercier R, Boccard F: **DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome**. *Mol Microbiol* 2008, **68**:1418–1427.
Time-lapse microscopy used to track fluorescent FROS markers revealed that the dynamics of genomic loci varies between macrodomains.

65. Bryant JA, Sellars LE, Busby SJW, Lee DJ: **Chromosome position effects on gene expression in *Escherichia coli* K-12**. *Nucleic Acids Res* 2014, **42**:11383–11392.

66. Brambilla E, Sclavi B: **Gene regulation by H-NS as a function of growth conditions depends on chromosomal position in *Escherichia coli***. *G3 Genes|Genomes|Genetics* 2015, **5**:605–614.

67. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li G-W, Park J, Blackburn EH, Weissman JS, Qi LS, *et al.*: **Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system**. *Cell* 2013, **155**:1479–1491.

68. Ma H, Tu L-C, Naseri A, Huisman M, Zhang S, Grunwald D, Pederson T: **Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow**. *Nat Biotech* 2016, **34**:528–530.

69. Zhou Y, Wang P, Tian F, Gao G, Huang L, Wei W, Xie XS: **Painting a specific chromosome with CRISPR/Cas9 for live-cell imaging**. *Cell Res* 2017, **27**:298–301.

70. Dreissig S, Schiml S, Schindele P, Weiss O, Rutten T, Schubert V, Gladilin E, Mette MF, Puchta H, Houben A: **Live-cell CRISPR imaging in plants reveals dynamic telomere movements**. *Plant J* 2017, **91**:565–573.

71. Fisher JK, Bourniquel A, Witz G, Weiner B, Prentiss M, Kleckner N: **Four-dimensional imaging of *E. coli* nucleoid organization and dynamics in living cells**. *Cell* 2013, **153**:882–895.

72. Binder D, Probst C, Grünberger A, Hilgers F, Loeschcke A, Jaeger KE, Kohlheyer D, Drepper T: **Comparative single-cell analysis of different *E. coli* expression systems during microfluidic cultivation**. *PLoS One* 2016, **11**:1–19.

73. Baltekin Ö, Boucharin A, Tano E, Andersson DI, Elf J: **Antibiotic susceptibility testing in less than 30 Min using direct single-cell imaging**. *Proc Natl Acad Sci* 2017, https://doi.org/10.1073/pnas.1708558114.

74. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, *et al.*: **Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells**. *Nature* 2013, **498**:236–240.

75. Grün D, Van Oudenaarden A: **Design and analysis of single-cell sequencing experiments**. *Cell* 2015, **163**:799–810.

76. Golding I, Cox EC: **RNA dynamics in live *Escherichia coli* cells**. *Proc Natl Acad Sci* 2004, **101**:11310–11315.

77. Golding I, Paulsson J, Zawilski SM, Cox EC: **Real-time kinetics of gene activity in individual bacteria**. *Cell* 2005, **123**:1025–1036.

78. Adamala KP, Martin-Alarcon DA, Boyden ES: **Programmable RNA-binding protein composed of repeats of a single modular unit**. *Proc Natl Acad Sci* 2016, **113**:E2579–E2588.

79. Yoshimura H, Inaguma A, Yamada T, Ozawa T: **Fluorescent probes for imaging endogenous β-actin mRNA in living cells using fluorescent protein-tagged Pumilio**. *ACS Chem Biol* 2012, **7**:999–1005.

80. Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, Doudna JA, Yeo GW: **Programmable RNA tracking in live cells with CRISPR/Cas9**. *Cell* 2016, **165**:488–496.

81. O'Connell MR, Oakes BL, Sternberg SH, East-Seletsky A, Kaplan M, Doudna JA: **Programmable RNA recognition and cleavage by CRISPR/Cas9**. *Nature* 2014, **516**:263–266.

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology and Molecular Biology Reviews®

REVIEW

The Divided Bacterial Genome: Structure, Function, and Evolution

George C. diCenzo, Turlough M. Finan

Department of Biology, McMaster University, Hamilton, Ontario, Canada

Published 9 August 2017

Citation diCenzo GC, Finan TM. 2017. The divided bacterial genome: structure, function, and evolution. Microbiol Mol Biol Rev 81:e00019-17. https://doi.org/10.1128/MMBR.00019-17.

Address correspondence to Turlough M. Finan, finan@mcmaster.ca.

SUMMARY Approximately 10% of bacterial genomes are split between two or more large DNA fragments, a genome architecture referred to as a multipartite genome. This multipartite organization is found in many important organisms, including plant symbionts, such as the nitrogen-fixing rhizobia, and plant, animal, and human pathogens, including the genera *Brucella*, *Vibrio*, and *Burkholderia*. The availability of many complete bacterial genome sequences means that we can now examine on a broad scale the

characteristics of the different types of DNA molecules in a genome. Recent work has begun to shed light on the unique properties of each class of replicon, the unique functional role of chromosomal and nonchromosomal DNA molecules, and how the exploitation of novel niches may have driven the evolution of the multipartite genome. The aims of this review are to (i) outline the literature regarding bacterial genomes that are divided into multiple fragments, (ii) provide a meta-analysis of completed bacterial genomes from 1,708 species as a way of reviewing the abundant information present in these genome sequences, and (iii) provide an encompassing model to explain the evolution and function of the multipartite genome structure. This review covers, among other topics, salient genome terminology; mechanisms of multipartite genome formation; the phylogenetic distribution of multipartite genomes; how each part of a genome differs with respect to genomic signatures, genetic variability, and gene functional annotation; how each DNA molecule may interact; as well as the costs and benefits of this genome structure.

KEYWORDS secondary replicons, genome analysis, genome organization, genomics, megaplasmids, population genetics, secondary chromosome

INTRODUCTION

In 1963, John Cairns reported autoradiographs of DNA from *Escherichia coli* that provided the first evidence that its genome consists of a single circular chromosome (1). Together with subsequent studies (see, for example, references 2 and 3), that work led to the generally accepted view that all bacterial genomes consist of a single circular chromosome, possibly including some smaller, nonessential, circular plasmids. However, that view had begun to change within 20 years. The identification of the first linear plasmid in *Streptomyces* in 1979 (4) and the determination that the *Borrelia burgdorferi* chromosome is linear in 1989 (5, 6) illustrated that bacterial DNA molecules need not be circular. Moreover, a *Sinorhizobium meliloti* plasmid with a molecular mass of $>300 \times 10^6$ Da (~460 kb), which the authors of that study termed a "megaplasmid," was identified in 1981 (7), challenging the notion that nearly the entire bacterial genome is located on the chromosome (8). Finally, in 1989, the report of a "second chromosome" in *Rhodobacter sphaeroides* (9) illustrated the potential for essential cell functions to be encoded by multiple replicons within the bacterium. There is also the peculiar case of an unusual clade within the *Aureimonas* genus that has the sole copy of the rRNA operon on a 9.4-kb plasmid (10). The recent explosion in complete genome sequencing has revealed that approximately 10% of bacterial genomes do not contain a single, circular chromosome like *E. coli* and instead contain several large and potentially essential replicons of either a linear or a circular nature (11). The genome architecture consisting of a chromosome plus one or more additional large replicons is referred to as a divided genome or a multipartite genome. Interestingly, studies have repeatedly observed for multipartite genomes that not only is each DNA molecule physically separate, but each molecule also has distinct properties, such as differences in codon usage (the ratio of synonymous codons that are used), GC content (percentage of the DNA consisting of guanine and cytosine), and dinucleotide relative abundance (the frequency with which each pair of nucleotides appears in the DNA sequence).

The organization of prokaryotic genomes is not stochastic, but instead, their organization reflects some functional or regulatory purpose (12–14). For example, enzymes for each step of a biosynthetic or catabolic pathway are generally encoded by a single operon and are often colocalized on the chromosome with their regulator (13). The chromosomal location of a gene can influence its expression level (15) and, at least in fast-replicating species, the copy number of the gene (16). Additionally, there is a general bias for bacterial genes to be enriched in the leading strand to avoid head-on collisions between the transcriptional and DNA replicative machineries (17). Given the structured nature of prokaryotic genomes, it is unlikely that the multipartite genome structure simply represents an evolutionary peculiarity, and instead, it is presumably shaped by selective pressures. Understanding the evolutionary forces driving the

emergence of the multipartite genome and the advantage of maintaining multiple replicons is particularly salient, as many important bacteria contain this genomic architecture. These bacteria include plant symbionts such as many of the rhizobia (18), plant pathogens such as *Agrobacterium* (19), and animal and human pathogens, including *Brucella* (20), *Vibrio* (21), and *Burkholderia* (22). Understanding the emergence and function of this genome structure may lead to generalizable insights into the biology of these diverse organisms that could lead to practical applications in promoting or suppressing these symbiotic and pathogenic relationships.

Purpose of This Review

The first goal of this review is to build upon previous reviews (11, 23–31) and to provide an unbiased assessment of the information available on the structure, function, and evolution of divided bacterial genomes. This consists of a comprehensive review of the relevant literature as well as an analysis of all complete genomes available through the National Center for Biotechnology Information (NCBI) genome database (accessed 21 March 2016) as a way of reviewing the abundant, untapped information present within these sequences. The second goal is to synthesize the data presented throughout this review into a generalized model explaining the evolution and function of the multipartite bacterial genome.

BACTERIAL REPLICON CLASSIFICATION

There are several terms that describe the different types of DNA molecules that are present within a multipartite genome. In this section, these terms are defined, and general characteristics of each replicon class are provided. It is important to keep in mind that many DNA molecules are likely to blur the boundaries of these classes and that the characteristics of DNA molecules are best thought of as belonging on a spectrum. However, just as the political spectrum is split into several discrete groups for descriptive purposes, it is important to split the spectrum of DNA molecules into discrete classes in order to easily portray the main characteristics of the replicon, even if such classifications may be an oversimplification in some cases.

Replicon and Secondary Replicon

We use the term "replicon" as a general term in reference to any DNA molecule regardless of its specific nature, and each replicon can be further classified based on specific characteristics, as described below. The term "secondary replicon" refers to any replicon that is not the primary chromosome of the cell. We suggest that each replicon be classified into one of the following five groups, as described below and in Fig. 1: chromosome, second chromosome, chromid, megaplasmid, and plasmid.

In the strictest sense of the term replicon, it should be used only in reference to DNA molecules with a single origin of replication. While this distinction is irrelevant in reference to bacterial genomes, this definition would exclude the chromosomes of some archaea that have a chromosome containing multiple origins of replication (32, 33). As such, while the classification system described here should be applicable to archaea, the term replicon should be avoided when describing the chromosomes of archaea. As this review is focused only on bacterial genomes, the term replicon is used throughout.

Chromosome

"Chromosome" refers to the primary replicon. As described by Harrison et al. (11), the chromosome is always the largest replicon in the genome and contains the majority of the core/essential genes. There is nearly a 100-fold distribution in the sizes of fully sequenced and assembled bacterial chromosomes, with average and median sizes of ~3.65 Mb and ~3.46 Mb, respectively (Fig. 2A). The average and median bacterial genome sizes are ~3.87 Mb and ~3.65 Mb, respectively (Fig. 2A), illustrating that the chromosome accounts for nearly all of the genetic material of most prokaryotic organisms. However, this is not universal. For example, the chromosomes of *Sinorhizobium meliloti* 1021 and *Burkholderia xenovorans* LB400 account for only 54.6% and

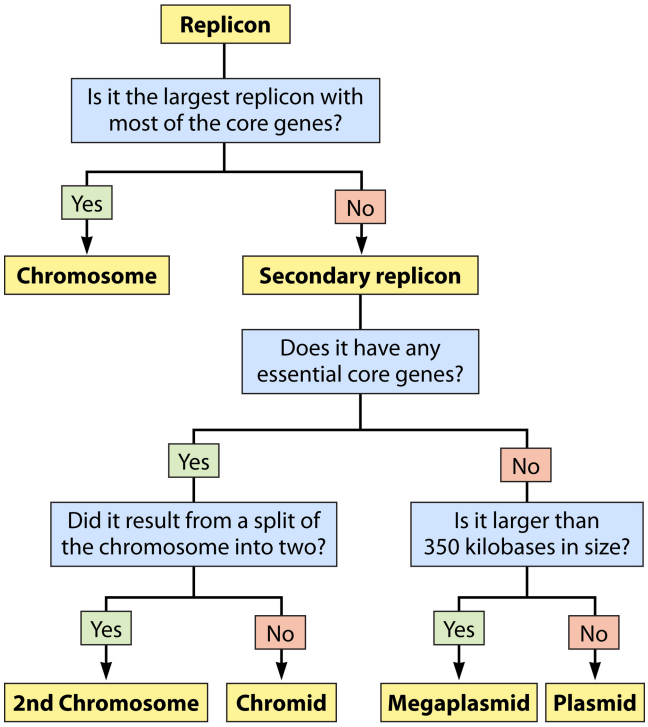


FIG 1 Decision chart for the classification of bacterial replicons. This flow chart illustrates the decisions involved in the classification of bacterial replicons.

50.3% of their genomes, respectively (18, 22). Nevertheless, 1,017 (~59.5%) of the 1,708 bacterial species with a complete genome available in the NCBI database contain a chromosome but no secondary replicon (chromid, megaplasmid, or plasmid), while only 192 (~11%) have a chromid and/or megaplasmid. The major features of genomes and all replicon classes are summarized in Table 1.

Plasmid and Megaplasmid

Many of the secondary replicons in bacterial genomes carry no core genes and are nonessential and thus dispensable for cell viability in most environments. The majority of the genes on these replicons were acquired through recent horizontal gene transfer (HGT), and their genomic signatures, such as GC content and dinucleotide composition, differ significantly from the chromosome (11). These types of replicons, defined by the lack of core genes, are referred to as "plasmids" and "megaplasmids." The distinction between plasmid and megaplasmid is currently based solely on size, although there is no established boundary between plasmid and megaplasmid in the literature. While any size limitation is essentially arbitrary, we suggest a lower cutoff of 350 kb for megaplasmid status, as this is equal to roughly 10% of the median bacterial genome size. Any nonessential replicon of <350 kb would therefore be a plasmid. When using this boundary to distinguish megaplasmids from plasmids, the average and median plasmid sizes are ~78.9 kb and ~46.2 kb, respectively. In contrast, the average and median megaplasmid sizes are approximately 10 times larger, at ~772 kb and ~558 kb, respectively, with the pSymA replicon of *S. meliloti* 2011 (1.35 Mb) and the third replicon of *Burkholderia lata* 383 (1.4 Mb) being the largest to have been experimentally demonstrated to be nonessential and thus megaplasmids (34, 35). It is also interesting to note that the sizes of chromosomes follow a bell-shaped distribution (Fig. 2A), whereas the size distributions of plasmids and megaplasmids are instead positively skewed (Fig. 2B and C). This is perhaps suggestive of evolutionary forces acting to limit the size of these nonessential replicons.

The term megaplasmid was originally coined in reference to a large *S. meliloti* plasmid (7), and since then, megaplasmid has been used simply as a way of referring

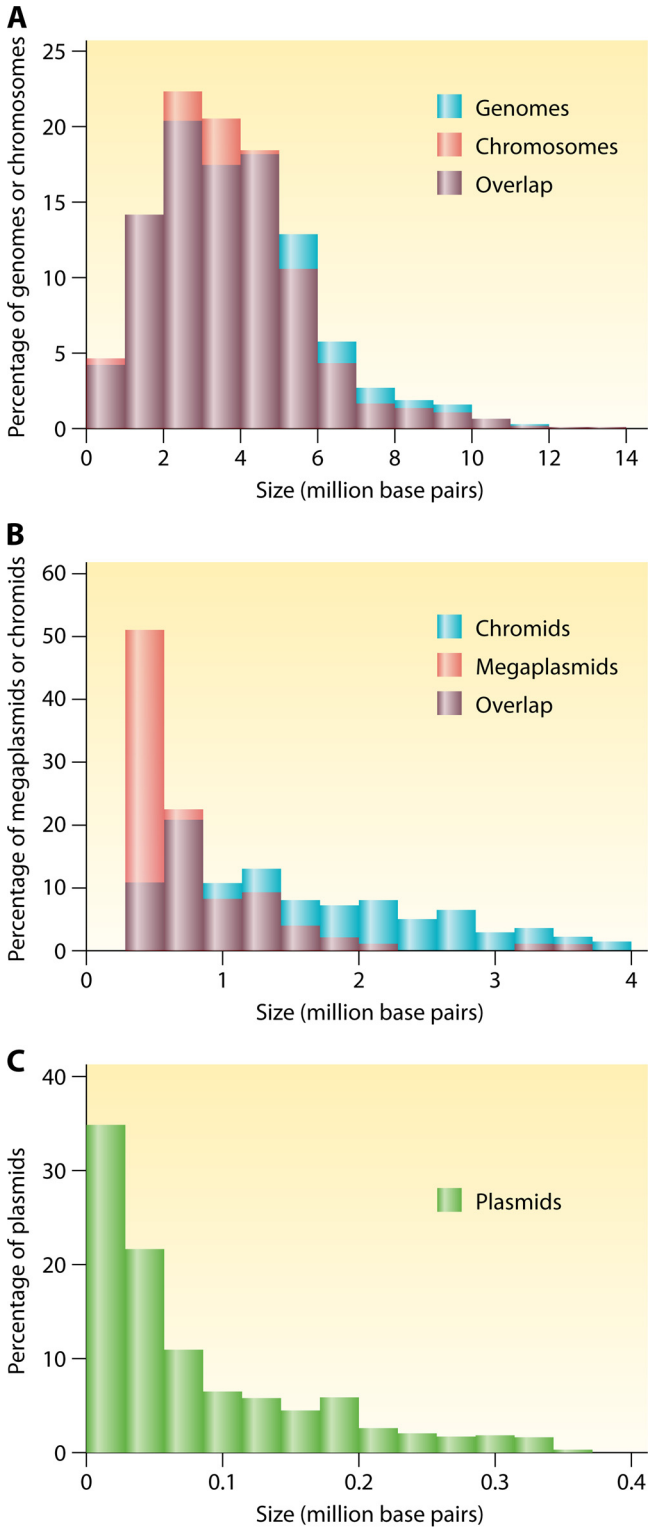


FIG 2 Size distributions of bacterial genomes and replicons. These histograms display the size distributions of all bacterial genomes and all bacterial chromosomes (A), chromids and megaplasmids (B), and plasmids (C). The dark reddish-purple color occurs as a result of the overlap between the red and blue bars. Histograms are based on the 1,708 bacterial species with a completed genome available in the NCBI genome database (accessed 21 March 2016). When more than one genome was available for a species, the genome and chromosome sizes were first averaged for each species, and a representative strain was chosen for analysis of the plasmids, megaplasmids, and chromids. Methods are provided in the supplemental material.

TABLE 1 Summary of genomic characteristics

Genome organization	Genome size (Mb)			Chromosomal GC content (%)			Chromosomal SCUO[a]		
	Median	Minimum	Maximum	Median	Minimum	Maximum	Median	Minimum	Maximum
Overall	3.64	0.16	13.1	49.04	14.55	74.91	0.28	0.13	0.7
Nonmultipartite	3.41	0.16	13.1	47.36	14.55	74.91	0.27	0.13	0.7
Multipartite	5.56	2.48	9.73	61.29	28.83	72.94	0.31	0.15	0.56

[a]SCUO (synonymous codon usage order) was calculated with CodonO (99) and is a measure of the extent of codon usage bias, with higher values indicating greater bias.

to a large plasmid. Similarly, size is used as the sole feature distinguishing megaplasmids from plasmids in this review. However, there may be a less arbitrary means of separating megaplasmids from plasmids that has yet to be elucidated. For example, megaplasmids often have a copy number equal or similar to that of the chromosome, they often encode their own partitioning systems, and the replication and partitioning of megaplasmids can be integrated into the cell cycle. It will be interesting to see if future research can identify a clear, functional distinction between plasmids and megaplasmids aside from an arbitrary size distinction.

Chromid

The term "chromid" itself is a combination of chromosome and plasmid (11) and underscores how chromid refers to a replicon that is an intermediate between a plasmid and a chromosome (11). The replication systems of chromids are similar to those of plasmids and megaplasmids (11) but may have additional regulatory controls that integrate their replication into the cell cycle (29, 36, 37). However, unlike plasmids and megaplasmids, chromids carry at least one gene that is essential for cell viability (i.e., a core gene whose loss would result in cell death) and generally have genomic signatures that better resemble those of the chromosome (11). Ideally, a replicon would be classified as a chromid based on experimental evidence that the replicon carries a nondispensable core gene and not just on genome annotation. In our set of putative chromids, the average (~1.52 Mb) and median (~1.26 Mb) sizes are around 2-fold larger than the average and median megaplasmid sizes (~0.77 Mb and ~0.56 Mb, respectively), despite the lower size cutoff of chromids and megaplasmids in our classification scheme being the same. Additionally, the size distribution of the putative chromids displays a weak positive skew (Fig. 2B). The larger size and weaker positive skew of the putative chromids than of megaplasmids may suggest weaker evolutionary pressure to limit replicon size.

It was suggested by Dziewit et al. that chromids be further subdivided into primary chromids and secondary chromids (38). In this classification scheme, primary chromids are absolutely essential for cell viability. In contrast, secondary chromids may be dispensable under some conditions but are expected to be required for survival in the organism's natural habitat. We accept that these subdivisions are potentially useful but would add that the replicon must be essential in the cell's native habitat to be considered a secondary chromid; for example, a secondary replicon in a soil-dwelling, opportunistic pathogen must be essential in the soil to be considered a secondary chromid. Similarly, many secondary replicons, including small plasmids, carry antibiotic or heavy metal resistance genes. These genes are required for growth in environments containing these antibiotics or heavy metals, but we consider environment-specific essentiality such as this to be insufficient for the chromid designation. For the sake of this review, primary and secondary chromids are not differentiated.

It is worth noting that the majority of chromids are considered essential solely on the basis of the genome annotation, and the expectation that the chromid carries a single-copy essential gene is largely without experimental support. These inferences are not always correct; the third-largest replicon of the *Burkholderia cepacia* complex species was thought to be essential based on genome annotations but was since shown to be a virulence megaplasmid (35). Similarly, the *minCDE* genes of the pSymB

replicon of *S. meliloti* were predicted to be essential (39), although follow-up experimentation revealed them to be dispensable (40). However, there are experimentally validated cases of essential genes existing on chromids, such as the *engA* and $tRNA^{arg}$ genes on the pSymB replicon of *S. meliloti* (41). It is therefore important to experimentally validate the essential nature of more putative chromids to develop a true understanding of the prevalence of this replicon type. However, an inability to remove a replicon from the genome should not be considered sufficient to confirm its essentiality, as there may be other explanations, such as the presence of plasmid addiction systems (42–44). For example, despite being a nonessential replicon (34), the pSymA megaplasmid of *S. meliloti* is nearly impossible to forcefully remove (cure) from the cell due in part to the presence of at least three active toxin-antitoxin loci (45, 46).

Second Chromosome

Historically, the term "second chromosome" was used in reference to a replicon that would now be described as a chromid. As nicely described by Harrison et al., chromid is a more appropriate term to describe such replicons, and the use of second chromosome in this respect should no longer be applied (11). As we describe below in Proposed Mechanisms of Chromid Formation, it is highly likely that nearly all secondary replicons carrying essential core genes evolved from plasmids. However, very rarely, a secondary replicon may form as a result of a split of an ancestral chromosome into two replicons, and we propose that the term second chromosome continue to be used to describe this rare occurrence. No documented cases of such an event are present in the literature; however, we found two examples by scanning the complete genomes available in the NCBI genome database. Assuming that these are not errors in the genome assembly, synteny analysis revealed that the ~0.73-Mb replicon of *Salmonella enterica* strain NCTC10384 and the ~2.66-Mb replicon of *Nocardia farcinica* NCTC11134 represent second chromosomes (see Fig. S1 in the supplemental material). Although it may be difficult to differentiate between a second chromosome and a chromid when the second chromosome was formed through a very ancient split, in general, differentiation between these replicon classes should be possible. Second chromosomes are expected to show high synteny to the chromosomes of related species, depending on the age of the split, and the distribution of core genes between the two replicons is expected to be random, unlike that for chromids. For the analyses presented in this review, second chromosomes were not differentiated from chromids due to their low abundance, and neither of the above-mentioned strains are included in our list of representative strains.

Classification of the Replicons Present in the NCBI Genome Database

As of 21 March 2016, the NCBI genome database contained 4,541 genomes, representing 1,708 bacterial species, that were marked as a "complete genome." As a way of reviewing the untapped information held within these genome sequences and to examine whether conclusions based on intensive research on a limited number of species were generalizable, we downloaded the RefSeq (47) version of all 4,541 complete genomes; annotated each replicon as "chromosome," "putative chromid," "megaplasmid," or "plasmid"; and performed several analyses on each class of replicon. By using the RefSeq version, we could be sure that all genomes were consistently annotated using the NCBI prokaryotic genome annotation pipeline. The complete set of replicons identified in the database is provided in File S2 in the supplemental material. In many of the subsequent analyses, we did not examine all 4,541 genomes but instead chose one random representative genome for each species in order to limit bias due to certain species being overrepresented in the database. We did not perform further controls for phylogenetic structure, such as controlling for genera or families that were overrepresented. We also did not attempt to determine whether two secondary replicons in related genomes shared a common ancestry, but if one were to attempt such an analysis, we feel that common ancestry should be based on phylo-

genetic analysis of the replication/partitioning proteins and not simply on gene content or synteny.

The annotation of the replicons was performed as described below, which largely follows the process outlined in Fig. 1, with some exceptions. First, the largest replicon in each genome was annotated as the chromosome. This step did not involve a search for essential genes due to the inability of us to do so on such a large scale; however, as the largest replicon is always the chromosome (11), chromosomes can be reliably annotated solely on the basis of size. Next, any replicon that was below 350 kb was classified as a plasmid, as these replicons were below our size threshold for megaplasmids and putative chromids. Megaplasmids and putative chromids were distinguished on the basis of GC content and dinucleotide relative abundance. A replicon with a GC content within 1% of the corresponding chromosome and with a dinucleotide relative abundance distance from the chromosome of $\leq$0.4 (Fig. S2) was considered a putative chromid. Otherwise, the replicon was called a megaplasmid. Each of the 1,708 species was considered to have a megaplasmid or a putative chromid as long as one was found in at least one genome available for that species. Second chromosomes were not differentiated from putative chromids due to the rarity at which second chromosomes occur and the impracticality of manually examining each putative chromid to determine if it was a second chromosome. As defined above, a replicon must have an essential core gene to be classified as a chromid; however, the massive computational requirements to perform this analysis on such a large scale prevented us from doing so. Therefore, putative chromids and megaplasmids were differentiated on the basis of genomic signatures, which are often good proxy measures for distinguishing these replicons types. The limitations of the methods used here are discussed further in the supplemental material. While the methods used here are certainly imperfect, we believe that a small number of misclassified replicons will have a limited influence on the analyses described below and will not significantly impact the conclusions that are drawn. Indeed, the ability to detect differences between the putative chromid group and megaplasmid group with respect to several nonselected characteristics, as described below, supports this notion.

REPLICATION AND SEGREGATION DYNAMICS IN MULTIPARTITE GENOMES

How cells ensure the orderly replication and segregation of each replicon in a multipartite genome has been studied to some extent for several species but is by far best studied using *Vibrio cholerae* as the model system. As this topic has been reviewed in depth elsewhere in recent years (28–31), here only key aspects are covered, and an update on the most recent literature is provided.

The currently available data suggest that chromids and megaplasmids generally have a low copy number similar to that of the chromosome. The chromids and/or megaplasmids whose copy numbers have been examined in the family *Rhizobiaceae* (48–50) and the genera *Burkholderia* (51) have a copy number approximately equal to that of the chromosome. Similarly, the copy numbers of the large plasmids of *Thermus thermophilus* (52) and *Sphingomonas wittichii* (53) are similar to those of the chromosomes. However, the copy number of the chromid of the fast-replicating organism *V. cholerae* can actually be lower than that of the chromosome depending on the growth medium (54).

In *V. cholerae*, replications of the chromosome and chromid are initiated at different time points of the cell cycle such that the termination of replication occurs simultaneously (55). Similarly, the replication of the chromosome of *S. meliloti* is initiated prior to that of the chromid and megaplasmid, although the timing of termination has not been examined (56). Initiations of the replication of the *V. cholerae* chromosome and chromid are controlled by distinct factors (57, 58). The integration of the chromid into the *V. cholerae* cell cycle is accomplished at least in part through chromosomal factors that regulate the initiation of chromid replication (59). In particular, the interaction of the chromid replication initiator protein RctB with the chromosomal *crtS* locus promotes RctB binding to iterons within the chromid origin of replication (59), with chromid

replication being initiated following the replication, and, thus, the duplication, of the *crtS* locus (60). The termini of replication of the chromosome and chromid physically interact at midcell (60). The chromosomal terminus remains at midcell until cell division, whereas the terminus of the chromid is segregated slightly before cell division (61, 62). Moreover, the chromid contains binding sites for SlmA, an inhibitor of FtsZ polymerization (63) that contributes to cell cycle control, although there does not appear to be a checkpoint to ensure that the replication of the chromid occurs before cell division begins (64).

The segregation of the three large replicons of *Burkholderia cenocepacia* follows an orderly manner, with the segregation of the origin of the newly replicated chromosome generally preceding that of the origin of the chromid, after which the segregation of the origins of the small chromid/megaplasmid occurs (51). In contrast, chromosomal segregation in *S. meliloti* is initiated first, followed by the segregation of the megaplasmid and, finally, the chromid (65). In *B. cenocepacia*, the segregation of the chromosome and chromid appears to be highly integrated into the cell cycle, while megaplasmid segregation is more variable (51). This is similar to *S. meliloti*, where the segregation of all three replicons appears to be highly integrated into the cell cycle (56, 65). Interestingly, all *S. meliloti* replicons contain genes whose transcription is cell cycle dependent (e.g., *groEL2* on pSymA, *minCDE* on pSymB, and *divK* on the chromosome), with most cell cycle-regulated genes on the chromosome, an intermediate number on the chromid, and the least on the megaplasmid, consistent with each element being integrated into the cell cycle to various extents (56). The segregation machinery of each replicon in *B. cenocepacia* is specific to the corresponding replicon (51, 66, 67). Similarly, the segregation machinery of each of the four secondary replicons in *Rhizobium leguminosarum* can distinguish between the self replicon and the others (68). Finally, the DnaA protein of *S. meliloti* is involved in the replication of the chromosome but not the chromid or megaplasmid (65).

Overall, it appears as though the replication and segregation of chromids and large megaplasmids can become integrated into the overall cell cycle of the host organism, although the precise details are likely to differ between species.

PROPOSED MECHANISMS OF CHROMID FORMATION

Two primary hypotheses describing the process through which an essential secondary replicon may be formed have been put forth: the schism hypothesis and the plasmid hypothesis (23, 24, 26, 31, 69). As described below, the plasmid hypothesis almost certainly represents the mechanism accounting for the formation of essential secondary replicons in most, if not essentially all, species examined to date. In this section, the data supporting and opposing these views are presented, why the data support the plasmid hypothesis is described, and two mechanisms for how a chromid may evolve from a megaplasmid are provided.

The Schism Hypothesis

The schism hypothesis states that a second essential replicon is formed as a result of a split of an ancestral chromosome into two replicons, a chromosome and a chromid. The schism hypothesis is the older of the two ideas and was initially proposed to describe the chromid formation of *Brucella suis* (70) and *Rhodobacter sphaeroides* (71). If the schism hypothesis is correct, it would predict that the properties of the two resulting replicons are highly similar, with an equal distribution of core genes between both replicons. The ability to produce viable *E. coli* or *Bacillus subtilis* strains that have had their single chromosome artificially split into two self-replicating chromosomes provides support to show that such a scenario is possible (72, 73). However, the strong enrichment of essential genes on the chromosomes of species with multiple replicons is inconsistent with this model (11). Additionally, evidence now indicates that the chromids of *B. suis* and *R. sphaeroides* did not result from a schism event.

In the case of *B. suis*, biovars 1, 2, and 4 contain a chromid, while biovar 3 has a single chromosome with a size equal to that of the chromosome plus the chromid of the other

biovars (70). It was originally proposed that the single-replicon structure was ancestral (70); however, phylogenetic analysis subsequently showed this was not true and that the single chromosome of biovar 3 resulted from a fusion of the chromosome and chromid in this lineage (23). For *R. sphaeroides*, the many genomic features being highly similar between the 3.2-Mb chromosome and the 0.94-Mb chromid (71), the large number of gene duplications between these replicons (74), and the large number of genes on the chromid predicted to be essential (75, 76) led to the suggestion that the chromid resulted from a split of an ancestral chromosome. However, the lower coding density of the chromid than of the chromosome (76, 77), gene functional biases as determined by Cluster of Orthologous Genes (COG) analyses (76), differences in evolutionary rates (78), and variations in gene content and size (78, 79) are inconsistent with this view. Overall, there is little evidence for the formation of a secondary essential replicon through the schism hypothesis occurring in nature except perhaps very rarely, and as mentioned above, we recommend that such replicons be referred to as second chromosomes.

The Plasmid Hypothesis

Whereas the schism hypothesis predicts that a secondary essential replicon evolves from a chromosome, the plasmid hypothesis states that it evolves from a megaplasmid. According to the plasmid hypothesis, the sustained coevolution of a megaplasmid with a chromosome will result in a regression of the genomic signature of the megaplasmid to that of the chromosome and the gain of essential genes potentially through transfer from the chromosome. As summarized above and as described by Harrison et al. (11), such replicons should be referred to as chromids.

In support of this model, the replication and partitioning machinery of chromids resembles that of megaplasmids (11), although in many cases, experimental evidence for the functionality of these systems is lacking (31). That said, replicons of the *repABC* family that carry essential genes, and are thus chromids, have *repABC* replication/partitioning genes that have a codon usage more similar to that of the chromosome than do *repABC* family members that do not carry essential genes (80), consistent with these replication systems being functional. Additionally, data from a phylogenetic analysis of the plasmid partitioning protein RepA were consistent with chromids evolving from preexisting megaplasmids in the *Alphaproteobacteria* (11).

There are two main observations that can be explained by the plasmid hypothesis but that cannot be accounted for by the schism hypothesis. Essential genes are strongly underrepresented on chromids (11), as would be expected if the chromid originated as a megaplasmid that subsequently gained a few core genes, for example, through interreplicon gene transfers. There is also a consistently observed bias in the functional annotation of genes present on chromids versus genes on chromosomes, as determined via COG analyses (see, for example, references 19, 21, 22, and 76). This is not surprising if the chromid and chromosome originated independently. In contrast, equal distributions of core genes and of functional annotations would be expected if the chromid formed as a result of a chromosomal schism. Hence, *in toto*, it appears as though the plasmid hypothesis is likely to explain the formation of most, if not all, chromids studied to date.

Conversion of a Megaplasmid into a Chromid

The transition of a replicon from a megaplasmid into a chromid requires two main conversions: the amelioration of the genomic signatures to that of the chromosome and the acquisition of core/essential genes. Genomic signatures such as codon usage and dinucleotide composition are shaped by a variety of factors and can have adaptive advantages (81, 82). Therefore, the similarity of the genomic signatures between chromosomes and chromids in the same species is presumably driven by evolutionary forces selecting for optimized genome function and can be caused by, for example, selection for improved translational efficiency or mutational biases of the cellular machinery (81, 82).

Less intuitive is how to explain the occurrence of essential genes on a chromid when the cell was fully capable of surviving without this replicon in the past. There are two possible mechanisms for this process. The primary process is through interreplicon translocations resulting in the transfer of essential genes from the chromosome to the secondary replicon. Perhaps the best-supported example of this mechanism is in *S. meliloti* (Fig. 3A). Two essential genes have been experimentally demonstrated to exist on the *S. meliloti* pSymB chromid, *engA* and an unique arginine tRNA, $^{ARG}tRNA_{CCG}$ (41). Computational analysis of the surrounding region demonstrated that their presence on pSymB is the result of the translocation of a contiguous 69-kb fragment, including *engA* and the tRNA, from the chromosome to the pSymB precursor in a recent *S. meliloti* ancestor (41, 83). In *V. cholerae*, there are four putatively essential genes (*dsdA*, *thrS*, L20, and L35) present in two clusters on the chromid, and all of these genes are chromosomally situated in related *Vibrio* species (31). Numerous other clusters of genes in the *Vibrio* and *Burkholderia* genera and the order *Rhizobiales* are predicted to have moved from the chromosome to a secondary replicon (84). Similarly, 25 to 30% of genes on the *S. meliloti* chromid have orthologs on the *Agrobacterium tumefaciens* chromosome (85), suggesting significant amounts of interreplicon gene flow, which is supported by the results of a genealogy study that are suggestive of recombination between the *S. meliloti* replicons in nature (86). Moreover, a phylogenetic analysis of individual genes between *Bacillus cereus* strains indicated the frequent transfer of genes between chromosomes and plasmids (87). The precise mechanism through which gene transfer from the chromosome to a secondary replicon occurs has not been studied. However, considering that the multiple replicons of a multipartite genome can naturally form cointegrants (88, 89), it may be that the integration of the replicons followed by an imprecise excision event results in interreplicon translocations (90). Alternatively, a recombination event, mediated, for example, by insertion sequence (IS) elements, may result in the excision of a chromosomal gene region that is subsequently captured by the secondary replicon.

The second putative mechanism through which secondary replicons can come to carry core genes is genetic redundancy. It was experimentally shown through transposon mutagenesis that >10% of chromosomal genes in *S. meliloti* may have a functionally redundant copy on one of the secondary replicons (91). Based on sequence similarity, there also appear to be many gene duplications between the chromosome and secondary replicons of *R. sphaeroides* (74, 92), *V. cholerae* (21), *B. cereus* (87), and *Burkholderia vietnamiensis* (93). Genetic redundancy between core genes on the chromosome and the chromid could be a result of an interreplicon duplication of a chromosome gene or the acquisition of an orthologous gene through horizontal gene transfer. If the copy of the gene on the secondary replicon is able to fully complement the disruption of the chromosomal version, the degeneration of the chromosomal copy would be fitness neutral, and the second copy of the gene would become the sole copy, in effect transferring a core gene to a secondary replicon.

PHYLOGENETIC DISTRIBUTION OF MULTIPARTITE GENOMES

Analyses of the distribution of multipartite genomes have focused mostly on the distribution of chromids, with little attention being paid to the distribution of megaplasmids, the evolutionary precursor of chromids. Based on the rationale described above, Harrison et al. reported in 2010 that ~10% of all complete bacterial genomes (1,086 genomes) contained a chromid (11). Organisms containing a chromid are enriched in the proteobacteria, including members of the alpha-, beta-, and gammaproteobacteria, but chromids can also be detected in phylogenetically distant genera, including, among others, *Prevotella*, *Leptospira*, and *Deinococcus* (11, 94). The number of complete genomes sequences has drastically increased since 2010, with 4,541 complete genomes now available (NCBI genome repository; accessed 21 March 2016). The distribution of both chromids and megaplasmids was therefore reexamined.

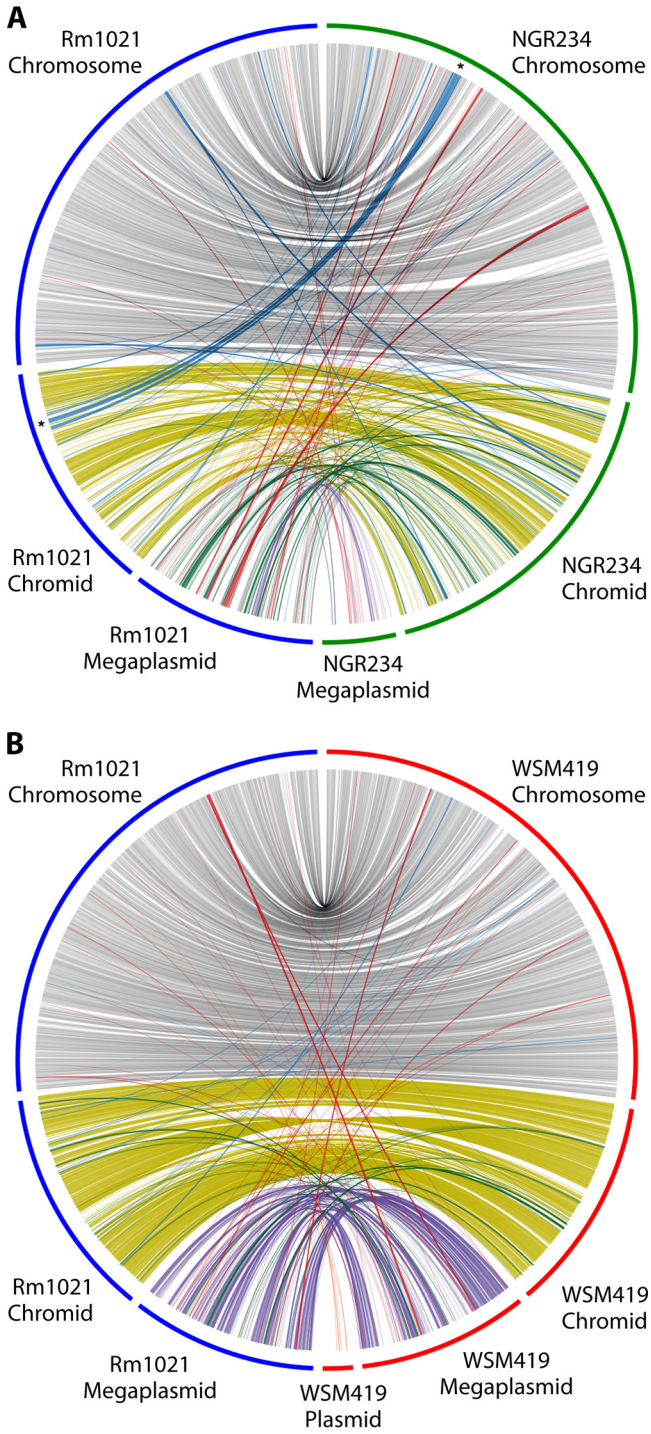


FIG 3 Synteny analysis of the *S. meliloti* genome. The *S. meliloti* Rm1021 genome was compared with the genomes of the more distantly related organism *S. fredii* NGR234 (A) and the closely related organism *S. medicae* WSM419 (B). Putative orthologous genes between species were identified by performing BLAST bidirectional best-hit analyses using the proteomes. BLAST bidirectional best hits with an E value of $\leq 1 \times 10^{-100}$ and $\geq$50% identity were linked to the corresponding gene, and their position was mapped on the genome. Each putative ortholog between genomes is connected by a line and color coded based on replicon type (black, chromosome to chromosome; yellow, chromid to chromid; purple, megaplasmid to megaplasmid; blue, chromosome to chromid; red, chromosome to megaplasmid; green, chromid to megaplasmid; orange, plasmid to anywhere. * indicates a 69-kb region of the *S. fredii* chromosome that translocated to an ancestor of the pSymB chromid through a single translocation event, resulting in the transfer of two essential genes to pSymB (41). Methods are provided in the supplemental material.

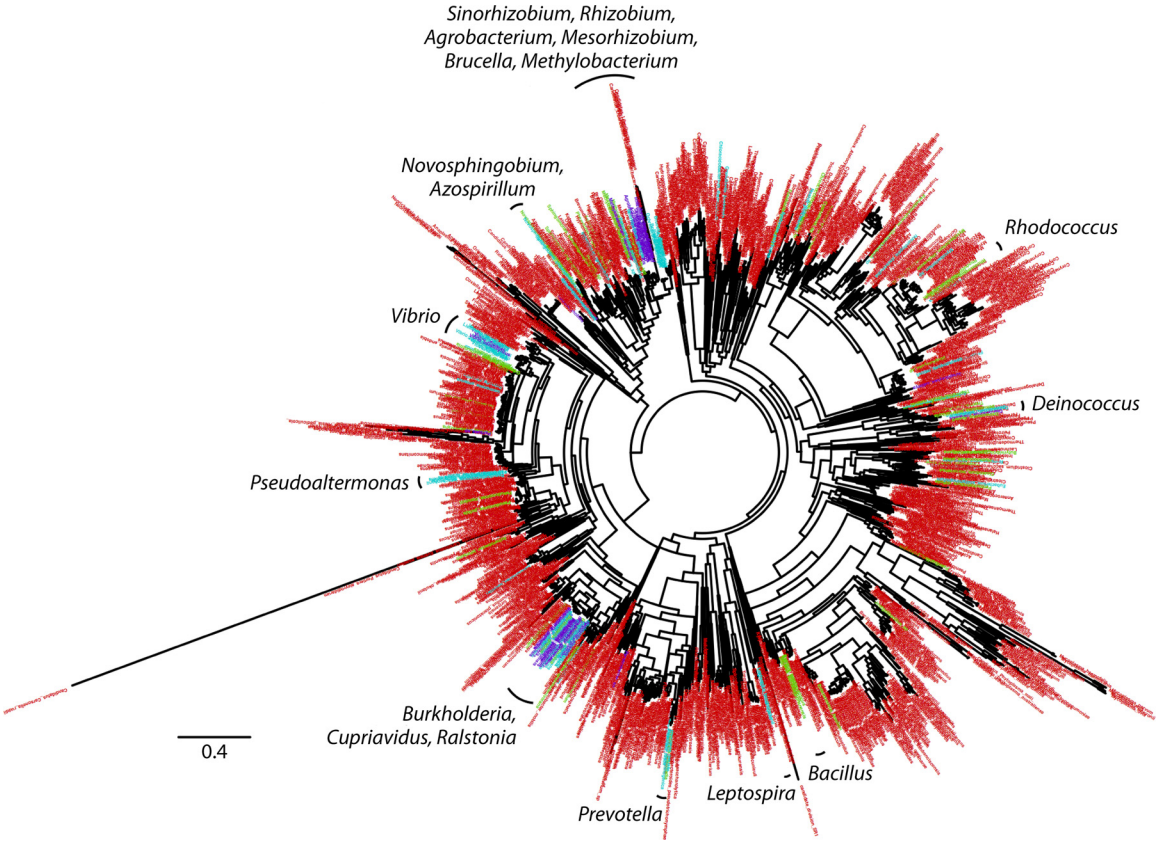


FIG 4 Distribution of multipartite genomes throughout the bacterial phylogeny. A phylogenetic distribution of 1,708 bacterial species with a complete genome available in the NCBI genome database is shown (accessed 21 March 2016). The taxon names are colored based on genome structure, with red for species with no megaplasmid or chromid, green for species with a megaplasmid(s) but no chromid, blue for species with a chromid(s) but no megaplasmid, and purple for species with both a megaplasmid(s) and a chromid(s). Several genera enriched for megaplasmids and/or chromids are labeled. For species with more than one completed genome available in the NCBI database, the species was considered to have a megaplasmid or chromid as long as it was present in at least one strain. For the construction of the phylogeny, 12 ribosomal proteins (RplA, RplC, RplE, RplF, RplN, RplP, RplT, RpsC, RpsE, RpsI, RpsK, and RpsM) present as a single copy in at minimum 1,704 species were identified with the help of the AMPHORA2 pipeline (215). Each set of proteins was aligned with Clustal Omega (216), and the alignments were trimmed with trimAl (217) and then concatenated. The phylogeny was produced based on the concatenated alignment using the RAxML BlackBox mirror site on the CIPRES Gateway Web server (218, 219), and the bootstrap best tree following 204 bootstrap replicates is shown. High-quality images of the phylogeny are provided in Fig. S5 and S6 in the supplemental material. A Newick-formatted tree with bootstrap values as well as an annotation file are available upon request. Methods are provided in the supplemental material.

Phylogenetic Distribution of Secondary Replicons

Of the 1,708 bacterial species examined, 11.2% included strains with a multipartite genome (megaplasmid and/or chromid), and 7.4% or 6.4% included at least one strain with at least one putative chromid or one megaplasmid, respectively (Fig. 4; see also the supplemental material). Moreover, there appeared to be an affinity for putative chromids to cooccur with megaplasmids, as ~2.5% of all the species examined had both a chromid and a megaplasmid (although not necessarily in the same strain). While some of this apparent cooccurrence of putative chromids and megaplasmids may reflect the difficulty in clearly distinguishing between these replicons, we note that the majority of species that appeared to have both putative chromids and megaplasmids were in the genus *Burkholderia* and the order *Rhizobiales*, which are known to carry both elements. The apparently higher prevalence of putative chromids than of megaplasmids in the bacterial phylogeny was surprising given that chromids appear to have evolved from megaplasmids. This may reflect a greater instability or more dynamic nature of megaplasmids than of chromids.

By using the *ace* function of the *ape* package in R (95), it was predicted that putative chromids arose 45 times in the bacterial phylogeny and were lost only twice (see the

supplemental material for methods and the limitations of this analysis). Of the 126 species containing a putative chromid, the large majority (~91%) of them contained only one. At most, five putative chromids were detected in a single species, and strikingly, all three species with five putative chromids belonged to the genus *Azospirillum*. Similar to species with a putative chromid, ~88% of the 109 species containing a megaplasmid had only a single megaplasmid. An additional ~11% of these species had two megaplasmids, and only the agarolytic marine bacterium *Persicobacter* sp. strain JZB09 had three. In contrast, ~51% of the 627 species with a plasmid contained more than 1 plasmid, ~12% had at least 5 plasmids, and nearly 2% had 10 or more plasmids. At most, 21 plasmids, accounting for ~40% of the total genome, were identified in a single genome; this was observed for *Borrelia burgdorferi*. In fact, all four species with >15 plasmids belonged to the genus *Borrelia*.

Multipartite genomes were dispersed throughout the bacterial phylogeny, but clear clusters of species with multipartite genomes are visible (Fig. 4). In particular, megaplasmids were observed to be common in genera that contain numerous soil and marine bacteria that interact with eukaryotic species in either a symbiotic or a pathogenic relationship. These genera included *Bacillus*, *Burkholderia*, *Sinorhizobium*, *Rhizobium*, *Mesorhizobium*, *Agrobacterium*, and *Methylobacterium*. Megaplasmids were also enriched in the genera *Rhodococcus* and *Novosphingobium*, which contain soil and marine organisms capable of inhabiting polluted environments and catabolizing the pollutants. Putative chromids were similarly found to be prevalent in several genera with species that enter into symbiotic or pathogenic relationships with eukaryotic organisms. These genera included *Sinorhizobium*, *Rhizobium*, *Agrobacterium*, *Burkholderia*, *Cupriavidus*, *Vibrio*, *Pseudoalteromonas*, *Azospirillum*, *Ralstonia*, and *Prevotella*. Putative chromids were also prevalent in a few genera containing species that are able to survive in extreme environments due to resistance to several stresses such as UV irradiation, metal ions, and aromatic compounds. These genera included *Ralstonia*, *Deinococcus*, and *Cupriavidus*. The lack of additional clusters in the phylogeny may simply reflect sequencing biases and the underrepresentation of genome sequences from certain taxa. For example, the only representative genome for each of the genera *Persicobacter* (agarolytic marine bacterium), *Tistrella* (soil and marine bacteria), *Chelatococcus* (marine moderate thermophiles), and *Chloracidobacterium* (marine moderate thermophiles) contained a putative chromid, and the sequencing of additional genomes from these genera may reveal that the presence of a multipartite genome is characteristic of these genera.

It was previously noted that chromids appear to contain genus-specific genes, and the presence of a chromid may correspond to the emergence of a new genus (11). This observation remained largely true in this expanded data set, although some exceptions were detected, where the presence of a putative chromid was not a defining characteristic of the genus. For example, *R. sphaeroides* contains a chromid, whereas *Rhodobacter capsulatus* does not; *Xanthomonas sacchari* contains a putative chromid, whereas the other seven *Xanthomonas* species do not; and only a few *Deinococcus* species have a putative chromid, but the species having a putative chromid did not form a monophyletic group in the phylogeny (Fig. 4). On the other hand, the acquisition of a chromid may also predate the emergence of a genus. For example, it was argued that the chromids of the genera *Sinorhizobium*, *Rhizobium*, and *Agrobacterium* were acquired by the common ancestor of these genera prior to their divergence (84).

In contrast to chromids, megaplasmids are rarely conserved at the genus level, although multiple species in a genus will often contain a megaplasmid. Even in the rare cases where megaplasmids are present in all species of a genus, different species may have unique megaplasmids. For example, all *Sinorhizobium* species have at least one strain with a megaplasmid, but analysis of the replication and partitioning proteins suggests that they do not share a common ancestry (96).

GENOMIC SIGNATURES OF BACTERIAL REPLICONS

Several genomic features vary between species. These features include codon usage (the ratio at which synonymous codons are present in a genome), GC content (percentage of the genome consisting of guanine and cytosine), and dinucleotide relative abundance (the frequency with which each pair of nucleotides appears in the genome). These same features have also been shown to differ between replicons of the same genome, with the extent of the differences being reflective of the type of replicon. While these differences are often not very strong, they are robustly and reliably observed. Here we review the relevant literature and provide an analysis of all replicons from a representative genome for each of the 1,708 bacterial species that we examined (1,708 chromosomes, 139 putative chromids, 99 megaplasmids, and 1,114 plasmids).

Codon Usage

The codon usage of a gene is correlated with the expression level of the gene; highly expressed genes have a codon usage that closely mimics the relative tRNA abundance, whereas lowly expressed genes often do not (97). Differences in codon usage bias between replicons have been reported for numerous species (11, 98). For example, Cooper et al. (98) examined codon usage bias in 22 species, including species of the *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, and *Deinococci*, and in all cases, the codon usage biases of the chromosomes were greater than those of chromids, which in turn were greater than those of additional chromids or megaplasmids.

An analysis of codon usage bias was performed on representative genomes from 1,708 bacterial species by using CodonO to calculate synonymous codon usage order (SCUO) (see the supplemental material) (99). It was found that 85.6% of putative chromids and 85.9% of megaplasmids had less codon usage bias (a lower SCUO value) than that of the corresponding chromosome, with a median SCUO difference between the chromosome and a putative chromid or megaplasmid of -0.02 or -0.06, respectively (Fig. 5A). Somewhat surprisingly, only 61.7% of plasmids had an SCUO value lower than that of the chromosome, with a median difference of -0.02 (Fig. 5A). Additionally, the difference in SCUO values between chromosomes and megaplasmids displayed an unexpected bimodal distribution, unlike the bell-shaped distribution for putative chromids (Fig. 5A). However, both of these unexpected results appear to be due to the inclusion of genomes with low codon usage bias in the analysis (Fig. S3). When the analysis was limited to genomes with a chromosomal SCUO value above the median (Fig. 5A, inset), the bimodal distribution for megaplasmids was largely eliminated, and 77.4%, 95.5%, and 82.3% of plasmids, megaplasmids, and putative chromids, respectively, displayed less codon usage bias than the chromosome. The average differences in SCUO values from the chromosome were -0.05, -0.08, and -0.02 for plasmids, megaplasmids, and chromids, respectively, and the means of all three differences were statistically different from zero (uncorrected P value of $<1e^{-15}$ by a one-sample t test) and from each other ($\alpha = 0.05$ by one-way analysis of variance [ANOVA] with Tukey's honestly significant difference [HSD] *post hoc* test). Overall, these results are consistent with secondary replicons generally displaying less codon usage bias than the chromosome and with the codon usage bias of chromids being higher than that of megaplasmids and plasmids, at least in genomes with high chromosomal codon usage bias.

GC Content

GC contents vary considerably in prokaryotic organisms and can range from ~15% (14.55% in "*Candidatus* Carsonella ruddii" HT) to ~75% (74.91% in *Anaeromyxobacter dehalogenans* 2CP-C). Several factors can influence the GC content of an organism, among which are environmental adaptation (100, 101) and possibly recombination (102). In addition to differing between species, GC contents can vary considerably within a genome and have often been used to identify genes recently acquired through horizontal gene transfer (103). The GC content of each replicon in a multipartite

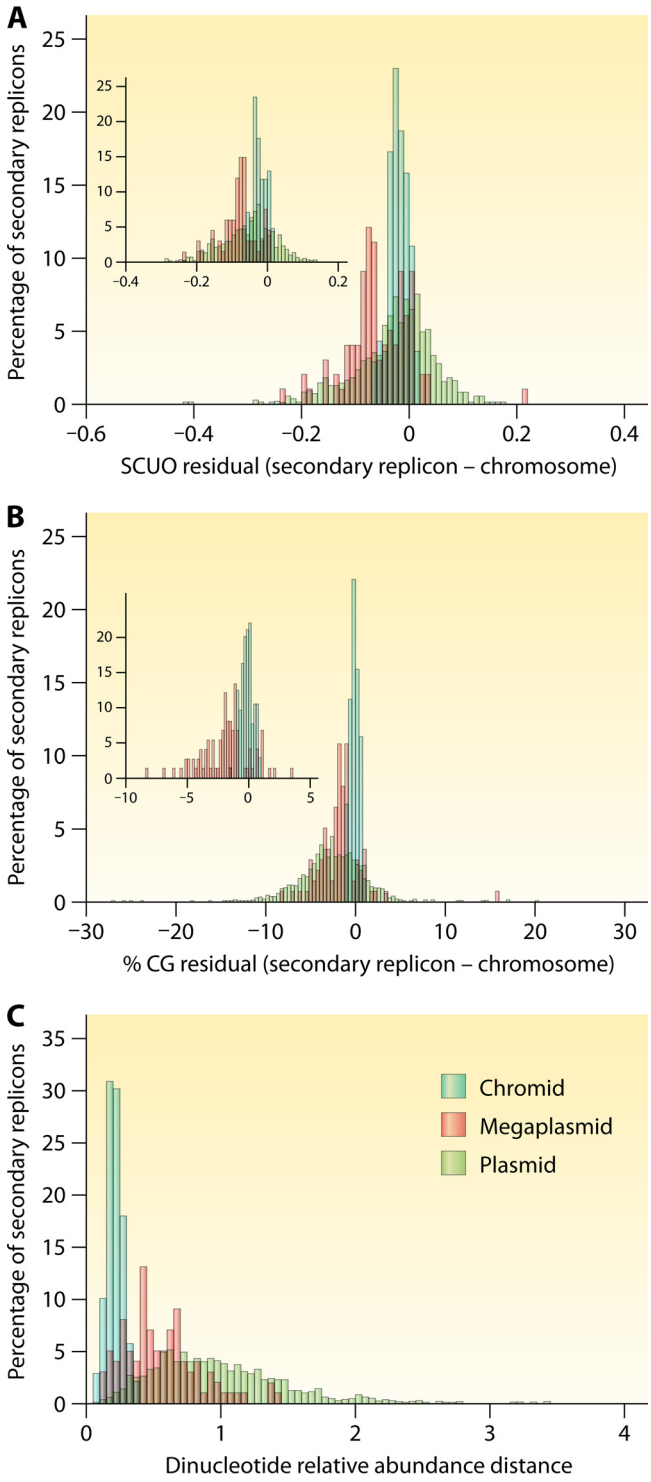


FIG 5 Genomic signatures of bacterial secondary replicons. The analysis was based on one representative genome from each of the 1,708 bacterial species (139 chromids, 99 megaplasmids, and 1,114 plasmids) with a completed genome available in the NCBI genome database (accessed 21 March 2016). (A) Codon usage bias as measured via synonymous codon usage order (SCUO) was determined for each replicon, the value of the chromosome was subtracted from the value of each secondary replicon, and the distribution of the resulting values are presented for plasmids (green), megaplasmids (red), and chromids (blue). The inset displays the results if only genomes containing a chromosome with an SCUO value above the median chromosomal SCUO value are examined. (B) The difference in GC contents of each secondary replicon compared to the corresponding chromosome was determined, and the distributions of the differences are presented for plasmids (green), megaplasmids (red), and chromids (blue). The inset displays an enlargement of the central region of the histogram and shows just the megaplasmids and

(Continued on next page)

genome is almost always different, and the extent of the difference is reflective of replicon type; the GC contents of a chromid and a megaplasmid usually differ by <1% and >1%, respectively, from that of the chromosome (11). In an analysis of one representative genome from 1,708 bacterial species, the median absolute GC content difference between a chromosome and a putative chromid was 0.34% (standard deviation [SD], 0.29%), that between a chromosome and a megaplasmid was 1.9% (SD, 2.0%), and that between a chromosome and a plasmid was 2.8% (SD, 3.1%) (Fig. 5B), with each difference being statistically different from zero (uncorrected *P* value of $<1e^{-15}$ by a one-sample *t* test) and from each other ($\alpha = 0.05$ by one-way ANOVA with Tukey's HSD *post hoc* test) (see the supplemental material).

Not only does the extent of differences in GC contents differ between putative chromids and megaplasmids/plasmids, the direction of the difference also appeared to differ. It was previously observed that the majority of plasmids have a GC content lower than that of the chromosome (104–106). Similarly, in the analysis presented here, the majority of plasmids (78.5%) and megaplasmids (89.4%) had a GC content lower than that of the chromosome (Fig. 5B), with the mean of each distribution being statistically different from zero (uncorrected *P* value of $<1e^{-8}$ by a one-sample *t* test). However, only little bias in the direction of the GC content difference for the putative chromids relative to chromosomes was observed; 58.2% of chromids had a GC content lower than that of the chromosome, while 41.2% had a higher GC content (Fig. 5B), with rather low statistical support for the mean of the distribution differing from zero (uncorrected *P* value of 0.013 by a one-sample *t* test). It was suggested that the lower GC content of plasmids is due to selection for reduced energy expenditure, as the maintenance of GC-rich sequences is energetically more expensive (104). Perhaps, as chromids cannot be lost from the genome, selection for reduced energy expenditure is largely absent, and the GC content is shaped almost solely by the same forces acting on the chromosome. Alternatively, evidence suggests a general mutation bias toward AT due to G/C to A/T transitions (107, 108). The reduced GC content of plasmids/megaplasmids, but not chromids, may be reflective of more relaxed selection acting on these replicons (98).

Dinucleotide Relative Abundance

The profiles of dinucleotide relative abundances in a genome have been shown to be distinct for each bacterial genome and are reflective of bacterial phylogeny (105, 109). Studies have also illustrated that dinucleotide relative abundances can be used to differentiate chromosomes from chromids and from plasmids (81). The dinucleotide relative abundance distance refers to the sum of the differences in the frequencies of each dinucleotide pair between two sources of DNA. In the current analysis of 1,708 representative genomes, it was seen that the median absolute difference between a chromosome and a putative chromid was 0.21 (SD, 0.06), that between a chromosome and a megaplasmid was 0.50 (SD, 0.28), and the difference between a chromosome and plasmid was 0.91 (SD, 0.50) (Fig. 5C), with all differences being statistically different from zero (uncorrected *P* value of $<1e^{-15}$ by a one-sample *t* test) and from each other ($\alpha = 0.05$ by one-way ANOVA with Tukey's HSD *post hoc* test) (see the supplemental material). Thus, as for GC content, putative chromids appeared the most like chromosomes, while plasmids appeared the least like chromosomes.

It is interesting to note that the dinucleotide relative abundance distances of the putative chromids compared to chromosomes appeared to follow a bell-shaped distribution centered away from zero, whereas the difference in the GC contents of the putative chromids compared to chromosomes appeared centered at around

FIG 5 Legend (Continued)
chromids. (C) The dinucleotide relative abundance distance of each replicon compared to the corresponding chromosome was calculated, and the distributions of the distances are presented for plasmids, megaplasmids, and chromids. Colors in addition to green, red, and blue occur as a result of the overlap of the bars. Methods are provided in the supplemental material.

zero (Fig. 5B and C). This suggests that whereas the GC content of chromids is continually ameliorated toward that of the chromosome, there is a constraint on the amelioration of the dinucleotide composition. However, whether this is simply a consequence of the original dinucleotide relative abundance difference between the two replicons or whether this reflects an adaptive function is unclear.

Conjugal Transfer and Interreplicon Genomic Signature Differences

For both GC content and dinucleotide relative abundance, it was observed that the difference between chromosomes and megaplasmids was less than that between chromosomes and plasmids despite both elements being nonessential replicons (Fig. 5). A possible explanation may be that the mobility of megaplasmids is lower than that of plasmids and that successful megaplasmid transfer to phylogenetically distant organisms occurs less frequently than it does for plasmids. Although megaplasmids can retain conjugative machinery (see, for example, references 110–113) and the transfer of megaplasmids between related organisms has been observed in nature (see references 114–117, among others), experimental studies have noted difficulties in promoting megaplasmid transfer between phylogenetically distant species (113). Like megaplasmids, at least some chromids retain conjugative properties and can be induced to transfer to naive cells under laboratory conditions (118). Moreover, the chromid of *S. meliloti* Rm41 is naturally transmissible (119). However, there is no evidence for successful horizontal transfer (transfer and maintenance) of chromids in nature (11), and transfer of the *S. meliloti* pSymB chromid to the related organism *A. tumefaciens* in the laboratory resulted in an obvious fitness decrease (120). Hence, it is likely that plasmids are highly mobile, and many of the plasmids detected by genome sequencing are relatively newly acquired. In contrast, the successful transfer (i.e., the transfer and maintenance) of megaplasmids, and more so for chromids, is less frequent (due to either poor maintenance following transfer or the inability of the replicon to conjugate), meaning that most of the detected megaplasmids/chromids were acquired less recently, providing more time for the amelioration of the genomic signatures.

EVOLUTIONARY TRAITS OF BACTERIAL REPLICONS

The patterns of evolution and the rates of genetic change of each replicon in a multipartite genome are unique. This can be clearly seen in *S. meliloti*, in which (i) the chromosome is structurally stable and primarily vertically transmitted, (ii) the chromid was formed by ancient horizontal gene transfer and is under greater positive selection (particularly in genes for environmental adaptation), and (iii) the megaplasmid is structurally fluid and formed by recent and ongoing horizontal transfer (121). In this section, we review the literature examining how the evolutionary characteristics of each replicon differ, specifically in terms of genetic variability and rates of evolution.

Genetic Variability

The levels of sequence and gene conservation between related bacterial strains and species are different for chromosomes, chromids, and megaplasmids. This can be visualized in Fig. 3, which illustrates how the *S. meliloti* chromosome is highly conserved within the *Sinorhizobium* genus, the *S. meliloti* chromid shows less conservation, and the *S. meliloti* megaplasmid is poorly conserved. Studies have shown that *S. meliloti* chromosomal genes show the highest level of conservation, followed by chromid genes and finally by megaplasmid genes, both when comparing different strains of *S. meliloti* and in an interspecies comparison with *Sinorhizobium medicae* (122, 123). Similarly, in both the *Vibrio* and *Burkholderia* genera, higher percentages of chromosomal genes are conserved between species than are chromid genes, and where applicable, genes on the megaplasmid or the smaller chromid were the least conserved (98, 124, 125). In *R. sphaeroides*, greater synteny between the chromosomes of related strains than between the chromids of the same strains was observed (78). Chromosome- and megaplasmid-specific pangenome analyses of 11 *Bacillus thuringiensis* strains with a megaplasmid by using Roary (126) revealed that whereas 5,153 chromosomal genes

were present in at least 5 genomes (1,984 in all 11 strains), only 163 megaplasmid genes were present in at least 5 of the strains (none in all 11 strains) (see methods and Fig. S4 in the supplemental material). Interestingly, comparisons between the Gram-negative *Cupriavidus* species showed greater ortholog conservation between chromosomes than between chromids, whereas comparison between strains belonging to the same *Cupriavidus* species showed only slightly greater ortholog conservation on the chromosome than on the chromid (27, 127, 128). Similarly, genes on all replicons in *B. cenocepacia* showed strong conservation between strains, but whereas the level of conservation of the chromosome remained high compared to those of other *Burkholderia* species, a low level of conservation of the smaller chromid/megaplasmid was seen, while the larger chromid is highly conserved only in closely related species (125).

The same general observations are detected when nucleotide conservation, instead of gene conservation, is examined. For 7 of the 9 examined species belonging to the genera *Brucella*, *Rhodobacter*, *Burkholderia*, and *Vibrio*, the level of nucleotide identity between the chromosomes of strains belonging to the same species was greater than the level of nucleotide identity between the chromids (129). The same pattern was observed for 9 of 10 intergenus comparisons of related species (129). In a population genomics study, Epstein et al. observed that similar percentages (~95%) of nucleotides of the chromosome and chromid of the reference *S. meliloti* and *S. medicae* genomes were conserved across 32 and 12 strains, respectively, while the level of conservation of the megaplasmid was much lower (<80%) (123).

All considered, these data suggest that chromosomes are the most genetically stable replicons, followed by chromids and finally by megaplasmids. Chromosomes display high synteny at both the species and genus levels. Chromids may be conserved nearly as strongly as chromosomes at the species level; however, the level of conservation drops off at the genus level. In contrast, megaplasmids can display high variability even between strains of the same species.

Evolutionary Rates

Several studies have observed different rates of evolution on each replicon in a multipartite genome. The substitution rate of the chromid of *Vibrio* species is higher than that of the chromosome, whereas purifying selection is weaker on the chromid (98). Similarly, the substitution rate of the chromid in *Burkholderia multivorans* is higher than that of the chromosome but lower than that of the megaplasmid, while purifying selection is greatest on the chromosome, then the chromid, and finally the megaplasmid (98). In *S. meliloti*, the rate of positive selection is highest for the chromid (121). A comparison of orthologous gene products of *Burkholderia xenovorans* to those of *Burkholderia cepacia* indicated that the percent amino acid identity was highest for the chromosome, intermediate for the chromid, and lowest for the small chromid/megaplasmid (22). In contrast, genes involved in rhizobium-legume symbiosis carried by the megaplasmid of *Sinorhizobium* species showed less divergence than those carried by the chromosome or chromid (130). While at first glance, this result conflicts with the *Burkholderia* observations, it is perhaps not surprising, as the megaplasmid is the primary replicon with respect to the symbiosis.

Mutation accumulation studies with *B. cenocepacia* (131, 132) indicated that the rates of the different types of substitution mutations differed across replicons. Additionally, the overall rate of substitution mutations, but not indels, was highest on the chromosome and lowest on the chromid (131). Given that the evolutionary rate of the chromosome is lower than those of the other replicons, it was suggested that the above-mentioned results are consistent with much stronger purifying selection on the chromosome (131). Somewhat different results were observed for *Vibrio* species. In *Vibrio fischeri*, the substitution mutation rate was higher on the chromid than on the chromosome, while no difference was detected in *V. cholerae* (133). The rates of the particular substitutions also varied between replicons (133).

Comparison of conserved sequences between *R. sphaeroides* strains suggested that the chromid is experiencing more rapid evolution than the chromosome (78). However,

when only duplicated genes were considered, there was only a non-statistically significant difference in selective constraint for genes where one duplicate was on the chromosome and the other was on the chromid compared to when both duplicates were on the same replicon (134). This may suggest that the higher rate of divergence of secondary replicons is not an intrinsic property of secondary replicons but instead reflects differences in the types of genes. However, a separate study determined that the elevated evolutionary rates of genes on secondary replicons were due to both an intrinsic property of secondary replicons as well as differences in the types of genes (98). This was done by comparing rates of evolution of genes in the multipartite *Burkholderia* and *Vibrio* genomes to conserved orthologs of related genomes (*Bordetella* and *Xanthomonas*, respectively) that lack secondary replicons (98). Overall, these data suggest that each replicon in a multipartite genome may experience different rates of evolution and unique types of evolutionary pressures and that these differences are at least partially independent of the differences in the gene content of each replicon.

FUNCTIONAL ANALYSIS OF BACTERIAL REPLICONS

Studies have repeatedly observed functional biases between each replicon in a multipartite genome. This is most commonly approached by using COG analysis (135). Core processes are consistently found to be enriched on the chromosome (21, 22, 76, 127, 136). Transport and metabolism, such as for inorganic ions, lipids, amino acids, and carbohydrates, are often enriched on chromids and megaplasmids (18, 21, 22, 38, 76, 127, 136, 137). Genes associated with transcription and regulatory functions, including signal transduction, are also commonly overrepresented on chromids and megaplasmids (18, 21, 22, 127, 136), as are motility-related functions (127, 136, 137). The functional biases of chromids and megaplasmids are likely to differ from that of plasmids; for example, plasmids in *B. cereus* are enriched in replication/recombination/repair, transcription, protein modification/turnover, and cellular trafficking (87). Hypothetical genes and genes of unknown function can also show skewed distributions between each replicon in a genome, in some cases being overrepresented on the chromosome (76) and in other cases being overrepresented on a secondary replicon (18, 21, 127).

Global Replicon Functional Biases

As functional analyses of multipartite genomes have focused on individual species, it is unclear whether the types of functions showing a biased distribution will vary between phylogenetically distant taxa. Therefore, a global COG analysis was performed (see the supplemental material). All genes for each replicon class from a single representative genome of 1,708 species (1,708 chromosomes, 139 chromids, 99 megaplasmids, and 1,114 plasmids) were pooled, regardless of whether the genome was multipartite, and COG analyses were performed. Indeed, several global biases were evident, as summarized in Table 2, consistent with each replicon class having specific functions enriched regardless of phylogeny.

Not surprisingly, core functions were enriched on chromosomes, such as COG classes J, A, and Z. Chromids also generally appeared to be enriched in some core functions compared to megaplasmids, although only the differences in COG class J (translation) and class M (cell wall/membrane/envelope biogenesis) were statistically significant. There was a large overlap in the functional groups enriched in chromids and megaplasmids, although the extents of enrichment often differed (Table 2). COG classes I, Q, and W were similarly enriched on chromids and megaplasmids, while a small but statistically significant difference in COG class K (transcription) was observed. The enrichment in COG class K likely represents a larger number of transcription factors present on these replicons allowing gene regulation in response to numerous environmental signals (e.g., carbon availability). The transport and metabolism of a few types of compounds (COG classes E, G, and P) as well as motility and signal transduction (COG classes N and T), both of which may be related to movement in response to

TABLE 2 Global replicon-specific functional analysis[a]

COG class	Description	Replicon enrichment (fold)				Total no. of genes	Significant comparison(s)
		Chromosome	Chromid	Megaplasmid	Plasmid		
A	RNA processing and modification	0.06	−1.64	−2.61	−2.37	2,212	AB, AC, AD
B	Chromatin structure and dynamics	0.04	−0.48	−1.28	−1.98	1,982	AD
C	Energy production and conversion	0.00	0.22	0.11	−1.24	311,749	AB, AC, AD, BD, CD
D	Cell cycle control, cell division, chromosome partitioning	0.01	−0.83	−0.76	0.95	53,815	AB, AC, AD, BD, CD
E	Amino acid transport and metabolism	0.00	0.46	0.04	−1.42	423,590	AB, AD, BC, BD, CD
F	Nucleotide transport and metabolism	0.05	−0.91	−1.39	−2.08	122,047	AB, AC, AD, BC, BD, CD
G	Carbohydrate transport and metabolism	−0.01	0.50	0.10	−1.07	327,915	AB, AC, AD, BC, BD, CD
H	Coenzyme transport and metabolism	0.03	−0.54	−0.77	−1.46	218,207	AB, AC, AD, BC, BD, CD
I	Lipid transport and metabolism	0.00	0.28	0.22	−1.18	192,881	AB, AC, AD, BD, CD
J	Translation, ribosomal structure, and biogenesis	0.06	−1.69	−2.25	−3.20	272,466	AB, AC, AD, BC, BD, CD
K	Transcription	−0.02	0.57	0.39	−0.35	390,373	AB, AC, AD, BC, BD, CD
L	Replication, recombination, and repair	0.00	−0.85	0.04	1.11	265,234	AB, AD, BC, BD, CD
M	Cell wall/membrane/envelope biogenesis	0.02	−0.13	−0.53	−0.88	294,750	AB, AC, AD, BC, BD, CD
N	Cell motility	0.01	0.16	−0.53	−0.90	97,783	AB, AC, AD, BC, BD, CD
O	Posttranslational modification, protein turnover, chaperones	0.03	−0.51	−0.59	−1.08	181,774	AB, AC, AD, BD, CD
P	Inorganic ion transport and metabolism	−0.01	0.36	0.08	−0.51	259,353	AB, AD, BC, BD, CD
Q	Secondary metabolite biosynthesis, transport, and catabolism	−0.03	0.60	0.65	−0.53	127,189	AB, AC, AD, BD, CD
R	General function prediction only	0.01	0.09	−0.07	−0.85	596,567	AB, AC, AD, BC, BD, CD
S	Function unknown	0.01	0.06	−0.24	−0.64	426,972	AB, AC, AD, BC, BD, CD
T	Signal transduction mechanisms	0.00	0.30	−0.03	−0.92	318,564	AB, AD, BC, BD, CD
U	Intracellular trafficking, secretion, and vesicular transport	0.00	−0.29	−0.16	0.49	121,189	AB, AD, BD, CD
V	Defense mechanisms	0.01	−0.15	−0.13	−0.32	83,369	AB, AD
W	Extracellular structures	−0.08	1.26	1.16	−1.18	324	None
Y	Nuclear structure	0.08	0.00	0.00	0.00	1	None
Z	Cytoskeleton	0.06	−2.53	−1.79	−0.54	936	AB
	Transposases	−1.10	1.22	1.83	6.23	196,590	AB, AC, AD, BC, BD, CD

[a]Results of the COG functional analysis and transposon identification are presented. One representative genome from each of the 1,708 species with complete genomes available in the NCBI database were chosen, and all genes from each replicon type were extracted (5,342,421 chromosome genes, 174,984 chromid genes, 62,606 megaplasmid genes, 79,077 plasmid genes, and 5,659,088 total genes). The genes were annotated with COG categories via WebMGA (220), and transposons were identified based on the RefSeq annotation of the protein fasta files. Enrichment (fold change of the observed compared to the expected values; e.g., a value of 2 indicates twice as many genes had the annotation than expected, whereas a value of −2 indicates half as many genes had the annotation than expected, and a value of 1 or −1 indicates no change from the expected value) for each category for each replicon is given, as is the total number of genes annotated for each class. The letters in the right column indicate which pairwise comparisons were statistically significant (A, chromosome; B, chromid; C, megaplasmid; D, plasmid). For example, AB indicates that the value for the chromosome is statistically different from the value for the chromid. Statistically significant comparisons were determined by using pairwise Fisher exact tests, with an adjusted P value of <0.05 following Bonferroni multiple-test correction. Additional methods are provided in the supplemental material.

external stimuli, were primarily enriched on chromids and less so, if at all, on megaplasmids. No classes were enriched specifically on megaplasmids. Plasmids were enriched in replication-related functions (COG classes D and L) and COG class U, which may be related to the replication and conjugal transfer of the plasmid and to resistance to toxic compounds.

In toto, the global functional analysis revealed that many functional categories of genes are universally enriched on secondary replicons, consistent with secondary replicons playing a conserved role in the biology of these organisms. It was also notable that functional biases could be detected between each of chromids, megaplasmids, and plasmids, supporting that these classifications are biologically relevant.

Distribution of Transposable Elements

As a proxy to examine the prevalence of transposable elements on each type of replicon, the RefSeq protein fasta files for each of the 1,708 bacterial genomes were searched for the term "transposase" (see the supplemental material). Approximately 3.1% of the chromosomal genes were annotated as a transposase, compared to 4.3% of putative chromid genes, 6.3% of megaplasmid genes, and 21.6% of plasmid genes, and all differences between replicon classes were statistically significant (Table 2). Thus, the pattern of prevalence of transposable elements on each replicon appeared to

mimic that of the genomic signatures of these replicons (Fig. 5); i.e., putative chromids appeared most like chromosomes, followed by megaplasmids, with plasmids being very different from the others. Given that the gain of insertion elements is generally deleterious (138–140), perhaps the biases in transposase prevalence reflect differences in the expendabilities of genes on each type of replicon and differences in purifying selection (98).

INTERREPLICON INTERACTIONS

Despite genes on each replicon in a multipartite genome being physically separated, in many cases, there may be interactions between their gene products. The enzymes involved in the multistep pantothenate and lipopolysaccharide biosynthetic pathways are encoded by multiple replicons in *Rhizobium etli* and *Rhizobium leguminosarum* (141, 142). In the case of pantothenate, this may be due to gene transfer from the chromosome to the secondary replicon (141). Similarly, complex biological processes can require genes situated on multiple replicons, as is the case for rhizobium-legume symbiosis (83, 143–145). Additionally, an *in silico* analysis of the seven replicons of *R. etli* predicted functional links between each of the replicons, with the two most recently acquired replicons showing the fewest connections to the others (146).

Interactions between replicons can also occur at a regulatory level. The replication of the chromid of *V. cholerae* is subjected to regulation by chromosomally encoded mechanisms (21, 29). It has also been noted that in *V. cholerae*, the chromosomally encoded RpoS protein regulates genes on both the chromosome and chromid, the chromid gene *hylA* is regulated by the chromosomally encoded HylU protein, and quorum-sensing genes are split between the chromosome and chromid (21). An *in silico* regulon analysis predicted that most *S. meliloti* transcriptional factors regulate genes on the same replicon (147). However, a subset was predicted to regulate genes across multiple replicons, and there was a bias for chromosomal regulators to modulate chromid/megaplasmid genes compared to the number of chromid/megaplasmid regulators predicted to regulate chromosomal genes (147). Consistent with this, the cell cycle regulator CtrA and the symbiotic nitrogen fixation regulator FixJ regulate genes on all three replicons in *S. meliloti* while preferentially regulating genes on the same replicon that they are encoded on (148, 149). In contrast, the *S. meliloti* RpoN sigma factor appears to preferentially regulate genes on other replicons (150, 151). Finally, the complete deletion of the *S. meliloti* chromid resulted in at least a 2-fold change in gene expression levels in ~5 to 10% of chromosomal genes, whereas no statistically significant changes in chromosomal gene expression were observed when the megaplasmid was absent from the genome (G. C. diCenzo, B. Golding, and T. M. Finan, unpublished data). Similarly, in *B. cenocepacia*, the expression of only 55 chromosomal or chromid genes was influenced by the removal of the megaplasmid (35).

COSTS ASSOCIATED WITH MULTIPARTITE GENOMES

As is discussed in the following section, the presence of megaplasmids and chromids may provide certain advantages to the host cell. However, these large replicons may also come with significant costs. Transfer of the *S. meliloti* chromid to *A. tumefaciens* resulted in a reduced growth rate, and the *A. tumefaciens* cells spontaneously lost the chromid (120). Conversely, an *S. meliloti* 2011 strain that lacks the megaplasmid appears to have a slightly higher growth rate than do *S. meliloti* 2011 strains containing the megaplasmid (83, 152). Similarly, an *A. tumefaciens* C58 strain lacking the pATC58 plasmid is able to outcompete the wild type under laboratory conditions (153). When the large megaplasmid of *Pseudomonas syringae* Pla107 is transferred to other *P. syringae* strains or more distantly related pseudomonads, both the growth rate and competitive fitness of the recipient cell were decreased, and the megaplasmid was spontaneously lost (113). In fact, the gain of this megaplasmid influenced an array of phenotypes, including biofilm formation, antibiotic resistance, and thermal tolerance, among others (154). However, despite this megaplasmid being recently acquired by *P.*

syringae Pla107 (155), it was difficult to construct a Pla107 derivative lacking the megaplasmid (113), which is potentially suggestive of rapid, partial adaptation to accommodate the costs of this replicon. Additionally, an experimental evolution study of populations of *Methylobacterium extorquens* AM1 identified the repeated loss of parts of a megaplasmid region accounting for up to 10% of the genome (156). While the deletion events resulted in more rapid growth under the growth conditions from which they were isolated, these mutants grew more slowly under alternate growth conditions. These results suggested that selection for the loss of environment-specific accessory genes, rather than genetic drift, dominated the genome reduction process (156).

Why exactly these fitness costs are observed is unclear, although several suggestions have been put forth. Streamlining theory suggests that the loss of the replicon could be favored as it reduces the amount of phosphorus tied up in DNA (157), although others have argued that there is little support for this hypothesis (158). Alternatively, loss of the replicon could be favored by reducing the energetic demands associated with DNA replication and/or gene expression (transcription and translation), particularly the expression of multiprotein ABC transport systems that are likely energetically expensive to synthesize and that are enriched on secondary replicons (113, 159, 160). Decreasing the number of transcripts of nonessential proteins could also free up ribosomes for the translation of core proteins, and/or decreasing the number of recently acquired genes whose gene products may be misfolded could also promote the loss of a secondary replicon (113). Finally, negative interactions between pathways encoded by the chromosome and secondary replicon could promote a loss of the secondary replicon, as could negative interactions between these replicons at the transcriptional level (113, 159). Likely, a combination of factors explains why secondary replicons confer fitness costs to the host and why their loss may be favored during growth in particular environments.

PUTATIVE ADVANTAGES OF MULTIPARTITE GENOMES

Several hypotheses have been put forth to describe why bacterial multipartite genomes have emerged and are maintained. In this section, the main putative advantages are discussed, and the data that support and contradict each hypothesis are described. A summary of these points is provided in Table 3.

Increased Genome Size

It has been suggested that multipartite genomes allow for further genome expansion once the chromosome has reached its maximal size (84). In support of this, it was noted that as of 2010, the mean total size of genomes lacking a chromid was 3.38 Mb (SD, 1.81 Mb), whereas the mean size of genomes with a chromid was 5.73 Mb (SD, 1.66 Mb) (11). In contrast, it was pointed out that some small genomes, like that of *Brucella melitensis*, are multipartite, whereas some large genomes have a single chromosome, such as the 9-Mb chromosome of *Myxococcus xanthus* (31). When 1,708 representative genomes from the NCBI genome database were examined, the mean and median total genome sizes of species containing a putative chromid/megaplasmid were ~3.67 and ~3.41 Mb, respectively, whereas the mean and median genome sizes for species with a megaplasmid and/or a putative chromid were ~5.72 and ~5.56 Mb, respectively (Fig. 6A). The difference in genome sizes between these two groups can be associated primarily with the secondary replicon, as there was little difference in the mean and median chromosome sizes (Fig. 6B). Hence, it is clear that multipartite genomes are, on average, larger than genomes lacking megaplasmids and chromids. That said, multipartite genomes are not a prerequisite for a large genome, and in fact, fewer than one-third of the genomes with a size of >6 Mb are multipartite (Fig. 6A), while none of the 26 largest genomes, and only 3 of the top 50, are multipartite. Thus, it seems unlikely that the multipartite genome organization evolved simply to allow increased gene accumulation, as the majority of large genomes are not multipartite. Additionally, causality has not been demonstrated; i.e., it has not been established whether genomes

TABLE 3 Summary of the four described hypotheses on the role and evolution of multipartite genomes[a]

Hypothesis	Main tenant	Support	Contradiction(s)
Increased genome size	Dividing the genome allows for a larger genome than if only a chromosome was present	Multipartite genomes are on avg larger than nonmultipartite genomes; difference in genome sizes is due to the size of the secondary replicons and not chromosomal differences	Some small genomes are multipartite, while some large genomes are not multipartite; only 3 of the largest 50 bacterial genomes are multipartite; unclear if being multipartite allows larger genomes or if genomes are larger because they are multipartite
Increased rate of bacterial growth	Dividing the genome allows a higher growth rate due to faster replication of the genome	Some of the fastest-growing species (e.g., *Vibrio*) have multipartite genomes; fast-growing rhizobia contain chromids, whereas slow-growing rhizobia do not	Many slower-growing species have a multipartite genome, and some fast-growing species (e.g., *Clostridium*) do not have a multipartite genome; no correlation between genome size and growth rate; chromosomes and chromids are not equally sized
Coordinated gene regulation	Localization of related genes on the same replicon facilitates their coordinated regulation	The replicon that the gene is on can influence gene dosage; individual replicons are often over- or underrepresented in genes up- or downregulated in different environments	Gene dosage effect is likely limited to fast-replicating species; unclear if coordinated gene regulation was a driving force of multipartite genome evolution or a by-product of the colocalization of related genes on 1 replicon
Adaptation to novel niches	The secondary replicons are specialized for colonization and fitness in new environments	Consistent with several features of secondary replicons, including genetic variability and evolutionary rates; different replicons can show environment-specific patterns of gene regulation; secondary replicons are often enriched in genes associated with environmental adaptation	Many organisms without multipartite genomes occupy the same niches as those with multipartite genomes and display equal levels of genetic variability

[a]See Putative Advantages of Multipartite Genomes for an expanded discussion of these points.

are multipartite to allow increased size or whether the increased size is a consequence of having chromids/megaplasmids.

Increased Rate of Bacterial Growth

A second consideration is that the multipartite genome organization may allow faster bacterial division by decreasing the time required to replicate the genome, as each replicon can replicate concurrently (55, 65). Indeed, some of the fastest-replicating species, such as those in the genera *Vibrio*, have a multipartite genome (31), and the "fast-growing" rhizobia contain chromids, whereas the "slow-growing" rhizobia do not (161). However, multipartite genomes can be found in many species with relatively long generation times (162), such as *R. sphaeroides* (31), and multipartite genomes are certainly not a requirement for fast growth. For example, nonmultipartite *Clostridium perfringens* strains can have generation times as low as 7 min (163). Moreover, an analysis of 214 genomes failed to detect a correlation between genome size and minimal generation time (164). If the emergence and maintenance of a multipartite genome are driven by selective pressure to reduce the time required to replicate the genome, coresident chromosomes and chromids should be equally sized. However, this is not observed (162). Chromids are always smaller than the chromosome (11), and although there is a large range of disparity in the sizes of chromids versus chromosomes, chromids are on average less than half the size of the coresident chromosome (Fig. 7). Overall, while the multipartite genome may impart an ability to replicate the genome more rapidly, the data on the whole do not support this as a driving force for the evolution of the multipartite genome. It may, however, help promote the maintenance of this genome organization in fast-replicating species once it has formed.

Coordinated Gene Regulation

A third hypothesis states that the division of genes between multiple replicons facilitates their coordinated regulation. This could be accomplished through the modulation of gene dosage. Initiation of the replication of the chromosome of *Vibrio* species

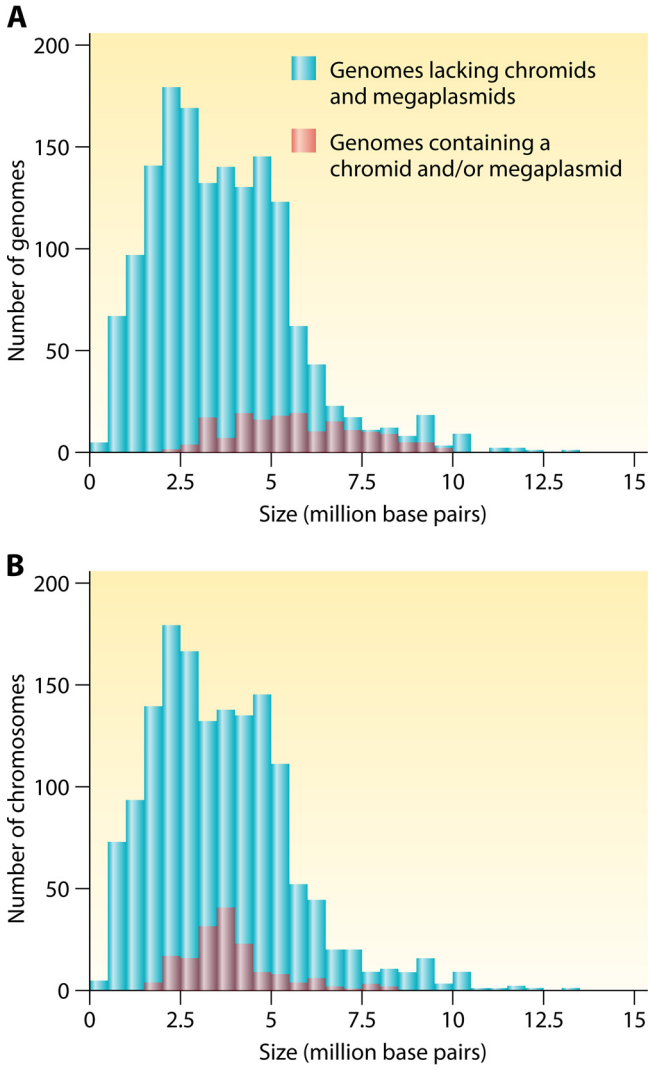


FIG 6 Size distribution of single chromosomes versus multipartite genomes. The histograms display the distributions of total genome sizes (A) and chromosomal sizes (B) for genomes lacking chromids and megaplasmids and for genomes containing a chromid and/or megaplasmid. The purple color occurs as a result of the overlap between the red and blue bars. Histograms are based on one representative genome of each of the 1,708 bacterial species with a completed genome available in the NCBI genome database (accessed 21 March 2016).

occurs prior to that of the chromid, resulting in a higher average gene dosage, and, thus, transcription, of chromosomal genes (16). However, this effect is expected to be limited to fast-growing bacteria (162), and indeed, evidence suggests that there is no gene dosage bias among the three replicons of the *S. meliloti* genome (165). Recently, a related suggestion was put forth, stating that the localization of genes to different replicons may facilitate their correct subcellular positioning (137). Although an exciting possibility, experimental support for this hypothesis is currently lacking.

Alternatively, the grouping of genes together on individual replicons may facilitate their coordinated regulation by transcription factors. This hypothesis has merit, considering that the transcription machinery is not equally distributed throughout the cell (166). Supporting this, an *in silico* regulon analysis of 41 *S. meliloti* transcription factors indicated a bias for transcription factors to regulate genes on the same replicon (147). Furthermore, multiple transcriptomic studies have observed that particular replicons are enriched in differentially regulated genes during niche adaptation. Comparison of the *V. cholerae* transcriptome under laboratory growth conditions to that under intes-

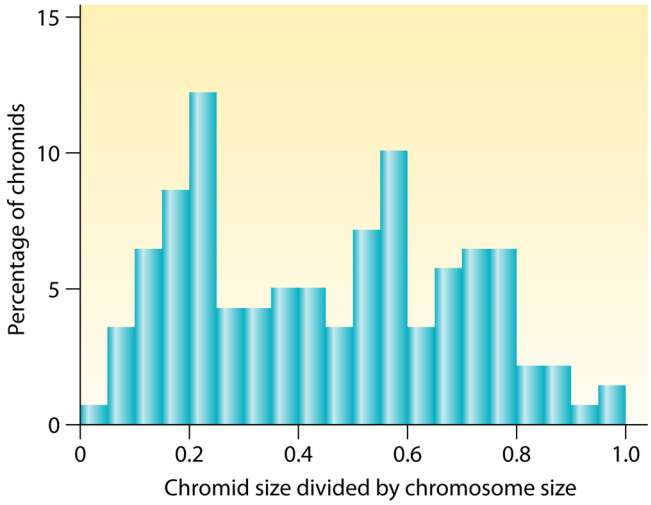


FIG 7 Comparison of chromid and chromosome sizes. All chromids from one representative genome of each of the 1,708 bacterial species with a completed genome available in the NCBI genome database (accessed 21 March 2016) were analyzed. The size of each chromid was divided by the size of the coresident chromosome, and the distribution of the results is shown.

tinal growth conditions illustrated that many more chromid genes are expressed in the intestine than under laboratory conditions (167). In a comparative transcriptomics study of *B. cenocepacia* J2315, the number of expressed genes during soil colonization was biased toward the larger chromid, whereas the number of genes expressed under *in vitro* cystic fibrosis conditions was biased toward the chromosome (168). Additionally, transcriptional differences between *B. cenocepacia* strain J2315 and a related strain were biased toward the smaller chromid of J2315 (168). The pRL8 replicon of *R. leguminosarum* is enriched for genes upregulated during growth in the pea rhizosphere (169), and similarly, secondary replicons of *Rhizobium phaseoli* are enriched in genes upregulated during rhizosphere colonization (170). During symbiosis with legumes, genes downregulated in *S. meliloti* are overrepresented on the chromosome, whereas upregulated genes are overrepresented on the megaplasmid (151, 171). These studies clearly demonstrate that replicon-specific patterns of gene expression can be observed. However, causation has not been demonstrated, and thus, it is unclear if the multipartite genome evolved to facilitate coordinated gene regulation or whether these transcriptional patterns are a by-product of functionally related genes being colocalized to different replicons.

Adaptation to Novel Niches

To account for many of the observations related to multipartite genomes, including replicon-specific regulation patterns, functional biases, and distinct evolutionary patterns, it was suggested previously that individual replicons within multipartite genomes contribute to adaptation to unique environments (22, 121, 152). It is our opinion that the primary advantage of the multipartite genome architecture is to mediate adaptation to novel niches. In this section, we present a generalized evolutionary model based on niche adaptation that attempts to explain the observations reported throughout this review. This model is summarized in Fig. 8 and is an extension of ideas proposed previously (22, 121, 152).

The main tenant of the proposed model is that secondary replicons act as specialized entities for adaptation to unique environments. This implicitly states that the primary chromosome is sufficient for growth and survival in nonspecialized soil or aquatic environments. Computational analyses suggested that the ancestor of the *Alphaproteobacteria* harbored ~3,300 genes, with lower and upper bounds of 3,000 and 5,000 proteins, respectively (172). The median bacterial chromosome size of 3.46 Mb is therefore likely similar to that of the ancestral alphaproteobacterial genome.

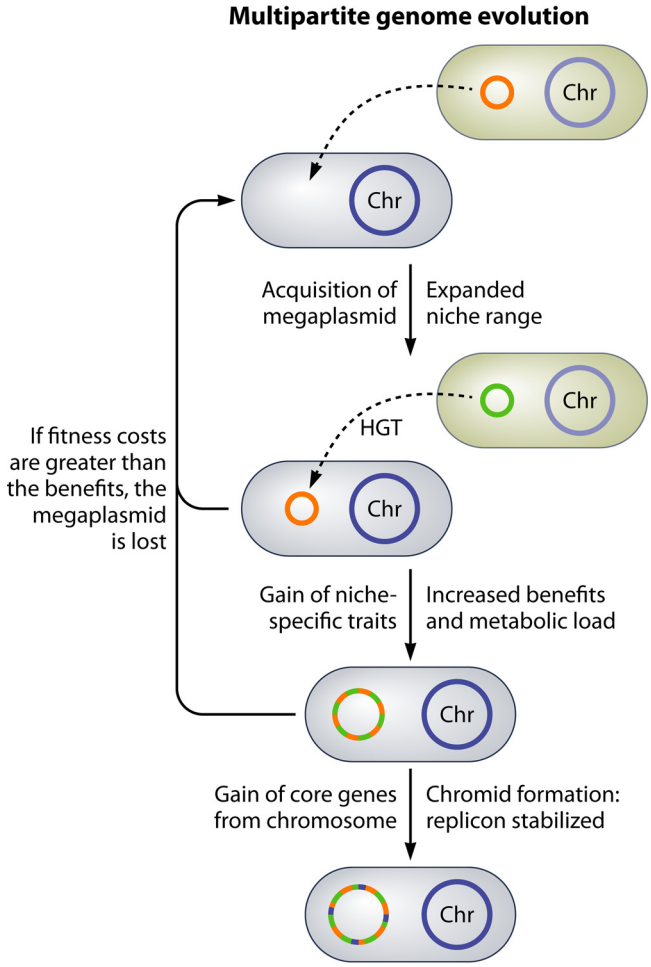


FIG 8 Described model of multipartite genome evolution. Shown is a schematic of the proposed model to explain the evolution and function of the bacterial multipartite genome. Multipartite genome evolution begins with the acquisition of a megaplasmid through horizontal gene transfer (HGT). It is hypothesized that the selective pressure for the maintenance of the megaplasmid is the ability to colonize a new niche. In this new environment, the megaplasmid size increases by HGT and the gain of new genes involved in adaptation to the newly inhabited niche. At any point during evolution, the megaplasmid may be lost if the costs start to outweigh the benefits or if the cell leaves the niche where the megaplasmid is beneficial. Coresidence of the megaplasmid with the chromosome facilitates interreplicon gene flow, which results in the transfer of core genes to the megaplasmid, resulting in the formation of a chromid and a replicon that is more integrated into the core biological networks of the cell.

Additionally, an *S. meliloti* strain with a genome reduced by 45% through the removal of the chromid and megaplasmid was capable of growing quite well in bulk soil mesocosms (152). Together, these observations are consistent with the chromosomal gene repertoire being sufficient for high fitness in nonspecialized environments.

Accepting that the plasmid hypothesis explains the formation of chromids, then all multipartite genomes must originate from the acquisition of a megaplasmid through horizontal gene transfer (HGT). Secondary replicons often carry the key determinants required for initial colonization of new environments, such as symbiotic and virulence factors, although they generally account for only a small portion of large secondary replicons (161, 173). The high level of gene variations of megaplasmids, as discussed above (27, 78, 98, 122–125, 127, 128), suggests that they are undergoing rapid gene loss and gene gain through HGT, as was observed for *S. meliloti* (121). Comparative genomics and metabolic modeling studies illustrated that most genes acquired through HGT are involved primarily in adaptation to different environments (174–177), with different genomic regions being responsible for different ecologies (178).

Other computational work illustrated that genome expansion within lineages of the *Alphaproteobacteria* and the order *Rhizobiales* was linked to an association with plants and the evolution of symbiosis, respectively (172, 179). Those genome expansions involved mainly the acquisition of transcriptional, transport, and metabolic functions (172, 179), which are the same functions commonly enriched on secondary replicons (Table 2). Hence, it is reasonable to suggest that the gain of secondary replicons first allows an association with a novel niche, after which the new replicon gains genes associated with environmental adaptation. In particular, we hypothesize that these environments often represent new niches formed due to the emergence of eukaryotic organisms, such as the rhizosphere and the animal gut. However, the association with eukaryotic organisms may reflect biases in the organisms chosen for whole-genome sequencing, and sequencing of more environmental isolates may reveal that secondary replicons may be generally associated with diverse niche colonization independent of eukaryotes.

As chromids are on average twice as large as megaplasmids (Fig. 2), the evolution of a megaplasmid must therefore involve significant gene accumulation. However, it has been argued that most HGT events, including those that are eventually fixed in a population, are initially deleterious (180, 181). As a result, most genes acquired through HGT are lost from the genome, as the costs outweigh the benefits (175, 180–183). The high variability of megaplasmids is consistent with most genes acquired by these replicons eventually being lost either because they provide no benefit or because the costs are too high. Only genes whose benefits outweigh the costs to the cell are eventually fixed in the population, and as described above, these genes are expected to be associated with adaptation to the new environment. This conclusion is supported by the transcriptional studies summarized above (151, 167–171), by an *in silico* metabolic modeling study that suggested that the metabolic abilities associated with the *S. meliloti* chromid are rhizosphere specialized (184), and by experimental work showing that the loss of the pATC58 megaplasmid decreases fitness in the plant rhizosphere (153).

As chromids evolve from megaplasmids, chromids could be considered a subclass of megaplasmids, and it might therefore be expected that megaplasmids would be more prevalent than chromids. However, chromids are found in approximately twice as many species as megaplasmids (Fig. 4). This may reflect two divergent fates of megaplasmids. The potentially high costs of megaplasmids (83, 113, 120, 152, 155, 159), as discussed above, may result in the loss of the megaplasmid in particular environments and may limit the ability of megaplasmids to be successfully maintained following horizontal transfer of the entire replicon. On the other hand, if the megaplasmid provides enough of a benefit to remain in the cell, rapid conversion into a chromid is likely to occur. The observation that megaplasmids are limited to more narrow taxonomic groups than are chromids (Fig. 4) supports that secondary replicons remain an evolutionarily stable component of the genome only if a conversion to a chromid occurs.

Genes recently acquired through HGT undergo rapid evolution (181). As described above, megaplasmids experience high evolutionary rates (22, 78, 98, 121, 130), which is possibly reflective of the rapid amelioration of the replicon to reduce the associated costs. This can include modification of genes and regulatory elements, such as through the amelioration of codon usage (185), and promoter modifications to better integrate the genes into existing transcriptional networks (186). At the same time, the gain of essential genes from the chromosome results in the formation of a chromid (41, 84). This interreplicon gene flow also contributes to the increased stability of the secondary replicon, in terms of reducing both the rate of gene loss and the loss of the entire replicon (11, 51, 84, 152), and contributes to the integration of the replicon into the cell's core metabolism (187). In this way, environmental adaptation can drive the emergence of a multipartite genome.

The primary observation opposing this hypothesis is that many species lack multipartite genomes, yet they can still show high genetic variability and colonization of multiple niches. Species of the genera *Sinorhizobium* and *Bradyrhizobium* are legume N_2-fixing endosymbionts that also colonize bulk soil and the plant rhizosphere. Both genera also contain large pangenomes, and thus, both genera have high genetic

variability (188). However, only species of the genus *Sinorhizobium* have a multipartite genome (18, 189). Moreover, the organisms of the genus *Prochlorococcus* have a genome size of just 2,000 genes but have an estimated pangenome size of >58,000 genes (190). Thus, even if the evolution of the multipartite genome is driven by niche adaptation, a multipartite genome is by no means a prerequisite for niche adaptation or genetic variability.

REMAINING QUESTIONS

Although many characteristics of multipartite genomes are known, there are several questions that remain unanswered. Here, five topics that require further study are outlined, and in most cases, potential answers are detailed.

Maintenance of the Multipartite Genome

Even if the evolution of the multipartite genome is driven by environmental adaptation, it remains unclear why the secondary replicon remains an independent unit and why it does not eventually become integrated into the chromosome, as appears to have occurred in a few cases (84). One possibility is that, at least in some species, the potential benefits of divided genomes, as summarized above, may help maintain the multipartite genome structure once it has formed. However, we hypothesize that the multipartite architecture is often an evolutionary relic limited by what came before. Megaplasmids are transferable (110–117). Hence, a gene on a megaplasmid may have higher fitness than a gene on a chromosome due to the increased frequency of HGT mediated through megaplasmid conjugation. We note that chromosomes can also carry mobile elements as well, but as the gain of a chromosomal mobile element can result in the disruption of important chromosomal genes (191, 192), plasmid-mediated HGT may be more efficient. Selection for increased HGT may be lost in larger megaplasmids and chromids due to their increased costs; however, integration into the chromosome may still be unfavorable due to the large size of such replicons. The chromosomal origin of replication and terminus region normally, but not always, split chromosomes into two roughly equal halves referred to as replichores (193–195). Genome rearrangements that significantly perturb this equal distribution appear to have a negative impact on fitness and can be selected against (13, 196–200), meaning that the integration of a 1.5-Mb chromid into a bacterial chromosome is likely to be unfavorable. Integration may be further selected against if the gene strand bias of the chromosome is not maintained (13, 17, 194, 201–203). Indeed, *S. meliloti* strains with all three replicons fused together display a fitness decrease (88), and it has been proposed that *V. cholerae* strains with both replicons fused together are less fit (89). Hence, the maintenance of the multipartite genome architecture may reflect selection for increased HGT early in its development and selective pressures against chromosome disruptions later during its maintenance.

Enrichment of Environmental Adaptation Genes on Secondary Replicons

If the niche adaptation model outlined above is correct, secondary replicon enlargement occurs as a result of the acquisition of niche-specific genes. However, it is unclear why these genes would be preferentially acquired by the secondary replicon and not equally acquired by the chromosome. It may be that secondary replicons more readily acquire new DNA (84) or because insertions into the chromosome are more likely to disrupt growth-promoting genes than are insertions into a megaplasmid, leading to greater selection against chromosomal insertions (191, 192). This is supported by the reduced purifying selection observed on secondary replicons (98) and the higher prevalence of transposable elements (Table 2). Additionally, and perhaps more importantly, megaplasmids can move through conjugation, and thus, genes that integrate into these replicons may have higher fitness, as they can more readily spread horizontally throughout the population. Moreover, the colocalization of related/interacting pathways on the same replicon results in their genomic linkage. If this replicon is

transferable, then both pathways will move together, and if this is beneficial to the recipient, evolution may select for the linkage of the pathways.

Fixation of Essential Gene Transfer Events

The key characteristic of a chromid is that it contains core biological functions. One way in which this occurs is through the transfer of genes from the chromosome to the secondary replicon (31, 41, 83–87). However, it is unclear why such a translocation event would become fixed in the population. Many secondary replicons within the *Alphaproteobacteria* belong to the *repABC* family and encode a partitioning system that helps ensure high stability and segregation of the replicon to both daughter cells (204, 205). However, segregation is sufficiently imperfect so that the replicon could be lost from the population within a few thousand generations (205). Thus, it may be that the transfer of essential genes results in the stabilization of the replicon that, combined with the loss of replicons without essential genes from the population through genetic drift, results in the fixation of essential genes on the chromid. This could also explain why chromids generally contain only a few essential genes; the first transfer of essential genes would stabilize the replicon, while additional transfers of essential genes would provide little additional advantage (51).

Multipartite Genome Topology

An exciting research direction that has so far remained largely unexplored is the study of the genome topology and DNA physical interactions in species with multipartite genomes through techniques such as chromosome conformation capture methods like Hi-C (206, 207) and multicolor fluorescence *in situ* hybridization (FISH) (208). The sole study examining three-dimensional (3D) genome topology in a multipartite genome was performed with *V. cholerae* (60). It would be interesting to use these techniques and various model systems to address general questions related to multipartite genome evolution. Potential research topics include examining whether there are interreplicon chromatin interactions and if such interactions are correlated with regions of increased interreplicon transcriptional interactions or gene flow. It would also be fascinating to examine whether the replicons are intermingled or if each replicon occupies a distinct and stable location in the cell, similar to how each chromosome occupies an unique nuclear territory in eukaryotic organisms (209), and if the removal of one or more replicons impacts the localization of the remaining replicons. In the case of *V. cholerae*, the data were consistent with the chromosome and chromid having very different organizations, with each replicon occupying distinct locations in the cell, and with there being direct physical interactions between the two replicons (60). However, as the data were analyzed with respect to one specific question, additional work is required to address general questions related to multipartite genome topology and the generalizability of the observations.

It is also worth noting that many secondary replicons encode nucleoid-associated proteins (NAPs) that can influence the topology of themselves as well as the chromosome, influence chromosomal transcription, and impact host fitness. This topic was recently reviewed by Shintani et al. (210), and we refer readers to that article for an in-depth discussion of this topic. Interestingly, they found that only a low percentage of plasmids carried NAPs, while over one-third of megaplasmids/chromids encoded NAPs (210); however, as chromids and megaplasmids were not differentiated, the relative frequencies of these two classes of replicons cannot be compared.

Loss of Conjugal Properties

It is also unclear why megaplasmids, but not chromids, appear to transfer via conjugation in nature (110–117) despite at least some replicons of both classes retaining conjugal properties in laboratory settings (11, 118, 119). This may simply reflect a high cost of acquisition. It has been argued that most genes acquired through HGT are maintained due to their low costs as opposed to their benefits (180), and thus, the costs associated with HGT (180, 211, 212) may mean that chromids are rapidly lost

if they are transferred to a new cell. Additionally, core "information" genes are less likely to be successfully transferred via HGT (213, 214), and such genes are found on chromids but not megaplasmids. Thus, the large size of chromids and the types of genes that they carry may lead to their rapid loss in the event that they are horizontally transferred.

CONCLUSIONS AND PERSPECTIVES

In this article, we review the available information related to bacterial multipartite genomes through a literature search and through a meta-analysis of complete bacterial genome sequences. The characteristics of the three main classes of large bacterial replicons (chromosomes, chromids, and megaplasmids) have been studied for a variety of species, and it has been found that regardless of which characteristic is examined, chromids and megaplasmids have a conserved set of features that set them apart from each other and from chromosomes. In future research, it will be important to experimentally validate the chromid designation of more replicons to ensure that the distribution of this replicon class is well understood, as designations based solely on data from informatics analyses or predictions can be misleading (35). It will also be valuable to completely remove chromids from numerous species by first moving just the essential functions to the chromosome, as was done for *S. meliloti* (152), in order to validate the overall biological role of these entities.

Nevertheless, currently available research on different taxonomic groups has allowed global comparisons of each replicon type and the elucidation of conserved characteristics. We look forward to seeing how such information can be applied in practical applications. The apparent role of secondary replicons in the colonization of specific niches implies that they serve as reservoirs for functions associated with adaptation to the corresponding environment. The mining of these replicons can therefore lead to the discovery of genes relevant to biotechnological applications, such as engineering improved plant bioinoculants or combating bacteria during pathogenic associations with humans or livestock.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/MMBR.00019-17.

SUPPLEMENTAL FILE 1, PDF file, 1.4 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.8 MB.

ACKNOWLEDGMENTS

This work was supported by the National Science and Engineering Council of Canada through grants to T.M.F. and an NSERC CSG-D award to G.C.D.

T.M.F. thanks Trevor Charles, Patrick Chain, Ivan Oresnik, Shawn MacLellan, Allyson MacLean, Branka Milunovic, Maryam Zamani, Richard Morton, and Brian Golding for contributions to the pSymB project over the years.

REFERENCES

1. Cairns J. 1963. The bacterial chromosome and its manner of replication as seen by autoradiography. J Mol Biol 6:208–213. https://doi.org/10.1016/S0022-2836(63)80070-4.
2. Wake RG. 1973. Circularity of the *Bacillus subtilis* chromosome and further studies on its bidirectional replication. J Mol Biol 77:569–575. https://doi.org/10.1016/0022-2836(73)90223-4.
3. Bode HR, Morowitz HJ. 1967. Size and structure of the *Mycoplasma hominis* H39 chromosome. J Mol Biol 23:191–199. https://doi.org/10.1016/S0022-2836(67)80026-3.
4. Hayakawa T, Tanaka T, Sakaguchi K, Ōtake N, Yonehara H. 1979. A linear plasmid-like DNA in *Streptomyces* sp. producing lankacidin group antibiotics. J Gen Appl Microbiol 25:255–260. https://doi.org/10.2323/jgam.25.255.
5. Baril C, Richaud C, Baranton G, Saint Girons I. 1989. Linear chromosome of *Borrelia burgdorferi*. Res Microbiol 140:507–516. https://doi.org/10.1016/0923-2508(89)90083-1.
6. Ferdows MS, Barbour AG. 1989. Megabase-sized linear DNA in the bacterium *Borrelia burgdorferi*, the Lyme disease agent. Proc Natl Acad Sci U S A 86:5969–5973. https://doi.org/10.1073/pnas.86.15.5969.
7. Rosenberg C, Boistard P, Dénarié J, Casse-Delbart F. 1981. Genes controlling early and late functions in symbiosis are located on a megaplasmid in *Rhizobium meliloti*. Mol Gen Genet 184:326–333.
8. Bonhoeffer F, Messer W. 1969. Replication of the bacterial chromosome. Annu Rev Genet 3:233–246. https://doi.org/10.1146/annurev.ge.03.120169.001313.
9. Suwanto A, Kaplan S. 1989. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. J Bacteriol 171:5850–5859. https://doi.org/10.1128/jb.171.11.5850-5859.1989.
10. Anda M, Ohtsubo Y, Okubo T, Sugawara M, Nagata Y, Tsuda M, Minamisawa K, Mitsui H. 2015. Bacterial clade with the ribosomal RNA operon

on a small plasmid rather than the chromosome. Proc Natl Acad Sci U S A 112:14343–14347. https://doi.org/10.1073/pnas.1514326112.
11. Harrison PW, Lower RPJ, Kim NKD, Young JPW. 2010. Introducing the bacterial "chromid": not a chromosome, not a plasmid. Trends Microbiol 18:141–148. https://doi.org/10.1016/j.tim.2009.12.010.
12. Junier I. 2014. Conserved patterns in bacterial genomes: a conundrum physically tailored by evolutionary tinkering. Comp Biol Chem 53:125–133. https://doi.org/10.1016/j.compbiolchem.2014.08.017.
13. Rocha EPC. 2008. The organization of the bacterial genome. Annu Rev Genet 42:211–233. https://doi.org/10.1146/annurev.genet.42.110807.091653.
14. Lawrence JG. 2003. Genome organization: selection, selfishness, and serendipity. Annu Rev Microbiol 57:419–440. https://doi.org/10.1146/annurev.micro.57.030502.090816.
15. Bryant JA, Sellars LE, Busby SJW, Lee DJ. 2015. Chromosome position effects on gene expression in *Escherichia coli* K-12. Nucleic Acids Res 42:11383–11392. https://doi.org/10.1093/nar/gku828.
16. Dryselius R, Izutsu K, Honda T, Iida T. 2008. Differential replication dynamics for large and small *Vibrio* chromosomes affect gene dosage, expression and location. BMC Genomics 9:559. https://doi.org/10.1186/1471-2164-9-559.
17. Rocha EPC. 2004. The replication-related organization of bacterial genomes. Microbiology 150:1609–1627. https://doi.org/10.1099/mic.0.26974-0.
18. Galibert F, Finan TM, Long SR, Pühler A, Abola AP, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dréano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernández-Lucas I, Hong A, Huizar L, Hyman RW, Jones T, Kahn D, Kahn ML, Kalman S, Keating DH, Kiss E, Komp C, Lelaure V, Masuy D, Palm C, Peck MC, Pohl T, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thébault P, Vandenbol M, et al. 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 293:668–672. https://doi.org/10.1126/science.1060966.
19. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S. 2001. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. Science 294:2323–2328. https://doi.org/10.1126/science.1066803.
20. DelVecchio VG, Kapatral V, Redkar RJ, Patra G, Mujer C, Los T, Ivanova N, Anderson I, Bhattacharyya A, Lykidis A, Reznik G, Jablonski L, Larsen N, D'Souza M, Bernal A, Mazur M, Goltsman E, Selkov E, Elzer PH, Hagius S, O'Callaghan D, Letesson J-J, Haselkorn R, Kyrpides N, Overbeek R. 2002. The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. Proc Natl Acad Sci U S A 99:443–448. https://doi.org/10.1073/pnas.221575398.
21. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleishmann RD, Nierman WC, White O, Salzberg SL, Smith HO, Colwell RR, Mekalanos JJ, Venter JC, Fraser CM. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483. https://doi.org/10.1038/35020000.
22. Chain PSG, Denef VJ, Konstantinidis KT, Vergez LM, Agulló L, Reyes VL, Hauser L, Córdova M, Gómez L, González M, Land M, Lao V, Larimer F, LiPuma JJ, Mahenthiralingam E, Malfatti SA, Marx CJ, Parnell JJ, Ramette A, Richardson P, Seeger M, Smith D, Spilker T, Sul WJ, Tsoi TV, Ulrich LE, Zhulin IB, Tiedje JM. 2006. *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proc Natl Acad Sci U S A 103:15280–15287. https://doi.org/10.1073/pnas.0606924103.
23. Moreno E. 1998. Genome evolution within the alpha *Proteobacteria*: why do some bacteria not possess plasmids and others exhibit more than one different chromosome? FEMS Microbiol Rev 22:255–275. https://doi.org/10.1111/j.1574-6976.1998.tb00370.x.
24. Prozorov AA. 2008. Additional chromosomes in bacteria: properties and origin. Mikrobiologiia 77:437–447.
25. Schwartz E (ed). 2009. Microbial megaplasmids. Springer, Berlin, Germany.
26. Choudhary M, Cho H, Bavishi A, Trahan C, Myagmarjav B-E. 2012. Evolution of multipartite genomes in prokaryotes, p 301–323. *In* Pontarotti P (ed), Evolutionary biology: mechanisms and trends. Springer, Berlin, Germany.
27. Van Houdt R, Mergeay M. 2012. Plasmids as secondary chromosomes, p 1–4. *In* Wells RD, Bond JS, Klinman J, Masters BSS, Bell E (ed), Molecular life sciences: an encyclopedic reference. Springer, Berlin, Germany.
28. Val M-E, Soler-Bistué A, Bland MJ, Mazel D. 2014. Management of multipartite genomes: the *Vibrio cholerae* model. Curr Opin Microbiol 22:120–126. https://doi.org/10.1016/j.mib.2014.10.003.
29. Ramachandran R, Jha J, Paulsson J, Chattoraj D. 2017. Random versus cell cycle-regulated replication initiation in bacteria: insights from studying *Vibrio cholerae* chromosome 2. Microbiol Mol Biol Rev 81:e00033-16. https://doi.org/10.1128/MMBR.00033-16.
30. Ramachandran R, Jha J, Chattoraj DK. 2015. Chromosome segregation in *Vibrio cholerae*. J Mol Microbiol Biotechnol 24:360–370. https://doi.org/10.1159/000368853.
31. Egan ES, Fogel MA, Waldor MK. 2005. Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol 56:1129–1138. https://doi.org/10.1111/j.1365-2958.2005.04622.x.
32. Barry ER, Bell SD. 2006. DNA replication in the archaea. Microbiol Mol Biol Rev 70:876–887. https://doi.org/10.1128/MMBR.00029-06.
33. Wu Z, Liu J, Yang H, Xiang H. 2014. DNA replication origins in archaea. Front Microbiol 5:179. https://doi.org/10.3389/fmicb.2014.00179.
34. Oresnik IJ, Liu SL, Yost CK, Hynes MF. 2000. Megaplasmid pRme2011a of *Sinorhizobium meliloti* is not required for viability. J Bacteriol 182:3582–3586. https://doi.org/10.1128/JB.182.12.3582-3586.2000.
35. Agnoli K, Schwager S, Uehlinger S, Vergunst A, Viteri DF, Nguyen DT, Sokol PA, Carlier A, Eberl L. 2012. Exposing the third chromosome of *Burkholderia cepacia* complex strains as a virulence plasmid. Mol Microbiol 83:362–378. https://doi.org/10.1111/j.1365-2958.2011.07937.x.
36. Orlova N, Gerding M, Ivashkiv O, Olinares PDB, Chait BT, Waldor MK, Jeruzalmi D. 2017. The replication initiator of the cholera pathogen's second chromosome shows structural similarity to plasmid initiators. Nucleic Acids Res 45:3724–3737. https://doi.org/10.1093/nar/gkw1288.
37. Venkova-Canova T, Chattoraj DK. 2011. Transition from a plasmid to a chromosomal mode of replication entails additional regulators. Proc Natl Acad Sci U S A 108:6199–6204. https://doi.org/10.1073/pnas.1013244108.
38. Dziewit L, Czarnecki J, Wibberg D, Radlinska M, Mrozek P, Szymczak M, Schlüter A, Pühler A, Bartosik D. 2014. Architecture and functions of a multipartite genome of the methylotrophic bacterium *Paracoccus aminophilus* JCM 7686, containing primary and secondary chromids. BMC Genomics 15:124. https://doi.org/10.1186/1471-2164-15-124.
39. Finan TM, Weidner S, Wong K, Buhrmester J, Chain P, Vorhölter FJ, Hernández-Lucas I, Becker A, Cowie A, Gouzy J, Golding B, Puhler A. 2001. The complete sequence of the 1,683-kb pSymB megaplasmid from the N_2-fixing endosymbiont *Sinorhizobium meliloti*. Proc Natl Acad Sci U S A 98:9889–9894. https://doi.org/10.1073/pnas.161294698.
40. Cheng J, Sibley CD, Zaheer R, Finan TM. 2007. A *Sinorhizobium meliloti minE* mutant has an altered morphology and exhibits defects in legume symbiosis. Microbiology 153:375–387. https://doi.org/10.1099/mic.0.2006/001362-0.
41. diCenzo G, Milunovic B, Cheng J, Finan TM. 2013. The $tRNA^{arg}$ gene and *engA* are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. J Bacteriol 195:202–212. https://doi.org/10.1128/JB.01758-12.
42. Hayes F. 2003. Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science 301:1496–1499. https://doi.org/10.1126/science.1088157.
43. Van Melderen L, Saavedra De Bast M. 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet 5:e1000437. https://doi.org/10.1371/journal.pgen.1000437.
44. Gerdes K, Christensen SK, Løbner-Olesen A. 2005. Prokaryotic toxin-antitoxin stress response loci. Nat Rev Microbiol 3:371–382. https://doi.org/10.1038/nrmicro1147.
45. MacLellan SR, Smallbone LA, Sibley CD, Finan TM. 2005. The expression of a novel antisense gene mediates incompatibility within the large *repABC* family of α-proteobacterial plasmids. Mol Microbiol 55:611–623. https://doi.org/10.1111/j.1365-2958.2004.04412.x.
46. Milunovic B, diCenzo GC, Morton RA, Finan TM. 2014. Cell growth inhibition upon deletion of four toxin-antitoxin loci from the megaplasmids of *Sinorhizobium meliloti*. J Bacteriol 196:811–824. https://doi.org/10.1128/JB.01104-13.

47. O'Leary NA, Wright MW, Brister JR, 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, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
48. Suzuki K, Iwata K, Yoshida K. 2001. Genome analysis of *Agrobacterium tumefaciens*: construction of physical maps for linear and circular chromosomal DNAs, determination of copy number ratio and mapping of chromosomal virulence genes. DNA Res 8:141–152. https://doi.org/10.1093/dnares/8.4.141.
49. Sibley CD, MacLellan SR, Finan T. 2006. The *Sinorhizobium meliloti* chromosomal origin of replication. Microbiology 152:443–455. https://doi.org/10.1099/mic.0.28455-0.
50. Döhlemann J, Wagner M, Happel C, Carrillo M, Sobetzko P, Erb TJ, Thanbichler M, Becker A. 2017. A family of single copy *repABC*-type shuttle vectors stably maintained in the alpha-proteobacterium *Sinorhizobium meliloti*. ACS Synth Biol 6:968–984. https://doi.org/10.1021/acssynbio.6b00320.
51. Du W-L, Dubarry N, Passot FM, Kamgoué A, Murray H, Lane D, Pasta F. 2016. Orderly replication and segregation of the four replicons of *Burkholderia cenocepacia* J2315. PLoS Genet 12:e1006172. https://doi.org/10.1371/journal.pgen.1006172.
52. Li H, Angelov A, Pham VTT, Leis B, Liebl W. 2015. Characterization of chromosomal and megaplasmid partitioning loci in *Thermus thermophilus* HB27. BMC Genomics 16:317. https://doi.org/10.1186/s12864-015-1523-3.
53. Hartmann EM, Badalamenti JP, Krajmalnik-Brown R, Halden RU. 2012. Quantitative PCR for tracking the megaplasmid-borne biodegradation potential of a model sphingomonad. Appl Environ Microbiol 78:4493–4496. https://doi.org/10.1128/AEM.00715-12.
54. Srivastava P, Chattoraj DK. 2007. Selective chromosome amplification in *Vibrio cholerae*. Mol Microbiol 66:1016–1028. https://doi.org/10.1111/j.1365-2958.2007.05973.x.
55. Rasmussen T, Jensen RB, Skovgaard O. 2007. The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. EMBO J 26:3124–3131. https://doi.org/10.1038/sj.emboj.7601747.
56. De Nisco NJ, Abo RP, Wu CM, Penterman J, Walker GC. 2014. Global analysis of cell cycle gene expression of the legume symbiont *Sinorhizobium meliloti*. Proc Natl Acad Sci U S A 111:3217–3224. https://doi.org/10.1073/pnas.1400421111.
57. Duigou S, Knudsen KG, Skovgaard O, Egan ES, Lobner-Olesen A, Waldor MK. 2006. Independent control of replication initiation of the two *Vibrio cholerae* chromosomes by DnaA and RctB. J Bacteriol 188:6419–6424. https://doi.org/10.1128/JB.00565-06.
58. Egan ES, Waldor MK. 2003. Distinct replication requirements for the two *Vibrio cholerae* chromosomes. Cell 114:521–530. https://doi.org/10.1016/S0092-8674(03)00611-1.
59. Baek JH, Chattoraj DK. 2014. Chromosome I controls chromosome II replication in *Vibrio cholerae*. PLoS Genet 10:e1004184. https://doi.org/10.1371/journal.pgen.1004184.
60. Val M-E, Marbouty M, de Lemos Martins F, Kennedy SP, Kemble H, Bland MJ, Possoz C, Koszul R, Skovgaard O, Mazel D. 2016. A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. Sci Adv 2:e1501914. https://doi.org/10.1126/sciadv.1501914.
61. Srivastava P, Fekete RA, Chattoraj DK. 2006. Segregation of the replication terminus of the two *Vibrio cholerae* chromosomes. J Bacteriol 188:1060–1070. https://doi.org/10.1128/JB.188.3.1060-1070.2006.
62. Demarre G, Galli E, Muresan L, Paly E, David A, Possoz C, Barre F-X. 2014. Differential management of the replication terminus regions of the two *Vibrio cholerae* chromosomes during cell division. PLoS Genet 10:e1004557. https://doi.org/10.1371/journal.pgen.1004557.
63. Galli E, Poidevin M, Le Bars R, Desfontaines J-M, Muresan L, Paly E, Yamaichi Y, Barre F-X. 2016. Cell division licensing in the multichromosomal *Vibrio cholerae* bacterium. Nat Microbiol 1:16094. https://doi.org/10.1038/nmicrobiol.2016.94.
64. Kadoya R, Chattoraj DK. 2012. Insensitivity of chromosome I and the cell cycle to blockage of replication and segregation of *Vibrio cholerae* chromosome II. mBio 3:e00067-12. https://doi.org/10.1128/mBio.00067-12.
65. Frage B, Döhlemann J, Robledo M, Lucena D, Sobetzko P, Graumann PL, Becker A. 2016. Spatiotemporal choreography of chromosome and megaplasmids in the *Sinorhizobium meliloti* cell cycle. Mol Microbiol 100:808–823. https://doi.org/10.1111/mmi.13351.
66. Dubarry N, Pasta F, Lane D. 2006. ParABS systems of the four replicons of *Burkholderia cenocepacia*: new chromosome centromeres confer partition specificity. J Bacteriol 188:1489–1496. https://doi.org/10.1128/JB.188.4.1489-1496.2006.
67. Passot FM, Calderon V, Fichant G, Lane D, Pasta F. 2012. Centromere binding and evolution of chromosomal partition systems in the *Burkholderiales*. J Bacteriol 194:3426–3436. https://doi.org/10.1128/JB.00041-12.
68. Żebracki K, Koper P, Marczak M, Skorupska A, Mazur A. 2015. Plasmid-encoded RepA proteins specifically autorepress individual *repABC* operons in the multipartite *Rhizobium leguminosarum* bv. *trifolii* genome. PLoS One 10:e0131907. https://doi.org/10.1371/journal.pone.0131907.
69. Fricke WF, Kusian B, Bowien B. 2009. The genome organization of *Ralstonia eutropha* strain H16 and related species of the *Burkholderiaceae*. J Mol Microbiol Biotechnol 16:124–135. https://doi.org/10.1159/000142899.
70. Jumas-Bilak E, Michaux-Charachon S, Bourg G, O'Callaghan D, Ramuz M. 1998. Differences in chromosome number and genome rearrangements in the genus *Brucella*. Mol Microbiol 27:99–106. https://doi.org/10.1046/j.1365-2958.1998.00661.x.
71. Choudhary M, Mackenzie C, Nereng K, Sodergren E, Weinstock GM, Kaplan S. 1997. Low-resolution sequencing of *Rhodobacter sphaeroides* 2.A.1T: chromosome II is a true chromosome. Microbiology 143:3085–3099. https://doi.org/10.1099/00221287-143-10-3085.
72. Liang X, Baek C-H, Katzen F. 2013. *Escherichia coli* with two linear chromosomes. ACS Synth Biol 2:734–740. https://doi.org/10.1021/sb400079u.
73. Itaya M, Tanaka T. 1997. Experimental surgery to create subgenomes of *Bacillus subtilis* 168. Proc Natl Acad Sci U S A 94:5378–5382. https://doi.org/10.1073/pnas.94.10.5378.
74. Choudhary M, Fu Y-X, Mackenzie C, Kaplan S. 2004. DNA sequence duplication in *Rhodobacter sphaeroides* 2.4.1: evidence of an ancient partnership between chromosomes I and II. J Bacteriol 186:2019–2027. https://doi.org/10.1128/JB.186.7.2019-2027.2004.
75. Choudhary M, Mackenzie C, Nereng KS, Sodergren E, Weinstock GM, Kaplan S. 1994. Multiple chromosomes in bacteria: structure and function of chromosome II of *Rhodobacter sphaeroides* 2.4.1T. J Bacteriol 176:7694–7702. https://doi.org/10.1128/jb.176.24.7694-7702.1994.
76. Mackenzie C, Choudhary M, Larimer FW, Predki PF, Stilwagen S, Armitage JP, Barber RD, Donohue TJ, Hosler JP, Newman JE, Shapleigh JP, Sockett RE, Zeilstra-Ryalls J, Kaplan S. 2001. The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1. Photosynth Res 70:19–41. https://doi.org/10.1023/A:1013831823701.
77. Kontur WS, Schackwitz WS, Ivanova N, Martin J, Labutti K, Deshpande S, Tice HN, Pennacchio C, Sodergren E, Weinstock GM, Noguera DR, Donohue TJ. 2012. Revised sequence and annotation of the Rhodobacter sphaeroides 2.4.1 genome. J Bacteriol 194:7016–7017. https://doi.org/10.1128/JB.01214-12.
78. Choudhary M, Zanhua X, Fu YX, Kaplan S. 2007. Genome analyses of three strains of *Rhodobacter sphaeroides*: evidence of rapid evolution of chromosome II. J Bacteriol 189:1914–1921. https://doi.org/10.1128/JB.01498-06.
79. Nereng KS, Kaplan S. 1999. Genomic complexity among strains of the facultative photoheterotrophic bacterium *Rhodobacter sphaeroides*. J Bacteriol 181:1684–1688.
80. Castillo-Ramírez S, Vázquez-Castellanos JF, González V, Cevallos MA. 2009. Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in *Alphaproteobacteria*: the *repABC* operon. BMC Genomics 10:536. https://doi.org/10.1186/1471-2164-10-536.
81. Wong K, Finan T, Golding B. 2002. Dinucleotide compositional analysis of *Sinorhizobium meliloti* using the genome signature: distinguishing chromosomes and plasmids. Funct Integr Genomics 2:274–281. https://doi.org/10.1007/s10142-002-0068-0.
82. Carbone A, Zinovyev A, Képès F. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19:2005–2015. https://doi.org/10.1093/bioinformatics/btg272.
83. diCenzo GC, Zamani M, Milunovic B, Finan TM. 2016. Genomic re-

sources for identification of the minimal N_2-fixing symbiotic genome. Environ Microbiol 18:2534–2547. https://doi.org/10.1111/1462-2920.13221.
84. Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ, Banta L, Dickerman AW, Paulsen I, Otten L, Suen G, Welch R, Almeida NF, Arnold F, Burton OT, Du Z, Ewing A, Godsy E, Heisel S, Houmiel KL, Jhaveri J, Lu J, Miller NM, Norton S, Chen Q, Phoolcharoen W, Ohlin V, Ondrusek D, Pride N, Stricklin SL, Sun J, Wheeler C, Wilson L, Zhu H, Wood DW. 2009. Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. J Bacteriol 191:2501–2511. https://doi.org/10.1128/JB.01779-08.
85. Wong K, Golding GB. 2003. A phylogenetic analysis of the pSymB replicon from the *Sinorhizobium meliloti* genome reveals a complex evolutionary history. Can J Microbiol 49:269–280. https://doi.org/10.1139/w03-037.
86. Sun S, Guo H, Xu J. 2006. Multiple gene genealogical analyses reveal both common and distinct population genetic patterns among replicons in the nitrogen-fixing bacterium *Sinorhizobium meliloti*. Microbiology 152:3245–3259. https://doi.org/10.1099/mic.0.29170-0.
87. Zheng J, Guan Z, Cao S, Peng D, Ruan L, Jiang D, Sun M. 2015. Plasmids are vectors for redundant chromosomal genes in the *Bacillus cereus* group. BMC Genomics 16:6. https://doi.org/10.1186/s12864-014-1206-5.
88. Guo X, Flores M, Mavingui P, Fuentes SI, Hernández G, Dávila G, Palacios R. 2003. Natural genomic design in *Sinorhizobium meliloti*: novel genomic architectures. Genome Res 13:1810–1817.
89. Val M-E, Kennedy SP, Soler-Bistué AJ, Barbe V, Bouchier C, Ducos-Galand M, Skovgaard O, Mazel D. 2014. Fuse or die: how to survive the loss of Dam in *Vibrio cholerae*. Mol Microbiol 91:665–678. https://doi.org/10.1111/mmi.12483.
90. Ng WV, Ciufo SA, Smith TM, Bumgarner RE, Baskin D, Faust J, Hall B, Loretz C, Seto J, Slagel J, Hood L, DasSarma S. 1998. Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? Genome Res 8:1131–1141. https://doi.org/10.1101/gr.8.11.1131.
91. diCenzo GC, Finan TM. 2015. Genetic redundancy is prevalent within the 6.7 Mb Sinorhizobium meliloti genome. Mol Genet Genomics 290:1345–1356. https://doi.org/10.1007/s00438-015-0998-6.
92. Bavishi A, Lin L, Schroeder K, Peters A, Cho H, Choudhary M. 2010. The prevalence of gene duplications and their ancient origin in *Rhodobacter sphaeroides* 2.4.1. BMC Microbiol 10:331. https://doi.org/10.1186/1471-2180-10-331.
93. Maida I, Fondi M, Orlandini V, Emiliani G, Papaleo MC, Perrin E, Fani R. 2014. Origin, duplication and reshuffling of plasmid genes: insights from *Burkholderia vietnamiensis* G4 genome. Genomics 103:229–238. https://doi.org/10.1016/j.ygeno.2014.02.004.
94. Landeta C, Dávalos A, Cevallos MÁ Geiger O, Brom S, Romero D. 2011. Plasmids with a chromosome-like role in rhizobia. J Bacteriol 193:1317–1326. https://doi.org/10.1128/JB.01184-10.
95. Popescu A-A, Huber KT, Paradis E. 2012. ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics 28:1536–1537. https://doi.org/10.1093/bioinformatics/bts184.
96. Österman J, Marsh J, Laine PK, Zeng Z, Alatalo E, Sullivan JT, Young JPW, Thomas-Oates J, Paulin L, Lindström K. 2014. Genome sequencing of two *Neorhizobium galegae* strains reveals a *noeT* gene responsible for the unusual acetylation of the nodulation factors. BMC Genomics 15:500. https://doi.org/10.1186/1471-2164-15-500.
97. Quax TEF, Claassens NJ, Söll D, van der Oost J. 2015. Codon bias as a means to fine-tune gene expression. Mol Cell 59:149–161. https://doi.org/10.1016/j.molcel.2015.05.035.
98. Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput Biol 6:e1000732. https://doi.org/10.1371/journal.pcbi.1000732.
99. Wan XF, Zhou J, Xu D. 2007. CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes. Int J Gen Syst 35:109–125. https://doi.org/10.1080/03081070500502967.
100. Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. EMBO Rep 6:1208–1213. https://doi.org/10.1038/sj.embor.7400538.
101. Mann S, Chen Y-PP. 2010. Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics 95:7–15. https://doi.org/10.1016/j.ygeno.2009.09.002.
102. Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet 11:e1004941. https://doi.org/10.1371/journal.pgen.1004941.
103. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. PLoS Comput Biol 11:e1004095. https://doi.org/10.1371/journal.pcbi.1004095.
104. Rocha EPC, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet 18:291–294. https://doi.org/10.1016/S0168-9525(02)02690-2.
105. van Passel MWJ, Bart A, Luyf ACM, van Kampen AHC, van der Ende A. 2006. Compositional discordance between prokaryotic plasmids and host chromosomes. BMC Genomics 7:26. https://doi.org/10.1186/1471-2164-7-26.
106. Nishida H. 2012. Comparative analyses of base compositions, DNA sizes, and dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. Int J Evol Biol 2012:342482. https://doi.org/10.1155/2012/342482.
107. Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115. https://doi.org/10.1371/journal.pgen.1001115.
108. Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6:e1001107. https://doi.org/10.1371/journal.pgen.1001107.
109. Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290. https://doi.org/10.1016/S0168-9525(00)89076-9.
110. Pérez-Mendoza D, Sepúlveda E, Pando V, Muñoz S, Nogales J, Olivares J, Soto MJ, Herrera-Cervera JA, Romero D, Brom S, Sanjuán J. 2005. Identification of the *rctA* gene, which is required for repression of conjugative transfer of rhizobial symbiotic megaplasmids. J Bacteriol 187:7341–7350. https://doi.org/10.1128/JB.187.21.7341-7350.2005.
111. He X, Chang W, Pierce DL, Seib LO, Wagner J, Fuqua C. 2003. Quorum sensing in *Rhizobium* sp. strain NGR234 regulates conjugal transfer (*tra*) gene expression and influences growth rate. J Bacteriol 185:809–822. https://doi.org/10.1128/JB.185.3.809-822.2003.
112. Yang JC, Lessard PA, Sengupta N, Windsor SD, O'Brien XM, Bramucci M, Tomb J-F, Nagarajan V, Sinskey AJ. 2007. TraA is required for megaplasmid conjugation in *Rhodococcus erythropolis* AN12. Plasmid 57:55–70. https://doi.org/10.1016/j.plasmid.2006.08.002.
113. Romanchuk A, Jones CD, Karkare K, Moore A, Smith BA, Jones C, Dougherty K, Baltrus DA. 2014. Bigger is not always better: transmission and fitness burden of ~1MB *Pseudomonas syringae* megaplasmid pMPPla107. Plasmid 73:16–25. https://doi.org/10.1016/j.plasmid.2014.04.002.
114. Brom S, García-de los Santos A, Cervantes L, Palacios R, Romero D. 2000. In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons. Plasmid 44:34–43. https://doi.org/10.1006/plas.2000.1469.
115. Herrera-Cervera JA, Caballero-Mellado J, Laguerre G, Tichy H-V, Requena N, Amarger N, Martínez-Romero E, Olivares J, Sanjuán J. 1999. At least five rhizobial species nodulate *Phaseolus vulgaris* in a Spanish soil. FEMS Microbiol Ecol 30:87–97. https://doi.org/10.1111/j.1574-6941.1999.tb00638.x.
116. Brom S, Girard L, García-de los Santos A, Sanjuan-Pinilla JM, Olivares J, Sanjuán J. 2002. Conservation of plasmid-encoded traits among bean-nodulating *Rhizobium* species. Appl Environ Microbiol 68:2555–2561. https://doi.org/10.1128/AEM.68.5.2555-2561.2002.
117. Young JPW, Wexler M. 1988. Sym plasmid and chromosomal genotypes are correlated in field populations of *Rhizobium leguminosarum*. Microbiology 134:2731–2739. https://doi.org/10.1099/00221287-134-10-2731.
118. Blanca-Ordóñez H, Oliva-García JJ, Pérez-Mendoza D, Soto MJ, Olivares J, Sanjuán J, Nogales J. 2010. pSymA-dependent mobilization of the *Sinorhizobium meliloti* pSymB megaplasmid. J Bacteriol 192:6309–6312. https://doi.org/10.1128/JB.00549-10.
119. Banfalvi Z, Kondorosi E, Kondorosi A. 1985. *Rhizobium meliloti* carries two megaplasmids. Plasmid 13:129–138. https://doi.org/10.1016/0147-619X(85)90065-4.
120. Finan TM, Kunkel B, De Vos GF, Signer ER. 1986. Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. J Bacteriol 167:66–72. https://doi.org/10.1128/jb.167.1.66-72.1986.
121. Galardini M, Pini F, Bazzicalupo M, Biondi EG, Mengoni A. 2013. Replicon-dependent bacterial genome evolution: the case of *Sinorhizobium meliloti*. Genome Biol Evol 5:542–558. https://doi.org/10.1093/gbe/evt027.

122. Guo H, Sun S, Eardly B, Finan T, Xu J. 2009. Genome variation in the symbiotic nitrogen-fixing bacterium *Sinorhizobium meliloti*. Genome 52:862–875. https://doi.org/10.1139/G09-060.
123. Epstein B, Branca A, Mudge J, Bharti AK, Briskine R, Farmer AD, Sugawara M, Young ND, Sadowsky MJ, Tiffin P. 2012. Population genomics of the facultatively mutualistic bacteria *Sinorhizobium meliloti* and *S. medicae*. PLoS Genet 8:e1002868. https://doi.org/10.1371/journal.pgen.1002868.
124. Holden MTG, Seth-Smith HMB, Crossman LC, Sebaihia M, Bentley SD, Cerdeño-Tárraga AM, Thomson NR, Bason N, Quail MA, Sharp S, Cherevach I, Churcher C, Goodhead I, Hauser H, Holroyd N, Mungall K, Scott P, Walker D, White B, Rose H, Iversen P, Mil-Homens D, Rocha EPC, Fialho AM, Baldwin A, Dowson C, Barrell BG, Govan JR, Vandamme P, Hart CA, Mahenthiralingam E, Parkhill J. 2009. The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol 191:261–277. https://doi.org/10.1128/JB.01230-08.
125. Holden MTG, Titball RW, Peacock SJ, Cerdeño-Tárraga AM, Atkins T, Crossman LC, Pitt T, Churcher C, Mungall K, Bentley SD, Sebaihia M, Thomson NR, Bason N, Beacham IR, Brooks K, Brown KA, Brown NF, Challis GL, Cherevach I, Chillingworth T, Cronin A, Crossett B, Davis P, DeShazer D, Feltwell T, Fraser A, Hance Z, Hauser H, Holroyd S, Jagels K, Keith KE, Maddison M, Moule S, Price C, Quail MA, Rabbinowitsch E, Rutherford K, Sanders M, Simmonds M, Songsivilai S, Stevens K, Tumapa S, Vesaratchavest M, Whitehead S, Yeats C, Barrell BG, Oyston PCF, Parkhill J. 2004. Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. Proc Natl Acad Sci U S A 101:14240–14245. https://doi.org/10.1073/pnas.0403302101.
126. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.
127. Janssen PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, Michaux A, Benotmane MA, Leys N, Vallaeys T, Lapidus A, Monchy S, Médigue C, Taghavi S, McCorkle S, Dunn J, van der Lelie D, Mergeay M. 2010. The complete genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh and anthropogenic environments. PLoS One 5:e10433. https://doi.org/10.1371/journal.pone.0010433.
128. Van Houdt R, Monsieurs P, Mijnendonckx K, Provoost A, Janssen A, Mergeay M, Leys N. 2012. Variation in genomic islands contribute to genome plasticity in *Cupriavidus metallidurans*. BMC Genomics 13:111. https://doi.org/10.1186/1471-2164-13-111.
129. Bavishi A, Abhishek A, Lin L, Choudhary M. 2010. Complex prokaryotic genome structure: rapid evolution of chromosome II. Genome 53:675–687. https://doi.org/10.1139/G10-046.
130. Guo HJ, Wang ET, Zhang XX, Li QQ, Zhang YM, Tian CF, Chen WX. 2014. Replicon-dependent differentiation of symbiosis-related genes in *Sinorhizobium* strains nodulating *Glycine max*. Appl Environ Microbiol 80:1245–1255. https://doi.org/10.1128/AEM.03037-13.
131. Dillon MM, Sung W, Lynch M, Cooper VS. 2015. The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. Genetics 200:935–946. https://doi.org/10.1534/genetics.115.176834.
132. Dillon MM, Cooper VS. 2016. The fitness effects of spontaneous mutations nearly unseen by selection in a bacterium with multiple chromosomes. Genetics 204:1225–1238. https://doi.org/10.1534/genetics.116.193060.
133. Dillon MM, Sung W, Sebra R, Lynch M, Cooper VS. 2017. Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. Mol Biol Evol 34:93–109. https://doi.org/10.1093/molbev/msw224.
134. Peters AE, Bavishi A, Cho H, Choudhary M. 2012. Evolutionary constraints and expression analysis of gene duplications in *Rhodobacter sphaeroides* 2.4.1. BMC Res Notes 5:192. https://doi.org/10.1186/1756-0500-5-192.
135. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41. https://doi.org/10.1186/1471-2105-4-41.
136. Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, Read TD, Dodson RJ, Umayam L, Brinkac LM, Beanan MJ, Daugherty SC, Deboy RT, Durkin AS, Kolonay JF, Madupu R, Nelson WC, Ayodeji B, Kraul M, Shetty J, Malek J, Van Aken SE, Riedmuller S, Tettelin H, Gill SR, White O, Salzberg SL, Hoover DL, Lindler LE, Halling SM, Boyle SM, Fraser CM. 2002. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci U S A 99:13148–13153. https://doi.org/10.1073/pnas.192319099.
137. Frank O, Göker M, Pradella S, Petersen J. 2015. Ocean's twelve: flagellar and biofilm chromids in the multipartite genome of *Marinovum algicola* DG898 exemplify functional compartmentalization. Environ Microbiol 17:4019–4034. https://doi.org/10.1111/1462-2920.12947.
138. Mahillon J, Chandler M. 1998. Insertion sequences. Microbiol Mol Biol Rev 62:725–774.
139. Elena SF, Ekunwe L, Hajela N, Oden SA, Lenski RE. 1998. Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica 102–103:349–358.
140. Wu Y, Aandahl RZ, Tanaka MM. 2015. Dynamics of bacterial insertion sequences: can transposition bursts help the elements persist? BMC Evol Biol 15:288. https://doi.org/10.1186/s12862-015-0560-5.
141. Villaseñor T, Brom S, Dávalos A, Lozano L, Romero D, García-de los Santos A. 2011. Housekeeping genes essential for pantothenate biosynthesis are plasmid-encoded in *Rhizobium etli* and *Rhizobium leguminosarum*. BMC Microbiol 11:66. https://doi.org/10.1186/1471-2180-11-66.
142. García-de los Santos A, Brom S. 1997. Characterization of two plasmid-borne *lpsβ* loci of *Rhizobium etli* required for lipopolysaccharide synthesis and for optimal interaction with plants. Mol Plant Microbe Interact 10:891–902. https://doi.org/10.1094/MPMI.1997.10.7.891.
143. Brom S, García-de los Santos A, Stepkowsky T, Flores M, Davila G, Romero D, Palacios R. 1992. Different plasmids of *Rhizobium leguminosarum* bv. *phaseoli* are required for optimal symbiotic performance. J Bacteriol 174:5183–5189.
144. Hynes MF, Simon R, Müller P, Niehaus K, Labes M, Pühler A. 1986. The two megaplasmids of *Rhizobium meliloti* are involved in the effective nodulation of alfalfa. Mol Gen Genet 202:356–362. https://doi.org/10.1007/BF00333262.
145. Hynes MF, McGregor NF. 1990. Two plasmids other than the nodulation plasmid are necessary for formation of nitrogen-fixing nodules by *Rhizobium leguminosarum*. Mol Microbiol 4:567–574. https://doi.org/10.1111/j.1365-2958.1990.tb00625.x.
146. González V, Santamaría RI, Bustos P, Hernández-González I, Medrano-Soto A, Moreno-Hagelsieb G, Janga SC, Ramírez MA, Jiménez-Jacinto V, Collado-Vides J, Dávila G. 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. Proc Natl Acad Sci U S A 103:3834–3839. https://doi.org/10.1073/pnas.0508502103.
147. Galardini M, Brilli M, Spini G, Rossi M, Roncaglia B, Bani A, Chiancianesi M, Moretto M, Engelen K, Bacci G, Pini F, Biondi EG, Bazzicalupo M, Mengoni A. 2015. Evolution of intra-specific regulatory networks in a multipartite bacterial genome. PLoS Comput Biol 11:e1004478. https://doi.org/10.1371/journal.pcbi.1004478.
148. Pini F, De Nisco NJ, Ferri L, Penterman J, Fioravanti A, Brilli M, Mengoni A, Bazzicalupo M, Viollier PH, Walker GC, Biondi EG. 2015. Cell cycle control by the master regulator CtrA in *Sinorhizobium meliloti*. PLoS Genet 11:e1005232. https://doi.org/10.1371/journal.pgen.1005232.
149. Bobik C, Meilhoc E, Batut J. 2006. FixJ: a major regulator of the oxygen limitation response and late symbiotic functions of *Sinorhizobium meliloti*. J Bacteriol 188:4890–4902. https://doi.org/10.1128/JB.00251-06.
150. Ronson CW, Nixon BT, Albright LM, Ausubel FM. 1987. *Rhizobium meliloti ntrA* (*rpoN*) gene is required for diverse metabolic functions. J Bacteriol 169:2424–2431. https://doi.org/10.1128/jb.169.6.2424-2431.1987.
151. Barnett MJ, Toman CJ, Fisher RF, Long SR. 2004. A dual-genome symbiosis chip for coordinate study of signal exchange and development in a prokaryote-host interaction. Proc Natl Acad Sci U S A 101:16636–16641. https://doi.org/10.1073/pnas.0407269101.
152. diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. 2014. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet 10:e1004742. https://doi.org/10.1371/journal.pgen.1004742.
153. Morton ER, Platt TG, Fuqua C, Bever JD. 2014. Non-additive costs and interactions alter the competitive dynamics of co-occurring ecologically distinct plasmids. Proc Biol Sci 281:20132173. https://doi.org/10.1098/rspb.2013.2173.
154. Dougherty K, Smith BA, Moore AF, Maitland S, Fanger C, Murillo R, Baltrus DA. 2014. Multiple phenotypic changes associated with large-scale horizontal gene transfer. PLoS One 9:e102170. https://doi.org/10.1371/journal.pone.0102170.

155. Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, Roach J, Grant SR, Jones CD. 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. PLoS Pathog 7:e1002132. https://doi.org/10.1371/journal.ppat.1002132.
156. Lee M-C, Marx CJ. 2012. Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet 8:e1002651. https://doi.org/10.1371/journal.pgen.1002651.
157. Hessen DO, Jeyasingh PD, Neiman M, Weider LJ. 2010. Genome streamlining and the elemental costs of growth. Trends Ecol Evol 25:75–80. https://doi.org/10.1016/j.tree.2009.08.004.
158. Vieira-Silva S, Touchon M, Rocha EPC. 2010. No evidence for elemental-based streamlining of prokaryotic genomes. Trends Ecol Evol 25:319–320. https://doi.org/10.1016/j.tree.2010.03.001.
159. Morton ER, Merritt PM, Bever JD, Fuqua C. 2013. Large deletions in the pAtC58 megaplasmid of *Agrobacterium tumefaciens* can confer reduced carriage cost and increased expression of virulence genes. Genome Biol Evol 5:1353–1364. https://doi.org/10.1093/gbe/evt095.
160. Mauchline TH, Fowler JE, East AK, Sartor AL, Zaheer R, Hosie AHF, Poole PS, Finan TM. 2006. Mapping the *Sinorhizobium meliloti* 1021 solute-binding protein-dependent transportome. Proc Natl Acad Sci U S A 103:17933–17938. https://doi.org/10.1073/pnas.0606673103.
161. MacLean AM, Finan TM, Sadowsky MJ. 2007. Genomes of the symbiotic nitrogen-fixing bacteria of legumes. Plant Physiol 144:615–622. https://doi.org/10.1104/pp.107.101634.
162. Couturier E, Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol 59:1506–1518. https://doi.org/10.1111/j.1365-2958.2006.05046.x.
163. Labbe RG, Huang TH. 1995. Generation times and modeling of enterotoxin-positive and enterotoxin-negative strains of *Clostridium perfringens* in laboratory media and ground beef. J Food Prot 58:1303–1306. https://doi.org/10.4315/0362-028X-58.12.1303.
164. Vieira-Silva S, Rocha EPC. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet 6:e1000808. https://doi.org/10.1371/journal.pgen.1000808.
165. Cowie A, Cheng J, Sibley CD, Fong Y, Zaheer R, Patten CL, Morton RM, Golding GB, Finan TM. 2006. An integrated approach to functional genomics: construction of a novel reporter gene fusion library for *Sinorhizobium meliloti*. Appl Environ Microbiol 72:7156–7167. https://doi.org/10.1128/AEM.01397-06.
166. Weng X, Xiao J. 2014. Spatial organization of transcription in bacterial cells. Trends Genet 30:287–297. https://doi.org/10.1016/j.tig.2014.04.008.
167. Xu Q, Dziejman M, Mekalanos JJ. 2003. Determination of the transcriptome of *Vibrio cholerae* during intraintestinal growth and midexponential phase *in vitro*. Proc Natl Acad Sci U S A 100:1286–1291. https://doi.org/10.1073/pnas.0337479100.
168. Yoder-Himes DR, Konstantinidis KT, Tiedje JM. 2010. Identification of potential therapeutic targets for *Burkholderia cenocepacia* by comparative transcriptomics. PLoS One 5:e8724. https://doi.org/10.1371/journal.pone.0008724.
169. Ramachandran VK, East AK, Karunakaran R, Downie JA, Poole PS. 2011. Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. Genome Biol 12:R106. https://doi.org/10.1186/gb-2011-12-10-r106.
170. López-Guerrero MG, Ormeño-Orrillo E, Acosta JL, Mendoza-Vargas A, Rogel MA, Ramírez MA, Rosenblueth M, Martínez-Romero J, Martínez-Romero E. 2012. Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. Plasmid 68:149–158. https://doi.org/10.1016/j.plasmid.2012.07.002.
171. Becker A, Bergès H, Krol E, Bruand C, Rüberg S, Capela D, Lauber E, Meilhoc E, Ampe F, de Bruijn FJ, Fourment J, Francez-Charlot A, Kahn D, Küster H, Liebe C, Pühler A, Weidner S, Batut J. 2004. Global changes in gene expression in *Sinorhizobium meliloti* 1021 under microoxic and symbiotic conditions. Mol Plant Microbe Interact 17:292–303. https://doi.org/10.1094/MPMI.2004.17.3.292.
172. Boussau B, Karlberg EO, Frank AC, Legault B-A, Andersson SGE. 2004. Computational inference of scenarios for α-proteobacterial genome evolution. Proc Natl Acad Sci U S A 101:9722–9727. https://doi.org/10.1073/pnas.0400975101.
173. Johnson TJ, Nolan LK. 2009. Pathogenomics of the virulence plasmids of *Escherichia coli*. Microbiol Mol Biol Rev 73:750–774. https://doi.org/10.1128/MMBR.00015-09.
174. Pál C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37:1372–1375. https://doi.org/10.1038/ng1686.
175. Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A 95:9413–9417. https://doi.org/10.1073/pnas.95.16.9413.
176. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304. https://doi.org/10.1038/35012500.
177. Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x.
178. Niehus R, Mitri S, Fletcher AG, Foster KR. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun 6:8924. https://doi.org/10.1038/ncomms9924.
179. Pini F, Galardini M, Bazzicalupo M, Mengoni A. 2011. Plant-bacteria association and symbiosis: are there common genomic traits in *Alphaproteobacteria*? Genes 2:1017–1032. https://doi.org/10.3390/genes2041017.
180. Park C, Zhang J. 2012. High expression hampers horizontal gene transfer. Genome Biol Evol 4:523–532. https://doi.org/10.1093/gbe/evs030.
181. Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res 16:636–643. https://doi.org/10.1101/gr.4746406.
182. Kurland CG, Canback B, Berg OG. 2003. Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A 100:9658–9662. https://doi.org/10.1073/pnas.1632870100.
183. Hao W, Golding GB. 2010. Inferring bacterial genome flux while considering truncated genes. Genetics 186:411–426. https://doi.org/10.1534/genetics.110.118448.
184. diCenzo GC, Checcucci A, Bazzicalupo M, Mengoni A, Viti C, Dziewit L, Finan TM, Galardini M, Fondi M. 2016. Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. Nat Commun 7:12219. https://doi.org/10.1038/ncomms12219.
185. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397. https://doi.org/10.1007/PL00006158.
186. Lercher MJ, Pál C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25:559–567. https://doi.org/10.1093/molbev/msm283.
187. Fei F, diCenzo GC, Bowdish DME, McCarry BE, Finan TM. 2016. Effects of synthetic large-scale genome reduction on metabolism and metabolic preferences in a nutritionally complex environment. Metabolomics 12:23. https://doi.org/10.1007/s11306-015-0928-y.
188. Tian CF, Zhou YJ, Zhang YM, Li QQ, Zhang YZ, Li DF, Wang S, Wang J, Gilbert LB, Li YR, Chen WX. 2012. Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. Proc Natl Acad Sci U S A 109:8629–8634. https://doi.org/10.1073/pnas.1120436109.
189. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S, Watanabe A, Idesawa K, Iriguchi M, Kawashima K, Kohara M, Matsumoto M, Shimpo S, Tsuruoka H, Wada T, Yamada M, Tabata S. 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. DNA Res 9:189–197. https://doi.org/10.1093/dnares/9.6.189.
190. Baumdicker F, Hess WR, Pfaffelhuber P. 2012. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol 4:443–456. https://doi.org/10.1093/gbe/evs016.
191. Rankin DJ, Rocha EPC, Brown SP. 2011. What traits are carried on mobile genetic elements, and why? Heredity (Edinb) 106:1–10. https://doi.org/10.1038/hdy.2010.24.
192. Lerat E, Ochman H. 2004. Psi-Phi: exploring the outer limits of bacterial pseudogenes. Genome Res 14:2273–2278. https://doi.org/10.1101/gr.2925604.
193. Song J, Ware A, Liu S-L. 2003. Wavelet to predict bacterial *ori* and *ter*: a tendency towards a physical balance. BMC Genomics 4:17. https://doi.org/10.1186/1471-2164-4-17.
194. Hendrickson H, Lawrence JG. 2006. Selection for chromosome architecture in bacteria. J Mol Evol 62:615–629. https://doi.org/10.1007/s00239-005-0192-2.
195. Morton RA, Morton BR. 2007. Separating the effects of mutation and

selection in producing DNA skew in bacterial chromosomes. BMC Genomics 8:369. https://doi.org/10.1186/1471-2164-8-369.
196. Darling AE, Miklós I, Ragan MA. 2008. Dynamics of genome rearrangement in bacterial populations. PLoS Genet 4:e1000128. https://doi.org/10.1371/journal.pgen.1000128.
197. Lesterlin C, Pages C, Dubarry N, Dasgupta S, Cornet F. 2008. Asymmetry of chromosome replichores renders the DNA translocase activity of FtsK essential for cell division and cell shape maintenance in *Escherichia coli*. PLoS Genet 4:e1000288. https://doi.org/10.1371/journal.pgen.1000288.
198. Hill CW, Gray JA. 1988. Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. Genetics 119:771–778.
199. Liu G-R, Liu W-Q, Johnston RN, Sanderson KE, Li S-X, Liu S-L. 2006. Genome plasticity and *ori-ter* rebalancing in *Salmonella typhi*. Mol Biol Evol 23:365–371. https://doi.org/10.1093/molbev/msj042.
200. Esnault E, Valens M, Espéli O, Boccard F. 2007. Chromosome structuring limits genome plasticity in *Escherichia coli*. PLoS Genet 3:e226. https://doi.org/10.1371/journal.pgen.0030226.
201. Mao X, Zhang H, Yin Y, Xu Y. 2012. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res 40:8210–8218. https://doi.org/10.1093/nar/gks605.
202. Koonin EV. 2009. Evolution of genome architecture. Int J Biochem Cell Biol 41:298–306. https://doi.org/10.1016/j.biocel.2008.09.015.
203. Srivatsan A, Tehranchi A, MacAlpine DM, Wang JD. 2010. Co-orientation of replication and transcription preserves genome integrity. PLoS Genet 6:e1000810. https://doi.org/10.1371/journal.pgen.1000810.
204. Cevallos MA, Cervantes-Rivera R, Gutiérrez-Ríos RM. 2008. The *repABC* plasmid family. Plasmid 60:19–37. https://doi.org/10.1016/j.plasmid.2008.03.001.
205. MacLellan SR, Zaheer R, Sartor AL, MacLean AM, Finan TM. 2006. Identification of a megaplasmid centromere reveals genetic structural diversity within the *repABC* family of basic replicons. Mol Microbiol 59:1559–1575. https://doi.org/10.1111/j.1365-2958.2006.05040.x.
206. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293. https://doi.org/10.1126/science.1181369.
207. Le TBK, Imakaev MV, Mirny LA, Laub MT. 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342:731–734. https://doi.org/10.1126/science.1242059.
208. Ried T, Schröck E, Ning Y, Wienberg J. 1998. Chromosome painting: a useful art. Hum Mol Genet 7:1619–1626. https://doi.org/10.1093/hmg/7.10.1619.
209. Cremer T, Cremer M. 2010. Chromosome territories. Cold Spring Harb Perspect Biol 2:a003889. https://doi.org/10.1101/cshperspect.a003889.
210. Shintani M, Suzuki-Minakuchi C, Nojiri H. 2015. Nucleoid-associated proteins encoded on plasmids: occurrence and mode of function. Plasmid 80:32–44. https://doi.org/10.1016/j.plasmid.2015.04.008.
211. San Millan A, Toll-Riera M, Qi Q, MacLean RC. 2015. Interactions between horizontally acquired genes create a fitness cost in *Pseudomonas aeruginosa*. Nat Commun 6:6845. https://doi.org/10.1038/ncomms7845.
212. Baltrus DA. 2013. Exploring the costs of horizontal gene transfer. Trends Ecol Evol 28:489–495. https://doi.org/10.1016/j.tree.2013.04.002.
213. Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A 95:6239–6244. https://doi.org/10.1073/pnas.95.11.6239.
214. Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801–3806. https://doi.org/10.1073/pnas.96.7.3801.
215. Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033–1034. https://doi.org/10.1093/bioinformatics/bts079.
216. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.
217. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
218. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
219. Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees, p 1–8. *In* Proceedings of the Gateway Computing Environments Workshop, New Orleans, LA. ACM, New York, NY.
220. Wu S, Zhu Z, Fu L, Niu B, Li W. 2011. WebMGA: a customizable Web server for fast metagenomic sequence analysis. BMC Genomics 12:444. https://doi.org/10.1186/1471-2164-12-444.

George C. diCenzo began work on this topic in 2010 as a Molecular Biology and Genetics undergraduate student. He obtained his B.Sc. in 2012 from McMaster University and defended his Ph.D. at the same institution in late 2016 under the supervision of Turlough Finan in the Department of Biology. He is currently a postdoctoral fellow at the University of Florence, Italy, under the supervision of Alessio Mengoni. During his Ph.D. work, he combined genome reduction approaches with molecular genetic, genomics, system-level, and *in silico* analyses to study the contributions of each replicon of *Sinorhizobium meliloti* to the biology of the free-living and symbiotic forms of the bacterium. His current work is focused on using *in silico* genome-scale metabolic network reconstruction and flux balance analysis, together with genetic studies, to characterize the genetics and metabolism of *S. meliloti* during symbiosis with alfalfa in order to develop strategies for manipulating this interaction.

Turlough M. Finan is a professor of Biology at McMaster University, Hamilton, Ontario, Canada. His interest in secondary replicons originated during his B.Sc. and M.Sc. in Microbiology under the supervision of Kieran Dunican, National University, Galway, Ireland. He obtained his Ph.D. in 1981 under the supervision of Carl Jordan, Microbiology Department, University of Guelph, Canada. Following a year at the Connaught Research Institute studying the neutralizing antigens of poliovirus, he then performed postdoctoral studies on *Sinorhizobium* under the supervision of Ethan Signer in the Department of Biology, Massachusetts Institute of Technology. He teaches undergraduate and graduate courses on Microbiology, Molecular Genetics, and Environmental Microbiology. For over 30 years, he has examined genomic and metabolic aspects of the interaction between *Sinorhizobium meliloti* and alfalfa. Initial studies focused on the detection and analysis of symbiotic loci and expanded to the biology of the 1.7-Mb pSymB replicon, including carbon and phosphate metabolism.

Genome Size and Structure, Bacterial

H Ochman and A Caro-Quintero, University of Texas, Austin, TX, USA

Glossary

Effective population size The size of the population that makes a genetic contribution to the subsequent generation – usually much smaller than the actual population size.
Genetic drift The change in gene frequencies in a population due to random sampling.
Okazaki fragments Short DNA fragments synthesized on the lagging strand during DNA replication.
Paralogs Genes whose shared ancestry can be traced to a duplication event.
Plasmids Small, circular, autonomously replicating, double-stranded DNA molecules that are distinct from the cell's chromosome.
Pseudogenes Previously functional regions that have been inactivated by mutation.
Xenologs Genes whose shared ancestry can be traced to a horizontal transfer event.

Abbreviations

bp basepair
kb kilobase
Mb megabase
sRNA Small RNA

The Architecture of Bacterial Genomes

Early work on diverse model systems, such as *Escherichia coli* and *Bacillus subtilis*, led to the view that all bacteria possessed a single circular chromosome. However, physical mapping techniques revealed that species from divergent bacterial phyla have linear chromosomes, including *Borrelia burgdorferi* (a spirochete), *Agrobacterium tumefaciens* (a proteobacterium), and *Streptomyces coelicolor* (an actinomycete), indicating that this chromosome structure arose multiple times independently (Casjens, 1998; Ochman, 2002). The prevalence of circular chromosomes seems to reside in the fact that linear chromosomes must solve the problem of fully replicating their chromosome ends. Otherwise, the advantage of one configuration over the other is unknown (Marri *et al.*, 2008): in experiments in which the normally circular chromosome of *E. coli* was linearized synthetically, showed no obvious changes in growth rate, gene expression, or cell morphology (Cui *et al.*, 2007).

Several bacterial species have genomes that are partitioned into multiple chromosomes, including at least one instance, that of *A. tumefaciens*, where the genome contains one linear and one circular chromosome. The presence of multiple chromosomes of different genetic content within a cell – as opposed to those bacteria that maintain multiple copies of their single chromosome – is not surprising given that bacteria often harbor additional replicons in the form of plasmids. Distinguishing between chromosomes and plasmids is not always straightforward, and differentiation between the two has been based on size, contents, copy number, dispensability, transmissibility, and mode of replication (Ochman, 2002). In the majority of cases, the origins of a second chromosome in a bacterial genome – and indeed, it appears that a single circular chromosome is the ancestral state – can be traced to an extrachromosomal accessory element that enlarged and acquired essential genes as opposed to the duplication or dissolution of the single ancestral chromosome.

Bacteria Have Compact Genomes

Bacterial genomes are notable in that they are tightly packed with protein-coding genes (**Figure 1**), which average about 1 kb in length and do not contain introns. Typically, 80–90% of a bacterial chromosome encodes proteins, and a large fraction of intergenic DNA is devoted to regulatory sequences and large numbers of other functional noncoding elements, such as sRNAs (as well as the structural RNAs required for protein assembly). This greatly contrasts the situation in the human genome, in which protein-coding regions are nearly a hundred times longer (owing mostly to the presence of introns) and noncoding regions constituted nearly 98% of the genome (Ahnert *et al.*, 2008). The paucity of nonfunctional DNA is one of the hallmarks of bacterial genomes, and their high-coding densities means that there is a strong association between genome size and gene number (Mira *et al.*, 2001).

Insights into the basis for the high density of functional sequences within bacterial genomes came from several sources, but particularly from those genomes that were anomalous in that they contained large numbers of pseudogenes. The first sequenced genomes shown to have substantial numbers of inactivated genes was *Mycobacterium leprae* (the etiological agent of leprosy), in which over half of its encoded sequences were pseudogenes, which had intact counterparts in its congener *Mycobacterium tuberculosis* (Cole *et al.*, 2001). When comparing nonfunctional regions, such as pseudogenes, to their functional counterparts across a wide array of bacterial lineages, the nonfunctional regions display an excess of deletional events, implying that the mutational process in bacteria

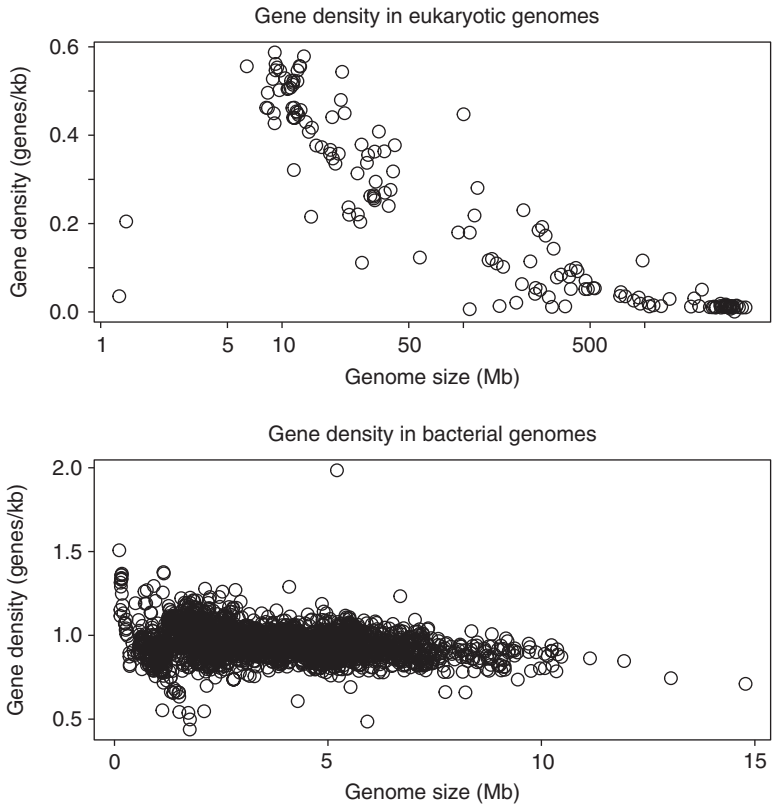


Figure 1 Relationship between gene density and genome size. Gene densities are calculated as the number of genes (in a genome of a given size) per kilobase. Note that gene densities in bacteria are similar regardless of genome size (such that gene numbers increase with genome size), whereas in eukaryotes, there is an inverse relationship between genome size and gene density (such that larger genomes do not necessarily encode additional genes).

is biased toward deletions over insertions. This mutational bias has also been observed in experimental populations, in which evolved strains can harbor individual deletions up to 200 kb in length (Nilsson *et al.*, 2005).

The deletional bias maintains the high density of functional genes observed in bacterial genomes: when inactivating mutations occur in genes that are no longer required, the nonfunctional regions gradually erode through deletions and are eventually eliminated (Andersson and Andersson, 2001; Mira *et al.*, 2001). The presence of a deletional process that removes nonfunctional sequences implies that virtually all genes in a bacterial genome are functional and maintained in the genome by natural selection (Kuo and Ochman, 2009).

Determinants of Genome Size in Bacteria

Among bacteria, genome sizes varies over two orders of magnitude, which, due to the relationship between genome size and gene contents in bacteria means that lineages can differ by as much as a 100-fold in their gene numbers. (In contrast, humans and the budding yeast, *Saccharomyces cerevisiae*, differ by only fourfold in gene number.) Even after the sequencing of only a dozen bacterial genomes, there was a notable association between genome size and bacterial lifestyle: those bacteria possessing small genomes were host-associated pathogens and symbionts, whereas those bacteria with large genomes were either free-living or environmental isolates. Currently, the bacteria with the smallest genomes are obligate symbionts of insects (Moran and Bennett, 2014), with the tiniest genome yet sequenced – only 112 kb – belonging to *Nasuia deltocephalinicola*, a symbiont of leafhoppers (Bennett and Moran, 2013). At the other end of spectrum are free-living bacteria, and a soil-associated species, *Ktedonobacter racimifer*, which at 15.6 Mb has the largest sequenced genome to date (Chang *et al.*, 2011).

Phylogenetic analyses indicate that host-associated bacterial lineages are derived from ancestors with larger genomes, which is not surprising considering that bacteria as a group are much more ancient than their potential eukaryotic hosts. Pathogen and symbiont genomes can possess fewer functional genes because many nutrients and biochemical pathways are supplied by their hosts. Considering all of these factors together, it is possible to trace the evolutionary progression toward highly reduced genomes and to account for the variation in the gene repertoires observed among contemporary species. Initially, (ancestral) free-living bacteria become associated with a (nutrient-rich) eukaryotic host. This association renders many bacterial genes superfluous in the host environment, and these unnecessary genes accumulate mutations, becoming pseudogenes. The inactivated genes are eventually removed by the pervasive mutational bias toward deletions, resulting in a compact genome harboring only functional genes. Note that this scenario accounts not only for the high gene densities

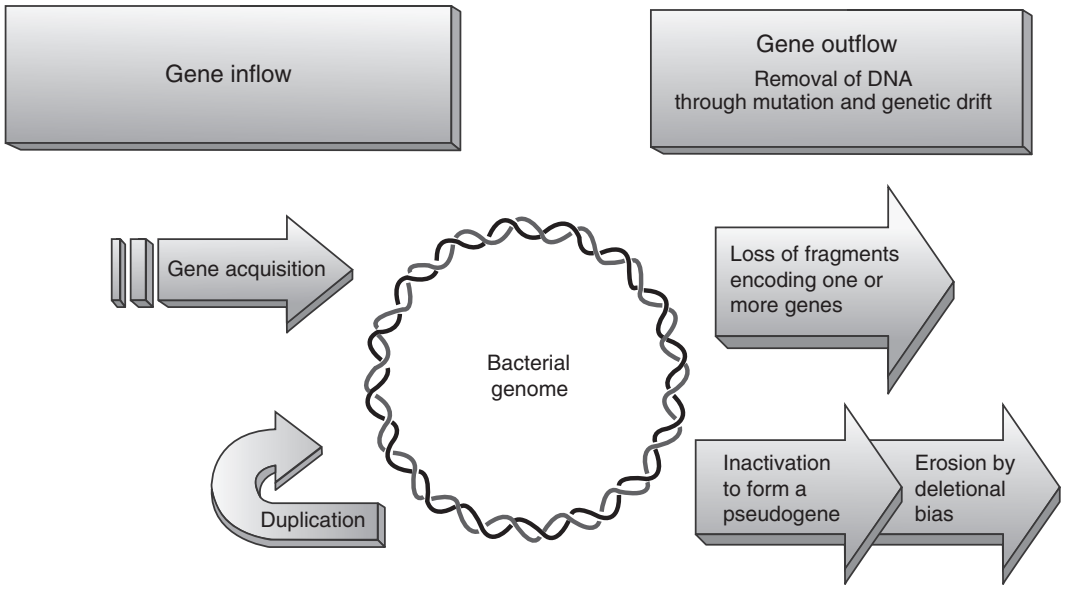


Figure 2 Interplay of factors impacting bacterial genome size. New sequences are acquired by gene transfer and gene duplication, whereas DNA loss occurs by large deletions that remove multiple genes in a single event and by the erosion of pseudogenes and other nonfunctional sequences. Modified from Mira, A., Ochman, H., Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics 10, 589–596.

observed in the largest and smallest genomes but also for the presence of large numbers of pseudogenes observed in recent pathogens and other host-associated lineages.

The compactness of bacterial genomes has traditionally been thought to result from 'streamlining' and considered to be an adaptation to increase replication speed and growth rates. For the most part, this view is incorrect. When looking across bacteria, there is no clear association between generation time and genome size (Mira *et al.*, 2001). In tests within species, natural strains of *E. coli* that differed by 15% in genome size did not differ in growth rates (Bergthorsson and Ochman, 1998), and isogenic strains of *Salmonella* engineered to harbor duplications up to 700 bp in length did not replicate significantly slower than the parental strain in either minimal or nutrient-rich media (Matthews and Maloy, 2010). Although bacterial genome size is not generally driven by selection for replication efficiency, a few species in nutritionally poor environments have reduced genomes in order to decrease the metabolic burden associated with replicating extra DNA. The marine planktonic bacteria, *Prochlorococcus* (Dufresne *et al.*, 2005) and *Pelagibacter ubique* (Giovannoni *et al.*, 2005) occupy environments where nitrogen and phosphorous concentration are limited, and both possess the smallest and most gene-dense genomes among free-living bacteria.

The contrasting view is that bacterial genome size is governed by a nonadaptive process (Kuo *et al.*, 2009). Due to the association between genome size and gene number in bacteria, the evolutionary forces that act on individual genes will have profound effects on overall genome size. As a result, those species with small effective population sizes, such as those whose populations are severely restricted during transmission between hosts, are subject to genetic drift, which reduces the efficacy of selection and allows the accumulation of deleterious mutations, even in useful genes. Thus, bacteria with small effective population sizes, such as symbionts and pathogens, will have the smallest genomes because they are more susceptible to gene decay and loss, even for genes that are usually considered beneficial.

The view that small genome size in bacteria is largely a consequence of a nonadaptive process, i.e., genetic drift, runs counter to the situation in eukaryotes. Lynch (2007) has posited that reductions in the efficacy of selection in eukaryotic species with small effective population sizes have allowed the proliferation of deleterious sequences in the form of introns and transposable elements, both of which serve to increase genome size. Therefore, increased genetic drift has caused the size of eukaryotic genomes to increase but bacterial genomes to decrease.

Offsetting the deletional processes that serve to reduce genome size, bacterial genomes can grow by gene duplication and gene acquisition (*a.k.a.* lateral/horizontal gene transfer) (**Figure 2**). To determine the relative contributions of these two processes to the emergence of new bacterial genes, the numbers of paralogs (genes arising by duplication) and xenologs (genes acquired by transfer) were compared in bacterial groups across a wide range of genome sizes (Treangen and Rocha, 2011). In general, gene transfer played a larger role in the expansion of protein families and provided bacteria with genes of new function, whereas gene duplication usually led to an increase in gene dosage.

Genome Size Variation within Species

DNA content was considered to be invariant within a species, hence the application of the term 'C(onstant)-value' to denote the amount of DNA contained in a gamete or genome. Whereas the sexual system of eukaryotes has a homogenizing effect on chromosome size, no such constraints are imposed on organisms that reproduce by binary fission. But because the

Figure 3 Genome size variation within bacterial species. Size ranges are shown for species for which complete genome sequences are available for at least 10 strains.

members of a bacterial species are metabolically and phenotypically similar – in fact, species assignment and differentiation were based largely on metabolic capabilities before DNA typing became routine – the genomes of bacteria typed to the same species were thought to be alike in their sizes and contents, with variation, if any, attributable to the sporadic distribution of extrachromosomal elements among genomes.

There is early evidence, based on DNA reassociation experiments, that there was substantial variation in the genome sizes among clinical isolates of *E. coli* – variation could not be attributed to extrachromosomal sequences since many strains had smaller genomes than a control strain known to lack plasmids (Brenner *et al.*, 1972). However, it was not until the broad-scale application of pulsed-field gel electrophoresis – a method that allows resolution of very large DNA fragments – that the true extent of genome size diversity within bacterial species was fully appreciated (Cole and Saint Girons, 1994).

The sequencing and assembly of complete bacterial genomes has led to comparisons of the contents of genomes among closely related bacteria, and scrutiny of the extent of variation and the specific changes that contributed to variation in genome size (**Figure 3**). The first in-depth analysis of multiple sequenced members of bacterial species was of strains of *E. coli* that differed in their pathogenic potential (Welch *et al.*, 2002). The most pronounced source of the differences among these strains of *E. coli* was the integration of large DNA segments, mostly by phage-mediated events, forming islands of genes that were unique to a particular strain. This seems to be a common theme among bacteria: although changes in the numbers of repetitive and translocatable elements can inflate and deflate genomes, the acquisition of large regions by horizontal gene transfer is the major contributor to the genome size variation among members of a bacterial species. The within-species variation in genome size and contents has led

to the concept of the 'core-genome,' which comprises those genes that are present in (virtually) all members of a species, and the 'pan-genome,' which is the entire set of genes encoded by the species (Tettelin *et al.*, 2005).

Genome Organization

Genes encoded are not positioned at random along the bacterial chromosome. It has long been known that certain sets of genes are situated in proximity because they constitute a single functional unit (i.e., an operon); however, the location, arrangement, and distribution of genes in bacterial genome are influenced by several gene- and genome-level factors, many of which result from the manner in which replication occurs. Recall that in bacteria, there is a single replication origin, and that replication of the circular chromosome proceeds bidirectionally to a terminus that is situated transversely, at approximately 180°, from the origin. This replication process is asymmetric, generating leading strands, which are synthesized continuously from the origin to the terminus, and lagging strands, which are replicated discontinuously through the synthesis and joining of short Okazaki fragments.

Maintenance of Chromosome Balance

Early comparisons of the genetic maps of *E. coli* and *Salmonella typhimurium*, related enteric species thought to have diverged about hundred million years ago, displayed a high degree of concordance but differ with respect to a large inversion that spanned the replication terminus (Krawiec and Riley, 1990). Additionally, other *Salmonella* strains possess different inversions that were similarly symmetric with respect to the replication terminus, suggesting that inversions that offset the positions of the origin and terminus relative to one another are deleterious (Liu *et al.*, 1993). Although experimentally disrupting the positions of the origin and terminus through the introduction of large duplications on one side of the chromosome showed no appreciable effect on growth rates (Matthews and Maloy, 2010), there is evidence that selection acts to maintain chromosome balance over evolutionary timescales. Using physical maps to reconstruct the history of the *Salmonella paratyphi* genome exposed that an insertion that disrupted the balance between origin and the terminus was succeeded by an inversion that restored their positions (Liu and Sanderson, 1995).

Gene Location and Orientation

Inversions that maintain chromosome balance, i.e., those that are symmetric with respect to replication origin and terminus, preserve not only the positions of the origin and terminus but also the distances of all genes relative to the origin and terminus. All other inversions alter the positions of some genes relative to the origin. Moreover, those inversions that do not include the replication origin or terminus reverse gene orientation, such that genes that were originally encoded on the leading become lagging-strand genes and vice versa. Comparing the relative positions of genes in pairs of related genomes revealed a remarkable pattern: most genes retain the same orientation and the same relative distance to the replication origin despite the occurrence of numerous rearrangement events, yielding whole-genome alignment plots that display a characteristic X-shaped pattern (Eisen *et al.*, 2000; Tillier and Collins, 2000; Suyama and Bork, 2001; **Figure 4**).

How can proximity to the replication origin exert a selective effect on genes and chromosome organization? This can occur in two ways: Firstly, both mutation rates (Mira and Ochman, 2002) and recombination events (Louarn *et al.*, 1994) increase with distance from the replication origin, so it is possible that selection might favor the positioning of highly conserved, essential genes nearer to the replication origin as is observed in *B. subtilis* and *E. coli* (Rocha and Danchin, 2003). Secondly, because new rounds of replication can be initiated before the previous round is complete, genes closer to the replication origin exist in higher dosage and are more highly expressed than those near the replication terminus (Segall *et al.*, 1988; Liu and Sanderson, 1996). Therefore, the relocation of genes or operons by inversions or transpositions can alter their expression patterns and have a detrimental effect on the cell.

How can strand orientation exert a selective effect on gene and chromosome organization? The difference in the processes

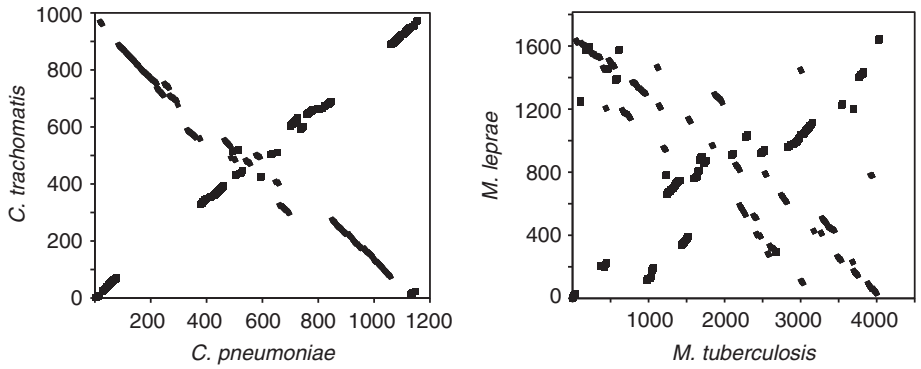


Figure 4 Correspondence of positions of homologous genes in pairs of sequenced genomes in *Chlamydia* (*C. pneumoniae* and *C. trachomatis*) and *Mycobacterium* (*M. tuberculosis* and *M. leprae*). Genomes are represented linearly beginning at the replication origin (position 0) with gene positions, indicated in kb from the replication origin, proceeding clockwise around the chromosome. Scatterplots redrawn from Tillier, E.R., Collins, R.A., 2000. Genome rearrangement by replication-directed translocation. Nature Genetics 26, 195–197 and Eisen, J.A., Heidelberg, J.F., White, O., Salzberg, S.L., 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biology 1, research0011.1–0011.9.

used to replicate the leading and lagging strands causes asymmetries in the rates and patterns of mutations, and hence, the nucleotide contents of each strand (Francino and Ochman, 1997; Frank and Lobry, 1999). In general, there is an excess of guanine residues on the leading strand – termed 'GC skew,' which calculated as degree of the difference in the number of guanines and cytosines on a given strand relative to the total number of these guanines and cytosine residues (i.e., G − C/G + C) – and many genome sequences show an abrupt change in the direction of skew at the replication origin and terminus (Lobry, 1996).

Because the same DNA strands are used as templates for replication and transcription, both processes will occur simultaneously; however, DNA polymerase proceeds at over 10 times the speed of RNA polymerase causing collisions between the two polymerization machineries. On the leading strand, replication and transcription proceed in the same direction, such that collisions between DNA polymerase and RNA polymerase are co-oriented and produce fully formed transcripts. On the lagging strand, however, DNA polymerase and RNA polymerase have head-on collisions that abort transcription (Rocha and Danchin, 2003; Merrikh *et al.*, 2012). Owing to the deleterious nature of head-on collisions, there is an asymmetric distribution of genes between the two strands. In most genomes, the majority of genes are preferentially encoded on the leading strand (Mao *et al.*, 2012; Chen and Zhang, 2013), indicating that there is selection against inversions and translocations that alter the orientation of genes or operons.

Genomic Base Composition

One of the most highly variable features of bacterial genomes is overall base composition, which among sequenced genomes ranges from 13% to 75% G + C. The fact base composition is relatively homogeneous within a genome but varies greatly among genomes led to the view, developed in the 1960s, that the variation was caused by differences in the underlying patterns of mutations, such that high G + C and low G + C organisms differed in their spectra of replication errors (Freese, 1962; Sueoka, 1962). (It is interesting to note that this attribution of molecular variation to a strictly mutational process preceded Kimura's neutral theory of molecular evolution by several years.) An alternative explanation is that the variation in genomic base compositions is adaptive, such that the differences are due to a selective force that favors certain base compositions under some particular environment conditions (McEwan *et al.*, 1998; Naya *et al.*, 2002; Rocha and Danchin, 2002; Romero *et al.*, 2009). In that G/C basepairing is more thermally stable than A/T basepairing, it has been suggested repeatedly that genome base compositions reflect growth conditions; however, there is no association between growth temperature and genomic base composition when analyzed across diverse taxa (and in fact, many thermophiles have genomic G + C contents that are substantially lower than those of mesophiles) (Galtier and Lobry, 1997; Hurst and Merchant, 2001; Wang *et al.*, 2006).

To date, no fewer than a dozen factors have been proposed as the source of the base-compositional variation in bacteria (Rocha and Feil, 2010); however, the lack of a single unifying explanation has led many to adopt the decades-old view that the variation is neutral and caused by differences in the mutational process. Recent analyses performed on multiple taxa suggest that, even in bacterial groups with intermediate and high genomic G + C contents, the mutations are biased toward A + T (Hershberg and Petrov, 2010; Hildebrand *et al.*, 2010). These results, based on sequence comparisons, are most readily explained by the action of natural selection favoring individual base substitutions that increase the G + C composition of the genome; however, the manner in which selection is operating is still unknown.

See also: Genome Organization, Evolution of. Genome Plasticity, Bacterial. Mutation and Genome Evolution

References

Ahnert, S.E., Fink, T.M., Zinovyev, A., 2008. How much non-coding DNA do eukaryotes require? Journal of Theoretical Biology 252, 587–592.
Andersson, J.O., Andersson, S.G.E., 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Molecular Biology and Evolution 18, 829–839.
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, 1675–1688.
Bergthorsson, U., Ochman, H., 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Molecular Biology and Evolution 15, 6–16.
Brenner, D.J., Fanning, G.R., Skerman, F.J., Falkow, S., 1972. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. Journal of Bacteriology 109, 953–965.
Casjens, S., 1998. The diverse and dynamic structure of bacterial chromosomes. Annual Review of Genetics 32, 339–377.
Chang, Y.-J., Land, M., Hauser, L., *et al.*, 2011. Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium *Ktedonobacter racemifer* type strain (SOSP1-21T). Standards in Genomic Science 5, 97–111.
Chen, X., Zhang, J., 2013. Why are genes encoded on the lagging strand of the bacterial genome? Genome Biology and Evolution 5, 2436–2439.
Cole, S.T., Eiglmeier, K., Parkhill, J., *et al.*, 2001. Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011.
Cole, S.T., Saint Girons, I., 1994. Bacterial genomics. FEMS Microbiology Reviews 14, 139–160.
Cui, T., Moro-Oka, N., Ohsumi, K., *et al.*, 2007. *Escherichia coli* with a linear genome. EMBO Reports 8, 181–187.
Dufresne, A., Garczarek, L., Partensky, F., 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biology 6, R14.
Eisen, J.A., Heidelberg, J.F., White, O., Salzberg, S.L., 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biology 1, research0011.1–0011.9.
Francino, M.P., Ochman, H., 1997. Strand asymmetries in DNA evolution. Trends in Genetics 13, 240–245.
Frank, A.C., Lobry, J.R., 1999. Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. Gene 238, 65–77.
Freese, E., 1962. On the evolution of the base composition of DNA. Journal of Theoretical Biology 3, 82–101.
Galtier, N., Lobry, J.R., 1997. Relationships between genomic G + C content, RNA secondary structures and optimal growth temperature in prokaryotes. Journal of Molecular Evolution 44, 632–636.
Giovannoni, S.J., Tripp, H.J., Givan, S., *et al.*, 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242–1245.
Hershberg, R., Petrov, D.A., 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genetics 6, e1001115.
Hildebrand, F., Meyer, A., Eyre-Walker, A., 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genetics 6, e1001107.

Hurst, L.D., Merchant, A.R., 2001. High guanine–cytosine content is not an adaptation to high temperature: A comparative analysis amongst prokaryotes. Proceedings of the Royal Society B: Biological Sciences 268, 493–497.
Krawiec, S., Riley, M., 1990. Organization of the bacterial chromosome. Microbiological Reviews 54, 502–539.
Kuo, C.-H., Moran, N.A., Ochman, H., 2009. The consequences of genetic drift for bacterial genome complexity. Genome Research 19, 1450–1454.
Kuo, C.-H., Ochman, H., 2009. Deletional bias across the three domains of life. Genome Biology and Evolution 1, 145–152.
Liu, S.-L., Hessel, A., Sanderson, K.E., 1993. The *Xba*I–*Bln*I–*Ceu*I genomic cleavage map of *Salmonella enteritidis* shows an inversion relative to *Salmonella typhimurium* LT2. Molecular Microbiology 10, 655–664.
Liu, S.-L., Sanderson, K.E., 1995. Rearrangements in the genome of the bacterium *Salmonella typhi*. Proceedings of the National Academy of Sciences of the United States of America 92, 1018–1022.
Liu, S.-L., Sanderson, K.E., 1996. Highly plastic chromosomal organization in *Salmonella typhi*. Proceedings of the National Academy of Sciences of the United States of America 93, 10303–10308.
Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology and Evolution 13, 660–665.
Louarn, J., Cornet, F., François, V., Patte, J., Louarn, J.-M., 1994. Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: Possible relation to nucleoid organization. Journal of Bacteriology 176, 7524–7531.
Lynch, M., 2007. The Origins of Genome Architecture, first ed. Sunderland, MA: Sinauer Associates.
Mao, X., Zhang, H., Yin, Y., Xu, Y., 2012. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Research 40, 8210–8218.
Matthews, T.D., Maloy, S., 2010. Fitness effects of replichore imbalance in *Salmonella enterica*. Journal of Bacteriology 192, 6086–6088.
Marri, P.R., Harris, L.K., Houmiel, K., Slater, S.C., Ochman, H., 2008. The effect of chromosome geometry on genetic diversity. Genetics 179, 511–517.
McEwan, C.E., Gatherer, D., McEwan, N.R., 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas 128, 173–178.
Merrikh, H., Zhang, Y., Grossman, A.D., Wang, J.D., 2012. Replication–transcription conflicts in bacteria. Nature Reviews Microbiology 10, 449–458.
Mira, A., Ochman, H., 2002. Gene location and bacterial sequence divergence. Molecular Biology and Evolution 19, 1350–1358.
Mira, A., Ochman, H., Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics 10, 589–596.
Moran, N.A., Bennett, G.M., 2014. The tiniest tiny genomes. Annual Review of Microbiology 68, 195–215.
Naya, H., Romero, H., Zavala, A., Alvarez, B., Musto, H., 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. Journal of Molecular Evolution 55, 260–264.
Nilsson, A.I., Koskiniemi, S., Eriksson, S., *et al*., 2005. Bacterial genome size reduction by experimental evolution. Proceedings of the National Academy of Sciences of the United States of America 102, 12112–12116.
Ochman, H., 2002. Chromosome arithmetic and geometry. Current Biology 12, R427–R428.
Rocha, E.P.C., Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. Trends in Genetics 18, 291–294.
Rocha, E.P.C., Danchin, A., 2003. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Research 31, 6570–6577.
Rocha, E.P.C., Feil, E.J., 2010. Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? PLoS Genetics 6, e1001104.
Romero, H., Pereira, E., Naya, H., Musto, H., 2009. Oxygen and guanine-cytosine profiles in marine environments. Journal of Molecular Evolution 69, 203–206.
Segall, A.M., Mahan, M.J., Roth, J.R., 1988. Rearrangement of the bacterial chromosome: Forbidden inversions. Science 241, 1314–1318.
Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proceedings of the National Academy of Sciences of the United States of America 48, 582–592.
Suyama, M., Bork, P., 2001. Evolution of prokaryotic gene order: Genome rearrangements in closely related species. Trends in Genetics 17, 10–13.
Tettelin, H., Masignani, V., Cieslewicz, M.J., *et al*., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. Proceedings of the National Academy of Sciences of the United States of America 102, 13950–13955.
Tillier, E.R., Collins, R.A., 2000. Genome rearrangement by replication-directed translocation. Nature Genetics 26, 195–197.
Treangen, T.J., Rocha, E.P.C., 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genetics 7, e1001284.
Wang, H.-C., Susko, E., Roger, A.J., 2006. On the correlation between genomic G + C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. Biochemical and Biophysical Research Communications 342, 681–684.
Welch, R.A., Burland, V., Plunkett, G., *et al*., 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America 99, 17020–17024.