Microbial genomes

Biosciences 741: Genomics
Fall, 2013
Week 15

Microbes are single-celled organisms

- Archaebacteria have small genomes, high gene density, usually one circular chromosome. However, mechanisms of DNA replication & gene expression are more similar to eukaryotes… They are now regarded as a separate kingdom of life, but are also included in the category of "microbes". Examples include many (but not all) thermophilic bacteria.
- Eubacteria (most bacteria) have small genomes, high gene density, usually one circular chromosome. Examples include *Salmonella* and *E. coli*. The term "bacteria" is often used to refer to Eubacteria.
- Eukaryotes have moderate to large genomes, moderate to low gene density, introns, and a nucleus. Examples include yeast and humans. Yeast are microbes; humans are not.
- Metazoans (multicellular animals) are eukaryotes but not microbes. All metazoans share a single common ancestor.

Bacterial Genomes

- Most bacterial genomes contain about one gene per kb (i.e., little non-coding DNA). This makes whole genome shotgun assembly relatively easy.
- Gene identification is also relatively easy - pseudogenes are rare, and most long open reading frames do turn out to be functional genes.
- Synteny maps undergo rapid rearrangements, even between closely related species. However, a subset of genes tend to remain relatively close to the origin of DNA replication.
- Individual genes are frequently not conserved between closely related species, or even sub-species (strains). This is due to frequent gene loss plus lateral gene transfer. Much new biology can be inferred from each new bacterial genome sequence.
- Species are classified at present mostly by of the basic pathways that underlie the central dogma (DNA replication, transcription, translation). These genes appear to be rarely or never lost in bacteria, and hence rarely or never acquired by lateral gene transfer.

Larger prokaryotic genomes devote a higher percentage of their genes to regulatory functions

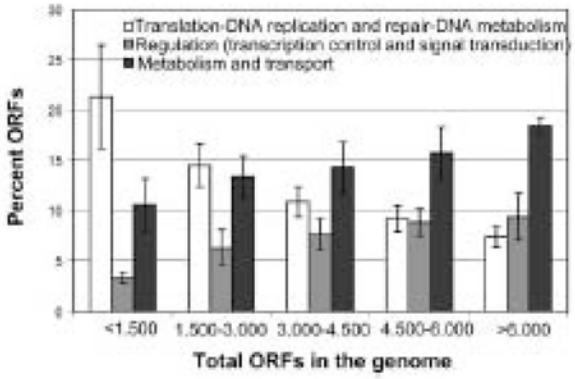


Fig. 5. Summary of the shifts in gene content with genome size in prokaryotic genomes. The bars represent the sum of the COG functional categories, which showed strong correlation with genome size and are involved in the same major cellular processes. Only normalized genomes (represented by solid squares in Fig. 1) have been included. Errors bars represent the standard deviation from the mean except for the last genome size class, where error bars represent data range due to a small number of normalized genomes in this class (three genomes).

Number of ORFs and amount of noncoding DNA in prokaryotic genomes

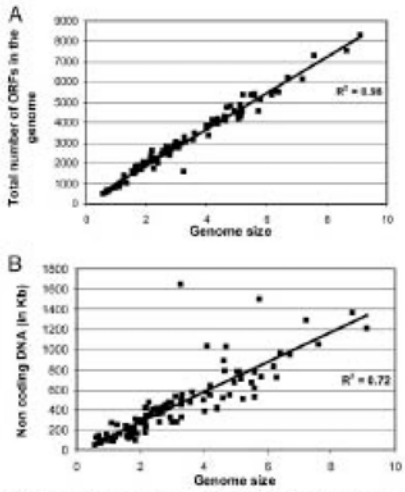


Fig. 2. Correlation among total number of ORFs in the genome, noncoding DNA, and genome size for prokaryotic genomes. (A) The total number of ORFs in the genome vs. the genome size for 115 completed prokaryotic genomes. (B) The total amount of noncoding DNA in the genome vs. genome size.

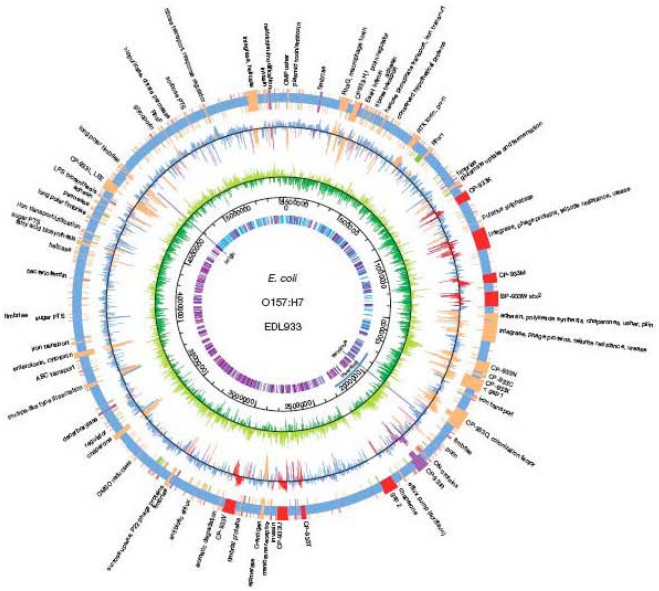


Figure 1 Circular genome map of EDL933 compared with MG1655. Outer circle shows the distribution of islands: shared co-linear backbone (blue); position of EDL933-specific sequences (O-islands) (red); MG1655-specific sequences (K-islands) (green); O-islands and K-islands at the same locations in the backbone (tan; hypervariable (purple). Second circle shows the G+C content calculated for each gene longer than 100 amino acids, plotted around the mean value for the whole genome, colour-coded like outer circle. Third circle shows the GC skew for third-codon position, calculated for each gene longer than 100 amino acids: positive values, lime; negative values, dark green. Fourth circle gives the scale in base pairs. Fifth circle shows the distribution of the highly skewed octamer Chi (GCTGGTGG), where bright blue and purple indicate the two DNA strands. The origin and terminus of replication, the chromosomal inversion and the locations of the sequence gaps are indicated. Figure created by Genvision from DNASTAR.

Codon bias in 18 species of bacteria, analyzed by self-organizing maps (Kanaya et al., 2001)

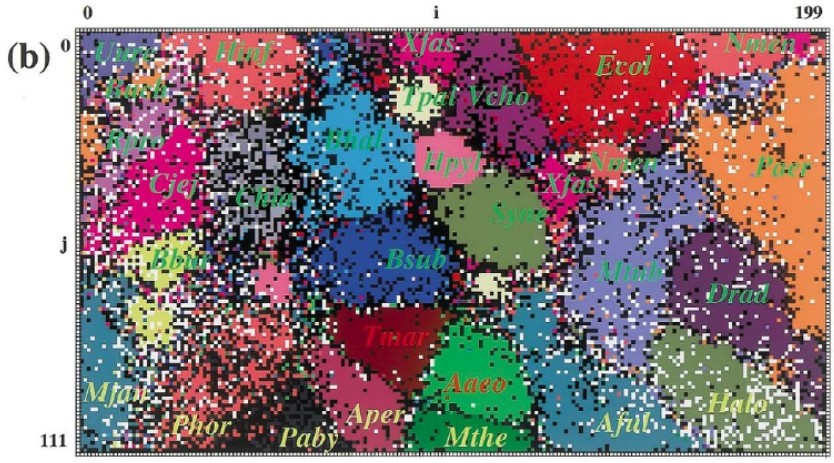


Fig. 1. Gene classification by (a) initial weights and (b) final weights: *A. aeolicus* (abbreviated as Aaeo), *A. fulgidus* (Aful), *A. pernix* (Aper), *B. subtilis* (Bsub), *B. halodurans* (Bhal), *B. burgdorferi* (Bbur), *Buchnera* sp. (Buch), *C. jejuni* (Cjej), *C. trachomatis* and *C. pneumonia* (Chla), *D. radiodurans* (Drad), *E. coli* (Ecol), *H. influenzae* (Hinf), *Halobactgerium* sp. (Halo), *H. pylori* (Hpyl), *M. jannaschii* (Mjan), *M. thermoautotrophicum* (Mthe), *M. tuberculosis* (Mtub), *N. meningitidis* (Nmen), *P. aeruginosa* (Paer), *P. abyssi* (Paby), *P. horikoshii* (Phor), *R. prowazekii* (Rpro), *Synechocystis* sp. (Syne), *T. maritima* (Tmar), *T. pallidum* (Tpal), *U. urealyticum* (Uure), *V. cholerae* (Vcho), and *X. fastidiosa* (Xfas). Archaea, eubacteria and two thermophilic bacteria are denoted by yellow, green, and blue letters. The configuration of bacterial species in (b) is depicted in Fig. 2. These SOM results are available on the Xanagen Inc. web (URL http://www.xanagen.com).

Horizontally-transferred gene islands

- Horizontally-transferred gene islands typically have a base composition and codon bias that is initially distinctly different from their new host.
- Over time, mutation of the third codon position causes the codon bias to equilibrate to the new host bias (in 1-2 million years).
- Mutation of the first and second codon positions to adopt the host base composition requires much longer, approximately 50-100 million years. Thus comparison of first, second, and third codon positions allows estimation of the age of gene islands, even though the species of origin may be unknown.
- Computational analysis of gene islands in E. coli and other species suggests that: (i) young islands are much more common than old; and (ii) young islands are significantly larger than older islands.

Lineage-specific segments: many are about the size of a gene, but many lineage-specific segments contain about 5-50 genes

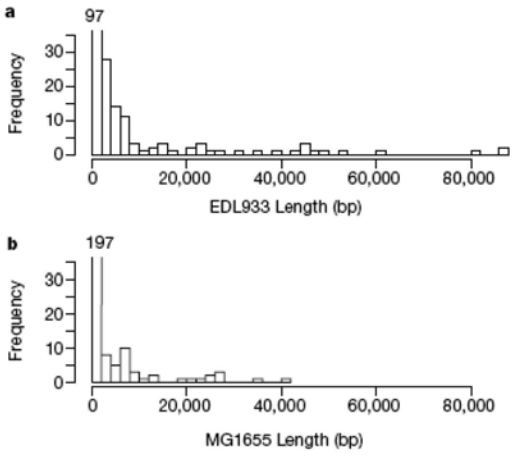


Figure 3 Histograms of lineage-specific segment lengths. **a**, EDL933; **b**, MG1655. The frequencies for the smallest length class are truncated to emphasize the distribution of longer clusters.

Emerging infectious diseases

- Approximately 160 newly emerging infectious diseases caused by bacteria have been discovered in the past 70 years.
- This is believed to be a real increase in EID rate, driven primarily by population growth, as well as agricultural and travel practices.
- Climate change and improved detection are doubtless also involved.
- Pathogens often arise by transfer of "pathogenicity islands", which may be plasmids, bacteriophage, transposons, and/or site-specific recombining "cassette chromosomes", particularly in *Staphylococcus*.
- The advent of whole-genome bacterial genome sequencing has allowed the forensic identification of sources of emerging diseases and even individual infections, based on unique SNPs.

Pathogenicity islands

Pathogenicity islands may include genes for toxins, that increase the virulence of bacteria that carry it. However, virulence carries both costs and benefits to the bacterial cell, as hurting or killing its host may reduce its own reproduction.

Pathogenicity islands often include genes for antibiotic resistance, which can be beneficial even to non-pathogenic bacteria.

In fact, MRSA, one of the most serious emerging diseases, originated from the animal commensal S. sciuri, which contains a penecillin-binding protein with a high degree of similarity to that found in MRSA.

Use of antibiotics in animal feed is likely to have long-term consequences for world health.

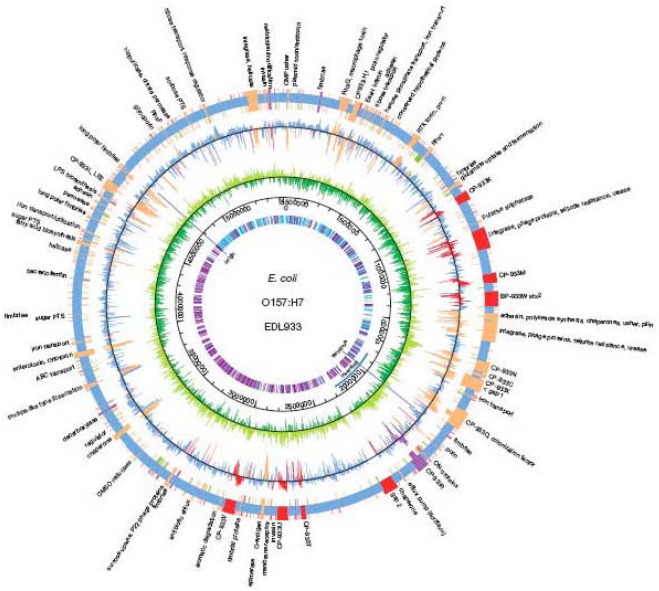


Figure 1 Circular genome map of EDL933 compared with MG1655. Outer circle shows the distribution of islands: shared co-linear backbone (blue); position of EDL933-specific sequences (O-islands) (red); MG1655-specific sequences (K-islands) (green); O-islands and K-islands at the same locations in the backbone (tan; hypervariable (purple). Second circle shows the G+C content calculated for each gene longer than 100 amino acids, plotted around the mean value for the whole genome, colour-coded like outer circle. Third circle shows the GC skew for third-codon position, calculated for each gene longer than 100 amino acids: positive values, lime; negative values, dark green. Fourth circle gives the scale in base pairs. Fifth circle shows the distribution of the highly skewed octamer Chi (GCTGGTGG), where bright blue and purple indicate the two DNA strands. The origin and terminus of replication, the chromosomal inversion and the locations of the sequence gaps are indicated. Figure created by Genvision from DNASTAR.

Leading *vs*. lagging strands at the origin of replication in *E. coli*

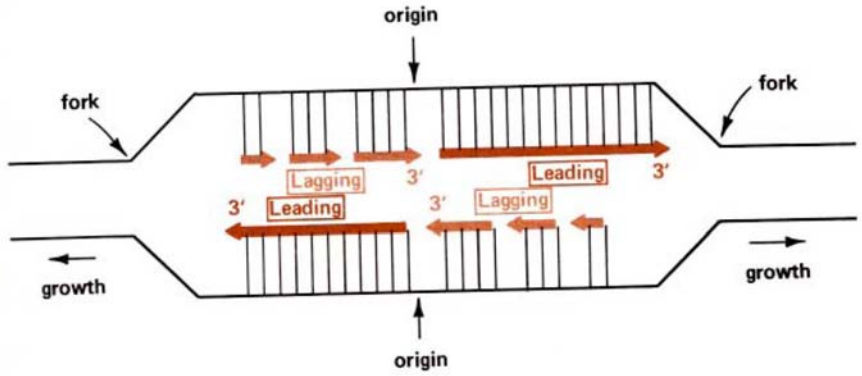
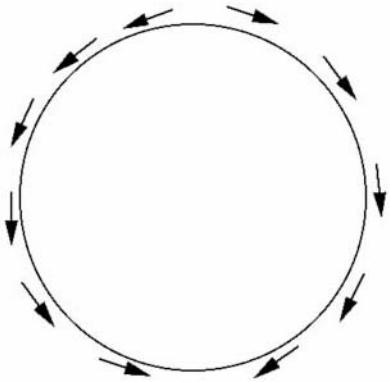


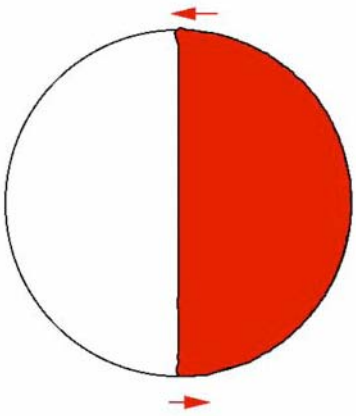
FIGURE 11-2
Bidirectional fork movement from the origin. Replication is semidiscontinuous: continuous on the leading strand and discontinuous on the lagging strand.

From "DNA Replication" by A. Kornberg (1980)

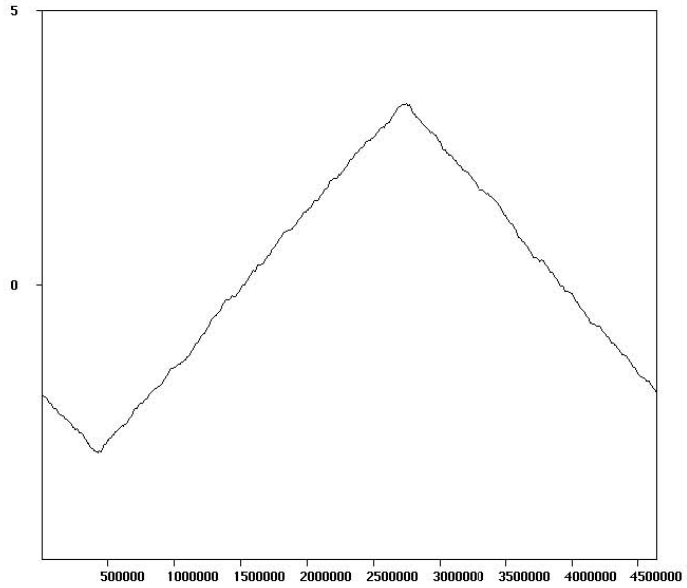
Bacterial strand assymetries are oriented with respect to a single origin of replication,



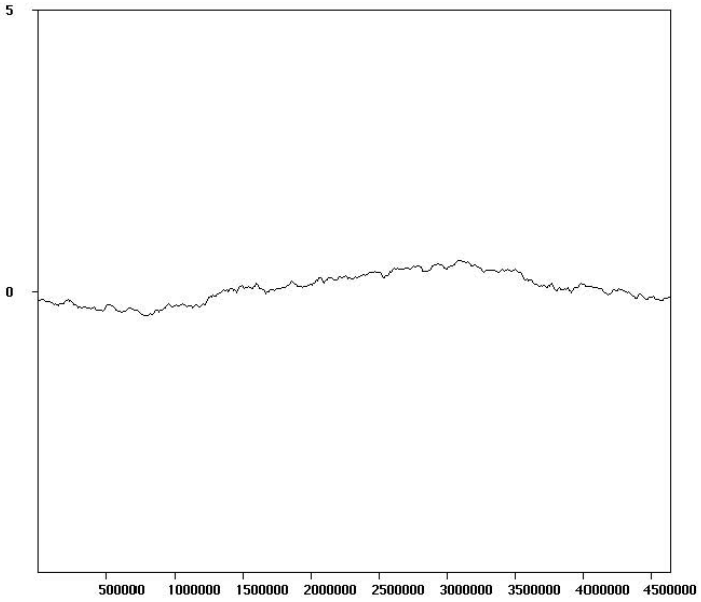
and are therefore best discovered by rotating half-genome length windows around a circular DNA sequence.



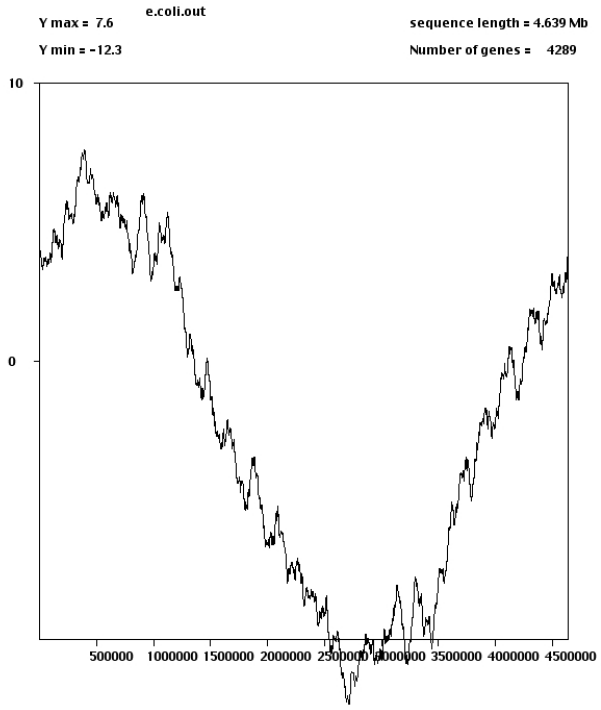
CG skew in *E. coli*



AT skew in *E. coli*



Bias in gene orientation ("transcription skew"), in the *E. coli* genome

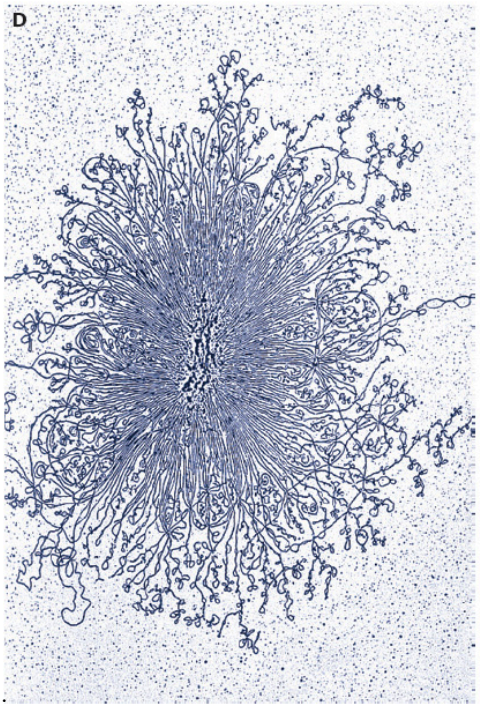


Supercoiling of bacterial chromosomes

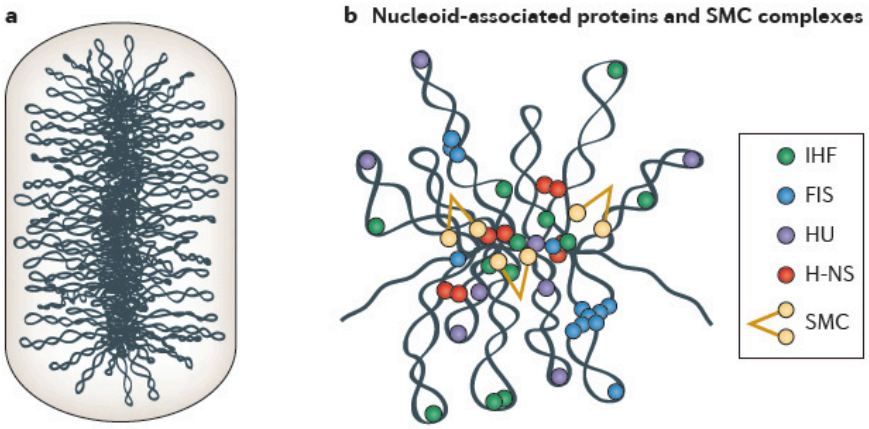
- Bacterial DNA is generally maintained in a negatively-supercoiled conformation.
- The supercoiling of a plasmid can be relaxed by one single-strand break (nick) (how?).
- However, relaxation of a bacterial chromosome requires numerous nicks, suggesting that the chromosomal DNA is organized into domains that are topologically insulated from each other.
- These domains average about 10 kb in length, in other words a few genes or operons.
- Normal maintenance of supercoiling is governed by DNA gyrase, which introduces negative supercoils, and topoisomerase I, which relaxes them.

Bacterial chromosomes are organized into loops *in vivo*.

Bacillus subtilis nucleoid. stained with Giemsa, using acid-treated cells.



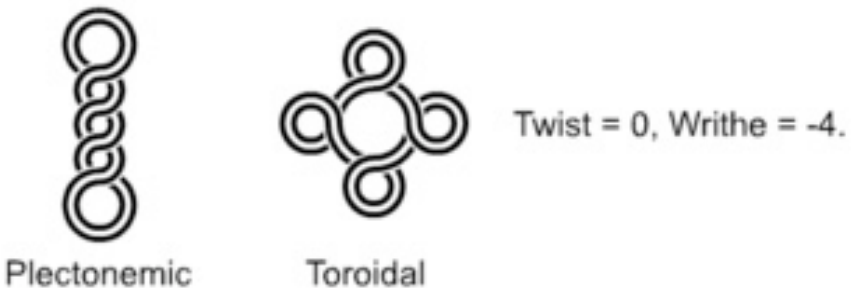
Wang et al. (2013) *Nat. Rev. Genet*. 14, 191-203.

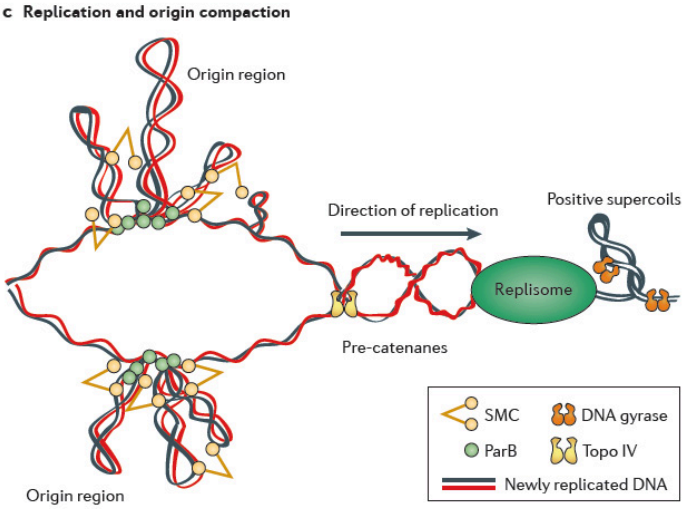


a | Schematic representation of the bottlebrush model of the nucleoid. The topologically isolated domains are on average 10 kb and therefore are likely to encompass several branched plectonemic loops. **b** | Schematic representation of the small nucleoid-associated proteins and the structural maintenance of chromosome (SMC) complexes. These proteins introduce DNA bends and also function in bridging chromosomal loci.

Wang et al. (2013) *Nat. Rev. Genet*. 14, 191-203.

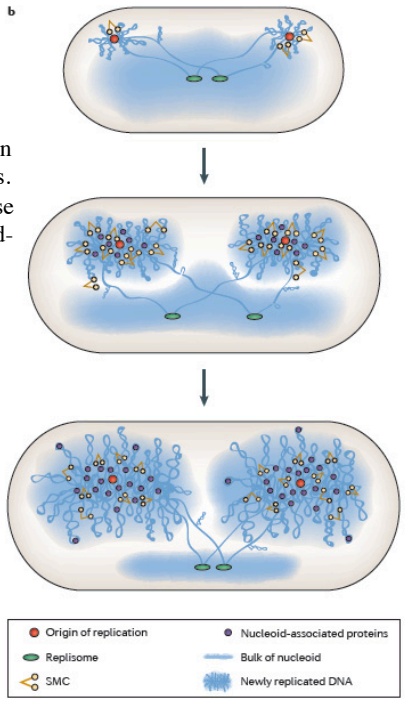
Illustration of a "plectonemic" supercoil (left)





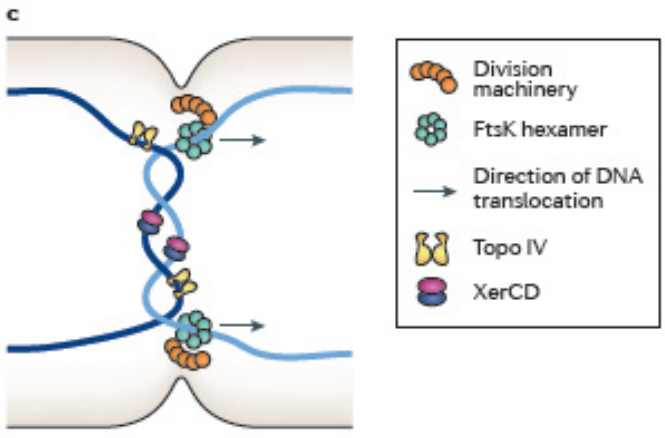
c | The diagram depicts replication fork progression and compaction of the origin region. Replication generates positive supercoils ahead of the fork, which can diffuse behind the replisome, producing pre-catenanes. Positive supercoils are removed by DNA gyrase and topoisomerase IV (Topo IV), and pre-catenanes are unlinked by Topo IV. Newly replicated origin regions are compacted by the SMC complexes that are recruited to the origin by ParB.

Wang et al. (2013) *Nat. Rev. Genet.* 14, 191-203.



b | A schematic model of bulk chromosome segregation that may be conserved among various bacterial species. After newly replicated origins are separated, lengthwise condensation mediated by supercoiling, small nucleoid-associated proteins and the structural maintenance of chromosome (SMC) complexes, in lockstep with replication, drive disentanglement and segregation of the sister chromosomes.

Wang et al. (2013) *Nat. Rev. Genet.* 14, 191-203.



c | A schematic of terminus segregation in *Escherichia coli*. The replicated terminus is translocated to appropriate daughter cell by the FtsK DNA translocase, whereas topoisomerase IV (Topo IV) and XerCD resolve catenanes and chromosome dimers, respectively. FtsK specifically localizes at the division septum, where it participates in cytokinesis and DNA segregation

Wang et al. (2013) *Nat. Rev. Genet.* 14, 191-203.

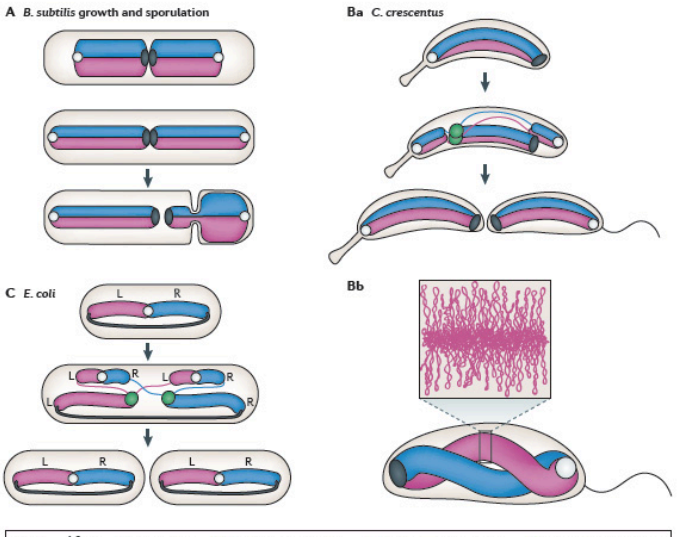


Figure 3 | **Spatial organization of bacterial chromosomes.** Cellular organization of the chromosome in *Bacillus subtilis* during growth and spore formation (**A**); slow-growing *Caulobacter crescentus* (**B**) and *Escherichia coli* (**C**) during vegetative cell cycles. During sporulation in *B. subtilis*, an asymmetric division traps ~25% of the chromosome in the smaller spore compartment[77,139]. A DNA translocase (not shown) pumps the remaining 75% of the chromosome into the spore after cytokinesis[140]. **Bb** shows a model for the spatially separate but gently twisted arms of the *C. crescentus* chromosome on the basis of chromosome conformation capture[81]. The inset shows the plectonemic loops within one of the arms. This figure is adapted, with permission, from REF. 157 © (2012) US National Academy of Science.

Discussion questions

- Discuss the role of chromosome structure in organizing DNA replication and chromosome segregation in bacteria. Your answer should include the replication of plectonemic loops, and the orderly segregation of DNA replication origins and termination points.
- Briefly discuss the role of lateral gene transfer in the evolution of bacteria and pathogenicity. Where do these genes come from - the same species? The same genus? Why or why not? Are these transfer events frequent or rare? How (if at all) does this affect the classification of phylogenetic relationships between bacterial species?
- Briefly discuss the structural characteristics of gene islands, pathogenicity islands, and mobile genetic elements in bacteria. Define each of these terms. How are they identified in bacterial genomes? Under what circumstances would they be expected to have unusual base composition at the first codon position? Third codon position? Codon bias?