Annu. Rev. Genet. 2004. 38:771–91
doi: 10.1146/annurev.genet.38.072902.094318

First published online as a Review in Advance on August 13, 2004

COMPARATIVE GENOMIC STRUCTURE OF PROKARYOTES

Stephen D. Bentley and Julian Parkhill
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom; email: sdb@sanger.ac.uk; parkhill@sanger.ac.uk

Key Words bacteria, genome, structure, rearrangement, evolution

■ **Abstract** Recent advances in DNA-sequencing technologies have made available an enormous resource of data for the study of bacterial genomes. The broad sample of complete genomes currently available allows us to look at variation in the gross features and characteristics of genomes while the detail of the sequences reveal some of the mechanisms by which these genomes evolve. This review aims to describe bacterial genome structures according to current knowledge and proposed hypotheses. We also describe examples where mechanisms of genome evolution have acted in the adaptation of bacterial species to particular niches.

CONTENTS

INTRODUCTION

The 1998 edition of the *Annual Review of Genetics* included an outstanding review by Sherwood Casjens entitled "The Diverse and Dynamic Structure of Bacterial Genomes" (13). This current review aims to update some of the ideas covered by Casjens. Indeed, it is fascinating to see just how rapidly this field has advanced in such a short time, largely due to the surge in whole genome sequencing, and the

associated wealth of new data, over the period. By the end of 1998, 17 complete bacterial genome sequences were available. By the end of 2003, that figure had risen to 149 [for a comprehensive record of genome sequencing projects, both complete and in progress, see http://www.genomesonline.org/ (10)].

Prior to the advent of whole genome sequencing, our knowledge of prokaryotic genome structure was largely limited to what we could determine from polyacrylamide gel electrophoresis and classical genetic maps (9). From the early 1970s through to the twenty-first century, rapid advances in the technology associated with DNA sequencing meant that obtaining the complete DNA sequence for a bacterial genome went from being very challenging to relatively simple, at least for certain genomes. By the end of 2003, the public sequence databases held the complete sequence of 99 bacterial and 12 archaeal genomes covering 69 and 12 different species, respectively. This represents an unprecedented dataset with which to investigate prokaryotic genome structure. Studying each genome in isolation has brought novel insights, and this new knowledge has often been further enriched by comparison of genomes from related strains. For the study of genome structure, the importance of having complete and accurate genome sequences from related bacteria cannot be overstated. With such information we can observe rearrangement and evolution of genomes from the multi-mega base-pair level down to single-nucleotide resolution. This review aims to convey the major themes in prokaryotic genome structure, focusing for the most part on advances in the past decade.

GENOME STRUCTURE

The obvious parameters for describing a bacterial genome are size, geometry, replicon number, and G + C content; variation in these alone ensures that there is no such thing as a typical bacterial genome. Actually defining a bacterial genome is complicated by the variety of DNA replicons found within bacterial cells. For this review the genome includes only those replicons that can be considered to be "chromosomes." This allows us to disregard transient replicons such as plasmids but the definition of "chromosome" is in itself controversial (see below).

Genome Size

Prokaryotic genome sizes vary across more than a twentyfold range. Figure 1*A* shows how this size range lies relative to all forms of life, overlapping with the smallest eukaryotes and the largest viruses. It is now clear that archaeal genomes, previously thought to cover a relatively narrow range (13), vary to a similar degree as those of eubacteria. Within the prokaryote group, different phyla cover broadly overlapping size ranges (Figure 1*B*). Even within species large-scale variation can be seen; genomes of *Escherichia coli*, *Prochlorococcus marinus*, and *Streptomyces coelicolor* have all been seen to vary by more than 1,000,000 bp (46, 56, 75). Bacterial genome size is the sum of different genetic events, such as gene duplication,

horizontal acquisition, and lineage-specific gene loss and is therefore not a good indicator of evolutionary lineage.

Although distantly related, the proteobacteria and actinobacteria both appear to reach the extremes for bacteria, suggesting that the factors limiting genome size may be common to all. The exact nature of these limiting factors is unclear. The "minimal genome" concept is attracting great interest, with the seductive aim of defining the minimal set of genes required for cellular life, though the minimal set will of course vary depending upon the niche (33, 35). Definition of a "maximal genome" is perhaps less intuitive. Whereas a minimal genome is limited by the ability to live, the maximal boundaries are more difficult to draw. Complex issues relating to generation time, replication rate, and energy supply need to be considered as well as simple practical issues such as the physical space taken up by the genome within the cell sacculus.

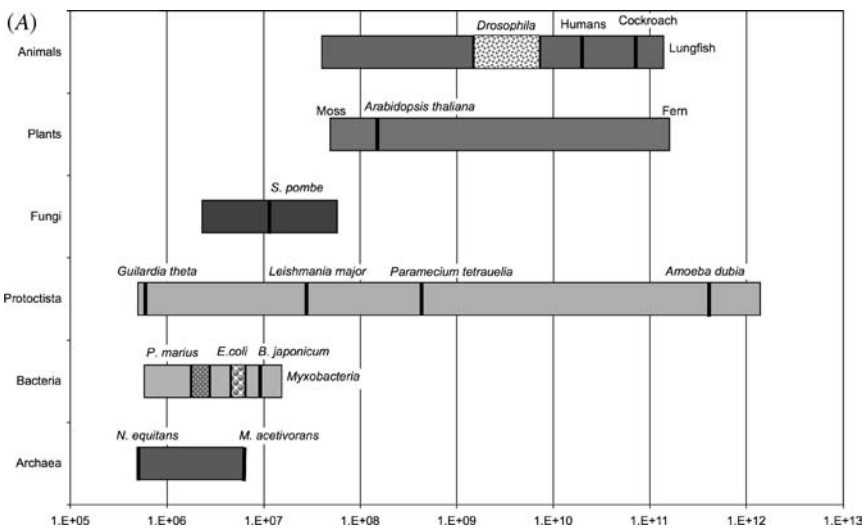

Figure 1 (*A*) Estimates of genome sizes, based on data from DOGS (Database of Genome Sizes; http://www.cbs.dtu.dk/databases/DOGS/) (72). A selection of genome sizes and size ranges from specific species are indicated. For simplicity this figure includes one line for the kingdom Protoctista, which is defined as nucleated microorganisms and their descendants, exclusive of fungi, animals, and plants. (*B*) Size ranges of sequenced prokaryotic genomes. (*C*) GC content ranges of sequenced prokaryotic genomes. For (*B*) and (*C*) the data are shown as "box and whiskers" plots, where the shaded box represents the middle 50% of the observations (interquartile range) and the vertical line in each box is the median value. The "whiskers" represent plus or minus 1.5 times the interquartile range. Outlier points are designated with circles. *n* is the number of sequenced genomes for the given phylum or group. (*D*) GC content versus genome length for 146 sequenced prokaryotic genomes. Figures supplied by D.W. Ussery & P.F. Hallin (71, 72).

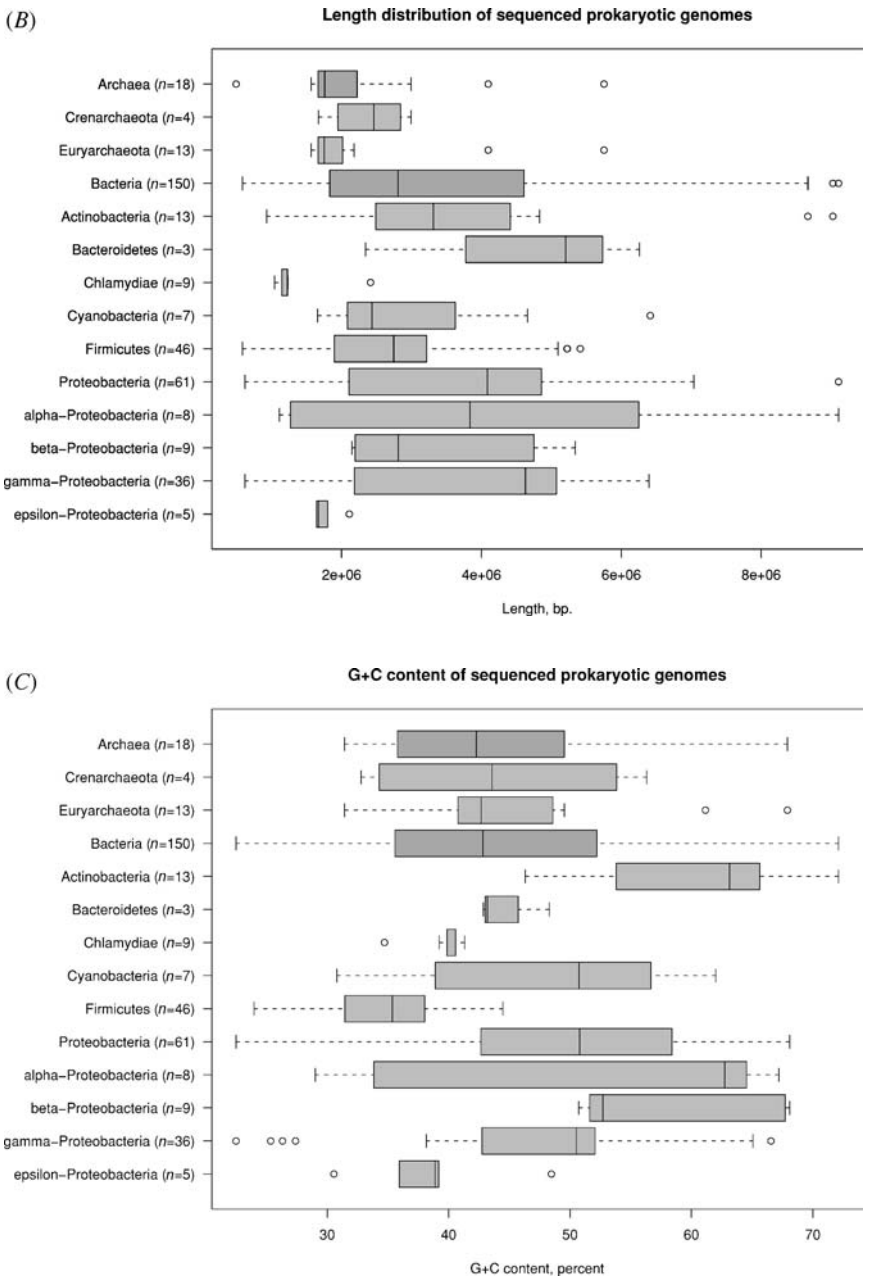

Figure 1 (*Continued*)

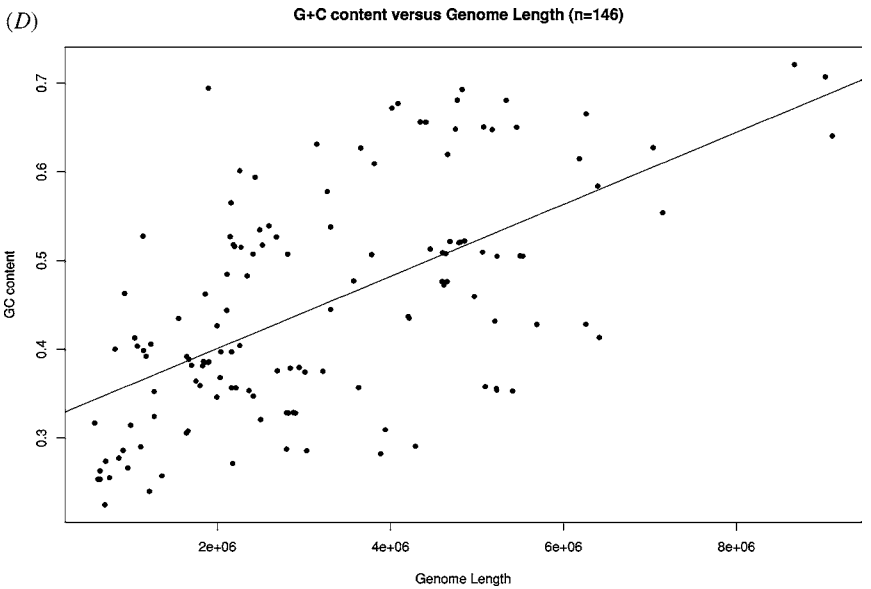

Figure 1 (*Continued*)

The smallest prokaryote genomes tend to belong to organisms restricted to a stable niche, often in association with a host organism, whereas the bacteria with genomes at the larger end of the scale tend to occupy highly complex and variable environments such as soil. At the time of writing, the sizes of sequenced prokaryotic genomes ranges from 490,885 bp for *Nanoarchaeum equitans*, an obligate symbiont growing only in coculture with crenarchaeon *Ignicoccus* (74), to 9,1058,28 bp for *Bradyrhizobium japonicum*, a metabolically adept soil-dweller capable of colonizing plant root nodules (31).

Generally the coding density of bacterial genomes does not vary much, with most approximating one gene per kilobase of DNA. Clearly then, genome size is directly proportional to number of genes, a trend not true for eukaryotes. An increase in gene numbers allows for two possibilities: (*a*) an increase in the number of protein families encoded and (*b*) an increase in the number of encoded members of each protein family. Comparison of the predicted proteome of *S. coelicolor* with those of a range bacteria showed that both of these occur (7). Such a broadened repertoire of genes presents a management problem. Cells containing larger genomes are still limited by energy supply, which in turn limits gene expression to levels equivalent to those for cells with smaller genomes. Accordingly, larger genomes require more complex regulation of gene expression via an increased numbers of genes encoding regulatory proteins. Stover et al. (63) first noted that, rather than being directly proportional, the percentage of regulatory genes in a genome appeared to increase with increasing genome size. This observation was

based on comparison of the handful of genomes available at the time but was still gaining support years later (7, 12). Two recent comprehensive studies have expanded on this type of observation.

Ranea et al. (54) looked at the distribution of 816 structural protein superfamilies across 56 prokaryotic genomes including 10 archaea. Each of the superfamilies were selected from the CATH structural protein database (47) as being present in at least 70% of the genomes and thus considered to be "universally" distributed. Correlation with genome size revealed two major groups; size dependent and size independent. Within the size-independent group was a subgroup of proteins evenly distributed across all genomes. These proteins are almost exclusively involved in translation, ribosome structure, and protein biogenesis. The size-dependent group could be subdivided into families with a linear distribution relative to genome size and those with a nonlinear relationship. The majority of the linear group was concerned with cellular metabolism while the nonlinear group mostly comprised overrepresented families involved in regulation of gene expression. Notably, none of the superfamilies in the nonlinear group was underrepresented.

Konstantinidis & Tiedje (34) used a fundamentally different approach but still arrived at compatible major conclusions. They used the COG database (67) to place the protein products of 115 bacterial genomes into functional categories and then looked to see how these groups correlate with genome size. Negative correlation with genome size was observed for proteins involved in translation, ribosome structure, protein biogenesis, DNA replication and repair, cell division, chromosome partitioning, and nucleotide transport and metabolism. This is equivalent to the size-independent group of Ranea et al. (54). Positive correlation with genome size was seen for transcription and its control, signal transduction, cell motility, secondary metabolism, and energy production and conversion. It was also noted that the proportion of noncoding DNA and genes encoding proteins of unknown function (hypotheticals) remains constant. These data fit well with the idea that genome size and content are largely dictated by environmental pressures.

Genome Geometry and Replicon Arithmetic

By far the majority of bacterial genomes exist as a single circular chromosome. Indeed, only relatively recently were deviations from this model found to exist in the form of bacteria with linear and/or multiple replicons. So far, linear chromosomes have been found in *S. coelicolor* (32), *Borrelia burgdorferi* (19), *Agrobacterium tumefaciens* (2), and related species. Streptomycetes have a single linear chromosome whereas agrobacterial and borrelial genomes comprise a mixture of linear and circular replicons. Linear plasmids have been isolated from these three groups as well as from others where only circular chromosomes have been found (2, 6, 55, 79). Phylogenetic distance alone strongly suggests that linear chromosomes arose independently from a circular progenitor in the three taxonomic groups. This is further supported by their differing mechanisms for maintenance of the chromosome ends or telomeres.

In *Borrelia*, the telomeres exist as closed hairpins formed by a process known as telomere resolution (69). This process is part of the general DNA replication mechanism for bacterial and phage replicons with covalently closed hairpin telomeres. Bidirectional replication from an internal origin forms a circular head-to-head, tail-to-tail dimer with the two DNA monomers covalently linked at the telomere. The circular dimer is processed by a DNA breakage and reunion reaction generating daughter replicons with hairpin telomeres. In *Borrelia*, this reaction is catalyzed by ResT, an enzyme with similarity to recombinases and topoisomerases. The telomeres of *A. tumefaciens* are also thought to have a covalently closed structure (23), although the mechanism of maintenance has not been confirmed.

The replication of *Streptomyces* linear plasmids also proceeds divergently from a central origin but instead of forming a circular molecule, a 3′-leading-strand overhang is generated at the telomere. This is followed by extension of the recessed 5′ ends to produce full-length duplex DNA molecules (14). As *Streptomyces* linear chromosomes and linear plasmids have similar termini (29, 53), and the linear chromosomes also replicate bidirectionally from an internal origin of replication (44), linear chromosome telomeres are presumed to employ the same mechanism. Both linear plasmids and linear chromosomes of *Streptomyces* have neighboring genes known as *tpg* (terminal protein gene) and *tap* (terminal associated protein). Both genes are essential for replication of *Streptomyces* chromosomes and plasmids in a linear form (3, 80). Tpg protein is covalently attached to the 5′ DNA ends but there is no evidence that it may function as a primer for DNA synthesis (3, 80). Tap protein recruits Tpg to the telomere termini by interacting with both Tpg and specific sequences on the 3′ overhang of telomeric DNA (4).

Although the evolutionary triggers and advantages for linear chromosomes are open to speculation, reversion to a circular form has been observed in both *Borrelia* (20) and *Streptomyces* (38).

Physical mapping techniques and whole genome sequencing have revealed several cases where bacterial genomes have more than one large replicon (Table 1). Although it is not surprising to find multiple replicons, a situation where the second largest replicon approaches the size of the largest replicon creates a problem of classification. Has the chromosome split to form a secondary chromosome or has a smaller plasmid accumulated extra DNA to form a megaplasmid? Often the designation has been influenced by historical precedent but the availability of complete genomes sequences in a growing number of cases has allowed a more detailed and objective assessment. One simple test may be whether the bacterium can grow without the second replicon. The replicon-deleted strain can be difficult to create but the presence of a known essential gene on the second replicon is good evidence of essentiality, particularly if it is a unique copy. Table 1 shows that simply looking at distribution of ribosomal RNA operons can be a convenient test. On this criterion, the designation of most replicons is straightforward and correlates well with the distribution of the overwhelming majority of housekeeping functions to the chromosomes. However, the fluid nature of bacterial genomes means that there

TABLE 1

Species	Appellation	Size (kb)	Shape	rDNA no.
Streptomyces coelicolor	Chromosome	8667	Linear	6
	Plasmid	356	Linear	0
	Plasmid	31	Circular	0
Agrobacterium tumefaciens	Chromosome	2842	Circular	2
	Chromosome	2057	Linear	2
	Plasmid	543	Circular	0
	Plasmid	214	Circular	0
Borellia burgdorferi	Chromosome	911	Linear	1
	Plasmid (n = 11)	9–54	Circular/Linear	0
Brucella melitensis	Chromosome	2117	Circular	2
	Chromosome	1178	Circular	1
Clostridium acetobutylicum	Chromosome	3941	Circular	11
	Megaplasmid	192	Circular	0
Deinococcus radiodurans	Chromosome	2649	Circular	3
	Chromosome	412	Circular	0
	Megaplasmid	177	Circular	0
	Plasmid	46	Circular	0
Ralstonia solanacearum	Chromosome	3716	Circular	3
	Megaplasmid	2095	Circular	1
Salmonella typhi	Chromosome	4809	Circular	7
	Plasmid	218	Circular	0
	Plasmid	107	Circular	0
Sinorhizobium meliloti	Chromosome	3654	Circular	3
	Megaplasmid	1683	Circular	0
	Megaplasmid	1354	Circular	0
Vibrio cholerae	Chromosome	2941	Circular	8
	Chromosome	1072	Circular	0
Yersinia pestis	Chromosome	4654	Circular	6
	Plasmid (n = 3)	10–96	Circular	0

From Ochman (45).

will inevitably be exceptions to such a simple test; *Deinococcus radiodurans* and *Vibrio cholerae* both have designated chromosomes that lack rDNA.

Vibrio cholerae is an interesting case (26). The larger of its two chromosomes (chromsome 1) contains most of the genes essential for growth and pathogenicity, whereas the smaller chromosome (chromosome 2) contains the majority of the genes for DNA repair and damage response. However, all eight ribosomal RNA operons and most tRNA genes are on chromosome 1, with a few redundant copies

on chromosome 2. Regardless of the essentiality of the replicon, the origin of genes for functions such as replication, partitioning, and conjugal transfer can be good indicators of the origin of the replicon. *V. cholerae* chromosome 2 contains an integron island, typically found on plasmids, that contains a gene capture system and host addiction genes. The gene capture system seems to have facilitated the acquisition of several coding sequences for potential virulence and drug-resistance proteins. Analysis of the *parA* locus on each chromosome is also revealing. The ParA protein is essential for partition and distribution of replicons to daughter cells during cell division. The chromosome 1-encoded ParA is most related to other chromosomally encoded ParAs, whereas chromosome 2-encoded ParA is most related to those encoded by plasmids and plasmid prophages. *Agrobacterium tumefaciens* shows similar inconsistencies with its linear chromosome displaying several features reminiscent of plasmids, including conjugative proteins and a RepABC-type replication system (78).

Ralstonia solanacearum is another exceptional case, with the second largest replicon designated as a megaplasmid despite harboring a ribosomal RNA operon (61). Also, although the chromosome encodes many housekeeping functions, the megaplasmid carries several important genes for which there are no counterparts on the chromosome and for which deletion results in auxotrophy. The megaplasmid also carries genes for flagella and exopolysaccharide production that, although not essential for the survival of the bacterium, are fundamental to its lifestyle and identity.

The complete genome of *Burkholderia pseudomallei*, a close relative of *R. solanacearum*, has recently been sequenced revealing a similar replicon organization (27). The complete genome of *B. pseudomallei* strain K96243 consists of two circular replicons that have been designated chromosome 1 and chromosome 2. Chromosome 1 encodes the typical prokaryotic chromosomal replication machinery while the chromosome 2 replication mechanism appears to be of plasmid origin. However, the authors designated the second replicon as a chromosome because of the presence of several likely essential genes, most notably a unique seryl tRNA (Ser^{GGA}) necessary for translation of genes on both chromosomes. Alignment of replicons from *B. pseudomallei* and *R. solanacearum* reveals extensive synteny (conservation of DNA sequence and gene order) between the two largest replicons but almost none between the secondary replicons (27). This implies either that the emergence of the second replicon occurred independently in each lineage or that the second replicon has diverged at a faster rate than the primary replicon (see below).

Where there are two chromosomes in a genome it may seem sensible, from a practical viewpoint, for them to merge to form a single chromosome. For *V. cholerae*, Heidelberg et al. (26) describe two possible explanations for why this does not occur: (*a*) differential replicon copy numbers may be required for specific gene expression levels in certain conditions; (*b*) aberrant segregation may be desirable such that "drone" daughter cells may be produced with only a single chromosome. Such drone cells may represent the "viable but non culturable state"

and have been implicated in biofilm development. Whatever the case, the distinction between chromosome and megaplasmid is unlikely to be clarified without the addition of many more multireplicon genome sequences to the analysis.

BASE COMPOSITION

It was noted as early as 1962 that "Among bacteria. . .the mean GC. . .content of DNA varies approximately from 25 to 75%, and this range extends over the range of the mean GC content of DNA of higher organisms" (64). The current GC content range for sequenced bacterial genomes is from 72.1% [*S. coelicolor* (7)] to 26.5 [*Wigglesworthia glossinidia* (1)]. Figure 1*C* shows that although the bacterial phyla occupy differing ranges of GC content, the actinobacteria tend to be GC-rich, whereas the firmicutes are mostly AT-rich. The epsilon group of proteobacteria stand out as AT-rich compared with the other proteobacterial subgroups. There is a clear correlation between genome size and GC content (Figure 1*D*) (25, 42, 58), with the large genome soil-dwellers tending to have GC-rich genomes and the reduced genome host obligates tending toward AT-rich genomes. The mean genome GC content for free-living bacteria is around 49% and the equivalent figure for host obligates is 38% (58). Again, looking within the different phyla shows further compelling trends. The small AT-rich genomes belong to species such as *W. glossinidia* [proteobacteria; 0.7Mb; 22% G + C (1)], *Mycoplasma mycoides* [firmicute; 1.2 Mb; 24% G + C (77)] and *Tropheryma whipplei* [actinobacteria; 0.9 Mb, 46% G + C (8)], whereas the GC-rich large genomes belong to *B. japonicum* (proteobacteria; 9.1 Mb; 60.1 G + C), *Bacillus cereus* (firmicute; 5.4 Mb; 35.3% G + C), and *S. coelicolor* (actinobacteria; 8.7 Mb; 72.1% G + C).

Although the correlation between genome size and niche complexity seems logical, explanations for the linear relationship with GC content are more equivocal. One consideration is the fact that GTP and CTP nucleotides are more energetically expensive than ATP and UTP (58). Also, the central role of ATP in energy metabolism leads to a greater availability relative to other nucleotides. In circumstances where resources are limited, these factors could drive the drift toward AT-richness. Rocha & Danchin (58) suggest that these factors may result in a selective pressure toward AT-richness in the small genome organisms and extend this hypothesis to explain that the comparatively high AT content of plasmids, phages, and insertion elements may also be due to differential cost and availability of the relevant metabolites in the cell. However, this proposal does not take into account the fact that host obligate bacteria vary in their external supply of nucleotides as well as in their ability to synthesize them. Another factor likely to be relevant is DNA repair. Every reduced genome appears to have lost some genes involved in DNA recombination and repair pathways, though the precise set discarded varies (42). These genes are important for bacteria generally but clearly not essential in such circumstances. Their absence allows for unchecked point mutations to accumulate. Experiments have shown that the most frequent random mutation occurring in cells is C to T (or G to A), due to the deamination of Cytosine to form

Uracil, which is subsequently replicated as Thymidine (22). Thus in the absence of DNA repair, genomes tend to become more AT-rich.

Figure 1*C* represents the mean values for whole genomes, but GC content across chromosomes is far from uniform and shows some interesting trends. For reduced-size genomes the drift toward AT is more pronounced in intergenic, noncoding sequences and in the codon third base position in coding sequences. For all genomes, upstream promoter sequences tend to be AT-rich in order to allow a curved, rigid conformation that unwinds more easily (52). For similar reasons, the region a few hundred bases around the origin of replication is AT-rich; however, this does not explain the decrease in GC content in intergenic regions that do not contain a promoter. On a whole genome scale there appears to be a slight preference for GC toward the origin and AT toward the terminus (71), and this may be due to structural constraints on the genome, or to a physical or functional partitioning of the chromosome. Due to codon redundancy, AT bias at the codon third base position is less likely to change the encoded amino acid; however, AT drift is random so first and second codon positions are inevitably affected, unless reversed by selection. These changes cause a drift in amino acid usage and a concomitant increase in average isoelectric point for the proteome (62) for AT-rich organisms. Obviously, AT-biased substitutions also have the potential to reduce or ablate gene function. The highest level of AT bias in *Ureaplasma urealyticum* occurs in genes with orthologues shown to be nonessential in the close relative *Mycoplasma genitalium*, reflecting the opposing pressures of AT bias and preservation of gene function (22). A similar observation in the highly degraded genome of *Mycobacterium leprae* showed that the nonfunctional pseudogenes had a lower G + C content overall than the functional genes (15).

The GC equilibrium of each genome is presumably imposed and constrained by a whole host of functional and environmental influences. Accordingly, foreign DNA incorporated into a genome may have a different GC composition. Over time, such DNA is subjected to a process of amelioration where directional mutation pressures act to alter the base composition of the incoming DNA to match that of the whole genome. Modeling of this phenomenon has been used to predict the rate of amelioration for a particular genome, which can then be used to calculate the time since a foreign region of DNA was first integrated (36, 37).

A curious base composition feature of bacterial genomes is the measure of G-C/G + C for leading and lagging strand, commonly referred to as GC skew. The ubiquitous pattern in prokaryote genomes is for a bias toward G over C on the leading strand of DNA resulting in a biphasic pattern across the genome, which is particularly useful in locating the origin and terminus of replication (Figure 2) (39). The mechanism responsible for creating this pattern remains elusive but it seems likely that some strand-specific mutation frequency due to replication asymmetry is crucial. Preferential gene order has been shown not to be a major influencing factor but it has been observed that, as for AT bias, the GC skew is more pronounced in "neutral" DNA such as the third codon position of coding sequences (41). Studies in *Escherichia coli* indicate that functional DNA motifs may be relevant. The distribution of Chi and Rag motifs, both G-rich, is more skewed to the leading strand

than would be predicted from the GC skew itself. Chi sequences are recognized by the RecBCD complex and may serve to facilitate the restarting of aborted replication forks (24). Rag motifs may be involved in resolution of chromosome dimers and clearing of the closing septum. Together these two motifs can only account for up to 14% of the total GC skew, but other processes may possibly rely on such polarity. GC skew is also subject to amelioration such that recent chromosomal rearrangements or insertions can be detected due to aberrant GC skew patterns (50).

GENE ORIENTATION

Genes tend to be oriented in the direction of replication, transcribing away from the origin of replication. The strength of the bias varies from around 52% to 83% and tends to be most pronounced in the low GC firmicutes (57). The more extremely biased genomes tend to encode a PolC orthologue, a subunit of the replication complex thought to act asymmetrically (57). In a study of 59 bacterial genomes, those containing a PolC orthologue carried 78% of their genes on the leading strand whereas those lacking a *polC* gene had a 58% strand bias. Importantly, the presence of *polC* does not correlate with GC-skew.

Clear gene strand bias in bacteria that do not encode PolC indicates that other factors are involved. One popular theory suggests that this arrangement has evolved in order to minimize the number of collisions between replication and transcription complexes as they move along the DNA (11, 21). Such collisions cause replication to stall and can cause transcription to either pause or abort (21). This in turn would be influenced by the level of expression such that highly expressed genes tend to be on the leading strand. This fits well with the systematic location of rDNA and ribosomal protein operons on the leading strand (41). However, a positive correlation between expression and strand bias suggests that the bias would be stronger for faster-growing bacteria but this does not seem to be the case. For example, the slow-growing *Mycobacterium tuberculosis* (59%) is more biased than the fast-growing *E. coli* (55%). More recently, this has been refined in a study that concludes that essentiality rather than expression level drives gene strand bias, with the majority of essential genes across a cross-section of bacterial genomes shown to be orientated away from the origin of replication. It was postulated that this may be due to the toxicity of truncated translation products of essential genes (59).

SYNTENY: CONSERVATION OF GENE ORDER

The word synteny was originally used to describe the occurrence of gene orthologues on equivalent chromosomes in two different eukaryotic genomes (51). Over recent years the usage has been adapted by genomicists and is now widely accepted as referring to multigene regions where the DNA sequence and gene order are conserved between genomes. The detailed genetic maps of *E. coli* and *Bacillus subtilis* indicated that genes did not necessarily occur at the same

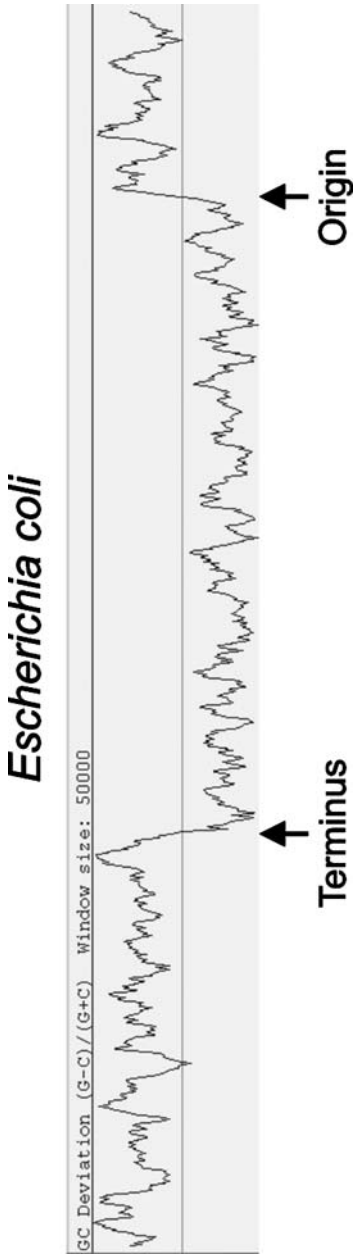

Figure 2 GC deviation (or GC skew) across the whole genome of *Escherichia coli* K12. This simple measure of base composition shows a distinct correlation with chromosome replication.

relative position in all bacterial genomes, but it was known that certain gene clusters were syntenic. Genome sequencing allowed a more detailed assessment, and it was soon concluded that although there seems to be a positive selection for clustering of physically interacting proteins, there is no absolute requirement for juxtaposition of any genes in a bacterial genome and synteny is lost at a much faster rate than sequence similarity (43). Nevertheless, synteny remains a useful indicator in assessment of genome evolution (65) and prediction of gene function (16). Gene clusters that appear resistant to dispersal include the ribosomal protein operon, the *nuo* operon (NADH dehydrogense), and the dcw cluster (66).

The genomes of close relatives show extensive synteny and reveal a striking feature of bacterial genome evolution. Using dot plots to represent similarity between aligned genomes, it has been shown that large-scale symmetrical inversions centered on the origin of replication are common in bacteria (18, 68). In such plots collinear genomes are represented as an unbroken diagonal. A single symmetrical inversion around the origin results in a counterdiagonal with multiple inversions appearing as a broken X pattern (18, 27). Using this method, synteny can be easily detected in between genomes as phylogenetically distant as *S. coelicolor* and *M. tuberculosis* (7). The prevalence of these inversions over other types of rearrangements is intriguing and has provoked several plausible explanations. An unresolved question is whether such rearrangements are more likely to occur per se, or whether they are more readily fixed by selection, with other rearrangements being more deleterious (60). One factor may be that the inversion does not disturb the orientation or distance of a gene relative to the origin. Since, under certain circumstances, multiple replication forks may be in operation, distance from the origin would have a gene-dosage effect (40). Furthermore, such reciprocal inversions do not disturb the equality of replichore lengths, and the differential mutational pressures on the leading and lagging strands (see above) may have a detrimental effect on genes whose strand is switched by other types of inversion (40). A link with replication seems likely with the unwound and unpackaged DNA within replication forks serving as hotspots for reciprocal recombination (17, 68). *Mycoplasma* and *Chlamydia* genomes have a slower rate of inversions relative to phylogenetic distance, possibly due to absence of proteins from their replication machinery that are involved in recombination (65).

The apparently gradual reduction of synteny due to reciprocal inversions and other rearrangements implies that conserved gene order may be useful as a phylogenetic measure for studying the relationship between genomes, but other factors need to be taken into account such as the potentially catastrophic effect on synteny of insertion element expansion (49).

EXPANDING AND CONTRACTING GENOMES

The classical view of evolution says that change occurs slowly via the accumulation of point mutations in particular genes. For bacteria the process is much more dynamic, with various mechanisms allowing the rapid gain, loss, and rearrangement

of substantial portions of the genome. This state of flux has allowed prokaryotes to evolve rapidly in response to environmental pressures and is the main reason why they exist and persist in almost every niche on the planet. Comparative genome analysis has brought enormous insights into the fluid evolution of bacterial genomes. The scenarios for genome evolution are as numerous as the number of possible niches, and here we consider a selection of interesting examples arbitrarily classified as expanding and contracting.

Expanding Genomes

As described in the section on genome size, the largest bacterial genomes tend to belong to bacteria that dominate in complex environments such as the soil or rhizosphere. In these cases, large numbers of genes have been acquired that may only be occasionally advantageous. Unlike host obligate or host-associated bacteria, the soil-dwellers must be able to persist and prosper in a broad range of physical, nutritional, and biological conditions, and it is the possession of a large inventory of genes that makes this possible. Another feature of this environment seems to be that there is little penalty for slow growth so replication time does not seem to be critical. One typical example is *S. coelicolor*, described as a "boy-scout" bacterium because it seems prepared for anything (7, 28). The *S. coelicolor* genome is rich in genes for degradation of complex carbohydrates, making it capable of exploiting the decaying matter from plants, animals, insects, and fungi as well as other bacteria. It also dedicates a large portion of the proteome to the production of specialist compounds, known as secondary metabolites, which can function as antibiotics, protect against desiccation and low temperature stress, and aid in scavenging of iron from the environment. Addition of the genes necessary for mycelial growth and sporulation results in a genome clearly expanded to enrich complexity. Several mechanisms for genome expansion are apparent. Twenty potentially laterally acquired regions have been detected in the genome, the largest of which spans 153 kb (148 genes). There is also likely to have been extensive gene duplication. The study by Ranea and colleagues (54) showed that gene duplication and genome size are strongly connected. Indeed, it is suggested that lineage-specific gene expansion is positively correlated with genome size and may account for up to 33% of coding capacity (30).

The most remarkable feature of the *S. coelicolor* genome is its spatial and functional compartmentalization. It has a single linear chromosome that can be roughly divided into three portions, a central core and two flanking arms. Most essential and housekeeping genes reside in the core while the arms contain most of the "occasionally useful" functions. The extreme ends are rich in insertion elements and pseudogenes, suggesting a regional tolerance to insertion events. A major factor in the expansion of the *S. coelicolor* genome from the ancestral chromosome has likely been the insertion of genes, from lateral acquisition and intragenomic duplication, into the chromosome arms. Strong evidence for this hypothesis comes from the observation that the core region shows extensive synteny with the entire

3.9-Mb genome of *M. tuberculosis*, but this synteny does not extend into the arms.

Spatial and functional compartmentalization seems to be a theme for larger genomes though general mechanisms vary. Other bacteria with large genomes such as *A. tumefaciens*, *R. solanacearum*, and *S. loti* have multiple replicons. Particularly interesting is the situation for the Burkholderiaceae. Each of these genomes has two replicons with a distinct partitioning of core and accessory functions to the larger and smaller replicon, respectively (27).

"Balanced" and Contracting Genomes

Reduction in genome size can be seen at many different levels in different genomes. Generally, overall reduction in genome size can be seen as an adaptation, or response, to a simplified or more stable environment. This adaptation should probably be differentiated from the loss of genes that occurs as a corollary to gene acquisition and niche change in free-living organisms. As originally postulated by Lawrence & Ochman (37), in these organisms gene acquisition is balanced in the longer term by gene loss, such that genome sizes tend to remain relatively stable within taxa. The genomes of enteric pathogens such as *E. coli* and *Salmonella* show very clear evidence of this. There is ample evidence for large recent insertions of self-mobile DNA, termed pathogenicity islands, and for the presence in some strains of smaller islands and single genes not present in others (17, 48, 50). Despite this, the overall size of the genomes within the Enteric group remains reasonably similar, at around 4–5 Mb. Gene acquisition must therefore have been balanced by gene loss over evolutionary timescales.

The consequent balancing gene loss can be seen most easily within the genomes of organisms that have recently changed niche, and is clearest in human pathogens such as *Salmonella enterica* serovar Typhi (48), *Shigella flexneri* (76), and *Yersinia pestis* (50). Such genomes often contain a larger number of pseudogenes than close relatives, around 5% or more of their total coding capacity. These pseudogenes have been interpreted as being due to accelerated genetic drift caused by the evolutionary bottlenecks that these organisms have gone through on changing niche. Many genes that have been inactivated in this way were likely to have been involved in adaptation to the old niche and are no longer required (or disadvantageous) in the new niche; specific examples include the flagellar apparatus and various adhesins of *Y. pestis* (50); these are required for survival as a gut pathogen but are unnecessary for a systemic pathogen. However, some genes may also have been inactivated by drift during a bottleneck, even though these mutations may be neutral or mildly deleterious. Many of these pseudogenes are caused by point mutations incorporating stop codons, or causing frameshifts; however, they can also be due to an expansion of Insertion Sequence (IS) elements. These selfish mobile elements can expand in number as a consequence of relaxed intraspecific competition during evolutionary bottlenecks, and are often associated with chromosomal rearrangement and gene loss by deletion (50), as well as inactivation of genes by insertion.

Acquisition of pseudogenes can also be seen as the first stage in a larger-scale genome reduction, and again this can be exemplified by human pathogens such as *Bordetella pertussis* (49). Like the organisms described above, *B. pertussis* also seems to have recently changed its niche, and, again like *S.* Typhi, it has become restricted to a single host. However, the scale of gene inactivation, loss, and IS element expansion in *B. pertussis* is considerably greater. Compared with *B. bronchiseptica*, which appears to be closely similar to its immediate ancestor, *B. pertussis* has lost over 20% of its chromosomal DNA by deletion, and nearly 10% of the genes that remain have been inactivated. Much of this inactivation has been mediated by IS element expansion; *B. pertussis* has nearly 240 copies of a single IS element, IS481, scattered around its genome, and when compared with *B. bronchiseptica*, there are nearly 150 chromosomal rearrangements. Given that the two species are 99.8% identical at the 16S rRNA level, this is an unprecedented level of recent genome decay. It is a moot point as to whether this is representative of balancing gene loss, or whether *B. pertussis* is on the route to a permanent reduction in genome size; there is almost no evidence of recent gene acquisition in this organism.

Other organisms that appear to be host dependant and display visible signs of gene decay include the Rickettsias, where the different stages of gene loss can be traced through different species or strains (33), and *Mycobacterium leprae*, where a massive loss of gene function has not yet been caught up with by DNA deletion; with a 3.4-Mb genome, but only 1604 functional genes, *M. leprae* has 1116 identifiable pseudogenes, many heavily degraded (15).

Although many reduced genomes show clear signs of gene loss, as described above, others do not. The organisms with some of the smallest genomes, such as *Mycoplasma* (5), appear to have extremely compact and streamlined genomes, with almost no evidence for nonfunctional genes. This is despite the fact that their genome size is apparently a derivative state and not representative of a small-genomed common ancestor (35). In these cases, the selection for DNA loss after gene inactivation appears to be strong, which does not seem to be the case with, for example, *M. leprae*. One possibility is that this difference is correlated with the fact that *Mycoplasma*s are extracellular pathogens, whereas *M. leprae* is an obligate intracellular pathogen. However, this supposition will require further investigation.

CONCLUSIONS AND FUTURE PERSPECTIVES

Our knowledge of bacterial genome structure has clearly benefited from the recent advances in genome sequencing. We now have detailed information relating to over 150 bacterial genomes and this figure continues to rise. This large dataset helps draw a useful picture of bacterial genomes in terms of size, geometry, and replicon number and the variation within these parameters. The detail of the data has brought new insights into genome structure and evolution and has provoked many hypotheses to explain observed patterns. The onus now is for those hypotheses to

be proven experimentally, though in many cases they may simply be strengthened or weakened by the accumulation of further genome sequences.

Since genome sequencing requires considerable funding, the focus of bacterial genomics thus far has been motivated primarily by medical and commercial considerations. The priority has been to sequence the genomes of pathogens along with a handful of model organisms and important environmental species. This strategy has greatly advanced our understanding of the mechanisms by which bacteria evolve and adapt to best fit their niche and, for pathogens, evade immune responses.

The prioritization of research resources has inevitably meant that the current dataset is biased, but continued sequencing at decreasing costs should help us move toward a representative view of all bacterial genomes regardless of our ability to culture them. The emerging field of metagenomics, the analysis of all genomes within a niche, will be crucial in this endeavor. Recent studies have carried out random shotgun sequencing of DNA representing ecological communities from an acid mine drainage (AMD) site (70) and from the Sargasso Sea (73). The dominance of just a few species within the AMD ecosystem meant that two near-complete and three partially complete genomes could be constructed. The Sargasso Sea sample was much more complex, so the billion bases of sequence generated did not assemble into many whole genomes but did indicate the presence of over 1800 species. These are the first steps for metagenomics but, if the progress of genomics is anything to go by, we can expect to see hundreds (if not thousands) of new complete bacterial genomes in the next decade. Eventually, our ability to interpret genome sequences will mean that complete characterization of metagenomes in terms of the nature and interactions of the individual species will be a realistic and exciting task.

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

LITERATURE CITED

1. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, et al. 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat. Genet.* 32:402–7
2. Allardet-Servent A, Michaux-Charachon S, Jumas-Bilak E, Karayan L, Ramuz M. 1993. Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J. Bacteriol.* 175:7869–74
3. Bao K, Cohen SN. 2001. Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. *Genes Dev.* 15:1518–27
4. Bao K, Cohen SN. 2003. Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. *Genes Dev.* 17:774–85
5. Barré A, de Daruvar A, Blanchard A, Blanchard A. 2004. MolliGen, a database dedicated to the comparative genomics of mollicutes. *Nucleic Acids Res.* 32:D307–10
6. Bentley SD, Brown S, Murphy LD, Harris DE, Quail MA, et al. 2004. SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). *Mol. Microbiol.* 51:1615–28

7. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417:141–47
8. Bentley SD, Maiwald M, Murphy LD, Pallen MJ, Yeats CA, et al. 2003. Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet* 361:637–44
9. Berlyn MKB. 1998. Linkage map of *Escherichia coli* K-12, edition 10: the traditional map. *Microbiol. Mol. Biol. Rev.* 62: 814–984
10. Bernal A, Ear U, Kyrpides N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* 29:126–27
11. Brewer BJ. 1988. When polymerases collide—replication and the transcriptional organization of the *Escherichia coli* chromosome. *Cell* 53:679–86
12. Cases I, de Lorenzo V, Ouzounis CA. 2003. Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.* 11:248–53
13. Casjens S. 1998. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* 32:339–77
14. Chang PC, Cohen SN. 1994. Bidirectional replication from an internal origin in a linear *Streptomyces* plasmid. *Science* 265:952–54
15. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. 2001. Massive gene decay in the leprosy bacillus. *Nature* 409:1007–11
16. Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23:324–28
17. Deng W, Burland V, Plunkett G, Boutin A, Mayhew GF, et al. 2002. Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* 184:4601–11
18. Eisen JA, Heidelberg JF, White O, Salzberg SL. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 1: research0011.1-.9
19. Ferdows MS, Barbour AG. 1989. Megabase-sized linear DNA in the bacterium *Borrelia burgdorferi*, the Lyme-disease agent. *Proc. Natl. Acad. Sci. USA* 86: 5969–73
20. Ferdows MS, Serwer P, Griess GA, Norris SJ, Barbour AG. 1996. Conversion of a linear to a circular plasmid in the relapsing fever agent *Borrelia hermsii*. *J. Bacteriol.* 178:793–800
21. French S. 1992. Consequences of replication fork movement through transcription units in vivo. *Science* 258:1362–65
22. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH. 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* 407:757–62
23. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, et al. 2001. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* 294:2323–28
24. Gruss A, Michel B. 2001. The replication-recombination connection: insights from genomics. *Curr. Opin. Microbiol.* 4:595–601
25. Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P. 1998. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G + C content of an endocytobiotic DNA. *J. Mol. Evol.* 47:52–61
26. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, et al. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406:477–83
27. Holden MTG, Titball RW, Peacock SJ, Cerdeño-Tárraga AM, Atkins T, et al. 2004. Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci. USA*. In press
28. Hopwood DA. 2003. The *Streptomyces* genome—be prepared! The completion of the sequence of a second *Streptomyces*

chromosome further establishes these soil-dwelling bacteria as nature's most prolific producers of potentially useful pharmaceuticals. *Nat. Biotechnol.* 21:505–6

29. Huang CH, Lin YS, Yang YL, Huang SW, Chen CW. 1998. The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. *Mol. Microbiol.* 28:905–16
30. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11:555–65
31. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, et al. 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* 9:189–97
32. Kieser HM, Kieser T, Hopwood DA. 1992. A combined genetic and physical map of the *Streptomyces coelicolor* A3(2) chromosome. *J. Bacteriol.* 174:5496–507
33. Klasson L, Andersson SGE. 2004. Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol.* 12: 37–43
34. Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. USA* 101: 3160–65
35. Koonin EV. 2000. How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* 1:99–116
36. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44:383–97
37. Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* 95: 9413–17
38. Lin YS, Kieser HM, Hopwood DA, Chen CW. 1993. The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol. Microbiol.* 10:923–33
39. Lobry JR, Sueoka N. 2002. Asymmetric directional mutation pressures in bacteria. *Genome Biol.* 3(10):research0058
40. Mackiewicz P, Mackiewicz D, Gierlik A, Kowalczuk M, Nowicka A, et al. 2001. The differential killing of genes by inversions in prokaryotic genomes. *J. Mol. Evol.* 53:615–21
41. McLean MJ, Wolfe KH, Devine KM. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* 47:691–96
42. Moran NA. 2002. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 108:583–86
43. Mushegian AR, Koonin EV. 1996. Gene order is not conserved in bacterial evolution. *Trends Genet.* 12:289–90
44. Musialowski MS, Flett F, Scott GB, Hobbs G, Smith CP, Oliver SG. 1994. Functional evidence that the principal DNA-replication origin of the *Streptomyces coelicolor* chromosome is close to the *dnaA-gyrB* region. *J. Bacteriol.* 176:5123–25
45. Ochman H. 2002. Bacterial evolution: chromosome arithmetic and geometry. *Curr. Biol.* 12:R427–28
46. Ochman H, Jones IB. 2000. Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J.* 19:6637–43
47. Orengo CA, Pearl F, Lee D, Bray J, Todd I, Thornton JM. 2001. Insights into protein evolution and function from the CATH structural database. *Biophys. J.* 80:140
48. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica serovar* Typhi CT18. *Nature* 413:848–52
49. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.* 35:32–40
50. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MTG, et al. 2001. Genome

sequence of *Yersinia pestis*, the causative agent of plague. *Nature* 413:523–27
51. Passarge E, Horsthemke B, Farber RA. 1999. Incorrect use of the term synteny. *Nat. Genet.* 23:387
52. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW. 2000. A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.* 299:907–30
53. Qin ZJ, Cohen SN. 1998. Replication at the telomeres of the *Streptomyces* linear plasmid pSLA2. *Mol. Microbiol.* 28:893–903
54. Ranea JAG, Buchan DWA, Thornton JM, Orengo CA. 2004. Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.* 336:871–87
55. Ravin N, Lane D. 1999. Partition of the linear plasmid N15: interactions of N15 partition functions with the *sop* locus of the F plasmid. *J. Bacteriol.* 181:6898–906
56. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042–47
57. Rocha EPC. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.* 10:393–95
58. Rocha EPC, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18:291–94
59. Rocha EPC, Danchin A. 2003. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* 31: 5202, 6570–77
60. Roth JR, Benson N, Galitski T, Haack K, Lawrence JG, Miesel L. 1996. Rearrangements of the bacterial chromosome: formation and applications. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. FC Neidhardt, pp. 2256–76. Washington, DC: ASM Press. 2nd ed.
61. Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, et al. 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* 415:497–502
62. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp APS. *Nature* 407:81–86
63. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, et al. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 406:959–64
64. Sueoka N. 1962. On genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* 48:582–92
65. Suyama M, Bork P. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* 17:10–13
66. Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* 2:research0020.1–.11
67. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformat.* 4:41
68. Tillier ERM, Collins RA. 2000. Genome rearrangement by replication-directed translocation. *Nat. Genet.* 26:195–97
69. Tourand Y, Kobryn K, Chaconas G. 2003. Sequence-specific recognition but position-dependent cleavage of two distinct telomeres by the *Borrelia burgdorferi* telomere resolvase, ResT. *Mol. Microbiol.* 48:901–11
70. Tyson GW, Hugenholtz P, Allen EE, Ram RJ, Banfield JF, et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43
71. Ussery DW, Hallin PF. 2004. Genome update: AT content in sequenced prokaryotic genomes. *Microbiology* 150:749–52
72. Ussery DW, Hallin PF. 2004. Genome update: length distributions of sequenced prokaryotic genomes. *Microbiology* 150: 513–16
73. Venter JC, Remington K, Hoffman J, Baden-Tillson H, Pfannkoch C, et al. 2004.

Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74

74. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, et al. 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. USA* 100: 12984–88
75. Weaver D, Karoonuthaisiri N, Tsai HH, Huang CH, Ho ML, et al. 2004. Genome plasticity in *Streptomyces*: identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. *Mol. Microbiol.* 51:1535–50
76. Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, et al. 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect. Immun.* 71:2775–86
77. Westberg J, Persson A, Holmberg A, Goesmann A, Lundeberg J, et al. 2004. The genome sequence of *Mycoplasma mycoides* subsp *mycoides* SC type strain PG1(T), the causative agent of contagious bovine pleuropneumonia (CBPP). *Genome Res.* 14:221–27
78. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* 294:2317–23
79. Xu Y, Bruno JF, Luft BJ. 2003. Detection of genetic diversity in linear plasmids 28-3 and 36 in *Borrelia burgdorferi sensu stricto* isolates by subtractive hybridization. *Microb. Pathog.* 35:269–78
80. Yang CC, Huang CH, Li CY, Tsay YG, Lee SC, Chen CW. 2002. The terminal proteins of linear *Streptomyces* chromosomes and plasmids: a novel class of replication priming proteins. *Mol. Microbiol.* 43:297–305

Annual Review of Genetics
Volume 38, 2004

CONTENTS