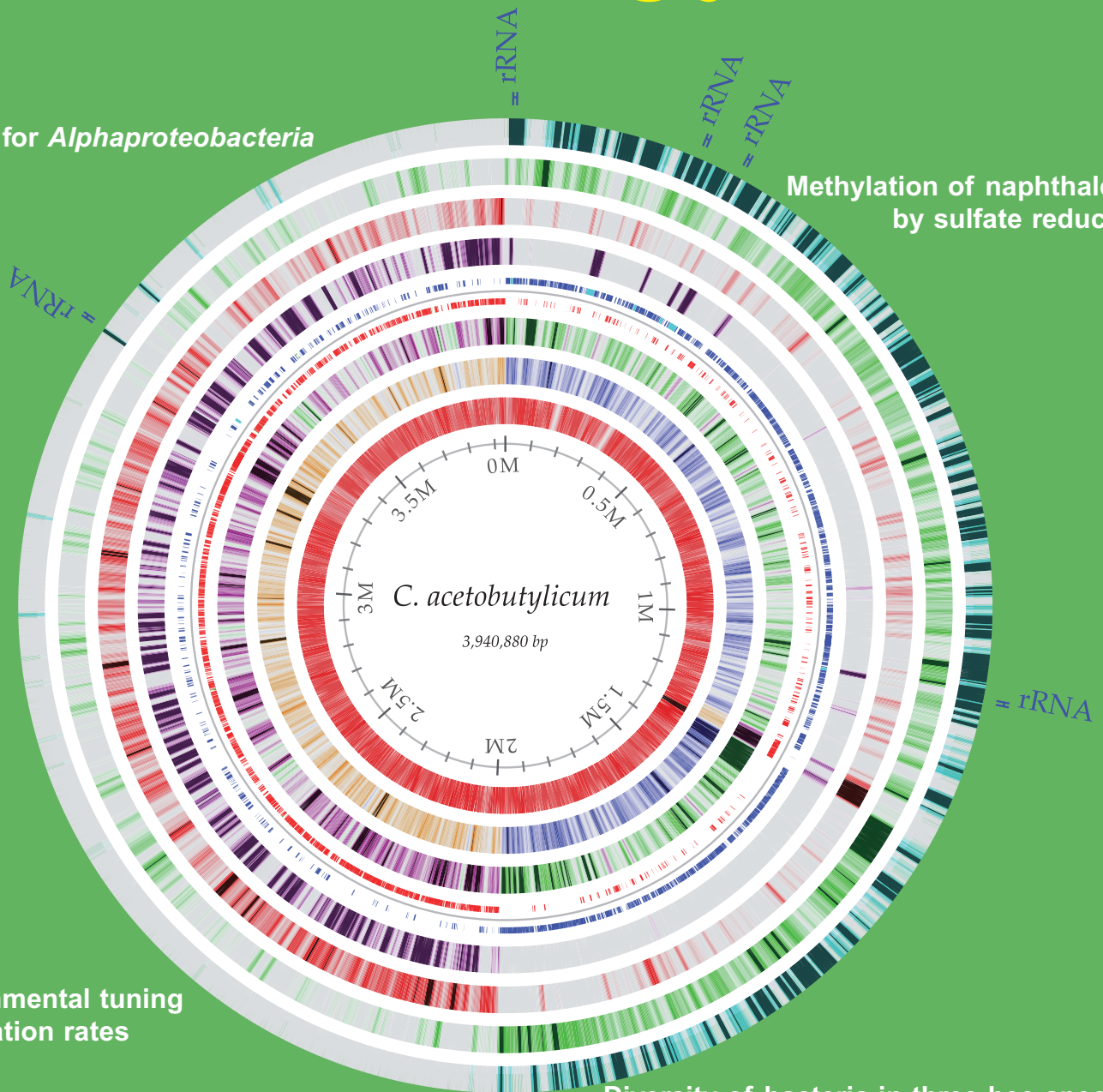


environmental microbiology

Microarray for *Alphaproteobacteria*

Methylation of naphthalene by sulfate reducers



Environmental tuning of mutation rates

Diversity of bacteria in three bee species

Origin of replication in circular prokaryotic chromosomes

Peder Worning,^{1,3†} Lars J. Jensen,^{2,3†} Peter F. Hallin,³
Hans-Henrik Stærfeldt³ and David W. Ussery^{3*}

¹Biological Sciences, AstraZeneca R and D Lund, S-221 87 Lund, Sweden.

²European Molecular Biology Laboratory, D-69117 Heidelberg, Germany.

³Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby, Denmark.

Summary

To predict origins of replication in prokaryotic chromosomes, we analyse the leading and lagging strands of 200 chromosomes for differences in oligomer composition and show that these correlate strongly with taxonomic grouping, lifestyle and molecular details of the replication process. While all bacteria have a preference for Gs over Cs on the leading strand, we discover that the direction of the A/T skew is determined by the polymerase- α subunit that replicates the leading strand. The strength of the strand bias varies greatly between both phyla and environments and appears to correlate with growth rate. Finally we observe much greater diversity of skew among archaea than among bacteria. We have developed a program that accurately locates the origins of replication by measuring the differences between leading and lagging strand of all oligonucleotides up to 8 bp in length. The program and results for all publicly available genomes are available from <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin>.

Introduction

The replication of bacterial chromosomes typically starts at a well-defined site, the origin of replication, from which two replication forks proceed in opposite directions. Replication continues until it is either stopped at a termination signal or the two replication forks meet and the whole genome is duplicated. While the origin is a unique site where replication is initiated in both directions, several

termination sites often exist within a chromosome. Each of these sites can only stop replication forks moving in one direction (Baker, 1995).

Because the two DNA strands making up the double helix are antiparallel and nucleotides are only added to the 3' end of the growing chain, the two strands are synthesized differently. One strand, the leading strand, is made continuously in the same direction as the replication fork is moving. Synthesis of the other strand, the lagging strand, must take place in the opposite direction to the movement of the replication fork. The lagging strand is therefore synthesized as smaller chains, Okazaki fragments, which are subsequently joined (Kornberg and Baker, 1992). The discontinuous synthesis of the lagging strand creates long stretches of single-stranded DNA during replication. This difference in synthesis of the two strands gives rise to a mutational bias between the two strands. It is a well known observation that bacterial chromosomes exhibit strand-specific biases, both in terms of strand-specific oligonucleotide sequences and also gene orientation (Lobry, 1996a; Rocha, 2004).

The replication of DNA is carried out by a functionally highly conserved protein complex, the DNA polymerase holoenzyme. The elongation of the growing DNA chain is catalysed by the polymerase α subunit, which exists in two forms. One is homologous to the *dnaE* gene from *Escherichia coli*, and the other is homologous to the *polC* gene from *Bacillus subtilis*. In *E. coli*, the two polymerase α subunits of the pol III holoenzyme are both encoded by the *dnaE* gene (Kornberg and Baker, 1992) and are not prededicated to the leading or the lagging strand (Yuzhakov *et al.*, 1996). In *B. subtilis*, the two α subunits are not identical and are encoded by the essential *polC* and *dnaE* genes. The PolC polymerase replicates the leading strand while the DnaE replicates the lagging strand (Dervyn *et al.*, 2001; le Chatelier *et al.*, 2004). The difference between the *polC* and *dnaE* encoded polymerase subunits is in the proofreading function. In *E. coli* the proofreading capacity of the polymerase resides in the interaction between the α and the θ subunits encoded by the *dnaE* and the *dnaQ* genes respectively (Kornberg and Baker, 1992). The PolC protein from *B. subtilis* is larger than the DnaE protein from *E. coli* and its C-terminal is 30% homologous to the θ subunit mentioned above. For *B. subtilis* and the Firmicutes in general, the leading strand is replicated by a polymerase subunit where proof-

Received 12 January, 2005; accepted 1 August, 2005. *For correspondence. E-mail dave@cbs.dtu.dk; Tel. +45 4525 2488; Fax +45 4593 1585. †These two authors contributed equally.

reading capacity resides in one polypeptide chain, while for the lagging strand the proofreading capacity involves a complex of two polypeptides. The *Thermotoga maritima* genome encodes both a PolC and a DnaE homologue (Huang and Ito, 1998). In the present work, we show that the *Fusobacterium nucleatum* and *Aquifex aeolicus* genomes also encode both a PolC and a DnaE homologue.

Several computational methods have been devised to locate the origin and terminus of replication in microbial genomes (reviewed in Rocha, 2004). The vast majority of these methods rely on analysing so-called skews which represent the difference between leading and lagging strand. In many bacterial chromosomes the leading strand contains more G's than C's and the origin can be identified by the G/C skew (Lobry, 1996a,b; Frank and Lobry, 2000). A similar albeit usually weaker strand bias is often seen for adenine and thymine where the leading strand normally contains more T's than A's (Rocha, 2000). The G/C and A/T skews can be combined as either the purine-skew (G,A vs. C,T) or the keto-skew (G,T vs. A,C), which can provide a better origin prediction than the single nucleotide skews (Freeman *et al.*, 1998). The information in the different single nucleotide skews can be combined into a three-dimensional curve, the Z-curve, which has been used to predict origins in both bacterial and archaeal chromosomes (Zhang and Zhang, 2002). Going beyond mononucleotide skews, a method based on skewed octamers has proven valuable in predicting origins in both bacterial and archaeal chromosomes (Salzberg *et al.*, 1998).

The origin and the terminus are turning points in a circular chromosome where the leading strand continues directly into the lagging strand. Like other methods, the method we describe here works by seeking the positions in the sequence that maximize the difference between the leading and the lagging strand. However, we go beyond generic base skews and instead search for chromosome specific oligomer skews involving all oligomers up to a length of eight nucleotides. We show that our method is more sensitive than existing ones based on mononucleotide skews or the octamer skews. Furthermore, it provides a quantitative measure for the difference between the leading and the lagging strand. Finally, we show that the direction of the A/T skew is determined by the type of DNA polymerase- α subunit, that is involved in proof-reading of the leading strand.

Results and discussion

Out of more than 200 circular bacterial chromosomes we have analysed, our measure of strand bias, the signal-to-noise ratio, varies between 44 for the Firmicute *Clostridium perfringens* and 0.08 for the cyanobacterium *Gloeo-*

bacter violaceus. While every computational method calculating strand skew would probably succeed in locating the origin in the *C. perfringens* genome, we doubt that it would be possible to locate the origin by comparing the two strands in *G. violaceus*, as the strand bias of this genome is only marginally above the 0.07 observed for randomly generated sequences.

The signal-to-noise ratio gives a very good indication of how easily the origin of replication is to locate: it can be done without difficulty when the signal-to-noise ratio is above 2, it is more difficult but still possible for values between 1 and 2, and for values lower than 1 it is almost impossible to locate the origin. Most of the genomes in the latter category are thermophiles, with the exceptions being Cyanobacteria and *Deinococcus radiodurans* (see Fig. 1). Most of the Cyanobacteria have an exceptionally low strand bias, which suggests that the process of DNA replication in these species is somehow different from other bacteria. Indeed, experimental analysis of the DNA replication in *Synechocystis* shows that the *dnaA* gene is not required for DNA replication (Richter *et al.*, 1998).

DNA polymerases influence skews

Our analysis of bacterial chromosomes shows that while the tendency of more G's on the leading strand and more C's on the lagging strand is a general trend, the direction of the A/T skew varies between species and phyla. The Firmicutes, Thermotogales, Aquificae and Fusobacteria have a surplus of G's and A's on the leading strand, which is in sharp contrast to Actinobacteria, Proteobacteria, Chlamydiae, Bacteroidetes and Spirochetes that all have a surplus of G's and T's on the leading strand. These two different skew patterns are exemplified by *B. subtilis* and *Bacteroides thetaiotaomicron* (see Fig. 2A and B). Typically, the G/C skew is stronger than the A/T skew, which in some species is almost absent (see Fig. 1).

The genomes with a positive A/T skew (more A's on the leading strand) contain both a *polC* and a *dnaE* homologue, while the other genomes only have a *dnaE* homologue. These skew patterns across distantly related species suggest that the skews reflect differences in the proofreading system. For *dnaE* the proofreading resides in the interaction between the θ and α subunits (le Chatelier *et al.*, 2004), while *polC* contains both the elongation and the proofreading functions in the α subunit (McHenry, 2003). In Firmicutes the leading strand is replicated by *polC* while the lagging strand is replicated by *dnaE* (Dervyn *et al.*, 2001).

We used the *polC* gene from *B. subtilis* and the *dnaE* gene from *E. coli* to make BLAST searches against all sequenced bacterial genomes. In *F. nucleatum*, *T. maritima*, *A. aeolicus* and all Firmicutes, we found full length matches to both sequences. In all other bacterial

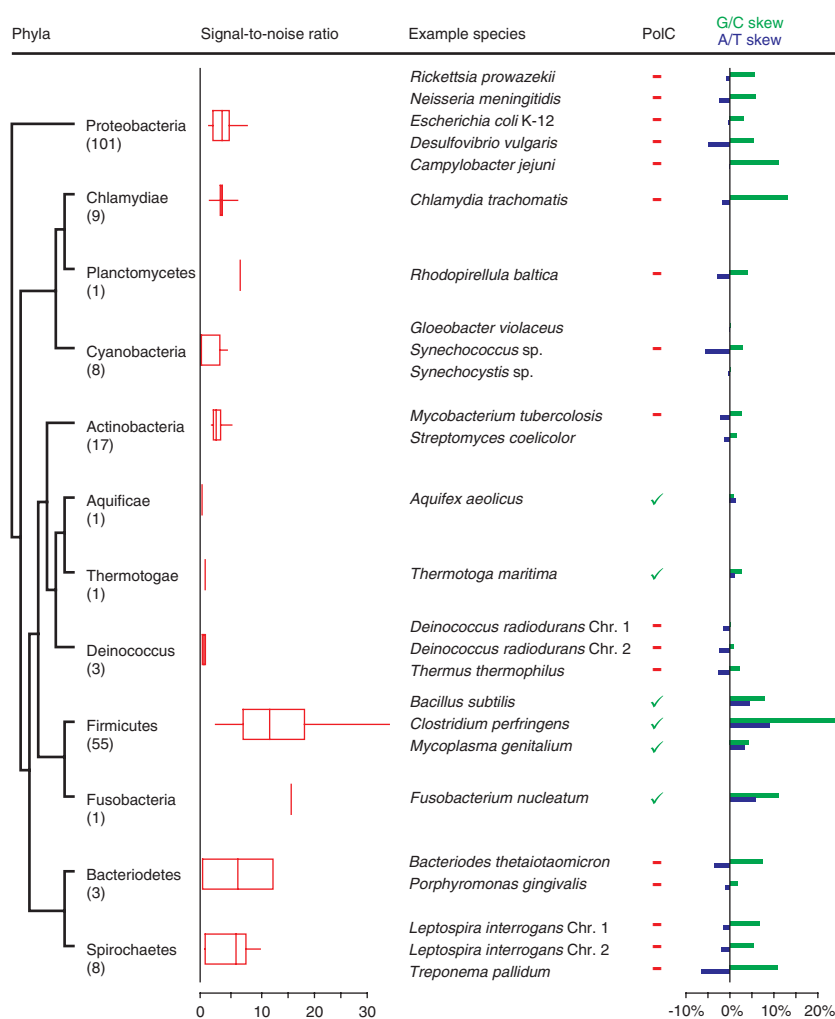


Fig. 1. Phylogenetic overview of skews. For each phyla, the distribution of signal-to-noise ratios is represented by a box-and-whiskers plot based on the 10, 25, 50, 75 and 90 percentiles. Within each phyla the presence/absence of a PoIC homologue is shown for selected species along with the G/C and A/T skews, shown as blue and green bars respectively. Note that the A/T skew is positive for all species with a PoIC homologue and vice versa.

genomes, including Proteobacteria and Cyanobacteria we found only a good match to the *dnaE* sequence. We thus observed perfect agreement between the direction of the A/T skew and the presence/absence of a *poIC* homologue (see Figs 1 and 2).

Skew strength and growth rate

The *T. maritima* origin of replication has proven difficult to find from skews due to the low signal-to-noise ratio. However, our method predicts an origin at position 164 kb (see Fig. 2D), close to where a consensus sequence for bacterial origins has been found (Lopez *et al.*, 2000). In contrast, the origin is easily located within the *F. nucleatum* chromosome as the signal-to-noise ratio is high (see Fig. 2C). Nonetheless, the origin of replication appears to have been misplaced in the published sequence (Kapatral *et al.*, 2002). *Bacteriodes thetaiotaomicron* is another recent example where our origin prediction method provides a reliable prediction, which differs considerably from

position zero in the published sequence (see Fig. 2B) (Xu *et al.*, 2003).

The five sequenced *Clostridia* genomes stand out by having an extraordinarily strong strand bias. The signal-to-noise ratios are about 40 for all five genomes, for which reason it is surprising that the origin was misplaced by 50 kb in the recent publication of the *Clostridium tetani* genome (Brüggemann *et al.*, 2003). The *Clostridium* group contains the fastest replicating organisms known, e.g. *C. perfringens* that has a minimal generation time of eight minutes (Shimizu *et al.*, 2002), which is likely to put severe strains on the genome architecture. In contrast, the smallest signal-to-noise ratios among the Firmicutes are observed for Mollicutes, i.e. *Mycoplasma* and *Ureaplasma*. These organisms have long doubling time despite their small genomes, e.g. 94 min for *Mycoplasma capricolum* where replication takes place at a rate of only ~ 100 bp s^{-1} (Seto and Miyata, 1998). From these observations we speculate that high growth rate may in general result a strong strand bias, which makes sense given that

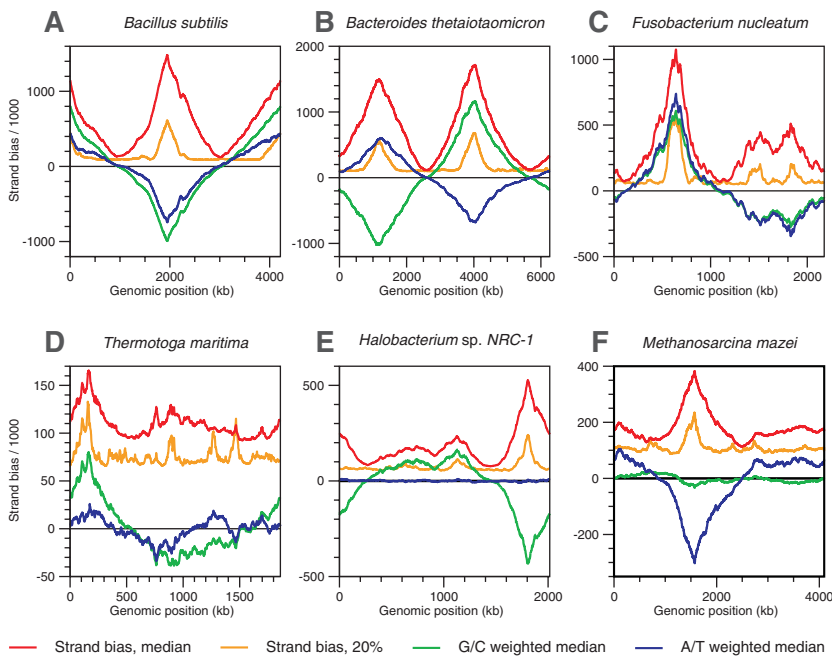


Fig. 2. The strand bias curves for six genomes. The median oligo skew curves (red) were calculated from five window sizes (50%, 55%, 60%, 65% and 70%; see *Experimental procedures* for details).

A. *B. subtilis* has a typical bacterial genome in that the G/C skew is positive and the A/T skew is negative.

B. *B. thetaiotaomicron* shows the same skew pattern. The origin of replication is predicted at position 4040 kb rather than at position zero.

C. For the *F. nucleatum* genome, both the G/C and the A/T skews are positive. The strand bias curves clearly suggest an origin at position 641 kb rather than at position zero.

D. In *T. maritima*, the skews are very weak, yet the origin is predicted at position 164 kb, very close to the likely true origin. As in *F. nucleatum* and Firmicutes, both the G/C and the A/T skews are positive, which is consistent with these genomes containing both a *polC* and a *dnaE* homologue.

E. *Halobacterium* NRC-1 is a rare example of a genome with a strong, negative G/C skew around the origin of replication.

F. No G/C skew is present around the origin in the *M. mazei* genome, however, a strong negative A/T skew allows the origin to be localized.

the genome must be replicated more often in faster dividing cells.

Two mechanisms for termination of replication

The Firmicutes are characterized by a large difference between leading and lagging strands (Rocha and Dancin, 2001). *B. subtilis* is a typical example of a Firmicute, where the origin is easily located from the G/C skew (see Fig. 2A). Our method shows two very strong, well-defined peaks that represent the origin and terminus of replication; the G/C-weighted curve shows which of the two peaks that corresponds to the origin. Both peaks remain strong even for a window size of 20%, which shows that the termination of the DNA replication is tightly controlled. Indeed, the terminus position at 1950 kb is in the middle of several mapped termination sites. Similarly, termination of replication appears to be tightly controlled in *B. thetaiotaomicron* but not in *F. nucleatum* or *T. maritima* (see Fig. 2B–D).

One or multiple origins of replication in archaea?

The mechanism of DNA replication is at present not firmly established in archaea. It is not known whether archaea have a single origin, like bacteria, or multiple origins, like eukarya. Evidence of a single origin of replication has been found in some archaeal genomes (Salzberg *et al.*, 1998; Claverys *et al.*, 1999; Lopez *et al.*, 1999; Mylykallio *et al.*, 2000).

Seven of the archaeal genomes analysed show one or two peaks in the strand bias curve that are relatively stable towards changes in the window size, which suggests that they have a single origin of replication. The variety of skewed oligomers is much greater among the several archaeal genomes than among the bacterial ones: the three *Pyrococcus* genomes have positive G/C skews, the two *Methanosarcina* genomes have no G/C skew but strong, negative A/T skews (see Fig. 2F), and *Halobacterium* NRC-1 has a very strong, negative G/C skew (see Fig. 2E).

Halobacterium NRC-1 is one of the few non-thermophilic, archaeal genomes that we have analysed. With a signal-to-noise ratio of 5.85, its strand bias is more than twice that of other archaeal genomes. Two putative origins of replication have previously been published based on the Z-curve method (Zhang and Zhang, 2003); however, consistent with our results only a single origin at position 1807 kb was identified in a recent targeted genetic screen for autonomously replicating sequences (Berquist and DasSarma, 2003).

For *Archaeoglobus fulgidus*, *Methanococcus jannaschii*, *Aeropyrum pernix* and the *Sulfolobus* genomes the number and position of the peaks change with the window size (data not shown), which could be caused by multiple origins of replication. Experimental evidence supports multiple origins of replication in the main chromosomes of *Sulfolobus solfataricus* and *M. jannaschii*, whereas *A. fulgidus* appears to have only a single origin (Maisnier-Patin *et al.*, 2002; Robinson *et al.*, 2004).

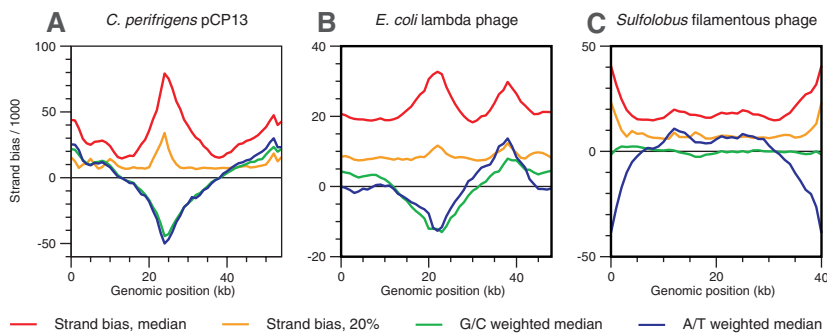


Fig. 3. Plasmids and phages.
A. *C. perfringens* pCP13.
B. *E. coli* lambda phage.
C. *Sulfolobus islandicus* filamentous phage.

Plasmids and phage genomes

Several different mechanisms exist for DNA replication of plasmids. Many small plasmids are replicated by the rolling-circle mechanism (Khan, 2000), where the two strands are replicated sequentially. For larger plasmids, both strands are typically replicated simultaneously by a theta-type mechanism, which can be either unidirectional or bidirectional. Our method, as well as all other methods based on skews, only works for plasmids that are replicated by the bidirectional theta-type mechanism. Of the 150 plasmids we have analysed, pCP13 from *C. perfringens* has the strongest strand bias (see Fig. 3A). Since the main chromosome of *C. perfringens* also showed the strongest bias in our analysis, this emphasizes that the signal-to-noise ratio reflects general properties of the DNA replication in a given organism, rather than intrinsic properties of the individual chromosome.

As in the case of plasmids, DNA replication of viral genomes is performed either by rolling circle or by the theta-type mechanism. The circular bacteriophage lambda genome is replicated bidirectionally from a single origin of replication (Kornberg and Baker, 1992), which has been experimentally mapped (Denniston-Thompson *et al.*, 1977). We correctly locate the origin of replication as shown in Fig. 3B. Our method is also able to find the origin of replication in some archaeal phage genomes, e.g. in a filamentous phage from *Sulfolobus islandicus* (see Fig. 3C). In bacteriophage and eukaryotic virus

genomes, the A/T skew tends to be stronger than the G/C skew, as opposed to the bacterial chromosomes where the G/C skew is usually stronger than the A/T skew.

Although short sequences, such as plasmids and viral genomes, usually give low signal-to-noise ratios, the signals are in some cases very clear as shown in Fig. 3.

Variation of oligomer skew within different *Yersinia pestis* strains

Yersinia pestis strains can be divided into three major subtypes, or biovars, which correspond to major pandemics: Antiqua, Mediaevalis and Orientalis. At the time of writing, the genome sequences of three *Y. pestis* strains have been published, namely strain CO-92 biovar Orientalis (Parkhill *et al.*, 2001), strain Kim10+ biovar Mediaevalis (Deng *et al.*, 2002) and strain 91001 biovar Mediaevalis has been sequenced (Song *et al.*, 2004). The latter is avirulent to humans yet it can kill mice (Song *et al.*, 2004).

As can be seen from Fig. 4, the three different *Y. pestis* strains exhibit quite different skew patterns, with one of the biovar Mediaevalis strains having very strong peaks (strain Kim10+), and the other Mediaevalis strain (91001) having essentially no skew. Indeed, comparative genomics reveals that the three strains are known to have undergone substantial genomic rearrangements due to IS elements (Deng *et al.*, 2002; Song *et al.*, 2004). Moreover, comparative analysis using microarrays indicate that

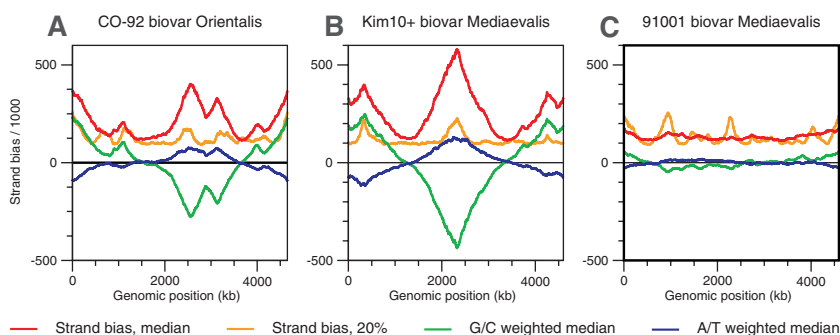


Fig. 4. Oligo skew measures for three different strains of *Yersinia pestis*.

A. *T. pestis* strain CO-92 biovar Orientalis.
B. *Y. pestis* strain Kim10+ biovar Mediaevalis.
C. *Y. pestis* strain 91001 biovar Mediaevalis.

strain 91001 should perhaps be classified as a new biovar, Microtus (Song *et al.*, 2004). At present, it is not known how often such rearrangements occur in nature; Parkhill and colleagues (2001) observed rearrangements in *Y. pestis* during growth in the laboratory. It is thus obviously dangerous to use a single genome as representative of a species (let alone a much larger taxonomic class such as phyla), considering that even global properties like skews may vary between strains (see Fig. 4).

Whether the differences in the *Y. pestis* strains is reflective of what might be expected for other genomes is not known presently. Among Firmicutes where multiple genomes of the same genus have been sequenced (e.g. *Bacillus* and *Streptococcus*), the skew patterns are consistently quite large and similar (data not shown, but see <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/> for the complete list).

Symmetry of inversions

Our method gives an overview of the genome architecture, i.e. how the origin and the terminus are located relative to each other and how the strand difference is varying within the genome. We see clear evidence for symmetric chromosomal inversions around the origin of replication between closely related species as suggested by several authors (Eisen *et al.*, 2000; Tillier and Collins, 2000). Figure 5 illustrates this by comparing the two *Bacillus* and two *Pseudomonas* genomes; the origin is located at position zero in all four genomes.

The *B. subtilis* and *Bacillus halodurans* termini are located 150 kb before/after the position opposite to the origin respectively (see Fig. 5A). A dot plot of the two genomes reveals that while the region around the origin has the same direction in the two genomes, a large region around the terminus is inverted in one genome relative to the other (see Fig. 5B). The precise match of the terminus positions in the *B. subtilis* genome with the reverse complement of the *B. halodurans* genome shows that the inversion has been symmetric relative to the origin (see Fig. 5A). In the *Pseudomonas* genomes in Fig. 5C, the termini of replication again only match if one looks at the reverse complement of one of the genomes. While this could in principle be explained by a single inversion symmetric to the origin of replication, the clear X-pattern in the dot plot reveals that numerous such inversions must have taken place in *Pseudomonas aeruginosa* and *Pseudomonas putida* (Fig. 5D). These observations provide independent evidence for inversions being symmetric relative to the origin as previously suggested by others (Eisen *et al.*, 2000; Tillier and Collins, 2000).

Conclusions

The oligomer skew method presented here is a very sensitive method for predicting boundaries between leading strand and lagging strand. It gives a general measure of the difference between the two stands around a number of putative origin positions. This method is much more sensitive than a cumulative G/C skew or the skewed

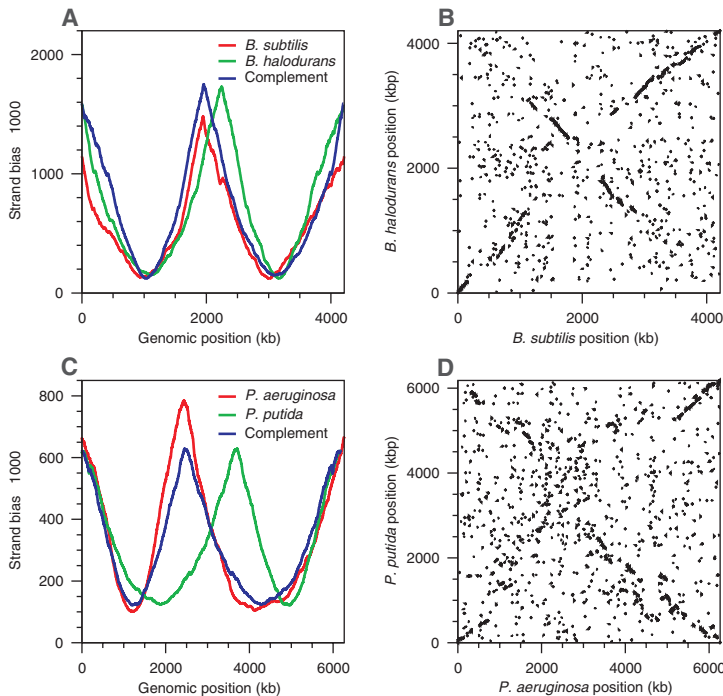


Fig. 5. Symmetry of inversions relative to replication origin. In the dot plots (B and D), each dot corresponds to two protein coding genes with a BLASTP *E*-value smaller than 10^{-40} and a match length at least 95% of the longest protein. Only the best match to each protein was retained.

A. Strand bias curves for *B. subtilis* and *B. halodurans*.

B. Dot plot of *B. subtilis* versus *B. halodurans*.

C. Strand bias curves for *P. aeruginosa* and *P. putida*.

D. Dot plot of *P. aeruginosa* versus *P. putida*.

octamer method. Furthermore, the signal-to-noise ratio provides a quantitative measure of the difference between leading and lagging strands. We have demonstrated that the direction of the A/T skew is correlated with the type of polymerase- α subunit performing the DNA elongation. This finding is in agreement with the observed proofreading capabilities of the different DNA polymerase- α subunit complexes. The signal-to-noise ratios of the different genomes correlates with both phylogeny and the lifestyle of the species. The species with low strand bias are dominated by thermophiles, although *Cyanobacteria* and *D. radiodurans* also have exhibit a weak strand bias. It is possible that the low mutational bias between leading and lagging strand is overruled by the selectional restrictions of the extreme environments inhabited by both thermophiles and *D. radiodurans*. The reason for the total lack of strand bias in some *Cyanobacteria* is still unclear, although it may be related to the known peculiarity in how DNA polymerase proceeds in *Synechosystis* (Richter *et al.*, 1998). The archaeal chromosomes are replicated by a type of polymerase very different from the bacterial polymerases, which results in much more varied skew patterns than in bacteria. Furthermore, some archaeal chromosomes use multiple origins of replication while others only use one.

Experimental procedures

A number of hypothetical origin positions, p , are chosen, equally spaced throughout the sequence. For whole chromosomes we use a 1000 bp spacing between the hypothetical origin positions and for the smaller viral genomes or plasmid sequences we use a spacing of 100 bp. Within a window centred around every such position, the number of occurrences of all oligonucleotides up to a length of $n = 8$ nucleotides are counted in both the leading and the lagging strand, assuming that p is the replication origin. The strand bias between leading and lagging strands is then calculated for every position.

The strand bias

When a hypothetical origin position p is chosen, an oligonucleotide, i , is encountered $N_{lead,p}^i$ times in the leading strand and $N_{lagg,p}^i$ in the lagging strand relative to p . This difference can be expressed as a weighted double Kullback-Leibler distance, D_p^i , which is calculated as:

$$D_p^i = (N_{lead,p}^i - N_{lagg,p}^i) \log_2 \left(\frac{N_{lead,p}^i + r}{N_{lagg,p}^i + r} \right)$$

where r is a re-normalization number, which prevents divergence when the count is very low in one of the strands. We have tried several values for r and we have chosen to use $r = 5$ in the present work. The strand difference, D_p , is the weighted double Kullback-Leibler distance between leading and lagging strand for the origin position p . It is calculated as the sum over all oligonucleotides:

$$D_p = \sum_i D_p^i$$

The strand difference provides a measure of the difference between leading and lagging strand, assuming that the origin (or the terminus) is located at the position p . By plotting the strand difference (D_p) as a function of the position (p), a curve with two peaks corresponding to the origin and terminus of replication is typically obtained (see Fig. 2A–C). The two peaks mark the positions that provide the maximal difference between leading and lagging strand.

Influence of window size

In order to see the difference in strand bias around the two peaks we only use a fraction, w , of the genome around the hypothetical origin position, p , when D_p is calculated. This fraction is called the window size and it can be varied between 0 and 1. Figure 2A shows an example of how the strand difference curve evolves when the window size is changed for genomes with a strong strand bias. The base width of the peaks corresponds to the window size.

A circular bacterial genome has one origin position while there are several possible terminus positions. This could imply that the border between leading and lagging strands is sharper at the origin than at the terminus and we expect in this case that the origin peak is sharper and more stable than the terminus peak when we vary the window size. This pattern is shown for the bacteria *F. nucleatum* and *T. maritima* as well as the Archaea *Methanosarcina mazei* and *Halobacterium* NRC-1 (Fig. 2C–F).

When the difference between leading and lagging strand is weak, the position of the peaks in the strand difference curve can vary with the window size. To obtain a robust estimate we first calculate the strand bias with five different window sizes: 50%, 55%, 60%, 65% and 70%. These curves are then scaled to have the same area as the 60% curve and the median of the five scaled curves is calculated at every position in the genome to use the information from all five curves.

Recognizing origin from terminus

We use two strategies to distinguish the origin from the terminus. One is based on how the shape and the position of the two peaks changes when the window size is varied as explained above. The other is based on a weighted strand difference, which is defined as:

$$I_{p,f} = \sum_i f^i \cdot I_p^i,$$

where f^i is the weighting factor for oligonucleotide i , which is calculated as the sum of the weights for the nucleotides in the oligo divided by the length of the oligo. We define two weighting schemes:

G/C where the weights are +2 for G, –2 for C, and 0 for A and T

A/T where the weights are +2 for A, –2 for T, and 0 for G and C

Plots of the strand difference and the weighted strand differences are shown in Fig. 2. The peak with the positive G/C-

weighted strand difference curve represent the origin. The discrimination between origin and terminus by this approach is generally reliable provided that the shape and amplitude of the G/C-weighted curve is similar to that of the raw strand bias curve.

The signal-to-noise ratio

When every oligo in a genome of several million base pairs is counted and compared between leading and lagging strands there will always be a strand bias at every position. The interesting quantity to observe is how much the strand difference at a special position rises above the background level. The signal strength is the difference between the maximal and minimal strand difference I_p of the whole genome. The size of the genome will influence the signal strength and to get a measure of the signal quality we calculate a signal-to-noise ratio, S/N , as the ratio between the signal strength and the minimal strand difference:

$$S/N = \frac{I_{p,max} - I_{p,min}}{I_{p,min}}$$

Randomization of genomes

In order to find out which signal-to-noise ratio to expect from a genome without any large scale structures, we made random DNA sequences using four different genomes as templates and calculated the signal-to-noise ratio for each of those. The four genomes have been chosen to represent very different cases: *C. perfringens* and *G. violaceus* having the largest and smallest signal-to-noise ratio, respectively, *Mycoplasma genitalium* being smallest genome, and *Methanobacterium thermoautotrophicum* an archaeal genome with a very unusual oligomer skew. For each, 10 randomized genomes of the same length and of the same octamer composition were generated using a seventh order Markov model and their average signal-to-noise ratio was calculated. Very similar signal-to-noise ratios were obtained for the four genomes: 0.062 (*G. violaceus*), 0.066 (*M. thermoautotrophicum*), 0.070 (*C. perfringens*) and 0.074 (*M. genitalium*). The standard deviation across randomizations was ± 0.01 for all four genomes, showing that 0.07 can be regarded as a universal lower limit for detectable skews.

Acknowledgements

The authors would like to thank Hanno Teeling from the Max-Planck-Institut für Marine Mikrobiologie in Bremen, Germany, and people at CBS for valuable propositions and conversations. This work was supported by grants from the Danish National Research Foundation and the Danish Natural Science Research Council.

References

Baker, T. (1995) Replication arrest. *Cell* **80**: 521–524.
Berquist, B.R., and DasSarma, S. (2003) An archaeal chromosomal autonomously replicating sequence element

from an extreme halophile, *Halobacterium* sp. strain NRC-1. *J Bacteriol* **185**: 5959–5966.
Brüggemann, H., Bäumer, S., Fricke, W.F., Wiezer, A., Liesegang, H., Decker, I., et al. (2003) The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc Natl Acad Sci USA* **100**: 1316–1321.
le Chatelier, E., Bécherel, O.J., d'Alençon, E., Canceill, E., Ehrlich, S.D., Fuchs, R.P., and Jannièrè, L. (2004) Involvement of DnaE, the second replicative DNA polymerase from *Bacillus subtilis*, in DNA mutagenesis. *J Biol Chem* **279**: 1757–1767.
Claverys, J.-P., Granadel, C., Berry, A.M., and Paton, J.C. (1999) Identification of putative chromosomal origins of replication in Archaea. *Mol Microbiol* **32**: 883–886.
Deng, W., Burland, V., Plunkett, G., III, Boutin, A., Mayhew, G.F., Liss, P., et al. (2002) Genome sequence of *Yersinia pestis* KIM. *J Bacteriol* **184**: 4601–4611.
Denniston-Thompson, K., Moore, D.D., Kruger, K.E., Furth, M.E., and Blattner, F.R. (1977) Physical structure of the replication origin of bacteriophage lambda. *Science* **198**: 1051–1056.
Dervyn, E., Suski, C., Daniel, R., Gruand, C., Chapuis, J., Errington, J., et al. (2001) Two essential DNA polymerases at the bacterial replication fork. *Science* **294**: 1716–1719.
Eisen, J.A., Heidelberg, J.F., White, O., and Salzberg, S.L. (2000) Evidence for symmetric chromosomal inversions around the replication origin in Bacteria. *Genome Biol* **1**: 1–9.
Frank, A.C., and Lobry, J.R. (2000) OriLoc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16**: 560–561.
Freeman, J.M., Plasterer, T.N., Smith, T.F., and Mohr, S.C. (1998) Patterns of genome organization in Bacteria. *Science* **279**: 1827a.
Huang, Y.P., and Ito, J. (1998) The hyperthermophilic bacterium *Thermotoga maritima* has two different classes of family c dna polymerases: evolutionary implications. *Nucleic Acids Res* **26**: 5300–5309.
Kapatal, V., Anderson, I., Ivanova, N., Reznik, G., Los, T., Lykidis, A., et al. (2002) Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J Bacteriol* **184**: 2005–2018.
Khan, S.A. (2000) Plasmid rolling-circle replication: recent developments. *Mol Microbiol* **37**: 477–484.
Kornberg, A., and Baker, T. (1992) *DNA Replication*, 2nd edn. New York, USA: Freeman.
Lobry, J.R. (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**: 660–665.
Lobry, J.R. (1996b) Origin of replication of *Mycoplasma genitalium*. *Science* **272**: 745–746.
Lopez, P., Hervé, P., Myllykllio, H., and Forterre, P. (1999) Identification of putative chromosomal origins of replication in archaea. *Mol Microbiol* **32**: 883–886.
Lopez, P., Forterre, P., le Guyader, H., and Philippe, H. (2000) Origin of replication of *Thermotoga maritima*. *Trends Genet* **16**: 59–60.
McHenry, C. (2003) Chromosomal replicases as asymmetric dimers: studies of subunit arrangement and functional consequences. *Mol Microbiol* **49**: 1157–1165.
Maisnier-Patin, S., Malandrin, L., Birkeland, N.-K., and BERNANDER, R. (2002) Chromosome replication patterns in the

- hyperthermophilic euryarchaea *Archaeoglobus fulgidus* and *Methanocaldococcus* (*Methanococcus*) *jannaschii*. *Mol Microbiol* **45**: 1443–1450.
- Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., *et al.* (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* **288**: 2212–2215.
- Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T., Prentice, M.B., *et al.* (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**: 523–527.
- Richter, S., Hagemann, M., and Messer, W. (1998) Transcriptional analysis and mutation of a *dnaA*-like gene in *Synechocystis* sp. strain PCC 6803. *J Syst Bacteriol* **180**: 4946–4949.
- Robinson, N.P., Dionne, I., Lundgren, M., Marsh, V.L., Bernander, R., and Bell, S.D. (2004) Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell* **116**: 25–38.
- Rocha, E.P.C. (2000) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol* **10**: 393–395.
- Rocha, E.P.C. (2004) The replication-related organization of bacterial genomes. *Microbiology* **150**: 1609–1627.
- Rocha, E., and Dancin, A. (2001) Ongoing evolution of strand composition in bacterial genomes. *Mol Biol Evol* **18**: 1789–1799.
- Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., and Tomb, J.-F. (1998) Skewed oligomers and the origin of replication. *Gene* **217**: 57–67.
- Seto, S., and Miyata, M. (1998) Cell reproduction and morphological changes in *Mycoplasma capricolum*. *Mol Microbiol* **180**: 256–264.
- Shimizu, T., Ohtani, K., Hirakawa, H., Ohshima, K., Yamashita, A., Shiba, T., *et al.* (2002) Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc Natl Acad Sci USA* **99**: 996–1001.
- Song, Y., Tong, Z., Wang, J., Wang, L., Guo, Z., Han, Y., *et al.* (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* **11**: 179–197.
- Tillier, E.R.M., and Collins, R.A. (2000) Genome rearrangement by replication directed translocation. *Nat Genet* **26**: 195–197.
- Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., *et al.* (2003) A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science* **299**: 2074–2076.
- Yuzhakov, A., Turner, J., and O'Donnell, M. (1996) Replisome assembly reveals the basis for asymmetric function in leading and lagging strand replication. *Cell* **86**: 877–886.
- Zhang, R., and Zhang, C.-T. (2002) Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z curve method. *Biochem Biophys Res Comm* **297**: 396–400.
- Zhang, R., and Zhang, C.-T. (2003) Multiple replication origins of the archaeon *Halobacterium* species NRC-1. *Biochem Biophys Res Comm* **302**: 728–734.