Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 333 (2004) 121–133
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes

Pedro Bernaola-Galván[a], José L. Oliver[b], Pedro Carpena[a], Oliver Clay[c], Giorgio Bernardi[c,*]

[a] *Dpto. de Física Aplicada II, Universidad de Málaga, E-29071, Málaga, Spain*
[b] *Dpto. de Genética, Inst. de Biotecnología, Universidad de Granada, Spain*
[c] *Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy*

Received 29 August 2003; received in revised form 14 November 2003; accepted 10 February 2004

Available online 6 May 2004

Received by Wentian Li

Abstract

The sequencing of prokaryotic genomes covering a wide taxonomic range has sparked renewed interest in intrachromosomal compositional (GC) heterogeneity, largely in view of lateral transfers. We present here a brief overview of some methods for visualizing and quantifying GC variation in prokaryotes. We used these methods to examine heterogeneity levels in sequenced prokaryotes, for a range of scales or stringencies. Some species are consistently homogeneous, whereas others are markedly heterogeneous in comparison, in particular *Aeropyrum pernix*, *Xylella fastidiosa*, *Mycoplasma genitalium*, *Enterococcus faecalis*, *Bacillus subtilis*, *Pyrobaculum aerophilum*, *Vibrio vulnificus* chromosome I, *Deinococcus radiodurans* chromosome II and *Halobacterium*. As we discuss here, the wide range of heterogeneities calls for reexamination of an accepted belief, namely that the endogenous DNA of bacteria and archaea should typically exhibit low intrachromosomal GC contrasts. Supplementary results for all species analyzed are available at our website: http://bioinfo2.ugr.es/prok.

Keywords: Compositional heterogeneity; Lateral gene transfer; Segmentation; DNA correlations; Isochores; Analytical ultracentrifugation

1. Introduction

A widespread and long-standing belief has persisted since the 1950s, namely that the intrinsic compositional (GC) contrasts within prokaryotic chromosomes should be low (see, e.g., Lawrence and Ochman, 1997 and references therein). In order to assess the contrasts that are actually present in archaea and bacteria, we present here an overview of intrachromosomal GC heterogeneities, as calculated from complete DNA sequences using different methods. This study extends a much earlier compilation of heterogeneity data from ultracentrifuge work on *Clostridium perfringens*, *Haemophilus influenzae*, *Micrococcus luteus*, *Streptococcus pneumoniae*, *Bacillus subtilis* and *Escherichia coli* (Cuny et al., 1981). Using a variety of measures and stringencies or scales, we consistently observe a wide range of heterogeneities.

Before quantifying the GC variabilities that exist within prokaryotic genomes, it is important to place them in their proper context. The dramatic GC contrasts in prokaryotic chromosomes are often generated by relatively short regions ($\ll$100 kb). At scales above 5–10 kb, contrasts tend to become small, compared to the huge GC contrasts found within chromosomes of warm-blooded vertebrates at such scales. Indeed, GC levels of adjacent isochores in human can differ by as much as 20–25% GC and extend over a megabase or more.

A traditional way of quantifying the compositional heterogeneity of a genome is to calculate the distribution of GC levels of its fragments (segments, windows) for a fixed size, and then to calculate the standard deviation of this distribution. The GC distribution can be obtained either from the entire genomic sequence, where this is available, or experimentally by analytical ultracentrifugation of the ge-

Abbreviations: bp, base; kb, kilobase; Mb, megabase pairs; GC, molar fraction of guanine and cytosine (G+C) in DNA, guanine-cytosine base pair; rRNA, rDNA, ribosomal RNA/DNA; SCC, sequence compositional complexity.

* Corresponding author. Tel.: +39-81-583-3300; fax: +39-81-245-5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.02.042

nomic DNA in a CsCl density gradient. In prokaryotes, GC distributions become narrower as the fragment size increases, so heterogeneity must be specified with respect to a scale, or stringency.

In view of one or two recent misunderstandings, it should be pointed out that random sequences, consisting of a sequence or chain of independent and identically distributed ("i.i.d.") nucleotides, cannot be used as a reference or null model for homogeneity (see Bernardi, 2001 and Li et al., 2003 for discussions). No natural DNA except highly repetitive satellite DNA is as homogeneous as such a random sequence. This observation was made by several authors already in the 1950s and early 1960s. Essentially all genomic DNA exhibits positive autocorrelations and/or systematic GC differences within genes, notably among the three codon positions. Such dependencies and systematic fluctuations in mean GC increase intrachromosomal heterogeneity to levels that are well above the random sequence level.

2. Materials and methods

2.1. Measures derived via segmentation (segmentation algorithm, sequence compositional complexity)

Entropic compositional segmentation (Bernaola-Galván et al., 1996; Román-Roldán et al., 1998) has proven to be a powerful method to partition a DNA sequence into fairly homogeneous, isochore-like segments. We applied here a recently developed segmentation variant which proceeds by maximizing the GC contrast at each step of a recursion. Early versions of this algorithm were used to find isochore boundaries in mammalian genomes (Oliver et al., 2001, 2002) and a detailed description of the recent improvements will be provided elsewhere (Oliver et al., 2004). In brief, the algorithm works as follows. First, short-scale sequence heterogeneity below a certain minimum length is filtered out using a coarse-graining (or denoising) technique to eliminate short-scale variations of GC content. We have chosen this minimum length to be 3 kb, which corresponds to a homogeneity criterion for mammalian isochores, derived from ultracentrifugation of DNA at different molecular weights. Once coarse-grained, the original nucleotide sequence is transformed into an array of real numbers corresponding to GC values in fixed-length, non-overlapping windows 3 kb in length.

Second, we partition the array of GC values, which is composed of many segments with different mean values, in such a way as to maximize the GC contrast between adjacent segments. In doing so, we move a sliding pointer from left to right along the array of GC values. At each position of the pointer, we compute the mean of the subset of GC values to the left and to the right of the pointer. To measure the significance of the difference between left and right mean values, we use the t statistic. We next determine the position of the pointer for which t reaches its maximum value, t_{max}, and compute the statistical significance of t_{max} by means of Monte Carlo simulations, following Bernaola-Galván et al. (2001). The significance level $P(\tau)$ of a possible cutting point with $t_{max} = \tau$ is defined as the probability of obtaining the value τ or lower values within a random sequence. Thus, a series of N random numbers of fixed mean would remain unsegmented with probability $P(\tau)$. Related change-point problems have also been discussed elsewhere (see e.g. Gionis and Mannila, 2003; Page, 1955).

Third, we check if this significance exceeds a selected threshold P_0, usually taken to be 95%. If so, then the sequence is cut at this point into two subsequences; otherwise the sequence remains undivided. If the sequence is cut, the procedure continues recursively for each of the two resulting subsequences created by each cut. All resulting segments have a statistically significant difference in their means. The process stops when none of the possible cutting points has a significance exceeding P_0, and we say that the sequence has been segmented at the "significance level P_0". Our method leads to partitioning of a DNA sequence into compositional domains with well-defined mean GC levels, each significantly different from the mean GC level of the adjacent domains.

Once a sequence has been segmented, a natural measure of the GC heterogeneity, or complexity, among the segments of a DNA sequence, is the sequence compositional complexity, or SCC (Román-Roldán et al., 1998). The SCC of a segmented sequence is the difference between the sequence's informational (Shannon) entropy and the sum of all its segments' entropies, weighted by their respective lengths. Here, the entropy of a segment is $-\mu\log_2\mu-(1-\mu)\log_2(1-\mu)$, where μ is the mean GC level of the segment, expressed as a fraction of 1. An important feature of the SCC measure is that it is essentially independent of the choice of starting point in a circular chromosome (see Li et al., 2002).

2.2. Autocorrelations and GC differences among codon positions

We first consider, for simplicity, a hypothetical sequence in which the base pairs are identically distributed, although not independent. In this case, the mean GC level remains the same at all locations along the sequence (as in many intergenic sequences). The standard deviation among the sequence's segments (fragments, windows) of any fixed length is then determined by the serial autocorrelations that may be present in the sequence, i.e., by the sequence's correlation function. This function specifies, for any distance d along the sequence, the autocorrelation among two base pairs that are separated by this distance (see Bernaola-Galván et al., 2002): $c_d = \langle(u_i-\mu)(u_{i+d}-\mu)\rangle/(\mu(1-\mu))$. Here, the angular brackets $\langle\ldots\rangle$ denote the average (mean,

expectation), taken over all possible locations i in the sequence, u_i is the value (AT = 0, GC = 1) of the base pair at the i-th location, and μ is the mean GC level. In this case, the autocorrelations for different distances contribute to the standard deviation via an exact sum. The sum expresses the vector of standard deviations of segments having lengths $l = 1,2,3\ldots$ as a product of a lower triangular matrix and the vector of autocorrelations for distances $d = 1,2,3\ldots$ (see Clay, 2001). For a given GC level and segment length, higher positive autocorrelations will typically lead to higher standard deviations. By inverting the matrix, conversely, one can obtain an estimate of the autocorrelation function (or correlogram) from a vector of standard deviations for different lengths. It is, however, simpler to calculate the correlogram directly by sampling over the pairs of base pairs in the sequence.

In real sequences, which include genes, base pairs are not identically distributed: GC levels in coding regions have systematically different means in the three codon positions. The (often large) differences among the three codon positions will contribute to the correlogram, as calculated formally via the above definition, and to the heterogeneity.

2.3. Heterogeneity among GC_3 levels of genes

In our study we report analyses of GC levels in chromosomal DNA, but the basic results for scales around 1 kb (e.g. for heterogeneity ranking of species) are similar, mutatis mutandis, to those obtained for GC or GC_3 (third codon position GC) levels of genes, as we could verify by checks (not shown). The explanation is that prokaryotic (but not higher eukaryotic) chromosomes consist mostly, or almost entirely, of coding DNA, their genes are typically compact, and GC levels at different codon positions are related, to a good approximation, by conserved linear equations (major axes; see Cruveiller et al., 2003 and references therein).

2.4. Taxa analyzed

The archaeal and bacterial chromosomes analyzed are those that had been sequenced and were available at EBI ftp://ftp.ebi.ac.uk/pub/databases/embl/genomes/) in August 2003. They are listed here in abbreviated form: number used to represent the species in Fig. 3, species (first four letters of genus, first three letters of species, chromosome/strain), and sequence compositional complexity (SCC in bits/kb, for a stringency set at the 3 kb/95% level; see Section 2.1). A full table that also lists GC levels, sequence lengths and accession numbers of the chromosomes (and of sequenced plasmids and phages) is available at our web site, http://bioinfo2.ugr.es/prok/, (Supplementary Table 1) together with segmentation details and compositional maps for each species.

Archaea: 1 *Aero_per* 4.83, 2 *Arch_ful* 0.34, 3 *Halo_sp* 1.39, 4 *Meth_jan* 1.03, 5 *Meth_kan* 1.01, 6 *Meth_ace* 0.99, 7 *Meth_maz* 0.27, 8 *Meth_the* 0.34, 9 *Pyro_aer* 1.77, 10 *Pyro_aby* 0.80, 11 *Pyro_fur* 0.66, 12 *Pyro_hor* 0.89, 13 *Sulf_sol* 0.76, 14 *Sulf_tok* 0.07, 15 *Ther_aci* 0.42, 16 *Ther_vol* 0.10.

Bacteria: 1a *Agro_tum_c* 0.17, 1b *Agro_tum_l* 0.29, 2a *Agro_tum_c* 0.17, 2b *Agro_tum_l* 0.30, 3 *Aqui_aeo* 0.00, 4 *Baci_ant* 0.74, 5 *Baci_cer* 0.94, 6 *Baci_hal* 0.61, 7 *Baci_sub* 2.20, 8 *Bact_the* 1.38, 9 *Bifi_lon* 1.10, 10 *Borr_bur* 0.09, 11 *Brad_jap* 1.22, 12a *Bruc_mel_1* 0.58, 12b *Bruc_mel_2* 0.00, 13a *Bruc_sui_1* 0.66, 13b *Bruc_sui_2* 0.00, 14 *Buch_aph* 0.00, 15 *Buch_aph* 0.00, 16 *Buch_aph* 0.16, 17 *Camp_jej* 0.00, 18 *Cand_Blo* 1.58, 19 *Caul_cre* 0.23, 20 *Chla_mur* 0.23, 21 *Chla_tra* 0.15, 22 *Chla_cav* 0.14, 23 *Chla_pne* 0.53, 24 *Chla_pne* 0.38, 25 *Chla_pne* 0.51, 26 *Chla_pne* 0.51, 27 *Chlo_tep* 0.23, 28 *Clos_ace* 0.64, 29 *Clos_per* 0.64, 30 *Clos_tet* 0.70, 31 *Cory_eff* 0.82, 32 *Cory_glu* 0.75, 33 *Cory_glu* 0.75, 34 *Coxi_bur* 0.00, 35a *Dein_rad_1* 0.00, 35b *Dein_rad_2* 4.19, 36 *Ente_fae* 2.53, 37 *Esch_col* 0.74, 38 *Esch_col_K12* 0.55, 39 *Esch_col_O157* 0.09, 40 *Esch_col_O157s* 0.04, 41 *Fuso_nuc* 0.69, 42 *Haem_inf* 0.08, 43 *Heli_hep* 0.12, 44 *Heli_pyl* 0.00, 45 *Heli_pyl* 0.61, 46 *Lact_pla* 1.27, 47 *Lact_lac* 0.28, 48a *Lept_int_1* 0.06, 48b *Lept_int_2* 0.00, 49 *List_inn* 0.44, 50 *List_mon* 0.87, 51 *Meso_lot* 0.89, 52 *Mycob_bov* 0.00, 53 *Mycob_lep* 0.18, 54 *Mycob_tub* 0.00, 55 *Mycob_tub* 0.00, 56 *Mycop_gal* 0.00, 57 *Mycop_gen* 2.34, 58 *Mycop_pen* 0.06, 59 *Mycop_pne* 1.28, 60 *Mycop_pul* 0.00, 61 *Neis_men_B* 1.89, 62 *Neis_men_A* 0.12, 63 *Nitr_eur* 0.47, 64 *Nost_sp* 0.00, 65 *Ocea_ihe* 0.96, 66 *Past_mul* 0.33, 67 *Pire_sp* 0.17, 68 *Proc_mar* 0.94, 69 *Pseu_aer* 0.31, 70 *Pseu_put* 1.83, 71 *Pseu_syr* 0.15, 72 *Rals_sol* 0.19, 73 *Rick_con* 0.00, 74 *Rick_pro* 0.00, 75 *Salm_ent* 0.86, 76 *Salm_ent* 0.51, 77 *Salm_typ* 1.04, 78 *Shew_one* 0.53, 79 *Shig_fle* 0.31, 80 *Shig_fle* 0.34, 81 *Sino_mel* 0.47, 82 *Stap_aur* 0.56, 83 *Stap_aur* 0.38, 84 *Stap_aur* 0.39, 85 *Stap_epi* 0.19, 86 *Stre_aga* 1.08, 87 *Stre_aga* 0.94, 88 *Stre_mut* 0.55, 89 *Stre_pne* 0.09, 90 *Stre_pne* 0.42, 91 *Stre_pyo* 1.14, 92 *Stre_pyo* 1.06, 93 *Stre_pyo* 1.09, 94 *Stre_pyo* 0.76, 95 *Stre_ave* 0.27, 96 *Stre_coe* 0.52, 97 *Syne_sp* 0.00, 98 *Ther_ten* 1.27, 99 *Ther_elo* 0.00, 100 *Ther_mar* 0.26, 101 *Trep_pal* 0.31, 102 *Trop_whi* 0.28, 103 *Trop_whi* 0.38, 104 *Urea_ure* 0.38, 105a *Vibr_cho_1* 1.11, 105b *Vibr_cho_2* 1.16, 106a *Vibr_par_1* 0.55, 106b *Vibr_par_2* 0.84, 107a *Vibr_vul_1* 1.61, 107b *Vibr_vul_2* 0.54, 108 *Wigg_glo* 1.02, 109 *Xant_axo* 0.29, 110 *Xant_cam* 0.09, 111 *Xyle_fas* 4.06, 112 *Xyle_fas* 1.07, 113 *Yers_pes* 0.50, 114 *Yers_pes* 0.28.

3. Results

3.1. Traditional fixed-length window analyses and CsCl gradient ultracentrifugation

The main points mentioned in Introduction are illustrated in Fig. 1, which shows intrachromosomal GC heterogeneity

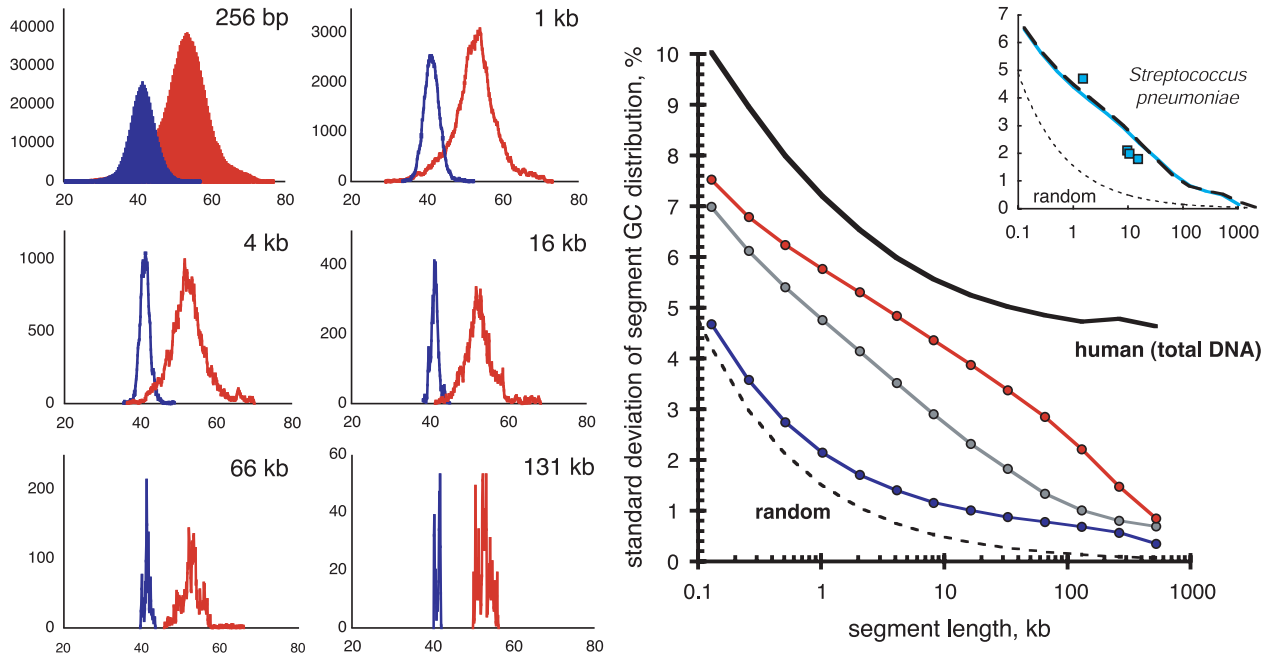

Fig. 1. Fixed-length window GC level histograms (left panel) and their standard deviations (right panel) for different species. Window sizes chosen for the histograms are successive powers of 2 bp, from 256 bp to>131 kb. A homogeneous bacterium (*C. trachomatis*; blue curves at left/bottom) and a heterogeneous bacterium (*X. fastidiosa*; red curves at right/top) are shown in both representations. The standard deviations are shown also for human (highest, black curve with plateau due to isochore structure) and for a random DNA sequence consisting of independent, identically distributed nucleotides (lowest, dashed black curve; valid for GC levels between 30% and 70%). Two historically important bacteria, *E. coli* (grey middle curve) and *S. pneumoniae* (inset) are also included for comparison. In the inset, the two entirely sequenced strains of *S. pneumoniae* are shown by the almost coinciding turquoise and long dashed lines; early experimental results (Guild, 1963) are shown by four turquoise squares.

at different scales for four bacterial species, and, for comparison, the heterogeneity of the human genome. The heterogeneities shown here were assessed using the traditional fixed-window approach. The difference between a homogeneous (*Chlamydia trachomatis*) and a heterogeneous genome (*Xylella fastidiosa*) is striking. Whereas *C. trachomatis* is only slightly more heterogeneous than a random sequence, at all scales from 100 bp to 100 kb, *X. fastidiosa* is much more heterogeneous, attaining standard deviations close to those of human DNA at scales around 5 kb.

The widely accepted belief that all bacteria tend to exhibit a low level of intrinsic GC heterogeneity (cf. Lawrence and Ochman, 1997; Ochman et al., 2000 and references therein) may have become fixed by early comparisons between prokaryotes and mammals. Such comparisons unfortunately gave an exaggerated picture of the difference between bacteria and mammals. Following tradition, comparisons typically chose calf thymus to represent mammals (Chargaff et al., 1953; Meselson et al., 1957; Sueoka, 1959, 1961, 1962; Sueoka et al., 1959; Kit, 1962), and scales (molecular weights) of or above 10–15 kb. At these scales, the intermediate heterogeneities of *E. coli* and *S. pneumoniae* paled in comparison to the calf thymus. Later analyses showed that calf had been a fortuitous choice, since it has one of the widest CsCl profiles of all mammals (Thiery et al., 1976; Sabeur et al., 1993; C. Douady et al., unpublished results). Furthermore, its excessive heterogeneity is not a property of its single-copy DNA, but instead a consequence of unusual quantities of highly repetitive satellite DNA, which account for one quarter of its genome (Filipski et al., 1973; Macaya et al., 1978).

Fig. 1 illustrates the behavior of GC heterogeneity of fixed-length fragments when the fragment length is increased. The inset shows separately the behavior for a bacterial genome that was studied over many decades, and of which two strains have been sequenced, *S. pneumoniae* (also called pneumococcus or *Diplococcus pneumoniae*). Already its analysis by early density gradient experiments (squares; Guild, 1963) showed that its standard deviation, which can be seen to follow a roughly exponential decrease (i.e. a straight line on a semi-logarithmic plot), remains well above the curve expected for a random sequence of independent, identically distributed, or "i.i.d.", base pairs. This difference from a random sequence persists until one reaches very large (Mb) scales. The early results from ultracentrifugation experiments, obtained within a few years after Meselson, Stahl and Vinograd had introduced CsCl gradient ultracentrifugation in 1957, already demonstrated that the compositional homogeneity of real DNA cannot be assessed by taking an extrinsic scholastic reference such as the i.i.d. sequence (see, e.g., Rolfe and Meselson, 1959). Homogeneity or heterogeneity must instead be measured relative to intrinsic standards, such as the narrowest and the

widest GC distributions in a relevant group of genomes. For example, *X. fastidiosa* is very heterogeneous when the relevant group is the set of prokaryotic genomes, yet becomes much more homogeneous when the group includes the eukaryotic genomes.

The sequencing of entire genomes gave a clear confirmation that GC contrasts have different forms and extents in different prokaryotic taxa. In some cases, such as *Mycoplasma*, large differences in GC heterogeneity and/or in the organization of GC variation are observed within a single genus (Kerr et al., 1997).

3.2. Generalizing the fixed-length window approach

Using the standard deviation of GC levels among fixed-length segments to measure compositional heterogeneity has two important advantages. First, the standard deviation is a measure of dispersion that has been extensively studied for over a century, and its biases and anomalies are well understood (e.g. sensitivity to outlier values, such as are sometimes caused by satellites or rDNA). Second, when sequence information is lacking, heterogeneities can be measured experimentally using CsCl gradient ultracentrifugation. CsCl absorbance profiles yield good estimates of the fixed-length fragment distribution of a genome's GC levels, and of its standard deviation.

Every measure of genomic heterogeneity has drawbacks, however. As well as being sensitive to spurious outlier DNA, the standard deviation does not do justice to chromosomes that are organized into relatively homogeneous regions of very different sizes. In this case, a plot of standard deviation vs. window length, as in Fig. 1, gives only a very indirect picture. We therefore used several measures. They allowed us to explore different ways of visualizing GC heterogeneity, to confirm robustness of the wide range of intrachromosomal heterogeneities observed, and to see that a fair number of species consistently occur among the most heterogeneous (or homogeneous) ones observed.

The usual fixed-length window approach can be generalized in at least three ways. A first way is to keep fixed windows, but to present all relevant window sizes in the same plot (wavelets), and/or to allow weighting of the nucleotides inside each window according to their locations within the window. A second way is to partition the entire sequence, not into fixed-length segments, but into variable-length segments whose lengths are chosen by the GC variation itself. A third way is to find a representation of heterogeneity that reveals the contribution from codon position differences, and the contribution from heterogeneities above the scale of genes.

We sketch the first way here. Traditional moving-window plots use a box filter: each GC base pair within a fixed length window is counted once. From such windows' GC levels one can recover their standard deviation, the traditional measure of heterogeneity. Alternative filters can be designed, which weight the GC counts according to their location in the window. The weighting can be made to follow a Gaussian form, in order to emphasize the center of the window (Fig. 2, top panel), or it can be made to follow the Gaussian's first derivative, in order to emphasize local contrast (Fig. 2, bottom panel; wavelet method). When window counts are represented by color or brightness rather than by the height of a line plot, a single plot can accommodate the full range of possible window sizes. From such plots one can quickly locate the homogeneous and heterogeneous regions present in a genome at any scale, and assess their stability when the window size changes. Such plots allow quick comparisons between the overall contrast levels present in different genomes, as can be seen in Fig. 2. They also localize the regions that cause the contrasts, and indicate if they are confined to a single locus or ubiquitous. *X. fastidiosa* (right panel) contains one striking, very GC-rich (blue) locus, but also shorter, similarly GC-rich regions throughout its genome. Even after excluding these regions, the overall heterogeneity remains distinctly higher than in *C. trachomatis*. The same conclusion was reached via segmentation analysis (see below).

3.3. Segmentation methods

A second alternative to traditional fixed-length window scans is a rigorous partitioning, using criteria that involve the GC variation within the sequence, into relatively homogeneous segments of different sizes.

A genome can be partitioned in this way by recursive top-down segmentation (Bernaola-Galván et al., 1996). First, the entire sequence is searched to find a partition into two parts that would maximize their GC contrast. A chosen significance threshold or stopping criterion then decides if the contrast is large enough to justify the proposed partitioning (see Materials and methods). The process is iterated until the contrast between adjacent (sub)segments becomes weaker than the threshold. Although the stringency of the stopping criterion is related to the average lengths of the segments, the segment lengths can span a wide range (see Oliver et al., 2001, 2002 for examples). This approach therefore yields a quite different picture of a genome's compositional variation than when fixed segment lengths are used. It also leads to a natural measure of the GC heterogeneity or complexity for the sequence's segments, called the sequence compositional complexity or SCC (Román-Roldán et al., 1998; see Materials and methods).

We applied here a recently developed segmentation variant (Oliver et al., 2004), which incorporates a key improvement: short-scale sequence heterogeneity below 3 kb is filtered out. This makes the method suitable for locating boundaries between fairly homogeneous, isochore-like regions. The significance level for which we report results here (95%; see Materials and methods) corresponds to what we consider a reasonable tradeoff between

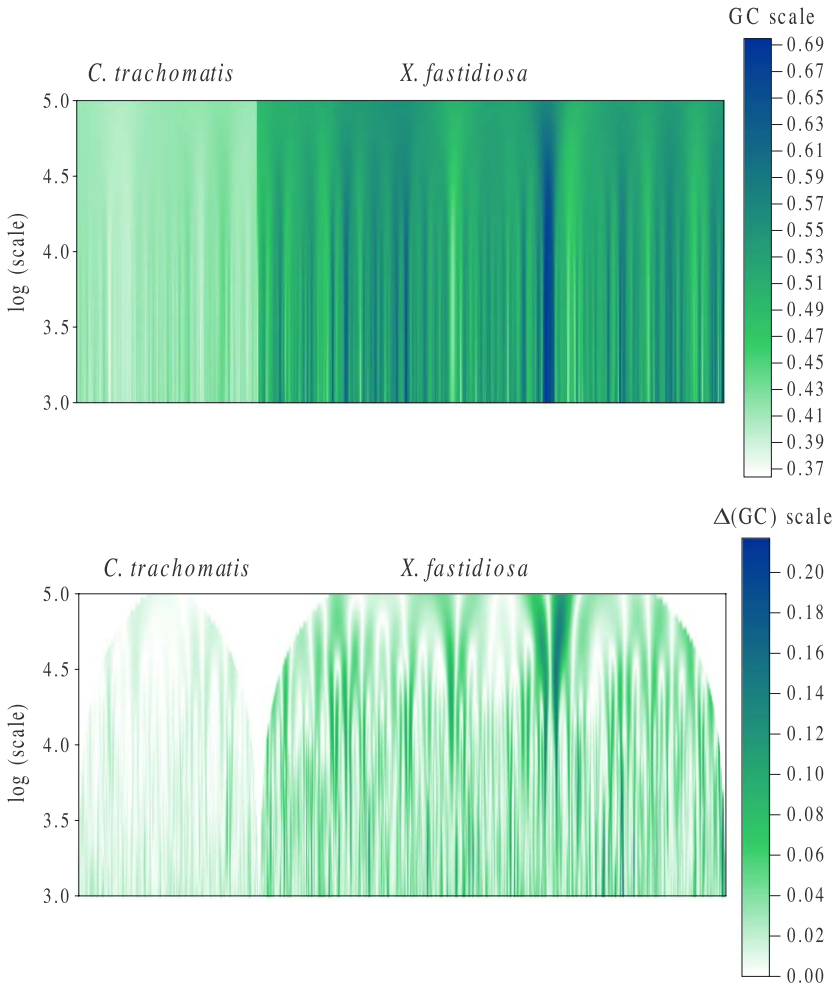

Fig. 2. Fixed-length moving window plots of the GC heterogeneity of two completely sequenced bacterial genomes, *C. trachomatis* and *X. fastidiosa*, using alternatives to the traditional box filter. Top panel: Moving window plots using a Gaussian filter. GC counts are given smaller weights if they are located farther from the window's midpoint, according to a Gaussian weight curve. GC content (expressed as a fraction of 1) is shown as color (dark blue in GC-rich regions). Each horizontal transect is a moving-window plot for a different window size; along the vertical axis, window sizes from 1 to 100 kb are labelled by their logarithms to base 10. Bottom panel: Moving window plots using wavelets to highlight local contrasts. The weight curve is the first derivative of a Gaussian (Arneodo et al., 1996), so that dark blue vertical transects indicate sharp drops or rises in GC level within the window.

sensitivity and specificity: at this stringency, none of the archaeal chromosomes, and 20 of the 123 bacterial chromosomes, remained unsegmented (SCC = 0).

Compositional segmentation maps are shown in Fig. 3 (left) for eight heterogeneous and two homogeneous genomes. They consolidate the result that was exemplified above for *X. fastidiosa*: in several heterogeneous genomes, the heterogeneity is not just due to one or two small regions having exceptional GC levels, but it is instead a genome-wide phenomenon. The figure also shows a scatterplot of the corresponding SCC heterogeneity values versus genomic GC (right). With respect to SCC, the species with the most heterogeneous chromosomes include *Aeropyrum pernix* (triangle 1), *Deinococcus radiodurans*, chromosome II (square 35b), *X. fastidiosa* (square 111), *Enterococcus faecalis* (square 36), *Mycoplasma genitalium* (square 57), *B. subtilis* (square 7), *Neisseria meningitidis* (square 61), *Pseudomonas putida* (square 70), *Pyrobaculum aerophilum* (triangle 9), *Vibrio vulnificus* (chromosome I, square 107a) and *Halobacterium* (triangle 3).

It can be seen that heterogeneity shows no clear relation to overall chromosomal GC level, at these stringencies, either in archaea or in bacteria. Furthermore, there is no obvious grouping of hyperthermophiles. These general observations are also valid when the standard deviation of fixed-length windows is used as the heterogeneity measure.

3.4. Correlation methods: GC correlograms as a representation of compositional heterogeneity

Positive compositional (auto)correlations are ubiquitous in DNA, and, where long-ranged, they can appreciably broaden the distributions of GC levels, compared to sequences of independent and identically distributed base pairs. The heterogeneity of a genome is thus related to the

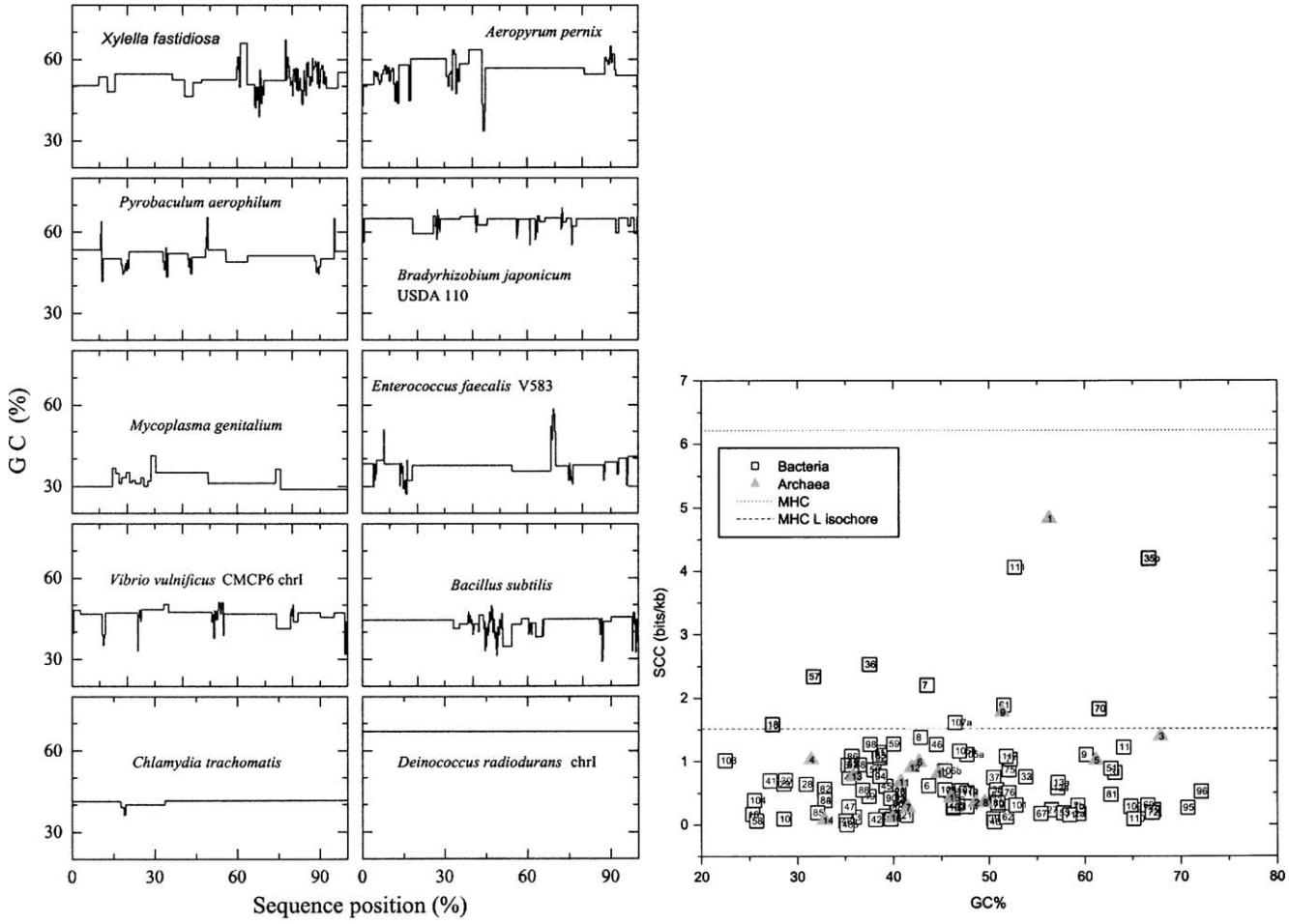

Fig. 3. Compositional segmentations of sequenced prokaryotic genomes and their associated GC heterogeneities as measured by sequence compositional complexity (SCC). Left: Segmentations for eight heterogeneous and two homogeneous genomes (bottom panels). Right: SCC values for segmentations of sequenced archaeal (triangles) and bacterial genomes (squares). Numbering of genomes is as in Materials and methods. The SCC of a segmented sequence is the difference between the sequence's informational (Shannon) entropy and the sum of all its segments' entropies, weighted by their respective lengths (Román-Roldán et al., 1998). Horizontal lines in the scatterplots indicate SCC values of human sequences: the very heterogeneous MHC region in chromosome 6, spanning three adjacent isochores (top line), and the GC-poor (L) central, classical MHC class II isochore (bottom line). Note that the SCC levels of several heterogeneous prokaryotic genomes are higher than those of human L isochores.

correlations among its base pairs: the longer the range of the positive correlations, the wider the range of scales at which heterogeneity is increased. Correlograms show the correlation between base pairs as a function of the distance between them. They can be estimated via indirect methods such as Fourier analysis or, preferably, they can be calculated directly.

Fig. 4 shows the correlograms for the genomic sequences of *A. pernix* and *Borrelia burgdorferi*. This representation of genomes' heterogeneities is in principle equivalent to the standard deviation representation, exemplified in Fig. 1 (see Materials and methods; an explicit check confirmed that the standard deviations calculated from these two correlograms exactly match the values from the GC distributions). Correlograms present, however, a different view of the GC heterogeneity, and can dissect it to reveal the contribution from codon position differences.

From Fig. 4, it can be seen that the correlograms split into two branches. The top branch shows correlations between locations separated by multiples of 3 bp; the bottom branch shows the correlations for all other distances. Unlike mammalian genomes, where the proportion of coding regions is only about 3%, prokaryotic genomes consist largely or almost entirely of coding DNA. The two-branch structure of the correlograms has, therefore, a simple interpretation: at short (intragenic) scales, most of the genome's GC heterogeneity is explained by positive correlations among base pairs at identical codon positions (1–1, 2–2 and 3–3), which decrease exponentially as the distance increases (see right insets in Fig. 4). This behavior is found also in other sequenced prokaryotes.

The lower branch of the correlogram represents the (codon-averaged) correlations between different codon positions (1–2, 1–3, 2–3, 2–1, 3–1 and 3–2), and typically yields mildly negative values for small distances. Excep-

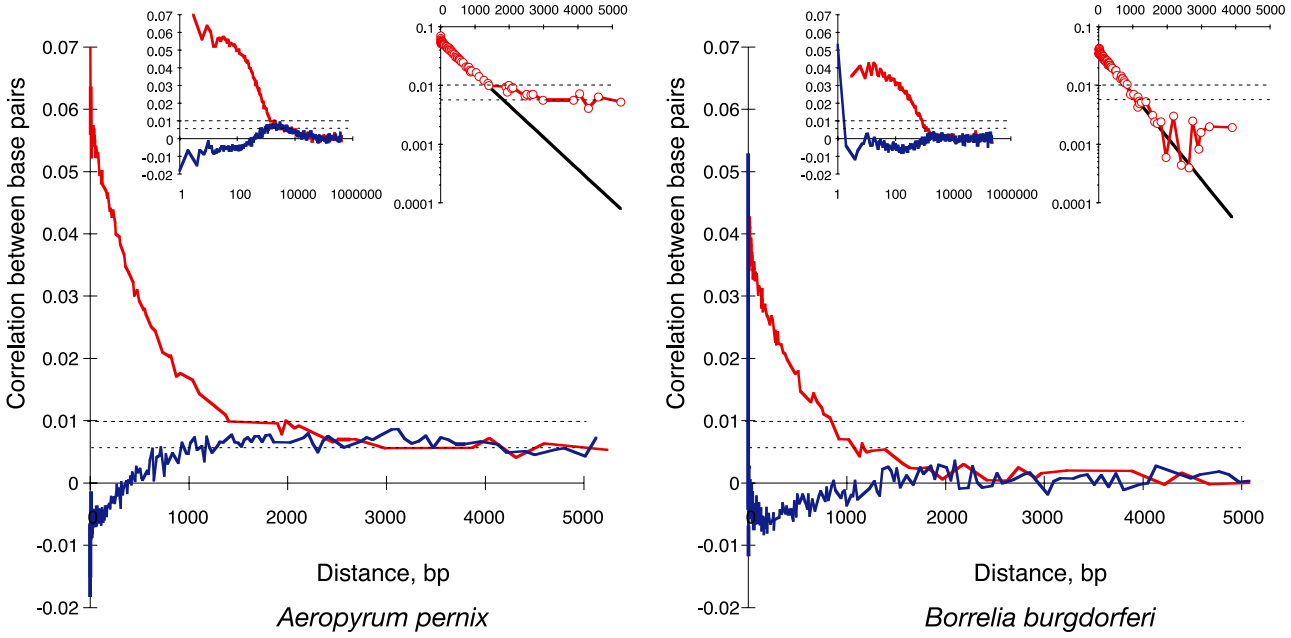

Fig. 4. Inter-base pair (GC/AT) autocorrelation functions (correlograms) at distances from 1 bp to 1 Mb, for a relatively heterogeneous (*A. pernix*; left panel) and a relatively homogeneous prokaryotic genome (*B. burgdorferi*; right panel). The top branch in each plot shows correlations between locations separated by integral multiples of 3 bp, which correspond to identical codon positions or phases; the bottom branch shows correlations for the other distances. Inset plots use logarithmic scales for distances (abscissa), or, alternatively, for the correlation values (ordinate). The latter plots show the top branch only, together with the regression line, the distance where the top branch deviates from an exponential decrease, and the correlation at that distance (horizontal dashed lines show the limits of the correlation estimate for *A. pernix*). At a farther distance, the two branches merge. This distance estimates the extent of the heterogeneity caused by differences among codon positions. The corresponding correlation value quantifies heterogeneity above the genic scale.

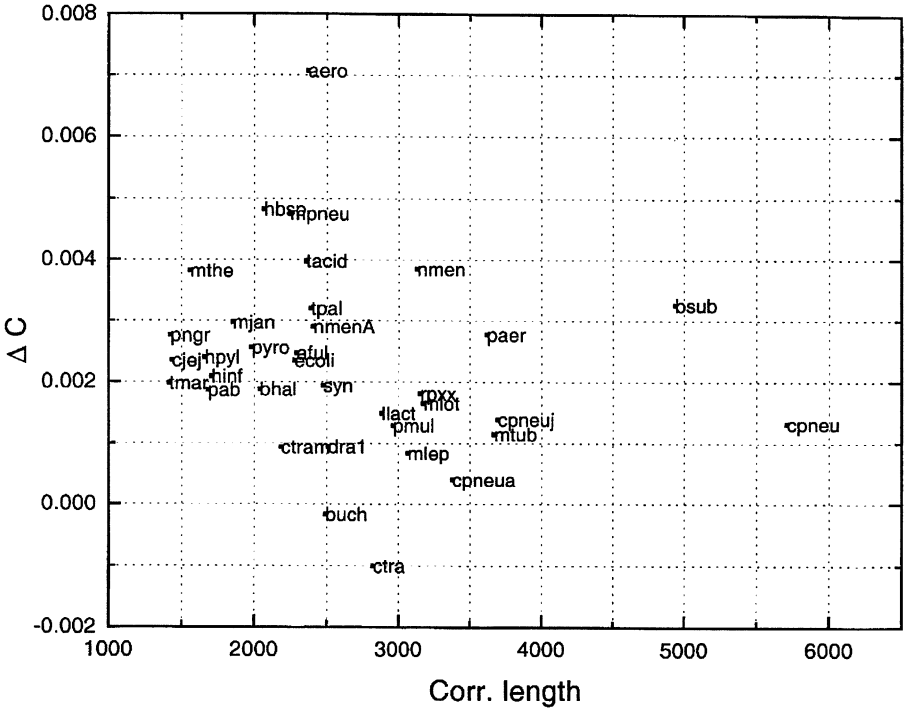

Fig. 5. Scatterplot showing two heterogeneity measures derived from correlograms, for a sample of 30 sequenced genomes. GC correlograms such as those of Fig. 4 define a distance at which the two branches of the correlogram meet. This distance or length (horizontal axis), and the corresponding correlation value (vertical axis), are parameters that can be used to assess and interpret GC heterogeneity. A comparison with human, as shown in other figures, would not be relevant here because of the low coding content of mammalian genomes. Thermophilic archaea: *Aeropyrum pernix*, aero; *Archaeoglobus fulgidus*, aful; *Halobacterium* sp., hbsp; *Methanobacterium thermoautotrophicum*, mthe; *Methanococcus jannaschii*, mjan; *Pyrococcus abyssi*, pab; *Pyrococcus horikoshii*, pyro; *Thermoplasma acidophilum*, tacid; Thermophilic bacterium: *Thermotoga maritima*, tmar.

tions are the transiently positive correlations between adjacent base pairs ($d=1$) and/or between base pairs that are separated by a single turn of the double helix ($d=10,11$).

The two branches merge where the overall influence of codon position differences becomes negligible. At that location they will indicate the distance at which the codon influence becomes small, and the correlation at that distance. Different genomes show different behaviors in this respect. In some compositionally homogeneous species, the exponential decrease exhaustively characterizes the top branch of the correlogram, so that the correlations vanish for distances longer than the typical lengths of genes ($\sim$ 2 kb). This is not the case for heterogeneous species such as *A. pernix*: in this example, the top branch of the correlogram departs from an exponential decrease at approximately 1.5 kb, yet appreciable positive correlations persist until at least 20 kb (see the left insets in Fig. 4, which show an expanded horizontal axis).

By plotting the distance at which the branches meet on the abscissa of a scatterplot, and the autocorrelation at that distance on the ordinate, we obtain a third view of GC heterogeneity that is not captured by the other methods presented above. Fig. 5 shows such a scatterplot for a sample of 30 sequenced species. In this sample, the top left sector contains all 9 (hyper)thermophiles in the sample, but only 7 of the 21 mesophilic species; more sequencing should reveal if such thermophile clustering is a general tendency.

The locations of the species on the scatterplots are determined by at least two factors: the lengths of coding regions in the genomes and the differences among the GC levels of the codon positions. Other effects co-determine the two measures displayed. Indeed, the distance at which the branches meet is apparently not just a simple function of typical gene lengths: the hyperthermophile *M. janaschii* has a quite different distribution of coding lengths from that of two other hyperthermophiles, *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus* (Li, 1999), yet it clusters with them in the scatterplot. Conversely, neither coding lengths nor codon position differences are likely to vary much among strains of *Chlamydia pneumoniae*, but the branching point is apparently at a farther distance for one of these strains (cpneu) than for the other two (cpneua, cpneuj), although this branching distance becomes difficult to estimate when it is large.

4. Discussion

4.1. Robust heterogeneity is found in archaea and bacterial chromosomes spanning a wide compositional range

We have used several measures of intragenomic GC heterogeneity: the variation among fixed-length windows, the sequence compositional complexity at different stringency levels, and the values of the GC correlograms where top and bottom branches meet.

The results obtained show that whereas some prokaryotes are consistently homogeneous (such as *C. trachomatis*, the two *Rickettsia* species or *Campylobacter jejuni*), others are consistently heterogeneous (such as *X. fastidiosa*, *A. pernix* or *Halobacterium*), as assessed by different measures at several scales and stringencies.

Already early ultracentrifugation results on bacterial GC heterogeneity (Sueoka, 1959; Guild, 1963; Yamagishi, 1974) had indicated that long regions of distinct base composition should exist within the genomes of some bacteria. We can now confirm the existence of such regions at the sequence level.

At very small scales ($\ll$1 kb), prokaryotes' GC heterogeneities depend on the mean GC level: the dependence follows a bow-shaped curve with a maximum at 50% GC (see also Liò, 2002). At scales of a few base pairs, such GC dependence is essentially trivial (the standard deviation approaches that of a binomial distribution). A question is what happens at larger scales. Early data sets consisting of very few species had suggested a dependence of this type also at scales of 10 kb or more (Cuny et al., 1981). The large sample of sequenced bacteria and archaea no longer suggests such a systematic GC dependence at these scales, whether we use the SCC or the traditional standard deviation as our heterogeneity measure. This situation is not what is found in mammals and birds, for example: at the same scales, their GC heterogeneities increase rapidly as GC levels rise from 30% to 50% (Cuny et al., 1981; Bernardi, 2001 and references therein).

There is not an obvious relation between genome size and genome compositional complexity (SCC) at our chosen stringency level (3 kb/95%), i.e., at a level that corresponds to GC variation above the typical lengths of genes. For example, *Mesorhizobium loti* has one of the largest genome sizes among the sequenced species (above 7 Mb) and contains about 8000 genes (Kaneko et al., 2000), yet its SCC, 0.89, is not among the highest.

A number of factors may influence intrachromosomal heterogeneities, to different extents, so it may remain difficult to generalize about their relative contributions until more genomes have been sequenced. For example, the high complexity (SCC = 2.53) of *E. faecalis* may be linked to the fact that it has one of the highest proportions of mobile elements observed in a bacterial genome (Paulsen et al., 2003). Segmentation differences among different strains of some species (e.g. in *E. coli*, *Salmonella enterica*, *Streptococcus agalactiae*, *Streptococcus pyogenes* and *Yersinia pestis*) may be due to insertion sequences or pathogenicity islands. On the other hand, the presence of such pathogenicity islands (e.g. in *Helicobacter pylori* or *Pseudomonas aeruginosa*), or of symbiotic islands (e.g. in *M. loti*), does not always lead to exceptionally high SCC values. The complexity differences between the two strains of *X. fastidiosa* may be related to phage-associated chromosomal rearrangements and deletions in this species, to its unusually broad host range, and to the diverse disease

phenotypes it induces (Van Sluys et al., 2003). The high compositional heterogeneity in the archaeon *A. pernix* (4.83 bits/kb), and its correlations extending until 20 kb (see Fig. 4), may be related to the large number of duplicated ORFs in this genome (Kawarabayasi et al., 1999).

The chromosomes within a genome often show very different compositional complexities. Compare for example the chromosome pairs within *Agrobacterium tumefaciens*, *Brucella suis*, *Vibrio cholerae*, *Vibrio parahaemolyticus* or *V. vulnificus*. The apparently extreme example of such genomic diversity is given by *D. radiodurans*: while chromosome I remains unsegmented (SCC = 0), chromosome II shows one of the highest compositional complexities (SCC = 4.19) we have observed in a prokaryotic chromosome.

In the context of *Deinococcus*, it should however be pointed out that our chosen coarse-graining and segmentation parameters will not always recognize all large-scale compositional heterogeneity. Thus, genes and their third positions in chromosome I of *D. radiodurans* reveal a significantly lower GC level of one large segment, when compared with the rest of the chromosome (this example was pointed out by a referee). The GC-poorer segment, at about 700–1250 kb from the replication origin, is missed by the segmentation algorithm unless the coarse-graining is substantially reduced. The example illustrates that large-scale genomic heterogeneities will tend, if anything, to be higher than the SCC values given here, and that although the stringency criteria that we used will work well for many genomes, they may not be optimal for all genomes.

4.2. How much intrachromosomal heterogeneity is due to integrated prophages?

Differences in heterogeneity between genomes can result when phages of different GC levels integrate into a chromosome. Such contributions to overall intrachromosomal heterogeneity do not appear to be dramatic, as a general rule: excising DNA of known, sequenced phages in the list of Rocha and Danchin (2002; http://abraxa.snv.jussieu.fr/~erocha/scarceGC/phages.htm) from their sequenced host chromosomes had, in general, relatively little influence on the segmentation or heterogeneity (SCC), at the stringency used here. More exhaustive studies will, however, be needed to quantify the heterogeneity contribution of phages.

4.3. How much intrachromosomal heterogeneity is due to intergeneric lateral transfers?

A principle, postulated to be generally valid for bacteria and archaea, has been discussed in much detail in recent literature. The proposal is that intrachromosomal heterogeneity, at genic (Lawrence and Ochman, 1997, 1998, Ochman et al., 2000) and/or supragenic scales (Garcia-Vallvé et al., 2000), should correlate with the proportion of DNA that the genome has recently acquired via lateral transfer from genomes having a different (mean or modal) GC level. In other words, if there is appreciable GC heterogeneity in some chromosomes, the idea is that it is largely caused by lateral transfer from other species or genera. The proposal is that most compositionally anomalous DNA will eventually "ameliorate", so that its GC approaches the modal GC of the new host. Although this is a stimulating idea, its hypothetical nature should not be forgotten: estimates of amelioration rates can be calculated, yet they rely on assumptions that have been suggested, not proved, by tracking orthologous genes in a few taxa. Variants of this hypothesis might include 'regional amelioration', in which foreign DNA might acclimatize to the local GC of a region in which it incorporates, or selection and/or excision when invading DNA disrupts the local GC (and, e.g., hampers transcriptional or other processes for which the local GC may be optimal).

The idea that chromosome-wide GC heterogeneity is primarily caused by intergeneric lateral transfer was an ex novo hypothesis. In particular, it could not have followed from, or been strongly suggested by, any feature that was evident from the initial characterizations of bacterial genomes via buoyant density gradients in the late 1950s or early 1960s (cf. Lawrence and Ochman, 1997): such initial measurements of species' GC ranges via ultracentrifugation necessarily included both native and laterally acquired DNA, and the experiments could not distinguish the two types of DNA.

Some advocates of GC contrast as a tool for detecting lateral transfers have, in essence, suggested that recently acquired foreign DNA is responsible, e.g., for most of the large, GC-poor tail of the *E. coli* genome's asymmetric GC distribution. This asymmetry is evident at a wide range of scales, up to about 70 kb. The idea is that if the foreign DNA were absent, the genome would be much more homogeneous. More precisely, the GC or genic GC_3 distribution would be thinner and almost symmetric (Lawrence and Ochman, 1998; ftp://ftp.pitt.edu/dept/biology/lawrence/eco.txt). Estimates, based on GC contrasts, of the amount of recent transfer DNA in *E. coli* range from 12.8% to over 25%, depending on the criteria or extrapolations used (Ochman et al., 2000; Lawrence and Ochman, 2002; Garcia-Vallvé et al., 2000; Ragan, 2002; see, however, Daubin et al., 2003a for another viewpoint).

To identify foreign DNA via GC contrasts, authors often proceed by first fixing limits beyond which GC is considered 'anomalous'. Limits are occasionally tailored to a particular genome such as *E. coli* (Lawrence and Ochman, 1997). Relative criteria are more often used, and aim for species-independence by taking the genomes' compositional heterogeneity into account. They are typically defined in terms of standard deviations or standard errors (Lawrence and Ochman, 1998; Garcia-Vallvé et al., 2000). Such relative criteria are unfortunately not self-consistent, since they will give rise to an infinite regress. When recursively applied to any genome, they will continue to erode DNA

that had been labelled as native DNA by previous iterations, and the process will not converge to a stable, native genome core until no native DNA is left.

The extent to which lateral transfers can be recognized via GC contrasts is still largely unsolved. Few GC-based predictors have been subjected to the kind of quantitative scrutiny that is routinely used when evaluating gene prediction algorithms (although Lawrence and Ochman, 2002 or Doolittle, 1999 contains ideas for cross-evaluation tests that could be used, e.g., together with traditional parameters such as specificity and sensitivity). If some bacteria frequently exchange large parts of their genomes, this would have important implications, not only for the understanding of GC variation. The fate or history of a species would be very sensitive to its potential donor genomes, and physical environment or ecology could be strong determinants of a genome's evolution (Doolittle, 1999; see, however, Daubin et al., 2003b).

4.4. How much intrachromosomal heterogeneity was created by intrageneric 'transfer' of extrachromosomal (plasmid) DNA?

It has been noted before that plasmids, and genes that evolve in them, often have quite different (typically lower) GC levels than chromosomes of the same species, i.e., than another element that is present in the same cell (Rocha and Danchin, 2002; White et al., 1999; Falkow, 1975). Interestingly, the likely transfers that Lawrence and Ochman (1997) discussed in detail, and used to consolidate their intergeneric transfer interpretation of GC heterogeneity, were not of the chromosome-to-chromosome type, but of the plasmid-to-chromosome type. Furthermore, the *Shigella* plasmid DNA in question would have been a GC-poor anomaly, even if it had been found in the main chromosome of its own species. In other words, typical GC-based recognition criteria for lateral transfers would probably have misclassified this DNA as arriving from an alien species, whereas it would in fact have been just plasmid DNA from the same species. Another example of plasmid-to-chromosome transfer may be the GC-poor regions in chromosome II of *D. radiodurans* (see the segmentation map at our website). Much attention has been given to intrachromosomal GC contrasts arising from possible lateral transfers of alien (extra-generic) DNA. Less attention has been given to GC contrasts created by chromosomal DNA recombining with native plasmid DNA, or with plasmid DNA received from closely related species in the same genus.

Do species that more frequently undergo conjugation, and/or exchanges between plasmid and chromosome, tend to be more heterogeneous in GC? Conversely, do other species tend to be more homogeneous? In full generality, these questions seem difficult to address at present. More quantitative estimates of conjugation frequencies may be necessary before we can answer them. In particular, we cannot yet expect to categorically rank the sequenced taxa by conjugation frequency. This being said, the enterobacteria include well-studied species that are prone to undergo conjugation, and they (together with other Section 5 taxa of the current Bergey classification) tend to exhibit higher intrachromosomal heterogeneities than many other groups of bacteria. Alternatively, such heterogeneities in enterics may be a result of a shared property other than high conjugation frequency or recombinogenicity (sexuality).

4.5. How much intrachromosomal heterogeneity may be truly endogenous?

One question has been only occasionally addressed in the recent literature: could some of the factors shaping GC contrasts in prokaryote chromosomes be truly endogenous? In particular, could their causes be similar to those shaping and maintaining the much more pronounced GC contrasts (i.e., isochores) in vertebrate chromosomes? Such contrasts have been conserved for hundreds of Myr, yet lateral transfer from remote species is apparently very rare and could not possibly explain the contrasts, for example, in human (see, e.g., Stanhope et al., 2001). The contrasts in human are also too large to be explained by contributions from repetitive DNA (Pavlícek et al., 2001). Could many of the intrachromosomal contrasts in bacteria and archaea have arisen endogenously as well?

Along these lines, Sueoka (1992) has offered the following perspective: "The small[er] intragenomic heterogeneity of G + C contents in unicellular organisms such as *E. coli* and yeast [compared to some higher metazoans] . . . may be regarded as a simple form of isochores (domains of unique G + C contents). . . . It is a likely possibility that the intragenomic heterogeneity of DNA G + C content is a ubiquitous phenomenon from bacteria to mammals and that the only difference is the extent of heterogeneity between unicellular and multicellular organisms".

For instance, there could be regional selection to keep GC high, or low, in certain parts of a genome, or in certain genes or gene cassettes. One well-known, exemplary case is given by the (albeit non-coding) rDNA modules in several species. Hyperthermophiles apparently experience strong selective pressure to keep their ribosomal RNA (rRNA) genes at high GC levels. The high GC levels ensure the correct secondary structures for the rRNA at high temperatures, and are maintained almost without regard for the GC level of the rest of the genome. Indeed, GC levels of rRNA genes in hyperthermophiles correlate with temperature rather than with the overall genomic GC content (Galtier and Lobry, 1997; Hurst and Merchant, 2001; see also Mandel, 1969 and references therein). rDNA often has a much higher GC than the rest of the genome in which it resides, so that a high compositional contrast must be maintained between very GC-rich regions of about 3 kb, containing rRNA genes, and GC-poorer flanking regions. GC distributions of 1-kb fragments then show a distinct rDNA peak to the right of the 'main band'. This phenom-

enon can be seen in archaeal hyperthermophiles such as *P. abysii*, *A. fulgidus* or *Methanococcus jannaschii*, and in bacterial hyperthermophiles such as *A. aeolicus* and *T. maritima*. Similarly, lower GC levels of the region surrounding the terminus of replication, in several prokaryotes (Daubin and Perrière, 2003; Guindon and Perrière, 2001; Deschavanne and Filipski, 1995), might simply be a footprint of repair-related GC erosion as these authors suggest, or they might represent some advantage, e.g. for more efficient replication. Temperature-dependent constraints, to allow correct structures, have been shown experimentally for origin of replication (*ori*) sequences in bacteriophage G4 and in the mitochondrial genome of yeast (see Goursot et al., 1988 and references therein). There appears to be no reason, a priori, why GC constraints could not also conserve the GC content of some regions that contain protein-coding genes.

Interestingly, some phages and plasmids (i.e., episomes) display a striking GC mosaicism inside their sequence. The best-known example is perhaps coliphage lambda, but a similar mosaicism of a 11–18-kb GC-rich region contrasting with a (≈ 10%) poorer GC-poor region is found in R, F and other plasmids (including the virulence plasmid O157:H7) of *E. coli* and of the GC-poorer enteric genus *Proteus*. This general observation was already made about 40 years ago (see Falkow and Cowie, 1968, Falkow, 1975 and references therein), and can now be confirmed in detail for several entirely sequenced episomes. The mosaicism in these plasmids and phages is apparently not just random, but corresponds to their functional partitioning into DNA segments or blocks of genes having different roles or expression/emergence times (e.g., capsid assembly genes in coliphage lambda, leading region in plasmids). Such striking mosaicism led to an early suggestion (Falkow and Cowie, 1968) that evident intramolecular heterogeneity may be a general property of episomes, and may provide clues to their evolutionary history. Our segmentation results indicate that, where sequences are available, phages and plasmids tend to be more heterogeneous than the much larger chromosomes (see our website, http://bioinfo2.ugr.es/prok/).

In summary, it is conceivable that compositional constraints, having identifiable functional correlates, may turn out to be responsible for a sizeable proportion of the larger GC contrasts found in archaeal and bacterial chromosomes. It will be a challenging task to tease apart the roles that such compositional constraints, intrageneric plasmid-chromosome exchanges, intergeneric transfers and other factors are likely to play in shaping intrachromosomal GC variation in prokaryotes. Tools such as the ones presented here may render this task feasible.

Acknowledgements

We thank Guy Perrière and Eduardo Rocha for helpful comments and suggestions when these results were first presented in Sorrento, as well as three anonymous referees. We also thank Wentian Li for helpful discussions. P.B. and J.L.O. thank Giorgio Bernardi for his warm hospitality at the Laboratorio di Evoluzione Molecolare in Naples, which allowed us to begin this collaboration. P.B., J.L.O and P.C. also acknowledge the financial support of the Spanish Government (Grant nos. BIO99-0651-CO2-01, BIO2002-04014-C03 and BFM2002-00183).

References

Arneodo, A., Aubenton-Carafa, Y., Bacry, E., Graves, P., Muzy, J., Thermes, C., 1996. Wavelet based fractal analysis of DNA sequences. Physica, D 96, 291–320.

Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys. Rev., E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics 53, 5181–5189.

Bernaola-Galván, P., Ivanov, P.Ch., Amaral, L.A.N., Stanley, H.E., 2001. Scale invariance in the nonstationarities of human heart rate. Phys. Rev. Lett. 87, 168105.

Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L., 2002. Study of statistical correlations in DNA sequences. Gene 300, 105–115.

Bernardi, G., 2001. Misunderstandings about isochores. Part I. Gene 276, 3–13.

Chargaff, E., Crampton, C., Lipshitz, R., 1953. Separation of calf thymus deoxyribonucleic acid into fractions of different composition. Nature 211, 289–292.

Clay, O., 2001. Standard deviations and correlations of GC levels in DNA sequences. Gene 276, 33–38.

Cruveiller, S., Jabbari, K., Clay, O., Bernardi, G., 2003. Compositional features of vertebrate genomes for checking predicted genes. Brief. Bioinform. 4, 43–52.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes. I. Preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 115, 227–233.

Daubin, V., Perrière, G., 2003. G+C3 structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. 20, 471–483.

Daubin, V., Lerat, E., Perrière, G., 2003a. The source of laterally transferred genes in bacterial genomes. Genome Biol. 4, R57.

Daubin, V., Moran, N.A., Ochman, H., 2003b. Phylogenetics and the cohesion of bacterial genomes. Science 301, 829–832.

Deschavanne, P., Filipski, J., 1995. Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. Nucleic Acids Res. 23, 1350–1353.

Doolittle, W.F., 1999. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. Science 286, 1443.

Falkow, S., 1975. Infectious Multiple Drug Resistance. Pion, London.

Falkow, S., Cowie, D.B., 1968. Intramolecular heterogeneity of the deoxyribonucleic acid of temperate bacteriophages. J. Bacteriol. 96, 777–784.

Filipski, J., Thiery, J., Bernardi, G., 1973. An analysis of the bovine genome by $Cs_2SO_4^-$ Ag^+ density gradient centrifugation. J. Mol. Biol. 80, 177–197.

Galtier, N., Lobry, J.R., 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. 44, 632–636.

Garcia-Vallvé, S., Romeu, A., Palau, J., 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10, 1719–1725.

Gionis, A., Mannila, H., 2003. Finding recurrent sources in sequences. In: RECOMB 2003 (Proceedings, Berlin, Germany), ACM, 123–130.

Goursot, R., Goursot, R., Bernardi, G., 1988. Temperature can reversibly modify the structure and the functional efficiency of *ori* sequences of the yeast mitochondrial genome. Gene 69, 141–145.

Guild, W., 1963. Evidence for intramolecular heterogeneity in pneumococcal DNA. J. Mol. Biol. 6, 214–229.

Guindon, S., Perrière, G., 2001. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. Mol. Biol. Evol. 18, 1838–1840.

Hurst, L.D., Merchant, A.R., 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. R. Soc. Lond., B 268, 493–497.

Kaneko, T., Nakamura, Y., Sato, S., et al., 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res. 6, 331–338.

Kawarabayasi, Y., Hino, Y., Horikawa, H., et al., 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res. 6, 83–101, 145–152.

Kerr, A., Peden, J., Sharp, P., 1997. Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Mol. Microbiol. 25, 1177–1184.

Kit, S., 1962. Species differences in animal deoxyribonucleic acids as revealed by equilibrium sedimentation in density gradients. Nature 193, 274–275.

Lawrence, J.G., Ochman, H., 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44, 383–397.

Lawrence, J.G., Ochman, H., 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. U. S. A. 95, 9413–9417.

Lawrence, J.G., Ochman, H., 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10, 1–4.

Li, W., 1999. Statistical properties of open reading frames in complete genome sequences. Comput. Chem. 23, 283–301.

Li, W., Bernaola-Galván, P., Haghighi, F., Grosse, I., 2002. Applications of recursive segmentation to the analysis of DNA sequences. Comput. Chem. 26, 491–510.

Li, W., Bernaola-Galván, P., Carpena, P., Oliver, J.L., 2003. Isochores merit the prefix 'iso'. Comput. Biol. Chem. 7, 5–10.

Liò, P., 2002. Investigating the relationship between genome structure, composition, and ecology in eukaryotes. Mol. Biol. Evol. 19, 789–800.

Macaya, G., Cortadas, J., Bernardi, G., 1978. An analysis of the bovine genome by density gradient centrifugation. Eur. J. Biochem. 84, 179–188.

Mandel, M., 1969. New approaches to bacterial taxonomy: perspective and prospects. Annu. Rev. Microbiol. 23, 239–274.

Meselson, M., Stahl, F., Vinograd, J., 1957. Equilibrium sedimentation of macromolecules in density gradients. Proc. Natl. Acad. Sci. 43, 581–588.

Ochman, H., Lawrence, J., Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299–304.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Oliver, J.L., Carpena, P., Román-Roldán, R., Mata-Balaguer, T., Mejías-Romero, A., Hackenberg, M., Bernaola-Galván, P., 2002. Isochore chromosome maps of the human genome. Gene 300, 117–127.

Oliver, J.L., Carpena, P., Hackenberg, M., Bernaola-Galván, P., 2004. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res. (in press).

Page, E.S., 1955. A test for a change in a parameter occurring at an unknown point. Biometrika 42, 523–527.

Paulsen, I.T., et al., 2003. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science 299, 2071–2074.

Pavlícek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Ragan, M.A., 2002. Reconciling the many faces of lateral gene transfer: response from Ragan. Trends Microbiol. 10, 4.

Rocha, E.P.C., Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18, 291–294.

Rolfe, R., Meselson, M., 1959. The relative homogeneity of microbial DNA. Proc. Natl. Acad. Sci. U. S. A. 45, 1039–1043.

Román-Roldán, R., Bernaola-Galván, P., Oliver, J., 1998. Sequence compositional complexity of DNA through an entropic segmentation algorithm. Phys. Rev. Lett. 80, 1344–1347.

Sabeur, G., Macaya, G., Kadi, F., Bernardi, G., 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. J. Mol. Evol. 37, 93–108.

Stanhope, M., Lupas, A., Italia, M., Koretke, K., Volker, C., Brown, J., 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411, 940–944.

Sueoka, N., 1959. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc. Natl. Acad. Sci. U. S. A. 45, 1480–1490.

Sueoka, N., 1961. Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. J. Mol. Biol. 3, 31–40.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U. S. A. 48, 582–592.

Sueoka, N., 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. 34, 95–114.

Sueoka, N., Marmur, J., Doty, P., 1959. Heterogeneity in deoxyribonucleic acids: II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine content. Nature 183, 1429–1433.

Thiery, J., Macaya, G., Bernardi, G., 1976. An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. 108, 219–235.

Van Sluys, M.A., et al., 2003. Comparative analyses of the complete genome sequences of Pierces disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J. Bacteriol. 185, 1018–1026.

White, O., Eisen, J.A., Heidelberg, J.F., et al., 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286, 1571–1577.

Yamagishi, H., 1974. Nucleotide distribution in bacterial DNA's differing in G+C content. J. Mol. Evol. 3, 239–242.