BMC Genomics

RESEARCH ARTICLE Open Access

Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands

Jon Bohlin[1*], Mark WJ van Passel[2], Lars Snipen[3], Anja B Kristoffersen[4], David Ussery[5] and Simon P Hardy[4]

Abstract

Background: We sought to assess whether the concept of relative entropy (information capacity), could aid our understanding of the process of horizontal gene transfer in microbes. We analyzed the differences in information capacity between prokaryotic chromosomes, genomic islands (GI), phages, and plasmids. Relative entropy was estimated using the Kullback-Leibler measure.

Results: Relative entropy was highest in bacterial chromosomes and had the sequence chromosomes > GI > phage > plasmid. There was an association between relative entropy and AT content in chromosomes, phages, plasmids and GIs with the strongest association being in phages. Relative entropy was also found to be lower in the obligate intracellular *Mycobacterium leprae* than in the related *M. tuberculosis* when measured on a shared set of highly conserved genes.

Conclusions: We argue that relative entropy differences reflect how plasmids, phages and GIs interact with microbial host chromosomes and that all these biological entities are, or have been, subjected to different selective pressures. The rate at which amelioration of horizontally acquired DNA occurs within the chromosome is likely to account for the small differences between chromosomes and stably incorporated GIs compared to the transient or independent replicons such as phages and plasmids.

Background

Horizontal gene transfer in microbial communities has been recognized as a key driver of evolutionary change in microbes [1,2]. In addition to plasmids and phages, regions within the bacterial chromosomes are assumed to have been horizontally acquired [3]. Such putatively horizontally transferred regions are termed Genomic Islands (GI). GIs originate from different sources [4] including plasmids and phages (prophages) and carry traits that have important biological phenotypes such as virulence determinants and antibiotic resistance genes. Genetic material is most readily exchanged between related genetic elements, [5]*i.e.* chromosomes exchange DNA with chromosomes, plasmids with plasmids, and phages with phages. However, this exchange is not entirely restrictive with low frequency transfer occurring between chromosomes on one hand and plasmids and phages on the other [5]. Mathematical models predict plasmids to be the predominant means of genetic variation among bacteria [5]. Based on findings from genomic signatures (and analyses of CRISPSs in bacteria [6]), phages, and viruses in general, have been found to co-evolve with their hosts [7]. Plasmids on the other hand, although sharing some similarities with their hosts, have a more different DNA composition than what would be expected compared to the hosts chromosome [8]. In fact, genomic signatures based methods reveal prokaryotic plasmid-host similarity to correlate with genomic GC content, *i.e.* the more GC rich an organism is the more compositionally similar it tends to be with its plasmid(s) [9]. GC content has also been associated with genome wide rates of mutation, where organisms of low GC content tend to have more random genomes than GC rich ones [10,11], *i.e.* the signal-to-noise ratio is lower in AT rich genomes. An organism's DNA sequence that has been subjected to numerous random mutations is

* Correspondence: jon.bohlin@nvh.no
[1]Norwegian School of Veterinary Science, EpiCentre, Department of Food Safety and Infection biology, Ullevålsveien 72, Oslo, Norway
Full list of author information is available at the end of the article

assumed to possess less information than the DNA of an organism under strong selective pressure. In other words, due to more accumulated mutations, it appears as if less information is carried by the DNA sequences of AT rich microbes compared to GC rich microbes. Thus, to test the assertion that accumulated mutations lower the information capacity we explored the use of information theory as a means of measuring information capacity in DNA sequences.

The concept of information theory was originally introduced by Claude E. Shannon as a tool to systematically analyze data flow in general communication systems [12]. The theory has been extended and subsequently applied to many fields including DNA sequence analysis [13-15]. Methods of Information theory focusing on DNA sequence compression have found differences between coding and non-coding sequences as well as between prokaryotic and eukaryotic organisms [16].

These results led us to apply information theoretical methods to examine the extent to which information content differed between the genomes of bacterial chromosomes, plasmids, phages and GIs, and whether such differences could be related to distinct genomic properties of bacterial chromosomes and mobile genomic elements. We used the Kullback-Leibler divergence measure (D_{KL}) of tetranucleotide frequencies within genomic DNA sequences, similar to that descried by Sadovsky [15], but using tetranucleotide frequencies and a zero order Markov model instead of a second order Markov model. These alterations increase the sensitivity of detection [17]. The zero order Markov model assumes the simplest possible dependence structure between neighboring nucleotides. This means that D_{KL} will be higher than in models that do account for dependence between adjacent nucleotides, like the first or second order Markov models [17]. The expected tetranucleotide frequencies, statistically speaking, are thus calculated from mononucleotide frequencies implying that the bases are independent of each other. Thus, D_{KL} reflects relative entropy in the sense that the genomic sequences are compared to a random sequence sharing only the same AT content. Low D_{KL} means low relative entropy and high D_{KL} means high relative entropy [18]. Since the DNA sequence from the biological entity is compared to a random, 0^{th} order Markov based sequence (sharing only total AT content), a lower D_{KL} reflects a greater independence between nucleotides in the corresponding tetranucleotides, and hence that less information is carried by the DNA sequence. Conversely, higher D_{KL} is taken to mean that more information is carried by the DNA sequence since the adjacent nucleotides in the corresponding tetranucleotides are more dependent on each other.

We sought to use methods from information theory to examine information capacity (relative entropy) in chromosomes, plasmids, phages and GIs. We investigated possible influences affecting relative entropy in the different types of DNA sequences and how relative entropy varies along bacterial chromosomes, focusing particularly on the AT rich *Bacillus cereus*, the medium AT:GC *Escherichia coli* and the GC rich *Mycobacterium tuberculosis*. We also examined the relative entropy of highly conserved genes in two closely related species (*M.tuberculosis* and *M.leprae*) of which one has presumably undergone considerable genome reduction [19,20].

Results

A note on the calculation of D_{KL}

The relative entropy of a DNA sequence, which we refer to as D_{KL}, is measured as the divergence between observed tetranucleotide frequencies from approximated tetranucleotide frequencies using a zero order Markov model. The zero order Markov model assumes that every base in the sequence is occurring with a probability independent of all other neighboring bases. It is reasonable to assume that in regions of high mutation activity this is a good description [11]. We compare the computed tetranucleotide frequencies from the zero order Markov model to counted tetranucleotide frequencies from each DNA sequence. So the information capacity in a DNA sequence is positively associated with the magnitude of the divergence from the approximated sequence. Hence, the higher the divergence between observed and expected (approximated) tetranucleotide frequencies the more information potential in the DNA sequence, and vice versa.

D_{KL} differences between chromosomes, GIs, phages and plasmids

We examined whether information capacity varied between chromosomes and two potential 'vectors': *i.e.* phages and plasmids, as well as GIs. Figure 1 shows that the D_{KL} was slightly lower amongst GIs than chromosomes ($p \sim 0.004$, see the Methods section for more details on the statistical methods). Phages were in turn found to have a lower D_{KL} than GIs ($p < 0.001$), and plasmids had slightly lower D_{KL} than phages ($p \sim 0.004$). Hence, the largest difference in D_{KL} (the most divergent tetranucleotide frequencies compared to a random sequence) was between chromosomes and plasmids ($p < 0.001$). In other words, chromosomes were, on average, the most biased DNA sequences while the plasmids had the most random (least biased) DNA composition.

Relative entropy vs AT content

An association between information capacity and AT content has been found for chromosomes in previous studies using slightly different methods than those described here (see Methods section) [10,11]. Since

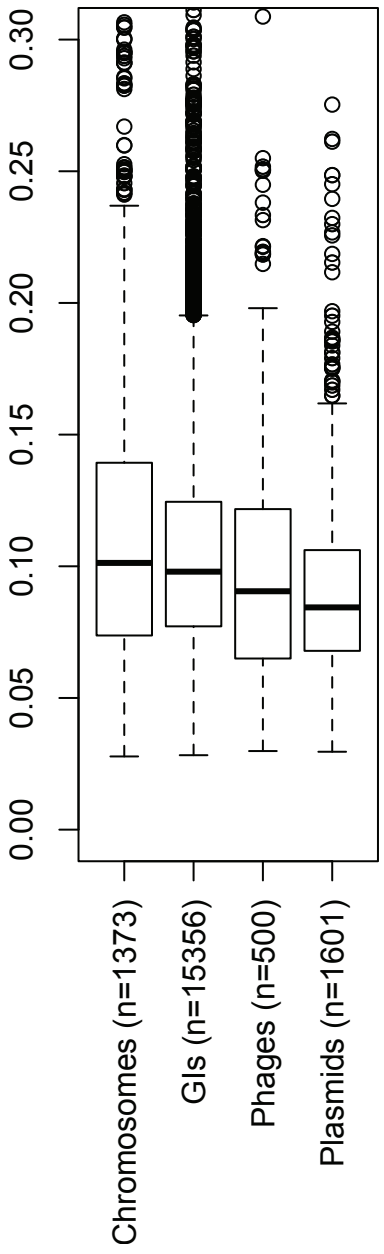

Figure 1 Boxplot of D_{KL}. The box and whisker plot shows D_{KL} differences between chromosomes, GIs, phages and plasmids. The boxes contain 50% of the values surrounding the median (thick horizontal line), while the lower and upper whiskers represent approximately the 25% lowest and highest D_{KL} values, respectively. The open circles above the upper whiskers are considered as outlying values. Although differences were small between the types of DNA sequence families, chromosomes were found to be, on average, the most biased (highest information potential) while plasmids the least.

there was a statistical significant difference in relative entropy between vectors (plasmids and phages) and chromosomes we explored whether similar associations could be found between the vectors and AT content. Figure 2 shows that relative entropy, D_{KL}, in chromosomes, plasmids, phages and GIs is negatively correlated with AT content: D_{KL} tends to decrease with increasing AT content. Regression analyses with D_{KL} as the response and AT content as the predictor gave $R^2 = 0.33$ for chromosomes, $R^2 = 0.21$ for plasmids, $R^2 = 0.56$ for phages, and $R^2 = 0.22$ for GIs. A likelihood ratio test between ANOVA models with size plus AT content versus AT content alone did not improve the correlation. All statistical results mentioned were significant, $p < 0.001$.

Relative entropy comparisons of shared genes between *M. tuberculosis* and *M. leprae*

It has been shown that the genomes of intracellular microbes have a tendency to reduce in size due in part to more mutations and eventual loss of DNA repair genes [21,22]. We examined whether these changes are reflected in relative entropy of the genomes of *M. tuberculosis*, a facultative intracellular pathogen, and *M. leprae*, an obligate intracellular pathogen considered to be in a transitional state between free living and intracellular lifestyles [19,20]. *M. leprae* has a smaller genome than *M. tuberculosis* (3.3 mb vs. 4.4 mb) and it is more AT rich (42.3% vs 24.4%). Figure 3 shows that D_{KL} taken from highly conserved coding regions was also lower in *M. leprae* than for *M. tuberculosis*, implying that *M. leprae* has a more random base composition, possibly due to an increased number of accumulated mutations. The fact that relative entropy was taken from shared functional genes between the two organisms supports the existing model of genome decay in intracellular microbes [21] resulting in increased randomness amongst the protein coding regions.

Phylogenetic influence on relative entropy

Using comparable methods to D_{KL}, Reva and Tümmler argued that DNA sequence bias appears to be a taxon-specific phenomenon within bacteria [10]. To assess whether D_{KL} was influenced by taxonomy (Figure 4) we picked out one strain from each species to decrease bias from multiple strains, reducing the dataset to 709 chromosomes. We found that phylogenetic relationship did significantly influence D_{KL}, but only slightly ($R^2 = 0.21$) and comparable to that of GC content ($R^2 = 0.22$). The phyla and %GC factors did, however, not interact and a model including both GC content and phyla as predictors explained approximately 40% ($R^2 = 0.4$) of the

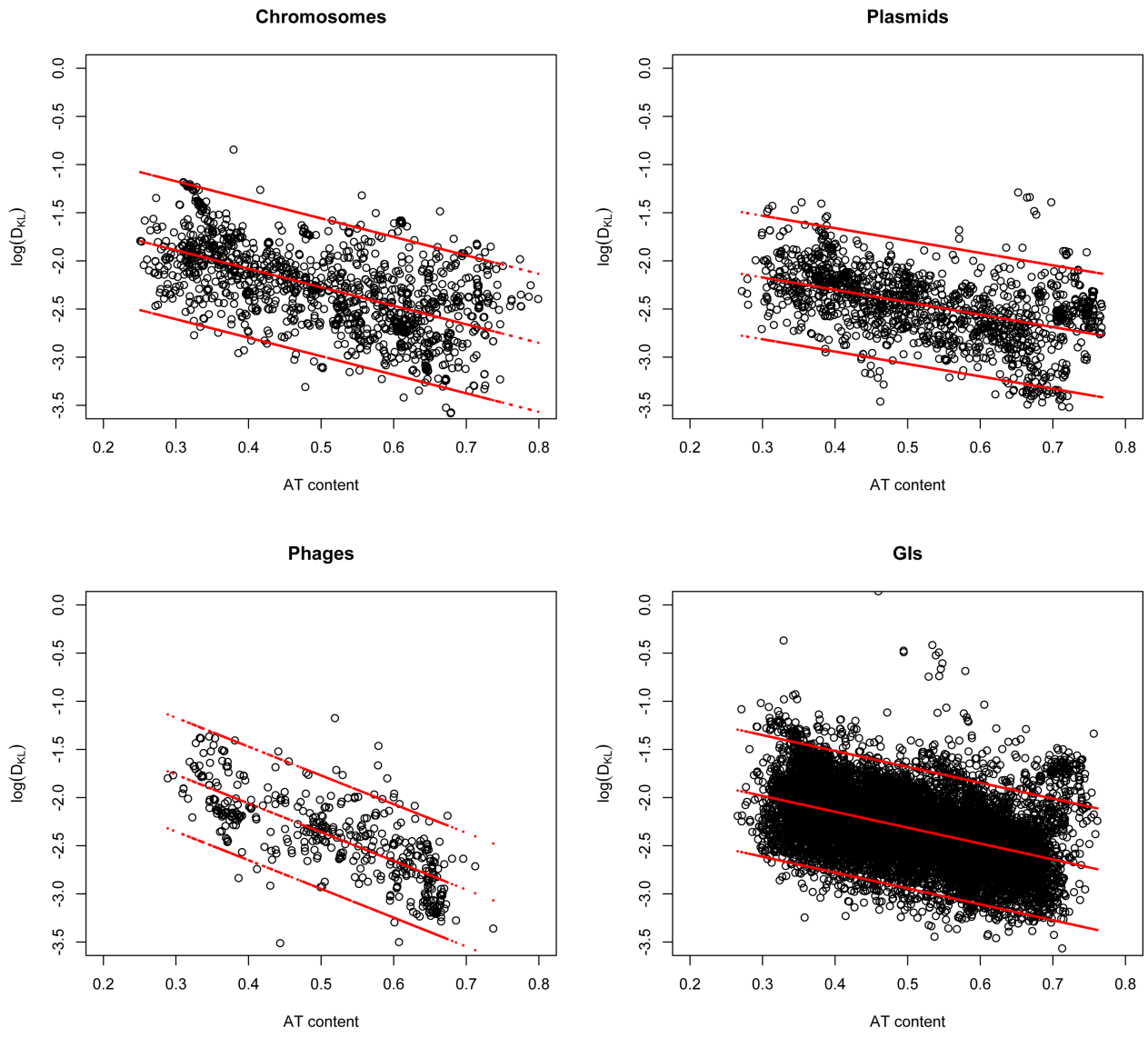

Figure 2 D_{KL} vs AT content. Log-transformed D_{KL} (vertical axis) is plotted against AT content (horizontal axis) with accompanying regression lines and 99% prediction intervals for chromosomes, plasmids, phages, and GIs. A clear correlation can be observed for all DNA sequences between (log-transformed) D_{KL} and AT content meaning that randomness in DNA sequences increases with AT content. The highest correlation was observed between relative entropy in phages and AT content.

variance observed. All results were statistically significant with $p < 0.001$. No significant difference ($p \sim 0.87$, Welch two-sample T test) in relative entropy was found between archaea and bacteria.

D_{KL} changes within genomes

To assess how relative entropy varied within bacterial chromosomes we examined the chromosomes of GC-rich *Mycobacterium tuberculosis* (65% GC), *Escherichia coli* K-12 with approximately 50% AT/GC, and AT rich *Bacillus cereus* (65% AT) using a sliding window of 5 kbp with D_{KL} from each window compared to D_{KL} for the whole chromosome. The aim was to examine whether D_{KL} could be regarded as a stable measure within bacterial chromosomes, similar to the genome signature [23]. Figure 5 shows how D_{KL} changed within the three species compared to a randomly constructed 50% GC chromosome of equivalent size to *E.coli* (5 Mbp). Notice that although D_{KL} varied within the chromosomes the level of variance was stable, indicating that average D_{KL} is a robust property for the whole DNA sequence.

In addition, Figure 5 shows that although *M. tuberculosis* and *E. coli* had similar D_{KL} measures throughout

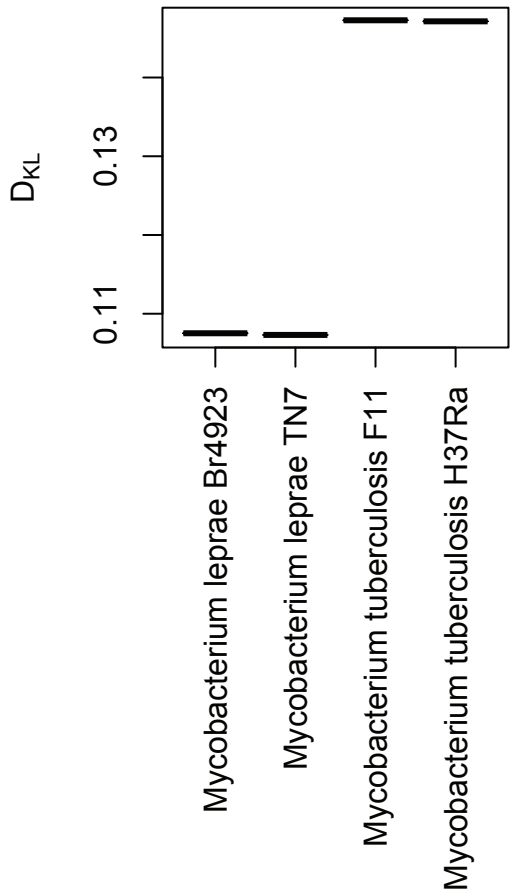

Figure 3 D_{KL}differences in *M. leprae* and *M. tuberculosis*. The figure show D_{KL} (vertical axis) for shared highly conserved genes of *M. leprae* and *M. tuberculosis* (horizontal axis).

the chromosome, the *B. cereus* chromosome exhibited considerably lower D_{KL}. This was especially pronounced in the middle of the chromosome. The accompanying BLAST atlas (Figure 6) [24] shows that the DNA molecule in this area was more AT rich, had more pronounced intrinsic curvature, increased stacking energy (making the double stranded DNA string easier to melt), higher position preference, and a higher occurrence of quasi- and perfect palindromes.

Size vs AT content

Although it has been demonstrated that AT content and chromosome sizes are inversely correlated in prokaryotes, we carried out additional tests for plasmids, phages, GIs as well as chromosomes. From Figure 7 it can be seen, as expected, that we found an association between chromosome size and AT content R^2~0.22, $p < 0.001$. In addition, we found a significant association between plasmid size and AT content, albeit low (R^2~0.16), which could be due to the increased variance. With an R^2~0.01 or less, the size of both phages and GIs were not associated with AT content. All results were statistically significant ($p < 0.001$).

Size vs. relative entropy

Since the correlation between DNA sequence size and GC content is well established [25,26] we examined whether D_{KL} was affected by DNA sequence size. We performed regression analyses with D_{KL} of chromosomes, GIs, phages and plasmids as the response and the corresponding sequence size as the predictor variable, measuring, in effect, the correlation between D_{KL} and sequence size. In all instances R^2 (the coefficient of determination) was found to be lower than 0.05, meaning that less than 5% ($p < 0.001$) of the variance observed in the data was explained by the regression models. A regression analysis with GC content as outcome indicated that variance explained increased additively as DNA sequence size (21% and 15% ($p < 0.001$) for bacterial chromosomes and plasmids, respectively) and D_{KL} (48% and 29% ($p < 0.001$) chromosomes and plasmids, respectively) was added to the model. Hence, AT content has an independent effect on DNA sequence size and relative entropy in bacterial chromosomes and plasmids, while D_{KL} was not affected by DNA sequence size regardless of DNA sequence type examined. It should be noted that for the combined regression model including both D_{KL} and DNA sequence size the %-variance explained metrics (i.e. R^2) were slightly different from the individual models discussed in the above sections due to the different types of transformations used (see Materials section for further details).

Discussion

Relative entropy in chromosomes, plasmids, phages and GIs

Chromosomes were, on average, the most biased sequences (*i.e.* least similar to a random sequence) and therefore presumably the most subjected to selective pressures of the sequences examined here. In terms of D_{KL} there was a small, but significant difference between GIs and chromosomes. This difference is expected since GIs are found within chromosomes and have ameliorated over time, which, in base compositional terms, tend towards that of the host chromosome [27]. Hence, a number of studies indicate that GIs consist of horizontally acquired mobile genetic fragments [22,28], but our data does not identify what type of vector has brought these GIs to their respective chromosomes.

The reduced D_{KL} of phages compared to plasmids was small but statistically significant. In contrast to phages, plasmids exist independently of the host chromosome and are generally non-lethal [29]. When the phenotypic features of the plasmid are not required for bacterial

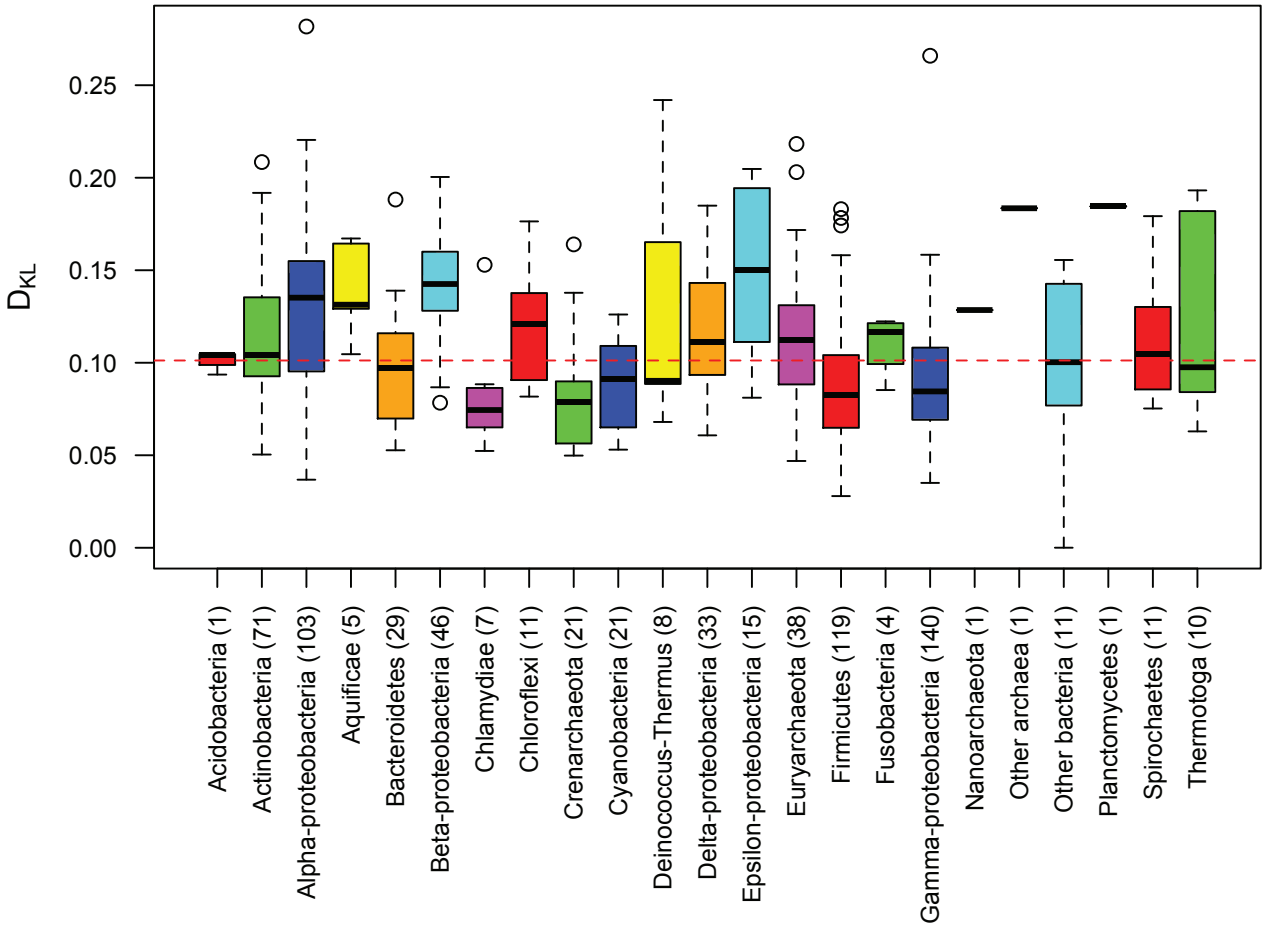

Figure 4 D_{KL} vs Phyla. The figure shows a boxplot of D_{KL} (vertical axis) plotted against phylogenetic groups (horizontal axis, number of genomes from each phylogenetic group in parenthesis). The dashed horizontal red line is the average D_{KL} for all groups.

survival, the plasmid will exist only in a small minority of the total microbial population [30]. In this way the forces of selective pressure are reduced compared to the host chromosome. Phages also exist independently of bacterial chromosomes but rely on the bacterial machinery for replication [29,30]. However, those phages that are lytic will be under greater selective pressure than plasmids. What particular features of phages that result in the reduced information content remains to be clarified.

It should be noted that the comparisons were between all deposited DNA sequences, which means that the results reflect the distributions of chromosomes, GIs, phages and plasmids that initially have been originally selected and sequenced for a purpose. The effect of this bias is not clear.

Association between D_{KL} and AT content

Figure 2 shows that decreased relative entropy (D_{KL}) is associated with increasing AT content. An example of this was demonstrated in Figure 3, where the more AT rich *M. leprae* was found to have lower D_{KL} in genes that are also shared with the more GC rich *M. tuberculosis*.

Although the coefficient of determination, R^2, varied between GIs, phages, plasmids and chromosomes, Figure 2 shows that the trend remained for all DNA sequences examined. Phages obtained a surprisingly high coefficient of determination, $R^2 = 0.56$, implying that relative entropy was more linked to changes in AT content in these organisms.

D_{KL} variation within chromosomes

The D_{KL} profile of the *B. cereus* chromosome may imply that areas of low relative entropy (low D_{KL}) might be indicators of genetic regions especially prone to rearrangement. This propensity for re-arrangements may be due to the increased stacking energy, position preference and amount of quasi-palindromes observed in the region, all of which are determinants of genomic re-

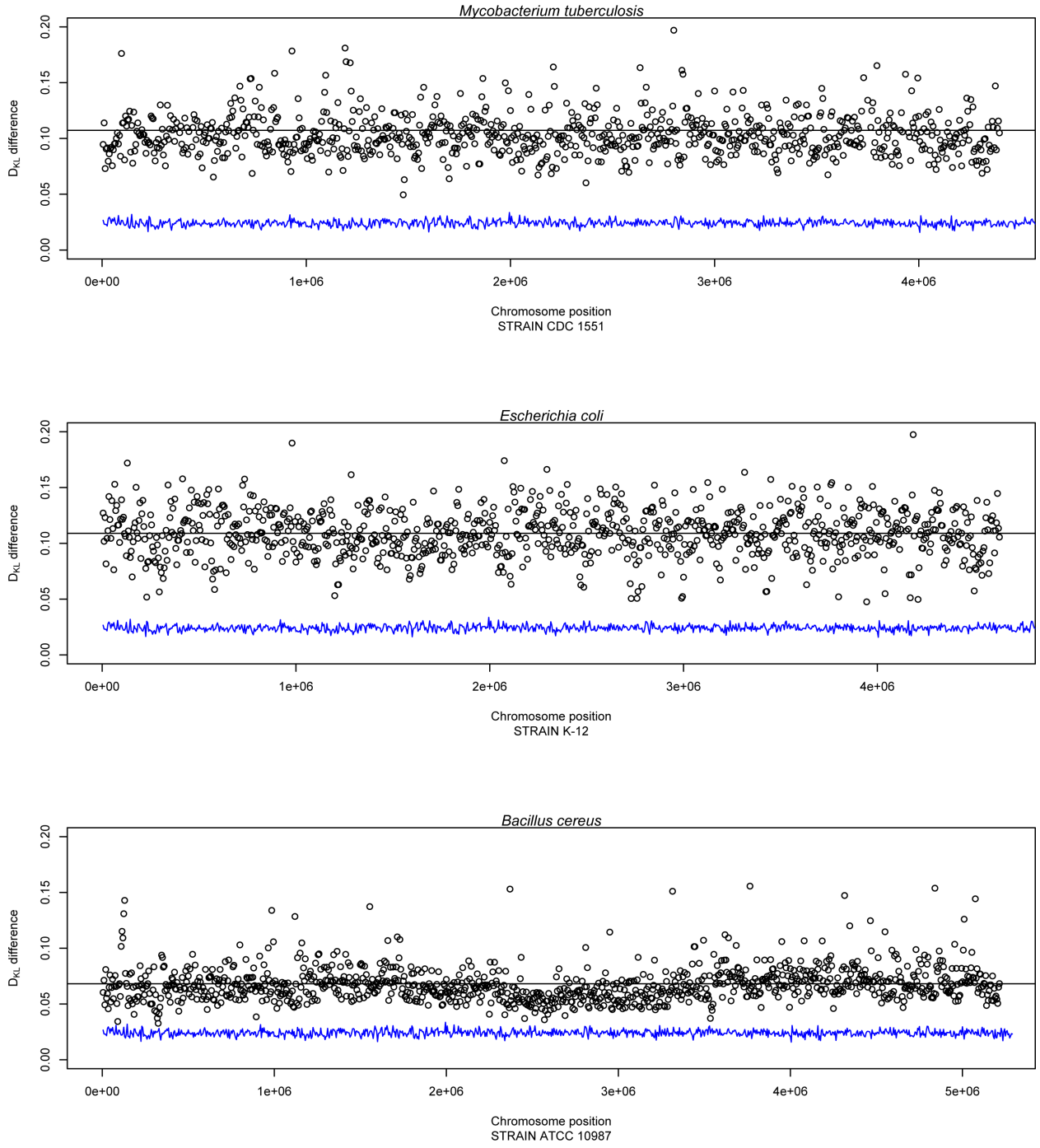

Figure 5 Profiles of D_{KL} differences within *M. tuberculosis, E. coli* and *B. cereus*. Profiles made from the D_{KL} values of non-overlapping sliding windows in *M. tuberculosis, E. coli* and *B. cereus*. It can be seen that D_{KL} values within the chromosomes are remarkably stable. *B. cereus* has noticeably lower D_{KL} values than the other genomes indicating that the chromosome has a comparably more random base composition. The D_{KL} values of a 50% GC content random genome are also included for comparison. For all chromosomes, the black horizontal line represents mean D_{KL}.

arrangement. The relatively high occurrence of both palindromes and quasi-palindromes in the region of *B. cereus* with low relative entropy may indicate that the mechanisms leading to quasi palindrome correction have not been operating properly in these regions as compared to the chromosome in general [31] possibly resulting also in a higher number of accumulated mutations [17]. A similar region has been found for all

sequenced members of the *B. cereus*-group, which implies that the genetic region has been selected and kept possibly due to some unknown advantage. As can be seen from Figure 6, the region is predominantly gene coding. Since the genomes of the *B. cereus* group are relatively large compared with the distantly related *B. subtilis* it can be speculated that the region is an acquired phage or plasmid.

Connections between DNA sequence and structure

Although relative entropy has some mathematical associations with thermodynamics the two concepts are, in general, independent of each other [18]. However, it is known that greater energy is required to melt GC rich sequences than AT rich sequences [32]. Considering our results found a negative correlation between D_{KL} and AT content it is possible that DNA structure energetics and DNA sequence relative entropy may be connected and provides a link between DNA structure and sequence. This is supported by the findings shown in Figure 6 where a genetic region of low relative entropy was found to have more intrinsic DNA structural curvature, increased stacking energies and higher position preference. Hence, our findings may point to possible DNA structural differences between bacterial chromosomes, plasmids and phages that could have implications for how these biological entities are integrated into their hosts.

Phylogenetic influences on relative entropy

Our measure of relative entropy revealed that approximately 21% of the variation in D_{KL} could be explained by a close phylogenetic relationship. This value compares well with the 22% in variation that is explained by GC content. Thus, D_{KL} appears to be as much influenced by phyla as GC content is, while almost 80% is accounted for by other factors. Using a method that is strongly associated with relative entropy (OUV, oligonucleotide usage variance), 55% of the variance could be explained by environment, phyla and AT content [17]. If non-coding regions were excluded 67% of the variance could be explained using environment, phylum and AT

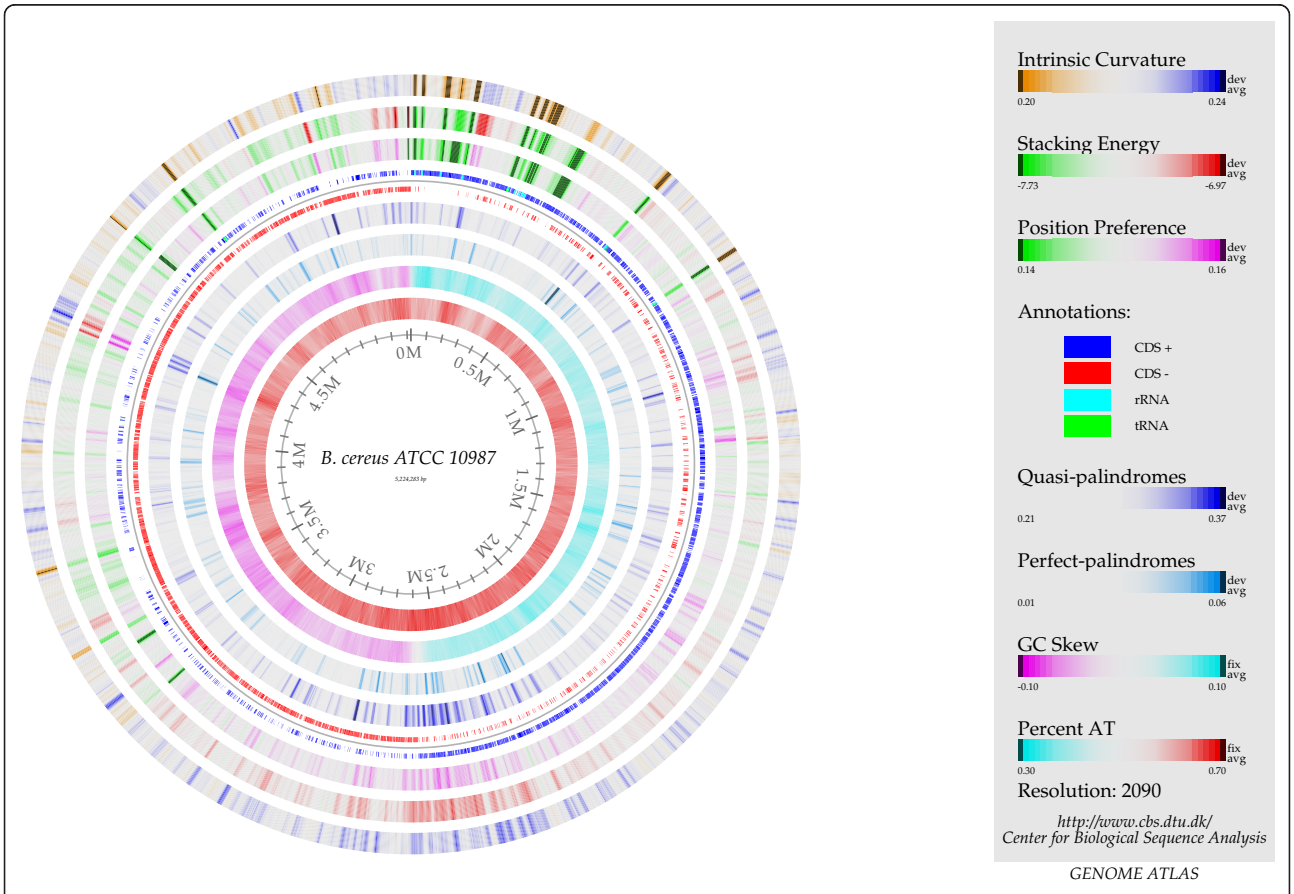

Figure 6 Blast atlas of *B. cereus* ATCC 10987. Blast atlas for *B. cereus* ATCC 10987 depicting several structural and sequential features. Of special interest is the region approximately between 2 mb and 3 mb, which has a substantial higher occurrence of both quasi- and perfect palindromes, increased stacking energy and intrinsic curvature, as well as higher position preference than the rest of the chromosome.

content. The above mentioned study also discusses possible influences between environmental factors and possible implications of high and low OUV for a number of microbes that is relevant to the present exposition. The difference between OUV and relative entropy is explained in the Methods section.

Relation between relative entropy and DNA sequence size

Although a possible link between plasmid size and ecology has been reported [29], and a correlation between microbial chromosome size and GC content has been established previously, to the best of our knowledge no such correlation has been reported between plasmid size and GC content. It can also be seen from Figure 7 that plasmid sizes vary considerably more with respect to AT content than chromosomes, which could indicate that the DNA sequences of plasmids are less stable and more prone to genetic exchange than the DNA sequences of chromosomes.

Lack of correlation between relative entropy and DNA sequence size

Although a correlation between DNA sequence size and D_{KL} in bacterial chromosomes and plasmids could be expected due to the correlation found between these factors and genomic AT content, no such correlation was found. This may imply that the relation between genomic AT content and DNA sequence size is independent of the relation between genomic AT content and relative entropy. In other words, genomic AT content may be differently related to DNA sequence size than to relative entropy in bacterial chromosomes and plasmids (no correlation was found between AT content and DNA sequence size in GIs and phages). This claim

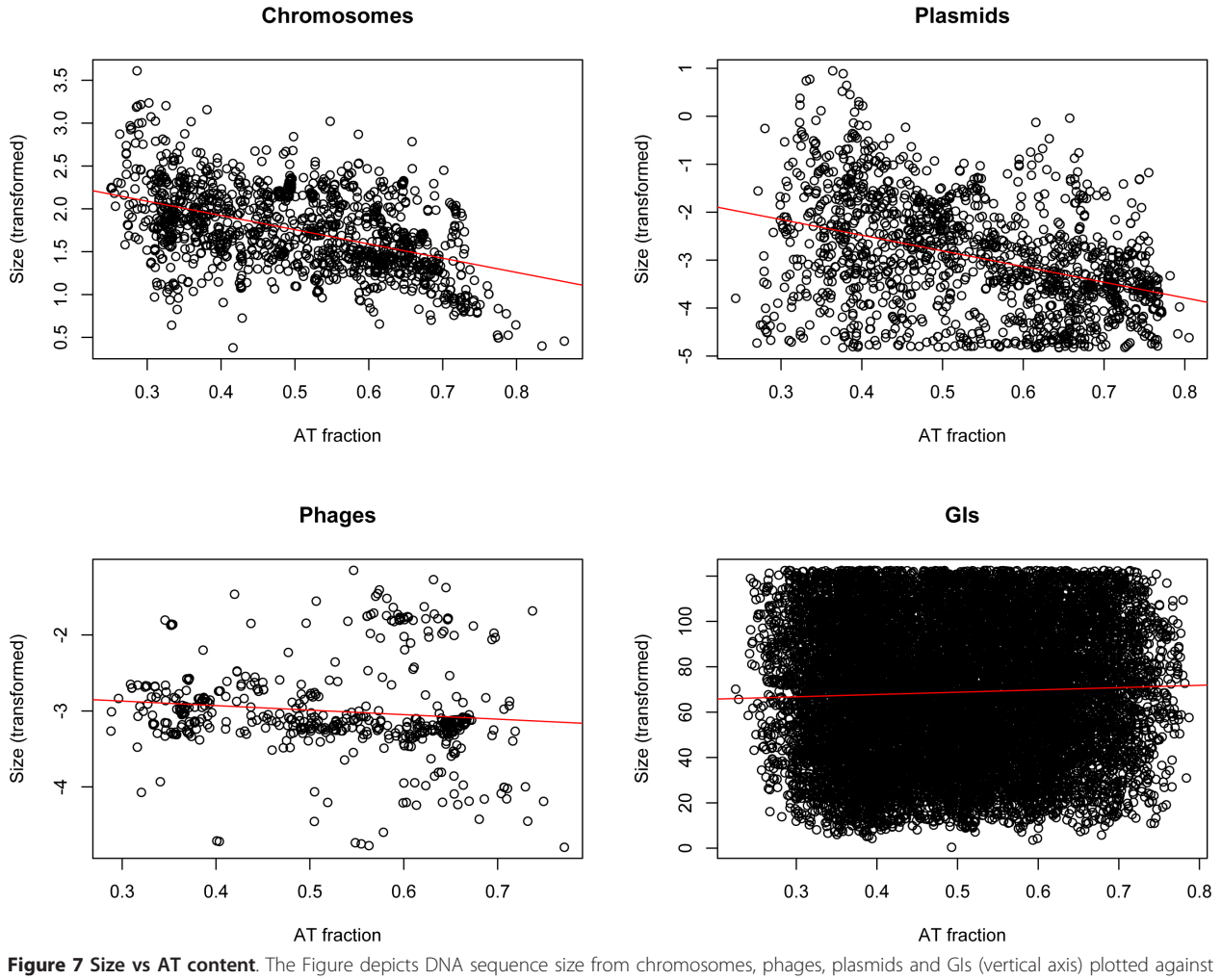

Figure 7 Size vs AT content. The Figure depicts DNA sequence size from chromosomes, phages, plasmids and GIs (vertical axis) plotted against their corresponding AT contents (horizontal axis). Associations between DNA sequence size and AT content can be observed for chromosomes and plasmids.

was further strengthen by a linear regression analysis, which indicated that the variance explained increased additively with DNA sequence size and relative entropy added as predictors. Hence, our models indicate that the mechanisms connecting AT with DNA sequence size are unrelated and different to the mechanisms linking AT content with relative entropy.

Connections to other studies

By using BLAST and graph/network analyses it has been found that the different groups, *i.e.* chromosomes, plasmids and phages, share, in the majority of cases, DNA amongst themselves. In other words, chromosomes share DNA with chromosomes, plasmids share DNA with plasmids and phages share DNA with phages [5]. Variation among bacterial chromosomes however is predominantly mediated by genetic exchange from plasmids and only transiently so by phages [5]. Our results indicated that plasmids, on average, had significantly lower D_{KL} than any of the other types of DNA sequences. This could mean that plasmids are more tolerant to genetic alterations something that may be crucial to maximize host range [33]. A previous study has reported a correlation between plasmid-host similarity and GC content, *i.e.* the more similar the plasmids-hosts were in terms of genomic signatures, the more GC rich they tended to be [9]. Phages have been found to have a narrow host range, in fact even more so than plasmids [5] in spite of their larger numbers (estimations go as high as 5-10 phages for each bacterium on earth [34-36]), which may indicate that they have been subjected to increased selective pressures resulting, in turn, in significantly higher D_{KL} than for plasmids. Due to the possible link between relative entropy and DNA sequence mutations it can be speculated whether phages are more vulnerable to genetic rearrangements than plasmids, resulting in higher D_{KL}, on average in phages.

Conclusions

In conclusion, we find that GIs and chromosomes have similar relative entropy (D_{KL}), which may be due to amelioration of the foreign DNA towards the base composition of the host chromosome. Both plasmids and phages had significantly lower relative entropy than GIs and chromosomes. Plasmids had the lowest D_{KL} of all types of DNA sequences examined, meaning that plasmids contained, on average, the most mutated DNA sequences. Relative entropy decreased in all types of DNA sequences in concordance with increasing AT content, possibly implying that the number of accumulated mutations appear to increase with AT content regardless of the (prokaryotic) biological entity. This was also demonstrated on a shared set of highly conserved genes from *M. tuberculosis* and *M. leprae*, of which the latter, known to have undergone considerable genome reduction, was found to have significantly lower relative entropy (*i.e.* more random DNA sequences possibly due to mutation) in the protein coding genes. AT content and D_{KL} association was especially pronounced for phages, which may reflect an evolutionary strategy that associates the number of accumulated mutations with AT content to a substantially larger extent in phages than bacteria.

Methods

Chromosomes, plasmids and phages were downloaded from the NCBI website http://www.ncbi.nlm.nih.gov/genome/, while the GIs were downloaded from the Islandviewer website http://www.pathogenomics.sfu.ca/islandviewer/query.php. Only DNA sequences larger than 10 kb were considered due to limitations of the method. Single copy orthologs were assigned by OrthoMCL [37] for the genomes of *Mycobacterium tuberculosis* F11 (CP000717.1), *M. tuberculosis* H37Ra (AL123456.2), *M. leprae* Br4923 (FM211192.1) and *M. leprae* TN7 (AL450380.1). Statistical analyses were carried out with R http://www.r-project.org/, which was also used to create all figures except the BLAST atlas (Figure 6). The BLAST atlas was made using CBS in-house software [24,38].

The Kullback-Leibler divergence (D_{KL}, also referred to as the relative entropy) is a measure of difference between two discrete probability mass functions [18]. Let $\mathbf{s}$ be a DNA sequence, and $\mathbf{z}_1,...,\mathbf{z}_{256}$ be all possible tetramers of the DNA alphabet ($4^4 = 256$). The observed frequencies of tetranucleotides from DNA sequence $\mathbf{s}$ is written as $O(\mathbf{z}_i|\mathbf{s})$. The expected frequencies of tetranucleotides from DNA sequence $\mathbf{s}$ found using a zero order Markov model is written as $E(\mathbf{z}_i|\mathbf{s})$. The KL divergence for the sequence $\mathbf{s}$ is given as:

$$D_{KL}(\mathbf{s}) = \sum_{i=1}^{256} O(\mathbf{z}_i|\mathbf{s}) \log\left(\frac{O(\mathbf{z}_i|\mathbf{s})}{E(\mathbf{z}_i|\mathbf{s})}\right)$$

A lower D_{KL} is interpreted as lesser information potential is carried by the DNA sequence $\mathbf{s}$ due to lesser dependence between the nucleotides in the corresponding tetranucleotides. Conversely, a higher D_{KL} is taken to mean that higher information potential is carried by the DNA sequence (higher relative entropy), since the nucleotides in the corresponding tetranucleotides are more dependent on each other. The OUV measure [17] described in the Discussion section and compared to relative entropy is calculated as follows (O, E, $\mathbf{z}_i$ and $\mathbf{s}$ are the same as above):

$$D_{OUV}(\mathbf{s}) = \frac{1}{256}\sum_{i=1}^{256}\frac{O(\mathbf{z}_i|\mathbf{s})}{E(\mathbf{z}_i|\mathbf{s})}$$

Although the OUV measure is similar to relative entropy, we use the latter here due to the larger theoretical framework and tools available from information theory [12,18].

Comparisons between D_{KL} and factors such as phyla, AT content, DNA sequence size, etc. were carried out using linear regression with transformations applied to correct for non-normality where needed.

D_{KL} was computed for each DNA sequence (chromosome, plasmid, phage and GI) and compared to AT content, size and phyla using linear regression:

$$Y = a + bX + \varepsilon$$

For comparisons between chromosome, plasmid, GI and phage size ($Y = Y_{size}$) versus D_{KL} (X_{KL}) no transformation was used.

To examine the relationship between D_{KL}, DNA sequence size and AT content for bacterial chromosomes and plasmids, a linear regression model was used without transformations on the response:

$$Y_{AT} = a + bX_{KL} + cX_{Size} + d(X_{Size})^2 + \varepsilon$$

Linear regression between D_{KL} as outcome ($Y = Y_{KL}$) and AT content as response ($X = X_{AT}$) was log-transformed:

$$LogY_{KL} = a + bX_{AT} + \varepsilon$$

Several transformations were used to assess associations between chromosome, plasmid, phage and GI size (Y_{Size}) vs AT content (X_{AT}) using the following regression equation:

$$Y_{Size} = a + bX_{AT} + \varepsilon$$

A square root transform was used when the response was sequence sizes for chromosomes; log transformations for both phage and plasmid sizes; and ($1/ Y_{Size}$) transform for GI sizes as outcome.

Comparison of D_{KL} between chromosomes, plasmids, phages and GI, as seen in Figure 1, were carried out using the non-parametric Wilcoxon (Mann-Whitney) test due to skewed (but similar) distributions.

All statistical results presented as results were found to be statistically significant with $p < 0.001$, if not otherwise stated in the text.

All D_{KL} measurements of DNA sequences were carried out using in-house software. The profiles measuring D_{KL} changes within bacterial chromosomes as seen in Figure 5 were performed using non-overlapping sliding windows of 5 kbp compared to average chromosomal D_{KL}.

Acknowledgements

The authors wish to thank the referees as well as Hilde Mellegård and Torunn Dønsvik for their helpful comments. MWJvP is funded by the Netherlands Organization for Scientific Research (NWO) via a VENI grant.

Author details

[1]Norwegian School of Veterinary Science, EpiCentre, Department of Food Safety and Infection biology, Ullevålsveien 72, Oslo, Norway. [2]Systems and Synthetic Biology, Wageningen University, Wageningen, the Netherlands. [3]Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Sciences, Ås, Norway. [4]National Veterinary Institute, Ullevålsveien 68, Pb 750 Sentrum, N-0106 Oslo, Norway. [5]Center for Biological Sequence Analysis, Department of Systems Biology, Comparative genomics unit, Technical University of Denmark, DK-2800 Lyngby, Denmark.

Authors' contributions

JB, LS and ABK carried out statistical analyses. JB, MWJvP, SPH, and DU contributed to data analyses and discussion. All authors participated in the writing of the manuscript. The study was initiated by JB. All authors have read and approved the final manuscript.

Received: 15 November 2011 Accepted: 10 February 2012
Published: 10 February 2012

References

1. van Passel MW, Marri PR, Ochman H: **The emergence and fate of horizontally acquired genes in Escherichia coli.** *PLoS Comput Biol* 2008, **4(4)**:e1000059.
2. Roos TE, van Passel MW: **A quantitative account of genomic island acquisitions in prokaryotes.** *BMC Genomics* 2011, **12**:427.
3. Fournier PE, Drancourt M, Raoult D: **Bacterial genome sequencing and its use in infectious diseases.** 2007, **7(11)**:711-723.
4. Langille MG, Hsiao WW, Brinkman FS: **Evaluation of genomic island predictors using a comparative genomics approach.** *BMC Bioinformatics* 2008, **9**:329.
5. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E: **Network analyses structure genetic diversity in independent genetic worlds.** *Proc Natl Acad Sci USA* 2010, **107(1)**:127-132.
6. Haerter JO, Trusina A, Sneppen K: **Targeted bacterial immunity buffers phage diversity.** *J Virol* 2011, **85(20)**:10554-10560.
7. Pride DT, Wassenaar TM, Ghose C, Blaser MJ: **Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses.** *BMC Genomics* 2006, **7**:8.
8. van Passel MW, Bart A, Luyf AC, van Kampen AH, van der EA: **Compositional discordance between prokaryotic plasmids and host chromosomes.** *BMC Genomics* 2006, **7(1)**:26.
9. Bohlin J, Skjerve E, Ussery DW: **Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes.** *BMC Genomics* 2008, **9**:104.
10. Reva ON, Tummler B: **Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns.** *BMC Bioinformatics* 2004, **5**:90.
11. Bohlin J, Skjerve E, Ussery DW: **Investigations of oligonucleotide usage variance within and between prokaryotes.** *PLoS Comput Biol* 2008, **4(4)**: e1000057.
12. Shannon CE: **The mathematical theory of communication. 1963.** *MD Comput* 1997, **14(4)**:306-317.
13. Yockey HP: **Origin of life on earth and Shannon's theory of communication.** *Comput Chem* 2000, **24(1)**:105-123.
14. Schneider TD: **Information content of individual genetic sequences.** *J Theor Biol* 1997, **189(4)**:427-441.
15. Sadovsky MG: **Information capacity of nucleotide sequences and its applications.** *Bull Math Biol* 2006, **68(4)**:785-806.
16. Menconi G, Marangoni R: **A compression-based approach for coding sequences identification. I. Application to prokaryotic genomes.** *J Comput Biol* 2006, **13(8)**:1477-1488.
17. Bohlin J, Skjerve E: **Examination of genome homogeneity in prokaryotes using genomic signatures.** *PLoS One* 2009, **4(12)**:e8113.
18. Cover TM, Thomas JA: Elements of Information Theory: Wiley; 1991.
19. Vissa VD, Brennan PJ: **The genome of *Mycobacterium leprae*: a minimal mycobacterial gene set.** *Genome Biol* 2001, **2(8)**:REVIEWS1023.
20. Gomez-Valero L, Rocha EP, Latorre A, Silva FJ: **Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction.** *Genome Res* 2007, **17(8)**:1178-1185.
21. Moran NA: **Microbial minimalism: genome reduction in bacterial pathogens.** *Cell* 2002, **108(5)**:583-586.

22. Rocha EP, Danchin A: **Base composition bias might result from competition for metabolic resources.** *Trends Genet* 2002, **18(6)**:291-294.
23. Karlin S, Campbell AM: **Which bacterium is the ancestor of the animal mitochondrial genome?** *Proc Natl Acad Sci USA* 1994, **91(26)**:12842-12846.
24. Hallin PF, Binnewies TT, Ussery DW: **The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology.** *Mol Biosyst* 2008, **4(5)**:363-371.
25. Mitchell D: **GC content and genome length in Chargaff compliant genomes.** *Biochem Biophys Res Commun* 2007, **353(0006-291; 1)**:207-210.
26. Musto H, Naya H, Zavala A, Romero H, varez-Valin F, Bernardi G: **Genomic GC level, optimal growth temperature, and genome size in prokaryotes.** *Biochem Biophys Res Commun* 2006, **347(0006-291; 1)**:1-3.
27. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44(4)**:383-397.
28. Langille MG, Hsiao WW, Brinkman FS: **Detecting genomic islands using bioinformatics approaches.** *Nat Rev Microbiol* 2010, **8(5)**:373-382.
29. Slater FR, Bailey MJ, Tett AJ, Turner SL: **Progress towards understanding the fate of plasmids in bacterial communities.** *FEMS Microbiol Ecol* 2008, **66(1)**:3-13.
30. Bahl MI, Hansen LH, Sorensen SJ: **Persistence mechanisms of conjugative plasmids.** *Methods Mol Biol* 2009, **532**:73-102.
31. van Noort V, Worning P, Ussery DW, Rosche WA, Sinden RR: **Strand misalignments lead to quasipalindrome correction.** *Trends Genet* 2003, **19(7)**:365-369.
32. Sinden RR: DNA Structure and Function: Academic Press; 1994.
33. Kirzinger MW, Stavrinides J: **Host specificity determinants as a genetic continuum.** *Trends Microbiol* 2012, **20(2)**:88-93.
34. Brussow H, Hendrix RW: **Phage genomics: small is beautiful.** *Cell* 2002, **108(1)**:13-16.
35. Paul JH, Sullivan MB, Segall AM, Rohwer F: **Marine phage genomics.** *Comp Biochem Physiol B Biochem Mol Biol* 2002, **133(4)**:463-476.
36. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R: **Reticulate representation of evolutionary and functional relationships between phage genomes.** *Mol Biol Evol* 2008, **25(4)**:762-777.
37. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ Jr: **Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups.** *Curr Protoc Bioinformatics* 2011, , **Chapter 6:** Unit 6.12.1-19.
38. Hallin PF, Ussery DW: **CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data.** *Bioinformatics* 2004, **20(18)**:3682-3686.

doi:10.1186/1471-2164-13-66
Cite this article as: Bohlin *et al.*: Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. *BMC Genomics* 2012 **13**:66.