Preprint of: *Gene* 410: 18-25, doi:10.1016/j.gene.2007.11.006

Distribution and evolution of short tandem repeats in closely related bacterial genomes

Keywords: Genome evolution; Microsatellite; SSR; *Chlamydia*; *Escherichia*; Prokaryotes

Edit Kassai-Jáger[a], Csaba Ortutay[b], Gábor Tóth[c], Tibor Vellai[a], Zoltán Gáspári[d,*]

[a]*Department of Genetics, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary,*

[b]*Institute of Medical Technology, FI-33014 University of Tampere, Tampere, Finland,*

[c]*Bioinformatics Group, Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, H-2100 Gödöllő, Hungary,*

[d]*Institute of Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, Hungary.*

Abbreviations: SSR, simple sequence repeat; bp, base pair(s); Mbp, megabase(s) or 1 million bp.

[*] Corresponding author. Institute of Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, Hungary, Tel.: +36-1-2090555 ext. 1408, Fax.: +36-1-3722620.

E-mail address: szpari@chem.elte.hu (Z. Gáspári).

Abstract

Simultaneous identification and comparison of perfect and imperfect microsatellites within a genome is a valuable tool both to overcome the lack of a consensus definition of SSRs and to assess repeat history. Detailed analysis of the overall distribution of perfect and imperfect microsatellites in closely related bacterial taxa is expected to give new insight into the evolution of prokaryotic genomes. We have performed a genome-wide analysis of microsatellite distribution in four *Escherichia coli* and seven *Chlamydial* strains. *Chlamydial* strains generally have a higher density of SSRs and show greater intra-group differences of SSR distribution patterns than *E. coli* genomes. In most investigated genomes the distribution of the total lengths of matching perfect and imperfect trinucleotide repeats are highly similar, with the notable exception of *C. muridarum*. Closely related strains show more similar repeat distribution patterns than strains separated by a longer divergence time. The discrepancy between the preferred classes of perfect and imperfect repeats in *C. muridarum* implies accelerated evolution of SSRs in this particular strain. Our results suggest that microsatellites may play an important role in the evolution of prokaryotic genomes and several gene families.

1. Introduction

Microsatellites – simple sequence repeats (SSRs) – are of great practical and theoretical importance in eukaryotes (Ellegren 2004; Kashi and King, 2006). In prokaryotes, their abundance is relatively low (van Belkum et al., 1998; Eckert and Yan, 2000; Metzgar et al.

2001; Schlotterer et al., 2006; Mrazek et al., 2007,), they nevertheless contribute to genome polymorphism in bacteria (Lindstedt, 2005). Escherichia coli O157:H7 VNTR repeats have been recently monitored (Noller et al., 2006). Since there is no consensus definition of microsatellites (Ellegren, 2004), it is not straightforward to compare SSRs identified in different studies. We have recently introduced a new approach, the separate identification and subsequent comparison of perfect and imperfect SSRs (Gáspári et al., 2007) to overcome this difficulty. Our approach is also expected to yield information about the history of the repeats if we assume that the majority of imperfect repeats containing a perfect core is a remnant of a longer perfect stretch. In this paper we apply our approach to related bacterial taxa to assess the intra- and inter-group similarities of genomic repeat distributions. Parallel investigation of related genomes using multiple SSR detection methods, combined with standardized SSR classification (Jurka and Pethiyagoda, 1995; Tóth et al., 2000) is expected to yield a biologically relevant picture of the significance of SSRs in the bacterial strains under study.

The two bacterial groups selected for the present survey comprise *E. coli* and *Chlamydial* strains. *Chlamydiales* comprise a monophyletic group that is phylogenetically well separated from other bacterial taxa (Stephens et al., 1998; Kalman et al., 1999; Read et al., 2000, 2003; Shirai et al., 2000; Chen et al., 2007). Their genome evolution has recently been investigated by bioinformatic methods (Ortutay et al., 2003; McNally et al., 2007). These features make these genomes ideal for a comparative analysis.

We chose *Escherichia coli* genoems as our other target group because these bacteria are among the most widely studied prokaryotes, with well-described genetics. The increasing number of *E. coli* strains with known genomes offers a unique opportunity to analyze SSR evolution in these very closely related bacteria (Blattner et al., 1997; Hayashi et al., 2001; Perna et al., 2001; Welch et al., 2002). In this case we analyzed 4 *Escherichia* genomes used also in other studies (Azad et al., 2007; McNally et al., 2007).

2. Material and Methods

2.1. Genomes used for this study

Complete genome sequences for various strains of *E. coli*, *Chlamydia muridarum*, *C. trachomatis*, *Chlamydophyla pneumoniae* and *C. caviae* were downloaded from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The genome sequences used for this study are summarized in **Table 1**.

2.2. SSR extraction and classification

SSR extraction and classification was performed as described previously (Gáspári et al., 2007), using in-house programs and `Tandem Repeats Finder` (`TRF`; Benson, 1999). Repeats with 1-6 bp units and with a minimum length of 12 bp were considered. Repeat unit classes were standardized as described earlier (Jurka and Pethiyagoda, 1995; Tóth et al., 2000; Gáspári et al., 2007), *e.g.* the class ‘**`acg`**’ represents all of its permutant and/or reverse complement sequences (**`acg=cga=gac=cgt=gtc=tcg`**). To identify imperfect repeats corresponding to perfect ones (i.e. to select perfect and imperfect repeats at identical loci within a selected genome), all imperfect repeats found around the location of each perfect repeat were selected. If there were multiple imperfect repeats matching the perfect one, priority was given to repeats with a repeat class identical to that of the perfect repeat. If no such imperfect repeat was found, the lengths of the repeated units were considered in a way that one of them must be a multiple of the other (e.g., a perfect repeat with unit length 6 can match an imperfect one with unit length 3). It is important to stress that in this context, ‘matching’ perfect and imperfect repeats are located in the same genome, and no systematic attempt was made to find homologous repeats in related genomes.

2.3. Data processing and evaluation

All data were stored in `MySQL` tables for subsequent analysis. Matching perfect and imperfect repeats were identified as being located at the same chromosomal position within a genome. Therefore, these repeats were identified by two independent methods. SSRs were assigned to coding or non-coding regions according to the ‘CDS’ records in the NCBI annotation. Orthologous genes in related strains were identified using the KEGG Database (*http://www.genome.jp/kegg*, Kanehisa et al., 2006). Gene sequences were aligned using `ClustalW` (Thompson et al., 1994). All other computation was performed using in-house `PERL` programs. Amino acid repeats encoded by trinucleotide SSRs were identified by mapping the repeat position onto the coding nucleotide and the corresponding translated protein sequences in the annotation of the GenBank files. Identical amino acids coded by at least 50% of the bases in the repeat sequence were included in the statistics. This was important to characterize imperfect repeats and also to account for the fact that repeat units may not coincide with codons (even a perfect repeat may code for more than one amino acid types).

2.4. Comparison of repeat distributions

To assess the differences between perfect and imperfect repeat distribution patterns and to compare the differences of such repeat distributions among bacterial strains, we applied the chi-square contingency analysis which tests the equality of two observed distributions as described in Gáspári et al. (2007). Here, the term ‘distribution pattern’ refers to the relative abundance of repeats in different classes in whole genomes (data as presented in **Tables 2 and 3** and in **supplementary Table S1**). We emphasize that the probability value obtained in the analysis is not evaluated here as in a conventional statistical test, but rather as an indicator that can measure the similarity of the distributions (Gáspári et al., 2007). Moreover, when used for the comparison of matching perfect and imperfect repeats within a genome, it is not expected to give statistically significant differences since the data compared come from different interpretations of the very same sequence. Therefore, any discrepancy detected suggests a difference of biological relevance. Our approach is highly similar to the `PRIDE` method applied for protein structure comparison (Carugo and Pongor, 2002) where contingency analysis is also applied as a measure for similarity and not as a conventional statistical test. Details of the contingency analysis calculations (excluded classes, χ^2 values etc.) can be found in **supplementary Table S2**.

3. Results

3.1. Overall microsatellite distribution in the selected strains

We found a considerable excess of tri- and hexanucleotide repeats with much less mono-, di-, tetra- and pentanucleotide SSRs. This is consistent with coding sequences comprising a large fraction of the genomes investigated, as only repeats with unit lengths of multiples of three do not cause frameshift upon expansion in coding regions. Indeed, the vast majority of repeats are in coding regions (**supplementary Table S1**). Most of the perfect repeats identified are matched by an imperfect one. Exceptions can arise when the imperfect repeat is identified with a different repeat unit. Typical examples are repeats of $\mathbf{a_n x}$ units (e.g. **aaag** and **aaaaat**) that can be identified as imperfect polyA stretches by TRF.

The length of the vast majority of the perfect repeats found is 12 nt (e.g. four 3-mer units), which was the minimum detectable SSR length allowed in our analysis. The longest average perfect repeat lengths are observed for the *E. coli* strains O157:H7 and O157:H7 EDL933, where the **agagcc** repeats are over 34 and 46 nt long on average, respectively.

Although we detected SSRs up to hexamer units, in accordance with the genomes under study being packed with protein-coding sequences, we focus primarily on trinucleotide repeats in this paper. Cumulated length per megabase data for the matching perfect and imperfect trinucleotide repeats (here the term ‘matching’ means those with strictly the same repeat class) were calculated for the four *E. coli* and seven *Chlamydial* genomes (**Tables 2 and 3 and supplementary Figure S1**).

3.2. Overview of trinucleotide repeats in the investigated genomes

Although the number of identified repeats is generally low (there are repeat classes represented by less than five or even only a single 12 nt copy in all genomes), several trends can clearly be identified. *Chlamydial* genomes tend to exhibit higher density of trinucleotide repeats than the *E. coli* strains (**Tables 2 and 3**). The difference can be more than twofold (*C. caviae*, ~345 bp/Mbp vs. *E. coli O157:H7 EDL33*, ~142 bp/Mbp for perfect repeats). (Considering all identified repeats, there are no striking differences between the overall repeat abundances but the *E. coli* genomes contain about 1.5 times more types of repeats (classes) than the *Chlamydial* ones (**supplementary Table S1**)). *E. coli* strains have a clear preference for **acc, agc, atc** and **ccg** repeats. As expected from their close evolutionary relationship, the two *O157:H7* strains exhibit highly similar repeat distributions. The highest imperfect/perfect ratio (≈ 1.75) is also found in these two genomes for **aac** repeats. This means that the imperfect microsatellites are on average 1.75 times longer than their matching perfect counterparts. All of the *Chlamydial* genomes are rich in **aag** repeats, although their SSR distribution is less homogeneous than that of *E. coli*. In *C. muridarum* **agc** is the most abundant repeat (**Table 3** and **supplementary Fig. S1**) and it is almost as abundant as **aag** in *C. caviae*, while **ccg** repeats were not detectable in *Chlamydiales* except for *C. caviae*. The three *C. pneumoniae* strains show highly similar repeat distributions. The largest imperfect/perfect ratio is found in the *C. pneumoniae* strains for **aac** repeats (>2.7) and in *C. muridarum* for the class **agc** (1.9).

The observed abundance of perfect trinucleotide repeats for all (i.e. coding plus non-coding) sequences is roughly two (for *Chlamydiales*) and three times (for *E. coli* strains) less than that found in all sequences in eukaryotes using similar recognition criteria (Tóth et al., 2000). Although this difference increases considerably (with a factor of two or three, depending on the taxon) when comparing bacterial data to eukaryotic coding sequences, this indicates that overall trinucleotide repeat density is not dramatically different in prokaryotic and eukaryotic genomes as a whole. Different abundance of repeats with identical GC-content (e.g. **aac** and **aag**, **agc** and **acg**) suggests that the overall repeat distribution is affected by selective forces and is not a result of nonspecific stochastic processes.

There are several interesting differences in the amino acid sequences coded by trinucleotide repeats when comparing C*hlamydiales* and *E. coli* strains (**supplementary Table S3**). Perfect **aag** repeats mostly code for glutamic acid residues in *Chlamydia* but for phenylalanine in *E. coli.* Although the former is the most abundant residue in all strains when considering imperfect repeats, the preference for phenylalanine in the *E. coli* genomes is still apparent. In the case of both perfect and imperfect **agc** repeats, the difference in the coded sequences can clearly be observed, as alanine is the most abundant in *Chlamydiales* and leucine in *E. coli* strains.

3.3. Comparison of perfect and imperfect trinucleotide repeats

Genomic distributions of all perfect and imperfect trinucleotide repeats found are available in **supplementary Fig. S2A and S2B**. As expected, imperfect repeats are detected at many more locations in the genomes than perfect ones. It is apparent that the much smaller genome size results in the absence of certain repeat classes (**aat**, **act**, **ccg**) from most strains of *Chlamydiales*, even though the normalized total abundance of SSRs is higher in *Chlamydiales* than in *E. coli*. To obtain a robust measure of the similarity of perfect and imperfect repeat distribution patterns, contingency analysis was performed on the abundance (number and total length of repeats per megabase) data of the matching perfect and imperfect trinucleotide repeats (**Tables 2 and 3**). *P* values around 1 correspond to high similarity of the two compared distributions while lower values indicate discrepancies, meaning that the relative abundance of imperfect repeat classes does not follow that of matching perfect ones (**Table 4**). While the perfect/imperfect distributions are highly similar in most genomes, they are largely different in *Chlamydia muridarum*. Closer inspection of the repeat data shows that this is due to the discrepancy in the total lengths of **agc** repeats. The total length of imperfect **agc** stretches is nearly twice as long as that of perfect ones, while other classes show no or only subtle differences in terms of perfect and imperfect repeat lengths (**supplementary Fig. S1**). We note that this strain also contains imperfect **aag** repeats in high abundance but the majority of them is not matched by perfect repeats with the same class. It is therefore very unlikely that they stem from perfect **aag** stretches. This example illustrates the importance of our approach of analyzing SSRs identified independently as both perfect and imperfect ones.

3.4. Comparison of repeat distributions in different bacteria

Distributions of matching perfect and imperfect trinucleotide repeats were also compared across bacterial genomes (**Table 5 and supplementary Table S2**). In general, intra-group similarities of *Chlamydial* and *E. coli* genomes are higher than the inter-group similarities between members of the two taxonomic groups, which are practically zero. The only notable exception is *C. muridarum*, which shows higher similarity to three *E. coli* genomes. This is only observable for the length per megabase, but not for the number per megabase measure. Within *Chlamydia*, the various *C. pneumoniae* strains are considerably more similar to each other in terms of repeat abundance than to other species within the group. These observations indicate different SSR preferences for the two phylogenetic groups investigated. Low probability values are also observed comparing different *Chlamydia* species.

3.5. Genes with high repeat content

The genes with the highest SSR content were identified and investigated in the eleven analyzed genomes. Only repeats identified both as perfect and imperfect SSRs were considered, but – in contrast to the analysis of trinucleotide repeats – SSRs with different repeat units were included (e.g. a long imperfect trinucleotide repeat can match multiple perfect hexanucleotide stretches, and in this case the imperfect repeat was not discarded). A number of the identified genes code for polymorphic outer membrane proteins or hypothetical proteins. In **supplementary Tables S4 and S5** we give details for the ten genes with the highest repeat content from all studied genomes.

When considering functional categories for the genes with high perfect repeat content (using the classification of the KEGG database, see section 2.3), it is apparent that genes in connection with Metabolism are by far the most abundant ones. Within this group, most genes are in connection with carbohydrate metabolism. Only eleven genes are classified as participating in information processing (environmental or genetic) and two as being involved in human diseases.

The picture becomes different when considering the genes with high imperfect repeat content. Here, genes involved in environmental information processing comprise the most abundant functional group, leaving metabolism-connected ones the second place. These observations can be rephrased as the relative length of perfect stretches in imperfect repeats of genes responsible for environmental information processing is lower than in metabolism-related coding sequences. It should be noted, however, that most genes are not yet classified in KEGG, thus cannot be included in this analysis (**Table 6**).

3.6. Gene-specific comparisons

We have selected some functionally well-characterized genes containing SSRs in at least two of the investigated genomes to highlight the importance of simultaneous analysis of perfect and imperfect repeats in the interpretation of evolutionary processes.

3.6.1. Gene tolA, coding for a cell envelope protein

The longest imperfect repeats found are located in the *tolA* gene that encodes a cell envelope protein in the four *E. coli* strains. This protein is not present in *Chlamydiales*. TRF classifies this repeat as an imperfect one with **agc** trinucleotide unit, while there is no matching perfect repeat in the same class. However, short perfect hexanucleotide repeat stretches can be found in this region. The sequence of the repeated 6-bp unit always contains the subsequence **agc** (e.g., **aagcag**). In two strains the imperfect repeat is 171 bp long while in the two *O157:H7* strains the repeated length is 291 bp. The *tolA* gene sequences from these four strains are highly similar with two isoforms, the shorter one being carried by the *O157:H7* strains. Although the long **agc** repeat is located around the 27-residue insertion, the longer SSR variants are found in the shorter isoforms. Closer inspection of the sequences reveals that a highly similar repeat region is detected differently in the two cases **(supplementary Fig. S3A and S4)**. Since *tolA* genes have been selected for biased amino acid frequency according to other studies (Rooney 2003) and no perfect repeats with matching repeat class were found in this gene, we suggest that the imperfect repeat at this locus was not derived from an ancestral perfect one but resulted from selection for codon frequency, although a scenario with repeat expansion cannot be completely ruled out.

3.6.2. Gene ftsK, coding for a cell division protein

The FtsK cell division protein is a DNA segregation ATPase. Its gene, *ftsK*, is among the most repeat-rich ones in terms both of perfect and imperfect SSRs in *E. coli* strains but not in *Chlamydiales*. Similarly to the *tolA* gene, *ftsK* is dominated by a long **agc** repeat matched only by perfect repeats of different hexanucleotide sequences **(supplementary Fig. S3B and S4).**

3.6.3. Gene coding for 50S (large subunit) ribosomal protein L7/L12

The *rplL* gene encoding the 50S (large subunit) ribosomal protein L7/L12 of *E. coli* contains a 45-bp long imperfect **agc** repeat with a 12-bp perfect stretch. As the nucleotide sequence is almost fully conserved, the repeats are identical in the four *E. coli* strains and correspond to the alanine-rich protein sequence AAAAVAVAAGPVEAA. In *Chlamydiales,* this gene also contains imperfect repeats of class **agc** or **aacagc**, generally not matched by perfect ones. Further data (including name, genome, NCBI-GI and ACCESSION, length of repeat) of genes with high repeat content is available in **supplementary Table S5**.

4. Discussion

Distribution of short tandem repeats in 11 bacterial genomes show various patterns characteristic for one or more strains under study. Although most perfect SSRs are only 12 nt long, longer imperfect repeats flanking them often indicate past expansion and demolition events. Other characteristic phenomena are the presence of preferred SSR classes and variations in imperfect and perfect SSR content in the investigated strains. This may reflect different evolutionary pressures. Low similarity of repeat distributions, especially in *Chlamydiales*, shows that SSR preferences can differ even between closely related strains, comprising a new feature which changes during speciation. Our results are consistent with the notion that although SSRs are less frequent in prokaryotes than in eukaryotes, they can nevertheless influence gene evolution (Ellegren, 2004).

When comparing imperfect and perfect repeats, the simplest assumption is that they follow the same overall genomic distribution, i.e. abundance data obtained for only perfect repeats represent well those of imperfect ones. This is true for most of the genomes used in our study with the notable exception of *C. muridarum* where the imperfect/perfect ratio of **agc** repeats is exceptionally high compared to other repeat classes. This discrepancy may reflect recent evolutionary changes if we assume that the majority of imperfect repeats with a perfect ‘core’ arose as a perfect repeat, and the present imperfect part is the result of disrupting mutations still allowing the recognition of the original SSR. It is also interesting that in terms of trinucleotide repeat distribution *Chlamydia muridarum* bears some resemblance to the studied *E. coli* genomes (**supplementary Table S2**), even though this weak similarity is only detectable at the total lengths of the repeats. It seems that the identified **agc** repeats are capable of expansion, 9 of the 13 affected genes contain both perfect and imperfect **agc** repeats (**supplementary Table S5**). The affected genes are either connected to translation (*infB* and *rplL*), or code for metabolism-related proteins (or have a yet unknown function).

Microsatellites can contribute to genetic variation by coding for amino acid repeats and also by being located in sequences that control gene expression (Kashi and King, 2006). The high fraction of coding sequences (and the corresponding relative abundance of tri- and hexanucleotide repeats) suggests that the former mechanism can be prevalent in bacteria. We have shown that the coding preferences of trinucleotide repeats are different among strains, and these differences are observable for both perfect and imperfect repeats (**supplementary Table S3**). This indicates active roles for SSRs in bacterial polymorphisms, consistent with earlier suggestions (van Belkum et al., 1998). We note that prevalence of tri- and hexanucleotide repeats in coding regions was also recently reported for *Mycoplasma hyopneumoniae* (Mrazek, 2006), which suggests that our findings have more general relevance in prokaryotes.

Polymorphisms caused by repeated sequences were previously assessed for several *Chlamydial* genomes (Rocha et al., 2002). Our results obtained using a different approach confirmed those findings. Among the genes with high SSR density, those for which function could be assigned are either responsible for essential cellular processes or code for polymorphic membrane proteins. For example, the FtsK protein is the essential bacterial ATPase that is responsible for the correct segregation of daughter chromosomes during cell division (Iyer et al., 2004). The presence of multiple SSRs in this gene might allow rapid evolution of some segments of the protein in *E. coli*, but not in *Chlamydial* strains.

Our observation that among genes with high repeat content, those involved in environmental information processing have higher imperfect/perfect ratio than genes with other functions (**Table 6**), raises the possibility of accelerated evolution of these coding sequences. This suggestion is supported by the polymorphic outer membrane proteins in *Chlamydial* genomes. In *Chlamydiales*, genes with relatively large repeat content in more than one strain almost exclusively encode polymorphic outer membrane proteins (Pmps) (**supplementary Table S5**) that are often associated with pathogenicity (Gomes et al., 2004; Carlson et al., 2005). It is interesting that *E. coli* TolA, also a membrane-associated protein, contains a very long, although imperfect repeat that is polymorphic in the investigated strains. In this case the presence of the repeat can most probably be attributed to codon selection pressure, although it cannot be ruled out that this repeat was capable of expansion/reduction during the evolutionary history of these genes. In accordance with other analyses concerning tandem repeats (de Castro et al., 2006) we suggest that SSR polymorphisms contribute to the variability of membrane proteins in prokaryotes.

In order to investigate the effect of horizontal gene transfer (HGT) events on the distributions of the repeats, we examined all genes introduced from different donor species into the common ancestor (Ortutay et al. 2003) of the *Chlamydial* strains and investigated their repeat content (SSR data are shown in **supplementary Table S6**). We note that no differences regarding matching perfect and imperfect trinucleotide repeats are detected (i.e. differences are only observed for repeats not matched by the other type). Thus, we believe that our analysis presented in the manuscript does not suffer from biases caused by HGT events.

We have shown that *E. coli* and *Chlamydial* strains exhibit characteristic SSR distribution with a marked relative abundance of tri- and hexanucleotide repeats. Simultaneous analysis of perfect and imperfect SSRs revealed discrepancies in repeat class distribution in *C. muridarum* indicative of recent evolutionary events. We suggest that microsatellites contribute to the genomic variability of prokaryotes even in closely related genomes, as supported by the observed difference in SSR distribution patterns even between related genomes. Genes related to environmental information processing, and in particular, the *Chlamydia*-specific Pmp proteins are candidates of sequences where SSRs can contribute to genetic variation in the strains examined.

Acknowledgements

G.T. was supported by a János Bolyai postdoctoral fellowship from the Hungarian Academy of Sciences.

References

Azad, R.K., Lawrence, J.G., 2007. Detecting laterally transferred genes: use of entropic clustering methods and genome position. Nucleic Acids Res. 35, 4629-4639.

Blattner, F.R., et al., 1997. The complete genome sequence of Escherichia coli K-12. Science 277, 1453-1474.

van Belkum, A., Scherer, S., van Alphen, L., Verbrugh, H., 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev. 62, 275-293.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573-580.

Carlson, J.H., Porcella, S.F., McClarty, G., Caldwell, H.,D., 2005. Comparative genomic analysis of Chlamydia trachomatis oculotropic and genitotropic strains. Infect Immun. 73, 6407-6418.

Carugo, O., Pongor, S., 2002. Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. J. Mol. Biol. 315, 887-898.

de Castro, L.A., Rodrigues, Pedroso T., Kuchiishi, S.S., Ramenzoni, M., Kich, J.D., Zaha, A., Henning Vainstein, M., Bunselmeyer Ferreira, H., 2006. Variable number of tandem aminoacid repeats in adhesion-related CDS products in Mycoplasma hyopneumoniae strains. Vet. Microbiol. 116, 258-269.

Chen, Y., Timms, P., Chen, Y.P., 2007. CIDB: Chlamydia Interactive Database for cross-querying genomics, transcriptomics and proteomics data. Biomol Eng. *in press*

Eckert, K.A., Yan, G., 2000. Mutational analyses of dinucleotide and tetranucleotide microsatellites in Escherichia coli: influence of sequence on expansion mutagenesis. Nucleic Acids Res. 28, 2831-2838.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5, 435-445.

Gáspári, Z., Ortutay, C., Tóth, G., 2007. Divergent microsatellite evolution in the human and chimpanzee lineages. FEBS Lett. 581, 2523-2526.

Gomes, J.P., Bruno, W.J., Borrego, M.J., Dean, D., 2004. Recombination of the genome of Chlamydia trachomatis involving the polymorphic membrane protein C gene relative to ompA and evidence for horizontal gene transfer. J. Bacteriol. 186, 4295-4306.

Hayashi, T., et al., 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8, 11-22.

Iyer, L.M., Makarova, K.S., Koonin, E.V., Aravind, L., 2004. Comparative genomics of the FtsK - HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res. 32, 5260-5279.

Jurka, J, Pethiyagoda, C., 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. 40, 120-126.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W. and Stephens, R.S., 1999. Comparative genomes of Chlamydia pneumoniae and C. trachomatis. Nature Genet. 21, 385-389.Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354-357.

Kashi, Y., King, D., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253-258.

Lindstedt, B.A., 2005. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis 26, 2567-2582.

McNally, D., Fares, M.A., 2007. In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae. BMC Evol. Biol. 7:81.

Metzgar, D., Thomas, E., Davis, C., Field, D. Wills, C., 2001. The microsatellites of Escherichia coli: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. Mol. Microbiol. 39, 183-190.

Mrazek, J., 2006. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Mol. Biol. Evol. 23, 1370-1385.

Mrazek, J., Guo, X., Shah, A., 2007. Simple sequence repeats in prokaryotic genomes. Proc. Natl. Acad. Sci. USA 104, 8472-8477.

NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/)

Noller, A.C., McEllistrem, M.C., Pacheco, A.G.F., Boxrud, D.J., Harrison, L.H., 2003. Multilocus Variable-Number Tandem Repeat Analysis distinguishes outbreak and sporadic Escherichia coli O157:H7 isolates. J.Clin. Microbiol. 41, 5389-5397.

Ortutay, C., Gáspári, Z., Tóth, G., Jáger, E., Vida, G., Orosz, L., Vellai, T., 2003. Speciation in Chlamydia: genomewide phylogenetic analyses identified a reliable set of acquired genes. J. Mol. Evol. 57, 672-680.

Perna, N.T., et al., 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 409, 529-533.

Read, T.D., et al., 2000. Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. Nucleic Acids Res. 28, 1397-1406.

Read, T.D., et al., 2003. Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. Nucleic Acids Res. 31, 2134-2147.

Rocha, E.P., Pradillon, O., Bui, H., Sayada, C., Denamur, E., 2002. A new family of highly variable proteins in the Chlamydophila pneumoniae genome. Nucleic Acids Res. 30, 4351-4360.

Rooney, A.P., 2003. Selection for highly biased amino acid frequency in the TolA cell envelope protein of Proteobacteria. J. Mol. Evol. 5, 731-736.

Schlotterer, C., Imhof, M., Wang, H., Nolte, V., Harr, B., 2006. Low abundance of Escherichia coli microsatellites is associated with an extremely low mutation rate. J. Evol. Biol. 19, 1671-1676.

Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S., Nakazawa, T., 2000. Comparison of whole genome sequences of Chlamydia pneumoniae J138 from Japan and CWL029 from USA. Nucleic Acids Res. 28, 2311-2314.

Stephens, R.S., Kalman, S., Lammel, C.J., Fan, J., Marathe, R., Aravind, R., Mitchell, W.P., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W., 1998. Genome sequence of an obligate intracellular pathogen of humans, Chlamydia trachomatis. Science 282, 754-759.

Thompson, D., Higgins, D.G., Gibson, T.J., 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penaltiesand weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967-981.

Welch, R.A., et al., 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. USA 99, 17020-17024.

Tables

Table 1

Genomes used for this study.

Strain	**Accession (GenBank)**	**RefSeq identifier**	**Total length (bp)**	**Length of coding regions (bp)**	**GN[*]**
Chlamydia muridarum Nigg	**AE002160**	NC_002620	1072950	961248	***cmu***
Chlamydia trachomatis ser. D	**AE001273**	NC_000117	1042519	936164	***ctr***
Chlamydophila caviae GPIC	**AE015925**	NC_003361	1173390	1046055	***cca***
Chlamydophila pneumoniae AR39	**AE002161**	NC_002179	1229858	1090813	***cpa***
Chlamydophila pneumoniae CWL029	**AE001363**	NC_000922	1230230	1085960	***cpn***
Chlamydophila pneumoniae J138	**BA000008**	NC_002491	1226565	1097297	***cpj***
Chlamydophila pneumoniae TW-183	**AE009440**	NC_005043	1225935	1102622	***cpt***
Escherichia coli K12	**U00096**	NC_000913	4639221	4048916	***eco***
Escherichia coli O157:H7	**BA000007**	NC_002695	5498450	4819150	***ecs***
Escherichia coli O157:H7 EDL933	**AE005174**	NC_002655	5528445	4820481	***ece***
Escherichia coli CFT073	**AE014075**	NC_004431	5231428	4600495	***ecc***

[*]GN: Genome identifier according to the KEGG Database

Table 2

Total length per megabase of matching perfect and imperfect trinucleotide repeats in four complete *E. coli* genomes (bp/Mbp).

	eco[*]		***ecc***		***ecs***		***ece***	
	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect
aac	12.93	19.18	6.88	7.46	9.28	16.19	9.22	16.10
aag	12.93	14.87	13.76	16.06	8.73	10.37	8.68	10.31
aat	10.35	10.35	6.88	6.88	10.91	10.91	10.85	10.85
acc	28.45	29.10	38.99	43.01	19.64	24.55	19.54	24.42
acg	12.93	12.93	6.88	6.88	6.55	8.18	6.51	8.14
act	2.59	2.59	2.29	2.29	2.18	2.18	2.17	2.17
agc	28.45	37.51	25.23	32.69	26.19	35.47	28.22	37.44
atc	31.04	42.46	32.11	45.69	32.74	42.38	32.56	42.69
ccg	32.33	42.46	38.42	50.85	24.55	28.92	24.42	28.76
ALL	**172.00**	**211.45**	**171.44**	**211.81**	**140.77**	**179.15**	**142.17**	**180.88**

[*] The 3-letter codes correspond to the genome identifiers (GN) of the KEGG Database, see Table 1.

Only matching repeats with identical repeat class are summarized.

Table 3

Total length per megabase of matching perfect and imperfect trinucleotide repeats in seven complete *Chlamydial* genomes (bp/Mbp).

	cmu[*]		***ctr***		***cca***		***cpa***		***cpn***		***cpj***		***cpt***	
	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect	perfect	imperfect
aac	47.53	55.92	34.53	34.53	40.91	46.02	9.76	26.83	9.75	26.82	9.78	26.90	9.79	26.92
aag	55.92	61.51	103.59	151.56	132.95	168.74	82.94	107.33	102.42	137.37	102.73	137.78	102.78	137.85
aat	0.00	0.00	25.90	31.65	10.23	39.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
acc	0.00	0.00	0.00	0.00	30.68	35.79	9.76	9.76	9.75	9.75	9.78	9.78	9.79	9.79
acg	13.98	13.98	14.39	14.39	10.23	10.23	9.76	9.76	9.75	9.75	9.78	9.78	9.79	9.79
act	11.18	11.18	23.02	23.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
agc	114.64	218.09	48.92	48.92	43.46	53.69	29.27	39.03	29.26	39.02	29.35	39.13	29.36	39.15
agg	22.37	22.37	23.02	23.02	10.23	10.23	51.23	53.66	51.21	53.65	51.36	53.81	51.39	53.84
atc	11.18	11.18	34.53	61.39	56.25	63.92	39.03	46.35	39.02	46.33	39.13	46.47	39.15	46.49
ccg	0.00	0.00	0.00	0.00	10.23	10.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	276.80	**394.23**	**307.90**	**388.48**	**345.17**	**438.05**	**231.75**	**292.72**	**251.16**	**322.69**	**251.91**	**323.65**	**252.05**	**323.83**

[*] The 3-letter codes correspond to the genome identifiers (GN) of the KEGG Database, see Table 1.

Only matching repeats with identical repeat class are summarized.

Table 4

Probability of identity values of matching perfect/imperfect trinucleotide repeats in all regions of the genomes.

Values below 0.5 are marked bold.

Strain*	***cmu***	***ctr***	***cca***	***cpa***	***cpn***	***cpj***	***cpt***	***ecc***	***eco***	***ecs***	***ece***
length per megabase	**0.30**	0.89	1.00	0.94	0.92	0.92	0.92	0.98	0.99	0.99	0.99
number per megabase	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

*The 3-letter codes correspond to the genome identifiers (GN) of the KEGG Database, see Table 1. Higher values indicate closer similarity on a scale of 0 to 1 (see section 2.4 for details).

Table 5

Similarity of repeat distribution patterns in different genomes measured by probability of identity values.

A). All regions

	cmu	ctr	cca	cpa	cpn	cpj	cpt	ecc	eco	ecs	ece
cmu		0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.045	0.554	0.412
cmu		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ctr	0.000		0.627	0.033	0.064	0.064	0.064	0.000	0.000	0.000	0.000
ctr	0.001		0.730	0.043	0.097	0.097	0.097	0.000	0.000	0.000	0.000
cca	0.000	0.992		0.257	0.162	0.162	0.162	0.000	0.000	0.000	0.000
cca	0.000	0.730		0.206	0.158	0.158	0.158	0.000	0.000	0.000	0.000
cpa	0.000	0.150	0.960		0.908	0.908	0.908	0.000	0.000	0.000	0.000
cpa	0.000	0.043	0.206		0.895	0.895	0.895	0.000	0.000	0.000	0.000
cpn	0.000	0.256	0.900	0.946		1.000	1.000	0.000	0.000	0.000	0.000
cpn	0.000	0.097	0.158	0.895		1.000	1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.256	0.900	0.946	1.000		1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.097	0.158	0.895	1.000		1.000	0.000	0.000	0.000	0.000
cpt	0.000	0.256	0.900	0.946	1.000	1.000		0.000	0.000	0.000	0.000
cpt	0.000	0.097	0.158	0.895	1.000	1.000		0.000	0.000	0.000	0.000
ecc	0.083	0.000	0.000	0.000	0.000	0.000	0.000		0.890	0.362	0.297
ecc	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.890	0.369	0.304
eco	0.144	0.000	0.000	0.000	0.000	0.000	0.000	0.833		0.957	0.945
eco	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.861		0.953	0.940
ecs	0.265	0.000	0.000	0.000	0.000	0.000	0.000	0.442	0.955		1.000
ecs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.301	0.953		1.000
ece	0.331	0.000	0.000	0.000	0.000	0.000	0.000	0.366	0.935	1.000	
ece	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.249	0.940	1.000	

B) Coding regions

	cmu	ctr	cca	cpa	cpn	cpj	cpt	ecc	eco	ecs	ece
cmu		0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.177	0.152
cmu		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ctr	0.000		0.479	0.033	0.050	0.050	0.050	0.000	0.000	0.000	0.000
ctr	0.000		0.549	0.043	0.070	0.070	0.070	0.000	0.000	0.000	0.000
cca	0.000	0.629		0.257	0.213	0.213	0.213	0.000	0.000	0.000	0.000
cca	0.000	0.549		0.206	0.190	0.190	0.190	0.000	0.000	0.000	0.000
cpa	0.000	0.125	0.960		0.985	0.985	0.985	0.000	0.000	0.000	0.000
cpa	0.000	0.043	0.206		0.983	0.983	0.983	0.000	0.000	0.000	0.000
cpn	0.000	0.186	0.970	0.997		1.000	1.000	0.000	0.000	0.000	0.000
cpn	0.000	0.070	0.190	0.983		1.000	1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.186	0.970	0.997	1.000		1.000	0.000	0.000	0.000	0.000
cpj	0.000	0.070	0.190	0.983	1.000		1.000	0.000	0.000	0.000	0.000
cpt	0.000	0.186	0.970	0.997	1.000	1.000		0.000	0.000	0.000	0.000
cpt	0.000	0.070	0.190	0.983	1.000	1.000		0.000	0.000	0.000	0.000
ecc	0.218	0.000	0.000	0.000	0.000	0.000	0.000		0.829	0.494	0.497
ecc	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.827	0.510	0.507
eco	0.544	0.000	0.000	0.000	0.000	0.000	0.000	0.780		0.781	0.848
eco	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.867		0.772	0.836
ecs	0.890	0.000	0.000	0.000	0.000	0.000	0.000	0.497	0.904		1.000
ecs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.451	0.772		1.000
ece	0.941	0.000	0.000	0.000	0.000	0.000	0.000	0.498	0.954	1.000	
ece	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.475	0.836	1.000	

Perfect repeats are compared in the upper triangle, imperfect ones in the lower triangle of the matrices. Lines with grey background contain statistics concerning the total length per megabase, lines with white background refer to the number of repeats per megabase. Higher values indicate closer similarity on a scale of 0 to 1 (see section 2.4 for details). 3-letter codes correspond to the genome identifiers in KEGG (Table 1.)

Table 6

Distribution of the ten genes containing the longest repeats from the 11 investigated genomes by functional categories according to KEGG categorization.

KEGG Functional category	**Number of genes with perfect repeats**	**Number of genes with imperfect repeats**
Category Metabolism	**29**	**8**
Subcategory Carbohydrate metabolism	9	2
Subcategories Carbohydrate metabolism and Amino acid metabolism	4	1
Subcategories Carbohydrate metabolism and Environmental information processing	3	1
Subcategories Carbohydrate metabolism and Energy metabolism	1	
Subcategories Carbohydrate metabolism and Starch and sucrose metabolism	1	
Subcategories Amino acid metabolism and Genetic information processing	5	1
Subcategories Ezyme families and Genetic information processing	4	
Subcategory Glycan biosynthesis and metabolism	4	
Subcategory Lipid metabolism	2	2
Subcategories Nucleotide (purine and pyrimidine) metabolism and Replication and repair	2	
Subcategory Metabolism of cofactors and vitamins	2	
Subcategories Energy metabolism	1	1
Category Environmental information processing	**6**	**15**
Subcategories Environmental information processing and Cellular processes	5	
Subcategory Environmental information processing	1	
Subcategories Environmental information processing and Cellular processes (cell division)		4
Subcategory Membrane Transport/ABC transporters		4
Subcategory Membrane Transport/Pores ion channels		4
Subcategory Membrane Transport/Other ion-coupled transporters		2
Subcategory Membrane Transport//Phosphotransferase system		1
Category Genetic information processing	**5**	**4**
Subcategories Genetic information processing and Replication and repair	3	
Subcategories Genetic information processing and Translation	3	3
Subcategory Genetic information processing	1	1
Category Human Diseases	**2**	**2**
Unclassified in KEGG	**68**	**81**

Supplementary material

Table S1: Repeat distribution statistics for the investigated bacteria. [xls]

A: *Escherichia coli CFT073,* **B:** *Escherichia coli K12,* **C:** *Escherichia coli O157:H7 EDL933,* **D:** *Escherichia coli O157:H7* **E:** *Chlamydia_muridarum,* **F:** *Chlamydia trachomatis,* **G:** *Chlamydophila caviae,* **H:** *Chlamydophila pneumoniae AR39***,**
I: *Chlamydophila pneumoniae CWL029,* **J:** *Chlamydophila pneumoniae J138,*
K: *Chlamydophila pneumoniae TW 183.*

Table S2: Repeat distribution similarity calculations for perfect-imperfect repeats in each bacterium and all-against-all comparison of bacterial repeats. [xls]

A: Summary of the probability values obtained, **B:** Details of the calculations (excluded classes, Chi2 values etc.)

Table S3: Statistics of amino acid repeats coded by the trinucleotide repeats identifie [xls]

A: perfect trinucleotide repeats, **B:** imperfect trinucleotide repeats

Table S4: Data of top 10 genes with highest repeat content [xls]

A: top 10 genes with highest perfect repeat content, **B:** top 10 genes with highest imperfect repeat content

Table S5: Details of genes with high repeat content. [xls]

Table S6: SSRs detected genes introduced into the common ancestor of the Chlamydial strains investigated [xls]

Fig. S1: Total length per megabase of trinucleotide repeats [pdf]

Fig. S2: Graphical summary of the genomic distributions of the individual trinucleotide repeat classes [pdf]

A: *E. coli* strains, **B:** *Chlamydia* strains

Fig. S3: Alignment of the **agc** repeat region of selected genes in the 4 *E. coli* genomes [pdf]

A: *tolA* gene, **B**: *ftsK* gene

Fig. S4: Full sequence alignments of the *tolA* and *ftsK* genes in the 4 *E. coli* strains. [pdf]