

Measuring Genome Divergence in Bacteria: A Case Study Using Chlamydian Data

Daniel A. Dalevi,¹ Niklas Eriksen,² Kimmo Eriksson,³ Siv G.E. Andersson¹

¹ Department of Molecular Evolution, University of Uppsala, Norbyvagen 18C, SE-751 36, Uppsala, Sweden

² Department of Mathematics, Royal High School of Technology, SE-100 44, Stockholm, Sweden

³ Department of Mathematics and Physics, Mälardalens högskola, SE-721 23, Västerås, Sweden

Received: 16 December 2000 / Accepted: 27 November 2001

Abstract. We have studied the relative contribution of inversions, transpositions, deletions, and nucleotide substitutions to the evolution of *Chlamydia trachomatis* and *Chlamydia pneumoniae*. The minimal number of rearrangement events required for converting the gene order structure of one genome into that of the other was estimated to 59 ± 6 events, including 13% inversions, 38% short inversions, and 49% transpositions. In contrast to previous findings, no examples of horizontal gene transfer subsequent to species divergence were identified, nor any evidence for an excessive number of tandem gene duplications. A statistical model was used to compare nucleotide frequencies for a set of genes uniquely present in one species to a set of orthologous genes present in both species. The two data sets were not significantly different, which is indicative of a low frequency of horizontal gene transfer events. This is based on the assumption that a foreign gene of different nucleotide content will not have become completely ameliorated, as verified by simulations of the amelioration rate at twofold and fourfold degenerate codon sites. The frequencies of nucleotide substitutions at twofold and fourfold degenerate sites, deletions, inversions, and translocations were estimated to 1.42, 0.62, 0.18, 0.01, and 0.01 per site, respectively.

Key words: Amelioration — Chlamydia — Deletions — Inversions — Rearrangements — Substitutions — Transpositions

Introduction

Questions concerning bacterial evolution have so far primarily been attacked by phylogenetic reconstructions based on individual gene sequences. The results of such studies have provided information about the evolutionary histories of the analyzed genes, but these may not necessarily reflect the history of the organisms—if such a history can be defined. The increasing number of fully sequenced genomes now enables a more thorough analysis of gene flows and mutations affecting genomic architectures. Important questions concern the causes and consequences of genomic rearrangements and whether horizontal gene transfer events have occurred to such an extent that phylogenetic reconstructions based on individual gene sequences are doomed to failure (Doolittle 1999).

Comparative genomics have already made it apparent that the evolution of microbial genomes cannot solely be modeled by small-scale mutations, such as substitutions and local insertions/deletions. Large-scale rearrangement events play a crucial role in shaping the sizes and architectures of genomes. By superimposing complete genomic structures onto established phylogenetic relationships, it may be possible to reconstruct the order of rearrangements

and/or to determine the relative frequencies at which transposition, deletion, insertion, and inversion events occur. Such measures may then be incorporated into global methods for phylogenetic reconstructions based on gene order structures rather than on individual gene sequences.

The simplest way of using genomic information for phylogenetic purposes is to identify shared, atypical gene order structures and use these as diagnostic tools for sets of strains, species, genera, or subdivisions. A more sophisticated approach is to align all genes in a segment or a genome and determine the minimal number of events (inversions, transpositions, deletions, and insertions) required for converting the order of genes in one genome into that of the other. This distance-based approach has so far mainly been applied to mitochondrial and chloroplast genomes, some of which are very recombinogenic and therefore ideal for this kind of analysis. For example, Sankoff and others (1992) reconstructed a mitochondrial phylogeny based on edit distances derived from pairwise comparisons of gene orders and insertion/deletion frequencies.

The number of breakpoints between two genomes is another general measure of genomic distances that does not require any prior understanding of the biological nature of the different types of rearrangement events (Watterson et al. 1982). This methodology has recently been applied to studies of animal phylogeny, as inferred from mitochondrial gene order structures (Blanchette et al. 1999). Previous attempts to find optimal methods for reconstructing the ancestral genome structures on a tree with a fixed topology have been made by iterative heuristics (Blanchette et al. 1997; Sankoff and Blanchette 1998). An unresolved task in all of these methods is to distinguish the true evolutionary pathway from the multitude of reconstructions that are equally optimal. A partial solution is to extract information about invariant aspects of the reconstructions (Blanchette et al. 1999).

Whereas inversions and transpositions can in principle be examined and quantified by methods such as those described above, horizontal gene transfer events are harder to trace. Unless an alignment of gene orders of several closely related species is available, recently introduced genes may only be detected if they are in some way different from the genes of the host genome in which they reside. The methods currently used to identify aliens are based on codon biases and nucleotide frequency statistics (Lawrence and Ochman 1998; GarciaValle et al. 1999; Garcia-Vallve et al. 2000). For example, Lawrence and Ochman (1998) estimated the proportion of horizontally transferred genes to be about 18% in *Escherichia coli* based on an analysis of nucleotide frequency statistics, codon usage patterns, and the codon adaptation index.

Pairs of closely related genomes represent the best source of data for studies of the frequencies at which rearrangement events occur. Fortunately, more and more genomes from closely related strains and species are being sequenced. For example, the 1.0 Mb genome of *Chlamydia trachomatis* and the 1.2 Mb genome of *Chlamydia pneumoniae* have recently been published (Stephens et al. 1998; Kalman et al. 1999; Benson et al. 2000). The genome of *C. pneumoniae* contains 1052 genes as compared to 894 genes in the genome of *C. trachomatis*. To date as many as three strains of *C. pneumoniae* (CWL039, AR39, J138) and two strains of *C. trachomatis* (MoPN, serovar D) have been sequenced (Read et al. 2000; Shirai et al. 2000). A dot plot analysis of *C. trachomatis* and *C. pneumoniae* has shown that numerous chromosomal rearrangements have occurred since the two genomes diverged from each other (Read et al. 2000). The most notable of these rearrangements is two large DNA inversions at the regions surrounding the origins of replication and termination (Fig. 1). The difference between the *C. trachomatis* strains is mainly confined to a 50 kb plasticity zone near the origin of replication (Read et al. 2000). Likewise, the sole difference between the *C. pneumoniae* strains CWL029 and AR39 is an inverted DNA segment upstream of the uridine kinase gene and AR39 contains a novel 4523 nucleotide circular single-stranded phage not found in CWL029 (Read et al. 2000).

The purpose of this paper was to quantify the frequencies of insertions, deletions, inversions, and transpositions in relation to the frequencies of nucleotide substitutions in the *Chlamydia* genomes. The results of our analysis indicate that the major factors influencing the structure of the *Chlamydia* genomes are nucleotide substitutions and deletions. We have found no evidence of frequent horizontal gene transfer events, nor for an excessive number of genes that are tandemly repeated in any of the two genomes.

Materials and Methods

Inference of genome rearrangements. The genome sequences of *C. trachomatis* (AE001273) and *Chlamydia pneumoniae* CWL029 (AE001363) were obtained from GenBank. A list of gene orders was obtained from the *Chlamydia* Genome Project (Stephen, personal communication). A gene order permutation is available on request. Gene order permutations were analyzed using the program Derange II (Blanchette et al. 1996). Derange II attempts to minimize a weighted sum of three operations: transpositions, inversions, and inverted transpositions, where different weights give different results. The program was extended to distinguish between short and long operations (Eriksen et al. unpublished). The optimization of weights was done as described (Eriksen et al. unpublished).

The binomial mutation model. Following Sueoka (1962), we assume that the choice of GC or AT at a given third codon position is made by an independent binomial experiment with the proba-

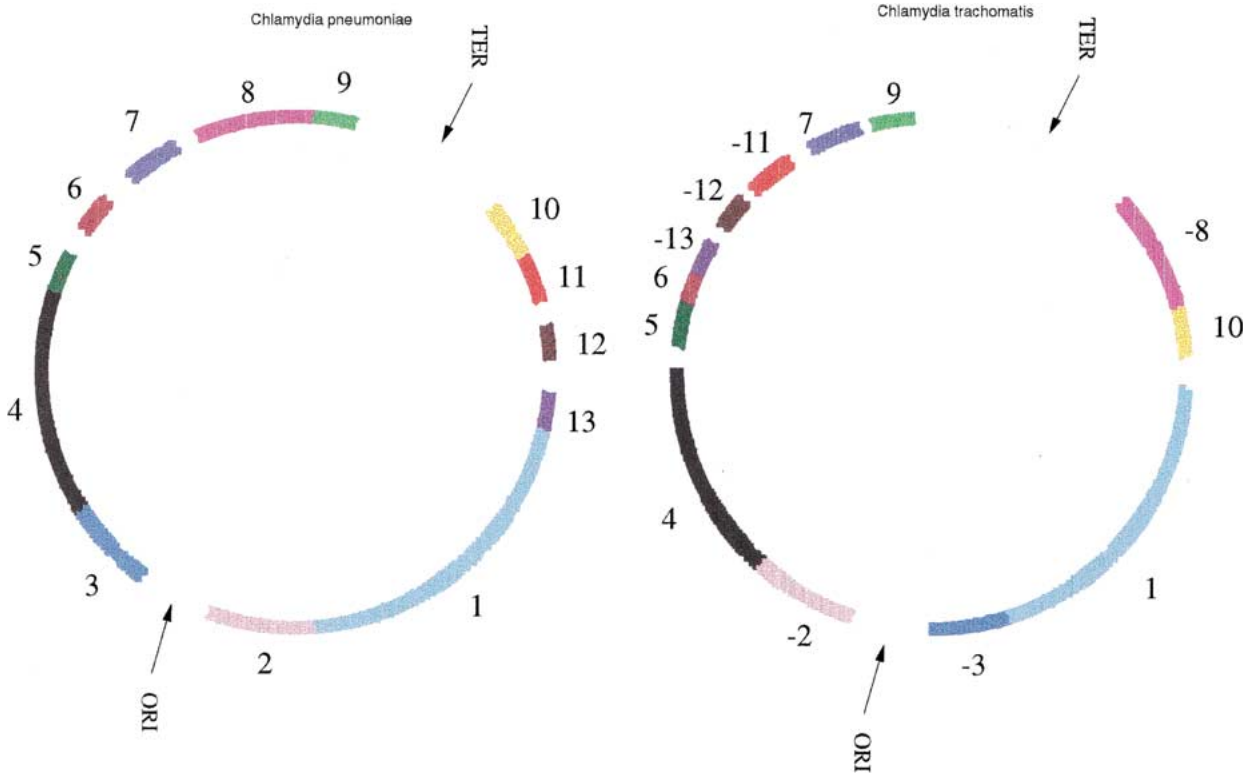


Fig. 1. Schematic illustration of gene order structures in *Chlamydia trachomatis* and *Chlamydia pneumoniae*. ORI = origin of replication; TER = termination of replication.

bility given by the GC3 content. We obtain the following statistical model: With x being the GC3 content of a certain gene of length n codons in a position where the probability of GC is p , the stochastic variable nx has $Bin(n, p)$ distribution. Since n is large (more than 100), we may approximate with the normal distribution $N(np, \sqrt{np(1-p)})$. We can normalize to an $N(0,1)$ distribution of “ q -values” by defining:

$$q = \frac{x - p}{\sqrt{p(1-p)/n}} \quad (1)$$

The GC3 content shows a significant large-scale variation over the genomes in question, so we estimate the probability of GC at a given position by the GC3 content in a window of 5000 nucleotides centered on the gene of interest. A smaller window may cause overfitting of our model to the data, while a larger may average out local effects.

For comparison of q -value distributions to each other, the Kolmogorov-Smirnov method (see for example p. 670, Edward and Mishra 1988) was used. In this test, a deviation measure K is defined as the maximal difference between the cumulative distributions, scaled by the square root of N , the number of sampled values:

$$K = \max |F(x) - F_H(x)|\sqrt{N}, \quad (2)$$

where $F(x)$ is the sampled distribution and $F_H(x)$ is the expected distribution. The higher K -value, the more significantly different are the distributions.

Substitution rates. The rates of synonymous substitutions were estimated from the synonymous divergences using standard models, such as the Kimura 2-parameter model (K2P). In highly A + T biased genomes, like *Chlamydia*, it is important to use substitution models that allow for nucleotide biases. With sufficient amount of data it is possible to apply the most general model where all substitution rates (k_{AC}, k_{AB}, \dots where k_{AC} is the rate with which A is

replaced by C) may be different, but assuming that AG and TC differences are equivalent (Berg 1995). For the twofold and the fourfold degenerate codons these rate matrices apply:

$$Q_{2f} = \begin{bmatrix} \cdot & \frac{k_{GA}+k_{CT}}{2} \\ \frac{k_{AG}+k_{TC}}{2} & \cdot \end{bmatrix}, \quad Q_{4f} = \begin{bmatrix} \cdot & k_{CA} & k_{GA} & k_{TA} \\ k_{AC} & \cdot & k_{GC} & k_{TC} \\ k_{AG} & k_{CG} & \cdot & k_{TG} \\ k_{AT} & k_{CT} & k_{GT} & \cdot \end{bmatrix} \quad (3)$$

Similar rate matrices can be calculated for the synonymous substitutions in the threefold and sixfold degenerate codons (Berg 1995). As in other substitution models, the matrices provide estimates of the substitutions rates based on the observed nucleotide differences in pairwise sequence alignments and considering all possible pathways (Berg 1995). Here, it is assumed that the rates are the same in both lineages, and if not, an average value is calculated. Thus, no information is provided about the direction of events from the ancestral node.

Results

Minimal Genome Rearrangement Scenarios

A gene order based alignment of the two *Chlamydia* genomes using a gap size of 10 genes identifies 43 segments that are located in different relative positions on the chromosomes. These have an average of 19 genes per segment, ranging from two genes to 170 genes. The locations and orientations of the 13 largest segments are schematically shown in Fig. 1. Within these rearranged segments, many local rearrange-

Table 1. Proportion of rearrangement operations separating the two *Chlamydia* genomes

Operation	Proportion (%)
Inversion	13
Short Inversion	38
Transposition	49

The rearrangement frequencies were inferred from a modified version of Derange II (Blanchette et al. 1996).

ments of different types have occurred. To estimate the contribution of inversions, transpositions and inverted transpositions to the observed rearrangements within and among segments, we calculated the number of architectural changes required for reconstructing one genome from the other. These measures provide a minimal estimate of the number of rearrangement events that have occurred since the two genomes diverged.

A gene order permutation was obtained by comparing the exact gene order information of the two genomes. It has previously been shown that short inversions, which only affect a single gene, occur predominantly at the recombination boundaries (Tillier and Collins 2000b). Short inversions were considered separately with the help of a modified version of the previously developed software Derange II (Blanchette et al. 1996). This program minimises a weighted sum of three operations: transpositions, inversions, and inverted transpositions, where different weights yield different results (Blanchette et al. 1996). The modified version of the program is also capable of distinguishing between local and long distance operations (Eriksen et al. unpublished). Weights were taken from the best results of a simulated data set of about 750 genes, i.e. comparable to the *Chlamydia* data set (Eriksen et al. unpublished).

The proportions of rearrangement events required to convert the order of genes in one genome into that of the other are presented in Table 1. The relative frequencies of inversion and transposition events are 51% and 49%, respectively. The uncertainty level in our experiment is about 5% of the true value for all operations, if the mean of the transpositions and the inverted transpositions are considered. The total number of events was estimated to 59 ± 6 events for the *Chlamydia* data (Table 1). Short rearrangements represented the majority of sorting events (Fig. 2), with single gene inversions ranging from 77 bp to 1610 bp, with an average of 444 bp in *C. trachomatis* and 454 bp in *C. pneumoniae*. This is not significantly different from the mean gene size in *C. trachomatis* (350 bp) and *C. pneumoniae* (345 bp), if taking into account that the data set with the inverted genes may be slightly biased due to a few single gene inversions of large sizes.

It has previously been shown that inversions and translocations occur in a symmetric manner (Read

et al. 2000; Tillier and Collins 2000b), as schematically illustrated by the symmetric location of fragments 3, 8, 11, 12, and 13 in Fig. 1. Indeed, inverted blocks of more than 10 genes and transposed blocks of more than five genes were predominantly sorted by Derange II in a symmetric manner during the conversion of one genome into the other, while shorter rearrangement events appeared to be more randomly distributed. For example, of 10 inversions of more than two genes, seven were sorted in a symmetric manner. Likewise, 19 out of 27 translocations were accounted for by symmetric reorganizations.

The neighborhoods and chromosomal locations of a representative set of single gene inversions identified by Derange II within segments and at rearrangement boundaries are schematically shown in Figs. 3 and 4. At least 14 of the 22 single gene inversions performed during the sorting procedure can unambiguously be identified as real single gene inversions, as inferred from their genomic neighborhoods. Three of these are flanked by genes with similar orientation (Fig. 3A), while four inversions are flanked by genes oriented in either a convergent or divergent manner (Fig. 3B and C). Five single gene inversions are located at the boundaries of rearranged segments (Fig. 3E), as also observed by Tillier and Collins for several recombination boundaries (Tillier and Collins 2000b).

Taken together, our analysis demonstrates that inferences of minimal rearrangement scenarios by automatic procedures not only provides good distance measures but also reconstructs events that are biologically meaningful. The overall rearrangement frequency since the divergence of *C. trachomatis* and *C. pneumoniae* was estimated to less than one large inversion event per 100 genes and one translocation event per 100 genes on the average.

Low Frequency of Tandemly Repeated Genes

It has been previously reported that the presence of tandemly repeated genes is a characteristic feature of the *Chlamydia* genomes (Read et al. 2000). To examine this in more detail, each genome was searched internally for paralogs using each individual gene as a query sequence. Indeed, hits with significant scores were often observed to neighbouring genes. However, a closer inspection revealed that most sequences (with the exception of *oppA1/2*, CPn0369/370, CPn0537/538, CPn0842/843, and CPn1006/1007 and the *pmp* genes) did not align over their entire length as would be expected for duplicated genes. The observed scores were found to be the result of short overlapping stretches of sequences at the ends of neighboring genes. Thus, contrary to previous reports (Read et al. 2000) we were unable to find any evidence for an

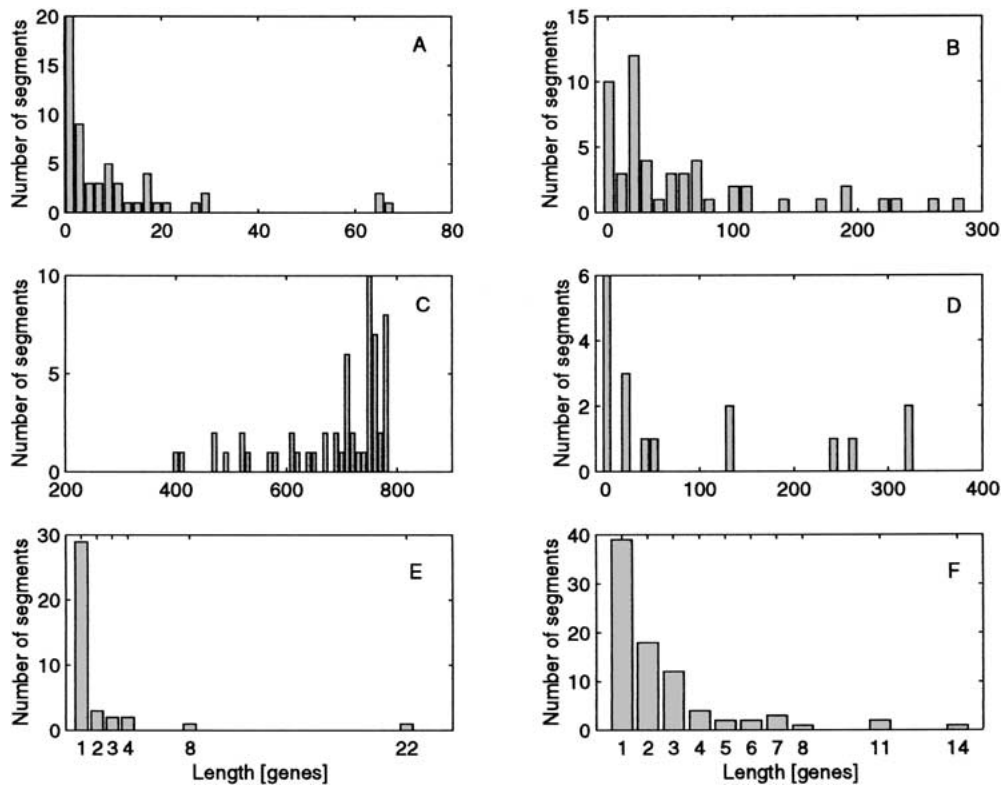


Fig. 2. Size distribution of (A–C) transpositions, (D) inversions, and (E–F) deletions. The transposition of each segment results in the creation of three novel fragments. The size distribution of (A) the shortest, (B) the middle, and (C) the longest of these novel fragments are shown. The size distribution of deletions is estimated from the size distribution of unique genes in (E) *C. trachomatis* and (F) *C. pneumoniae*.

unusually high frequency of tandemly repeated genes in the *Chlamydia* genomes.

In total, we have identified 11 genes that are duplicated twice or more in *C. pneumoniae* but represented by single gene sequences in *C. trachomatis*. The most highly duplicated gene is *pmp*, which codes for a polymorphic outer membrane protein. One of the *pmp* genes in *C. trachomatis* (CT871) is present in 14 copies in *C. pneumoniae*. All of the 14 copies exhibit a base composition pattern that is typical of *C. pneumoniae* genes (data not shown). However, in the absence of an out-group, intra-genomic duplications in one genome can not be distinguished from deletions in the other genome. Taken together, the data suggests that the duplication frequency is less than one gene duplication per 100 genes on the average since the divergence of *C. trachomatis* and *C. pneumoniae*.

Placing Lateral Gene Transfers in Perspective

Frequencies of lateral gene transfers have previously been inferred from analyses of nucleotide frequency statistics. For example, Lawrence and Ochman used G+C content values and codon usage data to estimate the fraction of lateral gene transfers in *E. coli* (1997, 1998) and other bacteria (Ochman et al. 2000).

Likewise, Garcia-Vallvé et al. (2000) used a statistical analysis of nucleotide patterns that also takes into account parameters such as gene position and amino acid content to identify horizontally transferred genes. However, a major limitation with these and other similar methods is the difficulty to define cut-off levels that distinguish the outliers from the typical genes since the overall distribution is not fully known (Koski et al. 2001).

The availability of complete genome sequence data from two or more closely related species offers a new perspective to this problem. Since it is unlikely that the same set of genes would have been acquired independently at homologous positions, candidates for recent transfers should be searched among the set of genes that are uniquely present in either of the two genomes (Koski et al. 2001). The task is then reduced to distinguishing between genes lost in one genome versus those introduced in the other genome. Our approach for solving this problem is to examine the extent to which the unique genes are equilibrated with the internal mutation bias of the orthologous genes.

There are 70 genes uniquely present in *C. trachomatis* and another 214 genes uniquely present in *C. pneumoniae*. These are quite randomly distributed around the genome (Fig. 5). A majority of the unique genes are singletons (45–65%), with the average number of genes per cluster being 2.3 in *C. tracho-*

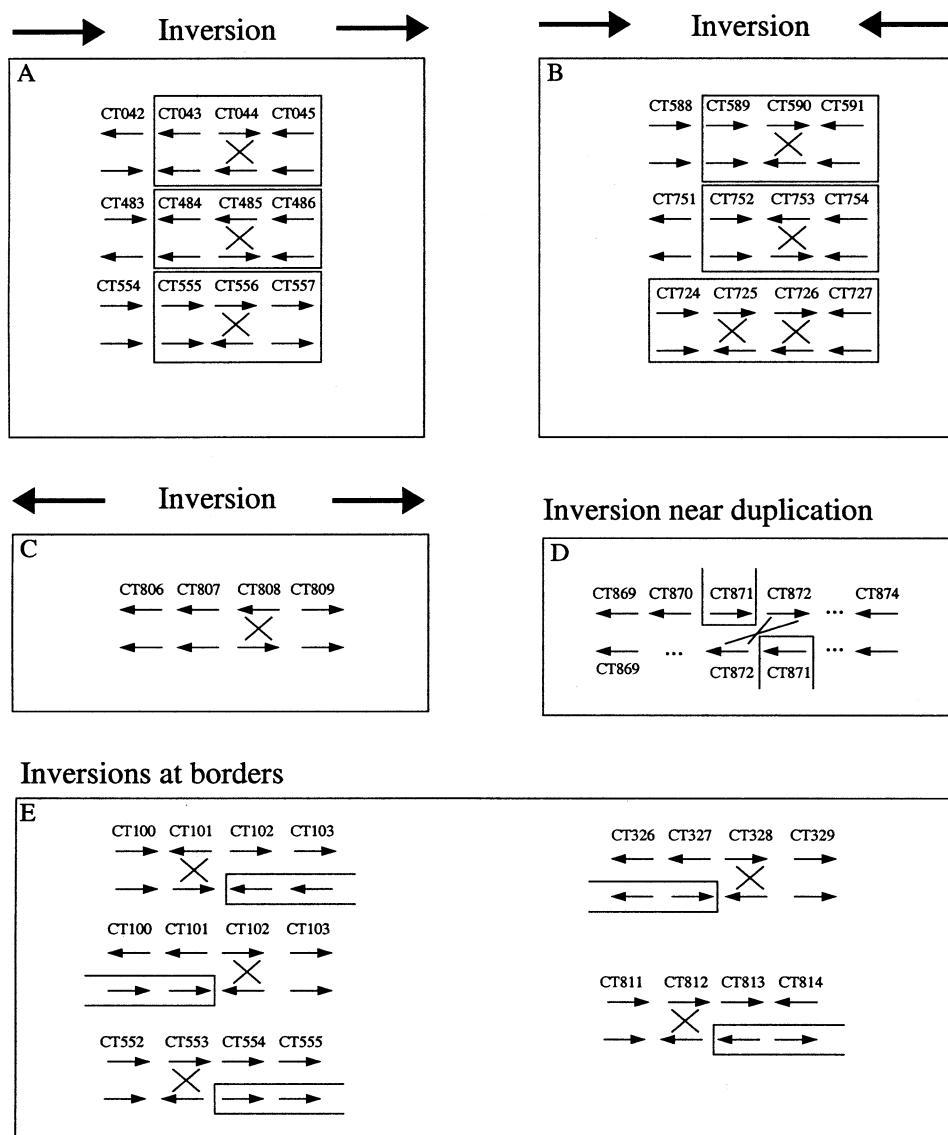


Fig. 3. Examples of short inversions in the *Chlamydia* genomes as inferred by Derange II. The arrows indicate the direction of transcription of neighboring genes.

matis and 2.5 in *C. pneumoniae* (Fig. 2). The largest cluster of unique genes contains 22 genes in *C. trachomatis* and 14 genes in *C. pneumoniae*, respectively. Correspondence analysis (Johnson and Wichern 1992) was used to investigate whether the non-orthologous genes share some unusual codon usage characteristics. No clusters were identified for those genes (Fig. 6A), suggesting that the unique genes as a group are not associated with abnormal nucleotide frequency statistics.

To quantify any deviations from genome-specific base composition features, we applied the binomial mutation model (Equation 1) to the data set of unique genes. Here, the orthologous genes were used as our reference data set for genes that have not been horizontally transferred since the two *Chlamydia* species diverged. The orthologous genes displayed essentially the same q -distribution as the unique genes

(Fig. 6B); the deviation measures $K = 0.95$ in *C. pneumoniae* and $K = 0.77$ in *C. trachomatis* as would appear by chance in 15% and 30% of the experiments, respectively. This suggests that a majority of the unique genes have a base composition pattern that is compatible with the overall genome-specific codon usage pattern.

However, this does not per se imply that the unique genes have been vertically transmitted from a shared ancestral genome, since the earliest gene transfers may have undergone complete amelioration. To approach this problem, we have studied substitution frequencies and estimated the time it takes for a foreign gene to become completely ameliorated at its synonymous sites. The mean K_a and K_s values for 474 orthologous gene sequences were estimated to be 0.21 and 1.22, respectively, with Li's method (1993). Since the K_s value is close to satu-

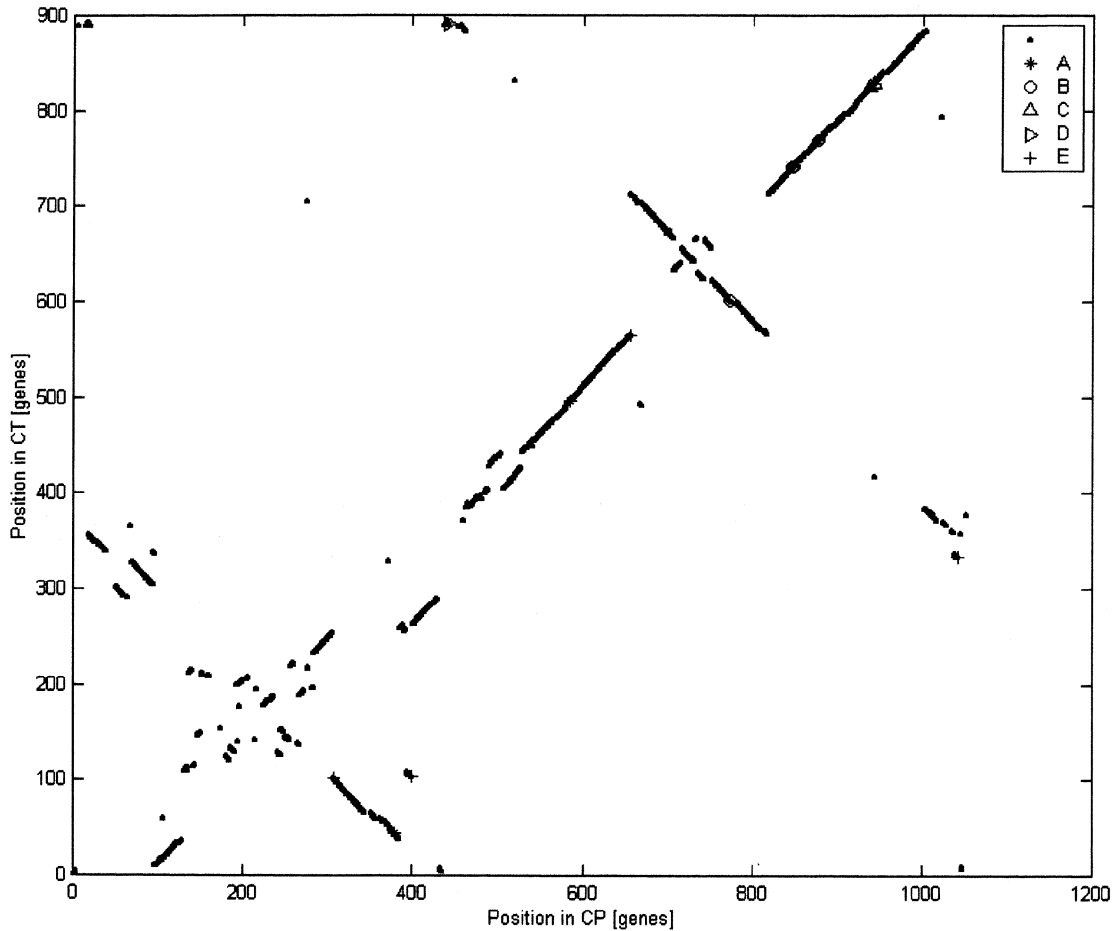


Fig. 4. Chromosomal location of the short inversions in *C. trachomatis* and *C. pneumoniae*. The x-axis represents the order of genes in *C. trachomatis* and the y-axis the order of genes in *C. pneumoniae*. The positions of short inversions are indicated. Letters refer to the categories of inversions described in Fig. 3.

ration, the complete substitution matrices at twofold and fourfold degenerate sites was estimated using a correction model that accounts for base composition variation (Berg 1995). For the twofold and fourfold degenerate sites in the 474 aligned orthologous gene pairs the rate matrices can be described as:

$$Q_{2ft} = \begin{bmatrix} \cdot & 0.89 \\ 0.48 & \cdot \end{bmatrix},$$

$$Q_{4ft} = \begin{bmatrix} \cdot & 0.21 & 0.61 & 0.43 \\ 0.14 & \cdot & 1.11 & 0.40 \\ 0.38 & 0.98 & \cdot & 0.12 \\ 0.55 & 0.74 & 0.25 & \cdot \end{bmatrix} \quad (4)$$

where t is the divergence time. From these (Berg 1995) the mean number of substitutions per synonymous site was estimated to $Ks^{(2f)} = 0.62$ and $Ks^{(4f)} = 1.42$. The profile of substitutions in the rate matrix indicates that there is a strong bias for conversions between G and C nucleotides, as also observed by Rocha and Danchin (2001).

To minimize the effects of strand-specific mutational biases, the orthologous gene set was divided

into two sets of genes located on either of the two strands. For the set of genes located on the leading strands (253 alignments) and the lagging strands (181 alignments) in both genomes the following matrices were found:

$$Q_4^{\text{leading}}{}_t = \begin{bmatrix} \cdot & 0.18 & 0.62 & 0.43 \\ 0.11 & \cdot & 0.99 & 0.43 \\ 0.41 & 1.07 & \cdot & 0.14 \\ 0.57 & 0.72 & 0.28 & \cdot \end{bmatrix},$$

$$Q_{4f}^{\text{lagging}}{}_t = \begin{bmatrix} \cdot & 0.20 & 0.63 & 0.42 \\ 0.16 & \cdot & 1.34 & 0.51 \\ 0.35 & 0.92 & \cdot & 0.08 \\ 0.53 & 0.80 & 0.17 & \cdot \end{bmatrix} \quad (5)$$

The exchange rates for G and C nucleotides are very high and near to saturation on both strands. As a result, inverted genes are expected to equilibrate more rapidly with the GC skew values of their new host strand (Tillier and Collins 2000a, 2000c) than the horizontally inserted genes will equilibrate with the overall G + C content.

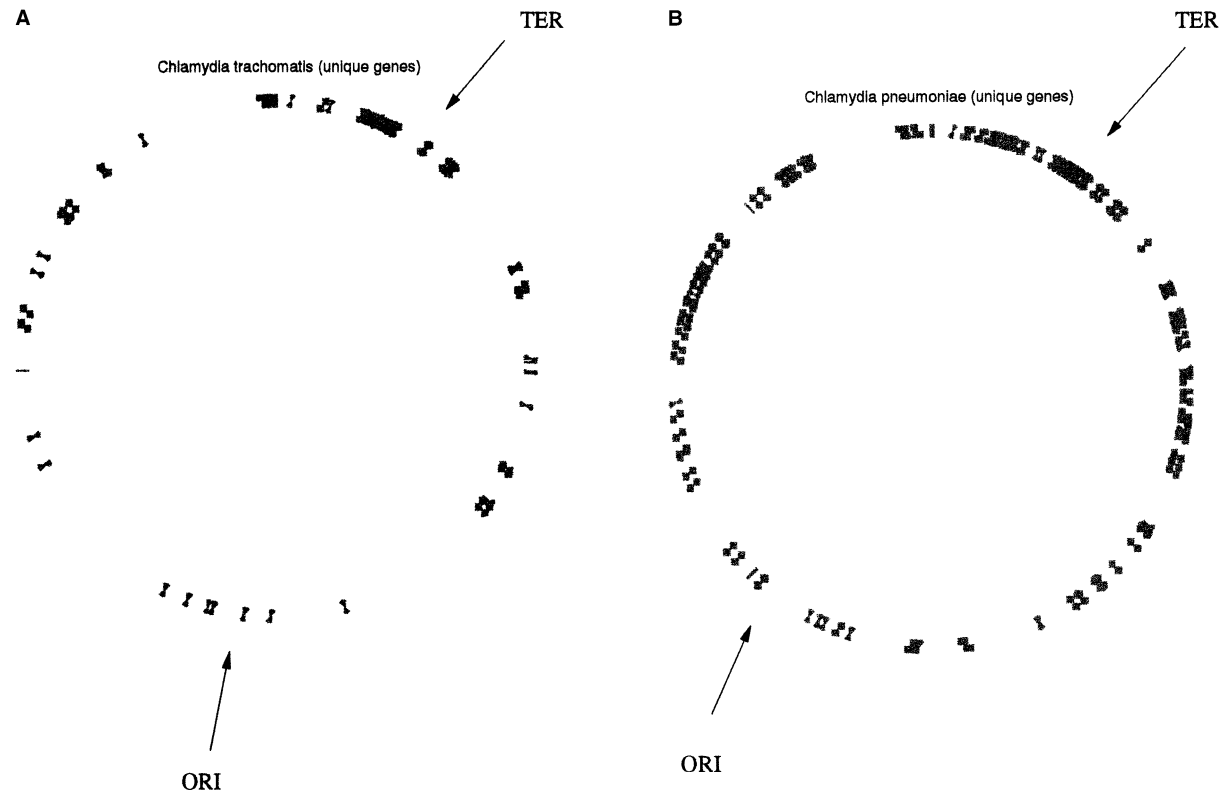


Fig. 5. Schematic figure showing the chromosomal locations of genes uniquely present in (A) *C. trachomatis* and (B) *C. pneumoniae*.

To search for selective constraints on nucleotide substitutions, we compared the substitution matrices for the highly expressed set of genes to the complete set of genes. The twofold degenerate set is particularly useful for this purpose since they are not sensitive to strand specific effects on nucleotide substitutions. Based on a set of putatively highly expressed genes (ribosomal protein genes and elongation factor genes) we found a real-value answer for the twofold degenerated codons ($Ks^{(2f)} = 0.84$), which is slightly higher than the mean frequency of substitutions per synonymous site for the complete set of genes ($Ks^{(2f)} = 0.62$). Similar results were also obtained for the fourfold degenerate sites using the Kimura two-parameter model [$Ks^{(2f)} = 1.14$ (full data set), $Ks^{(4f)} = 1.13$ (leading strand), and $Ks^{(4f)} = 1.05$ (highly expressed genes)]. Thus, in contrast to previous findings (Romero et al. 2000), we found no significant difference in synonymous substitution frequencies for genes with putatively different expression profiles. This suggests that selection plays only a weak, if any role in determining synonymous substitution frequencies.

Finally, we used the substitution matrices to estimate the time before a foreign gene is completely ameliorated at its synonymous sites. The following differential equation describes the mutation model:

$$\frac{dP}{dt} = \frac{1}{2}QP(t) \quad (6)$$

which has the solution:

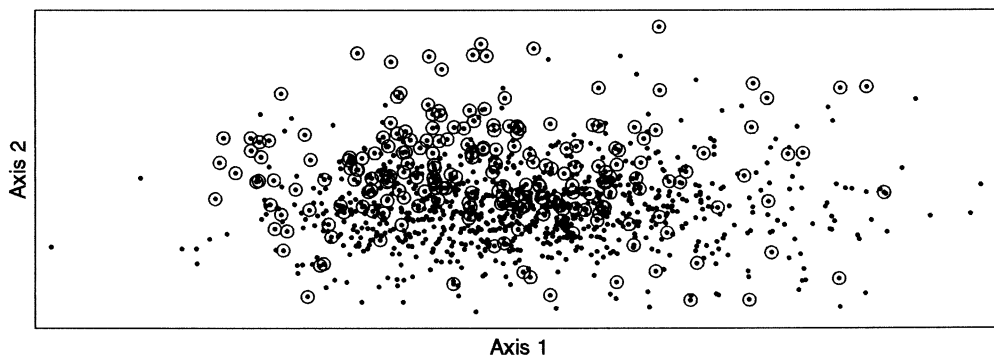
$$P(t) = e^{\frac{1}{2}Qt} P(0) \quad (7)$$

where $P(t)$ is a vector describing nucleotide concentrations and $P(0)$ is the initial value from when the gene enters the cell (Berg 1995). Using this information the following relation was found,

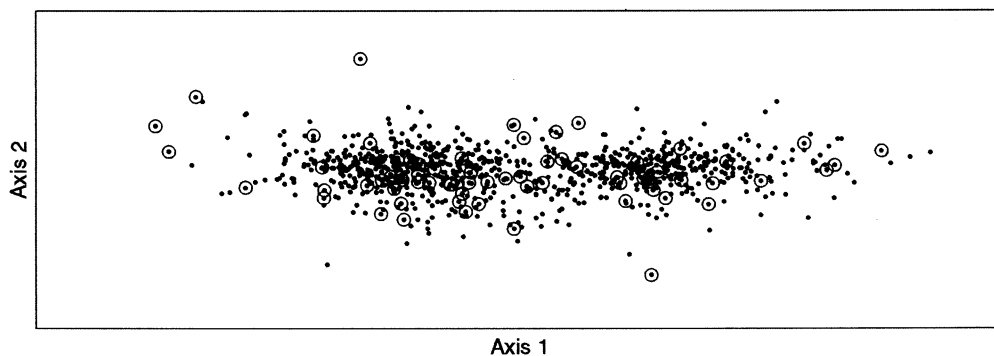
$$GC_{3s}(t) = GC_{3s}^{t \rightarrow \infty} + (GC_{3s}^{t=0} - GC_{3s}^{t \rightarrow \infty})e^{-t/\tau} \quad (8)$$

All times are in units of the time of divergence and $GC_{3s}^{t \rightarrow \infty}$ represents the G + C content at third codon position of the host, i.e. the value that all foreign genes will equilibrate towards, and $1/\tau$ is a combination of the rate constants for GC and AT nucleotide substitutions. To determine the value of τ we have simulated the changes in G + C content from a variety of starting conditions ($GC^{t=0}$) using the data in the substitution matrices for twofold and fourfold degenerate sites (see above). The simulations provide a series of GC content values at any given t for each starting point. By inserting these values and plotting the logarithm of the equation against time, the numerical value of $1/\tau$ can be estimated from the slope.

A



B



C

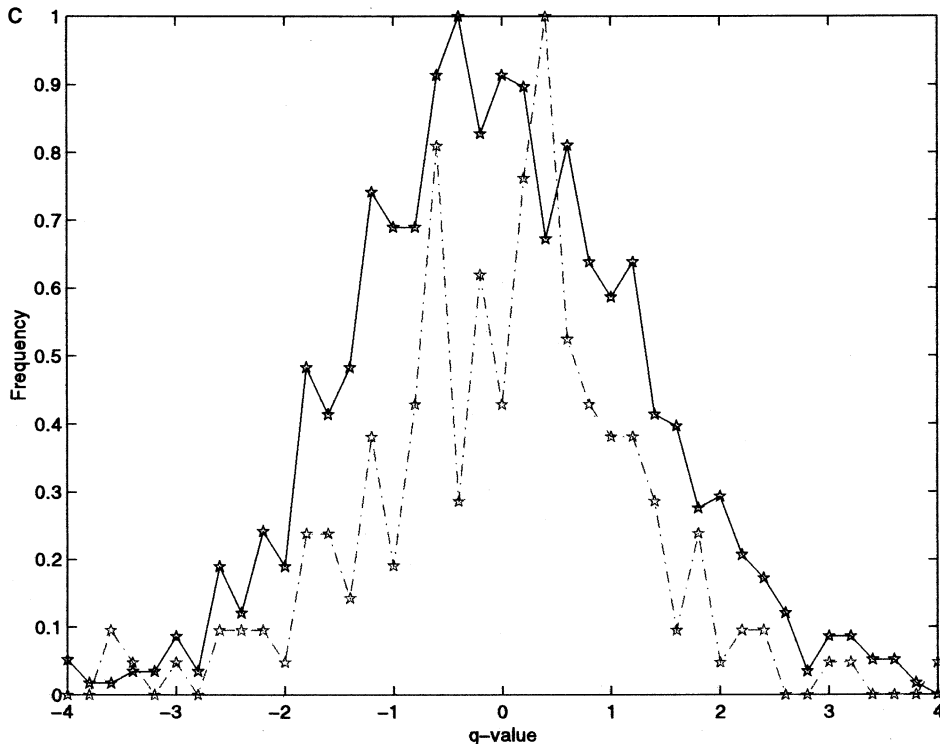


Fig. 6. Comparison of codon usage features in unique and orthologous genes. Plot of the two most prominent axes generated by the COA of the RSCU values from the (A) *C. pneumoniae* and (B) *C. trachomatis* genomes. The open circles correspond to genes

uniquely present in one species. (C) Distribution of codon usage patterns in unique genes versus orthologous genes. The q -values (Equation 1) are plotted on the x-axis for both set of genes and the frequency on the y-axis.

The standard deviation of $1/\tau$ was determined by a jack-knifing procedure with 100 replication steps. Here, the substitution matrix was recalculated 100

times based on a random set (50%) of the alignment. From this simulation procedure the following values were obtained:

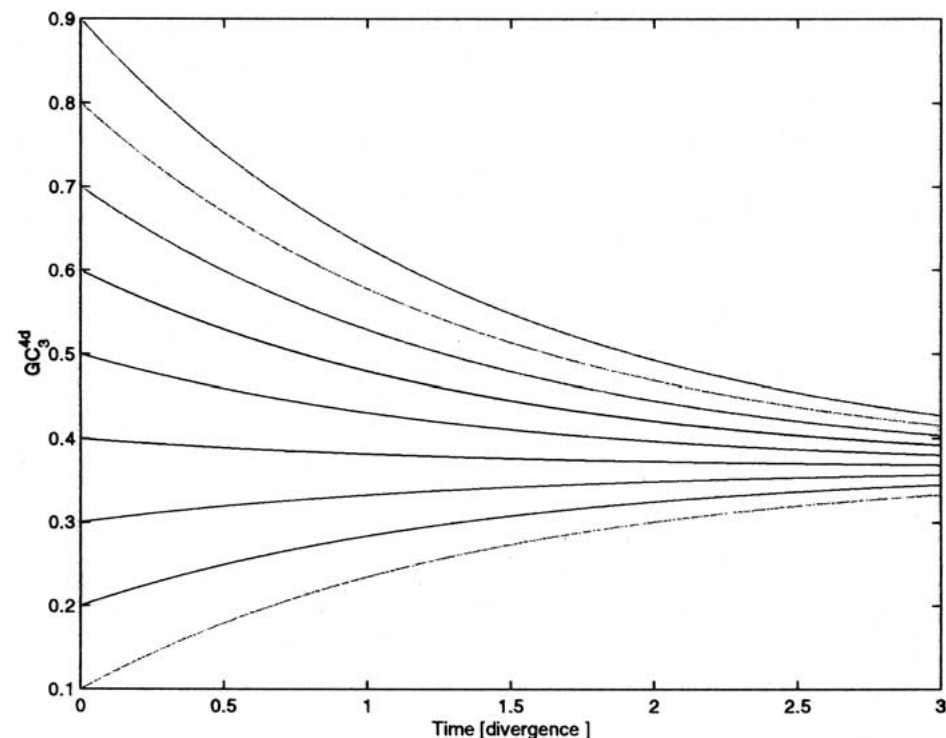


Fig. 7. Rate of amelioration in G + C content at fourfold degenerate sites for sequences with different initial G + C content values.

$$\begin{cases} \tau_{2f} = 1.4 \pm 0.07 \\ \tau_{4f} = 1.4 \pm 0.07 \end{cases} \quad (9)$$

Figure 7 shows the approximate rate of amelioration in G + C content for a set of different initial values ($GC_{3f}^{i=0}$).

The results suggest that genes with significantly different G + C content values prior to insertion would have been detected as foreign genes in the modern *Chlamydia* genomes.

Since we were unable to obtain any evidence for horizontal gene transfers in *Chlamydia*, we suggest that most of the genes unique to one species are the result of deletions in the other genome. If so, approximately 225,000 nucleotides may have been deleted, which corresponds to a deletion frequency of more than 10 genes per 100 genes per genome since the divergence of *C. trachomatis* and *C. pneumoniae*.

Discussion

The fixation rates for nucleotide substitutions are well characterised for numerous genes in many organisms, including several bacterial pathogens. The availability of complete genome sequence data for closely related bacterial species now enables us to measure with the same precision the frequencies at which other genomic mutations occur. How do the rates of rearrangements (inversions and translocations) and indels (insertions and deletions) compare to the rates

of nucleotide substitutions? To what extent do horizontal gene transfer events contribute to the changes of microbial genomes? Can rates of genomic mutations be used to describe the divergence of any two genomes in a way that takes into account all of the different processes whereby microbial genomes evolve?

A major problem when trying to approach these questions concerns the reliability of the methods used for inferring rearrangement and horizontal gene transfer events. For example, it has often been argued that sorting algorithms such as those used by Derrange are of limited biological value since they only provide minimal estimates of divergences and the sorting steps may be quite different from the actual rearrangement events. Indeed, it should be stressed that several rearrangement scenarios can be reconstructed for genomes such as those of *Chlamydia* that are separated by 50 or more rearrangement events. A particular problem is that we do not know a priori what kind of weights to use for the sorting operations. For the analysis of the *Chlamydia* dataset, we have devised a simulation scheme to find appropriate weights (Eriksen et al. unpublished). Using these weights, we estimated the frequencies of large inversions and translocations to one event per 100 genes, respectively.

So, how does this automatic sorting procedure correlate with putative rearrangement events inferred by manual comparison of the two *Chlamydia* genomes? The lessons learnt from our case study shows

that inferences of minimal rearrangement scenarios based on circa 50 sorting steps not only provides good distance measures but also reconstructs rearrangements events that are biologically meaningful. For example, derange II successfully sorted a majority of the differently organized gene blocks in a symmetrical manner, in accordance with the suggestion that the DNA at the replication fork is particularly vulnerable to recombination events (Tillier and Collins 2000b). Likewise, the analysis provides evidence for a strong bias towards short inversions, in accordance with previous suggestions (Tillier and Collins 2000b).

Horizontal gene transfer is another process that induces major genomic alterations. Phylogenetic methods have been used to search for and confirm selected examples of horizontal transmission (Koski et al. 2001), based on the assumption that the transferred genes will have an evolutionary history that is different from that of the host genome. The recent development of tools for automatic, large-scale phylogenetic reconstructions has simplified these analyses, making it possible to trace the evolutionary history not just of a few selected genes, but of all genes in a genome (Sicheritz-Ponten and Andersson 2001). Such a “phylogenomic” approach has shown that a majority of bacterial genes display a phylogenetic relationship that is characteristic of the domain in which they reside, suggesting that the separation of Bacteria and Archaea into two domains is robust (Sicheritz-Ponten and Andersson 2001).

However, there are many problems associated with these large-scale phylogenetic approaches concerning the accuracy of the methods. For example, many gene comparisons violate some of the fundamental principles required for phylogenetic reconstructions, such that the functional constraints should be the same for all genes being compared, which is far from always the case. The interpretations of the resulting trees are therefore very complicated and atypical placements of genes in phylogenetic trees are not necessarily an indication of horizontal gene transfer events. Vice versa, the lack of phylogenetic support for transfer is not proof that it did not occur, since transfers between closely related species can not be detected by this method. Not surprisingly, there is currently a heated debate about the extent to which horizontal gene transfer events have occurred in microbial evolution (Doolittle 1999; Kurland 2000).

An alternative method that has also been used to search for recent horizontal gene transfer events is to identify genes with base composition features that deviate from the normal pattern. This is based on the assumption that recently introduced genes are likely to have codon usage patterns that reflects the genome from which they were derived, rather than the host genome in which they reside. Unfortunately, also this

method has its problems and limitations, as discussed in a recent reclassification of horizontal gene transfer events in *Escherichia coli* and *Salmonella typhimurium* based on positional orthologs, codon features, and phylogenetic methods (Koski et al. 2001). For example, it is extremely difficult to define a cutoff level a priori that distinguishes genes with a “normal” pattern from genes with an “abnormal” pattern. And even if such a cutoff level can be defined, there may be many different explanations for any observed deviations, such as different selective pressures and/or temporary gene inactivation events. Another complication is that, although the frequency of nucleotides is relatively homogenous over bacterial genomes, the variation may still be on a scale that affects the results. Genes also have different lengths and short genes have greater stochastic variation than long genes. For all of these reasons, identifying horizontal gene transfer events by using measures based on codon preference statistics is not a trivial task.

This is illustrated by the contradictory results of our study and a recent study of the frequencies of horizontal gene transfers in a number of microbial genomes, including *Chlamydia* (Garcia-Vallvé et al. 2000). Our analysis is based on the assumption that the best candidates for recent horizontal gene transfers should be found in the sets of genes that are uniquely present in one of the two *Chlamydia* genomes. Most of these are hypothetical genes or genes that are not found outside the genus *Chlamydia*, making it unlikely that they have been introduced from a foreign source by horizontal gene transfer. Indeed, we have shown that the G + C content values at third codon positions for the unique genes follow the same distribution function as for the orthologous genes. Furthermore, a study of amelioration rates suggests that alien genes with a significantly different G + C content prior to the insertion would have been identified as foreign genes, even if the insertion occurred shortly after the divergence of the two species (Fig. 7). Based on these results we conclude that no or only few horizontal gene transfer events have occurred into the *Chlamydia* genomes from other distant genomes.

This result is in striking contrast to that of Garcia-Vallvé et al. (2000) who identified 55 genes in *C. pneumoniae* and 36 genes in *C. trachomatis* with atypical base composition features, which they interpret as recent horizontal gene transfer events. A closer inspection of the gene sets identified as putative transfers shows that 41 of the 55 genes in *C. pneumoniae* have orthologs in *C. trachomatis*, and vice versa for 31 of the 36 putative transfers in *C. trachomatis*. In other words, 75% of the putative transfers in *C. pneumoniae* and 86% in *C. trachomatis* have orthologs in the other species. These values are similar to the overall genomic fraction of orthologous

genes (75% and 90%, respectively), suggesting that the genes identified as putative transfers represent a random set of the total gene pool. Indeed, out of the 31 and 41 putative transfers with orthologs in both genomes only five are shared between the two, which we interpret as a negative result of the integrity of the method used and/or to insufficient data.

Garcia-Vallvé et al. (2000) suggests that the results of their analysis should be taken as conservative predictions of lateral gene transfers. Our results suggest just the opposite. Likewise, only a few putative transfers were identified in an analysis of dinucleotide frequencies in the *Chlamydia* genomes (Hooper and Berg 2002). Considering the lifestyle of *Chlamydia*, low frequencies of horizontal gene transfer events are to be expected since the interior of eukaryotic cells represents a very isolated growth habitat and the probability that other bacteria are present in the same cells as those invaded by *Chlamydia* should be very low. For all of these reasons, it seems more likely that a majority of the genes uniquely present in one species represent recent gene loss in the other species. This result is also consistent with the observed mutation bias for deletions observed in studies of neutral sequence evolution in the obligate intracellular parasite *Rickettsia prowazekii* and its close relatives (Andersson and Andersson 1999a, 1999b, 2001). Indeed, most obligate intracellular bacteria are thought to have undergone reductive genome evolution with massive losses of genetic information and extensive rearrangements (Andersson and Kurland 1998).

If two genomes have not gained a single gene since their divergence, searches for genes with atypical base composition features may be seriously flawed due to the absence of representative outliers with a significantly different pattern. Thus, intra-genomic comparison of nucleotide frequency statistics is a badly suited method for identifying horizontal gene transfer events (Koski et al. 2001), particularly in genomes with low frequencies of such events. We suggest that such inferences should be based on comparative analyses of base composition features of orthologous and unique gene sets in pairs of closely related genomes, as done for *E. coli* and *S. typhimurium* (Koski et al. 2001) as well as for *Chlamydia* in this case study.

Acknowledgments. We thank Otto Berg for many helpful discussions. This work was supported by grants from the Natural Sciences Research Council and the Foundation for Strategic Research via the National Graduate School in Scientific Computing.

References

Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. *Trends Microbiol* 6:263–268

- Andersson JO, Andersson SGE (1999a) Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol* 16:1178–1191
- Andersson JO, Andersson SGE (1999b) Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev* 9:664–671
- Andersson JO, Andersson SGE (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol Biol Evol* 18:829–839
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000) Genbank. *Nucleic Acids Res* 28:15–18
- Berg OG (1995) Kinetics of synonymous codon change for an amino acid of arbitrary degeneracy. *J Mol Evol* 41:345–352
- Blanchette M, Kunisawa T, Sankoff D (1996) Parametric genome rearrangements. *Gene* 172:GC11–17
- Blanchette M, Bourque G, Sankoff D (1997) Breakpoint phylogenies. *Genome Inform Ser Workshop Genome Inform* 8:25–34
- Blanchette M, Kunisawa T, Sankoff D (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J Mol Evol* 49:193–203
- Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 25:2124–2129
- Edward JD, Mishra SN (1988) *Modern mathematical statistics*. John Wiley, New York
- Garcia-Vallvé S, Palau J, Romeu A (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol Biol Evol* 16:1125–1134
- Garcia-Vallvé S, Romeu A, Palau P (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* 10:1719–1725
- Hooper SD, Berg OG (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. *J Mol Evol*, accepted for publication
- Johnson RA, Wichern DW (1992) *Applied multivariate statistical analysis*. Simon & Schuster, Englewood Cliffs, CA
- Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat Genet* 21:385–389
- Koski LB, Morton RA, Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* 18:404–412
- Kurland CG (2000) Something for everyone. *Horizontal gene transfer in evolution*. *EMBO Rep* 1:92–95
- Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44:383–397
- Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA* 95:9413–9417
- Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 36:201–206
- Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 18:299–304
- Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, Bass S, Linher K, Weidman J, Khouri H, Craven B, Bowman C, Dodson R, Gwinn M, Nelson W, DeBoy R, Kolonay J, McClarty G, Salzberg SL, Eisen J, Fraser CM (2000) Genome sequence of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res* 28:1397–1406
- Rocha EPC, Danchin A (2001) Ongoing evolution of strand composition in bacterial genomes. *Mol Biol Evol* 18:1789–1799
- Romero HA, Zavala A, Musto H (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res* 28:2084–2090

- Sankoff D, Blanchette M (1998) Phylogenetic invariants for metazoan mitochondrial genome evolution. *Genome Inform Ser Workshop Genome Inform* 9:22–31
- Sankoff D, Leduc G, Antoine N, Paquin B, Land B, Cedergren R (1992) Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc Natl Acad Sci USA* 89:6575–6579
- 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
- Sicheritz-Ponten T, Andersson SGE (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* 29:545–552
- Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 23:754–759
- Sueka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 48:582–592
- Tillier ERM, Collins RA (2000a) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* 50:249–257
- Tillier ERM, Collins RA (2000b) Genome rearrangement by replication-directed translocation. *Nat Genet* 26:195–197
- Tillier ERM, Collins RA (2000c) Replication orientation affects the rate and direction of bacterial gene evolution. *J Mol Evol* 51:459–463
- Waterson G, Ewens W, Hall T, Morgan A (1982) The chromosome inversion problem. *J Theor Biol* 99:1–7