Genome rearrangement by replication-directed translocation

Elisabeth R.M. Tillier & Richard A. Collins

Gene order in bacteria is poorly conserved during evolution[1–3]. For example, although many homologous genes are shared by the proteobacteria *Escherichia coli*, *Haemophilus influenzae* and *Helicobacter pylori*, their relative positions are very different in each genome, except local functional clusters such as operons[3–6]. The complete sequences of the more closely related bacterial genomes, such as pairs of *Chlamydia*[7–9], *H. pylori*[10,11] and *Mycobacterium* species[12], now allow identification of the processes and mechanisms involved in genome evolution. Here we provide evidence that a substantial proportion of rearrangements in gene order results from recombination sites that are determined by the positions of the replication forks. Our observations suggest that replication has a major role in directing genome evolution.

We determined plots for the relative positions of unique orthologous pairs of genes in pairwise comparisons of bacterial genomes (Fig. 1). Comparison of *H. pylori* 26696 with *Campylobacter jejuni*[13] (Fig. 1*a*) showed a high degree of rearrangement of the order of 748 identified orthologues. In contrast, a plot of the relative positions of 732 pairs of orthologues in *Chlamydia pneumonia* CWL029 and *Chlamydia trachomatis* serovar D showed that most (458) are located at the same relative position and orientation in both genomes, as indicated by their lying on a diagonal line with a slope of approximately 1 (Fig. 1*b*, filled squares). The 274 genes that are located in a different relative position comprise at least 22 distinct clusters, and are not randomly distributed throughout the genome. Instead, the genes that have moved form a perpendicular diagonal line with a slope of –1 (Fig. 1*b*, open circles) that intersects the first line at approximately the position of the termination of replication. A similar pattern has been noted independently in a comparison of *C. pneumonia* AR39 and *C. trachomatis* MoPn (ref. 9). This pattern indicates that almost all of the non-collinear genes have been inverted and translocated to the 'opposite side' of the genome: a mirror-image position across an axis defined by the putative origin and termination of replication (the 'replication axis'; Fig. 2).

We also found this pattern of gene translocation for 6 clusters containing 147 of 1,145 orthologues identified in the *H. pylori* strains 26695 and J99 (Fig. 1*c*). An analysis of homologous sequences in the *Mycobacterium tuberculosis* genome and the recently completed, but not yet annotated, *Mycobacterium leprae* sequence revealed that the gene order in the two *Mycobacterium* genomes is substantially more rearranged than that in the *Chlamydia* or *Helicobacter* genomes (Fig. 1*d*). Even with approximately 35% difference in genome length, however, the pattern of perpendicular diagonals is still apparent, indicating that many of the gene-order differences involve exchange across the replication axis.

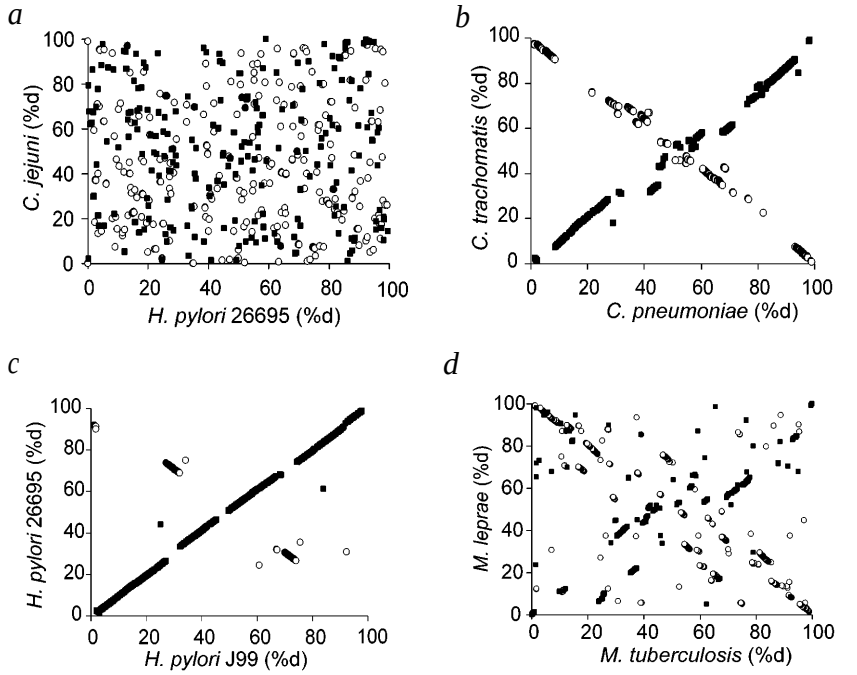

Fig. 1 Plots of the position of genes in related genomes. The *x* and *y* axes represent the chromosomes, linearized at the origin of replication and scaled to a common length. The positions of orthologous genes are plotted clockwise along the chromosome as a percentage of their distance from the origin (%d). Filled squares and open circles indicate genes that are in the same orientation in both genomes, or in inverted orientation in one relative to the other, respectively. **a**, *H. pylori* 26695 versus *C. jejuni*. **b**, *C. pneumoniae* versus *C. trachomatis*. **c**, *H. pylori* 26695 versus *H. pylori* J99. **d**, *M. tuberculosis* versus *M. leprae*.

At the scale of the whole genome, the translocation of genes across the replication axis appears to be quite symmetrical, forming almost straight lines (Fig. 1*b–d*). We found three instances in *Chlamydia* (Fig. 3*a*) and one in *Helicobacter* (data not shown) of a block of genes from one side of the genome exchanging with a block from the other side, consistent with two reciprocal recombination events, either simultaneous or consecutive.

A closer examination of the genome rearrangements revealed that only a few are this straightforward (Figs 2 and 3*a*). Many cases appear to result from a series of recombination events with slightly asymmetrical breakpoints. We diagrammed the arrangement of the regions of the genome flanked by *xerD* and *htrB*, and by *hflX* and *ycbF* (Fig. 3*b*). Although these two blocks of genes appear to have been reciprocally translocated (Fig. 3*a*), multiple rearrangements of gene order and loss/gain of genes have occurred within

Department of Molecular and Medical Genetics, University of Toronto, Toronto, Canada. Correspondence should be addressed to E.R.M.T. (e-mail: e.tillier@utoronto.ca).

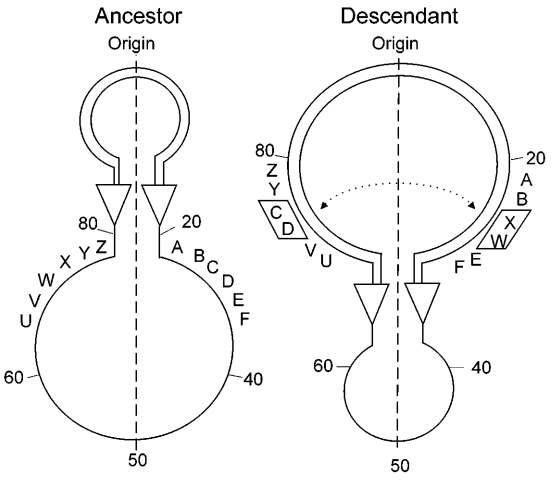

Fig. 2 Rearrangement of gene order by translocation of genes across the replication axis. A hypothetical ancestral gene order is indicated (left). After passage of the replication forks (triangles), genes C and D have exchanged positions with W and X by translocation across the replication axis (vertical dashed line) in the descendant genome. For simplicity, the diagram shows a reciprocal translocation that might occur in a single round of replication through two reciprocal recombination events. The diagram does not specify a mechanism for the translocation of genes, which may also occur in several steps as a series of recombination events in separate rounds of replication through intermediate genome organizations. The two replication forks are proposed to be across the replication axis and physically close together, promoting translocation of sequences at the forks. Numbers indicate the percentage of distance from the origin.

the blocks. For instance, genes 'F' and 'G' (Fig. 3*b*) constitute an example of local reordering that could be explained by consecutive translocations involving different recombination sites that are located approximately across the replication axis. This would involve a putative intermediate organization in which 'F' or 'G' had translocated independently of the other, followed by a second translocation back to the original side of the genome, but to a slightly different position.

For many genome rearrangements, we could not identify the recombination junctions with certainty due to an inversion or a loss or gain of one or more genes at the boundaries of the repositioned blocks of genes. Single-gene inversions occurred at 12 of 44 boundaries identified in *Chlamydia*, compared with only 13 local inversions adjacent to any of the 643 genes not at a boundary, indicating that inversions are frequently associated with translocation events ($P < <0.001$). We also found that 15 of the 44 translocation boundaries are flanked by a gene (or group of genes) occurring in one genome, but without a homologue in the other. Loss (or, less likely, gain) of a gene might occur by non-homologous recombination, causing truncation and inactivation of non-essential genes, or fusion of domains. An inversion or deletion of sequences adjacent to recombination sites might also be observed if the recombination was mediated by certain types of transposons[14]. The observation that loss or gain of genes is frequently observed adjacent to recombination sites ($P < <0.001$) is consistent with a role for illegitimate recombination.

In bacteria, homologous recombination between repeated sequences (such as rRNA genes) on opposite sides of the replication axis[3,15–18] results in the inversion of a large segment of the genome. Two consecutive inversions at different sites[15] would result in the apparent reciprocal translocation of genes between the recombination sites. The rearrangement breakpoints in the *Chlamydia* genomes do not correspond to homologous sequences (data not shown). Although sequence divergence might have obscured any homology that may have guided the recombination, our observations provide no indication for a role of homologous recombination in these genome rearrangements.

One interpretation of the observed pattern of genome rearrangements in *Chlamydia* is that many more rearrangement events actually occurred, but that natural selection has eliminated most of them[19–21]. Nonetheless, it appears that selection against genome rearrangement is not insurmountable, because long-range gene collinearity decays rapidly with divergence time, indicating that most genes can function adequately at many positions[1–5].

We propose an alternative explanation for the observed pattern of rearrangements: translocations occur preferentially across the replication axis as a result of the processes involved in genome replication. In genomes that replicate bi-directionally from a single origin, the two replication forks will be approximately equidistant from the origin (Fig. 2). Because the translocated sequences are also approximately equidistant from the origin, this suggests that the replication forks are important in genome rearrangement, and that the forks may be in close physical proximity to each other during replication. Single-stranded DNA or double-stranded breaks (such

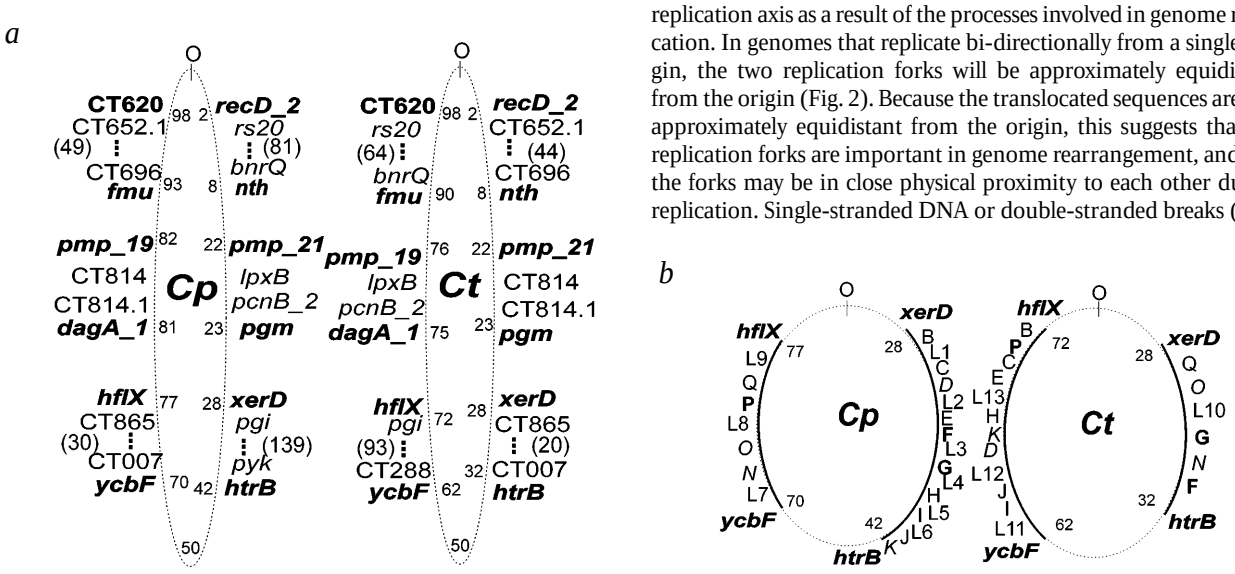

Fig. 3 Examples of translocated genes in *Chlamydia*. Genes indicated in bold are in the same relative positions in both genomes, but genes between them have been translocated (the number of genes not shown is given in parenthesis). Numbers inside the circles indicate the distance (% of the genome length) of the indicated gene from the origin (O). For clarity, where a gene in *C. pneumoniae* (*Cp*) and its orthologue in *C. trachomatis* (*Ct*) had been assigned a different name in each genome, we have used the *C. pneumoniae* name. The affected *C. trachomatis* genes (with the *C. pneumoniae* name in parentheses) are *pmpD* (*pmp21*), *mrsA* (*pgm*), CT409 (*dagA_1*), pmpA (*pmp_19*) and *pcnB_1* (*pcnB_2*). For hypothetical proteins, the name of the hypothetical gene product (CT′) is given because they are the same in both genomes. ***a***, These examples show reciprocal exchanges. One of them has complex rearrangements of the translocated genes shown in (***b***). Letters can indicate genes, or groups of genes. Bold letters indicate genes that are on the same side of the genome in both *Chlamydia* species. Italic letters indicate a group of genes in which a terminal gene was inverted (*D*, *K*, *O* and *N*). L1–L13 indicate genes or groups of genes for which no identifiable homologue was found in the other genome.

as those induced by topoisomerases at the replication forks) have been implicated in illegitimate recombination[9,22–25], and a close proximity of the forks would increase the probability of reciprocal recombination or transposition between sequences at the two forks. That the forks are near each other is also consistent with the 'replication factory' model based on immunolocalization of components of the replication machinery in *Bacillus subtilis*[26,27]. The plots of gene position in related genomes (Fig. 1) suggest that replication forks are preferred sites of DNA exchange, resulting in a preference for translocations to be generated across the replication axis.

This model may also explain more complex patterns of gene-order rearrangement. Any differences in the rates of replication at each fork would position different sequences near each other in different rounds of replication. Multiple rounds of replication with recombination occurring at slightly different positions in each round may lead to local re-ordering, such as described earlier for the genes 'F' and 'G' (Fig. 3*b*). A sufficient number of such quasi-symmetric recombination events might eventually lead to the more general situation among the proteobacterial genomes, where gene order has been almost completely randomized[1–5] (Fig. 1*a*).

Because few, if any, of the translocated genes in *Chlamydia* and *Helicobacter* have moved to a position off the perpendicular diagonal (Fig. 1), replication-directed rearrangements must be more common than other rearrangements that move a gene to a random position in the genome. Gene order appears to deteriorate non-randomly by consecutive, imperfectly symmetrical translocations across the replication axis. Irrespective of the exact processes involved, the observation that almost all translocations can be explained by a series of recombination events across the replication axis indicates that replication has a central role in targeting the position of a translocation. The replication-directed translocation process can explain not only genome-wide rearrangements, but also local rearrangements and possibly even gene gain and loss in the evolutionary divergence of genomes.

Methods

We used Fasta[28] analysis to identify homologues between the following genomes: *C. pneumoniae* CWL029 (ref. 7) versus *C. trachomatis* serovar D (ref. 8), *H. pylori* 26695 (ref. 10) versus *H. pylori* J99 (ref. 11), *H. pylori* 26695 versus *C. jejuni* NCTC11168 (ref. 13). Two coding regions were considered orthologous if they shared more than 30% amino acid identity and the Fasta calculated E() value was <0.01. To eliminate paralogues that might confuse the analysis, homologues were only considered if they were unique (that is, if no other gene meeting the above criteria was found that had at most half the z-score of the better match).

We used Tfasta[28] to identify positions in the translated sequence of the unannotated *M. leprae* genome (ftp://ftp.sanger.ac.uk/pub/pathogens/leprae/ML_assembly.dbs) homologous to annotated protein-coding genes in *M. tuberculosis*[11]. The positions of genes in *M. tuberculosis* that have a homologue in *M. leprae* with at least 50% protein sequence identity over the length of the query sequence are plotted in Fig. 1*d*.

The differences in base composition between the leading and lagging strands have been used to infer the location of the replication origin in bacterial genomes[29]. We inferred the location of the origin and termination of replication in the *Chlamydia* and *Helicobacter* genomes from the position of change in direction of the GC skew at the third positions of codons[30]. The location of the origin and total length of the genome was used to calculate the percentage distance from the origin for each gene (%d). The inferred locations of the origin and termination also correspond to the points that define the axis of symmetry regarding the translocations described in Fig. 1; therefore, we refer to this axis as the replication axis.

We estimated the statistical significance of the gene inversions and gene losses at translocation boundaries by χ^2 tests. For the tests, the probabilities of inversions and of deletions at any random position were estimated from the total number observed between the two genomes.

Accession numbers. *C. pneumoniae* CWL029, NC_000922; *C. trachomatis* serovar D, NC_000117; *H. pylori* 26695, NC_000915; *H. pylori* J99, NC_000921; *C. jejuni* NCTC11168, NC_002163; *M. tuberculosis*, NC_000962.

Acknowledgements

We thank W.F. Doolittle for discussion and B. Funnell for critical reading of the manuscript. R.A.C. is a fellow of the Canadian Institute for Advanced Research (CIAR). This work was funded by the National Sciences and Engineering Research Council (NSERC) grant to R.A.C.

Received 27 March; accepted 10 July 2000.

1. Mushegian, A.R. & Koonin, E.V. Gene order is not conserved in bacterial evolution. *Trends Genet.* **12**, 289–290 (1996).
2. Huynen, M.A. & Bork, P. Measuring genome evolution. *Proc. Natl Acad. Sci. USA* **95**, 5849–5856 (1998).
3. Casjens, S. The diverse and dynamic structures of bacterial genomes. *Annu. Rev. Genet.* **32**, 339–377 (1998).
4. Tatusov, R.L. *et al.* Metabolism and evolution of *Hemophilus influenzae* deduced from a whole genome comparison with *Escherichia coli*. *Current Biol.* **3**, 279–291 (1996).
5. Kolsko, A.B. Dynamic bacterial genome organization. *Mol. Microbiol.* **24**, 241–248 (1997).
6. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73 (1997).
7. Kalman, S. *et al.* Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genet.* **21**, 385–389 (1999).
8. Stephens, R.S. *et al.* Genome sequence of an obligate intracellular pathogen of humans, *Chlamydia trachomatis*. *Science* **282**, 754–759 (1998).
9. Read, T.D. *et al.* Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic Acids Res.* **28**, 1397–1406 (2000).
10. Tomb, J.F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
11. Alm, R.A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
12. Cole, S.T. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
13. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuna* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
14. Shapiro, J.A. Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proc. Natl Acad. Sci. USA* **76**, 1933–1937 (1979).
15. Liu, S.-H. & Sanderson, K.E. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl Acad. Sci. USA* **92**, 1018–1022 (1995).
16. Shmid, M.B. & Roth, J.R. Selection and endpoint distribution of bacterial inversion mutations. *Genetics* **105**, 539–557 (1983).
17. Rebollo, J.-E., François, V. & Louar, J.-M. Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome. *Proc. Natl Acad. Sci. USA* **85**, 9391–9395 (1988).
18. Itaya, M. Physical map of the *Bacillus subtilis* 166 genome: evidence for the inversion of an approximately 1900 kb continuous DNA segment, the translocation of an approximately 100kb segment and the duplication of a 5kb segment. *Microbiology* **143**, 3723–3732 (1997).
19. Caro, L.G. & Berg, C.M. Chromosome replication in some strains of *Escherichia coli* K12. *Cold Spring Harb. Symp. Quant. Biol.* **33**, 559–573 (1968).
20. Schmid, M.B & Roth, J.R. Gene location affects expression level in *Salmonella typhimurium*. *J. Bacteriol.* **169**, 2872–2875 (1987).
21. Brewer, B.J. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**, 679–686 (1988).
22. Ikeda H., Moriya, K. & Matsumoto, T. *In vitro* study of illegitimate recombination: involvement of DNA gyrase. *Cold Spring Harb. Symp. Quant. Biol.* **45**, 399–408 (1980).
23. Michel, B., Ehrlich, S.D. & Uzest, M. DNA double-strand breaks caused by replication arrest. *EMBO J.* **16**, 430–438 (1997).
24. Bierne H., Ehrlich, S.D. & Michel, B. Deletions at stalled replication forks occur by two different pathways. *EMBO J.* **16**, 3332–3340 (1997).
25. Kuzminov, A. & Stahl, F.W. Double-strand end repair via the RecBC pathway in *Escherichia coli* primes DNA replication. *Genes Dev.* **13**, 345–356 (1999).
26. Newport, J. & Yan, H. Organization of DNA into foci during replication. *Curr. Opin. Cell Biol.* **8**, 365–368 (1996).
27. Lemon, K.P. & Grossman, A.D. Localization of bacterial DNA polymerase: evidence for a factory model of replication. *Science* **28**, 1516–1519 (1998).
28. Pearson, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63 (1990).
29. Lobry, J.R. Origin of replication of *Mycoplasma genitalium*. *Science* **272**, 745–746 (1996).
30. Tillier, E.R.M. & Collins, R.A. The contributions of replication orientation, gene direction and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50**, 249–257 (2000).