Directional Evolution of Chlamydia trachomatis Towards Niche-specific Adaptation

Running title: Chlamydia trachomatis adaptive evolution

Vítor Borges a, Alexandra Nunes a, Rita Ferreira a, Maria J. Borrego a, and João P. Gomes a,*

a Department of Infectious Diseases, National Institute of Health, Av. Padre Cruz, 1649-016 - Lisbon, Portugal

* Corresponding author: J.P. Gomes

Address: Department of Infectious Diseases, National Institute of Health, Av. Padre Cruz, 1649-016 - Lisbon, Portugal; Tel.: (+315) 217 519 241; fax: (+351) 217 526 400; E-mail address: j.paulo.gomes@insa.min-saude.pt
ABSTRACT

On behalf of host–pathogen arms race, a cutting-edge approach for elucidating genotype-phenotype relationships relies on the identification of positively selected loci involved in the pathoadaptation. We studied the obligate intracellular bacterium Chlamydia trachomatis, for which same-species strains display a nearly identical core and pan genome, while presenting a wide range of tissue tropism and ecological success. We aimed to evaluate the evolutionary patterns underlying species separation (divergence) and C. trachomatis serovar radiation (polymorphism), and to establish genotype/phenotype associations. By analyzing 60 Chlamydia strains, we detected traces of Muller’s ratchet as a result of speciation, and identified positively selected genes and codons hypothetically involved in infection of different human cell types: columnar epithelial cells of ocular or genital mucosae, and mononuclear phagocytes; and also, events likely driving pathogenic and ecological success dissimilarities. In general, these genes code for proteins involved in immune response elicitation, proteolysis, subversion of host-cell functions, and also proteins with unknown function. Several genes are potentially involved in more than one adaptive process, suggesting multiple functions or a distinct modus operandi for a specific function, and thus should be considered as crucial research targets. Additionally, six out of the nine genes encoding the putative antigens/adhesins polymorphic membrane proteins seem to be under positive selection along specific serovars, which sustains an essential biological role of this extra-large paralogues family in chlamydial pathobiology. This study provides insight into how evolutionary inferences illuminate ecological processes such as adaptation to different niches, pathogenicity, or ecological success driven by arms races.
INTRODUCTION

Genomic changes of microbial pathogens are directly linked to the evolutionary arms race that takes place between microbe and host during the infectious process, as a result of the antagonistic interaction, and they are a consequence of polymorphisms accumulated after selective pressure from the host’s inflammatory or immune response (32). However, the majority of coding genes present a higher number of synonymous rather than non-synonymous substitutions, which indicates that purifying selection is operating to preserve the current function and structure of the protein, and only a small fraction of the genes are expected to be positively selected where diversification is favored through increased fitness (11). In order to understand the evolutionary forces that act on gene variation, major challenges are to identify loci that might have been under selection, and to determine the type of natural selection that has influenced their evolutionary history (59). In the field of infectious diseases, site-specific inferences regarding positive selection on loci involved in drug resistance (19) or in the interaction with the host immune system have been proposed as complementary approaches for the development of vaccines against HIV and other viruses (33), and also to predict the evolution of virulent strains of the influenza virus (12). Also, it has been shown that core genes are equally subjected to positive selection as pathogen specific accessory genes (4), suggesting that blind genomic-scale analysis should be performed.

For a species such as Chlamydia trachomatis with a wide range of tissue tropism and ecological success, but presenting a nearly identical core and pan genome, and a DNA sequence similarity of > 98% (39), the few existing polymorphisms are expected to be extremely informative of the adaptive evolution process. However, an excess of nonsynonymous substitutions alone is not sufficient to invoke positive selection, as it requires an increase in fitness caused by the corresponding amino-acid replacement. Otherwise, it may represent the accumulation of slightly deleterious mutations (not severe
deleterious as these will not become fixed because they render their bearers non-viable) to the pathogen on behalf of the Muller’s ratchet theory (27, 60). This is predicted to operate in intracellular replicating bacteria (as *C. trachomatis*, which replicates within a host vacuole named inclusion) that are subject to recurrent bottlenecks and replicate in small populations, with little opportunity for recombination and few back or compensatory mutations (2). Although it was recently shown (39) that recombination events affect much more chromosome regions than previous suspected in *C. trachomatis*, the frequency and the relative weight of recombination and mutation calculated for this pathogen (ρ/θ < 0.07 and r/m < 0.71, respectively) (28, 49) indicates the point mutation events as the major evolutionary driving force.

In the present study, we used comparative genomics over 59 *C. trachomatis* strains (comprising all serovars) to clarify the mutational dynamics underlying both the separation of *C. trachomatis* as species, and the pathoadaptation driven by arms race. We identified positively selected genes and codons that are hypothetically involved in the evolutionary adaptation of *C. trachomatis* serovars to different cell types: mucosal cells from the eye conjunctiva (responsible for trachoma) (serovars A-C), from the genitalia (primarily yielding cervicitis) (serovars D-K), and mononuclear phagocytes (yielding invasive diseases such as hemorrhagic proctitis and suppurative lymphadenitis) (serovars L1-L3). Finally, we also detected positive selection events likely driving pathogenic and ecological success dissimilarities.

**MATERIAL AND METHODS**

*C. trachomatis* strains, cell culture and DNA extraction. The present study encompasses data from 59 *C. trachomatis* strains and the *Chlamydia muridarum* Nigg strain (also called, Mouse Pneumonitis strain – MoPn) (listed in the Table S1). These include in silico data from recently
analyzed fully-sequenced *C. trachomatis* strains (39) and eight historical prototype strains (Ba/Apache-2, C/TW3, F/IC-Cal3, G/UW57, H/UW43, I/UW12, J/UW36 and K/UW31) in order to enroll all the 15 major serovars. Those additional eight strains were propagated in HeLa 229 cell monolayers and at 48 to 72 hours post-infection, cells were harvested and a bacteria-enriched pellet was obtained and resuspended in 200 µl of PBS, as previously described (9). DNA was extracted using the QIAamp® DNA Mini Kit (Qiagen) according to manufacturer's instructions, and stored at -80 ºC until use.

**Selection of loci and sequencing.** Based on available *in silico* full-genome sequence data, we searched for polymorphic genes among *C. trachomatis* strains through the progressiveMauve algorithm (23) of the Mauve software v2.3.1. A detailed evaluation of polymorphism of each locus was further performed by using Lasergene® 9.0. (DNASTAR, Madison, Wisconsin, USA) and MEGA5 (76). Chromosome loci revealing an extremely low polymorphism were discarded from the present analysis as their use would hamper the accurate application of likelihood tests. We ended up with 75 top-ranked polymorphic genes (listed in the Table S2). These were categorized according to their functional role, involving 20 housekeeping genes (HKs), 14 genes encoding well-known cell envelope proteins (CEPs), 31 genes coding for secreted proteins (SECs), and 10 genes coding for proteins with unknown function or for which the biological role is not consensual. The SEC category involves proteins secreted [either by the Type III Secretion System (T3SS) - a machinery used by many bacterial pathogens to manipulate eukaryotic host cells by injecting virulence proteins - or by an undefined mechanism], into the cytosol of the host cells or to the inclusion membrane. For analyses enrolling divergence *versus* polymorphism, the corresponding orthologous genes of the *Chlamydia muridarum* Nigg strain were identified (by NCBI-BLAST search) and sequences were collected from the full-genome annotated in the GenBank database (accession number NC_002620) (65). For the strains that we needed to propagate as no *in silico* data was available, the 75 genes were amplified and sequenced by using
standard procedures (36). The sequences and location of primers, as well as the amplicon sizes are listed in the Table S3. Automated sequencing was achieved using BigDye Terminator v1.1 Cycle Sequencing chemistry, according to the manufacturer’s instructions (Applied Biosystems) in an Applied Biosystems 3130xl Genetic Analyzer. Sequence reads were assembled using SeqBuilder software (DNASTAR) and alignments were generated using the ClustalW algorithm implemented in both the MegAlign software (DNASTAR) and MEGA5. A concatenated alignment of the 75 genes was also constructed for all C. trachomatis and C. muridarum strains. As the ClustalW program generates alignment artifacts in the presence of insertion/deletion (indel) events by disrupting codons, we edited “by hand” the amino acid alignments rather than only automate the process before editing the corresponding nucleotide sequences. When strain-exclusive single nucleotide polymorphisms (SNPs), indel events and pseudogenes were identified, resequencing was performed from a newly extracted DNA, and new sequences reads were generated for comparative purposes.

**Phylogenomic analysis.** Analyses of genetic diversity and phylogeny were conducted for each gene by using MEGA5. Briefly, we computed overall mean distances (number of differences and p-distance) and matrices of pairwise comparisons at both nucleotide and amino acid levels. For phylogenetic analysis, individual trees were generated using the Neighbor-Joining method with bootstrapping (67) and the evolutionary distances were computed using the Kimura 2-parameter method (52). For all these analyses, the pairwise-deletion option was selected as it excludes sites containing alignment gaps or missing data from the analysis only when necessary in the pairwise distance estimation. Truncated genes, which are expected to encode non-functional proteins, were excluded from the phylogenetic and evolutionary analyses, except for the strains with non-disrupted sequences, as their biological role may be phenotype specific.
Global analysis of molecular evolution. The nonsynonymous/synonymous substitution rate ratio (d_N/d_S) among related protein-coding DNA sequences, where d_N refers to the number of nonsynonymous substitutions per nonsynonymous site and d_S is defined as the number of synonymous substitutions per synonymous site, may be suggestive of the selective pressures driving the mutational trends (86). Initially, for a global analysis of these trends, we estimated d_N and d_S values with MEGA5 by using the Kumar model (61). For each gene, d_N/d_S was calculated over all C. trachomatis sequence pairs and between the sequences of the two species (C. trachomatis and C. muridarum). More, in order to reinforce the comparison between the amount of evolutionary variation within the C. trachomatis species (polymorphism) and the variation between C. trachomatis and C. muridarum (divergence), we also applied the McDonald-Kreitman (MK) test (26, 55). However, as it has been assumed that the results from the MK test cannot directly discriminate the type of selection acting on genes (62), the subjacent MK test algorithm was only used to clarify the neutral and amino acid-altering mutational trends underlying the C. trachomatis speciation process. This kind of analysis is suitable for tracing the Muller’s ratchet phenomenon, which is commonly observed in niche-restricted pathogens.

Evaluation of the directionality in C. trachomatis evolution. In order to search for genes on which positive selection putatively operates, two distinct approaches were applied. First, as a statistical support of the d_N and d_S estimations within C. trachomatis strains, the codon-based Z-test of selection was computed by MEGA5 using the Kumar method (61), where bootstrapping (1000 replicates) was used for estimation of the variation in the statistic test. This test calculates the probability of rejecting the null hypothesis of strict-neutrality (d_N = d_S) in favor of one of two alternative hypothesis: positive selection (d_N > d_S) or purifying selection (d_N < d_S). Results with p-values less than 0.05 were considered significant at the 5% level. On a second approach, the branch-site test of positive selection (branch-site test 2) (85, 88) was employed using the codeml application from the Phylogenetic Analysis
Alignments of nucleotide sequences from the 59 *C. trachomatis* strains and *C. muridarum* (built and corrected on MEGA5) were converted into the “interleaved” PHYLIP format using the BioEdit package (version 7.0.0) (http://www.mbio.ncsu.edu/bioedit/bioedit.html), where stop codons were removed from sequences.

The branch-site test is a robust bioinformatic approach (84) that is recommended to infer positive selection in a lineage of interest (called foreground lineage) when several lineages in the phylogeny may have been subjected to distinct selective pressures (85, 88). The statistical significance of the presence of positive selection along the branch of interest was addressed by the likelihood-ratio test (LRT) (82). In the branch-site test 2, the LRT compares the twice of the log likelihood difference ($2\Delta l$) between two models (alternative and null models) with the chi-square distribution with one degree of freedom for p-value calculation (88). The alternative model allows positive selection ($d_N/d_S \geq 1$) for the foreground branch, whereas the null model assumes the $d_N/d_S$ ratios < 1 or = 1 for all site classes in all branches in the phylogeny. When positive selection acting on a specific gene was suggested by a significant LRT (p-value less than 0.05), the Bayes empirical Bayes analysis (87) was used to identify the specific positively selected sites within that gene along the foreground branches. Therefore, the branch-site model requires an *a priori* definition and labeling of the foreground branches to be tested for positive selection, which should rely on well-defined biological hypotheses (85). Thus, based on the assumption that some genes might be involved in *C. trachomatis* phenotypic dissimilarities as a result of targeted positive selective pressures, we created six comprehensible biological hypotheses (H1-H6). The hypotheses evaluate the existence of genes under positive selection that may be involved in the following biological processes: specific cell-appetence to columnar epithelial cells of ocular (H1) or genital mucosae (H2), and to mononuclear phagocytes (H3); pathogenic diversity among strains causing ocular disease (H4), genital disease (H5), or hemorrhagic proctitis and suppurative
lymphadenitis (H6). Only the genes for which the phylogeny supported any of these scientific hypotheses were tested.

Finally, as recombination may bias the results of positive selection, we used published data on recombination analysis enrolling all *C. trachomatis* genes (39, 41, 49) to inspect whether the genes selected for the present study showed evidences of recombination. Consequently, for the genes showing incongruent trees where unequivocal recombination was detected within a specific branch, the analysis of positive selection was excluded *a priori* for the corresponding biological hypothesis. On the other hand, genes yielding congruent trees but for which recombination had been previously detected (39, 41, 49) were still subjected to positive selection analysis and are properly identified in the present study.

**Nucleotide sequences accession number.** The nucleotide sequences determined in the present study were submitted to the GenBank database ([http://www.ncbi.nlm.nih.gov/Genbank/index.html](http://www.ncbi.nlm.nih.gov/Genbank/index.html)) and are currently available for consulting through the accession numbers: JQ066324 - JQ066722.

**RESULTS**

**Polymorphism significance of the selected genes.** Distribution of point mutations in *C. trachomatis* chromosome is highly heterogeneous. Although the selected genes (Fig. 1 and Table S2) represent 11% of the coding region length, they encompass about 55% of all chromosomal SNPs occurring within coding regions, which corresponds to a total of 5083 polymorphic sites among the 59 strains. In fact, we found that any given chromosomal SNP has 10.0 (odds ratio, 95% CI: 9.3 - 10.7) times higher probability to belong to the pool of genes under evaluation than to show up in any other gene (Fisher's exact test, *p* < 10^{-7}).
Divergence versus polymorphism – detection of Muller’s ratchet. Considering that \textit{C. trachomatis} and \textit{C. muridarum} species evolved from a last common ancestor (65), the comparison between the amount of evolutionary variation within the \textit{C. trachomatis} species (polymorphism) and between \textit{C. trachomatis} and \textit{C. muridarum} (divergence) may shed some light on the evolutionary mutational dynamics that drove the \textit{C. trachomatis} speciation. Accordingly, the divergence of the two species was evaluated through estimation of the \(d_N/d_S\) ratio between orthologous genes. All genes revealed \(d_N/d_S\) values lower than one, where the mean value was 0.21 (standard deviation (SD), 0.14) (Fig. 2). This observation suggests an unequivocal higher weight of synonymous than nonsynonymous changes on the species divergence, in agreement with the neutral theory of molecular evolution (53), which postulates that the fixation of selectively neutral mutations by random genetic drift is the major factor responsible for species divergence.

Subsequently, we compared these results with the \(d_N/d_S\) values obtained solely within \textit{C. trachomatis} species (i.e., among the 59 strains). We observed that there was a high, but dissimilar, decrease of both \(d_S\) and \(d_N\) values after the separation of the two species from a common ancestor (Fig. 2). In fact, the observed mean of the decay of \(d_S\) values \([147.2 \text{ (SD} \pm 243.7)]\) was 5.5 times higher than the one observed for \(d_N\) \([26.9 \text{ (SD} \pm 19.9)]\). This was a consistent trend as the McDonald-Kreitman test algorithm yielded a similar decays ratio of 4.7. Globally, this suggests that, since \textit{C. trachomatis} was established as species, the nonsynonymous changes increased their relative weight to synonymous changes in contrast with the evolutionary process that originated the separation of the two species. This observation is consistent with the Muller’s ratchet theory (27, 60) which assumes an accumulation of slightly deleterious non-silent mutations on microbial populations repeatedly subjected to genetic bottlenecks. In fact, the obligate intracellular lifestyle of \textit{C. trachomatis}, which is characterized by niche-restricted and low-size populations, and expected low frequency of recombination relative to
mutation events (28, 49), may lead to less effective elimination and consequent accumulation of non-
silent mutations (2).

**Evolutionary trends within the *C. trachomatis* species.** We further focused on studying the
accumulation of mutations after *C. trachomatis* speciation, with special emphasis on protein-altering
changes that have contributed for phenotype divergences, which may help to clarify the *C. trachomatis*
niche-specific adaptation. For each gene, the $d_\text{N}/d_\text{S}$ ratio was evaluated over all *C. trachomatis*
sequence pairs (Fig. 3), where genes exhibiting $d_\text{N}/d_\text{S}$ ratios above 1 and a significant p-value ($<0.05$) in
the codon-based Z-test of positive selection ($d_\text{N}/d_\text{S} > 1$) were considered as putative targets of positive
selection. Twenty-seven genes exhibited overall $d_\text{N}/d_\text{S}$ values higher than one, in which 15 (including
14 SECs) revealed a significant Z-test p-value. As predicted, all housekeeping genes presented $d_\text{N}/d_\text{S}$
ratios below one, which indicates that the genes involved in regulatory/metabolic functions are less
likely targeted by diversifying selection (47). On the other hand, 22 out of the 31 SECs support an
opposite scenario, which is relevant as these proteins contact directly with the host, and thus are more
prone to be involved in pathoadaptation. We also investigated if the types of mutation are dependent on
the degree of genetic variability by evaluating the relationship between $d_\text{N}$ and $d_\text{S}$ values and the
nucleotide polymorphism (p-distance) among *C. trachomatis* strains (Fig. 4). Globally, as observed
above for species divergence, we found that the increment in polymorphism is essentially driven by
fixation of silent mutations, which presented an increase rate about four-fold higher than non-silent
changes.

**Distribution of $d_\text{N}$ and $d_\text{S}$ versus disease outcomes and ecological success.** We also aimed to
understand if the general distribution of both silent and non-silent mutations among *C. trachomatis*
corresponds to strains clustering by disease outcomes. We used the concatenated sequences
encompassing all genes under evaluation to calculate the $d_\text{N}$ and $d_\text{S}$ distances between each strain and
the different groups of strains (i.e., three disease groups) (Fig. 5). Our results sustain a non-random accumulation of mutations where strains with the same cell-appetence are unequivocally clustered either by silent mutations or protein-changing alterations. Indeed, the genetic distances between strains with dissimilar tropism are 1.8- to 10.5-fold (for \( d_S \)) and 1.7- to 6.7-fold (for \( d_N \)) higher than the ones between strains with similar cell-appetence. Additionally, the highly ecological succeeded strains causing non-invasive genital infections (mostly from serovars E and F) are slightly separated from the remainder genital strains (Fig. 5). This analysis clearly supports that our approach for detecting positively selected genes (and codons) relying on rationally established biological hypotheses (see methods) may be an useful step for understanding the molecular basis underlying \( C. trachomatis \) phenotypic differences.

**Positive selection driving bacterial specific appetite to different human cell types.** The phylogenetic analysis revealed genes whose trees cluster all strains that preferentially infect the same human cell type in a single branch. Thus, genes (and codons) targeted by positive selection along those branches may be involved on specific host-cell interactions. To evaluate this, we conducted the branch-site test of positive selection under the biological hypotheses H1 to H3 (see methods). All genes and the inferred positively selected codons found to be putatively involved in specific adaptive processes are described in Fig. 6 and Table 1. Five genes were found to be under positive selection in the evolutionary process that drove the segregation of ocular strains (branch H1). These include genes encoding two polymorphic membrane proteins (Pmps) (CT869/pmpE and CT870/pmpF), one Pmp-like protein (CT050), one inclusion membrane protein (Inc) (CT115/incD) and the translocated actin recruiting phosphoprotein (CT456/tarp). The Pmps and Incs are among the most promising research targets for which there is cumulative evidence of their involvement in biological mechanisms such as adhesion, immune response elicitation, or subversion of intracellular trafficking (see Table 1 for
details) (21, 25, 35, 38, 66, 68, 78). For example, PmpF was predicted in silico to contain T-cell epitopes that bind HLA class I and II alleles (15). On the other hand, Tarp is a chlamydial effector of the T3SS associated with the chlamydial invasion of the host cells (44) by mediating host actin polymerization and inclusion development (20, 42).

For the biological hypothesis H2, we found significant evidence supporting a fixation of adaptive mutations driving a better appetcence to the columnar epithelial cells of genital mucosa (branch H2) solely for the gene CT105. Although its function is unknown, a previous study of heterologous expression in yeast (71) suggested that CT105 may be involved in modulation of host cellular functions. Nevertheless, it is worth to note that CT105 is a pseudogene for ocular strains, thus we may be facing a scenario of a gene strictly needed for tropism functions other than those involving the ocular conjunctiva. The frequent tree incongruence involving genital strains hampered the evaluation of several genes for this specific biological hypothesis (Table 1).

Regarding the branch that clusters all strains infecting the mononuclear phagocytes (branch H3), we detected 18 genes likely under positive selection. This set includes one HK, and genes encoding 11 SECs (seven Incs), four CEPs (including three Pmps), and two proteins with unknown function (Table 1). Besides the general relevance of Incs and Pmps (explained above), we would highlight the SEC CT223, an Inc protein for which it was suggested a role in subversion of host cell functions, either by containing SNARE-like (eukaryotic soluble N-ethylmaleimide-sensitive attachment protein receptors) motifs (impact in intracellular trafficking) (25), or by blocking host cell cytokinesis (1). The SECs CT622 is an antigen putatively secreted by the T3SS (21, 30, 37), whereas CT867 and CT868 are proteases that possess deubiquitinating and dendeddylating activities (57), which may suggest a role in virulence. The invasive infection pattern of L1-L3 strains along with the expectation that these strains were the first to diverge from a common C. trachomatis ancestor (73) may justify the
high number of genes detected under diversifying selection on branch H3 rather than on branches H1 and H2. These genes may play a role in the specificity of mononuclear phagocytes-bacteria interactions that yield invasive infections with L1-L3 strains.

**Positive selection driving pathogenic diversity among strains infecting the same human cell type.** The detection of positive selection acting on specific genes along branches of strains causing similar disease outcomes may be useful for understanding adaptive alterations underlying niche-specific pathogenic dissimilarities. Among the ocular strains, we detected two genes under positive selection (branches H4) (Fig. 6): one pmp (CT413/pmpB) and CT456/tarp (Table 1). CT456/tarp had been already indicated as potentially involved in distinct pathogenic patterns displayed by two ocular strains (both of serovar A) when infecting cynomolgus monkeys (51).

Regarding the detection of positive selection driving pathogenic differences among strains causing non-invasive urogenital disease (branches H5), we found the above described virulence factors CT223 and CT456/tarp, and one pmp (CT872/pmpH). Once more, the analysis of positive selection underlying pathogenicity among non-invasive genital strains was impaired by the described recombination events involving these strains (Table 1) (39, 41, 49). Of notice was the detection of positive selection events governing the evolutionary segregation of the most succeeded genital strains (mostly from E and F serovars) for the gene coding CT694. This protein is an immunodominant antigen (69), and it was also demonstrated its secretion into the host cytoplasm by the T3SS at early time-points after infection (as CT456/Tarp), where it localizes to host cell membranes and interacts with eukaryotic AHNAK, an actin-binding protein; it is believed that CT694 may act by regulating membrane fluidity or by remodeling actin filaments during invasion or early stages of *C. trachomatis* development (10, 40). Thus, differences in immune evasion strategies or in host-cell manipulation during invasion may be crucial biological processes underlying ecological success.
Despite the remarkable genomic homogeneity of *C. trachomatis* strains that infect the mononuclear phagocytes (Fig. 5), we found 11 positively selected genes along the branches embodying SNPs that distinguish these strains (branches H6). These include four *pmps* (CT413/pmpB, CT414/pmpC, CT871/pmpG and CT872/pmpH), two genes encoding Pmp-like proteins (CT050 and CT051), three genes coding for Incs (CT147, CT233/incC and CT442/crpA), CT456/tarp and CT868. Besides what was generally described above for these proteins, it is worth to note that the Inc CrpA (6) is a T3SS substrate (75) that may play a role in immune evasion as it was found to be targeted by CD8+ T cells in response to infection in murine (72). Also notable is that all adaptive codons inferred for the CT233/incC (Table 1) correspond to the IncC N-terminal domain, and, more specifically within the first 15 residues, where it is known to reside the secretion signals recognized by T3SSs (75).

**Evolutionary inferences and associated bias.** The branch-site test may generate some erroneous detected positively selected genes. Although this was not assessed in the present study, recent robust evaluations estimated a range of 0-5% of false positives on this test (84). Nevertheless, this is more problematic when performing inter-species analyses as it involves highly divergent sequences (5), which is not the case of the intra-species analysis performed in the present study. Furthermore, to guide against violations of model assumptions, we applied very conservative criteria to calculate p-values in the LRT by using $\chi^2_1$ as the null distribution (5, 88). Recombination is another critical factor that may bias the estimation of positive selection. As previous data based on full-genome sequences (39, 41, 49) detected recombination for some of the genes enrolled in the present study, some specific biological hypothesis could not be subjected to the branch-site test of positive selection. These specific exclusions are indicated in the Table 1. The remainder biological hypotheses were tested as recombination is not observed in the corresponding branches. For example, for CT147 tree, where some ocular strains are shown within a genital branch (hampering the analysis of the hypotheses H1,
H2, H4, and H5), the hypotheses H3 and H6 could still be validated. In another scenario, when recombination is known to occur in genes presenting strong congruent trees (as for CT870/pmpF), all hypothesis were evaluated. For these specific cases (indicated in Fig. 6) the results should be eyed with caution.

DISCUSSION

A well-known metaphor in evolutionary biology is the adaptive landscape represented by a two-dimensional plot of all genotypes in a specific environment, with their fitness represented by the height of the landscape. For each new environment, in order to climb the fitness peak, bacteria will have to acquire new beneficial mutations, which will likely be differentially spread among different genotypes (80). Presumably, the radiation of *C. trachomatis* species into strains with different cell-appetence may be explained by this scenario. Indeed, the different environments are represented by the dissimilar human tissues that strains preferentially infect (ocular, genital tract and lymph nodes), which present heterogeneous properties in terms of competing flora, immune response, and physiological characteristics (such as pH and hormonal concentration). Also, strains present dissimilar fitness as serovar E and F together represent more than 40% of all genital infections worldwide (64), and serovars A and L2 clearly predominate in ocular (3) and lymphogranuloma (79) infections, respectively. On the other hand, *C. trachomatis* is an obligate intracellular bacterium with low doubling time (9) and population size (34), and is thus subjected to transmission bottlenecks which make this pathogen a target for the accumulation of deleterious mutations on behalf of the Muller’s ratchet theory. The validity of the Muller’s ratchet has been evaluated either in RNA viruses (18), which present high mutation rates, are subjected to recurrent bottlenecks and the rate of compensatory back-mutations is low, or even in large free-living bacteria such as *Salmonella typhimurium*, where these
contributing factors are clearly attenuated (2). In our study when the values of \( d_N \) and \( d_S \) are compared independently, it is noticeable that, after the \( C. \) muridarum/\( C. \) trachomatis separation from a common ancestor, the values of \( d_S \) show a 147-fold decrease whereas the values of \( d_N \) only decrease 27-times (Fig. 2), which evidences the existence of an accumulation of deleterious mutations due to genetic bottleneck, as postulate by Muller. We speculate that, besides this scenario, some non-silent changes may reflect adaptive mutations to different niches rather than deleterious mutations specific of same-niche infecting strains. Thus, the detection of positive selection events acting on particular genomic regions may help to elucidate genotype-phenotype relationships. Unfortunately, there are scarce cases where genotype and phenotype are unequivocally linked in \( C. \) trachomatis, because no straightforward tools to genetically manipulate this pathogen are available so far. One of the few examples is illustrated by the mutational pattern of the \( trpBA \) operon (encodes tryptophan synthase, which uses indole as substrate), where genital strains possess an intact and active operon whereas it is truncated by point mutations or small indels in strains infecting the ocular conjunctiva (where indole is rare) (13). Thus, mutations that are beneficial in one genetic background are not necessarily beneficial in another background. In our study, a similar scenario may stand for CT105, which is a pseudogene for the ocular strains whereas our results suggest that it may be involved in the strains’ appentence to the genital epithelium (Fig. 6 and Table 1).

Our results showed that non-silent changes differentiate strains with different cell-appetence or pathogenesis (Fig. 5) and involve genes whose functions may underlie distinct phenotypes. In fact, among the 25 genes identified as positively selected along specific lineages (Table 1), we found genes encoding proteins implicated in immune response elicitation (such as CT147, CT442/crpA, CT529, CT694, and \( pmps \)) (21, 30, 31, 69, 72, 78), proteolytic activity (such as CT867 and CT868) (57), and subversion of host-cell functions (such as CT223 and CT456/tarp) (25, 44). Some of these genes were...
also identified in a previous study (49), but no genotype/phenotype associations could be established because only six serovars were evaluated, contrarily to the present study which constitutes a considerable scale-up in terms of genetic variability (enrolling all major 15 serovars represented by 59 strains). A detailed view of the positively selected loci that we have detected revealed 11 genes (CT050, CT051, CT115/incD, CT147, CT223, CT413/pmpB, CT456/tarp, CT868, CT870/pmpF, CT871/pmpG and CT872/pmpH) supporting two or more biological hypotheses for adaptive changes (Fig. 6). Although this seems intriguing in terms of evolutionary directionality, experimental evidences suggest multiple functions for some of them (Table 1) or a distinct modus operandi for a specific function. The most striking example is illustrated by CT456 which codes for Tarp. Strong experimental evidences showed that this T3SS effector is associated with the recruitment of host-cell actin observed at early stages of invasion, involving a C-terminal actin binding domain (ABD) and a proline-rich region (43). Whereas the invasive serovar L2 contains a single functional ABD and it is believed that the proline rich domain plays also a role in actin nucleation, Tarp from strains with different cell-appetence contain multiple ABD sites that are able to nucleate actin without the need of the respective proline-rich domain (44). These data suggest that strains may use Tarp distinctly for actin nucleation. Also, Tarp harbors an N-terminal tyrosine-rich repeat domain (the number of repeats are serovar-dependent) that is tyrosine phosphorylated by host cell kinases (42). Curiously, some positively selected sites found to be associated with infection of mononuclear phagocytes (Table 1) are located precisely within the tyrosine-rich repeat domain. Moreover, there seems to be a pattern of amino acid substitution, where positive selection is operating on exactly the same amino acid positions within the repeated regions, involving always the exchange between aspartic acid (D) and glycine (G) for seven of the positively selected sites (Table 1). In support of recently published data by Mehlitz and colleagues (56), it can be speculated that, as for the Tarp C-terminal region, dissimilar modus operandi of Tarp N-
terminal region may underlie distinct phenotype properties of *C. trachomatis* strains. Also of relevance is CT868, a deubiquitinating and de neddylating enzyme that likely interferes in multiple cellular processes. Indeed, a recent study demonstrated that CT868 is capable of inhibiting the host inflammatory responses by blocking the nuclear factor-kB pathway, a known mechanism by which pathogenic microorganisms evade the host immune responses (54). Facing these data and considering the privileged representation of these 11 loci in the Fig. 6, they should be considered as crucial research targets to improve our knowledge in the pathobiology of *C. trachomatis*.

It is also worth notable that, among the nine-member Pmp paralogues family, six genes seem to be under positive selection for specific phenotypes. Cumulative evidences indicate that Pmps may function as fine-tune determinants of *C. trachomatis* pathobiology, either by antigenic variation (15, 21, 63, 78), or host-cell adhesion (22, 36, 38), hypothetically through a shut-off mechanism at the inclusion level (77). It is posited that the accelerated evolution between paralogues is common and constitutes a mechanism for the generation of new genes and new biochemical functions (46).

Darwinian evolution generally relies on the existence of an adaptive pathway in which intermediate steps provide a gradual improvement of fitness. Thus, adaptive changes should not completely rule out synonymous mutations. In fact, the latter may alter the immediate protein adaptive landscape (by changing the proximal amino acids), providing the protein with new opportunities to evolve (14). Also, they can change RNA secondary structure and influence its stability (17) as well as originate codons with different frequency usage (associated with tRNA abundance), which was already shown to affect the translation efficiency in several microorganisms (50, 70). We have previously shown synonymous changes to more favorable codons for the *C. trachomatis* major antigen (64), and it is reasonable to expect that several other loci present synonymous changes that, in a camouflaged way, become adaptive. Our results support a non-random accumulation of synonymous mutations in *C.
trachomatis. In fact, we found that strains infecting the same human cell type are clearly the most closely related through dS analysis (Fig. 5), which suggests a non-stochastic fixation of synonymous mutations.

As concluding remarks, our results support a directional evolution of C. trachomatis towards niche-specific adaptation besides a background of Muller's ratchet deleterious mutations. Whereas the molecular basis for organ/cell-appetence is likely complex, these data suggest that population genetics and evolutionary inferences may be key factors to a comprehensive understanding of the resulting phenotypes, by guiding subsequent experimental procedures to specific targets.

ACKNOWLEDGEMENTS

This work was supported by a grant, PTDC/SAU-MII/099623/2008, from Fundação para a Ciência e a Tecnologia (FCT). VB and RF are recipients of Ph.D. fellowships (SFRH/BD/68527/2010 and SFRH/BD/68532/2010, respectively) from FCT. AN is a recipient of a post-doctoral fellowship (SFRH/BPD/75295/2010) from FCT.

REFERENCES

1. Alzhanov DT, Weeks SK, Burnett JR, Rockey DD. 2009. Cytokinesis is blocked in mammalian cells transfected with Chlamydia trachomatis gene CT223. BMC Microbiol. 9:


Chlamydia trachomatis polymorphic membrane protein D is a species-common pan-neutralizing 


characterization of chlamydiae-specific type III secretion substrates: the Inc proteins. BMC 

SNARE protein mimicry by an intracellular bacterium. PLoS Pathog. 4:e1000022.

website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res. 


nature on estimating recombination and mutation rates in Chlamydia trachomatis. G3 (Bethesda) 
2:761–768.

29. Fields KA, Mead DJ, Dooley CA, Hackstadt T. 2003. Chlamydia trachomatis type III secretion: 

Agnusdei M, Nardelli F, Bartolini E, Scarselli M, Caproni E, Laera D, Zedda L, Skibinski D, 


FIGURE LEGENDS

FIG 1 Chromosomal mapping of loci involved in the directional evolution of Chlamydia trachomatis.
From the outside in, the first and second circles (light blue lines) refer to forward and reverse coding regions, respectively, according to the published genome of the C. trachomatis strain D/UW3. The 75 evolutionary informative genes evaluated in the present study are highlighted by dark blue lines. These loci encompass about 55% of SNPs occurring within the chromosomal coding regions. The third (orange lines) and forth circles (red lines) illustrate genes found to be under positive selection by the codon-based Z-test of positive selection (MEGA5) or the branch-site test of positive selection (PAML), respectively. Circle five shows the GC skew plot. The origin of replication (ORI) and the termination region (TER) are also marked. The figure was built using DNAPlotter (16).

FIG 2 Evidence for Muller’s ratchet phenomenon. These graphs show nonsynonymous versus synonymous mutational dynamics on the C. trachomatis/C. muridarum separation process. The scattering plot depicts the results concerning the evaluation of nonsynonymous and synonymous substitutions within the C. trachomatis species (open circles) (reflecting polymorphism) and between the species C. trachomatis and C. muridarum (crosses) (reflecting divergence). Neutrality line is also shown. The box plots display the dispersion of the overall decays in the dN and dS values [i.e., dN(divergence)/dN(polymorphism) and dS(divergence)/dS(polymorphism), respectively]. Outliers and extreme values are marked with open circles and asterisks, respectively. The considerable lower decay
values for \(d_N\) suggests that the accumulation of deleterious mutations among strains from *C. trachomatis* species results from genetic bottleneck due to niche restriction (Muller’s ratchet effect).

**FIG 3** \(d_N\) *versus* \(d_S\) by gene functional category. This graph represents \(d_N\) and \(d_S\) values estimated for each gene for all 59 *C. trachomatis* strains. Housekeeping genes (HKs), genes encoding well-known cell envelope proteins (CEPs), genes coding for proteins secreted into the cytosol of the host cells or to the inclusion membrane (SEC), and “other genes” (see methods and Table S2 for details) are represented by squares, circles, triangles, and crosses, respectively. Neutrality line is also shown.

**FIG 4** Genetic variability *versus* type of mutation. Distribution of \(d_N\) (triangles) and \(d_S\) (circles) values according to the nucleotide polymorphism (mean genetic p-distance) of the 75 genes under evaluation among *C. trachomatis* strains. Slope values of the trend lines show a near four-fold higher increase of \(d_S\) with p-distance than of \(d_N\).

**FIG 5** Non-random distribution of both nonsynonymous and synonymous mutations according to tropism and ecological success. The 3D scatter plot shows the genetic distances between each of the 59 strains and the three disease groups by both \(d_N\) (grey) and \(d_S\) (black) estimations. Values were estimated by using the concatenated sequences enclosing all genes under evaluation. Strains infecting mononuclear phagocytes (shown in the bottom of the cube), the columnar epithelial cells of ocular (shown in the left side) or genital mucosae compose three major clusters for both \(d_S\) and \(d_N\). Within the non-invasive genital strains, the more clinically prevalent strains (labeled with an ellipse) are clustered apart.
FIG 6 Positive selection driving the directional *C. trachomatis* evolution towards niche-specific adaptation. The figure represents a model tree encompassing all 59 *C. trachomatis* strains that was created to facilitate a proper visualization of all biological hypotheses. These evaluate the existence of genes under positive selection (through the branch-site test of positive selection) that may be involved in: specific cell-appetence to columnar epithelial cells of ocular (H1) (serovars A-C) or genital mucosae (H2) (serovars D-K), and to mononuclear phagocytes (H3) (serovars L1-L3); pathogenic diversity among strains causing ocular disease (H4), genital disease (H5), or hemorrhagic proctitis and suppurative lymphadenitis (H6). This test was applied to each individual gene tree in an independent manner for each hypothesis depending of the tree topology. The figure boxes show the genes found to be positively selected for each biological hypothesis. The likelihood-ratio test (LRT) was used for inferring the statistic significance (p-values) of positive selection in the foreground branches. ** and * indicate significance with p < 0.01 and p < 0.05, respectively. § refers to genes presenting congruent trees, but for which a specific biological hypothesis may be affected by recombination (49). See Table 1 for details on positively selected codons.

**TABLE 1** Positively selected genes and the inferred codons putatively involved in specific adaptive evolution based on the branch-site test of positive selection by PAML.

<table>
<thead>
<tr>
<th>ORF a (gene)</th>
<th>Biological hypothesis with positive selection (H1 to H6) b</th>
<th>p-value (LRT test) c</th>
<th>Specific codons under positive selection d, e</th>
<th>Putative function / experimental evidences</th>
<th>Specific biological hypothesis excluded due to putative recombination f</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT050</td>
<td>H1</td>
<td>p &lt; 10^-10</td>
<td>n/d</td>
<td>Pmp-like protein identified in the inclusion lumen (48, 71).</td>
<td>H2, H5</td>
</tr>
<tr>
<td></td>
<td>H3</td>
<td>0.0352</td>
<td>51K*, 80K*, 366K*, 523S*, 546K*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strain</td>
<td>p-Value</td>
<td>ID Variants</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT051</td>
<td>0.0004</td>
<td>563V*</td>
<td>Pmp-like protein (48, 71).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT105</td>
<td>0.0004</td>
<td>435R**</td>
<td>Function unknown.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT115</td>
<td>0.0004</td>
<td>11D**, 12G*</td>
<td>Inc (68). T3SS effector (75).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT147</td>
<td>0.0003</td>
<td>1407I*</td>
<td>Inc; human early endosomal antigen 1 (EEA1) homologue (8). Immunodominant antigen (69). Involvement in pathogenic differences in vivo (51).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT222</td>
<td>0.0051</td>
<td>124I*, 125S**, 126V**</td>
<td>Inc associated with host kinases in microdomains that interact with the host centrosomes (58).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT223</td>
<td>0.0082</td>
<td>99R*, 152S*</td>
<td>Inc (6). T3SS effector (75). Subversion of intracellular trafficking (25). Host cell cytokinesis blockage (1).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT249</td>
<td>0.0004</td>
<td>8Y*, 24N*, 80T**, 89I**</td>
<td>Inc (45). T3SS effector (24).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strain</td>
<td>H3</td>
<td>H4</td>
<td>H5</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>CT288</td>
<td>0.0104</td>
<td>n/d</td>
<td></td>
<td>Inc (6). T3SS effector (75).</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H1, H2, H4, H5</td>
<td></td>
</tr>
<tr>
<td>CT413</td>
<td>0.0015</td>
<td>n/d</td>
<td></td>
<td>Adhesin (38). Antigen (21, 35, 78).</td>
<td></td>
</tr>
<tr>
<td>(pmpB)</td>
<td></td>
<td></td>
<td></td>
<td>H2, H5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0198</td>
<td>58A*, 75T**, 235A*, 820A*, 947E*, 998A*, 1061N**, 1171T**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0017</td>
<td>n/d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT414</td>
<td>p &lt; 10^-6</td>
<td>145P*, 544I**, 598V**</td>
<td>Adhesin (38). Antigen (21, 35, 78).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(pmpC)</td>
<td></td>
<td></td>
<td></td>
<td>H2, H5</td>
<td></td>
</tr>
<tr>
<td>CT442</td>
<td>0.0154</td>
<td>24A*, 29K**, 48I*, 104I**, 133D*, 137V**</td>
<td>Inc (6). T3SS effector (75). Antigen (72).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(crpA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT456</td>
<td>0.0094</td>
<td>447H*, 978H*</td>
<td>Translocated actin-recruiting phosphoprotein / early T3SS effector (20). Contribution for the pathogen phagocytosis (44). Involvement in pathogenic differences in vivo (51). Antigen (30, 81).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(tarp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p &lt; 10^-24</td>
<td>n/d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0049</td>
<td>n/d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p &lt; 10^-8</td>
<td>189A**, 237G*, 407S*, 481A**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT529</td>
<td>0.0009</td>
<td>3A*</td>
<td></td>
<td>Inc (31). T3SS effector (75). Antigen (31, 81).</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
<p>| CT622 | H3 | 0.0012 | n/d | Involvement in pathogenic differences in vivo (51). T3SS effector (37). Antigen (21, 30). |
| CT694 | H? | 0.0190 | n/d | Immunodominant antigen (21, 69). Early T3SS effector (40). Modulation of host cell processes (10). |
| CT859 (ispH) | H3 | 0.0135 | n/d | 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (74). |
| CT867 | H3 | 0.0141 | n/d | Deubiquitinase and deneddylase (57). |
| CT868 | H3 | 0.0064 | n/d | Deubiquitinase and deneddylase (57). Involvement in pathogenic differences in vivo (51). Inhibition of a crucial pathway for host inflammatory responses (54). |
| CT869 | H1 | 0.0348 | 59N*, 139I*, 469A* | Adhesin (38). |</p>
<table>
<thead>
<tr>
<th>Strain</th>
<th>Gene</th>
<th>Hypothesis</th>
<th>p-value</th>
<th>n/d</th>
<th>Adhesin (38)</th>
<th>Antigen (15)</th>
<th>Antigen (21, 35, 78)</th>
<th>H2, H5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT870</td>
<td>pmpE</td>
<td>H1</td>
<td>$p &lt; 10^{-4}$</td>
<td>n/d</td>
<td>H1</td>
<td>Adhesin (38)</td>
<td>Antigen (15)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H3</td>
<td>0.0020</td>
<td>n/d</td>
<td>Adhesin</td>
<td>Antigen (15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT871</td>
<td>pmpG</td>
<td>H3</td>
<td>0.0082</td>
<td>n/d</td>
<td>H3</td>
<td>Adhesin (38)</td>
<td>H2, H5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H6</td>
<td>$p &lt; 10^{-4}$</td>
<td>258G*, 320K*, 324S*, 812Q**</td>
<td>H3</td>
<td>H3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT872</td>
<td>pmpH</td>
<td>H5</td>
<td>0.0303</td>
<td>n/d</td>
<td>H5</td>
<td>Adhesin (38)</td>
<td>Antigen (21, 78)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H6</td>
<td>0.0302</td>
<td>n/d</td>
<td>H6</td>
<td>Antigen (21, 78)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Open reading frame (ORF) numbers are based on genome annotation of the strain D/UW3 (GenBank No. NC_000117).

b The hypotheses were created to evaluate the existence of genes under positive selection involved particular biological processes: specific cell-appetence to columnar epithelial cells of ocular (H1) or genital mucosae (H2), and to mononuclear phagocytes (H3); pathogenic diversity among strains causing ocular disease (H4), genital disease (H5), or hemorrhagic proctitis and suppurative lymphadenitis (H6).

c The likelihood-ratio test (LRT) was used for inferring the statistic significance of positive selection in the foreground branches (p-value). The degree of freedom is 1 for the comparisons of alternative hypothesis versus the null hypothesis in the branch-site test 2.

d The posterior probabilities that each site belongs to the site class of positive selection on the foreground lineages are inferred by the Bayes Empirical Bayes (BEB) analysis. Positively selected sites are those with $** p > 0.99$ and $* p > 0.95$. n/d, not discriminated: an excess of positively selected codons hampered their discrimination by PAML, or the identified codons revealed $p < 0.95$.

e For simplification purposes, amino acid positions for biological hypotheses H1 and H4 are based on the protein sequence annotation for the strain A/Har13, whereas for H2 and H5 are based on the
annotation for strain D/UW3, and for H3 and H6 refer to strain L2/434. Within CT456/tarp and CT868, the position of the codons labeled with $^\text{§}$ referred to strain L1/1322/p2 as L2/434 is deleted in this region.

$^\text{f}$ Recombination was detected in previous studies (39, 41, 49).

$^\text{g}$ This gene was detected to be under positive selection specifically for the most clinically prevalent genital strains (mostly from E and F serovars).