Polymorphisms in Inc proteins and differential expression of inc genes among

*Chlamydia trachomatis* strains correlate with invasiveness and tropism of

lymphogranuloma venereum isolates

Running Title: *C. trachomatis* Incs and invasiveness of LGV strains

Filipe Almeida¹*, Vítor Borges²*, Rita Ferreira², Maria José Borrego², João Paulo Gomes², and Luís Jaime Mota¹,³#

¹Infection Biology Laboratory, Instituto de Tecnologia Química e Biológica, Universidade Nova
der Lisboa, Oeiras, Portugal.

²Department of Infectious Diseases, National Institute of Health, Lisbon, Portugal.

³Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de
Lisboa, Caparica, Portugal.

*These two authors equally contributed to this work.

#Corresponding author: jmota@itqb.unl.pt
Chlamydia trachomatis is a human bacterial pathogen that only multiplies within an intracellular membrane-bound vacuole, the inclusion. C. trachomatis includes ocular and urogenital strains, usually causing infections restricted to epithelial cells of the conjunctiva and genital mucosa, respectively, and lymphogranuloma venereum (LGV) strains, which can infect macrophages and spread into lymph nodes. However, C. trachomatis genomes display > 98% of identity at DNA level. In this work, we studied whether C. trachomatis Inc proteins, showing a bilobal hydrophobic domain that may mediate their insertion in the inclusion membrane, could be a factor determining these different types of infection and tropisms. Analyses of polymorphisms and phylogeny of 48 IncS from 51 strains encompassing the three disease groups showed significant amino acid differences that were mainly due to variations between IncS from LGV and ocular or urogenital isolates. Studies of the evolutionary dynamics of Inc genes suggested that 10 IncS are likely under positive selection and indicated that most non-silent mutations are LGV-specific. Additionally, real-time quantitative PCR analyses in prototype and clinical strains covering the three disease groups identified 3 Inc genes showing LGV-specific expression. We determined the transcriptional start sites of these genes and found LGV-specific nucleotides within their promoters. Thus, subtle variations in the amino acids of a subset of Inc proteins and in the expression of Inc genes may contribute for the unique tropism and invasiveness of C. trachomatis LGV strains.
INTRODUCTION

*Chlamydiae* are a large group of obligate intracellular bacteria. It includes *Chlamydia trachomatis*, which causes ocular and genital infections in humans. These infections are the leading cause of preventable blindness in developing countries (62), and the most prevalent cause of bacterially sexually transmitted diseases worldwide (4). *C. trachomatis* strains include trachoma and LGV (lymphogranuloma venereum) biovars (48). The trachoma biovar comprises ocular and urogenital strains, which cause localized infections of the epithelial surface of the conjunctiva and genital mucosa, respectively; strains of the LGV biovar cause invasive urogenital disease, due to their ability to infect macrophages and spread into lymph nodes. *C. trachomatis* strains can be further classified into ocular serovars A-C; urogenital serovars D-K; and LGV serovars L1-L3.

*C. trachomatis* genomic sequences of different ocular, urogenital and LGV strains exhibit >98% of identity and a high degree of synteny (12, 25, 29, 31, 50, 53, 55, 59, 60). Therefore, the determinants of the different types of infection (invasive or non-invasive) and tissue tropism (eyes, genitals, and lymph nodes) must rely on the few genes present in some strains but not in others, and on nucleotide differences which may either lead to proteins with disease group-specific amino acids or to differential gene expression. Some of these determinants have been suggested in previous studies: the tryptophan (*trpRBA*) operon (10, 19, 51), and genes encoding cytotoxin (11), phospholipase (40), polymorphic membrane proteins (Pmps) (24), and Tarp (34).

*Chlamydiae* are characterized by a developmental cycle involving the inter-conversion between an infectious form, the elementary body, and a non-infectious form, the reticulate body (1). Throughout development, the bacteria reside and multiply within a membranaceous compartment, known as the inclusion, and manipulate host cells by using a type III secretion
(T3S) system to translocate effector proteins into host cells (6, 61). Chlamydial T3S substrates have been found by the identification of an N-terminal secretion signal using Salmonella (26), Shigella (56, 57), or Yersinia (14, 15, 20, 21, 27) as heterologous hosts. These include inclusion membrane proteins (Incs), characterized by a bilobal hydrophobic motif thought to mediate their insertion into the inclusion membrane (16, 21, 45, 56, 57). Incs from the same chlamydial species are normally unrelated to each other (16, 35), and only a subset of ~25 Incs is conserved between species (16, 35). C. trachomatis Incs CT119/IncA, CT115/IncD, CT147, CT229, and CT813 have been shown or suggested to subvert host cell vesicular and non-vesicular transport (17, 18, 46). However, virtually nothing is known about the biological role of most Incs, which also reflects the lack of straightforward methods to genetically manipulate Chlamydiae.

In this work, we used phylogenetic, molecular evolution, and gene expression analyses to analyze if Incs could impact on the type of infection and tissue tropism associated with C. trachomatis. Our studies suggest that a subset of Incs might play a role in the unique capacity of C. trachomatis LGV strains to infect macrophages and disseminate into lymph nodes.

MATERIALS AND METHODS

Bacterial strains and growth conditions. C. trachomatis prototype strains B/Har36, C/TW3, E/Bour, L2/434, and L3/404 (from ATCC) and clinical strains F/CS465-95 and L2b/CS19-08 (from the collection of the Portuguese National Institute of Health) were used as detailed below, and propagated in HeLa 229 cells (from ATCC) using standard techniques (49).

Escherichia coli TOP10 (Invitrogen) was used for construction and purification of the plasmids. Yersinia enterocolitica ΔHOPEMT (MRS40 plML421 [yopHΔ1-352; yopOΔ65-558].

on July 15, 2017 by guest http://jb.asm.org/ Downloaded from

on July 15, 2017 by guest
yopP23, yopE21, yopM23, yopT135) (28), deficient for the Yersinia T3S effectors YopH, O, P, E, M, and T, but T3S-proficient, was used for T3S assays. To construct a T3S-deficient derivative of ΔHOPEMT, we deleted in this strain the complete coding sequence (codons 1-354) of the yscU gene, which encodes an essential component of the Y. enterocolitica T3S system (54). This was done by allelic exchange with the mutator plasmid pLY16 (54). The resulting Y. enterocolitica ΔHOPEMT ΔYscU strain was also used in T3S assays. The yscUΔ1-354 mutation had been previously shown to be non-polar (54). E. coli or Y. enterocolitica were routinely grown in liquid or solid Luria-Bertani medium with the appropriate antibiotics and supplements. Plasmids were introduced into E. coli or Y. enterocolitica by electroporation.

**Construction of plasmids.** Plasmids were constructed and purified using proof-reading Phusion DNA polymerase (Finnzymes), restriction enzymes (MBI Fermentas), T4 DNA Ligase (Invitrogen), DreamTaq DNA polymerase (MBI Fermentas), NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel), and GeneElute Plasmid Miniprep kit (Sigma), according to the instructions of the manufacturers. In brief, to analyze T3S signals we constructed plasmids harboring hybrid genes encoding the N-terminal 20 amino acids of each C. trachomatis Inc, or of the Y. enterocolitica T3S chaperone SycT (28), and the mature form of TEM-1 β-lactamase (TEM-1). These hybrids were made by PCR using plasmid pCX340 as template (13). Each forward primer contained an Ndel site followed by 60 nucleotides encoding the first 20 amino acids of each Inc (or of SycT) in frame with a sequence complementary to the 3’ extremity of the transcribed strand of the TEM-1-encoding gene; each reverse primer contained either a HindIII or a XhoI site followed by a sequence complementary to the 3’ extremity of the non-transcribed strand of the TEM-1-encoding gene. Digested PCR products were ligated into pLJM3, a low-
copy plasmid which enables expression of cloned genes driven by the promoter of the *Y. enterocolitica* yopE gene (36). A similar strategy was used to construct a plasmid encoding TEM-1 alone, except that the forward primer did not contain an *inc* gene sequence. For primer design, the DNA sequence of each *inc* gene in *C. trachomatis* strain L2/434 (Table S1) was used, except for *ct036*, *ct115/incD*, and *ct119/incA*, in which the sequence from strain D/UW3 was used (Table S1); the DNA sequence of *sycT* was from *Y. enterocolitica* pYVe227 (accession number AF102990). The sequence of all primers is available upon request. The accuracy of the nucleotide sequence of all the inserts in the constructed plasmids was checked by DNA sequencing.

**Y. enterocolitica** T3S assays. These analyses were done as previously described (54). Proteins in bacterial pellets and culture supernatants were analyzed by immunoblotting with mouse monoclonal anti-TEM-1 antibodies (QED Bioscience; 1:500), and with rabbit polyclonal anti-SycO antibodies (1:1000) (32), and detected with a ChemiDoc XRS+ system (BioRad). The amount of protein in the supernatant relative to the amount of total protein (% of secretion), was determined from immunoblot images with Image Lab (Bio-Rad).

**DNA sequences.** Regardless of the origin of the *C. trachomatis* gene, throughout this work we used the nomenclature of the annotated D/UW3 strain (Table S1). The nucleotide sequences of the *inc, pmp*, and housekeeping genes analyzed were from the available genomes of 51 *C. trachomatis* strains (12, 25, 29, 31, 50, 53, 55, 59, 60). The strains and corresponding genome accession numbers are listed in Table S1. The DNA sequences were retrieved from pairwise alignments obtained by BLAST. All sequences were manually inspected and corrected.
for accuracy and completeness. We excluded from further analysis a few DNA sequences containing ambiguous nucleotides. Whenever there were distinct annotations in GenBank for the start codon of the same inc gene in C. trachomatis archetype ocular (A/Har13), urogenital (D/UW3), and LGV (L2/434) strains, we used: for ct036 and ct119/incA, in all cases the start codons of each of the genes as annotated for D/UW3 and A/Har13; for ct115/incD and ct192, the start codons of each of the genes as annotated for A/Har13, D/UW3, or L2/434, in ocular, urogenital or LGV strains, respectively; for ct226, in all cases the start codon as annotated for A/Har13 (Tables S1 and S3).

Sequence alignments, and analyses of polymorphisms, phylogeny, and molecular evolution. Alignments of the amino acid sequences of the Inc, Pmp, and housekeeping proteins, deduced from the retrieved nucleotide sequences, were generated using the ClustalW algorithm in MEGA5 software (www.megasoftware.net) (58). All alignments were manually inspected and corrected for artifacts. We excluded all strain-specific pseudogenes (Table S3) from further phylogenomic and evolutionary analyses.

For the analyses of polymorphism, phylogeny, and molecular evolution, various tools present in MEGA5 were used, essentially as previously described (24). Briefly, for analyses of polymorphism, we computed pairwise, overall, within groups (ocular, urogenital, or LGV), and between groups (ocular vs urogenital, ocular vs LGV, or urogenital vs LGV) amino acid p-distances. For analyses of phylogeny, trees were generated using the Neighbor-Joining method (47). The generated phylograms of Inc proteins were inspected for separate branches (segregation) of all ocular, urogenital, or LGV strains. This analysis was supported by comparison of pairwise amino acid p-distances between and within disease groups. The
phylogeny of an Inc protein was considered to segregate a disease group if the maximum pairwise amino acid $p$-distance within groups was less than any of the pairwise amino acid $p$-distances between groups. For molecular evolution analyses, we used the Kumar method (39) to compute overall means of non-synonymous ($d_N$) and synonymous ($d_S$) substitutions (per non-synonymous or synonymous sites, respectively) and to find genes that may be under positive selection. By using the codon based Z-test of selection in MEGA5, genes were considered under positive selection if they showed a statistically significant value ($p < 0.05$) to reject the null hypothesis of strict-neutrality ($d_N = d_S$) in favour of both positive selection ($d_N > d_S$) and lack of neutrality ($d_N \neq d_S$). All these analyses were performed by Bootstrap with 1,000 replicates, and with the pairwise-deletion option selected.

Real-time quantitative PCR. The expression of $inc$ genes during the developmental cycle of $C. trachomatis$ B/Har36, C/TW3, E/Bour, F/CS465-95, L2/434, L2b/CS19-08, and L3/404 was estimated by determining $inc$ mRNA levels at different times post-infection by real-time quantitative PCR (RT-qPCR). These experiments were essentially done as previously described (7, 42). Briefly, for each strain, 6 tissue culture flasks of 25 cm$^2$ containing monolayers of HeLa 229 cells were inoculated at a multiplicity of infection of 1; cells were harvested at 2, 6, 12, 20, 30 and 42 h post-infection, by scraping in ice-cold phosphate buffered saline. The cell suspension was sonicated to disrupt mammalian cells and promote bacterial release, followed by low-speed centrifugation at 4°C. The supernatant was then frozen in liquid nitrogen and stored at -80°C. These samples were used for total RNA purification and generation of cDNA, as previously described (7). Primers (available upon request) were designed for each $inc$ gene using Primer Express (Applied Biosystems), based on identical $C. trachomatis$ sequences between
strains. The RT-qPCR assays were done using the ABI 7000 SDS, SYBR Green chemistry and optical plates (Applied Biosystems), as previously described (7, 42). At each time-point, raw RT-qPCR data of each inc gene was normalized against the data obtained for the 16S rRNA transcript, as it was previously demonstrated that this is a good endogenous control (7). The final results were based on at least two independent experiments.

Transcription linkage analysis and identification of transcriptional start sites. We searched for disease group-specific nucleotides within the promoter region of 3 inc genes (ct058, ct192, and ct214) that showed differential gene expression. For this, we determined their transcriptional start sites (TSSs). For ct058 and ct192, it was ambiguous whether their promoter would lie immediately upstream from their predicted start codons (Fig. 5A). Therefore, we used reverse transcription coupled with PCR (RT-PCR) to determine if ct058 and ct059, or ct192 and ct193, are part of the same transcriptional unit. For this, RNA was isolated from HeLa 229 cells infected for 30 h with C. trachomatis L2/434 using NZY Total RNA kit (NZYTech). cDNA was then generated by using random hexamers and iSCRIPT (BioRad). Primers (Table S2) were designed to generate PCR products containing ~300 bp upstream and downstream from the predicted start codons. PCR products were obtained using DreamTaq DNA polymerase (MBI Fermentas). As controls for the PCRs, we also used as template: the product of a typical reverse transcription reaction but without iSCRIPT; total DNA, isolated using the NZY Tissue gDNA kit (NZYTech), either from cells infected with strain L2/434 for 42 h or left uninfected.

The identification of the TSSs of ct059 (upstream from ct058, and in the same transcriptional unit; Fig. 5A), ct192, and ct214 in L2/434 was done by 5’ rapid amplification of cDNA ends (RACE), using the 5’/3’RACE kit, 2nd Generation (Roche). We used RNA isolated
as described above from HeLa 229 cells infected for 30 h with *C. trachomatis* L2/434 and the
primers listed in Table S2. Final PCR amplification of double stranded cDNA was done with
Phusion DNA polymerase (Finnzymes). PCR products were purified after agarose gel
electrophoresis, using High Pure PCR Purification Kit (Roche), and then subjected to DNA
sequencing. To analyze the determined TSSs in the context of the promoter regions of *ct059-
c058, ct192, and ct214* in all strains used in the RT-qPCR assays, the corresponding nucleotide
sequences were either retrieved from GenBank (E/Bour, L2/434, and L3/404; Table S1) or
determined by DNA sequencing (C/TW3, B/Har36, F/CS465-95, and L2b/CS19-08), as
previously described (24) and using the primers listed in Table S2. All these manipulations were
done according to instructions from the indicated manufacturers.

**Nucleotide sequence accession numbers.** The sequences of the promoter regions of
*ct192, and ct214* and of *ct059-ct058* in C/TW3, B/Har36, F/CS465-95, and L2b/CS19-08
determined in this study were submitted to GenBank and are available under accession numbers
JX451863 - JX451874.

**RESULTS**

**Identification of T3S signals in *C. trachomatis* Inc proteins.** We focused on 48
predicted *C. trachomatis* Inc proteins (i.e. possessing a bilobal hydrophobic domain), which have
been singled out and studied by Li et al (33) (Table 1). Thus far, 23 of these proteins have been
detected in the inclusion membrane by immunofluorescence microscopy using specific
antibodies (known Incs) [(16, 33, 38), and references therein], but not the other predicted Incs (putative Incs) (16, 35) (Table 1).

As Incs are believed to be transported into the inclusion membrane by a T3S mechanism (16, 21, 56, 57), we used *Y. enterocolitica* as heterologous bacteria to identify a possible N-terminal T3S signal in putative Incs of *C. trachomatis* by comparison to known Incs. We sought to obtain additional indirect evidence (besides the bilobal hydrophobic motif) that putative Incs localize to the inclusion membrane. We analyzed secretion of hybrid proteins comprising the first 20 amino acids of each putative or known Inc and TEM-1 by T3S-proficient (ΔHOPEMT) or T3S-deficient (ΔHOPEMT ΔYscU) *Y. enterocolitica* (Fig. 1). The different *Y. enterocolitica* strains were incubated in T3S-inducing conditions (54), followed by fractionation of the bacterial cultures into culture supernatants and bacterial pellets and subsequent immunoblotting analyses of the proteins in the two fractions (examples in Fig. 1A and B). In total, we have analyzed T3S signals in 24 putative Incs (18 have not been previously analyzed for T3S) and in 15 known Incs (7 have not been previously analyzed for T3S) (Fig. 1C and Table 1). The expression levels of the TEM-1 hybrid of putative Inc CT006 were extremely low, which hampered the analysis of a T3S signal in this protein (data not shown). These experiments led to the identification of a T3S signal in 18 putative Incs and in 12 known Incs (Fig. 1C and Table 1). This revealed 5 known and 14 putative Incs as novel *C. trachomatis* T3S substrates (Fig. 1C and Table 1). However, we did not detect a clear T3S signal in three known Incs (CT101, CT225, and CT850; Fig. 1 and Table 1). Their T3S signal could extend beyond the first 20 amino acids or might not be recognized by the *Y. enterocolitica* T3S system; alternatively, they may be transported into the inclusion membrane by a distinct mechanism. Regardless of the exact explanation, this implies that the lack of a detectable T3S signal could not be taken as a definitive indication that a
putative Inc does not localize to the inclusion membrane. Furthermore, the percentage of putative
and known Incs analyzed that displayed an N-terminal region recognized by the *Y. enterocolitica*
T3S machinery was nearly identical [75% (18 out of 24) and 80% (12 out of 15), respectively].
Overall, these analyses indicated that most of the putative Incs analyzed are T3S substrates.

Differences in the amino acid sequences of Inc proteins among *C. trachomatis* strains
correlate with the type of infection and with tissue tropism. The nucleotide sequences of the
genes encoding the selected 48 known and putative Inc proteins (hereafter named Incs) were
retrieved from 51 fully sequenced *C. trachomatis* genomes (7 ocular, 23 urogenital, and 21 LGV
strains; Table S1). In an initial analysis, 7 incs showed to be pseudogenes in different *C.
trachomatis* strains (*ct058, ct101, ct135, ct192, ct227, ct228, ct300; Table S3). *ct358* was
described as a pseudogene in L2/434 and L2/UCH-1(59), and its nucleotide sequence is 100%
identical in all LGV strains. However, analysis of its nucleotide sequence suggests that *ct358*
may encode a functional protein that is 7 amino acids shorter at its C-terminus than CT358 (178
amino acid residues long) in ocular and urogenital strains (Table S3). In all the 7 incs that we
identified as pseudogenes the full-length gene is disrupted by a mutation that leads to a
significantly truncated protein. The only disease group-specific correlation was observed with
*ct300*, which is a pseudogene in all LGV strains analyzed. Therefore, the encoded protein is
expendable for LGV infections. In addition, *ct058* and *ct101* revealed to be pseudogenes in
almost all analyzed ocular and urogenital strains, respectively (Table S3). *ct135, ct192, ct227,
and ct228*, are only pseudogenes in a few strains, with no obvious correlation with ocular,
urogenital, or LGV disease groups (Table S3). Furthermore, 12 inc genes showed small deletion
and insertion events (Table S3).
To understand if the amino acid sequence of Inc proteins varies among strains, we determined the overall mean genetic distance (amino acid $p$-distance) for each Inc among all 51 *C. trachomatis* strains (discarding strain-specific pseudogenes) (Fig. 2A). As reference, we also analyzed the 9 Pmps and 9 housekeeping proteins of *C. trachomatis* previously shown to be polymorphic (43) (Fig. 2A). The Pmps should localize to the bacterial outer membrane and the housekeeping proteins within the bacterial cell. The average $p$-distance was 0.017 (SEM, 0.002) for Incs, 0.020 (SEM, 0.007) for Pmps, and 0.013 (SEM, 0.003) for housekeeping proteins. Based on this, and considering the average $p$-distance for Incs as a cut-off value, we defined 19 (40% of the total) Incs as “polymorphic” ($p$-distance ≥ 0.017) (Fig. 2A). As comparison, 4 Pmps (44%) and 2 housekeeping proteins (22%) displayed a $p$-distance of ≥ 0.017. This showed that the overall degree of polymorphism in Inc proteins among *C. trachomatis* strains is similar to that of Pmps and is higher than that of known polymorphic housekeeping proteins.

To understand if the amino acid differences between Inc proteins were related with the type of infection and with tissue tropism, we determined the average $p$-distance of Incs within and between the three groups of strains (ocular, urogenital, and LGV) (Fig. 2B). This showed that the differences were largely due to variations between Incs from LGV strains and ocular (average $p$-distance = 0.031; SEM, 0.004) or urogenital strains (average $p$-distance = 0.029; SEM, 0.004) (Fig. 2B). These average $p$-distances were significantly higher (in all cases, $p < 0.0001$; two-tailed $t$-test) than those between Incs from ocular and urogenital strains (average $p$-distance = 0.011; SEM, 0.002) or within Incs from the same disease groups (average $p$-distance < 0.006) (Fig. 2B). Similar observations were made for Pmps except that, in contrast to Incs, the average $p$-distance between Pmps from ocular and urogenital strains was not significantly different than that between Pmps from LGV strains and ocular or urogenital strains (in both
cases, \( p > 0.05 \); two-tailed \( t \)-test) (Fig. 2B). Also in contrast to the Incs, the variation in the amino acid sequences of housekeeping proteins between each of the three groups were clearly not different (in all cases, \( p > 0.05 \); two-tailed \( t \)-test) (Fig. 2B). Furthermore, the amino acid sequences of housekeeping proteins varied nearly as much between or within groups (Fig. 2B).

The main exception was the LGV group, within which the housekeeping proteins, as Incs and Pmps, showed to be extremely conserved (average \( p \)-distance < 0.002 for the three cases) (Fig. 2B). Thus, the separation (\( p \)-distance) between LGV strains and ocular or urogenital strains is much more marked for Incs than for Pmps or housekeeping proteins.

We then made and analyzed phylogenetic reconstructions based in the amino acid sequence of Inc proteins. The phylograms of 5 Incs (10\%) showed tropism, i.e. segregation of the three disease groups, and those of 38 Incs (84\%) evidenced segregation of at least one disease group (Fig. 2C and Table S4). The phylograms of a total of 35 Incs (73\%) showed segregation of LGV strains, while only 12 (25\%) and 8 (17\%) displayed clustering of ocular and urogenital strains, respectively (Fig. 2C and Table S4). This scenario was mirrored by the phylograms of Pmps but was in contrast with the phylograms of housekeeping proteins, in which segregation by disease group was less often seen (Fig. 2C and Table S4).

In summary, there are significant differences in the amino acid sequences of Inc proteins among \( C. \) \( t r a c h o m a t i s \) strains and they correlate with the type of infection and with tissue tropism. In particular, differences are mostly between Incs from LGV strains and Incs from ocular or urogenital strains.

\textbf{inc} genes of \( C. \) \( t r a c h o m a t i s \) have distinct evolutionary dynamics and several incs are likely under positive selection. To understand the underlying evolutionary pressures that drive
amino acid changes in Inc proteins, we analyzed the molecular evolution of inc genes. We first
determined overall $d_N/d_S$ values for inc genes, by comparison to the 9 pmp genes and the 9
selected housekeeping genes of C. trachomatis. We found that 24 inc genes (50%) showed a
$ds/\delta s$ value which was $> 1$, and that in 4 inc genes all substitutions were non-synonymous (Fig.
3A and Table S5). In contrast, only 2 pmp genes (22%) showed a $d_N/d_S$ value that was $> 1$, and
all housekeeping genes displayed a $d_N/d_S$ value of $< 1$ (Fig. 3A and Table S5).

Analyses of the overall $d_N/d_S$ values using the codon based Z-test of selection (see
Materials and Methods) yielded statistically significant values for 10 inc genes (ct116/incE,
ct118/incG, ct119/incA, ct222, ct223, ct228, ct229, ct249, ct288, and ct813), but only for one
pmp gene and for no housekeeping gene (Fig. 3A and Table S5). This indicated that these 10 inc
genes are likely under positive selection. We then aimed to understand which group of strains
might cause the detection of this possible evolutionary trend. For this, we assessed the impact of
artificially removing ocular, urogenital, or LGV strains from the analyses (Fig. 3B). This showed
that discarding LGV strains caused major alterations on the $d_N$ and $d_S$ values and confined the
ability to detect Z-test-based likely positive selection to only one inc gene, whereas discarding
ocular or urogenital strains had less pronounced effects (Fig. 3B). In addition, within each of the
10 proteins encoded by inc genes likely under positive selection, we found 113 amino acid
residues that are disease group-specific (Table 2). Among these residues, 104 (92%) were in Inc
proteins from LGV strains, 77 (74%) of which localized in regions of the proteins predicted to be
on the cytoplasmic side of the inclusion membrane (Table 2).

Overall, this suggested that the evolutionary dynamics of incs is distinct from pmps or
housekeeping genes, and that LGV-specific amino acid residues in a subset of Inc proteins might
be involved in the unique ability of *C. trachomatis* LGV strains to infect macrophages and diseminate into lymph nodes.

**Disease group-specific expression of* C. trachomatis* inc genes.** To analyze if there were differences in the expression of *inc* genes between *C. trachomatis* strains that correlate with the type of infection or with tissue tropism, we used RT-qPCR to determine the mRNA levels of the 48 selected *incs* throughout the developmental cycle of *C. trachomatis*. We aimed to find *inc* genes showing differences in the highest mRNA levels during the cycle (peak of expression) or in the variation of mRNA levels throughout development (profile of expression) between *C. trachomatis* strains.

We first infected HeLa 229 cells with ocular (C/TW3), urogenital (E/Bour), or LGV (L2/434) prototype strains. Total RNA was isolated at 2, 6, 12, 20, 30, and 42 h post-infection, which was used to generate cDNA for RT-qPCR assays (complete data is shown in Table S6). Generally, the comparison of the peak of expression revealed differences from 30 to 60-fold (depending on the strain) between *inc* genes (Fig. 4A). However, the average of the peaks of expression of *inc* genes in C/TW3, E/Bour, or L2/434 were not significantly different between each strain (Fig. 4A). Regarding the profile of expression, essentially as previously described (5, 41, 52), we have identified: *incs* with highest expression at 2 or 6 h post-infection that then either decreased or remained constant throughout the cycle (“early-cycle genes”); *incs* with highest expression only at 12 or 20 h post-infection and which then decreased or remained constant at later time points (“mid-cycle genes”); and *incs* with highest expression only at 30 or 42 h post-infection (“late-cycle genes”) (Fig. 4B and examples in Fig. S1). We also identified 4 *inc* genes in E/Bour and 5 *inc* genes in L2/434 that were simultaneously “early- and late-cycle genes”,...
showing identically high mRNA levels both at 2 or 6 h post-infection and at 30 or 42 h post-infection but lower expression at mid-cycle (Fig. 4B). This was typically the case of ct214 and ct288 in L2/434 (Fig. S2). Generally, 35 inc genes showed the same profile of expression in the three C. trachomatis strains, and the majority of incs showed an “early-cycle” profile of expression (34 in C/TW3, 27 in E/Bour, and 32 in L2/434) (Fig. 4B). In spite of these common features, we identified 9 inc genes (ct005, ct058, ct192, ct214, ct232/incB, ct249, ct288, ct440, and ct442) whose peak of expression consistently showed differences of > 2-fold between strains and/or whose profile of expression displayed differences that could not be explained by distinct growth kinetics of the strains (Fig. S2). With the exception of ct214, all these genes showed consistently higher peaks of expression in L2/434 than in C/TW3 and/or E/Bour; ct442 also showed a higher peak of expression in E/Bour than in C/TW3 (Fig. S2).

To analyze whether the differences found in the expression of ct005, ct058, ct192, ct214, ct232/incB, ct249, ct288, ct440, and ct442 were strain-specific or disease group-specific, we determined the mRNA levels of these genes during the developmental cycle of additional C. trachomatis strains. For this, we infected HeLa 229 cells with ocular B/Har36 (prototype), urogenital F/CS465-95 (clinical isolate), LGV L2b/CS19-08 (clinical isolate), or LGV L3/404 (prototype) strains. The infected cells were processed for RT-qPCR assays and analyzed as described above (complete data is shown in Table S6). We only detected disease group-specific differences in gene expression for ct058, ct192, and ct214 (Fig. 4C): ct058 showed an “early-cycle gene” profile of expression in which mRNA levels were evident for LGV strains but only vestigial for ocular and urogenital strains; ct192 showed only a clear “early-cycle gene” profile of expression in LGV strains and its expression levels were generally higher in LGV strains than in ocular or urogenital strains; ct214 displayed an “early- and late-cycle gene” profile of
expression in LGV strains but a “late-cycle gene” profile of expression in ocular or urogenital 
strains. Therefore, we have identified 3 inc genes (ct058, ct192, and ct214) showing differences 
in gene expression between C. trachomatis strains that correlate with the type of infection and 
tissue tropism, in particular with LGV isolates.

Identification of LGV-specific nucleotides in the promoter regions of ct058, ct192, 
and ct214. We next attempted to obtain insights on the genetic basis for the disease group-
specific expression of ct058, ct192 and ct214 by analyzing the promoter region of these genes in 
C. trachomatis L2/434. The gene organization of these incs suggested that the promoter region of 
ct214 should lie between the start codons of ct214 and ct215 or within the first codons of ct215 
(Fig. 5A). However, the localization of the promoter regions of ct058 or ct192 was unclear, as 
these genes could be co-transcribed with ct059 or ct193, respectively (Fig. 5A). Therefore, we 
used RT-PCR with cDNA template generated from total RNA of HeLa 229 cells infected with C. 
trachomatis L2/434 to determine the possible transcriptional linkage between ct058 and ct059, 
and between ct192 and ct193. This indicated that ct058 is co-transcribed with ct059, which likely 
encodes a ferredoxin, and transcription of ct192 is unlinked from ct193 (Fig. 5A and B). We 
previously detected clear measurable mRNA levels of ct059 in C. trachomatis ocular and 
urogenital strains (7). We re-confirmed this, using the same biological samples in which the 
levels of ct058 mRNA were vestigial (data not shown). Therefore, we tentatively propose that 
expression of ct058 in ocular and urogenital strains might be down-regulated by specific 3’ to 5’ 
post-transcriptional processing of the ct059-ct058 transcript.

To precisely define the promoter region of ct059-ct058, ct192, and ct214, we determined 
their TSSs by RACE, using as template total RNA of cells infected with C. trachomatis L2/434
and primers complementary to the ct059, ct192, or ct214 mRNA (Fig. 5A and C-E). When using a primer complementary to the ct058 mRNA, we were unable to identify a ct058-exclusive TSS upstream from its start codon in L2/434 (data not shown). The TSS of ct214 matched the one previously identified by deep-sequencing in strain L2b/UCH-1 (2), while the TSSs of ct059-ct058 and ct192 have not been identified before.

*C. trachomatis* encodes three σ factors: σ66, the homolog of *E. coli* main σ factor (σ70); σ28, a minor σ factor; and σ54, an alternative σ factor (37). By inspecting the nucleotide sequences immediately upstream from the determined TSSs of ct059-ct058, ct192, and ct214 for σ66, σ54, and σ28-like promoters (37), we only identified σ66-like promoters (Fig. 5C-E and Fig. S3-S4). Finally, we analyzed the promoter regions of ct059-ct058, ct192, and ct214, and the ct059-ct058 transcript, for LGV-specific nucleotide differences, in all strains used in the RT-qPCR assays. We found LGV-specific nucleotides upstream from the predicted -35 region of ct059 (Fig. 5C and Fig. S3). Furthermore, we found 7 LGV specific nucleotides scattered within the coding region of ct059 and in between the stop codon of ct059 and the start codon of ct058 (Fig. 5C and Fig. S3). In particular, 3 of these LGV-specific nucleotide differences were clustered 30-35 nucleotides upstream from the start codon of ct058 (Fig. 5C and Fig. S3). However, it is unclear how these LGV-specific differences could explain the vestigial mRNA levels of ct058 in ocular and urogenital strains or have a discriminatory role in the proposed hypothetical degradation of the ct058 transcript in ocular and urogenital strains. The scenario was simpler for ct192 and ct214, as we identified discrete LGV-specific nucleotides within the promoter regions of these genes that may explain their disease group-specific expression (Fig. 5D and E, and Fig. S4).
We found that amino acid differences between Inc proteins and distinct mRNA levels among inc genes throughout the developmental cycle of C. trachomatis strains correlate with the specific invasiveness and tropism of LGV isolates. Thus, we propose the novel hypothesis that a subset of Inc proteins may contribute for the specificity of infection by LGV strains. In fact, the vast majority of amino acid differences between Inc genes are due to variations between proteins from LGV and trachoma biovars (Fig. 2). This could simply reflect the evolutionary history of C. trachomatis (25). However, most inc genes showed d_\text{N}/d_\text{S} values of > 1 among C. trachomatis strains and, according to the Z-test of selection, 10 incs are likely under positive selection. In contrast, pmps or selected housekeeping genes of C. trachomatis mostly showed d_\text{N}/d_\text{S} values of < 1 (Fig. 3). Moreover, polymorphisms in C. trachomatis genomes are essentially driven by fixation of silent mutations (8). This suggests that the amino acid differences between Inc proteins should not be solely explained by genetic drift. In addition, almost all disease group-specific amino acids of Inc genes encoded by genes likely under positive selection were found among proteins of LGV strains, and the majority of these amino acids localized in regions of the proteins predicted to face the host cell cytosol (Table 2). Yet, it must be clarified that in addition to Inc proteins other proteins are likely involved in the specificity of infection by LGV strains (8, 30). Moreover, the overall determinants of tissue tropism of C. trachomatis should be complex and multifactorial, and certainly also include e.g. the products of the trpRBA operon (10, 19, 51) or of the cytotoxin gene (11), and Tarp (34).

We have done a focused and in-depth study of the variability of Inc proteins and evolution and expression of inc genes among C. trachomatis serovars, encompassing almost all genomic information currently available. We reveal that the overall degree of variation in the
amino acid sequences of Inc proteins among strains is similar to that of a characteristic family of *C. trachomatis* polymorphic proteins (Pmps). Our results strengthen previous studies that suggested that some *C. trachomatis* Incs could contribute to tissue tropism (8, 9, 35, 43), and confirm recent data suggesting that many inc genes could be under positive selection (8, 30). However, almost all of these previous studies analyzed a limited number of sequences, as the majority of the 51 genomic sequences used in our work only became available very recently (25).

Our findings also revealed that differential gene expression could be a mechanism contributing to the different invasiveness and tissue tropism of *C. trachomatis* strains. A previous work have identified differences in expression between pmp genes from reference and clinical strains (42), but disease group-specific differences in expression of *C. trachomatis* genes have not been noticed before. The inc genes *ct058, ct192*, and *ct214* evidenced LGV-specific gene expression (Fig. 4), and we have further identified LGV-specific nucleotides in the promoter regions of *ct192* and *ct214* and within the *ct059-ct058* transcript (Fig. 5). This was directly analyzed for the strains used in RT-qPCR assays, but the specificity is maintained within all 51 *C. trachomatis* genomes (in Table S1) that we used in our studies (data not shown). As only 3 out of 48 inc genes showed LGV-specific gene expression, it is unlikely that this specificity is a common feature among *C. trachomatis* genes.

For *ct214*, we tentatively propose that the LGV-specific nucleotides could differentially affect its expression either at transcriptional or post-transcriptional level. For example, a recent study showed that a *Salmonella* small non-coding RNA can discriminate mRNA regions that differ in one single nucleotide (44). The picture is more complex for *ct058* and *ct192*. Regarding *ct058*, in L2/434 this gene is co-transcribed with *ct059* and mRNA levels of *ct059* were...
previously detected also in ocular and urogenital strains (7), which we re-confirmed (data not shown). We speculate that specific processing of the ct059-ct058 transcript selectively reduces the levels of ct058 mRNA in ocular and urogenital strains. Furthermore, ct058 is a pseudogene in many ocular strains (but not in those used in the RT-qPCR assays; Table S3 and data not shown) and ct192 is a pseudogene in at least ocular strain B/Jali20 (Table S3). This does not necessarily have an impact in the interpretation of the data, as for example the LGV-specific pseudogene ct300 showed similar mRNA levels during the developmental cycle of C/TW3, E/Bour or L2/434 (Table S6). Another issue is that there are different annotations in GenBank of the start codon of ct192 in the archetype ocular (A/Har13), urogenital (D/UW3) and LGV (L2/434) strains (Table S3). This alone could explain the differential expression of ct192, owing to dissimilar promoter regions. However, inspection of the nucleotide sequence immediately upstream from the annotated start codons of ct192 in each of these three strains only reveals a strong putative ribosome binding site in L2/434 (Table S3). Therefore, it is likely that the start codon of ct192 is conserved between C. trachomatis strains and corresponds to the one annotated in L2/434. In this situation, ct192 is a pseudogene in all ocular strains (Table S3) and the differences in the mRNA levels of ct192 between LGV and ocular or urogenital strains may be tentatively explained by the single nucleotide differences found within its promoter region (Fig. 5D and Fig. S4).

It is not possible to make a rigorous side-by-side comparison between our RT-qPCR data and previous analyses of inc gene expression by RT-PCR (52) and microarrays (5, 41), as the sensitivity of the methods used is quite different. In general, we have confirmed that inc genes display different profiles of expression but that they are mostly “early-cycle genes”, which supports the idea that many Incs should play a role in modifying the inclusion membrane early
We also found *inc* genes displaying an “early- and late-cycle” profile of expression. A similar profile has been observed in microarray studies (5), and it was suggested to result from “carry-over” mRNA of highly expressed late genes from the previous infectious cycle. However, the first time-point we analyzed was at 2 h post-infection, when “carry-over” mRNA would have been degraded. Furthermore, at 30 or 42 h post-infection, several “late-cycle” *inc* genes showed levels of expression at least comparable to those of “early- and late-cycle” *incs* (Table S6). This suggests that the expression of some *inc* genes could be induced early in the cycle, down-regulated at mid-cycle, and induced again at late-cycle.

Not all predicted *C. trachomatis* *incs* have been detected on the inclusion membrane. We addressed this by the analysis of T3S signals in 48 *C. trachomatis* *inc* proteins. This led to the identification of 19 novel T3S substrates, including 14 putative *incs* (Fig. 1 and Table 1). In particular, and importantly, putative *incs* CT058, CT192, and CT214 showed a T3S signal. Furthermore, all 10 *inc* genes likely under positive selection encode known *incs*. Our data for T3S signals in 14 *inc* proteins that were previously analyzed using *Shigella flexneri* as heterologous bacteria (16, 56) (Table 1) only differed for CT192 and CT850. This could be related to the distinct heterologous hosts used, or to the analysis of different N-terminal regions of CT192 (due to the different annotations of the start codon of *ct192* in GenBank).

Although human macrophages have strong antimicrobial activity against *C. trachomatis* ocular and urogenital strains, they support the growth of LGV strains (63). We hypothesize that specific amino acids in *inc* proteins, or their earlier and/or higher expression, could specifically enable LGV strains to inhibit phagolysosomal fusion in macrophages and/or prevent the formation of reactive oxygen or nitrogen species (22). Unfortunately, little is known about the function of the *inc* proteins that we have identified as potentially involved in these processes.
CT222 and CT118/IncG co-localize with kinases of the Src family in discrete regions of the
inclusion membrane that associate with host cell centrosomes (38), and CT223 could inhibit host
cell cytokinesis (3). However, it is unclear how this may relate to our hypothesis. On the other
hand, Incs that could manipulate intracellular membrane trafficking, such as those that have
SNARE-like motifs (CT119/IncA, CT223, and CT813) (17) or interact with Rab GTPases
(CT229) (46), are good candidates to selectively inhibit macrophage phagolysosomal fusion. Not
all \textit{C. trachomatis} Incs tentatively proposed to be specifically involved in inhibiting
phagolysosomal fusion have homologues in other chlamydial species, which also avoid this host
cell degradation pathway (16, 34). It is possible that each chlamydial species evolved particular
Incs, and other virulence proteins, to account for the specificity of each type of cell it infects.

ACKNOWLEDGEMENTS

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through
grants PEst-OE/EQB/LA0004/2011 and PTDC/SAU-MII/099623/2008, and by the European
Commission through a Marie Curie European Re-integration Grant (PERG03-GA-2008-230954)
to LJM. FA, VB, and RF hold PhD fellowships SFRH/BD/73545/2010, SFRH/BD/68527/2010,
and SFRH/BD/68532/2010, respectively, from FCT. We thank Irina Franco for critical review of
the manuscript.

REFERENCES

   FEMS Microbiol Rev \textbf{29}:949-959.


58. Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood,
evolutionary distance, and maximum parsimony methods. Molecular biology and
evolution 28:2731-2739.

Skipp, C. D. O'Connor, I. Goodhead, H. Norbertzeak, B. Harris, D. Ormond, R.
trachomatis: genome sequence analysis of lymphogranuloma venereum isolates. Genome

60. Unemo, M., H. M. Seth-Smith, L. T. Cutcliffe, R. J. Skilton, D. Barlow, D. Goulding,
K. Persson, S. R. Harris, A. Kelly, C. Bjartling, H. Fredlund, P. Olcen, N. R.
Thomson, and I. N. Clarke. 2010. The Swedish new variant of *Chlamydia trachomatis*:
genome sequence, morphology, cell tropism and phenotypic characterization.
Microbiology 156:1394-1404.


human mononuclear phagocytes against the human biovars of *Chlamydia trachomatis*. J
Immunol 139:1297-1302.
FIG. 1. Type III secretion (T3S) signals in *C. trachomatis* Inc proteins. *Y. enterocolitica* T3S-proficient (ΔHOPEMT) and T3S-defective (ΔHOPEMT ΔYscU) bacteria were used to analyse secretion of hybrid proteins comprising the first 20 amino acids of Inc proteins, or of *Y. enterocolitica* SycT, fused to the mature form of TEM-1 β-lactamase (TEM-1). (A) and (B) Immunoblots show the result of representative assays in which proteins in culture supernatants (S - secreted proteins), and in bacterial pellets (P - non-secreted proteins) from ~5x10^7 bacteria were loaded per lane. SycT and SycO are strictly cytosolic *Yersinia* T3S chaperones (28, 32). SycT20-TEM-1 was a negative control for the T3S assays. Immunodetection of SycO ensured that presence of TEM-1 hybrid proteins in culture supernatants was not a result of bacterial lysis or contamination. (C) The percentage (%) of secretion of each TEM-1 hybrid was calculated by densitometry, as the ratio between the amount of secreted and total protein. The threshold to decide whether a protein was secreted was set to 5% (dashed line), based on the % of secretion of SycT20-TEM-1. Data are the mean ± SEM from at least 3 independent experiments. Please note that in this work we did not analyse all described putative Inc proteins (see legend of Table 1) and that T3S signals were not analysed for all known Inc proteins.

FIG. 2. Polymorphisms in *C. trachomatis* Inc proteins. Polymorphic membrane proteins (Pmps) and housekeeping proteins (HKs) were analyzed as reference. (A) Overall mean genetic distance (polymorphisms) based on the p-distance between all possible pairs of amino acid (aa) sequences of Inc proteins, Pmps, and HKs among *C. trachomatis* strains. Proteins marked with an asterisk (*) have an aa p-distance that is equal or higher than the average value for Inc proteins (0.017; shown as a
dashed line). (B) Average mean genetic distance, based on the $p$-distance between all possible pairs of aa sequences within (Ocular, urogenital, or LGV; OC, UROG, or LGV, respectively) or between (OC-UROG, OC-LGV, or UROG-LGV) *C. trachomatis* disease groups. (C) Venn diagrams showing the phylogenetic segregation of *C. trachomatis* disease groups based on Neighbor-Joining trees of Incs, Pmps or HKs and on pairwise aa $p$-distances between and within disease groups, for all possible pairs of Incs, Pmps or HKs sequences from the *C. trachomatis* strains analysed. All these analyses were performed by Bootstrap with 1,000 replicates. The error bars represent SEM.

FIG. 3. Evolutionary dynamics of *inc* genes. Genes encoding polymorphic membrane proteins (*pmps*) or housekeeping (HK) proteins were used as reference. (A) Ratio of non-synonymous ($d_N$) to synonymous ($d_S$) substitutions among *C. trachomatis* strains. The dashed line indicates neutrality ($d_N/d_S = 1$). The arrows specify genes likely under positive selection, according to the codon based Z-test of selection (see Materials and Methods). (B) Distribution of $d_N$ vs $d_S$ values for the 10 *inc* genes likely under positive selection, comparing the impact of artificially discarding ocular, urogenital, or LGV strains relative to the analysis with all *C. trachomatis* strains. The straight line in each graph indicates neutrality ($d_N/d_S = 1$). In all cases, *inc* genes likely under positive selection (codon based Z-test of selection) are depicted as black circles, while *inc* genes for which statistical support of likely positive selection can no longer be detected after discarding a particular group of strains are depicted as white circles. All these analyses were performed by Bootstrap with 1,000 replicates. For sake of clarity, the SEM values of both (A) and (B) are only presented in Table S5.
FIG. 4. mRNA levels of *inc* genes during the developmental cycle of different *C. trachomatis* strains. The mRNA levels of 48 *inc* genes (A and B), or of *ct058, ct192*, and *ct214* (C), were analyzed by RT-qPCR throughout the developmental cycle of the indicated prototype (B/Har36, C/TW3, E/Bour, L2/434, and L3/404) and clinical (F/CS465-95, and L2b/CS19-08) strains. (A) Peak of expression (highest mRNA levels during the developmental cycle) of each *inc* gene. The *p* values were calculated by two-tailed *t*-tests. (B) Number of *inc* genes showing the indicated profiles of expression (variation of mRNA levels during the developmental cycle). (C) The expression values (mean ± SEM) result from the RT-qPCR raw data (x10^5) of each gene normalized against the *16SrRNA*, from at least two independent experiments. Complete data is shown in Table S6.

FIG. 5. Identification of LGV-specific nucleotides in the promoter regions of *ct192* and *ct214* and within the *ct059-ct058* transcript. (A) Genetic organization of *ct058, ct192*, and *ct214* (the nomenclature of D/UW3 is used) depicting the fragments amplified in the transcriptional linkage analysis and the approximate localization of the transcriptional start sites (TSSs) determined by RACE in L2/434. (B) Transcriptional linkage analysis in L2/434: gDNA +, PCR from total DNA isolated from cells infected with L2/434; cDNA +, PCR from cDNA generated with reverse transcriptase (RT) from total RNA isolated from cells infected with L2/434; cDNA −, as cDNA+ but without RT; gDNA −, PCR from DNA of uninfected cells. (C to E) Schematic view of the nucleotide sequences of the *ct059* promoter region, *ct059-ct058* intragenic region, and first codons of *ct058* (C), and of the promoter regions of *ct192* (D) and *ct214* (E), in L2/434. The TSSs are labelled +1 and the predicted -10 and -35 σ66-like hexamers are underlined. The
sequences of the 3 LGV strains used in RT-qPCR assays (Fig. 5) are 100% identical within the depicted regions (Fig. S3 and S4). The identified LGV-specific nucleotides are indicated with an arrow and below are the nucleotides present in those positions in the ocular and urogenital strains used in RT-qPCR assays (Fig. 5). The full nucleotide sequences of these regions in the strains used in RT-qPCR assays are shown in Fig. S3 and S4.
TABLE 1. Summary of type III secretion (T3S) signals found in known and putative Incs of *C. trachomatis* analysed in this work

<table>
<thead>
<tr>
<th>Inc protein*</th>
<th>T3S signal</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Known Incs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT101</td>
<td>No</td>
<td>This work</td>
</tr>
<tr>
<td>CT115/IncD</td>
<td>Yes</td>
<td>(56); this work</td>
</tr>
<tr>
<td>CT116/IncE</td>
<td>Yes</td>
<td>(56)</td>
</tr>
<tr>
<td>CT117/IncF</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT118/IncG</td>
<td>Yes</td>
<td>(56)</td>
</tr>
<tr>
<td>CT119/IncA</td>
<td>Yes</td>
<td>(56); this work</td>
</tr>
<tr>
<td>CT147</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CT222</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT223</td>
<td>Yes</td>
<td>(56)</td>
</tr>
<tr>
<td>CT225</td>
<td>No</td>
<td>This work</td>
</tr>
<tr>
<td>CT226</td>
<td>Yes</td>
<td>(16)</td>
</tr>
<tr>
<td>CT228</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT229</td>
<td>Yes</td>
<td>(56)</td>
</tr>
<tr>
<td>CT232/IncB</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT233/IncC</td>
<td>Yes</td>
<td>(21, 56); this work</td>
</tr>
<tr>
<td>CT249</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT288</td>
<td>Yes</td>
<td>(56)</td>
</tr>
<tr>
<td>CT358</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT440</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT442</td>
<td>Yes</td>
<td>(56)</td>
</tr>
<tr>
<td>CT618</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT813</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT850</td>
<td>Yes/No</td>
<td>(16); this work</td>
</tr>
<tr>
<td><strong>Putative Incs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT005</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT006</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CT036</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT058</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT134</td>
<td>No</td>
<td>This work</td>
</tr>
<tr>
<td>CT135</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT164</td>
<td>No</td>
<td>This work</td>
</tr>
<tr>
<td>CT179</td>
<td>No</td>
<td>This work</td>
</tr>
<tr>
<td>CT192</td>
<td>No/Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT195</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT196</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT214</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT224</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT227</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT300</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>Inc protein(^a)</td>
<td>T3S signal</td>
<td>Reference</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Putative Inc(^c)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT345</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT357</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT365</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT383</td>
<td>Yes</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT449</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT483</td>
<td>Yes</td>
<td>This work</td>
</tr>
<tr>
<td>CT484</td>
<td>No</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT565</td>
<td>No</td>
<td>(16); this work</td>
</tr>
<tr>
<td>CT728</td>
<td>No</td>
<td>This work</td>
</tr>
<tr>
<td>CT789</td>
<td>Yes</td>
<td>This work</td>
</tr>
</tbody>
</table>

\(^a\)Proteins containing a bilobal hydrophobic motif that were analyzed by Li et al (33), which we selected to study in this work. We did not consider proteins (Cap1, and CopN) that localize to the inclusion membrane but which do not possess the bilobal hydrophobic domain (20, 23). More recent bioinformatics-based analyses identified additional putative Inc\(s\) (CT018, CT079, CT081, CT244, CT324, CT326, CT556, CT578, CT616, CT618, CT642, CT645, CT788, CT789, CT814.1, CT819, CT837, CT846, and CT873) in \(C.\) trachomatis (16, 35), but these proteins were not analyzed in our study.

\(^b\)Known Inc\(s\): \(C.\) trachomatis proteins containing a bilobal hydrophobic motif and that have been localized to the inclusion membrane by immunofluorescence microscopy using specific antibodies [(16, 33, 38), and references therein].

\(^c\)Putative Inc\(s\): \(C.\) trachomatis proteins containing a bilobal hydrophobic motif but which have not yet been localized to the inclusion membrane.

\(^d\)Conflicting data between our observations and previous analyses using \(S.\) flexneri as a heterologous bacterial host.
TABLE 2. Amino acids residues within Inc proteins encoded by genes likely under positive selection that are specific of *C. trachomatis* disease groups

<table>
<thead>
<tr>
<th>Inc</th>
<th>Length (aa)a</th>
<th>TM segmentsb</th>
<th>LGVc</th>
<th>Urogenitalc</th>
<th>Ocularc</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT116/IncE</td>
<td>132</td>
<td>[36-59],</td>
<td>V37A, V38A,</td>
<td>D19G</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[64-87]</td>
<td>S39C, S64G,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT118/IncG</td>
<td>167</td>
<td>[33-57],</td>
<td>A33V, F77C,</td>
<td>D142G</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[63-88]</td>
<td>C79Y, N87S,</td>
<td></td>
<td>V161M</td>
</tr>
<tr>
<td>CT119/IncA</td>
<td>273</td>
<td>[35-59],</td>
<td>I75T, Q207E,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[64-84]</td>
<td>V211A,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V212A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT222</td>
<td>129</td>
<td>[39-63],</td>
<td>Y10C, 175,</td>
<td>A40V,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[69-93]</td>
<td>C89Y, L125L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V126S, Y127Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT223</td>
<td>270</td>
<td>[38-61],</td>
<td>G11R, A65T,</td>
<td>A75V,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[67-91]</td>
<td>K100R, I108L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>K127G, N130D</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P134R, C153Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E160R,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T166K, H200Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E204D,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N207R, L208M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R231L,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V241A, P251L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D261V,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G263F,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT228</td>
<td>196</td>
<td>[38-59],</td>
<td>I22T, A68V,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[65-86]</td>
<td>A72P, C90Y,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V119A, A143V</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I144M,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V146F,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT229</td>
<td>215</td>
<td>[42-65],</td>
<td>A39S, V99L,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[71-90]</td>
<td>G112E, K114E</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>F121S, Q125R</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V126A, H137Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Q144K, Y152H</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E154A, E164K</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G173R, S181N</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T201A,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H8Y, D24N,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V80T, A89I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT249</td>
<td>116</td>
<td>[51-72],</td>
<td>H8Y, D24N,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[78-97]</td>
<td>V80T, A89I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT288</td>
<td>563</td>
<td>[36-58],</td>
<td>A67T, I69V,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[65-88],</td>
<td>E102A, S121P</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[242-263],</td>
<td>A194T, S198T</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[269-291]</td>
<td>N203K, T209L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A258T, I260V</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I261V, L318W</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S346G, L385Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-47D, A516P</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V531L, A532T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT813</td>
<td>264</td>
<td>[41-61],</td>
<td>N7T, V76I,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[68-94]</td>
<td>E107K, K125E</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R128Q, T145A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E161K, A165G</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E168K, E171K</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E172Q, H181V,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aThe amino acid (aa) lengths and positions are based on the protein sequence annotation of the *C. trachomatis* strain D/UW3.
bThe numbers under brackets indicate the position of transmembrane (TM) domains within the aa sequence of the corresponding Inc, obtained from (16), except for CT223 where we only found the two indicated TM domains.
cThe aa specific of Inc proteins from each disease group are indicated in bold, relative to the aa in the same position in the other disease groups, and those in regions predicted to be on the cytoplasmic side of the inclusion membrane are underlined (we considered that the loop region within two TM segments faces the lumen of the inclusion); “-“ indicates a deletion or an insertion.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5