Rita Susana da Silva Ferreira

Mestre em Biologia Molecular e Humana

Molecular dynamics of *Chlamydia trachomatis* chromosome and plasmid

Dissertação para obtenção do Grau de Doutor em Biologia

Orientador: Doutor João Paulo Gomes, Investigador Auxiliar com Habilitação, Instituto Nacional de Saúde Dr. Ricardo Jorge

Co-orientador: Doutora Maria José Borrego, Investigadora Auxiliar, Instituto Nacional de Saúde Dr. Ricardo Jorge

Co-orientador: Doutora Ana Madalena Ludovice, Professora Auxiliar, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Júri:

Presidente: Presidente do Conselho Científico Prof. Doutor Fernando José Pires Santana.

Vogais: Doutora Maria Aida da Costa e Silva da Conceição Duarte, Professora Associada, com Agregação, da Faculdade de Farmácia da Universidade de Lisboa;

Doutor Luís Jaime Gomes Ferreira da Silva Mota, Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa;

Doutor João Paulo dos Santos Gomes, Investigador Auxiliar com Habilitação do Instituto Nacional de Saúde Dr. Ricardo Jorge;

Doutora Mónica Alexandra de Sousa Oleastro, Investigadora Auxiliar do Instituto Nacional de Saúde Dr. Ricardo Jorge.

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Agosto de 2016

Rita Susana da Silva Ferreira
Mestre em Biologia Molecular e Humana

Molecular dynamics of *Chlamydia trachomatis* chromosome and plasmid

"To give anything less than your best is to sacrifice the gift”

– Steve Prefontaine

Agradecimentos

A todos os que, de uma forma directa ou indirecta, me apoiaram, aconselharam e incentivaram na concretização do trabalho que culminou nesta Tese de Doutoramento, etapa extremamente importante da minha vida profissional, venho aqui expressar meu mais sincero agradecimento.

Manifesto um especial reconhecimento,

ao meu orientador Doutor João Paulo Gomes, a excelente orientação, toda a motivação e incentivo científico que me dedicou, e que tornaram os trabalhos que desenvolvi tão mais ricos e precisos. O amor e dedicação pela ciência que transparece, foi sempre para mim um motivo de inspiração. Obrigada!

à minha co-orientadora Doutora Maria José Borrego, a orientação, a precisão experimental e linguística que me ajudaram a crescer a nível profissional, bem como a motivação, as imensas palavras de incentivo e a boa disposição.

A todos os restantes membros da Comissão de Acompanhamento de Tese (CAT), pela discussão científica e por todas as sugestões concedidas. Em especial, à minha co-orientadora Dra. Ana Madalena Ludovice, Professora Auxiliar da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT/UNL) por ter aceitado ser minha orientadora, pela simpatia e disponibilidade; e ao Dr. Jaime Mota, Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT/UNL) pela colaboração científica, disponibilidade e simpatia. Agradeço, ainda, à Professora Marta Aires de Sousa, Professora Coordenadora da Escola Superior de Saúde da Cruz Vermelha Portuguesa, (ESSCVP).

ao meu colega Vítor Borges, a sua paciência inesgotável para comigo, o seu inestimável apoio em diversos ensaios experimentais e todas as suas críticas tão constructivas e oportunas que tantas vezes auxiliaram a fluência e a precisão dos trabalhos em curso. Muito obrigada Vítor, pelo teu profissionalismo, companheirismo e boa disposição!

à minha colega Alexandra Nunes, que implementou várias metodologias de análise bioinformática e que dedicou algum do seu tempo a transmitir-me parte desse conhecimento para que eu própria as pudesse realizar de forma independente. Agradeço-lhe também toda a disponibilidade demonstrada, quer na revisão dos vários trabalhos escritos, bem como em tudo o que eu fui necessitando durante a progressão dos trabalhos em curso.

Agradecimentos

a todos os meus restantes colegas de investigação do INSA, em particular à Vera Damião, pelo companheirismo, simpatia e amizade, e à Minia Antelo e ao Miguel Pinto a colaboração, a entreajuda e a boa disposição.

a todas as colegas do laboratório Dora Cordeiro, Inês João e Margarida Dinis a simpatia, disponibilidade, apoio e companheirismo.

a toda a equipa da Unidade de Tecnologia e Inovação do INSA. Em especial, ao seu responsável, Dr. Luís Vieira, pelo seu empenho na implementação e optimização das metodologias de Sequenciação de Nova Geração, em particular de RNA, as quais foram tão relevantes na fase final deste meu trabalho.

ao Instituto Nacional de Saúde Dr. Ricardo Jorge, em particular ao Departamento de Doenças Infecciosas, na pessoa do seu coordenador, Dr. Jorge Machado.

à Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT/UNL), em particular à Professora Isabel Sá Nogueira, Coordenadora do Programa Doutoral de Biologia.

à Fundação para a Ciência e Tecnologia, Ministério da Educação e Ciência (FCT/MEC), pelo financiamento concedido através da Bolsa Individual de Doutoramento (SFRH/BD/68532/2010).

Por fim, quero agradecer a todos os meus familiares e amigos que ocupam um lugar muito especial na minha vida e sem os quais a concretização desta ou de qualquer outra etapa na minha vida teria sido absolutamente impossível. São o meu pilar, o meu conforto, a minha energia e amo-vos profundamente por isso!

A todos, o meu mais sincero Obrigado!

Notes of the Author

The work described in this Ph.D. thesis was carried out at the National Reference Laboratory of Sexually Transmitted Infections, Department of Infectious Diseases, of the National Institute of Health, Lisbon, Portugal.

It encompasses experimental studies performed throughout the author Ph.D. project, which resulted in four independent manuscripts, which are presented here as individual chapters (Chapter II-V). Three of them have been published in international journals (with *peer review*), and one is in the process of publication. The order of presentation of these chapters does not reflect the chronological order of their publication. As they focus on the study of *Chlamydia trachomatis* (*C. trachomatis*) chromosome, plasmid and both, the order of their presentation in this thesis respects that same order of contents. Also, all chapters were formatted in a cohesive style, since they all are reproductions of individual manuscripts with different layouts. References where cited by sequential numbers, according to the order of their citation in the text, and listed in a single "Reference" section. Of note, all abbreviations were also uniformized. Finally, supplemental data of each chapter is also compiled in a single section designated by "Supplemental Material".

As each chapter contains a specific background and a detailed discussion of the results, this thesis contains only a succinct overview of the *C. trachomatis* bacterium biology (Chapter I) and a final overview of the main findings and conclusions, with regards to future perspectives and lines of work (Chapter VI).

Resumo

A bactéria *Chlamydia trachomatis* infeta exclusivamente o Homem, e pode ser classificada em 15 serótipos principais (A-L3). Estes infectam preferencialmente três locais anatómicos distintos: a conjunctiva ocular (A-C), os órgãos genitais (D-K) e os nódulos linfáticos (L1-L3). Embora apresentem <2% de variabilidade genética, desconhecem-se os mecanismos moleculares que estão na base deste tropismo diverso. O conhecimento do impacto da escassa variabilidade na patobiologia associada aos serótipos poderá culminar em formas específicas de tratamento ou profilaxia.

Utilizando a genómica, a transcriptómica e a bioinformática procedemos à análise do cromossoma e do plasmídeo de *C. trachomatis* para decifrar as dinâmicas moleculares que possam estar na base das diferenças fenotípicas dos seus serótipos.

Os nossos resultados revelaram que esta bactéria apresenta taxas de recombinação e de mutação compatíveis com a sua natureza enquanto microorganismo intracelular obrigatório e reforçam a relevância da variabilidade genética dos seus serótipos, nos diferentes fenótipos que lhes estão associados (tropismo, doença e sucesso ecológico). Encontrámos polimorfismos ao nível dos genes e das proteínas, específicos dos serótipos que causam o mesmo tipo de doença, em particular dos que causam o linfogranuloma venéreo. Observámos que o plasmídeo de *C. trachomatis* expressa preferencialmente dois genes e os seus dois RNAs "anti-sense". Verificámos ainda que vários dos genes mais expressos e mRNAs mais estáveis são comuns aos diferentes serótipos.

No geral, os resultados alcançados nesta tese revelaram aspectos específicos dos serótipos de *C. trachomatis*, essencialmente ao nível da sua variabilidade genética e das dinâmicas transcricionais, contribuindo para o esclarecimento da patobiologia associada de cada um deles. Os resultados reforçam ainda a necessidade de clarificar a função de proteínas que consideramos serem fundamentais às dissimilaridades das estirpes de *C. trachomatis*, em particular dos efectores do sistema de secreção tipo III, das proteínas da membrana da inclusão e das proteínas com função desconhecida.

Palavras-chave

Chlamydia trachomatis; Genómica; Transcriptómica; Bioinformática; Plasmídeo; Cromossoma.

Abstract

Chlamydia trachomatis is a strict human pathogenic bacterium, whose strains may be classified into 15 major serovars (A-L3). They preferentially infect three distinct anatomic locations: the ocular conjuntciva (A-C), the genitalia (D-K) and the lymph nodes (L1-L3). Although the serovars only display <2% genetic variability, the molecular mechanisms by which they proliferate in such different niches are still to be elucidated. Hence, understanding how the scarce variability shapes serovars' pathobiology will certainly grant ways of directing treatment or even developing prophylactic measures.

By using genomics, transcriptomics and bioinformatics we scrutinized the chromosome and the plasmid of *C. trachomatis* in order to decipher the molecular dynamics that could underlie the phenotypic differences displayed by the serovars.

Our results revealed that this bacterium shows recombination and mutation rates concordant to its obligatory intracellular nature and reinforce the relevance of such genetic variability among *C. trachomatis* serovars in the dissimilar phenotypes they display, namely, cell-appetence, disease outcome and ecological success. We found polymorphisms, at both gene and protein level, that were specific to the same disease-causing serovars, in particular to those causing the lymphogranuloma venereum. We observed that the *C. trachomatis* plasmid preferentially expresses two of its eight genes and the two anti-sense RNAs. We also found that different-serovar strains share several of the highest expressed genes and several of the most stable mRNAs.

Overall, the findings achieved throught this thesis revealed some specific features of *C. trachomatis* serovars, mainly regarding genetic variability and transcriptional dynamics, and hence contributing for the clarification of the serovars' pathobiology. They also emphasize the need for the complete clarification of the function of proteins which we believe to play a crucial role in the dissimilar phenotypes displayed by *C. trachomatis* strains, namely, type III secretion system effectors, inclusion membrane proteins and proteins with unknown function.

Keywords

Chlamydia trachomatis; Genomics; Transcriptomics; Bioinformatics.

Table of Contents

Table of Contents

Figure Index

Figure Index

Table Index

Abbreviations

Throughout this Ph.D. thesis, acronyms are expanded upon first usage and whenever believed necessary to improve reading clarity.

A260	Absorbance acquired at 260 nanometers
ATP	Adenosine Tri-Phosphate
bp	Base pair
CO_2	Carbon dioxide
°C	Celsius degree
CPU	Central Processing Unit
CDS	Coding Sequence
cDNA	Complementary Deoxyriboucleic Acid
CI	Confidence Interval
cm^2	Square centimeter
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide
dt	Doubling time
dN	Nonsynonymous substitutions *per* nonsynonymous sites
dS	Synonymous substitutions *per* synonymous sites
EB	Elementary body
ER	Endoplasmic reticulum
FCT/UNL	Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
FPKM	Fragments *per* Kilobase of CDS per Million mapped reads
FCT/MEC	Fundação para a Ciência e a Tecnologia, Ministério da Educação e Ciência
g	Gram
h	Hour
HK	Housekeeping gene
HIV	Human Immunodeficiency Virus
HP	Hypothetical protein
H_2O	Water
Inc	Inclusion membrane protein
IDO	Indolamine 2,3-Dioxygenase
Indel	Insertion/deletion
IFN-γ	Interferon-γ
IGR	Intergenic Region
Kb	Kilobase

Abbreviations

kDa	Kilo Dalton
K2P	Kimura 2-parameters
Log	Logarithm
LGV	Lymphogranuloma venereum
$MgCl_2$	Magnesium chloride
MOMP	Major Outer Membrane Protein
Mbp	Mega base
mol	Mole
MEGA	Molecular Evolutionary Genetics Analysis
MACPF	Membrane Attack Complex/Perforin
MSM	Men who have Sex with Men
mRNA	Messenger Ribonucleic Acid
mL	Mililitre
mM	Milimolar
M	Million
mya	Million Years Ago
MEM	Minimum essential medium
min	Minute
MW	Molecular Weight
MLST	Multi-Locus Sequence Typing
nm	Nanometre
nM	Nanomolar
NJ	Neighbour-Joining
nt	Nucleotide
No.	Number
ORF	Open Reading Frame
Ori	Origin of replication
PBS	Phosphate-Buffered Saline
PLD	Phospholipase D endonuclease
PZ	Plasticity Zone
PCR	Polymerase Chain Reaction
Pmp	Polymorphic membrane protein
PSG	Positively Selected Gene
pi	Post-infection
PDI	Protein Dissulfide Isomerase
qPCR	Quantitative real-time PCR

R^2	Correlation coefficient
RNA-seq	(High throughput) RNA sequencing
R	Recombination rate
RB	Reticulate Body
RT	Reverse Transcription
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
rpm	Rotations *per* minute
RS	Sequence whose size was used for reference purposes
s	Second
Serovar	Serological variant
STI	Sexually Transmitted Infection
SNP	Single Nucleotide Polymorphism
sRNA	Small anti-sense RNA
SD	Standard Deviation
SE	Standard Error
$t_{1/2}$	Half-life Time
TSS	Transcription Start Site
ts	Transition
TARP	Translocated Actin-Recruiting Protein
tv	Transversion
TE	Tris-Ethylenediaminetetraacetic acid
trpRBA	Tryptophan biosynthesis operon
T3SS	Type III Secretion System
U	Kunitz Unit
UK	United Kingdom
USA	United States of America
WHO	World Health Organization
µg	Microgram
µL	Microlitre
µm	Micrometer
µM	Micromolar
δ	Average tract length of a recombination event
θ	Mutation rate
ψ	Putative pseudogene
ν	Rate of new polymorphism introduced by recombination

Thesis Outline

This Ph.D. thesis is divided into several chapters, encompassing the following contents:

Chapter I: a general introduction that gives an overview of the bacterium *C. trachomatis* aiming to contextualize and substantiate the relevance of the studies performed throughout the thesis. It starts with a very succinct historical background and progresses with several phylogenetic, clinical and biological aspects of *C. trachomatis*. Finally, the main objectives of this Ph.D. thesis are presented in a more detailed and contextualized manner.

Capter II: a study designed to determine the mutation and the recombination rates of the *C. trachomatis* chromosome, and simultaneously evaluate how specific groups of genes affect those bioinformatic estimates. It should be noted that, at the time this study was designed and carried out, the number of *C. trachomatis* whole-genome sequences were limited. This chapter reproduces the contents of the publication: Ferreira R, Borges V, Nunes A, Nogueira PJ, Borrego MJ, Gomes JP. Impact of loci nature on estimating recombination and mutation rates in *Chlamydia trachomatis*. G3 (Bethesda). 2012 Jul;2(7):761-8. doi: 10.1534/g3.112.002923.

Chapter III: like the previous chapter, this was also a study foccused on the chromosome of *C. trachomatis*. This unprecedented study, intended to evaluate the genetic diversity of all ~900 chromosome-encoded genes and to look for a potencial polymorphism pattern (at both gene and protein levels) that could contribute to a particular clinical outcome of the strains. This chapter reproduces the contents of the publication: Ferreira R, Antelo M, Nunes A, Borges V, Damião V, Borrego MJ, Gomes JP. 2014. *In silico* scrutiny of genes revealing phylogenetic congruence with clinical prevalence or tropism properties of *Chlamydia trachomatis* strains. G3 (Bethesda); 5(1):9-19. doi: 10.1534/g3.114.015354.

Chapter IV: a study focused entirely on the *C. trachomatis* single plasmid, mainly regarding the clarification of its copy number and the assessment of the transcription profiles of each plasmid-encoded ORF and of the two known plasmid anti-sense RNAs. We also attempted to check for a potential correlation between each of those observations and the dissimilar tropism of different-serovar strains. This chapter reproduces the contents of the publication: Ferreira R, Borges V, Nunes A, Borrego MJ, Gomes JP. Assessment of the load and transcriptional dynamics of *Chlamydia trachomatis* plasmid according to strains' tissue tropism. Microbiol Res. 2013 Jul 19;168(6):333-9. doi: 10.1016/j.micres.2013.02.001.

Chapter V: a study aiming the evaluation of the stability and abundance of all *C. trachomatis* RNAs, encoded by both its chromosome and its plasmid, by using a novel approach, the high

troughput RNA-sequencing methodology. By comparing the abundance and stability of the transcripts observed among different-biovar strains, we intended to clarify their importance on disease outcome and/or tissue tropism. This chapter reproduces the contents of the paper (in the submition process): Ferreira R, Borges V, Borrego MJ, Gomes JP. Global survey of mRNA levels and decay rates in the two biovars of the obligate intracellular *C. trachomatis*.

Chapter VI: a global overview of the subjects addressed throughout the previous chapters, where the main results and the conclusions achieved in this Ph.D. thesis are highlighted. It also includes possible subsequent research approaches for the study of the intracellular pathogen *C. trachomatis*, enabled by the emergence of novel methodologies, and also from the outputs and the questions raised during the course of the studies performed throughout this Ph.D. thesis.

Chapter I

General Introduction

Chapter I

1. General Introduction

1.1. Historical background

While working with Giemsa-stained conjunctival scrapings from trachoma cases, Halberstaedter and von Prowazek described for the first time, in 1907 [1], the causative agent of this pathology as being a "mantled protozoan", for referring cytoplasmatic vesicles full of those microorganisms. Chlamydozoa, the designation attributed to this pathologic agent, derives from the greek word "Chlamys/Khlamus", meaning mantle. From then, similar vesicles were found to be associated with several other diseases, like urethritis, cervicitis, conjunctivitis and also in lymphogranuloma venereum (LGV) cases [2,3]. Around 1930, during the worldwide pandemic of an atypical and acute pneumonia, resultant from the contact with psittacine birds (parrots), similar microorganisms were found in samples collected from both infected birds and humans [4-6]. In 1935, Miyagawa and collegues [7] misconsidered the ethiological agents of the psittacosis-LGV group as viruses because they could be passed through bacterial filters and were unable to grow on artificial media. Only later, with the advent of the electron microscopy, Chlamydiae were classified as bacteria because they were found to possess DNA and RNA, ribossomes and a cell wall resembling that of Gram-negative bacteria [8]. Currently, these bacteria are known as *Chlamydia*, a misnomer derived from their first designation, back in 1907.

1.2. Taxonomy and phylogeny

Chlamydiae is the term used to designate the members of the order Chlamydiales, which are bacteria characterized by their obligate growth within eukaryotic cells, distinct from other bacteria at both the phylogenetic (Figure 1.1) and the phenotypic level [9]. By using the *16SrRNA* as time-scale calibrator, it was estimated that the evolutionary divergence of Chlamydiales and the *Parachlamydia amoebophila* (a *Chlamydia*-related endosymbiont of free-living amoebas) from their commom ancestor occurred at about 700 mya [10,11].

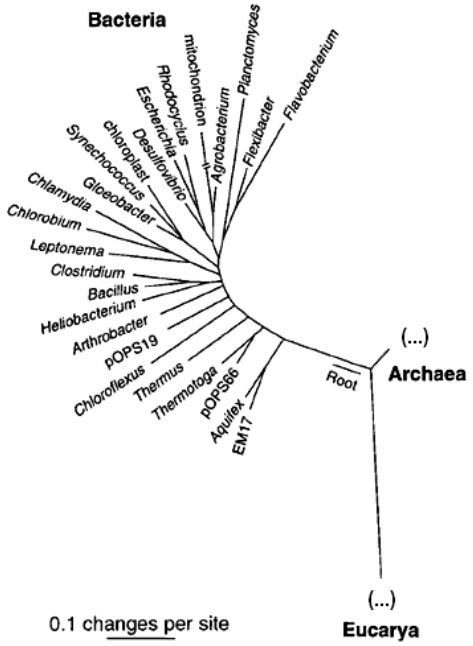


Figure 1.1. Universal phylogenetic tree, based on small-subunit rRNA sequences, with particular focus on the Bacteria branching. The original tree [9] was constructed with 64 rRNA sequences representative of the phylogenetic diversity of the three domains (Bacteria, Archaea and Eucarya), and genetic differences between pairs of sequences were considered to be the measure of evolutionary distance. For simplification purposes, the branches of both Archaea and Eucarya were cut where the respective radiation started within each domain.

Since Chlamydiae were recognized as bacteria, several attempts were made to define the genus. Ultimately, the single order Chlamydiales was proposed to be composed of a single family, Chlamydiaceae, with only one genus, *Chlamydia*. However, in 1999, Everett and colleagues [12] proposed the existence of two separate genera, the genus *Chlamydia* and the genus *Chlamydophila*. The former was composed by three species (*C. trachomatis*, *C. muridarum* and *C. suis*), whereas the latter was composed of six species (*C. abortus*, *C. psittaci*, *C. caviae*, *C. felis*, *C. pneumonia* and *C. pecorum*). Nevertheless, this very recent classification was not accepted by the scientific community [13], as it was only based on the similarity degree of the 16SrRNA encoding gene (threshold > 95%) for clustering species in the same genus. This was not considered the correct criteria because: *i*) despite the value of the *16SrRNA* in evolutionary studies, its sole use in speciation studies may not be discriminatory enough, and other molecular markers should also be included in those analyses [13]; and *ii*) species from *Chlamydia* and *Chlamydophila* genera often share ~97% of *16SrRNA* sequence similarity [14]. Moreover, some genomic features (chromosome distribution of the protein encoding genes, the similarity of the plasmid sequence and de *Chlamydia*-specific indel events) [15,16], together with the unique and highly conserved biology shared by these organisms, are not recognized when they are grouped into separate genera, also contradicting the applicability of this taxonomic classification.

With the continuous accumulation of genomic data, in 2009 [14] it was proposed that Chlamydiacea members should be regrouped into the former genus *Chlamydia*, composed of all the nine species, even though they exhibit major differences in host range (human and animal), tissue tropism, and disease pathology (reviewed in detail in [17]), and this makes the currently accepted taxonomy (Figure 1.2): *C. trachomatis*, *C. muridarum*, *C. suis*, *C. abortus*, *C. psittaci*, *C. caviae*, *C. felis*, *C. pneumonia* and *C. pecorum*. It is worth noting that, although *C. muridarum* is a mouse pneumonia agent, due to the high phylogenetic relatedness with *C. trachomatis* (~82% of chromosome homology) [18,19] it has been often used to model disease caused by the latter (e.g. [20-23]).

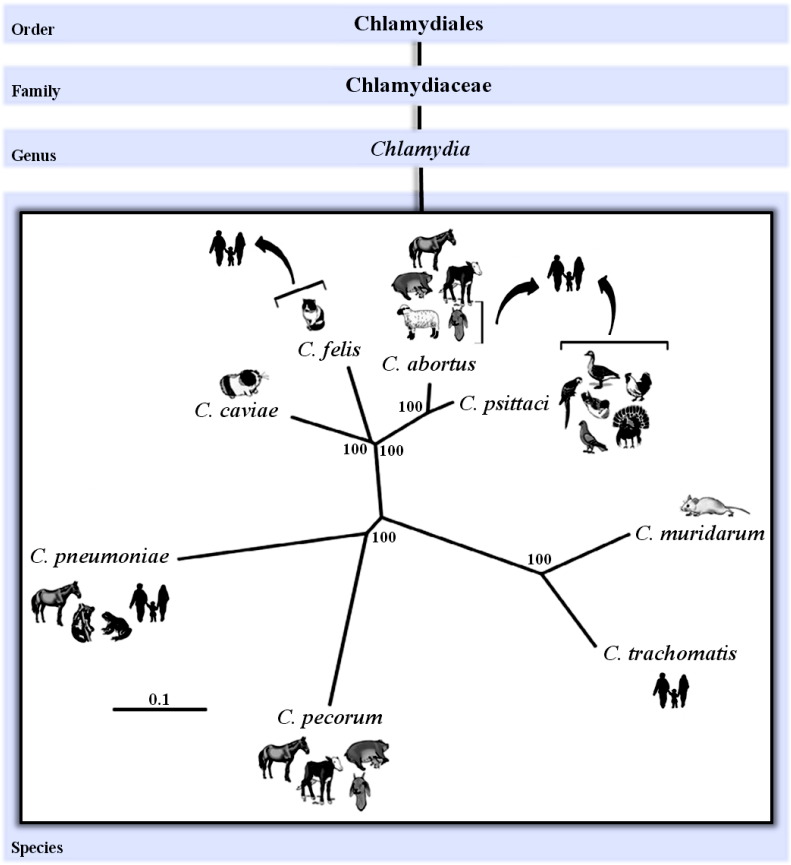


Figure 1.2. Current Chlamydiae taxonomy and phylogenetic reconstruction of the species from the genus *Chlamydia*. Taxonomy of Chlamydiae is structured in one order (Chlamydiales), one family (Chlamydiaceae), one genus (*Chlamydia*) and nine species (*C. trachomatis*, *C. muridarum*, *C. pneumonia*, *C. suis*, *C. psittaci*, *C. abortus*, *C. felis*, *C. caviae* and *C. pecorum*). Phylogenetic relation of eight species of *Chlamydia* genus is represented in the "species pannel" (bottom). Natural hosts are shown for each chlamydial species, and confirmed zoonotic transmission to humans is indicated in brackets (*C. psittaci*, *C. abortus* and *C. felis*). *C. suis* is not represented because there is still a lack of genomic data for this species. Adapted from [17].

Upon the emergence of modern human lineages at about 6 mya [24], one may suppose that the strict human pathogen *C. trachomatis*, had diverged from the remainder Chlamydiaceae somewhere

during this period of time. From then, this well adapted intracellular pathogen had also undergone a process of radiation, with different strains displaying differences in tissue tropism, disease outcome and prevalence.

The strains of *C. trachomatis* (Figure 1.3) may be classified into 15 major variants (serovars) – A-K, L1, L2 and L3 – based on the serological reactivity of monoclonal antibodies directed to the Major Outer Membrane Protein (MOMP) [25], which constitutes ~60% of the dry weight of the outer membrane [26]. Later on, the advent of molecular biology methodologies evidenced that genotypes defined by *ompA* sequence variabilility, perfectly correlated with the prior defined serovars. MOMP contains genus-, species- and serovar-specific epitopes, where the latter are found in the four variable regions of the protein and the remainder are encoded in its conserved regions [27,28].

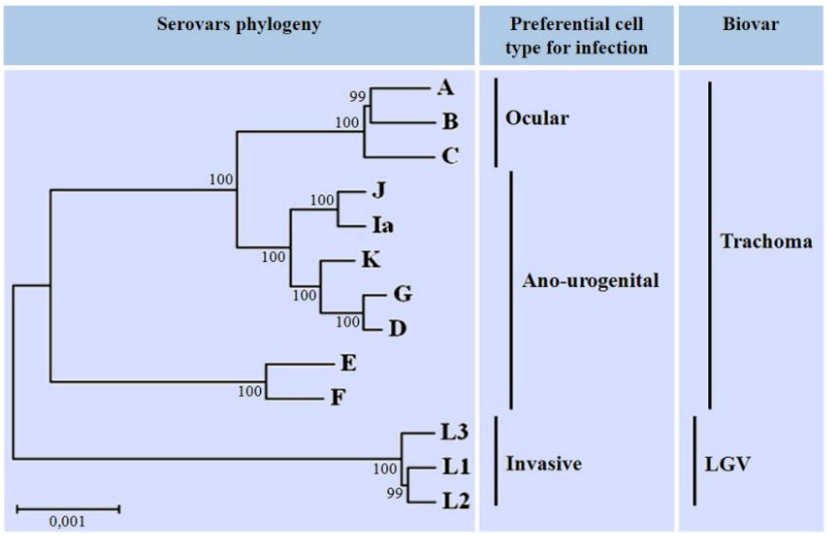


Figure 1.3. *C. trachomatis* serovars' classification based on tissue tropism and phylogeny. Representation of the phylogenetic relation between *C. trachomatis* serovars (left panel) and their classification according to the biovar and the type of infection they cause (right and middle panels, respectively). Not all the 15 main serovars are represented because, of these, two (serovars Da and H) were not fully-sequenced yet. For that same reason, the serovar I is represented by the Ia variant. The sequences used to construct this phylogeny, and the respective GenBank access numbers, are as follows: A/Har13 (NC007429), B/Jali20-OT (NC012687), C/TW3-ATCC (CP006945), D/UW3-CX (NC000117), E/Bour (HE601870), F/SW5 (NC017953), G/SotonG1 (HE601807), Ia/SotonIa1 (HE601808), J/6276 (ABYD01000001), K/SotonK1 (HE601794), L1/440-LN (HE601950), L2/434-BU (NC010287) and L3/404-LN (HE601955).

Serovars A-C are capable of infecting the ocular conjunctiva, and can lead to the development of blinding trachoma, while serovars D-K preferentially infect the epithelial cells of the genitalia but are also able to infect the ocular conjunctiva (although not leading to trachoma) and also spread to distant anatomical locations, i.e. joint epithelia and liver. Finally, serovars L1-L3 are able to infect the macrophages and therefore invade the inguinal lymph nodes (see section "1.3. Clinics and

Epidemiology" for details). According to the type of cells they infect, serovars may be classified into two "biovars" (biological variants): the trachoma biovar (serovars A-K) and LGV biovar (L1-L3) (Figure 1.3) [29]. Phylogenetic constructions using genomic sequences of the serovars clearly reflect their segregation according to tissue tropism, disease outcome and prevalence (Figure 1.3), which corroborates the fact that such phenotypic differences must have a genetic base. It also indicates that the radiation of the *C. trachomatis* serovars begun with the segragation of the LGV group, followed by the segragation of the most prevalent serovars (E and F), and finally the segregation of the ocular serovars, meaning that the latter share a recent genital ancestor with the less-prevalent genital serovars [30-32].

1.3. Clinics and Epidemiology

C. trachomatis is the leading cause of bacterial sexually transmitted infections (STI) worldwide and is the causative agent of both ocular and genital (and also anal or pharyngeal, depending on type of sexual contact) infections with several and serious complications. In 2008, the World Heath Organization (WHO) estimated that the incidence of STI due to this pathogen was 105.7 million new cases *per* year and that the infection rate has been increasing over the years [33,34]. Because the great majority of patients with *C. trachomatis* urogenital infections do not exhibit any symptoms (75-90% of patients), or exhibit slight non-specific clinical manifestations, a significant fraction of them fail to be diagnosed and, therefore, remain untreated [35,36], constituting a reservoir of individuals capable of recurrently transmitting the infection to their sexual partners. Moreover, untreated infected individuals can develop chronic infections with serious clinical sequelae, characterized by inflammation and scarring (Figure 1.4), in particular if repeated infection episodes occur, which results in significant damage of reproductive system of the host (in particular women) [35,37,38]. Also, vertical transmition may occur when the newborn passes through an infected birth canal, causing neonatal conjunctivitis and pneumonia [39,40].

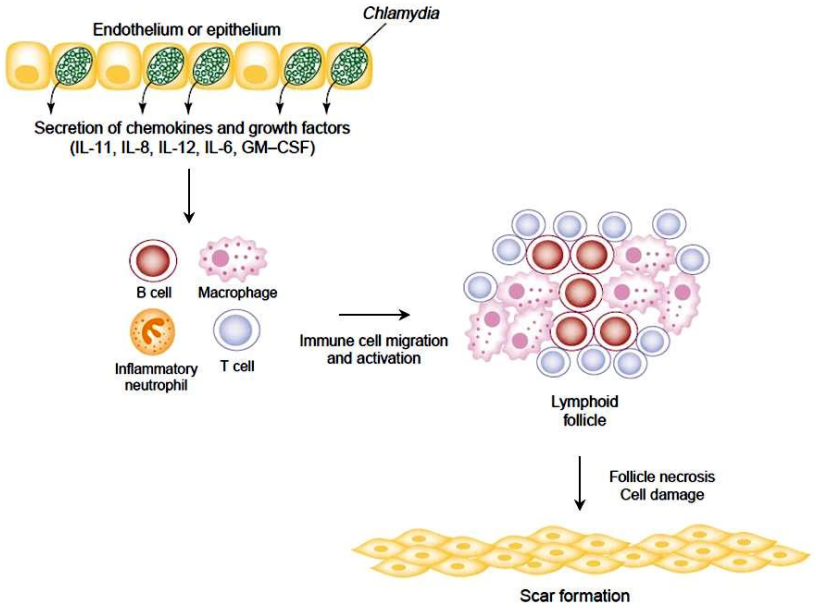


Figure 1.4. Cellular paradigm of chlamydial pathogenesis. Host's nonimmune epithelial cells are thought to be the prime factors of the inflammatory process during infection with chlamydial organisms, as they can initiate and sustain immunological responses [41,42]. Briefly, upon infection, infected host cells release cytokines and chemokines that lead to the recruitment and activation of innate (in the first place) and adaptative cells (later on). The resulting inflammatory response ultimately resolves the infection but the tissue on site eventually ends up damaged (scared) [41,43]. Adapted from [41].

Women with genital infections may develop cervicitis (with mucopurulent vaginal discharge) [44], endometritis, salpingitis and urethritis, while infections in men may progress to epididymitis and urethritis [36,43,45-49]. In 20-40% of the infected women, *C. trachomatis* progression to the upper genital tract cause serious sequelae like pelvic inflammatory disease, ectopic pregnancy and even infertility [35,43,50,51]. Also, anal infections with this bacterium can cause proctitis characterized by rectal pain, discharge and bleeding [35]. In 1-3% of the individuals, *C. trachomatis* genital tract infection disseminates, causing reactive arthritis [52].

C. trachomatis serovars A-C are able to infect the host's conjunctiva, which may result in the development of trachoma, the leading cause of irreversible, yet preventable, blindness worldwide [53,54]. Trachoma is particularly common in many of the poorest and most rural areas of Africa, Asia, Central and South America, Australia and the Middle East (according to the [55]), where inadequate hygiene routines, crowded households and water shortage are some of the risk factors that help sustain the endemic character of the disease. Active trachoma has an estimated prevalence of 21 million people infected [56], and is most commonly found among children [57]. Recurrent infection, especially within endemic regions, causes intense conjuntival inflammation episodes, that in turn results in scarring, distortion and inturning of the eye lid (trichiasis), which causes cornea damage and leads to blindness (0.5 million irreversibly blind individuals worldwide due to chlamydial infections) [55-57].

Finally, L1 to L3 serovar strains are able to mainly infect monocytes and macrophages and spread to regional lymph nodes, where they may cause the LGV [29]. This disease may present as one or more genital ulcers followed by the development of painful inguinal lymphoadenopathy (inguinal buboes) [58,59]. The LGV is endemic in tropical and subtropical regions, while it used to be considered rare in developed countries, essentially due to the regular use of antibiotics and easy access to medical care [60], but from 2004 several outbreaks have been reported in North America, Europe and Australia, especially among men who have sex with men (MSM) [61-66]. LGV, specially in MSM, seems to favor the acquisition of HIV infection, and coinfections with other STI (gonorrhoea, hepatitis C, syphilis, and *Herpes simplex*) are frequent [60,67-69].

1.4. Biology

1.4.1. Morphologycal features

Like all other members of the genus *Chlamydia, C. trachomatis* presents a unique and specialized biphasic developmental cycle of 48-72h, unparalleled among prokaryotes. The alternation among two morphologically distinct forms was first observed and described by Bedson and Bland [70], who noticed the existence of larger and dividing particles among the known smaller ones. The former are called the Reticulate Bodies (RB), which are ~1 µm in size and are capable of replication through binary fission (Table 1.1). The latter, the Elementary Bodies (EB) are smaller particles (~0.3 µm) capable of infecting the host cells (Table 1.1). Because this form possesses a highly condensed nucleoid and a relatively high resistance to osmotic or physical stress [71], it has been considered to be inert (spore-like), packed only with the proteins and nucleic acids needed for a consequent infection [37]. However, a recent and revolutionary metabolic study, performed on the EBs and the RBs separately, showed that both forms undergo *de novo* protein and ATP synthesis [72], and therefore both should be considered metabolically active forms.

Table 1.1. List of biological features of the two *C. trachomatis* forms.

Biological features	EB	RB
Diameter	~0.3 µm	~1 µm
Shape	round	round
Capable of infection	Yes	No
Capable of replication	No	Yes
Nucleoid	Highly compacted, eccentric	Diffuse, fibrillar

Both biological forms of this bacterium possess a cell envelop similar to that of Gram-negative bacteria [8], with inner and outer membranes [73,74] and a peptidoglycan layer inbetween. In the majority of free-living bacteria, this sugar amino acid polymer aids in the cell division process [75], in

the maintenance of osmotic pressure and also helps stabilize integral membrane proteins and transmembrane complexes. However, the detection of this "typical" peptidoglycan layer has been challenging in chlamydial organisms [76] and, therefore, it was proposed that chlamydiae lack peptidoglycan. Several attempts to disprove this "chlamydial anomaly" [77] have been performed. In particular, *in vitro* assays using antibiotics (β-lactams and D-cycloserin), which target the peptidoglycan synthesis, led to the emergence of aberrant RB morphologies, with no apparent RBs division nor differenciation back into EBs [77-82]. With the sequencing of the first *C. trachomatis* genomes [83,84], genes of the peptidoglycan synthesis pathway were found to be encoded in its chromosome and also that, at least some of these genes, are actively transcribed [85,86] and translated [81,87], specially during the EB to RB differenciation and RB division stages of the developmental cycle. Altogether, the genetic, transcriptomic, proteomic and antibiotic susceptibility findings seemed to indicate that *C. trachomatis* indeed synthesize peptidoglycan or a peptidoglycan-like polymer, speculations only confirmed very recently by Liechti and colleagues [88].

1.4.2. Developmental cycle

The *C. trachomatis* developmental cycle (Figure 1.5) is initiated with the EB attachment to the host cell and entry by invagination of the cell membrane. As *Chlamydia* can invade several non-phagocytic and most cultured cells, it suggests that the host receptor is ubiquitous or that more than one receptor may be used. Initially, the interaction between the EB and the host cell is thought to be reversible and mediated by heparan sulphate glycosaminoglycans [89-92]. The mannose receptor, the mannose 6-phosphate receptor, and the estrogen receptor were also proposed to play a role at this stage (reviewed in [93]. Subsequently, it is established a high-affinity and irreversible binding with secondary unidentified host cell receptors [94]. Recent studies have demonstrated the involvement of the cell surface-exposed protein dissulfide isomerase (PDI) [95] as well as growth factors and their receptors (reviewed in [96].

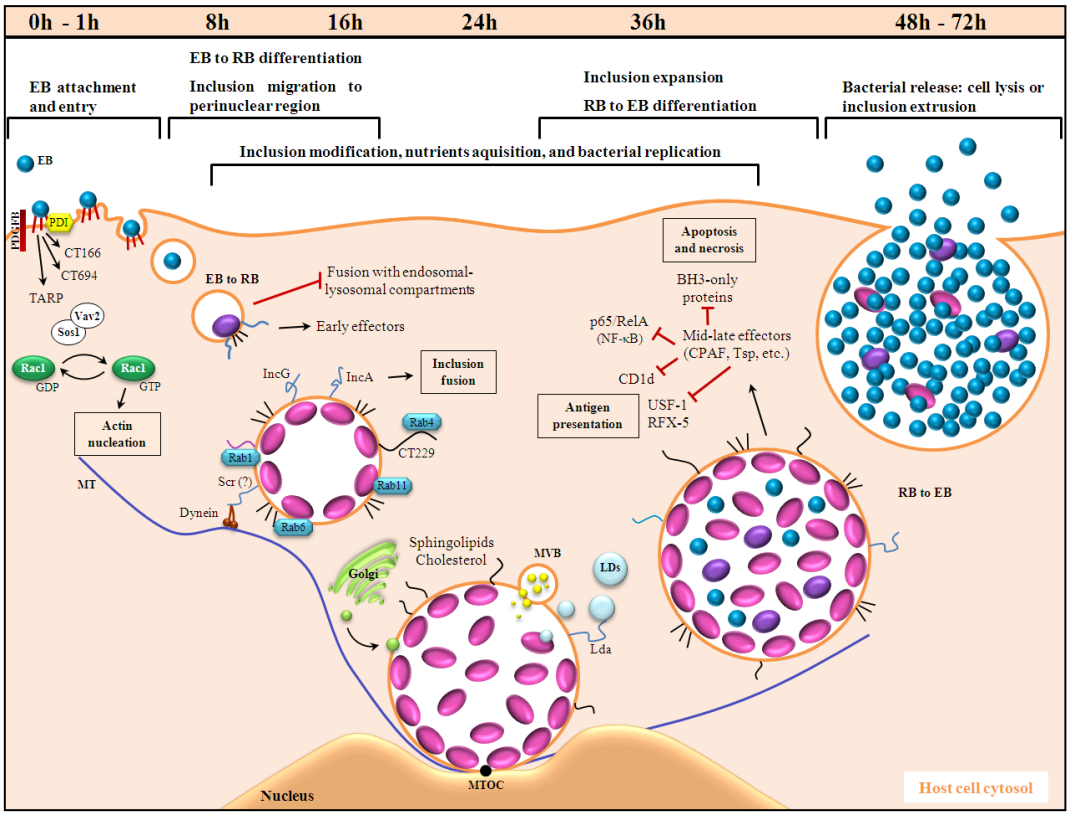


Figure 1.5. Developmental cycle of *C. trachomatis*. Schematic representation of the 48-72h developmental cycle of the *C. trachomatis* bacterium, which may be divided mainly into five main stages: 1) EB attachment and entry – with injection of several effectors into the host cytoplasm and inclusion formation; 2) EB to RB differentiation and inclusion migration to perinuclear region – with evasion from the endolysosomal pathway and RB replication; 3) inclusion modification and nutrients acquisition from the host; 4) Inclusion expansion and RB to EB differentiation – with blockage of several host pathways (e.g., apoptosis and antigen presentation) mediated by chlamydial effectors; and 5) Release from host cell by lysis or inclusion extrusion – with newly produced EBs set to infect neighbouring cells. Adapted from [93].

On the bacterial side, multiple adhesins and ligands have been proposed to mediate EB entry like, for example, glycosaminoglycans [97], MOMP [98] and polymorphic membrane protein (Pmp)-D [99]. Upon contact with the host cell membrane, the type III secretion system (T3SS) of the EB (reviewed in [100] immediately discharges effectors into the host cell cytosol. One of them is the polymorphic TARP (translocated actin-recruiting protein), which contributes to the bacterial internalization, through its ability to directly nucleate actin polymerization [101,102] and by recruiting at least two guanine nucleotide exchange factors (Sos1 and Vav2), which in turn activate Rac1 GTPases and signal to the actin machinery [103,104]. The actin rearrangements, that culminate in bacterial entry into the host cell, are transient and may be terminated by the secreted chlamydial effectors CT166 and CT694, where the former glucosylates the Rac1 GTPase [105] and the latter interacts with the actin binding protein AHNAK [106]. This process creates an EB-containing vesicle, termed the inclusion, which is able to

escape from the endolysomal pathway and to migrate towards a perinuclear location (reviewed in [107]) along the microtubule network, in a dynein-dependent but dynactin-independent manner [108,109]. During this stage, EBs quickly differentiate into RBs, which engage in repeated cycles of replication by binary fission. Also, the inclusion initiates a series of interactions with host molecules and organelles involved in: *i*) the endoplasmic reticulum (ER)-Golgi trafficking (Rab GTPases), for the acquisition of essential host-derived nutrients; *ii*) the subversion of host-innate immune response; *iii*) the inhibition of host cell death (apoptosis); and *iv*) the occasional homotypic fusion with other inclusions through IncA (reviewed in [96] and in [110]; [111]. Following the period of cell replication, RBs begin to redifferentiate back into EBs. The stimulus for the initiation of this process is still unknown, but it was proposed that the dissociation of the dividing RBs from the inner side of the inclusion membrane may work as a trigger [37]. Finally, EBs may be released from the host cell by one of two mechanisms, cell lysis or inclusion extrusion, for infecting neighboring cells and initiate another cycle.

Under *in vitro* stress growth conditions, imposed by immunological factors (e.g., IFN (interferon)-γ) [112], presence of antibiotics [113] or nutrient deprivation [114], the normal developmental cycle of *C. trachomatis* is disrupted and chlamydial organisms may enter a state of "persistence". This "persistent infection" has been defined as a long-term association between viable chlamydiae and their hosts, in which the former remains in a culture-negative state, with non-typical morphology and no evident growth. Therefore, it is assumed that this situation usually translates into mild or absent clinical manifestations [115]. Although reversible, this *in vitro* persistence is characterized by altered chlamydial growth features, which include loss of infectivity and the development of smaller inclusions containing few aberrant RBs (enlarged and with multiple chromosomes), that neither undergo binary fission nor differenciate into EBs (reviewed in [38] and [37]). While the great majority of researchers attribute the subtle nature of some chronic chlamydial infections (occurring *in vivo*), partially, to these phenotipically abnormal RBs (observed *in vitro*) [115-117], this association has also been argued against [118]. A comprehensive overview on this subject was published by Wyrick [119], which finalizes with an appeal for more research on samples from infected patients, to help shed some light on this, still to prove but, probable association between *in vitro* persistence with (lack of) *in vivo* symptoms.

1.5. Genomics

Since the sequencing of the first genome [83], more than 100 have become available at free-access databases on the internet (http://www.ncbi.nlm.nih.gov/genome/genomes/471; http://www.sanger.ac.uk/resources/downloads/bacteria/chlamydia-trachomatis.html; http://www.ebi.ac.uk/ebisearch/search.ebi?query=chlamydia+trachomatis&db=allebi&requestFrom=ebi_index). Nonetheless, this increase in genomic data only corroborated the first observations of the main characteristics of *C. trachomatis* genome. In particular, that this bacterium is known to own one

of the smallest chromosomes among bacteria (1.04 – 1.05 Mb) [83,120,121], with a GC content of ~41% [83,84,122,123], and about 900 predicted coding sequences (CDS) [83], which translates to a coding content of ~89.5%, very similar to what is found among other bacterial genomes on average [124], but contrasting to other obligate intracellular pathogens [125]. Also, this bacterium often carries a single [126], but multicopy [120], double-stranded DNA plasmid of ~7.5 Kb in length, which encodes only eight proteins [127].

Comparative analysis of several strains' genomes showed that they possess a high degree of sequence similarity (>98%), nearly identical synteny, and that both the chromosome and the plasmid have similar sizes among strains [14,30,128].

1.5.1. Chromosome

The small chromosome size of the *C. trachomatis* species is thought to be the result of a reductive evolution, characterized by irreversible gene loss, which has been associated to the long adaptation process to the intracellular lifestyle [124]. Moreover, and unlike the majority of other bacteria, *C. trachomatis* chromosome presents no pathogenicity islands nor other mobile genetic elements, with the exception of some putative remnant fragments of excised insertion sequence-like elements with flanking direct repeats, within several loci [129,130].

As referred above, the *C. trachomatis* strains present a similarity degree of >98% at the genome level but manifest significantly different phenotypes. Therefore, it is assumed that those <2% of variability hold the genetic basis of the dissimilar virulence, tissue/cell tropism and ecological success displayed by the strains. Comparative genomics already pointed out the regions of the genome more prone to vary but there is still a lack of studies linking those genetic differences with *C. trachomatis* strains' dissimilar biological properties. Some of the main variable loci, and whose phylogeny segregate one or more groups of same-disease causing strains, are located within the "plasticity zone" (PZ), which is a ~50 Kb long (45–49 genes) region [19,83], known to be hypervariable even among Chlamydiaceae [19,131], while the others exist scattered around the chlamydial chromosome. The latter set of genes encompasses important antigens, structural proteins and T3SS effectors and transporters. Of note, they are: *i*) the *ompA*, traditionally used for typing purposes, that codes for the MOMP, the major surface exposed chlamydial antigen presenting distinct antigenic profiles for B- and T-cell epitopes [27,28], also assumed to function as an adhesin [132], and a porin [133,134]; *ii*) the *tarp*, which encodes a T3SS effector secreted into the host cytoplasm at a very early stage of the infection to induce actin polymerization [102]; *iii*) the *pmp* family, comprised by nine *Chlamydiae*-specific genes, constituting ~13.6% of the *C. trachomatis* genome coding capacity [83,84,122,135], and coding for autotransporter proteins [136-138], which may also mediate niche-specific adhesion [139,140] and provide antigenic diversity essential for immune evasion [122,130,141]; *iv*) the *inc* family, a wide [142,143] and heterogeneous group of T3SS effectors [144] that share a 40-60 long aminoacid bilobed hydrophobic secondary motif, and with only a few members with characterized function; and *v*) a significant number

of genes (30-35% of the genome) encoding proteins whose functions are yet to be disclosed (termed hypothetical proteins, HPs), of which ~87% are chlamydial-specific and show no similarity to hypothetical proteins from any other bacteria [83], probably due to their involvement in *C. trachomatis*/host-specific interactions.

On the other hand, inside the PZ is worth highlighting: *i*) *CT153*, coding for the membrane attack complex/perforin (MACPF) domain protein, located at the inclusion membrane and whose expression was associated with the ability of *C. trachomatis* to accumulate lipid droplets inside the inclusion, in a strain-specific manner [145]; *ii*) the phospholipase D (PLD) family, assumed to be involved in the aquisition of host-derived lipids and to play a strain-specific role in chlamydial pathogenesis [146]; *iii*) the cytotoxin gene, homologous to other known bacterial toxin encoding genes, that codes for a protein that acts very early in the infection process and causes morphological and cytoskeletal changes in the epithelial cells [147-149]; and *iv*) the tryptophan biosynthesis operon (*trpRBA*), coding for a bifunctional enzyme that catalyzes some of the metabolic reactions for producing tryptophan from indole, which allows chlamydial growth in an IFN-γ-rich environment [149,150].

Although the genetic variability in all these (and some other) genes among the *C. trachomatis* strains are known, to date only the genetic variability of the *trpRBA* operon was directly associated with different tissue tropism, i.e., only strains with an intact *trpRBA* operon are able to infect the genitalia-epithelium, likely due to the induction of tryptophan-degrading enzyme indoleamine 2,3-dioxygenase (IDO) by the IFN-γ present in this epithelium [151], which in turn causes the starvation of strains with a defective thryptophan biosynthesis pathway [152]. The discovery of other reliable associations between genotype and clinically relevant phenotype, has been mainly sustained by the unique characteristics of *C. trachomatis* developmental cycle, i.e., EBs are spore-like structures, exhibiting a highly compact cell-wall [153,154], and because RBs' chromosome is enclosed within three distinct membranes (-self, -inclusion, and -host). Also, attempts to genetically manipulate the genome of this bacterium have been unsuccessfull, until the present decade, when Wang and colleagues [155] developed a plasmid-based transformation system. Since then, several methods have been established and applied to introduce different genes into *C. trachomatis* [156-161], enabling, for instance, the functional characterization of the genes encoded by the single plasmid of this bacterium (see section "1.5.2. Plasmid" for details), in a study associating the deletion mutagenesis of plasmid genes with subsequent chlamydial transformation [162]. Also, potential associations between genotype and phenotype have been confirmed using these novel approaches, which will certainly help dechiphering the genetics underlying clinical relevant serovars' dissimilarities.

Despite the recent *in vitro* advances in growing *C. trachomatis* ouside its strick host [72], three distinct membranes separate, *in vivo*, its chromosome from the extracellular environment for most of the developmental cycle (as stated above). Therefore, horizontal gene transfer with foreign DNA was considered unlikely to occur, corroborated by the fact that the core- and pan- genomes of this bacterium were found to be almost identical. Even intra-species recombination was a very recently accepted event

given the constraints created by the "bottleneck" effect of the developmental cycle on effective population size, and also because co-infections by different-serovar strains are expected to occur only at a 1% frequency [10,163], and the fusion of both inclusions would be mandatory. However, not only genome-dispersed recombination (involving some well-known "hot-spots") between *C. trachomatis* strains, regardless of their serovar, has been evidenced [30,164-171], as Chlamydiae also have all the recombination apparatus required to do so [83]. It has been also observed that the exchange of genetic material is more frequent among strains sharing the same niche-tropism [30,31], with particular emphasis between the ano-urogenital strains, when comparing to the exchange among the more niche strict ocular or the LGV stains [31]. On the other hand, although recombination has been continuously acknowleged as a natural and somehow frequent phenomenon, shaping the evolution of *C. trachomatis* strains, this pathogen remains recognized as a low recombining organism [31,172,173], for which mutation events were found to occur at a much higher frequency, likely constituting its major evolutionary driving force [32].

1.5.2. Plasmid

Despite the extensive reductive evolution of the *C. trachomatis* chromosome [124], this bacterium still keeps the small [~7.5 Kb and with only eight open reading frames (ORFs)] and conserved plasmid (less than 1% of nucleotide sequence variation among serovars) [174,175] (Figure 1.6), which suggests a strong selective pressure acting towards its maintenance. The presence of the plasmid was shown not to be essential for chlamydial survival [176] and for several years it was considered to be cryptic, although "plasmidless" strains displayed a slight decrease in growth efficiency [177]; this phenomenon may relate to the reduced pathogenicity observed during *in vivo* infectivity assays by using plasmidless and plasmid-bearing strains of *C. muridarum* as a model [23,178,179]. Another observed phenotype associated to the plasmid's absence was the inability of strains to accumulate glycogen within the inclusion [177,180], but the biological impact of this feature on chlamydial survival or virulence lacks experimental confirmation and can only be hypothesized [180]. Moreover, the frequency of the natural occurrence of plasmidless strains in a population [176,181,182] was at ~1% [177], corroborating the notion that plasmid-bearing *C. trachomatis* organisms hold selective advantage over the plasmidless ones.

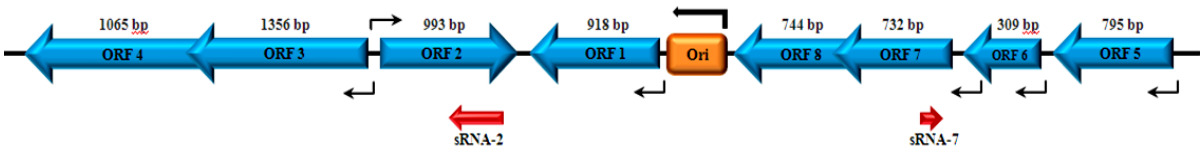


Figure 1.6. Schematic representation of the *C. trachomatis* plasmid. The orange box, represents the location of the origin of replication (Ori) of this plasmid, and the direction of the replication is indicated by the black arrow above it. In thick blue arrows are represented the eight genes encoded by the plasmid and their respective orientation. The designation of each gene (shown inside the blue

arrows) is the same adopted by Thomas and collegues [183], where the numbering of genes was given according to the plasmid's replication direction, starting from the gene immediatly downstream the Ori (ORF1) and continuing to the last gene upstream the Ori (ORF8). The numbers above each gene correspond to the respective length in base pairs (bp) (based on sequence of plasmid pCTA (NC_007430) of A/HAR-13 strain). Moreover, the red arrows indicate the approximate location and orientation of the two anti-sense sRNA found to date. Finally, all the other thin black arrows intend to indicate the genes' known transcription start sites and direction [184].

Over the years, several studies were conducted with the intent of characterizing this intriguing plasmid. It was already demonstrated that all plasmid genes are transcribed and translated into proteins [185-187], and further studies indicated that the plasmid may function as a virulence factor and a transcriptional regulator of chromosomal genes [162,180]. Furthermore, and despite the *C. trachomatis* plasmid has been considered a virulence factor, it does not seem to be horizontally transferred during infections, nor it has conjugative or integrative capabilities [175], unlike plasmids of other bacteria carrying several different virulence factors [188]. As phylogenetic data showed, the segregation of strains based on plasmid sequence, reflects the one based on their chromosome, suggesting a parallel evolution of both molecules within strains [189].

On the other hand, genetic characteristics and putative functions have been pointed out for the majority of the plasmid genes and proteins, respectively. For instance, the ORF2-encoded protein possesses conserved domains belonging to the "recombinase-like" family, and also shares ~32-35% of amino acid identity with the plasmid ORF1-encoded protein, which in turn has been implicated in plasmid replication, due to its size, net positive charge and location in the plasmid [183,190]. Therefore, it is assumed that both proteins may be related to each other and together play a role in the plasmid replication process [183]. However, the product of the ORF2 has been designated as the "major replication protein", because ORF1 is truncated in the *C. pneumoniae* species but the replication of its plasmid still occurs [183,191]. The protein encoded by the downstream ORF3 is also presumed to play a role on the replication as well, due to its homology to an helicase, the DnaB, of other bacteria (*Escherichia coli* and the *Salmonella typhimurium*), involved in the unwinding of the DNA strands during the replication process [183]. The ORF5 encodes a 28 kDa protein (Pgp3), which was found within the inclusion, as well as in the cytosol of the infected host cells [185]. This is the most immunogenic antigen of all the eight plasmid-encoded proteins [18], which may correlate with the fact that its sequence is the most variable among them [174,189]. Moreover, it was shown that Pgp3 protects against "chlamydial challenges" in murine experimental models [192,193], rendering it the status of a vaccine candidate. On the other hand, this trimeric protein [185,187] also stimulates macrophages to release cytokines, which may contribute to the pathology by inducing inflammation [18]. ORF7 was shown to have partial sequence homology with other genes coding for partitioning proteins, like SopA and ParA [183], and given the common transcription start site (TSS) with ORF8, shown to be essential

for the stable maintenance of the plasmid in tissue culture [162], it has been speculated that these two ORFs may constitute an operon [184] similar to *sopA/sopB* and *parA/parB* operons of other bacteria [194,195], responsible for the partitioning of plasmids during cell division. This argument is corroborated by the fact that the *C. trachomatis* plasmid is found in a low copy number [189,191,196,197] within bacterial cells, and that only a small fraction of the population (~1%) fails to harbour it; therefore, its segregation seems not to rely on chance but rather on an active and well-regulated partitioning system. Finally, both ORF4 and ORF6 have no homology to known genes, and while the protein encoded by the former has no attributed function, the protein encoded by the latter seems to be required for the regulation of the transcription of both chromosomal and plasmid genes [162].

Apart from the genes carried by the plasmid, it was shown that it possesses an Ori located between the ORF1 and the ORF8, characterized by 4 tandem repeats of 22 bp, preceded by a cluster of AT repeats [183,197]. Moreover, the *C. trachomatis* plasmid was also found to encode two antisense mRNAs in the complementary strand of the ORF2 and ORF7, respectively, implying a putative role in plasmid transcription regulation by a complementary base-pairing mechanism [183,186,190]; nonetheless, their function has not been precisely elucidated.

1.6. Scope of the thesis

C. trachomatis is a strict human pathogen with a huge impact on human heath and so, the management of the the infection burden and sequelae should reduce the impact of the pathologies caused by this bacterium and also their associated economical costs. Hence, elucidating the role of "virulence factors" and/or the mechanisms by which *C. trachomatis* survive, persist, evade the immune system, sustain the inflammatory response, and ultimately, cause great damage to the host, remain of extreme importance. Also, understanding the molecular dissimilarities between strains that cause different pathologies will also help the development of more specific prophylactic measures and therapies. However, the obligatory nature of *C. trachomatis* intracellular developmental cycle, the inexistence of a suitable animal model for *in vivo* studies, and its (up until very recently) genetically non-tractable character have hampered the applicability of commonly used research approaches in the study of this unique bacterium with high clinical interest. As a result, genomics, transcriptomics and bioinformatics gained special relevance for identifying and possibly deciphering the inter-strains' variability at the genome and transcriptome levels that could translate in specific pathological phenotypes.

Considering this, the ultimate goal of the present thesis was to contribute for the understanding of *C. trachomatis* chromosome and plasmid molecular dynamics underlying relevant phenotypic

differences. As the subsequent chapters represent independently published studies, the detailed objectives of each one are as follows:

In chapter II, we determined the mutation and recombination rates of *C. trachomatis* chromosome. We also performed a parallel evaluation of the biasing effect caused by genes with different polymorphic charecteristics and biological roles on such evolutionary parameters estimates, because the use of conserved *versus* polymorphic genes for these determinations were still debatable at the time this study was carried out. During these analyses, the reproducibility of the results using different numbers of iterations (a software feature), as well as the convergence of twin bioinformatic runs, were evaluated, as they may also impact the evolutionary estimations.

The study presented in chapter III intended to be a complete and systematic polymorphism and evolutionary analysis of all chromosome genes, by using all the genome sequences available at the time. Ultimately, the information generated through this work constitutes a valuable and unprecedented database of the variability (gene-by-gene) found within the *C. trachomatis* species.

In chapter IV, the subject of the study was the plasmid of *C. trachomatis*. The main objectives of this chapter were to: *i*) determine the precise number of plasmids harboured by this bacterium, and if it varies among strains; *ii*) determine if the number of plasmids *per* strain, and the expression profiles (of both the plasmid genes and the sRNAs) differentiate strains by their tropism; *iii*) try to assess if there is a relation between genes and sRNAs expression levels with plasmid load during *C. trachomatis* developmental cycle; and *iv*) perform a comprehensive polymorphism analysis of each plasmid gene.

Finally, in chapter V we aimed to evaluate the transcriptome of the obligate intracellular bacterium *C. trachomatis* and to assess the decay rate of all mRNAs from different-biovar strains. We also evaluated a potential relation between mRNAs stability with their abundance and phenotypic dissimilarities between the strains included in this study.

Chapter II

Impact of Loci Nature on Estimating Recombination and Mutation Rates in *C. trachomatis*

This chapter corresponds to a manuscript (with discrete changes) with the following reference:
Ferreira R, Borges V, Nunes A, Nogueira PJ, Borrego MJ and Gomes JP (2012) *G3 (Bethesda)*. 2:761-768.

Personal contribution
RF performed most of the experiments, analyzed the data, and wrote the paper.

2. Impact of loci nature on estimating recombination and mutation rates in *C. trachomatis*

2.1. Abstract

The knowledge of the frequency and relative weight of mutation and recombination events in evolution is essential for understanding how microorganisms reach fitted phenotypes. Traditionally, these evolutionary parameters have been inferred by using data from multilocus sequence typing (MLST), which is known to have yielded conflicting results. In the near future, these estimations will certainly be performed by computational analyses of full-genome sequences. However, it is not known whether this approach will yield accurate results as bacterial genomes exhibit heterogeneous representation of loci categories, and it is not clear how loci nature impacts such estimations. Therefore, we assessed how mutation and recombination inferences are shaped by loci with different genetic features, using the bacterium *C. trachomatis* as the study model. We found that loci assigning a high number of alleles and positively selected genes yielded nonconvergent estimates and incongruent phylogenies and thus are more prone to confound algorithms. Unexpectedly, for the model under evaluation, housekeeping genes and noncoding regions shaped estimations in a similar manner, which points to a nonrandom role of the latter in *C. trachomatis* evolution. Although the present results relate to a specific bacterium, we speculate that microbe-specific genomic architectures (such as coding capacity, polymorphism dispersion, and fraction of positively selected loci) may differentially buffer the effect of the confounding factors when estimating recombination and mutation rates and, thus, influence the accuracy of using full-genome sequences for such purpose. This putative bias associated with in silico inferences should be taken into account when discussing the results obtained by the analyses of full-genome sequences, in which the "one size fits all" approach may not be applicable.

2.2. Keywords

Mutation rate; Recombination rate; Evolutionary inference; ClonalFrame

2.3. Introduction

The ecological success of bacteria relies on their constant ability to diversify their genetic background to reach better-fitted phenotypes through selection. In this regard, point mutations and recombination events are especially relevant as they may be the basis for antigenic polymorphism, virulence dissimilarities, and differential tissue tropism [28,198,199]. As for mutation events, in which

bacteria range from monomorphic (e.g. *Yersinia pestis*) to highly polymorphic (e.g. *Helicobacter pylori*) [200], recombination is not equally important for all microorganisms. Indeed, they range from strictly clonal (lack or extremely low rates of recombination), such as *Mycobacterium* species or *Staphylococcus aureus* [172,201,202], to typical recombinants, such as *Helicobacter pylori* or *Neisseria gonorrheae* [203,204]. In the middle, there are microorganisms with a moderate recombination background that generate new genomic mosaic structures more fitted to deal with the environment, yielding new successful clones through a never-ending evolutionary process.

The influence of allelic exchange in the evolution of bacterial pathogens has been measured by calculating the relative weight of recombination and mutation rates. Traditionally, these calculations have been performed on MLST data resulting from the analysis of housekeeping genes (HK). However, the use of MLST data has yielded strikingly different results within the same species when estimations are performed with dissimilar MLST loci, strain samples, or analytical methodologies [205]. The rational for using this strategy relies on several arguments. On the one hand, large data sets are available for molecular typing purposes, and HKs are commonly dispersed around the chromosome, which prevents more than one gene from being affected by a single recombination event. Moreover, the use of HKs intends to avoid biased results because the accumulation of mutations may be confounded with the exchange of alleles by recombination when we employ loci that are either "highly polymorphic" or "too conserved", multicopy or under positive selection [206]. Nevertheless, this may not be a straightforward assumption as, except for the fixation of beneficial mutations through positive selection, the occurrence of point mutations exactly in the same genomic position simultaneously for several strains (homoplasy) likely results from recombination within the population [207]. Another question when employing MLST data to infer recombination is the use of a low number of HKs (usually seven), which may not accurately represent the genomic variability. Indeed, a previous study on bacteria found no justifiable reason for applying HKs when inferring intraspecies phylogenetic relationships, and it pointed out that the major concern when choosing candidate loci should rely on their genetic variability [208]. Thus, a wider approach based on using full-genome sequences has been recently applied, as it is expected that biasing effects from "inconvenient" loci are diluted. However, there is a multiplicity of bacterial species in which genomes have a highly heterogeneous representation of loci with different traits, such as polymorphism degree, size of intergenic regions, and selective pressures. Thus, it should be assessed how loci nature shapes the estimation parameters for understanding microbial evolution. One microorganism that may constitute a good model for evaluating the bias associated with the calculation of recombination and mutation rates through the analysis of different types of loci is the obligate intracellular human pathogen *C. trachomatis* due to its singular genomic features. Indeed, the core and the pan genomes of the 15 serological variants (serovars) of this pathogen are nearly identical, indicating that horizontal gene transfer is not relevant in *C. trachomatis* evolution. Moreover, the genome similarity among serovars is about 99%, in which major polymorphism is provided by few highly variable loci dispersed throughout the chromosome [84], with evidence of positive selection for some of them

[173,209]. Also, *C. trachomatis* is under the final stages of the evolutionary process of genome reduction [124], containing few nonessential genes and pseudogenes. Therefore, intergenic regions (IGR) likely contain regulatory domains of essential genes, which make IGRs putative targets of selection. In fact, it has been shown that several IGRs exhibit the same phylogenetic signal as neighboring genes [32]. Finally, although mutation events likely constitute the *C. trachomatis* major evolutionary driving force [32], phenomena of genome-dispersed recombination have been recently described, seemingly related to tissue tropism and antigenic variability [167,169,170]. Accordingly, we applied the widely used robust bioinformatic platform ClonalFrame [210] to several data sets encompassing loci that may differently impact the estimation of recombination and mutation rates, namely, *i*) HKs from a recently developed MLST scheme [211]; *ii*) positively selected genes (PSG); *iii*) five groups of loci strictly ranked by their number of alleles; and *iv*) intergenic regions. The results from these data sets were compared with data generated through a wide genomic approach. The present study gets insights on the bias introduced when loci with different genetic features are used to estimate recombination and mutation rates. Our approach differs from previous evaluations [208,212,213] as we have assessed the individual weight of each group of loci. We believe our results may help to elucidate how the evolutionary parameters are shaped, which will certainly be essential for the comprehension and validation of the data generated through the computational analyses of full-genome sequences.

2.4. Materials and methods

2.4.1. Chlamydial culture

By the time this work was performed, only four (A/Har13, B/Jali20, D/UW3, and L2/434) out of the 15 *C. trachomatis* prototype strains (representing the 15 existing serovars) had been fully sequenced [83,84,122,189]. To obtain sequences for *in silico* analysis, we propagated prototype strains from the remaining serovars (Ba/Apache-2, C/TW3, E/Bour, F/IC-Cal3, G/UW57, H/UW43, I/UW12, J/UW36, K/UW31, L1/440, and L3/404). Our strategy relied on using the 15 prototype strains representing all serovars because tropism differences are well defined at the serovar level, and recent phylogenetic analysis showed that the chosen strains are likely representative of the major genetic variability within the species [30]. Indeed, it is known that differences between same-serovar strains may be as low as 20 single nucleotide polymorphisms (SNP) [10]. Cell culture was performed through standard techniques as previously described [214]. Briefly, T_{25} cm^2 flasks of confluent HeLa 229 cell monolayers were independently inoculated with each strain, and cultures were allowed to grow at 37°C, 5% CO_2 for about 48h. After bacterial growth, infected cells were harvested by scraping, sonicating, and centrifuging, and the obtained bacterial pellet was subjected to DNA extraction by using the QIAamp DNA Mini Kit (Qiagen) according to manufacturer's instructions, and then stored at -80°C until use. We then amplified and sequenced the selected genomic regions (see below) for the propagated serovars. PCR primers are

listed in supporting information, Supplemental Table 2.1. Sequencing was performed as previously described [130].

2.4.2. Loci selection and grouping strategies

Considering the high genomic similarity among the *C. trachomatis* serovars (about 99%) [84], we used comparative genomics over the four fully sequenced serovars to select informative genomic regions for inferring evolutionary parameters. We were able to select a set of 136 chromosome-scattered and functionally diverse genomic regions (see Supplemental Table 2.2), which include 56 IGRs and 80 genes. The selected genomic regions are highly representative of the *C. trachomatis* serovar variability as they comprise about 55% of the total SNPs in just one tenth of the chromosomal length ($P < 10^{-7}$) (see Supplemental Table 2.3). These regions were then differently grouped according to specific characteristics. First, for each serovar, we created a group encompassing all 136 regions by compiling their sequences while maintaining the relative order of loci in the *C. trachomatis* chromosome. Throughout the text, the strategy using this first data set will be referred to as the wide genomic approach. The second data set, termed HK-MLST, is constituted by the seven HKs that compose a MLST system [211]. Subsequently, we created five additional data sets by dividing the 80 selected genes according to the number of alleles that each gene defines among the 15 *C. trachomatis* serovars: 1 to 5 (17 genes), 6 and 7 (17 genes), 8 and 9 (18 genes), 10 and 11 (15 genes), and 12 to 15 alleles (13 genes) (see Supplemental Table 2.2). Finally, we intended to evaluate the impact of using PSGs and IGRs, which are loci categories commonly not recommended when performing this type of analysis, although their potential confounding effects lack experimental support. Thus, we created two data sets composed of 11 PSGs and 56 IGRs, respectively. The use of the IGR data set also relies on recent evidence indicating that noncoding regions may also be affected by selection [215,216] and recombination [167], which suggests that there is no apparent reason to completely rule out their use for evolutionary inferences. All studied loci are represented in Figure 2.1.

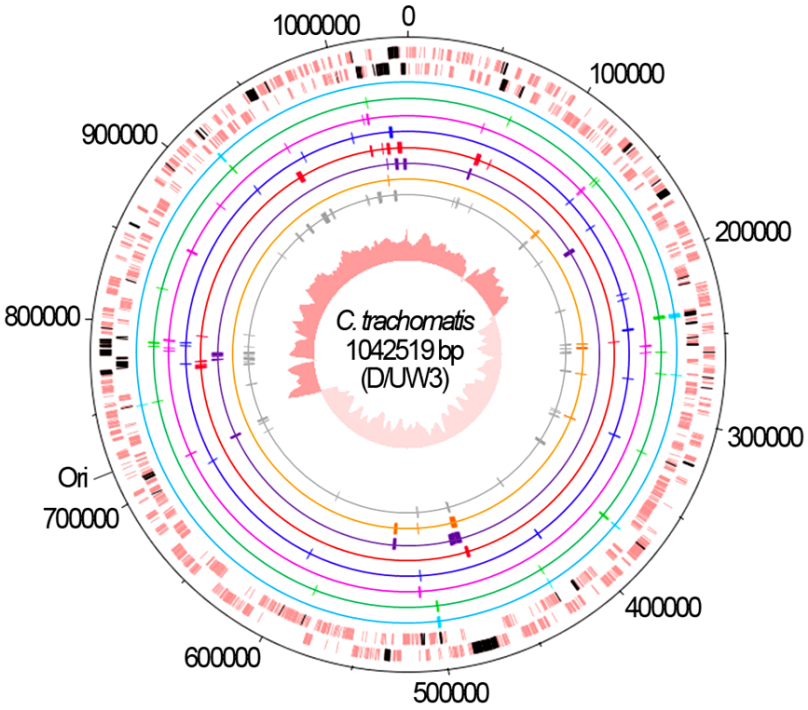


Figure 2.1. Chromosomal mapping of studied loci. The two outer lanes represent the DNA strands of the *C. trachomatis* chromosome of D/UW3 strain (GenBank accession number NC000117), where the 80 genes (from the total 136 genomic regions evaluated) are shown in black. Each data set is represented by inner circles: HK-MLST (light blue), alleles 1 to 5 (green), alleles 6 and 7 (pink), alleles 8 and 9 (dark blue), alleles 10 and 11 (red), alleles 12 to 15 (purple), PSG (orange) and IGR (gray). The central circle shows the G/C skew plot. The precise identification of the loci is shown in Supplemental Table 2.2.

2.4.3. progressiveMauve alignments

Mauve software (http://asap.ahabs.wisc.edu/mauve/) allows the construction of multiple genome alignments for the identification of conserved regions, SNPs, indel events, inversions, and other rearrangements (and their breakpoints location) across the aligned genomes [217]. We aligned the sequences of the 15 prototype strains of each data set through the progressiveMauve algorithm [218] of the Mauve software v2.3.1. As the sequences length of different data sets were below 1 Mbp, we used a conservative seed weight value (match seed weight = 11) to improve the alignment by reducing noisy matching. The resulting alignments were manually confirmed, and the output files were subsequently used in ClonalFrame software. Although Mauve is particularly useful for aligning full-genome sequences, we used this application as it generates reliable alignments in a compatible format for ClonalFrame.

2.4.4. ClonalFrame analysis

ClonalFrame (http://www.xavierdidelot.xtreemhost.com/clonalframe.htm) is a widely applied software for inferring the bacterial evolutionary parameters and events underlying DNA sequence

variation either from full genomes or from independent regions (such as MLST data sets). The computational cost of the analysis is greatly reduced when the inference is applied to unlinked regions rather than to full genomes, by reconstructing the clonal genealogy and further analyzing each region separately. This is a viable strategy as unlinked regions of the genome are assumed approximately independent given the clonal genealogy of a sample. The ClonalFrame inference is performed in a Bayesian framework, assuming a standard neutral coalescent model [210]. In this study, the ClonalFrame software v1.2 was used for estimating mutation and recombination rates of dissimilar data sets to evaluate the impact of loci nature on these estimations. Considering the aim of the present study, the ClonalFrame options were selected to: *i*) estimate the mutation rate (θ), the rate of new polymorphism introduced by recombination (ν), the average tract length of a recombination event (δ), and the recombination rate (R) during each run; *ii*) construct a uniformly chosen coalescent tree; *iii*) assume a constant population size model; *iv*) generate a random seed value for each independent run; and *v*) perform the branch swapping attempts in at least half of the time of each iteration. For each data set, two independent ClonalFrame runs were performed. When alignment artifacts hampered the correct function of the software, we manually removed the gap regions while maintaining the genetic variability among *C. trachomatis* serovars, and both new Mauve alignments and ClonalFrame runs were performed. All simulations were carried out using a Linux server.

As different numbers of iterations may yield deviating results, we conducted an analysis of the ClonalFrame reproducibility by performing two independent runs of the wide genomic data set, using a wide range of iterations (30,000, 100,000, 300,000, 500,000, and 1,000,000). For all runs, the first half of the iterations was discarded as burn-ins, and parameters were sampled every 100 iterations during the second half. The optimal number of iterations determined was applied for the subsequent analyses.

We also assessed the convergence of the estimated parameters (θ, R, δ, and ν) from independent runs on the same data set and with the same options by applying the method of Gelman and Rubin [219] implemented in the Graphical User Interface of the ClonalFrame software. We assumed replicate runs to be convergent only when the calculated test statistic was adequate (i.e. below 1.1) for all parameters. Additionally, we performed a fine-tune analysis using the ClonalFrame phylogenetic tree comparison tool, which allows the visualization of the level of confidence (based on a color scale) in each node of the consensus tree of a first run according to the output data of a second run. Each node is given a color code according to the level of confidence; white and black indicate no confidence or total confidence, respectively. On this basis, we attributed a score to each node [ranging from zero (white nodes) to three (black nodes)] (see Supplemental Figure 2.1) to achieve a numerical comparison between the runs of different data sets. The sum of the scores of all nodes of each tree was then divided by the respective number of nodes to calculate an average concordance score. Finally, we evaluated the confidence on the estimates of r/m (measure of the weight of recombination on diversification relative to mutation) and ρ/θ (measure of the frequency of occurrence of recombination relative to mutation events) obtained for each data set.

2.4.5. Nucleotide sequence accession numbers

The sequences of all *C. trachomatis* loci determined in this study were submitted to GenBank under the accession numbers JQ066324–JQ066356 and JQ066367–JQ066722.

2.5. Results and discussion

The analysis of the evolutionary history of bacteria relies on deciphering genetic differences that arose from several mechanisms, of which point mutations and recombination events are among the most relevant driving forces. The knowledge of the frequency and the relative weight of these two mechanisms is crucial for understanding the biology and the genealogy of microorganisms. This is generally achieved by calculating the ratio ρ/θ, which determines the relative frequency of occurrence of recombination and mutation events, and the ratio r/m, which measures the relative impact of recombination and mutation in genetic diversification. In fact, the estimation of these basic population parameters for microbial pathogens has proved useful, for instance, in explaining the dynamics of drug resistance and pathogenicity and may indicate which epidemiological process should be targeted for disease control [207,220]. Nevertheless, identifying and determining the exact extent of recombination events is not a simple and straightforward procedure, as there is no ideal methodology for establishing relationships for all bacteria, from strictly clonal to highly recombining microorganisms [221]. Didelot and Falush [210] developed a robust computational platform, ClonalFrame, which has yielded valuable results in the inference of both the population structure and the role of the recombination process in several microorganisms, such as *H. pylori* [212], *Listeria monocytogenes* [222], and *Salmonella enterica* [223]. Although most inferences have been generated by using MLST data, it is expected that the analysis of full-genome sequences will be the most applied strategy in the near future. However, loci of different natures are heterogeneously represented in bacterial genomes, and it is not known if they differently impact evolutionary inferences. In the present study, we evaluated how loci nature shapes ρ/θ and r/m estimates, and we used the generated data to speculate about the validity of using full-genome sequences as the approach to estimate such parameters.

2.5.1. Wide genomic approach

We compiled loci sequences for all 15 existing serovars, encompassing about 55% of all chromosome SNPs (see Supplemental Table 2.3), which is expected to better represent the *C. trachomatis* intraspecies genetic variability. This wide genomic data set was preliminarily used for the assessment of the accuracy of the ClonalFrame analysis by evaluating whether different numbers of iterations (i.e., different durations of the simulation period) yield variable results. In fact, the optimization of the number of iterations is a critical step when performing ClonalFrame analysis. The

software was run with 30,000, 100,000, 300,000, 500,000, and 1,000,000 iterations for evaluating their impact in both r/m and ρ/θ ratio estimations. We found that the highest dispersion of the estimates of both parameters was obtained for the runs using 30,000 and 100,000 iterations, which noticeably affected the mean values, revealing that for a low number of iterations, small variations may markedly bias the estimation of the evolutionary parameters (Figure 2.2). By increasing the number of iterations, there was a tendency toward the stability of the results, as similar values were detected when using 500,000 and 1,000,000 iterations. These runs were also the most reproducible and reliable; thus, all subsequent analyses were run by using 1,000,000 iterations to decrease the putative bias strictly associated with simulation duration. We believe that a preliminary step of optimization is critical and mandatory, despite its large computational cost (50% of the 972 CPU hours dispended in all performed simulations).

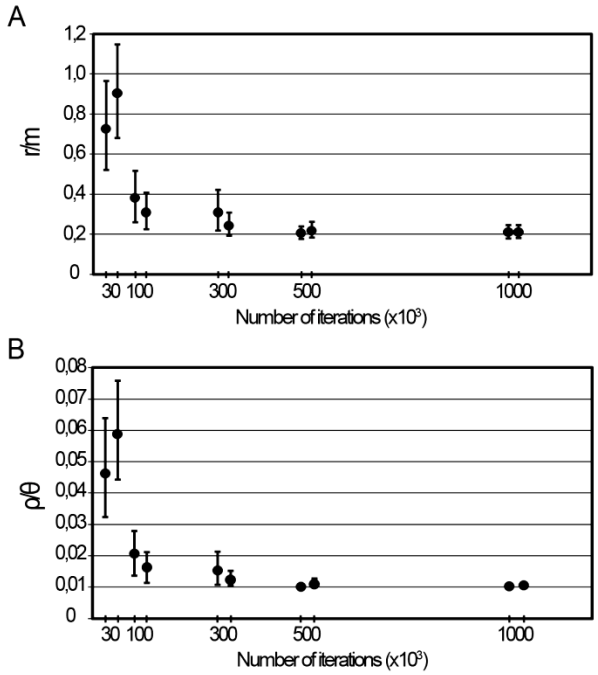


Figure 2.2. Accuracy assessment of r/m and ρ/θ estimations by varying the number of iterations. The figure illustrates the impact of the number of iterations on the estimations of the ratios r/m (A) and ρ/θ (B) inferred from the wide genomic data set. The graphs present the values and respective 95% confidence intervals of the two independent runs performed with the same number of iterations. The stability (graph plateau), reproducibility (the proximity of the mean estimates from replicate runs), and high levels of confidence (narrower error bars) of both r/m and ρ/θ values were reached only for runs using 500,000 and 1,000,000 iterations.

Another critical stage when estimating r/m and ρ/θ relies on ensuring that independent runs yield convergent estimates for all parameters (θ, R, δ, and ν) and thus sustain similar results. For the wide genomic data set, we observed a convergence scenario for all estimated parameters by using the Gelman-Rubin test implemented in the software (Figure 2.3, Supplemental Table 2.4).

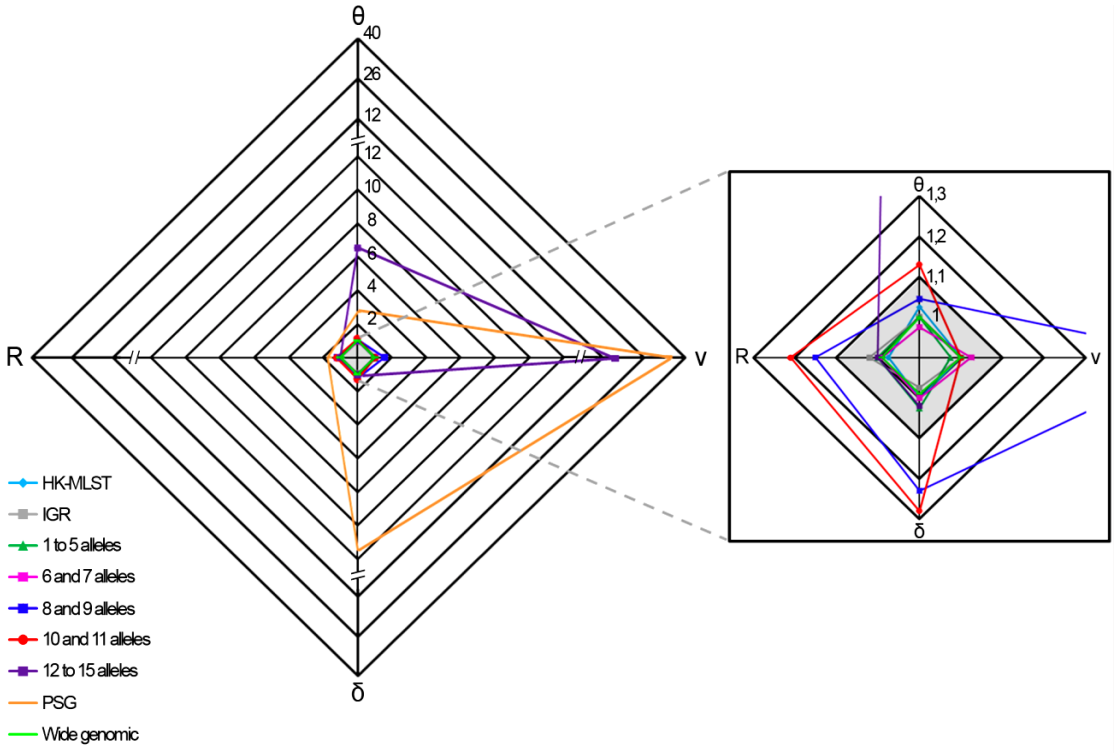


Figure 2.3. Convergence assessment of the parameters θ, ν, δ, and R. For each data set, the graph shows the convergence values from two independent simulations for the estimated parameters θ, ν, δ, and R. The shaded region of the graph (amplified on the right) indicates the satisfactory range of values (below 1.1) of the test statistic for all parameters according to the Gelman-Rubin test. For the data sets PSG (orange), "8 and 9 alleles" (dark blue), "10 and 11 alleles" (red), and "12 to 15 alleles" (purple), convergence was not observed for at least one parameter.

As a fine-tune evaluation of convergence, we also used the phylogenetic tree comparison tool, which assesses the degree of concordance between trees from replicate runs (Figure 2.4, Supplemental Table 2.4). It is worth noting that the inferred tree for the wide genomic data set had total confidence in all nodes (average concordance score = 3), which, in addition to the accuracy (Figure 2.2) and convergence assessment steps, supports that the ratios r/m and ρ/θ were correctly inferred through the analysis of this data set. The mean estimates of r/m and ρ/θ ratios were 0.21 and 0.01, respectively (Figure 2.5, Supplemental Table 2.4), which seem plausible concerning the unique biology of this bacterium. The low ρ/θ value was expected due to the obligate intracellular life style of *C. trachomatis*. Thus, recombination requires a host-cell coinfection by distinct strains (which is expected to occur at a frequency of 1% [10]) followed by the fusion of the inclusion vacuoles where this pathogen replicates. With respect to the low r/m value, the high genomic similarity degree of different serovars (about 99%) implies that, except for well-described situations [130,166,167,169,170], a recombinant fragment introduces little diversity in the recipient microorganism.

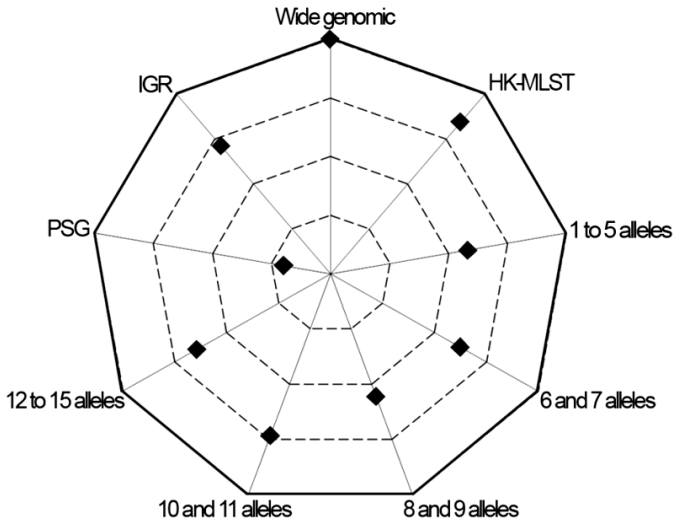


Figure 2.4. Concordance score between phylogenetic trees. The chart presents the average concordance scores between trees of replicate runs calculated for each data set. More external values correspond to higher concordance between trees, and the outer line represents the maximum average score (score = 3). Values were obtained by using the tree comparison tool of the ClonalFrame, which ranks each node of the first consensus tree according to the level of confidence found between the respective nodes of both trees from replicate runs. The color-based qualitative representation of this tool (see Supplemental Figure 2.1) was converted into a quantitative approach as described in Materials and methods to permit the concordance evaluation at the whole-tree level. Only the wide genomic data set reached the maximum average concordance score.

Our estimates using 15 prototype strains are similar to those obtained by Joseph and colleges [173] based on four prototype and eight clinical strains ($r/m = 0.71$ and $\rho/\theta = 0.07$), in which the minor differences may be due to the dissimilar sample sets. Indeed, both results place *C. trachomatis* in the same position (among organisms with low recombination rates) of a r/m "scale" (from 0.02 to 63.6) presented in a previous study that focused on a broad set of bacteria and archaea [172].

2.5.2. HK-MLST

Although the MLST data has been widely used for estimating recombination rates of several bacteria, nonconsensual results have been published [224-228], and they may be strikingly conflicting, as illustrated for *Bacillus cereus* in which different studies reported recombination rates differing up to two orders of magnitude [213,229]. For *C. trachomatis*, a previous study determined a r/m mean estimate of 0.3 based on MLST data [172], which is in agreement with our estimation using the wide genomic approach, although the authors reported wide 95% CIs (0.0–1.8). Therefore, we decided to test a more recent MLST system [211] for comparison purposes. We obtained r/m mean values of 1.09 and 0.91, and ρ/θ mean values of 0.12 and 0.14 (Figure 2.5, Supplemental Table 2.4) from convergent and reproducible replicate runs according to both the Gelman-Rubin test (Figure 2.3, Supplemental Table

2.4) and the phylogenetic tree comparison (Figure 2.4, Supplemental Table 2.4), despite the wide CIs that hamper precise estimations (Figure 2.5, Supplemental Table 2.4).

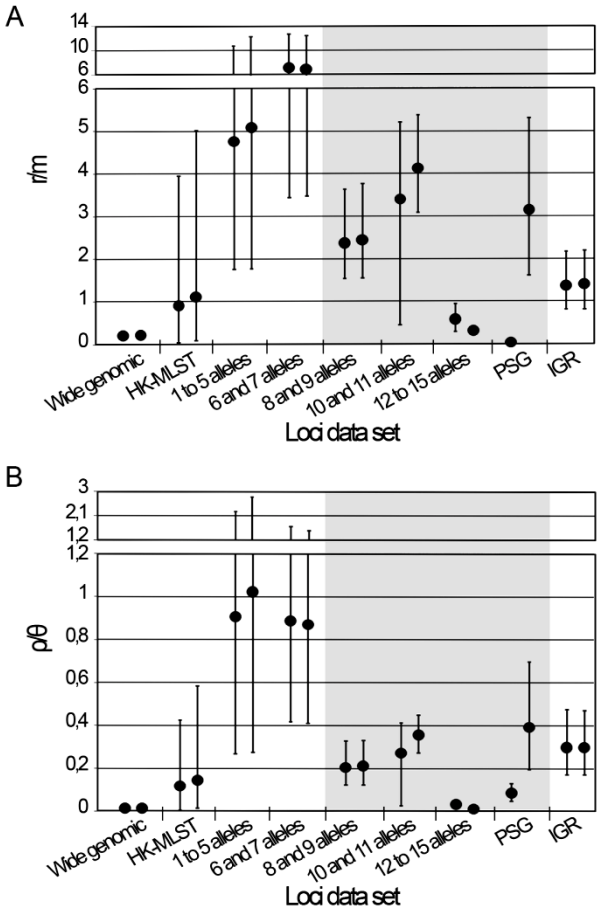


Figure 2.5. Estimates of r/m and ρ/θ. The graphs show the estimates of r/m (A) and ρ/θ (B) ratios calculated by the ClonalFrame software. For each data set, the results (mean and respective 95% CIs) of the two independent runs performed with 1,000,000 iterations are shown. The data sets that yielded nonconvergent runs assessed by the Gelman-Rubin test (see Figure 2.3) are shaded in gray.

Three major issues may underlie the dissimilarity between MLST-based analyses: analytical methodology, strain sampling, and loci selection. As these two analyses using MLST schemes were performed based on ClonalFrame and employed the same set of serovars, we speculate that the loci nature is the major factor influencing estimations. Therefore, MLST data should be applied with prudence when performing this type of evolutionary inference [200], as only a residual proportion of the genome is analyzed (usually 6 to 10 loci of approximately 400 to 600 bp in length [206]), which implies that the whole genetic diversity may not be guaranteed [205]. This is especially relevant in monomorphic organisms, in which the maximum level of variability is extremely low [200]. Nevertheless, the relevance of the application of MLST systems for the characterization of bacterial isolates at the molecular level remains unquestionable.

2.5.3. Allelic profile

MLST systems usually employ genes that assign a low number of alleles. Therefore, we evaluated the impact of increasing the number of alleles per locus on the estimation of mutation and recombination rates, as the level of polymorphism could shape the results differently. Independent runs were not convergent with the three data sets involving loci that define the highest number of alleles (8 and 9, 10 and 11, and 12 to 15) (i.e. Gelman-Rubin statistic above 1.1 for at least one parameter) (Figure 2.3, Supplemental Table 2.4), and thus the parameters are poorly estimated by the software, resulting in inaccurate inferences of r/m and ρ/θ ratios (Figure 2.5, Supplemental Table 2.4). For the two groups of genes assigning a low number of alleles (1 to 5, and 6 and 7), the replicate simulations were convergent and reproducible, but they yielded a high dispersion of both ratios estimates. Moreover, these results contrasted with our estimations using the wide genomic data set and pointed to an implausible scenario of an excessive weight of recombination on genetic diversity of *C. trachomatis* (r/m mean ratios higher than 4 for the two groups) (Figure 2.5, Supplemental Table 2.4). Globally, we found that the level of polymorphism definitely affects the estimations of r/m and ρ/θ at both heterogeneity of results and confidence level. In particular, loci presenting high mutation rates are more prone to confound the estimations, which makes sense considering that an excessive polymorphism is expected to mask the haplotype structures that have evolved over time, making it difficult to analyze the presence or absence of recombination [207].

2.5.4. Positively selected genes

The detection of genes under positive selection has been of great importance for clarifying the evolutionary history of bacteria, as they encrypt adaptive signatures that may underlie phenotypic differences, such as those related to pathogenicity [209,230]. However, it has been assumed that PSGs should not be used to infer recombination rates, in spite of the fact that their unsuitability has not been validated experimentally. The rational for their exclusion is that PSGs likely present an unusual number of changes, and the fixation of mutations due to selection could be confounded with their acquisition through a transferred recombining fragment [206]. In fact, recombining fragments may bring together beneficial mutations that allow a faster increase in fitness in the presence of major environmental changes instead of solely accumulating point mutations through positive selection [231]. It is also known that recombination is increased in the proximity of positively selected regions [230,231], as demonstrated, for instance, for the genus *Streptococcus* [232]. In the present study, we tested a data set composed exclusively of genes putatively under positive selection [173,209]. The evaluation of accuracy revealed lack of convergence for all parameters (values highly above the acceptable cut-off) (Figure 2.3, Supplemental Table 2.4), and the PSG data set was the bottom-ranked group in analysis of the concordance between trees from independent runs (Figure 2.4, Supplemental Table 2.4). Consequently, we found that this data set presented unreliable (wide 95% CIs) and the least reproducible

results, which is reflected by the discrepant mean estimate values between runs differing up to two orders of magnitude (Figure 2.5, Supplemental Table 2.4). These results suggest that, for genomes subjected to strong selective pressures, estimations of recombination rates may be biased by the presence of a high fraction of PSGs. Nevertheless, because it is known that PSGs are also targets of recombination [231], we believe that, for the majority of the bacterial genomic contexts, the use of wide genome approaches will likely buffer the confounding effects of PSGs on estimations. In fact, in the present study, the inclusion of PSGs in the wide genomic approach did not hamper the accurate inferences of the evolutionary parameters.

2.5.5. Intergenic regions

The IGRs have been excluded for inferring evolutionary histories of organisms, although they are known to carry promoter regions, ribosome binding sites, as well as transcription factor and regulator binding regions, which play critical roles in regulation of gene transcription. Recent studies demonstrated that noncoding regions are subject to significant selective constraints [215,216]. For *C. trachomatis*, we previously detected recombination hotspots involving IGRs [167], and we observed phylogenies of IGRs revealing the clustering of strains with the same disease outcomes [32], which suggest selection or hitchhiking events [233] involving these regions. This evidence, together with the knowledge that the small genome of *C. trachomatis* likely retains only the indispensable genes [124], points to a relevant role of IGRs in *C. trachomatis* evolution. Thus, we estimated rates of recombination and mutation using 56 IGRs because the accumulation of mutations in these regions may not be a random process and because they are heterogeneously represented in different genomes. We obtained ~90% of concordance between trees, and a Gelman-Rubin test statistic below 1.1 for all parameters (Figures 2.3 and 2.4, Supplemental Table 2.4), indicating convergence. The r/m and ρ/θ mean estimates (Figure 2.5, Supplemental Table 2.4) are about 1-log above those obtained for the wide genomic data set, but they are similar to the HK-MLST data set estimates, which suggests that this large set of noncoding regions and these specific HKs shape these evolutionary parameters in a similar fashion for the model under evaluation.

2.6. Conclusion

We used a specific human pathogen with well-defined genomic characteristics as a model to study bias associated with the estimation of evolutionary parameters by computational simulations. Our results show that the estimation of mutation and recombination rates in *C. trachomatis* is influenced by the characteristics of the loci used for such calculations. Although the use of full-genome sequences to infer recombination and mutation rates is suitable for most microorganisms, we anticipate that soon a greater proportion of highly polymorphic or positively selected loci can make it an inaccurate approach. Thus,

the correctness of the final output will depend on the dilution effect of these confounding factors by the remaining portions of the genome with dissimilar architectures. As data from population genetics has contributed to a better understanding of the biology and pathogenicity of organisms, the clarification of the putative bias associated with *in silico* inferences is of great interest for deciphering evolutionary traits.

Acknowledgments

This work was supported by a grant, ERA-PTG/0004/2010, from Fundação para a Ciência e a Tecnologia (FCT/MEC) (to J.P.G.), in the frame of ERA-NET PathoGenoMics. R.F. and V.B. are recipients of Ph.D. fellowships (SFRH/BD/68532/2010 and SFRH/BD/68527/2010, respectively) from FCT/MEC. A.N. is a recipient of a post-doctoral fellowship (SFRH/BPD/75295/2010) from FCT/MEC. We are grateful to Karol Dobrzanski for providing the Linux server.

Chapter III

In Silico Scrutiny of Genes Revealing Phylogenetic Congruence with Clinical Prevalence or Tropism Properties of *C. trachomatis* Strains

This chapter corresponds to a manuscript (with discrete changes) with the following reference:

Ferreira R, Antelo M, Nunes A, Borges V, Damião V, Borrego MJ and Gomes JP (2015) *G3 (Bethesda)*. 5:9-19.

This study was also presented as a section of the Master's thesis developed by Antelo, M.

Personal contribution

RF performed the majority of the bioinformatic analyses, and wrote the paper.

3. *In silico* scrutiny of genes revealing phylogenetic congruence with clinical prevalence or tropism properties of *C. trachomatis* strains

3.1. Abstract

Microbes possess a multiplicity of virulence factors that confer them the ability to specifically infect distinct biological niches. Contrary to what is known for other bacteria, for the obligate intracellular human pathogen *C. trachomatis*, the knowledge of the molecular basis underlying serovars' tissue specificity is scarce. We examined all ~900 chromosomal genes to evaluate the association between individual phylogenies and cell-appetence or ecological success of *C. trachomatis* strains. Only ~1% of the genes presented a tree topology showing the segregation of all three disease groups (ocular, urogenital, and lymphatic) into three well-supported clades. Approximately 28% of the genes, which include the majority of the genes encoding putative type III secretion system effectors and Inc proteins, present a phylogenetic tree where only lymphogranuloma venereum strains form a clade. Similarly, an exclusive phylogenetic segregation of the most prevalent genital serovars was observed for 61 proteins. Curiously, these serovars are phylogenetically cosegregated with the lymphogranuloma venereum serovars for ~20% of the genes. Some clade-specific pseudogenes were identified (novel findings include the conserved hypothetical protein CT037 and the predicted α-hemolysin CT473), suggesting their putative expendability for the infection of particular niches. Approximately 3.5% of the genes revealed a significant overrepresentation of nonsynonymous mutations, and the majority encode proteins that directly interact with the host. Overall, this *in silico* scrutiny of genes whose phylogeny is congruent with clinical prevalence or tissue specificity of *C. trachomatis* strains may constitute an important database of putative targets for future functional studies to evaluate their biological role in chlamydial infections.

3.2. Keywords

Chlamydia trachomatis; Genomics; Clinical prevalence; Tropism; Loci.

3.3. Introduction

The observation that there are pathogenic and nonpathogenic microbes has compelled investigators to search for traits underlying their phenotypic differences. This search for the so called "virulence factors" has greatly contributed to the understanding of pathogenicity and to the elucidation

of the genetic mechanisms underlying microbes' capability to infect different cell types or organs. The notion that microbial pathogenicity relies on the interaction between a pathogen and its host (or a specific tissue), and that a virulence factor is either a microbial product or a strategy capable of causing damage to a susceptible host, can be broadly applied [234]. In this perspective, virulence factors may involve an endless list of products and mechanisms, such as toxins, adhesins, motility structures like flagella and pili, immune evasion determinants, capsules, biofilms, secretion systems, and signal transduction mechanisms (reviewed in [234]. Usually, microbes carry several of these virulence factors, which work together in the process of host invasion and microbe survival. Among pathogenic agents, several bacteria present intracellular lifestyles (obligatory or facultative). Their host-cell targets range from epithelial cells to phagocytes, like macrophages and neutrophils [235], which implies that these pathogens have been developing specialized strategies that allow them, for instance, to survive within or avoid the adverse environment of the macrophage phagosome (membrane-bound vacuole) [236,237]. Whereas some bacteria (e.g., *Salmonella* spp, *Coxiella burnetii*, and *Cryptococcus neoformans*) are able to reside within the lysosomal vacuole, others (e.g., *C. trachomatis* and *Mycobacterium* spp) need to "remodel" it to allow their survival, whereas others (e.g., *Listeria monocytogenes* and *Shigella* spp) degrade the vacuole membrane to gain access to the host-cell cytosol, where they may complete their developmental cycle [237,238]. Moreover, some pathogenic bacteria are also able to infect different cell types or organs of a given host. For example, *L. monocytogenes* can cross the intestinal epithelium, the blood–brain, and fetoplacental barriers [239] and may cause severe septicaemia and meningoencephalitis [240], whereas *Streptococcus pneumoniae* is capable of infecting the lung, the blood, and the nasopharynx [241].

Another example of bacteria capable of infecting different cell types is *C. trachomatis*, an obligate intracellular human pathogen that can be classified into 15 main serovars, according to the polymorphism of the gene (*ompA*) encoding the MOMP. Serovars A-C cause ocular infections that can progress to trachoma, the leading cause of preventable blindness worldwide [242,243], whereas serovars D-K cause ano-urogenital infections that can evolve to cervicitis, urethritis, epididymitis (men), or pelvic inflammatory disease (women), the latter of which can lead to significant long term sequelae such as infertility and ectopic pregnancy [244]. Finally, serovars L1-L3 are responsible for an invasive disease, the lymphogranuloma venereum (LGV), through the infection of macrophages and dissemination to regional draining lymph nodes [245]. Despite the huge phenotypic differences among *C. trachomatis* serovars regarding tissue tropism, virulence and ecological success, little is known about the molecular factors underlying serovars' biological uniqueness. This is mostly due to the lack of suitable animal models that mirror the human chlamydial infection *in vivo* and because *C. trachomatis* has been genetically intractable until very recently [155-157,246]. Probably the only unequivocal demonstration of the association of a virulence factor with tropism was provided by Caldwell and colleges [152], who showed that an active tryptophan operon (*trpRBA*) is mandatory for any *C. trachomatis* strain to infect the genitalia. This observation also was valid for genital strains harbouring an "ocular" *ompA* gene (likely inherited by recombination), excluding the serovar status as a possible

tropism determinant. Nevertheless, a revision concerning the genetics beyond tropism was recently published [129].

Recent phylogenetic analysis [30] using the complete genome of several *C. trachomatis* strains found: *i*) the segregation of strains by their cell-appetence, suggesting a coevolution with the infected tissue; *ii*) the separation of the LGV strains before the separation of the ocular and the epithelial-genital strains; *iii*) that the most prevalent serovars (E and F), which account for ~50% of all chlamydial genital infections among the heterosexual population [28], clearly segregate apart from the remainder epithelial-genital strains; and *iv*) that the ocular strains probably derived from a nonprevalent genital serovar. On the other hand, the small genome (~1 Mbp) of *C. trachomatis* reveals a high degree of conservation among serovars (98%), with nearly identical pan- and core-genomes, a high coding density, and no evidence of recent horizontal gene transfer besides allelic recombination, which suggests a likely complete genetic reduction process as a result of a long-term intracellular niche adaptation process [11,247]. Considering this, one may speculate that the phenotypic disparities (tissue tropism, virulence and ecological success) among strains are encoded in a small number of variable genes along the *C. trachomatis* genome. Thus, given the recent availability of dozens of *C. trachomatis* fully sequenced genomes, our main goal was to scrutinize all the ~900 genes at the phylogenetic and evolutionary level in order to better understand the relationship between strains' genetic diversity and phenotypic disparities. In this regard, after analyzing the global trends of polymorphism, we performed a detailed analysis of each gene tree topology to assess the degree of concordance between strains' segregation and their clinical outcome and prevalence. This approach intends to identify the genes that phylogenetically contribute for the main branches (LGV, prevalent genital, nonprevalent genital, and ocular serovars) of the species tree [30].

3.4. Materials and methods

3.4.1. Alignments generation

For the polymorphism and evolutionary analyses, different alignment strategies were conducted. First, the whole-genome sequences of the 53 studied *C. trachomatis* strains were retrieved from the GenBank (Supplemental Table 3.1) and aligned using progressiveMauve from Mauve software, version 2.3.1 [218]. Orthologous genes were identified by Mauve and individual alignments of each one of the 896 genes (considering the total number of annotated genes on the available D/UW-3/CX sequence) were extracted from the whole-genome alignment. These alignments were subsequently uploaded into the Molecular Evolutionary Genetics Analysis software, version 5 (MEGA 5; http://www.megasoftware.net) [248] and visually inspected for further correction (whenever needed) prior to evolutionary and genetic diversity analyses. A core-alignment was also extracted by keeping regions where the 53 genomes aligned over at least 500 bp (corresponding of ~97% of the *C. trachomatis*

chromosome), and aligned segments were concatenated into a single-core genome alignment to be further used in the construction of the species phylogenetic tree. This alignment was then exported and directly uploaded into MEGA 5 for whole-genome analyses purposes.

3.4.2. Exclusion criteria

Among all the 53 strains, variability in start codon predictions of homologous genes was removed by trimming each start site prediction to the innermost common start codon. This was not applied when an upstream codon was annotated as a consequence of a mutation in the codon correspondent to the translation initiation codon of the other sequences. We also observed that, for some other genes, there were strains that had more than one coding sequence annotated at the same region. These cases were treated as pseudogenes and the respective strains were removed from the analysis. There were also genes for which a single frameshift yielded a biased polymorphism, and for this reason they were not considered has truly polymorphic. Nevertheless, some of them (CT120, CT160, CT162, CT172, CT172.1, CT358, CT480.1, CT793, and CT852) constitute interesting cases as the frameshift occurred solely for the strains of the same disease group. Moreover, for 22 chromosomal genes, it was not possible to obtain an accurate alignment (Supplemental Table 3.2) mainly because of accentuated gene size differences, hampering the analyses.

3.4.3. Polymorphism and evolutionary analyses

Each alignment (core-genome and individual genes) was analyzed according to previously described methods [32,142]. Concerning the individual alignments of all homologous genes, we first removed from each analysis the strains' sequences that were considered as putative pseudogenes or had annotation issues (see the section Exclusion criteria). By using the algorithms available in MEGA 5, we determined the overall mean distances (number of differences and *p*-distance) and matrices of pairwise comparisons at both nucleotide and amino acid level, along with the respective standard error estimates (bootstrap = 1000). Then, for each gene, the number of synonymous substitutions *per* synonymous site (dS) as well as the number of nonsynonymous substitutions *per* nonsynonymous site (dN) were determined by using the Kumar model [249] and the standard error estimates were obtained by a bootstrap procedure of 1000 replicates. dN/dS ratios were determined and the Z-test of positive selection was applied for the genes revealing dN/dS > 1. The probability of rejecting the null hypothesis of strict-neutrality (dN = dS) in favor of the alternative hypothesis of positive selection (dN > dS) was considered significant when $P < 0.05$ (bootstrap = 1000) [249]. We also assessed the existence of correlation between p-distance and dN, dS, or dN/dS by using the Pearson's Product Moment Correlation coefficient (P), which measures the strength and direction of a linear relationship between two variables [250].

Phylogenetic trees for both the whole-genome and individual genes sequences were inferred by using the Neighbor-Joining method (bootstrap = 1000) [251,252]. For the nucleotide sequences, the

evolutionary distances were computed using the Kimura 2-parameter method (K2P) [253], whereas for the amino acid sequences (for individual genes solely), the evolutionary distances were computed based on the number of differences [249]. A gene was considered to segregate a specific group of strains (ocular, genital and LGV serovars) by taking into account both the tree topology and the number of differences between sequences of different taxa. Additionally, phylogenies were also inspected for the segregation of the strains from the most prevalent genital serovars.

3.4.4. Characterization of the mosaic structure of the strains from the most prevalent serovars

We started by comparing the genome sequences of both D(s)/2923 and D/SotonD1 with that of the F/SW5 strain (because this strain was found to be the most closely related to both – see Results section of this chapter) using the DNA polymorphism tool of the DnaSP software, version 5 [254], with a window size and step size of 1000 each. Chromosomal regions with high SNP density, which may indicate the occurrence of recombination events, were further analyzed by SimPlot/BootScan (http://sray.med.som.jhmi.edu/SCRoftware/simplot/) [255,256] for a precise determination of potential mosaic structures. These analyses were performed as previously described [167], using a sliding window size of 200 bp moved across the alignment in a step size of 30 bp for estimating pairwise genetic distances with Neighbor-Joining method (Kimura 2-parameter method; bootstrap = 500; gaps strip off; ts/tv of 2.0). For BootScan analyses, the likelihood that the observed distribution of informative sites [257] favoring specific phylogenetic groupings might occur randomly was assessed using the maximum χ^2 test. A *P*-value for any specified breakpoint was determined by the Fisher's exact test (two-tailed). A Bonferroni multiple correction testing was applied to evaluate the significance of the *P*-values at 95% confidence.

3.5. Results

3.5.1. Polymorphism and molecular evolution analysis

Overall, we were able to analyze ~97.5% (874/896) of all the *C. trachomatis* chromosomal genes. The 22 genes excluded from the analysis (see the section Exclusion criteria of this chapter) comprise five housekeeping genes, the cytotoxin locus, genes encoding 13 hypothetical proteins, two of the phospholipase D endonuclease superfamily gene members (PLDs), and CT081 (Supplemental Table 3.2).

Besides well-known polymorphic genes (CT870/*pmpF*, CT872/*pmpH*, CT681/*ompA*, CT049-CT051), the polymorphism analyses highlighted *CT619* (Table 3.1 and Supplemental Table 3.2) [coding for a putative type III secretion system (T3SS) secreted protein with unknown function] that, to our knowledge, had never been considered before as polymorphic.

Table 3.1. Top five ranking of the most polymorphic *C. trachomatis* chromosomal genes.

Rank	Nucleotide		Amino acid	
	No. differences	*p*-distance	No. differences	*p*-distance
1	CT870/*pmpF* (217.3)	CT681/*ompA* (0.121)	CT870/*pmpF* (72.4)	CT681/*ompA* (0.107)
2	CT681/*ompA* (143.7)	CT051 (0.07)	CT619 (48.4)	CT051 (0.093)
3	CT619 (124.2)	CT870/*pmpF* (0.069)	CT051 (46.6)	CT049 (0.08)
4	CT872/*pmpH* (109)	CT049 (0.048)	CT681/*ompA* (42.1)	CT870/*pmpF* (0.071)
5	CT050 (104.9)	CT619 (0.047)	CT049 (38.9)	CT050 (0.058)

The numbers in parenthesis refer to the respective number of differences and *p*-distance value.

To understand the underlying evolutionary pressures that drove amino acid changes of all 874 analyzed chromosomal proteins, we evaluated their molecular evolution by determining the dN/dS values of the respective genes. We verified that 150 genes (~17%) revealed a dN/dS > 1, but only 31 (3.5%) showed a significant Z-test of positive selection (Supplemental Table 3.2) and were thus considered as putative targets of positive selection. Twenty-three of the latter encode 11 Inclusion Membrane Proteins (Incs), 10 T3SS effectors, and two putative membrane proteins, which are proteins expectedly involved in interactions with the host. We also found three hypothetical proteins encoding genes, one PLD encoding gene, and four housekeeping genes that are likely under positive selection. We have no reasonable explanation for the latter finding, as housekeeping genes are usually highly conserved and expected to be under purifying selection.

Furthermore, we evaluated the correlation between nucleotide polymorphism and evolutionary parameters, such as dN, dS, and dN/dS, for all 874 chromosomal genes. From the inspection of the genomic distribution of *p*-distance and dN/dS (Figure 3.1, A and B) and by determining the Pearson's product moment correlation coefficient, we observed no correlation between them (P = 0.02), besides minor coincident peaks. Figure 3.1C highlights the 25 top ranked loci for both parameters. On the other hand, a strong positive linear correlation between *p*-distance and both dN (P = 0.92) and dS (P = 0.9) was found (Figure 3.1D).

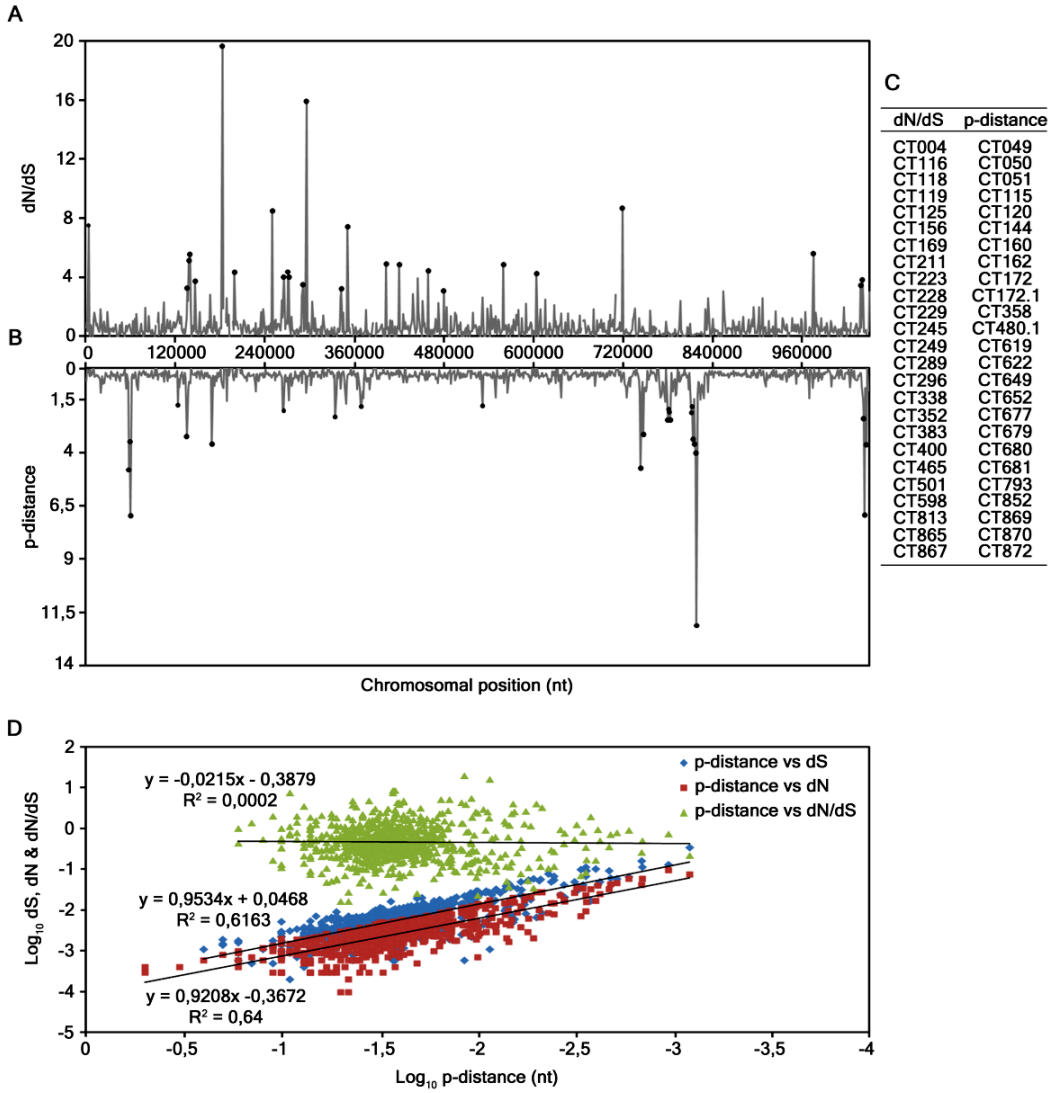


Figure 3.1. Evaluation of the association between polymorphism and dN, dS and dN/dS. (A and B) Distribution of dN/dS and *p*-distance values, respectively, obtained from the analyses of all the 874 genes from the 53 strains. The horizontal axis represents the *C. trachomatis* chromosomal positions where genes are placed in their chromosomal order, from the *CT001* to the *CT875* (genes names and positions according to D/UW-3/CX strain annotation). C) 25 genes (ordered by their relative chromosomal position) that display the greater values for both analyses, which are representative of the lack of correlation between dN/dS and polymorphism. D) Scatter plots of *p*-distance vs. dN, dS, and dN/dS, on a log-scale for clarity. The Pearson's product moment correlation coefficient for *p*-distance vs. dN, dS, and dN/dS are P = 0.92, P = 0.9, and P = 0.02, respectively.

3.5.2. Species polymorphism vs. number of taxa

Considering that the genetic diversity among same-serovar or same-disease group strains was recently pointed out to be higher than expected [30], we wonder whether both the polymorphism and selective pressure results are impacted by the number of sequences used. Thus, besides using all 53 strains, we also selected a group of 17 strains representative of the major branches of the phylogenetic tree constructed with the whole-genome sequences (Figure 3.2).

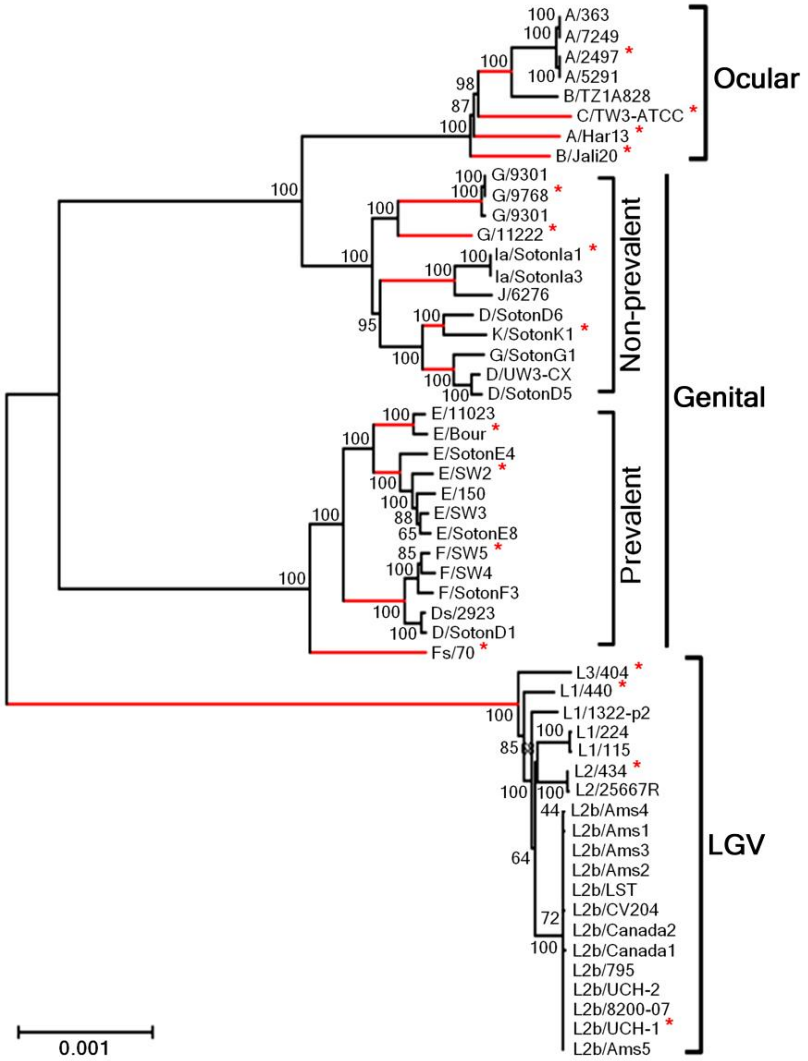


Figure 3.2. Phylogenetic reconstruction of *C. trachomatis* species. The tree was constructed using the whole genome of 53 strains encompassing the majority of the CT681/*ompA* serovars. The asterisks indicate the 17 strains representative of the major tree branches (in red) that were used to evaluate the relation between species polymorphism and the number of taxa (see the Results section of this chapter for details).

Both groups of strains (17 vs. 53) encompass the same set of 13 *C. trachomatis* serovars. We then used the 100 most polymorphic genes (as they provide the vast majority of informative sites) and compared the distribution of polymorphism and dN/dS obtained from the analysis of the two groups (Figure 3.3). The *P*-values (paired two-tailed t-test) calculated for the *p*-distance and the dN/dS results were 0.91 and 0.13, respectively, which indicates that these parameters do not depend on the number of same-serovar sequences that are used. Although the validity of the traditional CT681/*ompA* typing has been strongly questioned (as its tree does not segregate strains by tissue tropism properties and disease outcomes) [30], it is worth noting that a small group of strains encompassing the majority of the *C. trachomatis* serovars represent the main genetic variability of this bacterium.

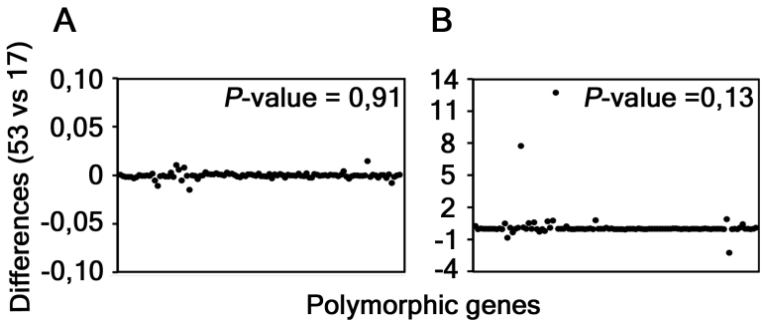


Figure 3.3. Differences obtained during the analyses using 53 and 17 strains. The graphs show the differences obtained between the results of the *p*-distance (A) and the dN/dS (B) analyses of all the 53 and the set of 17 strains (representative of the majority of the tree branches). Each black dot represents one of the 100 polymorphic genes selected for these comparisons. *P*-values were calculated through the paired two-tailed t-test.

3.5.3. Gene-based phylogenetic analysis

To evaluate the concordance between strains' segregation and their clinical outcome and prevalence, we first performed a detailed analysis of the recombination phenomena involving the two D strains (D(s)/2923 and D/SotonD1) that phylogenetically cluster with the most prevalent serovars (E and F) and apart from the other D strains [30,169], in order to define their true genomic backbone. We verified that those D strains differ from the same serovar prototype strain (D/UW3-CX) by ~5500 nucleotides, but differ from a serovar F strain (F/SW5) by only ~300 nucleotides, with ~50% of these mutations concentrated at the CT681/*ompA* region (Figure 3.4A). SimPlot and BootScan analyses identified the exact location of the two breakpoints underlying the recombination event (identical for both strains) (Supplemental Figure 3.1). One breakpoint is located at the beginning of CT680/*rpsB* (P = 9.28 x 10^{-44}) (Figure 3.4B), whereas the other is located at the beginning of CT681/*ompA* (P = 6.65 x 10^{-19}) (Figure 3.4C). These results clearly indicate that both recombinant D strains have a genome backbone of a serovar F strain, whereas solely the region spanning between the two recombination breakpoints was inherited from a serovar D strain. Therefore, from now on these two D serovar strains will be included into the cluster of the most clinically prevalent serovars.

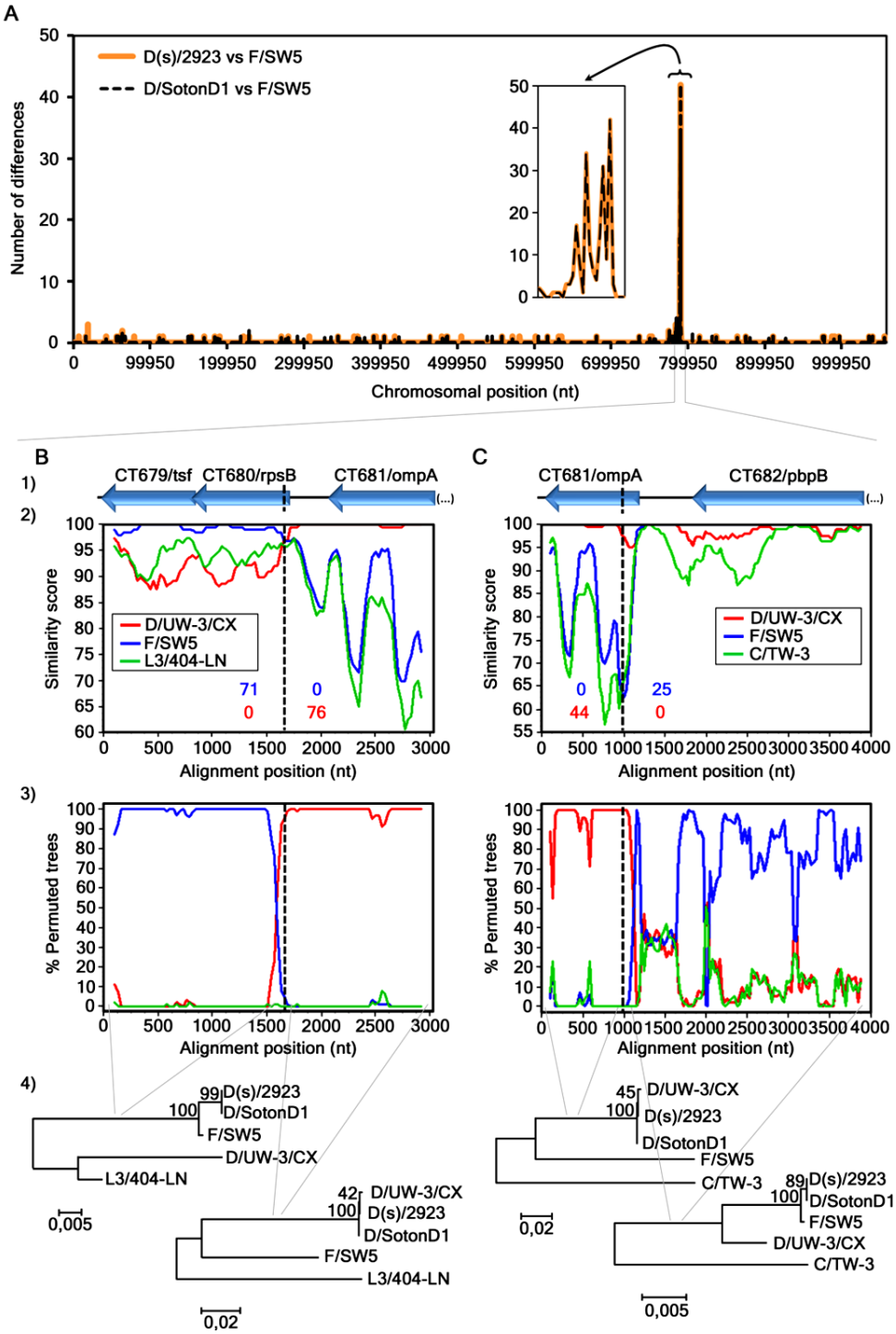


Figure 3.4. Recombination analyses of the D(s)/2923 and D/SontonD1 strains. (A) Number of nucleotide differences (vertical axis) that exist between the genomic sequence of D(s)/2923 or D/SotonD1 and F/SW5. This polymorphism assessment was performed by using the DnaSP software, v5, with a window size and a step size of 1000 bp each. The smaller graph represents an enlarged view of the detected highly polymorphic region. In panel B (first crossover) and panel C (second crossover) are shown the genes in each analyzed region (1) and also the results of the SimPlot (2), the BootScan (3), and the phylogenetic (4) analyses. Recombination breakpoints were individually analyzed because they were better mapped when a different outgroup strain was used for each one, i.e., the L3/404-LN for the first (B) and the C/TW-3 for the second (C) breakpoint.

SimPlot graphs (2) show the level of similarity between the recombinant sequences and the respective parental strains (the number of informative sites supporting this relatedness are colored according to the graph legend box), whereas the BootScan graphs (3) show the phylogenetic relatedness (% of permuted trees) between those same sequences. Both analyses were obtained with a sliding window size of 200 bp and a step size of 30 bp. The sequence of the recombinant D strains was used as query. The vertical dashed black lines indicate the location of the estimated crossovers, shown in detail in Supplemental Figure 3.1. Seventy-one informative sites support the similarity between the recombinant strain and F/ SW5, whereas 76 support its similarity with D/UW-3/CX ($P = 9.28 \times 10^{-44}$). Forty-four informative sites support the similarity between the recombinant strain and D/UW-3/CX, whereas 25 support its similarity with F/SW5 ($P = 6.65 \times 10^{-19}$). In these defined regions there are no informative sites supporting the alternative hypotheses. The phylogenetic trees (4) were constructed with the nucleotide sequences adjacent to each estimated breakpoint region (NJ method; K2p method; bootstrap = 1000) and support the recombination event.

To identify loci that phylogenetically contribute for the main branches of the species tree [30], we performed a detailed analysis of each gene phylogenetic tree. For clarification purposes, a gene/protein was considered to segregate a group of strains sharing a specific phenotype (ocular, prevalent genital, non-prevalent genital and LGV serovars) when the genetic differences among them are lower than the differences to any other strain. Overall, we found that 136, 14, 431, and 695 genes phylogenetically segregate the ocular, genital, prevalent genital and LGV groups, respectively (Figure 3.5A, Table 3.2, and Supplemental Table 3.2).

Table 3.2. Number of genes/proteins that segregate *C. trachomatis* strains according to distinct phenotypes.

	Segregation by phenotype[a]						**Exclusive Segregation by phenotype[b]**			
	Full-tropism[c]	Ocular	Genital[d]	Prevalent Genital	Prevalent Genital + LGV[e]	LGV	Ocular	Genital[d]	Prevalent Genital	LGV
Nucleotide	11 (1.3%)	136 (15.6%)	14 (1.6%)	431 (49.3%)	173 (19.8%)	695 (79.5%)	7 (0.8%)	0 (0%)	47 (5.4%)	245 (28%)
Amino acid	12 (1.4%)	105 (12%)	15 (1.7%)	302 (34.6%)	146 (16.7%)	531 (60.8%)	21 (2.4%)	1 (0.1%)	61 (7%)	240 (27.5%)

The numbers in parenthesis refer to the proportion of genes/proteins, found in each category, relative to the 874 analyzed genes/proteins. LGV, lymphogranuloma venereum; [a] Genes/proteins for which the phylogenetic tree differentiates at least one group of strains in a nonexclusive manner; [b] Genes/proteins for which the phylogenetic tree differentiates only one particular group of strains whereas the remainder are mixed; [c] Genes/proteins for which the phylogenetic tree shows three clades (ocular, genital, and LGV serovars); [d] Refers to all genital strains (prevalent plus non-prevalent serovars); [e] Genes/proteins for which the phylogenetic tree clusters the strains from prevalent genital and LGV serovars in the same clade.

The low number of genes segregating the group of genital serovars reflects the high heterogeneity within this group as a direct consequence of the recombination background affecting mostly these strains

[30] and the existence of distinct polymorphism signatures. An example of the latter stands for the F(s)/70 strain, which was isolated from the cervix and frequently showed a rather unusual polymorphism pattern that did not resemble any of the other 52 strains. Therefore, only 11 (1.3%) of nucleotide trees and 12 (1.4%) of protein trees were found to segregate strains by full-tropism (Figure 3.5A and Table 3.2), where ocular, LGV and all genital (prevalent and nonprevalent) serovar strains are segregated into three main clusters. *In silico* studies have already implicated some of these genes in the different cell-appetence of the strains, namely CT456/*tarp*, CT870/*pmpF*, CT872/*pmpH*, CT115/*incD*, CT116/*incE*, two PLD (CT156 and CT157), and one MACPF domain family protein (CT153) [84,130,143,209]. The remainders include three housekeeping genes (CT106/*yceC*, CT110/*groEL1*, and CT703/*engA*), and genes encoding one T3SS effector (CT161) [258] and one putative inclusion membrane protein (Inc) (CT383) [144] (Supplemental Table 3.2).

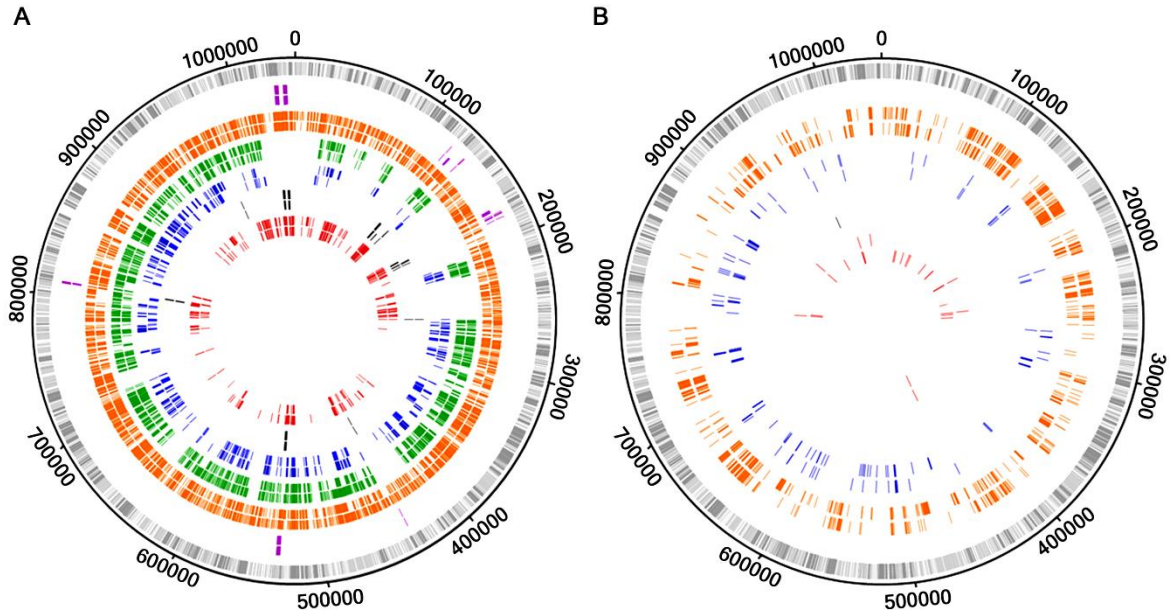


Figure 3.5. Genes that segregate strains according to their biological characteristics. The outer circle in both panels represents the genome of *C. trachomatis* D/UW-3/CX strain, where each bar represents a gene at its respective genomic position (light gray bars, forward strand; dark gray bars, reverse strand). A) The tracks' color scheme represent genes whose phylogeny segregates at least a group of strains according to their biological characteristics, i.e., each color illustrates a particular segregation (that may not be exclusive): full-tropism (purple), LGV strains (orange), strains from prevalent genital serovars (green), cosegregation of LGV and prevalent genital serovar strains (blue), genital strains (prevalent and nonprevalent serovars) (black), and ocular strains (red). B) The tracks' color scheme was maintained for the different groups of strains and represent genes that exclusively segregate a unique group of strains. For both panels, the outer and inner tracks of each color correspond to nucleotide and amino acid results, respectively.

We also detected events of exclusive phylogenetic segregation, i.e., the clustering of a particular group of strains sharing the same phenotype, whereas the remainder strains (regardless of their phenotype) are mixed together. For instance, the group of strains from the most prevalent genital serovars (E, F, and recombinant D strains) are exclusively segregated by 61 proteins, which may contain molecular features that contribute for their higher ecological success. We also observed that the most prevalent and the LGV serovars share hundreds of mutations, particularly in 173 genes (Table 3.2) revealing a major tree branch where these two groups co-segregate apart from the remaining strains. Concerning the LGV group, ~28% of all chromosomal genes exclusively segregate these strains (Figure 3.5B), conferring this group a unique genetic make-up within the species diversity.

Also, based on either the presence of nonsense mutations or the considerable differences in gene size, we scrutinized the genome for the existence of genes that are putative pseudogenes exclusively for a specific disease group (Table 3.3). This set includes: *i*) CT058 (a putative Inc [142]), CT105 (a T3SS effector possibly involved in the cell-appetence of the genital strains [209,258]), *trpRBA* operon [152], and CT374/*aaxC* [259], which are pseudogenes for most ocular strains; *ii*) CT101 (Inc [142]) is a pseudogene for the majority of the genital strains; *iii*) CT473 (predicted a-hemolysin) is a pseudogene for the prevalent genital serovar strains; *iv*) CT373/*aaxB* [259] and CT300 (putative Inc [144]) are pseudogenes for LGV strains [142] (for CT300, this occurs only if one considers the same start codon as that annotated for ocular and genital strains); and *v*) CT037 (conserved hypothetical protein) is a pseudogene for both prevalent genital and LGV serovars strains. This scenario suggests that these genes may be expendable for the *C. trachomatis* infection of specific biological niches.

Table 3.3. *C. trachomatis* known and putative pseudogenes for a particular disease group and genes that present differences in gene length among strains from different disease groups.

Gene	Functional Category	Strains group				Observations	Reference
		Ocular	Non-prevalent genital	Prevalent genital	LGV		
CT037	HP	=	RS	Ψ	Ψ	A/2497, A/363, A/5291 and A/7249 are smaller than the non-prevalent genital serovars.	This study
CT052	Coproporphyrinogen III oxidase	>	>	=	RS		This study
CT058	Putative inclusion membrane protein	Ψ	=	=	RS	A/Har13 and C/TW-3 are not Ψs.	[142]
CT101	Inclusion membrane protein	=	Ψ	Ψ	RS	E/Bour, E/11023 and D/UW3 are not Ψs.	This study and [142]
CT105	T3S effector	Ψ	=	=	RS		[209]
CT106	Predicted pseudouridine synthetase family	=	>	>	RS		This study
CT160	HP	>	>	>	RS	B/Jali20 is a Ψ and F(s)/70 is smaller.	This study
CT161	HP	=	<	<	RS	B/Jali20 and E/SotonE8 are Ψs.	This study
CT162	HP	<	<	<	RS	E/SotonE8, F(s)/70, J/6276, Ia/SotonIa1, Ia/SotonIa3 are Ψs.	This study

CT171	Tryptophan synthase (alpha chain)	Ψ	=	=	RS	B/TZ is not a Ψ.	[152]
CT172	HP	<	<<	<<	RS		This study
CT234	Membrane transport protein from the major facilitator superfamily	=	=	<	RS		This study
CT300	Putative inclusion membrane protein	RS	=	=	Ψ		[142]
CT358	HP	>	>	>	RS	B/Jali20 is smaller.	[142] and [84]
CT373	HP	RS	=	=	Ψ		This study and [259]
CT374	Arginine/ornithine antiporter	Ψ	=	=	RS		This study and [259]
CT392	HP	>	>	>	RS		This study
CT441	Tail-specific protease	<	<	=	RS	Ia/SotonIa1, Ia/SotonIa3 and J/6276 have the size of the LGV and prevalent genital serovars sequences.	This study
CT470	HP	=	=	>	RS		This study
CT473	HP	=	=	Ψ	RS		This study
CT480.1	HP	>	>	>	RS	G/9301, G/9768, G/11074, J/6276, Ia/SotonIa1 and Ia/SotonIa3 are smaller than the remainder strains' sequences.	This study
CT522	S3 ribosomal protein	=	=	<	RS		This study
CT605	HP	>	>	=	RS		This study
CT793	HP	>	>	>	RS	G/9301, G/11074 and G/9768 are Ψs.	This study
CT807	Glycerol-3-P acyltransferase	<	<	=	RS		This study
CT809	HP	<	<	=	RS		This study
CT833	Initiation factor 3	<	<	<	RS		This study
CT852	YhgN family	<	<	<	RS		This study
CT868	Membrane thiol protease (predicted)	>	>	>	RS		This study

The differences in sequence length shown only refer to differences in termination between strains. Genes with discordant 5' annotation, for which the correct start codon lacks confirmation, were not included. The differences in length do not contemplate indel events; "Ψ" – Sequences considered as pseudogenes; "RS" – The sequence whose size was used for reference purposes. LGV sequences were used by default except for LGV pseudogenes; "=" – Gene of the same size as the reference; ">" – Gene larger than the reference; "<" – Gene smaller than the reference; "<<" – Gene with the smallest size. Three sequence sizes were observed for CT172, depending on the disease group.

3.6. Discussion

Phylogenetic studies in *C. trachomatis* have been extensively performed on dozens of genes. Given the recent availability of more than 50 genomes, we sought to perform comparative genomics to examine all the ~900 *C. trachomatis* chromosomal genes. We aimed to evaluate the degree of concordance between strains' segregation and their clinical outcome and prevalence. In fact, the molecular basis underlying tissue specificity in *C. trachomatis* remains to be elucidated, although it is believed that it may rely on SNPs or small indel events in specific genes [129], given the tremendous

genome similarity (> 98%) among sequenced strains. It is known that there are biases associated with phylogenetic-based inferences ("phylogenetic dependence"), such as the weight of neutral mutations in the tree topology. Nevertheless, there are well-built examples in the literature where tree topology of *C. trachomatis* genes seems to be associated with niche specificity. This is the case of Tarp (Translocated actin-recruiting protein), for which distinct functional domains that are variable in number across serovars from different disease groups likely lead to differences in the host-cell actin-driven uptake of *Chlamydia* and to differential activation of diverse signaling pathways (like the Rac/WAVE2/Arp2/3 cascade and the humoral and cellular immune signaling pathways) [101,260,261]. Another relevant example is provided by most Incs, which may be associated with infection of mononuclear phagocytes due to the existence of specific mutational patterns leading to the phylogenetic segregation of LGV strains and to the higher expression of some Incs in these strains [142,143]. In this regard, although our phylogenetic approach certainly carries associated biases, the identification of genes that phylogenetically contribute for the main branches (LGV, prevalent genital, non-prevalent genital and ocular serovars) of the species tree may be highly relevant for future functional studies.

Overall, only ~1.4% (12/874) of the proteins was found to present a plain segregation of strains according to their tissue tropism (ocular conjunctiva, genital epithelium, and lymph nodes). This low number is probably due to the existence of intra- and intergenic recombination events that take place during mixed infections (believed to occur at a frequency of approximately 1% [10]), essentially involving the genital strains [30]. Although *C. trachomatis* is known to have a low population-level recombination rate based on the frequency and relative weight of recombination and mutation events [31,173,262], recombination has been detected, even among different disease-causing strains, and hotspots have been identified [30,167]. The biological role of some of these proteins has already been assessed [145,146,263], but with exception of the above cited TARP, only a single serovar was tested, hampering any implication in tropism. On the other hand, it is possible that each of the corresponding genes is simply evolving more quickly than the genome average (quite likely due to host pressures). A radically different scenario is found for the lymph nodes niche, as the majority of the genes (~80%) segregate the LGV strains and 28% (245/874) segregate them in an exclusive manner (Figure 3.5 and Table 3.2). This corroborates the early divergence of these strains [14] and/or their fastest evolutionary nature. The latter hypothesis may rely on the fact that the LGV strains must be capable of establishing a wider set of molecular interactions and be subject to additional selective pressures, given their ability of infecting two distinct cell-types (epithelial and mononuclear phagocytes). It is worth noting that the majority of the genes encoding T3SS effectors and Incs (known and putative) segregate the LGV strains. One interesting example is CT144 that codes for a putative substrate of the T3SS [258] and is likely involved in the "men who have sex with men" epidemiological sexual network [264], for which most of LGV-specific polymorphisms are concentrated in ~150 bp on the first half of the gene [32], highlighting this specific region as the one hypothetically involved in the interaction with the host cell. Another example comes from the well-studied T3SS effector TARP for which the enhanced

phosphorylation found in LGV strains was shown to additionally promote high affinity interactions with proteins associated with the immune signaling pathways [260], likely explaining the capacity of these strains to cross the mucosa epithelium and to infect mononuclear phagocytes.

We also observed that half of the *C. trachomatis* genes segregate strains of the most prevalent genital serovars, where 61 encode proteins displaying a mutational pattern that is exclusive of these strains. The majority of these genes (33/61) encode proteins that mediate basic cellular functions, like some redox reactions (CT078/*folD*, CT278/*nqrB*, CT539/*trxA*, and CT745/*hemG*), structural ribosomal proteins (CT125/*rplM*, CT506/*rplQ*, CT511, CT523/*rplV*, CT525/*rplB*, and CT787/*rpsN*) and proteins intervenient in the translation process (CT193/*tgt*, CT437/*fusA*, and CT851/*map*). However, given the high representation of these functional categories in *C. trachomatis* genome, we can hardly assume that specific metabolic functions underlie the higher clinical prevalence of strains from serovars E and F. Nevertheless, it seems clear that these serovars share a singular genomic makeup. In fact, two recombinant strains classified as serovar D that cluster in the same branch as E and F are actually F-like strains, and so, the branch of the most ecological succeeded serovars involve exclusively taxa with E or F backbone.

Curiously, we also found that 173 genes (19.8%) cosegregated the strains from the most prevalent genital serovars and the LGV strains. Some relevant examples refer to CT651, a possible virulence factor since it is under the regulation of *C. trachomatis* plasmid [162], and CT338 and CT619, two T3SS substrates [258,265]. Possible explanations for the existence of hundreds of shared polymorphisms between these two groups include: *i*) incomplete lineage sorting, where several unresolved polymorphisms would have been accumulated in the common ancestral before the clades' separation of the current species tree [266]; *ii*) recombination, although it cannot fully explain this scenario as the genetic exchange between these two groups has been recently demonstrated to be restricted to limited genomic regions [30]; and *iii*) short coevolutionary process between E/F and LGV strains before the separation of the latter. One may speculate that some of the shared polymorphisms could endow "invasive" properties to the most prevalent serovars strains. If that would be the case, it would mirror for instance the infection scenario of *L. monocytogenes*, which is capable of surviving within macrophages and also replicating in a variety of nonphagocytic cells [267]. Therefore, one could hypothetically identify E or F strains during recent LGV outbreaks in Europe and USA. However, full-genome sequencing was not performed for all strains identified in those outbreaks and, to our knowledge, no E and F strains were identified so far. Thus, no immediate assumption can be made concerning specific phenotypes conferred by the related mutational pattern in those 173 genes between E/F and invasive strains.

We also identified several putative pseudogenes occurring in different strains (Table 3.3). The most interesting cases were the genes that were truncated only for strains of the same disease group, as it may be an indication of their expendability for the infection of a specific niche. We highlight the CT473, a lipid droplet-associated protein (Lda3) found to be translocated into the host cell cytoplasm

and capture lipid droplets [268], which is likely being lost on the course of the evolutionary process of the strains from prevalent genital serovars, and the CT037 (conserved hypothetical protein), which is a pseudogene in both the prevalent genital and LGV serovar strains. Although we have no clues about the importance of maintaining a functional protein in the clades where these genes are not pseudogenes, it was already demonstrated for example that a functional *trpRBA* operon is mandatory for any strain to infect the genitalia [152]. Also, we have previously shown that the positively selected gene CT105 (a pseudogene for ocular strains) has a significant overrepresentation of nonsynonymous mutations when comparing sequences between urogenital and LGV strains [209], indicating that it has been evolutionarily diverging toward niche-specific adaptation. The identification of niche-specific pseudogenes may be indicative that further genome reduction may still be ongoing in *C. trachomatis*, leading to the future disappearance of those sequences from the chromosome. We also identified genes with differences in sequence length according to strains phenotype. For instance, both CT833 (translation initiation factor) and CT852 (integral membrane component) have longer sequences for all LGV strains, making them interesting targets for future evaluation, as the gene length may have a differential impact on the protein functionality. Additional analyses are now being performed at our lab in order to better characterize this complete set of genes (Table 3.3).

The analysis of polymorphism and dN/dS revealed no correlation between the two parameters, indicating that positive selection is highly targeted on specific genes or gene regions, or acts on strains with specific cell-appetence [209]. Although our analysis was focused on whole genes (leading to an underestimation of positive selection), it is notable that the genes with significant dN/dS > 1 were mainly *incs* and T3SS effectors encoding genes, whereas the most polymorphic ones code essentially for membrane and hypothetical proteins. This seems to corroborate the assumption that proteins involved in strict pathogen-host interactions during the infection process are more prone to fix nonsynonymous mutations, as previously reported in smaller scale studies [142,209]. On the other hand, polymorphism seems to be more pronounced in genes of other functional categories and may be due to discrete genetic drift, as most of the polymorphism is given by dS.

Finally, despite the controversial use of the traditional *ompA*-based typing method, it is worth noting that the main genetic variability within the *C. trachomatis* species is given by the different serovars, where additional strains from the same serovar contribute with few novel polymorphisms (driven either by drift or positive selection) that may impact the individual gene phylogenies (Figure 3.3).

As concluding remarks, our approach allowed the identification of genes that phylogenetically segregate strains according to specific phenotypes, namely the infection of the ocular tissue, the genitalia, the lymph nodes, as well as their clinical prevalence. It will certainly constitute an important database for prioritizing the targets for functional studies that are mandatory to clarify both their biological role and their involvement in tissue tropism, virulence and ecological success.

Chapter III

Acknowledgments

This work was supported by a grant, ERA-PTG/0004/2010, from Fundação para a Ciência e a Tecnologia (FCT/MEC) (to J.P.G.), in the frame of ERA-NET PathoGenoMics. A.N. is recipient of a FCT post-doctoral fellowship (SFRH/BPD/75295/2010), V.B. and R.F. are recipients of Ph.D. fellowships (SFRH/BD/68527/2010 and SFRH/BD/68532/2010, respectively) from FCT/MEC, and V.D. is a recipient of fellowship on behalf of the grant ERA-PTG/0004/2010.

Chapter IV

Assessment of the load and transcriptional dynamics of *C. trachomatis* plasmid according to strains' tissue tropism

This chapter corresponds to a manuscript (with discrete changes) with the following reference:
Ferreira R, Borges V, Nunes A, Borrego MJ and Gomes JP (2013) *Microbiol Res*. 168:333-339.

Personal contribution
RF performed most of the experimental work, contributed to the data interpretation and wrote the paper.

4. Assessment of the load and transcriptional dynamics of *C. trachomatis* plasmid according to strains' tissue tropism

4.1. Abstract

C. trachomatis maintain a conserved plasmid, which is a primary regulator of chromosomal genes, but there is no experimental evidences associating it with the strains' differential tissue tropism (ocular and genital mucosae, and lymph nodes). We investigated if the number of plasmids *per* strain correlate with expression profiles of plasmid ORFs and sRNAs, and also if these molecular features underlie tropism dissimilarities. We performed absolute and relative qPCR to determine both the plasmid load and expression throughout *C. trachomatis* development. Our findings suggest that plasmid load (never exceeding 8 copies) is not a function of expression needs and does not reflect tissue tropism. However, for most ORFs, ocular strains presented lower expression than genital or lymphogranuloma venereum (LGV) strains, and ORF6/*pgp4* (transcriptional regulator of virulence associated genes) presented the highest mean expression among strains, followed by the virulence factor ORF5/*pgp3* (also regulated by ORF6/*pgp4*). More, the mean expression levels of the sRNA-2 (anti-sense to ORF2/*pgp8*) were up to 100-fold higher than those of the ORFs, and up to 12-fold higher than that of sRNA-7 (anti-sense to ORF7/*pgp5*) for the LGV strains. Overall, besides the known regulatory role of *C. trachomatis* plasmid, its transcriptional dynamics sustains tropism differences.

4.2. Keywords

Chlamydia trachomatis; Expression; Plasmid; sRNA

4.3. Introduction

Bacterial plasmids are naturally occurring self-replicating, extrachromosomal genetic elements that are stably maintained at their characteristic copy number (one or several hundred per cell [269] within a given host under fixed bacterial growth conditions [270]), due to their capacity of controlling their concentration by regulating the replication rate [271,272]. Usually, the small high copy number plasmids are randomly segregated into daughter cells, during host cell division, while the stably partition of the large low copy number plasmids is reached by mechanisms of plasmid multimer resolution, active partitioning, and postsegregational killing [273-277]. Bacterial plasmids bear genes considered not to be essential for host cell survival, but confer selective advantages for survival in specialized

environments [195] such as resistance to antibiotic and toxic heavy metals, virulence (turning the host into a pathogenic bacterium), novel ecological interactions, and host enhanced nutritional ability [269,278].

C. trachomatis is an obligate intracellular bacterium that is responsible for several diseases in humans, where serovars A to C are associated with trachoma, serovars D to K infect the urogenital tract, and serovars L1–L3 are generally more invasive and cause the LGV. It presents a unique biphasic developmental cycle of up to 48 h that alternates between the infectious EB and the metabolically active RB, which replicates inside a host-vacuole termed inclusion. Apart from the highly conserved chromosome (>98% of similarity among strains) [30], *C. trachomatis* naturally harbours a ~7.5 kb plasmid, which was firstly identified in 1980 [126]. This plasmid is also highly conserved among strains [30,174] and possesses eight ORFs that are known to be transcribed and translated [184,190,196], as well as two sRNAs [186,190]. The presence of the highly conserved plasmid among several chlamydial species suggests that it was acquired early in the evolution of Chlamydiae and must be subjected to a strong selective pressure for its maintenance by these bacteria that shows a reduced chromosome size, given their intracellular nature [124,125]. In fact, the occurrence of *C. trachomatis* plasmidless strains, both *in vivo* and *in vitro* is considered a rare phenomenon [176,177,181,279]. Recent infectivity studies in *C. muridarum*, its most closely related organism (~82% of chromosome homology; [18] revealed that the plasmid enhances the pathogenesis of infection [23,178]. In *C. trachomatis*, it was shown that the plasmid is associated with the accumulation of glycogen in the inclusion [177], and in a nonhuman primate model of trachoma, ocular infection with plasmidless strains elicited higher protective immunity when compared with that of plasmid-bearing strains [280]. Moreover, the plasmid is a transcriptional regulator of virulence associated genes [162,180,281], where unprecedented deletion mutagenesis assays pointed ORF6/*pgp4* as the primary intervenient [162].

Despite these significant contributions for the elucidation of the plasmid biological role, it is not known if the plasmid contributes to the dissimilar tropism presented by *C. trachomatis* strains. Considering that its high degree of genetic similarity does not seem to be instructive about such phenotypic differences, we aimed to study this issue on a transcriptomic-based approach. In particular, we investigated if the expression profiles of plasmid ORFs and sRNAs, as well as the number of plasmids *per* strain differentiate strains with dissimilar tropism. We also evaluated the putative relationship between the expression levels and the plasmid load.

4.4. Materials and methods

4.4.1. Polymorphism analyses of plasmid ORFs

Although it is assumed that the sequence of *C. trachomatis* plasmid is highly conserved among strains, we performed an overall polymorphism analysis of plasmid ORFs by using all 44 sequences

available at GenBank, in order to evaluate the polymorphism distribution. The analysis was assessed through the Molecular Evolutionary Genetics Analysis software, version 5 (MEGA 5; Tamura et al., 2011; http://www.megasoftware.net), according to previously described models [209]. We determined the *p*-distance for all ORFs based on both the nucleotide and amino acids differences, and we also estimated the number of synonymous substitutions *per* synonymous sites (dS), and the nonsynonymous substitutions *per* nonsynonymous sites (dN) (Kumar model; [249]. For dN/dS > 1, the Z-test of positive selection was applied and values of *P* less than 0.05 were considered significant.

4.4.2. Culture and harvesting of C. trachomatis strains

We used seven *C. trachomatis* strains representing the three disease groups: two ocular strains (prototype strains B/Har36 and C/TW3), two genital strains (the prototype strain E/Bour and the clinical isolate F/CS465-95), and three LGV strains (prototype strains L2/434 and L3/404, and the clinical isolate L2b/CS19-08). All strains were cultured by standard methods. Briefly, each strain was inoculated into five T_{25} cm^2 flasks of confluent HeLa 229 cell monolayers by centrifuging for 1h at 34 °C, followed by an incubation stage of 1h at 37 °C, in a 5% CO_2 atmosphere. Cell culture medium was then discarded and fresh enriched medium (MEM 10% foetal bovine serum, vitamins, non-essential aminoacids, glucose and 0.5 µg/ml cycloheximide) was added to the cultures prior to incubation at 37 °C with 5% CO_2 for bacterial growth. For each strain, developmental cycle was monitored by phase-contrast microscopy and was interrupted at time-points 4, 12, 20, 30 and 42h post-infection (pi), for RNA and DNA extraction. Culture medium was rejected and cells were scraped with 1400 µl of PBS at 4 °C; the suspension was then subjected to sonication for 30 s (Vibra Cell, Bioblock Scientific) for eukaryotic cell disruption and chlamydial release. The harvested cells were subsequently subjected to centrifugation (1500 rpm for 2 min) at 4 °C and the supernatant (bacterial suspension) was recovered, homogeneized and rigorously divided into two twin aliquots. One of these aliquots was stored at -20 °C for further DNA extraction and the other was subjected to immediate RNA extraction to guarantee minimal RNA degradation. We opted by using the twin aliquot approach as, according to our experience, simultaneous extraction of RNA and DNA yields non-accurate results (e.g. reduced RNA protection). To ensure accuracy, three biological replicates of each strain were used.

4.4.3. RNA and DNA extraction

Total RNA was extracted from each time-point, as previously described [282]: first, bacterial suspensions were subjected to high-speed centrifugation (14,000 rpm) for 10 min at 4 °C, to obtain a *Chlamydia*-containing pellet which was subsequently resuspended in lysozyme-containing TE buffer. The RNeasy® Mini Kit (Qiagen) was then used according to manufacturer's instructions. During the RNA extraction assay, an on-column DNA digestion, using 30 U of RNase-free DNase (Qiagen), was performed for further removal of residual contaminant DNA. RNA was eluted in 50 µl of RNase-free water, quantified at A_{260} nm, and stored at -80 °C in RNase-free tubes.

DNA was extracted by centrifuging the initially stored (-20 °C) aliquot at 14,000 rpm for 10 min (at 4 °C). The pellet was then resuspended in 200 µl PBS for QIAamp® DNA Mini Kit (Qiagen) extraction, according to manufacturer's instructions. DNA was eluted in 50 µl of AE buffer and stored at -20 °C after A_{260} nm quantification.

4.4.4. Quantification of plasmid copy number and bacterial genomes

For determining the number of chlamydial genomes and plasmids throughout development, we generated standard curves as previously described [214]. Briefly, we cloned an amplified fragment of single-copy genes, *ompA* (from the chromosome) and ORF2/*pgp8* (from the plasmid), into TOPO vectors using the TOPO® TA technology for PCR products (see Supplemental Table 4.1 for primer information), according to manufacturer's instructions. DH5α competent cells (Invitrogen, Carlsbad, CA, USA) were transformed with each construct independently, and were propagated prior to purification of the cloned vector using the Quick Plasmid Miniprep Kit (Invitrogen, Carlsbad, CA, USA), according to package inserts. RNase A was used in order to guarantee the maximal removal of eventual contaminating RNA that could interfere with the absorbance readings. For confirming cloning success, PCR amplifications (using the vector primers – see Supplemental Table 4.1) were performed and amplicons were sequenced. The copy number of each cloned vector was subsequently determined by A_{260} nm measurements, according to the formula: No. vectors/µl = [Avogadro's no. × vector conc. (g/µl)]/MW of 1 mol of vector(g). Finally, for generating standard curves for qPCR assays, eight-serial vector dilutions (from 10^2 to 10^7 vector copies/µl) were prepared for each construct. qPCR was then performed by using the ABI 7000 SDS equipment, SYBR Green chemistry and optical plates and caps (Applied Biosystems, USA). The qPCR mixture consisted of 1× SYBR® Green PCR Master Mix (Applied Biosystems, UK), 400 nM of each primer and 5 µl of each DNA sample, in a final volume of 25 µl. Plates for the absolute quantification of bacterial genomes and plasmids contained both vector standard curves and duplicates of DNA extracted at each time-point, for each determination. All qPCR plates also included "no template" controls. The thermocycling profile was: 10 min at 95 °C followed by 40 cycles of 15s at 95 °C and 1 min at 60 °C. The specificity of the amplifications was confirmed by the observation of the dissociation melting curves, generated by stepwise increase of the temperature from 60 to 95 °C. The mean value obtained in the qPCR assays of plasmid quantification was divided by the respective mean value obtained in the qPCR assays for chlamydial genome quantification in order to determine the number of plasmids/genome (plasmid load) at each time-point for each strain. Finally, by using the bacterial genomes quantification data obtained during the exponential period of the developmental cycle, we determined the growth rate and estimated the doubling time of each strain.

4.4.5. Expression analysis

In order to study the expression profile of each plasmid ORF and sRNA during the *C. trachomatis* life-cycle, we produced cDNA from 2 µl of each RNA sample as previously described [214]. Briefly,

TaqMan® RT Reagents (Applied Biosystems, Branchburg, NJ, USA) were used as follows: 2.5 µM of random hexamers, 5.5 mM $MgCl_2$, 500 µM of each dNTP, 1× RT Buffer, 0.8 U/µl RNase inhibitor and 1.25 U/µl MultiScribe RT, in a final reaction volume of 50 µl. Cycling conditions were: 10 min at 25 °C, 15 min at 42 °C and 5 min at 99 °C. Although we have previously detected similar efficiencies of random hexamers and target specific primers while studying expression of specific genes, we opted to use the former in the present study in order to avoid bias associated with putative dissimilarities in the RT-efficiency of target-specific primers (which could hamper accurate comparison between loci). cDNA was stored at -20 °C until use in DNase-free tubes.

For transcriptional analysis we generated DNA standard curves. Therefore, the DNA from a 48h (pi) chlamydial culture was extracted and subjected to eight two-fold serial dilutions in DNase-free water. The use of DNA standard curves allows the cross-comparison of the expression levels among genes at each time-point that could not be achieved by using cDNA standard curves, given that DNA represents equal amounts of each single copy gene. Primers for the *16SrRNA* and for each plasmid ORF and sRNA (Supplemental Table 4.1) were design using the Primer Express software (Applied Biosystems). The qPCR assays were performed using the reaction mixtures and amplification profiles described above with the following differences: 5 µl of the generated cDNA were used, and the plates comprised a DNA standard curve for each ORF and duplicates of cDNA from each time-point. Also, all qPCR plates included "no RT" controls. Normalization of the gene expression values, was achieved by dividing the raw qPCR data of the plasmid genes (multiplied by a factor of 10^6) at each time-point by the respective *16SrRNA* transcription values, as this gene was previously shown to be an accurate normalizing gene for gene expression analysis in *C. trachomatis* [214]. The amount of each sRNA was determined by subtracting the expression of the respective ORF (quantified by other primer sets), as the primers used for quantifying the sRNAs also target the ORFs. The final quantification results were based on triplicate experiments for all strains.

4.5. Results

4.5.1. Polymorphism analysis of plasmid ORFs

The genetic analysis revealed that all plasmid ORFs presented a similarity degree of ≥ 99.1% at the nucleotide level, among all 44 strains used in the analyses, an even higher value than the one found for the highly conserved chromosome [30] (Supplemental Table 4.2). In general, mutations only distinguished LGV strains from the remainder. This trend is also seen at the protein level, except for Pgp3 that showed a *p*-distance value of 0.021 (SE ± 0.006) (i.e., 2.1%), which mirrors the values found for high polymorphic proteins encoded by the chromosome [130,142,209]. A brief survey of the evolutionary profile revealed that only ORF5/*pgp3* showed dN/dS > 1, but this value was not statistically significant on the Z-test of positive selection. Globally, these results are in agreement with previous

analyses that highlighted the sequence conservation of *C. trachomatis* plasmid [30,174] among strains of different disease-groups. Thus, this tiny genetic variability unlikely contributes to the dissimilar cell-appetence among *C. trachomatis* strains, which suggest the need for evaluating other approaches, such as the one presented in this study.

4.5.2. Plasmid copy number per genome

For both the absolute quantification and expression assays, the analysis of the melting dissociation curves of all qPCR plates showed a single peak for each PCR product confirming the high specificity of the reactions, which was corroborated by the lack of amplification of the "no template" and "no RT" controls (data not shown). The number of plasmids *per* genome throughout the *C. trachomatis* developmental cycle ranged from 1.0 (SD ± 0.3) (F/CS465-95 at 4h pi) to 7.4 (SD ± 2.3) (L2/434 at 20h pi) (Figure 4.1). In addition, the plasmid copy number/genome was higher at 20h pi for almost all strains, which corresponds to the RBs replication stage. The exception occurred for the clinical strain F/CS465-95 that showed the higher value at 42h pi but also presented the highest associated standard deviation. Considering the two clinical strains, it is noteworthy that they presented lower mean values of plasmid load than the reference strains sharing the same tropism. Still, if one considers the full chlamydial developmental cycle, these differences are not statistically significant (two tailed Student's t-test; $P > 0.068$).

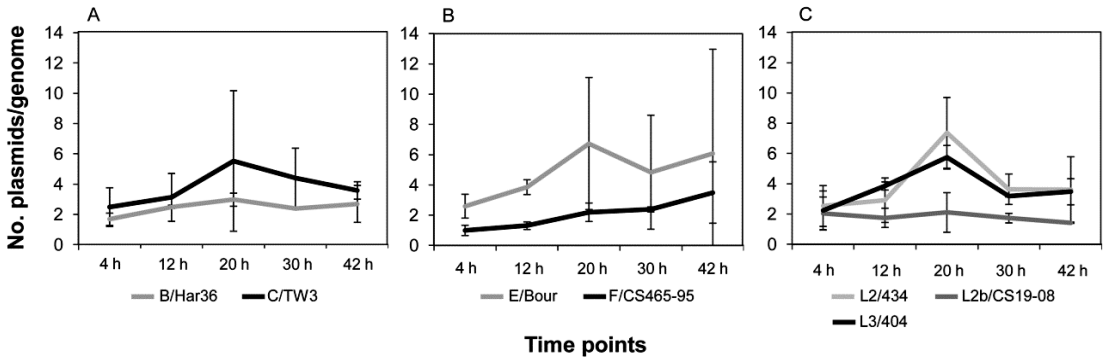


Figure 4.1. Plasmid load *per C. trachomatis* genome throughout development. Panels A–C refer to ocular, genital, and LGV strains, respectively. For each strain, and for each time-point, the number of plasmids *per C. trachomatis* chromosome was calculated by absolute qPCR (see *Section 4.4.4* for details), and was based on three independent biological replicates. Vertical lines represent standard deviations to the mean values at each time-point of the bacterial developmental cycle.

The quantification of the number of genomes throughout development allowed us to further determine the growth rate of each strain and the respective doubling time (dt). We observed that the strain with the lowest growth rate was B/Har36 (~$0.135h^{-1}$, dt = 5.13h) and the fastest growing strain was L2b/CS19-08 (~$0.27h^{-1}$, dt = 2.57h), but no association could be established with the plasmid load.

4.5.3. Transcription analyses

Globally, we observed that the mean expression level of plasmid ORFs were about 5-fold lower than the one observed for tenths of chromosomal ORFs that we have previously analyzed by using precisely the same methodology and calibration procedures [142,283]. In the present study, we first analyzed if some plasmid ORFs are more expressed than others, regardless the strain and their developmental stage, and also if the relative expression of each ORF differentiates strains with dissimilar tissue tropism. We observed that ORF6/*pgp4* presented the highest mean expression value (followed by ORF5/*pgp3*, ORF7/*pgp5* and ORF2/*pgp8*), which was ~6-fold higher than the one of the lowest expressed (ORF3/*pgp1*) (Figure 4.2). For all ORFs, it could be observed that the ocular strains present lower mean expression values than the genital or LGV strains. Statistically significant differences between disease groups (two tailed Student's t-test; $P < 0.05$) were found for several ORFs, where the most prominent difference was seen for ORF5/*pgp3* that was ~5-fold more expressed in LGV strains than in ocular strains.

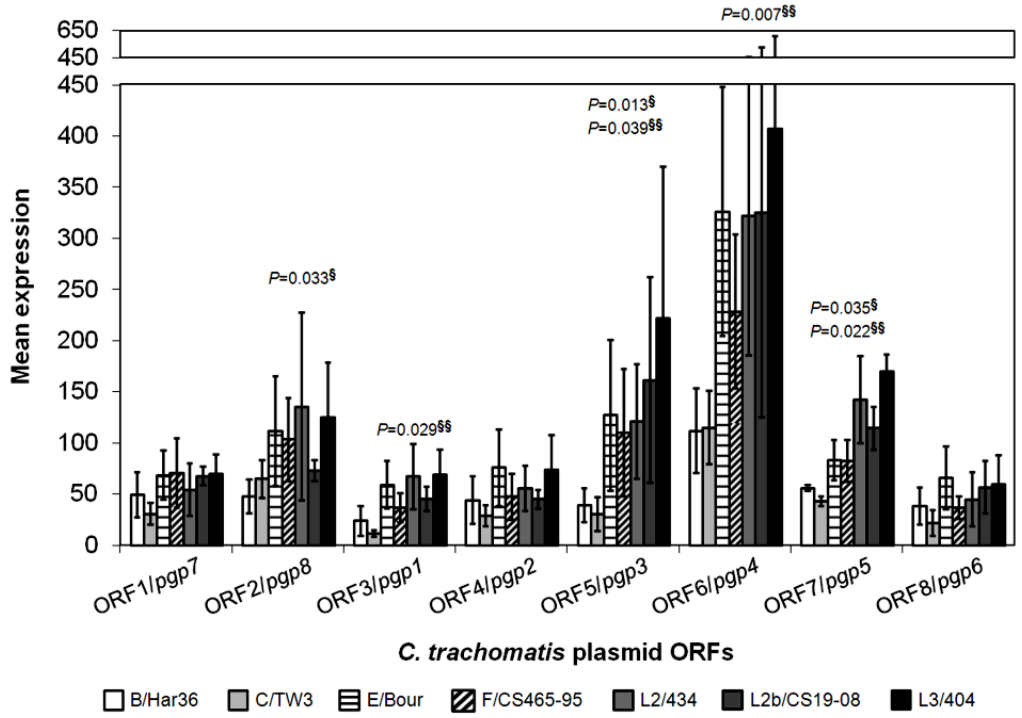


Figure 4.2. Expression-based relevance of each plasmid ORF. The graph represents the mean expression for each ORF calculated from the expression values of all time-points. The bars represent the mean expression for each strain and vertical lines represent standard deviations. *P* values (two tailed Student's t-test) concerning the expression differences among the three disease groups were calculated based on the mean values for each strain, and are shown above the respective ORF bars: § and §§ represent the statistically significant expression differences of ocular versus genital strains, and ocular versus LGV strains, respectively. No statistically significant expression differences were found between the genital and the LGV strains.

We also evaluated if the abundance of ORF transcripts throughout development, regardless the individual contribution of each ORF, is associated with the plasmid load. Although the level of ORFs transcripts *per* plasmid seems to be quite stable for most strains (Supplemental Figure 4.1), a strinkingly opposite scenario was observed for the genital strain F/CS465-95 and the LGV strain L2b/CS19-08 that presented higher expression levels at the developmental stages with the lowest plasmid load.

Subsequently, we intended to perform a fine tune analysis of the expression of each ORF at the five different time-points (4, 12, 20, 30 and 42h) to assess if possible differences could be specific of strains with the same tropism. We observed almost identical expression profiles among the two ocular strains (Figure 4.3), where ORF5/*pgp3* and ORF6/*pgp4* presented a rather unusual profile, suggesting that these genes could be induced early in the cycle, down-regulated at mid-cycle, and induced again at late cycle, as previously observed for some inclusion membrane protein coding genes [142]. In general, the expression profiles were heterogeneous among strains hampering any association with tissue tropism.

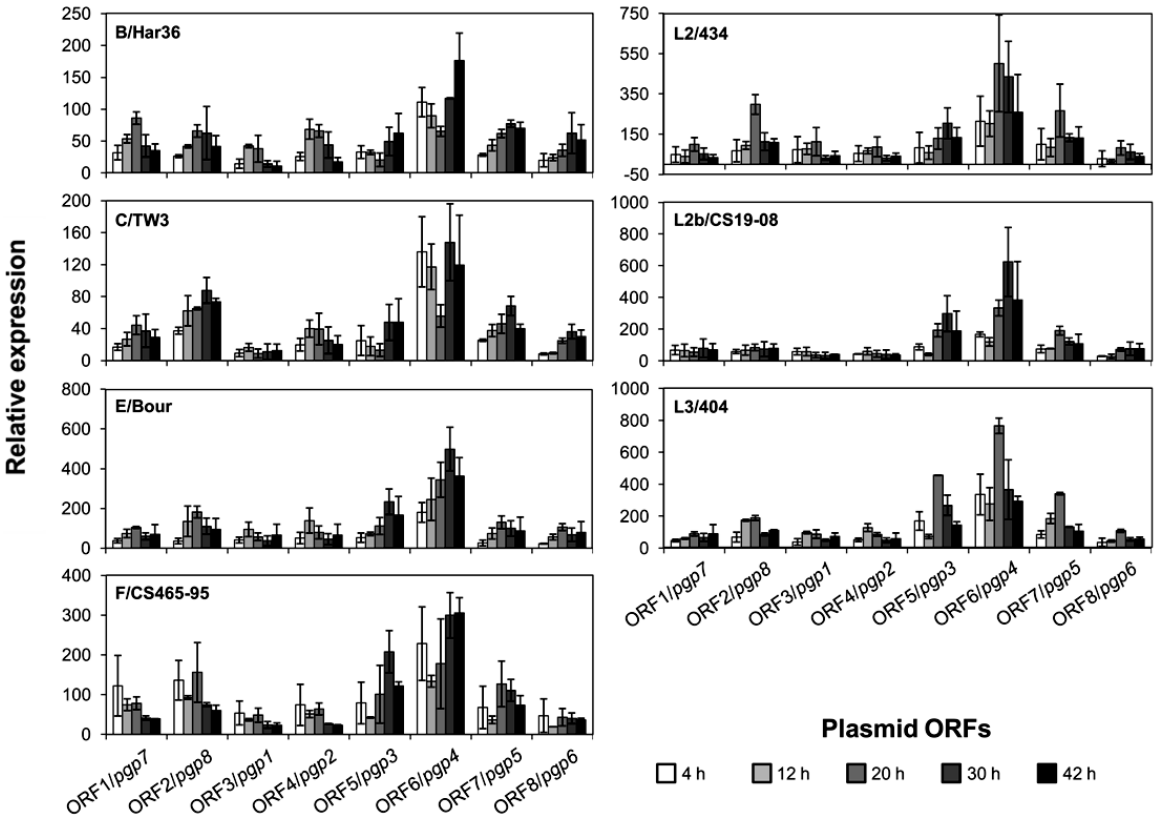


Figure 4.3. Individual expression profile of plasmid ORFs. In each graph is shown the expression levels and profiles of all eight plasmid ORFs for each strain at five time-points of the developmental cycle. qPCR determinations were normalized against the *16SrRNA* expression. Vertical lines represent standard deviations.

Finally, we also aimed to determine the transcription levels of two plasmid sRNAs, sRNA-2 and sRNA-7, which are anti-sense to ORF2/*pgp8* and ORF7/*pgp5*, respectively. In particular, the sRNA-2 was found to be generally more expressed than previously analyzed chromosomal ORFs [142,283] and also than plasmid ORFs (this study), where the observed differences can exceed two orders of magnitude. For example, for C/TW3, the expression peak of sRNA-2 was about 100-fold higher than the one for the lowest expressed ORF3/*pgp1*. sRNA-2 showed a mean expression value always higher than that of sRNA-7 for all strains (differences ranging from 1.4- to 11.9-fold) with more marked differences for LGV strains. The expression peak of sRNA-2 was observed at 12h pi for five out of the seven strains (Figure 4.4A), regardless strains tissue tropism. On the other hand, for six out of the seven strains the 4h pi time-point is among the two which present the lowest expression values. This contrasts with the observation for the sRNA-7 (Figure 4.4B), where the 4h pi was among the time-points with the highest expression levels throughout development for all strains.

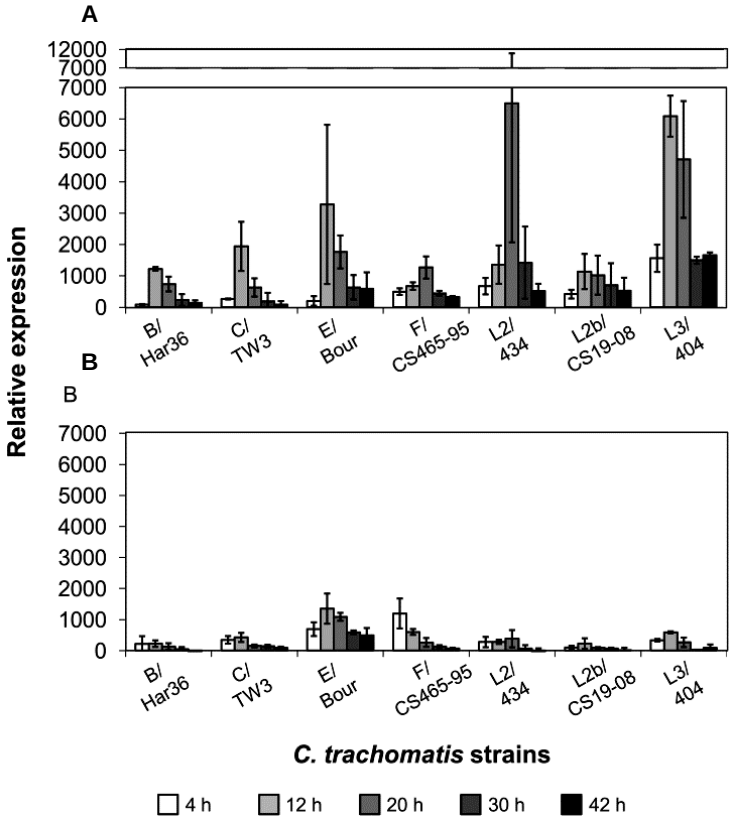


Figure 4.4. Expression of two plasmid anti-sense sRNAs during *C. trachomatis* developmental cycle. Panel A and panel B depict the expression levels and profiles of the sRNA-2 and sRNA-7, respectively, throughout the life-cycle of each strain. For comparison purposes the same scale was used for both sRNA. All expression determinations were normalized against the *16SrRNA* and vertical lines represent standard deviations.

4.6. Discussion

Despite the ongoing chromosomal size-reduction of pathogenic Chlamydiaceae [11,124], almost all species retain a highly conserved plasmid [175], for which the biological role is not well understood. Diverse studies have been focused on the plasmid molecular characterization, such as the evaluation of gene content and organization [174,183,197], transcriptional activity [184,186,187,190], and proteins function [18,162,180,185,284-286]. Efforts have also been done to clarify its implication on bacterial phenotype with the ultimate goal of elucidating its role in bacterial pathogenicity [23,162,177-180]. In particular, comparative infectivity studies enrolling plasmidless and plasmid-bearing strains conducted in *C. muridarum* showed that plasmid-cured strains retained the infectious phenotype but displayed a reduced ability in causing disease when compared to the plasmid-bearing strains [23,178,179]. In *C. trachomatis*, naturally occurring plasmidless strains were found to emerge in just ~1% of the inclusions in a bacterial culture [177]. In the present study, we verified that all *C. trachomatis* strains, regardless their tissue tropism, harbour the plasmid at a copy number of 1.0–7.4 *per* bacterial genome (Figure 4.1), where the highest number of copies was observed during the exponential-phase (~20h) of the developmental cycle, in six out of the seven strains. These results are in agreement with previous studies referring that *C. trachomatis* carries a low copy number plasmid [191,196,197] and with the general assumption that a strong selective pressure is likely acting towards plasmid maintenance. The higher number of plasmids/genome observed at RBs replication stage is likely crucial to ensure the transmission of this low copy number plasmid to the daughter-cells, where a tight regulation of the plasmid segregation process during cell division must be mandatory. This regulation is believed to be carried out by a hypothetical active partitioning system, given the sequence homology of the plasmid ORF7/*pgp5* with the *sopA* and *parA* genes of other bacteria encoding partitioning proteins [183,195,287]. Intriguingly, deletion mutagenesis assays suggest that ORF7/*pgp5* is not essential for plasmid maintenance [162], contrarily to ORF8/*pgp6* which is likely co-transcribed with the former [184]. Considering that partitioning systems are known to involve co-transcribed genes (e.g. *sopA*/*sopB* and *parA*/*parB*) [194], and that both ORF7/*pgp5* and ORF8/*pgp6* presented higher expression levels mostly at the mid stage of the developmental cycle (Figure 4.3) (where the products of these ORFs are likely needed for plasmid segregation accomplishment), we speculate that ORF8/*pgp6* may play a role in plasmid transmission.

Other studies associated the absence of the *C. trachomatis* plasmid with the inability of strains to accumulate glycogen within the inclusion [177], and showed that the plasmid is a trans-acting transcriptional regulator of chromosomal genes [162,180], which implicates the plasmid in the pathogenesis of this bacterium *in vivo*. Still, there is a lack of experimental evidences directly associating the plasmid with other phenotypes, specifically the differential tissue tropism exhibited by *C. trachomatis* serovars. At the genetic level no ORF fully reflects tropism, but it is worth noting that for ORF5/*pgp3*, which encodes the immunodominant antigen Pgp3 [18,187], most nucleotide differences

are nonsynonymous (Supplemental Table 4.2). The Pgp3 is secreted into the inclusion lumen as well as into the host cytosol [18], and the overrepresentation of nonsynonymous alterations had been previously observed for chromosomal genes whose products are also known to be secreted [142,209]. We believe that this limited information given by the genomic analysis may be enriched by the analysis of the transcriptional activity of the *C. trachomatis* plasmid, which unveiled some important findings concerning tropism dissimilarities. In general, ocular strains presented lower expression of plasmid genes than the genital or LGV strains (Figure 4.3). In fact, significant expression differences were found for five out of the eight plasmid ORFs between the ocular and at least one of the other disease groups (Figure 4.2). Although this suggests an apparent lower importance of the plasmid in the ocular strains, this is pure speculation that lacks experimental evidence. When comparing the expression levels of all eight plasmid ORFs, we observed that ORF6/*pgp4*, followed by ORF5/*pgp3*, displayed the highest amount of transcripts for all strains (Figures 4.2 and 4.3). Curiously, these two genes were found to be non-essential for plasmid maintenance [162], a trait common to many plasmid-encoded virulence factors [195,269,278]. Experiments using a plasmid with ORF6/*pgp4* deleted [162] resulted in a phenotype virtually identical to that of a plasmidless strain, implicating this gene as the transcriptional regulator of ORF5/*pgp3* and several chromosomal genes including one involved in the glycogen biosynthesis (*glgA*) and other putative virulence factors, such as CT049, CT084 and CT144. In support of the ORF6/*pgp4* high relevance in *C. trachomatis* biology, previous studies [184,187] showed that this gene is transcribed from two different promoters, which likely boosts the generation of Pgp4.

We also observed that the plasmid load is not associated with strains' tropism and that *C. trachomatis* does not regulate the number of plasmids throughout development according to the ORFs expression needs, as a higher number of plasmids (Figure 4.1) does not reflect a higher transcriptional activity of the plasmid ORFs (Figure S4.1). In fact, although the global amount of ORFs mRNA *per* plasmid copy was found to be nearly constant (Figure S4.1), the two clinical isolates revealed a clear absence of correlation between expression and plasmid load.

Regarding the two plasmid sRNAs, we observed that their mean expression levels were generally higher than those of the ORFs, where in some cases up to 100-fold differences were detected (Figures 4.2 and 4.4). Contrarily to a previous study [186] showing sRNA-7 (anti-sense to ORF7/*pgp5*) as the most abundant plasmid transcript, we verified that the sRNA-2 (anti-sense to ORF2/*pgp8*) was the most expressed locus, presenting a mean expression value always higher than that of sRNA-7 for all strains (up to 12-fold). It is notable that these differences were always higher than 7-fold in the LGV strains and never exceeded 2-fold in genital strains. Moreover, the expression peak for both sRNAs occurs considerably earlier than for the plasmid ORFs, which is particularly notable for sRNA-7 that is required at very early stages of the life cycle. Considering that the deletion mutagenesis assays [162] did not target the sRNA-2, the putative regulatory role of this abundant sRNA remains to be elucidated.

In summary, by investigating the load and transcriptional dynamics of *C. trachomatis* plasmid among strains with dissimilar tropism, we found that, although plasmid copy number was not instructive,

ocular strains generally exhibit significantly lower expression than genital or LGV strains (with special emphasis for the transcriptional regulator ORF6/*pgp4*), and that the biological relevance of sRNA-2 relative to sRNA-7 seems to be higher for LGV strains. We believe that the extension of the very recent and unprecedented deletion mutagenesis-based study [162] to *C. trachomatis* strains causing different disease outcomes and also targeting the hugely expressed sRNA-2 will certainly clarify the putative association between the plasmid role and strains' tropism.

Acknowledgments

This study was supported by grant ERA-PTG/0004/2010 from the Fundação para a Ciência e a Tecnologia (FCT/MEC) in the frame of ERA-NET PathoGenoMics (to JPG). RF and VB are recipients of Ph.D. fellowships (SFRH/BD/68532/2010 and SFRH/BD/68527/2010, respectively) from the FCT/MEC. AN is a recipient of a postdoctoral fellowship (SFRH/BPD/75295/2010) from the FCT/MEC.

Chapter V

Global survey of mRNA levels and decay rates in the two biovars of the obligate intracellular *C. trachomatis*

Manuscript in the submition process

Ferreira R, Borges V, Borrego MJ and Gomes JP (2016)

Personal contribution

RF performed the great majority of the experimental procedures, analysed and interpreted the data and wrote the paper.

5. Global survey of mRNA levels and decay rates in the two biovars of the obligate intracellular *C. trachomatis*

5.1. Abstract

Interpreting the intricate bacterial transcriptomics implies understanding the dynamical relationship established between de novo transcription and degradation rate of transcripts. Here, we aimed to evaluate the transcriptome and to assess the decay rate of mRNAs from different-biovar strains of the obligate intracellular bacterium *C. trachomatis* (D/CS637-11, E/CS1025-11, Ia/CS190-96 and L2b/CS19-08). By using RNA-sequencing at mid developmental stage, we observed that: *i*) 60-70% of the top-50 expressed genes encode proteins with unknown function (accounting for 44% in L2b/CS19-08) and proteins involved in "Translation, ribosomal structure and biogenesis" (accounting for 42% in Ia/CS190-96); *ii*) 22 genes out of that top-50 expressed were common to strains from different biovars; *iii*) with few exceptions, the expression ranking by genes' functional categories was concordant among strains, regardless their biovar, where "Plasmid genes" and genes from the "Secondary metabolites biosynthesis, transport, and catabolism" categories being found among the most expressed; *iv*) the median of the half-life time ($t_{1/2}$) values were 15 min and 17 min for L2b/CS19-08 and E/CS1025-11, respectively, contrasting with the considerably lower values of any other bacteria studied so far; *v*) as observed for other organisms, there was a lack of correlation between transcripts' expression and decay rate; and that *vi*) the majority of the 100 most stable transcripts were essentially members of four functional categories, of which the HPs were the most represented, with CT016, CT035, CT360, CT577, CT578, and CT865 being found to be common to the L2b/CS19-08 and E/CS1025-11 strains. Despite the general high complexity of the mechanisms of RNA production and decay in bacteria (which difficults robust and straightforward associations) and of the unique biological characteristics of *C. trachomatis*, the present study provides an overview of the bacterium transcriptome dynamics, highlighting the dissimilitudes and concurrences among different-biovar strains in terms of mRNAs abundance and decay rates.

5.2. Keywords

Chlamydia trachomatis; RNA-seq; transcript stability; gene expression

Chapter V

5.3. Introduction

For the past few years, several studies have shown that bacteria present complex transcriptional activity, showing intricate gene expression regulation, and encoding multiple classes of transcripts (including sRNAs), which play a major part in those processes [288-291]. In this context, understanding such transcriptomic scenery is essential to get insights on the pathways of a cell physiology, metabolism, and adaptation to changing environments. To start addressing this issue, it is vital to assess the expression profile of each transcript, which in turn is the result of the balance established between de novo transcription and degradation of existing transcripts at a particular moment, under certain conditions. A multitude of studies have focused on the regulation and mechanisms of transcription initiation and/or on measuring steady-state transcript levels in different bacteria, but only a handful evaluated and put in context the global trends of transcripts' molecular stability (some examples: [292-299]). Although these studies generally revealed lack of association between several characteristics of a transcript and its decay rate over time (e.g., secondary structures, G+C content, codon composition and transcript length), they further pointed that transcripts' stability could be species-specific and influenced by bacterial growth rates [297], and/or by a variety of cell signals and external stimuli (reviewed in [292]). However, mRNA decay studies have been performed only in extracellular bacteria, which are more undemanding to handle, resulting in a general lack of information for obligate intracellular bacteria.

Besides the straightforward gene expression quantification, the versatility and accuracy of the high-throughput RNA sequencing (RNA-seq) technology [300], have already proven to be valuable for the characterization of bacterial transcriptomes, including TSS mapping, the definition of operons, the annotation of 3' ends, the detection of antisense transcription, the characterization of small RNAs and non-coding RNAs, the uncovering of gene fusion, and even the determination of mRNA decay rates [301-305]. For the obligate intracellular bacteria *C. trachomatis*, the RNA-seq technology has been successfully applied, for instance, to map all TSSs and to identify novel sRNAs [186], to simultaneously assess the gene's expression of both the bacterium and the host in response to an infection scenario [306], and also to evaluate the impact of propagation-derived mutations on the transcriptome [307]. The complete and detailed analysis and understanding of the transcriptome of *C. trachomatis* holds great interest as this is a highly prevalent human pathogen and a major public health concern. In fact, it is the etiologic agent of the blindness trachoma and constitutes the major bacterial cause of sexually transmitted infections worldwide [242,243]. As a member of the Chlamydiaceae family, this bacterium displays a unique biphasic developmental cycle of about 36-96h [308] during which, the extracellular and infectious form, the EB, enters the host cell and differentiates into the intracellular and replicative form, the RB. After several rounds of RB binary fission within a membrane-bound vacuole known as the inclusion, these differentiate back into the EBs, which are released from the host cell, by lysis or inclusion extrusion, to carry on new infections in neighbouring cells.

Overall, the strains of *C. trachomatis* can be classified into 2 biovars and 15 major serovars according to the polymorphism of the MOMP encoding gene (*ompA*). The trachoma biovar include strains from serovars A-C and D-K, which preferentially infect epithelial cells of the ocular conjunctivae and the ano-urogenital tract, respectively, thus causing organ-specific diseases [244,308]. On the other hand, the LGV biovar include strains from the serovars L1-L3, which also colonize the host through the ano-urogenital tract, but are able to disseminate to the regional draining lymph nodes, via infection of the macrophages [309]. To date, investigators are struggling to understand which factors underlie the phenotypic differences, namely, the huge discrepancies in growth rates, routs of infection, cell tropism and disease-outcomes ([30,84,130,143,144,209,258,310,311] and reviewed in [129]). Considering that the dissimilarity among the *C. trachomatis* strains is less than 2% at the genome level, it has been speculated that differences in the regulation of gene expression also contribute to the above mentioned phenotypic discrepancies displayed by strains. Some studies aiming the evaluation of gene expression differences among *C. trachomatis* strains have been performed, namely, whole transcriptomic analysis focused on a single strain [85,186,306], or multi-strain analyses focused on a restricted set of genes [27,142,162,214,312]. Still, one must take into account that the intra-strain gene expression dissimilarities observed so far reflect transcripts abundance at discrete moments, and that this abundance is completely dependent on the joint action of both the transcription initiation and decay rates [313-316]. To fullfil the gap of knowledge on *C. trachomatis* gene transcriptional dynamics, we aimed to undertake a global survey regarding the expression level and the decay rate of transcripts in strains representing the two biovars, in an attempt to unveil putative differences that may underlie phenotypic differences of such biovars.

5.4. Materials and Methods

5.4.1. Cell culture, rifampicin treatment and harvesting

C. trachomatis E/CS1025-11 and L2b/CS19-08 clinical strains (subjected to minimal culture passages) were each inoculated into 15 T_{25} flasks of confluent HeLa229 cells monolayers by centrifuging at 2200 rpm for 1h (34 °C), followed by an incubation stage of 1h at 37 ºC, in a 5% CO_2 atmosphere. Cell culture medium was then discarded and fresh enriched medium (MEM 10% foetal bovine serum, vitamins, non-essential aminoacids, glucose and 0.5 µg/ml cycloheximide) was added to the cultures. Bacterial cells were allowed to grow at 37 °C with 5% CO_2, until the mid-phase of the developmental cycle was achieved, because the majority of *C. trachomatis* genes are being actively transcribed at this point [85,306] and also because, at this stage carry over mRNA, which would be quickly degraded biasing the decay rate calculations, was found to be no longer present [306].

Three sets of five flasks, for each strain, were submitted to different periods of rifampicin treatment (10 µg/ml) – 0 min (T_0, no treatment), 10 min (T_{10}) and 30 min (T_{30}) – in order to stop *de novo*

RNA synthesis [317,318], and thus, allow the evaluation of the mRNA decay. One T_{25} flask of T_0, T_{10} and T_{30} sets, was immediately harvested by using glass beads, the cell suspension was sonicated for 7 min (Vibra Cell, Bioblock Scientific), for host cells disruption and chlamydial release, submitted to a centrifugation of 7 min at 700 rpm, for cell debris deposit, and finally the supernatant was stored at -20ºC for further DNA extraction. The remaining four T_{25} flasks of T_0, T_{10} and T_{30} culture sets, were treated with the rifampicin solution, as previously mentioned. Immediately after the respective treatment period, the solution was replaced with a 2:1 solution of RNAprotect™ Bacteria Reagent (Qiagen) and PBS buffer and left for 2 min, to ensure diffusion into the cells for the bacterial RNA preservation during subsequent handling. Again, cultures were harvested, sonicated and centrifuged as previously referred, and the supernadant was immediately submitted to RNA extraction. Note that we intentionally did not treat the cell culture from which DNA would be extracted, because a preliminary assay showed that the RNAprotect™ Bacteria Reagent (Qiagen) degrades DNA (data not shown).

5.4.2. DNA and RNA extraction

Both nucleic acids were extracted as previously referred [312]. Briefly, DNA was extracted by centrifuging the stored (-20 °C) supernatant at 14,000 rpm for 10 min (at 4 °C) and the pellet was resuspended in 200 µl PBS for QIAamp® DNA Mini Kit (Qiagen) extraction, according to manufacturer's instructions. DNA was eluted in 50 µl of AE buffer and stored at -20 °C after A260 nm quantification in a NanoDrop 1000 spectrophotometer (Thermo Scientific). For total RNA the RNeasy® Mini Kit (Qiagen, CA, USA) was used according to manufacturer's instructions. Briefly, the culture supernadants were subject to a high-speed centrifugation (14,000 rpm) for 10 min at 4 °C, the pellets were suspendend in lysozyme-containing TE buffer and treated with RLT buffer with 1% β-mercaptoethanol for cell lysis. An on-column DNase treatment (RNase-free DNase, Qiagen) was included in the procedure to remove contaminant DNA, and RNA was eluted in 40 µl of RNase-free water. RNA yield and purity were determined by absorbance measurement at 260 nm and 280 nm using the NanoDrop 1000 spectrophotometer (Thermo Scientific). For a parallel qPCR quantification of the expression levels (and decay kinetics) along the time frame used, 5 µl of this total RNA was further used for cDNA generation (see below).

5.4.3. Bacterial mRNA preparation/purification

Eukaryotic, mitochondrial and bacterial rRNA were depleted by using the Ribo-Zero™ Gold rRNA Removal Kit (Epidemiology) (Illumina, CA, USA), according to manufacturer's instructions. Subsequently, the Dynabeads® mRNA Purification Kit for mRNA Purification from Total RNA preps (LifeTechnologies) was used, with and adapted protocol, to pull the eukaryotic mRNA out of the samples, leaving only the bacterial mRNA. The obtained bacterial mRNA was concentrated in a final volume of 13 µl, by using the RNeasy® MinElute™ Cleanup Kit (Qiagen, CA, USA), according to manufacturer's instructions. Bacterial mRNA quality and concentration were determined by the

Bioanalyzer (Agilent) equipment, where the absence of both 18S and 28S rRNA readings reflects the purity of the mRNA.

5.4.4. RNA-seq

Bacterial mRNA-enriched samples were subjected to library construction (TruSeq Stranded mRNA sample preparation kit, Illumina) and sequencing on an Illumina MiSeq sequencer using a paired-end (2x75bp) strategy (at least 15M reads were dedicated *per* sample), according to manufacturer's instructions. FastQC analysis was used to assess reads quality and Bowtie2 was applied for mapping reads of each strain to the respective chromosome and plasmid DNA sequences obtained in a previous study [307]. Cufflinks (version 2.1.1; http://cufflinks.cbcb.umd.edu/) tools were further applied to quantify the amount of transcripts of each chromosome and plasmid CDS normalized as fragments *per* kilobase of CDS *per* million mapped reads (FPKM), as previously described [307]. Calculations of transcript amounts for each time point (before and after 10 and 30 min of rifampicin treatment) were based on two biological replicates. The use of a high throughput technology, such as the Illumina technology, is essential for the quantification of decaying mRNAs' expression levels [296] as it potentiates the capture of low expressed genes and extremely labile transcripts. In this study, transcript counts were obtained for the three experimental conditions (T_0, T_{10} and T_{30}) for >99% of all annotated CDSs in both genomes.

5.4.5. cDNA generation and qPCR assays

In order to validate the obtained RNA-seq RNA decay kinetics we performed parallel qPCR assays for which cDNAs were generated from 2 µl of each T_0, T_{10} and T_{30} samples, from both replicates of both strains, as previously described [214,312]. Briefly, TaqMan® RT Reagents (Applied Biosystems, Branchburg, USA) were used as follows: 2.5 µM of random hexamers, 5.5 mM MgCl2, 500 µM of each dNTP, 1× RT Buffer, 0.8 U/µl RNase inhibitor and 1.25 U/µl MultiScribe RT, in a final reaction volume of 50 µl. Cycling conditions were: 10 min at 25 °C, 15 min at 42 °C and 5 min at 99 °C. cDNA was stored at -20 °C until use.

DNA from a 48h (pi) chlamydial culture was extracted (as described above) and subjected to eight two-fold serial dilutions in DNase-free water, in order to produce the DNA standard curves for the expression quantification. The use of DNA standard curves allows the cross-comparison of the expression levels among genes at each time-point that could not be achieved by using cDNA standard curves, given that DNA represents equal amounts of each single copy gene. Primers for the 16SrRNA and 21 other genes selected for quantification purposes were design using the Primer Express software (Applied Biosystems). The qPCR assays were performed using the LightCycler® 480 equipment, SYBR Green chemistry and optical plates and caps (Roche). The qPCR mixture consisted of 1× LightCycler® 480 SYBR Green I Master (Roche), 400 nM of each primer and 5 µl of each DNA (standard curves) or

cDNA (samples), in a final volume of 25 µl. Plates included one standard curve for each gene, cDNA duplicates, and "no RT" controls.

Gene expression was normalized by dividing the mean value of raw qPCR duplicates by the respective mean value of the *16SrRNA*, for each condition (T_0, T_{10} and T_{30}), of both replicates of both strains. This strategy was used because *16SrRNA* was previously shown to be a stable gene, being a valuable resource for expression normalization in *C. trachomatis* [214].

5.4.6. mRNA half-life time ($t_{1/2}$) analysis

Each mRNA half-life time ($t_{1/2}$) was calculated by using an adaptation of the "two-fold" decay step method [319] based on the fit of an exponential decay between values obtained at the first time-point (before rifampicin addition; T_0) and the values calculated t minutes (10 min and/or 30 min) after the transcriptional arrest (T_1), using the formula: $t_{1/2} = -\ln 2/k$; where the rate of decay rate (k) was estimated as follows: $k = \ln(T_1/T_0)/t$. Therefore, it was also required to establish the appropriate time interval to correctly apply the formula, which was determined by performing an overall screening of the "decay profile" of each transcript, by considering its mean value of FPKM for the T_0 (no treatment), T_{10} (10 min of antibiotic treatment) and T_{30} (30 min of antibiotic treatment). In that regard, the criteria applied were: *i*) if expression at $T_0 > T_{10} > T_{30}$, we used the interval with the highest slope (usually was the T_0-T_{10}); *ii*) if $T_0 > T_{10} < T_{30}$, we used the T_0-T_{10} interval, whether $T_0 > T_{30}$ or $T_0 < T_{30}$, because we assumed that at 30 min after rifampicin treatment the blockage effect of this antibiotic may not be as absolute as after 10 min; *iii*) if $T_0 < T_{10} > T_{30}$, we used T_0-T_{30}, but only if $T_0 > T_{30}$, otherwise we considered that there was no overall mRNA decay; and *iv*) if $T_0 < T_{10} < T_{30}$, we were not able to calculated mRNA decay at any interval, and therefore those transcripts were removed from the analysis.

Throughout the text, results were analyzed either in a gene-by-gene manner (assuming the gene designation of the D/UW3-CX genome annotation; NC_000117) or by functional categories (according to Heizer and collegues [320], with some exceptions; Supplemental Table 5.1).

5.5. Results and Discussion

The main goal of the present study was to evaluate the whole transcriptome dynamics of *C. trachomatis* strains from two biovars by using the cutting-edge RNA-seq technology in multiple strains. Specific goals include: *i*) comparison of the gene expression levels between four (D/CS637-11, E/CS1025-11, Ia/CS190-96 and L2b/CS19-08) different-serovar strains; *ii*) comparison of the decay rate of all mRNAs of two different-biovar strains (E/CS1025-11 and L2b/CS19-08); *iii*) identification of putative relations between expression level and half-life time ($t_{1/2}$) at both gene and functional category levels.

The reliability of the RNA-seq data was preliminarily assessed through qPCR analysis of 22 transcripts, ensuring that not only the initial mRNA levels of both strains were accurately determined by RNA-seq (T_0), but also that the decay pattern of each transcript (T_0 to T_{10} to T_{30}) was also correctly determined. We obtained high reproducibility of the gene expression quantification results generated with the two distinct quantitative methods ($R^2 > 0.88$ for all comparisons) (Supplemental Figure 5.1), which corroborates the suitability of the normalization strategy applied to the raw RNA-seq data.

5.5.1. Expression analysis between four strains from two biovars

In this analysis, we compared the expression level of transcripts of *C. trachomatis* E/CS1025-11 and L2b/CS19-08 strains, and also of two additional strains (D/CS637-11 and Ia/CS190-96), for which RNA-seq data was obtained from a previous study [307]. All four expression data sets were acquired at the mid-phase stage of the developmental cycle to minimize sample variation and to ensure a dynamic expression of the majority of the *C. trachomatis* genes [85,306].

Firstly, we looked at the genes presenting the highest levels of expression at the mid-phase within each strain, and found a considerable parallelism between strains. In fact, CT001, CT267, CT443/*omcB*, CT444/*omcA*, CT500, CT681/*ompA*, and ORF2/*pgp8* were found among the 10 most expressed genes regardless of the strain. This set includes three genes (CT443/*omcB*, CT444/*omcA* and CT681/*ompA*) encoding major constituents of the outer membrane of this species, whose high expression at replication stage is expectable. The high level detected for ORF2/*pgp8* reflects the high abundance of its anti-sense sRNA [186,312]. By extending the investigation to the top-50 most expressed genes, we observed that about 60-70% of those top ranked genes encode "HPs" and proteins from the "Translation, ribosomal structure and biogenesis" functional category (Figure 5.1A). The proportion of genes from each of the remainder functional categories within the top-50 ranking never exceeded 6%.

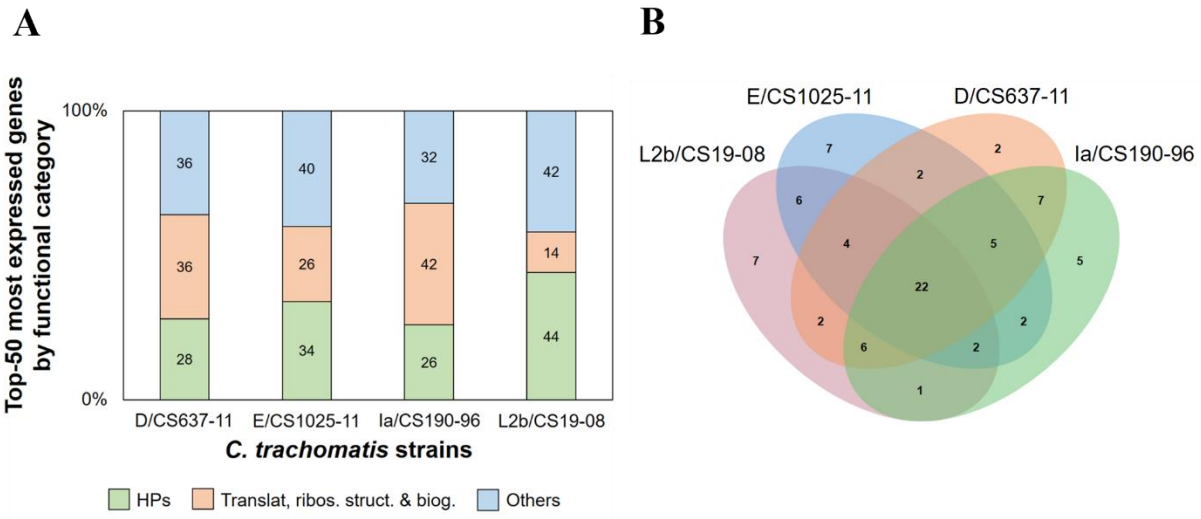


Figure 5.1. Distribution of the top-50 most expressed genes of each *C. trachomatis* strain. A) Composition of the top-50 most expressed genes of the four strains. The two most represented functional categories, for all strains, are represented in green ("HPs") and in orange ("Translat, ribos. struct. & biog."), whereas the remainder are clustered together ("Others") and represented in blue. The proportion (in percentage) of each functional category is displayed within the vertical bars. B) Venn's diagram depicting the distribution, among strains, of their top-50 most expressed genes. L2b/CS19-08 is represented in light red, E/CS1025-11 in blue, D/CS367-11 in orange and Ia/CS190-96 in green. Each number represents the absolute quantity of genes shared by the indicated strains in each comparison.

Furthermore, 22 genes were found to be shared by all four top-50 ranks (Figure 5.1B), where again the "HPs" were the most represented genes (7/22), followed by the "Translation, ribosomal structure and biogenesis" genes (5/22). All these 22 genes had already been pointed out as being highly expressed (FPKM ≥ 1.0 and a minimum of 50 mapped reads) at the mid-phase stage of the developmental cycle of a trachoma biovar serovar E strain in a previous RNA-seq study [306]. Considering our results on different serovar (and biovar) strains, these "core" genes certainly play a role in very conserved and essential mechanisms in *C. trachomatis*, at the mid-phase stage of the developmental cycle. Whereas for the members of the "Translation, ribosomal structure and biogenesis" functional category, this result may not be surprising, as at this developmental stage there must be a tremendous demand of protein translation and overall metabolism intervenients, the high representation of HPs coding genes among the highly expressed warrant further investigation. In fact, *C. trachomatis* genes with unknown function are mostly genus- or species-specific [15,83] and certainly play relevant roles during the developmental cycle of this bacterium [285,321], as corroborated by the present study.

Subsequently, we evaluated the global gene expression differences between strains by calculating the median of the gene-by-gene expression differences for each strain pair. We verified that the Ia/CS190/96 had the highest median value of expression difference, being 36% and 32% more expressed then the D/CS637-11 and the L2b/CS19-08 strains, respectively. On the other hand, these two latter strains presented only ~1% variation in the median of expression differences between each other. We then assessed if there are functional categories for which the observed trend (Ia/CS190-96 > E/CS1025-11 > L2b/CS19-08 > D/CS637-11) is not applied. Thus, the median of expression values was determined for each functional category, for each strain (Figure 5.2).

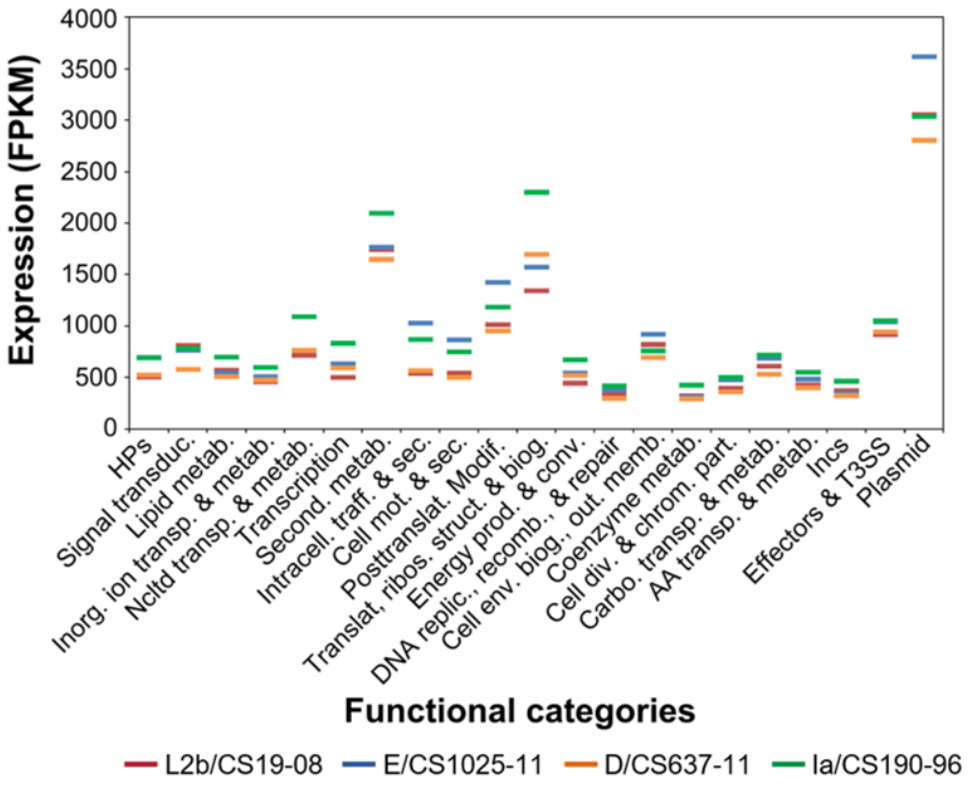


Figure 5.2. Medians of gene expression in twenty-one functional categories, for the four *C. trachomatis* strains used. In the horizontal axis are indicated all the defined twenty-one functional categories, and the vertical axis specifies the expression level. Coloured bars represent *C. trachomatis* strains expression medians: L2b/CS19-08 in red, E/CS1025-11 in blue, D/CS637-11 in orange, and Ia/CS190-96 in green.

Considering the previous result, the Ia/CS190-96 expectedly presented the highest expression median for the majority of the functional categories (15/21). Accordingly, both the L2b/CS19-08 and the D/CS637-11 strains presented the lowest median values of expression for almost all functional categories. However, these trends had exceptions, i.e., the L2b/CS19-08 was the strain with the highest expression median value in the "Signal transduction mechanisms" category, while the E/CS1025-11 strain had the highest expression medians in five functional categories, in particular in the "Plasmid genes" and "Cell envelope biogenesis, outer membrane" categories where the Ia/CS190-96 was the strain with the second lowest median value. It is also clear from Figure 5.2 that four functional categories ("Secondary metabolites biosynthesis, transport, and catabolism", "Posttranslational modification, protein turnover, chaperones", "Translation, ribosomal structure and biogenesis", and "Plasmid genes") present higher medians than the remainder, regardless of the strain. On the other hand, there were also functional categories for which the medians were systematically low in all strains, namely, "DNA replication, recombination, and repair", "Coenzyme metabolism", "Cell division and chromosome partitioning", "Amino acid transport and metabolism", and "Incs or putative Incs". Of note, the expression differences observed between functional categories, within each strain, should be eyed with

caution, as they may be a consequence of the time point at which the developmental cycle was arrested and the growth conditions applied, which most certainly impact the biological needs of the bacterium at a given moment. For instance, the observation of *incs* among the less expressed genes at the mid-stage of the developmental cycle likely reflects the fact that the majority of those genes shows an early-cycle profile of expression [142]. On the other hand, expression differences between strains, within the same functional category, may suggest that dissimilar transcriptional regulation of different sets of genes is probably a determinant factor in the biological dissimilarities of the strains.

We further evaluated the expression profile of each plasmid-encoded gene, as they also comprise one of the most expressed functional categories, and because the expression of chromosomal genes pointed to be regulated by the plasmid [162] (Supplemental Figure 5.2A).

As ORF2/*Pgp8* is known to encode a highly expressed antisense sRNA [186,312], the elevated expression values obtained for this gene (Supplemental Figure 5.2A), in all strains, are the result of the combined expression of both the ORF2/*Pgp8* and its antisense sRNA, and hence, the medians determined for the "Plasmid genes" functional category are overestimated. With that in mind, we considered ORF6/*Pgp4* to be clearly the most expressed plasmid-encoded gene regardless of the *C. trachomatis* serovar (Supplemental Figure 5.2A) at the mid-phase stage of *C. trachomatis* developmental cycle, corroborating previous qPCR data acquired throughout development [312] and substantiates its role as a regulator of several chromosomal genes [162,186]. However, we were not able to establish a direct relation between the expression of ORF6/*Pgp4* and the majority of its putative chromosomal target genes (Supplemental Figure 5.2B). The only exception was the CT798/*glgA*, which presented higher levels of expression and an expression profile among strains that seems to mimic the profile of the ORF6/*Pgp4*, i.e., the highest expression of both genes was verified for the E/CS1025-11 strain, and the lowest value was obtained for the Ia/CS190/96 strain (Supplemental Figure 5.2). Curiously, the CT798/*glgA* is one of the five genes involved in the glycogen metabolic pathway [83,84,122] and was previously shown to be significantly more expressed in the plasmid-bearing L2/434-BU strain than in the L2/25667R, its plasmidless equivalent, specially at the mid-stage of the developmental cycle [180]. Once again, these combined results appear to imply the direct positive regulation of the CT798/*glgA* by the plasmid-encoded ORF6/*Pgp4*, as previously suggested by Song and colleagues [162], which in turn leads to the accumulation of glycogen within the inclusion.

5.5.2. Half-life time analysis between different biovar strains

Presently, it is recognized the essential role of transcripts' degradation in the regulation process of a gene's expression level. The coordinated action between the transcription initiation rate and the decay rate of each RNA, results in its steady-state abundance at a particular moment, under certain environmental conditions. Therefore, unravelling transcripts stability may ultimately allow the inference of their role in the biology of bacteria. To date, several studies have already conducted global surveys of mRNA half-life times in several microorganisms [292-299,322], but none of them included an

obligate intracellular pathogen like *C. trachomatis*. As such, we determined the $t_{1/2}$ of the transcripts of two different-biovar *C. trachomatis* strains (L2b/CS19-08 and E/CS1025-11) by measuring the transcript levels at 10 and 30 min post-treatment with the RNA polymerase blocking agent rifampicin. After defining the most accurate time interval to determine the decay rate of all mRNAs (see Materials and Methods), we observed that the global mRNA stability between these two strains was very similar (Figure 5.3). In particular, 50% of the $t_{1/2}$ values determined for the L2b/CS19-08 and E/CS1025-11 fall within the 9.7 min – 25.17 min (median = 15 min) and 10.12 min – 31.26 min (median = 17 min) intervals, respectively. A previous study [297] speculated that, for different species of the same genus, the differences of the mean $t_{1/2}$ values observed could be linked to the length of the developmental cycle, i.e., slower growths would correlate with higher mRNA stability. Rustad and colleagues [297] argued that, the extended developmental cycle is a consequence of the strain's inability to quickly regulate mRNAs abundance. As such, our observation that the slightly higher $t_{1/2}$ determined for the E/CS1025-11 genes, comparing to those found for the L2b/CS19-08 genes, could eventually be explained by the fact that LGV strains typically grow faster than the trachoma biovar strains. However, we observed no significant differences in the growth rate of these strains during our experiments. Further analyses with multiple strains with dissimilar growth rates is needed to verify if this Rustad and colleagues [297] assumption applies to *C. trachomatis* species.

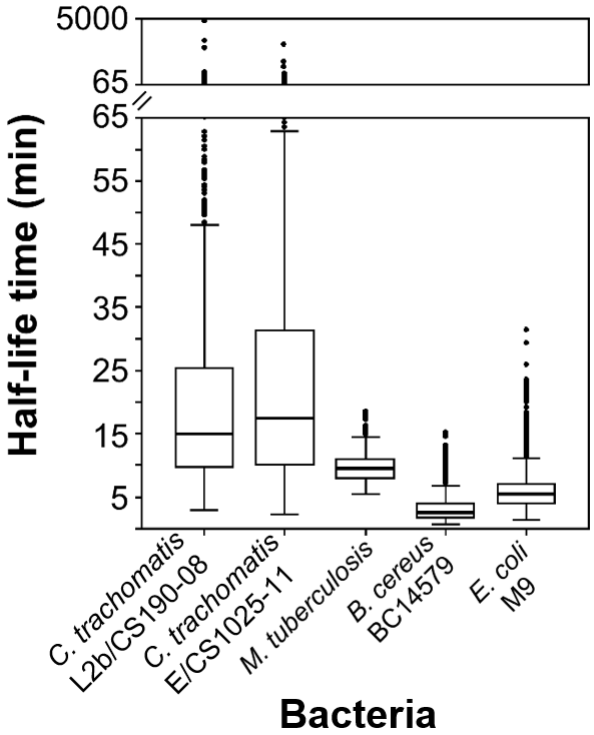


Figure 5.3. Boxplots showing the distribution of mRNA $t_{1/2}$ determined for different bacterial species: *Escherichia coli* [293], *Bacillus cereus* [296], *Mycobacterium tuberculosis* [297], and *C. trachomatis*, E/CS1025-11 and L2b/CS19-08 strains (this study).

The range of $t_{1/2}$ values obtained for the obligate intracellular *C. trachomatis* strains was strikingly wider than those previously obtained for other bacteria [293,296,297], for which the $t_{1/2}$ were also much lower, with medians of 9.3 min, 5.4 min and 2.6 min for *Mycobacterium tuberculosis* (facultative intracellular), *Escherichia coli* and *Bacillus cereus* (both extracellular), respectively (Figure 5.3). In fact, the latter not only has the lowest median, but also the narrowest range of $t_{1/2}$ values (Figure 5.3). Previous studies regarding the Gram-negative *E. coli* and the Gram-positive *B. cereus* attributed the $t_{1/2}$ differences to key ribonucleases that each bacteria possesses, a feature believed to distinguish the two Gram groups [296,323]. However, even larger differences between the transcripts stability can be observed when comparing bacteria with rather different life-styles (Figure 5.3). Free-living bacteria possess very low transcripts $t_{1/2}$ medians ($\leq$ 5.4 min), whereas the obligate intracellular *C. trachomatis* possesses much higher transcripts $t_{1/2}$ medians ($\geq$ 15 min). Values of this magnitude had previously been reported for 5/10 genes with reliable $T_{1/2}$ determination for the same-genus species, *C. pneumoniae* [324]. Curiously, the facultative intracellular bacterium used in this comparison (*M. tuberculosis*) presented a transcripts' $t_{1/2}$ median that lies in between those two values (~9.3 min), and which contrasts with the bulk mRNA $t_{1/2}$ (5.2 min) of its closely related saprophyte *Mycobacterium smegmatis* [297]. Taken all together, although the presence/absence of particular ribonucleases can ultimately determine differences in mRNA decay dynamics, our observations suggest that a major factor underlying the $t_{1/2}$ differences among bacteria is their life-style. For instance, the fact that obligate intracellular organisms may face less drastic environmental changes than free-living bacteria could underlie the slower turn-over rate of the transcripts abundance observed for *C. trachomatis*. However, data from future studies using more bacterial species, both Gram-positive and -negative, with distinct life-styles, will certainly help elucidate if this observed trend is a sampling bias or if it mirrors the actual association between bacterial life-style (and consequently gene function) and mRNA stability.

From the detailed analysis of $t_{1/2}$ detected for the two *C. trachomatis* strains, it was also striking that both strains presented a large number of highly stable genes (Figure 5.3). In an attempt to investigate the nature of those genes, we looked in detail for the 100 genes with the highest $t_{1/2}$ values for each strain. For both strains, this set of genes presented $t_{1/2}$ above 45 min and the majority (> 55%) belong to only four out of the 21 previously defined functional categories, where three (HPs, "Incs or putative Incs", and "Amino acid transport and metabolism") are shared (Figure 5.4).

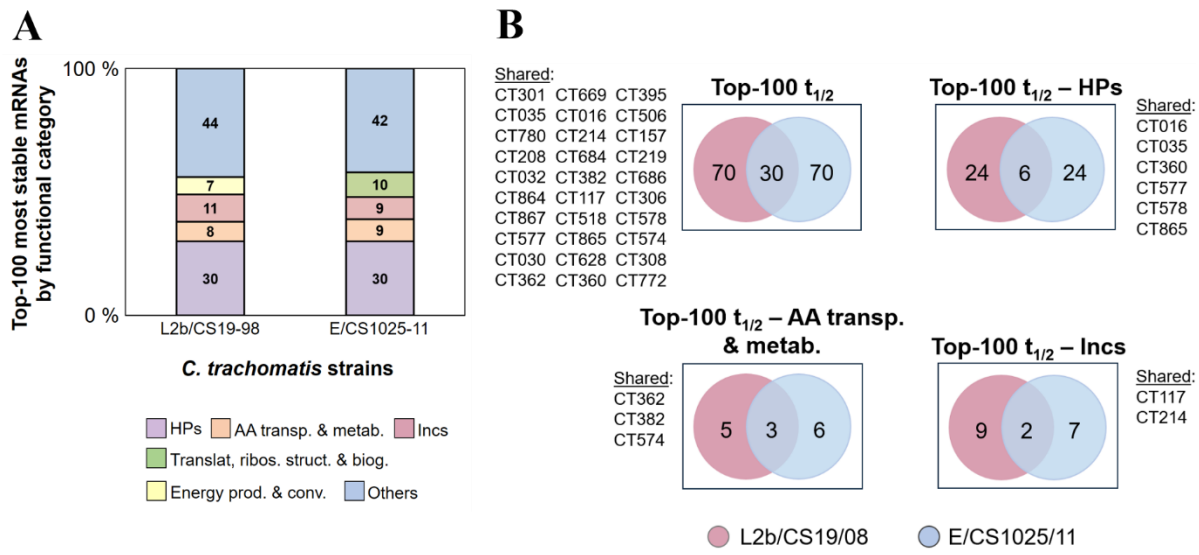


Figure 5.4. Composition of the top-100 most stable mRNAs of the two different-biovar *C. trachomatis* strains, L2b/CS19-08 and E/CS1025-11. A) The four most represented functional categories, of each strain, are represented in purple ("HPs"), orange ("AA transp. & metab."), light red ("Incs"), yellow ("Energy prod. & conv." for L2b/CS19-08), and green ("Translat, ribos. struct. & biog." for E/CS1025-11). The remainder are grouped together and represented in blue for both strains. The proportion (in percentage) of each functional category is displayed within the vertical bars. B) Venn's diagram depicting the distribution, between the L2b/CS19-08 (light red) and E/CS1025-11 (blue) strains, of their top-100 most stable mRNAs, and of the mRNAs of the three functional categories most represented in both their top-100: HPs, "AA transp. & metab., and "Incs". Each number represent the absolute quantity of genes shared by the indicated strains in each comparison.

Although we found this parallelism at the functional category level (Figure 5.4A), the overlap was not so obvious at gene level within each functional category (Figure 5.4B). For instance, only six (CT016, CT035, CT360, CT577, CT578, and CT865) out of the 30 highly stable "HPs" of each strain are shared by both *C. trachomatis* strains. Of note, it must be stated that, despite the overrepresentation (30%) of "HPs" among the 100 most stable genes in both strains, this result must be interpreted with caution, as the "HPs" category is mostly composed by the unique *C. trachomatis* genes for which the biological function is yet to be established. For that reason, it is certainly highly heterologous in terms of gene function, encompassing representatives of several other categories. As such, any association between the huge mRNA stability and the scarce data available for some of them would be likely speculative. However, for instance for CT577, which was already defined as a late expressed gene [85,186] and its product found to be one of the most expressed proteins in EBs [325], our results showing a tremendous stability of its mRNA molecule at mid developmental stage may point that CT577 is also relevant during RB replication.

Given our result on the gene-by-gene comparison for the most stable mRNAs, we further focused our analysis on transcripts presenting $t_{1/2}$ lower than 30 min in both strains (n = 525), i.e., transcripts revealing more than two-fold decay within the studied time interval (0 – 30 min post-treatment). Once again, the global pairwise comparison of the transcripts' $t_{1/2}$ revealed no correlation between the two strains (Pearson correlation coefficient, P = 0.546; R^2 = 0.2978) (Figure 5.5).

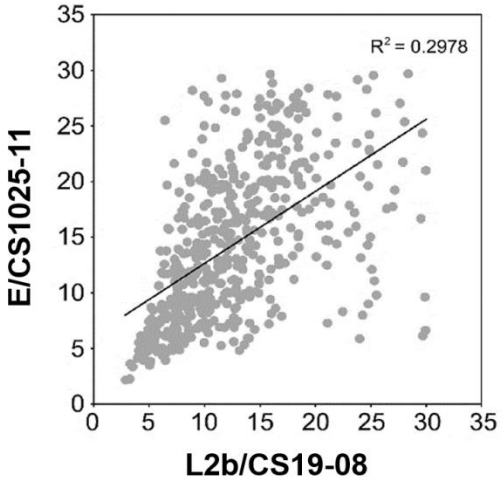


Figure 5.5. Global pairwise comparison of transcripts' $t_{1/2}$ between L2b/CS19-08 (horizontal axis) and E/CS1025-11 (vertical axis) strains. Linear correlation is also displayed (R^2).

As this can eventually be a result of the excessive sensitivity of the mathematical extrapolation to determine the $t_{1/2}$ values, we again focused on mRNA stability trends at functional category level. In fact, this approach is corroborated by previous genome-wide studies, which suggested some degree of correlation between gene function and mRNAs' $t_{1/2}$ in several organisms [292,322,326]. We observed statistically significant differences of $t_{1/2}$ values between strains for three functional categories (Figure 5.6): "Translation, ribosomal structure and biogenesis" ($P < 0.001$), "Energy production and conversion" ($P < 0.05$), and "DNA replication, recombination, and repair" ($P < 0.05$).

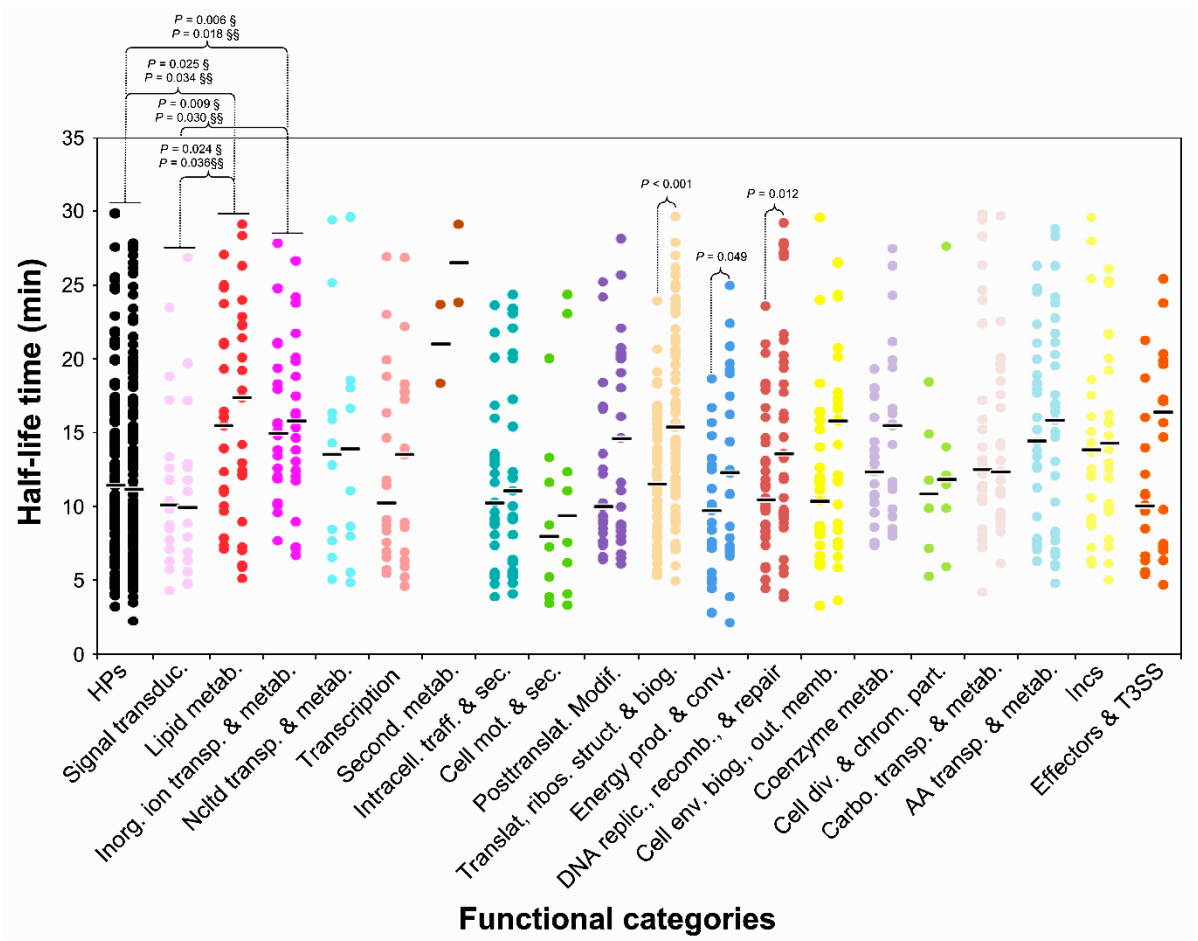


Figure 5.6. Representation of the $t_{1/2}$ of the 525 genes, grouped according to their functional category (in different colours), for the L2b/CS19-08 (left) and E/CS1025-11 (right) strains. The horizontal small dashes represent the median of the $t_{1/2}$ values for that functional category, for each strain. Statistically significant differences between groups of $t_{1/2}$ are indicated in the graph with the respective *P*. Within strains comparisons are discriminated with § (L2b/CS19-08) and §§ (E/CS1025-11).

Interestingly, for both strains, the "DNA replication, recombination, and repair" was also found to be one of the categories with the lowest medians of gene expression, whereas the "Translation, ribosomal structure and biogenesis" category was found to be one with the highest medians of gene expression, (see section "Expression analysis between strains" of this chapter). However, both categories present statistically significant $t_{1/2}$ differences between *C. trachomatis* strains from the two biovars, meaning that, despite the mRNA abundance, both groups of housekeeping genes are differently regulated by these strains at the mid-stage of the developmental cycle. Taking into account the properties of these functional categories, the observed differences may account for dissimilar metabolic performance between two biovars' strains potentially impacting the well-known development and pathogenicity dissimilarities. We also checked for significant differences in the $t_{1/2}$ values between functional categories within each strain, and found that, regardless of the biovar, "Lipid metabolism"

and “Inorganic ion transport and metabolism” displayed $t_{1/2}$ values significantly higher than “HPs” and “Signal transduction mechanisms” (Figure 5.6). This could suggest that transcripts acting in specific networks need to be maintained for longer periods, hypothetically in order to guarantee that the production of specific proteins is not interrupted. This feature seems to be contingent on the species rather than on the biovar.

5.5.3. Comparison between expression level and $t_{1/2}$

Finally, we looked at the relation between the expression level at the mid-stage of the developmental cycle (T_0) and the $t_{1/2}$ calculated for each transcript of both strains (L2b/CS19-08 and E/CS1025-11) (Supplemental Figure 5.3).

As shown in Supplemental Figure 5.3, and also revealed by the Pearson correlation coefficients calculated for the same pairs of comparisons, it is evident the lack of correlation between the stability of a transcript and its level of expression. Even when taken into account only the genes with the lowest $t_{1/2}$ (< 10 min) (Supplemental Figure 5.3A.3 and Supplemental Figure 5.3B.3), i.e., those less prone to be biased by both the sensitivity of the formula applied for the $t_{1/2}$ calculations and the period of time used, there was still no observed relation with the level of expression. Previous studies with different bacteria had also found negative or no correlation between these two parameters (expression and $t_{1/2}$) [293,296,297]. The negative correlation was suggested to be a consequence of the bacterial need for a quick turnover of highly abundant transcripts. We raise the hypothesis that, for an organism with a unique biphasic and multi-stage life-cycle as *C. trachomatis*, the overall lack of correlation may be just a consequence of the tight transcriptional control that is continuously operating during such a complex development. Another fact contributing to this trend may be the intricate host-pathogen interactions that take place for an obligate intracellular pathogen, which implies its continuous and dynamic adaptation. Nevertheless, we found sporadic cases in which a negative correlation was observed, with only four genes being common to both biovars. Whereas CT421.2 and CT559 presented high expression levels and low mRNA stability, CT219 and CT157 revealed the opposite scenario (analysis strictly focused on the first 100 upper and bottom ranked genes). On the other hand, Bernstein and colleagues [293] also found a group of highly abundant transcripts that were highly stable, which does not fit the mentioned “negative correlation”. Curiously, although we found some transcripts following this pattern, none of them was shared by the two biovars (data not shown). Overall, considering that a highly abundant mRNA may or may not be as susceptible to degradation as a scarce mRNA, it implies that abundance alone is not a trigger for their quick-degradation. Noteworthy, previous studies were also unsuccessful in establishing independent associations between the stability of the transcripts and other natural mRNA characteristics, like the existence of secondary structures, G+C content, density of the ribonucleases cleavage sites, several 5’ and 3’ UTRs’ features, protective effect of ribosomes, codon composition, transcript length, and the above mentioned level of expression [292,293,296,297]. Altogether, our results reinforced the fact that the interpretation of the transcripts’ degradation process is not

straightforward as it does not rely on general basic features but on accurate, complex and specific mechanisms whose regulation still needs to be elucidated. One example of this complex dynamics is the guided transcripts' degradation mediated by specific small RNAs, a notion that is becoming more and more consolidated over the past few years (reviewed in [327].

5.6. Conclusion

The present study constitutes the first comparative overview of expression levels and decay rates of the transcripts using different-biovar *C. trachomatis* strains, shedding some light on the transcriptome molecular dynamics of its obligatory developmental cycle. We observed that a substantial proportion of the highest expressed genes at the mid-phase stage of the developmental cycle is shared by all strains (regardless their biovar), reinforcing their importance on behalf of such unique biology that characterizes this obligate intracellular pathogen. We also observed that the degree of transcripts' stability seems to correlate with the bacterial intracellular life-style, as *C. trachomatis* revealed lower transcript decay rates than facultative intracellular and free-living bacteria. Regarding biovar comparisons, only at very few instances (for the functional categories-based analysis) was possible to unveil dissimilarities potentially underlying phenotypic differences. Altogether, interpreting the transcriptome dynamics of *C. trachomatis* is far from being a straightforward task, mostly due to the particularities of this bacterium regarding the unique biphasic developmental cycle and the intricate host-pathogen interactions, which probably exacerbate the observed complexity of its regulation process.

Acknowledgments

RF is the recipient of a Ph.D. fellowship (SFRH/BD/68532/2010) from the FCT/MEC.

Chapter VI

Final overview, concluding remarks and future perspectives

6. Final overview, concluding remarks and future perspectives

The study of the strict obligate intracellular human pathogen *C. trachomatis* holds great interest given its huge impact on public health. However, unveiling the biological idiossincrasies of this bacterium throught the use of conventional microbiology and molecular biology techniques have proven to be very hard. To overcome such difficulties, genomics, transcriptomics and bioinformatics were the approaches of choice throughout the period of this thesis to shed some light on the molecular dynamics of the chromosome and the plasmid of *C. trachomatis*, which potentially underlie the different-serovar strains' specific phenotypes.

As the large scale whole-genome sequencing of several organisms, in particular *C. trachomatis*, is a fairly recent advent, few portions of a genome, traditionally MLST, were used as the prime subjects, at the time of the study described in the chapter II. This procedure allowed us to infer evolutionary parameters, such as mutation and recombination rates, essential for understanding how microorganisms, in particular *C. trachomatis*, achieve fitted phenotypes. However, this approach was not only restrictive of the genome variability evaluated but was also found to yield different results within same-species. Thus, and by using *C. trachomatis* as a study model, we assessed how computational mutation and recombination estimations are shaped by loci with different genetic features, as bacterial genomes exhibit a highly heterogeneous representation of those loci. We found that the estimation of mutation and recombination rates, in *C. trachomatis*, is indeed influenced by the characteristics of the selected loci. In particular, highly polymorphic and positively selected genes revealed to be more prone to confound algorithms as they yielded non-reproductible estimates and incongruent serovars phylogenies. Also, noncoding regions were found to shape estimations similarly to the housekeeping genes, probably due to the promotors and regulatory regions they often carry. These observations rose awereness to the indiscriminate applicability of alghoritms that were being used at that time. Moreover, and with the novel availability of new whole-genome sequences, we anticipated their subsequent use in such estimations, as they encompass the complete species' genetic variability information and thus allow more robust calculations, like the ones we further obtained using the "wide genomic approach". Nevertheless, and taking into account our results, we consider that evolutionary parameters' estimations should always be looked with caution as bacteria-specific genomic architectures may differentially buffer the effect of the confounding factors that each genome contains, and thus the "one size fits all" approach may not always be applicable.

Even though this first study (Chapter II) mainly intended to evaluate the relative weight of different sets of genes on the estimations of evolutionary parameters, it also allowed the confirmation of *C. trachomatis* low recombination and mutation rates, which could be expected considering the unique biology of this bacterium. Being an obligate intracellular pathogen that replicates within an inclusion, any recombination event requires a host-cell coinfection by distinct microorganisms (estimated

frequency of 1%), followed by the fusion of both inclusion vacuoles. Additionally, recombination would also introduce little diversity in the recipient microorganism, given the high genomic similarity degree of different-serovar strains, in which the polymorphism (< 2%) is provided by few highly variable loci dispersed throughout the chromosome. Even new mutations may easily become deleterious and disappear before being accounted for, as *C. trachomatis* is considered to be under the final stages of the evolutionary process of genome reduction, resulting in a highly compact chromosome, specialized in dealing with the challenges of intracellular survival. In this regard, the low genetic variability that different serovars exhibit must determine the dissimilar phenotypes they display in terms of cell-appetence, disease outcome and ecological success. However, the knowledge of the molecular basis underlying serovars' specificities is scarce. Knowing that previous specific mutational patterns of some genes seemed to be associated with serovars' niche specificity, and given that more than 50 *C. trachomatis* genomes had become available by that time, we took the opportunity to perform comparative genomics to examine all the ~900 genes of this bacterium to evaluate the putative association of gene's individual phylogenies with clinical outcome, cell-appetence and/or ecological success of the serovars (Chapter III). We believed that the identification of genes that contribute for the main branches of the species phylogenetic tree would be highly relevant for directing future functional studies. Surprisingly, we found a very low proportion of genes (~1.4%) whose proteins presented a tree topology with a plain segregation of the strains according to their cell-appetence (ocular conjunctiva, genital epithelium, and lymph nodes). We speculated that this could possibly be a consequence of the occurrence of intra- and intergenic recombination events during mixed infections, although they take place at a very low frequency, or possibly because these genes could have evolved more quickly than the remaining genome due to host pressure. On the other hand, when taking into accont only the segregation of same disease-causing serovars, we observed a higher proportion of genes. This was strikingly evident for the LGV serovars, for which almost all chromosomic genes (~80%) segregated them as a distinct clade, corroborating their early divergence from the remainder serovars and/or their fastest evolutionary nature. Notably, about ~28% of the genes segregated LGV strains in an exclusive manner, and mostly code for proteins that are responsible for establishing interactions with the host (like putative T3SS effectors and Inc proteins), which seems consistent with the fact that, unlike other serovars, the LGVs would have to be equipped with specific means for interacting with both epithelial cells and mononuclear phagocytes. Furthermore, serovars E and F were also expected to have a particular genetic makeup, as they display a distinctive higher prevalence than the remainder genital serovars, and were indeed found to be exclusively segregated by several genes. However, no specific associations with prevalence could be made because the majority of those genes encode intermediate of basic cellular functions. Curiously, the most prevalent genital serovars and the LGV were found to share hundreds of polymorphisms that resulted in co-segregation of the two groups by several genes, which could be explained by incomplete lineage sorting, recombination and/or coevolutionary process, although the true impact of such shared mutational patterns in one or the other group remains unclear.

Another interesting finding was the identification of clade-specific gene lengths and pseudogenes, which was suggestive of their putative expendability for the infection of particular niches and that further genome reduction may still be ongoing in *C. trachomatis*. Of note was also the fact that the most polymorphic genes code essentially for membrane and hypothetical proteins, whereas the genes displaying significant dN/dS > 1 mainly code for Incs and T3SS effectors. These findings seem to corroborate the assumption that proteins involved in strict pathogen-host interactions are more prone to fix nonsynonymous mutations, and that polymorphism may be due to discrete genetic drift, as it was mostly given by the synonymous substitutions *per* synonymous sites. Overall, this was an extensive study that allowed the identification of several genetic features that distinguish serovars in terms of cell-appetence and prevalence, in a gene-by-gene manner. As it encompassed all *C. trachomatis* chromosomic genes, the results certainly constitute an important database of putative targets for developing future functional studies, aiming the clarification of their biological role on chlamydial infection in terms of tissue tropism, virulence and ecological success.

Still, the study of *C. trachomatis* genome would not be complete without the inclusion of its sole plasmid. As *C. trachomatis* maintains this highly conserved DNA molecule, despite the ongoing process of chromosomal size-reduction, its significance is unquestionable for the chlamydial infection. In fact, this plasmid had been pointed out as a primary regulator of chromosomal genes, but there were no experimental evidences, at the time the study described in Chapter IV was performed, associating the plasmid with strains' differential tissue tropism, although it was already shown that both plasmid and chromosome display a parallel evolution within same-serovar strains. Therefore, we decided to investigate whether plasmid characteristics, and putative associations between them, namely plasmid copy number *per* cell, and expression profile of the eight plasmid genes and its two sRNAs, would underlie serovars tropism. We verified that, regardless the tissue tropism of the strains, 1.0–7.4 plasmid copies were always present *per* bacterial genome, and the highest number of copies was observed at the replicative stage of the developmental cycle. This suggests that plasmid load does not reflect expression needs and does not reflect tissue tropism, but this rather ensures its transmission to the daughter-cells. Like for some chromosomal genes whose products are also known to interact with the host, the plasmid encoded immunodominant antigen Pgp3, which is secreted into the host cytosol, presented an overrepresentation of nonsynonymous mutations. In terms of gene expression, despite there were overall significant expression differences between different disease-causing serovars, we found that ORF6/*pgp4* presented the highest mean expression among all strains. This interesting observation fits the fact that this gene is transcribed from two different promoters and highlights its importance as a transcriptional regulator of virulence-associated genes. Moreover, both sRNAs were highly expressed relative to the genes, specially at very early stages of the *C. trachomatis* developmental cycle. Overall, the study of the plasmid revealed that its transcriptional dynamics appears to sustain serovars' tropism differences, although plasmid copy number was not informative in that regard.

Chapter VI

In Chapter V we intended to undertake an unprecedented comparative overview of expression levels and decay rates of all *C. trachomatis* mRNAs. In fact, transcriptomics had proved valuable for deciphering the biological role of several chlamydial genes, and the interpretation of bacterial transcriptomics should contribute for understanding the dynamics between de novo transcription and degradation rate of transcripts. The expression analysis showed that, at the mid-phase stage of the developmental cycle, strains (regardless the biovar) shared several of the highest expressed genes, which suggests their significance for the biological uniqueness of the obligate intracellular *C. trachomatis*. On the other hand, we also observed that this group of the most expressed genes was mainly composed of HPs for the LGV-biovar strain, whereas for the trachoma-biovar strains (specially the non-prevalent genital) the representation of HPs was much lower. Furthermore, transcripts revealed an unusually high overall stability when compared to other bacteria, which led us to extrapolate a putative association between the degree of transcripts' stability with the bacterial intracelluler life-style. We also found that the genes encoding hypothetical proteins were highly represented amongst the most stable transcripts, althought only six were common between the two different-biovar strains compared. On the other hand, when considering the most unstable transcripts, we verified that there was no statistically significant correlation between strains. However, when those same transcripts were grouped according to their biological function, three of these groups revealed statistically significant differences between strains. Overall, the study described in Chapter V intended to highlight differences and similarities between different-serovar strains at transcriptome level, but due to *C. trachomatis* unique biological features, the complex interpretation of the transcriptome dynamics raises singular challenges, hampering straightforward associations and conclusions.

In conclusion, we believe that the findings presented throughout this Ph.D. thesis contribute to a better understanding of the *C. trachomatis* chromosome and plasmid molecular dynamics sustaining strains phenotypic differences. However, much is still to unveil in order to completely understand the ways this unique human pathogen survives and persists, and it will demand steady steps in the bumpy ride of chlamydial research. The extended knowledge on *C. trachomatis* genomic and transcriptomic backgrounds, and considering the ongoing progress in the fields of molecular biology and technology, deciphering such questions may be imminent.

Meanwhile, several interesting observations arose during the course of this Ph.D. thesis which we believe would be worth exploring. Taking advantage of the recent progress in genetic manipulation and mutagenesis technologies of *C. trachomatis*, together with the genomic and transcriptomic information compiled so far, such specific lines of work would include:

- Clarifying the biological role of each of the proteins that directly interact with the host, like Incs and T3SS effectors. Immediate candidates should encompass the genes identified during this Ph.D.

thesis as being highly polymorphic and/or presenting dN/dS > 1, specially if their genetic variability sustains phylogenetic tree branches of a given clade (ocular, genital or LGV serovars), as these may be the genes modulating the specific serovars' cell-appetence and disease outcome;

- Understanding the biological role of the genes that were found to be putative pseudogenes for all the strains of a specific disease-group (chapter III), and further evaluate the true extent of their expendability in those strains;
- Full characterization of the two sRNAs encoded by the plasmid, as they displayed intriguing expression features;
- Elucidating the biological role of HPs, as they constitute an enigmatic group of proteins (genus- or species-specific), whose genes are highly represented amongst the most polymorphic ones and whose mRNAs are amongst the most stable of *C. trachomatis*. Altogether, these observations seem to indicate that HPs may not only play a relevant role in the unique biology of *C. trachomatis* but, may also contribute to the different pathobiology displayed by the different serovars.

References

1. Halberstaedter L, von Prowazek S: **Zur atiologie des trachoms.** *Deutsche Medizinische Wochenschrift* 1907, **33**:1285–1287.

2. Lindner K: **Zur atiologie der gonokokken-freien urethritis**. *Klin Wochenschr* 1910, **8**:283–284.

3. Durand NJ, Nicolas J, Favre M: **Lymphogranulomatose inguinale subaiguë d'origine génitale probable, peut-être vénérienne**. *Bulletin de la Société des Médecins des Hôpitaux de Paris* 1913, **35**:274-288.

4. Bedson SP, Western GT, Simpson SL: **Observations on the aetiology of psittacosis**. *Lancet* 1930, **1**:235-236.

5. Coles AC: **Micro-organisms in psittacosis**. *Lancet* 1930, **215**:1011–1012.

6. Lillie RD: **PsittacosisS: Rickettsia-like inclusions in man and in experimental animals**. *Public Health Reports* 1930, **45**:773-778.

7. Miyagawa Y, Mitamura T, Yaoi H, Ishii N, Okanishi J: **Fourth report: studies on the virus of lymphogranuloma inguinale Nicolas, Favre and Durand. Cultivation of the virus on the chrorioallantoic membrane of the chicken embryo**. *The Japanese journal of experimental medicine* 1935, **13** 733-738.

8. Moulder JW: **Relation of psittacosis group (Chlamydiae) to bacteria and viruses**. *Annual Review of Microbiology* 1966, **20**:107-&.

9. Pace NR: **A molecular view of microbial diversity and the biosphere**. *Science* 1997, **276**:734-740.

10. Clarke IN: **Evolution of *Chlamydia trachomatis***. *Evolution of Infectious Agents in Relation to Sex* 2011, **1230**:E11-E18.

11. Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D, et al.: **Illuminating the evolutionary history of chlamydiae**. *Science* 2004, **304**:728-730.

12. Everett KDE, Bush RM, Andersen AA: **Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms**. *International Journal of Systematic Bacteriology* 1999, **49**:415-440.

13. Schachter J, Stephens RS, Timms P, Kuo C, Bavoil PM, Birkelund S, Boman J, Caldwell H, Campbell LA, Chernesky M, et al.: **Radical changes to chlamydial taxonomy are not necessary just yet**. *International Journal of Systematic and Evolutionary Microbiology* 2001, **51**:249-249.

14. Stephens RS, Myers G, Eppinger M, Bavoil PM: **Divergence without difference: phylogenetics and taxonomy of *Chlamydia* resolved**. *Fems Immunology and Medical Microbiology* 2009, **55**:115-119.

15. Griffiths E, Ventresca MS, Gupta RS: **BLAST screening of chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, *Chlamydophila* and *Chlamydia* groups of species**. *Bmc Genomics* 2006, **7**.

16. Gupta RS, Griffiths E: **Chlamydiae-specific proteins and indels: novel tools for studies**. *Trends in Microbiology* 2006, **14**:527-535.

17. Nunes A, Gomes JP: **Evolution, phylogeny, and molecular epidemiology of *Chlamydia***. *Infection Genetics and Evolution* 2014, **23**:49-64.

18. Li Z, Zhong Y, Lei L, Wu Y, Wang S, Zhong G: **Antibodies from women urogenitally infected with *C. trachomatis* predominantly recognized the plasmid protein Pgp3 in a conformation-dependent manner**. *Bmc Microbiology* 2008, **8**.

19. Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, et al.: **Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39**. *Nucleic Acids Res* 2000, **28**:1397-1406.

20. Barron AL, White HJ, Rank RG, Soloff BL, Moses EB: **A new animal model for the study of *Chlamydia trachomatis* genital infections - Infection of mice with the agent of mouse pneumonitis**. *Journal of Infectious Diseases* 1981, **143**:63-66.

21. Farris CM, Morrison RP: **Vaccination against *Chlamydia* genital infection utilizing the murine *C. muridarum* Model**. *Infection and Immunity* 2011, **79**:986-996.

22. Yu H, Jiang X, Shen C, Karunakaran KP, Brunham RC: **Novel *Chlamydia muridarum* T Cell Antigens Induce Protective Immunity against Lung and Genital Tract Infection in Murine Models**. *Journal of Immunology* 2009, **182**:1602-1608.

23. O'Connell CM, Ingalls RR, Andrews CW, Jr., Scurlock AM, Darville T: **Plasmid-deficient *Chlamydia muridarum* fail to induce immune pathology and protect against oviduct disease**. *Journal of Immunology* 2007, **179**:4027-4034.

24. Steiper ME, Young NM: **Primate molecular divergence dates**. *Mol Phylogenet Evol* 2006, **41**:384-394.

25. Grayston JT, Wang SP: **New knowledge of Chlamydiae and diseases they caus**. *Journal of Infectious Diseases* 1975, **132**:87-105.

26. Caldwell HD, Kromhout J, Schachter J: **Purification and partial characterization of the Major Outer-Membrane Protein of *Chlamydia trachomatis***. *Infection and Immunity* 1981, **31**:1161-1176.

27. Nunes A, Borrego MJ, Nunes B, Florindo C, Gomes JP: **Evolutionary Dynamics of ompA, the Gene Encoding the *Chlamydia trachomatis* Key Antigen**. *Journal of Bacteriology* 2009, **191**:7182-7192.

28. Nunes A, Nogueira PJ, Borrego MJ, Gomes JP: **Adaptive evolution of the *Chlamydia trachomatis* dominant antigen reveals distinct evolutionary scenarios for B- and T-cell epitopes: worldwide survey**. *PloS one* 2010, **5**.

29. Schachter J: **Infection and disease epidemiology**. In *Chlamydia*. Edited by R S: ASM Press; 1999:139-169.

30. Harris SR, Clarke IN, Seth-Smith HMB, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW, et al.: **Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing**. *Nature Genetics* 2012, **44**:413-U221.

31. Joseph SJ, Didelot X, Rothschild J, de Vries HJC, Morre SA, Read TD, Dean D: **Population Genomics of *Chlamydia trachomatis*: Insights on Drift, Selection, Recombination, and Population Structure**. *Molecular Biology and Evolution* 2012, **29**:3933-3946.

32. Nunes A, Nogueira PJ, Borrego MJ, Gomes JP: ***Chlamydia trachomatis* diversity viewed as a tissue-specific coevolutionary arms race**. *Genome Biology* 2008, **9**.

33. Bebear C, de Barbeyrac B: **Genital *Chlamydia trachomatis* infections**. *Clinical Microbiology and Infection* 2009, **15**:4-10.

34. World Health Organization: **Global incidence and prevalence of selected curable sexually transmitted infections - 2008**. Edited by. (World Health Organization, Geneva, Switzerland, 2012) 2012.

35. Manavi K: **A review on infection with *Chlamydia trachomatis***. *Best Practice & Research in Clinical Obstetrics & Gynaecology* 2006, **20**:941-951.

36. Vasilevsky S, Greub G, Nardelli-Haefliger D, Baud D: **Genital *Chlamydia trachomatis*: Understanding the Roles of Innate and Adaptive Immunity in Vaccine Research**. *Clinical Microbiology Reviews* 2014, **27**:346-370.

37. AbdelRahman YM, Belland RJ: **The chlamydial developmental cycle**. *Fems Microbiology Reviews* 2005, **29**:949-959.

38. Hogan RJ, Mathews SA, Mukhopadhyay S, Summersgill JT, Timms P: **Chlamydial persistence: beyond the biphasic paradigm**. *Infection and Immunity* 2004, **72**:1843-1855.

39. Beem MO, Saxon EM: **Respiratory-Tract Colonization and a Distinctive Pneumonia Syndrome in Infants Infected with *Chlamydia trachomatis***. *New England Journal of Medicine* 1977, **296**:306-310.

40. Schachter J, Grossman M, Sweet RL, Holt J, Jordan C, Bishop E: **Prospective-study of perinatal transmission of *Chlamydia trachomatis***. *Jama-Journal of the American Medical Association* 1986, **255**:3374-3377.

41. Stephens RS: **The cellular paradigm of chlamydial pathogenesis**. *Trends in Microbiology* 2003, **11**:44-51.

42. Rasmussen SJ, Eckmann L, Quayle AJ, Shen L, Zhang YX, Anderson DJ, Fierer J, Stephens RS, Kagnoff MF: **Secretion of proinflammatory cytokines by epithelial cells in response to *Chlamydia* infection suggests a central role for epithelial cells in chlamydial pathogenesis**. *J Clin Invest* 1997, **99**:77-87.

43. Darville T, Hiltke TJ: **Pathogenesis of Genital Tract Disease Due to *Chlamydia trachomatis***. *Journal of Infectious Diseases* 2010, **201**:S114-S125.

44. Brunham RC, Paavonen J, Stevens CE, Kiviat N, Kuo CC, Critchlow CW, Holmes KK: **Mucopurulent cervicitis - the ignored counterpart in women of urethritis in men**. *New England Journal of Medicine* 1984, **311**:1-6.

45. Cates W, Wasserheit JN: **Genital chlamydial infections - epidemiology and reproductive sequelae**. *American Journal of Obstetrics and Gynecology* 1991, **164**:1771-1781.

46. El Hakim EA, Gordon UD, Akande VA: **The relationship between serum *Chlamydia* antibody levels and severity of disease in infertile women with tubal damage**. *Archives of Gynecology and Obstetrics* 2010, **281**:727-733.

47. Malhotra M, Sood S, Mukherjee A, Muralidhar S, Bala M: **Genital *Chlamydia trachomatis*: An update**. *Indian Journal of Medical Research* 2013, **138**:303-316.

48. Miller KE: **Diagnosis and treatment of *Chlamydia trachomatis* infection**. *American Family Physician* 2006, **73**:1411-1416.

49. Mylonas I: **Female genital *Chlamydia trachomatis* infection: where are we heading?** *Archives of Gynecology and Obstetrics* 2012, **285**:1271-1285.

50. Mardh PA, Ripa T, Svensson L, Westrom L: ***Chlamydia trachomatis* infection in patients with acute salpingitis**. *New England Journal of Medicine* 1977, **296**:1377-1379.

51. Paavonen J, Eggert-Kruse W: ***Chlamydia trachomatis*: impact on human reproduction**. *Human Reproduction Update* 1999, **5**:433-447.

52. Zeidler H, Kuipers J, Köhler L: ***Chlamydia*-induced arthritis**. *Curr Opin Rheumatol* 2004, **16**:380-392.

53. Hu VH, Holland MJ, Burton MJ: **Trachoma: Protective and Pathogenic Ocular Immune Responses to *Chlamydia trachomatis***. *Plos Neglected Tropical Diseases* 2013, **7**.

54. Mariotti SP, Pascolini D, Rose-Nussbaumer J: **Trachoma: global magnitude of a preventable cause of blindness**. *Br J Ophthalmol* 2009, **93**:563-568.

55. World Health Organization: **Trachoma Fact sheet N°382**. Edited by World Health Organization (2015).

56. Taylor HR, Burton MJ, Haddad D, West S, Wright H: **Trachoma**. *Lancet* 2014, **384**:2142-2152.

57. Alemayehu M, Koye DN, Tariku A, Yimam K: **Prevalence of Active Trachoma and Its Associated Factors among Rural and Urban Children in Dera Woreda, Northwest Ethiopia: A Comparative Cross-Sectional Study**. *BioMed research international* 2015, **2015**:570898-570898.

58. Mabey D, Peeling RW: **Lymphogranuloma venereum**. *Sex Transm Infect* 2002, **78**:90-92.

59. McLean CA, Stoner BP, Workowski KA: **Treatment of lymphogranuloma venereum**. *Clin Infect Dis* 2007, **44 Suppl 3**:S147-152.

60. Rönn MM, Ward H: **The association between lymphogranuloma venereum and HIV among men who have sex with men: systematic review and meta-analysis**. *BMC Infect Dis* 2011, **11**:70.

61. Ahdoot A, Kotler DP, Suh JS, Kutler C, Flamholz R: **Lymphogranuloma venereum in human immunodeficiency virus-infected individuals in New York City**. *J Clin Gastroenterol* 2006, **40**:385-390.

62. de Vries HJ, Zingoni A, Kreuter A, Moi H, White JA, Infections EBotIUaST, Venereology EAoDa, Forum ED, Diseases ESoCMaI, Specialists UoEM, et al.: **2013 European guideline on the management of lymphogranuloma venereum**. *J Eur Acad Dermatol Venereol* 2015, **29**:1-6.

63. de Vrieze NH, de Vries HJ: **Lymphogranuloma venereum among men who have sex with men. An epidemiological and clinical review**. *Expert Rev Anti Infect Ther* 2014, **12**:697-704.

64. Götz HM, Ossewaarde JM, Nieuwenhuis RF, van der Meijden WI, Dees J, Thio B, de Zwart O, van de Laar MJ: **A cluster of lymphogranuloma venereum among homosexual men in Rotterdam with implications for other countries in Western Europe**. *Ned Tijdschr Geneeskd* 2004, **148**:441-442.

65. Kropp RY, Wong T, Group CLW: **Emergence of lymphogranuloma venereum in Canada**. *CMAJ* 2005, **172**:1674-1676.

66. Morton AN, Fairley CK, Zaia AM, Chen MY: **Anorectal lymphogranuloma venereum in a Melbourne man**. *Sex Health* 2006, **3**:189-190.

67. Ceovic R, Gulin SJ: **Lymphogranuloma venereum: diagnostic and treatment challenges**. *Infect Drug Resist* 2015, **8**:39-47.

68. Danta M, Brown D, Bhagani S, Pybus OG, Sabin CA, Nelson M, Fisher M, Johnson AM, Dusheiko GM, group HaAHH: **Recent epidemic of acute hepatitis C virus in HIV-positive men who have sex with men linked to high-risk sexual behaviours**. *AIDS* 2007, **21**:983-991.

69. Fleming DT, Wasserheit JN: **From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection**. *Sex Transm Infect* 1999, **75**:3-17.

70. Bedson SP, Bland JOW: **A morphological study of psittacosis virus, with the description of a developmental cycle**. *The British Journal of Experimental Pathology* 1932, **13**:461–466.

71. Hackstadt T, Todd WJ, Caldwell HD: **Disulfide-mediated interactions of the chlamydial major outer-membrane protein - role in the differentiation of chlamydiae**. *Journal of Bacteriology* 1985, **161**:25-31.

72. Omsland A, Sager J, Nair V, Sturdevant DE, Hackstadt T: **Developmental stage-specific metabolic and transcriptional activity of *Chlamydia trachomatis* in an axenic medium**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:19781-19785.

73. Schachter J, Caldwell HD: **Chlamydiae**. *Annu Rev Microbiol* 1980, **34**:285-309.

74. Newhall WJ, Jones RB: **Disulfide-linked oligomers of the major outer membrane protein of chlamydiae**. *J Bacteriol* 1983, **154**:998-1001.

75. Egan AJ, Vollmer W: **The physiology of bacterial cell division**. *Ann N Y Acad Sci* 2013, **1277**:8-28.

76. Fox A, Rogers JC, Gilbart J, Morgan S, Davis CH, Knight S, Wyrick PB: **Muramic acid is not detectable in *Chlamydia psittaci* or *Chlamydia trachomatis* by gas chromatography-mass spectrometry**. *Infect Immun* 1990, **58**:835-837.

77. Moulder JW: **Why is *Chlamydia* sensitive to penicillin in the absence of peptidoglycan?** *Infect Agents Dis* 1993, **2**:87-99.

78. How SJ, Hobson D, Hart CA: **Studies *in vitro* of the nature and synthesis of the cell wall of *Chlamydia trachomatis*** *Current Microbiology* 1984, **10**:269-274.

79. Hammerschlag MR, Gleyzer A: ***In vitro* activity of a group of broad-spectrum cephalosporins and other beta-lactam antibiotics against *Chlamydia trachomatis***. *Antimicrob Agents Chemother* 1983, **23**:493-494.

80. Johnson FW, Hobson D: **The effect of penicillin on genital strains of *Chlamydia trachomatis* in tissue culture**. *J Antimicrob Chemother* 1977, **3**:49-56.

81. Storey C, Chopra I: **Affinities of beta-lactams for penicillin binding proteins of *Chlamydia trachomatis* and their antichlamydial activities**. *Antimicrob Agents Chemother* 2001, **45**:303-305.

82. Chopra I, Storey C, Falla TJ, Pearce JH: **Antibiotics, peptidoglycan synthesis and genomics: the chlamydial anomaly revisited**. *Microbiology* 1998, **144 (Pt 10)**:2673-2678.

83. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao QX, et al.: **Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis***. *Science* 1998, **282**:754-759.

84. Thomson NR, Holden MTG, Carder C, Lennard N, Lockey SJ, Marsh P, Skipp P, O'Connor CD, Goodhead I, Norbertzcak H, et al.: ***Chlamydia trachomatis*: Genome sequence analysis of lymphogranuloma venereum isolates**. *Genome Research* 2008, **18**:161-171.

85. Belland R, Zhong G, Crane D, Hogan D, Sturdevant D, Sharma J, Beatty W, Caldwell H: **Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis***. *Proceedings of the National Academy of Sciences* 2003, **100**:8478-8483.

86. Nicholson T, Olinger L, Chong K, Schoolnik G, Stephens R: **Global stage-specific gene regulation during the developmental cycle of *Chlamydia trachomatis***. *Journal of Bacteriology* 2003, **185**:3179-3189.

87. Barbour AG, Amano K, Hackstadt T, Perry L, Caldwell HD: ***Chlamydia trachomatis* has penicillin-binding proteins but not detectable muramic acid**. *J Bacteriol* 1982, **151**:420-428.

88. Liechti GW, Kuru E, Hall E, Kalinda A, Brun YV, VanNieuwenhze M, Maurelli AT: **A new metabolic cell-wall labelling method reveals peptidoglycan in *Chlamydia trachomatis***. *Nature* 2014, **506**:507-510.

89. Chen JCR, Stephens RS: **Trachoma and lgv biovars of *Chlamydia trachomatis* share the same glycosaminoglycan-dependent mechanism for infection of eukaryotic cells**. *Molecular Microbiology* 1994, **11**:501-507.

90. Chen JCR, Stephens RS: ***Chlamydia trachomatis*** **glycosaminoglycan-dependent and independent attachment to eukaryotic cells**. *Microbial Pathogenesis* 1997, **22**:23-30.

91. Chen JCR, Zhang JP, Stephens RS: **Structural requirements of heparin binding to *Chlamydia trachomatis***. *Journal of Biological Chemistry* 1996, **271**:11134-11140.

92. Wuppermann FN, Hegemann JH, Jantos CA: **Heparan sulfate-like glycosaminoglycan is a cellular receptor for *Chlamydia pneumoniae***. *Journal of Infectious Diseases* 2001, **184**:181-187.

93. Cocchiaro JL, Valdivia RH: **New insights into *Chlamydia* intracellular survival mechanisms**. *Cellular Microbiology* 2009, **11**:1571-1578.

94. Dautry-Varsat A, Subtil A, Hackstadt T: **Recent insights into the mechanisms of *Chlamydia* entry**. *Cellular Microbiology* 2005, **7**:1714-1722.

95. Abromaitis S, Stephens RS: **Attachment and Entry of *Chlamydia* Have Distinct Requirements for Host Protein Disulfide Isomerase**. *Plos Pathogens* 2009, **5**.

96. Bastidas RJ, Elwell CA, Engel JN, Valdivia RH: **Chlamydial Intracellular Survival Strategies**. *Cold Spring Harbor Perspectives in Medicine* 2013, **3**.

97. Menozzi FD, Pethe K, Bifani P, Soncin F, Brennan MJ, Locht C: **Enhanced bacterial virulence through exploitation of host glycosaminoglycans**. *Molecular Microbiology* 2002, **43**:1379-1386.

98. Su H, Raymond L, Rockey DD, Fischer E, Hackstadt T, Caldwell HD: **A recombinant *Chlamydia trachomatis* major outer membrane protein binds to heparan sulfate receptors on epithelial cells**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**:11143-11148.

99. Moelleken K, Schmidt E, Hegemann JH: **Members of the Pmp protein family of *Chlamydia pneumoniae* mediate adhesion to human cells via short repetitive peptide motifs**. *Molecular Microbiology* 2010, **78**:1004-1017.

100. Beeckman DSA, Vanrompay DCG: **Bacterial Secretion Systems with an Emphasis on the Chlamydial Type III Secretion System**. *Current Issues in Molecular Biology* 2010, **12**:17-41.

101. Clifton DR, Fields KA, Grieshaber SS, Dooley CA, Fischer ER, Mead DJ, Carabeo RA, Hackstadt T: **A chlamydial type III translocated protein is tyrosine-phosphorylated at the site of entry and associated with recruitment of actin**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:10166-10171.

102. Jewett TJ, Fischer ER, Mead DJ, Hackstadt T: **Chlamydial TARP is a bacterial nucleator of actin**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:15599-15604.

103. Lane BJ, Mutchler C, Al Khodor S, Grieshaber SS, Carabeo RA: **Chlamydial entry involves TARP binding of guanine nucleotide exchange factors**. *Plos Pathogens* 2008, **4**.

104. Carabeo RA, Dooley CA, Grieshaber SS, Hackstadt T: **Rac interacts with Abi-1 and WAVE2 to promote an Arp2/3-dependent actin recruitment during chlamydial invasion**. *Cellular Microbiology* 2007, **9**:2278-2288.

105. Thalmann J, Janik K, May M, Sommer K, Ebeling J, Hofmann F, Genth H, Klos A: **Actin Re-Organization Induced by *Chlamydia trachomatis* Serovar D - Evidence for a Critical Role of the Effector Protein CT166 Targeting Rac**. *Plos One* 2010, **5**.

106. Hower S, Wolf K, Fields KA: **Evidence that CT694 is a novel *Chlamydia trachomatis* T3S substrate capable of functioning during invasion or early cycle development**. *Molecular Microbiology* 2009, **72**:1423-1437.

107. Hackstadt T: **Redirection of host vesicle trafficking pathways by intracellular parasites**. *Traffic* 2000, **1**:93-99.

108. Clausen JD, Christiansen G, Holst HU, Birkelund S: ***Chlamydia trachomatis* utilizes the host cell microtubule network during early events of infection**. *Molecular Microbiology* 1997, **25**:441-449.

109. Grieshaber SS, Grieshaber NA, Hackstadt T: ***Chlamydia trachomatis* uses host cell dynein to traffic to the microtubule-organizing center in a p50 dynamitin-independent process**. *Journal of Cell Science* 2003, **116**:3793-3802.

110. Damiani MT, Tudela JG, Capmany A: **Targeting eukaryotic Rab proteins: a smart strategy for chlamydial survival and replication**. *Cellular Microbiology* 2014, **16**:1329-1338.

111. Hackstadt T, Scidmore-Carlson MA, Shaw EI, Fischer ER: **The *Chlamydia trachomatis* IncA protein is required for homotypic vesicle fusion**. *Cellular Microbiology* 1999, **1**:119-130.

112. Beatty WL, Byrne GI, Morrison RP: **Morphologic and antigenic characterization of interferon gamma-mediated persistent *Chlamydia trachomatis* infection *in vitro***. *Proc Natl Acad Sci U S A* 1993, **90**:3998-4002.

113. Matsumoto A, Manire GP: **Electron microscopic observations on the effects of penicillin on the morphology of *Chlamydia psittaci***. *J Bacteriol* 1970, **101**:278-285.

114. Coles AM, Reynolds DJ, Harper A, Devitt A, Pearce JH: **Low-nutrient induction of abnormal chlamydial development: a novel component of chlamydial pathogenesis?** *FEMS Microbiol Lett* 1993, **106**:193-200.

115. Beatty WL, Morrison RP, Byrne GI: **Persistent chlamydiae: from cell culture to a paradigm for chlamydial pathogenesis**. *Microbiol Rev* 1994, **58**:686-699.

116. Wyrick PB, Knight ST: **Pre-exposure of infected human endometrial epithelial cells to penicillin *in vitro* renders *Chlamydia trachomatis* refractory to azithromycin**. *Journal of Antimicrobial Chemotherapy* 2004, **54**:79-85.

117. Gieffers J, Fullgraf H, Jahn J, Klinger M, Dalhoff K, Katus HA, Solbach W, Maass M: ***Chlamydia pneumoniae* infection in circulating human monocytes is refractory to antibiotic treatment**. *Circulation* 2001, **103**:351-356.

118. Workowski KA, Lampe MF, Wong KG, Watts MB, Stamm WE: **Long-term eradication of *Chlamydia trachomatis* genital-infection after antimicrobial therapy - evidence against persistent infection**. *Jama-Journal of the American Medical Association* 1993, **270**:2071-2075.

119. Wyrick PB: ***Chlamydia trachomatis* Persistence *In Vitro*: An Overview**. *Journal of Infectious Diseases* 2010, **201**:S88-S95.

120. Kingsbury DT: **Estimate of the genome size of various microorganisms**. *J Bacteriol* 1969, **98**:1400-1401.

121. Sakharkar KR, Dhar PK, Chow VT: **Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis**. *Int J Syst Evol Microbiol* 2004, **54**:1937-1941.

122. Carlson JH, Porcella SF, McClarty G, Caldwell HD: **Comparative genomic analysis of *Chlamydia trachomatis* oculotropic and genitotropic strains**. *Infection and Immunity* 2005, **73**:6407-6418.

123. Kari L, Whitmire WM, Carlson JH, Crane DD, Reveneau N, Nelson DE, Mabey DCW, Bailey RL, Holland MJ, McClarty G, et al.: **Pathogenic diversity among *Chlamydia trachomatis* ocular strains in nonhuman primates is affected by subtle genomic variations**. *Journal of Infectious Diseases* 2008, **197**:449-456.

124. Zomorodipour A, Andersson SGE: **Obligate intracellular parasites: *Rickettsia prowazekii* and *Chlamydia trachomatis***. *Febs Letters* 1999, **452**:11-15.

125. Andersson SGE, Kurland CG: **Reductive evolution of resident genomes**. *Trends in Microbiology* 1998, **6**:263-268.

126. Lovett M, Kuo CC, Holmes K, Falkow S: **Plasmids of the genus *Chlamydia***. In *Current Chemotherapy and Infectious Diseases*. Edited by Nelson J, Grassi C: American Society for Microbiology; 1980:1250-1252.

127. Comanducci M, Ricci S, Ratti G: **The structure of a plasmid of *Chlamydia trachomatis* believed to be required for growth within mammalian cells**. *Mol Microbiol* 1988, **2**:531-538.

128. Seth-Smith HMB, Thomson NR: **Whole-genome sequencing of bacterial sexually transmitted infections: implications for clinicians**. *Current Opinion in Infectious Diseases* 2013, **26**:90-98.

129. Nunes A, Borrego MJ, Gomes JP: **Genomic features beyond *Chlamydia trachomatis* phenotypes: What do we think we know?** *Infection Genetics and Evolution* 2013, **16**:392-400.

130. Gomes JP, Nunes A, Bruno WJ, Borrego MJ, Florindo C, Dean D: **Polymorphisms in the nine polymorphic membrane proteins of *Chlamydia trachomatis* across all serovars: Evidence for serovar Da recombination and correlation with tissue tropism**. *Journal of Bacteriology* 2006, **188**:275-286.

131. Voigt A, Schöfl G, Saluz HP: **The *Chlamydia psittaci* genome: a comparative analysis of intracellular pathogens**. *PLoS One* 2012, **7**:e35097.

132. Su H, Watkins NG, Zhang YX, Caldwell HD: ***Chlamydia trachomatis*** **host-cell interactions - role of the chlamydial major outer-membrane protein as an adhesin**. *Infection and Immunity* 1990, **58**:1017-1025.

133. Bavoil P, Ohlin A, Schachter J: **Role of disulfide bonding in outer-membrane structure and permeability in *Chlamydia trachomatis***. *Infection and Immunity* 1984, **44**:479-485.

134. Wyllie S, Ashley RH, Longbottom D, Herring AJ: **The major outer membrane protein of *Chlamydia psittaci* functions as a porin-like ion channel**. *Infection and Immunity* 1998, **66**:5202-5207.

135. Rockey DD, Lenart J, Stephens RS: **Genome sequencing and our understanding of chlamydiae**. *Infection and Immunity* 2000, **68**:5473-5479.

136. Henderson IR, Lam AC: **Polymorphic proteins of *Chlamydia spp*. - autotransporters beyond the Proteobacteria**. *Trends in Microbiology* 2001, **9**:573-578.

137. Swanson KA, Taylor LD, Frank SD, Sturdevant GL, Fischer ER, Carlson JH, Whitmire WM, Caldwell HD: ***Chlamydia trachomatis*** **Polymorphic Membrane Protein D Is an Oligomeric Autotransporter with a Higher-Order Structure**. *Infection and Immunity* 2009, **77**:508-516.

138. Vandahl BB, Birkelund S, Demol H, Hoorelbeke B, Christiansen G, Vandekerckhove J, Gevaert K: **Proteome analysis of the *Chlamydia pneumoniae* elementary body**. *Electrophoresis* 2001, **22**:1204-1223.

139. Grimwood J, Stephens RS: **Computational analysis of the polymorphic membrane protein superfamily of *Chlamydia trachomatis* and *Chlamydia pneumoniae***. *Microbial & comparative genomics* 1999, **4**:187-201.

140. Becker E, Hegemann JH: **All subtypes of the Pmp adhesin family are implicated in chlamydial virulence and show species-specific function**. *Microbiologyopen* 2014, **3**:544-556.

141. Grotenbreg GM, Roan NR, Guillen E, Meijers R, Wang JH, Bell GW, Starnbach MN, Ploegh HL: **Discovery of CD8+ T cell epitopes in *Chlamydia trachomatis* infection through use of caged class I MHC tetramers**. *Proc Natl Acad Sci U S A* 2008, **105**:3831-3836.

142. Almeida F, Borges V, Ferreira R, Borrego MJ, Gomes JP, Mota LJ: **Polymorphisms in Inc Proteins and Differential Expression of inc Genes among *Chlamydia trachomatis* Strains Correlate with Invasiveness and Tropism of Lymphogranuloma Venereum Isolates**. *Journal of Bacteriology* 2012, **194**:6574-6585.

143. Lutter EI, Martens C, Hackstadt T: **Evolution and conservation of predicted inclusion membrane proteins in chlamydiae**. *Comparative and functional genomics* 2012, **2012**:362104-362104.

144. Dehoux P, Flores R, Dauga C, Zhong G, Subtil A: **Multi-genome identification and characterization of chlamydiae-specific type III secretion substrates: the Inc proteins**. *Bmc Genomics* 2011, **12**.

145. Taylor LD, Nelson DE, Dorward DW, Whitmire WM, Caldwell HD: **Biological Characterization of *Chlamydia trachomatis* Plasticity Zone MACPF Domain Family Protein CT153**. *Infection and Immunity* 2010, **78**:2691-2699.

146. Nelson DE, Crane DD, Taylor LD, Dorward DW, Goheen MM, Caldwell HD: **Inhibition of chlamydiae by primary alcohols correlates with the strain-specific complement of plasticity zone phospholipase D genes**. *Infection and Immunity* 2006, **74**:73-80.

147. Belland RJ, Scidmore MA, Crane DD, Hogan DM, Whitmire W, McClarty G, Caldwell HD: ***Chlamydia trachomatis* cytotoxicity associated with complete and partial cytotoxin genes**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:13984-13989.

148. Carlson JH, Hughes S, Hogan D, Cieplak G, Sturdevant DE, McClarty G, Caldwell HD, Belland RJ: **Polymorphisms in the *Chlamydia trachomatis* cytotoxin locus associated with ocular and genital isolates**. *Infection and Immunity* 2004, **72**:7063-7072.

149. McClarty G, Caldwell HD, Nelson DE: **Chlamydial interferon gamma immune evasion influences infection tropism**. *Current Opinion in Microbiology* 2007, **10**:47-51.

150. Fehlner-Gardiner C, Roshick C, Carlson JH, Hughes S, Belland RJ, Caldwell HD, McClarty G: **Molecular basis defining human *Chlamydia trachomatis* tissue tropism. A possible role for tryptophan synthase**. *J Biol Chem* 2002, **277**:26893-26903.

151. Taylor MW, Feng GS: **Relationship between interferon-gamma, indoleamine 2,3-dioxygenase, and tryptophan catabolism**. *Faseb Journal* 1991, **5**:2516-2522.

152. Caldwell HD, Wood H, Crane D, Bailey R, Jones RB, Mabey D, Maclean I, Mohammed Z, Peeling R, Roshick C, et al.: **Polymorphisms in *Chlamydia trachomatis* tryptophan synthase genes differentiate between genital and ocular isolates**. *Journal of Clinical Investigation* 2003, **111**:1757-1769.

153. Hatch T, Miceli M, Sublett J: **Synthesis of disulfide-bonded outer membrane proteins during the developmental cycle of *Chlamydia trachomatis***. *Journal of Bacteriology* 1986, **165**:379-385.

154. Everett KD, Hatch, TP: **Architecture of the cell envelope of *Chlamydia psittaci* 6BC**. *Journal of Bacteriology* 1995, **177**:877-882.

155. Wang Y, Kahane S, Cutcliffe LT, Skilton RJ, Lambden PR, Clarke IN: **Development of a Transformation System for *Chlamydia trachomatis*: Restoration of Glycogen Biosynthesis by Acquisition of a Plasmid Shuttle Vector**. *Plos Pathogens* 2011, **7**.

156. Kari L, Goheen MM, Randall LB, Taylor LD, Carlson JH, Whitmire WM, Virok D, Rajaram K, Endresz V, McClarty G, et al.: **Generation of targeted *Chlamydia trachomatis* null mutants**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:7189-7193.

157. Nguyen BD, Valdivia RH: **Virulence determinants in the obligate intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic approaches**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:1263-1268.

158. Agaisse H, Derré I: **Expression of the effector protein IncD in *Chlamydia trachomatis* mediates recruitment of the lipid transfer protein CERT and the endoplasmic reticulum-resident protein VAPB to the inclusion membrane**. *Infect Immun* 2014, **82**:2037-2047.

159. Wickstrum J, Sammons LR, Restivo KN, Hefty PS: **Conditional gene expression in *Chlamydia trachomatis* using the tet system**. *PLoS One* 2013, **8**:e76743.

160. Wang Y, Kahane S, Cutcliffe LT, Skilton RJ, Lambden PR, Persson K, Bjartling C, Clarke IN: **Genetic transformation of a clinical (genital tract), plasmid-free isolate of *Chlamydia trachomatis*: engineering the plasmid as a cloning vector**. *PLoS One* 2013, **8**:e59195.

161. Johnson CM, Fisher DJ: **Site-specific, insertional inactivation of *incA* in *Chlamydia trachomatis* using a group II intron**. *PLoS One* 2013, **8**:e83989.

162. Song L, Carlson JH, Whitmire WM, Kari L, Virtaneva K, Sturdevant DE, Watkins H, Zhou B, Sturdevant GL, Porcella SF, et al.: ***Chlamydia trachomatis* Plasmid-Encoded Pgp4 Is a Transcriptional Regulator of Virulence-Associated Genes**. *Infection and Immunity* 2013, **81**:636-644.

163. Wang Y, Skilton RJ, Cutcliffe LT, Andrews E, Clarke IN, Marsh P: **Evaluation of a High Resolution Genotyping Method for *Chlamydia trachomatis* Using Routine Clinical Samples**. *Plos One* 2011, **6**.

164. Brunelle BW, Sensabaugh GF: **Nucleotide and phylogenetic analyses of the *Chlamydia trachomatis* ompA gene indicates it is a hotspot for mutation**. *BMC research notes* 2012, **5**:53-53.

165. Dean D, Stephens RS: **Identification of individual genotypes of *Chlamydia trachomatis* from experimentally mixed serovars and mixed infections among trachoma patients**. *Journal of Clinical Microbiology* 1994, **32**:1506-1510.

166. Gomes JP, Bruno WJ, Borrego MJ, Dean D: **Recombination in the genome of *Chlamydia trachomatis* involving the polymorphic membrane protein C gene relative to *ompA* and evidence for horizontal gene transfer**. *Journal of Bacteriology* 2004, **186**:4295-4306.

167. Gomes JP, Bruno WJ, Nunes A, Santos N, Florindo C, Borrego MJ, Dean D: **Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots**. *Genome Research* 2007, **17**:50-60.

168. Hayes LJ, Yearsley P, Treharne JD, Ballard RA, Fehler GH, Ward ME: **Evidence for naturally-occurring recombination in the gene encoding the major outer-membrane protein of lymphogranuloma-venereum isolates of *Chlamydia trachomatis***. *Infection and Immunity* 1994, **62**:5659-5663.

169. Jeffrey BM, Suchland RJ, Quinn KL, Davidson JR, Stamm WE, Rockey DD: **Genome Sequencing of Recent Clinical *Chlamydia trachomatis* Strains Identifies Loci Associated with Tissue Tropism and Regions of Apparent Recombination**. *Infection and Immunity* 2010, **78**:2544-2553.

170. Millman KL, Tavare S, Dean D: **Recombination in the *ompA* gene but not the *omcB* gene of *Chlamydia* contributes to Serovar-specific differences in tissue tropism, immune surveillance, and persistence of the organism**. *Journal of Bacteriology* 2001, **183**:5997-6008.

171. Somboonna N, Wan R, Ojcius DM, Pettengill MA, Joseph SJ, Chang A, Hsu R, Read TD, Dean D: **Hypervirulent *Chlamydia trachomatis* Clinical Strain Is a Recombinant between Lymphogranuloma Venereum (L-2) and D Lineages**. *Mbio* 2011, **2**.

172. Vos M, Didelot X: **A comparison of homologous recombination rates in bacteria and archaea**. *Isme Journal* 2009, **3**:199-208.

173. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD: **Interplay of recombination and selection in the genomes of *Chlamydia trachomatis***. *Biology Direct* 2011, **6**.

174. Comanducci M, Ricci S, Cevenini R, Ratti G: **Diversity of the *Chlamydia trachomatis* common plasmid in biovars with different pathogenicity**. *Plasmid* 1990, **23**:149-154.

175. Rockey DD: **Unraveling the basic biology and clinical significance of the chlamydial plasmid**. *Journal of Experimental Medicine* 2011, **208**:2159-2162.

176. Peterson EM, Markoff BA, Schachter J, Delamaza LM: T**he 7.5-kb plasmid present in *Chlamydia trachomatis* is not essential for the growth of this microorganism**. *Plasmid* 1990, **23**:144-148.

177. Matsumoto A, Izutsu H, Miyashita N, Ohuchi M: **Plaque formation by and plaque cloning of *Chlamydia trachomatis* biovar trachoma**. *Journal of Clinical Microbiology* 1998, **36**:3013-3019.

178. O'Connell CM, Nicks KM: **A plasmid-cured *Chlamydia muridarum* strain displays altered plaque morphology and reduced infectivity in cell culture**. *Microbiology-Sgm* 2006, **152**:1601-1607.

179. Russell M, Darville T, Chandra-Kuntal K, Smith B, Andrews CW, Jr., O'Connell CM: **Infectivity Acts as *In Vivo* Selection for Maintenance of the Chlamydial Cryptic Plasmid**. *Infection and Immunity* 2011, **79**:98-107.

180. Carlson JH, Whitmire WM, Crane DD, Wicke L, Virtaneva K, Sturdevant DE, Kupko JJ, III, Porcella SF, Martinez-Orengo N, Heinzen RA, et al.: **The *Chlamydia trachomatis* plasmid is a transcriptional regulator of chromosomal genes and a virulence factor**. *Infection and Immunity* 2008, **76**:2273-2283.

181. Farencena A, Comanducci M, Donati M, Ratti G, Cevenini R: **Characterization of a new isolate of *Chlamydia trachomatis* which lacks the common plasmid and has properties of biovar trachoma**. *Infection and Immunity* 1997, **65**:2965-2969.

182. An Q, Radcliffe G, Vassallo R, Buxton D, O'Brien WJ, Pelletier DA, Weisburg WG, Klinger JD, Olive DM: **Infection with a plasmid-free variant *Chlamydia* related to *Chlamydia trachomatis* identified by using multiple assays for nucleic acid detection**. *J Clin Microbiol* 1992, **30**:2814-2821.

183. Thomas NS, Lusher M, Storey CC, Clarke IN: **Plasmid diversity in *Chlamydia***. *Microbiology-Uk* 1997, **143**:1847-1854.

184. Ricci S, Ratti G, Scarlato V: **Transcriptional regulation in the *Chlamydia trachomatis* pCT plasmid**. *Gene* 1995, **154**:93-98.

185. Li Z, Chen D, Zhong Y, Wang S, Zhong G: **The chlamydial plasmid-encoded protein Pgp3 is secreted into the cytosol of *Chlamydia*-infected cells**. *Infection and Immunity* 2008, **76**:3415-3428.

186. Albrecht M, Sharma CM, Reinhardt R, Vogel J, Rudel T: **Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome**. *Nucleic Acids Research* 2010, **38**:868-877.

187. Comanducci M, Cevenini R, Moroni A, Giuliani MM, Ricci S, Scarlato V, Ratti G: **Expression of a plasmid gene of *Chlamydia trachomatis* encoding a novel 28-kDa antigen**. *Journal of General Microbiology* 1993, **139**:1083-1092.

188. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification**. *Annu Rev Microbiol* 2001, **55**:709-742.

189. Seth-Smith HMB, Harris SR, Persson K, Marsh P, Barron A, Bignell A, Bjartling C, Clark L, Cutcliffe LT, Lambden PR, et al.: **Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain**. *Bmc Genomics* 2009, **10**.

190. Ricci S, Cevenini R, Cosco E, Comanducci M, Ratti G, Scarlato V: **Transcriptional analysis of the *Chlamydia trachomatis* plasmid pCT identifies temporally regulated transcripts, antisense rna and sigma-70-selected promoters**. *Molecular & General Genetics* 1993, **237**:318-326.

191. Pickett MA, Everson JS, Pead PJ, Clarke IN: **The plasmids of *Chlamydia trachomatis* and *Chlamydophila pneumoniae* (N16): accurate determination of copy number and the paradoxical effect of plasmid-curing agents**. *Microbiology-Sgm* 2005, **151**:893-903.

192. Donati M, Sambri V, Comanducci M, Di Leo K, Storni E, Giacani L, Ratti G, Cevenini R: **DNA immunization with pgp3 gene of *Chlamydia trachomatis* inhibits the spread of chlamydial infection from the lower to the upper genital tract in C3H/HeN mice**. *Vaccine* 2003, **21**:1089-1093.

193. Li Z, Wang S, Wu Y, Zhong G, Chen D: **Immunization with chlamydial plasmid protein pORF5 DNA vaccine induces protective immunity against genital chlamydial infection in mice**. *Sci China C Life Sci* 2008, **51**:973-980.

194. Hiraga S: **Chromosome and plasmid partition in *Escherichia coli***. *Annual Review of Biochemistry* 1992, **61**:283-306.

195. Williams DR, Thomas CM: **Active partitioning of bacterial plasmids**. *Journal of General Microbiology* 1992, **138**:1-16.

196. Palmer L, Falkow S: **A common plasmid of *Chlamydia trachomatis***. *Plasmid* 1986, **16**:52-62.

197. Tam JE, Davis CH, Thresher RJ, Wyrick PB: L**ocation of the origin of replication for the 7.5-kb *Chlamydia trachomatis* plasmid**. *Plasmid* 1992, **27**:231-236.

198. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation**. *Nature* 2000, **405**:299-304.

199. Spratt BG, Hanage WP, Feil EJ: **The relative contributions of recombination and point mutation to the diversification of bacterial clones**. *Current Opinion in Microbiology* 2001, **4**:602-606.

200. Achtman M: **Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens**. *Annual Review of Microbiology* 2008, **62**:53-70.

201. Smith NH, Dale J, Inwald J, Palmer S, Gordon SV, Hewinson RG, Smith JM: **The population structure of *Mycobacterium bovis* in Great Britain: Clonal expansion**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:15271-15275.

202. Supply P, Warren RM, Banuls AL, Lesjean S, van der Spuy GD, Lewis LA, Tibayrenc M, van Helden PD, Locht C: **Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area**. *Molecular Microbiology* 2003, **47**:529-538.

203. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, Suerbaum S: **Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: Estimates of clock rates, recombination size, and minimal age**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:15056-15061.

204. Feil EJ, Spratt BG: **Recombination and the population structures of bacterial pathogens**. *Annual Review of Microbiology* 2001, **55**:561-590.

205. Didelot X, Maiden MCJ: **Impact of recombination on bacterial evolution**. *Trends in Microbiology* 2010, **18**:315-322.

206. Maiden MCJ: **Multilocus sequence typing of bacteria**. *Annual Review of Microbiology* 2006, **60**:561-588.

207. Awadalla P: **The evolutionary genomics of pathogen recombination**. *Nature Reviews Genetics* 2003, **4**:50-60.

208. Cooper JE, Feil EJ: **The phylogeny of *Staphylococcus aureus* - which genes make the best intra-species markers?** *Microbiology-Sgm* 2006, **152**:1297-1305.

209. Borges V, Nunes A, Ferreira R, Borrego MJ, Gomes JP: **Directional Evolution of *Chlamydia trachomatis* towards Niche-Specific Adaptation**. *Journal of Bacteriology* 2012, **194**:6143-6153.

210. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data**. *Genetics* 2007, **175**:1251-1266.

211. Dean D, Bruno WJ, Wan R, Gomes JP, Devignot S, Mehari T, de Vries HJC, Morre SA, Myers G, Read TD, et al.: **Predicting Phenotype and Emerging Strains among *Chlamydia trachomatis* Infections**. *Emerging Infectious Diseases* 2009, **15**:1385-1394.

212. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, et al.: ***Helicobacter pylori* genome evolution during human infection**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:5033-5038.

213. Perez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA: **Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data**. *Infection Genetics and Evolution* 2006, **6**:97-112.

214. Borges V, Ferreira R, Nunes A, Nogueira P, Borrego MJ, Gomes JP: **Normalization strategies for real-time expression data in *Chlamydia trachomatis***. *Journal of Microbiological Methods* 2010, **82**:256-264.

215. Andolfatto P: **Adaptive evolution of non-coding DNA in *Drosophila***. *Nature* 2005, **437**:1149-1152.

216. Bush EC, Lahn BT: **Selective constraint on noncoding regions of hominid genomes**. *Plos Computational Biology* 2005, **1**:593-598.

217. Darling ACE, Mau B, Blattner FR, Perna NT: **Mauve: Multiple alignment of conserved genomic sequence with rearrangements**. *Genome Research* 2004, **14**:1394-1403.

218. Darling AE, Mau B, Perna NT: **progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement**. *Plos One* 2010, **5**.

219. Gelman A, Rubin DB: **Inference from iterative simulation using multiple sequences**. *Statistical Science* 1992, **7**:457-511.

220. Conway DJ, Cavanagh DR, Tanabe K, Roper C, Mikes ZS, Sakihama N, Bojang KA, Oduola AMJ, Kremsner PG, Arnot DE, et al.: **A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses**. *Nature Medicine* 2000, **6**:689-692.

221. Stumpf MPH, McVean GAT: **Estimating recombination rates from population-genetic data**. *Nature Reviews Genetics* 2003, **4**:959-968.

222. den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M: **Lineage specific recombination rates and microevolution in *Listeria monocytogenes***. *Bmc Evolutionary Biology* 2008, **8**.

223. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, et al.: **Recombination and Population Structure in *Salmonella enterica***. *Plos Genetics* 2011, **7**.

224. Whittam TS: **Genetic population structure and pathogenicity in enteric bacteria**. *Population Genetics of Bacteria* 1995, **52**:217-245.

225. Feil EJ, Maiden MCJ, Achtman M, Spratt BG: **The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis***. *Molecular Biology and Evolution* 1999, **16**:1496-1502.

226. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, et al.: **How clonal is *Staphylococcus aureus*?** *Journal of Bacteriology* 2003, **185**:3307-3316.

227. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NPJ, Enright MC, Goldstein R, Hood DW, Kalla A, Moore CE, et al.: **Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:182-187.

228. Meats E, Feil EJ, Stringer S, Cody AJ, Goldstein R, Kroll JS, Popovic TJ, Spratt BG: **Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing**. *Journal of Clinical Microbiology* 2003, **41**:1623-1636.

229. Hanage WP, Fraser C, Spratt BG: **The impact of homologous recombination on the generation of diversity in bacteria**. *Journal of Theoretical Biology* 2006, **239**:210-219.

230. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R: **Genes under positive selection in *Escherichia coli***. *Genome Research* 2007, **17**:1336-1343.

231. Vos M: **Why do bacteria engage in homologous recombination?** *Trends in Microbiology* 2009, **17**:226-232.

232. Lefebure T, Stanhope MJ: **Evolution of the core and pan-genome of *Strepococcus*: positive selection, recombination, and genome composition**. *Genome Biology* 2007, **8**.

233. Kaplan NL, Hudson RR, Langley CH: **The hitchhiking effect revisited**. *Genetics* 1989, **123**:887-899.

234. Casadevall A, Pirofski L-a: **Virulence factors and their mechanisms of action: the view from a damage-response framework**. *Journal of Water and Health* 2009, **7**:S2-S18.

235. Wilson JW, Schurr MJ, LeBlanc CL, Ramamurthy R, Buchanan KL, Nickerson CA: **Mechanisms of bacterial pathogenicity**. *Postgraduate Medical Journal* 2002, **78**:216-224.

236. Garciadelportillo F, Finlay BB: **The varied life-styles of intracellular pathogens within eukaryotic vacuolar compartments**. *Trends in Microbiology* 1995, **3**:373-380.

237. PizarroCerda J, Moreno E, Desjardins M, Gorvel JP: **When intracellular pathogens invade the frontiers of cell biology and immunology**. *Histology and Histopathology* 1997, **12**:1027-1038.

238. Ernst RK, Guina T, Miller SI: **How intracellular bacteria survive: Surface modifications that promote resistance to host innate immune responses**. *Journal of Infectious Diseases* 1999, **179**:S326-S330.

239. Cossart P: **Illuminating the landscape of host-pathogen interactions with the bacterium *Listeria monocytogenes***. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:19484-19491.

240. Allerberger F, Wagner M: **Listeriosis: a resurgent foodborne infection**. *Clinical Microbiology and Infection* 2010, **16**:16-23.

241. Hava D, Camilli A: **Large-scale identificationof serotype 4 *Streptococcus pneumoniae* virulence factors**. *Molecular Microbiology* 2002, **45**:1389-1405.

242. Burton MJ: **Trachoma: an overview**. *British Medical Bulletin* 2007, **84**:99-116.

243. Wright HR, Turner A, Taylor HR: **Trachoma**. *Lancet* 2008, **371**:1945-1954.

244. Peipert JF: **Genital chlamydial infections**. *New England Journal of Medicine* 2003, **349**:2424-2430.

245. Schachter J: **Chlamydial infections .1**. *New England Journal of Medicine* 1978, **298**:428-435.

246. Mishra MK, Gerard HC, Whittum-Hudson JA, Hudson AP, Kannan RM: **Dendrimer-Enabled Modulation of Gene Expression in *Chlamydia trachomatis***. *Molecular Pharmaceutics* 2012, **9**:413-421.

247. Read TD, Joseph SJ, Didelot X, Liang B, Patel L, Dean D: **Comparative analysis of *Chlamydia psittaci* genomes reveals the recent emergence of a pathogenic lineage with a broad host range**. *mBio* 2013, **4**.

248. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**. *Molecular Biology and Evolution* 2011, **28**:2731-2739.

249. Nei M, Kumar S: *Molecular Evolution and Phylogenetics*. New York: Oxford University Press 2000.

250. Rodgers JL, Nicewander WA: **13 ways to look at the correlation-coefficient**. *American Statistician* 1988, **42**:59-66.

251. Felsenstein J: **Confidence-limits on phylogenies - an approach using the bootstrap**. *Evolution* 1985, **39**:783-791.

252. Saitou N, Nei M: **The neighbor-joining method - a new method for reconstructing phylogenetic trees**. *Molecular Biology and Evolution* 1987, **4**:406-425.

253. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences**. *Journal of Molecular Evolution* 1980, **16**:111-120.

254. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data**. *Bioinformatics* 2009, **25**:1451-1452.

255. Salminen MO, Carr JK, Burke DS, McCutchan FE: **Identification of breakpoints in intergenotypic recombinants of hiv type-1 by bootscanning**. *Aids Research and Human Retroviruses* 1995, **11**:1423-1425.

256. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC: **Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination**. *Journal of Virology* 1999, **73**:152-160.

257. Robertson DL, Hahn BH, Sharp PM: **Recombination in aids viruses**. *Journal of Molecular Evolution* 1995, **40**:249-259.

258. da Cunha M, Milho C, Almeida F, Pais SV, Borges V, Mauricio R, Borrego MJ, Gomes JP, Mota LJ: **Identification of type III secretion substrates of *Chlamydia trachomatis* using *Yersinia enterocolitica* as a heterologous system**. *Bmc Microbiology* 2014, **14**.

259. Giles TN, Fisher DJ, Graham DE: **Independent inactivation of arginine decarboxylase genes by nonsense and missense mutations led to pseudogene formation in *Chlamydia trachomatis* serovar L2 and D strains**. *Bmc Evolutionary Biology* 2009, **9**.

260. Mehlitz A, Banhart S, Maeurer AP, Kaushansky A, Gordus AG, Zielecki J, MacBeath G, Meyer TF: **Tarp regulates early *Chlamydia*-induced host cell survival through interactions with the human adaptor protein SHC1**. *Journal of Cell Biology* 2010, **190**:143-157.

261. Carabeo R: **Bacterial subversion of host actin dynamics at the plasma membrane**. *Cellular Microbiology* 2011, **13**:1460-1469.

262. Ferreira R, Borges V, Nunes A, Nogueira PJ, Borrego MJ, Gomes JP: **Impact of Loci Nature on Estimating Recombination and Mutation Rates in *Chlamydia trachomatis***. *G3-Genes Genomes Genetics* 2012, **2**:761-768.

263. Derre I, Swiss R, Agaisse H: **The Lipid Transfer Protein CERT Interacts with the *Chlamydia* Inclusion Protein IncD and Participates to ER-*Chlamydia* Inclusion Membrane Contact Sites**. *Plos Pathogens* 2011, **7**.

264. Christerson L, Bom RJM, Bruisten SM, Yass R, Hardick J, Bratt G, Gaydos CA, Morre SA, Herrmann B: ***Chlamydia trachomatis* Strains Show Specific Clustering for Men Who Have Sex with Men Compared to Heterosexual Populations in Sweden, the Netherlands, and the United States**. *Journal of Clinical Microbiology* 2012, **50**:3548-3555.

265. Muschiol S, Boncompain G, Vromman F, Dehoux P, Normark S, Henriques-Normark B, Subtil A: **Identification of a Family of Effectors Secreted by the Type III Secretion System That Are Conserved in Pathogenic Chlamydiae**. *Infection and Immunity* 2011, **79**:571-580.

266. Galtier N, Daubin V: **Dealing with incongruence in phylogenomic analyses**. *Philosophical Transactions of the Royal Society B-Biological Sciences* 2008, **363**:4023-4029.

267. Pizarro-Cerda J, Kuhbacher A, Cossart P: **Entry of *Listeria monocytogenes* in mammalian epithelial cells: an updated view**. *Cold Spring Harbor perspectives in medicine* 2012, **2**.

268. Kumar Y, Cocchiaro J, Valdivia RH: **The obligate intracellular pathogen *Chlamydia trachomatis* targets host lipid droplets**. *Current Biology* 2006, **16**:1646-1651.

269. Petersen J: **Phylogeny and compatibility: plasmid classification in the genomics era**. *Archives of Microbiology* 2011, **193**:313-321.

270. del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, Diaz-Orejas R: **Replication and control of circular bacterial plasmids**. *Microbiology and Molecular Biology Reviews* 1998, **62**:434-+.

271. Maloy SR, Cronan JE, Freifelder D: **Plasmids**. In *Microbial Genetics*. Edited by Campbell JH, Schopf JW: Jones and Bartlett Publishers; 1994:213-238.

272. Nordstrom K, Dasgupta S: **Copy-number control of the *Escherichia coli* chromosome: a plasmidologist's view**. *Embo Reports* 2006, **7**:484-489.

273. Nordstrom K, Austin SJ: **Mechanisms that contribute to the stable segregation of plasmids**. *Annual Review of Genetics* 1989, **23**:37-69.

274. Gerdes K, Moller-Jensen J, Jensen RB: **Plasmid and chromosome partitioning: surprises from phylogeny**. *Molecular Microbiology* 2000, **37**:455-466.

275. Hayes F: **The partition system of multidrug resistance plasmid TP228 includes a novel protein that epitomizes an evolutionarily distinct subgroup of the ParA superfamily**. *Molecular Microbiology* 2000, **37**:528-541.

276. Moller-Jensen J, Jensen RB, Gerdes H: **Plasmid and chromosome segregation in prokaryotes**. *Trends in Microbiology* 2000, **8**:313-320.

277. Sengupta M, Austin S: **Prevalence and Significance of Plasmid Maintenance Functions in the Virulence Plasmids of Pathogenic Bacteria**. *Infection and Immunity* 2011, **79**:2502-2509.

278. Bennett PM: **Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria**. *British Journal of Pharmacology* 2008, **153**:S347-S357.

279. Stothard DR, Williams JA, Van Der Pol B, Jones RB: **Identification of a *Chlamydia trachomatis* serovar E urogenital isolate which lacks the cryptic plasmid**. *Infection and Immunity* 1998, **66**:6010-6013.

280. Kari L, Whitmire WM, Olivares-Zavaleta N, Goheen MM, Taylor LD, Carlson JH, Sturdevant GL, Lu C, Bakios LE, Randall LB, et al.: **A live-attenuated chlamydial vaccine protects against trachoma in nonhuman primates**. *Journal of Experimental Medicine* 2011, **208**:2217-2223.

281. O'Connell CM, AbdelRahman YM, Green E, Darville HK, Saira K, Smith B, Darville T, Scurlock AM, Meyer CR, Belland RJ: **Toll-Like Receptor 2 Activation by *Chlamydia trachomatis* Is Plasmid Dependent, and Plasmid-Responsive Chromosomal Loci Are Coordinately Regulated in Response to Glucose Limitation by *C. trachomatis* but Not by *C. muridarum***. *Infection and Immunity* 2011, **79**:1044-1056.

282. Gomes JP, Hsia RC, Mead S, Borrego MJ, Dean D: **Immunoreactivity and differential developmental expression of known and putative *Chlamydia trachomatis* membrane proteins for biologically variant serovars representing distinct disease groups**. *Microbes and Infection* 2005, **7**:410-420.

283. Nunes A, Gomes JP, Mead S, Florindo C, Correia H, Borrego MJ, Dean D: **Comparative Expression Profiling of the *Chlamydia trachomatis pmp* Gene Family for Clinical and Reference Strains**. *Plos One* 2007, **2**.

284. Comanducci M, Manetti R, Bini L, Santucci A, Pallini V, Cevenini R, Sueur JM, Orfila J, Ratti G: **Humoral immune-response to plasmid protein Pgp3 in patients with *Chlamydia trachomatis* infection**. *Infection and Immunity* 1994, **62**:5491-5497.

285. Valdivia RH: ***Chlamydia* effector proteins and new insights into chlamydial cellular microbiology**. *Current Opinion in Microbiology* 2008, **11**:53-59.

286. Chen D, Lei L, Lu C, Galaleldeen A, Hart PJ, Zhong G: **Characterization of Pgp3, a *Chlamydia trachomatis* Plasmid-Encoded Immunodominant Antigen**. *Journal of Bacteriology* 2010, **192**:6017-6024.

287. Bignell C, Thomas CM: **The bacterial ParA-ParB partitioning proteins**. *Journal of Biotechnology* 2001, **91**:1-34.

288. Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity.** *Nature Reviews Genetics* 2010, **11**:9-16.

289. Georg J, Hess W: **Cis-antisense RNA, another level of gene regulation in bacteria.** *Microbiology and Molecular Biology Reviews* 2011, **75**:286-300.

290. Thomason M, Storz G: **Bacterial antisense RNAs: how many are there, and what are they doing?** *Annual Review of Genetics* 2010, **44**:167-188.

291. Lasa I, Toledo-Arana A, Gingeras T: **An effort to make sense of antisense transcription in bacteria.** *RNA Biology* 2012, **9**:1039-1044.

292. Wang Y, Liu C, Storey J, Tibshirani R, Herschlag D, Brown P: **Precision and functional specificity in mRNA decay**. *Proceedings of the National Academy of Sciences* 2002, **99**:5860-5865.

293. Bernstein J, Khodursky A, Lin P, Lin-Chao S, Cohen S: **Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays**. *Proceedings of the National Academy of Sciences* 2002, **99**:9697-9702.

294. Hambraeus G, von Wachenfeldt C, Hederstedt L: **Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs.** *Molecular Genetics and Genomics* 2003, **269**:706-714.

295. Barnett T, Bugrysheva J, Scott J: **Role of mRNA stability in growth phase regulation of gene expression in the group A *Streptococcus*.** *Journal of Bacteriology* 2007, **189**:1866-1873.

296. Kristoffersen S, Haase C, Weil M, Passalacqua K, Niazi F, Hutchison S, Desany B, style="font-size:12.0pt slE-U, line-height:, 105%, et al.: **Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium**. *Genome Biology* 2012, **13**:R30.

297. Rustad T, Minch K, Brabant W, Winkler J, Reiss D, Baliga N, Sherman D: **Global analysis of mRNA stability in *Mycobacterium tuberculosis*** *Nucleic Acids Research* 2012, **41**:509-517.

298. Liu B, Deikus G, Bree A, Durand S, Kearns D, Bechhofer D: **Global analysis of mRNA decay intermediates in *Bacillus subtilis* wild-type and polynucleotide phosphorylase-deletion strains.** *Molecular Microbiology* 2014, **94**:41-55.

299. Chen H, Shiroguchi K, Ge H, Xie X: **Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*** *Molecular Systems Biology* 2015, **11**:781.

300. Goodwin S, McPherson J, McCombie W: **Coming of age: ten years of next-generation sequencing technologies**. *Nature Reviews Genetics* 2016, **17**:333-351.

301. Ozsolak F, Milos P: **RNA sequencing: advances, challenges and opportunities**. *Nature Reviews Genetics* 2011, **12**:87-98.

302. Creecy J, Conway T: **Quantitative bacterial transcriptomics with RNA-seq**. *Current Opinion in Micribiology* 2015, **23**:133-140.

303. Croucher N, Thomson N: **Studying bacterial transcriptomes using RNA-seq**. *Current Opinion in Microbiology* 2010, **13**:619-624.

304. Sharma C, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, et al.: **The primary transcriptome of the major human pathogen *Helicobacter pylori*** *Nature* 2010, **464**:250-255.

305. Sharma C, Vogel J: **Differential RNA-seq: the approach behind and the biological insight gained.** *Current Opinion in Microbiology* 2014, **19**:97-105.

306. Humphrys M, Creasy T, Sun Y, Shetty A, Chibucos M, Drabek E, Fraser C, Farooq U, Sengamalay N, Ott S, et al.: **Simultaneous transcriptional profiling of bacteria and their host cells.** *PLoS One* 2013, **8**:e80597.

307. Borges V, Pinheiro M, Antelo M, Sampaio D, Vieira L, Ferreira R, Nunes A, Almeida F, Mota L, Borrego M, et al.: ***Chlamydia trachomatis In Vivo* to *In Vitro* Transition Reveals Mechanisms of Phase Variation and Down-Regulation of Virulence Factors**. *PLoS One* 2015, **10**:e0133420.

308. Miyairi I, Mahdi O, Ouellette S, Belland R, Byrne G: **Different growth rates of *Chlamydia trachomatis* biovars reflect pathotype.** *The Journal of Infectious Diseases* 2006, **194**:350-357.

309. Schachter J: **Chlamydial infections (third of three parts)**. *N Engl J Med* 1978, **298**:540-549.

310. Ferreira R, Antelo M, Nunes A, Damião V, Borrego MJ, Gomes JP: ***In silico* scrutiny of genes revealing phylogenetic congruence with clinical prevalence or tropism properties of *Chlamydia trachomatis* strains**. *G3 (Bethesda)* 2015, **5**:9-19.

311. Abdelsamed H, Peters J, Byrne GI: **Genetic variation in *Chlamydia trachomatis* and their hosts: impact on disease severity and tissue tropism**. *Future Microbiol* 2013, **8**:1129-1146.

312. Ferreira R, Borges V, Nunes A, Borrego MJ, Gomes JP: **Assessment of the load and transcriptional dynamics of *Chlamydia trachomatis* plasmid according to strains' tissue tropism**. *Microbiological Research* 2013, **168**:333-339.

313. Grunberg-Manago M: **Messenger RNA stability and its role in control of gene expression in bacteria and phages**. *Annual Reviews of Genetics* 1999, **33**:193-227.

314. Steege D: **Emerging features of mRNA decay in bacteria.** *RNA* 2000, **6**:1079-1090.

315. Keene J: **The global dynamics of RNA stability orchestrates responses to cellular activation**. *BMC Biology* 2010, **8**:95.

316. Pérez-Ortín J, Alepuz P, Moreno J: **Genomics and gene transcription kinetics in yeast**. *Trends in Genetics* 2007, **23**:250-257.

317. Hartmann G, Honikel K, Knusel F, Nuesch J: **The specific inhibition of the DNA-directed RNA synthesis by rifamycin**. *Biochim Biophys Acta* 1967, **145**:843-844.

318. Levin M, Hatfull G: ***Mycobacterium smegmatis* RNA polymerase: DNA supercoiling, action of rifampicin and mechanism of rifampicin resistance**. *Molecular Microbiology* 1993, **8**:277-285.

319. Selinger D, Saxena R, Cheung K, GM C, Rosenow C: **Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation**. *Genome Research* 2003, **13**:216-223.

320. Heizer EJ, Raiford D, Raymer M, Doom T, Miller R, Krane D: **Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis**. *Molecular Biology and Evolution* 2006, **23**:1670-1680.

321. Subtil A, Delevoye C, Balañá ME, Tastevin L, Perrinet S, Dautry-Varsat A: **A directed screen for chlamydial proteins secreted by a type III mechanism identifies a translocated protein and numerous other new candidates**. *Mol Microbiol* 2005, **56**:1636-1647.

322. Andersson A, Lundgren M, Eriksson S, Rosenlund M, Bernander R, Nilsson P: **Global analysis of mRNA stability in the archaeon *Sulfolobus***. *Genome Biology* 2006, **7**:R99.

323. Commichau F, Rothe F, Herzberg C, Wagner E, Hellwig D, Lehnik-Habrink M, Hammer E, Volker U, Stulke J: **Novel activities of glycolytic enzymes in *Bacillus subtilis*: interactions with essential proteins involved in mRNA processing.** *Molecular & Cellular Proteomics* 2009, **8**:1350-1360.

324. Engström P, Bailey L, Onskog T, Bergström S, Johansson J: **A comparative study of RNA and DNA as internal gene expression controls early in the developmental cycle of *Chlamydia pneumoniae*.** *FEMS Immunology & Medical Microbiology* 2010, **58**:244-253.

325. Shaw AC, Gevaert K, Demol H, Hoorelbeke B, Vandekerckhove J, Larsen MR, Roepstorff P, Holm A, Christiansen G, Birkelund S: **Comparative proteome analysis of *Chlamydia trachomatis* serovar A, D and L2**. *Proteomics* 2002, **2**:164-186.

326. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE: **Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes**. *Genome Res* 2003, **13**:1863-1872.

327. Waters L, Storz G: **Regulatory RNAs in Bacteria**. *Cell* 2009, **136**:615-628.

328. Borges V, Pinheiro M, Vieira L, Sampaio DA, Nunes A, Borrego MJ, Gomes JP: **Complete Genome Sequence of *Chlamydia trachomatis* Ocular Serovar C Strain TW-3**. *Genome Announc* 2014, **2**.

329. Unemo M, Seth-Smith HM, Cutcliffe LT, Skilton RJ, Barlow D, Goulding D, Persson K, Harris SR, Kelly A, Bjartling C, et al.: **The Swedish new variant of *Chlamydia trachomatis*: genome sequence, morphology, cell tropism and phenotypic characterization**. *Microbiology* 2010, **156**:1394-1404.

Supplemental Material

Supplemental Table 2.1. Oligonucleotide primers used for PCR and sequencing.

ORF[a]	Primer[a]	Primer sequence (5' to 3')[a]	Primer location[a]	Amplicon size (bp)[a]
CT048 (*yraL*)	CT048-1[b]	GAGCCGGCTCTTTTAAATGGTTT	53104 - 53126	1059
	CT048-2[b]	GTCGACGGAACAGACGAAGAAA	54141 - 54162	
CT050/CT051	CT050/51-1A[c, d]	TGGGCGCTGGTTATTAACTATTTG	377806 - 377829	3704
	CT050/51-2A[b, d]	GACCCCATCCCCTTTGGAGT	381490 - 381509	
	CT050/51-1D[b]	ACAAAGCGCTTTCAGAACATACAT	55452 - 55475	3589
	CT050/51-2D[b]	AGGGCGTCTTTTTCATGATTCTAT	59017 - 59040	
	CT050/51-1[e]	CTAAGAGTTATGTAGCTATC	55542 - 55561	
	CT050/51-1S[e]	AGTTAAGGGAGAGAATCTC	55593 - 55611	
	CT050/51-3[e]	TTGTAGTGTGCAAGATTGTC	55857 - 55876	
	CT050/51-4[e]	TTGTGCCACTACAATACCTT	58726 - 58745	
	CT050/51-5[e]	AACCTTTCCAATATCACCGT	56566 - 56585	
	CT050/51-6[e]	GCACAGATCGCCAATATCAA	58076 - 58095	
	CT050/51-7[e]	AGTCACTCCAGACAATTCTA	57699 - 57718	
	CT050/51-8[e]	TTAGTGAGACAGGCATTGA	57160 - 57178	
CT058/CT059	CT058/9-1[b]	AGTACCGGCCGAATCTCTTTCTCC	67315 - 67338	1634
(*fer*)	CT058/9-2[b]	GTCGGGGGTTCGAATCCCTCTA	68927 - 68948	
	CT058/9-3[e]	TCTCATTACTTCTCTTGCGT	68432 - 68451	
CT147[f]	CT147-1[b]	GGGGAAAGTGAGCTCTTCGGTATC	165430 - 165453	2819
	CT147-2[b]	CGCCGCTACAACAGCTTTAGTGA	168226 - 168248	
	CT147-3[b]	ATTGCGTCCCAAGATATACGACAG	167928 - 167951	2294
	CT147-4[b]	CACGCCAACCCAGAATCCTT	170202 - 170221	
	CT147-5[e]	TAATCATCCACTAGAAGCG	166229 - 166247	
	CT147-6[e]	TGTTCTAGCTGCTCTTGAAT	167649 - 167668	
	CT147-7[e]	TGGATGGTGTTGCAGAATTA	168511 - 168530	
	CT147-8[e]	TACCTCTAGATGTTTTGCGT	169786 - 169805	
CT192	CT192-1[b]	ATATGCGCAAGCACACCTTCC	215717 - 215737	1016
	CT192-2[b]	CTGGGCGTCCATTCACAACA	216713 - 216732	
	CT192-3[e]	CGTATCGATTCCTTCTTCTA	215998 - 216017	
	CT192-4[e]	CTCCTCTTATTGAAGAAGCT	2156196 - 216215	
CT195	CT195-1[b]	CCTCCGCCTAATCCTCGACTACAT	219655 - 216678	1320
	CT195-2[b]	CCAGCGGTTGATATTTCTTGATTA	220951 - 220974	
CT214	CT214-1[b]	AGGGCTTCTATTCCTCAAACAGTA	241463 - 241486	1797
	CT214-2[b]	TTCCCCGTTCTAAAGATCAGTTAT	243236 - 243259	
	CT214-3[e]	AACAGCCTGGATCTATATCA	242027 - 242046	
CT223	CT223-1[b]	GCAACGCATATCGCTCCTCA	251102 - 251121	1308
	CT223-2[b]	GTGCGCCCCTTCTCGTAAAG	252390 - 252409	
CT228/CT229	CT228/9-1[b]	CGGTCCCGGATTATCAAAACAAGT	254667 - 254690	1816
	CT228/9-2[b]	ATGCGGCCATCCCAGAAGC	256464 - 256482	
	CT228/9-3[e]	AGATTACGCAAACGTTGCTC	255033 - 255052	
	CT228/9-4[e]	GTTGTGATTGCAGCAGTAG	255972 - 255990	
CT232/CT233	CT232/3-1[b]	GATTAGGCGGAGGGGTTCTCTT	259223 - 259244	1262
(*incB*/*incC*)	CT232/3-2[b]	CTCTCCGCGACGCAAACTAAG	260464 - 260484	
CT249	CT249-1[b]	ACCACCCTTTAGCCATCCATTCC	279170 - 279192	704
	CT249-2[b]	AATTGCGCCGCCTCCTTGTA	279854 - 279873	
CT288	CT288-1[b]	TTTTACGCACAATGAACCCAGAAA	321582 - 321605	2063
	CT288-2[b]	CGGGCTCCTCGGGAACAG	323627 - 323644	
	CT288-3[e]	TTACCTGACCTCAGACACC	322164 - 322182	

	CT288-4[e]	GTCAGCTCGTCGTTTATTG	323121 - 323139	
CT293 (*accD*)	CT293-1[b]	TGCGCCAGAAGCTCCAGAAGTAGC	326322 - 326345	1350
	CT293-2[b]	AGGATCTGGCTGGGGATGGTTAGC	327648 - 327671	
CT365	CT365-1[b]	AAATTCGCAAACTTGCTCTTTTTC	416102 - 416125	2213
	CT365-2[b]	GATCGGGATTCCCCTGGATA	418295 - 418314	
	CT365-3[e]	CTAACTTCCAAGTTTCCTCT	416572 - 416591	
	CT365-4[e]	CTCATTGCAGGTATTGTTGT	417720 - 417739	
CT442 (*crpA*)	CT442-1[b]	CTCCTCCCTTTCCATACATCATCT	511256 - 511279	783
	CT442-2[b]	AAGCGATTCTTTCTTCCGATACAT	512015 - 512038	
CT456 (*tarp*)	CT456-1[b]	ACAAACGTTACCCGGTATGCTGTT	530723 - 530746	3362
	CT456-2[b]	TTGCGCCTTGTCGATTGTGAT	534064 - 534084	
	CT456-3[e]	TACCTCATCAAGCGATCATA	531252 - 531271	
	CT456-4[e]	CCACCAGTTGTTATTATGTC	533470 - 533489	
	CT456-5[e]	AGACATGTCTCTTCCTTCAT	531867 - 531886	
	CT456-6[e]	TACATCAGAGATTACGTCTC	532889 - 532908	
	CT456-7[e]	GAGTTTCATTGGAGAAGGAA	532413 - 532432	
	CT456-8[e]	CGTTACCCGGTATGCTGTT	530728 - 530746	
	CT456-9[e]	TACAAACACTACTGCCTTCA	533363 - 533382	
	CT456-10[e]	TTGTTACTACCTACGTCATC	531328 - 531347	
	CT456-11[e]	CTAATTAAATCGGCTGTTG	530869 - 530887	
CT529	CT529-1[b]	ACGCGGCTCCTTAAAGCAAACAA	596464 - 596486	1659
	CT529-2[b]	CGCGCATATCCGGGGAGTCT	598103 - 598122	
	CT529-3[e]	TCTCGCAAGCATTTTCCTCT	596984 - 597003	
CT618	CT618-1[b]	TCCCGATATGCCTCCTTTTGAGTC	698080 - 698103	1360
	CT618-2[b]	ATGCGCACGCAAGCCAATC	699421 - 699439	
CT622	CT622-1[b]	GGCTCCCCCTCAATTCACAAACTT	707046 - 707069	2319
	CT622-2[b]	GGTCGCGGAAACCAAATGAAATA	709342 - 709364	
	CT622-3[e]	TGATTGCTTGTATTTCGGCT	707784 - 707803	
	CT622-4[e]	TTCAGCATCGTCCTCTGTAA	708885 - 708904	
	CT622-5[e]	AGAAGAGATTATGCAGAAGC	707298 - 707317	
CT623	CT623-1[b]	TTTGCCCATTAATAATTGGATTCA	708957 - 708980	1516
	CT623-2[b]	CATGGGTCGGTTGTATGAGATGT	710450 - 710472	
CT653 (*yhbG*)	CT653-1[b]	AGCCGCGATAGCTAACGAAGTG	750194 - 750215	922
	CT653-2[b]	GAAGGCGGAATGAAAGTCCTCTC	751093 - 751115	
CT674 (*yscC*)	CT674-1[b]	TTTCAAGCGGAATCGCAAGGAAT	770234 - 770256	3229
	CT674-2[b]	CCGGGATCGAACCGACGAC	773444 - 773462	
	CT674-3[e]	AGAGCCATCAGATTTTCTCT	770846 - 770865	
	CT674-4[e]	AGAGGAAGAGAACTGAGTAA	772996 - 773015	
	CT674-5[e]	ATGACTTGAAAGTCGTTGAA	771470 - 771489	
	CT674-6[e]	TCCGTACATCATATCACTGA	772505 - 775524	
	CT674-7[e]	GATATCGGAGTCAATCTTGTT	772771 - 772791	
CT675 (*karG*)/	CT675/6-1[b]	CCCGGCTTTGGGCATTCC	773458 - 773475	1933
CT676	CT675/6-2[b]	TCATTCCGGTAACAGGGGTTCG	775369 - 775390	
	CT675/6-3[e]	CGTGATCAGATTAATCAGCT	774669 - 774688	
CT677/CT678/	CT677/9-1[b]	ATGGGGCCAGGACGGGTCTA	775420 - 775439	2393
CT679	CT677/9-2[b]	AAATTTTATCTCCGGTGCGTCCTG	777789 - 777812	
(*frr, pyrH, tsf*)	CT677/9-4[e]	GAAGATATCACTGTACCAAC	776088 - 776107	
	CT677/9-5[e]	TGGCTAAAGACATTGCTATG	777107 - 777126	
CT682 (*pbpB*)[f]	CT682-1[b]	GCATTGTGATCGCGCAGGAGTA	780841 - 780862	3149
	CT682-2[b]	TTTCCGCCTCTTCCATAGTCGTTC	783966 - 783989	
	CT682-3[e]	TTGATAGGCAAGCGATCTAT	781484 - 781503	

	CT682-4[e]	ATATTCTCCAGGAAGTCCTA	783318 - 783337	
	CT682-5[e]	TGGAAATGTTTGAGTGTGAA	782079 - 782098	
CT683 (*TPR-*	CT683-1[b]	TCGCTGCGGTAGGATATGAAGATG	783756 - 783779	1702
motif protein)	CT683-2[b]	TCGCCGCGTAAATGAACCAAT	785437 - 785457	
	CT683-3[e]	GACTGTCAAACAGCCCTTAA	785108 - 785127	
CT684[f]	CT684-A1[b]	TGGGCATTACAATCTTGGGTTATG	784615 - 784638	1471
	CT684-A2[b]	AACAGCGGCATGCAGTTGATG	786065 - 786085	
	CT684-B1[b]	ATTTCGGGAGGCAATCCACAAT	785785 - 785806	1398
	CT684-B2[b]	TCCCGGGAATCCATATACCTCTTC	787159 - 787182	
	CT684-A3[e]	TCATCAATAGGGATGCCTAA	785651 - 785670	
	CT684-A4[e]	TCTGGATCTGCGTCTTCTAA	785609 - 785628	
CT685/CT686	CT685/6-1[b]	GCTCGGGAAGCAACCAAGTTATTA	786722 - 786745	2386
	CT685/6-2[b]	AAGCGAGTTCCCATGATACGAGAT	789084 - 789107	
	CT685/6-3[e]	TCTTGATGTCGATGCTTTGA	787308 - 787327	
	CT685/6-4[e]	ATGCAGAAGCTCGATTACTT	788502 - 788521	
CT694	CT694-1[b]	TACAGGGGGAGGCGCTTCCTTA	796071 - 796092	1541
	CT694-2[b]	CGCGCTCTTCTAGCTCTCCCTCTT	796528 - 797551	
	CT694-3[e]	GATAACTCTTAACCCCATTG	796267 - 796286	
CT760 (*ftsW*)	CT760-1[b]	CTTGGGCCATTGCATTGAGTAAT	892695 - 892717	1417
	CT760-2[b]	CCCCAGAGAACATCCGATTGAC	894090 - 894111	
	CT760-3[e]	AGGATGTAGGTAAACTTGCA	893445 - 893464	
CT783	CT783-1[b]	AGCGGGGATTCAGCATTCCT	919328 - 919347	1427
	CT783-2[b]	TGCCCCTCGCCTCTTCATC	920676 - 920694	
	CT783-3[e]	TGTCAATACCTTCCCTAGTT	919913 - 919932	
CT813	CT813-1[b]	CTGCGTGTCGCTCTGGAAAATAAT	954988 - 955011	1460
	CT813-2[b]	AGGCCGAGCCCTACTCAAAAACT	956365 - 956387	
CT818 (*tyrP_2*)	CT818-1[b]	CCTGGCGGGAAAGGGACTCT	961600 - 961619	1531
	CT818-2[b]	GCGCATAATCGCGATCATACAATC	963107 - 963130	
	CT818-3[e]	GAATAGCACGTTTAACCTCA	962512 - 962531	
CT852 (*yhgN*)	CT852-1[b]	CTGCCGCACCAGCAAGGAT	1001213 - 1001231	768
	CT852-2[b]	TAGGCGCTCAACTTCTGGTATCTG	1001957 - 1001980	
CT859 (*ispH*)/	CT856/60-1[b]	GAGGGGGCTTTGCGGATTTAT	1011689 - 1011709	2779
CT860	CT856/60-2[b]	CCGGAATGCTTGGCTTGACA	1014448 - 1014467	
	CT856/60-3[e]	ACGACATTGAGTATGGATGA	1012253 - 1012272	
	CT856/60-4[e]	ACCCCGATATCTCATAAATC	1013829 - 1013848	
	CT856/60-5[e]	GTATTTCAGTTGCCTAAGGA	1012658 - 1012677	
CT861/CT862	CT861/2-1[c]	GAGGGCAGAGGCTTCTTCACAAG	1014122 - 1014144	2400
(*lcrH_2*)	CT861/2-2[b]	CTAGGCGTCCCAATTGGAGACTC	1016499 - 1016521	
	CT861/2-3[e]	AATAGCTCTCCAACCATCAA	1014640 - 1014659	
	CT861/2-4[e]	GACTATGAGGAAAGTTCTAC	1016071 - 1016090	
	CT861/2-5[e]	TCTCCTGTTGCTATTGTTTG	1014159 - 1014178	
	CT861/2-6[e]	TTGACTTTCTCCTGATGCTT	1015102 - 1015121	
CT867/CT868	CT867/8-1[c]	TCCCGACTGCTGGGGCTTAGA	1022902 - 1022922	2889
	CT867/8-2[c]	CATCGCGTCATGCCATGTCCTAT	1025768 - 1025790	
	CT867/8-3[e]	TCCTTCAGCTACTCTGATT	1022935 - 1022953	
	CT867/8-4[e]	TGGTAAGCGGATTACAGAT	1023641 - 1023659	
	CT867/8-5[e]	TTCATGGCGTTCTACAGAAT	1024302 - 1024321	
	CT867/8-6[e]	AATGTCAGAATCCCAAGCA	1024923 - 1024941	
	CT867/8-7[e]	AGAGGGAAGTCTTAATTTCC	1025542 - 1025561	

[a] Open reading frame (ORF) numbers, primer location (genome coordinates), primer sequences and amplicon size are based on the D/UW3-CX strain genome annotation (GenBank accession number NC_000117).
[b] Amplification primers also used for automated sequencing.
[c] Primers exclusively used for PCR amplification.
[d] Primer sequences, location and amplicon size refers to the L2/434 strain genome annotation (GenBank accession number NC_010287) as these primers were designed for amplification and automated sequencing in LGV and ocular strains, and they have no homology in the D/UW3-CX genome sequence.
[e] Primers exclusively used for automated sequencing.
[f] Due to the large gene size (for CT147 and CT682) or for PCR optimization (for CT684), two PCR primer pairs were designed to generate two overlapping amplicons for the entire gene. The primer pair for the amplification of the first region of the CT682 gene was previously described in [167] (data not shown).

Supplemental Table 2.2. List of the studied loci.

Loci [a]	Alleles Number [b]	d_N/d_S >1 (Z-test; p <0,05) [c]
IGR (CT044/CT045)	-	
IGR (CT047/CT048)	-	
CT048 (*yraL*)	6	
IGR (CT048/CT049)	-	
CT049	13	
CT050	11	
CT051	11	
CT058	11	
IGR (CT058/CT059)	-	
CT059 (*fer*)	5	
CT115	6	
IGR (CT115/CT116)	-	
CT116	8	+
CT117	7	
IGR (CT117/CT118)	-	
CT118	5	+
IGR (CT118/CT119)	-	
CT121 (*araD*) [d]	3	
CT144	9	
IGR (CT144/CT145)	-	
CT147 (*EEA1*)	12	
IGR (CT191/CT192)	-	
CT192	8	
IGR (CT192/CT193)	-	
CT195	8	
CT209 (*leuS*) [d,e]	5	
CT214	8	
IGR (CT222/CT223)	-	
CT223	11	+
IGR (CT223/CT224)	-	
CT228	6	+
IGR (CT228/CT229)	-	
CT229	7	+
IGR (CT229/CT230)	-	
IGR (CT231/CT232)	-	
CT232 (*IncB*)	5	
IGR (CT232/CT233)	-	
CT233 (*IncC*)	7	
IGR (CT233/CT234)	-	
CT245 (*pdhA*) [d,e]	2	
IGR (CT248/CT249)	-	
CT249	9	+
IGR (CT287/CT288)	-	
CT288	9	+
IGR (CT288/CT289)	-	

IGR (CT292/CT293)	-	
CT293 (*accD*)	4	
IGR (CT293/CT294)	-	
IGR (CT315/CT316)	-	
CT332 (*pykF*) [d,e]	5	
CT365	8	
IGR (CT365/CT366)	-	
CT376 (*mdhC*) [d,e]	4	
CT412 (*pmpA*)	10	
CT413 (*pmpB*)	13	+
IGR (CT413/CT414)	-	
CT414 (*pmpC*)	13	
CT432 (*glyA*) [d,e]	3	
IGR (CT441/CT442)	-	
CT442 (*crpA*)	8	+
IGR (CT442/CT443)	-	
CT443 (*omcB*)	7	
CT456 (*tarp*)	13	+
CT505 (*gapA*) [d]	3	
CT529	9	
IGR (CT529/CT530)	-	
CT618	8	
IGR (CT618/CT619)	-	
CT622	12	
IGR (CT622/CT623)	-	
CT623	6	
IGR (CT624/CT625)	-	
IGR (CT652.1/CT653)	-	
CT653 (*yhbG*) [e]	3	
IGR (CT653/CT654)	-	
CT674 (*yscC*)	11	
IGR (CT674/CT675)	-	
CT675 (*karG*)	10	
CT676	8	
IGR (CT676/CT677)	-	
CT677 (*frr*)	10	
CT678 (*pyrH*)	10	
IGR (CT678/CT679)	-	
CT679 *(tsf*)	12	
CT680 (*rs2* or *rpsB*)	12	
IGR (CT680/CT681)	-	
CT681 (*ompA*)	15	
IGR (CT681/CT682)	-	
CT682 (*pbpB*)	14	
IGR (CT682/CT683)	-	
CT683 (*TPR-motif protein*)	7	
IGR (CT683/CT684)	-	
CT684	6	

CT685	5	
CT686	9	
CT687 (*yfhO_1*)	5	
IGR (CT687/CT688)	-	
CT688 (*parB*)	8	
CT689 (*dppF*)	6	
CT690 (*dppD*) [d]	6	
CT694	10	+
IGR (CT698/CT699)	-	
CT713 (*porB*) [d]	4	
16SrRNA	6	
IGR (CT759/CT760)	-	
CT760 (*ftsW*)	6	
CT781 (*lysS*) [d,e]	4	
IGR (CT782/CT783)	-	
CT783	8	
IGR (CT783/CT784)	-	
IGR (CT796/CT797)	-	
CT812 (*pmpD*)	10	
IGR (CT812/CT813)	-	
CT813	7	+
IGR (CT813/CT814)	-	
IGR (CT817/CT818)	-	
CT818 (*tyrP_2*)	9	
IGR (CT818/CT819)	-	
IGR (CT851/CT852)	-	
CT852 (*yhgN*)	9	
IGR (CT852/CT853)	-	
CT859 (*ispH*)	6	
IGR (CT859/CT860)	-	
CT860	10	
IGR (CT860/CT861)	-	
CT861	7	
IGR (CT861/CT862)	-	
CT862 (*lcrH_2*)	3	
CT867	10	+
CT868	13	+
CT869 (*pmpE*)	11	
CT870 (*pmpF*)	9	
IGR (CT870/CT871)	-	
CT871 (*pmpG*)	12	
CT872 (*pmpH*)	10	
CT874 (*pmpI*)	12	

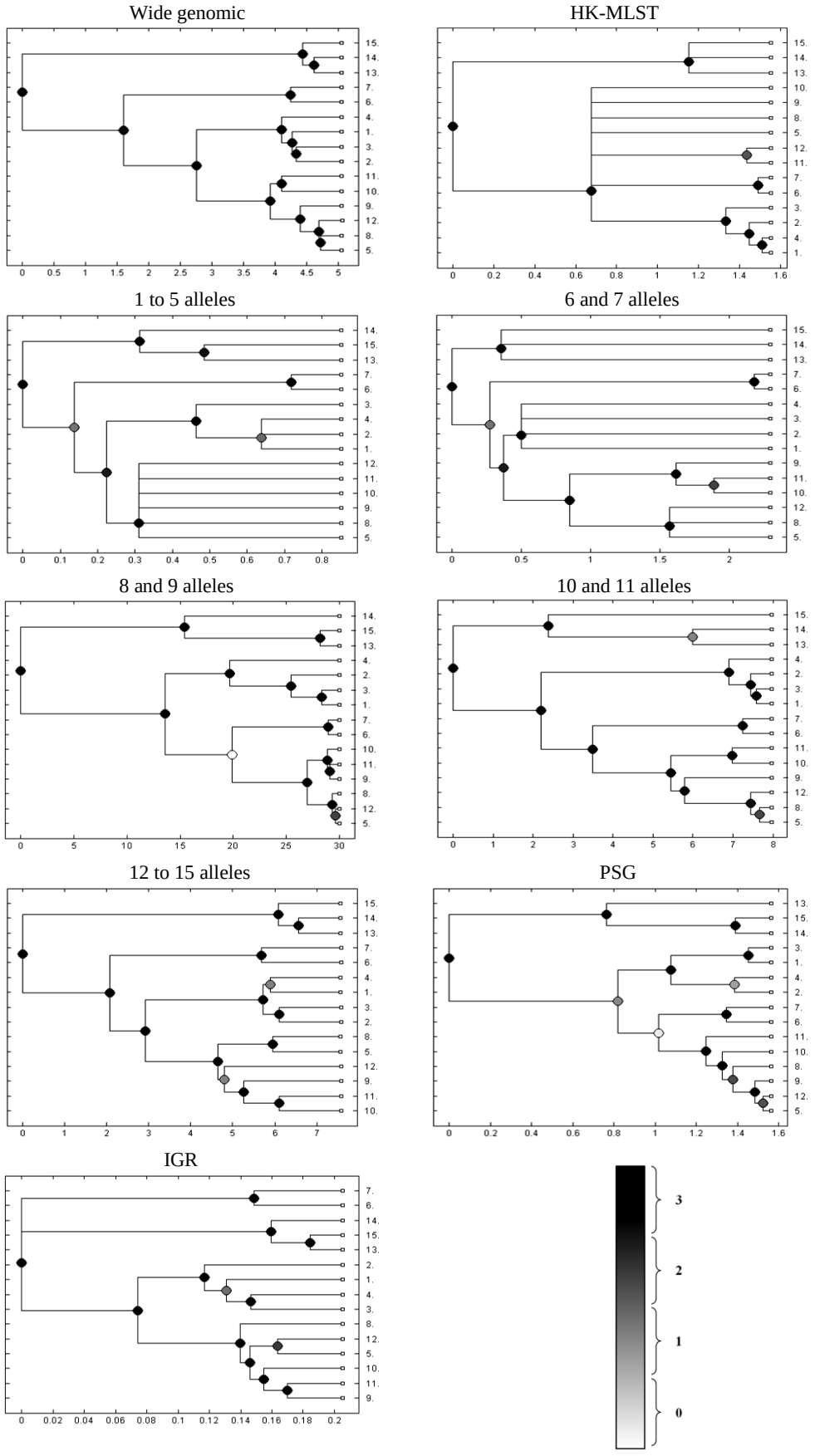
[a] Open reading frame (ORF) numbers are based on the D/UW3-CX strain genome annotation (GenBank accession number NC_000117).
[b] Number of alleles that each gene assigns were determined based on MEGA5 evolutionary analysis (distance matrices and phylogenetic trees) of each gene (for more details see the Materials and methods section of Chapter II).
[c] Genes putatively under positive selection that compose the PSG data set [173,209].

[d] For these genes, the partial sequences available at the GenBank were used.
[e] Genes that constitute a previously defined MLST scheme for *C. trachomatis* molecular characterization [211]. The partial sequences used for typing purposes compose the HK-MLST data set.

Supplemental Table 2.3. Contingency table for estimating the significance of the polymorphism present in the loci studied.

	Inside Region (bp) [a]	**Outside Region (bp)**	
SNPs	1802	1552	3354 [b]
Without SNPs	112530	926635	1039165
	114332	928187	1042519 [c]

[a] Chromosomal region evaluated in the present study (composed by the 136 genomic regions: 80 genes and 56 intergenic regions).
[b] Total number of SNPs identified between the *C. trachomatis* serovars A/Har13 (NC_007429) and D/UW3-CX (NC_000117) [122].
[c] Total length of the chromosome of *C. trachomatis* serovar D/UW3-CX (NC_000117). The polymorphism significance ($p < 10^{-7}$) of the analyzed regions was calculated using the Fisher's exact test.



Supplemental Figure 2.1. Trees generated by the tree comparison tool of ClonalFrame. This tool attributes a colored circle to each node of the first consensus tree according to the level of confidence

found between both trees of replicate runs (ranging from white, which indicates no confidence, to black, which indicates total confidence). For ranking purposes, a score was given to each node of the resultant tree, ranging from 0 (white nodes) to 3 (black nodes) in order to calculate the average concordance score for each data set. Strains A/Har13, B/Jali20, Ba/Apache2, C/TW3, D/UW3-CX, E/Bour, F/IC-Cal3, G/UW57, H/UW43, I/UW12, J/UW36, K/UW31, L1/440, L2/434-BU and L3/404 are numbered 1 to 15, respectively.

Supplemental Table 2.4. Accuracy results and r/m and ρ/θ estimates for all loci data sets.

Loci Data Sets	Accuracy Assessments			ClonalFrame Results			
	Trees concordance score	Gelman-Rubin test		r/m (mean; [95% CI])		ρ/θ (mean; [95% CI])	
		Test statistics	Convergence	1st run	2nd run	1st run	2nd run
Wide genomic	3	1.00 (θ, ν, δ); 0.99 (R)	+	0.21; [0.18-0.25]	0.21; [0.18-0.24]	0.01; [0.01-0.01]	0.01; [0.01-0.01]
HK-MLST	2.875	1.00 (θ, ν, δ, R)	+	0.91; [0.05-3.95]	1.09; [0.07-5.02]	0.12; [0.004-0.42]	0.14; [0.01-0.58]
1 to 5 alleles	2.667	1.00 (θ, δ, R); 0.99 (ν)	+	4.77; [1.77-10.72]	5.09; [1.76-12.36]	0.91; [0.27-2.27]	1.02; [0.27-2.87]
6 and 7 alleles	2.7	1.00 (δ, R); 0.99 (θ, ν)	+	7.04; [3.45-12.67]	6.98; [3.47-12.57]	0.89; [0.42-1.71]	0.87; [0.41-1.62]
8 and 9 alleles	2.643	1.04 (θ); 1.15 (R); 1.58 (ν); 1.23 (δ)	-	2.38; [1.54-3.65]	2.44; [1.55-3.77]	0.21; [0.12-0.33]	0.21; [0.12-0.33]
10 and 11 alleles	2.786	1.13 (θ); 1.21 (R); 1(ν); 1.28 (δ)	-	3.41; [0.47-5.21]	4.12; [3.09-5.38]	0.27; [0.03-0.41]	0.35; [0.27-0.45]
12 to 15 alleles	2.714	6.51 (θ); 1(R); 16.99 (ν); 1.02 (δ)	-	0.58; [0.3-0.96]	0.29; [0.21-0.38]	0.03; [0.01-0.05]	0.01; [0.01-0.01]
PSG	2.357	2.70 (θ); 1.73 (R); 37.74 (ν); 11.45 (δ)	-	0.05; [0.02-0.1]	3.15; [1.6-5.31]	0.08; [0.05-0.13]	0.39; [0.2-0.7]
IGR	2.769	1.00 (θ, ν, δ, R)	+	1.38; [0.82-2.17]	1.38; [0.8-2.19]	0.3; [0.17-0.48]	0.29; [0.17-0.47]

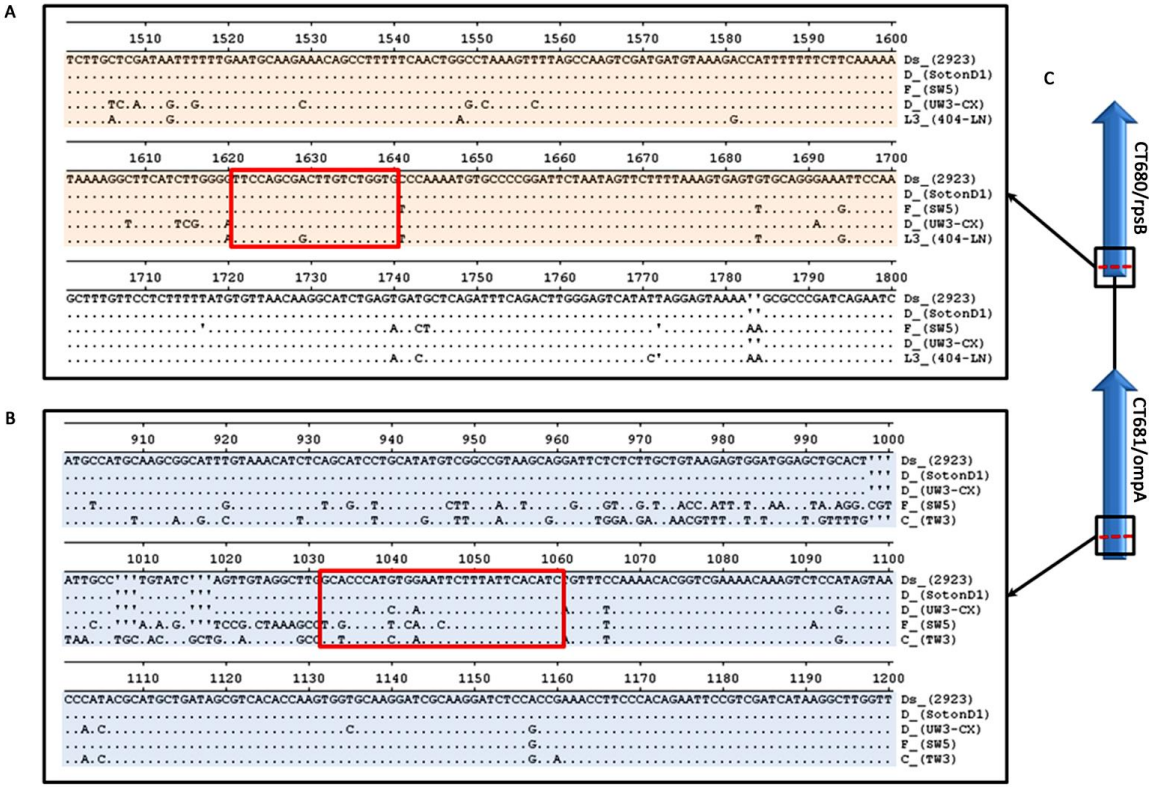
CI – Confidence interval; θ – mutation rate; *ν* – rate of new polymorphism introduced by recombination; δ – average tract length of a recombination event; R – recombination rate; r/m – measure of the weight of recombination on the diversification relative to mutation; ρ/θ – measure of the frequency of occurrence of recombination relative to mutation events.

Supplemental Table 3.1. *C. trachomatis* strains used in Chapter III.

Strains	Accession nº.	Isolation location	Reference
A/Har13	CP000051	Conjunctiva	[122]
A/2497	FM872306	Ocular	[30]
A/363	HE601796	Ocular	[30]
A/5291	HE601810	Ocular	[30]
A/7249	HE601797	Ocular	[30]
B/TZ1A828/OT	FM872307	Ocular	[189]
B/Jali20/OT	FM872308	Ocular	[189]
C/TW-3	CP006945	Conjunctiva	[328]
D/UW-3/CX	AE001273	Cervix	[83]
D(s)/2923	ACFJ01000001	Cervix	[169]
D/SotonD1	HE601798	Endocervix	[30]
D/SotonD5	HE601799	Endocervix	[30]
D/SotonD6	HE601800	Endocervix	[30]
E/Bour	HE601870	Ocular	[30]
E/SW2	FN652779	Urethra	[329]
E/SW3	HE601801	Cervix	[30]
E/SotonE4	HE601802	Endocervix	[30]
E/SotonE8	HE601803	Endocervix	[30]
E/11023	CP001890	Cervix	[169]
E/150	CP001886	Rectum	[169]
F/SW4	HE601804	Cervix	[30]
F/SW5	HE601805	Cervix	[30]
F/SotonF3	HE601806	Endocervix	[30]
F(s)/70	ABYF01000001	Cervix	[169]
G/9301	CP001930	Urethra	[169]
G/9768	CP001887	Rectum	[169]
G/11222	CP001888	Cervix	[169]
G/11074	CP001889	Rectum	[169]
G/SotonG1	HE601807	Endocervix	[30]
J/6276	ABYD01000001	Cervix	[169]
Ia/SotonIa1	HE601808	Endocervix	[30]
Ia/SotonIa3	HE601809	Endocervix	[30]
K/SotonK1	HE601794	Endocervix	[30]
L1/440/LN	HE601950	Lymph node	[30]
L1/1322/p2	HE601951	Genital ulcer	[30]
L1/115	HE601952	LGV patient	[30]
L1/224	HE601953	LGV patient	[30]
L2/434/Bu	AM884176	Lymph node	[84]
L2/25667R	HE601954	Rectal biopsy	[30]
L2b/UCH-1	AM884177	Rectum	[84]
L2b/8200/07	HE601795	Rectum	[30]
L2b/UCH-2	HE601956	Rectum	[30]
L2b/Canada1	HE601963	Rectum	[30]
L2b/Canada2	HE601957	Rectum	[30]
L2b/LST	HE601958	Rectum	[30]
L2b/CV204	HE601960	Rectum	[30]
L2b/795	HE601949	Rectum	[30]
L2b/Ams1	HE601959	Penile ulcer	[30]
L2b/Ams2	HE601961	Anus	[30]
L2b/Ams3	HE601962	Anus	[30]
L2b/Ams4	HE601964	Anus	[30]
L2b/Ams5	HE601965	Anus	[30]
L3/404/LN	HE601955	Lymph node	[30]

Supplemental Table 3.2. Bioinformatical results of all *C. trachomatis* ORFs with detailed information of putative pseudogenes, strains' segregation, overall mean distances and dN/dS values. The genes' annotation corresponds to the D/UW3-CX strain.

Due to the massive extent of this table, it was impossible to present it in a printable format. Please access the link above to review the table in the original paper:
http://www.g3journal.org/content/suppl/2014/11/05/g3.114.015354.DC1/TableS2.xlsx



Supplemental Figure 3.1. Nucleotide sequences of crossovers for strains D(s)/2923 and D/SotonD1. Crossover regions (red boxes) are delimitated by informative sites from SimPlot/BootScan analysis. Panels A and B represent the partial alignments used for the determination of the crossovers in CT680/*rpsB* and the CT681/*ompA*, respectively. Panel C shows the genetic localization of those partial alignments.

Supplemental Table 4.1. Primers used in the qPCR assays.

ORF	Product Description	Primers	Primer Sequence (5' to 3')	Primer Location[a]	Amplicon Size (bp)
Cloning:					
lacZ (pCR®2.1 vector)	β-galactosidase (reporter gene)	M13F M13R	GTAAAACGACGGCCAG CAGGAAACAGCTATGAC	404-389[b] 205-221[b]	200[c]
Chromosome:					
CT681[d]	Major outer membrane protein (OmpA)	OmpA-9[e] OmpA-10[e]	TGCCGCTTTGAGTTCTGCTT GTCGATCATAAGGCTTGGTTCAG	33-52[f] 108-86[f]	76
CTr01/CTr04[d]	16S ribosomal RNA (16SrRNA)	16SRNA-9[e] 16SRNA-10[e]	GCGAAGGCGCTTTTCTAATTTAT CCAGGGTATCTAATCCTGTTTGCT	734-756[f] 809-786[f]	76
Plasmid:					
ORF1[g]	Integrase/recombinase homologue (pGP7)	pCTA_1-C pCTA_1-D	TCTTTGCGCACAGACGATCT GCGAAAGGAAATCTGATTGGAT	508-527[h] 558-537[h]	51
ORF2[g]	Integrase/recombinase homologue (pGP8)	pCTA_8-A pCTA_8-B	GCGGTCCAATGCATAATAACTTC GGAAACGCATGAAAAGCTTCTC	149-127[h] 98-119[h]	52
ORF3[g]	DnaB like helicase (pGP1)	pCTA_7-1 pCTA_7-2	CGGCTTGGGAAGAGCTTTT GCAACATTAACCCGAGATACGAT	200-218[h] 268-246[h]	69
ORF4[g]	Hypothetical protein (pGP2)	pCTA_6-1 pCTA_6-2	AAAGTCCTATCCACCTTGAAAATCA TTTGAGCCAATTTGGGAGATATC	32-56[h] 112-90[h]	81
ORF5[g]	Hypothetical protein (pGP3)	pCTA_5-1 pCTA_5-2	CAAAAGCTCTGGGAGCATGTT CAACGCCGCCTTCCATT	468-488[h] 535-519[h]	68
ORF6[g]	Hypothetical protein (pGP4)	pCTA_4-1 pCTA_4-2	GGCTTTGATTATGCTTATCTTGTCTAGA TCAGCCTTGGAAAACATGTCTTT	255-282[h] 305-283[h]	51
ORF7[g]	Partitioning protein homologue (pGP5)	pCTA_3-E pCTA_3-F	CGGTCGGAAAACCTGAAGAAG CCCAAAAAGACAAAGCTATTCCAA	437-457[h] 490-467[h]	54
ORF8[g]	Hypothetical protein (pGP6)	pCTA_2-1 pCTA_2-2	ATAAACCTCCCCAACCAAACTCT CGGCCAAAATATATGCGGATT	397-419[h] 469-449[h]	73
sRNA-2	Small anti-sense RNA to ORF2/*pgp8*	pCTA_8-1 pCTA_8-2	ATTTTTCCGGAGCGAGTTACG GTACATCGGTCAACGAAGAGGTT	758-738[h] 708-730[h]	51
sRNA-7	Small anti-sense RNA to ORF7/*pgp5*	pCTA_3-1 pCTA_3-2	CTGACCTAGACCCGCAATCC TGACACTAGCCCCCAATCCA	47-66[h] 97-78[h]	51

[a] Primer locations in the respective gene.
[b] Locations according to TA Cloning® Kit guide, version V (Invitrogen).
[c] Amplicon size without the gene fragment cloned.
[d] Open reading frame (ORF) numbers are based on the D/UW3-CX strain genome annotation (GenBank No. NC_000117).
[e] Previously described in [282].
[f] Based on the sequence of L2/434-BU strain (GenBank No. NC_010287).
[g] ORF numbers according to [183].
[h] Based on the sequence of the plasmid ORFs of A/Har13 strain (GenBank No. CP000052).

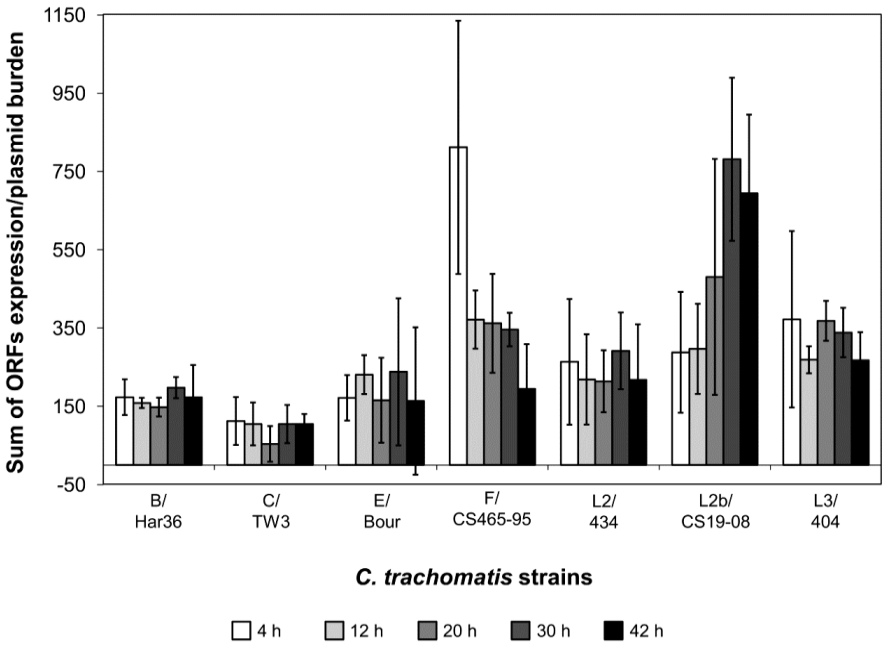
Supplemental Table 4.2. Polymorphism analysis of the eight *C. trachomatis* plasmid ORFs by using 44 available plasmid sequences.

ORF[a]	ORF Size (bp)[b]	**Overall Mean Distance** (pairwise deletion; bootstrap = 1000 replicates)													
		Nucleotide				**Amino Acid**				**Kumar Method** (*p*-distance; Non- and Synonymous substitutions)					
		No. of differences	[SE]	*p*-distance	[SE]	No. of differences	[SE]	*p*-distance	[SE]	Mean dS	[SE]	Mean dN	[SE]	dN/dS[c]	[SE]
ORF1/*pgp7*	918	5.02	1.33	0.005	0.001	1.94	0.82	0.006	0.003	0.01	0.004	0.003	0.001	0.3	0.156
ORF2/*pgp8*	993	4.45	1.36	0.004	0.001	0.28	0.28	0.001	0.001	0.012	0.004	0.0004	0.0004	0.03333	0.035
ORF3/*pgp1*	1356	5.3	1.52	0.004	0.001	1.92	0.85	0.004	0.002	0.008	0.003	0.002	0.001	0.25	0.156
ORF4/*pgp2*	1065	5.64	1.51	0.005	0.001	1.31	0.69	0.004	0.002	0.011	0.004	0.002	0.001	0.18181	0.112
ORF5/*pgp3*	795	7.05	1.76	0.009	0.002	5.54	1.48	0.021	0.006	0.006	0.003	0.01	0.003	1.6667	0.972
ORF6/*pgp4*	309	0.5	0.49	0.002	0.002	0.5	0.49	0.005	0.005	0	0	0.002	0.002	-	-
ORF7/*pgp5*	732	2.29	0.86	0.003	0.001	1.47	0.64	0.006	0.003	0.003	0.002	0.003	0.001	1	0.745
ORF8/*pgp6*	744	3.3	1.14	0.004	0.002	1.72	0.84	0.007	0.003	0.006	0.003	0.004	0.001	0.6667	0.471

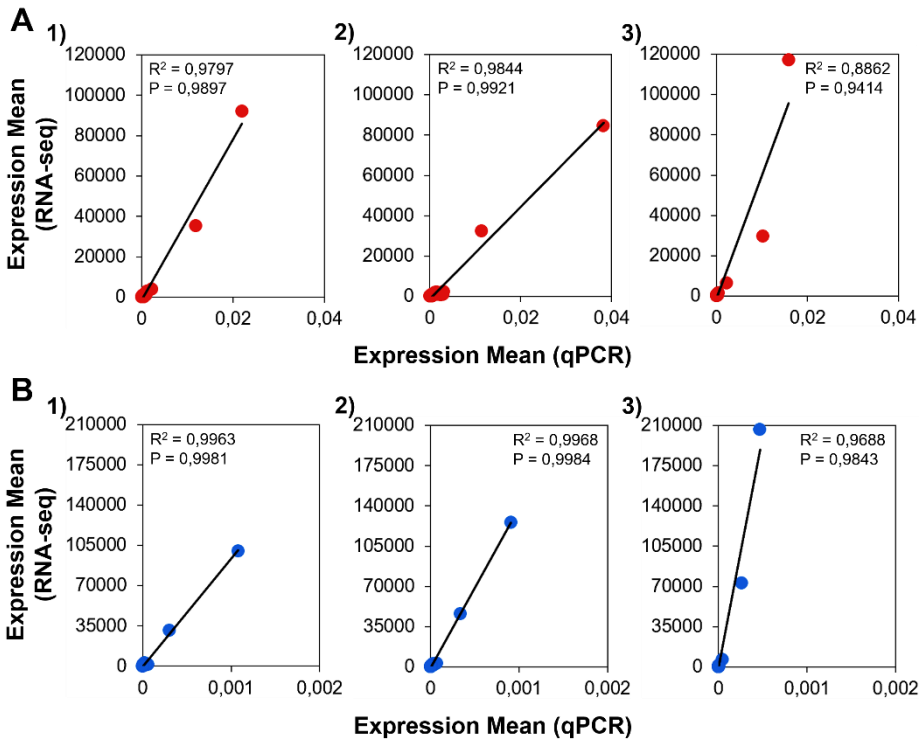
[a] ORFs numbers according to [183].

[b] Based on the sequence of the plasmid of A/Har13 strain (GenBank No. CP000052).

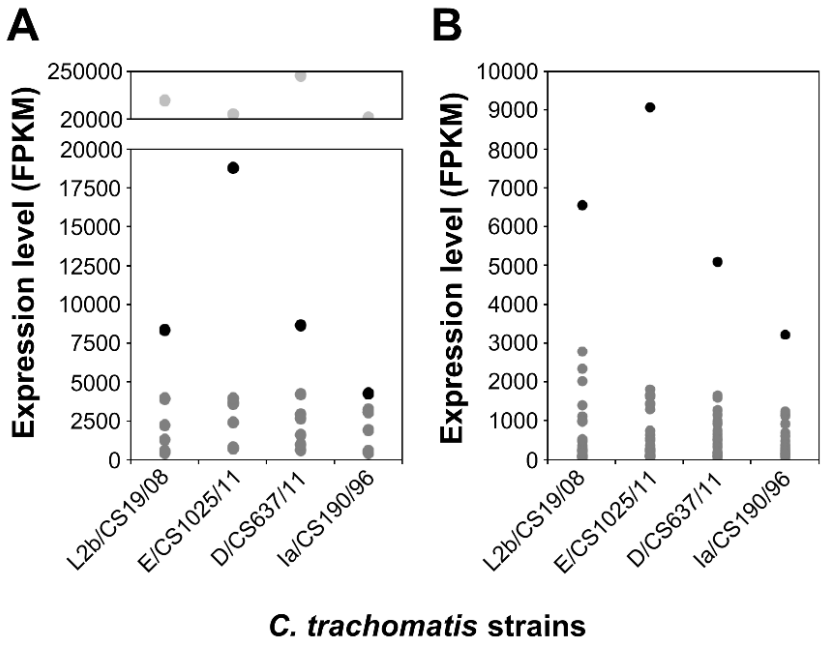
[c] The value of dN/dS > 1 obtained for ORF5 was not statistically significant by the Z-test of positive selection.



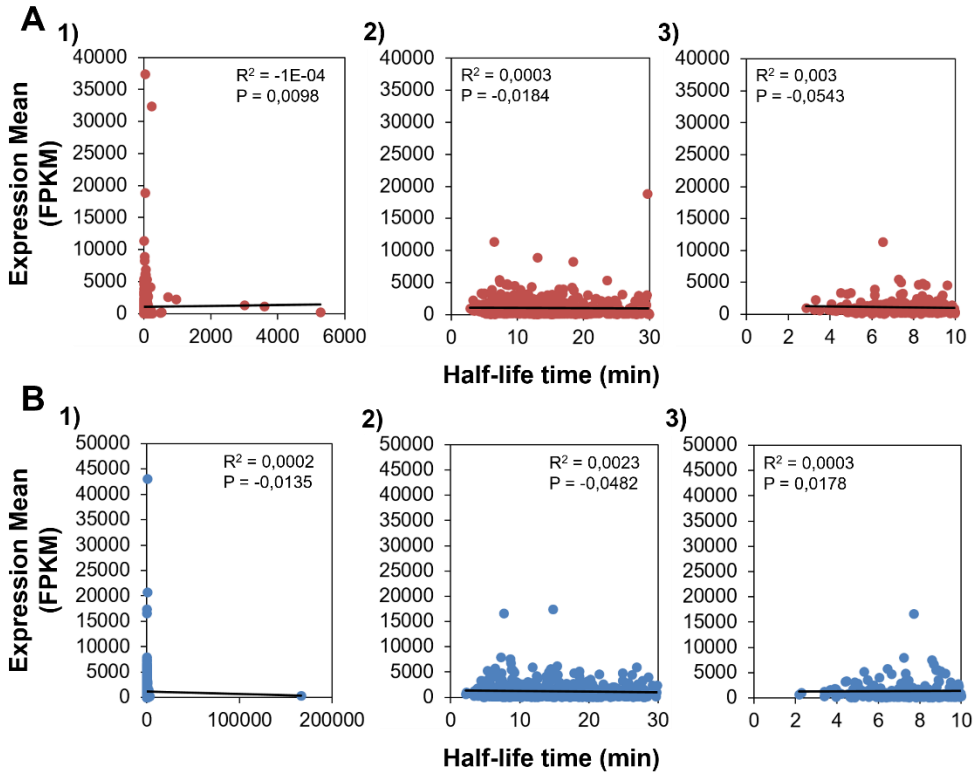
Supplemental Figure 4.1. Global transcriptional activity of the plasmid ORFs *per* plasmid throughout *C. trachomatis* development. The graph represents, for each time-point, the sum of the mean expression values of all ORFs obtained at the respective time-point. Vertical lines represent the final standard deviations.



Supplemental Figure 5.1. Relation of the expression levels acquired by qPCR (horizontal axis) and RNA-seq (vertical axis), for two *C. trachomatis* strains: L2b/CS19/08 (A) and E/CS1025/11 (B). Each dot represents the relation of the expression values obtained by the two methods, for the same transcript, of that particular strain. For each strain is shown that relation regarding three different sets of transcripts: 1) All transcripts quantified for each strain; 2) transcripts with a calculated half-life time ≤ 30 min; and 3) transcripts with a calculated half-life time ≤ 10 min. Both the linear correlation coefficient (R^2) and the Pearson correlation coefficient (P) are shown in each graph, and indicating strong correlation.



Supplemental Figure 5.2. Expression levels of plasmid-encoded transcripts (A) and chromosomal genes putatively regulated by the plasmid-encoded gene ORF6/*pgp4* (B). In panel A, the plasmid-encoded transcripts are represented by different colours in order to facilitate the discrimination of the expression differences between the strains: ORF2/*pgp8* in light grey, ORF6/*pgp4* in black, and the remainder (ORF1/*pgp7*, ORF3/*pgp1*, ORF4/*pgp2*, ORF5/*pgp3*, ORF7/*pgp5*, and ORF8/*pgp6*) in dark grey. In panel B, the CT798/*glgA* is shown as black dots also for a better perception of its expression profile between strains, while the other chromosomal genes evaluated are represented by dark grey dots.



Supplemental Figure 5.3. Pairwise relation between the genes' expression level, determined at the mid-stage of the developmental cycle (T_0), and their $t_{1/2}$. The three graphs, both in panel A and panel B, represent the relation obtained for three different groups of genes of L2b/CS19/08 and E/CS1025/11 strains, respectively. Those three groups of genes were defined according to the $t_{1/2}$ they exhibited: 1) all transcripts with calculated $t_{1/2}$; 2) transcripts with $t_{1/2} \leq 30$ min; and 3) transcripts with $t_{1/2} \leq 10$ min. Both the linear correlation coefficient (R^2) and the Pearson correlation coefficient (P) are shown in each graph.