Alexandra Isabel Cardoso Nunes

Genomic and Transcriptomic Features of Chlamydia trachomatis: Tracking the Basis for the Ecological Success

Lisboa
2010
The work described in this Ph.D. thesis was carried out at the National Reference Laboratory of Bacterial Sexually Transmitted Infections of the Infectious Diseases Department from the National Institute of Health, Lisbon, Portugal.

All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, without the prior written permission of the author.

ALEXANDRA ISABEL CARDOSO NUNES

Genomic and Transcriptomic Features of Chlamydia trachomatis: Tracking the Basis for the Ecological Success

Dissertation presented to obtain a Ph.D. degree in Biology, speciality in Molecular Genetics, by Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.

Supervisors: Doctor João P. Gomes & Doctor Maria J. Borrego
Instituto Nacional de Saúde Dr. Ricardo Jorge

Co-supervisor: Prof. Doctor Ana Madalena Ludovice
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Lisboa
2010
Não há ninguém com quem deseje mais partilhar o encerrar desta difícil etapa na minha vida, senão com o meu querido maninho. Ricardo (que saudades em escrever o teu nome...), mesmo estando longe, sinto que tens estado sempre presente. A promessa que te fiz tem-me acompanhado constantemente e, é nela em que me apoio quando preciso de força e coragem. Tens sido o meu modelo ao longo destes anos. Foste, és e sempre serás a pessoa especial da minha vida, a minha inspiração, a minha rocha e, sobretudo, o meu incansável protector... Não houve ninguém que tivesse acreditado em mim como tu... Este 'triunfo' é teu também! Não me posso sentir mais orgulhosa, honrada e privilegiada por saber que continuas presente no coração das pessoas com quem conviveste e, no quanto elas ainda recordam a tua humildade, integridade, bondade e excelente sentido de humor. Não foste esquecido, maninho... e, ao fim destes anos, eu sei que a tua passagem aqui será eternamente lembrada com uma saudade imensa, não só por mim e pelos pais, mas por todos. Isso é a maior dádiva que alguém pode ter... e, para nós, tua família, é um conforto mínimo por a vida nos ter sido tão madrasta. Como mana, é um orgulho enorme ter-te como irmão!!

Sinto falta daqueles teus abraços, maninho...
Fazes falta... muita falta...

Esta tese é para ti...
ACKNOWLEDGEMENTS / AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer aos meus orientadores, JP e Zézita, sem os quais nada disto teria sido possível. Obrigada, não só pela excelente orientação, mas também pela paciência, disponibilidade e constante incentivo ao longo destes seis anos. Obrigada por terem acreditado em mim…

Não poderia deixar de agradecer à Professora Ana Madalena pela total disponibilidade e apoio sempre demonstrados ao longo destes anos, mesmo durante a faculdade. Embora nunca lho tenha dito, a Prof.ª foi, sem dúvida, um dos motivos do meu interesse pelo mundo da investigação; a emoção que transmitia nas suas aulas, era simplesmente fascinante…

Gostaria de agradecer à D. Albertina e à D. Arminda pela simpatia com que me acolheram e, total disponibilidade, paciência e sobretudo gentileza demonstradas. Foram as minhas ‘mães’ substitutas… O meu sincero obrigado por terem cuidado de mim!

Ao Vítor e à Rita, pela disponibilidade nos últimos tempos, sobretudo por me terem permitido dedicar exclusivamente à escrita da tese ao assegurarem por completo os trabalhos em execução. Ao meu colega Carlitos, pela sua constante boa disposição!

À Silvita pelas conversas longas e motivação!

Ao Instituto Nacional de Saúde Dr. Ricardo Jorge, em particular, a todas as pessoas do Departamento de Doenças Infecciosas pela simpatia e apoio prestados ao longo destes anos. Em especial, não podia deixar de agradecer à minha querida D. Goretti pela preocupação e amizade, ao Sr. Belém, Sr. Ezequiel, D. Jacinta e D. Alice por me terem tratado sempre bem, ao Nuno Verdasca pela ajuda fundamental na formatação da tese e, a toda a equipa de segurança pela companhia, simpatia e apoio durante as minhas múltiplas noitadas no Instituto.

À Fundação para a Ciência e Tecnologia, pela bolsa de doutoramento atribuída (SFRH/BD/25651/2005) e restante apoio financeiro.

À minha ‘sofiita’, por teres estado sempre presente… mesmo separadas por quilómetros de distância, obrigada por me ouvires desabafar vezes e vezes sem conta sobre o mesmo!! A tua companhia via Skype (nos últimos tempos, já diária!), foi fundamental para manter a lucidez e aligeirar o sacrifício que estes quatro anos exigiram, sobretudo nesta última fase. Obrigada, miga, pelos inúmeros Pdfs!!! Sem eles, esta tese seria tão pobre… :D Adoro-te linda!!!
Ao meu Nuninho, que mesmo do outro lado do mundo, sei o quanto te preocupas e estás a torcer por mim. Obrigada pelas longas conversas às 2h da manhã… Tens sido um verdadeiro amigo! És lindo!!!

A todos os meus amigos, pelo incentivo constante e apoio demonstrados… principalmente àqueles (Martinha e Ruizinho, João e Nês, Carlitos…) que têm impressora e me oferecem cházinho, caipirosca, ginginhas ou jantares!!! :D Obrigada por terem compreendido a minha ausência, sobretudo nesta última fase, sem nunca se terem esquecido de mim ou deixado de se preocupar comigo. A todos, o meu sincero obrigado!!

Aos meus segundos pais, Titi e Gracias, por se terem sempre preocupado comigo…

Aos meus queridos avós, ‘Janeta’ e ‘Merrau’, que desejaram tanto assistir ao meu ‘sucesso’ mas, com as voltas da vida, infelizmente já não estão entre nós…

Finalmente e, seguramente com a maior importância, quero agradecer às duas pessoas que mais me apoiaram, incentivaram e, sobretudo sofreram ao longo destes anos: aos meus queridos pais. Obrigado por serem o meu ‘porto de abrigo’ e, pela influência que tiveram na minha vida. Meu Ti-no-ni (ou Gugu, dependendo dos dias!) e minha Glorinha, vocês são tudo para mim… Os dois, por igual!!!!! :D Não poderia ter maior orgulho em vós ter como papas!! Sem vocês, nada disto teria sido possível… este ‘sucesso’ é vosso também! Desculpem os sacrifícios que foram muitas vezes obrigados a fazer por minha causa…
NOTES OF THE AUTHOR

- The work reported in this thesis encompasses several studies performed throughout the author’s Ph.D. project, constituting the subject of six independent publications that are presented here as individual chapters. For normalization purposes, chapters were formatted in a unique style since almost all of the described studies are reproductions of published refereed papers with different layouts. Also, to eliminate any confusion arising from the differential use in these publications of various terms, such as ‘serovar’ or ‘genotype’ to designate different types of Chlamydia trachomatis (C. trachomatis) strains, or disparate abbreviations of the same term, each chapter was modified accordingly.

- The order of presentation of these chapters does not necessarily reflect a chronological order, as some of the studies described below were performed simultaneously, with the results obtained during one particular study influencing the progress of the others and vice-versa. Also, the time between the submission of an article and its publication date largely depends on the journal as well as on possible necessary revisions. As a consequence, the prototype strains used as baselines to represent C. trachomatis genotypes may slightly differ among studies/chapters.

- As each chapter contains a specific background and a detailed discussion of the results, this dissertation involves only a first brief overview of the unique C. trachomatis biology and a final global discussion of the major conclusions with regards to recent findings, in order to put into context and emphasize all the performed studies of this thesis.

- For normalization purposes, a single section of ‘Supplemental Material’ and ‘References’ is presented at the end of this dissertation. As two out of the six presented studies were published in PLoS ONE journal, all references were formatted according to the PLoS journals style (using the EndNote programme), with citations indicated by a reference number in square brackets throughout the thesis.
SUMMARY

*Chlamydia trachomatis* is an obligate intracellular pathogen, comprehending 18 genotypes responsible for ocular, urogenital, and inguinal lymph node infections worldwide. Genotypes present an intriguing biological uniqueness given their colossal genomic similarity, and the molecular basis underlying those disparities is still elusive.

The scope of this thesis was to search for genomic and transcriptomic features that distinguish *C. trachomatis* genotypes by analyzing several polymorphic loci potentially important in the chlamydial biology.

We started by validating normalization strategies for real-time expression data in *C. trachomatis*. Then, we performed transcriptomic and immunoreactivity analyses of the nine-member polymorphic membrane protein gene (*pmp*) family for prototype and clinical strains. Both differential immunoreactivity and expression inter-*pmps* and inter-strains suggest a Pmp variable surface expression according to strain-specific needs, which may promote an important phenotypic diversity in terms of antigenicity, virulence, tissue tropism, and ecological success.

To better understand the impact of the host pressure on *C. trachomatis*, we evaluated the evolutionary mutational dynamics of the gene encoding the chlamydial key antigen (MOMP). We found that MOMP variability emerges from intrinsic trends likely driven by its complex pathogenesis-related functions. Both the rampant B-cell antigenic variation and the high conservation of T-cell epitope clusters evidence the existence of distinct adaptive evolutionary antigenic scenarios that may benefit the pathogen. Moreover, the apparent MOMP conservation among strains from the two most prevalent genotypes worldwide (E and F) suggests the existence of more fitted antigenic profiles.

Finally, we performed a high-scale phylogenomic analysis to study the evolution of *C. trachomatis* genotypes. We found that their genetic variability reflects an evolutionary adaptation to each infected tissue, and also an independent co-evolutionary pathway for E and F. We showed that radiation of genotypes sharing the same cell appetite involved primarily the accumulation of mutations on the whole chromosome, where, beyond surface-exposed protein genes (like *pmps*), several hypothetical protein genes appear to be important.

Overall, the findings presented in this Ph.D. thesis suggest that *C. trachomatis* diversity in terms of tissue tropism, pathogenesis and ecological success relies on complex and likely strain-specific genomic and transcriptomic features that probably reflect a global chromosomal dynamics. In particular, the unique genomic make-up of genotypes E and F
suggests the emergence of successful clones, well-adapted to play the ‘arms race game’ with the host.

Key-words: Chlamydia trachomatis, evolution, polymorphism, antigen, tissue tropism, ecological success, genomics, transcriptomics.
RESUMO

*Chlamydia trachomatis* é um patogénio intracelular obrigatório, compreendendo 18 genótipos responsáveis por infecções oculares, urogenitais e dos glânglios linfáticos inguiniais. Os genótipos apresentam uma singularidade fenotípica enigmática, dada a elevada similaridade genómica, sendo ainda desconhecida a base molecular responsável por tal disparidade.

O âmago desta tese consistiu na pesquisa de características genómicas e transcriptómicas que diferenciam os diversos genótipos, através da análise de múltiplos loci polimórficos potencialmente importantes para a biologia de clamídia.

Inicialmente, foram validadas estratégias de normalização de dados de expressão genética por PCR em tempo real em *C. trachomatis*. Em seguida, foram efectuadas análises transcriptómicas e de imunoreactividade para os nove genes (*pmp* s) que codificam proteínas membranares polimórficas, em estirpes protótipo e clínicas. Tanto as diferenças de imunoreactividade como de expressão inter-*pmp* s e inter-estirpes sugerem uma expressão variável das Pmps à superfície da bactéria consoante as necessidades específicas de cada estirpe, podendo promover uma diversidade fenotípica importante.

Para uma melhor compreensão do impacto da pressão imunitária do hospedeiro em *C. trachomatis*, foi avaliada a dinâmica mutacional do gene que codifica o principal antigénio (MOMP). Foi possível verificar que a variabilidade na MOMP resulta de pressões selectivas específicas, provavelmente associadas às suas complexas funções patogénicas. A extensa variabilidade nos epitopos das células B e a elevada conservação dos epitopos das células T sugerem a existência de cenários evolutivos distintos potencialmente benéficos para o patogénio. Além disso, a aparente conservação da MOMP entre estirpes pertencentes aos dois genótipos mundialmente mais prevalentes (E e F) sugere que os respectivos perfis antigénios estejam mais adaptados.

Por fim, foi realizada uma mega análise genómica e filogenética para estudar a evolução dos genótipos, tendo revelado uma variabilidade genética que reflecte uma adaptação evolutiva a cada tecido infectado bem como uma co-evolução dos genótipos E e F. A radiação das estirpes com apetência celular semelhante parece ter ocorrido por acumulação de mutações dispersas no cromossoma, onde, para além dos genes que codificam proteínas expostas à superfície, vários genes que codificam proteínas hipotéticas parecem ter relevância.

Globalmente, estes resultados sugerem que a diversidade de tropismo, patogénese e sucesso ecológico de *C. trachomatis* deriva de características genómicas e transcriptómicas complexas, reflectindo uma dinâmica cromossomal global, onde as Pmps, a MOMP e
algumas proteínas hipotéticas parecem desempenhar um papel importante. Em particular, a singularidade genómica dos dois genótipos com maior sucesso ecológico sugere a emergência de clones dominantes com um *make-up* genómico favorável à interacção com o hospedeiro.

*Palavras-chave: Chlamydia trachomatis*, evolução, polimorfismo, antigénio, tropismo celular, sucesso ecológico, genómica, transcriptómica.
THESIS OUTLINE

This Ph.D. dissertation is divided into four main sections that encompass:

i) a general introduction (CHAPTER 1) that intend to briefly describe the unique biology of *C. trachomatis*, in particular the current knowledge about the molecular factors underlying the distinct tissue tropism, pathogenesis, and ecological success of the 18 *C. trachomatis* serovars. The issues focused on this chapter aim to put into context and to emphasize the relevance of the studies described throughout the thesis. Ultimately, the detailed aims of the present Ph.D. thesis are presented.

ii) a set of six genomic and transcriptomic studies over several polymorphic loci (CHAPTERS 2 to 7) in order to get further insights for deciphering the molecular basis of *C. trachomatis* pathobiologic diversity and ecological success. This section constitutes the main body of this thesis, and reproduces the contents of the following publications:

iii) an overall discussion and final conclusions (CHAPTER 8), where the major findings of each performed study are highlighted and discussed in a more general context, taking into account the scope of this thesis.

iv) future perspectives (CHAPTER 9) that intend to briefly describe the subsequent research steps that emerge from the findings obtained as well as from the questions raised during the course of the thesis.
LIST OF ABBREVIATIONS

General abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ab</td>
<td>Antibody</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>BEB</td>
<td>Bayes empirical Bayes</td>
</tr>
<tr>
<td>Biovar</td>
<td>Biological variant</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CD</td>
<td>Constant domain</td>
</tr>
<tr>
<td>CEP</td>
<td>Cell envelope protein gene</td>
</tr>
<tr>
<td>CPAF</td>
<td>Protease-like activity factor</td>
</tr>
<tr>
<td>Ct</td>
<td>Threshold cycle</td>
</tr>
<tr>
<td>CTL</td>
<td>Cytotoxic T lymphocyte</td>
</tr>
<tr>
<td>dN</td>
<td>Nonsynonymous substitution rate</td>
</tr>
<tr>
<td>dS</td>
<td>Synonymous substitution rate</td>
</tr>
<tr>
<td>EB</td>
<td>Elementary body</td>
</tr>
<tr>
<td>EEA1</td>
<td>Human endosomal antigen 1</td>
</tr>
<tr>
<td>ESSTI</td>
<td>European Surveillance of Sexually Transmitted Infections</td>
</tr>
<tr>
<td>FD</td>
<td>Fold-difference</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic DNA</td>
</tr>
<tr>
<td>HGT</td>
<td>Horizontal gene transfer</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td>HK</td>
<td>Housekeeping gene</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigen</td>
</tr>
<tr>
<td>HP</td>
<td>Hypothetical/unclassified protein gene</td>
</tr>
<tr>
<td>HPV</td>
<td>Human papilloma virus</td>
</tr>
<tr>
<td>IAP</td>
<td>Inhibitor of apoptosis protein</td>
</tr>
<tr>
<td>IDO</td>
<td>Indoleamine-2,3-dioxygenase</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>Interferon-gamma</td>
</tr>
<tr>
<td>IGR</td>
<td>Intergenomic region</td>
</tr>
<tr>
<td>Inc</td>
<td>Inclusion membrane protein</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion/deletion event</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>κ</td>
<td>Ratio of Ts to Tv</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>K2P</td>
<td>Kimura two-parameter</td>
</tr>
<tr>
<td>LD</td>
<td>Lipid droplet</td>
</tr>
<tr>
<td>LGV</td>
<td>Lymphogranuloma venereum</td>
</tr>
<tr>
<td>LPS</td>
<td>Lipopolysaccharide</td>
</tr>
<tr>
<td>M</td>
<td>Reference gene stability measure</td>
</tr>
<tr>
<td>MHC</td>
<td>Major histocompatibility complex</td>
</tr>
<tr>
<td>MOMP</td>
<td>Major outer membrane protein</td>
</tr>
<tr>
<td>MP</td>
<td>Maximum parsimony</td>
</tr>
<tr>
<td>MSM</td>
<td>Men who have sex with men</td>
</tr>
<tr>
<td>MVB</td>
<td>Multivesicular body</td>
</tr>
<tr>
<td>NF</td>
<td>Nuclear factor</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institute of Health</td>
</tr>
<tr>
<td>NJ</td>
<td>Neighbor-joining</td>
</tr>
<tr>
<td>nMOMP</td>
<td>Native MOMP</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>OD</td>
<td>Optical density</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate buffered saline</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDI</td>
<td>Protein disulfide isomerise</td>
</tr>
<tr>
<td>pi</td>
<td>Post-infection</td>
</tr>
<tr>
<td>PID</td>
<td>Pelvic inflammatory disease</td>
</tr>
<tr>
<td>Pmp</td>
<td>Polymorphic membrane protein</td>
</tr>
<tr>
<td>pt</td>
<td>Post-treatment</td>
</tr>
<tr>
<td>r</td>
<td>Recombinant</td>
</tr>
<tr>
<td>RB</td>
<td>Reticulate body</td>
</tr>
<tr>
<td>RBS</td>
<td>Ribosome binding sequence</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse transcription</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>Real-time quantitative reverse transcription PCR</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SE</td>
<td>Standard error</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SPG</td>
<td>Sucrose-phosphate-glutamate</td>
</tr>
<tr>
<td>STD</td>
<td>Sexually transmitted disease</td>
</tr>
<tr>
<td>STI</td>
<td>Sexually transmitted infection</td>
</tr>
</tbody>
</table>
Tarp  Translocated actin recruiting protein
Th   T helper
TN   Tamura-Nei
Ts   Transition rate
T3S  Type III secretion
T3SS Type III secretion system
Tv   Transversion rate
VD   Variable domain
ω    Distribution of the dN/dS ratio
WHO World Health Organization

Amino acid abbreviations

Ala (A) Alanine
Arg (R) Arginine
Asn (N) Asparagine
Asp (D) Aspartic acid
Cys (C) Cysteine
Gln (Q) Glutamine
Glu (E) Glutamic acid
Gly (G) Glycine
His (H) Histidine
Ile (I) Isoleucine
Leu (L) Leucine
Lys (K) Lysine
Met (M) Methionine
Phe (F) Phenylalanine
Pro (P) Proline
Ser (S) Serine
Thr (T) Threonine
Trp (W) Tryptophan
Tyr (Y) Tyrosine
Val (V) Valine
# TABLE OF CONTENTS

Acknowledgements / Agradecimentos................................................................. i
Notes of the Author .......................................................................................... iii
Summary .............................................................................................................. v
Resumo ............................................................................................................... vii
Thesis Outline ................................................................................................... ix
List of Abbreviations........................................................................................... xi
  General abbreviations ...................................................................................... xi
  Amino acid abbreviations ............................................................................... xiii
Table of Contents ............................................................................................. xv
List of Figures ................................................................................................... xix
List of Tables ..................................................................................................... xxi

## CHAPTER 1 ..................................................................................................... 1
1. General Introduction .................................................................................... 3
  1.1. *Chlamydia trachomatis* biology ............................................................ 3
  1.2. *C. trachomatis* pathobiologic diversity ............................................... 9
  1.3. *C. trachomatis* vaccine ....................................................................... 12
  1.4. Scope of the thesis ................................................................................ 12

## CHAPTER 2 .................................................................................................. 15
2. Normalization Strategies for Real-Time Expression Data in *C. trachomatis* .... 17
  2.1. Abstract ............................................................................................... 17
  2.2. Introduction ......................................................................................... 17
  2.3. Materials and Methods ....................................................................... 19
  2.4. Results ................................................................................................ 23
  2.5. Discussion ......................................................................................... 31

## CHAPTER 3 .................................................................................................. 35
3. Comparative Expression Profiling of the *C. trachomatis* pmp Gene Family for
   Clinical and Prototype Strains .................................................................... 37
  3.1. Abstract ............................................................................................... 37
  3.2. Introduction ......................................................................................... 37
3.3. Materials and Methods ................................................................. 39
3.4. Results ................................................................................ 44
3.5. Discussion........................................................................... 50

CHAPTER 4 .......................................................................................... 55

4. Lymphogranuloma Venereum in Portugal: Unusual Events and New Variants
During 2007.................................................................................. 57
4.1. Abstract.............................................................................. 57
4.2. Introduction ....................................................................... 57
4.3. Materials and Methods ...................................................... 58
4.4. Results ............................................................................... 59
4.5. Discussion ........................................................................... 61

CHAPTER 5 .......................................................................................... 65

5. Evolutionary Dynamics of \textit{ompA}, the Gene Encoding the \textit{C. trachomatis} Key
Antigen ........................................................................................... 67
5.1. Abstract.............................................................................. 67
5.2. Introduction........................................................................ 67
5.3. Materials and Methods ...................................................... 69
5.4. Results ............................................................................... 72
5.5. Discussion ........................................................................... 83

CHAPTER 6 .......................................................................................... 87

6. Adaptive Evolution of the \textit{C. trachomatis} Dominant Antigen Reveals Distinct
Evolutionary Scenarios for B- and T-cell Epitopes: Worldwide Survey ................. 89
6.1. Abstract.............................................................................. 89
6.2. Introduction....................................................................... 90
6.3. Material and Methods ........................................................ 91
6.4. Results ............................................................................... 95
6.5. Discussion ........................................................................... 104

CHAPTER 7 .......................................................................................... 109

7. \textit{C. trachomatis} Diversity Viewed as a Tissue-specific Coevolutionary Arms Race. 111
7.1. Abstract.............................................................................. 111
7.2. Introduction....................................................................... 111
7.3. Materials and Methods ........................................................ 113
7.4. Results ............................................................................... 116
7.5. Discussion ........................................................................... 125

CHAPTER 8 .......................................................................................... 131
LIST OF FIGURES

Figure 1.1 - Schematic illustration of the *C. trachomatis* developmental cycle ..................... 6

Figure 2.1 - Growth profile of C/TW3, E/Bour and L2/434 prototype strains ....................... 25

Figure 2.2 - Global view of the expression patterns of the 10 genes for *C. trachomatis* .......... 26

Figure 2.3 - Evaluation of expression stability of the 10 genes for C/TW3, E/Bour and L2/434 prototype strains ................................................................. 27

Figure 2.4 - RT-qPCR results of 16SrRNA gene by using RT target-specific priming ........... 30

Figure 3.1 - Expression profile of the nine *pmp* genes and *ompA* throughout the development of *C. trachomatis* ................................................................................ 45

Figure 3.2 - Expression profile of *pmp* and *ompA* genes throughout the development of *C. trachomatis* clinical strains ................................................................. 46

Figure 3.3 - Predicted *pmpF* promoter sequence for prototype and clinical strains .......... 48

Figure 3.4 - Distribution/Location of the putative RNase E cleavage sites within the *pmpFE* operon coding sequence .................................................................................. 49

Figure 3.5 - Dot-Blot of serum immunoreactivity against rPmpD and rPmpF ............... 50

Figure 4.1 - Partial *ompA* and MOMP sequences of the two L2 Portuguese 2007 variants ... 60

Figure 5.1 - Distribution of the 795 *C. trachomatis* clinical specimens by *ompA* genotype... 73

Figure 5.2 - Evolutionary relationship between all *C. trachomatis* prototype strains and the 43 types of genetic variants .................................................................................... 74

Figure 5.3 - Variability of *C. trachomatis* MOMP ................................................................ 78

Figure 5.4 - Number and type of substitutions involving a specific amino acid .................. 82

Figure 6.1 - *ompA* phylogeny of the 251 *C. trachomatis* genetic variants ......................... 96

Figure 6.2 - Worldwide mutational pattern of MOMP ......................................................... 98

Figure 6.3 - Topology sketch of MOMP mutational hotspots ............................................. 99

Figure 6.4 - Selective pressure acting on *ompA* ................................................................. 100

Figure 6.5 - Evolutionary dynamics of amino acid gains and losses .................................. 103

Figure 7.1 - *Loci* distribution in the ~1.04 Mb *C. trachomatis* circular chromosome ......... 117

Figure 7.2 - Phylogenomics of individual *loci versus* strain segregation ......................... 118
Figure 7.3 - Identification of loci domains characteristic of strains infecting a specific biological niche

Figure 7.4 - C. trachomatis evolutionary history

Figure 7.5 - Impact of indel events on C. trachomatis evolution and ecological success

Figure 8.1 – Schematic diagram presenting the hypothetical evolutionary scenario of C. trachomatis

Figure 8.2 - Putative clonal evolution of the two most prevalent C. trachomatis genotypes (E and F)

Figure S7.1 - Overall mean genetic distances among the 15 main C. trachomatis genotypes for the 51 loci

Figure S7.2 - Pathogen evolutionary history by loci category

Figure S.1 - MOMP B- and T-cell core epitopes mapped for all prototype strains representing the C. trachomatis genotypes

Figure S.2 - Putative chromosomal mosaic structures for the 29 recombinant C. trachomatis clinical isolates
LIST OF TABLES

Table 2.1. Gene expression stability evaluation throughout normal growth cycle..........29

Table 2.2. Gene expression stability evaluation throughout growth cycle under D-cycloserine treatment...................................................................................................................................31

Table 2.3. Top-ranked genes according to geNorm and Normfinder applications........31

Table 3.1. Oligonucleotide primers used for RT-qPCR ................................................41

Table 3.2. Clinical and microbiologic characteristics of female adolescents from whom sera were used for determining the immunoreactivity against rPmpD and rPmpF .........................49

Table 4.1. LGV infection cases detected in the Portuguese NIH during 2007...........59

Table 5.1. Nucleotide sequence variation in \(ompA\) genotype variants compared with the respective prototype strain.................................................................................................................................76

Table 5.2. Likelihood ratio test statistics for testing positive selection.........................80

Table 6.1. Description of the 17 baseline prototype strains representing the \(C. trachomatis\) genotypes.................................................................................................................................92

Table 7.1. Contingency table for estimating polymorphism significance.......................114

Table 7.2. Genetic polymorphism for the concatenated sequences.............................120

Table 7.3. Overall mean genetic distances within and between disease groups............122

Table 7.4. Polymorphic \(loci\) among strains that infect the same biological niche..........123

Table S2.1. Oligonucleotide primers used in qPCR assays ........................................153

Table S5.1. Description of all prototype strains as well as of one representative of each type of genetic variant found among the 232 Portuguese variant clinical specimens........154

Table S5.2. PAML results for genotype G .....................................................................156

Table S5.3. PAML results for genotype Ia .....................................................................156

Table S5.4. PAML results for genotype Da.................................................................157

Table S7.1. Cellular role of the studied 51 \(loci\) ............................................................158

Table S7.2. Primers used for PCR and sequencing .....................................................160

Table S7.3. Deletion events among the 15 main \(C. trachomatis\) prototype strains.........162
CHAPTER 1

General Introduction
1. General Introduction

1.1. Chlamydia trachomatis (C. trachomatis) biology

*Chlamydiae* are obligate intracellular bacteria characterized by a highly specialized biphasic developmental cycle that is unique among prokaryotes. These parasites infect a diverse array of vertebrates (including humans, birds, ruminants and amphibians) and belong to the family *Chlamydiaceae*, which encompass the genera *Chlamydia* and *Chlamydophila* [1]. The later comprises the species *C. psittaci, C. abortus, C. caviae, C. felis, C. pecorum*, and *C. pneumoniae*, while the former contains the species, *C. trachomatis, C. muridarum*, and *C. suis*.

As other *Chlamydiae*, the human pathogen *C. trachomatis* has undergone massive genome reduction after acquiring an intracellular lifestyle, resulting in the disappearance of several biosynthesis pathways. Consequently, this pathogen has one of the smallest genomes among bacteria [~1.04 million base pairs (bp)], corresponding to just about a quarter of the size of the genome of common free-living bacteria such as *Bacillus subtilis* or *Escherichia coli*. Interestingly, of its ~900 open reading frames (ORFs), ~30% have no homology with other bacterial proteins. *C. trachomatis* also harbours a 7.5 kbp plasmid, whose function is still unknown [2].

This pathogen can be divided into 18 serovars or genotypes, based on the differential immunoreactivity of its major outer membrane protein (MOMP) or polymorphism of *ompA* (that encodes MOMP) [3], respectively, which match 100%. Genotypes A-C and Ba usually cause ocular infections, genotypes D-K, Da, Ia and Ja are normally associated with ano-urogenital infections (with E and F representing ~50% of them), while genotypes L1-L3 are the causative agents of lymphogranuloma venereum (LGV). Genotypes can be further grouped into two distinct biological variants (biovars) based on the infected cell-type and on their *in vitro* infection properties: trachoma biovar (genotypes A-K, Ba, Da, Ia and Ja) is noninvasive, is mucosae-restricted and epitheliotropic, whereas LGV biovar is invasive and infects primarily lymphatic tissue, presenting faster and more vigorous growth in cell culture [4].

Traditionally, *C. trachomatis* genotypes are represented by 18 prototype strains, which have been used as ‘study-models’ worldwide since their isolation ~60 years ago. So far, only genomes of prototype strains from genotypes A [5], B [6], D [2], and L2 [7] were fully-sequenced.
1.1.1. Diseases

Infections caused by *C. trachomatis* genotypes are a serious public health problem worldwide. Ocular infections can result in trachoma, a chronic inflammatory disease that affects 300-500 million people and leads to conjunctival scarring and blindness, especially in resource-poor nations [8-12]. Recently, this world’s leading cause of preventable infectious blindness [13] was recognized as one of the seven major neglected diseases of the 21st century [14], and it is currently the target of a World Health Organization (WHO) eradication campaign to the year 2020 (http://www.who.int/pbd/publications/trachoma/en/get_jan1998.pdf). *C. trachomatis* ano-urogenital infections are the primary cause of bacterial sexually transmitted diseases (STDs) in both industrialized and developing countries, causing cervicitis and urethritis [15]. The asymptomatic character of many of these infections (~75% in women and ~50% in men) often promotes their progression to the upper genital tract, causing epididymitis in men and pelvic inflammatory disease (PID) in women, which can lead to significant long-term sequelae, such as infertility and ectopic pregnancy [16-19]. *C. trachomatis* is also responsible for more-invasive and systemic diseases, such as LGV, through dissemination via infection of macrophages to regional draining lymph nodes, where it establish a chronic granulomatosis disease [20]. After a recent outbreak in Europe and the United States, LGV has been emerging as a significant problem among men who have sex with men (MSM) [21-23]. Surprisingly, despite aggressive antibacterial control measures over the past decade, the prevalence of *C. trachomatis* STDs has been increasing [19], with an estimated 92 million new cases occurring worldwide each year (http://www.who.int/), constituting a enormous morbidity and socioeconomic burden. This gains even greater public health impact as these infections may also increase the risk of acquiring or transmitting human immunodeficiency virus (HIV) [24,25] and hepatitis C [26] as well as of developing invasive cervical carcinoma induced by human papilloma virus (HPV) [27,28].

1.1.2. Morphology

*Chlamydiae* exist as two morphologically distinct forms: a metabolically inert infectious elementary body (EB) and a metabolically active replicative but noninfectious reticulate body (RB) [29]. The former is small (~0.3 μm) and round, contains a central and dense nucleoid, and functions as a ‘spore-like’ form with a rigid cell wall that permits chlamydial survival in extracellular environment [30-34]. As EB lacks a detectable peptidoglycan layer, its structural rigidity is thought to result from a mesh of disulfide cross-linked cysteine-rich outer membrane proteins [35,36]. Enigmatically, *Chlamydiae* are susceptible to drugs that inhibit
peptidoglycan synthesis (such as penicillin and D-cycloserine) and possess a complete set of genes for the peptidoglycan biosynthesis [2,5-7,37] that are expressed in the metabolically active stage of the chlamydial cycle [38] and are thought to be involved in the bacterial cell division [39]. RBs are larger than EBs (~1 μm), structurally flexible and osmotically fragile (due to the paucity of cross-linked membrane proteins), richer in RNA and contain diffuse and fibrillar DNA. These properties enable intracellular replication, uptake and transport of nutrients, protein synthesis as well as several other metabolic activities [32-34].

Both EB and RB cell walls encompass two sets of trilaminar membranes: an inner cytoplasmic membrane and an outer membrane [4,40], whose surface is covered with projections (especially in RBs) that are suspected to be type III secretion system (T3SS) ‘needle’ structures similar to those described for Salmonella spp., Yersinia spp. and Shigella spp. [41-44]. T3SS is a virulence mechanism used by several pathogenic Gram-negative bacteria to translocate effector proteins directly into host cell cytoplasm to manipulate specific cellular functions at distinct stages of infection [45]. It was already reported that Chlamydiae express an active T3SS throughout the infectious cycle, although the identity of the complete set of chlamydial type III secretion (T3S) effectors, which is predicted to encompass >10% of the bacterial chromosome, is still elusive [43,46,47]. Nevertheless, several studies [43,46,48,49] have been showing that this arsenal may be required for efficient cell invasion and inhibition of phagocytosis, establishment of the inclusion, acquisition of nutrients, modulation of intracellular trafficking, early inhibition and late induction of apoptosis, and avoidance of innate immune responses.

1.1.3. Developmental cycle

Chlamydiae undergo a unique biphasic developmental cycle that may be divided into five major steps (Fig. 1.1): i) EB attachment and entry into the host cell yielding the formation of an inclusion/vacuole; ii) EB to RB differentiation inside the inclusion; iii) inclusion modification and bacterial replication; iv) inclusion expansion and RB to EB transition; and v) bacterial release [32,33,40].
1.1.3.1. Attachment and entry

Chlamydial infection begins with the attachment of EBs to the surface of susceptible host cells in a two-stage process: an initial electrostatic-mediated chlamydial contact with host microvilli [50,51], followed by an irreversible secondary binding to unidentified host receptors that culminate with EB internalization [52,53]. A number of chlamydial ligands, including MOMP [54,55], glycosaminoglycans [56,57], heat shock protein 70 [58], OmcB [59,60], and PmpD [61,62] have been implicated in this process. Although the detailed mechanisms for EB attachment and entry are still unknown, it was recently demonstrated that a host multi-functional factor, the protein disulfide isomerase (PDI), is structurally required for EB attachment, while its enzymatic activity is necessary for EB entry through targeting multiple bacterial and host surface proteins [63,64]. PDI may be essential to activate the translocation of several T3S effectors into host cells, including the translocated actin recruiting protein (Tarp), which was found to play a central role in the remodeling of the host cytoskeleton at attachment sites to facilitate bacterial engulfment [65-69]. In addition, it has...
been proposed that a functional clathrin-mediated pathway may also participate in *C. trachomatis* entry and may operate together with the Tarp-based mechanism, where Tarp may alternatively mediate clathrin coat formation [70].

1.1.3.2. Inclusion modification and intracellular growth

Shortly after uptake, EB starts to differentiate into RB, while the RB-containing vacuole is rapidly modified to a membrane-bound inclusion, resulting in trafficking of the inclusion to the host perinuclear region and avoidance of phagolysosomal fusion [71]. RBs then undergo repeated cycles of binary fission, synthesizing their own DNA, RNA and proteins while obtain energy, nutrients and biosynthetic precursors from the host cell. As the chlamydial inclusion membrane is impervious to the diffusion of molecules >520 Da [72], *Chlamydiae* are adept at acquiring host metabolites through selective interaction with host organelles and subversion of membrane trafficking [49]. For instance, it has been showed that chlamydiaal recruitment of host-derived lipids, including cholesterol [73], sphingolipids [74,75], and glycerophospholipids [76,77] results from interception with Golgi-derived exocytic vesicles [74,78] and multivesicular bodies (MVBs) [79], or direct uptake of lipid droplets (LDs) [80]. Cumulative evidences indicate that chlamydial T3S-secreted inclusion membrane proteins (Incs) might have a central role in orchestrating inclusion modification through recruitment of key regulators of host cell membrane trafficking to the inclusion membrane, hence rerouting host cell vesicular trafficking and modulating inclusion fusion events [81]. During this process, chlamydial inclusion grows to accommodate the increasing number of RBs. Roughly midway through infection, chlamydial replication becomes asynchronous as RBs begin redifferentiating into EBs (by remodelling bacterial outer membrane and condensation of chromosome), which accumulate within the inclusion while the remainder RBs continue to multiply. Meanwhile, effector proteins required for both bacterial release and infection of new cells are assembled and preloaded onto EBs [43].

1.1.3.3. Persistence

Under adverse growth conditions, *Chlamydiae* may also enter in a persistent phase, where the development is disrupted, resulting in the appearance of enlarged and morphologically aberrant RBs [82,83]. During this phase, bacterial cell division and differentiation to EBs are inhibited, despite DNA replication continues, and elevated levels of stress-response proteins are expressed. Persistent infection can be induced by nutrient deprivation (such as iron [84] and amino acids [85,86]), antibiotics [87,88], interferon-gamma (IFN-γ) exposure,
temperature [89], infection of monocytes [90], or concomitant herpes infection [91]. Although the *in vitro* persistence is considered reversible upon removal of the stressful factor, aggressive inflammatory responses to repeated and persistent infections *in vivo* are thought to initiate irreversible pathogenic events (including promotion of cellular proliferation, tissue remodelling and scarring) that ultimately lead to debilitating chronic sequelae of blinding trachoma and tubal infertility [92,93].

1.1.3.4. Host cell exit
At about 48-72 h post-infection (pi), the expanded inclusion ultimately occupies almost of the entire host cell cytoplasm, and *Chlamydiae* are released by two mutually exclusive independent mechanisms: host cell lysis and extrusion. The former is a destructive process that consists of sequential rupture of bacterial inclusion, host nucleous and cell membranes, resulting in host cell death. In contrast, extrusion is a slow process dependent on actin polymerization, in which a portion of the inclusion is released by a membranous protrusion, leaving both the inclusion and host cell intact, although left-over inclusion remnants remain inside the host cell [94], which may contribute for persistent infection. The released multitude chlamydial infectious forms will then infect neighboring host cells or new hosts, initiating subsequent rounds of infection.

1.1.3.5. Evasion of host surveillance
Considering the unique biphasic life-cycle of *Chlamydiae*, these bacteria must be able to overcome various sets of host immune defence during both the extracellular and intracellular phases of their development in order to survive: *i*) before invading the host cells, extracellular EBs must be able to avoid attack by host pre-existing humoral and cellular effector mechanisms (such as neutralizing antibodies (Abs) and antigen-presenting cells), and *ii*) during infection, *Chlamydiae* have not only to block the early host phagolysosomal action but also to maintain the integrity and viability of the infected cells until complete their development. It has been suggested that the ability of chlamydial inclusion to block lysosomal fusion [95] may result from the presence of a chlamydial analogue of EEA1 (human endosomal antigen 1) in the inclusion membrane [96], as it lacks the binding domain that in the human protein is essential to regulate endosomal fusion [97]. Furthermore, besides the apparent protective shielding of the inclusion itself, it is believed that several chlamydial evasion strategies may simultaneously operate through hijacking certain signaling pathways of the innate and adaptive host immunity systems [98,99]. Indeed, cumulative evidences have
been showing that the early chlamydial antiapoptotic activity encompass the inhibition of mitochondrial cytochrome c release and downstream caspase activation [100], stabilization of inhibitor of apoptosis proteins (IAPs) [101], sequestration of proapoptotic proteins [102], and inhibition of transcription nuclear factor (NF)-kB activation pathway [103]. These antiapoptotic activities are prolonged during chlamydial persistence [104]. Interestingly, apoptosis also seems to have an immunomodulatory role in chlamydial infections as *C. trachomatis* appear to induce T-cell apoptosis by paracrine effects [104], which might suppress host immune responses.

Alternatively, *Chlamydiae* may also avoid host immune recognition through degradation of specific host transcription factors required for IFN-γ-induced major histocompatibility complex (MHC) class I and II expression and lipid antigen presentation, both by secreting the protease-like activity factor (CPAF) [105,106].

### 1.2. *C. trachomatis* pathobiologic diversity

*In vivo*, *C. trachomatis* genotypes present biological uniqueness, exhibiting significant differences in tissue tropism, pathogenesis and ecological success. To date, relatively little is known about the molecular factors underlying pathobiologic specificities of genotypes. However, the colossal genomic similarity (~99%) observed among the few currently fully-sequenced *C. trachomatis* genomes [2,5-7] suggests that the polymorphism of a relatively small number of loci likely determine *C. trachomatis* diversity.

Relevant contributions to explain the colonization of a particular tissue (eyes/genitalia) or cell-type (epithelial/lymph cells) came from the putative virulence factors type III effector Tarp [2,5-7], cytotoxin gene [107], and mainly trpRBA operon [108,109]. For instance, the presence of twice more tyrosine-rich tandem repeats within the Tarp N-terminus of LGV genotypes than that of noninvasive genotypes may function as a putative virulence mechanism that influence the colonizing capacity of the former [5]. In fact, as tyrosine phosphorylation of this domain is critical for chlamydial entry into host cells [110], the putative increased LGV Tarp phosphorylation may result in a more efficient entry process, and consequently in an enhanced ability to avoid the early host phagolysosomal action by rapid modification of chlamydial inclusion membrane.

Also, it was found that only epithelial-genital genotypes encode an intact cytotoxin that contains both the functional glycosiltransferase and UDP-glucose binding domains, required for enzymatic activity of well-known bacterial toxin homologs [107,111]. The ocular genotypes (except B) encode only the UDP-glucose binding domain, while the LGV
genotypes have both domains entirely deleted [107]. As chlamydial cytotoxic activity appears to be correlated with a high multiplicity of infection, this putative virulence factor may allow epithelial-genital genotypes to circumvent host immune system by inactivating specific IFN-γ-mediated antimicrobial effectors at attachment sites, which may lead to disruption of host cytoskeleton and alteration of early vesicular trafficking [111].

Moreover, it was recently demonstrated that a fully active trpRBA operon is a mandatory condition for any strain to infect the genital but not the ocular tract [108]. Indeed, only genital-infecting strains retain the ability to encode a functional Trp synthase [109] that catalyzes the final steps in the biosynthesis of Trp from endogenous sources such as indole, which is produced by other microorganisms existing solely on genital flora. This ability is particularly important in an IFN-γ-rich environment (that is a result of a anti-chlamydial immune response), where an IFN-γ-induced tryptophan-catabolizing enzyme (indoleamine-2,3-dioxygenase, IDO) depletes intracellular tryptophan pools, thereby starving the pathogen from this essential amino acid [108]. Consequently, both the bacterial replication and secondary differentiation are inhibited, which may induce a persistent phase. Thus, the ability to synthesize tryptophan from endogenous indole, produced in a mixed microbial environment, is therefore a potential virulence factor for genital-infecting strains, allowing them to avoid host IFN-γ-mediated eradication.

However, none of these putative virulence factors fully explain the existence of three major tropism groups (eyes, epithelial-genitalia, and lymph nodes) neither the distinct invasiveness (mucosotropic and lymphotropic) nor the ecological success of genotypes. This becomes even more complex as, in contrast to the LGV strains, the preference for infecting mucosae-epithelial cells is not exclusive, and therefore ocular strains can occasionally be found in the urogenital tract and vice-versa [108]. Moreover, several studies [112-116] have been showing that, despite of its intracellular condition, this pathogen may undergo intragenic recombination and even extensive interstrain genetic exchanges (during mixed infections) involving polymorphic and immunogenic loci, which indicate some pathogen plasticity to deal with host niche-specific environmental changes in order to survive. Some of these loci belong to the nine-member polymorphic membrane protein gene family (pmpA to pmpI).

1.2.1. Pmps

Over the last decade, Pmps have been intense research targets as they are unique to Chlamydiaceae and encompass ~3.2% of the C. trachomatis chromosome (13.6% of the specific coding capacity) [2], which, in the light of the pathogen’s reductive evolution,
suggests their importance in the chlamydial biology. Supporting this, a recent comparative genomic analysis between prototype strains from genotypes A and D [5] has shown that 20% of the total single nucleotide polymorphisms (SNPs) observed within coding regions derived from the nine \textit{pmp} genes.

In \textit{C. trachomatis}, this superfamily was found to encompass three distinct chromosomal regions of two gene clusters (\textit{pmpA} to \textit{pmpC} and \textit{pmpE} to \textit{pmpI}) and the genetically isolated \textit{pmpD} [2]. Pmp grouping as a family relies on the presence of multiple repeated GGAI (glycine-glycine-alanine-isoleucine) motifs in their N-terminal half, which so far were only seen in three nonchlamydial proteins, where these motifs are involved in adherence processes [117,118]. In support of a putative adhesion role for Pmps, cumulative studies have been showing that these extra-large proteins (96 to 187 kDa) [117] are surface-exposed [119,120], presenting structural features of outer membrane proteins, such as a cleavable \textit{sec}-dependent N-terminal leader (except for PmpA), a C-terminal phenylalanine (Phe) residue, and a Trp-rich C-terminus [121]. Plus, both \textit{in silico} [122] and experimental [61,123-125] evidences imply Pmps as autotransporters that mediate the translocation of the N-terminus to the bacterial surface through beta-barrel folding of the C-terminal domain. Also, the polymorphism exhibited by some Pmps was found to phylogenetically segregate genotypes according to their tissue tropism (PmpB, PmpC, PmpF, PmpG, PmpH and PmpI), to be concentrated at distinct protein domains exclusive of genotypes infecting a specific tissue (PmpE, PmpF and PmpH), or even to co-segregate the two most ecological succeeded genotypes E and F distantly from the remaining urogenital genotypes (PmpB, PmpC, PmpD, PmpH and PmpI) [113,121]. At least for PmpD, adhesin function was already experimentally demonstrated, since anti-PmpD Abs neutralized chlamydial infectivity \textit{in vitro} [62]. Furthermore, Pmps were found to be immunogenic when using sera from \textit{C. trachomatis} infected patients [126,127], which imply Pmps as antigenic targets of host immune system, holding potential as vaccine candidates. Supporting this, PmpD was shown to generate species-specific neutralizing Abs against \textit{C. trachomatis} genotypes [62] as well as to induce protective \textit{C. trachomatis}-specific CD4+ T-cell responses [128]. Also, PmpF was shown to contain numerous SNPs located within predicted HLA (human leukocyte antigen) class I and II T-cell epitopes of the N-terminal domain [5]. Recently, both the PmpF and PmpG orthologs in \textit{C. muridarum} were found to be immunodominant, possessing a N-terminal MHC class II epitope highly conserved between \textit{C. trachomatis} genotypes and \textit{C. muridarum} [129] that is able to induce significant protective immunity against lung and genital tract infections in animal models [130].
Although much more work still needs to be done to completely understand the nature, function and processing of Pmps, all the exposed findings suggest that Pmps may be important virulence factors with a dual role of promoting niche-specific pathogenesis-related adhesion function while providing antigenic diversity for chlamydial host immune evasion.

1.3. *C. trachomatis* vaccine

It is believed that the effective management of human chlamydial diseases will likely require the development of an efficacious prophylactic or therapeutic vaccine. Over the last two decades, there has been a gradual shift from the use of inactivated or attenuated intact pathogens (due to the existence of immunopathogenic components) to peptide or subunit vaccines [131]. However, despite of the intensive efforts, attempts to develop a chlamydial vaccine have been unsuccessful to date, conferring only a partial short-lived genotype-specific protective immunity [132].

Nevertheless, contemporary vaccine research continues to seek not only potent adjuvants and effective delivery vehicles but also a stable optimal antigen cocktail able to strongly elicit, at different stages of cycle, both the humoral and cellular immune responses required for a vaccine to confer long-lasting multi genotype-broad protection. The major chlamydial membrane component MOMP [133] is regarded as one of the leading candidates as it is immunodominant, immunoaccessible, neutralizing, highly variable among genotypes, and possesses species and genotype-specific epitopes that elicit both humoral and cellular immune responses [134-140]. However, it has been suggested that the success of a MOMP-based vaccine may depend on the presence of other antigens with neutralizing effects [132,141,142], such as the promising Pmps.

1.4. Scope of the thesis

It is assumed that a detailed understanding of the molecular factors that determine the differential pathogenesis of *C. trachomatis* genotypes, in particular their tissue tropism, virulence and immune evasion strategy, is vital for deciphering the basis of their ecological success as well as for defining and evaluating protective antichlamydial immune responses to foster vaccine design.

Contrarily to other microorganisms, the progress in *C. trachomatis* research has been clearly limited by hurdles related to its obligate intracellular development (in particular, its asynchronous life-cycle and its inability to grow in a host-free environment), its genetically nontractable character, as well as the inexistence of suitable animal and in vitro models that
mirror the chlamydial infection \textit{in vivo}. Therefore, due to the inapplicability of many of the traditional research methodologies used for other bacteria, genomics, transcriptomics and bioinformatics gain an enhanced relevance in \textit{Chlamydiae}, being crucial tools for identifying \textit{loci}, whose genetic and transcriptomic polymorphisms may get further insights in deciphering the molecular basis of the pathobiologic diversity and ecological success of \textit{C. trachomatis}.

In order to contribute for this knowledge, the scope of the present Ph.D. thesis was to evaluate the genomic and transcriptomic features of several polymorphic \textit{loci}, given especial relevance to the nine \textit{pmp} family and the chlamydial key antigen (MOMP) as they are predicted to play an important role in \textit{C. trachomatis} biological diversity. In particular, this thesis intended to scrutinize the putative biological role of the nine \textit{pmp} family, to provide a deeper understanding of the impact of the host pressure on MOMP, to study the evolution of the \textit{C. trachomatis} genotypes by analyzing a panel of heterogeneous \textit{loci} scattered throughout the chromosome, and ultimately to outline the putative molecular aspects that may underline genotype ecological success. To achieve robust and reliable conclusions, both the traditional prototype strains (‘study-models’) and current clinical isolates were used in parallel whenever it was logical, since the later are not laboratory-adapted and reflect the immune selection that results from the infection of several hosts. Therefore, the following detailed objectives were pursued and constituted the subject of six independent chapters that can be read separately:

\(i\) to evaluate several normalization strategies for validation of real-time expression data in \textit{C. trachomatis} (CHAPTER 2);

\(ii\) to determine the expression profile of the nine-member \textit{pmp} gene family throughout development for prototype and clinical strains representing the two chlamydial biovars, as well as to evaluate their immunoreactivities (CHAPTER 3);

\(iii\) to perform a genetic characterization of the chlamydial major antigen (MOMP) of the strains isolated in Portugal on behalf of the recent LGV outbreak observed worldwide (CHAPTER 4);

\(iv\) to perform an evolutionary mutational trend analysis of MOMP using all variant strains reported to date in Portugal (CHAPTER 5) as well as in the rest of the globe (CHAPTER 6);

\(v\) and finally, to perform a high-scale evolutionary concatenation-based phylogenomic survey that comprised about one third of all chromosome SNPs of the 15 main \textit{C. trachomatis} genotypes (CHAPTER 7), in order to understand the molecular evolution behind the genotype radiation.
In the end, the major findings of these six studies are discussed in a more general context (CHAPTER 8), taking into account the scope of this thesis.
CHAPTER 2

Normalization Strategies for Real-Time Expression Data in C. trachomatis

Published in

Author Contributions
MJB and JPG conceived and designed the experiments; VB, RF, AN, MJB and JPG performed the experiments; PJN performed the statistics; VB, RF, AN and JPG analyzed the data; VB, RF and JPG wrote the paper; AN and MJB reviewed the paper.
2. Normalization Strategies for Real-Time Expression Data in C. trachomatis

2.1. Abstract

C. trachomatis is a widespread obligate intracellular pathogen genetically nontractable for which transcriptomics is a fundamental tool to better understand its biology. However, the suitability of endogenous controls for normalization of transcriptomic data in this bacterium still needs validation. We aimed to assess the stability of 10 genes for their potential use as endogenous controls in real-time quantitative polymerase chain reaction (PCR) assays at both normal and stress (D-cycloserine treatment) growth conditions throughout the developmental cycle of three C. trachomatis strains with different tissue tropism. Normalization was performed by real-time absolute quantification of the bacterial genomes. We also tested the applicability of two widely used softwares (geNorm and Normfinder) to our data. For all strains, we found that 16SrRNA was the most stably expressed gene throughout the chlamydial normal developmental cycle, which indicates its potential use as endogenous control in relative expression assays. However, it was highly unstable under D-cycloserine treatment (where oppA_2 was top-ranked), suggesting prudence when using ribosomal genes in expression experiments involving stress conditions. The geNorm and Normfinder algorithms revealed contrasting results and seem inappropriate for the selected pool of genes. Considering the multiplicity of experimental conditions, there should be an in loco validation of endogenous controls, where 16SrRNA gene appears to be in the front line. Alternatively, normalization of expression data against genomic DNA, which is less influenced by experimental constraints that are especially relevant for intracellular organisms, likely constitutes a good option. Moreover, the number of genomes also seems to be less subject to variation than expression of endogenous controls when working under stress conditions. The present study constitutes the first evaluation of putative endogenous controls for real-time expression assays in C. trachomatis.

2.2. Introduction

C. trachomatis is a widespread obligate intracellular bacterium, where genotypes A–C are the causative agents of trachoma, genotypes D–K infect the urogenital tract and the invasive genotypes L1–L3 cause LGV. It has a unique biphasic developmental cycle (~40–45 h) that alternates between an infectious form (the EB) and a metabolically active form (the RB)
Normalization of C. trachomatis Expression

which replicates within a host cell inclusion. The normal developmental cycle can be disturbed by several stress conditions such as nutrient starvation, temperature, host immune response or antibiotic treatment, which are known to affect transcription [143,144].

Gene expression in Chlamydia has been mostly performed using real-time quantitative reverse transcription PCR (RT-qPCR), which presents well-known advantages over traditional mRNA quantification methods [145,146]. However, the requirement of a proper normalization strategy is probably the major problem when performing relative expression assays, where inappropriate methodologies can lead to inaccurate data and incorrect conclusions [147,148]. To date, numerous strategies for RT-qPCR data normalization have been applied, such as the use of the total RNA mass [148], the number of cells determined by quantitative culture [149] and the use of an external reference of known amount added to the cultures prior to sample processing [146,150]. Besides the inherent and well-known disadvantages of these three approaches, additional drawbacks limit their application to obligate intracellular organisms like C. trachomatis: i) the total RNA that is extracted is composed not only by variable proportions of bacterial rRNA, tRNA, mRNA, small non-coding RNA but also by eukaryotic RNA [150]; ii) quantitative cultures have low sensitivity and are difficult to reproduce [145,151]; iii) the amount of external reference that enters or stays out the host cell cannot be controlled, hampering the determination of bacteria-specific mRNA.

The most commonly employed method involves the use of mRNA of an endogenous control, usually a housekeeping gene (HK), to normalize the target mRNA ensuring that both mRNAs are subjected to exactly the same processing procedure. Thus, differences associated with input RNA amount and variable RT efficiencies are minimized [146,152]. Nevertheless, it is mandatory that the mRNA levels of the endogenous control gene do not vary significantly throughout the biological process under study (cellular growth stages, stress conditions, etc.) [152]. In fact, remarkable changes in the expression levels of HKs were observed for both eukaryotic [147,153] and prokaryotic organisms [149,154] and for that reason, various attempts have been made in order to validate endogenous control genes [155-158]. Validation of the suitability of commonly used endogenous controls, such as 16SrRNA, for normalization in prokaryotes is still lacking [146], in particular for intracellular microorganisms for which the traditional protocols [159] cannot be applied. Some software applications have been developed to rank genes according to their stability based on the mathematical transformation of RT-qPCR raw data, allowing the use of more than one gene for normalization [160-162].

18
Considering the lack of a validated method for selecting genes as endogenous controls in *C. trachomatis*, we aimed to assess the stability of 10 chromosome-spread genes that revealed an apparent stability during development through microarray analysis [96]. These genes are not likely metabolically related to each other and include a ribosomal RNA gene (16SrRNA), housekeeping genes (*yraL, fer, oppA_2, radA, yaeI, hemN_2, tyrP_2* and *map*), and CT147 that codes for a hypothetical “early endosomal antigen 1 homologue”. To achieve this, the number of chlamydial genomes, determined by cloning-based real-time absolute quantification, was used as the normalizing factor of gene expression. We tested three strains presenting distinct tissue tropism and growth rate. The gene expression stability was assessed in normal growth and under stress condition induced by D-cycloserine, an antibiotic that is known to interfere with bacterial cell wall synthesis, preventing RB differentiation into EB [163,164]. We further evaluated the accuracy of using statistical algorithms to validate endogenous control genes in this intracellular pathogen. Finally, we discuss the reliability of normalizing gene expression against the number of bacterial genomes in opposition to using the experimentally and/or mathematically determined top-ranked endogenous control genes.

### 2.3. Materials and methods

#### 2.3.1. Preliminary assays

**2.3.1.1. Preliminary assay 1: Evaluation of D-cycloserine toxicity**

In order to determine the D-cycloserine concentration to be used in the expression assays under stress conditions, we first evaluated the cytotoxic effect of this antibiotic on cell monolayers and *C. trachomatis*. Therefore, the old enriched medium in the T25 cm² flasks containing HeLa 229 cells not infected and infected with this bacterium (see below details of chlamydial culture) was replaced by fresh enriched medium with a concentration range of 0.001 μg/ml to 200 μg/ml of D-cycloserine at 18 h (pi). The morphologic alterations of HeLa 229 cells and *C. trachomatis* inclusions were then examined by phase-contrast microscopy every hour post-treatment (pt), in order to select the antibiotic concentration that induces the abnormal RB development without interfering with normal growth of HeLa 229 cells.

**2.3.1.2. Preliminary assay 2: Evaluation of nucleic acid extraction reproducibility**

Considering that the efficiency of the nucleic acids extraction methods may affect the final gene expression results, we assessed the reproducibility of both the DNA and RNA extractions. Thus, three groups of five twin samples (each group collected from *C. trachomatis* cultures containing different amounts of biological material) were submitted to
Normalization of *C. trachomatis* Expression

independent DNA and RNA extraction using QIAamp® DNA Mini Kit (Qiagen) and RNeasy Mini Kit (Qiagen), respectively, according to manufacturer's instructions. After elution, the DNA and RNA yields were determined by optical density (OD) measured at $A_{260}$ nm. The reproducibility of both extraction methods was statistically evaluated by calculating the variation coefficient within each group of five samples, and then by the determination of the mean value of the three independent assays.

2.3.1.3. Preliminary assay 3: Evaluation of RT efficiency using random hexamers versus target-specific primer

In order to evaluate if different RT methodologies influence gene expression data, we assessed if the levels of cDNA obtained by using random hexamers or a target-specific primer reflect the initial quantities of RNA. This assay was performed solely for 16SrRNA as it was the most stably expressed gene (see results for details). cDNA was generated from four-fold serial dilutions of the same RNA, ranging from about 50 pg to 50 ng of total RNA. Subsequently, 5 μl of each cDNA were used in qPCR reaction. Both RT and qPCR reaction conditions are described below. To evaluate the performance of both assays we calculated the respective sum of the square residuals calculated for each concentration. Assuming that each residual approximates to the standard normal distribution, their sum was considered to follow a chi-square distribution enabling the probability calculation of the observed deviation values being equal to zero.

2.3.2. Chlamydial culture and harvesting

The *C. trachomatis* prototype strains used in the present study represent the three disease groups: C/TW3 (ocular), E/Bour (urogenital) and L2/434 (LGV). Cell culture of each strain was performed as previously described [165]. Briefly, eight T25 cm² flasks of confluent HeLa 229 cell monolayer were inoculated with each strain by centrifuging for 1 h at 34 °C and subsequent incubation for 1 h at 37 °C in 5% CO₂. Cell medium was then replaced by enriched medium (MEM with 10% fetal bovine serum, vitamins, nonessential amino acids, glucose and 0.5 μg/ml cycloheximide) and cultures were allowed to grow at 37 °C in 5% CO₂. For *C. trachomatis* growth under normal conditions, five T25 cm² flasks were used and the developmental cycle was interrupted at time-points 4, 12, 20, 30 and 42 h (pi) for RNA and DNA extraction. The remaining three flasks were used for D-cycloserine treatment, which was added at a concentration of 30 μg/ml (see results) at 18 h (pi), and disruption of the developmental cycle was performed at 6, 16 and 36 h (pt). All flasks were subjected to the
same cell harvesting protocol, as previously described [166]: medium was removed and cells were scraped with 1400 μl of phosphate buffered saline (PBS) at 4 °C; for disruption of eukaryotic cells and bacterial release, the suspension was sonicated for 30 s and the harvested cells were submitted to low-speed centrifugation (1500 rpm for 2 min) at 4 °C; finally, the supernatant was collected, homogenized and rigorously divided into two identical aliquots. One of these was stored at −20 °C until DNA extraction and the other aliquot was subjected to immediate RNA extraction due to its very short half-life. The use of two aliquots is a mandatory step as RNA quantification demands exclusion of any contaminant DNA and, on the other hand, DNA extraction procedure does not protect RNA. To ensure accuracy, three biological replicates of each strain were used.

2.3.3. Nucleic acid extraction

From each time-point, extraction of total RNA was carried out as previously described [127]. Basically, lysates were centrifuged at 14,000 rpm for 10 min (at 4 °C) to collect the Chlamydia-containing pellet which was subsequently resuspended in lysozyme-containing TE buffer. Then, RNeasy Mini Kit (Qiagen) was used according to the package inserts. Since double-stranded DNA is more effectively amplifiable than the single stranded cDNA that is generated from RNA, an on-column DNA digestion using 30 U of RNase-free DNase (Qiagen) was performed to ensure the complete removal of residual contaminant DNA. In fact, DNase treatment has proven to be highly effective by removing >98% of DNA [149]. RNA was eluted in 50 μl of RNase-free water and stored at −80 °C in RNase-free tubes after A_{260} nm quantification.

For DNA extraction, the respective −20 °C aliquot was centrifuged at 14,000 rpm for 10 min (at 4 °C), the pellet was resuspended in 200 μl PBS and the QIAamp® DNA Mini Kit (Qiagen) was used, according to manufacturer's instructions. DNA was eluted in 50 μl of AE buffer, quantified at A_{260} nm and stored at −80 °C until use.

2.3.4. RT assays

cDNA was generated from 2 μl of RNA sample from each time-point using TaqMan RT Reagents (Applied Biosystems), as previously described [127,166]. Briefly, 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 were used in a final reaction volume of 50 μl, although cycling conditions were modified: 10 min at 25 °C, 15 min at 42 °C and 5 min at 99 °C. As the results of preliminary assay 3 revealed identical RT performances using either
random hexamers or target-specific primer (see results), we decided to evaluate if the expression profile throughout development for the most stable gene (16SrRNA) would be similar using both methodologies. Thus, we have also generated cDNA for 16SrRNA by using 0.75 μM of 16SrRNA-10 primer (Table S2.1) and keeping the remaining RT conditions. cDNA were stored in DNase-free tubes at −80 °C, until use.

2.3.5. Generation of standard curves

For quantification of the number of bacterial genomes, a plasmid standard curve was generated as previously described [167]. Briefly, an amplified fragment of the single-copy of C. trachomatis ompA (Table S2.1) was cloned into the TOPO vector using the TOPO TA technology for PCR products prior to transformation into DH5α competent cells (Invitrogen), according to manufacturer's instructions. Plasmid DNA was purified using the QIAprep Spin Miniprep Kit Protocol (Qiagen), according to package inserts. To ensure elimination of any contaminating RNA that can interfere with plasmid quantification, RNase A was used. Confirmation of the cloning and transformation success was performed by EcoRI restriction digestion of the plasmids as well as by PCR amplification (using the same primers used to generate the fragments for cloning) and sequencing of each cloned fragment. The plasmid copy number was determined by A260 nm measurement of the plasmid concentration, according to the formula: No. plasmid/μl = (Avogadro's No. × Plasmid conc.(g/μl)) / MW of 1 mol of plasmids (g). Finally, eight-serial plasmid dilutions (concentration ranging from about 30 to 1×10^6 plasmid copies/μl) were prepared in order to generate the external standard curve for qPCR.

For transcripts quantification, a 48 h (pi) chlamydial culture was used for chromosomal DNA extraction. Subsequently, this stock DNA was subjected to eight two-fold serial dilutions in DNase-free water to generate the standard curves. The use of DNA standard curves allowed us to simultaneously compare the expression levels among different genes at each time-point, as DNA represents equal amounts of each single copy gene. This cross-comparison could not be achieved using cDNA standard curves.

2.3.6. Real-time quantitative PCR

For each gene, primers were designed using Primer Express (Applied Biosystems) (Table S2.1) based on constant regions determined through comparison of sequences available in GenBank.
The real-time quantification was achieved using the ABI7000 SDS, SYBR Green chemistry and optical plates (Applied Biosystems), as previously described [166]. The qPCR reagents consisted of 1× SYBR Green PCR Master Mix (Applied Biosystems), 400 nM of each primer and 5 µl of sample DNA or cDNA, in a final volume of 25 µl.

For each strain, plates included a plasmid standard curve and duplicates of DNA extracted at each time-point (for absolute quantification of bacterial genomes), as well as DNA standard curves and duplicates of cDNA obtained at each time-point (for quantification of transcripts). “No template” and “no RT” controls were also included in every qPCR assays. The thermocycling profile was: 10 min/95 °C followed by 40 cycles of 15 s/95 °C and 1 min/60 °C. The dissociation melting curves of the PCR products, generated by stepwise increase of the temperature from 60 °C to 95 °C, were used to verify the specificity of the amplified products. Finally, raw qPCR data at each time-point was normalized against the number of chlamydial genomes determined for the corresponding T25 cm² flask. The final qPCR results were based on three independent experiments for all strains.

2.3.7. Software applications (geNorm and Normfinder)

Gene expression stability was evaluated using two widely used MS Excel applications, geNorm version 3.5 [160] and Normfinder version 0.953 [161]. Briefly, the geNorm application assumes that the expression ratio of two stable genes will remain identical in all samples, independently of experimental conditions and cell-type, and computes a reference gene stability measure (M) for all genes included in a study. It also allows the stepwise elimination of the gene with the highest M value (less stable) and the recalculation of new M values for the remaining genes, and so on, until the two most stable are left. geNorm further determines the minimal number of genes required for normalization, taking into account a normalization factor based on the geometric mean of these best-scoring reference genes. With respect to Normfinder analysis, it assumes groups of samples and intra/intergroup variations to calculate the stability value for all genes tested, allowing the identification of the best endogenous control candidate and also the selection of the most stable pair of reference genes.

2.4. Results

2.4.1. Preliminary assays

D-cycloserine preliminary assays revealed that antibiotic exposure above 50 µg/ml induced morphologic alterations on HeLa 229 cells preventing their normal development. Thus, for
the subsequent experiments, we used a D-cycloserine concentration of 30 μg/ml, which is in agreement with previous studies [168,169]. During D-cycloserine treatment no apparent motion inside inclusions was observed at 6 h (pt) for all strains and both inclusions and RB showed abnormal size and shape (data not shown), without visible cytotoxic effects on HeLa cells. Contrarily, typical chlamydial inclusions and bacterial Brownian movements were detected (noticeably from 20 h pi) during the normal growth cycle.

The reproducibility of the nucleic acids extraction is a crucial aspect of the protocol that could bias the final normalized expression results. According to the normalization strategy adopted in this study, the number of chlamydial genomes found for each T25 cm² flask and the corresponding transcript amounts are obtained independently from one another, since both DNA and RNA were subjected to different extraction methods. In our preliminary assay, a mean variation coefficient of 22.1% (standard deviation (SD)±1.0%) and 26.9% (SD±2.1%) was obtained for DNA and RNA, respectively. Thus, it is reasonable to expect that the final normalized gene expression data may be slightly biased.

Evaluation of RT efficiency using random hexamers versus target-specific primer (for 16SrRNA) showed that the final proportions of cDNA obtained mirrored the initial amounts of RNA used in RT reaction. Indeed, qPCR assays revealed differences of about two threshold cycle (Ct) values between every two consecutive four-fold dilutions, where the probability of the deviation between the theoretical and the experimental obtained Ct being similar to zero was found to be 0.999 for both assays (data not shown). Curiously, the A260 nm determined for all cDNA samples was similar instead of reflecting their four-fold difference between consecutive dilutions. Thus, the A260 nm determination is not a reliable method for cDNA quantification because the nonconsumed RT reagents (nucleotides, primers, and enzyme) are the major responsible for the A260 nm measurements rather than the generated cDNA.

2.4.2. Real-time quantitative PCR

The melting curve analysis of all qPCR runs revealed a single-peak for each PCR product, indicating a high specificity of the methodology. For all qPCR assays, no amplification products were detected in both the “no template” and “no RT” controls. Plasmid and DNA standard curves presented high qPCR efficiency values for each gene (89.6% to 115.4%) (data not shown).
2.4.2.1. Chlamydial growth

The growth profile for C/TW3, E/Bour and L2/434 (Fig. 2.1) revealed an overall increase in chlamydial genomes of about 43-fold, 22-fold and 188-fold (from 4 h to 42 h pi), respectively. However, one E/Bour replicate showed less bacterial load for 12 h (pi) than for 4 h (pi), which could be due to the ~22% variation coefficient obtained in the DNA extraction procedure and/or to the inherent losses during cell harvesting. Surprisingly, C/TW3, which is assumed to have slow growth characteristics as any ocular strain [170], presented a doubling time of 2.87 h (SD±0.25) and the exponential phase starting at 12 h, although this period ended up at 20 h (pi). On the other hand, for E/Bour and L2/434, an extended exponential period (12–30 h pi) was observed with doubling times of 4.14 h (SD±1.05) and 2.36 h (SD±1.65), respectively.

![Figure 2.1 - Growth profile of C/TW3, E/Bour and L2/434 prototype strains.](image)

**Figure 2.1 -** Growth profile of C/TW3, E/Bour and L2/434 prototype strains. For each strain, the results at each time-point are represented as a mean±SD of the logarithmic fold-difference of the genome copy number (in each T_{25} cm^2 flask) compared to 4 h (pi). Black bars refer to normal growth conditions (4, 12, 20, 30 and 42 h pi), while grey bars represent bacterial growth under D-cycloserine treatment. Antibiotic was added at 18 h (pi) and harvesting was done at 18+6, 18+16 and 18+36 h (pi). Results are based on three biological replicates of each strain.

2.4.2.2. Gene expression

Figure 2.2 illustrates a global view of the expression pattern of each gene based on the log2 value of the mean expression at each time-point. In a rough estimation, genes of L2/434 presented expression levels about two-fold (not visible in Fig. 2.2) and four-fold higher than C/TW3 and E/Bour, respectively. The remarkable lower expression for E/Bour when compared to L2/434 supports previous findings observed for *pmp* genes [166]. For all strains, 16SrRNA was clearly the most expressed gene, showing expression values of >10³-fold than those of the remaining genes throughout development. Moreover, some genes seemed to be co-expressed during the three major temporal classes of gene expression (early, mid and late).
Normalization of *C. trachomatis* Expression

For instance, for C/TW3, five genes (*yraL*, *fer*, CT147, *radA* and *yaeI*) were simultaneously expressed at both the early- and mid-stages (4 h to 20 h pi) of chlamydial development. However, this chart does not allow an evaluation of the expression stability, as variations of about eight-fold can be represented by the same color.

**Figure 2.2** - Global view of the expression patterns of the 10 genes for *C. trachomatis*. Strains C/TW3, E/Bour and L2/434 were grown under normal conditions (4, 12, 20, 30 and 42 h pi). Expression values were normalized against the number of bacterial genomes. Each interval represents the log₂ values of the mean expression levels at each time-point, where the same color can represent variations up to eight-fold. Results are based on three biological replicates of each strain.

Thus, in order to accurately assess their expression stability, we carried out another strategy by calculating the absolute fold-difference (FD) between each time-point and the mean expression (determined for the entire cycle under normal growth conditions) (Fig. 2.3). Globally, most genes showed variation above four-fold to the respective mean. Exceptions were observed for: *oppA_2*, *hemN_2* and 16SrRNA for C/TW3, *oppA_2* and 16SrRNA for L2/434, and all genes except *fer* and CT147 for E/Bour. Moreover, 16SrRNA was the only gene for which expression variations never exceeded two-fold for all strains. However, *oppA_2* and *hemN_2* also exhibited variations below two-fold except for one time-point for each strain.
Figure 2.3 - Evaluation of expression stability of the 10 genes for C/TW3, E/Bour and L2/434 prototype strains. (A) Includes yraL, fer, CT147, oppA_2 and radA, whereas (B) includes yaeI, hemN_2, tyrP_2, map and 16SrRNA. Each graph represents the mean±SD absolute FD of expression values of each time-point compared to the mean expression value calculated for the entire developmental cycle under normal growth conditions. Expression values were normalized against the number of bacterial genomes. Black circles represent normal growth (4, 12, 20, 30 and 42 h pi), and white triangles refer to D-cycloserine exposure (18+6, 18+16 and 18+36 h pi) which was added at 18 h (pi).
Additionally, for each *C. trachomatis* strain, the 10 genes were ranked according to their expression stability, by calculating the sum of the absolute FDs between each time-point and the mean for the whole cycle as well as the sum of the corresponding SD (Table 2.1). Therefore, the best-ranked gene was 16SrRNA followed by oppA_2 for all strains. hemN_2 was ranked in third position but only for C/TW3 and L2/434. The remaining genes presented a wide variation in their rank positions.
Table 2.1. Gene expression stability evaluation throughout normal growth cycle

<table>
<thead>
<tr>
<th>Genes</th>
<th>C/TW3</th>
<th>E/Bour</th>
<th>L2/434</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank</td>
<td>Σ FDa</td>
<td>Σ SDb</td>
</tr>
<tr>
<td>16SrRNA</td>
<td>1</td>
<td>6.27</td>
<td>0.94</td>
</tr>
<tr>
<td>oppA_2</td>
<td>2</td>
<td>8.13</td>
<td>1.27</td>
</tr>
<tr>
<td>hemN_2</td>
<td>3</td>
<td>8.81</td>
<td>2.31</td>
</tr>
<tr>
<td>map</td>
<td>4</td>
<td>11.56</td>
<td>2.51</td>
</tr>
<tr>
<td>yaeI</td>
<td>5</td>
<td>11.97</td>
<td>2.49</td>
</tr>
<tr>
<td>radA</td>
<td>6</td>
<td>13.28</td>
<td>2.38</td>
</tr>
<tr>
<td>yraL</td>
<td>7</td>
<td>14.87</td>
<td>3.98</td>
</tr>
<tr>
<td>CT147</td>
<td>8</td>
<td>15.03</td>
<td>4.16</td>
</tr>
<tr>
<td>fer</td>
<td>9</td>
<td>20.93</td>
<td>5.07</td>
</tr>
<tr>
<td>tyrP_2</td>
<td>10</td>
<td>26.15</td>
<td>7.20</td>
</tr>
</tbody>
</table>

a Sum of the absolute FD between each time-point and the mean calculated for the whole cycle.
b Sum of the corresponding SD.

Since 16SrRNA was the most stable gene in the expression study using random hexamers in the RT reaction, we determined its expression profile throughout development of all strains using a target-specific primer. Matching results were obtained, where 16SrRNA gene did not exceed two-fold variation relatively to the mean expression value for all time-points (Figs. 2.3B & 2.4A).

2.4.2.3. D-cycloserine assays

As expected, the influence of D-cycloserine treatment on bacterial growth did not seem to affect the normal increase of genome copy number (Fig. 2.1). In fact, this antibiotic is known to block the peptidoglycan synthesis rather than interfering with chromosomal replication [163,164,168].

The effect of D-cycloserine in gene expression was also assessed (Fig. 2.3). For most genes, the largest expression variation occurred at 36 h (pt), when compared to the mean expression calculated from mRNA levels at time-points with no antibiotic treatment. In general, for all strains, the prolonged exposure to D-cycloserine resulted in a gradual decrease of gene expression (not shown in Fig. 2.3). For E/Bour, all genes except fer, CT147 and yaeI showed no expression variations above four-fold to the mean value, suggesting a lower influence of this antibiotic on the gene expression of this strain compared to that of C/TW3 and L2/434 where greater expression fluctuations were observed for all genes. Globally, oppA_2 and hemN_2 were the least variable genes under this stress condition with no variations larger than four-fold to the mean value, for all strains (except hemN_2 for L2/434).
Figure 2.4 - RT-qPCR results of 16SrRNA gene by using RT target-specific priming. (A) Represents the mean±SD absolute FD of expression values of each time-point compared to the mean expression value calculated for the entire development cycle under normal growth conditions. (B) Refers to the 16SrRNA gene expression profile throughout development, where the logarithm transformation enables a symmetrical distribution of the values around zero allowing the representation of up- or down-regulation phenomena for each time-point. Black circles in (A) and black bars in (B) refer to normal growth conditions (4, 12, 20, 30 and 42 h pi), while white triangles in (A) and grey bars in (B) represent bacterial growth under D-cycloserine treatment (18+6, 18+16 and 18+36 h pi).

As above, we also ranked all genes according to their expression stability under antibiotic treatment (Table 2.2). The results showed that oppA_2 was clearly the most stable gene presenting the lower sum of the absolute FD values as well as the lower sum of the corresponding SD for all strains, whereas the rank position of the other genes was incoherent among strains.

Regarding these analyses, all genes seemed to be down-regulated by D-cycloserine treatment (data not shown). However, oppA_2 was the least affected, and will be a good endogenous control candidate to be used in upcoming gene expression studies under D-cycloserine treatment. Although 16SrRNA was the most stable gene under normal growth conditions, its expression was clearly down-regulated in the antibiotic-treated cultures (Fig. 2.4B).
Table 2.2. Gene expression stability evaluation throughout growth cycle under D-cycloserine treatment

<table>
<thead>
<tr>
<th>Genes</th>
<th>C/TW3 Rank</th>
<th>C/TW3 Σ FDa</th>
<th>C/TW3 Σ SDb</th>
<th>E/Bour Rank</th>
<th>E/Bour Σ FDa</th>
<th>E/Bour Σ SDb</th>
<th>L2/434 Rank</th>
<th>L2/434 Σ FDa</th>
<th>L2/434 Σ SDb</th>
</tr>
</thead>
<tbody>
<tr>
<td>oppA_2</td>
<td>1</td>
<td>6.06</td>
<td>1.33</td>
<td>1</td>
<td>6.34</td>
<td>1.95</td>
<td>1</td>
<td>6.64</td>
<td>2.81</td>
</tr>
<tr>
<td>hemN_2</td>
<td>2</td>
<td>8.27</td>
<td>1.94</td>
<td>5</td>
<td>8.14</td>
<td>3.35</td>
<td>4</td>
<td>18.89</td>
<td>13.07</td>
</tr>
<tr>
<td>16SrRNA</td>
<td>3</td>
<td>11.10</td>
<td>3.65</td>
<td>7</td>
<td>8.63</td>
<td>3.54</td>
<td>2</td>
<td>13.39</td>
<td>4.85</td>
</tr>
<tr>
<td>radA</td>
<td>4</td>
<td>17.17</td>
<td>6.19</td>
<td>4</td>
<td>8.05</td>
<td>3.91</td>
<td>3</td>
<td>15.05</td>
<td>6.21</td>
</tr>
<tr>
<td>map</td>
<td>5</td>
<td>18.79</td>
<td>5.89</td>
<td>2</td>
<td>6.96</td>
<td>3.38</td>
<td>6</td>
<td>30.34</td>
<td>20.16</td>
</tr>
<tr>
<td>yaeI</td>
<td>6</td>
<td>20.31</td>
<td>4.40</td>
<td>8</td>
<td>14.22</td>
<td>12.14</td>
<td>8</td>
<td>40.42</td>
<td>35.03</td>
</tr>
<tr>
<td>yraL</td>
<td>7</td>
<td>22.07</td>
<td>6.15</td>
<td>6</td>
<td>8.14</td>
<td>3.68</td>
<td>5</td>
<td>22.07</td>
<td>9.33</td>
</tr>
<tr>
<td>fer</td>
<td>8</td>
<td>28.35</td>
<td>6.01</td>
<td>9</td>
<td>15.16</td>
<td>12.46</td>
<td>9</td>
<td>49.59</td>
<td>30.77</td>
</tr>
<tr>
<td>CT147</td>
<td>9</td>
<td>42.60</td>
<td>12.41</td>
<td>10</td>
<td>25.45</td>
<td>20.53</td>
<td>7</td>
<td>31.28</td>
<td>19.63</td>
</tr>
<tr>
<td>tyrP_2</td>
<td>10</td>
<td>66.18</td>
<td>19.61</td>
<td>3</td>
<td>8.03</td>
<td>2.99</td>
<td>10</td>
<td>55.00</td>
<td>44.66</td>
</tr>
</tbody>
</table>

a sum of the absolute FD between each time-point and the mean calculated for the whole cycle.

b sum of the corresponding SD.

2.4.3. Software applications (geNorm and Normfinder)

The results obtained through the application of both mathematical algorithms (Table 2.3) clearly diverged from those obtained experimentally. The top-ranked genes described above (16SrRNA and oppA_2) were classified as unstable by both software tools for the three strains. In fact, geNorm and Normfinder always ranked 16SrRNA and oppA_2 among the five most unstable genes. Moreover, for the top-ranked genes, no agreement was found to exist between the results of both softwares for each strain (Table 2.3).

| Table 2.3. Top-ranked genes according to geNorm and Normfinder applications |
|-----------------------------|-----------------------------|-----------------------------|
|                            | C/TW3                      | E/Bour                      | L2/434                     |
| geNorm NormFinder          | geNorm NormFinder          | geNorm NormFinder          |
| Top -                      | yraL radA                  | radA radA                  |
| Ranked                     | fer yraL                   | tyrP_2 map                 |
| Genes                      | tyrP_2 yael/map             | map yraL                   |

2.5. Discussion

Genomics and transcriptomics have been fundamental approaches to get insights into the biology of the genetically non-tractable C. trachomatis. Hence, the progressive improvement of RT-qPCR allows a better understanding of the expression profile of chlamydial genes. However, the absence of a validated normalization strategy for C. trachomatis remains an obstacle for the acquirement of accurate expression data. The use of genomic DNA (gDNA) as normalizing factor for gene expression data has proved to be a reliable method [145,171],
Normalization of *C. trachomatis* Expression

yielding even more precise results than the use of quantitative culture [145]. As the later cannot be accurately applied to obligate intracellular pathogens, such as *C. trachomatis*, we have performed absolute qPCR to determine the number of bacterial genomes, which was used as denominator to evaluate the expression stability of potential endogenous controls.

So far, no such evaluation was done in *C. trachomatis*, although the results of a previous study [172] suggested that 16SrRNA gene could be a good endogenous control candidate for qPCR, and it has been clearly the most widely used endogenous control in *Chlamydia* to date [96,166,172-174]. Our results showed that 16SrRNA was the most stably expressed gene throughout *C. trachomatis* developmental cycle even for strains with dissimilar growth rate and representing different disease groups. This consistent 16SrRNA gene expression profile together with the extremely high mRNA levels when compared to the other genes, likely mirrors the homogeneous need of this ribosomal component throughout *C. trachomatis* normal developmental cycle. In fact, *C. trachomatis* is an obligate intracellular pathogen that presents a complex and unique developmental cycle where protein production is an endless process which may be associated with early stages (attachment/invasion, T3S effectors, disarm innate immune responses, disrupting apoptosis, etc.), mid-stages (replication, development of inclusion membrane, T3S effectors, etc.), or late-stages (reorganization of the replicating form into the infectious form, cell lysis/extrusion, etc.). 16SrRNA showed an expression of about $10^4$-fold higher than that of the other two experimentally best-ranked genes, *oppA_2* and *hemN_2* (Fig. 2.2), which code for an oligopeptide binding protein permease and a coproporphyrinogen III oxidase, respectively. For these later genes, transcript levels were only detected around 30th Ct, where higher variance is expected to occur considering the Monte Carlo effect [175], which postulates that artifacts associated with the amplification of small amounts of template within complex nucleic acid mixtures (as for intracellular pathogens like *Chlamydia*) may compromise PCR reproducibility.

The regulation of housekeeping metabolic activity may be a very dynamic process in response to variation of external/environmental conditions, which is particularly applicable to pathogens considering the permanent “arms race” with the host [165]. Thus, it is expected that in response to hostile environmental conditions, the amount of ribosomes declines rapidly in order to conserve energy for other metabolic processes [176]. Considering this, and although no other stress condition was assayed, the observed decline of the 16SrRNA gene expression under D-cycloserine treatment is expected, and show the potential inaccuracy of using this gene for normalization in expression assays when stress conditions are implicated. On the other hand, *oppA_2* was found to be the least unstable gene, although we have no reasonable
explanation for that, as any association with its biological role is merely speculative. Nevertheless, oppA_2 should be part of a pool of potential endogenous controls in future qPCR evaluations under other stress conditions.

We have also evaluated the suitability of geNorm and NormFinder (software applications widely used for other organisms) [157,160,161,177] for calculating the expression stability of the *C. trachomatis* selected genes. In the present study, the experimentally best-ranked genes, 16SrRNA and oppA_2, were never placed among the most stable genes according to geNorm and Normfinder. This was not surprising as these applications select the most stable genes within a pool of genes assumed to be relatively stable, which our experimental assays subsequently revealed not to be the case. Moreover, the results generated by geNorm require that the genes under evaluation are not co-regulated, because a strong bias of this software is that it top ranks co-expressed genes disregarding their stability [178]. Although we have selected genes that are dispersed throughout the chromosome and not likely related in terms of metabolic function, the hypothesis of co-regulation for some of them cannot be discarded because there is a lack of information about the expression profiles for the ~900 *C. trachomatis* ORFs. In fact, the transcriptome was only evaluated for D/UW3 and L2/434 strains [38,96] and it is well-known that gene expression may vary considerably even between same-genotype *C. trachomatis* strains [166]. The postulation that genes are not co-regulated only because they are chromosome spread and involved in different metabolic functions, which has been assumed by some authors for other bacteria [155,177], may be dangerous and may lead to biased results as clearly observed in the present study. Indeed, considering the association between gene expression and the three main bacterial developmental stages (lag, exponential, and stationary) and also the existence of the three major temporal classes of *Chlamydia* gene expression [179,180], we may roughly estimate that any gene is co-regulated with about 30% of all the other genes, which was experimentally observed (Fig. 2.2) for the 10 genes under evaluation. In the case of Normfinder, which overcomes the co-regulation problem that affects the pairwise comparison approach of geNorm, we can only speculate that the lack of stability of the pool of genes was the main reason for this results disagreement. Thus, geNorm and NormFinder are likely ineffective in the present model and they should be used with caution for any other organism for which no “guiding” transcriptome is available yet.

Although gDNA was used for evaluation of endogenous controls in the present study, it may be an accurate optional approach for general normalization of qPCR data instead of using endogenous controls. In fact, gDNA provides an indirect estimate of the number of bacteria (some fluctuations may exist throughout development in the number of genomes *per*
bacterium [146]) that is generating the target mRNA. Contrarily to the relative quantification assays, the most critical point of this normalization strategy is associated with the fact that the determination of mRNA (numerator) and gDNA (denominator) are experimentally independent. Thus, a lack of reproducibility on both RNA and DNA extraction procedures unequivocally influences the expression profiles. In fact, in the preliminary assay, the reproducibility of the repeated RNA and DNA extractions of the same sample revealed a variation coefficient of 26.9% (SD±2.1%) and 22.1% (SD±1.0%), respectively. Considering these results, it is plausible that the gene expression differences between time-points are overestimated. Another possible drawback of using gDNA as denominator is that it does not correct for variability in RT efficiency. However, our results showed a tremendous RT steadiness using both random hexamers and target-specific primer. Also, we speculate that calibration using gDNA would clearly be more appropriate for evaluating gene expression under stress conditions, as chromosomal copies likely have little fluctuations when compared to the huge variation observed in mRNA levels.

Globally, we believe that our results constitute a valuable guide for chlamydial transcriptomics, which is an essential tool for studying the pathobiology of this genetically non-tractable obligate intracellular pathogen. Besides the validation of 16SrRNA gene as a potential endogenous control for qPCR assays, our strategy of using gDNA can also be considered an accurate alternative, especially in studies involving intracellular organisms as well as stress conditions, enabling more reliable cross-comparison of results from different laboratories. However, there is not a single strategy that fits every experimental situation, and it remains up to researchers the validation of the most suitable method for each specific experiment.
CHAPTER 3

Comparative Expression Profiling of the *C. trachomatis pmp* Gene Family for Clinical and Prototype Strains

*Published in*


*Author Contributions*

NA, JPG, MJB and DD conceived and designed the experiments; AN, CF, HC, JPG, MJB and SM performed the experiments; AN, JPG and DD analyzed the data; AN, JPG and DD wrote the paper.
3. Comparative Expression Profiling of the *C. trachomatis* *pmp* Gene Family for Clinical and Prototype Strains

3.1. Abstract

*Chlamydia trachomatis*, an obligate intracellular pathogen, is a leading worldwide cause of ocular and urogenital diseases. Advances have been made in our understanding of the nine-member *pmp* family of *C. trachomatis*. However, there is only limited information on their biologic role, especially for biological variants (biovar) and clinical strains. We evaluated expression for *pmps* throughout development for prototype strains E/Bour and L2/434, representing different biovars, and for clinical E and L2 strains. Immunoreactivity of patient sera to recombinant (r)Pmps was also determined. All *pmps* were expressed at two hours. *pmpA* had the lowest expression but was up-regulated at 12 h for all strains, indicating involvement in reticulate body development. For *pmpD*, expression peaked at 36 h. Additionally, 57.7% of sera from infected and 0% from uninfected adolescents were reactive to rPmpD (*P* = 0.001), suggesting a role in immunogenicity. *pmpF* had the highest expression levels for all clinical strains and L2/434 with differential expression of the *pmpFE* operon for the same strains. Sera were nonreactive to rPmpF despite immunoreactivity to rMOMP and rPmpD, suggesting that PmpF is not associated with humoral immune responses. *pmpFE* sequences for clinical strains were identical to those of the respective prototype strains. We identified the putative *pmpFE* promoter, which was, surprisingly, 100% conserved for all strains. Analyses of ribosomal binding sites, RNase E, and hairpin structures suggested complex regulatory mechanism(s) for this >6 Kb operon. The dissimilar expression of the same *pmp* for different *C. trachomatis* strains may explain different strain-specific needs and phenotypic distinctions. This is further supported by the differential immunoreactivity to rPmpD and rPmpF of sera from patients infected with different strains. Furthermore, clinical E strains did not correlate with the E prototype strain at the gene expression level, reinforcing the need for expansive studies of clinical strains.

3.2. Introduction

*Chlamydia trachomatis* is an obligate intracellular pathogen that is responsible for significant worldwide morbidity associated with ocular and STD. The developmental cycle of the organism is biphasic beginning with the adhesion of the EB, an infectious and metabolically inert form, to the host cell. After endocytosis, the EB differentiates ~2 to 12 h (pi) into a
larger, non-infectious and metabolically active RB, which initiates intracellular replication by binary fission within a vacuole called an inclusion body. At ~30 to 36 h (pi), RBs reorganize into new EBs, which are released by host cell lysis or exocytosis at 48 to 72 h (pi) that initiates another infectious cycle [29,181].

The disease spectrum of \textit{C. trachomatis} ranges from conjunctivitis and ocular trachoma to tubal factor infertility, ectopic pregnancy and infant pneumonitis [182,183]. \textit{C. trachomatis} serological variants (serovars) are grouped within two human biovars according to characteristics of disease presentation: the trachoma biovar, including serovars A to C and Ba, which cause conjunctivitis and trachoma, and serovars D to K and Ba, Da, Ia and Ja, which cause conjunctivitis, urogenital infections and infant pneumonitis, and the LGV biovar (serovars L1 to L3 and L2a). The later are more invasive, causing genital ulceration, lymphadenitis and proctitis [4,182]. However, serotyping of MOMP and phylogenetic reconstructions of this protein and the corresponding gene (\textit{ompA}) [184,185] do not group serovars by trachoma, non-invasive urogenital or invasive LGV disease groups.

The molecular mechanisms behind these biological differences among serovars (or strains) are not well understood. Genome sequences of prototype strains D/UW3 [2] and A/Har13 [5], as well as ongoing \textit{C. trachomatis} genomic sequencing are providing information on specific genes and proteins that may explain tissue tropism and virulence differences for the three disease groups. \textit{C. trachomatis} contains a partial tryptophan operon where urogenital strains, but not trachoma strains, can synthesize tryptophan from mucosal substrates [108]. The toxin gene possesses an intact N-terminal region that encodes an active enzymatic domain for the urogenital strains but not for trachoma or LGV strains [107].

Research on the nine member polymorphic membrane protein (Pmp) gene (\textit{pmp}) family has revealed phylogenetic reconstructions where six pmps (\textit{pmpB}, \textit{pmpC}, \textit{pmpF}, \textit{pmpG}, \textit{pmpH} and \textit{pmpI}) form clades that correspond to the three disease groups [113,121,186]. At the gene expression level, previous RT-PCR analyses of prototype strains D/UW3 and L2/434 [187], and microarray analysis of D/UW3 [96] found that all nine \textit{pmps} were transcribed starting at 8 h (pi). Yet, based on RT-qPCR, we found expression as early as 2 h (pi) for \textit{pmpC} for prototype strains Ba/Apache2, G/UW57 and L2/434, and a differential expression profile with earlier up-regulation of \textit{pmpC} for L2/434 [127]. Also, another study based on RT-qPCR, Kiselev \textit{et al.} [123] detected \textit{pmpD} expression as early as 2h (pi) for L2/434. Proteomics analyses have also shown that all Pmps of L2/434 are detected as outer membrane constituents [119,120,188,189]. There is also evidence that some Pmps are antigenic for human sera [126,187]. We observed a heterogeneous immunoreactivity to rPmpC using sera from patients infected with different \textit{C. trachomatis} strains, suggesting a
role for PmpC in antigenic variation [127]. More recently, Pmps have been considered autotransporters based on bioinformatics analyses [122,190]. Wehrl et al. [61] has experimentally demonstrated the autotransporter model for the *C. pneumonia* ortholog of *C. trachomatis* PmpD, Pmp21. Further, using immunofluorescence microscopy, Western blotting and penicillin treatment, the results of Kiselev et al. [123] for L2/434 PmpD are in general agreement with the autotransporter model for this protein. PmpD has also been shown to be a species-common neutralizing antigen [62], while PmpF has been implicated as a potential target of the host immune response as it contains several predicted MHC epitopes within the N-terminal domain [5].

Despite the potential importance of Pmps in chlamydial biology, there is a lack of expression data for the *pmp* genes as well as an insufficient understanding of the host immune response to their proteins. Here, we profile the expression of all *pmps* throughout development for prototype strains E/Bour and L2/434, representing the two *C. trachomatis* biovars. We chose E/Bour because it is the most prevalent strain worldwide, although the mechanisms of its ecological success are not yet understood. L2/434 was selected as it has been widely studied with a plethora of biological information for comparative analyses. The biological uniqueness of these two strains *in vivo* is reflected in their differential tissue tropism, virulence and disease presentation. In light of our recent findings that prototype strains do not represent the same genetic composition of clinical strains that are circulating among human populations today [114], we also compared the nine *pmp* expression levels for four *C. trachomatis* clinical strains, representing *ompA* genotypes of E and L2. Further, we examined the immunoreactivity of sera from adolescents with and without *C. trachomatis* urogenital infections against rPmps to further define their potential importance in human disease.

### 3.3. Materials and methods

#### 3.3.1. *C. trachomatis* cell culture of prototype strains and clinical strains

*C. trachomatis* prototype strains E/Bour and L2/434, three clinical strains belonging to *ompA* genotype E (designated as E/537C-05, E/S-141 and E/CS-500-96) and one clinical strain belonging to *ompA* genotype L2 were evaluated in this study. E/537C-05 and E/S-141 were collected from patients with vaginal discharge, E/CS-500-96 from a patient with cervicitis, and L2 from a patient with proctitis. Each was propagated in HeLa229 cell monolayers using standard techniques as previously described [183,191]. EBs were harvested at 48–72 h (pi) and purified by discontinuous density centrifugation in Renografin [133].
Confluent HeLa cells were either mock-infected or infected with a multiplicity of infection of one for each prototype strain or clinical strain in sucrose-phosphate-glutamate (SPG) buffer prior to incubation with culture medium [183,191]. Eight T25 cm² flasks (one for each time point of 2, 6, 12, 18, 24, 36, and 48 h and mock-infected) per strain were inoculated and placed at 37 ºC in 5% CO₂ [127]. Cultured cells were harvested at each time point, and total RNA was extracted as previously described [127].

3.3.2. Reverse Transcription and Real-Time Quantitative PCR

RNA was quantified by OD measured at A260 nm. cDNA was generated from 500 ng of each RNA sample using TaqMan RT Reagents (Applied Biosystems) and random hexamers, and was quantified at A260 nm.

Quantitation of pmp expression was achieved using the ABI7000 SDS (Applied Biosystems), SYBR Green chemistry, and the standard curve method for relative quantitation, using reagents and thermocycling as previously described [127]. 16SrRNA gene was used as the endogenous control since normalizing the data against 16SrRNA gene provides a control for the number of organisms (EBs and RBs) and, therefore, for the differential growth rate of each strain. ompA was included as a quality control for RT-qPCR results since it has been widely used for gene expression studies [96,127].

Primers for each of the nine pmps (Table 3.1) were designed using Primer Express (Applied Biosystems). Primers for ompA, 16SrRNA, and pmpC were used as previously designed [127].
Table 3.1. Oligonucleotide primers used for RT-qPCR

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primers</th>
<th>Primer sequence (5’ to 3’)</th>
<th>Gene Location</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>pmpA</em></td>
<td><em>pmpA-3</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGCTAGGGAAGATGTTGCAAATAG</td>
<td>1434-1457</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td><em>pmpA-4</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGACGGGCTTTGCTTAAATACG</td>
<td>1484-1463</td>
<td>102</td>
</tr>
<tr>
<td><em>pmpB</em></td>
<td><em>pmpB-5</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CGACTATCAGCAAACACTGCAA</td>
<td>2120-2144</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td><em>pmpB-6</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TACCGAGTCTCTGAGATATTCGTT</td>
<td>2221-2295</td>
<td>51</td>
</tr>
<tr>
<td><em>pmpC</em></td>
<td><em>pmpC-11</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TTAGGTCCCTCTACAGACTCATCA</td>
<td>4150-4174</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td><em>pmpC-12</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CCCGTCAGTACTATTTTCTGAGCTT</td>
<td>4205-4181</td>
<td>51</td>
</tr>
<tr>
<td><em>pmpD</em></td>
<td><em>pmpD-3</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>GCCGGTCTGCTTCTTAAATAT</td>
<td>4455-4476</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td><em>pmpD-4</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ACTGTCGTAGAAGAACTACCAGTTC</td>
<td>4505-4480</td>
<td>51</td>
</tr>
<tr>
<td><em>pmpE</em></td>
<td><em>pmpE-1</em>&lt;sup&gt;c&lt;/sup&gt;</td>
<td>CATATCCGTCTTCCGGAATAC</td>
<td>2140-2160</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td><em>pmpE-2</em>&lt;sup&gt;c&lt;/sup&gt;</td>
<td>GTGTGTCTGCTCCGTCTACAT</td>
<td>2190-2169</td>
<td>51</td>
</tr>
<tr>
<td><em>pmpF</em></td>
<td><em>pmpF-5</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TCTCATGTTTGATCGCATTGCT</td>
<td>2520-2541</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td><em>pmpF-6</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CTCCGAGTCTCTGAGATATTCGTT</td>
<td>2588-2566</td>
<td>69</td>
</tr>
<tr>
<td><em>pmpG</em></td>
<td><em>pmpG-1</em>&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>TGGGTGCTTTGGGAGTTTTTGAAT</td>
<td>2221-2242</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td><em>pmpG-2</em>&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>ACTTAAAGGATGTCGGGCTTCAT</td>
<td>2271-2252</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td><em>pmpG-3</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGGGCCTCTGCTCAGAATTCTCAGT</td>
<td>1165-1187</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td><em>pmpG-4</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>AAATCGCTCCACCATCATTAGC</td>
<td>1216-1195</td>
<td>52</td>
</tr>
<tr>
<td><em>pmpH</em></td>
<td><em>pmpH-15</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGCAATAGCGACTTATACTTGAATCA</td>
<td>2486-2511</td>
<td>61</td>
</tr>
<tr>
<td><em>pmpI</em></td>
<td><em>pmpI-1</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGCATAGGGAAGATGTTGCAAATAG</td>
<td>2546-2525</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td><em>pmpI-2</em>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>GAAAGTCTGCGCAGCATTGCT</td>
<td>2176-2195</td>
<td>51</td>
</tr>
<tr>
<td><em>ompA</em></td>
<td><em>OmpA-9</em>&lt;sup&gt;c&lt;/sup&gt;</td>
<td>TCCCGGCTTTGAGATGTTGCT</td>
<td>33-52</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td><em>OmpA-10</em>&lt;sup&gt;c&lt;/sup&gt;</td>
<td>GTCGATCATAAGCGTTGCTTCA</td>
<td>108-86</td>
<td>75</td>
</tr>
<tr>
<td>16SrRNA</td>
<td>16SRNA-9&lt;sup&gt;d&lt;/sup&gt;</td>
<td>GCCAGCGCCGCTTTTCTAATATTCAT</td>
<td>734-756</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>16SRNA-10&lt;sup&gt;d&lt;/sup&gt;</td>
<td>CCAGGATGTCTACTTCTTCTGTTTCT</td>
<td>809-786</td>
<td>76</td>
</tr>
</tbody>
</table>

<sup>a</sup> Primers designed based on each *pmp* sequence of prototype strains E/Bour and L2/434 [121].

<sup>b</sup> Primers designed only for strain L2.

<sup>c</sup> Primers designed based on the *ompA* sequence of prototype strains E/Bour and L2/434 (GenBank: X52557 and M14738, respectively).

<sup>d</sup> Primers designed based on the 16SrRNA gene sequence of prototype strains E/Bour and L2/434 (GenBank: D85722 and U68443, respectively).

Each plate contained two replicates of each sample cDNA, three different negative controls and standard curves for each gene as previously described [127]. For all experiments, the amount of target and control gene was determined from the respective standard curve by conversion of the mean threshold cycle values. Normalization was obtained by dividing the quantity of the target gene by the quantity of the control gene. The specificity of the amplified products was verified by analysis of the dissociation curves generated by the ABI7000 software based on the specific melting temperature for each amplicon. The results were based on three independent experiments for prototype strains E/Bour and L2/434, and for the four clinical strains.
3.3.3. **Genetic analysis of the pmpFE operon for *C. trachomatis* prototype and clinical strains**

Based on the considerable expression disparities between *pmpF* and *pmpE* (which belong to the same operon) for prototype strain L2/434 and mostly for the clinical strains (see results below), we sequenced the *pmpFE* operon as well as the upstream 164 bp *pmpG/pmpF* intergenomic region (IGR) that likely contains the operon regulatory region. In the *C. trachomatis* chromosome, *pmpF* and *pmpE* are located on the minus strand; *pmpF* is located upstream of *pmpE*, with a 2 bp IGR. *pmpE* was sequenced for the six strains (Genbank: EF490370 for E/537C-05, EF490371 for E/CS-500-96, EF490372 for E/S-141, EF490373 for L2, EF490374 for E/Bour, and EF490375 for L2/434), while *pmpG/pmpF* IGR and *pmpF* were sequenced only for the clinical strains (Genbank: EF490366 for L2, EF490367 for E/CS-500-96, EF490368 for E/S-141, and EF490369 for E/537C-05), as the sequences for the prototype strains were available from our previous study (GenBank: AY887650 for E/Bour and AY887660 for L2/434) [121]. The amplification and sequencing strategies were performed as previously described [121], except for the *pmpG/pmpF* IGR, where we used primer 5’-ACTCGGATCTCCTATAACAG-3’ for sequencing.

Since the transcription process can be strongly affected by the structure and sequence variability of promoter regions [192-195], a putative promoter search for the *pmpFE* operon was performed using EditSeq software (DNASTAR) for sequences described in the literature and also by using two promoter prediction programs: [http://www.fruitfly.org/seq_tools/promoter.html](http://www.fruitfly.org/seq_tools/promoter.html) and [http://www.prodoric.de/vfp/vfp_promoter.php](http://www.prodoric.de/vfp/vfp_promoter.php). Given the expression dissimilarities obtained in this study for *pmpF* and *pmpE*, we searched for putative Shine-Dalgarno ribosome binding sequences (RBS) [196] as well as previously described chlamydial RBS [197-200] within this operon, since ribosomes can either prolong or shorten the lifetime of mRNA in response to events that occur during translation or termination processes [201]. The existence of putative consensus cleavage sites for RNase E [202-204], the major endonuclease that generally initiates mRNA degradation in most bacteria [205], was also examined within the *pmpFE* operon. These two analyses were performed using EditSeq software (DNASTAR). Putative stemloop structures were searched throughout the *pmpFE* operon using GeneQuest software (DNASTAR) and RNAstructure software version 4.4 ([http://rna.urmc.rochester.edu/rnastructure.html](http://rna.urmc.rochester.edu/rnastructure.html)) due to the regulatory or processing role of stem-loop structures in premature transcription termination as well as in mRNA degradation and maturation mechanisms [206-210], respectively.
3.3.4. Immunoreactivity of patient sera against Pmp fusion proteins.

We generated fusion proteins for PmpD and PmpF because the latter displayed such high mRNA expression for L2/434 and the clinical strains, and the former was expressed late in development for all strains under study, being the last up-regulated protein for four of the six strains analyzed. Also, PmpD has been associated with neutralizing epitopes [62]. The rMOMP fusion protein was available from a previous study [211]. The PET30 expression system (EMD Biosciences) was used for cloning PCR products containing pmpD or pmpF generated from strain E/Bour genomic DNA as we have described [127]. The forward and reverse primers were 5’-GACGACGACAAGATGAGTTCCGAGAAAGATATA-3’ and 5’-AATGCTGGATTGCCATTCTCTTCACCCGGCTTCTCCTC-3’ for pmpD, respectively, and 5’-GACGACGACAAGATGATTAAAAGAACTTCTCTA-3’ and 5’-AATGCAGGAGGAGCTCTGGTCTTTTAACCGGGCTTCTCCTC-3’ for pmpF, respectively. Sequencing confirmed that the insert was in frame with the S-tag and His-tag as we have described previously for rPmpC [127]. The clones were transformed into E. coli BL21 and induced using 0.1 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) during the exponential growth phase. Ni-agarose (Sigma) was used for fusion protein purification according to the package insert. Recombinant proteins were determined to be the correct molecular weight (calculated at ~160.6 kDa for rPmpD and ~112.3 kDa for rPmpF) by immunoblot using AP-conjugated S-protein, which binds to the S-tag peptide with a distinct band at the correct molecular weight for each. Optimal protein concentrations were determined and standardized using BCA (Pierce) before analyzing the clinical sera. The optimal protein concentration for rPmpD was 50 ng and 100 ng for rPmpF. Sera from 39 consented female adolescents 14 to 19 years of age attending clinics in Oakland (USA) were used at a 1:50 dilution for immunoblotting as described previously [127]. The Institutional Review Board of Children’s Hospital Oakland (USA) approved the study, and all patients provided written consent for all clinical samples that were obtained and used in this study. The blots were blocked with Blotto prior to reacting with patient sera and alkaline phosphatase-conjugated anti-human IgG (R&D Systems). The chemiluminescent substrate ECF (Amersham) was used to visualize reactive bands. Twenty six (67%) of the 39 adolescents were infected with a single ompA genotype: 3 Ba, 3 D, 8 E, 5 F, 1 G, 1 Ia, 2 J, and 3 K.

There was no evidence for mixed infections. The original cervical samples were used for sequencing [127] to best determine the presence of a mixed infection, since propagation may result in one strain overgrowing another. On inspection of the electropherograms, none of the samples had ambiguous results. All nucleotides (nts) were represented by single, clear
peaks with extremely low background and without evidence for double peaks in a nt position where different *ompA* genotypes differ, which might suggest a mixed infection.

## 3.4. Results

### 3.4.1. Real-time quantitation for *pmp* expression

The results of specificity assays revealed no nonspecific products, and indicated the presence of the expected amplicon for each gene. Standard curves for all 11 genes had slope values between -3.1 and -3.5, which represents efficiencies between 93 and 100%. There were only minor variations in the slope for each standard curve among independent experiments, indicating a highly reproducible qPCR as we have also shown in previous experiments [127].

We defined the gene expression profile as the qualitative gene expression pattern throughout development where quantitative values are not considered. For example, one expression profile would show increasing expression up to a peak with tapering down of the expression after the peak. An expression peak was defined as the time point of the highest relative mRNA value. All quantitative expression comparisons refer to differences between the expression peaks of each gene, even those occurring at different time-points (pi).

### 3.4.2. Expression profile of the nine *pmp* genes throughout development for E/Bour and L2/434

L2/434 had strikingly different mRNA levels among some *pmps* and also between different time points for the same *pmp* (Fig. 3.1A). *pmpF* had the highest relative mRNA expression, up to 11.5-fold higher than for *pmpA*, the least expressed gene. mRNA levels were detected at 2 h (pi) for all *pmps*, including *pmpA* where the scale limits visualization of the low mRNA expression, and peaked at different time points. For all *pmps* except *pmpA*, mRNA levels decreased consistently after the peak until 48 h.

The *pmp* expression profiles for E/Bour were more homogeneous than for L2/434, and, in some cases, mRNA levels were lower than for the corresponding L2/434 *pmp* (Fig. 3.1B). Similar to L2/434, mRNA levels were detected at 2 h (pi) for all *pmp* genes. *pmpE*, *pmpF* and *pmpG* showed the highest expression levels for strain E/Bour with up to 3.1-fold higher mRNA levels than for *pmpA* and *pmpI*, the least expressed genes. As for L2/434, *pmpA* and *pmpD* were the earliest and the latest up-regulated genes, respectively. In contrast to L2/434, all *pmp* genes except *pmpA* and *pmpI* had stable mRNA levels after the expression peak until 48 h. For this reason, the expression peak for E/Bour *pmps* was defined as the time point at which a noticeable expression increase occurred.
For both prototype strains, \(ompA\) had remarkably higher mRNA values at all time points than for the \(pmp\)s (Fig. 3.1). In contrast to most \(pmp\)s, \(ompA\) revealed a similar gene expression profile for both prototype strains.

3.4.3. Expression profile of the nine \(pmp\) genes throughout development for \(C.\ trachomatis\) clinical strains

The four clinical strains had a similar \(pmp\) expression profile (Fig. 3.2), which showed decreasing mRNA levels after the expression peak to 48 h. mRNA levels were detected at 2 h (pi) for all \(pmp\)s, although the scale limits visualization. Overall, \(pmp\)s peaked at 18 h for L2 (Fig. 3.2A) and E/CS-500-96 (Fig. 3.2D), and at 36 h for clinical strains E/537C-05 (Fig. 3.2B) and E/S-141 (Fig. 3.2C). Similar to prototype strains E and L2, \(pmpA\) was the first up-regulated gene for all clinical strains. In addition, \(pmpD\) was the last up-regulated gene for
Clinical strains L2 and E/CS-500-96 (Figs. 3.2A & 3.2D), and was also expressed late in development (together with other pmps) for the other two clinical E strains under study.

Figure 3.2 - Expression profile of pmp and ompA genes throughout the development of C. trachomatis clinical strains. Strain L2 (A) shares the sameompA genotype as L2/434, while strains E/537C-05 (B), E/S-141 (C) and E/CS-500-96 (D) share the sameompA genotype as E/Bour. Values represent the mean±SE based on three independent experiments for time points of 2, 6, 12, 18, 24, 36, and 48 h post infection. See methods for details.
pmpF had the highest expression among all of the pmps for the clinical strains (Fig. 3.2). In fact, there was a 27-fold higher expression of pmpF compared with the least expressed gene (pmpD) for L2. For clinical E strains, there was a 19.2- and 22.6-fold higher expression of pmpF compared with the least expressed gene (pmpI) for E/537C-05 and E/CS-500-96 respectively, and a 54.2-fold higher expression than pmpA for E/S-141. Although no relevant dissimilarities were observed for pmpF between clinical L2 and L2/434, there were considerable expression differences among the clinical strains and E/Bour with up to 11.7-fold higher mRNA values for E/CS-500-96 than for E/Bour.

For ompA, mRNA levels peaked at 36 h for E/537C-05 and E/S-141, and at 18 h for E/CS-500-96 and L2, declining thereafter (Fig. 3.2). The most striking example of differential mRNA levels between ompA and pmps occurred for E/S-141, where ompA had a 232.0-fold higher value compared with the least expressed gene, pmpA. However, all clinical strains except E/CS-500-96 had lower ompA expression levels for all time points compared with the corresponding prototype strains.

3.4.4. Genetic analysis of pmpFE operon for C. trachomatis prototype strains and clinical strains

The pmpF, pmpE and pmpG/pmpF IGR sequences for the three clinical E strains were 100% similar to the corresponding E/Bour sequences, while L2 showed 4 nt differences to L2/434 but only for pmpE. Compared to both L2 strains, the four E strains showed 317 (10.2%) nt and 106 (10.3%) amino acid differences for pmpF as well as 56 (1.9%) nt and 21 (20 to L2) (2.1%) amino acid differences for pmpE. For the pmpG/pmpF IGR, which comprises the ~164 bp upstream region of pmpF, there were 5 nt differences between the L2 and the E strains, although none of them fell within the putative promoter region for the pmpFE operon (Fig. 3.3). The putative promoter is located within a 100% conserved stretch of the pmpG/pmpF IGR for both prototype and all clinical strains (Fig. 3.3). The -10 promoter element (TAAAAT) identified in this study was identical to the one that was previously characterized for the L2/434 and D/UW3 ltuB promoter, while the -35 region (TTGCAT) was 100% similar to the hctA promoter of the same chlamydial prototype strains [193].
C. trachomatis pmp Expression

Figure 3.3 - Predicted pmpF promoter sequence for prototype and clinical strains. Sequences are for prototype strains E/Bour and L2/434, and clinical strains E/537C-05, E/S-141, E/CS-500-96, and L2. The predicted transcription promoter for pmpF is located within a 100% conserved region of the pmpG/pmpF IGR, where putative -10 and -35 elements are in blue characters and boxed. Potential A/T spacer region is underlined, and the predicted transcription start site is shown in a larger font below a red asterisk. The putative RBS for pmpF is in orange characters, and the putative RNase E cleavage sites are highlighted in grey. Numbers represent positions relative to the start codon of pmpF (highlighted in yellow). The start codon of pmpG is highlighted in blue.

Analysis of the pmpFE operon sequence revealed several putative hairpin loop structures although the actual RNA folding in those regions functioning as a classic rho-independent type transcriptional terminator [210] cannot be assumed. At least 41 putative RNase E cleavage sites were identified throughout the pmpFE operon, 13 of which were not conserved between L2/434 (and L2) and the four E strains (Fig. 3.4). One of these non-conserved sites involved the pmpF/pmpE IGR, and is specific for the E strains. The search for an RBS revealed a perfect prokaryotic Shine-Dalgarno sequence (AGGAGG) located 17 nts upstream of the start codon of pmpE, which is approximately 3000 bp below the last bp in Figure 3.3 and, therefore, is not shown. This RBS is in close proximity to the above-described putative RNase E cleavage site shared only by the four E strains. However, the best approach for a putative RBS sequence for pmpF has two mismatches when compared with the ones described in the literature, and is unusually distant from the start codon (Fig. 3.3). Two additional putative RNase E cleavage sites, one of which was in close proximity to this RBS, were identified within the pmpF regulatory region (Fig. 3.3).
Figure 3.4 - Distribution/Location of the putative RNase E cleavage sites within the *pmpFE* operon coding sequence. The sequence is for prototype strains E/Bour and L2/434 and for clinical strains E/537C-05, E/S-141, E/CS-500-96 and L2. Black vertical lines represent all RNase E cleavage sites conserved among all strains under study; green vertical lines show the ones only conserved among the four E strains; orange vertical lines represent those specific solely for both L2 strains. Numbers represent nt positions relative to the start codon of *pmpF*.

3.4.5. **Immunoreactivity of patient sera with Pmp fusion proteins**

Table 3.2 shows the clinical and microbiologic characteristics of the 39 adolescents enrolled in the study and the results of their serum immunoreactivity to rPmpD and rPmpF. All sera from patients infected with chlamydial clinical strains Ba, E, F and K (n = 15; 57.7%), but none with D, Ia, J or G (n = 11; 42.3%), were reactive to rPmpD while sera from uninfected patients were nonreactive with rPmpD (P = 0.001).

<table>
<thead>
<tr>
<th><em>ompA</em> genotype&lt;sup&gt;a&lt;/sup&gt; (n)</th>
<th>Clinical diagnosis&lt;sup&gt;b&lt;/sup&gt; (n)</th>
<th>Immunoreactivity of sera against recombinant fusion proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ba (3)</td>
<td>Cervicitis</td>
<td>rPmpD (%)&lt;br&gt;3/3 (100) rPmpF (%)&lt;br&gt;0/3 (0)</td>
</tr>
<tr>
<td>D (3)</td>
<td>Cervicitis Discharge&lt;sup&gt;c&lt;/sup&gt; (1/3)</td>
<td>0/3 (0) rPmpD (%)&lt;br&gt;0/3 (0) rPmpF (%)&lt;br&gt;0/3 (0)</td>
</tr>
<tr>
<td>E (8)</td>
<td>Cervicitis</td>
<td>rPmpD (%)&lt;br&gt;8/8 (100) rPmpF (%)&lt;br&gt;0/8 (0)</td>
</tr>
<tr>
<td>F (5)</td>
<td>Cervicitis Discharge&lt;sup&gt;c&lt;/sup&gt; (4/5)</td>
<td>1/5 (20) rPmpD (%)&lt;br&gt;0/5 (0) rPmpF (%)&lt;br&gt;0/5 (0)</td>
</tr>
<tr>
<td>G (1)</td>
<td>Cervicitis Discharge&lt;sup&gt;c&lt;/sup&gt; (1/1)</td>
<td>0/1 (0) rPmpD (%)&lt;br&gt;0/1 (0) rPmpF (%)&lt;br&gt;0/1 (0)</td>
</tr>
<tr>
<td>Ia (1)</td>
<td>Cervicitis</td>
<td>rPmpD (%)&lt;br&gt;1/1 (100) rPmpF (%)&lt;br&gt;0/1 (0)</td>
</tr>
<tr>
<td>J (2)</td>
<td>Cervicitis</td>
<td>rPmpD (%)&lt;br&gt;1/2 (50) rPmpF (%)&lt;br&gt;0/2 (0)</td>
</tr>
<tr>
<td>K (3)</td>
<td>Cervicitis</td>
<td>rPmpD (%)&lt;br&gt;3/3 (100) rPmpF (%)&lt;br&gt;0/3 (0)</td>
</tr>
<tr>
<td>Uninfected (13)</td>
<td>No clinical signs or symptoms</td>
<td>rPmpD (%)&lt;br&gt;0/13 (0) rPmpF (%)&lt;br&gt;0/13 (0)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Patients were adolescents 14–19 years of age who had a *C. trachomatis* infection with only one *ompA* genotype as described in methods or were uninfected;

<sup>b</sup> The diagnosis of cervicitis was based on physical findings consistent with cervicitis as determined by the examining physician; all adolescents infected with *C. trachomatis* had cervicitis, and none of these patients complained of any symptoms;

<sup>c</sup> A cervical discharge was noted by the examining physician; none of these patients had clinical signs or symptoms consistent with upper genital tract disease.
Figure 3.5 shows the immunoblot results of representative sera from patients infected with Ba, D, E, F, G, Ia, J and K to rPmpD. Because pmpD is highly conserved among all prototype strains [121], constructing rPmpD using the pmpD sequence of prototype strain E/Bour should not have contributed to the observed differences in immunoreactivity. Further, cross-reactivity between strains was unlikely since the patients were infected with only a single strain, and sera that were reactive to rPmpD were not reactive to rPmpF. In our previous study, sera from the same individuals infected with clinical strains D, E and G reacted with rPmpC [127]. Surprisingly, none of the sera reacted to rPmpF (Fig. 3.5), not even sera from the eight patients infected with strain E, although all sera from infected patients and one uninfected patient reacted with rMOMP as previously shown [127].

![Image](image.png)

**Figure 3.5** - Dot-Blot of serum immunoreactivity against rPmpD and rPmpF. Sera were obtained from adolescents singly infected and uninfected with a different *C. trachomatis* clinical strain as described previously [127] (see also methods). rPmpD and rPmpF concentrations were standardized for use on the blots. Immunoreactivity to each fusion protein for sera from patients infected with strain Ba (n = 3), D (n = 3), E (n = 8), F (n = 5), G (n = 1), Ia (n = 1), J (n = 2) or K (n = 3) are shown. Of note is that immunoreactivity was consistent for sera from patients infected with the same clinical strain except for strain F (Table 3.2); all eight patients infected with strain E were reactive to rPmpD.

### 3.5. Discussion

In this study, we determined the gene expression profile of the nine *pmps* throughout development for prototype strains L2/434 and E/Bour, and four clinical strains belonging to *ompA* genotypes E and L2. The prototype strains had significant gene expression differences where E/Bour had relatively lower mRNA levels and generally sustained expression from 24 to 48 h compared with L2/434 (Fig. 3.1). Surprisingly, in contrast to clinical L2, the three clinical E strains showed a dissimilar *pmp* expression profile to E/Bour (Fig. 3.2). These remarkable expression dissimilarities are generally supported by our recent comparative genomics findings where the laboratory adapted prototype strains did not reflect the same
genetic make-up of strains that are circulating among human populations today and currently exposed to immune selection [114].

It is well known that the developmental stages for E strains occur at later time-points than for L2 strains [170]. This is supported by the *ompA* expression for E/Bour, which is shifted ~6 h later than for L2/434. However, the differential growth rate between these two prototype strains does not explain the dissimilar *pmp* expression, as most *pmps* were up-regulated at the same time point for both (Fig. 3.1). For E/Bour, almost all of the *pmps* had increased expression during the second half of development with comparable mRNA levels at these stages, suggesting a similar involvement in RB division and RB to EB transformation. For L2/434, although most *pmps* showed a general up-regulation of transcription at the exponential growth phase of RB division when new membranes are being formed, *pmpC*, *pmpE* and *pmpF* appeared to play a more important role during this phase. Thus, the gene expression results of both E/Bour and L2/434 suggest their potential importance in membrane integrity. However, some Pmps may have more specific functions than others, depending on the chlamydial strain. In support of this, a proteomics study by Shaw *et al.* [120] detected five Pmps among prototype strains A/Har13, D/UW3, and L2/434, where PmpF was only detected for L2/434. In another proteomics study, only *pmpE*, *pmpG*, and *pmpH* were detected for L2/434 [189]. However, it is possible that these studies reflect a lack of sensitivity in detecting Pmps since a recent study was able to detect all Pmps for L2/434 [119].

Interestingly, *pmpA* had, in general, the lowest expression levels of all *pmps* at each time point except that it had one of the highest levels at 12 h (pi) (Figs. 3.1 & 3.2), suggesting a greater importance of PmpA during early stages of development. This is supported by shotgun proteomics where Skipp *et al.* [119] identified PmpA exclusively in RBs, whereas all other Pmps were detected in both RBs and EBs for L2/434. Additionally, for PmpD, the late up-regulation at 36 h corresponds to RB transformation into EBs, suggesting a role in EB outer membrane structure. In support of this, PmpD has a cysteine content considerably higher than any other Pmp [121]. There are 26 conserved cysteine (Cys) residues in PmpD for all 18 *C. trachomatis* prototype strains, while the mean for all other Pmps is only 13.9 [SE 2.3]. Cysteine residues are responsible for the highly disulfide cross-linked proteins of the outer membrane complex of EBs. Previous studies found that PmpD is surface located and cross-linked in the chlamydial outer membrane complex through disulfide bonds [123]. Furthermore, the N-terminal domain of *C. pneumoniae* Pmp21, the *C. trachomatis* PmpD ortholog, was shown to be non-covalently bound to other components of the EB surface [61]. Additionally, PmpD has shown species-specific neutralizing activity [62]. These collective data are supported by our findings that sera from *C. trachomatis* infected patients were
reactive to rPmpD (Fig. 3.5). Our results were remarkably consistent for sera from patients infected with the same strain. For example, sera from all eight patients infected with strain E were reactive as were sera from three patients infected with strain Ba and three infected with strain K, although only one of the five patients with strain F were reactive; none of patients with strains D, Ia, J or G were reactive. Additional research is required to determine epitopes on PmpD that may correlate with the differential immune responses we observed.

Overall, considering both prototype and clinical strains, pmpA and pmpI were the least expressed genes, while pmpF was the most highly expressed, although pmpE and pmpG also had similar expression levels for E/Bour. We previously found that PmpF is the most polymorphic protein among the C. trachomatis Pmps for both prototype and clinical strains [114,121]. Consistent with the observed protein diversity, phylogenetic analyses of PmpF grouped C. trachomatis strains by tissue tropism properties [121]. Further, comparative analyses of PmpF reveal distinct domains that may be associated with a specific disease group.

The outer membrane exposure of the N-terminus has been experimentally demonstrated for some C. pneumoniae Pmps [61,124], suggesting that these proteins may be subjected to host immune pressure. The N-terminal half for C. trachomatis PmpF also contains numerous nonsynonymous amino acid changes at locations of predicted MHC epitopes [5], indicating that it may be involved in eliciting a cellular immune response. Our findings that none of the sera from infected patients reacted with rPmpF suggest that this protein is not associated with the humoral immune response. Strain origin (E/Bour) of rPmpF did not seem to be an issue as sera from the eight patients infected with strain E were nonreactive. Furthermore, sensitivity was unlikely to be an issue given the immunoreactivity of the same sera with rPmpC and rMOMP, as we have previously described [127], and with rPmpD in this study. The occurrence of highly repeated GGAI motifs in the N-terminus suggests that Pmps may be associated with cell adhesion [117], which has been reported for Pmp21 of C. pneumoniae [61]. These cumulative findings suggest that Pmps are expressed with a differential immune response for patients infected with a specific strain. These findings and the remarkable pmpF expression dissimilarities among L2/434, E/Bour and the clinical strains suggest that there may be differential biological functions across strains and within the same strain for PmpF, either as a structural component to maintain membrane integrity, as part of a large pool of polymorphic antigens to elicit cellular immunity, or as an adhesin.

In our study, the pmpF sequences for the three clinical E strains were found to be 100% similar to the E/Bour sequence as was L2 to L2/434. Since it is highly unlikely that identical proteins have diverse functions, we hypothesized that there may be differential
regulation at the promoter level or regulation involving variations in mRNA processing and/or degradation, which would yield distinct mRNA amounts according to strain-specific needs. It is well known that point mutations in regulatory regions, such as promoter regions and RBS, can affect transcription and translation levels. However, analysis of the putative promoter region and RBS for \( pmpF \) showed that they are 100% conserved for both prototype and the clinical strains (Fig. 3.3), suggesting that the observed \( pmpF \) expression heterogeneity may result from variations in mRNA processing and/or degradation. In fact, regulatory systems of gene expression acting at both the transcriptional and translational levels are well represented in the chlamydial genome, including homologues of endoribonucleases E, III, G and P, exoribonucleases II and PNPase, and oligoribonuclease [2,5]. These are known to control mRNA stability and processing as well as translational efficiency in other bacteria, such as \( E. coli \) and \( Staphylococcus aureus \) [205,207,212-214]. The susceptibility of mRNA to ribonuclease attack may be influenced by events occurring not only at any stage during ribosome binding, but also during translation elongation or termination [201].

We identified two conserved putative RNase E cleavage sites in the \( pmpG/pmpF \) IGR, one of which is in close proximity to the putative RBS (Fig. 3.3). It is known that RBS sequence variability and sequestering by competitive regulatory proteins or conformational impediments can affect ribosome binding/loading and, thus, mRNA lifetime [201]. Considering this, a hypothetical initial cleavage by RNase E could reduce the affinity of the \( pmpF \) translation initiation region for ribosomes, thereby allowing subsequent mRNA degradation/processing by endo- and exonucleases, preferentially for E/Bour when compared to the other strains. A similar regulation has already been reported for \( sodB \) mRNA of \( E. coli \) at low iron concentrations [215]. However, this hypothetical mechanism, although possible, is speculative and lacks experimental evidence.

\( pmpF \) and \( pmpE \) belong to the same operon, yet had remarkably dissimilar mRNA levels for L2/434, and more so for all clinical strains with up to 8.4-fold higher expression for \( pmpF \) than for \( pmpE \) (Figs. 3.1A & 3.2). This did not occur for the \( pmpGH \) operon. We speculated that the expression heterogeneity within the \( pmpFE \) operon may be generated by premature termination of transcription, rapid mRNA processing, or mRNA degradation primarily of the downstream gene (\( pmpE \)) of this large operon transcript (>6 Kb). Similar regulatory mechanisms have already been suggested to explain the existence of multiple transcripts within other bacterial policistronic operons [213], such as those of \( B. subtilis ara \) [216], \( Nitrosomonas europaea cbb \) [217], and \( Borrelia burgdorferi ospAB \) and \( bmpAB \) [218,219].
Although we cannot assume that the putative stem-loop structures found within the \textit{pmpFE} operon sequence may act as classic rho-independent type transcriptional terminators [210], the possibility of hairpin formation (a common phenomenon in mRNA, mainly on large transcripts) cannot be ignored nor can its hypothetical processing role in mRNA degradation and maturation be discounted. Furthermore, several putative RNase E cleavage sites were identified throughout the \textit{pmpFE} operon (Fig. 3.4), which is expected for policistronic operons, although it is well known that RNase E cleaves mRNA only at a limited number of sites [213]. Interestingly, some of the RNase E sites were not conserved between L2 and E strains, suggesting that targeted mRNA degradation or rapid processing events may occur in this large transcript. Curiously, one of these nonconserved recognition sites involved solely the \textit{pmpF/pmpE} IGR of the four E strains. Thus, if RNase E uses this cleavage site, subsequent degradation or processing events from this point would only occur for E strains and could hypothetically yield an mRNA decay of \textit{pmpE}. Yet, as above, this mechanism is speculative and lacks experimental evidence. However, in a previous study, differential transcript quantities were reported for the MMSO genes of \textit{E. coli} that contained a consensus RNase E cleavage site in the intergenic regions of the operon, suggesting complex mRNA processing [220].

Overall, the heterogeneous expression levels among \textit{pmps} and among strains highlight the importance of this gene family in chlamydial biology. In particular, the unique expression disparity for the \textit{pmpFE} operon with relatively high \textit{pmpF} mRNA levels for 5 of the 6 strains under study, as well as the differential immunoreactivity of patient sera to rPmpD, suggest that some Pmps may explain phenotypic differences among strains for antigenicity, virulence and tissue tropism. Furthermore, our findings that clinical E strains do not correlate with prototype strain E/Bour at the gene expression level are supported by our previously reported genomic data [114], reinforcing the need to examine clinical along with prototype strains to advance our understanding of the role of \textit{pmps} in chlamydial biology and disease pathogenesis.
CHAPTER 4

Lymphogranuloma Venereum in Portugal: Unusual Events and New Variants During 2007

Published in


Author Contributions

JPG and MJB conceived and designed the experiments; IS and JA contributed with patient population and clinical data; AN, CF, MAF, and MJB performed the experiments; AN, JPG and MJB analyzed the data; MJB wrote the paper; AN and JPG reviewed the paper.

4.1. Abstract

Several European countries identified an ongoing LGV outbreak, particularly among MSM. In Portugal, no particular surveillance measures were launched. Nonetheless, circulating LGV strains could eventually be detected through the routine *C. trachomatis* *ompA* genotyping procedure held in the Portuguese National Institute of Health (NIH). During 2007, 178 *C. trachomatis* specimens were genotyped through amplification and automated-sequencing of *ompA*. Sequences of 891bp (nt142-nt1032) were aligned with currently available chlamydial sequences from GenBank to identify the corresponding genotype. Eight *C. trachomatis* specimens matched LGV genotypes (7 ‘L2’ and 1 mixed E+L2 undetermined variant). These specimens were identified in samples collected from 4 women and 4 men. One HIV(+) MSM presented LGV related symptoms, while the other infected persons were either asymptomatic or presented no clear LGV symptoms. All samples revealed *ompA* sequences different from the L2/434 prototype strain and from the L2b/144276, which is the most frequently described genotype during the recent LGV outbreak. The detection of 7 LGV specimens during 2007 is in contrast with their absence over the previous 5 years. The LGV infected individuals do not seem to be related to any sexual networks of MSM, contrarily to those described in other European countries. Moreover, all Lisbon LGV specimens revealed unusual *ompA* sequences that differentiate them from the currently reported LGV infections in Europe. The results of the current study further justify an attentive surveillance of LGV strains infecting different populations and the study of their relation with clinical aspects and disease patterns.

4.2. Introduction

The *Chlamydia trachomatis* variety that causes LGV has always been significant in Africa, Southeast Asia, Central, and South America, and in the Caribbean [221]; however, it was rarely reported in industrialized countries. The first recent LGV outbreak in Europe (The Netherlands-Rotterdam) among MSM [222] justified an international alert launched by the European Surveillance of Sexually Transmitted Infections (ESSTI) network in 2004. Subsequently, there have been epidemiologic investigations on LGV disease in some European countries, namely in The Netherlands [222,223], the United Kingdom [224,225],
Germany [226], Sweden [227], and France [228]. In Portugal, no particular epidemiologic surveillance measures were taken to perceive a potential LGV disease outbreak. Nevertheless, all C. trachomatis specimens routinely detected in the NIH of Portugal are systematically genotyped, which potentially enables the detection of Portuguese circulating LGV strains.

The current study reports the unusual number of LGV specimens observed during 2007, and analyses clinical and epidemiologic data of LGV infected individuals. The ompA sequence of LGV specimens was compared with both L2/434 prototype strain and L2 variants described in literature.

4.3. Materials and methods

The Portuguese NIH not only acts as a research and prototype laboratory but also provides routine laboratory diagnosis as a service to the community. The NIH Chlamydia laboratory has been providing (since 1990) routine laboratory diagnosis of bacterial (and also Trichomonas vaginalis) urogenital infections (in particular the sexually transmitted) to attendees of general practitioners, dermatologists, gynaecologists, obstetricians, and urologists from the Lisbon area, according to their request. Sexual orientation data are always absent. Also, although clinical data are not usually provided, most patients refer to themselves as asymptomatic. On the other hand, our laboratory collaborates (for more than 20 years) with the major Portuguese STD clinic (Lapa Health Centre, Lisbon) providing laboratory diagnosis to its attendees (all systematically screened for C. trachomatis, constituting ~60% of our samples), comprising equally symptomatic and asymptomatic individuals. In this STD clinic, since 2007, MSM reporting unprotected anal intercourse have a rectal swab collection for a preliminary evaluation of LGV genotype carriage among Portuguese MSM.

During the year 2007, C. trachomatis infection was tested over 2482 biologic samples (1824 urines, 607 cervical, urethral or lesion/ulcer exudates, and 51 rectal exudates), collected from 1031 men and 1451 women. The 51 rectal swabs (from asymptomatic men) were taken from receptive MSM at the STD clinic of the Lapa Health Centre.

Briefly, each sample was analyzed for the presence of C. trachomatis by the Cobas-Amplicor PCR method (Roche) according to manufacturer’s instructions, and 236 (9.5%) samples yielded a positive result. DNA was extracted from 3 ml of C. trachomatis-positive first void urine or from 500 μl of C. trachomatis-positive exudates transport media (Cobas-Amplicor collection kit or SPG buffer) using the QIAamp DNA Mini Kit (Qiagen) according to the manufacturer’s instructions. The final elution was performed using 200 μl of 10 mMTris, pH 8.5, and DNA was stored at -20°C. C. trachomatis genotype
determination was achieved by amplification and automated-sequencing of the \textit{ompA} gene as previously described [127,167]. \textit{ompA} gene sequences of 891bp (nt142-nt1032) were aligned with currently available chlamydial sequences from GenBank using LaserGene software (DNASTAR) to identify the corresponding \textit{ompA} genotype of each clinical strain.

4.4. Results

A total of 178/236 (75.4\%) \textit{C. trachomatis} specimens could be \textit{ompA} genotyped and 8/178 (4.5\%) matched LGV-genotypes (7 ‘L2’ and 1 mixed E+L2 undetermined variant). LGV genotypes were equally observed in men and women. In fact, these 8 specimens were obtained from 2 urines and 2 cervical samples collected from 4 women [aged 16–34 years; all heterosexual and HIV(-)], and from 3 urines and 1 rectal ulcer sample collected from 4 men [38–52 years; 2 HIV(+) MSM and 2 heterosexual HIV(-)] (Table 4.1). One HIV(+) MSM presented LGV related symptoms (proctitis and rectal lesions), while the other infected persons were either asymptomatic or presented no clear LGV symptoms (one man with confirmed herpetic penile ulcers).

\textbf{Table 4.1.} LGV infection cases detected in the Portuguese NIH during 2007

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Gender</th>
<th>Clinical data</th>
<th>Sexual status</th>
<th>Biological sample</th>
<th>Other STD</th>
<th>Nucleotide differences to</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS-182/07</td>
<td>38</td>
<td>M</td>
<td>Asymptomatic</td>
<td>MSM</td>
<td>Urine</td>
<td>Syphilis (1992), HIV</td>
<td>1 1 2</td>
</tr>
<tr>
<td>CS-183/07</td>
<td>39</td>
<td>F</td>
<td>Asymptomatic</td>
<td>Heterosexual</td>
<td>Urine</td>
<td>Syphilis (2006), \textit{N. gonorrhoeae}</td>
<td>1 1 2</td>
</tr>
<tr>
<td>CS-185/07</td>
<td>42</td>
<td>M</td>
<td>Asymptomatic</td>
<td>Heterosexual</td>
<td>Urine</td>
<td>None</td>
<td>1 1 2</td>
</tr>
<tr>
<td>128C/07</td>
<td>20</td>
<td>F</td>
<td>Asymptomatic</td>
<td>Heterosexual</td>
<td>Cervix</td>
<td>Unknown</td>
<td>1 1 2</td>
</tr>
<tr>
<td>134C/07</td>
<td>30</td>
<td>F</td>
<td>Asymptomatic</td>
<td>Heterosexual</td>
<td>Cervix</td>
<td>Unknown</td>
<td>1 1 2</td>
</tr>
<tr>
<td>614C/07</td>
<td>52</td>
<td>M</td>
<td>Proctitis, Anal lesions</td>
<td>MSM</td>
<td>Anorectal lesions</td>
<td>HIV, Syphilis</td>
<td>3 1 2</td>
</tr>
<tr>
<td>CS-698/07</td>
<td>49</td>
<td>M</td>
<td>Herpetic Penile lesion</td>
<td>Heterosexual</td>
<td>Urine</td>
<td>Unknown</td>
<td>E-L2*</td>
</tr>
<tr>
<td>CS-909/07</td>
<td>16</td>
<td>F</td>
<td>Asymptomatic</td>
<td>Heterosexual</td>
<td>Urine</td>
<td>Unknown</td>
<td>1 1 2</td>
</tr>
</tbody>
</table>

* Undetermined variant.

Only 3 out of the 8 LGV-infected individuals evidenced concurrent sexually transmitted infections (STIs) (Table 4.1). In fact: \textit{i}) 1 woman was additionally infected with \textit{Neisseria gonorrhoeae} (also detected in her partner that was \textit{C. trachomatis} negative) and suffered a syphilis episode during 2006; \textit{ii}) 1 HIV(+) MSM reported having had syphilis 15 years ago and reported more than 10 partners during the last 6 months, and \textit{iii}) 1 MSM was co-infected with \textit{Treponema pallidum} and reported more than 4 partners in the last 6 months.
All 8 LGV samples revealed *ompA* sequences different from the L2/434 prototype strain and from the L2b/144276, which is the most frequently reported LGV genotype, and seemingly is circulating among MSM since the 1980s [226,228-232]. In fact, 6 specimens presented 1 nt difference regarding the L2/434 (C->G at nt position 470), shared by the L2b/144276. However, the L2b/144276 mutation (to L2/434) (A->G at nt position 485) that encodes an amino acid change (Asn->Ser at amino acid position 162) [232] was not present in those specimens (Fig. 4.1). The other LGV specimen shared the 2 *ompA* mutations that characterize the L2b/144276 strain, but presents an additional mutation to both L2/434 and L2b/144276 (C->T at nt position 517) encoding 1 amino acid change, Leu->Phe at position 173. Curiously, this is the strain infecting the single individual presenting clinical signs of LGV disease.

**Figure 4.1** - Partial *ompA* and MOMP sequences of the two L2 Portuguese 2007 variants. L2 variants (in bold) (GenBank: EU676180 & EU676181) are aligned against strains L2/434 (GenBank: M14738), L2b/144276 (GenBank: DQ217607), and L2c/Pt5 (GenBank: EF460796). The *ompA* gene of variant 128C/07 is 100% similar to the remainder 5 L2 variants (CS-182/07, CS-183/07, CS-185/07, CS-909/07, and 134C/07). Boxed regions comprehend variable domain II (VDII). Highlighted in grey are the nucleotide and amino acid differences to prototype strain L2/434. The peptide region highlighted in black represents a known antigenic domain for L2. Numbers on top of nucleotide and amino acid sequences of L2/434 MOMP represent positions relative to the start codon of *ompA* and first methionine, respectively.

The screening of LGV genotypes carriage among 51 Portuguese MSM detected 11.8% (6/51) positive *C. trachomatis* samples. Five of these specimens were *ompA* genotyped (*ompA* could not be amplified for 1 specimen), none revealed sequences matching L1-L3 genotypes, and common urogenital genotypes were found: Da (1), E (1), F (1), and G (2 samples collected from nonrelated men).
4.5. Discussion

From 1991 to 2002, *C. trachomatis* infection was evaluated on ~13,800 biologic samples, yielding 596 positive results. These *C. trachomatis* specimens were further genotyped in the Portuguese NIH through PCR amplification of *ompA*, and subsequent analysis of the restriction profiles of amplification products, digested mostly by *AluI*, *EcoRI*, and *DdeI* restriction enzymes [233]. During this period, only 1.1% (5/463) of the whole typed specimens (n=463) matched LGV genotypes, corresponding to 5 putative L2 strains (4 L2 and 1 mixed E+L2 infections). On present days, the lack of DNA from these early L2 genotypes discards further confirmation by sequencing. The specimens were identified from 2 men (1 urine sample and 1 urethral swab) and 3 women (cervical swabs) with no clinical signs of LGV disease.

As the restriction pattern did not enable discrimination of all genotypes, our laboratory changed the *C. trachomatis* *ompA* typing procedure during 2002. Consequently, all *C. trachomatis ompA* PCR positive products detected since 2002 have been typed through automated-sequencing followed by the alignment of the obtained sequences against *ompA* sequences of prototype strains, as described in the Materials and Methods section. No *C. trachomatis* LGV-genotypes were detected between 2002 and 2006 (~9500 samples tested, 674 positive for *C. trachomatis*, 0 LGV over 503 typed specimens). Surprisingly, during the year 2007 (Table 4.1), 4.5% of the typed specimens matched LGV-genotypes (7 ‘L2’ and 1 mixed E+L2 undetermined variant). Curiously, LGV genotypes were observed equally in men and women, which constitute a unique scenario in the recent LGV outbreak and suggest the need for further evaluation of gender as an independent risk factor to acquire LGV genotypes. However, we did not find any suggestive factors for an ongoing outbreak in Portugal; although the LGV cases were geographically clustered (Lisbon), we could not relate them to any social venues or to European locations. Peculiarly, only 1 LGV strain was identified from a MSM rectal ulcer swab.

Other STI, in particular HIV, have been diagnosed in LGV infected patients [227,229], contributing for the assumption of sexual risk behaviors among LGV infected people. Indeed, in our study, 3 out of the 8 individuals evidenced concurrent STI (Table 4.1), and other 3 (CS-185/07, CS-698/07, and CS-909/07) reported sexual risk behaviour (unprotected sex with several partners).

A very recent study held in Austria [234] identified 3 new L2 genetic variants (designated “new strains” L2c, L2d, and L2e), suggesting more than one source for LGV infection in Vienna. Considering our results and this highly recent report, we could name our variants as L2f and L2g, represented in Figure 4.1 by the clinical specimens 128C and 614C,
respectively (GenBank: EU676181 & EU676180, respectively) (Fig. 4.1). However, the likely emergence of new variants implies a reflection on the relevance of these new designations as it may turn into a never-ending process. Interestingly, the L2 variants described to date (a to g) include 3 neutral and 5 nonsilent SNPs, and 4 of this 5 amino acid changes occur inside a VDII core serovar-specific epitope, that has been described as a common antigenic domain for L2 and trachoma strains (Fig. 4.1) [134,135]. We could speculate that these successful L2 variants reflect a bacterial skill for evading the immune system. Likewise, more gene changes might have occurred in the remnant genome [114], comprehending other putative chlamydial antigens that may influence LGV disease outcome, emphasizing the interest of studying the whole genome of clinical specimens. Supporting this, Kari et al. [37] demonstrated that subtle genetic differences resulted in different virulence properties of genotype A trachoma strains, which suggest that the same phenomena may occur in LGV specimens.

Curiously, L1 is less frequently described than L2, and L3 seems to be rare; in fact, it was not described during the recent LGV outbreak [226,228-232]. In our laboratory, L1 and L3 were never detected.

Recent studies suggested that as C. trachomatis detection is not routinely performed on rectal swabs, LGV infections could remain undetected among MSM [229,235], which is further supported by the high percentage of C. trachomatis detection when studying rectal swabs from MSM: 7% [236] or 14.8% [230,231]. Moreover, a study performed over C. trachomatis positive rectal swabs showed that 33.3% of the 982 typeable specimens confirmed to be genotype L2 [229].

Some studies reported a high percentage of LGV genotypes among anorectal chlamydial specimens [229,230,236] but large evaluations held in the United Kingdom and The Netherlands showed respectively 66.7% (of the typeable specimens among symptomatic MSM) and 81.3% (of the typeable specimens among both asymptomatic and symptomatic MSM) anorectal carriage of C. trachomatis by genotypes D–K [229,231]. So, apparently, anal intercourse spreads exactly the same genotypes that are transmitted through vaginal intercourse, and not distinctively LGV genotypes. However, Klint et al. [227] proposed that the distribution of D–K genotypes isolated from anorectal samples would be unusual, with a special prevalence of genotypes G and D in opposition to the typical E and F predominance. They also suggested that genetic variants would be more frequent among anorectal isolates [227]. Our preliminary study concerning 51 MSM only determined 5 D–K genotypes, which unable further speculations about changes on Portuguese C. trachomatis genotype distribution among anorectal specimens in comparison with urogenital specimens. Curiously, we observed
the G and Da *ompA* genetic variants described by Klint *et al.* [227] (identical to G1/JS4481 and to D1/IS5643) in anorectal samples; however, the same variants have already been observed among urogenital specimens in our laboratory (personal observation). Thus, further evaluations are needed to confirm and understand the genetic basis and the significance of the observation of these variants.

In Portugal, there is no laboratory based surveillance system for STI and LGV disease is no longer a mandatory reportable disease; consequently, the 2007 Portuguese LGV infection episodes passed unnoticed for national health authorities that could not launch any specific surveillance/control measures or implement information campaigns concerning LGV. To our knowledge, the LGV infected individuals reported in this study are not related to any sexual networks of MSM, contrarily to the generally described in other European countries. All Lisbon LGV specimens revealed unusual *ompA* sequences that differentiate them from the currently reported LGV infections in Europe. Globally, the majority of the amino acid changes on the L2 variants described in this study and in the other literature reports occurs in a well-defined antigenic domain, strongly suggesting a chlamydial strategy for host immune evasion.

All together, these results further justify an attentive surveillance of LGV strains infecting different populations and the study of their relation with clinical aspects and disease patterns.
CHAPTER 5

Evolutionary Dynamics of *ompA*, the Gene Encoding the *C. trachomatis* Key Antigen

*Published in*


*Author Contributions*

AN, JPG and MJB conceived and designed the experiments; AN, CF, and MJB performed the experiments; AN and BN performed the statistics; AN and JPG analyzed the data; AN and JPG wrote the paper; MJB reviewed the paper.
5. Evolutionary Dynamics of ompA, the Gene Encoding the C. trachomatis Key Antigen

5.1. Abstract

*Chlamydia trachomatis* is the trachoma agent and causes most bacterial sexually transmitted infections worldwide. Its major outer membrane protein (MOMP) is a well-known porin and adhesin and is the dominant antigen. So far, investigation of MOMP variability has been focused mainly on molecular epidemiological surveys. In contrast, we aimed to evaluate the impact of the host pressure on this key antigen by analyzing its evolutionary dynamics in 795 isolates from ano-urogenital infections, taking into account the MOMP secondary structure and the sizes/positions of antigenic regions. One-third of the specimens showed a mutational drift from the corresponding genotype, where ~42% of the mutations had never been described. Amino acid alterations were six-fold more frequent within B-cell epitopes than in the remaining protein \((P = 0.027)\), and some mutations were also found within or close to T-cell antigenic clusters. Interestingly, the two most ecologically successful genotypes, E and F, showed a mutation rate 60.3-fold lower than that of the other genotypes \((P < 10^{-8})\), suggesting that their efficacy may be the result of a better fitness in dealing with the host immune system rather than of specific virulence factors. Furthermore, the variability exhibited by some genetic variants involved residues that are known to play a critical role during the membrane mechanical movements, contributing to a more stable and flexible porin conformation, which suggests some plasticity to deal with environmental pressure. Globally, these MOMP mutational trends yielded no mosaic structures or important phylogenetic changes, but instead yielded point mutations on specific protein domains, which may enhance pathogen’s infectivity, persistence, and transmission.

5.2. Introduction

*Chlamydia trachomatis* is an obligate intracellular human pathogen that can be classified into 18 variants based on the immunoreactivity of MOMP or *ompA* (which encodes MOMP) polymorphism: genotypes A to C and Ba are responsible for trachoma; genotypes D to K, Da, Ia, and Ja are normally associated with infection of the urogenital tract; and genotypes L1 to L3 cause LGV [4]. This preference for particular cell types is not exclusive, and therefore ocular strains can occasionally be found in the urogenital tract and vice versa. However, it is
thought that only L1 to L3 strains possess the ability to invade the inguinal lymph nodes. MOMP has been implicated in the mechanisms of attachment, infection, and/or pathogenesis due to its variability, surface exposure, and antigenic properties. Previous studies have shown that MOMP may act as a putative cytoadhesin by promoting nonspecific interactions with host cells [55]. This major chlamydial membrane component, which constitutes about 60% of the membrane dry weight [133], is also thought to play a role in maintaining structural integrity of the organism [133,237] by forming a trimeric structure [238]. Also, during chlamydial replication, MOMP may act as a porin [239] that is folded into a β-barrel structure containing five constant domains (CDI to CDV) of transmembrane β-strands and periplasmic turns and four highly variable surface-exposed domains (VDI to VDIV) [240-242]. Furthermore, MOMP possesses species- and serovar-specific epitopes [135,243-245] that are able to elicit both the humoral (B-cell mediated through the production of antibodies) and cellular (T-cell mediated and also influencing the B-cell response) immune responses, making this dominant chlamydial antigen a potential candidate for the development of vaccines and therapeutic strategies [185,246-248]. Indeed, although no efficacious chlamydial vaccine has been developed so far, the use of inactivated or live-attenuated pathogens has been replaced by peptide or subunit vaccines, and MOMP is definitely one of the leading candidates [249].

To improve our knowledge of the effects of the host pressure on MOMP and also of the molecular epidemiology of the circulating C. trachomatis strains, it is imperative to investigate genetic variability in ompA. Here, we performed a sequence-based analysis of the ompA mutational trends in clinical isolates that were collected from patients with sexually transmitted C. trachomatis infections. So far, most studies have been limited to a small number of strains with variations in ompA [184,250-258] or were restricted to the analysis of VDs [112,259-262], discarding the CDs, which contain numerous cytotoxic T lymphocyte (CTL) and T helper (Th) cell epitopes [137-140,263]. We performed a detailed bioinformatic and statistical analysis of the mutational dispersion on both VDs and CDs, based on MOMP structure and on the mapping of all the B- and T-cell epitopes reported in the literature. We present statistically validated genomic evidence of the adaptation of this pathogen’s key antigen to the host pressure, which strongly indicates a strategy to evade the human immune system.
5.3. Materials and methods

5.3.1. Study population and ompA genotyping

From 2001 to 2007, a total of 15,135 first-void urine samples as well as cervical, urethral, and rectal swab samples were collected from patients attending general practice, family planning, and sexually transmitted disease clinics in the Lisbon area, and these samples were analyzed for the presence of *C. trachomatis* at the Portuguese NIH. This assessment was performed by DNA amplification methods (Amplicor PCR and Cobas-Amplicor) (Roche Molecular Systems) as per the manufacturer’s instructions. For the 1,103 clinical specimens (7.3%) that were positive for *C. trachomatis*, DNA was extracted from 3 ml of first-void urine samples or 500 μl of exudate transport medium (Cobas-Amplicor collection kit or SPG buffer) using the QIAamp DNA mini kit (Qiagen) according to the manufacturer’s instructions. Identification of clinical specimens was performed using ompA sequencing-based genotyping with subsequent BLAST comparison with the available GenBank sequences from *C. trachomatis* prototype strains, as previously described [113]. Primers ompA-S (5’-TGCTGAACCAAGCCTTATGA-3’) and sero2A (5’-TTTCTAGATTTTCATTTTTGTT-3’) were used for the automated sequencing in the present survey. To avoid any epidemiological linkage that may lead to biased results, specimens collected from sexual partners during the same year were excluded from this study if they presented the same ompA sequence. All specimens presenting ompA variant sequences were confirmed by resequencing newly extracted DNA in order to overcome any amplification or sequencing artifact. The 18 prototype strains (Table S5.1) that were used as the baseline for genotype comparisons were A/Har13, B/TW5, Ba/Apache2, C/TW3, D/UW3, Da/TW448, E/Bour, F/IC-Cal3, G/UW57, H/UW43, I/UW12, Ia/UW202, J/UW36, Ja/IUA795, K/UW31, L1/440, L2/434, and L3/404. Except for Ja/IUA795, the designation of these strains as prototype strains is consensual. Nevertheless, Ja/IUA795, which was isolated in 1986, has been the Ja strain most used as a prototype. Additionally, other strains also referred to in the literature as prototype strains (Table S5.1) were used for comparative purposes in the present study.

5.3.2. Genomic and phylogenetic analyses

For each ompA genotype, nucleotide and protein alignments of all clinical and prototype strains were generated using the ClustalW method of the LaserGene (DNASTAR) and MEGA 4.0.2 (http://www.megasoftware.net) software. MEGA 4.0.2 was also used to create matrices of pairwise comparisons and to estimate the number of variable sites. For all clinical specimens, the existence of any recombination event in ompA was evaluated using the
SimPlot/BootScan software (http://sray.med.som.jhmi.edu/SCRoftware/), as previously described [114]. Phylogenies were generated using the neighbor-joining (NJ) method [264] with both nucleotide differences and Kimura two-parameter (K2P) [265] or gamma [266] models to estimate evolutionary distances at the nucleotide or protein level, respectively. For all genomic and phylogenetic analyses, the pairwise-deletion option was chosen to remove all sites containing missing data or alignment gaps from all distance estimations, only when the need arose and not prior to the analyses.

5.3.3. Analysis of molecular evolution

We used the Nei-Gojobori method [267] of MEGA 4.0.2 to calculate the overall mean of synonymous (dS) and nonsynonymous (dN) substitution rates for each ompA genotype, as previously described [121]. Given the degeneracy of the genetic code, the p-distance model was used to normalize the computed differences against the number of potential synonymous and nonsynonymous sites.

Specific evaluation of the selective pressure (neutral [dN/dS = 1], purifying [dN/dS < 1], or positive [dN/dS > 1] selection) acting on ompA was performed for the genotypes Da, G, and Ia, as these genotypes presented the highest number of variant clinical specimens (Fig. 5.1). As the selective pressures vary among different regions of the protein, this analysis was performed using a maximum-likelihood model of codon substitution along the phylogeny of ompA [268,269] implemented by the codeml program in the PAML v4 software (http://abacus.gene.ucl.ac.uk/software/paml.html), which allows dN/dS to differ across different codons. Briefly, the distribution of the dN/dS ratio (ω) across sites was estimated using the following codon substitution models [269]: M0 (one ratio), M1a (nearly neutral), M2a (positive selection), M3 (discrete), M7 (beta distribution), and M8 (beta distribution + ω). The existence of amino acid sites under positive selection was tested with a likelihood ratio test by comparing the nested models that differ by only one parameter (the class of sites with ω > 1), i.e., M2a versus M1a and M8 versus M7. The null distribution of nonexistence of sites under positive selection was evaluated by the likelihood ratio test statistic (2Δl, where Δl is the difference between the log-likelihood scores of the two methods) assuming that under null hypothesis the 2Δl follows a χ² distribution where the degrees of freedom are the difference in the number of free parameters between the two methods. The Bayes empirical Bayes (BEB) method was used calculate the posterior probabilities of each site falling into the class of positively selected sites (i.e., class with ω>1)
as well as the posterior distribution of the $\omega$ parameter describing that class. Thus, sites likely under positive selection are the ones with higher probabilities of belonging to this class.

5.3.4. Analysis of protein features

To shed some light on the putative impact of each mutation on the MOMP sequence, the Protean program of LaserGene (DNASTAR) was used to perform a comparative analysis of the protein sequences of all variant and prototype strains for the following features: charge density [271], secondary structure (method of Garnier et al. [272]), hydropathy (methods of Kyte and Doolittle [273], Hopp and Woods [274], and Engelman et al. [275]), antigenicity (Jameson-Wolf [276] and AMPHI [277] methods), amphiphilicity (method of Eisenberg et al. [278]), surface probability (method of Emini et al. [279]), and flexibility (Karplus-Schultz method [280]). Basically, for all these protein features, with the exceptions of the secondary structure and antigenicity, values were assigned for all residues and were then averaged over a sliding window of a specified range of amino acids for each method. For these analyses, the default parameters of each method were used, except for the Kyte-Doolittle and Hopp-Woods methods, where a window size of seven amino acids was used.

5.3.5. Statistical analysis

For each ompA genotype, only the variable sites where clinical specimens presented nucleotide changes to the corresponding prototype strain (Table 5.1) were analyzed. Globally, a total of 80 variable sites were considered, as follows: genotype A, 1 site; B, 15; C, 5; D, 2; Da, 6; E, 5; F, 6; G, 7; H, 4; I, 1; Ia, 14; J, 10; Ja, 2; and L2, 2. To investigate the variability inside both the VDs and CDs of ompA, we evaluated whether the 31 nonsynonymous mutations found in VDs are overrepresented relative to the ones that occurred in CDs. We found that there were 23 nonsynonomous substitutions in CDs, 2 synonymous substitutions in VDs, and 24 synonomous substitutions in CDs.

We also investigated the mutational trend in MOMP B-cell antigenic regions (Fig. S.1) for all genotypes containing variant clinical specimens. To achieve this, we estimated the total length of the MOMP B-cell antigenic regions for each genotype, where all the epitope sequences described for each VD were overlapped and the lengths of the resulting regions were summed. Thereafter, we calculated the mean of the final value for all genotypes, which yielded an antigenic region encompassing a length of 105 bp per ompA (35 amino acids per MOMP). Since it was experimentally demonstrated [134] that a change in the amino acid immediately adjacent to a defined epitope may alter its immunoreactivity, we also considered
all variable sites occurring one amino acid before and after each epitope. We then evaluated whether 20 variable sites restricted to the 105 bp coding for the B-cell antigenic regions are overrepresented relative to the 60 variable sites found in the rest of \textit{ompA}. With a restricted mean gene sequence of 900 bp (the \textit{ompA} sequence length available for all clinical specimens) for each strain, we found 85 nonvariable sites for the B-cell antigenic regions and 735 nonvariable sites for the nonantigenic regions.

Also, given that 18 of the above 20 variable sites occurring in B-cell antigenic regions yielded amino acid replacement, we also evaluated whether these sites restricted to the superimposed antigenic regions are overrepresented relative to the 36 sites found outside that region. We found 2 synonymous sites in the B-cell antigenic regions and 24 synonymous sites in the nonantigenic regions.

For all these statistical analyses, the SPSS Base version 15.0 (SPSS Inc. Chicago) was used to estimate the \( P \) values by Fisher’s exact test as well as the odds ratios with a 95% confidence interval.

5.3.6. Nucleotide sequence accession numbers

The sequence data for all types of genetic variants (Table S5.1) were submitted to GenBank under accession numbers DQ116393, DQ116396 to DQ116398, DQ116400, DQ116402, EU676180, and FJ943511 to FJ943546.

5.4. Results

5.4.1. Analysis of \textit{ompA} genotypes and genetic variants

We initially characterized \textit{C. trachomatis} isolates by analyzing \textit{ompA} nucleotide variation. Of the 15,135 ano-urogenital samples that were received at the Portuguese NIH, 1,103 were positive for \textit{C. trachomatis}, with 795 specimens (493 from women, 294 from men, and 8 for which the gender of the donor was unknown) successfully typed. These represented 15 \textit{ompA} genotypes; Ba, L1, and L3 were not found (Fig. 5.1). Genotype E was the most prevalent type (39.8%), followed by genotypes F (17.3%), G (11.2%), Da (8.2%), Ia (6.2%), D and J (4.9%), H (3.0%), Ja (1.9%), L2 (1.0%), K (0.8%), B and C (0.4%), and finally A and I (0.1%). Globally, 232 clinical specimens (29.2%) of all identified genotypes (with the exception of genotype K) were found to present \textit{ompA} nucleotide changes compared with the respective prototype strain (Fig. 5.1). The later were isolated up to 50 years ago (Table S5.1) and have been used worldwide as prototypes for comparative purposes. The highest number of clinical variants was presented by genotypes Da, G, and Ia, with 65, 54, and 49 strains, respectively.
Moreover, all specimens typed as A, B, C, Da, I, and Ia presented \textit{ompA} sequences that varied from that of the respective prototype strain. Interestingly, for the two most prevalent genotypes, E and F, which together represent almost 57\% of all typed specimens, only 7 out of 317 E strains (2.2\%) and 6 out of 138 F strains (4.4\%) had \textit{ompA} variant sequences. This is intriguing as it suggests that conservation of the major antigen favors ecological success. In support of this divergent \textit{ompA} molecular trend for these genotypes, we have previously shown a mutational dynamic involving the entire chromosome that separated E and F from the remaining genotypes [165].

![Figure 5.1](image.png)

**Figure 5.1** - Distribution of the 795 \textit{C. trachomatis} clinical specimens by \textit{ompA} genotype. The bars represent the absolute numbers of nonvariant (in white) and variant (in gray) specimens.

Overall, 43 types of genetic variants (i.e., strains with a distinct \textit{ompA} mutational pattern) were identified among our 232 variant clinical specimens, involving 14 \textit{ompA} genotypes (Fig. 5.2). Twenty-nine out of the 43 genetic variants have not been reported before, and these represent a total of 49 strains. Curiously, 44.9\% of these belong to genotype Ia. Also of note is that all genetic variants of genotypes A, C, I, and Ja found for the Portuguese population were never described for any other country. No insertion or deletion (indel) event was found in \textit{ompA} for any of our 43 genetic variants compared to the respective prototype strain (data not shown). Furthermore, based on SimPlot/BootScan analyses, none of them revealed any trace of recombination within \textit{ompA}, presenting only slight dissimilarities to the corresponding prototype and/or other prototype strains (data not shown). Supporting this, both the nucleotide (Fig. 5.2) and amino acid (data not shown) phylogenetic analyses showed strain segregation according to three main clusters corresponding to the traditional B, C, and Intermediate \textit{ompA} serogroups with maximum bootstrap values (98 to 100\%). In fact, within each serogroup, closely related clusters were found for all variant genotypes, where genetic distances among strains varied from 0.1\% [SE 0.1\%] for genotypes A, I, and Ja to
0.9% [SE 0.2%] for genotype B (data not shown). Interestingly, all the Da and Ia variants showed higher similarity to prototype strains Da/IU1554 and Ia/IU4168, respectively, than to the corresponding traditional prototype strains (Da/TW448 and Ia/UW202). Altogether, this indicates that in the population studied, \textit{ompA} diversity is strictly a consequence of the occurrence of point mutations rather than recombination or indel events.

### Figure 5.2 - Evolutionary relationship between all \textit{C. trachomatis} prototype strains and the 43 types of genetic variants

The NJ phylogenetic tree is based on the number of nucleotide differences observed among the \textit{ompA} sequences to better illustrate the genetic distances between genetic variants. K2P and gamma models yielded similar topologies. The tree was rooted using the \textit{ompA} sequence from \textit{C. muridarum} strain MoPn. Bootstrap values (1,000 replicates) are shown next to the branch nodes. The traditional \textit{ompA} serogroups (B, C, and Intermediate) are shown above the three major branches of the tree. All types of genetic variants and prototype strains belonging to the same genotype are represented by the same color together with the respective prototype strain. Values in parentheses indicate the number of variant specimens represented by that type of genetic variant and the total number of variant specimens found for the same \textit{ompA} genotype. Types of genetic variants that had never been described before are represented by yellow circles.

#### 5.4.2. Analysis of \textit{ompA} variable sites per genotype

We proceeded with a detailed analysis of \textit{ompA} variability within same-genotype strains. Table 5.1 describes the nucleotide sequence variation in \textit{ompA} for each genotype, showing all mutations relative to the corresponding prototype strain. A total of 80 nucleotide variable sites
were found, where B, Ia, and J presented the highest number of sites (with 15, 14, and 10 changes, respectively), while genotypes A and I were the least variable ones (both with 1 change). Each variable site was represented by no more than two types of nucleotides, with exception of genotype G, which presented a parsimony-informative site, where two different nonsynonymous changes were found for the same variable site at position 1003. Interestingly, for the vast majority of genotypes, variable sites were dispersed throughout \textit{ompA} (Table 5.1), reflecting a general mutational trend on both VDs and CDs. One notable exception occurred for B strains, where 11 of the 15 mutations were located at CDI ($P < 10^{-7}$), suggesting a highly restricted mutational clustering for genotype B. Also interestingly, all six mutations displayed by F genetic variants were found solely in CDs.

Overall, from the 80 variable sites found for all types of genetic variants (Table 5.1), 33 occurred in VDs and 47 occurred in CDs, with CDI encompassing almost half (44.7%) of them. Moreover, 93.9\% (31/33) of the mutations that occurred in VDs yielded amino acid changes, while the reverse was seen for CDs, with 51.1\% (24/47) of the mutations resulting in silent substitutions. In fact, we found a global nonsynonymous mutation rate that was 16.2-fold higher for VDs than for CDs ($P < 10^{-4}$), indicating a nonsynonymous variability targeted mainly on MOMP surface-exposed domains, which are involved in the interaction with the host.

5.4.3. Impact of mutations on MOMP antigenic regions

Considering that well-defined B- and T-cell epitopes were already described in MOMP VDs and CDs (Fig. S.1), we also investigated the specific mutational trend in these antigenic regions for all genotypes presenting variant strains in order to evaluate the host immune pressure on the chlamydial major antigen. Therefore, we constructed a comprehensive “map” of \textit{C. trachomatis} B- and T-cell (CTL and Th) epitopes described in the literature for all 18 prototype strains [134,136-140,263,281-285] (Fig. S.1). B-cell epitopes encompassing full VDs were not considered, as immunoreactivity is generally due to small epitopes inside those large regions [134]. Also, although the Th-stimulatory peptides reported in the literature are known to contain multiple core Th epitopes, only a minority were already mapped and only these were considered for the present study. This conservative approach was used to minimize any bias that could result from considering an overrepresentation of antigenic regions in MOMP, which would substantially favor the statistical analysis.
Table 5.1. Nucleotide sequence variation in *ompA* genotype variants compared with the respective prototype strain

<table>
<thead>
<tr>
<th>Genotype (no. of variants)</th>
<th>No. of variable sites</th>
<th>MOMP region</th>
<th>Nucleotide position&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Amino acid change&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Type of change&lt;sup&gt;c&lt;/sup&gt;</th>
<th>No. of isolates (x/y)</th>
<th>dN/dS&lt;sup&gt;d&lt;/sup&gt; [SE]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (1)</td>
<td>1</td>
<td>CD III</td>
<td>736</td>
<td>Val-&gt;Ile</td>
<td>x x</td>
<td>1/1</td>
<td>---&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>B (3)</td>
<td>15</td>
<td>CD I</td>
<td>154</td>
<td>Thr-&gt;Ile</td>
<td>2/3</td>
<td>0.35</td>
<td>[0.19]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>163/164</td>
<td>Val-&gt;Cys</td>
<td>E 3/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>184/186</td>
<td>Met-&gt;Val</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>195</td>
<td>---</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>198</td>
<td>---</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>216</td>
<td>---</td>
<td>3/3</td>
<td>0.35</td>
<td>[0.19]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>228</td>
<td>---</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>246</td>
<td>---</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>249</td>
<td>---</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>268</td>
<td>Thr-&gt;Ile</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>286/287</td>
<td>Val-&gt;Thr</td>
<td>2/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>514</td>
<td>Ala-&gt;Thr</td>
<td>3/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C (3)</td>
<td>5</td>
<td>CD I</td>
<td>155</td>
<td>Thr-&gt;Ile</td>
<td>x</td>
<td>1/3</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>183</td>
<td>---</td>
<td>x</td>
<td>1/3</td>
<td>[0.51]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>284</td>
<td>Gly-&gt;Asp</td>
<td>x x E</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>693</td>
<td>---</td>
<td>x C</td>
<td>1/3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1003</td>
<td>Ala-&gt;Ser</td>
<td>x</td>
<td>3/3</td>
<td></td>
</tr>
<tr>
<td>D (12)</td>
<td>2</td>
<td>VD II</td>
<td>488</td>
<td>Glu-&gt;Gly</td>
<td>x E</td>
<td>1/12</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>976</td>
<td>Ala-&gt;Thr</td>
<td>x</td>
<td>12/12</td>
<td></td>
</tr>
<tr>
<td>Da (65)</td>
<td>6</td>
<td>VD II</td>
<td>485</td>
<td>Asn-&gt;Ser</td>
<td>11/65</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>636</td>
<td>---</td>
<td>46/65</td>
<td>[0.25]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>939</td>
<td>---</td>
<td>x</td>
<td>61/61</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>976</td>
<td>Ala-&gt;Thr</td>
<td>x</td>
<td>11/58</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>997</td>
<td>Thr-&gt;Ala</td>
<td>x</td>
<td>57/57</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1050</td>
<td>---</td>
<td>x</td>
<td>1/1</td>
<td></td>
</tr>
<tr>
<td>E (7)</td>
<td>5</td>
<td>VD II</td>
<td>512</td>
<td>Asn-&gt;Ser</td>
<td>x x</td>
<td>1/7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>514</td>
<td>Ser-&gt;Ala</td>
<td>x x C</td>
<td>1/7</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>568</td>
<td>Ala-&gt;Thr</td>
<td>x</td>
<td>2/7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>995</td>
<td>Ser-&gt;Asn</td>
<td>x x</td>
<td>1/7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>997</td>
<td>Ala-&gt;Thr</td>
<td>x</td>
<td>2/7</td>
<td></td>
</tr>
<tr>
<td>F (6)</td>
<td>6</td>
<td>CD I</td>
<td>167</td>
<td>Asp-&gt;Gly</td>
<td>x E</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>180</td>
<td>Met-&gt;Ile</td>
<td>x</td>
<td>1/6</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>388</td>
<td>Ile-&gt;Val</td>
<td>x</td>
<td>1/6</td>
<td>[1.60]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>703</td>
<td>Phe-&gt;Leu</td>
<td>x</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>780</td>
<td>---</td>
<td>1/6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>802</td>
<td>Asp-&gt;Asn</td>
<td>x P</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>G (54)</td>
<td>7</td>
<td>CD I</td>
<td>228</td>
<td>---</td>
<td>15/54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>374</td>
<td>Cys-&gt;Tyr</td>
<td>x P</td>
<td>1/54</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>487</td>
<td>Gly-&gt;Ser</td>
<td>24/54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>700</td>
<td>Glu-&gt;Gln</td>
<td>P 15/54</td>
<td>2.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>857</td>
<td>Thr-&gt;Ile</td>
<td>x W</td>
<td>1/54</td>
<td>[2.55]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1003</td>
<td>Ser-&gt;Ala</td>
<td>29/49</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1015</td>
<td>Ile-&gt;Val</td>
<td>x</td>
<td>1/45</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Nucleotide position

<sup>b</sup> Amino acid change

<sup>c</sup> Type of change: New B-cell, T-cell, Loop

<sup>d</sup> No. of isolates: x/y

<sup>e</sup> dN/dS: [SE]
<table>
<thead>
<tr>
<th>Genotype</th>
<th>Variable Site</th>
<th>Amino Acid Change</th>
<th>Frequency</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD I</td>
<td>4</td>
<td>Ile-&gt;Val</td>
<td>22/49</td>
<td></td>
</tr>
<tr>
<td>CD II</td>
<td>327</td>
<td>Asn-&gt;Asp</td>
<td>49/49</td>
<td></td>
</tr>
<tr>
<td>CD IV</td>
<td>840</td>
<td>Lys-&gt;Arg</td>
<td>46/46</td>
<td></td>
</tr>
<tr>
<td>CD V</td>
<td>1188</td>
<td>Ala-&gt;Thr</td>
<td>1/1</td>
<td></td>
</tr>
<tr>
<td>J (8)</td>
<td>10</td>
<td>Ile-&gt;Thr</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>CD I</td>
<td>213</td>
<td>Asn-&gt;Ser</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>CD II</td>
<td>369</td>
<td>Thr-&gt;Ala</td>
<td>1/1</td>
<td></td>
</tr>
<tr>
<td>CD III</td>
<td>684</td>
<td>Ala-&gt;Glu</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>CD IV</td>
<td>782</td>
<td>Thr-&gt;Ala</td>
<td>1/7</td>
<td></td>
</tr>
<tr>
<td>Ja (3)</td>
<td>2</td>
<td>Asn-&gt;Asp</td>
<td>2/3</td>
<td></td>
</tr>
<tr>
<td>L2 (2)</td>
<td>2</td>
<td>Asn-&gt;Ser</td>
<td>2/2</td>
<td></td>
</tr>
</tbody>
</table>

*a* Based on the alignment of the strains belonging to the same genotype.

*b* Amino acid change presented by the variant clinical specimen compared with the prototype strain of the respective *ompA* genotype.

*c* Mutations that were observed only for the Portuguese clinical specimens or that occurred within well-defined B-cell epitopes (x), mutations that occurred within (W) or in close proximity (C) to T-cell (CTL and Th) epitopes, and mutations found in MOMP external loops (E) or periplasmic turns (P). For this last type of mutation, only the residues known to be involved in MOMP structural constraints were considered for this analysis.

*d* x represents the total number of clinical specimens sharing the same variable site; y represents the total number of clinical specimens for which the nucleotide sequence involving the variable site is available.

*e* —, calculation of the dN/dS ratio was not feasible for these genotypes as they presented solely nonsynonymous mutations.

We found that 20 out of the 80 mutations (25%) occurred within B-cell epitopes with species, B-serogroup, and serovar specificities \(P < 10^{-3}\) (Table 5.1 & Fig. 5.3). This phenomenon was observed for all genotypes except F and Ja, where mutations were found solely outside these antigenic regions. Additionally, we also found a global nonsynonymous
mutation rate six-fold higher for the B-cell antigenic regions than for the rest of *ompA* (*P* = 0.027). Indeed, 90% (18/20) of the B-cell epitope mutations resulted in amino acid replacements. Furthermore, for six genotypes (C, E, G, H, Ia, and J), mutations also occurred within or close to the few core CTL and/or Th-cell epitopes that were already mapped on MOMP (Table 5.1 & Fig. 5.3).

Figure 5.3 - Variability of *C. trachomatis* MOMP. Panels (A) to (E) correspond to genotypes B, C, Da, E, and Ia, respectively. Within each panel, the upper row represents the prototype strain for the respective genotype. Each row in the alignment consists of a single type of genetic variants that may represent several specimens. For comparative proposes, sequences from other prototype strains are shown below the prototype sequence. Types of genetic variants (yellow circles) and variable sites (blue circles) that had never been described before are represented. Silent mutations are highlighted in gray. Positions 62 in panels (A) and (E) constitute a simultaneous occurrence of one silent and one nonsilent mutation on the respective codon. Protein VDs are highlighted in yellow. B-cell, CTL, and Th-cell antigenic regions (which may contain various overlapping epitopes) are underlined in red, purple, and green, respectively.
An interesting example occurred for genotype E, where two nonsynonymous mutations at positions 514 and 568 were found to flank a CTL epitope that spans VDII and CDIII and is specifically recognized by subjects infected with *C. trachomatis* genotype E [138]. Curiously, this epitope is within a serovar E-specific Th epitope-containing peptide [139,140] and is adjacent to three well-defined B-cell core epitopes [136] (Fig. 5.3). In addition, some silent mutations found among all H (at position 850) and Ia (at positions 837 and 840) strains as well as one nonsynonymous mutation found for one G variant clinical specimen (at position 857) occurred within a cluster of five HLA class I-restricted minimal CTL epitopes, which are located at CDIV and are *C. trachomatis* species specific [138]. In addition, for all genotypes (except G), these CTL epitopes fully overlap at least six known HLA class II-restricted core Th-cell epitopes [140]. Thus, this remarkable concentration of amino acid alterations in MOMP antigenic regions points to a pathogen’s strategy for host immune evasion, where these changes may lead to disruption of functional/structural epitopes.

### 5.4.4. Global and codon-based evaluation of selective pressure

Considering the targeted occurrence of nonsilent mutations in MOMP antigenic regions, we used bioinformatics to examine the existence of codons under positive selection. The analysis of the molecular evolution of the whole *ompA* gene showed dN/dS values of <1 for genotypes B, C, Da, H, Ia, J, and Ja, while genotypes F and G showed dN/dS values of >1 (1.53 [SE 1.60] and 2.48 [SE 2.55], respectively) (Table 5.1), suggesting that the *ompA* genes of the later may be subject to positive selection. In fact, 83.3% and 85.7% of the mutations displayed by all F and G strains, respectively, yielded amino acid alterations. In addition, the dN/dS ratio for genotypes A, D, E, I, and L2 is expectedly high (>1), as all the mutations displayed by their genetic variants were nonsynonymous. However, the nonexistence of silent mutations for these genotypes rules out the estimation of dN/dS. Still, a calculation of dN/dS for the whole gene may be biased by both the existence of a nonrepresentative number of variants for each genotype and the nondiscrimination of codons with distinct evolutionary signatures. In fact, it is known that when very few amino acids are under positive selection, the global dN/dS statistics are not very sensitive to that selection. If synonymous substitutions accumulate at a higher rate than nonsynonymous substitutions, the few amino acid replacements that occur due to positive selection will not influence dN enough to cause it to be significantly greater than dS.
Thus, we applied the codon substitution model to evaluate positive selection. Considering software constraints, we studied genotypes Da, G, and Ia, as they presented the highest number of variant clinical specimens. A proportion of 0.5% of sites under positive selection and 99.5% under neutral selection ($\omega=1.00$) was estimated for genotype G (including all variant and nonvariant G strains), considering models M2a and M8 (Table S5.2). The proportion of sites under positive selection was not statistically significant (M2a versus M1a; $P = 0.064$ and M8 versus M7: $P = 0.063$) (Table 5.2). Both models (M2a and M8) identified six sites (125, 163, 234, 286, 335, and 339) with a posterior probability of belonging to a class of sites under positive selection higher than 0.5. From these, only position 335 presented a probability of higher than 0.87, showing a posterior $\omega$ of 5.6 [SE 3.1]. Interestingly, site 286 is within the above-cited cluster of five HLA class I-restricted CTL species-specific epitopes overlapping six known HLA class II-restricted Th-cell epitopes [138] (Fig. S.1).

For genotype Ia (including all variant and nonvariant Ia strains), the hypothesis of the nonexistence of sites under positive selection was rejected from the comparison of models M8 versus M7 ($P = 0.042$) (Table 5.2). Estimates from model M8 showed that 7.3% of the sites were under positive selection and 92.7% of sites were under purifying selection, with $\omega=0.0052$ (Table S5.3). Eight sites (62, 93, 94, 176, 191, 193, 255, and 311) were identified as belonging to the positive selection class ($\omega>1$, posterior probability $>0.5$), although only positions 62 ($\omega=5.7$; SE 3.2) and 176 ($\omega=5.9$; SE 3.1) presented posterior probabilities higher than 0.89. Curiously, sites 93, 94, and 255 are within or close to B-cell epitopes (Figs. 5.3E & S.1).

With respect to genotype Da (including all variant and nonvariant Da strains), although the M8 model identified three sites (162, 326, and 333) as being positively selected

---

**Table 5.2. Likelihood ratio test statistics for testing positive selection**

<table>
<thead>
<tr>
<th>Genotype and Models</th>
<th>$2\Delta\ell^a$</th>
<th>df$^b$</th>
<th>$p^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>G</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2a versus M1a</td>
<td>5.509764</td>
<td>2</td>
<td>0.064</td>
</tr>
<tr>
<td>M8 versus M7</td>
<td>5.526772</td>
<td>2</td>
<td>0.063</td>
</tr>
<tr>
<td><strong>Ia</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2a versus M1a</td>
<td>3.347866</td>
<td>2</td>
<td>0.188</td>
</tr>
<tr>
<td>M8 versus M7</td>
<td>6.345914</td>
<td>2</td>
<td>0.042</td>
</tr>
</tbody>
</table>

*a* Log-likelihood difference between compared models.

*b* Degrees of freedom between compared models.

*c* $p$ of the log-likelihood test.
(ω > 1, posterior probability > 0.5) (Table S5.3), the likelihood ratio test of a positive selection was not statistically significant (data not shown). Once more, sites 326 and 333 are within B-cell epitopes with specificity for serogroup B and genotype Da, respectively (Figs. 5.3C & S.1).

Collectively, despite the lack of statistical significance for a positive selection, the six described sites (one for G, three for Ia, and two for Da) within MOMP antigenic regions were also found to be variable in *C. trachomatis* strains collected in other regions of the world (see Discussion). Thus, this strongly suggests that they are under positive selection, and that the lack of statistical significance may be a matter of population size.

### 5.4.5. Types of amino acid substitutions

In order to evaluate the existence of mutations that are likely becoming fixed in the population, we investigated the frequency of each amino acid substitution per nonsynonymous variable site, involving all 232 variant clinical specimens (Table 5.1). The change of Thr to Ala was the most frequent, occurring 63 times (15.5%), 94.5% of which were for genotype Da. The Asn-to-Asp, Ile-to-Thr, and Val-to-Met changes were also highly common, with 51, 48, and 47 occurrences, respectively, and the vast majority involved genotype Ia. These high frequencies observed for both genotypes Da and Ia indicate a fixation of these mutations among the corresponding strains, where they likely play an important structural or functional role.

We also evaluated the frequency of each amino acid substitution per nonsynonymous variable site but considering only the 43 types of genetic variants, in order to avoid any bias due to the number of same-genotype variant specimens presenting that substitution. Overall, from the 30 distinct types of amino acid substitutions that took place, the change of Ala to Thr was the most frequent, occurring seven times for seven different genotypes (B, D, Da, E, Ia, J, and L2), followed by the Thr-to-Ala change, with five occurrences (Table 5.1). Moreover, among all types of genetic variants, Ala and Thr were also the most frequently replaced amino acids throughout MOMP (11 times for eight genotypes and 7 times for six genotypes, respectively), as well as the two most acquired ones (7 times for six genotypes and 11 times for nine genotypes, respectively) (Fig. 5.4). This is not surprising considering that these two amino acids are the most common among all chlamydial prototype strains and the 43 genetic variants under study (data not shown).
Figure 5.4 - Number and type of substitutions involving a specific amino acid. Comparative analysis of all types of genetic variants for all genotypes revealed 52 nonsynonymous variable sites in MOMP, representing 30 distinct types of amino acid substitutions. The size of each bar represents the number of times that a residue is replaced (in gray) or acquired (in black) among solely the 43 types of genetic variants (this excludes any bias due to the heterogeneous number of same-genotype variant specimens presenting that substitution). The number of ompA genotypes presenting that specific amino acid substitution is shown in parentheses above each bar.

5.4.6. Analysis of protein features

To shed some light on the putative impact of each mutation on MOMP, a protein sequence-based comparative analysis was also performed for the 43 types of genetic variants and all prototype strains. In general, although 80% of the 30 types of amino acid substitutions occurred between residues with dissimilar properties, no significant differences were observed for charge density, secondary structure, hydrophathy, antigenicity, amphiphilicity, surface probability, or flexibility (data not shown). Interestingly, although only 23.3% of all substitutions are indicated to cause a gain or loss of charge density, a consequent reversal of charge density was frequently observed for their neighbour amino acids (data not shown).

Considering the porin role of MOMP, we also analyzed the loss or gain of amino acids that were shown to influence porin structure [242]. Six of the nine changes involving residues Asp and Glu (Table 5.1) were on the MOMP external loops, leading to an alteration of the negative charge in these regions. Moreover, these substitutions occurred for specimens of several genotypes (C, D, F, J, and Ja). Similarly, a targeted addition of Asp, Glu, and Tyr residues in the MOMP periplasmic turns was also observed for all Ia (Asp) as well as for 15 G (Glu) and one G (Tyr) variant specimens (Table 5.1). Therefore, although the impact of
the loss or gain of these specific amino acids is not known, the nonrandomness of these occurrences suggests some plasticity to deal with environmental changes/pressure.

5.5. Discussion

*Chlamydia trachomatis* MOMP has been implicated as a porin, adhesin, and key antigen. In the present study, we performed a mutational trend analysis of this major membrane component from 795 clinical isolates from ano-urogenital infections. In common with previous surveys for other countries [251,255,256,286,287], E and F were the most prevalent genotypes among the Portuguese population (Fig. 5.1), representing almost 57% of all *C. trachomatis* infections (39.8% E and 17.3% F), which indicates little fluctuation in genotypic prevalence worldwide. Although the secret of this apparent worldwide success still remains to be elucidated, it has been speculated that the biological advantage of E and F may reside in a more effective host immune evasion [112] as well as in the existence of specific adhesins or virulence factors that may give them some functional or structural facility in terms of infection and transmission [165,167]. However, further experimental evidence will be needed to support these assumptions.

Globally, 232 clinical specimens (29.2%) were found to present *ompA* variant sequences compared to the respective prototype strain (Fig. 5.1). Interestingly, the two most successful genotypes, E and F, were the least variable among the studied population, showing an *ompA* mutation rate 60.3-fold lower than that of the other genotypes ($P < 10^{-8}$). In agreement with this, several studies from different geographic regions [251,253,287,288] have reported a high level of conservation of both E and F strains. The apparent lack of *ompA* variability in these strains may indicate a better fitness for E and F, where the resultant MOMP may be less immunogenic, which could favor longer E/F “camouflage,” further dissemination, and, consequently, higher success. Supporting this, by testing serovar-specific lymphoproliferative responses to several MOMP synthetic peptides, Arno *et al.* [289] found that individuals infected with genotype E recognized a peptide containing a known CDIII T-cell epitope less frequently than did non-E-infected individuals. Moreover, although the lack of clinical data has hampered us from making any correlation with clinical symptoms, genotype E has been frequently associated with asymptomatic infections [259,262,290], reinforcing the speculation that the most prevalent genotypes may elicit a less vigorous host immune response. However, some weakly supported associations with clinical symptoms were previously found for genotype F [259,290-292], which makes any assumption merely controversial and inconclusive.
Overall, 43 distinct types of genetic variants (involving 14 genotypes) were found (Fig. 5.2), and 67.4% of them had not been reported before. It is known that genetic variability in MOMP may arise from point mutations, recombination events, or indels as a result of host selective pressure and bacterial adaptation. However, no visible trace of recombination or any indel event was found in \textit{ompA} in our genetic variants compared to the respective prototype strain, suggesting that the genetic variability observed is strictly a consequence of the occurrence of point mutations. Interestingly, of the total 80 nucleotide-variable sites exhibited by the 43 genetic variants (Table 5.1), 65% resulted in amino acid alterations, which suggest that they are evolutionarily nonneutral. In support of this, dN/dS values of $>1$ were found, at least for variant specimens F and G, suggesting the existence of positive selection. More specifically, various codons (some of them located in MOMP antigenic regions) were likely positively selected in Da, G, and Ia variant specimens, and statistical significance was achieved for the later ($P = 0.042$). Curiously, 46 out of the 80 nucleotide changes were shared by other same-genotype strains already described in other regions of the world. This suggests a fixation of these mutations, likely as a biological advantage in terms of infectivity and transmission. Although no visible mutational “hot spots” were found among genetic variants of different genotypes, we observed a few cases where the occurrence of mutations for different \textit{ompA} genotypes was in exactly the same gene location (Fig. 5.3 & Table 5.1), which may suggest the existence of spots that are more prone to change. We speculate that this may be due either to fitness advantages or simply to MOMP structural constraints.

For most genotypes, mutations were found in both the \textit{ompA} VDs and CDs, indicating a general mutational trend throughout the entire gene. A notable exception occurred for B strains, which exhibited a 20.4-fold higher probability of undergoing changes in CDI than in the remainder of the gene ($P < 10^{-7}$). Nevertheless, we have no reasonable explanation for this highly restricted mutational clustering for genotype B, considering that CDI is the protein region that is thought to be more conserved and one of the less immunogenic. In contrast to what was expected, a greater number of mutations (1.42-fold higher) was seen for the CDs than for the VDs. However, 51.1% of the mutations in CDs were silent, suggesting that the conservation of these domains may be important to maintain either the MOMP porin function related to its monomeric form (34, 59) or the chlamydial membrane structural integrity, which is associated with the trimeric form of MOMP [238]. In contrast, VDs showed a global nonsynonymous mutation rate 16.2-fold higher than that of the CDs, where 93.9% of the changes resulted in amino acid replacements ($P < 10^{-4}$). This seems to be remarkable evidence of the existence of positive selection almost exclusively targeted to the MOMP
surface-exposed domains. Considering that MOMP is the major chlamydial membrane component and unquestionably its dominant antigen, these findings point to an evolutionary dynamic used by 
\textit{C. trachomatis} to deal with the environment and the host immune pressures. In support of a chlamydial host evasion strategy, 25\% (20/80) of the mutations occurred within well-defined B-cell epitopes for most genotypes (Table 5.1), which is statistically significant as these epitopes encompass only about 12\% of the total MOMP length ($P < 10^{-3}$).

Furthermore, 90\% of the mutations found within B-cell antigenic regions resulted in amino acid replacements. In fact, these regions showed a global nonsynonymous mutation rate six-fold higher than that for the rest of \textit{ompA} ($P = 0.027$), which suggests a targeted amino acid variability in MOMP B-cell antigenic regions. Curiously, some of the nonsynonymous mutations reported in the present study were previously shown to abrogate monoclonal Ab binding and neutralization of infectivity of several prototype and/or genovariant strains \textit{in vitro} [136,281]. For example, the VDIV Ala -> Thr change found in all D and 11 Da variant specimens (Table 5.1) was previously shown to prevent the antibody binding at both the serogroup B \textit{322TIAGAGD328} and \textit{323IAGAG327} epitopes for the known genovariant D' [281]. Also, one Portuguese genovariant, C, showed a Gly -> Asp alteration in VDI that was found to prevent the antibody binding to the \textit{92DVAGL96} epitope (specific for genotypes A, C, I, J, and Ja) in a previous complete-replacement analysis of this antigenic site [134]. Although it would be interesting to extend this kind of \textit{in vitro} experimental demonstration to all mutations occurring within B-cell epitopes, their reflection \textit{in vivo} remains speculative, as the true epitope conformation cannot be ensured and potential differences among genotypes cannot be evaluated.

Interestingly, for some genotypes, mutations were also found to occur within or close to two MOMP T-cell antigenic clusters (Table 5.1 & Fig. 5.3) [136,138-140]. Although the impact of mutations in these epitopes remains unknown, they may indicate an evasion strategy where single mutations are used to disrupt or affect clusters of multiple epitopes that simultaneously elicit different host protective immune responses, especially when they are serovar specific [137]. This strategy would promote the persistence of the infection and/or disease progression, which could significantly increase transmission and/or reinfection. However, Kim and DeMars [137] have suggested, based on a previous study [184], that the existence of mutations in cluster epitopes with species specificity (and, thus, highly conserved among genotypes) would be deleterious and greatly reduce pathogen fitness. This negative selection theory finds some support in our data. In fact, of the three mutations occurring inside that region, only one yielded an amino acid replacement, and it was exhibited solely by one specimen out of the 795 strains under study. Nevertheless, 64 variant strains from different
genotypes (all 18 H and 46 Ia) presented silent mutations within that cluster, which points to a fixation event and may influence the subsequent evolutionary landscape of ompA. In fact, it was recently demonstrated that the fixation of silent mutations in open reading frames may influence both the transcriptional pattern (due to the modification of the primary mRNA structure) and the type of amino acids that can be added in a second round of mutations [293]. Although we cannot evaluate the impact of this mutational clustering in B- and T-cell epitopes of the chlamydial key antigen on vaccine success, new antigenic profiles may be expected. Accordingly, acquisition of knowledge about the ompA evolutionary mutational pattern and the identification of new types of variants that are arising emerge as important tasks.

Besides the remarkable evidences of a MOMP mutational dynamics strongly focused on a host evasion strategy, our data also suggest that the variability exhibited by some genetic variants may have a direct influence on the protein structure. In fact, some replacements involved amino acids that are thought to be important to the MOMP conformation. For instance, most of the changes involving the negatively charged residues Asp and Glu have occurred on the protein external loops (Table 5.1), where, according to the literature [242], they are normally abundant and participate in binding to the lipopolysaccharide (LPS) [294]. Since the substitution of charged amino acids has a direct mimic effect on the charge density of the neighbour amino acids (data not shown), it is expected that changes altering the number of these residues may affect the MOMP porin topology. Similarly, the additional presence of Asp, Glu, and Tyr residues in the MOMP periplasmic turns of some genetic variants (Table 5.1) may also contribute to a more stable and flexible porin conformation. In fact, it was previously reported that residues with a higher bend potential (such as Asp, Glu, and Tyr) are well represented in the small periplasmic turns of several bacterial porins [242], where they likely play a critical role during the membrane mechanical movements. However, an evaluation of the conformational impact of these alterations on MOMP is likely an unfeasible task, considering the extremely complex putative protein structure [241,242]. Collectively, these results show a global mutational evolutionary dynamics of C. trachomatis ompA that is associated with MOMP functional/structural constraints and mainly with evading host immune surveillance. In the light of the continuous coevolution between this pathogen and its human host [165] that takes place during infection, this evolutionary scenario is expectedly beneficial in terms of infectivity, persistence, and transmission.
CHAPTER 6

Adaptive Evolution of the *C. trachomatis* Dominant Antigen Reveals Distinct Evolutionary Scenarios for B- and T-cell Epitopes: Worldwide Survey

Published in

Author Contributions
AN and JPG conceived and designed the experiments; AN performed the experiments; AN and PJN performed the statistics; AN and JPG analyzed the data; AN and JPG wrote the paper; MJB reviewed the paper.

6.1. Abstract

*Chlamydia trachomatis* is one of the most disseminated human pathogens, for which no vaccine is available yet. Understanding the impact of the host pressure on pathogen antigens is crucial, but so far it was only assessed for a very restricted geographic area. We aimed to evaluate the evolutionary picture of the chlamydial key antigen (MOMP), which is one of the leading multi-subunit vaccine candidates, in a worldwide basis. Using genetic and molecular evolution methods and mathematical modelling, we analyzed all MOMP sequences reported worldwide, composed by 5026 strains from 33 distinct geographic regions dispersed by five continents. Overall, 35.9% of variants were detected. The evolutionary pattern of MOMP amino acid gains/losses was found to differ from the remaining chromosome, reflecting the demanding constraints of this porin, adhesin and dominant antigen. Amino acid changes were 4.3-fold more frequent in host-interacting domains (*P* < 10\(^{-12}\)), specifically within B-cell epitopes (*P* < 10\(^{-5}\)), where 25% of them are at fixation (*P* < 10\(^{-5}\)). According to the typical pathogen-host arms race, this rampant B-cell antigenic variation likely represents natural neutralization escape mutants, as some of the mutations were previously shown to abrogate neutralization of chlamydial infectivity in vitro. In contrast, T-cell clusters of diverse HLA specificities are under purifying selection, suggesting a strategy that may lead to immune subversion. Moreover, several silent mutations were found to be at fixation, generating preferential codons that may influence expression, and may also reflect recombination-derived 'hitchhiking effect' from favourable nonsilent changes. Interestingly, the most prevalent *C. trachomatis* genotypes, E and F, showed a mutation rate 22.3-fold lower than that of the remainder (*P* < 10\(^{-20}\)), suggesting more fitted antigenic profiles. Globally, the adaptive evolution of the *C. trachomatis* dominant antigen is likely driven by its complex pathogenesis-related function and reflects distinct evolutionary antigenic scenarios that may benefit the pathogen, and thus should be taking into account in the development of a MOMP-based vaccine.
6.2. Introduction

*Chlamydia trachomatis* is an obligate intracellular pathogen that causes ocular-genital infections in humans. Trachoma (chlamydial genotypes A-C and Ba) is the world’s leading cause of preventable blindness with special impact in resource-poor nations, which has been recently placed on the WHO's priority list for intervention [295]. Also, the asymptomatic character of most genital chlamydial infections (genotypes D-K, Da, Ia, Ja and L1-L3) makes this pathogen the major cause of bacterial sexually transmitted infections worldwide [16]. Thus, *C. trachomatis* constitute a major public health problem, and the development of effective preventive strategies, such as a vaccine, are urgently needed. So far, vaccine attempts failed to provide broad coverage and conferred limited protection [296,297].

One of the leading multi-subunit vaccine candidates is the *C. trachomatis* MOMP, coded by *ompA*, whose variations underlie strain classification into serogroups (B, C and Intermediate) or genotypes [3]. It is the dominant antigen with tenths of well-defined species and serovar-specific epitopes, eliciting both the humoral and cellular immune responses [134,136,137,244,296]. MOMP constitutes about 60% of the membrane dry-weight [133] and is a trimer stable under reducing conditions [238]. Each monomer is predicted to form a 16-stranded β-barrel structure resembling a porin [242], where the four highly variable domains (VDI to VDIV) of the protein are surface exposed. Also, the charge properties of VDs support an adhesion role for MOMP [55], both at the time of attachment to the host cell as well as to the host inclusion membrane to interact with the mitochondria and endoplasmatic reticulum [298,299]. Moreover, this protein is likely an important virulence factor as it has been demonstrated that subtle mutations may yield profound distinct strain-specific neutralizing Ab responses in a nonhuman primate trachoma model [300], and it presents an immune decoy function by blocking the binding of broadly protective species-common pan-neutralizing Abs [62].

It is believed that *C. trachomatis* evolution involved the loss of huge portions of its genome upon becoming an obligate intracellular parasite [301]. Subsequently, the evolutionary accumulation of mutations and indel events resulted in the formation of pseudogenes [7,302], the phylogenetic segregation of strains according to their cell-appetence [6,165], and the generation of specific evolutionary pictures for heterogeneous loci categories [165]. Recently, a geographic-restricted study showed an adaptation of the *C. trachomatis* key antigen to the host immune pressure [303], shedding some lights about the MOMP adaptive evolution. Here, we aimed to determine MOMP evolutionary picture in a worldwide basis by using all available *ompA* sequences reported in the literature, which totalize 5026 strains isolated in 33 distinct geographic regions dispersed by five continents. We mapped MOMP
positions that are more prone to change as well as specific mutations in epitopes that are at fixation around the globe. Globally, the adaptive evolution of the *C. trachomatis* dominant antigen likely reflects its complex pathogenesis-related function as an adhesion, porin, and major antigen, showing distinct evolutionary scenarios for B- and T-cell epitopes. The worldwide coverage of this assessment is unprecedented to date.

### 6.3. Material and Methods

#### 6.3.1. Study population

An exhaustive and complete literature search was performed to look for all sequence-based MOMP published surveys containing both variant and nonvariant *ompA* sequences compared to the respective prototype strains. In general, the study population from the 56 surveys performed to date, encompassed patients attending to general practice, family planning, obstetrics, gynaecology and STD clinics, as well as individuals from specific and restrict communities (such as homosexual networks or resource-poor villages where LGV and trachoma are relevant, respectively). To avoid bias potentially arising from the duplication of specimens, the following precautions were taken: 

1. specimens collected from sexual partners during the same year were excluded; and
2. for longitudinal studies where more than one sample can be collected from the same patient, only one stage of the study was considered as it would be impossible to distinguish if samples of further stages were from re-infection or a merely persistence episode. To our knowledge, none of the strains was propagated in cell culture after a plaque assay (or limiting dilution techniques) in order to obtain clones. Accordingly, one cannot tell whether the determined *ompA* genotype refers to a single clone or to the predominant clone (if a mixed infection occurred).

For genotype comparisons, the traditional prototype strains representing all *C. trachomatis* genotypes were used as baseline. Considering that for some prototype strains inconsistencies were observed among *ompA* sequences that were obtained in the 80s and the beginning of 2000 (when automated sequencing by capillary electrophoresis started), only the later were selected because of their likely higher reliability. However, the possibility of these inconsistencies could also be due to different clones of the same prototype strain cannot be discarded. Therefore, the prototype strains used in this study are described in Table 6.1. Although the use of Ba/Apache2 as prototype strain is consensual, it was considered as a B-variant in the present study due to its closer proximity to the prototype B/TW5 than some well-known B strains as well as to its similar *ompA* mutational pattern with several previously classified B-variant strains, which makes any B/Ba-genotype distinction, in our opinion,
inaccurate. Moreover, some specimens reported in the literature as D-, I-, and J-variants were found to be identical to prototype strains Da, Ia and Ja, and thus were not considered as variants in the present study.

### Table 6.1. Description of the 17 baseline prototype strains representing the *C. trachomatis* genotypes.

<table>
<thead>
<tr>
<th>Genotype/Strain</th>
<th>Year</th>
<th>Location</th>
<th>Biological Sample</th>
<th>GenBank Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/Har13</td>
<td>1958</td>
<td>Egypt</td>
<td>Conjunctiva</td>
<td>NC007429</td>
</tr>
<tr>
<td>B/TW5</td>
<td>1959</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td>AF304856</td>
</tr>
<tr>
<td>C/TW3</td>
<td>1959</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td>AF352789</td>
</tr>
<tr>
<td>D/UW3</td>
<td>1965</td>
<td>Washington</td>
<td>Cervix</td>
<td>AE001338</td>
</tr>
<tr>
<td>Da/TW448</td>
<td>1985?</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td>X62921</td>
</tr>
<tr>
<td>E/Bour</td>
<td>1959</td>
<td>California</td>
<td>Conjunctiva</td>
<td>X52557</td>
</tr>
<tr>
<td>F/IC-Cal3</td>
<td>1960?</td>
<td>California</td>
<td>Conjunctiva</td>
<td>X52080</td>
</tr>
<tr>
<td>G/UW57</td>
<td>1971</td>
<td>Washington</td>
<td>Cervix</td>
<td>AF063199</td>
</tr>
<tr>
<td>H/UW43</td>
<td>1965</td>
<td>Washington</td>
<td>Cervix</td>
<td>X16007</td>
</tr>
<tr>
<td>I/UW12</td>
<td>1966</td>
<td>Washington</td>
<td>Urethra</td>
<td>AF063200</td>
</tr>
<tr>
<td>Ia/TU4168</td>
<td>1987</td>
<td>Indiana</td>
<td>Urethra</td>
<td>AF063201</td>
</tr>
<tr>
<td>J/UW36</td>
<td>1971</td>
<td>Washington</td>
<td>Cervix</td>
<td>AF063202</td>
</tr>
<tr>
<td>Ja/TUA795</td>
<td>1986</td>
<td>n.a.</td>
<td>Cervix</td>
<td>AF063203</td>
</tr>
<tr>
<td>K/UW31</td>
<td>1973?</td>
<td>Washington</td>
<td>Cervix</td>
<td>AF063204</td>
</tr>
<tr>
<td>L1/440</td>
<td>1968</td>
<td>California</td>
<td>Lymph Nodes</td>
<td>M36533</td>
</tr>
<tr>
<td>L2/434</td>
<td>1968</td>
<td>California</td>
<td>Lymph Nodes</td>
<td>M14738</td>
</tr>
<tr>
<td>L3/404</td>
<td>1967</td>
<td>California</td>
<td>Lymph Nodes</td>
<td>X55700</td>
</tr>
</tbody>
</table>

6.3.2. Genomic and phylogenetic analyses

Genomic and phylogenetic analyses were performed as previously described [303]. Briefly, for each *ompA* genotype, the MEGA 4.0.2 software ([http://www.megasoftware.net](http://www.megasoftware.net)) was used to produce alignments, to create matrices of pairwise comparisons, to estimate the number of variable sites, to compute overall mean genetic distances within and between serogroups, and to generate phylogenies at the nucleotide level. For the later, both the NJ method [with K2P and Tamura-Nei (TN) models] and the Maximum Parsimony (MP) method (using the max-min branch-and-bound algorithm) were used as previously described. As more than two-thirds of the mutations are nonsynonymous, these analyses were repeated at the protein level to check the existence of ambiguities. In addition, the mean genetic diversity within each *ompA* genotype-population (π) was estimated using the Maximum Composite Likelihood method. Considering that many genes display a nonrandom usage of synonymous codons for specific amino acids, the SWAAP 1.0.3 software ([http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm](http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm)) was used to compute the codon frequency and the relative synonymous codon usage [304] for all genotypes. Because of the different lengths of the sequences, the pairwise-deletion option was chosen to remove all sites
containing missing data or alignment gaps, only when the need arose and not prior to the analyses. Standard error estimates were obtained by a bootstrap procedure (1,000 replicates).

Since recombination is a mechanism of genetic variability, the SimPlot/BootScan software (http://sray.med.som.jhmi.edu/SCRoftware/) was used to evaluate the existence of any recombination event in *ompA* of each specimen, as previously described [114].

**6.3.3. Analysis of molecular evolution**

For each *ompA* genotype, the Nei-Gojobori method [267] of MEGA was used to calculate the overall mean of dS and dN, as previously described [121]. The p-distance model was used to normalize the computed differences against the number of potential synonymous and nonsynonymous sites. dS and dN were also evaluated over a sliding window (window size=15; step size=3) using the Nei-Gojobori method of the SWAAP. In order to evaluate the statistical significance of the operating selective pressure on MOMP specific domains, a codon-based Z-test of selection was performed to calculate the probability (95% confidence interval) for rejecting the null hypothesis of strict-neutrality (dN=dS) in favor of the positive (dN>dS) or purifying selection (dN<dS). An exhaustive evaluation of the exact codons under positive selection, using a more conservative and powerful bioinformatic approach (maximum likelihood analyses), would required the knowledge of all nonvariant specimens circulating worldwide, which is unfeasible as studies involving solely nonvariant specimens were not included in the present survey.

Since estimation of the ratio of transitions rate (Ts) to transversions rate (Tv) provides insight into the process of molecular evolution of a locus, the Tamura 3-Parameter method [305] of MEGA was used to compute the overall mean of Ts/Tv ratio (κ) for all examined *ompA* sequences, while SWAAP was used to evaluate R over a sliding window (window size=60; step size=1), in order to identify gene regions in which functional constraints are tolerated.

**6.3.4. Analysis of protein features**

To shed some light on the putative impact of each mutation on MOMP, a comparative analysis of the protein sequences of all variant and prototype strains was performed using the Protean program of LaserGene (DNASTAR). Basically, the following protein features were evaluated: charge density, secondary structure, hydropathy, antigenicity, amphiphilicity, surface probability and flexibility. We used a sliding window of a specified range of amino acids with the default parameters of each method, as previously described [303].
6.3.5. Mathematical modelling

To examine the mutational trend of amino acid composition in MOMP evolution, we used the *ompA* sequences of prototype strains as baseline. We developed a mathematical approach to estimate the ‘expected’ frequency of each amino acid considering all possible SNPs for each position of the ~397 *ompA* triplets. Depending on genotype, the novel triplets (*t*<sub>i</sub>) were determined according to the distribution:

\[
(t_i, w_i), \quad i = 1, \cdots, N
\]

where *w*<sub>i</sub> represents the relative weight of each type of SNP. This correction was performed because transitions typically occur in nature much more often than transversions. For comparative purposes, two different overall mean \(\kappa\) values were used: one relative to *ompA* (\(\kappa = 1.2\)) and the other relative to ~50 loci of the *C. trachomatis* chromosome (\(\kappa = 3.6\)) [165]. Thus, for each possible transition, it was assigned a relative weight of 1.2 or 3.6, while a weight of 1 was assigned for all possible transversions. All amino acids (*a*<sub>j</sub>) resulting from the novel triplets generated, were determined according to the following distribution:

\[
(a_j, s_j), \quad j = 1, \cdots, n
\]

where each *j* represents the set of indices \(I_j = i_{1j}, i_{2j}, \cdots, i_{kj}\) of all triplets \((t_i, w_i)\) that code a certain amino acid, and

\[
s_j = \sum_{i \in I_j} w_i
\]

represents the relative weight of each novel amino acid. Considering that one amino acid may be encoded by several synonymous triplets, the relative weight of each amino acid was calculated by summing the weights of all same-amino acid triplets. The final ‘expected’ frequency for each amino acid was calculated according to the formula:

\[
f(a_j) = \frac{s_j}{\sum_{i=1}^{n} s_i}
\]

All calculations were performed using the free software environment for statistical computing and graphics R (http://www.r-project.org/).
6.3.6. Statistic analyses

All statistics was performed using the SPSS Base version 15.0 (SPSS Inc. Chicago) in order to estimate the $P$ values by Fisher’s exact test as well as the odds ratios with a 95% confidence interval.

6.4. Results

6.4.1. Characterization of C. trachomatis specimens

A total of 5026 C. trachomatis strains isolated in 33 distinct geographic regions from five continents were analyzed in this worldwide survey. These encompassed all $ompA$ genotypes with exception of L3, which was not found. Genotype E was the most prevalent type (30.0%), followed by genotypes F (12.9%), and L2 (9.8%). Globally, 1802 (35.9%) specimens were found to present $ompA$ variant sequences when compared to that of the respective prototype strain (that define each $ompA$ genotype), which have been used worldwide as comparative-baseline since their isolation up to 60 years ago (Table 6.1). Genotypes L2 and G presented the highest number of $ompA$ variant specimens ($n=462$ and $n=309$, respectively), while genotypes E and F, which together represent almost half (42.3%) of all analyzed specimens, were the least variable, where only 83 out of 1464 E strains (5.6%) and 40 out of 640 F strains (6.3%) were variants. In fact, these two most worldwide prevalent types showed an $ompA$ mutation rate 22.3-fold lower than that of the other genotypes ($P < 10^{-20}$), which suggest that the dominant antigenic profile of E and F MOMP may be better fitted to deal with the host immune system, which consequently may favor their ecological success.

The 1802 variant specimens constitute 234 distinct types of genetic variants (i.e., with a distinct $ompA$ mutational pattern that may be shared by several strains), for which no insertion/deletion or any trace of recombination event was found in $ompA$. Supporting this, robust phylogenies (Fig. 6.1) were observed with maximum bootstrap support (97-100%) in the three main branches corresponding to the traditional chlamydial serogroups. In fact, nucleotide genetic distances among strains within the same genotype did not exceed 0.7% [SE 0.2%]. In contrast to the other genotypes, most types of genetic variants for the least variable genotype E are represented by a single specimen, which suggests that the corresponding mutations are not being fixed.
6.4.2. Worldwide ompA variability

Globally, a total of 511 variable sites (corresponding to the sum of all mutations between each type of genetic variants and the respective prototype strain) were found throughout ompA, where 56% of them occurred in VDs and 44% occurred in CDs. About 68% of mutations yielded amino acid changes (with two thirds occurring in VDs), while ~32% resulted in silent substitutions (with two thirds occurring in CDs).

Figure 6.1 - ompA phylogeny of the 251 C. trachomatis genetic variants. The represented tree was generated using the NJ method (with K2P model) but similar topologies were obtained with other methodologies (see Methods). Genotypes are represented in different colors. Thickness of solid triangles is proportional to the number of taxa from each genotype (n). The mean genetic diversity within each ompA genotype-population (π) is shown. The ompA sequence from C. muridarum strain MoPn (GenBank: NC002620) was used to root the tree. Bootstrap values (1,000 replicates) are shown next to the branch nodes.
Figure 6.2 illustrates the most relevant variation found in MOMP for the 17 *C. trachomatis* genotypes, showing 70.3% (n=359) of all variable sites found in the *ompA*. One fifth of them were exhibited by same-genotype strains isolated in two to 20 (for nonsynonymous mutations) or ≥5 (for silent mutations) different geographic regions (Fig. 6.2), which may indicate a fixation of these mutations in those genotypes likely conferring them some structural or antigenic advantage. For example, the VDII Ala-to-Thr change occurring in the B-cell epitope AFV[172]P[175] [284] seems to become fixed within genotype B, as it was exhibited by 137 out of the 157 B variant specimens from 18 distinct geographic regions (data not shown).

Moreover, of the 12 silent mutations that may be at fixation (Fig. 6.2), 75% of them generated synonymous codons with high frequency usage in *C. trachomatis* [306]. For instance, for genotype J, the CDII C-to-T third-position change in the low frequent Asn codon AAC yielded the more frequent synonymous codon AAT (that is used almost two times more often than AAC) in 101 of the 113 (76%) J variant isolates from 10 distinct geographic regions (data not shown). Curiously, this AAT codon is conserved among all genotypes of B- and Intermediate-serogroups.
Figure 6.2 - Worldwide mutational pattern of MOMP. Panels (A), (B), and (C) represent protein alignments of prototype strains according to serogroup B, Intermediate, and C, respectively. Amino acid changes that naturally occur among prototype strains are in grey characters (they do not represent changes among variant specimens and the respective prototype strains). Mutations resulting from the comparison between the 5026 strains isolated in 33 different geographic regions from five continents and the respective prototype strain are highlighted in different colors, corresponding to dissimilar degrees of evolutionary fixation. Green, orange and red represent amino acid changes occurring in strains isolated in a single, 2-4, and ≥5 geographic regions, respectively. Only silent mutations that occurred in strains isolated in ≥5 geographic regions are shown (highlighted in grey). Well-defined B-cell (blue), Th-cell (purple) and CTL (black) antigenic regions are underlined. Degenerated variable sites are marked with an asterisk.

In general, considering the MOMP genetic variability observed for all C. trachomatis genotypes, we found that 148 (37.3%) different sites of the protein were already subject to amino acid alterations (Fig. 6.3). About 44% of the sites suffered mutations that occurred simultaneously among strains from two to ten different genotypes. Also, more than half of these ‘hotspot’ sites involved changes that seem to be at fixation, and were found to occur 2.4-fold more frequently within B-cell epitopes (P < 10⁻²). Altogether, these results suggest that MOMP variability is not random as there are protein sites that are more prone to change, likely due to fitness advantages or simply to MOMP functional constraints.
6.4.3. Evaluation of selective pressure on MOMP

To examine the selective pressure acting on MOMP, only the 234 types of genetic variants together with the 17 prototype strains were considered. This conservative approach avoids any bias arising from the existence of an unequal number of specimens represented by each type of genetic variant, resulting from the considerably different population sizes of the studies performed to date. The analysis of the molecular evolution of whole $ompA$ sequence for all the 251 taxa showed a global mean $dN/dS$ of 0.26 (SE 0.03). The analysis per genotype showed $dN/dS$ values >1 for genotypes A, G, Ia, and K (data not shown), suggesting a phenomenon of positive selection. However, due to the nondiscrimination of codons with distinct evolutionary signatures, this global $dN/dS$ statistics may generate artifactual trends of synonymous and nonsynonymous rate variation when very few amino acids are under positive selection.

In order to investigate if any portion of $ompA$ is undergoing strong selection as well as which regions are not bound by strong functional constraints, a codon-based sliding-window analysis was also used to evaluate $dN/dS$ throughout the entire gene. Although $ompA$ showed a general trend of $dN/dS <1$, three distinct gene regions were identified with a statistically significant ratio >1 (Fig. 6.4A), indicating that they are likely under positive selection.
Interestingly, these regions are precisely located inside the VDI, VDII and VDIV, which (together with VDIII) showed a global nonsynonymous rate 4.3-fold higher than that of the CDs ($P < 10^{-12}$), suggesting a targeted amino acid variability in MOMP surface-exposed domains. In support of this, several putative positively selected codons were found within the VDI165-TQSSNF170 epitope or within a VDIV region containing multiple core B-cell epitopes for genotypes Ia and E, respectively. Almost all of these codons suffered nonsynonymous mutations that may lead to disruption of those epitopes in some E and Ia genetic variants. For all genotypes, almost all of the $dN$ peaks observed occurred within or adjacent to well-defined MOMP epitopes, and also involved mutations that seem to become fixed among same-genotype strains (Fig. 6.4B).

Figure 6.4 - Selective pressure acting on *ompA*. (A) Ratio of $dN/dS$ throughout *ompA*. $dN/dS$ was estimated over a nucleotide sliding window (window size=15; step size=3), considering the 234 types of genetic variants and the 17 prototype strains. VDs are shown in horizontal red lines, while dashed line represents the traditional cut-off for neutral evolution. Only $P$-values showing statistical significance are shown. For positive and purifying selection $P$-values are displayed above and below the cut-off dashed line, respectively. Region containing MOMP CDIV species-specific cluster of five CTL and six Th epitopes (see text) is marked with (*). (B) $dN$ values throughout *ompA* for genotype A (here, as an example). Calculations were performed over a sliding window (window size=15; step size=3), considering all types of genetic variants of genotype A. VDs (red boxes) display a global nonsynonymous rate 4.3-fold higher than that of the CDs ($P < 10^{-12}$). $dN$ peaks involving mutations occurring within/adjacent to B-cell epitopes and/or mutations that are likely at fixation in this genotype are marked with (*) and (†), respectively.
Globally, it is possible that these results of positively selected regions may clearly be underestimated. In fact, it was recently demonstrated that for short time-scale models, the anticipated signature of adaptive evolution \((dN/dS > 1)\) is violated because \(dN/dS < 1\) was found to occur under both negative and positive selection [307].

6.4.4. Is variability in MOMP a host evasion strategy?

The mutational trend in MOMP antigenic regions was investigated taking into account all B- and T-cell (CTL and Th) epitopes already described in the literature for the corresponding prototype strains (Fig. S.1). For this analysis, only minimal epitopes fully mapped were considered in order to avoid any bias arising from an overrepresentation of MOMP antigenic regions. Overall, of the total 511 mutations, 176 (34.4%) were found to occur within or adjacent to MOMP B- and T-cell epitopes with species, serogroup or genotype specificities. From these, 88.1% solely involved B-cell epitopes, which is statistically significant as they encompass only about 19% of the total protein length \((P < 10^{-15})\). We observed that 82.6% of these resulted in amino acid replacements, yielding a global nonsynonymous mutation rate almost three-fold higher for these B-cell antigenic regions than for the rest of the \(ompA\) gene \((P < 10^{-5})\), which suggests targeted amino acid variability. More, 25% of these mutations were found to become worldwide fixed within some genotypes \((P < 10^{-5})\) (Fig. 6.2). Also, some of them were previously shown to abrogate monoclonal Ab binding and neutralization of the infectivity of several prototype and/or variant strains \textit{in vitro} [134,136,281]. For instance, the VDIV Ala-to-Thr change observed for genotypes D and Da was previously shown to prevent the antibody binding at the serogroup B 323IAGAG327 epitope for the known genovariant D' [281]. This mutation seems to become fixed within genotypes D and Da, as it was found in strains from five and eight different geographic regions, respectively (Fig. 6.2). Moreover, the VDI Ala-to-Val change found in 25% of the C variant strains from three distinct geographic regions was shown to prevent antibody binding to the 92DVAGL96 epitope (Fig. 6.2), which is known to be highly intolerant to substitution [134]. Although the impact of these changes also depends on the epitope conformation that may vary among genotypes, the existence of such disrupting mutations in MOMP B-cell epitopes clearly evidence a \textit{C. trachomatis} strategy to evade recognition and neutralization by host Abs, allowing thereby the infection to evolve.

Of the 21 mutations occurred within or adjacent to T-cell antigenic regions, 66.7% involved clusters of CTL and Th epitopes. For example, several nonsynonymous mutations exhibited by D, Da, G, H, Ia and J specimens were found within the CDIV species-specific cluster of five HLA class I-restricted minimal CTL epitopes [138] that superimposes (except
for G) a cluster of at least six HLA class II-restricted core Th epitopes [140] (Fig. 6.2). Another interesting example occurred for genotype E, where four nonsynonymous mutations were found within the E-specific CTL epitope $^{177}\text{SLDQSVVEL}^{185}$ [138] that overlaps an E-specific Th-epitope-containing peptide [139,140] and is adjacent to three well-defined B-cell core epitopes [136] (Fig. 6.2A). Curiously, none of the above cited mutations seems to become fixed as they were exhibited solely by one strain of a single genotype or were restricted to a single geographic region, suggesting their nonfixation. In support of this, we found a statistically significant low dN/dS value for MOMP CDIV T-cell epitope region suggestive of purifying selection (Fig. 6.4A).

6.4.5. Evolutionary dynamics of amino acid gains and losses

Considering the above evidenced strong selective pressure acting on this dominant chlamydial antigen, we examined the mutational trend of amino acid gains and losses in MOMP evolution. Therefore, for each prototype strain, a mathematical approach was used to estimate the ‘expected’ frequency of each amino acid considering all possible SNPs for each codon position. Depending on genotype, about 3550 different SNPs may randomly occur in ompA. Curiously, no significant differences were found for the amino acid frequencies calculated using different overall mean $\kappa$ values. The evolution of the amino acid composition of MOMP was determined by comparing these ‘expected’ frequencies (mathematical approach) with the ones ‘observed’ for all types of genetic variants isolated worldwide (Fig. 6.5A). The amino acids Ala, Val, Thr and Ser were the most accrued in all scenarios, while Trp and Met revealed the lowest frequencies. However, frequency discrepancies were found for Met, Phe, Tyr and Cys, suggesting that evolution of amino acid composition in MOMP may be ruled by intrinsic trends that emerge from specific mutational and selective pressures. For example, despite Met is one of the less frequent amino acids, it is being acquired up to 4 times less than the mathematically expected. Also, the observed frequencies for Phe, Tyr and Cys are about 2 times less than the expected. Interestingly, these amino acids were found to belong to the top five ‘gainers’ (His, Arg, Pro, Tyr and Val) or ‘losers’ (Met, Trp, Phe, Cys and Asp) among the worldwide genetic variants, as they presented the highest or the lowest ratios of create/remove substitutions and thus are being evolutionary accrued or lost in MOMP, respectively (Fig. 6.5B). For instance, for each His that is lost, nine are being acquired ($P = 0.007$), while for each Met accrued, 10 are being removed ($P = 0.01$). In contrast to what is traditionally assumed, this nonrandom mutational dynamics suggests that the amino acid composition of MOMP is far from a total equilibrium.
Figure 6.5 - Evolutionary dynamics of amino acid gains and losses. (A) Ratio between the amino acid ‘expected’ frequency (calculated by mathematical modeling using two different overall mean \( \kappa \) values, 1.2 (upper row) and 3.6 (lower row) – see methods) and the ‘observed’ frequency for the 251 genetic variants. Normalized ratio intervals (\( \log_2 \)) are color-coded. For Trp, which was the unique amino acid that was never created from the total 511 variable sites found in \( ompA \), the ‘observed’ frequency was calculated by determining the mid-point of the respective exact 95% confidence interval using the ratio between the F and Binomial distribution [308]. (B) Empirical data showing the ratio of created/removed amino acid substitutions that were ‘observed’ among the 251 genetic variants. Normalized ratio intervals (\( \log_2 \)) are color-coded, and each amino acid is shown below to the correspondent interval. Statistical significance \( (P \leq 0.02) \) was found for the amino acids marked with an asterisk. For panels (A) and (B), logarithmic transformation of the ratios is required for symmetrical distribution of the data around zero.

Although the impact of these mutational trends in MOMP evolution is not known, the gain or loss of certain amino acids were already shown to influence the structure of several bacterial porins [242]. In agreement with this, we found that about two thirds of the 83 substitutions involving the negatively charged residues Asp and Glu occurred in MOMP external loops, where they are known to be involved in the binding to the LPS [294]. Bioinformatically, these substitutions did not reveal significant differences in protein chemical properties, suggesting that only minor porin structural alterations are expected on behalf of adaptive environmental changes. Also interesting was the preservation of the seven Cys residues, which are 100% conserved among all \( C. trachomatis \) genotypes. More, the apparent fixation of the CDI Val-to-Cys change in all B variant specimens (isolated from...
12 distinct geographic regions) indicates that this Cys residue, which is conserved among the remaining genotypes, is important at this specific protein position. Considering their ability to form intra- and intermolecular disulfide bonds, the conservation of a higher number of Cys residues in MOMP reinforces their role in maintaining the membrane structural integrity of a pathogen that lacks the typical peptidoglycan layer found in other Gram-negative bacteria [35].

6.5. Discussion

The detailed understanding of the evolutionary pathway of pathogen’s antigens is of the utmost importance, especially when no efficacious preventive strategies (such as a vaccine) are available yet. In this worldwide survey, we intended to provide a more complete picture of the mutational trends of the *C. trachomatis* leading multi-subunit vaccine candidate.

Contrarily to other chlamydial loci, where polymorphism reflects genotype’s adaptation to different biological niches and ecological success [165], no compelling association was established at molecular, phylogenetic or immunological level for MOMP. Our data revealed that amino acid composition of MOMP is far from a detailed equilibrium as certain residues seem to be consistently accrued (His, Arg, Pro, Tyr and Val) or lost (Met, Trp, Phe, Cys and Asp) in protein evolution (Fig. 5B). This MOMP substitution pattern diverge from the ongoing changes of amino acid frequencies of the remaining *C. trachomatis* chromosome, where Cys, Phe, Arg, Met and His as well as Lys, Leu, Glu Pro and Tyr appear as the five strong gainers and losers, respectively [309]. More, the frequency fluctuations relative to mathematical modelling (Fig. 6.5A) corroborate the singular nature of this key antigen, ruled out by specific mutational and selective pressures. They likely drive protein evolution towards better antigenic fitness, as amino acid changes were almost three-fold more frequent in B-cell antigenic regions (with 25% likely at fixation worldwide) than in the remaining protein (*P* < 10^-5), similar to that previously found in Portuguese isolates [303]. Although the *in vivo* implications of these alterations in MOMP are still unknown, some of them were previously shown to prevent recognition and neutralization by host Abs *in vitro* [134,136,281]. Moreover, this immunodominant protein may be strategically used as reservoir for antigenic variability in a chlamydial decoy-like host immune evasion, as MOMP was found to block the binding of more broadly protective species-common pan-neutralizing antibodies, which have been suggested as potential alternatives for a chlamydial vaccine [62].

It has been argued that both the immunogenicity and adhesion role of MOMP may also depend on the protein native conformation (nMOMP) (still unknown) [283]. Consonant
with this, we have found a nonrandom mutational dynamics involving amino acids (ex. Asp, Glu and Cys) that are thought to be critical for the nMOMP structure [241,242], which may consequently affect the conformation and immunoaccessibility of antigenic regions. Recently, by using a nonhuman primate trachoma model, Kari et al. [300] found that immunization with the trimeric nMOMP yielded striking different neutralizing Ab titers against two genotype A strains that differ solely by four amino acids in MOMP. Although none of these changes fall within known B-cell epitopes, it suggests that any apparently minor alteration may influence the nMOMP structure [300]. Accordingly, we speculate that the 18 nonsilent alterations occurring outside known VD antigenic regions that seem to be at fixation worldwide (Fig. 6.2), may have important conformational immunogenic implications in nMOMP, with likely more impact in exposed B-cell epitopes than in the transmembrane linear T-cell epitopes.

It has been shown that humoral immune responses facilitate antibody-dependent cellular cytotoxicity and, most importantly, boost the induction of an optimal memory Th1 response through rapid uptake, processing and presentation of antigens [310]. Thus, it is expected that evolutionary fixation of mutations in MOMP B-cell antigenic regions may have pathological consequences at the concertedly humoral and cellular immune responses, which may contribute to chlamydial recurrence or persistence, as previously demonstrated for several virus [311]. In MOMP, CTL and Th epitope clusters of diverse HLA specificities are localized in nonpositively selected CDs (Fig. 6.4), which could apparently constitute promising components for a successful multi-subunit vaccine, enabling widespread protective immunization for a genetically diverse human population [137]. An interesting candidate is the MOMP CDIV species-specific cluster of five CTL and six Th epitopes binding three HLA class I and four HLA class II allotypes [138,140], in which variability is thought to disrupt vital protein functions and greatly reduce pathogen fitness [137]. This is supported by our data as none of the nonsynonymous mutations occurred in these cluster epitopes seems to become fixed (as shown by the low dN/dS value for this region – see Fig. 6.4A). However, this lack of antigenic variation on these MOMP T-cell clusters might indicate a pathogen’s evolutionary mechanism that may lead to immune subversion, analogous to the one found for Salmonella enterica serovar Typhimurium [312] and Mycobacterium tuberculosis [313], which show a similar intracellular lifestyle. It is believed that both recognition and consequent cellular immune response to these epitopes might actually provide a net benefit to the pathogen, while may be detrimental to the host [313], as they may maximize the likelihood of pathogen’s transmission or persistence. Reinforcing this, it has been shown that MOMP alone is unable to elicit a Th1 response (predominantly by IFN-γ secretion) sufficiently strong to
resolve chlamydial infection and confer protective immunity [297]. Thus, it could be speculated that this pathogen may have developed ways of avoiding IFN-γ mediated effectors, clearly compromising the Th1-way of reaching effective protection, which seems mandatory for a vaccine against *Chlamydia*.

Intriguingly, a high level of conservation of MOMP seems to be associated with a higher ecological success, as we have previously suggested [303]. Indeed, the two most prevalent *C. trachomatis* genotypes, E and F, showed an *ompA* mutation rate 22.3-fold lower than that of the other genotypes (*P* < 10^-20). As the existence of distinct antigenic profiles among genital genotypes was previously demonstrated [139,140,289], we speculate that the about 12% MOMP amino acid differences among genotypes [165] may confer E and F antigenic profiles better fitted to deal with the host immune system, which would be less prone to change in the light of the Darwinian evolutionary theory. Considering the uniqueness of the chromosomal genetic make-up and evolutionary course of these two genotypes [165], we speculate that E/F specific virulence factors may also contribute for their ecological success, which does not seem to be correlated with the intracellular multiplication rate [167].

In *C. trachomatis*, no weight has usually been given to silent mutations that take place in *ompA*, although they were found to constitute 49.3% and 18.5% of all mutations observed in CDs and VDs, respectively, and some of them seem to be fixed worldwide within some genotypes (Fig. 6.2). However, it has been demonstrated that the generation of synonymous codons may influence the mRNA structure [314], the translation efficiency [315], and the protein folding and tertiary structure [316]. For example, it was shown that some highly expressed genes of *E. coli* [317], *Saccharomyces cerevisiae* [304] and *B. subtilis* [318] have a strong preference for codons recognized by the most abundant tRNA species, which promotes translational efficiency. Similarly, we have found that 75% of all silent mutations that are likely at evolutionary fixation resulted in synonymous codons with high frequency usage in *C. trachomatis* [306]. This “travelling without moving” phenomenon [293] permit the immediate MOMP adaptive landscape to differ among strains without changing its sequence, function or fitness, allowing the protein to modify its evolutionary pathway according to environmental and host immune pressures.

Another interesting but cautious consideration about beneficial or deleterious mutations in *ompA* relates to the “hitchhiking effect“, where an evolutionarily neutral or even deleterious mutation may spread through the population due to its proximity to a beneficial mutation (being dragged in the same recombinant fragment) [319]. This effect may be expected to occur in *ompA* due to its well-described recombinant character [112,116], and it may explain the high frequency of some apparent neutral mutations, such as the seven silent
mutations in close proximity to three potential advantageous nonsynonymous mutations in CDI of numerous B variant specimens from five to 11 geographic regions (Fig. 6.2). One of these favourable mutations was found to provide all B variant specimens with a Cys residue, which is conserved among the remaining *C. trachomatis* genotypes and thus may play an important role in maintaining the membrane structural integrity.

Although the extant variant specimens reported in this worldwide review may have suffered the effects of the immune system of hundreds of different hosts and may represent several thousands of bacterial generations, still, they constitute a flash in the ~100 million-year chlamydial evolution [320]. Even with this extremely short time-scale (about six decades), our results clearly evidence the existence of intrinsic mutational trends for MOMP, a crucial antigen, adhesin, and porin. We observed the existence of distinct evolutionary scenarios in MOMP B- and T-cell epitopes that may benefit to the pathogen. Despite the hypothetical accumulation of some mutations in prototype strains (that were here used as “baselines” for the clinical variants) due to *in vitro* passage (which was never demonstrated for *Chlamydia*), there is no immune pressure associated with the cell-lines used for their passage. Thus, even if some mutations occur in laboratory prototype strains they are driven by genetic drift, and it is not credible that they could mask the statistically supported B-cell targeted fixation of mutations observed worldwide for clinical strains. This hitherto unrevealed extensive variation in B-cell antigenic regions suggests a still undisclosed complex scenario of dynamic antigenic profiles for this unavoidable pathogen’s antigen, likely representing new neutralization escape mutants that continue to co-evolve with the human host. Together with the apparent conserved T-cell scenario, they likely constitute an obstacle for the development of an efficacious MOMP-based vaccine, whose success will require the presence of other protective antigens and/or the identification of novel broadly cross-reacting conformational-dependent MOMP neutralizing Abs, whose specificity and protective function are not affected by this mutational scenario.
CHAPTER 7

*Helicobacter pylori* Diversity Viewed as a Tissue-specific Coevolutionary Arms Race

*Published in*

*Author Contributions*
AN and JPG conceived and designed the experiments; AN performed the experiments; AN and PJN performed the statistics; AN and JPG analyzed the data; AN and JPG wrote the paper; MJB reviewed the paper.
7. **C. trachomatis** Diversity Viewed as a Tissue-specific Coevolutionary Arms Race

7.1. Abstract

The genomes of pathogens are thought to have evolved under selective pressure provided by the host in a coevolutionary arms race (the 'Red Queen's Hypothesis'). Traditionally, adaptation by pathogens is thought to rely not on whole chromosome dynamics but on gain/loss of specific genes, yielding differential abilities to infect distinct tissues. Thus, it is not known whether distinct host organs differently shape the genome of the same pathogen. We tested this hypothesis using *C. trachomatis* as model species, looking at 15 genotypes that infect different organs: eyes, genitalia and lymph nodes. We analyzed over 51,000 base pairs from all genotypes using various phylogenetic approaches and a non-phylogenetic indel-based algorithm to study the evolution of individual and concatenated loci. This survey comprised about 33% of all single nucleotide polymorphisms in *C. trachomatis* chromosomes. We present a model in which genome evolution indeed correlates with the cell type (epithelial versus lymph cells) and organ (eyes versus genitalia) that a genotype infects, illustrating an adaptation to physiologically distinct niches, and discarding genetic drift as the dominant evolutionary driving force. We show that radiation of genotypes occurred primarily by accumulation of single nucleotide polymorphisms in IGRs, HKs, and genes encoding hypothetical and cell envelope proteins. Furthermore, genotype evolution also correlates with ecological success, as the two most successful genotypes showed a parallel evolution. We identified a single nucleotide polymorphism-based tissue-specific arms race for strains in the same species, reflecting global chromosomal dynamics. Studying such tissue-specific arms race scenarios is crucial for understanding pathogen-host interactions during the course of infectious diseases, in order to dissect pathogen biology and develop preventive and therapeutic strategies.

7.2. Introduction

When two species interact with each other, such as a pathogen and human, a never-ending reciprocal and dynamic adaptation process takes place. Whereas the 'goal' of the human being is to try to avoid, solve or minimize the infection, the 'goal' of the pathogen is to deal with this constant host environmental and immune pressure, through genomic evolutionary changes, in order to win this arms race [321-324]. Typically, genome evolution within same-species
Evolutionary History

strains of a pathogen has been studied mainly in the light of horizontal gene transfer (HGT) at specific chromosome loci [325,326], as for *E. coli* [327,328], *S. aureus* [329], *Streptococcus pyogenes* [329], *S. enterica* [330], *Shigella flexneri* [331], and *Pseudomonas syringae* [321]. An extreme example is provided by the well-studied *E. coli*, where strains K-12 and O157 differ by more than 1 million base pairs [332], and same-genotype strains were found to present profound differences in gene content [333,334]. Globally, these targeted HGT events reflect different pathoadaptation processes for microorganisms with reversible genome size-plasticity; depending on the transitory 'cassette-genes' carried at any specific time, the pathogenecity or ability of these microorganisms to infect different tissues may vary [327]. Thus, generally, these processes rely on gain/loss of virulence/colonization factors rather than reflect whole chromosomal dynamics, the evaluation of which remains complex. Indeed, assessment of tissue-specific adaptive evolution at the whole genome level demands that same-species strains of a pathogen specifically and nontransitorily infect different tissues. Therefore, on behalf of the arms race theory assumed by the evolutionary Red Queen's Hypothesis [335,336], one question arises: do distinct host organs differently shape the genome of the same pathogen? No microorganism is more suitable than *C. trachomatis*, the most prevalent sexually transmitted bacterial pathogen worldwide, to test this hypothesis, as the species comprises several genotypes with a wide range of specific human tissue tropism. This pathogen is mainly classified into 15 genotypes based on the differential immunoreactivity of MOMP, constituting three disease groups [4]: genotypes A-C and Ba are commonly associated with ocular trachoma; genotypes D-K infect the epithelial cells of genitalia and are normally found in non-invasive STIs (where genotype E represents about one-third of all infections, and together with genotype F constitute up to 50% of them); genotypes L1-L3 are also sexually transmitted but are invasive and disseminate into the local lymph nodes causing LGV. However, in the context of this classification system, the evaluation of adaptive evolution becomes enigmatic because there is no correlation between it and *C. trachomatis* tropism nor with the ecological success of the different genotypes, as strains with different organ specificities are placed within the same classification group.

As occurred for *Mycobacterium leprae* [337], *Rickettsia prowazekii* [338], and the aphid endosymbiont *Buchnera aphidicola* [339], the first stages of *Chlamydia* evolution consisted of a massive genome reduction upon becoming an obligate intracellular parasite [2,340]. However, comparative genomics over the few currently fully sequenced *C. trachomatis* genomes [2,5,7,37] revealed that gene decay is not involved in the more recent evolutionary stages. Indeed, contrary to most pathogens, the core- and the pan-genome [326]
of this microorganism are near identical, indicating that the factors involved in the differential organ specificity among genotypes are not acquired by gene transfer [7].

To evaluate if distinct arms races occur between different infected human organs and this pathogen's serovars, we performed high-scale concatenation-based phylogenomics, using about one-third of all chromosome SNPs. So far, in contrast to the ocular group, only one strain from the epithelial-genital and LGV groups has been fully sequenced [2,5,7,37], making our multiple-loci scrutiny of all 15 genotypes the ideal tool to track the evolutionary diversity of a microorganism characterized for its distinct infection niches. Here, we show a matchless model of SNP-based adaptive evolution of same-species strains to each infected cell-type and organ that relies on whole chromosome evolutionary dynamics, unlike previous reports for other pathogens focused on specific gene gain/loss.

7.3. Materials and methods

7.3.1. Culture of C. trachomatis prototype strains

We used the most common prototype strains representing the 15 main C. trachomatis genotypes: A/Har13, B/TW5, Ba/Apache2, C/TW3, D/UW3, E/Bour, F/IC-Cal3, G/UW57, H/UW43, I/UW12, J/UW36, K/UW31, L1/440, L2/434 and L3/404. McCoy cell culture of all strains plated in T25 cm² flasks was performed as previously described [341]. At 48-72 h (pi), EBs were harvested, and DNA was extracted using QIAamp® DNA Mini Kit (Qiagen) according to the manufacturer's instructions. Genotype confirmation of each prototype strain was performed using ompA genotyping with BLAST comparison of the available GenBank sequences.

7.3.2. Selection of loci

A GenBank search was performed to look for genomic regions that had been sequenced for at least one C. trachomatis prototype strain from each of the three disease groups. Up to 93 loci were found, comprising about 84,000 bp of the chromosome, and involving IGRs, HKs, and genes encoding hypothetical or unclassified proteins (HPs) and cell envelope proteins (CEPs). Only nonconstant loci were selected (51 of the 93; Fig. 7.1 & Table S7.1) for sequencing the other prototype strains if their sequences were not available yet. Automated sequencing was performed as previously described [121]. The DNA sequence data have been deposited in a public database (GenBank: EU239694-EU239702, GenBank: EU239705-EU239712, and GenBank: EU247618-EU247753). Primer sequences are given in Table S7.2. For all strains, five types of concatenated sequences were created in a head-to-tail fashion: one for each loci.
category (IGRs, HKs, HPs and CEPs) and a global concatenated sequence involving all *loci* (approximately 50,000 bp for each taxon).

7.3.3. Polymorphism significance

We used data from the fully sequenced genomes A/Har13 and D/UW3 for this evaluation. Thus, considering the 3,354 SNPs identified between these two genomes [5], we evaluated whether 1,099 SNPs restricted to the 51,074 bp analyzed in this study are overrepresented relative to the 2,255 SNPs found in the rest of the chromosome. We framed this as a contingency table (Table 7.1) with a restricted sequence of 1,042,519 bp for each strain (corresponding to the length of the D/UW3 chromosome), and we estimated P-values using the Fisher's exact test as well as the odds ratios with a 95% confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>Inside region</th>
<th>Outside region</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>1,099</td>
<td>2,255</td>
</tr>
<tr>
<td>Without SNPs</td>
<td>49,975</td>
<td>1,465,713</td>
</tr>
</tbody>
</table>

Table 7.1. Contingency table for estimating polymorphism significance

7.3.4. Genomic analysis

For all individual *loci* and concatenated sequences, alignments of all strains were generated using LaserGene (DNASTAR) and MEGA 3.1 (http://www.megasoftware.net). MEGA 3.1 was also used to create matrices of pairwise comparisons and to estimate the number of variable sites, the number of parsimony informative sites and overall mean genetic distances. The pairwise-deletion option was chosen to remove all sites containing missing data or alignment gaps from all distance estimations, only when the need arose and not prior to the analysis.

In order to search for distinct regions that may be associated with strains belonging to a specific disease group, SimPlot 3.5.1 (http://sray.med.som.jhmi.edu/SCRoftware/) was used on all 51 *loci*. For each similarity plot, genotypes were grouped according to the cell-type/organ that they infect, and nucleotide pairwise distances were calculated using the K2P method (gaps excluded; ts/tv of 2.0) in a sliding window size of 160 bp moved across the alignment in a step size of 10 bp. Additionally, for all *loci* where genotypes E and F clustered apart from the other epithelial-genital genotypes was observed, a SimPlot analysis was also performed to evaluate if the E/F nucleotide differences compared to the other genital genotypes were clustered in specific domains of each *locus*.
To evaluate the existence of foreign genetic material [342], SWAAP 1.0.2 ([http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm](http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm)) was used to calculate the percentage of GC content for all sequences of each taxon.

7.3.5. Phylogenetic analysis

Prior to the phylogenetic reconstructions, and in order to select the appropriate evolutionary models, we evaluated the homogeneity of substitution patterns between sequences by calculating the Monte Carlo test-based Disparity Index per site [343]. This gives the probability of rejecting the null hypothesis that sequences have evolved with the same pattern of substitution. The NJ method [264] was used with K2P [265], Jukes-Cantor [344] and TN [345] models to generate phylogenies. For the concatenated sequences, in order to examine the accuracy of the major conclusions reached from the NJ analysis, trees were also constructed under the MP criterion [266], using the max-min branch-and-bound algorithm.

Considering that recombination disturbs a phylogenetic signal since the two parts of the recombined region may have different evolutionary histories [346,347], one locus (ompA) was excluded from the phylogenetic concatenated analysis, as its highly recombinant nature has already been demonstrated [112,116]. The use of outgroup sequences was discarded in the present study because no rooted trees were needed to achieve the objectives defined above. Also, the most suitable strain for use as an outgroup, C. muridarum (MoPn strain), has several loci that vary greatly in size and diverge from those in the C. trachomatis strains, which would entail the removal of a huge portion of the sequences being analyzed.

7.3.6. Evolutionary $\gamma$-distance

We used a non-phylogenetic method for estimating the evolutionary distance between each pair of homologous DNA sequences, which is given by the parameter $\gamma$ [348]:

$$\gamma = -2 \log_e P$$

where

$$P = \frac{n_{xy}}{\sqrt{n_x n_y}}$$

where $n_{xy}$ is the number of nucleotides shared by the two sequences, and $n_x$ and $n_y$ are the number of nucleotides of each sequence. For comparative purposes, we used the same set of loci as for the phylogenetic concatenated analyses, that is, we excluded the recombinant ompA gene. The $\gamma$ variability was estimated by Monte Carlo using the alignments of each individual locus through the statistical platform R 2.5.1 ([http://www.r-project.org/](http://www.r-project.org/)). Each time, 20 loci
were randomly selected with replacement and $\gamma$-distances were calculated by repeating this procedure 50 times.

7.3.7. Statistical analysis

The statistical association between genetic and phylogenetic variables was performed using the analysis of variance (ANOVA) test by comparing groups' population means. We considered as genetic variables the overall mean values of percent GC content, p-distance and absolute SNPs obtained for each of the selected loci. The phylogenetic variables were: clustering of strains according to tropism properties; co-segregation of E/F strains; segregation of a LGV cluster or an ocular cluster; and the 'weight' of each locus in the final concatenated tree. The homogeneity of variances was tested using the Levene's test. Whenever the hypothesis of homogeneity of variances was rejected, the nonparametric Mann-Whitney test was used to compare distributions among groups. A $P$-value of 0.05 or less was considered significant.

7.4. Results

7.4.1. Evaluation of the degree of polymorphism for the selected loci

Considering that the strain radiation yielding the present-day chlamydial genotypes likely occurred over millions of years [320], the use of prototype strains is an accurate strategy as they were isolated only a few decades ago. Thus, in this evolutionary survey, we used the traditional prototype strains that represent the 15 main C. trachomatis genotypes. We selected 51 polymorphic loci (approximately 51,000 bp) dispersed throughout the chromosome (Figs. 7.1 & S7.1) that represent the following loci categories: 16 IGRs, 16 CEPs, 13 HKs, and 6 HPs (Table S7.1).
Figure 7.1 - Loci distribution in the ~1.04 Mb C. trachomatis circular chromosome. Gene names and ORF numbers are based on the C. trachomatis D/UW3 genome annotation (GenBank: AE001273). Loci categories are illustrated by different colors. Only the first nucleotide of each locus is marked on the figure.

In order to evaluate the degree of polymorphism of these loci in comparison with the whole chromosome, we used the data generated from two of the five fully sequenced genomes, A/Har13 (ocular) [5] and D/UW3 (epithelial-genital) [2]. We observed in the studied 51 loci a global mutation rate 9.7-fold higher than in the remaining chromosome regions ($P < 0.001$). Moreover, we found 1,099 SNPs in these 51 loci between A/Har13 and D/UW3, which is greater than 200-fold more than what has been studied to date through concatenation [349], and comprises about 33% of the whole chromosome SNPs, indicating that our results could be scaled up to the full-chromosome level.

Additionally, a global overview of GC content revealed a mean value for all loci categories (data not shown) that is similar to the total mean GC content of approximately 41% observed for the fully sequenced genomes [2,5,7,37] with a SD of 2.9%, which is not indicative of any putative HGT event.

7.4.2. Correlation of individual loci with tissue-specific strain radiation

We used phylogenomics to correlate each individual locus with tissue-specific strain radiation. Only four (25.0%) CEPs ($incD, incE, pmpF$ and $pmpH$) and
one (6.3%) IGR \((incD/incE)\) comprehensively grouped the strains according to their cell-type/organ appetite (that is, revealed a larger evolutionary distance between strains with different niche appetencies than between strains infecting the same niche; Fig. 7.2A). This clustering seems to be associated with loci revealing a higher p-distance-based polymorphism \((P = 0.025)\). A full segregation by cell-type/organ appetence was not seen for most of the remaining CEPs due to the heterogeneity among the genital strains, where genotypes E and F frequently form a separate cluster for 62.5% of CEPs (Fig. 7.2A).

**Figure 7.2 -** Phylogenomics of individual loci versus strain segregation. (A) Phylogenetic strain segregation. Loci categories are illustrated by different colors. Numbers on the top of each bar show the percentage of loci, within each category, that generate a tree where a full tissue tropism, or a particular cluster of strains, or an E/F co-segregation is observed. (B) Percentage of loci (within each functional category) for which the majority of SNPs yield an amino acid change. The color scheme for the represented loci categories is the same as (A).

Globally, 77.6% of loci belonging to different functional categories grouped strains that invade the lymph nodes as an individual cluster (LGV cluster), and the clustering of strains infecting the ocular tissue (ocular cluster) was also frequent (Fig. 7.2A). As above, we identified a significant association between a higher absolute number of SNPs and both the occurrence of a LGV cluster and an ocular cluster for each locus \((P = 0.037\) and \(P = 0.045,\) respectively). Interestingly, from the loci that better illustrate adaptation to lymph nodes, 80% of HPs and 53% of CEPs, compared with only 29% of HKs, show >50% non-synonymous SNPs (Fig. 7.2B). Considering the DNA replication process, all SNPs on one strand that may imply strain segregation will also have the same impact on the other DNA strand. However, from the 51 loci that we used, only 4 pairs of loci overlap and the overlapping region never exceeds 10 bp (data not shown), which makes this effect negligible. Overall, these results suggest that the distinct genetic variability of strains infecting a specific celltype/organ likely reflects an evolutionary adaptation process.

By performing intra-locus analysis, we observed that three HPs (CT049, CT144 and CT622) and two IGRs \((rs2/ompA and ompA/pbpB)\) revealed distinct domains in which SNPs
are concentrated, instead of being randomly distributed, and are associated with strains that infect a specific cell-type/organ (Fig. 7.3).

**Figure 7.3** - Identification of *loci* domains characteristic of strains infecting a specific biological niche. SimPlot graphs show the nucleotide similarity between the ocular, epithelial-genital and LGV strains for (A) CT049, (B) CT144, (C) CT622, (D) rs2/ompA IGR and (E) ompA/pbpB IGR. Epithelial-genital (pink) and LGV (blue) strains are compared to the ocular strains (represented in the upper x-axis). For CT622 (C) and rs2/ompA IGR (D), where an E/F clustering apart from the other epithelial-genital strains was observed, SimPlot analysis has also involved genotypes E/F (green). For each panel, the *loci* domains that are specific to LGV, epithelial-genital, ocular or E/F strains are bordered by boxes in blue, pink, yellow and green, respectively. For panels (C) and (D), LGV and E/F specific domains partially or completely overlap, respectively. The represented domains correspond to a non-random fixation of SNPs, yielding clusters of amino acid changes.
For these HPs, the SNP domains correspond to clusters of amino acid changes in the protein sequence (data not shown), mirroring the previous findings for some polymorphic membrane protein genes [121]. Unfortunately, there is no assigned role for these open reading frames, which rules out any speculation about the functional implications of these specific clustered amino acid alterations. Nevertheless, this tissue-specific amino acid clustering points to a targeted fixation of mutations that may reflect the host-pathogen specific interaction within each organ.

7.4.3. Genomic analysis of the concatenated loci

We evaluated the nucleotide sequence variation in each concatenated loci category (Table 7.2). We highlight the multi-loci concatenation approach as a powerful tool to generate robust phylogenomic inferences, even when individual loci have evolved with different substitution patterns [350-352]. Overall, the HPs exhibit the highest number of variable sites (10.3%), whereas the HKs are the least variable (3.3%), which is supported by the mean p-distance values. Curiously, the IGRs show polymorphism similar to the CEPs. Globally, concatenation of all 51 loci yielded a ‘super’ sequence of up to 51,074 bp for each of the 15 prototype strains, showing a mean of 1,032.1 [SE 17.2] nucleotide differences.

<table>
<thead>
<tr>
<th>Loci Categories</th>
<th>IGRs</th>
<th>HKs</th>
<th>HPs</th>
<th>CEPs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>3834</td>
<td>7984</td>
<td>3906</td>
<td>35350</td>
<td>51074</td>
</tr>
<tr>
<td>Variable sites&lt;sup&gt;a&lt;/sup&gt;</td>
<td>192 (5.0%)</td>
<td>267 (3.3%)</td>
<td>402 (10.3%)</td>
<td>1800 (5.1%)</td>
<td>2662 (5.2%)</td>
</tr>
<tr>
<td>Parsimony informative sites&lt;sup&gt;b&lt;/sup&gt;</td>
<td>147</td>
<td>233</td>
<td>372</td>
<td>1656</td>
<td>2408</td>
</tr>
<tr>
<td>Overall mean distance (nt)</td>
<td>[SE 4.7]</td>
<td>[SE 5.3]</td>
<td>[SE 7.6]</td>
<td>[SE 15.6]</td>
<td>[SE 17.2]</td>
</tr>
<tr>
<td>Overall mean p-distance (nt)</td>
<td>0.0172 [SE 0.0013]</td>
<td>0.0121 [SE 0.0007]</td>
<td>0.0430 [SE 0.0020]</td>
<td>0.0199 [SE 0.0004]</td>
<td>0.0202 [SE 0.0003]</td>
</tr>
</tbody>
</table>

Calculations based on the alignment of the 15 strains.
<sup>a</sup> The % value is relative to the respective size of each loci category.
<sup>b</sup> The % value is relative to the respective number of variable sites of each loci category.

7.4.4. Evolutionary history of C. trachomatis

Due to the speed and efficiency of the NJ method in inferring large phylogenies [353,354], we used this approach on concatenated data. The NJ phylogenies inferred from the four concatenated loci categories (Fig. S7.2) are consistent with most of the respective individual loci trees. Although only the CEP category clearly segregates strains by the disease they
cause, the other categories show a notable segregation of at least one disease group, suggesting that heterogeneous loci categories are involved in the arms race process. The global phylogenetic tree presented in Figure 7.4 (where each taxon is represented by about 50,000 bp) reveals the putative final picture of *C. trachomatis* evolution, showing strain grouping according to the cell-type (epithelial and lymph cells) and organ (eyes and genitalia) that they infect.

**Figure 7.4** - *C. trachomatis* evolutionary history. The global phylogenetic tree (NJ, K2P model) is based on about 50,000 bp/taxa. Bootstrap values (1,000 replicates) are shown next to the branch nodes. Ocular, epithelial-genital and LGV strains are represented within yellow, pink and blue boxes, respectively. Charts show the loci contributing to taxa segregation for the assigned tree branches, where the most prominent ones (genetic variability >4%) are highlighted with the corresponding color. Within these highlighted loci, the ones revealing polymorphism (defined as ≥10 SNPs, or >50% amino acid changes when <10 SNPs) among strains infecting the same organ (eyes or lymph nodes), may be involved in pathogenesis (marked with asterisks). Loci without polymorphism within strains infecting the same organ likely reveal the final stages of adaptive evolution (underlined).
C. trachomatis Evolutionary History

These distinct segregations are supported by maximum bootstrap values (99-100%) in the nodes that separate disease groups, reinforcing that the targeted and distinct fixation of nucleotide changes on strains infecting a specific cell-type/organ are likely adaptive and barely the consequence of genetic drift. In fact, the genetic distance matrix (Table 7.3) shows that all strains that preferentially infect the eyes revealed only 0.27% [SE 0.02%] differences among them, but shows a mean genetic distance 7.4- and 11.2-fold higher (corresponding to 983 [SE 20] and 1,484 [SE 42] nucleotides) to strains infecting the epithelial-genital and lymph node tissues, respectively. Also, the LGV strains differ by only 69 [SE 8] nucleotides, whereas their distance to the epithelial-genital strains is 1,226 [SE 34] nucleotides. A separate main branch involving all epithelial-genital strains was not comprehensively seen for any individual loci (except for the CEPs pmpF and pmpH; data not shown) due to the separation of E and F strains. Indeed, the later has a mean genetic distance of 673 [SE 16] nucleotides to the other epithelial-genital strains (Table 7.3). Similar NJ tree topologies were obtained for the three models used to estimate evolutionary distances (K2P, Jukes-Cantor or Tamura-Nei) as well as for the maximum parsimony method (data not shown), with only slight variations in the bootstrap values, which supports the robustness of these distinct arms race scenarios.

### Table 7.3. Overall mean genetic distances within and between disease groups

<table>
<thead>
<tr>
<th></th>
<th>Nucleotide differences</th>
<th>Genetic distance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within-group means</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ocular</td>
<td>133 [SE 10]</td>
<td>0.27 [SE 0.02]</td>
</tr>
<tr>
<td>Genital (w/o E/F)</td>
<td>309 [SE 14]</td>
<td>0.63 [SE 0.03]</td>
</tr>
<tr>
<td>Genital (with E/F)</td>
<td>460 [SE 13]</td>
<td>0.93 [SE 0.03]</td>
</tr>
<tr>
<td>LGV</td>
<td>69 [SE 8]</td>
<td>0.14 [SE 0.02]</td>
</tr>
<tr>
<td><strong>Between-group means</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ocular/Genital (w/o E/F)</td>
<td>942 [SE 21]</td>
<td>1.91 [SE 0.04]</td>
</tr>
<tr>
<td>Ocular/Genital (with E/F)</td>
<td>983 [SE 20]</td>
<td>2.00 [SE 0.04]</td>
</tr>
<tr>
<td>Ocular/LGV</td>
<td>1484 [SE 42]</td>
<td>3.02 [SE 0.08]</td>
</tr>
<tr>
<td>LGV/Genital (w/o E/F)</td>
<td>1241 [SE 36]</td>
<td>2.52 [SE 0.07]</td>
</tr>
<tr>
<td>LGV/Genital (with E/F)</td>
<td>1226 [SE 34]</td>
<td>2.49 [SE 0.07]</td>
</tr>
<tr>
<td>Genital/E-F</td>
<td>673 [SE 16]</td>
<td>1.37 [SE 0.03]</td>
</tr>
</tbody>
</table>

*Genetic distances were estimated in the concatenated ~50,000 bp/taxa.

We also highlight the loci that most contribute to the final tree topology (Fig. 7.4), as they may be relevant for the evolutionary adaptation to each specific niche. Among these loci, we have found either highly conserved or polymorphic loci for strains infecting the same cell-type/organ. The former may represent a step forward in the evolutionary process by revealing the final stages [323] of this tissue-specific adaptive evolution, while the later may also be involved in pathogenic differences between strains infecting the same tissue [37].
most extreme case is given by the CEP \textit{pmpF}, where all the strains that infect the lymph nodes are 100\% similar but show a mean distance of 312 and 421 SNPs to strains infecting the epithelial-genital and ocular tissues, respectively. In contrast, the epithelial-genital strains reveal up to 129 SNPs among them (data not shown). Although less markedly, CT049 is polymorphic among the LGV strains but near 100\% identical among the ocular strains.

Additionally, we identified \textit{loci} that do not seem to have influenced adaptation to each niche, since they generate an incongruent strain radiation (Table 7.4), and whose polymorphism may thus be a consequence of genetic drift. However, previous results have demonstrated the involvement of some of these \textit{loci} (CT622, \textit{tsf}, \textit{rs2} and \textit{pbpB}) in the pathogenesis of trachoma [37]. As expected because of the genotype multiplicity, the epithelial-genital group revealed a higher number of polymorphic \textit{loci}, and, overall, these \textit{loci} belong to different categories. In contrast, strains infecting the lymph nodes constitute the most homogeneous group.

\begin{table}[h]
\centering
\caption{Polymorphic \textit{loci} among strains that infect the same biological niche}
\begin{tabular}{llll}
\hline
& Ocular & Epithelial-genital & LGV \\
\hline
\textit{rs2/ompA} IGR & \textit{rs2/ompA} IGR & \textit{tsf} ompA \\
\textit{ompA/pbpB} IGR & \textit{karG} & \textit{ompA} \\
\textit{tsf} & & \\
\textit{ompA} & CT049 & CT144 \\
\textit{pbpB} & CT622 & CT622 \\
\textit{pmpC} & pmpA & \\
\textit{ompA} & & \\
\textit{pmpE} & & \\
\hline
\end{tabular}
\end{table}

The represented \textit{loci} may hypothetically be involved in the pathogenesis of ocular trachoma, genital infections or LGV disease, but do not contribute to the adaptive evolution of strains to each correspondent biological niche (Fig. 7.4). Polymorphism was defined as >4\% nucleotide differences or >40 SNPs (for genes >3 Kb).

\textbf{7.4.5. Impact of small indels on tissue-specific strain radiation}

In order to have a more complete picture of the evolution of the genotypes, we studied the chromosomal occurrence of small indel events, which are non-phylogenetic parameters. We observed 84 small indel events (from 1-43 bp) inside the global concatenated \textit{loci} for all strains, which mainly occurred within the IGR and CEP categories (Table S7.3). None of these events was found to disrupt the coding sequence of the respective \textit{loci}, indicating the absence of gene decay in the studied regions.
For the global concatenated data, we estimated the evolutionary distances using the indel-based parameter $\gamma$ [348], which computes the number of gap nucleotides per nucleotide site between those sequences, while SNPs are not considered. The $\gamma$-distances (Fig. 7.5A) are highly concordant with phylogenomic analyses, showing high heterogeneity within the epithelial-genital strains, and remarkable homogeneity among the LGV strains. Also, they revealed a segregation of strains by their cell-type/organ appetencies, which supports the tissue-specific arms race scenario.

![Figure 7.5 - Impact of indel events on C. trachomatis evolution and ecological success.](image)

(A) Evolutionary $\gamma$-distances for the global concatenated data within (colored boxes) and between (grey boxes) disease groups (ocular, epithelial-genital and LGV). Boxes represent the variability of all distance estimates, while the vertical line within each box divides 50% of all values. The minimum and maximum distances are represented by the extremes of each horizontal line. (B) Impact of indel events on C. trachomatis's ecological success. The mean $\gamma$-distances from any ocular (yellow), epithelial-genital (pink) or LGV (blue) strain to E/F strains are represented in parentheses. Each evolutionary distance was normalized against the distance between E and F. The relative length of each line is represented in the correct scale.

7.4.6. Evolutionary inferences on the ecological success

Analysis of the global phylogenetic tree (Fig. 7.4) also shows that the two most prevalent genital genotypes worldwide, E and F, are closely related and separated from the other epithelial-genital strains. This segregation is observed for the majority of loci, with the exception of the HPs (Fig. 7.2A). From all these loci, 70% of CEPs show an amino acid replacement for >50% of SNPs, compared to only 20% of HKs (Fig. 7.2B). Curiously, the most remarkable segregation of E and F was seen for two IGRs ($rs2/ompA$ and $yfh0_1/parB$) and three HKs ($karG$, $tsf$ and $rs2$) (Fig. 7.4). Furthermore, for the still unclassified protein gene CT622 and for the $rs2/ompA$ IGR, we observed a non-random distribution of SNPs that
are present in genotypes E and F but not in the other epithelial-genital strains (Figs. 7.3C & 7.3D). Finally, the mean $\gamma$-distance from any epithelial-genital strain to genotype E or F was from 3.4-fold (between G and E/F) to 4.7-fold (between I and E/F) higher than the distance between E and F (Fig. 7.5B), which supports this close relationship between the two most ecologically successful genotypes.

7.5. Discussion

We have hypothesized that distinct arms races may occur inside the same host when the same pathogen is able to infect different organs. In contrast to free living bacteria, where HGT is strongly associated with a pathogen's adaptive evolution [321,325-331], Chlamydia has been characterized by genetic isolation and, while cumulative studies suggest that HGT has almost certainly occurred in Chlamydiaceae [355-357], there is no report to date of transferable mobile elements in C. trachomatis. Here, we demonstrate that C. trachomatis strains that preferentially infect the eyes, the epithelial-genital cells or the lymph nodes present a distinct evolutionary pattern likely illustrating a SNP-based tissue-specific arms race.

In order to develop a more compelling argument for a causal link between genome profile and cell/organ appetite, the use of genetic modification and especially the use of animal models are appealing approaches. However, C. trachomatis is genetically non-tractable and, except for the cynomolgus monkey (accurate for studying the trachoma pathology) [37], no suitable animal model exists for the three types of C. trachomatis disease. Also, there is no in vitro model, such as cell culture, that mirrors the chlamydial infection in vivo, and it has been previously demonstrated that intensive serial passaging of chlamydial strains yielded no mutations on the most variable chlamydial gene (ompA) [358]. Furthermore, it would be inconceivable that these approaches could represent millions of years of chlamydial evolution.

It is believed that the LGV biovar was the first to diverge from a common C. trachomatis ancestor when new primate hosts evolved after the dinosaur extinction, whereas separation of genital and ocular genotypes might have occurred with the appearance of early humanoid primate hosts [320]. The skill to colonize different organs and cell-types likely developed through indel events and SNP accumulation on virulence/colonization factors. So far, chlamydial putative virulence factors, such as the type III effector tarp [5], the cytotoxin gene [107], and especially the tryptophan operon [108,109], are the best candidates for providing that skill. In particular, while the first of these factors differentiates the LGV strains from the other groups, the other two differentiate the strains colonizing the genitalia.
from the strains colonizing other niches. For example, it was clearly demonstrated that only strains possessing a functional trpRBA operon are able to colonize the genital tract [108]. With respect to T3S effectors, although their role in C. trachomatis tropism is not clear, it was shown that evolutionary genetic diversification of the T3S effector HopZ family, via horizontal transfer, had clear implications for Pseudomonas syringae host specificity [321]. However, none of the chlamydial putative virulence factors fully explain the existence of the three major tropism groups made up from the different genotypes. Also, the putative emergence of tissue-specific adhesins cannot be discarded.

With regard to our results (Fig. 7.4), strain radiation within each disease group likely occurred because of accumulation of mutations throughout the chromosome caused by environmental and immune pressure in each niche, giving rise to the contemporary genotypes. Within the genitalia, the higher genotype multiplicity and radiation of epithelial-genital strains compared to the LGV strains would be unexpected in the light of the earlier evolutionary divergence of the later [320]. However, besides the different host immune responses in those niches, the epithelial-genitalia environment presents pH and hormonal fluctuations that are variable among individuals, and also an abundant nutrient-competing flora, which could have strongly influenced the evolutionary pathway of the infecting strains. In support of this, nutrient-competing flora were shown to be a major factor in the successful pathoadaptation of S. enterica serovar Typhimurium to the intestinal tract, as the inflammatory process induced by this pathogen was shown to make a negative impact on mainly the other colonizing microrganisms and, thus, a positive impact on its arms race with the host [359].

Globally, we have observed that the loci that most contribute to strain segregation by cell-type/organ are spread throughout the chromosome (Fig. 7.1) and belong to different functional categories, suggesting that this dynamic evolutionary adaptation is a general trait of the entire genome. Whereas the contribution of CEPs is likely associated with putative structural, antigenic or host-adhesion roles, no assumption can be made for the HPs. However, we found that HPs were the most variable among the genotypes, with an overall polymorphism 2.2-fold higher than the CEPs (Table 7.2), which suggests a higher involvement in chromosomal dynamics. With respect to IGRs, we speculate that their contribution to strain segregation may be associated with recombination events that may promote genetic variability, as we recently described [114]. Nevertheless, the high variability of IGRs was surprising, as they commonly involve regulatory regions that are expected to be conserved; thus, the existence of random genetic drift may also be considered for IGRs. Finally, although the HKs are involved in strain segregation, the vast majority of them
showed <50% nonsynonymous mutations (Fig. 7.2B), which is consistent with their role in essential biological functions.

It is known that in populations without HGT and with bottlenecks, as is the case for *C. trachomatis*, random genetic drift can play a major role in evolution, being responsible for the fixation of unfavorable mutations [360]. However, our results suggest that chlamydial strain segregation according to tropism properties occurred mainly through an adaptive evolutionary process and not through dominant genetic drift. Several arguments point in this direction: the statistical association found between most polymorphic loci (number of SNPs/loci and p-distance/loci) and the strain clustering according to their tissue specificity; *Chlamydiae* presents a relatively high ratio of nonsynonymous to synonymous changes when compared, for example, to *E. coli* and *Buchnera* [320], further supported by our findings where the majority of HPs and CEPs involved in the segregation of the LGV strains showed >50% non-synonymous SNPs (Fig. 7.2B); for at least eight loci (CT049, CT144, CT622, *pmpE*, *pmpF*, *pmpH*, rs2/ompA IGR and *ompA/phpB* IGR), we observed a nonrandom fixation of SNPs exclusive of same niche-infecting strains (Fig. 7.3), corresponding to specific clusters of amino acid changes in coding sequences; the extremely robust global phylogenetic tree with maximum bootstrap support (99-100%) in the branch nodes where strains are separated by their cell-type/organ specificity (Fig. 7.4); 20 out of the 22 loci that contribute to the segregation of strains that preferentially infect the eyes are also involved in the segregation of strains that colonize the lymph nodes (Fig. 7.4) by presenting a dissimilar and specific SNP pattern; and finally, the well-known differences in environmental and immune pressure as well as competing flora and physiological specificities between ocular, epithelial-genital and lymph node tissues.

Within all the loci that are more likely to be involved in the adaptive evolution to each specific niche, we have found either highly conserved or polymorphic loci among strains infecting the same cell-type/organ (Fig. 7.4), where the most remarkable examples are *pmpF* and CT049 (see Results). We hypothesize that *pmpF* and CT049 may be good representatives of a final stage of the adaptive evolution to the lymph nodes and the eyes, respectively, considering their extreme conservation among the corresponding strains. On the other hand, these genes may be responsible for pathogenic differences among epithelial-genital and LGV strains, respectively, based on their strong polymorphism among the corresponding strains. While PmpF has been implicated as a potential target for the host immune response, as it contains several putative MHC epitopes [5], biological information for CT049 is lacking.

Additionally, we found several loci that are polymorphic among strains infecting the same cell-type/organ that seem not to have been involved in the adaptation to each niche, but
which may have been involved in the pathogenesis of trachoma, genital infections or LGV disease (Table 7.4). Indeed, 4 of these loci (CT622, tsf, rs2 and pbpB) belong to a pool of 22 genes that are responsible for profound differences in virulence among two *C. trachomatis* ocular strains in nonhuman primates [37].

Interestingly, we also observed a clear evolutionary co-segregation of the two most ecologically successful genotypes (E and F). This is intriguing as there is a 15% difference between them in the gene coding for the major antigen (MOMP), which constitutes about 60% of the membrane dry-weight [133] and is a putative cytoadhesin [54]. Although it is not known why genotypes E and F are the most prevalent worldwide, their ecological success seems not to be associated with intracellular multiplication rate [167], indicating that it is likely defined at the host cell adhesion and entry steps. However, the existence of E/F specific virulence factors or adhesins cannot be addressed in this study. Even so, tarp is the unique virulence factor that distinguishes genotype E from the other epithelial-genital genotypes (including F), as it presents fewer repeat motifs in the 5' region [5], but its phenotypic consequences are not known. Moreover, a more successful host immune evasion could also be speculated for genotypes E and F considering the well-known different antigenic profile among epithelial-genital genotypes [112].

Regarding the loci that most markedly contribute to the segregation of genotypes E and F, we highlight the tsf, rs2 and rs2/ompA IGR (Fig. 7.4). The first two of these may be involved in hypothetical differences in strain growth [37], while the last involves the regulatory region of rs2. This IGR includes specific domains where most SNPs are exclusive of strains E and F (Fig. 7.3D), suggesting a potential impact on the rs2 regulation and, thus, on strain growth. Also, the rs2/ompA IGR is a recombination hotspot for the generation of mosaic structures within chlamydial strains [114], suggesting that recombination may contribute to the ecological success of the two genotypes. However, as most SNPs of the CEPs involved in the E/F segregation confer amino acid replacements (Fig. 7.2B), we suggest that the positive selection for the membrane proteins may also be a driving force for the E/F evolutionary divergence, likely through antigenic variability.

It is not surprising that bacterial populations that evolved in different ecological niches have different profiles of genetic variability. However, contrary to all previous reports for other pathogens focused on HGT events and gene decay, we present evidence of SNP-based, tissue-specific evolutionary adaptation relying on whole chromosome dynamics, as a consequence of the occurrence of dissimilar arms races between the pathogen and diverse host organs. Answering the proverbial question of ‘which came first’ (tropism or SNPs), the scenario presented here suggests that while some SNPs, on very few and specific loci, are
likely responsible for tropism differences, the vast majority of SNPs throughout the chromosome are a consequence of different tissue tropisms and are expected to be involved in maintaining organ appetite, as *per* the Red Queen's Hypothesis. Mirroring bacterial virulence [326], we present evidence that a 'one size fits all' approach cannot be applied to adaptive evolution. This phenomenon is illustrated by a pathogen believed to infect 140 million people, where the incidence rate can be as high as 30% among adolescent females (http://www.who.int). We believe that grasping a pathogen's genetic trends with regard to its interaction with the host will be an essential tool in deciphering the molecular genetic aspects of infectious diseases.
CHAPTER 8

General Discussion
8. General Discussion

It is quite impressive how a pathogen like *C. trachomatis*, with a very small genome and unavoidably restricted to an intracellular development, exists as multiple genotypes that exhibit differential tissue tropism, virulence and ecological success while possessing a colossal genomic similarity among them. Despite of the efforts and fruitful research performed in the later years, researchers are still far away from deciphering the molecular factors responsible for these phenotypic differences, and hence some puzzles remain to be disclosed such as: *i*) why are LGV genotypes more invasive than both ocular and genital genotypes?; *ii*) why are LGV genotypes unable to infect the eye?; *iii*) why are genital genotypes unable to induce trachoma despite their ability to colonize the eye?; *iv*) which virulence factors determine *C. trachomatis* pathogenesis?; and *v*) what’s the secret underlying the ecological success of genotypes E and F?. The general goal of the present thesis was to search for genomic and transcriptomic features that differentiate *C. trachomatis* genotypes, in order to contribute for the understanding of their pathobiological disparities. For this purpose, we evaluated several polymorphic loci potentially important in the *C. trachomatis* biology, giving especial attention to the nine *pmp* family and the chlamydial key antigen (MOMP).

We started with transcriptomics to profile gene expression throughout chlamydial development using RT-qPCR, which is the most sensitive, specific, and reproducible method for rapidly quantifying transcript levels over time (even of low copy number) with minimal intra-assay variation [146,361,362]. Considering the lack of normalization strategies for validating gene expression in *C. trachomatis* as well as its pitfalls, we intended to assess, for the first time, the suitability of putative endogenous controls for chlamydial real-time expression assays (CHAPTER 2). From a set of 10 chromosome-spread housekeeping genes, 16SrRNA was clearly the most homogeneously and highly expressed gene throughout *C. trachomatis* normal development, even for strains with distinct cell-appetence (eyes, genitalia and lymph nodes) and *in vitro* growth rate (Figs. 2.2 & 2.3). Curiously, due to its high expression levels in non/semiquantitative methods, this ribosomal gene has been the standard control used in real-time PCR studies (although without any validation of its suitability [146,152]) for most bacteria, including *Chlamydia* [96,127,172-174,363]. The observed expression of 16SrRNA gene is in concordance with the permanent need of ribosomes during the well-known temporal synthesis of early, mid and late proteins required for the specific developmental stages of the complex chlamydial life-cycle [38,96,179], which points 16SrRNA gene as a good normalizing control in *C. trachomatis*. Supporting this,
16SrRNA gene also proved to be the most suitable internal control among other housekeeping genes in RT-qPCR assays for other prokaryotes [149,154].

However, the 16SrRNA gene expression was clearly down-regulated under stress conditions induced by antibiotic treatment (Fig. 2.4B). Similar lack of stability was also reported for *S. epidermidis*, where the 16SrRNA gene expression varied after heat shock or glucose challenge [149]. We hypothesize that these variable expressions likely reflect an adaptation of the bacterial metabolic activity (mediated by the ribosome content) to multiple environmental changes, which may be a *conditio sine qua non* for bacterial virulence. It is known that, under adverse conditions, *Chlamydiae* are able to disrupt their development and enter in a quite quiescent persistence phase, likely as a mechanism to enhance host immune evasion and transmission [83,364]. Both the aberrant RBs and the lack of motion inside inclusions observed during antibiotic treatment clearly evidence a scenario of persistent *in vitro*, where bacterial cell division is inhibited while genome replication continues [82,83]. In this situation, the normalization of expression data against gDNA (which is an indirect estimate of the number of bacteria) may constitute an accurate good alternative, as it was less prone to vary under stress conditions than endogenous controls (Fig. 2.1). We also evaluated the suitability, in *C. trachomatis*, of two widely used statistical algorithms for validating endogenous controls in expression assays [157,160,161,177], but the contrasting results obtained suggest prudence when using these applications in biological systems for which accurate transcriptome information is lacking.

After validation of normalization strategies for real-time expression assays, we proceeded to the second objective of this thesis, which was to evaluate the relative expression of the entire *pmp* family throughout normal development of both prototype and current circulating E and L2 strains, which represent the two *C. trachomatis* biovars, and the former is the most succeeded genotype worldwide (CHAPTER 3). So far, expression of all nine *pmps* had only been evaluated for a single prototype strain through RT-PCR [187] and microarray [96] analyses. However, contrarily to our [127] and previous [123,365] RT-qPCR studies, where *pmp* expression was visible as early as 2 h (pi), the above methodologies revealed low sensibility and weak quantitative accuracy, as they only detected *pmp* transcripts since 8 h (pi), which highlights the advantage of using RT-qPCR over semi-quantitative or hybridization-based array approaches.

Overall, heterogeneous expression levels were observed not only for the same *pmp* among strains, but also between *pmps* for each strain, even within operons (Figs. 3.1 & 3.2). Such divergences were also seen in recent RT-qPCR studies for *C. abortus* strain S26/3 [366] and for different strains of *C. felis* [367], which suggest a variable *pmp* expression within
Chlamydiae. Therefore, there may be differential biological relevance for each Pmp, where strains exhibit dissimilar needs for the same Pmp in the various developmental stages, in particular during the RB division and RB to EB conversion. Considering the putative dual role of Pmps in bacterial adherence and antigenicity (see below), we speculate that C. trachomatis strain-specific needs may be determined either by the different infected cell-types (epithelial versus lymph cells) and organs (eyes versus genitalia), or by their associated immune responses rather than other environmental signals, such as bacterial density and energy/nutrient availability. Similar modulation was already reported for B. burgdorferi, where both the in vitro and in vivo expressions of several genes encoding immunogenic proteins involved in tissue invasion, immune evasion and protective immunity (such as the chlamydial Pmps) were found to be temporal and tissue-related [368-370], and likely mediated by host immunity [370,371].

Curiously, these expression dissimilarities do not seem to result from the nucleotide polymorphism among genotypes or even among same-genotype strains, as demonstrated for the pmpFE operon, where the highest pmp expression discrepancies were seen (particularly for pmpF, the most polymorphic pmp [121]). Thus, we hypothesize that there may be complex regulatory mechanisms acting at both transcriptional and translational levels for pmps, like those reported for multiple bacterial genes [217,219,372-377]. Indeed, several lines of evidence point in that direction: i) the differential occurrence of deletion events among genotypes E and L2 for pmpB, pmpD, pmpE, pmpF and pmpG [121], which seem to correlate with the tissue-specific evolution of C. trachomatis genotypes (see CHAPTER 7); ii) the in silico identification of several putative hairpin structures and RNase E cleavage sites, some of them in close proximity to putative RBSs and/or not conserved between different-genotype strains throughout the pmpFE operon (Fig. 3.4) and its regulatory region (Fig. 3.3); and iii) the differential presence of remnant IS-like elements among genotypes for both pmpB [121] and pmpC [113], as insertion/excision of transposable genetic elements was found to modulate gene expression of some virulence capsular genes of Neisseria meningitidis [378] and Citrobacter freundii [379].

So far, detection of the produced Pmps, which may get insights of possible protein post-translational modifications and surface variation, has been a difficult issue to address, leading to inconsistent results among Pmps and/or among genotypes [119,120,188,189]. Some of these difficulties are related to the obligate intracellular nature of C. trachomatis, the biological uniqueness of each genotype, and the large size of Pmps (96 to 187 kDa), but also to inherent drawbacks of proteomics (requirement of protein solubility, difficulties in protein purification and in detection of low expressed proteins, etc.), which result in a weak
quantitative accuracy and sensitivity [380]. However, it has been suggested that Pmps may be modified upon translocation, as demonstrated for both the *C. trachomatis* [123,125] and *C. pneumoniae* [61] PmpD, where various temporal-dependent proteolytic processing phenomena to generate membrane-associated and soluble forms were observed. Moreover, it was recently shown (at least for one *C. trachomatis* E strain) that the *in vitro* surface expression of each Pmp is subjected to an independent high-frequency on/off switching at the inclusion level [381], phenotypically similar to that seen in other gene families of several pathogens [382]. In fact, it is known that the opacity or pilin proteins of *Neisseria* sp. [383,384], the variable surface proteins of both *Borrelia* sp. [385] and *Trypanosoma brucei* [386], and the LPS of *Haemophilus influenza* [387], all use phase/antigenic variation mechanisms for altering its surface coat in order to promote multiple antigenic and/or adherence phenotypes. We hypothesize that Pmps may have an analogous role in *Chlamydiae*, being important fine tune determinants for the heterogeneous antigenicity, virulence and tissue tropism among strains [388].

In agreement with this, cumulative immunoreactivity studies over PmpC [127], PmpD and PmpF (Table 3.2), and more recently over all nine Pmps [389], have been showing that different *C. trachomatis*-infected patients display variable and distinct antibody specificity profiles against individual or multiple Pmps, at both the quantitative and qualitative levels. Curiously, the most conserved Pmps (PmpB, PmpC, PmpD and PmpI) seem to be more frequently and strongly recognized than the most polymorphic ones (PmpF, PmpH and PmpE) [121,166,389], which suggest differential involvements in host humoral immune response. For instance, while PmpD seems to elicit conformation-dependent species-common neutralizing Abs [62,125] which makes it an attractive vaccine candidate, the high polymorphism of PmpF may provide antigenic diversity to evade humoral immunity (as supported by *pmpF* highest and disparate expression). More, the existence of numerous SNPs located within predicted HLA class I and II T-cell epitopes of the N-terminal domain of PmpF [5], also implies it in the cellular immune evasion, thus facilitating transmission or the development of chronic or repeated infections. Analogously, the relatively conserved PmpI was shown to elicit a CD8+ T-cell response in *C. trachomatis* D- and L2-infected mice, with cytotoxic activity only seen against the invasive L2 strain [390]. However, it would be interesting to perform a detailed B- and T-cell epitope characterization for all Pmps of all genotypes, in order to identify the most relevant targets of each arm of the host immunity, which could be useful for the understanding of the infection success of each genotype as well as for the design of a vaccine against *C. trachomatis*. 

136
In addition, Pmps may also provide an important functional diversity for *C. trachomatis* pathogenesis. In fact, we previously suggested that the differential occurrence and close proximity of multiple repetitive GGAI motifs (also present in homologous autotransporter adhesins [391]) and Cys residues in the N-terminus [121] likely promotes the formation of different ‘receptors’ that may be involved in host cell interaction. For some Pmps, especially for PmpE, PmpF and PmpH, such adhesion disparities may also be widened whether these host-interacting regions fall within specific protein domains that are highly polymorphic and are involved in genetic segregation of strains by their cell-appetence [121]. Recently, PmpD was implied as an important virulence factor with pleiotropic actions within and distal to infected-cells due to an infection-dependent proteolytic segregation of two putative eukaryote-interacting functional motifs [125], which could result in the induction of apoptosis of neighboring uninfected cells [392] and/or in the suppression of the memory CD8+ T-cell response [393]. In fact, like other autotransporter adhesins (including the *E. coli* Ag43 [394] and the *Bordetella pertussis* Pertactin [395], BrkA [396] and Tef [397]), PmpD contains a predicted RGD (Arginine-Glycine-Aspartate) motif that is recognized by many integrins, a family of major cell adhesion receptors that mediate a variety of biological processes by activating complex signaling pathways [398,399]. On the other hand, the PmpD nuclear localization signal (NLS) is involved in the regulation of host transcription, similar to the *Helicobacter pylori* VacA [400]. Nevertheless, it would be interesting to evaluate whether these or other eukaryote-interacting functional motifs exist in the remaining Pmps in order to scrutinize the putative virulence potential of this family.

To allow worldwide cross-comparisons of methods and results, chlamydial research has been traditionally addressed to prototype strains, which have been laboratory adapted since their isolation ~60 years ago. However, the heterogeneous expression observed between each prototype and the respective clinical strains (Figs. 3.1 & 3.2), clearly showed that the use of prototype strains as unique models for transcriptomic studies should be reevaluated. Such divergences were also seen for PmpD in a recent RT-qPCR study over both prototype and current circulating D and L2 strains [365]. Concordant with this, it has been shown that this lack of correlation extends at the genomic level, as these study-models also do not seem to represent the same genomic make-up of recent circulating isolates [114,401]. More, based on the differential immunoreactivity among Pmps and among strains, it is expected that these dissimilarities are also reflected at phenotypic level. As recent isolates are a result of the successive exposure to the immune selection of present-day chlamydial infections, they might reflect the evolutionary mechanisms used by *C. trachomatis* to infect a specific niche, evade host surveillance, and ultimately survive intracellularly. Thus, we believe that the examination
of clinical isolates along with the corresponding prototype strains is crucial to advance our understanding in the chlamydial biology and disease pathogenesis.

This led us to pursue other objectives of this thesis, which were to evaluate, first in a Portugal-restricted study (CHAPTERs 4 & 5) and then in a worldwide basis (CHAPTER 6), the impact of the host immune pressure on *C. trachomatis* strains by analyzing the evolutionary dynamics of the chlamydial dominant antigen (MOMP). From 2001 to 2007, a total of 795 isolates from ano-urogenital infections were successfully typed at the Lisbon NIH. Heterogeneous strain distributions per *ompA* genotype were observed (Fig. 5.1), with ~57% of the strains belonging to genotypes E and F, while Ba, L1 and L3 or A, B, C and I strains were either absent or rare in ano-urogenital infections, respectively. Given the consistency with previous epidemiological reports for other countries dispersed around the globe [251,254-256], we may speculate little fluctuation in genotype distribution among ano-urogenital infections worldwide. An interesting case was found for the L2 strains, which had never been confirmed in Portugal until 2007. However, both the unique strain-type and lack of relation to any sexual network of MSM (Table 4.1) deviate these unusual Portuguese LGV infections from the recent ongoing LGV outbreak in Europe and United States [21-23]. As in previous reports [227,229,231], our results also showed that anorectal infections in MSM as well as their severity are not exclusive of LGV strains and may involve other urogenital genotypes. So far, cumulative molecular, phylogenetic and immunological studies assessing the relationship of *C. trachomatis* serovars or genotypes to disease severity or other covariants (such as age, gender, race/ethnicity, sexual behavior, prevalence of infection, etc.) were inconclusive, nonreproducible and/or contradictory (reviewed in [402,403]). Interestingly, none of these features seem to be associated with the intracellular multiplication rate of strains *in vitro* as well [167], which suggests that the disparate virulence and pathogenic potential among genotypes *in vivo* is likely defined by other chlamydial factors (beyond the *ompA* paradigm) or even host genetic or immune factors.

Considering that MOMP is the major chlamydial surface component [133], and a well-characterized porin [242], adhesin [55] and highly variable immunodominant antigen that elicits both arms of the host immune response [134,136,137,244], we cogitate an unique evolutionary pathway for MOMP that likely reflects the demanding constraints of its complex quadruple role. Our evaluation of the mutational trends of all MOMP sequences reported to date in the literature, which constitutes an evolutionary short-scale (~60-years) model composed by 5026 strains isolated in 33 worldwide-dispersed geographic regions (including Portugal), supports this. In fact, we showed that: i) amino acid composition of MOMP (Fig. 6.5) is evolving through intrinsic mutational and selective pressures, while it diverges
from the ongoing substitution pattern of the remaining chromosome [309]; ii) the evolutionary adaptive landscape of MOMP appears to be also influenced by *ompA*-specific codon usage; iii) there are sites of the protein more prone to change (Fig. 6.3), with more than half involving mutations at evolutionary fixation, specially within host-interacting domains; and iv) the gain/loss of certain amino acids thought to be critical for the nMOMP trimeric structure [241,242] is not random, which may provide the pathogen with some environmental-adaptive plasticity. As recently shown in a nonhuman primate trachoma model [300], these subtle mutations, even outside known antigenic regions, may yield profound distinct antigenic MOMP profiles (with consequences in the elicited immune response) by affecting both the conformation and immunoaccessibility of epitopes. However, they must equally maintain the global porin and adhesin functionality of MOMP by allowing for a conserved structure that is important for MOMP-eukaryotic receptor(s) interaction.

Overall, our results are likely suggestive of two distinct evolutionary antigenic scenarios in MOMP. Based on some *in vitro* neutralization studies [134,136,281] and on the acting strong positive selection (Fig. 6.4), we hypothesize that the rampant variation found at fixation in B-cell epitopes (Fig. 6.2) may represent natural neutralization escape variants that successfully continue to co-evolve with the human host. This might consequently impact the putative *C. trachomatis* decoy-like immune evasion strategy active *in vivo*, whereby natural immunodominant type-specific surface antigens (such as MOMP) block the neutralizing ability of species-common Abs to minor antigens (as demonstrated for PmpD), hence avoiding a broadly heterotypic host protection to reinfection [62]. Although this decoy function could explain the observed *in vitro* genotype-specific immunity to reinfection, more convincing clinical evidence is needed to elucidate if it mimics the natural human infection [404]. Nevertheless, it has been suggested that the acquired protective immunity likely affects *in vivo* replication instead of averting chlamydial entry, as chlamydial infectious load seem to decrease during consecutive reinfection episodes [167]. Such incomplete protection may outcome from a potent chlamydial T-cell suppressive activity during secondary infection, as a result of the development of memory CD8+ T-cells functionally defective [393]. In fact, consequent pathological effect on antibody-regulated cell-mediated immunity (ADCC or protective Th1 memory response) is also expected from chlamydial escape to host humoral response [310,405]. Corroborating this, similar humoral response avoidance through SNP selection in determinants of neutralizing envelope proteins was already found for various viruses, particularly during conditions of CTL deficiency, leading to virus persistence *in vivo* [311,406,407].
In contrast, the high conservation of T-cell epitopes in MOMP CDs (Fig. 6.2), in particular of species-specific CTL and Th epitope clusters with diverse HLA specificities [137-140], might indicate an additional evolutionary pathway that likely leads to host immune subversion, as observed for other intracellular pathogens (such as *Salmonella Typhi* [312] and *M. tuberculosis* [313]). Indeed, our data showed that these MOMP epitopes are under strong selective pressure to be maintained (Fig. 6.4), suggesting that *C. trachomatis* might benefit more from human T-cell recognition than from its avoidance. Based on several *in vitro* and *in vivo* evidences [83,364], we hypothesize that the elicited cellular response (predominantly by IFN-γ secretion [408]) may allow the pathogen to establish a persistent infection, hence maximizing the likelihood of its infectivity and cell-to-cell transmission. Indeed, such strategy was already shown to function as a potential virulence mechanism of genital-infecting strains, which exclusively retain the ability to synthesize its own Trp (an essential amino acid for both eukaryotic cells and *Chlamydiae*) from endogenous sources of the local genital flora, therefore avoiding host IFN-γ-mediated killing (mainly through Trp depletion) [108]. Also, the cytotoxic activity exclusively exhibited by epithelial-genital strains is thought to be critical for circumventing IFN-γ-mediated immune responses at attachment sites and promoting infection in this highly competing flora [107,111]. Moreover, the IFN-γ-secreting Th1 response elicited by MOMP was found to be insufficient to resolve chlamydial infection and to confer protective immunity [297,409-415], which clearly supports that this pathogen may have developed ways of avoiding IFN-γ-mediated effectors. Nevertheless, we cannot exclude the possibility that functional constraints (that are independent of T-cell recognition) may also contribute to this high conservation of T-cell epitopes in MOMP CDs, similar to the already reported for the Nef, Gag, Env and Tat proteins of HIV and simian immunodeficiency virus (SIV) [416-419].

Besides SNPs, cumulative evidences have been showing that recombination is also involved in determining the repertoire of MOMP variability, which makes its evolutionary pathway even more complex. Indeed, although this was not the focus of our worldwide study, multiple *ompA* variants showing mosaic structures derived from two or three parental strains have been repeatedly detected [112,115,116,254,260,420], as a result of superinfection of the host with various *C. trachomatis* genotypes. Despite the chlamydial obligate intracellular sequestering in a cytoplasmic vacuole, recombination is biologically plausible, as DNA repair and recombination systems are well represented in the *C. trachomatis* genome [2,5-7,37,401], and fusion of chlamydial inclusions (favoring genetic exchange between strains) in superinfected host cells have been documented [421-423]. As some of these mosaic structures were found to involve the exchange of antigenic or attachment-involved regions among
genotypes with different cellular appetite [116,254], it is tempting to speculate that such intragenic recombination in \textit{ompA} may be a mechanism used by the pathogen to rapidly engender the required genetic diversity in order to improve its fitness. In fact, recombination may: \emph{i}) disrupt or create new combinations of epitopes in ways that might affect immune recognition of this key antigen, ultimately leading to persistence in the host; and \emph{ii}) provide MOMP with tissue-specific adhesion motifs that confer strains the capability of infecting other cell-types. Nevertheless, the differential developmental expression of \textit{ompA} observed not only among distinct prototype strains (Fig. 3.1; and [127]) but also among same-genotype strains with 100% conserved sequences (where a surprisingly up to five-fold expression disparity was found) (Fig. 3.2), suggests different strain-specific biological needs for MOMP that may also underline phenotypic distinctions, likely as a result of complex regulatory mechanisms at both transcriptional and translational levels [200].

Considering the findings obtained for the surface-exposed Pmps and MOMP, highly suggestive of their relevance for \textit{C. trachomatis} biology, we next moved forward to a broader objective of this thesis, which consisted on studying the evolution of the chlamydial genotypes through the analysis of a panel of heterogeneous polymorphic \textit{loci} (IGRs, HKs, HPs, and CEPs) scattered throughout the chromosome (CHAPTER 7). Interestingly, we found that the evolutionary history of the 15 main \textit{C. trachomatis} genotypes reflects a global chromosomal adaptation to each infected tissue (eyes, epithelial-genitalia, and lymph nodes) (Fig. 7.4). Based on our results, we presume a scenario where, after an ancestor strain common to all genotypes had developed the skill to colonize the different tissues (see CHAPTER 7 for details), the subsequent stages of \textit{C. trachomatis} evolution involved primarily the accumulation of mutations (and small indels) throughout the chromosome, as an outcome of the environmental and immune pressures on each niche (Fig. 8.1). Such targeted fixation resulted in distinct profiles of genetic variability, leading to the radiation of the contemporary genotypes according to their cell-appetence. As the studied \textit{loci} were found to be either highly conserved or polymorphic among strains infecting the same niche (Fig. 7.4), we speculate that the former likely represents the final stages of the tissue-specific adaptive evolution, while the later may either be involved in strain pathogenicity [37] or simply represent the continuous search for the best fitness in order to keep survival and transmission in that niche, on the behalf of the coevolutionary Red Queen’s theory [335,336].
Figure 8.1 – Schematic diagram presenting the hypothetical evolutionary scenario of *C. trachomatis*. The evolutionary and biological processes are represented in parallel. Phase I illustrates the pathogen lifestyle change from extracellular to obligate intracellular [2,340]. Phase II represents the skill to infect different cell-types/organs [5,107-109,424]. Phase III illustrates strain radiation due to niche-specific environmental and immune pressure (supported by our study – CHAPTER 7). Ocular, epithelial-genital and LGV strains are represented in yellow, pink and blue boxes, respectively.

Curiously, despite hypothetical proteins represent less than 12% of all studied *loci*, their highest polymorphism among genotypes (almost twice-fold higher than cell envelope proteins) (Table 7.1) suggests an important involvement in tissue-specific strain segregation. Although there is no assigned role for these ORFs, we speculate a primary function in adherence to the host cells, inclusion development or antigenic polymorphism. In support of this, three of them (CT049, CT144, and CT622) were found to contain, like some Pmps [121], distinct protein domains in which strains infecting different tissues vary up to 65% among them (Fig. 7.3). Considering the known acidic character of CT049 [425], we hypothesize that these domains may be responsible for different protein topologies, providing tissue-specific host-interaction motifs likely important for either the attachment to the host cells or the modulation of intracellular trafficking. Curiously, CT049 was recently found to be paralogously related to the passenger domain of PmpC [425], although it does not seem to follow the typical autotransporter secretion pathway [425,426]. Nevertheless, like some Pmps [113], CT049 exhibits differential indel events among genotypes (Table S7.3), which were shown to correlate with tissue-specific strain radiation (Fig. 7.5A), and to exist as both soluble and differentially inclusion membrane-associated forms that may be important for...
chlamydial pathogenesis, particularly in regulating inclusion expansion [425]. This polymorphic protein, recently identified as one of the 20 loci subject to strong selective pressure in murine genital model [427], was also found to be accessible to specific Abs *in vitro* [425], which may imply a role in evading cytoplasmic innate immune surveillance mechanisms or antigen presentation pathways. Another interesting aspect related with the hypothetical proteins is that the genetic variability of CT622 was shown to be responsible for profound virulence differences (that may also contribute to disease severity) among same niche-infecting strains in non-human primates [37]. More, polymorphisms in CT144 were recently found to be statistically associated with rectal tropism in genotype G [401].

Collectively, this thesis also sheds some light on the secret behind the worldwide success of *C. trachomatis* genotypes E and F. Indeed, our high-scale genomic concatenation results revealed that E and F possess a similar chromosomal genetic make-up that distinguishes them from the remaining genotypes (Fig. 7.4). Based on the loci linked to this independent co-segregation (Fig. 7.4), we speculate about the existence of specific adhesins or still-undefined virulence attributes (for example, factors related to bacterial growth) that may confer E and F strains some functional and/or structural advantage in terms of infection and/or transmission. Supporting the former, we have reported [121] that some *pmps* display, nearby to conserved GGAI motifs, exclusive Cys residues for E and F, which may yield unique E/F conformational motifs for interaction with the host. Also, both the two-codon insertion solely present in *pmpB* of E and F strains [121] and the E/F-specific intra-loci domains found in CT622 (Fig. 7.3) may promote the formation of exclusive host-interacting regions not yet identified. Additionally, it has been reported [60,428-431] that, in contrast to other genotypes, the adhesion of OmcB from genotype E is completely independent of heparin-like glycosaminoglycan structures on the host epithelial cells, which have been implicated in cell attachment and invasion by many pathogens, including chlamydial species [56]. Such independence was recently shown to be specifically determined by a single leucine residue in the vicinity of the OmcB conserved heparin-binding motif, probably by generating a distinct conformation in this region [60]. As our comparative genomics over the 15 main chlamydial genotypes (CHAPTER 7) revealed that this leucine is also exhibited by genotype F, we hypothesize that OmcB may influence host cell invasion by E and F strains.

Moreover, in contrast to what is traditionally assumed, the highest success of E and F strains is not associated with a higher *in vivo* chlamydial infectious load [167]. However, the greatest involvement of housekeeping genes in the co-segregation of E and F than in the chlamydial tissue-specific evolutionary adaptation (when compared with the other functional loci categories) (Fig. 7.4), likely implies a distinct bacterial metabolism for these genotypes.
For instance, the significant E/F-specific polymorphism exhibited by both the rs2/ompA and yfh0_1/parB IGRs may influence the well-known temporal protein synthesis and the bacterial replication, respectively. In fact, the later encompasses the gene regulatory region of the partitioning protein ParB, which is essential for efficient chromosome and plasmid segregation during replication of many bacterial species [432], likely affecting the time of replication. On the other hand, the rs2/ompA IGR contains a tRNA gene as well as the regulatory region of the putative rs2-tsf operon that encodes two essential components of the translational apparatus [433,434], the ribosomal protein S2 and the elongation factor Ts. Interestingly, the genetic variability exhibited by these proteins also distinguishes E and F genotypes from the remainder, which may contribute to advantageously phenotypic differences. Additionally, the distinct arginine kinase KarG displayed by E and F also suggests differences in maintaining the demand chlamydial energy metabolism [435].

Intriguingly, our results indicated that conservation of the major chlamydial antigen (MOMP) may also favor the success of E and F strains. According to Darwin’s natural selection theory [436], their remarkable low mutation rate [22.3-fold lower than that of other genotypes (CHAPTER 6)] likely implies more fitted MOMP antigenic profiles to handle host immunity, which would be less prone to evolutionary change. In opposition to the remaining genotypes, where MOMP variability seems to represent a strategy to escape recognition by the host immune surveillance, we speculate that E and F advantage in preserving these MOMP antigenic profiles may relies on a lesser overall immunogenicity that enables them to go unnoticed for longer time, favoring their infectivity and dissemination.

Preliminary data from an ongoing study (in preparation) to evaluate the impact of recombination on C. trachomatis population, using a sampling of multiple recent isolates that reflects the worldwide distribution of each genotype, seem to evidence a likely clonal trait for E and F strains, as seen in Salmonella enterica, E. coli, S. aureus, and Neisseria meningitidis [437,438]. In fact, ~88% and ~63% of all analyzed E and F strains were 100% genetically similar to the corresponding prototype E/Bour and F/IC-Cal3 (Fig. 8.2B), respectively, which suggest the existence of a predominant evolutionary adaptive clone for both genotypes. Moreover, the likelihood of E and F strains to undergo genetic recombination is about 12-fold lower than that of the other genotypes (P < 10^-2) (Fig. 8.2A), supporting the clonal expansion of E and F. Considering both the randomness of every recombination process and the high prevalence of E and F strains (and, consequently, their higher frequency in mixed infections), it is implausible that these strains are less biologically susceptible to genomic rearrangements. Thus, there may be a strong in vivo selection to maintain these particular genomic compositions, whose advantageous adaptive fitness to deal with the host, likely
promote the clonal-expansion-based E/F genotype success in detriment of other E and F less favorable clones and, ultimately, of other chlamydial genotypes. Similar to the predictable ‘epidemic’ structure population of *N. menigitidis* and *S. aureus* [437-439], such superimposed favourable E and F clones may persist for decades until recombination has time to homogenize the population genetic background. However, for some well studied pathogens (such as *S. aureus* [440]), not always emergent clones are more abundant, showing that, in contrast for *C. trachomatis* E and F genotypes, this trait is not necessarily the key to success.

![Figure 8.2](image)

Figure 8.2 - Putative clonal evolution of the two most prevalent *C. trachomatis* genotypes (E and F). (A) Impact of recombination on current circulating *C. trachomatis* isolates (n=56). To detect mosaic genomic structures, 11 loci (*yraL, yraLCT049 IGR, CT049, pmpC, rs2, rs2/ompA IGR, ompA, ompA/pbpB IGR, pbpB and pmpF*) involving two statistically-confirmed recombination hotspots and representing four well-separated regions of the *C. trachomatis* chromosome [114] were used. For each locus, the sequence of each isolate was compared with that of the respective prototype strain. Mosaic structures of all recombinants are displayed in detail in Figure S2. The strain sampling used reflects the worldwide distribution of each of the 12 chlamydial genotypes evaluated, which are colored-coded. (B) Phylogenetic tree showing the putative clonal evolution of the 26 non-recombinant isolates. Dominant E and F clones are colored-boxed. The concatenated tree (NJ method, K2P model) is based on about 13,000 bp/taxa. Bootstrap values (1,000 replicates) are shown next to the branch nodes. Due to its recombinant nature [112,115,116,254,262,420], *ompA* was excluded from the phylogenetic analysis.
8.1. Concluding Remarks

Overall, the findings presented in this Ph.D. thesis suggest that the biological diversity of *C. trachomatis* genotypes may be determined by complex molecular mechanisms that probably reflect a global chromosomal dynamics, where Pmps, MOMP and some hypothetical proteins seem to play an important role. Indeed, through the genomic approach, we showed that the differential tissue tropism and pathogenic potential among genotypes likely reside on the genetic polymorphism displayed by several *loci*, which might be crucial for promoting multiple antigenic and adherence phenotypes important for host immune evasion and host interaction, respectively. On the other hand, through the transcriptomic approach, the heterogeneous expressions observed intra- and inter-strains (even from the same genotype), suggest that the molecular basis of *C. trachomatis* biological diversity may be defined beyond genotype level, and might be strain-specific, which likely hinders the development of vaccinal strategies against *C. trachomatis*. Understanding the impact of the resultant functional or antigenic phenotypic diversity on pathogen’s interaction with the host will be crucial for comprehending the differential tissue tropism, pathogenic potential and virulence observed among *C. trachomatis* strains, as well as for deciphering the basis of their ecological success. Ultimately, the unique genomic make-up of the two most ecological succeed genotypes worldwide suggests that they may have already found their own advantageous evolutionary pathway to play the ‘arms race game’ with the host, where a noticeable lack of chromosomal mosaicism with a strikingly low mutational rate of the dominant antigen (MOMP) (when compared with strains from the remainder genotypes) may be essential determinant factors for a higher *C. trachomatis* ecological success. Time will tell if this linkage disequilibrium will last, or, in agreement with other pathogens, if the recombination background, known to exist in *C. trachomatis* [114], will dilute this evolutionary effect leading to a subsequent emergence of new successful “clones” [437,441].
CHAPTER 9

Future Perspectives
9. Future Perspectives

The findings obtained during the course of this Ph.D. thesis opened avenues for future studies that will soon constitute main research lines of our group. In particular:

i) To scrutinize the biological role of a specific group of hypothetical proteins. This novel study constitutes the Ph.D. proposal of a member of our team.

ii) To identify putative host factors that participate in chlamydial attachment and/or entry into host cells. This subject constitutes my Pos-Doc proposal.

The rational for the former is the notable polymorphism displayed by the hypothetical protein genes among *C. trachomatis* genotypes as well as their apparent involvement in tissue tropism, adhesion, and ecological success (see CHAPTERS 7 & 8). Thus, it would be interesting to:

- Perform the genomic and transcriptomic approach for the set of hypothetical protein genes described throughout the thesis that seem important in chlamydial biology.

- Assess the HP secretion through T3SS, using a surrogate host (such as Yersinia spp., Shigella spp., or Salmonella spp. [43]). This step would allow us to identify potential T3S effectors, which are known to function as virulence factors that modulate specific host cellular functions during chlamydial infection [43,46,48]. We will have the collaboration of Luis Jaime Mota from ITQB, who is proficient in these methodologies.

- Evaluate the HP immunoreactivity against a panel of sera from patients infected with different *C. trachomatis* genotypes, by using a fusion protein array approach for profiling the specific human Ab responses elicited. This would allow us to assess whether these proteins are host immune targets, and if so, whether they yield differential immunoreactivity profiles among strains with different or similar tissue tropism.

- Evaluate the HP putative role as adhesins by performing *in vitro* neutralization assays of chlamydial infectivity with Abs specific to these proteins, as already performed for MOMP [55], PmpD [62] and OmcB [59].
The rational for the second future research line is the identification of the host factors involved in *C. trachomatis* attachment and entry, whose inhibition may contribute for the development of prophylactic and therapeutic strategies through the inhibition of chlamydial infectivity. To achieve this, microarrays of the human cell line (HeLa229) (used in our laboratory for chlamydial cell culture) will be performed to evaluate the host up- and down-regulated genes during the *C. trachomatis* entry (first 2 h pi). Several chlamydial strains from genotypes exhibiting differential tissue tropism and ecological success, will be used in these assays. Also, we aim to further identify host factors that are specifically regulated when the human cell line is infected with chlamydial strains displaying similar cell-appetence. Ultimately, the genetic knockdown of these factors (given the existence of commercial kits for this particular human cell line, where genes may be knock-out under request) will constitute a confirmatory step forward in the identification of these host ‘receptors’.
Supplemental Material
Table S2.1. Oligonucleotide primers used in qPCR assays

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene function</th>
<th>Primers</th>
<th>Primer sequence (5’ to 3’)</th>
<th>Primer location</th>
<th>Amplicon (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT048</td>
<td>Sam-dependent methyltransferase (yraL)</td>
<td>CT048-C</td>
<td>CCAAAATCCTAAGGAAGATCTCGTTA</td>
<td>432-458^b</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT048-D</td>
<td>GGTATCCGAGGCTATCATTACAG</td>
<td>521-495^b</td>
<td>145</td>
</tr>
<tr>
<td>CT059</td>
<td>Ferredoxin (fer)</td>
<td>CT059-A</td>
<td>GAAATGCTCGGTCGTTT</td>
<td>82-101^b</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT059-B</td>
<td>GCCTTCCAAGAATTCTACAG</td>
<td>154-135^b</td>
<td>145</td>
</tr>
<tr>
<td>CT147</td>
<td>Hypothetical ‘EEA1 homologue’</td>
<td>CT147-A</td>
<td>CAGGTAAAGTCTGTTAAGGACCTTTT</td>
<td>673-699^p</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT147-B</td>
<td>TGAATGCTCGGTTACCGTTAAGGCTTAACAC</td>
<td>723-700^b</td>
<td>51</td>
</tr>
<tr>
<td>CT175</td>
<td>Oligopeptide binding protein permease (oppA_2)</td>
<td>CT175-E</td>
<td>GAGCGCGGAAAGGATATTTAATATGTTAAT</td>
<td>640-665^b</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT175-F</td>
<td>CTTGGAATACCTACAGAGGTTT</td>
<td>690-668^b</td>
<td>51</td>
</tr>
<tr>
<td>CT298</td>
<td>Sms protein ATPase (radA)</td>
<td>CT298-E</td>
<td>GGAATGCTCGGTTAAGGACCTTTT</td>
<td>1189-1211^b</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT298-F</td>
<td>AATGGTTTCAATTCTATCAGGTGTTT</td>
<td>1239-1216^b</td>
<td>51</td>
</tr>
<tr>
<td>CT461</td>
<td>Phosphohydrolase (yaeI)</td>
<td>CT461-A</td>
<td>GGTGTTCTGATTGACATTT</td>
<td>151-172^b</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT461-B</td>
<td>CAGGGCTCGGATGTTT</td>
<td>202-183^b</td>
<td>51</td>
</tr>
<tr>
<td>CT746</td>
<td>Coproporphyrinogen III oxidase (hemN_2)</td>
<td>CT746-C</td>
<td>CACCATTCGCTGACTGCAAT</td>
<td>69-88^b</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT746-D</td>
<td>TACCCAGGGCTAGCATAATAT</td>
<td>119-98^b</td>
<td>51</td>
</tr>
<tr>
<td>CT818</td>
<td>Tyrosine specific transport protein (tyrP_2)</td>
<td>CT818-C</td>
<td>CTGTTGATCTGATGTTCTTTT</td>
<td>62-84^b</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT818-D</td>
<td>AAGCGGCAAGTACATAGCA</td>
<td>140-119^b</td>
<td>79</td>
</tr>
<tr>
<td>CT851</td>
<td>Methionine aminopeptidase (map)</td>
<td>CT851-C</td>
<td>GGGAGGCACGAACTGTA</td>
<td>776-794^b</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT851-D</td>
<td>TAGAGATTTCCCATGTGCAAATA</td>
<td>826-804^b</td>
<td>51</td>
</tr>
<tr>
<td>CT681</td>
<td>Major outer membrane protein (ompA)</td>
<td>16S RNA-9^c</td>
<td>GCGAGGGCTGTTTCTATTTTT</td>
<td>734-756^d</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16S RNA-10^c</td>
<td>CCCAGGTTATCTAAACCTGTTTGC</td>
<td>809-786^d</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OmpA-9^d</td>
<td>TGCCGCTTTGAGTTCTGCTT</td>
<td>33-52^d</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OmpA-10^d</td>
<td>CGCTGATCAAGAAGGTTT</td>
<td>108-86^d</td>
<td>76</td>
</tr>
</tbody>
</table>

^a ORF numbers are based on the D/UW3 strain genome annotation (GenBank: NC000117).
^b Based on the sequence of D/UW3 strain.
^c Previously described in [127].
^d Based on the sequence of L2/434 strain (GenBank: NC010287).
Table S5.1. Description of all prototype strains as well as of one representative of each type of genetic variant found among the 232 Portuguese variant clinical specimens

<table>
<thead>
<tr>
<th>Genotype/Strain</th>
<th>Year</th>
<th>Location</th>
<th>Biological Sample</th>
<th>No. of same-type variants (x, y)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>GenBank No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genotype A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Har13</td>
<td>1958</td>
<td>Egypt</td>
<td>Conjunctiva</td>
<td></td>
<td>NC007429</td>
</tr>
<tr>
<td>Har1</td>
<td>n.a.</td>
<td>n.a.</td>
<td></td>
<td></td>
<td>DQ064279</td>
</tr>
<tr>
<td>CS-805/04</td>
<td>2004</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/1</td>
<td>FJ943511</td>
</tr>
<tr>
<td><strong>Genotype B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TW5</td>
<td>1959</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td></td>
<td>AF304856</td>
</tr>
<tr>
<td>Har36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DQ064297</td>
</tr>
<tr>
<td>Tunis864</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DQ064280</td>
</tr>
<tr>
<td>IU1226</td>
<td>1987</td>
<td>n.a.</td>
<td>Urine</td>
<td></td>
<td>AF063208</td>
</tr>
<tr>
<td>CS-720/06</td>
<td>2006</td>
<td>Portugal</td>
<td>Urine</td>
<td>2/3</td>
<td>FJ943512</td>
</tr>
<tr>
<td>CS-1038/06</td>
<td>2006</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/3</td>
<td>FJ943513</td>
</tr>
<tr>
<td><strong>Genotype Ba</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apache2</td>
<td>1960</td>
<td>Arizona</td>
<td>Conjunctiva</td>
<td></td>
<td>AF063194</td>
</tr>
<tr>
<td><strong>Genotype C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TW3</td>
<td>1959</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td></td>
<td>AF352789</td>
</tr>
<tr>
<td>Har32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DQ064298</td>
</tr>
<tr>
<td>CS-110/03</td>
<td>2003</td>
<td>Portugal</td>
<td>Cervix</td>
<td>1/3</td>
<td>FJ943514</td>
</tr>
<tr>
<td>CS-428/03</td>
<td>2003</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/3</td>
<td>FJ943515</td>
</tr>
<tr>
<td>CS-434/03</td>
<td>2003</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/3</td>
<td>FJ943516</td>
</tr>
<tr>
<td><strong>Genotype D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW3</td>
<td>1965</td>
<td>Washington</td>
<td>Cervix</td>
<td></td>
<td>AE001338</td>
</tr>
<tr>
<td>181C/91</td>
<td>1991</td>
<td>Portugal</td>
<td>Urethra</td>
<td>11/12</td>
<td>FJ943517</td>
</tr>
<tr>
<td>1-953</td>
<td>2005</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/12</td>
<td>FJ943518</td>
</tr>
<tr>
<td><strong>Genotype Da</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TW448</td>
<td>1985?</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td></td>
<td>X62921</td>
</tr>
<tr>
<td>IU1554</td>
<td>1997</td>
<td>n.a.</td>
<td>Urine</td>
<td></td>
<td>AF063209</td>
</tr>
<tr>
<td>464C/93</td>
<td>1993</td>
<td>Portugal</td>
<td>Urethra</td>
<td>35/64</td>
<td>FJ943519</td>
</tr>
<tr>
<td>CS-431/04</td>
<td>2004</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/64</td>
<td>DQ116402</td>
</tr>
<tr>
<td>CS-1228/06</td>
<td>2006</td>
<td>Portugal</td>
<td>Urine</td>
<td>11/64</td>
<td>FJ943520</td>
</tr>
<tr>
<td>CS-908/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Urine</td>
<td>17/64</td>
<td>FJ943521</td>
</tr>
<tr>
<td><strong>Genotype E</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bour (1990)</td>
<td>1959</td>
<td>California</td>
<td>Conjunctiva</td>
<td></td>
<td>X52557</td>
</tr>
<tr>
<td>DK20</td>
<td>1967?</td>
<td>Denmark</td>
<td>Conjunctiva</td>
<td></td>
<td>FJ943547</td>
</tr>
<tr>
<td>7456/05</td>
<td>2005</td>
<td>Portugal</td>
<td>Urine</td>
<td>2/7</td>
<td>FJ943522</td>
</tr>
<tr>
<td>CS-165/94</td>
<td>1994</td>
<td>Portugal</td>
<td>Cervix</td>
<td>2/7</td>
<td>FJ943523</td>
</tr>
<tr>
<td>CS-433/05</td>
<td>2005</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/7</td>
<td>FJ943524</td>
</tr>
<tr>
<td>CS-1267/06</td>
<td>2006</td>
<td>Portugal</td>
<td>Cervix</td>
<td>1/7</td>
<td>FJ943525</td>
</tr>
<tr>
<td>CS-1221/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/7</td>
<td>FJ943526</td>
</tr>
<tr>
<td><strong>Genotype F</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IC-Cal3</td>
<td>1960?</td>
<td>California</td>
<td>Conjunctiva</td>
<td></td>
<td>X52080</td>
</tr>
<tr>
<td>150C/01</td>
<td>2001</td>
<td>Portugal</td>
<td>Cervix</td>
<td>1/6</td>
<td>FJ943527</td>
</tr>
<tr>
<td>725C/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/6</td>
<td>FJ943528</td>
</tr>
<tr>
<td>30C/08</td>
<td>2008</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/6</td>
<td>FJ943529</td>
</tr>
<tr>
<td>BS-332</td>
<td>1999</td>
<td>Portugal</td>
<td>Urine</td>
<td>1/6</td>
<td>FJ943530</td>
</tr>
<tr>
<td>CS-48/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Cervix</td>
<td>1/6</td>
<td>FJ943531</td>
</tr>
<tr>
<td>CS-503/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Cervix</td>
<td>1/6</td>
<td>FJ943532</td>
</tr>
<tr>
<td><strong>Genotype G</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW57</td>
<td>1971</td>
<td>Washington</td>
<td>Cervix</td>
<td></td>
<td>AF063199</td>
</tr>
<tr>
<td>392</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td></td>
<td>DQ064288</td>
</tr>
<tr>
<td>Accession</td>
<td>Year</td>
<td>Country</td>
<td>Sample Type</td>
<td>Genotype</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------</td>
<td>-------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>1156C/06</td>
<td>2006</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-212/06</td>
<td>2006</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-68/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-149</td>
<td>1996</td>
<td>Portugal</td>
<td>n.a.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-1131</td>
<td>1999</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-19/08</td>
<td>2008</td>
<td>Portugal</td>
<td>Anorectal lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-19/08</td>
<td>2008</td>
<td>Portugal</td>
<td>Anorectal lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW4</td>
<td>1965</td>
<td>Washington</td>
<td>Cervix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-190/96</td>
<td>1996</td>
<td>Portugal</td>
<td>Urethra</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BS-903</td>
<td>1999</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-139</td>
<td>1996</td>
<td>Portugal</td>
<td>n.a.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW12</td>
<td>1966</td>
<td>Washington</td>
<td>Urethra</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-24</td>
<td>1995</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype Ia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW202</td>
<td>1987</td>
<td>Indiana</td>
<td>Urethra</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IU1588</td>
<td>1997</td>
<td>n.a.</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>870</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-19/08</td>
<td>2008</td>
<td>Portugal</td>
<td>Urethra</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype J</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW36</td>
<td>1971</td>
<td>Washington</td>
<td>Cervix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-190/96</td>
<td>1996</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS-95/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype Ja</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IU17542</td>
<td>1987</td>
<td>n.a.</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IU11789</td>
<td>1997</td>
<td>n.a.</td>
<td>Urine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-91</td>
<td>1997</td>
<td>Portugal</td>
<td>Cervix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>634C/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Cervix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UW31</td>
<td>1937?</td>
<td>Washington</td>
<td>Cervix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>440</td>
<td>1968</td>
<td>California</td>
<td>Lymph Nodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype L2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>434</td>
<td>1968</td>
<td>California</td>
<td>Lymph Nodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2a/UW396</td>
<td>1985?</td>
<td>US/Europe</td>
<td>Lymph Nodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2b/44276</td>
<td>2005?</td>
<td>n.a.</td>
<td>Lymph Nodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>614C/07</td>
<td>2007</td>
<td>Portugal</td>
<td>Anorectal lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype L3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>404</td>
<td>1967</td>
<td>California</td>
<td>Lymph Nodes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*x* represents the total number of clinical specimens sharing the same *ompA* variant sequence (that is represented by one of the strains in the left column); *y* represents the total number of clinical specimens presenting any SNP in the full *ompA* sequence when compared with the prototype strain of the same genotype.
Table S5.2. PAML results (log-likelihood scores, parameter estimates and positively selected sites of \( \text{ompA} \)) for genotype G

<table>
<thead>
<tr>
<th>Model</th>
<th>( m )</th>
<th>( \mathcal{L} )</th>
<th>( \kappa )</th>
<th>Parameter estimates(^d)</th>
<th>Positive selected sites (BEB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: one-ratio</td>
<td>1</td>
<td>-1258.824626</td>
<td>2.06530</td>
<td>( \omega = 2.28790 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M1a: nearly neutral</td>
<td>2</td>
<td>-1259.185328</td>
<td>1.91654</td>
<td>( p_0 = 0.00000, \omega_0 = 1.00000, ) ( p_1 = 1.00000 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M2a: positive selection</td>
<td>4</td>
<td>-1256.430446</td>
<td>2.10785</td>
<td>( p_0 = 0.99527, \omega_0 = 1.00000, ) ( p_1 = 0.00000, p_2 = 0.00473, ) ( \omega_2 = 237.25045 )</td>
<td>125, 163, 234, 286, 335, 339</td>
</tr>
<tr>
<td>M3: discrete (K=3)</td>
<td>5</td>
<td>-1256.337305</td>
<td>2.18137</td>
<td>( p_0 = 0.00000, \omega_0 = 0.00000, ) ( p_1 = 0.99555, \omega_1 = 1.58688, ) ( p_2 = 0.00445, \omega_2 = 362.45695 )</td>
<td>---</td>
</tr>
<tr>
<td>M7: ( \beta )</td>
<td>2</td>
<td>-1259.193545</td>
<td>1.89242</td>
<td>( p = 0.04825, q = 0.00500, E[\omega] = 0.094 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M8: ( \beta &amp; \omega )</td>
<td>4</td>
<td>-1256.430159</td>
<td>2.10798</td>
<td>( p_0 = 0.99527, p = 10.16834, ) ( q = 0.00500, E[\omega_0] = 0.9995, ) ( p_1 = 0.00473, \omega_8 = 237.18850 )</td>
<td>125, 163, 234, 286, 335, 339</td>
</tr>
</tbody>
</table>

\(^a\) \( m \) is the number of free parameters for the \( \omega \) ratios.
\(^b\) \( \mathcal{L} \) is the log-likelihood value.
\(^c\) \( \kappa \) is the Ts/Tv.
\(^d\) Posterior probability was obtained by the BEB method.

Table S5.3. PAML results (log-likelihood scores, parameter estimates and positively selected sites of \( \text{ompA} \)) for genotype Ia

<table>
<thead>
<tr>
<th>Model</th>
<th>( m )</th>
<th>( \mathcal{L} )</th>
<th>( \kappa )</th>
<th>Parameter estimates(^d)</th>
<th>Positive selected sites (BEB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: one-ratio</td>
<td>1</td>
<td>-1282.402015</td>
<td>6.36147</td>
<td>( \omega = 1.10856 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M1a: nearly neutral</td>
<td>2</td>
<td>-1282.091783</td>
<td>5.90223</td>
<td>( p_0 = 0.36970, \omega_0 = 0.00000, ) ( p_1 = 0.63030 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M2a: positive selection</td>
<td>4</td>
<td>-1280.417850</td>
<td>10.40838</td>
<td>( p_0 = 0.90735, \omega_0 = 0.00000, ) ( p_1 = 0.00656, p_2 = 0.08608, ) ( \omega_2 = 13.37649 )</td>
<td>62, 176</td>
</tr>
<tr>
<td>M3: discrete (K=3)</td>
<td>5</td>
<td>-1278.922747</td>
<td>6.28627</td>
<td>( p_0 = 0.92696, \omega_0 = 0.00000, ) ( p_1 = 0.07304, \omega_1 = 18.16888, ) ( p_2 = 0.00000, \omega_2 = 18.36151 )</td>
<td>---</td>
</tr>
<tr>
<td>M7: ( \beta )</td>
<td>2</td>
<td>-1282.095640</td>
<td>5.85873</td>
<td>( p = 0.00751, q = 0.00500, E[\omega] = 0.004 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M8: ( \beta &amp; \omega )</td>
<td>4</td>
<td>-1278.922683</td>
<td>6.28642</td>
<td>( p_0 = 0.92696, p = 0.00500, ) ( q = 0.96315, E[\omega_0] = 0.0052, ) ( p_1 = 0.07304, \omega_8 = 18.17007 )</td>
<td>62, 93, 94, 176, 191, 193, 255, 311</td>
</tr>
</tbody>
</table>

\(^a\) \( m \) is the number of free parameters for the \( \omega \) ratios.
\(^b\) \( \mathcal{L} \) is the log-likelihood value.
\(^c\) \( \kappa \) is the Ts/Tv.
\(^d\) Posterior probability was obtained by the BEB method.
Table S5.4. PAML results (log-likelihood scores, parameter estimates and positively selected sites of *omp*A) for genotype Da

<table>
<thead>
<tr>
<th>Model</th>
<th>( m^a )</th>
<th>( \ell^b )</th>
<th>( \kappa^c )</th>
<th>Parameter estimates(^d)</th>
<th>Positive selected sites (BEB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0: one-ratio</td>
<td>1</td>
<td>-1208.747883</td>
<td>7.11424</td>
<td>( \omega = 0.50740 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M1a: nearly neutral</td>
<td>2</td>
<td>-1208.747883</td>
<td>7.11420</td>
<td>( p_0 = 1.00000, \omega_0 = 0.50740, ) ( p_1 = 0.00000 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M2a: positive selection</td>
<td>4</td>
<td>-1208.747883</td>
<td>7.11421</td>
<td>( p_0 = 1.00000, \omega_0 = 0.50741, ) ( p_1 = 0.00000, p_2 = 0.00000, ) ( \omega_2 = 1.00000 )</td>
<td>---</td>
</tr>
<tr>
<td>M3: discrete (K=3)</td>
<td>5</td>
<td>-1208.748168</td>
<td>7.11415</td>
<td>( p_0 = 0.00000, \omega_0 = 0.00000, ) ( p_1 = 0.00000, \omega_1 = 0.00000, ) ( p_2 = 1.00000, \omega_2 = 0.50741 )</td>
<td>---</td>
</tr>
<tr>
<td>M7: ( \beta )</td>
<td>2</td>
<td>-1208.747985</td>
<td>7.11419</td>
<td>( p = 99.00000, q = 96.09810, E[\omega] = 0.50744 )</td>
<td>Not allowed</td>
</tr>
<tr>
<td>M8: ( \beta ) &amp; ( \omega )</td>
<td>4</td>
<td>-1208.769984</td>
<td>7.14572</td>
<td>( p_0 = 0.47322, p = 0.00500, q = 1.30894, E[\omega] = 0.00381, p_1 = 0.52678, \omega_s = 1.00000 )</td>
<td>162, 326, 333</td>
</tr>
</tbody>
</table>

\(^a m\) is the number of free parameters for the \( \omega \) ratios.

\(^b \ell\) is the log-likelihood value.

\(^c \kappa\) is the Ts/Tv.

\(^d\) Posterior probability was obtained by the BEB method.
### Table S7.1. Cellular role of the studied 51 loci

<table>
<thead>
<tr>
<th>Loci</th>
<th>Designation</th>
<th>Cellular Role</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HKs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yraL</td>
<td>SAM-dependent methyltransferase protein synthesis: tRNA and rRNA base modification</td>
<td></td>
</tr>
<tr>
<td>araD</td>
<td>Ribulose-P epimerase energy metabolism</td>
<td></td>
</tr>
<tr>
<td>accD</td>
<td>AcCoA carboxylase/Carboxyl transferase beta fatty acid and phospholipid metabolism</td>
<td></td>
</tr>
<tr>
<td>gapA</td>
<td>Glyceraldehyde-3-P dehydrogenase energy metabolism</td>
<td></td>
</tr>
<tr>
<td>karG</td>
<td>Arginine kinase regulatory function</td>
<td></td>
</tr>
<tr>
<td>tsf</td>
<td>Translation and elongation factor TS protein synthesis</td>
<td></td>
</tr>
<tr>
<td>rs2</td>
<td>S2 ribosomal protein protein synthesis</td>
<td></td>
</tr>
<tr>
<td>CT686</td>
<td>ABC transporter-associated protein biosynthesis of cofactors, prosthetic groups and carriers</td>
<td></td>
</tr>
<tr>
<td>yfh0_1*</td>
<td>NifS-related enzyme central intermediary metabolism: cysteine desulfurase</td>
<td></td>
</tr>
<tr>
<td>parB</td>
<td>Chromosome partitioning protein cell process: cell division</td>
<td></td>
</tr>
<tr>
<td>dppF</td>
<td>ABC transport ATPase transport and binding protein</td>
<td></td>
</tr>
<tr>
<td>dppD</td>
<td>ABC Transport ATPase transport and binding protein</td>
<td></td>
</tr>
<tr>
<td>16SrRNA</td>
<td>16S ribosomal RNA rRNA synthesis</td>
<td></td>
</tr>
<tr>
<td><strong>HPs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT049</td>
<td>hypothetical protein biological process and molecular function unknown</td>
<td></td>
</tr>
<tr>
<td>CT143</td>
<td>hypothetical protein biological process and molecular function unknown</td>
<td></td>
</tr>
<tr>
<td>CT144</td>
<td>hypothetical protein biological process and molecular function unknown</td>
<td></td>
</tr>
<tr>
<td>CT676</td>
<td>ClpC ATPase unknown function</td>
<td></td>
</tr>
<tr>
<td>CT683</td>
<td>TPR-motif protein unknown function</td>
<td></td>
</tr>
<tr>
<td>CT622</td>
<td>CHLPN 76kDa protein-like unclassified: role not assigned</td>
<td></td>
</tr>
<tr>
<td><strong>CEPs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incD</td>
<td>Inclusion membrane protein D cell envelope</td>
<td></td>
</tr>
<tr>
<td>incE</td>
<td>Inclusion membrane protein E cell envelope</td>
<td></td>
</tr>
<tr>
<td>incF</td>
<td>Inclusion membrane protein F cell envelope</td>
<td></td>
</tr>
<tr>
<td>incG</td>
<td>Inclusion membrane protein G cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpB</td>
<td>Polymorphic membrane protein B cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpC</td>
<td>Polymorphic membrane protein C cell envelope</td>
<td></td>
</tr>
<tr>
<td>omcB</td>
<td>60kDa cysteine-rich outer membrane protein</td>
<td></td>
</tr>
<tr>
<td>ompA</td>
<td>Major outer membrane protein cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpB</td>
<td>PBP2-transglycolase/transpeptidase cell envelope: biosynthesis and degradation of murein sacculus and peptidoglycan</td>
<td></td>
</tr>
<tr>
<td>porB</td>
<td>Outer membrane protein B cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpD</td>
<td>Polymorphic membrane protein D cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpE</td>
<td>Polymorphic membrane protein E cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpF</td>
<td>Polymorphic membrane protein F cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpG</td>
<td>Polymorphic membrane protein G cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpH</td>
<td>Polymorphic membrane protein H cell envelope</td>
<td></td>
</tr>
<tr>
<td>pmpI</td>
<td>Polymorphic membrane protein I cell envelope</td>
<td></td>
</tr>
<tr>
<td><strong>IGRs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssb/pepA</td>
<td>IGR between ssDNA binding protein and Leucyl aminopeptidase A genes</td>
<td></td>
</tr>
<tr>
<td>yraL/CT049</td>
<td>IGR between yraL and CT049 genes</td>
<td></td>
</tr>
<tr>
<td>CT114/incD</td>
<td>IGR between hypothetical protein</td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Material

CT114 and incD genes
incD/IncE IGR between incD and incE genes
incF/IncG IGR between incF and incG genes
incG/IncA IGR between incG and incA genes
CT144/CT145 IGR between hypothetical protein CT144 and CT145 genes
rpoB/rl7 IGR between RNA polymerase beta and L7 ribosomal protein genes
pmpB/pmpC IGR between pmpB and pmpC genes
rs2/ompA IGR between rs2 and ompA genes
ompA/pbpB IGR between ompA and pbpB genes
CT683/CT684 IGR between TPR-motif protein CT683 and hypothetical protein CT684 genes
yfh0_1/parB IGR between yfh0_1 and parB genes
thdF/psdD IGR between Thiophene/Furan oxidation protein and Phosphatidylserine decarboxylase genes
glyQ/pgsA IGR between Glycyl tRNA synthetase and CDP-diacylglycerol-glycerol-3-P 3-phosphatidyltransferase genes
pmpF/pmpG IGR between pmpF and pmpG genes

Gene names and ORF numbers are based on the C. trachomatis D/UW3 genome annotation (GenBank: AE001273). Cellular roles are designated according to TIGR annotation (http://cmr.tigr.org). * The sequence of this gene is 100% similar to the one of yfh0_2 for the C. trachomatis A/Ha13 strain (GenBank: CP000051), which points to an annotation incongruence among the two strains.
Table S7.2. Primers used for PCR and sequencing

<table>
<thead>
<tr>
<th>Locus</th>
<th>Primers</th>
<th>Primer sequence (5’ to 3’)</th>
<th>Amplicon size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssb/pepA IGR</td>
<td>IGR 44/45-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TCCGGATCTTTCTCGTCTCTGATG</td>
<td>915</td>
</tr>
<tr>
<td></td>
<td>IGR 44/45-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>GCACCCCTGCTGCCAAGTTCA</td>
<td></td>
</tr>
<tr>
<td>CT114/incD IGR, incD, incD&lt;sub&gt;E&lt;/sub&gt; IGR, incE, incF, incF&lt;sub&gt;E&lt;/sub&gt; IGR, incG &amp; incG&lt;sub&gt;E&lt;/sub&gt; IGR</td>
<td>inc-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>AGCTCAGGAGGGCTCTGAAGA</td>
<td>2387</td>
</tr>
<tr>
<td></td>
<td>inc-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>GGCTCAGCGATAAGCTCTGAAGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>inc-4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>GCTTCATTGACGGCTCTGAAGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>inc-5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CGCCGCTGATATAGTTCA</td>
<td></td>
</tr>
<tr>
<td>CT143, CT144, &amp; CT144/CT145 IGR</td>
<td>CT144-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CTGGCGGACCACCCGTAAT</td>
<td>1531</td>
</tr>
<tr>
<td></td>
<td>CT144-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TCAACTGCGGCTCGCTCCTTC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CT144-3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ATAGGGCTGAAAGCTCTTA</td>
<td></td>
</tr>
<tr>
<td>rpoB/rl7 IGR</td>
<td>rpoB 315/316-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TCCCCGAGAATACTACTCTTC</td>
<td>799</td>
</tr>
<tr>
<td></td>
<td>rpoB 315/316-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ATCCGCTCAGAGTATAGTTCA</td>
<td></td>
</tr>
<tr>
<td>CT686, yfh0_1 &amp; yfh0_1/parB IGR</td>
<td>yfh0-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TCCGCGGAGCGCTGATATCTTA</td>
<td>1795</td>
</tr>
<tr>
<td></td>
<td>yfh0-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TCCCATGAGCAGCTCTGCTAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>yfh0-3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>GCCACATGCTTAACGTTCA</td>
<td></td>
</tr>
<tr>
<td>parB</td>
<td>parB-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGCGGAGGTTGCTCAGATTTC</td>
<td>1308</td>
</tr>
<tr>
<td></td>
<td>parB-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CTGGGCTCAGAGCTCAGATC</td>
<td></td>
</tr>
<tr>
<td>dppF &amp; dppD</td>
<td>dppF-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>AACGGTAAAGGCTCAGGGAAGACG</td>
<td>1210</td>
</tr>
<tr>
<td></td>
<td>dppF-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGATCGAGGCTCAGGGAAGAT</td>
<td></td>
</tr>
<tr>
<td>porB</td>
<td>porB-1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TGGAAGCGGCTTAGATATGTTCA</td>
<td>712</td>
</tr>
<tr>
<td></td>
<td>porB-2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ATGCCGATTACCCGAGCTGCTAT</td>
<td></td>
</tr>
</tbody>
</table>

Primers were designed based on published genome sequences of prototype strains A/Har13 and D/UW3 (GenBank: CP000051 & AE001273, respectively).

<sup>a</sup> Primers also used for automated sequencing.

Figure S7.1 - Overall mean genetic distances among the 15 main C. trachomatis genotypes for the 51 loci. The nucleotide p-distances are represented by vertical bars with the respective SE. Loci categories are illustrated by different bar colors. Gene names and ORF numbers are based on the C. trachomatis D/UW3 genome annotation (GenBank: AE001273).
Figure S7.2 - Pathogen evolutionary history by *loci* category. Phylogenetic reconstructions (NJ, K2P model) are based on concatenated nucleotide sequences from the 4 *loci* categories: (A) IGRs, (B) HKs, (C) HPs and (D) CEPs. This analysis was performed without the recombinant *ompA* gene (see methods of CHAPTER 7 for details). Branch lengths are proportional to distances between strains. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) is shown next to the branch nodes. Ocular, epithelial-genital and LGV strains are represented within yellow, pink and blue boxes, respectively.
Table S7.3. Deletion events among the 15 main *C. trachomatis* prototype strains

<table>
<thead>
<tr>
<th>Loci</th>
<th>Deletion events</th>
<th>Position (bp)</th>
<th>Size (bp)</th>
<th>Prototype strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssb/pepA IGR</td>
<td></td>
<td>29-30</td>
<td>2</td>
<td>All but LGV</td>
</tr>
<tr>
<td>yraL</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>yraL/CT049 IGR</td>
<td></td>
<td>62</td>
<td>1</td>
<td>All but H, J</td>
</tr>
<tr>
<td>CT049</td>
<td></td>
<td>64-90</td>
<td>27</td>
<td>All but H, J</td>
</tr>
<tr>
<td></td>
<td></td>
<td>172-177</td>
<td>6</td>
<td>All but F, H, I, J</td>
</tr>
<tr>
<td></td>
<td></td>
<td>379-387</td>
<td>9</td>
<td>H, I, J</td>
</tr>
<tr>
<td></td>
<td></td>
<td>595-597</td>
<td>3</td>
<td>All but H, I, J</td>
</tr>
<tr>
<td></td>
<td></td>
<td>883-894</td>
<td>12</td>
<td>H, I, J</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1228-1233</td>
<td>6</td>
<td>D, G, H, K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1462-1467</td>
<td>6</td>
<td>All but D, G, H, K</td>
</tr>
<tr>
<td>CT114/incD IGR</td>
<td></td>
<td>22</td>
<td>1</td>
<td>A-C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28</td>
<td>1</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>86</td>
<td>1</td>
<td>LGV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89-131</td>
<td>43</td>
<td>A-C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>121</td>
<td>1</td>
<td>LGV</td>
</tr>
<tr>
<td>incD</td>
<td></td>
<td>1-28</td>
<td>28</td>
<td>All but A-C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>1</td>
<td>A-C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>37-39</td>
<td>3</td>
<td>All but A-C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>437</td>
<td>1</td>
<td>LGV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>458-473</td>
<td>16</td>
<td>All but LGV</td>
</tr>
<tr>
<td>incD/incE IGR</td>
<td></td>
<td>1-16</td>
<td>16</td>
<td>LGV</td>
</tr>
<tr>
<td>incE</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>incF</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>incF/incG IGR</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>incG</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>incG/incA IGR</td>
<td></td>
<td>55-57</td>
<td>3</td>
<td>A, H, J, L1, L3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>56-57</td>
<td>2</td>
<td>B-C, D, F, G, I, K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>122</td>
<td>1</td>
<td>D, G, I, K</td>
</tr>
<tr>
<td>araD</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>CT143</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>CT144</td>
<td></td>
<td>289-297</td>
<td>9</td>
<td>All but J, LGV</td>
</tr>
<tr>
<td>CT144/CT145 IGR</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>accD</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>rpoB/rl7 IGR</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>pmpB</td>
<td></td>
<td>1158-1163</td>
<td>6</td>
<td>A-D, G-K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1161-1163</td>
<td>3</td>
<td>LGV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1971-1973</td>
<td>3</td>
<td>E, F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2844-2852</td>
<td>9</td>
<td>LGV</td>
</tr>
<tr>
<td>pmpB/pmpC IGR</td>
<td></td>
<td>98</td>
<td>1</td>
<td>LGV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>118</td>
<td>1</td>
<td>A-C, L2</td>
</tr>
<tr>
<td>pmpC a</td>
<td></td>
<td>296-337</td>
<td>42</td>
<td>D, G-K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2365-2394</td>
<td>30</td>
<td>LGV</td>
</tr>
<tr>
<td>omcB</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>gapA</td>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>CT622</td>
<td></td>
<td>244-249</td>
<td>6</td>
<td>LGV</td>
</tr>
<tr>
<td></td>
<td></td>
<td>253-270</td>
<td>18</td>
<td>D, I, K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>262-267</td>
<td>6</td>
<td>A-C, G, H</td>
</tr>
<tr>
<td>Gene/IGR</td>
<td>Start</td>
<td>End</td>
<td>Coverage</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>----</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>karG</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>CT676</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>tsf</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>rs2</td>
<td>733-744</td>
<td>12</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>799-781</td>
<td>3</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>799-807</td>
<td>9</td>
<td>All but B, C</td>
<td></td>
</tr>
<tr>
<td>rs2/ompA IGR</td>
<td>17</td>
<td>1</td>
<td>F, L1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>1</td>
<td>E, F, LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>83-84</td>
<td>2</td>
<td>A-C, D, G-K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>221-222</td>
<td>2</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>222</td>
<td>1</td>
<td>A, D, H-K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>233</td>
<td>1</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>238</td>
<td>1</td>
<td>A, C, D, H-K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>255</td>
<td>1</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>257</td>
<td>1</td>
<td>H, J</td>
<td></td>
</tr>
<tr>
<td></td>
<td>259-262</td>
<td>4</td>
<td>A, C, D, H-K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>272-273</td>
<td>2</td>
<td>A, C, D, H-K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>286-287</td>
<td>2</td>
<td>A, C, D, H-K, L3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>287</td>
<td>1</td>
<td>B, Ba, E, L1, L2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>329-337</td>
<td>9</td>
<td>All but LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>364-369</td>
<td>6</td>
<td>B, Ba, E-G, LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>391</td>
<td>1</td>
<td>A, C, D, H-K, L3</td>
<td></td>
</tr>
<tr>
<td>ompA/tpB IGR</td>
<td>493-513</td>
<td>21</td>
<td>B, Ba, F</td>
<td></td>
</tr>
<tr>
<td>phpB</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>CT683</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>CT683/CT684 IGR</td>
<td>204-205</td>
<td>2</td>
<td>E, F, LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>205</td>
<td>1</td>
<td>A-C, D, G, I-K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>227</td>
<td>1</td>
<td>All but J</td>
<td></td>
</tr>
<tr>
<td>CT686</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>yfh0_1</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>yfh0_1/parB IGR</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>parB</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>dppF</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>dppD</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>thdF/psdD IGR</td>
<td>126</td>
<td>1</td>
<td>All but LGV</td>
<td></td>
</tr>
<tr>
<td>porB</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>16SrRNA</td>
<td>835</td>
<td>1</td>
<td>All but A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>1</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>glyQ/pgsA IGR</td>
<td>170</td>
<td>1</td>
<td>D, K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>310-317</td>
<td>8</td>
<td>G, LGV</td>
<td></td>
</tr>
<tr>
<td>pmpD</td>
<td>3124-3126</td>
<td>3</td>
<td>E, F, LGV</td>
<td></td>
</tr>
<tr>
<td>pmpE</td>
<td>355-357</td>
<td>3</td>
<td>D-H, K</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1048-1053</td>
<td>6</td>
<td>A-C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1159-1161</td>
<td>3</td>
<td>A-C, I, J</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1411-1419</td>
<td>9</td>
<td>A-C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1426-1428</td>
<td>3</td>
<td>A-C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1540-1551</td>
<td>12</td>
<td>All but A-C</td>
<td></td>
</tr>
<tr>
<td>pmpF</td>
<td>1292-1294</td>
<td>3</td>
<td>LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2104-2106</td>
<td>3</td>
<td>H, I, J</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2115-2117</td>
<td>3</td>
<td>LGV</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Start (bp)</td>
<td>Length</td>
<td>Region</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
<td>--------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>pmpF/pmpG IGR</td>
<td>28</td>
<td>1</td>
<td>D-G, K, LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>1</td>
<td>D-G, K, LGV</td>
<td></td>
</tr>
<tr>
<td>pmpG</td>
<td>1046-1048</td>
<td>3</td>
<td>L2</td>
<td></td>
</tr>
<tr>
<td>pmpH</td>
<td>484-519</td>
<td>36</td>
<td>LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>484-489</td>
<td>6</td>
<td>I, J</td>
<td></td>
</tr>
<tr>
<td></td>
<td>691-693</td>
<td>3</td>
<td>LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>700-708</td>
<td>9</td>
<td>All but LGV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>868-870</td>
<td>3</td>
<td>L1, L3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>886-888</td>
<td>3</td>
<td>All but A-C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1675-1677</td>
<td>3</td>
<td>A-C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1885-1890</td>
<td>6</td>
<td>All but A-C</td>
<td></td>
</tr>
</tbody>
</table>

*For pmpC the information presented in this table refers to remnant fragments of the putative IS-like elements [121].
Genotype B - (TW5)

MKKLLKSVLVFAALSSASSLQALPVGNPAEPSLMIDGILWEGFGGDPCDPCTTWIDASMQYVDPWFVRXKTQVANKEMQMGAPPTTQYGNAPSTDAZVLPWQGRHHQDAEMTFNAAQCAKMWQDRFVCTGASDCKGKNSASF
NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Genotype Ba - (Apache2)

MKKLLKSVLVFAALSSASSLQALPVGNPAEPSLMIDGILWEGFGGDPCDPCTTWIDASMQYVDPWFVRXKTQVANKEMQMGAPPTTQYGNAPSTDAZVLPWQGRHHQDAEMTFNAAQCAKMWQDRFVCTGASDCKGKNSASF
NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Genotype D - (UW3)

LAEAILDVTTLNPTIAGKGSVVSAGTDNELADTMQIVSLQLNKMKSRKSCGIAVGTTIVDADKYAVTVETRLIDERAAHVNAQFRF.

Genotype E - (Boor)

MKKLLKSVLVFAALSSASSLQALPVGNPAEPSLMIDGILWEGFGGDPCDPCTTWIDASMQYVDPWFVRXKTQVANKEMQMGAPPTTQYGNAPSTDAZVLPWQGRHHQDAEMTFNAAQCAKMWQDRFVCTGASDCKGKNSASF
NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Genotype F - (IC-Cal2)

LAEAILDVTTLNPTIAGKGSVVSAGTDNELADTMQIVSLQLNKMKSRKSCGIAVGTTIVDADKYAVTVETRLIDERAAHVNAQFRF.

Genotype D3 - (TW448)

MKKLLKSVLVFAALSSASSLQALPVGNPAEPSLMIDGILWEGFGGDPCDPCTTWIDASMQYVDPWFVRXKTQVANKEMQMGAPPTTQYGNAPSTDAZVLPWQGRHHQDAEMTFNAAQCAKMWQDRFVCTGASDCKGKNSASF
NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Genotype G - (UW57)

NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Genotype A - (Har13)

LAEAILDVTTLNPTIAGKGSVVSAGTDNELADTMQIVSLQLNKMKSRKSCGIAVGTTIVDADKYAVTVETRLIDERAAHVNAQFRF.

Genotype C - (TW3)

NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Genotype H - (UW4)

NLVOLFEDONQVIYKQFVQLNLGVELVSTFTTQAGWAVVIIGAACECQGQASDFOYDQPKVGSHELWNAOQAFNEMRAVOYDQPKVGSHELWNAOQAFNEMRAVO9

Supplemental Material
**Figure S.2** - Putative chromosomal mosaic structures for the 29 recombinant *C. trachomatis* clinical isolates. The 11 loci, involving two statistically-confirmed recombination hotspots and representing four well-separated regions of the *C. trachomatis* chromosome [114], are represented by colored boxes and arrows (showing the coding strand direction of each gene), according to the putative parental prototype strain(s). The presence of recombination was evaluated using phylogenetic reconstructions, compatibility matrices, and statistically based recombination programs (SimPlot/BootScan and RIP).
REFERENCES


References


174

References


References


231. Van der Bij AK, Spaargaren J, Morre SA, Fennema HS, Mindel A, et al. (2006) Diagnostic and
clinical implications of anorectal lymphogranuloma venereum in men who have sex with men: a
trachomatis L2 strains identified in a recent outbreak of lymphogranuloma venereum in Vienna,
Austria. Sex Transm Dis 35: 377-382.
235. van de Laar MJ (2006) The emergence of LGV in western Europe: what do we know, what can
chlamydial major outer membrane proteins: VS4 domains are not required for pore formation but
components of MOMP of Chlamydia trachomatis serovar F. Protein Sci 15: 122-134.
of the membrane-spanning beta-strands of the major outer membrane protein of Chlamydia.
Protein Sci 11: 1854-1861.
antibodies recognize epitopes located on the major outer membrane protein of Chlamydia
trachomatis serovar- and serogroup-specific major outer membrane protein determinants. Infect
Immun 57: 636-638.
245. Morrison RP, Manning DD, Caldwell HD (1992) Immunology of Chlamydia trachomatis
infections: immunoprotective and immunopathogenic responses. In: Quinn TC, editor. Sexually
3: S521-523.
248. de la Maza MA, de la Maza LM (1995) A new computer model for estimating the impact of
vaccination protocols and its application to the study of Chlamydia trachomatis genital infections.
Chlamydia trachomatis omp1 genotypes among sexually transmitted disease patients in Sweden. J
Chlamydia trachomatis omp1 gene detected in samples from men tested in male-only saunas in
References


