
Sandeep J. Joseph, Xavier Didelot, James Rothschild, Henry J.C. de Vries, Servaas A. Morré, Timothy D. Read, and Deborah Dean

1Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine
2Department of Infectious Disease Epidemiology, Imperial College, London, United Kingdom
3Center for Immunobiology and Vaccine Development, Children's Hospital Oakland Research Institute, California
4Department of Dermatology, Academic Medical Centre, University of Amsterdam, The Netherlands
5STI Outpatient Clinic, Infectious Diseases Cluster, Public Health Service Amsterdam, The Netherlands
6Centre for Infection and Immunity, Academic Medical Centre, University of Amsterdam, The Netherlands
7Department of Pathology, Laboratory of Immunogenetics, VU University Medical Center, Amsterdam, The Netherlands
8Department of Genetics and Cell Biology, Institute of Public Health Genomics, School for Public Health and Primary Care and School for Oncology & Developmental Biology, Faculty of Health, Medicine & Life Sciences, University of Maastricht, The Netherlands
9Department of Human Genetics, Emory University School of Medicine
10Department of Medicine, University of California, San Francisco
11Joint Graduate Program in Bioengineering, University of California, San Francisco and Berkeley

*yThese authors contributed equally to this work.

*Corresponding author: E-mail: ddean@chori.org.

Associate editor: Daniel Falush

Abstract

The large number of sexually transmitted diseases and ocular trachoma cases that are caused globally each year by Chlamydia trachomatis has made this organism a World Health Organization priority for vaccine development. However, there is no gene transfer system for Chlamydia to help identify potential vaccine targets. To accelerate discoveries toward this goal, here we analyzed the broadest diversity of C. trachomatis genomes to date, including 25 geographically dispersed clinical and seven reference strains representing 14 of the 19 known serotypes. Strikingly, all 32 genomes were found to have evidence of DNA acquisition by homologous recombination in their history. Four distinct clades were identified, which correspond to all C. trachomatis disease phenotypes: lymphogranuloma venereum (LGV; Clade 1); noninvasive urogenital infections (Clade 2); ocular trachoma (Clade 3); and protocolitis (Clade 4; also includes some noninvasive urogenital infections). Although the ancestral relationship between clades varied, most strains acted as donor and recipient of recombination with no evidence for barriers to genetic exchange. The niche-specific LGV and trachoma clades have undergone less recombination, although the opportunity for mixing with strains from other clades that infect the rectal and ocular mucosa, respectively, is evident. Furthermore, there are numerous occasions for gene conversion events through sequential infections at the same anatomic sites. The size of recombinant segments is relatively small (~357 bp) compared with in vitro experiments of various C. trachomatis strains but is consistent with in vitro estimates for other bacterial species including Escherichia coli and Helicobacter pylori. Selection has also played a crucial role during the diversification of the organism. Clade 2 had the lowest nonsynonymous to synonymous ratio (dN/dS) but the highest effect of recombination, which is consistent with the widespread occurrence of synonymous substitutions in recombined genomic segments. The trachoma Clade 3 had the highest dN/dS estimates, which may be caused by an increased effect of genetic drift from niche specialization and a reduced effective population size. The degree of drift, selection, and recombination in C. trachomatis suggests that the challenge will remain to identify genomic regions that are stable and cross protective for the development of an efficacious vaccine.

Key words: recombination, Chlamydia trachomatis, population genomics, selection.

Introduction

Chlamydia trachomatis is an ancient human pathogen that was first described as a cause of the chronic eye disease referred to as trachoma. Trachoma can result in blindness following the onset of trichiasis, defined as in-turned eyelashes that touch the globe of the eye. Trichiasis was first described in the 27th century BC in China and later in 1550 BC in the Ebers Papyrus of Egypt (Dean 2010). C. trachomatis is also responsible for lymphogranuloma venereum (LGV), a sexually transmitted disease (STD) that was recognized in the late
C. trachomatis of various genomics of multiple strains of the organism. Knowledge toward these goals can be acquired through comparative genomics (Wang et al. 2011), the lack of a reliable gene transfer system (Brunham and Rey-Ladino 2005). However, valuable knowledge toward these goals can be acquired through comparative genomics of multiple strains of the organism.

Despite this long history and public health importance, C. trachomatis is extremely poorly understood largely because of its obligate intracellular growth, limited appropriate animal models, and a lack of genetic systems for experimental manipulation (Belland et al. 2004; Brunham and Rey-Ladino 2005; Editorial—Nat. Rev. Microbiol. 2005; Hafner et al. 2008; Hafner and McNeilly 2008). Without a gene transfer system, alternative species for DNA mobilization have been tried with limited success (e.g., yeast, mammalian cells, other bacteria); RNA interference is also possible for pseudogenetic knock-out studies, although collateral effects are not properly understood (Fields et al. 2003; Alzhanov et al. 2004; Delevoye et al. 2004; Sisko et al. 2006; Cortes et al. 2007; Li et al. 2008). Although a recent study showed that C. trachomatis can be transfected with a chlamydial plasmid to restore glycoprotein synthesis (Wang et al. 2011), the lack of a reliable gene transfer system limits our ability to understand disease pathogenesis and gene function linked to virulence, protective immunity and host tissue specificity. Consequently, previous efforts at designing a vaccine have repeatedly failed (Brunham and Rey-Ladino 2005). However, valuable knowledge toward these goals can be acquired through comparative genomics of multiple strains of the organism.

Over the last decade, comparative genetics and genomics of various C. trachomatis strains has shown that recombination is frequent, causing alteration in certain critical genes (Brunham et al. 1994; Dean et al. 1995; Millman et al. 2001; Gomes et al. 2006, 2007; Joseph et al. 2011; Somboonna et al. 2011; Joseph and Read 2012). A prime example of this is the recent whole-genome evidence for recombination occurring between virulent LGV and nonvirulent C. trachomatis strains co-infecting the same rectal mucosal niche (Somboonna et al. 2011). The recombinant strain, referred to as L2/C, was isolated from a man who presented with hemorrhagic proctitis but no inguinal syndrome, which occurs following lymphatic spread of LGV strains to the inguinal lymph nodes. This lack of spread was likely due to the strain’s cytotoxicity, which was caused by the acquisition of a functional toxin gene from a D strain. Indeed, in culture, the isolate developed marked cellular toxicity unlike LGV causing strains described to date. The genomic data, therefore, provided the first evidence for the emergence of a hypervirulent C. trachomatis strain following recombination. However, the evolutionary mechanisms of recombination in C. trachomatis remain elusive. In this study, we compared 32 genomes representing 14 serotypes and 25 clinical strains to assess drift, recombination and selection in C. trachomatis and to understand the major evolutionary forces acting on the genome of this bacterium.

Materials and Methods

C. trachomatis Strains, Clonal Purification, and Generation of Genomic DNA

The 32 strains of C. trachomatis used in this study are summarized in Table 1. Available published C. trachomatis genomes included 19 reference and clinical strains. We sequenced 13 additional strains, eight of which were recent clinical isolates from STD populations that were identified as recombinants based on MLST (Dean et al. 2009). The 32 strains represented 14 serotypes from seven different countries worldwide isolated between 1958 and 2010. Ocular strains were from trachoma patients in Egypt, Gambia, Taiwan, and Tanzania, whereas urogenital strains were from STD populations in The Netherlands, Sweden, United Kingdom, and the United States.

Each of the 13 new strains was individually plaque purified in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for the ompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for the ompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for the ompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for the ompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for the ompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for the ompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for theompA and MLST genes as previously detailed (Somboonna et al. 2011). Single clones were used for purification of genomic DNA and for propagation in HeLa 229 cells, and clonal purity was confirmed by sequencing 10 clones of each strain for theompA and MLST genes as previously detailed (Somboonna et al. 2011).
Ortholog Retrieval, Core Cluster Alignment, and Phylogenetic Inference

The complete predicted proteome from all the genomes annotated in this study along with the re-annotated protein sequences of the previously published genomes was searched against itself using BLASTP with an e-value cutoff of 1e-05.

The best BLASTP scores were converted into a normalized similarity matrix using the OrthoMCL (Li et al. 2003) algorithm, which utilizes an additional step of the Markov Clustering algorithm (MCL) to improve sensitivity and specificity of the orthologous sequences identified. Core genes are defined as the protein-coding gene clusters that are shared by all C. trachomatis strains.

The program MUSCLE (Edgar 2004) was used for multiple sequence alignment (MSA) of the core protein-coding genes using default settings. These protein alignments were reverse-translated to codon-based nucleotide alignments using PAL2NAL (Suyama et al. 2006), which also used the corresponding DNA sequences for positive selection analysis (discussed later). Another set of the MSA was filtered for uninformative characters by GBLOCKS (Castresana 2000) using default settings. Phylogenetic analyses for gene families were conducted using UPGMA (Felsenstein 1989) under the Jones–Taylor–Thornton (JTT) substitution model. The support for the data from each of the internal branch of the phylogeny was estimated using 100 bootstraps. Whole-genome phylogenies were performed on a concatenate of aligned individual core genes followed by neighbor joining (NJ) and maximum parsimony (MP) using NEIGHBOR in the PHYLIP package and the extended majority rule consensus tree was inferred. The MP tree was constructed using PROTPARS in the PHYLIP package with 100 randomizations of input order, and four clades were defined.

Analysis of Positive Selection

Genes under positive selection were identified using codeml as implemented in PAML version 4.4 (Yang 2007). We applied the branch-site test 2 (Zhang et al. 2005) to identify genes

Table 1. C. trachomatis Strains Used in This Study.

<table>
<thead>
<tr>
<th>Strain Name</th>
<th>Strain Notation</th>
<th>Serotype</th>
<th>Country</th>
<th>Anatomic Source</th>
<th>Year Isolated</th>
<th>Accession No.</th>
<th>Genome Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>B/TZ1A882/OT</td>
<td>B/Tz</td>
<td>B</td>
<td>Tanzania</td>
<td>Ocular</td>
<td>1998</td>
<td>FM872308</td>
<td>Seth-Smith et al. (2009)</td>
</tr>
<tr>
<td>C/TW-3/OT*</td>
<td>C</td>
<td>C</td>
<td>Taiwan</td>
<td>Conjunctiva</td>
<td>1959</td>
<td>SRA051538.1</td>
<td>This publication</td>
</tr>
<tr>
<td>D/2932</td>
<td>D/2932</td>
<td>D</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>NA</td>
<td>ACF001000001</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>D/84s</td>
<td>D/84s</td>
<td>D</td>
<td>USA (SF Bay Area)</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051539.1</td>
<td>This publication</td>
</tr>
<tr>
<td>D/2s</td>
<td>D/2s</td>
<td>D</td>
<td>USA (SF Bay Area)</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051544.1</td>
<td>This publication</td>
</tr>
<tr>
<td>D/43nL</td>
<td>D/43nL</td>
<td>D</td>
<td>The Netherlands</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051526.1</td>
<td>This publication</td>
</tr>
<tr>
<td>D_EC</td>
<td>D/EC</td>
<td>D</td>
<td>USA (Montana)</td>
<td>Genital tract</td>
<td>2010</td>
<td>CP002054</td>
<td>Sturdevant et al. (2010)</td>
</tr>
<tr>
<td>D_LC</td>
<td>D_LC</td>
<td>D</td>
<td>USA (Montana)</td>
<td>Genital tract</td>
<td>2010</td>
<td>CP002052</td>
<td>Sturdevant et al. (2010)</td>
</tr>
<tr>
<td>E/SW2</td>
<td>E/Swed</td>
<td>E</td>
<td>Sweden</td>
<td>Urethra</td>
<td>2001</td>
<td>FN652779</td>
<td>Unemo et al. (2010)</td>
</tr>
<tr>
<td>E/11023</td>
<td>E/11023</td>
<td>E</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>NA</td>
<td>CP001890</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>E/150</td>
<td>E/150</td>
<td>E</td>
<td>USA (Seattle)</td>
<td>Rectum</td>
<td>NA</td>
<td>CP001886</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>E/5s</td>
<td>E/5s</td>
<td>E</td>
<td>USA (SF Bay Area)</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051547.1</td>
<td>This publication</td>
</tr>
<tr>
<td>F/1</td>
<td>F/1</td>
<td>F</td>
<td>USA (SF Bay Area)</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051547.1</td>
<td>This publication</td>
</tr>
<tr>
<td>F/70</td>
<td>F/70</td>
<td>F</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>NA</td>
<td>ABYF001000001</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>F/38nL</td>
<td>F/38nL</td>
<td>F</td>
<td>The Netherlands</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051469.2</td>
<td>This publication</td>
</tr>
<tr>
<td>G/UW-57*</td>
<td>G</td>
<td>G</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>1971</td>
<td>SRA051545.1</td>
<td>This publication</td>
</tr>
<tr>
<td>G/9301</td>
<td>G/9301</td>
<td>G</td>
<td>USA (Seattle)</td>
<td>Urethra</td>
<td>NA</td>
<td>CP0001930</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>G/9768</td>
<td>G/9768</td>
<td>G</td>
<td>USA (Seattle)</td>
<td>Rectum</td>
<td>NA</td>
<td>CP001887</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>G/11222</td>
<td>G/11222</td>
<td>G</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>NA</td>
<td>CP001888</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>G/11074</td>
<td>G/11074</td>
<td>G</td>
<td>USA (Seattle)</td>
<td>Rectum</td>
<td>NA</td>
<td>CP001889</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>H/UW-4/CL*</td>
<td>H</td>
<td>H</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>1965</td>
<td>SRA051548.1</td>
<td>This publication</td>
</tr>
<tr>
<td>H/18s</td>
<td>H/18s</td>
<td>H</td>
<td>USA (SF Bay Area)</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051541.1</td>
<td>This publication</td>
</tr>
<tr>
<td>Ia/UW-202a</td>
<td>Ia</td>
<td>Ia</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>1985</td>
<td>SRA051537.1</td>
<td>This publication</td>
</tr>
<tr>
<td>J/6276</td>
<td>J/6276</td>
<td>J</td>
<td>USA (Seattle)</td>
<td>Cervix</td>
<td>NA</td>
<td>ABYD01000001</td>
<td>Jeffrey et al. (2010)</td>
</tr>
<tr>
<td>Ja/47nL</td>
<td>Ja/47nL</td>
<td>Ja</td>
<td>The Netherlands</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051542.1</td>
<td>This publication</td>
</tr>
<tr>
<td>Ja/26s</td>
<td>Ja/26s</td>
<td>Ja</td>
<td>USA (SF Bay Area)</td>
<td>Cervix</td>
<td>2000s</td>
<td>SRA051540.1</td>
<td>This publication</td>
</tr>
<tr>
<td>L1/434/BUC*</td>
<td>L1</td>
<td>L1</td>
<td>USA (California)</td>
<td>Lymph node</td>
<td>1968</td>
<td>AM884176</td>
<td>Thomson et al. (2008)</td>
</tr>
<tr>
<td>L1b/UCH-1</td>
<td>L1b</td>
<td>L1b</td>
<td>UK</td>
<td>Rectum</td>
<td>1968</td>
<td>AM884177</td>
<td>Thomson et al. (2008)</td>
</tr>
<tr>
<td>L1C</td>
<td>L1C</td>
<td>L1C</td>
<td>USA (SF Bay Area)</td>
<td>Rectum</td>
<td>2000s</td>
<td>NC_015744</td>
<td>Somboonna et al. (2011)</td>
</tr>
</tbody>
</table>

Note.—NA, not available.

*Reference strains are defined as strains that were originally isolated in the 1950s or 1960s and have been laboratory adapted over the last decades.
that are under positive selection in each of the four clades of the whole-genome tree. Initially, the inferred whole-genome tree was used for all PAML analyses. For all genes that were identified as being under positive selection, PAML was re-run to check whether the positive selection results obtained using gene-specific trees differed from that of the whole genome tree. For each test, the likelihood of a model that does not allow positive selection (null model) was compared with a model that allows positive selection (alternative model) using a Likelihood Ratio Test with one degree of freedom. Correction for multiple testing was performed using the method of Benjamini and Hochberg (1995) implemented in the software Q-value (Storey 2002).

Analysis of Homologous Recombination
ClonalFrame (Didelot and Falush 2007) version 1.2 was applied to the genomic regions found by MAUVE (Darling et al. 2004, 2010) to be homologous in all 32 genomes. ClonalFrame was run for a total of 40,000 iterations, of which the first half was discarded as MCMC burn-in. Four runs were performed independently and in parallel and were found to be highly congruent in terms of the phylogenies reconstructed and recombination events detected. For each branch of the reconstructed genealogy, the number of mutation events, recombination events, and substitutions introduced by recombination were estimated. The relative effect of recombination and mutation (r/m) in the whole sample and for each clade was calculated by forming the ratio of the number of substitutions introduced by recombination and mutation for the relevant branches of the phylogeny. The phylogenetic tree produced by ClonalFrame has branch lengths measured in coalescent units of time, which are equal to the effective population size Ne times the duration of a generation. During one coalescent unit of time, the expectation is that there are \( \theta/2 = N_e \mu \) mutation events and \( \rho/2 = N_e r \) recombination events, where \( \theta \) and \( \rho \) are the scaled rates of mutation and recombination estimated by ClonalFrame, and \( \mu \) and \( r \) are the per-generation rates of mutation and recombination (Didelot et al. 2011).

Attribution of Origins to the Recombination Events
For each branch of the tree reconstructed by ClonalFrame, we defined recombined fragments as genomic intervals with a posterior probability of recombination above 50% at every site and reaching 95% in at least one site (Didelot et al. 2011). Each such recombined fragment was searched for using BLAST against the whole database containing all the “finished” genome and plasmid sequences of Chlamydiaceae bacterial species minus the strains of the clade affected by the import. The hits with the highest normalized BLASTN score along with a percent identity of at least 98% were kept. If all these hits were with strains belonging to the same clade, the origin of the event was attributed to this clade, and otherwise the origin was called ambiguous.

Structure Analysis
The Bayesian analysis method STRUCTURE (Pritchard et al. 2000) version 2.3 was used to identify the underlying population structure present in our data. The linkage model of STRUCTURE was used, which accounts for the correlation between nearby sites arising in admixed populations (Falush et al. 2003). Four independent runs were performed for each value of the number of ancestral populations \( K \) ranging from 2 to 10. Each run consisted of 100,000 MCMC iterations, of which the first half was discarded as MCMC burn-in. Convergence and mixing of the program were found to be acceptable by manual comparison of independent runs with the same value of \( K \). The optimal value was found to be \( K = 4 \) by comparing the posterior probabilities of the data given each value of \( K \) from 2 to 10, and identifying the value of \( K \) where the posterior probabilities plateau (Pritchard and Falush 2009). We also applied the analytical method based on the second-order rate of change of the likelihood function with respect to \( K \) as described previously (Evanno et al. 2005), which also resulted in the estimate \( K = 4 \).

Substitution Rate (dN/dS) Calculations
To calculate the nonsynonymous (dN) and synonymous (dS) substitutions for an ortholog in a pair of C. trachomatis strains, we aligned their amino acid sequences using MUSCLE (Edgar 2004), and the resulting protein alignments were converted to nucleotide alignments using PAL2NAL (Suyama et al. 2006). We applied the YN00 method (Yang and Nielsen 2000) implemented in the PAML package to calculate the dN/dS ratios (Rocha et al. 2006). For each pair of strains, we estimated the median value of dN/dS and further disentangled the contribution of each strain to the pairwise dN/dS using ANOVA (analysis of variance) and the nonparametric Kruskal–Wallis test.

Results
Genome Sequencing
We sequenced 13 novel C. trachomatis genomes comprising one strain from serotype C (ocular), three from serotype D (urogenital), one from serotype E (urogenital), two from serotype F (urogenital), one from serotype G (urogenital), two from serotype H (urogenital), one from serotype I (urogenital), and two from serotype J (urogenital). Genomes for serotypes C, H, and J had never been previously sequenced. The estimated redundancy of coverage for the sequenced genomes was 15- to 45-fold. We added to this new genomic data 19 previously published C. trachomatis genomes, making a total of 32 genomes analyzed in this study as detailed in table 1.

Ortholog Identification and Whole-Genome Phylogeny for C. trachomatis
We identified 786 core genes present among all of the 32 genomes used in this study, which is between 85 and 90% of the total for any strain. This considerable number confirms the high level of sequence conservation and genome
sequence conservation of *C. trachomatis* (Thomson and Clarke 2010). There were 184 Locally Collinear Blocks (LCBs), which are basically homologous regions identified by the MAUVE alignment of all 32 genomes. These blocks represent 987,264 bp of the 1 Mbp average genome size of *C. trachomatis*, indicating a higher proportion of homologous regions. Based on the reference genome, D/UW3/CX, MAUVE identified 18,045 variable sites that include both SNPs and indels. The whole-genome phylogeny was inferred using two different methods. The first method utilized a concatenated alignment (super alignment) of amino acid residues of the 786 translated core genes followed by NJ and MP phylogenetic construction. The second approach was based on ClonalFrame analysis. Both methods yielded trees with the same topology, based on which we defined with confidence the following four clades: Clade 1 (yellow) contained L2, L2b, and L2c (LGV strains); Clade 2 (dark blue) contained all the E, F, J, and two D strains; Clade 3 (light blue) contained all A, B, and C strains (trachoma strains); and Clade 4 (red) contained all the H, J, G, Ia, and five D strains (fig. 1). Clade 4 comprised both noninvasive prevalent (D/UW3) and nonprevalent (G, H, Ia, and J) strains. One serotype (D) was split between two clades (2 and 4), and several others did not form subclades (e.g., E and F intermingled within Clade 2), highlighting the limitations of serotyping as a phylogenetic marker.

Population Structure of *C. trachomatis*

The population structure of *C. trachomatis* was reconstructed using STRUCTURE (Pritchard et al. 2000; Falush et al. 2003).

The ancestry among the different strains was analyzed based on the patterns of polymorphisms they share using the linkage model of STRUCTURE. The number of ancestral populations (K) needed to explain the current population structure was estimated to be equal to K = 4 based on previously described techniques (Evanno et al. 2005) (supplementary fig. S1, Supplementary Material online). The proportion of ancestry from each of these ancestral populations is shown in color for each strain in figure 2. The strains were classified according to which ancestral population provided the highest proportion of genetic material for each strain, and this classification corresponded exactly to the clades defined using a phylogenetic approach (fig. 1). However, none of the 32 genomes was entirely from one of the four STRUCTURE populations, suggesting that they have all acted frequently as donor and recipient of DNA exchanges.

ClonalFrame (Didelot and Falush 2007) reconstructs the clonal relationships between members of a sample from a population using a Bayesian phylogenetic framework which accounts not only for the mutation events but also for the recombination events. Figures 3 and 4 show, respectively, the clonal genealogy inferred from our whole genome alignment and the distribution of recombination events on the branches of the clonal genealogy along the whole genome alignment of *C. trachomatis*. The four clusters identified by STRUCTURE corresponded to clades of the ClonalFrame tree. Based on the combined evidence from STRUCTURE and ClonalFrame analyses, these four groups can confidently be called clades (lineages) of *C. trachomatis*. The parameter estimates from the ClonalFrame analysis for these four clades are summarized in table 2.

We estimated the age of the four clades relative to the age of *C. trachomatis* based on the ClonalFrame output by dividing the estimated ages (in coalescent units) of the nodes corresponding to the ancestors of the four clades by that of the root. The common ancestor of Clade 1 was the most recent, followed by Clade 3 and then Clade 2. Clade 4 was found to be the eldest, with an estimated age of almost a third of the age of *C. trachomatis*. A previous study estimated that the split between *C. trachomatis* and *C. pneumoniae* happened 250 Ma (Horn et al. 2004). Because the highest genetic distance between two *C. trachomatis* genomes is ~1%, which is approximately a fifth of the distance between *C. trachomatis* and *C. pneumoniae*, we deduce that the common ancestor of *C. trachomatis* existed approximately 50 Ma. Consequently, the age of each of the four clades would be between 10 and 15 Ma (table 2).

Recombination Analysis Using ClonalFrame

Genome-wide recombination has previously been reported in *C. trachomatis*, and we wanted to assess the role it played in the evolution of this intracellular bacterial pathogen. ClonalFrame estimates two values, namely π̂/θ and r/m, where the former measures the frequency of occurrence of recombination relative to mutation, whereas the latter measures how important the effect of recombination is in genetic diversification relative to mutation. ClonalFrame estimated

![Fig. 1. Phylogeny of Chlamydia trachomatis. Thirty-two C. trachomatis strains are represented in this tree. The phylogeny was constructed based on the concatenated alignment of all individual core genes by both NJ and MP procedures. The tree inferred by ClonalFrame also showed the same topology. Clade 1 is shown in yellow, Clade 2 in dark blue, Clade 3 in light blue, and Clade 4 in red.](image)
the 95% credibility interval of $\rho/\theta$ to be 0.10–0.23 (mean = 0.12), indicating that recombination happened less frequently than mutation. The 95% credibility interval of $r/m$ in this study was 0.85–1.60 (mean = 1.14), consistent with the fact that unlike mutation, recombination often affects several nucleotides at each occurrence. In a previous study based on only 12 genomes (Joseph et al. 2011), our estimates of both $\rho/\theta$ (mean = 0.07) and $r/m$ (mean = 0.71) were lower but with
overlapping credibility intervals (0.05–0.11 and 0.56–1.01, respectively). Recombination was found to affect segments with a length of 357 bp on average (95% CI [298–470]) compared with 202 bp when using only 12 genomes (Joseph et al. 2011).

We also looked into the specific role and patterns of recombination occurring within each of the four clades of C. trachomatis. The role played by recombination seems to be uneven across these four clades based on the STRUCTURE results where the strains had acquired genetic material from a different ancestral population at varying proportions. To confirm this observation, we extracted from the ClonalFrame output the number of mutation events, recombination events, and substitutions introduced by recombination for each of the four clades (table 2). Recombination was found to play a more important role relative to mutation in Clades 2 and 4 ($r/m = 1.95$ and $r/m = 0.85$, respectively) than in the niche-specific LGV Clade 1 and trachoma Clade 3 ($r/m = 0.31$ and 0.77, respectively).

We were also able to determine the genes with evidence for recombination shared across subsets of the four clades (supplementary fig. S2, Supplementary Material online). Supplementary table S1, Supplementary Material online, provides a list of genes with evidence of recombination in each of the four clades, compared across all four clades, based on the ClonalFrame analysis. The majority of recombinant genes were imported by only one clade. There were only six genes recombinant in all four clades, three of which encode hypothetical proteins of unknown function. However, the other
three genes were ompA, karG, and the serine/threonine protein kinase. ompA has been well reported to undergo recombination involving part of or the whole gene with noted changes in immune recognition, tissue tropism and persistence for different strains (Millman et al. 2001). karG has also been shown to be involved in recombination (Joseph et al. 2011) and gene switching may enhance ATP activity, an essential molecule for chlamydial metabolism.

Ten recombinant genes were shared across the non-LGV Clades 2–4. One-half of these code hypothetical proteins. The remaining genes were pbpB, which is immediately upstream of ompA and known to be involved in recombination (Gomes et al. 2007), the two autotransporter genes pmpB and pmpE of unclear functional significance, pfkA that is involved in metabolism, and yscC, which probably plays an important role in the type III secretion system as in other bacteria such as Yersinia (Diepold et al. 2010). An additional 14 genes were recombinant in both Clades 2 and 4, most of which were involved in metabolism when their function was known.

**Interclade Recombination Flux**

We assigned the origin of each recombination event identified by ClonalFrame in the four clades by postprocessing the output as described in the Materials and Methods section. Figure 5 summarizes the flow of recombination imports with an unambiguous origin that affected members of each of the clades. Clade 1 had the smallest number of resolved recombination events, with four from Clade 2 and 10 from Clade 3. Clade 2 had the highest number of resolved recombination events but their origins were imbalanced, with 152 imports from Clade 4 compared with one and eight from Clades 1 and 3, respectively. The ocular strains (Clade 3) received recombination imports roughly evenly from other clades with 41 events from Clade 1, 17 events from Clade 2, and 24 events from Clade 4. Finally, in Clade 4, 51 events were found to come from the urogenital Clade 2, 62 from the ocular Clade 3 and only 23 from Clade 1.

**Genetic Drift and Selection**

We estimated the level of genetic drift and selection occurring in this intracellular and niche-specific bacteria and correlated the results with the amount of recombination as well as the divergence age of each of the four clades. We measured the effect of genetic drift acting on these 32 C. trachomatis genomes by estimating the dN and dS substitutions for each of the 32 combinations of pair-wise orthologous genes and assessed the dN/dS ratio for each strain. Figure 6 shows a comparison between the mean pairwise dN/dS ratios estimated for each of the genomes. There were significant variations in the dN/dS ratios among the strains (ANOVA test \( P \) value = 6.85e–42; Kruskal–Wallis test \( P \) value = 4.97e–38). The overall mean of the dN/dS estimates of all the genomes was 0.4021, which was similar to a previous estimate based on four genomes of Chlamydia pneumoniae (Rocha et al. 2006). All the urogenital Clade 2 strains (except E/Swed) had lower than average dN/dS, whereas all ocular Clade 3 strains had higher than average dN/dS with Clades 1 and 4 showing no clear deviation either way.

We also used PAML to compare the likelihood of models with and without positive selection using a likelihood ratio test. We applied this test gene-by-gene and clade-by-clade in order to find clade-specific genes that are under positive selection. The highest number of genes that showed evidence for positive selection was from the urogenital clades, Clade 4 (urogenital and rectal; 53 genes) followed by ocular Clade 3 (49 genes). The Clade 1 (LGV) and Clade 2 (urogenital) showed signs of positive selection acting on 1 and 44 genes, respectively (Supplementary table S2, Supplementary Material online). The strains in Clade 4 (8 strains out of 14) and Clade 3 (all strains), which have the highest number of genes under positive selection, also showed higher genome-wide dN/dS estimates compared with the overall mean (fig. 6). Supplementary figure S3, Supplementary Material online, shows the overlap across clades in the genes under positive selection. The majority of the genes are not shared, and the findings are consistent with our previous analysis of 12 genomes (Joseph et al. 2011).

**Discussion**

We present the most comprehensive analysis of the broadest diversity of C. trachomatis genomes to date, providing new insight into the evolution of this obligate intracellular pathogen. Although there was an expected high level of sequence conservation among the genomes, recombination and selection have had a significant effect on C. trachomatis evolution and diversification. The fact that the common ancestor of this human pathogen predates the arrival of Homo sapiens, at about 195,000 years ago (McDougall et al. 2005), by millions of years suggests that C. trachomatis may have a long history of adaptation and evolution in nonhuman primates.

**A Population Structure Composed of Four Clades**

Combined analyses using PHYLIP, ClonalFrame, and STRUCTURE strongly suggest the existence of four lineages (clades) of C. trachomatis. The members of each clade were similar with regards to phenotypic characteristics. Clade 1 contained only the invasive LGV biological variants (biovars) whereas Clade 3 was comprised exclusively of trachoma strains. Clade 2 contained prevalent urogenital D, E, and F strains as well as the less prevalent Ja strains but none that infect the rectal mucosa. Clade 4 contained prevalent D strains and nonprevalent G, H, Ia, and J, including strains that cause proctitis (G/9768 and G/11074). The distribution of the strains also makes sense from a clinical perspective in that both the LGV and ocular strains reside within their own specific niche with infrequent contact with other C. trachomatis strains. LGV strains commonly cause a painless ulcer, which can go unnoticed and thus untreated (Richardson and Goldmeier 2007), that does not support the growth of other noninvasive strains requiring an intact epithelial surface.
**FIG. 5.** Circos plot of interclade recombination flux reconstructed among the four clades. The flux is represented by colors that correspond to genetic exchange among the clades. The outer ring represents the color for that clade. Each of the color-coded outer segments in the plot represents the number of fully resolved recombination events involving one of the four clades. The left side of each segment indicates the number of recombination events with respect to the particular clade (with a small segment indicating the color of the recipient clades) and the right side of the segment indicates recombination events counted from the other clades. For example, considering the Clade 1 segment (yellow), Clades 4, 3, and 2 received, respectively, 23, 41, and 1 fully resolved recombination events from Clade 1. At the same time, Clade 1 received 10 recombination imports from Clade 3 and 4 imports from Clade 2. Segments are arranged clockwise. Circos software was used to produce the plot (Altschul et al. 2009; Krzywinski et al. 2009).

**FIG. 6.** Boxplot showing the comparison of the pairwise dN/dS for each of the 32 *C. trachomatis* genomes. Notches in each box indicate the 95% confidence interval around the mean dN/dS of each genome. The color of boxes indicates the clade of the genomes: yellow, Clade 1; dark blue, Clade 2; light blue, Clade 3; and red, Clade 4. The horizontal line is the overall mean of all the mean pairwise dN/dS ratios (0.4021).
Relative Importance of Recombination between Clades

Recombination events were identified using ClonalFrame, which presents the advantage to rely on a defined evolutionary model and, therefore, to make clearly stated and testable assumptions (Didelot and Falush 2007). Recombination was found to play a greater role relative to mutation for the urogenital Clades 2 and 4 compared with the LGV Clade 1 and ocular Clade 3 (table 2). The lower r/m ratio for Clade 1 might be attributed to the fact that there were low numbers of LGV strains in our analysis. Yet, a recent study that included 20 LGV genomes supports our findings (Harris et al. 2012). Clade 2 had the highest number of imports that come from the urogenital strains in Clade 4, indicating that genetic exchange is likely enhanced between strains with the same tissue tropism. Not surprisingly, the majority of recombination events affecting Clade 4 came from Clade 2 with some also from Clade 3.

These results for the relative importance and flux of recombination between clades correlate well with what can be deduced about opportunities for recombination based on the biological properties of the clades. Rectal LGV infections offer multiple opportunities for genetic exchange with noninvasive urogenital strains, such as D, G, and J and occasionally E, F, and K, that can infect the rectal mucosa (Barnes et al. 1987; Boisvert et al. 1999; Klint et al. 2006; Dang et al. 2009; Bax et al. 2011; Quint et al. 2011; Twin et al. 2011). Indeed, rectal infections among women and men attending STD clinics are more prevalent than previously thought (9–17.5% and 9–14%, respectively) (van der Helm et al. 2009; Hunte et al. 2010). Mixed C. trachomatis infections are also high among STD populations (20–35%) (Batteiger et al. 1989; Dean et al. 1992, 1995, 2000; Brunham et al. 1994, 1996; Molano et al. 2005). For trachoma patients, urogenital strains have been found to infect the eyes as single (2.5%) or mixed (6.5%) infections (Dean et al. 2008), although the frequency of these events among different geographic populations is not fully known. Although many C. trachomatis ocular and sexually transmitted infections (STI) are cleared by antibiotics, follow-up screening shows that a substantial number of people develop re-infection, treatment failure or persistence (Dean et al. 2000; Hogan et al. 2004; Atik et al. 2006). Moreover, the high rates of asymptomatic STIs among both males (~50%) and females (~70%) provide extensive occasions for unchecked transmission and contact with multiple strains.

Absence of Absolute Barrier to Genetic Exchange

Our genomic analyses found no absolute barrier to genetic exchange among serotypes or biovars of C. trachomatis. These results are consistent with our previous analyses of recombination involving theompA gene for 27 reference and clinical strains, including all serotypes of C. trachomatis (Millman et al. 2001). A lack of barriers to recombination may seem surprising at first since replication occurs within an isolated cytoplasmic vacuole called an inclusion. However, there are at least three opportunities for genetic exchange. A number of studies have shown that more than one C. trachomatis strain can infect the same cell (Ridderhof and Barnes 1989; Rockey et al. 2002). Although individual inclusions form for each strain, most fuse into a larger inclusion where the genomic contents are essentially pooled. In addition, many strains, primarily G, D, K, F, and E (in decreasing order), form fibers that extend from the primary inclusion to other cells, forming secondary inclusions (Suchland et al. 2005). The replicating reticulate bodies of the organism have been shown to be transported to these secondary inclusions. This provides an additional opportunity for mixing if another strain has infected the cell that contains the secondary inclusion. Finally, the EB in many Chlamydia species has been shown to transport bacteriophages into the cell during infection (Hisao, Ohayon, et al. 2000; Hsia, Ting, et al. 2000; Everson et al. 2002). There is also evidence that these bacteriophages have been incorporated into the genomes of C. pneumoniae and C. caviae (Hsia, Ting, et al. 2000; Read et al. 2003) and perhaps other species as well. These findings are important because a similar mechanism may allow DNA from a prior C. trachomatis infection to hitchhike its way into the cell along with the EB or to be taken up via traditional transformation (Dubnau 1999).

Length of Recombined Fragments

We also evaluated the size of the genomic segments involved in recombination. The segments involved a length of 357 bp on average which is similar to what has been identified in other natural bacterial populations (Didelot and Maiden 2010), although considerably smaller than what was observed in vitro experiments with C. trachomatis (Demars et al. 2007; Demars and Weinfurter 2008). It is possible that large genomic rearrangements occur in the host but are unstable and thus remain undetected (Joseph et al. 2011). This argument would be consistent with the differences observed between natural and in vitro finding for other bacteria such as Escherichia coli (Touchon et al. 2009).

Natural Selection in C. trachomatis as a Whole

As bacteria diversify by point mutation and homologous recombination, most nucleotide changes make the organism less competitive and are destined to be removed from the population. However, this purging process does not happen instantaneously, and this allows slightly deleterious mutations to remain in the genomes at early stages of diversification. Kuo et al. (2009) showed that an increased level of genetic drift, resulting from reduced effective population size (N_e) and/or genome-wide relaxation of selection, can result in an increased incidence of slightly deleterious amino acid replacements and consequently an increase in the genome-wide dN/dS ratio. However, Castillo-Ramirez et al. (2011) examined recently emerged clones of methicillin resistant Staphylococcus aureus (MRSA) and Clostridium difficile and noted a high proportion of synonymous substitutions (reduced dN/dS ratio) in genes affected by recombination. In this study, the overall mean of the dN/dS estimates for the 32 genomes was 0.4021, which is high in comparison with free-living bacteria (<0.06); obligate pathogens were in the
range of 0.04–1.2 (Kuo et al. 2009). Previously, two separate studies (Jordan et al. 2002; Rocha et al. 2006) also observed high dN/dS ratios when closely related genomes of Chlamydia pneumoniae were compared. The opportunities for genetic exchange discussed above may provide a high effect from recombination with reduced drift as for the more broadly tissue infecting urogenital strains, which would be more similar to free-living bacteria, but an increased effect of genetic drift from niche specialization and a small population size as for the trachoma strains (discussed later). These effects may explain the overall higher dN/dS mean estimate for Chlamydia compared with free-living bacteria such as Escherichia and Bacillus species that reside in advantageous environments (e.g., gastrointestinal tract and soil, respectively) with larger pools of like species with the increased opportunity for higher rates of recombination and acquisition of synonymous substitutions (Rocha et al. 2006).

Variations in the Effect of Selection between Clades

We observed statistically significant differences in dN/dS ratios across all four clades, which might indicate clade-specific levels of evolutionary forces. The urogenital Clades 4 and 2 had the highest and third highest number of genes, respectively, with evidence for positive selection. The higher dN/dS estimates for the ocular strains (fig. 6) could be an indication of an increased effect of genetic drift due to niche specialization, which in turn might be due to the reduced effective population size. This would affect all genes equally, explaining why this clade had the highest level of dN/dS overall and yet only a relatively modest number of genes (49) with evidence of positive selection. The urogenital Clade 2 had the lowest dN/dS ratio (fig. 6) indicating reduced effect of genetic drift along with the highest extent of recombination (table 2), resulting in increased synonymous mutations, which is in agreement with the results of Castillo-Ramirez et al. (2011). The high dN/dS ratio only in the L2b strain combined with the lowest rate of recombination in Clade 1 may be due to the increased effect of genetic drift acting on this recently emerged L2b strain that has undergone a clonal expansion in Europe.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org).

Acknowledgments

The authors thank Brian Desany and Mark Driscoll at Roche/454 for access to the GS Junior sequencer and help with the data. This work was supported in part by Public Health Service grant from the National Institutes of Health R01 AI059647 to D.D. and National Science Foundation grant 2009-65109-05760 to D.D.

References


Quint KD, Bom RJ, Quint WG, Bruisten SM, van der Looef MF, Morre SA, de Vries HJ. 2011. Anal infections with concomitant Chlamydia trachomatis genotypes among men who have sex with men in Amsterdam, the Netherlands. BMC Infect Dis. 11:63.


