Population Genomics of *Chlamydia trachomatis*: Insights on Drift, Selection, Recombination, and Population Structure

Sandeep J. Joseph,[1] Xavier Didelot,[2] James Rothschild,[3] Henry J.C. de Vries,[4,5,6] Servaas A. Morré,[7,8] Timothy D. Read,[†,1,9] and Deborah Dean*[†,3,10,11]

[1]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine

[2]Department of Infectious Disease Epidemiology, Imperial College, London, United Kingdom

[3]Center for Immunobiology and Vaccine Development, Children's Hospital Oakland Research Institute, California

[4]Department of Dermatology, Academic Medical Centre, University of Amsterdam, The Netherlands

[5]STI Outpatient Clinic, Infectious Diseases Cluster, Public Health Service Amsterdam, The Netherlands

[6]Centre for Infection and Immunity, Academic Medical Centre, University of Amsterdam, The Netherlands

[7]Department of Pathology, Laboratory of Immunogenetics, VU University Medical Center, Amsterdam, The Netherlands

[8]Department 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

[9]Department of Human Genetics, Emory University School of Medicine

[10]Department of Medicine, University of California, San Francisco

[11]Joint Graduate Program in Bioengineering, University of California, San Francisco and Berkeley

[†]These authors contributed equally to this work.

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

Associate editor: Daniel Falush

Research article

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 (d*N*/d*S*) 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 d*N*/d*S* 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

18th century (Schwartz 1997). An expanding number of LGV causing serovars (L_{1-3}, L_2a, L_2b, and L_2c) have been discovered in the last 50 years (Spaargaren et al. 2005; Van der Bij et al. 2006; Somboonna et al. 2011). These strains can invade the basal layers of the epidermis and disseminate via regional lymphatics to inguinal lymph nodes. They, therefore, represent a biological variant (biovar) distinct from the noninvasive urogenital (D-K, Da, Ia, and Ja) and ocular (A, B, Ba, and C) serovars of the organism. Not until the mid 20th century was the organism cultured and recognized as a major global cause of STDs. Over 100 million cases of *C. trachomatis* STDs are estimated to occur annually (World Health Organization 2011).

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; Hafne 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 glycogen 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 L_2c, 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 to maintain stock cultures of the clonal strain. *C. trachomatis* elementary bodies (EB) were purified by density gradient centrifugation followed by DNase treatment to remove contaminating human DNA and subsequent genomic DNA purification using the High Pure PCR template preparation kit (Roche Diagnostics, Indianapolis, IN) as previously described (Somboonna et al. 2008, 2011; Dean et al. 2009).

Genome Sequencing

Genomes were sequenced using GS-FLX as well as GS-Junior (454 Life Sequencing Inc., Branford, CT). Libraries for sequencing were prepared from 1 to 5 μg of genomic DNA. The sequencing reads for each strain were assembled de novo using the Newbler program (Margulies et al. 2005) with default parameters. Because of the small genome size and nonrepetitive nature of the *C. trachomatis* genome as well as longer read lengths from the 454 sequencing, all the assemblies produced only a small number of contigs. The contigs were aligned against the *C. trachomatis* reference D/UW3/CX genome sequence using MUMmer (Kurtz et al. 2004) to create concatenated ordered "pseudocontigs." Although there are genome sequences of other reference strains available (table 1), D/UW3/CX has most commonly been used for these types of genome comparisons. These pseudocontigs were considered as the bacterial chromosome for each strain and were annotated using the ISGA bacterial annotation pipeline (Hemmerich et al. 2010). All the published genomes were also re-annotated using the same pipeline. The genome and plasmid sequences have been deposited at NCBI under the accession numbers given in table 1.

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

Strain Name	Strain Notation	Serotype	Country	Anatomic Source	Year Isolated	Accession No.	Genome Publication
A/HAR-13[a]	A	A	Egypt	Conjunctiva	1958	CP000051	Carlson et al. (2004)
B/TZ1A828/OT	B/Tz	B	Tanzania	Ocular	1998	FM872308	Seth-Smith et al. (2009)
B/Jali20	B/Jali	B	Gambia	Ocular	1985	FM872307	Seth-Smith et al. (2009)
C/TW-3/OT[a]	C	C	Taiwan	Conjunctiva	1959	SRA051538.1	This publication
D/UW3/CX[a]	D	D	USA (Seattle)	Cervix	1965	AE001273	Stephens et al. (1998)
D/2932	D/2932	D	USA (Seattle)	Cervix	NA	ACFJ01000001	Jeffrey et al. (2010)
D/84s	D/84s	D	USA (SF Bay Area)	Cervix	2000s	SRA051539.1	This publication
D/2s	D/2s	D	USA (SF Bay Area)	Cervix	2000s	SRA051544.1	This publication
D/43nl	D/43nl	D	The Netherlands	Cervix	2000s	SRA051526.1	This publication
D_EC	D/EC	D	USA (Montana)	Genital tract	2010	CP002054	Sturdevant et al. (2010)
D_LC	D/LC	D	USA (Montana)	Genital tract	2010	CP002052	Sturdevant et al. (2010)
E/SW2	E/Swed	E	Sweden	Urethra	2001	FN652779	Unemo et al. (2010)
E/11023	E/11023	E	USA (Seattle)	Cervix	NA	CP001890	Jeffrey et al. (2010)
E/150	E/150	E	USA (Seattle)	Rectum	NA	CP001886	Jeffrey et al. (2010)
E/5 s	E/5s	E	USA (SF Bay Area)	Cervix	2000s	SRA051547.1	This publication
F/1	F/1	F	USA (SF Bay Area)	Cervix	2000s	SRA051574.1	This publication
F/70	F/70	F	USA (Seattle)	Cervix	NA	ABYF01000001	Jeffrey et al. (2010)
F/38nl	F/38nl	F	The Netherlands	Cervix	2000s	SRA051469.2	This publication
G/UW-57[a]	G	G	USA (Seattle)	Cervix	1971	SRA051545.1	This publication
G/9301	G/9301	G	USA (Seattle)	Urethra	NA	CP001930	Jeffrey et al. (2010)
G/9768	G/9768	G	USA (Seattle)	Rectum	NA	CP001887	Jeffrey et al. (2010)
G/11222	G/11222	G	USA (Seattle)	Cervix	NA	CP001888	Jeffrey et al. (2010)
G/11074	G/11074	G	USA (Seattle)	Rectum	NA	CP001889	Jeffrey et al. (2010)
H/UW-4/CX[a]	H	H	USA (Seattle)	Cervix	1965	SRA051548.1	This publication
H/18 s	H/18s	H	USA (SF Bay Area)	Cervix	2000s	SRA051541.1	This publication
Ia/UW-202[a]	Ia	Ia	USA (Seattle)	Cervix	1985	SRA051537.1	This publication
J/6276	J/6276	J	USA (Seattle)	Cervix	NA	ABYD01000001	Jeffrey et al. (2010)
Ja/47nl	Ja/47nl	Ja	The Netherlands	Cervix	2000s	SRA051542.1	This publication
Ja/26 s	Ja/26s	Ja	USA (SF Bay Area)	Cervix	2000s	SRA051540.1	This publication
L_2/434/BU[a]	L_2	L_2	USA (California)	Lymph node	1968	AM884176	Thomson et al. (2008)
L_2b/UCH-1	L_2b	L_2b	UK	Rectum	1968	AM884177	Thomson et al. (2008)
L_2C	L_2C	L_2C	USA (SF Bay Area)	Rectum	2000s	NC_015744	Somboonna et al. (2011)

NOTE.—NA, not available.

[a]Reference strains are defined as strains that were originally isolated in the 1950s or 1960s and have been laboratory adapted over the last decades.

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 of the data for 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

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 N_e 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 θ and ρ are the scaled rates of mutation and recombination estimated by ClonalFrame, and μ 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 (d*N*/d*S*) Calculations

To calculate the nonsynonymous (d*N*) and synonymous (d*S*) 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 d*N*/d*S* ratios (Rocha et al. 2006). For each pair of strains, we estimated the median value of d*N*/d*S* and further disentangled the contribution of each strain to the pairwise d*N*/d*S* 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 Ia (urogenital), and two from serotype Ja (urogenital). Genomes for serotypes C, H, and Ja 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 L_2, L_2b, and L_2c (LGV strains); Clade 2 (dark blue) contained all the E, F, Ja, 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).

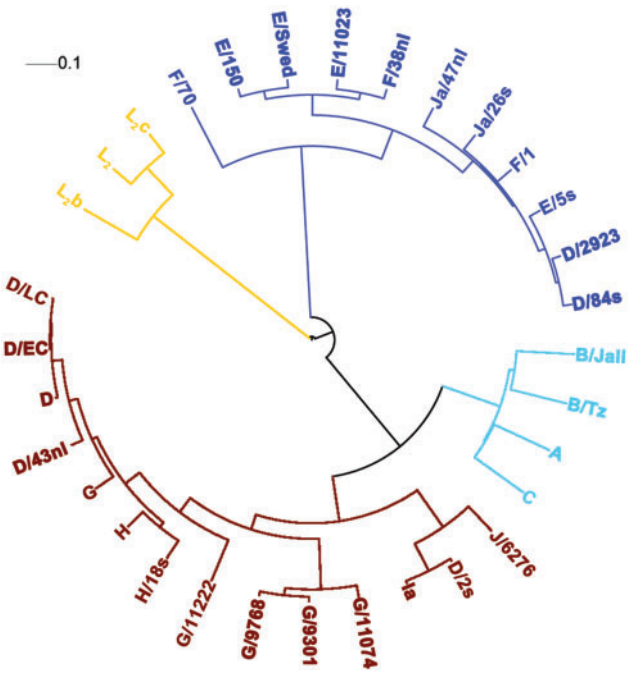

Fig. 1. Phylogeny of *Chlamydia trachomatis*. Thirty-two C. *trachomatis* strains are represented in this tree. The phylogeny was constructed based on the concatenated alignment of all individual core genes by both NJ and MP procedures. The tree inferred by ClonalFrame also showed the same topology. Clade 1 is shown in yellow, Clade 2 in dark blue, Clade 3 in light blue, and Clade 4 in red.

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 ρ/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

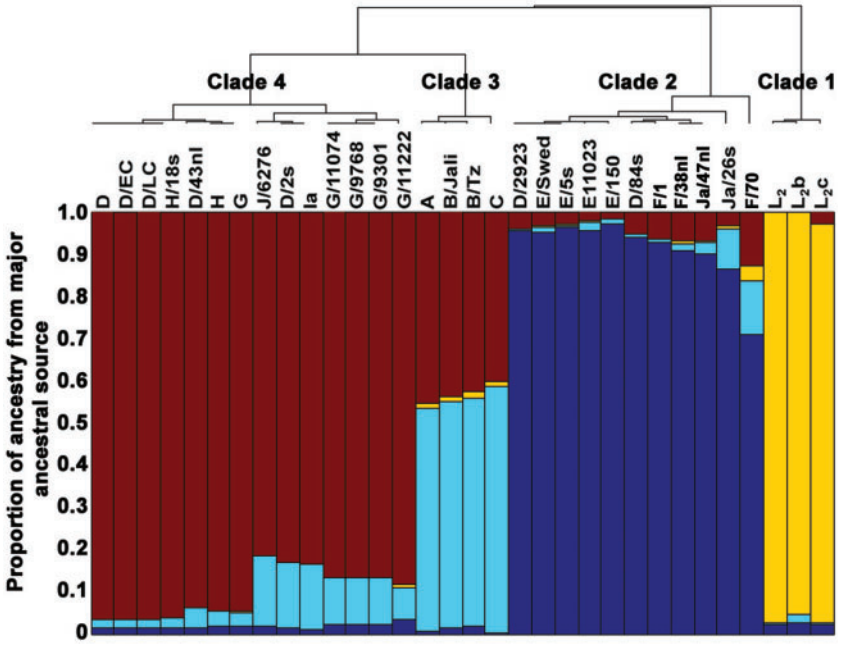

Fig. 2. Bar graph from the application of the linkage model in STRUCTURE to the whole genome data for 32 *C. trachomatis* strains. Each vertical line represents one of the 32 strains. The *y* axis shows the proportion of ancestry from each of the four ancestral populations (colored). Single letters on the top *x* axis are reference strains, including L_2. The tree at the top represents a complete linkage clustering of the strains based on their proportions of ancestry from each population.

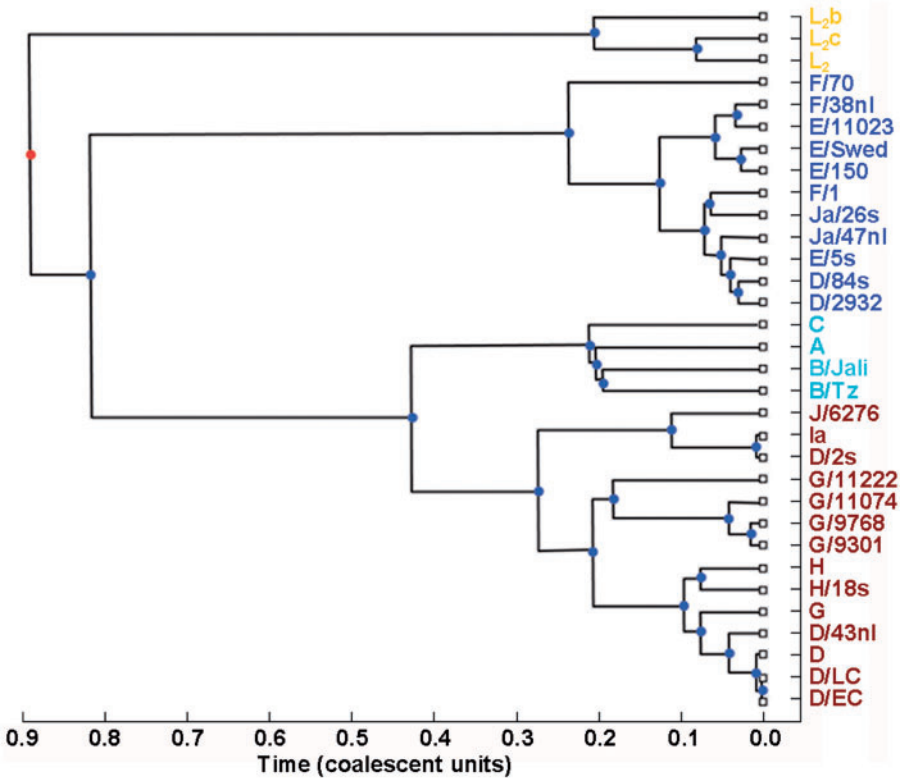

Fig. 3. Clonal geneology inferred by ClonalFrame using whole genome alignment data for the 32 *C. trachomatis* strains. The branch lengths are shown in coalescent units.

the 95% credibility interval of ρ/θ 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 ρ/θ (mean = 0.07) and r/m (mean = 0.71) were lower but with

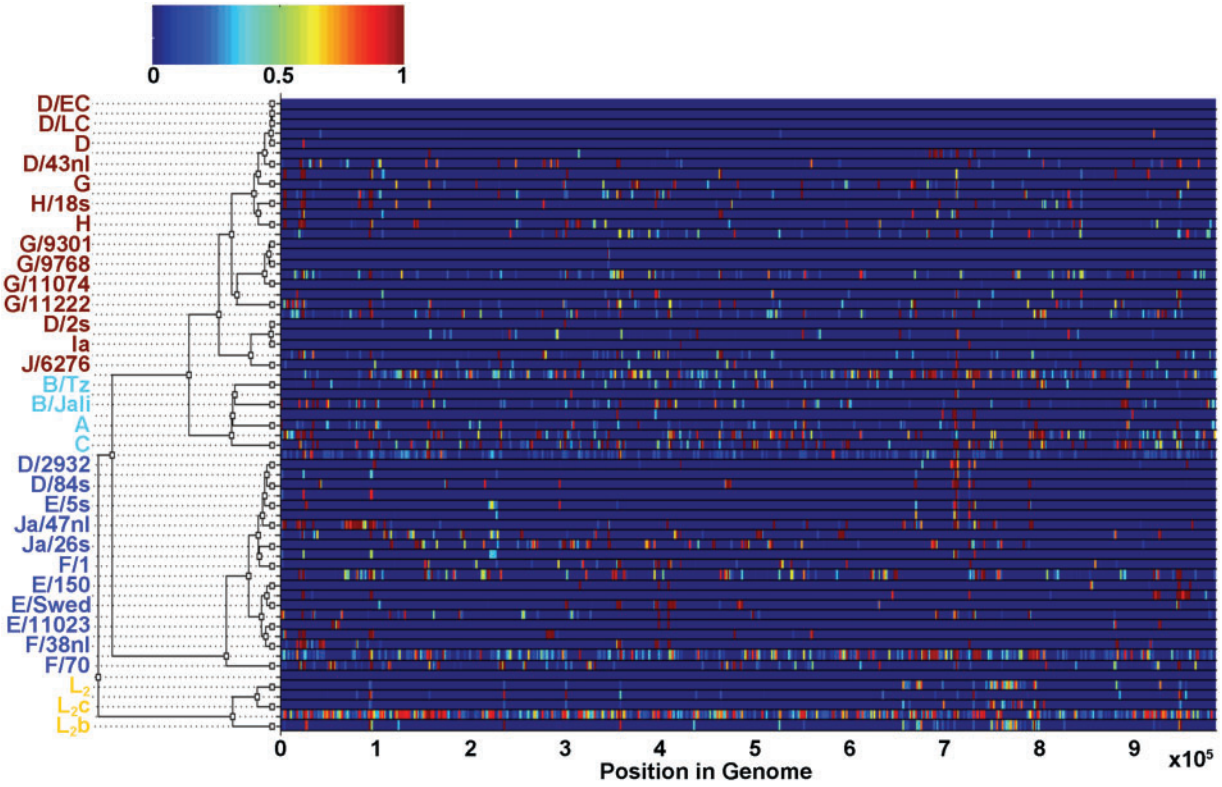

Fig. 4. Results of the ClonalFrame analysis on an alignment of the 32 *C. trachomatis* genomes. The inferred clonal genealogy is shown on the left. Each branch of the tree corresponds to a row of the heat map, which is horizontally aligned with it. Each row of the heat map shows the posterior probability of recombination estimated by ClonalFrame on the corresponding branch (*y* axis) and along the positions of the alignment (*x* axis; $\times 10^5$ bp).

Table 2. Results from ClonalFrame Analysis.

	Clade 1	Clade 2	Clade 3	Clade 4	All
Genomes	3	11	4	14	32
Mutation events	284	1,845	1,713	2,788	6,630
Recombination events	20	392	135	216	763
Substitutions introduced by recombination	89	3,602	1,328	2,370	7,389
Relative rate of mutation and recombination (ρ/θ)	0.07042254	0.21246612	0.07880911	0.07747489	0.11508296
Relative effect of mutation and recombination (r/m)	0.31338028	1.95230352	0.7752481	0.85007174	1.11447964
Age in coalescent units	0.27	0.21	0.23	0.2	0.88
Age relative to age of ancestor of all genomes	0.31	0.24	0.26	0.23	1
Age (My)	15.34	11.93	13.07	11.36	50

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 d*N* and d*S* substitutions for each of the 32 combinations of pair-wise orthologous genes and assessed the d*N*/d*S* ratio for each strain. Figure 6 shows a comparison between the mean pairwise d*N*/d*S* ratios estimated for each of the genomes. There were significant variations in the d*N*/d*S* ratios among the strains (ANOVA test P value = 6.85e−42; Kruskal–Wallis test P value = 4.97e−38). The overall mean of the d*N*/d*S* 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 d*N*/d*S*, whereas all ocular Clade 3 strains had higher than average d*N*/d*S* 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 d*N*/d*S* 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.

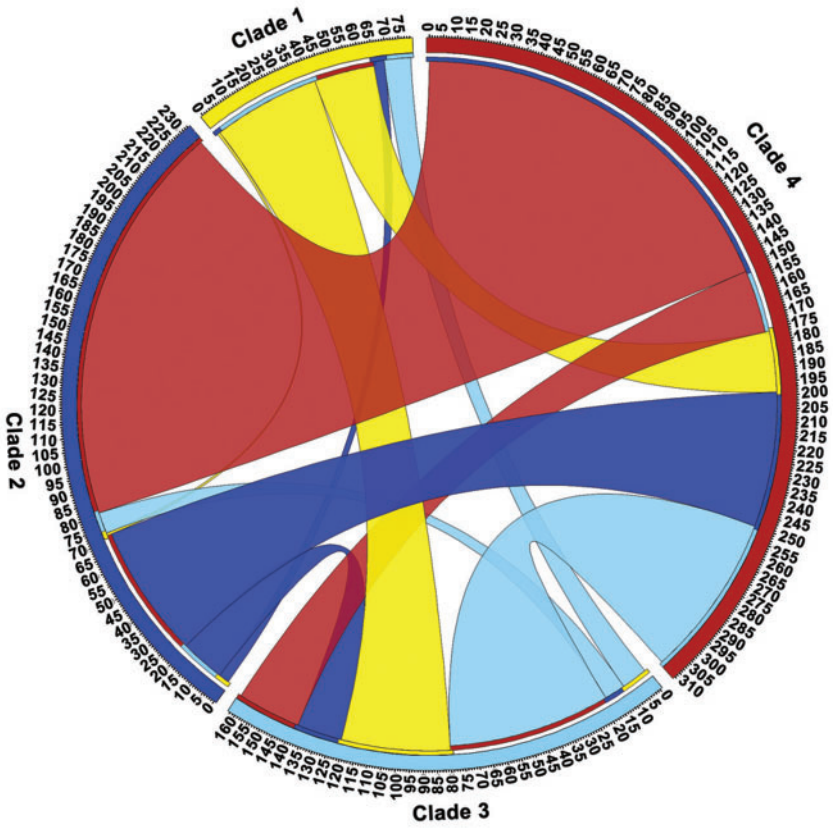

Fig. 5. Circos plot of interclade recombination flux reconstructed among the four clades. The flux is represented by colors that correspond to genetic exchange among the clades. The outer ring represents the color for that clade. Each of the color-coded outer segments in the plot represents the number 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).

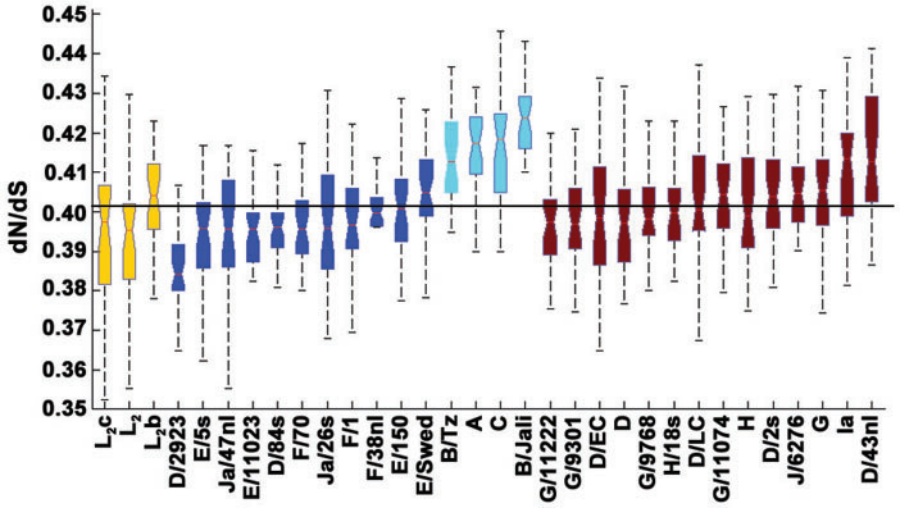

Fig. 6. Boxplot showing the comparison of the pairwise d*N*/d*S* for each of the 32 *C. trachomatis* genomes. Notches in each box indicate the 95% confidence interval around the mean d*N*/d*S* of each genome. The color of boxes indicates the clade of the genomes: yellow, Clade 1; dark blue, Clade 2; light blue, Clade 3; and red, Clade 4. The horizontal line is the overall mean of all the mean pairwise d*N*/d*S* 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 the *ompA* 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 (Hsia, 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 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 pneumonia* 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 effect 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 L_2b 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 L_2b 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

Altschul SF, Gertz EM, Agarwala R, Schaffer AA, Yu YK. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res.* 37:815–824.

Alzhanov D, Barnes J, Hruby DE, Rockey DD. 2004. Chlamydial development is blocked in host cells transfected with *Chlamydophila caviae* incA. *BMC Microbiol.* 4:24.

Atik B, Thanh TT, Luong VQ, Lagree S, Dean D. 2006. Impact of annual targeted treatment on infectious trachoma and susceptibility to reinfection. *JAMA.* 296:1488–1497.

Barnes RC, Rompalo AM, Stamm WE. 1987. Comparison of *Chlamydia trachomatis* serovars causing rectal and cervical infections. *J Infect Dis.* 156:953–958.

Batteiger BE, Lennington W, Newhall WJ, Katz BP, Morrison HT, Jones RB. 1989. Correlation of infecting serovar and local inflammation in genital chlamydial infections. *J Infect Dis.* 160:332–336.

Bax CJ, Quint KD, Peters RP, et al. (14 co-authors). 2011. Analyses of multiple-site and concurrent *Chlamydia trachomatis* serovar infections, and serovar tissue tropism for urogenital versus rectal specimens in male and female patients. *Sex Transm Infect.* 87:503–507.

Belland R, Ojcius DM, Byrne GI. 2004. Chlamydia. *Nat Rev Microbiol.* 2:530–531.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol.* 57:289–300.

Boisvert JF, Koutsky LA, Suchland RJ, Stamm WE. 1999. Clinical features of *Chlamydia trachomatis* rectal infection by serovar among homosexually active men. *Sex Transm Dis.* 26:392–398.

Brunham R, Yang C, Maclean I, Kimani J, Maitha G, Plummer F. 1994. *Chlamydia trachomatis* from individuals in a sexually transmitted diseases core group exhibit frequent sequence variation in the major outer membrane protein (*omp1*) gene. *J Clin Invest.* 94:458–463.

Brunham RC, Kimani J, Bwayo J, et al. (11 co-authors). 1996. The epidemiology of *Chlamydia trachomatis* within a sexually transmitted diseases core group. *J Infect Dis.* 173:950–956.

Brunham RC, Rey-Ladino J. 2005. Immunology of *Chlamydia* infection: implications for a *Chlamydia trachomatis* vaccine. *Nat Rev Immunol.* 5:149–161.

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

Castillo-Ramirez S, Harris SR, Holden MT, He M, Parkhill J, Bentley SD, Feil EJ. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog.* 7:e1002129.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Cortes C, Rzomp KA, Tvinnereim A, Scidmore MA, Wizel B. 2007. *Chlamydia pneumoniae* inclusion membrane protein Cpn0585 interacts with multiple Rab GTPases. *Infect Immun.* 75:5586–5596.

Dang T, Jaton-Ogay K, Flepp M, et al. (12 co-authors). 2009. High prevalence of anorectal chlamydial infection in HIV-infected men who have sex with men in Switzerland. *Clin Infect Dis.* 49:1532–1535.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14:1394–1403.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One.* 5:e11147.

Dean D. 2010. Pathogenesis of chlamydial ocular infections. In: Tasman W, Jaeger EA, editors. Duane's foundations of clinical

ophthalmology. Philadelphia (PA): Lippincott Williams & Wilkins. p. 678–702.

Dean D, Bruno WJ, Wan R, et al. (11 co-authors). 2009. Predicting phenotype and emerging strains among *Chlamydia trachomatis* infections. *Emerg Infect Dis.* 15:1385–1394.

Dean D, Kandel RP, Adhikari HK, Hessel T. 2008. Multiple Chlamydiaceae species in trachoma: implications for disease pathogenesis and control. *PLoS Med.* 5:e14.

Dean D, Oudens E, Bolan G, Padian N, Schachter J. 1995. Major outer membrane protein variants of *Chlamydia trachomatis* are associated with severe upper genital tract infections and histopathology in San Francisco. *J Infect Dis.* 172:1013–1022.

Dean D, Schachter J, Dawson CR, Stephens RS. 1992. Comparison of the major outer membrane protein variant sequence regions of B/Ba isolates: a molecular epidemiologic approach to *Chlamydia trachomatis* infections. *J Infect Dis.* 166:383–392.

Dean D, Suchland RJ, Stamm WE. 2000. Evidence for long-term cervical persistence of *Chlamydia trachomatis* by *omp1* genotyping. *J Infect Dis.* 182:909–916.

Delevoye C, Nilges M, Dautry-Varsat A, Subtil A. 2004. Conservation of the biochemical properties of IncA from *Chlamydia trachomatis* and *Chlamydia caviae*: oligomerization of IncA mediates interaction between facing membranes. *J Biol Chem.* 279: 46896–46906.

Demars R, Weinfurter J. 2008. Interstrain gene transfer in *Chlamydia trachomatis* in vitro: mechanism and significance. *J Bacteriol.* 190: 1605–1614.

Demars R, Weinfurter J, Guex E, Lin J, Potucek Y. 2007. Lateral gene transfer in vitro in the intracellular pathogen *Chlamydia trachomatis*. *J Bacteriol.* 189:991–1003.

Didelot X, Bowden R, Street T, et al. (11 co-authors). 2011. Recombination and population structure in *Salmonella enterica*. *PLoS Genet.* 7:e1002191.

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

Didelot X, Maiden MC. 2010. Impact of recombination on bacterial evolution. *Trends Microbiol.* 18:315–322.

Diepold A, Amstutz M, Abel S, Sorg I, Jenal U, Cornelis GR. 2010. Deciphering the assembly of the Yersinia type III secretion injectisome. *EMBO J.* 29:1928–1940.

Dubnau D. 1999. DNA uptake in bacteria. *Annu Rev Microbiol.* 53: 217–244.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32: 1792–1797.

Editorial—*Nat Rev Microbiol*. 2005. STIs: not just for women. *Nat Rev Microbiol.* 3:94.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* 14:2611–2620.

Everson JS, Garner SA, Fane B, Liu BL, Lambden PR, Clarke IN. 2002. Biological properties and cell tropism of Chp2, a bacteriophage of the obligate intracellular bacterium *Chlamydophila abortus*. *J Bacteriol.* 184:2748–2754.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (Version 3.2). *Cladistics.* 5:164–166.

Fields KA, Mead DJ, Dooley CA, Hackstadt T. 2003. *Chlamydia trachomatis* type III secretion: evidence for a functional apparatus during early-cycle development. *Mol Microbiol.* 48:671–683.

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

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

Hafner L, Beagley K, Timms P. 2008. *Chlamydia trachomatis* infection: host immune responses and potential vaccines. *Mucosal Immunol.* 1:116–130.

Hafner LM, McNeilly C. 2008. Vaccines for *Chlamydia* infections of the female genital tract. *Future Microbiol.* 3:67–77.

Harris SR, Clarke IN, Seth-Smith HM, et al. (24 co-authors). 2012. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nat Genet.* 44:413–419.

Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q. 2010. An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics* 26:1122–1124.

Hogan RJ, Mathews SA, Mukhopadhyay S, Summersgill JT, Timms P. 2004. Chlamydial persistence: beyond the biphasic paradigm. *Infect Immun.* 72:1843–1855.

Horn M, Collingro A, Schmitz-Esser S, et al. (13 co-authors). 2004. Illuminating the evolutionary history of chlamydiae. *Science* 304: 728–730.

Hsia R, Ohayon H, Gounon P, Dautry-Varsat A, Bavoil PM. 2000. Phage infection of the obligate intracellular bacterium, *Chlamydia psittaci* strain guinea pig inclusion conjunctivitis. *Microbes Infect.* 2: 761–772.

Hsia RC, Ting LM, Bavoil PM. 2000. Microvirus of *Chlamydia psittaci* strain guinea pig inclusion conjunctivitis: isolation and molecular characterization. *Microbiology* 146:1651–1660.

Hunte T, Alcaide M, Castro J. 2010. Rectal infections with chlamydia and gonorrhoea in women attending a multiethnic sexually transmitted diseases urban clinic. *Int J STD AIDS.* 21:819–822.

Jeffrey BM, Suchland RJ, Quinn KL, Davidson JR, Stamm WE, Rockey DD. 2010. Genome sequencing of recent clinical *Chlamydia trachomatis* strains identifies loci associated with tissue tropism and regions of apparent recombination. *Infect Immun.* 78:2544–2553.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Microevolutionary genomics of bacteria. *Theor Popul Biol.* 61:435–447.

Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD. 2011. Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. *Biol Direct.* 6:28.

Joseph SJ, Read TD. 2012. Genome-wide recombination in *Chlamydia trachomatis*. *Nat Genet.* 44:364–366.

Klint M, Lofdahl M, Ek C, Airell A, Berglund T, Herrmann B. 2006. Lymphogranuloma venereum prevalence in Sweden among men who have sex with men and characterization of *Chlamydia trachomatis ompA* genotypes. *J Clin Microbiol.* 44: 4066–4071.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19:1639–1645.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19: 1450–1454.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–2189.

Li Z, Chen C, Chen D, Wu Y, Zhong Y, Zhong G. 2008. Characterization of fifty putative inclusion membrane proteins encoded in the *Chlamydia trachomatis* genome. *Infect Immun.* 76:2746–2757.

McDougall I, Brown FH, Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 433: 733–6.

Margulies M, Egholm M, Altman WE, et al. (56 co-authors). 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.

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

Molano M, Meijer CJ, Weiderpass E, Arslan A, Posso H, Franceschi S, Ronderos M, Munoz N, van den Brule AJ. 2005. The natural course of *Chlamydia trachomatis* infection in asymptomatic Colombian women: a 5-year follow-up study. *J Infect Dis.* 191:907–916.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. *Proc Natl Acad Sci U S A.* 96:12638–12643.

Pritchard JK, Falush D. 2009. Documentation for structure software [Internet]. Version 2.3 Chicago: University of Chicago [cited 2012 Aug 16]. Available from: http://pritch.bsd.uchicago.edu/structure.html.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Quint KD, Bom RJ, Quint WG, Bruisten SM, van der Loeff 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.

Read TD, Myers GS, Brunham RC, et al. (21 co-authors). 2003. Genome sequence of *Chlamydophila caviae* (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the *Chlamydiaceae*. *Nucleic Acids Res.* 31:2134–2147.

Richardson D, Goldmeier D. 2007. Lymphogranuloma venereum: an emerging cause of proctitis in men who have sex with men. *Int J STD AIDS.* 18:11–14, quiz 15.

Ridderhof JC, Barnes RC. 1989. Fusion of inclusions following superinfection of HeLa cells by two serovars of *Chlamydia trachomatis*. *Infect Immun.* 57:3189–3193.

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol.* 239:226–235.

Rockey DD, Viratyosin W, Bannantine JP, Suchland RJ, Stamm WE. 2002. Diversity within *inc* genes of clinical *Chlamydia trachomatis* variant isolates that occupy non-fusogenic inclusions. *Microbiology* 148: 2497–2505.

Schwartz DA. 1997. *Lymphogranuloma Venereum*. In: Connor DH, Schwartz DA, Chandler FW, editors. Pathology of infectious diseases. Stamford (CT): Appleton and Lange Publishers. p. 491–507.

Seth-Smith HM, Harris SR, Persson K, et al. (20 co-authors). 2009. Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain. *BMC Genomics* 10:239.

Sisko JL, Spaeth K, Kumar Y, Valdivia RH. 2006. Multifunctional analysis of *Chlamydia*-specific genes in a yeast expression system. *Mol Microbiol.* 60:51–66.

Somboonna N, Mead S, Liu J, Dean D. 2008. Discovering and differentiating new and emerging clonal populations of *Chlamydia trachomatis* with a novel shotgun cell culture harvest assay. *Emerg Infect Dis.* 14:445–453.

Somboonna N, Wan R, Ojcius DM, Pettengill MA, Joseph SJ, Chang A, Hsu R, Read TD, Dean D. 2011. Hypervirulent *Chlamydia trachomatis* clinical strain is a recombinant between lymphogranuloma venereum (L(2)) and D lineages. *MBio.* 2: e00045–00011.

Spaargaren J, Schachter J, Moncada J, de Vries HJ, Fennema HS, Pena AS, Coutinho RA, Morre SA. 2005. Slow epidemic of lymphogranuloma venereum L2b strain. *Emerg Infect Dis.* 11:1787–1788.

Stephens RS, Kalman S, Lammel C, et al. (12 co-authors). 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282:754–759.

Storey JD. 2002. A direct approach to false discovery rates. *J R Stat Soc Ser B Stat Methodol.* 64:479–498.

Sturdevant GL, Kari L, Gardner DJ, et al. (11 co-authors). 2010. Frameshift mutations in a single novel virulence factor alter the in vivo pathogenicity of *Chlamydia trachomatis* for the female murine genital tract. *Infect Immun.* 78:3660–3668.

Suchland RJ, Rockey DD, Weeks SK, Alzhanov DT, Stamm WE. 2005. Development of secondary inclusions in cells infected by *Chlamydia trachomatis*. *Infect Immun.* 73:3954–3962.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Thomson NR, Clarke IN. 2010. *Chlamydia trachomatis*: small genome, big challenges. *Future Microbiol.* 5:555–561.

Thomson NR, Holden MT, Carder C, et al. (17 co-authors). 2008. *Chlamydia trachomatis*: genome sequence analysis of lymphogranuloma venereum isolates. *Genome Res.* 18:161–171.

Touchon M, Hoede C, Tenaillon O, et al. (41 co-authors). 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5:e1000344.

Twin J, Moore EE, Garland SM, Stevens MP, Fairley CK, Donovan B, Rawlinson W, Tabrizi SN. 2011. *Chlamydia trachomatis* genotypes among men who have sex with men in Australia. *Sex Transm Dis.* 38: 279–285.

Unemo M, Seth-Smith HM, Cutcliffe LT, et al. (14 co-authors). 2010. The Swedish new variant of *Chlamydia trachomatis*: genome sequence, morphology, cell tropism and phenotypic characterization. *Microbiology* 156:1394–1404.

Van der Bij AK, Spaargaren J, Morre SA, Fennema HS, Mindel A, Coutinho RA, de Vries HJ. 2006. Diagnostic and clinical implications of anorectal lymphogranuloma venereum in men who have sex with men: a retrospective case-control study. *Clin Infect Dis.* 42: 186–194.

van der Helm JJ, Hoebe CJ, van Rooijen MS, Brouwers EE, Fennema HS, Thiesbrummel HF, Dukers-Muijrers NH. 2009. High performance and acceptability of self-collected rectal swabs for diagnosis of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men who have sex with men and women. *Sex Transm Dis.* 36: 493–497.

Wang Y, Kahane S, Cutcliffe LT, Skilton RJ, Lambden PR, Clarke IN. 2011. Development of a transformation aystem for *Chlamydia trachomatis*: restoration of glycogen biosynthesis by acquisition of a plasmid shuttle vector. *PLoS Pathog.* 7: e1002258.

World Health Organization. 2011. Initiative for vaccine research. Geneva: WHO [cited 2012 Aug 16]. Available from: http://www.who.int/vaccine_research/diseases/soa_std/en/index1.html.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.