Hypervirulent *Chlamydia trachomatis* Clinical Strain Is a Recombinant between Lymphogranuloma Venereum (L$_2$) and D Lineages

Naraporn Somboonna,[a]* Raymond Wan,[a] David M. Ojcius,[b] Matthew A. Pettengill,[b] Sandeep J. Joseph,[c] Alexander Chang,[d] Ray Hsu,[a] Timothy D. Read,[c,d] and Deborah Dean[a,e,f]

Center for Immunobiology and Vaccine Development, Children's Hospital Oakland, Research Institute, Oakland, California, USA[a]; Health Sciences Research Institute and School of Natural Sciences, University of California, Merced, Merced, California, USA[b]; Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, Georgia, USA[c]; Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA[d]; Department of Medicine, University of California, San Francisco, San Francisco, California, USA[e]; and Joint Graduate Program in Bioengineering, University of California, Berkeley, Berkeley, California, USA[f]

* Present address: Department of Microbiology, Faculty of Science, Chulalongkorn University, Bangkok, Thailand.

ABSTRACT *Chlamydia trachomatis* is an obligate intracellular bacterium that causes a diversity of severe and debilitating diseases worldwide. Sporadic and ongoing outbreaks of lymphogranuloma venereum (LGV) strains among men who have sex with men (MSM) support the need for research on virulence factors associated with these organisms. Previous analyses have been limited to single genes or genomes of laboratory-adapted reference strain L$_2$/434 and outbreak strain L$_2$b/UCH-1/proctitis. We characterized an unusual LGV strain, termed L$_2$c, isolated from an MSM with severe hemorrhagic proctitis. L$_2$c developed nonfusing, grape-like inclusions and a cytotoxic phenotype in culture, unlike the LGV strains described to date. Deep genome sequencing revealed that L$_2$c was a recombinant of L$_2$ and D strains with conserved clustered regions of genetic exchange, including a 78-kb region and a partial, yet functional, toxin gene that was lost with prolonged culture. Indels (insertions/deletions) were discovered in an *ftsK* gene promoter and in the *tarp* and *hctB* genes, which encode key proteins involved in replication, inclusion formation, and histone H1-like protein activity, respectively. Analyses suggest that these indels affect gene and/or protein function, supporting the *in vitro* and disease phenotypes. While recombination has been known to occur for *C. trachomatis* based on gene sequence analyses, we provide the first whole-genome evidence for recombination between a virulent, invasive LGV strain and a noninvasive common urogenital strain. Given the lack of a genetic system for producing stable *C. trachomatis* mutants, identifying naturally occurring recombinants can clarify gene function and provide opportunities for discovering avenues for genomic manipulation.

IMPORTANCE Lymphogranuloma venereum (LGV) is a prevalent and debilitating sexually transmitted disease in developing countries, although there are significant ongoing outbreaks in Australia, Europe, and the United States among men who have sex with men (MSM). Relatively little is known about LGV virulence factors, and only two LGV genomes have been sequenced to date. We isolated an LGV strain from an MSM with severe hemorrhagic proctitis that was morphologically unique in tissue culture compared with other LGV strains. Bioinformatic and statistical analyses identified the strain as a recombinant of L$_2$ and D strains with highly conserved clustered regions of genetic exchange. The unique culture morphology and, more importantly, disease phenotype could be traced to the genes involved in recombination. The findings have implications for bacterial species evolution and, in the case of ongoing LGV outbreaks, suggest that recombination is a mechanism for strain emergence that results in significant disease pathology.

Address correspondence to Deborah Dean, ddean@chori.org.

*C*hlamydia trachomatis is responsible for a broad spectrum of diseases in males and females of all age groups worldwide. *C. trachomatis* is the leading cause of preventable blindness in tropical developing countries and the leading global cause of bacterial sexually transmitted diseases (STDs) (1). Over 92 million individuals are infected with *C. trachomatis* in the urogenital tract annually (1). In the United States alone, more than 1 million cases are reported each year, although the actual rates are estimated to be 2.8 million due to the deficiencies in screening and reporting

(2). The Centers for Disease Control and Prevention (CDC) have reported that these infections result in an annual cost to Americans of over $10 billion.

The complications of urogenital infection include tubal factor infertility, ectopic pregnancy, and chronic pelvic pain. The lymphogranuloma venereum (LGV) strains of *C. trachomatis* are considered biological variants of the organism and cause invasive disease such as genital ulcers, hemorrhagic proctitis, rectal fistulae, and suppurative lymphadenitis (3). LGV is prevalent in develop-

ing countries and represents a neglected tropical disease. To date, five different strains of LGV have been identified: L_1 to L_3, L_2a, and L_2b.

Infection is initiated at the mucosal epithelia by host cell endocytosis of the infectious but metabolically inert elementary body (EB). Following cell entry, the EB expands into the replicative form termed the reticulate body (RB) within a nonacidified inclusion. Over 30 *C. trachomatis* proteins are secreted into the membrane of the inclusion, including a family of inclusion proteins of which IncA is the best described (4, 5). IncA is thought to be critical for fusion of independent inclusions that form when more than one EB enters the same host cell.

The mechanism for ascension to upper genital tract tissues remains ill described. While LGV strains have not been identified in endometrial or fallopian tube tissue, they are known to cause ulceration, invade the basal layers, and travel via the lymphatic system to regional draining lymph nodes (3). There have been sporadic and ongoing outbreaks of LGV among men who have sex with men (MSM) in many developed countries worldwide where, historically, the rates of LGV have been extremely low. The first reported outbreak occurred in Rotterdam, Netherlands, in 2003 (6). Additional outbreaks were reported to occur in the United States, Europe, and Australia (7). The majority of these infected men had engaged in high-risk behavior, were HIV positive, and presented with painful hemorrhagic proctitis, discharge, and, in some cases, constipation. There was a surprising absence of genital ulcers and the inguinal syndrome.

In 2005, a novel LGV strain, termed L_2b, which was associated with both symptomatic and asymptomatic disease, was identified (8); the genome was subsequently sequenced (9). While a similar strain based on *ompA* genotyping was detected among isolates dating back to the 1980s in San Francisco, CA (10), the degree of genome homology between the two is not known. L_2b strains have also recently been reported to occur in urethritis cases (11). These cumulative findings, especially the association of L_2b with different disease presentations, and the availability of only two LGV genomes ($L_2/434$ and L_2b/UCH-1/proctitis [9]), suggest that further analyses of the morphological, molecular, and genetic characteristics of clinical LGV strains are warranted to identify virulence factors and understand the microbial nature of the outbreaks. The purpose of the current study was to advance these objectives by characterizing the cellular biology and genomics of an LGV isolate from an HIV-negative MSM who had severe hemorrhagic proctitis.

We isolated a unique LGV strain from the MSM that was discovered to be a recombinant of invasive L_2 and noninvasive D strains. We present the first computational and statistical whole-genome evidence for recombination in this pathogen. Our findings have implications for bacterial evolution and, in the case of ongoing LGV outbreaks, suggest that recombination is a mechanism for strain emergence that results in significant disease pathology. While we do not know the molecular clock for these events, the results suggest either that the recombinant was produced in the rectal mucosa of our patient prior to sampling or that a clonal population of L_2 and D recombinants had already emerged and was circulating among the patient's core sexual group. Methods for direct genome sequencing of the organisms present in various tissues will be required to fully ascertain the diversity and recombinant nature of *C. trachomatis* strain infections and their associated disease phenotypes.

RESULTS

**Morphological characteristics of reference and clinical LGV strains.** We compared the morphological characteristics of reference strains D/UW3, $L_1/440$, $L_2/434$, L_2a/TW-396, and $L_3/404$, clinical strain L_2b, and a clinical isolate from the rectal mucosa of an MSM from this study, termed L_2c, in culture using HeLa229 cells and confocal fluorescence microscopy (see Materials and Methods). Reference strains refer to *C. trachomatis* organisms that were isolated many decades ago, have been propagated since then, and are considered laboratory adapted. At 24 h, $L_2/434$ (Fig. 1A) and $L_1/440$, L_2a/TW-396, $L_3/404$, and L_2b (not shown) produced large inclusions with IncA-containing fibers extending to secondary inclusions (Fig. 1A, arrowheads). D/UW3 had smaller inclusions than $L_2/434$ at 24 h but a similar-sized inclusion at 48 h (Fig. 1B), with IncA staining at all time points. In contrast, L_2c formed multiple smaller inclusions that did not fuse (Fig. 1C, arrows); there was no evidence for IncA staining, IncA-containing fibers, or fusion of inclusions during the course of development at 12, 18, or 24 h for L_2c (Fig. 1C and 1D) (data not shown for 12 and 18 h). We found that *incA* mRNA was expressed for strains $L_2/434$, D/UW3, and L_2c (Fig. 1E), despite the absence of IncA staining for L_2c.

L_2c was more difficult to grow and had 20% fewer infected and surviving cells than LGV strains at 24 and 36 h with the use of the same multiplicity of infection (MOI). The growth curve (Fig. 1F) showed peak 16S rRNA relative expression at 24 h, with a taper to 36 and 48 h for L_2. The peak was also at 24 h for L_2c but was much lower than for L_2, while the peak for D/UW3 was at 48 h and lower than for $L_2/434$ but higher than for L_2c.

Cytotoxic characteristics of reference LGV, reference D/UW3, and clinical L_2c strains. Because of the difference in morphological characteristics between LGV and L_2c strains and previous reports of nonfusing strains being less virulent (12), we hypothesized that other LGV strains would be more cytotoxic than L_2c. We also assayed the D/UW3 strain given the probability that the partial toxin gene was acquired from a D strain (described below). The cytopathic effects of each strain were determined by infecting monolayers of HeLa cells with each strain independently using a toxicity assay (see Materials and Methods). Infections were visualized by light microscopy after 4 h. Surprisingly, the cellular effects of $L_2/434$ showed far less cytotoxicity (Fig. 2A) than those of L_2c (Fig. 2B) or D/UW3 (Fig. 2C), the latter two of which had significantly higher numbers of cells that underwent rounding, detachment, and lysis (score, 3+) than $L_2/434$ (100% versus <10%; $P < 0.01$ for both). By Western blot anlaysis, we show toxin protein production for D/UW3 and L_2c but not for L_2 (Fig. 2G).

Comparative genomic analyses. Genomic DNA was extracted from L_2c grown in HeLa cells after plaque purification and propagation (13) (see Materials and Methods). To verify clonal purity, *ompA* and MLST genes (14) for 10 clones were Sanger sequenced following amplification by PCR and cloning using a TOPO TA kit (Invitrogen, Carlsbad, CA) (see Materials and Methods). The sequences of the eight genes were identical for each clone.

Genome sequencing was performed using 454 pyrosequencing (15). The genome sequences for the L_2c chromosome (1,038,313 nucleotides [nt]) and plasmid (7,499 nt) were obtained following assembly and *in silico* closure (see Materials and Methods). The genome was annotated by the Integrative Services for Genomics
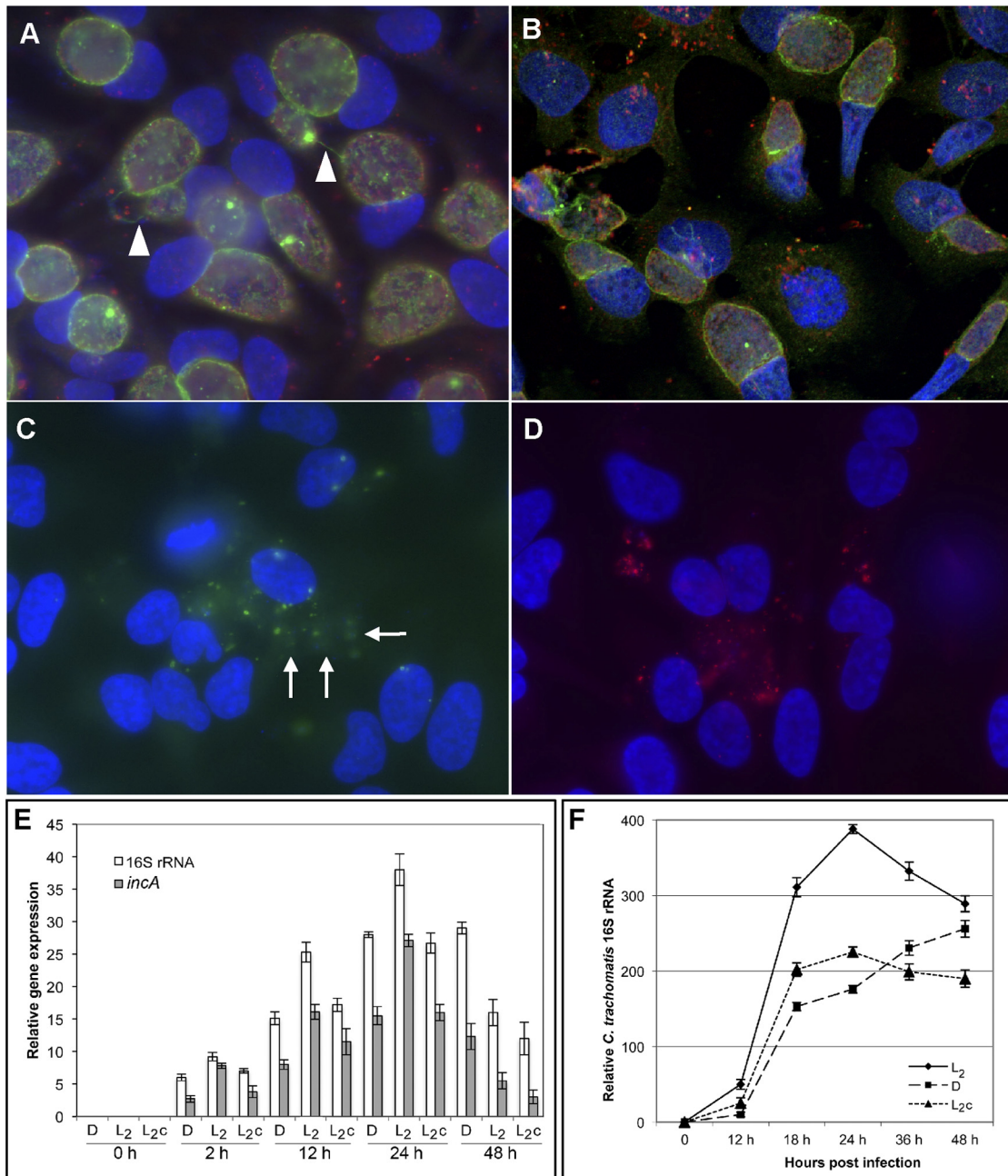
FIG 1 IncA staining, *incA* mRNA expression, and inclusion morphology of D/UW3-, L_2-, and L_2c-infected HeLa229 cells. Cells were infected at an MOI of 1.0 and stained at 24 or 48 h as noted postinfection (see Materials and Methods). (A) L_2/434. Green, IncA-specific monoclonal antibody (fluorescein-conjugated secondary antibody); red, *C. trachomatis*-specific LPS monoclonal antibody (Cy3-conjugated secondary antibody); blue, DAPI (DNA). Arrowheads, IncA-stained fibers to secondary inclusions. (B) D/UW3. Evidence for IncA at 48 h; staining as in (A). (C) There was no observed IncA staining for L_2c. Arrows denote multiple inclusions for a cell. Green, *C. trachomatis*-specific heat shock protein 60 monoclonal (fluorescein secondary antibody); red, IncA-specific monoclonal (Cy3 secondary antibody); blue, DAPI. (D) L_2c, same staining as for panel A, with no evidence of IncA. (E) *incA* mRNA expression. Cells were infected with L_2/434, D/UW3, or L_2c at an MOI of 5, and total RNA was extracted at the designated time points. The graph represents the relative *incA* gene expression based on quantitative real-time PCR of *incA* and 16S rRNA. Values represent the mean ± standard deviation (SD) based on three independent experiments (see Materials and Methods). (F) Growth curve. Cells were infected with L_2/434, D/UW3, or L_2c at an MOI of 1.0. Total RNA was harvested at designated time points for quantitative PCR of *C. trachomatis* 16S rRNA. GAPDH was used for normalization. Values represent the mean ± SD based on three independent experiments (see Materials and Methods).

Analysis pipeline (16). There were 1,005 putative protein-coding genes. While the overall sequence was similar to that of L_2/434, we noticed unusual patterns of localized variation. Using an in-house program called Q-plotGenome (see Materials and Methods) and

BLAST score ratio analysis (17), we identified a recombinant region of 78 kb and other smaller regions that were similar in sequence to D/UW3 (Fig. 3 and Table 1, region 1; see also Fig. S1 in the supplemental material). Phylogenetic reconstructions (Fig. 4)
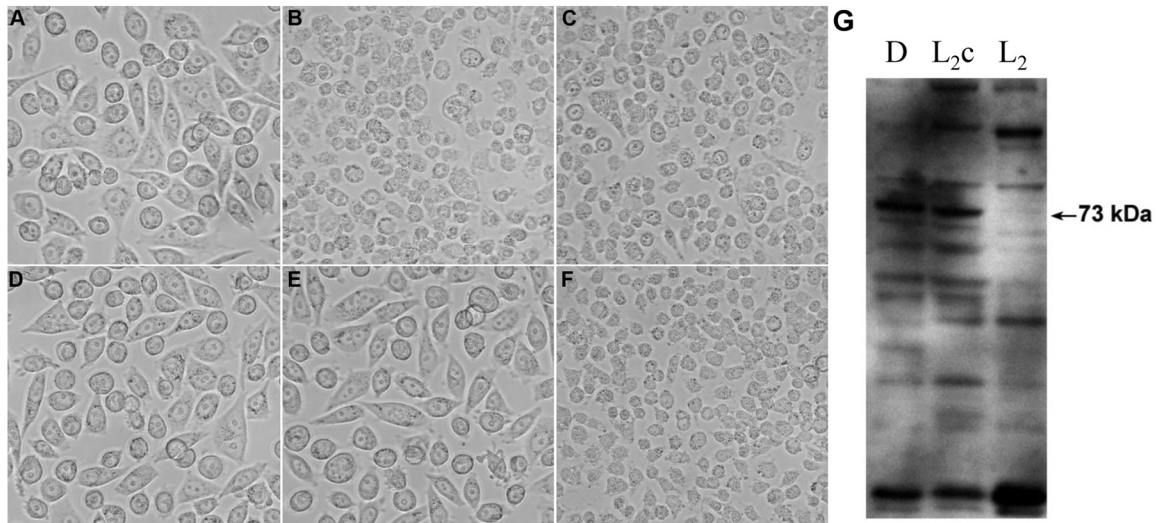
FIG 2 Optical phase-contrast microscopy demonstrating different cytopathic effects and protein expression of $L_2/434$-, D/UW3-, and L_2c-infected HeLa229 cells. Cells were infected with an MOI of 100 and imaged at 4 hours postinfection. Experiments were performed in triplicate (see Materials and Methods). (A) There were no cytopathic effects for $L_2/434$ (original magnification, ×400). (B) A significantly higher number of cells with rounding, detachment, and lysis were observed for L_2c (original magnification, ×400). (C) The findings for D/UW3 were similar to those for L_2c as shown in panel B (original magnification, ×400). After prolonged passage (see Materials and Methods), the cytopathic effects of L_2c were lost (E) in comparison to those for D/UW3 (F). The morphology now resembled that of $L_2/434$ (D). (G) Toxin protein expression demonstrated by Western blot analysis. D/UW3, L_2c, and $L_2/434$ were grown in HeLa cells at an MOI of 100, purified, run on a precast 4 to 12% gel, transferred to a PVDF membrane, and reacted with polyclonal antiserum against recombinant CT166 (see Materials and Methods). Lanes represent purified EBs. Both D/UW3 and L_2c showed strong reactivity to a protein at ~73 kDa that is close to the predicted 74.8 kDa size of CT166; there was no protein expression for $L_2/434$.

confirmed the recombinant regions to have originated recently from an ancestor similar to D/UW3, while the rest of the chromosome and plasmid were closely related to L_2 sequences. Using Q-plotGenome, we identified indels for the cytotoxin region, the *tarp* and *hctB* genes, and the intergenic region (IGR) upstream of the *ftsK* gene (Fig. S1).

A puzzling feature of the recombinant regions was a persistent background of reads with bases similar to those of $L_2/434$. We plotted the number of nucleotide variants at each of the 7,982 single nucleotide polymorphisms (SNPs) between D/UW3 and

$L_2/434$ (see Materials and Methods). At 6,527 (82%) of the known SNP positions, all nucleotide reads had the $L_2/434$ variant base rather than that of D/UW3. At the other 18% of the known SNP positions, D-like variants were either approximately 50% or significantly above zero (Fig. 3, red bars, and Table 1, regions 2, 3, 5, and 6). It is probable that these SNPs were introduced through gene conversion of localized segments with a D-like ancestor. The Circos plot highlights candidate recombinant regions of L_2c that match to D/UW3 or $L_2/434$ sequences (Fig. 5). A further 235 variant positions in two regions were not known variants between
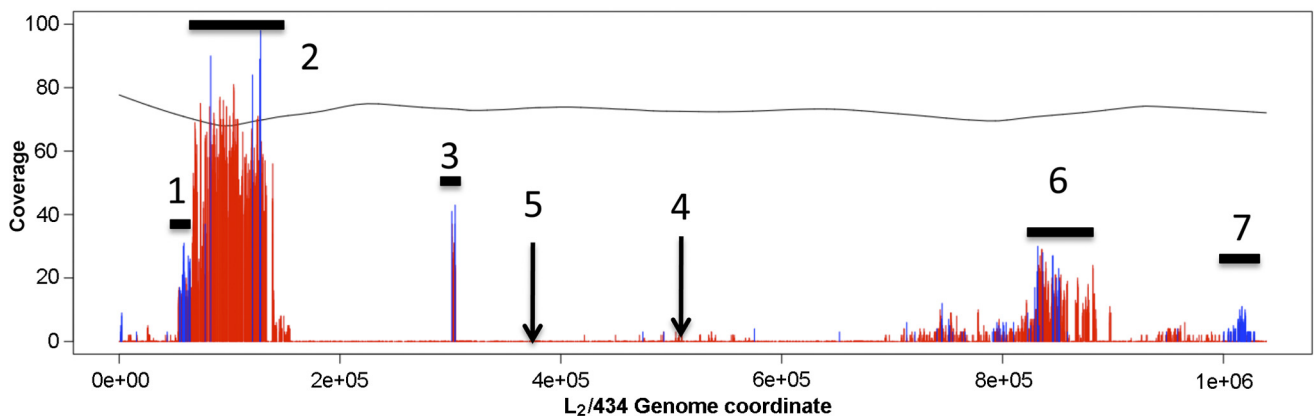
FIG 3 Plot of the sequence read variation within the L_2c chromosome. The x axis shows the coordinates of the L_2c consensus sequence; the y axis shows the level of 454 read coverage, expressed as the number of reads per base; the gray line is a smoothed curve of overall read redundancy (produced using the *R* "lowess" function, with the *f* smoothing parameter set to 0.2). Red bars indicate the number of variant reads that match D/UW3 at the 7,982 known SNP positions between D/UW3 and $L_2/434$. Blue bars indicate all variant positions within the assembly that do not coincide with D/UW3 SNP but that may represent an outlying strain within the D cluster. The horizontal bars represent the recombination regions defined in Table 1, including the location of the chromosomal partial toxin insertion (region 4) and the *hctB* deletion (region 5).

TABLE 1 Seven putative recombinant regions in clinical strain L$_2$c

Region	L$_2$c coordinates[a]	Gene(s)	Polymorphism level (%)	Strain of origin	Notes
1	53700–65500	CTL$_2$c_349–475 (*sufB*)	15–25	Unknown	
2	65500–139200	CTL$_2$c_475-1002	70–90	D	Includes 16S rRNA operon, *tsK*, and IGR upstream of *ftsK*
3	301100–304500	CTL$_2$c_838 (*glgB*)	30–50	D	Deletion specific to L$_2$ in membrane thiol protease
4	510400	Intergenic (CT166)[b]	0[b]	D	Toxin gene insertion
5	373606–374061	498 (CT046)[c]	100	NA[d]	Deletion in *hctB*
6	829000–855400	CTL$_2$c_920 (*omcB-secD*)	30–50	D	Includes CT456/CTL$_2$c_721 (*tarp* gene)
7	1019400–1019700	CTL$_2$c_125	15	Unknown	Homolog of TTSS secreted protein CT621

[a] Approximated to the nearest round hundred.
[b] Although the genome sequence did not contain the partial cytotoxin gene, the initial plaque-purified isolate (prior to further propagation) did contain the gene.
[c] The CT designation refers to the gene locus for D/UW3 genome.
[d] NA, not applicable; there was no source strain because this was an intrinsic deletion.

L$_2$/434 and D/UW3 (Fig. 3, blue bars, and Table 1, regions 1 and 7) and may have arisen through recombination with a strain significantly different from D/UW3 or with a D strain that has variant sequences in those regions. However, comparison of L$_2$c with another D strain that was recently genome sequenced [D(s)/2923; GenBank accession no. NZ_ACFJ00000000] (18) showed similar results. The distribution of SNPs was almost identical to that for D/UW3 (see Fig. S2 in the supplemental material). In fact, fewer D(s)/2923 SNPs match (7,565 SNPs, versus 7,982 for D/UW3).

The likely explanation for these results is that an L$_2$ strain acquired DNA segments from a D strain and either one other *C. trachomatis* strain or a D strain variant through gene conversion in the patient. The patterns of SNP discovery could not have been produced by accidental mixing of strain D and L$_2$ genomic DNA in the laboratory before genome sequencing. If that were the case, all D-like SNPs would have had the same frequency in the chromosome and plasmid. Instead, more than 80% of the SNPs were not seen at all, and 18% of the variants were found at frequencies of greater than 15%. Furthermore, the extent of recombination and conservation of the recombinant regions would have had to have occurred in relatively few culture generations along with elimination of any contaminating D or other donor strains, since donor strains were not seen even with high-redundancy sequencing (75-fold). Also, all SNPs between L$_2$c and L$_2$/434 were within the boundaries of the seven discrete recombinant regions (Table 1). Outside these regions, there was no intrinsic genetic variation in the genome backbones of L$_2$c and L$_2$/434-LGV strains isolated in California more than 30 years apart (10). Finally, all 10 plaque-purified clones screened by *ompA* and typed by multilocus sequence typing (MLST) as L$_2$ had no minor contaminating base peaks.

Analyses of recombinant regions. As described above, we identified seven putative recombinant regions within the L$_2$c genome (Table 1). The regions were independently analyzed and confirmed using Sanger sequencing (see Materials and Methods).

Region 1 contained SNPs in ~20% of the reads that match a mixture of D-like and non-D-like variants, suggesting that the DNA source was not a D strain. Comparison of L$_2$c with D(s)/2923 also did not reveal a match in this region. However, since only two D strains [reference D/UW3 and clinical D(s)/2923] have been genome sequenced to date and the distributions of SNPs are almost identical to each other, it is possible that other strains within the D cluster may match this region.

In region 2, from the middle of the *sufB* gene, the proportion of reads with SNPs rose to 70 to 90% for variants entirely D-like, continuing for 78 kb and including a 16S rRNA operon. The *ftsK* gene, a complement gene involved in cytokinesis, was in this region, reading from the 3′-to-5′ side of the genome origin of replication. The IGR sequence upstream of *ftsK* was annotated using the Bacterial PROMoter and BDGP programs; two putative promoters were identified at alignment coordinates 45 to 168 and 310 to 484, respectively, upstream of the start codon (see Fig. S3 in the supplemental material). Both predictions had significant scores. A 33-nt deletion in the promoter at nt 406 was present for L$_2$c, D/UW3 and ocular strains compared to what was observed for L$_2$/434 and L$_2$b/UCH-1/proctitis (Fig. S3), although the significance of these findings in terms of gene regulation is unknown.

Region 3 was a small recombinant region centered on the *glgB* gene, part of the glycogen biosynthetic pathway.

Region 4 contained a large insertion of 1,915 nt representing a partial toxin gene (CT166 in D/UW3) at the 5′ end not present in any LGV strains (Fig. 6). ClustalW alignment revealed that the region of CT165 to CT168 was almost identical to that of D/UW3. *In vitro* experiments revealed expression of the toxin with a cytotoxic phenotype in HeLa cells for L$_2$c (Fig. 2B) but not for the other LGV strains (Fig. 2A) (see Materials and Methods). After 20 passages of L$_2$c in tissue culture, the toxin insertion was lost. The cytotoxicity assay was repeated for the passaged isolate, and there was no observed cytopathic effect for L$_2$c (Fig. 2E) (see Materials and Methods) or L$_2$ (Fig. 2D); D/UW3 maintained a cytotoxic effect (Fig. 2F). However, other genes that could have contributed to the change in phenotype may have also been lost.

Region 5 contained a deletion of 216 nt (72 amino acids) located in a region of the *hctB* gene, which encodes a histone H1-like protein, Hc2, and differentiates ocular, noninvasive, and LGV disease groups from one another (19, 20) (see Fig. S4 in the supplemental material). Compared to LGV strains, D/UW3 and D(s)/2923 have a 20-amino-acid indel, while ocular strains have a 56-amino-acid deletion (Fig. S5). The basic amino acids arginine and lysine (Fig. S5, blue and purple letters), as well as pentapeptide

A

B
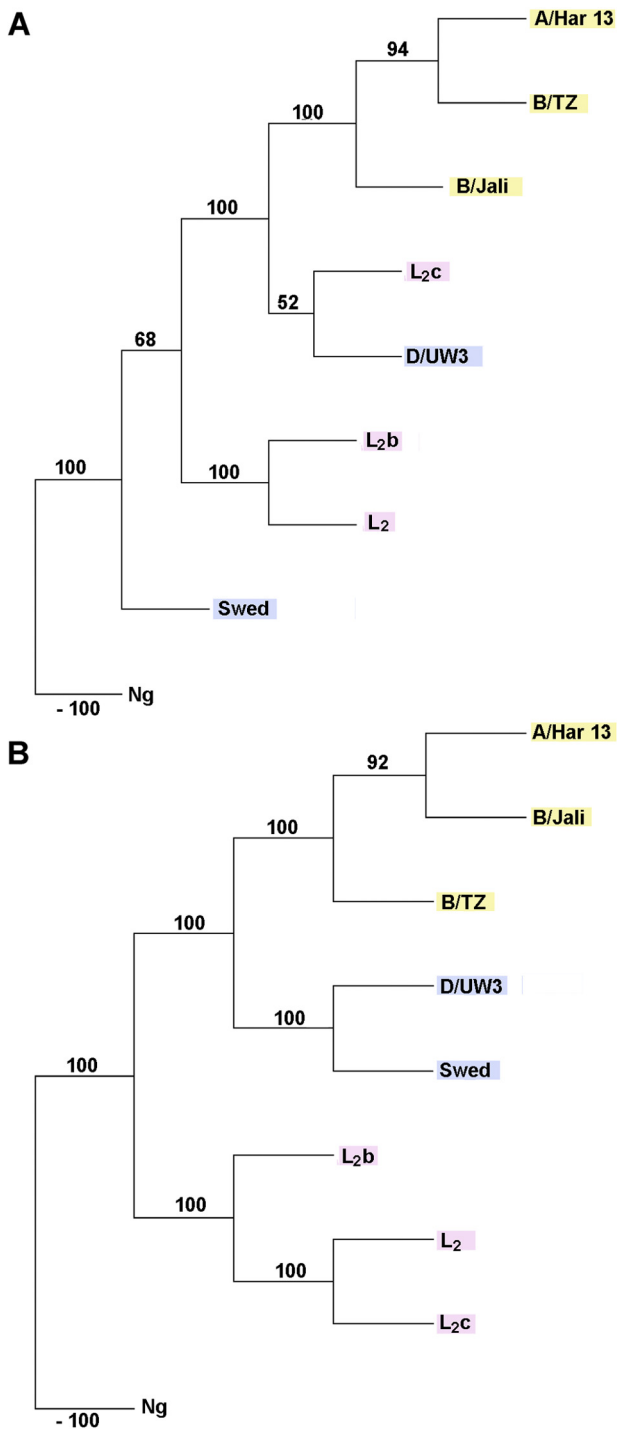
FIG 4 Phylogenetic comparison of the 75-kb recombinant region and a random 75-kb region of the L₂c chromosome. The trees for the major recombinant region (coordinates 65K to 140K) (A) and a random region of the chromosome of the same 75-kb size (coordinates 500K to 575K) (B) are shown. Colors denote disease grouping: yellow, trachoma; pink, invasive urogenital disease; blue, noninvasive urogenital disease. The trees were calculated from MUSCLE alignments using the PHYLIP programs DNApars, Seqboot, and Consense; 1,000 bootstrap replicates were performed to determine the significance of the support for each node in the tree (see Materials and Methods). The *Chlamydia muridarum* Nigg (Ng) strain was used as the outgroup (GenBank accession no. AE002160). All other *C. trachomatis* strains are described in the text except for the Sweden strain (Swed; GenBank accession no. FN652779).
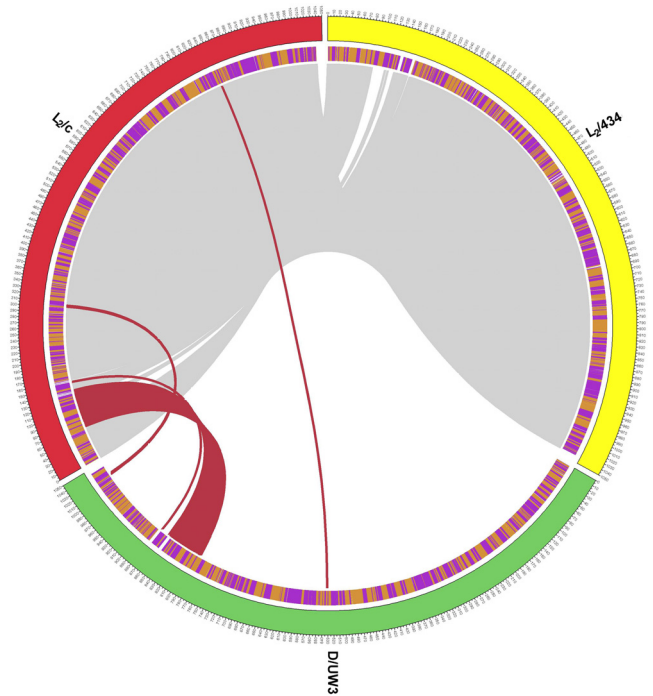

FIG 5 Circos plot mapping of the consensus L₂c sequence against L₂/434 and D/UW3. The figure highlights candidate recombinant regions of L₂c that match to D/UW3 or L₂/434. Mapping was performed using 100-nt windows; sequence comparisons used BLAST score ratio (see Materials and Methods). Note that the coordinates of the D/UW3 genome follow the original GenBank accession number and do not start at the origin of replication as for L₂/434 and L₂b/UCH-1/proctitis. Fragments of L₂c that have a stronger match to L₂/434 are shown in grey; those in red have a stronger match to D/UW3. The inner track with purple and orange bands represents protein-coding genes on the negative strand (purple) and positive strand (orange). The figure was produced using Circos software, a tool for graphical representation of genome data (58).

motifs (e.g., TAARK, VAAKK, and TVAKR), are highly repetitive (Fig. S5, red stars, denoting separation of pentapeptides) and consist primarily of three aliphatic residues followed by two basic residues.

Region 6 contained two in-frame deletions in the *tarp* gene, which encodes the translocated actin-recruiting phosphoprotein TARP (21), at residues 379 to 687 and at residues 1084 to 1239, based on the alignment with L₂/434 (see Fig. S6 in the supplemental material). Both deletions are similar to D/UW3 and represent areas of sequence divergence from LGV strains (Fig. S6). The deletion might have occurred from a crossover event, as suggested by SimPlot informative sites and phylogenetic and bootscan analyses (Fig. 7A to D) (see Materials and Methods). With the use of SimPlot, recombination breakpoints were located at residues 990 to 991 and 1218 to 1219, equivalent to residues 996 to 997 and 1689 to 1690 when gaps are included (Fig. 7B and C; Fig. S6).

The deletions represent an area of tyrosine-rich repeats in Tarp (21). While L₂/434 and L₂b/UCH-1/proctitis contain six tandems of tyrosine-rich repeat regions, L₂c, A/2497, B/HAR36, C/TW3, and D/UW3 contain three partial tandems of tyrosine-rich regions (see Fig. S7, purple letters, tyrosine in bold, in the supplemental material). Other polymorphisms include indels, SNPs, and C-terminal repeat regions in B/HAR36, C/TW3, and D/UW3 that differentiate these strains from L₂/434, L₂b/UCH-1/proctitis,
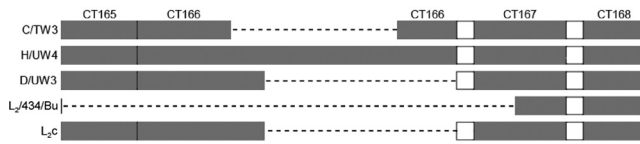
FIG 6 Schematic alignment of the toxin loci for clinical strain L$_2$c and reference strains. Reference strains C/TW-3, H/UW4, D/UW3, and L$_2$/434 and clinical strain L$_2$c were aligned using ClustalW. The toxin loci span genes CT165 to CT168, although only CT166 is considered the toxin B-related protein. The other genes represent hypothetical proteins of unknown function. The white boxes show the intergenic regions (IGRs). The grey rectangles represent the genes but are not drawn to scale. Dashed lines represent indels.

and L$_2$c (Fig. S7, cyan and grey highlights, respectively). No polymorphisms in the proline-dense and actin-binding domains were identified.

Region 7 contained non-D-like SNPs at a frequency of 10 to 20% of the total reads. Comparison of D(s)/2923 with L$_2$c in this region showed similar results.

Other significant individual L$_2$c genes. Although neither IncA expression nor large fused inclusions were observed *in vitro* for L$_2$c (Fig. 1), all LGV strains, including L$_2$c, had identical *incA* sequences (data not shown). The promoter region encompassing 1,500 nt upstream of the *incA* start codon showed a high degree of sequence conservation, with ≥99% nucleotide identity among reference strains A to K, Ba, Da, Ia, Ja, and L$_1$ to L$_3$ and clinical strains, L$_2$b and L$_2$c but not with D(s)/2923. Moreover, the promoter and encoding sequences of five hypothetical proteins (CTL0475 to CTL0540) in L$_2$/434, which might function like IncA proteins due to the characteristic IncA protein domains and family, were also highly conserved (data not shown).

The sequence of the L$_2$c *ompA* gene differed from the L$_2$/434 sequence in an SNP upstream of variable segment 4 (VS4) that is conserved among L$_2$b/UCH-1/proctitis and L$_2$′ (see Fig. S8 in the supplemental material). The mutation results in a synonymous codon change.

DISCUSSION

The last decade has seen a cumulative increase in the evidence for recombination and horizontal gene transfer (LGT) among intracellular bacteria, and *Chlamydia* has been no exception to this (14, 22–31). Data on recombination for *C. trachomatis* have come solely from extensive comparative analyses of multiple genes dispersed throughout the genome for reference and clinical strains. However, the authors of two recent publications of different *C. trachomatis* clinical isolates that were genome sequenced concluded that there were regions of the genomes that were consistent with interstrain recombination (9, 18). Here, we provide the first analytically confirmed whole-genome evidence for recombination between *C. trachomatis* strains and, surprisingly, between invasive (LGV) and noninvasive urogenital strains. This provokes a reassessment of how we think about *C. trachomatis* infections and their evolution *in vivo*. We know that there are many diverse microbes that inhabit the rectum. However, it has not been considered that *C. trachomatis* strains coinfect the rectal mucosal with any relevant frequency, unlike the urethra or cervix (32), and undergo LGT to produce recombinants.

Our study was also informative in showing how laboratory processing of *C. trachomatis* strains after isolation from the patient skewed understanding of the nature of the infection. Initially, the
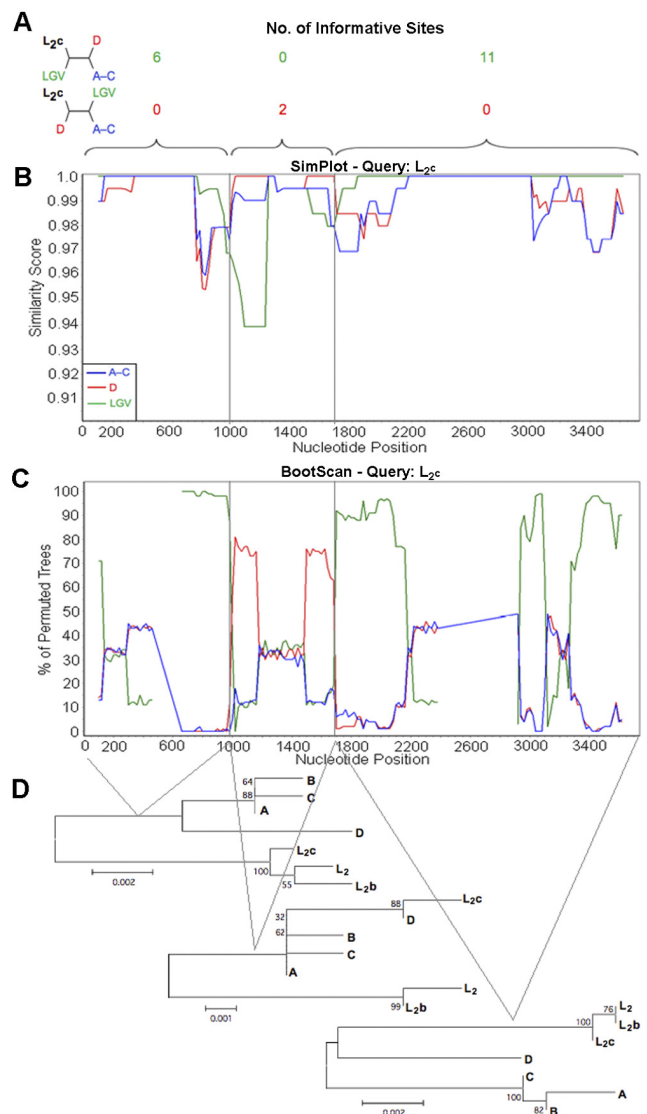
FIG 7 SimPlot representation and maximum chi-square analyses of putative recombinant crossover regions in the *tarp* gene for clinical strain L$_2$c compared with those of parental strains L$_2$/434, L$_2$b/UCH-1/proctitis, and D/UW3. The outgroup was the collective sequence of the ocular strains A/2497, B/HAR36, and C/TW3, as they were identical for this gene. (A) The number of informative sites shared by the recombinant L$_2$c *tarp* gene sequence with parental LGV (green) and D/UW3 (red) *tarp* gene sequences is shown. The ocular outgroup is denoted in blue. Note that L$_2$c is more similar to D in the first ~996 bp and to LGV strains in the middle ~800 bp but similar again to D for the remainder of the gene. (B) The similarity plot is shown using the L$_2$c *tarp* gene sequence as the query sequence in comparison with those of the LGV (green), ocular group (blue), and D/UW3 (red) strains (similarity score on the *y* axis). (C) The bootscan analyses are shown, with percent similarity and phylogenetic relatedness (percentage of permuted trees on the *y* axis), with significance of support for each node in the tree between recombinant and parental sequences as denoted by the colors. (D) The phylogenetic reconstructions are shown for each region bounded by recombination breakpoints supporting each crossover along with the results for 1,000 bootstrap replicates as supported by the analysis represented in panel C. Left tree, base pairs 0 to 996; middle tree, base pairs 997 to 1689; right tree, base pairs 1690 to 3735. See Materials and Methods.

L$_2$c culture exhibited toxicity to HeLa cells and expressed the toxin protein. In addition, an intact D-like toxin gene was recovered by PCR and Sanger sequencing of the L$_2$c genome. These data sug-

gested a correlation with the clinical severity of disease. After multiple passage in cell culture, the cytotoxic phenotype was lost (Fig. 2), likely representing selection against the cytopathic clone in culture, and we could not detect the toxin gene in the genome sequencing project using the passaged isolate. The conclusion from this experience is that ideally, in order to fully grasp the complexity of *C. trachomatis* infections, including the frequency and diversity of recombinants, techniques for direct genome sequencing (without prior amplification) of the organisms present in infected tissue will be needed.

IncA is an important constituent of the inclusion membrane, facilitating the fusion of inclusions within the cell. Variants of *C. trachomatis* strains B, D to H, Ia, and J that completely lack *incA* or lack a portion of the gene produce multiple inclusions that do not fuse (5, 33). The possible benefits of fusogenic inclusions include interaction with host cell vesicle trafficking and genetic exchange between the DNA from different RBs (4), although it has been reported that nonfusogenic strains can undergo recombination (18), as has been demonstrated *in vitro* between nonfusogenic and wild-type strains (27). Interestingly, patients infected with naturally occurring *incA* "knockouts" or mutants have fewer signs and symptoms, lower proliferative capacity, and fewer inclusion-forming units (IFUs) in culture than wild-type fusing strains (12). L_2c failed to express IncA at any time during development and produced many small nonfusogenic inclusions (Fig. 1). In contrast to the majority of asymptomatic cases caused by other nonfusogenic strains (12), L_2c was hypervirulent in terms of clinical signs and symptoms, producing severe hemorrhagic proctitis, although the patient did not exhibit an inguinal syndrome, which may in part be due to the presence of a functional cytotoxin (Fig. 2) (discussed below). The L_2c sequence of *incA* was identical to that of $L_2/434$, and *incA* mRNA expression levels were similar for both strains (Fig. 1E), which suggests that there may be a disruption in regulation for protein processing. Recent studies of the transcription expression profile and cell culture kinetics of naturally occurring IncA knockout and wild-type strains suggest that the IncA-negative phenotype may arise from multistep events, involving a decrease in transcription level and/or a partial or complete inactivation of translation (5). Alternatively, host environmental clues may be necessary for regulation, as has been discovered for certain proteins among other human Gram-negative pathogens, such as *Burkholderia* (34).

While the genomes of *C. trachomatis* strains to date are relatively conserved and share a high degree of synteny, an exception to this is found in the plasticity zone (PZ) (28). The PZ is typically rich in heterogeneity, with evidence for genome rearrangement as well as LGT for many bacterial species, including *Bartonella grahamii*, *Helicobacter pylori*, and *Shigella flexneri* (28, 35–37). The chlamydial PZ contains metabolic and virulence factors associated with tissue tropism and immune evasion, including the toxin loci (38, 39). Here, we consider it likely that the partial toxin gene was acquired from coinfection with a D strain, since no LGV sequences ($n = 6$) to date contain a partial or complete toxin gene. Indeed, D strains are prevalent in rectal infections among MSM. A recent study in Sweden of *C. trachomatis* infections among 197 MSM identified high prevalences of strains G (45%), D (27%), and J (26%) in the rectum (40), although there was no information on the presence of mixed infections or coinfection with LGV strains. Importantly, the genomic uptake of DNA by transformation can occur during coinfection or sequential infection (41).

Transformation is a likely mechanism employed by *C. trachomatis*, which would provide vast opportunities for genetic exchange.

We also found that the toxin acquired by L_2c was functional (Fig. 2). In our cytotoxicity assay, L_2c had a profound effect on cell morphology and death, as did D/UW3, compared with $L_2/434$. This effect was similar to that noted previously for strain D (42). The *C. trachomatis* toxin is known to play an important role in damaging host cell actin microfilaments, likely facilitating growth of the intracellular inclusion (38, 42). Analyses of the toxin loci of *Chlamydia muridarum*, a mouse pathogen, have suggested that the toxin may function during an early phase of infection to inactivate GTPase near sites of EB entry, resulting in innate immune evasion (28, 39). Consequently, introduction of the functional toxin into L_2c may support a mechanism for survival through escaping immune surveillance and allowing sufficient replication in nonfusing inclusions to cause severe localized mucosal disease as in our patient. Moreover, the relatively high degree of cytotoxicity may result in barriers to dissemination to regional lymph nodes and, thereby, a lack of an inguinal syndrome. While L_2c does not appear to be more cytotoxic than D/UW3, some degree of cytotoxicity may limit the ability to successfully culture these and other strains with a partial or complete toxin gene, which could potentially hinder our abilities to detect emerging LGV strains that contain the toxin and to further characterize them.

Tarp is secreted via the type III secretion system (TTSS), present in both EBs and RBs (43), at sites of entry into the host cell for purposes of pathogen-directed actin polymerization and cytoskeleton rearrangement, an event that coincides with EB entry into the cell (44). While the L_2c *tarp* gene sequence alignment and phylogenetic tree imply a close genetic relationship with $L_2/434$ and $L_2b/UCH-1/proctitis$ strains, L_2c contains a large in-frame deletion. Based on our analyses, the L_2c *tarp* gene is likely a recombinant of L_2 and D strains (Fig. 7).

According to functional studies of the chlamydial Tarp protein (21), tyrosine phosphorylation has been associated with actin recruitment and inclusion development, while the number of tyrosine-rich repeat regions has been associated with functionality. But inhibition of Tarp tyrosine kinase activity had no effect on EB entry (21). Thus, the presence of only two partial and two complete regions of tyrosine-rich repeats in the L_2c Tarp, compared to six complete regions in other LGV strains (see Fig. S7 in the supplemental material), would not prevent pathogen entry but might affect cytoskeletal rearrangement that could impair inclusion development and result in smaller inclusions, as observed for L_2c (Fig. 1C). If this is found to be a common recombinant region for clinical strains, it would highlight an evolutionary mechanism for diversifying the number of tyrosine residues to affect intracellular growth and infection outcomes. Indeed, recent sequence analysis of the *tarp* gene from numerous clinical strains found mutations that were similar among strains causing the same disease, and phylogenetic analysis suggested that this is one of the few genes that are responsible for *C. trachomatis*-specific disease phenotypes (45). Inferior actin recruitment may also be correlated with a lack of IncA expression and function, which would affect inclusion fusion as in L_2c.

During the late stage in the developmental cycle, *C. trachomatis* expresses two histone H1 homologues that are involved in RB-to-EB transition through nucleoid compaction and downregulation of gene expression (19, 46). As shown in Fig. S5 in the supplemental material, the molecular mass and repetitive penta-

peptide motifs of Hc2 are inversely correlated with the size of the deletion; thus, there are variable numbers among the *C. trachomatis* strains (19, 20). The deletion in L$_2$c Hc2 resulted in a substantial decrease in pentapeptide motifs and also in the number of positively charged amino acids in two-thirds of the Hc2 amino terminus. According to *in vitro* studies showing the DNA-binding activity of L$_2$/434 Hc2 expressed in *Escherichia coli* (20), the lower number of repetitive motifs in L$_2$c Hc2 may weaken electrostatic and hydrogen-bonding interactions of Hc2 with DNA, which may affect transcriptional regulation. Moreover, while ocular and D/UW3 strains contain a proline at coordinate 153 (Fig. S5, denoted in bold and underlined), which creates a "kink" in protein structure allowing Hc2 to participate in stronger interactions (20, 47), the lack of proline in L$_2$c Hc2 (P153T) suggests weaker DNA-binding activity. Consequently, inefficient repression of gene expression would likely impede RB-to-EB differentiation and may in part explain the poor propagation rate and morphological characteristics of L$_2$c in culture.

A significant fraction of *C. trachomatis* genomes have unknown function, and the implied function of many genes is based on sequence similarity to homologous genes from other prokaryotes and a few eukaryotes. While there have been numerous attempts to produce stable mutants of *C. trachomatis* without success (48), which has hindered our ability to obtain unambiguous information about gene function, the discovery of naturally occurring recombinants, such as L$_2$c, can help clarify the functional importance of specific genes and disease phenotypes and pave the way towards identifying the underlying mechanisms of LGT for developing a gene transfer system for *Chlamydia*.

MATERIALS AND METHODS

Clinical sample source. The study was approved by the Institutional Review Board of Children's Hospital and Research Center at Oakland in accordance with the Declaration of Helsinki. The clinical sample was obtained from the rectal mucosa of a male who had a history of sex with men and presented with severe hemorrhagic proctitis. Briefly, a 26-year-old male presented to a San Francisco Bay area clinic with a complaint of severe rectal pain with blood on defecation. He described a series of encounters with homosexual men and unprotected anal receptive intercourse during the prior month until the onset of rectal pain 5 days prior to the clinic visit. The rectal bleeding had commenced 1 day earlier. The man had a history of gonorrhea in the past but no other known STDs, no known exposure to men with known STDs, and no history of illicit drug use. He was reported to be HIV negative and had no other medical conditions, was not on any over-the-counter or prescription medications, and appeared to be in excellent health. On the physical exam, he was a well-appearing male, afebrile, and normotensive, with no evidence for an ulcerative lesion on the glans or shaft of the penis and no inguinal adenopathy. The anus was inflamed, and on proctoscopy, there was extensive bleeding of the mucosa, with evidence of a purulent discharge. Four swabs from each quadrant of the rectum were obtained and placed in transport medium. Three swabs were sent to the clinical laboratory for standard detection of *Neisseria gonorrhoeae* by commercial PCR and culture and of *C. trachomatis* by commercial PCR. The fourth swab was sent to the *Chlamydia* Research Laboratory at CHORI for in-house *C. trachomatis* culture.

***C. trachomatis* strains and plaque assay.** Reference strains D/UW3, L$_1$/440, L$_2$/434, L$_2$a/TW-396, and L$_3$/404 (49), a clinical L$_2$b strain from Amsterdam (a kind gift from Servaas Morré), and the clinical sample, referred to as L$_2$c, from the above-described case, were analyzed. Each strain was propagated in the human cervical adenocarcinoma cell line HeLa229 using our previously described protocols (13, 50, 51). The clinical sample was diluted and directly plaque purified using sequential

plaque purifications per our referenced protocols (13, 50, 51). To verify the clonal purity of the plaques, *ompA* and MLST genes (14) were amplified by PCR and cloned separately using a TOPO TA cloning kit (Invitrogen) (52); 10 clones of each were randomly selected for Sanger sequencing using techniques we have described previously (24). The confirmed clonal plaques were then individually propagated in tissue culture. A total of two passages were performed for L$_2$c to generate sufficient gDNA for genome sequencing. The EBs for each *C. trachomatis* isolate were purified from contaminating human cells using DNase treatment followed by gradient ultracentrifugation, and genomic DNA was purified from each isolate using a High Pure PCR template preparation kit (Roche Diagnostics, Indianapolis, IN) as we previously described (13, 53).

For determination of MOI based on IFUs, duplicate serial 10-fold dilutions of purified EBs were used to infect HeLa cells in 24-well plates. After 24 to 48 h, the cells were fixed in methanol, washed with phosphate-buffered saline (PBS), and stained using the Pathfinder *Chlamydia* culture confirmation monoclonal antibody (MAb) (Kallestad Diagnostics, Chaska, MN) in accordance with the manufacturer's directions. The number of IFUs per well was divided by the number of cells per well; an average was taken for duplicate wells to arrive at the MOI per strain.

Morphological characterization, *incA* expression, and growth curve. HeLa cell monolayers grown in minimal essential medium (MEM) containing 10% fetal bovine serum (UCSF Cell Culture Facility, San Francisco, CA) and 1 μg/ml gentamicin (MP Biomedicals, Solon, OH) at a confluence of 80% on 12-mm coverslips (Fisher Scientific, Pittsburgh, PA) in 24-well plates were infected with either reference strains L$_1$/440, L$_2$/434, L$_2$a/TW-396, L$_3$/404, and D/UW3, clinical strain L$_2$b from Amsterdam, or the clinical strain L$_2$c from this study in sucrose-phosphate-glutamine (219 mmol/liter sucrose, 3.82 mmol/liter KH$_2$PO$_4$, 8.59 mmol/liter Na$_2$HPO$_4$, 4.26 mmol/liter glutamic acid, 10 μg/ml gentamicin [MP Biomedicals], 100 μg/ml vancomycin [Acros Organics, Morris Plains, NJ], and 25 U/ml nystatin [MP Biomedicals] in distilled water, pH 7.4) at an MOI of 1, unless indicated, for 2 h on an orbital shaker at room temperature. The inocula were aspirated, and the infected monolayers were cultured in a humidified incubator at 37°C with 5% CO$_2$ in Dulbecco's modified MEM (Cellgro, Manassas, VA) with GlutaMAX-1 (Life Technologies, Rockville, MD) supplemented with 10% fetal bovine serum (UCSF Cell Culture Facility), 0.45% glucose solution (Cellgro), 20 mM HEPES (UCSF Cell Culture Facility), 0.08% NaHCO$_3$, and 1 μg/ml cycloheximide (13). At 12, 18, 24, 36, and 48 h (48 h for D/UW3 only) postinfection, the coverslips were fixed with methanol for 10 min, rinsed in PBS, and incubated for 30 min with anti-*C. trachomatis* specific lipopolysaccharide (LPS) MAb (Virostat, Portland, ME) and anti-IncA MAb 3H7 (gift from Daniel D. Rockey) or polyclonal anti-IncA (gift from Ted Hackstadt). The secondary antibodies were Cy-3 conjugated IgG (Jackson ImmunoResearch, West Grove, PA) for LPS and fluorescein isothiocyanate (FITC)-labeled IgG (Jackson ImmunoResearch) or Alexa 488 (Invitrogen) for IncA and chlamydial heat shock protein 60. DAPI (4′,6-diamidino-2-phenylindole dihydrochloride) (Vector Laboratories, Burlingame, CA) was used to stain DNA. The inclusions formed by the reference LGV and D/UW3 strains and the clinical L$_2$c strain were visualized on a Zeiss 510 confocal microscope. Light microscopy was used to examine the D/UW3, LGV, and clinical L$_2$c strains at 36 h, but no inclusions were observed for any LGV strains at this time point.

We examined *incA* expression using our previously described techniques, with slight modifications (54, 55). HeLa cells were infected at 0, 2, 12, 24, and 48 h with D/UW3, L$_2$/434, and L$_2$c at an MOI of 5. Total RNA was extracted using an RNeasy minikit (Qiagen, Valencia, CA) per the manufacturer's instructions; on-column DNase (Qiagen) treatment was performed to remove contaminating DNA. cDNA was generated from 2 μg of total RNA using TaqMan reverse transcriptase (RT) reagents and random hexamers (Applied Biosystems, Foster City, CA). Quantitative real-time PCR was performed in replicate using SYBR green chemistry, reagents, primers (see Table S1 in the supplemental material), thermocycling, and standard curves as we previously described (54, 55). 16S rRNA

was used for normalization. Negative controls and the standard curves for each gene were used as previously described (54, 55). Analysis of the dissociation curves was used to verify the specificity of the amplified products. The conversion of the mean threshold cycle values determined the relative amounts of target and control gene from the respective standard curve. Three independent experiments were performed.

Growth curves for strains D/UW3, L$_2$/434, and L$_2$c were generated using quantitative PCR as we previously described (52, 56). Briefly, each strain at an MOI of 1 was grown in HeLa cells and harvested at 0, 12, 18, 24, 36, and 48 h, and total RNA was extracted and reverse transcribed as described above. Each quantitative PCR consisted of 1× SYBR green master mix (Applied Biosystems), 1 μl of cDNA, and 5 pmol of each primer (16S rRNA and glyceraldehyde-3-phosphate dehydrogenase [GAPDH] as described in reference 56) in a total reaction volume of 25 μl run on an ABI 7900 (Applied Biosystems) using our thermocycling profile as previously described (56); standard curves were generated, and negative controls for each primer pair were included in each run as we previously described (56). GAPDH was used for normalization. Melting curves were used to verify the specificity of the reactions and the absence of primer dimmers. Samples were amplified in triplicate where the mean was used for analysis. Three independent experiments were performed.

Cytotoxicity assay. Reference strains D/UW3 and L$_2$/434 and the clinical L$_2$c strain from this study were analyzed in a cytotoxicity assay adapted from the method of Belland et al. (42). HeLa cells were grown to 80% confluence as described above in 24-well plates. The monolayers were treated with 1 ml DEAE-dextran (45 μg/ml) in Hanks' balanced salt solution (HBSS) for 15 min at 37°C. The cells were infected at an MOI of 100 (~5 × 10^7 IFU) with each strain and mock infected in duplicate at room temperature for 4 h on an orbital shaker. The inocula were removed, and the cells were washed with HBSS and incubated for an additional 4 h in growth medium as described above except for the addition of 10 μg/ml gentamicin (MP Biomedicals), 25 μg/ml vancomycin (Fisher) to prevent any nonchlamydial bacterial growth. The monolayers were visualized under light microscopy at ×400 magnification. The cells were scored for rounding, detachment, and lysis compared to the levels for uninfected control cells by the following metric: (−), same as cell control; 3+, 100% of cells affected; 2+, 75% of cells affected; and 1+, 25% of cells affected.

The above-described experiments were repeated using 1 μg/ml of doxycycline to pretreat cells before infection. In addition, the assay was repeated using L$_2$/434 and a higher passage number (passage no. 20) for L$_2$c. A total of two independent experiments were performed for each of these studies.

To detect the toxin protein, a Western blot analysis was performed as we have previously described, with modifications (55). Briefly, L$_2$/434, D/UW3, and L$_2$c were independently grown in HeLa cells as described above at an MOI of 100. Purified EBs from each isolate were solubilized in Laemmli buffer, run on NuPAGE Novex 4 to 12% bis-Tris precast gels (Invitrogen) with a BenchMark prestained protein ladder (Invitrogen), and electrophoretically transferred onto a BioTrace polyvinylidene difluoride (PVDF) membrane (Pall Life Science, Port Washington, NY) at 120 V for 2 h in 1× NuPAGE transfer buffer (Invitrogen) with 10% methanol. The membrane was incubated with 5% skim milk in 1× Tris-buffered saline (TBS) and 0.1% Tween 20, washed with 1× TBS-0.1% Tween 20, and reacted overnight with a polyclonal antiserum against recombinant CT166 (gift from Harlan Caldwell) at a 1:500 dilution. The membrane was washed, and secondary antibody (goat anti-rabbit IgG conjugated to horseradish peroxidase [HRP]; Bio-Rad, Hercules, CA) was applied at a 1:1,000 dilution, incubated at room temperature, and washed in 1× TBS prior to detection using the SuperSignal West Pico chemiluminescent substrate (enhanced chemiluminescence [ECL] detection system; Thermo Scientific, Rockford, IL) and CL-XPosure Film (Thermo Scientific).

Genome sequencing and assembly. The purified genomic DNA from the clinical case was shotgun sequenced using a combination of the 454/Roche GS-FLX Titanium and GS Junior instruments (454 Life Sciences,

Branford, CT). The 809,874 reads were generated with an average read length of 404 nt. Preliminary analysis suggested an unusually high proportion of contamination of DNA from the HeLa cell culture. Therefore, 618,844 reads of human origin were removed (identified by mapping to the hs19 golden path release of the genome [http://genome.ucsc.edu/cgi-bin/hgGateway]) using Newbler gsMapper 2.5 software.

We assembled a random subset of 50,000 of the remaining reads *de novo* using the 454 gsAssembler software program, version 2.0.01.14, with default parameters. The order of the contigs in the final genome sequence was determined using a combination of the possible connections between contigs suggested by overlapping reads (57) and information from mapping contigs against the L$_2$/434 reference genome (9). After determination of the correct path through the contig graph, the contigs were assembled to form the final sequence. To accomplish this, the ends of adjoining contigs were matched using BLAST (58) against the database of reads. Reads found in both contigs were assembled into a consensus sequence that bridged the two contigs to form a single, larger contig. The majority of gaps were small enough to be spanned by a single read. The plasmid was a single contig with reads overlapping the beginning and end, indicating the expected circular redundancy.

For verification of the sequences, we aligned the original reads to the consensus chromosomal and plasmid templates using gsMapper 2.3; 91.6% of the 191,030 human screened reads (average length, 436 nt) mapped to the chromosome, and 6.8% mapped to the plasmid. The remaining unmapped reads were found not to assemble *de novo* into any contigs consisting of more than 2 reads, suggesting that these were mostly contamination that had not matched the human reference sequence. The mapped assemblies were inspected to remove a small number of errors. The final assembly, with an average redundancy of coverage of approximately 75-fold, was used for annotation. The identity of each base in the final consensus sequence was determined by majority vote. The gsMapper software revealed two possible structural variants in a minority of sequence reads. The first was a deletion between coordinates 133542 and 155549 present in 4 sequence reads. The second was a deletion between bases 849581 and 849942 present in 19 sequence reads.

Genome annotation. The L$_2$c consensus chromosome and plasmid sequences were annotated automatically using the Integrative Services for Genomics Analysis pipeline (16). Promoter 2.0 prediction (http://www.cbs.dtu.dk/services/Promoter/), Bacterial PROMoter prediction (http://www.softberry.ru/berry.phtml), and the Berkeley *Drosophila* Genome Project neural network (http://www.fruitfly.org/seq_tools/promoter.html) were used to identify the *ftsK* putative promoter region, including the initiation site, −10 and −35 promoter elements, and ribosome binding sites. The default parameters were used for prediction scores for significance.

Verification of indels. To verify the regions of diversity (L$_2$-D hybrids), single amplicons of these regions (the *tarp* gene, *incA*, *hctB*, and toxin locus genes; the IGR upstream of *ftsK*) that were identified by the SAS program Q-plotGenome (see below) and genomic analyses (see below) as divergent from L$_2$/434 were amplified, cloned using a TOPO TA cloning kit (Invitrogen), and Sanger sequenced using primers (see Table S1 in the supplemental material) designed to amplify and sequence the full length of each as previously described (16). Ten clones were sequenced for each. The sequences were compared with those of the following available reference strains (GenBank accession numbers are given in parentheses): A/2497 (EU121607), B/HAR36 (EU121608), C/TW3 (EU121609), D/UW3 (AE001273), L$_2$/434/Bu (AM884176), and L$_2$b/UCH-1/proctitis (AM884177) for the *tarp* gene; A/HAR13 (CP000051), B/Jali20/0T (FM872308), C/TW3 (EU121596), D/UW3 (AE001273), L$_2$/434/Bu (AM884176), and L$_2$b/UCH-1/proctitis (AM884177) for *hctB*; A/HAR13 (CP000051), B/TW5 (DQ064209), B/TZ1A828/OT (FM872307), Ba/Apache-2 (DQ064210), C/TW3 (DQ064211), D/UW3, E/Bour (DQ064214), F/IC-CAL3 (DQ064215), G/UW57 (EU247624), H/UW4 (EU247625), i/UW12 (DQ064218), Ia/870 (DQ064219), J/UW36 (DQ064220), K/UW31 (DQ064221), L$_1$/440 (DQ064222), L$_2$/

434/Bu, L_2b/UCH-1/proctitis, and L_3/404 (DQ064224) for *incA*; C/TW3 (AY647994), D/UW3, H/UW4 (AY647999), and L_2/434 for the cytotoxin locus; and A/HAR13, B/Jali20/0T, D/UW3, L_2/434, and L_2b/UCH-1/ proctitis for the IGR upstream of *ftsK*.

Comparative genomics and phylogenetic analysis. To detect recombination, we developed a program called Q-plotGenome that was written as a set of macros in the SAS software 9.2 language (SAS Institute, Inc., Cary, NC) to compare the genome sequences of L_2c with those of reference strains L_2/434 and D/UW3. The core of the program is similar to that of SimPlot (59) except that it covers the entire 1-Mb genome instead of a discrete sliding window of limited size; it samples a series of fixed-length subsequences (windows) from one genome and then compares them to windows from the other genome. Q-plotGenome is, thus, a tool for comparing DNA sequences in the range of 1 Mb plus and for displaying results graphically. The description of Q-plotGenome is detailed in Text S1 in the supplemental material.

To identify statistically significant recombination breakpoints or regions, SimPlot software 3.5.1 (http://sray.med.som.jhmi.edu/SCRoftware/) was utilized as we have previously described (25). The parameters included a window size of 200 base pairs (bp), a step of 20 bp, neighbor-joining trees calculated using the Kimura-2-parameter distance model (60), and 1,000 bootstrap replicates to determine confidence for each branch (Fig. 7).

For a global analysis, we used a version of the BLAST score ratio approach (17) to identify recombinants. We aligned the databases of both (i) raw L_2c sequences and (ii) 100-nt windows along the L_2c chromosome sequence separated by 50 nt against the L_2/434 genome as a reference and the other complete genomes listed in the previous section. Matches with a BLAST score ratio of <0.95 were plotted using Circos software (Fig. 5) (61).

For analysis of specific regions, nucleotide sequences were aligned by ClustalW 1.8, MUSCLE (62), or MAUVE (63). Gaps were removed using GBLOCKS (64) with default parameters. Phylogenetic inference and tree plotting were performed using the MEGA 3.1 (65) package described previously (13, 24, 25) or PHYLIP (66) programs DNApars, Seqboot, and Consensus. Neighbor-joining trees were calculated using the Kimura 2-parameter model that assumes constant nucleotide frequencies and their rates of substitution among sites (60) and 1,000 bootstrap replicates.

For detection of strain D-specific SNPs within the L_2c data, we first identified SNPs between the D/UW3 and L_2/434 genomes using the show-snps tool of the MUMmer package (67). We then mapped L_2c against L_2/434 using gsMapper 2.3 and extracted the number of "D"-like and "L_2"-like nucleotides at each variant position from the 454AllDiffs.txt output using a custom script. Similar analyses were performed for D(s)/ 2923.

Nucleotide sequence accession numbers. The complete chromosome and plasmid sequences were submitted to the NCBI GenBank database (accession no. CP002024). The NCBI genome project identification number is 47581. The raw data from the 191,030 human screened L_2c reads were deposited in the NCBI short read archive (sra; accession no. SRP002231). The sequence of the L_2c toxin gene from the initial plaque purification was submitted to GenBank (accession no. 2981755).

ACKNOWLEDGMENTS

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org /lookup/suppl/doi:10.1128/mBio.00045-11/-/DCSupplemental.

Text S1, PDF file, 0.057 MB.
Table S1, PDF file, 0.060 MB.
Figure S1, PDF file, 2.139 MB.
Figure S2, PDF file, 0.161 MB.
Figure S3, PDF file, 0.088 MB.
Figure S4, PDF file, 0.086 MB.
Figure S5, PDF file, 0.069 MB.
Figure S6, PDF file, 0.122 MB.
Figure S7, PDF file, 0.089 MB.
Figure S8, PDF file, 0.095 MB.

REFERENCES

1. **World Health Organization.** 2009, posting date. Sexually transmitted infections http://www.who.int/topics/sexually_transmitted_infections /en/ Accessed 23 December 2009.
2. **Centers for Disease Control and Prevention.** 2009. *Chlamydia* screening among sexually active young female enrollees of health plans—United States, 2000-2007. MMWR Morb. Mortal. Wkly. Rep. **58**:362–365.
3. **White JA.** 2009. Manifestations and management of lymphogranuloma venereum. Curr. Opin. Infect. Dis. **22**:57–66.
4. **Hackstadt T, Fischer ER, Scidmore MA, Rockey DD, Heinzen RA.** 1997. Origins and functions of the chlamydial inclusion. Trends Microbiol. **5**:288–293.
5. **Suchland RJ, et al.** 2008. Identification of concomitant infection with *Chlamydia trachomatis* IncA-negative mutant and wild-type strains by genomic, transcriptional, and biological characterizations. Infect. Immun. **76**:5438–5446.
6. **Centers for Disease Control and Prevention.** 2004. A cluster of lymphogranuloma venereum among men who have sex with men—Netherlands, 2003-2004. MMWR Morb. Mortal. Wkly. Rep. **53**:985–988.
7. **Martin-Iguacel R, et al.** 2010. Lymphogranuloma venereum proctocolitis: a silent endemic disease in men who have sex with men in industrialised countries. Eur. J. Clin. Microbiol. Infect. Dis. **29**:917–925.
8. **Spaargaren J, Fennema HS, Morré SA, de Vries HJ, Coutinho RA.** 2005. New lymphogranuloma venereum *Chlamydia trachomatis* variant, Amsterdam. Emerg. Infect. Dis. **11**:1090–1092.
9. **Thomson NR, et al.** 2008. *Chlamydia trachomatis*: genome sequence analysis of lymphogranuloma venereum isolates. Genome Res. **18**: 161–171.
10. **Spaargaren J, et al.** 2005. Slow epidemic of lymphogranuloma venereum L_2b strain. Emerg. Infect. Dis. **11**:1787–1788.
11. **Herida M, Kreplack G, Cardon B, Desenclos JC, de Barbeyrac B.** 2006. First case of urethritis due to *Chlamydia trachomatis* genovar L_2b. Clin. Infect. Dis. **43**:268–269.
12. **Geisler WM, Suchland RJ, Rockey DD, Stamm WE.** 2001. Epidemiology and clinical manifestations of unique *Chlamydia trachomatis* isolates that occupy nonfusogenic inclusions. J. Infect. Dis. **184**:879–884.
13. **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.
14. **Dean D, et al.** 2009. Predicting phenotype and emerging strains among *Chlamydia trachomatis* infections. Emerg. Infect. Dis. **15**:1385–1394.
15. **Margulies M, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**:376–380.
16. **Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q.** 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics **26**:1122–1124.
17. **Rasko DA, Myers GS, Ravel J.** 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics **6**:2.
18. **Jeffrey BM, et al.** 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.
19. **Hackstadt T, Brickman TJ, Barry CE III, Sager J.** 1993. Diversity in the *Chlamydia trachomatis* histone homologue Hc2. Gene **132**:137–141.
20. **Brickman TJ, Barry CE III, Hackstadt T.** 1993. Molecular cloning and expression of *hctB* encoding a strain-variant chlamydial histone-like protein with DNA-binding activity. J. Bacteriol. **175**:4274–4281.
21. **Jewett TJ, Dooley CA, Mead DJ, Hackstadt T.** 2008. *Chlamydia trachomatis* tarp is phosphorylated by src family tyrosine kinases. Biochem. Biophys. Res. Commun. **371**:339–344.
22. **Fitch WM, Peterson EM, de la Maza LM.** 1993. Phylogenetic analysis of the outer-membrane-protein genes of *Chlamydia*, and its implication for vaccine development. Mol. Biol. Evol. **10**:892–913.

23. **Millman K, Tavaré 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.

24. **Gomes JP, et al.** 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.

25. **Gomes JP, et al.** 2007. Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots. Genome Res. **17**:50–60.

26. **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.

27. **Suchland RJ, Sandoz KM, Jeffrey BM, Stamm WE, Rockey DD.** 2009. Horizontal transfer of tetracycline resistance among *Chlamydia* spp. in vitro. Antimicrob. Agents Chemother. **53**:4604–4611.

28. **Read TD, et al.** 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. **28**:1397–1406.

29. **Read TD, et al.** 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.

30. **Dugan J, Rockey DD, Jones L, Andersen AA.** 2004. Tetracycline resistance in *Chlamydia suis* mediated by genomic islands inserted into the chlamydial inv-like gene. Antimicrob. Agents Chemother. **48**:3989–3995.

31. **Binet R, Maurelli AT.** 2009. Transformation and isolation of allelic exchange mutants of *Chlamydia psittaci* using recombinant DNA introduced by electroporation. Proc. Natl. Acad. Sci. U. S. A. **106**:292–297.

32. **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.

33. **Suchland RJ, Rockey DD, Bannantine JP, Stamm WE.** 2000. Isolates of *Chlamydia trachomatis* that occupy nonfusogenic inclusions lack IncA, a protein localized to the inclusion membrane. Infect. Immun. **68**:360–367.

34. **O'Grady EP, Viteri DF, Malott RJ, Sokol PA.** 2009. Reciprocal regulation by the CepIR and CciIR quorum sensing systems in *Burkholderia cenocepacia*. BMC Genomics **10**:441.

35. **Berglund EC, et al.** 2009. Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*. PLoS Genet. **5**:e1000546.

36. **Fischer W, et al.** 2010. Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer. Nucleic Acids Res. **38**:6089–6101.

37. **Wei J, et al.** 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect. Immun. **71**:2775–2786.

38. **Carlson JH, et al.** 2004. Polymorphisms in the *Chlamydia trachomatis* cytotoxin locus associated with ocular and genital isolates. Infect. Immun. **72**:7063–7072.

39. **McClarty G, Caldwell HD, Nelson DE.** 2007. Chlamydial interferon gamma immune evasion influences infection tropism. Curr. Opin. Microbiol. **10**:47–51.

40. **Klint M, et al.** 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.

41. **Dubnau D.** 1999. DNA uptake in bacteria. Annu. Rev. Microbiol. **53**:217–244. doi:10.1146/annurev.micro.53.1.217.

42. **Belland RJ, et al.** 2001. *Chlamydia trachomatis* cytotoxicity associated with complete and partial cytotoxin genes. Proc. Natl. Acad. Sci. U. S. A. **98**:13984–13989.

43. **Beeckman DS, Vanrompay DC.** 2010. Bacterial secretion systems with an emphasis on the chlamydial type III secretion system. Curr. Issues Mol. Biol. **12**:17–42.

44. **Clifton DR, et al.** 2005. Tyrosine phosphorylation of the chlamydial effector protein Tarp is species specific and not required for recruitment of actin. Infect. Immun. **73**:3860–3868. doi:10.1128/IAI.73.7.3860-3868.2005.

45. **Lutter EI, et al.** 2010. Phylogenetic analysis of *Chlamydia trachomatis* Tarp and correlation with clinical phenotype. Infect. Immun. **78**:3678–3688.

46. **Murata M, et al.** 2007. Chlamydial SET domain protein functions as a histone methyltransferase. Microbiology **153**:585–592. doi:10.1099/mic.0.29213-0.

47. **Pabo CO, Sauer RT.** 1992. Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Biochem. **61**:1053–1095. doi:10.1146/annurev.bi.61.070192.005201.

48. **Heuer D, Kneip C, Mäurer AP, Meyer TF.** 2007. Tackling the intractable—approaching the genetics of Chlamydiales. Int. J. Med. Microbiol. **297**:569–576.

49. **Yuan Y, Zhang YX, Watkins NG, Caldwell HD.** 1989. Nucleotide and deduced amino acid sequences for the four variable domains of the major outer membrane proteins of the 15 *Chlamydia trachomatis* serovars. Infect. Immun. **57**:1040–1049.

50. **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.

51. **Dean D, Powers VC.** 2001. Persistent *Chlamydia trachomatis* infections resist apoptotic stimuli. Infect. Immun. **69**:2442–2447.

52. **Gomes JP, et al.** 2006. Correlating *Chlamydia trachomatis* infectious load with urogenital ecological success and disease pathogenesis. Microbes Infect. **8**:16–26.

53. **Caldwell HD, Kromhout J, Schachter J.** 1981. Purification and partial characterization of the major outer membrane protein of *Chlamydia trachomatis*. Infect. Immun. **31**:1161–1176.

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

55. **Nunes A, et al.** 2007. Comparative expression profiling of the *Chlamydia trachomatis* pmp gene family for clinical and reference strains. PLoS One **2**:e878.

56. **Pettengill MA, Lam VW, Ojcius DM.** 2009. The danger signal adenosine induces persistence of chlamydial infection through stimulation of A2b receptors. PLoS One **4**:e8299.

57. **Nagarajan N, et al.** 2010. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. BMC Genomics **11**:242.

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

59. **Lole KS, et al.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73**:152–160.

60. **Kimura M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

61. **Krzywinski M, et al.** 2009. Circos: an information aesthetic for comparative genomics. Genome Res. **19**:1639–1645.

62. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.

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

64. **Talavera G, Castresana J.** 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. **56**:564–577.

65. **Kumar S, Tamura K, Nei M.** 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5**:150–163.

66. **Felsenstein J.** 1989. PHYLIP—Phylogeny inference package (version 3.2). Cladistics **5**:164–166.

67. **Delcher AL, Phillippy A, Carlton J, Salzberg SL.** 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. **30**:2478–2483.