Title

Whole-Genome Enrichment and Sequencing of Chlamydia trachomatis Directly from Patient Clinical Vaginal and Rectal Swabs

Journal

mSphere, 6(2)

ISSN

1556-6811

Authors

Bowden, Katherine E
Joseph, Sandeep J
Cartee, John C
et al.

Publication Date

2021-04-28

DOI

10.1128/msphere.01302-20

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE

Whole-Genome Enrichment and Sequencing of *Chlamydia trachomatis* Directly from Patient Clinical Vaginal and Rectal Swabs

Katherine E. Bowden,[a]* Sandeep J. Joseph,[a] John C. Cartee,[a] Noa Ziklo,[d] Damien Danavall,[a] Brian H. Raphael,[a] Timothy D. Read,[b] Deborah Dean[c,d]

[a]Division of STD Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[b]Emory University School of Medicine, Atlanta, Georgia, USA
[c]University of California San Francisco School of Medicine, San Francisco, California, USA
[d]University of California San Francisco Benioff Children's Hospital Oakland Research Institute, Oakland, California, USA

ABSTRACT *Chlamydia trachomatis*, an obligately intracellular bacterium, is the most prevalent cause of bacterial sexually transmitted infections (STIs) worldwide. Numbers of U.S. infections of the urogenital tract and rectum have increased annually. Because *C. trachomatis* is not easily cultured, comparative genomic studies are limited, restricting our understanding of strain diversity and emergence among populations globally. While Agilent SureSelect^{XT} target enrichment RNA bait libraries have been developed for whole-genome enrichment and sequencing of *C. trachomatis* directly from clinical urine, vaginal, conjunctival, and rectal samples, public access to these libraries is not available. We therefore designed an RNA bait library (34,795 120-mer probes based on 85 genomes, versus 33,619 probes using 74 genomes in a previous one) to augment organism sequencing from clinical samples that can be shared with the scientific community, enabling comparison studies. We describe the library and limit of detection for genome copy input, and we present results of 100% efficiency and high-resolution determination of recombination and identical genomes within vaginal-rectal specimen pairs in women. This workflow provides a robust approach for discerning genomic diversity and advancing our understanding of the molecular epidemiology of contemporary *C. trachomatis* STIs across sample types, geographic populations, sexual networks, and outbreaks associated with proctitis/proctocolitis among women and men who have sex with men.

IMPORTANCE *Chlamydia trachomatis* is an obligate intracellular bacterium that is not easily cultured, which limits our understanding of urogenital and rectal *C. trachomatis* transmission and impact on morbidity. To provide a publicly available workflow for whole-genome target enrichment and sequencing of *C. trachomatis* directly from clinical urine, vaginal, conjunctival, and rectal specimens, we developed and report on an RNA bait library to enrich the organism from clinical samples for sequencing. We demonstrate an increased efficiency in the percentage of reads mapping to *C. trachomatis* and identified recombinant and identical *C. trachomatis* genomes in paired vaginal-rectal samples from women. Our workflow provides a robust genomic epidemiologic approach to advance our understanding of *C. trachomatis* strains causing ocular, urogenital, and rectal infections and to explore geo-sexual networks, outbreaks of colorectal infections among women and men who have sex with men, and the role of these strains in morbidity.

KEYWORDS *Chlamydia trachomatis*, SNPs, recombination, whole-genome enrichment

Chlamydia trachomatis is the most common cause of bacterial sexually transmitted infections (STIs) worldwide and the most notifiable disease in the United States (1,

Citation Bowden KE, Joseph SJ, Cartee JC, Ziklo N, Danavall D, Raphael BH, Read TD, Dean D. 2021. Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from patient clinical vaginal and rectal swabs. mSphere 6:e01302-20. https://doi.org/10.1128/mSphere.01302-20.

Editor Angela L. Rasmussen, Georgetown University

Address correspondence to Katherine E. Bowden, KBowden@cdc.gov.

* Present address: Katherine E. Bowden, Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

Received 18 December 2020
Accepted 10 February 2021
Published 3 March 2021

2). Although *C. trachomatis* infection can present with conjunctivitis, pharyngitis, urethritis, vaginal discharge, proctitis, or inguinal syndrome, most infections are asymptomatic, which can lead to reproductive morbidity in women and proctitis or proctocolitis in women and men who go untreated (2).

C. trachomatis strains are classified by genotype based on the outer membrane protein gene (*ompA*), which encodes the major outer membrane protein (MOMP) and is typically linked to clinical presentation (3). Genotypes are grouped according to disease association: ocular disease (A to C and Ba), urogenital and anorectal disease (D to K, Da, Ga, Ia, and Ja), and lymphogranuloma venereum (LGV) (L_1 to L_3, L_2a, L_2b, and L_2c) (2, 4–7). Several studies have reported that the prevalence of these genotypes differs by anatomical site and sexual network. Genotypes D and G are more commonly detected in the anorectal tract and, along with genotype J, are prevalent in women and men who have sex with men (MSM) (8–15). Genotypes D, E, and F are found in the majority of urogenital infections and are common among heterosexuals (8, 16). LGV genotypes are prevalent in MSM with and without HIV and are associated with an anorectal infection and the inguinal syndrome (17–21).

A 2016 meta-analysis of extragenital *C. trachomatis* and *Neisseria gonorrhoeae* infections in women, MSM, and men who have sex only with women (MSW) demonstrated median prevalences of 8.7%, 8.9%, and 7.7%, respectively, for rectal *C. trachomatis*, with infection often being asymptomatic (22). Further, a number of studies have shown that rectal infections outnumber those in the urogenital tract of women and are on the increase among MSM (23–28). Although common, rectal *C. trachomatis* transmission and its impact on morbidity are not well understood, likely due to the lack of routine screening of populations other than MSM (1, 22, 29). In May 2019, the FDA cleared two diagnostic tests, the Aptima Combo 2 assay (Hologic, Inc.) and the Xpert CT/NG (Cepheid), for use with extragenital specimens in the detection of *C. trachomatis* and *N. gonorrhoeae*. This recent diagnostic advancement will improve screening and surveillance capacity while offering an opportunity to better understand transmission of rectal *C. trachomatis* and its role in morbidity.

Transmission and molecular epidemiologic studies of *C. trachomatis* rely on *ompA* genotyping, multilocus sequence typing (MLST), and multilocus variable-number tandem-repeat analysis (MLVA) (10, 30–39). Unfortunately, these techniques are challenging and laborious when performed on clinical specimens and, except for *ompA* genotyping, often require tissue culture to generate sufficient DNA. Further, due to low genetic resolution, these methods fail to demonstrate precise inter- and intrastrain recombination events across the genome that contribute to strain diversity (40–44). Recombination has been important in creating emerging strains of *C. trachomatis*, such as L_2b, among MSM in many countries of the world and recombinant strains L_2/D (termed L_2c) and L_2b/D-Da in the United States and Portugal, respectively (28, 40–47).

Enrichment of the low copy number of *C. trachomatis* in clinical specimens presents the greatest challenge for culture-independent genome sequencing (48). Initially, this method employed immunomagnetic separation (IMS) for enrichment, followed by genome amplification using multiple displacement amplification (MDA), but the demonstrated success rate (15 to 30%) was low across clinical specimens (49–51). Other methods currently in use include depletion-enrichment, cell sorting-MDA, and multiplexed microdroplet PCR (48). In 2014, Christiansen et al. sequenced *C. trachomatis* from urine and vaginal swabs with an 80% (8/10) success rate (≥95 to 100% coverage of the respective reference genome) using custom RNA baits to enrich for *C. trachomatis* during library preparation (52). The same RNA bait library was subsequently used for conjunctival samples with a 60% (12/20) success rate and similar genome coverage. More recently, another Agilent custom RNA bait library was developed for *C. trachomatis* enrichment from rectal samples (47). This RNA bait method has therefore been used to understand genomic diversity in circulating *C. trachomatis* ocular, urogenital, and LGV lineages from clinical specimens but with varying success (47, 53–56).

In an effort to make direct sequencing of all *C. trachomatis* strains causing clinical

infections more efficient and publicly available, we designed an RNA bait library based on 85 *C. trachomatis* complete genomes with 34,795 120-mer probes, compared to 74 genomes with 33,619 120-mer probes as for previous bait libraries (48, 52) and optimized the experimental protocol. We used paired vaginal and rectal specimens from four women as a novel comparator but also because the latter sample types have become more clinically relevant due to increased numbers of these infections among heterosexual women (23–28). The new system will increase our ability to sequence *C. trachomatis* genomes of current circulating strains causing ocular, urogenital, and rectal infections in diverse populations, providing a more robust data set to understand current sexual networks and transmission within and among anatomic sites. From these data, we will be able to differentiate clusters and sporadic cases during outbreaks and potentially identify novel markers for typing *C. trachomatis*, in addition to exploring the role of these strains in ocular, genital, and colorectal morbidity.

RESULTS

Limit of detection. The limit of detection (LOD) was determined using various genomic copy numbers of a *C. trachomatis ompA* genotype D strain mapped to the D reference strain. Genomic libraries were prepared, enriched for *C. trachomatis*, and sequenced from spiked serial dilutions of genomic DNA (gDNA) bulked with human DNA for a total input of 3 μg DNA (for fragmentation and library preparation) using the expanded RNA bait library and Agilent SureSelect^{XT} protocol. Total *C. trachomatis* genome copy input ranged from 265 to 7,854,406 copies, which is similar to the range for genome copies from clinical samples (see below), with 1.33 to 99.28% of the quality-controlled reads binned as *Chlamydia* species, along with a mean mapping read depth to the *C. trachomatis* reference genome ranging from 0.57 to 562.67, respectively (Table 1; Fig. 1; also, see Fig. S1 in the supplemental material). With the quality control (QC) criteria for efficiency set at $\geq$98% genome coverage at a $\geq$5$\times$ read depth, genotype D had an LOD of 16,945 total genome copies (Table 1; Fig. 1).

Enrichment and genomic sequencing of *C. trachomatis* from genotype L_2b and patient specimen sets. To ensure efficiency of target enrichment from a *C. trachomatis ompA* genotype prevalent in anorectal infections in HIV-infected MSM, genomic libraries from spiked mock samples of genotype L_2b were prepared, enriched for *C. trachomatis*, and sequenced. Total *C. trachomatis* genome copy input ranged from 9,000 to 900,000 copies, with 97.22 to 98.77% of the quality-controlled reads binned as *Chlamydia* species along with a mean mapping read depth to the *C. trachomatis* reference genome ranging from 531.81 to 550.78, respectively (Table 2; Fig. S2).

To determine efficiency of target enrichment from vaginal and rectal clinical specimens, *C. trachomatis* was directly sequenced from 4 sets of patient-matched vaginal and rectal swabs. The ranges of copy numbers by qPCR were 328 to 2,218 genomes per μl for the vaginal samples and 1,664 to 43,905 genomes per μl for the rectal samples. Total input ranged from 20,764 to 3,336,780 *C. trachomatis* genome copies. More genome copies were present in the rectal swab than in the vaginal swabs in 3 of the 4 patient specimen sets (Table 2). As with the spiked gDNA serial dilutions, the proportion of reads classified as *Chlamydia* spp. from clinical samples was dependent on total genome copy input (Table 2; Fig. 2).

For three of the four patient specimen sets (sets 107, 192, and 98), 18.07%, 30.82%, and 17.13% more reads were classified, respectively, as *Chlamydia* spp. in the rectal swab than the respective vaginal swab (Table 2; Fig. 2). Interestingly, for patient specimen set 72, there were 32.94% more *Chlamydia* reads in the vaginal swab, which contained 68,201 more genome copies than the rectal swab (Table 2; Fig. 2). Mean read depth was on average 3.5-fold higher in rectal swabs from patient specimen sets 107, 192, and 98, while patient specimen set 72 demonstrated a 3.3-fold-higher mean read depth in the vaginal swab (Table 2; Fig. S2). For all patient specimen sets, the percentage of the *C. trachomatis* genome covered with at least 5$\times$ coverage was $>$98%, with the exception of 72R, which had only 96.11% of the genome covered at a minimum of 5$\times$ read depth.

TABLE 1 Sequence data analysis of *C. trachomatis* gDNA from spiked serial dilutions of genotype D

Sample ID	*ompA* genotype	C_T[a]	Total no. of genome copies input	No. of:		Nonhuman read pair ratio	No. of:		% selected read pairs[b,c,d]	Mean read depth[e]	% of genome with coverage	
				Raw read pairs	Nonhuman read pairs		Total read pairs after QC[b,c]	Selected read pairs[b,c,d]			≥5×	≥10×
D1	D	18.29	7,854,406	2,136,708	2,118,719	0.99	2,095,052	2,079,951	99.28	562.67	99.02	99.01
D2	D	22.01	942,398	1,482,054	1,382,594	0.93	1,335,421	1,318,461	98.73	359.60	99.02	99.01
D3	D	27.58	39,392	1,115,066	729,165	0.65	667,101	624,035	93.54	169.14	99.02	98.95
D4	D	29.055	16,945	1,117,188	277,491	0.25	237,908	155,934	65.54	42.22	98.69	97.66
D5	D	32.915	1,876	1,955,298	219,077	0.11	162,471	28,011	17.24	7.47	70.29	29.16
D6	D	36.345	265	1,509,294	189,690	0.13	159,750	2,132	1.33	0.57	0.73	0.006

[a]C_T, cycle threshold based on *C. trachomatis ompA* RT-PCR (62). Cycle threshold is the cycle number at which the fluorescence generated within a reaction crosses the fluorescence threshold determined for the assay.
[b]With a quality score of ≥15.
[c]After quality control, i.e., deduplication and adapter trimming.
[d]Selected for *Chlamydia* spp. and *C. trachomatis* read pairs.
[e]With rRNA masked.

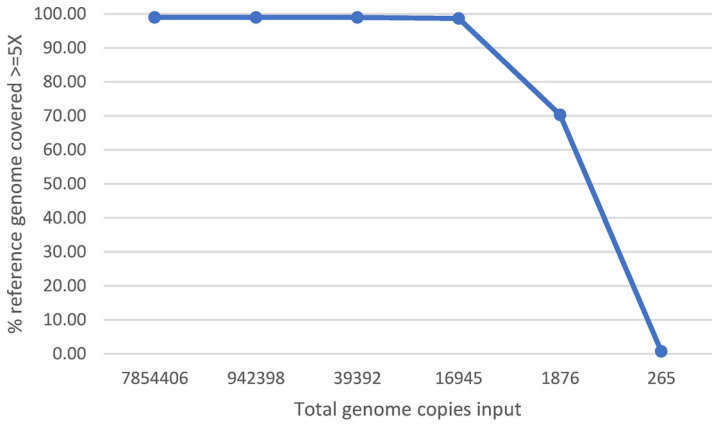


FIG 1 LOD of genome copies using the SureSelect^{XT} target enrichment workflow for spiked serial dilutions of reference strain D/UW-3/Cx gDNA. The data are the percent coverage of the reference genome for each serial dilution.

Phylogeny and single nucleotide polymorphisms among patient specimens. To ensure that genomes generated from the four patient specimen sets clustered to any of the four deep-branching monophyletic *C. trachomatis* lineages, a whole-genome phylogenetic analysis was constructed with 94 *C. trachomatis* single-contig genome sequences that were available from GenBank in December 2019 (Table S2) (40, 42, 54). Both the genomes from patient specimen sets 107 and 192 clustered within the "prevalent urogenital and rectal strain" clade, while patient specimen sets 98 and 72 clustered within the "nonprevalent urogenital and rectal strain" clade (Fig. 3).

Patient genomes 107 and 192 clustered in the same clade as the E genotype genomes (prevalent urogenital and anorectal lineage), whereas patient genomes 98 and 72 clustered in the same clade as G *ompA* genotypes (nonprevalent urogenital and anorectal lineage) (Fig. 3). For all four patients, the genomes derived from the two body sites formed distinct monophyletic clades within their respective urogenital lineage with a mean difference of 3 single nucleotide polymorphisms (SNPs) (1 to 5 SNPs), indicating that each patient likely carried the same strain within her rectum and vagina (Fig. 3; Table 3). Interestingly, five within-host SNPs were identified in patient 72, and 2 of these SNPs were within the highly recombinogenic *ompA* (2 SNPs) and *pmpF* (2 SNPs) genes. Genomic comparison of all the plasmids against all reference plasmid sequences showed that there was 100% sequence similarity within each patient specimen set (e.g., sample sets 107 and 192 had E plasmids in both anatomic sites) with the exception that the vaginal plasmid from patient 192 had a single nucleotide deletion at nucleotide position 5241 compared to the rectal plasmid. This deletion was within a gene that encodes a hypothetical protein. No other indels or SNPs were noted in any of the plasmids.

Detection of recombination events from genomes derived from patient specimens. Patient specimen sets harbored a total of 14 putative recombination blocks that contained between 21 and 419 homoplasic SNPs thought to have been introduced via homologous recombination. The number of putative recombination blocks varied between 1 and 6 within a patient specimen set, covering an average 12.2-kb region per specimen (Table S3). All the putative recombination blocks identified and described were shared within and detected only in each of the patient specimen sets, indicating that these recombination events were ancestral and acquired through clonal descent. Among the patient specimen sets, 107R and 107V (urogenital prevalent lineage) had the highest rates of recombination (ρ/θ = 0.111) as well as increased effects of recombination over point mutations (r/m = 5.777) followed by patient specimen sets 72R and 72V (urogenital nonprevalent lineage; ρ/θ = 0.053; r/m = 3.803). The lowest number of recombination events were observed among the

TABLE 2 Sequence data analysis of *C. trachomatis* gDNA extracted from spiked mock samples of genotype L_2b and patient specimen sets

Sample ID[a]	*ompA* genotype	Total no. of genome copies input	No. of:		Nonhuman read pair ratio	No. of:		% selected read pairs[b,c,d]	Mean read depth[e]	% of genome with coverage	
			Raw read pairs	Nonhuman read pairs		Total read pairs after QC[b,c]	Selected read pairs[b,c,d]			≥5×	≥10×
L1	L_2b	900,000	2,099,167	2,079,997	0.99	2,067,413	2,042,024	98.77	550.78	98.98	98.97
L2	L_2b	90,000	2,239,718	2,206,788	0.99	2,189,496	2,142,872	97.87	564.40	98.98	98.97
L3	L_2b	9,000	2,230,573	2,108,624	0.95	2,045,830	1,988,919	97.22	531.81	98.97	98.94
107R	Ja	1,378,275	2,302,152	2,051,954	0.89	1,924,399	705,294	36.65	190.25	98.91	98.89
107V	Ja	177,440	1,347,716	1,048,957	0.78	942,944	175,198	18.58	47.15	98.84	98.67
192R	Ja	3,336,780	2,175,904	1,937,403	0.89	1,793,255	1,184,836	66.07	313.09	98.92	98.91
192V	Ja	28,536	1,390,989	1,040,223	0.75	940,352	331,250	35.23	90.23	98.91	98.90
72R	G	121,472	1,764,699	1,652,600	0.94	1,429,554	53,096	3.71	13.99	96.11	74.60
72V	G	189,673	1,568,358	556,417	0.35	463,667	169,916	36.65	46.16	98.82	98.55
98R	G	2460,822	1,930,840	1,822,417	0.94	1,715,197	962,269	56.10	261.91	98.94	98.86
98V	G	20,764	1,259,385	856,804	0.68	791,372	307,678	38.88	84.58	98.88	98.64

[a]R, rectal swab; V, vaginal swab.
[b]With a quality score of ≥15.
[c]After quality control, i.e., deduplication and adapter trimming.
[d]Down-selected for *Chlamydia* spp. and *C. trachomatis* read pairs.
[e]With rRNA masked.

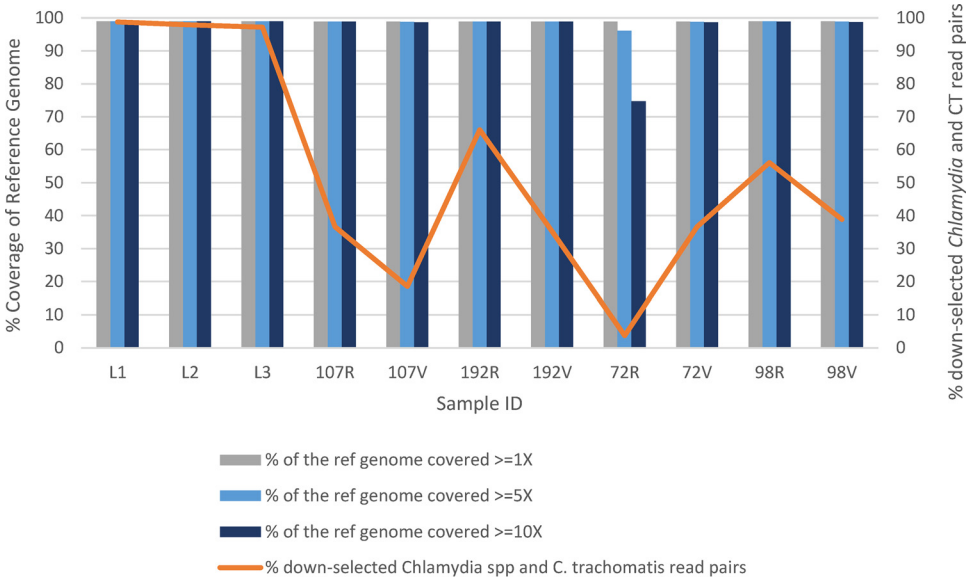


FIG 2 Percent coverage of reference genome and reads mapping to the respective reference genome for spiked mock samples of genotype L_2b and patient specimen sets. The percent coverage of the reference genome for each patient specimen is represented by bars, with bar coloring based on average mean read depth as indicated. R, rectal sample; V, vaginal sample. The orange line represents the percent down-selected *Chlamydia* spp. and *C. trachomatis* read pairs.

98V and 98R patient specimen sets (nonprevalent lineage; $\rho/\theta = 0.031$; $r/m = 0.6562$) (Table S3).

Some of the previously identified genomic regions of higher homologous recombination were also identified in this study. The *ompA* gene has undergone recombination in both prevalent urogenital patient specimen sets 107 and 192. *ompA* genotyping by both Sanger sequencing and whole-genome data indicated that the *ompA* genotype of specimens from patients 107 and 192 was Ja within an E genome backbone. Our recombination detection analysis showed that homologous recombination might have mediated the transfer of an ~12.9-kbp fragment containing CT_681/*ompA* along with the neighboring genes (from type III secretion system protein gene [CT_672] to *pbpB* [CT_682]) into patient 192 and a larger fragment of 17.6 kbp from CT_672 to *pbpB* (CT_682) along with *ompA* into patient 107, likely from a Ja strain (Fig. 3; Table S3). The *pmpE* and *mrsA_1* genes were estimated to be recombinant, respectively, in sets 107 and 192 (40, 54). The inclusion membrane protein gene *incD* was predicted to be recombinant only in the nonprevalent urogenital patient specimen set 98. (Table S3) (40, 54).

DISCUSSION

A SureSelect^{XT} workflow with a 2.698 Mbp RNA bait library with 34,795 120-mer probes was developed from 85 GenBank *C. trachomatis* reference genomes encompassing all four lineages of *C. trachomatis*, compared to the previous RNA bait library developed from 74 GenBank *C. trachomatis* reference genomes with 33,619 120-mer probes (52). By including more *C. trachomatis* reference genomes in our bait library design, which generated 1,000+ more probes, we likely captured more genetic diversity of the *C. trachomatis* strains circulating in the population than the RNA baits generated previously by Christiansen et al. (52). Moreover, having >98% of the *C. trachomatis* reference genome covered with at least 10 reads per nucleotide position for clinical specimens (with an exception for specimen 72R) indicates that the probes were uniformly covered along the *C. trachomatis* genome. Compared to a previous study that obtained 98% coverage with a total input of 4,800 *C. trachomatis* strain F/SW4 genome

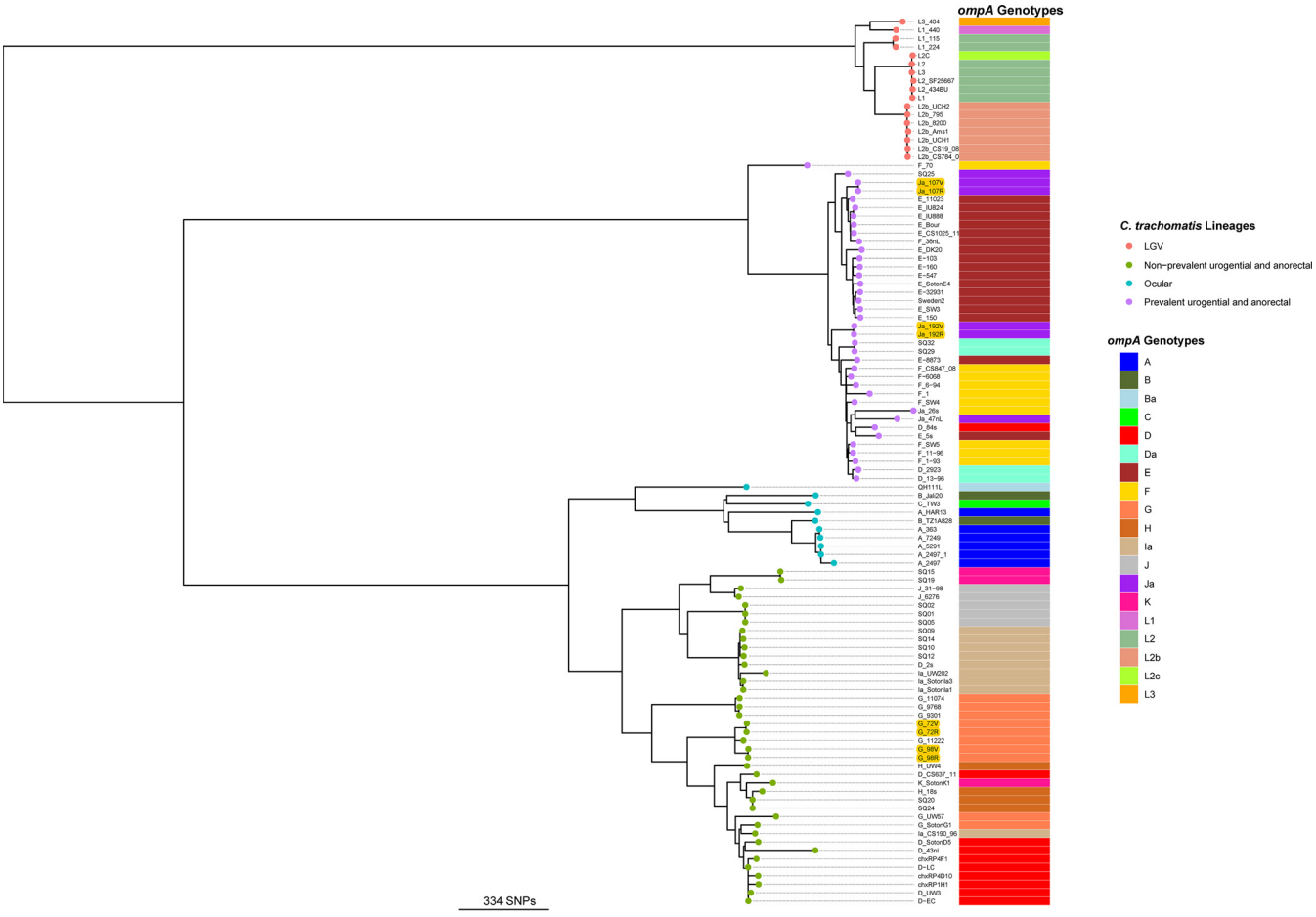


FIG 3 Global phylogeny of patient specimen sets and 94 *C. trachomatis* complete genomes. The four major lineages of *C. trachomatis* are highlighted with circular tip shapes in four distinct colors. The associated *ompA* genotypes for each genome derived from the whole-genome data are also shown with the color code on the right. Patient specimen sets are highlighted in yellow.

copies (52), we report an LOD of around 16,000 total genomes at 98% coverage for *C. trachomatis* reference strain D/UW-3/Cx. This LOD was within the same order of magnitude as that in the previous study, although we utilized an expanded RNA bait library, a different assay to determine copy number (i.e., quantitative real-time PCR [qRT-PCR] versus qPCR), and gDNA from a spiked sample versus a sample that had been propagated in tissue culture.

Here, it was useful to calculate comparable mean read depths and the number of down-selected read pairs among the spiked samples of L_2b, as this demonstrates success in enrichment for LGV strains, which have become prevalent among MSM and MSW (23, 25, 27, 28). Phylogenetic analysis of the 94 reference genomes (of which 85 genomes were used to develop the RNA bait library) showed genome representation from all four *C. trachomatis* lineages (Fig. 3). Overall, this workflow was successful in enrichment and sequencing of *C. trachomatis* strains that are prevalent in anorectal infections from two different populations: the MSM population, from which L_2b originated, and the female heterosexual population, from which the paired vaginal and rectal samples originated.

Sequence data analysis of patient specimen vaginal and rectal swabs sets revealed successful enrichment and genome sequencing of *C. trachomatis* from both clinical specimen types. In three of the four patient specimen sets, sequencing was more efficient for the rectal swabs, likely due to the higher genome copy number of *C. trachomatis* for those specimen sets. With a total input of 3,336,780 *C. trachomatis* genome copies, rectal specimen 192 demonstrated the best enrichment, with 89% nonhuman

TABLE 3 SNPs detected within patient specimen sets

Patient specimen set[a]	Gene containing within-host SNPs	No. of SNPs	Genome location(s) (SNP and nucleotide coverage for each variant)	Reference genome used for calling the SNPs[b]
107R and 107V	Gene for hypothetical protein (FSW4_RS00260)	1	54906 (C→T; T:62, C:1)	F/SW4 (NC_017951.1)
	Gene for hypothetical protein (FSW4_RS00270)	2	59520 (T→C; C:33, T:0), 59523 (C→T; T:33, C:0)	F/SW4 (NC_017951.1)
	Intergenic region	1	79002 (G→A; A:77, G:0)	F/SW4 (NC_017951.1)
	Gene for amino acid ABC transporter ATP-binding protein	1	146249 (C→T; T:145, C:0)	F/SW4 (NC_017951.1)
	Intergenic region	1	437237 (T→C; C:36, T:0)	F/SW4 (NC_017951.1)
192R and 192V	Gene for PTS sugar transporter subunit IIA	1	326567 (G→A; A:75, G:5)	F/SW4 (NC_017951.1)
72R and 72V	*ompA*	2	779230 (G→A; A:33, G:1), 779235 (C→A; A:31, G:2)	D/UW-3/CX (NC_000117.1)
	Gene for hypothetical protein (CT_744)	1	864903 (T→C; C:50, T→0)	D/UW-3/CX (NC_000117.1)
	pmpF	2	1029641 (G→T; T:75, G:0), 1029643 (A→G; G:75, A:0)	D/UW-3/CX (NC_000117.1)
98R and 98V	Gene for hypothetical protein (CT_049)	1	54281 (C→T; T:23, C:0)	D/UW-3/CX (NC_000117.1)

[a]R, rectal; V, vaginal.
[b]SNPs were called by comparing the genomes of patient specimen sets to reference genomes that clustered in the whole-genome phylogenetic tree.

reads, of which 66.07% belonged to *Chlamydia* species that mapped to 98.92% of the *C. trachomatis* reference E strain genome with at least 10 reads per nucleotide position and with a mean read depth coverage of 313.09. This is an improvement over a previous study that subjected clinical urine and vaginal samples to the prior Agilent bait library, where the best sample had 49.57% of the reads belonging to *Chlamydia* species, with 99.9% mapping to the reference D genome at a mean read depth of 410 from a total input of 68,864,400 *C. trachomatis* genome copies (52). Unfortunately, the number of reads mapped per nucleotide position to achieve the 99.9% coverage was not provided. Overall, enrichment of *C. trachomatis* from the samples in that study was successful in only eight (80%) of 10 samples, for a genome coverage of ≥95 to 100% for the respective reference genome; the percentage of reads mapping to *C. trachomatis* ranged from 0.07 to 49.57% across the specimens, possibly due to hybridization of the RNA bait library primers with human DNA. Without a reporting of the number of reads mapped per reference nucleotide position to achieve the genome coverage of ≥95 to 100%, we cannot make a direct comparison of the efficiencies of the bait library in this study compared to the previous *C. trachomatis* bait library.

In another study using the prior Agilent bait library, only 12 (60%) of 20 clinical ocular samples reached ≥95 to 100% genome coverage (55), again without reads mapped per reference nucleotide position. In our study, all eight (100%) clinical samples reached ≥95 to 100% genome coverage, with reads mapping to *C. trachomatis* ranging from 3.71 to 98.77%, indicating that this probe library—which further excludes baits with human homology—can achieve the desired efficiency.

The higher *C. trachomatis* bacterial load detected in rectal specimens in three of the four patient specimen sets conflicts with two studies that demonstrated similar loads across sets of vaginal and anorectal specimens collected from the same women with and without anal intercourse visiting an STI clinic in the Netherlands and in a high-HIV-prevalence area in South Africa (57, 58). For our study, the trend may be a characteristic of the assay used to determine copy number or the study population itself or, because of the small sample size, may not be representative of the study population as a whole. Nevertheless, Dirks et al. (59) pointed out the ineffectiveness of comparing load across *C. trachomatis* surveillance studies due to the lack of standardization for load determination and the presence of inflammatory cells which can artificially lower the number of *C. trachomatis* load. However, it is useful to determine the bacterial load in patient specimens to ensure that enough gDNA is present to successfully enrich for *C. trachomatis* in a target enrichment sequencing workflow, although conclusions should not be drawn from this estimated value about severity of infection, associated symptoms, or transmission.

Genomes derived from each patient specimen set were located phylogenetically within the representative lineages of their respective strain genotype as determined by Sanger sequencing, demonstrating success of the workflow and bioinformatic analysis described here. Although overall genome coverage was determined using conservative read depths (5× and 10×), a small number of SNPs were identified within patient sample sets at a range of 23 to 145× coverage, demonstrating low genome diversity within a single patient with a low probability of false positives and negatives. Interestingly, patient specimen sets 107 and 192 were associated with the prevalent strain lineage, a recent urogenital lineage that has been suggested to have been derived from recombination (40, 54). Further, the detection of recombination involving *ompA*, a known hot spot for recombination in the *C. trachomatis* genome, in these samples as part of an approximately 12-kb exchange event, in addition to four and five other recombination blocks, respectively, could be identified only by whole-genome sequencing (WGS) and not by Sanger sequencing alone or with traditional molecular typing methods such as *ompA* genotyping, MLST, or MLVA (40, 43, 54). In contrast, patient specimen sets 72 and 98 had three and one recombination blocks, respectively. These data highlight the genetic resolution achieved by our workflow. Indeed, the circulation of two different *C. trachomatis* lineages with various degrees of recombination across the genomes in a single population is of interest and displays the complexity of *C. trachomatis* strain evolution and transmission. Furthermore, the novel finding that vaginal and rectal pairs have the same genomes suggests within-host transmission, indicating the need for larger studies to further explore this possibility.

Here, we describe a higher-efficiency target enrichment bait library and a workflow that streamlines the molecular characterization of *C. trachomatis* from rectal as well as vaginal specimens. This type of high-resolution data can be used to understand the genetic diversity of current *C. trachomatis* strains causing genital and rectal infections and provide a robust molecular epidemiologic approach to advance our understanding of geo-sexual networks, outbreaks of colorectal infections among women and men who have sex with men, and the role of these strains in morbidity. The bioinformatic pipeline can further be used to potentially identify novel markers for typing *C. trachomatis* and to examine the microbiome to determine the role it plays in susceptibility, transmission and clearance of urogenital and rectal *C. trachomatis* infections, especially given the need for a longer duration of therapy for rectal infections than for most uncomplicated urogenital infections (60, 61). Finally, the bait library is publicly available, which will support comparative genome studies going forward.

MATERIALS AND METHODS

Spiked serial dilutions of *C. trachomatis* reference strain D/UW-3/Cx. Genomic DNA (gDNA) from *C. trachomatis* reference strain genotype D/UW-3/Cx (ATCC VR-885D) was used to determine the limit of detection (LOD) for this workflow. The LOD was set at the minimal genome copy number required to generate a ≥5× read depth with ≥98% genome coverage compared to the reference strain of the same *ompA* genotype. Six 100-μl serial dilutions (10^{-1} to 10^{-6}) were prepared by spiking into 1× phosphate-buffered saline (PBS). A standard curve based on ATCC's reported copy number for genotype L_2b (ATTC VR-902BD) was generated using real-time PCR (RT-PCR) targeting the *C. trachomatis* single-copy polymorphic membrane protein gene *pmpH* (62). This standard curve ($y = 1 \times 10^{e(-0.602x)}$; $R^2 = 0.987$) was then used to calculate a more precise genome copy number for each serial dilution of genotype D.

Spiked mock samples of *C. trachomatis* genotype L_2b. gDNA from *C. trachomatis* clinical strain genotype L_2b (ATCC VR-902BD) was used to ensure success of this workflow with a *C. trachomatis* strain prevalent in anorectal infections in HIV-infected MSM. Three 100-μl serial dilutions (9,000 to 900,000 total genome copies) were prepared by spiking into 1× PBS, all within the LOD established using *C. trachomatis* reference strain genotype D.

***C. trachomatis* clinical specimens and determination of *C. trachomatis* genome copy number.** Clinical urogenital and rectal samples were obtained from women aged 18 to 40 years who were at high risk for STIs after giving informed consent as part of a separate study that was approved by the institutional review board of UCSF Benioff Children's Hospital Oakland Research Institute. For this study, the samples were stripped of all personal identifiers with no trace to the patient names. FLOQswab vaginal and rectal swabs (Copan, Murrieta, CA) had been collected using standard techniques by trained clinic staff and screened for *C. trachomatis* using the Xpert CT/NG test (Cepheid, Sunnyvale, CA). Four clinical vaginal samples and the four paired rectal swabs from the same four women (randomly selected using a

table of random numbers from over 200 women positive for *C. trachomatis* at both sites) were used in this study.

Approximately 200 μl of remnant swab collection buffer that had not been run in the Xpert test was lysed with 59 μl of a cocktail consisting of 50 μl lysozyme (10 mg/ml; MilliporeSigma, St. Louis, MO), 3 μl of lysostaphin (4,000 U/ml in sodium acetate; MilliporeSigma), and 6 μl of mutanolysin (25,000 U/ml; MilliporeSigma) for 1 h at 37°C as described elsewhere (63). gDNA was then purified from the lysate using the QIAamp DNA minikit (Qiagen) according to the manufacturer's instructions. For rectal swabs collected in M4 medium (Thermo Fisher, South San Francisco, CA), 200 μl was treated as described above.

gDNA was subjected to an in-house qPCR to quantitate the genomic copy number of *C. trachomatis* as we described previously (64). Briefly, a standard curve was calculated based on 10-fold serial dilutions of a linearized plasmid containing the single-copy *ompA* gene. *C. trachomatis* genomic copy number of the clinical samples was determined based on the standard curve.

***C. trachomatis ompA* genotyping and plasmid sequencing.** *ompA* genotyping was performed as previously described (65). Briefly, primers flanking the *ompA* gene were used for PCR. The PCR product was purified by exoSAP-IT (Thermo Fisher) and subjected to Sanger sequencing using the PCR primers (65). Forward and reverse sequences were aligned using MAFFT v7.450 (66) to create a consensus sequence that was aligned to all reference *ompA* genotypes. The reference strains included A/HAR-13, B/TW-5/OT, Ba/Apache-2, C/TW-3/OT, D/UW-3/Cx, Da/TW-448, E/Bour, F/IC-Cal-13, G/UW-57/Cx, H/UW-4/Cx, I/UW-12/Ur, Ia/UW-202, J/UW-36/Cx, Ja/UW-92, K/UW-31/Cx, L_1/440, L_2/434, L_2a/UW-396, L_2b/UCH-1/proctitis, L_2c, and L_3/404.

Five overlapping PCR primer pairs were designed using the IDT PrimerQuest tool (https://www.idtdna.com/pages/tools/primerquest) to amplify the entire plasmid (Table S1). The thermocycling parameters were 3 min at 95°C followed by 40 cycles of 95°C for 30 s, 56°C for 30 s, and 72°C for 1 min 10 s with a final incubation at 72°C for 7 min. The PCR product was purified and sequenced as described above, and the consensus sequence was aligned to all reference plasmid sequences as described above using MAFFT v7.450.

Quantification and fragmentation. Samples were quantified using a Qubit 2.0 fluorometer, and human gDNA (Promega, San Luis Obispo, CA) was added to reach a total input of 3 μg/130 μl for fragmentation and library prep. Samples were sheared on a Covaris LE220-plus instrument using the 8 microTUBE strip V1 (PN 520053; Covaris, Woburn, MA) with the base pair mode set to 250 to 300 bp following the manufacturer's instructions.

RNA bait library design. A 2.698 Mbp RNA bait library consisting of 34,795 120-mer probes spanning 85 GenBank *C. trachomatis* reference genomes was designed using Agilent SureDesign. No plasmid probes were included in the RNA bait library construction. The bait library was synthesized by Agilent Technologies. Although Agilent Technologies prevents publication of the probe sequences for the RNA bait libraries they design, the custom-designed RNA bait library (ELID 3173001) used in this study can be retrieved by contacting Agilent Technologies, Inc. (Santa Clara, CA). The sequences for the bait library developed by Christiansen et al. (52) and the bait library itself are not publicly available, resulting in the inability to compare it with the RNA bait library designed in this study. Sequencing of the RNA bait library itself was not necessary, as Agilent provided the probe sequences, which were analyzed using BLAST to determine that they represented all *C. trachomatis* genovars.

Library prep. After shearing, the SureSelect^{XT} target enrichment system for Illumina paired-end multiplexed sequencing library (VC2; December 2018) and all recommended quality control steps were performed on all gDNA samples. A 16-h incubation at 65°C was performed for RNA bait library hybridization. Postcapture PCR cycling was set at 12 cycles based on a capture library size of >1.5 Mb.

Illumina MiSeq sequencing. The eight clinical samples were multiplexed for two runs of paired end sequencing on an Illumina MiSeq instrument using a 300-bp v2 reagent kit. For the final multiplexed library pool, libraries were diluted to 2 nM/3 μl in low-EDTA Tris-EDTA buffer (TE) for a final concentration of 10 pM; 12.5 pM PhiX was added to the final pool that was loaded onto the MiSeq sequencer.

Sequence and phylogenetics analysis. Host genome sequences were first filtered from the raw sequencing data set using Bowtie2 version 2.2.9 (67), which removed any contaminating human sequences using the h19 human reference genome (68). Cutadapt version 1.8.3 (69) was used to trim specified primers and adapters and to filter out reads below Phred quality scores of 15 and read length below 50 bp. Deduplication of the reads was performed using Clumpify (sourceforge.net/projects/bbmap/) with the dedupe=t option to prevent biased coverage of genomic regions. *C. trachomatis* sequencing reads were selected using K-SLAM (70), a k-mer-based metagenomics taxonomic profiler, which used a database containing all bacterial and archaeal reference nucleotide sequences. The presence of *C. trachomatis* sequences was also confirmed using Metaphlan2 (71). We generated a custom version of the *C. trachomatis* D/UW-3/CX reference sequence (NC_000117.1) from which we masked 6 rRNA genes, CT_r01 (16SrRNA_1), CT_r02 (23SrRNA_1), CT_r03 (5SrRNA_1), CT_r04 (16SrRNA_2), CT_r05 (23SrRNA_2), and CT_r06 (5SrRNA_2), present in the repeated rRNA operons using the bedtools v2.17.0 (72) tool "maskfasta." Prefiltered chlamydial reads were mapped against this custom reference genome using BWA mem v2.12.0 (73) (MapQ $\geq$ 20), followed by consensus sequence generation and estimation of sequencing depth and mapping statistics using SAMtools (74) (options "depth" and "mpileup") and bcftools v1.9. The prefiltered *C. trachomatis* sequencing reads were also used to generate *de novo* short-read assemblies using SPAdes 3.7.0 (75) with the "careful" option. To genotype the patient samples, *de novo* contigs were used to extract and compare the *ompA* genes against a customized BLAST (76) database of the 21 reference *ompA* sequences (see above). The equation used to calculate mean read depth was: (number of mapped reads $\times$ average bp read length)/(bp length of CT reference genome).

For phylogenetic analysis, apart from the patient genome sequences ($n = 8$), we also included all *C. trachomatis* genomes (without plasmid sequences) available in NCBI ($n = 94$) and used a reference mapping approach with the above-mentioned custom version of the *C. trachomatis* D/UW-3/CX reference genome sequence. In short, full-length whole-genome alignments were generated using Snippy v3.1 (https://github.com/tseemann/snippy), which identifies variants using Freebayes v1.0.2 (77) with a minimum 10× read coverage and 90% read concordance at a locus for each SNP. Regions of increased density of homoplasious SNPs introduced by possible recombination events were predicted iteratively and masked using Gubbins (78). The final phylogenetic tree was reconstructed using RaxML (79) on the recombination removed alignment using the general time-reversible (GTR) model. The genes located within the putative recombination blocks for the patient samples were identified by comparing the alignment genomic coordinates for the predicted recombination blocks to the gene annotations of the reference genome. Within-host SNP differences were derived from the alignment before masking the predicted recombination events.

Data availability. All sequencing data associated with this study were submitted to the National Center for Biotechnology Information's sequence read archive (SRA) under the BioProject accession ID PRJNA609714.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.2 MB.

FIG S2, PDF file, 0.2 MB.

TABLE S1, XLSX file, 0.01 MB.

TABLE S2, XLSX file, 0.01 MB.

TABLE S3, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

This work was made possible through support from CDC's Advanced Molecular Detection (AMD) program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We thank Sankhya Bomanna for excellent technical assistance.

K. E. Bowden transcribed the manuscript and generated the genomic libraries from the serial dilutions and clinical specimens for sequencing and analysis. S. J. Joseph performed the bioinformatics analysis, conducted phylogenetic analysis of the genomic data, and contributed to the writing of the manuscript. J. C. Cartee pooled and sequenced the genomic libraries on the MiSeq system. N. Ziklo prepared and purified the gDNA from the clinical samples and performed the qPCR and fragmentation for library preparation. D. Danavall prepared spiked gDNA serial dilutions for LOD determination. B. H. Raphael oversaw the project within CDC and provided technical support and subject matter expertise on the genome sequencing workflow and bioinformatic analysis. T. D. Read conducted initial phylogenetic analysis while providing continued technical support and subject matter expertise for the data analysis. D. Dean collected and provided the patient specimens and provided technical support and subject matter expertise on the genome sequencing workflow and contributed to the writing of the manuscript. All authors reviewed, edited, and contributed to the manuscript.

We declare no competing interests.

REFERENCES

1. Centers for Disease Control and Prevention. 2014. Recommendations for the laboratory-based detection of *Chlamydia trachomatis* and Neisseria gonorrhoeae–2014. MMWR Recomm Rep 63:1–19.
2. Stamm WE. 2011. *Chlamydia trachomatis* infections of the adult. *In* Holmes KK (ed), Sexually transmitted diseases, 4th ed, p 575–606. McGraw Hill Professional, New York, NY.
3. 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. https://doi.org/10.1093/infdis/166.2.383.
4. Isaksson J, Carlsson O, Airell A, Stromdahl S, Bratt G, Herrmann B. 2017. Lymphogranuloma venereum rates increased and *Chlamydia trachomatis* genotypes changed among men who have sex with men in Sweden 2004-2016. J Med Microbiol 66:1684–1687. https://doi.org/10.1099/jmm.0.000597.
5. Spaargaren J, Fennema HS, Morre SA, de Vries HJ, Coutinho RA. 2005. New lymphogranuloma venereum *Chlamydia trachomatis* variant,

Amsterdam. Emerg Infect Dis 11:1090–1092. https://doi.org/10.3201/eid1107.040883.
6. 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. https://doi.org/10.3201/eid1111.050821.
7. Christerson L, de Vries HJ, de Barbeyrac B, Gaydos CA, Henrich B, Hoffmann S, Schachter J, Thorvaldsen J, Vall-Mayans M, Klint M, Herrmann B, Morre SA. 2010. Typing of lymphogranuloma venereum *Chlamydia trachomatis* strains. Emerg Infect Dis 16:1777–1779. https://doi.org/10.3201/eid1611.100379.
8. Versteeg B, van Rooijen MS, Schim van der Loeff MF, de Vries HJC, Bruisten SM. 2014. No indication for tissue tropism in urogenital and anorectal *Chlamydia trachomatis* infections using high-resolution multilocus sequence typing. BMC Infect Dis 14:464. https://doi.org/10.1186/1471-2334-14-464.
9. Bom RJ, van der Helm JJ, Schim van der Loeff MF, van Rooijen MS, Heijman T, Matser A, de Vries HJ, Bruisten SM. 2013. Distinct transmission networks of *Chlamydia trachomatis* in men who have sex with men and heterosexual adults in Amsterdam, The Netherlands. PLoS One 8:e53869. https://doi.org/10.1371/journal.pone.0053869.
10. Lysen M, Osterlund A, Rubin CJ, Persson T, Persson I, Herrmann B. 2004. Characterization of ompA genotypes by sequence analysis of DNA from all detected cases of *Chlamydia trachomatis* infections during 1 year of contact tracing in a Swedish County. J Clin Microbiol 42:1641–1647. https://doi.org/10.1128/jcm.42.4.1641-1647.2004.
11. 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. https://doi.org/10.1186/1471-2334-11-63.
12. Bax CJ, Quint KD, Peters RP, Ouburg S, Oostvogel PM, Mutsaers JA, Dorr PJ, Schmidt S, Jansen C, van Leeuwen AP, Quint WG, Trimbos JB, Meijer CJ, Morre SA. 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. https://doi.org/10.1136/sti.2010.048173.
13. Dewart CM, Bernstein KT, DeGroote NP, Romaguera R, Turner AN. 2018. Prevalence of rectal chlamydial and gonococcal infections: a systematic review. Sex Transm Dis 45:287–293. https://doi.org/10.1097/OLQ.0000000000000754.
14. Andersson N, Boman J, Nylander E. 2017. Rectal chlamydia—should screening be recommended in women? Int J STD AIDS 28:476–479. https://doi.org/10.1177/0956462416653510.
15. Waalboer R, van der Snoek EM, van der Meijden WI, Mulder PG, Ossewaarde JM. 2006. Analysis of rectal *Chlamydia trachomatis* serovar distribution including L2 (lymphogranuloma venereum) at the Erasmus MC STI clinic, Rotterdam. Sex Transm Infect 82:207–211. https://doi.org/10.1136/sti.2005.018580.
16. Morre SA, Rozendaal L, van Valkengoed IG, Boeke AJ, van Voorst Vader PC, Schirm J, de Blok S, van Den Hoek JA, van Doornum GJ, Meijer CJ, van Den Brule AJ. 2000. Urogenital *Chlamydia trachomatis* serovars in men and women with a symptomatic or asymptomatic infection: an association with clinical manifestations? J Clin Microbiol 38:2292–2296. https://doi.org/10.1128/sti.38.6.2292-2296.2000.
17. 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. https://doi.org/10.1128/JCM.00574-06.
18. Koper NE, van der Sande MA, Gotz HM, Koedijk FD. 2013. Lymphogranuloma venereum among men who have sex with men in the Netherlands: regional differences in testing rates lead to underestimation of the incidence, 2006-2012. Euro Surveill 18:20561. https://doi.org/10.2807/1560-7917.es2013.18.34.20561.
19. Ward H, Alexander S, Carder C, Dean G, French P, Ivens D, Ling C, Paul J, Tong W, White J, Ison CA. 2009. The prevalence of lymphogranuloma venereum infection in men who have sex with men: results of a multicentre case finding study. Sex Transm Infect 85:173–175. https://doi.org/10.1136/sti.2008.035311.
20. Stark D, van Hal S, Hillman R, Harkness J, Marriott D. 2007. Lymphogranuloma venereum in Australia: anorectal *Chlamydia trachomatis* serovar L2b in men who have sex with men. J Clin Microbiol 45:1029–1031. https://doi.org/10.1128/JCM.02389-06.
21. Ronn MM, Ward H. 2011. The association between lymphogranuloma venereum and HIV among men who have sex with men: systematic review and meta-analysis. BMC Infect Dis 11:70. https://doi.org/10.1186/1471-2334-11-70.
22. Chan PA, Robinette A, Montgomery M, Almonte A, Cu-Uvin S, Lonks JR, Chapin KC, Kojic EM, Hardy EJ. 2016. Extragenital infections caused by *Chlamydia trachomatis* and *Neisseria gonorrhoeae*: a review of the literature. Infect Dis Obstet Gynecol 2016:5758387. https://doi.org/10.1155/2016/5758387.
23. van Liere GA, van Rooijen MS, Hoebe CJ, Heijman T, de Vries HJ, Dukers-Muijrers NH. 2015. Prevalence of and factors associated with rectal-only chlamydia and gonorrhoea in women and in men who have sex with men. PLoS One 10:e0140297. https://doi.org/10.1371/journal.pone.0140297.
24. Chandra NL, Broad C, Folkard K, Town K, Harding-Esch EM, Woodhall SC, Saunders JM, Sadiq ST, Dunbar JK. 2018. Detection of *Chlamydia trachomatis* in rectal specimens in women and its association with anal intercourse: a systematic review and meta-analysis. Sex Transm Infect 94:320–326. https://doi.org/10.1136/sextrans-2017-053161.
25. Danby CS, Cosentino LA, Rabe LK, Priest CL, Damare KC, Macio IS, Meyn LA, Wiesenfeld HC, Hillier SL. 2016. Patterns of extragenital chlamydia and gonorrhea in women and men who have sex with men reporting a history of receptive anal intercourse. Sex Transm Dis 43:105–109. https://doi.org/10.1097/OLQ.0000000000000384.
26. Kong FY, Tabrizi SN, Law M, Vodstrcil LA, Chen M, Fairley CK, Guy R, Bradshaw C, Hocking JS. 2014. Azithromycin versus doxycycline for the treatment of genital chlamydia infection: a meta-analysis of randomized controlled trials. Clin Infect Dis 59:193–205. https://doi.org/10.1093/cid/ciu220.
27. Foschi C, Gaspari V, Sgubbi P, Salvo M, D'Antuono A, Marangoni A. 2018. Sexually transmitted rectal infections in a cohort of 'men having sex with men'. J Med Microbiol 67:1050–1057. https://doi.org/10.1099/jmm.0.000781.
28. de Vrieze NH, de Vries HJ. 2014. Lymphogranuloma venereum among men who have sex with men. An epidemiological and clinical review. Expert Rev Anti Infect Ther 12:697–704. https://doi.org/10.1586/14787210.2014.901169.
29. Bachmann LH, Johnson RE, Cheng H, Markowitz L, Papp JR, Palella FJ, Hook EW. 2010. Nucleic acid amplification tests for diagnosis of *Neisseria gonorrhoeae* and *Chlamydia trachomatis* rectal infections. J Clin Microbiol 48:1827–1832. https://doi.org/10.1128/JCM.02398-09.
30. Labiran C, Marsh P, Zhou J, Bannister A, Clarke IN, Goubet S, Soni S. 2016. Highly diverse MLVA-ompA genotypes of rectal *Chlamydia trachomatis* among men who have sex with men in Brighton, UK and evidence for an HIV-related sexual network. Sex Transm Infect 92:299–304. https://doi.org/10.1136/sextrans-2015-052261.
31. Labiran C, Rowen D, Clarke IN, Marsh P. 2017. Detailed molecular epidemiology of *Chlamydia trachomatis* in the population of Southampton attending the genitourinary medicine clinic in 2012-13 reveals the presence of long established genotypes and transitory sexual networks. PLoS One 12:e0185059. https://doi.org/10.1371/journal.pone.0185059.
32. Dean D, Bruno WJ, Wan R, Gomes JP, Devignot S, Mehari T, de Vries HJ, Morre SA, Myers G, Read TD, Spratt BG. 2009. Predicting phenotype and emerging strains among *Chlamydia trachomatis* infections. Emerg Infect Dis 15:1385–1394. https://doi.org/10.3201/eid1509.090272.
33. Kapil R, Press CG, Hwang ML, Brown L, Geisler WM. 2015. Investigating the epidemiology of repeat *Chlamydia trachomatis* detection after treatment by using *C. trachomatis* OmpA genotyping. J Clin Microbiol 53:546–549. https://doi.org/10.1128/JCM.02483-14.
34. Herrmann B, Isaksson J, Ryberg M, Tangrot J, Saleh I, Versteeg B, Gravningen K, Bruisten S. 2015. Global multilocus sequence type analysis of *Chlamydia trachomatis* strains from 16 countries. J Clin Microbiol 53:2172–2179. https://doi.org/10.1128/JCM.00249-15.
35. Pannekoek Y, Morelli G, Kusecek B, Morre SA, Ossewaarde JM, Langerak AA, van der Ende A. 2008. Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria *Chlamydia trachomatis*. BMC Microbiol 8:42. https://doi.org/10.1186/1471-2180-8-42.
36. Bom RJ, Christerson L, Schim van der Loeff MF, Coutinho RA, Herrmann B, Bruisten SM. 2011. Evaluation of high-resolution typing methods for *Chlamydia trachomatis* in samples from heterosexual couples. J Clin Microbiol 49:2844–2853. https://doi.org/10.1128/JCM.00128-11.
37. Klint M, Fuxelius HH, Goldkuhl RR, Skarin H, Rutemark C, Andersson SG, Persson K, Herrmann B. 2007. High-resolution genotyping of *Chlamydia trachomatis* strains by multilocus sequence analysis. J Clin Microbiol 45:1410–1414. https://doi.org/10.1128/JCM.02301-06.
38. Lan J, Walboomers JM, Roosendaal R, van Doornum GJ, MacLaren DM, Meijer CJ, van den Brule AJ. 1993. Direct detection and genotyping of *Chlamydia trachomatis* in cervical scrapes by using polymerase chain reaction and restriction fragment length polymorphism analysis. J Clin Microbiol 31:1060–1065. https://doi.org/10.1128/JCM.31.5.1060-1065.1993.

39. Smelov V, Vrbanac A, van Ess EF, Noz MP, Wan R, Eklund C, Morgan T, Shrier LA, Sanders B, Dillner J, de Vries HJC, Morre SA, Dean D. 2017. *Chlamydia trachomatis* strain types have diversified regionally and globally with evidence for recombination across geographic divides. Front Microbiol 8:2195. https://doi.org/10.3389/fmicb.2017.02195.
40. Harris SR, Clarke IN, Seth-Smith HM, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW, Lewis DA, Spratt BG, Unemo M, Persson K, Bjartling C, Brunham R, de Vries HJ, Morre SA, Speksnijder A, Bebear CM, Clerc M, de Barbeyrac B, Parkhill J, Thomson NR. 2012. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet 44:413–419. https://doi.org/10.1038/ng.2214.
41. 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. https://doi.org/10.1186/1745-6150-6-28.
42. Joseph SJ, Read TD. 2012. Genome-wide recombination in *Chlamydia trachomatis*. Nat Genet 44:364–366. https://doi.org/10.1038/ng.2225.
43. 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. https://doi.org/10.1101/gr.5674706.
44. 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:e00045-11. https://doi.org/10.1128/mBio.00045-11.
45. Castro R, Baptista T, Vale A, Nunes H, Prieto E, Araujo C, Mansinho K, da Luz Martins Pereira F. 2010. Lymphogranuloma venereum serovar L2b in Portugal. Int J STD AIDS 21:265–266. https://doi.org/10.1258/ijsa.2009.009134.
46. Martin-Iguacel R, Llibre JM, Nielsen H, Heras E, Matas L, Lugo R, Clotet B, Sirera G. 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. https://doi.org/10.1007/s10096-010-0959-2.
47. Borges V, Cordeiro D, Salas Al, Lodhia Z, Correia C, Isidro J, Fernandes C, Rodrigues AM, Azevedo J, Alves J, Roxo J, Rocha M, Corte-Real R, Vieira L, Borrego MJ, Gomes JP. 2019. Chlamydia trachomatis: when the virulence-associated genome backbone imports a prevalence-associated major antigen signature. Microb Genom 5:e000313. https://doi.org/10.1099/mgen.0.000313.
48. Taylor-Brown A, Madden D, Polkinghorne A. 2018. Culture-independent approaches to chlamydial genomics. Microb Genom 4:e000145. https://doi.org/10.1099/mgen.0.000145.
49. Putman TE, Suchland RJ, Ivanovitch JD, Rockey DD. 2013. Culture-independent sequence analysis of *Chlamydia trachomatis* in urogenital specimens identifies regions of recombination and in-patient sequence mutations. Microbiology (Reading) 159:2109–2117. https://doi.org/10.1099/mic.0.070029-0.
50. Hedrum A, Lundeberg J, Pahlson C, Uhlen M. 1992. Immunomagnetic recovery of *Chlamydia trachomatis* from urine with subsequent colorimetric DNA detection. PCR Methods Appl 2:167–171. https://doi.org/10.1101/gr.2.2.167.
51. Seth-Smith HM, Harris SR, Skilton RJ, Radebe FM, Golparian D, Shipitsyna E, Duy PT, Scott P, Cutcliffe LT, O'Neill C, Parmar S, Pitt R, Baker S, Ison CA, Marsh P, Jalal H, Lewis DA, Unemo M, Clarke IN, Parkhill J, Thomson NR. 2013. Whole-genome sequences of *Chlamydia trachomatis* directly from clinical samples without culture. Genome Res 23:855–866. https://doi.org/10.1101/gr.150037.112.
52. Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR, Holdstock J, Holland MJ, Stevenson S, Dave J, Tong CY, Einer-Jensen K, Depledge DP, Breuer J. 2014. Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from clinical samples. BMC Infect Dis 14:591. https://doi.org/10.1186/s12879-014-0591-3.
53. Last AR, Pickering H, Roberts CH, Coll F, Phelan J, Burr SE, Cassama E, Nabicassa M, Seth-Smith HMB, Hadfield J, Cutcliffe LT, Clarke IN, Mabey DCW, Bailey RL, Clark TG, Thomson NR, Holland MJ. 2018. Population-based analysis of ocular *Chlamydia trachomatis* in trachoma-endemic West African communities identifies genomic markers of disease severity. Genome Med 10:15. https://doi.org/10.1186/s13073-018-0521-x.
54. Hadfield J, Harris SR, Seth-Smith HMB, Parmar S, Andersson P, Giffard PM, Schachter J, Moncada J, Ellison L, Vaulet MLG, Fermepin MR, Radebe F, Mendoza S, Ouburg S, Morre SA, Sachse K, Puolakkainen M, Korhonen SJ, Sonnex C, Wiggins R, Jalal H, Brunelli T, Casprini P, Pitt R, Ison C, Savicheva A, Shipitsyna E, Hadad R, Kari L, Burton MJ, Mabey D, Solomon AW, Lewis D, Marsh P, Unemo M, Clarke IN, Parkhill J, Thomson NR. 2017. Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. Genome Res 27:1220–1229. https://doi.org/10.1101/gr.212647.116.
55. Alkhidir AAl, Holland MJ, Elhag WI, Williams CA, Breuer J, Elemam AE, El Hussain KMK, Ournasseir MEH, Pickering H. 2019. Whole-genome sequencing of ocular *Chlamydia trachomatis* isolates from Gadarif State, Sudan. Parasit Vectors 12:518. https://doi.org/10.1186/s13071-019-3770-7.
56. Pickering H, Chernet A, Sata E, Zerihun M, Williams CA, Breuer J, Nute AW, Haile M, Zeru T, Tadesse Z, Bailey RL, Callahan EK, Holland MJ, Nash SD. 2020. Genomics of ocular *Chlamydia trachomatis* after 5 years of SAFE interventions for trachoma in Amhara, Ethiopia. Journal Infecti Dis jiaa615. https://doi.org/10.1093/infdis/jiaa615.
57. Dirks JAMC, van Liere GAFS, Hoebe CJPA, Wolffs P, Dukers-Muijrers NHTM. 2019. Genital and anal *Chlamydia trachomatis* bacterial load in concurrently infected women: a cross-sectional study. Sex Transm Infect 95:317–321. https://doi.org/10.1136/sextrans-2018-053678.
58. Dubbink JH, de Waaij DJ, Bos M, van der Eem L, Bébéar C, Mbambazela N, Ouburg S, Peters RPH, Morré SA. 2016. Microbiological characteristics of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections in South African women. J Clin Microbiol 54:200–203. https://doi.org/10.1128/JCM.02848-15.
59. Dirks JAMC, Hoebe CJPA, van Liere GAFS, Dukers-Muijrers NHTM, Wolffs PFG. 2019. Standardisation is necessary in urogenital and extragenital Chlamydia trachomatis bacterial load determination by quantitative PCR: a review of literature and retrospective study. Sex Transm Infect 95:562–568. https://doi.org/10.1136/sextrans-2018-053522.
60. Dukers-Muijrers NH, Speksnijder AG, Morre SA, Wolffs PF, van der Sande MA, Brink AA, van den Broek IV, Werner MI, Hoebe CJ. 2013. Detection of anorectal and cervicovaginal *Chlamydia trachomatis* infections following azithromycin treatment: prospective cohort study with multiple time-sequential measures of rRNA, DNA, quantitative load and symptoms. PLoS One 8:e81236. https://doi.org/10.1371/journal.pone.0081236.
61. Kong FY, Hocking JS. 2015. Treatment challenges for urogenital and anorectal *Chlamydia trachomatis*. BMC Infect Dis 15:293. https://doi.org/10.1186/s12879-015-1030-9.
62. Chen CY, Chi KH, Alexander S, Ison CA, Ballard RC. 2008. A real-time quadriplex PCR assay for the diagnosis of rectal lymphogranuloma venereum and non-lymphogranuloma venereum *Chlamydia trachomatis* infections. Sex Transm Infect 84:273–276. https://doi.org/10.1136/sti.2007.029058.
63. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108:4680–4687. https://doi.org/10.1073/pnas.1002611107.
64. Sharma M, Recuero-Checa MA, Fan FY, Dean D. 2018. *Chlamydia trachomatis* regulates growth and development in response to host cell fatty acid availability in the absence of lipid droplets. Cell Microbiol 20:cmi.12801. https://doi.org/10.1111/cmi.12801.
65. Batteiger BE, Wan R, Williams JA, He L, Ma A, Fortenberry JD, Dean D. 2014. Novel *Chlamydia trachomatis* strains in heterosexual sex partners, Indianapolis, Indiana, USA. Emerg Infect Dis 20:1841–1847. https://doi.org/10.3201/2011.140604.
66. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
67. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
68. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933. https://doi.org/10.1038/35057149.
69. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet J 17:10. https://doi.org/10.14806/ej.17.1.200.
70. Ainsworth D, Sternberg MJE, Raczy C, Butcher SA. 2017. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. Nucleic Acids Res 45:1649–1656. https://doi.org/10.1093/nar/gkw1248.
71. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic

taxonomic profiling. Nat Methods 12:902–903. https://doi.org/10.1038/nmeth.3589.
72. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
73. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv
74. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
75. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
76. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
77. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv.
78. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196.
79. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

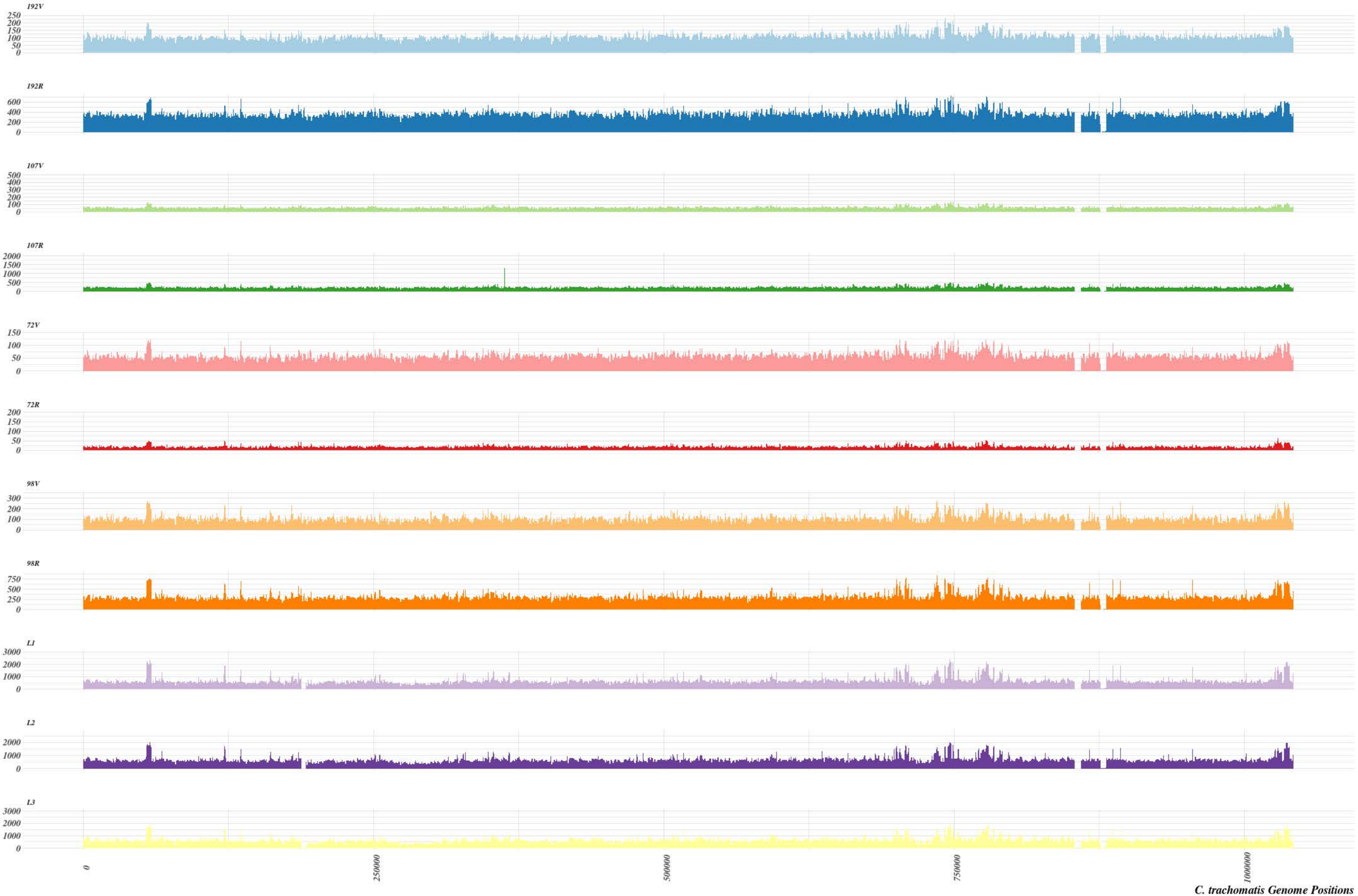
Supplemental Material

- **FIG S1**

  Coverage plots for reference strain D/UW-3/Cx. All the rRNA genes were masked before mapping the reads in order to avoid nonspecific read mapping of the reads because of the usage of RNA bait capture technology for enrichment and direct sequencing. The two distinct gaps seen are the two repeated 23S rRNA genes, and that gap is visible because the lengths of the 23S RNA genes are greater than those of 5S and 16S rRNAs. Download FIG S1, PDF file, 0.2 MB.

- **FIG S2**

  Coverage plots for L_2b strain and patient specimen sets. All the rRNA genes were masked before mapping the reads in order to avoid nonspecific read mapping of the reads because of the usage of RNA bait capture technology for enrichment and direct sequencing. The two distinct gaps seen are the two repeated 23S rRNA genes, and that gap is visible because the lengths of the 23S RNA genes are higher than those of 5S and 16S rRNAs. Download FIG S2, PDF file, 0.2 MB.


192V
250
200
150
100
50
0
192R
600
400
200
0
107V
500
400
300
200
100
0
107R
2000
1500
1000
500
0
72V
150
100
50
0
72R
200
150
100
50
0
98V
300
200
100
0
98R
750
500
250
0
L1
3000
2000
1000
0
L2
2000
1000
0
L3
3000
2000
1000
0
0
250000
500000
750000
1000000
C. trachomatis Genome Positions


D1
2000
1500
1000
500
0
0
250000
500000
750000
1000000
D2
1200
800
400
0
0
250000
500000
750000
1000000
D3
600
400
200
0
0
250000
500000
750000
1000000
e Depth
D4
150
100
50
0
0
250000
500000
750000
1000000
D5
40
30
20
10
0
0
250000
500000
750000
1000000
D6
12.5
10.0
7.5
5.0
2.5
0.0
0
250000
500000
750000
1000000
CT Genome Positions

Supplementary Table 1: Primers used for PCR amplification and sequencing of the entire *C. trachomatis* plasmid for each patient sample set.

Fragment	PCR primer	PCR primer sequence (5'-3')	Fragment size (bp)	Sequencing primer	Sequencing primer sequence (5'-3')
Fragment 1	F1_Forward	AACTCTTGGTGGTAGACTTTGCA	1751	F1_Reverse_seq1	TGACTTGTTGTTACAGGAATCCCT
	F1_Reverse	AACAAGTTCGAGCAGCAAGC		Same as F1_Reverse	
Fragment 2	F2_Forward	GCTTGCTGCTCGAACTTGTT	1750	Same as F2_Forward	
	F2_Reverse	TGTAATACCGAAGAGAAAACCGA		Same as F2_Reverse	
Fragment 3	F3_Forward	CATGGATCGGTTTTCTCTTCGG	1900	Same as F3_Forward	
	F3_Reverse	CCCATACCACACCGCTTTCT		Same as F3_Reverse	
Fragment 4	F4_Forward	CGTATTCATTACGTGTAGGCGG	1474	Same as F4_Forward	
	F4_Reverse	TCGATCCAAACTCTGACTTTCCT		Same as F4_Reverse	
Fragment 5	F5_Forward	ACTCTCAGAGGACAACGTGA	828	Same as F5_Forward	
	F5_Reverse	TCTCTCTCGTAAAATCAAATCCCT		Same as F5_Reverse	

Supplementary Table 2: List of *Chlamydia trachomatis* genomes and their NCBI/SRA accession IDs used for reconstructing the global phylogenetic tree and subsequent putative recombination detection analysis

Sample Id	***omp* A genotype**	**NCBI/SRA Accession Ids**
Ja_107R	Ja	PRJNA609714
Ja_107V	Ja	PRJNA609714
Ja_192R	Ja	PRJNA609714
Ja_192V	Ja	PRJNA609714
G_72R	G	PRJNA609714
G_72V	G	PRJNA609714
G_98R	G	PRJNA609714
G_98V	G	PRJNA609714
A_2497	A	nc_017437.1
A_2497_1	A	nc_016798.1
A_363	A	nc_020966.1
A_5291	A	nc_020939.1
A_7249	A	nc_020944.1
A_HAR13	A	nc_007429.1
B_Jali20	B	nc_012686.1
B_TZ1A828	B	nc_012687.1
C_TW3	C	nc_023060.1
chxRP1H1	D	nz_cp020537.1
chxRP4D10	D	nz_cp020536.1
chxRP4F1	D	nz_cp020535.1
D_13-96	Da	nc_022119.1
D_2923	Da	ACFJ01000001
D_2s	D	SRA051544.1
D_43nl	D	SRA051526.1
D_84s	D	SRA051539.1
D_CS637_11	D	nz_cp007131.1
D_SotonD5	D	nc_020943.1
D_UW3	D	AE001273
D-EC	D	nc_017434.1
D-LC	D	nc_017436.1
E_11023	E	nc_017431.1
E_150	E	nc_017439.1
E_5s	E	SRA051547.1
E_Bour	E	nc_020971.1
E_CS1025_11	E	nz_cp010567.1
E_DK20	E	nz_cp015304.1
E_IU824	E	nc_020511.1
E_IU888	E	nc_020512.1
E_SotonE4	E	nc_020969.1
E_SW3	E	nc_017952.1
E-103	E	nz_cp015294.1
E-160	E	nz_cp015296.1
E-32931	E	nz_cp015302.1
E-547	E	nz_cp015298.1
E-8873	E	nz_cp015300.1

F_1	F	SRA051574.1
F_11-96	F	nc_022107.1
F_1-93	F	nc_022117.1
F_38nL	E	SRA051469.2
F_6-94	F	nc_022118.1
F_70	F	ABYF01000001
F_CS847_08	F	nz_cp010569.1
F_SW4	F	nc_017951.1
F_SW5	F	nc_017953.1
F-6068	F	nz_cp015306.1
G_11074	G	nc_017440.1
G_11222	G	nc_017430.1
G_9301	G	nc_017432.1
G_9768	G	nc_017429.1
G_SotonG1	G	nc_020941.1
G_UW57	G	SRA051545.1
H_18s	H	SRA051541.1
H_UW4	H	SRA051548.1
Ia_CS190_96	Ia	nz_cp010571.1
Ia_SotonIa1	Ia	nc_020970.1
Ia_SotonIa3	Ia	nc_020940.1
Ia_UW202	Ia	SRA051537.1
J_31-98	J	nc_022121.1
J_6276	J	nc_021892.1
Ja_26s	F	SRA051540.1
Ja_47nL	Ja	SRA051542.1
K_SotonK1	K	nc_020965.1
L1_115	L2	nc_020929.1
L1_224	L2	nc_020973.1
L1_440	L1	nc_020937.1
L2_434BU	L2	nc_021052.1
L2_SF25667	L2	nc_020930.1
L2b_795	L2b	nc_020938.1
L2b_8200	L2b	nc_020945.1
L2b_Ams1	L2b	nc_020933.1
L2b_CS19_08	L2b	nz_cp009923.1
L2b_CS784_08	L2b	nz_cp009925.1
L2b_UCH1	L2b	nc_010280.2
L2b_UCH2	L2b	nc_020931.1
L2C	L2c	nc_015744.1
L3_404	L3	nc_020974.1
QH111L	Ba	nz_cp018052.1
SQ01	J	nz_cp017740.1
SQ02	J	nz_cp017741.1
SQ05	J	nz_cp017742.1
SQ09	Ia	nz_cp017736.1
SQ10	Ia	nz_cp017737.1
SQ12	Ia	nz_cp017738.1

SQ14	Ia	nz_cp017739.1
SQ15	K	nz_cp017745.1
SQ19	K	nz_cp017746.1
SQ20	H	nz_cp017732.1
SQ24	H	nz_cp017733.1
SQ25	Ja	nz_cp017743.1
SQ29	Da	nz_cp017731.1
SQ32	Da	nz_cp017730.1
Sweden2	E	nc_017441.1

Table S3. Genes present in putative recombination blocks for patient specimen sets.

Lineages/Clinical sample	Total # of SNPs	Total # of SNPs outside putative recombination sites	r/m	rho/teta	Putative Recombinant sites	Start genomic cordinates	End genomic cordinates	homoplasious snps within the putative recombinant sites
Prevalent - 192R-192V	419	139	3.014	0.0431	Block 1	90931	98123	21
					Block 2	313243	366721	222
					Block 3	404379	407297	13
					Block 4	439342	446179	16
					Block 5	755205	756077	11
					Block 6	768836	781753	136
Prevalent -107R 107V	260	45	5.777	0.1111	Block1	425626	426421	7
					Block 2	732506	738362	10
					Block 3	754335	756077	14
					Block 4	764505	782188	157
					Block 5	1018158	1026863	72
Non-prevalent - 72R 72V	213	56	3.8035	0.05357	Block 1	790402	811640	95
					Block 2	818795	820024	7
					Block 3	825543	853694	111
Non-prevalen - 98V 98R	21	32	0.6562	0.0312	Block 1	120752	135523	21

Genes present within the predicted putative recombinant sites

smpB (CT_076), Thiamine biosynthesis lipoprotein (CT_077), folD(CT_078), Hypothetical Protein (CT_079),ltuB (CT_080), hypothetical protein (CT_081), hypothetical protein (CT_082),hypothetical protein (CT_083), phospholipase (CT_084) and decarboxylase (CT_085)

gcsH(CT_282),hypothetical protein (CT_283), phospholipase D (CT_284) lplA_1 (CT_285), clpC (CT_286), MnmA (CT_287), hypothetical protein (CT_288), hypothetical protein (CT_289), plsN_1 (CT_290), dut (CT_292), accD (CT_293), sodM (CT_294), mrsA_1 (CT_295), hypothetical protein (
def (CT_353), kgsA (CT_354), hypothetical protein (CT_355), yyaL (CT_356)

yclF (CT_385), hypothetical protein (CT_386), hypothetical protein (CT_387), hypothetical protein (CT_388), hypothetical protein (CT_389), aspC (CT_390), ABC transporter substrate-binding protein (CT_391)

DNA topoisomerase IV subunit A

Type III secretion system protein (CT_672), pkn5 (CT_673), yscC (CT_674), karG (CT_675), protein-arginine kinase activator protein (CT_676), rrf (CT_677), pyrH (CT_678), tsf (CT_679), rs2 (CT_680), ompA (CT_681) and pbpB (CT_682)

porin (CT_372), pyruvoyl-dependent arginine decarboxylase (CT_373)

recC (CT_640), ygeD (CT_641), hypothetical protein (CT_642), topA (CT_643)

sfhB (CT_658), hypothetical protein (CT_659), DNA topoisomerase IV subunit A (CT_660),

hypothetical protein (CT_668), yscN (CT_669), hypothetical protein (CT_670), hypothetical protein (CT_671), Type III secretion system protein (CT_672), pkn5 (CT_673), yscC (CT_674), karG (CT_675), protein-arginine kinase activator protein (CT_676), rrf (CT_677), pyrH (CT_678), tsf (CT_679)

xerD (CT_864), hypothetical protein (CT_865), glgB (CT_866), deubiquitinase (CT_867), deubiquitinase (CT_868), pmpE (CT_869)

parB (CT_688), dppF (CT_689), dppD (CT_690), hypothetical protein (CT_691), phosphate permease (CT_691), pgk (CT_693), hypothetical protein (CT_694), hypothetical protein (CT_695), hypothetical protein (CT_696), nth (CT_697), tRNA modification GTPase (CT_698), psdD (CT_699), hypo
mreB (CT_709), pckA (CT_710)

porB (CT_713), gpdA (CT_714), UDP-N_acetylglucosamine pyrophosphorylase (CT_715), hypothetical protein (CT_716), Type III secretion system ATP synthase (CT_717), hypothetical protein (CT_718), type III secretion system protein (CT_719), hypothetical protein (CT_720), yfhO_2 (CT_721),p

hypothetical protein (CT_105), yceC (CT_106), muLY (CT_107), hypothetical protein (CT_108), hypothetical protein (CT_109), groEL_1 (CT_110), groES (CT_111), pepF (CT_112), clpB (CT_113), hypothetical protein (CT_714), inclusison membrane protein D (CT_115)

:in (CT_300), pknD (CT_301), valS (CT_302), hypothetical protein (CT_303), alpK (CT_304), alpI (CT_305), alpD (CT_306), alpB (CT_307), alpA (CT_308), hypothetical protein (CT_309), alpE (CT_310), hypothetical protein (CT_311), hypothetical protein (CT_312), Lal (CT

GTPase Der (CT_703), pcnB_2 (CT_704), clpX (CT_705)

5), cell division protein (CT_726), zntA (CT_727), hypothetical protein (CT_728), serS (CT_729), ribD (CT_730), ribA/ribB (CT_731), ribE (CT_732), hypothetical protein (CT_733), lipoprotein (CT_734), dagA_2 (CT_735), hypothetical protein (CT_736), hypothetical protein (

l'_313), rpoC (CT_314), rpoB (CT_315), rl7 (CT_316), rl10(CT_317), rl1 (CT_318), rl11 (CT_319), nusG (CT_320), secE (CT_321),tufA (CT_322), infA (CT_323), hypothetical protein (CT_324), SufE (CT_325), hypothetical protein (CT_326), trpC (CT_327).

CT_737), yycJ (CT_738), ftsK (CT_739),