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 Deanc,d

aDivision of STD Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
bEmory University School of Medicine, Atlanta, Georgia, USA
cUniversity of California San Francisco School of Medicine, San Francisco, California, USA
dUniversity 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 SureSelectXT 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 work now provides a robust approach for discerning genomic diversity and advancing our understanding 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

C. trachomatis is the most common cause of bacterial sexually transmitted infections (STIs) worldwide and the most notifiable disease in the United States (1, 3).
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 men and women 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) (L1 to L3, L2a, L2b, and L2c) (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 L1b, among MSM in many countries of the world and recombinant strains L2/D (termed L2c) and L1b/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 SureSelectXT 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 ≥98% genome coverage at a ≥5× 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 L2b 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 L2b 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× coverage was ≥98%, with the exception of 72R, which had only 96.11% of the genome covered at a minimum of 5× read depth.
**TABLE 1** Sequence data analysis of C. trachomatis gDNA from spiked serial dilutions of genotype D

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>ompA genotype</th>
<th>C&lt;sub&gt;T&lt;/sub&gt;</th>
<th>Total no. of genome copies input</th>
<th>No. of:</th>
<th>Nonhuman read pair ratio</th>
<th>No. of:</th>
<th>% selected read pairs</th>
<th>Mean read depth</th>
<th>% of genome with coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total read pairs</td>
<td>Raw read pairs</td>
<td>Nonhuman read pairs</td>
<td>Total read pairs after QC&lt;sup&gt;b&lt;/sup&gt;&lt;sup&gt;e&lt;/sup&gt;</td>
<td>Selected read pairs&lt;sup&gt;b&lt;/sup&gt;&lt;sup&gt;c&lt;/sup&gt;&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>D</td>
<td>18.29</td>
<td>7,854,406</td>
<td>2,136,708</td>
<td>2,118,719</td>
<td>0.99</td>
<td>2,095,052</td>
<td>2,079,951</td>
<td>99.28</td>
</tr>
<tr>
<td>D2</td>
<td>D</td>
<td>22.01</td>
<td>942,398</td>
<td>1,482,054</td>
<td>1,382,594</td>
<td>0.93</td>
<td>1,335,421</td>
<td>1,318,461</td>
<td>98.73</td>
</tr>
<tr>
<td>D3</td>
<td>D</td>
<td>27.58</td>
<td>39,392</td>
<td>1,115,066</td>
<td>729,165</td>
<td>0.65</td>
<td>667,101</td>
<td>624,035</td>
<td>93.54</td>
</tr>
<tr>
<td>D4</td>
<td>D</td>
<td>29.055</td>
<td>16,945</td>
<td>1,117,188</td>
<td>277,491</td>
<td>0.25</td>
<td>237,908</td>
<td>155,934</td>
<td>65.54</td>
</tr>
<tr>
<td>D5</td>
<td>D</td>
<td>32.15</td>
<td>1,876</td>
<td>1,955,298</td>
<td>219,077</td>
<td>0.11</td>
<td>162,471</td>
<td>28,011</td>
<td>17.24</td>
</tr>
<tr>
<td>D6</td>
<td>D</td>
<td>36.345</td>
<td>265</td>
<td>1,509,294</td>
<td>189,690</td>
<td>0.13</td>
<td>159,750</td>
<td>2,132</td>
<td>1.33</td>
</tr>
</tbody>
</table>

<sup>a</sup>C<sub>T</sub>, 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.

<sup>b</sup>With a quality score of ≥15.

<sup>c</sup>After quality control, i.e., deduplication and adapter trimming.

<sup>d</sup>Selected for *Chlamydia* spp. and *C. trachomatis* read pairs.

<sup>e</sup>With rRNA masked.
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 “prevailing urogenital and rectal strain” clade, while patient specimen sets 98 and 72 clustered within the “nonprevailing urogenital and rectal strain” clade (Fig. 3).

Patient genomes 107 and 192 clustered in the same clade as the E genotype genomes (prevailing urogenital and anorectal lineage), whereas patient genomes 98 and 72 clustered in the same clade as G ompA genotypes (nonprevailing 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 ($\rho/\theta = 0.111$) 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; $\rho/\theta = 0.053$; $r/m = 3.803$). 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 L2b and patient specimen sets

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>ompA genotype</th>
<th>Total no. of genome copies input</th>
<th>Raw read pairs</th>
<th>Nonhuman read pairs</th>
<th>Nonhuman read pair ratio</th>
<th>Total read pairs after QC</th>
<th>Selected read pairs</th>
<th>% selected read pairs</th>
<th>Mean read depth</th>
<th>% of genome with coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>L2b</td>
<td>900,000</td>
<td>2,099,167</td>
<td>2,079,997</td>
<td>0.99</td>
<td>2,067,413</td>
<td>2,042,024</td>
<td>98.77</td>
<td>550.78</td>
<td>98.98</td>
</tr>
<tr>
<td>L2</td>
<td>L2b</td>
<td>90,000</td>
<td>2,239,718</td>
<td>2,206,788</td>
<td>0.99</td>
<td>2,189,496</td>
<td>2,142,872</td>
<td>97.87</td>
<td>564.40</td>
<td>98.98</td>
</tr>
<tr>
<td>L3</td>
<td>L2b</td>
<td>9,000</td>
<td>2,230,573</td>
<td>2,108,624</td>
<td>0.95</td>
<td>2,045,830</td>
<td>1,988,919</td>
<td>97.22</td>
<td>531.81</td>
<td>98.97</td>
</tr>
<tr>
<td>107R</td>
<td>Ja</td>
<td>1,378,275</td>
<td>2,302,152</td>
<td>2,051,954</td>
<td>0.89</td>
<td>1,924,399</td>
<td>705,294</td>
<td>36.65</td>
<td>190.25</td>
<td>98.91</td>
</tr>
<tr>
<td>107V</td>
<td>Ja</td>
<td>177,440</td>
<td>1,347,716</td>
<td>1,048,957</td>
<td>0.78</td>
<td>942,944</td>
<td>175,198</td>
<td>18.58</td>
<td>47.15</td>
<td>98.84</td>
</tr>
<tr>
<td>192R</td>
<td>Ja</td>
<td>3,336,780</td>
<td>2,175,904</td>
<td>1,937,403</td>
<td>0.89</td>
<td>1,793,255</td>
<td>1,184,836</td>
<td>66.07</td>
<td>313.09</td>
<td>98.92</td>
</tr>
<tr>
<td>192V</td>
<td>Ja</td>
<td>28,536</td>
<td>1,390,989</td>
<td>1,040,223</td>
<td>0.75</td>
<td>940,352</td>
<td>331,250</td>
<td>35.23</td>
<td>90.23</td>
<td>98.91</td>
</tr>
<tr>
<td>72R</td>
<td>G</td>
<td>121,472</td>
<td>1,764,699</td>
<td>1,652,600</td>
<td>0.94</td>
<td>1,429,554</td>
<td>53,096</td>
<td>3.71</td>
<td>13.99</td>
<td>96.11</td>
</tr>
<tr>
<td>72V</td>
<td>G</td>
<td>189,673</td>
<td>1,568,358</td>
<td>556,417</td>
<td>0.35</td>
<td>463,667</td>
<td>169,916</td>
<td>36.65</td>
<td>46.16</td>
<td>98.82</td>
</tr>
<tr>
<td>98R</td>
<td>G</td>
<td>2460,822</td>
<td>1,930,840</td>
<td>1,822,417</td>
<td>0.94</td>
<td>1,715,197</td>
<td>962,269</td>
<td>56.10</td>
<td>261.91</td>
<td>98.94</td>
</tr>
<tr>
<td>98V</td>
<td>G</td>
<td>20,764</td>
<td>1,259,385</td>
<td>856,804</td>
<td>0.68</td>
<td>791,372</td>
<td>307,678</td>
<td>38.88</td>
<td>84.58</td>
<td>98.88</td>
</tr>
</tbody>
</table>

---

*a, 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.*
98V and 98R patient specimen sets (nonprevalent lineage; \( p/\theta = 0.031; \rho/m = 0.6562 \)) (Table S3).

Some of the previously identified genomic regions of higher homologous recombination were also identified in this study. The \( \text{ompA} \) gene has undergone recombination in both prevalent urogenital patient specimen sets 107 and 192. \( \text{ompA} \) genotyping by both Sanger sequencing and whole-genome data indicated that the \( \text{ompA} \) 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/\( \text{ompA} \) along with the neighboring genes (from type III secretion system protein gene [CT_672] to \( \text{pbpB} \) [CT_682]) into patient 192 and a larger fragment of 17.6 kbp from CT_672 to \( \text{pbpB} \) (CT_682) along with \( \text{ompA} \) into patient 107, likely from a Ja strain (Fig. 3; Table S3). The \( \text{pmpE} \) and \( \text{mrsA}_1 \) genes were estimated to be recombinant, respectively, in sets 107 and 192 (40, 54). The inclusion membrane protein gene \( \text{incD} \) was predicted to be recombinant only in the nonprevalent urogenital patient specimen set 98. (Table S3) (40, 54).

**DISCUSSION**

A SureSelectXT workflow with a 2.698 Mbp RNA bait library with 34,795 120-mer probes was developed from 85 GenBank \( C. \text{trachomatis} \) reference genomes encompassing all four lineages of \( C. \text{trachomatis} \), compared to the previous RNA bait library developed from 74 GenBank \( C. \text{trachomatis} \) reference genomes with 33,619 120-mer probes (52). By including more \( C. \text{trachomatis} \) reference genomes in our bait library design, which generated 1,000+ more probes, we likely captured more genetic diversity of the \( C. \text{trachomatis} \) strains circulating in the population than the RNA baits generated previously by Christiansen et al. (52). Moreover, having \(-98\%\) of the \( C. \text{trachomatis} \) 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 \( C. \text{trachomatis} \) genome. Compared to a previous study that obtained 98% coverage with a total input of 4,800 \( C. \text{trachomatis} \) strain F/SW4 genome.

---

**FIG 2** Percent coverage of reference genome and reads mapping to the respective reference genome for spiked mock samples of genotype L2b 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 indicated. R, rectal sample; V, vaginal sample. The orange line represents the percent down-selected \( \text{Chlamydia} \) spp. and \( C. \text{trachomatis} \) read pairs.
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 L2b, 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 L2b 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...
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.

### TABLE 3 SNPs detected within patient specimen sets

<table>
<thead>
<tr>
<th>Patient specimen set</th>
<th>Gene containing within-host SNPs</th>
<th>No. of SNPs</th>
<th>Genome location(s) (SNP and nucleotide coverage for each variant)</th>
<th>Reference genome used for calling the SNPs&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>107R and 107V</td>
<td>Gene for hypothetical protein (FSW4_RS00260)</td>
<td>1</td>
<td>54906 (C→T; T:62, C:1)</td>
<td>F/SW4 (NC_017951.1)</td>
</tr>
<tr>
<td></td>
<td>Gene for hypothetical protein (FSW4_RS00270)</td>
<td>2</td>
<td>59520 (T→C; C:33, T:0), 59523 (C→T; T:33, C:0)</td>
<td>F/SW4 (NC_017951.1)</td>
</tr>
<tr>
<td></td>
<td>Intergenic region</td>
<td>1</td>
<td>79002 (G→A; A:77, G:0)</td>
<td>F/SW4 (NC_017951.1)</td>
</tr>
<tr>
<td></td>
<td>Gene for amino acid ABC transporter ATP-binding protein</td>
<td>1</td>
<td>146249 (C→T; T:145, C:0)</td>
<td>F/SW4 (NC_017951.1)</td>
</tr>
<tr>
<td>192R and 192V</td>
<td>Gene for PTS sugar transporter subunit IIA</td>
<td>1</td>
<td>437237 (T→C; C:36, T:0)</td>
<td>F/SW4 (NC_017951.1)</td>
</tr>
<tr>
<td>72R and 72V</td>
<td><em>ompA</em></td>
<td>2</td>
<td>779230 (G→A; A:33, G:1), 779235 (C→A; A:31, G:2)</td>
<td>F/SW4 (NC_017951.1)</td>
</tr>
<tr>
<td></td>
<td>Gene for hypothetical protein (CT_744)</td>
<td>1</td>
<td>864903 (T→C; C:50, T:0), 864904 (C→T; T:51, C:0)</td>
<td>D/UW-3/CX (NC_000117.1)</td>
</tr>
<tr>
<td></td>
<td><em>pmpF</em></td>
<td>2</td>
<td>1029641 (G→T; T:75, G:0), 1029643 (A→G; G:36, A:0)</td>
<td>D/UW-3/CX (NC_000117.1)</td>
</tr>
<tr>
<td>98R and 98V</td>
<td>Gene for hypothetical protein (CT_049)</td>
<td>1</td>
<td>54281 (C→T; T:23, C:0)</td>
<td>D/UW-3/CX (NC_000117.1)</td>
</tr>
</tbody>
</table>

<sup>a</sup>R, rectal; V, vaginal.

<sup>b</sup>SNPs were called by comparing the genomes of patient specimen sets to reference genomes that clustered in the whole-genome phylogenetic tree.
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 10⁻⁹) were prepared by spiking into 1× phosphate-buffered saline (PBS). A standard curve based on ATCC's reported copy number for genotype L2b (ATCC 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 × 10⁻⁰·⁶⁰₂⁰x; R² = 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 L2b. gDNA from C. trachomatis clinical strain genotype L2b (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
Assemblies using SPAdes 3.7.0 (75) with the D/UW-3/Cx reference sequence (NC_000117.1) from which we masked 6 rRNA genes, CT_r01 contigs were used to extract and compare the lysed with 59 tis sequences was also conbase containing all bacterial and archaeal reference nucleotide sequences. The presence 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 exonSAP-IT (Thermo Fisher) and subjected to Sanger sequencing using the PCR primers (65). Forward and reverse sequences were aligned using MAFFT v7.450 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/Uc, J/UW-202, J/UW-36/Cx, Ja/UW-92, K/UW-31/Cx, L_440, L_434, L_a/UW-396, L_b/UCH-1/ proctitis, L_c, and L_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 SureSelectXT target enrichment system for Illumina paired-end multiplexed sequencing library (V2, December 2018) and all recommended quality control steps were performed on all RNA libraries. 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. Duplication of the reads was performed using Clumpify (sourceforge.net/projects/bbmap/) with the dedupe=1 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_01 (16SrRNA_1), CT_02 (23SrRNA_1), CT_03 (5SrRNA_1), CT_04 (16SrRNA_2), CT_05 (23SrRNA_2), and CT_06 (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 ≥ 20), followed by consensus reference 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 × 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**


5. Spaargaren J, Fennema HS, Morre SA, de Vries HJ, Coutinho RA. 2005. New lymphogranuloma venereum \(Chlamydia\) trachomatis variant,


