Title

Whole-genome automated assembly pipeline for Chlamydia trachomatis strains from reference, in vitro and clinical samples using the integrated CtGAP pipeline.

Journal

NAR Genomics and Bioinformatics, 7(1)

Authors

Olagoke, Olusola
Aziz, Ammar
Zhu, Lucile
et al.

Publication Date

2025-03-01

DOI

10.1093/nargab/lqae187

Peer reviewed

NAR Genomics and Bioinformatics, 2025, **7**, lqae187
https://doi.org/10.1093/nargab/lqae187
Advance access publication date: 7 January 2025
Pipelines and Workflows for Biological Data Analysis

OXFORD

Whole-genome automated assembly pipeline for *Chlamydia trachomatis* strains from reference, *in vitro* and clinical samples using the integrated CtGAP pipeline

Olusola Olagoke[1], Ammar Aziz[2], Lucile H. Zhu[3], Timothy D. Read[4,†] and Deborah Dean[1,3,5,6,7,*,†]

[1]Departments of Medicine and Pediatrics, Division of Infectious Diseases and Global Health, University of California San Francisco School of Medicine, 550 16th Street, 4th Floor Mission Hall, San Francisco, CA, 94158, USA
[2]Victorian Infectious Diseases Reference Laboratory, 792 Elizabeth Street, Melbourne, Victoria, 3000, Australia
[3]Department of Bioengineering, University of California San Francisco and Berkeley School of Engineering, 306 Stanley Hall, Berkeley, CA, 94720, USA
[4]Departments of Medicine and Genetics, Division of Infectious Diseases, Emory University School of Medicine, 100 Woodruff Circle, Atlanta, GA, 30322, USA
[5]Bixby Center for Global Reproductive Health, University of California San Francisco, 1001 Potrero Ave, San Francisco, CA, 94110, USA
[6]Benioff Center for Microbiome Medicine, University of California San Francisco, 513 Parnassus Avenue, S357, San Francisco, CA, 94143, USA
[7]University of California San Francisco Institute of Global Health Sciences, 550 16th Street, 3rd Floor Mission Hall, San Francisco, CA, 94158, USA

[*]To whom correspondence should be addressed. Tel: +1 510 450 7655; Fax: +1 510 450 7190; Email: deborah.dean@ucsf.edu
[†]The last two authors should be regarded as Joint Last Authors.

Abstract

Whole genome sequencing (WGS) is pivotal for the molecular characterization of *Chlamydia trachomatis* (*Ct*)—the leading bacterial cause of sexually transmitted infections and infectious blindness worldwide. *Ct* WGS can inform epidemiologic, public health and outbreak investigations of these human-restricted pathogens. However, challenges persist in generating high-quality genomes for downstream analyses given its obligate intracellular nature and difficulty with *in vitro* propagation. No single tool exists for the entirety of *Ct* genome assembly, necessitating the adaptation of multiple programs with varying success. Compounding this issue is the absence of reliable *Ct* reference strain genomes. We, therefore, developed CtGAP— *Chlamydia trachomatis* Genome Assembly Pipeline—as an integrated 'one-stop-shop' pipeline for assembly and characterization of *Ct* genome sequencing data from various sources including isolates, *in vitro* samples, clinical swabs and urine. CtGAP, written in Snakemake, enables read quality statistics output, adapter and quality trimming, host read removal, *de novo* and reference-guided assembly, contig scaffolding, selective *omp*A, multi-locus-sequence and plasmid typing, phylogenetic tree construction, and recombinant genome identification. Twenty *Ct* reference genomes were also generated. Successfully validated on a diverse collection of 363 samples containing *Ct*, CtGAP represents a novel pipeline requiring minimal bioinformatics expertise with easy adaptation for use with other bacterial species.

Introduction

Chlamydia trachomatis (*Ct*), an obligate intracellular bacterium, is the leading cause of bacterial sexually transmitted infections (STIs) with ∼130 million cases occurring worldwide each year (1). *Ct* is also the leading cause of blindness—referred to as trachoma—with over 200 million individuals at risk of irreversible blindness (https://www.who.int/health-topics/trachoma#tab=tab_1). These figures are likely a gross underestimation of the true burden of disease given the inability to effectively screen at-risk global populations, thus highlighting the importance of *Ct* as a major public health concern.

Traditionally, the molecular basis for *Ct* strain typing was attributed to the major outer membrane protein encoded by *omp*A (2). Typing evolved to sequencing the *omp*A gene (3) and then seven housekeeping genes, known as multiple locus sequence typing (MLST) (4,5). While these methods were largely useful in linking clinical presentation and tissue tropism along with socio-demographic and epidemiologic information, utilizing only one or a small number of genes can obscure the true nature of *Ct* genomes (6,7–9), including indels and recombination. For instance, the L_2c (also known as L_2-D/SF/L_2c) strain, which is a recombinant of a lymphogranuloma venereum (LGV) strain L_2 and a urogenital strain D (10), has an MLST that is identical to all other non-recombinant LGV strains (6).

Whole genome sequencing (WGS) has become an essential and economically beneficial tool for the molecular characterization and surveillance of bacterial pathogens (7) especially in clinical samples. As postulated by Simar and colleagues (8), WGS-based bacterial strain typing will ultimately lead to more effective infection control and interventions globally. Assembling quality *Ct* genomes for downstream applica-

Received: August 13, 2024. **Revised:** December 10, 2024. **Editorial Decision:** December 16, 2024. **Accepted:** December 18, 2024

tions, however, remains a major challenge. Currently, no single tool or pipeline has been developed to handle the entire process of *Ct* genome assembly. But the lack of a standardized workflow is not unique to *Ct* studies. A recent review highlighted, 'differing interpretations, quality control issues and decreased reproducibility' (8) as common problems arising from bacterial WGS studies. As an obligate intracellular pathogen, *Ct* must be grown in host cells, further complicating genome assembly if host read contamination is not successfully depleted/extracted. As such, there is an urgent need for the creation and validation of standardized workflows for *Ct* WGS that will facilitate the comparison of WGS data from different laboratories throughout the world for various research applications and improved public health surveillance (9).

The choice of a reference genome for mapping during genome assembly could introduce biases mainly through single nucleotide polymorphisms (SNPs) in the final assembly, thus leading to false discovery of new variants or incorrect phylogenetic trees (11). The choice of reference strains for *Ct* genome studies appears to have been loosely based on lab preference. For instance, while some labs have utilized reference strain D/UW-3/Cx (12,13), others have used reference strain L_2/434/LN for WGS mapping or scaffolding (14). However, these genomes vary in key regions such as the Plasticity Zone (15). In addition, while some have performed mapping against a single reference genome (12,16), others have created a consensus reference sequence from multiple available reference strains for mapping (17,18). But the source of these 'reference strains' vary and may have several SNPs compared across the same strains. This lack of consistency in reference sequence selection likely hampers meaningful comparisons of *Ct* WGSs from different labs globally. This problem is compounded by a lack of a reliable set of *Ct* reference strain genomes for genomic studies.

To address these issues, we developed the *Chlamydia trachomatis* Genome Assembly Pipeline (CtGAP) as a 'one-stop-shop' pipeline for genome assembly and initial characterization of *Ct* sequencing reads generated from DNA purified from *Ct* isolates, clinical swabs, *in vitro* studies and urine samples. Consensus genomes for *Ct* reference strains were developed as well as a plurality consensus sequence of the 21 reference strains to enable reference-guided assembly. Here, CtGAP was used to process 363 *Ct* samples from the NCBI sequence read archives (SRAs), European Nucleotide Archive (ENA) and GenBank databases representing geographically diverse trachoma and STI populations. Comparative analysis performed on the dataset include construction of whole genome phylogenetic trees, and genotyping with *ompA* and MLST as cross-comparators. CtGAP can also be customized for other *Chlamydia* spp. and bacteria.

Materials and methods

Generation of *C. trachomatis* reference strain genomes using Illumina SureSelect and nanopore

Ct reference strains summarized in Table 1 were propagated, and genomic (g)DNA was extracted and *ompA* genotyped to verify the strain as we described (12,19) prior to WGS. Illumina sequencing used the SureSelect RNA bait capture methodology as we described (12,13). Briefly, 3μg of gDNA in 130μl of TE_{low} was sheared on a Covaris M220 instrument (Covaris, Woburn, MA) followed by magnetic bead purification. SureSelectXT Target Enrichment employed an upgraded RNA bait library consisting of 35 996, 120-mer probes spanning 86 GenBank *Ct* reference and clinical chromosomes and plasmids (Agilent Technologies, INC, Santa Clara, CA, reference: ELID: 3325141) (12). Capture libraries were sequenced as 150 bp paired end reads on an Illumina NovaSeq instrument for at least 100× coverage per genome (~1.05 Mb *Ct* genome; ~7 kb plasmid).

For nanopore sequencing, the same reference strain gDNA was used as for Illumina sequencing but without the need for the SureSelectXT Target Enrichment step. Sample libraries were prepared for gDNA using the Oxford Nanopore Technologies Native Barcoding Kit 24 V14 (SQK-NBD114.24) per manufacturer's specifications. All samples were run on Nanopore R10.4.1 flow cells on either a MinION, Mk1B or GridION. Samples were subsequently demultiplexed using Guppy (V6). Prior to preprocessing, reads were filtered to keep the top 80% reads with the highest quality using the FiltLong package (https://github.com/rrwick/Filtlong) with a minimum 100× coverage generated per sample.

Both nanopore and SureSelect sequences were preprocessed using a suite of tools available in the bbtools package (20) for adapter trimming and removal of contaminating human reads. *De novo* genome assembly was performed using SPAdes (21) and Flye (22) using default settings on sequences from Illumina and nanopore platforms, respectively. Assembled contigs were subsequently aligned using BLASTn (23) against all *Ct* reference strain genomes present in NCBI database to confirm their *Ct* identity with a threshold of >99% to call a match. An additional step of scaffolding contigs generated by SPAdes was performed using RagTag (24).

For 10 reference strains (Table 1), genome sequences had previously been generated using the Roche 454 sequencing platform, which has now been discontinued. The genomes from the 454 platform were assembled using the 454 gsAssembler software (V 2.0.01.14) with default parameters and had been closed using primers flanking gaps to generate PCR products for Sanger sequencing.

Curating consensus *C. trachomatis* reference genome assemblies

Each reference genome was sequenced and assembled *de novo* by at least two technology platforms except for L_2b (see Table 1). To generate a consensus genome assembly, assembled genomes from the different technologies were aligned using the progressiveMauve option in Geneious Prime (Version 2023.2.1). The alignment was then manually inspected to identify regions where the assemblies differed and resolve potential sequencing or assembly errors. Where assemblies were generated from reads from the three platforms ($n = 10$), an agreement between at least two assemblies was required to resolve each observed discrepancy. For assemblies from only two platforms, a 'third' publicly available *Ct* reference sequence, if available, was added to the comparison. In general, this approach allowed us to confidently determine each base throughout the genome for the reference strain.

Assemblies from nanopore and Illumina platforms were available for 20 reference genomes; we did not have access to DNA from reference strain L_2b/UCH-2. Resolution of single nucleotide indels present in homopolymer regions, within which nanopore sequencing technology is known to be less accurate than Illumina (25), was done using the nucleotide (s) from the Illumina assembly for the final consensus genome. A BLASTn alignment against similar strains was also performed

Table 1. Characteristics of *Ct* reference strains and associated *omp*A genotypes, MLST and sequencing platform used

Strain	Year isolated	Anatomic site	Geographic origin	*omp*A allele	MLST[a]	Sequencing platform[b]
A/Sa-1	1957	Conjunctiva	Saudi Arabia	A	50	I, N, R
B/TW-5/OT	1959	Conjunctiva	Taiwan	B	6	I, N[c]
Ba/Apache-2	1960	Conjunctiva	Arizona	Ba	18	I, N, R
C/TW-3/OT	1959	Conjunctiva	Taiwan	C	11	I, N[c]
D/UW-3/Cx	1965	Cervix	Washington	D	48	I, N[c]
Da/TW-448	1985	Conjunctiva	Taiwan	Da	37	I, N, R
E/bour	1959	Cervix	California	E	39	I, N[c]
F/IC-Cal-13	1960	Cervix	California	F	34	I, N, R
G/UW-57/Cx	1971	Cervix	Washington	G	30	I, N, R
H/UW-4/Cx	1965	Cervix	Washington	H	19	I, N, R
I/UW-12/Ur	1966	Urethra	Washington	I	19	I, N[c]
Ia/UW-202	1985	Urethra	Washington	Ia	23	I, N, R
J/UW-36/Cx	1971	Cervix	Washington	J	9	I, N, R
Ja/UW-92	1992	Cervix	Washington	Ja	39	I, N[c]
K/UW-31/Cx	1973	Cervix	Washington	K	19	I, N, R
L_1/440/LN	1968	Lymph node	California	L_1	1	I, N[c]
L_2/434/Bu	1968	Bubo	California	L_2	1	I, N[c]
L_2a/UW-396/LN	1985	Lymph node	Washington	L_2a	1	I, N[c]
L_2b/UCH-2[d]	NA	Rectum	London	L_2b	1	I[d]
L_2-D/SF/L_2c	2010	Rectum	San Francisco	L_2c	1	I, N, R
L_3/404/LN	1967	Lymph node	California	L3	1	I, N[c]

The assembled genomes and SRAs have been deposited in NCBI under BioProject ID: PRJNA1137892 (URL: https://dataview.ncbi.nlm.nih.gov/object/PRJNA1137892?reviewer=oq7dbk66mg4bgvk6fr54msog3h) (Supplementary Table S1). NA, not available.
[a]MLST scheme (*C. trachomatis*) (4,6).
[b]I, Illumina SureSelect; N, nanopore; R, Roche 454.
[c]*Ct* reference genomes from NCBI were used to inform the consensus sequences.
[d]Sequence was obtained from a published work (14).

to confirm the call with a threshold of 100% identity. Larger sequence deletions were resolved by retaining the information from the other assembly. This was particularly helpful in resolving instances of Illumina assemblies missing 1–2 copies of known multicopy genes.

Genome assembly pipeline and development of a plurality consensus sequence

To perform both *de novo* and reference-guided *Ct* genome assemblies, we developed a workflow—referred to as CtGAP— utilizing open-source Linux-based packages (Figure 1, Supplementary Figure S1). The workflow was designed to handle WGS read data derived from gDNA extracted from *Ct* isolates and clinical samples and sequenced on the Illumina sequencing platform. CtGAP takes in raw DNA sequence reads in fastq format as input and performs an initial quality control including adaptor removal and read trimming using the Fastp package (26). This is followed by the removal of contaminating human reads using Scrubby (https://github.com/esteinig/scrubby) by: (i) specifying the depletion of all reads that are classified as Archaea, Eukaryota, Holozoa and Nucletmycea with kraken2 (27); and (ii) aligning leftover reads against human reference sequence GRCh38 with minimap2 (28). Scrubby is then utilized to extract all remaining reads mapping to the *Chlamydiales* order, as well as read statistics post-cleanup.

When working with reads from *Ct* isolates where the reference strain is known, CtGAP can perform a reference-guided genome assembly starting with reads mapping to the selected reference genome with Bowtie2 (29) and conversion of the samtools (30) sorted bam file into fastq with bedtools (31) followed by genome assembly with Shovill (https://github.com/tseemann/shovill). CtGAP also performs a *de novo* genome assembly using Shovill followed by *omp*A genotyping with custom BLAST (23), MLST and plasmid typing (4,32,33), and scaffolding with ragtag (24) and gapfiller (34) for the assembled contigs.

A consensus plurality reference genome was created from all 21 reference *Ct* strains (Table 1). These genomes were oriented with Dnaapler (35) and aligned with Mugsy (36) using default settings. Thereafter, goalign (37) was used to create a plurality consensus sequence to capture the variants present in the 21 reference *Ct* strains. CtGAP comes pre-equipped with this consensus sequence and can be used to perform a reference-guided genome assembly when working with clinical samples that have no prior strain information or when the user would prefer to perform this type of assembly.

CtGAP can also compare mapping of the query fastq to the reference genome database by mapping the filtered reads against all 21 *Ct* reference genomes using Bowtie2 (29) and generating genome coverage statistics using Samtools (30). The coverage output can be visualized in any text editing software. The reference strain with the most covered bases along with the highest mean mapping quality is recommended to be used in a second iteration of CtGAP for reference-guided genome assembly. The information from the coverage statistics may also be used in conjunction with other *de novo* assembly metrics—*omp*A genotyping, MLST and phylogenetic tree—to infer if the clinical sample is a recombinant. Genome quality assessment using QUAST (38) is performed on the genome output. Where both *de novo* and reference guided assemblies are performed, CtGAP also performs a comparison of both assemblies using MuMmer (39). A .tree file is generated by CtGAP that can be imported into other phylogenetic software such as Figtree or iTOL (40) for the downstream processing of choice. To ease piping into other bacterial genome analysis pipelines such as Bactopia (41), CtGAP outputs key intermediate data such as the processed fastq reads. A compre-

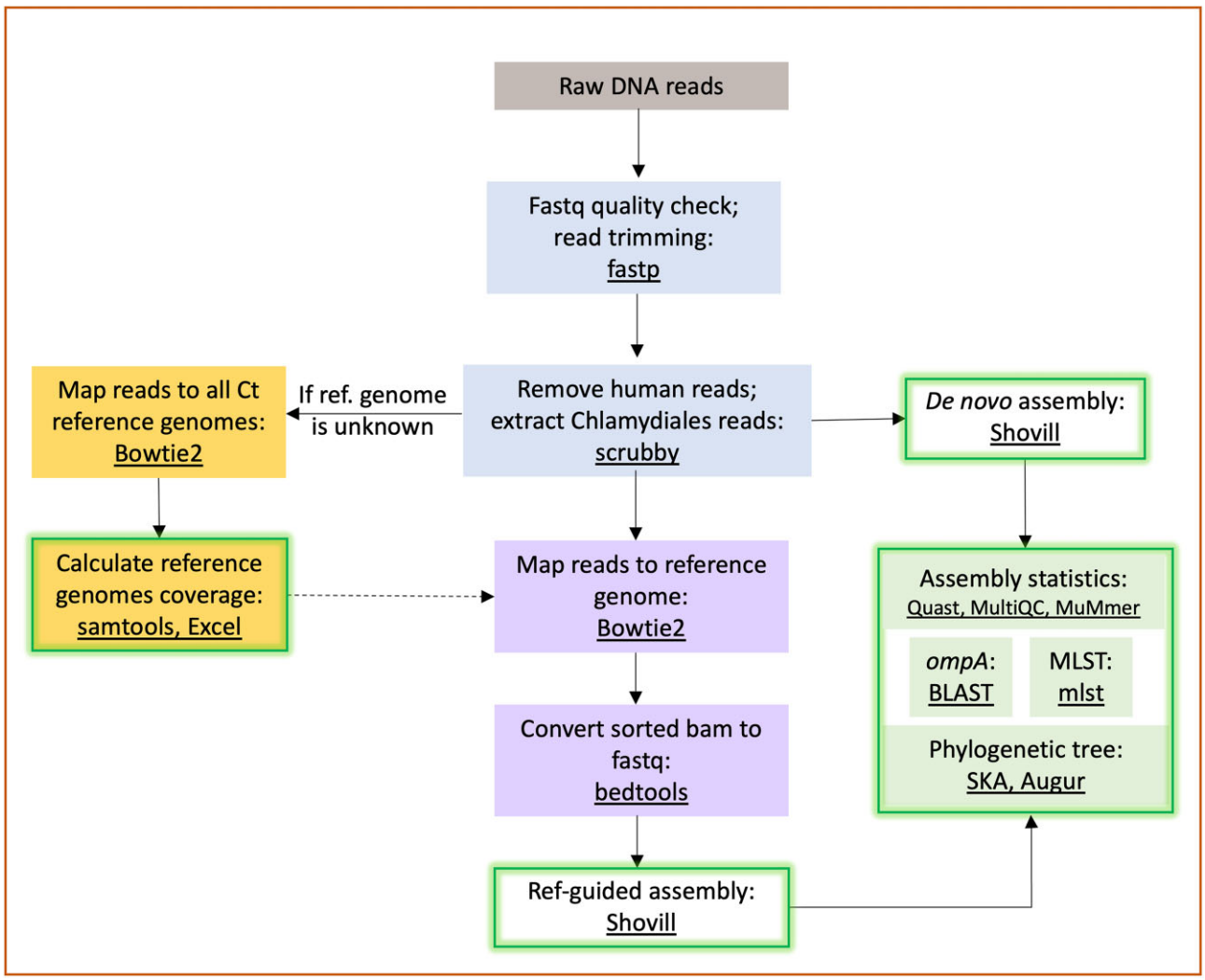

Figure 1. Simplified CtGAP workflow for generating *C. trachomatis* genomes. CtGAP provides the opportunity for *de novo* and reference-guided genome assembly with other outputs such as the *omp*A genotype, MLST and phylogenetic tree construction.

hensive CtGAP how-to documentation along with the source code is available at https://github.com/D-Dean-Lab/CtGAP.

Data sets and comparative genomics

To assess the functionality of the CtGAP workflow in comparative genomic analyses of *Ct*, raw reads from three data sets were tested: (i) *Ct* reference strains sequenced on the Illumina platform as part of this study ($n = 20$) (Supplementary Table S1); (ii) Complete *Ct* genomes from NCBI [$n = 92$; where no raw reads were available, synthetic reads were generated using ART (42)] (Supplementary Tables S2 and S3); and (iii) available SRAs in NCBI and ENA for clinical samples ($n = 271$) (Supplementary Table S4). A whole-genome phylogeny of all reconstructed genomes assembled using the *de novo* approach (unless otherwise stated) was generated using the Augur package (43).

Results

Manually polished *Ct* reference genomes and plurality consensus sequence generation

DNA extracted from 20 *Ct* reference strains were sequenced on both the Illumina and nanopore platforms and assembled as described in Methods. For 10 of these strains, Roche 454 sequenced genomes with gap closure using Sanger sequencing were also available. While programs like SPAdes (21) allow for hybrid assembly (44) of Illumina short reads with long nanopore reads, in our hands, we noticed that the hybrid assembly output was identical to the default assembly mode thus making the nanopore data redundant. Also, the SPAdes assembler cannot handle Roche 454 genomes. For each reference strain, therefore, genome assemblies from the available platforms [Illumina + Nanopore ($n = 20$), or Illumina + Nanopore + Roche 454 ($n = 10$)] were used to create a manually polished final assembly by comparing each assembly base-by-base and resolving any observed sequencing errors. When only two assemblies were available, WGSs of reference strains, if available in the public databases, were used as comparators (Table 1).

CtGAP training datasets and phylogenetic analyses

The CtGAP workflow (Figure 1) successfully incorporates a suite of open-source packages to facilitate the assembly and characterization of *Ct* WGS data. An important output from CtGAP is genome quality statistics obtained by incorporating QUAST into our pipeline. Examples of key information from

this output includes genome length, number of contigs and GC content. This information is provided for every CtGAP-assembled genome (data not shown). In the current study, the performance of CtGAP in correctly assembling *Ct* was tested on 92 available published reference and clinical genomes from NCBI. CtGAP automatically computes a phylogenetic tree by comparing the sequence of the assembled samples to the 21 reference genomes. Figure 2 shows the phylogenetic reconstructions of the 92 published genomes assembled using the *de novo* approach in CtGAP. All 92 genomes were successfully reassembled and typed using CtGAP.

Of the 92 publicly available *Ct* genomes above, we also ran CtGAP on 42 *Ct* samples that had both their raw reads and genome assemblies available in NCBI databases. Accession numbers for the 42 raw reads and assembled genomes are in Supplementary Table S3. CtGAP returned identical genomes to those available in NCBI's genome database based on analysis using MuMmer's dnadiff option (39) (Figure 3).

The performance of CtGAP was also tested on 271 clinical samples downloaded from NCBI's SRA database (Supplementary Table S4). These samples represent a diversity of *Ct* strains from various worldwide sexually transmitted and ocular trachoma populations. Examples of the *omp*A genotype, MLST, plasmid type and reference genome coverage outputs of CtGAP for these samples are shown in Supplementary Tables S5–S8, respectively. The CtGAP-generated *omp*A genotypes were determined by comparison against complete *omp*A gene sequences in GenBank using BLASTn (Table 1). Similarly, the CtGAP MLST designations for both genomes and plasmids were determined by using the *C. trachomatis* typing scheme in the PubMLST database (https://pubmlst.org/organisms/chlamydiales-spp) (4,32). However, in some cases, the NCBI SRAs did not contain any plasmid sequences. We therefore designated the samples without plasmid sequences as 'no plasmid profile' (Supplementary Table S7).

Four assemblies were obtained for each of the 271 samples using the *de novo*, plurality consensus sequence, and a single reference assembly option in the pipeline—a generic reference strain (D/UW-3/Cx) and the reference strain with the most mapped reads and mapping quality from the Bowtie2 mapping step (i.e. the 'top mapped' reference strain). A representative phylogenetic tree for all 271 samples using *de novo* assembly along with the associated MLST, and *omp*A genotypes is shown in Figure 4.

To assess the performance of the assembly methods, assemblies for the 271 samples were compared for the number of SNPs between any pair of assemblies using MuMmer's dnadiff option (39). The results showed that no sample had identical assemblies with all four assembly options and that the SNPs between the different assembly pairs varied between 1 and 495 (Figure 5, Supplementary Table S9, Supplementary Figure S2). The comparison of *de novo* and top mapped reference guided assemblies showed fewer SNPs compared to other assembly pairs suggesting that these two methods were more similar. Both *de novo* and top mapped reference methods also had far fewer failed assemblies compared to the other assembly methods as shown in Supplementary Table S9.

To assist in the discovery of recombination events where the *omp*A genotype and the genome backbone are different, information from the mapping coverage and quality, and *omp*A genotyping outputs (Supplementary Tables S8 and S5, respectively) can be informative. For this analysis, the reference strain with the top mapped reads and mapping quality is

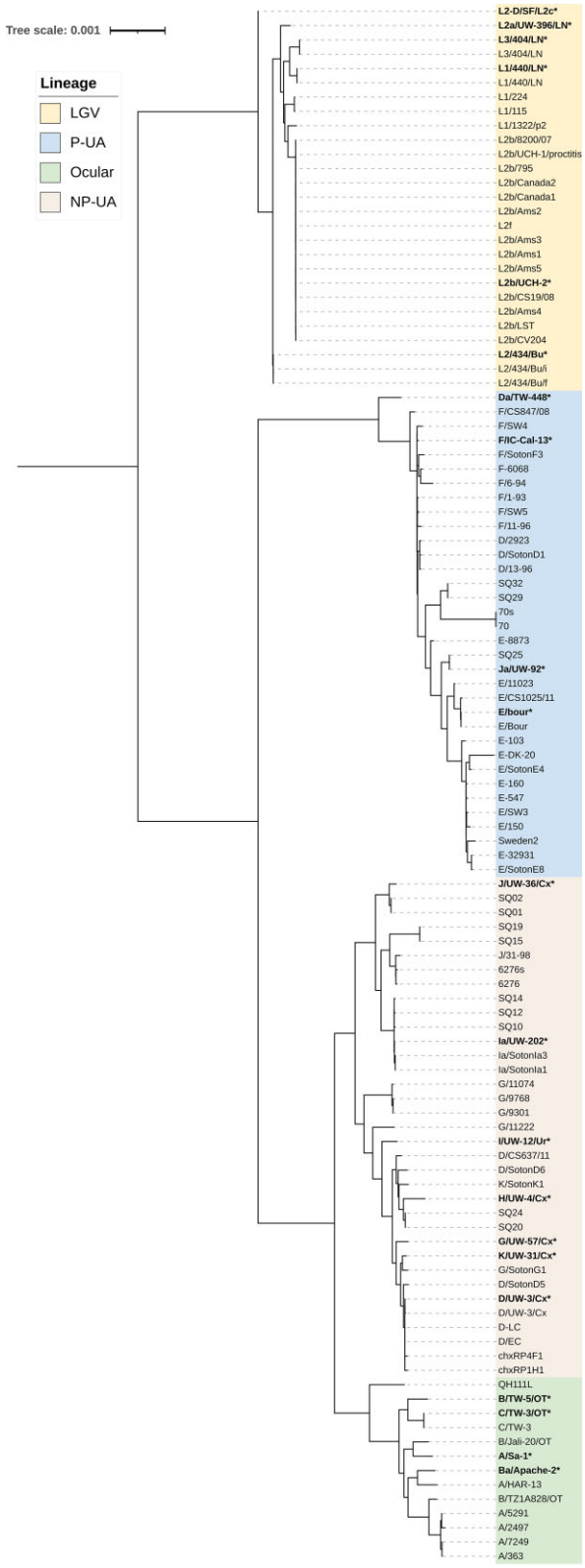

Figure 2. Whole genome phylogeny of *C. trachomatis* sequences available from NCBI ($n = 92$) in addition to the 21 *Ct* reference genomes from this study. The CtGAP generated tree was visualized, and metadata added in iTOL (40). *Ct* reference strains are shown in bold followed by an asterisk. Highlighted lineages: light yellow, LGV, lymphogranuloma venereum strains; light blue, P-UA, prevalent urogenital and anorectal strains; beige, NP-UA, non-prevalent urogenital and anorectal strains; and light green, ocular strains. The scale bar represents the substitutions per site.

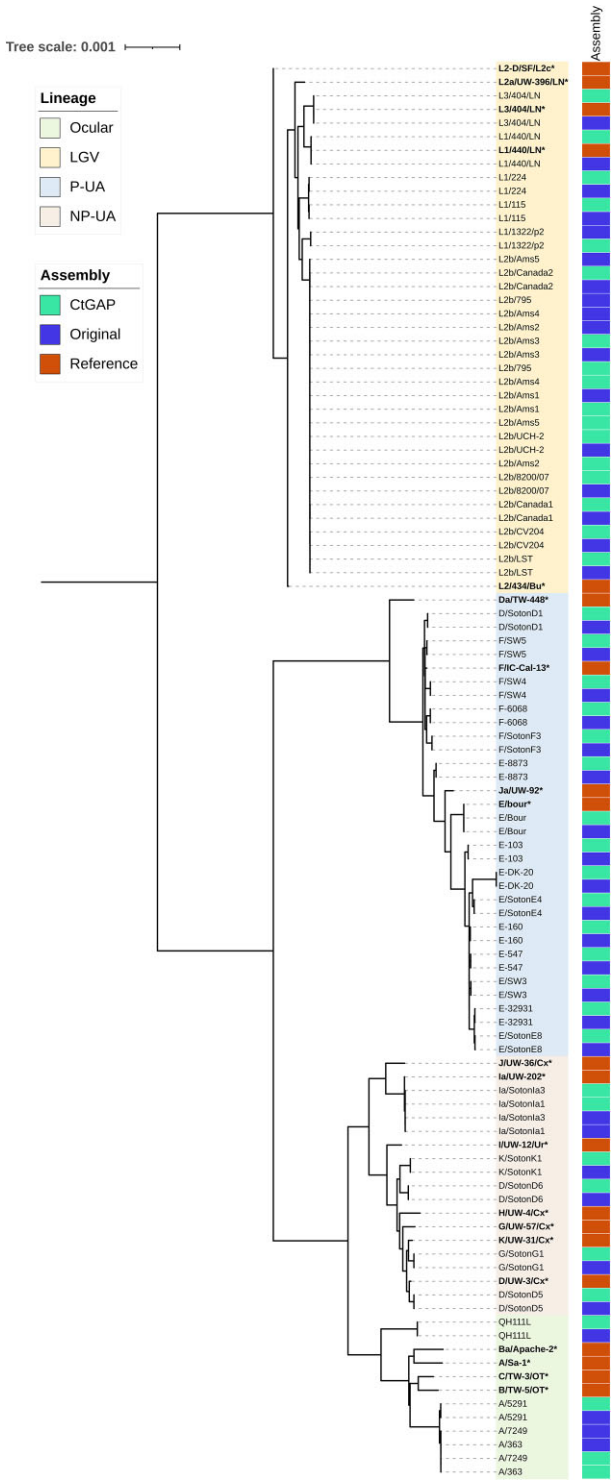

Figure 3. Whole genome phylogeny comparison of 42 CtGAP *de novo* assembled genomes to their original publicly available genome assemblies using the respective SRA from NCBI. The comparison of CtGAP generated genomes to those available in NCBI's genome database was performed using MuMmer's dnadiff option (39). The CtGAP generated tree was visualized, and metadata added in iTOL (40). Assemblies in red represent the *Ct* reference strains from this study; those in bright green are the CtGAP assembled genomes; and the ones in dark blue are the original genome assemblies from NCBI. Highlighted lineages: light yellow, LGV, lymphogranuloma venereum strains; light blue, P-UA, prevalent urogenital and anorectal strains; beige, NP-UA, non-prevalent urogenital and anorectal strains; and light green, ocular strains. The scale bar represents the substitutions per site.

recorded and matched with the *omp*A genotype. For example, genomes previously described as recombinants with *omp*A genotypes different from their genome backbones were confirmed by CtGAP, including SRR25447214, SRR25447288–SRR25447289 and SRR25447265–SRR25447267, where the *omp*A genotype was a Ja but the genome backbone was an strain E (13), and ERR3288031–ERR3288035 with a Da *omp*A genotype and an strain L_2b genome backbone (45).

Discussion

The lack of a standardized workflow for assembling *Ct* genomes continues to limit WGS adoption in the field. This study therefore describes the development of CtGAP, a simple-to-use open-source and easily customizable tool for the assembly of *Ct* genomes and subsequent analyses, thus reducing the level of hands-on time required for processing and characterizing *Ct* genomes.

To effectively process WGS paired end read data from *Ct* isolates and clinical samples sequenced on the Illumina sequencing platform, CtGAP harnesses the power of several standalone genomic tools (i.e., Fastp, Scrubby, Kraken2 Bowtie2, Samtools, BEDTools, Shovill) into a single easy-to-use pipeline. CtGAP also integrates key sequence analysis tools such as SKA, Augur, BLAST and pyMLST to reduce the need for transferring data across multiple platforms thus simplifying downstream processing of the assembled genomes. By incorporating all the above tools into a single pipeline, users can eradicate potential tool-based assembly/analysis bias.

The inclusion of two rounds of Fastp in the CtGAP pipeline was intended to allow users to obtain raw read statistics preprocessing and after adapter trimming and depletion of unwanted reads. This provides the user comparative metrics of the raw reads compared to reads available for assembly. Both outputs are available in either text or html formats. An important feature of the CtGAP pipeline is the incorporation of Scrubby for the depletion of unwanted reads. Scrubby was used to facilitate a fast k-mer based depletion of all reads classified as Archaea, Eukaryota, Holozoa and Nucletmycea with Kraken2. This was then followed up with a further alignment of the remaining reads against the human reference genome (GRCh38) with minimap2 to further deplete any possible left-over human genome contaminants. Scrubby was also used to extract all reads mapping to the Chlamydiales order. CtGAP stores its intermediate products. such as host-depleted reads, in appropriately named directories to facilitate their usage in other genome analysis or assembly tools such as Bactopia (41).

CtGAP is uniquely equipped with the ability to produce two types of genome assemblies per sample: *de novo* and reference-guided assemblies. For the reference-guided assembly, the user can choose between a plurality consensus sequence or specific reference strain-based assembly. With the understanding that a reference based genome assembly is only as good as the reference utilized in the mapping step (11) and to ensure optimum quality reference-based assembly, we re-sequenced 20 reference *Ct* strains using Illumina and nanopore platforms, 10 of which also had a Roche 454 genome assembly available with gap closure. All available assemblies—with the addition of any reference strains from NCBI when only two assemblies were available—were then used to manually generate a quality sequence for each reference strain. These sequences are available for use in CtGAP as is (for single reference strain-

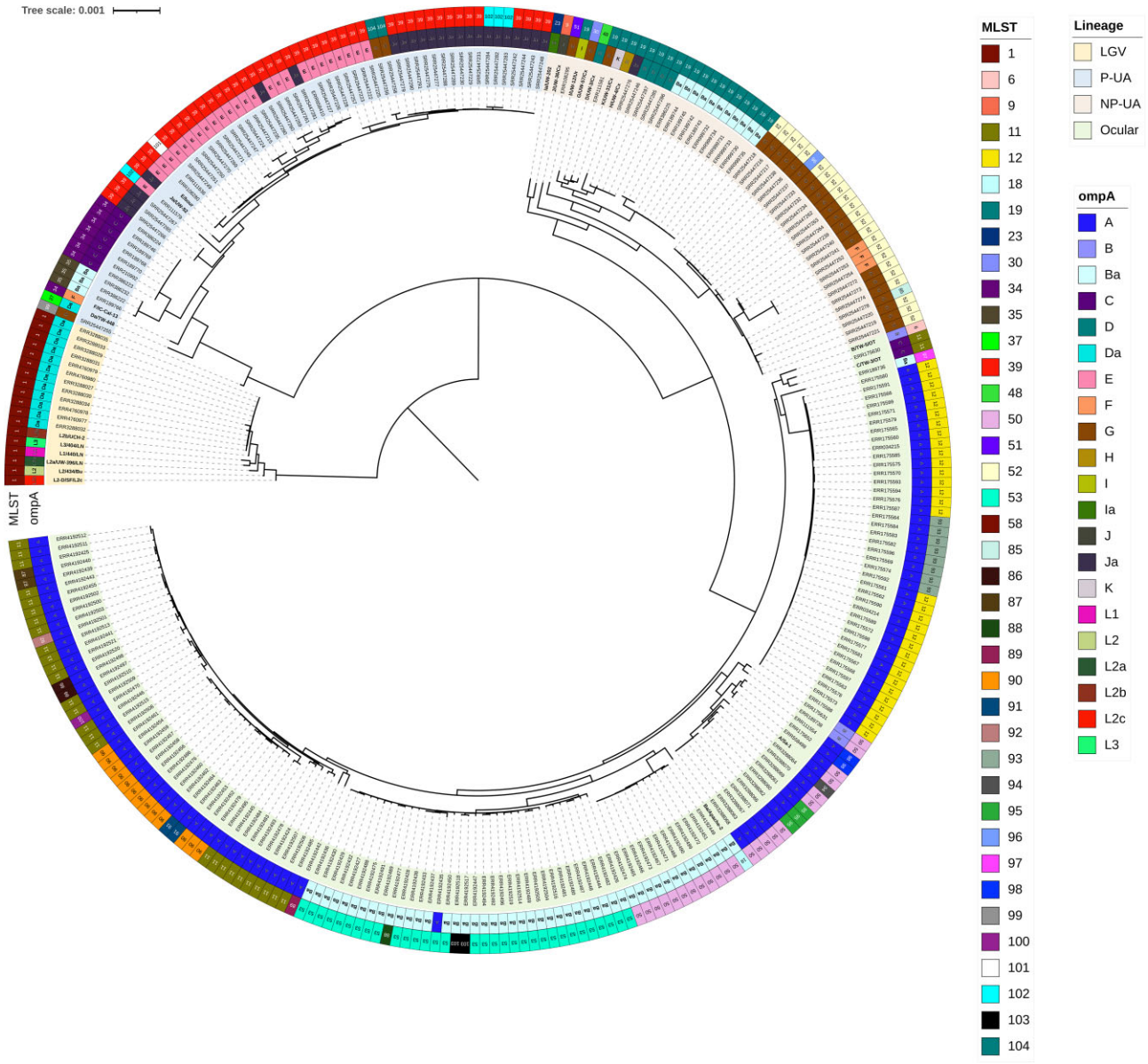

Figure 4. Phylogeny of *C. trachomatis* genomes generated by CtGAP ($n = 271$) using *de novo* assembly and showing MLST (outer circle) and *omp*A genotype (2nd circle from outside) metadata (see Supplementary Table S6). The CtGAP generated tree was visualized, and metadata added in iTOL (40). The color coding for the *omp*A genotypes and MLSTs are to the right of the tree in separate columns. *Ct* reference genome strains are in bold font (3rd circle from outside). Highlighted lineages (3rd circle from outside): light yellow, LGV, lymphogranuloma venereum strains; light blue, P-UA, prevalent urogenital and anorectal strains; beige, NP-UA, non-prevalent urogenital and anorectal strains; and light green, ocular strains. The scale bar represents the substitutions per site.

based assembly) or in the form of a plurality consensus-based assembly.

In addition to the *de novo* assembly mode, CtGAP can perform three different reference guided assemblies using the plurality consensus, a predetermined/generic reference, or a CtGAP informed reference. The first two can be performed alongside the *de novo* assembly. As a deliverable, CtGAP provides coverage report statistics of the samples against 21 reference *Ct* strains. This information can facilitate an informed selection of a reference genome (the top-mapped reference) that closely matches the genome backbone of the sample for an optional second iteration of reference-guided assembly on CtGAP. This latter method is especially useful for assembling unknown clinical *Ct* strains and identifying recombinants where there can be two top-matched reference strains that, through two interactions, ultimately optimize the assembly.

Assembly statistics such as total number of bases, contigs and GC content is also reported for every genome. When both *de novo* and reference-guided assemblies are performed, a report comparing both assemblies is produced by CtGAP and contains key information such as sequence length, identity, number and position of SNPs between assemblies. These statistics allow users to easily determine the quality of the assembled genomes without the need to export the genome to standalone QC tools.

With our testing dataset, the different assembly methods do not produce identical genome assemblies with differences between assembly pairs varying between one to as many as

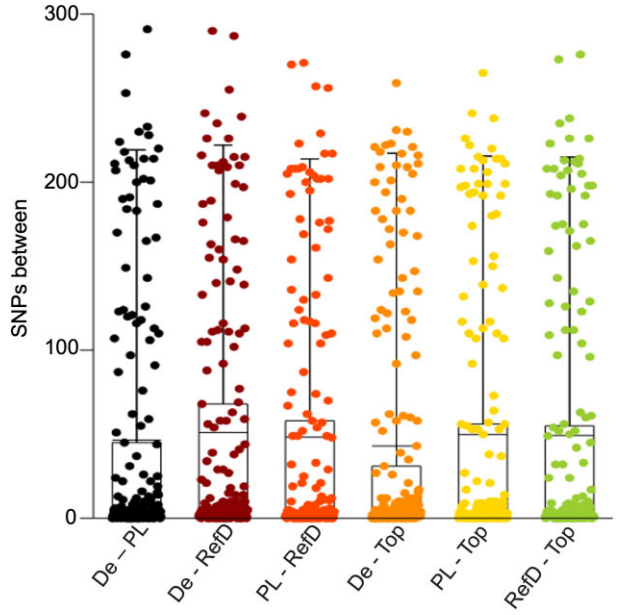

Figure 5. Comparison of SNPs between different assembly pairs for 271 samples. The four assemblies performed in CtGAP for the 271 samples included *de novo*, reference-guided using the top mapped *Ct* reference genome, reference-guided using only the *Ct* reference strain D/UW-3/Cx and the plurality sequence generated from the 21 *Ct* reference genome strains in this study. SNPs between any pair of assemblies were identified using MuMmer's dnadiff option tool (39). The Boxplot shows the 5–95% confidence intervals of data distribution with mean values indicated by horizontal bars. Generic ref, *Ct* reference genome strain D/UW-3/Cx; top mapped ref, *Ct* reference strain with the most mapped reads and quality using the Bowtie2 mapping tool in CtGAP. De–PL, de novo—plurality; De–RefD, de novo—reference strain D/UW-3/CX; PL–RefD, plurality—reference strain D/UW-3/CX; De–Top, de novo—top mapped reference strain; PL–Top, plurality-top mapped reference strain; and RefD–Top, reference strain D/UW-3/CX—top mapped reference strain.

495 SNPs. The *de novo* and top-mapped reference guided assemblies were more similar than any other pair of assemblies, suggesting that the best assembly will be the use of either the *de novo* assembly method or the top-mapped reference-guided assembly method. The differences between the assembly methods further illustrates the need to select the appropriate *Ct* reference strain and not just a commonly used strain such at D/UW-3/Cx, L_2/434/Bu or A/HAR-13. This will provide a more robust assembly to guarantee reproducibility globally. As CtGAP is equipped with the resources to simultaneously generate both *de novo* and reference-guided assemblies for *Ct* genomes, it is a promising tool to study the effects of assembly methods on *Ct* genomes in future studies.

A limitation of the CtGAP workflow is the lack of capacity to handle NGS reads generated from the nanopore sequencing platform. The use of nanopore in sequencing *Ct* genomes is currently in its infancy with ∼5 *Ct* nanopore genome assemblies currently available in online databases. The lack of a robust training dataset made it impossible to include nanopore capabilities in the current iteration of CtGAP. In addition, the pipeline can't assemble *Ct* strains that occur as mixed infections as we recently reported (12). However, these mixed infections can be identified based on QUAST statistics.

In summary, the adoption of NGS data analysis in the study of *Ct* pathogenesis and epidemiology is increasing rapidly. However, there are currently no *Ct*-specific tools designed to handle the first and most important step—genome assembly. CtGAP provides an effective workflow to assemble *Ct* genomes from raw NGS reads. CtGAP also provides the user with initial data for comparative analyses on the assembled genomes and on typing strategies for comparison with current global data that are only resolved to the single *omp*A gene or MLST gene level. While CtGAP is currently designed specifically for *Ct* genomes, it is easily customizable to handle other *Chlamydia* species and bacteria.

Data availability

The genome sequence data generated as part of this study are freely available at the NCBI's Genome database BioProject PRJNA1137892. The accession codes for all data described in this study are provided in the Supplementary Tables S1–S4. A comprehensive CtGAP how-to documentation along with the source code is available at https:doi.org/10.5281/zenodo.14511460 and https://github.com/D-Dean-Lab/CtGAP.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors would like to acknowledge Parul Sharma and Morgan Dehdashti for their feedback on the manuscript.

Funding

National Institutes of Health [R01AI151075 and R01AI158527 to D.D. and T.D.R.].

Conflict of interest statement

None declared.

References

1. World Health Organization (2016) Department of Reproductive Health and Research. In: *Global Health Sector Strategy on Sexually Transmitted Infection 2016-2021: Towards ending STIs*. https://www.who.int/publications/i/item/WHO-RHR-16.09.
2. Wang,S.P., Kuo,C.C., Barnes,R.C., Stephens,R.S. and Grayston,J.T. (1985) Immunotyping of *Chlamydia trachomatis* with monoclonal antibodies. *J. Infect. Dis.*, **152**, 791–800.
3. Millman,K., Black,C.M., Johnson,R.E., Stamm,W.E., Jones,R.B., Hook,E.W., Martin,D.H., Bolan,G., Tavaré,S. and Dean,D. (2004) Population-based genetic and evolutionary analysis of *Chlamydia trachomatis* urogenital strain variation in the United States. *J. Bacteriol.*, **186**, 2457–2465.
4. Dean,D., Bruno,W.J., Wan,R., Gomes,J.P., Devignot,S., Mehari,T., de Vries,H.J., Morré,S.A., Myers,G., Read,T.D., *et al.* (2009) Predicting phenotype and emerging strains among *Chlamydia trachomatis* infections. *Emerg. Infect. Dis.*, **15**, 1385–1394.
5. Pannekoek,Y., Morelli,G., Kusecek,B., Morré,S.A., Ossewaarde,J.M., Langerak,A.A. and van der Ende,A. (2008) Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria *Chlamydia trachomatis*. *BMC Microbiol.*, **8**, 42.
6. Smelov,V., Vrbanac,A., van Ess,E.F., Noz,M.P., Wan,R., Eklund,C., Morgan,T., Shrier,L.A., Sanders,B., Dillner,J., *et al.* (2017)

Chlamydia trachomatis strain types have diversified regionally and globally with evidence for recombination across geographic divides. *Front. Microbiol.*, **8**, 2195.
7. Price,V., Ngwira,L.G., Lewis,J.M., Baker,K.S., Peacock,S.J., Jauneikaite,E. and Feasey,N. (2023) A systematic review of economic evaluations of whole-genome sequencing for the surveillance of bacterial pathogens. *Microb. Genom.*, **9**, mgen000947.
8. Simar,S.R., Hanson,B.M. and Arias,C.A. (2021) Techniques in bacterial strain typing: past, present, and future. *Curr. Opin. Infect. Dis.*, **34**, 339–345.
9. Uelze,L., Grützke,J., Borowiak,M., Hammerl,J.A., Juraschek,K., Deneke,C., Tausch,S.H. and Malorny,B. (2020) Typing methods based on whole genome sequencing data. *One Health Outlook*, **2**, 3.
10. Somboonna,N., Wan,R., Ojcius,D.M., Pettengill,M.A., Joseph,S.J., Chang,A., Hsu,R., Read,T.D. and Dean,D. (2011) Hypervirulent *Chlamydia trachomatis* clinical strain is a recombinant between lymphogranuloma venereum (L (2)) and D lineages. *mBio*, **2**, e00045-11.
11. Valiente-Mullor,C., Beamud,B., Ansari,I., Francés-Cuesta,C., García-González,N., Mejía,L., Ruiz-Hueso,P. and González-Candelas,F. (2021) One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput. Biol.*, **17**, e1008678.
12. Joseph,S.J., Bommana,S., Ziklo,N., Kama,M., Dean,D. and Read,T.D. (2023) Patterns of within-host spread of *Chlamydia trachomatis* between vagina, endocervix and rectum revealed by comparative genomic analysis. *Front. Microbiol.*, **14**, 1154664.
13. Bowden,K.E., Joseph,S.J., Cartee,J.C., Ziklo,N., Danavall,D., Raphael,B.H., Read,T.D. and Dean,D. (2021) Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from patient clinical vaginal and rectal swabs. *mSphere*, **6**, e01302-20.
14. Harris,S.R., Clarke,I.N., Seth-Smith,H.M., Solomon,A.W., Cutcliffe,L.T., Marsh,P., Skilton,R.J., Holland,M.J., Mabey,D., Peeling,R.W., *et al.* (2012) Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nat. Genet.*, **44**, 413–419.
15. Read,T.D., Brunham,R.C., Shen,C., Gill,S.R., Heidelberg,J.F., White,O., Hickey,E.K., Peterson,J., Utterback,T., Berry,K., *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.*, **28**, 1397–1406.
16. Alkhidir,A.A.I., Holland,M.J., Elhag,W.I., Williams,C.A., Breuer,J., Elemam,A.E., El Hussain,K.M.K., Ournasseir,M.E.H. and Pickering,H. (2019) Whole-genome sequencing of ocular *Chlamydia trachomatis* isolates from Gadarif State, Sudan. *Parasit Vectors*, **12**, 518.
17. Hadfield,J., Harris,S.R., Seth-Smith,H.M.B., Parmar,S., Andersson,P., Giffard,P.M., Schachter,J., Moncada,J., Ellison,L., Vaulet,M.L.G., *et al.* (2017) Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. *Genome Res.*, **27**, 1220–1229.
18. Andersson,P., Harris,S.R., Smith,H., Hadfield,J., O'Neill,C., Cutcliffe,L.T., Douglas,F.P., Asche,L.V., Mathews,J.D., Hutton,S.I., *et al.* (2016) *Chlamydia trachomatis* from Australian aboriginal people with trachoma are polyphyletic composed of multiple distinctive lineages. *Nat. Commun.*, **7**, 10688.
19. Somboonna,N., Mead,S., Liu,J. and Dean,D. (2008) Discovering and differentiating new and emerging clonal populations of *Chlamydia trachomatis* with a novel shotgun cell culture harvest assay. *Emerg. Infect. Dis.*, **14**, 445–453.
20. Bushnell,B., Rood,J. and Singer,E. (2017) BBMerge—accurate paired shotgun read merging via overlap. *PLoS One*, **12**, e0185056.
21. Prjibelski,A., Antipov,D., Meleshko,D., Lapidus,A. and Korobeynikov,A. (2020) Using SPAdes de novo assembler. *Curr. Protoc. Bioinformatics*, **70**, e102.
22. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.
23. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
24. Alonge,M., Lebeigle,L., Kirsche,M., Jenike,K., Ou,S., Aganezov,S., Wang,X., Lippman,Z.B., Schatz,M.C. and Soyk,S. (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.*, **23**, 258.
25. Delahaye,C. and Nicolas,J. (2021) Sequencing DNA with nanopores: troubles and biases. *PLoS One*, **16**, e0257521.
26. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
27. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.*, **20**, 257.
28. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
29. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
30. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
31. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
32. Jolley,K.A., Bray,J.E. and Maiden,M.C.J. (2018) Open-access bacterial population genomics: bIGSdb software, the PubMLST.Org website and their applications. *Wellcome Open Res.*, **3**, 124.
33. Versteeg,B., Bruisten,S.M., Pannekoek,Y., Jolley,K.A., Maiden,M.C.J., van der Ende,A. and Harrison,O.B. (2018) Genomic analyses of the *Chlamydia trachomatis* core genome show an association between chromosomal genome, plasmid type and disease. *BMC Genomics*, **19**, 130.
34. Boetzer,M. and Pirovano,W. (2012) Toward almost closed genomes with GapFiller. *Genome Biol.*, **13**, R56.
35. Bouras,G., Grigson,S.R., Papudeshi,B., Mallawaarachchi,V. and Roach,M.J. (2024) Dnaapler: a tool to reorient circular microbial genomes. *J. Open Source Software*, **9**, 5968.
36. Angiuoli,S.V. and Salzberg,S.L. (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334–342.
37. Lemoine,F. and Gascuel,O. (2021) Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genom.Bioinform.*, **3**, lqab075.
38. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
39. Marçais,G., Delcher,A.L., Phillippy,A.M., Coston,R., Salzberg,S.L. and Zimin,A. (2018) MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.*, **14**, e1005944.
40. Letunic,I. and Bork,P. (2021) Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
41. Petit,R.A. and Read,T.D. (2020) Bactopia: a flexible pipeline for complete analysis of bacterial genomes. *mSystems*, **5**, e00190-20.
42. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2011) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
43. Huddleston,J., Hadfield,J., Sibley,T.R., Lee,J., Fay,K., Ilcisin,M., Harkins,E., Bedford,T., Neher,R.A. and Hodcroft,E.B. (2021)

Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *J. Open Source Softw.*, **6**, 2906.
44. Antipov,D., Korobeynikov,A., McLean,J.S. and Pevzner,P.A. (2015) hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, **32**, 1009–1015.
45. Borges,V., Cordeiro,D., Salas,A.I., Lodhia,Z., Correia,C., Isidro,J., Fernandes,C., Rodrigues,A.M., Azevedo,J., Alves,J., *et al.* (2019) *Chlamydia trachomatis*: when the virulence-associated genome backbone imports a prevalence-associated major antigen signature. *Microb. Genom.*, **5**, e000313.

Received: August 13, 2024. **Revised:** December 10, 2024. **Editorial Decision:** December 16, 2024. **Accepted:** December 18, 2024