Global Multilocus Sequence Type Analysis of *Chlamydia trachomatis* Strains from 16 Countries

Björn Herrmann, a Jenny Isaksson, a Martin Ryberg, b Jeanette Tångrot, c Isam Saleh, a Bart Versteeg, d Kirsten Gravningen, e Sylvia Bruisten, d on behalf of The *Chlamydia trachomatis* MLST Study Group

Section of Clinical Bacteriology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden; Department of Organismal Biology, Uppsala University, Uppsala, Sweden; Bioinformatics Infrastructure for Life Sciences (BILS) and Computational Life Science Cluster (CLiC), Department of Molecular Biology, Umeå University, Umeå, Sweden; Public Health Laboratory, Public Health Service of Amsterdam, Amsterdam, the Netherlands; Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway

The Uppsala University *Chlamydia trachomatis* multilocus sequence type (MLST) database (http://mlstdb.bmc.uu.se) is based on five target regions (non-housekeeping genes) and the *ompA* gene. Each target has various numbers of alleles—hctB, 89; CT058, 51; CT144, 30; CT172, 38; and *pbpB*, 35—derived from 13 studies. Our aims were to perform an overall analysis of all *C. trachomatis* MLST sequence types (STs) in the database, examine STs with global spread, and evaluate the phylogenetic capability by using the five targets. A total of 415 STs were recognized from 2,089 specimens. The addition of 49 *ompA* gene variants created 459 profiles. ST variation and their geographical distribution were characterized using eBURST and minimum spanning tree analyses. There were 609 samples from men having sex with men (MSM), with 4 predominating STs detected in this group, comprising 63% of MSM cases. Four other STs predominated among 1,383 heterosexual cases comprising, 31% of this group. The diversity index in ocular trachoma cases was significantly lower than in sexually transmitted chlamydia infections. Predominating STs were identified in 12 available *C. trachomatis* whole genomes which were compared to 22 *C. trachomatis* full genomes without predominating STs. No specific gene in the 12 genomes with predominating STs could be linked to successful spread of certain STs. Phylogenetic analysis showed that MLST targets provide a tree similar to trees based on whole-genome analysis. The presented MLST scheme identified *C. trachomatis* strains with global spread. It provides a tool for epidemiological investigations and is useful for phylogenetic analyses.

*Chlamydia trachomatis* is one of the most common sexually transmitted infections (STIs) worldwide (1), and besides urogenital infections, it also causes lymphogranuloma venereum (LGV), which is a rare but more invasive sexually transmitted disease. In addition, *C. trachomatis* causes the eye infection trachoma, which is the major infectious cause of preventable blindness worldwide. Severe sequelae from urogenital chlamydia infections include ectopic pregnancy and infertility (2). In spite of testing, treatment, partner notification, and counseling, huge public health efforts have not been able to control urogenital chlamydia infections. Current knowledge about the role of repeated infections and transmission in sexual networks is still limited and needs to be extended to achieve a reduction in the rate of infections.

In this context, it is important to have adequate tools for genotyping to understand the epidemiology of chlamydia infections. Traditional typing of *C. trachomatis* was based on serotyping of the major outer membrane protein (MOMP) and, later on, genotyping of the *ompA* gene, which encodes MOMP. However, neither MOMP nor *ompA* provides sufficient discriminatory power for epidemiological purposes (3). In most countries, almost half of all urogenital chlamydia infections are of serotype E, and within this serotype the *ompA* E/Bour genotype predominates (4–8). Therefore, other typing methods were developed, such as multilocus variable-number tandem-repeat analysis (MLVA) (8, 9) and multilocus sequence typing (MLST) (10–12).

MLST relies on PCR amplification and DNA sequencing of several genomic loci. There are three such schemes described for *C. trachomatis*. Two of them are based on housekeeping genes, have a resolution similar to that of *ompA* sequencing, and are suitable for evolutionary studies (11, 12). The third scheme was developed by Klint et al. (10); it is intended for short-term clinical epidemiology and outbreak investigations and is based on five highly variable genomic loci (*hctB* [CT046], CT058, CT144, CT172, and *pbpB* [CT682]). The MLST scheme has been slightly modified to facilitate rational processing with only 1% loss of discriminatory capacity (13), and the current setup can be found at the Uppsala University *C. trachomatis* MLST database (http://mlstdb.bmc.uu.se).

Since its development in 2007, the scheme has been applied to a variety of clinical specimens, including samples from cases of urogenital chlamydia (7, 13–21), LGV (22, 23), and trachoma (24). It has been useful for different purposes: (i) identification of clonal spread of LGV (22) and the new variant of *C. trachomatis*
(16, 17); (ii) differentiation of strains within the predominating serotype, E (7, 17); (iii) differentiation of strains infecting men who have sex with men (MSM) and heterosexuals (15, 19); (iv) investigation of the role of tissue tropism (21); and (v) molecular epidemiology and antibiotic treatment of trachoma (24). At present, our database comprises 415 unique sequence types (STs) derived from 2,089 C. trachomatis specimens.

The aims of the current study were to (i) perform an overall analysis of all C. trachomatis MLST STs in the current database, (ii) identify and examine STs with global spread, and (iii) evaluate the use of the MLST targets for phylogenetic analysis.

MATERIALS AND METHODS

Specimens. A total of 2,089 entries were obtained from 13 studies with different objectives and are available at the Uppsala University C. trachomatis MLST database (http://mlstdb.bmc.uu.se).

Study populations. Samples from studies with specific objectives to investigate chlamydia infections in MSM are here designated samples and comprised samples from trachoma patients (24).

PCR amplification, sequencing, and genotyping. The currently used protocol for PCR amplification, sequencing, and genotyping of the five highly variable target regions included in this MLST scheme is available at the database above. Study sites could also use their own amplification and sequencing protocols as long as the trimmed sequences of the loci were included from primer to primer. In addition to the five MLST targets, ompA sequence data were also included.

D. The diversity index (D) of a typing method refers to the probability that two unrelated strains sampled from the test population will be placed into different typing groups. The D was determined for entries in the MLST database using Hunter and Gaston’s modification of Simpson’s diversity index (25). A cluster was defined as a group of STs differing by not more than one locus from another ST.

eBURST analysis. eBURST v.3 (http://eburst.mlst.net/; Department of Infectious Disease Epidemiology, Imperial College London, United Kingdom) software was used to identify founders among the sequence types. A founder was defined as the ST with the largest number of single-locus variants in a group. An ST that appears to have diversified to produce multiple single-locus variants is called a subgroup founder. A list of all STs was inserted into the single data set function at the eBURST website, and the number of loci was set to 5 (ompA not included). The analysis was computed generating groups and predicted founders, and for each larger group a diagram was drawn.

MFST analysis. BioNumerics software (version 7.0; Applied Maths, Sint-Martens-Latem, Belgium) was used to construct a minimum spanning tree (MST) of all entries in the database. The sequence data for all 2,089 samples (ompA not included) was entered into a database within the BioNumerics software. As the algorithm, we used the predefined template “MST for categorical data” plug-in, which uses the categorical coefficient to calculate the similarity matrix. This calculates a standard MST with single- and double-locus variants.

Whole-genome analysis. C. trachomatis genomes were downloaded from GenBank (6 draft genomes and 122 complete genomes and plasmids) and from the Ribosomal Multilocus Sequence Typing (rMLST) database (http://pubmlst.org/rmlst/), containing 397 genomes). MLST was performed on all genomes using the approach by Larsen et al. (26), with source code kindly provided by the authors and using our own MLST scheme. We selected a subset of genomes for further investigation. One to three representative full genomes were chosen for each of 5 founder STs (see Results), for a total of 12 genomes. The number of genomes for each ST depended on the number of genomes including this particular ST. Complete genomes were selected preferably instead of draft genomes. We also selected 22 genomes with other STs, in order to have a large number of profiles represented for comparative analyses. See Table S1 in the supplemental material for a list of genomes in the subset. For the genomes lacking annotations, coding sequences were predicted using Prodigal version 2.6 (27).

Identifying clusters of genes. The program OrthoMCL version 2.0 (28) was used to cluster genes based on all-against-all blastp searches. Since the genomes of C. trachomatis are highly conserved between strains, a high level of similarity is expected within groups. The blast comparison was performed without filtering for sequences with low information content, to avoid artificially short matches of nearly identical sequences. We required at least 90% sequence identity within groups and used an inflation index of 1.5 for the clustering. Members of the groups containing only sequences from genomes with founder STs were searched against the NCBI nucleotide database using blastp and standard settings. Typical best matches for each group, with 100% sequence identity in the aligned area when not otherwise stated, are listed in Table S2 in the supplemental material.

Phylogenetic analyses. Two data sets based on the MLST genes were analyzed and compared with the phylogenomic analysis in the work of Harris et al. (29). The first data set consisted of the MLST STs from the same strains as in the work of Harris et al. (referred to here as the Harris et al. data set). The second data set also included all unique STs in the MLST database (http://mlstdb.bmc.uu.se; referred to here as the high-resolution [hr]-CT-MLST data set). For both data sets the individual genes were aligned by eye using Seaview (30). Potential recombinations between strains within genes were detected using RDP4 at default settings. To reduce the number of ambiguously aligned regions and number of gap positions, the poly(G) and intron regions of CT172, and the repetitive elements of hctB were removed. CT144 was removed from further analysis since it included a large amount of potential recombinations in both data sets. A phylogenetic tree was constructed for each gene separately using RAxML (31) with the GTR plus gamma model. Support values were estimated with 100 rapid bootstraps. The trees were rooted using the L strains as the outgroup. Sequences in clades separated from the rest by long branches (>0.05) were excluded to reduce long-branch attraction, and the phylogenetic analysis was redone. For the Harris et al. data set, conflicts between gene trees that had bootstrap support (BS) above 70 (in both gene trees) were identified and resolved by removing the least number of gene sequences. The same was done for the hr-CT-MLST data set, but using a cutoff of 65 to include apparent conflicts that were not identified using the cutoff of 70. The gene alignments were then concatenated and analyzed with RAxML. The GTR model and gamma distribution were used with parameters estimated separately for each gene. Support values were estimated using 10,000 rapid bootstraps. For the hr-CT-MLST data set, this analysis is referred to as “no conflict.”

Ethics. Ethic permissions were obtained for each individual study. The present study includes no information that could identify individual patients.

RESULTS

The 2,089 entries in our database comprised 13 studies from 16 countries: the Netherlands (39%), Sweden (16%), Norway (12%), Suriname (8%), Tunisia (5%), China (4%), the United States (4%), Argentina (4%), Gambia and Senegal (4%), and France, Chile, Denmark, Spain, Australia, and Germany (all ≤1%). The collection of clinical specimens spanned more than 19 years (1992 to 2011), and reference strains were isolated from 1958 onwards. Overall, 609 (29%) of the database samples originated from MSM (75% of samples from rectum), 1,383 (66%) urogenital samples were assumed to be from heterosexuals, and 75 (4%) ocular sam-
FIG 1 (A and B) Minimum spanning tree analysis of the entire database based on the five MLST target regions visualized by *ompA* genovar distribution (A) and by sexual behavior (B). Sphere sizes indicate the numbers of samples in each sphere. Solid branches show single-locus variants, and dashed branches show double-locus variants. The most common STs are indicated by number. (C) Minimum spanning tree showing the geographical ST distribution of 1383 heterosexual samples, 75 trachoma samples (Gambia and Senegal), and 15 reference strains in the MLST database. Sphere size is proportional to the number of samples of each sequence type. Solid branches show single-locus variants, dashed branches show double-locus variants, and dotted branches show triple-locus variants.
The present work is a compilation of 13 studies with different design and sampling strategies, and the study populations are not representative for the general sexually active population in the 16 countries. To compare the ompA genotype distribution in our study with the ompA distribution in previous studies, we performed ompA sequencing of the 1,383 samples assumed to be from heterosexuals, with the following results: E, 45% (n = 617); F, 16% (n = 217); D, 13% (n = 173); G, 10% (n = 136); K, 5.8% (n = 80); I, 5.4% (n = 75); J, 3.9% (n = 54); H, 2.3% (n = 32); and B, 0.01% (n = 10).

Differences between C. trachomatis typing results from MSM and heterosexuals are summarized in Table 2.

The D for 75 trachoma specimens was lower (0.772 [0.742 to 0.803]) than for urogenital samples (0.968 [0.966 to 0.970]). The number of ompA variants was 10, and the number of STs using the five target regions in MLST was 12, while a combination of ompA and MLST resulted in 19 STs.

Whole-genome analysis of strains with predominating MLST STs. The distribution pattern of STs clearly shows that a few strains have been successful in global spread in different populations. Concurrently, many strains are single-locus variants to the

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MSM population</th>
<th>Heterosexual population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most predominant ompA genovars</td>
<td>D (26%, n = 160)</td>
<td>D (12%, n = 170); P &lt; 0.001 for the five serotypes</td>
</tr>
<tr>
<td>No. clusters</td>
<td>A few large clusters</td>
<td>Multiple clusters of various size</td>
</tr>
<tr>
<td>Overlap with clusters in the other risk group</td>
<td>Almost absent</td>
<td>A small proportion</td>
</tr>
<tr>
<td>Subpopulations within clusters</td>
<td>No</td>
<td>Ethnicity</td>
</tr>
<tr>
<td>Geographic variation within clusters</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

a ompA data missing for 4 MSM.

b See Fig. 1B.

c Clusters in MST analysis with specimens from MSM (red in Fig. 1B) have almost no specimens from the suggested heterosexuals, while in clusters predominated by heterosexuals (blue), a smaller proportion of specimens come from MSM.

d Exemplified by studies described in references 18 and 20.

e See Fig. 1C.
founders and subgroup founders identified by eBURST analysis and seem to be genetically closely related to the predominating strains. Twelve *C. trachomatis* genomes containing any of the predominating STs (ST3, ST12, ST52, ST56, and ST108), here designated founders, were compared with 22 other available *C. trachomatis* genomes to investigate if specific genes or proteins could be identified in the strains with founder STs. No genome containing ST109 was yet available in GenBank.

After OrthoMCL analysis, where the proteins of the 34 selected genomes were clustered into groups of orthologues, 17 groups of proteins were identified as unique for strains with any of the founder STs. However, no group contained proteins from all genomes with founder STs. There were 13 groups containing protein sequences from more than one genome. Four of these groups consisted of sequences that are fragments of longer proteins. Most of the groups are similar to hypothetical proteins; only three of these are from *C. trachomatis*, while the others are from eukaryotic organisms (see Table S2 in the supplemental material). The analysis could not identify any specific gene that was common for strains with predominating STs but missing in other strains.

**Phylogenetic analysis.** With regard to the Harris et al. data set, B/TZ1A828/OT was on a long branch in *pbpB*, indicating highly divergent sequences, and was therefore excluded for this gene. Three cases with conflicts between gene trees were identified. The placement of L1 1322-p2 was inconsistent between CT058 and CT172 and was removed from CT172. D/SotonD6 and G11222 clustered differently in CT058 and *pbpB* and were removed from *pbpB*.

The concatenated tree was largely in agreement with the full genome tree of Harris et al. (29), but with less resolution and lower bootstrap support values (Fig. 2). There are, however, some differences supported by bootstrap values of >70. In the Harris et al. data set, F(s)/70 is the sister to a clade including E, F, and D strains, while it is nested within that clade in the MLST tree. A/HAR-13 and B/jali20 form a clade for the MLST genes, while they constitute the basal grade to a clade dominated by A and B strains in the Harris et al. data set. There are several differences in the clade composing G, D, K, Ia, and J strains. In the clade with L strains, the MLST tree has a clade with L1/224, L1/115, L1/440/LN, and L1/404/LN, while these are part of a basal grade in the phylogeny of Harris et al. (29).

The hr-CT-MLST data set resulted in a tree that has essentially the same topology as obtained when using the Harris et al. data set but with considerably more branches (see Fig. S1 in the supplemental material).

**DISCUSSION**

This report is a unique summary of the most frequently used high-resolution genotyping scheme for *C. trachomatis* to date with the database including more than 2,000 samples from 16 countries worldwide. The distribution of STs clearly shows that a few strains have been successful in spreading in many countries among different populations and thus appears to predominate in both space and time. Different STs predominate among MSM and heterosexual individuals. Currently, multiple strains are single-locus variants of the founders and subgroup founders identified by eBURST.
were not representative for samples from the general population. However, the traits that contribute to successful spread may not be directly linked to the 5 MLST targets or the ompA gene, but these targets may be biomarkers for other genes or traits in strains that are successfully spread worldwide. Our comparison of whole genomes from strains with and without founder STs could not identify any gene that was common for strains with predominating STs but missing in other strains.

The MLST scheme has been used in study populations and provides high resolution for differentiation of C. trachomatis strains (32). A major concern about a typing method is its reliability. Since the 5 MLST targets are not housekeeping genes, but were selected from the most divergent genes in the few genomes available in 2005, it has been questioned if the targets are stable. In previous studies the stability of the included target regions has been demonstrated in laboratory experiments (33) as well as in longitudinal studies of the new variant of C. trachomatis and lymphogranuloma venereum (16, 17, 22, 34). The longitudinal studies also indicated that ompA was more unstable, which is in agreement with recent whole-genome studies that have shown extensive recombination in ompA (29, 35, 36). Our MLST has therefore proved suitable for use in epidemiological studies. Identification of founders in eBURST analysis showed that ST12 was detected in a reference strain collected as early as 1960 (F/IC-cal3), a common and widespread genotype. This further indicates the stability of the MLST targets.

This MLST scheme has shown up to a 5-fold-higher resolution than ompA (7). The D was 0.975 (95% CI, 0.974 to 0.976) when calculated on all entries in the database. In previous studies of sexually transmitted chlamydia infections, the D varied between 0.84 and 0.97 depending on the degree of epidemiological relatedness between samples in the study populations (7, 13, 17, 37). By calculating an overall D for all 2,089 samples, epidemiological relatedness was eliminated. The obtained index was well above 0.95, an “ideal” cutoff value for a molecular typing method (38). The lower D for the predominating clusters in Table 1 is explained by the inclusion of some epidemiologically related samples, especially among MSM who may be connected via extended international sexual networks (15, 19).

Regarding trachoma, the MLST has been used only in a limited number of samples where many cases were epidemiologically linked, explaining why a low D (0.77) was obtained. The number of ompA variants was almost equal to the number of STs based on the five MLST targets (24). This is in contrast to the case with sexually transmitted chlamydia infections, for which up to 5-fold more STs have been found compared to ompA variants (7).

As most of the 13 studies included STI clinics, these samples were not representative for samples from the general population. However, the distribution of ompA genotypes among the 1,383 assumed heterosexual samples is similar to that in other studies (3) and indicates that our study reflects strain distribution in most countries where genotyping of C. trachomatis has been performed.

Previous ompA typing studies have shown that specific genotypes are frequently found in rectal samples from MSM (39–42), and tissue tropism has been suggested as an explanation (35, 43–46), mainly based on genetic and cell biological findings. However, recent MLST data show that differing sexual network structures and distributions of C. trachomatis strains in MSM versus heterosexual networks may be an important reason for the findings of certain genotypes in rectum (21). This also explains the findings in our data set of mostly different STs in MSM and heterosexuals.

Maximum likelihood analysis showed that obtained trees were similar to trees obtained from whole-genome analysis by Harris et al. (29), but bootstrap support values were lower. The phylogeny based on the MLST targets is also in agreement with the four identified lineages in whole-genome analysis by Joseph et al. (36). Analyses of MLST targets are easier to perform than obtaining whole genome sequences and thereafter performing analyses. Using MLST may therefore facilitate phylogenetic studies. We also applied phylogenetic analysis to the entire hr-CT-MLST data set and confirmed the results of the limited data set, but it also gave rise to more clades. A limitation is the occurrence of recombination that disturb the phylogenetic analysis. Therefore, identified recombinations were removed which influenced the tree topology, but the major structure remained intact.

The MLST target genes used in this study were originally selected because of their sequence variation, which was thought to provide high resolution and to be suitable for short-term investigations. Considering the conserved nature of the C. trachomatis genomes, we expected that this MLST system would be saturated with a limited number of STs. However, in each new study population, several new STs have been identified. The number of detected alleles provides a theoretical number of $8.8 \times 10^9$ STs, assuming no linkage between targets, indicating that only a fraction of possible STs have been identified. More refined genetic analysis is needed to elucidate why a few STs predominate in space and time. To provide a better understanding of evolutionary changes for C. trachomatis and link it to short-term variations, it would be rewarding to combine our MLST system with data from a housekeeping-based MLST. Furthermore, the role of the host response to different infecting strains is not well understood.

The strength of this study is the size of the data set. Although the geographic representation is patchy and the separate studies have different objectives, the results provide an overall picture of strain distribution for C. trachomatis in time and space and in populations with different sexual behavior that has not been observed previously.

In summary, the presented MLST system provides high resolution for C. trachomatis strains, is useful for epidemiological investigations, has identified a few predominating strains that have spread successfully in several countries, and is beneficial for phylogenetic analysis.

ACKNOWLEDGMENTS

The study was financed by local funds from Uppsala University Hospital, Sweden.

Additional members of The Chlamydia trachomatis MLST Study Group include Bertille de Barbeyrac and Olivia Peuchant (University of Bordeaux, Bordeaux, France), Reiner Bom (Municipal Health Service Amsterdam, Amsterdam, the Netherlands), Hans Fredlund (Örebro University Hospital, Örebro, Sweden), Houda Gharsallah and Adene Hamami (Habib Bourguiba University, Sfax, Tunisia), Charlotte A Gaydos (Johns Hopkins University, Baltimore, MD), David Mabey (London School of Hygiene and Tropical Medicine, London, United Kingdom), Servaas A Morré (VU University Medical Center, Amsterdam, the Netherlands), Marcelo Rodriguez Ferrempin (University of Buenos Aires, Buenos Aires, Argentina), Julius Schachter (University of California, San Francisco, CA), Henry J de Vries (Municipal Health Service Amsterdam, Amsterdam, the Netherlands), Qianqiu Wang (Institute of Dermatology,
REFERENCES


C. trachomatis MLST Analysis