Chlamydia trachomatis

Insights on genetics, molecular epidemiology, diagnostics and treatment, and host-pathogen interactions

Versteeg, A.J.

Publication date
2018
Document Version
Other version
License
Other

Link to publication

Chapter 4

Genomic analyses of the *Chlamydia trachomatis* core genome show an association between chromosomal genome, plasmid type and disease

Versteeg B, Bruisten SM, Pannekoek Y, Jolley KA, Maiden MCJ, van der Ende A, Harrison OB

Abstract

Background
Chlamydia trachomatis (Ct) plasmid has been shown to encode genes essential for infection. We evaluated the population structure of Ct using whole-genome sequence data (WGS). In particular, the relationship between the Ct genome, plasmid and infection pathology was investigated.

Results
WGS data from 157 Ct isolates deposited in the Chlamydiales pubMLST database (http://pubMLST.org/chlamydiales/) were annotated with 902 genes including the core and accessory genome. Plasmid associated genes were annotated and a plasmid MLST scheme was defined allowing plasmid sequence types to be determined. Plasmid allelic variation was investigated. Phylogenetic relationships were examined using the Genome Comparator tool available in pubMLST. Phylogenetic analyses identified four distinct Ct core genome clusters and six plasmid clusters, with a strong association between the chromosomal genotype and plasmid. This in turn was linked to *ompA* genovars and disease phenotype. Horizontal genetic transfer of plasmids was observed for three urogenital-associated isolates, which possessed plasmids more commonly found in isolates resulting from ocular infections. The *pgp3* gene was identified as the most polymorphic plasmid gene and *pgp4* was the most conserved.

Conclusion
A strong association between Ct chromosomal genotype and plasmid was observed, consistent with previous studies. This suggests co-evolution of the Ct chromosome and their plasmids, but we confirmed that plasmid transfer can occur between isolates. These data provide a better understanding of the genetic diversity occurring across the Ct genome in association with the plasmid content.

Background

Chlamydia trachomatis (Ct) is responsible for the majority of bacterial sexually transmitted infections worldwide [1]. In addition, ocular Ct infections (trachoma) are the world's leading cause of preventable blindness [2,3]. Although there are few documented reports of antibiotic resistance in Ct and infections can be easily treated, the not declining incidence of Ct infections globally, makes this infection an important public health priority.

Ct isolates can be grouped into 15 main genovars based on sequence data of *ompA,* the gene encoding the major outer membrane protein [4-6]. Specific genovars have been strongly associated with distinct disease pathologies: genovars A-C are associated with conjunctival epithelia; genovars D-K with urogenital, pharyngeal and anorectal epithelia; and genovars L1-L3 with submucosal connective tissue layers resulting in dissemination to locoregional lymph nodes and lymphogranuloma venereum (LGV) [7]. Ct clonal groups identified through different multilocus sequence typing (MLST) schemes based on 7 housekeeping genes as well as phylogenetic analyses of whole-genome sequence (WGS) data [8-11], have also shown an association between Ct strains and tissue tropism.

Ct isolates possess multiple copies of highly conserved small 7.5-kb plasmids containing non-coding RNA of which the function is unknown and 8 open reading frames (ORFs), designated *pgp1* to *pgp8* [12,13]. The plasmid of Ct has been shown to encode genes essential for infection and transmission, consistent with the rare occurrence of plasmid-deficient clinical isolates [14-16]. The essential role of the plasmid in virulence and inflammatory responses was further demonstrated using animal models, where plasmid-deficient Ct strains or those with mutated plasmids were found to exhibit reduced pathology and decreased inflammatory responses [17-23]. Putative functions have been assigned to the 8 plasmid genes, based on homology to known proteins [13,24-26], with *pgp1* exhibiting homology to a *DnaB* like helicase, *pgp5* to a partitioning protein which may regulate expression of a set of chromosomal genes, and *pgp7* and *-8* identified as integrase/recombinase homologues. In contrast, *pgp*2, and *pgp6* are unique to the chlamydia genus since homology to other proteins has not been identified to-date. The product of *pgp*3 is secreted into the host cell cytosol and is the most studied plasmid gene, both as a serological marker for past infection and as virulence factor, as

4

it was demonstrated to play an important role in the induction of inflammatory responses [21,27,28]. Finally, *pgp4* is a transcriptional regulator of *pgp3* and of some chromosomal genes that are likely to be important for chlamydial virulence [25]. The *pgp7* gene is not essential for plasmid maintenance as it was found to be interrupted in naturally occurring Ct strains resulting in the emergence and rapid spread of a new Ct variant [29]. This new variant originated in Sweden in 2006 and had a 377 bp deletion in the *pgp7* gene that prevented detection of infections using plasmid based PCR diagnostics targeting the *pgp7* gene [13,29]. The potential of spread and emergence of new strains due to genetic variation indicates the need for more comprehensive studies to better understand the Ct genetic population structure.

Although it is clear that the plasmid plays an important role in the pathogenesis of Ct infection, limited data is available on its genetic diversity and whether distinct plasmid types are associated with different tissue tropisms and pathologies. This study set out to characterise the population structure of Ct in association with the plasmid using WGS data from 157 isolates available in the *Chlamydiales* pubMLST database (http://pubMLST.org/chlamydiales/). A better understanding of genetic diversity across the Ct genome in association with plasmid content may elucidate Ct epidemiology, ultimately reducing the burden of infection.

Methods

Ct isolate collections and WGS methods

Whole-genome sequence (WGS) data from 157 Ct isolates were analysed. WGS data were obtained from published isolate collections for which plasmid sequence data were also available [9,26,30-32]. Short reads were obtained from the European Nucleotide Archive (ENA) and assembled *de novo* using VELVET in combination with VELVETOPTIMISER as previously described [33]. The resulting contigs were uploaded to the Bacterial Isolate Genome Sequence (BIGSdb) genomics platform hosted on www.pubMLST.org/chlamydiales.

The Chlamydiales pubMLST platform consists of two types of database: i) a sequence definition database that contains sequences of known alleles for loci as well as allelic profiles for specific schemes such as MLST; and, ii) an isolate

database that contains isolate provenance and other metadata along with nucleotide sequences associated with that isolate [34]. Sequence definitions have been established for 902 protein-encoding genes, annotated with the CHLAM prefix, and the majority of these have been organised into schemes dependent on function. Chromosomal genes were defined using the annotated genome from Ct strain D/UW-3/CX (accession number NC_000117, [35]). The BIGSdb software includes 'autotagger' and "autodefiner" tools which scan deposited WGS against defined loci identifying alleles ≥98% sequence identity. This process runs in the background and automatically updates isolate records with specific allele numbers, marking regions on assembled contigs for any of the defined loci. Loci with sequence identity <98% are manually checked and curated.

OmpA, which encodes the major outer membrane protein, was annotated in WGS data as CHLAM0681. Nucleotide sequence data from CHLAM0681 was extracted from all WGS. Bionumerics software (version 7.5, Applied Maths, Sint-Martens-Latem, Belgium) was used to import all extracted *ompA* sequences to a local offline reference database of *ompA* sequences that had been identified in previous Ct studies [10]. Based on sequence similarity of the *ompA* variable domains 1 and 2, *ompA* genovars were assigned to all isolates. Genovars A-C were considered to be ocular isolates, genovars D-K, urogenital isolates and genovar L, LGV isolates. Genovar B is known to cause both urogenital and ocular infections.

Phylogenetic analyses
Relationships among isolates were established using the Genome Comparator tool implemented within the *Chlamydiales* pubMLST database [34]. Genome Comparator compares groups of shared genes among isolates with any number of loci predefined in the *Chlamydiales* database or a reference genome. For each locus, allele sequences, designated by integers, are compared and used to generate a distance matrix that is based on the number of variable loci across the genome generating a wgMLST profile. Genome Comparator output provides lists of loci that are: i) identical, ii) variable, iii) missing, or incomplete between data sets, rapidly resolving bacterial population structures and relationships, and identifying loci that belong to the core of a particular data set [33].

Using the Genome Comparator tool, all chromosomal genes identified in the automated annotation process were compared. The set of 888 genes shared

between all Ct isolates was referred to as the 'core genome'. Genome Comparator was used to compare the core genome among all isolates and to generate a distance matrix based on the number of variable loci. In addition, this was used to compare previously identified plasmid genes among all isolates. The generated distance matrix for the core genome and plasmid genes were further analysed using the Neighbor Joining algorithm in SplitsTree version 4.14, to investigate the phylogenetic clustering of Ct isolates according to both their core genome and plasmid [36]. In addition, to each isolate a unique ID, the corresponding *ompA* genovar and the plasmid sequence type (pST) were linked to each isolate.

Ct plasmid

Sequences from the plasmid belonging to Ct strain D/SotonD6 were retrieved from plasmid pSotonD6 (HE603231) and designated as CHLAM0895 through to CHLAM0902 encoding the genes *pgp1* to *pgp8*. Using Blast, all WGS sequence data deposited in pubMLST were annotated for these loci as described previously [34,37,38].

The eight Ct plasmid genes were grouped into a plasmid MLST typing scheme and pSTs were assigned based on identified allele variants for isolates with sequence data on all eight plasmid genes. The number of polymorphic sites per plasmid gene, was assessed using the locus explorer tool in the database (http://pubMLST. org/chlamydiales/). Molecular Evolutionary Genetics Analysis software, version 6 (MEGA 6; http://www.megasoftware.net) was used to calculate average pairwise diversity between isolates [39]. For each gene the *p*-distance was estimated, both on the nucleotide and amino acid level, with pairwise deletion option selected and standard error (SE) determined with 1,000 bootstrap replications. The average numbers of synonymous substitutions per synonymous site (dS) and non-synonymous substitutions per non-synonymous site (dN) were calculated by using the overall mean Kumar model [40,41]. For dN/dS>1 the Z-test of positive selection was applied and values of $P<0.05$ were considered significant.

Results

Ct core genome analyses

Ct WGS data available in the *Chlamydiales* pubMLST database, were filtered to identify those which included complete plasmid sequence data, resulting in 157 isolates. Isolates dated from 1959-2011 and were from diverse geographical locations.

A total of 31 allelic *ompA* variants corresponded to 13 genovars: A, 9.6% (n=15 isolates); B, 3.2% (n=5); C, 1.9% (n=3); D, 8.9% (n=14); E, 26.8% (n=42); F, 8.9% (n=14); G, 8.3% (n=13); H, 2.5% (n=4); I, 4.5% (n=7); J, 1.9% (n=3); K, 9.6% (n=15); L1, 3.8% (n=6), and L2b, 10.2% (n=16).

A total of 888 out of 902 loci (98.4%) were found to be shared among the 157 included isolates and represented the Ct core genome (cgMLST). Based on the diversity, a distance matrix was calculated from which a Neighbor Joining tree was generated (Figure 1). Four phylogenetically distinct clusters were observed, consistent with previous studies [9]. These included: Cluster I, comprising the ocular genovars A, B and C; Cluster II, the clinically more prevalent urogenital genovars D, E and F; Cluster III, the LGV genovars L1 and L2b and, Cluster IV, the rarer urogenital genovars B, D, G, H, I, J and K. WGS from two trachoma isolates (708 and 840, both genovar C), however, clustered with urogenital Ct strains in Cluster II (Figure 1) indicative of horizontal transfer.

The number of alleles present for each core gene varied from 1 to 44. CHAM0061 (*fliA*) encoding a Sigma-28/WhiG family protein had the lowest number of allelic variants (n=1), while the highest number of alleles (n=44) was observed for CHLAM0147 encoding a hypothetical protein associated with the type III secretion system (T3SS) [42]. Overall, the most diverse genes were *ompA* (CHLAM0681), the polymorphic outer membrane proteins (CHLAM0412, CHLAM0413, CHLAM0414, CHLAM0812, CHLAM0869, CHLAM0870, CHLAM0871, CHLAM0872, and CHLAM0874), 11 genes associated with the T3SS and 4 genes located in the plasticity zone (CHLAM0153, CHLAM0154, CHLAM0157 and CHLAM0166) in accordance to what was described previously [42,43]. The more conserved genes were found to be hypothetical proteins, RNA associated genes, genes involved in DNA replication and nucleotide excision repair.
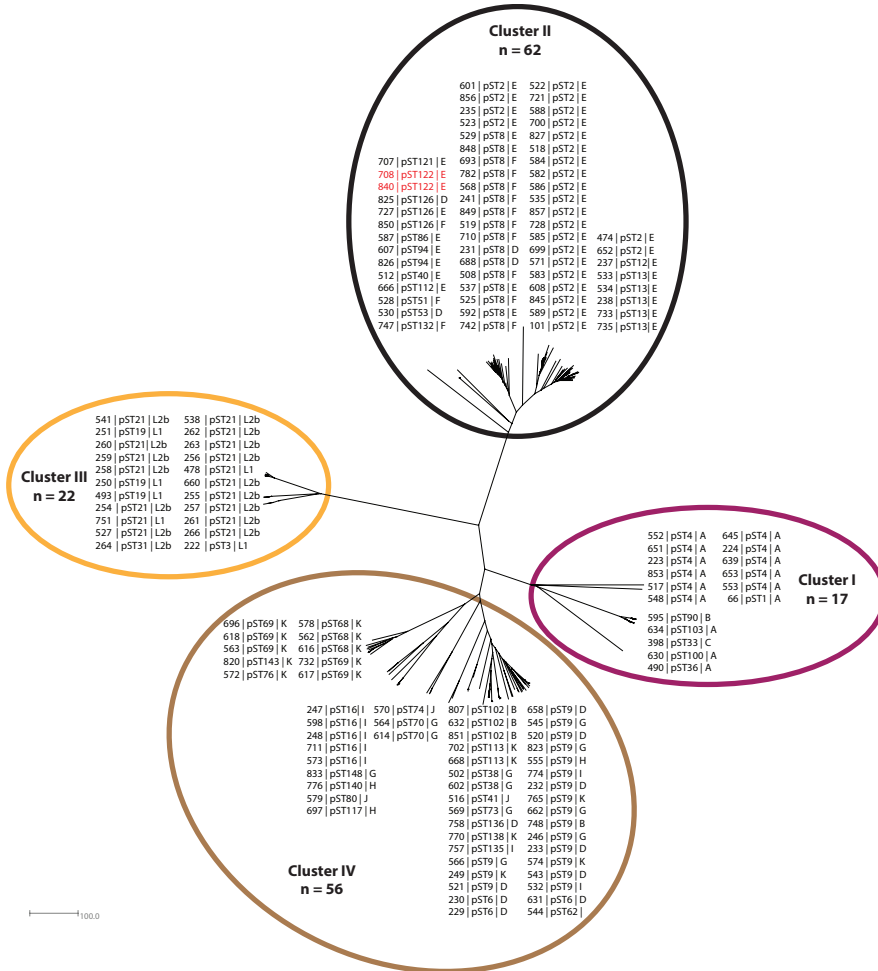
Figure 1: Neighbor Joining tree showing the core genome clustering of 157 *Chlamydia trachomatis* isolates. Coloured halos indicate the distinct clusters, isolates are indicated by the Chlamydiales pubMLST ID numbers, plasmid sequence type (pST) and ompA genovar. Isolates shown in red (in cluster II) indicate potential recombinants.

The remaining 14 genes not included in the core genome were considered to be accessory genes. Of these, 8 comprised the plasmid genes *pgp1-pgp8* that were considered accessory to assess the association between the core genome and plasmid types. The remaining 6: CHLAM0165, CHLAM0166, CHLAM0167, CHLAM0173, and CHLAM0174 encoded hypothetical proteins, while CHLAM0456 encoded the translocated actin recruiting phosphoprotein (*tarp*), a T3SS effector in Chlamydia [43]. *Tarp* has previously been reported as a highly variable gene correlating with ocular, urogenital and LGV disease phenotypes [42-44]. Among

isolates included in this study, *tarp* alleles were present in 93/157 isolates (59.2%) with a total of 36 unique allelic variants (Supplementary Table 1). Analysis of allelic variation of *tarp* revealed that 24/36 allelic variants were associated with disease phenotype (ocular, urogenital or LGV). Of the remaining 8 alleles 25, 33 and 36 were found in WGS data from ocular, urogenital and LGV isolates, alleles 20, 22 and 44 were shared between both ocular and urogenital isolates, allele 1 was found in urogenital and LGV isolates while allele 13 was associated with an ocular and LGV isolate. Overall, the majority of *tarp* alleles were specific to isolates belonging to a certain disease phenotype, although some *tarp* alleles were also exchanged between isolates belonging to different disease phenotypes.

Analysis of *p*-distance values across the core genome revealed that the DNA-binding protein CHLAM0046 (*hctB*, *p*-distance=0.451), ribulose-phosphate 3-epimerase CHLAM0121 (*araD*, *p*-distance=0.500), rRNA methylase CHLAM0133 (*p*-distance=0.063), the plasticity zone genes: CHLAM0155 (*p*-distance=0.011), CHLAM0157 (*p*-distance=0.099), CHLAM0171 (*p*-distance=0.064) and hypothetical protein CHLAM0326 (*p*-distance=0.039) were the most diverse core genome loci (Figure 2).

Ct plasmid analyses

The gene *pgp4* (CHLAM0900) was the most conserved plasmid gene with 3 alleles containing only 2 polymorphic sites (Table 1). In contrast, *pgp6* (CHLAM902) was the most diverse with 17 allelic variants, followed by: *pgp2* with 14, *pgp3* and *pgp1* with 12, *pgp5* and *pgp7* with 13 and *pgp8* with 11 unique allelic variants. In addition, all plasmid genes had between 15 and 23 polymorphic sites except for *pgp4*. Further analysis revealed that *pgp3* (CHLAM0899), which has also been described as a virulence-associated gene [21,27,28], was the most polymorphic (*p*-distance=0.008; SE±0.002 and amino acid level 0.016; SE±0.004). To estimate whether plasmid genes were subject to selection, the ratio between the rate of non-synonymous (dN) and synonymous (dS) substitutions per (non-)synonymous nucleotide site was determined. This ranged from 0.083 (*pgp8*) to 2.00 (*pgp5*) indicating that plasmid genes were not subject to strong positive selection. Gene *pgp5* which is suggested to regulate the expression of some chromosomal genes, showed a dN/dS=2, but this value was not statistically significant on the Z-test of positive selection.
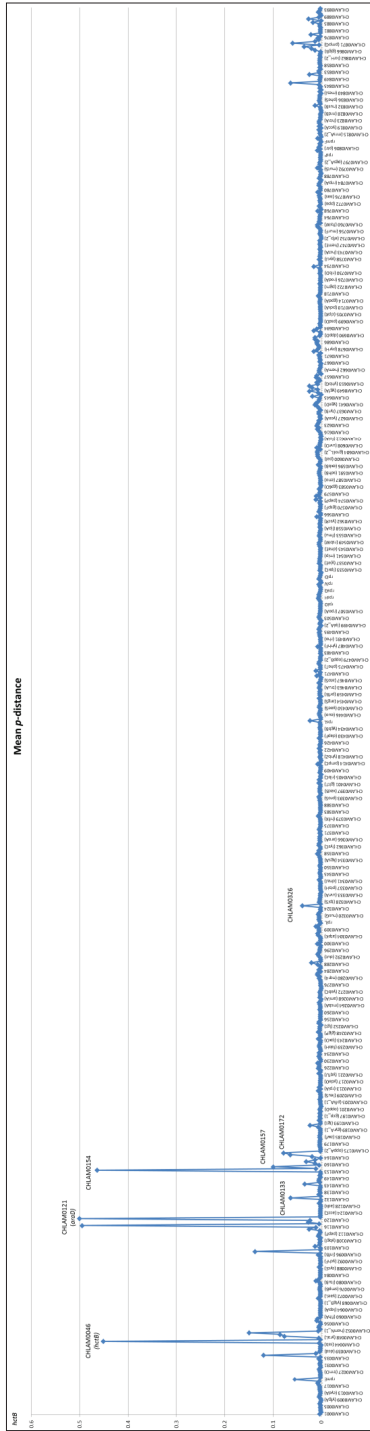
Figure 2. Genomic diversity for 851 core loci of 157 *Chlamydia trachomatis* isolates. The most variable loci are labelled with their gene name

Table 1: Polymorphism analysis and allelic variance of the plasmid genes from 157 *Chlamydia trachomatis* isolates

Gene (ORF)[a]	Length (bp)[b]	No. alleles	No. polymorphic sites	No. of differences	[SE]	Overall Mean Distance (pairwise deletion; bootstrap = 1000 replicates)									
						Nucleotide				Amino Acid			P-distance; Non- and Synonymous substitutions		
						No. of differences	[SE]	P-distance	[SE]	P-distance	[SE]	Mean dS	[SE]	Mean dN	[SE]	dN/dS[c]
pgp1 (ORF3)	1356	12	18	4.965	1.402	1.602	0.682	0.004	0.001	0.004	0.002	0.007	0.002	0.002	0.001	0.286
pgp2 (ORF4)	1065	14	20	4.420	1.125	1.043	0.553	0.004	0.001	0.003	0.002	0.009	0.003	0.001	0.001	0.111
pgp3 (ORF5)	795	12	23	6.133	1.488	4.172	1.122	0.008	0.002	0.016	0.004	0.007	0.003	0.007	0.002	1.000
pgp4 (ORF6)	309	3	2	0.363	0.262	0.243	0.239	0.001	0.001	0.002	0.002	0.001	0.001	0.001	0.001	1.000
pgp5 (ORF7)	795	13	15	2.720	0.934	2.111	0.872	0.003	0.001	0.008	0.003	0.002	0.001	0.004	0.002	2.000
pgp6 (ORF8)	744	17	20	3.003	1.020	1.415	0.648	0.004	0.001	0.006	0.003	0.006	0.003	0.003	0.001	0.500
pgp7 (ORF1)	918	13	23	4.523	1.227	1.715	0.739	0.005	0.001	0.006	0.002	0.010	0.004	0.002	0.001	0.200
pgp8 (ORF2)	993	11	18	4.346	1.223	0.476	0.428	0.004	0.001	0.001	0.001	0.012	0.004	0.001	0.001	0.083

[a] ORFs designation according to Thomas et al, 1997. Microbiology, 143, 1847-1854.

[b] Relative to the sequence of the plasmid of the E/Bour strain (GenBank No. HE603212).

[c] The value of dN/dS > 1 obtained for pgp5 was not statistically significant by the Z-test of positive selection.

A total of 47 unique pSTs were randomly assigned and (Figure 1). Sequence comparison of plasmid genes identified six phylogenetic clusters numbered 1 to 6 (Figure 3). They were associated with core genome clusters; all isolates of plasmid cluster 2, -4, and -6 fell into core genome clusters II, III and IV, respectively (Figure 4). All isolates except three, of plasmid cluster 1 grouped with those from core genome cluster I. The three exceptions grouped with isolates of core genome cluster IV, indicating horizontal plasmid transfer. In addition, plasmid clusters 3 and 6 formed sub-branches in core genome cluster IV. Overall, these results showed a strong association between Ct core genome and plasmid, but in a minority of isolates exchange of plasmids was observed. Overall, each core genome cluster had one dominating pST (Figure 1 and 4).

4

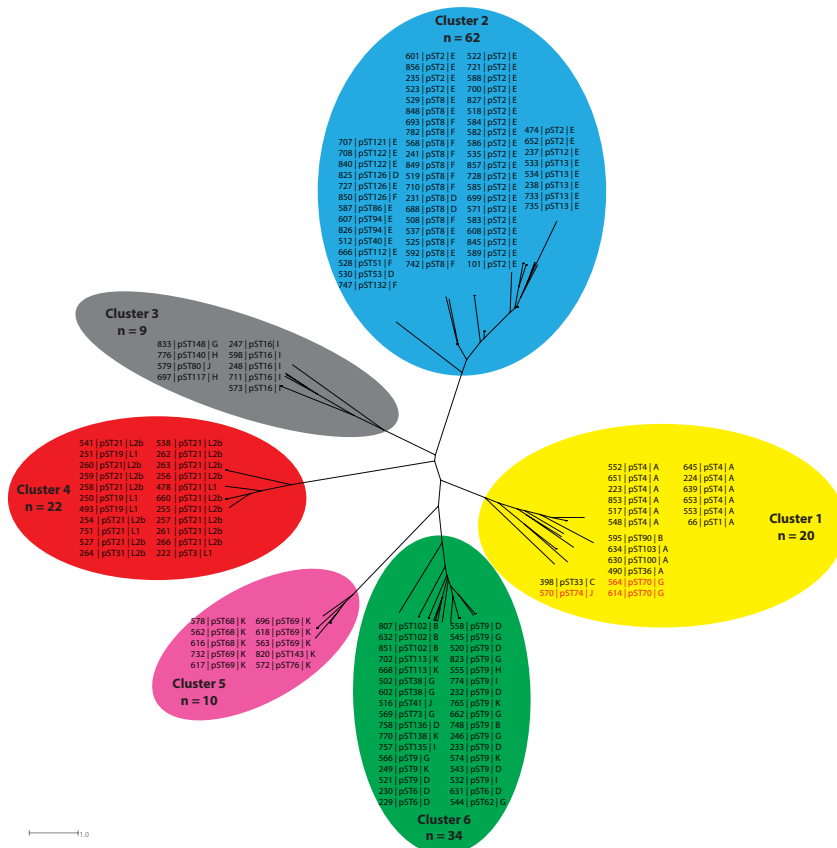**Figure 3:** Neighbor Joining tree showing clustering of the plasmid sequence types (pST) generated from plasmid genes pgp1 to pgp8 of 157 *Chlamydia trachomatis* isolates. Coloured halos indicate the distinct clusters, isolates are indicated by the Chlamydiales pubMLST ID numbers, pST and ompA genovar. Isolates shown in red (cluster 1) indicate potential plasmid exchange compared to the core genome clusters.
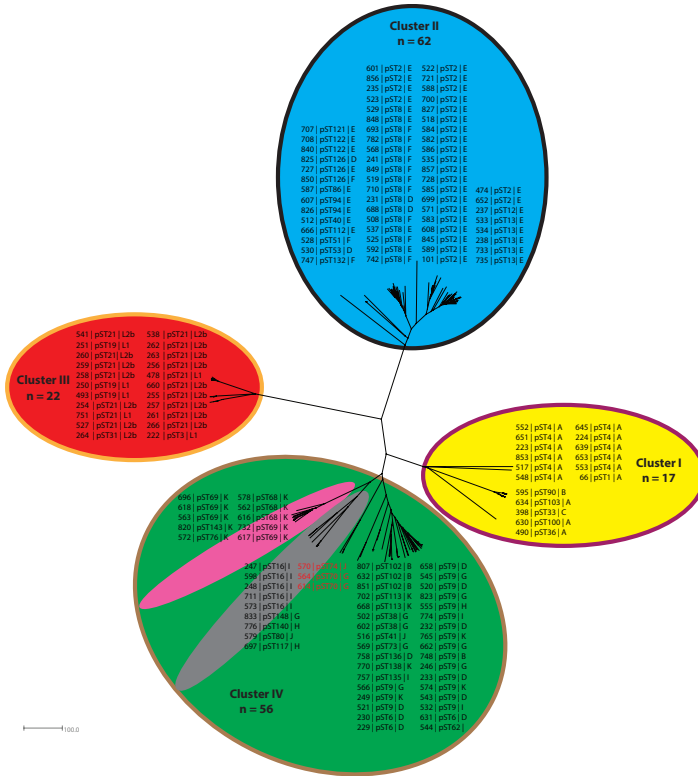
Figure 4: Neighbor Joining tree showing the comparison between core genome and plasmid clustering of 157 *Chlamydia trachomatis* isolates. Coloured halos indicate the distinct clusters from Figure 1, isolates are indicated by the Chlamydiales pubMLST ID numbers, plasmid sequence type (pST) and ompA genovar. Isolates shown in red indicate potential plasmid exchanges compared to the plasmid clusters in Figure 2.

Analysis of the allelic variation in each single plasmid gene (*pgp1-8*) with respect to the observed plasmid clusters revealed that exchange of alleles between plasmid clusters was limited. Plasmid clusters possessed specific *pgp1, -2, -3*, and *-5* alleles that were unique to each cluster (Figure 5, Supplementary Table 2). For example, plasmids in cluster 1 contained *pgp1* alleles 3 and 17 which were not found in any of the other plasmid clusters, while cluster 3 had *pgp1* alleles 9 and 10. The allelic variation of *pgp6, -7* and *-8* was specific for the clinically most prevalent urogenital cluster (cluster 2), one cluster containing some of the less frequently occurring genital genovars (cluster 3) and the LGV cluster (cluster 4), but exchange of allelic variants was observed among clusters 1, 5 and 6. Finally, the allelic variation of the *pgp4* gene was only specific for the LGV cluster (cluster 4) and a urogenital cluster containing the less frequently occurring genovar K isolates (cluster 5), but exchange of *pgp4* allelic variants was observed among clusters 1, 2, 3 and 6.
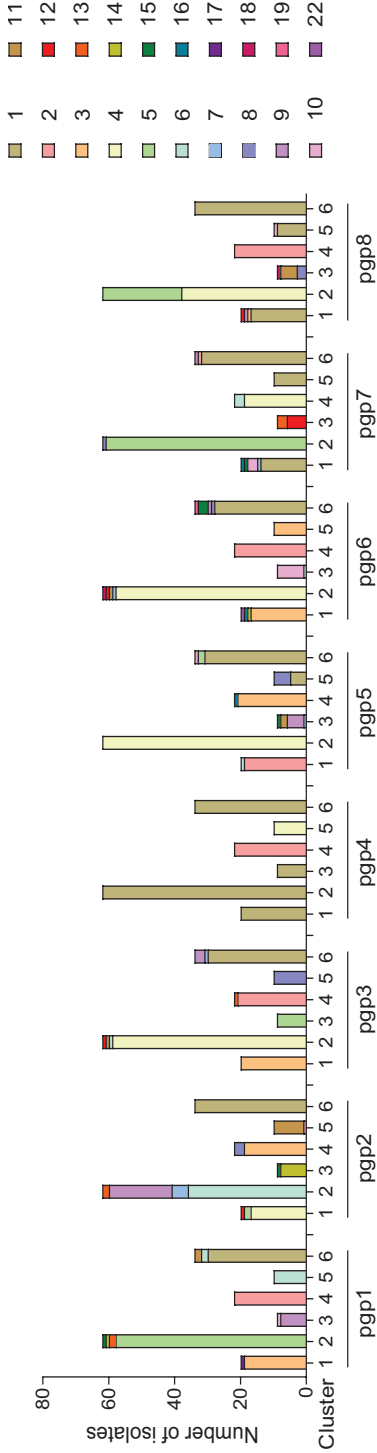
Figure 5. Graph showing the allelic variants for each single plasmid gene (pgp1-8) with respect to the observed plasmid clusters (cluster 1 to 6) among 157 *Chlamydia trachomatis* isolates. Each allelic variant received a unique colour as shown in the colour legend. Allelic numbers were randomly assigned to each unique allelic variant of a single plasmid gene (pgp1-8). Specific allelic numbers can be assigned to multiple plasmid genes.

4

Discussion

Advances in sequencing technologies and increasing availability of WGS data provide unique opportunities for improving our understanding of Ct infection and epidemiology. Pivotal to this is the ability to rapidly extract strain information from WGS data including *ompA* genovar, plasmid type and MLST sequence type such that global surveillance of Ct infections can be achieved. In this study, a catalogue of genes both core and accessory to the Ct genome was generated on the web-accessible http://pubMLST.org/chlamydiales/ website providing tools for Ct surveillance in an open database. Data presented here reveal that Ct core genomes were strongly associated with distinct Ct plasmid types (Figures 1, 3 and 5). Four core genome clusters were apparent following cgMLST analyses consistent with a previous study [9], however, two recombinant isolates were also apparent. These were two trachoma isolates with urogenital backbones that had been identified by Andersson et al. [30].

A total of 888 genes were found to be core, with only 14 accessory genes consistent with the effects of genome reduction known to have occurred in Ct [45]. All of the accessory genes encoded hypothetical proteins with the exception of the plasmid genes and CHLAM0456, which encodes *tarp*. None of the hypothetical genes correlated with disease phenotype. However, subsequent analysis of *tarp* allelic variants revealed that the majority of alleles (59.2%) were associated with distinct disease phenotypes (ocular, urogenital, or LGV), although some alleles were also shared between isolates from different disease phenotypes. This suggests that although *tarp* contributes to tissue tropism, it is not the sole gene driving these phenotypes. The most variable gene was CHLAM0147 encoding a T3SS effector involved in endosomal trafficking by recruiting nutrient-rich endocytic vesicles via a non-fusogenic pathway to the chlamydial inclusion [42]. Ct T3SS are activated when Ct attaches to a host cell, after which T3SS are used to deliver an arsenal of bacterial gene-encoded effector proteins into the cytosol of the host cell [42,44,46]. The exact molecular mechanisms of T3SS remains to be elucidated, however, it is probable that these genetically diverse T3SS genes function together to favour specific tissue tropisms [42,43].

The most conserved gene was *fliA* encoding sigma-28, for which the exact function remains to be elucidated. It has been suggested that expression of

sigma-28 is in response to cellular stress, such as nutrient deprivation within the chlamydial inclusion [47,48]. Overall, the most variable core genes were *ompA*, the polymorphic outer membrane proteins, genes associated with T3SS and the plasticity zone, which all have been suggested to contribute to tissue tropism and disease severity due to their high polymorphic variation [42,43]. The more conserved genes encoded hypothetical proteins for which no known functions have been described, but have been suggested to play an important role in the complex Ct-host interactions [42,43]. The highly-conserved nature of these genes suggests that mutations may have deleterious effects on biological fitness. Polymorphic genes included *hctB* (CHLAM0046), *araD* (CHLAM121), CHLAM133, CHLAM0155, CHLAM0157, CHLAM172 and CHLAM326. CHLAM0046 (*hctB)* encodes a DNA-binding protein thought to mediate the chromatin compaction [49]. This gene is described to vary among Ct genovars due to internal deletions from a region of the *hctB* gene encoding lysine- and alanine-rich pentameric repeats [50]. The genes CHLAM0155, CHLAM0157 and CHLAM0172 are all part of a 20.3-kb highly polymorphic genomic region encoding toxin-like genes known as the plasticity zone [43]. The chlamydia plasticity zone has also been known to vary among genovars in accordance with known phylogenetic tissue tropism (urogenital, ocular and LGV) [42,43]. The CHLAM0326 gene encodes a hypothetical protein without a known function, but this gene has been associated with rectal tropism of Ct genovar G isolates [51]. Finally, the genes *araD* (CHLAM0121) and CHLAM0133 are known to encode a ribulose-phosphate 3-epimerase and rRNA methylase, but variation has not been linked to specific Ct strains [52].

A total of 47 unique pSTs were identified (Figure 2, Table 1) which aggregated into six distinct plasmid clusters, five of which were comparable to those of Harris et al. [9]. The additional cluster consisted solely of genovar K isolates and was likely the result of Ct strains that had previously not been sampled. In comparison to the clusters observed following cgMLST analyses, plasmid analyses showed that the rarer *ompA* genovars (genovars B, D, G, H, I, J, and K) formed three distinct plasmid clusters. Moreover, three urogenital isolates with genovars G and J were identified that contained plasmids clustering with ocular isolates (genovars A-C), suggesting that exchange of plasmids may have occurred between urogenital and trachoma isolates. Overall, our findings are in agreement with previous studies that suggested co-evolution of Ct plasmids and their chromosome and demonstrated that, although rare, exchange of a plasmid can occur [9,12]. In addition, allelic

variation in some of the plasmid genes was specific and distinct for each cluster, while variation in other genes was mainly specific for the LGV cluster since the allelic variants of the remaining clusters were shared.

The observed polymorphic variation was in agreement with previous analyses that highlighted the sequence conservation of Ct [9,13,53]. Although the plasmid was highly conserved, much of the diversity appeared to be restricted to one plasmid gene known to be associated with chlamydial virulence. This gene, *pgp3*, which is associated with increased Ct inflammatory responses [21,27,28], was the most polymorphic gene (*p*-distance=0.008) and is known to be secreted in the host cell cytosol and its diversity is possibly a result of immune selection [12,21,27,28]. In contrast, *pgp4* which functions as a transcriptional regulator of *pgp3* and some chromosomally encoded genes was identified as the most conserved gene. The fact that this small gene (309 bp) was highly conserved suggests that *pgp4* is an indispensable factor for chlamydial virulence and infection, although previous studies have demonstrated that *pgp4* is dispensable for growth in vitro [25,54]. Finally, we observed that exchange of alleles between plasmid clusters was limited. Some alleles were specific to each plasmid cluster, with, in particular, distinct plasmid alleles among LGV isolates As Ct harbours multiple plasmid copies, this knowledge may be useful to design plasmid assays which can distinguish clinically relevant Ct strain types, for instance to detect the Swedish truncated plasmid variant or to differentiate between LGV/non-LGV isolates. The latter is of clinical importance since LGV infections are more invasive and need extended treatment [9,12,55].

A limitation of our study is that it was conducted using previously sequenced and stored isolates, which unfortunately contained limited epidemiological or geographical data. Since these isolates were originally selected for different objectives, the data available could not be extrapolated to all included isolates. Previous molecular epidemiological studies on Ct using MLST showed no association between Ct strain types and symptoms, anatomical locations, gender or geographical location. Most strain types were also globally distributed with similar *ompA* genovar distributions [10,56-58]. These studies therefore suggest that the population structure of Ct is comparable among different human populations and that the effect of selection bias would be minimal. However, MLST is limited by the fact that only a small (but polymorphic) fraction of the genome is used for

typing, so samples that are indistinguishable with respect to MLST target regions may still have considerable variation in the rest of the genome. Future studies should therefore include a well-defined population with known epidemiological and clinical data to gain a better understanding of the Ct epidemiology in relation to the population structure.

In conclusion, a strong association between Ct core genome and plasmid types was observed, consistent with co-evolution of Ct plasmids and their chromosome. Moreover, we confirmed that, although rare, plasmid exchange can occur between isolates. Finally, we showed that exchange of alleles between plasmid clusters was limited. Future research should apply the gene-by-gene approach to a well-defined population with known epidemiological and clinical data, as this will enhance our understanding of chlamydial transmission and disease.

4

References

1. Newman L, Rowley J, Vander HS, Wijesooriya NS, Unemo M, Low N, Stevens G, Gottlieb S, Kiarie J, Temmerman M: Global Estimates of the Prevalence and Incidence of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review and Global Reporting. *PLoS One* 2015, 10:e0143304.

2. Mariotti SP, Pascolini D, Rose-Nussbaumer J: Trachoma: global magnitude of a preventable cause of blindness. *Br J Ophthalmol* 2009, 93:563-568.

3. Taylor HR, Burton MJ, Haddad D, West S, Wright H: Trachoma. *Lancet* 2014, 384:2142-2152.

4. Barnes RC, Wang SP, Kuo CC, Stamm WE: Rapid immunotyping of Chlamydia trachomatis with monoclonal antibodies in a solid-phase enzyme immunoassay. *J Clin Microbiol* 1985, 22:609-613.

5. Batteiger BE, Newhall WJ, Terho P, Wilde CE, III, Jones RB: Antigenic analysis of the major outer membrane protein of Chlamydia trachomatis with murine monoclonal antibodies. *Infect Immun* 1986, 53:530-533.

6. Caldwell HD, Schachter J: Antigenic analysis of the major outer membrane protein of Chlamydia spp. *Infect Immun* 1982, 35:1024-1031.

7. de Vries HJ, Smelov V, Middelburg JG, Pleijster J, Speksnijder AG, Morre SA: Delayed microbial cure of lymphogranuloma venereum proctitis with doxycycline treatment. *Clin Infect Dis* 2009, 48:e53-e56.

8. Dean D, Bruno WJ, Wan R, Gomes JP, Devignot S, Mehari T, de Vries HJ, Morre SA, Myers G, Read TD et al.: Predicting phenotype and emerging strains among Chlamydia trachomatis infections. *Emerg Infect Dis* 2009, 15:1385-1394.

9. Harris SR, Clarke IN, Seth-Smith HM, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW et al.: Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. *Nat Genet* 2012, 44:413-9, S1.

10. Herrmann B, Isaksson J, Ryberg M, Tangrot J, Saleh I, Versteeg B, Gravningen K, Bruisten S: Global Multilocus Sequence Type Analysis of Chlamydia trachomatis Strains from 16 Countries. *J Clin Microbiol* 2015, 53:2172-2179.

11. Pannekoek Y, Morelli G, Kusecek B, Morre SA, Ossewaarde JM, Langerak AA, van der Ende A: Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria Chlamydia trachomatis. *BMC Microbiol* 2008, 8:42.

12. Seth-Smith HM, Harris SR, Persson K, Marsh P, Barron A, Bignell A, Bjartling C, Clark L, Cutcliffe LT, Lambden PR et al.: Co-evolution of genomes and plasmids within Chlamydia trachomatis and the emergence in Sweden of a new variant strain. *BMC Genomics* 2009, 10:239.

13. Thomas NS, Lusher M, Storey CC, Clarke IN: Plasmid diversity in Chlamydia. *Microbiology* 1997, 143 (Pt 6):1847-1854.

14. Stothard DR, Williams JA, Van Der Pol B, Jones RB: Identification of a Chlamydia trachomatis serovar E urogenital isolate which lacks the cryptic plasmid. *Infect Immun* 1998, 66:6010-6013.

15. Farencena A, Comanducci M, Donati M, Ratti G, Cevenini R: Characterization of a new isolate of Chlamydia trachomatis which lacks the common plasmid and has properties of biovar trachoma. *Infect Immun* 1997, 65:2965-2969.

16. Peterson EM, Markoff BA, Schachter J, de la Maza LM: The 7.5-kb plasmid present in Chlamydia trachomatis is not essential for the growth of this microorganism. *Plasmid* 1990, 23:144-148.

17. Carlson JH, Whitmire WM, Crane DD, Wicke L, Virtaneva K, Sturdevant DE, Kupko JJ, III, Porcella SF, Martinez-Orengo N, Heinzen RA et al.: The Chlamydia trachomatis plasmid is a transcriptional regulator of chromosomal genes and a virulence factor. *Infect Immun* 2008, 76:2273-2283.

18. Kari L, Whitmire WM, Olivares-Zavaleta N, Goheen MM, Taylor LD, Carlson JH, Sturdevant GL, Lu C, Bakios LE, Randall LB et al.: A live-attenuated chlamydial vaccine protects against trachoma in nonhuman primates. *J Exp Med* 2011, 208:2217-2223.

19. O'Connell CM, Ingalls RR, Andrews CW, Jr., Scurlock AM, Darville T: Plasmid-deficient Chlamydia muridarum fail to induce immune pathology and protect against oviduct disease. *J Immunol* 2007, 179:4027-4034.

20. Olivares-Zavaleta N, Whitmire W, Gardner D, Caldwell HD: Immunization with the attenuated plasmidless Chlamydia trachomatis L2(25667R) strain provides partial protection in a murine model of female genitourinary tract infection. *Vaccine* 2010, 28:1454-1462.

21. Ramsey KH, Schripsema JH, Smith BJ, Wang Y, Jham BC, O'Hagan KP, Thomson NR, Murthy AK, Skilton RJ, Chu P et al.: Plasmid CDS5 influences infectivity and virulence in a mouse model of Chlamydia trachomatis urogenital infection. *Infect Immun* 2014, 82:3341-3349.

22. Russell M, Darville T, Chandra-Kuntal K, Smith B, Andrews CW, Jr., O'Connell CM: Infectivity acts as in vivo selection for maintenance of the chlamydial cryptic plasmid. *Infect Immun* 2011, 79:98-107.

23. Sigar IM, Schripsema JH, Wang Y, Clarke IN, Cutcliffe LT, Seth-Smith HM, Thomson NR, Bjartling C, Unemo M, Persson K et al.: Plasmid deficiency in urogenital isolates of Chlamydia trachomatis reduces infectivity and virulence in a mouse model. *Pathog Dis* 2014, 70:61-69.

24. Hatt C, Ward ME, Clarke IN: Analysis of the entire nucleotide sequence of the cryptic plasmid of Chlamydia trachomatis serovar L1. Evidence for involvement in DNA replication. *Nucleic Acids Res* 1988, 16:4053-4067.

25. Song L, Carlson JH, Whitmire WM, Kari L, Virtaneva K, Sturdevant DE, Watkins H, Zhou B, Sturdevant GL, Porcella SF et al.: Chlamydia trachomatis plasmid-encoded Pgp4 is a transcriptional regulator of virulence-associated genes. *Infect Immun* 2013, 81:636-644.

26. Thomson NR, Holden MT, Carder C, Lennard N, Lockey SJ, Marsh P, Skipp P, O'Connor CD, Goodhead I, Norbertzcak H et al.: Chlamydia trachomatis: genome sequence analysis of lymphogranuloma venereum isolates. *Genome Res* 2008, 18:161-171.

27. Chen D, Lei L, Lu C, Galaleldeen A, Hart PJ, Zhong G: Characterization of Pgp3, a Chlamydia trachomatis plasmid-encoded immunodominant antigen. *J Bacteriol* 2010, 192:6017-6024.

28. Li Z, Chen D, Zhong Y, Wang S, Zhong G: The chlamydial plasmid-encoded protein pgp3 is secreted into the cytosol of Chlamydia-infected cells. *Infect Immun* 2008, 76:3415-3428.

29. Unemo M, Seth-Smith HMB, Cutcliffe LT, Skilton RJ, Barlow D, Goulding D, Persson K, Harris SR, Kelly A, Bjartling C et al.: The Swedish new variant of Chlamydia trachomatis: genome sequence, morphology, cell tropism and phenotypic characterization. *Microbiology* 2010, 156:1394-1404.

30. Andersson P, Harris SR, Smith HMBS, Hadfield J, O'Neill C, Cutcliffe LT, Douglas FP, Asche LV, Mathews JD, Hutton SI et al.: Chlamydia trachomatis from Australian Aboriginal people with trachoma are polyphyletic composed of multiple distinctive lineages. *Nat Commun* 2016, 7.

31. Hadfield J, Harris SR, Seth-Smith HMB, Parmar S, Andersson P, Giffard PM, Schachter J, Moncada J, Ellison L, Vaulet MLG et al.: Comprehensive global genome dynamics of Chlamydia trachomatis show ancient diversification followed by contemporary mixing and recent lineage expansion. *Genome Res* 2017.

32. Seth-Smith HM, Harris SR, Skilton RJ, Radebe FM, Golparian D, Shipitsyna E, Duy PT, Scott P, Cutcliffe LT, O'Neill C et al.: Whole-genome sequences of Chlamydia trachomatis directly from clinical samples without culture. *Genome Res* 2013, 23:855-866.

33. Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC: A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes. *BMC Genomics* 2014, 15:1138.

34. Maiden MC, Harrison OB: Population and Functional Genomics of Neisseria Revealed with Gene-by-Gene Approaches. *J Clin Microbiol* 2016, 54:1949-1955.

35. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q et al.: Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science* 1998, 282:754-759.

36. Huson DH, Bryant D: Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006, 23:254-267.

37. Jolley KA, Maiden MC: BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010, 11:595.

38. Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC: Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods. *J Clin Microbiol* 2012, 50:3046-3053.

39. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011, 28:2731-2739.

40. Nei M KS: Molecular evolution and phylogenetics. *Oxford: Oxford University Press* 2000.

41. Ferreira R, Borges V, Nunes A, Borrego MJ, Gomes JP: Assessment of the load and transcriptional dynamics of Chlamydia trachomatis plasmid according to strains' tissue tropism. *Microbiol Res* 2013, 168:333-339.

4

42. Nunes A, Borrego MJ, Gomes JP: Genomic features beyond Chlamydia trachomatis phenotypes: what do we think we know? *Infect Genet Evol* 2013, 16:392-400.

43. Abdelsamed H, Peters J, Byrne GI: Genetic variation in Chlamydia trachomatis and their hosts: impact on disease severity and tissue tropism. *Future Microbiol* 2013, 8:1129-1146.

44. Lutter EI, Bonner C, Holland MJ, Suchland RJ, Stamm WE, Jewett TJ, McClarty G, Hackstadt T: Phylogenetic analysis of Chlamydia trachomatis Tarp and correlation with clinical phenotype. *Infect Immun* 2010, 78:3678-3688.

45. Nunes A, Gomes JP: Evolution, phylogeny, and molecular epidemiology of Chlamydia. *Infect Genet Evol* 2014, 23:49-64.

46. Abdelrahman YM, Belland RJ: The chlamydial developmental cycle. *FEMS Microbiol Rev* 2005, 29:949-959.

47. Yu HH, Tan M: Sigma28 RNA polymerase regulates hctB, a late developmental gene in Chlamydia. *Mol Microbiol* 2003, 50:577-584.

48. Yu HH, Kibler D, Tan M: In silico prediction and functional validation of sigma28-regulated genes in Chlamydia and Escherichia coli. *J Bacteriol* 2006, 188:8206-8212.

49. Klint M, Thollesson M, Bongcam-Rudloff E, Birkelund S, Nilsson A, Herrmann B: Mosaic structure of intragenic repetitive elements in histone H1-like protein Hc2 varies within serovars of Chlamydia trachomatis. *BMC Microbiol* 2010, 10:81.

50. Hackstadt T, Brickman TJ, Barry CE, III, Sager J: Diversity in the Chlamydia trachomatis histone homologue Hc2. *Gene* 1993, 132:137-141.

51. Jeffrey BM, Suchland RJ, Quinn KL, Davidson JR, Stamm WE, Rockey DD: Genome sequencing of recent clinical Chlamydia trachomatis strains identifies loci associated with tissue tropism and regions of apparent recombination. *Infect Immun* 2010, 78:2544-2553.

52. Joseph SJ, Didelot X, Rothschild J, de Vries HJ, Morre SA, Read TD, Dean D: Population genomics of Chlamydia trachomatis: insights on drift, selection, recombination, and population structure. *Mol Biol Evol* 2012, 29:3933-3946.

53. Comanducci M, Ricci S, Cevenini R, Ratti G: Diversity of the Chlamydia trachomatis common plasmid in biovars with different pathogenicity. *Plasmid* 1990, 23:149-154.

54. Gong S, Yang Z, Lei L, Shen L, Zhong G: Characterization of Chlamydia trachomatis plasmid-encoded open reading frames. *J Bacteriol* 2013, 195:3819-3826.

55. de Vries HJ, Zingoni A, Kreuter A, Moi H, White JA: 2013 European guideline on the management of lymphogranuloma venereum. *J Eur Acad Dermatol Venereol* 2015, 29:1-6.

56. Bom RJ, van den Hoek A, Wang Q, Long F, de Vries HJ, Bruisten SM: High-resolution typing reveals distinct Chlamydia trachomatis strains in an at-risk population in Nanjing, China. *Sex Transm Dis* 2013, 40:647-649.

57. Versteeg B, van Rooijen MS, Schim van der Loeff MF, de Vries HJ, Bruisten SM: No indication for tissue tropism in urogenital and anorectal Chlamydia trachomatis infections using high-resolution multilocus sequence typing. *BMC Infect Dis* 2014, 14:464.

58. Versteeg B, Himschoot M, van den Broek IV, Bom RJ, Speksnijder AG, Schim van der Loeff MF, Bruisten SM: Urogenital Chlamydia trachomatis strain types, defined by high-resolution multilocus sequence typing, in relation to ethnicity and urogenital symptoms among a young screening population in Amsterdam, The Netherlands. *Sex Transm Infect* 2015, 91:415-422.

Supplementary files

Supplementary Table 1: *The number of allelic variants for the CHALM0456 (Tarp) gene in respect to disease phenotype (ocular, urogenital, LGV). The samples are sorted by allelic variant.*

Gene	Allelic variation	Disease phenotype		
		Ocular	Urogenital	LGV
Tarp	1	-	7	1
	6	-	1	-
	7	-	2	-
	13	1	-	7
	14	-	-	1
	15	-	-	1
	16	-	-	1
	18	-	1	-
	19	1	-	-
	20	3	3	-
	21	-	1	-
	22	1	1	-
	23	-	1	-
	24	-	3	-
	25	2	18	1
	26	-	1	-
	27	-	-	1
	28	-	4	-
	29	-	1	-
	30	-	-	1
	31	-	-	1
	32	-	1	-
	33	3	2	1
	34	-	1	-
	35	-	1	-
	36	1	2	1
	37	-	1	-
	38	-	1	-
	39	-	1	-
	40	-	2	-
	41	-	1	-
	42	-	2	-
	44	1	1	-
	46	-	1	-
	47	-	1	-
	48	1	-	-

4

Supplementary Table 2: *The number of allelic variants for each plasmid gene (pgp1 to -8) in respect to the observed plasmid clusters among 157 C. trachomatis isolates. The samples are sorted by plasmid gene and allelic variant.*

Gene	Allelic variant	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pgp1	1	-	-	-	-	-	32
	2	-	-	-	22	-	-
	3	19	-	-	-	-	-
	5	-	58	-	-	-	-
	6	-	-	-	-	10	-
	9	-	-	8	-	-	-
	10	-	-	1	-	-	-
	11	-	-	-	-	-	2
	13	-	2	-	-	-	-
	14	-	1	-	-	-	-
	15	-	1	-	-	-	-
	17	1	-	-	-	-	-
pgp2	1	-	-	-	-	-	34
	2	-	-	-	-	1	-
	3	-	-	-	19	-	-
	4	17	-	-	-	-	-
	5	2	-	-	-	-	-
	6	-	36	-	-	-	-
	7	-	5	-	-	-	-
	8	-	-	-	3	-	-
	9	-	19	-	-	-	-
	11	-	-	-	-	9	-
	12	1	-	-	-	-	-
	13	-	2	-	-	-	-
	14	-	-	8	-	-	-
	15	-	-	1	-	-	-
pgp3	1	-	-	-	-	-	30
	2	-	-	-	21	-	-
	3	20	-	-	-	-	-
	4	-	59	-	-	-	-
	5	-	-	9	-	-	-
	6	-	1	-	-	-	-
	7	-	-	-	-	-	1
	8	-	-	-	-	10	-
	9	-	-	-	-	-	3
	11	-	1	-	-	-	-
	12	-	1	-	-	-	-
	13	-	-	-	1	-	-
pgp4	1	20	62	9	-	-	34
	2	-	-	-	22	-	-
	4	-	-	-	-	10	-

Supplementary Table 2: *continued*

Gene	Allelic variant	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pgp5	1	-	-	-	-	5	31
	2	19	-	-	-	-	-
	3	-	-	-	21	-	-
	4	-	62	-	-	-	-
	5	-	-	-	-	-	2
	6	1	-	-	-	-	-
	7	-	-	1	-	-	-
	8	-	-	-	-	5	-
	9	-	-	5	-	-	-
	10	-	-	-	-	-	1
	11	-	-	2	-	-	-
	15	-	-	1	-	-	-
	16	-	-	-	1	-	-
pgp6	1	-	-	-	-	-	28
	2	-	-	-	22	-	-
	3	17	-	-	-	10	-
	4	-	58	-	-	-	-
	6	-	-	1	-	-	-
	7	-	1	-	-	-	-
	8	-	-	-	-	-	1
	9	-	-	-	-	-	1
	10	-	-	8	-	-	-
	11	-	1	-	-	-	-
	12	-	1	-	-	-	-
	14	1	-	-	-	-	-
	15	-	-	-	-	-	3
	16	1	-	-	-	-	-
	18	-	1	-	-	-	-
	19	-	-	-	-	-	1
	22	1	-	-	-	-	-
pgp7	1	14	-	-	-	10	32
	2	-	-	-	-	-	1
	4	-	-	-	19	-	-
	5	-	61	-	-	-	-
	6	-	-	-	3	-	-
	7	1	-	-	-	-	-
	8	-	1	-	-	-	-
	9	-	-	-	-	-	1
	10	3	-	-	-	-	-
	12	-	-	6	-	-	-
	13	-	-	3	-	-	-
	15	1	-	-	-	-	-
	16	1	-	-	-	-	-

Supplementary Table 2: *continued*

Gene	Allelic variant	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pgp8	1	17	-	-	-	9	34
	2	-	-	-	22	-	-
	3	1	-	-	-	-	-
	4	-	38	-	-	-	-
	5	-	24	-	-	-	-
	8	-	-	3	-	-	-
	9	1	-	-	-	-	-
	10	-	-	-	-	1	-
	11	-	-	5	-	-	-
	12	1	-	-	-	-	-
	18	-	-	1	-	-	-