Supplementary data

Global multilocus sequence type (MLST) analysis of *Chlamydia trachomatis* strains from 16 countries

MLST from whole genome analysis

Materials and Methods

Suppl Table 1: List of genomes used for MLST analysis

ST	Short name	Accession number
108	ST108-1	ERR386231_58082
12	ST12-1	NZ_ACFJ01000001
12	ST12-2	NC_017953
12	ST12-3	ERR164690_9528
3	ST3-1	ERR108280_9613
52	ST52-1	NC_017432
52	ST52-2	NC_017440
52	ST52-3	ERR108295_9615
56	ST56-1	NC_017952
56	ST56-2	NC_017439
59	ST59-1	ERR111561_10581
59	ST59-2	ERR210999_42767
136	ST136-1	6276s_689
unknown	UNK-1	B-Jali20-OT_693
35	ST35-1	D-SotonD6_10483
unknown	UNK-2	E-Bour_10504
unknown	UNK-3	ERR021962_9459
unknown	UNK-4	ERR026568_9527
unknown	UNK-5	ERR034339_9538
24	ST24-1	ERR111610_9600
55	ST55-1	ERR211055_42800
unknown	UNK-6	E-SW3_10471
unknown	UNK-7	G-SotonG1_9499
20	ST20-1	NC_000117
unknown	UNK-8	NC_007429
278	ST278-1	NC_017430
unknown	UNK-9	NC_020929
144	ST144-1	NC_020930
unknown	UNK-10	NC_020944
unknown	UNK-11	NC_020970
unknown	UNK-12	NC_021888
141	ST141-1	NC_021899
46	ST46-1	NC_023060
unknown	UNK-13	NZ_ABYG01000001

STs in red denotes founders that predominate in the database.

Suppl table 2: Orthologous groups containing only genomes with founder profiles, groups representing more than one genome [a]

Group number	Genomes represented	Similar to
I1.5_977	ST108-1, ST12-1, ST12-3, ST52-1, ST52-2	hypothetical protein G9768_00845 [Chlamydia trachomatis G/9768] Genovar G, male urethra specimen.
I1.5_1011 [b]	ST108-1, ST52-3, ST59-1	putative gag-pol protein (mouse), RNase_HI_RT_Bel, integrase (98% id)
I1.5_1033	ST52-1, ST52-2, ST52-3	hypothetical protein CTG9301_02525 [Chlamydia trachomatis G/9301] Genovar G, male rectal specimen.
I1.5_1054	ST108-1, ST52-3	hypothetical protein [Plasmodium yoelii yoelii 17XNL] (90% id)
I1.5_1055	ST108-1, ST52-3	hypothetical protein [Plasmodium berghei ANKA] (96% id)
I1.5_1058	ST108-1, ST52-3	small fragment, (part of) integrase [Plasmodium yoelii yoelii 17XNL]
I1.5_1059	ST108-1, ST59-1	hypothetical protein [Plasmodium berghei ANKA] (98% id)
I1.5_1061 [b]	ST12-3, ST59-2	penicillin-binding protein [Chlamydia trachomatis E/150] Genovar E, rectum specimen.
I1.5_1079	ST3-1, ST59-2	hypothetical protein [Macaca fascicularis] (87% id)
I1.5_1084 [b]	ST52-3, ST59-1	endonuclease/exonuclease/phosphatase [Plasmodium yoelii yoelii 17XNL] (99% id)
I1.5_1085	ST52-3, ST59-1	mCG147802 [Mus musculus] (96% id)
I1.5_1086	ST52-3, ST59-1	Endonuclease domain (L1-EN) + RT_nLTR, ORF2 gb|AAC72797.1| [Mus musculus domesticus]
I1.5_1087 [b]	ST52-3, ST59-1	putative gag-pol protein [Mus musculus]

[a] In addition to the 17 groups of orthologous proteins presented in the study 9 other groups were detected, but were only present in the draft genome ERR386231 (having ST108). These groups were caused by fragmented gene predictions and were excluded from further analyses.

[b] Sequence is part of a protein found in other genomes.

Phylogenetic analysis

Methods

For the *hr-CT-MLST dataset* three additional methods were used to account for conflicts in the concatenated alignment. Putative recombinant sequences were identified using the RDP, MaxChi, and BootScan statistics with default parameter settings in RDP4. In sequences indicated as recombinants the sequence indicated as coming from the minor parent was removed and the RAxML analysis redone as above. This analysis will be referred to as no recombinations. As an alternative rogue taxa, i.e. taxa lowering the support values

disproportionately [1] were identified using RogueNaRok at default settings [2]. The rogue taxa were then excluded from the phylogeny and support values recalculated. This analysis will be referred to as no rogues.
Conflicts between genes were removed as described in paper. This analysis is referred to as no conflict.

Results

In the *hr-CT-MLST dataset* three genotypes were removed from *pbpB* due to long branches. Thirteen potential recombinations were identified out of 30 genotypes in CT144. This gene also had two regions suspected to have non-homologous introns and was therefore excluded from the concatenated analysis. No recombinations were detected in the other genes.

There were 414 unique MLST profiles in the database. To resolve conflicts between genes *pbpB* was removed for 75 strains, CT058 from 70 strains, *hctB* from 8 strains, and CT172 from 3 strains. No single strain had more than one gene removed. After removing conflicts the tree is in large agreement with Harris et al. [3], see Fig. S1. The only conflict supported with BS >70 is the same conflict described above for the L strain clade in the Harris et al. dataset. Like in the Harris et al. dataset, there are also several differences in the clade composing G, D, K, Ia, and J strains, but these differences does not have high support from the Database dataset. Since the *hr-CT-MLST dataset* include many more strains it is not surprising that it also includes clades that are not present in the Harris dataset or in Harris et al. [3]. Six strains were found in intermediate position between clades, and were identified by at least one of the three methods as putative recombinant or rogue taxa and their phylogenetic position may therefore be an artifact of unresolved conflicts between genes.

RDP4 identified 230 recombinant strains from the concatenated alignment. The no-recombinations tree identified many of the same clusters as the no-conflict tree but the relationships between the clusters and the patterns of nestedness were very different between the two. The no-recombinations tree is in much less agreement with Harris et al. [3] than the no-conflict tree (data not shown).

RogueNaRok identified 293 rogue taxa which were removed from the tree. The no rogue tree is in large agreement with the no-conflict tree (data not shown). The most notable difference was in the position of strains that had conflict between genes but were not identified as rogue taxa. The no rogue tree has the same basic disagreements with Harris et al. [3] as the no-conflict tree although many of the strains from Harris et al. were removed as rogue taxa.

References

1. Thomson RC, Shaffer HB. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst Biol 2010; 59: 42–58.

2. Aberer AJ, Krompass D, Stamatakis A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst Biol 2013; 62: 162–6.

3. Harris SR, Clarke IN, Seth-Smith HMB 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.

Figure legend to Fig. S1

Phylogenetic tree of the hr_CT-MLST dataset by maximum likelihood analysis after removal of recombination events. The tree comprised 418 unique MLST STs including 52 STs from Harris et al (3). Numbers on nodes represent percentage of bootstrap replicates supporting each clade. The tree was rooted using the L strains as outgroup. Code for branch designations: Specimen number_genovar_allele numbers for hctB_CT 058 _CT 172_pbpB_ST5 number.

Fig. S1.