Novel overlapping coding sequences in *Chlamydia trachomatis*

Klaus T. Jensen[1], Lise Petersen[2], Søren Falk[1], Pernille Iversen[2], Peter Andersen[1], Michael Theisen[1] & Anders Krogh[2]

[1]Laboratory of Infectious Diseases Immunology, Statens Serum Institut, Artillerivej, Copenhagen, Denmark; and [2]The Bioinformatics Centre, Institute of Molecular Biology and Physiology, University of Copenhagen, Universitetsparken, Copenhagen, Denmark

Correspondence: Klaus T. Jensen, Statens serum institute, Chlamydia Vaccine Unit, Department of Infectious Disease Immunology, 5 Artillerivej, DK-2300 Copenhagen S, Denmark. Tel.: +45 32683771; fax: +45 32683148; e-mail: ksa@ssi.dk

Received 31 May 2006; revised 31 August 2006; accepted 11 September 2006.
First published online 13 October 2006.

DOI:10.1111/j.1574-6968.2006.00480.x

Editor: Michael Galperin

Keywords
Chlamydia trachomatis; gene prediction; overlapping genes.

Abstract

Chlamydia trachomatis is the aetiological agent of trachoma and sexually transmitted infections. The *C. trachomatis* genome sequence revealed an organism adapted to the intracellular habitat with a high coding ratio and a small genome consisting of 1.042-kilobase (kb) with 895 annotated protein coding genes. Here, we repredict the protein-coding genes of the *C. trachomatis* genome using the gene-finder EasyGene that was trained specifically for *C. trachomatis*, and compare it with the primary *C. trachomatis* annotation. Our work predicts 15 genes not listed in the primary annotation and 853 that are in agreement with the primary annotation. Forty two genes from the primary annotation are not predicted by EasyGene. The majority of these genes are listed as hypothetical in the primary annotation. The 15 novel predicted genes all overlap with genes on the complementary strand. We find homologues of several of the novel genes in *C. trachomatis* Serovar A and *Chlamydia muridarum*. Several of the genes have typical gene-like and protein-like features. Furthermore, we confirm transcriptional activity from 10 of the putative genes. The combined evidence suggests that at least seven of the 15 are protein coding genes. The data suggest the presence of overlapping active genes in *C. trachomatis*.

Introduction

Genital infections with *Chlamydia trachomatis* are the most common sexually transmitted bacterial infections worldwide, with more than 90 million cases annually. The infection can result in chronic changes in the genital tract causing infertility, abdominal pain and ectopic pregnancies. *C. trachomatis* is also the causative agent of trachoma, a chronic infection of the conjunctiva characterized by extensive scarring and blindness. The genus *Chlamydia* belongs to the taxonomic family *Chlamydiaceae*, which also contains the genus *Chlamydophila*. (Everett *et al.*, 1999; Bush & Everett, 2001) The genus *Chlamydia*, which includes *C. trachomatis* and the mouse pathogenic *Chlamydia muridarum*, was the first genus to diverge from the *Chlamydiaceae* tree. The divergence of the *Chlamydophila* genus occurred later and stepwise with the emergence of *Chlamydophila pneumoniae* before most other *Chlamydophila* members. The division of the two genera is based on 16S and 23S rRNA gene sequence homology as well as other genotypic and phenotypic markers (Everett *et al.*, 1999; Bush & Everett, 2001). The distribution of virulence traits such as host and tissue tropism does not reflect the phylogenies, but might reflect adaptation to the environment.

The genome of *C. trachomatis* serovariant D was sequenced in 1998 (Stephens *et al.*, 1998) and this had great impact on the field. The usability of a genome is very dependent on the annotation of the genomic sequence. The annotation is in turn very closely connected to the quality of the databases used for similarity searches with the predicted genes and other tools used for annotation. New genes are often identified by several layers of successive experimental testing. The lab-work requires time and resources and is not feasible for the majority of genes at the time of completion of the genome sequencing. The alternative is computational gene finding, in which protein-coding genes are predicted based on general features such as the ORF length, codon usage statistics and strength of the ribosome-binding site (Fickett, 1996). Very long ORFs are unlikely to occur at random in a DNA sequence and are easy to identify. On the other hand, noncoding small ORFs are frequent even in a random sequence of bases, and one of the major problems encountered by gene-finders is discrimination of small genes (shorter than 300–600 nts) from random ORFs. Thus, small

ORFs tend to be annotated with a high ratio of false-positives (Skovgaard *et al.*, 2001). Another problem is management of overlapping genes. In overlapping genes, the same DNA sequence encodes multiple proteins from different reading frames. A short overlap of genes (1–10 bp) within the same transcription unit is common in some genomes (Salgado *et al.*, 2000; de Hoon *et al.*, 2005), but longer overlaps are quite rare in prokaryotes. Overlapping genes are by many gene finders assumed not to exist and only the best of two overlapping gene predictions is reported (Frishman *et al.*, 1998; Lukashin & Borodovsky, 1998; Salzberg *et al.*, 1998). The primary annotation of *C. trachomatis* strain D/UW-3/Cx involved a combination of the genefinders GeneMark (Borodovsky & McIninch, 1993; Lukashin & Borodovsky, 1998) and Glimmer version 1.0 (Salzberg *et al.*, 1998). Both genefinders use hidden Markov models. For overlapping genes, both genefinders use rule-based criteria. The GeneMark system used the Viterbi algorithm as the core procedure, which basically did not take overlapping genes into account (Lukashin & Borodovsky, 1998) or tended to predict them as short as possible. To deal with overlapping genes, a post-processing procedure was used to search for ribosome-binding sites to strengthen the prediction. The Glimmer annotation system discards the shorter of the overlapping ORFs if the overlap regions scores the highest in the longest ORF (Delcher *et al.*, 1999). A newer version of Glimmer has further improvements in resolving overlapping genes (http://www.cbcb.umd.edu/software/glimmer/). Glimmer was first developed at The Institute for Genomic Research (TIGR) where it is used as the primary microbial genefinder. The novel genefinder EasyGene matches or exceeds other newer gene-finders in performance, and has proven superior, especially in limiting the number of short 'random' ORFs that are predicted as real genes by considering the probability that the ORF could be random (Larsen & Krogh, 2003; Nielsen & Krogh, 2005). In EasyGene, there is no postprocessing of overlapping predictions, so if an ORF is significant, it is reported. In this work, we train and run the EasyGene application on the genomic sequence of *C. trachomatis* strain D/UW-3/Cx. We compare the gene predictions with the primary annotation (Stephens *et al.*, 1998) and find a good overall agreement between them with 95% identity (853 genes) between the two gene sets. The EasyGene prediction includes 15 novel genes not included in the primary annotation, while 42 mainly hypothetical genes included in the primary annotation are excluded from the EasyGene prediction. The 15 novel potential protein-coding sequences (CDSs) are located on the complementary strand of other genes included in the primary annotation, either completely overlapping (eight ORFs) or partially (seven ORFs) overlapping the other gene. Several of the novel ORFs have a high degree of homology to ORFs in *C. muridarum* and *C. trachomatis* Serovar A and contain signal peptide sequences and Shine–Dalgarno (SD) sequences characteristic of protein-coding sequences. Transcriptional activity from several of the novel genes is shown by strand-specific RT-PCR. The combined evidence suggests that seven of the 15 predicted ORFs are protein-coding sequences.

Materials and methods

Gene prediction

The *C. trachomatis* Serovar D/UW-3/Cx (Ct-SvD) (NC000117) genome was run through the gene prediction tool EasyGene (Larsen & Krogh, 2003; Nielsen & Krogh, 2005) (http://servers.binf.ku.dk/cgi-bin/easygene/search). For comparison and determination of homologies of novel genes, the genomes of *C. muridarum* NIGG (Cm) (NC002620), *C. pneumoniae* AR39 (Cp) (NC002179) and *C. trachomatis* Serovar A/HAR-13 (Ct-SvA) (NC007429) were also repredicted with the EasyGene algorithm. Gene prediction with EasyGene was performed as described (Larsen & Krogh, 2003; Nielsen & Krogh, 2005). Briefly, all ORFs longer than 120 bp were extracted from the genome and compared using BLAST against Swiss-Prot (Release 41) (Altschul *et al.*, 1990; Boeckmann *et al.*, 2003) to generate a training set for the model. EasyGene is based on a Hidden Markov Model (HMM) (Lukashin & Borodovsky, 1998) and incorporates various types of genome-specific signals for prediction of gene location. Overlapping genes are predicted with high confidence (Larsen & Krogh, 2003; Nielsen & Krogh, 2005). The probability of finding a high-scoring ORF is highly dependent on length, and short ORFs will frequently obtain a high score randomly, simply because there are many short ORFs in a genome. EasyGene introduces the *R*-value score for each gene prediction. The *R*-value is the number of ORFs of the same length as the gene prediction that will obtain the same score or higher, in 1 Mb of random sequence. We used an *R*-value cutoff of 2, meaning that one can expect two false-positive predictions in 1 Mb of genomic sequence.

The gene prediction HMM algorithm includes separate models for start and stop codons, genome statistics and ribosome-binding sites (RBS). The parameters for the RBS model are trained from the individual genomes and do not rely on a consensus sequence. The algorithm is therefore capable of predicting genes with nonconsensus ribosome-binding sites or without SD motifs. The SD motifs preceding the predicted novel CDSs were inspected manually with regard to upstream distance and sequence. The sequences were matched against the anti-SD core-motif CCUCC in the 16S rRNA gene (Osada *et al.*, 1999; Saito & Tomita, 1999; Ma *et al.*, 2002) and only slight variations were accepted

(Schurr *et al.*, 1993). The aligned upstream distance of an SD sequence was defined as the spacing in bases from the first base in the start codon to the U in the core anti-SD motif CCUCC after duplex formation.

Database searches

Predictions of transmembrane helices in the novel CDSs were performed with TMHMM (Krogh *et al.*, 2001) (http://www.cbs.dtu.dk/services/TMHMM/). Predictions of signal peptides in the novel CDSs were performed with SignalP (Bendtsen *et al.*, 2004) (http://www.cbs.dtu.dk/services/SignalP/). All novel CDSs were submitted to PFAM (Bateman *et al.*, 2004) (http://www.sanger.ac.uk/Software/Pfam/) to search for domain or protein family homology.

General patterns concerning overlapping genes were analysed. Distances between adjacent genes were extracted from the primary annotation of Ct-SvD. Also, distances between pairs of genes with adjacent 3′ ends or 5′ ends were extracted. Only overlapping genes were considered. The lengths of the overlaps were measured in base pairs. The relative phases of the overlapping reading frames (the reading frame offset) were extracted in the analysis. Overlap phases and reading frames were named as described elsewhere (Boldogkoi & Barta, 1999; Krakauer, 2000; Johnson & Chisholm, 2004). Briefly, opposite-strand overlaps can exist in phase 0, phase -1 or phase -2. Phase 0 denotes in-phase reading frames on opposite strands ($0+3n$ shared bases). In phase -1 ($2+3n$ shared bases), codon position 3 of the sense and antisense frames are complementary. In phase -2 ($1+3n$ shared bases), codon position 2 of the sense and antisense frames are complementary. Same-strand overlaps can exist in phase $+1$ or phase $+2$. In phase $+1$ ($2+3n$ shared bases), reading frames 1 and 2 are overlapping. In phase $+2$ ($1+3$n shared bases), frame 1 and 3 are overlapping. In-phase same-strand overlapping rading frames require a stop codon read-through and are very rare (Keese & Gibbs, 1992).

Homology searches

The protein sequences of the 16 novel ORFs predicted by the EasyGene algorithm were matched against a low complexity filtered database containing ORFs from Cm, Cp and Ct-SvA using BLASTP (basic local alignment search tool) (Altschul *et al.*, 1990) with an e-value cutoff of 0.001. Low complexity filtering was performed using the SEG algorithm with standard settings (Wootton & Federhen, 1993). The matching ORFs were inspected manually, and homologous sequences were identified according to current guidelines (Mount, 2004). Homologous sequences were further compared with the EasyGene reprediction of the Cm, Cp and Ct-SvA genomes. In addition, the protein sequences of the novel ORFs were matched against the genome sequences of Cm, Cp and Ct-SvA using TBLASTN with an e-value cutoff of 0.001.

The protein sequences for the 42 annotated genes missing from the EasyGene prediction were matched against a database containing all annotated proteins from Ct-SvA, Cm and Cp using BLASTP with an e-value cutoff of 0.001. The procedure was repeated with a database containing either an old version or the newest available version of Swiss-Prot (version 41 and 50.4). The Swiss-Prot databases were modified as mentioned elsewhere (Nielsen & Krogh, 2005). Briefly, the content of the ID line in the FASTA version of Swiss-Prot was changed to include keyword information for further processing. The databases was filtered using SEG (Wootton & Federhen, 1993). The BLAST reports from the analyses can be downloaded from http://binf.ku.dk/∼krogh/chlamydia.

Chlamydia, cell lines and infections

McCoy cells (ATCC CRL-1696) and Hela-229 (Hela) (ATCC CCL-2.1) were grown in complete media (RPMI-1640 supplemented with 5% FCS, 10 mM Hepes, 2 mM glutamate, Gentimycin (50 g mL^{-1})) and maintained by passage every 3–4 days at subconfluency.

Chlamydia trachomatis serovar D strain UW3/Cx (Ct-SvD) (ATCC VR-885) were maintained by propagation in Hela cells. Briefly, the Hela cells were passaged overnight in T175 cell culture flasks (Nunc, Denmark) to confluency. The cells were treated with DEAE dextran (45 g mL^{-1}) for 15 min, washed in PBS, followed by addition of the Ct-SvD at a multiplicity of infection (MOI) of 1 diluted in complete media with glucose (0.5%) and cycloheximide (1 g mL^{-1}) and incubated for 2 h at 37 °C. Ct-SvD elementary bodies (EB) were harvested as described (Schachter & Wyrick, 1994). Briefly, the infected Hela cells were washed with 37 °C PBS and detached in 5 mL SPG (sucrose, phosphate, glutamate) with a cell scraper. The bacteria were releashed from the infected Hela cells by a brief sonication (31 000 J), washed by pelleting the EB's at 30 000 ***g*** and resuspended in SPG. The EBs were further purified by discontinous gradient centrifugation. The EBs were layered on top of a discontinous gradient (30% Renografin) and centrifuged for 2 h at 18 000 ***g*** using an AH-629 rotor. Several pellets (3–6 depending on the outcome) were merged and resuspended in SPG and stored at -80 °C.

Infectivity of the Ct-SvD stock was determined by titred infections of McCoy and Hela cells. The cells were passaged overnight in 6-well plates (Nunc, Denmark) to confluency in complete media. After inoculation with dilutions of the Ct-SvD stock in SPG, the media were supplemented with glucose and cycloheximide and the plates were centrifuged for 1 h at 750 ***g*** at room temperature (RT). After incubation

for 2 h at 35 °C, the supplemented media were replaced with fresh complete media supplemented with glucose and cycloheximide and the infected cells were incubated at 37 °C in 5% CO_2. The Ct-SvD infected cells were after 70 h of infection fixed with 90% ethanol for 15 min at 4 °C, stained with 5% propidium iodide for 5 min, followed by incubation with a rabbit anti-MOMP serum for 45 min at room temperature. After three washes in PBS, a secondary FITC-conjugated swine-antirabbit Ig antibody was added and incubated for 45 min at RT.

For isolation of Ct-SvD RNA, McCoy cells passaged overnight to subconfluency in 6-well plates and infected with various MOIs: the cells were incubated for 12 h (MOI 10), 24 h or 48 h (MOI 1).

RNA purification and RT-PCR

Total RNA from infected McCoy cultures (2×10^6 McCoy cells plated in 6-well plates) were isolated at 12, 24 and 48 h. The cells were harvested in SPG by rolling glass beads, washed by pelleting at 500 ***g*** and resuspended in SPG. Total RNA was isolated with the RNAeasy kit (Qiagen) according to the manufacturer's protocol. The integrity of the RNA was checked by agarose gel-electrophoresis before downstream applications (data not shown). The presence of transcription from the predicted CDSs was analysed by RT-PCR. To ensure a strict strand-specific approach, several steps were taken. The RNA was cleared of residual genomic DNA by a treatment with DNAse I (Boehringer Mannheim), followed by phenol:chloroform extraction to remove the DNAse. Specific primers (Table 1) were used for priming the cDNA synthesis. Between 2 and 8 μg of total RNA was used as a template for the reverse transcription (Omniscript, Qiagen). After reverse transcription, residual reverse transcriptase was removed by phenol–chloroform extraction, followed by removal of residual RT-primer using DNA spin-columns (Microspin S-300, Amersham Biosciences). The cDNA was used as a template for PCR. All primer-pairs (Table 1) were generated using Primer3 (Rozen & Skaletsky, 2000). For increasing the PCR-specificity, the 3′ primer was different from the RT-primer and placed upstream. All sets of primers were checked on Chlamydia genomic DNA to give unique DNA bands of correct sizes, excluding unspecific reactions. The negative control included parallel reactions with the

Table 1. Specific RT primers, PCR primers and amplicon length

ORF	RT primer	5′ primer	3′ primer	Product length (nt)
EG02	TTCTCCCTGTCGATAGATCA	AGCTACTGTTTTGCGGGCTA	CACAAGAAAAACTGCTGCCA	124
EG03	GACTCCGTCTACAGATCCTATTT	TGCTTGCACGGGTATTTGTA	AGTCACCCCTCCAACAACAG	253
EG04	CGAAGATATTGATATGAGCAGTAA	ATTCTGTGCGTTGCTGTCAC	ATCACCTCTCCACAAATGGC	237
EG05	ATGCTTCGACACCTCATCA	GCTTCCCAAGACCAAGATGA	TGGAGACGCTTCAGCAACTA	183
RG06	ACTTCCTGTTTTGACTCCTAGAC	ATCCAAAAAGCATTTCACCG	CCTCCTTTCCTGCCCTAATC	249
EG07	TCTTGGAGCTTCAGGATTATG	CTTACGGTAGGAGCTCGACG	GGAGCTACTGCCCTGTCAAC	245
EG08	AGAACAAAATATTAAGCTCTCTATCAA	TGCAAGAGACTGCTGTTTGG	CGAGCCTGTGCTTTAGCTCT	225
EG10	TTCTAAAACGGTTACTCTATCTAACC	TGTTAGCCGAACTTCCTTCTG	TCAAGTCTACCTTCACTTTTGCAG	150
EG11	CCCTGGCGACAATGG	CCTGTGCCCGTTCTTAATGT	AAGGGCCTCTCCAAGCTAAC	202
EG12	GCAGAGTTTGATTTCAGGG	GTGGCTTTAGCTCCAACTGC	GATTCGCGGTCTCTGTAGGA	187
EG13	AAACTGAAAAAAATAGTTTAAAACAAC	TACGAACTCTTTGTGCTGCG	GCGCTAAAAGATACGGCAAA	106
EG14	TTCTGAAAGTGCACTTGCA	GATGGATCGAGTTGTTGCTG	GGAGCTTCTTTATCCGGAGC	105
EG15	CTAGTCAACAAGTTGAGGAGAAG	ACAGATCAATCCTGCGATCC	TTTTGGTCCATCTGCATGAC	181
EG16	CTAAGATCTGCGACGAGGT	AATCCTTTTGGGGATGCTTT	TCGCGCTTCTCTTGAAAACT	235
EG17	TTATGCCTTCATTTCAGCA	TGTAGCAGACCCACTACCCC	CTCCTGCTGCTTCAAGCTCT	197
Ct046	TCTAGCGACTAATTTCATTAATTGT	CACAAGAAAAACTGCTGCCA	GCTACAGGCTTACGAGCTGC	156
Ct050	AGGCTGTGATGGTGTCG	AGTCACCCCTCCAACAACAG	AGGAGTGGGTGATGGAGTTG	166
Ct051	AGGTGTTTGTCCTTCTGTTG	TGGAGACGCTTCAGCAACTA	GGTGATGGAGACGAAGGTGT	261
Ct058	TAAATTCACGGGTTGAGGG	GAGAATGCTCTTCTGCCACC	ATCTCCGAAAGCAGCGATTA	245
Ct082	TGAATCGCCTCCTGCA	TGCACGAGGAGGCTTCTTAT	CTAACACGGCCATCCCTAAA	234
Ct247	ATTGAGAAGTAAAACAGAAGGAGC	CGAGCCTGTGCTTTAGCTCT	TGCAAGAGACTGCTGTTTGG	225
Ct414	GAATGTCATACGAGCACCG	ATCCGTGAACGAAACAAAGG	CGAATGCTAACCCCATAGGA	232
Ct456	TCCTACGGTATCAATCAGTGC	GAAATGACGGACCTTCTGGA	CCTGTGCCCGTTCTTAATGT	292
Ct579	GTTAAACATAGAGGCTGTCGTT	GCAGGAGGATTATTCGGTCA	CTGCCATACCAGCCATTCTT	238
Ct743	TTTTTTTGTTGAGCGAGTTT	CGCAGCACAAAGAGTTCGTA	TTGTTTTAGGAGCGGCTTTT	248
Ct804	TGTAGCTGCAAGTGCACTT	CAGCAACTGCATTGTTTGCT	ACAGTCCGAAGGTTGTACGG	251
CT813	TATCGAACCACGTCTTCCT	CATCGTAGCCGTCGGTTTAT	ATCATCGCTCCAACCTTTTG	196
Ct864	GATATCTCTAGGATGGTAGGTGTG	TCAGCATCAAAGCCAAACAA	GATCCGCGTGGTTATTGAGT	237
Ct872	AAAGATTCTATTCAAGCCCATG	TGACAAAACGCAGAAGTTCG	AAGGAGGAGATCTTCGCACA	293

reverse transcriptase omitted. The positive control included parallel RT-PCR against HctB (Ct046) (data not shown). For PCR, the 25 μL reaction [primers (2 μM), Taq-polymerase (Qiagen)(1.3 U/reaction), dNTP's (200 μM), $MgCl_2$ (4 mM)] was heated to 94 °C for 5 min, followed by amplification using 35 temperature cycles of (45 s at 94 °C+45 s at 58 °C+45 s at 72 °C) using Taq polymerase (Qiagen). The completed reaction was size separated by electrophoresis on 1.5% agarose containing Ethidium Bromide and visualized under UV-light.

Results

Gene prediction

Running EasyGene on the Ct-SvD genome resulted in a data set very similar to the primary annotation by Stephens *et al.* (1998). The primary annotation contains 895 protein-CDSs compared with 869 CDSs predicted by EasyGene, of which 853 (95.3%) were in agreement with the primary annotation. EasyGene also predicted 15 CDSs that were not included in the primary annotation. The 15 novel ORFs, named EG02-EG08 and EG10-EG17, showed a characteristic pattern as all of them were completely (EG03, EG04, EG05, EG06, EG07, EG08, EG10, EG11, EG12, EG17) or partially (EG02, EG13, EG14, EG15, EG16) overlapping with genes located on the opposite strand (Table 2). While the novel ORFs were unique to the EasyGene prediction, the genes on the opposite strand were included in both the primary annotation and the EasyGene predictions. All the novel predicted genes were in phase −2 relative to the opposite strand overlap.

Although EG14 was not included in the primary annotation, the translational product of this gene, rbp_7 (AJ_421777), was subsequently demonstrated by proteome analysis (Shaw *et al.*, 2002). The coding sequence for rbp_7 is also included in the TIGR annotation of the Ct-SvD genome, with the same startcodon (NT01CT0860). Aligning the novel CDSs against the Ct-SvD TIGR annotation (http://tigrblast.tigr.org/cmr-blast/) resulted in an additional, but imperfect match of EG13 with NT01CT0793. EG13 and NT01CT0793 share the same stop codon; however, the NT01CT0793 start codon is predicted 33 bases upstream of the EG13 start codon.

The algorithm used in the present work predicted that four of the novel ORFs, EG11, EG12, EG13 and EG14, have a consensus-like SD sequence within a distance of up to 14 nucleotides upstream from the start codon (Table 2). The remaining 11 novel ORFs were not predicted to contain SD sequences in the upstream noncoding region.

Predictions of transmembrane helices and signal peptides in the potential proteins were performed with TMHMM and SignalP. The predictions showed that three proteins (EG02, EG04 and EG13) were predicted to contain transmembrane helices, and four proteins (EG02, EG04, EG12 and EG16) were predicted to contain signal peptides (Table 2). None of the novel ORF sequences, except EG14, generated matches in PFAM.

Our work also showed that the primary annotation included 42 coding sequences that were not predicted by

Table 2. ORFs predicted exclusively by EasyGene

Predicted gene	Start position	Stop position	Distance: SD to start codont)	SD motif	*R*-value	Strand	Predicted transmembrane helices	Signal peptide	Phase	Overlap with:
EG02	51940	51449			3.95e-07	−	5	+	−2	Ct046
EG03	55669	56133			5.3e-05	+	0	−	−2	Ct050
EG04	56269	56730			1.1	+	3	+	−2	Ct050
EG05	57446	57868			0.215	+	0	−	−2	Ct051
EG06	68086	68481			1.59	+	0	−	−2	Ct058
EG07	95174	94713			0.85	−	0	−	−2	Ct082
EG08	276806	276312			0.549	−	0	−	−2	Ct247
EG10	481707	481528			0.422	−	0	−	2	Ct414
EG11	533474	533196	10	CACCAGT	0.555	−	0	−	−2	Ct456
EG12	652163	651885	14	TTCGAGG	1.32	−	0	+	−2	Ct579
EG13	862743	863306	10	AAGGAGG	0.0101	+	4	−	−2	Ct743
EG14	940879	940676	9	AAGGAGC	0.565	−	0	−	−2	Ct804
EG15	955612	955247			0.679	−	0	−	−2	Ct813
EG16	1018301	1018606			0.698	+	0	+	−2	Ct864
EG17	1035622	1035161			0.00267	−	0	−	−2	Ct872

Shine–Dalgarno (SD) motifs predicted by the genefinder were inspected manually and matched against the 16S rRNA gene anti-SD CCUCC motif and only slight variations were accepted. The aligned upstream distance of a sequence was defined as the spacing in bases from the first base in the start-codon to the U in the core anti-SD motif. The phase refers to the reading frame offset of the opposite-strand overlapping genes. The phases are phase 0 (an in-phase overlap on opposite strands), phase −1 (2+3*n* shared bases) and phase −2 (1+3*n* shared bases).

EasyGene. Most of these are assigned as hypothetical proteins (31 out of 42). Four of the remaining 11 have function assignments as ribosomal proteins (Ct150 (rl33), Ct786 (rl36), Ct802 (rs18) and Ct810 (rl32)). The 42 annotated genes that were not predicted by EasyGene were in general short (mean 86.2 aa, SD 32.1) compared with the genes uniquely found by EasyGene (mean 127 aa, SD 38.5) or with all annotated *C. trachomatis* genes (mean 348.9 aa, SD 244.4) (supplementary Table S1).

Analysis of overlapping genes

To evaluate the presence of overlapping genes in the Ct-SvD chromosome, we analysed the presence of same-strand and opposite-strand overlapping genes in the primary annotation and recorded the position and length of the overlapping regions (supplementary Table S2). Same-strand overlaps are frequent and the primary annotation predicts 136 same-strand overlaps. (Fig. 1a). Most of these overlaps are in the +2 phase (Table 3). Opposite-strand overlaps are less frequent. The primary annotation predicts 36 overlaps on opposite strands in the Ct genome, 22 of them longer than 4 bp (Fig. 1b). These overlaps were mostly in phase −2 (Table 3). All the overlapping genes in the primary annotation, both same-strand and opposite-strand, are partial overlaps, with no finding of one gene contained completely within another.

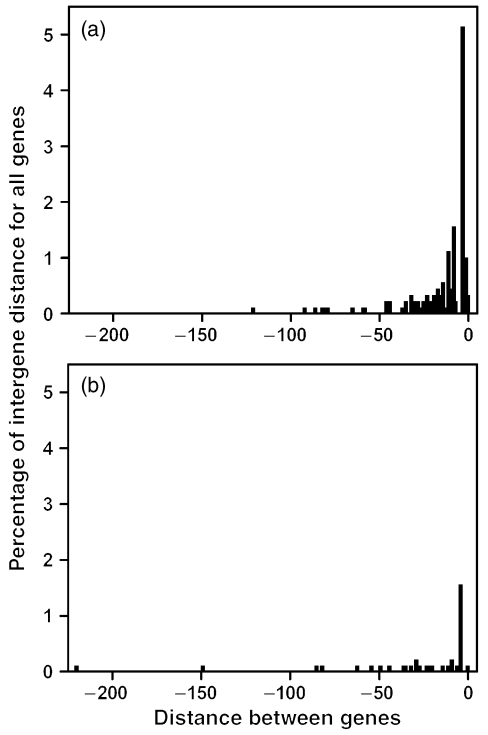

Fig. 1. Histograms of overlapping gene-regions. Positions of start and stop codons of genes included in the primary *Chlamydia trachomatis* serovar D annotation were used to determine overlapping regions of genes. The length in base pairs of the overlapping regions is plotted in a histogram for either same-strand (a) or opposite-strand (b) overlaps.

Homology searches and reprediction of additional *Chlamydiaceae* genomes

The Ct-SvA, Cm and the Cp genomes were reanalysed with EasyGene, to compare them with the novel Ct genes. In Ct-SvA, the EasyGene algorithm identified the homologues of EG02, EG08, EG10, EG11, EG12, EG13 and EG16. The remaining novel genes, EG03, EG04, EG05, EG06, EG07, EG14, EG15 and EG17, had low scores in the Ct-SvA genome, with *R*-values exceeding 2 (Table 4). EG16 is included in the primary Ct-SvA annotation named CTA_0943.

Running EasyGene on the Cm genome identified the homologues of EG02, EG14 and EG16 with *R*-scores below 2. The remaining novel predicted genes had no corresponding ORFs in the Cm genome. (Table 4) The EG16 Cm homologue is included in the primary Cm annotation as Tc0254 (Read *et al*., 2000). Reanalysis of the Cp genome did not predict any of the 16 novel potential genes predicted in Ct-SvD. Furthermore, matching against the Cm genome sequence revealed sequences with homology to EG05, EG07, EG08, EG13 and EG17, but the Cm sequences were not flanked by in-frame start and stop codons. Matching against the Cp genome sequence revealed a sequence with homology to EG08, but not flanked by in-frame start and stop codons.

For 39 of the 42 annotated genes missing from the EasyGene prediction homologues were found in at least one of the other annotated *Chlamydiaceae* genomes (Ct-SvA, Cm and Cp). The BLAST search against Swiss-Prot reported matching proteins for nine of the sequences (supplementary Table S1).

Transcriptional activity from the novel ORFs

To further assess the validity of the novel predicted genes, we investigated whether they were transcriptionally active. A

Table 3. Reading frame phases between overlapping genes included in the primary annotation

	Phase −2	Phase −1	Phase 0	Phase +1	Phase +2
Primary annotation	20	10	6	53	83
EasyGene prediction	15	0	0	0	0

The number of overlaps within codon overlap phases were counted. The phases are phase +2 (1+3*n* shared bases on the same strand), phase +1 (2+3*n* shared bases on the same strand), phase 0 (an in-phase overlap on opposite strands), phase −1 (2+3*n* shared bases on opposite strands) and phase −2 (1+3*n* shared bases on opposite strands).

Table 4. EasyGene prediction of the novel genes in *Chlamydia trachomatis* Serovar A and in *Chlamydia muridarum*

EasyGenes	Ct-SvA EasyGene *R*-value	Ct-SvA coordinates	Ct-SvA primary annotation	Cm EasyGene *R*-value	Cm coordinates	Cm primary annotation
EG02	1.62×10^{-3}	52048–52458		0.0137	377701–373735	
EG03	4.46×10^{-3}	56601–56930				
EG04	3.27×10^{-2}	57021–57632				
EG05	11.78	58294–58602				
EG06	3.77	68810–69205				
EG07	4.38	95432–95686				
EG08	0.34	278564–279058				
EG10	1.13	483927–484106				
EG11	0.46	535902–536180				
EG12	1.55	654595–654873		3.44×10^{-4}	1004761–1005312	
EG13	3.72×10^{-4}	865489–866052				
EG14	3.40	943423–943626		1.21	220026–220229	
EG15	233.00	957998–958177				
EG16	1.47	1020906–1021211	CTA_0943	0.271	297233–297550	TC0254
EG17	2.42×10^{-3}	1037763–1038224				

An EasyGene gene prediction analysis was performed on the Ct-SvA and Cm genomic sequences. *R*-values, genomic coordinates (start and stop positions) and primary annotation numbers are listed in the table for the genes corresponding to the novel ORFs in Ct-SvD.

Table 5. Summary of predictive and transcriptional findings for the novel genes

Novel genes	Predicted Shine–Dalgarno	Predicted signalpeptide	Predicted transmembrane domain	*C. muridarum* homologue	*C. muridarum* reprediction	*C. trachomatius* SvA homologues	Reprediction *C. trachomatis* SvA	RT-PCR
EG02		x	x	x	x	x	x	x
EG03								x
EG04		x	x			x		x
EG05						x		
EG06						x		x
EG07						x		
EG08						x	x	
EG10						x	x	
EG11	x					x	x	x
EG12	x	x					x	x
EG13	x		x			x	x	x
EG14	x			x	x	x		x
EG15						x		x
EG16		x		x	x	x	x	x
EG17						x		

transcript will indicate whether the ORF is a putative gene and that the prediction was correct. To analyse the novel CDSs at the transcriptional level, we analysed the presence of RNA by reverse transcriptase polymerase chain reaction (RT-PCR). The novel CDSs were all sharing regions with genes included in the primary annotation, located on the complementary strand. A strict strand-specific RT-PCR approach was therefore adopted, to ensure specific detection of potentially overlapping RNA. In summary, we observed positive RT-PCR bands of the correct size from 10 of 15 novel CDSs (Table 5, Fig. 2a) and from 10 of 15 genes included in the primary annotation (Fig. 2a). For maps of the regions including the novel genes, see Fig. 2b. Of the 10 RT-PCR-positive novel CDSs, seven (EG02, EG03, EG04, EG06, EG12, EG13, EG14) cotranscribed with the overlapping gene within a Chlamydial life cycle. The remaining three RT-PCR-positive novel CDSs (EG11, EG15, EG16) had opposite strand overlaps that were RT-PCR negative. A total of five novel CDSs were RT-PCR negative. Three of them (EG07, EG08, EG10) had transcriptionally active opposite-strand overlaps. The remaining two (EG05, EG17) had RT-PCR negative opposite-strand overlaps. Of the 15 CDSs included in the primary annotation, 5 (Ct051, Ct456, Ct813, Ct864, Ct872) were negative in the RT-PCR even though

they have confirmed transcriptional activity (Belland *et al*., 2003). In addition, Ct456 and Ct872 also have confirmed translational products and the negative RT-PCR presented here may be caused by target RNA copy numbers below the detection level. This suggests in addition that a similar effect might result in the RT-PCR-negative predicted novel genes being below the detection level and therefore false negatives.

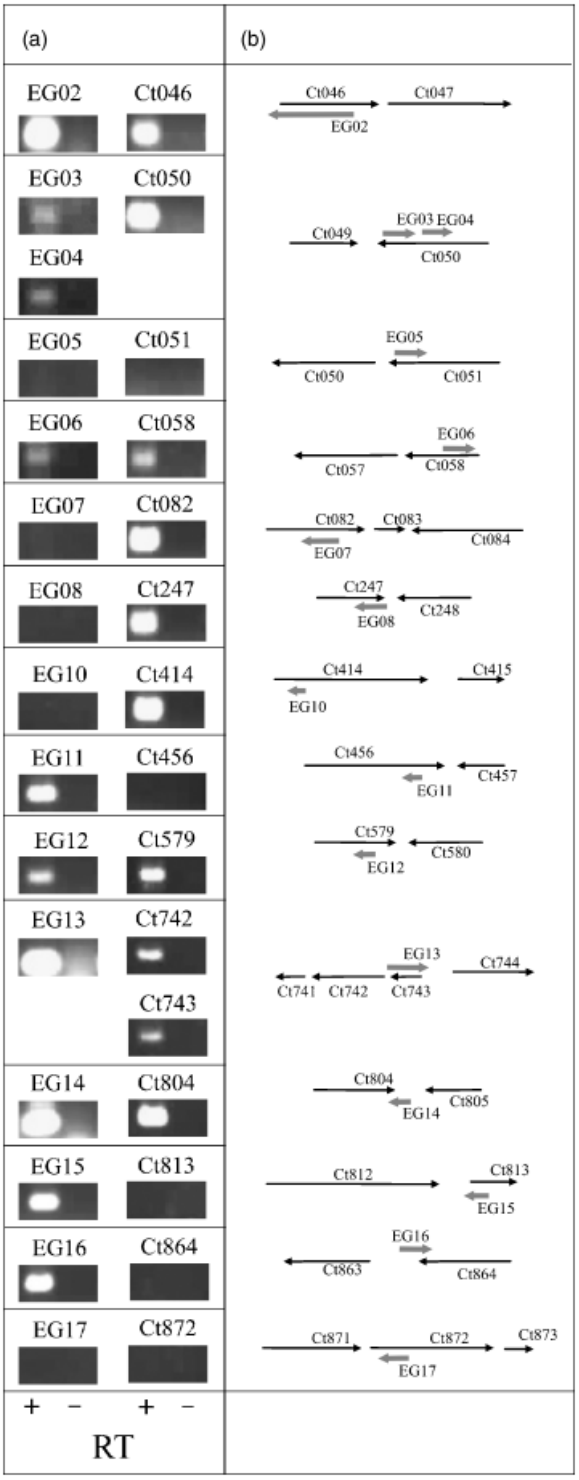

Fig. 2. Transcriptional activity of the novel ORFs and overlapping genes. To assess the transcriptional activity of the novel ORFs, we analysed the presence of RNA by strand-specific RT-PCR. After the specifically primed reverse transcription, any residual reverse transcriptase was removed by phenol–chloroform extraction. In addition, any unused RT-primer was removed on spin-columns. The negative controls included parallel reactions with the reverse transcriptase omitted. The positive controls included parallel PCR against Ct genomic DNA target and RT-PCR against HctB (Ct046). After the reaction, positive bands were visualized by agarose gel electrophoresis and the presence of bands of the right size was recorded. (a) RT-PCR ordered by EG-names (left column) and overlapping genes in the right left column. Each column shows RT-PCR amplicons to the left and minus reverse-transcriptase control to the right. (b) Map of Ct-ORFs (black arrows) and EG-ORFs (grey arrows). The off-set phases of the opposite strand overlapping ORFs are listed in Table 2 and summed up in Table 3.

Discussion

Our knowledge about microbial genomes is constantly increasing and in turn making it necessary to reinterpret genomic data. In this work, we apply advances made in computational genefinding in an attempt to reanalyse the *C. trachomatis* Serovar D/UW-3/Cx (Ct-SvD) genome with the gene prediction algorithm EasyGene. We find an overall good agreement with the primary annotation (Stephens *et al*., 1998) with 853 genes in common. Our work predicts, in addition, a set of 15 novel genes that are not included in the primary annotation. To further increase the strength of the predictions, the 15 novel CDSs were analysed for the presence of homologues in other *Chlamydiaceae* genomes, the presence of sequence motifs characteristic for protein-encoding genes and for transcriptional activity across the ORFs (Table 5). Seven of the novel CDSs were in particular promising with regard to these predictive and transcriptional features. These seven putative genes are EG02, EG04, EG11, EG12, EG13, EG14, and EG16. Three of them, EG02, EG04 and EG16, were all RT-PCR positive, had homologues in Ct-SvA and had predicted signal peptide sequences. EG02 and EG16 were also repredicted in the Cm genome by EasyGene. In addition, EG16 was included in the primary Ct-SvA and Cm annotations (named CTA_0943 and Tc0254, respectively). An SD sequence within the correct upstream distance of the start codon is a strong predictive finding, suggesting that a gene is expressed. SD-like sequences with the correct upstream distance were found for EG11, EG12, EG13 and EG14. These predicted genes were all RT-PCR positive, strongly suggesting protein-encoding ORFs. In addition, all except EG12 had homologues in the

Ct-SvA genome. EG14 was predicted by EasyGene in the Cm genome, but with an *R*-score slightly above the cut-off and therefore not included in the Cm gene prediction. This is interesting, as this gene is known to be expressed in Ct-SvD (Shaw *et al.*, 2002). The combined evidence is in favour of the gene predictions and strongly suggests EG02, EG04, EG11, EG12, EG13, EG14, EG16 to be protein-expressing genes.

Three other novel predicted genes (EG03, EG06 and EG15) were positive in RT-PCR with homologues sequences located in the Ct-SvA genome, but the EasyGene predictions were not supported by any additional predictive features characteristic for protein-coding sequences. The same argument makes it difficult to confirm the RT-PCR-negative CDSs predictions, EG05, EG07, EG08, EG10 and EG17, as protein-encoding genes. The present data therefore do not support the predictions of EG03, EG05, EG06, EG07, EG08, EG10, EG15 and EG17 to be protein-expressing genes. Owing to the lack of sequence motifs necessary for protein-expressing genes especially SD sequences, these genes may be functionless ORFs, or pseudogenes (Mira & Pushker, 2005). Adaptation to an intracellular habitat is known to cause a considerable genome reduction with deletion of obsolete genes. *Chlamydiae* are highly specialized organisms and the adaptation to the intracellular niche is not a recent event. For Ct-SvD, this is observed as an advanced degree of genome decay (Sakharkar *et al.*, 2004), with a high proportion of coding sequence and very few pseudogenes (Liu *et al.*, 2004). A pseudogene overlapping with an active gene on the complementary strand will most likely show a decreased tendency to be deleted. Therefore, the EasyGene predicted genes EG03, EG05, EG06, EG07, EG08, EG10, EG15 and EG17 may be pseudogenes that escaped deletion due to a localization that is overlapping an active gene.

In addition to the 15 novel genes already mentioned, one additional novel predicted gene was added to the list at a late stage. This gene, named EG09, was therefore not subjected to the same range of predictive and experimental analyses as the other 15 novel genes, and was therefore excluded from this work. Despite the incomplete data, EG09 is promising. EG09 (start position at 474565, stop position at 474305) has an *R*-score at 0.9, which is below the cut-off at 2, has a clear SD sequence (GAGGAGT) 11 bases upstream of the start codon and is RT-PCR positive, suggesting transcription across the reading frame (data not shown). In addition, EG09 is overlapping with Ct413 in phase -2, which is the same phase as the 15 novel overlapping genes. Additional work is needed for this putative gene.

In the gene prediction, 42 genes from the primary annotation are missing. Seven of these genes are assigned a function in the primary annotation. Four Additional genes have been assigned functions later (Iliopoulos *et al.*, 2003). The remaining 31 genes are labelled as hypothetical genes, which is a large fraction compared with the whole-genome annotation. The 42 genes not predicted by EasyGene were in general short (mean 86.2 aa, SD 32.1) compared with the genes uniquely found by EasyGene (mean 127 aa, SD 38.5) or to all annotated Ct-SvD genes (mean 348.9 aa, SD 244.4). The majority of the 42 nonpredicted genes have homologues annotated in at least one of the other *Chlamydiaceae* genomes. This in itself cannot be considered a validation of the protein as errors in annotation tend to propagate (Doerks *et al.*, 1998; Galperin & Koonin, 1998). When aligning the 42 nonpredicted genes against the well-curated Swiss-Prot database (Version 50.4) using BLASTP, the result was matches for nine of the 42 genes. Among the nine genes are all the genes annotated as ribosomal proteins (Ct150, Ct786, Ct802, Ct810) as well as four other genes with assigned function (Ct080, Ct377, Ct444, Ct473). EasyGene is, to a large extent, dependent on the database used for generation of the training set. In this work, version 41 of Swiss-Prot was used for the EasyGene prediction. When rerunning the EasyGene prediction for confirmation purposes on the Ct-SvD genome sequence using a recent version of Swiss-Prot (Version 50.4), EasyGene predicts 8 additional genes that were previously left out (data not shown). These eight genes were all included in the set of nine genes found by BLASTP (Ct150, Ct786, Ct802, Ct810, Ct080, Ct377, Ct444, Ct473). The addition of the eight genes was most likely the result of a larger, update SwissProt database, now containing more genes. Therefore, the prediction will most likely change further as the databases develop and more experimental data become available. Whether the remaining 34 genes are false negatives or not real genes cannot be determined based on the present data. However, conclusions made elsewhere (Kyrpides & Ouzounis, 1999) suggest that the Ct-SvD genome may be overannotated.

When evaluating the output from the *C. trachomatis* gene prediction, the 15 novel genes in the otherwise densely packed Ct-SvD genome were somewhat surprising. In support of the validity of the gene prediction was the inclusion of EG14 in the set of novel predicted genes. EG14 is identical to rbp_7, a gene not included in the primary annotation, but discovered later by mass spectrometry (Shaw *et al.*, 2002). Identification of entirely new genes is frequent along with the development of new gene prediction tools (Bocs *et al.*, 2002) and a given gene prediction is always a current interpretation of a genomic sequence. These findings emphasize the concept that new gene prediction algorithms may contribute to new information.

The seven putative genes, which are supported by additional predictive and transcriptional parameters, are all overlapping with genes on the complementary strand. These complementary genes are all included in the primary annotation as well as in the EasyGene prediction. Four of the overlapping regions are partial, where each overlapping

gene has nonoverlapping regions. The remaining 3 putative genes have overlapping regions that are complete, with one of the genes being contained completely within the boundaries of the other (see Fig. 2). Five of the putative genes (EG02, EG04, EG12, EG13 and EG14) have RT-PCR-positive opposite-strand overlapping genes (Ct046, Ct050, Ct579, Ct743 and Ct804, respectively). The remaining two (EG11 and EG16) overlap with genes that are RT-PCR negative in our assay (Ct456 and Ct864). For Ct456, the negative RT-PCR most likely indicates a false-negative result as this protein is well characterized (Clifton *et al.*, 2004).

The concept of overlapping genes is not uncommon in prokaryotes, and overlapping genes may be the result of evolutionary pressure to minimize genome size (Sakharkar *et al.*, 2005). Overlapping transcripts may as well be involved in *cis*-encoded antisense RNA regulation (Storz *et al.*, 2005). Overlapping genes are described in Chlamydia both in the plasmid and on the chromosome. In the plasmid (Fahr *et al.*, 1992), anti-sense transcripts are most likely involved in RNA-mediated regulation of ORF8. In the chromosome, the primary annotation predicts both same-strand and opposite-strand overlaps. Same-strand overlaps are the most common (supplementary Table S2). The primary annotation predicts 136 same-strand overlaps, with 82 of them longer than 4 bp and same-strand overlaps are a common general finding in bacterial genomes. Opposite-strand overlaps are less frequent. We counted 36 overlaps on opposite strands in the Ct-SvD genome based on the primary annotation, with 22 of them longer than 4 bp. Some of the opposite-strand overlapping regions in the Ct-SvD chromosome can be relatively large, for instance the 62 bp overlapping 3′-ends of Ct473 and Ct474. However, none of the existing overlapping regions included in the primary annotation matches the long overlapping regions predicted in the present work. Even though the novel EasyGene predicted genes are predicted with *R*-scores convincingly below the *R*-value cutoff, the genes were not included in earlier gene prediction work carried out on Ct-SvD. The reason for this may be the fact that overlapping genes are actively avoided in gene-prediction. The in-phase ORF on the complementary strand (the 0-phase) will often be an ORF and therefore appears as a gene, a so-called shadow (Merino *et al.*, 1994; Silke, 1997). ORFs overlapping existing genes on the complementary strand in phase -1 and phase -2 are less frequent and will likely be short (Silke, 1997). The tendency of 0-phase codons to form ORFs was one of the reasons why early gene prediction tools included only the best prediction when dealing with overlapping ORFs, simply removing the least optimal prediction of the overlapping genes in a postprocessing step.

Another observation made from the present data is the convincing tendency for the 7 overlapping putative genes to be in phase -2 (EG02, EG04, EG11, EG12, EG13, EG14, EG16). None of the putative genes has overlapping regions in phase 0 or -1. The same tendency for genes with opposite-strand overlaps to be in phase -2 is observed for the complete genome where more than 50% of the opposite-strand overlaps are in phase -2. This clear bias cannot be explained by preferences or bias within the gene-prediction algorithm. An explanation may be found in the codon degeneracy, which causes the selective independence of an overlapping gene to depend on the phase of the overlap. In silico predictions based on the information content shared between 2 overlapping reading frames suggest that phase -2 is more abundant than phase -1 overlaps (Krakauer, 2000). The -1 phase is the phase under the most restricted evolutionary constraints, because the third codon positions of both genes are complementary so that a nonsynonymous mutation in one gene most likely will also affect the complementary gene nonsynonymously. Phase -2 is the overlap phase with the least evolutionary constraints, as codon-position 3 on one strand is complementary with codon-position 2 on the opposite strand, which is the most conserved of the codon-positions. Mutations on one strand of phase -2 overlaps will therefore for several possible base-substitutions leave the complementary strand synonymously unaffected.

The gene-pairs involved in phase -1 overlaps are therefore selectively dependent, while gene-pairs involved in phase -2 overlaps are selectively less dependent. Gene-pairs involved in phase -2 are therefore more likely to appear and persist by coevolution. EG14 and EG16 only share short overlapping regions with their complementary gene and the evolutionary constraints imposed from the overlap are probably limited. EG02, EG04, EG11, EG12 and EG13 share longer overlapping regions and these genes may have coevolved with their complementary genes. The concept of the shared information content suggests that unnecessary genes involved in phase -1 overlaps are more likely to persist as remnants of functionless genes. This is more unlikely to occur for genes involved in phase -2 overlaps. When gene pairs overlap in phase -2, it therefore suggests a necessary function for both genes. That all the individual novel predicted genes overlap their complementary genes in phase -2, in several situations involving major or complete parts of one of the involved genes, therefore suggests that both genes are under positive selection.

Several of the novel predicted genes might be so-called singletons, which are sequences with no detectable similiarity to sequences in any database and therefore represents a gene sequence unique to the organism. The number of singletons larger than 150 nt was initially quite high in Ct-SvD (23% of total ORFs), but decreased along with the development of the sequence databases. On re-evaluating the Ct-SvD genome against a larger database containing 60 microbial genomes, the number of singletons was reduced

(2.6% of total ORFs) (Siew & Fischer, 2003). It is likely that completion of additional microbial genomes will reveal matches for the novel predicted genes.

The family Chlamydiaceae is comprised of the two genera Chlamydia and Chlamydophila. Of the 15 novel ORFs, only EG08 has homology to a sequence in the Cp genome. This sequence is not flanked by in-frame start and stop codons and thus cannot be an actively transcribed gene in Cp. Cp belongs to the genus *Chlamydophila,* while Ct-SvD, Ct-SvA and Cm belong to the *Chlamydia* genus (Bush & Everett, 2001), The present data therefore suggest that the absence of the majority of the putative genes in the Cp genome reflects selective adaptation to a specific environment (Bush & Everett, 2001). The overlapping genes may therefore contribute to the different biology displayed by the species. Alternatively, the novel putative genes could have emerged after the splitting of the two genera, for example by horizontal gene transfer, which is evident in Chlamydia (Ortutay *et al.*, 2003).

Acknowledgements

Kathryn Wattam and Lars G.T. Jørgensen are thanked for technical assistance. Claus S. Aagaard and Frank Follmann are thanked for scientific discussions and for carefully reading through the manuscript. P.I. was supported by the Danish National Research Foundation and EU grant no. QLRI-CT-2001-00015.

References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bateman A, Coin L, Durbin R *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res* **32**: D138–D141.

Belland RJ, Zhong G, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL & Caldwell HD (2003) Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. *Proc Natl Acad Sci USA* **100**: 8478–8483.

Bendtsen JD, Nielsen H, von Heijne G & Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**: 783–795.

Bocs S, Danchin A & Medigue C (2002) Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes. *BMC Bioinformatics* **3**: 5.

Boeckmann B, Bairoch A, Apweiler R *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**: 365–370.

Boldogkoi Z & Barta E (1999) Specific amino acid content and codon usage account for the existence of overlapping ORFS. *Biosystems* **51**: 95–100.

Borodovsky M & McIninch J (1993) GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry* **17**: 123–133.

Bush RM & Everett KD (2001) Molecular evolution of the *Chlamydiaceae*. *Int J Syst Evol Microbiol* **51**: 203–220.

Clifton DR, Fields KA, Grieshaber SS, Dooley CA, Fischer ER, Mead DJ, Carabeo RA & Hackstadt T (2004) A chlamydial type III translocated protein is tyrosine-phosphorylated at the site of entry and associated with recruitment of actin. *Proc Natl Acad Sci USA* **101**: 10166–10171.

de Hoon MJ, Makita Y, Nakai K & Miyano S (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* **1**: e25.

Delcher AL, Harmon D, Kasif S, White O & Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–4641.

Doerks T, Bairoch A & Bork P (1998) Protein annotation: detective work for function prediction. *Trends Genet* **14**: 248–250.

Everett KD, Bush RM & Andersen AA (1999) Emended description of the order *Chlamydiales*, proposal of *Parachlamydiaceae* fam. nov. and *Simkaniaceae* fam. nov., each containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. *Int J Syst Bacteriol* **49**: (Part 2): 415–440.

Fahr MJ, Sriprakash KS & Hatch TP (1992) Convergent and overlapping transcripts of the *Chlamydia trachomatis* 7.5-kb plasmid. *Plasmid* **28**: 247–257.

Fickett JW (1996) Finding genes by computer: the state of the art. *Trends Genet* **12**: 316–320.

Frishman D, Mironov A, Mewes HW & Gelfand M (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* **26**: 2941–2947.

Galperin MY & Koonin EV (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol* **1**: 55–67.

Iliopoulos I, Tsoka S, Andrade MA *et al.* (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* **19**: 717–726.

Johnson ZI & Chisholm SW (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Res* **14**: 2268–2272.

Keese PK & Gibbs A (1992) Origins of genes: "big bang" or continuous creation? *Proc Natl Acad Sci USA* **89**: 9489–9493.

Krakauer DC (2000) Stability and evolution of overlapping genes. *Evolution Int J Org Evolution* **54**: 731–739.

Krogh A, Larsson B, von Heijne G & Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

Kyrpides NC & Ouzounis CA (1999) Whole-genome sequence annotation: 'Going wrong with confidence'. *Mol Microbiol* **32**: 886–887.

Larsen TS & Krogh A (2003) EasyGene–a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4**: 21.

Liu Y, Harrison PM, Kunin V & Gerstein M (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* **5**: R64.

Lukashin AV & Borodovsky M (1998) GeneMark hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107–1115.

Ma J, Campbell A & Karlin S (2002) Correlations between Shine–Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* **184**: 5733–5745.

Merino E, Balbas P, Puente JL & Bolivar F (1994) Antisense overlapping open reading frames in genes from bacteria to humans. *Nucleic Acids Res* **22**: 1903–1908.

Mira A & Pushker R (2005) The silencing of pseudogenes. *Mol Biol Evol* **22**: 2135–2138.

Mount D (2004) Sequence Database Searching for Similar Sequences. *Bioinformatics: Sequence and Genome Analysis*. (Mount D, ed.). pp. 227–280. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Nielsen P & Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21**: 4322–4329.

Ortutay C, Gaspari Z, Toth G, Jager E, Vida G, Orosz L & Vellai T (2003) Speciation in Chlamydia: genomewide phylogenetic analyses identified a reliable set of acquired genes. *J Mol Evol* **57**: 672–680.

Osada Y, Saito R & Tomita M (1999) Analysis of base-pairing potentials between 16S rRNA and 5′ UTR for translation initiation in various prokaryotes. *Bioinformatics* **15**: 578–581.

Read TD, Brunham RC, Shen C *et al.* (2000) Genome Sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. *Nucleic Acids Res* **28**: 1397–1406.

Rozen S & Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365–386.

Saito R & Tomita M (1999) Computer analyses of complete genomes suggest that some archaebacteria employ both eukaryotic and eubacterial mechanisms in translation initiation. *Gene* **238**: 79–83.

Sakharkar KR, Dhar PK & Chow VT (2004) Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int J Syst Evol Microbiol* **54**: 1937–1941.

Sakharkar KR, Sakharkar MK, Verma C & Chow VT (2005) Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int J Syst Evol Microbiol* **55**: 1205–1209.

Salgado H, Moreno-Hagelsieb G, Smith TF & Collado-Vides J (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA* **97**: 6652–6657.

Salzberg SL, Delcher AL, Kasif S & White O (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**: 544–548.

Schachter J & Wyrick PB (1994) Culture and isolation of *Chlamydia trachomatis*. *Methods Enzymol* **236**: 377–390.

Schurr T, Nadir E & Margalit H (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res* **21**: 4019–4023.

Shaw AC, Larsen MR, Roepstorff P, Christiansen G & Birkelund S (2002) Identification and characterization of a novel *Chlamydia trachomatis* reticulate body protein. *FEMS Microbiol Lett* **212**: 193–202.

Siew N & Fischer D (2003) Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* **53**: 241–251.

Silke J (1997) The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage. *Gene* **194**: 143–155.

Skovgaard M, Jensen LJ, Brunak S, Ussery D & Krogh A (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* **17**: 425–428.

Stephens RS, Kalman S, Lammel C *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**: 754–759.

Storz G, Altuvia S & Wassarman KM (2005) An abundance of RNA regulators. *Annu Rev Biochem* **74**: 199–217.

Wootton J & Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computational Chemistry* **17**: 149–163.

Supplementary material

The following supplementary material is available for this article online

Table S1. The open reading frames not included in the EasyGene annotation.

Table S2. Overlap lengths and phases of same-strand and opposite-strand overlaps in *Chlamydia trachomatis*.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1574-6968.2006.00480.x (This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.