

QUT Digital Repository:
<http://eprints.qut.edu.au/>



Grech, Brian J. and Maetschke, Stefan and Mathews, Sarah A. and Timms, Peter (2007) Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint. *Research in Microbiology* 158(8-9):pp. 685-693.

© Copyright 2007 Elsevier

1 Genome wide analysis of chlamydiae for promoters that phylogenetically footprint

2

3 Brian Grech^a, Stefan Maetschke^b, Sarah Mathews^a, Peter Timms^{a*}

4

5 ^aInstitute of Health and Biomedical Innovation, Queensland University of Technology,

6 Corner Musk Avenue and Blamey Street, Kelvin Grove, Brisbane, Queensland, 4059,

7 Australia

8

9 ^bFaculty of Information Technology, Queensland University of Technology, 126

10 Margaret Street, Brisbane, Queensland, 4001, Australia

11

12 Running title: Phylogenetic footprinting chlamydial promoters

13

14 Email: p.timms@qut.edu.au *Correspondence and reprints

15 Tel: +61 7 3138 6199

16 Fax: +61 7 3138 6030

17 Abstract

18 Currently there is a lack of phylogenetic footprinting programmes that can take
19 advantage of multiple whole genome sequences of different species within the same
20 bacterial genus. Therefore, we have developed and tested a position weight matrix
21 based programme called *Footy*, that performs genome wide analysis of bacterial
22 genomes for promoters that phylogenetically footprint. When *Footy* was used to
23 analyse the non-coding regions upstream of genes from three chlamyidal species for
24 promoters that phylogenetically footprint, it predicted a total of 42 promoters, of which
25 41 were new. Ten of the 41 new promoters predicted by *Footy* were biologically
26 assayed in *Chlamydia trachomatis* by mapping the 5' end of the transcripts for the
27 associated genes. The primer extension assay validated seven of the 10 promoters.
28 When *Footy* was compared to two other accepted methods for genome wide
29 prediction of promoters in bacteria (the *standard PWM method* and MITRA), *Footy*
30 performed equally as well or better than these programmes. This paper, therefore,
31 shows the value of a bioinformatics programme able to perform genome wide
32 analysis of bacteria for promoters that phylogenetically footprint.

33

34 Key Words: bioinformatics, *Chlamydia*, phylogenetic footprinting, promoter

35 1. Introduction

36 The post-genomic era has generated an interest in developing theoretical
37 approaches to discover as much information as possible from the available genome
38 sequences [16]. One area of intense research is modelling and predicting bacterial
39 promoters. However, the structure of bacterial promoters makes it difficult to devise a
40 general prediction algorithm [8, 10, 11]. Most of the currently available programmes
41 for the prediction of bacterial promoters exhibit poor specificity, generating many false
42 positive predictions. One approach to filter out false positives predicted by the current
43 methods is phylogenetic footprinting [8].

44 Phylogenetic footprinting is a computational method for predicting homologous
45 promoters in the equivalent non-coding regions (NCRs) upstream of a gene family (or
46 genes with a common function), from evolutionarily related species [6]. (NCRs
47 upstream of genes will be referred to as *upstream regions* in this paper.)
48 Phylogenetic footprinting predicts promoters by assuming that: (1) the upstream
49 regions of homologous genes from different species are regulated by homologous
50 promoters, and (2) spacers between promoters within the equivalent upstream
51 regions are free from evolutionary constraints. Therefore, substitution of another base
52 within the spacer can be accepted at any position, whereas the homologous
53 promoters can only accept certain substitutions, as they must be recognizable to the
54 cognate σ factor or transcription factor.

55 The increasing number of complete bacterial genome sequences now allows
56 genome wide analysis for promoters that phylogenetically footprint. This is because it
57 is possible to find a dataset of equivalent upstream regions from two or more
58 bacterial species separated by an appropriate evolutionary distance for phylogenetic

59 footprinting of promoters [6, 7, 34]. The appropriate evolutionary distance is observed
60 within a particular evolutionary time frame. For example, a dataset of equivalent
61 upstream regions from too *closely related species* will produce a high rate of false
62 positives, when phylogenetically footprinting promoters. Since the spacers
63 surrounding the promoters have not had enough evolutionary time to mutate, so that
64 the conservation of the spacers across the different species is poorer when
65 compared to the conservation of the promoters across the same species. At
66 evolutionary distances that are too great, only *well conserved* promoters, upstream of
67 well conserved genes can be phylogenetically footprinted. The appropriate
68 evolutionary distance is when there is an observable difference between the
69 conservation of the spacers and the conservation of the promoters across the
70 different species [6, 34].

71 Bacteria of the genus *Chlamydia* are ideal organisms for finding a species set at
72 an appropriate evolutionary distance for phylogenetically footprinting promoters. This
73 is because chlamydiae are phylogenetically isolated, which has resulted in a high
74 level of conservation of genes and gene order between the species [1, 22] and
75 *Chlamydia* is one of the most sequenced organisms with 10 genome sequences
76 available for six species [1, 4, 9, 13, 21, 22, 26, 27, 32]. Consequently, many of the
77 promoters would be expected to be well conserved across the different chlamydial
78 species and the probability of finding these well conserved promoters would be high.

79 Chlamydial σ^{66} promoters are probably the best choice of promoters to be
80 identified by phylogenetic footprinting. This is because σ^{66} promoters have a greater
81 likelihood to be found upstream of the majority of chlamydial operons. (Since, the σ^{66}
82 factor (RpoD) of *Chlamydia*, is the principal σ factor [14, 18].) Promoter mutagenesis
83 and *in vitro* transcription assays has shown that σ^{66} have the greatest affinity for two

84 motifs that are identical to the -35 and -10 hexamers of the *Escherichia coli* σ^{70}
85 consensus sequence and are, therefore, σ^{70} -like [30].

86 The accepted method for predicting *E. coli* σ^{70} promoters is to use a pair of
87 position weight matrices (PWMs) to predict the -35 and -10 hexamers [28]. A PWM
88 is a two dimensional array of values representing the information content (IC) of a
89 motif. The IC is a measure of the bit rate, i.e. bits per base. However, a phylogenetic
90 footprinting algorithm that uses PWMs to analyse the upstream regions of multiple
91 whole bacterial genomes for promoters that phylogenetically footprint has not been
92 published.

93 The work presented in this paper has developed and tested, a bioinformatic
94 programme that can perform genome wide analysis of bacteria for promoters that
95 phylogenetically footprint. This programme is called *Footy* and is based on the
96 *standard PWM method*, with an extension that can analyse multiple bacterial
97 genomes for phylogenetically conserved promoters. When Footy was applied to the
98 genomes of *Chlamydia trachomatis*, *Chlamydia pneumoniae* and *Chlamydia caviae*,
99 42 σ^{66} promoters were predicted, of which 41 were new.

100 **2. Material and methods**

101 *2.1. Sequence data*

102 The plus strand of whole genomes of *C. trachomatis* serovar D, *C. pneumoniae*
103 strain AR39, *C. caviae* biovar GPIC and *Chlamydia muridarum* biovar MoPn, and the
104 corresponding annotation table were downloaded from GenBank
105 (<http://www.ncbi.nlm.nih.gov>; accession no.: AE001273, AE002160, AE002161 and
106 AE015925; respectively). The coordinates used for each predicted gene and
107 *structural RNAs* (rRNAs and tRNAs) were based on the annotated start and stop
108 positions [21, 22, 27]. Two datasets of regions were generated from the four
109 genomes: (1) upstream regions and (2) *downstream regions* (or NCRs between
110 convergently transcribed genes). The 3' ends of the upstream regions were reduced
111 by 10 bp from the start of the associate gene. The maximum length of each upstream
112 region was limited to 390 bp and the minimum length was restricted to 35 bp. The
113 maximum length of each downstream region was not limited, but the minimum length
114 was restricted to 35 bp.

115 *2.2. Identification of equivalent upstream and downstream regions*

116 A *Table of predicted homologous genes* (TOPHG) was constructed for *C.*
117 *trachomatis*, *C. muridarum*, *C. pneumoniae* and *C. caviae* using prototype tables of
118 homologous genes calculated by “TIGR Comprehensive Microbial Resource Total
119 Protein Hit” search engine (<http://www.tigr.org/tigr-scripts/CMR2/>). The parameters
120 used for BLAST analysis were as follows: similarity $\geq 40.0\%$, identity $\geq 10.0\%$ and P-
121 value ≤ 0.05 . For duplicate entries with the same gene names, the set of homologous
122 genes with the lowest P-value was selected. The sets of structural RNAs were

123 identified by comparing the location of the genes downstream of the structural RNAs
124 in *C. trachomatis* with their homologs in *C. muridarum*, *C. pneumoniae* and *C. caviae*.
125 Candidate upstream and downstream regions were determined to be equivalent if
126 their associated genes were predicted to be homologous, using the above
127 parameters.

128 2.3. Footy

129 A flow chart of the Footy algorithm is shown in Fig. 1. The promoter model
130 consisting of two PWMs and variable spacer was calculated. The PWMs were
131 derived from an alignment of the –35 and –10 hexamers of 300 *E. coli* σ^{70} promoters
132 taken from Lissner and Margalit [17]. The weights of the PWMs were calculated using
133 equations (Equations S1, S2 and S3, supplementary material) based on the
134 equations developed by Stormo and Hartzell [29]. The first stage of Footy scans the
135 upstream regions of the chlamydial genomes for patterns that are similar to the
136 model. The first stage is the same as the standard PWM method. The second stage
137 of Footy phylogenetically footprints promoters with homologous promoters in the
138 other chlamydial species. Footy with instruction on usage is available at
139 <http://eresearch.fit.qut.edu.au/Footy/>.

140 2.4. Reduction of false positives

141 The second stage of Footy eliminates many of the promoters predicted that did
142 not phylogenetically footprint by aligning the predicted –35 and –10 hexamers in *C.*
143 *trachomatis* with the predicted hexamers for the homologous genes (where available
144 based on the TOPHG) in the other chlamydial species (Fig. 2). To do this Footy
145 performed un-gapped pair-wise alignments (PWAs) of the predicted hexamers.
146 These regions were aligned at the first base of each predicted hexamer. The

147 predicted hexamers were reported as *conserved* if the number of mismatches
148 between the reference species and the other species were equal to or less than the
149 pre-set mismatch threshold. If all of the PWAs reported conserved hexamers equal to
150 or less than the mismatch threshold, then the hexamers were reported as well
151 conserved (Fig. 2A). Once this process was completed, the next set of equivalent
152 upstream regions were analysed (Fig. 2B).

153 To further filter out false positives, the multiple sequence alignments (MSAs)
154 calculated by Footy were inspected to decide which, if any, of the predicted well
155 conserved promoters could be eliminated. The position of each predicted promoter
156 with respect to the start site of the associated gene was compared between species.
157 If the distances varied more than 200 bp between different species, the promoters
158 were eliminated. If multiple promoters were predicted in the same upstream regions,
159 the highest bit scoring and lowest mismatch MSA of conserved promoters was
160 selected.

161

162 *2.5. Validation of a subset of predicted promoters*

163 To validate the predicted promoters, a subset were chosen to have the 5' end of
164 the RNA of the associated genes mapped by primer extension in *C. trachomatis*
165 serovar L2/434/Bu (Table S1) [19]. The predicted promoter was considered to be
166 correct if the spacer between the -10 hexamer and the mapped 5' end of the RNA
167 was from 4 to 12 bp [12].

168 3. Results

169 3.1. Footy predicted 42 promoters that were phylogenetically conserved in *Chlamydia*

170 To determine the number of species chosen at an appropriate evolutionary
171 distance, IC threshold and mismatch threshold, a dataset of equivalent downstream
172 regions of *Chlamydia* was analysed for false positives using a σ^{70} promoter model.
173 The analysis of the downstream regions revealed a lack of false positives (data not
174 shown); therefore, the downstream regions could not be used to provide an
175 appropriate negative control. While performing this analysis it became clear that a 16
176 to 18 bp spacer between the two hexamers rather than a 15 to 21 bp spacer, which is
177 the allowable spacer length for *E. coli* σ^{70} promoters [17], substantially reduced the
178 number of false positives predicted (data not shown).

179 The number of species chosen, IC threshold and mismatch threshold were set so
180 that the maximum number of promoters were predicted and no false positives were
181 detected in the upstream regions of a dataset of 15 positive controls. The 15 positive
182 controls chosen were, the *C. trachomatis* 16S rDNA P1, CT602, *hctA*, *infA*, *ItuA*, *ItuB*,
183 *omcA*, *ompA* P1 and P2, *pkn5*, *rpoD*, *secA*, *sctU*, *srp* and tRNAThr2 (B.J. Grech
184 unpublished data), [20]. These σ^{66} promoters were determined to be a suitable set of
185 positive control promoters for setting parameters and testing the performance of
186 Footy, because they were located in NCRs on the *C. trachomatis* chromosome. The
187 analysis revealed that using the species of *C. trachomatis*, *C. pneumoniae* and *C.*
188 *caviae*; with an IC threshold of 3.0 bits (or 1.5 bits per hexamer) and a mismatch
189 threshold of two resulted in the maximum number of promoters predicted with none of
190 the 15 positive controls reported as false positives. Table 1 shows the number of

191 false positives for different combination of species, IC thresholds and mismatch
192 thresholds.

193 Using these parameters and after applying the rules discussed in the Materials
194 and methods, Footy predicted 42 promoters that were conserved in the dataset of
195 305 equivalent upstream regions extracted from the three chlamydiae. One of the 15
196 positive control promoters the promoter of *infA* was predicted correctly by Footy.
197 Footy did, however, predict three new promoters in the upstream regions of *euo*,
198 *groES* and *rpsA*, with homologs that have been biologically confirmed in chlamydial
199 species not analysed in this study. The -35 and -10 hexamers of the predicted
200 promoters of *groES* and *rpsA* were 100% conserved with the -35 and -10 hexamers
201 of the biological confirmed promoters of *groES* and *rpsA* of *C. muridarum* [31]. The
202 predicted promoter of *euo* was 4 bp upstream of the nucleotide in *C. trachomatis* that
203 corresponds to the 5' end of the *euo* P1 transcript, mapped in *C. psittaci* 6BC by
204 Wichlan and Hatch [35] (Table 2).

205 An analysis was conducted to determine why 14 of the 15 positive controls were
206 missed by Footy. The *C. trachomatis*, *C. pneumoniae* and *C. caviae* genomes were
207 visually inspected for patterns similar to the 14 false negatives. Analysis of the
208 equivalent regions in *C. pneumoniae* and *C. caviae* for patterns similar to the 14 false
209 negatives, identified patterns with no more than two mismatches from the promoters
210 of *C. trachomatis* CT602, *hctA*, *omcA*, *rpoD* and *sctU*. The promoters CT602, *hctA*,
211 *rpoD* and *sctU* were missed because they were below the (3.0 bit) IC threshold and
212 the promoter for *omcA* was missed because the analogous pattern in the equivalent
213 upstream region of *C. pneumoniae* was located within the open reading frames
214 (ORFs) (Table 3).

215 The total number of mismatches that the –35 and –10 hexamers of each of the
216 42 promoters had from the σ^{70} consensus sequence (TTGACA and TATAAT, for –35
217 and –10, respectively) were determined. The number of mismatches ranged from
218 zero to five out of a possible 12, with a statistical mode of four mismatches. Fifteen
219 (35%) of the promoters had four mismatches, 34 (80%) of the promoters had three to
220 five mismatches and 41 (98%) of the promoters had one to six mismatches from the
221 σ^{70} consensus sequence.

222 *3.2. Transcription start site mapping confirms an additional seven promoters* 223 *predicted by Footy*

224 Since only one of the 42 predicted promoters predicted by Footy had been
225 biological confirmed, more experimental data was needed to assess the performance
226 of Footy. Therefore, 10 of the of the 42 promoters predicted by Footy, the promoters
227 of *tyrS*, *gcp1*, *clpC*, *rs12*, CT547, *sctJ*, *exbB*, *snf*, *greA* and *elp2*, were chosen for
228 primer extension. The 10 genes were selected since they are highly expressed at 24
229 h post infection by micro-array analysis [2], thus ensuring gene specific RNA would
230 be isolated.

231 The 5' end for *tyrS*, *clpC*, *rs12*, *sctJ*, *exbB*, *snf* and *elp2* transcripts correctly
232 mapped to the promoters predicted by Footy (Table 2 and Fig. S1, supplementary
233 material), hence confirming an additional seven promoters. The promoters of the
234 remaining three genes (*gcp1*, CT547 and *greA*) were unable to be confirmed,
235 because the 5' end of RNA could not be mapped or was mapped elsewhere (data not
236 shown). Since *C. trachomatis* was grown in HEp2 monolayers there was the
237 possibility of non-specific binding of the gene-specific primers to HEp2 RNA.

238 Therefore, primer extension was also performed on RNA extracted from uninfected
239 HEp2 cell lines and all results were negative (data not shown).

240 The predicted -35 and -10 hexamers of the seven newly confirmed promoters
241 (*tyrS*, *clpC*, *rs12*, *sctJ*, *exbB*, *snf* and *elp2*) were analysed and compared to *E. coli* σ^{70}
242 promoters. The bit scores for both hexamers ranged from 4.4 to 9.6 bits, out of a
243 range of 3.0 to 14.4 bits and the number of mismatches for both hexamers across the
244 three chlamydiae ranged from zero to the maximum mismatch threshold of two. The
245 promoters of *elp2*, *rs12* and *sctJ* had nine nucleotides; the promoters of *snf* and *tyrS*
246 had eight nucleotides and the promoters of *clpC* and *exbB* had seven nucleotides
247 identical to the σ^{70} consensus sequence. Since σ^{70} promoters can have as few as
248 five nucleotides identical to the consensus sequence [17], the seven newly identified
249 promoters are σ^{70} -like and classified as σ^{66} promoters.

250 2.3. Footy performs better than the Standard PWM method on *C. trachomatis*

251 To compare Footy to the standard PWM method [24], the upstream regions of *C.*
252 *trachomatis* were analysed for promoters similar to *E. coli* σ^{70} promoters using the
253 standard PWM method. The promoters and equations used to calculate the PWMs
254 were the same as Footy and the model had a spacer length of 15 to 19 bp. The IC
255 threshold of 10.0 bits (or 5.0 bits per hexamer) was determined by analysing the 536
256 upstream regions and the 122 downstream regions of *C. trachomatis* for promoters
257 and false positives, respectively, and by determining the statistical significance of the
258 promoters predicted in the upstream regions. The statistical significance of the
259 promoters predicted was determined using the χ^2 test on two contingency tables, one
260 corresponding to the predictions in the upstream regions and the other the false
261 positives in the downstream regions. At an IC threshold of 10.0 bits, 10 promoters

262 were predicted in the upstream regions (Table S2, supplementary material). Of the 15
263 positive control promoters (described above) one was predicted correctly by the
264 standard PWM method, the promoter of *infA*.

265 The 10 promoters predicted by the standard PWM method all showed high
266 homology to the *E. coli* σ^{70} consensus sequence [17]. The number of mismatches
267 that each predicted promoter deviated from the σ^{70} consensus sequence ranged from
268 zero to two, with a statistical mode of two mismatches. The spacer between the -35
269 and -10 hexamers ranged from 16 to 18 bp, with a mode of 17 bp.

270 4. Discussion

271 Footy predicted 42 promoters that phylogenetically footprinted across three
272 species of *Chlamydia*. A computational method that is capable of genome wide
273 phylogenetic footprinting of promoters across the multiple genome sequences of a
274 bacterial genus has not been previously reported; thus demonstrating the usefulness
275 of Footy.

276 Comparison of Footy to the standard PWM method shows that Footy performs
277 better when analysing *Chlamydia* for σ^{66} promoters. The 42 promoters predicted by
278 Footy contained up to five mismatches from the *E. coli* σ^{70} consensus sequence,
279 whereas the standard PWM method predicted 10 promoters with up to two
280 mismatches from the σ^{70} consensus sequence (Table 4). Footy used an IC threshold
281 that was 7.0 bits lower than the IC threshold used by the standard PWM method in
282 this study. This increases the likelihood of finding promoters.

283 The standard PWM method did however predict three promoters (CT016, CT763
284 and *glyQ*) not predicted by Footy. This is because the standard PWM method keeps
285 promoters that either do not have a homolog in the equivalent upstream regions of *C.*

286 *pneumoniae* or *C. cavaie*, or where the equivalent upstream regions in *C.*
287 *pneumoniae* or *C. cavaie* cannot be identified. Consequently, the standard PWM
288 method analysed 231 more upstream regions of *C. trachomatis* than Footy.
289 Therefore, the standard PWM method shows better sensitivity than Footy at IC
290 thresholds above 10 bits when analysing *C. trachomatis* for σ^{66} promoters, hence
291 Footy will not replace the standard PWM based programmes, such as “ScanACE”
292 (<http://arep.med.harvard.edu/mrnadata/mrnasoft.html>), [3] and “patser” [33].
293 However, Footy will be very effective when used in conjunction with these
294 programmes to predict more promoters within an organism.

295 Eskin and colleagues [5] analysed 120 and 136 regions between divergently
296 transcribed genes of *C. muridarum* and *C. pneumoniae*, respectively, for statistically
297 significant over-represented patterns homologous to the *E. coli* σ^{70} consensus
298 sequence. The analysis used a promoter model of two hexamers separated by a 3 to
299 23 bp spacer, using the programme, MITRA. The authors reported that MITRA was
300 unable to extract an over-represented pattern from the upstream regions of *C.*
301 *muridarum* that was homologous to the σ^{70} consensus sequence. When MITRA was
302 applied to *C. pneumoniae* it discovered the over-represented pattern, TTGACA N₁₉
303 ATAATT, which was made up of 27 hits. If the –10 hexamer of this over-represented
304 pattern is shifted 1 bp upstream, this pattern is identical at 11 of the 12 positions to
305 the σ^{70} consensus sequence. Therefore, MITRA predicted 27 σ^{66} promoters in *C.*
306 *pneumoniae*, some of which contained up to three mismatches from the σ^{70}
307 consensus sequence. Interestingly, only one of these promoters (designated as
308 RCPX0664_RCPX0066, [5]), was also predicted by Footy (CP0079) (Table 4).

309 The analysis of *Chlamydia* with the standard PWM method, MITRA and Footy
310 show that by analysing equivalent upstream regions from multiple species for

311 promoters that phylogenetically footprint, more promoters were predicted and some
312 had fewer matches to the σ^{70} consensus sequence.

313 MITRA predicted promoters not predicted by Footy, because it is able to predict
314 promoters without homologs in the other chlamydial species. Therefore, the
315 performance of MITRA will not suffer if the promoters or associated genes are not
316 conserved across the different bacterial genomes analysed. Pattern discovery
317 programmes such as MITRA could also be used in conjunction with Footy to increase
318 the number of promoters predicted within an organism.

319 Given the hypothesis, that chlamydial promoters are expected to be well
320 conserved in the equivalent upstream regions, the low number of promoters predicted
321 as phylogenetically conserved across chlamydial species by Footy is surprising.
322 Possible explanations are: (1) that homologous promoters may have been eliminated
323 from the dataset of equivalent upstream regions because of promoters in one or more
324 species overlapped or are located within the coding regions of genes, (2) there is a
325 low level of conserved σ^{66} promoters across different species of *Chlamydia*, and/or
326 (3) that many of the σ^{66} promoters are dissimilar to the σ^{70} consensus sequence and
327 therefore below the detection levels of Footy.

328 The major outcome of this study is the development of a new programme that
329 predicts conserved promoters on a genome wide scale across multiple bacteria, while
330 performing equally as well or better than the current methods. For example, when
331 analysing *Chlamydia*, Footy predicted more promoters with some having fewer
332 matches to the *E. coli* σ^{70} consensus sequence and maintained a level of sensitivity
333 and specificity comparable with other promoter prediction programmes. Finally, the
334 increased number of σ^{66} promoters predicted by Footy in *Chlamydia* will be of
335 significant value to researchers studying this organism.

336 **Acknowledgments**

337 We gratefully acknowledge Karl Eisler and Melinda Ziino from the Australian
338 Genome Research Facility, (Melbourne, Australia) for performing the fragment
339 analysis on cDNA; Anthony Rasmussen from the High Performance Computing and
340 Research Support, (Queensland University of Technology, Brisbane, Australia) for his
341 technical assistance and Michael Towsey from the Faculty of Information Technology
342 Innovation, (Queensland University of Technology, Brisbane, Australia) for his
343 technical assistance.

344 **References**

- 345 [1] Azuma, Y., Hirakawa, H., Yamashita, A., Cai, Y., Rahman, M.A., Suzuki, H.,
346 Mitaku, S., Toh, H., Goto, S., Murakami, T., Sugi, K., Hayashi, H., Fukushi, H.,
347 Hattori, M., Kuhara, S., Shirai, M., (2006) Genome sequence of the cat
348 pathogen, *Chlamydophila felis*, DNA Res., 13, 15-23.
- 349 [2] Belland, R.J., Zhong, G., Crane, D.D., Hogan, D., Sturdevant, D., Sharma, J.,
350 Beatty, W.L., Caldwell, H.D., (2003) Genomic transcriptional profiling of the
351 developmental cycle of *Chlamydia trachomatis*, Proc Natl Acad Sci U S A,
352 100, 8478-83.
- 353 [3] Berg, O.G., von Hippel, P.H., (1987) Selection of DNA binding sites by
354 regulatory proteins. Statistical-mechanical theory and application to operators
355 and promoters, J Mol Biol, 193, 723-50.
- 356 [4] Carlson, J.H., Porcella, S.F., McClarty, G., Caldwell, H.D., (2005) Comparative
357 genomic analysis of *Chlamydia trachomatis* oculotropic and genitotropic
358 strains, Infect Immun, 73, 6407-18.
- 359 [5] Eskin, E., Keich, U., Gelfand, M.S., Pevzner, P.A., (2003) Genome-wide
360 analysis of bacterial promoter regions, Pac Symp Biocomput, 29-40.
- 361 [6] Gelfand, M.S., (1999) Recognition of regulatory sites by genomic comparison,
362 Res Microbiol, 150, 755-71.
- 363 [7] Gelfand, M.S., Koonin, E.V., Mironov, A.A., (2000) Prediction of transcription
364 regulatory sites in Archaea by a comparative genomic approach, Nucleic Acids
365 Res, 28 695-705.

- 366 [8] Gelfand, M.S., Novichkov, P.S., Novichkova, E.S., Mironov, A.A., (2000)
367 Comparative analysis of regulatory patterns in bacterial genomes, Brief
368 Bioinform, 1, 357-71.
- 369 [9] Geng, M.M., Schuhmacher, A., Muehldorfer, I., Bensch, K.W., Schaefer, K.P.,
370 Schneider, S., Pohl, T., Essig, A., Marre, R., Melchers, K., (2003) The genome
371 sequence of *Chlamydia pneumoniae* TW183 and comparison with other
372 *Chlamydia* strains based on whole genome sequence analysis, Byk Gulden
373 Pharmaceuticals.
- 374 [10] Gershenzon, N.I., Stormo, G.D., Ioshikhes, I.P., (2005) Computational
375 technique for improvement of the position-weight matrices for the DNA/protein
376 binding sites, Nucleic Acids Res., 33, 2290-301.
- 377 [11] Gordon, J.J., Towsey, M.W., Hogan, J.M., Mathews, S.A., Timms, P., (2006)
378 Improved prediction of bacterial transcription start sites, Bioinformatics., 22,
379 142-8.
- 380 [12] Harley, C.B., Reynolds, R.P., (1987) Analysis of *E. coli* promoter sequences,
381 Nucleic Acids Res, 15, 2343-61.
- 382 [13] Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W.,
383 Olinger, L., Grimwood, J., Davis, R.W. and Stephens, R.S., Comparative
384 genomes of *Chlamydia pneumoniae* and *Chlamydia trachomatis*, Nat Genet,
385 21 (1999) 385-9.
- 386 [14] Koehler, J.E., Burgess, R.R., Thompson, N.E., Stephens, R.S., (1990)
387 *Chlamydia trachomatis* RNA polymerase major sigma subunit. Sequence and
388 structural comparison of conserved and unique regions with *Escherichia coli*
389 sigma-70 and *Bacillus subtilis* sigma-43, J Biol Chem, 265, 13206-14.

- 390 [15] Lambden, P.R., Everson, J.S., Ward, M.E., Clarke, I.N., (1990) Sulfur-rich
391 proteins of *Chlamydia trachomatis*: developmentally regulated transcription of
392 polycistronic mRNA from tandem promoters, *Gene*, 87, 105-12.
- 393 [16] Lee, P.S., Lee, K.H., (2000) Genomic analysis, *Curr Opin Biotechnol.*, 11, 171-
394 5.
- 395 [17] Lisser, S., Margalit, H., (1993) Compilation of *E. coli* mRNA promoter
396 sequences, *Nucleic Acids Res*, 21, 1507-16.
- 397 [18] Lonetto, M., Gribskov, M., Gross, C.A., (1992) The sigma-70 family: sequence
398 conservation and evolutionary relationships, *J Bacteriol*, 174, 3843-9.
- 399 [19] Mathews, S.A., Timms, P., (2000) Identification and mapping of sigma-54
400 promoters in *Chlamydia trachomatis*, *J Bacteriol*, 182, 6239-42.
- 401 [20] Mathews, S.A., Timms, P., (2006) *In Silico* Identification of Chlamydial
402 Promoters and their Role in Regulation and Development. In P.M. Bavoil and
403 P.B. Wyrick (Eds.), *Chlamydia: Genomics and Pathogenesis*, Horizon
404 Bioscience, 133-56.
- 405 [21] Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O.,
406 Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K.,
407 Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M.,
408 Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J.,
409 Fraser, C.M., (2000) Genome sequences of *Chlamydia trachomatis* MoPn and
410 *Chlamydia pneumoniae* AR39, *Nucleic Acids Res*, 28, 1397-406.
- 411 [22] Read, T.D., Myers, G.S., Brunham, R.C., Nelson, W.C., Paulsen, I.T.,
412 Heidelberg, J., Holtzapple, E., Khouri, H., Federova, N.B., Carty, H.A.,
413 Umayam, L.A., Haft, D.H., Peterson, J., Beanan, M.J., White, O., Salzberg,
414 S.L., Hsia, R.C., McClarty, G., Rank, R.G., Bavoil, P.M., Fraser, C.M., (2003)

- 415 Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC):
416 examining the role of niche-specific genes in the evolution of the
417 *Chlamydiaceae*, *Nucleic Acids Res*, 31, 2134-47.
- 418 [23] Sardinia, L.M., Engel, J.N., Ganem, D., (1989) Chlamydial gene encoding a 70-
419 kilodalton antigen in *Escherichia coli*: analysis of expression signals and
420 identification of the gene product, *J Bacteriol*, 171, 335-41.
- 421 [24] Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A., (1986) Information
422 content of binding sites on nucleotide sequences, *J Mol Biol*, 188, 415-31.
- 423 [25] Shen, L., Shi, Y., Douglas, A.L., Hatch, T.P., O'Connell, C.M., Chen, J.M.,
424 Zhang, Y.X., (2000) Identification and characterization of promoters regulating
425 *tuf* expression in *Chlamydia trachomatis* serovar F, *Arch Biochem Biophys*,
426 379, 46-56.
- 427 [26] Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba,
428 T., Ishii, K., Hattori, M., Kuhara, S., Nakazawa, T., (2000) Comparison of
429 whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and
430 CWL029 from USA, *Nucleic Acids Res*, 28, 2311-4.
- 431 [27] Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L.,
432 Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W.,
433 (1998) Genome sequence of an obligate intracellular pathogen of humans:
434 *Chlamydia trachomatis*, *Science*, 282, 754-9.
- 435 [28] Stormo, G.D., (2000) DNA binding sites: representation and discovery,
436 *Bioinformatics*, 16, 16-23.
- 437 [29] Stormo, G.D., Hartzell, G.W., (1989) Identifying protein-binding sites from
438 unaligned DNA fragments, *Proc Natl Acad Sci U S A*, 86, 1183-7.

- 439 [30] Tan, M., Gaal, T., Gourse, R.L., Engel, J.N., (1998) Mutational analysis of the
440 *Chlamydia trachomatis* rRNA P1 promoter defines four regions important for
441 transcription *in vitro*, J Bacteriol, 180, 2359-66.
- 442 [31] Tan, M., Wong, B., Engel, J.N., (1996) Transcriptional organization and
443 regulation of the *dnaK* and *groE* operons of *Chlamydia trachomatis*, J
444 Bacteriol, 178, 6983-90.
- 445 [32] Thomson, N.R., Yeats, C., Bell, K., Holden, M.T., Bentley, S.D., Livingstone,
446 M., Cerdeno-Tarraga, A.M., Harris, B., Doggett, J., Ormond, D., Mungall, K.,
447 Clarke, K., Feltwell, T., Hance, Z., Sanders, M., Quail, M.A., Price, C., Barrell,
448 B.G., Parkhill, J., Longbottom, D., (2005) The *Chlamydophila abortus* genome
449 sequence reveals an array of variable proteins that contribute to interspecies
450 variation, Genome Res, 15, 629-40.
- 451 [33] van Helden, J., (2003) Regulatory sequence analysis tools, Nucleic Acids Res,
452 31, 3593-6.
- 453 [34] Wels, M., Francke, C., Kerkhoven, R., Kleerebezem, M. and Siezen, R.J.,
454 (2006) Predicting *cis*-acting elements of *Lactobacillus plantarum* by
455 comparative genomics with different taxonomic subgroups, Nucleic Acids
456 Res., 34, 1947-58.
- 457 [35] Wichlan, D.G., Hatch, T.P., (1993) Identification of an early-stage gene of
458 *Chlamydia psittaci* 6BC, J Bacteriol, 175, 2936-42.
- 459

Figure 1. Schematic representation of the steps of Footy. The inner-boxed section (– · –) shows the steps of the standard PWM method.

Figure 2. Representation of how the filtering algorithm of Footy performs PWAs of all the predicted promoters within the equivalent upstream regions. **A.** Diagrammatic representation of promoters predicted within a dataset of three equivalent upstream regions. Two boxes (–35 and –10 hexamers) represent each predicted promoter, which are numbered in order of their prediction (eg. pR1 and pR2), within each upstream region. **B.** Illustration of the order in which the un-gapped PWA were performed. Hexamers predicted in part A were transferred to part B and pasted together. The line arrows represent the order in which the PWAs occurred within each loop. The boxed arrows represent the order in which different combinations of PWAs occurred between loops.

Table 1. Results of the analysis of the upstream regions of different combination of chlamydial species for positive control promoters.

1 IC threshold (bits)	7.5												7.0																							
2 Species	D & MoPn			D & GPIC			D & AR39			D & MoPn			D & GPIC			D & AR39			D, GPIC & AR39			D, MoPn, GPIC & AR39														
3 Mismatches	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3				
4 False positives																																				
5 Number of positive controls predicted		3		3									3		3						3		3		4		4						4		4	
6 Number of promoters predicted		30		31									35		44						14		14		18		24						29		16	19

IC threshold (bits)	6.5												3.0																							
Species	D & MoPn			D & GPIC			D & AR39			D, GPIC & AR39			D, MoPn, GPIC & AR39			D & MoPn			D & GPIC			D & AR39			D, GPIC & AR39			D, MoPn, GPIC & AR39								
Mismatches	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
False positives																																				
Number of positive controls predicted																					3		4		4		6		3		5	5	5		5	
Number of promoters predicted																					33		65		24		53		19		35	44	23		46	

1 IC threshold (bits)	2.5															
2 Species	D & MoPn			D & GPIC			D & AR39			D, GPIC & AR39			D, MoPn, GPIC & AR39			
3 Mismatches	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
4 False positives																
5 Number of positive controls predicted																
6 Number of promoters predicted																

Row 1: the sum of the IC threshold of the –35 and –10 hexamers used for the analysis of the dataset upstream regions of chlamydiae. Row 2: species analysed; species abbreviations are as follows: AR39, *C. pneumoniae*; D, *C. trachomatis*; GPIC, *C. caviae*; and MoPn, *C. muridarum*. Row 3: the mismatch threshold used for the analysis. Row 4: indicates the IC threshold (Row 1) species combination (Row 2), and mismatch threshold (Row 3) where the first false positive was reported. Row 5: the number of positive controls predicted. Row 6: the number of promoters predicted. (Data shown in Rows 5 and 6 are at an IC threshold 0.5 bits more than the IC threshold at which the first false positive was reported.)

Table 2. σ^{66} promoters predicted by Footy in the upstream regions of *C. trachomatis* D.

Name	Location	Sequence	-35 hexamer	-10 hexamer	IC (btis)	Mismatches	Putative product
CT017	18418	ATCGAACAGTCGCAGTTGACTTTTTTCCTTT	TTGACTTTTTTCCTTT	AAGTCAATAATAATTCCTCTCTA	9.0	1	hypothetical protein
CT043	48617	TCGCGGCTCTAATCATTACTAACAACCT	TTACTAACAACCT	GCTTATGCTAGGTTTAAAAAAC	5.5	2	hypothetical protein
CT062-tyrS	71848	CTGCTATCGCTTGCCTTGCTATAAAAAGAAC	TTGCTATAAAAAGAAC	AGGATAGATAAGATGTTGCTAGAT	7.5	2	throsyl tRNA synthetase
CT066	799156	AAAGATAAACACTAATTGATTTTTATTTT	TTGATTTTTATTTT	ACTGAACATTAATCGAAAAAAC	5.9	2	hypothetical protein
CT098-rs1	115695	AGTCAAGGGAAATCTTGCCCTTTTTAAGG	TTTGCCCTTTTTAAGG	TGAATATTTACTACTCTcttTTG	5.9	0	30S ribosomal protein S1
CT111-groES	128417	CAACTGCTAAACCAGTTGCAAAAAAGCGAG	TTGCAAAAAAGCGAG	GACTTTGCTATCGTTCTTCCtCTG	7.5	1	10 kD chaperonin (heat shock protein GroES)
tRNAAla_1	202535	TTTGATAATCTTTTCTTGTCCTTAAATCGC	TTTGTCCTTAAATCGC	TCTTGGATTAAGATGGCGCTTTGT	5.1	2	tRNA Ala
CT197-gcp1	221387	CTTGGCATTAAACGCTTGCTTGATTAACAA	TTGCTTGATTAACAA	TCTCATGATACGATCCCTCTCCTC	5.8	2	O-sialoglycoprotein endopeptidase
CT265-accA	297412	TAAGAGAAATTAATTTGTTGCGTGAAA	TTGTTGCGTGAAA	AAGGTCAATATAATCAAATAGTTG	8.7	2	acetyl-CoA carboxylase transferase (α subunit)
CT266	298721	TCAGCGTAAGCAAGCTTGACTCTAAATTTT	TTGACTCTAAATTTT	CTCAAGATTAATTTTTGCCATTGG	8.7	2	hypothetical protein
CT267-ihfA	299179	AAGATAAAAAAGTCTTGAATCCAAAGGA	TTGAATCCAAAGGA	TGAATGCATATATACGCATATAT	8.6	1	histone-like DNA binding proteins, IHFA, IHFB or DBH
CT269-murE	301529	GTTAGTCGACAAAGCTTGACAACGAATAT	TTGACAACGAATAT	GTGTATAGTAACTATTTGAGAAA	11.5	2	UDP-N-acetylmuramoylalanylglutamyl DAP Ligase
CT273	305141	TCCCCTCTACCATACTTGACTTTTTCCCT	TTGACTTTTTCCCT	CCCCCGATTATGATTGAGATTGTG	11.3	1	Chlamydia-specified hypothetical protein (basic)
CT286-clpC	317907	TCCCTTTACGAAAAGTTGCATCATTATCAT	TTGCATCATTATCAT	AAATGTCGTATATGCTTGAaaaAT	4.4	0	CLP protease ATP-binding subunit (CLPA/CLPB/CLPC)
CT297-rnc	330617	AACTCGAAAAGTACTATAGACTTTAAGATT	TTAGACTTTAAGATT	TTCCCGCTATAAAAAACCCGATTG	5.4	0	ribonuclease III
CT323-infA	363851	TTTTTGACAAGTTGTTTGACATTTTCTGT	TTTGACATTTTCTGT	TTAGTCGATATAATCGCTCTcTCG	14.4	1	translation initiation factor IF-1
CT342-rs21	391021	CAACTTAAGTATCTCTTGAAGCTAAAATAAA	TTGAAGCTAAAATAAA	AGTGGTGTACAAATCCCCGTCTC	9.4	0	30S ribosomal protein S21
CT393-proS	447959	AAAAAATCACAGAGATTGATCTGATAAACAC	TTGATCTGATAAACAC	TCCTATGCTAAGATGCTCTCCAC	7.2	0	Prolyl tRNA synthetase
CT398	156056	TTAAACAAAACGTGCTTACTTCTTGCAGA	TTACTTCTTGCAGA	AAAATCGGTAAACTTGCCGTTTCG	7.6	1	conserved hypothetical protein
CT439-rs12	508566	CCTAGAAATAACCCCTTGCAAACAAGATAT	TTGCAAACAAGATAT	TCTTATCTATATTTCCCTGtTTG	9.6	1	30S ribosomal protein S12
CT446-euo	517160	TTTTTAACAAACCGCTTGATTAATAAGTTT	TTGATTAATAAGTTT	TTTGTGGGAAATgttacCTTCT	5.8	0	CHLPS Euo protein
CT455-murA	530787	TATTTTTATTTGTTTTTAAAAACAACAAT	TTTAAAAACAACAAT	GTCTCTTTGTTAATAAGATGTTTT	4.0	2	UDP-N-Acetylglucosamine Transferase
CT475-pheT	548337	CTCCCCTCAAATACTTGCTACTATACACG	TTGCTACTATACACG	CCACTTCGTAaaATCTACCAAAAA	9.1	1	phenylalanyl tRNA synthetase beta
CT496-pgsA1	574775	AGCTAAACTCTCTGCTTGCTTTTGGAGTGT	TTGCTTTTGGAGTGT	CTATGTTTCATAATATGTGTCAAT	6.9	2	CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase
CT528-r13	596666	GCTTAGCTTTTCTTATTGTAAAAATCGTCT	TTTGTAAAAATCGTCT	TCCTTTGATAACTGTCCCTTTAA	5.3	1	L3 Ribosomal protein
CT546	617372	AACAAAAAATTTATTGGCATTGCTGTTT	TTTGGCATTGCTGTTT	TTATTTTATAAAATAAAATAAAAAG	4.3	2	Chlamydia-specific hypothetical protein
CT547	617442	AAAGCTATAAGAGATTGACAAATTTCTTTT	TTGACAAATTTCTTTT	TTCTTTTTTATGATGACGCTTTGT	12.9	0	Chlamydia-specific hypothetical protein
CT559-sctJ	631398	AAAAAATGTTCCGATTGGCACTAATCTCC	TTGCACTAATCTCC	CCATTTGCTATGGTGAGTGaaaAG	8.8	0	flagellar M-ring protein (YopJ translocation protein)
CT596-exbB	676817	ATACCAAAAAGGATCTTGGTTCTATAACAAG	TTGATCTATAACAAG	AAATTTGTTAGGATCGTCTAGGAA	5.6	1	polysaccharide transporter
CT619	701690	TTATAAAAACAAACATAGAAAAAACTTTTT	TTGAAAAAACTTTTTT	TTAAATAAGAAAAATAAAACATAA	4.2	0	hypothetical protein
CT626-rs4	714150	AATCTAGGAAATCCCTTGTAGAAAATTGGA	TTGAGAAAATTGGA	AATAGAACTAGAAATGCTCTTTTTGT	8.0	1	30S ribosomal protein S4
CT636-greA	723227	TTATAAAAACAAACATAGAAAAAACTTTTT	TTGAAAAAACTTTTTT	TTAAATAAGAAAAATAAAACATAA	7.8	0	transcription elongation factor (GreA)
CT646	741250	TAATTAATGTTTTTCTTGAaaaAGATGTTT	TTTGAaaaAGATGTTT	TTATTTTTTAAATGAGCGCTCTT	10.9	0	Chlamydia-specific hypothetical protein
CT681-ompA	780229	GTTTTTCTTATCAACTTTACGAGAATAAGAA	TTTACGAGAATAAGAA	AATTTTGTTATGGTCTCGAGCATT	7.1	2	Major outer membrane protein
tRNAGly_2	778678	TTCTCAAAGAAAGATTGCATAAAAACTCTT	TTGCAATAAAAACTCTT	GCTTCCAGTACTATATCGGTCTAC	5.5	2	tRNA Gly
CT706-clpP2	813229	ATCGCAGGAAAACGCTTGACCCAAGAGACA	TTGACCCAAGAGACA	CTTAAACATAGAAATTCATCATTTT	11.4	0	ATP-dependent ClpP endopeptidase subunit
CT708-snf	814796	GGGTCAAAATTTTCATTGATTAGCGGAAG	TTTTCATTGATTAGCGGAAG	TAAAAAGGTACAAGTAACAGaTcT	5.2	0	probable helicase

Table 2. (Continued)

Name	Location	Sequence	-35 hexamer	-10 hexamer	IC (btis)	Mismatches	Putative product
CT752- <i>efp2</i>	884446	TTCGCGACATTCTTCT GGACA AGCTTAGAA	GGACA	GAGAACGATA ACAT AGATGGaGAA	8.1	0	Elongation factor P (EF-P)
CT768	901360	GATCCATAAAAACCG TTGACG AATAATGCAT	TTGACG	TGCCAGAG CAACT TTGACTACCA	7.9	1	hypothetical protein
CT769- <i>jbeB</i>	903504	CTTTAGAAAAAAGCT CGAC CTTATCTTAGA	CGAC	TAATCGGG TAT TCTCAGGCCAGTT	6.8	1	iojap superfamily ortholog
CT827- <i>nrdA</i>	974155	TATGCTATTTTCAAT TGCAG GAAACGTTG	TGCAG	CTAGCTT CTATAT TGGTATACAA	4.6	1	ribonucleoside diphosphate reductase alpha chain
CT837	984553	TATAAAATAAAATAT TTGAA AGCTAATTCAT	TTGAA	TTATAAAATA AACT AGAAGACAAT	9.6	2	hypothetical protein

Table 2. (Continued)

Promoters were identified by the name of the associated gene. The location of promoters refers to the site on the *C. trachomatis* D genome of the first base of the predicted -35 hexamer. The bold uppercase nucleotides represent the predicted -35 and -10 hexamers and the bold lower case nucleotides represent the mapped TSSs (if available). Promoters aligned at the predicted -35 and -10 hexamers. Shaded nucleotides represent the nucleotides that were identical when the equivalent upstream regions of *C. trachomatis* D, *C. pneumoniae* AR39 and *C. psittaci* GPIC were aligned. The corresponding nucleotide/s in *C. trachomatis* D are marked, for TSSs determined in *C. muridarum* MoPn for CT098-*rs1* and CT111-*groES* [23, 31]; *C. trachomatis* serovar F for CT323-*infA* [25]; *C. psittaci* biovar 6BC for CT446-*euo* [35] and *C. trachomatis* L2 for CT062-*tyrS*, CT286-*clpC*, CT439-*rs12*, CT444-*omcA* [15], CT559-*sctJ*, CT596-*exbB*, CT708-*snf* and CT752-*efp2*. The information content (IC) of the promoters was the sum of the IC scores for the individually predicted -35 and -10 hexamers. Mismatches are the maximum number of mismatches the predicted promoters in *C. pneumoniae* AR39 or *C. psittaci* GPIC were from the predicted promoter in *C. trachomatis* D. Putative product is as described by GenBank.

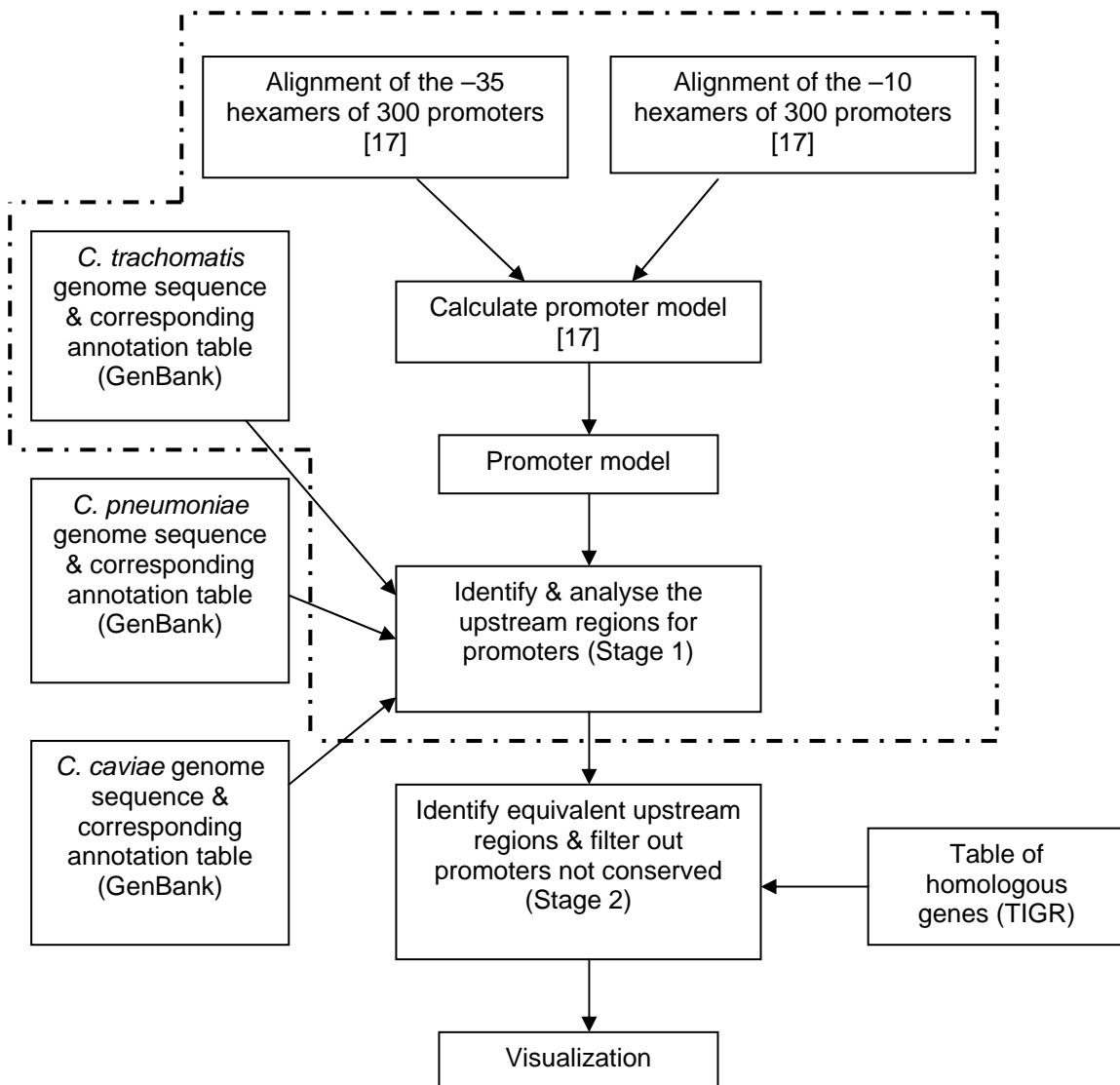
Table 3. Analysis of the 15 positive control promoters predicted and not predicted by Footy.

Promoter name	True positives ^a		False negatives ^b	
		Promoter located partly or fully within a gene	Promoter eliminated by stage 1 of Footy ^c	Promoter eliminated by stage 2 of Footy ^c
<i>infA</i> (F)	✓			
<i>omcA</i> (L1, L2, 6BC, EAE & IOL207)		✓		
16S rDNA1 P1 (L2 & MoPn)			✓	✓
CT602 (L2)			✓	
<i>hctA</i> (L2 & MN)			✓	
<i>ltuA</i> (L2)			✓	✓
<i>ltuB</i> (L2)			✓	✓
<i>ompA</i> P1 (L2)			✓	✓
<i>ompA</i> P2-P3 (L2, GPIC & MN)			✓	✓
<i>pkn5</i> (L2)			✓	✓
<i>rpoD</i> (L2)			✓	
<i>sctU</i> (L2)			✓	
<i>secA</i> (L2)			✓	✓
<i>srp</i> (L1)			✓	✓
tRNAThr2 (F)				✓

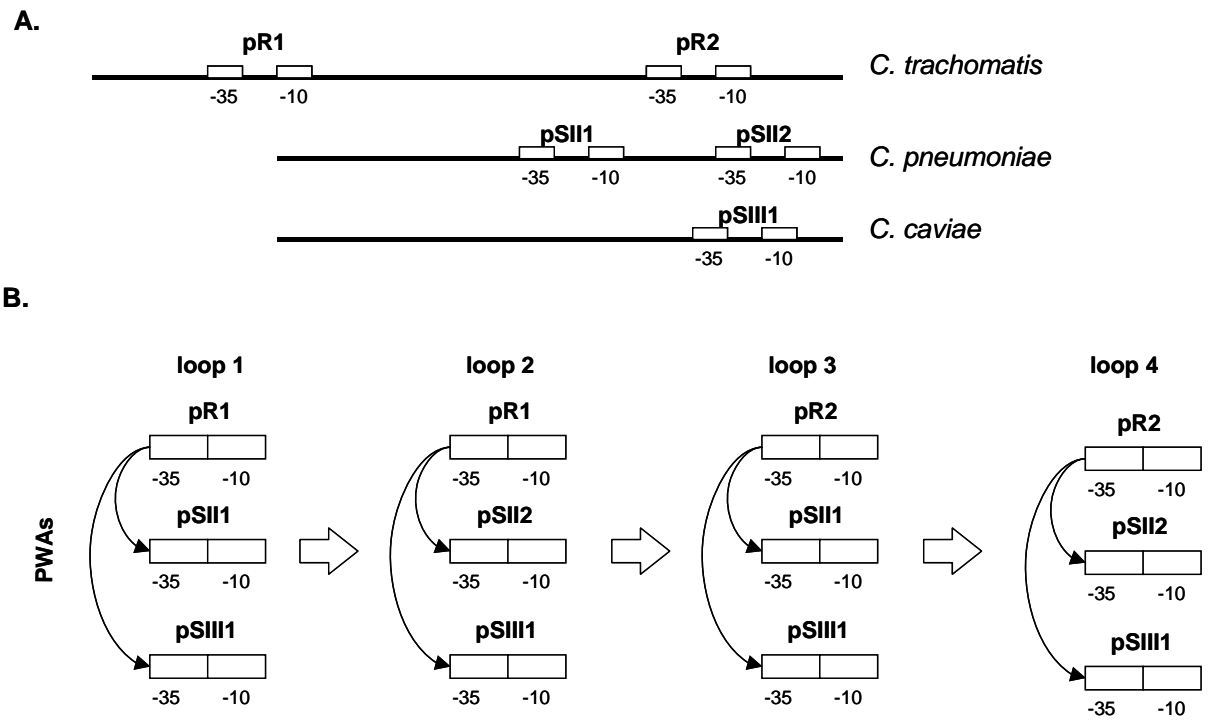
^aTrue positives refer to known promoters predicted by Footy. ^bFalse negatives refer to known promoters not predicted by Footy. ^cIndicates at which stage (1 or 2) of Footy the positive control promoters were eliminated. Parentheses contain strain names. Strain abbreviations are as follows: 6BC, *C. psittaci* strain 6BC; EAE, *C. psittaci* strain EAE/A22/M; F, *C. trachomatis* serovar F; GPIC, *C. caviae* biovar GPIC; IOL207, *C. pneumoniae* strain IOL-207; L1, *C. trachomatis* serovar L1; L2, *C. trachomatis* serovar L2; MN, *C. psittaci* sub-species meningo-pneumonitis; and MoPn, *C. muridarum* biovar MoPn.

Table 4. Comparison of Footy to some of the preferred promoter prediction programmes in bacteria.

Programme	Organism/s	Regions analysed	No. of predictions in genome/s	Positive controls predicted	Maximum no. of mismatches from the σ^{70} consensus sequence	No. also predicted by Footy
Standard PWM method	<i>C. trachomatis</i>	Upstream regions	10	1	2	7
MITRA	<i>C. muridarum</i>	Regions between divergently transcribed genes	0	N/A	N/A	N/A
MITRA	<i>C. pneumoniae</i>	Regions between divergently transcribed genes	27	N/A	3	1
Footy	<i>C. trachomatis</i> <i>C. pneumoniae</i> <i>C. caviae</i>	Equivalent upstream regions	42	1	5	N/A



(Fig. 1)



(Fig. 2)

Supplementary material

Genome wide analysis of chlamydiae for promoters that phylogenetically footprint

Brian Grech, Stefan Maetschke, Sarah Mathews, Peter Timms

Figure S1. Determination of 5' RNA ends of 16S rRNA (control), *tyrS*, *clpC*, *rs12*, *sctJ*, *exbB*, *snf* and *efp2*. The upper box under the peaks is the peak position (bp), the middle box is the size of peak in fluorescent units and the lower box is the area under the peak. Two sets of boxes corresponding to the two peaks are present for *snf*. For all of the electropherograms no peaks were observed between 300 and 500 bp.

$$IC_{ij} = \text{Log}_2 \left(\frac{f_{ij}}{p_i} \right)$$

(Equation S1)

Where:

 IC_{ij} : information content of base i in column j (bits) i : A, C, G, T j : 1 to length of the sequences f_{ij} : relative frequency of base i in column j of the sequences p_i : relative frequency of base i in the reference genome (relative background frequency)

$$f_{ij} = \frac{(c_{ij} + \alpha)}{(n + 4\alpha)}$$

(Equation S2)

Where:

 f_{ij} : relative frequency of base i in column j i : A, C, G, T j : 1 to length of the sequences c_{ij} : count of base i in column j α : constant to handle zero counts (set to 1)

$$IC_p = \sum_{\substack{j=1 \\ i=A,C,G,T}}^L \text{Log}_2 \left(\frac{f_{ij}}{p_i} \right)$$

 n : number of aligned sequences (assuming $\sum c_{iA}, c_{iC}, c_{iG}, c_{iT}$ is the same for all i)

(Equation S3)

Where:

 IC_p : information content of the pattern (bits) L : combined length of all the elements of the model (excluding gaps) f_{ij} : relative frequency of base i in column j of the sequences p_i : relative frequency of base i in the *C. trachomatis* D genome (relative background frequency)

Table S1. Oligonucleotide sequences used for primer extension analysis.

Gene	Primer name	Sequence	Site (5') on the genome (bp) ^b	Reaction temperature (°C)
CT062- <i>tyrS</i>	<i>tyrS</i>	5' 6-FAM TGGATCGAACCCTAAATAGGCAGAAACAGG	72 017	55
CT197- <i>gcp1</i>	<i>gcp1</i>	5' 6-FAM AGATTTTCCCCTTCTGGACAAGAGAACAGG	221498	65
CT286- <i>clpC</i>	<i>clpC</i>	5' 6-FAM TCCTAGATAGTTGTGATTGAGTCGTTGAGC	318028	55
CT439- <i>rs12</i>	<i>rs12</i>	5' 6-FAM GGAGTTTTTAGTTTTTACCTGAAGACAGACC	508 404	55
CT547	CT547	5' 6-FAM TTCAAAAGAGGGCACTCGTGCATAACATCC	617553	55
CT559- <i>sctJ</i>	<i>sctJ</i>	5' 6-FAM TAATCATGGAACGACTATCACAAAGCCGAGC	631 506	55
CT569- <i>exbB</i>	<i>exbB</i>	5' 6-FAM AAGGACTGTCCATGTACATATCGAAAGAGC	676 993	55
CT636- <i>greA</i>	<i>greA</i>	5' 6-FAM TTAACGACATCATTA AACAGTACTCTTCC	723 347	55
CT708- <i>snf</i>	<i>snf</i>	5' 6-FAM CCTTGAGCAAATAACTCTTTTCCATCTTGC	814909	55
CT708- <i>snf</i>	<i>snf-2</i>	5' 6-FAM CCATTCATAGATAAGATTTTGGCACTAACC	814945	65
16S rDNA1 ^a	16SrRNA	5' 6-FAM GAACCAAGATCAAATTCCTCAG	854 187	55
16S rDNA2 ^a	16SrRNA	5' 6-FAM GAACCAAGATCAAATTCCTCAG	876 203	55
16S rDNA1 ^a	16SrRNA-2	5' 6-FAM AATATATACTTTGATTTATTAACGGGTTC	854 153	65
16S rDNA2 ^a	16SrRNA-2	5' 6-FAM AATATATACTTTGATTTATTAACGGGTTC	876 169	65
CT752- <i>efp2</i>	<i>efp2</i>	5' 6-FAM TTTGACTTTGATTCTATTA AAAGCCTGTCC	884 639	55

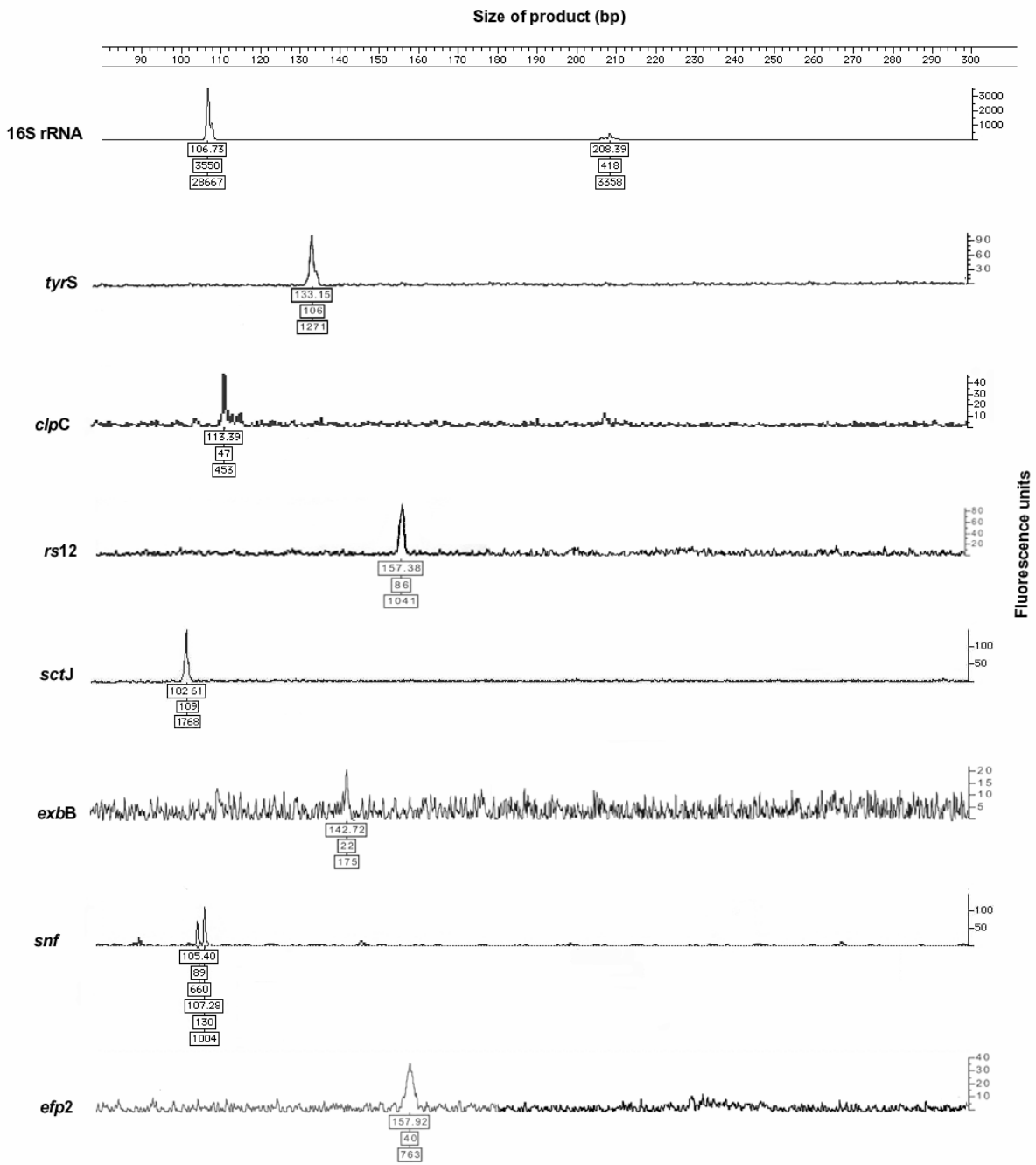
^a16S rRNA was used as a control for the primer extension. ^b*C. trachomatis* D genome [27]

Table S2. σ^{66} promoters predicted by the standard PWM method in the upstream regions of *C. trachomatis* D.

Name	Location	Sequence	IC (bits)	Putative product	
CT016	17572	TGTGCTAATCTGTT TTGTCA AAAAATGT	ACCCCTTAAC TACAAT GCCGAGGAA	10.86	hypothetical protein
CT273	305141	CCCTCTACCATACT TTGACT TTTTTCCC	TCCCCCGATT TATGAT TGAGATTGT	11.32	Chlamydia-specified hypothetical protein (basic)
CT323- <i>infA</i>	363851	TTTGACAAGTTGTT TTGACA TTTTCTG	TTTAGTCGAT TATAAT CGCTCTcTC	14.55	translation initiation factor IF-1
CT394- <i>hrcA</i>	449801	AGCGCTAAAAATC TTGACC AGTGGAG	ACGGTTTTCT TATAAT GACACCGAC	13.00	HTH Transcriptional Repressor
CT547	617442	AGCTATAAGAGAT TTGACA AATCTCT	TTTTCTTTTT TATGAT GACGCTTTG	12.96	hypothetical protein
CT617- <i>rs20</i>	697518	ATCTACTATGATAT TTGGCA ACTACTG	AAACTTCCTT TAAAA TAGGTCTCTT	10.76	S20 ribosomal protein
CT646	741250	ATTAATGTTTTTC TTGAAA AAGATGT	TTTTATTTTT TAAAA TGAGCGCTCT	10.71	Chlamydia-specific hypothetical protein
CT706- <i>clpP2</i>	813229	CGCAGGAAAAACG CTTGACC CAAGAGA	CACTTAAACAT TAGAAT TCATCATTT	11.46	ATP-dependent ClpP endopeptidase subunit
CT763	897915	ATAAGCGAAAAT TTGACG CTTTTTT	AGAATTCAT TATATT CTTCCCACA	11.46	hypothetical protein
CT796- <i>glyQ</i>	903504	TTTT TTGAAA AAGTCAG	CGCCACATG TATGAT CTATCCGGC	11.10	glycyl-tRNA synthetase alpha chain

The bold upper case nucleotides represent the predicted –35 and –10 hexamers, respectively. The bold lower case nucleotides represent the mapped TSSs.

The location of promoters refers to the site on the *C. trachomatis* D genome of the first base of the predicted –35 hexamer. The TSSs were determined in other strains of *C. trachomatis* and the corresponding nucleotides in *C. trachomatis* D are marked. The TSSs were determined in *C. trachomatis* serovar F for CT323-*infA* [25]. Promoters are aligned at the predicted –35 and –10 hexamers. The information content (IC) of the promoter is the sum of the IC scores for the individual –35 and –10 hexamers.



(Fig. S1)