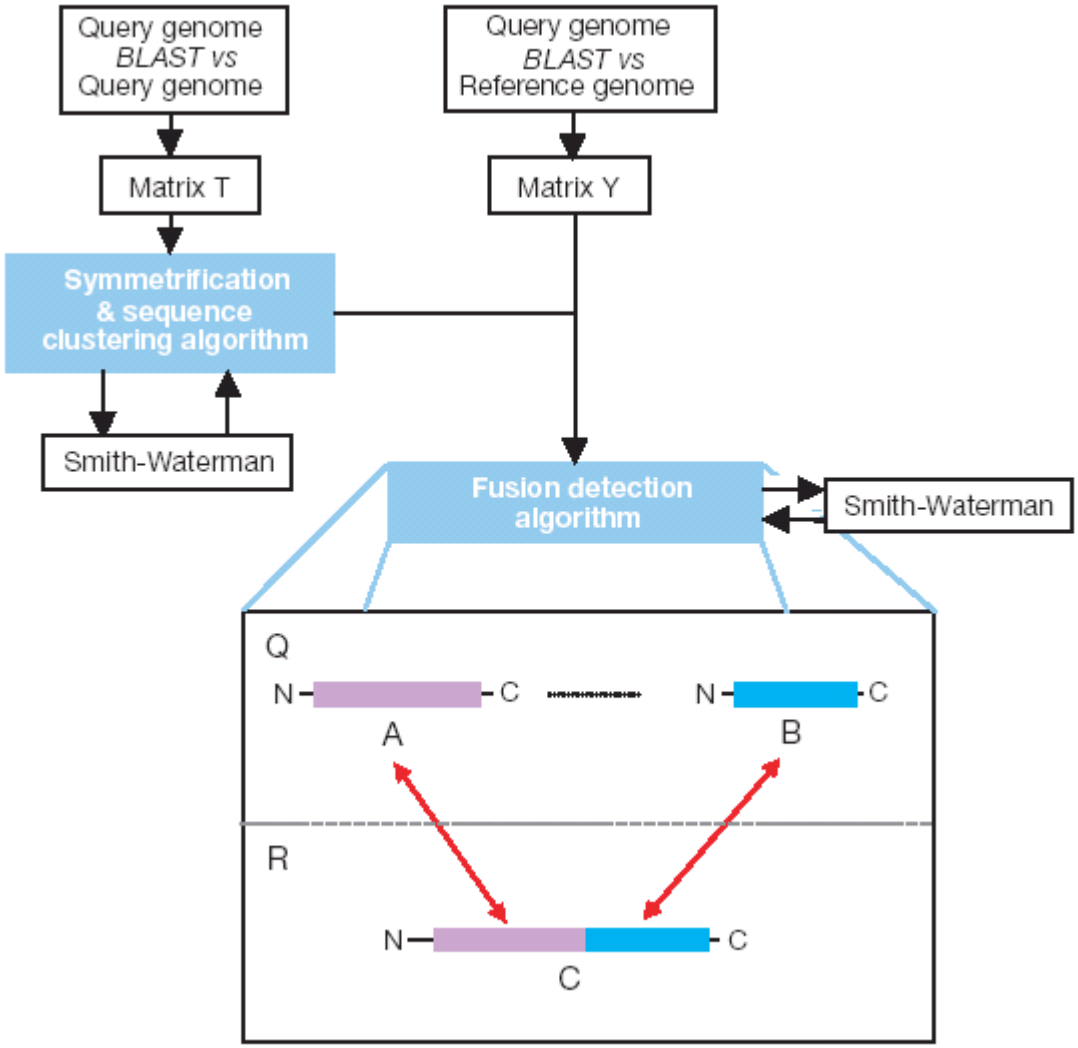
ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΙΙ

Συγκριτική Γονιδιωματική

Παντελής Μπάγκος

Protein interaction maps for complete genomes based on gene fusion events



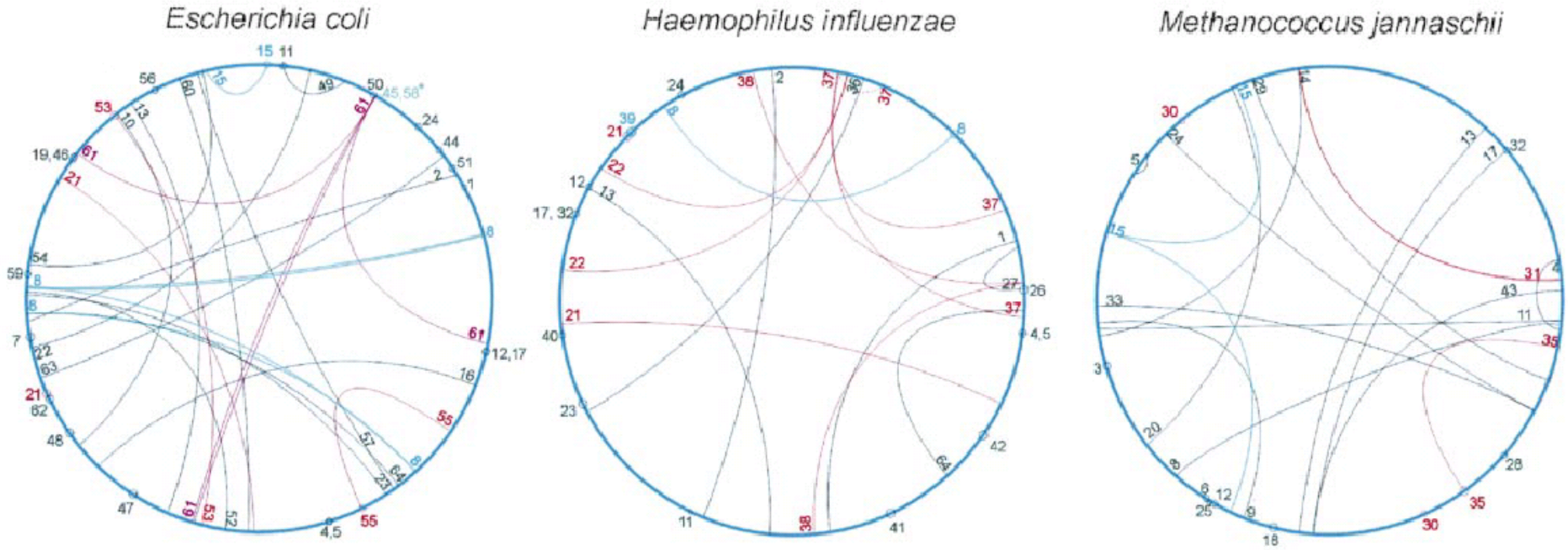


Figure 2 Representation of protein interaction maps for the most likely interactions predicted for *E. coli*, *H. influenzae* and *M. jannaschii*. In the large blue circles, which represent the three genomes, 0° corresponds to the first base pair, and 360° the last base pair of the genome. Predicted interactions are indicated by linking the circular map positions of the genes involved. In cases of neighbouring genes (<5°), a small circle indicates the predicted interaction between two genes at that region; otherwise, an arc links the two genes in question. Multiple interactions are not cross-labelled. Some paralogous cases are resolved and only the most likely case is indicated by an arc. All cases are numbered according to Table 1. Predictions are colour coded: black, pairwise interactions; blue, multiple interactions; red/purple, cases where, due to paralogy, more than one pairwise interaction is possible (red, two possibilities; purple, more than two possibilities); green (marked by asterisk), because of a large number of paralogues, no interaction can be easily resolved. The source of the prediction (composite protein from a given species) is not indicated.

Enright et al, NATURE,402, 1999

Σύνοψη

- Χρήση μόνο ομοιότητας ακολουθιών
- 88 Fusion events σε 3 γονιδιώματα
- Αναγνωρίζονται πολλές μακρινές (στο γονιδίωμα) αλληλεπιδράσεις
- Ορισμένες περιοχές στο γονιδίωμα έχουν μεγάλη τάση για fusions

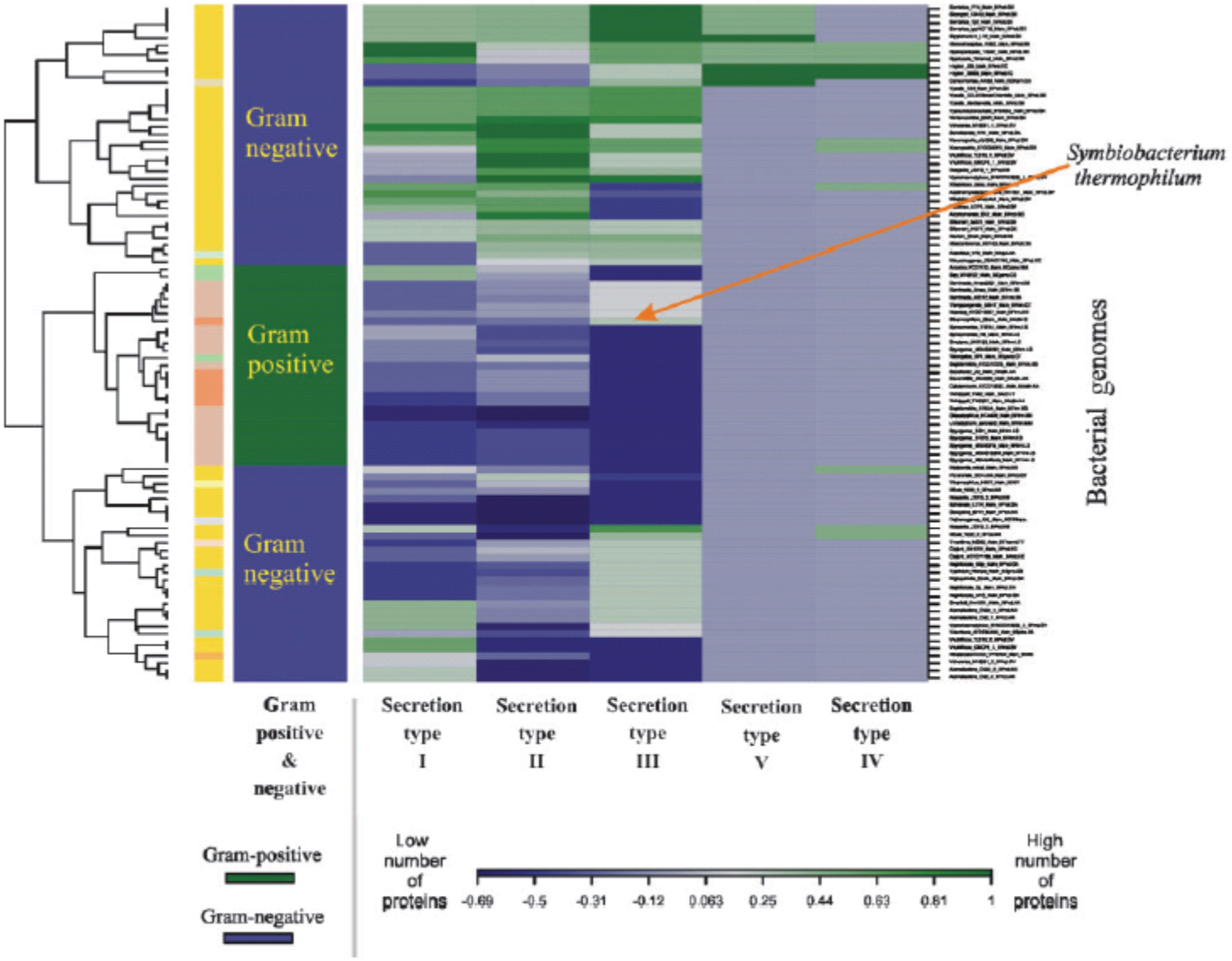


Fig. 1. Two-dimensional clustering (Willenbrock *et al.*, 2005) of bacterial genome sequences versus secretion systems type I–V. Dark blue indicates that a low number of the selected proteins is present for the specific secretion type; dark green represents cases where we find that most of the proteins for a given secretion system are present. It should be noted that data within each column are normalized around the centre using minimum and maximum values.

Whole genome alignment

- http://mummer.sourceforge.net/
- http://bioperl.org/wiki/LAGAN
- http://www.ncbi.nlm.nih.gov/BLAST/ (mega-BLAST)
- http://athena.bioc.uvic.ca/techDoc/basebybase/
- http://gel.ahabs.wisc.edu/mauve/index.php
- http://bibiserv.techfak.uni-bielefeld.de/mga/
- http://genome.lbl.gov/vista/index.shtml

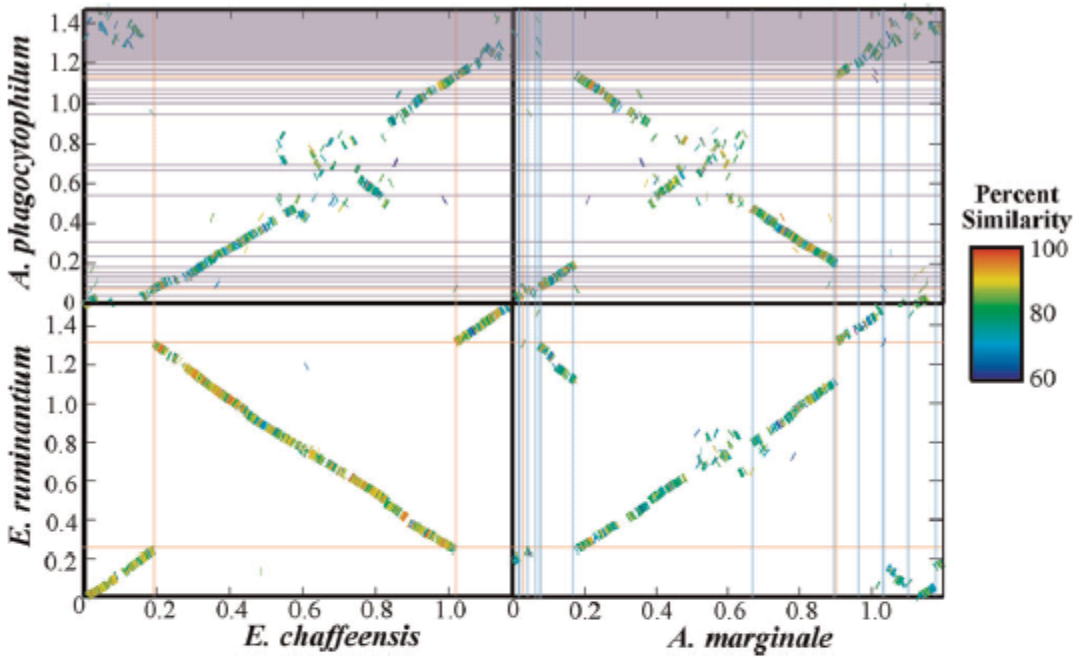


Figure 3. Synteny between *Anaplasma* spp. and *Ehrlichia* spp.

Anaplasma spp. and *Ehrlichia* spp. share conserved gene order (synteny) across their chromosomes. *E. ruminantium* and *E. chaffeensis* have a single symmetrical inversion near two duplicate Rho termination factors (approximate positions shown in pink). Genomic rearrangements between these Rho termination factors are also apparent in *A. marginale* (pink). In addition to the synteny breaks near the Rho termination factors, *A. marginale* has rearrangements located near the *msp2*- and *msp3*-expression locus and pseudogenes (approximate positions shown in light blue). Likewise, in *A. phagocytophilum,* numerous changes in genome arrangement are located near the homologous *p44* expression locus and silent genes (approximate positions shown in lavender).

DOI: 10.1371/journal.pgen.0020021.g003

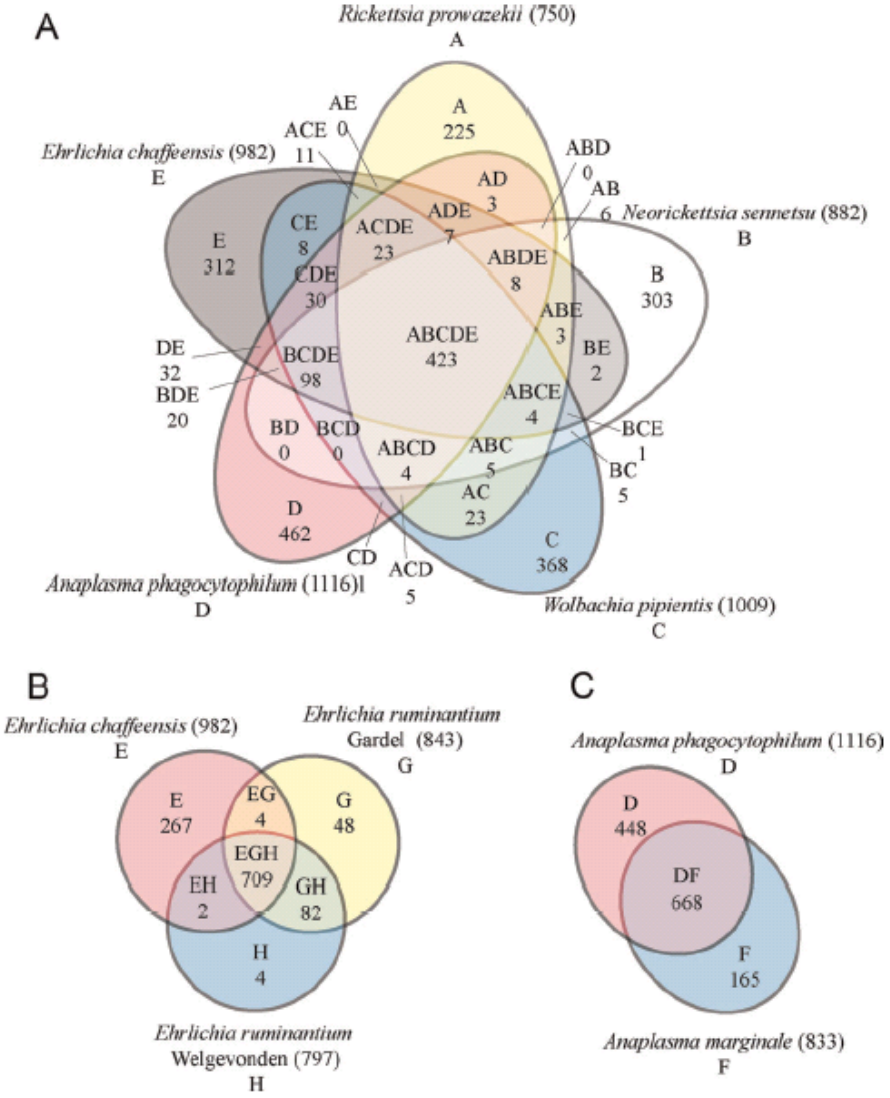


Figure 4. Comparison of the Rickettsiales Gene Sets

The composition of ortholog clusters (see Materials and Methods) of representative Rickettsiales (A), *Ehrlichia* spp. (B), and *Anaplasma* spp. (C) were compared. Numbers within the intersections of different ovals indicate ortholog clusters shared by 2, 3, 4, or 5 organisms. Species compared are indicated in diagram intersections as follows. A, *R. prowazekii;* B, *N. sennetsu;* C, *W. pipientis;* D, *A. phagocytophilum;* E, *E. chaffeensis;* F, *A. marginale;* G, *E. ruminantium* Gardel; and H, *E. ruminantium* Welgevonden.

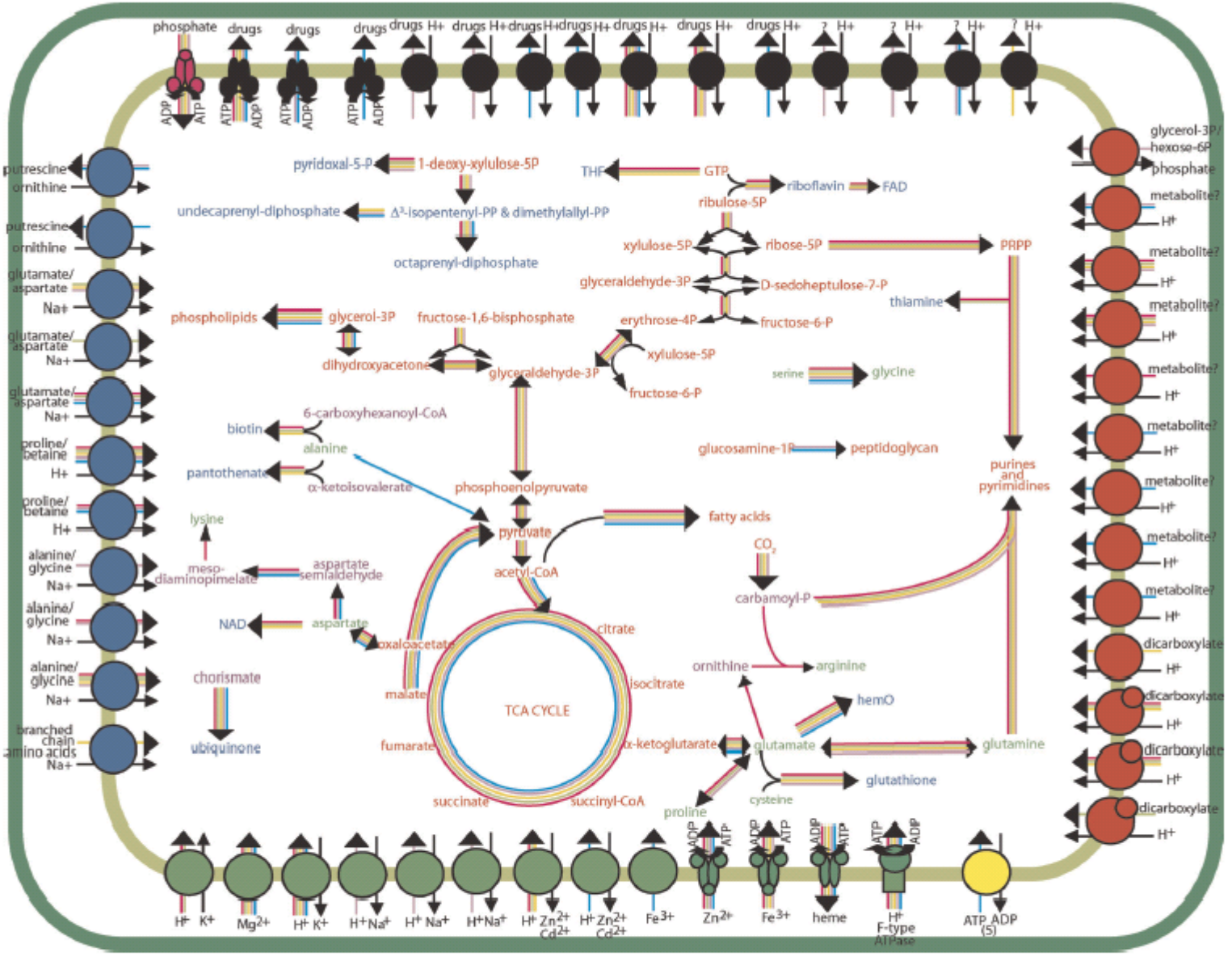


Figure 6. Comparative Metabolic Potential of Select Rickettsiales

Metabolic pathways of *E. chaffeensis* (magenta arrows), *A. phagocytophilum* (green arrows), *N. sennetsu* (gold arrows), *W. pipientis* (lavender arrows), and *R. prowazekii* (cyan arrows) were reconstructed and compared. The networks of some of the more important pathways are shown with metabolites color coded: red and purple, central and intermediary metabolites; blue, cofactors; green, amino acids; and black, cell structures. Transporters are shown in the membrane and are grouped by predicted substrate specificity: green, inorganic cations; magenta, inorganic anions; red, carbohydrates and carboxylates; blue, amino acids/peptides/amines; yellow, nucleotides/nucleosides; and black, drug/polysaccharide efflux or unknown.

DOI: 10.1371/journal.pgen.0020021.g006

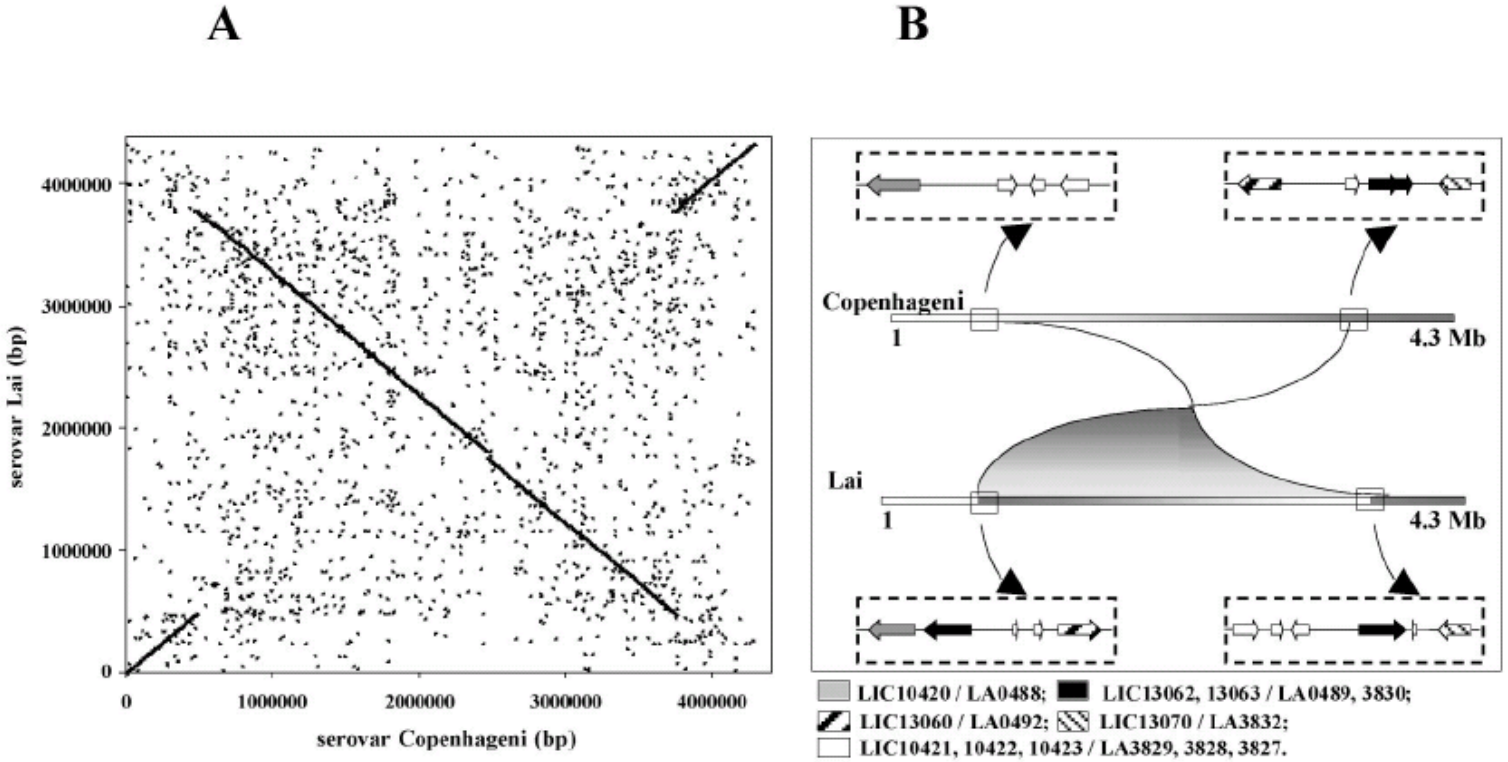


FIG. 1. Inversion of *L. interrogans* serovars Copenhageni and Lai CI chromosomes. (A) Nucleotide alignment obtained by using MUMmer, which relies on exact matches of at least 20 bp. Each dot in the figure is one such match. The dark lines on the two main diagonals result from the high density of points with sequence identity along chromosome I of the two serovars. The scattered points outside the main diagonals represent other short regions of sequence identity. (B) Scheme showing predicted genes flanking the inversion breakpoints. Pairs of ortholog genes have the same pattern code. The black arrows represent IS elements.

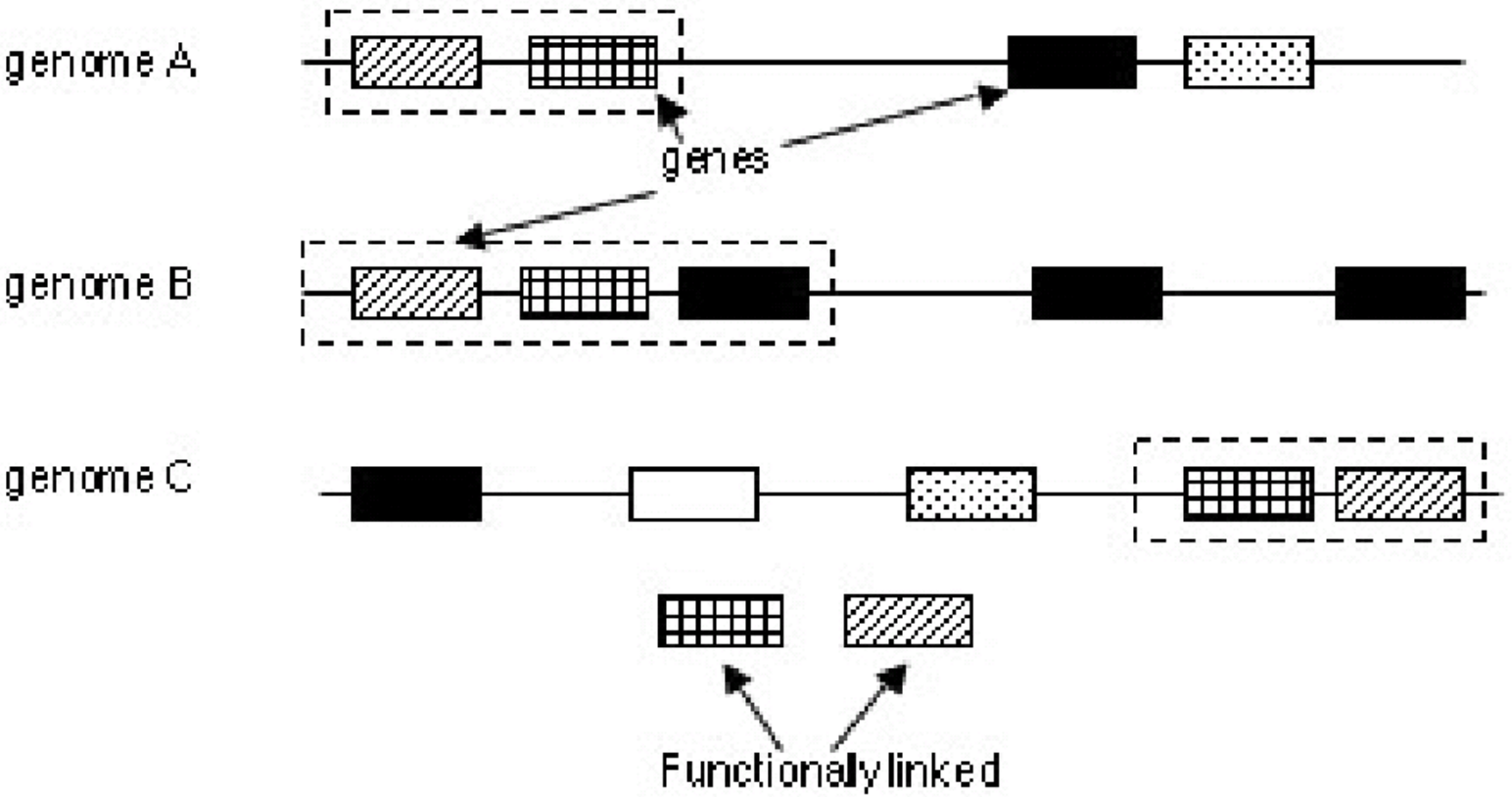
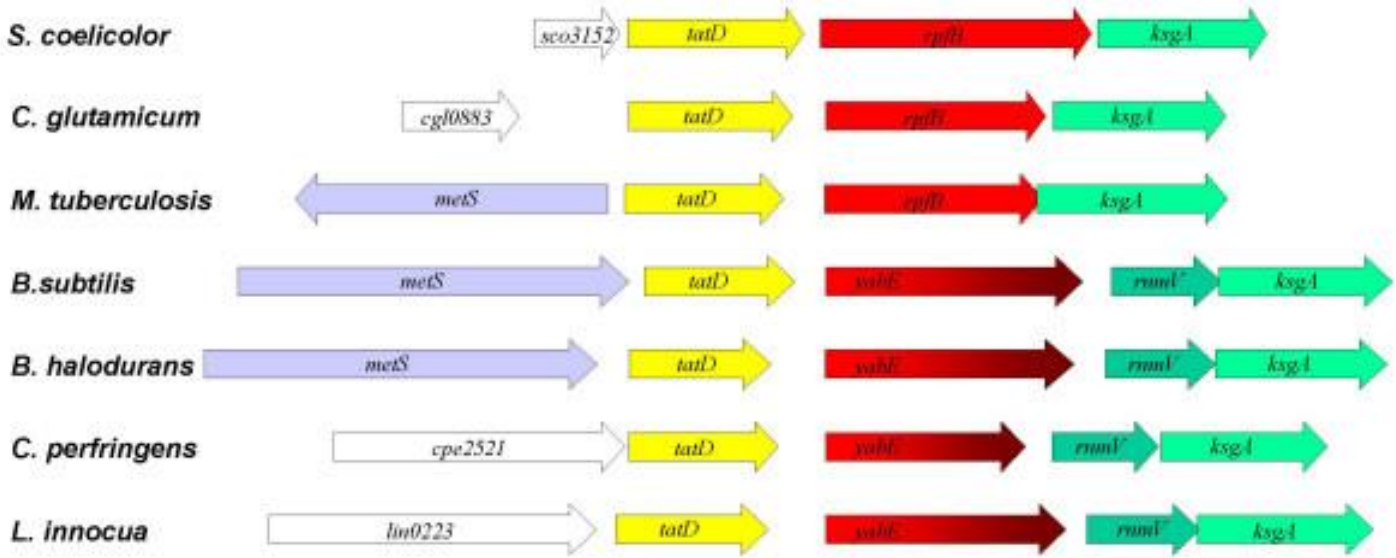
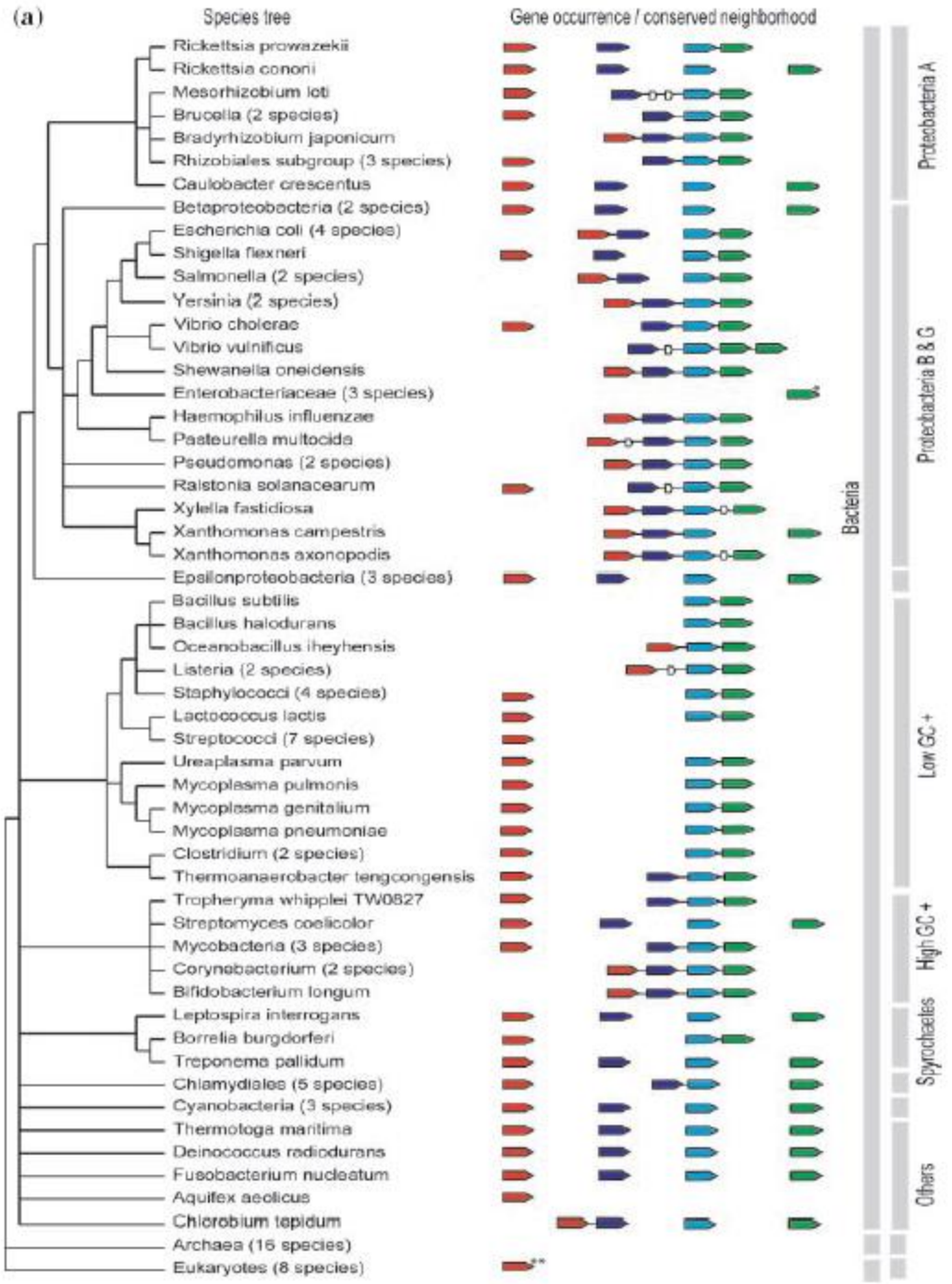


Fig. 2. Conservation of gene order/neighborhood. Genes that are consistent neighbors across multiple genomes may be function-


S. coelicolor
sco3152
tatD
rpfB
ksgA
C. glutamicum
cgl0883
tatD
rpfB
ksgA
M. tuberculosis
metS
tatD
rpfB
ksgA
B.subtilis
metS
tatD
rnmV
ksgA
B. halodurans
metS
tatD
rnmV
ksgA
C. perfringens
cpe2521
tatD
rnmV
ksgA
L. innocua
lin0223
tatD
rnmV
ksgA


(a)
Species tree
Gene occurrence / conserved neighborhood
Rickettsia prowazekii
Rickettsia conorii
Mesorhizobium loti
Brucella (2 species)
Bradyrhizobium japonicum
Rhizobiales subgroup (3 species)
Caulobacter crescentus
Betaproteobacteria (2 species)
Escherichia coli (4 species)
Shigella flexneri
Salmonella (2 species)
Yersinia (2 species)
Vibrio cholerae
Vibrio vulnificus
Shewanella oneidensis
Enterobacteriaceae (3 species)
Haemophilus influenzae
Pasteurella multocida
Pseudomonas (2 species)
Ralstonia solanacearum
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
Epsilonproteobacteria (3 species)
Bacillus subtilis
Bacillus halodurans
Oceanobacillus iheyhensis
Listeria (2 species)
Staphylococci (4 species)
Lactococcus lactis
Streptococci (7 species)
Ureaplasma parvum
Mycoplasma pulmonis
Mycoplasma genitalium
Mycoplasma pneumoniae
Clostridium (2 species)
Thermoanaerobacter tengcongensis
Tropheryma whipplei TW0827
Streptomyces coelicolor
Mycobacteria (3 species)
Corynebacterium (2 species)
Bifidobacterium longum
Leptospira interrogans
Borrelia burgdorferi
Treponema pallidum
Chlamydiales (5 species)
Cyanobacteria (3 species)
Thermotoga maritima
Deinococcus radiodurans
Fusobacterium nucleatum
Aquifex aeolicus
Chlorobium tepidum
Archaea (16 species)
Eukaryotes (8 species)
**
Bacteria
Proteobacteria A
Proteobacteria B & G
Low GC +
High GC +
Spyrochaetes
Others

(b)

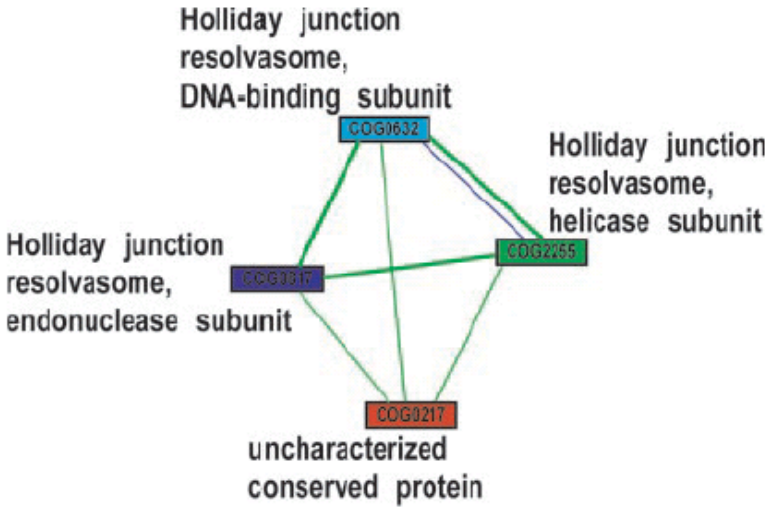


Figure 2. (**a**) Case study: evidence linking COG0217 to well-annotated proteins. Species tree showing conserved operon architecture and co-occurrence of genes coding for subunits of Holliday junction resolvasome (COG0217: uncharacterized conserved protein (red), COG0632: Holliday junction resolvasome, DNA-binding subunit (light blue), COG2255: Holliday junction resolvasome, helicase subunit (dark blue), COG0817: Holliday junction resolvasome, endonuclease subnit (green). Single asterisk (*), not present in Buchnera species and double asterisks(**), not present in Encephalitozoon cuniculi. (**b**) Network representation of evidence related to COG0217 (red). The network edges represent the predicted functional associations. An edge may be drawn with up to three different colour lines—these lines represent the existence of the three types of evidence used in predicting the associations. A red line indicates the presence of fusion evidence; a green line represents the neighbourhood evidence; and a blue line the co-occurrence evidence. Line thickness correlates linearly with STRING scores.

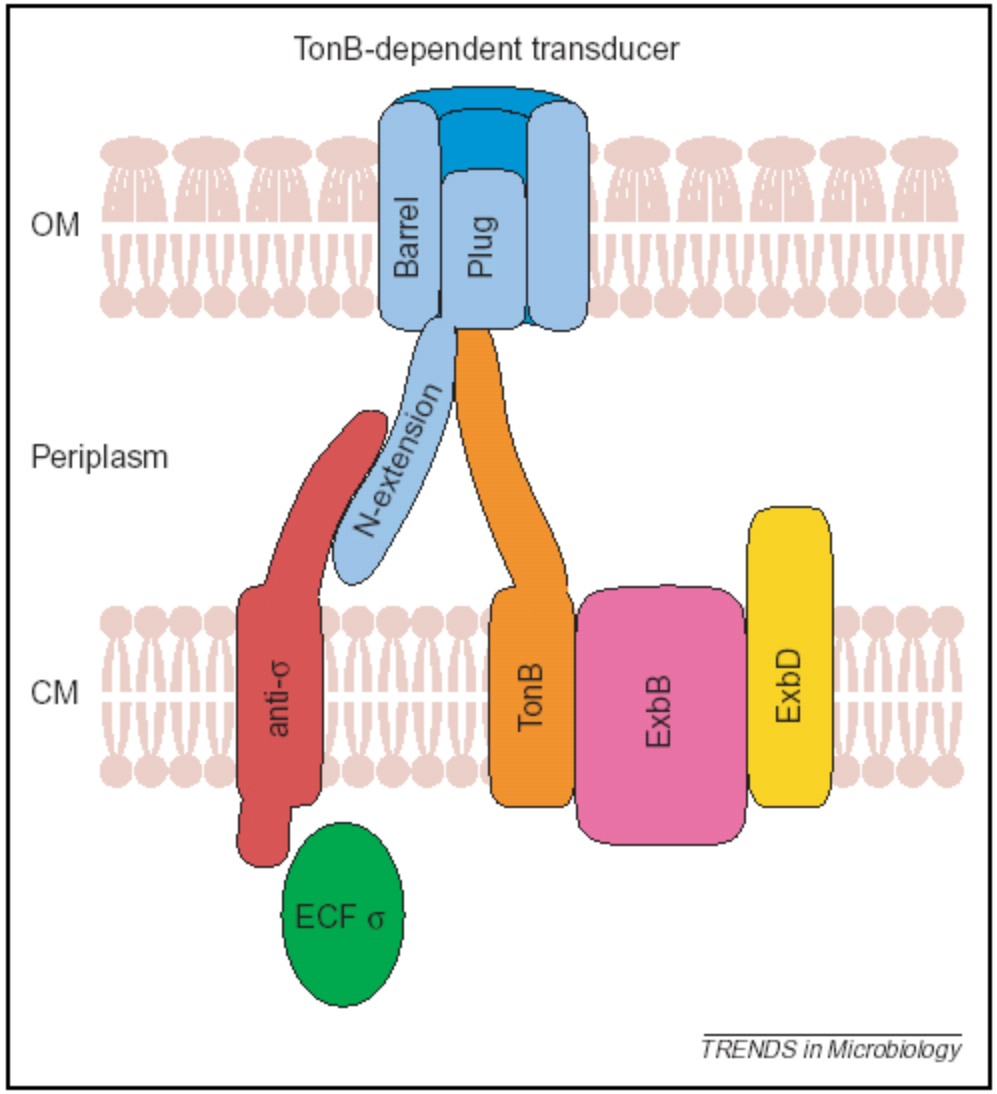


Figure 1. Structural organization of TonB-dependent regulatory systems. TonB-dependent regulatory systems consist of six components, an outer membrane TonB-dependent transducer (blue) in interplay with its energizing TonB–ExbBD protein complex (orange, pink and yellow), a cytoplasmic membrane-localized anti-sigma factor (red) and an ECF-subfamily sigma factor (green). Abbreviations: CM, cytoplasmic membrane; OM, outer membrane.

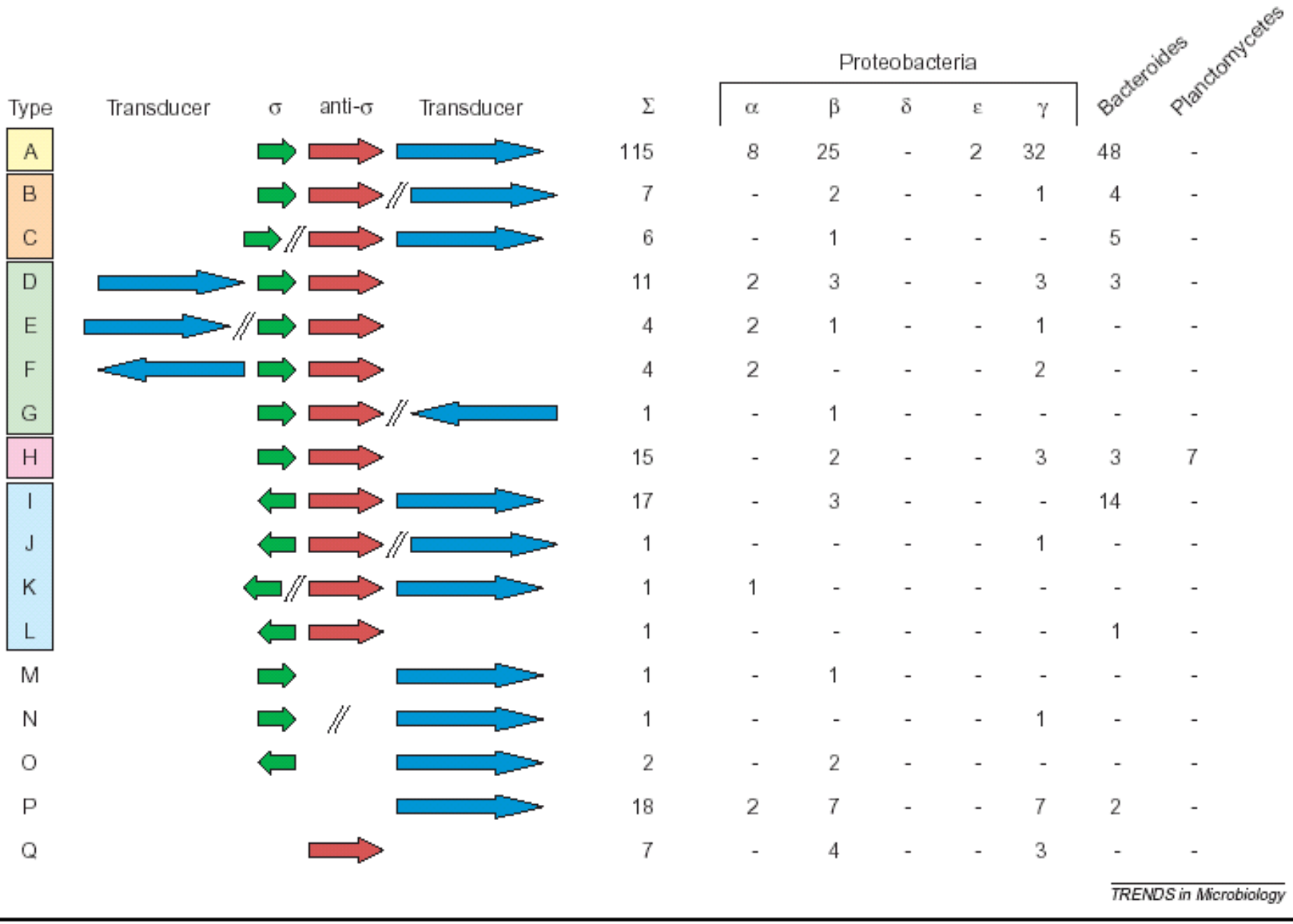


Figure 2. Genetic organization of regulatory systems consisting of a TonB-dependent transducer (blue), an anti-sigma factor (red) and an ECF-subfamily sigma factor (green). The occurrence of the different genetic organizations in proteobacteria (α, β, δ, ε and γ), in *Bacteroides*, in planctomycetes and in total (Σ) is shown. A diagonal double line indicates that one or a few additional genes are present between the TonB-dependent regulatory genes. Pseudogenes have been included in this representation. The color code of the type (left) corresponds to that of Supplementary Table 2.

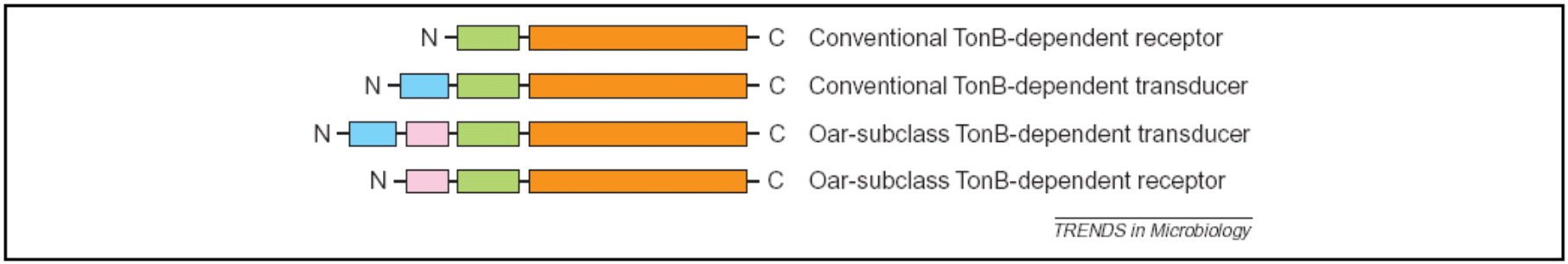


Figure 3. Domain structure of TonB-dependent receptors. All TonB-dependent receptors consist of a C-terminal β-barrel (orange) and a plug domain (green), which seals the barrel (see also Figure 1). TonB-dependent transducers have an N-terminal extension (blue) of ~70 amino acids. Receptors from *Bacteroides* often have another additional domain in the N-terminal region (pink). A related protein domain is also found in the Oar protein from *M. xanthus* and in a few receptors from *Xanthomonas* and *Xylella* species.

Minimal gene set

- *Mycoplasma genitalium* (468 identified protein-coding genes)
- *Haemophilus influenzae* (1703 genes)
- 240 *M. genitalium* genes have orthologs among the genes of *H. influenzae*.
- 22 nonorthologous displacements

Last universal common ancestor (LUCA)

- 1000 gene families, of which more than 90% are also functionally characterized.
- when only prokaryotes are considered, the number varies between 1006 and 1189 gene families while when eukaryotes are also included, this number increases to between 1344 and 1529 families
- the common belief that the hypothetical genome of LUCA should resemble those of the smallest extant genomes of obligate parasites is not supported by recent advances in computational

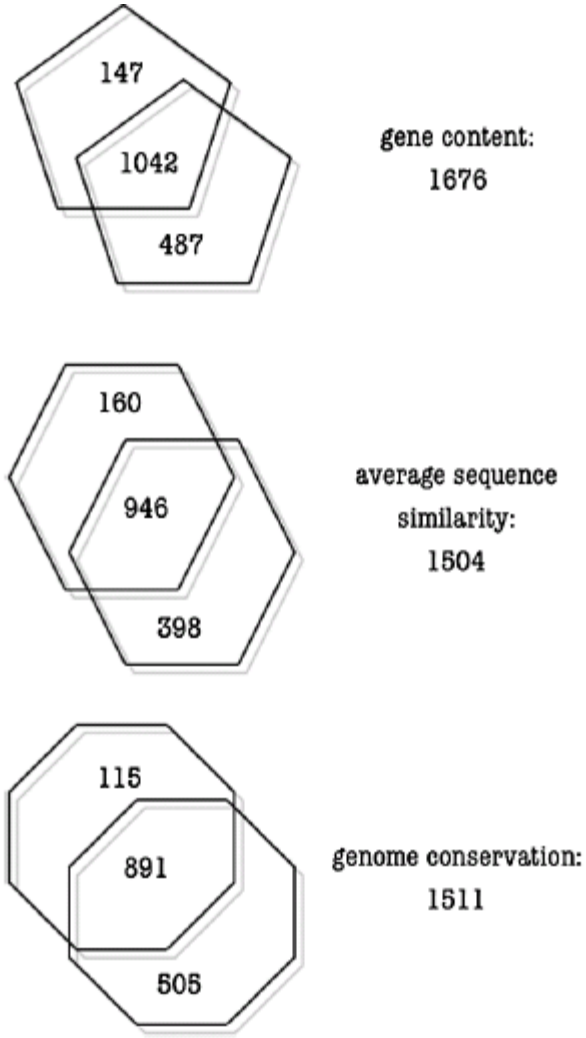


Fig. 1. A representation of the minimal gene content for LUCA. Upper diagrams represent analyses without eukaryotes, lower diagrams represent analyses with eukaryotes; pentagons represent gene content (CT), hexagons represent average sequence similarity (AS), octagons represent genome conservation (GC)—see Section 2. The number of unique (outside the intersection) and common (inside the intersection) gene families per category are given in the diagrams; the number of total unique families is also provided (listed below the corresponding method).

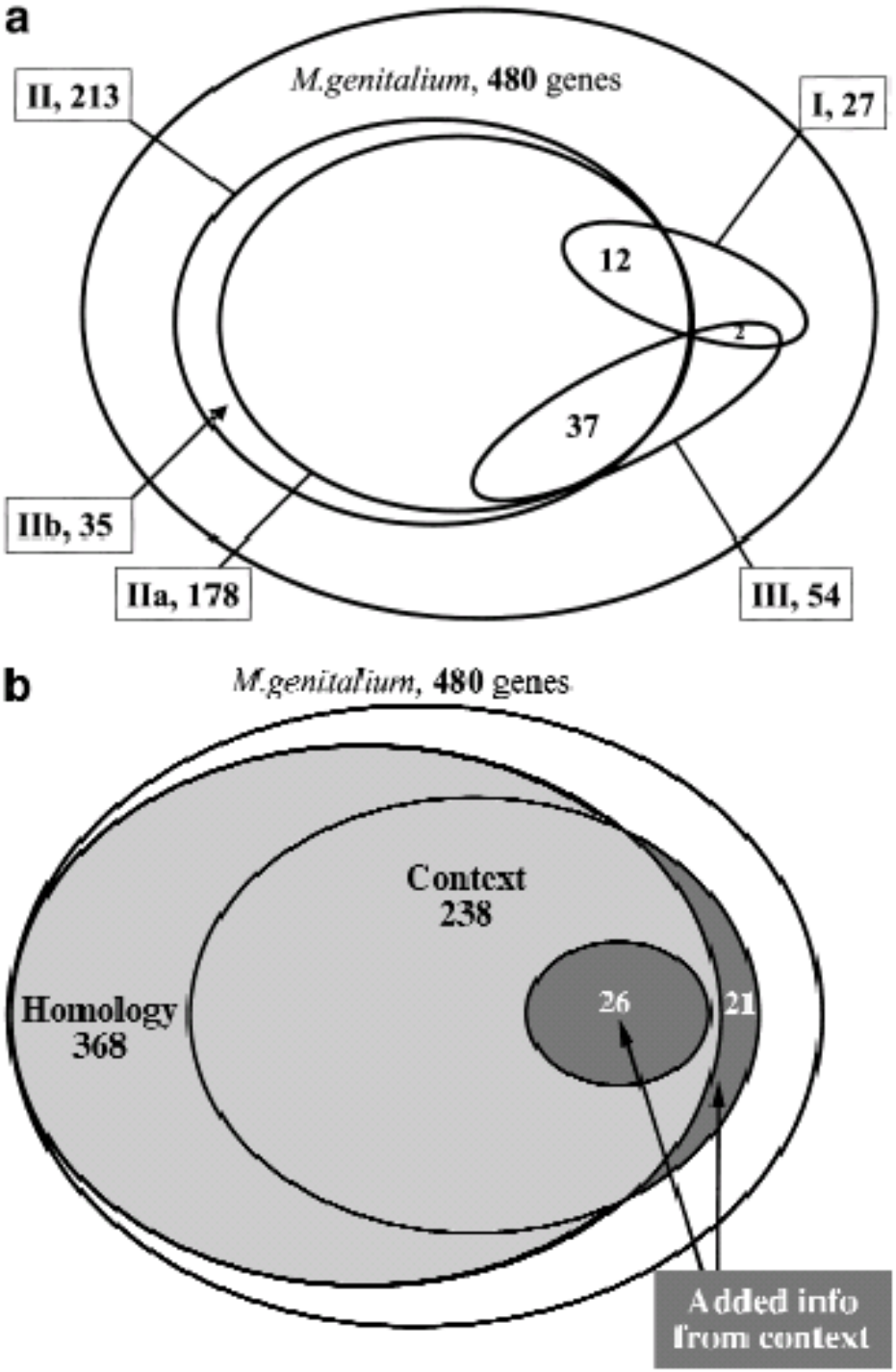


Figure 1 (a) Coverage of and overlap between various types of genomic context for *M. genitalium* genes. Type I is gene-fusion. Type II is the conservation local gene neighborhood, which is separated in type IIa (the conservation of gene order) and type IIb (the co-occurrence of genes within potential operons in absence of the conservation of gene order). Type III is the co-occurrence of genes in genomes. (b) Overlap between genes for which significant genomic context is available and genes for which functional features can be predicted by homology searches. For the latter, only genes that are homologous to genes with known molecular functions were included, which were determined by manual inspection. The dark gray areas in the figure are genes for which new functional features can be predicted by genomic context. They can be homologous to proteins with a known molecular function, in which case the context can indicate in which process this function plays a role (see text for specific examples). A complete list of genes for which new functional features could be predicted by genomic context and, if available, homology to proteins with known function, is available from http://dove.embl-heidelberg.de/MG/Context.

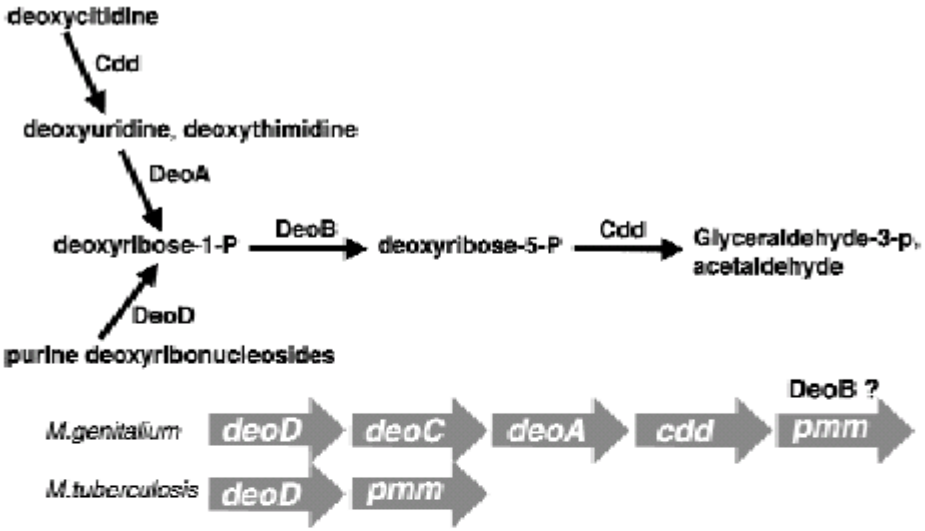


Figure 3 Genomic context predicts substrate specificity of proteins involved in a nucleoside salvage pathway in *M. genitalium*. A cluster of five genes in *M. genitalium* encodes four genes of a nucleoside salvage pathway. The "standard" gene for this fifth reaction in the pathway, phosphoribomutase (*deoB*), is absent. The fifth gene in the operon is homologous to phosphomannomutases and phosphoglucomutases. *M. genitalium* does not contain any other candidate for a phosphoribomutase. The most likely candidate for the phosphoribomutase is thus MG053. The significance of the location of a homolog of MG053 in a run with *deoD* is supported by the location of a homolog of the *M. genitalium* gene MG053 beside *deoD* in *Mycobacterium tuberculosis*.

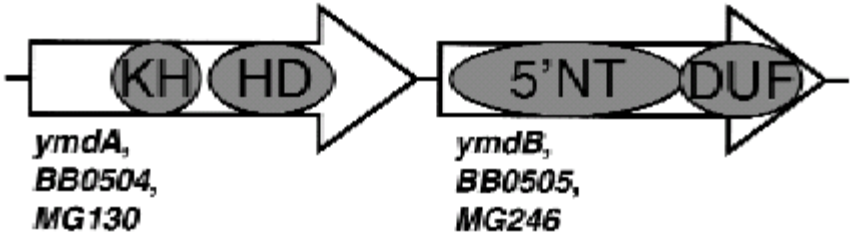


Figure 4 Domain organization of two proteins that are encoded by neighboring genes on *B. subtilis* (*ymdA* and *ymdB*) and *B. burgdorferi* (BB0504 and BB0505), and that are both present in *M. genitalium* (MG130 and MG246). The three domains that have functionally been characterized, KH, HD, and 5′NT, can all be related to ribonucleotide metabolism. KH binds (single-stranded) RNA; HD hydrolyzes phosphates from nucleotides; and 5′NT hydrolyzes NMP to nucleosides. A fourth, uncharacterized sequence domain (DUF) is present at C-terminus of MG246 and its orthologs.

Βιβλιογραφία

- Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA.Nature. **Protein interaction maps for complete genomes based on gene fusion events.**1999 Nov 4;402(6757):86-90.[PDF]
- Alfonso Valencia and Florencio Pazos. **Computational methods for the prediction of protein interactions.** Current Opinion in Structural Biology 2002, 12:368–373 [PDF]
- Tsoka S, Ouzounis CA. **Recent developments and future directions in computational genomics**.FEBS Lett. 2000 Aug 25;480(1):42-8.[PDF]