The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria

Siv G. E. Andersson*, Alireza Zomorodipour*, Jan O. Andersson*, Thomas Sicheritz-Pontén*, U. Cecilia M. Alsmark*, Raf M. Podowski*, A. Kristina Näslund*, Ann-Sofie Eriksson*, Herbert H. Winkler† & Charles G. Kurland*

* *Department of Molecular Biology, University of Uppsala, Uppsala S-75124, Sweden*
† *Department of Microbiology and Immunology, University of South Alabama, Mobile, Alabama 36688, USA*

Nature © Macmillan Publishers Ltd 1998

We describe here the complete genome sequence (1,111,523 base pairs) of the obligate intracellular parasite *Rickettsia prowazekii*, the causative agent of epidemic typhus. This genome contains 834 protein-coding genes. The functional profiles of these genes show similarities to those of mitochondrial genes: no genes required for anaerobic glycolysis are found in either *R. prowazekii* or mitochondrial genomes, but a complete set of genes encoding components of the tricarboxylic acid cycle and the respiratory-chain complex is found in *R. prowazekii*. In effect, ATP production in *Rickettsia* is the same as that in mitochondria. Many genes involved in the biosynthesis and regulation of biosynthesis of amino acids and nucleosides in free-living bacteria are absent from *R. prowazekii* and mitochondria. Such genes seem to have been replaced by homologues in the nuclear (host) genome. The *R. prowazekii* genome contains the highest proportion of non-coding DNA (24%) detected so far in a microbial genome. Such non-coding sequences may be degraded remnants of 'neutralized' genes that await elimination from the genome. Phylogenetic analyses indicate that *R. prowazekii* is more closely related to mitochondria than is any other microbe studied so far.

The *Rickettsia* are α-proteobacteria that multiply in eukaryotic cells only. *R. prowazekii* is the agent of epidemic, louse-borne typhus in humans. Three features of this endocellular parasite deserve our attention. First, *R. prowazekii* is estimated to have infected 20–30 million humans in the wake of the First World War and killed another few million following the Second World War (ref. 1). Because it is the descendent of free-living organisms[2–4], its genome provides insight into adaptations to the obligate intracellular lifestyle, with probable practical value. Second, phylogenetic analyses based on sequences of ribosomal RNA and heat-shock proteins indicate that mitochondria may be derived from the α-proteobacteria[5,6]. Indeed, the closest extant relatives of the ancestor to mitochondria seem to be the *Rickettsia*[7–10]. That modern *Rickettsia* favour an intracellular lifestyle identifies these bacteria as the sort of organism that might have initiated the endosymbiotic scenario leading to modern mitochondria[11]. Finally, the genome of *R. prowazekii* is a small one, containing only 1,111,523 base pairs (bp). Its phylogenetic placement and many other characteristics identify it as a descendant of bacteria with substantially larger genomes[2–4]. Thus *Rickettsia*, like mitochondria, are good examples of highly derived genomes, the products of several types of reductive evolution.

The genome sequence of *R. prowazekii* indicates that these three features may be related. For example, prokaryotic genomes evolving within a cell dominated by a much larger, eukaryote genome and constrained by bottle-necked population dynamics will tend to lose genetic information[12,13]. Predictable sets of expendable genes will tend to disappear from the prokaryotic genome when they are made redundant by the activities of nuclear genes. Likewise, non-essential sequences and otherwise highly conserved gene clusters may be obliterated by deleterious mutations that are fixed in clonal parasite or organelle populations because they cannot be eliminated by selection. This process is ongoing in the *Rickettsia* genomes, as shown by the identification of sequences that have recently become pseudogenes. Also, a large fraction (~25%) of non-coding sequences in this genome may be gene remnants that have been degraded by mutation and have not yet been removed from the genome. Finally, transfer of genes from a mitochondrial ancestor to the nucleus of the host would both reduce the mitochondrial genome size and stabilize the symbiotic relationship. Phylogenetic reconstructions that identify genes in the *Rickettsia* genome as sister clades to eukaryotic homologues found in the nucleus or the organelle support this interpretation. *Rickettsia* and mitochondria probably share an α-proteobacterial ancestor and a similar evolutionary history.

General features of the genome

The circular chromosome of *R. prowazekii* strain Madrid E has 1,111,523 bp and an average G+C content of 29.1% (Figs 1, 2). The genome contains 834 complete open reading frames with an average length of 1,005 bp. Protein-coding genes represent 75.4% of the genome and 0.6% of the genome encodes stable RNA. We have assigned biological roles to 62.7% of the identified genes and pseudogenes; 12.5% of the identified genes match hypothetical coding sequences of unknown function and the remaining 24.8% represent unusual genes with no similarities to genes in other organisms (Table 1). Multivariate statistical analysis has shown that there is no major variation in codon-usage patterns among genes that are expressed in different amounts, indicating that codon-usage patterns in *R. prowazekii* may be dominated mainly by mutational forces[14]. G+C-content values at the three codon positions average 40.4, 31.2 and 18.6%, and these values are similar at different positions in the genome. We classified the open reading frames with significant sequence-similarity scores to gene sequences in the public databases into functional categories (Table 1) that allow comparisons with the metabolic profiles of other bacterial genomes[15–23].

Non-coding DNA. The coding content of previously sequenced bacterial genomes is, on average, 91%, ranging from 87% in *Haemophilus influenzae* to 94% in *Aquifex aeolicum*. In comparison, a large fraction of the *R. prowazekii* genome, 24%, represents non-coding DNA (Fig. 3). A small fraction of this corresponds to

Nature © Macmillan Publishers Ltd 1998

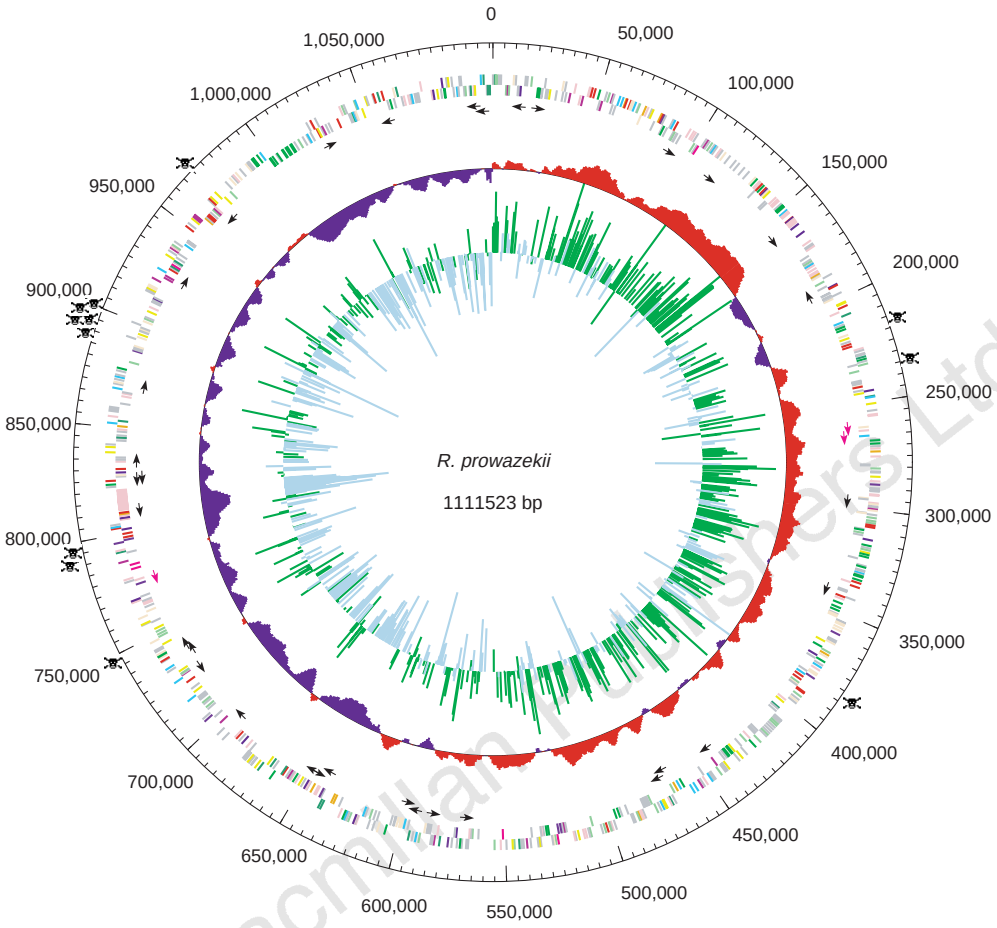

Figure 1 Overall structure of the *R. prowazekii* genome. The putative origin of replication is at 0 kb. The outer scale indicates the coordinates (in base pairs). The positions of pseudogenes are highlighted with death's heads. The distribution of genes is shown on the first two rings within the scale. The location and direction of transcription of rRNA are shown by pink arrows and of tRNA genes by black arrows. The next circle in shows GC-skew values measured over all bases in the genome. Red and purple colours denote positive and negative signs, respectively. The window size was 10,000 nucleotides and the step size was 1,000 nucleotides. The central circles shows GC-skew values calculated for third positions in the codon only. GC-skew values were calculated separately for genes located on the outer strand (green) and on the inner strand (blue). To allow easier visual inspection, the signs of the values calculated for genes located on the inner strand have been reversed.

Nature © Macmillan Publishers Ltd 1998

pseudogenes (0.9% of the genome) and less than 0.2% of the genome is accounted for by non-coding repeats. The remaining 22.9% contains no open reading frames of significant length and it has the low G+C content (mean 23.7%) that is characteristic of spacer sequences in the *R. prowazekii* genome[14]. A region of 30 kilobases (kb) located at position 886–916 kb contains as much as 41.6% non-coding DNA and 11.5% pseudogenes. The non-coding DNA in this region has a small, but significantly higher, G+C content (mean 27.3%) than non-coding DNA in other areas of the genome (mean 23.7%) ($P < 0.001$), indicating that it may correspond to inactivated genes that are being degraded by mutation (Fig. 3).

Origin of replication. The origin of replication has not been experimentally identified in the *R. prowazekii* genome, but we identified *dnaA* at ~750 kb. However, the genes flanking the *dnaA* gene differ from the conserved motifs found in *Escherichia coli* and *Bacillus subtilis* (*rnpA–rpmH–dnaA–dnaN–recF–gyrB*). In *R. prowazekii*, the genes *rnpA* and *rpmH* are located in the vicinity of *dnaA*, but in the reverse orientation compared to the consensus motif, and *dnaN*, *recF* and *gyrB* are located elsewhere.

The origin and end replication in microbial genomes are often associated with transitions in GC skew (G − C/G + C) values[24]. In *R. prowazekii* we observe transitions in the GC skew values at around 0 and 500–600 kb (Fig. 1). There is a weak asymmetry in the distribution of genes in the two strands, such that the first half of the genome has a 1.6-fold higher gene density on one strand and the second half of the genome has a 1.6-fold higher gene density on the other strand. The shift in coding-strand bias correlates with the shift in GC-skew values. As most genes are transcribed in the direction of replication in microbial genomes, the origin of replication may correspond to the shift in GC-skew values at the position that we have chosen as the start point for numbering. Indeed, several short sequence stretches that are characteristic of *dnaA*-binding motifs are found in the intergenic region of genes *RP001* and *RP885* at 0 kb, supporting this interpretation.

Stable RNA sequences and repeat elements. We identified 33 genes encoding transfer RNA, corresponding to 32 different isoacceptor-tRNA species. There is a single copy of each of the rRNA genes, with *rrs* located more than 500 kb away from the *rrl–rrf* gene cluster

Figure 2 Linear map of the *R. prowazekii* chromosome. The position and orientation of known genes are indicated by arrows. Coding regions are colour-coded according to their functional roles. The positions of tRNA genes are indicated (inverted triangle on stalk). For additional information, see http://evolution.bmc.uu.se/~siv/gnomics/Rickettsia.html.

Nature © Macmillan Publishers Ltd 1998

(Fig. 1). Comparison of the sequences from ten different *Rickettsia* species indicates that the disruption of the rRNA gene operon preceded the divergence of the typhus group and spotted fever group *Rickettisia* (S.G.E.A. *et al.*, unpublished observations). In addition, the genome contains a short sequence with similarity to a 213-nucleotide RNA molecule in *Bradyrhizobium japonicum* that may regulate transcription[25].

There are unusually few repeat sequences in this genome. We identified four different types of repeat sequence: all of these are located in intergenic regions. There is a sequence of 80 bp that is repeated seven times downstream of *rpmH* and *rnpA* in the *dnaA* region. A repetitive sequence of 325 bp is found at two intergenic regions that are more than 80 kb apart, downstream of the genes *ksgA* and *rnh*, respectively. A 440-bp-long repetitive sequence has been identified at two intergenic sites, 140 kb apart; one of these sites is downstream of *rrf* and the others downstream of *pdhA* and *pdhB*. Finally, two similar sequences of 730 bp are located immediately next to each other at 850 kb.

Paralogous families. We have identified 54 paralogous gene families comprising 147 gene products. Of these, 125 have an assigned function. Most paralogues encode proteins with transport functions, such as the ABC transporters, the proline/betaine transporters and the ATP/ADP transporters. Five paralogous genes located next to each other at 115 kb encode putative integral membrane proteins with unknown functions.

Biosynthetic pathways

A striking feature of the *R. prowazekii* genome is the small proportion of biosynthetic genes compared with free-living proteobacterial relatives (such as *Haemophilus influenzae*, *Helicobacter pylori* and *E. coli*)[15,19,20]. This scarcity of biosynthetic functions is also seen in diverse endocellular and epicellular parasites[16–18,23]. This scarcity of biosynthetic functions is also seen in diverse endocellular and epicellular parasites[16–18,23].

Amino-acid metabolism. As many as 43 and 69 genes required for amino-acid biosynthesis are found in *Helicobacter pylori* and *Haemophilus influenzae*, respectively. In contrast, *Mycoplasma genitalium* and *Borrelia burgdorferi* contain only *glyA*, which encodes serine hydroxymethyltransferase. This gene is also found in *R. prowazekii* (Table 1). Serine hydroxymethyltransferase catalyses the conversion of serine and tetrahydrofolate into glycine and methylenetetrahydrofolate, respectively. A role in tetrahydrofolate metabolism may account for the ubiquity of *glyA* in bacteria.

Seven genes normally associated with lysine biosynthesis (*lysC*, *asd*, *dapA*, *dapB*, *dapD*, *dapE* and *dapF*) are also present in *R. prowazekii*. The biosynthetic pathways leading to lysine, methionine and threonine share the first two of these (*lysC* and *asd*). However, none of the downstream genes for threonine biosynthesis are found in *R. prowazekii*. Likewise, the lysine pathway is incomplete, and *lysA*, which encodes the enzyme that converts meso-diaminopimelate to lysine, is missing. The likely role of the upstream genes of this pathway in *R. prowazekii* is the biosynthesis of diaminopimelate, an essential envelope component. We have therefore classified these genes as 'cell-envelope' genes (Table 1).

We have identified other genes that are superficially involved in the metabolism of amino acids, but which apparently function in deamination pathways that divert amino acids into the tricarboxylic acid (TCA) cycle. For example, there is *aatA*, encoding aspartate aminotransferase, which catalyses the degradation of aspartate to oxaloacetate and glutamate. *tdcB* encodes threonine deaminase, which converts threonine into α-ketobutyrate. Another gene (*ilvE*) encodes branched-chain-amino-acid aminotransferase, which converts leucine, isoleucine or valine into glutamate. *pccA* and *pccB* encode propionyl-CoA carboxylase, which converts propionyl-CoA, an intermediate in the breakdown of methionine, valine and isoleucine, into succinyl-CoA. The *pccA* and *pccB* gene products show greatest similarity to the eukaryotic proteins that are located in the mitochondrial matrix.

Nucleotide biosynthesis. No genes required for the *de novo* syntheses of nucleosides have been found in the *R. prowazekii* genome. However, four genes required for the conversion of nucleoside monophosphates into nucleoside diphosphates (*adk*, *gmk*, *cmk* and *pyrH*) are present. There are also two genes encoding ribonucleotide reductase, which converts ribonucleoside diphosphates into deoxyribonucleoside diphosphates. Nucleoside diphosphate kinase (encoded by *ndk*), which converts NDPs and dNDPs to NTPs and dNTPs, is also present in *R. prowazekii*. Finally, there is a complete set of genes for the conversion of dCTP and dUTP into TTP, including *thyA*, which codes for thymidylate synthase. Thus, the *R. prowazekii* genome encodes all of the enzymes required for the interconversion of nucleoside monophosphates into all of the other required nucleotides. The nucleoside monophosphates are probably imported from the eukaryotic host.

Table 1 Asterisks indicate putative pseudogenes. Abbreviations of species names are: Bacteria: *Acinetobacter calcoaceticus* (B-*Aca*), *Actonobacillus actinomycetemcomitans* (B-*Aac*), *Acyrthosiphon condii* (B-*Aco*), *Agrobacterium tumefaciens* (B-*Atu*), *Alcaligenes eutrophus* (B-*Aeu*), *Anabena sp.* PCC7120 (B-*Asp*), *Anabena variabilis* (B-*Ava*), *Anacystis nidulans* (B-*Ani*), *Azorhizobium caulinodans* (B-*Aca*), *Azosprillium brasiliense* (B-*Abr*), *Azotobacter vinelandii* (B-*Avi*), *Bacillus caldotenax* (B-*Bca*), *Bacillus stereothermophilus* (B-*Bst*), *Bacillus subtilis* (B-*Bsu*), *Bartonella bacilliformis* (B-*Bba*), *Bartonella henselae* (B-*Bhe*), *Bordetella pertussis* (B-*Bpe*), *Borrelia burgdorferi* (B-*Bbu*), *Bradyrhizobium japonicum* (B-*Bja*), *Brucella abortus* (B-*Bab*), *Brucella ovis* (B-*Bov*), *Caulobacter crescentus* (B-*Ccr*), *Chlamydia trachomatis* (B-*Ctr*), *Chloroflexus aurantiacus* (B-*Cau*), *Chromatum viosum* (B-*Cvi*), citrus-greening-disease-associated bacterium (B-*Cgr*), *Clostridium acetobutylicum* (B-*Cac*), *Clostridium pasteurianum* (B-*Cpa*), *Clostridium thermosaccharolyticum* (B-*Cts*), *Coxiella burnetii* (B-*Cbu*), *Erwinia chrysanthemi* (B-*Ech*), *Escherichia coli* (B-*Eco*), *Haemophilus influenzae* (B-*Hin*), *Helicobacter pylori* (B-*Hpy*), *Klebsiella pneumoniae* (B-*Kpn*), *Legionella pneumophila* (B-*Lpn*), *Leucothrix mucor* (B-*Lmu*), *Liberobacter africanum* (B-*Laf*), *Methylobacterium extorquens* (B-*Mex*), *Micrococcus luteus* (*Mlu*), *Moraxella catarrhalis* (*Mca*), *Mycobacterium leprae* (*Mle*), *Mycobacterium smegmatis* (B-*Msm*), *Mycobacterium tuberculosis* (B-*Mtu*), *Mycoplasma capricolum* (B-*Mca*), *Mycoplasma genitalium* (B-*Mge*), *Mycoplasma pneumoniae* (B-*Mpn*), *Paracoccus denitrificans* (B-*Pde*), *Pasteurella haemolytica* (B-*Pha*), *Plectonema boryanum* (B-*Pbo*), *Proteus mirabilis* (B-*Pmi*), *Proteus vulgaris* (B*Pvu*), *Pseudomonas aeruginosa* (B-*Pae*), *Pseudomonas fluorescens* (B-*Pfl*), *Pseudomonas putida* (B-*Ppu*), *Pseudomonas syringae* (B-*Psy*), *Rhizobium meliloti* (B-*Rme*), *Rhizobium sp.* NGR234 (B-*Rsp*), *Rhodobacter capsulatus* (B-*Rca*), *Rhodobacter sphaeroides* (B-*Rsp*), *Rhodobacter sulfidophilus* (B-*Rsu*), *Rhodopseudomonas blastica* (B-*Rbl*), *Rhodospirillum rubrum* (B-*Rru*), *Rickettsia japonicum* (B-*Rja*), *Rickettsia rickettsii* (B-*Rri*), *Rickettsia typhi* (B-*Rty*), *Salmonella typhi* (B-*Sti*), *Salmonella typhimurium* (B-*Sty*), *Shigella flexneri* (B-*Sfl*), *Sprioplasma citri* (B-*Sci*), *Staphylococcus aureus* (B-*Sau*), *Staphyloccus carnosus* (B-*Sca*), *Streptococcus pneumoniae* (B-*Spn*), *Streptomyces clavuligerus* (B-*Scl*), *Streptomyces coelicor* (B-*Sco*), *Synechocystis* PCC 6803 (B-*Syn*), *Thermus aquaticus* (B-*Taq*), *Thermus thermophilus* (B-*Tth*), *Thiobacillus cuprinus* (B-*Tcu*), *Treponema hyodysenteriae* (B-*Thy*), *Vibrio alginolytics* (B-*Val*), *Vibrio cholera* (B-*Vch*), *Vibrio parahaemolyticus* (B-*Vpa*), *Vibrio proteolyticus* (B-*Vpr*), *Wolbachia* sp. (B-*Wsp*), *Yersinia entercolitica* (B-*Yen*), *Zooglea ramigera* (B-*Zra*), *Zymomonas mobilis* (B-*Zmo*). Archaea: *Methanococcus jannaschii* (A-*Mja*), *Sulfulobus acidocaldarius* (A-*Sac*). Eukaryotes: *Apis mellifera* (E-*Ame*), *Arabidopsis thaliana* (E-*Ath*), *Atratylodes japonica* (E-*Aja*), *Bos taurus* (E-*Bta*), *Candida albicans* (E-*Cal*), *Caenorhabidits elegans* (E-*Cel*), *Dictylostelium discoideum* (E-*Ddi*), *Flaveria trinervia* (E-*Ftr*), *Giardia theta* (E-*Gth*), *Glycine max* (E-Gma), *Haematobia irritans* (E-*Hir*), *Homo sapiens* (E-*Hsa*), *Marchantia polymorpha* (E-*Mpa*), *Mus musculum* (E-*Mmu*), *Prototheca wickerhamii* (E-*Pwi*), *Petunia hybrida* (E-*Phy*), *Pisum sativum* (E-*Psa*), *Porphyra purpurea* (E-*Ppu*), *Odontella sinensis* (E-*Osi*), *Reclinomonas americana* (E-*Ram*), *Rattus novergicus* (E-*Rno*), *Rhizopus oryzae* (E-*Ror*), *Saccharomyces cerevisiae* (E-*Sce*), *Schizosaccharomyces pombe* (E-*Spo*), *Solanum tuberosum* (E-*Stu*), *Spinacia oleracea* (E-*Sol*).

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998

Energy metabolism

Early in its infectious cycle, *R. prowazekii* uses the ATP of the host with the help of membrane-bound ATP/ADP translocases. However, *R. prowazekii* is also capable of generating ATP, which may compensate for the gradual depletion of cytosolic ATP later in the infection. *R. prowazekii*'s repertoire of genes involved in ATP production and transport include determinants for the TCA cycle, the respiratory-chain complexes, the ATP-synthase complexes and the ATP/ADP translocases (Table 1). Genes to support anaerobic glycolysis are absent.

Pyruvate dehydrogenase. Pyruvate is imported into mitochondria directly from the cytoplasm and converted into acetyl-CoA by pyruvate dehydrogenase. The genes encoding three components (E1–E3) of the pyruvate dehydrogenase complex are found in *R. prowazekii*, indicating that it too uses cytosolic pyruvate. Pyruvate dehydrogenase (E1) consists of two subunits (α and β) in *R. prowazekii*, mitochondria and Gram-positive bacteria; the corresponding genes are clustered in the genome. In contrast, proteobacteria such as *E. coli*, *Haemophilus influenzae* and *Helicobacter pylori* have a single subunit for the E1 component and these have little similarity to the α and β subunits of the E1 component in *R. prowazekii* and mitochondria (data not shown).

Two paralagous genes code for the dihydrolipoamide dehydrogenase (E3) in *R. prowazekii*. One of these most resembles mitochondrial homologues, whereas the other is most similar to bacterial homologues (data not shown). The presence of several paralogous gene families for pyruvate dehydrogenases complicates attempts to reconstruct a genome phylogeny based on these genes.

ATP production. Genes encoding all enzymes in the TCA cycle are found in *R. prowazekii*. Proton translocation is mediated by NADH dehydrogenase (complex I), cytochrome reductase (complex III) and cytochrome oxidase (complex IV). Several clusters of genes code for components of the NADH dehydrogenase complex. Seven of these genes (*nuoJKLM* and *nuoGHO*) are located near to each other, but the order of genes is inverted relative to the order of this cluster in *E. coli*. An additional set of five genes is grouped in the order *nuoABCDE*, but the single genes *nuoF* and *nuoN* are distant from both of these clusters. Several proteins in the cytochrome bc_1 reductase complex, such as ubiquinol–cytochrome *c* reductase (encoded by *petA*), cytochrome *b* (encoded by *cytb*) and cytochrome c_1 (encoded by *fbhC*), are present, as are several subunits of the cytochrome oxidase complex.

The ATP-synthesizing complex is composed of the ATP synthase F_1 component (comprising five polypeptides, α, β, γ, ϵ and δ) and the F_o component, a hydrophobic segment that spans the inner mitochondrial membrane. The genes encoding these components are normally clustered in one of the most highly conserved operon structures in microbial genomes. In *R. prowazekii*, however, the ATP-synthase genes encoding the α, β, γ, δ and ϵ subunits of the F_1 complex (*atpH*, *atpA*, *atpG*, *atpD* and *atpC*) are clustered in the common order, but *atpB*, *atpE* and *atpF*, encoding the A, B and C chains of the F_o complex, are split from this cluster.

Replication, repair and recombination

R. prowazekii has a smaller set of genes involved in DNA replication than do free-living bacteria such as *E. coli*, *Haemophilus influenzae* and *Heliocobacter pyrlori*. Four genes have been identified that code for the core structure of DNA polymerase III, which includes the α (*dnaE*), ϵ (*dnaQ*), β (*dnaN*), γ and θ (*dnaX*) subunits. Extra subunits present in the *E. coli* DNA polymerase III are missing from *R. prowazekii*, as well as from *M. genitalium* and *B. burgdorferi*.

Genes encoding DNA-repair mechanisms are similar in the small genomes of the parasites *R. prowazekii*, *M. genitalium* and *B. burgdorferi*. Thus, genes involved in the repair of ultraviolet-induced DNA damage (*uvrABCD*) have been identified in all three genomes. In *R. prowazekii*, DNA-excision repair probably occurs by a pathway involving endonuclease III, polI and DNA ligase, as in *B. burgdorferi*.

The *R. prowazekii* genome has a limited capacity for mismatch repair. The DNA-mismatch-repair enzymes encoded by *mutL* and *mutS* are present, but *mutH* and *mutY* are not. There is a complete lack of *mut* genes in *M. genitalium*, but *mutL* and *mutHLY* have been identified in *B. burgdorferi* and *Chlamydia trachomatis*. The transcription-repair coupling factor (encoded by *mfD*) is found in *R. prowazekii*, *B. burgdorferi* and *C. trachomatis* but not in *M. genitalium*.

The *R. prowazekii* genome contains several genes involved in homologous recombination, such as *recA*, *recF*, *recJ*, *recN* and *recR*. A similar set of genes has been found in *A. aeolicus*[21]. The *rec* genes

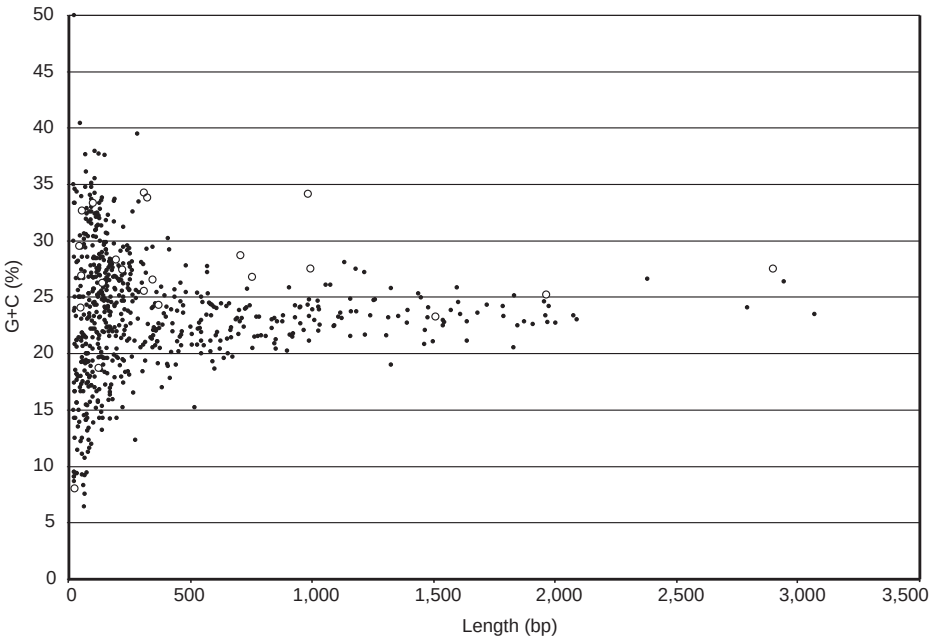

Figure 3 G+C content in intergenic regions longer than 20 bp in the *R. prowazekii* genome. The empty circles correspond to spacer sequences located at 886 to 916 kb, a region with an unusually large fraction of non-coding DNA and pseudogenes.

Nature © Macmillan Publishers Ltd 1998

are scattered in the other small genomes of parasites. The *RecBCD* complex is missing in *R. prowazekii*, *M. genitalium* and *Helicobacter pylori* but it has been identified in *B. burgdorferi*.

Transcription and translation

R. prowazekii has three subunits (α, β and β′) of the core RNA polymerase, as well as σ^{70} and one alternative σ factor, σ^{32}, which controls transcription of the genes encoding heat-shock proteins in *E. coli*. Genes involved in transcription elongation and termination, *nusA*, *nusB*, *nusG*, *greA* and *rho*, are also present. The gene encoding σ^{32} is absent in most other small genomes, such as those of *B. burgdorferi*, *Helicobacter pylori*, *M. genitalium* and *C. trachomatis*, although genes for heat-shock proteins are present.

An unusually large number of genes involved in RNA degradation are found in *R. prowazekii*. Of these, only four appear to be common to the bacterial genomes analysed so far (those encoding polyribonucleotide nucleotidyltransferase and ribonucleases HII, III and P). Four more ribonucleases (D, E, HI and PH) are present in *R. prowazekii*, but in none of the other small parasites.

Of the 33 identified tRNA genes, which code for 32 different tRNA isoacceptor species, two code for $tRNA^{Phe}$. There are two tRNA species for most of the amino acids that are encoded by four-codon boxes; the exceptions are the four-codon boxes for proline and valine, for which we have identified only one isoacceptor-tRNA species, with U in the first anticodon position. *selC*, which codes for $tRNA^{Sec}$, and *selABD* are missing. *R. prowazekii* has a set of genes coding for tRNA modifications (*tgt*, *queA*, *trmD*, *truA*, *truB* and *miaA*) which resembles that of *Helicobacter pylori*, *C. trachomatis* and *B. burgdorferi*; *M. genitalium* has only *trmD* and *truA*.

In *R. prowazekii*, 21 genes encode 18 of the 20 aminoacyl-tRNA synthetases normally required for protein synthesis. There are two genes (*gltX*) encoding glutamyl-tRNA synthetase. As seen in several bacterial genomes[25], the gene coding for glutaminyl-tRNA synthetase, *glnS*, is missing. Three genes encoding subunits of the glutamyl-tRNA amidotransferase are present, indicating that a glutamyl-tRNA charged with glutamic acid may be transamidated to generate Gln-tRNA. The gene coding for asparaginyl-tRNA synthetase, *asnS*, is also missing from the *R. prowazekii* genome as well as from *Helicobacter pylori*, *C. trachomatis* and *A. aeolicus*[26]. A transamidation process to form Asn-$tRNA^{Asn}$ from Asp-$tRNA^{Asn}$ has been proposed for the archaeon *Haloferax volcanii*[27] and this reaction may also occur in *R. prowazekii*. The valyl-tRNA synthetase is 38.3% identical to its homologue in *Methanococcus jannaschii*, but only 27.6% identical to its most similar homologue in bacteria, which is found in *Bacillus stearothermophilus*, possibly indicating a horizontal transfer event. The lysyl-tRNA synthetase (encoded by *lysS*) in *R. prowazekii* is a class I enzyme with no resemblance to the conventional class II lysyl-tRNA synthetases. Class I type of lysyl-tRNA synthetases have been observed previously in only *B. burgdorferi*, *Pyrococcus woesii*, *Methanococcus jannaschii* and a few other methanogens[26].

Regulatory systems

As in other genomes of small parasites, *R. prowazekii* has a reduced set of regulatory genes. There are a few members of two-component regulatory systems, such as the proteins encoded by *barA*, *envZ*, *ntrY*, *ntrX*, *ompR* and *phoR*. *spoT*, which is involved in the stringent response, has been identified in *B. burgdorferi*, *Helicobacter pylori* and *M. genitalium*. Only remnants of genes coding for amino-terminal fragments of proteins similar to those encoded by *spoT* and *relA* are identifiable in *R. prowazekii*. No fragments of *spoT* encoding the carboxy-terminal segments of the protein have been identified in the genome.

Cell division and protein secretion

Proteins involved in detoxification, such as superoxide dismutase, and those involved in thiophen and furan oxidation are present in *R. prowazekii*. Two genes encoding haemolysins have also been identified, and an *R. typhi* homologue of *tlyC* exhibits haemolytic activities when expressed in *E. coli* (S. Radulovic, J. M. Troyer, B. Noden, S.G.E.A. and A. Azad, unpublished observations).

The data indicate that the basic mechanisms of cell division and secretion in *R. prowazekii* are similar to those in free-living proteobacteria. There is a common set of bacterial chaperones (encoded by *dnaK*, *dnaJ*, *hslU*, *hslV*, *groEL*, *groEL*, *groES* and *htpG*) and genes involved in the *secA*-dependent secretory system (*secABDEFGY*, *ffH* and *ftsY*). *R. prowazekii* has a significantly larger set of genes involved in peptide secretion than does *M. genitalium*.

Membrane-protein analysis

Many studies of *R. prowazekii* have focused on outer-surface membrane proteins because of their potential importance in bacterial detection and vaccination. The superficial lipopolysaccharide (LPS) molecule is important in the pathogenesis of *R. prowazekii*. LPS consists of a polysaccharide that is covalently linked to lipid A, the biosynthesis of which is catalysed by products of *lpxABCD*, all of which are present in the *R. prowazekii* genome. These genes are clustered in *E. coli*, but *lpxA* and *lpxD* are separate from *lpxB* and *lpxC* in *R. prowazekii*. Three genes involved in the biosynthesis of the 3-deoxy-D-manno-octulosonic acid (KDO) residues reside in the *R. prowazekii* genome (*kdsA*, *kdsB* and *kdtA*). Only one gene (*rfaJ*) with a putative function in outer-core biosynthesis has been identified.

We have identified a set of genes involved in the biosynthesis of murein and diaminopimelate and a set involved in the biosynthesis of fatty acids. These includes: *fabD*, which is involved in the last step of the initiation phase of fatty-acid biosynthesis; four genes involved in the elongation cycle of fatty-acid biosynthesis (*fabFGHI*); and three genes involved in the first three steps of the synthesis of polar head groups (*cdsA*, *pssA* and *pgsA*). Finally, post-translational processing and addition of lipids to an N-terminal cysteine require the gene products prolipoprotein diacylglycerol transferase (*lgt*), prolipoprotein signal peptidase (*lspA*) and apolipoprotein:phospholipid *N*-acyl transferase (*lnt*). These are found in the genome with several genes involved in the degradation of fatty acids, such as *fadA* which encodes the 3-ketoacyl-CoA thiolase.

Virulence

The *R. prowazekii* genome contains several homologues of the *VirB* gene operon found in *Agrobacterium tumefaciens*. This gene family encodes proteins that direct the export of the T-DNA–protein complex across the bacterial envelope to the plant nuclei[28]. *R. prowazekii* has two homologues of *VirB4* and one homologue each of *VirB8*, *VirB9*, *VirB10*, *VirB11* and *VirD4*. The latter five genes are clustered with the gene *trbG*, which is involved in conjugation in *Agrobactrium tumefaciens*. Homologues of the single-stranded DNA-binding proteins VirD2 and VirE2 are missing. In *Agrobacterium tumefaciens*, these proteins are bound to the transferred T-DNA, indicating different functions for the homologues of the *VirB* genes in *R. prowazekii*. Indeed, VirB proteins are homologous to components of the *E. coli* transport system for plasmids, as well as to components of the Pt1 transport machinery in *Bordetella pertussis*, which exports pertussis toxin[28]. A set of genes coding for VirB4 and several other VirB proteins has been identified in the *cag* pathogenicity island of *Helicobacter pylori*. In this species, the VirB proteins facilitate export of a factor that induces interleukin-8 secretion in gastric epithelial cells[28]. Thus, *R. prowazekii* may encode components of a transport system for both conjugal DNA transfer and protein export.

The virulence of *Staphylococcus aureus* has been correlated with the production of capsular polysaccharides in phagocytic assays and mouse lethality assays[29,30]. A cluster of ten capsule genes (*capA–M*) is involved in capsule biosynthesis in *S. aureus* strain M[31]. We have identified three *R. prowazekii* genes with sequence similarities to *S. aureus cap* genes. Two of these (*capD* and *capM*) are separated by ten

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998

genes, most of which are unknown genes or genes involved in the biosynthesis of LPS or techoic acid. Thus, *R. prowazekii* may produce components of a microcapsular layer that is involved in virulence.

Reductive evolution

Genome sequences of organisms enjoying an endosymbiotic lifestyle are at risk. The activities of homologous nuclear genes may render genes of the endosymbiont expendable and as a consequence they become vulnerable to obliteration by mutation. Good candidates for such purged genes in *Rickettsia* and mitochondria are genes required for amino-acid biosynthesis, nucleoside biosynthesis and anaerobic glycolyis. These and other genes would have been deleted when an ancestral genome first lived in a nucleated cell. Once genes essential to a free-living mode are lost, the endosymbiont becomes an obligate resident of its host.

Likewise, small, bottle-necked populations of bacteria infecting a eukaryotic cell will tend to accumulate deleterious mutations because selection cannot remove them from such clonal populations[13]. The accumulation of such harmful but non-lethal mutations is referred to as 'Muller's ratchet'[32] or 'near-neutral evolution'[33,34]. The consequence of accumulation of these mutations will be the inactivation and eventual deletion of non-essential genes.

The first mutation that inactivates an expendable gene is likely to initiate a sequence of events in which subsequent mutations freely transform it, by degrees, from a pseudogene, to unrecognizable sequence, to small fragments, to extinction. In this sequence, mutations are released from amino-acid-coding constraints. Thus nucleotide substitutions will reflect the mutation bias of the genome. This bias can be estimated roughly by frequencies of third-position bases in the codons. For *R. prowazekii*, the bias of the third-position bases is 18% G+C rather than the 29% G+C average for the genome. So, as sequences age in *R. prowazekii*, their composition should gradually approach the low G+C content of third codon positions. Nearly one-quarter of the *R. prowazekii* genome is composed of non-coding sequences, with a G+C content lower than that of coding sequences (25% G+C compared to 30%; $P < 0.001$). Thus, much of the non-coding sequence may be remnants of coding sequences that are in the process of being eliminated from the genome.

The gene encoding *S*-adenosylmethionine synthetase (*metK*), which catalyses the biosynthesis of *S*-adenosylmethionine (SAM), illustrates the initiation of this process. The *metK* sequence in the strain of *R. prowazekii* studied here has a termination codon within a region of the gene that is otherwise highly conserved among bacterial species[35]. However, a closely related strain does not have the termination codon. Many other defects, such as termination codons, insertions, and a preponderance of small deletions, have also been observed in the *metK* genes in several members of the spotted fever group *Rickettsia* (J.O.A. and S.G.E.A., unpublished observations). This random distribution of lethal mutations among some *metK* alleles from different *Rickettsia* species indicates that the gene may have just entered the extinction process. This distribution, and the identification of 11 more pseudogenes for carboxypeptidase (*ypwA*), penicillin-binding protein (*pbpC*), succinyl CoA-transferase (*scoB*), transposase (*tra3*), resolvase (*pin*), conjugative transfer protein (*taxB*), a hypothetical protein (*yfc*1) and four different fragmented open reading frames for (p)ppGpp 3′-pyrophosphohydrolase, indicates that the *R. prowazekii* genome continues to eliminate genes.

Genome sequences can be purged by a more abrupt mechanism. This consists of intrachromasomal recombination at duplicated sequences, which can result in the deletion of intervening sequences, the loss of a sequence duplication and the rearrangement of flanking

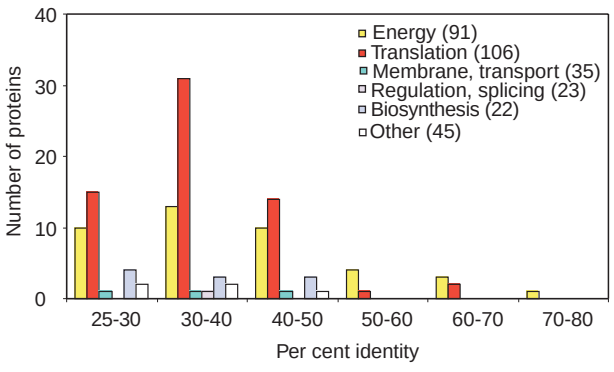

Figure 4 Histogram representation of the similarity of predicted *R. prowazekii* proteins to yeast proteins targeted to the mitochondria. Only protein pairs with per cent identity values greater than 25% are shown. Numbers in parentheses represent the total number of yeast mitochondrial proteins within each category. The yeast mitochondrial protein sequences have been taken from http://www.proteome.com.

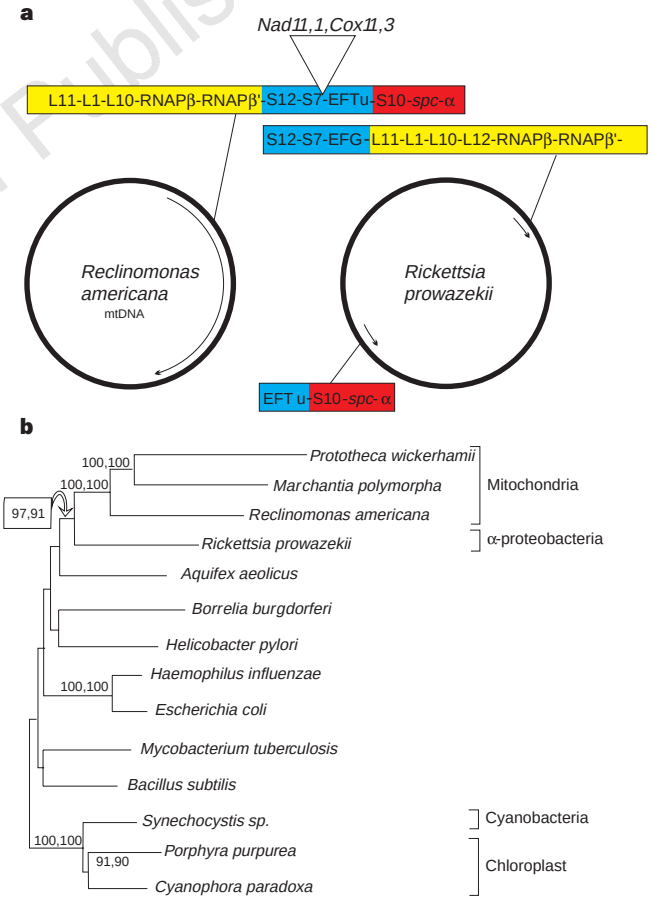

Figure 5 The organization and phylogenetic relationships of gene encoding ribosomal protein from *R. prowazekii* and the mitochondrial genome of *Reclinomonas americana*. **a**, The organization of ribosomal-protein genes. The S10, *spc* and α-operons are organized similarly in these two genomes, except that several ribosomal-protein genes[38] have been deleted from the mitochondrial genome of *Reclinomonas americana*. **b**, The phylogenetic relationships of mitochondria and bacteria were derived from the combined amino-acid sequences of ribosomal proteins S2, S3, S7, S10, S11, S12, S13, S14, S19, L5, L6 and L16. Neighbour-joining and maximum-parsimony methods gave identical topologies. Branch lengths are proportional to those reconstructed by using the neighbour-joining method. Values at nodes are bootstrap values indicating the degree of support for individual clusters under each method (neighbour-joining, maximum parsimony). Only bootstrap values >90% are shown.

Nature © Macmillan Publishers Ltd 1998

sequences. Such a mechanism will account for the presence in *R. prowazekii* of one, unlinked copy of *rrs* and *rrl*, both of which are surrounded by new flanking sequences[36]. Likewise, *R. prowazekii* has one *tuf* gene and one *fus* gene in atypical clusters that seem to have been created by intrachromosomal recombination between the two *tuf* genes that are normally found in Gram-negative bacteria[37]. Indeed, rearranged gene operon structures encoding ribosomal proteins are characteristic of all members of the genus *Rickettsia* (H. Amiri, C.A. and S.G.E.A., unpublished observations).

Conserved operons that are found in free-living bacteria are often dispersed throughout the *Rickettsia* genome (see above). The *R. prowazekii* genome contains an unusually small fraction of repeat sequences (<10% of that observed in free-living bacteria). We suggest that the repeat sequences found in the ancestor to the *Rickettsia* have been 'consumed' by the intrachromosomal-recombination mechanism that generated some of the deletions and rearrangements seen in *R. prowazekii*. Such intrachromosomal recombinants arise at a substantial rate in bacteria growing in culture, but here they are eliminated from the populations by selection. That such remnants of intrachromosomal recombination are retained in *R. prowazekii* indicates that purifying selection has been attenuated in this organism.

Mitochondrial affinities

The reduction in genome size in mitochondria and *Rickettsia* is likely to have occurred independently in the two lineages. Most of the genes supporting mitochondrial activities are nuclear. Many of the 300 proteins encoded in the nucleus of the yeast *Saccharomyces cerevisiae* but destined for service within the mitochondrion are close homologues of their counterparts in *R. prowazekii*. Nearly one-quarter of these proteins are required for bioenergetic processes and another one-third of them are required for the expression of the genes encoded in the mitochondrial genome. In total, more than 150 nucleus-encoded mitochondrial proteins share significant sequence homology with *R. prowazekii* proteins (Fig. 4).

Another group of 58 nucleus-encoded mitochondrial proteins represents components of the mitochondrial transport machinery and regulatory system (Fig. 4). These include proteins found in the mitochondrial outer membrane and others involved in splicing reactions. Such proteins have probably been secondarily recruited to mitochondria from genomes not necessarily related to that of the α-proteobacterial ancestor.

The mitochondrial genome of the early diverging, freshwater protozoan *Reclinomonas americana* is more like that of a bacterium than any other mitochondrial genome sequenced so far[38]. This genome contains 67 protein-coding genes, most of which provide components of genetic processes and the bioenergetic system[38]. Several gene clusters in this mitochondrial genome are reminiscent of those in bacteria (Figs 5a, 6a). Most similarities represent retained, ancestral traits present in the common ancestor of bacteria and mitochondria. For example, the genes *rplKAJL* and *rpoBC* are identically organized in *R. prowazekii* and the mitochondrial genome of *Reclinomonas americana*. Likewise, the genes encoding the S10, *spc* and the α-ribosomal protein operons are organized similarly in the two genomes. The immediate proximity of these two clusters in the *Reclinomonas americana* mitochondrial DNA is reminiscent of the arrangement in free-living bacteria, whereas the physical separation of the two clusters in the *R. prowazekii* genome is atypical. A further rearrangement event is indicated by the fact that the *rpsLrpsGfus* cluster is located upstream of the *rplKAJLrpoBC* cluster in *R. prowazekii*, rather than downstream as it is in the *Reclinomonas americana* mtDNA. Phylogenetic reconstructions based on ribosomal proteins within each of these two clusters indicate that there is a close evolutionary relationship between *R. prowazekii* and mitochondria (Fig. 5b).

Mitochondria and *R. prowazekii* have a similar repertoire of proteins involved in ATP production and transport, including genes encoding components of the TCA cycle, the respiratory-chain complexes, the ATP-synthase complexes and the ATP/ADP translocases. There are some similarities in the gene orders of some functional clusters (Fig. 6a). There are also some rearrangements of clusters that are specific to *Rickettsia*. One example is the inversion of segments corresponding to *nuoJKLM* and *nuoGHI*. Another is the scattered displacement of genes involved in the biogenesis of cytochrome *c*. Nevertheless, phylogenetic reconstructions based on components of the NADH dehydrogenase complexes indicate that there is a close evolutionary relationship between *R. prowazekii* and mitochondria (Fig. 6b).

We have identified as many as five genes coding for ATP/ADP transporters, all of which are expressed (R.M.P. *et al.*, unpublished observations). The *Rickettsia* ATP/ADP translocases are monomers with 12 transmembrane regions each, whereas the mitochondrial translocates are dimers with six transmembrane regions per dimer. We found no relationship between the primary structures of the mitochondrial and *Rickettsia* ATP/ADP translocases, indicating that these transport systems may have originated independently.

The study of the *R. prowazekii* genome sequence supports the idea that aerobic respiration in eukaryotes originated from an ancestor of the *Rickettsia*, as indicated previously by phylogenetic reconstructions based on the rRNA gene sequences[7,9]. Phylogenetic analyses of the *petB* and *coxA* genes indicate that the respiration systems of *Rickettsia* and mitochondria diverged ~1,500–2,000 million years ago[10], shortly after the amount of oxygen in the atmosphere began

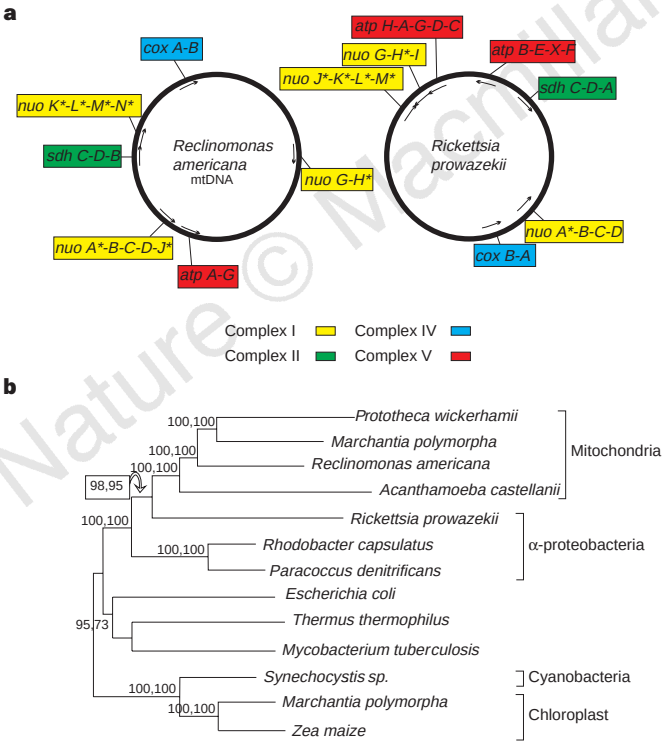

Figure 6 The organization and phylogenetic relationships of genes involved in ATP synthesis from *R. prowazekii* and the mitochondrial genome of *Reclinomonas americana*. **a**, The organization of bioenergetic genes. **b**, The phylogenetic relationships of mitochondria and bacteria were derived from the combined amino-acid sequences of NADH dehydrogenase I chains A, J, K, L, M and N, which are encoded by the genes *nuoA*, *J*, *K*, *L*, *M*, *N*. These genes are highlighted by asterisks in **a**. Neighbour-joining and maximum-parsimony methods gave identical topologies. Branch lengths are proportional to those reconstructed using the neighbour-joining method. Values at nodes are bootstrap values indicating the degree of support for individual clusters under each method (neighbour-joining, maximum parsimony). Only bootstrap values >90% are shown.

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998

to increase. The finding that the ATP/ADP translocases in *R. prowazekii* and mitochondria are of different evolutionary origin is problematic (R.M.P. *et al.*, unpublished observations). Free-living bacteria do not seem to have homologues of ATP/ADP translocases, which are found only in organelles and in two obligate intracellular parasites, *Rickettsia* and *Chlamydia*. Thus it is not known whether the original endosymbiont was capable of efficient exchange of adenosine nucleotides with its host cell. More detailed comparative analysis of the genomes of α-proteobacteria may refine our understanding of the origins of mitochondria. □

Methods

Genome sequencing. We prepared genomic DNA from the Madrid E strain of *R. prowazekii*, which was originally isolated in Madrid from a patient who died in 1941 with epidemic typhus. We propagated *R. prowazekii* in the yolk sac of embryonated hen eggs and purified DNA according to standard procedures[39]. We sequenced the *R. prowazekii* genome by a whole-genome shotgun approach in combination with shotgun sequencing of a selected set of clones from a cosmid library (A.Z. *et al.*, unpublished observations). Genomic and cosmid DNA was sheared by nebulization to an average size of ~2 kb. The random fragments were cloned into a modified M13 vector using the double adaptor method[40]. We collected 19,078 sequence reads during the random sequencing phase using Applied Biosystems 377 DNA sequencers (Perkin-Elmer).

The sequences were assembled and the consensus sequence was edited using the STADEN program[41]. We verified the structure of the assembled sequence by end-sequencing of 3-kb-insert λ Zap II clones[36], 10-kb λ clones and 30-kb cosmid clones. More than 97% of the genome was covered by clones from the three different libraries (A.Z. *et al.*, unpublished observations). Gaps between contigs were closed by direct sequencing of clones from the three libraries or of polymerase chain reaction (PCR) products. The final four gaps were closed by direct sequencing of PCR products generated with the Long Range PCR system (Gene Amp). Regions of ambiguity were identified by visual inspection of the assembly and resequenced. The final assembly contains ~20,000 sequences. The genome sequence has eightfold coverage on average and no single region has less than twofold coverage. We estimate the overall error frequency to be $< 1 \times 10^{-5}$.

Informatics. Sequence analysis and annotation was managed by CapDB (T.S.-P. *et al.*, unpublished observations). We identified open reading frames of more than 50 codons as genes on the basis of their characteristic patterns in nucleotide-frequency statistics[14] using BioWish[42]. The identified genes were analysed using the program BLASTX[43] to search for sequence similarities in EMBL, TREMBL, SwissProt and in-house databases. We identified tRNA genes with the program tRNA scan-SE[44]. Remaining frameshifts were considered to be authentic and annotated as pseudogenes. Families of paralogues were constructed using BLAST to search for sequence similarities within the *R. prowazekii* genome. Multiple alignments and phylogenetic trees for genes with significant sequence similarities to genes in the public databases were constructed automatically using CLUSTAL-W[45], Phylo_win[46] and GRS[47]. The final annotation was based on manual inspection of the phylogenetic placement of *R. prowazekii* in the resulting gene trees.

Received 21 July; accepted 24 September 1998.

1. Gross, L. How Charles Nicolle of the Pasteur Institute discovered that epidemic typhus is transmitted by lice: reminiscences from my years at the Pasteur Institute in Paris. *Proc. Natl Acad. Sci. USA* **93,** 10539–10540 (1996).
2. Weisburg, W. G., Woese, C. R., Dobson, M. E. & Weiss, E. A common origin of Rickettsiae and certain plant pathogens. *Science* **230,** 556–558 (1985).
3. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51,** 221–227 (1987).
4. Weisburg, W. G. *et al.* Phylogenetic diversity of the rickettsias. *J. Bacteriol.* **171,** 4202–4206 (1989).
5. Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. Mitochondrial origins. *Proc. Natl Acad. Sci. USA* **82,** 4443–4447 (1985).
6. Gray, M. W., Cedergren, R., Abel, Y. & Sankoff, D. On the evolutionary origin of the plant mitochondrion and its genome. *Proc. Natl. Acad. Sci. USA* **86,** 2267–2271 (1989).
7. Olsen, G. J., Woese, C. R. & Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176,** 1–6 (1994).
8. Viale, A. & Arakaki, A. K. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett.* **341,** 146–151 (1994).
9. Gray, M. W. & Spencer, D. F. in *Evolution of Microbial Life* (eds Roberts, D. M., Sharp, P. M., Alderson, G. & Spencer, D. F.) 109–126 (Cambridge Univ. Press, Cambridge, 1996).
10. Sicheritz-Pontén, T., Kurland, C. G. & Andersson, S. G. E. A phylogenetic analysis of the cytochrome *b* and cytochrome *c* oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. *Biochim. Biophys. Acta* **1365,** 545–551 (1998).
11. Margulis, L. *Origin of Eukaryotic Cells* (Yale Univ. Press, New Haven, 1970).
12. Kurland, C.. G. Evolution of mitochondrial genomes and the genetic code. *Bioessays* **14,** 709–714 (1992).
13. Andersson, S. G. E. & Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6,** 263–268 (1998).
14. Andersson, S. G. E. & Sharp, P. M. Codon usage and base composition in *Rickettsia prowazekii*. *J. Mol. Evol.* **42,** 525–536 (1996).
15. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–511 (1995).
16. Fraser, C. M. *et al.* The *Mycoplasma genitalium* genome reveals a minimal gene complement. *Science* **270,** 397–403 (1995).
17. Himmelreich, R. *et al.* Complete genome sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **3,** 109–136 (1996).
18. Fraser, C. M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390,** 580–586 (1997).
19. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547 (1997).
20. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1462 (1997).
21. Deckert, G. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392,** 353–358 (1998).
22. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).
23. Fraser, C. M. *et al.* Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281,** 375–388 (1998).
24. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13,** 660–665 (1996).
25. Ebeling, S., Kündig, C. & Hennecke, H. Discovery of a rhizobial RNA that is essential for symbiotic root nodule development. *J. Bacteriol.* **173,** 6373–6382 (1991).
26. Koonin, E. V. & Aravind, L. Genomics: re-evaluation of translation machinery evolution. *Curr. Biol.* **8,** 266–269 (1998).
27. Curnow, A. W., Ibba, M. & Soll, D. tRNA-dependent asparagine formation. *Nature* **382,** 589–590 (1996).
28. Christie, P. J. *Agrobacterium tumefaciens* T-complex transport apparatus: a paradigm for a new family of multifunctional transporters in eubacteria. *J. Bacteriol.* **179,** 3085–3094 (1997).
29. Melly, M. A., Duke, L. J., Liau, D.-F. & Hash, J. H. Biological properties of the encapsulated *Staphylococcus aureus*. *M. Infect. Immun.* **10,** 389–397 (1974).
30. Peterson, P. K., Wilkinson, B. J., Kim, Y., Schmeling, D. & Quie, P. G. Influence of encapsulation on staphylococcal opsonization and phagocytosis by human polymorphonuclear leukocytes. *Infect. Immun.* **19,** 943–949 (1978).
31. Lin, W. S., Cunnen, T. & Lee, C. Y. Sequence analysis and molecular characterization of genes required for the biosynthesis of type I capsular polysaccharide in *Staphylococcus aureus*. *J. Bacteriol.* **176,** 7005–7016 (1994).
32. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78,** 157–159 (1977).
33. Ohta, T. & Kimura, M. On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1,** 18–25 (1971).
34. Ohta, T. Evolutionary rate of cistrons and DNA divergence. *J. Mol. Evol.* **1,** 150–157 (1972).
35. Andersson, J. O. & Andersson, S. G. E. Genomic rearrangements during evolution of the obligate intracellular parasite *Rickettsia prowazekii* as inferred from an analysis of 52 015 bp nucleotide sequence. *Microbiology* **143,** 2783–2795 (1997).
36. Andersson, S. G. E., Zomorodipour, A., Winkler, H. H. & Kurland, C. G. Unusual organization of the rRNA genes in *Rickettsia prowazekii*. *J. Bacteriol.* **177,** 4171–4175 (1995).
37. Andersson, S. G. E. & Kurland, C. G. Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.* **73,** 775–787 (1995).
38. Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387,** 493–497 (1997).
39. Winkler, H. H. Rickettsia permeability: an ATP/ADP transport system. *J. Biol. Chem.* **251,** 389–396 (1976).
40. Andersson, B. *et al.* A 'double adaptor' method for improved shotgun library construction. *Anal. Biochem.* **236,** 107–113 (1996).
41. Staden, R. The Staden sequence analysis package. *Mol. Biotech.* **5,** 233–241 (1996).
42. Sicheritz-Pontén, T. BioWish: a molecular biology command extension to Tcl/Tk. *Comput. Appl. Biosci.* **13,** 621–622 (1997).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
44. Low, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
45. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
46. Galtier, N., Gouy, M. & Gautier, C. SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12,** 543–548 (1996).
47. Sicheritz-Pontén, T. & Andersson, S. G. E. GRS: a graphic tool for genome retrieval and segment analysis. *Microb. Comp. Genomics* **2,** 123–139 (1997).

Acknowledgements. We thank C. Woese for discussions; M. Andersen for computer system support; and B. Andersson, K. Andersson, I. Tamas, B. Canbäck, A. Jamal, H. Amiri and S. Jossan for technical advice and assistance. This work was supported by the Swedish Foundation for Strategic Research, the Swedish Natural Sciences Research Council, the Knut and Alice Wallenberg Foundation and the European Commission.

Correspondence and requests for materials should be addressed to C.G.K. (e-mail: chuck@xray.bmc.uu.se).

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998

Table 1. Functional classification of *Rickettsia prowazekii* protein-coding genes. Gene numbers correspond to those in Fig. 2. Percentages represent per cent identitites.

AMINO ACID METABOLISM 6

Amino acid biosynthesis

RP743	*glyA*	serine hydroxymethyl transferase (B-*Mex*, 60.9%)
RP428	*ilvE*	branched-chain amino acid aminotransferase (B-*Eco*, 36.8%)

Amino acid degradation

RP091	*aatA*	aspartate aminotransferase (B-*Rme*, 55.6%)
RP618	*pccA*	propionyl-CoA carboxylase α chain (E-*Rno*, 45.0%)
RP619	*pccB*	propionyl-CoA carboxylase β chain (E-*Hsa*, 63.3%)
RP449	*tdcB*	threonine dehydratase (E-Yeast, 35.3%)

BIOSYNTHESIS OF COFACTORS. 25

Folic acid

RP536	*folC*	folylpolyglutamate synthetase (B-*Bsu*, 34.5%)
RP515	*folD*	methylenetetrahydrofolate dehydrogenase (*Bsu*, 46.3%)
RP383	*folE*	GTP cyclohydrolase I (B-*Syn*, 48.1%)

Haem and porphyrins

RP841	*hemA*	delta-aminolevulinate synthase (B-*Bja*, 49.1%)
RP539	*hemB*	delta-aminolevulinate dehydratase (B-*Bja*, 53.3%)
RP466	*hemC*	prophobilinogen deaminase (B-*Pmi*, 39.1%)
RP885	*hemE*	uroporhyrinogen decarboxylase (B-*Rca*, 42.6%)
RP882	*hemF*	coproporhyrinogen III oxidase (E-*Gma*, 43.0%)
RP884	*hemH*	ferrochelatase (B-*Syn*, 40.4%)
RP847	*hemK*	protoporphyrinogen oxidase (B-*Eco*, 44.3%)
RP175	*hemN*	oxygen-independent coproporphyrinogen II (B-*Bsu*, 34.4%)

Lipoate

RP742	*lipA*	lipoic acid synthetase (B-*Hin*, 50.5%)
RP876	*lipB*	lipoic acid ligase (B-*Mtu*, 35.6%)

Menaquinone and ubiquinonens

RP190	*coq7*	ubiquinone biosynthesis prt Coq7 (E-*Rno*, 36.9%)
RP479	*ispB*	octaprenyl-diphosphate synthase (B-*Eco*, 36.5%)
RP886	*ubiA*	4-hydroxybenzoate octaprenyltransferase (B-*Eco*, 36.1%)
RP541	*ubiX*	3-octaprenyl-4-hydroxybenzoate carboxylase (B-*Eco*, 53.2%)
RP680	*ubiE*	ubiquinone biosynthesis methyltransferase (B-*Eco*, 44.0%)
RP622	*ubiG*	3-demethylubiquinone methyltransferase (B-*Eco*, 39.1%)

Thio- and glutaredoxin

RP204	*grxC1*	gluteredoxin 3 (B-*Eco*, 50.0%)
RP745	*grxC2*	gluteredoxin 3 (B-*Syn*, 50.0%)
RP327	*tdpX1*	thioredoxin-peroxidase (B-*Hpy*, 54.0%)
RP002	*trxA*	thioredoxin (B-*Ani*, 52.8%)
RP445	*trxB1*	thioredoxin reductase (B-*Hin*, 52.0%)
RP514	*trxB2*	thioredoxin reductase (B-*Cpa*, 28.4%)

CELL ENVELOPE 59

Diaminopimelate

RP316	*asd*	aspartate-semialdehyde dehydrogenase (B-*Vch*, 43.3%)
RP429	*dapA*	dihydrodipicolinate synthase (A-*Mja*, 39.6%)
RP148	*dapB*	dihydrodipicolinate reductase (B-*Hpy*, 37.7%)
RP194	*dapD*	tetrahydrodipicolinate N-succinyltransferase (B-*Eco*, 57.9%)
RP874	*dapE*	succinyl-diaminopimelate desuccinylase (B-*Hin*, 37.5%)
RP415	*dapF*	diaminopimelate epimerase (B-*Hin*, 35.0%)
RP753	*lysC*	aspartokinase (B-*Bst*, 37.3%)

Membranes and lipoproteins

RP347	*asmA*	outer membrane assembly protein (B-*Eco*, 19.3%)
RP446	*lgtD*	prolipoprotein diacylglyceryl transferase (B-*Vch*, 29.1%)
RP366	*lnt*	apolipoprotein N-acyltransferase (B-*Hin*, 29.1%)
RP300	*nlpD*	lipoprotein (B-*Hin*, 22.4%)
RP390	*rlpA*	rare lipoprotein A (B-*Hin*, 23.9%)
RP224	*tolC*	outer membrane protein (B-*Eco*, 22.9%)
RP048	*yidC*	inner membrane protein, 60 kDa (B-*Hin*, 30.4%)

Murein sacculus

RP095	*alr*	alanine racemace (B-*Hin*, 29.5%)
RP389	*dacF*	penicillin binding protein precursor (B-*Bsu*, 33.8%)
RP249	*ddlB*	D-alanine:D-alanine ligase (B-*Hin*, 32.8%)
RP454	*glmU*	UDP-N-acetylglucoseamine pyrophosphorylase (B-*Hin*, 34.3%)
RP595	*mraY1*	phospho-N-acetylmuramoyl-pentapeptide-transferase (B-*Hin*, 49.9%)
RP825	*mraY2*	phospho-N-acetylmuramoyl-pentapeptide-transferase (B-*Sac*, 22.0%)
RP807	*mrcA*	penicillin binding protein 1A (B-*Eco*, 35.6%)
RP579	*murA*	UDP-N-acetylglucosamine 1-carboxyvinyltransferase (B-*Aca*, 51.6%)
RP248	*murB*	UDP-N-acetylenolyruvoylglucosamine reductase (B-*Bsu*, 35.7%)
RP247	*murC*	UDP-N-acetylmuramoylalanine ligase (B-*Hin*, 41.5%)
RP410	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase (B-*Hin*, 32.9%)
RP597	*murE*	UPD-MurNac-tripeptide synthetase (B-*Bsu*, 35.2%)
RP596	*murF*	UDP-MurNac-pentapeptide synthetase (B-*Eco*, 30.6%)
RP412	*murG*	UDP-MurNac-pentapeptide transferase (B-*Bsu*, 28.8%)
RP565	*pbpA1*	penicillin binding protein (B-*Hin*, 34.3%)
RP567	*pbpA2*	penicillin binding protein (B-*Bsu*, 30.7%)
RP195	*pbpC**	penicillin-binding protein (B-*Eco*, 26.7%)
RP259	*pbpE*	penicillin binding protein (B-*Bsu*, 25.2%)
RP400	*slt*	lytic murein transglycosidase (B-*Hin*, 21.4%)

Surface polysaccharides, lipopolysaccharides and antigens

RP333	*capD*	capsular polysaccharide biosynthesis protein CapD (B-*Sau*, 34.6%)
RP344	*capM1*	capsular polysaccharide biosynthesis protein CapM (B-*Sau*, 24.9%)
RP414	*capM2*	capsular polysaccharide biosynthesis protein CapM prt (B-*Sau*, 23.7%)
RP339	*ggaB*	galactosamine-containing techoic acid biosynthesis (B-*Bsu*, 23.3%)
RP718	*htrB*	lauroyl acetyltransferase (B-*Hin*, 21.5%)
RP007	*lpxA*	UDP-N-acetylglucosamine acyltransferase (B-*Rri*, 90.4%)
RP321	*lpxB*	lipid A disaccharide synthetase (B-*Hin*, 27.3%)
RP254	*lpxC*	UDP-3-*O*-acyl N-acetylglucosamine deacetylase (B-*Eco*, 44.4%)
RP009	*lpxD*	UDP-3-O-(R-3-hydroxymyristoyl)-glucosamine N-acetyltransferase (B-*Rri*, 92.4%)
RP062	*kdsA*	3-deoxy-d-manno-octulosonic acid 8-phosphate synthetase (B-*Pha*, 45.1%)
RP379	*kdsB*	CTP:CMP-3-deoxy-manno-octulosonate-cytidylyl transferase (B-*Hin*, 34.8%)
RP089	*kdtA*	3-deoxy-D-manno-octulosonic acid transferase (B-*Eco*, 28.9%)
RP505	*kpsF*	polysiliac acid capsule expression protein (B-*Eco*, 36.9%)
RP833	*omp*	cell surface antigen, 17 kD (B-*Rty*, 46.9%)
RP160	*omp1*	OMP1 precursor (B-*Bab*, 29.5%)
RP771	*pal*	peptidoglycan-associated lipoprotein (B-*Eco*, 37.9%)
RP476	*rfaJ*	lipopolysaccharide 1,2-glucosyltransferase (B-*Sty*, 26.1%)
RP004	*rfbA*	O-antigen export system permease (B-*Kpn*, 22.3%)
RP003	*rfbE*	O-antigen ABC export system, ATP-binding protein (B-*Yen*, 34.0%)
RP334	*rffE*	UDP-N-acetylglucosamine 2-epimerase (B-*Bsu*, 26.8%)
RP018	*sca1*	cell surface antigen (B-*Rri*, 24.9%)
RP081	*sca2*	cell surface antigen (B-*Rri*, 27.4%)
RP451	*sca3*	cell surface antigen (B-*Rri*, 27.6%)
RP498	*sca4*	cell surface antigen (B-*Rja*, 67.4%)
RP704	*sca5*	cell surface antigen (B-*Rri*, 72.5%)
RP779	*udg*	UDP-glucose 6-dehydrogenase (B-*Pae*, 31.8%)

CELLULAR PROCESSES 44

Cell division

RP251	*ftsA*	cell division protein FtsA (B-*Hin*, 29.5%)
RP043	*ftsH*	cell division protein FtsH (B-*Eco*, 54.0%)
RP163	*ftsJ*	cell division protein FtsJ (B-*Eco*, 44.4%)
RP823	*ftsK*	cell division protein FtsK (B-*Cbu*, 41.5%)
RP250	*ftsQ*	cell division protein FtsQ (B-*Hin*, 17.9%)
RP411	*ftsW*	cell division protein FtsW (B-*Eco*, 33.2%)
RP775	*ftsY*	cell division protein FtsY (B-*Hin*, 43.2%)
RP666	*ftsZ*	cell division protein FtsZ (B-*Wsp*, 65.3%)
RP056	*gidA*	glucose inhibited division protein A (B-*Eco*, 48.8%)
RP057	*gidB*	glucose inhibited division protein (B-*Ppu*, 26.8%)
RP815	*maf*	MAF protein (B-*Bsu*, 38.1%)
RP042	*mesJ*	cell cycle protein MesJ (B-*Eco*, 22.1%)
RP768	*mreB*	rod shape-determining protein (B-*Eco*, 60.5%)
RP767	*mreC*	rod shape-determining protein (B-*Eco*, 23.1%)
RP280	*rodA*	rod shape-determining protein (B-*Eco*, 38.1%)

Cell killing

RP555	*tlyA*	hemolysin (B-*Thy*, 34.3%)
RP740	*tlyC*	hemolysin (B-*Thy*, 28.8%)

Chaperones and stress-induced proteins

RP670	*cspA*	cold shock protein (B-*Scl*, 57.6%)
RP816	*dksA*	DnaK suppressor protein (B-*Hin*, 38.8%)
RP184	*dnaJ*	heat shock protein (B-*Bov*, 49.7%)
RP185	*dnaK*	heat shock protein 70 (B-*Rme*, 72.7%)
RP626	*groEL*	heat shock protein GroEL (B-*Rme*, 69.4%)
RP627	*groES*	heat shock protein GroES (B-*Rty*, 98.9%)
RP629	*grpE*	heat shock protein GrpE (B-*Ccr*, 39.5%)
RP200	*hscA*	heat shock protein A (B-*Hin*, 39.6%)
RP320	*hslU*	heat shock protein HslU (B-*Bsu*, 54.8%)
RP319	*hslV*	heat shock prt HslV (B-*Hin*, 54.1%)
RP273	*hsp22*	heat shock protein (E-*Phy* cp, 29.0%)
RP840	*htpG*	heat shock protein C62.5 (B-*Eco*, 43.1%)

Detoxification

RP535	*sodB*	superoxide dismutase (B-*Lpn*, 53.4%)
RP759	*thdF*	thiophene and furan oxidizer (B-*Hin*, 34.7%)

Protein and peptide secretion

RP315	*aprD*	protease secretion ATP-binding protein (B-*Pae*, 40.0%)
RP314	*aprE*	protease secretion ATP-binding protein (B-*Pae*, 32.4%)
RP173	*ffh*	signal recognition particle receptor protein (B-*Eco*, 49.6%)
RP275	*lepA*	GTP-binding membrane protein (B-*Bsu*, 57.0%)
RP116	*lepB*	signal peptidase (B-*Sty*, 37.3%)
RP575	*secA*	preprotein translocase SecA subunit (B-*Rca*, 51.8%)
RP070	*secB*	preprotein translocase SecB subunit (B-*Hin*, 30.7%)
RP586	*secD*	protein-export membrane protein (B-*Eco*, 40.4%)
RP134	*secE*	preprotein translocase SecE subunit (B-*Bsu*, 37.3%)
RP114	*secF*	protein-export membrane protein (B-*Hin*, 37.7%)
RP079	*secG*	protein-export membrane protein (B-*Hpy*, 32-0%)
RP639	*secY*	preprotein translocase SecY subunit (B-*Eco*, 50.0%)
RP842	*tig*	trigger factor (B-*Eco*, 32.0%)

ENERGY METABOLISM 67

ATP-proton motive force interconversion

RP803	*atpA*	ATP synthase F1 alpha subunit (B-*Rru*, 66.2%)
RP023	*atpB*	ATP synthase F0 subunit a (B-*Rru*, 51.5%)
RP800	*atpC*	ATP synthase F1 epsilon subunit (B-*Rru*, 24.5%)
RP801	*atpD*	ATP synthase F1 beta subunit (B-*Rca*, 77.0%)
RP022	*atpE*	ATP synthase F0 subunit c (E-*Ram* mt, 43.2%)
RP020	*atpF*	ATP synthase F0 subunit b (B-*Rru*, 21.1%)
RP802	*atpG*	ATP synthase F1 gamma subunit (B-*Rbl*, 38.0%)
RP804	*atpH*	ATP synthase F1 delta chain (E-*Osi* cp 26.4%)
RP021	*atpX*	ATP synthase F0 subunit b′ (E-*Ram* mt, 28.6%)

Electron transport

RP588	*ccmE*	cytochrome *c* biogenesis protein (B-*Hin*, 33.2%)
RP703	*ccmF*	cytochrome *c* biogenesis protein (E-*Ram* mt, 33.6%)
RP405	*coxA*	cytochrome *c* oxidase subunit I (E-*Mpo* mt, 68.9%)
RP406	*coxB*	cytochrome *c* oxidase subunit II (E-*Mpo* mt, 48.0%)
RP191	*coxC*	cytochrome *c* oxidase subunit III (E-*Pwi* mt, 48.9%)
RP257	*coxW*	cytochrome *c* oxidase assemly (E-*Sce*, 35.2%)
RP304	*cox11*	cytochrome *c* oxidase assembly (E-*Ram* mt, 42.9%)
RP346	*ctaB*	cytochrome *c* oxidase assembly factor (B-*Pde*, 40.6%)
RP253	*cycM*	cytochrome c (B-*Bja*, 35.6%)
RP216	*cydA*	cytochrome oxidase *d* subunit I (B-*Avi*, 34.0%)
RP217	*cydB*	cytochrome oxidase *d* subunit II (B-*Eco*, 30.0%)
RP272	*fbcH*	ubiquinol cytochrome *c* oxidoreductase, cytochrome c1 subunit (B-*Bja*, 47.8%)
RP829	*fdxA*	ferredoxin (*Rca*, 57.5%)
RP357	*nuoA*	NADH dehydrogenase I chain A (E-*Pwi* mt, 64.6%)
RP356	*nuoB*	NADH dehydrogenase I chain B (E-*Ram* mt, 73.2%)
RP355	*nuoC*	NADH dehydrogenase I chain C (B-*Pde*, 42.1%)
RP354	*nuoD*	NADH dehydrogenase I chain D (E-*Ram* mt, 71.4%)
RP353	*nuoE*	NADH dehydrogenase I chain E (E-*Rno*, mt, 55.1%)
RP115	*nuoF*	NADH dehydrogenase I chain F (B-*Pde*, 69.1%)
RP797	*nuoG*	NADH dehydrogenase I chain G (E-*Bta* mt, 49.3%)
RP796	*nuoH*	NADH dehydrogenase I chain H (E-*Ram* mt, 63.5%)
RP795	*nuoI*	NADH dehydrogenase I chain I (E-*Ram* mt, 71.4%)
RP790	*nuoJ*	NADH dehydrogenase I chain J (E-*Ram* mt 42.3%)
RP791	*nuoK*	NADH dehydrogenase I chain K (B-*Pde*, 61.4%)
RP792	*nuoL1*	NADH dehydrogenase I chain L (E-*Ram* mt, 45.5%)
RP283	*nuoL2*	NADH dehydrogenase I chain L (E-*Aja* cp, 27.0%)
RP282	*nuoL3*	NADH dehydrogenase I chain L (E-*Ame* mt, 17.0%)
RP793	*nuoM*	NADH dehydrogenase I chain M (E-*Ram* mt, 46.6%)
RP537	*nuoN1*	NADH dehydrogenase I chain N (E-*Ram* mt, 34.6%)
RP284	*nuoN2*	NADH dehydrogenase I chain N (E-*Ram* mt, 23.2%)
RP270	*petA*	Rieske-I iron sulphur protein (B-*Bja*, 58.3%)
RP271	*petB*	cytochrome *b* (B-*Rru*, 65.4%)
RP863	*pntAA*	NAD(P) transhydrogenase α subunit (B-*Eco*, 37.7%)
RP862	*pntAB*	NAD(P) transhydrogenase α subunit (B-*Hin*, 44.7%)
RP074	*pntB*	NAD(P) transhydrogenase β subunit (B-*Hin*, 51.5%)

Fermentation

RP110	*ackA*	acetate kinase (B-*Cts*, 38.4%)

Glycolysis

RP492	*ppdK*	pyruvate, orthophosphate dikinase (E-*Ftr*, 48.8%)

Phosphate

RP589	*ppa*	inorganic pyrophosphatase (B-*Eco*, 59.3%)

Pyruvate dehydrogenase

RP261	*pdhA*	pyruvate dehydrogenase E1 component, α subunit (E-*Ath*, 44.0%)
RP262	*pdhB*	pyruvate dehydrogenase E1 component, β subunit (E-*Sce*, 59.7%)
RP530	*pdhC*	dihydrolipoamide acetyltransferase E2 component (E-*Rno*, 45.1%)
RP460	*pdhD*	dihydrolipoamide dehydrogenase E3 component (E-*Psa*, 54.7%)
RP805	*pdhD*	dihydrolipoamide dehydrogenase E3 component (*Zym*, 51.1%)

TCA cycle

RP799	*acnA*	aconitate hydratase (B-*Lpn*, 59.1%)
RP665	*fumC*	fumarate hydratase (B-*Ror*, 63.5%)
RP844	*gltA*	citrate synthase (B-*Rty*, 97.8%)
RP265	*icd*	isocitrate dehydrogenase (B-*Tth*, 38.6%)
RP376	*mdh*	malate dehydrogenase (B-*Can*, 51.5%)
RP128	*sdhA*	succinate dehydrogenase, flavoprotein subunit (B-*Bja*, 70.0%)
RP044	*sdhB*	succinate dehydrogenase, iron-sulphur protein (E-*Ram* mt, 69.0%)
RP126	*sdhC*	succinate dehydrogenase, cytochrome b_{556} subunit (E-*Ram* mt, 39.5%)
RP127	*sdhD*	succinate dehydrogenase, subunit IV (E-*Ram* mt, 25.6%)
RP180	*sucA*	2-oxoglutarate dehydrogenase (B-*Hin*, 44.3%)
RP179	*sucB*	dihydrolipoamide succinyltransferase (B-*Eco*, 48.7%)
RP433	*sucC*	succinyl-CoA synthetase, β subunit (B-*Eco*, 52.1%)
RP432	*sucD*	succinyl-CoA synthetase, α subunit (E-*Ddi*, 70.7%)

Sugars

RP509	*exoC*	phosphomannomutase (B-*Abr*, 42.7%)
RP299	*lacA*	galactosidase acetyltransferase (B-*Mpn*, 44.4%)

FATTY ACID AND PHOSPHOLIPID METABOLISM. 25

RP620	*aas*	2-acyl-glycerol-phosphate-ethanolamine (B-*Eco*, 39.9%)
RP038	*aco1*	acyl-CoA desaturase (E-Yeast, 27.6%)
RP763	*acpP*	acyl carrier protein (B-*Lmu*, 52.6%)
RP577	*acpS*	holo-[acyl carrier protein] synthase (B-*Eco*, 38.5%)
RP533	*birA*	biotin Ac-CoA carboxylase synthase (B-*Pde*, 33.6%)
RP424	*cdsA*	phosphatidate cytidylyltransferase (B-*Eco*, 31.5%)
RP735	*fabD*	malonyl-CoA:Acyl carrier protein transacylase (B-*Bsu*, 40.3%)
RP764	*fabF*	3-oxoacyl-[acyl-carrier-protein] synthase II(B-*Eco*, 53.5%)
RP762	*fabG*	3-oxoacyl-[acyl-carrier-protein] reductase (B-*Rme*, 54.8%)
RP772	*fabH*	3-oxoacyl-[acyl-carrier-protein] synthase (B-*Rca*, 47.3%)
RP365	*fabI*	enoyl-[acyl-carrier-protein] reductase (B-*Asp*, 49.0%)
RP008	*fabZ*	3R-hydroxymyristoyl acyl carrier protein dehydratase (B-*Rri*, 91.7%)
RP737	*fadA*	acetyl-CoA acetyltransferase (B-*Pde*, 54.5%)
RP560	*fadB*	fatty oxidation complex a subunit (E-*Cel*, 30.6%)
RP442	*gpsA*	glycerol-3-phosphate dehydrogenase (B-*Eco*, 32.1%)
RP046	*lgt*	prolipoprotein diacylglycerol (B-*Eco*, 39.1%)
RP035	*phbB*	acetoacetyl-CoA reductase (B-*Zra*, 52.9%)
RP738	*phbC1*	poly-beta-hydroxybutyrate polymerase (B-*Cvi*, 22.7%)
RP820	*phbC2*	poly-beta-hydroxybutyrate polymerase (B-*Mex*, 37.4%)
RP750	*pgpA*	phosphatidylglycerophosphatase A (B-*Hin*, 31.9%)
RP049	*pgsA*	phospatidylglycerophosphate synthase (B-*Bsu*, 40.1%)
RP469	*plsC*	1-acyl-glycerol-3-phosphate acyltransferase (E-*Sce*, 23.6%)
RP242	*pssA*	phosphatidylserine synthase (B-*Hpy*, 28.8%)
RP373	*tme*	malic enzyme (B-*Hin*, 45.5%)
RP093	*vacJ*	VacJ lipoprotein precursor (B-*Sfl*, 33.8%)

PURINES, PYRIMIDINES 14

Deoxyribonucleotide metabolism

RP069	*dcd*	deoxycytidine triphosphate deaminase (B-*Hin*, 27.8%)
RP064	*dgtP*	deoxyguanosine triphosphate triphosphohydrolase (B-*Eco*, 30.1%)
RP399	*dut*	deoxyuridine 5′ triphosphate nucleotidohydrolase (B-*Hin*, 43.1%)
RP055	*ndk*	nucleoside diphosphate kinase (B-*Rsu*, 56.4%)
RP513	*nrdA*	ribonucleotide redutase α chain (E-*Sce*, 33.7%)
RP512	*nrdB*	ribonucleotide reductase β chain (E-*Mmu*, 28.5%)
RP301	*thyA*	thymidylate synthase (E-*Ddi* 47.9%)
RP684	*tmk*	thymidylate kinase (B-*Bsu*, 38.7%)

Purine ribonucleotide metabolism

RP638	*adk*	adenylate kinase (B-*Sco*, 33.8%)
RP765	*gmk*	guanylate kinase (B-*Hin*, 49.0%)
RP220	*purC*	phosphoribosylaminoimidazole-succinocarboxamide synthase (B-*Spn*, 34.0%)

Pyrimidine ribonucleotide biosynthesis

RP522	*cmk*	cytidylate kinase (B-*Eco*, 34.9%)
RP378	*pyrG*	CTP synthase (B-*Abr*, 54.7%)

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998

RP155 *pyrH* uridylate kinase (B-*Syn*, 53.3%)

REGULATORY FUNCTIONS 14
RP229 *barA* histidine kinase sensor protein (B-*Eco*, 23.2%)
RP071 *czcR* transcriptional activator protein (B-*Aeu*, 35.1%)
RP426 *envZ* histidine kinase osmolarity sensor protein (B-*Sti*, 23.6%)
RP294 *gppA* pppGpp phosphohydrolase (B-*Hpy*, 23.3%)
RP011 *nifR3* transcriptional activator nitrogen assimilation protein (B-*Abr*, 59.0%)
RP614 *ntrY* histidine kinase nitrogen sensor protein (B-*Aca*, 30.6%)
RP562 *ntrX* transcriptional activator nitrogen assimilation protein (B-*Aca*, 45.2%)
RP427 *ompR* transcriptional activator protein OmpR (B-*Rca*, 42.9%)
RP465 *phoR* histidine kinase phosphatase synthesis sensor protein (B-*Bsu*, 24.4%)
RP312 *spoT** (p)ppGpp 3´-pyrophosphohydrolase (B-*Eco*, 29.9%)
RP624 *spoT** (p)ppGpp 3´-pyrophosphohydrolase (B-*Mpn*, 27.8%)
RP625 *spoT** (p)ppGpp 3´-pyrophosphohydrolase (B-*Eco*, 48.7%)
RP705 *spoT** (p)ppGpp 3´-pyrophosphohydrolase (B-*Eco*, 31.7%)
RP517 *yhbH* sigma 54 modulation protein (B-*Bja*, 26.2%)

REPLICATION 46

Degradation of DNA
RP734 *addA* ATP-dependent nuclease (B-*Bsu*, 23.7%)
RP260 *xthA1* exodeoxyribonuclease III (B-*Eco*, 30.1%)
RP676 *xthA2* exodeoxyribonuclease III (B-*Eco*, 33.2%)
RP675 *xseA* exodeoxyribonuclease large subunit (B-*Eco*, 31.7%)
RP350 *xseB* exodeoxyribonuclease small subunit (B-*Eco*, 32.5%)

DNA replication, restriction, modification, recombination and repair
RP601 *dnaA* chromosomal replication initiation protein DnaA (B-*Eco*, 44.1%)
RP542 *dnaB* DNA helicase (E-*Osi* cp, 40.9%)
RP778 *dnaE* DNA polymerase III alpha subunit (B-*Sty*, 37.2%)
RP859 *dnaG* DNA primase (B-*Sty*, 29.0%)
RP419 *dnaN* DNA polymerase III beta subunit (B-*Ppu*, 29.9%)
RP732 *dnaQ* DNA polymerase III epsilon subunit (B-*Sty*, 46.7%)
RP865 *dnaX* DNA polymerase III gamma chain (B-*Eco*, 31.4%)
RP206 *gyrA* DNA gyrase A subunit (B-*Rsp*, 49.4%)
RP227 *gyrB1* DNA gyrase B subunit (B-*Sci*, 42.0%)
RP580 *gyrB2* DNA gyrase B subunit (B-*Ppu*, 51.5%)
RP172 *holB* DNA polymerase III, delta prime subunit (B-*Pae*, 22.3%)
RP171 *hupA* DNA binding protein HU (B-*Vpr*, 47.8%)
RP720 *lig* DNA ligase (B-*Zmo*, 45.7%)
RP777 *metK** S-adenosylmethionine synthetase (B-*Eco*, 66.3%)
RP598 *mfd* transcription-repair coupling factor (B-*Hin*, 33.9%)
RP351 *mpg* DNA-3-methyladenine glycosidase (E-*Hsa*, 29.7%)
RP880 *mutL* DNA mismatch repair protein MutL (B-*Spn*, 35.4%)
RP298 *mutS* DNA mismatch repair protein MutS (B-*Bsu*, 39.0%)
RP746 *nth* endonuclease III (B-*Eco*, 50.7%)
RP067 *parC* DNA topoisomerase IV subunit A (B-*Hin*, 39.0%)
RP711 *pin** invertase/recombinase (B-*Eco*, 38.0%)
RP776 *polA* DNA polymerase I (B-*Bca*, 37.2%)
RP540 *priA* primosomal protein replication factor (B-*Rru*, 39.7%)
RP546 *radA* DNA repair (B-*Bsu*, 46.5%)
RP761 *recA* recombination protein RecA (B-*Pde*, 71.2%)
RP029 *recF* DNA repair protein, ATP binding protein (B-*Ccr*, 30.4%)
RP593 *recG* ATP-dependent DNA helicase (B-*Eco*, 34.1%)
RP528 *recJ* single-stranded DNA-specific exonuclease (B-*Eco*, 32.8%)
RP182 *recN* recombination protein RecN (B-*Hin*, 31.6%)
RP438 *recR* recombination protein RecR (B-*Bsu*, 36.9%)
RP385 *ruvA* Holliday junction DNA helicase (B-*Pae*, 35.0%)
RP386 *ruvB* Holliday junction DNA helicase (*Pae*, 51.5%)
RP119 *ruvC* Holliday junction endodeoxyribonuclease (B-*Eco*, 36.1%)
RP836 *ssb* single-stranded binding protein (B-*Bab*, 52.6%)
RP326 *topA* DNA topoisomerase I (B-*Bsu*, 44.9%)
RP835 *uvrA* repair excision nuclease subunit A (B-*Eco*, 57.7%)
RP203 *uvrB* repair excision nuclease subunit B (B-*Hin*, 56.0%)
RP572 *uvrC* repair excision nuclease subunit C (B-*Pfl*, 36.9%)
RP447 *uvrD* DNA helicase (B-*Sau*, 43.5%)
RP817 *xerC* integrase/recombinase (B-*Bsu*, 32.2%)
RP361 *xerD* integrase/recombinase (B-*Eco*, 37.6%)

TRANSCRIPTION 20

Degradation of RNA
RP504 *pnp* polyribonucleotide nucleotidyltransferase (B-*Eco*, 48.9%)
RP117 *rnc* ribonuclease III (B-*Hpy*, 40.2%)
RP462 *rnd* ribonuclease D (B-*Eco*, 28.5%)
RP256 *rne* ribonuclease E (B-*Eco*, 35.9%)
RP726 *rnhA* ribonuclease HI (B-*Msm*, 43.4%)
RP202 *rnhB* ribonuclease HII (B-*Eco*, 44.7%)
RP611 *rnpA* ribonuclease P (B-*Mca*, 28.4%)
RP628 *rph* ribonuclease PH (B-*Hin*, 55.05%)

RNA synthesis and modification
RP861 *greA* transcription elongation factor GreA (B-*Hin*, 61.4%)
RP553 *nusA* transcription termination factor NusA (B-*Eco*, 36.9%)
RP162 *nusB* transcription termination factor NusB (B-*Eco*, 32.9%)
RP135 *nusG* transcription antitermination protein NusG (B-*Eco*, 42.2%)
RP015 *pcnB* poly (A) polymerase I (B-*Bsu*, 26.3%)
RP669 *rhlE* ATP-dependent RNA helicase (B-*Eco*, 38.3%)
RP526 *rho* transcription termination factor Rho (B-*Rsp*, 72.0%)
RP635 *rpoA* RNA polymerase alpha subunit (B-*Bpe*, 47.2%)
RP140 *rpoB* RNA polymerase beta subunit (B-*Rty*, 87.4%)
RP141 *rpoC* RNA polymerase beta´ subunit (B-*Eco*, 58.8%)
RP303 *rpoH* RNA polymerase sigma-32 factor (B-*Atu*, 52.0%)
RP858 *rpoD* RNA polymerase sigma-70 factor (B-*Rca*, 50.5%)

TRANSLATION 118

Aminoacyl-tRNA synthetases
RP856 *alaS* alanyl-tRNA synthetase (B-*Bba*, 52.7%)
RP065 *argS* arginyl-tRNA synthetase (B-*Hpy*, 33.0%)
RP145 *aspS* aspartyl-tRNA synthetase (B-*Syn*, 43.3%)
RP085 *cysS* cysteinyl-tRNA synthetase (B-*Hin*, 46.0%)
RP325 *gltX1* glutamyl-tRNA synthetase (B-*Abr*, 45.6%)
RP623 *gltX2* glutamyl-tRNA synthetase (B-*Hpy*, 40.3%)
RP850 *glyQ* glycyl-tRNA synthetase (B-*Mca*, 60.6%)
RP849 *glyS* glycyl-tRNA synthetase (B-*Bsu*, 32.9%)
RP308 *hisS* histidyl-tRNA synthetase (B-*Eco*, 38.3%)
RP620 *ileS* isoleucyl-tRNA synthetase (B-*Mtu*, 48.6%)
RP421 *leuS* leucyl-tRNA synthetase (B-*Eco*, 45.3%)
RP371 *lysS* lysyl-tRNA synthetase (B-*Bbu*, 26.3%)
RP683 *metS* methionyl-tRNA synthetase (B-*Bsu*, 46.9%)
RP417 *pheS* phenylalanyl-tRNA synthetase alpha sub (B-*Hin*, 49.2%)
RP418 *pheT* phenylalanyl-tRNA synthetase beta sub (B-*Hin*, 33.9%)
RP384 *proS* proline-tRNA synthetase (B-*Zmo*, 51.8%)
RP783 *serS* seryl-tRNA synthetase (B-*Cbu*, 47.2%)
RP221 *thrS* threonyl-tRNA synthetase (B-*Hin*, 50.6%)
RP468 *trpS* tryptophanyl-tRNA synthetase (B-*Syn*, 48.5%)
RP556 *tyrS* tyrosyl-tRNA synthetase (B-*Bca*, 38.7%)
RP687 *valS* valyl-tRNA synthetase (A-*Mja*, 38.3%)

tRNA and amino acyl-tRNA modification
RP208 *def* methionyl-tRNA deformylase (B-*Eco*, 49.4%)
RP209 *fmt* methinoyl-tRNA formyltransferase (B-*Hin*, 41.9%)
RP152 *gatA* glutamyl-tRNA (Gln) amidotransferase subunit A (B-*Mca*, 48.6%)
RP151 *gatB* glutamyl-tRNA (Gln) amidotransferase subunit B (B-*Mca*, 46.9%)
RP153 gatC glutamyl-tRNA (Gln) amidotransferase subunit C (B-Bsu, 24.7%)
RP672 *ksgA* dimethyladenosine transferase (B-*Bsu*, 35.7%)
RP510 *miaA* tRNA delta-2-isopentenylpyrophosphate (IPP) transferase (B-*Atu*, 30.7%)
RP605 *pth* peptidyl-tRNA hydrolase (B-*Hin*, 40.5%)
RP213 *queA* S-adenosylmethionine:tRNA ribosyltransferase-isomerase (B-*Hin*, 43.3%)
RP721 *tgt* tRNA-guanine transglycosylase (B-*Zmo*, 61.2%)
RP111 *trmD* tRNA (guanine-N1)-methyltransferase (B-*Eco*, 44.7%)
RP857 *truA* pseudouridylate synthase I (B-*Eco*, 40.1%)
RP501 *truB* tRNA pseudouridine 5S synthase (B-*Hin*, 37.6%)

Degradation of proteins, peptides and glycopeptides
RP036 *clpB* ATP-dependent protease, ATP binding subunit (B-*Hin*, 54.3%)
RP520 *clpP* ATP-dependent Clp protease (B-*Yen*, 67%)
RP692 *clpX* ATP-dependent protease, ATPase subunit (B-*Eco*, 62.8%)
RP228 *ctp* tail-specific protease precursor (B-*Bba*, 42.6%)
RP037 *gcp* sialoglycoprotein endopeptidase (B-*Hin*, 42.2%)
RP123 *hflC* lambda cII stability-governing protein (B-*Eco*, 33.9%)
RP122 *hflK* lambda cII stability-governing protein (B-*Vpa*, 30.3%)
RP124 *htrA* serine protease (B-*Bhe*, 37.7%)
RP186 *htrA* protease DO (E-*Sce*, 26.7%)
RP450 *lon* ATP-dependent protease LA (B-*Ccr*, 53.1%)
RP408 *lspA* lipoprotein signal peptidase (B-*Bsu*, 27.9%)
RP824 *map* methionyl aminopeptidase (B-*Sty*, 55.3%)
RP219 *mpp* mitochondrial protease (B-*Bsu*, 35.4%)
RP142 *pepA* aminopeptidase A (B-*Pae*, 36.6%)
RP174 *ppcE* peptidase II (B-*Rsn*, 32.5%)
RP281 *ptrB* protease II (B-*Eco*, 34.2%)
RP398 *sohB* protease IV (B-*Mja*, 23.9%)
RP525 *sppA* protease IV (B-*Hin*, 27.6%)

Protein modification and translation factors
RP238 *efp* elongation factor P (B-*Bsu*, 39.5%)
RP132 *fusA* elongation factor G (B-*Atu*, 68.7%)
RP814 *infA* initiation factor IF-1 (B-*Hin*, 67.1%)
RP552 *infB* initiation factor IF-2 (B-*Hin*, 42.6%)
RP531 *infC* initiation factor IF-3 (B-*Pvu*, 47.7%)
RP529 *prfA* peptide chain release factor RF-1 (B-*Bsu*, 50.1%)
RP274 *prfB* peptide chain release factor RF-2 (B-*Eco*, 50.4%)
RP435 *rbfA* ribosome binding factor A (B-*Bsu*, 31.6%)
RP693 *rimJ* ribosome protein alanine acetyltransferase (B-*Eco*, 23.2%)
RP154 *rrf* ribosome recycling factor (B-*Hin*, 43.3%)
RP397 *tlpA* thiol:disulphide interchange protein (B-*Bja*, 27.4%)
RP661 *tuf* elongation factor Tu (B-*Tcu*, 81.5%)
RP087 *tsf* elongation factor Ts (B-*Sci*, 40.7%)

Ribosomal proteins; synthesis and modification
RP137 *rplA* ribosomal protein L1 (B-*Cgr*, 50.2%)
RP656 *rplB* ribosomal protein L2 (E-*Ram* mt, 61.5%)
RP659 *rplC* ribosomal protein L3 (E-*Sce*, 44.1%)
RP658 *rplD* ribosomal protein L4 (B-*Bst*, 39.3%)
RP647 *rplE* ribosomal protein L5 (B-*Aco*, 53.6%)
RP644 *rplF* ribosomal protein L6 (B-*Bst*, 45.4%)
RP041 *rplI* ribosomal protein L9 (E-*Ppu* cp, 33.6%)
RP138 *rplJ* ribosomal protein L10 (B-*Laf*, 36.7%)
RP136 *rplK* ribosomal protein L11 (E-*Ram* mt, 45.5%)
RP139 *rplL* ribosomal protein L7/L12 (B-*Bab*, 66.9%)
RP233 *rplM* ribosomal protein L13 (B-*Sca*, 52.8%)
RP649 *rplN* ribosomal protein L14 (B-*Eco*, 69.6%)
RP640 *rplO* ribosomal protein L15 (B-*Bst*, 46.5%)
RP652 *rplP* ribosomal protein L16 (B-*Aac*, 53.3%)
RP634 *rplQ* ribosomal protein L17 (B-*Eco*, 57.5%)
RP643 *rplR* ribosomal protein L18 (B-*Bst*, 58.6%)
RP112 *rplS* ribosomal protein L19 (B-*Eco*, 58.8%)
RP609 *rplT* ribosomal protein L20 (B-*Psy*, 61.5%)
RP751 *rplU* ribosomal protein L21 (E-*Sol*, 42.3%)
RP654 *rplV* ribosomal protein L22 (B-*Eco*, 50.0%)
RP657 *rplW* ribosomal protein L23 (B-*Bst*, 46.3%)
RP648 *rplX* ribosomal protein L24 (B-*Bst*, 55.9%)
RP606 *rplY* ribosomal protein L25 (B-*Mtu*, 26.9%)
RP752 *rpmA* ribosomal protein L27 (E-*Ram* mt, 62.9%)
RP099 *rpmB* ribosomal protein L28 (B-*Mge*, 43.7%)
RP651 *rpmC* ribosomal protein L29 (B-*Bst*, 39.4%)
RP641 *rpmD* ribosomal protein L30 (B-*Mlu*, 33.3%)
RP100 *rpmE* ribosomal protein L31 (B-*Mtu*, 31.6%)
RP773 *rpmF* ribosomal protein L32 (B-*Hin*, 49.1%)
RP879 *rpmG* ribosomal protein L33 (B-*Hin*, 51.8%)
RP610 *rpmH* ribosomal protein L34 (B-*Ppu*, 65.9%)
RP608 *rpmI* ribosomal protein L35 (E-*Ppu* cp, 38.5%)
RP456 *rpmJ* ribosomal protein L36 (E-*Gth* cp, 65.8%)
RP521 *rpsA* ribosomal protein S1 (B-*Rme*, 48.6%)
RP086 *rpsB* ribosomal protein S2 (B-*Mtu*, 41.5%)
RP653 *rpsC* ribosomal protein S3 (B-*Bst*, 54.2%)
RP345 *rpsD* ribosomal protein S4 (B-*Bsu*, 43.2%)
RP642 *rpsE* ribosomal protein S5 (B-*Bst*, 50.6%)
RP039 *rpsF* ribosomal protein S6 (B-*Hin*, 30.2%)
RP131 *rpsG* ribosomal protein S7 (E-*Ram* mt, 45.2%)
RP645 *rpsH* ribosomal protein S8 (B-*Bsu*, 42.0%)
RP234 *rpsI* ribosomal protein S9 (B-*Bst*, 48.8%)
RP660 *rpsJ* ribosomal protein S10 (B-*Hin*, 60.8%)
RP636 *rpsK* ribosomal protein S11 (B-*Syn*, 53.5%)
RP130 *rpsL* ribosomal protein S12 (B-*Tth*, 60.5%)
RP637 *rpsM* ribosomal protein S13 (B-*Bst*, 58.8%)
RP646 *rpsN* ribosomal protein S14 (B-*Syn*, 47.0%)
RP503 *rpsO* ribosomal protein S15 (B-*Bst*, 53.4%)
RP878 *rpsP* ribosomal protein S16 (B-*Hin*, 45.1%)
RP650 *rpsQ* ribosomal protein S17 (B-*Tth*, 61.8%)
RP040 *rpsR* ribosomal protein S18 (B-*Syn*, 50.7%)
RP655 *rpsS* ribosomal protein S19 (B-*Eco*, 58.2%)
RP633 *rpsT* ribosomal protein S20 (B-*Rme*, 40.9%)
RP615 *rpsU* ribosomal protein S21 (B-*Bbu*, 33.3%)

TRANSPORT AND BINDING PROTEINS 38

General
RP060 *abcT1* ABC transporter, ATP-binding protein (B-*Hin*, 55.7%)
RP508 *abcT2* ABC transporter, ATP-binding protein (B-*Rme*, 48.8%)
RP214 *abcT3* ABC transporter, ATP-binding protein (B-*Hin*, 33.6%)
RP387 *msbA1* ABC transporter, ATP-binding protein (B-*Eco*, 26.2%)
RP696 *msbA2* ABC transporter, ATP-binding protein (B-*Eco*, 28.2%)

Amino acids
RP307 *atrc1* cationic amino acid transporter (E-*Mmu*, 29.7%)
RP129 *glnP* glutamine transport system permease (B-*Bsu*, 48.6%)
RP700 *glnQ1* glutamine ABC transporter, ATP-binding protein (B-*Eco*, 39.1%)
RP868 *glnQ2* glutamine ABC transporter, ATP-binding protein (B-*Bst*, 51.0%)
RP176 *gltP* glutamate-aspartate transporter (B-*Bca*, 35.2%)
RP483 *potE* putrescine-ornithine transporter (B-*Hin*, 26.9%)
RP369 *potG* putrescine ABC transporter, ATP-binding protein (B-*Mpn*, 29.2%)
RP077 *proP1* proline/betain transporter (B-*Eco*, 26.7%)
RP313 *proP2* proline/betain transporter (B-*Eco*, 24.9%)
RP375 *proP3* proline/betain transporter (B-*Eco*, 21.2%)
RP685 *proP4* proline/betain transporter (B-*Eco*, 24.0%)
RP755 *proP5* proline/betain transporter (B-*Eco*, 27.8%)
RP852 *proP6* proline/betain transporter (B-*Eco*, 34.8%)
RP881 *proP7* proline/betain transporter (B-*Eco*, 28.7%)
RP150 *yqiX* amino acid ABC transporter (B-*Bsu*, 32.4%)

Nucleosides and nucleotides
RP097 *mkl* ribonucleotide ABC transporter, ATP-binding protein (B-*Mle*, 36.2%)
RP053 *tlc1* ATP/ADP translocase (B-*Ctr*, 43.3%)
RP377 *tlc2* ATP/ADP translocase (B-*Ctr*, 35.2%)
RP477 *tlc3* ATP/ADP translocase (B-*Ctr*, 39.6%)
RP500 *tlc4* ATP/ADP translocase (B-*Ctr*, 36.3%)
RP739 *tlc5* ATP/ADP translocase (B-*Ctr*, 34.7%)

Carbohydrates, organic alcohols and acids
RP054 *glpT* glycerol-3-phosphate permease (B-*Bsu*, 37.1%)

Cations
RP834 *afuC* iron ABC transporter, ATP-binding protein (B-*Eco*, 33.9%)
RP810 *kefB* glutathionine-regulated potassium-efflux system protein (B-*Eco*, 33.9%)
RP583 *mgtE* magnesium transporter (B-*Syn*, 27.0%)

Other
RP205 *atm1* mitochondrial ABC transporter, ATP-binding protein (E-*Sce*, 43.3%)
RP794 *ccmA* haem ABC transporter A, ATP-binding protein (B-*Hin*, 35.5%)
RP268 *ccmB* haem exporter protein B (E-*Ram* mt, 20.9%)
RP830 *ccmC* haem exporter protein C (B-*Bja*, 43.7%)
RP571 *panF* panthotenate permease (B-*Hin*, 20.5%)
RP630 *perM* permease PerM homologue (B-*Hin*, 25.0%)
RP374 *sec7* transport protein Sec7 (B-*Hsa*, 29.6%)
RP576 *prsA* protein export (B-*Bsu*, 28.9%)

OTHER CATEGORIES 52

Adaptations to atypical conditions
RP708 *himA* integration host factor α (B-*Eco*, 29.5%)
RP236 *invA* invasion protein A (B-*Bba*, 42.8%)
RP590 *mviN* virulence factor Mvin protein (B-*Sty*, 32.4%)
RP717 *taxB** conjugative DNA processing (B-*Eco*, 33.5%)
RP286 *trbG* conjugal transfer (B-*Rsn*, 24.7%)
RP103 *virB4* virulence protein VIRB4 (B-*Atu*, 30.9%)
RP784 *virB4* virulence protein VIRB4 (B-*Atu*, 20.3%)
RP287 *virB8* virulence protein VIRB8 (B-*Atu*, 20.4%)
RP290 *virB9* virulence protein VIRB9 (B-*Atu*, 24.8%)
RP291 *virB10* virulence protein VIRB10 (B-*Atu*, 20.3%)
RP292 *virB11* virulence protein VIRB11 (B-*Atu*, 29.6%)
RP293 *virD4* virulence protein VIRD4 (B-*Atu*, 31.3%)

Drug and analogue sensitivity
RP170 *acrD* acriflavin resistance protein D (B-*Eco*, 31.3%)
RP475 *ampG1* AMPG protein (B-*Eco*, 31.4%)
RP668 *ampG2* AMPG prt (B-*Eco*, 26.3%)
RP781 *ampG3* AMPG prt (B-*Eco*, 27.6%)
RP603 *bcr1* bicyclomycin resistance (B-*Eco*, 21.7%)
RP698 *bcr2* bicyclomycin resistance (B-*Eco*, 18.8%)
RP243 *emrA* multidrug resistance protein A (B-*Eco*, 26.9%)
RP157 *emrB* multidrug resistance protein B (B-*Hin*, 29.3%)
RP786 *terC* tellurite resistance protein (B-*Eco*, 35.3%)

Colicin-related functions
RP302 *tolB* colicin tolerance protein (B-*Hin*, 29.8%)
RP309 *tolQ* inner membrane protein (B-*Eco*, 39.7%)
RP310 *tolR* inner membrane protein (B-*Pae*, 40.1%)

Uncategorized
RP493 *addA* adducin alpha subunit (E-*Hsa*, 32.6%)
RP199 *adx1* adrenodoxin precursor (E-*Spo*, 57.1%)
RP714 *ank2* ankyrin (E-*Hsa*, 32.7%)
RP245 *bolA* BolA protein (B-*Val*, 34.2%)
RP181 *ctaQ** thermostable carboxypeptidase (B-*Pho*, 29.1%)
RP297 *cysQ* sulphite synthesis pathway protein (B-*Eco*, 31.2%)
RP323 *cyaY* CyaY protein (B-*Ech*, 31.1%)
RP118 *era* GTP binding protein Era (B-*Bsu*, 33.6%)
RP063 *hesB1* HesB protein (B-*Ava*, 37.0%)
RP484 *hesB2* HesB protein (B-*Pbo*, 40.2%)
RP212 *n2B* N2B, ATPase protein (E-*Hir*, 27.2%)
RP485 *nifU* nitrogen fixation protein (B-*Avi*, 43.0%)
RP832 *p34* P34 protein (B-*Rri*, 91.3%)
RP602 *pat1* patatin B1 precursor protein (E-*Stu*, 22.9%)
RP317 *pkcI* protein kinase C inhibitor (B-*Abr*, 38.6%)
RP109 *ptb* phosphate buturyltransferase (B-*Cab*, 36.4%)
RP594 *scoB** succinyl-CoA:3-ketoacid-CoA transferase subunit B (B-*Mtu*, 22.0%)
RP031 *sco2* Sco yeast precursor protein (E-*Sce*, 32.6%)
RP587 *sco2* Sco yeast precursor protein (E-*Sce*, 36.6%)
RP846 *sfhB* SfhB protein (B-*Zmo*, 40.6%)
RP430 *smpB* small protein (B-*Syn*, 46.7%)
RP058 *soj* SOJ protein (B-*Bsu*, 50.4%)
RP486 *spl1* tRNA splicing protein (E-*Cal*, 58.9%)
RP487 *spl1* tRNA splicing protein (E-*Cal*, 32.3%)
RP059 *spoOJ* sporulation protein (B-*Bsu*, 40.4%)
RP239 *suhB* supressor protein (B-*Eco*, 22.8%)
RP733 *surf1* SurF1 protein (E-*Hsa*, 23.9%)
RP710 *tra3** transposase (B-*Rme*, 34.0%)

HYPOTHETICAL PROTEINS 110
Integral membrane proteins 19

NO SIMLARITY 208
Integral membrane proteins 21

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998