

# The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*

Gerard Deckert\*†, Patrick V. Warren\*†, Terry Gaasterland‡, William G. Young\*, Anna L. Lenox\*, David E. Graham§, Ross Overbeek‡, Marjory A. Snead\*, Martin Keller\*, Monette Aujay\*, Robert Huber||, Robert A. Feldman\*, Jay M. Short\*, Gary J. Olsen§ & Ronald V. Swanson\*

\* *Diversa Corporation, 10665 Sorrento Valley Road, San Diego, California 92121, USA*

‡ *Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*

§ *Department of Microbiology, University of Illinois, Urbana, Illinois 61801, USA*

|| *Lehrstuhl für Mikrobiologie, Universität Regensburg W-8400, Regensburg W-8400, Germany*

***Aquifex aeolicus* was one of the earliest diverging, and is one of the most thermophilic, bacteria known. It can grow on hydrogen, oxygen, carbon dioxide, and mineral salts. The complex metabolic machinery needed for *A. aeolicus* to function as a chemolithoautotroph (an organism which uses an inorganic carbon source for biosynthesis and an inorganic chemical energy source) is encoded within a genome that is only one-third the size of the *E. coli* genome. Metabolic flexibility seems to be reduced as a result of the limited genome size. The use of oxygen (albeit at very low concentrations) as an electron acceptor is allowed by the presence of a complex respiratory apparatus. Although this organism grows at 95 °C, the extreme thermal limit of the Bacteria, only a few specific indications of thermophily are apparent from the genome. Here we describe the complete genome sequence of 1,551,335 base pairs of this evolutionarily and physiologically interesting organism.**

Complete genome sequences have been determined for a number of organisms, including Archaea<sup>1</sup>, Bacteria<sup>2–7</sup>, and Eukarya<sup>8</sup>. Here we present and explore the genome sequence of *Aquifex aeolicus*. With growth-temperature maxima near 95 °C, *Aquifex pyrophilus* and *A. aeolicus* are the most thermophilic bacteria known. Although isolated and described only recently<sup>9</sup>, these species are related to filamentous bacteria first observed at the turn of the century, growing at 89 °C in the outflow of hot springs in Yellowstone National Park<sup>10,11</sup>. The observation of these macroscopic assemblages would later be instrumental in the drive to culture hyperthermophilic organisms<sup>12</sup>.

The *Aquificaceae* represent the most deeply branching family within the bacterial domain on the basis of phylogenetic analysis of 16S ribosomal RNA sequences<sup>13,14</sup>, although analyses of individual protein sequences vary in their placement of *Aquifex* relative to other groups<sup>15–18</sup>. The genera in this group, *Aquifex* and *Hydrogenobacter*, are thermophilic, hydrogen-oxidizing, microaerophilic, obligate chemolithoautotrophs<sup>9,19–21</sup>. *A. aeolicus* (isolated by R.H. and K. O. Stetter) was cultured at 85 °C under an H<sub>2</sub>/CO<sub>2</sub>/O<sub>2</sub> (79.5:19.5:1.0) atmosphere in a medium containing only inorganic components. *A. aeolicus* does not grow on a number of organic substrates, including sugars, amino acids, yeast extract or meat extract. Unlike its close relative *A. pyrophilus*, *A. aeolicus* has not been shown to grow anaerobically with nitrate as an electron acceptor in the laboratory.

From study of the physiology of the organism, several predictions can be made. As an autotroph, *A. aeolicus* must have genes encoding proteins for one or more modes of carbon fixation and a complete set of biosynthetic genes. As autotrophy is a feature that is distributed throughout the Archaea and Bacteria, most of the associated genes are expected to be of ancient origin and clearly related to those characterized elsewhere. The obligate autotrophy suggests a biosynthetic rather than a degradative character. Oxygen respiration

implies the presence of corresponding utilization and tolerance genes. The early divergence of the *Aquificaceae* inferred from ribosomal RNA sequences leads to several questions. Are the machineries for oxygen usage and tolerance homologous to those found in mitochondria and well studied organisms such as *Escherichia coli*, or were they invented separately? If there was far less oxygen when the lineage originated, is there evidence for use of alternative oxidants?

## Genome

General features of the *A. aeolicus* genome are listed in Box 1. We classified 1,512 open-reading frames (ORFs) into one of three categories, namely, identified (Table 1), hypothetical, or unknown. Identified ORFs were further classified into one of 57 cellular role categories adapted from Riley<sup>22</sup> (Table 1). The relatively high G + C content of the two 16S-23S-5S rRNA operons (65%) is characteristic of thermophilic bacterial rRNAs<sup>23</sup>. The genome is densely packed: most genes are apparently expressed in polycistronic operons and many convergently transcribed genes overlap slightly. Nonetheless, many genes that are functionally grouped within operons in other organisms, such as the tryptophan or histidine biosynthesis pathways, are found dispersed throughout the *A. aeolicus* genome or appear in novel operons. Even when they encode subunits of the same enzyme, the genes are often separated on the chromosome (for example, *gltB* and *gltD*, the genes encoding the large and small subunits of glutamate synthase). Operon organization of genes for the biosynthesis of amino acids is found in both Archaea and Bacteria but it is not universal in either group. *A. aeolicus* is extreme in that no two amino acid biosynthetic genes are found in the same operon. In contrast, genes required for electron transport, hydrogenase subunits, transport systems, ribosomal subunits, and flagella are often in functionally related operons in *A. aeolicus* (Fig. 1). No introns or inteins (protein splicing elements) were detected in the genome.

A single extrachromosomal element (ECE) was identified during sequencing. Sequence redundancy for the total project was calculated to be 4.83. The ECE, however, is significantly over-represented

† Present addresses: Codex Bioinformatics Services, PO Box 90273, San Diego, California 92169, USA (G.D.); Department of Bioinformatics, SmithKline Beecham Pharmaceuticals, Collegeville, Philadelphia 19426, USA (P.V.W.)

relative to the chromosome; when calculated independently for the final assemblies, redundancies are 4.73 and 8.76 for the chromosome and for the ECE, respectively. The ECE therefore appears to be present at roughly twice the copy number of the chromosome. Although no ORFs on the ECE can be assigned a function with confidence, except for a transposase, two of the predicted proteins show similarity to hypothetical proteins in the *Methanococcus jannaschii* genome<sup>1</sup>. One ORF on the ECE is also present in two identical copies on the *A. aeolicus* chromosome, providing evidence of genetic exchange between the chromosome and the ECE.

### Reductive tricarboxylic acid cycle

As an autotroph, *A. aeolicus* obtains all necessary carbon by fixing CO<sub>2</sub> from the environment. An assay for activity of the reductive tricarboxylic acid (TCA) cycle in *A. pyrophilus* cell extracts showed *in vitro* activities for each proposed reaction<sup>24</sup>. The reductive (reverse) TCA cycle fixes two molecules of CO<sub>2</sub> to form acetyl-coenzyme A (acetyl-CoA) and other biosynthetic intermediates<sup>25</sup>. The *A. aeolicus* genome contains genes encoding malate dehydrogenase, fumarate hydratase, fumarate reductase, succinate-CoA ligase, ferredoxin oxidoreductase, isocitrate dehydrogenase, aconitase and citrate synthase, which together could constitute the TCA pathway. There is no biochemical evidence for alternative carbon-fixation pathways in *A. pyrophilus*<sup>24,25</sup> nor is there sequence evidence for such pathways in *A. aeolicus*.

The TCA cycle is vital as it provides the substrates of many biosynthetic pathways. (It is beyond the scope of this report to detail these biosynthetic pathways, but they seem to be typically bacterial, and candidate genes for all or most of the enzymes have been identified in *A. aeolicus*.) The central role of the TCA cycle is emphasized by duplication of many of its constituent genes in *A. aeolicus*. Two genes encode proteins that are similar to malate dehydrogenase (in addition to a lactate dehydrogenase). The fumarate hydratase is split into amino- and carboxy-terminal subunits, as is the case in *M. jannaschii*<sup>1</sup>. Unlinked genes encoding two iron-sulphur proteins of fumarate reductase (alternatively succinate dehydrogenase) accompany a single flavoprotein subunit. Two sets of genes resembling succinate-CoA ligase (both the  $\alpha$ - and  $\beta$ -subunits) are present. *A. aeolicus* has two putative operons encoding four-subunit ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ) 2-acid ferredoxin oxidoreductases; members of this family catalyze reversible carboxylation/decarboxylation of pyruvate, 2-isoketovaleate, or 2-oxoglutarate with varying specificity<sup>26</sup>. These duplicated genes may encode paralogous proteins with unique substrate specificity, as opposed to redundant functions. For example, a paralogue of succinate-CoA ligase may activate citrate with coenzyme A to form citryl-CoA, which citrate synthase can cleave to produce oxaloacetate and acetyl-CoA.

### Gluconeogenesis through the Embden–Meyerhof–Parnas pathway

Growing autotrophically, *A. aeolicus* must synthesize pentose and hexose monosaccharides from products of the reductive TCA cycle. Pyruvate produced by pyruvate ferredoxin oxidoreductase or by pyruvate carboxylase (oxaloacetate decarboxylase)<sup>24</sup> may enter the Embden–Meyerhof–Parnas pathway of glycolysis and gluconeogenesis. Genes encoding fructose-1,6-bisphosphatase, an essential gluconeogenic enzyme in *E. coli*, have not been identified in the genomes of the autotrophs *A. aeolicus* or *M. jannaschii*<sup>1</sup>, suggesting that an unidentified pathway may exist. The *A. aeolicus* genome also encodes enzymes of the pentose-phosphate pathway and enzymes for glycogen synthesis and catabolism. We found neither (phospho) gluconate dehydrase nor 2-keto-3-deoxy-(6-phospho)gluconate aldolase of the Entner–Doudoroff pathway.

### Respiration

*Aquifex* species are able to grow by using oxygen concentrations as low as 7.5 p.p.m. (R.H. and K. O. Stetter, unpublished observations).

The enzymes for oxygen respiration are similar to those of other bacteria: ubiquinol cytochrome *c* oxidoreductase (*bc*<sub>1</sub> complex), cytochrome *c* (three different genes) and cytochrome *c* oxidase (with two different subunit I genes and two different subunit II genes). The alternative system, with cytochrome *bd* ubiquinol oxidase, is also present. Clearly, the *Aquifex* lineage did not independently invent oxygen respiration. This leaves at least three possibilities: consistent with the ability of *Aquifex* to use very low levels of oxygen, the oxygen-respiration system was highly developed when oxygen had only a small fraction of its present concentration before the advent of oxygenic photosynthesis; contrary to what is implied by the 16S phylogeny, the lineage including *Aquifex* originated after the rise in atmospheric oxygen; or oxygen respiration developed once, and was then laterally transferred among bacterial lineages and acquired by *Aquifex*.

Many other oxidoreductases are present in addition to those obviously involved in oxygen respiration. The physiological role of most of these oxidoreductases is unknown or ambiguous, but two deserve comment. There is a putative nitrate reductase in the genome, although *A. aeolicus* has not been observed to perform NO<sub>3</sub><sup>-</sup> respiration, unlike the closely related *A. pyrophilus*. The nitrate reductase gene is adjacent to a nitrate transporter, and may be involved in nitrogen assimilation rather than respiration. It is also possible that *A. aeolicus* has a latent ability to respire with nitrate but that the conditions required have not been found. Two gene sequences show strong similarities to Rieske proteins, even though the rest of the ubiquinol cytochrome *c* oxidoreductase subunits appear only once in the genome. One of these Rieske protein genes is adjacent to a sulphide dehydrogenase subunit, suggesting a role in sulphur respiration.

### Oxidative stress

*A. aeolicus* grows optimally under microaerophilic conditions and consequently possesses various protective enzymes to counter reactive oxygen species, particularly superoxide and peroxide. The genome contains three genes encoding superoxide dismutases, two of the copper/zinc family and one of the iron/manganese family. The latter has also been noted in *A. pyrophilus*<sup>27</sup>. One of the copper/zinc superoxide dismutase genes is located in a large gene cluster encoding formate dehydrogenase.

No catalase genes were identified. There are several genes in the genome that might encode proteins that catalyze the detoxification of H<sub>2</sub>O<sub>2</sub>, including cytochrome *c* peroxidase, thiol peroxidase, and two alkyl hydroperoxide reductase genes. All of these enzymes require an exogenous reductant and therefore do not evolve O<sub>2</sub>. However, treatment of *A. pyrophilus*<sup>9</sup> or *A. aeolicus* biomass with H<sub>2</sub>O<sub>2</sub> results in the rapid evolution of gas bubbles. This catalase activity may result from a novel enzyme that cannot yet be identified by sequence similarity.

### Motility

Like *A. pyrophilus*<sup>9</sup>, *A. aeolicus* is motile and possesses monopolar polytrichous flagella. More than 25 genes encoding proteins involved in flagellar structure and biosynthesis have been identified in *A. aeolicus* (Box 1). However, no homologues of the bacterial chemotaxis system were identified. In enteric bacteria, membrane-bound receptors bind chemoattractants and repellents and mod-

**Figure 1** Linear map of the *A. aeolicus* circular chromosome. Genes are shown as arrows which denote the direction of transcription and are coloured to denote functional categorization according to the key below the figure. The sequences of the two rRNA gene clusters are identical. Here, the first base of the coding sequence of *fusA* was arbitrarily assigned as base number 1 as no origin of replication has been identified. ORF numbers are discontinuous because some ORFs representing 100 amino acids or more are not predicted to be coding and are not shown.

ulate the activity of the histidine kinase CheA<sup>28</sup>. Phosphoryl groups from CheA are transferred to CheY, which then binds to the flagellar switch, altering the direction of flagellar rotation. Homologous chemotaxis systems are present in the archaea *Halobacterium salinarum*<sup>29</sup> and *Pyrococcus* sp. OT3 (H. Sizuya, personal communication), although the bacterial and archaeal flagellar apparatuses are not homologous<sup>30</sup>. The *M. jannaschii* genome also lacks homologues of known genes required for chemotaxis. Thus, either motility in *A. aeolicus* and *M. jannaschii* is undirected or input for controlling taxis is mediated through another, unidentified system. The most studied chemotaxis systems respond to sugars and amino acids, although responses to other inputs (for example, metals, redox potential, and light) may also occur. In contrast to all the organisms known to possess the classical chemotactic signal-transduction pathways, both *A. aeolicus* and *M. jannaschii* are obligate chemoautotrophs. Chemoautotrophs may respond to a different set of factors, such as concentrations of dissolved gas (CO<sub>2</sub>, H<sub>2</sub> or O<sub>2</sub>) or another critical parameter such as temperature.

In *E. coli*, the flagellar switch is essential for flagellar structure and function and coupling of chemotaxis signals. But the *A. aeolicus* genome encodes homologues of only two of the three *E. coli* proteins that make up the switch, FliG and FliN. Biochemical<sup>31</sup> and genetic<sup>32</sup> studies implicate the missing FliM protein as the receptor for phosphorylated CheY, the switch signal. The absence of both FliM and CheY in *A. aeolicus* supports the identification of FliM as the receptor for phosphorylated CheY in *E. coli*. This result also argues against a direct role for FliM in torque generation.

### DNA replication and repair

The *A. aeolicus* primary replicative DNA polymerase, corresponding to the DNA polymerase III holoenzyme in *E. coli*, probably consists

**Figure 2** Histogram representation of the similarity of selected classes of predicted proteins to predicted proteins from the *E. coli* (EC) and *M. jannaschii* (MJ) genomes. Predicted *A. aeolicus* proteins representing each category were independently compared to sets of all potential polypeptides (≥100 amino acids) from the two genomes using FASTA<sup>44</sup>. If the top scoring alignment covered ≥80% of the length of the *A. aeolicus* protein, the score was plotted. There were more positives found in the *E. coli* genome in nearly every category. Hypothetical proteins (those identified by database match but of unknown function) are very similarly represented by *M. jannaschii* and *E. coli*. There are a small number of very highly conserved hypotheticals that are shared between *A. aeolicus* and *M. jannaschii*. Generally, biosynthetic categories show less discrimination than information-processing categories, which are clearly more *E. coli*-like. The variation in the apparent rates of evolution in different categories suggests that different phylogenies may be inferred depending on the sequence analysed. Within each graph, correspondence to *E. coli* is shown in white and *M. jannaschii* is shown in black. Avg id, average identity; count, number of proteins analysed.

### Box 1 *Aquifex aeolicus* genome features

#### General

Length 1,551,335 bp  
 G + C content 43.4%  
 Protein-coding regions 93%  
 Stable RNA 0.8%  
 Non-coding repeats (none significant)  
 Intergenic sequences 6.2%

#### RNA

Ribosomal RNA Chromosome coordinates  
 16S-23S-5S 572785-567770  
 16S-23S-5S 1192069-1197084  
 Transfer RNA  
 44 species (7 clusters, 28 single genes)  
 Other RNAs Chromosome coordinates  
 tmRNA 1153844-1153498

#### Chromosomal coding sequences

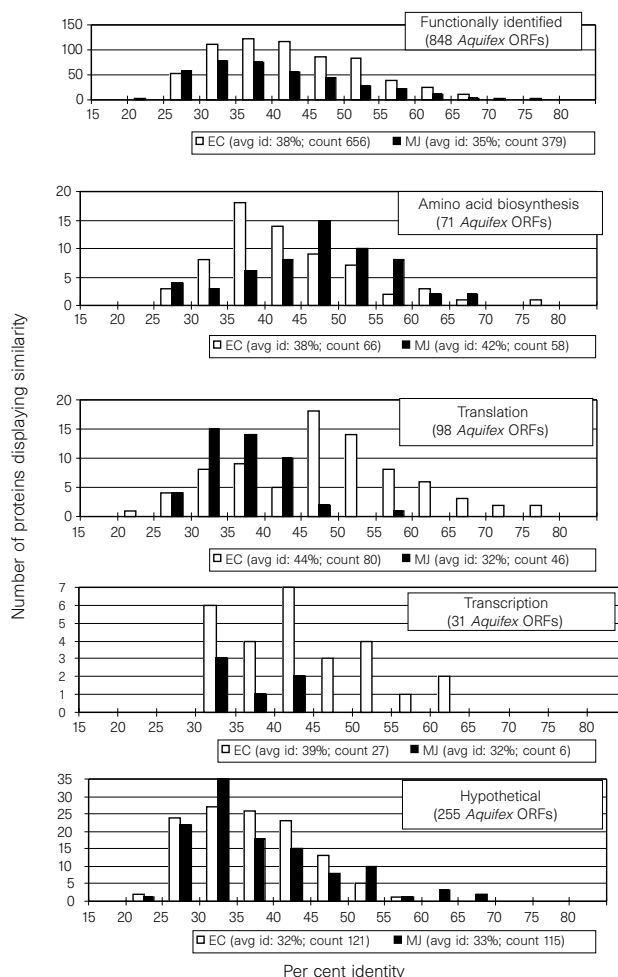
849 similar to protein of known function (average length 1,066 bp)  
 256 similar to protein of unknown function (average length 898 bp)  
 407 unknown coding regions (average length 762 bp)  
 1,512 total (average length 956)

#### Extrachromosomal element (ECE)

Length 39,456 bp  
 G + C content 36.4%  
 Protein-coding regions 53.5%

#### ECE-coding sequences

1 similar to proteins of known function (length 948 bp)  
 4 similar to proteins of unknown function (average length 667 bp)  
 27 unknown coding regions (average length 648 bp)



of a core structure containing  $\alpha$ - and  $\epsilon$ -subunits, a  $\gamma$ - $\tau$ -subunit and an additional member of the  $\gamma$ - $\tau$ / $\delta'$ -family. A gene encoding a protein homologous to the  $\beta$ -sliding clamp was also found. This minimalistic complex lacks homologous  $\theta$ -,  $\delta$ -,  $\chi$ - and  $\psi$ -subunits, as does the *Mycoplasma genitalium* holoenzyme<sup>3</sup>. Translation of the 54K (relative molecular mass)  $\gamma$ - $\tau$ -ATPase subunit may proceed without a programmed frameshift to produce a protein similar to the N-terminal region of the *E. coli*  $\gamma$ -subunit. DNA polymerase I is present as separate Klenow fragment and 5'  $\rightarrow$  3' exonuclease subunits, encoded by two non-adjacent ORFs. Although the repair polymerase, DNA polymerase II, has not been found in *A. aeolicus*, one ORF (Aq1422) encodes a protein similar to the eukaryotic DNA repair polymerase- $\beta$ . A member of the same family has been identified in *Thermus aquaticus*<sup>33</sup> and *Bacillus subtilis*.

**Transcriptional and translational apparatuses**

The transcriptional apparatus of *A. aeolicus* is similar to that of *E. coli* and lacks any components specific to the Eukarya or Archaea (Fig. 2). In addition to the core RNA polymerase  $\alpha$ -,  $\beta$ -, and  $\beta'$ -subunits, four  $\sigma$ -factors which determine promoter specificity are present (Table 1). Several different families of bacterial transcriptional regulators were also identified, including two-component systems. All of the ribosomal proteins and elongation factors common to other bacteria are present, indicating that all bacteria-specific ribosomal proteins were present in the common ancestor of *Aquifex* and other bacteria. Also present are the four *sel* genes required for the cotranslational incorporation of selenocysteine. These latter genes are clustered in a 15-kilobase-pair segment that also encodes the biosynthetic and structural proteins for formate dehydrogenase, the only selenocysteine-containing protein identified. The gene that encodes selenocysteine transfer RNA, *selC*, is apparently cotranscribed with the genes encoding the formate dehydrogenase structural proteins.

*A. aeolicus* lacks glutamyl-tRNA and asparaginyl-tRNA synthetases. The genes required for transamidation of glutamyl-tRNA<sup>Gln</sup> are present<sup>34</sup>. Charging of asparaginyl-tRNA is likely to proceed through the analogous reaction, as shown in halobacteria<sup>35</sup>, although the genes(s) for that transamidase are unknown. The canonical methionyl- and leucyl-tRNA synthetases have only been seen previously as single polypeptide enzymes; however, in *A.*

*aeolicus* the homologues appear fragmented into two subunits. In both cases, the genes that encode the N- and C-terminal portions are widely separated on the chromosome. No complete three-dimensional structural data are available for either methionyl- or leucyl-aminoacyl tRNA synthetases, but the subunit organization in the *A. aeolicus* aminoacyl-tRNA synthetases may reflect domain organization in the homologous proteins.

**Thermophily**

The *A. aeolicus* genome is the second completely sequenced genome of a hyperthermophile. By comparing the *A. aeolicus* and *M. jannaschii* genomes and contrasting them with the complete genomes of mesophiles, we can discover whether there are aspects of the genome or the encoded information that are diagnostic of hyperthermophiles. The G + C content of the stable RNAs is clearly indicative of the high growth temperature of the organism. This property can be used to identify stable RNAs against the relatively low G + C background of the *A. aeolicus* genome. The gene encoding tmRNA (or 10Sa RNA)<sup>36</sup>, an RNA involved in tagging polypeptides translated from incomplete messenger RNAs for degradation, was located in this way.

Two genes for reverse gyrase are present in the genome. This is the only protein known to be present only in thermophiles. Other proteins, currently described as hypotheticals, may be diagnostic of hyperthermophiles but the data sets are not yet large enough to decide this with confidence.

Although features of stabilization may not be apparent in any given protein<sup>37</sup>, a large enough data set may reveal general trends in amino-acid usage that are informative. Particularly important in this regard is inclusion of multiple genomes of hyperthermophiles so as not to allow the idiosyncracies of a single organism to bias the conclusions. As shown in Table 2, comparison of the amino-acid composition encoded by six genomes shows that use of individual amino acids can vary significantly from genome to genome. The data suggest trends that may be correlated with the thermostability of the encoded proteins. One apparent trend is that the hyperthermophile genomes encode higher levels of charged amino acids on average than mesophile genomes<sup>38</sup>, primarily at the expense of uncharged polar residues. Glutamine in particular seems to be significantly discriminated against in the hyperthermophiles. Although this observation might be rationalized on the basis of

**Table 2 Comparison of relative amino acid compositions (in percentages) of mesophiles and thermophiles**

Amino acid	Mesophiles				Thermophiles	
	<i>H. influenzae</i>	<i>H. pylori</i>	<i>E. coli</i>	<i>Synechosystis</i>	<i>A. aeolicus</i>	<i>M. jannaschii</i>
A	8.21	6.83	9.55	9.07	5.90	5.54
C	1.03	1.09	1.11	1.01	0.79	1.27
D	4.98	4.77	5.20	5.07	4.32	5.52
E	6.48	6.88	5.91	6.20	9.63	8.67
F	4.46	5.41	3.87	3.75	5.13	4.20
G	6.65	5.76	7.42	7.77	6.75	6.41
H	2.05	2.12	2.26	1.93	1.54	1.43
I	7.10	7.20	5.95	6.31	7.32	10.45
K	6.32	8.94	4.48	4.26	9.40	10.36
L	10.50	11.18	10.56	10.93	10.57	9.38
M	2.44	2.28	2.86	2.12	1.92	2.33
N	4.89	5.83	3.88	3.76	3.60	5.24
P	3.72	3.28	4.41	5.09	4.07	3.38
Q	4.64	3.70	4.42	5.26	2.04	1.44
R	4.47	3.46	5.58	5.18	4.91	3.85
S	5.84	6.81	5.67	5.46	4.79	4.46
T	5.20	4.37	5.35	5.53	4.21	4.06
V	6.68	5.59	7.11	7.10	7.93	6.85
W	1.12	0.70	1.48	1.30	0.93	0.71
Y	3.12	3.68	2.83	2.78	4.13	4.33
	Mesophiles				Thermophiles	
Charged residues (DEKRH)	24.11				29.84	
Polar/uncharged residues (GSTNQYC)	31.15				26.79	
Hydrophobic residues (LMIVWPAF)	44.74				43.36	

an increased rate of deamidation of this residue at higher temperatures, asparagine does not appear subject to similar discrimination.

### Phylogeny

The placement of the *Aquifex* lineage as one of the earliest divergences in the eubacterial tree<sup>13,14</sup> is interesting because of the insights it could provide into the ancestral eubacterial phenotype, including the hypothesized thermophilic nature of the first bacteria. Protein-based phylogenies often do not support the original rRNA-based placement<sup>15,16,18</sup>. Thus, the availability of some 1,500 genes from an *Aquifex* species would seem to offer a definitive resolution of the phylogeny. However, our analyses of ribosomal proteins, aminoacyl-tRNA synthetases, and other proteins do not do so, showing no consistent picture of the organism's phylogeny. We cannot make a more complete analysis and discussion here, but some observations can be made. These proteins do not yield a statistically significant placement of the *Aquifex* lineage or of other major eubacterial lineages. This situation partially reflects the inadequacy of some protein sequences as indicators of distant molecular genealogy because of their particular evolutionary dynamic, including the patterns and rates of amino-acid replacements. In some cases (such as the aminoacyl-tRNA synthetases for arginine, cysteine, histidine, proline and tyrosine), the analyses are further complicated by the presence of paralogous genes and/or apparent lateral gene transfers. It seems that a more extensive survey of genes and a better sampling of major eubacterial taxa will be required to confidently confirm or refute an early divergence of the *Aquifex* lineage.

### Conclusions

Advances in sequencing techniques have allowed us to move beyond studies of single genes to studies of complete genomes only recently<sup>2</sup>. This rapid advance has created the opportunity to begin to characterize an organism with the full knowledge of the genome in hand. The complete genome summarized in this report represents our first view of *A. aeolicus*. The challenge now is to ask specific questions in ways which take advantage of the whole-genome data.

Beyond studies of any single organism in isolation, complete genomes allow comprehensive comparisons between organisms. For instance, comparisons of the similarity of genes can be made that reveal that genes in different categories vary in their relative conservation (Fig. 2). In addition, genome-wide trends are apparent. For example, why is there not more of a tendency to group functionally related genes (for example, biosynthetic pathways) into operons in *A. aeolicus*? This was also seen in the genome sequence of the autotroph *M. jannaschii*<sup>1</sup>. Is this because the autotrophic lifestyle decreases the need for selective regulation? There also seem to be a few multifunctional, fused proteins in *A. aeolicus* and *M. jannaschii*. Although this seems unlikely to be related to autotrophy, it might be associated with extreme thermophily. The large number of diverse genome sequences that will become available in the coming years will allow more detailed correlation of global genomic properties with particular physiologies. □

### Methods

**Sequencing strategy.** The sequencing strategy used to assemble the complete genome was based on the whole genome random (or 'shotgun') approach, which has been successfully used for other genomes of similar size<sup>1-4</sup>. Shotgun sequencing projects are characterized by two phases: an initial completely random phase in which the bulk of the data is collected, followed by a closure phase where directed techniques are used to close gaps and complete the assembly. By pursuing a strategy where only 97% coverage was initially achieved, we were able to limit the number of sequences needed for the random phase to only 10,500 (ref. 39).

Sequences were generated from a small insert library constructed in  $\lambda$  ZAP II vectors<sup>40,41</sup> (average insert length 2.9 kilobase pairs). Two different methods were used for sequencing: first, dye-primer M13-21 and M13 reverse primer ABI Prism CS<sup>+</sup> ready reaction kits, analysed on 48-cm 4% polyacrylamide

gels; and second, dye-terminator (ABI Prism FS<sup>+</sup>) reactions using two pBluescript-specific primers. These reactions were analysed on 36-cm 5% Long-Ranger gels.

The sequence fragments were assembled on an Apple Power Macintosh computer using Sequencher (Gene Codes, Ann Arbor, MI), an assembly and editing program. Assembly was typically performed in batches of roughly 200–400 sequences, and was followed by inspection and editing of the assemblies. All sequences in the set were compared with all others through this process. After assembly, the sequences comprised ~750 contigs at the end of the random phase. Sequences were obtained from both ends of ~200 randomly chosen clones from a fosmid library<sup>42,43</sup>. These sequences were then assembled with consensus sequences derived from the contigs of random-phase sequences using Sequencher. Gaps between contigs were closed by direct sequencing on fosmids not wholly contained within a contig. The fosmid library thus served a purpose analogous to that of the  $\lambda$ -scaffold in other projects<sup>1-4</sup>. The final eight gaps were closed by direct sequencing of polymerase chain reaction (PCR) products generated with the TaqPlus Long PCR System (Stratagene Cloning Systems, La Jolla, CA).

Consequences of reducing the number of sequences in the random phase are the large number of gaps that remain to be closed in the directed phase, and the reduction in overall coverage. To ensure that reduced coverage did not compromise accuracy, ~200 oligonucleotide primers were synthesized to resequence regions of ambiguity identified by visual inspection of the entire assembly. 13,785 sequences, with an average edited read length of 557 base pairs, constitute the final assembly. On the basis of a relatively small number of errors identified during the annotation process, we estimate the error frequency to be <0.01%, comparable to other published genomic sequence estimates.

### Gene (ORF + RNA) identification and functional assignment approaches.

Coding regions of the *A. aeolicus* genome were analysed and assigned using primarily the programs BLASTP<sup>44</sup> and FASTA<sup>45</sup> to search against a non-redundant protein database. Many analyses were carried out within the context of MAGPIE<sup>46,47</sup>, an integrated computing environment for genome analysis. The results of these analyses are available for user interpretation, validation, and categorization. Additional ORFs were identified and start sites refined using the program CRITICA (J. H. Badger and G.J.O., unpublished program). Finally, all presumed 'intergenic regions' were examined with BLASTX for similarities to known protein sequences<sup>48</sup>. Transfer RNA genes were identified with the program tRNAscan-SE<sup>49</sup>.

Received 26 August 1997; accepted 3 February 1998.

- Bult, C. et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073 (1996).
- Fleischmann, R. D. et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–511 (1995).
- Fraser, C. M. et al. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
- Tomb, J.-F. et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
- Himmelreich, R. et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420–4449 (1996).
- Kaneko, T. et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC7803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136 (1996).
- Blattner, F. R. et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
- Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546 (1996).
- Huber, R. et al. *Aquifex pyrophilus* gen. nov. sp. nov. represents a novel group of marine hyperthermophilic hydrogen oxidizing bacteria. *Arch. Microbiol.* **15**, 340–351 (1992).
- Reysenbach, L., Wickham, G. S. & Pace, N. R. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* **60**, 2113–2119 (1994).
- Setchell, W. A. The upper temperature limits of life. *Science* **17**, 934–937 (1903).
- Brock, T. D. The road to Yellowstone—and beyond. *Annu. Rev. Microbiol.* **49**, 1–28 (1995).
- Burggraf, S., Olsen, G. J., Stetter, K. O. & Woese, C. R. A phylogenetic analysis of *Aquifex pyrophilus*. *Syst. Appl. Microbiol.* **15**, 353–356 (1992).
- Pitulle, C. et al. Phylogenetic position of the genus *Hydrogenobacter*. *Int. J. Syst. Bacteriol.* **44**, 620–626 (1994).
- Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**, 7749–7754 (1996).
- Klenk, H.-P., Palm, P. & Zillig, W. in *Molecular Biology of the Archaea* (eds Pfeifer, E., Palm, P. & Scleifer, K. H.) 139–147 (Vch Pub, 1994).
- Bocchetta, M. et al. Arrangement and nucleotide sequence of the gene (*fus*) encoding elongation factor G (EF-G) from the hyperthermophilic bacterium *Aquifex pyrophilus*: phylogenetic depth of hyperthermophilic bacteria inferred from analysis of the EF-G/*fus* sequences. *J. Mol. Evol.* **41**, 803–812 (1995).
- Wetmur, J. G. et al. Cloning, sequencing, and expression of RecA proteins from three distantly related thermophilic eubacteria. *J. Biol. Chem.* **269**, 25928–25935 (1994).
- Kawasumi, T., Igarashi, Y., Kodama, T. & Minoda, Y. *Hydrogenobacter thermophilus* gen. nov., sp. nov.

- an extremely thermophilic, aerobic, hydrogen-oxidizing bacterium. *Int. J. Syst. Bacteriol.* **34**, 5–10 (1984).
20. Kristjansson, J., Ingason, A. & Alfredsson, G. A. Isolation of thermophilic obligately autotrophic hydrogen-oxidizing bacteria, similar to *Hydrogenobacter thermophilus*, from Icelandic hot springs. *Arch. Microbiol.* **140**, 321–325 (1985).
  21. Kryukov, V. R., Savel'eva, N. D. & Pusheva, M. A. *Calderobacterium hydrogenophilum* gen. nov., sp. nov. an extreme thermophilic bacterium and its hydrogenase activity. *Microbiology (Engl. Trans. Mikrobiologiya)* **52**, 611–618 (1983).
  22. Riley, M. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952 (1993).
  23. Weisburg, W. G., Giovannoni, S. J. & Woese, C. R. The *Deinococcus-Thermus* phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* **11**, 128–134 (1989).
  24. Beh, M., Strauss, G., Huber, R., Stetter, K. O. & Fuchs, G. Enzymes of the reductive citric acid cycle in the autotrophic eubacterium *Aquifex pyrophilus* and in the archaeobacterium *Thermoproteus neutrophilus*. *Arch. Microbiol.* **160**, 306–311 (1993).
  25. Fuchs, G. in *Autotrophic Bacteria* (eds Schegel, H. G. & Bowein, B.) 365–382 (Springer, New York, 1987).
  26. Mai, X. & Adams, M. W. Characterization of a fourth type of 2-keto acid-oxidizing enzyme from a hyperthermophilic archaeon: 2-ketoglutarate ferredoxin oxidoreductase from *Thermococcus litoralis*. *J. Bacteriol.* **178**, 5890–5896 (1996).
  27. Lim, J. H. *et al.* Cloning and expression of superoxide dismutase from *Aquifex pyrophilus*, a hyperthermophilic bacterium. *FEBS Lett.* **406**, 142–146 (1997).
  28. Bourret, R. B., Borkovich, K. A. & Simon, M. I. Signal transduction pathways involving protein phosphorylation in prokaryotes. *Annu. Rev. Biochem.* **60**, 401–441 (1991).
  29. Rudolph, J., Tolliday, N., Schmitt, C., Schuster, S. C. & Oesterhelt, D. Phosphorylation in halobacterial signal transduction. *EMBO J.* **14**, 4249–4257 (1995).
  30. Jarrell, K. F., Bayley, D. P. & Kostyukova, A. S. The archaeal flagellum: a unique motility structure. *J. Bacteriol.* **178**, 5057–5064 (1996).
  31. Welch, M., Oosawa, K., Aizawa, S. I. & Eisenbach, M. Effects of phosphorylation, Mg<sup>2+</sup>, and conformation of the chemotaxis protein CheY on its binding to the flagellar switch protein FlIM. *Biochemistry* **33**, 10470–10467 (1994).
  32. Sockett, H., Yamaguchi, S., Kihara, M., Irikura, V. M. & Macnab, R. M. Molecular analysis of the flagellar switch protein FlIM of *Salmonella typhimurium*. *J. Bacteriol.* **174**, 793–806 (1992).
  33. Motoshima, H. *et al.* Molecular cloning and nucleotide sequence of the aminopeptidase T gene of *Thermus aquaticus* YT-1 and its high-level expression in *Escherichia coli*. *Agric. Biol. Chem.* **54**, 2385–2392 (1990).
  34. Curnow, A. W. *et al.* Glu-tRNA<sup>Gln</sup> amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl Acad. Sci. USA* **94**, 11819–11826 (1997).
  35. Curnow, A. W., Ibba, M. & Söll, D. tRNA-dependent asparagine formation. *Nature* **382**, 589–590 (1996).
  36. Tu, G. F., Reid, G. E., Zhang, J. G., Moritz, R. L. & Simpson, R. J. C-terminal extension of truncated proteins in *Escherichia coli* with a 10Sa decapeptide. *J. Biol. Chem.* **270**, 9322–9326 (1995).
  37. Böhm, G. & Jaenicke, R. Relevance of sequence statistics for the properties of extremophilic proteins. *Int. J. Pept. Protein Res.* **43**, 97–106 (1994).
  38. Choi, I.-G. *et al.* Random sequence analysis of genomic DNA of a hyperthermophile: *Aquifex pyrophilus*. *Extremophiles* **1**, 125–134 (1997).
  39. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
  40. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Res.* **16**, 7583–7600 (1988).
  41. Altling-Mees, M. A. & Short, J. M. pBluescript II: gene mapping vectors. *Nucleic Acids Res.* **17**, 9494 (1989).
  42. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89**, 8794–8797 (1992).
  43. Kim, U.-J., Shizuya, H., de Jong, P. J., Birren, B. & Simon, M. I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* **20**, 1083–1085 (1992).
  44. Altschul, S. F., Fish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
  45. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988).
  46. Gaasterland, T. & Sensen, C. W. MAGPIE: automated genome interpretation. *Trends Genet.* **12**, 76–78 (1996).
  47. Gaasterland, T. & Sensen, C. W. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* **78**, 302–310 (1996).
  48. Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nature Genet.* **3**, 266–272 (1993).
  49. Lowe, 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).

**Acknowledgements.** This work was supported in part by Department of Energy Microbial Genome Program grants (to R.V.S., C. R. Woese and G.J.O.). We thank C. Woese for his cooperation in the analysis of the genome and interest in the project; K. Stetter for continuing interest; G. Frey, J. Holaska, S. Peralta, D. Hafenbrandl, S. Delk, T. Robinson, and J. Arnett for technical assistance; and D. Robertson, J. Stein, I. Sanyal, T. Richardson, G. Hauska, and K. Williams for discussions.

Correspondence should be addressed to R.V.S. (e-mail: rswanson@diversa.com). Requests for *Aquifex aeolicus* should be addressed to R.H. (e-mail: Robert.huber@biologie.uni-regensburg.de). The sequences have been deposited with GenBank and assigned accession numbers AE000657 (chromosome) and AE000667 (extrachromosomal element).





**Table 1 *Aquifex aeolicus* Open Reading Frame Identifications. Gene numbers (Aq) correspond to those in Fig. 1. Percentages refer to the identity found in the best FASTA alignment. The percentage of the sequence covered by the alignment is displayed with bullets as follows 20–40% ••, 40–60% •••, 60–80% ••••, 80–100% •••••**

<b>Amino Acid Biosynthesis</b>		
<b>Aromatic amino acids</b>		
Aq1536	aroA	5-enolpyruvylshikimate-3-phosphate synthetase
Aq081	aroC	chorismate synthase
Aq021	aroD	3-dehydroquinate dehydratase
Aq901	aroE	shikimate 3-dehydrogenase
Aq2177	aroK	shikimate kinase
Aq951	pheA	chorismate mutase/prephenate dehydratase
Aq1548	trpA	tryptophan synthase alpha subunit
Aq706	trpB1	tryptophan synthase beta subunit
Aq1410	trpB2	tryptophan synthase beta subunit
Aq1787	trpC	indole-3-glycerol phosphate synthase
Aq196	trpD1	phosphoribosylanthranilate transferase
Aq209	trpD2	phosphoribosylanthranilate transferase
Aq582	trpE	anthranilate synthase component I
Aq2076	trpF	phosphoribosyl anthranilate isomerase
Aq549	trpG	anthranilate synthase component II
Aq1755	tyrA	prephenate dehydrogenase
<b>Aspartate family</b>		
Aq1866	asd	aspartate-semialdehyde dehydrogenase
Aq1969	aspC1	aspartate aminotransferase
Aq2094	aspC2	aminotransferase (AspC family)
Aq421	aspC3	aminotransferase (AspC family)
Aq473	aspC4	aminotransferase (AspC family)
Aq1143	dapA	dihydrodipicolinate synthase
Aq916	dapB	dihydrodipicolinate reductase
Aq547	dapF	succinyl-diaminopimelate desuccinylase
Aq1838	dapF	diaminopimelate epimerase
Aq1208	lysA	diaminopimelate decarboxylase
Aq1152	lysC	aspartokinase
Aq1710	metE	tetrahydropteroylriglutamate methyltransferase
Aq1812	thrA	homoserine dehydrogenase
Aq1309	thrB	homoserine kinase
Aq608	thrC1	threonine synthase
Aq425	thrC2	threonine synthase
<b>Branched-chain family</b>		
Aq451	ilvB	acetolactate synthase large subunit
Aq1245	ilvC	acetohydroxy acid isomeroeductase
Aq837	ilvD	dihydroxyacid dehydratase
Aq1893	ilvE	branched-chain amino acid aminotransferase
Aq1851	ilvH	acetolactate synthase
Aq356	leuA1	2-isopropylmalate synthase
Aq2090	leuA2	2-isopropylmalate synthase
Aq244	leuB	3-isopropylmalate dehydrogenase
Aq940	leuC	large subunit of isopropylmalate isomerase
Aq1398	leuD	3-isopropylmalate dehydratase
<b>Glutamate family</b>		
Aq2068	argB	acetylglutamate kinase
Aq1879	argC	N-Acetyl-gamma-glutamylphosphate reductase
Aq023	argD	N-acetylmethionine aminotransferase
Aq1711	argF	ornithine carbamoyltransferase
Aq1140	argG	argininosuccinate synthase
Aq1372	argH	argininosuccinate lyase
Aq970	argJ	glutamate N-acetyltransferase
Aq111	glnA	glutamine synthetase
Aq109	glnB	nitrogen regulatory PII protein
Aq1774	glnE	glutamate ammonia ligase adenyllyl-transferase
Aq1565	glnF	glutamate synthase large subunit
Aq2064	glnD	glutamate synthase small subunit
Aq1071	proA	gamma-glutamyl phosphate reductase
Aq1134	proB	glutamate 5-kinase
Aq166	proC	pyrroline carboxylate reductase
<b>Histidine</b>		
Aq1303	hisA	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase
Aq039	hisB	imidazoleglycerolphosphate dehydratase
Aq2084	hisC	histidinol-phosphate aminotransferase
Aq782	hisD	histidinol dehydrogenase
Aq181	hisF	HisF (cyclase)
Aq1613	hisG	ATP phosphoribosyltransferase
Aq732	hisH	amidotransferase HisH
Aq1968	hisE	phosphoribosyl-ATP pyrophosphohydrolase
<b>Selenocysteine</b>		
Aq1031	selA	L-seryl-rRNA (ser) selenium transferase
Aq1030	selD	selenophosphate synthase
<b>Serine family</b>		
Aq1556	cysM	cysteine synthase, O-acetylserine (thiol) lyase B
Aq479	glyA	serine hydroxymethyl transferase
Aq1905	serA	D-3-phosphoglycerate dehydrogenase
<b>Cell Envelope</b>		
<b>Pili and fimbriae</b>		
Aq1433	fimZ	minor pilin
Aq1432	ppdD1	pilin
Aq1434	ppdD2	pilin
Aq1435	ppdD3	pilin
<b>Lipoproteins and porins</b>		
Aq270	lgt	prolipoprotein diacylglycerol transferase
Aq819	lnt	apolipoprotein N-acyltransferase
Aq652	nlpD1	lipoprotein
Aq1753	nlpD2	lipoprotein NlpD fragment
Aq529	optC	outer membrane protein c
Aq2147	pal	peptidoglycan associated lipoprotein
Aq1370	rlpA1	rare lipoprotein A
Aq1174	rlpA2	rare lipoprotein A
Aq2166	schA	adhesion protein
Aq619	yfcA	adhesion B precursor
<b>Peptidoglycan</b>		
Aq1827	alr	alanine racemase
Aq1681	amiB	N-acetylmuramoyl-L-alanine amidase
Aq2195	bacA	undecaprenol kinase
Aq1798	cphA1	beta lactamase precursor
Aq974	cphA2	beta lactamase precursor
Aq521	ddlA	D-alanine:D-alanine ligase
Aq301	glmS	glucosamine-fructose-6-phosphate aminotransferase
Aq607	glmU	UDP-N-acetylglucosamine pyrophosphorylase
Aq053	mraY	phospho-N-acetylmuramoyl-pentapeptide-transferase
Aq624	mrcA	penicillin binding protein 1A
Aq1281	mura	UDP-N-acetylglucosamine
Aq520	murB1	1-carboxyvinyltransferase
Aq511	murB2	UDP-N-acetylenolpyruvylglucosamine reductase
Aq1360	murC	UDP-N-acetylmuramate-alanine ligase
Aq2075	murD	UDP-N-acetylmuramoylalanine-D-glutamate ligase
Aq1747	murE	UDP-MurNac-tripeptide synthetase
Aq821	murF	UDP-MURNAC-pentapeptide synthetase
Aq1177	murG	phospho-N-acetylmuramoyl-pentapeptide-transferase
Aq325	murI	glutamate racemase
Aq1189	pbpA1	penicillin binding protein 2
Aq556	pbpA2	penicillin binding protein 2
Aq185	tagD1	glycerol-3-phosphate cytidylyltransferase
Aq1368	tagD2	glycerol-3-phosphate cytidylyltransferase
<b>Surface polysaccharides and lipopolysaccharides</b>		
Aq1684	alg	alginate synthesis-related protein
Aq1641	cap	capsular polysaccharide biosynthesis protein
Aq1899	dmt	dolichol-phosphate mannosyltransferase
Aq1772	envA	UDP-3-O-acyl-N-acetylglucosamine deacetylase
Aq1757	exbB	biopolymer transport exbB
Aq1839	exbD	biopolymer transport ExbD
Aq1069	galE	UDP-glucose-4-epimerase
Aq1705	galF	UDP-glucose pyrophosphorylase
Aq908	gmlA	phosphohexose isomerase
Aq085	kdsA	3-deoxy-D-manno-octulosonate cytidylyltransferase
Aq326	kdtA	3-deoxy-D-manno-2-octulosonic acid transferase
Aq253	kdtB	lipopolysaccharide core biosynthesis protein
Aq1546	kpsF	polysialic acid capsule expression protein
Aq692	kpsU	3-deoxy-manno-octulosonate cytidylyltransferase
Aq1742	lgtI	beta 1,4 glucosyltransferase
Aq604	lpxA	acyl-acyl-carrier-protein-UDP-N-acetylglucosamine acyltransferase
Aq1427	lpxB	lipid A disaccharide synthetase
Aq538	lpxD	UDP-3-O-(3-hydroxymyristoyl) glucosamine N acyltransferase
Aq718	mpg	mannose-1-phosphate guanylyltransferase
Aq1096	mtfA	mannosyltransferase A
Aq515	mtfB	mannosyltransferase B
Aq516	mtfC	mannosyltransferase C
Aq1335	nse	nucleotide sugar epimerase
Aq505	otnA	polysaccharide biosynthesis protein
Aq504	otnA'	polysaccharide biosynthesis protein (fragment)
Aq1543	rfaC1	ADP-heptose:PS heptosyltransferase
Aq145	rfaC2	ADP-heptose:PS heptosyltransferase
Aq344	rfaD	ADP-1-glycero-D-manno-heptose-6-epimerase
Aq565	rfaE	ADP-heptose synthase
Aq2115	rfaG	glucosyl transferase I
Aq1082	rfaH	GDP-D-mannose dehydratase
Aq519	rfe	undecaprenyl-phosphate-alpha-N-acetylglucosaminyltransferase
Aq1367	spj	glucose-1-phosphate thymidyltransferase
Aq518	spk	spore coat polysaccharide biosynthesis protein
Aq589	xanB	Spk
<b>Cellular Processes</b>		
<b>Cell division</b>		
Aq698	acrE	acriflavin resistance protein AcrE
Aq1275	cafA	cytoplasmic axial filament protein
Aq523	ftsA	cell division protein FtsA
Aq936	ftsH	cell division protein FtsH
Aq1139	ftsW	cell division protein FtsW
Aq920	ftsY	cell division protein FtsY
Aq525	ftsZ	cell division protein FtsZ
Aq761	gidA1	glucose inhibited division protein A
Aq691	gidA2	glucose inhibited division protein A
Aq1582	gldB	glucose inhibited division protein B
Aq1718	maf	MAP protein
Aq1887	mesJ	cell cycle protein MesJ
Aq878	minC	septum site-determining protein MinC
Aq1217	minD1	septum site-determining protein MinD
Aq877	minD2	septum site-determining protein MinD
Aq845	mreB	rod shape determining protein MreB
Aq025	rodA	rod shape determining protein RodA
Aq1130	sufI	periplasmic cell division protein (SufI)
<b>Chaperones</b>		
Aq154	ctaB	cytochrome c oxidase assembly factor
Aq1735	dnaJ1	chaperone DnaJ
Aq703	dnaJ2	chaperone DnaJ
Aq996	dnaK	Hsp70 chaperone DnaK
Aq483	grpE	heat shock protein GrpE
Aq192	hslU	chaperone HslU
Aq1283	hspC	small heat shock protein (class I)
Aq1991	hspX	heat shock protein X
Aq2200	mopA	GroEL
Aq1999	mopB	GroES
<b>Detoxification</b>		
Aq486	ahpC1	alkyl hydroperoxide reductase
Aq858	ahpC2	alkyl hydroperoxide reductase
Aq685	arsC	arsenate reductase
Aq136	cpx	cytochrome c peroxidase
Aq1005	cutA	periplasmic divalent cation tolerance protein
Aq1499	cutB	superoxide dismutase (Fe/Mn family)
Aq1050	sodC1	superoxide dismutase (Cu/Zn)
Aq238	sodC2	superoxide dismutase (Cu/Zn)
Aq488	tpx	thiol peroxidase
<b>Motility</b>		
Aq833	flgA	flagellar protein FlgA
Aq1184	flgB	flagellar basal body rod protein FlgB
Aq1183	flgC	flagellar biosynthesis FlgC
Aq1859	flgE	flagellar hook protein FlgE
Aq2051	flgG1	flagellar hook basal-body protein FlgG
Aq834	flgG2	flagellar hook basal-body protein FlgG
Aq1714	flgH	flagellar L-ring protein FlgH
Aq1713	flgI	flagellar P-ring protein FlgI
Aq1662	flgK	flagellar hook associated protein FlgK
Aq1663	flgL	flagellar hook associated protein FlgL
Aq1212	flhA	flagellar export protein
Aq2014	flhB	flagellar biosynthetic protein FlhB
Aq1214	flhF	flagellar biosynthesis FlhF
Aq1998	flhC	flagellin





Aq1708	pkA	phosphofructokinase	49.4% ....	Aq046	pyrD	dihydroorotate dehydrogenase	50.5% ....
Aq750	pgi	glucose-6-phosphate isomerase	37.8% ....	Aq1305	pyrDB	dihydroorotate dehydrogenase electron transfer subunit	34.7% ....
Aq118	pgk	phosphoglycerate kinase	54.5% ....	Aq1580	pyrF	orotidine-5'-phosphate decarboxylase	37.2% ....
Aq1990	pgmA	phosphoglycerate mutase	33.2% ....	Aq1334	pyrG	CTP synthetase	57.5% ....
Aq501	pnu	phosphoglucomutase/phosphomannomutase	37.8% ....	Aq713	pyrH	UMP kinase	62.1% ....
Aq2142	ppsA	phosphoenolpyruvate synthase	56.3% ....	Aq640	thy	thymidylate synthase complementing protein	30.5% ....
Aq1520	pycA	pyruvate carboxylase c-terminal domain	46.6% ....	Aq969	tmk	thymidylate kinase	35.1% ....
Aq1517	pycB	pyruvate carboxylase n-terminal domain	57.1% ....	Aq1907	umpS	uridine 5-monophosphate synthase	42.1% ....
Aq360	timA	triose phosphate isomerase	52.2% ....	Aq2163	uraP	uracil phosphoribosyltransferase	42.0% ....
<b>Hydrogenase</b>							
Aq665	hozZ	Ni/Fe hydrogenase B-type cytochrome subunit	40.4% ....	<b>Regulation</b>			
Aq667	hupD	HupD hydrogenase related function	40.9% ....	Aq1058	acrR1	transcriptional regulator (TetR/AcrR family)	34.1% ..
Aq666	hupE	HupE hydrogenase related function	38.3% ....	Aq2179	acrR2	transcriptional regulator (TetR/AcrR family)	31.0% ....
Aq1021	hysA	hydrogenase accessory protein HysA	39.8% ....	Aq281	acrR3	transcriptional regulator (TetR/AcrR family)	29.7% ....
Aq671	hysB	hydrogenase expression/formation protein B	50.6% ....	Aq1387	arsR	transcriptional regulator (ArsR family)	35.3% ....
Aq1157	hupD	hydrogenase expression/formation protein HypD	56.1% ....	Aq1724	degT	transcriptional regulator (DegT/DnrT/EryC family)	34.1% ....
Aq662	mbhL1	hydrogenase large subunit	50.6% ....	Aq534	draG	ADP-ribosylglycohydrolase	32.1% ....
Aq960	mbhL2	hydrogenase large subunit	44.3% ....	Aq831	exsB	trans-regulatory protein ExsB	38.5% ....
Aq804	mbhL3	hydrogenase large subunit	27.9% ....	Aq490	fmr	transcriptional regulator (Crp/Fnr family)	29.5% ....
Aq660	mbhS1	hydrogenase small subunit	66.6% ....	Aq1207	furR1	transcriptional regulator (FurR family)	37.9% ....
Aq965	mbhS2	hydrogenase small subunit	51.3% ....	Aq1418	furR2	transcriptional regulator (FurR family)	34.6% ....
Aq802	mbhS3	hydrogenase small subunit	36.7% ....	Aq213	ghnB	PH-like protein GhnB	48.0% ..
Aq1591	shyS	soluble hydrogenase small subunit	41.6% ....	Aq1908	hIX	GTP-binding protein HIX	40.3% ....
<b>Sugar metabolism</b>							
Aq968	cbfE2	ribulose-5-phosphate 3-epimerase	47.2% ....	Aq1115	hksP1	histidine kinase sensor protein	27.7% ..
Aq1658	fucA1	fuculose-1-phosphate aldolase	31.8% ....	Aq316	hksP2	histidine kinase sensor protein	28.1% ....
Aq1979	fucA2	fuculose-1-phosphate aldolase	29.7% ....	Aq905	hksP3	histidine kinase sensor protein	23.6% ....
Aq498	gnd	6-phosphogluconate dehydrogenase	45.2% ....	Aq231	hksP4	histidine kinase sensor protein	28.2% ....
Aq497	gsdA	glucose-6-phosphate 1-dehydrogenase	32.3% ....	Aq156	hozX	hydrogenase regulator HozX	46.7% ....
Aq1138	rpiB	ribose 5-phosphate isomerase B	54.5% ....	Aq093	hth	transcriptional regulator (H-T-H)	50.2% ....
Aq1119	talC	transaldolase	71.1% ....	Aq1019	hypE	hydrogenase expression/formation protein	44.3% ....
Aq1765	tkkA	transketolase	52.4% ....	Aq672	hypF	transcriptional regulatory protein HypF	44.8% ....
<b>NADH dehydrogenase</b>							
Aq1385	nuoA1	NADH dehydrogenase I chain A	42.0% ....	Aq764	icrR	transcriptional regulator (IcR family)	30.4% ....
Aq1310	nuoA2	NADH dehydrogenase I chain A	44.9% ....	Aq638	lysR1	transcriptional regulator (LysR family)	32.8% ....
Aq1312	nuoB	NADH dehydrogenase I chain B	60.1% ....	Aq1038	lysR2	transcriptional regulator (LysR family)	28.9% ....
Aq551	nuoD1	NADH dehydrogenase I chain D	37.7% ....	Aq702	merR	transcriptional regulator (MerR family)	32.8% ....
Aq1314	nuoD2	NADH dehydrogenase I chain D	42.2% ....	Aq218	nifA	transcriptional regulator (NifA family)	42.8% ....
Aq574	nuoE	NADH dehydrogenase I chain E	38.9% ....	Aq1117	nitC1	transcriptional regulator (NtrC family)	41.0% ....
Aq573	nuoF	NADH dehydrogenase I chain F	20.5% ....	Aq1792	nitC2	transcriptional regulator (NtrC family)	40.2% ....
Aq437	nuoG	NADH dehydrogenase I chain G	35.4% ..	Aq1493	nitC3	transcriptional regulator (NtrC family)	38.3% ....
Aq1315	nuoH1	NADH dehydrogenase I chain H	41.0% ....	Aq164	nitC4	transcriptional regulator (NtrC family)	35.3% ....
Aq1373	nuoH2	NADH dehydrogenase I chain H	42.1% ....	Aq2069	obg	GTP-binding protein	54.9% ..
Aq1374	nuoH3	NADH dehydrogenase I chain H	38.9% ....	Aq319	phoB	transcriptional regulator (PhoB-like)	41.6% ....
Aq1317	nuoI1	NADH dehydrogenase I chain I	30.5% ....	Aq906	phoU	transcriptional regulator (PhoU-like)	41.9% ....
Aq1375	nuoI2	NADH dehydrogenase I chain I	29.2% ..	Aq844	pplE	(p)ppGpp 3-phosphorylhydrolase	57.2% ....
Aq1318	nuoI1	NADH dehydrogenase I chain I	35.4% ....	Aq1496	xyrR	transcriptional regulator (NagC/XyIR family)	29.3% ....
Aq1377	nuoI2	NADH dehydrogenase I chain I	30.6% ....	<b>DNA Replication and Repair</b>			
Aq1319	nuoK1	NADH dehydrogenase I chain K	51.1% ....	Aq358	dinG	ATP-dependent helicase (DinG family)	27.9% ....
Aq1378	nuoK2	NADH dehydrogenase I chain K	48.4% ....	Aq322	dnaA	chromosome replication initiator protein DnaA	36.5% ....
Aq1320	nuoL1	NADH dehydrogenase I chain L	39.0% ....	Aq1472	dnaB	replicative DNA helicase	40.3% ....
Aq866	nuoL2	NADH dehydrogenase I chain L	30.2% ..	Aq910	dnaC	DNA replication protein DnaC	26.4% ....
Aq1379	nuoL3	NADH dehydrogenase I chain L	43.1% ....	Aq1008	dnaE	DNA polymerase III alpha subunit	21.9% ....
Aq1321	nuoM1	NADH dehydrogenase I chain M	43.6% ....	Aq1493	dnaG	DNA primase	39.8% ....
Aq1382	nuoM2	NADH dehydrogenase I chain M	36.9% ....	Aq1882	dnaN	DNA polymerase III beta chain	32.1% ....
Aq1322	nuoN1	NADH dehydrogenase I chain N	34.1% ....	Aq932	dnaQ	DNA polymerase III epsilon subunit	40.0% ....
Aq1383	nuoN2	NADH dehydrogenase I chain N	32.8% ....	Aq1855	dnaX	DNA polymerase III gamma subunit	36.6% ....
<b>Lipid metabolism</b>							
Aq2058	aas	2-acylglycerophosphoethanolamine acyltransferase	37.1% ....	Aq1422	dpbF	DNA polymerase beta family	39.1% ....
Aq1206	accA	acetyl-CoA carboxylase alpha subunit	57.1% ..	Aq1693	dpfE	N-terminus of phase SPO1 DNA polymerase	37.3% ....
Aq1363	accB	biotin carboxyl carrier protein	44.6% ....	Aq980	gyrA	DNA gyrase A subunit	43.6% ....
Aq1664	accC1	biotin carboxylase	54.4% ....	Aq1026	gyrB	gyrase B	55.2% ..
Aq1470	accC2	biotin carboxylase	56.5% ....	Aq2057	helX	DNA helicase	49.7% ....
Aq445	accD	acetyl-CoA carboxyltransferase beta subunit	56.9% ....	Aq1484a	himA	DNA binding protein HU	40.2% ....
Aq1717a	accP	acyl carrier protein	71.2% ....	Aq2174	ihfB	integration host factor beta subunit	35.8% ....
Aq813	acsB	holo-[acyl-carrier protein] synthase	30.8% ....	Aq1394	lig	DNA ligase (ATP dependent)	50.8% ....
Aq2104	acs	acetyl-coenzyme A synthetase	54.0% ....	Aq633	ligA	DNA ligase (NAD dependent)	45.7% ....
Aq2103	acs'	acetyl-coenzyme A synthetase c-terminal fragment	61.2% ....	Aq1578	mutL	DNA mismatch repair protein MutL	72.3% ....
Aq1249	cds	phosphatidate cytidylyltransferase	29.2% ....	Aq308	mutS1	DNA mismatch repair protein MutS	77.5% ....
Aq1737	cfi	cyclopropane-fatty-acyl-phospholipid synthase	37.5% ....	Aq1242	mutS2	DNA mismatch repair protein MutS	37.0% ....
Aq892	fabD	malonyl-CoA:Acyl carrier protein transacylase	42.1% ....	Aq1449	mutT	8-Oxo-dGTPase domain (mutT domain)	45.8% ....
Aq1717	fabF	3-oxoacyl-[acyl-carrier-protein] synthase II	58.4% ....	Aq282	mutY1	endonuclease III	53.6% ....
Aq1716	fabG	3-oxoacyl-[acyl-carrier-protein] reductase	52.9% ....	Aq172	mutY2	endonuclease III	51.8% ....
Aq1099	fabH	3-oxoacyl-[acyl-carrier-protein] synthase III	47.0% ....	Aq496	mutY3	endonuclease III	43.4% ....
Aq1552	fabI	enoyl-[acyl-carrier-protein] reductase (NADH)	49.6% ....	Aq1629	nfo	deoxyribonuclease IV	39.0% ....
Aq566	fabZ	(3R)-hydroxymyristoyl-acyl carrier protein dehydratase	58.7% ....	Aq10	nucI	thermoscissal nuclease homolog	36.4% ....
Aq999	fadD	long-chain-fatty-acid CoA ligase	30.0% ..	Aq1495	ogt	O-6-methylguanine-DNA-alkyltransferase	36.9% ....
Aq1638	lplA	lipote-protein ligase A	28.1% ..	Aq1628	pol	DNA polymerase I 3'-5' exo domain	43.2% ....
Aq958	pgsA	phosphatidylglycerophosphate synthase	37.3% ..	Aq1967	polA	DNA polymerase I (Poll)	30.5% ....
Aq2154	pgsA	phosphatidylglycerophosphate synthase	38.9% ....	Aq1610	radC	DNA repair protein RadC	39.0% ....
Aq1101	plxK	PlxK protein	43.7% ....	Aq2150	recA	recombination protein RecA	85.5% ....
<b>Purines, Pyrimidines, Nucleotides and Nucleosides</b>							
Aq094	mrDA	ribonucleotide reductase alpha chain	35.0% ....	Aq2053	recG	ATP-dependent DNA helicase RecG	38.9% ....
Aq1505	mrDF	ribonucleotide reductase beta chain	36.2% ..	recJ	single-strand-DNA-specific exonuclease RecJ	31.8% ....	
Aq568	deoD	purine nucleoside phosphorylase	33.1% ....	recN	recombination protein RecN	27.7% ....	
Aq236	guaA	GMP synthase	58.4% ....	Aq1478	recR	recombination protein RecR	38.3% ....
Aq2023	guaB	inosine monophosphate dehydrogenase	65.4% ....	Aq793	rep	ATP-dependent DNA helicase REP	33.4% ....
Aq544	hpt	hypoxanthine-guanine phosphoribosyltransferase	48.2% ....	Aq1886	sbcD	ATP-dependent dDNA exonuclease	29.9% ....
Aq078	hcaA1	adenylate kinase	50.0% ....	Aq064	ssb	single stranded DNA-binding protein	39.4% ....
Aq1590	ndk	nucleoside diphosphate kinase	48.2% ....	Aq657	topA	topoisomerase I	39.6% ....
Aq1636	prs	phosphoribosylpyrophosphate synthetase	55.2% ....	Aq886	topG1	reverse gyrase	41.6% ....
Aq1290	purA	adenylosuccinate synthetase	49.2% ....	Aq1159	topG2	reverse gyrase	35.1% ....
Aq597	purB	adenylosuccinate frase	52.4% ....	Aq686	uvrA	repair excision nuclease subunit A	61.0% ....
Aq2117	purC	phosphoribosylaminoimidazole-succinocarboxamide synthase	52.5% ....	Aq1856	uvrB	repair excision nuclease subunit B	53.9% ....
Aq742	purD	phosphoribosylamine-glycine ligase	54.2% ....	Aq2126	uvrC	repair excision nuclease subunit C	32.5% ....
Aq1178	purE	phosphoribosylaminoimidazole carboxylase	64.6% ....	<b>Transcription</b>			
Aq1175	purF	amidophosphoribosyltransferase	42.7% ....	RNA polymerase and transcription factors			
Aq1963	purH	phosphoribosylaminoimidazolecarboxamide formyltransferase	48.2% ....	Aq613	deaD	ATP-dependent RNA helicase DeaD	42.3% ....
Aq245	purK	phosphoribosyl aminoimidazole carboxylase	35.6% ....	Aq357a	flgM	anti sigma factor FlgM	20.6% ....
Aq1836	purL	phosphoribosylformylglycinamide synthase II	49.3% ....	Aq2118	fla	RNA polymerase sigma factor FlhA	37.2% ....
Aq769	purM	phosphoribosylformylglycinamide cyclo-ligase	50.0% ....	Aq259	nusA	transcription termination NusA	45.4% ....
Aq857	purN	phosphoribosylglycinamide formyltransferase	48.3% ....	Aq133	nusB	transcription termination NusB	32.3% ....
Aq1105	purQ	phosphoribosylformylglycinamide synthase I	51.1% ....	Aq1931	nusG	transcription antitermination protein NusG	46.3% ..
Aq1818	purU	formyltetrahydrofolate deformylase	56.3% ....	Aq873	rho	transcriptional terminator Rho	59.6% ....
<b>Pyrimidines</b>							
Aq410	carA	carbamoyl phosphate synthetase small subunit	52.2% ....	Aq070	rpoA	RNA polymerase alpha subunit	40.4% ....
Aq1172	carB	carbamoyl-phosphate synthase large subunit	60.7% ....	Aq1939	rpoB	RNA polymerase beta subunit	44.0% ....
Aq2101	carB	carbamoyl-phosphate synthase, large subunit	63.1% ....	Aq1945	rpoC	RNA polymerase beta prime subunit	46.9% ....
Aq2153	cmk	cytidylate kinase	38.5% ....	Aq1490	rpoD	RNA polymerase sigma factor RpoD	41.6% ....
Aq1607	dcl	deoxycytidine triphosphate deaminase	39.5% ....	Aq599	rpoN	RNA polymerase sigma factor RpoN	30.6% ....
Aq220	dut	deoxyuridine 5-triphosphate nucleotidylhydrolase	39.5% ....	Aq1452	rpoS	RNA polymerase sigma factor RpoS	40.5% ....
Aq409	pyrB	aspartate carbamoyltransferase catalytic chain	42.0% ....	<b>RNA modification</b>			
Aq806	pyrC	dihydroorotase	37.3% ....	Aq1816	ksgA	dimethyladenosine transferase	36.1% ....
<b>Regulation</b>							
RNA, delta-2-isopentenylpyrophosphate (IPP) transferase							
poly A polymerase							
poly A polymerase							
polyribonucleotide nucleotidyltransferase							
queuosine biosynthesis protein							
RNase III							
RNase HII							
RNase PH							
rRNA methylase SpoU							
queuine tRNA-ribosyltransferase							
N2,N2-dimethylguanosine tRNA							

