and tree bisection-reconnection (TBR) branch swapping. Bootstrap analyses used 100 heuristic searches, with one random taxon addition sequence per replicate, and with MAXTREES limited to 10. Based on previous analyses[29], the heterobasidiomycetes *Auricularia* and *Dacrymyces* were used for rooting purposes. Terminal taxa were coded as ectomycorrhizal, non-ectomycorrhizal or uncertain, and ancestral state reconstructions were performed on all trees with equally weighted parsimony using MacClade 3.0 (ref. 11). Maximum likelihood estimates of support for alternative states at selected nodes of one tree selected at random (with branch lengths estimated from molecular data using parsimony) were calculated using Discrete[12]. Because Discrete requires that all terminal taxa be scored for the character of interest, the estimated states from parsimony optimizations were assigned for the eight species that were coded as uncertain (Fig. 1). The 'local' method for estimating support for alternative ancestral states was used, with a difference of two units in log likelihood ($\Delta \log L > 2$)taken as an approximate criterion of significance[12].

Received 14 March; accepted 20 July 2000.

1. Price, P. W. in *Symbiosis As A Source Of Evolutionary Adaptation* (eds Margulis, L. & Fester, R.) 262–272 (MIT Press, Cambridge, MA, 1991).
2. Bronstein, J. L. Conditional outcomes in mutualistic interactions. *Trends Ecol. Evol.* **9,** 214–217 (1994).
3. Thompson, J. N. *The Coevolutionary Process* (Univ. Chicago Press, Chicago, 1994).
4. Herre, E. A., Knowlton, N., Mueller, U. G. & Rehner, S. A. The evolution of mutualisms: exploring the paths between conflict and cooperation. *Trends Ecol. Evol.* **14,** 49–53 (1999).
5. Pellmyr, O. & Huth, C. J. Evolutionary stability of mutualism between yuccas and yucca moths. *Nature* **372,** 257–260 (1994).
6. Pirozynski, K. A. & Malloch, D. W. The origin of land plants: a matter of mycotrophism. *Biosystems* **6,** 153–164 (1975).
7. Malloch, D. W., Pirozynski, K. A. & Raven, P. H. Ecological and evolutionary significance of mycorrhizal symbioses in vascular plants (a review). *Proc. Natl Acad. Sci. USA* **77,** 2113–2118 (1980).
8. Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* 2nd edn (Academic, San Diego, 1997).
9. Hawksworth, D. L., Kirk, P. M., Sutton, B. C. & Pegler, D. N. *Dictionary Of The Fungi* 8th edn (CAB International, Wallingford,, 1996).
10. Swofford, D. L. *PAUP* 4. 0b2a* (Sinauer, Sunderland, 1999).
11. Maddison, W. P. & Maddison, D. R. *MacClade version 3* (Sinauer, Sunderland, 1992).
12. Pagel, M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48,** 612–622 (1999).
13. Fitter, A. H. & Moyersoen, B. Evolutionary trends in root-microbe symbioses. *Philos. Trans. R. Soc.* (Ser. 351) **1345,** 1367–1375 (1996).
14. Trappe, J. M. Fungus associates of ectotrophic mycorrhizae. *Bot. Rev.* **28,** 538–606 (1962).
15. Ashton, D. H. Studies on the mycorrhizae of *Eucalyptus regnans* F. Muell. *Aust. J. Bot.* **24,** 723–741 (1976).
16. Malajczuk, N., Molina, R. & Trappe, J. M. Ectomycorrhiza formation in *Eucalyptus*. I. Pure culture syntheses, host specificity and mycorrhizal compatibility with *Pinus radiata*. *New Phytol.* **91,** 467–482 (1982).
17. Smits, W. T. M. *Dipterocarpaceae: Mycorrhizae and regeneration* (Tropenbos Foundation, Wageningen, 1994).
18. Nuhamara, S. T., Hadi, S. & Bimaatmadja, E. I. in *Proceedings of the 6th North American Conference on Mycorrhizae* (ed. Molina, R.) 439 (Forest Research Laboratory, Oregon State Univ., 1985).
19. Danielson, R. M. Ectomycorrhizal associations in jack pine stands in north-eastern Alberta. *Can. J. Bot.* **62,** 932–939 (1984).
20. Molina, R. & Trappe, J. M. Patterns of ectomycorrhizal host specificity and potential among pacific northwest conifers and fungi. *Forest Sci.* **28,** 423–458 (1982).
21. Kropp, B. R. & Trappe, J. M. Ectomycorrhizal fungi of *Tsuga heterophylla*. *Mycologia* **74,** 479–488 (1982).
22. Bruns, T. D. Thoughts on the processes that maintain local species diversity of ectomycorrhizal fungi. *Plant Soil* **170,** 63–73 (1995).
23. Hibbett, D. S. & Thorn, R. G. in *The Mycota* Vol. VII *Systematics and Evolution* (eds McLaughlin, D. J., McLaughlin, E. G. & Lemke, P. A.) (Springer, in the press).
24. Cooke, R. C. & Rayner, A. D. M. *Ecology of Saprotrophic Fungi* (Longman, New York, 1984).
25. Leake, J. R. & Read, D. J. in *The Mycota* Vol. IV *Environmental and Microbial Relationships* (eds Wicklow, D. T. & Söderström, B. E.) 281–301 (Springer, Berlin, 1997).
26. Pellmyr, O., Leebens-Mack, J. & Huth, C. J. Non-mutualistic yucca moths and their evolutionary consequences. *Nature* **380,** 155–156 (1996).
27. Clay, K. in *Coevolution of Fungi with Plants and Animals* (eds Pirozynski, K. A. & Hawksworth, D. L.) 79–105 (Academic, San Diego, 1988).
28. Lamb, R. J. Effect of D-glucose on utilization of single carbon sources by ectomycorrhizal fungi. *Trans. Br. Mycol. Soc.* **63,** 295–306 (1974).
29. Swann, E. C. & Taylor, J. W. Higher taxa of basidiomycetes: an 18S rRNA perspective. *Mycologia* **85,** 923–936 (1993).

Supplementary information is available on *Nature's* World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank M. Pagel for providing a copy of Discrete; USDA, DAOM and other sources for fungal samples; J.-M. Moncalvo and R. Vilgalys for access to unpublished sequences; and J. Bronstein for helpful comments. This work was supported by grants from the NSF.

Correspondence and requests for materials should be addressed to D.S.H. (e-mail: dhibbett@black.clarku.edu). Sequences have been deposited in GenBank (accession numbers AF287817–AF287891), and the data matrix and tree have been deposited in TreeBASE (http://phylogeny.harvard.edu/treebase).

The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*

Andreas Ruepp*, Werner Graml*, Martha-Leticia Santos-Martinez*, Kristin K. Koretke†, Craig Volker†, H. Werner Mewes‡, Dmitrij Frishman‡, Susanne Stocker‡, Andrei N. Lupas† & Wolfgang Baumeister*

* *Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany*
† *Bioinformatics, Smith Kline Beecham Pharmaceuticals, Collegeville, Pennsylvania 19426, USA*
‡ *GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for Protein Sequences, Am Klopferspitz 18a, D-82152 Martinsried, Germany*

Thermoplasma acidophilum **is a thermoacidophilic archaeon that thrives at 59 °C and pH 2, which was isolated from self-heating coal refuse piles and solfatara fields[1,2]. Species of the genus *Thermoplasma* do not possess a rigid cell wall, but are only delimited by a plasma membrane. Many macromolecular assemblies from *Thermoplasma*, primarily proteases and chaperones, have been pivotal in elucidating the structure and function of their more complex eukaryotic homologues[3,4]. Our interest in protein folding and degradation led us to seek a more complete representation of the proteins involved in these pathways by determining the genome sequence of the organism. Here we have sequenced the 1,564,905-base-pair genome in just 7,855 sequencing reactions by using a new strategy. The 1,509 open reading frames identify *Thermoplasma* as a typical euryarchaeon with a substantial complement of bacteria-related genes; however, evidence indicates that there has been much lateral gene transfer between *Thermoplasma* and *Sulfolobus solfataricus*, a phylogenetically distant crenarchaeon inhabiting the same environment. At least 252 open reading frames, including a complete protein degradation pathway and various transport proteins, resemble *Sulfolobus* proteins most closely.**

Two basic approaches have been taken to genome sequencing. The statistical approach ('shotgun sequencing') relies on the determination of a highly redundant set of random genomic DNA sequences, which are assembled in the computer, the gaps remaining to be closed by other methods[5]. This approach rapidly yields 90–98% of the genomic information, but requires an extensive robotic infrastructure. The directed approach relies on a complete mapping of the genome, followed by the sequencing of large overlapping genomic fragments by shotgun sequencing or primer walking[6]. This approach reduces the infrastructure requirements but is slowed down by the need for a genetic map.

One of our aims was to establish a strategy for sequencing microbial genomes in reasonable time without extensive infrastructure. This strategy ('shotgun primer walking') combines features of the statistical and directed methods. After construction of several phagemid libraries and one cosmid library, inserts from randomly chosen clones were sequenced from the ends by primer walking (see Supplementary Information). Sequencing was stopped in regions that were redundant with already determined sequences. In total, 400 phagemids, covering 850 kilobase-pairs (kbp) (54%), and 469 cosmids, covering 1,533 kbp (98%), were partially or fully sequenced. With an average insert length of 40 kbp, the clones of the cosmid library statistically covered the genome 12 times. Because cosmids are susceptible to recombination events, the reverse strand of cosmid DNA was always sequenced with DNA templates from other cosmid clones covering the same region, or

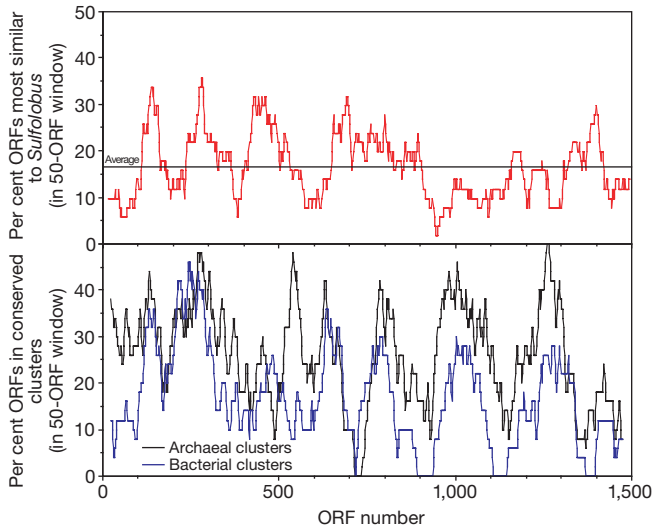

Figure 1 Chromosomal distribution of *Thermoplasma* ORFs most similar to *Sulfolobus* (top panel), and of ORFs found in conserved gene clusters (bottom panel). ORFs were considered to occur in a conserved cluster if their chromosomal vicinity was also observed in at least 1 of 13 complete reference genomes (6 archaeal, 7 bacterial; see Methods). The approximate locations of proteins most similar between *Thermoplasma* and *Sulfolobus* mentioned in the text are indicated in the top panel,

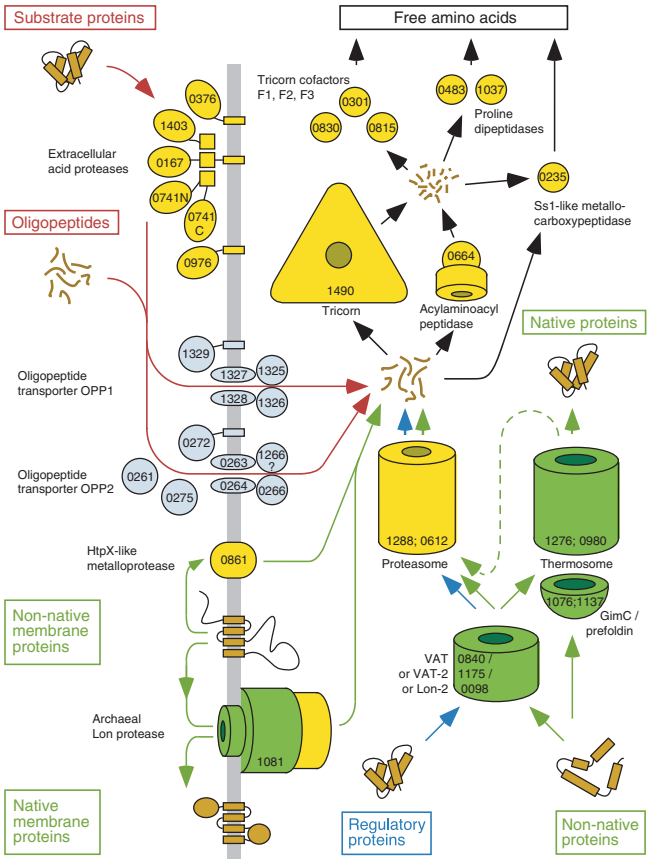

Figure 2 Overview of the main proteolytic pathways in *Thermoplasma*. Red arrows indicate the metabolic pathway, green arrows the 'quality control' pathway for damaged proteins and blue arrows the regulatory pathway. Chaperones are shown in green, proteases in yellow and transporters in blue. The numbers indicate the ORF code; that is, 1490 corresponds to Ta1490.

from polymerase chain reaction (PCR) fragments generated from genomic DNA. As the project advanced, fragments were assembled into progressively larger contigs, until 4 gaps of 33 kbp in total were left. These were closed by four PCR fragments, which were sequenced by primer walking. By this method, we completed the 1.56 Mbp genome of *T. acidophilum* with 7,855 sequencing reactions. For comparison, shotgun sequencing of the 1.86 Mbp *Thermotoga maritima* genome, which has the same G+C content, required over 30,000 sequencing reactions.

During the annotation process, we detected ten open reading frames (ORFs) with putative frameshifts. The corresponding genomic regions were amplified by PCR and resequenced with the primers obtained during primer walking. Four of the frameshifts were confirmed, and six had to be corrected. All errors occurred in phagemids, which, unlike cosmids, were sequenced on both strands but without using template DNA from other sources. This suggests that in genome projects operating with low sequence coverage the two strands should be sequenced with template DNA from different sources. It also shows, however, that high-quality data can be obtained with only twofold sequence coverage.

Beyond the advantages discussed here, shotgun primer walking generates an extensive primer set, covering the whole genome. This represents a powerful tool, which can be used to determine rapidly from PCR fragments differences in related DNA sequences, for example between strains of pathogenic bacteria or between individuals in human populations.

The genome of *T. acidophilum* DSM 1728, one of the smallest among free-living organisms, consists of a single circular chromosome of 1,564,905 bp (Table 1). We did not detect any plasmids, either by biochemical methods or through DNA sequencing. A single plasmid of 15.2 kbp has previously been reported in other isolates of *T. acidophilum*[7]. The origin of replication was estimated using the cumulative skew of the orientation of ORFs as well as different nucleotide skews[8]. The centre of a 477-bp intergenic region in this area was chosen as nucleotide 1.

Thermoplasma contains one copy of each of the ribosomal RNAs, which are dispersed in the genome[9] and separated by at least 52 kbp. This is different from all other known archaea, where at least two of the three genes are contiguous on the chromosome.

We identified 1,509 ORFs in the genome (see Supplementary Information), of which over one-third have homologues in all three domains of life (see Supplementary Information). After manual curation, 823 ORFs (55%) were considered similar to functionally annotated proteins, 446 (29%) resembled hypothetical ORFs in other organisms, and 240 (16%) were not recognizably similar to other sequences in current databases ('singletons'). In total, we could assign 685 ORFs (46%) to categories of the MIPS Functional Catalogue (see Supplementary Information). The only category that deviates significantly from other archaea is transport facilitation, which occupies a larger proportion of the genome than in any other archaeon, but is not exceptional when compared with some bacteria (*Thermotoga*, *Deinococcus*). By conservative search methods, we could match 620 domains of 537 ORFs (35.6%) to proteins of known structure (Table 1; Supplementary Information). The distribution of structural classes resembles that of other prokaryotic genomes, but shows a striking overrepresentation of three-layer (αβα) sandwich folds (43% of all matches), which primarily encompass enzymes.

About one-third of the ORFs (484; 32%) occurred in 139 conserved gene clusters, as judged relative to a set of 7 bacterial and 6 archaeal genomes (see Methods). Twenty-two of these clusters occurred only in *Thermoplasma* and bacteria and are presumably due to lateral transfer. The clusters occur in 'hot spots' on the chromosome (Fig. 1), suggesting that entire gene regions may have been acquired in discrete genetic events. The largest detected clusters encode ribosomal proteins (Ta1249–Ta1271), NADH dehydrogenase subunits (Ta0959–Ta0970), precorrin biosynthesis

proteins (Ta0653–Ta0660) and flagellar proteins (Ta0553–Ta0560).

In the thermoacidophiles *Thermoplasma* and *Sulfolobus*, glucose degradation proceeds by a non-phosphorylated variant of the Entner–Doudoroff pathway[10] (see Supplementary Information), in which the first step is catalysed by glucose dehydrogenase. The acetyl-CoA produced in this pathway enters the oxidative tricarboxylic acid (TCA) cycle. A complete set of TCA enzymes is recognizable in *Thermoplasma*, many of which have been studied experimentally. In the glycolysis/gluconeogenesis pathway (Embden–Meyerhof–Parnas; EMP), homologues are recognizable for most enzymes, but not for phosphofructokinase and fructose bisphosphate aldolase, so the presence of a working EMP pathway[11] cannot be confirmed.

Thermoplasma is microaerophilic and contains several respiratory chain proteins, such as electron transfer flavoproteins (Ta0429, Ta0329 and Ta0212) and cytochrome *b* homologues (Ta1222, Ta1228 and Ta1003). Two of the latter are found in gene clusters with Rieske iron–sulphur proteins (Ta1223 and Ta1229). In addition, we identified two homologues of a cytochrome *bd* quinol oxidase[12] (Ta1484 and Ta0992).

Thermoplasma is also able to gain energy anaerobically by sulphur respiration[2]. Unexpectedly, no homologues of the genes that mediate sulphur respiration in *Archaeoglobus fulgidus*[13] were found. Instead, *Thermoplasma* contains homologues of AsrA (Ta0046) and AsrB (Ta0047), which mediate dissimilatory sulphur reduction in *Salmonella typhimurium*[14]. A sulphide-quinone reductase homologue (Ta1129) may also be involved in sulphur respiration.

Most studied archaea are motile and previous work suggested that they use a chemotaxis signal transduction pathway resembling that of bacteria[15]. However, *Thermoplasma* is flagellated and motile[2], yet lacks recognizable chemotaxis proteins. This situation is also found in other archaea (*Methanococcus jannaschii*, *Aeropyrum pernix*) and bacteria (*Aquifex aeolicus*), suggesting the presence of an alternative, unrelated signal transduction pathway. Another puzzle concerning motility is the lack of a cell wall, which anchors the flagella in prokaryotes. It is unclear what structure could function as the stator for flagellar rotation in *Thermoplasma*.

Two chaperones of *Thermoplasma* have been studied experimentally, the thermosome[3] and the protein 'unfoldase' VAT[16,17]. Additional chaperones recognizable in the genome (Table 2) include three Hsp20-like proteins (Ta0437, Ta0471 and Ta0864) and a complete Hsp60 system, consisting of two thermosome subunits (α, Ta0980; and β, Ta1276) and two prefoldin/Gim subunits (α, Ta1076; and β, Ta1137). Unlike many archaea, *Thermoplasma* also contains an Hsp70 system (DnaK, Ta1087; DnaJ, Ta1088; and GrpE, Ta1086), whose genes are clustered on the chromosome in contrast to the endogenous Hsp60 genes.

Thermoplasma contains several 'unfoldases' (AAA+ proteins)[17], which are frequently associated with proteolytic degradation[18]. These are the Cdc48 homologues VAT (Ta0840) and VAT-2 (Ta1175), an archaeal-type Lon protease (Ta1081) and a Lon-related ATPase lacking the protease domain (Ta0098). Surprisingly, *Thermoplasma* is missing the proteasomal AAA ATPase (PAN; ref. 19), which has hitherto been found in all archaea. It is

Table 1 Genome features

General features of the *Thermoplasma acidophilum* genome	
Length of sequence	1,564,905 bp
G+C ratio	46%
Open reading frames	1509
Average ORF length	909 bp
Protein coding regions	87%
Ribosomal RNAs	1 5S; 1 16S; 1 23S
tRNAs	45
7S RNA	1
Repetitive elements	Coordinates
Short, non-coding repeats	846,047–849,104
gtaaaatagaaccttaataggattgaaag (29 nt; 46 copies)	
Long repetitive elements	
Ire1 (185 nt; 2 copies; 5′ flanking sequence of Ta0698	739,187–739,371
and Ta0708)	and 751,166–751,350
Ire2 (329 nt; 2 copies; 3′ flanking sequence of Ta0698	739,451–739,779
and Ta0708)	and 751,430–751,758
Ire3 (654 nt; 2 copies, contains Ta1170/1171	1,243,940–1,244,593
and Ta1414/Ta1415)	and 1,464,457–1,465,110
Long, inverted repeats	
Iir (169 nt; 2 copies)	184,427–184,595
	and 397,089–397,257
Chromosomal coding sequences	
No. similar to known proteins	823 (55%)
No. similar to proteins of unknown function	446 (29%)
No. without a database match	240 (16%)
Inteins	1 (Ta004)
No. of protein domains similar to known structures	620 (in 537 proteins)
Mostly α (CATH classfication)	19%
Mostly β (CATH classification)	12%
α and β (CATH classification)	69%
No. predicted to contain one transmembrane helix	114 (7.6%)
No. predicted to contain two or more transmembrane helices	268 (17.8%)
No. predicted to contain coiled-coil domains	34 (2%)

nt, nucleotide.

Table 2 The chaperone complement of *T. acidophilum*

Class	Type	ORF	Annotation	Remarks
Hsp20	Hsp20/sHSP	Ta0437*	ATPase containing hsp20 domain	Hsp20-like C-terminal domain absent in other archaea; ATPase domain similar to *E. coli* ArsA
		Ta0471	Small heat-shock protein (hsp20) related	
		Ta0864*	Small heat-shock protein (hsp20) related	
Hsp60	Thermosome	Ta0980	Thermosome α-chain	
		Ta1276	Thermosome β-chain	
	GimC	Ta1076	GimC, α-subunit related	
		Ta1137	GimC, β-subunit related	
Hsp70	Hsp70/DnaK	Ta1087*	Probable DnaK-type molecular chaperone	
	Hsp40/DnaJ	Ta1088*	Heat-shock protein DnaJ related	
	Hsp23/GrpE	Ta1086*	Heat-shock protein GrpE related	
AAA+†	Cdc48	Ta0840	VAT ATPase (VCP-like ATPase)	Ta1064 and Ta1217 resemble the double-psi barrel domain of VAT contains only one AAA domain; N-domain does not resemble that of other Cdc48 proteins
		Ta1175	VAT-2 protein	
	Archaeal Lon	Ta1081	ATP-dependent proteinase La (Lon) related	Membrane-anchored by a helical hairpin inserted into the ATPase domain consists of only the ATPase domain; C-terminal protease domain missing
		Ta0098	Lon-related ATPase	
PDI	Thioredoxin	Ta0866*	Thioredoxin related	
		Ta0125	Glutaredoxin related	Contains two glutaredoxin-like domains
PPIases	FKBP	Ta1011	Peptidyl-prolyl *cis-trans* isomerase related	
Other	CsaA	Ta1046*	Probable chaperone (csaA)	
		Ta1284	Chaperone (csaA) related	Similar to the C-terminal domain of archaeal and bacterial Met tRNA synthesases

* ORFs of probable bacterial origin.
† *Thermoplasma* contains four other AAA+ proteins with potential chaperone-like activity: Ta0576 and Ta0920, which may act as enzyme-specific factors involved in the maturation of cofactor-containing enzymes; and Ta0799 and Ta1285, which may act in the assembly and/or disassembly of replication complexes.

conceivable that Ta0098, VAT-2 or VAT, which has protein unfolding activity[17], can substitute for this function.

Other chaperones include two protein disulphide isomerases (Ta0125 and Ta0866), a peptidyl-prolyl isomerase of the FKBP family (Ta1011) and two homologues of cold-shock protein CsaA (Ta1284, Ta1046). Ta1284 is most similar to the carboxy-terminal domain of archaeal methionyl-tRNA synthetases (MetRS); this domain is missing from *Thermoplasma* MetRS. Overall, a surprising number of chaperones from *Thermoplasma* appear to have been acquired from bacteria (Table 2).

The pathway for polypeptide degradation in *Thermoplasma* is well understood[20]. Unfolded proteins are cleaved into fragments of 6–15 residues by the proteasome (α, Ta1288; and β, Ta0612), whose structure and activity have been extensively studied[4]. These fragments are further degraded by the tricorn peptidase (Ta1490) to di-, tri- and tetrapeptides, which are hydrolysed to free amino acids by three tricorn-interacting factors F1 (Ta0830), F2 (Ta0301) and F3 (Ta0815). It is, however, unclear how proteins are targeted to this degradation pathway, because ubiquitin, whose existence in *T. acidophilum* had been proposed on the basis of peptide sequences[21], is not present in this genome (or indeed in any other archaeon). It is also unclear how proteins are unfolded before degradation by the proteasome, as *Thermoplasma* is missing a proteasomal AAA ATPase.

In addition to the 6 proteins involved in this degradation pathway, *Thermoplasma* contains at least 23 putative proteases (Table 3, Fig. 2), as well as several α/β hydrolases of unknown specificity and a broad-spectrum deacetylase with potential glutamate carboxypeptidase activity (Ta0934). Few proteases seem to be of bacterial origin; however, nine are most similar to proteins of the phylogenetically distant crenarchaeon *Sulfolobus solfataricus* (http://niji.imb.nrc.ca/sulfhome/results.html). These proteins are tricorn, its interacting factors and five extracellular, multidomain acid proteases (Ta1403, Ta0167, Ta0376, Ta0741 and Ta0976), three of which are membrane-anchored by C-terminal transmembrane sequences. It is unclear how *Thermoplasma* retains unanchored secreted proteins without a cell wall, but we note that the two secreted proteases and one of the anchored ones (Ta0167) contain a domain conserved in many archaeal S-layer proteins, which may serve as a scaffold.

Overall, ten *Thermoplasma* proteases are probably extracellular. These include two membrane-anchored signal peptidase I homologues (Ta0151 and Ta0378), which are closely related to their eukaryotic homologues and may yield useful model systems for the more elaborate eukaryotic complex, a membrane-anchored serine peptidase of the S49 family (Ta1083), a probable *O*-sialoglycoprotein endopeptidase (Ta0324) and a metalloprotease of the M6 family (Ta0728). The two latter proteins are probably secreted, on the basis of experimental findings from other organisms. Like their homologues[22], however, they lack recognizable signal sequences. In total, 11 *Thermoplasma* proteases are membrane-associated, including the archaeal Lon protein, which is anchored to the membrane from the cytoplasmic side by a helical hairpin, a structure present in all archaea but missing in bacterial proteins. As the membrane-bound protease FtsH, which is essential in bacteria, is absent from archaea, its function may be assumed by the archaeal Lon protease.

As discussed, *Thermoplasma* surprisingly lacks ubiquitin, a proteasomal ATPase and an archaeal-type sulphur respiration system. We searched systematically for further differences to other archaeal genomes using clusters of orthologous groups (COGs; http://www.ncbi.nlm.nih.gov/COG). The first four published archaeal genomes share a 'stable core' of 543 COGs; thirty-six of these appear to be missing in *Thermoplasma* (http://www.biochem.mpg.de/baumeister/genome). The most notable are DNA topoisomerase VI and eukaryal-like histones. Topoisomerase VI (ref. 23) is a divergent type II topoisomerase present only in archaea. In its place, *Thermoplasma* contains a bacterial-type DNA gyrase (GyrA, Ta1054; and GyrB, Ta1055).

Table 3 The protease complement of *T. acidophilum*

Clan	Family	ORF	Annotation	Cellular location	Remarks
AX	A5	Ta1403*	Thermopsin related	ext	
		Ta0167*	Thermopsin related	ext/anc	Anchored by C-terminal transmembrane helix
		Ta0741*	Acid protease	ext	Protease domain 1, residues 78–386
	A7	Ta0376*	Pseudomonapepsin related	ext/anc	Anchored by C-terminal transmembrane helix
		Ta0741*	Acid protease	ext	Protease domain 2, residues 810–1,257; family A7 probably belongs to clan SB
		Ta0976*	Pseudomonapepsin related	ext/anc	Anchored by C-terminal transmembrane helix
MA	M1	Ta0301*	Tricorn cofactor F2	cyt	F2 and F3 nearest sequence neighbours
		Ta0815*	Tricorn cofactor F3	cyt	
	M6	Ta0728	Metalloprotease	ext?	Divergent family member, but active-site signature most similar to M6
	M48	Ta0861	Heat-shock protein (HtpX) related	mem	
MG	M24A	Ta1439	Methionine aminopeptidase I related	cyt	
	M24B	Ta1037	Proline dipeptidase related	cyt	
		Ta0483	Proline dipeptidase related	cyt	
MH	M40	Ta0235†	Carboxypeptidase Ss1 related	cyt	
MK	M22	Ta0324	*O*-sialoglycoprotein endopeptidase-related	ext?	Contains a C-terminal protein kinase domain
MX	M50	Ta1274	S2P protease-related	mem	Families M50 and M51 probably belong to clan MA
	M51	Ta1344	SpoIVFB metalloprotease-related	mem	
		Ta0562	SpoIVFB metalloprotease-related	mem	
SC	S9C	Ta0664	Acylaminoacyl-peptidase related	cyt	
	S33	Ta0830*	Prolyl iminopeptidase/Tricorn cofactor F1	cyt	
SF	S26B	Ta0151	Signal peptidase I related	ext/anc	Ta0151 and Ta0378 nearest sequence neighbours; anchored by N- and C-terminal transmembrane helices
		Ta0378	Signal peptidase I related	ext/anc	
SK	S49	Ta0074	Protease IV related	cyt	
		Ta1083	Serine protease	ext/mem	Contains C-terminal integral membrane domain
SX	S16	Ta1081	ATP-dependent proteinase La (Lon) related	cyt/anc	Transmembrane helical hairpin inserted in the ATPase domain
	S41	Ta1490*	Tricorn core protease	cyt	
PB	T1A	Ta1288	Proteasome α-subunit	cyt	Non-peptidase homologue of the β-subunit
		Ta0612	Proteasome β-subunit	cyt	
	T3	Ta0994†	Glutamyltransferase related	cyt	
UX	U46	Ta0465	PfpI endopeptidase related	cyt	Probably belongs to cysteine protease family C26

Proteases classified according to the MEROPS classification[22] (http://www.merops.co.uk). anc, anchored; cyt, cytoplasmic; ext, extracellular; mem, membrane.
* ORFs most similar to proteins of *Sulfolobus solfataricus*.
† ORFs of probably bacterial origin.

Thermoplasma also contains proteins not present in any other archaeal genome (68, after elimination of singletons; http://www.biochem.mpg.de/baumeister/genome). Most noteworthy was Hta (Ta0093), the first archaeal DNA-binding protein to be investigated[24], which is closely related to bacterial proteins and appears to substitute functionally for the missing histones. Other proteins include an *egghead* homologue (Ta0213), a membrane-anchored protein kinase (Ta0488) and a Ras-like GTPase (Ta1192).

In the past, a *Thermoplasma*-like organism has been debated as a possible ancestor of the eukaryotic cytoplasm. The genome sequence, however, shows that *Thermoplasma* is a typical archaeon with a fairly large protein complement of bacterial origin. We could not identify any of the proteins peculiar to eukaryotes, such as subunits of the nuclear pore complex, the exocyst or the cytoskeleton. However, *Thermoplasma* has been reported to contain a cytoskeleton[25], and we identified in genome searches a cytosolic coiled-coil protein (Ta1488), which is conserved in all archaea and seems analogous to intermediate filament proteins.

During the comparison to other archaea we noted that 252 *Thermoplasma* ORFs (17%) were most similar to proteins of *S. solfataricus*, a higher number than for any other organism in our reference set—even though this genome sequence is still incomplete. Of these, 60 appear to function in transport and are frequently clustered (for example, Ta1325–Ta1329, Ta0261–Ta0275 and Ta0126–Ta0146). Many others are involved in metabolism (including energy metabolism) or in proteolysis, or are secreted or membrane-bound ORFs of unknown function. With a few notable exceptions, such as DNA polymerase (Ta0450) or Rad50 (Ta0157), hardly any are involved in information processes (replication, transcription and translation). Although it has been observed previously that in prokaryotes core information processing generally tracks organism phylogeny, whereas metabolism is strongly affected by lateral transfer, it is highly unusual to see such a large number of genes transferred between two phylogenetically distant organisms. We propose that the adaptation to an extreme environment shared by few other organisms has led to substantial genetic exchange. This may have proceeded by only a few large genetic events, resulting in the skewed distribution of *Sulfolobus*-like genes in the *Thermoplasma* genome (Fig. 1).

Thermoplasma inhabits a hot and highly acidic environment, sometimes as low as pH 0.5, in which few organisms are viable. It has adapted to scavenging nutrients from the decomposition of organisms killed by the extreme acidity and requires yeast, bacterial or meat extract when grown in culture. There is an absolute growth requirement for basic oligopeptides[26] and analysis of the genome shows the presence of a complete degradation chain for exogenous polypeptides, starting with a set of large extracellular proteases, over two oligopeptide transport systems, to a cytosolic degradation machine consisting of tricorn and its cofactors (Fig. 2). All proteins of this degradation chain are most similar to cognate proteins of *Sulfolobus*, as are many other proteins involved in transport and metabolism. Thus, it would seem that two classes of genes can be distinguished in *Thermoplasma*: One is primarily composed of constitutive, 'housekeeping' genes, which generally reflect the phylogenetic origin of the organism. The other mostly contains 'life-style' genes, including but not limited to metabolism, which are tailored to a specific environment and are shared between phylogenetically distant organisms within one ecological niche.

Methods

Genome sequencing analysis methods

Analysis methods are given in the Supplementary Information.

Annotation

In a first approximation, gene prediction was performed automatically by the Orpheus software[27] allowing for ORFs longer than 150 bases and for overlaps between ORFs no longer than 30% of the length of the shorter overlapping ORF. As an additional precaution we modified the algorithm to keep all ORFs longer than 450 bases as part of the preliminary ORF set destined for manual inspection. Selection of putative gene starts was assisted by ribosome-binding-site detection; however, the information content of these sites in *T. acidophilum* proved to be insufficient for improving prediction results. The predicted ORF complement was then manually refined on the basis of extrinsic evidence.

The main vehicle for automatic and manual annotation of gene products was the PEDANT software suite[28]. Complete annotation of the genome, including DNA and protein viewers, extensive protein reports, multiple functional and structural categories and search capabilities is available at http://pedant.mips.biochem.mpg.de.

Global comparisons of gene complement, searches for conserved gene clusters, and in-depth annotation of protein families were performed in the SmithKline Beecham AiBi Toolkit, a suite of software tools running on top of a relational database containing 68 partial and complete genomes of archaeal (8), bacterial (55) and eukaryotic origin (5), in addition to the non-redundant public sequence database from NCBI (http://www.ncbi.nlm.nih.gov). Sequence comparisons were made using BLAST2 and PSI-BLAST[29], with low-complexity regions (including coiled coils and transmembrane regions) masked out and a significance threshold of e^{-3}. Conserved gene clusters were identified by comparing pairs of ORFs separated by at most three ORFs between *Thermoplasma* and a reference set of six archaeal and seven bacterial genomes, chosen to broadly represent archaea and deep-branching bacteria. The archaeal genomes were *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus*, *Pyrococcus abyssi*, *Pyrococcus furiosus* and *Aeropyrum pernix*; and the bacterial genomes were *Escherichia coli*, *Bacillus subtilis*, *Clostridium acetobutylicum*, *Thermotoga maritima*, *Aquifex aeolicus*, *Deinococcus radiodurans* and *Synechocystis sp*. For opening a cluster, both members of the ORF pair had to have BLAST2 matches in chromosomal vicinity at e^{-10} in at least one of the reference genomes. Clusters were then extended by relaxing the significance cutoff to e^{-3} for proteins satisfying the distance cutoff. Finally, clusters sharing at least one ORF were merged.

Protein fold predictions were made using the PSI-Blast-based search routine SENSER[30], and results were compared against the CATH classification (see Supplementary Information). We report only the PSI-Blast results because, with the chosen settings, they have an estimated error rate of less than 5%. However, SENSER returned predictions for an additional 152 proteins with stringent settings, and 205 proteins with relaxed settings, for a total of 727 proteins (48% of the ORF complement; http://www.biochem.mpg.de/baumeister/genome). Coiled coils were predicted with the COILS program (http://www.ch.embnet.org/software/COILS_form.html), and signal sequences with SignalP (http://www.cbs.dtu.dk/services/SignalP/) and PHYSEAN.

Received 13 April; accepted 6 July 2000.

1. Darland, G., Brock, T. D., Samsonoff, W. & Conti, S. F. A thermophilic acidophilic mycoplasm isolated from a coal refuse pile. *Science* **170,** 1416–1418 (1970).
2. Segerer, A. & Stetter, K. O. in *The Prokaryotes* (eds Balows, A., Trüper, H. G., Dworkin, M., Harder, W. & Schleifer, K. H.) 712–718 (Springer, New York, 1992).
3. Gutsche, I., Essen, L. O. & Baumeister, W. Group II chaperonins: New TRiC(k)s and Turns of a Protein Folding Machine. *J. Mol. Biol.* **293,** 295–312 (1999).
4. Voges, D., Zwickl, P. & Baumeister, W. The 26S proteasome: A molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* **68,** 1015–1068 (1999).
5. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
6. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24,** 4420–4449 (1996).
7. Yasuda, M., Yamagishi, A. & Oshima, T. The plasmids found in isolates of the acidothermophilic archaebacterium *Thermoplasma acidophilum*. *FEMS Microbiol. Lett.* **128,** 157–161 (1995).
8. Lopez, P., Philippe, H., Myllykallio, H. & Forterre, P. Identification of putative chromosomal origins of replication in Archaea. *Mol. Microbiol.* **32,** 881–891 (1999).
9. Ree, H. K. R. & Zimmermann, R. A. Organization and expression of the 16S, 23S and 5S ribosomal RNA genes from the archaebacterium *Thermoplasma acidophilum*. *Nucleic Acids Res.* **18,** 4471–4478 (1990).
10. Budgen, N. & Danson, M. J. Metabolism of glucose via a modified Entner–Doudoroff pathway in the thermoacidophilic archaebacterium *Thermoplasma acidophilum*. *FEBS Lett.* **196,** 207–210 (1986).
11. Searcy, D. G. & Whatley, F. R. *Thermoplasma acidophilum:* Glucose degradative pathways and respiratory activities. *Syst. Appl. Microbiol.* **5,** 30–40 (1984).
12. Luebben, M. Cytochromes of archaeal electron transfer chains. *Biochim. Biophys. Acta* **1229,** 1–22 (1995).
13. Klenk, H. P. *et al.* The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390,** 364–370 (1997).
14. Huang, C. J. & Barret, E. L. Sequence Analysis and Expression of the *Salmonella typhimurium asr* Operon Encoding Production of Hydrogen Sulfide from Sulfite. *J. Bacteriol.* **173,** 1544–1553 (1991).
15. Rudolph, J. & Oesterhelt, D. Chemotaxis and phototaxis require a CheA histidine kinase in the archaeon *Halobacterium salinarium*. *EMBO J.* **14,** 667–673 (1995).
16. Coles, M. *et al.* The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple beta alpha beta beta element. *Curr. Biol.* **9,** 1158–1168 (1999).
17. Golbik, R., Lupas, A. N., Koretke, K. K., Baumeister, W. & Peters, J. The janus face of the archaeal Cdc48/p97 homologue VAT: Protein folding versus unfolding. *Biol. Chem.* **380,** 1049–1062 (1999).
18. Lupas, A., Flanagan, J. M., Tamura, T. & Baumeister, W. Self-compartmentalizing proteases. *Trends Biochem. Sci.* **22,** 399–404 (1997).
19. Zwickl, P., Ng, D., Woo, K. M., Klenk, H. P. & Goldberg, A. L. An archaebacterial ATPase, homologous to ATPases in the eukaryotic 26 S proteasome, activates protein breakdown by 20 S proteasomes. *J. Biol. Chem.* **274,** 26008–26014 (1999).
20. Tamura, N., Lottspeich, F., Baumeister, W. & Tamura, T. The role of Tricorn protease and its aminopeptidase-interacting factors in cellular protein degradation. *Cell* **95,** 637–648 (1998).
21. Wolf, S., Lottspeich, F. & Baumeister, W. Ubiquitin found in the archaebacterium *Thermoplasma acidophilum*. *FEBS Lett.* **326,** 42–44 (1993).

22. Barrett, A. J., Rawlings, N. D. & Woessner, J. F. *Handbook of proteolytic enzymes* (Academic, San Diego, CA, 1999).
23. Bergerat, A. *et al.* An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature* **386,** 414–417 (1997).
24. Stein, D. B. & Searcy, D. G. Physiologically important stabilization of DNA by a prokaryotic histone-like protein. *Science* **202,** 219–221 (1978).
25. Hixon, W. G. & Searcy, D. G. Cytoskeleton in the archaebacterium *Thermoplasma acidophilum*? Viscosity increase in soluble extracts. *BioSystems* **29,** 151–160 (1993).
26. Smith, P. F., Langworthy, T. A. & Smith, M. R. Polypeptide nature of growth requirement in yeast extract for *Thermoplasma acidophilum*. *J. Bacteriol.* **124,** 884–892 (1975).
27. Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26,** 2941–2947 (1998).
28. Frishman, D. & Mewes, H. W. PEDANTic genome analysis. *Trends Genet.* **13,** 415–416 (1997).
29. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
30. Koretke, K. K., Russell, R. B., Copley, R. R. & Lupas, A. N. Fold recognition using sequence and secondary structure information. *Proteins Struct. Funct. Genet.* **37,** 141–148 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank P. Forterre and P. Lopez for helping to define the origin of replication; M. Boicu and C. Czoppelt for sequencing; G. Mannhaupt for annotating part of the ORFs; I. Echabe for preparing template DNA and sequencing; and B. Marshall for developing software for gene cluster analysis and for data management.

Correspondence and requests for materials should be addressed to W.B. (e-mail: baumeist@biochem.mpg.de). The *Thermoplasma acidophilum* genome sequence has been deposited in the EMBL database (accession number AL139299).

An SNP map of the human genome generated by reduced representation shotgun sequencing

David Altshuler*†, Victor J. Pollara*, Chris R. Cowles*, William J. Van Etten*, Jennifer Baldwin*, Lauren Linton* & Eric S. Lander*‡

* *Whitehead Institute/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
† *Diabetes Unit, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA*
‡ *Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

Most genomic variation is attributable to single nucleotide polymorphisms (SNPs), which therefore offer the highest resolution for tracking disease genes and population history[1–3]. It has been proposed that a dense map of 30,000–500,000 SNPs can be used to scan the human genome for haplotypes associated with common diseases[4–6]. Here we describe a simple but powerful method, called reduced representation shotgun (RRS) sequencing, for creating SNP maps. RRS re-samples specific subsets of the genome from several individuals, and compares the resulting sequences using a highly accurate SNP detection algorithm. The method can be extended by alignment to available genome sequence, increasing the yield of SNPs and providing map positions. These methods are being used by The SNP Consortium, an international collaboration of academic centres, pharmaceutical companies and a private foundation, to discover and release at least 300,000 human SNPs. We have discovered 47,172 human SNPs by RRS, and in total the Consortium has identified 148,459 SNPs. More broadly, RRS facilitates the rapid, inexpensive construction of SNP maps in biomedically and agriculturally important species. SNPs discovered by RRS also offer unique advantages for large-scale genotyping.

To discover SNPs, several copies of each locus must be sampled from a population and compared for sequence differences. Two methods have been described. The first, locus-specific polymerase chain reaction (PCR) amplification (LSA), requires the synthesis of oligonucleotide primers for each locus, limiting it to regions of known sequence and making it expensive for large-scale approaches. Moreover, LSA produces diploid genotypes; this requires identification of SNPs as heterozygotes, which is technically challenging. The second, whole-genome shotgun[7], sequences random clones from the genomes of many individuals. It does not require previous knowledge of genomic sequence nor PCR, and provides haploid genotypes. Whole-genome shotgun is inflexible, however, requiring several-fold coverage of the genome before SNPs are discovered. For greater flexibility and efficiency, we sought to use 'reduced representations'—reproducibly prepared subsets of the genome, each containing a manageable number of loci to facilitate re-sampling. This approach presented three key challenges: creating reduced representations, finding orthologous matches and obtaining sufficient accuracy for automated calling of SNPs.

Many properties could be used to prepare reduced representations (for example, binding to a particular protein or functioning as an origin of replication). One of the simplest is to purify restriction fragments in a given size range. For example, *Bgl*II sites occur on average every ~3,100 base pairs (bp) in human DNA, which means that restriction fragments with length between 500 and 600 bp should number ~26,000 and comprise ~ 0.5% of the human genome. Computer analysis of 517 megabases (Mb) of finished human genomic sequence (17% of the estimated 3.1 gigabases) yielded 3,847 *Bgl*II fragments in this range—within 10% of the expected value. Thus, SNPs could be discovered by mixing DNA from many individuals, preparing a library of appropriately sized restriction fragments, and randomly sequencing clones. In this example, 52,000 sequences would provide, on average, twofold coverage of each locus, yielding thousands of SNPs.

We developed rules to align only those sequences representing the same genomic locus, excluding spurious matches arising from repeats. These rules eliminated known repeats, partial alignments that failed to extend across the entire sequence, matches showing excessive sequence divergence (compared with orthologous loci), and sequences re-sampled more often than expected for a single-copy locus (see Methods). Re-sequencing experiments confirmed that these rules successfully eliminated most spurious matches owing to paralogous repeats.

In comparing single-pass sequences for SNPs, base-calling errors can dominate the low rate of true polymorphism. With true SNPs occurring at a rate of 1 in 1,300 bp, base-calling errors must be less than 1 in 52,000 to achieve less than 5% false positive SNPs. Computer programs that estimate sequence accuracy (such as PHRED[8,9]) judge only a small fraction of single-pass bases to be this accurate. However, we noted that many base-calling 'errors' occur adjacent to easily detected artefacts (that is, compressions and stops), or are attributable to poor alignment. We hypothesized that bases surrounded by perfectly aligned, consistently high-quality sequence (termed 'good neighbourhoods') might be more accurate than predicted by PHRED. We defined a neighbourhood quality standard (NQS) to identify such bases (see Methods).

To test the NQS, we examined 3.8 Mb of single-pass human DNA sequence obtained from bacterial artificial chromosomes (BACs), and compared the base calls to the highly accurate finished sequence of each BAC. Because BAC DNA is clonal, there are no polymorphisms, and any 'SNPs' represent base-calling errors. As expected, PHRED quality scores accurately reflect the overall likelihood of error (see below). Critically, base calls within 'good neighbourhoods' were much more accurate than predicted by PHRED: 85% of bases with PHRED scores more than 20 satisfied the NQS, and displayed an error rate of 1 in 36,000. In contrast, 15% of such bases failed the NQS, and these had a 40-fold higher error rate of 1 in