

Comparative genomics of archaea: how much have we learned in six years, and what's next?

Kira S Makarova and Eugene V Koonin

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: Eugene V Koonin. E-mail: koonin@ncbi.nlm.nih.gov

Published: 16 July 2003

Genome Biology 2003, **4**:115

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/8/115>

© 2003 BioMed Central Ltd

Abstract

Archaea comprise one of the three distinct domains of life (with bacteria and eukaryotes). With 16 complete archaeal genomes sequenced to date, comparative genomics has revealed a conserved core of 313 genes that are represented in all sequenced archaeal genomes, plus a variable 'shell' that is prone to lineage-specific gene loss and horizontal gene exchange. The majority of archaeal genes have not been experimentally characterized, but novel functional pathways have been predicted.

"A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (i) the eubacteria, comprising all typical bacteria; (ii) the archaeobacteria, containing methanogenic bacteria; and (iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells."

CR Woese and GE Fox, 1977 [1]

Archaea before and after genomes

The quotation above neatly summarizes what is arguably one of the most important scientific discoveries of the twentieth century (rather remarkably, this quote is the entire abstract of Woese and Fox's groundbreaking article [1]). So profound are its implications that the debate rages to this day: did Carl Woese and George Fox really discover a new domain of life, which is equal in status to bacteria and eukaryotes [2,3], or is it 'merely' an unusual branch of bacteria [4-7]? Additional discussion of this debate is available with the complete version of this article, online.

In the years following Woese and Fox's breakthrough [1], many unique features of archaea have become apparent. To begin with, many of these organisms thrive under conditions

that, by the usual standards of biology, seem unimaginable, such as in the water in the vicinity of the hydrothermal vents called 'black smokers' heated to over-boiling temperatures and saturated with hydrogen sulfide, or in extreme salinity [11-13]. In the most extreme hyperthermophilic habitats, archaea are, in fact, the only detectable life forms. In more moderate environments, archaea coexist with bacteria and eukaryotes, and their ecological importance is being increasingly recognized [14]. The first molecular biological studies showed that archaea are highly unusual and clearly distinct from bacteria at the molecular level. In particular, the structure of the membrane glycerolipids in archaea is different from that of bacterial and eukaryal cells, and archaea do not contain murein, the predominant component of bacterial cell walls [15,16].

But the most striking differences between archaea and bacteria are seen in the organization of their information-processing systems. The structures of ribosomes and chromatin, the presence of histones, and sequence similarity between proteins involved in translation, transcription, replication and DNA repair all point to a closer relationship between archaea and eukaryotes than between either of these and bacteria [17-21]. Moreover, the key components of the DNA replication machinery - such as the polymerases involved in elongation and initiation and the replicative helicases - are

not homologous, or at least not orthologous, in archaea and eukaryotes on the one hand, and bacteria on the other [17,22]. This observation led to the hypothesis that replication of double-stranded DNA as the principal form of replication of the genetic material was 'invented' twice, independently: once in bacteria and once in the ancestor of archaea and eukaryotes [22,23]. In contrast many - although not all - of the metabolic pathways of archaea more closely resemble their bacterial rather than eukaryotic counterparts [24-26]. These studies support the status of archaea as a distinct domain of life with specific connections to eukaryotes, and emphasize the unusual and unique nature of archaeal genomes.

The new age of archaea began in 1996 with the whole-genome shotgun sequencing of the first archaeal genome, that of *Methanococcus* (now *Methanocaldococcus*) *jannaschii* [27]. The *Methanococcus* 'genomescape' at first looked largely mysterious, with clear functional assignments produced for only 38% of the genes [27]. A more detailed computational analysis that pushed the methodology available at the time to its limits yielded general functional predictions for up to 70% of the genes, showing that a solid connection between the genomes of archaea and those of other, better known forms of life did exist [24]. Nevertheless, the fact remained that, more than anything, the first sequenced archaeal genome revealed the depth of our ignorance of the biology of this remarkable group of organisms. Subsequent genome sequencing, while certainly less extensive than the devoted 'archaeologists' would wish, produced a rich sampling of genomes of taxonomically diverse archaea (Table 1). This set of completely sequenced genomes includes multiple representatives of the two major divisions of the archaea established by phylogenetic analysis of rRNA, namely the Euryarchaeota and the Crenarchaeota [3], as well as the principal ecological types of archaea, such as hyperthermophiles, moderate thermophiles, and mesophiles, as well as halophiles and methanogens; autotrophic and heterotrophic forms, and anaerobes and aerobes are also represented by multiple species (Table 1).

Some potentially important branches of archaea are still missing from sequence databases, however, such as the mysterious Korarchaeota, which might have branched off the trunk of the phylogenetic tree prior to the divergence of the remainder of the archaea [28], and the equally intriguing Nanoarchaea that so far seem to have the smallest genomes of all known cellular life forms [29,30]. These lacunae notwithstanding, the available sampling of archaeal genomes is substantial and is complemented by an even greater diversity of bacterial and eukaryotic genomes that are available for comparative analysis. This article critically assesses the contribution of comparative genomics to our understanding of the functional systems of archaeal cells and their evolution (more details of the evolutionary implications are given in the complete version of this article, online). We pose the following question: what have we learned from comparisons

of archaeal genomes that could not easily have been learned by other, more traditional approaches? We suggest some tentative answers, as we see them. What follows is a viewpoint from behind a computer terminal; we realize that, from the experimenter's bench, the perspective might be somewhat different.

From genome comparisons to functional and structural genomics of the archaea

In the era of comparative genomics, experimental studies on a genomic scale lag woefully behind computational studies. The great majority of the genes in most species will never be studied experimentally, and our understanding of the biochemistry and physiology of the respective organisms therefore depends on the transfer of information from functionally characterized orthologs [26,72]. For both bacteria and eukaryotes, such transfer is facilitated by the availability of a vast body of experimental data on model organisms, such as *Escherichia coli*, *Bacillus subtilis*, the yeast *Saccharomyces cerevisiae* or the fruit fly *Drosophila melanogaster*. The situation is quite different for archaea because, some genetic studies of mesophilic archaeal species notwithstanding [73], there is, so far, no satisfactory model system; this results primarily from the fact that most of these organisms grow slowly and are hard to cultivate. The functions of most of the archaeal genes have therefore been predicted by sequence analysis. Moreover, on many occasions the similarity between an archaeal protein and its functionally characterized homolog is so low that computational methods for sequence analysis have to be extended to the limit of their power.

A substantial fraction of the functional predictions for archaeal proteins appear 'trivial' in the sense that the respective proteins are highly conserved orthologs of well-characterized proteins from model organisms and, for all practical purposes, the validity of the prediction is beyond reasonable doubt (which is not to say that there are no important details of the functions of these proteins that can be uncovered only by experiment). For many other proteins, however, the prediction remains only a pointer to the probable biochemical function while the biology remains a mystery. A rough breakdown of the state of functional characterization of several archaea with sequenced genomes is given in Figure 4. The substantial fraction of genes for which only general, typically biochemical, prediction is available, is testimony to the current limited understanding of archaeal biology (this information is captured in the database of Clusters of Orthologous Groups of proteins, COGs [37]). Moreover, even some of the more definitive predictions only serve to emphasize the biological differences between archaea and the bacterial or eukaryotic models from which the predictions are inferred (Table 3). A good example is the archaeal ortholog of the bacterial DNA primase (DnaG), which is a highly conserved protein present in all archaea [24]. The

Table 1**Completely sequenced archaeal genomes**

Species	Abbreviation	Optimal growth temperature (°C)	Lifestyle and other features	Number of proteins*	Number (%) proteins in COGs	Date of genome release	Reference
Euryarchaeota							
<i>Archaeoglobus fulgidus</i> DSM	Afu	83	Anaerobic, sulfate-reducing chemolitho- or chemorgano-autotroph, motile	2,420	1,953 (81%)	1997	[124]
<i>Halobacterium</i> sp. NRC-1	Hsp	37	Aerobic chemorganotroph, obligate halophile, with a cell envelope; motile; two extrachromosomal elements	2,622	1,809 (69%)	2000	[125]
<i>Methanocaldococcus jannaschii</i>	Mja	85	Chemolithoautotroph, strict anaerobe, methanogen, motile; two extrachromosomal elements	1,758	1,448 (82%)	1996	[27]
<i>Methanopyrus kandleri</i> AV19	Mka	110	Chemolithoautotroph, strict anaerobe, methanogen, with high cellular salt concentration	1,691	1,253 (74%)	2002	[45]
<i>Methanosarcina acetivorans</i> C2A	Mac	37	Chemolithoautotroph, anaerobe possibly capable of aerobic growth; nitrogen-fixing, versatile methanogen; motile, and able to form multicellular structures	4,540	3,142 (69%)	2002	[55]
<i>Methanosarcina mazei</i> Goel	Mma	37	As for Mac	3,371	N/A	2002	[54]
<i>Methanothermobacter thermoautotrophicus</i> delta H	Mth	65	Chemolithoautotroph, strict anaerobe, nitrogen-fixing, methanogen	1,873	1,500 (80%)	1997	[126]
<i>Pyrococcus horikoshii</i>	Pho	96	Anaerobic heterotroph, sulfur enhances growth; motile	1,801	1,425 (79%)	1998	[127]
<i>Pyrococcus abyssi</i>	Pab	96	As for Pho	1,769	1,506 (85%)	2001	[128]
<i>Pyrococcus furiosus</i> DSM 3638	Pfu	96	As for Pho	2,065	N/A	2001	[129]
<i>Thermoplasma acidophilum</i>	Tac	59	Facultative anaerobe, chemorganotroph, thermoacidophilic, anaerobically able to metabolize sulfur; motile, with a plasma membrane	1,482	1,261 (85%)	2000	[96]
<i>Thermoplasma volcanium</i>	Tvo	60	As for Tac	1,499	1,277 (85%)	2000	[130]
Crenarchaeota							
<i>Pyrobaculum aerophilum</i>	Pae	100	Facultative nitrate-reducing anaerobe	1,840	1,236 (67%)	2002	[131]
<i>Aeropyrum pernix</i>	Ape	90	Aerobic chemorganotroph; sulfur enhances growth	2,605	1,529 (59%)	1999	[132]
<i>Sulfolobus solfataricus</i>	Sso	80	Aerobe metabolizing sulfur; thermoacidophilic chemorganotroph; motile	2,977	2,207 (74%)	2001	[97]
<i>Sulfolobus tokodaii</i>	Sto	80	As for Sso	2,826	N/A	2001	[133]

*According to the original genome annotation.

discovery of a predicted bacterial-type primase in archaea was unexpected, given that the archaeal replication system is orthologous to that of eukaryotes and, in particular, archaea encode the two subunits of the eukaryotic-type primase (COG1467 and COG2219; it should be noted parenthetically

that detection of the large primase subunit itself required extremely careful sequence analysis due to the low similarity to the eukaryotic ortholog [22]). Given that the niche of the replicative primase seems to be occupied by the eukaryotic-type enzyme [74,75], the DnaG ortholog is likely to

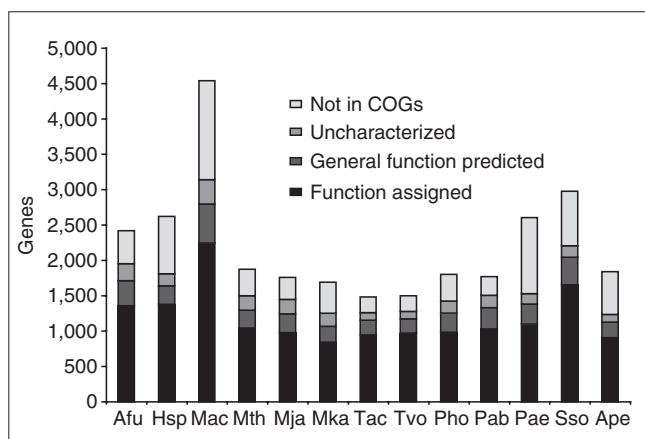


Figure 4
Functional breakdown of genes in each of the sequenced archaeal genomes. The data are from COGs; species name abbreviations are as in Table 1.

have a critical role in repair, but beyond this general idea its function has yet to be determined by direct experimentation; such experiments have the potential to reveal completely new repair systems and pathways. Other proteins implicated in repair as a result of exhaustive sequence analysis, such as the putative nucleases encoded by COG1833 and COG1628 (Table 3), illustrate the same point: the biochemical activities are predicted but the biology remains to be investigated experimentally.

Some of the other functional predictions inferred from sequence analysis directly help filling glaring gaps in otherwise well-characterized pathways of archaeal metabolism. A good example of such focused prediction is the identification of an archaeal fructose-1,6-bisphosphate aldolase, an indispensable glycolytic enzyme, which was first predicted computationally to be a member of the Dhna family of aldolases by our group [76] and subsequently identified experimentally [77]. In the same vein, during work for this article, we predicted the missing archaeal aconitase, an essential enzyme of the tricarboxylic acid cycle (Table 3; K.S.M. and E.V.K., unpublished observations).

The identities of a considerable number of proteins responsible for essential functions in archaea remain a mystery. Perhaps the most notable case is the missing cysteinyl-tRNA synthetase of thermophilic methanogens. Cysteine is incorporated into the proteins of these organisms as readily as in any others, but they lack an ortholog of cysteinyl-tRNA synthetase. Two different solutions for this paradox have been proposed, one involving an uncharacterized protein that has been proposed to be a 'third class' of aminoacyl-tRNA synthetases [78], and the other based on the apparent ability of the archaeal prolyl-tRNA synthetase to couple tRNA^{Cys} with cysteine [79]. The first hypothesis has been

refuted by our group upon more detailed sequence analysis [80], however, and the second did not seem to be compatible with subsequent structural studies [81]. The real cysteinyl-tRNA synthetase of methanogens seems still to be hiding among uncharacterized proteins. Gaping holes also remain in archaeal pathways of isoleucine biosynthesis [82], heme biosynthesis [83], biotin biosynthesis [26], and several others.

Beyond straightforward (even if highly sensitive) sequence analysis, a powerful approach to the prediction of functions involves analysis of various forms of genomic context, or establishing 'guilt by association' [26,84-87]. The associations employed to infer gene functions may be manifest at different levels, including the phyletic patterns discussed in the complete version of this article, online, juxtaposition of domains in multidomain proteins, clustering of genes in (predicted) operons, co-expression, and protein-protein interaction. The last two of these types of data, obtained through transcriptomic and proteomic efforts, are becoming increasingly important in the functional genomics of eukaryotes and, to a somewhat lesser extent, bacteria, but are so far unavailable for archaea. The main type of context information in archaea has therefore been obtained by analyzing conserved elements of gene order and multidomain proteins. Only a relatively small fraction (10-15%) of each archaeal genome is covered by evolutionarily conserved gene strings that can be predicted to form operons [87]. Nevertheless, by comparing gene orders in multiple genomes, partially conserved gene neighborhoods can be reconstructed and examination of some of these leads to predictions of functional systems whose existence has not previously been suspected (Table 3).

The most notable illustrations of this approach (both from our own group) are the prediction of the archaeal exosome [88] and a potential new repair system typical of archaeal and bacterial thermophiles [59]. The eukaryotic exosome is a multisubunit complex that consists of RNases and RNA-binding proteins and is involved in the exonucleolytic degradation of various classes of RNA [89-91]. During comparative analysis of gene order in prokaryotic genomes, it was observed that a distinct set of genes, some of which encode orthologs of eukaryotic exosome components, form a partially conserved predicted superoperon, which includes in total over 15 genes (although none of the archaeal genomes contains every one of these within the predicted superoperon). In addition to RNases and RNA-binding proteins (with an RNA helicase apparently encoded in a separate operon), the exosomal superoperon also encodes a proteasome subunit and a subunit of prefoldin, a co-translational molecular chaperone ([88] and Figure 5a). Thus, these observations point to the existence of a multifunctional macromolecular complex that could couple post-translational protein folding with regulated, ATP-dependent degradation of RNA and proteins. This complex remains to be discovered

Table 3**Examples of computational and experimental discovery of unexpected functions in archaea**

COG numbers [37,38]	Function and comments	References
Computational predictions		
0012, 1325, 1603, 1369, 0638, 1500, 1097, 689, 2123, 1996, 2136, 2892, 0618, 1782, 1096, 3286, 1761 and more	Archaeal exosome. Orthologs of eukaryotic exosome subunits form the largest conserved superoperon in archaea, after the ribosomal superoperon, suggesting the existence of a physical complex	[88]
1769, 1336, 3337, 1583, 1367, 1604, 1517, 1857, 1688, 1203, 1468, 1518, 2254, 1343, 1353, 1421, 1337, 1567, 1332, 4343	DNA repair system represented primarily in thermophiles	[59]
0358	Bacterial-type DNA primase (DnaG orthologs)	[24]
1311	Small subunit of euryarchaeal DNA polymerase II, predicted PHP family phosphohydrolase (probably phosphatase); eukaryotic homologs appear to be inactivated	[123]
1833	Uri superfamily endonuclease	[136]
1628	Endonuclease V homologs	K.S.M. and E.V.K., unpublished observations
1679,1786	Aconitase catalytic core and an interacting 'swiveling domain'	K.S.M. and E.V.K., unpublished observations
1711	Possible subunit of the DNA replication machinery	K.S.M. and E.V.K., unpublished observations
1310	Zn ²⁺ -dependent hydrolase homologous to the eukaryotic ubiquitin isopeptidase contained in the proteasome and COP9 signalosome	[137,138]
Computational predictions validated by experiments		
1708	'Minimal' nucleotidyltransferases	[100,139]
1830	Fructose-1,6-bisphosphate aldolases (DhnA family)	[76,77]
1351	Thymidylate synthase	[61,64]
1685	Shikimate kinase (predicted on the basis of operon organization)	[140]
3635	Phosphoglycerate mutase	[24,141]
Experimental discovery of unexpected protein functions in archaea		
1384	Class I lysyl-tRNA synthetase	[62]
1933	DNA polymerase II	[104]
1980	Fructose 1,6-bisphosphatase	[142]
1630	NurA, a novel 5'-3' nuclease encoded next to Rad50 and MreI I orthologs; present in all sequenced archaeal genomes and some bacteria	[143] and K.S.M. and E.V.K., unpublished observations
1812	S-adenosylmethionine synthetase, was identified by mass tags	[144]
1591	Holliday junction resolvase	[101]
1581	Alba, a major DNA-binding chromatin protein in Crenarchaeota	[106]
1945	Pyruvoyl-dependent arginine decarboxylase (PvlArgDC), involved in polyamine biosynthesis	[145]

experimentally, and the potential implications for new functional and physical interactions in eukaryotes are also open to experimental study.

A more sophisticated comparison of gene orders, which required special algorithms for delineation of partially

conserved genomic neighborhoods [92], led us to predict a distinct DNA repair system that is most prevalent in thermophiles and includes genes for a predicted novel DNA polymerase, a helicase, two nucleases and several uncharacterized genes, at least one of which could encode a novel nuclease ([59] and Figure 5b). Furthermore, this neighborhood

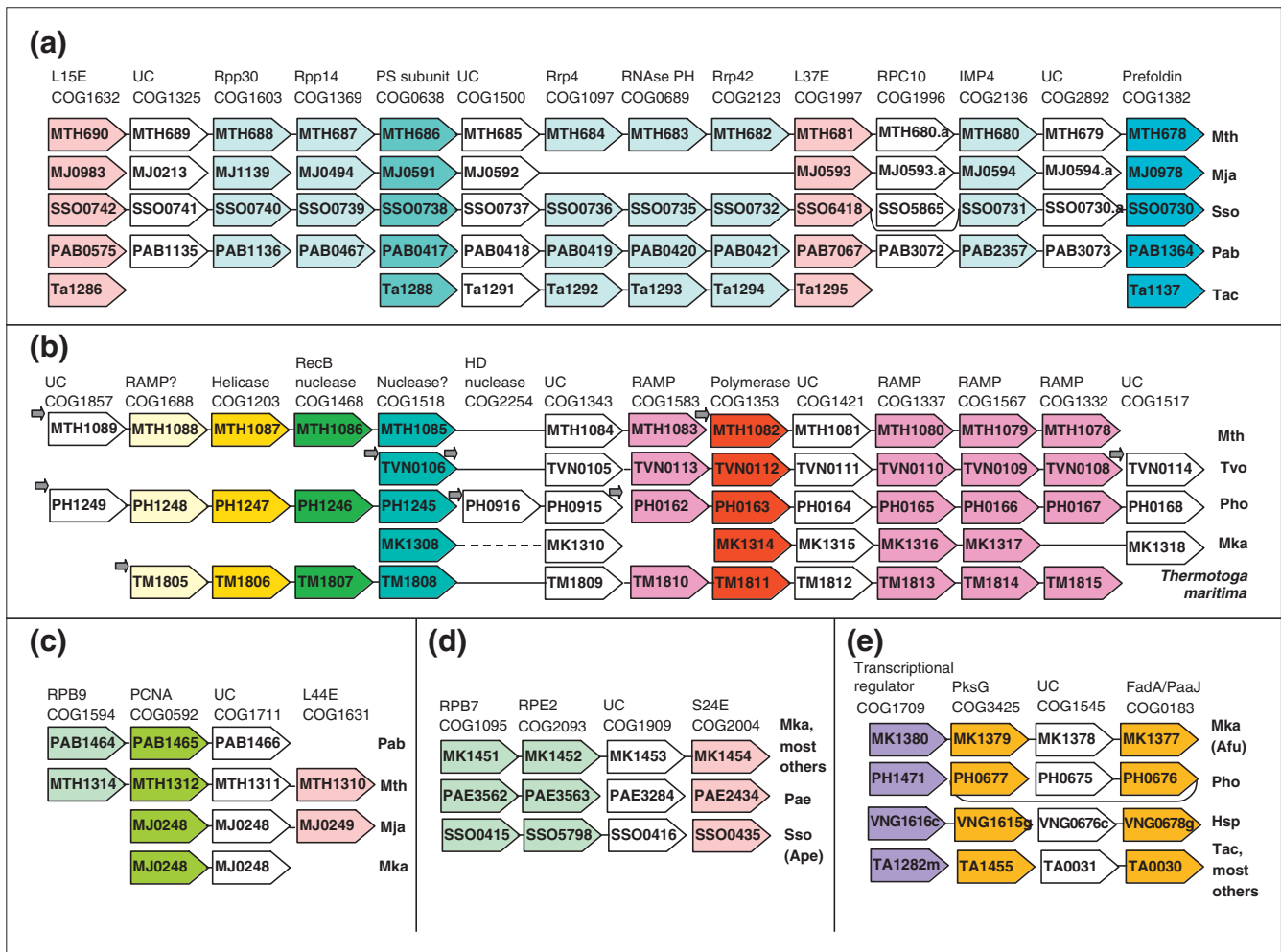


Figure 5
 Prediction of gene functions in archaea by genomic context analysis. **(a)** The superoperon coding for the predicted archaeal exosome (see [88]). **(b)** The partially conserved gene neighborhood coding for the predicted repair system found in archaeal and bacterial thermophiles (see [59] for details). **(c-e)** Predicted operons containing uncharacterized genes in the neighborhood of genes from the following COGs: COG1594, DNA-directed RNA polymerase, subunit M, and transcription elongation factor TFIIIS (RPB9); COG0592, encoding a DNA polymerase sliding clamp subunit (PCNA ortholog); COG1631, ribosomal protein L44E; COG1095, DNA-directed RNA polymerase, subunit E' (RPB7); COG2093, DNA-directed RNA polymerase, subunit E'' (RPE2); COG2004, ribosomal protein S24E; COG1709, transcriptional regulator; COG3425, 3-hydroxy-3-methylglutaryl CoA synthase (PksG); COG0183, acetyl-CoA acetyltransferase (Fad A/PaaJ orthologs). UC, uncharacterized, shown by white arrows. Species abbreviations are as in Table 1. Genes are shown not to scale and are denoted by their respective genes names (some are discussed further in the text); arrows indicate the direction of transcription. A solid line connects genes in a predicted operon. Species that have the same operon organization as the listed species are indicated in parentheses. Orthologous genes are aligned. Genes with similar general functions are shown by the same shading. Broken lines show that genes are in the same predicted operon but are not adjacent. Small arrows indicate the presence of additional functionally related genes in the same predicted operon; these genes are not shown for lack of space.

contains multiple, diverged versions of a gene coding for a protein with a probable structural role dubbed RAMP (repair-associated mysterious protein). The proliferation of RAMP genes (Figure 5b) is an example of a potentially adaptive lineage-specific expansion of a gene family; such expansions are discussed below in greater detail.

Additional, simpler cases of functional prediction via 'guilt by association' are illustrated in Figure 5c-e. The gene for the uncharacterized protein represented by COG1711 (Figure 5c)

forms an evolutionarily highly conserved gene pair with the gene for the clamp subunit of DNA polymerase (ortholog of the eukaryotic PCNA). The orthologs of COG1711 proteins are conserved in all eukaryotes, and this protein might be an essential but still uncharacterized component of the archaeo-eukaryotic DNA replication machinery (K.S.M. and E.V.K., unpublished observations). The gene represented by uncharacterized COG1909 is squeezed between genes for RNA polymerase subunits and that for a ribosomal protein (Figure 5d). Examination of the multiple alignments that

lead to this COG shows conservation of polar residues compatible with an enzymatic function (K.S.M. and E.V.K., unpublished observations). There are no readily detectable eukaryotic orthologs for this protein, which is therefore likely to be an archaea-specific enzyme with a house-keeping function.

Finally, uncharacterized COG1545 consists of genes encoding putative zinc-ribbon-containing proteins that form a stable gene pair with the gene for acetyl-CoA acetyltransferase, a central enzyme of fatty acid biosynthesis (Figure 5e). Both these genes show remarkable paralogous expansion in several archaea, probably as a result of a series of duplications of the gene doublet. Further discussion of lineage-specific expansion of paralogous genes can be found with the complete version of this article, online. It appears likely that proteins from COG1545 form a complex with acetyl-CoA acetyltransferase, with the zinc-ribbon protein regulating and/or stabilizing the enzyme. The predictions depicted in Figure 5c-e and other similar ones ([87]; and K.S.M. and E.V.K., unpublished observations) are not particularly precise, even in terms of the biochemical activity of the respective proteins. Nevertheless, guilt by association implicates each of these proteins in specific biological functions, and the evolutionary conservation of both the proteins themselves and the gene order all but proves that their functions are essential. Thus, these proteins appear to be excellent targets for experimental studies, which have the potential to reveal new facets of central cellular processes in archaea.

Archaeal comparative genomics is a young field and so far, as we have seen, largely predictive. But a few experimental studies have already been instigated as a result of comparative-genomic predictions. The discovery of the archaeal fructose-1,6-bisphosphate aldolase mentioned above [76,77] is a case in point, and several other examples of experimental validation of predictions are given in Table 3. It does not seem to be chance that these examples all involve metabolic enzymes for which the specific reaction could be predicted precisely. Validation is likely to be much more difficult for proteins of other functional groups, such as putative repair enzymes, for which the actual substrates are harder to predict.

For some conserved archaeal proteins, functions cannot be predicted computationally despite considerable effort. Several important discoveries have been made by experimental characterization of such mysterious proteins. The most notable cases include the archaeal Holliday-junction resolvase, which is not related to its functional analog in bacteria [101-103], and DNA polymerase II, a highly conserved euryarchaeal protein that is not found outside this lineage and shows no detectable sequence similarity to any other proteins [104,105]. Additional examples of direct experimental determination of the functions of archaeal proteins that could not be predicted by computational techniques (at

least not before the experiment had been reported) are given in Table 3.

Especially notable is the story of the Alba protein, a DNA-binding component of chromatin in Crenarchaeota [106,107]. As noted above, crenarchaea lack histones and in these organisms Alba appears to be the main chromatin protein, in a striking case of non-orthologous gene displacement. But orthologs of Alba are also present in thermophilic Euryarchaeota and in some eukaryotic lineages, where its functions remain to be elucidated. The most remarkable discovery regarding Alba is the regulation of its interaction with DNA and with the chromatin-associated protein deacetylase Sir2 via lysine acetylation and deacetylation [106,108]. In eukaryotes, regulation of chromatin dynamics via acetylation and deacetylation occurs through histone tails [109]. Thus, a special case of non-orthologous gene displacement seems to have taken place whereby the regulation mechanism is conserved but the actual substrates are different in archaea and eukaryotes. To add an extra twist to the story, *Thermoplasma* lacks both histones and Alba but has the bacterial DNA-binding protein HU, pointing to three distinct solutions to the problem of chromatin organization in archaea [107].

What's around the corner?

The first sequenced archaeal genome was a veritable *terra incognita*. Six years after that sequence appeared, the archaeal genomescape looks quite different. The principal landmarks have been mapped and now, when a new archaeal genome is released, we largely know what to expect from it. Computational approaches to comparative genomics, combining in-depth sequence and structure comparison with genome context analysis, have led to the reconstruction of the central functional systems of archaeal cells. But these approaches have also produced numerous isolated predictions of biochemical activities of archaeal proteins that remain to be fitted into a general picture, and this can be done only through 'wet' experiments, although new genome sequences will substantially help by enriching the genomic context. A shrinking but still notable set of archaeal genes includes those that encode highly conserved proteins without any clue to function; solving these mysteries has the potential to bring out truly new biology. Furthermore, in this article we have not even touched upon important aspects of archaeal genomics, such as the in-depth studies of the translation system, which have revealed several highly unusual, remarkable mechanisms and enzymatic systems [63,119] or the identification of regulatory sites in DNA and patterns of transcription regulation [120,121]. The latter avenue of research is still in its infancy but will certainly grow in scale once more archaeal genomes, and in particular closely related ones, are sequenced. The complete version of this article, online, includes discussion of the structural genomics initiatives that are shedding some light on archaeal protein functions.

Because of the lack of established model systems for archaeal experimental biology and the resulting difficulty with large-scale experimentation, clues from genome comparison are even more crucial for archaeal functional genomics than they are in the case of bacteria or eukaryotes. So far, the input of comparative genomics into actual experiments has been less prominent than we would hope. Simply put, it is not often that experimenters rush to test predictions produced by *in silico* genome comparison and, furthermore, it is even rarer that targets for functional characterization are carefully prioritized on the basis of how unusual and fundamental the predictions are. As discussed above, however, the few cases when such tests have been performed are encouraging. It is our hope that the future belongs to a much tighter integration of comparative, structural and functional genomics.

Beyond functional studies, archaeal genomics is fundamental to our understanding of two critical transitions in the evolution of life. The first is the primary split between the bacterial and archaeo-eukaryotic lineages, which might have involved the origin of the DNA-replication machinery and of the large, double-stranded DNA genomes themselves [22,23], and the second is the origin of eukaryotes [122]. With regard to the latter problem, archaea are a particularly valuable source of information because, on many occasions, they seem to have retained primitive traits while eukaryotes have undergone major changes. A characteristic example is the small DNA polymerase subunit, which has all the hallmarks of an active phosphatase in archaea, but not in eukaryotes, in which the phosphatase activity is predicted to be inactivated [123]. Indubitably, archaea resemble the common ancestor of the archaeo-eukaryotic line of descent more closely than eukaryotes do, so archaeal genomics is our best chance to reconstruct this critical intermediate in the evolution of life. We are confident that comparative archaeogenomics has a bright future, with major progress in both the functional and the evolutionary avenues of research expected within the next few years.

Additional data file

The list of genes in the reconstructed gene set of the last common ancestor of archaea is available with the complete version of this article, online.

Acknowledgements

We thank Boris Mirkin for producing the data used for Figure 3 (available with the complete version of this article, online) and Stephen Bell, Michael Galperin, Dieter Söll and Yuri Wolf for useful discussions.

References

1. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74**:5088-5090.
2. Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, et al.: **The phylogeny of prokaryotes.** *Science* 1980, **209**:457-463.
3. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci USA* 1990, **87**:4576-4579.
4. Woese CR, Gupta R: **Are archaeobacteria merely derived 'prokaryotes'?** *Nature* 1981, **289**:95-96.
5. Mayr E: **Two empires or three?** *Proc Natl Acad Sci USA* 1998, **95**:9720-9723.
6. Woese CR: **Default taxonomy: Ernst Mayr's view of the microbial world.** *Proc Natl Acad Sci USA* 1998, **95**:11043-11046.
7. Gupta RS: **Life's third domain (Archaea): an established fact or an endangered paradigm?** *Theor Popul Biol* 1998, **54**:91-104.
11. Stetter KO: **Extremophiles and their adaptation to hot environments.** *FEBS Lett* 1999, **452**:22-25.
12. Segerer AH, Burggraf S, Fiala G, Huber G, Huber R, Pley U, Stetter KO: **Life in hot springs and hydrothermal vents.** *Orig Life Evol Biosph* 1993, **23**:77-90.
13. DeLong EF: **A phylogenetic perspective on hyperthermophilic microorganisms.** *Methods Enzymol* 2001, **330**:3-11.
14. DeLong EF, Pace NR: **Environmental diversity of bacteria and archaea.** *Syst Biol* 2001, **50**:470-478.
15. Hanford MJ, Peeples TL: **Archaeal tetraether lipids: unique structures and applications.** *Appl Biochem Biotechnol* 2002, **97**:45-62.
16. Engelhardt H, Peters J: **Structural research on surface layers: a focus on stability, surface layer homology domains, and surface layer-cell wall interactions.** *J Struct Biol* 1998, **124**:276-302.
17. Edgell DR, Doolittle WF: **Archaea and the origin(s) of DNA replication proteins.** *Cell* 1997, **89**:995-998.
18. Sandman K, Pereira SL, Reeve JN: **Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome.** *Cell Mol Life Sci* 1998, **54**:1350-1364.
19. Sandman K, Bailey KA, Pereira SL, Soares D, Li WT, Reeve JN: **Archaeal histones and nucleosomes.** *Methods Enzymol* 2001, **334**:116-129.
20. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O: **Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale.** *Nucleic Acids Res* 2002, **30**:5382-5390.
21. Forterre P, Brochier C, Philippe H: **Evolution of the Archaea.** *Theor Popul Biol* 2002, **61**:409-422.
22. Leipe DD, Aravind L, Koonin EV: **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27**:3389-3401.
23. Forterre P: **The origin of DNA genomes and DNA replication proteins.** *Curr Opin Microbiol* 2002, **5**:525-532.
24. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25**:619-637.
25. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96**:3801-3806.
26. Koonin EV, Galperin MY: *Sequence - Evolution - Function. Computational Approaches in Comparative Genomics.* New York: Kluwer Academic; 2002.
27. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, et al.: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*.** *Science* 1996, **273**:1058-1073.
28. Pace NR: **A molecular view of microbial diversity and the biosphere.** *Science* 1997, **276**:734-740.
29. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO: **A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont.** *Nature* 2002, **417**:63-67.
30. Huber H, Hohn MJ, Stetter KO, Rachel R: **The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life.** *Res Microbiol* 2003, **154**:165-171.
37. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22-28.
45. Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, et

- al.: **The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens.** *Proc Natl Acad Sci USA* 2002, **99**:4644-4649.
54. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Baumer S, Jacobi C, et al.: **The genome of *Methanobacillus mazeri*: evidence for lateral gene transfer between bacteria and archaea.** *J Mol Microbiol Biotechnol* 2002, **4**:453-461.
 55. Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, FitzHugh W, Calvo S, Engels R, Smirnov S, Atnoor D, et al.: **The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity.** *Genome Res* 2002, **12**:532-542.
 59. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30**:482-496.
 61. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18**:609-613.
 62. Ibba M, Morgan S, Curnow AW, Pridmore DR, Vothknecht UC, Gardner W, Lin W, Woese CR, Soll D: **A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases.** *Science* 1997, **278**:1119-1122.
 63. Praetorius-Ibba M, Ibba M: **Aminoacyl-tRNA synthesis in archaea: different but not unique.** *Mol Microbiol* 2003, **48**:631-637.
 64. Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U: **An alternative flavin-dependent mechanism for thymidylate synthesis.** *Science* 2002, **297**:105-107.
 72. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol* 2000, **297**:233-249.
 73. Luo Y, Wasserfallen A: **Gene transfer systems and their applications in Archaea.** *Syst Appl Microbiol* 2001, **24**:15-25.
 74. Liu L, Komori K, Ishino S, Bocquier AA, Cann IK, Kohda D, Ishino Y: **The archaeal DNA primase: biochemical characterization of the p41-p46 complex from *Pyrococcus furiosus*.** *J Biol Chem* 2001, **276**:45484-45490.
 75. Bocquier AA, Liu L, Cann IK, Komori K, Kohda D, Ishino Y: **Archaeal primase: bridging the gap between RNA and DNA polymerases.** *Curr Biol* 2001, **11**:452-456.
 76. Galperin MY, Aravind L, Koonin EV: **Aldolases of the Dhna family: a possible solution to the problem of pentose and hexose biosynthesis in archaea.** *FEMS Microbiol Lett* 2000, **183**:259-264.
 77. Siebers B, Brinkmann H, Dorr C, Tjaden B, Lilie H, van der Oost J, Verhees CH: **Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase.** *J Biol Chem* 2001, **276**:28710-28718.
 78. Fabrega C, Farrow MA, Mukhopadhyay B, de Crecy-Lagard V, Ortiz AR, Schimmel P: **An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes.** *Nature* 2001, **411**:110-114.
 79. Stathopoulos C, Li T, Longman R, Vothknecht UC, Becker HD, Ibba M, Soll D: **One polypeptide with two aminoacyl-tRNA synthetase activities.** *Science* 2000, **287**:479-482.
 80. Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, Koonin EV: **Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences.** *Genome Biol* 2001, **2**:research0051.1-0051.11
 81. Kamtekar S, Kennedy WD, Wang J, Stathopoulos C, Soll D, Steitz TA: **The structural basis of cysteine aminoacylation of tRNAPro by prolyl-tRNA synthetases.** *Proc Natl Acad Sci USA* 2003, **100**:1673-1678.
 82. Xie G, Forst C, Bonner C, Jensen RA: **Significance of two distinct types of tryptophan synthase beta chain in Bacteria, Archaea and higher plants.** *Genome Biol* 2002, **3**:research0004.1-0004.13
 83. Panek H, O'Brian MR: **A whole genome view of prokaryotic haem biosynthesis.** *Microbiology* 2002, **148**:2273-2282.
 84. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10**:366-370.
 85. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10**:1204-1210.
 86. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10**:1074-1077.
 87. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11**:356-372.
 88. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11**:240-252.
 89. Decker CJ: **The exosome: a versatile RNA processing machine.** *Curr Biol* 1998, **8**:R238-R240.
 90. van Hoof A, Parker R: **The exosome: a proteasome for RNA?** *Cell* 1999, **99**:347-350.
 91. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D: **The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'→5' exoribonucleases.** *Cell* 1997, **91**:457-466.
 92. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV: **Connected gene neighborhoods in prokaryotic genomes.** *Nucleic Acids Res* 2002, **30**:2212-2223.
 96. Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W: **The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*.** *Nature* 2000, **407**:508-513.
 97. She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, Chan-Weiher CC, Clausen IG, Curtis BA, De Moors A, et al.: **The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2.** *Proc Natl Acad Sci USA* 2001, **98**:7835-7840.
 100. Lehmann C, Lim K, Chalamasetty VR, Krajewski W, Melamud E, Galkin A, Howard A, Kelman Z, Reddy PT, Murzin AG, Herzberg O: **The HI0073/HI0074 protein pair from *Haemophilus influenzae* is a member of a new nucleotidyltransferase family: structure, sequence analyses, and solution studies.** *Proteins* 2003, **50**:249-260.
 101. Komori K, Sakae S, Shinagawa H, Morikawa K, Ishino Y: **A Holliday junction resolvase from *Pyrococcus furiosus*: functional similarity to *Escherichia coli* RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea.** *Proc Natl Acad Sci USA* 1999, **96**:8873-8878.
 102. Aravind L, Makarova KS, Koonin EV: **Survey and summary: Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.** *Nucleic Acids Res* 2000, **28**:3417-3432.
 103. Daiyasu H, Komori K, Sakae S, Ishino Y, Toh H: **Hjc resolvase is a distantly related member of the type II restriction endonuclease family.** *Nucleic Acids Res* 2000, **28**:4540-4543.
 104. Ishino Y, Komori K, Cann IK, Koga Y: **A novel DNA polymerase family found in Archaea.** *J Bacteriol* 1998, **180**:2232-2236.
 105. Ishino Y, Ishino S: **DNA polymerases from euryarchaeota.** *Methods Enzymol* 2001, **334**:249-260.
 106. Bell SD, Botting CH, Wardleworth BN, Jackson SP, White MF: **The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation.** *Science* 2002, **296**:148-151.
 107. White MF, Bell SD: **Holding it together: chromatin in the Archaea.** *Trends Genet* 2002, **18**:621-626.
 108. Wardleworth BN, Russell RJ, Bell SD, Taylor GL, White MF: **Structure of Alba: an archaeal chromatin protein modulated by acetylation.** *EMBO J* 2002, **21**:4654-4662.
 109. Kurdستاني SK, Grunstein M: **Histone acetylation and deacetylation in yeast.** *Nat Rev Mol Cell Biol* 2003, **4**:276-284.
 119. Woese CR: **Translation: in retrospect and prospect.** *RNA* 2001, **7**:1055-1067.
 120. Gelfand MS, Koonin EV, Mironov AA: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **28**:695-705.
 121. Rodionov DA, Mironov AA, Gelfand MS: **Conservation of the biotin regulon and the BirA regulatory signal in eubacteria and archaea.** *Genome Res* 2002, **12**:1507-1516.
 122. Dacks JB, Doolittle WF: **Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help.** *Cell* 2001, **107**:419-425.
 123. Aravind L, Koonin EV: **Phosphoesterase domains associated with DNA polymerases of diverse origins.** *Nucleic Acids Res* 1998, **26**:3746-3752.
 124. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, et al.: **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 1997, **390**:364-370.

125. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, et al.: **Genome sequence of *Halobacterium* species NRC-1.** *Proc Natl Acad Sci USA* 2000, **97**:12176-12181.
126. Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, et al.: **Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics.** *J Bacteriol* 1997, **179**:7135-7155.
127. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, et al.: **Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3.** *DNA Res* 1998, **5**:55-76.
128. Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Querellou J, Ripp R, et al.: **An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*.** *Mol Microbiol* 2003, **47**:1495-1512.
129. Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB, Dunn DM: **Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology.** *Methods Enzymol* 2001, **330**:134-157.
130. Kawashima T, Amano N, Koike H, Makino S, Higuchi S, Kawashima-Ohya Y, Watanabe K, Yamazaki M, Kanehori K, Kawamoto T, et al.: **Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*.** *Proc Natl Acad Sci USA* 2000, **97**:14257-14262.
131. Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH: **Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*.** *Proc Natl Acad Sci USA* 2002, **99**:984-989.
132. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, et al.: **Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1.** *DNA Res* 1999, **6**:83-101.
133. Kawarabayasi Y, Hino Y, Horikawa H, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, Kosugi H, Hosoyama A, et al.: **Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7.** *DNA Res* 2001, **8**:123-140.
136. Aravind L, Walker DR, Koonin EV: **Conserved domains in DNA repair proteins and evolution of repair systems.** *Nucleic Acids Res* 1999, **27**:1223-1242.
137. Cope GA, Suh GS, Aravind L, Schwarz SE, Zipursky SL, Koonin EV, Deshaies RJ: **Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cull1.** *Science* 2002, **298**:608-611.
138. Verma R, Aravind L, Oania R, McDonald WH, Yates JR, 3rd, Koonin EV, Deshaies RJ: **Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome.** *Science* 2002, **298**:611-615.
139. Aravind L, Koonin EV: **DNA polymerase beta-like nucleotidyl-transferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27**:1609-1618.
140. Daugherty M, Vonstein V, Overbeek R, Osterman A: **Archaeal shikimate kinase, a new member of the GHMP-kinase family.** *J Bacteriol* 2001, **183**:292-300.
141. van der Oost J, Huynen MA, Verhees CH: **Molecular characterization of phosphoglycerate mutase in archaea.** *FEMS Microbiol Lett* 2002, **212**:111-120.
142. Rashid N, Imanaka H, Kanai T, Fukui T, Atomi H, Imanaka T: **A novel candidate for the true fructose-1,6-bisphosphatase in archaea.** *J Biol Chem* 2002, **277**:30649-30655.
143. Constantinesco F, Forterre P, Elie C: **NurA, a novel 5'-3' nuclease gene linked to rad50 and mre11 homologs of thermophilic Archaea.** *EMBO Rep* 2002, **3**:537-542.
144. Graham DE, Bock CL, Schalk-Hihi C, Lu ZJ, Markham GD: **Identification of a highly diverged class of S-adenosylmethionine synthetases in the archaea.** *J Biol Chem* 2000, **275**:4055-4059.
145. Graham DE, Xu H, White RH: ***Methanococcus jannaschii* uses a pyruvoyl-dependent arginine decarboxylase in polyamine biosynthesis.** *J Biol Chem* 2002, **277**:23500-23507.

This reference list includes only those references cited in this printed article. For the complete bibliography, see the complete version of this article, online.