linkage analysis in families recruited through 2 asthmatic sibs. Collaborative Study on the Genetics of Asthma (CSGA). *J. Allergy Clin. Immunol.* 102, 436–442
9 Mathias, R.A. *et al.* (2001) Genome-wide linkage analyses of total serum IgE using variance components analysis in asthmatic families. *Genet. Epidemiol.* 20, 340–355
10 Dizier, M.H. *et al.* (2000) Genome screen for asthma and related phenotypes in the French EGEA study. *Am. J. Respir. Crit. Care Med.* 162, 1812–1818
11 Laitinen, T. *et al.* (2001) A susceptibility locus for asthma-related traits on chromosome 7 revealed by genome-wide scan in a founder population. *Nat. Genet.* 28, 87–91
12 Hakonarson, H. *et al.* (2002) A major susceptibility gene for asthma maps to chromosome 14q24. *Am. J. Hum. Genet.* 71, 483–491
13 Koppelman, G.H. *et al.* (2002) Genome-wide search for atopy susceptibility genes in Dutch families with asthma. *J. Allergy Clin. Immunol.* 109, 498–506
14 Haagerup, A. *et al.* (2002) Asthma and atopy – a total genome scan for susceptibility genes. *Allergy* 57, 680–686
15 Libert, F. *et al.* (1998) The Δccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. *Hum. Mol. Genet.* 7, 399–406
16 Suarez, B.K. *et al.* (1994) Problems of replicating linkage claims in psychiatry. *Genetic Approaches to Mental Disorders* (Gershorn, E.S., Cloninger, C.R., *et al.* eds), pp. 23–46, American Psychiatric Press
17 Ober, C. *et al.* (1998) Genome-wide search for asthma susceptibility loci in a founder population. The Collaborative Study on the Genetics of Asthma. *Hum. Mol. Genet.* 7, 1393–1398
18 Cookson, W.O. *et al.* (2001) Genetic linkage of childhood atopic dermatitis to psoriasis susceptibility loci. *Nat. Genet.* 27, 372–373
19 Nair, R. *et al.* (1997) Evidence for two psoriasis susceptibility loci (HLA and 17q) and two novel candidate regions (16q and 20p) by genome-wide scan. *Hum. Mol. Genet.* 6, 1349–1356
20 Tosh, K. *et al.* (2002) A region of chromosome 20 is linked to leprosy susceptibility in a South Indian population. *J. Infect. Dis.* 186, 1190–1193
21 Drazen, J.M. and Weiss, S.T. (2002) Genetics: inherit the wheeze. *Nature* 418, 383–384
22 Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature* 405, 847–856
23 Abecasis, G.R. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68, 191–197
24 Patil, N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 1719–1723
25 Lander, E. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11, 241–247
26 Stone, A.L. *et al.* (1999) Structure–function analysis of the ADAM family of disintegrin-like and metalloproteinase-containing proteins. *J. Protein Chem.* 18, 447–465

0168-9525/03/$ - see front matter
doi:10.1016/S0168-9525(03)00025-8

Genome Analysis

Potential genomic determinants of hyperthermophily

Kira S. Makarova, Yuri I. Wolf and Eugene V. Koonin

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

We searched for genes that could be important for hyperthermophily using a flexible approach to phyletic pattern analysis. We identified 290 clusters of orthologous groups of proteins (COGs) that are preferentially present in archaeal and bacterial hyperthermophiles. Of these, 58 COGs include proteins from at least one bacterium and two archaea, and these were considered to be the best candidates for a specific association with the hyperthermophilic phenotype. Detailed sequence and genome-context analysis of these COGs led to functional predictions for several previously uncharacterized protein families, including a novel group of putative molecular chaperones and a unique transcriptional regulator.

The molecular basis of the hyperthermophilic phenotype of numerous prokaryotes remains unclear, although comparisons of orthologous proteins from hyperthermophiles and mesophiles pointed to potential adaptations to the hyperthermophilic environments; in particular, excess of electrostatic interactions [1,2]. Complete genome sequences of 11 hyperthermophiles were available as of August 1, 2002, including eight archaea from six distinct lineages, and three bacteria from diverse phyla. (Hyperthermophiles are defined as organisms with optimal growth temperature >75°C; thermophiles are those with optimal growth temperature of 55–75°C.) With this amount of data at hand, it is tempting to pursue a different avenue of search for potential determinants of this unique phenotype, namely comparative-genomic analysis aimed at the identification of genes that occur exclusively or primarily in hyperthemophiles.

Recent analysis of phyletic patterns (Box 1) in the database of clusters of orthologous groups of proteins (COGs) [3] showed that the only protein encoded in the genomes of all hyperthermophiles, and not in any other genomes, is reverse gyrase [4]. Reverse gyrase consists of a helicase and a Type I topoisomerase, and it introduces positive supercoiling into circular DNA, thus preventing excess local unwinding of the double helix at high temperatures [5]. Although reverse gyrase is, in all likelihood, necessary for hyperthermophily, it is hard to imagine that it alone could account for this phenotype [4]. The absence of other strictly hyperthermophile-specific COGs is not particularly unexpected in view of the substantial horizontal gene flow between thermophiles and mesophiles [6,7] and numerous non-orthologous gene displacements [8], which result in scattered phyletic patterns for most orthologous sets ([3,9]; Box 1). Here, we describe an attempt to search for potential genomic

Corresponding author: Eugene V. Koonin (koonin@ncbi.nlm.nih.gov).

Box 1. Some important definitions and concepts of evolutionary genomics

Phyletic pattern: Also called phylogenetic pattern. Pattern of representation of a set of orthologous genes in genomes of different species [24–26]. Table I shows the phyletic patterns for two COGs discussed in this article, COG2250 and COG2361. Examination of these patterns immediately shows that COG2250 is specific for hyperthermophiles (and missing in only one genome of a hyperthermophile), whereas COG2361 is scattered among thermophiles and mesophiles alike. Establishing a connection between a gene and a phenotype is the application of phyletic patterns that is central to this article. Another important application is identification of cases of non-orthologous gene displacement (see below).

Non-orthologous gene displacement: The situation when the same essential function is performed by unrelated or at least not orthologous proteins [8]. Non-orthologous gene displacement tends to result in phyletic patterns that are partially complementary; the complementarity is rarely perfect because some organisms often have both proteins, resulting in functional redundancy [27].

Lineage-specific expansion of a paralogous gene family: An increase in the number of paralogs as a result of one or more duplications that have occurred after the separation of a given lineage from other compared lineages. Lineage-specific expansions often reflect adaptations to a specific ecological niche [27,28].

Genome-context analysis: An approach in computational genomics whereby functional inferences are made on the basis of various associations between functionally characterized proteins or domains and uncharacterized ones. These associations include fusion of domains within the same protein, juxtaposition of genes in a (predicted) operon, co-expression and phyletic profiles. The context information is particularly reliable when supported by evolutionary conservation of the associations in question [29].

For a recent discussion of these and other aspects of evolutionary genomics, see [30].

Table I. Phyletic patterns for COG2250 and COG2361[a]

	COG2250	COG2361
Archaea		
Archaeoglobus fulgidus	+	−
Methanocaldococcus jannaschii	+	+
Methanopyrus kandleri	+	−
Methanothermobacter thermoautotrophicus	−	−
Methanosarcina acetivorum	+	+
Pyrococcus abyssi	+	−
Pyrococcus horikoshii	+	−
Thermoplasma acidophilum	−	−
Thermoplasma volcanii	−	−
Aeropyrum pernix	+	−
Pyrobaculum aerophilum	+	−
Sulfolobus solfataricus	+	−
Bacteria		
Aquifex aeolicus	−	−
Thermotoga maritima	+	−
Thermoanaerobacter tengcongensis	+	−
Bacillus subtilis	−	−
Clostridium acetobutilicum	−	−
Escherichia coli	−	−
Ralstonia solanaraceum	−	−
Helicobacter pylori	−	−
Caulobacter crescentus	−	+
Synechocystis sp.	−	+
*Nostoc*sp.	−	−
Deinococcus radiodurans	−	+
Corynebacterium glutamicum	−	−
Mycobacterium tuberculosis	−	−
Fusobacterium nucleatum	−	−
Treponema pallidum	−	−
Chlamydia trachomatis	−	−

[a]Plus (+) shows presence and minus (−) shows absence of a member of a COG in the given genome. Hyperthermophiles are shown in purple and thermophiles are shown in red. Only a sampling of sequenced bacterial genomes was included.

determinants of hyperthermophily using a flexible strategy of phyletic pattern analysis.

Recently, using a combination of detailed sequence analysis, structure prediction and gene order comparison, we predicted a previously undetected DNA-repair system, which consists of >20 COGs and appears to be largely specific for hyperthermophilic archaea and bacteria [10]. Although markedly enriched in hyperthermophiles, the COGs comprising this system showed considerable variability of phyletic patterns, which probably reflects major effects of lineage-specific gene loss and horizontal gene transfer. We used these COGs as a template to formulate criteria for phyletic pattern search and identify other COGs with a similar preference for hyperthermophiles, which could be considered potential genomic determinants of hyperthermophily. Specifically, the following criteria were employed: (1) a COG should include proteins from at least three hyperthermophiles, out of 11 sequenced genomes; (2) the number of hyperthermophiles in a COG should be greater than the number of other species; and (3) thermophiles (14 sequenced genomes altogether) should comprise more than half of the COG members.

Phyletic patterns that met the above criteria were recorded for 290 COGs (see Supplementary Material at http://archive.bmn.com/supp/tig/April2003-Makarova_etal.pdf), including 15 COGs (7% of the total), which represented the majority of the components of the predicted thermophile-specific repair system described earlier [10]. Given that among the 11 available genomes of hyperthermophiles eight were from archaea, it was not unexpected that a substantial fraction of the selected COGs were archaea-specific (Fig. 1). The 107 archaea-specific COGs included three components of the predicted DNA repair system (COG2254, COG2462, COG4343) and probably other proteins that are specifically important for hyperthermophily. However, this group of COGs is likely also to include archaeal proteins that are not directly

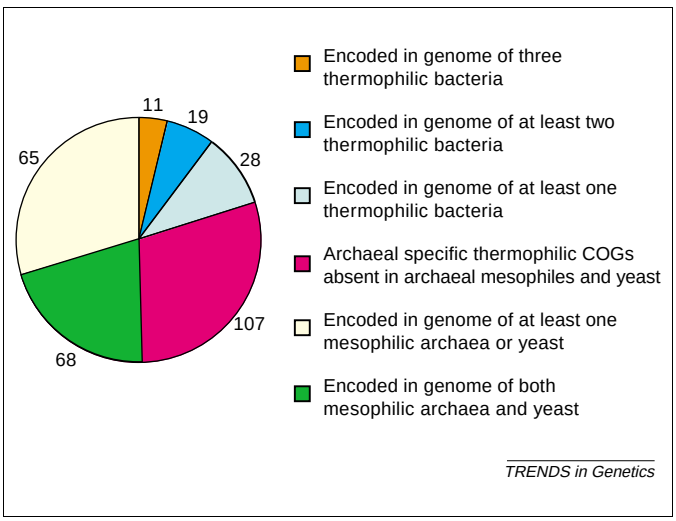

Fig. 1. Phylogenetic classification of COGs enriched in hyperthermophiles (the 290 COG set).

Table 1. COGs enriched in hyperthermophiles and shared by at least two hyperthermophilic bacteria

COG number[a]	Name and comments	HTP (TP)[b]	Non-hTP (MP)[c]
1110	Reverse gyrase	11 (11)	0 (0)
2250[d]	Homologous to COG1895, putative chaperones	10 (10)	0 (0)
1980[d]	Fructose 1,6-bisphosphatase	10 (13)	3 (0)
1688[d]	RAMP superfamily protein	10 (11)	4 (3)
1618[d]	Predicted nucleotide kinase	10 (13)	4 (1)
1313[d]	Homologs of pyruvate formate lyase activating protein PflX	10 (11)	3 (2)
1468[d]	RecB family exonuclease	10 (11)	5 (4)
3635[d]	Predicted phosphoglycerate mutase	10 (13)	5 (2)
1318[d]	Transcriptional regulator, often encoded next to RecA-superfamily ATPases implicated in signal transduction (COG0467)	9 (9)	0 (0)
1350[d]	Predicted alternative tryptophan synthase β-subunit	9 (12)	4 (1)
1353[d]	Predicted DNA polymerase	9 (11)	5 (3)
1144[d]	Pyruvate:ferredoxin oxidoreductase, δ subunit	9 (12)	5 (2)
1578	Predicted acyl-binding protein	8 (9)	2 (1)
1149	P-loop ATPase of the MinD superfamily, contains an inserted ferredoxin domain	7 (8)	2 (0)
1237	Metal-dependent hydrolase of the β-lactamase superfamily	7 (8)	4 (3)
1583	RAMP superfamily protein	7 (9)	3 (1)
1568	Predicted methyltransferases	6 (6)	0 (0)
1906	Predicted membrane transporter	6 (6)	0 (0)
2152	Predicted glycosylase	6 (6)	2 (2)
1336	RAMP superfamily protein	6 (7)	2 (1)
1148	Heterodisulfide reductase subunit A and related polyferredoxins	6 (7)	2 (1)
2516	Biotin synthase-related enzyme	5 (5)	0 (0)
1856	Biotin synthase-related enzyme	5 (5)	0 (0)
1604	RAMP superfamily protein	5 (5)	1 (1)
2406	Predicted hemoprotein distantly related to bacterioferritin	5 (6)	2 (1)
1059	Thermostable 8-oxoguanine DNA glycosylase	5 (7)	3 (1)
1769	RAMP superfamily protein	4 (4)	0 (0)
1542	Conserved protein, contains coiled-coil domains	4 (4)	0 (0)
1367	RAMP superfamily protein	4 (4)	1 (1)
2000	Predicted Fe–S proteins, hydrogenase component	4 (5)	1 (0)

[a]Clusters of orthologous groups (COGs) including components of the predicted thermophile-specific DNA repair system [10] are shown by bold type.
[b]Number of genomes of hyperthermophiles (hTP) and of thermophiles (TP, in parentheses) in the given COG; the COGs in the table are sorted in the descending order of the number of hyperthermophiles represented.
[c]Number of genomes of non-hyperthermophiles (non-hTP; moderate thermophiles and mesophiles) and of mesophiles (MP, in parentheses).
[d]Because these COGs are represented in a significant majority of hyperthermophiles, the remaining genomes of hyperthermophiles were searched using the TBLASTN program to detect potential unannotated COG members; however, such new COG members were not found.

relevant for the hyperthermophilic phenotype, and there is no obvious way, in this case, to differentiate between these two categories of proteins.

Therefore, we concentrated on the 58 COGs, which included, along with archaea, at least one of the three available genomes of hyperthermophilic bacteria, reasoning that, for proteins shared by the phylogenetically disjointed archaeal and bacterial hyperthermophiles, a direct functional link to hyperthermophily was particularly likely. Notably, of the 58 COGs in this set, 12 (21%) belong to the predicted DNA repair system (Table 1), which is compatible with the notion that this group of COGs might be enriched in proteins functionally linked to hyperthermophilicity. For the majority of the COGs in the '58 COG' set, no functional prediction or only a general prediction was available (Fig. 2 and Table 1). We investigated these COGs in detail using analysis of genome context combined with extensive, in-depth database searches [11–13], which resulted in a variety of new functional predictions.

The '58 COG' set included only a few COGs related to general metabolism (Fig. 2). Some of these are essential enzymes of pyruvate metabolism (COG1144), glycolysis (COG1980, COG3635) and amino acid biosynthesis (COG1350), which apparently substitute for analogous enzymes present in other organisms. In addition, there are still several gaps in central metabolic pathways of hyperthermophiles, for which candidate enzymes so far could not be predicted confidently [14–16]. Several COGs

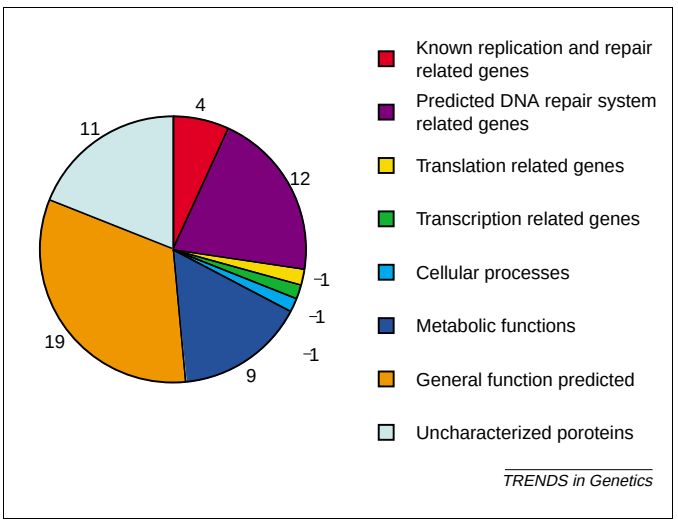

Fig. 2. Functional classification of COGs enriched in hyperthermophiles and represented in at least one hyperthermophilic bacterium (the 58 COG set).

in the '58 COG' list consist of predicted redox enzymes of a recently described superfamily, which utilize iron–sulfur clusters and *S*-adenosylmethionine for radical catalysis (SAM-radical enzymes) [17] and might be involved in diverse metabolic pathways. These enzymes include distant homologs of biotin synthase (COG2516, COG1856) and homologs of pyruvate formate-lyase activating enzyme (COG1313), which are present exclusively in thermophiles. In addition, in two other, larger COGs that consist of SAM-radical enzymes, namely COG0502 (biotin synthase) and COG1180 (pyruvate formate-lyase activating enzyme), most of the thermophiles are represented by several paralogs. We hypothesize that the unique catalytic mechanism of these enzymes is advantageous under low oxygen and high temperature conditions and could be employed in numerous redox reactions in the central metabolism of thermophiles.

COG1318 – a hyperthermophile-specific transcriptional regulator?

The only transcriptional regulator in the '58 COG' set (COG1318) is remarkable in that it nearly fits the definition of a 'genomic signature' of hyperthermophily (i.e. this COG is present exclusively in hypertermophiles and, among these, is missing only from *Sulfolobus* and *Thermoanaerobacter tengcongensis*). In the genomes of the pyrococci, *Thermotoga maritima*, *Pyrobaculum aerophilum* and *Archaeoglobus fulgidus*, the genes of COG1318 are associated with genes for KaiC-like RecA-superfamily ATPase (COG0467). KaiC is one of the three gene products (*kaiABC*) that are responsible for circadian oscillation regulation in cyanobacteria [18]. Recently, it was shown that this cassette is amplified in cyanobacteria under environmental stress [19]. Furthermore, an archaea-specific expansion (Box 1) is notable among COG0467 members, and all genomes of hyperthermophiles encode at least one member of this COG. Because archaea do not have homologs of *kaiA* and *kaiB* genes, which regulate the *kai* operon expression in cyanobacteria, transcriptional regulators of COG1318 probably regulate the expression of the *kaiC* homologs in hyperthermophiles; together, these two genes might have an important role in signal transduction in these organisms.

COG2250 and COG1895 – putative molecular chaperones important for hyperthermophily

COG2250 comes even closer to being a true 'hyperthermophilic signature' because it is represented in all hyperthermophiles, with the single exception of *Aquifex aeolicus*. Moreover, most genomes of hyperthermophiles, particularly crenarchaea, encode several paralogous members of this COG. Extensive PSI-BLAST searches using different queries [13,20] showed that COG2250 proteins were homologous to those in COG1895, another member of the '58 COG' set. Furthermore, a weaker but statistically significant sequence similarity was detected between these proteins and the uncharacterized C-terminal domain of the mammalian protein sacsin, the product of the gene mutated in a distinct form of spastic ataxia [21]. Sacsin also contains a J domain shared with DnaJ-family protein and an HSP90-like ATPase domain, which suggests a molecular chaperone function for this protein [22].

Most members of COG2250 are either encoded adjacent to or are fused within the same polypeptide with 'minimal' nucleotidyltransfereases (MNTs), another protein family expanded in archaea [23]. MNTs are often associated (within the same protein or within an operon) with another small protein (COG2361), which has been proposed to function as a substrate recognition domain or as a molecular chaperone aiding the folding of the MNT [23]. It appears most likely that small proteins from COG2250, COG1895 and COG2361 have similar functions and probably the same structure (secondary structure prediction indicates that they are all-α-helical proteins [22]), although only the former two COGs are linked to hyperthermophily. The biological functions of MNTs and the associated small α-helical proteins (domains) are not known. However, the observations that these α-helical domains are linked with MNTs, which are thought to be incapable of independent, stable folding [13], as well as the predicted molecular chaperone sacsin and some other chromatin-associated domains [22], all suggest a chaperone-like function for these domains. More specifically, we hypothesize that these proteins belong to a previously undetected class of molecular chaperones, which could be involved in chromatin remodeling, particularly in hyperthermophiles.

Conclusions

We believe that the examples discussed above, along with other predictions included in Table 1 and the Supplementary Material (http://archive.bmn.com/supp/tig/April2003-Makarova_etal.pdf), are sufficient to demonstrate the utility of flexible phyletic pattern search for producing experimentally testable predictions of protein functions that correlate with a particular phenotype. Furthermore, even those 11 COGs proteins in the '58 COG' set, for which there was no specific prediction, have an increased likelihood of contributing to the hyperthermophilic phenotype and could therefore be interesting experimental targets. Because of the wide spread of lineage-specific gene loss and horizontal gene transfer in the evolution of prokaryotes, phyletic pattern search for genomic signatures of unique phenotypes cannot be expected to produce many unequivocal results. Nevertheless, with the parallel increase in the number of available genome sequences and the functionally characterized genes linked to a particular biological feature, this approach is expected to become progressively more accurate.

References

1 Vieille, C. and Zeikus, G.J. (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1–43
2 Sterner, R. and Liebl, W. (2001) Thermophilic adaptation of protein. *Crit. Rev. Biochem. Mol. Biol.* 36, 39–106
3 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
4 Forterre, P. (2002) A hot story from comparative genomics: reverse

gyrase is the only hyperthermophile-specific protein. *Trends Genet.* 18, 236–237
5 Rodriguez, A.C. and Stock, D. (2002) Crystal structure of reverse gyrase: insights into the positive supercoiling of DNA. *EMBO J.* 21, 418–426
6 Koonin, E.V. *et al.* (2001) Horizontal gene transfer in prokaryotes – quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742
7 Ragan, M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.* 11, 620–626
8 Koonin, E.V. *et al.* (1996) Non-orthologous gene displacement. *Trends Genet.* 12, 334–336
9 Snel, B. *et al.* (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25
10 Makarova, K.S. *et al.* (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* 30, 482–496
11 Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11, 356–372
12 Aravind, L. (2000) Guilt by association: contextual information in genome analysis. *Genome Res.* 10, 1074–1077
13 Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* 287, 1023–1040
14 Slesarev, A.I. *et al.* (2002) The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4644–4649
15 Smit, A. and Mushegian, A. (2000) Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res.* 10, 1468–1484
16 Selkov, E. *et al.* (1997) A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* 197, GC11–GC26
17 Cheek, J. and Broderick, J.B. (2001) Adenosylmethionine-dependent iron–sulfur enzymes: versatile clusters in a radical new role. *J. Biol. Inorg. Chem.* 6, 209–226
18 Ishiura, M. *et al.* (1998) Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science* 281, 1519–1523
19 Dvornyk, V. *et al.* (2002) Long-term microclimatic stress causes rapid adaptive radiation of kaiABC clock gene family in a cyanobacterium, *Nostoc linckia*, from 'Evolution Canyons' I and II, Israel. *Proc. Natl. Acad. Sci. U. S. A.* 99, 2082–2087
20 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
21 Engert, J.C. *et al.* (2000) ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nat. Genet.* 24, 120–125
22 Makarova, K.S. *et al.* Putative novel chaperone domains expanded in hyperthermophiles and associated with the prokaryotic minimal nucleotidyltransferase and human spastic ataxia protein. *BMC Evolutionary Biol.* (in press)
23 Aravind, L. and Koonin, E.V. (1999) DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucleic Acids Res.* 27, 1609–1618
24 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
25 Ragan, M.A. and Gaasterland, T. (1998) Microbial genescapes: a prokaryotic view of the yeast genome. *Microb. Comp. Genomics* 3, 219–235
26 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
27 Galperin, M.Y. and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* 18, 609–613
28 Jordan, I.K. *et al.* (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11, 555–565
29 Huynen, M.A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.* 54, 345–379
30 Koonin, E.V. and Galperin, M.Y. (2002) *Sequence – Evolution – Function. Computational Approaches in Comparative Genomics*, Kluwer Academic

0168-9525/03/$ - see front matter. Published by Elsevier Science Ltd.
doi:10.1016/S0168-9525(03)00047-7

Why are the genomes of endosymbiotic bacteria so stable?

Francisco J. Silva, Amparo Latorre and Andrés Moya

Institut Cavanilles de Biodiversitat i Biologia Evolutiva and Departament de Genètica, Universitat de València, Apartado 22085, 46071 València, Spain

The comparative analysis of three strains of the endosymbiotic bacterium *Buchnera aphidicola* has revealed high genome stability associated with an almost complete absence of chromosomal rearrangements and horizontal gene transfer events during the past 150 million years. The loss of genes involved in DNA uptake and recombination in the initial stages of endosymbiosis probably underlies this stability. Gene loss, which was extensive during the initial steps of *Buchnera* evolution, has continued in the different *Buchnera* lineages since their divergence.

Bacterial genomes are continuously being modified by the gain and loss of genes, and movement of genes within or between the different DNA molecules that compose the genome. Inversions, translocations and other chromosomal rearrangements are frequently fixed in the genomes of free-living bacteria, and the incorporation of foreign genes through horizontal gene transfer (HGT) is one of the most important aspects of bacterial evolution. Species with the highest HGT rates have a higher probability of incorporating genes that will help them to adapt to their environments or to conquer new niches. Thus, a significant proportion of genetic diversity is obtained through the acquisition of sequences from distantly related organisms [1,2].

Recently however, Tamas and colleagues [3] reported an extreme case of genome stability. They sequenced the genome of *Buchnera aphidicola* BSg, an endosymbiotic γ-proteobacteria that lives in vesicles present in the

Corresponding author: Francisco J. Silva (Francisco.Silva@uv.es).