Supporting Information

Vieira-Silva et al. 10.1073/pnas.1110972108

SI Materials and Methods

Ortholog Alignments. The 61 families of orthologs in the 74 proteobacteria were aligned with MUSCLE (1). Poorly aligned regions were removed using Gblocks with nonstringent parameters, allowing a maximum of 10 nonconserved contiguous positions and a minimum block size of 8 and allowing for gaps in half the sequences (2).

Deep-Phylogenetic Analyses. We used the HOGENOM4 database (3) to identify homolog families present in at least six organisms of at least 10 of the following 15 clades: α-proteobacteria (αP, $n = 47$), β-proteobacteria (βP, $n = 32$), δ-proteobacteria (δP, $n = 12$), ε-proteobacteria (εP, $n = 7$), γ-proteobacteria (γP, $n = 63$), actinobacteria (Ac, $n = 28$), bacillales/lactobacillales (Ba, $n = 34$), bacteroidetes/chlorobi (Bt, $n = 9$), chlamydiae/verrucomicrobia (Ch, $n = 7$), clostridia (Cl, $n = 7$), cyanobacteria (Cy, $n = 13$), spirochaetes (Sp, $n = 7$), tenericutes (Te, $n = 14$), crenarchaeota ($n = 11$), euryarchaeota ($n = 25$).

All deep-phylogenetic trees were reconstructed with RAXML (model WAG+8Γ+I, 100 rapid bootstraps) (4). To assess the topological differences between deep-phylogenies reconstructed using lowly (LEP) or highly expressed proteins (HEP), we counted the number of nodes from the clade's most recent common ancestor to the root using ade4 in R (5, 6). We inferred the minimal generation time of each clade's ancestor using generalized least squares (GLS) with the Brownian motion model (R package *ape*) (7).

Expressivity Indices. We used several expressivity indices based on experimental data (mRNA and protein concentrations), or predictions (codon usage bias) for *Escherichia coli*. The Codon Adaptation Index (CAI) measures the fit between the codon usage of a gene and that of a reference set of highly expressed genes (8). For the latter we used the genes encoding the ribosomal proteins. The mRNA index is a weighted average of three experimental transcriptomics datasets for *E. coli*, normalized to the same average expression among common genes. The corrective factors were: 1.10 (9), 1.05 (10), 2.45 (11). Similarly, the Protein index is a weighted average of two experimental proteomics datasets with corrective factors 0.702 (12) and 1.989 (11, 13). Even though not all *E. coli* genes were analyzed in all proteomics or transcriptomics analysis, all of the 61 orthologs common to the 74 proteobacteria have a corresponding mRNA index, and 60 have a corresponding Protein index. Because the mRNA index is more complete, based on more experiments and independent of sequence composition (contrary to CAI) we used it throughout the work.

Evolutionary Rate Difference. For each pair of protein trees (p_1, p_2), for each taxon (sp), we compute the normalized difference of the terminal branches lengths (r_i) as $(r_{p1} - r_{p2})/(r_{p1}+r_{p2})$. We then computed the correlation between evolutionary rate variation and minimum generation times. For each pair of protein trees, we then computed the nonparametric Spearman correlation [$\rho_{(p1,p2)}$] between the evolutionary rate variation and the minimum generation time attributed to the taxon. The distribution of the coefficients of correlation of the set of pairs of proteins with very different expression levels was then used to test our hypothesis.

We also controlled this analysis for multiple comparisons and phylogenetic nonindependence of the data. To use exclusively independent pairwise comparisons, we sampled 1,000 times, without replacement, the set of 61 ortholog families for the pairs of proteins with significantly different expression levels. We thus obtained 1,000 sets of pairs of protein trees with a fivefold difference in expressivity and where one protein tree is represented at most once in each set. Then we tested for phylogenetic independent contrasts in each set of pairs of proteins. For each pair of protein trees we computed the phylogenetic independent contrasts (14) for the minimum generation times and for the evolutionary rate variation between proteins with ape (7), using the reference tree for the 74 proteobacteria. Finally, we calculated the median of the values of Spearman correlations for each set. The significance of the negative relation between evolutionary rate heterogeneity and minimum generation time is estimated as the number of sets with negative medians (e.g., if all medians are negative as proposed by H_1, then the control (PD-*P* value < 0.001). In short, PD-*P* value refers to the *P* value of the test on the median of the distribution of Spearman correlations using phylogenetically independent contrasts and controlling for multiple comparisons.

Estimation of Effective Population Size. We estimated the effective population size scaled by mutation rate (Ne.u) of 38 species with published minimum generation time by using published data on their genetic diversity (15).

Given the average number of substitutions between silent sites in two randomly sampled allelic sequences (H), Ne.u was computed: Ne.u = 3H/(3–4H) (16).

Assuming similar mutation and recombination rates, the composite parameter Ne.u will correlate to the degree of influence of genetic drift in evolutionary process in different lineages. It is thought that mutation rate is largely determined by genome size (17), and we have previously shown that there is no significant association between minimal generation times and genome size (18).

Skewness. The skewness of the distribution (γ_1) is estimated as follows. A positive coefficient ($\gamma_1 > 0$) indicates a distribution skewed to the left, with the bulk of values falling to the left of the mean. A Gaussian distribution has a null coefficient ($\gamma_1 = 0$).

$$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{sd(x)} \right)^3; \; with \; sd(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

1. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
2. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56:564–577.
3. Penel S, et al. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10(Suppl 6):S3.
4. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
5. Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package I: One-table methods. *R news* 4(1):5–10.
6. R Development Core Team (2009) R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria) http://www.R-project.org. Accessed September, 2009.
7. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
8. Sharp PM, Li WH (1987) The codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281–1295.

9. Allen TE, et al. (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: Model-driven analysis of heterogeneous data sets. *J Bacteriol* 185: 6392–6399.
10. Corbin RW, et al. (2003) Toward a protein profile of *Escherichia coli*: Comparison to its transcription profile. *Proc Natl Acad Sci USA* 100:9232–9237.
11. Masuda T, Saito N, Tomita M, Ishihama Y (2009) Unbiased quantitation of *Escherichia coli* membrane proteome using phase transfer surfactants. *Mol Cell Proteomics* 8: 2770–2777.
12. Lopez-Campistrous A, et al. (2005) Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol Cell Proteomics* 4: 1205–1209.
13. Ishihama Y, et al. (2008) Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* 9:102.
14. Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1.
15. Lynch M (2006) The origins of eukaryotic gene structure. *Mol Biol Evol* 23:450–468.
16. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404.
17. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88:7160–7164.
18. Vieira-Silva S, Touchon M, Rocha EP (2010) No evidence for elemental-based streamlining of prokaryotic genomes. *Trends Ecol Evol* 25:319–320, author reply 320–321.

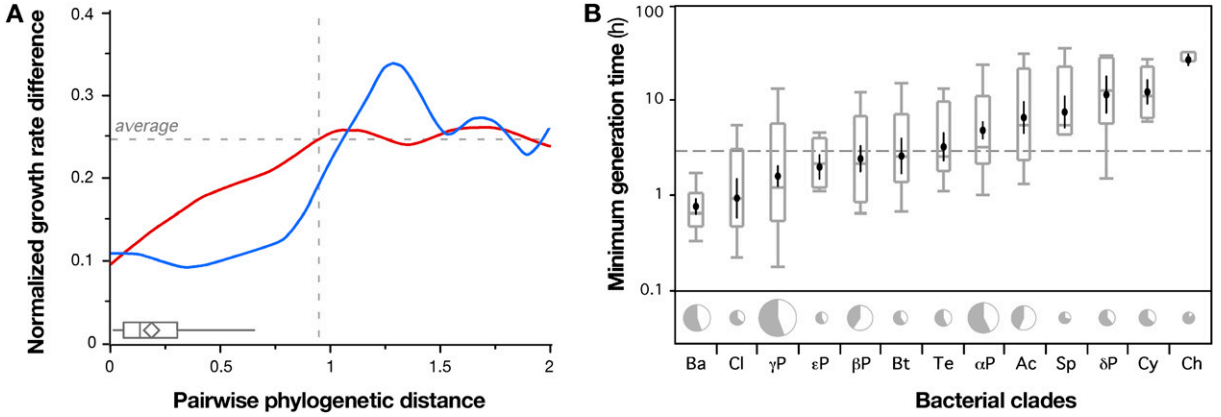

Fig. S1. Minimum generation times in bacteria and archaea. (*A*) Phylogenetic inertia associated to minimum generation time in 154 bacteria and archaea. We computed the pairwise differences in the log-transformed minimum generation times (average: horizontal dashed line). Pairwise phylogenetic distances were computed from the reference phylogeny. A flexible spline was fitted to pairwise comparisons between proteobacteria (red) and between all other taxa except proteobacteria (blue). The latter exhibit higher phylogenetic inertia (i.e., closely related taxa tend to have more similar growth rates). The boxplot on the *Lower Left* represents the distribution of terminal branch lengths in the reference tree of our dataset of 74 proteobacteria, where whiskers span from the minimum to the maximum. These are all below the threshold of significant phylogenetic inertia, represented by the dashed vertical line ($x = 0.95$ substitutions per site). In branches larger than this length in proteobacteria, the probability of an extant fast (slow) grower to remain fast (slow)-grower is the same as of having changed. (*B*) Distribution of characterized minimum generation times in bacterial clades. The surface of the pie-charts is proportional to the number of representatives from each clade in the reference tree, with the gray portion corresponding to those with characterized minimum generation times in the primary literature (1). The distribution of the generation times within each clade is represented by a gray boxplot, where the central line of the box is the median, the edges of the box are the quartiles, and the whiskers extend from the ends of the box to the outermost datapoint that falls within the distances computed: quartiles ± 1.5× interquartile range. The average and SD of the distribution are represented by a black dot and line. The dashed line is the overall average of characterized minimum generation times.

1. Vieira-Silva S, Rocha EPC (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* 6:e1000808.

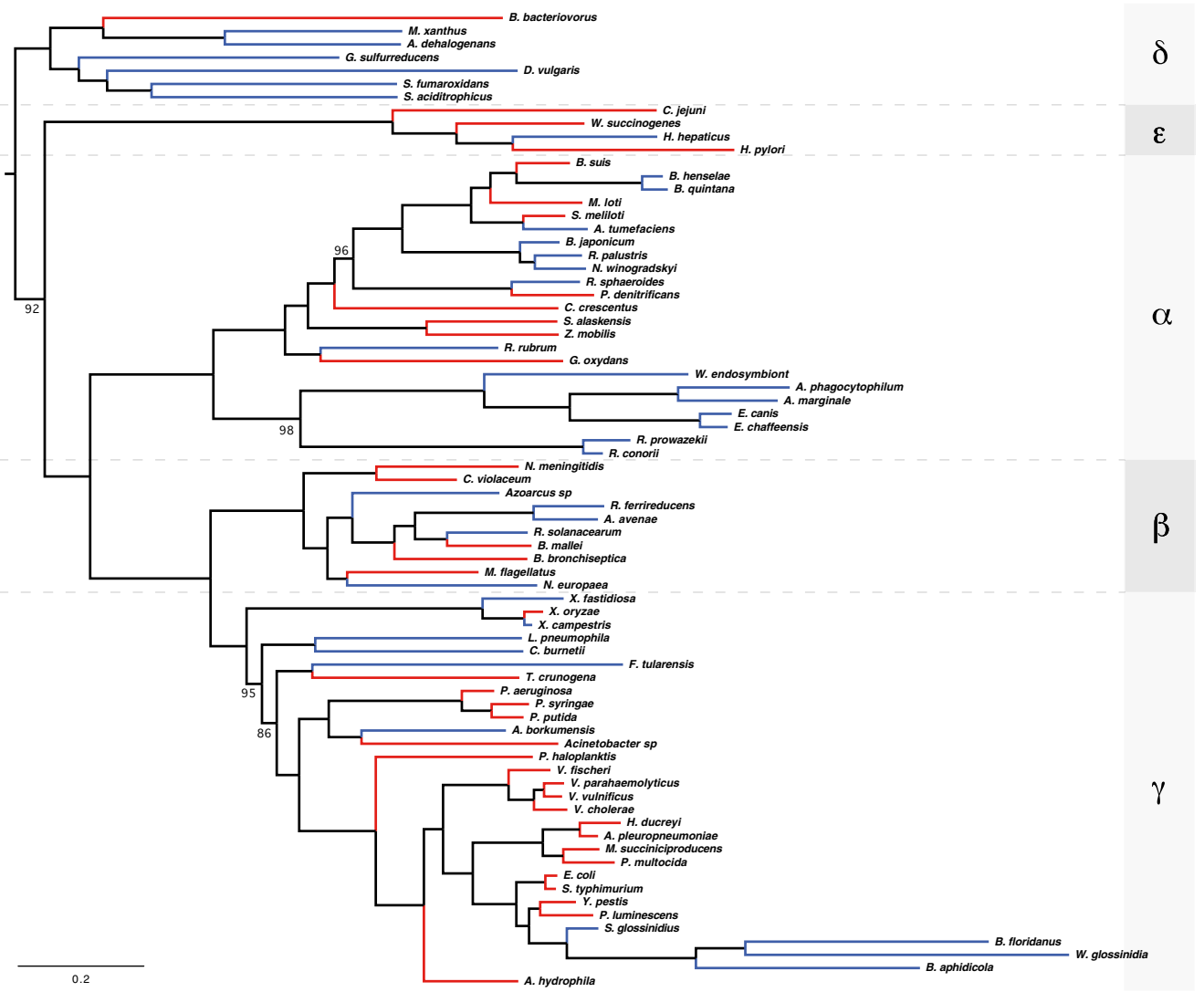

Fig. S2. Reference tree of the 74 proteobacteria with published minimum generation times. Branch supports lower than 99% are reported on the figure, as estimated by rapid bootstrap (4). Branch lengths represent protein evolutionary rates (scale: 0.2 substitutions per site). Red and blue branches correspond to the fast and slow (g > 2.5 h) growers. Classes within proteobacteria are identified by the corresponding greek letter.

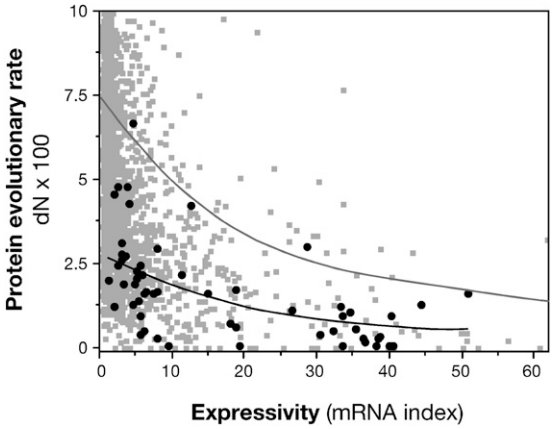

Fig. S3. Protein evolutionary rate variation according to expressivity in *E. coli*. Evolutionary rates for orthologs of *E. coli* and *Salmonella enterica* were retrieved from a previous publication (1). The black datapoints correspond to the subset of *E. coli* orthologs included in the 61 ortholog families shared by the 74 proteobacteria. A smoothing spline was fitted for the whole dataset ($\rho = -0.47$; $P < 10^{-4}$; $n = 2{,}060$) and for the 61 ortholog families ($\rho = -0.68$; $P < 10^{-4}$; $n = 61$). The figure was truncated for values of substitution rates higher than 10%.

1. Rocha EPC, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* 21:108–116.

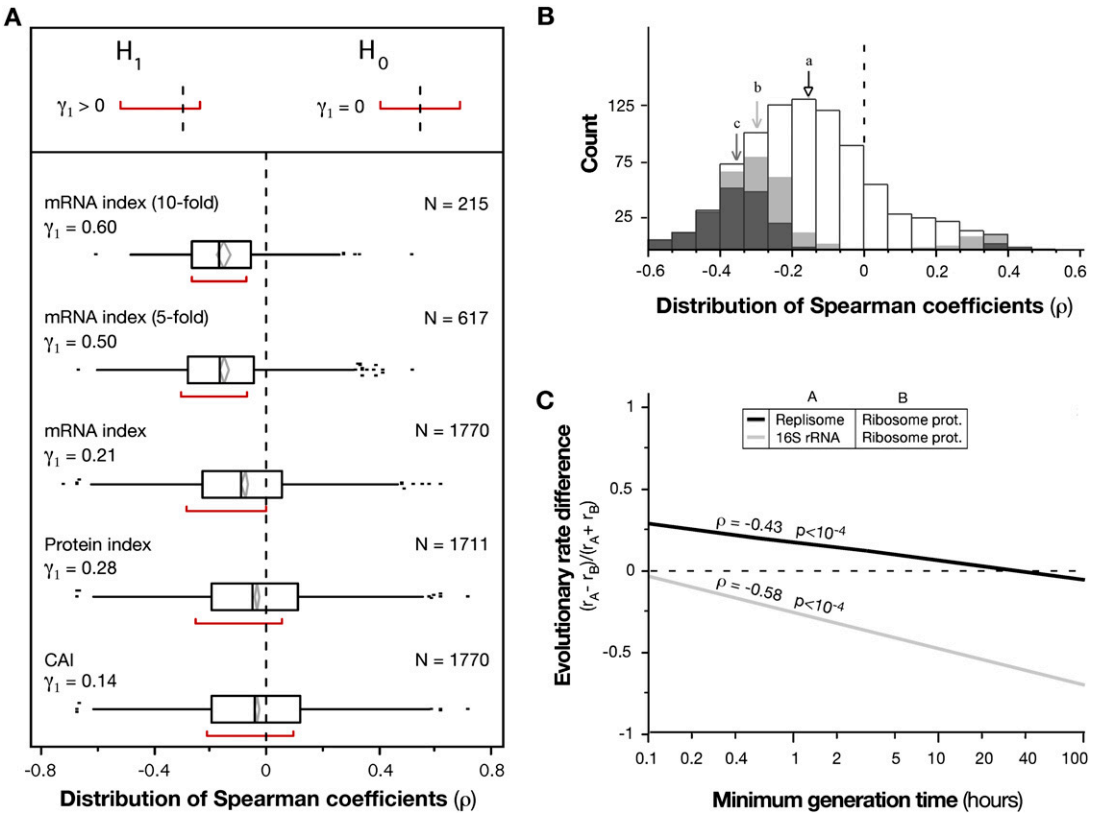

Fig. S4. Analysis of evolutionary rate differences on diverse datasets. (*A*) Distribution of Spearman coefficients of the association between the organism's generation time (g) and evolutionary rate variation between HEP and LEP (Δ) using different expressivity indexes (*SI Materials and Methods*). For any pair of protein where the first has a lower expressivity than the second, we calculated the Spearman correlation between g and Δ. We expect the distribution to be centered on zero under H_0 (no correlation between growth and evolutionary rate variations) or skewed toward negative values under H_1 (negative correlation between growth and evolutionary rate variations between LEP and HEP). The median and quartiles of the distributions are represented in boxplots. The average and 95% confidence interval are represented by a gray diamond. The red bracket identifies the most-dense 50% of the observations. All averages are significantly lower than zero ($P < 10^{-4}$) and the positive skewness coefficients (γ_1) indicate that the bulk of values lie to the left of the mean. (*B*) Distribution of Spearman coefficients of all pairwise comparisons between protein families with a fivefold difference in expressivity (mRNA index): (a) Median over the entire distribution ($n = 617$); (b and c) Median of the pairs that show individually statistically significant correlations (respectively at $P < 0.05$ and $P < 0.01$). (*C*) Ribosomal proteins evolve slower than the replisome proteins and the 16S rDNA with decreasing generation times of 74 proteobacteria.

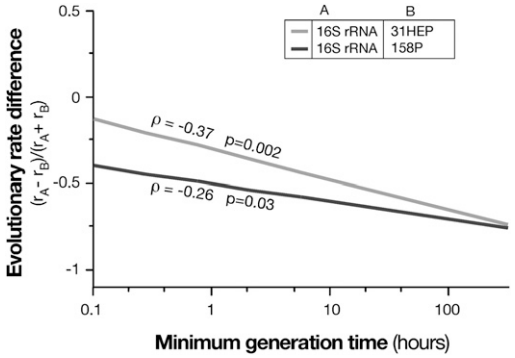

Fig. S5. Evolutionary rate variation in deep-phylogenies according to the minimum generation time. We compared the terminal branch lengths in deep phylogenies reconstructed with 31 HEP (1), 158 homolog families (158P), and 16S rRNA sequences for the set of species with congruent topology in all trees (72 bacteria and an outgroup of 10 archaea).

1. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283−1287.

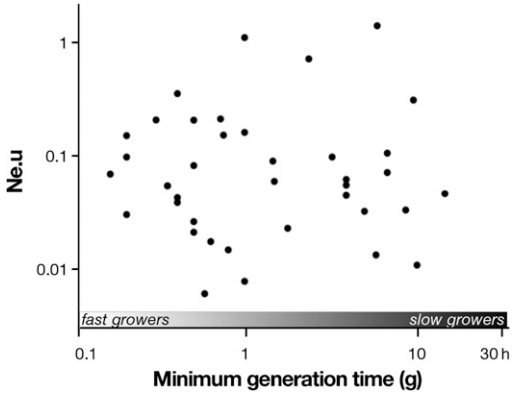

Fig. S6. Effective population size scaled by mutation rate (Ne.u) according to the minimum generation time. We estimated Ne.u for 38 species with published minimum generation time using published genetic diversity data (15). No significant correlation was found between the two variables ($\rho = -0.042$, *P* value = 0.80). From the original dataset in ref. 15, we excluded *Buchnera* (already controlled for when excluding endomutualists) and *Mycobacterium tuberculosis* (marked as an outlier in ref. 15). Inclusion of these two points does not change the conclusions ($\rho = -0.18$, *P* value = 0.27).

Table S1. Negative correlation between the expressivity of the 61 proteins in *E. coli* and their substitution rates between *Escherichia coli MG1655* and *Salmonella enterica LT2* (*E.coli/S.enterica* dN) or the average subtitution rate of the proteins in the 74 proteobacteria tree (average dN)

		E. coli/S. enterica dN		Average dN	
Expressivity	*n*	Spearman ρ	*P* value	Spearman ρ	*P* value
mRNA index	61	$\rho = -0.68$	$< 10^{-4}$	$\rho = -0.32$	0.01
CAI	61	$\rho = -0.39$	0.003	$\rho = -0.45$	0.0003
Protein index	60*	$\rho = -0.52$	$< 10^{-4}$	$\rho = -0.41$	0.001

See *SI Materials and Methods* for a definition of the expressivity indices.
*Protein concentration unknown for one protein.

Table S2. Proteins IDs and function included in the ribosome and replisome concatenates and in the concatenates of HEP and LEP of the HOGENOM4 reference set

Concatenate	Protein name/ID	Function
Proteobacterial orthologs		
Ribosome (2,446 sites)	rplA	50S Ribosomal subunit protein L1
	rplB	50S Ribosomal subunit protein L2
	rplC	50S Ribosomal subunit protein L3
	rplD	50S Ribosomal subunit protein L4
	rplF	50S Ribosomal subunit protein L6
	rplJ	50S Ribosomal subunit protein L10
	rplK	50S Ribosomal subunit protein L11
	rplL	50S Ribosomal subunit protein L7/L12
	rplM	50S Ribosomal subunit protein L13
	rplO	50S Ribosomal subunit protein L15
	rplP	50S Ribosomal subunit protein L16
	rplR	50S Ribosomal subunit protein L18
	rpsA	30S Ribosomal subunit protein S1
	rpsC	30S Ribosomal subunit protein S3
	rpsD	30S Ribosomal subunit protein S4
	rpsG	30S Ribosomal subunit protein S7
	rpsM	30S Ribosomal subunit protein S13
Replisome (2,199 sites)	dnaB	Replicative DNA helicase
	dnaE	DNA polymerase III α subunit
	dnaN	DNA polymerase III, β subunit
	gyrA	DNA gyrase type II topoisomerase subunit A
	gyrB	DNA gyrase type II topoisomerase subunit B
HOGENOM4 homolog families		
HEP (8,131 sites)	HBG000348	50S Ribosomal protein L19
	HBG001285	DNA directed RNA polymerase β chain
	HBG001445	DNA directed RNA polymerase α chain
	HBG003915	Phosphoglycerate kinase triosephosphate isomerase
	HBG011746	Elongation factor TS
	HBG023340	50S Ribosomal protein L11 2 L11 3
	HBG046904	30S Ribosomal protein S12
	HBG055795	50S Ribosomal protein L4
	HBG072316	50S Ribosomal protein L18
	HBG079273	50S Ribosomal protein L5
	HBG083688	Trigger factor 2
	HBG110521	50S Ribosomal protein L20
	HBG114333	30S Ribosomal protein S16
	HBG119370	30S Ribosomal protein S11
	HBG142539	Translation initiation factor IF 3
	HBG154396	50S Ribosomal protein L3
	HBG155458	50S Ribosomal protein L2 RPLB
	HBG174293	Nucleoside diphosphate kinase b
	HBG180703	50S Ribosomal protein L27
	HBG187258	16S rRNA processing protein rimM
	HBG187568	50S Ribosomal protein L28 2
	HBG192816	30S Ribosomal protein S8
	HBG211133	50S Ribosomal protein L13
	HBG232228	50S Ribosomal protein L22
	HBG232265	30S Ribosomal protein S17 2
	HBG236852	30S Ribosomal protein S10
	HBG249765	50S Ribosomal protein L21
	HBG297151	50S Ribosomal protein L6
	HBG301585	Phenylalanyl trna synthetase β chain
	HBG301748	nusA antitermination factor
	HBG313666	30S Ribosomal protein S6
	HBG313694	50S Ribosomal protein L7 L12
	HBG323189	50S Ribosomal protein L24
	HBG331018	50S Ribosomal protein L17
	HBG338914	30S Ribosomal protein S18 2
	HBG341023	50S Ribosomal protein L16
	HBG345344	50S Ribosomal protein L9
	HBG347903	Alanyl tRNA synthetase protein

Table S2. Cont.

Concatenate	Protein name/ID	Function
	HBG353111	30S Ribosomal protein S9
	HBG369991	50S Ribosomal protein L10
	HBG378671	30S Ribosomal protein S7 2
	HBG380059	50S Ribosomal protein L15
	HBG381711	30S Ribosomal protein S13
	HBG382195	30S Ribosomal protein S19
	HBG384371	50S Ribosomal protein L14
	HBG386975	30S Ribosomal protein S5
	HBG396698	30S Ribosomal protein S3
	HBG399262	tRNA guanine methyltransferase
	HBG415562	30S Ribosomal protein S20
	HBG433393	50S Ribosomal protein L32 3
	HBG448673	Preprotein translocase secy
	HBG468280	50S Ribosomal protein L35
	HBG481547	30S Ribosomal protein S2
	HBG507162	Hypothetical protein ybaB
	HBG515224	50S Ribosomal protein L34
	HBG529802	50S Ribosomal protein L1
	HBG532383	50S Ribosomal protein L23
LEP (4,469 sites)	HBG000055	Regulation factor membrane-associated
	HBG000124	Predicted regulation factor
	HBG001073	Pantetheine-phosphate adenylyltransferase
	HBG001081	Predicted rRNA methylase
	HBG001135	Hemolysin protein
	HBG001183	Holliday junction ATP dependent DNA helicase ruvB
	HBG001394	Glucose inhibited cell-division protein
	HBG060350	Heat inducible transcription repressor
	HBG230304	Hypothetical protein yqeL
	HBG123577	Putative methyltransferase SAM dependent
	HBG144394	Regulation factor cell partioning and DNA repair
	HBG184931	Primosomal protein N′ superfamily II helicase
	HBG185190	Protein yggJ
	HBG187703	Protein ygcC aminodeoxychorismate lyase
	HBG194406	Holliday junction resolvase
	HBG236550	Predicted o sialoglycoprotein endopeptidase
	HBG302091	Hypothetical protein yjeE
	HBG328898	ATP dependent DNA helicase recG
	HBG377949	DNA mismatch repair protein mutS
	HBG410043	Segregation and condensation protein A
	HBG457716	DNA repair protein recN
	HBG505437	Chromosome segregation protein smc

Table S3. Clades branching closest to the archaeal outgroup in deep-phylogenetic reconstruction based on HEP and LEP markers

			Closest†	Among closest‡	Closest†	Among closest‡
Clade	Ancestral log(g)*	Average log(g)	HEP trees		LEP trees	
Cl: clostridia	−0.33	−0.41	0	7.4	**13.8**	**42.6**
Ba: bacillales	−0.08	−0.11	0	1.8	0.2	0.2
εP: ε-proteobacteria	0.31	0.29	4.2	5.8	0	0
Te: tenericutes	0.33	0.51	7.2	9	**11.6**	**12.8**
βP: β-proteobacteria	0.35	0.38	0	2.6	0	8.6
γP: γ-proteobacteria	0.37	0.14	0.2	2.8	0	8.6
Ac: actinobacteria	0.49	0.84	0.2	9.6	**29.6**	**30**
αP: α-proteobacteria	0.68	0.68	0	3	3.4	**14.8**
Bt: bacteroidetes	0.8	0.43	0.6	4.8	0	0
δP: δ-proteobacteria	0.82	1.01	0	3.4	0	2.4
Sp: spirochaetes	0.87	0.88	**24.2**	**29**	0	0
Cy: cyanobacteria	1.17	1.13	**16.8**	**22.8**	0	0.4
Ch: chlamydiae	1.33	1.43	**28.4**	**31.4**	0	0

See Fig. S1*B* for the distribution of minimal generation times (g).

*Clades ordered according to inferred ancestral minimum generation time using GLS with the Brownian motion model (R package *ape*).

†Trees (%) where the clade's ancestor is the node closest to the archaeal outgroup. Boldface numbers: Trees (%) > 10.

‡Trees (%) where the clade's ancestor is part of several ancestor nodes equidistant to the root. Boldface numbers: Trees (%) > 10.