TA Isenbarger, CE Carr, SS Johnson, M Finney, GM Church, W Gilbert, MT Zuber, and G Ruvkun. The most conserved genome segments for life detection on Earth and other planets. *Origins of Life and Evolution of Biospheres* [Full citation TBD]

Electronic Supplementary Material

Supplementary tables. Table S1 summarizes the conserved regions of ribosomal 16S and 23S genes identified using the genome filtering procedure. Table S2 summarizes the distribution of low-scoring conserved sequences within NCBI COG groups.

Supplementary data. Summary of all conserved sequences identified at a cutoff value of 28. All 782 sequences identified at the score cutoff of 28 are identified by the coordinates of the first and last nucleotides covered by the conserved sequence (*E. coli* begin and *E. coli* end, respectively), the length of the sequence (length), the gene or genomic region covered by the sequence (*E. coli* gene), a brief description of that region (description), and the COG classification for sequences within genes (COG). Also, 3 scores are listed for each sequence resulting from blastn comparison to each of the12 genomes used for the search (other than *E. coli*): "bits mean" and "E mean" are the means of the 12 bits scores or 12 E values reported, and "normalized bits mean" is the "bits mean" divided by the blastn score of the sequence compared against itself using blastn (a "BLAST score ratio" similar to (Rasko et al. 2005)). A slash between two or more gene names indicates that the sequence overlaps multiple open reading frames; if these genes are in different COGs, this is indicated by multiple COG designators in the COG column. Intergenic regions are noted by "IG" and the genes between which the sequence falls. Significant sequence flanking an open reading frame is noted by a "<" or ">" before or after the gene name and an indication that the sequence may be in an untranslated region (3' or 5' UTR). For sequences that cover overlapping open reading frames, this is indicated by ">>", "<<", "<>", or "><" to indicate the transcriptional directions of the overlapping reading frames. The NCBI *E. coli K12* (Blattner et al. 1997) online sequence browser was used to assign annotations and COG designations.

Table S1. Conserved regions of ribosomal 16S and 23S genes identified.

			Regions of 16S and 23S genes identified						
C	F	A	*rrnH*	*rrnG*	*rrnD*	*rrnC*	*rrnA*	*rrnB*	*rrnE*
60	229	12	1867-1999 2496-2599	1867-1999 2496-2599	–	1867-1999 2496-2595	1873-1999 2496-2599	1867-1999 2496-2599	1867-1999 2496-2599
			–	–	–	–	–	–	–
55	784	20	1856-2002 2413-2638	1856-2002 2413-2638	1874-1994 2413-2537	1856-2002 2445-2638	1858-2002 2413-2638	1856-2002 2413-2638	1856-2002 2413-2638
			871-986	871-986	871-986	–	871-986	871-986	871-986
50	2065	35	415-541 1846-2010 2409-2652	415-545 1846-2010 2409-2652	415-545 1868-1998 2409-2652	415-545 1846-2010 2415-2652	415-545 1848-2010 2409-2652	415-545 1846-2010 2409-2652	415-545 1846-2010 2409-2652
			861-986 1433-1542	861-986 1433-1542	861-986 1433-1542	884-987 1433-1542	861-986 1433-1542	861-986 1433-1542	861-986 1433-1542
45	2915	35	386-546 1843-2013 2409-2671	386-546 1843-2013 2409-2671	386-546 1853-2001 2409-2671	383-546 1843-2013 2410-2671	386-546 1845-2013 2409-2671	386-546 1843-2013 2409-2671	386-546 1843-2013 2409-2671
			858-986 1307-1542	858-986 1307-1542	858-986 1307-1542	869-987 1307-1542	858-986 1307-1542	858-986 1307-1542	858-986 1307-1542
40	3734	48	381-546 1603-1708 1838-2022 2179-2304 2409-2676	381-546 1603-1708 1838-2022 2179-2304 2409-2676	381-546 1603-1708 1838-2008 2179-2304 2409-2676	374-546 1603-1708 1838-2022 2408-2676	381-546 1603-1708 1840-2022 2179-2304 2409-2676	381-546 1603-1708 1838-2022 2179-2304 2409-2676	381-546 1603-1708 1838-2022 2179-2304 2409-2676
			849-986 1299-1542	849-986 1299-1542	849-986 1299-1542	860-987 1299-1542	849-986 1299-1542	849-986 1299-1542	849-986 1299-1542
38	4088	56	380-546 1602-1708 1835-2025 2178-2305 2409-2677	380-546 1602-1708 1835-2025 2178-2305 2409-2677	380-546 1602-1708 1835-2009 2178-2305 2409-2677	373-546 1602-1708 1835-2025 2178-2303 2408-2677	380-546 1602-1708 1837-2025 2178-2305 2409-2677	380-546 1602-1708 1835-2025 2178-2305 2409-2677	380-546 1602-1708 1835-2025 2178-2305 2409-2677
			485-614 848-986 1299-1542	485-613 848-986 1299-1542	485-614 848-986 1299-1542	485-614 859-987 1299-1542	485-614 848-986 1299-1542	485-614 848-986 1299-1542	485-614 848-986 1299-1542

From left to right, columns indicate the cutoff value (C), number of 100 bp fragments identified (F), number of contiguous sequences assembled from overlapping 100 bp fragments (A) at each cutoff value, and conserved regions within each of the *E. coli* ribosomal gene operons (*rrnH*, *rrnG*, *rrnD*, *rrnC*, *rrnA*, *rrnB*, *rrnE*). Each score cutoff row is divided by a horizontal line—regions within the 23S and 16S genes are indicated in the top and bottom halves, respectively. Score cutoff values from 60-38 identified only regions within ribosomal RNA genes. At score cutoff values less than 38, other conserved sequences were identified in addition to ribosomal RNA sequences.

Table S2. Distribution of low-scoring conserved sequences within NCBI COG groups.

COG	*E. coli*	%	all	%	Δ	Description
G	368	8.7	61	13.1	4.4	Carbohydrate transport and metabolism
S	308	7.3	59	12.7	5.4	Function unknown
E	350	8.3	41	8.8	0.5	Amino acid transport and metabolism
L	220	5.2	38	8.2	3.0	Replication, recombination, repair
J	171	4.0	29	6.2	2.2	Translation
C	275	6.5	29	6.2	–0.3	Energy production and conversion
K	280	6.6	27	5.8	–0.8	Transcription
P	191	4.5	26	5.6	1.1	Inorganic ion transport and metabolism
R	338	8.0	25	5.4	–2.6	General function prediction only
T	134	3.2	22	4.7	1.5	Signal transduction mechanisms
M	235	5.5	17	3.7	–1.8	Cell well/membrane biogenesis
F	87	2.0	17	3.7	1.7	Nucleotide transport and metabolism
O	128	3.0	16	3.4	0.4	Posttranslational modification, protein turnover, chaperones
I	83	2.0	13	2.8	0.8	Lipid transport and metabolism
V	48	1.1	11	2.4	1.3	Defense mechanisms
N	107	2.5	8	1.7	–0.8	Cell motility
H	123	2.9	7	1.5	–1.4	Coenzyme transport and metabolism
D	34	0.8	5	1.1	0.3	Cell cycle control, mitosis, meiosis
Q	68	1.6	5	1.1	–0.5	Secondary metabolites biosynthesis, transport, catabolism
U	37	0.9	4	0.9	0.0	Intracellular trafficking and secretion
Total	4240		466			

The number of sequences encoding products classified in each COG is shown for the entire *E. coli* genome (column 2, "*E. coli*") and for the 466 low-scoring sequences (column 3, "all"). The percentages of sequences in each category were calculated for the *E. coli* genome and for the identified sequences, and the difference in the two percentages is listed in the Δ column. Positive Δ values denote enrichment of a COG class with sequences identified by the search.