Orig Life Evol Biosph
DOI 10.1007/s11084-008-9148-z

ASTROBIOLGY

The Most Conserved Genome Segments for Life Detection on Earth and Other Planets

Thomas A. Isenbarger · Christopher E. Carr · Sarah Stewart Johnson · Michael Finney · George M. Church · Walter Gilbert · Maria T. Zuber · Gary Ruvkun

Received: 17 June 2008 / Accepted: 23 September 2008

Abstract On Earth, very simple but powerful methods to detect and classify broad taxa of life by the polymerase chain reaction (PCR) are now standard practice. Using DNA primers corresponding to the 16S ribosomal RNA gene, one can survey a sample from any environment for its microbial inhabitants. Due to massive meteoritic exchange between Earth and Mars (as well as other planets), a reasonable case can be made for life on Mars or other planets to be related to life on Earth. In this case, the supremely sensitive technologies used to study life on Earth, including in extreme environments, can be applied to the search for life on other planets. Though the 16S gene has become the standard for life detection on Earth, no genome comparisons have established that the ribosomal genes are, in fact, the most conserved DNA segments across the kingdoms of life. We present here a computational comparison of full genomes from 13 diverse organisms from the Archaea, Bacteria, and Eucarya to identify genetic sequences conserved across the widest divisions of life. Our results identify the 16S and 23S ribosomal RNA genes as well as other universally conserved nucleotide sequences in genes encoding particular classes of transfer RNAs and within the nucleotide binding domains of ABC transporters as the most conserved DNA

Christopher E. Carr and Thomas A. Isenbarger have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11084-008-9148-z) contains supplementary material, which is available to authorized users.

T. A. Isenbarger · M. Finney · G. Ruvkun (✉)
Department of Molecular Biology, Massachusetts General Hospital, and Microbial Sciences Initiative, Harvard University, Cambridge, MA, USA
e-mail: ruvkun@molbio.mgh.harvard.edu

C. E. Carr · S. S. Johnson · M. T. Zuber
Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

G. M. Church
Department of Genetics, Harvard Medical School, Cambridge, MA, USA

W. Gilbert
Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

sequence segments across phylogeny. This set of sequences defines a core set of DNA regions that have changed the least over billions of years of evolution and provides a means to identify and classify divergent life, including ancestrally related life on other planets.

Keywords Polymerase chain reaction · DNA · RNA · Ribosome · Panspermia · Mars

Introduction

Increasing evidence, including theoretical calculations (Melosh 1988, 1993; Horneck et al. 1994; Gladman and Burns 1996; Gladman et al. 1996; Mileikowsky et al. 2000), low temperature transfer of meteorite cores (Weiss et al. 2000), and microbe survival during hypervelocity shocks (Horneck et al. 2008), suggests that objects capable of carrying life have been transferred between solar system bodies with significant frequency. In addition, microbial life has been discovered in Earth environments exposed to high radiation and freezing temperatures. While these environments are not as extreme as those on Mars and other planets, such findings demonstrate the incredible adaptability of microbes and suggest that habitable zones are much broader than previously thought. Together these facts raise the possibility that life could have been transferred between Earth and Mars early in the history of the solar system and could survive on Mars today.

Strategies for detecting life on other planets have sought to avoid the assumption it would share any particular features with life on Earth. The most general strategies—seeking informational polymers, structures of biogenic origin, or chemical or isotopic signatures of enzymatic processes—look for features that all life is expected to exhibit. This generality comes at a cost: the strategies are not particularly sensitive and, more importantly, there are abiological routes to these signatures. However, if life on Earth is related to life on other planets, we can use a far more powerful and information-rich technique, one currently used to detect and classify the most deeply divergent and extreme forms of life on Earth.

The current standard method for life detection on Earth uses the polymerase chain reaction, or PCR, to sample and classify microbial life in a wide range of terrestrial environments, ranging from the extreme and nutrient-poor (for example, acidic hot springs) to the temperate and nutrient-rich (for example, soil). PCR is a thermal cycling technique that is used to amplify genetic material, or DNA, to produce sufficient DNA for detection and analysis. When amplifying DNA in a specific, nonrandom manner, one must know the DNA sequences that flank a given region of interest, but not the DNA sequence between these regions.

The PCR strategy for life detection emerged from comparing nucleotide sequences of one of the first types of genes to be sequenced, the ribosomal RNA (rRNA) genes. These genes have been known for many years to vary much less across the tree of life than most genes. Ribosomal RNAs are the main structural and catalytic components of the ribosome, a molecular machine that translates RNA into proteins (Barns et al. 1994). It is the slow rate of change of certain regions of ribosomal genes that makes them the contemporary standard for life detection. For example, within the ˜1,500 nucleotides of the 16S (or 18S in eucarya; herein, for simplicity, we use 16S to represent both 16S and 18S) rRNA gene, there are multiple 15–20 nucleotide segments that are nearly exactly the same in all known organisms (Giovannoni et al. 1990). These segments are essential for ribosomal catalytic activity and have remained unchanged over billions of years (Fox et al. 1980).

The technology of PCR involves stable 15–25 nucleotide long DNA primers, a stable enzyme, nucleotide triphosphate monomers, and a simple heat pump that thermally cycles 20–30 times in about 2–3 h. To amplify 16S genes from a crude sample, universal DNA primers that are complementary to the ribosomal RNA gene and oriented towards each other are added to crudely purified DNA isolated from an environmental sample (for example, a few liters of sea water filtrate or 1 g of soil). As an example, for the ribosomal genes, the DNA primer 5′ GTGBCAGCMGCCGCGGTAA 3′, which corresponds to nucleotides 515–533 of the 16S gene, and 5′ GGHWACCTTGTTACGACTT 3′, which corresponds to the base pairing complement of nucleotides 1492–1510, are commonly used (B, H, M, and W are codes for mixtures of nucleotides necessary to capture all 16S genes; the numbering corresponds to that of *E. coli*, a convenient reference). Thirty thermal cycles will produce one billion (2^{30}) copies of the original sequences, and this amplified DNA can be analyzed for size and DNA sequence. PCR will even amplify complex mixtures of 16S ribosomal RNA genes from communities of organisms in environmental samples. Thus, PCR with DNA primers corresponding to the conserved elements can be used to amplify DNA from any species more than a billion fold, without need to isolate, culture, or grow the organism in any way (Giovannoni et al. 1990). PCR instruments involve simple and efficient heating and cooling of 0.1–100 mL samples. The technology is mature: thermal cycling machines are used in small labs all over the world and field PCR thermal cyclers are used, for example, in the military to detect biological warfare agents.

Currently, hundreds of research groups use 16S rRNA PCR to prospect for new archaeal and bacterial species from a wide range of environments (Friedmann and Ocampo-Friedmann 1984; Giovannoni et al. 1990; Barns et al. 1994; Fry et al. 1997; Pedersen 2000), and thousands of microbial species have been described in this way. Most of the life that is detected by PCR cannot be grown in the laboratory, suggesting either very slow growth rates or very particular growth conditions not met in the laboratory (for example, growth in complex communities). In this way it has been established that the previous culture-based exploration of microbial diversity missed 99% of the living world. Importantly, these molecular methods have led not only to the discovery of numerous previously unknown organisms, but also to the definition of entirely new high-level taxa (Woese 1987; Woese et al. 1990; Rappe and Giovannoni 2003; Isenbarger et al. 2008). Use of this molecular framework has provided biologists with a more dynamic view of the ancestry and taxonomic organization of all life. Such surveys of extreme environments have expanded what we consider to be habitable zones to below 0°C and over 110°C, from acidic hot springs to highly radioactive nuclear reactor pools to deep into the crust of the Earth, and has allowed particular 16S rRNA gene sequences to be assigned, with some caveats, to particular metabolic strategies (Friedmann and Ocampo-Friedmann 1984; Fry et al. 1997; Pedersen 2000).

Yet, the 16S ribosomal gene emerged as the standard for such studies before the recent determination of dozens of full genome sequences from widely divergent microbial and other taxa (Fox et al. 1980; Woese et al. 1990). It is now possible to assess whether in fact the 16S ribosomal gene is the most conserved DNA segment for such life detection strategies. While a large number of other studies have investigated the distribution and conservation of universal *protein* coding genes (Harris et al. 2003; Koonin 2003; Iyer et al. 2004; Santos and Ochman 2004), there has not been an exploration of conserved DNA elements that could reveal other RNA genes or universal *cis* regulatory elements.

We report here a comparison of 13 full genomes from disparate organisms to identify exceptionally conserved nucleotide sequences. Conserved sequences from protein-coding regions, structural RNAs, and non-coding regions were identified. This analysis demonstrates that the 16S and 23S ribosomal genes are the most conserved genomic regions across diverse microbial and non-microbial taxa; our analysis also reveals that particular tRNA genes, nucleotide binding domains of ATP-binding cassette transporters, and other sequences are also near universal.

Materials and Methods

We identified conserved genomic regions by quantifying sequence similarities in the complete genomes of 13 organisms representing disparate branches of the evolutionary tree of life; a second set of 13 genomes was used as a control to verify that the results were independent of the specific genomes selected. Organisms were selected to provide sufficient coverage of major divisions within the three phylogenetic domains of life (Fig. 1, Table 1), including, for each dataset, 6–7 Bacteria, 4–5 Archaea, and two Eucaryotes, and to represent a wide range of environments and metabolic capabilities. Full genome sequences were obtained from NCBI (Table 1).

Genome Filtering Method

To search for conserved sequences, we developed an iterative 'genome filtering' method (Fig. 2) to identify sequences shared by 13 disparate genomes. Searches were initiated with the *E. coli* K12 genome because it is well annotated and facilitates the reliable and accurate identification of the resulting sequences. Score cutoffs from 28–60 bits were used and conserved sequences were ranked according to the highest score that allowed passage through the entire filtering process.

The 'genome filtering' method was implemented using Perl (Wall et al. 1996) to perform the following steps (Fig. 2): 1) the program was provided a query genome sequence, a set of target genome sequences, and a bits score cutoff; 2) a tiled set of 100 bp overlapping sequences, offset by 1 bp (99 bp overlap), was created from the query genome; 3) the 100 bp query genome sequences were aligned to a target genome using blastn; 4) query genome sequences were selected if they satisfied the score cutoff, and sequences that did not pass the cutoff were removed from the query set and not used in further iterations—the entire 100 bp sequence was retained or removed, not only the high-scoring region identified by blastn; 5) the remaining 'filtered' set of selected 100 bp sequences was passed on to be aligned to the next target genome and the comparison and selection process repeated using the same cutoff value. After each iteration, only those 100 bp sequences that passed the previous selection were used in the next comparison.

After each target genome comparison, a final set of 100 bp sequences remained that was distinguished by the score cutoff used to select that particular set. For each final set of 100 bp sequences at a particular score cutoff, overlapping sequences were merged into contiguous sequences that were distinguished by the score cutoff that resulted in the set of 100 bp sequences from which the contiguous sequences were created.

For the genome filtering algorithm, NCBI blastn (2.2.5 and 2.2.6) was used for sequence alignments (Altschul et al. 1997). Default parameters were used except the word size was decreased to nine (–W 9), queries were not pre-filtered for sequence complexity (–F F), and a mismatch penalty of –1 (–q –1) was used.

To confirm the reproducibility of the genome filtering method, two control experiments were performed: 1) a search was performed at a cutoff of 30 bits using the *Pseudomonas aeruginosa* genome for reference instead of the *E. coli* genome, and 2) a search was performed at a cutoff of 28 using a second set of 13 genomes that provided similar coverage of the universal tree of life (see Methods). These searches used a lower resolution tiling (25 bp offset; 75 bp overlap) to shorten the search times for controls; however, this tiling also likely resulted in some sequences being missed, as indicated by preliminary results comparing tiling resolutions.

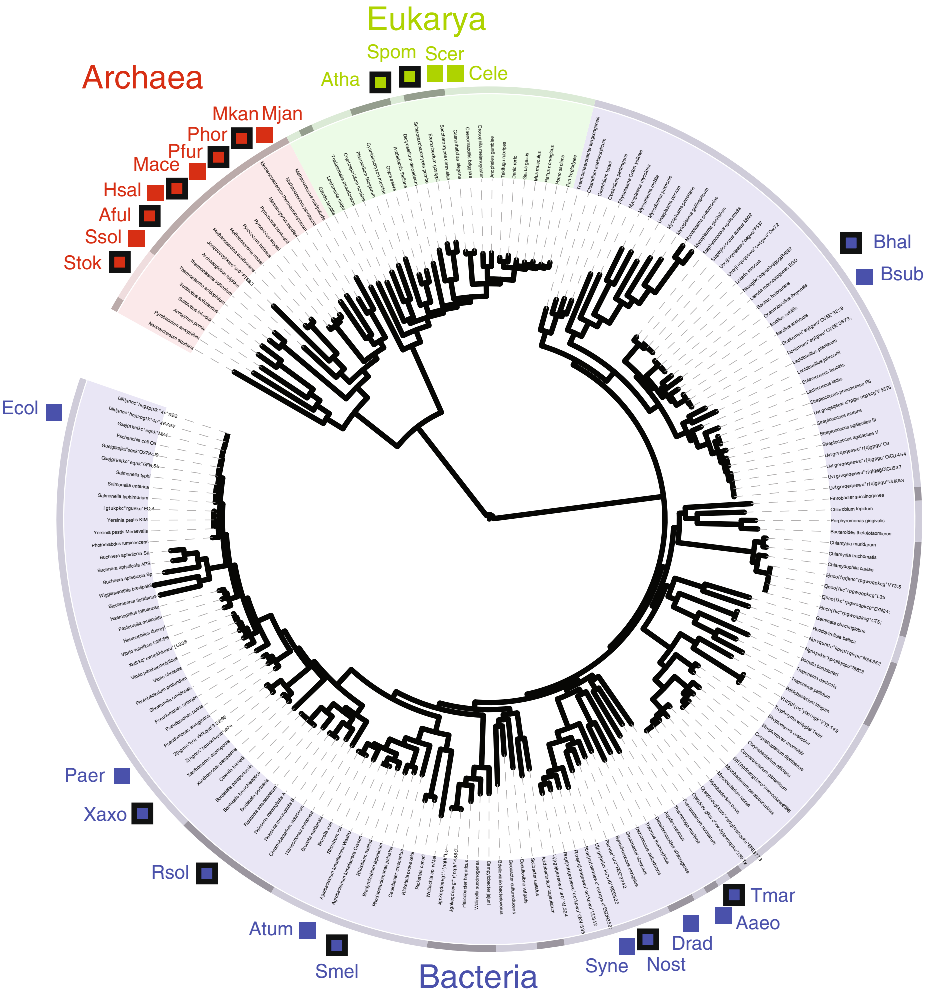

Fig. 1 Whole genome sequences used in this study. Sequences, overlaid on the tree of life, represent the Bacteria (*blue*), Archaea (*red*), and Eukarya (*green*). *Black borders* indicate sequences in the second (control) set. For abbreviations and details, see Table 1. Tree of life generated using the Interactive Tree of Life (iTOL; Letunic and Bork 2007)

Except for sequences known to encode RNAs, low-complexity sequences were identified with the DUST algorithm (Tatusov and Lipman, unpublished) using default parameters. A sequence was eliminated from further analysis if blastn alignments to a chosen subset of bacterial and archaeal genomes (*B. subtilis*, *P. aeruginosa*, *P. furiosus*, and *S. solfataricus*) did not all overlap; this subset of genomes was also used for detailed analyses of particular sequences where noted. For these alignments, WU-BLAST blastn 2.0MP 30 Jan 2004 (Gish, W., personal communication) was used with a word size of 1 (–W –1).

The tRNA identities are based on the assignments of the Genomic tRNA Database at http://rna.wustl.edu/GtRDB (Lowe and Eddy 1997). In most cases, these assignments agree

Table 1 Whole genomes in primary and control data sets

Dataset	Species	Abbreviation	ID	Domain	Notes
Primary	*Aquifex aeolicus* VF5	Aaeo	NC_000918	B	
	Agrobacterium tumefaciens str. C58 Wash. Univ.	Atum	NC_003062	B	
	Bacillus subtilis subsp. subtilis str. 168	Bsub	NC_000964	B	
	Caenorhabditis elegans	Cele	NC_003279	E	Chr I
	Deinococcus radiodurans R1	Drad	NC_001263	B	Chr I
	Escherichia coli K12 substr. MG1655	Ecol	NC_000913	B	
	Halobacterium sp. NRC-1	Hsal	NC_002607	A	
	Methanocaldococcus jannaschii DSM2661	Mjan	NC_000909	A	
	Pseudomonas aeruginosa PA01	Paer	NC_002516	B	
	Pyrococcus furiosus DSM3638	Pfur	NC_003413	A	
	Saccharomyces cerevisiae	Scer	NC_001144	E	Chr XII
	Sulfolobus solfataricus P2	Ssol	NC_002754	A	
	Synechocystis sp. PCC 6803	Syne	NC_000911	B	
Control	*Archaeoglobus fulgidus* DSM 4304	Aful	NC_000917	A	
	Arabidopsis thaliana	Atha	NC_003074	E	Chr 3
	Bacillus halodurans C-125	Bhal	NC_002570	B	
	Methanosarcina acetivorans C2A	Mace	NC_003552	A	
	Methanopyrus kandleri AV19	Mkan	NC_003551	A	
	Nostoc sp. PCC 7120	Nost	NC_003272	B	
	Pyrococcus horikoshii OT3	Phor	NC_000961	A	
	Ralstonia solanacearum GMI1000	Rsol	NC_003295	B	
	Sinorhizobium meliloti 1021	Smel	NC_003047	B	
	Schizosaccharomyces pombe	Spom	NC_003421	E	Chr III
	Sulfolobus tokodaii str. 7	Stok	NC_003106	A	
	Thermotoga maritima MSB8	Tmar	NC_000853	B	
	Xanthomonas axonopodis pv. citri str. 306	Xaxo	NC_003919	B	

Domains include Archaea (A), Eukarya (E), and Bacteria (B). When whole genomes included multiple chromosomes, a specific chromosome (Chr) is cited, corresponding to its identifier

with the NCBI *E. coli* genome annotations. However, in instances where the annotations are not consistent, the tRNA Database results were used.

Assessment of High-Scoring Genomic Regions

To assess conservation at the base pair level in detail, we performed multiple alignments, using the MAVID website (Bray and Pachter 2004), for genomic regions showing the highest level of conservation. For each of the 26 whole genomes (Table 1), we used the first annotation corresponding to each genomic region, except for Scer. Here, NCBI sequence DQ88227 was used because NC_001144 lacked specific annotations for the top-scoring genomic regions.

Results

We searched for sequence similarities in the complete genomes of 13 organisms representing disparate branches of the evolutionary tree of life. A second set of 13 genomes was used as a control to verify that the results were independent of the specific genomes selected. Each dataset included 6–7 Bacteria, 4–5 Archaea, and two Eucaryotes, and represent a range of environments and metabolic capabilities. Similar sequences and

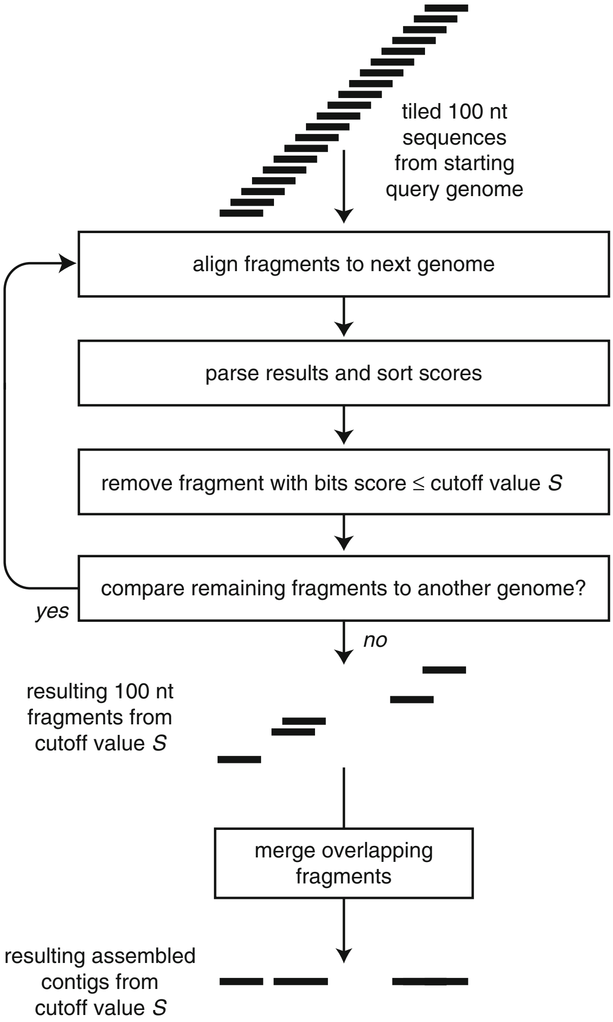
tiled 100 nt sequences from starting query genome
align fragments to next genome
parse results and sort scores
remove fragment with bits score ≤ cutoff value S
compare remaining fragments to another genome?
yes
no
resulting 100 nt fragments from cutoff value S
merge overlapping fragments
resulting assembled contigs from cutoff value S

Fig. 2 Genome filtering algorithm. The genome filtering algorithm begins with a tiled set of 100 nt fragments from the starting query genome, a set of genomes for comparison, and a score cutoff value. After each genome comparison, fragments that do not satisfy the score cutoff are removed and the process repeated with the remaining fragments and the next genome of the comparison set. After all genome comparisons are completed, overlapping 100 nt fragments in the resulting set are merged to create contigs, each representing a sequence conserved at the level of the cutoff value

rankings were obtained for the controls as for the primary dataset search with *E. coli*. Both primary and control searches identified regions of the 16S and 23S ribosomal genes, tRNAs, sequences that code for nucleotide binding domains of ATP-binding cassette (ABC) transporters, and regions of genes encoding products such as ATP synthase subunit *b*, topoisomerase, and ribosomal proteins. Thus, our findings were not a result of the order in which the genomes were searched or the genomes selected.

16S & 23S Ribosomal RNA Genes

The 16S and 23S ribosomal RNA genes scored as the most highly conserved sequences (Fig. 3a, Table 2), with the highest-scoring non-ribosomal genetic regions scoring only 37 bits (Fig. 3b, dotted line). The highest score cutoffs identified two regions of the 23S gene and one region of the 16S rRNA gene, with the 23S regions scoring slightly higher (Table 2). As the score cutoff was decreased from 60 to 38, the conserved regions grew

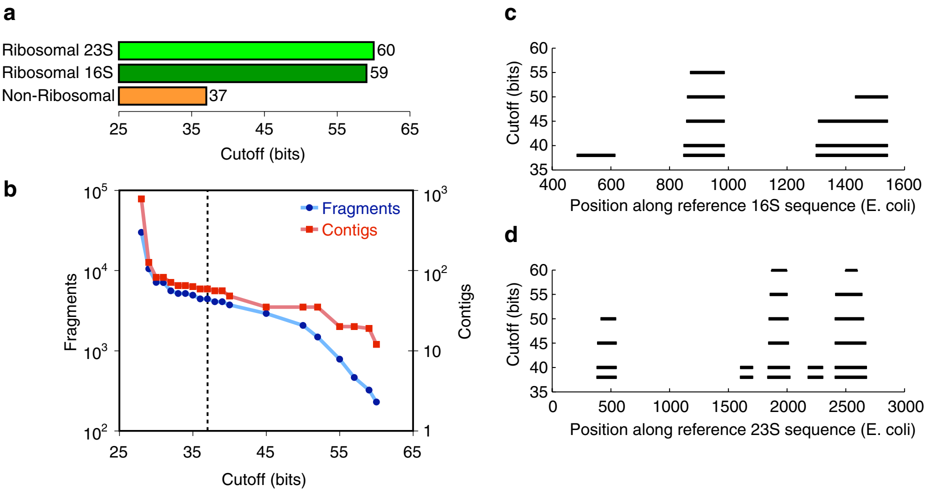

Fig. 3 Ribosomal 16S and 23S genes score highest of all gene regions with conserved sequences. **a** Maximum bit scores for gene regions in which conserved sequences appear. **b** Number of 100 bp fragments identified, and number of contiguous sequences assembled from overlapping 100 bp fragments, as a function of the score cutoff. Above 37 bits (*dotted line*), no conserved sequences were identified in non-ribosomal genes. Genes associated with conserved regions are listed in Table 2. **c** Conserved regions within ribosomal 16S gene as a function of cutoff score. **d** Conserved regions within ribosomal 23S gene. In **c** and **d**, regions shown correspond to those identified at the cutoff scores in Supplementary Table S1 for rrnH in *E. coli*

both in size and quantity to total five regions of the 23S and three regions of the 16S genes (Fig. 3c–d, Supplementary Table S1). The conserved regions identified within the 16S gene include sequences targeted by PCR primers used to amplify 16S genes from environmental samples, *e.g.* 515F, 906F, 926R, 1391R, and 1492R (Fig. 3c, Supplementary Table S1),

Table 2 Gene regions associated with conserved sequences

Cutoff	Gene regions (*E. coli* annotations)
60	rrlH, rrlG, rrlC, rrlA, rrlB, rrlE
59	rrlD, rrsH, rrsG, rrsD, rrsA, rrsB, rrsE
52	rrsD
37	**yecC**, *alaX*, *metZ/metW/metV*
35	*alaV*, *valV/valW*, *alaU*, *alaT*
34	*metY*, *argX*
32	**glnQ**, **ssuB**, **hisP**, *alaW*, *argQ*, *argZ*, **yhdZ**, *ileU*, *ileT*
31	*ileV*, *argU*, envY+, *metU/glnW*, *metT*, *glyW*, *proL*, *ileY*, *argY*, *argV*, *ileX*, *hisR*, *glyT/thrT*, *glyY*
29	*proM*, *gltT*, *thrT*, *gltV*, *gltU*, *aspT*, *trpT*, *leuT*, *gltW*, *serV*, *glyU*, *aspU*, *lysT*, *lysZ*, *tyrT*, *aspV*, *valT*, *lysW*, *lysQ*, *leuZ*, *valZ*, *lysY*, *lysV*, **ddpF**, nuoM, ygfG, +yhaO, argR, damX, selB/selA, yihI, degQ, upp+
28	See Supplementary data

Gene regions within which the conserved sequences, or contigs, were identified are listed for their corresponding highest score cutoff. The number of fragments and contigs identified as a function of cutoff score are shown in Fig. 2. ABC transporter NBD sequences are bold and sequences from tRNAs are italic. A "+" appended before or after the gene identifier denotes that the sequence identified includes significant sequence outside the 5′ or 3′ end, respectively, of the open reading frame. A "/" between gene names indicates that the sequence identified overlaps multiple neighboring open reading frames. A detailed list of the regions within the large (rrlX) and small (rrsX) ribosomal subunit rRNA genes is found in Supplementary Table S1.

though some regions targeted by commonly used bacterial-specific primers (*e.g.*, 8F or 27F) scored poorly because this search was directed toward finding sequences conserved in all three kingdoms. It is important to note that at primer-relevant length scales both the 16S and 23S genes show very high conservation (Fig. 4). Exceptionally, within 23S there exists a >60 bp region (1901–1962) that is nearly identical for all 13 bacterial genomes. The 23S gene is also nearly twice as long as the 16S gene and contains many highly variable regions

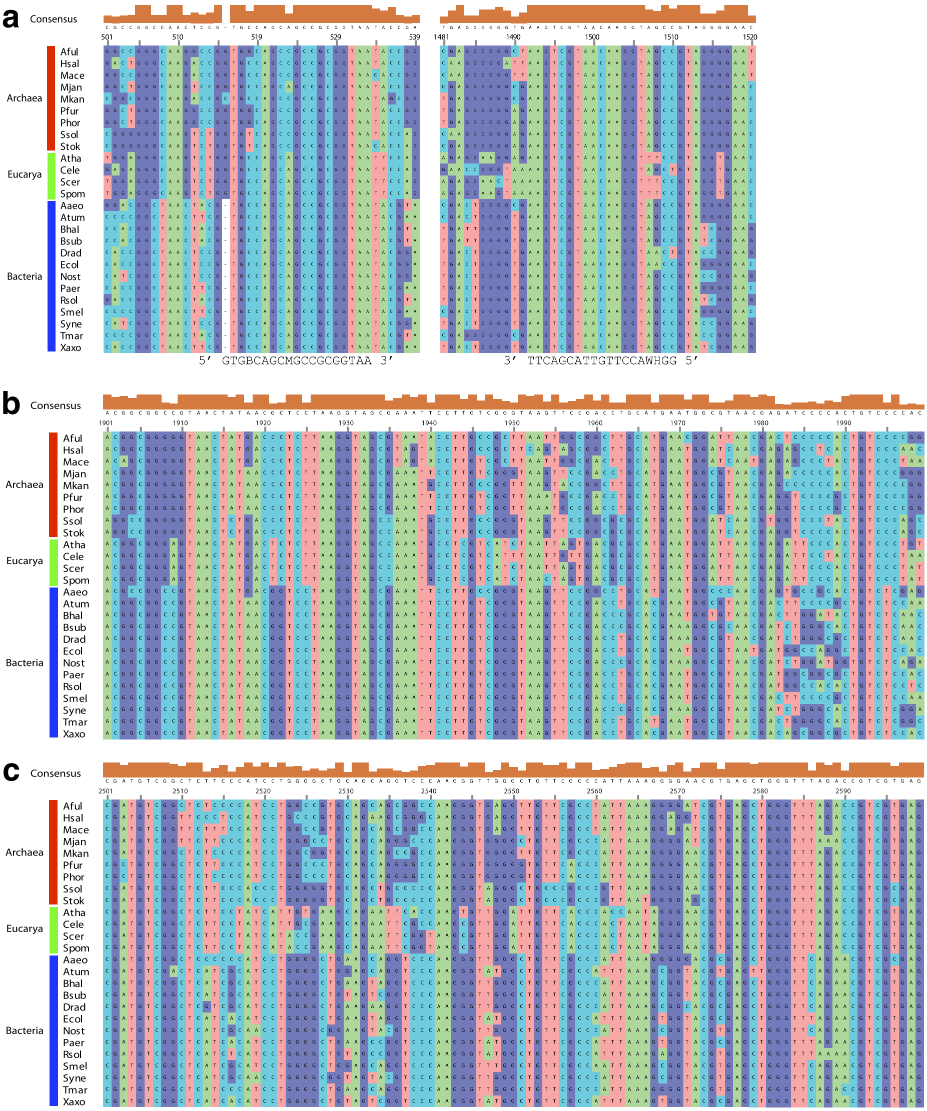

Fig. 4 Conservation within selected high-scoring regions of ribosomal genes. Multiple alignments across all 26 species for ribosomal genes: **a** 16S regions 501–539 (*left*) and 1481–1520 (*right*) show little variance across species or domain, making the primers (5′ GTGB… and 3′ TTCA…) suitable for near-universal amplification. **b** 23S region 1901–1999. **c** 23S region 2501–2599. In all panels numbering corresponds to *E. coli* reference sequence

that may be used to provide novel taxonomic information. However, on length scales of 10–20 bp, 16S regions near 515, 1391, and 1492 are as or better conserved than regions within 23S; unlike the 23S regions, which are closely spaced together, these 16S regions (e.g. those in Fig. 4) can provide amplicons up to ˜1 kb in length, suitable for a reasonably detailed and accurate phylogenetic analysis.

The sequences identified within the 16S and 23S genes are known to mediate several important interactions with transfer RNAs in the ribosome A, P, and E sites: with elongation factors involved in translation; and between regions of the 30S and 50S subunits (Green and Noller 1997; Yusupov et al. 2001). These important intramolecular interactions greatly reduce the rate of evolutionary change within these rRNA sequences and likely underlie their high degree of conservation. This search validated the use of the 16S and 23S genes as conserved elements by which to capture divergent genomes, and also suggests that the 23S gene may be useful in designing PCR primers to capture more divergent organisms.

tRNAs and ABC Transporter Binding Domains

Two classes of sequences were identified to be the next most-conserved sequences after the rRNA sequences (Table 2). The first group comprised sequences with homology to transfer RNAs (tRNA), and the members of the second group were similar to sequences that encode nucleotide-binding domains (NBD) of ATP-binding cassette (ABC) transporters (Linton and Higgins 1998; Schneider and Hunke 1998; Kerr 2002).

Of these two groups, the tRNA group was larger and generally the higher scoring of the two (Table 1). The search identified a total of 46 tRNA-associated sequences, each of which spanned one or more tRNA encoding regions (Table 1 and Supplementary data). With the sole exception of cysteine, all amino acids were represented by at least one tRNA sequence. At the high end (30–37) of the range of cutoff values that identified tRNA sequences, the tRNAs identified are for the amino acids Ala, Arg, Gln, Gly, His, Ile, Met, Pro, Thr, and Val, with the tRNAs for Ala, Met, and Arg scoring the highest. Some of these tRNAs are charged by type I tRNA synthetases and some by type II tRNA synthetases; thus, there does not appear to be a trend for one type to be more conserved than the other. A well-conserved sequence from the 3′ UTR of the *E. coli envY* gene scored within this range due to its sequence similarity to tRNA sequences.

Within the second group are sequences that are known or predicted to encode a conserved region of ABC transporter NBDs. The most conserved of these sequences were within the homologous genes *glnQ*, *ssuB*, *yecC*, *hisP*, and *yhdZ*, and one lower-scoring ABC transporter NBD sequence was identified within *ddpF*, also known as *yddO*. The most-conserved sequence of this group was the *yecC* sequence, which scored as high as the highest-ranking tRNA sequences. The top matches of the five higher-scoring sequences (glnQ, ssuB, yecC, hisP, and yhdZ) to *B. subtilis*, *P. aeruginosa*, *P. furiosus*, and *S. solfataricus* were 19 ABC transporter NBDs and one RecA-like NBD.

The regions most similar to *yecC* in the microbial and *C. elegans* genomes were selected using blastn and aligned with clustalw to assess the nucleotide sequence conservation of the regions (Fig. 5). The alignments demonstrated a high degree of sequence identity at the nucleotide level and revealed a pattern of conservation suggesting selective pressure is exerted on the amino acid sequence (Fig. 5). In the aligned genomes, approximately 60% (65/108 nucleotides) of the identified NBD region demonstrated a high level of nucleotide sequence conservation. The 'ABC signature' and 'Walker B' regions were enriched with conserved sites as was expected for these conserved motifs (Fig. 5). Conservation was typically present in the first two codon positions and was less common in the third position;

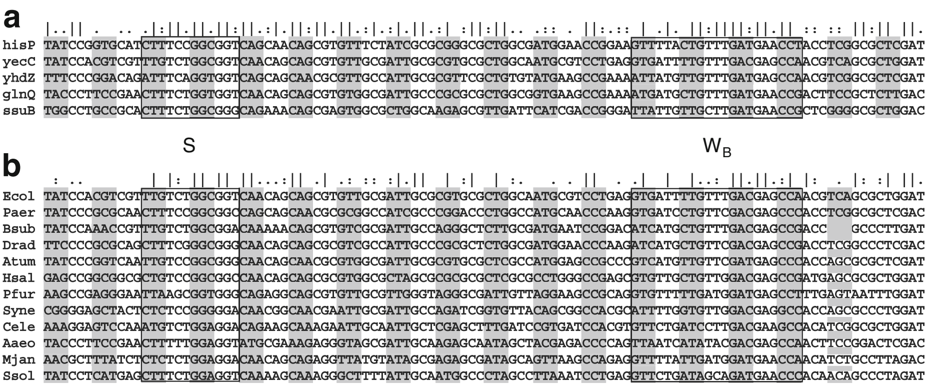

Fig. 5 Nucleotide sequence conservation within known or putative ABC transporter nucleotide binding domains (NBD). The nucleotide sequences of the five *E. coli* ABC transporter NBD regions identified in the search (**a**) and of the *yecC* best matches from 12 genomes (**b**) are shown as aligned by clustalw. Above the alignments, columns with 90%, 80%, and 60% conservation are marked with |, :, and •, respectively. Boxes denote the regions encoding the conserved ABC signature (S) and Walker B (WB) motifs

this observation suggested the high level of nucleotide conservation was due to selection at the level of the protein sequence, which was then reflected in unusually high conservation of the genome sequence. When translated, the NBD sequences identified here do not possess large numbers of low degeneracy amino acids, so it is unlikely that the detected nucleotide sequence conservation has resulted from simple sequence constraints imposed by selection for protein function plus low degeneracy.

ABC transporters have been found in all known organisms and constitute one of the most abundant and phylogenetically widespread protein superfamilies. Phylogenetic analyses of ABC transporter sequences have suggested that they share an evolutionarily ancient origin and began to specialize before the division of Archaea, Bacteria, and Eukarya (Saurin et al. 1999). Although their widespread and ancient nature can obscure identifying orthologs and can complicate phylogenetic arguments regarding sequence divergence, comparative analyses of organisms may profit from using ABC transporter NBD sequences because of their ubiquitous presence within genomes. Moreover, different suites of ABC transporters are found in different genomes, and are induced under particular growth conditions to mediate the transport of metabolites, and may thus reflect the environment in which organisms evolved. Finally, although the use of ABC transporter NBDs has been relatively unexplored as a means to identify organisms, PCR targeting an ABC transporter NBD sequence in the region identified by this search has been used to clone a putative macrolide drug transporter from *Saccaropolyspora erythraea*, an erythromycin-producing organism (O'Neill et al. 1995). A more detailed analysis of ABC transporter sequences may provide helpful tools for molecular identification and phylogenetic characterization of organisms, especially microbes.

Other Moderately Conserved Sequences

Besides the high-scoring tRNA and ABC transporter NBD sequences, several other sequences scored moderately high in this search. These sequences were annotated to possess various functionalities. However, in most cases the resulting alignments were to sequences that encode products that interact with nucleotides, nucleic acids, or related molecules (Table 3).

Table 3 Moderately conserved non-RNA sequences in two Bacteria and two Archaea

Gene	E. coli	B. subtilis	P. aeruginosa	P. furiosus	S. solfataricus
ddpF	ABC NBD (peptide)	ABC NBD (peptide)	Probable ABC NBD (multidrug/peptide/lipid)	(cobalt)	ABC NBD (phosphate)
nuoM	NADH dehydrogenase chain M	Related to sulfur transferases	NADH dehydrogenase chain M	Hypothetical membrane protein	GroEL/Hsp60
ygfG	Putative enoyl-CoA hydratase/isomerase	Unknown DNA repair ATPase	Conserved membrane virulence protein	DEXX-box ATPase	Hypothetical
degQ	Serine protease	Unknown integrase	Multidrug efflux	Putative ATP-dependent RNA helicase	Hypothetical
selB/selA	Selenocysteine elongation factor and synthase	EF-Tu	Selenocysteine elongation factor	eIF-2 gamma subunit GTPase	Hypothetical
yihI	Hypothetical ribonucleoprotein	Predicted HTH transcriptional regulator	Probable 2-component His kinase	Proteasome subunit alpha	Putative transcription factor

The *E. coli* genomic regions and known or putative functions are listed in the first two columns, and the known or putative functions of best matches in two Bacteria (*Bacillus subtilis* and *Pseudomonas aeruginosa*) and two Archaea (*Pyrococcus furiosus* and *Sulfolobus solfataricus*) are listed in the remaining columns.

Other Low-Scoring Sequences

Seven hundred and eighty two sequences were identified at the lowest score cutoff used (28, see Supplementary data). After accounting for RNA sequences and sequences identified at higher score cutoffs, 466 nucleotide sequences remained. Among these sequences were many phylogenetic markers used previously for the identification and classification of organisms, including topoisomerase (*gyrB*), the ATP synthetase subunit beta (*atpD*), a heat shock protein (*hsp70*), elongation factor Tu (*tufA*, *tufB*), and two sigma factors (*rpoD* and *rpoN*) (Ward-Rainey et al. 1997; Gupta 2004; Picard et al. 2004).

Because of the large number of sequences present within this group, they were analyzed for general composition and trends. First, sequences were grouped into clusters based on sequence similarity to identify families of closely related nucleotide sequences. This treatment resulted in two large groups of sequences, all with >90% identity to one another: one group contained ten transposase-encoding sequences and the second group contained six insertion sequences. Two other nearly identical (>90% identity) pairs of sequences were identified: one duplicated sequence (*tufA* and *tufB*) encodes protein chain elongation factor EF-Tu, and the other is an intergenic sequence found both between the genes *ybcW* and *nohB* and between *nohA* and *ydfO*. Four other less-similar groups of seven, four, four, and three members consisted of sequences known or predicted to encode ABC transporter NBDs.

Based on their known or predicted products in *E. coli*, the 466 sequences were partitioned into functional categories (Supplementary Table S2) defined by the Clusters of Orthologous Genes (COG) classification at NCBI (Tatusov et al. 2003). Of these sequences, 42 were wholly or partially in intergenic regions. Most of these intergenic sequences overlapped the ends of known or predicted coding regions in *E. coli* and may thus be untranslated regulatory regions of genes. These results demonstrated an enrichment of sequences encoding products involved in translation; replication, recombination, and repair; cell cycle control; defense mechanisms; signal transduction; posttranslational modification, protein turnover, and chaperones; and transport of various substances. Interestingly, there was also considerable enrichment for sequences of unknown function.

Discussion

The genome filtering search described here has identified widely conserved sequences that may be valuable for microbial identification and classification. The results support the use of 16S sequences as the best life detection sensors, although they also suggest that further development of 23S methods is warranted. Because 23S genes are approximately twice as long as 16S genes and contain both highly variable and conserved regions, the information they contain will be valuable for future research into life detection and microbial diversity. The results presented here serve as a guide for selecting 16S and 23S regions to use for such studies, depending on the level of conservation required. In addition, our results indicate that the large family of tRNA sequences is widely conserved. Although their small length and high conservation may not be useful for providing detailed phylogenetic resolution, they may be valuable for detection or for characterizing gene sequences that flank them. For instance, sequences flanking tRNA may be used to differentiate between strains with nearly identical 16S sequences or as sequence anchors to characterize DNA cloned directly from environmental samples. Importantly, the large conserved superfamily of ABC transporter NBD sequences provides a relatively unexplored space of homology that may not only provide a means to identify organisms and classify them, but may also provide information about their environments and metabolic functionalities.

Because this search demanded the presence of particular sequences in all 13 genomes tested, some near-universal sequences may have been missed. Additional useful sequences may be identified by computational approaches to find sequences present in particular subsets of taxa. For example, the search could be reconfigured to collect those sequences present in 90% of the genomes surveyed, or, by careful selection of the genomes searched, sequences common to a particular taxonomic group (e.g., all archaea or the alpha-proteobacteria) or organisms sharing particular metabolic capabilities (e.g., photosynthesis or methanotrophy).

Inheritance from a common ancestor is believed to be responsible for the presence of conserved sequences in genomes when horizontal gene transfer is not thought to play a role. In this investigation, the phylogenetic diversity of the 13 selected genomes suggests that the identified sequences were present in an ancient ancestor of all life. Thus, the conservation of functional RNA molecules (tRNA and rRNA) is not surprising considering they are ancient molecules with tight functional constraints on their evolution (Eigen et al. 1989). However, some intriguing points arise from trends observed in the other results of this search. First, the prevalence of sequences from intergenic regions or with unknown functions confirms that much is to be learned regarding the content of genomes, what functions are important or necessary for life, and how organisms are related. Second, there is an abundance of sequences encoding proteins that interact with DNA, RNA, nucleotides or structurally related molecules; the most notable example is the group of ABC transporters, some of which are as highly conserved as tRNA sequences, and have previously been identified as highly conserved (Delaye et al. 2005). This finding most likely results from nucleotides (i.e., ATP, GTP) and nucleotide-like cofactors (i.e., coenzyme A, NAD, NADH, FAD, SAM) holding ancient and central roles in biological reactions, thus contributing to their interaction with a multitude of proteins. Conservation of NBDs may be overestimated due to the large number of NBDs per genome; however, the ubiquity of these sequences still makes them potential life detection targets.

If any interplanetary transfer of life occurred after the evolution of conserved DNA segments, our analysis suggests that PCR targeting the 16S and 23S ribosomal genes and other conserved sequences could detect this life with extreme sensitivity, and could be readily incorporated into an instrument suitable for use, for example, on Mars. A typical small thermal cycler not optimized for space flight weighs 5–10 Kg and uses 100 W of power. Only tiny amounts of energy are needed to cyclically heat and cool a typical PCR reaction and to detect the product of that amplification. The fluid handling components to process the soil samples and move them to the PCR module could weigh less than 500 g, and an entire instrument could weigh ˜2 kg. A reverse transcriptase step could facilitate capture of RNA world organisms, and both random and specific amplification based on universal primers could be used to amplify DNA. Sequencing methods suitable for miniaturization can currently yield read-lengths of hundreds of bases (Shendure et al. 2004; Leamon et al. 2007); a positive PCR result could therefore be sequenced in-situ to rule-out contamination and directly assess sequence divergence from all known Earth sequences.

The Power of PCR and the DNA Sequence Biosignature

The DNA sequence that is amplified between universally conserved primers is key to classification of any PCR positives. For example, the DNA sequence between the universal 16S gene primers contains so much information that organisms detected only by their 16S gene sequences are routinely classified based only on that information. This DNA sequence carries information about the organism from which the ribosomal RNA gene is derived, and can allow a new organism to be fit into the tree of life. Thus the detected product is a unique biosignature.

Any Martian biota has been geographically as well as genetically more isolated from the Earth biota since the dramatic decline in meteoritic exchange 3.5 billion years ago, and thus is likely to be more deeply branching in the phylogenetic tree of 16S or 23S RNA sequences. Even though there may continue to be exchange, Mars is more likely to be more isolated than any possible Earth ecosystem. Therefore, we expect that any 16S or 23S gene detected on Mars will branch deeply compared to the phylogeny of 16S or 23S sequences of known Earth life. Detection of such a deeply branching ribosomal RNA gene would be a compelling signature of life on Mars. Even if highly conserved genes like 16S and 23S evolved after meteoritic transfer of life, PCR can still be readily used to detect both DNA- and RNA-based life using non-specific amplification (e.g. isothermal amplification using phage ϕ29 polymerase and random hexamers) and an initial reverse-transcriptase step, respectively.

It is also possible that any ancestrally related life on Mars is now extinct. We do not expect to detect microbial fossils using PCR; it has been used to detect DNA in samples hundreds of thousands of years old but not millions of years old (though viable halophilic bacteria have been reported in 250 million year old salt crystals (Vreeland et al. 2000), these results have been heavily contested (Graur and Pupko 2001)). Even if life did transfer between Mars and Earth four billion years ago and thrive in the early Mars environment, it may now only thrive in very particular refugia, for example deep in the crust where the temperature rises above that of the frozen surface or at particular volcanic thermal vents. There may be regions on Mars where liquid water is in ample supply, for example near the polar ice caps (Zuber et al. 1998). In addition, the D/H ratio of water in Martian meteorites suggests a much larger reservoir of water in the crust that is not in equilibrium with the atmosphere (Donahue 1995; Leshin 2000). This water is predicted to be liquid a few kilometers into the crust, where temperatures rise above surface temperature. In addition, there is evidence for recent Martian volcanism, suggesting sources of temperature gradients and fluid flows just below the Martian surface near these sites (Malin and Edgett 2000) with local hydrothermal systems (Farmer 1996). However, the negative (albeit insensitive) Viking seismic data and the build up of massive volcanic cones do not favor active tectonics, a key element for hydrothermal vents and the biology that thrives in their energy and nutrient flows (Anderson et al. 1977). PCR, unlike other detection strategies, could detect the dispersal of a few microbes or microbial genomes from remote refugia, such as might be dispersed by global dust storms.

Competing life detection technologies detect biosignatures, such as lipids, proteins, or nucleic acids, with a sensitivity of a few thousand monomer subunits of polymer per gram of input soil. However, there are abiogenic mechanisms that explain the presence of amino acids, including meteoritic deposition (Engel and Macko 1997). Proposals to search for enantiomeric organic molecules as evidence of life are also haunted by possible abiological sources of asymmetry. It is doubtful that isotopic fractionation observations could be so sensitive, though they do have the advantage of being able to observe fossil biosignatures.

No other current life detection technology rivals PCR for sensitivity. PCR can detect a single DNA strand in a pure preparation (Zhang et al. 1992). Rich environments on Earth such as soil in temperate zones have 10^8–10^9 microbes per gram of soil. The energy flux at the surface would allow the sorts of densities found on Earth, but the failure to detect organic molecules on the Martian surface suggests low microbial densities (Biemann et al. 1977). Thus, the central challenge of a PCR detector is to achieve a suitably low false positive rate; this can be achieved through minimization of forward contamination and by the ability to distinguish contaminants from a true signal of life on Mars. Modern microfludic PCR techniques obtain single molecule sensitivity with low false positive and false negative rates (Warren et al. 2006). Moreover, sequence analysis and other approaches allow for immediate identification of chimeric products or Earth-derived contaminants. Current databases (>180,000 sequences of length > 1,250 nt (DeSantis et al. 2006)) can exactly identify any particular likely contaminant sequence, and high-throughput DNA sequencing and metagenomic analysis may expand known rDNA sequences by a million-fold by the time of landing of a PCR instrument on Mars. In addition, primers designed to target a putative Martian sequence can be used to prospect for similar organisms on Earth, allowing terrestrial confirmation that the DNA sequence detected on Mars is not terrestrial contamination. Given these advantages and the significant possibility of meteoritic microbial transfer, DNA detection is a critical part of any comprehensive life detection strategy.

Acknowledgements TAI thanks Professor Jo Handelsman for support during the preparation of this manuscript, Ilya Ruvinsky and members of the Ruvkun laboratory for helpful comments and review of the manuscript, Lance Davidow for assistance with computing resources, and Qinghu Ren at TIGR for helpful discussions regarding ABC transporters. This work was funded by National Aeronautics and Space Administration grant NNG05GK27G.

References

Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anderson DL, Miller WF, Latham GV et al (1977) Seismology on Mars. J Geophys Res 82:4524–4546

Barns SM, Fundyga RE, Jeffries MW et al (1994) Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci USA 91:1609–1613

Biemann K, Oro J, Toulmin P et al (1977) The search for organic substances and inorganic volatile compounds in the surface of Mars. J Geophys Res (Scientific results of the Viking project) 82:4641–4658

Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699

Delaye L, Becerra A, Lazcano A (2005) The last common ancestor: what's in a name? Orig life evol biosph 35:537–554

DeSantis TZ, Hugenholtz P, Larsen N et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072

Donahue TM (1995) Evolution of water reservoirs on Mars from D/H ratios in the atmosphere and crust. Nature 374:432–434

Eigen M, Lindemann BF, Tietze M et al (1989) How old is the genetic code? statistical geometry of tRNA provides an answer. Science 244:673–679

Engel MH, Macko SA (1997) Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. Nature 389:265–268

Farmer JD (1996) Hydrothermal systems on Mars: An assessment of present evidence. In: Bock GR, Goode JA (eds) Evolution of hydrothermal ecosystems on Earth (and Mars?). Wiley & Sons, Chichester, pp 273–295

Fox GE, Stackebrandt E, Hespell RB et al (1980) The phylogeny of prokaryotes. Science 209:457–463

Friedmann EI, Ocampo-Friedmann R (1984) The Antarctic cryptoendolithic ecosystem: relevance to exobiology. Orig Life 14:771–776

Fry NK, Fredrickson JK, Fishbain S et al (1997) Population structure of microbial communities associated with two deep, anaerobic, alkaline aquifers. Appl Environ Microbiol 63:1498–1504

Giovannoni SJ, Britschgi TB, Moyer CL et al (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345(6270):60–63

Gladman BJ, Burns JA (1996) Mars meteorite transfer: simulation. Science 274:161–165

Gladman BJ, Burns JA, Duncan M et al (1996) The exchange of impact ejecta between terrestrial planets. Science 271:1387–1392

Graur D, Pupko T (2001) The Permian bacterium that isn't. Mol Biol Evol 18(6):1143–1146

Green R, Noller HF (1997) Ribosomes and translation. Annu Rev Biochem 66:679–716

Gupta RS (2004) The phylogeny and signature sequences characteristics of *Fibrobacteres*, *Chlorobi*, and *Bacteroidetes*. Crit Rev Microbiol 30:123–143

Harris JK, Kelley ST, Spiegelman GB et al (2003) The genetic core of the universal ancestor. Genome Res 13:407–412

Horneck G, Bucker H, Reitz G (1994) Long-term survival of bacterial spores in space. Adv Space Res 14:41–45

Horneck G, Stöffler D, Ott S et al (2008) Microbial rock inhabitants survive hypervelocity impacts on Mars-like host planets: first phase of lithopanspermia experimentally tested. Astrobiology 8(1):17–44

Isenbarger TA, Finney M, Ríos-Velázquez C et al (2008) Miniprimer PCR, a new lens for viewing the microbial world. Appl Environ Microbiol 74(3):840–849

Iyer LM, Leipe DD, Koonin EV et al (2004) Evolutionary history and higher order classification of AAA+ ATPases. J Struct Biol 146:11–31

Kerr ID (2002) Structure and association of ATP-binding cassette transporter nucleotide-binding domains. Biochim Biophys Acta 1561:47–64

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1:127–136

Leamon J, Braverman M, Rothberg J (2007) High-throughput, massively parallel DNA sequencing technology for the era of personalized medicine. Gene Therapy and Regulation 3:15–31

Leshin LA (2000) Implications for the origin and history of water on Mars from analyses of martian meteorite QUE94201. Geophys Res Lett 27:2017–2020

Letunic I, Bork P (2007) Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23(1):127–128

Linton KJ, Higgins CF (1998) The *Escherichia coli* ATP-binding cassette (ABC) proteins. Mol Microbiol 28:5–13

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Malin MC, Edgett KS (2000) Evidence for recent groundwater seepage and surface runoff on Mars. Science 288:2330–2335

Melosh HJ (1988) The rocky road to panspermia. Nature 332:687–688

Melosh HJ (1993) Blasting rocks off planets. Nature 363:498

Mileikowsky C, Cucinotta FA, Wilson JW et al (2000) Natural transfer of viable microbes in space. Icarus 145:391–427

O'Neill MP, Eady EA, Radford A et al (1995) The use of PCR to isolate a putative ABC transporter from *Saccharopolyspora erythraea*. FEMS Microbiol Lett 131:189–195

Pedersen K (2000) Exploration of deep intraterrestrial microbial life: current perspectives. FEMS Microbiol Lett 185:9–16

Picard FJ, Ke D, Boudreau DK et al (2004) Use of tuf sequences for genus-specific PCR detection and phylogenetic analysis of 28 streptococcal species. J Clin Microbiol 42:3686–3695

Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394

Santos SR, Ochman H (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. Environ Microbiol 6:754–759

Saurin W, Hofnung M, Dassa E (1999) Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. J Mol Evol 48:22–41

Schneider E, Hunke S (1998) ATP-binding-cassette (ABC) transport systems: functional and structural aspects of the ATP-hydrolyzing subunits/domains. FEMS Microbiol Rev 22:1–20

Shendure J, Mitra RD, Varma C et al (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5:335–44

Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics. doi: 10.1186/1471-2105-4-41

Vreeland RH, Rosenzweig WD, Powers DW (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. Nature 407:897–900

Wall L, Schwartz RL, Christiansen T (1996) Programming perl. a nutshell handbook. O'Reilly Inc., Sebastopol, California

Ward-Rainey N, Rainey FA, Stackebrandt E (1997) The presence of a dnaK (HSP70) multigene family in members of the orders Planctomycetales and Verrucomicrobiales. J Bacteriol 179:6360–6366

Warren L, Bryder D, Weissman IL et al (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc Natl Acad Sci USA 103:17807–17812

Weiss BP, Kirschvink JL, Baudenbacher FJ et al (2000) A low temperature transfer of ALH84001 from Mars to Earth. Science 290:791–795

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Yusupov MM, Yusupova GZ, Baucom A et al (2001) Crystal structure of the ribosome at 5.5 Å resolution. Science 292:883–896

Zhang L, Cui X, Schmitt K et al (1992) Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci USA 89:5847–5851

Zuber MT, Smith DE, Solomon SC et al (1998) Observations of the north polar region of Mars from the Mars orbiter laser altimeter. Science 282:2053–2060