NUCLEIC-ACID SEQUENCING FOR LIFE DETECTION AND CHARACTERIZATION. C. E. Carr[1*], C. S. Lui[1], H. Rowedder[2-3], M. T. Zuber[1] and G. Ruvkun[2-3], [1]Massachusetts Institute of Technology, Department of Earth, Atmospheric and Planetary Sciences, [2]Massachusetts General Hospital, Department of Molecular Biology, [3]Harvard Medical School, Department of Genetics. [*]Correspondence: chrisc@mit.edu

Introduction: Life on Mars, if it exists, may share a common ancestry with life on Earth due to meteoritic transfer of microbes between the planets [1, 2]. Recent discoveries of nucleic acids or their precursors within meteorites [3, 4] and in interstellar space [5] could steer the development of life towards these biomolecules; thus, one may want to search for RNA- or DNA-based life in habitable zones even outside the context of meteoritic exchange, such as in the probable liquid water oceans on Europa and Enceladus [6, 7]. We are building an instrument, the Search for Extraterrestrial Genomes (SETG)[8], to search for life via nucleic-acid sequencing in-situ on Mars.

Here we first discuss the detection limits of nucleic acid sequencing and the benefits of a metagenomic sequencing approach. We also propose a specific technology for in-situ sequencing that is small, robust, relies on hydrogen ion sensing instead of optics, and allows massively parallel sequencing. Initial sequencing results suggest it may be possible to generate enough reads to determine whether Earth and Mars organisms have a common ancestry and whether this split occurred before fixation of the genetic code, and to use any similarity in derived protein sequences to study how Mars organisms make their living. Finally, we discuss the relevance of in-situ sequencing to missions like Mars Sample Return (MSR).

Detection approach: A typical nucleic acid detection strategy involves extraction of nucleic acids while rejecting contaminants, followed by amplification via polymerase chain reaction (PCR), which amplifies DNA between two known regions (e.g. between highly conserved regions within the ribosomal 16S gene). Sequencing provides information on the (often unknown) region between the known regions. Because PCR and related amplification approaches such as isothermal whole genome amplification [9] can achieve single molecule detection, the detection limit is typically determined by limited yield during nucleotide extraction, with inhibitors playing a strong role in particularly vexing samples. Few studies have characterized absolute yield of DNA extraction, but the physical lysis (e.g. bead beating) required to break tough cell membranes (spores) achieves higher yields at the cost of increased DNA fragmentation.

Detection limits for specific gene targets: One experiment [10] found a 37% yield by bead-beating of filamentous fungi-spiked soils with median DNA size 0.5 /* 2 kb (e.g. log(size) has standard deviation of ~log(2)). Here, a 1kb region of the ribosomal 16S gene, assuming one copy per 2 megabase (Mb) genome, would be intact ~13% of the time (based on 1000 in-silico trials of assigning a target region to a random genome location, randomly sampling fragment lengths along that genome, checking for breaks within the target region). Thus, one viable DNA molecule might be expected for every ~20 cells at 37% yield. For fragment sizes much greater than the target region, the ideal detection limit is approximately equal to the yield (e.g. 10 /* 2 kb gives 93% intact, with estimated detection limit of 34% for a yield of 37%). This size distribution is typical of kits for isolation of DNA from soil, e.g. MoBio PowerSoil; here absolute yield dominates fragmentation in contrast to degraded (ancient DNA) samples, which can have median sizes of 100 bases or below. In lean samples, yields typically drop precipitously; however, an electrophoresis-related approach called synchronous coefficient of drag alteration (SCODA)[11] maintains >60% yield down to zeptomolar concentrations, while achieving ~10^3 better contaminant rejection than other approaches. Cartridge based detection systems manage to achieve detection limits down to 10 CFUs in some cases [12, 13]. Balancing quality and yield, a detection limit of perhaps 10^2 viable cells may be feasible in-situ. A typical *E. coli* cell weighs 1 pg; at a density of 10^2 cells/gram of soil, the cell mass is 100 parts per trillion (ppt). If 30% of the cell mass is organic material, an equivalent organics detector would need 30 ppt sensitivity.

Metagenomic approaches: Metagenomic sequencing [14], or sequencing any DNA molecule, should yield higher sensitivity that specific target-based approaches: First, metagenomic approaches do not depend upon extant organisms sharing specific conserved regions (like those within the ribosome). Second, highly sheared DNA can be utilized, allowing for sample preparation approaches that emphasize yield and achieve even better sensitivity than a specific gene target approach. An additional benefit is that metagenomics permits broad functional characterization of a microbial community through comparative sequence analysis in nucleotide and protein space. However, this analysis depends upon massively parallel sequencing. Specific informational sequences may be rare; for example, ribosomal sequences could be expected to represent 0.1% of metagenomic sequences.

RNA-seq: RNA can be reverse-transcribed to DNA and then sequenced. While RNA is easily degraded and thus not expected to be found except in viable organisms, RNA-seq can enhance the sensitivity of detecting ribosomal sequences; rRNA may represent up to 95% of total RNA in bacteria (e.g. ~10000 ribosomes in growing *E. coli*). rRNA count can be substantial even in slow growing organisms: The archaeal Richmond Mine acidophilic nanoorganisms (ARMAN) have ~1 Mb genomes and ~92 ribosomes in a 0.03 um^3 cell [15]. Thus, by performing metagenomic sequencing of RNA we can directly target ribosomal sequences, measure gene expression of DNA-based organisms, and detect a possible RNA world.

In-situ sequencing: Few high-throughput sequencing approaches [16] are compatible with in-situ sequencing due to their size and complexity. One exception is the Personal Genome Machine (PGM) from Ion Torrent [17], for which the key element is a 2.5 cm x 2.5 cm sequencing chip. Here, adaptors are ligated to fragmented DNA to produce library molecules, which are emulsion-PCR amplified at the single molecule level on beads. These beads are then loaded into the chip to achieve an ideal one bead per well. When nucleotides flow over a well and are incorporated, hydrogen ions are released, which are detected by an ISFET as a pulse of voltage, eliminating the need for optical detection. We sequenced *E. coli* DH10B (**Fig. 1**, 314 chip, 1.2 million wells), obtaining 52 Mb of data, over 40 Mb at 1% error. After ~30 runs to date, a typical good run gives 30-50 Mb of data (our record is over 80 Mb). Future sequencing chips with 10x more wells, combined with read lengths of 200-400 bp, may yield >1Gb/run. Thus, *if a single organism with moderate genome size dominated an environment on Mars, it may be possible to sequence its entire genome in-situ*. Follow up analysis may include nucleotide comparisons (placing an organism into the tree of life using ribosomal sequences) as well as searches for similarity in protein sequence space through translation of sequence reads. Even if Mars organisms have a different genetic code it may be possible to identify elements of this code through statistical analysis.

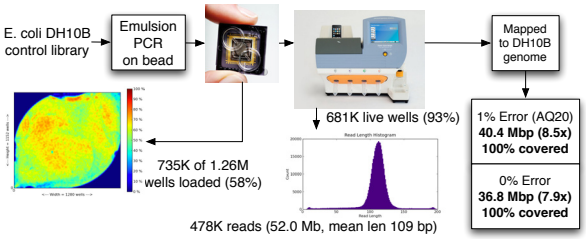

Fig. 1 Sequencing *E. coli* on the Ion Torrent PGM.

Relevance to MSR: Advances in sequencing may drive the specific approach used for future returned samples (for example, nanopore sequencing may enable robust single-molecule sequencing). However, there are several reasons to consider in-situ nucleic acid sequencing as part of MSR: First, a positive result would provide a strong incentive to complete the mission and return samples to Earth. Second, comparative sequence analysis of putative Martian organisms may help evaluate potential toxicity (or lack thereof). Third, such analysis could be used to prioritize what samples to return (with proper allocation of samples towards non-biological goals as well). Waiting to search for nucleic acids may also result in degradation due to exogenous or endogenous processes (e.g. cell lysis followed by degradation of RNA). Samples stored for years on the surface of Mars will receive large proton, heavy ion, and neutron doses; limited exposures conducted for SETG suggest DNA may survive such conditions for at least a couple years (data not shown).

Conclusions: Life detection through sequencing provides a highly sensitive and specific approach despite some obvious limits, e.g. inability to detect non-nucleotide based life or nucleotide-based life that uses incompatible nucleobases. It is now technologically feasible to pursue massively parallel sequencing in-situ, which could enhance the value of sample return missions. Given the possibility of shared ancestry between life on Earth and Mars, and the potential for RNA or DNA-based life elsewhere, searching for life *as we know it* is a critical part of any comprehensive life detection approach.

References: [1] Gladman B. J. and J. A. Burns (1996) Science, 274(5285), 161b-165. [2] Mileikowsky C., et al. (2000) Icarus, 145(2), 391-427. [3] Martins Z., et al. (2008) Earth Plan Sci Let, 270(1-2), 130-136. [4] Callahan M. P., et al. (2011) PNAS. [5] Hollis J., et al. (2000) Astrophys J Let, 540(2), L107-L110. [6] Carr M. H., et al. (1998) Nature, 391(6665), 363-365. [7] Postberg F., et al. (2011) Nature, 474(7353), 620-622. [8] Lui C., et al. *IEEE Aerospace Conference*. 2011. p. 1-12. [9] Asiello P. J. and A. J. Baeumner (2011) Lab Chip, 11(8), 1420-1430. [10] Kabir S., et al. (2003) J Biosci Bioeng, 96(4), 337-43. [11] Broemeling D., et al. (2008) JALA, 13(1), 40-48. [12] Lutz S., et al. (2010) Lab Chip, 10(7), 887-93. [13] Mahalanabis M., et al. (2010) Biomed Microdevices, 12(2), 353-9. [14] Wooley J. C., et al. (2010) PLoS Comp Bio, 6(2), e1000667. [15] Comolli L. R., et al. (2009) ISME journal, 3(2), 159-167. [16] Metzker M. L. (2010) Nat Rev Genet, 11(1), 31-46. [17] Rothberg J. M., et al. (2011) Nature, 475(7356), 348-352.