Approaches for Identifying Subsurface-Specific Taxa in Molecular Sequence Data. C. N. Thornton[1] and W. J. Brazelton[1], [1]1257 South 1400 East, Department of Biology, University of Utah, Salt Lake City, UT 84103

When sampling for studies in microbial ecology, it is inevitable that some of the organisms that get detected at a given site had originated elsewhere from the sampled location and are not capable of surviving under the conditions prevalent there. This scenario may be especially common when surveying microbial communities *in situ* as there is risk of unintentionally introducing, through the act of investigation, foreign material into the site being investigated. The potential for contamination in molecular sequence data is something that all such studies should consider; but special consideration for this issue should be given when investigating subsurface habitats in particular, due to the increased and unique risk of introducing contamination during sample collection. Introduced contamination is not restricted to the sampling process, however, as it can also occur during any of the subsequent stages of processing and laboratory handling. Reagents used in DNA extraction and sequencing can harbor their own sets of microbes [1], for example, and their DNA will be sequenced along with the sample DNA.

The picture is further complicated in that contamination need not be restricted to what gets introduced by the investigator. In the context of microbiome surveys, a contaminant can be thought of as an organism that is non-native to a habitat hostile to its lifestyle, but is present there nonetheless either transiently or in a dormant form. In this scenario, contamination becomes a special case in dispersal [2] – specifically, migration without establishment – and will occur naturally due to the mixing of materials in the environment. Consider a hypothetical case in which fluid samples are extracted from drill wells embedded in a stream. Inevitably such a solution will contain a mixture of subsurface fluids and stream water, as the well will serve as a transition zone for these two endmember sources. True subsurface fluids are nearly impossible to obtain in practice and unavoidably contain some form of 'contamination'.

Many molecular-based surveys of microbial communities in nature have either ignored the issue of contamination altogether or have attempted to address it solely through the establishment of strict sampling and handling protocols. Some studies have attempted to introduce more rigor to their investigations through the use of control samples [3]. Traditionally the standard treatment for a sequence variant, representing the cumulative instances of a single sequence, that is present in both the control and the sample of interest would be its removal from further analysis. But inevitably a situation will arise in which the variant will be found in high abundance in the study sample but low abundance in the control. In such cases, it is more probable that the direction of movement was from the sample of interest to the control, but the standard approach would compel the investigator to throw out the data anyway. Recognizing the overly "conservative" nature of this treatment, some attempts have been made to develop and use statistically-informed approaches instead [3][4].

Techniques that have previously been used for detecting endemic subsurface microorganisms in deep-biosphere studies have been inconsistently applied and so far no attempt has been made to understand their appropriateness to a given contamination "scenario", or the advantages and disadvantages associated with their use. The goal of this project is to compare and evaluate these various methods. Evaluations will be made on data simulated under different contamination scenarios, in addition to data obtained from a perennial stream that obtains most of its flow from a series of aquifers – a system ideal for addressing these questions due to the presence of drill wells that can serve as a "window" into the subsurface and a clear directionality to dispersal. Determining when a given approach is best suited to a situation will be an invaluable resource to not only investigations of the subsurface but could potentially be adapted to studies of non-terrestrial life, if ever applicable.

References:

[1] Salter S. J. et al. (2014) *BMC Bio., 12*, 87. [2] Nemergut D. R. et al. (2013) *MMBR, 77, 342-256*. [3] Inagaki F. et al. (2015) *Science*, 349, 420-424. [4] Brazelton W. J. et al. (2017) *PeerJ*, 5:e2945.