Abstract for presentation at Society for Social Studies of Science (4S) Annual Meeting, Washington, D.C., Oct. 28 - Nov. 1, 2009.

Cyberinfrastructure and Scientific Validity in Metagenomics Research

Matthew J. Bietz
Human Centered Design & Engineering
University of Washington, USA
mbietz@u.washington.edu

Charlotte P. Lee
Human Centered Design & Engineering
University of Washington, USA
cplee@u.washington.edu

At the same time that large computation infrastructures are enabling new kinds of science, they are also having second-order effects on conceptions of the scientific method. Computational technologies are becoming integral parts of scientific practice [1], but both their technical complexity and the social circumstances of their creation and use are challenging traditional assumptions about scientific validity. This paper reports findings from a qualitative study of the development of cyberinfrastructures for metagenomics and makes contributions to our understanding of the interactions among scientific practice and computational technologies.

Metagenomics is a "new science" that uses genetic analysis to study populations of microorganisms and their relationship with their environment [2], and depends heavily on new DNA sequencing and computational technologies. Metagenomic researchers use large public DNA sequence databases to perform comparative and statistical analyses. These databases are populated "by the community," and both social norms and contractual provisions encourage researchers to submit their data. There are a number of widely-used analysis tools, but it is also common for scientists to modify someone else's software, or write their own custom algorithms and analysis scripts.

Interviews with scientists suggest a dawning awareness that replicability, the traditional benchmark of a scientific fact in this field, is being compromised by the computational tools. The amount of available sequence data in public databases is doubling approximately every twelve to eighteen months, and at that pace, a comparison against the database done today could yield significantly different results from one performed just a few months in the future. Many sequence databases are known to have errors including redundant data, misclassified data, and errors in the DNA sequences themselves. And while two different database systems may claim to contain the same data, in practice they rarely do. Some analysis tools are "black boxes" where the actual algorithms and parameters are hidden from the

scientists. But even when the source code is available, bugs, version differences, and tweaks to parameters can make it difficult if not impossible to fully replicate an analysis.

This paper examines not only how the participants in the study view these problems, but also how they are responding to them. Some metagenomic researchers (often in cooperation with computer scientists) are working to create technological and social "fixes." For example, data standards, curation and review processes, open-source licensing of analysis software, and repositories for standard operating procedures are hoped to increase the ability to replicate an experiment. At the same time, other scientists are rethinking the relationship between the laboratory bench and the computer, and exploring the ramifications of moving from a hypothesis-driven science to a more interpretive science.

1. Hine, C., *Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work.* Social Studies of Science, 2006. **36**(2): p. 269-298.
2. National Research Council (U.S.). Committee on Metagenomics: Challenges and Functional Applications, *New science of metagenomics : revealing the secrets of our microbial planet*. 2007, National Academies Press: Washington, D. C. p. xii, 158 p. : ill.