Human Microbiome Project Sampling Workshop
July 25, 2007

David Relman, Chair

Welcome and Introduction

David Relman opened the workshop by describing the Human Microbiome Project (HMP) as a hypothesis-generating endeavor for establishing resources and tools that will allow researchers to further explore the role of the human microbiome. He set out the following questions that must be answered in the creation of a reference data set providing these resources and tools:

- The definition of a normal or healthy human being;
- The variability of the donor;
- A relevant spatial scale for sampling;
- A relevant time scale for sampling;
- The importance of rare community members;
- Relevant statistical markers for analysis.

Fiscal Year 2007 Jumpstart Funding Plan

The sequencing centers provided an overview of their current efforts directed towards generating genomic sequences from human-associated microbes. The initial goal for the HMP is to produce 1000 sequenced human-associated bacterial genomes that will be deposited in public databases as a reference against which whole genome shotgun metagenomic sequence data can be compared. Additionally, the centers plan to generate 16S rDNA sequence survey data from a set of body regions, to be defined by this workshop.

Ethical Issues of Risk Associated with HMP Sampling

The area of ethical, legal, and social implications of this research was described, highlighting the issue of respect or disrespect across social boundaries that could result from this work. The use of previously collected specimens to avoid the challenges of de novo sampling was discussed, but it was noted that few existing sample sets are sufficiently consented for public online data distribution as required for this project.

GI Tract Sampling Discussion

Experts in human gastrointestinal tract microbes provided an overview of sampling issues and discussed current data. Recent studies have shown that there is significant variation among individuals at the level of species and strains, but less variation on a grosser taxonomic scale. Within an individual, community composition is consistent over long periods of time, despite changes in the relative abundance of individual species. Although sampling the gut can be quite complex and varying levels of invasive sample collection were described, stool samples provide excellent material for a high level survey and are easy to obtain. This ease in sampling led to the suggestion of sampling stool samples from a large number of individuals, thereby obtaining an overview of microbial diversity in the gut.

Oral Cavity Sampling Discussion

The discussion of sampling the oral cavity began with reflections on the difference between the healthy and the normal individual. The prevalence of various forms of oral disease, many influenced by bacterial communities, renders orally healthy individuals atypical of the United States population. For this reason, although careful metadata collection by a dental specialist was encouraged, experts suggested sampling representative individuals rather than those with perfect oral health.

Experience to date in this field is that nucleic acid extraction must occur prior to sample storage. Efforts are currently underway to create draft sequences of large numbers of oral microbes as a database against which sequence data from these nucleic acid samples can be compared. Previously demonstrated correlations between oral and gut microbiota strongly support obtaining both GI and oral samples from the same individuals. Because of the goal of the HMP is to generate a reference data set, the suggestion was made to sample as many young adults as possible instead of sampling a smaller cohort more heavily or more frequently.

Vagina Sampling Discussion

Experts in vaginal sampling presented a tiered sampling approach, involving high-level characterization of microbial diversity to allow the use of statistical analyses to determine when sampling has revealed all of the microbes present or above a set frequency. Once the sampling reaches this point, clustering of the initial data identifies the most representative samples for follow-up sequencing, thus reducing the burden of the more detailed analysis. It was suggested that this approach be implemented in the sampling for the HMP, instead of establishing a set number of individuals to sample at the outset.

Several specific criteria for sampling individuals were also discussed, particularly the need to link sampling between the vagina, the GI tract, and the oral cavity. The problem of defining healthy individuals was raised because reported symptoms, or lack thereof, often do not correlate with clinical measures of vaginal disease. Although self-sampling of the vagina is often effective in clinical settings, the consensus was that it would be suboptimal for this project because samples collected by the donor do not come from consistent and well-defined regions within the vagina.

Skin Sampling Discussion

Experts in sampling the skin emphasized the diversity of the skin as a microbial habitat, both within an individual and among individuals. In particular, the skin microbiota is strongly dependent upon external environmental factors such as the workplace, with healthcare workers showing very different microbial profiles from outdoor laborers. Comparison of sampling mechanisms, on the other hand, has demonstrated that the dominant species returned by each of the three main techniques, swabbing, shaving, and punching, are the same, although rare species were differentially represented between sampling methods. For this reason, simple swabbing techniques were suggested as sufficient for screening large numbers of individuals.

Statistical Considerations

Key factors for sampling communities were discussed from a statistical perspective, particularly the need to take a rigorous and thoughtful approach. Rigorous sampling requires satisfying and testing statistical assumptions of independence and unbiasedness, while thoughtful sampling requires fundamental decisions about experimental design and the goal of the project. For the HMP, sampling strategies will depend on whether the project seeks to discover as many novel microorganisms as possible or to understand the broader ecology of human-associated microbial ecosystems. For this reason, the goal of the Roadmap project must be explicitly defined and consistently followed. From a statistical perspective, the group was reminded that the smaller the scope of the question being asked, the greater the statistical power to provide an answer. Because of this fundamental tradeoff that the smaller the target of inference, the more that can be known about it, the suggestion was made to involve plant and animal ecologists, taking advantage of their experience balancing breath with depth in ecological metagenomic studies.

Breakout Group Discussions

Following the morning presentations, workshop attendees divided into two breakout sessions charged with assimilating the region-specific recommendations into a sampling plan for the establishment of a HMP data resource. Their task was to address how many samples to collect, what samples to collect, and how to obtain the necessary samples. Following the sessions, the attendees reconvened to present and discuss their recommendations.

Breakout Group I

In defining the scope of the project, the first breakout group saw the greatest scientific relevance in defining overall human-associated microbial representation, not in seeking microbial novelty. On this basis, the group described a stratified cohort sampling plan with three nested sampling levels:

- A large number of individuals (>>100) who would undergo basic non-invasive sampling of all regions;
- A subset of approximately 100 individuals who would undergo more invasive sampling;
- And a final subset of approximately 10 individuals who would undergo the most extensive sampling.

This strategy would require the development of specific sampling protocols for each subset of individuals in each region, which should be completed by working groups of region-specific experts.

Although the breakout group did not have the opportunity to discuss the definition of a healthy versus a normal individual, the participants agreed that this distinction was critical for the project. They noted that self-reporting would be the least expensive way to collect health-related metadata, but were uncomfortable with the limitations this imposes in regions such as the oral cavity where disease often remains unnoticed if not identified by a clinician.

Breakout Group II

The second breakout group focused their discussion on a number of major issues affecting the overall structure of sampling. The group made the following recommendations:

- All research participants should be equally consented and sampled from as many regions as possible for both scientific and practical reasons.
- The sample archive should preserve DNA and RNA, not raw sample material.
- Host genetic material should be collected to retain the possibility of obtaining genotypic or sequence data at a later date, even if there are no immediate plans for its use.
- Professional sampling of subjects, as opposed to self-sampling, should be done in order to maintain aseptic technique and accurate sample tracking.
- Relevant metadata will be critical to the project, the most important of which are age, gender, occupation, social and economic status, zip code, diet, medications, smoking status, family health history, BMI, and asymptomatic pathology.
- Experts in sampling each region should be convened as “islands of expertise” to define standard operating procedures for sample collection.
- Although no definitive numbers were discussed for sampling, there was consensus that 100 individuals would be insufficient.