NOTE: This white paper was approved by the Sequencing Advisors excluding the sampling of urine for metagenomic sequencing.

Pilot Experiments for Metagenomic Sample Sequencing

May 3, 2007

George Weinstock, Richard Gibbs
Human Genome Sequencing Center, Baylor College of Medicine

Richard Wilson, Jeff Gordon, Sandra Clifton
Genome Sequencing Center, Washington University

Bruce Birren, Doyle Ward, Chad Nusbaum
The Broad Institute of MIT and Harvard

The human microbiome project (HMP) will require metagenomic sample sequencing data as a primary resource for analysis of the relationship between microbial communities and the individuals they inhabit. How this community impacts (or is impacted by) individuals who may vary in diet, geography, culture, age, and disease is a profoundly important topic. In addition to the human variables, the complexity of the microbial community itself makes sample sequencing challenging, often incomplete, and likely to be very sensitive to sampling methodologies that are still in their infancy. The purpose of this pilot work is to generate data needed to: i) evaluate sequencing instrumentation, laboratory procedures and data handling processes; ii) enable development of new analytic tools; and iii) provide a strong scientific foundation for subsequent metagenomic experiments.

Two broad types of metagenomic approaches are rDNA sequencing (ribotyping) and whole genome shotgun (WGS) methods. Most work to date has used the rDNA approach in which degenerate primers amplify signature portions of rDNA that are then cloned and sequenced. This method takes a census of the community, identifying species that are present by comparison to a database of 16S sequences. Since a single bacterial species can itself be diverse, comprising many subspecies that share rDNA sequences but differ in genome content by as much as 25% (amounting to >1 Mb and >1,000 genes for a medium sized genome such as *E. coli*), this method gives a coarse view of the functional genomic potential of the community. An advantage of ribotyping by PCR amplification of bacterial rDNA is that it is not affected by contaminating host DNA. Ribotyping has been the major method used for metagenomic sampling to date. Virtually all of the human microbiome sampling has been done using ribotyping.

WGS approaches have been used mainly in environmental studies from various sites (e.g., sea water, abandoned mines) where non-bacterial contamination is less of a concern. WGS sequencing is performed by sequencing the ends of clone libraries created as small plasmids or occasionally cosmids. The individual sequences are compared to sequence databases to identify species and genes. WGS studies may include ribotyping in parallel to allow comparison of the species censuses produced by the two methods. In some cases the shotgun reads are assembled to produce contigs (partial assemblies) of the most prevalent genomes. Assembly of even short contigs provides higher base quality for database comparisons and longer sequences that aid mapping larger regions of genome structure and identification of gene and operon relationships for valuable functional analysis.

Before we can proceed with assurance in developing the robust methodologies required to conduct metagenomic experiments, we must derive reasonable models of the population structures of the varied communities we intend to investigate. There is currently little data on which to base such modeling. In what follows, we discuss two recent experiments, which despite their limited data are some of the more extensive that have been performed.

Current pictures of the intestinal microbiome are illustrated by two recent rDNA metagenomic studies (Eckburg et al., Science 308:1635 (2005) “Diversity of the Human Intestinal Microbial Flora”; Gill et al., Science 312:1355 (2006) “Metagenomic Analysis of the Human Distal Gut Microbiome”). These papers are appended since they provide useful models for this proposal. In

the Eckburg study, 3 people were sampled at 6 sites along the colon and with fecal specimens. Over 13,000 rDNA sequences were produced. Nearly 90% of these were bacterial and clustered into 395 phylotypes (figure 1). The bacterial phylotypes were dominated by the *Firmicutes* and *Bacteroidetes*, although nine other phyla were represented at lower abundance. The remaining 10% of the rDNA sequences were from archaea and, in contrast, represented a single species, *Methanobrevibacter smithii*.

How comprehensive was this sampling? Figure 2 shows Collector's curves of the three individuals being sampled. From these it was concluded that at least 500 phylotypes were represented in these individuals, although this is a lower bound. Moreover, the 395 phylotypes observed represented 99% coverage, and each new phylotype would only be discovered after sequencing of 100 additional clones (i.e. $\sim 10^4$ sequences needed to identify the 100 additional phylotypes).

Significantly, Eckburg calculated that inter-subject variability was much greater than between stool and mucosal specimens from any single individual, and that variability along the intestinal mucosa was low. It appeared that no gradient of organisms occurred along the colonic sites but rather there were patches of organisms in different individuals. Moreover, the stool samples represented a combination of mucosal organisms and other, presumably non-adherent luminal organisms.

The study by Gill *et al*. sampled feces from two subjects and produced over 2,000 rDNA sequences. These showed a similar diversity (151 phylotypes were observed, all but one were *Firmicutes*, with up to 300 phylotypes expected with extrapolated sampling), although *Bacteroidetes* were underrepresented, possibly due to differences in DNA preparations. However as with Eckburg et al., the *Firmicutes* dominated, as did the *M. smithii* archaea.

The Gill *et al*. study also performed whole genome shotgun sequencing, obtaining about 140,000 reads. When these were assembled, 60% of the reads formed 18,000 contigs and 15,000 scaffolds covering 34 Mb at a coverage of slightly over 2x. These are likely to represent the most prevalent organisms. However it was also seen that the contigs included heterogeneous sequences, implying that these were the merging of closely related, but not identical organisms. A span of 34 Mb is about 10 bacterial genomes, but given the multiplicity of closely related organisms this may represent, it is likely that the number of prevalent organisms is in the tens and possibly over a hundred.

The remaining 40% (56,000) of the reads, comprising 45 Mb, did not assemble, presumably representing the low abundance organisms.. It is apparent that there is a great diversity in genome sequences in these samples in contrast to the corresponding rDNA observation that 150/151 phylotypes were *Firmicutes*. This underscores the lack of resolution obtained with limited sampling by rDNA methods.

We use these data to make some very rough estimates of the composition of the microbiome. If the 45 Mb of unassembled sequence represents low prevalence organisms that had achieved 0.1x genome coverage (hence they did not assemble), this would suggest those organisms contained 450 Mb of genome sequence, or about 100 bacterial species. Of course this is a low estimate

since the coverage could be much lower than this. Thus it seems reasonable to estimate there are at least hundreds of low prevalence organisms. This estimate is consistent with the rDNA sampling.

Given that the study of Gill et al. achieved these results with 140,000 reads, comprising a little over 100 Mb, we need to increase the amount of sequence we have to work with by a minimum of 10 fold (1.4M Sanger reads or ~1Gb of sequence) to test some of these quantitative assumptions . Specifically, we need perform assemblies using different amounts of the total available sequence to establish the rate at which contigs are formed as a function of depth of coverage. However given the uncertainty about the complexity of the target community and the relative frequency of it's members, we would be better served to further increase the data set. We propose to generate14 million Sanger reads, or 10 Gb, to ensure we reach the needed answers even in the face of many low abundance genomes.

Our inability to more accurately model the effort required to obtain the expected results reflects our fundamental uncertainty about the complexity of the underlying data; this uncertainty can only be dispelled with the larger data set this pilot will produce. Based on existing data we believe that the scale of this pilot is justified and will quickly move the field forward. .

The current efforts of the NHGRI human microbiome pilot project (HMPP) will generate reference genomes and thus enable a greater proportion of the metagenomic WGS reads to be identified. One question that arises is how to assess the impact this effort is having on sequence identification and whether the genomes being sequenced are the most informative. We note that there is no existing data set of WGS metagenomic sequences for the human microbiome comprehensive enough to be used for this purpose. The data of Gill et al. are useful, but as noted above, these may not sample less prevalent organisms very deeply and their may be bias in the DNA preparation.

It is also important to assess the relative performance of various sequencing platforms. For example, there could be differences in database hit rates of single 30b, 250b, and 750b reads from Solexa, 454, and Sanger sequencing, respectively. Moreover, the characteristics of the (partial) assemblies from each data type are also likely to be different. Thus, although there may be significant cost benefits between platforms, the utility of the data requires further exploration. Thus, an important question to be addressed before embarking on a large-scale metagenomic project is the utility and cost-benefit analysis of the sequencing platform to be used.

Additional technical questions outside of sequencing platforms include sample preparation methods and reproducibility within and between centers. As seen in the comparison of the Eckburg et al. and Gill et al. studies, sample preparation can have a profound effect on the results (in this case, the absence of a major phylotype, the *Bacteroidetes*, in the Gill et al. data set). Before any large-scale metagenomic experiment should commence, we would like to establish that sample preparation and sequence generation methods are robust and reproducible between centers.

Once these types of technical issues are addressed, we will be in a position to move forward confidently with experiments that focus on the basic questions of how complex the human

microbiome is, how many reference genomes are needed, how does species diversity impact the depth of WGS sampling required, what degree of variation exists between individuals, and what other issues will be critical for experimental design in the HMP. As noted above, approaching these questions is in an early stage and production of reasonable data sets will be critical for benchmarking the analyses and developing future experiments to resolve the issues.

Goals.

The longterm goal of the Human Microbiome Project is to perform metagenomic experiments that will study the relation of the human microbiome to health and disease. The goal of this pilot is to clarify key issues that will underlie metagenomic experimental design.

1. *Produce shared reference WGS data sets for the intestinal and urogenital tracts.* The key element in this proposal is production of a larger data set than previously available, which would be analyzed to address the various issues above. This would be accomplished by large-scale WGS sequencing of samples. These would be reference data sets in the same spirit as the production of reference genomes already ongoing: they would be made publicly available and would facilitate the NHGRI pilot project (below), be of benefit to the research community in general, and would be an important benchmark data set for planning the broader, future activities of the HMP.

The data set would be produced from more than one sequencing platform, and multiple individuals would be sampled. The scale of sequencing would establish this as the major WGS data set currently available. The amount of sequencing would be modeled based on the current rDNA data sets of these microbial communities. The plan would unfold in stages as discussed in Goals 2 and 3 below.

2. *Evaluate different sequencing platforms for their value in metagenomic sequencing.* The HMPP is already engaged in evaluating platforms for producing whole genome sequences of bacteria. However, the distinct challenges in producing and using metagenomic data require a similar evaluation for metagenomic applications. This will be a valuable complement to this activity and provide a fuller picture of what each platform can contribute to the HMP. Combining these more complete data evaluations with the ongoing cost and pipeline performance studies will allow future investments in instrumentation for the HMP to be made wisely. Moreover, the developmental activities in informatics for the HMP will benefit at this early stage from a clearer view of these data types. This evaluation can be performed early in the data production for Goal 1.

As an example of how this could be approached, we would produce data sets from fecal samples in the manner of Gill et al. but of larger overall size. Their 100 Mb data set allowed about 60% of the sequences to assemble into contigs, and about 50,000 ORFs were predicted which produced 20,000 unique database hits. Data of this scale or slightly larger can readily fit with either Sanger, 454, or Solexa sequencing runs (1-3 454 runs, 1-2 Solexa strips), and evaluated for database hits, assembly, and ORF predictions. These would be the important criteria for comparing performance and cost (cost-benefit analysis) of the platforms.

3. *Obtain initial information on sampling and reproducibility*. In addition to establishing a quantitative basis for sequencing methodology, we will also develop robust and reproducible metagenomic sampling. While we will ultimately want to perform sampling of internal body sites, we will initially aim to minimize host contamination and establish standard methods for sample handling, DNA preparation, and sequencing at the centers. To do this, sampling will be mainly from fecal and urine samples. In this way we will begin to define the factors related to technical reproducibility using the least invasive sampling procedures. We will include urine samples to interrogate low complexity communities. From these populations, smaller amounts of sequence can provide a depth of data that would be vastly more expensive if obtained, for example, from fecal or mucosal samples. In any case, these samples will provide a means to measure variability between centers. We are aware that variability between individuals being sampled will introduce additional complexity in this exercise, but we will mainly be looking for large differences (as between Eckburg et al. and Gill et al. *Bacteroidetes* populations) that greatly exceed that of inter-subject variation. Moreover, part of the exercise can include exchanging DNA samples to ensure that sequencing itself is reproducible.

Following establishing these methods, we will address more invasive sampling methods from both intestinal and urogenital tracts. In the end, this will provide another benchmark for the HMP in determining how to set up robust, reproducible DNA extraction and sample handling protocols.

4. *Assess our progress in establishing reference genomes to aid species identification and functional genomic analysis in future human metagenomic sampling.* This would be performed using the data sets in Goal 1. A critical parameter to measure is fraction of sequences that hit the reference genomes produced in the HMPP, but do not hit any other genomic sequences. This result will be important for evaluating strategies to select additional organisms for reference genome sequencing.

In addition to these four major Goals, the data set described in Goal 1 should allow further analysis of the complexity of the human microbiome as well as its variability. This will be an invaluable yardstick for design of the HMP.

Work plan and cost.

In providing these cost estimates we are mindful that we do not yet know exactly what proportion of the sequencing in this pilot will be performed on the various sequencing instruments and that the cost of sequencing with these different instruments is in rapid flux. A goal of this pilot is to remove the present uncertainty about the cost and value of each of these data types for metagenomic studies. Based on the discussion above, our general plan is to first compare sequencing platforms using the scale of data production of Gill et al. as a guideline. The Gill study produced over 100 Mb of data, and was very low coverage of the samples they used. We would aim for 300 Mb of data from each platform. The estimated cost is about $300,000.

Then we will produce a large data set for the subsequent analyses. The early stages of this activity will be aimed at achieving robust and reproducible protocols between the centers. As with Gill et al., we will focus on fecal samples, with additional sampling of urine. As discussed

above, we would aim for of the order of a 100-fold increase in data over Gill et al., or about 10 Gb of data. This will likely not be from a single platform, but will be apportioned based on the performance and cost-benefit results above. We can estimate an upper limit on this cost by assuming it will not cost more than $0.30/kb, for an upper limit of $3 million (including the cost of testing sequencing platforms). This is overwhelmingly for data production, with minor costs for sample collection and informatics. This will be split equally between the centers.

A possible scenario for this pilot would be to perform WGS sampling of 20 individuals, with an average sampling of 10 times that of the Gill study (which sampled two subjects). In addition, we will perform a limited amount of rDNA sequencing to provide a benchmark for comparison.

Informatics.

The papers by Eckburg et al. and Gill et al. illustrate the basic techniques for analyzing sampling, coverage, and diversity (between subjects, within subjects).

Reads will also be compared (e.g. BLAST) to GenBank and the reference genomes in order to build a picture of the structure of the microbial communities, as well as to ascertain the number of reads that do match any known sequences. In addition individuals would be compared to determine variability in community structure and which organisms may constitute the “core” microbiome and which are peripheral. Some samples (both before and after DNA extraction) would be shared between centers to allow determination of variability between centers. Comparison of results from wgs and rDNA will be performed in all of these cases.

Assembly of reads would also be performed using the different assembly programs in use at each center. This will allow both an assessment of assemblies of data from different sequencing platforms, as well as how assembly software handles the different data sets. The contigs resulting from assembly will be compared to GenBank and reference genomes and evaluated for presence of novel reads, possibly allowing these to be assigned to organisms when they co-assemble with reads with database matches, but taking note of the presence of repeats that give misassemblies. Moreover, the consistency of database matches of contigs and their component reads will be evaluated. Finally, a tally of novel vs known genes will be produced from the database comparisons.

Comparison of reads and contigs will also be to various derivative databases (KEGG, COGS, etc.) and this information will be used to build a picture of metabolic and other functional aspects of the communities and how these may vary between individuals.

Timeline.

Data production would be performed in one year to be on a faster time schedule than the production of reference sequences (two year project). Analysis of data will extend into a second year.

Consent issues.

Samples will be obtained for this pilot with IRB approval

Data and sample repositories.

Sequence data will be deposited in the NCBI Trace Archive, as with all other projects, subject to any constraints raised by IRBs. It is not yet clear whether a sample and/or DNA repository will be available. This topic is under discussion and evaluation. In any case, some metadata as to the source of the sample will be collected, such as age, sex, health status, antibiotic use, etc. of the source.

Figure 1 (Fig. S1 in Eckburg et al.). From Eckburg et al.: "Phylogenetic tree based on the combined human intestinal 16S rDNA sequence data set. The label for each clade includes, in order, the total number of recovered sequences, phylotypes, and novel phylotypes (in parentheses). The angle where each triangle joins the tree represents the relative abundance of sequences, and the lengths of the two adjacent sides indicate the range of branching depths within that clade. The colors used to represent each phylum are also used in Figs.1 and 3. Inset: The domain *Bacteria* (modified with permission from (*32*), ASM Journals, Washington, D.C.). Six of the 7 phyla represented by sequences recovered in this study are shown in red; the unclassified clade near *Cyanobacteria* is not pictured in the inset."

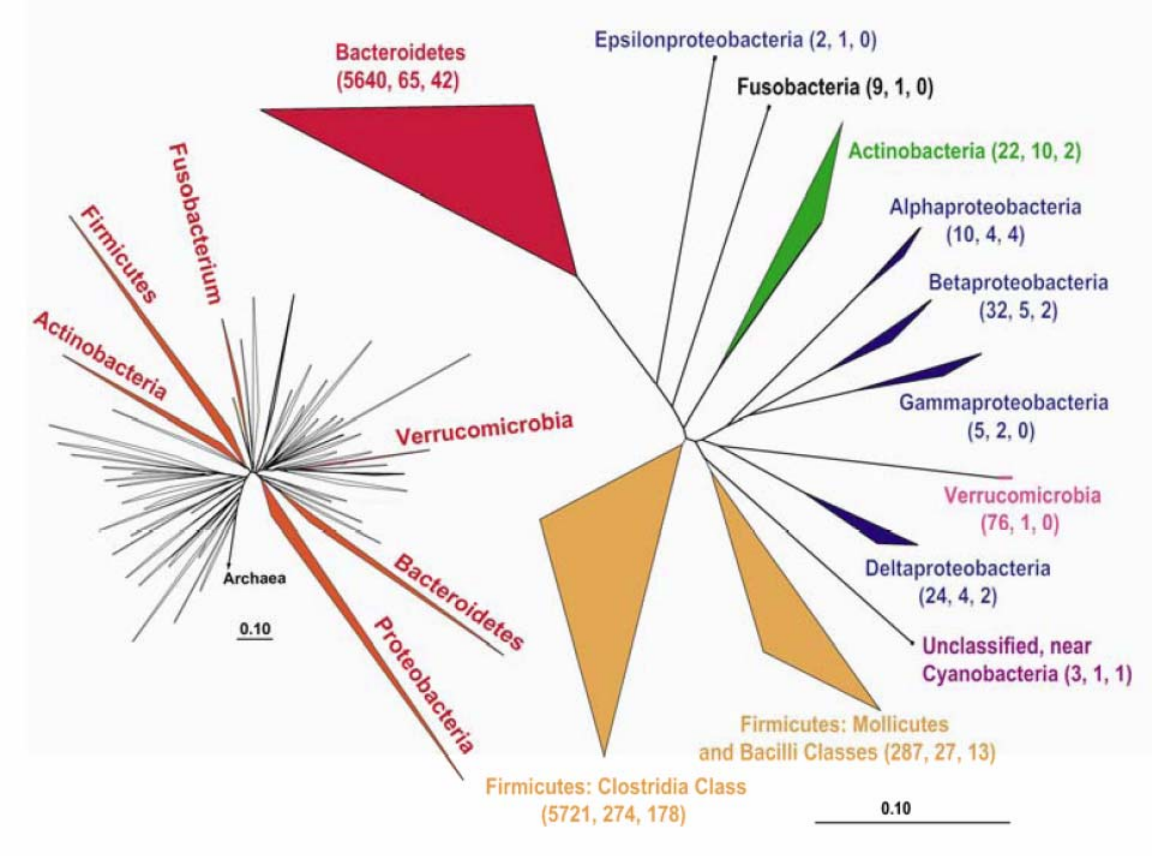

Figure 2 (Fig. S3 in Eckburg et al.) From Eckburg et al.: “Collector’s curves of observed and estimated phylotype richness. Each curve reflects the series of observed or estimated richness values obtained as clones are added to the dataset. The addition of clones along the X axis is nonrandom, ordered by anatomic site (cecum, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum) and stool for subjects A, B and C. The curves rise sharply with the first clones added from a new subject (clone number 4393 for subject B and 7999 for subject C), and then flatten as a majority of clones in that subject are identified. The relatively constant estimate of the number of unobserved phylotypes in each subject (the gap between observed and estimated richness) indicates that both observed and estimated richness may increase with additional sampling. Note the similarity between the different richness estimator curves (Chao1 and abundance-based coverage estimator [ACE]). Phylotypes were defined using the 99% OUT (operational taxonomic unit) cutoff.