Pilot Project to Expand the Number of Sequences of Culturable Microbes from the Human Body

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

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

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

Summary

We propose a Human Microbiome Pilot Project (HMPP) to generate key data that are needed to design a cost effective large scale Human Microbiome Project. Specifically, this pilot will:

1. Create reference genome sequences of representatives of divisions (superkingdoms) of bacteria and archaea in targeted human host habitats to assist assembly and/or interpretation of metagenomic sequence data.
2. Define the quality of the sequence and annotation required for these references and the appropriate technologies with which to produce it.
3. Collect data needed to design effective large-scale metagenomic sampling methods, including:
4. Determine the extent and nature of diversity within members of individual ‘species’ in a single host’s habitats, and between different hosts
5. Define the variation in prevalence of these different phylotypes and hence the required sensitivity of sampling methods.
6. Establish the relative levels of bacterial, archaeal and eukaryotic microbes and viruses in these communities.
7. Explore new methods for storing, displaying, and analyzing these data as necessary to enable rapid progress to a full scale Human Microbiome Project.

The project will be performed by the three NHGRI-funded Genome Centers with coordination and advice from the NHGRI staff and an HMPP Advisory Panel. The project will also coordinate with other groups engaged in human microbiome work and with data repositories that are involved in collecting and presenting human microbiome data.

Introduction

The collection of trillions of microorganisms resident in the human body form complex communities primarily concentrated in a few body sites (Table 1) (Paustian, 2006; Tierno, 2001). These communities contribute to our normal postnatal development, play a significant role in defining our physiology, and vary according to gender, age, environment, diet, and disease states. The key determinants and the degree of variation are not well understood.

Table 1. The Normal Human Microbiota

Body Site	**bacteria/ml or gram**	**# species**
Respiratory		?
Nose	10^3-10^4	
Oral	10^{10} total	>700
Saliva	10^8-10^{10}	
Gingival crevice	10^{12}	
Tooth surface	10^{11}	
Gastrointestinal Tract	10^{14} total	>1000
Stomach	10^0-10^4	
Small intestines	10^4-10^7	
Colon (feces)	10^{11}-10^{12}	
Skin	10^{12} total	?
Surface	10^5	
Urogenital		?
Vagina	10^9	
Human cells	10^{13} total	

While most thinking about the microbial origins of causes of disease has focused on invasion by pathogens, increasing attention is being paid to the idea that the normal microbiota affects predisposition to and can be a critical contributor to a number of pathologic states. Some of the better understood examples, such as dental caries, gingivitis, and vaginitis, result from changes in the composition of the oral and vaginal microbiota with associated host inflammatory responses. These can be thought of as "ecological diseases" which result from alteration of the normal balance of microbes in a community.

More novel mechanisms are being discovered as well. For example, recent studies (Turnbaugh et al., 2006; Ley et al., 2006) show a correlation between the abundance of different bacterial divisions in the gut and obesity. In this case, the amount of energy harvested from the diet is affected by the microbial community structure. Additional discussion of the importance of the gut microbiota was presented in a previous white paper to the NHGRI (Gordon et al., 2005). One hopes that an understanding of the connection between microbial communities and human health will lead to new therapies and diagnostics. Therapies could involve manipulating the composition of microbial communities, or the human genes and gene products whose expression and activities are found to be regulated by the microbiota, or using the chemical entities produced by microbial communities as therapeutic agents. Diagnostics could involve the ideas borrowed from ecology, such as "sentinel species" - organisms whose population dynamics are particularly sensitive to environmental changes (eg, changes in host physiology) and indicative of impending conditions.

As recognition of the importance of the normal microbiota of the human body has grown in recent years, several concepts have emerged. First is the idea that a microbial community is an important operational element itself. Although many organisms within a community cannot be cultured in isolation, they exist in the community environment, which includes not only microbial neighbors but also the host habitat. This community ("biome"), can be thought of as a unit analogous to an organ, and its genome as a "metagenome" (using meta- to denote something of a higher order) comprising the genomes of the included community organisms. This microbial metagenome (microbiome) has been referred to as the "second human genome" and the concept has been suggested of the human as a "superorganism" whose genetic and metabolic landscape is an amalgamation of microbial and human components (Gill et al., 2006). The interactions between components of this superorganism are poorly understood and offer an important area for future research.

The Human Microbiome Project (HMP) has been proposed to further our understanding of this aspect of human biology. In the broadest terms the HMP would produce reference sequences for representative genomes of the human microbiota, and perform metagenomic analyses (sequencing of microbial communities as mixtures of genomes) of samples obtained from individuals representing different aspects of the human lifecycle (i.e. males and females of varying ages with various lifestyles and living conditions, and with physiologic and pathophyisologic states, etc.). The reference sequences would allow the component organisms to be identified in the metagenomic samples and comparison of the metagenomic data would allow correlations to be made between the content of organismal and gene lineages and the human condition, as in the obesity example cited above.

The HMP is a very complex project, given the genetic/environmental/lifestyle variations that exist among humans, the variations in the composition of microbial communities that exist between individuals, the importance of considering space and time when sampling diversity in a given host habitat, the question of whether the species concept can be meaningfully applied to the microbial world, technical challenges in deep community sampling given the enormous range of abundance of different microbial phylotypes within some host habitats, the inability to culture most members of a microbiota using current methods, and many other issues. Nevertheless, while still at the conceptual stage, the HMP has gained much momentum over the last year, in part because of the introduction of a new generation of highly parallel DNA sequencers, and in part by the development of new computational tools for comparing microbial communities whose composition has been described by 16S rRNA gene sequence-based enumeration. Within the NIH, several discussions have been held and most recently the HMP has been selected for further development as one of the next Roadmap initiatives. In Europe, a conference was held addressing the idea of an international HMP, and recently the European Union announced funding for this initiative (Eichinger, 2007). Outside of the EU, there has been a commitment to the project at the Wellcome Trust Sanger Institute, and genome centers in China, Japan, and elsewhere have indicated their enthusiasm to participate with local funding. Despite a need to clarify many technical details, the broad enthusiasm for the concept indicates that it will eventually become a reality.

Establishing clear pilot studies at the appropriate scale to address key questions is critically important for the successful design of the HMP. NIH is currently funding several metagenomic projects. These include a RFA from the National Institute for Dental and Craniofacial Research (NIDCR) for metagenomic research of the oral cavity (currently being reissued) and a project at The Institute for Genomic Research (TIGR) funded by the National Institute for Allergies and Infectious Diseases (NIAID) for metagenomic studies of the vaginal flora. Last, but not least, the NHGRI approved the Human Gut Microbiome Initiative at the Genome Sequencing Center at Washington University in 2005 (Gordon et al., 2005).

These initial projects are important, but not of comparable scale to earlier NHGRI pilot studies for the Human Genome Project (during 1996-1998) or the ENCODE project (currently being completed). In this proposal we seek to expand the existing NHGRI pilot project, expanding it to include all three NHGRI-funded Genome Centers and broaden the scope. The need to obtain more information for the HMP through such projects is becoming pressing as the NIH Roadmap and EU initiatives move forward.

Goals of the Human Microbiome Pilot Project (HMPP).

Because the Human Microbiome is extremely complex, much baseline information is needed in order to properly design the HMP. The lack of this information makes it challenging to decide on an appropriate scope for a pilot project that will contribute significantly. We propose a Human Microbiome Pilot Project (HMPP) to provide some benchmarks, explore some key issues in studying the Human Microbiome, and add to the set of reference sequences needed for metagenomic sequencing studies. Some of the questions that will be addressed in the HMPP are:

- what is the range of variation of gene content in organisms belonging to a given 16S rRNA gene lineage (species-level phylotype) (i) within a given habitat of its human host, (ii) between different habitats within that host, and (iii) in the same or different habitats of different hosts;
- how much variation is there in the microbial communities of genetically identical human individuals, and of their mothers;
- what is the content of eukaryotic and archaeal microorganisms and viruses in a microbiota.

How many organisms must be sequenced to answer these questions? This answer itself is not known, but the initial data set from the HMPP aims to either answer these questions or at least provide enough information to design experiments to answer them.

HMPP Proposal.

The Genome Centers at Baylor College of Medicine, the Broad Institute, and Washington University propose to sequence the genomes of at least 300 culturable bacteria and archaea from the Human Microbiome. This would significantly expand our knowledge of microbial genomes, let alone those associated with the human body, since at present there are 436 completed bacterial genome sequences (www.genomesonline.org) with many of these representing organisms of importance in environment communities, but without obvious direct connection to human biology.

The criteria for selection of these organisms are described below, with an understanding that specific selections will be made with the involvement of the microbial research community, the HMPP Advisors, and NHGRI staff. The number of organisms is based on ideas that (1) a significant number of genomes need to be sequenced to address the goals, (2) the financial and experimental resources invested should be kept in line with the exploratory nature of the work and the early phase of the project, (3) the number of organisms needs to be realistic given potential issues in obtaining samples and analyzing data at a fast enough pace to complete the work in a reasonable time frame, and (4) improvements in technology and a larger HMP project in the future lead to the expectation of a significant expansion and continuation of this early work, thus it need not stand alone. We consider the target of 300 could likely increase as new sequencing technologies improve throughput and reduce costs during the lifetime of this project.

In addition to the production of reference genome sequences, we will also perform a limited amount of 16S rRNA- and WGS metagenomic sampling to assess the microbial and viral genomes that should be addressed in the future, as well as to address the diversity that exists between habitats and between individuals.

Selection of Bacteria and Archaea to Sequence.

There will be several criteria in selecting bacteria to sequence. We emphasize that this will be limited to culturable bacteria. In our set, we are targeting taxa belonging to the domains Bacteria and Archaea. We anticipate that part of the NIH Roadmap HMP will be technology development to both cultivate the currently non-culturable organisms as well as to sequence non-culturable organisms. Thus we feel these targets should be left to the Roadmap project.

Body sites. Six major body sites are being considered in the Roadmap HMP: gut, urogenital, skin, oral, ear, and nasopharyngeal. In order to avoid spreading the project too thin, we will concentrate on only two of these, the gut and urogenital tract. These are selected because of their known high content and diversity of bacteria (Table 1), the ongoing studies of the gut microbiome at the WU-GSC, current and previous work on sexually transmitted pathogens and vaginal flora at the BCM-HGSC, and interest of all centers in these tissues.

In addition, there is a NIAID-funded project at TIGR to study the vaginal microbiota which takes a different approach and would be complemented by the approach proposed here. In the NIAID study, an emphasis is made on taking a bacterial census by sequencing 16S rRNA, and only five reference genomes will be sequenced, one each of the five predominant species. As an aside, in most of the metagenomic studies previously performed, the emphasis has been on identifying community structure by sampling 16S rRNA gene sequences, and much work needs to be done on whole genome sampling.

We also note that the NIDCR has posted a RFA for oral metagenomic studies, and thus it seems prudent to defer any work on the oral microbiota until knowledge of the NIDCR projects is available. We expect that by limiting our genome analyses to two host habitats we will be able to have deeper sampling for the questions under study. However as technology and cost returns improve, we would consider addressing other sites.

Selection of organisms.

General considerations: 16 rRNA phylogenies say little about the functional capabilities encoded in the genomes of various taxa. For the purposes of this project, we will use 'species' to refer to named types available in culture (e.g, *Bacteroides thetaiotaomicron*), and 'phylotype' (phylogenetic type) to refer to clusters of related 16S rRNA gene sequences characterized by levels of pairwise sequence identity (≥97% ID is the commonly used threshold used for a "species"). Currently, there are close to 50 bacterial species for which more than one genome sequence is available: the results reveal that a bacterial genome is a dynamic entity, shaped by multiple forces including reassortment, gene duplication and functional diversification, plus gene loss and gain via lateral gene transfer (Fraser-Liggett, 2005). The ability of a bacterial species to acquire and stably incorporate foreign DNA provides an advantage in niche adaptation and may be responsible in large part for the emergence of new strains with unique capabilities. One of the consequences of gene loss and gain via lateral gene transfer (LGT) is that no single genome sequence can describe a species, and that in order to understand the nature of a microbial species we must focus on the "pan-genome", the sum of all genes present in all members of a species (Tettelin *et al.*, 2005; Medini *et al.* 2005). Pan-genome size can be vastly larger than the genome of any single isolate.

With these considerations in mind, there are three principal criteria that would qualify an organism for sequencing:

(1) The organism belongs to a 16S rRNA gene lineage representative of a major phylogenetic group in a given human microbial community, that has not been previously sequenced. In this case the sequence will make a significant contribution to the database of reference sequences that is necessary for interpreting comparative metagenomic studies of microbiomes . We will not limit the sequencing of these reference genomes to those from healthy individuals. It is likely that there are organisms that are present in very low numbers in healthy individuals, but undergo dramatic increases in their abundance in a disease state. Thus we will consider new reference genomes from these sources as well.

(2) Multiple strains (operationally defined as ≥99% 16S rRNA ID) of the same 16S rRNA lineage for comparison. There are several cases for this.

(a) The major groups in a given habitat (e.g. *Bacteroides in* the gut) should be more deeply sampled by whole genome sequencing to begin to define their genomic diversity. How many isolates need to be sequenced to capture this diversity (i.e. how open is the pan-genome)? The number of strains required to constrain/describe the genomic diversity within a lineage will be assessed on a case by case basis using initial data.

(b) Members of the same species but from different habitats should be sequenced and compared to determine if the degree to which variation between sites is different than within sites.

(3) A set of the organisms to be sequenced satisfying the above criteria should come from a reference set of humans. These individuals will thus be more thoroughly sampled than in most experiments. An attractive more controlled study design is to use sets of female monozygotic twins and their mothers. In this way microbial variation within and between individuals with shared early environmental exposures can be explored, with host genotype held constant.

Other sequencing. In addition to the sequencing of the 300 genomes, a limited amount of metagenomic sequencing will be performed to assess how many other genomes, archaeal, eukaryotic and viral, are present in the sites being sampled. This proposed component of the sequencing will seek to combine 16S/18S rRNA gene sequence surveys with whole community shotgun sequencing, and to compare the datasets to the NR database, as well as to draft or finished genomes, to determine how many sequences in the microbiome do or do not match existing entries. The extent of this work will be limited, and dependent on the platform. To give some examples, with 454 sequencing, producing ~500,000 reads/run, performing 10 runs/site would give 5 million reads for the depth of sampling. With Solexa sequencing producing 40 million reads/run, a single run would give eight-fold greater depth. Balancing this is the difference in read length between 454 (200 bases) and Solexa (25 bases) which will affect the yield of reliable/interpretable database hits. In any case, these questions can be explored with relatively few runs for these and other platforms while yielding significant information about the organismal content of these sites. Moreover, sampling of different individuals would provide additional benchmarks for variation.

Quality of sequences. The type of genome sequence produced will depend on the nature of the questions being addressed.

(1) A standard high quality, finished sequence will be produced for the subset of genomes that represent new reference sequences. Finishing criteria will be established among the Centers and NHGRI staff/HMPP Advisors so that difficult regions, such as rRNA operons, that are not informative, could be left in an unfinished state. Typically rRNA operons in bacteria can be 8kb in length, and present in nearly identical multiple copies, and thus pose a challenge to finish.

Finishing and directed sequencing will also be performed during the pan-genome sampling, so that novel regions in different isolates can have the same high quality as the original reference genome.

(2) A high quality draft (whole genome shotgun) will be produced for closely related organisms where the goal is to define diversity of gene content. For some of these organisms, it may be necessary to introduce a round of autofinishing to improve assemblies, particularly if it is necessary to determine if there are genome rearrangements or misassemblies, or if sequence differences are biological or due to sequencing errors.

(3) Finally, some methods will be explored such as low coverage draft whole genome shotgun and sequencing pools of similar organisms, both aimed at screening larger collections of similar organisms to identify those that merit further sequencing or the extent of pan-genomic variation.

Sequencing methodology. The centers will use a variety of sequencing methods and technologies, including existing and new sequencing platforms, understanding that each is likely to have advantages for certain of the genomes and goals of the pilot. It is also possible that different Centers, or different parts of this project, will use different methods. Some representative results obtained at Baylor College of Medicine from the use of 454 sequencing on bacteria are shown in Appendix 1 to demonstrate the feasibility of using new platforms for reference sequences. The designated methods for sequencing specific organisms will be based upon a regular review of costs and data quality among the Centers, NHGRI Staff and the HMPP advisory group.

Other activities. Outside of sequencing, the principal effort will be in the informatics area. This will primarily be for annotation (e.g. gene predictions) and analysis (e.g. comparisons) of genomes. At present, each Center has its own methods for these processes that are used in microbial genome projects (see below). One important discussion item between the Centers will be to develop standardization in this area.

In addition to annotation and analysis, the Centers realize it will be important to consider methods of presentation of these data to make them most useful to the community. The repositories for public access to the genome sequences (below) are somewhat limited at present in this area. For example, they do not provide online methods to access complete pan-genome information (only the component individual genomes can be accessed and these would need to be manually combined). Similarly, meta-data such as the site the organism came from, what individual it came from, etc. is not readily available. A microbiome database that captured these and other attributes would make the data being produced much more useful to researchers than in the current genome databases. Mindful of this, the HMPP will initially discuss these issues with NCBI, EBI, the JGI or other repositories to see if a joint solution can be developed (this is particularly important given the fact that the environmental microbiology community is devoting considerable resources to comparable efforts). In the event that this will be delayed or not possible, the HMPP will pilot tools to support the user community.

Data release. This will follow the familiar paths used by the Centers on other projects. Raw data will be deposited in the NCBI Trace archive, with immediate release. Assemblies and annotations will be submitted to GenBank according to a specified timeline and posted on the Genome Centers' sites.

The HMPP is sensitive to the possibility that clinical or other sample identifiers must be respected and this will be handled in a manner consistent with informed consent documentation. Any such consent issues will be addressed in consultation with the NHGRI.

Time frame. The HMPP will aim for a two-year project, representing the release of about one genome/week/center. This should allow the conclusions from the HMPP to be available in a time frame that is consistent with the development of the larger HMP. We note that ultimately the release schedule for genome data is subject to sample availability.

Coordination of the project between centers and external advisors. Inter-center coordination will also follow familiar routines, and will draw on experiences with the current Tumor Sequencing Project, as well as the genome project consortia led by the Centers previously. Communication will rely on a listserver, wiki, and probably other web-based tools, and will also include conference calls between the Centers, NHGRI staff, and HMPP advisors, expected to be at least monthly, but likely more frequent at the start of the project. The main topics for discussion will be ascertainment and allocation of samples, technology and analysis methods, evaluation of results, and tracking progress.

The HMPP Advisory Panel will be formed in consultation with the NHGRI staff and the Centers. Panel members will be drawn from the research community. In addition, the status of the project will be communicated to other institutes within the NIH and other agencies internationally that have an interest in ultimately supporting and contributing to the larger HMP.

Other activities, present and future, and the larger context for this project. As noted, the WU-GSC is already approved and engaged in sequencing 100 genomes from cultured representatives of the divisions represented in the human gut microbiota. The development of an NIH-wide HMP is likely to be a part of the NIH Roadmap, and the European Union is supporting related metagenomic activities. Elsewhere within the NIH there is an NIDCR oral metagenomics RFA, and an NIAID award to TIGR for a vaginal metagenomics project. The proposed HMPP is designed to complement, rather than duplicate, these activities and will be proactive in interacting with other groups.

Budget. The following budget estimate is based on the approved WU-GSC white paper, scaled to a target of 300 total genomes. The resulting budget is:

Draft sequences: $20k/genome = $6M (300 genomes)
Finishing: extra $30k/genome = $1.35M (15% of the genomes = 45)
Informatics $1.05M (annotation and databases; $350k x 3 Centers)
Total = $8.4M (one-third of this has already been approved/allocated for Wash U)

The metagenomic sequencing is not included since the extent of this work will be adjusted according to the initial results. However this is not expected to be more than 5% of the budget presented. It is not possible to be more precise than this at present since the sequencing platforms are still in flux, as are their costs. However, it is certain that these costs will be reduced, since these are numbers from last year. Thus this should be considered an upper limit. Moreover, the reduced costs would allow more genomes to be sampled if the total commitment of $8.4M remains.

Current center activities in microbial genomics. While the NHGRI experience with the Genome Centers has not included prokaryotic microbial sequencing until recently, it is worth pointing out that each Center has an extensive microbial sequencing program. These are described below, and form the basis of the expertise in microbial sequencing and analysis that the HMPP will draw upon.

BCM-HGSC: Previous and current microbial genome projects at the BCM-HGSC are listed at www.hgsc.bcm.tmc.edu/microbial. These include 45 prokaryotic genomes and 5 eukaryotic genomes, mainly supported outside of NHGRI. Draft sequencing previously was based on Sanger methodology but is currently almost exclusively using the 454 platform (see Appendix 1 for some examples). Genome finishing uses a pipeline that is being modified to accommodate 454 data. Several genomes have been finished by combining 454 and Sanger data. Annotation and analysis uses a web-based framework, CONAN, developed at the BCM-HGSC that allows remote community involvement in the process. Gene predictions are based on GLIMMER and GeneMark and then manually curated. CONAN provides a variety of other functions for comparative genomics (e.g. pre-computed searches and alignments) and functional annotation (searches of specialized databases for cellular location, metabolic function, etc.).

Broad:

The Broad Institute develops and carries out sequencing projects that produce draft and finished sequence from a wide range of prokaryotes, eukaryotes and the archaea, as well as bacteriophage and eukaryotic viruses. To date the Broad has released more than 35 genome sequences from fungi and oomycetes, more than 20 from bacteria, 2 from protozoan parasites and 1 archaeon. Approximately 80 funded bacterial and fungal genomes are currently in the Broad's sequencing pipeline. In addition, the Broad has established a production pipeline for sequencing entire viral genomes, such as hepatitis C and Dengue for which over 3500 genomes are in the pipeline, as well as methods for sequencing oceanic phage from very small amounts (e.g., 10 ng) of starting material. Ongoing metagenomic projects involve environmental samples from air and soil. To develop microbial sequencing projects that will have the maximum desired impact, the Broad's Project Development Scientists work closely with members of the research community, partner sequencing centers and funding agencies to identify key biological questions and select the appropriate organisms for sequencing, to establish the desired level of coverage/quality of product, and to select the best suited of the many alternative approaches and sequencing technologies. The Broad has extensive experience using new sequencing methods for microbial sequencing, having released more than a dozen bacterial assemblies produced using 454 sequence data.

Our annotation system uses an automated rule-based protocol that mimics manual annotation to use a variety of evidence and multiple gene prediction tools to rapidly produce high-confidence gene sets. Along with our raw annotation data, we also provide a user-friendly web-based interface for querying, retrieving and analyzing results along with a number of annotation quality metrics and standard pre-computed analyses. This interface links to other public databases to make it easy for users to place our gene annotations in a greater biological context. Because many microbial genomes being sequenced are closely related to previously sequenced references we have developed synteny based-methods to transfer these annotations to new sequences and provide tracking between the annotations of different genomes. We have recently established web sites to release data from clusters of related organisms to facilitate comparative analysis.

WU-GSC:

The current status and data for the Human Gut Microbiome Initiative (HGMI) are reported at: http://genome.wustl.edu/sub_genome_group.cgi?GROUP=3&SUB_GROUP=4. (Also see **Appendix 2** for examples). A hybrid sequencing strategy that utilizes reads from both 454 GS-20 and ABI 3730xl sequencers has been devised and implemented to generate the draft genome sequences in HGMI. Genomic DNA is purified from liquid culture derived from a single bacterial colony. At least 15X sequence coverage of 454 reads and at least ~4X sequence coverage of 3730 reads are collected. 454 reads are assembled using Newbler (454 Life Sciences) into 454 *de novo* contigs. These *de novo* contigs are converted to 800 bp paired reads ('superreads') with 400 bp overlap with neighboring superreads, *in silico*. PCAP (Huang et al., 2003) is used to assemble the super-reads and the conventional 3730 reads into a 'hybrid' assembly. One round of automated, targeted sequence improvement is performed using customized primers (autofinish). This is followed by manual inspection and improvement by experienced finishers to ensure quality of the assembly.

Each project begins with assessing the identity and purity of genomic DNA, through sequencing and QC of a 16S rRNA library amplified using universal primers from the genomic DNA sample. Only those that pass the 16S rRNA QC enter the separate 454 pipeline and 3730xl pipeline for read production. Contamination screening at each stage of 454- and 3730xl- read production requires high throughput and high sensitivity, and needs to accommodate characteristics of 454-reads (Margulies et al., 2005), such as their shorter read length, higher error rate in read level, and susceptibility to contamination from environmental DNA during emulsion PCR. To tackle these challenges, an automated contamination screening system has been developed by combining evidence from GC content, sequence coverage, phylogenetic markers, and similarity to known sequences. This system is deployed at each stage of both 3730xl and 454 pipelines to ensure early detection and removal of sequence contamination. Upon completion of QC, the reads and sequence assemblies are immediately submitted to GenBank.

At this annotation phase, tRNA gene and other non-coding RNA genes are determined using Rfam v0.1 (Griffiths-Jones et al., 2005). Protein-coding genes are first predicted using GeneMark v3.3 (Lukashin and Borodovsky, 1998) and Glimmer2 v2.13 (Delcher et al., 1999). Intergenic regions not spanned by GeneMark and Glimmer2 predictions are searched against the NR protein database using BLASTX (Altschul et al., 1990) such that additional coding sequences are predicted based on the protein alignments. Protein-coding sequences are annotated using BLASTP (Altschul et al., 1990), Interpro (Mulder et al., 2005), KEGG (Kanehisa et al., 2004), COG (Tatusov et al., 2000), PSORT (Nakai and Kanehisa, 1991), SignalP (Bendtsen et al., 2004) and Gene Ontology (Ashburner et al., 2000). The results are submitted to GenBank after extensive QC. The various assemblies (draft, pre-finished and manually finished) are also available for downloading from our web site at
http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/
and for blasting at http://genome.wustl.edu/tools/blast/
Seventeen other completed or ongoing prokaryotic genome projects in WU-GSC are listed at: http://genome.wustl.edu/sub_genome_group_index.cgi?GROUP=3.

Salmonella enterica project

An additional microbial project soon to be funded at our center is the *Salmonella enterica* project. *S. enterica* subspecies I contains over 1500 different serovars with an extraordinary diversity of host ranges and pathogenic mechanisms. This is a continuing project that will generate a resource of sequenced genomes and corresponding phenotypic data of judiciously chosen *S. enterica* serovars to capture some of this diversity. In the previous funding period, reductions in sequencing costs allowed us to completely sequence three Enterobacterial genomes instead of the two originally proposed. A recent leap in sequencing technology, using the 454 sequencer, will allow the near complete sequencing of an additional 25 *S. enterica* genomes at the same cost. The genomes will be sequenced to at least 15X coverage on the 454 sequencer with 1 X fosmid coverage produced on the ABI 3730xl to be added to aid in determining contig order and orientation. A reference genome (one of the already sequenced Salmonella genomes) then will be used for scaffolding each newly sequenced and assembled genome to aid in further improvement of the assembly and to determine SNPS and indels. The model used for the Human Gut Microbiome Initiative being done at the GSC will be used for this project, as well.

References cited

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T*., et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J Mol Biol *340*, 783-795.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res *27*, 4636-4641.

Eichinger, N. (2007). European funding targets big biology. *Nature* **445,** 8-9.

Fraser-Liggett, C.M. Insights on biology and evolution from microbial genome sequencing. *Genome Research* 15, 1603-1610 (2005).

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355-1359.

Gordon, J. I., Ley, R. E., Wilson, R., Mardis, E., Xu, J., Fraser, C. M., and Relman, D. A. (2005). Extending Our View of Self: the Human Gut Microbiome Initiative (HGMI). *www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf*.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res *33*, D121-124.

Huang, X., Wang, J., Aluru, S., Yang, S. P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. Genome Res *13*, 2164-2170.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res *32*, D277-280.

Lukashin, A. V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res *26*, 1107-1115.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. Human gut microbes linked to obesity. **Nature** 444: 1022-1023 (2006)

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. The microbial pan genome. *Curr. Opinion Genetics Develop*. 15, 589-594 (2005).

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., *et al.* (2005). InterPro, progress and status in 2005. Nucleic Acids Res *33*, D201-205.

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram- negative bacteria. Proteins *11*, 95-110.

Paustian, T. (2006). The normal flora of humans. *www.bact.wisc.edu/Microtextbook*.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res *28*, 33-36.

Tettelin H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini D., Ward, N.L., Angiouli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc. Natl. Acad. Sci. USA 102, 13950-13955 (2005).

Tierno, P. M. (2001). "The Secret Life of Germs." Pocket Books, New York.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027-131.

Appendix 1. Example statistics from bacterial sequencing with the 454 platform at the BCM-HGSC.

Results from whole genome shotgun sequencing of human commensals and pathogens are shown. In the first three genomes both read pairs and wgs reads (each produced on the 454 GS20 platform, consisting of 100 base reads) were co-assembled to produce a scaffolded assembly. In the other cases no read pairs were added and only the wgs reads were assembled. The latter two genomes contained contaminating DNA which did not interfere with the assembly.

The typical bacterial gene is about 1 kb in length, so the continuity of the contigs produced is sufficient to capture most genes intact. In experiments where draft 454 assemblies are compared to finished sequences, the error rate of sequences in draft contigs is about $5x10^{-4}$ which is only slightly below the error rate that defines finished sequence ($1x10^{-4}$). Thus the expectation is that the 454 platform can deliver suitable draft sequences for the goals of the pilot project.

The longer range ordering following scaffolding is mainly necessary when the genome will be finished, and the high degree of ordering of contigs following scaffolding greatly simplifies the finishing task. Although there are no plasmid templates produced for finishing, most gap-filling can be accomplished from PCR products generated directly from genomic DNA.

We are currently using the FLX instrument from 454 which produces reads of >200 bases. Initial results with this instrument show that the data assembles into longer contigs the GS20 data shown in the table.

Organism	Scaffolds	N50 scaff.	Contigs	N50 ctgs.	Genome size	Coverage
Enterococcus faecalis	8	597 kb	227	28.8 kb	2.71 Mb	21x
Francisella tularensis	20	116 kb	155	23.1 kb	1.77 Mb	27x
Streptococcus iniae	29	140 kb	227	31.0 kb	1.98 Mb	29x
Lactobacillus reuteri	ND	ND	233	33.1 kb	1.92 Mb	30x
Helicobacter pylori	ND	ND	75	48.1 kb	1.55 Mb	41x
*Treponema pallidum**	ND	ND	1352*	76.9 kb	1.28 Mb	*
*Rickettsia prowazekii***	ND	ND	10**	892 kb	1.12 Mb	**

*must be grown in rabbits and is contaminated with host DNA – cannot be efficiently filtered out since the whole genome rabbit sequence is not yet complete.
** must be grown in human tissue culture cells and is contaminated with human DNA: 390 contaminating reads were removed.

Appendix 2. Human Gut Microbiome Initiative: statistics of assemblies generated with a hybrid 454 and 3730xl platform at WU-GSC (as of December 31, 2006)*.

*Only those projects that are in or beyond the manual inspection and improvement stage are listed. An additional 19 projects are in various stages of the sequencing pipeline outlined above.

Organism	Scaffolds	N50 scaff.	Contigs	N50 ctgs.	Genome size
Collinsella aerofaciens	27	175 kb	66	69.8 kb	2.37 Mb
Bacteroides caccae	94	76.7 kb	113	60.4 kb	4.57 Mb
Eubacterium ventriosum	22	291 kb	29	147 kb	2.86 Mb
Ruminococcus obeum	67	92.5 kb	77	84.0 kb	3.64 Mb
Ruminococcus torques	38	99.7 kb	42	99.7 kb	2.66 Mb
Dorea longicatena	34	117 kb	38	107 kb	2.90 Mb