"Most Wanted" Taxa from the Human Microbiome
The Broad Institute

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

1 Abstract

2 Introduction

The human body is home to an enormous number and diversity of microbes. These microbes, our microbiome, are increasingly thought to be required for normal human development, physiology, immunity, and nutrition. The mission of the Human Microbiome Project (HMP) is to generate resources to describe these microbial communities and to analyze their role in health and disease. As part of this effort, the HMP will sequence, at minimum, 3000 bacterial genomes that represent both major and minor constituents of the human microbiome.

Currently, there are over 1000 bacterial genomes at various stages of sequencing through HMP and, yet, we are still far from covering the breadth of phylogenetic diversity present within the microbiome of humans. While we continue to solicit feedback and strains for sequencing from the scientific community, not all human-associated microbes are currently cultivated or are capable of being cultivated using standard methods. However, recent advances in culture- and single cell-based techniques are making access to these hard-to-culture members of the microbiome possible.

We have identified not-yet-sequenced members of the microbiome using methods that incorporate the 16S-based metagenomic- surveys of 200 'healthy' volunteers. Data from each of 18 different body sites were examined and organisms to be targeted using culture- and single cell-based approaches were prioritized based on their distance from already sequenced strains and frequency among samples. The resulting HMP 'Most Wanted' list is a resource for the community interested in isolating and sequencing novel and previously un-sequenced organisms found in association with humans.

This SOP describes the methods process that was used to identify, isolate and sequence the "most wanted" organisms.

3 Requirements

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

4 Procedure

4.1 Clustering 454 data with Abundant OTU and chimera removal

V1-V3 and V3-V5 16S rRNA sequences were taken from the LQ HMP pipeline (from the files "hmp1.v13.lq.seq.summary" and "hmp1.v35.lq.seq.summary" provided in version 3.0 of the release to the 16S working group by Pat Schloss (ftp.hmpdacc.org; /16S/Production/Analysis/PPS-and-SRP002395-1.0/Schloss_Lab-3.0/finalData). The file "pds.metadata" within that release was used to assign subjects and body habitats. The program AbundantOTU v2.0 [1], with the default parameters, was used to cluster 16S sequences from the HMP into 1,440 V1-V3 and 1,258 V3-V5 HMP OTU's. Chimeric sequences were removed with the program UCHIME [2]. OTUs were considered chimeric if their consensus sequences were flagged by UCHIME in either de novo mode, in which the number of times each consensus sequence was observed was set to the number of reads which mapped to the corresponding OTU, or in the reference mode, where the reference was the GOLD database (downloaded from Greengenes on April 10, 2011- Table 1), which contains 16S sequences from fully sequenced genomes and therefore cannot contain chimeras. The website, http://hmpdacc.org/most_wanted/, includes links to AbundantOTU output files that enable retrieval of the individual 454 reads 'assigned' to each HMP OTU.

4.2 General alignment of OTUs

Consensus sequences for each HMP OTU (provided in the .cons output file from AbundantOTU, provided at 'download' link at http://hmpdacc.org/most_wanted/) were used to represent each taxa. Global alignments were performed against each reference database by using the program align.seqs in version 1.15 of Mothur [3]. RDP taxonomy was assigned with version 2.1 of the standalone version of the RDP classifier [4]. All of the databases for which the HMP OTUs were searched are listed in Table 1 and the results of these searches are available at http://hmpdacc.org/most_wanted/.

4.3 Generation of 16S reference data sets

Greengenes [5,6] holds publicly available 16S rRNA gene sequence records from NCBI >1250 bases in length and verified as 16S by NAST alignment [7]. Each reference data set was created as described in Table 1.

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

Table 1. The databases against which the HMP OTUs were compared

Comparative Data Sets	# sequences	Notes on creation	Download date
Silva-SSU_REF v. 104	512,037	[8]	
GOLD	5,441	Greengenes records where the gg 'strain' contains the GOLD 'ORGANISM_NAME' or the gg ncbi_tax_id equals the GOLD 'TAXON_ID'	**April 10, 2011**
GOLD-Human	2,839	Greengenes records in the GOLD subset where the GOLD 'HOST NAME' is "Homo sapiens"	**May 11, 2011**
HMP Strains	1,898	Greengenes records in the GOLD subset where the GOLD 'IMG_HMP_ID' is registered. In cases where a strain's 16S sequence was not found within a contig of the genome project, a sequence from the same species (else the same genus, else the same family) was substituted as surrogate.	**May 11, 2011**
Isolated named strains	117,101	Greengenes records derived from whole genome sequencing projects or records that contain binomial names in the genbank field 'description' and do not contain words such as "unnamed", "unidentified" or "clone" in the genbank fields 'description', 'source' nor 'title'.	**June 16, 2011**
Isolated unnamed strains	5,869	Greengenes records not in Named isolates and without binomial names in the genbank field 'description' and with genbank field 'strain' containing a designation.	**June 16, 2011**

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

4.4 Defining priority/most wanted taxa

We applied the following criteria to place each HMP OTU into a priority grouping:

HIGH PRIORITY (MOST WANTED):
Negative by UCHIME (de novo and reference mode to GOLD)
Global percent identity to either GOLD-Human or HMP strains database <90%
Present in at least 20% of samples from one body habitat (maxFraction >= 0.2)

MEDIUM PRIORITY:
Negative by UCHIME (de novo and reference mode to GOLD)
Global percent identity to either GOLD-Human or HMP strains database >=90% & <98%
Present in at least 20% of samples from one body habitat (maxFraction >= 0.2)

LOW PRIORITY:
Negative by UCHIME (de novo and reference mode to GOLD)
Global percent identity to either GOLD-Human or HMP strains database >=98%
Present in less than 20% of samples from one body habitat (maxFraction <0.2)

CHIMERIC – NO PRIORITY:
Positively identified as chimeric by UCHIME (de novo or reference mode to GOLD)

Using these criteria, each HMP OTU was assigned to a priority grouping (Table 2).

Table 2. The number of HMP OTUs determined to be "high priority", "medium priority" or "low priority" for full genome characterization (non-chimeric HMP OTUs only).

	V1-V3	V3-V5	Both V regions
High Priority (Most Wanted)	85	34	119
Medium Priority	168	170	338
Low Priority	518	489	1011
TOTAL	773	695	1468

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

5 Implementation

Results Access

All analysis results, including the Most Wanted list can be found at http://hmpdacc.org/most_wanted/. Header descriptions for the list can be found at the DACC Most Wanted site as well as in Table 3 of this document.

Linking 454 reads to AbundantOTU's

The website, http://hmpdacc.org/most_wanted/, includes links to AbundantOTU output files that enable retrieval of the individual 454 reads 'assigned' to each HMP OTU. By clicking on these links, you can download a zipped directory for each data set, labeled AbundantOTU_V13_LQ and AbundantOTU_V35_LQ. Each directory contains three files, the extensions of which are explained below:

.clust – provides the identity of all sequences found in each AbundantOTU
.cons – provides the consensus sequence for each AbundantOTU
.clustsize – provides the number of reads found in each AbundantOTU

The .clust file output can be used to retrieve sequences from "hmp1.v13.lq.seq.summary" and "hmp1.v35.lq.seq.summary" provided in version 3.0 of the release to the 16S working group by Pat Schloss (ftp.hmpdacc.org ; /16S/Production/Analysis/PPS-and-SRP002395-1.0/Schloss_Lab-3.0/finalData). If you do not have access to the FTP site, please use the following link to request access: http://www.hmpdacc.org/internal/project_access.php. Please contact Anthony Fodor, Heather Huot-Creasy (HHuot@som.umaryland.edu) and Ashlee Earl if you have trouble accessing any of these data.

Table 3. Headers and header descriptions for comprehensive list of 'most wanted' analysis available for download at HMP DACC (url). Grayed rows represent headers used in Main Table at website.

HEADERS	HEADER DESCRIPTIONS
otuID	Number assigned to AbundantOTU (Ordered from AbundantOTU containing the most reads to the least), including variable region represented by sequence

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

priorityGroup	Priority status for each AbundantOTU: HIGH PRIORITY/MOST WANTED = not chimeric, <90% ID to already sequenced bacterial strains that were isolated from humans and found in at least 20% of subjects at, at least, one body habitat; MEDIUM PRIORITY = not chimeric, >=90% and <98% ID to already sequenced bacterial strains that were isolated from humans and found in at least 20% of subjects at, at least, one body habitat; LOW PRIORITY = not chimeric, >=98% ID to already sequenced bacterial strains that were isolated from humans and/or found in <20% of subjects at all body habitats; CHIMERIC - NO PRIORITY = putative chimeric OTUs not considered for prioritization or most wanted list
variableRegion	16S variable regions captured by sequence
count454	Number of 454 reads assigned to the AbundantOTU
consensusLength	Length of AbundantOTU in bases
consensusSequence	AbundantOTU consensus sequence
rdpSummary	Taxonomic string assigned to AbundantOTU based on RDP database and classifier
UCHIMERefScore	Confidence score assigned by UCHIME in reference mode (GOLD); higher score = increased confidence that the consensus sequence is chimeric
UCHIMEDenovoScore	Confidence score assigned by UCHIME in de novo; higher score = increased confidence that the consensus sequence is chimeric
maxUCHIMEScore	Maximum confidence score assigned by Uchime using de novo or reference mode; higher score = increased confidence that the consensus sequence is chimeric
UCHIMEVerdict	Answers whether AbundantOTU was flagged as chimeric using Uchime in either de novo or reference (GOLD) mode
toGoldGlobalPercentIdentity	Percent identity between AbundantOTU and global best match in GOLD (align.seq, MOTHUR)
toGoldGlobalBestHit	GOLD database identifier for global best match to AbundantOTU (align.seq, MOTHUR)
toGoldHumanGlobalPercentIdentity	Percent identity between AbundantOTU and global best match in GOLD human database (align.seq, MOTHUR)
toGoldHumanGlobalBestHit	GOLD human database identifier for global best match to AbundantOTU (align.seq, MOTHUR)

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

toHMPStrainsGlobalPercentIdentity	Percent identity between AbundantOTU and global best match in HMP Strains database (align.seq, MOTHUR)
toHmpStrainsGlobalBestHit	HMP Strains database identifier for global best match to AbundantOTU (align.seq, MOTHUR)
toSilvaGlobalPercentIdentity	Percent identity between AbundantOTU and global best match in Silva database (align.seq, MOTHUR)
toSilvaGlobalBestHit	Silva database identifier for global best match to AbundantOTU (align.seq, MOTHUR)
toGreengenesNamedGlobalPercentIdentity	Percent identity between AbundantOTU and global best match in Unnamed Isolate database (align.seq, MOTHUR)
toGreengenesNamedGlobalBestHit	Named Isolate database identifier for global best match to AbundantOTU (align.seq, MOTHUR)
toGreengenesUnnamedGlobalPercentIdentity	Percent identity between AbundantOTU and global best match in Unnamed Isolate database (align.seq, MOTHUR)
toGreengenesUnnamedGlobalBestHit	Named Isolate database identifier for global best match to AbundantOTU (align.seq, MOTHUR)
maxFraction	Highest fraction of subjects observed with AbundantOTU
maxFractionBodyHabitat	Body habitat corresponding to the highest fracton of subjects with AbundantOTU
454_seqCounts_Stool	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Buccal mucosa	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Hard palate	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Keratinized gingiva	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Palatine Tonsils	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Saliva	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Subgingival plaque	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Supragingival plaque	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Throat	Total number of 454 reads from body habitat assigned to

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

	the AbundantOTU
454_seqCounts_Tongue dorsum	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Anterior nares	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_L_Antecubital fossa	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_L_Retroauricular crease	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_R_Antecubital fossa	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_R_Retroauricular crease	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Mid vagina	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Posterior fornix	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_Vaginal introitus	Total number of 454 reads from body habitat assigned to the AbundantOTU
454_seqCounts_positive	Total number of 454 reads from positive control samples assigned to the AbundantOTU
454_seqCounts_water	Total number of 454 reads from negative control samples assigned to the AbundantOTU
454_subjectCounts_Stool	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Buccal mucosa	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Hard palate	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Keratinized gingiva	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Palatine Tonsils	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Saliva	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Subgingival plaque	Number of subjects with AbundantOTU for body habitat

454_subjectCounts_Supragingival plaque	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Throat	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Tongue dorsum	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Anterior nares	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_L_Antecubital fossa	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_L_Retroauricular crease	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_R_Antecubital fossa	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_R_Retroauricular crease	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Mid vagina	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Posterior fornix	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_Vaginal introitus	Number of subjects with AbundantOTU for body habitat
454_subjectCounts_positive	Number of positive control samples with AbundantOTU
454_subjectCounts_water	Number of negative control samples with AbundantOTU
454_subjectfractions_Stool	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Buccal mucosa	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Hard palate	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Keratinized gingiva	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Palatine Tonsils	Fraction of subjects with AbundantOTU for body habitat

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

454_subjectfractions_Saliva	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Subgingival plaque	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Supragingival plaque	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Throat	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Tongue dorsum	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Anterior nares	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_L_Antecubital fossa	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_L_Retroauricular crease	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_R_Antecubital fossa	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_R_Retroauricular crease	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Mid vagina	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Posterior fornix	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_Vaginal introitus	Fraction of subjects with AbundantOTU for body habitat
454_subjectfractions_positive	Fraction of positive control samples with AbundantOTU
454_subjectfractions_water	Fraction of negative control samples with AbundantOTU
454_RelativeAbundanceStool	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceBuccal mucosa	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceHard palate	Relative abundance of AbundantOTU per body habitat

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

454_RelativeAbundanceKeratinized gingiva	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundancePalatine Tonsils	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceSaliva	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceSubgingival plaque	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceSupragingival plaque	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceThroat	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceTongue dorsum	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceAnterior nares	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceL_Antecubital fossa	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceL_Retroauricular crease	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceR_Antecubital fossa	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceR_Retroauricular crease	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceMid vagina	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundancePosterior fornix	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundanceVaginal introitus	Relative abundance of AbundantOTU per body habitat
454_RelativeAbundancepositive	Relative abundance of AbundantOTU per positive control sample
454_RelativeAbundancewater	Relative abundance of AbundantOTU per negative control sample

Authors: **Anthony Fodor and Ashlee Earl**
Version: V.3
Effective Date: March 26, 2012

6 Discussion

7 Related Documents & References

1. Ye Y (2011) Identification and Quantification of Abundant Species from Pyrosequences of 16S rRNA by Consensus Alignment. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2010: 153-157.
2. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.
3. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.
4. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261-5267.
5. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, et al. (2012) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J 6: 94-103.
6. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6: 610-618.
7. DeSantis TZ, Dubosarskiy I, Murray SR, Andersen GL (2003) Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. Bioinformatics 19: 1461-1468.
8. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188-7196.

8 Revision History

Version	Author/Reviewer	Date	Change Made
3	Anthony Fodor **afodor@uncc.edu** Ashlee Earl **aearl@broadinstitute.org**	3/26/2012	Establish SOP