HMP DACC DATA FLOW PROTOCOL
FOR 16S rRNA GENE LIBRARIES

INRODUCTION: The purpose of this document is to provide guidelines for the HMP sequencing Centers for submitting 16S rRNA gene data and metadata to the HMP DACC. This document will establish a working protocol so that clinicians, biologists, administrators, and IT staff at each Center and the NCBI will have clear roles and responsibilities. The scope of this document is defined by data types specific to 16S rRNA libraries. Therefore, upstream information (i.e. volunteer attributes, clinical specimen collection procedures) is not included.

GUIDING PRINCIPALS:

1. **The Centers will have maximum flexibility with their internal procedures, annotation, library review/withdrawal and release dates.**
2. **The international research community will be able to consume data in predictable segments and formats.**
3. **The international research community will have access to the data soon after it is generated.**
4. **Differentiation of data objects should reflect the workflow.**
5. **NCBI's Trace Archive, Short Read Archive, GenBank, and forthcoming Analysis Archive are separate entities and yet each of these archives must receive consistent data files with unified project and library ids.**
6. **Basic Inter-library analysis should be available in a single web application and should be responsive to user selections.**

SOME ISSUES TO RESOLVE:

1. **Validate a method to ensure each library is free of any human sequence that would otherwise allow the identification of the patient.**

REGISTER A NUCLEIC ACID PREPARATION (NAP)

- A. Definition of a NAP
    - i. A preparation of nucleic acids (DNA or RNA) extracted from a single clinical specimen.
    - ii. Examples
        1. DNA extracted from one volunteer's fecal sample at the initial time point utilizing physical disruption.
        2. rRNA extracted from one volunteer's serum sample at the second time point utilizing enzymatic lysis.
- B. Who
    - i. Clinical Site (or Site performing the nucleic acid isolation) registers the NAP with the DACC.

- C. Be Aware

    i. Multiple NAPs may be prepared from a single specimen. This occurs when first NAP attempt failed library construction, for instance, and a re-extraction is performed (more possible with stool, less possible with epidermal scrapings).

    ii. A NAP may initially be linked to a 16S library but later could be also linked to a metagenome.

    iii. During Jump Start stage, only two Clinical Sites will perform extractions, but in following years other institutes will prepare NAPs.

    iv. Nucleic acid extraction methods comprise one or more of the following procedures.

        1. Digestion

        2. Disruption

        3. Purification

- D. Timing

    i. Clinical Sites register a NAP after the clinical specimen is registered (i.e. after an **EMMES Global Trace℠** identifier is already assigned for a specimen) and after the nucleic acids are prepared.

    ii. Recommendation: Centers register NAPs in batches on day of extraction.

- E. Procedure

    i. Center creates one text document to describe a batch of one or more nucleic acid preparations (NAPs) where the only difference between NAPs within the batch is the specimen that was extracted. The contents of this file provide the bench protocol followed while conducting the extraction. A consumer of this file would not be expected to reproduce the extraction based solely on the file contents, thus some details are omitted. Rather, consumers should be able to compare multiple NAPs and discover the main similarities and differences among them.

    ii. File Format: Tab-delimited text (JSON or XML can be used an alternative. Contact Todd if you wish to use these alternatives.)

        1. Create a text file with a name that includes the project name, the Center's name and timestamp such as NAP_metadata_JGI_2009-03-15-1430hrs.txt

2. When the value is a list, commas separate the list elements. Spaces following commas are ignored.

3. See NAP_metadata_example.txt for an example of the file's data structure.

4. See NAP_metadata_description.xls for descriptions of each TAG and acceptable VALUES.

iii. Upload NAP_metadata using your choice of two methods (both methods return a verification to user upon success)

1. Use the web form at: http://greengenes-web.lbl.gov/nap_metadata

2. curl -H "Accept: text/x-json" -F nap_metadata=@NAP_metadata_example_2009-05-18.txt -F X-Username=myname@oggh.edu -F X-Password=mypassword http://greengenes-web.lbl.gov/rest/nap_metadata

iv. DACC stores NAP tags and values to be used for web queries and for data collation and submission to NCBI's Analysis Archive downstream.

v. DACC prepares a database view allowing Clinical Sites to examine all their NAPs registered and to review/edit the tags and descriptions they've previously uploaded.

vi. Clinical Site forwards the NAP tubes to a Sequencing Center.

II REGISTER A LIBRARY

A. Definition of a Library

i. A collection of 16S rRNA gene sequences amplified with one pair of broad-specificity primers from a defined nucleic acid preparation NAP then cloned or nebulized.

ii. Examples

1. 2,000 Sanger sequences derived from one nucleic acid preparation amplified using primers 27F to 1492R.

2. 20,000 454 sequences derived from one nucleic acid preparation amplified with bar-coded primers.

- B. Who

    i. Sequencing Center registers a library with the DACC

- C. Be Aware

    i. A single 454 run will often contain multiple libraries.

    ii. Multiple libraries may be generated from a single clinical specimen.

    iii. Some libraries may be constructed but never successfully sequenced (poor sequencing run, for instance). This problem is tolerated by the DACC system. Here's an example:

        1. Amplicons are created using PCR method X and clones are created.

        2. Center decides to register the library metadata with the DACC.

        3. Sequencing QC (see section III: Library Quality Control) reveals only 10% of the clones contain 16S DNA.

        4. Center has options:

            a. Submit the SCF files to NCBI's Trace Archive to document a problem with the procedure.

            b. Not submit SCF files.

            c. With either option, Center would likely create a new library with PCR method Y and register a new library_id.

- D. Timing

    i. Centers register a library after the NAP is registered but before the traces/flows are submitted to NCBI's TA/SRA.

    ii. Recommend Centers to register a library after a successful PCR reaction since some gDNA samples may not produce amplicons.

- E. Procedure

    i. Centers create one text document to describe the construction of a batch of one or more libraries. The contents of this file provide the bench protocol followed while conducting the experiment. Minimal tags must be filled with appropriate values, optional tags are useful but not necessary for composing a complete file. The only differences between the libraries in the batch are the nucleic acid prep (NAP) used as input and/or the primers and/or the

barcodes. A consumer of this file would not be expected to reproduce the experiment based solely on the file contents, thus some details are omitted. Rather, consumers should be able to compare multiple libraries and determine the similarities and differences among them.

ii. File Format: Text (JSON or XML can be used an alternative. Contact Todd if you wish to use these alternatives.)

1. The data in each section is of the format TAG (tab) VALUE (newline).
2. When the value is a list, commas separate the list elements.
3. See Library_Construction_metadata_example.xls for an example of the data structure and descriptions of each TAG and acceptable VALUES.

iii. DACC stores Library Construction metadata tags and values to be used for web queries and for data collation and submission to NCBI's Analysis Archive downstream.

iv. DACC prepares a database view allowing HMP centers to examine all their submitted libraries and review/edit the tags and descriptions they've previously uploaded.

III LIBRARY QUALITY CONTROL

- A. Definition of Library Quality Control:

  i. An *optional* Center-performed Quality Control procedure to aid in deciding if the library should be re-constructed before submission of scf or sff files to NCBI's TA or SRA, respectively.

- B. Who

  i. Sequencing Center can perform this Quality Control procedure.

- C. Be Aware

  i. Some 16S rRNA gene libraries may produce high-quality non-16S data.

  ii. An *entire library* should either pass or fail QC.

- D. Timing

  i. Library QC is optionally performed by the Center before submitting

the entire library's SCF/SFF data to NCBI's TA/SRA.

- E. Optional Procedure

    i. Download most recent 16S core set from http://add.web.address.here/current_16S_core_set.fasta.gz)

    ii. Use blastall to compare all fasta formatted reads from a library (they don't need to be assembled or even trimmed) against the current core set using parameters (-p blastn -q -1 –m 8)

    iii. Count a read as putative 16S when blastall finds >50% identity to a Core Set sequence along >50% of the read length.

    iv. Consider re-constructing the library if <50% of reads are not matching known 16S sequences.

IV CENTER SUBMITS TRACES TO NCBI TRACE ARCHIVE

- A. Definition of a Trace Batch:

    i. A collection of scf files associated with one ore more libraries.

    ii. Centers can submit multiple libraries as a batch submission where common fields (i.e. center_name, amplification_forward) are constant.

- B. Who

    i. Sequencing Centers

- C. Be Aware

    i. Submitting traces to NCBI is a public data release.

    ii. Formatting submissions in a uniform manner allows 3rd parties to consume all HMP 16S rRNA data more efficiently.

    iii. Submitting a library in its entirety allows future improvements in chimera detection and noise reduction to be applied to data.

    iv. The Center-assigned local_library_id and local_NAP_id uniquely links a set of reads to all patient and process metadata.

- D. Timing

    i. Sequencing Center submits traces to NCBI after Center's QC.

- E. Procedure

i. Create a top-level directory with a name that includes the project name, the Center's name and timestamp (year-mo-dy-time).

    1. Example: HMP16S_BSM_2009-03-27-1100hrs

ii. Create two files as children of the top-level directory.

    1. TRACEINFO.xml This is the main file describing the submission. It contains ancillary data and references to trace files.

        a. An example of this file can be found in the Trace_Archive_FTP_Example directory.

        b. Don't forget to update center_name, insert_size, amplification_forward, amplification_reverse, etc.

        c. Note: Oddly, "species_code" contains the value "human gut metagenome" at the request of the Trace Archive curator. NCBI is extending their vocabulary and we aim to change this specification when possible.

    2. README: free text reserved for describing this volume and preparation.

iii. Create a directory named "traces" and one or more subdirectories named by local_library_id containing the traces in .scf format

```
HMP16S_BSM_2009-03-27-1100hrs/
HMP16S_BSM_2009-03-27-1100hrs/TRACEINFO.xml
HMP16S_BSM_2009-03-27-1100hrs/README
HMP16S_BSM_2009-03-27-1100hrs/traces
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/HBBAA1U0001.scf
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/HBBAA1U0002.scf
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/HBBAA1U0003.scf
...
```

iv. Create the MD5 hashes inside the top-level directory:

    1. find * -type f -exec md5 {} \; > MD5

v. Verify directory structure

```
HMP16S_BSM_2009-03-27-1100hrs/
HMP16S_BSM_2009-03-27-1100hrs/TRACEINFO.xml
HMP16S_BSM_2009-03-27-1100hrs/MD5
HMP16S_BSM_2009-03-27-1100hrs/README
HMP16S_BSM_2009-03-27-1100hrs/traces
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/HBBAA1U0001.scf
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/HBBAA1U0002.scf
HMP16S_BSM_2009-03-27-1100hrs/traces/4530/HBBAA1U0003.scf
```

vi. Use tar and gzip to convert the top-level directory and all of its subordinates into a single compressed file. The final file name should reflect the top level directory:

1. tar -cf - HMP16S_GGLBL_2009-03-27-1100hrs/ | gzip -c > HMP16S_GGLBL_2009-03-27-1100hrs.tgz

vii. Transfer the gzipped file to ftp-trace.ncbi.nih.gov using your Center's secure FTP account assigned by NCBI.

V CENTER SUBMITS FLOWGRAMS TO NCBI SHORT READ ARCHIVE

- A. Definition of a Flowgram Batch:
  - i. A collection of sff files associated with one ore more libraries.
  - ii. DACC will submit multiple libraries as a batch submission when appropriate.
- B. Who
  - i. DACC
- C. Be Aware
  - i. Submitting flowgrams to NCBI is a public data release.
  - ii. Formatting submissions in a uniform manner allows 3rd parties to consume all HMP 16S rRNA data more efficiently.
- D. Timing
  - i. DACC submits a flowgram to NCBI only after release is approved by Sequencing Center.
- E. Procedure (Performed by DACC)
  - i. Will be finalized once SRA submission format is finalized at NCBI.
    1. From SRA RFC: "…users should not design systems or software to these specs…"
  - ii. More about SRA
    1. http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=concepts&m=doc&s=concepts
    2. http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=rfc&m=doc&s=rfc (excellent schema descriptions).

VI INTRA-LIBRARY PIPELINE PROCESSING BY DACC

- A. Definition

i. All steps performed of an individual library of SCF or SFF files.

- B. Who
    i. The DACC will perform this task automatically.
    ii. BUT any Center will be able to obtain released data from any other Center at any of the DACC pipeline stages (from raw reads through tables of classified counts).
- C. Be Aware
    i. Processing the flowgrams is still an area of active research and results *will* change: versioning of the analyses and metadata about how they were performed will be essential for interpreting results.
- C. Timing
    i. Pipeline will commence upon receiving traces/flowgrams
- D. Procedure (will need to settle on agreeable parameters)
    i. Call bases (phred)
    ii. Screen vector (cross_match -minmatch 10 -minscore 20)
    iii. Assemble (phrap -forcelevel 10, STITCH)
    iv. Orient (by pattern match for 16S PCR primer)
    v. Locate all primers/barcodes
    vi. Locate homopolymers (threshold 6-mer)
    vii. Determine high-quality span (90% phd 15 within 30-mer windows)
    viii. Align into standard 16S-HMP format (NAST v. current Greengenes core set)
    ix. Chimera assessments (Bellerophon window_size:300 window_search_depth:5 core_ident_thresh:95 par_to_frag_thresh:95 div_thresh:1.1)
    x. Classify

- A. Definition
    i. All steps performed on a collection of libraries of sequences.
- B. Who
    i. The DACC will perform this task automatically.
    ii. BUT any Center will be able to obtain released data from any other Center at any of the DACC pipeline stages (from raw reads through to distance matrices relating the samples).
- B. Be Aware
    i. Inter-library analyses can change substantially depending on intra-library processing, notably quality filtering and chimera checking.
- C. Timing
    i. Inter-library analysis occurs after intra-library analysis.
- D. Procedure
    i. Input to the inter-library pipeline will be (i) a table of counts of each sequence (or prefixes of that sequence) in each sample, (ii) taxonomy assignments for each sequence from intra-library analysis, and (iii) metadata about each sample.
    ii. Output will be a collection of pairwise distances between each library measured using phylogenetic metrics (UniFrac), taxon- and OTU-based metrics (Euclidean distance, Jaccard index, etc.), multivariate reductions of these distances (via PCA/PCoA, NMDS, etc. as appropriate), and visualizations of the multivariate reductions brushed by appropriate metadata. The system will also provide for on-the-fly visualization of subsets of the data (e.g. just the healthy skin samples) brushed by arbitrary metadata (e.g. sequencing technology).
    iii. Analyses of core versus variable components of the microbiota in each habitat according to the method of Turnbaugh et al. 2009 will also be provided, along with plots showing the prevalence of each type of sequence across libraries.

VIII SUBMIT GENBANK RECORD (SEQUIN, FOR INSTANCE)

- A. Definition
    i. SECTION UNDER DEVELOPMENT
- B. Be Aware

i. GenBank advises against creating accession numbers for each SRA read.

IX SUBMIT ANALYSIS ARCHIVE