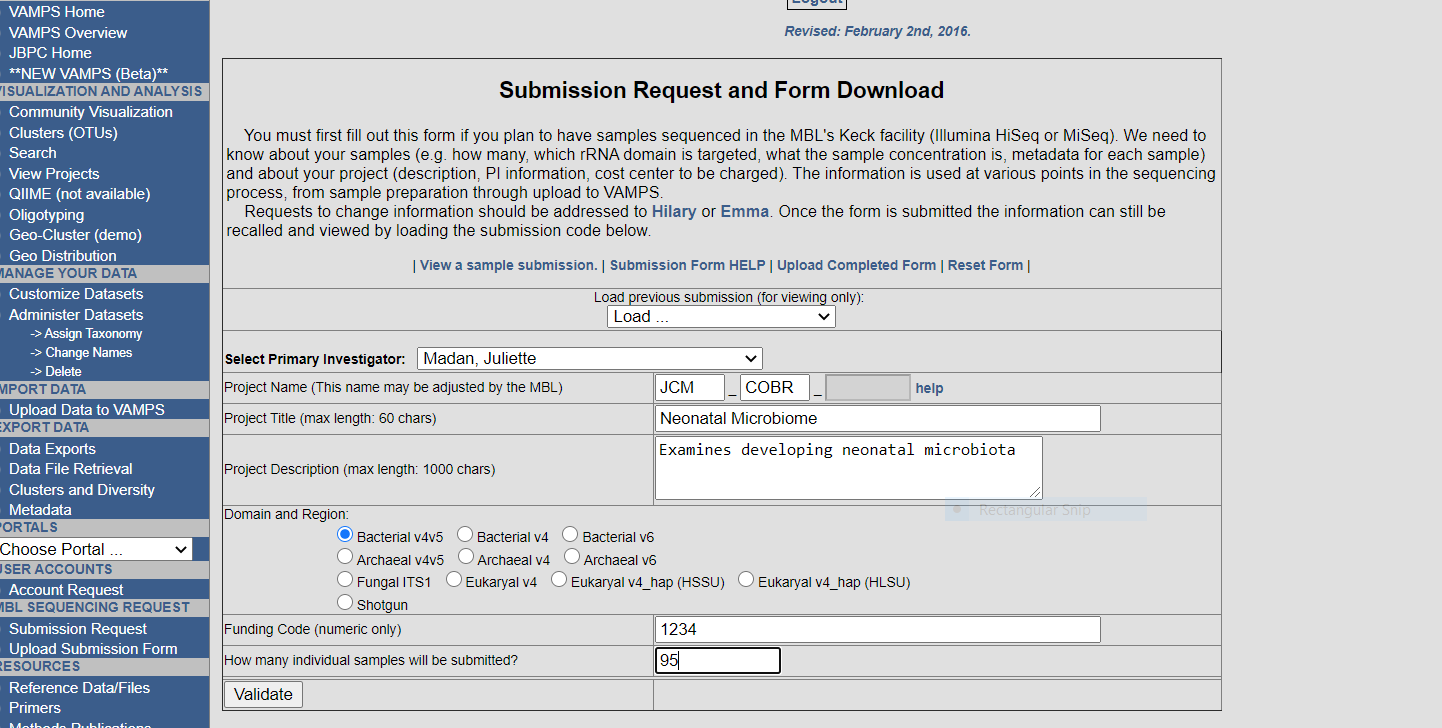
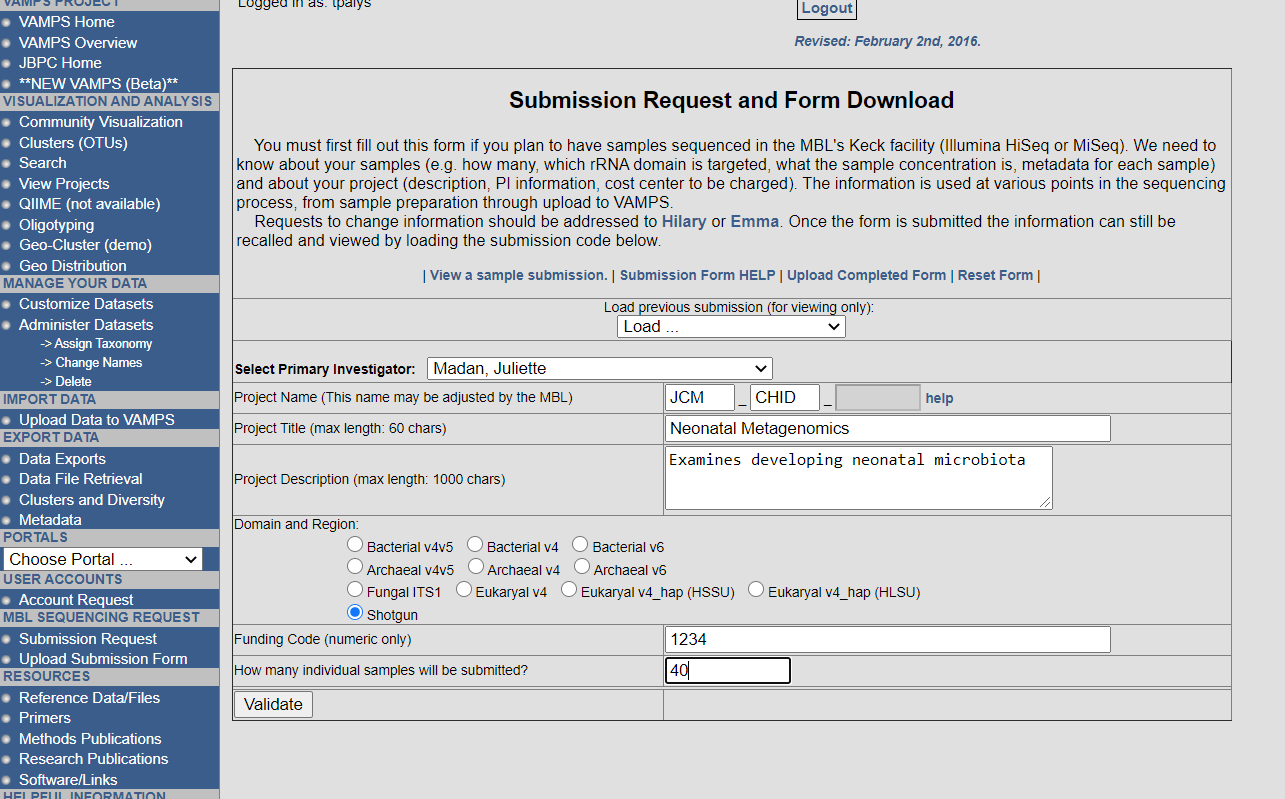
Meeting metagenomics protocol meeting 12/14/2020

Yuka Moroishi; Yuliya Halchenko; Thomas Palys, Scot Zens

1. Action items:
   1. Scot create a github archive for this project (Jan).
      1. Move this document to github in markdown
      2. Push all script copies into the github directory structure
   2. Scot fileserver space management
      1. Compression utility using parallel processing
      2. Mtg with Yuka to add this tool to her metaphlan processing
   3. Scot updating inventory
      1. Flow chart for sample inventory
      2. Reconcile sources
      3. Feedback and checking with MBL
   4. Scot processing kneaddata
      1. Bulk finished
      2. Still handful of stragglers
   5. Yuka update on metaphlan
      1. 500 done; 100 left
   6. Yuka to estimate human 3 process timing
      1. 2 nodes; 10 cores per node; 8 hours; 1 sample
      2. Possible test once on polaris
   7. Next meeting January
2. Draft protocol steps
   1. **Sample referral processes :** Manifest creation (Tom)
      1. Extraction DNA samples identified for referral are picked and Participant ID + Sample ID are scanned into a pre-referral worksheet. Aliquots generated for referral are transferred into new SFDID labeled tubes and are scanned in pre-referral worksheet along parent sample to record sample aliquot linkages. Control aliquots are added to pre-referral worksheet and aliquots are randomized. After randomize linkages are re-established and metadata regarding timepoint, sample type and samples concentration are added.
      2. Physical construction of manifest
      3. Scan label\_id into spreadsheets - S-drive
      4. Saved into DartFS MRKepistor7/nhbcs/LabResults/metagenomics/sample\_manifests/ (Tom)
   2. **Sample referral processes :** Linkage pre-shipping (Tom and Scot; backup Yuliya)
      1. Completed pre-referral sheets will be shared with data manager to check linkages, inadvertent duplicates/replacement samples , time period, sample type and other issues.
      2. For metagenomics, if aliquots had been previous shipped to MBL for 16S, the aliquots can be pulled from unused and returned 16S aliquots and resent for metagenomics. In this case linkages can be established from previous sent 16S manifests.
      3. NHBCS id – label\_type linked to masked labels (Tom creates and Scot verifies)
      4. What is the timeline we need? Involve Scot the day before shipment is sent.
      5. Need to address inadvertent duplicates submitted. (Scot identifies)
      6. Need to confirmation of sample identity, time period, type (Scot identifies)
      7. Batch level tracking; replacement samples are also batches; need explicit tracking of replacements to expedite unnecessary confusion and duplicate processing (Tom)
      8. Linked inventory in DartFS (MRKepistor7/nhbcs/LabResults/metagenomics/inventory/; access by AD group: rc-MRKepistor7-admin) and derived copy made available to lab staff through S-drive. (Tom and Scot)
      9. Delinked inventory in DartFS (access by AD group: rc-MRKepistor7-helpers). (Scot)
   3. **Sample referral processes :** Load Manifest into VAMPS (Tom)
      1. MBL submission request template and upload submission form is generated on VAMPS v1. located at <https://vamps.mbl.edu/>. Access to VAMPS may be obtained for select users by formal registration. VAMPS contact is Andrew Voorhis at [avoorhis@mbl.edu](mailto:avoorhis@mbl.edu).
      2. MBL submission request template is generated by clicking on “Submission Request” link and filling out template required information for desired shipment in Figure 1 and 2. 16S submissions utilize COBR designation and Bacterial v4v5 Domain and Region. Metagenomics submissions utilize CHLD designation and Shotgun Domain and Region.
      3. JCM-COBR – 16S projects for searching on site
      4. Fig 1. 16S template



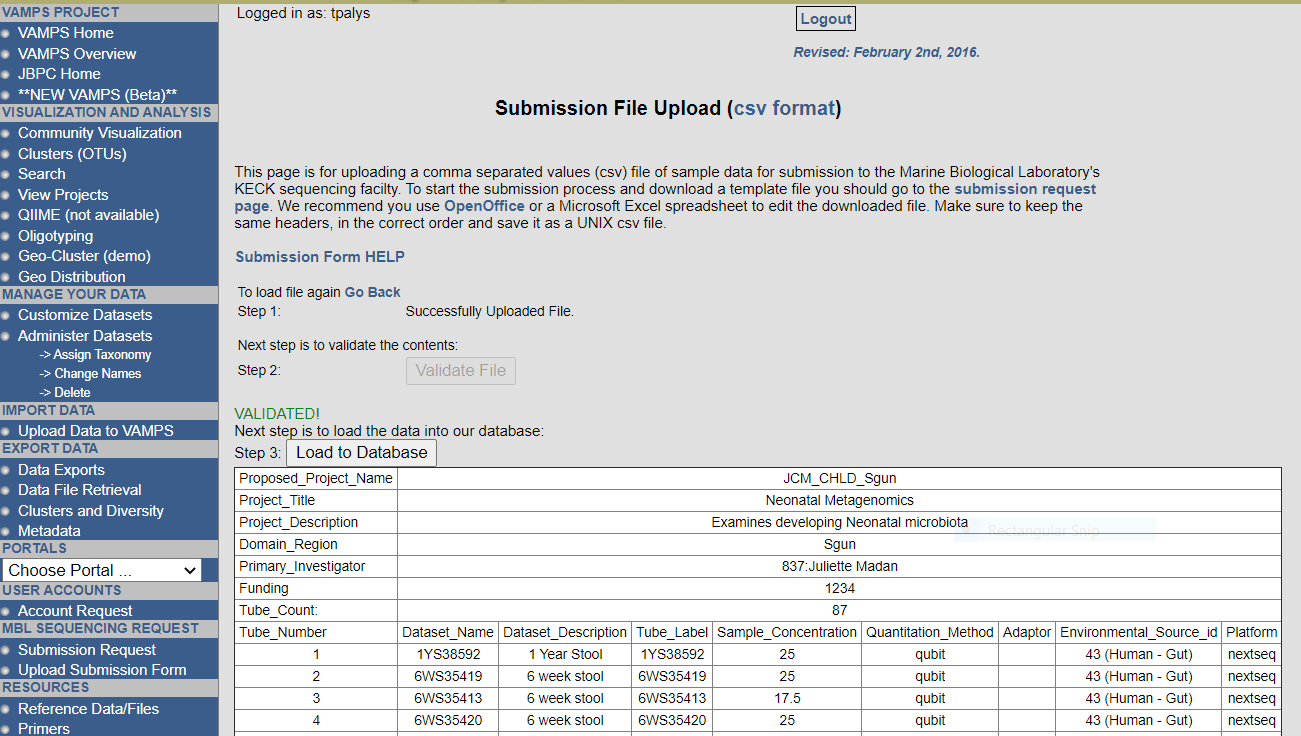
* + 1. JCM-CHLD – metagenomics projects for searching on site
    2. Fig 2. Metagenomics template



* + 1. After templated is validated, template file contain submission headers can be downloaded.
    2. Templated is completed with the prescribes submission metadata. 16S Platform is miseq. Metagenomics platform is nextseq. Stool Environmental source is 40 and non-stool source is 43.
    3. Masked label\_id –> VAMPs, along with metadata (shotgun, 16s, concentration , etc; Tom)
    4. Two submission forms are generated. The first form, which is the submission form uploaded into VAMPS, lacks the Participant ID. The second form is the NHBCS in-house form, which contains the NHBCS Participant ID, is stored in into DartFS MRKepistor7/nhbcs/LabResults/metagenomics/sample\_manifests/
    5. Label\_id = dataset name; used for replacement samples
    6. Replacement samples go out as another batch.
    7. Figure 3 demonstrates example information for both metagenomics and 16S samples.
    8. Fig 3. Example VAMPS Upload Submission Form



* + 1. Once completed, upload submission form is uploaded into VAMPS. This is done by loading and validation the file prior to uploading as shown in Figure 4. Upload submission form and NHBCS in-house submission form are both named with Batch number and approximate date of shipment.
    2. Fig 4. Upload Submission Form Validation examples



1. Data acquisition (Tom and Scot – with accounts on VAMPS and access to the MBL FTP server; backup Yuliya)
2. Metagenomics – sent an email link to MBL FTP server
3. 16S – must monitor the VAMPS website
4. Tom is in contact with Hilary Morrison ([hmorrison1981@gmail.com](mailto:hmorrison1981@gmail.com)) and Nicole Robichaud <[nrobichaud@mbl.edu](mailto:nrobichaud@mbl.edu) about all process scheduling and sample handling, and batch data ready for download.
5. Scot downloads actual data into MRKepistor7 (DartFS) compressed raw tarballs (access only to rc-MRKepistor7-admin).
6. Quality control and inventory with downloaded data from MBL FTP server (Scot; back-up Yuliya)
   * 1. All done on andes.dartmouth.edu; requires linux shell programming, python programming, large file and long job optimization techniques.
     2. Run scripts to check tarball contents and load into inventory.
     3. Extract individual datasets to raw compressed \*.fastq.gz directories and make read-only to rc-MRKepistor7-helpers.
     4. Check matching sequence data for each sample on manifest. Resolve discrepancies with Nicole Robichaud and Hillary Morrison.
7. Check for dataset duplication (inadvertent duplicates from samples sent twice, data reported back from MBL twice, etc.) and report to Tom and Modupe to identify the sample/set of data to use for subsequent processing.
8. Check for concatenation of files and implement in raw compressed \*.fastq.gz. (previous concatenation result files have names with \_S01\_)
9. Calculate sequence counts, quality read score (4th line alpha numeric PHRED Score - base33 or base64(older) – convert to integer (for the sample, average, min, max))
10. Check each dataset for completeness (count the reads) communicate if any files are not complete with Nicole Robichaud and Hilary Morrison.
11. Update inventory with sequence counts, quality read score (e.g. stool raw.seqcounts.qc.stool.897.csv produced by ~/SourceFiles/MBLmetagenomics/countFQreads\_w\_qc.sh).
12. Report back to MBL; Hilary Morrison SRA genebank ids and metadata.
13. Preprocessing (Yuka and Scot; back-up Yuliya)
    1. Install and maintain conda environment with software versions of python and biobakery to match the versions Modupe and Yuka use for Production
    2. Maintain the reference database – keep versions current.
    3. Produce kneadata (sample\_id\_kneaddata.fastq) files from raw files on MRKepistor7 one per sample).
14. Processing (Modupe; Yuka) (replace with Yuka’s long protocol document)
    1. Use discovery.dartmouth.edu
    2. Use knead files from MRKepistor7 (DartFS)
    3. Create conda environments
    4. HUMAnN2 – current processing path
    5. HUMAnN3 (with Kevin Bonham’s help; [kbonham@wellesley.edu](mailto:kbonham@wellesley.edu))
       * 1. Huttenhower group – Kevin develops here
15. Data requests (Yuliya; Scot)
    1. Create manifests for data request (based on inventory; Yuliya)
    2. Copy files, maintain permissions (Scot)
16. Establish DartFS directory structure to consolidate and represent pipeline
    1. Two types of disk used:
       1. DartFS-hpc [/dartfs-hpc/rc/lab/M/MRKepistor7/]
          1. 25TB
          2. faster for activity only where andes, polaris or discovery are processing
          3. premium cost and availability
          4. difficulty backing-up is creating resistance for further expansion on campus.
       2. DartFS [/dartfs/rc/lab/M/MRKepistor8/]
          1. 30TB
          2. still very fast access
          3. lower cost (1/2)
          4. easier to back up
       3. Set-up so that users can use one path [/dartfs-hpc/rc/lab/M/MRKepistor7/] and still see all the files.
       4. Keep more static files on standard DartFS.
    2. MRKepistor7 [/dartfs-hpc/rc/lab/M/MRKepistor7/nhbcs/LabResults/metagenomics/]
       1. ~/metagenomics/sample\_manifests/
          1. Directory is a symlink to MRKepistor8
          2. Tom Palys would save spreadsheets here
       2. ~/metagenomics/inventory/
          1. Directory is a symlink to MRKepistor8
          2. Existing inventory spreadsheets plus new consolidated inventory
          3. separate subdirectories for identified and unidentified inventory
       3. ~/metagenomics/raw\_tarballs/
          1. Directory is a symlink to MRKepistor8
          2. Downloaded tarballs
       4. ~/metagenomics/raw\_compressed/
          1. Directory is a symlink to MRKepistor8
          2. Detarred datasets
       5. ~/metagenomics/quality\_control/
          1. Directory is a symlink to MRKepistor8
          2. Shell scripts, python code used for cleaning, counting, checking, and updating inventory
       6. ~/metagenomics/reference\_datasets/
          1. datasets as needed for H2 and H2 pipelines, plus other qc processes
       7. ~/metagenomics/preprocessing/
          1. Directory is a symlink to MRKepistor8
          2. kneaddata processing
          3. output streams into MRKepistor7 subdirectories
       8. ~/metagenomics/processing/
          1. H2 and H3 subdirectories
    3. MRKepistor8 [/dartfs/rc/lab/M/MRKepistor8/nhbcs/LabResults/metagenomics/]
       1. ~/metagenomics/sample\_manifests/
       2. ~/metagenomics/inventory/
       3. ~/metagenomics/raw\_tarballs/
       4. ~/metagenomics/raw\_compressed/
       5. ~/metagenomics/quality\_control/
       6. ~/metagenomics/preprocessing/
17. Active directory groups for managing this project:
18. Group "rc-MRKepistor7-helpers"

Description: group who processes and maintains the unidentified data. Protected from all ID linkages.

Members of group "rc-MRKepistor7-helpers" are:

Michael S. Zens [d78039e; Epidemiology; DART; Staff]

Modupe O. Coker [f002sqn; Epidemiology; DART; Faculty]

[rc-MRKepistor7-admin] [Nested group]

Thomas J. Palys [f001d1b; Epidemiology; DART; Staff]

Yuka Moroishi [f00359t; GR; DART; Student]

Yuliya Halchenko [f000yzt; Biomed Data Sci; DART; Staff]

* 1. Group "rc-MRKepistor7-admin"

Description: group who maintains inventory, identified look-up and sample quality control.

Members of group "rc-MRKepistor7-admin" are:

Margaret R. Karagas [d31101k; Epidemiology; DART; Faculty]

Michael S. Zens [d78039e; Epidemiology; DART; Staff]

[rc-DartFSadmin] [Nested group]

Yuliya Halchenko [f000yzt; Biomed Data Sci; DART; Staff]

Thomas J. Palys [f001d1b; Epidemiology; DART; Staff]

* 1. A parallel group “rc-MRKepistor8-admin” is maintained and contains the group “rc-MRKepistor7-admin” changes to the dartfs-hpc access permissions will be automatically updated in the dartfs space to match.