**Methods and overviews**

**MR DNA provides free comprehensive taxonomic analysis of data. This analysis is provided in tab-delimited text format and excel sheets that can be imported into excel or other formats. This analysis is provided free for most 16s, 18s, and ITS projects. NOW also for many different functional genes.**

**Please respectfully, because this taxonomic analysis is a free service, we cannot provide extensive support of this data. We do offer wide range of bioinformatics and consulting if you do need extensive support. We are always improving and updating our functionality. This analysis is robust, but because it is what we consider developmental and always changing (database updates, process tests etc.) it will change. The following is the basic standard overview of the process only. We do strongly suggest that all data be analyzed or verified, even if you intend to use this free analysis, with secondary processes such as Qiime2 or Mothur when possible (Just good science to double check things). We do our best to ensure that data provided is accurate and substantial, but we do not accept any liability for errors or omissions in this free analysis. Please do not use the methods here verbatim, MR DNA is happy to review methods with you as needed for your specific analysis and sequencing project.**

**We do have a very extensive range of bioinformatics support available just ask for pricing For ANY type of analysis you need beyond what we provide above. Our prices are very fair, and we work very quickly to help you.**

**From advanced alpha and beta diversity, unifrac, rarefaction etc., qiime, analysis, biostatistical support all the way to full report, manuscript and interpretation service. We provide these additional services at a very low cost-recovery only, as a support for sequencing services.**

**MR DNA provides many free software modules to assist with data analysis. These can be found at** [**www.mrdnafreesoftware.com**](http://www.mrdnafreesoftware.com) **.**

**BASESPACE FOR ILLUMINA DATA. I can’t find my sequence data!**

Please note on basespace the raw data is found by clicking on the project ID ... then click on the link to the left which says “samples”. This will show you the raw sequence files. These files are fully raw and unbinned of course… (there is nothing of applicability in the “analysis” tab) please read further if you wish to use 16s data in applications such as qiime or mothur for instance. The fastq processor will create the types of files typically used with the qiime tutorials.

**USING RAW illumina data in qiime or mothur**. When using qiime on raw data I lose half my sequences.

FORMAT FOR QIIME OR MOTHUR: Typically our fastq files are fully RAW and untouched and unbinned etc. on basespace.. for applications like Qiime you may wish to have individual fastq and index fastq files.. EASY on our website [www.mrdnafreesoftware.com](http://www.mrdnafreesoftware.com) the **fastq processor** application can easily perform this task.. MR DNA also provides this service for $100. The **Fastq Processor** application which can take raw basespace files and convert them for use in the SRA e.g into individual fastq R1 and R2 for each sample along with index.fastq for each sample.

FASTQPROCESSOR IS VERY EASY: USE THE HELP FUNCTION IN THE SOFTWARE.. **Use the existing data from basespace R1 and R2 .. use the mapping.txt file that is found in your downloaded “fasta-qual-mapping-files” folder (this is one of the downloads that came with your original “data and invoice” email .**

Note BATCH SEQUENCE FILES: if sequences were run in batches (share barcodes with other samples) you will have multiple “fasta-qual-mapping-files” folders .. one for each batch all unique barcodes within a given batch. E.g. process each separately through first steps in qiime or mothur then merge the datasets (follow qiime applications guidance)

(note: each sample is sequenced as a paired-end set of reads (r1 and r2 files on basespace).. there are two paired read files.. that contain all of your samples .. the above method converts these to individual samples R1 and R2 and an index.fastq. ) **use the mapping2.txt file that is found in your downloaded “fasta-qual-mapping-files” folder (this is one of the downloads that came with your original “data and INVOICE” email ..**

**Troubleshooting 1) make sure your fastq files are unzipped and not still in .gz compressed format 2) make sure all the data and mapping files are located on your hardrive (not in a shared drive, not in a user directory e.g. desktop, download, documents, etc folder as these are not actual locations but symbolic links ) put the data onto your hardrive 3) make sure your computer has enough space to handle the amount of data (often terrabytes of free space may be necessary ) 4) if you have other problems send a screenshot of the problems.. keep in mind this software is used by thousands of researchers easily and successfully so it is usually something simple ie there is no problem with the software it is very good and functional.**

**MR DNA analysis pipeline overview.**

Data is subjected to an above average stringency quality control process. This process is good for most projects. Some projects may need higher stringency, while some projects can benefit from reducing stringency (e.g. chimera false positive and false negative rates). The following paragraph can be paraphrased to be included in publications. Please do not use verbatim! Please adapt and put into your own words!

bTEFAP®:

Barcoded amplicon sequencing processes described by many names and now performed by many technologies and by MR DNA® under the trademark service (bTEFAP®) was invented by Dr. Scot E Dowd back in 2007 and has been utilized in characterizing a wide range of environmental and health related microbiomes including the intestinal populations of a variety of sample types and environments, including cattle (SEE MANY REFERENCES BELOW). The bTEFAP® service was revamped and has continued to evolve and in a completely modified version of this original process and technologies including illumina, pac bio sequel, ion torrent.

**SRA submission and FASTQ files**

SRA also accepts individual fastq files. MR DNA provides free software to convert fasta/qual to fastq format allowing submission directly to SRA or other online data repository. To assist with this MR DNA has provided several free software modules available on [www.mrdnafreesoftware.com](http://www.mrdnafreesoftware.com) . One module will bin (using a mapping file) fasta and qual XXXXXXX-full.fasta (see below for details on all the files MR DNA provides) to produce individual fasta and qual files.. then a separate module can be used to convert these into individual fastq files.

There is also the Fastq Processor application OR fastq Splitter which can take raw basespace files and convert them for use the SRA into individual fastq for each sample. These are a proper format and FREE for submitting data to SRA. Both PGM and illumina data is submitted as fastq to NCBI also.

# Overviews of method for analysis: (this can roughly be paraphrased for the purpose of manuscripts and grants)

## MiSeq/Hiseq for amplicons

Methods of MiSeq or hiseq or novaseq6000 when run with amplicons!

The 16S rRNA gene V4 variable region PCR primers 515/806 (**OR whichever PRIMER set the investigator has SELECTED**) with barcode on the forward primer were used in a 30-35 (depends on primers and DNA..most studies use 30 cycles inquire from MR DNA as needed) PCR (5 cycles used on PCR products if you sent PCR products instead of gDNA) using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94°C for 3 minutes, followed by 30-35 cycles of 94°C for 30 seconds, 53°C for 40 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed. After amplification, PCR products are checked in 2% agarose gel to determine the success of amplification and the relative intensity of bands. Multiple samples are pooled together (e.g., 100 samples) in equal proportions based on their molecular weight and DNA concentrations. Pooled samples are purified using calibrated Ampure XP beads. Then the pooled and purified PCR product is used to prepare illumina DNA library. Sequencing was performed at MR DNA ([www.mrdnalab.com](http://www.mrdnalab.com), Shallowater, TX, USA) on a MiSeq following the manufacturer’s guidelines. Sequence data were processed using MR DNA analysis pipeline (MR DNA, Shallowater, TX, USA).  In summary, sequences are joined, sequences <150bp removed, and sequences with ambiguous base calls removed. Sequences are quality filtered using a maximum expected error threshold of 1.0 and dereplicated. The dereplicated or unique sequences are denoised; unique sequences identified with sequencing or PCR point errors are removed, followed by chimera removal, thereby providing a denoised sequence or zOTU. Final zOTUs were taxonomically classified using BLASTn against a curated database derived from NCBI (www.ncbi.nlm.nih.gov).

RAW DATA miseq and hiseq or novaseq: The raw data with Read1 and Read2 separate comes to you from a separate email .. this email comes from “basespace” not MR DNA.. If you do not get this email then often it is caught in a spam filter at your institution or in your email client.. PLEASE check here for this email at nearly the same time the data analysis and invoice were sent to you from MR DNA. If you do not receive these emails often they can be sent to a non-university email address such as gmail etc. JUST EMAIL US AND WE WILL HELP ☺

**USING RAW illumina data in qiime or mothur**.

Please note on basespace the raw data is found by clicking on the project ID .. then click on the link to the left which says “samples” .. this will show you the raw sequence files . these files are fully raw and unbinned of course.. please read further if you wish to use 16s data in applications such as qiime or mothur for instance.. the “fastq processor” will create the types of files expected by qiime and now Qiime2 (e.g. standard EMP format). Because we do not use the illumina basecalling software to bin (demultiplex) the raw data you will observe normal noise among the data. “fastq processor” can be used to create individual fastq R1 and R2 for each sample and index.fastq for each sample as noted below.

Typically our fastq files are untouched and un binned on basespace.. for applications like Qiime you may wish to have individual fastq and index files.. on our website [www.mrdnafreesoftware.com](http://www.mrdnafreesoftware.com) the fastq processor application can easily perform this task.. MR DNA also provides this service for $100. The Fastq Processor application which can take raw basespace files and convert them for use for functions such as submission to SRA, qiime, mothur etc.

(note: each sample is sequenced as a paired-end set of reads (r1 and r2 files on basespace).. there are two paired read files..

FORMAT OF RAW amplicon data on BASESPACE

1. To keep amplification bias to a minimum MR DNA does not use long concatamer primers as part of illumina data (ie 50bp of linker and barcode and a 20bp primer). We do create actual libraries out of each of our individual amplicons. This results in the amplicons being found in both 5’-3’ as usual and 3’-5’ orientation in the r1 and r2 files, this is normal for ligated libraries. Note the R1 and R2 are both in the 5’-3’ orientation as raw files.
   1. Forward primer format BARCODE-FORWARD PRIMER (can be found in R1 and R2)
   2. Reverse primer format REVERSE PRIMER (matched pair can be found in R1 and R2)

**Example of R1 and R2 format .. standard mixed pair format**

**R1 file**

**Sequence 1 barcode-forward primer-sequence**

**Sequence2 reverse primer-sequence**

**Sequence 3 barcode- forward primer- sequence**

**…etc**

**R2 file**

**Sequence1 reverse primer-sequence**

**Sequence 2 barcode- forward primer- sequence**

**Sequence 3 reverse- primer sequence**

**…etc**

1. For most assays MR DNA provides free joining of the reads as well as reorienting the 3’-5’ reads (after merging) into a uniform direction (these are the full.fasta and full.qual files found in the fasta-qual-mapping-files folders. Remember this is RAW data that has been joined using a standard method Usearch (for amplicons > 250 and < 570bp) e.g. 515-806 is <300 and ITS1-4 is variable with most sequences > 570) you can convert these to fastq as noted above.
2. To process the r1 and r2 files for amplicons >300bp and <570bp this is the process summarized that we use at MR DNA. There are other logical methods
   1. Join the reads together after q25 trimming of the ends (there are many publicly available softwares that join illumina paired end reads).
   2. Look for barcodes at the 5’ end, also find reverse compliment barcodes at the 3’ end of the joined reads.
   3. Reverse compliment the sequences containing barcodes at the 3’ end.
   4. The resulting file is our full.fasta and full.qual .. this has the joined reads all in the same 5’-3’ orientation. this is raw data just joined and reoriented.
   5. Please note you can convert the full.fasta and full.qual back to fastq using our free software [www.mrdnafreesoftware.com](http://www.mrdnafreesoftware.com)
3. To process the r1 and r2 files for amplicons <300bp.
   1. MR DNA in this case now **does** join the files we utilize the forward reads found in both r1 and r2 files. If you do not join the reads and utilize both the r1 and r2 by concatenating them or processing each in turn then all the barcodes do appear only in 5’-3’ orientation.
   2. Perform quality trimming.. keep all reads >250bp (subjective). And process as usual.
   3. The same basic process is used by MR DNA for amplicons too long to be joined effectively.
4. Please note that in your fasta-qual-mapping-files directory is the full.fasta and full.qual files

These files are simply merged and reoriented in the 5’-3’ direction and <200bp sequences removed. This is still (apart from joining the R1 and R2) raw data with the barcodes in them.. you can use thes files directly (similar to 454 or PGM data) along with the mapping2 file in secondary analysis software such as qiime or mothur. Or you can convert the full.fasta and full.qual to native fastq using free software found at [www.mrdnafreesoftware.com](http://www.mrdnafreesoftware.com) .

An issue with qiime and usually the 515F or ITS1 assays.. Qiime occasionally detects duplicate reads. This is usually a handful of definitions that cannot be distinguished. We have produced a software to clean the fasta and qual so qiime can get past this issue [www.mrdnafreesoftware.com](http://www.mrdnafreesoftware.com) . Our newer processes eliminate this need however so only older data has this problem.. We now offer free service to rectify this issue for qiime users ;-) just email Dr Dowd and we will clean the files for you no cost.

Remember if not joining the reads (which reorients the barcodes in the r2 files) both files can be used for finding indexes (barcodes) without reverse complementation. After binning the R1 file you can then bin the R2 file to identify all of the barcoded reads.. their pairs can be identified using the illumina definition line coding, reads can be joined with appropriate software if available.. MR DNA does now offer FREE joining of paired reads for amplicon sequencing as well as all of our comprehensive taxonomic analysis, advanced analysis and full report service. )

For assays that have amplicons >300bp and < 550bp joining is very efficient. For assays that are longer or shorter than this we will attempt joining. if joining is inefficient due to the size of the amplicon we will analyze unidirectional reads. **Note: MR DNA now DOES merge the 515-806 reads as well as many of our other assays < 300bp**

# Data processing (basic overview)

Do contact us if you want custom processing through out pipeline with other than our default settings and databases.. we are happy to help customize an analysis for your project

Please note this is a boiler plate generalize methods statement. Please paraphrase and contact us with questions of course!

The Q25 sequence data derived from the sequencing process was processed using the MR DNA ribosomal and functional gene analysis pipeline ([www.mrdnalab.com](http://www.mrdnalab.com) , MR DNA, Shallowater, TX). Sequences are depleted of primers, short sequences < 150bp are removed, and sequences with ambiguous base calls removed. Sequences are quality filtered using a maximum expected error threshold of 1.0 and dereplicated. The dereplicated or unique sequences are denoised; unique sequences identified with sequencing or PCR point errors are removed, followed by chimera removal, thereby providing a denoised sequence or zOTU. Final zOTUs were taxonomically classified using BLASTn against a curated database derived from NCBI ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) and compiled into each taxonomic level into both “counts” and “percentage” files. Counts files contain the actual number of sequences while the percent files contain the relative (proportion) percentage of sequences within each sample that map to the designated taxonomic classification. E.g. if there are 1000 sequences and 100 of the sequences are classified as *Staphylococcus* then we represent this as *Staphylococcus* being 10%.

We understand that taxonomic classification is both desirable for some and irrelevant to others, so we also define our data with operational taxonomic units. MR DNA provides qiime analysis of data for $10/sample (this is a price that basically helps us recover the cost of analysis, data storage and transfers).

The following folders and files are produced.  **NOTE: All files are tab-delimited text files. These can be imported or opened directly into excel. E.g. right-click on a file and using the “open-with” command in the right-click menu select open with “excel”.**

# Folders: (there is a base folder with ALL OTUs, there are also subfolders with more popular groups such as bacteria, archaea, fungi, separated out from the ALL OTUS)

1. **fasta-qual-mapping-files**: This folder contains the raw data for your sequencing project. The FASTQ data has been split into the respective fasta (FNA) and QUAL (quality) files. These have the typical extension ‘.fna’ and ‘.qual’. These are the text version of the sequence data and quality scores. There is also a mapping file in this directory. This is a normal text file which has each of the samples, barcode information, primer information and project name. Other in-house specific text files will be present, but the 4 main files will be the fasta file, the quality file, and the mapping file.
   * + 1. –full.fasta, -full.qual
          1. These files contain the raw sequence data information and still have primers and barcodes
       2. –pr.fasta, -pr.qual
          1. Sequencing information without the primer-barcode used in our analysis. The sample names from which each sequence is derived is encoded in the definition line of each sequence and short files < 150bp removed
       3. –mapping2.txt
          1. Contains:

Sample ID

Barcode

Primer

Barcode/Primer Name

ReversePrimer

Project Name

Description

1. **pipeline:** Within the pipeline folder are 6 different files, which are generated through various stages of our analysis pipeline.
   1. -pr.fastq
      1. Demultiplexed dataset stripped of forward and reverse primer sequences in fastq format.
   2. -filtered.fa
      1. Reads are filtered based on Q score and expected error probability and any read with a number of expected errors greater than 1.0 are discarded.
   3. -uniques.fa
      1. Dereplicated quality filtered reads
   4. -zotus.fa
      1. Denoised unique sequences; reads with sequencing or PCR errors are removed followed by the removal of chimeras
   5. -zotutab.txt
      1. A zOTU table created with the number of reads assigned to each zOTU in each sample. All reads, pre and post filtering are considered for zOTU table construction.
   6. -pr.fasta.readmap.uc
      1. zOTU read map
2. **AnalysisFiles**: The analysis folder of course contains the analysis or compilation files.

Key features: MR DNA does output many different types of files we like to point out a few of these

1. Species files - these are “force” to top hit species analyses. The 16s is not able to fully resolve classification to species and genera is often tentative as well. However, many want this type of output and so we do provide a force to closest or top hit species compilation
2. **IDENTITY files**- these are **a more valid type** of analysis file where the percent divergence dictates the level of classification reported in the compilation outputs. These are based on the classification divergence tables below.
3. SPECIAL NOTE on taxonomic level compilations .. files of interest
   * 1. SPECIES- this is a force to top hit species .. with short read data please do take special care using a species level analysis .. for publication this is not advised to rely on short read 16s data as a way to have valid species classification.. we provide this force to top hit as a specialized use only (due to popular request) .. GENUS level compilations are often okay to use .. but most relevant is the IDENTITY files which use the level of identity to judge at which taxonomic level the analysis for classification is most valid ..

# PRIMARY FOLDERS IN THE ANALYSIS FOLDER

## Directly in the analysis folder is the “ALL” hit or “All” OTUs data. Additionally, there will be a folder for each detected domain e.g. eukaryote, prokaryote, etc.

We do output compilations at each taxonomic level as well as zOTUs files with complete lineage and percent identity. It is important to note that upon opening the ‘AnalysisFiles’ folder you will find 31 files (**further explained below**) along with the individual domain folders. The main ‘AnalysisFiles’ folder contains overall counts files and percentages file from ***all detected taxonomic domains***. Increased specificity has been provided by dividing this data into individual domains where you will find these same 31 files for each specific taxonomic domain.

1. There are three sets of compiled files for each taxonomic level. These are the count (.txt) files, the percentage (.txt) files, and an excel workbook (.xlsx) that has both counts and percentages. The count file for kingdom level taxonomy for instance is “XXXXX-=pr.fasta.otus.fa.KINGDOM.TXT and XXXXX-=pr.fasta.otus.fa.KINGDOM.percentages.TXT” Obviously the file without percent is the counts file. This file has the number of sequences from which each taxonomic designation was derived. There are kingdom, phylum, family, order, class, genus, and species files. Each of these has a corresponding count and percent file associated with it..
2. There is also the OTU file which contains each zero-radius operational taxonomic units defined by the previously mentioned “XXXXXX-zotus.fa” pipeline file. These will have the full taxonomic lineage as well as the identity of the match to the BLASTn database.
3. There are also identity files. These will basically compile the zotus to the most relevant taxonomic level based upon the percent identity.

|  |  |
| --- | --- |
| **Identity to reference sequence** | **Identity Designation** |
| > 97% | Species |
| Between 97% and 95% | (unclassified Genus) |
| Between 95% and 90% | (unclassified Family) |
| Between 90% and 85% | (unclassified order) |
| Between 85% and 80% | (unclassified class) |
| Between 80% and 77% | (unclassified phylum) |
| < 77% | (unknown) |

* 1. ANALYSIS File List with description:
     1. .otus.fa.mapping.txt
        1. Extracted from the ucfile contains a list of
           1. Sample Name::Read ID
           2. Matching OTU
           3. OTU Homolog Reference Name
           4. OTU Homolog Percent ID
     2. .otus.fa.percentages.txt, .otus.fa.txt
        1. Contains:
           1. OTU Name
           2. Homolog information
           3. Sample Counts or Percentages
     3. .indentify.percentages.txt, .otus.fa.indentify.txt
        + 1. Homolog Information based on identity matrix
          2. Sample Counts or Percentages
          3. Taxonomic classification based upon identity as per table
     4. otus.fa.classs.percentages.txt, .otus.fa.classs.txt
     5. .otus.fa.family.percentages.txt, .otus.fa.family.txt
     6. .otus.fa.genus.percentages.txt, .otus.fa.genus.txt- to closes genera
     7. .otus.fa.kingdom.percentages.txt, .otus.fa.kingdom.txt
     8. .otus.fa.order.percentages.txt, .otus.fa.order.txt
     9. .otus.fa.phylum.percentages.txt, .otus.fa.phylum.txt
     10. .otus.fa.species.percentages.txt, .otus.fa.species.txt- to closest (best hit) species
         1. Contains:
            1. Individual’s (file name based) rank id
            2. Count or Percentage for Samples

REFERENCES

1. Changes in 16s RNA Gene Microbial Community Profiling by Concentration of Prokaryotic DNA. Glassing A, Dowd SE, Galandiuk S, Davis B, Jorden JR, Chiodini RJ.J Microbiol Methods. 2015 Dec;119:239-42. doi: 10.1016/j.mimet.2015.11.001. Epub 2015 Nov 10. PMID: 26569458

2. Natural mummification of the human gut preserves bacteriophage DNA. Santiago-Rodriguez TM, Fornaciari G, Luciani S, Dowd SE, Toranzos GA, Marota I, Cano RJ. FEMS Microbiol Lett. 2016 Jan;363(1). pii: fnv219. doi: 10.1093/femsle/fnv219. Epub 2015 Nov 11. PMID: 26564967

Select item 26467017

3. Gut Microbiome of an 11th Century A.D. Pre-Columbian Andean Mummy. Santiago-Rodriguez TM, Fornaciari G, Luciani S, Dowd SE, Toranzos GA, Marota I, Cano RJ. PLoS One. 2015 Sep 30;10(9):e0138135. doi: 10.1371/journal.pone.0138135. eCollection 2015. PMID: 26422376 Free PMC Article

4. Composition and Diversity of the Fecal Microbiome and Inferred Fecal Metagenome Does Not Predict Subsequent Pneumonia Caused by Rhodococcus equi in Foals. Whitfield-Cargile CM, Cohen ND, Suchodolski J, Chaffin MK, McQueen CM, Arnold CE, Dowd SE, Blodgett GP. PLoS One. 2015 Aug 25;10(8):e0136586. doi: 10.1371/journal.pone.0136586. eCollection 2015.

PMID: 26305682

5. Xylooligosaccharide supplementation alters gut bacteria in both healthy and prediabetic adults: a pilot study. Yang J, Summanen PH, Henning SM, Hsu M, Lam H, Huang J, Tseng CH, Dowd SE, Finegold SM, Heber D, Li Z. Front Physiol. 2015 Aug 7;6:216. doi: 10.3389/fphys.2015.00216. eCollection 2015. PMID: 26300782 F

6. Multiple Opportunistic Pathogens, but Not Pre-existing Inflammation, May Be Associated with Necrotizing Enterocolitis.Leach ST, Lui K, Naing Z, Dowd SE, Mitchell HM, Day AS. Dig Dis Sci. 2015 Dec;60(12):3728-34. doi: 10.1007/s10620-015-3830-6. Epub 2015 Aug 7. PMID: 26250832