

Table of Contents

- Methods and Overview
- ❖ A "How To" for BaseSpace and Illumina Data
 - Where Can I Find My Raw Data
 - Why Am I Receiving My Data From Dropbox Instead of BaseSpace?
- Frequently Asked Questions
 - Why Are There So Many Raw Files?
 - Why Are There So Many FASTQ Generation Folders?
 - ► How Do I remove the Forward and Reverse Primer Sequences from my Raw data??
 - ► How do I import my Raw data into Qiime2 or Mothur??
 - SRA submission and FASTQ files
- ❖ MR DNA analysis pipeline overview.
 - ▶ bTEFAP®
 - Methods of MiSeg or Novaseg6000 when run with amplicons!
- Data processing (basic overview)
- **❖** PRIMARY FOLDERS IN THE ANALYSIS FOLDER



Methods and Overview

MR DNA provides free comprehensive taxonomic analysis of data. This analysis is provided in tabdelimited 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, with secondary processes 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., Qiime2, 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 .



Where Can I Find My Raw Data???

RAW DATA NovaSeq / MiSeq / pac bio etc: comes in the data and invoice email under data downloads usually denoted as RAW data.

Why Am I Receiving My Data from Dropbox instead of MR DNA ftp?

- Under certain circumstances, it is beneficial for both MRDNA and our customer to provide the Raw Sequencing Data via DropBox instead of using ftp
- o If you have received data from MRDNA before, the DropBox link is similar to the Egnyte links you receive with your final analysis.
- There is no need for a DropBox account. By clicking the DropBox link, a web browser should appear and from there you will be able to download your data.



Frequently Asked Questions

Why are there so many Raw files for illumina??

- o The Raw Data you have received on is **demultiplexed**
- o A R1 and R2 fastq.gz file has been generated for each individual sample
 - All forward reads are binned into the R1 fastq.gz files
 - All reverse reads are binned into the R2 fastq.gz files
- Other than demultiplexing; you can consider the Raw Data on BaseSpace as untouched (The Forward and Reverse Primer Sequences e.g. 515F-806R have not been removed)

Why Are There So Many FASTQ_Generation Folders?

- Typically, after downloading your data, the user will see a single FASTQ_Generation folder containing one folder per sample with the raw R1 and R2 fastq files.
- Occasionally, a project will require additional sequencing. This additional sequencing is the reason for multiple FASTQ Generation folders.
 - Data is provided in a standard demux format (cassava single lane paired end for illumine). this is a standard input format for most software

How Do I remove the Forward and Reverse Primer Sequences from my Raw data??

Most analysis pipelines have a tool or method to remove primers (e.g. qiime2 has cutadapt)

How do I import my Raw data into Qiime2 or Mothur??

- o Qiime2
 - First Important Note: Your Data Is Already DEMULTIPLEXED!! Do NOT follow the "Moving Pictures" tutorial to import your data
 - Second Note: <u>It is recommended you remove the forward and reverse primer</u> sequences from your data
 - Select the "Learn More" button on the https://qiime2.org/ website
 - Under Tutorials, select Importing data
 - Scroll down to "Casava 1.8 paired-end demultiplexed fastq"
 - Follow the Importing Data instructions
 - The input path will be the path where your R1 and R2 fastq.gz files are located e.g. /home/.../qiime2/demux/
- Mothur



www.mrdnalab.com

806-789-7984

- Note: It is recommended you remove the forward and reverse primer sequences from your data
- Note: If you choose to remove primer sequences using an alternative method (); It is also important to take notice of the characters in your sample names.
 - Mothur does not accept sample names with colons, dashes, or forward slashes
 - Due to BaseSpace and In-House restrictions: MRDNA is limited to using dashes in place of spaces in sampleIDs i.e. Sample 1 -> Sample-1
 - It is necessary to replace the dashes in the sampleIDs before beginning Mothur
 - This process can be completed in addition to removing the primer sequences using the <u>FASTq Processor</u>
- Once your data is properly formatted, Follow the MiSeq tutorial found here:
 https://www.mothur.org/wiki/MiSeq SOP



SRA submission and FASTQ files?

SRA accepts submissions in paired-end FASTQ format directly as provided by MR DNA . MR DNA provides free software to convert a variety of other formats e.g. 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. One module will bin (using a mapping file) fasta and qual XXXXXXXX-pr.fasta or pr.qual (see below for details on all the files MR DNA provides) to produce individual fasta and qual files. Then a separate module (FASTA Qual & FASTQ Conversion) can be used to convert these into individual fastq files. There is also the Fastq Processor application which can take the raw demultiplexed basespace files, remove the forward and reverse primer sequences and thereby convert them for SRA Paired-End FASTQ upload. These are a proper format and FREE for submitting data to SRA.



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, PacBio Sequel, and Ion Torrent.

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

MiSeq/NovaSeq for amplicons:

Methods of MiSeq 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**) were used in a 30-35 cycle (depends on primers and DNA. Most studies use 30 cycles. Send inquiries to 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: 95°C for 5 minutes, followed by 30-35 cycles of 95°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 10 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. Samples are multiplexed using unique dual indices and 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 an Illumina DNA library. Sequencing was performed at MR DNA (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 were 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 and/or PCR point errors and 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).

Traditional OTUs vs zOTUs

Traditional OTUs are generated by a cluster of sequencing reads that have, in most cases, > 97% similarity to one another. A key reason many researchers are prompted to use zOTUs (zero-radius OTUs) or ASVs (Amplicon Single Variants), is the loss of biological information. In the case of zOTUs, all correct biological sequences are identified; distinguishing sequences with even a single difference. This level of specificity is not achievable when using a 97% identity threshold as with traditional OTUs, and as a result, different strains or species with closely matching sequences will be clustered together. Using zOTUs over traditional OTUs provides a greater resolution of all biological sequences.

Data processing (basic overview)

Do contact us if you would like to request custom processing throughout our pipeline outside of 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, 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) 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 qiime2 analysis of data for \$10/sample (this is a price that basically helps us recover the cost of analysis, data storage and transfers).

Summary of Data Folders:

fasta-qual-mapping-files:

This folder contains the raw data for your sequencing project. The FASTQ data has been split into the respective fasta (fasta) and QUAL (quality) files. These have the typical extension '.fasta' 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 sample IDs, barcode information, primer information and project name.

- 1. -full.fasta & full.qual
 - a. Full Sequencing Information: **INCLUDING** forward and reverse primer sequences
- 2. -pr.fasta & -pr.qual
 - a. **Primer Removed (PR)** Sequencing Information: reads after forward and reverse primer sequences have been removed.
 - The sample names from which each sequence is derived is encoded in the definition line of each sequence and short files < 150bp removed
- 3. -mapping.txt
 - a. Contains:
 - i. Sample ID
 - ii. Barcode
 - iii. Primer
 - iv. Barcode/Primer Name
 - v. ReversePrimer
 - vi. Project Name
 - vii. Description



pipeline:

Within the pipeline folder are 6 different files, which are generated through various stages of our analysis pipeline.

- b. -pr.fastq
 - i. Demultiplexed dataset stripped of forward and reverse primer sequences in fastq format.
- c. -filtered.fa
 - i. 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.
- d. -uniques.fa
 - i. Dereplicated quality filtered reads
- e. -zmap.txt
 - i. zOTU read map
- f. -zotus.fa
 - i. Denoised unique sequences; reads with sequencing or PCR errors are removed followed by the removal of chimeras
- g. -zotutab.txt
 - 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.



AnalysisFiles:

- The analysis folder of course contains the taxonomic analysis or compilation files.
- Folder Summary:
 - 1. AllDomains
 - a. This Folder contains ALL "hits" or ALL zOTU matches for each taxonomic domain e.g. Eukaryote AND Prokaryote
 - 2. Eukaryote
 - a. This folder contains "hits" / zOTU matches for the eukaryote domain ONLY
 - 3. Prokaryote
 - a. This folder contains "hits" / zOTU matches for the prokaryote domains ONLY
- Compilations are completed at each taxonomic level e.g. kingdom, phylum, class, etc.
- ❖ SPECIAL NOTE on taxonomic level compilations .. files of interest
 - 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 ..
- **
- ➤ There are 3 sets of complied files for each taxonomic level:
 - 1. Counts .txt file e.g. -zotus.fa.kingdom.txt (text file, open using notepad or excel)
 - a. Number of reads per sample per zOTU
 - Percentages .txt file e.g. -zotus.fa.kingdom.percentages (text file, open using notepad or excel)
 - a. Relative abundance of reads per sample per zOTU
 - Excel Workbook; Counts and Percentages e.g. -zotus.fa.kingdom.xlsx (open using MS excel)
- Also included in each folder: The OTU file (-zotus.fa.OTU...) which contains each zero-radius operational taxonomic unit defined by the previously mentioned "-zotus.fa" pipeline file.
 - These will have the full taxonomic lineage as well as the identity of the match to the BLASTn database.



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



- An alternative to Species files are the Identity files-
 - The identity files 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.

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)

- The zotus.fa.mapping.txt file is extracted from the -zmap.txt file (pipeline folder)
 - Contains a list of:
 - Sample Name::Read ID
 - Matching OTU
 - OTU Homolog Reference Name
 - OTU Homolog Percent ID



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



Additional References

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

Similar articles

Select item 26240200 9.

<u>Inability of Polymerase Chain Reaction, Pyrosequencing, and Culture of Infected and Uninfected Site Skin Biopsy Specimens to Identify the Cause of Cellulitis.</u>

Crisp JG, Takhar SS, Moran GJ, Krishnadasan A, **Dowd SE**, Finegold SM, Summanen PH, Talan DA; EMERGEncy ID Net Study Group.

Clin Infect Dis. 2015 Dec 1;61(11):1679-87. doi: 10.1093/cid/civ655. Epub 2015 Aug 3.

PMID:

26240200

Similar articles

Select item 26227608 10.

<u>Draft Genome Sequence of Streptomyces sp. Strain PBH53, Isolated from an Urban</u> Environment.

Gosse JT, Hill P, Dowd SE, Boddy CN.

Genome Announc. 2015 Jul 30;3(4). pii: e00859-15. doi: 10.1128/genomeA.00859-15.



PMID:

26227608

Free PMC Article

Similar articles

Select item 26222621 11.

<u>Microbial Population Differentials between Mucosal and Submucosal Intestinal Tissues in</u> Advanced Crohn's Disease of the Ileum.

Chiodini RJ, **Dowd SE**, Chamberlin WM, Galandiuk S, Davis B, Glassing A.

PLoS One. 2015 Jul 29;10(7):e0134382. doi: 10.1371/journal.pone.0134382. eCollection 2015.

PMID:

26222621

Free PMC Article

Similar articles

Select item 26021918 12.

<u>Draft Genome Sequence of Vibrio sp. Strain Vb278, an Antagonistic Bacterium Isolated from the Marine Sponge Sarcotragus spinosulus.</u>

Gonçalves AC, Franco T, Califano G, **Dowd SE**, Pohnert G, Costa R.

Genome Announc. 2015 May 28;3(3). pii: e00521-15. doi: 10.1128/genomeA.00521-15.

PMID:

26021918

Free PMC Article

Similar articles

Select item 26020315 13.



MOLECULAR RESEARCH, LP
RAPID DISCOVERY, MAINTAINED INTEGRITY
www.mrdnalab.com 806-789-7984
Evaluation of soluble corn fiber on chemical composition and nitrogen-corrected true
metabolizable energy and its effects on in vitro fermentation and in vivo responses in dogs.
Panasevich MR, Kerr KR, Serao MC, de Godoy MR, Guérin-Deremaux L, Lynch GL, Wils
D, Dowd SE , Fahey GC, Swanson KS, Dilger RN.
J Anim Sci. 2015 May;93(5):2191-200. doi: 10.2527/jas.2014-8425.
PMID:
T WILD.
26020315
20020313
Similar articles
Similar articles
Select item 25946040 14.
Laurinia alastabilias and multi anasia biofilm announties developed in burnar adius
L-arginine destabilizes oral multi-species biofilm communities developed in human saliva.
Kolderman E, Bettampadi D, Samarian D, Dowd SE , Foxman B, Jakubovics NS, Rickard AH.
PLoS One. 2015 May 6;10(5):e0121835. doi: 10.1371/journal.pone.0121835. eCollection 2015.
PMID:
25946040
Free PMC Article
Troo Filio Artiolo
Similar articles
Select item 25934152 15.
Differences in the feed microbiams of non diarrhapia clinically healthy dogs and gets
Differences in the faecal microbiome of non-diarrhoeic clinically healthy dogs and cats
associated with Giardia duodenalis infection: impact of hookworms and coccidia.
Slapeta J, Dowd SE , Alanazi AD, Westman ME, Brown GK.
Int J Parasitol. 2015 Aug;45(9-10):585-94. doi: 10.1016/j.ijpara.2015.04.001. Epub 2015 Apr 29.
PMID:
25934152

Similar articles

Select item 25903104 16.



RAPID DISCOVERY, MAINTAINED INTEGRITY
Loss of n-6 fatty acid induced pediatric obesity protects against acute murine colitis.
Nagy-Szakal D, Mir SA, Harris RA, Dowd SE , Yamada T, Lacorazza HD, Tatevian N, Smith
CW, de Zoeten EF, Klein J, Kellermayer R.
FASEB J. 2015 Aug;29(8):3151-9. doi: 10.1096/fj.14-267690. Epub 2015 Apr 22.
PMID:
05002404
25903104
Similar articles
Select item 25816691 17.
Select item 250 1009 i 17.
Comparative accessory gene fingerprinting of surface water Escherichia coli reveals genetically
diverse naturalized population.
Tymensen LD, Pyrdok F, Coles D, Koning W, McAllister TA, Jokinen CC, Dowd SE , Neumann
NF.
J Appl Microbiol. 2015 Jul;119(1):263-77. doi: 10.1111/jam.12814. Epub 2015 Apr 20.
PMID:
25816691
Similar articles
Select item 25663091 18.
Matawayayaia ayalyaia af tha ayat yaisyahiata af tha Tiyahay Dattleayaka Quatalya hayyidya
Metagenomic analysis of the gut microbiota of the Timber Rattlesnake, Crotalus horridus. McLaughlin RW, Cochran PA, Dowd SE .
Mol Biol Rep. 2015 Jul;42(7):1187-95. doi: 10.1007/s11033-015-3854-1. Epub 2015 Feb 7.
PMID:
25663091
Observation and the control of the c
Similar articles

Select item 25593266^{ll} 19.

<u>Draft Genome Sequence of Robinsoniella peoriensis Strain WTD, Isolated from the Fecal</u> Material of a Wood Turtle.



Braasch JL, Lapin CN, Dowd SE, McLaughlin RW.

Genome Announc. 2015 Jan 15;3(1). pii: e01444-14. doi: 10.1128/genomeA.01444-14.

PMID:

25593266

Free PMC Article

Similar articles

Select item 25587784 20.

<u>Listeria monocytogenes batch culture growth response to metabolic inhibitors.</u>

Milillo SR, Lungu B, O'Bryan CA, **Dowd SE**, Muthaiyan A, Johnson MG, Ricke SC.

 $\label{eq:control_scale} \mbox{J Environ Sci Health B. 2015;} 50(2):146-50. \mbox{ doi: } 10.1080/03601234.2015.975626.$

PMID:

25587784

Similar articles