# PROTOCOL FOR MULTIPLEX MICROBIOTA USING ILLUMINA MISEQ

GREG GLOOR

## 1. Background

1.1. **MiSeq.** The MiSeq instruments use the first 12 positions in the read for normalization, and appear to require significant complexity in order to separate spots efficiently. The primers used for the MiSeq are composed of 4 random nucleotides at the 5' end followed by 8mer barcodes that are sequence composition balanced with a minimum edit distance of 3. Primers for the V6 and V4 rRNA gene variable regions are given at the end of this document. With this strategy, we have successfully run as few as 4 multiplexed samples with 5% or less ΦX174 spiked in. We have not had success with 12mer barcodes containing equal nucleotide compositions in the absence of the 4 random nucleotides at the 5' end. We have not tested shorter or longer segments of random sequence. Figure 1 shows a schematic of the primer and barcode structure.

**Left-Illumina-adaptornnnnccaaggttLeft-primer**

**Right-Illumina-adaptornnnnccaaggttRight-primer**

FIGURE 1. Structure of the barcoded amplification primers. The 5‘ end of each primer contains the left- or right-side Illumina adaptor (black), this is followed by four degenerate nucleotides (dark blue), then by the 8-mer barcode (red) and finally the amplification primer (light blue).

## 2. Protocol

The protocol below is for primers that contain Illumina adapter sequences attached to the 5' end as given in the primer sequences.

*Important: do not size select the library without knowing the exact range of amplicon sizes. We do not size select our amplimers prior to loading on an Illumina MiSeq.*

2.1. **Amplification 1:** Taq is GoTaq hot start 2X colorless master mix from Promega (Catalogue numbers M5131, M5132, M5133).

Primer sequences for the rRNA V4 and V6 gene fragments are given at the end of this file. It is possible to replace these primer sequences with others specific to any desired amplicon. We have used this strategy to amplify single gene sequences from plasmids with high quality reads [1]. Primer stock solutions for long-term storage are made to 200 pMole/$\mu$l (200 $\mu$M) in deionized water and stored at -80°C. Prior to use, they are diluted to 3.2 pMole/$\mu$l in deionized water by adding 3.2 $\mu$l of concentrated stock to 197 $\mu$l of deionized water. The diluted stock is stable for several freeze-thaw cycles and several months at -20°C. PCR reactions are assembled as follows

(1) 50 $\mu$l light mineral oil
(2) 1 $\mu$l of input DNA
(3) 10 $\mu$l of primer A at 3.2 pMole/$\mu$l
(4) 10 $\mu$l of primer B at 3.2 pMole/$\mu$l
(5) heat 85°C prior to adding the GoTaq
(6) add 20 $\mu$l of GoTaq Master mix, heat to 95°C for 3 minutes to activate the GoTaq

cycling conditions are 1 minute each at 95°, 55° and 72° C, for the V6 primer. An annealing temperature of 52° is used for the V4 primers.

*We normally cycle for 25 cycles to reduce chimera formation and partial products. A test amplification should be conducted to ensure that plateau is reached with this number of cycles: in general 25 cycles is more than sufficient. Aliquots of random samples should be run on agarose gels to ensure that the reactions proceeded as planned.*

2.2. **Quantitation and pooling:** The most reliable method is to quantitate using the Qubit dsDNA kit. In this case the amplified product must be greater than 5X the negative control amplifications to be used. It is preferable to include two negative control reactions, one that was cycled, and one that was not. Both should have substantially the same reading. If the cycled negative control more than 25% greater than the non-cycled negative control, then steps to determine sources of contamination must be taken. The negative control readings are subtracted from each QuBit reading. Samples are pooled using their corrected relative concentration. The easiest way to do this is to add 1ul of the most concentrated sample and scale the volumes up for the other samples as needed. Pooled samples are mixed thoroughly, and then an aliquot (50-100 ul) is purified on a PCR cleanup column (Promega, Qiagen and Stratagene kits have been used successfully).

In the case of primers without these adapters the second amplification is not done immediately, instead the pooled, purified library has the Illumina adapters added by ligation using the Illumina paired-end protocol starting at the 3'A addition step. These primers are then amplified using the second stage amplification given below. This has been done at the the Centre for Applied Genomics (Toronto).

[1]McMurrough et. al. Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. PNAS June 10, 2014 vol. 111 no. 23 E2376-E2383

2.3. **Amplification 2:** The purified pooled libary is diluted 100 fold in water and amplified with primers:

```
OLJ139:
5AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA3
OLJ140:
5CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAAC3
```

Here the amplification is performed for 10 cycles using the same conditions as above, except the annealing temperature is increased to 60 degrees. If no discernible band is found it is acceptable to increase the number of cycles to 15.

2.4. **Sequencing instructions:** Samples are purified using the preferred PCR cleanup kit, quantitated vs. a negative control and sent to the genome centre for sequencing. Tell them that the library is already made and size selection is not required. They will want to know length of the amplimer and the attached adaptors.

For V6 amplimers, ask for a paired-end run with 2x100 cycles.

For V4 amplimers, ask for a paired-end run with 2x200 cycles using the 600 cycle kit.

## 3. Computational biology methods

current source is cjelli/git/miseq_bin.git. Working copy is in cjelli/Groups/LRGC/miseq_bin

3.1. **Requirements.** You will need an OS X or linux machine with 32 Gb of RAM for the later steps.

(1) bash and awk
(2) Pandaseq
(3) USEARCH: http://drive5.com/usearch/ #the latest 32 bit free version is fine
(4) mothur: http://http://www.mothur.org
(5) silva reference files in mothur format: http://www.mothur.org/wiki/Silva_reference_files

The directory structure on the machine that the scripts expect is below (the analysis directory will be created):

(1) Illumina_bin - location of all scripts and programs. you must know where this is and set it in the workflow.sh script
(2) reads - contains the raw fastq and the overlapped fastq made by pandaseq
(3) data_something - contains all intermediary data as outlined below, usually something is the variable region, or person's name, or experiment
(4) the samples.txt file must be in the same directory as the reads directory and workflow.sh
(5) analysis_something - contains the final read tables and OTU fasta files

The samples.txt file contains information about the sample IDs and the barcodes used. The format is tabbed, plain text, Unicode UTF-8 encoding.

| BC_L | BC_R | sample | Lpri | Rpri | Group |
|---|---|---|---|---|---|
| ccttggaa | ccaaggtt | Extraction Control | V4L5 | V5R1 | expt1 |
| ccttggaa | aaggttcc | KG04_01 | V4L5 | V5R2 | expt2 |

Step 1: Download and de-compress the MiSeq reads. This is best done from the Illumina Basespace site, ask for access when you do your run. Place the reads into the reads/directory. Reads are compressed with 7Zip: from the command line:

```
7z e  filename
```

Step 2: Overlap the reads with pandaseq. An example command for this with a minimum overlap of 30 nucleotides is below. This command is appropriate for the V4 amplimers:

```
pandaseq -g log.txt -T 8 -f L001_R1_001.fastq -r L001_R2_001.fastq -o 30
    -w ps_overlapped30.fastq -F &
```

Step 3: Run the workflow pipeline:

```
./workflow.sh name cluster_pid variable_region

./workflow_uc7.sh expt1 0.97 V4EMB
```

3.2. **What is happening behind the scenes:**
extract out the barcodes and primers associated with a particular samples.txt file. Output is a tabbed format file with the fields: read ID, sampleID primer sequence primer barcode q-score

```
$BIN/process_miseq_reads.pl $BIN samples.txt reads/overlapped.fastq $primer 8 0
    $name T > $rekeyedtabbedfile
```

make a fasta file of all identical sequences (ISU), and an index of those sequences

```
$BIN/group_gt1.pl $rekeyedtabbedfile $name
```

cluster at 97% identity using usearch (i.e., make OTU), also performs chimera filter singleton reads are excluded

```
$BIN/usearch7.0.1090_i86osx32 -cluster_otus ...
$BIN/usearch7.0.1090_i86osx32 -usearch_global ...
```

regenerate the tabbed reads file with each read tagged as to its OTU and ISU group membership

```
$BIN/map_otu_isu_read_us7.pl $c95file $reads_in_groups_file $rekeyedtabbedfile > $mapped
```

make two tables of counts in the analysis directory for OSU and ISU sequences gather the seed sequences for each OTU. Transpose the dataset for ease of import into QIIME

```
$BIN/get_tag_pair_counts_ps.pl $mappedfile $CUTOFF $name
$BIN/get_seed_otus_uc7.pl $c95file $groups_fa_file analysis_$name/OTU_tag_mapped.txt
     > analysis_$name/OTU_seed_seqs.fa
Rscript $BIN/OTU_to_QIIME.R analysis_$name
```

Use mother (must be installed separately) to annotate the OTU sequences against the silva database

```
$MOTHUR "#classify.seqs(fasta=analysis_$name/OTU_seed_seqs.fa, template=$TEMPLATE,
    taxonomy=$TAXONOMY, cutoff=70, probs=T, outputdir=analysis_$name, processors=4)"
$BIN/add_taxonomy_mothur.pl $TAX_FILE analysis_$name/td_OTU_tag_mapped.txt >
    analysis_$name/td_OTU_tag_mapped_lineage.txt
```

## 4. V6 barcodes, sets of 4 are balanced

V6Ll1 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnccaaggttCWACGCGARGAACCTTACC
V6Ll2 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnaaggttccCWACGCGARGAACCTTACC
V6Ll3 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnggttccaaCWACGCGARGAACCTTACC
V6Ll4 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnttccaaggCWACGCGARGAACCTTACC
V6Ll5 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnccttggaaCWACGCGARGAACCTTACC
V6Ll6 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnttggaaccCWACGCGARGAACCTTACC
V6Ll7 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnggaaccttCWACGCGARGAACCTTACC
V6Ll8 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnaaccttggCWACGCGARGAACCTTACC
V6Ll9 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnntccgttcgCWACGCGARGAACCTTACC
V6Ll10 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnngaatccatCWACGCGARGAACCTTACC
V6Ll11 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnncttaggtcCWACGCGARGAACCTTACC
V6Ll12 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnaggcaagaCWACGCGARGAACCTTACC
V6Ll13 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnntggcttcgCWACGCGARGAACCTTACC
V6Ll14 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnncaatggatCWACGCGARGAACCTTACC
V6Ll15 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnngttacctgCWACGCGARGAACCTTACC
V6Ll16 ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnaccgaacaCWACGCGARGAACCTTACC

V6Rl1 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnccaaggttACRACACGAGCTGACGAC
V6Rl2 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnaaggttccACRACACGAGCTGACGAC
V6Rl3 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnggttccaaACRACACGAGCTGACGAC
V6Rl4 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnttccaaggACRACACGAGCTGACGAC
V6Rl5 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnccttggaaACRACACGAGCTGACGAC
V6Rl6 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnttggaaccACRACACGAGCTGACGAC
V6Rl7 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnggaaccttACRACACGAGCTGACGAC
V6Rl8 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnaaccttggACRACACGAGCTGACGAC
V6Rl9 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnntccgttcgACRACACGAGCTGACGAC
V6Rl10 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnngaatccatACRACACGAGCTGACGAC
V6Rl11 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnncttaggtcACRACACGAGCTGACGAC
V6Rl12 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnaggcaagaACRACACGAGCTGACGAC
V6Rl13 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnntggcttcgACRACACGAGCTGACGAC
V6Rl14 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnncaatggatACRACACGAGCTGACGAC
V6Rl15 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnngttacctgACRACACGAGCTGACGAC
V6Rl16 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnnaccgaacaACRACACGAGCTGACGAC

## 5. V4 Earth Microbiome barcodes, sets of 4 are balanced

V4L1 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNccaaggttGTGCCAGCMGCCGCGGTAA
V4L2 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNaaggttccGTGCCAGCMGCCGCGGTAA
V4L3 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNggttccaaGTGCCAGCMGCCGCGGTAA
V4L4 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNttccaaggGTGCCAGCMGCCGCGGTAA
V4L5 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNccttggaaGTGCCAGCMGCCGCGGTAA
V4L6 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNttggaaccGTGCCAGCMGCCGCGGTAA
V4L7 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNggaaccttGTGCCAGCMGCCGCGGTAA
V4L8 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNaaccttggGTGCCAGCMGCCGCGGTAA
V4L9 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNtccgttcgGTGCCAGCMGCCGCGGTAA
V4L10 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNgaatccatGTGCCAGCMGCCGCGGTAA
V4L11 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNcttaggtcGTGCCAGCMGCCGCGGTAA
V4L12 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNaggcaagaGTGCCAGCMGCCGCGGTAA
V4L13 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNtggcttcgGTGCCAGCMGCCGCGGTAA
V4L14 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNcaatggatGTGCCAGCMGCCGCGGTAA
V4L15 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNgttacctgGTGCCAGCMGCCGCGGTAA
V4L16 ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNaccgaacaGTGCCAGCMGCCGCGGTAA

V5R1 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNccaaggttGGACTACHVGGGTWTCTAAT
V5R2 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNaaggttccGGACTACHVGGGTWTCTAAT
V5R3 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNggttccaaGGACTACHVGGGTWTCTAAT
V5R4 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNttccaaggGGACTACHVGGGTWTCTAAT
V5R5 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNccttggaaGGACTACHVGGGTWTCTAAT
V5R6 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNttggaaccGGACTACHVGGGTWTCTAAT
V5R7 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNggaaccttGGACTACHVGGGTWTCTAAT
V5R8 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNaaccttggGGACTACHVGGGTWTCTAAT
V5R9 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNtccgttcgGGACTACHVGGGTWTCTAAT
V5R10 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNgaatccatGGACTACHVGGGTWTCTAAT
V5R11 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNcttaggtcGGACTACHVGGGTWTCTAAT
V5R12 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNaggcaagaGGACTACHVGGGTWTCTAAT
V5R13 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNtggcttcgGGACTACHVGGGTWTCTAAT
V5R14 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNcaatggatGGACTACHVGGGTWTCTAAT
V5R15 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNgttacctgGGACTACHVGGGTWTCTAAT
V5R16 CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNaccgaacaGGACTACHVGGGTWTCTAAT