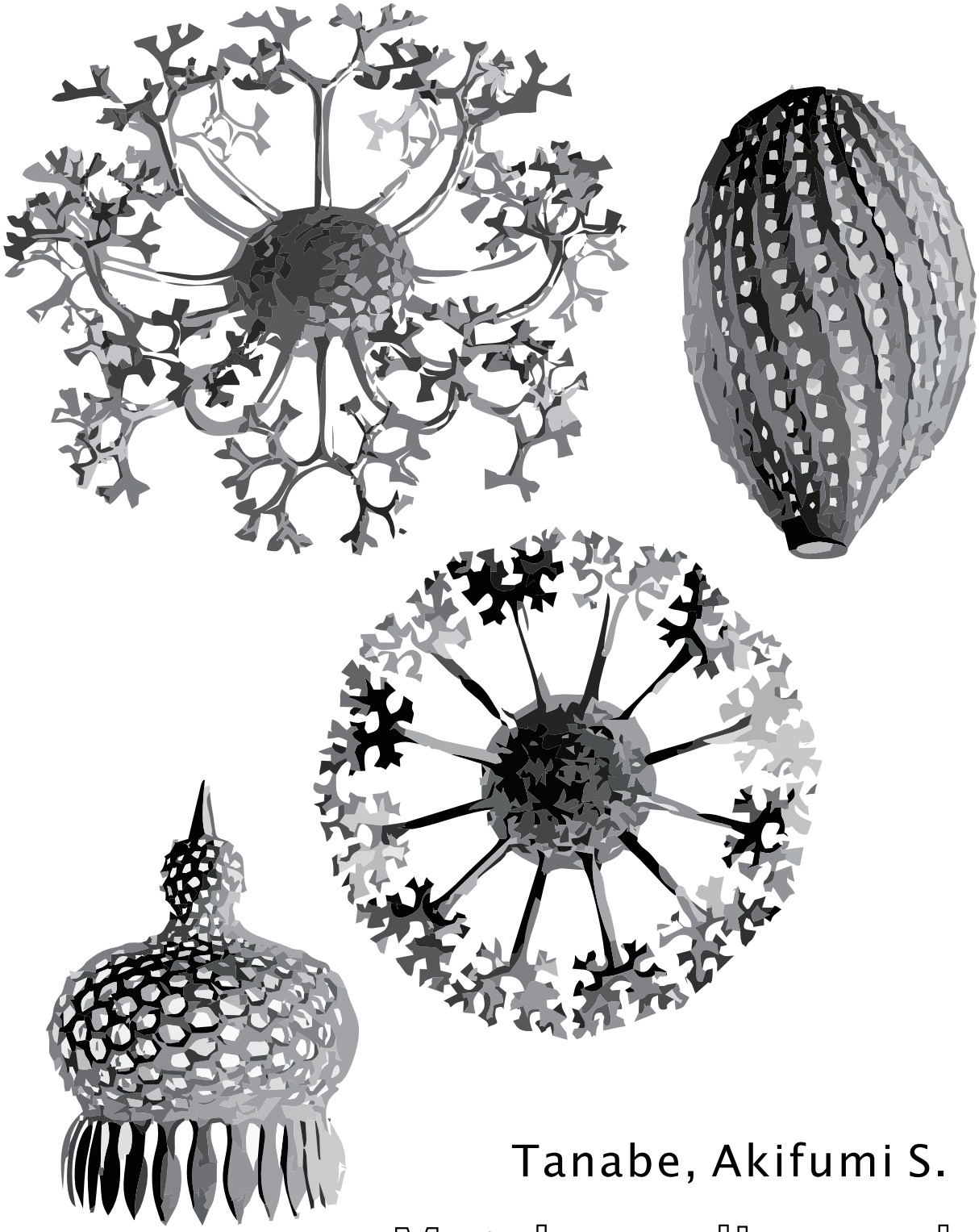

Tanabe, Akifumi S.

Metabarcoding and DNA barcoding for Ecologists

Matabarcoding and DNA barcoding for Ecologists

Akifumi S. Tanabe

September 16, 2016

Table of Contents

Preface

I started the writing of this text to promote amplicon sequence assember "Assams" and utility programs package "Claident" for sequence clustering and identification. In this text, I explain not only usage of these programs but also the methods and procedures of data collection and taxonomic identification by DNA barcoding.

Metabarcoding has been already widely used in bacterial reseaches, but it's utility is not limited to bacteria. Metabarcoding is also applicable to soil fungi, fungi inhabiting the bodies of animals and plants, aquatic planktons, and environmental DNA emitted from macro-organisms. I explain the methods of amplification of barcoding locus from environmental DNA or metagenomes, sequencing by high-throughput sequencers (a.k.a next-generation sequencers), taxonomic assignment of nucleotide sequences (a.k.a. DNA barcoding), and observation of presence/absence of operational taxonomic units.

I hope that this text helps you. I am grateful to Dr. Hirokazu Toju (Graduate Shool of Human and Environmental Sciences, Kyoto University), Dr. Satoshi Nagai (National Research Institute of Fisheries Science, Japan Fisheries Research and Education Agency), and you.

Legends

In this text, the input commands to terminals and display outputs are described as below.

```
# comments
> command argument1 \
argument2 \
argument3 ↓
output of command
> command argument1 argument2 argument3 ↓
output of command
```

In the above example, the same commands `command argument1 argument2 argument3` were executed twice. The outputs `output of command` were displayed after execution. The characters between # and line feed were comments and needless to input. > and space of line head indicate the prompt of terminal. Do not type these characters. ↓ means the end of input commands and arguments and needless to input, but you need to type Enter key to input line feed. I use line feed within commands or arguments for viewability. Such line feed is led by \. Therefore, the line feeds led by \ do not mean the end of commands or arguments, or designation to input Enter key. Involuntary line feeds may be generated by word wrap function depending on your read environment, but do not mean the end of commands or arguments, or designation to input Enter key.

The file content is shown as below in this text.

```
| The content of first line
| The content of second line
```

| and space of line head indicate the line head in the file, do not exist in the file and needless to input these characters. This code is written to help you to distinguish true line feeds and involuntary line feeds.

Chapter 0

Installing softwares and preparing analysis environment

In this text, I assume Debian GNU/Linux jessie (hereafter Debian) or Ubuntu Linux 14.04 LTS (hereafter Ubuntu) as operating system. If you use Windows PC, please install Debian or Ubuntu. Cygwin or Windows Subsystem for Linux provided for Windows 10 build 14393 or later can be used for the following analysis, but the programs run much more slowly. You can use CD, DVD or USB memory to boot installer of Linux. If your PC has only one storage device, you need to reduce Windows partition by using partition resizer software such as EaseUS Partition Master or using a partition resize function contained in the installer. You can also use newly added internal storage devices or external storage devices connected by USB. There are several variations of Ubuntu, and I recommend Xubuntu rather than normal Ubuntu.

Debian and Ubuntu can be installed to Mac. If there is no enough space, you need to resize OSX partition with the aid of Disk Utility or add storage device. The rEFIt or rEFInd boot selecter may be required to boot Debian, Ubuntu or the installer of them on Mac. If you install rEFIt or rEFInd to your Mac, you can boot the installer of Debian or Ubuntu from CD, DVD or USB memory. Do not delete existing partition of OSX. If you have enough free space, you don't need to use Disk Utility to resize existing partition. You can install Debian or Ubuntu to external storage devices on Mac.

I assume Intel64/AMD64 (x86_64) CPU machine as analysis environment. The other CPU machine can be used for analysis, but you need to solve problems by yourself. The 64 bits version of Debian or Ubuntu is also required because 32 bits version cannot use large memory.

0.1 Installation of Claident, Assams, databases, and the other required programs

Run the following commands in terminal or console as the user that can use `sudo`. Then, all of the required softwares will be installed. The installer will ask password to you when `sudo` is used.

```
> mkdir -p ~/workingdirectory ↓
> cd ~/workingdirectory ↓
> wget https://www.claident.org/installClaident_Debian.sh ↓
> sh installClaident_Debian.sh ↓
> wget https://www.claident.org/installOptions_Debian.sh ↓
> sh installOptions_Debian.sh ↓
> wget https://www.claident.org/installDB_Debian.sh ↓
> sh installDB_Debian.sh ↓
> wget https://www.claident.org/installUCHIMEDB_Debian.sh ↓
> sh installUCHIMEDB_Debian.sh ↓
> cd .. ↓
> rm -r workingdirectory ↓
```

By default, the softwares will be installed to `/usr/local`. In the installation, you will see `Permission denied` error and the installer ask password to you. If the installer continue after password input, you don't need to care about the error. The installer try to install without `sudo` at first and the installation output the above error. Then, the installer try to install using `sudo`.

If you need proxy to connect the internet, execute the following commands to set environment variables before execution of the installer.

```
> export http_proxy=http://server.address:portnumber ↓
> export https_proxy=$http_proxy ↓
> export ftp_proxy=$http_proxy ↓
> export all_proxy=$http_proxy ↓
```

If the proxy requires username and password, execute the following commands instead of the above commands.

```
> export http_proxy=http://username:password@server.address:portnumber ↓
> export https_proxy=$http_proxy ↓
> export ftp_proxy=$http_proxy ↓
> export all_proxy=$http_proxy ↓
```

0.1.1 Upgrading to new version

If you want to upgrade all of the softwares and the databases, run the same commands as initial installation. By this procedure, Assams, Claident, PEAR, VSEARCH, Metaxa and ITSx will be installed to `/usr/local`, and NCBI BLAST+, BLAST databases for molecular identification, taxonomy databases and the other required programs will be installed to `/usr/local/share/claident`. NCBI BLAST+ and BLAST databases used by Claident can co-exist system wide installation of NCBI BLAST+ and BLAST databases.

You can disable a part of upgrade like below.

```
> mkdir -p ~/workingdirectory ↓
> cd ~/workingdirectory ↓
# disable upgrade of Assams
> touch .assams ↓
```

```
# disable upgrade of Claident
> touch .claident ↓
# disable upgrade of PEAR
> touch .pear ↓
# disable upgrade of VSEARCH
> touch .vsearch ↓
# disable upgrade of NCBI BLAST+
> touch .blast ↓
# execute upgrade > wget https://www.claident.org/installClaident_Debian.sh ↓
> sh installClaident_Debian.sh ↓
# disable upgrade of sff_extract
> touch .sffextract ↓
# disable upgrade of HMMer
> touch .hmmer ↓
# disable upgrade of MAFFT
> touch .mafft ↓
# disable upgrade of Metaxa
> touch .metaxa ↓
# disable upgrade of ITSx
> touch .itsx ↓
# execute upgrade > wget https://www.claident.org/installOptions_Debian.sh ↓
> sh installOptions_Debian.sh ↓
# disable upgrade of ‘‘overall’’ BLAST and taxonomy databases
> touch .overall ↓
# execute upgrade > wget https://www.claident.org/installDB_Debian.sh ↓
> sh installDB_Debian.sh ↓
# disable upgrade of ‘‘rdp’’ reference database for chimera detection
> touch .rdp ↓
# disable upgrade of ‘‘unite’’ reference databases for chimera detection
> touch .unite ↓
# execute upgrade > wget https://www.claident.org/installUCHIMEDB_Debian.sh ↓
> sh installUCHIMEDB_Debian.sh ↓
> cd .. ↓
> rm -r workingdirectory ↓
```

0.1.2 Installing to non-default path

If you install the softwares based on the above procedure, the softwares will be installed to `/usr/local`. The executable commands will be installed to `/usr/local/bin`. You can change these install path for coexistence with the other programs such as older versions like below.

```
> mkdir -p ~/workingdirectory ↓
> cd ~/workingdirectory ↓
> export PREFIX=install_path ↓
> wget https://www.claident.org/installClaident_Debian.sh ↓
> sh installClaident_Debian.sh ↓
> wget https://www.claident.org/installOptions_Debian.sh ↓
> sh installOptions_Debian.sh ↓
> wget https://www.claident.org/installDB_Debian.sh ↓
> sh installDB_Debian.sh ↓
> wget https://www.claident.org/installUCHIMEDB_Debian.sh ↓
> sh installUCHIMEDB_Debian.sh ↓
> cd .. ↓
```

```
> rm -r workingdirectory ↓
```

In this case, the following commands need to be executed before analysis.

```
> export PATH=install_path/bin:$PATH ↓
```

You can omit above command if the above command is added to `~/.bash_profile` or `~/.bashrc`.

0.1.3 How to install multiple versions in a computer

If you install Claident and the other softwares to default install path of a computer to which Claident was already installed, all softwares will be overwritten. As noted above, multiple versions of Claident can coexist if you install Claident to non-default path. Note that a configuration file `.claident` placed at a home directory of login user (`/home/username`) or `/etc/claident` cannot coexist at the same path. You need to replace this file before changing the version of Claident. The configuration file at the home directory of login user will be used preferentially. To use multiple version, I recommend to make user account for each version and to install Claident to the home directory of each user. Then, the version of Claident can be switched by switching login user.

Chapter 1

Sequencing of multiple samples by next-generation sequencers

In this chaper, I explain brief overview of tagged multiplex sequencing method by Roche GS series sequencers, Ion PGM and Illumina MiSeq. These sequencers can read over-400bp contiguously and are suitable for metabarcoding and DNA barcoding. Note that MiSeq requires concatenation of paired-end reads. Therefore, PCR amplicons should be 500bp or shorter (400bp is recommended) in order to concatenate paired-end reads. Forward and reverse reads can be analyzed separetely, but I cannot recommend such analysis because reverse reads are usually low quality.

The next-generation sequencers output extremely large amount of nucleotide sequences in single run. Running costs of single run is much higher than Sanger method-based sequencers. To use such sequencers efficiently, multiplex sequencing method was developed. Multiplex identifier tag sequences are added to target sequences to identify the sample of origin, and the multiple tagged samples are mixed and sequenced in single run in this method. This method can extremely reduce per-sample sequencing costs. Multiplex identifier tag is also called as “barcode”. However, nucleotide sequence for DNA barcoding is called as “barcode sequence”. This is very confusing and “multiplex identifier tag” is too long. Thus, I call multiplex identifier tag sequence as just “tag” in this text. Please notice that tag is often called as “index”.

In the following analysis, chimera sequences constructed in PCR and erroneous sequences potentially causes misinterpretation of analysis results. If multiple PCR replicates are prepared, tagged and sequenced separately, shared sequences among all replicates can be considered as nonchimeric and less erroneous. This is because there are huge number of sequence combinations and joint points but no error sequence pattern is only one for one true sequence and nonchimeric and no error sequences likely to be observed at all replicates. Program cannot remove chimeras and errors enough but we can expect that the combination of PCR replicates and program improves removal efficiency of chimeras and errors. After removal of chimeras and errors, the number of sequences of PCR replicates can be summed up and used in subsequent analysis.

1.1 PCR using tag- and adapter-jointed-primers

In order to add tag to amplicon, PCR using tag-jointed primer is the easiest way. This method requires a set of tag-jointed primers. In addition, library preparation kits for next-generation sequencers usually presume that the adapter sequences specified by manufacturers are added to the both end of target sequences. Thus, the following tag- and adapter-jointed primer is used for PCR.

```
5' — [adapter] — [tag] — [specific primer] — 3'
```

If this kind of primers are used for the both forward and reverse primers, the following amplicon sequences will be constructed.

```
5' — [adapter-F] — [tag-F] — [specific primer-F] — [target sequence] — [specific primer-R (reverse
complement)] — [tag-R (reverse complement)] — [adapter-R (reverse complement)] — 3'
```

In the case of single-end read, tag-F leads specific primer-F and target sequence in the sequence data.

The supplement of Hamady *et al.* (2008) may be useful for picking tag sequences. In the case of single-end sequencing, 3'-side tag is not required, and tagless primer can be used for PCR. In the case of paired-end sequencing, single index (tag) can be applied, but dual index (tag) is recommended for detecting unlikely tag combinations which means that forward and reverse sequences are mispaired.

Using above primer sets for PCR, primers anneal to templates in "Y"-formation, and the amplicon sequences which have tags and adapters for both ends will be constructed. Then, the amplicon solutions are mixed in the same concentration and sequenced based on manufacturer's protocol. Spectrophotometer (including Nanodrop) is inappropreate for the measurement of the concentration of solution because measurement of dsDNA using spectrophotometer is likely to be affected by the other contaminants. I recommend Qubit (ThermoFisher) for measurement of dsDNA concentration. Quantitative PCR-based method can also be recommended but it's expensive and more time-consuming.

Primer annealing position sequence can also be used for recognizing the sample of origin. Therefore, the sequences of multiple loci, for example plant *rbcL* and *matK*, from same sample set tagged by same tag set can be multiplexed and sequenced. Of course, the sequences of multiple loci can also be recognized by themselves. Smaller number of cycles and longer extension time were recommended for PCR. Because the required amount of DNA for sequence sample preparation is not so high, the larger number of cycles of PCR amplification is not needed. The larger number of cycles and shorter extension time generates more incompletely extended amplicon sequences and the incompletely extended amplicon sequences are re-extend using different template sequences in next cycle. Such sequences are called as "chimeric DNA". Chimeric DNAs causes a discovery of non-existent novel species or a overestimation of species diversity. To reduce chimeric DNA construction, using high-fidelity DNA polymerase such as Phusion (Finnzymes) or KOD (TOYOBO) is effective. Stevens *et al.* (2013) reported that slowing cooling-down from denaturation temperature to annealing temperature reduced chimeric DNA construction. If your thermal cycler can change cooling speed, slowing cooling-down from denaturation temperature to annealing temperature can be recommended. Chimeric DNA sequences can also be eliminated by computer programs after sequencing. Because chimera removal by programs is incomplete and the nonchimeric

sequences shrink, we cannot do better than reduce chimeric DNA construction.

In the case of hardly amplifiable templates, using Ampdirect Plus (Shimadzu) for PCR buffer or crushing by homogenizer or beads before DNA extraction is recommended. Deep freezing before crushing can also be recommended. Removal of polyphenols or polysaccharides might be required if your sample contain those chemicals. If PCR amplification using tag- and adapter-jointed-primers fail, try two-step PCR that consist of primary PCR (20–30 cycles) using primers without tags and adapters, purification of amplicons by ExoSAP-IT, and secondary PCR (5–10cycles) using amplicons of primary PCR as templates and tag- and adapter-jointed-primers.

1.1.1 Decreasing costs by interim adapters

Tag- and adaper-jointed-primers are very long and expensive. In addition, we need to buy tag- and adaper-jointed-primers for each locus. To reduce cost of tag- and adaper-jointed-primers, interim adapter-jointed primers and two-step PCR is useful. The following primer set is used in primary PCR.

```
5' — [interim adapter] — [specific primer] — 3'
```

This PCR product have interim adapter sequences at the both ends. This PCR product is used as template in secondary PCR after purification. The following primer set is used in secondary PCR.

```
5' — [adapter specified by manufacturer] — [tag] — [interim adapter] — 3'
```

This two-step PCR enables us to reuse secondary PCR primers. However, this two-step PCR may increase PCR errors and PCR amplification biases, and decrease target sequence lengths. Note that final PCR product is constructed as the following style.

```
5' — [adapter-F specified by manufacturer] — [tag-F] — [interim adapter-F] — [specific primer-F] — [target
sequence] — [specific primer-R (reverse complement)] — [interim adapter-R (reverse complement)] — [tag-R
(reverse complement)] — [adapter-R (reverse complement) specified by manufacturer] — 3'
```

Illumina's multiplex sequencing method (Illumina corporation, 2013) using Nextera XT Index Kit is same as the above method. In the dual-index paired-end sequencing based on this method, the first read start from behind of interim adapter-F (i.e. head of specific primer-F) to target sequence. The second read start from behind of interim adapter-R and contains tag-R (index1) sequence. The third read start from behind of adapter-F and contains tag-F (index2) sequence. The last read start from ahead of interim adapter-R (i.e. tail of specific primer-R) to target sequence. The first, second, third and last reads are saved as `*_R1_*.fastq.gz`, `*_R2_*.fastq.gz`, `*_R3_*.fastq.gz` and `*_R4_*.fastq.gz`, respectively. The first, second and third reads are same strand, but last read is reverse strand. Because the sequencing primers for the first and the last reads are targeting interim adapter-F and interim adapter-R, respectively, the first and the last reads contains the sequences of specific primer-F and specific primer-R, respectively. Thus, the target sequences contained in the first and the last reads are shrinked. If the length of the target sequence is 500 bp or longer, there might be no overlap and paired-end reads cannot be concatenate. If specific primer-F and specific primer-R are used as sequencing primers for the first and the last reads, you can exclude the sequences of specific primer-F and specific primer-R from the first and the last reads. However, the following quality improvement method by insertion of N cannot be applied in such case.

1.1.2 Quality improvement by insertion of N

On the Illumina platform, luminescence of syntheses of DNA on a flowcell is detected by optical sensor. PCR amplicons of metagenomes are single locus and much more homogeneous than genome shotgun or RNA-seq library sequences. In such case, neighboring sequences on a flowcell is difficult to distinguish one from the other. In addition, if the nucleotide of the most sequences (especially first 12 nucleotides) are the same and nonluminescence, the Illumina platform sequencer will determined as failure and crash. To avoid this problem, insertion of `NNNNNN` between specific primer and interim adapter is effective. `NNNNNN` of the head of sequences enables sequencers to distinguish neighboring sequences and prevent black out, and the sequencing quality therefore will be improved (Nelson *et al.*, 2014). The varied length of `NNNNNN` causes artificial frameshift and also effective (Fadrosh *et al.*, 2014). PhiX control can be reduced by using the above methods, and the application sequences will increase.

Chapter 2

Preprocessing of nucleotide sequence data

Roche GS series sequencers and Ion PGM output raw sequencing data as `*.sff`. Illumina platform sequencers output `*.fastq` files. In this chapter, the procedures of demultiplexing, quality-trimming and quality-filtering. The `clsplitseq` command of Claident is recommended for demultiplexing because the programs provided by manufacturer ignores the quality of tag positions. The following commands should be executed in the terminal or console. Fundamental knowledge of terminal operations is required. If you are unfamiliar with terminal operations, you need to become understandable about the contents of appendix B.

2.1 Importing sequence data deposited to SRA/DRA/ERA or demultiplexed FASTQ

Claident assumes `SequenceID__RunID__TagID__PrimerID` for definition lines of sequences, and `RunID__TagID__PrimerID` for file names (without extension). Therefore, the sequence data deposited to SRA/DRA/ERA or demultiplexed FASTQ cannot be used as is. The `climportfastq` of Claident can convert such data. If your data is paired-end, you need to concatenate and filter the sequences before conversion (see section 2.3.3). The following plain text file is required for conversion.

```
| SequenceFileName1 RunID__TagID__PrimerID
| SequenceFileName2 RunID__TagID__PrimerID
| SequenceFileName3 RunID__TagID__PrimerID
```

Dummy RunID and PrimerID is acceptable. PrimerID need to be the same among the sample used the same primer set. TagID need to be different among the different sample files. TagID can be the same as the sequence file name.

After the above file was prepared, execute `climportfastq` like the following and the above file should be given as an input file.

```
> climportfastq \
--numthreads=NumberOfCPUs \
inputfile \
outputfolder ↓
```

Then, you can find converted files in the output folder. If your sequence data is single-end, quality filtering explained in section 2.2.3 is recommended.

2.2 For Roche GS series sequencers and Ion PGM

2.2.1 Converting SFF to FASTQ

First of all, conversion of raw SFF format file to FASTQ file is needed like the following.

```
> sff_extract -c inputfile(SFF) ↓
```

`-c` argument enables trimming of `TCAG` at the head of sequences. If you add `TCAG` to the head of tag sequences, do not use this argument. Assuming your SFF file name is `HOGEHOGE.sff`, `HOGEHOGE.fastq` will be saved as FASTQ file. `HOGEHOGE.xml` will also be generated, but this is not required. The output sequences have tag sequences at the beginning, followed by primer-F and target sequences, and primer-R (reverse complement) at the end. Note that all sequences are not completely read from the beginning to the end, the incomplete sequences are included. The `sff_extract` command is used in this book, but any other programs which can clip `TCAG` at the beginning can be used. If the SFF to FASTQ converter program cannot clip `TCAG` at the beginning, adding `TCAG` to the beginning of tag sequences to give to `clsplitseq` also works well, but the quality values will be strictly checked.

2.2.2 Demultiplexing of sequences

The FASTQ file that contain the sequences from multiple samples need to be demultiplexed based on tag sequences and primer sequences before the subsequent analysis. To do this process, a FASTA file which contain tag sequences and another FASTA file which contain primer-F sequences are required.

```
| >TagID
| [tag sequence]
| >examplesample1
| ACGTACGT
```

```
| >PrimerID
| [primer sequence]
| >exampleprimer1
| ACGTACGTACGTACGTACGT
```

Degenerate codes of nucleotides are not allowed for tag sequences, but those are allowed for primer sequences. Both of tag and primer FASTQ files can contain multiple sequences. If you use interim adapter explained in section 1.1.1, primer sequences should be written like the following.

```
| >PrimerID
| [interim adapter][primer sequence]
| >exampleprimer1
| TGATACTCGATACGTACGTACGTACGTACGTACGT
```

Thus, the sequences between tag and target sequences should be written in primer FASTA file.

All the above files are prepared, the following command can demultiplex nucleotide sequences to each sample FASTQ file.

```
> clsplitseq \
--runname=RunID \
--tagfile=TagSequenceFile \
--primerfile=PrimerSequenceFile \
--minqualtag=27 \
--numthreads=NumberOfCPUs \
inputfile \
outputfolder ↓
```

RunID must differ among different sequencing runs. RunID is given by sequencer in many cases, you can use such sequencer generated RunID. RunID is usually contained in sequence file name or sequence name in sequence file, but the naming rules are different among sequencing platforms. Therefore, `clsplitseq` requires RunID given by user. `--minqualtag` is an argument that specifies minimum quality threshold of tag position sequences. If 1 or more lower quality nucleotide than this threshold value is contained by a sequence, such sequence will be omitted from output sequences. 27 for minimum quality threshold is proposed by Kunin *et al.* (2010) for 3'-tail trimming of the sequences of Roche GS series sequencers. The different value might be more suitable for the other sequencers. In many cases, 30 is used for minimum quality threshold and can be recommended.

If multiplex sequencing technique is not used, `--tagfile` argument can be omitted. However, just omit of `--tagfile` generates incompatible FASTQ files for Claident. In such case, you should add identifier (dummy is acceptable) of tag sequences using `--indexname=TagID` argument.

The tag and primer position sequences are trimmed from the output sequences. Tag position sequence match is evaluated exactly and strictly. There are no arguments to tolerate a mismatch. Primer position sequence is aligned based on Needleman-Wunsch algorithm and evaluated allowing 14% of mismatches (the threshold can be changed). The output files are named as `RunID__TagID__PrimerID.fastq.gz` and saved in the output folder. `clsplitseq` can use multiple CPUs for faster processing. If your computer have 4 CPU cores, 4 should be speficied for `--numthreads` argument. Note that operating system and/or writing speed of storage devices might limit processing speed. By default, the output files are compressed by GZIP. Therefore, decompression is required to read/write by incompatible programs with GZIP-compressed FASTQ files. The commands of Claident used below can treat GZIP-compressed FASTQ files.

Before submission of manuscripts, sequence data need to be deposited to public database such as DDBJ Sequence Read Archive (DRA). GZIP-compressed FASTQ files in this step can be used for the data deposition.

If you sequenced a number of samples by multiple sequencing runs

Multiple demultiplexing by `clsplitseq` are required. However, `clsplitseq` cannot write already existing folder by default. The secondary run of `clsplitseq` requires `--append` argument like below.

```
> clsplitseq \
--runname=RunID \
--tagfile=TagSequenceFile \
--primerfile=PrimerSequenceFile \
--minqualtag=27 \
--numthreads=NumberOfCPUs \
inputfile1 \
outputfolder ↓
> clsplitseq \
--runname=RunID \
--tagfile=TagSequenceFile \
--primerfile=PrimerSequenceFile \
--minqualtag=27 \
--numthreads=NumberOfCPUs \
--append \
inputfile2 \
outputfolder ↓
```

If your tag sequence lengths are unequal

`clsplitseq` assumes that all tag sequence lengths are equal for faster processing. The unequal length tags must be splitted to multiple tag sequence files and multiple demultiplexing runs of `clcplitseq` are required as the following.

```
> clsplitseq \
--runname=RunID \
--tagfile=TagSequenceFile1 \
--primerfile=PrimerSequenceFile \
--minqualtag=27 \
--numthreads=NumberOfCPUs \
inputfile \
outputfolder ↓
> clsplitseq \
--runname=RunID \
--tagfile=TagSequenceFile2 \
--primerfile=PrimerSequenceFile \
--minqualtag=27 \
--numthreads=NumberOfCPUs \
--append \
inputfile \
outputfolder ↓
```

Recognition and elimination of reverse primer positions

In the above procedure, reverse primer position and subsequent sequences are not eliminated. Reverse primer position and subsequent sequences are artificial and should be eliminated if possible. To do so, reverse primer sequence file like the following is required.

```
| >PrimerID
| [primer sequence]
| >exampleprimer1
| TCAGTCAGTCAGTCAGTCAG
```

Multiple reverse primers can written in this file. Note that the N-th reverse primer sequence is assumed to associate with the N-th forward primer sequence. Therefore, the different number of primer sequences between forward and reverse primer sequence files causes an error. If there are the samples whose forward or reverse primer sequence is same but the other primer sequence is different, both combinations of forward and reverse primer sequences need to be written as different primers in the files.

After the preparation of the above file, perform `clsplitseq` as the following.

```
> clsplitseq \
--runname=RunID \
--tagfile=TagSequenceFile \
--primerfile=ForwardPrimerSequenceFile \
--reverseprimerfile=ReversePrimerSequenceFile \
--reversecomplement \
--minqualtag=27 \
--numthreads=NumberOfCPUs \
inputfile \
outputfolder ↓
```

In this processing, reverse-complement sequence of reverse primer is searched based on Needleman-Wunsch algorithm allowing 15% (this value can be changed) of mismatches and reverse primer position and subsequent sequence is eliminated in addition to the above process. If reverse-complement sequence of reverse primer is not found and the other requirement is fullfilled, the sequence will be saved to output file by default. The `--needreverseprimer` argument is required to filter out the sequence which does not contain reverse-complement sequence of reverse primer.

2.2.3 Trimming low quality tail and filtering low quality sequences

FASTQ sequences have read quality information. The low quality 3'-tail can be trimmed and the low quality sequences can be filtered out based on the quality values. The `clfilterseq` command can perform such processing as the following.

```
> clfilterseq \
--minqual=27 \
--minquallen=3 \
--minlen=350 \
--maxlen=400 \
--maxplowqual=0.1 \
--numthreads=NumberOfCPUs \
inputfile \
outputfile ↓
```

The values of `--minqual` and `--minquallen` indicate the minimum threshold of read quality value and size of sliding window, respectively. The above command trims 3'-tail positions until 3 bp long sequence whose read quality is 27 or higher in all 3 positions are observed. In addition, trimmed sequences shorter than `--minlen` will be filtered out and trimmed sequences longer than `--maxlen` will be trimmed to `--maxlen`. The remaining sequences containing `--maxplowqual` or more rate of lower quality positions than `--minqual` will also be filtered out. The output is a file by default, but can be saved to the file in the new folder using `--output=folder` argument. The output file name is same as the input file name in this case. If you want to save the output files to the existing folder, add `--append` argument.

If you want to apply `clfilterseq` to the all files in the output folder of `clsplitseq`, run the following command.

```
> for f in OutputFolderOfclsplitseq/*.fastq.gz ↓
do clfilterseq \
--output=folder \
--append \
--minqual=27 \
--minquallen=3 \
--minlen=350 \
--maxlen=400 \
--maxplowqual=0.1 \
--numthreads=NumberOfCPUs \
$f \
outputfolder ↓
done ↓
```

2.3 For Illumina platform sequences

2.3.1 Converting from BCL to FASTQ

The analysis software of Illumina platform sequences can demultiplex sequencing reads, but ignores read quality of tag positions. Therefore, the sequences have low quality tag positions possibly saved to demultiplexed FASTQ. To filering out such sequences, pre-demultiplexed FASTQ files are required and can be converted from BCL files with the aid of bcl2fastq. There are 1.x and 2.x series of bcl2fastq and both series can be used for Claident. However, the sequencers may be compatible to either 1.x or 2.x, you need to select proper version. Pre-demultiplexed FASTQ can be demultiplexed by `clsplitseq` in Claident. See appendix to install bcl2fastq.

To convert BCL to FASTQ, run data folder (superjacent folder of BaseCalls folder) need to be copied to the PC

installed bcl2fastq. If there is `SampleSheet.csv` in run data folder, this file must be renamed or deleted.

In the case of bcl2fastq 1.x, the following commands make FASTQ files from BCL files of 8 bp dual indexed 300PE sequencing data.

```
> cd RunDataFolder ↓
> configureBclToFastq.pl \
--fastq-cluster-count 0 \
--use-bases-mask Y300n,Y8,Y8,Y300n \
--input-dir BaseCalls \
--output-dir outputfolder ↓
> make -j4 ↓
```

The `--fastq-cluster-count 0` argument disable large output file splitting. The `--use-bases-mask Y300n,Y8,Y8,Y300n` is an argument to save forward 300 bp read (last base is trimmed), 8 bp index 1 (reverse-complement of tag-R), 8 bp index 2 (tag-F) and reverse 300 bp read (last base is trimmed) to `*_R1_001.fastq.gz`, `*_R2_001.fastq.gz`, `*_R3_001.fastq.gz` and `*_R4_001.fastq.gz`, respectively. The value of `--use-bases-mask` argument need to be changed for the other sequencing settings. For 6 bp single indexed 250SE and 8 bp dual indexed 300SE sequencing data, `--use-bases-mask Y250n,Y6` and `--use-bases-mask Y300n,Y8,Y8` should be suitable, respectively. `make -j4` executes the conversion using 4 CPUs. The output files will be compressed by GZIP. The extension `.gz` of output files indicates that the file is compressed by GZIP. Claident is compliant with GZIP-compressed FASTQ files and decompression is not required.

In the case of bcl2fastq 2.x, perform the following command.

```
> bcl2fastq \
--processing-threads NumberOfCPUs \
--use-bases-mask Y300n,Y8,Y8,Y300n \
--runfolder-dir RunDataFolder \
--output-dir outputfolder ↓
```

The `--processing-threads`, `--use-bases-mask` and `--runfolder-dir` indicate the number of processor used in conversion, masking option in common with 1.x and run data folder, respectively.

2.3.2 Demultiplexing of sequences

FASTA files containing tag (index) sequences and primer sequences like the following are needed for demultiplexing. FASTA files containing secondary tag (index) sequences and reverse primer sequences are also required for paired-end sequencing data.

```
| >TagID
| [tag sequence]
| >examplesample1
| ACGTACGT
```

```
| >PrimerID
| [primer sequence]
| >exampleprimer1
| ACGTACGTACGTACGTACGT
```

Degenerate code is not allowed for tag sequences, but can be used in primer sequences. Multiple tags and primers can be written in the files, but the N-th reverse tag/primer sequence is assumed to associate with the N-th forward tag/primer sequence. Therefore, the different number of tag/primer sequences between forward and reverse tag/primer sequence files causes an error. If there are the samples whose forward or reverse tag/primer sequence is same but the other tag/primer sequence is different, both combinations of forward and reverse tag/primer sequences need to be written as different tags/primers in the files. If you added `N` in front of primer, `N` need to be added in primer sequence. If your `N` length is unequal, only the longest `N` should be written in the file.

All the required files prepared, the following command demultiplex sequences to each sample file.

```
> clsplitseq \
--runname=RunID \
--index1file=Index1Sequence(tag-Rrevcomp)File \
--index2file=Index2Sequence(tag-F)File \
--primerfile=ForwardPrimerSequenceFile \
--reverseprimerfile=ReversePrimerSequenceFile \
--minqualtag=30 \
--numthreads=NumberofCPUs \
inputfile1 \
inputfile2 \
inputfile3 \
inputfile4 \
outputfolder ↓
```

The input files should be specified in the order of `*_R1_001.fastq.gz`, `*_R2_001.fastq.gz`, `*_R3_001.fastq.gz` and `*_R4_001.fastq.gz`. The `--index1file` and `--index2file` arguments requires the FASTA sequence files of index 1 (reverse-complement of tag-R) and index 2 (tag-F), respectively. By default, the acceptable mismatches are 14% and 15% for forward and reverse primers, respectively. If you added `N` in front of primer, the `--truncateN=enable` argument need to be given. This argument enables exclusion of `N` of primer and matched positions of sequences in calculation of the rate of mismatches. Therefore, only the longest `N` is required to find `N`-added primer even if the length of `N` is unequal. After the processing, the number of sequences in demultiplexed files should be compared with those in demultiplexed files generated by Illumina softwares. Correctly demultiplexed files should contain fewer sequences than demultiplexed files generated by Illumina softwares. If you used specific primers for sequencing primers, forward and reverse sequences do not contain specific primer positions. In such cases, `--primerfile` and `--reverseprimerfile` arguments are not required, but `--primername=PrimerID` argument need to be given for converting sequence names as compliant with Claident. Dummy PrimerID is acceptable but no PrimerID is not.

If you do not perform multiplex sequencing using tag/index, `--index1file` and `--index2file` arguments are unneeded, but `--indexname=TagID` argument must be given for converting sequence names as compliant with Claident. Dummy TagID is acceptable but no TagID is not.

After demultiplexing, `RunID__TagID__PrimerID.forward.fastq.gz` and `RunID__TagID__PrimerID.reverse.fastq.gz` will be generated. These GZIP-compressed FASTQ files can be used for data deposition to sequence read archive sites such as DDBJ Sequence Read Archive (DRA). In deposition process to DRA, it is required that the sequence lengths are equal or not. Because primer position sequences that can be unequal lengths even if only one primer set was used are eliminated from demultiplexed sequence files, do not specify that the sequence lengths are equal.

2.3.3 Concatenating forward and reverse sequences

In the case of overlapped paired-end

The `clconcatpair` command in Claident can be used for concatenating overlapped paired-end sequence data. The `clconcatpair` concatenate forward and reverse sequences based on overlap positions using VSEARCH by the following command.

```
> clconcatpair \
--mode=OVL \
--numthreads=NumberOfCPUs \
inputfolder \
outputfolder ↓
```

This command finds `*.forward.fastq` and `*.reverse.fastq` in inputfolder, and concatenate the pairs automatically. GZIP-compressed `.gz` and/or BZIP2-compressed `.bz2` files are also be found and concatenated. Concatenated sequence files will be generated as `*.fastq.gz` in outputfolder.

If input file names are not compliant with `*.forward.fastq` and `*.reverse.fastq`, the following command can be used for concatenating a pair of files.

```
> clconcatpair \
--mode=OVL \
--numthreads=NumberOfCPUs \
inputfile1 \
inputfile2 \
outputfile ↓
```

The forward and reverse sequence FASTQ files should be given as inputfile1 and inputfile2, respectively. Addition of `.gz` or `.bz2` is required for output file compression.

In the case of non-overlapped paired-end

If there are no overlaps between forward and reverse sequences, quality-trimming and quality-filtering using `clfilterseq` like the following should be performed at first.

```
> clfilterseq \
--minqual=30 \
--minquallen=3 \
--minlen=100 \
--maxplowqual=0.1 \
--numthreads=NumberOfCPUs \
inputfile1 \
inputfile2 \
outputfolder ↓
```

The values of `--minqual` and `--minquallen` indicate the minimum threshold of read quality value and size of sliding window, respectively. The above command trims 3'-tail positions until 3 bp long sequence whose read quality is 30 or higher in all 3 positions are observed. In addition, trimmed sequences shorter than `--minlen` will be filtered out. The remaining sequences containing `--maxplowqual` or more rate of lower quality positions than `--minqual` will also be filtered out. In this process, filtering out one of the sequence of a pair, the other sequence of the pair will also be filtered out. The output will be generated as the same name files in outputfolder. If you want to output to existing folder, you need to add `--append` argument. To apply the above command to all the pairs of `*.forward.fastq` and `*.reverse.fastq` in the current folder, execute the following commands.

```
> for f in `ls *.forward.fastq.gz | grep -P -o '^[^\.]+'` ↓
do clfilterseq \
--minqual=30 \
--minquallen=3 \
--minlen=100 \
--maxplowqual=0.1 \
--numthreads=NumberOfCPUs \
$f.forward.fastq.gz \
$f.reverse.fastq.gz \
outputfolder ↓
done ↓
```

After the quality-trimming and quality-filtering like above, perform sequence concatenation with the aid of `clconcatpair` like below.

```
> clconcatpair \
--mode=NON \
--numthreads=NumberOfCPUs \
inputfolder \
outputfolder ↓
```

In this process, the forward and reverse sequences like the following are assumed as input.

```
5' — forward sequence — 3'
5' — reverse sequence — 3'
```

The `clconcatpair --mode=NON` command will concatenate these sequence pairs and make sequences like the following.

```
5' — reverse sequence (reverse-complement) — ACGTACGTACGTACGT — forward sequence — 3'
```

Conduct removal of noisy and/or chimeric sequences in the same way as concatenated overlapped paired-end

sequence data. In the sequence clustering by `clclassseqv` and the raw reads mapping to centroid sequences by `clrecoverseqv`, add `--paddinglen=16` argument. The concatenated sequences like above causes overvaluation of sequence similarity because of artificial padding sequence `ACGTACGTACGTACGT`. The `--paddinglen=16` argument will offset such overvaluation by exclusion of `ACGTACGTACGTACGT` from sequence similarity calculation and cluster concatenated sequences based on correct sequence similarity.

In the estimation of host organism, split concatenated sequences based on `ACGTACGTACGTACGT`, assign taxonomy to forward and reverse sequences separately. Then, merge 2 taxonomy (see section 7.3). Generally speaking, forward sequences shows higher quality than reverse sequences, prefering forward sequence taxonomy is recommended if there is no *a priori* infomations about identification power and variability of forward and reverse sequences. Sequence division based on `ACGTACGTACGTACGT` can be applied to the sequences by the following command.

```
>cldivseq \
--query=ACGTACGTACGTACGT \
inputfile \
outputfile1 \
outputfile2
```

The outputfile1 and outputfile2 contain reverse-complement of reverse sequences and forward sequences, respectively.

Concatenating overlapping paired-end sequences using PEAR

PEAR (Zhang *et al.*, 2014) can also be used for concatenation of overlapped paired-end sequences. The following command will concatenate the pairs of sequences.

```
> pear \
-p 0.0001 \
-u 0 \
-j NumberOfCPUs \
-f RunID__TagID__PrimerID.forward.fastq.gz \
-r RunID__TagID__PrimerID.reverse.fastq.gz \
-o RunID__TagID__PrimerID ↓
```

If the processes correctly finished, the following files will be generated.

RunID__TagID__PrimerID.assembled.fastq Concatenated sequences.
RunID__TagID__PrimerID.unassembled.forward.fastq Unconcatenated forward sequences.
RunID__TagID__PrimerID.unassembled.reverse.fastq Unconcatenated reverse sequences.
RunID__TagID__PrimerID.discarded.fastq Discarded sequences by statistical test.

Only `RunID__TagID__PrimerID.assembled.fastq` is required in subsequent procedures. These output files are not compressed and consume large amount of storages, compression by GZIP or BZIP2 is recommended.

To apply concatenation to all the pairs of forward and reverse sequence files by PEAR, execute the following command.

```
> for f in `ls *.forward.fastq.gz | grep -P -o '^[^\.]+'` ↓
do pear \
-p 0.0001 \
-u 0 \
-j NumberOfCPUs \
-f $f.forward.fastq.gz \
-r $f.reverse.fastq.gz \
-o $f ↓
done ↓
```

Concatenating overlapping paired-end sequences using VSEARCH

VSEARCH can also be used directly for concatenation of overlapped paired-end sequences. The following command will concatenate the pairs of sequences.

```
> vsearch \
--threads NumberOfCPUs \
--fastq_mergepairs RunID_TagID_PrimerID.forward.fastq.gz \
--reverse RunID_TagID_PrimerID.reverse.fastq.gz \
--fastq_allowmergestagger \
--fastqout RunID_TagID_PrimerID.assembled.fastq ↓
```

If the amplicon sequences are shorter than read length, read tail of forward sequence possibly exceeds read head of reverse sequence or read tail of reverse sequence possibly exceeds read head of forward sequence. VSEARCH does not concatenate such sequences by default. The `--fastq_allowmergestagger` argument enables such sequence concatenation. The overhang positions which exceeded read head of the other sequence will be eliminated because such positions are artificial. The `--fastq_allowmergestagger` argument is not required if there is no such sequences. Using PEAR, the same processing as VSEARCH with `--fastq_allowmergestagger` will be performed. The unconcatenated forward and reverse sequences can be obtained by `--fastqout_notmerged_fwd outputfile` and `--fastqout_notmerged_rev outputfile` arguments, respectively. The minimum overlap length, the minimum length of concatenated sequence, the maximum length of concatenated sequence, the maximum number of allowed mismatches and the maximum allowed expected errors in concatenated sequence can be specified by `--fastq_minovlen`, `--fastq_minmergelen`, `--fastq_maxmergelen`, `--fastq_maxdiffs` and `--fastq_maxee`, respectively. In the concatenation of overlapped paired-end sequences by `clconcatpair`, `--fastq_minovlen 20 --fastq_maxdiffs 20` is used by default, but `--fastq_minovlen 10 --fastq_maxdiffs 5` is used by default of VSEARCH.

To apply concatenation by VSEARCH to all the files in current folder, execute the following command.

```
> for f in `ls *.forward.fastq.gz | grep -P -o '^[^\.]+'` ↓
do vsearch \
--threads NumberOfCPUs \
--fastq_mergepairs $f.forward.fastq.gz \
--reverse $f.reverse.fastq.gz \
--fastq_allowmergestagger \
--fastqout $f.assembled.fastq ↓
done ↓
```

2.3.4 Filtering potentially erroneous sequences

There are read quality values in the FASTQ files. Therefore, we can filter out potentially erroneous sequences using these quality values. To do so, the following command can conduct such processing.

```
> clfilterseq \
--minqual=30 \
--maxplowqual=0.1 \
--numthreads=NumberOfCPUs \
inputfile \
outputfile ↓
```

The sequences containing `--maxplowqual` or more rate of lower quality positions than `--minqual` will also be filtered out by the above command. The output is a file by default, but adding `--output=folder` argument changes to save output as the same name file in the outputfolder. If you want to save output file to existing folder, `--append` argument is needed. In the case of concatenated sequences of overlapped paired-end sequences generated by Illumina platform sequencers, positions close to the both end is usually high quality and overlapped positions is also high quality if the same positions of forward and reverse sequences are matched. Therefore, trimming low quality positions close to the both end is needless. Filtering out sequences containing low quality positions is recommended for concatenated overlapped paired-end sequences. The existing sequence filtering programs such as FastQC (Andrews, 2010) or PRINSEQ (Schmieder & Edwards, 2011) are also recommended.

To apply the same processing to the concatenated sequences by PEAR, execute the following command.

```
> for f in *.assembled.fastq ↓
do clfilterseq \
--output=folder \
--append \
--minqual=30 \
--maxplowqual=0.1 \
--numthreads=NumberOfCPUs \
$f \
outputfolder ↓
done ↓
```

Using quality values, we can calculate expected number of read errors. Quality-filtering based on the maximum allowed expected errors in input sequences can also be applied using VSEARCH. The following command can apply such quality-filtering to the input sequences.

```
> vsearch \
--threads NumberOfCPUs \
--fastq_filter inputfile \
--fastq_maxee 1.0 \
--fastqout outputfile ↓
```

For the single-end sequence data and unconcatenated sequence data, same quality-trimming and quality-filtering

as Roche GS series sequencers and Ion PGM can be recommended (see section 2.2.3). Note that thresholds for quality values and read lengths should be changed.

2.4 同じ鋳型を複数回 PCR または複数回シーケンスした場合について

キメラ配列は PCR によって生成されますので、同一 PCR 由来配列なのか、異なる PCR 由来配列なのかをプログラムに正しく認識させなくてはなりません。そのため、同じ鋳型を複数回 PCR または複数回シーケンスした場合には、プログラムがデータを正しく扱えるようにいくつかの操作が必要になります。なお、ここでは、Nested PCR や、増幅のための PCR とタグ配列付加のための PCR はまとめて 1 回と数えることに注意して下さい。つまり、ここで「複数回 PCR」とは、「増幅産物を鋳型にしてさらに PCR で増幅する」ことではなく、「同じ鋳型を使用した PCR」を複数回行うことです。

2.4.1 同じ鋳型を複数回 PCR して同じタグ配列を付加して同じランでシーケンスした場合

実験デザインが間違っており、複数回 PCR を活かしたキメラ検出・除去は行えません。通常の 1 回の PCR と同様にしか扱えません。

2.4.2 同じ鋳型を複数回 PCR して別のタグ配列を付加して同じランでシーケンスした場合

Claident ではサンプル ID=RunID__TagID__PrimerID ですから、同じ鋳型由来のサンプルが複数存在するデータとなります。わざと別サンプルのままにしておいて、1 つ以上のサンプルで出現しない OTU はキメラや読み間違い (PCR の合成ミスも含む) のある配列と判断するというのも良い考えだと思います。ただ、キメラの生成や読み間違いがランダムであればこれはベストな方法でしょうが、できやすいキメラや起きやすい読み間違いというものがもしあれば (あるはずです)、同一のキメラ配列・読み間違いのある配列が全サンプルに出現する可能性があります。この方法がどの程度正確かという情報はまだ十分ではありませんので、プログラムによるキメラ検出も併用した方が良いかもしれません。この方法の効果を調べた研究 (Lange *et al.,* 2015) も参照して下さい。この方法は Claident でサポートしているため実施方法は後述します。サンプルごとの各 OTU の配列数の集計表をこの後で作成しますが、集計表の加工コマンド `clfiltersum` により複数のサンプルを統合して 1 サンプルにすることが可能です (各 OTU の配列数は和になります)。

2.4.3 同じ鋳型を 1 回 PCR して同じタグ配列を付加して複数のランでシーケンスした場合

1 回のランでは、たまたま得られる配列が少ないことがあります。そのため同一サンプルを複数のランでシーケンスし、全ランのデータを使いたいことがあります。Claident ではサンプル ID=RunID__TagID__PrimerID ですから、同じ鋳型由来のサンプルが複数存在するデータとなります。そのような場合はそのまま別サンプルとして解析を進めることもできますし、1 サンプルにまとめてから解析を進めることもできます。別サンプルのまま処理を進める方を推奨します。そして、サンプルごとの各 OTU の配列数の集計表を作成する際に複数のサンプルを統合して 1 サンプルにまとめて下さい (各 OTU の配列数は和になります)。後者の方法では旧パイプラインによる解析でサンプルごとにキメラ検出を行うことが可能ですが、それ以外特にメリットはありません。

同じ鋳型由来の複数サンプルを別サンプルのまま解析を進めた場合、サンプルごとの各 OTU の配列数の集計表を作成後に集計表の加工コマンド `clfiltersum` により複数のサンプルを統合して 1 サンプルにまとめて下さい (各 OTU の配列数は和になります)。

同じ鋳型由来の複数サンプルを 1 つのサンプルに統合してからこの後の解析に進む場合は、以下のように処理して下さい。

```
> clsplitseq \
オプション省略 \
--runname=HOGEHOGE \
inputfile1 \
outputfolder ↓
> clsplitseq \
オプション省略 \
--runname=HOGEHOGE \
--append \
inputfile2 \
outputfolder ↓
```

このように`--runname` オプションで RunID を置換し、2 つの入力ファイルのラン名を同一にしてやることで、2 つのファイルからの同一サンプルからの配列が `HOGEHOGE__TagID__PrimerID.fastq.gz` という 1 つのファイルに書き出されます。

Claident では RunID と TagID と PrimerID が同じなら、同一 PCR 産物=同一サンプルの配列として扱います。PCR が 1 回でタグ配列ファイルとプライマー配列ファイルが同じなら、当然 TagID と PrimerID は同じになっているはずですので、RunID も同じにしてしまうことで、同一 PCR 産物の配列として認識させることができるようになります。

2.4.4 同じ鋳型を複数回 PCR して同じタグ配列を付加して別々のランでシーケンスした場合

複数ラン分の配列ファイルがあるはずですので、`clsplitseq` によるサンプルごとの配列の分別と `clfilterseq` による低品質配列の除去を別々に実行し、別々のフォルダに出力します。この後のノイジー配列とキメラ配列の除去も別々に行い、クラスタリングの際に複数ラン分のデータをまとめて入力ファイルとして与えて下さい。そして、サンプルごとの各 OTU の配列数の集計表を作成する際に複数のサンプルを統合して 1 サンプルにまとめて下さい (各 OTU の配列数は和になります)。

2.4.5 同じ鋳型を複数回 PCR して異なるタグ配列を付加して別々のランでシーケンスした場合

1 回のランでは、たまたま得られる配列が少ないことがあります。そのため同一サンプルを複数のランでシーケンスし、全ランのデータを使いたいことがあります。Claident ではサンプル ID=RunID__TagID__PrimerID ですから、同じ鋳型由来のサンプルが複数存在するデータとなります。そのような場合はそのまま別サンプルとして解析を進めることもできますし、1 サンプルにまとめてから解析を進めることもできます。別サンプルのまま処理を進め、ノイジー配列とキメラ配列の除去の際に「同じ鋳型由来の複数の PCR 産物に共通に出現しない塩基配列はキメラもしくはノイジー配列とみなす」方法を適用することを推奨します。そして、サンプルごとの各 OTU の配列数の集計表を作成する際に複数のサンプルを統合して 1 サンプルにまとめて下さい (各 OTU の配列数は和になります)。

clsplitseq は、複数の配列ファイルがあるはずですので以下のように実行して下さい。ここではファイルが 2 つの場合を示します。

```
> clsplitseq \
オプション省略 \
--tagfile=タグ配列ファイル 1 \
--primerfile=プライマー配列ファイル \
inputfile1 \
outputfolder ↓
> clsplitseq \
オプション省略 \
--tagfile=タグ配列ファイル 2 \
--primerfile=プライマー配列ファイル \
--append \
inputfile2 \
outputfolder ↓
```

プライマー配列ファイル、出力フォルダは同じものを与えて下さい。--runname オプションでは異なる RunID を指定します。また、タグ配列ファイルは個別に用意して与えて下さい。ただし、同じ鋳型由来の配列に付加した TagID は同じにして下さい。配列が同じタグであっても、異なる鋳型由来の配列に付加したタグの ID を同一にしてはいけません。例えば以下の 2 ファイルを用意したとします。

```
| >sample1
| ACGTACGT
| >sample2
| ATGCATGC
```

```
| >sample1
| ATGCATGC
| >sample2
| ACGTACGT
```

この場合、sample1 の鋳型には 1 ラン目には ACGTACGT を付加し、2 ラン目では ATGCATGC を付加したことになります。sample2 の鋳型では逆になります。もしも、ほぼ同じ領域を増幅してはいるがプライマー配列は異なる場合、タグと同様に同じ ID を付けた別々のファイルを用意して与えて下さい。この後の解析に関しては後述します。

Chapter 3

重複配列カウントによるノイジー配列とキメラ配列の除去

Claident では、配列のクラスタリング結果から読み間違いの多いノイジーな配列を推定して除去することができます。方法は CD-HIT-OTU (Li *et al.,* 2012) とほとんど同じです。また、旧パイプラインでは VSEARCH に実装された UCHIME (Edgar *et al.,* 2011) アルゴリズムを呼び出して結果を受け取ることで、キメラ配列の検出・除去も可能です。新パイプラインでは、クラスタリング処理が終わってからキメラ検出・除去処理を適用します。

3.1 VSEARCH を用いた新パイプラインの場合

以下のコマンドでノイジー配列の検出・除去が行えます。なお、多数のファイルを一度に与えることができますが、できるだけ 1 ラン分のファイルを与えて下さい。これは、ランごとに塩基配列決定の質にばらつきがあるためです。1 ラン分の配列が非常に多く (1000 万超) 時間がかかってしまう場合は、1 サンプルごとにこの処理を行って下さい。

```
> clcleanseqv \
--derepmode=PREFIX \
--primarymaxnmismatch=0 \
--secondarymaxnmismatch=1 \
--pnoisycluster=0.5 \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

オプションのうち、--derepmode は前処理において全長一致 (FULLLENGTH) で配列をまとめるか、前方一致 (PREFIX) で配列をまとめるかを指定します。オーバーラップのあるペアエンド配列を連結して得られた配列を処理する場合は FULLLENGTH を、シングルエンド配列を処理する場合は PREFIX を指定して下さい。--primarymaxnmismatch はプライマリクラスタリングで許容するミスマッチ数で通常は 0 にします。--secondarymaxnmismatch はセカンダリクラスタリングで許容するミスマッチ数で、通常は 1 にします。ただし、「先頭から末尾までに読み間違いが全くない配列」がほとんどないと思われるようなあまりにもノイズの多いデータでは、--primarymaxnmismatch を 1 に、--secondarymaxnmismatch を 3 にしてみて下さい。それでもダメなら--primarymaxnmismatch を 2 に、

--secondarymaxnmismatch を 5 にして試します。--secondarymaxnmismatch は--primarymaxnmismatch の 2 倍より 1 大きい値に設定して下さい。--pnoisycluster はノイジー配列と判定する感度に関するオプションです。0 以上 1 以下で指定して下さい。大きいほど感度が上がります。デフォルト値・推奨値は 0.5 です。クラスタリングを最終的に 97% 以下の閾値で行うのであればこれでいいと思いますが、99% 前後でクラスタリングする場合、--pnoisycluster は 0.9 くらいにするといいかもしれません。ただし、レア OTU を捨ててしまう可能性が高くなるので、1 サンプル当たりの配列数を多めに読んでおく必要があります。

outputfolder には、以下のファイル群が保存されます。

parameter.txt　プライマリクラスタをノイジーと判定するのに使用した所属配列数の下限値
primarycluster.denoised.fasta.gz　ノイジーと判定されなかったプライマリクラスタの代表配列
primarycluster.fasta.gz　プライマリクラスタの代表配列
secondarycluster.fasta.gz　セカンダリクラスタの代表配列
RunID__TagID__PrimerID.noisyreads.txt.gz　ノイジー配列と判定された配列のリスト
RunID__TagID__PrimerID.singletons.txt.gz　プライマリクラスタリングでシングルトンになった配列のリスト

他にも多くのファイルが出力されますが、後の解析に必要なものもありますので削除しないようにして下さい。

3.1.1 PCR レプリケート法によるノイジー配列・キメラ配列の除去

PCR レプリケートを用意してノイジー配列・キメラ配列を検出・除去するには、まず同一鋳型 DNA 由来の PCR レプリケートはどのサンプルとどのサンプルなのかを記したファイルを用意する必要があります。以下のような内容で用意して下さい。

```
| sample1  sample2  sample3
| sample4  sample5
| sample6  sample7
```

同一の行内にタブ区切りで書かれたサンプルが、同一鋳型由来の PCR レプリケートであることを表しています。異なる鋳型由来のサンプルは別の行に記述して下さい。PCR レプリケートは 3 つ以上あっても構いません。また、レプリケート数が鋳型ごとに異なっても構いません。

上記ファイルが用意できたら、以下のように clcleanseqv を実行することでレプリケート間で共通しないプライマリクラスタをキメラもしくはノイジーであると判断して除去します。

```
> clcleanseqv \
--replicatelist=PCR レプリケート指示ファイル \
--derepmode=PREFIX \
--primarymaxnmismatch=0 \
--secondarymaxnmismatch=1 \
--pnoisycluster=0.5 \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

レプリケートが 3 つ以上であっても、全てのレプリケートで検出されたプライマリクラスタのみがノイジーでもキメラ

でもないと判断されます。また、1 つのプライマリクラスタが複数の鋳型から検出されていた場合、どれか 1 つの鋳型でノイジーまたはキメラと判定されていれば、他の鋳型でもノイジーまたはキメラとみなして除去します。これらの設定は以下のオプションで変更することができます。

--minnreplicate　整数値で指定。この値以上のレプリケートで観測されたプライマリクラスタはそのサンプルではノイジーでもキメラでもないと見なす。デフォルト値は 2。
--minpreplicate　小数値で指定。プライマリクラスタが観測されたレプリケート数/サンプルの総レプリケート数がこの値以上であれば、そのサンプルではノイジーでもキメラでもないと見なす。デフォルト値は 1。
--minnpositive　整数値で指定。この値以上の配列 (サンプルではない) がノイジーまたはキメラと判定されていれば、全サンプルでノイジーまたはキメラと見なす。デフォルト値は 1。
--minppositive　小数値で指定。ノイジーまたはキメラと判定された配列数/プライマリクラスタの総配列数がこの値以上であれば、全サンプルでノイジーまたはキメラと見なす。デフォルト値は 0。
--runname　RunID を指定。サンプル名に含まれる RunID が全てこれに置換される。同一サンプル名になったサンプルは統合される。

--minnreplicate と--minpreplicate の値に基づく 1 サンプル内での判定と、--minnpositive と--minppositive の値に基づく全サンプルにまたがる判定が 2 段階で行われていることに注意して下さい。前者では 2 つの条件を両方満たすとノイジーでもキメラでもないと判定され、後者では 2 つの条件を両方満たすとノイジーまたはキメラであると判定されます。また、同一のプライマリクラスタが、あるサンプルではノイジーまたはキメラだが、別のサンプルではそうでない、といった判定は行えません。最終的な判定は全サンプル共通です。また、PCR レプリケート指示ファイルに記されていないサンプルでは判定が行われません。

上記のように clcleanseqv を実行すると、以下のようなファイルも出力フォルダに保存されます。

primarycluster.chimeraremoved.fasta.gz　キメラと判定されなかったプライマリクラスタの代表配列
primarycluster.cleaned.fasta.gz　ノイジーでもキメラでもないと判定されたプライマリクラスタの代表配列
RunID__TagID__PrimerID.chimericreads.txt.gz　キメラ配列と判定された配列のリスト

3.2 Assams を用いた旧パイプラインの場合

以下のコマンドでノイジー配列とキメラ配列の検出・除去が行えます。なお、多数のファイルを一度に与えることができますが、できるだけ 1 ラン分のファイルを与えて下さい。これは、ランごとに塩基配列決定の質にばらつきがあるためです 1 ラン分の配列が非常に多く (100 万超) 時間がかかってしまう場合は、1 サンプルごとにこの処理を行って下さい。

```
> clcleanseq \
uchime --minh 0.1 --mindiv 0.8 end \
--detectmode=N+C \
--pnoisycluster=0.5 \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

オプションのうち、uchime から end の間は UCHIME の実行オプションです。--minh と--mindiv はキメラ配列と判定する感度に関するオプションです。--minh は値が小さいほど感度が高くなります。UCHIME の作者は 0.1 以上 5 以下の範囲にするよう推奨しています。デフォルト値は 0.1 です。--mindiv の指定値が 0.8 の場合、0.8% 以下の違

いしかない配列間はキメラが形成されていても無視されます。最終的に 97% 一致規準で配列をクラスタリングするとしたら、2% 程度の違いしかない配列間のキメラはクラスタリング・コンセンサス配列作成によって潰されるため、この値を大きくすることで計算が速くなります。また、`--pnoisycluster` はノイジー配列と判定する感度に関するオプションです。0 以上 1 以下で指定して下さい。大きいほど感度が上がります。デフォルト値・推奨値は 0.5 です。クラスタリングを最終的に 97% 以下の閾値で行うのであればこれでいいと思いますが、99% 前後でクラスタリングする場合、`--pnoisycluster` は 0.9 くらいにするといいかもしれません。ただし、レア OTU を捨ててしまう可能性が高くなるので、1 サンプル当たりの配列数を多めに読んでおく必要があります。

各入力ファイルは同一の鋳型を同一のチューブ内で PCR した配列群を含んでいる必要があります。Claident がキメラ検出に利用している UCHIME では、元配列数が多いコンセンサス配列ほど PCR 当初から存在した＝キメラではないと仮定し、元配列数が少ないコンセンサス配列が、より元配列数の多いコンセンサス配列の部分配列の組み合わせで説明できるかどうかを検討します。したがって、UCHIME に「同一の鋳型を同一のチューブ内で PCR した配列群内でのコンセンサス配列の元配列数」を正確に伝える必要があるため、1 ファイルは同一の鋳型を同一のチューブ内で PCR した配列群でなくてはならないのです。これは 1 ラン分のファイルを一度に与えることよりもはるかに優先度が高いので、同一の鋳型を同一のチューブ内で PCR したサンプルを複数回のランで塩基配列決定した場合などは複数ラン分のデータを入力しても構いません。

UCHIME によるキメラ判定はそれぞれのサンプルごとに行われ、複数のサンプルで観測されている完全一致配列の場合、どれか 1 つのサンプルでキメラと判定されていれば、他のサンプルでもキメラとみなして除去します。ただし、下記のオプションでこの動作を変更することができます。両オプションの両方を満たせばキメラと判定されます。

`--minnpositive` 整数値で指定。この値以上の配列 (サンプルではない) がキメラと判定されていれば、全サンプルでキメラと見なす。デフォルト値は `1`。
`--minppositive` 小数値で指定。キメラと判定された配列数/完全一致した全配列数がこの値以上であれば、全サンプルでキメラと見なす。デフォルト値は `0`。

出力フォルダには、以下のファイル群が保存されます。

RunID__TagID__PrimerID.chimeraremoved.dereplicated.fastq.gz キメラ配列除去し完全一致配列をまとめた配列
RunID__TagID__PrimerID.chimeraremoved.fastq.gz キメラ配列除去した配列
RunID__TagID__PrimerID.denoised.dereplicated.fastq.gz ノイジー配列除去し完全一致配列をまとめた配列
RunID__TagID__PrimerID.denoised.fastq.gz ノイジー配列除去した配列
RunID__TagID__PrimerID.cleaned.dereplicated.fastq.gz キメラ配列とノイジー配列を除去し完全一致配列をまとめた配列
RunID__TagID__PrimerID.cleaned.fastq.gz キメラ配列とノイジー配列を除去した配列
RunID__TagID__PrimerID.chimericreads.txt.gz キメラ配列と判定された配列のリスト
RunID__TagID__PrimerID.noisyreads.txt.gz ノイジー配列と判定された配列のリスト
RunID__TagID__PrimerID.singletons.txt.gz 完全一致配列をまとめたときにシングルトンになった配列のリスト
RunID__TagID__PrimerID.uchime.txt.gz UCHIME によるキメラ判定結果のログ
RunID__TagID__PrimerID.parameter.txt まとまった完全一致配列をノイジーと判定するのに使用した所属配列数の下限値

他にも多くのファイルが出力されますが、後の解析に必要なものもありますので削除しないようにして下さい。キメラ配列除去を無効化してノイジー配列除去のみを行う場合は`--detectmode=noise`、逆にノイジー配列除去を無効化してキメラ配列除去のみを行う場合は`--detectmode=chimera` オプションを付けて実行して下さい。ただし、`--detectmode=noise` で処理した結果を`--detectmode=chimera` で処理しても、一度にノイジー配列・キメラ配列除去を行った場合 (`--detectmode=N+C` を指定した場合またはデフォルト設定) と結果は一致しませんのでご注意下さい。処理する場合は同時に行う必要があります。これは、キメラ配列もノイジー配列と判定するための情報を持っており、同時処理の場合にはそれを利用するためです。

3.2.1 PCR レプリケート法によるノイジー配列・キメラ配列の除去

PCR レプリケートを用意してノイジー配列・キメラ配列を検出・除去するには、まず同一鋳型 DNA 由来の PCR レプリケートはどのサンプルとどのサンプルなのかを記したファイルを用意する必要があります。以下のような内容で用意して下さい。

```
| sample1  sample2  sample3
| sample4  sample5
| sample6  sample7
```

同一の行内にタブ区切りで書かれたサンプルが、同一鋳型由来の PCR レプリケートであることを表しています。異なる鋳型由来のサンプルは別の行に記述して下さい。PCR レプリケートは 3 つ以上あっても構いません。また、レプリケート数が鋳型ごとに異なっても構いません。

上記ファイルが用意できたら、以下のように `clcleanseq` を実行することでレプリケート間で共通しないプライマリクラスタをキメラもしくはノイジーであると判断して除去します。

```
> clcleanseq \
--replicatelist=PCR レプリケート指示ファイル \
uchime --minh 0.1 --mindiv 0.8 end \
--detectmode=N+C \
--pnoisycluster=0.5 \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

レプリケートが 3 つ以上であっても、全てのレプリケートで検出されたプライマリクラスタのみがノイジーでもキメラでもないと判断されます。また、1 つのプライマリクラスタが複数の鋳型から検出されていた場合、どれか 1 つの鋳型でノイジーまたはキメラと判定されていれば、他の鋳型でもノイジーまたはキメラとみなして除去します。これらの設定は以下のオプションで変更することができます。

`--minnreplicate` 整数値で指定。この値以上のレプリケートで観測されたプライマリクラスタはそのサンプルではノイジーでもキメラでもないと見なす。デフォルト値は 2。
`--minpreplicate` 小数値で指定。プライマリクラスタが観測されたレプリケート数/サンプルの総レプリケート数がこの値以上であれば、そのサンプルではノイジーでもキメラでもないと見なす。デフォルト値は 1。
`--minnpositive` 整数値で指定。この値以上の配列 (サンプルではない) がノイジーまたはキメラと判定されていれば、全サンプルでノイジーまたはキメラと見なす。デフォルト値は 1。
`--minppositive` 小数値で指定。ノイジーまたはキメラと判定された配列数/プライマリクラスタの総配列数がこの値以上であれば、全サンプルでノイジーまたはキメラと見なす。デフォルト値は 0。
`--runname` RunID を指定。サンプル名に含まれる RunID が全てこれに置換される。同一サンプル名になったサンプルは統合される。

`--minnreplicate` と `--minpreplicate` の値に基づく 1 サンプル内での判定と、`--minnpositive` と `--minppositive` の値に基づく全サンプルにまたがる判定が 2 段階で行われていることに注意して下さい。前者では 2 つの条件を両方満たすとノイジーでもキメラでもないと判定され、後者では 2 つの条件を両方満たすとノイジーまたはキメラであると判定されます。また、同一のプライマリクラスタが、あるサンプルではノイジーまたはキメラだが、別のサンプルでは

そうでない、といった判定は行えません。最終的な判定は全サンプル共通です。また、PCR レプリケート指示ファイルに記されていないサンプルでは判定が行われません。

Chapter 4

塩基配列クラスタリングに基づく OTU ピッキング

4.1　サンプル内での配列クラスタリング

4.1.1　VSEARCH を用いた新パイプラインの場合

新パイプラインではこのステップはありません。

4.1.2　Assams を用いた旧パイプラインの場合

以下のコマンドで、同一の鋳型を同一のチューブ内で PCR した配列群のクラスタリングを行います。前節の結果得られる、完全一致配列をまとめた結果 (RunID__TagID__PrimerID.cleaned.dereplicated.fastq.gz) を入力します。

```
> clclassseq \
--minident=0 \
--strand=plus \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

前節の出力ファイルに対してこの処理を適用する場合、以下のようにコマンドを打てばいいでしょう。

```
> clclassseq \
--minident=0 \
--strand=plus \
--numthreads=NumberOfCPUs \
"前節の出力フォルダ/*.cleaned.dereplicated.fastq.gz" \
outputfolder ↓
```

オプションのうち、--minident=0 は不一致が 1 塩基以下の配列をまとめる設定です (実際には、類似度が (最短配列の長さ-11/最短配列の長さ-10) 以上の配列をまとめます。最長配列の長さが 400bp で最短配列の長さが 350bp なら、0.997 となり、400bp の配列でも 1 塩基しか不一致は許容されませんので意図通りになりますが、最短配列と最長配列の差が大きすぎると意図した通りになりません。そのため、最短配列は最長配列の半分より 11 塩基以上長くなくてはなりません)。最終的に 97% で配列をまとめる場合もこの時点では 0 にしておくのがおすすめです。最終的に 99% でまとめる場合にもここでは 0 にしておきます。--strand=plus は同一ストランドの比較だけでクラスタリングを行うオプションです。片側からしか読んでいないので、両ストランドの比較をする必要はなく、比較するとかえって異常なクラスタリングをしかねません。両側から読んでいる場合も、連結した後はストランドは揃っているのでこれでいいはずです。出力フォルダ内の*.assembled.fastq.gz がクラスタリング後の配列です。

4.2 サンプル間での配列クラスタリング

4.2.1 VSEARCH を用いた新パイプラインの場合

塩基配列のクラスタリングを行って OTU を作成するには、以下のコマンドを実行して下さい。

```
> clclassseqv \
--minident=0.97 \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

入力ファイルには、clcleanseqv の実行時に PCR レプリケート法によるノイジー配列・キメラ配列の除去を併用した場合は primarycluster.cleaned.fasta.gz を、併用しなかった場合には primarycluster.denoised.fasta.gz を与えて下さい。オーバーラップのないペアエンドデータを clconcatpair を用いて連結した配列では、--paddinglen=16 をオプションとして加えて実行して下さい。

出力フォルダには、以下のファイルが保存されています。

clustered.otu.gz　どの生配列がどの OTU に割り当てられたかを記録しているファイル
clustered.fasta.gz　OTU の代表配列

4.2.2 Assams を用いた旧パイプラインの場合

サンプル内クラスタリングデータを用いてサンプル間でのクラスタリングを行い、全体のクラスタリング結果を得るには以下のコマンドを実行します。1 サンプルしかない場合も、上記のサンプル内配列クラスタリングを行ってからこの処理を行って下さい。

```
> clclassclass \
--minident=0 \
```

```
--strand=plus \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder1 ↓
> clreclassclass \
--minident=0.99 \
--strand=plus \
--numthreads=NumberOfCPUs \
outputfolder1 \
outputfolder2 ↓
> clreclassclass \
--minident=0.97 \
--strand=plus \
--numthreads=NumberOfCPUs \
outputfolder2 \
outputfolder3 ↓
```

ここでの入力ファイルは前節の出力ファイル*.assembled.fastq.gz です。したがって、実際には以下のように入力することになるでしょう。

```
> clclassclass \
--minident=0 \
--strand=plus \
--numthreads=NumberOfCPUs \
"前節の出力フォルダ/*.assembled.fastq.gz" \
outputfolder1 ↓
> clreclassclass \
--minident=0.99 \
--strand=plus \
--numthreads=NumberOfCPUs \
outputfolder1 \
outputfolder2 ↓
> clreclassclass \
--minident=0.97 \
--strand=plus \
--numthreads=NumberOfCPUs \
outputfolder2 \
outputfolder3 ↓
```

1 つ目の clclassclass でサンプル内配列クラスタリングでまとめられた配列において、不一致が 1 塩基以下の配列をまとめます。2 つ目の clreclassclass で、99% 以上一致する配列をまとめ直します。さらに 3 つ目の clreclassclass で、97% 以上一致する配列をまとめ直します。こうすることで、より高速に、より正確にクラスタリングが行われます。clreclassclass を使用せずに clclassclass に 97% でまとめるよう指示してしまうと、アルゴリズム上「まとめすぎ」が発生しやすくなります。

出力フォルダには、以下のファイルが保存されています。

assembled.contigmembers.gz　どの生配列がどの OTU に割り当てられたかを記録しているファイル
assembled.fastq.gz　OTU のコンセンサス配列
assembled.fasta　OTU のコンセンサス配列

4.3 OTU 代表配列への生配列のマッピング

ノイジー配列・キメラ配列と判定されてしまった配列でも、OTU の代表配列に対して指定した類似度の閾値で貼り付けることが可能なのであれば、復活させることができます。これにより、データの減少を抑えることができます。

4.3.1 VSEARCH を用いた新パイプラインの場合

以下のコマンドを実行して下さい。

```
> clrecoverseqv \
--minident=0.97 \
--centroid=OTU 代表配列のファイル \
--numthreads=NumberOfCPUs \
inputfile1 \
中略 \
inputfileN \
outputfolder ↓
```

OTU 代表配列のファイルには clclassseqv の出力した clustered.fasta.gz を用います。入力ファイルとしては、clcleanseqv の出力した primarycluster.fasta.gz または clcleanseqv での処理や低品質配列の除去を行う前の配列ファイルを使用して下さい。出力フォルダには、clclassseqv と同様のファイル群が作成されます。なお、オーバーラップのないペアエンドデータを clconcatpair を用いて連結した配列では、--paddinglen=16 をオプションとして加えて実行して下さい。

4.3.2 Assams を用いた旧パイプラインの場合

旧パイプラインではこのステップはありません。

Chapter 5

OTU ピッキング結果の要約と後処理

5.1 集計表の作成

以下のコマンドで、クラスタリング結果から各サンプルにおける各 OTU に割り当てられた生配列数を表にすることができます。

```
> clsumclass \
--output=Matrix \
inputfile \
outputfile ↓
```

ここでの入力ファイルは、クラスタリング結果の出力フォルダに保存されている、`clustered.otu.gz` (新パイプライン) または `assembled.contigmembers.gz` (旧パイプライン) というファイルです。出力されるファイルは表 5.1 のようなタブ区切りのテキストファイルで、Excel などの表計算ソフトや R で読み込むことができます。ただし、表計算ソフトは巨大な行列の全てを読み込むことができない可能性がありますので注意して下さい。このファイルに基づいて、群集生態学的な解析を行うことができます。

samplename	amongsampleotu_1	amongsampleotu_2	amongsampleotu_3
sampleA	2371	0	0
sampleB	0	1518	0
sampleC	1398	0	0
sampleD	0	1436	0
sampleE	0	0	1360
sampleF	0	0	977

Table 5.1 集計表の例 — 数値は各サンプルにおける各 OTU の観測配列数です。

5.2 集計表からの特定の OTU・サンプルの除去

`clsumclass` では、1 回だけ出現した配列も含めて全て出力されますが、以下のコマンドを用いてサンプルや OTU を選別することができます。入力ファイルは `clsumclass` の出力ファイルです。

```
> clfiltersum \
オプション \
inputfile \
outputfile ↓
```

使用できるオプションは以下の通りです。

`--minnseqotu` 整数値で指定。この値に基づいて「割り当てられた生配列数がどのサンプルでも指定値未満の OTU」を除去します。
`--minpseqotu` 小数値で指定。この値に基づいて「割り当てられた生配列数/サンプルの生配列総数がどのサンプルでも指定値未満の OTU」を除去します。
`--minntotalseqotu` 整数値で指定。この値に基づいて「割り当てられた生配列総数が指定値未満の OTU」を除去します。
`--minnseqsample` 整数値で指定。この値に基づいて「どの OTU の生配列数も指定値未満のサンプル」を除去します。
`--minpseqsample` 小数値で指定。この値に基づいて「そのサンプルのその OTU に割り当てられた生配列数/OTU の生配列総数がどの OTU でも指定値未満のサンプル」を除去します。
`--minntotalseqsample` 整数値で指定。この値に基づいて「生配列総数が指定値未満のサンプル」を除去します。
`--otu` 残したい OTU 名をカンマで区切って指定します。
`--negativeotu` 除去したい OTU 名をカンマで区切って指定します。
`--otulist` ファイル名を指定。1 行に 1 つの OTU 名を記したテキストファイルとして残したい OTU 名を指定します。
`--negativeotulist` ファイル名を指定。1 行に 1 つの OTU 名を記したテキストファイルとして除去したい OTU 名を指定します。
`--otuseq` ファイル名を指定。FASTA 形式の配列ファイルとして残したい OTU 名を指定します。
`--negativeotuseq` ファイル名を指定。FASTA 形式の配列ファイルとして除去したい OTU 名を指定します。
`--sample` 残したいサンプル名をカンマで区切って指定します。
`--negativesample` 除去したいサンプル名をカンマで区切って指定します。
`--samplelist` ファイル名を指定。1 行に 1 つのサンプル名を記したテキストファイルとして残したいサンプル名を指定します。
`--negativesamplelist` ファイル名を指定。1 行に 1 つのサンプル名を記したテキストファイルとして除去したいサンプル名を指定します。
`--replicatelist` ファイル名を指定。PCR レプリケート関係にあるサンプルをタブ区切りで同一行に記述する。出力で 1 つのサンプルに統合される。
`--runname` RunID を指定。集計表中のサンプル名に含まれる RunID が全てこれに置換される。同一サンプル名になったサンプルは統合される。

後述するノイジー配列・キメラ配列の除去や他領域配列の除去を行った上で、残った配列を用いて`--otuseq` などのオプションを用いた OTU の選別を行う場合は、他の選別より先に行って下さい。

5.3 PCR レプリケート法によるノイジー配列・キメラ配列の除去

新パイプラインでは `clcleanseqv` の実行時に行っていますが、旧パイプラインでも PCR レプリケートを用意してノイジー配列・キメラ配列を検出・除去することが可能です。それには、まず同一鋳型 DNA 由来の PCR レプリケートはどのサンプルとどのサンプルなのかを記したファイルを用意する必要があります。以下のような内容で用意して下さい。

```
| sample1  sample2  sample3
| sample4  sample5
| sample6  sample7
```

同一の行内にタブ区切りで書かれたサンプルが、同一鋳型由来の PCR レプリケートであることを表しています。異なる鋳型由来のサンプルは別の行に記述して下さい。PCR レプリケートは 3 つ以上あっても構いません。また、レプリケート数が鋳型ごとに異なっても構いません。

上記ファイルが用意できたら、以下のコマンドでクラスタリング結果からレプリケート間で共通しない配列を除去し

ます。

```
> clfilterseq \
--contigmembers=*.contigmembers.gz \
--replicatelist=PCR レプリケート指示ファイル \
inputfile \
outputfile ↓
```

レプリケートが 3 つ以上であっても、全てのレプリケートで検出された配列のみがノイジーでもキメラでもないと判断されます。また、1 つの OTU が複数の鋳型から検出されていた場合、どれか 1 つの鋳型でノイジーまたはキメラと判定されていれば、他の鋳型でもノイジーまたはキメラとみなして除去します。これらの設定は以下のオプションで変更することができます。

--minnreplicate　整数値で指定。この値以上のレプリケートで観測された OTU はそのサンプルではノイジーでもキメラでもないと見なす。デフォルト値は 2。
--minpreplicate　小数値で指定。OTU が観測されたレプリケート数/サンプルの総レプリケート数がこの値以上であれば、そのサンプルではノイジーでもキメラでもないと見なす。デフォルト値は 1。
--minnpositive　整数値で指定。この値以上の配列 (サンプルではない) がノイジーまたはキメラと判定されていれば、全サンプルでノイジーまたはキメラと見なす。デフォルト値は 1。
--minppositive　小数値で指定。ノイジーまたはキメラと判定された配列数/OTU の総配列数がこの値以上であれば、全サンプルでノイジーまたはキメラと見なす。デフォルト値は 0。
--runname　RunID を指定。サンプル名に含まれる RunID が全てこれに置換される。同一サンプル名になったサンプルは統合される。

--minnreplicate と--minpreplicate の値に基づく 1 サンプル内での判定と、--minnpositive と--minppositive の値に基づく全サンプルにまたがる判定が 2 段階で行われていることに注意して下さい。前者では 2 つの条件を両方満たすとノイジーでもキメラでもないと判定され、後者では 2 つの条件を両方満たすとノイジーまたはキメラであると判定されます。また、同一の OTU が、あるサンプルではノイジーまたはキメラだが、別のサンプルではそうでない、といった判定は行えません。最終的な判定は全サンプル共通です。また、PCR レプリケート指示ファイルに記されていないサンプルでは判定が行われません。

ここで作成した FASTA 配列ファイルを第 5.2 節の clfiltersum の--otuseq オプションに指定して集計表を加工することで、この配列ファイルから除去された OTU を集計表からも除去することができます。また、--replicatelist オプションと--runname オプションを clfilterseq と同様に指定すれば、PCR レプリケート関係にあるサンプルは 1 つのサンプルに統合されます (セルの値は合算されます)。

5.4 UCHIME によるキメラ配列除去

旧パイプラインではクラスタリング前にも UCHIME でキメラ配列除去は行っていますが、クラスタリング後の配列に対してリファレンスなしでのキメラ配列除去とリファレンスありでのキメラ配列除去を追加で行うことができます。両方行う場合は、リファレンスなしでのキメラ配列除去を先に行ってから、リファレンスありでのキメラ配列除去を行います。両方のキメラ配列除去を行う場合は、クラスタリング前のノイジー配列・キメラ配列除去において、キメラ配列除去を無効にしてもいいかもしれません。クラスタリング後のキメラ配列除去の方がずっと高速なので、1 サンプル当たりの配列が膨大でクラスタリング前のキメラ配列除去では長時間を要する場合には、この方法の方が適しています。なお、PCR レプリケート法によるノイジー配列・キメラ配列の除去も併用する場合は、PCR レプリケート法によるノイジー配列・キメラ配列の除去を最初に行います。利用するコマンドは clrunuchime です。このコマンドでは、

UCHIME アルゴリズムを実装している VSEARCH というプログラムを呼び出して、キメラ配列除去を行います。

5.4.1 リファレンスなしでのキメラ配列除去

clrunuchime を以下のように実行して下さい。

```
> clrunuchime \
--contigmembers=*.contigmembers.gz \
--otufile=*.otu.gz \
inputfile \
outputfolder ↓
```

旧パイプラインを使用した場合は--contigmembers オプションを、新パイプラインを使用した場合は--otufile オプションを使用します。上記の例は一度に説明するためのもので、両方を指定してはいけません。なお、どちらも指定しなかった場合、*.otu.gz が入力ファイルと同じフォルダに存在すればそれを、*.contigmembers.gz が存在すればそれを見つけ出して使用しますので、普段は何も指定しなくてもいいでしょう。

5.4.2 リファレンスありでのキメラ配列除去

clrunuchime を以下のように実行して下さい。リファレンスなしでのキメラ配列除去を行った場合は、nonchimeras.fasta というファイルがあるはずですので、それを入力ファイルにすればいいでしょう。

```
> clrunuchime \
--referencedb=リファレンスデータベース名 \
inputfile \
outputfolder ↓
```

UCHIME 用のリファレンスデータベース名と内容は以下の通りです。

rdpgoldv9　細菌 16S 用の RDP Gold データベース (v9 版)
unite20160101　真菌 ITS 用の UNITE データベース (20160101 版)
unite20160101untrim　真菌 ITS 用の UNITE データベース (20160101 トリミングなし版)
unite20160101its1　真菌 ITS 用の UNITE データベース (20160101ITS1 版)
unite20160101its2　真菌 ITS 用の UNITE データベース (20160101ITS2 版)

5.4.3 出力フォルダの内容について

出力フォルダに保存されるファイルは以下の 4 つです。

chimeras.fasta　キメラと判定された配列
nonchimeras.fasta　キメラでないと判定された配列
uchimealns.txt　判定時のアライメント
uchimeout.txt　判定時の親配列やスコアなど

uchimeout.txt の各項目の意味は
http://drive5.com/usearch/manual/uchimeout.html
をご覧下さい。

5.5 配列からの低頻度出現配列の除去

クラスタリングの出力フォルダには、clustered.fasta.gz (新パイプライン) または assembled.fasta (旧パイプライン) というファイルとして、最終的に得られた代表配列もしくはコンセンサス配列が保存されています。これを次章の DNA バーコーディングに用いますが、全サンプルを通して少量しか出現しないものまで含めると数が多すぎることがあります。そこで、以下のようにして 5 配列以上出現するものだけを抽出することができます。配列数の閾値を適当に変えてもいいでしょう。この閾値は慎重に決定して下さい。ノイジー配列・キメラ配列の除去を行なっている場合には必要ないかもしれません。

```
> clfilterseq \
--contigmembers=*.contigmembers.gz \
--otufile=*.otu.gz \
--minnseq=5 \
inputfile \
outputfile ↓
```

旧パイプラインを使用した場合は--contigmembers オプションを、新パイプラインを使用した場合は--otufile オプションを使用します。上記の例は一度に説明するためのもので、両方を指定してはいけません。なお、どちらも指定しなかった場合、*.otu.gz が inputfile と同じフォルダに存在すればそれを、*.contigmembers.gz が存在すればそれを見つけ出して使用しますので、普段は何も指定しなくてもいいでしょう。ここで作成した FASTA 配列ファイルを第 5.2 節の clfiltersum の--otuseq オプションに指定して集計表を加工することで、この配列ファイルから除去された OTU を集計表からも除去することができます。

5.6 保存領域配列認識による領域分割

ここまでの配列が複数領域にまたがっている場合 (たとえば ITS1・5.8S rRNA・ITS2 といった風に)、領域ごとに分割した方が良いかもしれません。領域間かその付近に保存的な領域があれば、それを境界の目印として領域を分割できます。特にユニバーサルプライマー配列などを使うといいでしょう。これは以下のコマンドで行うことができます。

```
> cldivseq \
--query=保存領域の配列 \
--border=start \
inputfile \
outputfile1 \
outputfile2 ↓
```

このコマンドでは、指定された配列を Needleman-Wunsch アルゴリズムを用いたアライメントによって 15% まで不一致を許して探します。該当する部位が見つかったら、その左端を境界として配列を 2 つに分け、それぞれ出力ファイル 1 と出力ファイル 2 に保存します。該当する部位が見つからなかった場合は、そのまま出力ファイル 1 だけに保存

されます。信頼度配列ファイルがある場合には、こちらも塩基配列に合わせて分割されます。なお、指定する配列は対象配列と同一のストランドにしておいて下さい。ストランドが異なる場合は、`--reversecomplement` オプションを付加して実行することで、`--query` に指定された配列の逆相補配列をアライメントに用いるようになります。

`--border` オプションを `start` ではなく `end` にすることで、検索に一致した部位の右端を境界として分割することができます。`--border=both` にした場合は、検索に一致した部位はどちらのファイルにも出力されません (これがデフォルト設定です)。検索した配列が見つからなかった場合は出力ファイル 1 のみに配列が保存され、出力ファイル 2 には配列が保存されませんが、これでは都合が悪い場合は`--makedummy` オプションを付加することでダミーの塩基配列が出力ファイル 2 に保存されます。分割した各領域の配列で別個にホスト生物を同定した上で同定結果を統合する場合など、配列間で対応がとれていないと困る場合にご利用下さい。ここでは 2 分割の例を挙げましたが、分割後にさらに分割を繰り返すことで 3 分割や 4 分割も可能です。

5.7 ITSx や Metaxa による ITS・SSU rRNA 配列の抽出

ITSx は核 ITS1・ITS2 領域を認識してその部位だけを取り出すことができるプログラムです (Bengtsson-Palme *et al.*, 2013)。多くの分類群で ITS は非常に変異に富んでいますが、ずっと保守的な SSU・LSU rRNA と隣接しているため、これらの配列が残っていると、その配列の影響で ITS では明確に区別できる遠縁な生物の SSU・LSU 配列が BLAST 検索時にヒットしてしまい、うまく同定できなくなってしまうことがあります。ITSx で ITS 領域のみを切り出して同定することでこのような問題に対処できます。

Metaxa は ITSx と同様に SSU (12S/16S/18S) rRNA を区別してそれぞれの部位だけを取り出すことができるプログラムです (Bengtsson *et al.*, 2011)。SSU rRNA は真核生物のメタバーコーディング・DNA バーコーディングに広く利用されている領域ですが、核だけでなくミトコンドリアや葉緑体、混入した細菌にも含まれているため、核の SSU を標的とする PCR を行った場合にも多少はミトコンドリアや細菌の SSU rRNA が混入してしまいます。そのような配列は群集生態学的解析の障害となるため、このプログラムで前もって除去しておくと後の解析が楽になります。

いずれのプログラムもインストール方法、使用方法をここでは説明しません。それぞれの Web サイトとマニュアルをご覧下さい。これらのプログラムで作成した FASTA 配列ファイルを第 5.2 節の `clfiltersum` の`--otuseq` オプションに指定して集計表を加工することで、この配列ファイルに含まれている OTU だけの集計表を作成することができます。

5.8 非ターゲット領域の探索と削除

ITSx や Metaxa は ITS と SSU rRNA 領域にしか使用できませんが、もっと汎用的な非ターゲット領域の探索方法があります。一つは BLAST などで検索して配列がどの遺伝子のものかを判定してくれるプログラムにかけることで、もう一つは多重整列プログラムを利用することです。ClustalW2・ClustalX2 (Larkin *et al.*, 2007) や MAFFT (Katoh & Standley, 2013) には、多重整列した結果を系統的に近いものが近くになるように並び替えて出力することができます。この機能を使って並び替えた上で、多重整列の閲覧ソフトで表示して目で確認すればどれが非ターゲット領域かすぐわかります。非ターゲット領域の OTU を除去した FASTA 配列ファイルを第 5.2 節の `clfiltersum` の`--otuseq` オプションに指定して集計表を加工することで、この配列ファイルに含まれている OTU だけの集計表を作成することができます。

Chapter 6

DRA へのデータ登録

クラスタリング後の配列は従来通り DDBJ などに登録すればいいですが、新型シーケンサーの生データは DDBJ Sequence Read Archive (DRA) というところへ登録することになっています (サンガー法シーケンサーの生データは Trace Archive)。本書にあるように複数サンプルからの DNA にタグを付けてシーケンスした場合、サンプルごとのデータファイルに分割して登録することを求められます。これは `clsplitseq` によって分割した FASTQ を使えば問題ありません。同一鋳型由来配列ファイルが複数ある場合は、GZIP 圧縮を解除してから連結して再圧縮するだけで構いません。

また、サンプルについての情報 (メタデータ) を記述した XML ファイルを作成しなくてはなりません。データ登録を受け付けている側でも、XML 作成を補助するツールが用意されていますが大量サンプルを扱っている場合は用意するのは大変です。そこで、DRA 登録用 XML 作成を補助する `clmaketsv` および `clmakexml` というコマンドを用意しています。以下でこれらの使用法を説明します。

DRA への登録には、DDBJ が管理する D-way でのユーザーアカウント作成と公開鍵の登録が必要です。詳しくは DRA Handbook
http://trace.ddbj.nig.ac.jp/dra/submission.shtml
をご覧下さい。登録の際、Submission、Study、Experiment、Sample、Run の概念と対応関係を把握している必要がありますので、その部分を特に注意して読んで理解しておいて下さい。

DRA では、Study は BioProject という研究プロジェクトデータベースに登録して、それを参照することになります。BioProject への登録は BioProject Handbook
http://trace.ddbj.nig.ac.jp/bioproject/submission.html
を参照して予め済ませておいて下さい。

さらに、Sample は BioSample という研究サンプルデータベースに登録して、それを参照します。BioSample への登録は BioSample Handbook
http://trace.ddbj.nig.ac.jp/biosample/submission.html
を参照して予め済ませておいて下さい。なお、メタゲノムを鋳型としてユニバーサルプライマーで増幅を行ったシーケンスサンプルの場合、MIMarks-Survey を MIxS のタイプとして指定します。サンプルの採集された高度、水深、温度、湿度、pH など、様々な情報を付加しておくことができます。後世のためにも、面倒がらずにできるだけ多くの情報を付加しておいて下さい。なお、各項目の意味は、Genomic Standards Consortium から提供されているチェックリスト
http://wiki.gensc.org/index.php?title=MIMARKS

を確認して下さい。これを見てもわからない場合は DDBJ に問い合わせて下さい。特に、地下や海底下のサンプルでは、平均海水面からサンプリング地点までの高さ・深さ、地面からサンプリング地点までの高さ・深さ、平均海水面から地面までの高さ・深さを入れる項目があってややこしいのでご注意下さい。

6.1 XML 作成用タブ区切りテキストの作成

DRA に登録する XML には、登録する FASTQ ファイルごとにそのファイルに含まれている配列の情報を記述します。これは非常に複雑で大変なため、一旦単純なタブ区切りテキストに書き出した上で、Excel などを用いてそれを編集し、編集後のタブ区切りテキストから XML を作成します。タブ区切りテキストを生成するコマンドは `clmaketsv` です。以下のように使って下さい。

```
> clmaketsv \
inputfile1 \
中略 \
inputfileN \
outputfile ↓
```

入力ファイルは登録するものを指定して下さい。ワイルドカードも使えます。タブ区切りテキストができたら、表計算ソフトなどに読み込ませて、各セルを埋めていって下さい。なお、Excel では特定の文字列を勝手に日時などと認識して変換してしまう機能 (無効にする方法はありません) がありますので、十分注意して下さい。各セル内では、`[Hogehoge,Fugafuga]` とあった場合は `Hogehoge` または `Fugafuga` のどちらかを選んで括弧と選ばなかったものを消して下さい。`<Fill in this cell>`とあった場合は、<>内の指示に従ってセルを埋めて下さい (括弧は残さない)。その他のセルは適当に自分で考えて何とかして下さい。なお、BioProject ID や BioSample ID は、正式なアクセッション番号がまだ発行されていない場合、DDBJ に申請したときの Submission ID (BioProject では PSUB から始まり、BioSample では SSUB から始まるもの) を入力しておけば問題ありません。

6.2 タブ区切りテキストからの DRA 登録用 XML ファイルの生成

タブ区切りテキストの編集が終わったら、タブ区切りテキストとして保存した上で以下のように `clmakexml` を実行して下さい。登録に必要な XML ファイル群が生成されます。なお、タブ区切りテキストは一度に複数指定することも可能です。ただし、各ファイルの最初の 3 行に記述する内容は、最初のファイルのものしか利用されません。2 ファイル目以降の最初の 3 行は無視されます。

```
> clmakexml \
タブ区切りテキストファイル \
DRA から割り振られた submission-ID ↓
```

DRA では、「Create new submission(s)」によって submission-ID が割り振られます。**ユーザー ID-0001** などとなっているはずです。`DRA から割り振られた submission-ID` にはこれを指定します。すると、`submission-ID.*.xml` という名前のファイルが 3 つ生成されます。これを DRA の XML Upload 機能を利用して送信します。その他もろもろの処理が終わると、登録してあるメールアドレスに割り振られたアクセッション番号が送られてきます。ただし、論文には BioProject ID を書くことが多いと思います。

Chapter 7

DNA バーコーディングによる配列のホスト生物同定

DNA バーコーディングは、近年非常に多くの分野で応用が進められている、生物の同定方法です。しかし、既知配列データベースが不十分な上、それを前提とした同定アルゴリズムが欠けていました。そこで、筆者は「問い合わせ配列と最近隣配列との間の変異量＜分類群内の最大変異量」という規準を考案し、これを実現するアルゴリズム QCauto 法 (Tanabe & Toju, 2013) を Claident に実装しました。ただし、その配列のホストの可能性がある全生物が記載済みで、しかもそのバーコード配列もデータベースに登録済みである場合には、そのような難しいことを考えずに最近隣配列と同種とみなせばよいでしょう。そちらの方法も Claident には実装してあります。以下のコマンドは全てターミナルかコンソールで実行して下さい。基本的なターミナルの使い方の知識は持っているものとして話を進めます。

7.1 BLAST 検索による近隣既知配列群の取得

以下のコマンドで、「問い合わせ配列と最近隣配列との間の変異量＜分類群内の最大変異量」という条件を満たすために検討すべき近隣既知配列群 (の GenBank ID) を取得できます。

```
> clidentseq \
--blastdb=overall_genus \
--numthreads=NumberOfCPUs \
inputfile \
outputfile ↓
```

入力ファイルには FASTA 形式の塩基配列ファイルを指定します。`--blastdb` オプションでは、BLAST 検索に用いるデータベース名を指定します。筆者が用意しているデータベースは以下の通りです。

animals_COX1_genus　動物 (Metazoa) の mtDNA *COX1* 配列で属以下の情報があるもの
animals_COX1_species　同上だが種以下の情報があるもの
animals_mt_genus　動物 (Metazoa) の mtDNA 配列で属以下の情報があるもの
animals_mt_species　同上だが種以下の情報があるもの
eukaryota_LSU_genus　真核生物の LSU (28S) rRNA 配列で属以下の情報があるもの
eukaryota_LSU_species　同上だが種以下の情報があるもの
eukaryota_SSU_genus　真核生物の SSU (18S) rRNA 配列で属以下の情報があるもの
eukaryota_SSU_species　同上だが種以下の情報があるもの

fungi_ITS_genus 真菌の ITS 配列で属以下の情報があるもの
fungi_ITS_species 同上だが種以下の情報があるもの
overall_class NCBI nt の中で綱以下の情報があるもの
overall_order 同上だが目以下の情報があるもの
overall_family 同上だが科以下の情報があるもの
overall_genus 同上だが属以下の情報があるもの
overall_species 同上だが種以下の情報があるもの
plants_matK_genus 緑色植物の cpDNA *matK* 配列で属以下の情報があるもの
plants_matK_species 同上だが種以下の情報があるもの
plants_rbcL_genus 緑色植物の cpDNA *rbcL* 配列で属以下の情報があるもの
plants_rbcL_species 同上だが種以下の情報があるもの
plants_trnH-psbA_genus 緑色植物の cpDNA *trnH–psbA* 配列で属以下の情報があるもの
plants_trnH-psbA_species 同上だが種以下の情報があるもの
prokaryota_16S_genus 原核生物の 16S rRNA 配列で属以下の情報があるもの
prokaryota_16S_species 同上だが種以下の情報があるもの
prokaryota_all_genus 原核生物の配列で属以下の情報があるもの
prokaryota_all_species 同上だが種以下の情報があるもの
semiall_class `overall_class` から脊椎動物と *Caenorhabditis* と *Drosophila* を除いたもの
semiall_order `overall_order` から脊椎動物と *Caenorhabditis* と *Drosophila* を除いたもの
semiall_family `overall_family` から脊椎動物と *Caenorhabditis* と *Drosophila* を除いたもの
semiall_genus `overall_genus` から脊椎動物と *Caenorhabditis* と *Drosophila* を除いたもの
semiall_species `overall_species` から脊椎動物と *Caenorhabditis* と *Drosophila* を除いたもの

`overall_*`は非常に巨大なため、多くのメモリを必要としますが、これらは万能なのでどんな配列にも使えますし、想定外の分類群や想定外の遺伝子座の配列の混入があってもとんでもない誤同定を避けることができます。`overall_genus` では属以下の情報がある、つまりそれなりに分類群の情報が信頼できそうな配列を選別していますが、あまりにマイナーな分類群だとそれでは近隣配列が見つからないことがあります。その場合は、綱以下まで同定された配列のデータベース `overall_class` を使ってみて下さい。その他のデータベースは、`overall_*`よりもメモリ占有量がずっと少ないため、メモリの少ないコンピュータでも動作します。

7.1.1 キャッシュデータベースの構築による高速化

`clidentseq` は BLAST 検索を何度も繰り返しますが、それなりの時間と大量のメモリを必要とします。そこで、予め問い合わせ配列と類似した配列の上位 1 万本 (設定で変更可) のヒットした部位のみを BLAST データベースから取得してキャッシュとなる BLAST データベースを作成しておき、`clidentseq` ではこれを対象として検索することで、劇的な高速化が可能です。これを行うのが `clmakecachedb` で、以下のように用います。

```
> clmakecachedb \
--blastdb=overall_genus \
--numthreads=NumberOfCPUs \
inputfile \
outputfolder ↓
```

`clidentseq` の実行時には、`--blastdb` オプションに上記コマンドの出力フォルダを指定することで、キャッシュデータベースが使われます。ただし、`clmakecachedb` と `clidentseq` の入力ファイルは同一である必要があります。メモリの使用量も大幅に減少しますので、`overall_*`を使用してもメモリ 16GB 程度で済みます。

7.1.2 参照配列データベースが全種を網羅している場合

問い合わせ配列のホストの可能性がある全生物が記載済みで、しかもそのバーコード配列も全て参照配列データベースに登録済みである場合には、以下のようにして類似度 99% 以上の上位 1 位 (1 位タイを含む) を近隣配列とすればいいでしょう。

```
> clidentseq \
blastn -task megablast -word_size 16 end \
--method=1,99% \
--blastdb=overall_genus \
--numthreads=NumberOfCPUs \
inputfile \
outputfile ↓
```

類似度 99% 以上の上位 1 位の配列を探し出す場合、BLAST 検索オプションは`-task megablast -word_size 16` としても見逃すことはまずないでしょうから、ここでは高速化のためにこのように指定しています。

7.2 近隣既知配列群に基づく同定

以下のコマンドで、近隣既知配列群の所属分類群が同一になるまで分類階層を上げていくことで配列を同定します。これは lowest common ancestor (LCA) algorithm と呼ばれています (Huson *et al.*, 2007)。

```
> classigntax \
--taxdb=overall_genus \
inputfile \
outputfile ↓
```

inputfile には前節で作成した `clidentseq` の出力ファイルを指定して下さい。`--taxdb` は既知配列の所属分類群データベース指定オプションです。BLAST データベースと同じ名前のものがインストールされていますので、それを指定して下さい。

`classigntax` のデフォルト設定では、少なくとも 2 本以上の近隣既知配列がないと同定することができません。第 7.1.2 節のように `clidentseq` の実行時に`--method=1,99%` を指定していると、近隣既知配列が 1 本しかないために一つも同定できないことになります。この場合は以下のようにして必要な近隣既知配列数を 1 にします。

```
> classigntax \
--taxdb=overall_genus \
--minnsupporter=1 \
inputfile \
outputfile ↓
```

出力ファイルは表 7.1 のような形のタブ区切りのテキストになっています。

`classigntax` のデフォルト設定では、全ての近隣既知配列の所属分類群が同一になるまで分類階層を上げていきます

query	phylum	genus	species
seqA	Ascomycota	*Chloridium*	*Chloridium virescens*
seqB	Ascomycota	*Chloridium*	*Chloridium virescens*
seqC	Ascomycota	*Chloridium*	*Chloridium virescens*
seqD	Basidiomycota	*Amanita*	*Amanita fuliginea*
seqE	Basidiomycota	*Coltriciella*	*Coltriciella dependens*
seqF	Basidiomycota	*Filobasidium*	*Filobasidium uniguttulatum*
seqG	Basidiomycota	*Laccaria*	*Laccaria bicolor*
seqH	Basidiomycota	*Lactarius*	*Lactarius quietus*
seqI	Basidiomycota	*Russula*	*Russula densifolia*
seqJ	Basidiomycota	*Russula*	*Russula densifolia*
seqK	Basidiomycota	*Russula*	*Russula densifolia*
seqL	Basidiomycota	*Russula*	*Russula vesca*
seqM	Basidiomycota	*Agaricus*	
seqN	Basidiomycota	*Amanita*	
seqO	Basidiomycota	*Amanita*	
seqP	Ascomycota	*Bisporella*	
seqQ	Ascomycota	*Capronia*	
seqR	Ascomycota	*Capronia*	
seqS	Ascomycota	*Cenococcum*	

Table 7.1 同定結果の例 — 空欄は unidentified、つまり不明ということです。横幅を抑えるためいくつかの分類階層は省いてあります。

ので、誤同定された配列が混ざっていたりした場合には下位の分類階層はほとんど不明になってしまいます。これは「ほぼ正しいと思われる」結果を得る＝「間違えない」ことをデフォルト設定では優先しているためですが、近隣既知配列が多い場合、多少不一致を許容して誤同定の可能性を増やしてでもできるだけ下位の分類階層まで同定したい場合があります。そのような場合には以下のように`--maxpopposer`と`--minsoratio`を指定します。

```
> classigntax \
--taxdb=overall_genus \
--maxpopposer=0.05 \
--minsoratio=19 \
inputfile \
outputfile ↓
```

`classigntax`は結果として採用する分類群に該当する配列を supporter 配列とし、該当しない配列を opposer 配列とします。`--maxpopposer`には opposer 配列の存在を許容する割合を百分率で指定します。`--minsoratio`は supporter 配列数と opposer 配列数の比の下限値を指定します。supporter 配列数と opposer 配列数がこれを下回るような同定結果は許容せず分類階層を上げていきます。2 つのオプションが必要なのは、その分類階層の情報がない配列が存在し得るためです。そのような配列は supporter でも opposer でもないので、2 つのオプションが必要になります。上記の例では、近隣既知配列中、5% までの opposer 配列を許容し、opposer 配列の 19 倍以上の supporter 配列を必要とします。

7.3 複数の同定結果の統合

例えば植物の *rbcL* と *matK* など、複数のバーコード領域を併用しなければ同定できないこともありますし、overall_genus と overall_class での同定結果のいいとこ取りをしたい場合があります。また、厳密な LCA による同定結果と 5% まで opposer を許容する LCA の同定結果の組み合わせでいいとこ取りしたい場合もあるでしょう。そのようなことが、複数の同定結果を統合することで可能になります。

植物の *rbcL* と *matK* を併用して同定する場合は、より深くまで同定できている方の結果を採用するのがいいでしょう。これは以下のコマンドでできます。

```
> clmergeassign \
--priority=equal \
--preferlower \
rbcL での同定結果 \
matK での同定結果 \
outputfile ↓
```

より保守的に考えるなら、以下のようにして、両方の同定結果が同一か、一方では未同定の場合だけ採用することもできます。ただし、上の階層で不一致だった場合には、同定結果が一方で未同定でも採用されません。

```
> clmergeassign \
--priority=equal \
rbcL での同定結果 \
matK での同定結果 \
outputfile ↓
```

trnH–psbA も併用して、最も深くまで同定できているものを採用する場合は以下のようにします。

```
> clmergeassign \
--priority=equal \
--preferlower \
rbcL での同定結果 \
matK での同定結果 \
trnH-psbA での同定結果 \
outputfile ↓
```

overall_genus の結果を優先的に採用し、overall_genus で同定できている分類階層までは一致しているが overall_class の方がより下位まで同定できている場合だけ採用するには、以下のようにします。

```
> clmergeassign \
--priority=descend \
overall_genus での同定結果 \
overall_class での同定結果 \
outputfile ↓
```

同様に、厳密な LCA による結果を優先し、厳密な LCA で同定できている分類階層まで一致しているが制約を緩めた

LCA による結果の方がより下位まで同定できている場合にのみ採用するには、以下のようにします。

```
> clmergeassign \
--priority=descend \
厳密な LCA での同定結果 \
制約を緩めた LCA での同定結果 \
outputfile ↓
```

References

Andrews, S. (2010) Software distributed by the author at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bengtsson, J., Eriksson, K. M., Hartmann, M., Wang, Z., Shenoy, B. D., Grelet, G.-A., Abarenkov, K., Petri, A., Rosenblad, M. A., Nilsson, R. H. (2011) Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie Van Leeuwenhoek*, **100**, 471–475.

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., Veldre, V., Nilsson, R. H. (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data *Methods in Ecology and Evolution*, **4**, 914–919.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., Ravel, J. (2014) An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, **2**, 6.

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., Knight, R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, **5**, 235–237.

Huson, D. H., Auch, A. F., Qi, J., Schuster, S. C. (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Illumina corporation (2013) 16S metagenomic sequencing library preparation.

Katoh, K., Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.

Lange, A., Jost, S., Heider, D., Bock, C., Budeus, B., Schilling, E., Strittmatter, A., Boenigk, J., Hoffmann, D. (2015) AmpliconDuo: A Split-Sample Filtering Protocol for High-Throughput Amplicon Sequencing of Microbial Communities. *PLoS One*, **10**, e0141590.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Li, W., Fu, L., Niu, B., Wu, S., Wooley, J. (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, **13**, 656–668.

Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L., Graf, J. (2014) Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One*, **9**, e94249.

Schmieder, R., Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*,

27, 863–864.

Stevens, J. L., Jackson, R. L., Olson, J. B. (2013) Slowing PCR ramp speed reduces chimera formation from environmental samples. *Journal of Microbiological Methods*, **93**, 203–205.

Tanabe, A. S., Toju, H. (2013) Two new computational methods for universal DNA barcoding: a benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. *PLoS One*, **8**, e76910.

Zhang, J., Kobert, K., Flouri, T., Stamatakis, A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.

Appendix A

その他のプログラム・データベースのインストール

A.1　bcl2fastq のインストール

Illumina から提供されている、BCL 形式のベースコールデータから FASTQ を生成するプログラム bcl2fastq は MiSeq・HiSeq 用の v1.8.4 および NextSeq 500・HiSeq X 用の v2.17.1.14 が
http://support.illumina.com/downloads/bcl2fastq_conversion_software.html
からダウンロードできます。v1.8.4 は以下のようにインストールできます。

```
# alien のインストール
$ sudo apt-get install alien
# bcl2fastq v1.8.4 のダウンロード
$ wget \
ftp://webdata:webdata@ussd-ftp.illumina.com/Downloads/Software/bcl2fastq/bcl2fastq-1.8.4-Linux-x86_64.rpm
# .rpm から.deb への変換とインストール
$ sudo alien -i \
bcl2fastq-1.8.4-Linux-x86_64.rpm
# bcl2fastq に必要な追加パッケージのインストール
$ sudo apt-get install libxml-simple-perl xsltproc
```

Ubuntu・Xubuntu・Lubuntu の 14.04LTS では、Perl のバージョンが新しすぎ、より古いバージョンを前提として書かれた bcl2fastq が動作しません。そこで以下のようにして問題の部分を書き換えます。

```
$ sudo perl -i -npe \
's/qw\(ELAND_FASTQ_FILES_PER_PROCESS\)/\("ELAND_FASTQ_FILES_PER_PROCESS"\)/' \
/usr/local/lib/bcl2fastq-1.8.4/perl/Casava/Alignment/Config.pm
$ sudo perl -i -npe \
's/qw\(ELAND_GENOME\)/\("ELAND_GENOME"\)/' \
/usr/local/lib/bcl2fastq-1.8.4/perl/Casava/Alignment/Config.pm
```

テキストエディタで
`/usr/local/lib/bcl2fastq-1.8.4/perl/Casava/Alignment/Config.pm`
の 747 行目と 751 行目を手動で書き換えても構いません。`qw(HOGEHOGE)` を `("HOGEHOGE")` に書き換えて下さい。

v2.17.1.14 のインストールをする場合は以下のようにコマンドを実行して下さい。

```
# rpm2cpio と cpio のインストール
> sudo apt-get install rpm2cpio cpio ↓
# bcl2fastq2 v2.17 のダウンロード
> wget \
ftp://webdata2:webdata2@ussd-ftp.illumina.com/downloads/software/bcl2fastq/bcl2fastq2-v2.17.1.14-Linux-x86_64.zip
↓
# 圧縮ファイルの展開
> unzip -qq bcl2fastq2-v2.17.1.14-Linux-x86_64.zip ↓
# .rpm を展開してコマンドを抽出
> rpm2cpio \
bcl2fastq2-v2.17.1.14-Linux-x86_64.rpm | cpio -id ↓
# コマンドをインストール
> sudo mv usr/local/bin/bcl2fastq /usr/local/bin/↓
```

Appendix B

ターミナルコマンド集

B.1　配列を数え上げる

```
# FASTQ の場合
> grep -P -c '^\+$' inputfile ↓
# GZIP 圧縮 FASTQ の場合
> gzip -dc inputfile | grep -P -c '^\+$' ↓
# BZIP2 圧縮 FASTQ の場合
> bzip2 -dc inputfile | grep -P -c '^\+$' ↓
# XZ 圧縮 FASTQ の場合
> xz -dc inputfile | grep -P -c '^\+$' ↓
# FASTA の場合
> grep -P -c '^>' inputfile ↓
# GZIP 圧縮 FASTA の場合
> gzip -dc inputfile | grep -P -c '^>' ↓
# BZIP2 圧縮 FASTA の場合
> bzip2 -dc inputfile | grep -P -c '^>' ↓
# XZ 圧縮 FASTA の場合
> xz -dc inputfile | grep -P -c '^>' ↓
```

B.2　配列を閲覧する

```
# 無圧縮ファイル内容を画面に出力
> cat inputfile ↓
# 無圧縮ファイル内容をスクロールしながら見る
> less inputfile ↓
# GZIP 圧縮ファイル内容を画面に出力
> gzip -dc inputfile ↓
# GZIP 圧縮ファイル内容をスクロールしながら見る
> gzip -dc inputfile | less ↓
# FASTQ の最初の 1 配列を画面に出力
> head -n 4 inputfile ↓
# GZIP 圧縮 FASTQ の最初の 1 配列を画面に出力
> gzip -dc inputfile | head -n 4 ↓
# FASTQ の特定の配列を検索して画面に出力
> grep -P -A 3 '^\@配列名の正規表現' inputfile ↓
# GZIP 圧縮 FASTQ の特定の配列を検索して画面に出力
```

```
> gzip -dc inputfile | grep -P -A 3 '^\@配列名の正規表現' ↓
```

B.3 圧縮・展開

```
# フォルダ内の全.fastq.gz を展開
> for f in *.fastq.gz ↓
do gzip -d $f ↓
done ↓
# フォルダ内の全.fastq.gz を展開 (サブフォルダ含む)
> for f in `find .  -name *.fastq.gz` ↓
do gzip -d $f ↓
done ↓
# 4 つの CPU を使用してフォルダ内の全.fastq.gz を展開
> ls *.fastq.gz | xargs -L 1 -P 4 gzip -d ↓
# 4 つの CPU を使用してフォルダ内の全.fastq.gz を展開 (サブフォルダ含む)
> find .  -name *.fastq.gz | xargs -L 1 -P 4 gzip -d ↓
# フォルダ内の全.fastq を圧縮
> for f in *.fastq ↓
do gzip $f ↓
done ↓
# フォルダ内の全.fastq を圧縮 (サブフォルダ含む)
> for f in `find .  -name *.fastq` ↓
do gzip $f ↓
done ↓
# 4 つの CPU を使用してフォルダ内の全.fastq を圧縮
> ls *.fastq | xargs -L 1 -P 4 gzip ↓
# 4 つの CPU を使用してフォルダ内の全.fastq を圧縮 (サブフォルダ含む)
> find .  -name *.fastq | xargs -L 1 -P 4 gzip ↓
```

B.4 抽出・保存

```
# FASTQ から最初の 1 万配列を抽出
> head -n 40000 inputfile > outputfile ↓
# FASTQ から最後の 1 万配列を抽出
> tail -n 40000 inputfile > outputfile ↓
# GZIP 圧縮 FASTQ から最初の 1 万配列を抽出して GZIP 圧縮 FASTQ へ保存
> gzip -dc inputfile | head -n 40000 | gzip -c > outputfile ↓
# GZIP 圧縮 FASTQ から最後の 1 万配列を抽出して GZIP 圧縮 FASTQ へ保存
> gzip -dc inputfile | tail -n 40000 | gzip -c > outputfile ↓
# GZIP 圧縮 FASTQ の特定の配列を検索して GZIP 圧縮 FASTQ へ保存
> gzip -dc inputfile | grep -P -A 3 '^\@配列名の正規表現' | gzip -c > outputfile ↓
```