2017年　NGSハンズオン講習会

8月31日

メタゲノム解析

森　宙史 (Hiroshi Mori), Ph.D.
国立遺伝学研究所
生命情報研究センター
hmori@nig.ac.jp

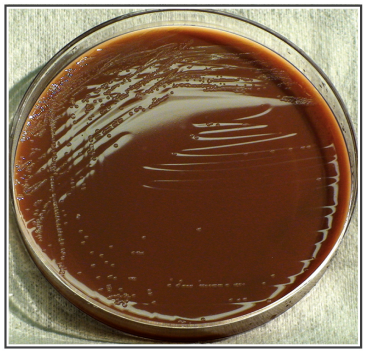


Figure 10.16 Microbiology: A Clinical Approach 2e (© Garland Science 2016)

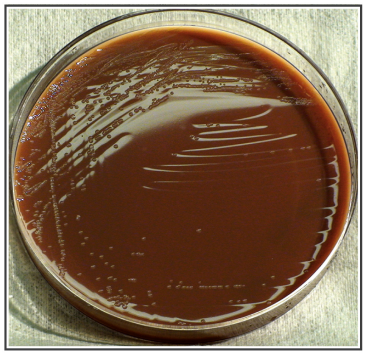


Figure 10.16 Microbiology: A Clinical Approach 2e (© Garland Science 2016)

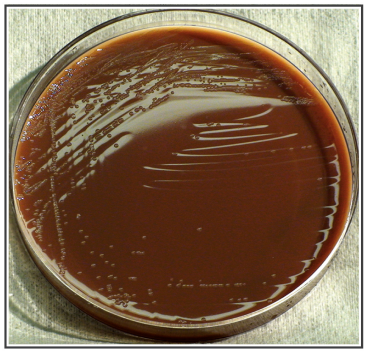


Figure 10.16 Microbiology: A Clinical Approach 2e (© Garland Science 2016)

数%ぐらいの菌しか
培養できない

細菌群集を解析するための様々な実験手法

細菌群集を解析するために使用されてきた実験手法

培養による細菌コロニーのカウント法 ・・・ 培養困難な細菌は解析出来ない

染色による細菌の数のカウント法 ・・・ 細菌の数しかわからない

FISH法による特定の細菌の染色法 ・・・ プローブ配列を設計する必要がある

DGGE法による細菌群集の解析法 ・・・ バンドパターンのみであり、

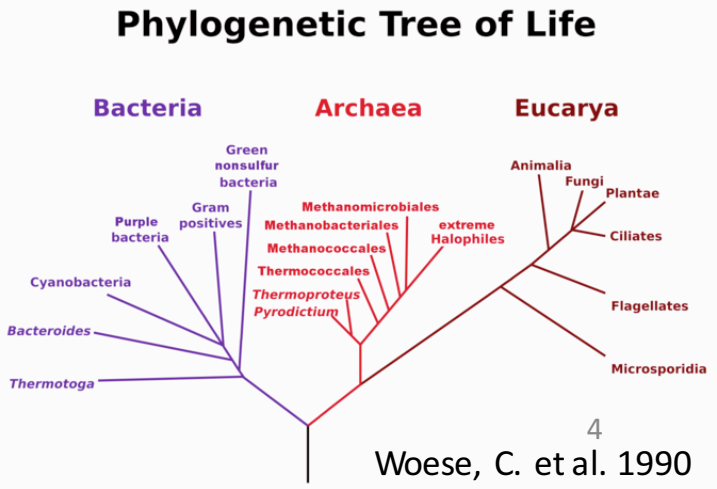
細菌群集の全体像をとらえるのは困難

細菌群集が形成するシステムを詳細に解析するためには、これらの手法では断片的な情報しか得られないため、細菌群集についての理解はあまり進んでいなかった

16S ribosomal RNA (16S rRNA)

- リボソームの核となるRNAの一つ
- 全ての細菌が所持
- 配列間の結合によって高次構造を形成
- 系統マーカー遺伝子の代表例
- 100万本以上の配列がデータベースに登録済み
- 全長1500 base

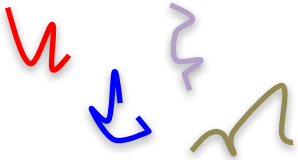
16S rRNA遺伝子は広範囲の細菌における
系統推定を行う上で最適な遺伝子



Woese, C. et al. 1990

16S rRNA gene amplicon sequencing analysis (メタ16S解析)

DNA extraction



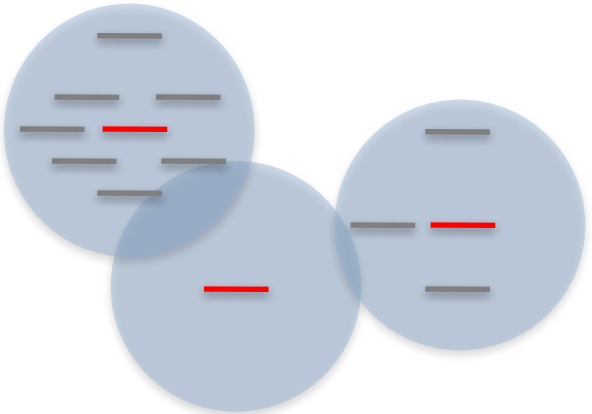
PCR amplification

DNA Sequencing

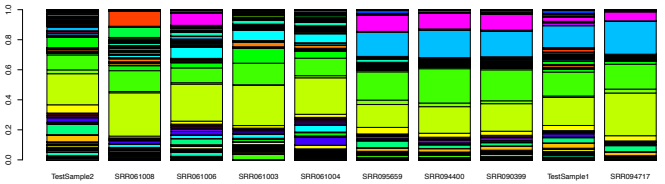
Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering with species level by CD-HIT-EST or UCLUST, etc.



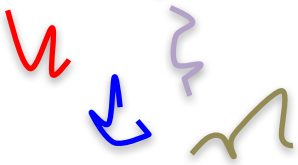
Taxonomic assignment and Comparison between samples



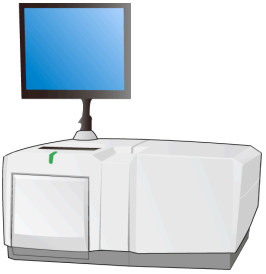
Who's there?

Metagenomic sequencing analysis (メタゲノム解析)

DNA extraction



DNA Sequencing



Metagenomic reads

Assemble — MEGAHIT, MetaPlatanus IDBA-UD etc.

Contig sets

Gene finding — MGA, MetaGeneMark

Gene sets

BLASTP

KEGG Kyoto Encyclopedia of Genes and Genomes

UniProt

Sample1 Metadata

Sample2 Metadata

Comparative metagenomics

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

Who's there?

What are they doing?

メタ16S解析

利点

・安価かつ少量のDNAから系統組成が得られる
・reference配列に依存しない解析も可能
・マシンパワーは少なくて済み、解析ツールも普及（QIIME・mothur等）

欠点

・PCRバイアスの存在
・種以下は分解能に問題あり
・個々の系統の機能が不明

メタゲノム解析

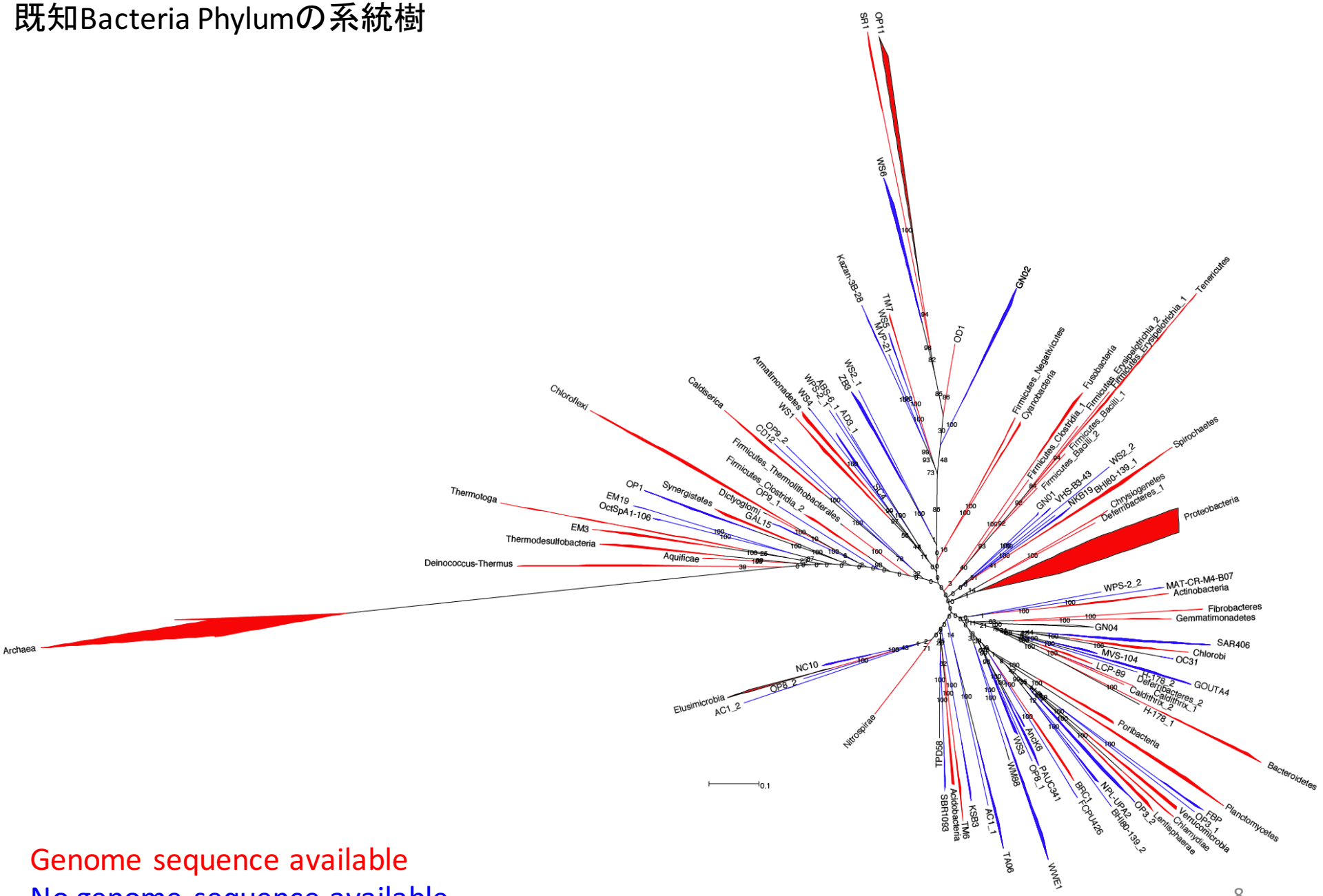
利点

・系統組成と遺伝子機能組成が得られる
・実験によるバイアスが少ない
・優占系統のドラフトゲノムの構築（条件が良ければ可能）

欠点

・reference配列に依存した解析
・目的依存で解析手法が変化し、マシンパワーも必要

16S rRNA遺伝子配列を用いた
既知Bacteria Phylumの系統樹

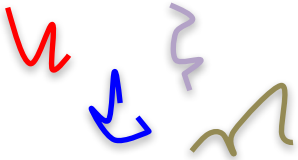


Genome sequence available

No genome sequence available

16S rRNA gene amplicon sequencing analysis (メタ16S解析)

DNA extraction



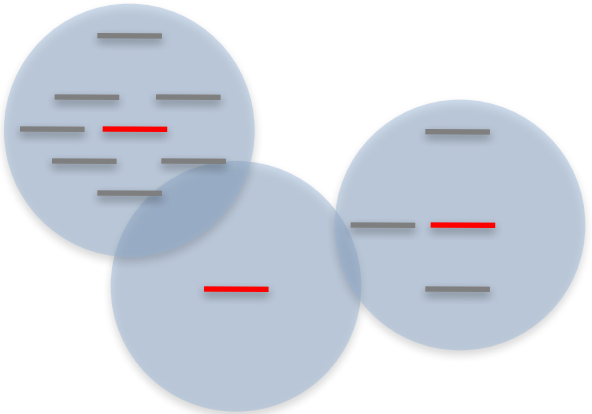
PCR amplification

DNA Sequencing

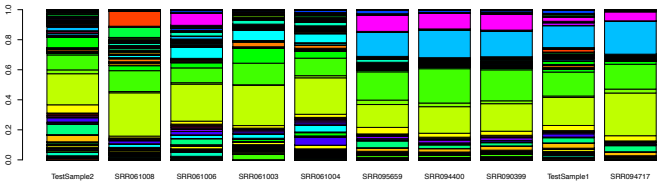
Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering with species level by CD-HIT-EST or UCLUST, etc.



Taxonomic assignment and Comparison between samples



Who's there?

PCRに使うプライマーは大きく分けて2種類

- 系統特異的なプライマー
- 系統Universalなプライマー

両者は何が違うのか？

系統(機能)特異的プライマー

例えば、

- 腸管出血性大腸菌とそうではない大腸菌を判別したい
- *Fusarium oxysporum*のレースを判別したい

などの病原性の有無の判定を迅速に行いたい場合に使われたりする。

遺伝子レベルで病原性のメカニズムがわかっている生物の場合、

- ある遺伝子を持っているか否か？
- ある遺伝子に一塩基置換があるか無いか？

が、病原性の有無に重要であるような場合には、非常に有効なプライマーが設計可能な場合が多い。

系統特異的プライマーの特徴

- 系統判定には増幅産物のシーケンスをしなくても大丈夫（PCR後の電気泳動でバンド出るか否か）
- どの遺伝子を使うかはバラエティに富む
- PCR条件を厳密に検討する必要がある（非特異的増幅の回避が重要）
- degenerate primerが少ない

degenerate primerとは？

Degenerate primer

例: 5′-GTGCCAGCMGCCGCGGTAA-3′

- 曖昧(縮重)塩基を使ったプライマー

曖昧塩基	塩基1	塩基2	塩基3	塩基4
R	A	G		
Y	C	T		
S	G	C		
W	A	T		
K	G	T		
M	A	C		
B	C	G	T	
D	A	G	T	
H	A	C	T	
V	A	C	G	
N	A	C	G	T

Universalプライマー

- 幅広い系統群を増幅できるプライマー

用途

- 系統の判別
- 群集組成を見る

特徴

- 増幅産物の系統判別には基本的にシーケンシングが必要
- degenerate primerが多い
- ターゲットになりうる遺伝子は少数

菌類(糸状菌･酵母など)

- 28S rRNA遺伝子
- 18S rRNA遺伝子
- COX1
- COX2
- rDNA–ITS1 (internal transcribed spacer), rDNA–ITS2

18S rDNA – ITS1 – 5.8S rDNA – ITS2 – 28S rDNA

Virus (RNA virusの場合)

- Tospovirus　Nタンパク質
- Comoviridae　RNA-dependent RNA polymerase
- Tombusviridae　RNA-dependent RNA polymerase
- Flexiviridae　RNA-dependent RNA polymerase、外被タンパク質（Coat protein)
- タバコモザイクウイルスなどの棒状ウイルス　外被タンパク質

細菌

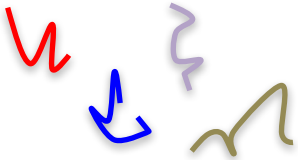
- 16S rRNA遺伝子
- rDNA–ITS

真核生物: 18S rDNA – ITS1 – 5.8S rDNA – ITS2 – 28S rDNA

原核生物: 16S rDNA – ITS – 23S rDNA – 5S rDNA

16S rRNA gene amplicon sequencing analysis (メタ16S解析)

DNA extraction



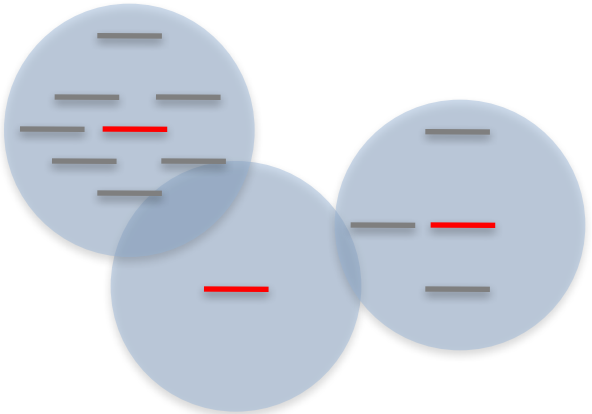
PCR amplification

DNA Sequencing

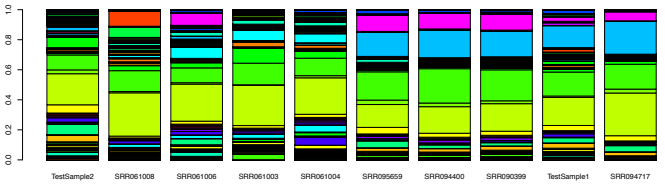
Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering with species level by CD-HIT-EST or UCLUST, etc.

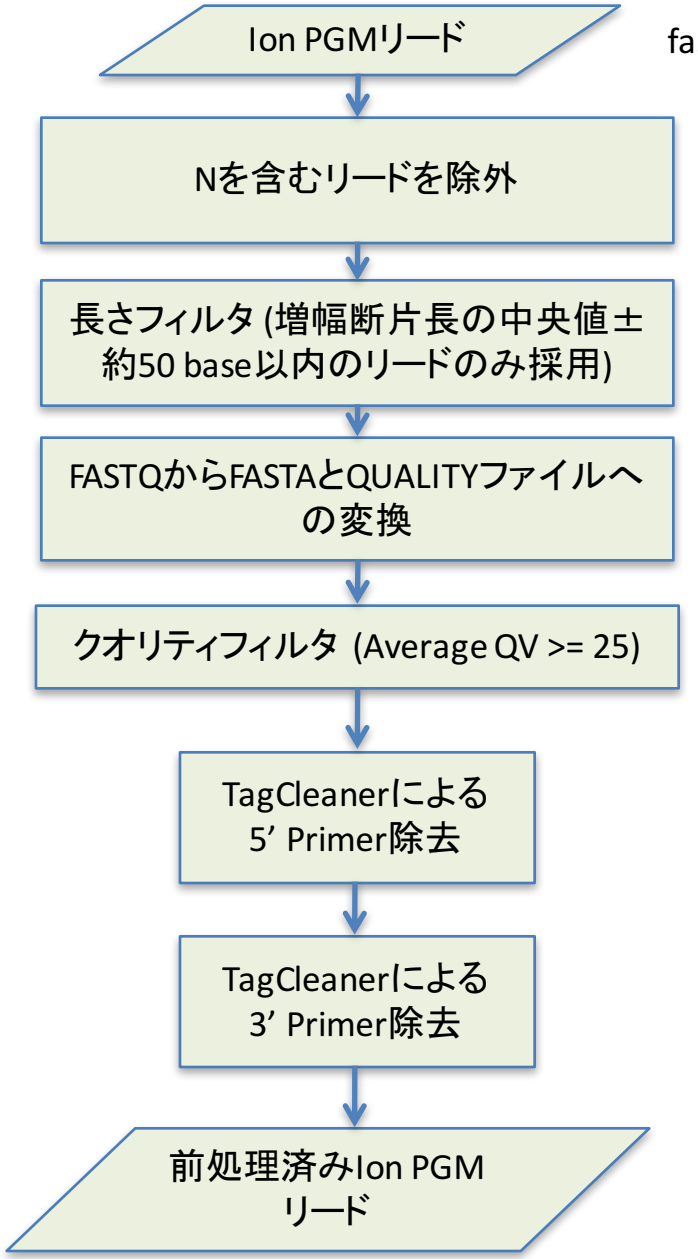


Taxonomic assignment and Comparison between samples

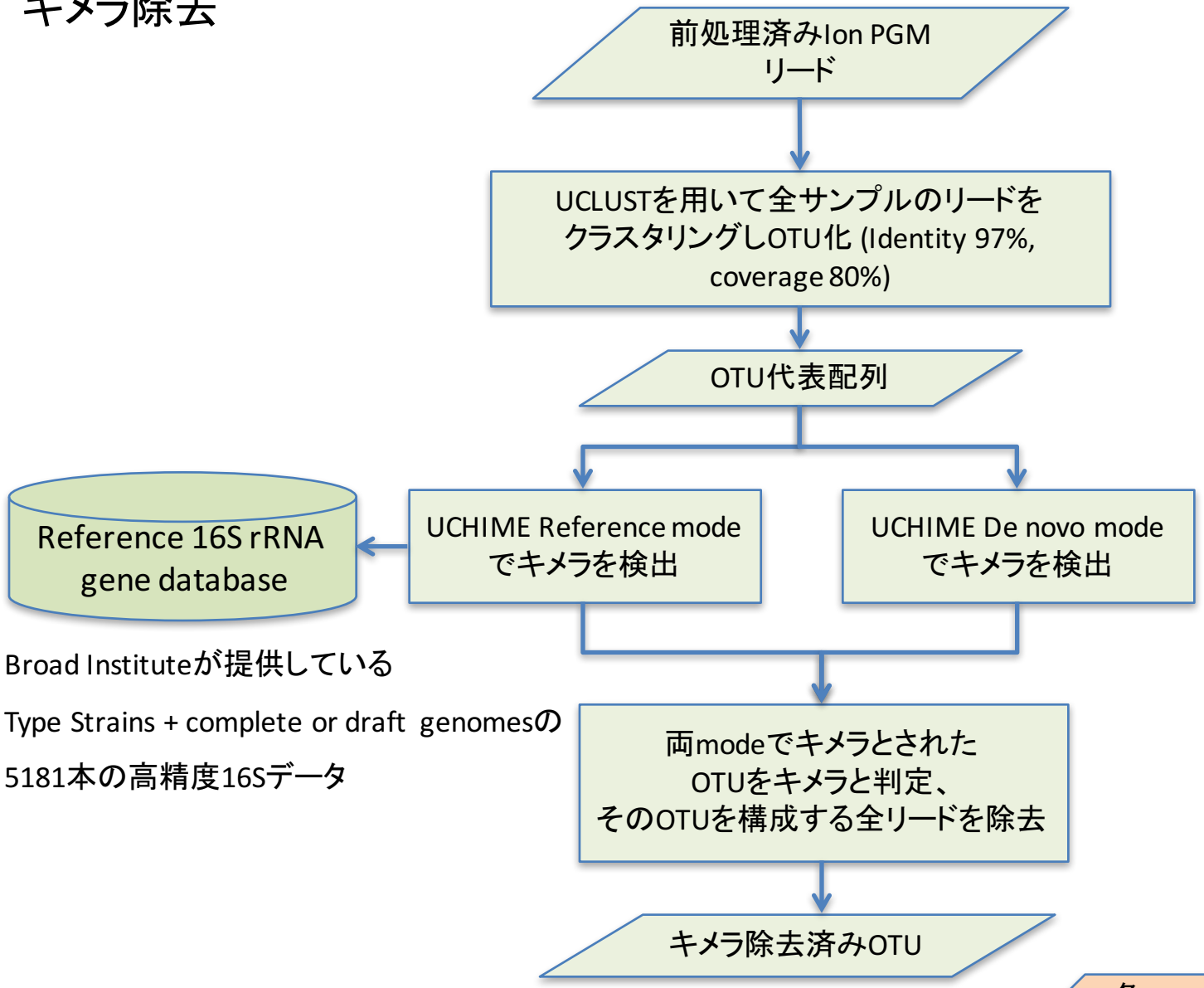


Who's there?

メタ16S解析パイプラインの例: Ion PGMリードの前処理



キメラ除去

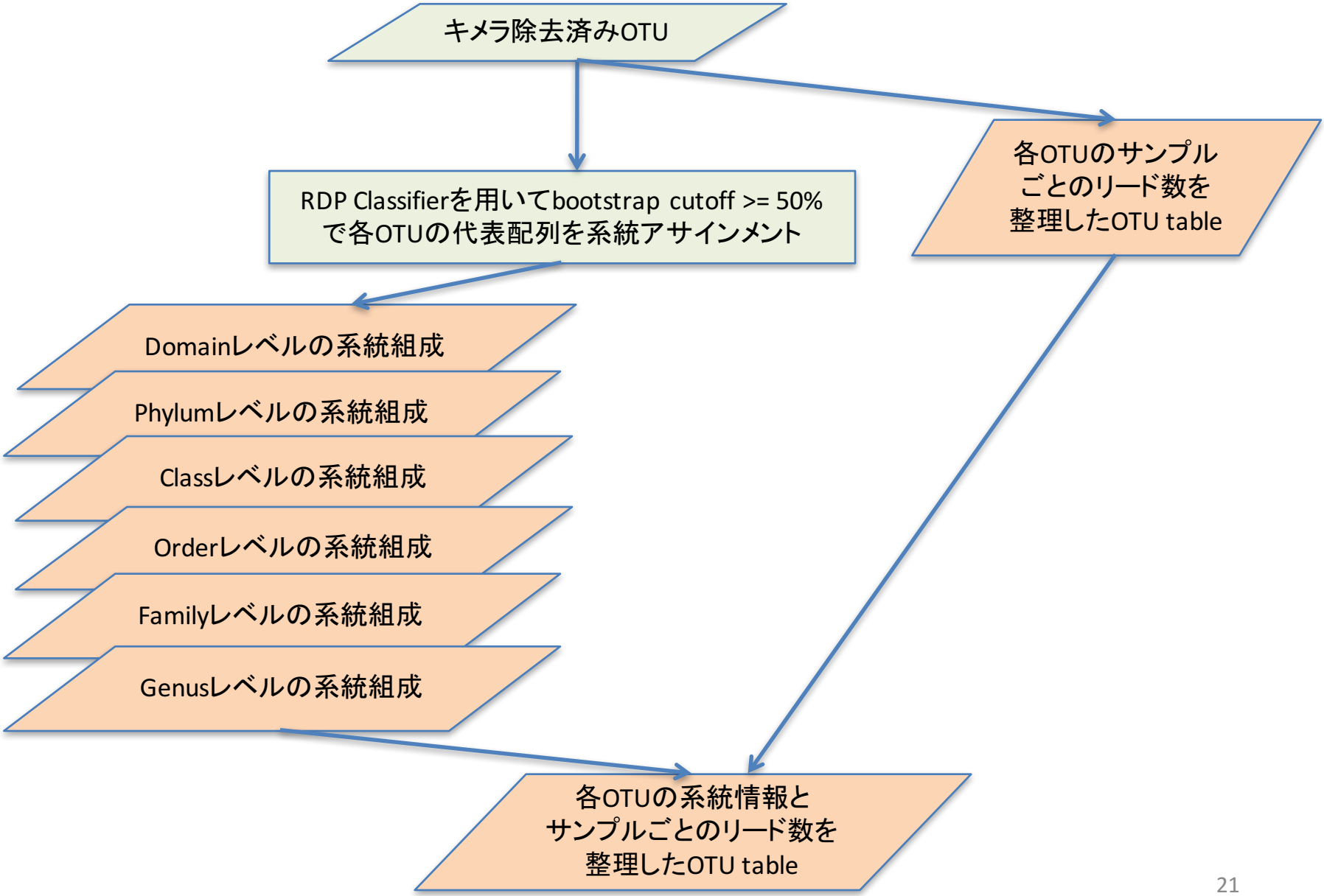


Broad Instituteが提供している

Type Strains + complete or draft genomesの

5181本の高精度16Sデータ

系統アサインメント


キメラ除去済みOTU
RDP Classifierを用いてbootstrap cutoff >= 50%
で各OTUの代表配列を系統アサインメント
各OTUのサンプル
ごとのリード数を
整理したOTU table
Domainレベルの系統組成
Phylumレベルの系統組成
Classレベルの系統組成
Orderレベルの系統組成
Familyレベルの系統組成
Genusレベルの系統組成
各OTUの系統情報と
サンプルごとのリード数を
整理したOTU table

http://qiime.org/

What is QIIME?

QIIME™ (canonically pronounced *chime*) stands for Quantitative Insights Into Microbial Ecology.

QIIME 2 will succeed QIIME 1 on January 1, 2018. QIIME 1 will no longer be supported at that time, as development and support effort for QIIME will be focused entirely on QIIME 2.

We recommend that existing QIIME users begin transitioning from QIIME 1 to QIIME 2 now. If you're new to QIIME, you should start by learning QIIME 2, not QIIME 1.

This site documents QIIME 1. To learn more about QIIME 2, see https://qiime2.org.

QIIME is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. QIIME is designed to take users from raw sequencing data generated on the Illumina or other platforms through publication quality graphics and statistics. This includes demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations. QIIME has been applied to studies based on billions of sequences from tens of thousands of samples.

QIIME citations since publication

Web of Science: ~3300

Google Scholar: ~5050

Available plugins

https://docs.qiime2.org/2017.7/plugins/available/

QIIME 2 microbiome analysis functionality is made available to users via plugins. The following official plugins are currently included in QIIME 2 train releases:

- alignment: Plugin for generating and manipulating alignments.
  - Methods
    - mafft: De novo multiple sequence alignment with MAFFT
    - mask: Positional conservation and gap filtering.
  - Visualizers
- composition: Plugin for compositional data analysis.
  - Methods
    - add_pseudocount: Add pseudocount to table
  - Visualizers
    - ancom: Apply ANCOM to identify features that differ in abundance.
- dada2: Plugin for sequence quality control with DADA2.
  - Methods
    - denoise_paired: Denoise and dereplicate paired-end sequences
    - denoise_single: Denoise and dereplicate single-end sequences
  - Visualizers
- deblur: Plugin for sequence quality control with Deblur.
  - Methods
    - denoise_16S: Deblur sequences using a 16S positive filter.
    - denoise_other: Deblur sequences using a user-specified positive filter.
  - Visualizers
    - visualize_stats: Visualize Deblur stats per sample.
- demux: Plugin for demultiplexing & viewing sequence quality.
  - Methods
    - emp_paired: Demultiplex paired-end sequence data generated with the EMP protocol.
    - emp_single: Demultiplex sequence data generated with the EMP protocol.
  - Visualizers
    - summarize: Summarize counts per sample.
- diversity: Plugin for exploring community diversity.

https://docs.qiime2.org/2017.7/tutorials/

qiime2 docs

Version: 2017.7

Table Of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
- Upgrading QIIME 2
- Tutorials
  - “Moving Pictures” tutorial
  - Fecal microbiota transplant (FMT) study: an exercise
  - “Atacama soil microbiome” tutorial
  - Importing data
  - Exporting data
  - Metadata in QIIME 2
  - Filtering data
  - Training feature classifiers with q2-feature-classifier
- Interfaces
- Plugins
- Semantic types
- Community
- Data resources
- Supplementary resources

Tutorials

Note

The tutorials assume you have installed the QIIME 2 Core distribution using one of the procedures in the **install documents**. The tutorials make extensive use of the QIIME 2 command-line interface so reviewing the **q2cli docs** is recommended.

- “Moving Pictures” tutorial
  - Sample metadata
  - Obtaining and importing data
  - Demultiplexing sequences
  - Sequence quality control and feature table construction
  - FeatureTable and FeatureData summaries
  - Generate a tree for phylogenetic diversity analyses
  - Alpha and beta diversity analysis
  - Taxonomic analysis
- Fecal microbiota transplant (FMT) study: an exercise
  - Obtain data files
  - Sequence quality control
  - Merging denoised data
  - Diversity analysis
  - Acknowledgements
- “Atacama soil microbiome” tutorial
  - Obtain the data
  - Paired-end read analysis commands
  - Questions to guide data analysis

https://mothur.org/wiki/Main_Page

Log

Page | Discussion | Read | View source | View history | Search

We will be hosting mothur, R, and reproducible research workshops throughout 2017. Learn more.

Main page
Recent changes
Random page
Help

Tools
What links here

Mothur manual

The goal of mothur is to have a single resource to analyze molecular data that is used by microbial ecologists. Many of these tools are available elsewhere as individual programs and as scripts, which tend to be slow or as web utilities, which limit your ability to analyze your data. mothur offers the ability to go from raw sequences to the generation of visualization tools to describe α and β diversity. Examples of each command are provided within their specific pages, but several users have provided several analysis examples, which use these commands. An exhaustive list of the commands found in mothur is available within the commands category index. If you have any questions, complaints, or praise, please do not hesitate to email Pat or to use the various discussion tabs:

Category:Commands

This is a listing of the various commands available within mothur.

Pages in category "Commands"

The following 146 pages are in this category, out of 146 total.

A

- Align.check
- Align.seqs
- Amova
- Anosim

B

- Bin.seqs

C

- Catchall
- Chimera.bellerophon
- Chimera.ccode
- Chimera.check
- Chimera.perseus
- Chimera.pintail
- Chimera.seqs
- Chimera.slayer
- Chimera.uchime
- Chimera.vsearch
- Chop.seqs
- Classify.otu
- Classify.rf

- Get.mimarkspackage
- Get.otulist
- Get.oturep
- Get.otus
- Get.rabund
- Get.relabund
- Get.sabund
- Get.seqs
- Get.sharedseqs

H

- Heatmap.bin
- Heatmap.sim
- Help
- Homova

I

- Indicator

K

- Kruskal.wallis

L

- Lefse

- Pcr.seqs
- Phylo.diversity
- Phylotype
- Pipeline.pds
- Pre.cluster
- Primer.design

Q

- Quit

R

- Rarefaction.shared
- Rarefaction.single
- Read.dist
- Read.tree
- Remove.dists
- Remove.groups
- Remove.lineage
- Remove.otulabels
- Remove.otus
- Remove.rare
- Remove.seqs
- Rename.file
- Rename.seqs

USEARCH has been cited by
5,305 papers
Google scholar
Last updated 27 Aug 2017

Buy 64-bit
Download 32-bit
Features
UPARSE OTU clustering
Documentation

10 - 1,250x BLAST
1 - 1,000x CD-HIT

New features in v10

High-throughput search and clustering
USEARCH is a unique sequence analysis tool with thousands of users world-wide. USEARCH offers search and clustering algorithms that are often orders of magnitude faster than BLAST.

35,834
registered users

64-bit users
Joint Genome Institute
MBL, Woods Hole
Cornell Univ.
CNRS (France)
La Jolla Institute

Improve insights
USEARCH algorithms outstandir support. T reduces th to take for compute t encourage enabling r new analy tried with

Example pipelines with test data

http://www.drive5.com/usearch/manual/pipe_examples.html

See also

OTU / denoising pipeline

Tutorials with data, scripts, and excercises with solutions **(new Aug 2017)**

Example	Description
Minimal, unpaired	OTUs from unpaired Illumina reads.
Minimal, paired	OTUs and ZOTUs from overlapping paired reads.
MiSeq 2x250 16S V4	OTUs and ZOTUs, OTU table, diversity analysis and taxonomy (16S).
MiSeq 2x300 ITS	OTUs and ZOTUs, OTU table, diversity analysis and taxonomy (ITS).
HMP 454 16S V5-V3	OTUs, OTU table, diversity analysis and taxonomy from 454 reads.

論文のMethodsの記述としてダメな例

- “sequence data were analyzed by QIIME”
- “sequence data were analyzed by mothur”
- “sequence data were analyzed by USEARCH”

としか書かないのはダメ

それらはコマンドの集合体であるため、必ず、
どのコマンドを使ってどのようなパラメータ設定で、
何をしたのかまで書く。

Supplementary Table 3. Analysis workflow and QIIME parameter settings used in this study

QIIME and Unix command [a]	Options
split_libraries.py [b]	-m MAP##.txt -f RunFile##.fna -q RunFile##.qual -l 200 -L350 -s 25 -a 0 -M 2 -e 1 -w 100 -b 10 -z disable --max-homopolymer 10 -o Run##/seqs.fna
truncate_reverse_primer.py [b]	-f Run##/seqs.fna -z truncate_remove -m SeqData/MAP##.txt -o Run##/ --primer_mismatches 2
cp [b]	Run##/seqs_rev_primer_truncated.fna seqs.Run##.fna
cat	*.fna > seqs.fna
pick_otus.py	-i seqs.fna -m usearch --db_filepath=/QIIME_database/gold_fa.fasta -g 2 --suppress_de_novo_chimera_detection
pick_rep_set.py	-f seqs.fna -i usearch_picked_otus/seqs_otus.txt -m most_abundant
align_seqs.py	-i seqs.fna_rep_set.fasta -m muscle
make_phylogeny.py	-i muscle_aligned/seqs.fna_rep_set_aligned.fasta -o rep_phylo.tre
assign_taxonomy.py	-i seqs.fna_rep_set.fasta -m rdp -c 0.5
make_otu_table.py	-i usearch_picked_otus/seqs_otus.txt -t rdp_assigned_taxonomy/seqs.fna_rep_set_tax_assignments.txt -o otu_table.biom
biom summarize-table	-i otu_table.biom -o reads_per_sample.txt
make_otu_heatmap_html.py	-i otu_table.biom -o otus/OTU_Heatmap/ -n 1 -m MAP.txt
summarize_taxa.py	-i otu_table.biom -o wf_taxa_summary/
plot_taxa_summary.py	-i wf_taxa_summary/otu_table_L5.txt -l Family -o graph_family -c bar --include_html_counts
multiple_rarefactions.py	-i otu_table.biom -o wf_arare/rarefaction/ --min 20 --max 5000 --num-reps 40 --step 20
alpha_diversity.py	-i wf_arare/rarefaction/ -m chao1,PD_whole_tree,observed_species,shannon -o wf_arare/alpha_div/ -t rep_phylo.tre
collate_alpha.py	-i wf_arare/alpha_div/ -o wf_arare/alpha_div_collated/

Matsuki T.†, Yahagi K.†, Mori H.†, *et al*, *Nature Commun.*, 2016

http://metagenomics.anl.gov/

MG-RAST

metagenomics analysis server

cite us

version 4.0.2

303,468 metagenomes containing 1,090 billion sequences and
145.71 Tbp processed for 23,610 registered users.

for programmatic access visit our API site

search string e.g. mgp128 or mgm4447970.3 | search

upload | download | analyze

hosted at
THE UNIVERSITY OF CHICAGO
and
Argonne NATIONAL LABORATORY

VITCOMIC2

Home
VITCOMIC2
Comparison

VITCOMIC1

VITCOMIC2 is a visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing.

Try VITCOMIC2

Metagenome/16S rRNA gene Amplicon Sequencing FASTA/FASTQ file: 選択... ファイルが選択されていません。
File format: FASTA flat FASTQ flat FASTA gzipped FASTQ gzipped
Conduct 16S rRNA gene Copy number normalization?: No Yes
Conduct 16S rRNA gene Assembly? (Shotgun metagenome only): No Yes
ID: (use [A-Za-z0-9-_])
Email:
upload&calculate clear

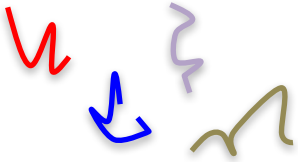
How to use

1. Input data

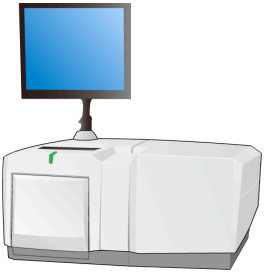
Both of a FASTA/FASTQ file and gzipped FASTA/FASTQ file are acceptable for the input data in the VITCOMIC2.
Example of a 16S rRNA gene Amplicon sequencing FASTA file from Turnbaugh et al 2009 and a shotgun metagenomic sequencing FASTQ.gz file from Nakamura et al 2016.

Metagenomic sequencing analysis (メタゲノム解析)

DNA extraction



DNA Sequencing



Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

Metagenomic reads

Assemble — MEGAHIT, MetaPlatanus IDBA-UD etc.

Contig sets

Gene finding — MGA, MetaGeneMark

Gene sets

BLASTP

KEGG Kyoto Encyclopedia of Genes and Genomes

UniProt

Sample1 Metadata

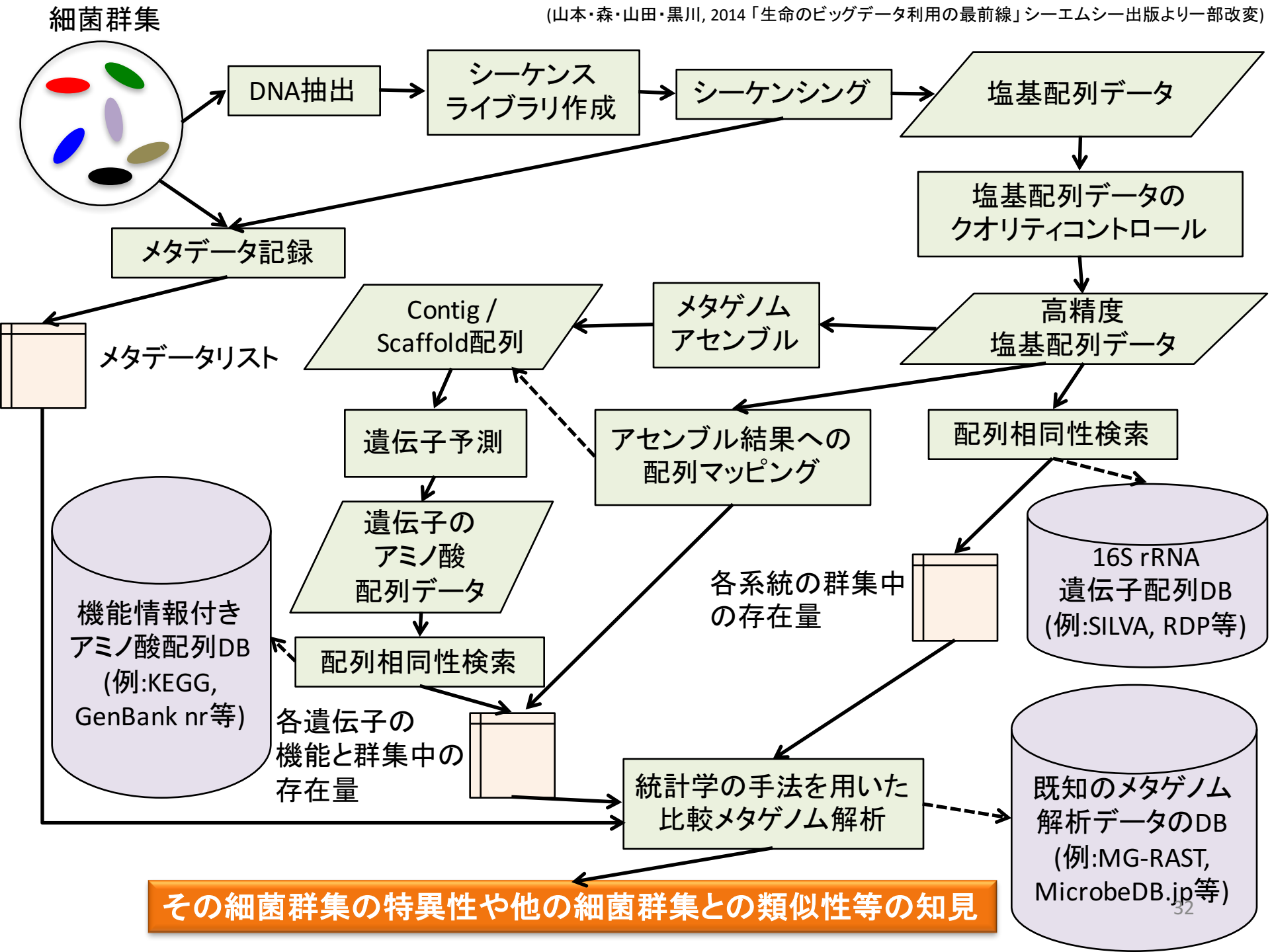
Sample2 Metadata

Comparative metagenomics

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

Who's there?

What are they doing?


細菌群集
(山本・森・山田・黒川, 2014「生命のビッグデータ利用の最前線」シーエムシー出版より一部改変)
DNA抽出
シーケンス
ライブラリ作成
シーケンシング
塩基配列データ
塩基配列データの
クオリティコントロール
メタデータ記録
メタデータリスト
Contig /
Scaffold配列
メタゲノム
アセンブル
高精度
塩基配列データ
遺伝子予測
アセンブル結果への
配列マッピング
配列相同性検索
遺伝子の
アミノ酸
配列データ
機能情報付き
アミノ酸配列DB
(例:KEGG,
GenBank nr等)
配列相同性検索
各系統の群集中
の存在量
16S rRNA
遺伝子配列DB
(例:SILVA, RDP等)
各遺伝子の
機能と群集中の
存在量
統計学の手法を用いた
比較メタゲノム解析
既知のメタゲノム
解析データのDB
(例:MG-RAST,
MicrobeDB.jp等)
その細菌群集の特異性や他の細菌群集との類似性等の知見

Next Generation Sequencer (NGS) (データはちょっと古いです)

Sequencer Name	Specific property	Read length (base)	Read number / run
ABI 3730xl	Sanger	500–1000	384
Ion Proton	Emulsion PCR	200	80,000,000
MiSeq	Bridge PCR	300	30,000,000
454 GS FLX+	Emulsion PCR	700-1,000	1,000,000
HiSeq 2500/4000	Bridge PCR	150 or 250	~3,200,000,000
PacBio RS II / Sequel	Single molecule	>7,000	>50,000
MinION	Single molecule	>数kb	>10,000

目的に応じたハードウェアのスペックの目安

原核・真核・メタゲノム・Transcriptomeアセンブル

特に重要なハードウェア: メモリ・CPU

- 原核生物のゲノムアセンブル　数十GBメモリ (50GB あれば十分)、1CPU 数コア
- Transcriptomeアセンブル　今のHiSeqのリード1億pairsなら、120GBあれば十分
  (例えばTrinityなら、100万pairsにつき1GBメモリが目安)
  (https://github.com/trinityrnaseq/trinityrnaseq/wiki/)
- メタゲノムのアセンブル　組成が単純な群集なら原核生物のアセンブルと同様
  群集が多様な場合、大量のリードが必要であり、
  数百GB-数TBメモリが必要な場合もある。 十数コア以上
- 真核生物のゲノムアセンブル　ゲノムサイズ、どれくらいheteroか、等に依存する
  数十GB-数TBメモリ、十数コア以上

少なくとも100GB以上のメモリをのせないと、アセンブルは辛い

目的に応じたハードウェアのスペックの目安

数千本以上の配列相同性検索

特に重要なハードウェア: CPU

- ゲノムやメタゲノムの遺伝子アノテーション
- 複数ゲノムの比較解析

入力配列が数千本以上になることが多く、計算を並列化して高速化する必要がある。

現在のnr相手のBLASTXやBLASTPは、並列化しても各プロセスで5GBほどメモリを使用するので、メモリもそれなりに必要。

CPU数、コア数が非常に重要、メモリもある程度必要

目的に応じたハードウェアのスペックの目安

マッピングツールでのReferenceゲノムへのマッピング

特に重要なハードウェア: ディスク

- Resequencing, RNA-Seq, ChIP-Seq解析におけるマッピング

リードのゲノムへのマッピングは高速でメモリ使用量も少ない。

結果のSAMファイルやBAMファイルが数十から数百GBになったりする。
samtools等でsortしたりすると、その規模のファイルが何個もできる。

マッピングを頻繁にするのなら、ディスクは少なくとも十数TBは必要

目的に応じたハードウェアのスペックの目安

1. 原核・真核・メタゲノム・Transcriptomeアセンブル

   特に重要なハードウェア: メモリ・CPU

2. 数千本以上の配列相同性検索

   特に重要なハードウェア: CPU

3. マッピングツールでのReferenceゲノムへのマッピング

   特に重要なハードウェア: ディスク

午後からの実習で用いるメタゲノムデータ

Backhed F. et al 2015の

4 daysの新生児の1サンプルを

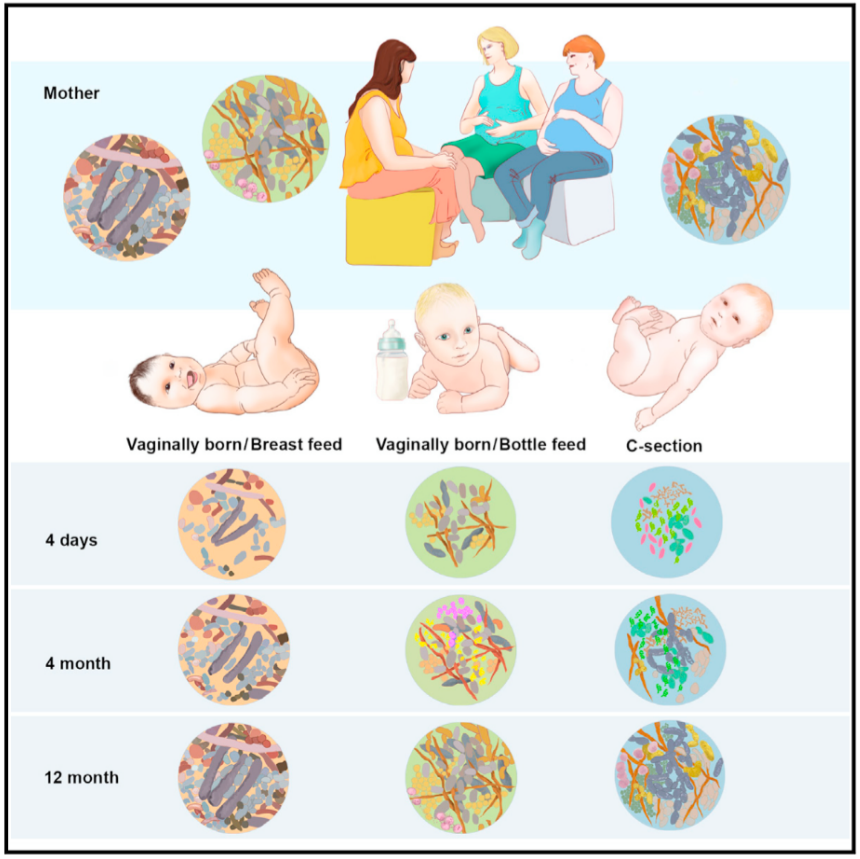
100万 pairに

ダウンサンプリングしたデータ

Cell Host & Microbe

Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life

Graphical Abstract



Authors

Fredrik Bäckhed, Josefine Roswall, ..., Jovanna Dahlgren, Jun Wang

Correspondence

fredrik.backhed@wlab.gu.se (F.B.),
jovanna.dahlgren@vgregion.se (J.D.),
wangj@genomics.org.cn (J.W.)

In Brief

Bäckhed et al. assessed the gut microbiomes of 98 Swedish mothers and their infants during the first year of life. Cessation of breast-feeding was identified as a major factor in determining gut microbiota maturation, with distinct shifts in signature species being hallmarks of its functional maturation.

Highlights

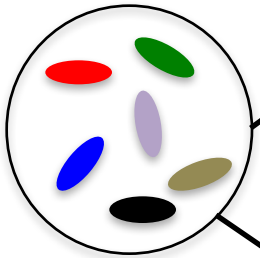
- Gut microbiomes of 98 mothers and their infants during the first year of life was assessed

Accession Numbers

ERP005989

ヒト乳児腸内細菌群集

(山本・森・山田・黒川, 2014「生命のビッグデータ利用の最前線」シーエムシー出版より一部改変)



DNA抽出

シーケンス
ライブラリ作成

シーケンシング

Illumina HiSeq 2000

塩基配列データ

100 万pairs

塩基配列データの
クオリティコントロール

Trimmomatic

高精度
塩基配列データ

MEGAHIT

メタゲノム
アセンブル

Contig配列

メタデータ記録

メタデータリスト

Prodigal

遺伝子予測

遺伝子の
アミノ酸
配列データ

アセンブル結果への
配列マッピング

Bowtie 2

配列相同性検索

BLASTN

16S rRNA
遺伝子配列DB
(例:SILVA, RDP等)

各系統の群集中
の存在量

機能情報付き
アミノ酸配列DB
(例:KEGG,
UniProt KB等)

BLASTP

配列相同性検索

各遺伝子の
機能と群集中の
存在量

統計学の手法を用いた
比較メタゲノム解析

既知のメタゲノム
解析データのDB
(例:MG-RAST,
MicrobeDB.jp等)

その細菌群集の特異性や他の細菌群集との類似性等の知見

今日の実習

- ほぼ全てターミナル上で行います
- cd
- cd Test
- TestがWorkingディレクトリです
- Test/には、Data/とProgram/があります
- コマンドは全てCommandMemo3.txtに書いてありますので、コピペで良いです
- マシンのメモリが8GB以下の場合は、Bio-Linuxは4GBで起動しましょう

今日の実習の参考資料

https://2017-ucsc-metagenomics.readthedocs.io/en/latest/

2017 Metagenomics workshop at UC Santa Cruz¶

Instructors: Harriet Alexander, Phillip Brooks, and C. Titus Brown

TAs: Luiz Irber, Shannon Joslin, Taylor Reiter

These are the online materials for a metagenomics workshop hosted by Marilou Sison-Mangus at UC Santa Cruz

Etherpad

Day 1: Noon to 5pm¶

Logging into the cloud (XSEDE Jetstream)

Evaluating your short-read data set quality

Assembling your short read data set with MEGAHIT

Day 2: Morning (9am - noon)¶

Software install for day 2

Binning genomes out of your metagenome

Quickly searching and comparing your samples with sourmash

Day 2: Afternoon (1:15pm - 5pm)¶

Annotating your short-read data set with Prokka

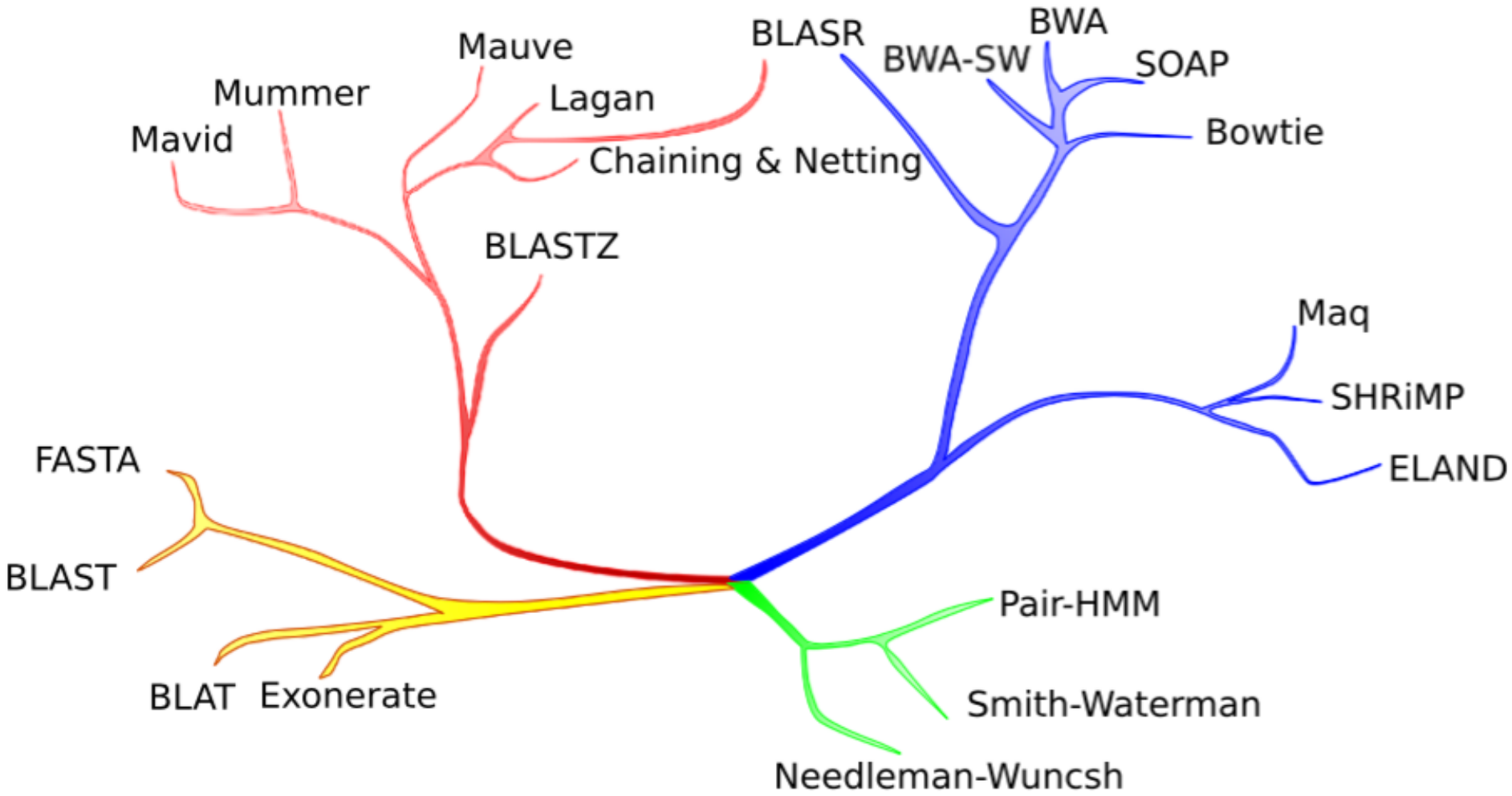
Quantifying abundance across samples with Salmon

A brief discussion of workflows & repeatability

PhiXについて

https://jp.illumina.com/content/dam/illumina-marketing/apac/japan/documents/pdf/2013_illumina_techsupport_session16.pdf

塩基配列の類似性検索ツールにも様々なものが存在する



Referenceと近ければMapper系 (Bowtie 2等) (Chaisson J. M. et al. 2012)
Referenceと遠いのなら、BLAST系

ゲノムアセンブルの二大戦略

- Overlap-Layout-Consensus
  - k-merの共有やローカルアラインメント等でリード間のoverlapを見つけて、短いContigを作成し、さらにContig間をoverlapをもとに結合（layout）。リードのoverlapを領域ごとに集めてマルチプルアラインメント等をしてconsensusをとることでアセンブルする
  - 例: Celera Assembler, Newbler, Mira, Canu
- de Bruijn Graph
  - リードをoverlapありのk-merに分割して、多数のリード間のk-merの共有をde Bruijn graphというグラフ構造で表現して、グラフ上で最短経路を見つける問題を解く

メタゲノムアセンブルツールの例

- IDBA-UD (Peng et al. 2012)
  - 短いk-merでアセンブルしてContig作成（Contig間のcoverageの差はある程度許容する）。そのContig群を用いて、もう少し長めのk-merでアセンブルしてContig作成。これを繰り返す。最後に、Contig間をまたがるpairリード（paired-endやmate pair)の情報をもとに、scaffoldingする。
  - 短いk-merでシーケンスエラー、長いk-merでリピートの問題に対処

- MEGAHIT (Li et al. 2015)
  - Contig作成の方法はIDBA-UDと類似しているが、de Bruijn graphの表現方法が簡素化されているため (http://alexbowe.com/succinct-debruijn-graphs/)、高速で省メモリ。また、coverageが小さいk-merの扱いについて色々と工夫している。scaffoldingはしない。

- metaSPAdes (Nurk et al. 2017)
  - Contig作成の方法はIDBA-UDと類似しているが、リードデータ中のstrainレベルの配列多様性をContig/Scaffoldにおいてもできるだけ保つために、サイトに多型があるとContigを分岐する傾向が強い。

今日のProkka

(https://github.com/tseemann/prokka/blob/master/README.md)

- MetaProdigal: CDS prediction
- Aragorn: tRNA

- CDSについては、機能アノテーションをするために、次の2 stepを行う。
  - BLASTPで、UniProt KB中のEvidence codeがreal protein or transcript evidenceとなっている、Prokaryote由来のproteinのアミノ酸配列に対して検索
  - HMMER3でPfamとTIGRFAMに対して検索

- 上記2 stepで閾値以上(BLASTならE-value <1e-6) でHitしなければ、hypothetical proteinとする。

メタゲノムデータからの系統組成推定

16S rRNA遺伝子を用いる手法

利点: Reference配列DBが充実（RDP, SILVA, GreenGenesなど)

欠点: ゲノム内コピー数の問題

系統マーカータンパク質の遺伝子(*rpoB*等)を用いる手法

利点: 1ゲノム1コピー

欠点: Reference配列DBが貧弱

リードやcontigのマッピング or k-mer組成を用いる手法

利点: Virus, 原核, 真核生物を同時解析可能

欠点

・Reference配列DBが貧弱

・Genus以下の精度がかなり落ちる

・水平伝播配列の扱い

パスウェイデータベース

- 先週行われた、AJACS河内の九州大の山西先生の資料が参考になります。

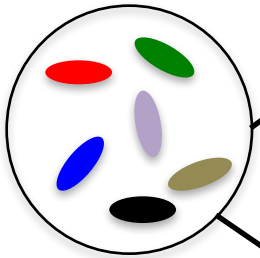
http://motdb.dbcls.jp/?AJACS66

http://motdb.dbcls.jp/?plugin=attach&pcmd=open&file=170824Pathway_DB_yamanishi_submit_r.pdf&refer=AJACS66

- メタゲノムでは、KEGGのKEGG Orthologyを遺伝子機能の単位として使うことが多い

ヒト乳児腸内細菌群集

(山本・森・山田・黒川, 2014「生命のビッグデータ利用の最前線」シーエムシー出版より一部改変)



DNA抽出

シーケンス
ライブラリ作成

シーケンシング

Illumina HiSeq 2000

塩基配列データ

100 万pairs

塩基配列データの
クオリティコントロール

Trimmomatic

高精度
塩基配列データ

MEGAHIT

メタゲノム
アセンブル

Contig配列

メタデータ記録

メタデータリスト

MetaProdigal

遺伝子予測

遺伝子の
アミノ酸
配列データ

アセンブル結果への
配列マッピング

Bowtie 2

配列相同性検索

BLASTN

16S rRNA
遺伝子配列DB
(例:SILVA, RDP等)

各系統の群集中
の存在量

機能情報付き
アミノ酸配列DB
(例:KEGG,
UniProt KB等)

BLASTP

配列相同性検索

各遺伝子の
機能と群集中の
存在量

統計学の手法を用いた
比較メタゲノム解析

既知のメタゲノム
解析データのDB
(例:MG-RAST,
MicrobeDB.jp等)

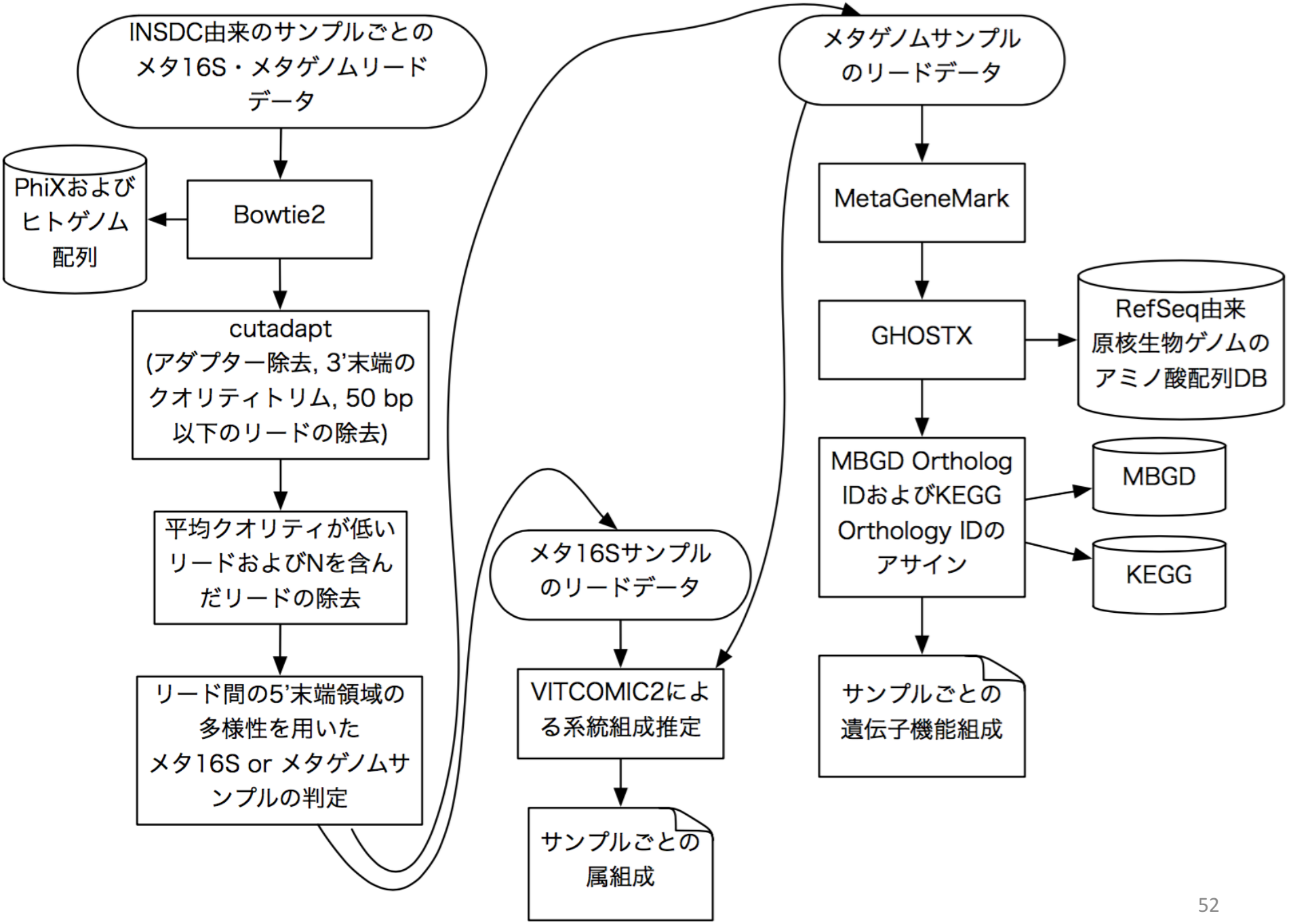
その細菌群集の特異性や他の細菌群集との類似性等の知見

代表的なメタゲノムデータベース

	運営者	配列データ	メタゲノム/メタ16S区別	系統組成	遺伝子機能組成	サンプル数 (2017 8月)	ゲノム等との統合化
NCBI Taxonomy + SRA https://www.ncbi.nlm.nih.gov/Taxonomy/	NCBI, USA	○	×	×	×	673,079	×
GOLD https://gold.jgi.doe.gov/	JGI, USA	×	×	×	×	24,922	×
IMG/M https://img.jgi.doe.gov/cgi-bin/m/main.cgi	JGI, USA	○（Reads + Contigs)	○（メタゲノムのみ収録）	○	○	7,982	△
MG-RAST http://metagenomics.anl.gov/index.html	Chicago U. USA	○	○	○	○	47,313 (303,594)	×
EBI-Metagenomics https://www.ebi.ac.uk/metagenomics/	EBI, EU	×	○	○	○	74,342	×
MicrobeDB.jp http://microbedb.jp	NIG, Japan	×	○	○	○	60,551 (173,359)	○

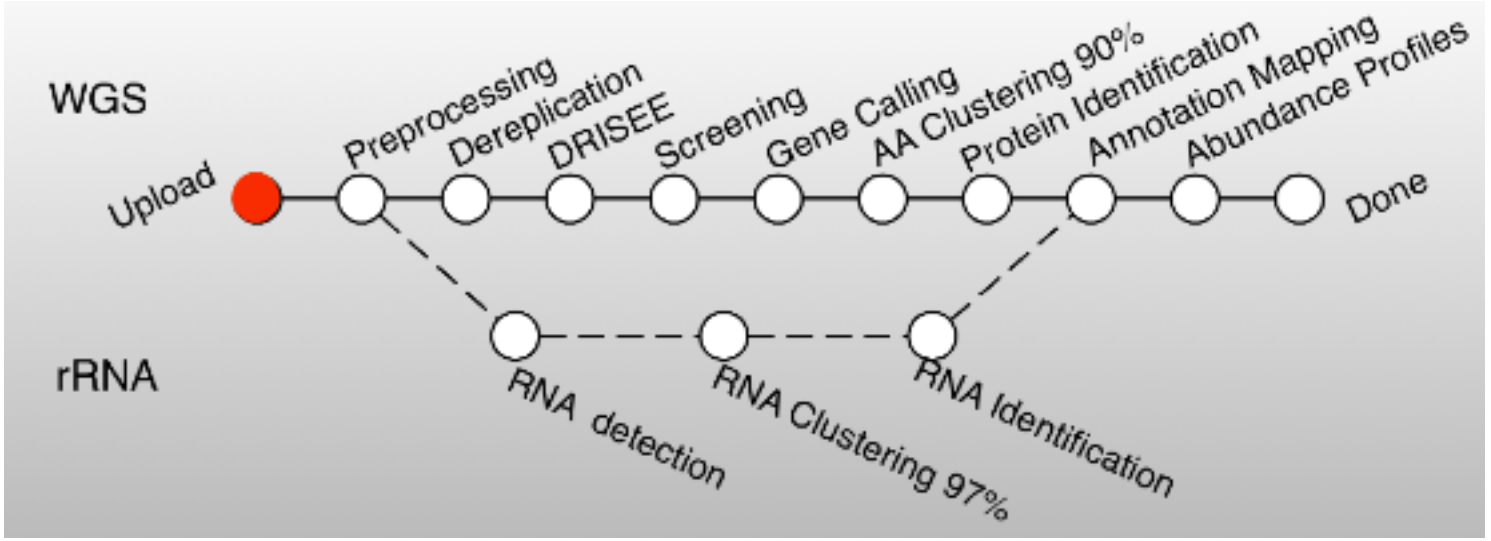
微生物統合DB「MicrobeDB.jp」

- 微生物に関するデータを系統・遺伝子・環境の３つの軸に沿って整理・統合し、フルRDFのDBを構築
- 約90億トリプルから構成
- 12種類のオントロジー＆ボキャブラリの開発
- 公開済みの約6万サンプルのメタ16S・メタゲノムデータ、約1万7千株のゲノム・ドラフトゲノムデータを収録
- 195種類のStanzaの開発
- 解析プロトコルの標準化および解析パイプラインの開発
- 単細胞の真菌類・藻類のゲノムデータも整理・統合
- 自動更新技術の開発


INSDC由来のサンプルごとのメタ16S・メタゲノムリードデータ
PhiXおよびヒトゲノム配列
Bowtie2
cutadapt (アダプター除去, 3'末端のクオリティトリム, 50 bp以下のリードの除去)
平均クオリティが低いリードおよびNを含んだリードの除去
リード間の5'末端領域の多様性を用いたメタ16S or メタゲノムサンプルの判定
メタ16Sサンプルのリードデータ
VITCOMIC2による系統組成推定
サンプルごとの属組成
メタゲノムサンプルのリードデータ
MetaGeneMark
GHOSTX
RefSeq由来原核生物ゲノムのアミノ酸配列DB
MBGD Ortholog IDおよびKEGG Orthology IDのアサイン
MBGD
KEGG
サンプルごとの遺伝子機能組成

MG-RAST v.3 Pipeline

ftp://ftp.metagenomics.anl.gov/data/manual/mg-rast-manual.pdf



Preprocessing: SolexaQA (Average QV, Length, N, 3' Trim)
Metagenome or Amplicon: Calculate Shannon entropy of first 20 sequence in reads
Dereplication and DRISEE: Identify duplicate in which first 50 bp identical reads >20 times
Screening: Bowtie2 against specified organism genome to remove host genome
Gene Calling: FragGeneScan (>75 bp)
AA Clustering: UCLUST (AA Identity 90%, representative sequence is the longest one)
Protein Identification: sBLAT (OpenMP parallelzation) against M5nr (GenBank, SEED, IMG, UniProt, KEGG, eggNOG)
Annotation Mapping: SEED Subsystems, IMG terms, COG, eggNOGs, GO
Abundance Profiles: E-value, Identity, Alignment length can be specified

rRNA pipeline
BLAT search against 90% clustered SILVA. Identified reads are then clustered at 97% identity. Longest sequence is the representative of the cluster. BLAT searched against the M5rna (SILVA, Greengenes, RDP)