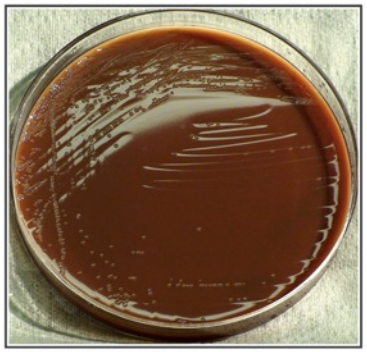
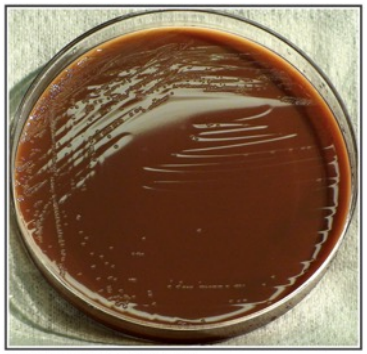
AJACS 十勝2

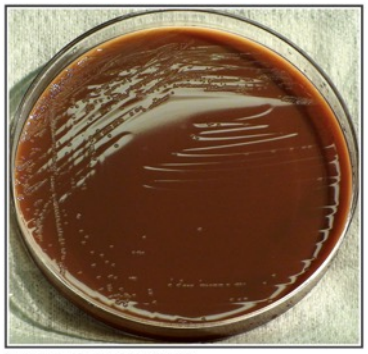
2019年9月25日

メタゲノム解析ツール

森　宙史, Ph.D.
国立遺伝学研究所
情報研究系
hmori@nig.ac.jp







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

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

これらの手法では細菌群集についての断片的な情報しか得られない

What's metagenomics?

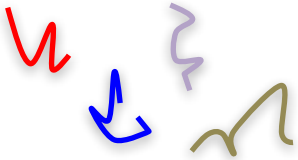
- **Microflora, microbiota, microbial community**: 微生物群集
  Total collection of microorganisms within a community

- **Metagenome**: ある群集の遺伝情報の総体
  Total genomic potential of a community
  [Handelsman et al. 1998, Chem. & Biol.]

- **Microbiome**: マイクロバイオーム
  Microbiota and metagenome in a microbial community

amplicon sequencing analysis
(アンプリコン解析, 16S rRNA遺伝子のアンプリコン解析)

DNA extraction



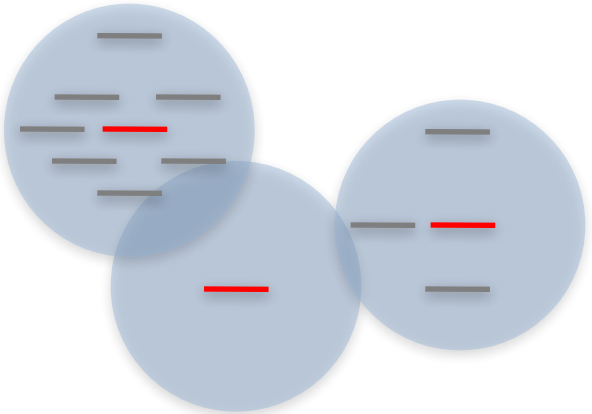
PCR amplification

DNA Sequencing

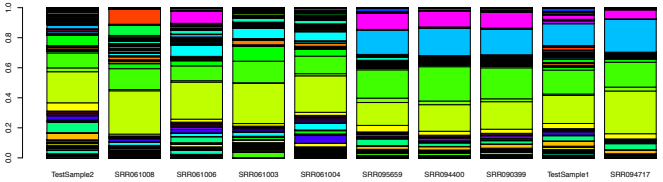
2016 DBCLS TogoTV / CC-BY-4.0

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



Taxonomic assignment and
Comparison between samples

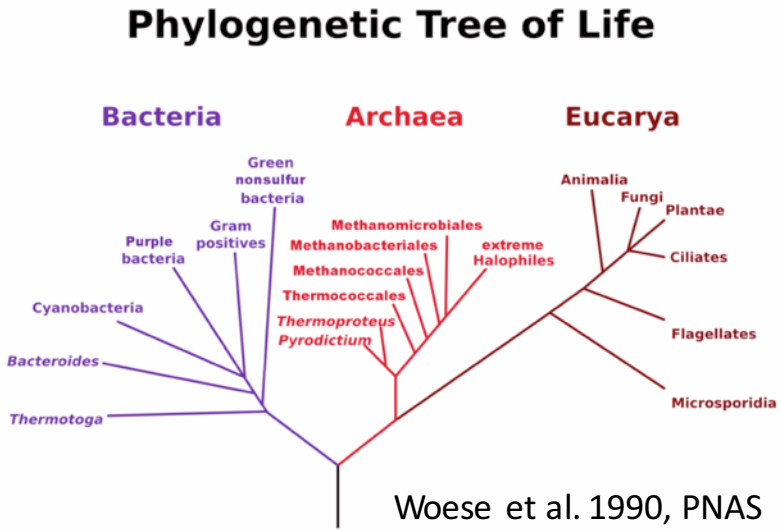


Who's there?

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 (保存されているサイトと多様なサイトがモザイク状に存在)
- 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
- 多くの細菌がゲノム内に複数の遺伝子コピーを所持
- 全長約1500 base

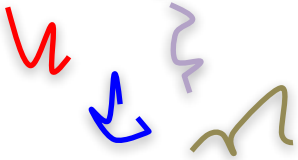
16S rRNA遺伝子は広範囲の細菌における系統推定を行う上で適した遺伝子



Woese et al. 1990, PNAS

Metagenomic sequencing analysis
(メタゲノム解析, ショットガンメタゲノム解析)

DNA extraction



DNA Sequencing

2016 DBCLS TogoTV / CC-BY-4.0

Metagenomic reads

Assemble — MEGAHIT, metaSPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

Sequence similarity search — BLASTP, DIAMOND, MMseqs2, etc.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

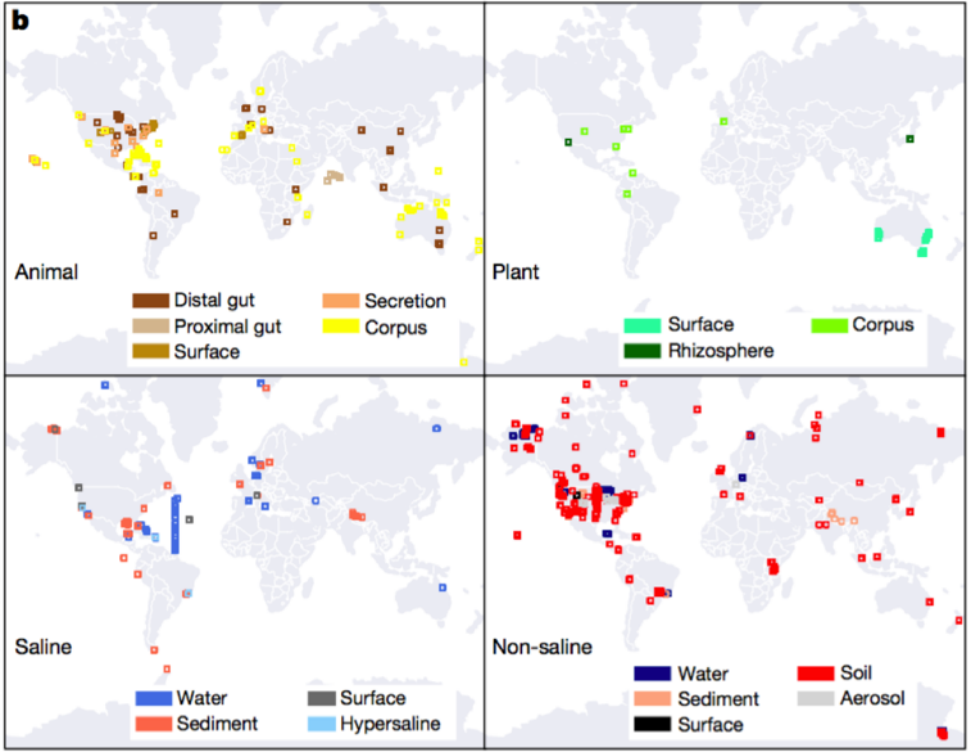
UniProt

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

Sample1 Metadata
Sample2 Metadata
Comparative metagenomics

Who's there?

What are they doing?



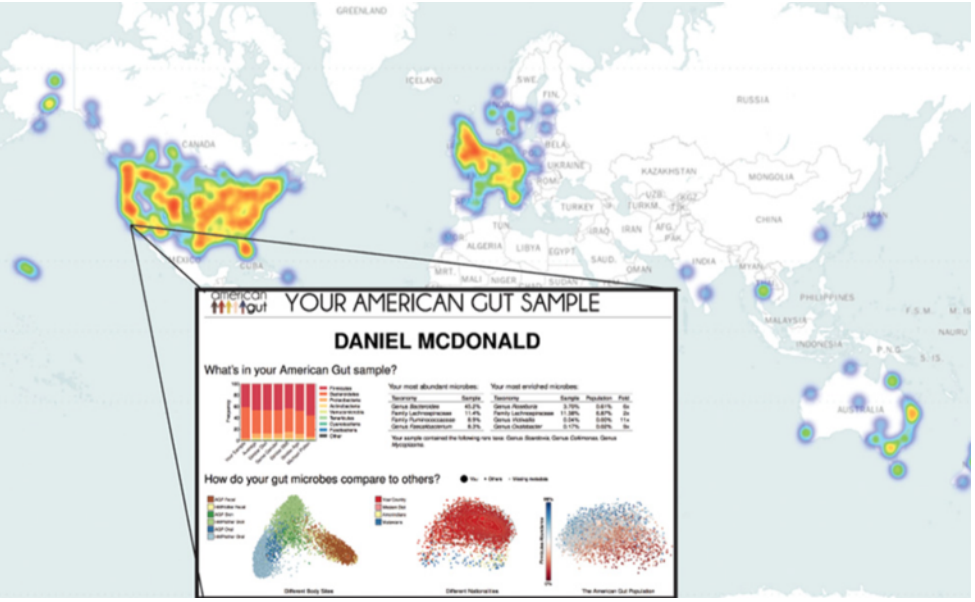
Earth Microbiome Project

Thompson et al. 2017, Nature

Soil, Sea, Pond, Animal, etc.

23,828 samples

Standardized experimental and bioinformatics procedure



American Gut Project

McDonald et al. 2018, mSystems

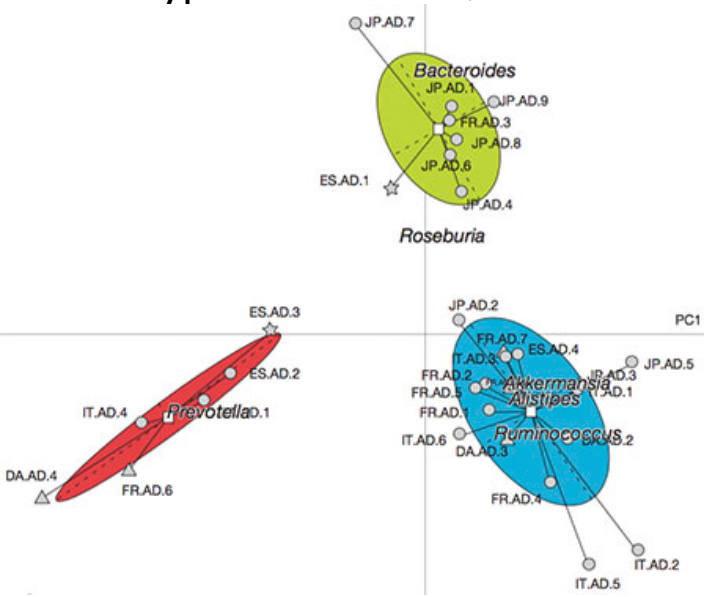
Human feces

9,511 samples

Standardized experimental and bioinformatics procedure.

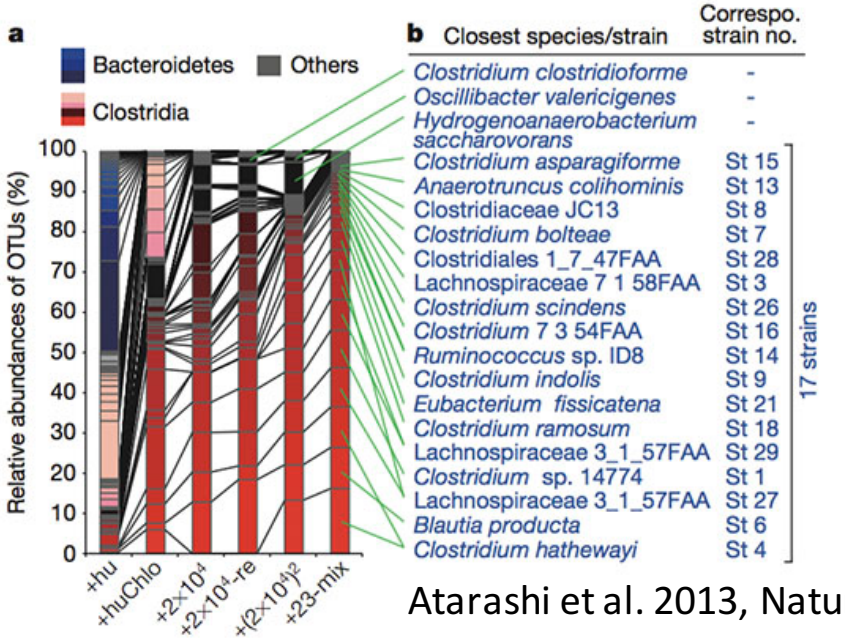
Cloud founding

Enterotype: コホート研究



Arumugam et al. 2011, Nature

腸管免疫誘導細菌群: 宿主との相互作用



Atarashi et al. 2013, Nature

糞便移植: 環境コントロール

The NEW ENGLAND JOURNAL of MEDICINE

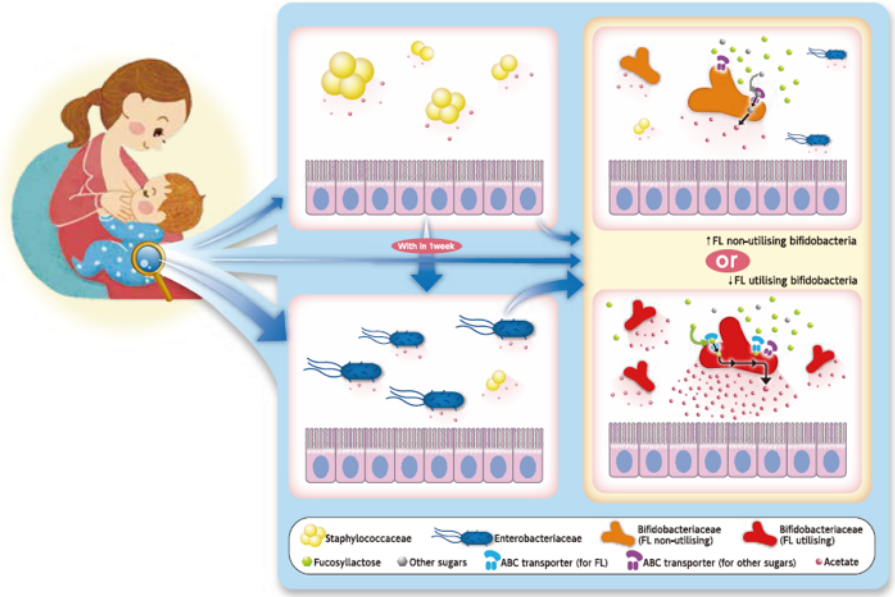
ESTABLISHED IN 1812 JANUARY 31, 2013 VOL. 368 NO. 5

Duodenal Infusion of Donor Feces for Recurrent *Clostridium difficile*

Els van Nood, M.D., Anne Vrieze, M.D., Max Nieuwdorp, M.D., Ph.D., Susana Fuentes, Ph.D., Erwin G. Zoetendal, Ph.D., Willem M. de Vos, Ph.D., Caroline E. Visser, M.D., Ph.D., Ed J. Kuijper, M.D., Ph.D., Joep F.W.M. Bartelsman, M.D., Jan G.P. Tijssen, Ph.D., Peter Speelman, M.D., Ph.D., Marcel G.W. Dijkgraaf, Ph.D., and Josbert J. Keller, M.D., Ph.D.

van Nood et al. 2013, NEJM

母乳オリゴ糖分解菌: 母親-乳児-細菌の相互作用



Matsuki et al. 2016, Nature Comm.

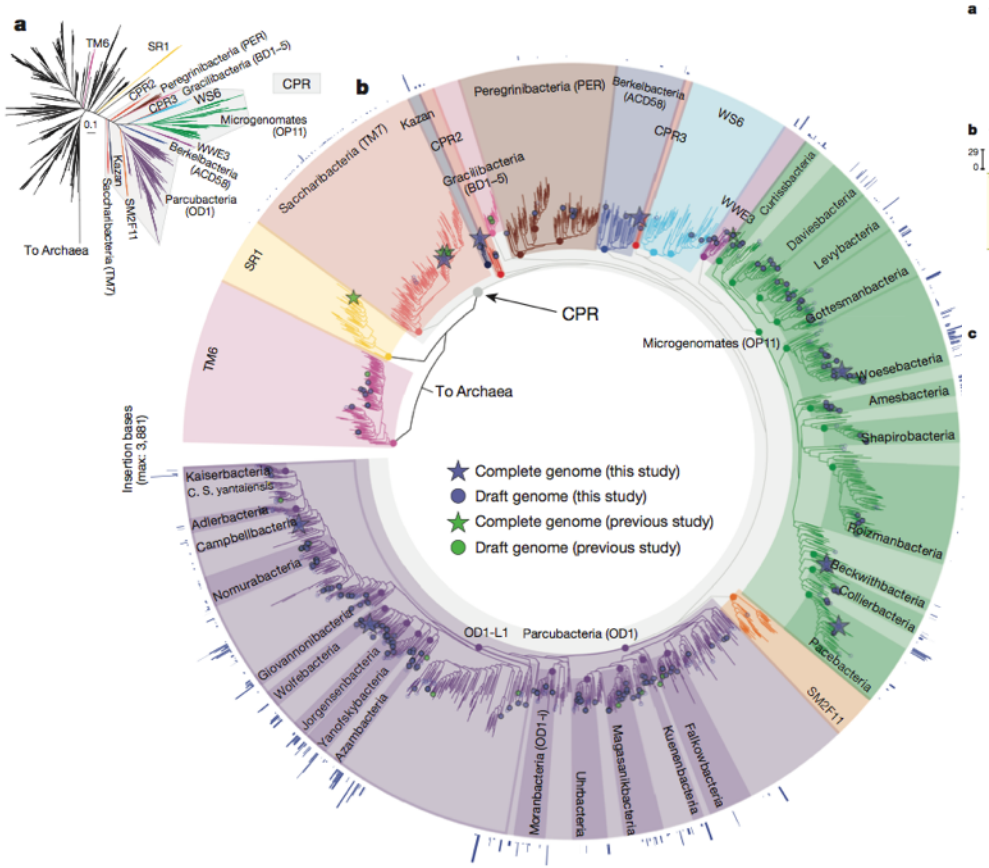
LETTER

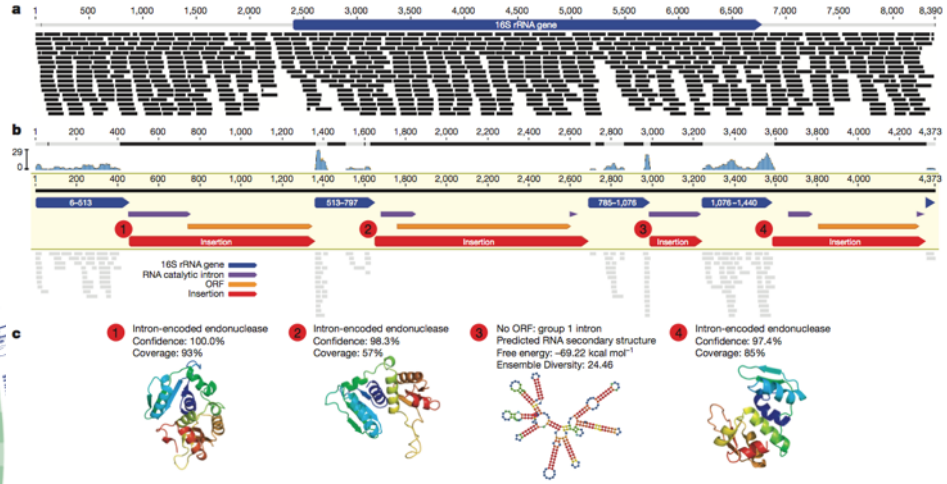
Nature, 2015

doi:10.1038/nature14486

Unusual biology across a group comprising more than 15% of domain Bacteria

Christopher T. Brown[1], Laura A. Hug[2], Brian C. Thomas[2], Itai Sharon[2], Cindy J. Castelle[2], Andrea Singh[2], Michael J. Wilkins[3,4], Kelly C. Wrighton[4], Kenneth H. Williams[5] & Jillian F. Banfield[2,5,6]





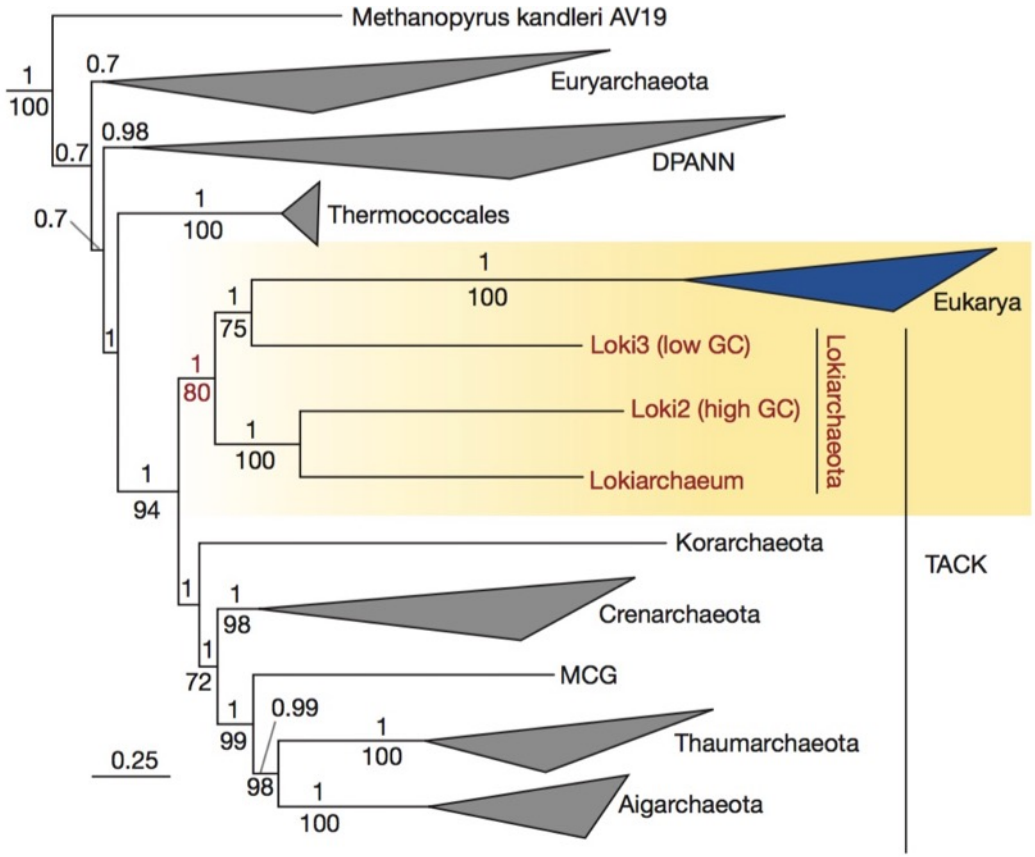
ARTICLE

Nature, 2015

doi:10.1038/nature14447

Complex archaea that bridge the gap between prokaryotes and eukaryotes

Anja Spang[1]*, Jimmy H. Saw[1]*, Steffen L. Jørgensen[2]*, Katarzyna Zaremba-Niedzwiedzka[1]*, Joran Martijn[1], Anders E. Lind[1], Roel van Eijk[1]†, Christa Schleper[2,3], Lionel Guy[1,4] & Thijs J. G. Ettema[1]



アンプリコン解析もメタゲノム解析も、Webブラウザ上やソフトウェア上でのマウスクリックのみでは解析困難

UNIXのスキルや自分のマシンでの解析スキル等が、意義ある結果を得るためにはほぼ必須

それらは、例えば先進ゲノム支援の情報解析初級者講習会の動画と資料等で勉強可能
https://www.genome-sci.jp/lecture201903

Difficulties of metagenome from a bioinformatics view point

Hypothesis testing or Data-driven?

Amplicon or Metagenome?

Short or Long reads?

Taxonomic assignment strategy?

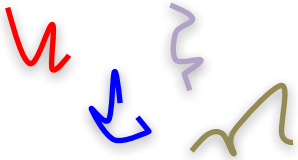
Metagenome-informatics analysis strategy?

Comparison between projects?

アンプリコン解析のツール

amplicon sequencing analysis
(アンプリコン解析, 16S rRNA遺伝子のアンプリコン解析)

DNA extraction



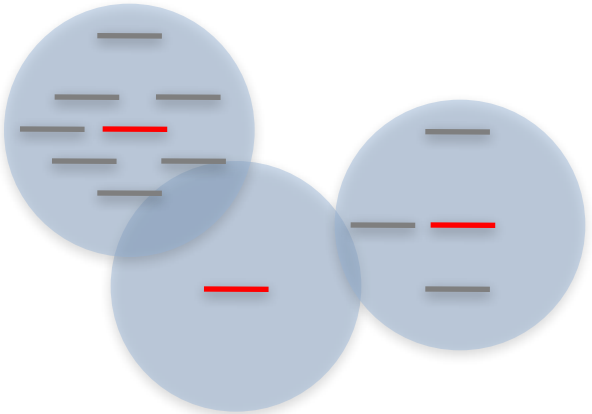
PCR amplification

DNA Sequencing

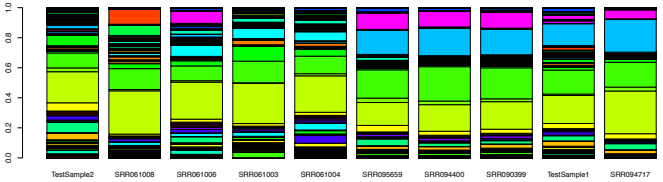
2016 DBCLS TogoTV / CC-BY-4.0

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



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

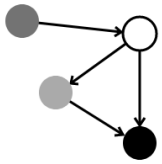
  原核生物: 16S rDNA – ITS – 23S rDNA – 5S rDNA

- 細菌の16S rRNA遺伝子のUniversal primerについては、万能なものは存在しない
- 341F-806R, 515F-806Rなどがよく使われている
- CpGメチル化の有無等で真核生物由来rRNA遺伝子と原核生物由来rRNA遺伝子を分画してくれるキットも存在

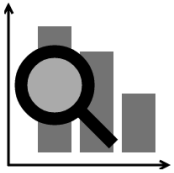
QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed.

https://qiime2.org/

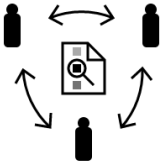
Learn more » Citing QIIME 2 »



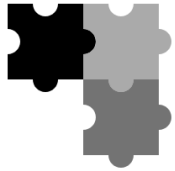
Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!



Interactively explore your data with beautiful visualizations that provide new perspectives.



Easily share results with your team, even those members without QIIME 2 installed.



Plugin-based system — your favorite microbiome methods all in one place.

Choose the interface that fits your needs

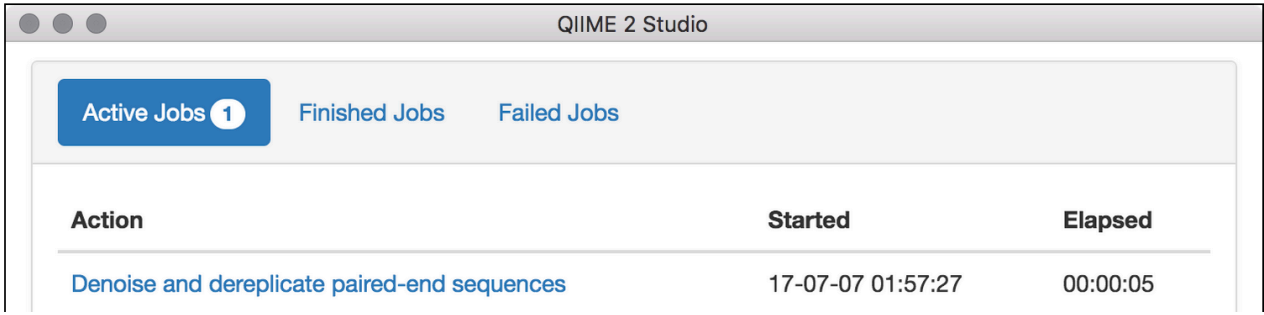
q2cli the command line interface

```
2. ~ (zsh)
$ qiime info
System versions
Python version: 3.5.3
QIIME 2 release: 2017.6
QIIME 2 version: 2017.6.0
q2cli version: 2017.6.0

Installed plugins
alignment 2017.6.0
composition 2017.6.0
dada2 2017.6.0
```

q2studio the graphical user interface (PROTOTYPE)

q2studio is a functional prototype of a graphical user interface for QIIME 2, and is not necessarily feature-complete with respect to q2cli and the Artifact API.



Quantitative Insights Into Microbial Ecology

https://docs.qiime2.org/2019.7/plugins/

Version: 2019.7

Table of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
- Tutorials
- Interfaces
- Plugins
  - Available plugins
  - Future plugins
  - Developing a QIIME 2 plugin
- Semantic types
- Community
- Data resources
- Supplementary resources
- Glossary
- Citing QIIME 2

Quick search

Go

Plugins

The following pages describe what QIIME 2 plugins are available and how to develop a new plugin.

- Available plugins
  - alignment: Plugin for generating and manipulating alignments.
  - composition: Plugin for compositional data analysis.
  - cutadapt: Plugin for removing adapter sequences, primers, and other unwanted sequence from sequence data.
  - dada2: Plugin for sequence quality control with DADA2.
  - deblur: Plugin for sequence quality control with Deblur.
  - demux: Plugin for demultiplexing & viewing sequence quality.
  - diversity: Plugin for exploring community diversity.
  - emperor: Plugin for ordination plotting with Emperor.
  - feature-classifier: Plugin for taxonomic classification.
  - feature-table: Plugin for working with sample by feature tables.
  - fragment-insertion: Plugin for extending phylogenies.
  - gneiss: Plugin for building compositional models.
  - longitudinal: Plugin for paired sample and time series analyses.
  - metadata: Plugin for working with Metadata.
  - phylogeny: Plugin for generating and manipulating phylogenies.
  - quality-control: Plugin for quality control of feature and sequence data.
  - quality-filter: Plugin for PHRED-based filtering and trimming.
  - sample-classifier: Plugin for machine learning prediction of sample metadata.
  - taxa: Plugin for working with feature taxonomy annotations.
  - types: Plugin defining types for microbiome analysis.
  - vsearch: Plugin for clustering and dereplicating with vsearch.

https://docs.qiime2.org/2019.7/

qiime 2 docs

Version: 2019.7

Table of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
- Tutorials
- Interfaces
- Plugins
- Semantic types
- Community
- Data resources
- Supplementary resources
- Glossary
- Citing QIIME 2

Quick search

Go

QIIME 2 user documentation

This site is the official user documentation for QIIME™ 2, including installation instructions, tutorials, and other important information. Visit http://qiime.org for information on QIIME™ 1.

Getting started

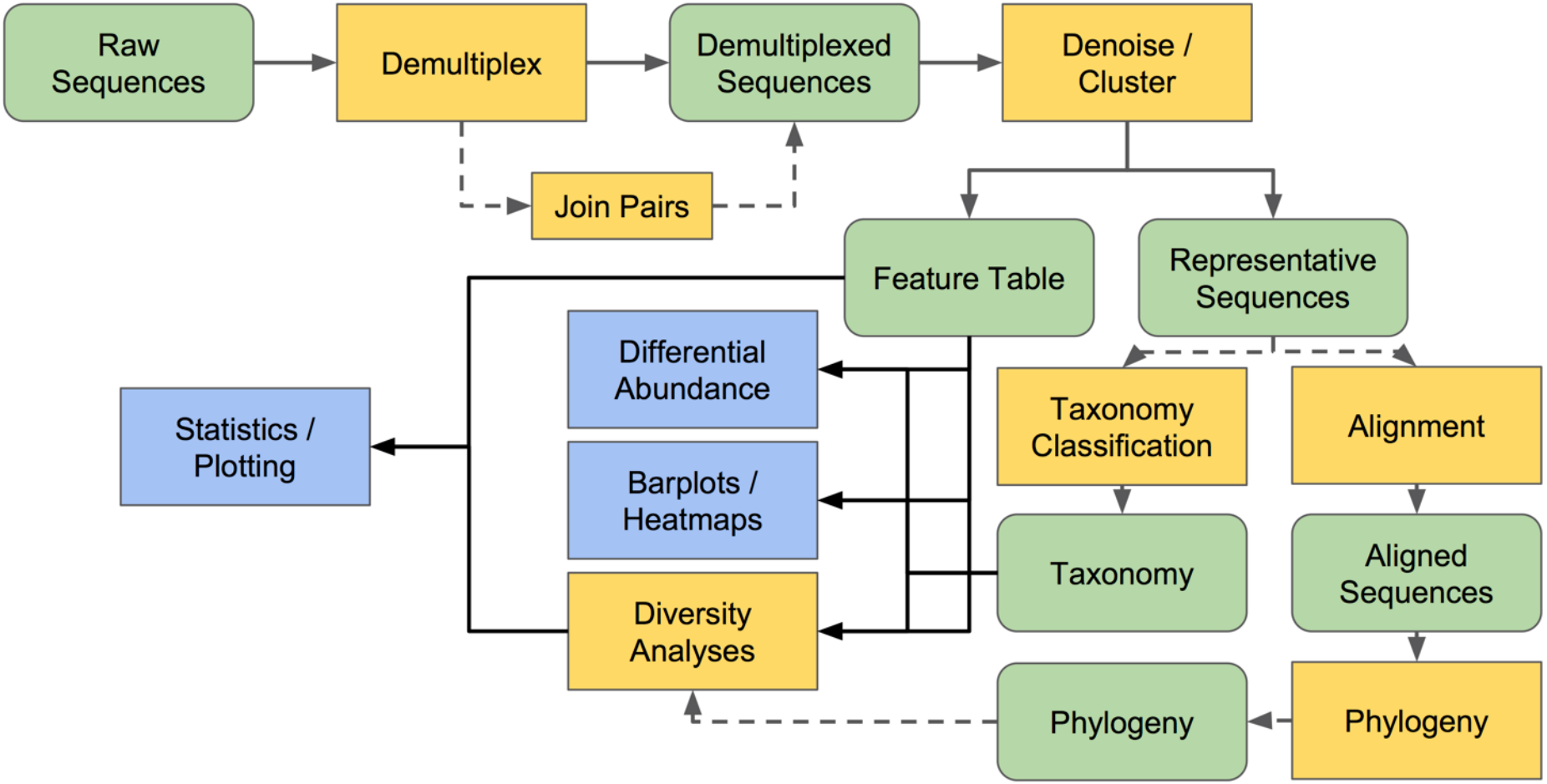
Check out the getting started guide to begin using QIIME 2.

Table of contents

- Getting started
- What is QIIME 2?
- Core concepts
  - Data files: QIIME 2 artifacts
  - Data files: visualizations
  - Semantic types
  - Plugins
  - Methods and visualizers
  - Next steps
- Installing QIIME 2
  - Natively installing QIIME 2
  - Installing QIIME 2 using Virtual Machines
  - QIIME 2 Core 2019.7 distribution
- Tutorials
  - Overview of QIIME 2 Plugin Workflows
  - QIIME 2 for Experienced Microbiome Researchers
  - “Moving Pictures” tutorial
  - Fecal microbiota transplant (FMT) study: an exercise
  - “Atacama soil microbiome” tutorial
  - Parkinson’s Mouse Tutorial
  - Differential abundance analysis with gneiss

QIIME2の解析ワークフローの基本

https://docs.qiime2.org/2019.7/tutorials/overview/



BIOF 089 | Microbiome Bioinformatics with QIIME2

Stay Connected

January 8, 2020 to January 10, 2020

Registration occurs on a first-come, first-served basis. The deadline for registration is one week before the first day of the course. If you are unable to register before the deadline, please email: **registrar@faes.org**or call 301-496-7977 for space availability.

NIH Only: Payment approval and authorization is done through your AO or lab manager. Receiving lab approval does not constitute enrollment. Fellows or those being sponsored by their lab can enroll using this **form** while waiting for authorization of payment.

Register Now

Contract: SF-182

Course Description:
Members of the QIIME development group will teach a three-day hands-on workshop on bioinformatics tools for microbial ecology. The workshop will include lectures covering basic QIIME usage and theory, and hands-on work with QIIME to perform microbiome analysis from raw sequence data through publication-quality statistics and visualizations. The workshop will also cover related bioinformatics tools including DADA2, Emperor, scikit-bio and an Introduction to Applied Bioinformatics. This workshop will provide the foundation on which students can begin using these tools to advance their own studies of microbiome analysis or microbial ecology.

This is a hands-on workshop. Participants must bring their laptop.

Early Bird Rates until September 16, 2019
General Training Rate
$745.00

Discounted Training Rates
$345.00-NIH Trainees(Fellows, PostDocs, PostBacs working at any NIH Campus ONLY)
$445.00 – Students and Postdocs (Non-NIH)
$545.00-NIH Community (Working, Appointed, or Assigned to any NIH Campus ONLY)
$645.00-Academia, Government, Military

On and After September 16, 2019
General Training Rate
$1045.00

https://faes.org/BIOF089.jan20

VITCOMIC2 http://vitcomic.org

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. Sample 16S rRNA gene Amplicon sequencing fastq data.

2. File format

File format is a file format identifier of your FASTA/FASTQ file. To reduce the size of your file, we strongly recommend that you compress your file with gzip. If you don't compress your file, please choose "flat file".

(Mori et al. 2018, BMC Systems Biol.)

Bioinformatics methodology difference of amplicon sequencing before and after NGS

Before NGS: (100-1,000 reads / sample)

multiple alignment & phylogenetic tree based analysis (e.g., UniFrac, DOTUR, ARB, Bellerophon)

Accurate but slow

After NGS: (10,000-100,000 reads / sample)

sequence clustering & sequence similarity search (e.g., CD-HIT, USEARCH/UCLUST/UPARSE, Fast UniFrac, PyNAST, RDP classifier)

Fast but accuracy?

いわゆるOperational taxonomic unit (OTU)クラスタリング

To avoid sequence clustering ambiguity problem ...

- **Only conduct taxonomic assignment approach**
  e.g., VITCOMIC2, MAPSeq, SortMeRNA
  1. conduct sequence similarity search of all reads against reference

- **Denoising & Derep. approach**
  e.g., DADA2 (R), Deblur
  1. remove sequencing error in reads by model-based approach
  2. dereplicate complete match reads to one representative read
  3. conduct taxonomic assignment

Denoising & Derep. approachの結果得られた完全マッチクラスタを
Amplicon sequence variant (ASV), Exact sequence variant (ESV),
Zero-radius OTU (ZOTU)等と呼んでOTUと区別する

DADA2 Pipeline Tutorial (1.12)

Here we walk through version 1.12 of the DADA2 pipeline on a small multi-sample dataset. Our starting point is a set of Illumina-sequenced paired-end fastq files that have been split (or "demultiplexed") by sample and from which the barcodes/adapters have already been removed. The end product is an **amplicon sequence variant (ASV) table**, a higher-resolution analogue of the traditional OTU table, which records the number of times each exact amplicon sequence variant was observed in each sample. We also assign taxonomy to the output sequences, and demonstrate how the data can be imported into the popular phyloseq R package for the analysis of microbiome data.

https://benjjneb.github.io/dada2/tutorial.html

Starting point

This workflow assumes that your sequencing data meets certain criteria:

- Samples have been demultiplexed, i.e. split into individual per-sample fastq files.
- Non-biological nucleotides have been removed, e.g. primers, adapters, linkers, etc.
- If paired-end sequencing data, the forward and reverse fastq files contain reads in matched order.

If these criteria are not true for your data (**are you sure there aren't any primers hanging around?**) you need to remedy those issues before beginning this workflow. See the FAQ for recommendations for some common issues.

Getting ready

First we load the `dada2` package. If you don't already have it, see the dada2 installation instructions.

```
library(dada2); packageVersion("dada2")
```

```
## [1] '1.12.1'
```

Older versions of this workflow associated with previous release versions of the dada2 R package are also available: 1.4, 1.6, 1.8.

The data we will work with are the same as those used in the mothur MiSeq SOP. To follow along, download the example data and unzip. These fastq files were generated by 2x250 Illumina Miseq amplicon sequencing of the V4 region of the 16S rRNA gene from gut samples collected longitudinally from a mouse post-weaning. For now just consider them paired-end fastq files to be processed. Define the following path variable so that it points to the extracted directory on **your** machine:

```
path <- "~/MiSeq_SOP" # CHANGE ME to the directory containing the fastq files after unzipping.
list.files(path)
```

Taxonomic reference data

The `assignTaxonomy` and `assignSpecies` functions require appropriately formatted fasta files describing the set of taxonomically assigned sequences to use as a training dataset. We provide appropriately formatted versions of several popular taxonomic databases, and describe the dada2-specific format for those who wish to use a custom database.

DADA2-formatted reference databases

We maintain reference fastas for the three most common 16S databases: Silva, RDP and GreenGenes. The dada2 package recognizes and parses the General Fasta releases of the UNITE project for ITS taxonomic assignment. Formatted versions of other databases can be "contibuted" and will be made available through this page if referencable by doi (eg. deposited at Zenodo or Figshare).

Please note that the files provided here are just derivative reformattings of these taxonomic databases. If using these files for taxonomic assignment, the source database should also be cited.

https://benjjneb.github.io/dada2/training.html

Maintained:

- **Silva version 132**, **Silva version 128**, Silva version 123 (Silva dual-license)
- **RDP trainset 16**, RDP trainset 14
- **GreenGenes version 13.8**
- **UNITE (use the General Fasta releases)**

Contributed:

- RefSeq + RDP (NCBI RefSeq 16S rRNA database supplemented by RDP)
  - Reference files formatted for `assignTaxonomy`
  - Reference files formatted for `assignSpecies`
- GTDB: Genome Taxonomy Database (More info: http://gtdb.ecogenomic.org/)
  - Reference files formatted for `assignTaxonomy`
  - Reference files formatted for `assignSpecies`
- HitDB version 1 (Human InTestinal 16S rRNA)
- RDP fungi LSU trainset 11
- Silva Eukaryotic 18S, v132 & v128
- PR2 version 4.7.2+. SEE NOTE BELOW.

16S rRNA gene reference DB

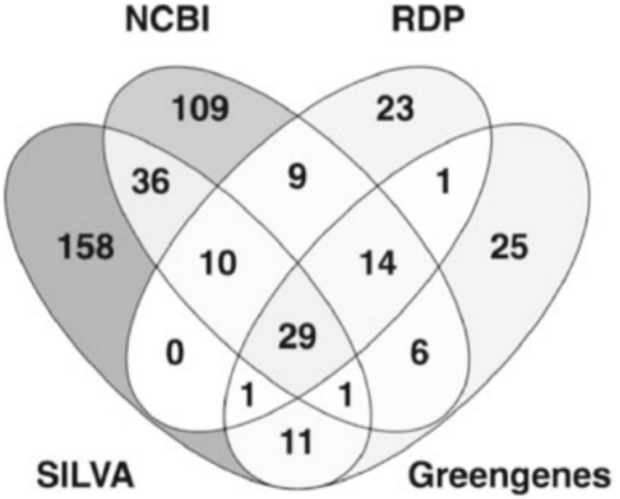
- **SILVA**
  (https://www.arb-silva.de/)
  Taxonomy: List of Prokaryotic Names with Standing in Nomenclature (LPSN)
  License: academic free
  Last update: Dec. 2017

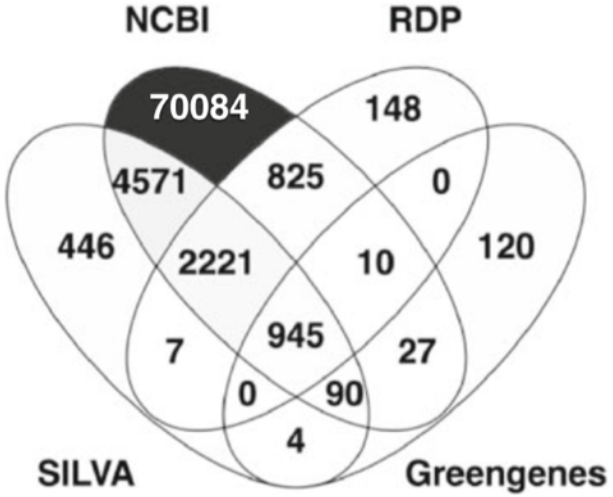
Phylum

- **Greengenes**
  (http://greengenes.secondgenome.com/)
  Taxonomy: NCBI Taxonomy (modified)
  License: CC BY SA
  Last update: Oct. 2013

- **RDP**
  (https://rdp.cme.msu.edu/)
  Taxonomy: Bergey's Manual
  License: CC BY SA
  Last update: Sep. 2016?

Genus

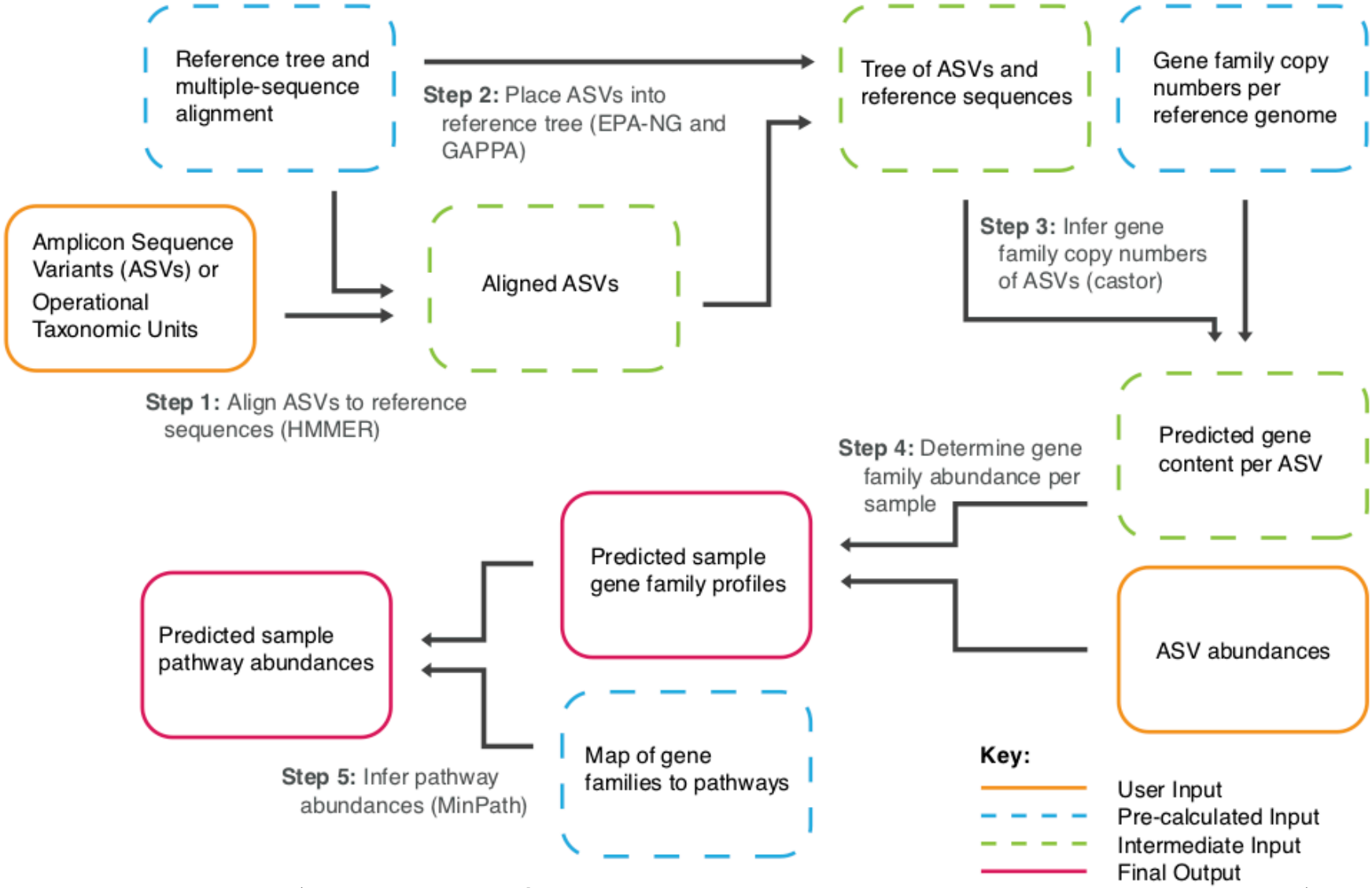




Balvociute & Huson, 2017, BMC Genomics

アンプリコン解析データからの遺伝子機能組成推定

https://github.com/picrust/picrust2/wiki

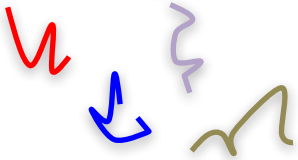


Reference genomeが決まっており遺伝子機能アノテーションもしっかりしている系統が多い環境ではある程度効果を発揮する

メタゲノム解析のツール

Metagenomic sequencing analysis
(メタゲノム解析, ショットガンメタゲノム解析)

DNA extraction



DNA Sequencing

2016 DBCLS TogoTV / CC-BY-4.0

Metagenomic reads

Assemble — MEGAHIT, metaSPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

Sequence similarity search — BLASTP, DIAMOND, MMseqs2, etc.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

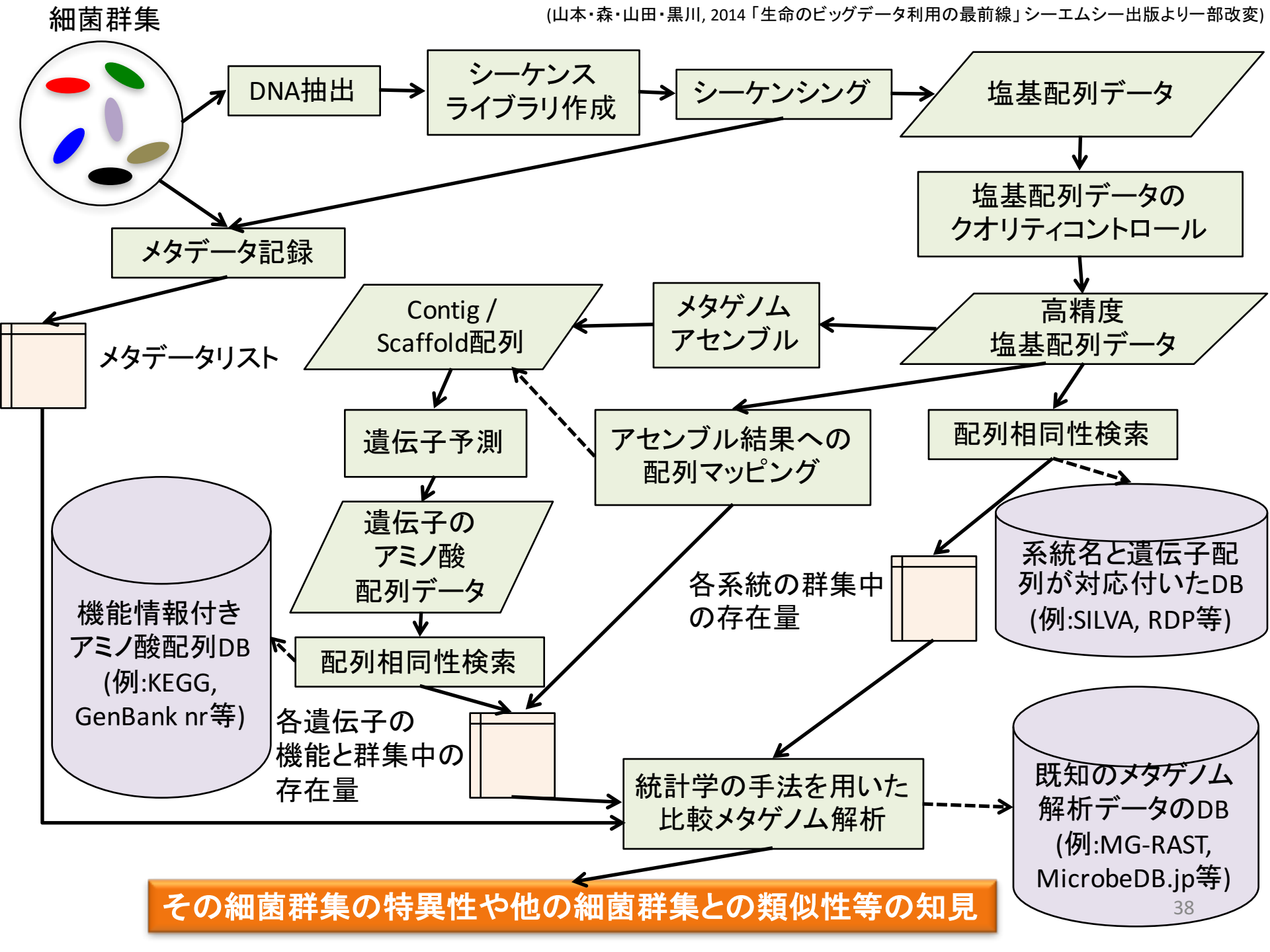
UniProt

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

Sample1 Metadata
Sample2 Metadata
Comparative metagenomics

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
系統名と遺伝子配
列が対応付いた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

- メタゲノムアセンブル

Read coverageがcontig, scaffold間で異なっていても良い

IDBA-UD, MEGAHIT, MetaVelvet, metaSPAdes, etc.

- メタゲノム遺伝子予測

コドン使用頻度がcontig, scaffold間で異なっていても良い

MetaGeneMark, MetaGeneAnnotator, Prodigal, etc.

ゲノムアセンブルの二大戦略

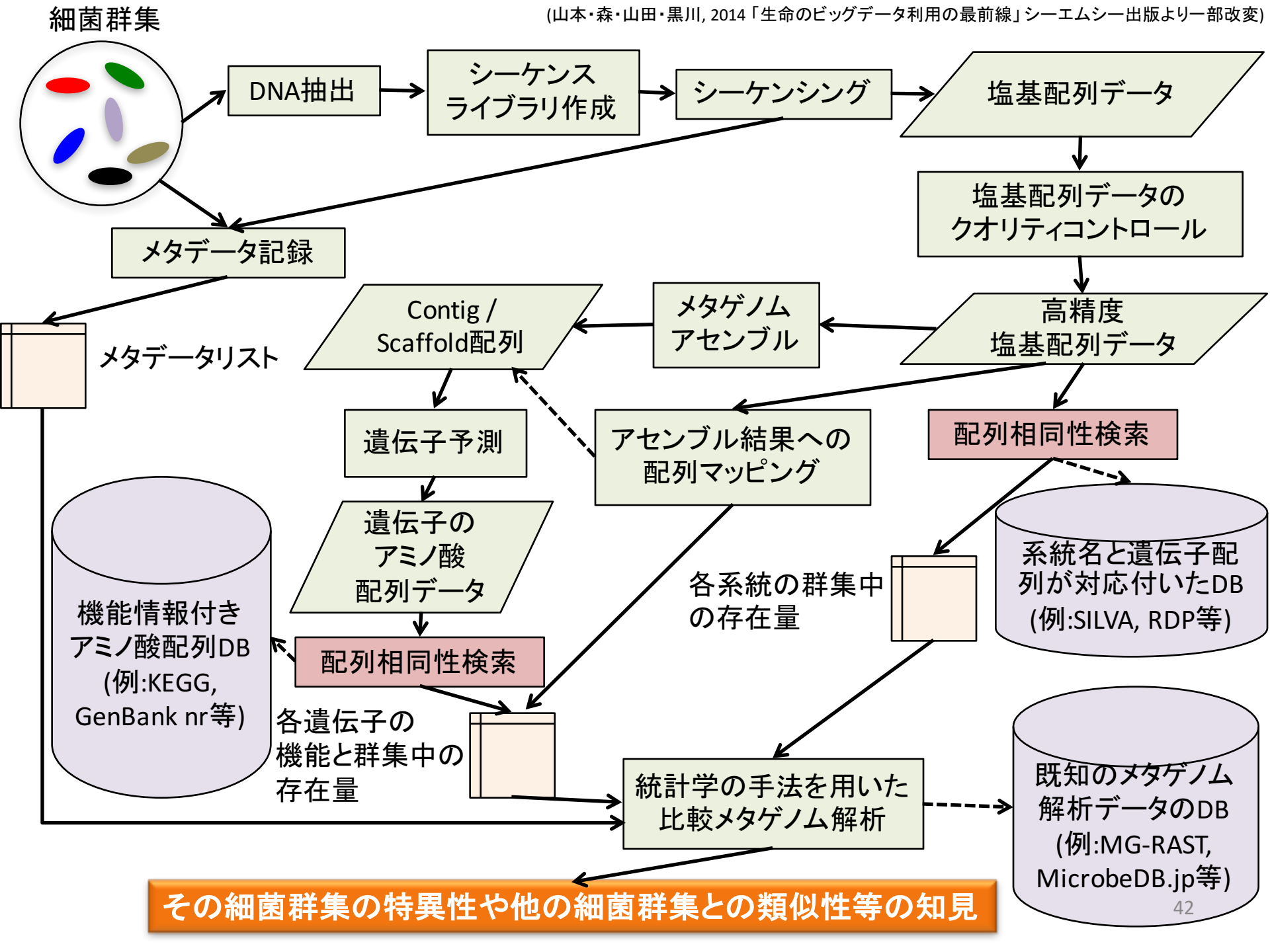
- Overlap-Layout-Consensus
  - k-merの共有やローカルアラインメント等でリード間のoverlapを見つけて、短いContigを作成し、さらにContig間をoverlapをもとに結合（layout）。リードのoverlapを領域ごとに集めてマルチプルアラインメント等をしてconsensusをとることでアセンブルする
  - 例: Celera Assembler, Newbler, Mira, Canu
- de Bruijn Graph
  - リードをoverlapありのk-merに分割して、多数のリード間のk-merの共有をde Bruijn graph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


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
系統名と遺伝子配
列が対応付いた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

Taxonomic assignment strategy?

	Coverage of ref. sequences	Single copy in genomes?	Can analyze eukaryotes and virus?	Robust against HGT?	Example of tools
16S rRNA genes	◯	×	×	◯	VITCOMIC2, MAPseq
Single copy genes	△	◯	×	◯	MAPLE, mOTU2
Unique marker genes	△	◯	×	◯?	MetaPhlAn2
Read mapping	△	×	◯	×	BWA-MEM, Centrifuge
k-mer	△	×	◯	×	Kraken, Mash

パスウェイデータベース

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

https://www.genome.jp/kegg/ko.html

KO (KEGG ORTHOLOGY) Database

Linking genomes to pathways by ortholog annotation

Menu PATHWAY BRITE MODULE KO Annotation ENZYME RModule BlastKOALA

Search KO for Go

KO Database of Molecular Functions

The **KO (KEGG Orthology)** database is a database of molecular functions represented in terms of functional orthologs. A functional ortholog is manually defined in the context of KEGG molecular networks, namely, KEGG pathway maps, BRITE hierarchies and KEGG modules. For example, when a pathway map is drawn, each box is given a KO identifier (called K number) and experimentally characterized genes and proteins in specific organisms are used to find orthologs in other organisms. The granularity of "function" is context-dependent, and the resulting KO grouping may correspond to a highly similar sequence group and a limited organism group or it may be a more divergent group.

The KO system is a network-based classification of KOs shown below:

KEGG Orthology (KO)

ただし、KEGGのアミノ酸配列データとKO IDとの対応関係を手軽に取得するためのKEGG FTPサイトへのアクセスは有料

GhostKOALA

Query Data Input

KEGG Automatic annotation and KEGG mapping service

https://www.kegg.jp/ghostkoala/

BlastKOALA | GhostKOALA | KofamKOALA

KOALA job status 2019/09/16 22:53:56 (GMT+9)

	Blast	Ghost	Kofam
Number of jobs in the queue	10	5	0
Submission of last completed job	2019/09/16 20:44:15	2019/09/16 21:10:01	2019/09/16 22:45:48

KOALA (KEGG Orthology And Links Annotation) is KEGG's internal annotation tool for K number assignment of KEGG GENES using SSEARCH computation. BlastKOALA and GhostKOALA assign K numbers to the user's sequence data by BLAST and GHOSTX searches, respectively, against a nonredundant set of KEGG GENES. KofamKOALA is a new member of the KOALA family available at GenomeNet using the HMM profile search, rather than the sequence similarity search, for K number assignment. See Step-by-step Instructions.

Reference: Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726-731. [pubmed] [pdf]

GhostKOALA accepts a larger dataset and is suitable for annotating metagenomes

Upload query amino acid sequences in FASTA format

Enter FASTA sequences

Or upload file: ファイルを選択 選択されていません

Your query data consisting of multiple amino acid sequences will be given K numbers by GhostKOALA. The file size of up to 300 MB (one million sequences with average length of 300 or three million sequences with average length of 100) may be uploaded.

Enter KEGG GENES database file to be searched

- ○ genus_prokaryotes
- ◉ genus_prokaryotes + family_eukaryotes
- ○ genus_prokaryotes + family_eukaryotes + viruses

The database files for GhostKOALA are somewhat different from those for BlastKOALA. For each group of KEGG organisms at the genus or family level, a nonredundant dataset is generated by taking all protein-coding genes from the representative genome and additional genes from the other genomes with two criteria. One is the same as in BlastKOALA, different K numbers, and the other is unique to GhostKOALA, different CD-HIT clusters, which are computed with 50% identity cutoff. In addition, the database file for viruses is created by CD-HIT wth 90% identitity cutoff from the viruses category of KEGG GENES. These additions are meant for analyzing taxonomic compositions of metagenomes.

Enter your email address

An email will be sent to you for confirmation of your input data. You will have to click on the link in the email to initiate your job. When the job is finished, you will receive another email for browsing the result and performing KEGG Mapper analysis. You cannot request another job until the current one is finished or canceled. *Notice: Your email address will not be used for any other purpose.*

Request for email confirmation | Clear

Genomaple (formerly MAPLE) -2.3.2
Genome Metabolic And Physiological potentiaL Evaluator

for gene mapping to the KEGG functional modules and calculation of module completion ratio (MCR)

trademark application pending

https://maple.jamstec.go.jp/maple/maple-2.3.1/

About MAPLE

MAPLE (**M**etabolic **A**nd **P**hysiological potentia**L** **E**valuator) is an automatic system for mapping genes in an individual genome and metagenome to the functional module and for calculating the module completion ratio (MCR) in each functional module defined by Kyoto Encyclopedia of Genes and Genomes (KEGG). The MCR calculation is performed based on a Boolean algebra-like equation defined by KEGG to each module. MAPLE first assigns a KO identifier (ID) to the query gene using KAAS, maps the KO-assigned genes to the KEGG functional modules, and calculates the MCR of each functional module and its abundance when the module is complete. There are two methods for KO assignment by KAAS: bidirectional best hit (BBH) and single-directional best hit (SBH). The BBH method is suitable for complete gene sets identified in complete genomes or contigs, while the SBH method is mainly for short-read sequences in metagenomes or incomplete genomes.

The result page displays the MCR, abundance of each KEGG module and the taxonomic information of the KO-assigned genes mapped to the module along with a mapping pattern. Also, a module list sharing the same KOs is shown. The results of KO assignment by KAAS, taxonomic information of the genes mapped to the KEGG modules, and calculated MCRs are downloadable in an Excel format. MAPLE can display the results of comparative analyses of mapping patterns, MCR results, and abundance of complete modules between different metagenomic samples. Generally, it is expected that the MCR is linked to the likelihood that the organisms perform the physiological function corresponding to the module. However, when the KOs used for a module are shared with the other modules, the MCR does not necessarily reflect the working probability of each functional module. To evaluate the working probability of the physiological function in the incomplete modules, we proposed the *Q*-value for determining the significance of module completeness. The *Q*-value, which implies the probability that a reaction module is identified by chance, is calculated based on the statistics of the sequence similarity score and KO abundance using the concept of multiple testing corrections according to the Boolean algebra-like equations.

- MAPLE Help

Metagenome or Partial Genome Sequences

KO assignment to short-read sequences (400–500 nt) produced by a high-throughput DNA sequencer is performed by KAAS using the single-directional best-hit (SBH) method. Query sequences must be translated into amino acid (aa) sequences before submission, and sequences longer than 100 aa are recommended for accurate KO assignment.

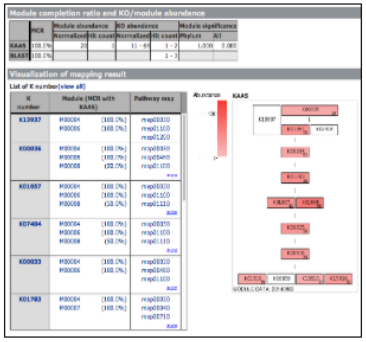
- MAPLE job request (SBH method)
- MAPLE job request (SBH method, KAAS result uploading)

Complete or Draft Genome Sequences

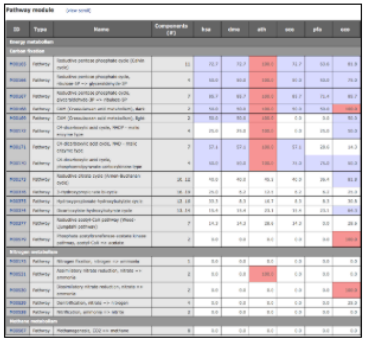
KO assignment for complete gene sets that were identified in the complete genome or contigs is performed by KAAS using the bidirectional best-hit (BBH) method.

Example of Results

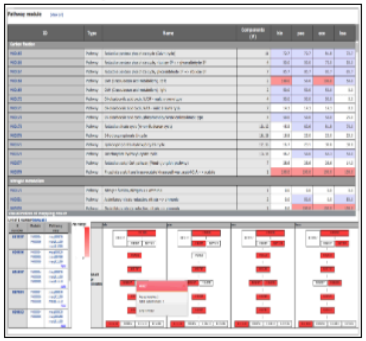
Mapping genes to the KEGG functional modules



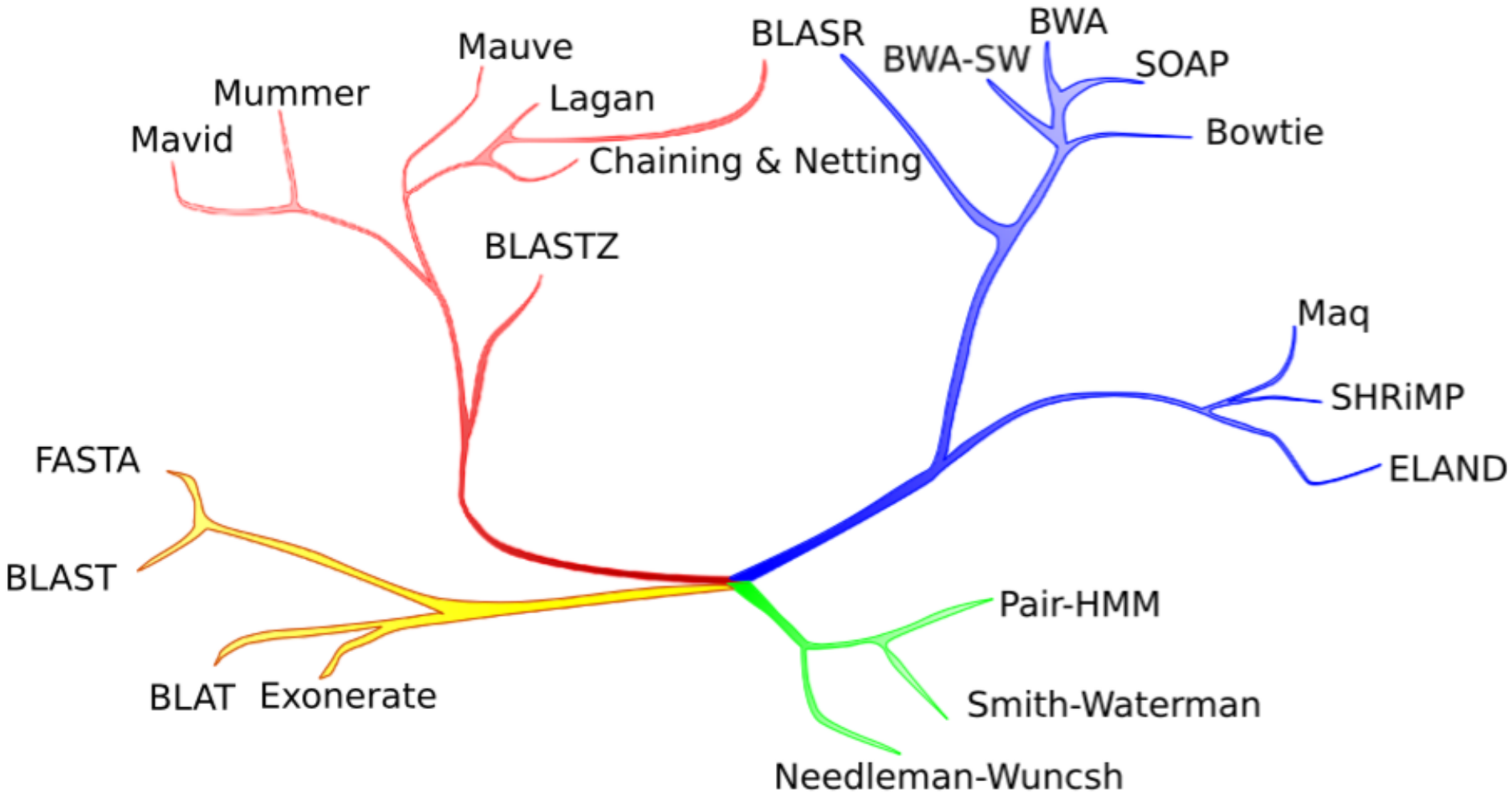
MCR calculation of each KEGG module



Results comparison



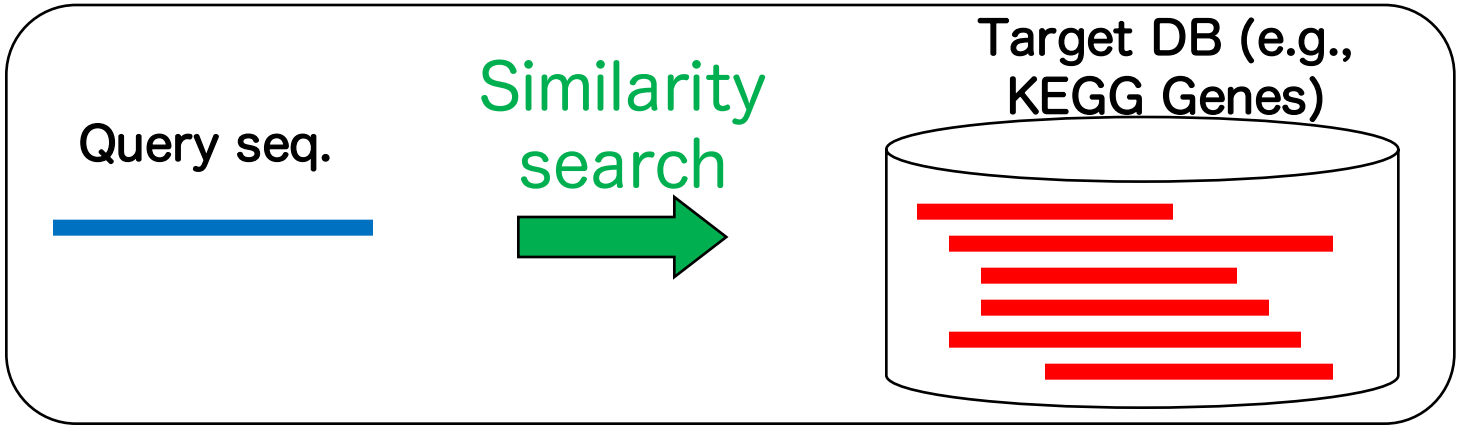
塩基配列の類似性検索ツールにも様々なものが存在する



(Chaisson & Tesler 2012, BMC Bioinformatics)

Referenceと近ければMapper系 (Bowtie 2, BWA-MEM等)
Referenceと遠いのなら、BLAST系

Sequence similarity search



Amplicon seq. query
- Seq. num. ~= 10,000 – 100,000
- Each seq. length ~= 150 – 500 base
- Total ~= 1 – 500MB

Metagenome seq. query
- Seq. num. ~= 1,000,000 – 40,000,000
- Each seq. length ~= 150 – 300 base
- Total ~= 150MB – 10GB

Target DB (for amplicon)
- Seq. num. ~= 300,000
- Each seq. len. ~= 1,500 base
- Total ~= 30MB

All sequences have some similarity

Target DB (for shotgun)
- Seq. num. ~= 17,384
- Each seq. len ~= 0.5M – 10M base
- Total ~= 35GB (AA: 10GB)

Small amount of sequences have similarity

ある程度遠縁なアミノ酸配列も探せる配列類似性検索ツール

Fast and sensitive protein alignment using DIAMOND

Benjamin Buchfink[1], Chao Xie[2,3] &
Daniel H Huson[1,2]

Nature Methods. 2015

MMseqs2: sensitive protein sequence searching for analysis of massive data sets

Martin Steinegger[1,2] & Johannes Söding[1]

[1]Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany; [2]Department for Bioinformatics and Computational Biology, Technische Universität München, 85748 Garching, Germany

e-mail: johannes.soeding@mpibpc.mpg.de; martin.steinegger@mpibpc.mpg.de

Nature Biotech. 2017

Faster sequence homology searches by clustering subsequences

Shuji Suzuki[1,2], Masanori Kakuta[1], Takashi Ishida[1] and Yutaka Akiyama[1,2,*]

Bioinformatics. 2014

MMseqs2: sensitive protein sequence searching for analysis of massive data sets

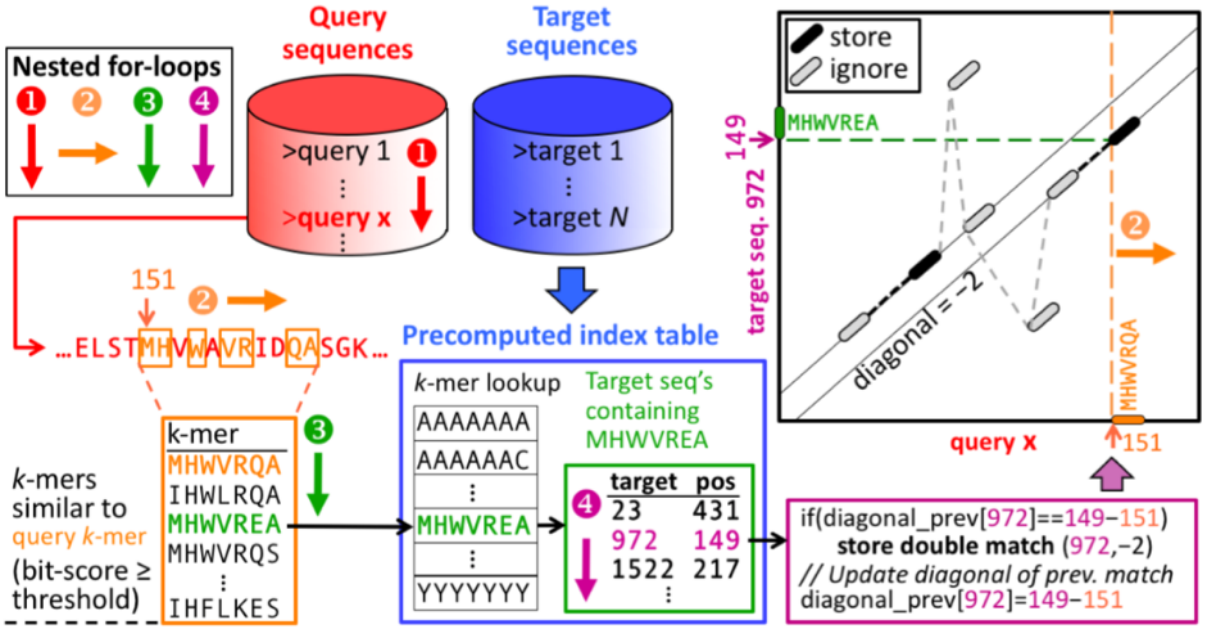
Martin Steinegger[1,2] **& Johannes Söding**[1]

[1]Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany; [2]Department for Bioinformatics and Computational Biology, Technische Universität München, 85748 Garching, Germany

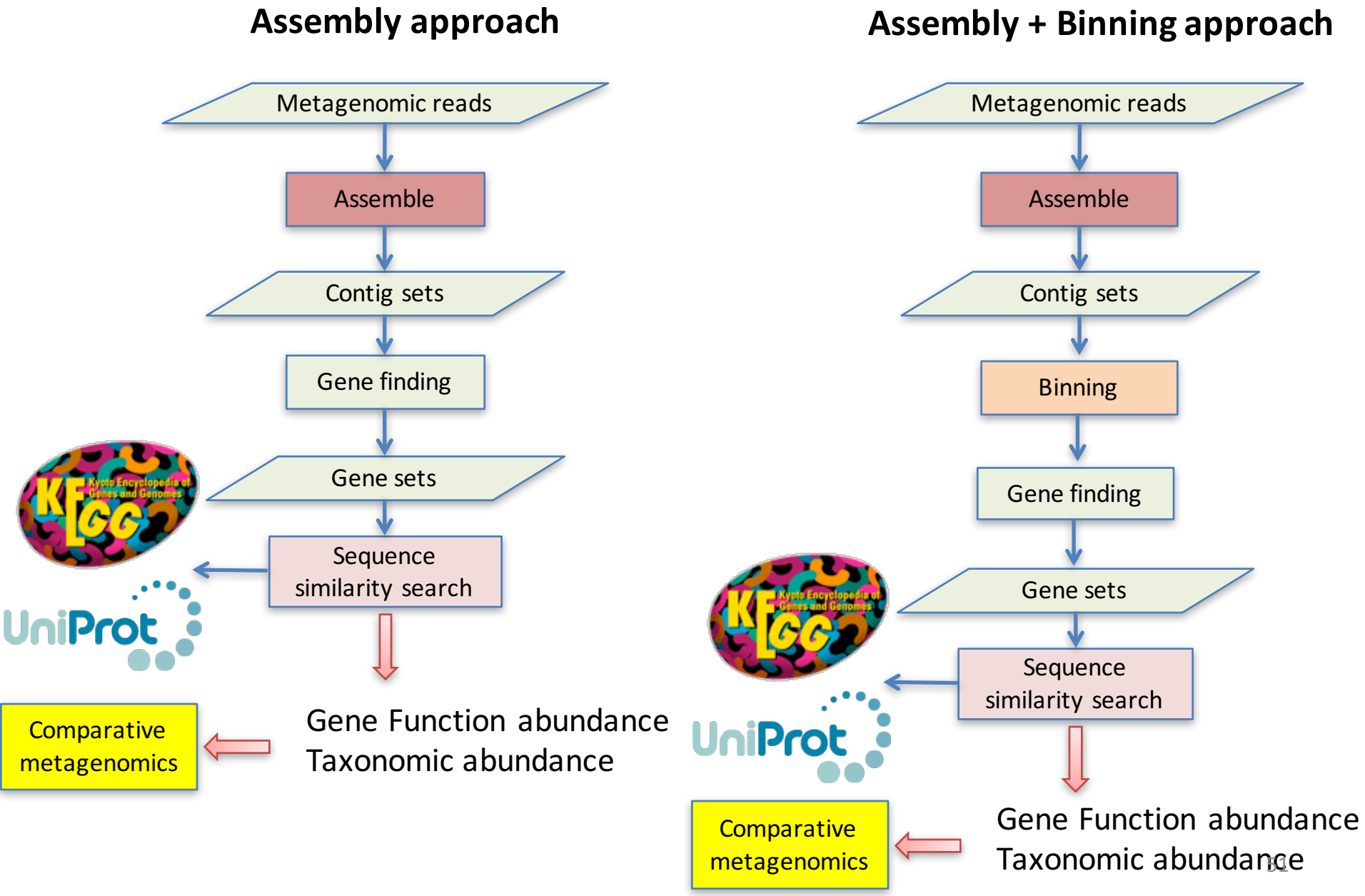
e-mail: johannes.soeding@mpibpc.mpg.de; martin.steinegger@mpibpc.mpg.de

Nature Biotech. 2017

- BLASTPの約4800倍速い
- seed探索、seed伸長、Gapありアラインメントの3ステップ
- seed探索は完全マッチではなく、類似k-merサーチ x 2
- seed長は7 AA (BLASTPは6 AA)で、spaced seedなので近隣に2つseedが見つかる必要あり
- 類似k-merサーチを用いているので、精度はそこそこ高い
- メモリを大量に使う
  （KEGGで60GB, nrで数百GB)



Bioinformatics analysis strategy?



k-mer組成やcoverageを用いてメタゲノムContigを分ける(binning)ツールの例

CONCOCT	Genome binner using differential coverage, tetranucleotide frequencies, paired-end linkage	Near complete (>95%) assignment of datasets at some cost for average genome purity and completeness.
MaxBin 2.0	Genome binner using multi-sample coverage, tentranucleotide frequencies	Largest average purity and completeness across entire abundance range. Recovery of 2nd most genomes with high purity and completeness.
MetaBAT	Genome binner using multi-sample coverage, tetranucleotide frequencies, paired-end linkage	Assignment of a large portion (>88%) of datasets at some costs for average genome purity and completeness.
MetaWatt-3.5	Genome binner using tetranucleotide frequencies	Recovery of the most genomes with high purity and completeness; near complete assignment of datasets at some cost for average genome purity and completeness.
MyCC	Genome binner using short k-mer frequencies, multi-sample coverage, and 40 universal phylogenetic marker genes	Near complete assignment of datasets at some cost for average genome purity and completeness.

Sczyrba et al. 2017, Nature Methods

メタゲノムアセンブル -> Binning -> ゲノムBinごとに系統・機能アノテーション

メタゲノムデータの全体像を議論するのではなく、優占系統の
ドラフトゲノム配列を抽出して各ゲノムが持つ機能について議論する

2017

nature microbiology

ARTICLES

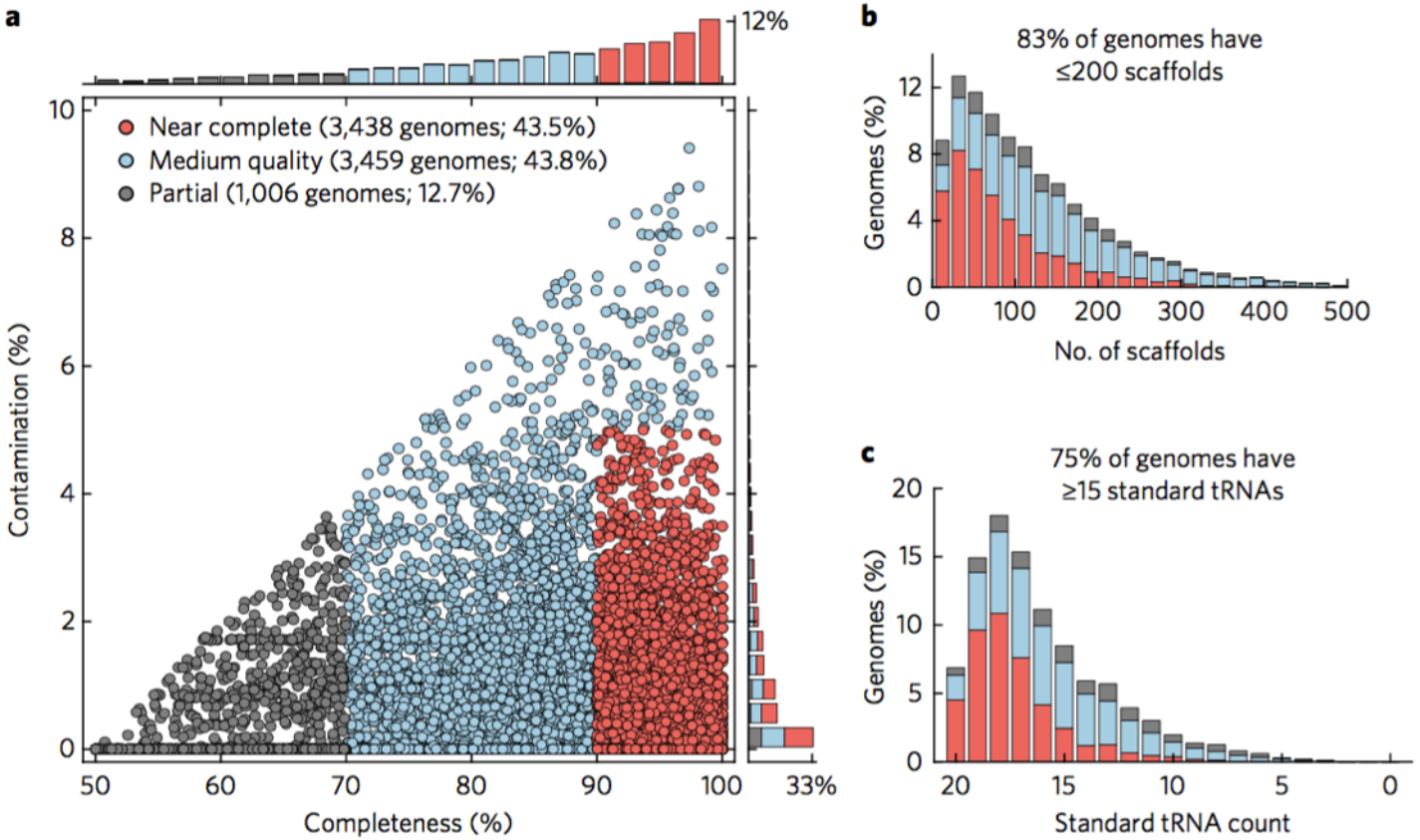
DOI: 10.1038/s41564-017-0012-7

OPEN

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

Metagenome Assembled Genomes (MAGs)



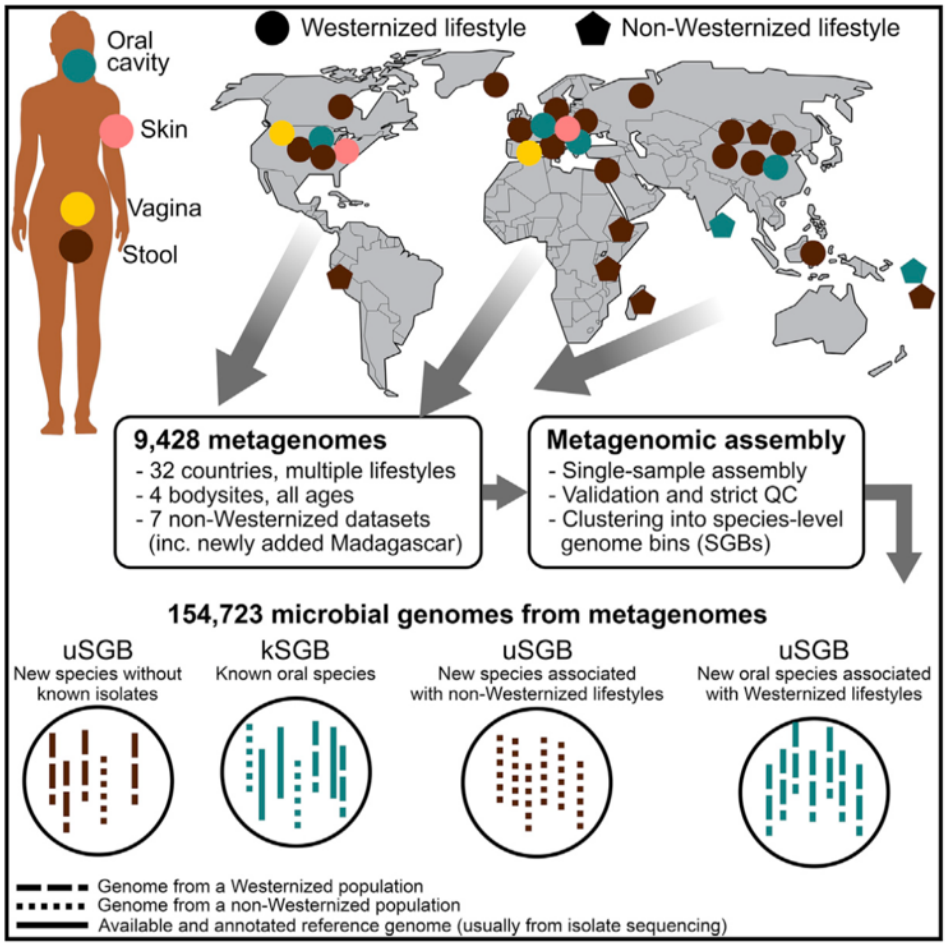
Cell

2019

Resource

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Graphical Abstract



Authors

Edoardo Pasolli, Francesco Asnicar, Serena Manara, ..., Christopher Quince, Curtis Huttenhower, Nicola Segata

Correspondence

nicola.segata@unitn.it

In Brief

The human microbiome harbors many unidentified species. By large-scale metagenomic assembly of samples from diverse populations, we uncovered >150,000 microbial genomes that are recapitulated in 4,930 species. Many species (77%) were never described before, increase the mappability of metagenomes, and expand our understanding of global body-wide human microbiomes.

2019

nature biotechnology

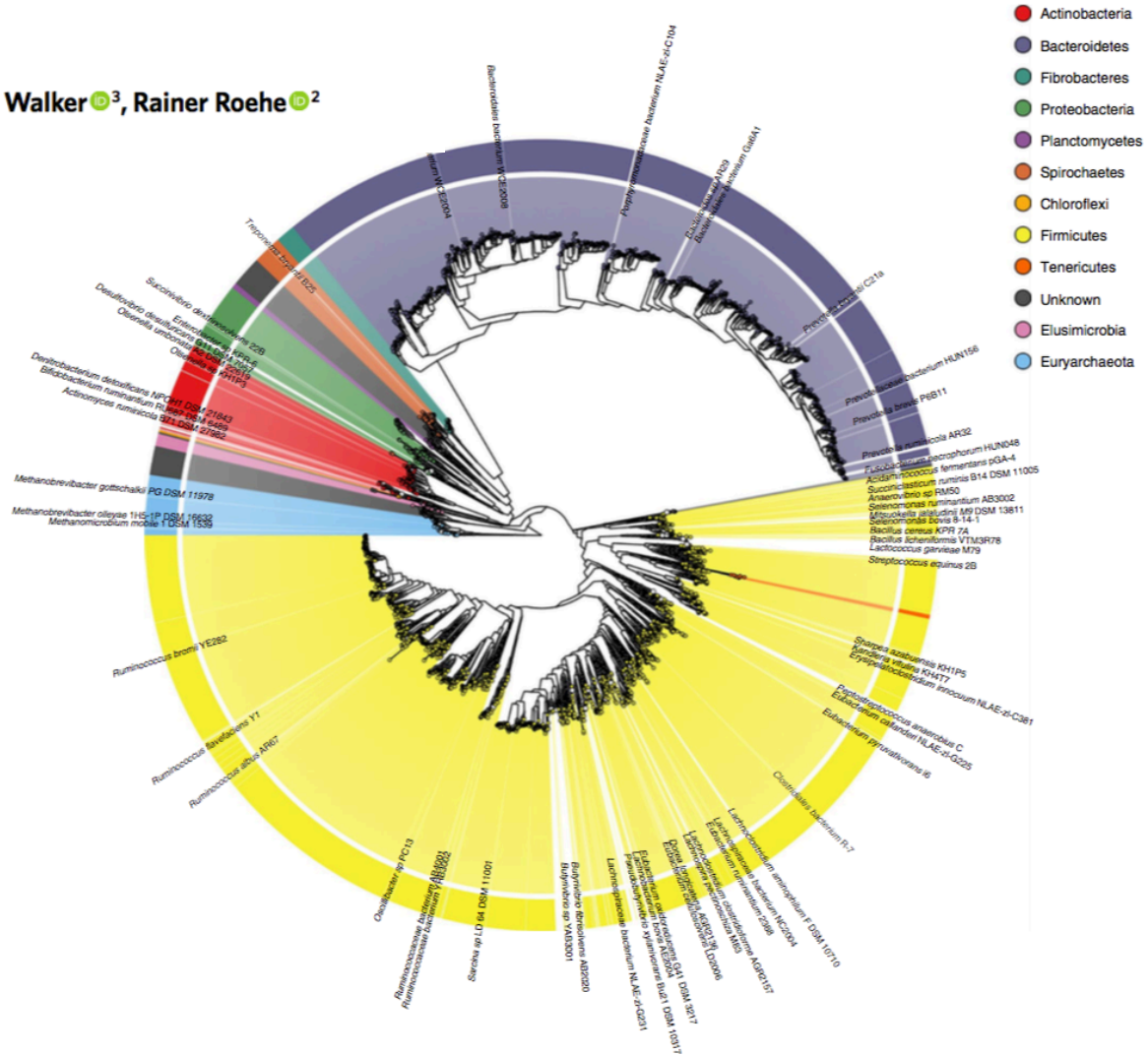
RESOURCE

https://doi.org/10.1038/s41587-019-0202-3

OPEN

Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery

Robert D. Stewart[1], Marc D. Auffret[2], Amanda Warr[1], Alan W. Walker[3], Rainer Roehe[2] and Mick Watson[1*]



Bioinformatics analysis strategy?

	Full length genes	Gene-neighbor	Draft genome	Minor taxa	HGT	Machine power	Sensitivity against ref. seq.
Assembly	○	○	△	×	○	high	high
Assembly + Binning	○	○	○	×	△	high	high
Read mapping	×	×	×	○	×	low	low
Read-based CDS	△	×	×	○	×	middle	high

Hypothesis testing or Data-driven?

Hypothesis testing

e.g., *Bacteroides*' glycosyl hydrolase gene family composition should be different between groups.

Conduct a statistical hypothetical test with the focused taxa/genes.

Data-driven

e.g., Something should be different between groups.
Groups should be exist in the samples.

Find patterns in something (taxonomic abundances, gene contents, SNVs, etc.) using multivariate analysis (PCA, k-means clustering, etc.).
Metadata description is important to understand results.

Amplicon or Metagenome?

	Cost	DNA input	Machine power	PCR bias	Strain level	Function	Common protocol
Amplicon	low	low	low	Y	N	N	Y
Metagenome	high	high	high	N	Y	Y	N

If you want gene function information,
and have enough money, machine power, bioinformatics skills;
you should conduct shotgun metagenome analyses.
Non-target taxonomic composition analysis can be performed by using metagenome.

Short or Long reads?

	Cost	DNA input	InDel error	Construct good ref.	gene-neighbor
Short reads	low	low	low	○	△
Long reads	high	high	high	◎	○

If you want to construct good reference sequences from metagenome data,
I will recommend you to conduct deep long read and shallow short read shotgun metagenome sequencing.

Suzuki *et al. Microbiome* (2019) 7:119
https://doi.org/10.1186/s40168-019-0737-z

Microbiome

RESEARCH Open Access

Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut

Yoshihiko Suzuki[1†], Suguru Nishijima[1,2,3†], Yoshikazu Furuta[4], Jun Yoshimura[1], Wataru Suda[1,5], Kenshiro Oshima[1], Masahira Hattori[1,3,5*] and Shinichi Morishita[1*]

Long readのみで遺伝子・系統の頻度情報を高精度に推定するのは困難なため、現状はShort readとの併用が必須

Comparison between projects?

	Admin	Database URL	Sequence data	Separate amplicon and shotgun?	Taxa	Function	Number of samples in Sep. 2019
NCBI Taxonomy + SRA	NCBI, USA		○	×	△	×	>1,600,000
GOLD	JGI, USA	https://gold.jgi.doe.gov/	×	×	×	×	50,821
IMG/M	JGI, USA	https://img.jgi.doe.gov/cgi-bin/m/main.cgi	○	×	○	○	16,170
MG-RAST	Chicago U. USA	http://www.mg-rast.org/	○	○	○	○	392,514
MGnify	EBI, EU	https://www.ebi.ac.uk/metagenomics/	×	○	○	○	176,360
MicrobeDB.jp v.2	NIG, Japan	http://microbedb.jp/MDB/	×	○	○	○	60,551

他の講習会の例

- https://bioinformaticsdotca.github.io/metagenomics_2018

Analysis of Metagenomic Data 2018
3 Day workshop

Module 1: Introduction to Metagenomics

Module 2: Marker Gene-Based Analysis

Module 3: PICRUSt

Module 4: Metagenomic Taxanomic and Functional Composition

Module 5: Pulling Genomes from Metagenomes

Module 6: Metatranscriptomics

Module 7: Statistical Tests for Metagenomics

Module 8: Biomarkers and Bringing It All Together