AJACS MAG (Metagenome-Assembled Genome) を
知って・学んで・使う」

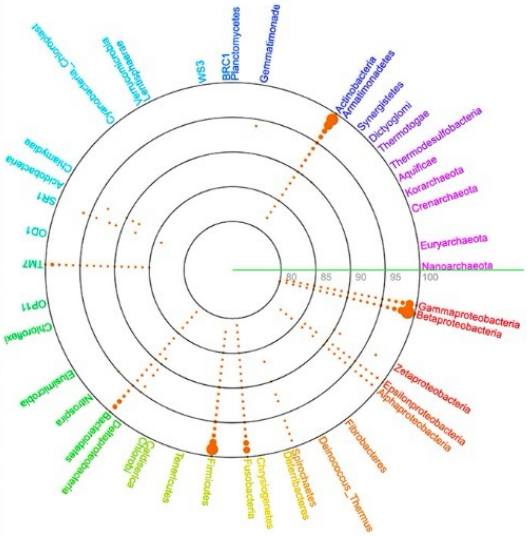
2024年7月25日

メタゲノムとMAG解析について

森　宙史, Ph.D.
(Hiroshi Mori)
国立遺伝学研究所
先端ゲノミクス推進センター
hmori@nig.ac.jp

HOME
Members
Research
Publications
DB&Tools
Resources
Link
Access&Contact

国立遺伝学研究所 ゲノム多様性研究室



本研究室では、バイオインフォマティクス技術を用いて微生物などが持つゲノムの多様性を解明する研究に取り組んでいます。メタゲノム解析技術の進展によって、培養が難しい微生物も含めてゲノム解読が可能になり、また、メタゲノムデータの一種であるAncient DNAデータを用いることで、数万年以上前に絶滅した生物のゲノム解析も可能になりました。我々は森が兼任する遺伝研の先端ゲノミクス推進センターと強固に連携し、最先端のゲノム解析技術とバイオインフォマティクス解析技術を武器に未だ未知な部分が多い生物のゲノムの多様性に関する幅広い研究を進めております。

- 微生物のゲノム解析
- 様々な環境のメタゲノム解析
- Ancient DNA解析

- これらに関わる情報解析ツール・DBの開発

研究室webページ https://www.genome.id

微生物について

微生物は小さい

- 単細胞真核生物 (Fungi, protozoa)　4–40 µm
- Prokaryotes　0.1–10 µm
- Viruses (微生物に含めない場合も多い)　0.03–0.3 µm

この小ささが、phenotype解析やsingle cell解析等
様々な解析を困難にしている

微生物を形で判別するのは困難

微生物の実験室での研究は、まず培養して増やすのが基本

培地の例

- LB培地 (Luria-Bertani培地)
- 化学的に定義された培地

https://www.jcm.riken.jp/cgi-bin/jcm/jcm_grmd?GRMD=58

58 INORGANIC SALTS-STARCH AGAR (ISP-4)

Soluble starch	10.0	g
K_2HPO_4	1.0	g
$MgSO_4 \cdot 7H_2O$	1.0	g
NaCl	1.0	g
$(NH_4)_2SO_4$	2.0	g
$CaCO_3$	2.0	g
Trace salts solution (see below)	1.0	ml
Agar	20.0	g
Distilled water	1.0	L

Unadjusted pH will be 7.0-7.4.

Trace salts solution:

$FeSO_4 \cdot 7H_2O$	0.1	g
$MnCl_2 \cdot 4H_2O$	0.1	g
$ZnSO_4 \cdot 7H_2O$	0.1	g
Distilled water	100.0	ml

Comment: The premixed powder is available from Becton Dickinson & Co. as ISP Medium 4.

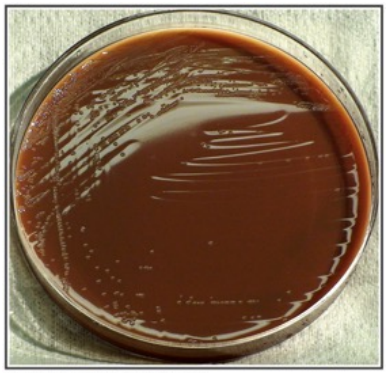
https://togomedium.org/

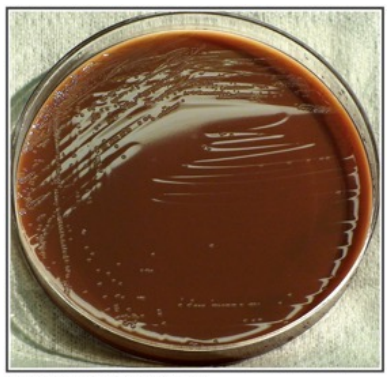
TOGO MEDIUM

Culture media database aggregated from various resources.

TogoMedium is a comprehensive knowledge base focused on culture media for microorganisms. The media available in TogoMedium have been compiled from information provided by diverse bioresource centers and research papers. All information in TogoMedium is described as RDF and the composition of these media is described with Growth Medium Ontology. This enables users to investigate the interconnectedness between organisms, media, and their ingredients, facilitating a deeper understanding of their relationships.

More about us >





数%ぐらいの菌しか
培養できない

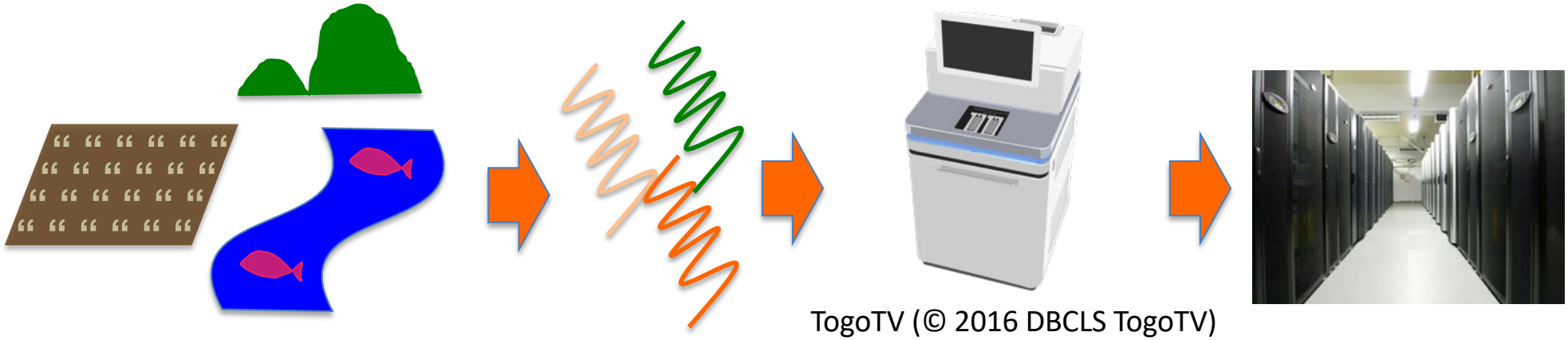
理由:
単一の環境由来でも
多数のニッチが存在し、
単一培地ではそれらを
再現できない

微生物群集、特に細菌群集をどのように解析するか？

メタゲノム解析概論

Metagenomics (since 1998)

Genome analysis against “Microbial community” to know member compositions and functions



TogoTV (© 2016 DBCLS TogoTV)

Microbiota sampling

DNA extraction from microbes

DNA sequencing

Bioinformatics analysis

What's metagenomics?

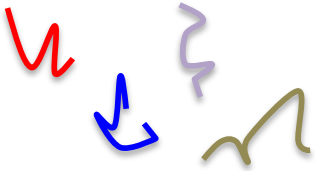
- **Microflora, microbiota, microbial community**: 微生物群集
  Total collection of microorganisms within a community

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

- **Microbiome**: マイクロバイオーム
  Micro+biome or Microbio + ome?
  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

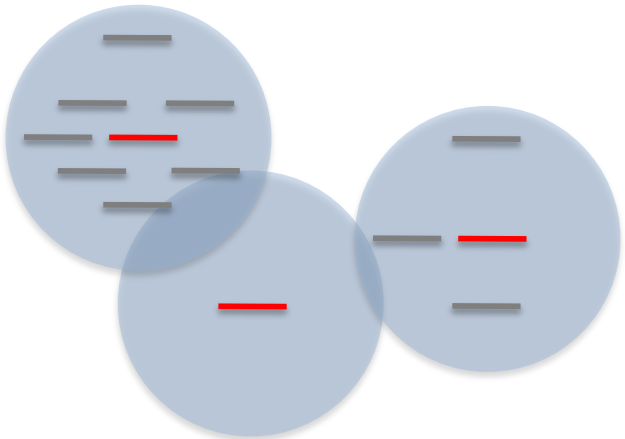


PCR amplification

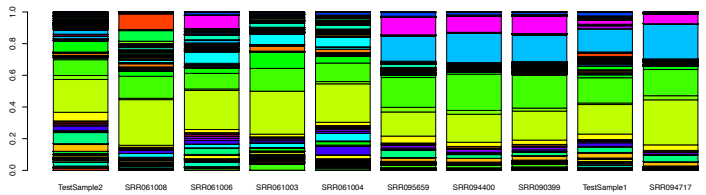
DNA Sequencing

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



Taxonomic assignment and
Comparison between samples

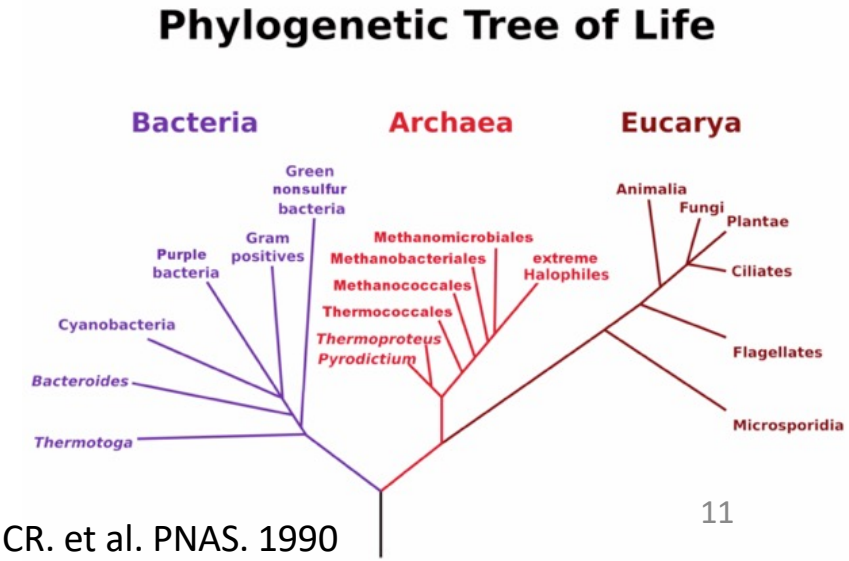


Who's there?

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
- 170万本以上のほぼ完全長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 CR. et al. PNAS. 1990

Metagenomic sequencing analysis
(メタゲノム解析, ショットガンメタゲノム解析)
(メタトランスクリプトーム解析)

DNA extraction / RNA extraction

Reverse transcription

DNA Sequencing

TogoTV (© 2016 DBCLS TogoTV)

Metagenomic reads

Assemble — MEGAHIT, SPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

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

KEGG, RefSeq等

Gene Function abundance
Pathway abundance
Taxonomic abundance

Comparative metagenomics

Sample1 Metadata

Sample2 Metadata

Who's there?

What are they doing?

アンプリコン解析

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（QIIME2・DADA2等）

欠点

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

メタゲノム解析

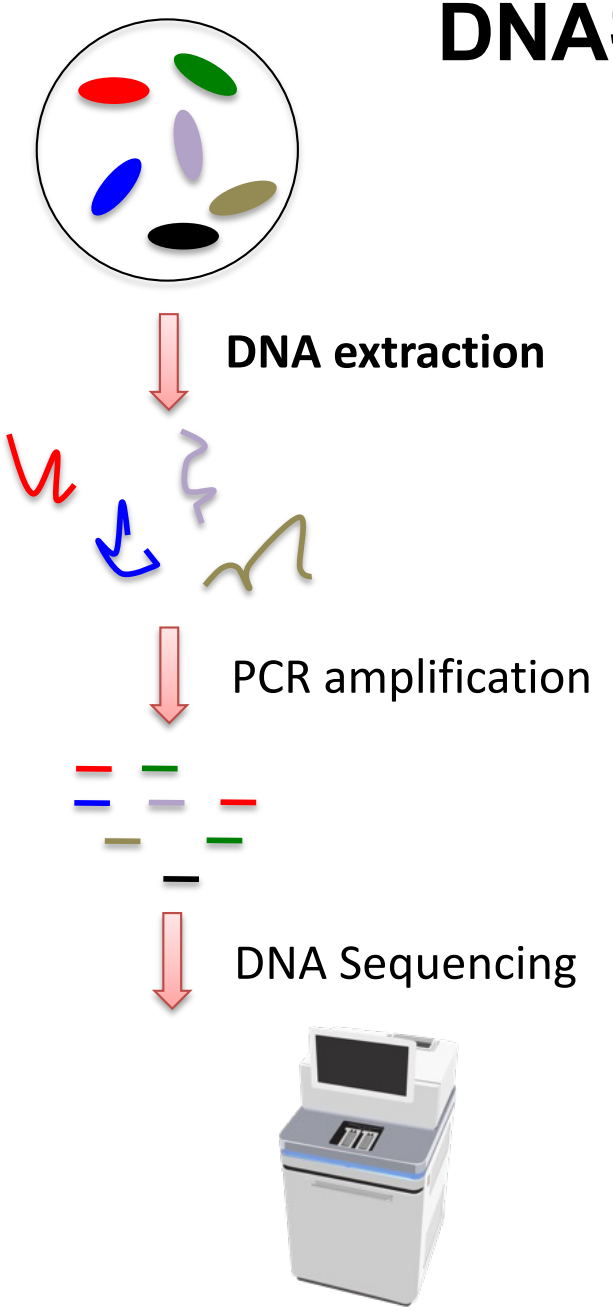
利点

・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）

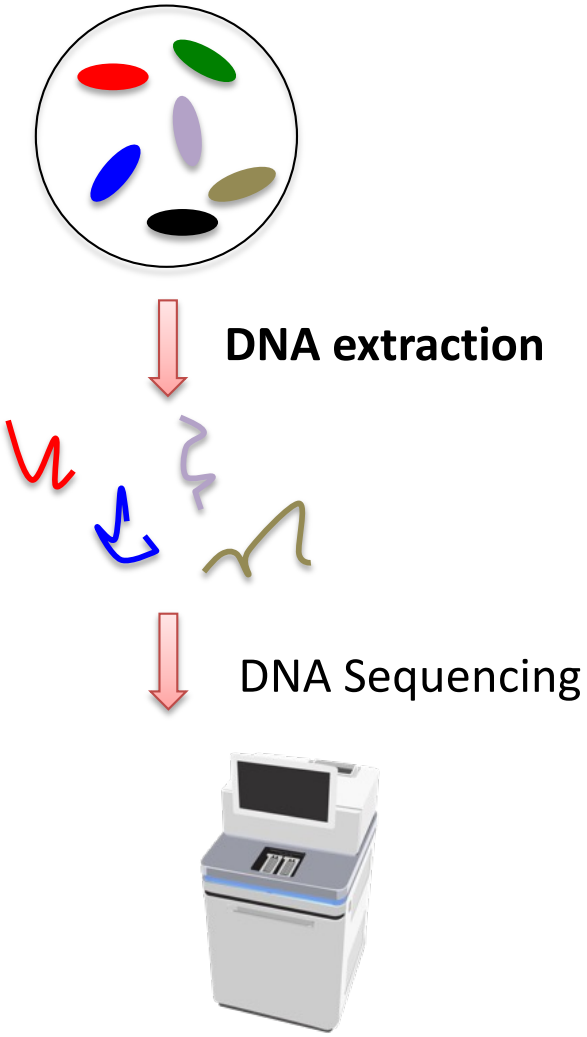
欠点

・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

アンプリコン解析もメタゲノム解析も、DNA抽出法が極めて重要

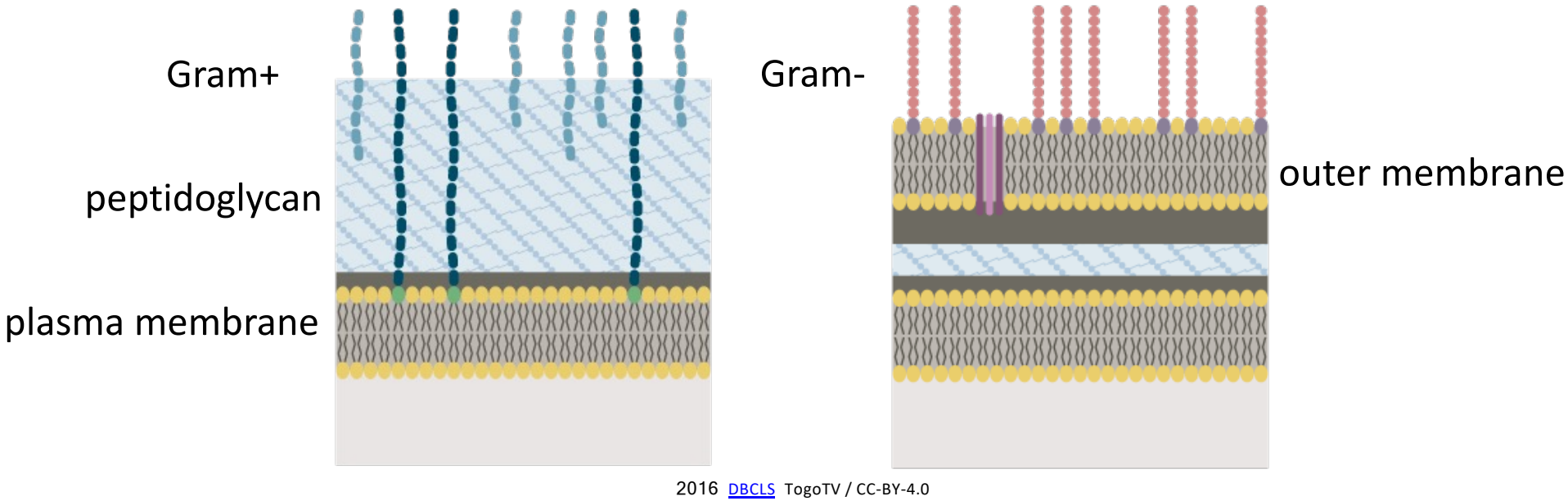


TogoTV (© 2016 DBCLS TogoTV)



TogoTV (© 2016 DBCLS TogoTV)

アンプリコン解析もメタゲノム解析も、DNA抽出法が極めて重要



- **微生物群集解析用の代表的なDNA抽出法**
  - 酵素法（Lysozyme, Proteinase, RNase等の組み合わせ)
  - ビーズ破砕法
  - フェノールクロロホルム法
  - シリカメンブレン法
  - スピンカラム法
- **短鎖型か長鎖型のシーケンサーを使うかによっても変わる**
  - せっかく長く読めるのに、元のDNAがズタズタだと意味が無い
  - DNA量も重要（長鎖型なら数マイクログラムは欲しい）

微生物の系統間でのDNA抽出バイアスの評価

ATCC Mock

20 Strain Even Mix Genomic Material

MSA-1002™

NGS Standards are mock microbial communities that mimic mixed metagenomic samples. This product comprises genomic DNA prepared f fully sequenced, characterized, and authenticated ATCC Genuine Cultures that were selected based on relevant phenotypic and genotypi attributes, such as Gram stain, GC content, genome size, and spore formation. Each order for this product is provided with access to run 20 analyses on One Codex, the leading bioinformatics platform for microbi genomics and metagenomics.

Specification Range: 2.69 ng/µL to 4.48 ng/µL*
*DNA concentrations indicate ATCC manufacturing specifications and ar provided as a reference only

99/100 Bioz Stars 11 Product Citations

Product category Bacteria

Product type Nucleic acid
NGS standard

Components Nucleic acids extracted from:
5% *Acinetobacter baumannii* (ATCC 17978)
5% *Bacillus pacificus* (ATCC 10987)
5% *Phocaeicola vulgatus* (ATCC 8482)
5% *Bifidobacterium adolescentis* (ATCC 15703)
5% *Clostridium beijerinckii* (ATCC 35702)
5% *Cutibacterium acnes* (ATCC 11828)
5% *Deinococcus radiodurans* (ATCC BAA-816)
5% *Enterococcus faecalis* (ATCC 47077)
5% *Escherichia coli* (ATCC 700926)
5% *Helicobacter pylori* (ATCC 700392)
5% *Lactobacillus gasseri* (ATCC 33323)
5% *Neisseria meningitidis* (ATCC BAA-335)
5% *Porphyromonas gingivalis* (ATCC 33277)
5% *Pseudomonas paraeruginosa* (ATCC 9027)
5% *Cereibacter sphaeroides* (ATCC 17029)
5% *Schaalia odontolytica* (ATCC 17982)
5% *Staphylococcus aureus* (ATCC BAA-1556)
5% *Staphylococcus epidermidis* (ATCC 12228)
5% *Streptococcus agalactiae* (ATCC BAA-611)
5% *Streptococcus mutans* (ATCC 700610)

ZymoBIOMICS Gut Mock

Species	Theoretical Abun. (%)
Faecalibacterium prausnitzii	14
Veillonella rogosae	14
Roseburia hominis	14
Bacteroides fragilis	14
Prevotella corporis	6
Bifidobacterium adolescentis	6
Fusobacterium nucleatum	6
Lactobacillus fermentum	6
Clostridioides difficile	1.5
Akkermansia muciniphila	1.5
Methanobrevibacter smithii	0.1
Salmonella enterica	0.01
Enterococcus faecalis	0.001
Clostridium perfringens	0.0001
Escherichia coli (JM109)	2.8
Escherichia coli (B-3008)	2.8
Escherichia coli (B-2207)	2.8
Escherichia coli (B-766)	2.8
Escherichia coli (B-1109)	2.8
Candida albicans	1.5
Saccharomyces cerevisiae	1.4

ZymoBIOMICS Mock

Species	Avg. GC (%)	Gram Stain	gDNA Abun. (%)
Pseudomonas aeruginosa	66.2	-	12
Escherichia coli	56.8	-	12
Salmonella enterica	52.2	-	12
Lactobacillus fermentum	52.8	+	12
Enterococcus faecalis	37.5	+	12
Staphylococcus aureus	32.7	+	12
Listeria monocytogenes	38.0	+	12
Bacillus subtilis	43.8	+	12
Saccharomyces cerevisiae	38.4	Yeast	2
Cryptococcus neoformans	48.2	Yeast	2

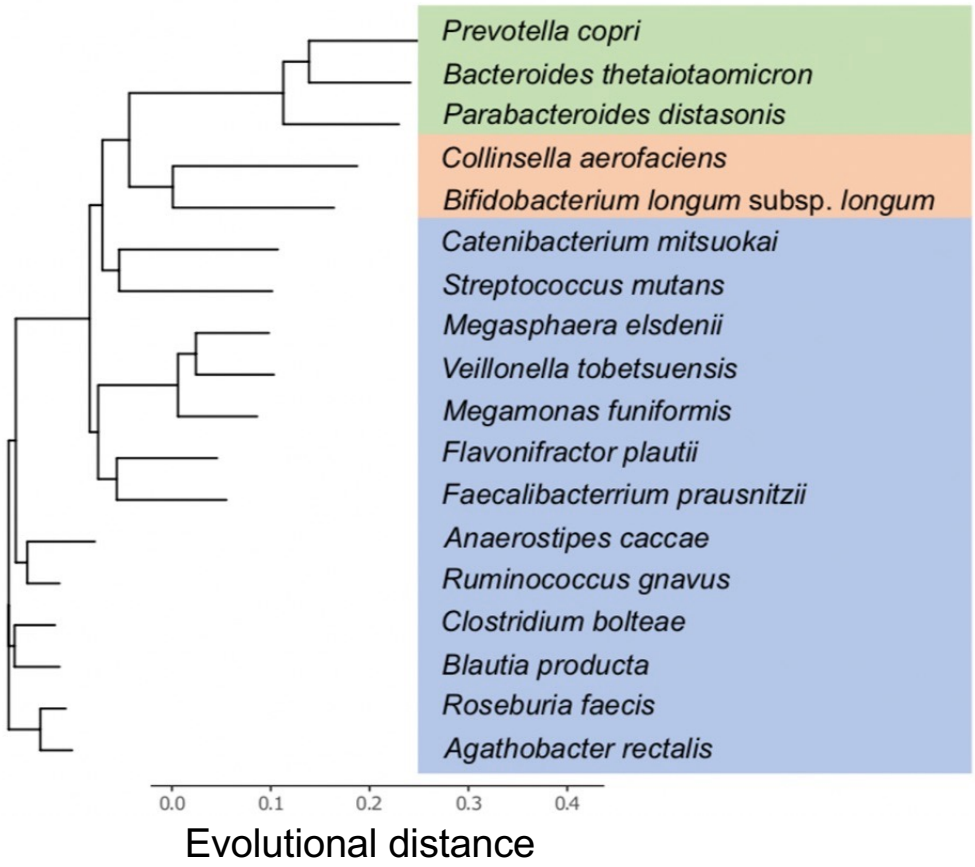
NBRC Mock

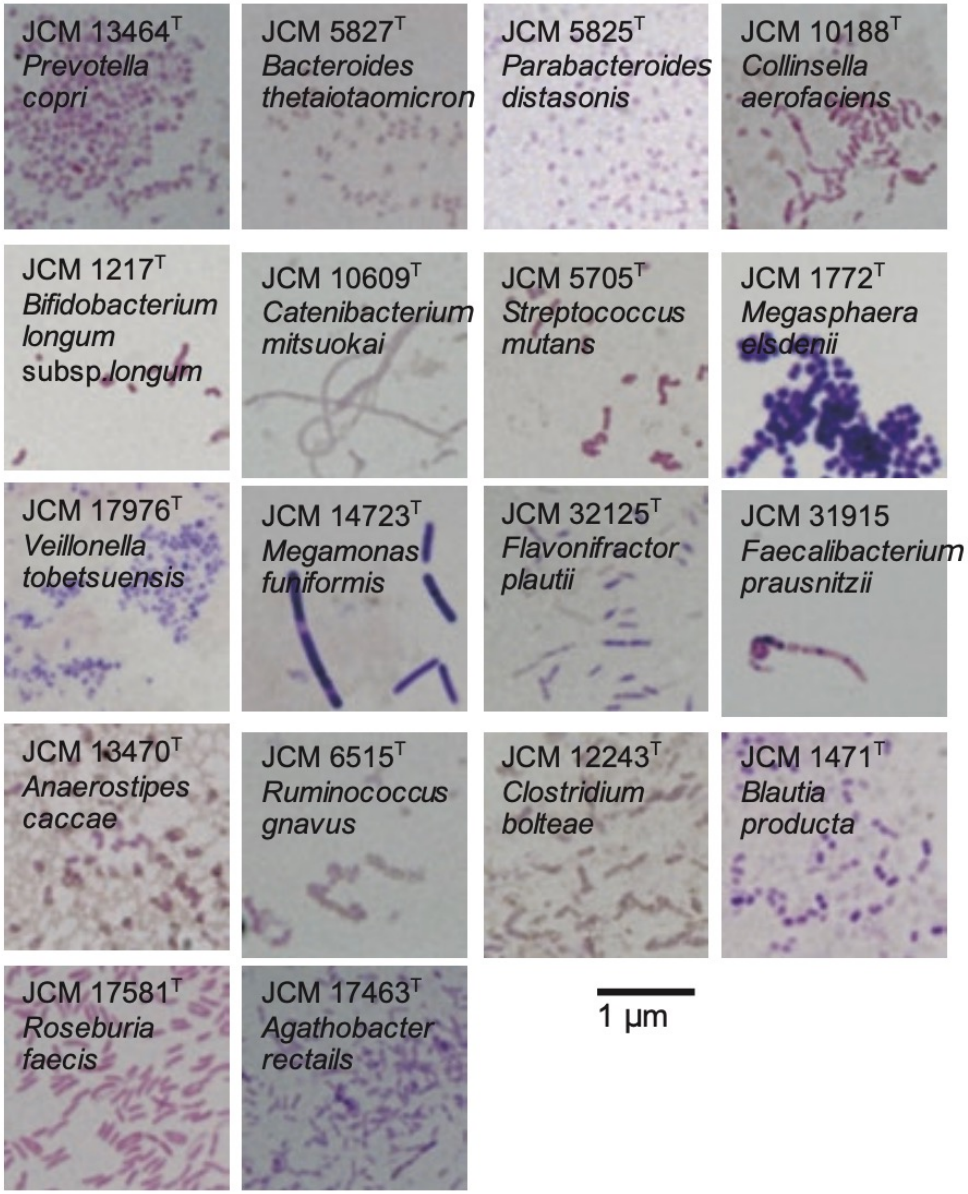
学名	NBRC 番号	理論値 (%)	各株の検出割合 (%) ※3
Bacillus subtilis subsp. *subtilis*	13719^T	6.7	6.8±0.2
Bifidobacterium pseudocatenulatum	113353	6.7	9.3±0.5
Clostridium butyricum	13949^T	6.7	7.7±0.2
Corynebacterium striatum	15291^T	6.7	7.3±0.3
Cutibacterium acnes subsp. *acnes*	107605^T	6.7	14.2±0.4
Lactobacillus delbrueckii subsp. *delbrueckii*	3202^T	6.7	10.3±0.5
Staphylococcus epidermidis	100911^T	6.7	5.4±0.2
Streptococcus mutans	13955^T	6.7	6.7±0.2
Acinetobacter radioresistens	102413^T	6.7	6.5±0.2
Bacteroides uniformis	113350	6.7	3.6±0.1
Enterocloster clostridioformis	113352	6.7	3.6±0.1
Comamonas terrigena	13299^T	6.7	5.8±0.4
Escherichia coli（K-12株）	3301	6.7	3.5±0.1
Parabacteroides distasonis	113806	6.7	5.2±0.3
Pseudomonas putida	14164^T	6.7	4.2±0.1

あくまで細胞混合物であり、実際のメタゲノム用サンプルとは異なる

18 species

- 18 species even cell-mix
- 18 species even DNA-mix

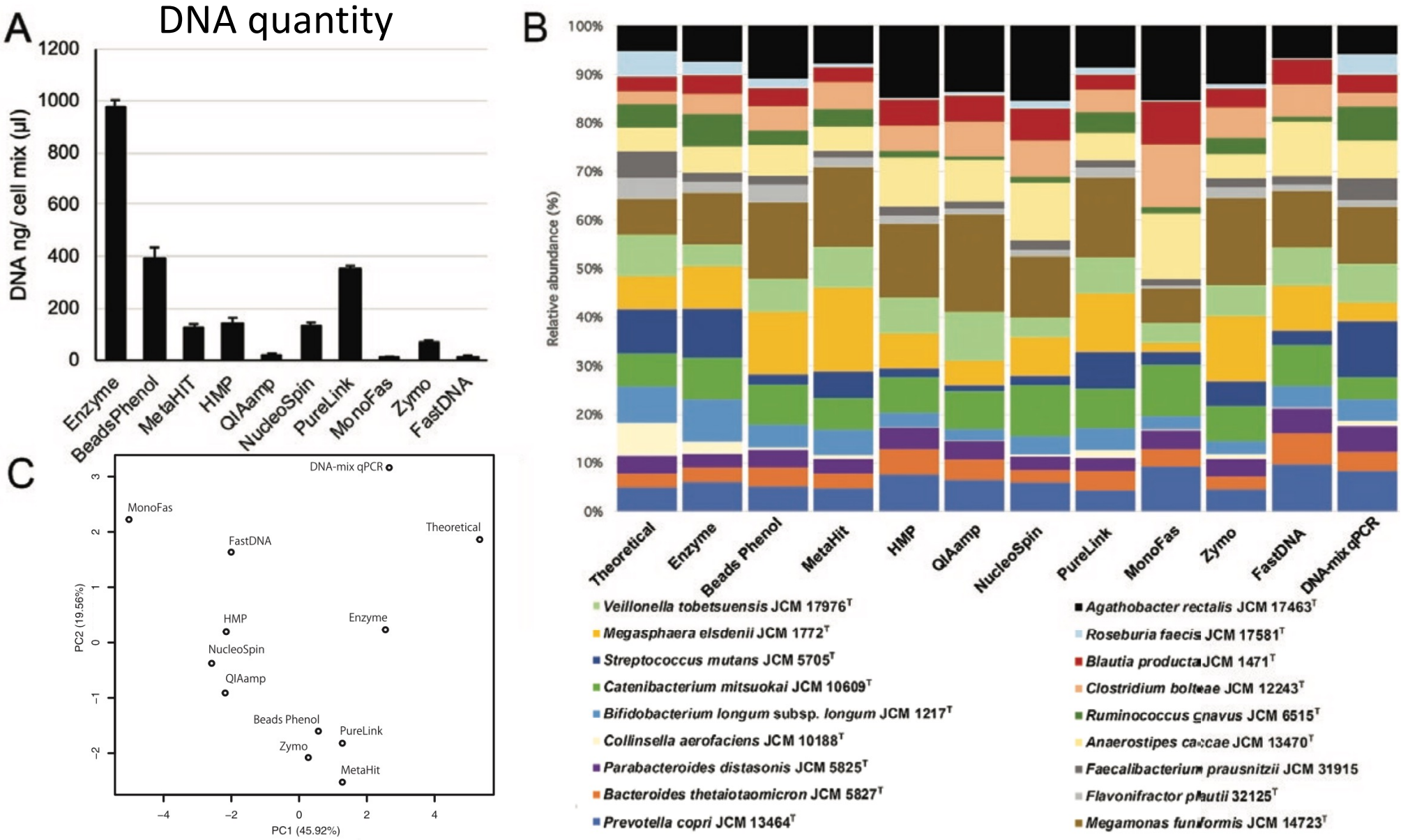




18種の細胞を等量混ぜた Mock communityを作成

Mori H. et al.
DNA Res. 2023

Comparison of 10 different DNA extraction methods using 18 species cell-mix



Mori H. et al.
DNA Res. 2023

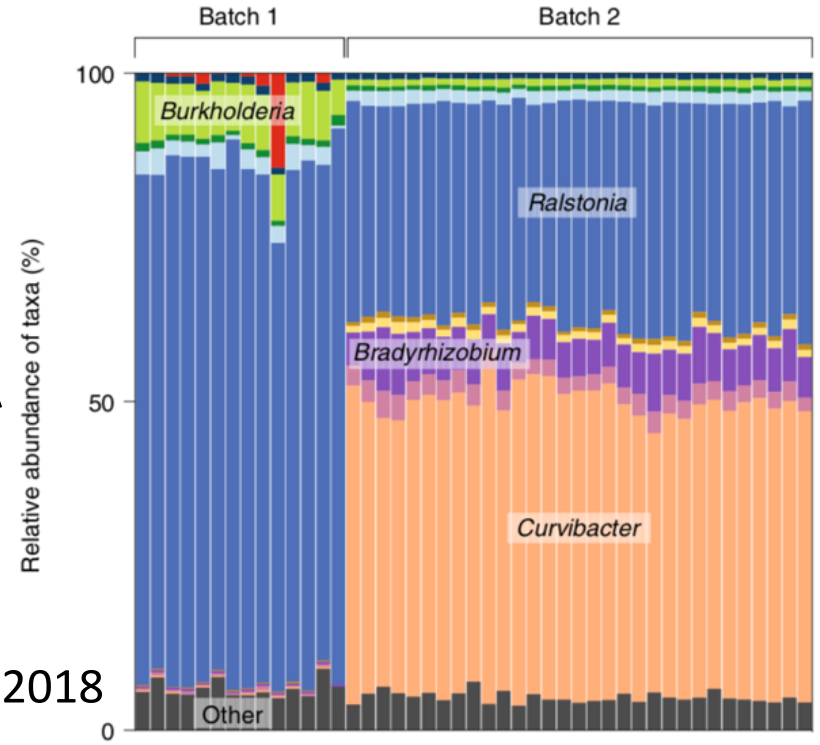
環境サンプルからのDNA抽出の難しさ

様々な夾雑物の存在

- 酵素反応阻害物質
  - フミン酸
  - DNase, RNase, proteinase
  - DNAの架橋促進物質

ターゲットのDNAが極少量の場合

- Host DNA/RNA
- 大量のRibosomal RNA
- キットや水、試薬中のコンタミDNA



de Goffau MC. et al. Nature Microbiol. 2018

自分で条件検討をするのは大変なので、先行研究を参考にすべき

メタゲノム解析の代表例

公共のマイクロバイオームサンプルの内訳 (2024年2月時点)

環境区分	プロジェクト数	サンプル数	平均サンプル数
自然環境	**69,628**	**2,185,756**	31.3
土壌	19,694	772,568	39.2
海水	7,753	341,949	44.1
宿主共生	**39,368**	**3,266,106**	82.9
ヒト	8,280	1,489,961	179.9
マウス	4,289	275,727	64.2

ショットガンメタゲノムに絞ると上記の約1/12

Human microbiome

- 消化管・皮膚・鼻腔・性器等に共生微生物が存在
  - 数百–数千種
  - 総重量約0.5 kg
  - 細胞数: ヒトの細胞数の1.5倍以上
  - 遺伝子の種類数: ヒトの遺伝子の数百倍以上

- ヒトの健康に大きく影響 (特に腸内細菌)
  - 様々な物質の代謝(ビタミン・脂肪酸・二次胆汁酸・アンモニア等)
  - 薬の代謝
  - 免疫の発達
  - 脳腸相関
  - 100以上の様々な病気と関連(ガン・自己免疫性疾患等)

ヒトマイクロバイオームについて、特に詳細にメタデータをキュレーション(2024年2月時点)

全世界

ヒトの体の部位	プロジェクト数	サンプル数
gut (feces)	4,633	847,414
oral cavity	1,195	153,980
respiratory system	755	72,700
skin	446	92,788
reproductive system	443	95,417
others	808	227,662

othersにはメタデータが無く分類できないものも含む

ヒトマイクロバイオームについて、特に詳細にメタデータをキュレーション(2024年2月時点)

日本人のみ

ヒトの体の部位	プロジェクト数	サンプル数
gut (feces)	144	15,294
oral cavity	26	4,187
respiratory system	5	168
skin	10	1,761
reproductive system	2	145
others (尿路、血液、眼、bile)	9	519

197 projects, 22,074 samples (shotgun 2,460 samples)

データサイズ: 約6 TB (shotgun 5.6TB)

世界のヒトマイクロバイオームサンプルの約1.4%

Summary of published human genome-microbiome association studies (2016~2022)

Paper	Population	Number of samples	Analysis method
Goodrich, J. K. et al. *Cell Host Microbe* 2016	UK	1126 pairs	16S rRNA
Wang, J. et al. *Nat. Genet.* 2016	German	1812	16S rRNA
Turpin, W. et al. *Nat. Genet.* 2016	Canadian	1561	16S rRNA
Bonder, M. J. et al. *Nat. Genet.* 2016	Dutch	1514	Shotgun
Rothschild, D. et al. *Nature* 2018	Israeli	1046	16S rRNA
Hughes, D. A. et al. *Nat. Microbiol.* 2020	Belgian, German	3890	16S rRNA
Xu, F. et al. *Microbiome* 2020	Chinese	1475	16S rRNA
Liu, X. et al. *Cell Discov.* 2020	Chinese	1295	Shotgun
Ishida, S. et al. *Commun. Biol.* 2020	Japan	1068	16S rRNA
Rühlemann, M. C. et al. *Nat. Genet.* 2021	German	8956	16S rRNA
Kurilshikov, A. et al. *Nat. Genet.* 2021	Many populations	18340	16S rRNA
Qin, Y. et al. *Nat. Genet.* 2022	Finnish	5959	Shotgun
Lopera-Maya, E. A. et al. *Nat. Genet.* 2022	Dutch	7738	Shotgun
Boulund, U. et al. *Cell Host Microbe* 2022	Many populations	4117	16S rRNA

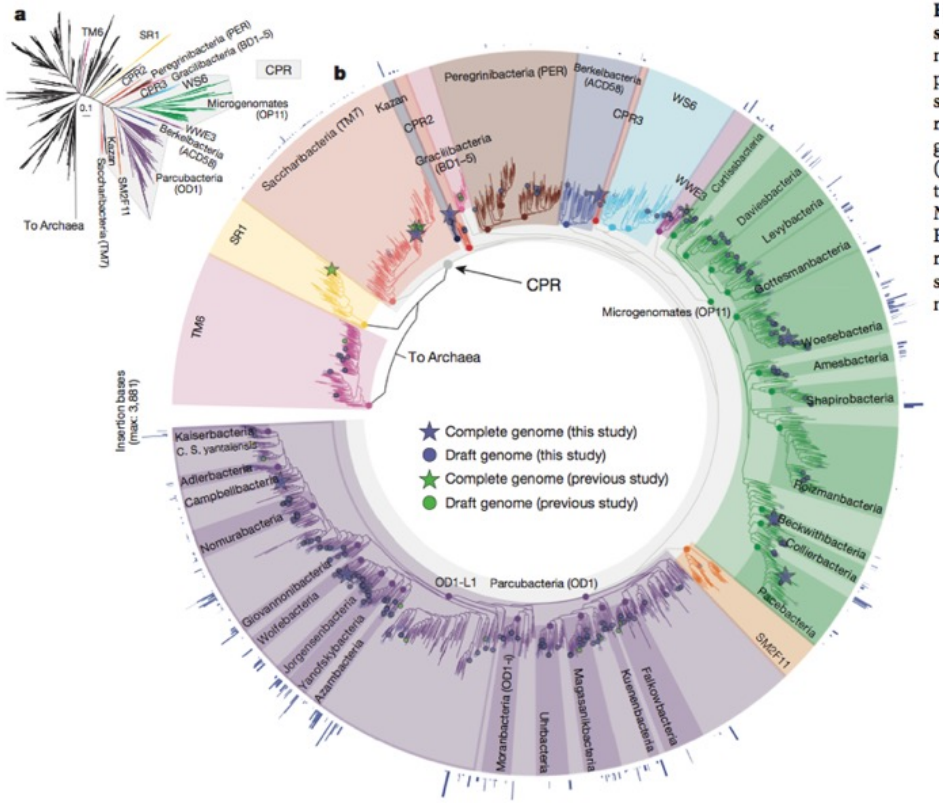
Sanna S. et al. Nat. Genet. 2022 modified

LETTER Nature, 2015

doi:10.1038/nature1448

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]

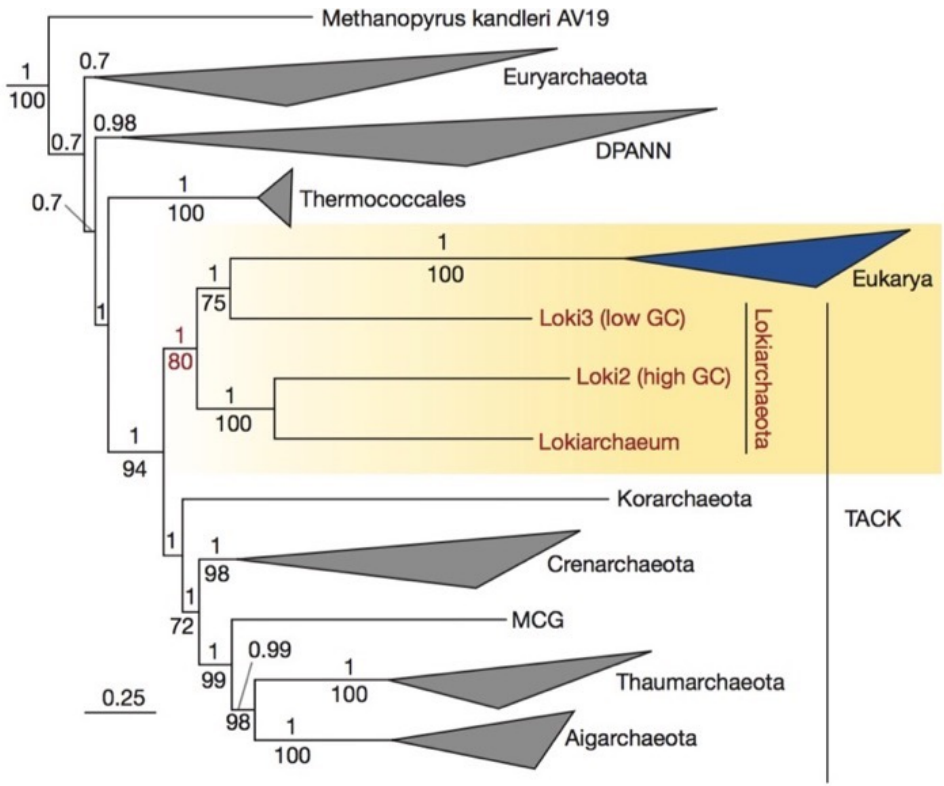


ARTICLE Nature, 2015

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]



Candidate Phyla Radiation (CPR)やAsgard clade等、培養を経ずに
メタゲノム配列データから機能や進化的な類縁関係が推定された系統が多数存在

メタゲノムの情報解析法

Metagenomic sequencing analysis
(メタゲノム解析, ショットガンメタゲノム解析)
(メタトランスクリプトーム解析)

DNA extraction / RNA extraction

Reverse transcription

DNA Sequencing

TogoTV (© 2016 DBCLS TogoTV)

Metagenomic reads

Assemble — MEGAHIT, SPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

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

KEGG, RefSeq等

Gene Function abundance
Pathway abundance
Taxonomic abundance

Sample1 Metadata

Sample2 Metadata

Comparative metagenomics

Who's there?

What are they doing?

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	△	◯	×	◯	mOTUs3
Unique marker genes	△	◯	×	◯?	MetaPhlAn4
Read mapping	△	×	◯	×	BWA-MEM, Centrifuge
k-mer	△	×	◯	×	Kraken2, sourmash

- 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, **SPAdes**, 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

Prodigal, MetaGeneMark, MetaGeneAnnotator, etc.

- メタゲノム遺伝子機能推定

Reference配列から遠い場合にも配列類似性を検出する必要がある

DIAMOND, **MMseqs2**, etc.

細菌の場合、基本的に変異は個体間で
独立に蓄積していき、個体間で均質化されない

少数のクローン個体から開始した実験室環境ならまだしも、
群集が形成されてから長い年月が経った自然環境では、
種内の多型が大量に存在するはず。

アセンブルの結果できた配列は、あくまで群集の
平均値であることに注意すべき。

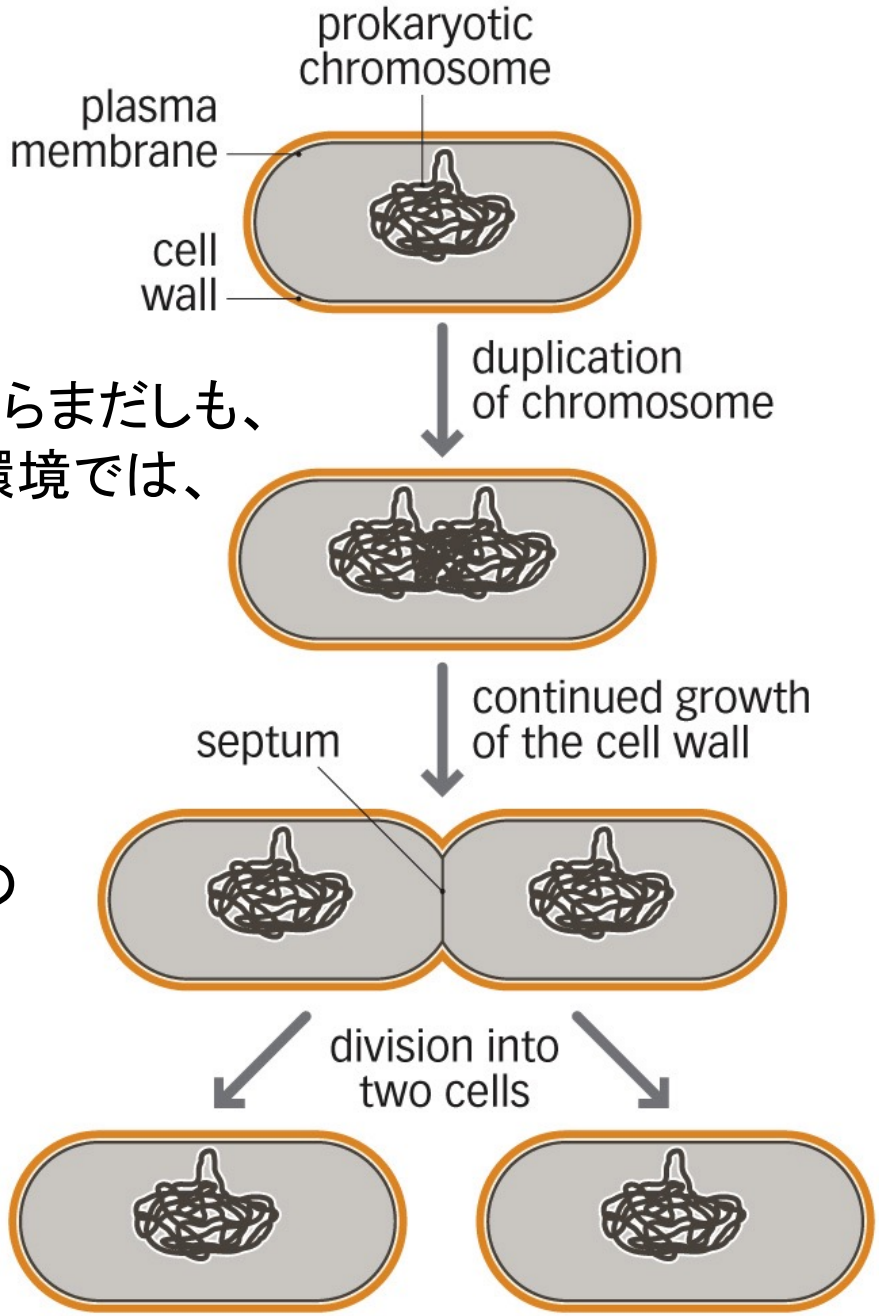


Figure 9.3 Microbiology: A Clinical Approach 2e (© Garland Science 2016)

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 (https://www.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。

- SPAdes (Nurk et al. 2017) メタゲノムオプション使用
  - 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。

Representative Next Generation Sequencer (NGS) in June 2024

Sequencer Name	Specific property	Read length (base)	Read number / run
ABI 3730xl (not NGS)	Sanger	500–1000	384
illumina iSeq	Bridge PCR	150	4,000,000
MiSeq	Bridge PCR	300	25,000,000
NextSeq 550/2000	Bridge PCR	150	400,000,000
NovaSeq 6000	Bridge PCR	150 or 250	~10,000,000,000
PacBio Sequel IIe / Revio	Single molecule	Average 20,000	~8,000,000 (Sequel IIe) ~25,000,000 (Revio)
Nanopore MinION(1), GridION (5), PromethION(48)	Single molecule	Average 10,000-30,000	~400,000 (1 flow cell in 12 hours) ~2,000,000 (1 PromethION flow cell)

nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-0422-6

Check for updates

2020

OPEN

Complete, closed bacterial genomes from microbiomes using nanopore sequencing

Eli L. Moss[1,3], Dylan G. Maghini[1,3] and Ami S. Bhatt[1,2]

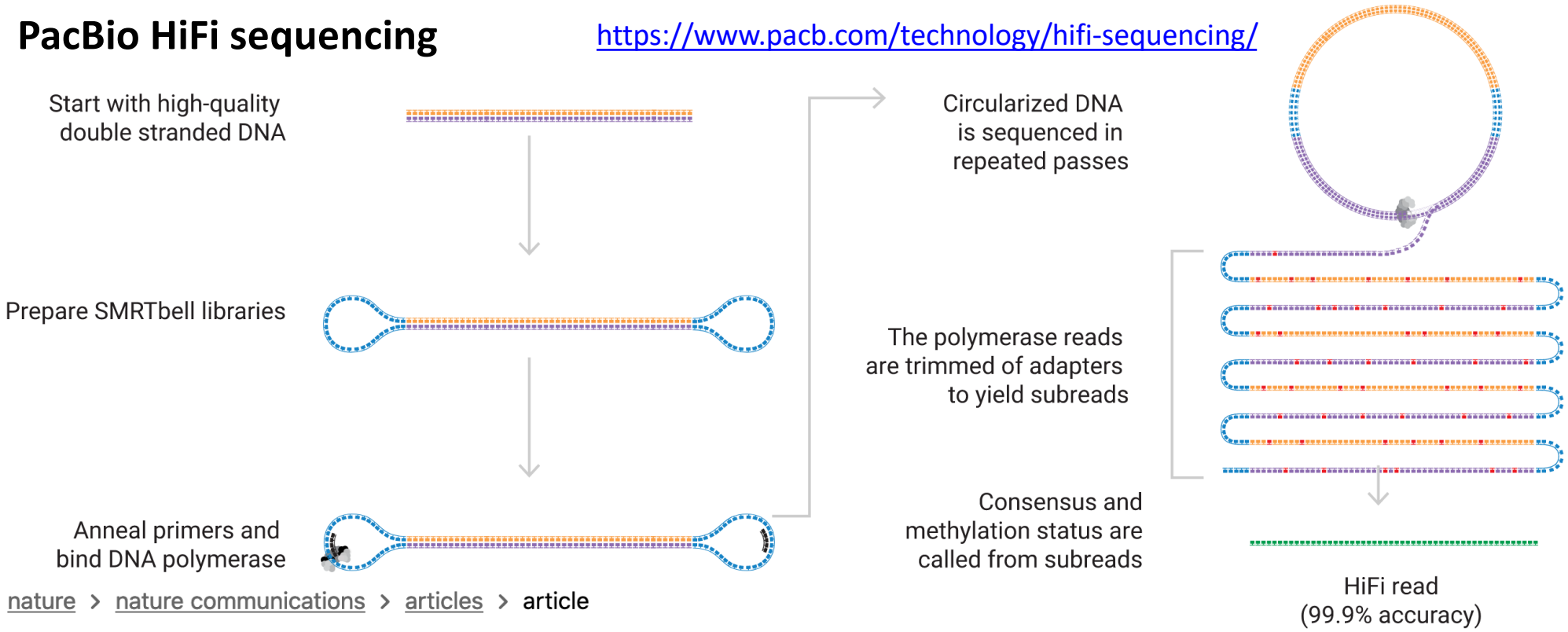
Table 1 | Circular bacterial genomes assembled from human stool samples

Genome	Sample	Assembler	Genome size (Mbp)	Genes	16S rRNA
Dialister sp.	P1	Canu	1.96	1,912	4
Dialister sp.	P2-A	Canu	1.89	1,803	4
Faecalibacterium prausnitzii[c]	P1	Canu	3.4	3,234	6
Oscillibacter sp.	P1	Canu	3.04	2,926	3
Phascolarctobacterium faecium	P2-B	Canu	2.35	2,307	5
P. copri	P2-A	Canu	3.71	3,324	5
Akkermansia muciniphila	I	Flye	3.01	2,906	3
Anaerotruncus sp.	F	Flye	2.11	2,156	2
Bacteroides sp.	F	Flye	3.04	2,467	3
Clostridales sp.[a]	D	Flye	2.05	1,971	2
Eubacterium siraeum[a]	F	Flye	3.12	2,894	3
Eubacterium sp.[a]	G	Flye	2.11	2,043	2
Methanobrevibacter smithii	B	Flye	1.78	3,579	2
Oscillibacter sp.[a]	G	Flye	3.29	3,169	3
Phascolarctobacterium faecium	I	Flye	2.35	2,481	5
Prevotella sp.	F	Flye	3.46	3,031	5
Roseburia sp.[a]	D	Flye	2.17	2,953	2
Ruminococcus bromii	G	Flye	2.21	2,820	3
Ruminococcus sp.	E	Flye	2.48	3,007	4
Sellimonas intestinales[a]	D	Flye	1.76	2,889	3

Short readsでの補正は必須

PacBio HiFi sequencing

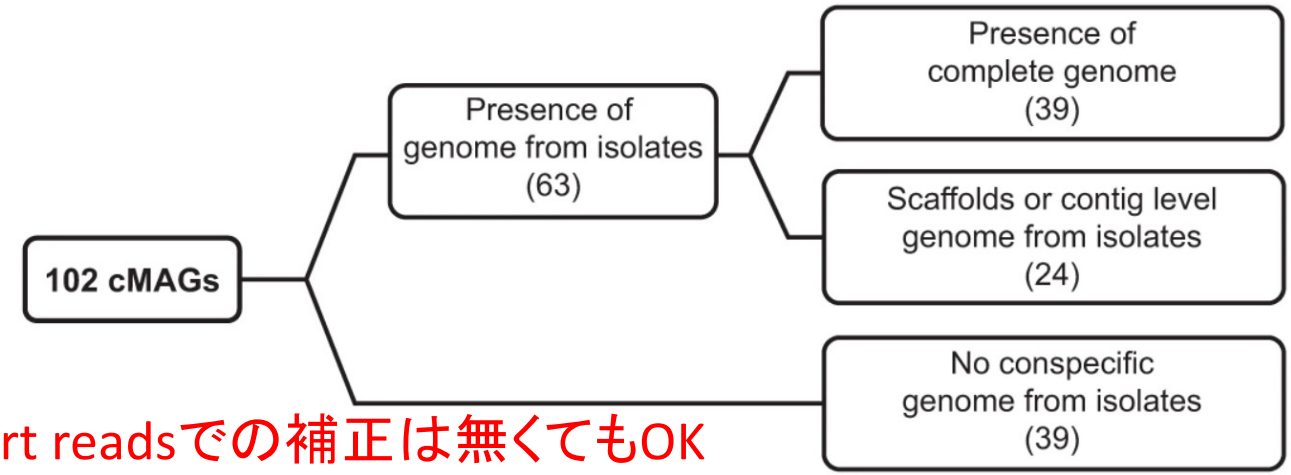
https://www.pacb.com/technology/hifi-sequencing/



Article | Open Access | Published: 26 October 2022

HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota

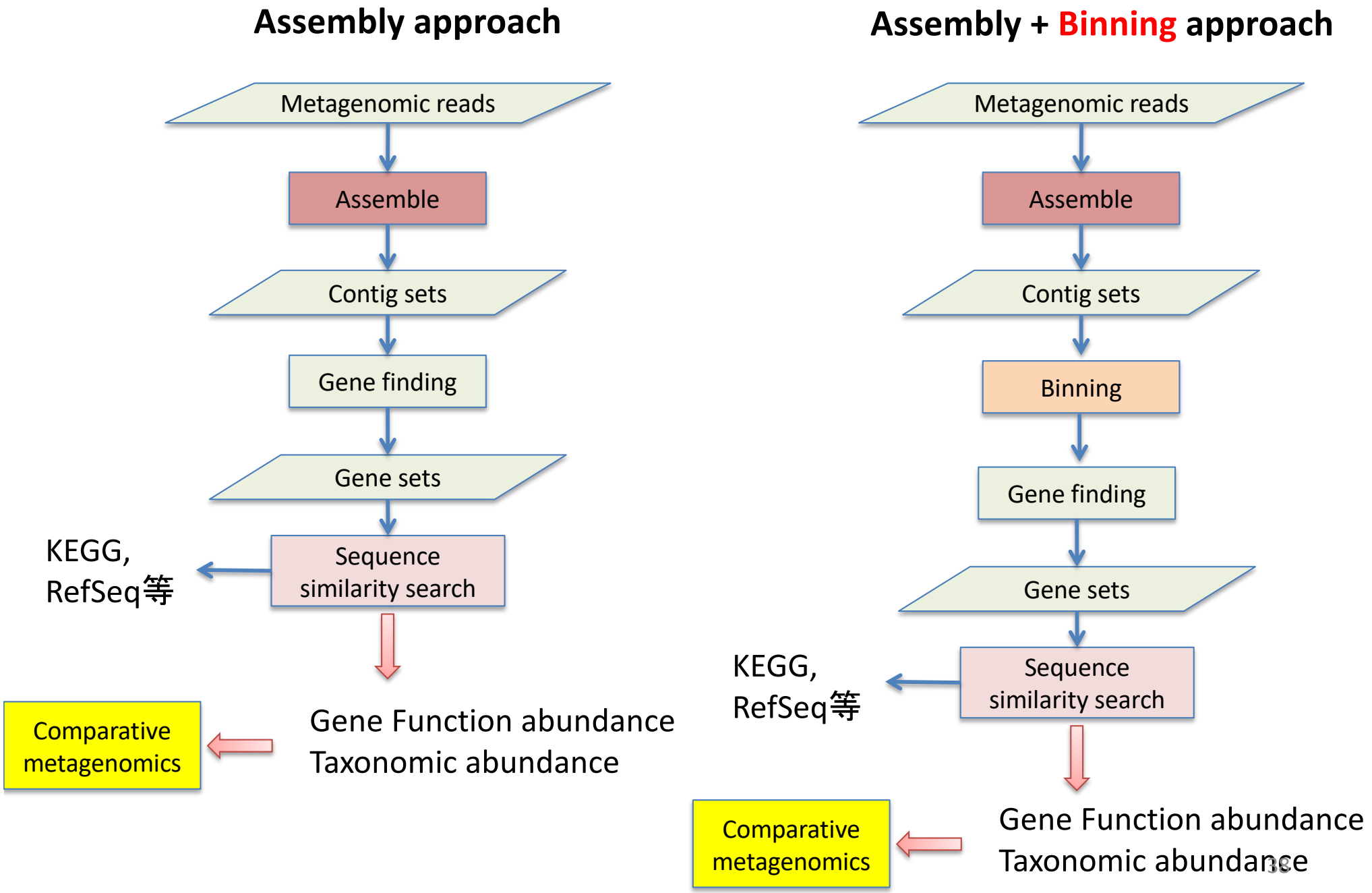
Chan Yeong Kim, Junyeong Ma & Insuk Lee



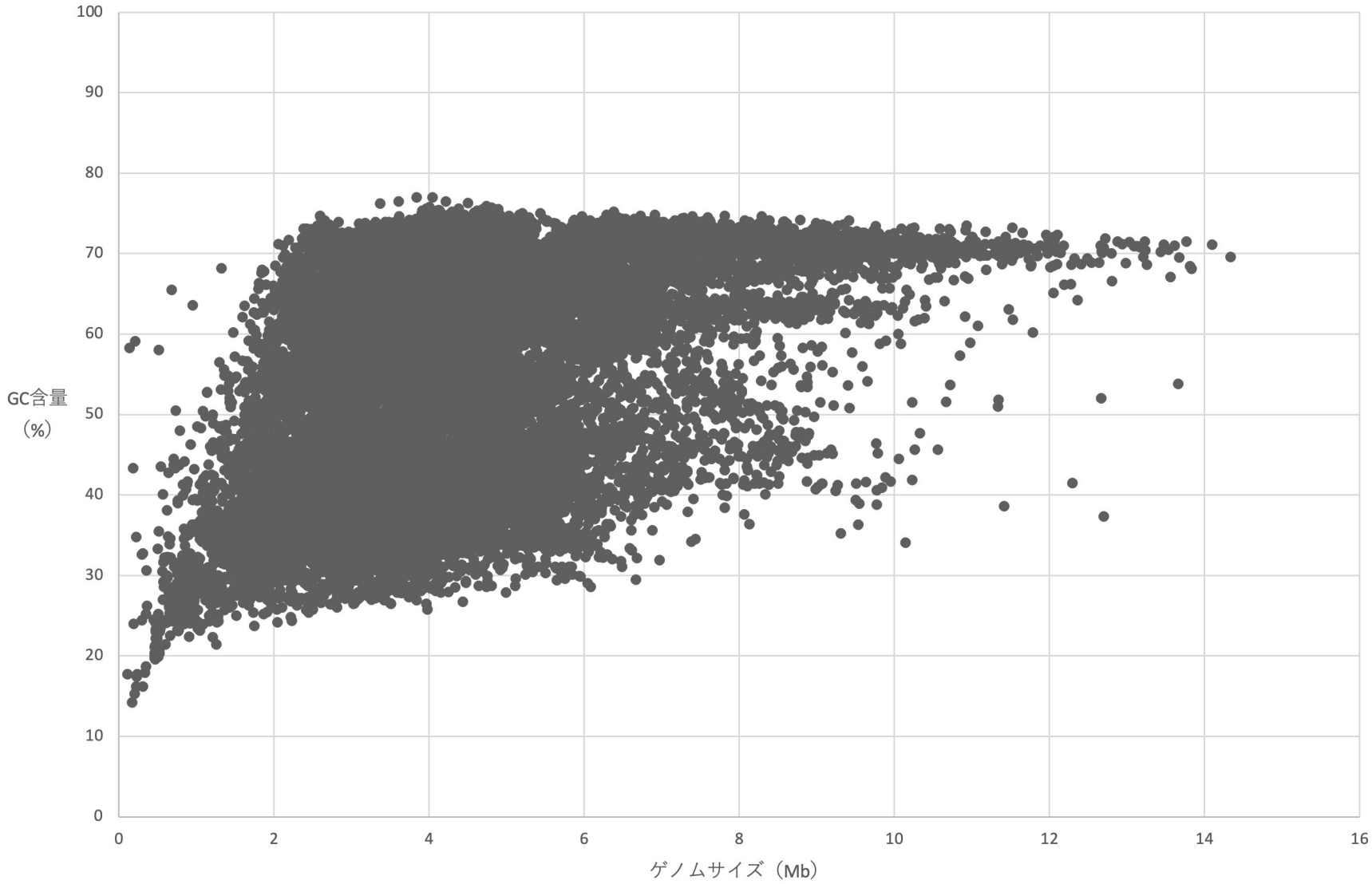
Short readsでの補正は無くてもOK

Metagenome Assembled Genome (MAG)

遺伝子組成解析とMAG解析



バクテリアゲノムのGC含量の多様性



NCBI Datasetsより原核生物18600ゲノムの統計情報をダウンロードして作成

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 A. et al. Nature Methods. 2017

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

大規模MAG研究の例

nature microbiology

Article | OPEN | Published: 11 September 2017

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 & Gene W. Tyson

Nature Microbiology **2**, 1533–1542 (2017) | Download Citation

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 16 January 2018

The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans

Benjamin J. Tully, Elaina D. Graham & John F. Heidelberg

Scientific Data **5**, Article number: 170203 (2018) | Download Citation

Cell

RESOURCE | VOLUME 176, ISSUE 3, P649-662.E20, JANUARY 24, 2019

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

Edoardo Pasolli • Francesco Asnicar [8] • Serena Manara [8] • ... Christopher Quince • Curtis Huttenhower • Nicola Segata [9] • Show all authors • Show footnotes

Open Access • Published: January 17, 2019 • DOI: https://doi.org/10.1016/j.cell.2019.01.001 •

nature
International journal of science

Article | OPEN | Published: 13 March 2019

New insights from uncultivated genomes of the global human gut microbiome

Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard & Nikos C. Kyrpides

Nature **568**, 505–510 (2019) | Download Citation

This article has been updated

nature
International journal of science

Article | OPEN | Published: 11 February 2019

A new genomic blueprint of the human gut microbiota

Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn

Nature **568**, 499–504 (2019) | Download Citation

scientific data

2022

OPEN

DATA DESCRIPTOR

The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments

Yosuke Nishimura [1,4] & Susumu Yoshizawa [1,2,3]

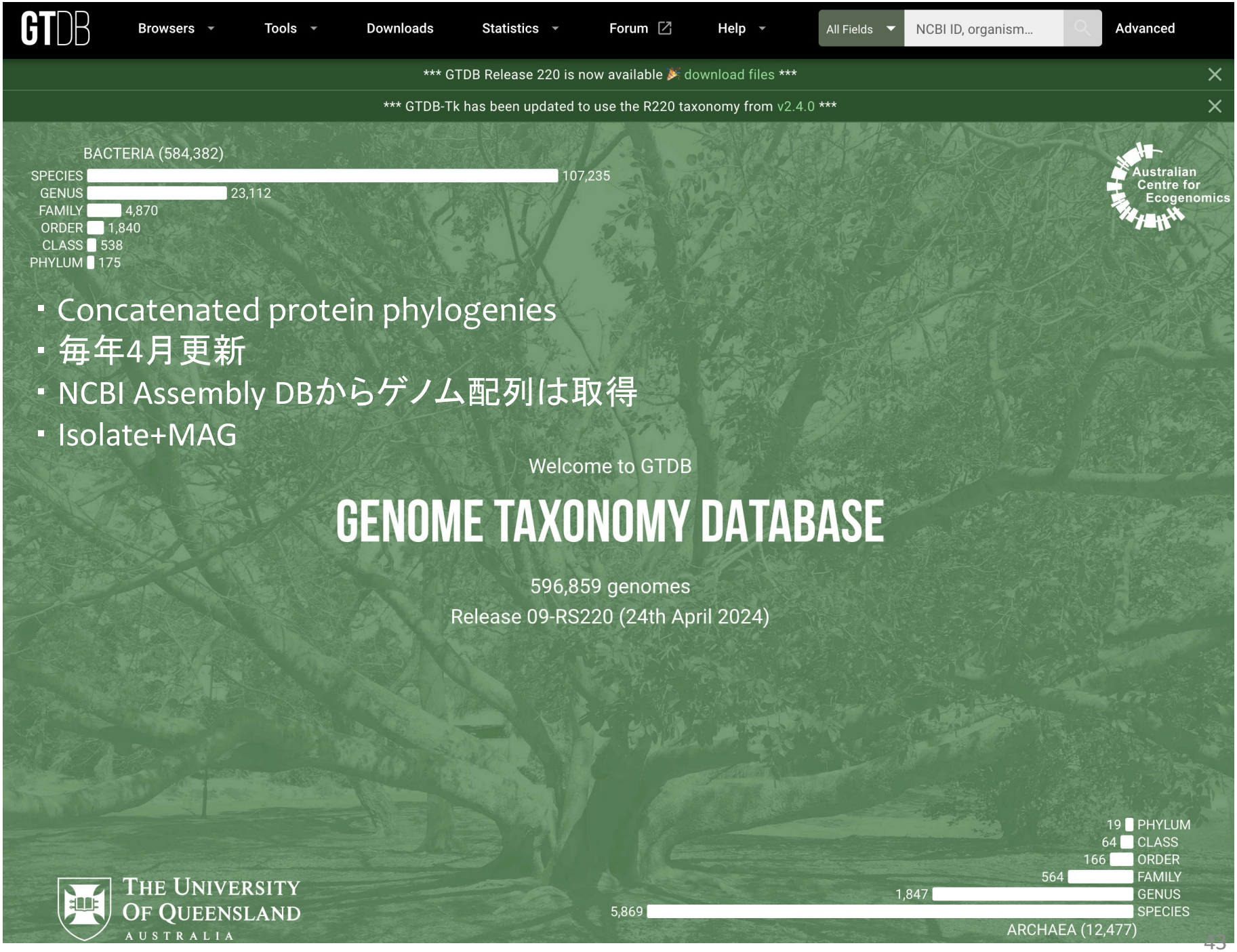
培養を経ずにメタゲノム配列データから
機能や進化的な類縁関係が推定された系統が多数存在

In February 2024

Taxonomic rank	Cultured and validly described taxa	Genome-based putative taxa including uncultured taxa
Phylum	49	181
Class	157	548
Order	294	1,772
Family	717	4,772
Genus	4,079	20,739
Species	24,363	85,205

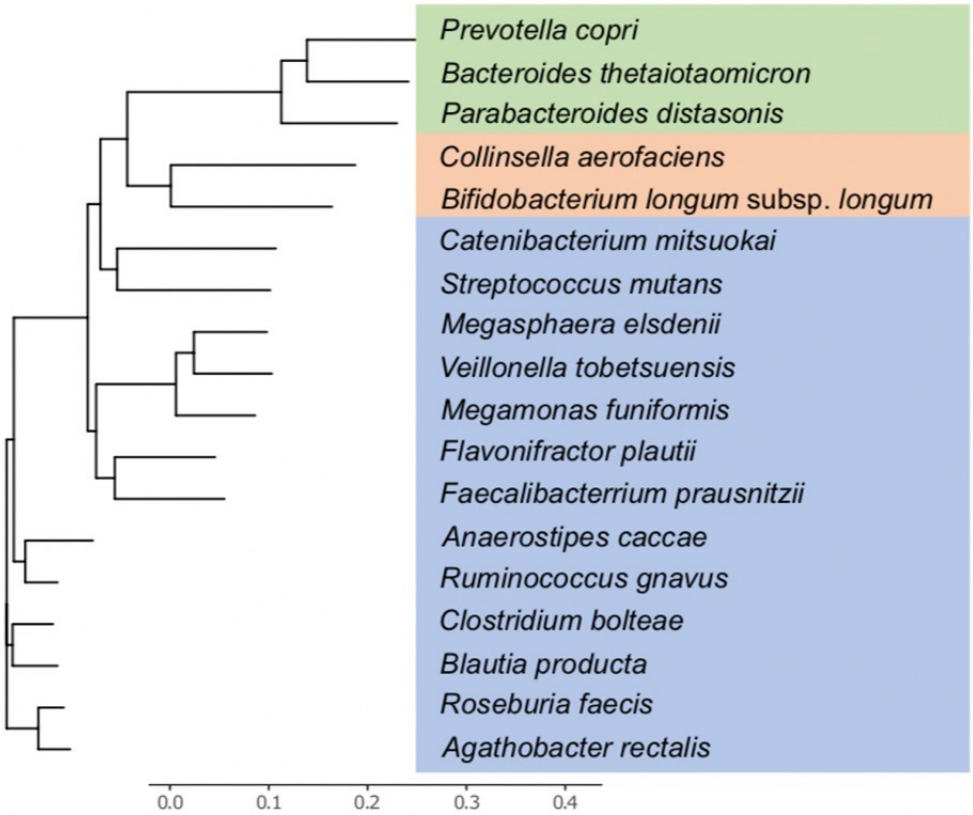
List of prokaryotic names with standing in nomenclature (https://lpsn.dsmz.de/text/numbers)
Genome Taxonomy Database (https://gtdb.ecogenomic.org/stats)

https://gtdb.ecogenomic.org/



MAG construction test using a metagenomic sequence data from an 18 species DNA even-mix sample

NovaSeq 6000 & MetaBAT2 & Gene content analysis



Completeness (%)	NovaSeq
Bacteroides thetaiotaomicron	94.28
Parabacteroides distasonis	92.02
Agathobacter rectalis	38.78
Anaerostipes caccae	22.49
Prevotella copri JCM	34.08
Collinsella aerofaciens	33.53
Catenibacterium mitsuokai	48.31
Bifidobacterium longum	50.25
Flavonifractor plautii	64.75
Megamonas funiformis	24.45
Blautia producta	38.87
Veillonella tobetsuensis	93.16
Roseburia faecis	34.11
Clostridium bolteae	50.85
Faecalibacterium prausnitzii	34.07
Megasphaera elsdenii	24.98
Streptococcus mutans	72.19
Ruminococcus gnavus	26.16

3–5 / 18 species' MAGs are very high quality

Mori H. et al. DNA Res. 2023

Short readでは群集を構成する全ての微生物のMAGが得られるわけでは無い

Short or Long reads?

	Cost	必要なDNA量	InDel error	Construct good ref.	gene-neighbor
Short reads	low	low	low	◯	△
Long reads	high	high	**high**	◎	◯

高分子DNAが数μgオーダーで必要になる点は課題だが、
Long readを用いてメタゲノムからの完全ゲノム構築が可能になりつつある

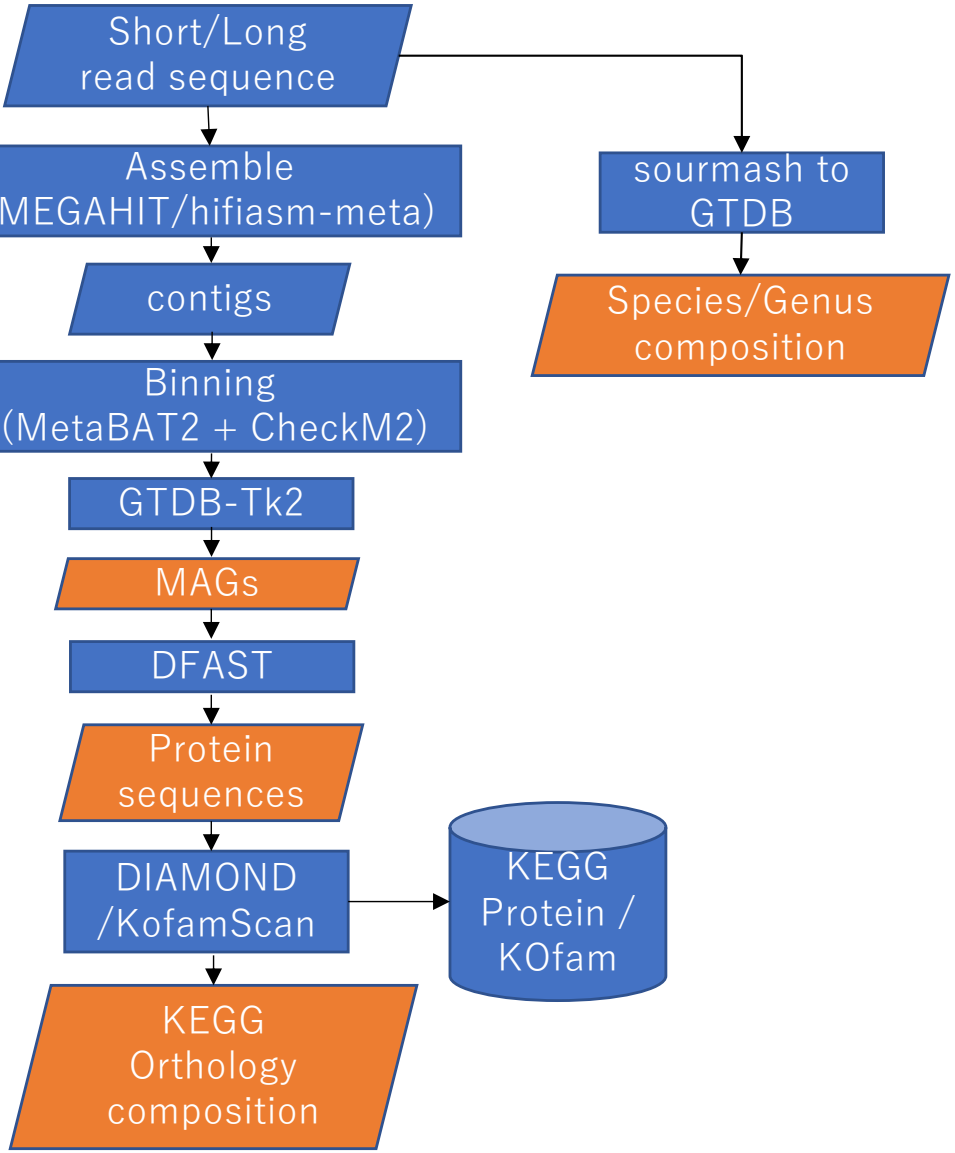
Nanopore GridION

Nanopore PromethION

PacBio Sequel IIe

PacBio Revio

MAG解析ワークフローの例



基本的にShort readとlong readではアセンブルツールが異なる
Long read専用の高精度なMAG構築binningツールは現状存在しない

Short read MAG構築の結果の例 (CheckM2の結果)

Bin Id	Marker lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination	Strain heterogeneity
bin.10	k__Bacteria (UID203)	5449	99	53	1	8	19	15	19	37	98.11	310.90	1.66
bin.24	p__Euryarchaeota (UID49)	95	228	153	4	223	1	0	0	0	97.39	0.13	0.00
bin.38	c__Deltaproteobacteria (UID3216)	83	247	155	12	227	8	0	0	0	96.77	4.35	25.00
bin.27	k__Bacteria (UID203)	5449	104	58	5	79	20	0	0	0	96.08	7.68	5.00
bin.7	c__Deltaproteobacteria (UID3216)	83	247	155	11	231	5	0	0	0	95.99	1.68	40.00
bin.31	c__Deltaproteobacteria (UID3217)	62	280	168	11	259	9	1	0	0	95.44	5.51	33.33
bin.69	c__Deltaproteobacteria (UID3217)	62	280	168	14	257	9	0	0	0	95.24	2.23	22.22
bin.72	k__Bacteria (UID3060)	138	335	243	20	303	11	1	0	0	93.80	3.70	35.71
bin.40	k__Bacteria (UID203)	5449	104	58	4	78	21	1	0	0	93.10	9.25	8.33
bin.17	k__Bacteria (UID2566)	525	208	136	17	187	4	0	0	0	92.28	1.96	25.00
bin.9	k__Bacteria (UID2495)	2993	147	91	10	134	3	0	0	0	92.11	0.71	66.67
bin.39	c__Deltaproteobacteria (UID3216)	83	247	155	17	225	5	0	0	0	90.16	1.36	60.00
bin.11	k__Bacteria (UID2565)	2921	143	88	11	131	1	0	0	0	89.71	1.14	0.00
bin.36	k__Bacteria (UID203)	5449	104	58	6	19	79	0	0	0	89.66	67.06	69.62
bin.28	k__Bacteria (UID3060)	138	335	244	55	90	90	56	32	12	87.39	112.01	21.93
bin.62	k__Bacteria (UID2495)	2993	147	91	18	120	9	0	0	0	85.49	6.24	55.56
bin.46	k__Bacteria (UID2566)	525	208	136	22	172	14	0	0	0	85.29	4.94	42.86
bin.70	k__Bacteria (UID203)	5449	104	58	12	26	47	18	1	0	85.19	69.87	54.21
bin.22	k__Bacteria (UID2569)	434	278	186	52	216	10	0	0	0	82.53	3.41	10.00
bin.52	k__Bacteria (UID3060)	138	335	244	81	228	25	1	0	0	81.72	8.21	46.43
bin.59	k__Bacteria (UID2566)	525	208	136	40	145	19	4	0	0	80.96	12.67	29.03
bin.43	k__Bacteria (UID2569)	434	278	186	43	226	8	1	0	0	80.63	3.51	18.18
bin.32	k__Bacteria (UID203)	5449	104	58	14	66	20	4	0	0	79.15	17.55	9.38
bin.26	k__Bacteria (UID2569)	434	278	186	45	202	30	1	0	0	78.85	8.29	36.36
bin.20	k__Bacteria (UID2495)	2993	147	91	26	113	7	1	0	0	77.47	5.79	20.00
bin.55	c__Deltaproteobacteria (UID3217)	62	280	168	80	186	14	0	0	0	77.41	5.32	35.71
bin.12	f__Flavobacteriaceae (UID2845)	53	548	298	126	366	53	3	0	0	76.85	10.83	3.23
bin.45	c__Deltaproteobacteria (UID3217)	62	280	168	67	169	38	6	0	0	75.54	18.82	1.79
bin.29	k__Bacteria (UID203)	5449	102	56	19	82	1	0	0	0	74.95	1.79	0.00
bin.30	k__Bacteria (UID203)	5449	104	58	39	25	26	8	4	2	74.11	68.85	47.87
bin.21	k__Bacteria (UID1452)	924	163	110	46	113	4	0	0	0	72.92	3.18	50.00
bin.63	k__Bacteria (UID1452)	924	163	110	36	122	5	0	0	0	72.27	2.64	40.00
bin.42	k__Bacteria (UID2570)	433	270	179	75	171	23	1	0	0	72.25	6.21	3.85
bin.49	k__Bacteria (UID3060)	138	335	244	94	226	14	1	0	0	70.11	5.61	29.41

・一般的にはCompleteness 90%以上、Contamination 10%以下が高精度なMAG
・各MAGのContig数は数十Contigs

パスウェイデータベース

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

KO (KEGG ORTHOLOGY) Database https://www.genome.jp/kegg/ko.html

Linking genomes to biological systems by functional orthologs

KEGG2 PATHWAY BRITE MODULE KO Annotation Taxonomy Synteny Mapper

Search KO for Go

Enter K numbers (Example) K00161 K00162 K00163 K00627 K00382

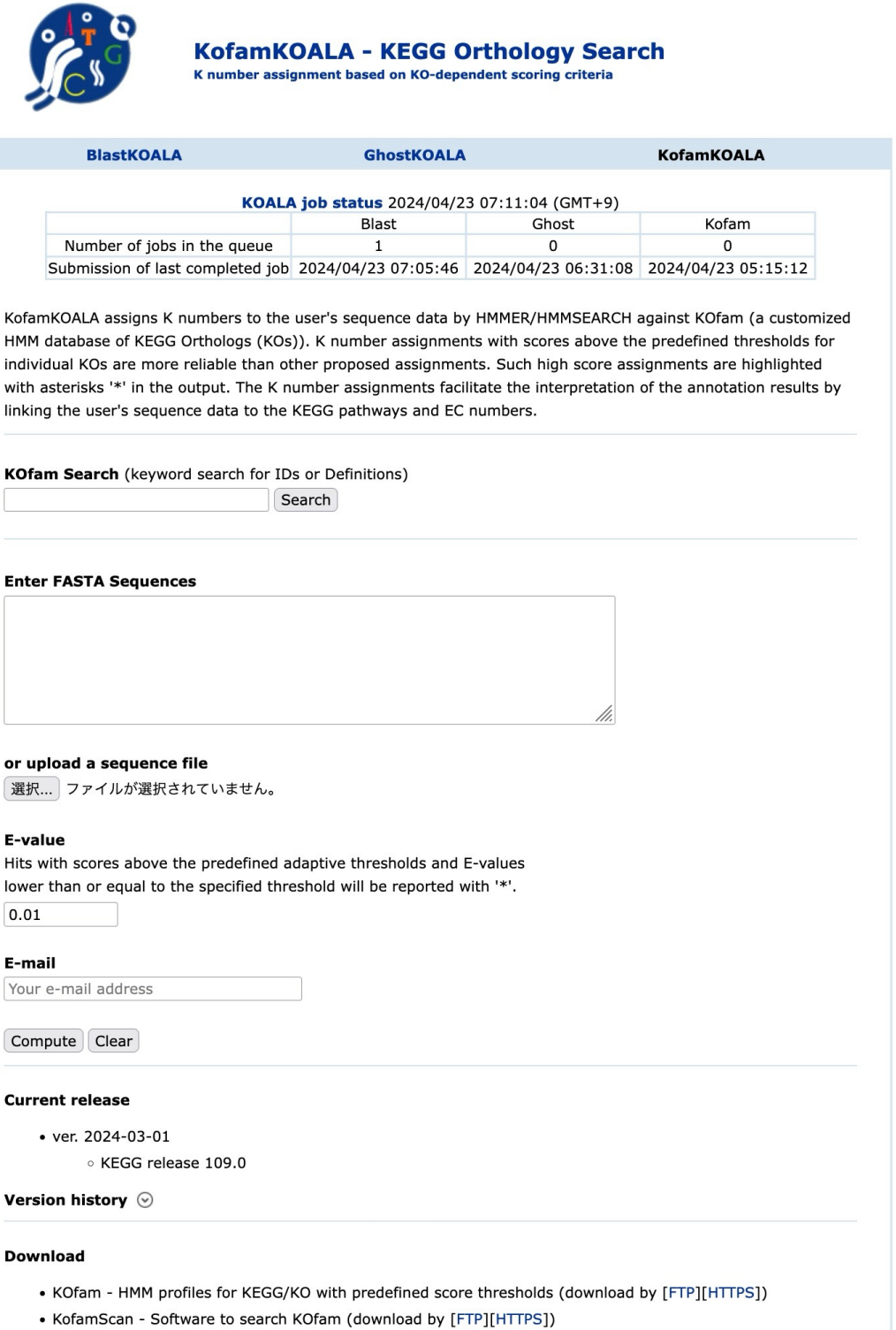
Filter Ortholog table Map pathway Map brite Map module Get title Get entry Clear

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. Each node of the network, such as a box in the KEGG pathway map, is given a KO identifier (called K number) as a functional ortholog defined from experimentally characterized genes and proteins in specific organisms, which are then used to assign orthologous genes in other organisms based on sequence similarity. The granularity of "function" is context-dependent, and the resulting KO grouping may correspond to a group of highly similar sequences within a limited organism group or it may be a more divergent group.

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

https://www.genome.jp/tools/kofamkoala/



KofamKOALA - KEGG Orthology Search

K number assignment based on KO-dependent scoring criteria

BlastKOALA | GhostKOALA | KofamKOALA

KOALA job status 2024/04/23 07:11:04 (GMT+9)

	Blast	Ghost	Kofam
Number of jobs in the queue	1	0	0
Submission of last completed job	2024/04/23 07:05:46	2024/04/23 06:31:08	2024/04/23 05:15:12

KofamKOALA assigns K numbers to the user's sequence data by HMMER/HMMSEARCH against KOfam (a customized HMM database of KEGG Orthologs (KOs)). K number assignments with scores above the predefined thresholds for individual KOs are more reliable than other proposed assignments. Such high score assignments are highlighted with asterisks '*' in the output. The K number assignments facilitate the interpretation of the annotation results by linking the user's sequence data to the KEGG pathways and EC numbers.

KOfam Search (keyword search for IDs or Definitions)

Search

Enter FASTA Sequences

or upload a sequence file

選択... ファイルが選択されていません。

E-value

Hits with scores above the predefined adaptive thresholds and E-values lower than or equal to the specified threshold will be reported with '*'.

0.01

E-mail

Your e-mail address

Compute Clear

Current release

- ver. 2024-03-01
  - KEGG release 109.0

Version history

Download

- KOfam - HMM profiles for KEGG/KO with predefined score thresholds (download by [FTP][HTTPS])
- KofamScan - Software to search KOfam (download by [FTP][HTTPS])

各機能タンパク質群 (KOfam)について、マルチプルアライメントから構築した配列保存プロファイルに対して隠れマルコフモデルを用いてプロファイル検索する

KEGG FTPサイトを使うKOのアミノ酸配列そのものを使うと有料だが、KOの配列保存プロファイルは無料でダウンロード可能。

配列類似性検索ツールや検索速度はアミノ酸配列そのものと配列保存プロファイルでは異なる。

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

OPEN

Sensitive protein alignments at tree-of-life scale using DIAMOND

Nature Methods. 2021

Benjamin Buchfink[1], Klaus Reuter[2] and Hajk-Georg Drost[1]✉

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

BLAST

1. 完全マッチのseed探索 (BLASTPならsize 6)
2. gap無しでseed延伸
3. gapありのSmith-Waterman アルゴリズムでアラインメント

DIAMOND ver. 2.x (BLASTPの1000倍以上高速)

1. クエリとリファレンスを11圧縮アミノ酸に変換
2. クエリとリファレンス配列両方に対して、シードと場所のペアのテーブルを作成。（実際はリファレンス側は事前に作成してファイルに保存済みなので、クエリのみ）
   シードは、spaced seed
   fast mode = weight 10 のspaced seedで2 shapes
   sensitive mode = weight 8で16 shapes
   very-sensitive mode = weight 7で14 shapes
   ultra-sensitive mode = weight 7で64 shapes
3. クエリ-リファレンス間でseedがマッチしたら、seedを中心に48 AAのwindowでペアのhamming distanceを計算。一定以下なら4に進む
4. gap無しでseedを前後に伸ばしていき、scoreが一定以上になったら5に進む
5. クエリごとに全リファレンス配列中で4を突破したgap無しアライメント群をscoreでソート
6. score高い順に400ペアを一単位 (chunk)としてgapありSmith-Watermanアルゴリズムでアラインメント。400ペア全てで一定のスコアを突破するアラインメントが無かったら、そのqueryについては以降のchunkは全部捨てる。

遺伝子組成解析とMAG解析