# ゲノムアッセンブリ・アノテーションと配列データベース

2020.05.12

国立遺伝学研究所　谷沢靖洋

# 自己紹介

**谷沢靖洋（国立遺伝学研究所・**
**大量遺伝情報研究室/生命情報・DDBJセンター）**

植物および微生物を中心としたゲノム解析
ウェブアプリケーション・DB の開発


## 現在進行中のプロジェクト

***Marchantia polymorpha***
(ゼニゴケ, 220Mbp genome)

***Eustoma spp.***
(トルコキキョウ)

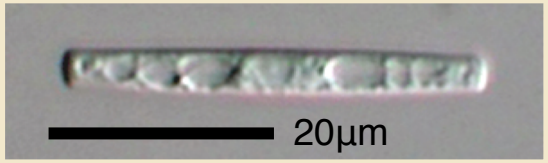



Figure (Ishii and Kamikawa, 2017, PBR)

***Nitzschia spp.***
(珪藻, 30-60 Mb genome)

# 次世代モデル植物　ゼニゴケ

***Marchantia polymorpha*** (liverwort, ゼニゴケ)

220 Mb genome,　常染色体 8 本　＋ 性染色体 1 本

京都大学・河内孝之教授、近畿大学・大和勝幸教授ら

との共同研究

~2017　ver3.1 genome.

Illumina, 454 などのNGSで解読したドラフトゲノム　(2,957 contigs)

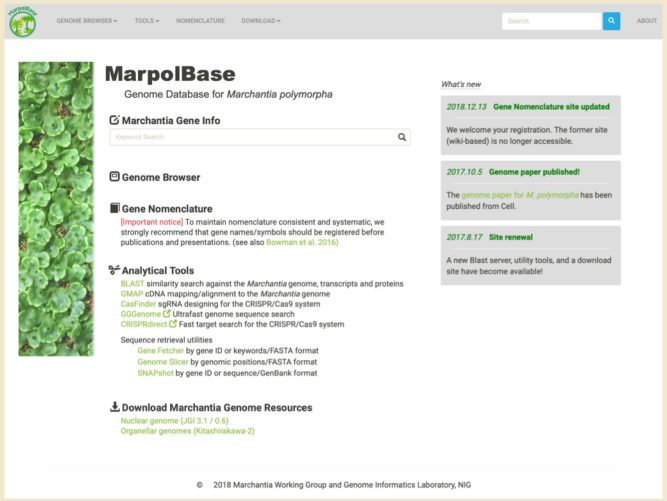
ゲノム解析論文　John Bowman et al. Cell 171(2) 287-304, 2017

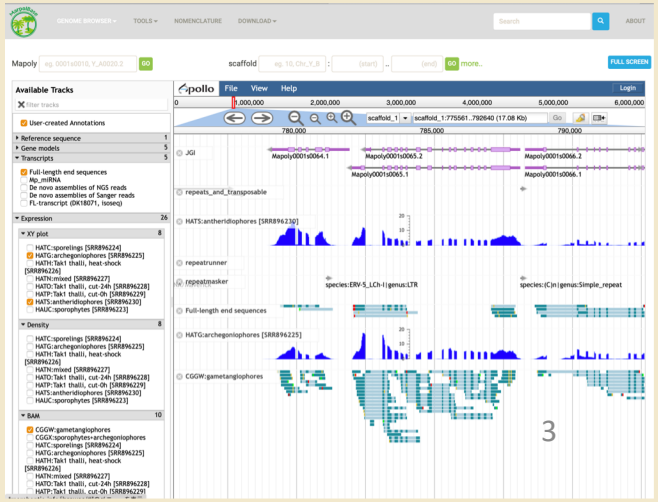
2018~　ver4 genome.

PacBio、HiCなどを使って構築した chromosome-level assembly

ゼニゴケゲノムDB

MarpolBaseを開発

# DDBJ　DNA Data Bank of Japan

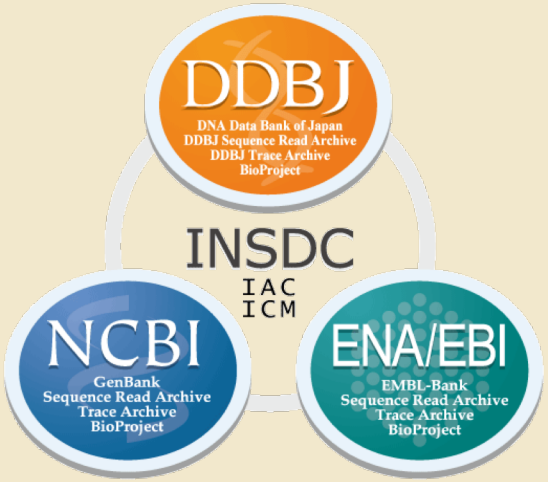
DDBJ
DNA Data Bank of Japan

INSDC (International Nucleotide Sequence Database Collaboration) の一員として、塩基配列データを収集・公開。

NCBI (米) および ENA/EBI (欧) との間でデータの共有。

生命科学における研究基盤としてスーパーコンピュータシステムを提供。

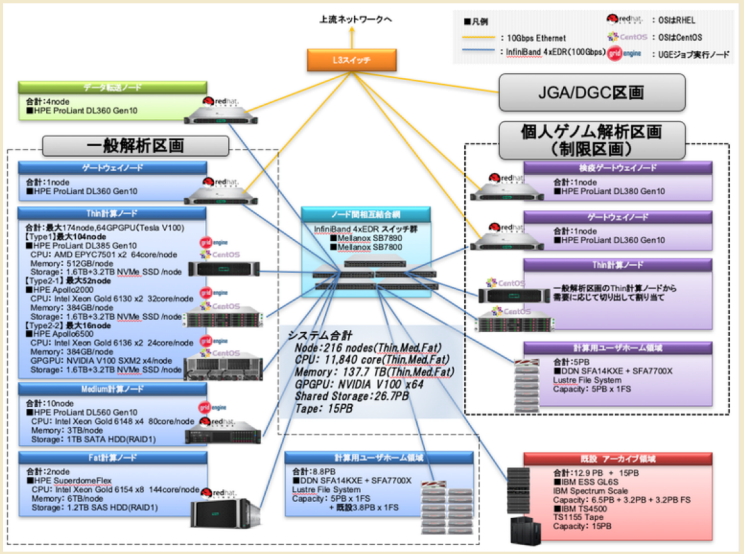
データ登録業務に従事するアノテータ16名 + システム運用のSE約20名




# 遺伝研スパコン

国際塩基配列DBのための計算基盤・ライフサイエンスのための共同計算基盤

無償で利用可能(一部課金サービスあり)




## 2019.3 リプレイス

Thin計算ノード

204台　合計約11,000 CPUコア
コアあたり8GB

Medium計算ノード

10台・80 CPUコア 3TBメモリ

Fat計算ノード

1台・144 CPUコア 12TBメモリ

一般計算用ストレージ　13.6PB
DB用ストレージ 30PB

個人ゲノム解析区画

2000以上の解析ツールの singularity コンテナ

# 本日の講義の内容

微生物ゲノムの*de novo*アッセンブリ手法

ゲノムアノテーションおよびその表現方法

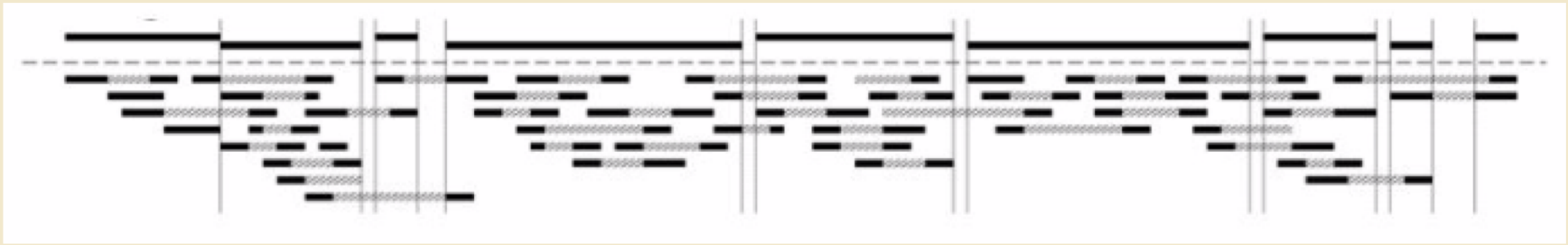
公共配列データベースの現状と諸問題

# 微生物ゲノムの*de novo* アセンブリ手法

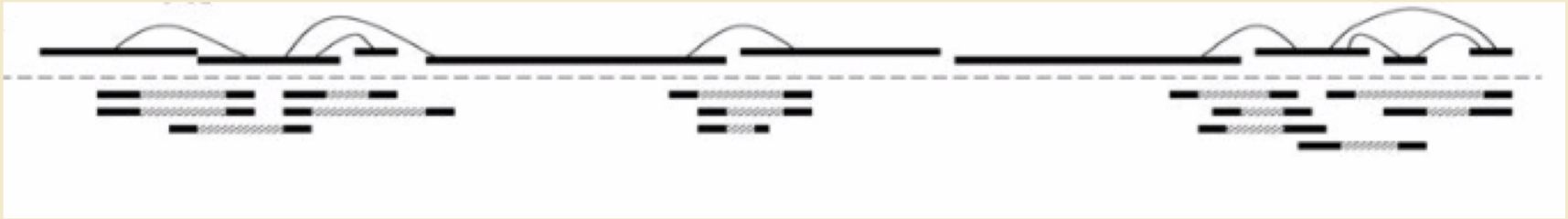
# ゲノムアッセンブリの一般的な流れ

DNAを断片化し、断片の両末端を読み取る（リード）

リードどうしの重なり合いを見つけてつなぎ合わせる（コンティグ）



リードのペア情報を利用して橋渡しをする（スキャフォールド）



図）バイオインフォマティクス研修 会 ゲノム配列解析および機能解析 2014.08.07

https://slidesplayer.net/slide/11110723/

# シーケンサーのリードデータはFASTQ 形式

４行で１件のデータを表す

クオリティは通常0-40までで、数値を文字に変換して表す。(“G”は38, “#”は2)

20 → 99%の信頼性 ($10^{-2}$のエラー率)　40 → 99.99%の信頼性 ($10^{-4}$のエラー率)

詳細は https://ja.wikipedia.org/wiki/Fastq

配列ID →
塩基配列 →
クオリティ →

```
@M00705:362:000000000-AL4WW:1:1101:20555:1000 1:N:0:3
NGAAAGCACTACTATTAACGGCTTAGGCATTAAGGCGGTGTTTAATGGCCAGCAGGTGCTCGTCGGCAACGAACTGCT...
+
#8ACCFC@EAG9DEGGGCG;FF:@8C,;6C,,C<9<FD6@,@,CC,,<E,;,@E@,;C8EFGGGGG@+@FF7,>,AC=...
@M00705:362:000000000-AL4WW:1:1101:17816:1001 1:N:0:3
GAACATATATGAAATATTTAAGACAAAAAATCTGACTCCGATAGAATATCGAAATCAGATTCTTAATCAGTAGTATAA...
+
#8BCCFGGGGGGGGGGGGGGGGGGGGFFFFFFFFFFFFGGGGGGGGFEFEGGGGGGGGGGGGGGGGGGGGFFGGG...
@M00705:362:000000000-AL4WW:1:1101:19096:1001 1:N:0:3
GATTCATATGAAGTTACGGCAAATTGACGATGGGACGATCACCATTGCCAATGTACCAGGACTGACTGAGAGTGAATT...
+
#8ABCGGGFGAFDC@@E7,@@CFD<CFFEGG8@@@EEGCFGGFGFG9<EFEE@,CFGF,CFGCEGG9@CA<ECFEFGG...
@M00705:362:000000000-AL4WW:1:1101:16996:1001 1:N:0:3
CAATTGGTCCGCAATTTCCACTTTATGTGAAATATTTGTTCGCAATTTTAACAAAATTTAAAAAGGCCAAAAGTCACA...
+
A8ACCGGGGGGGGGGGGGGGFEFEGGGGGGFfFGGFEFGGGGGGGGFCFGGFCFGGGFGGFDEEGGGGGGGGFGGGGG...
@M00705:362:000000000-AL4WW:1:1101:15425:1001 1:N:0:3
NTTCTTCAGGGTCGGTTTCTGGTACCTCTTGATGGTAGTCAGCCAACAATTCTTTACCATCGAGGCCGACAATTTGTG...
+
C8CCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEFGFGGGGGGGGFEFEGGGGGGFFFGGFEFGGGGGGGGGGG...
@M00705:362:000000000-AL4WW:1:1101:9600:1001 1:N:0:3
NTGCGAGCCCAATTGTTCAGTTGAACCACCTAACGGCATAATGCCGCCGCCGCGATGAGCAGCAATATTAGCTAAAAC...
+
A88CBF:CCFCFEFGGF<FAFF<FGGG<@CEFFFFGGG>FFGCFFECBEBE+:7@+@FEC,C,<,?FFGCGGGGGFFF...
```

# アッセンブルされたゲノムは通常FASTA形式で表す

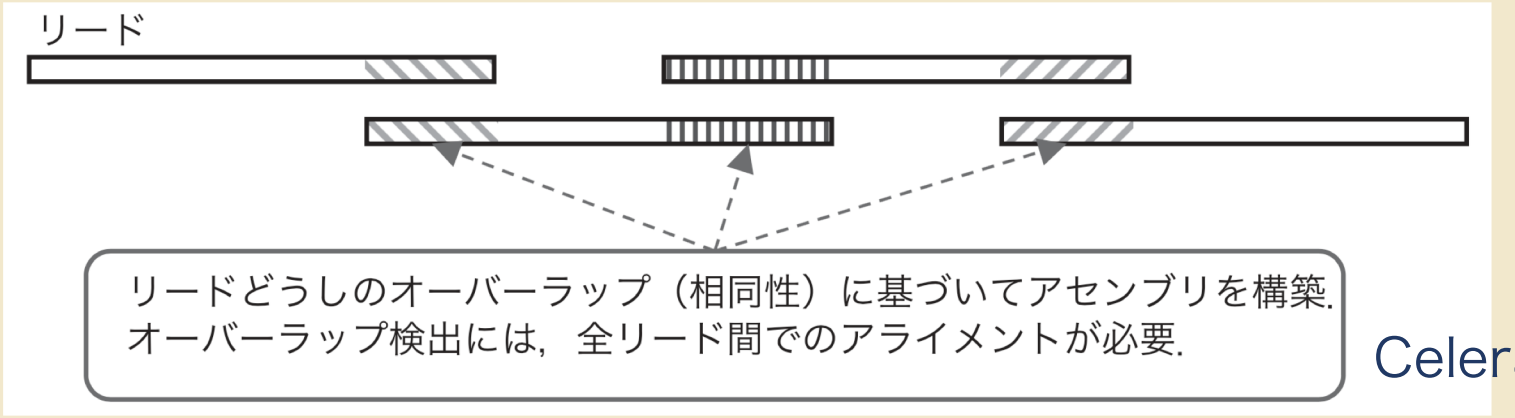
60~100文字で改行することが多い（改行しない場合もある）

配列が決定できなかった部分は未定塩基 N を用いてギャップとして扱われる

```
>scaffold_01
TGGTAATATTACTGTTGATTCATCAACGAGTAGCCCCATAGGGGCAATGGCAAAAGCATAAAGGTATCTA
TTAATTCGGATGTATAAATATTAAGTCGAATAAAAGGTATCTAGGAAAACTTGTGAGTACGTGAAGCTTA
AAAACGTCTGCTCGCGCTAAACGTCCTTGCTCTTTTTAAATGAAAAAGAGCCAAAGTCCATAAGGAGGTG
TTAATGGAACCAAAACGTGAAGCTTAGTTATGAAATTACGTACATCATTCGTCCTGACATGGATGAAGCT
ACAGCGCTTGTTGAACGATTTGACAAGATTGTGGCGCTAAACGTCAGATAATGGTGCTACGATCGTTGAT
GACTGGTCTACTCGTCGATTTGCTTATGAAATTGTGAAGCTTAGGTGATTACAACGAAGGTACTTACCAT
AATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGCGCTAAACG
ATCTTGCGTCACATGATTGTTAAGCGTGAAGCTTAATCACATGATTGTTCTAATCAATTTAAAGTTAAGA
TTAGAATCAAACGTGCTCGGATTATGGGTCTGCTACCATTCGTTGCAGAAGACTAATTTGGAAGACTGAC
TCCATATTGTATCTCTCGAGCCAATTAAATCAATTTTTGTGTTCTTAGGAAACTGCCAGAGGAGGGAAAT
CTCAACAAAGAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNTGTTTTGTGTTCTCACATGATTGTTTGTTTTATTAGTGCTATCG
AAACACGTTCACATCACATCACATGATTGTTAGGCGTTAAAATAATACATCGATTACAAAGATACTGATT
CGTTTTATTTCTGAACGCGGTAAGATTTTACCACGTCG...
>scaffold_02
AATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGCGCTAAACG
GACTGGTCTACTCGTCGATTTGCTTATGAAATTGTGAAGCTTAGGTGATTACAACGAAGGTACTTACCAT
GGATTATGGGTCTGCT...
```

# *de novo* アッセンブリアルゴリズムの二大潮流

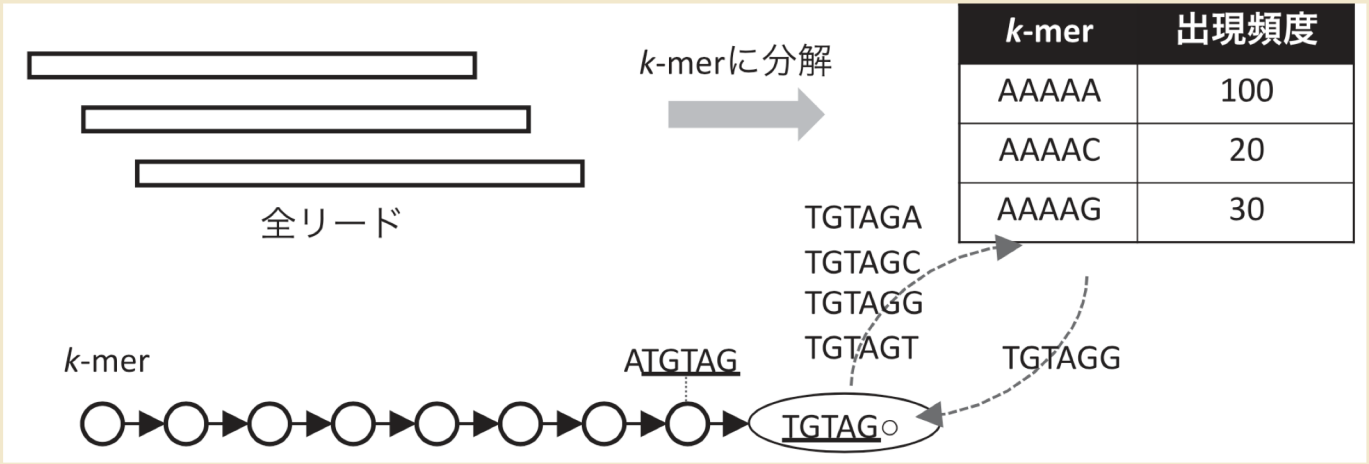
## Overlap Layout Concensus (OLC)




（例）
CAP
PHRAP
Celera Assembler

ロングリードの時代になり再評価

## de Bruijn graph (DBG)




（例）
EULER
Velvet
Trinity
Platanus

ショートリードのアセンブリに最適



図. 野口, 生物工学 95(9), 2017

# de Bruijn グラフによるアッセンブルの考え方

GCCGTAGCTACCTTTACAATA　ゲノム配列

GCCGTAGCT　AGCTACC　GCTACCTTT　CCTTTAC　CTTTACAATA　リード配列

GCCG
CCGT
CGTA
GTAG
TAGC
AGCT

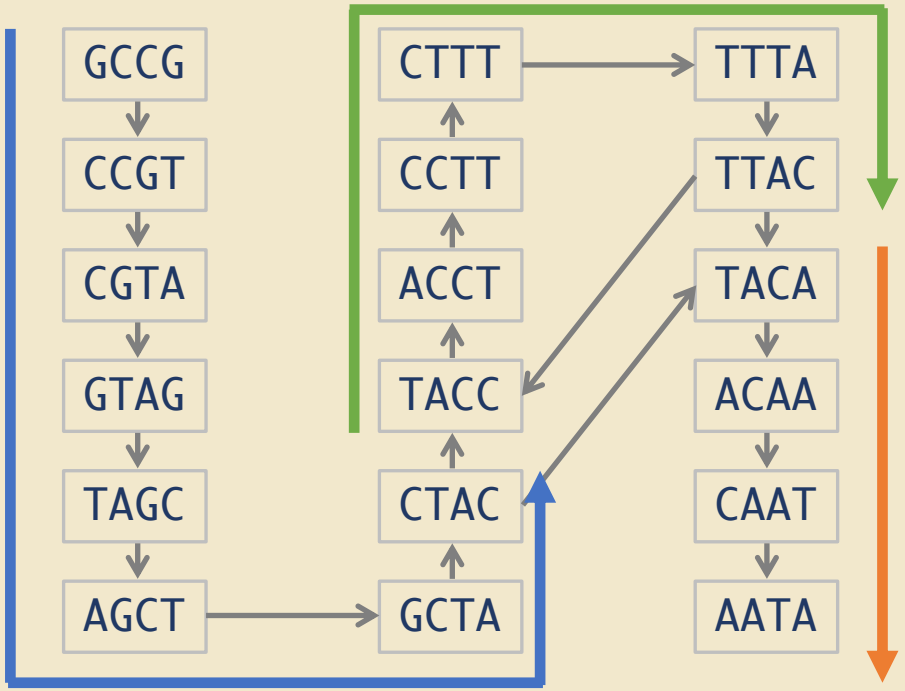
AGCT
GCTA
CTAC
TACC

GCTA
CTAC
TACC
ACCT
CCTT
CTTT

CCTT
CTTT
TTTA
TTAC

CTTT
TTTA
TTAC
TACA
ACAA
CAAT
AATA

*K-mer* に分解
(*K*=4)

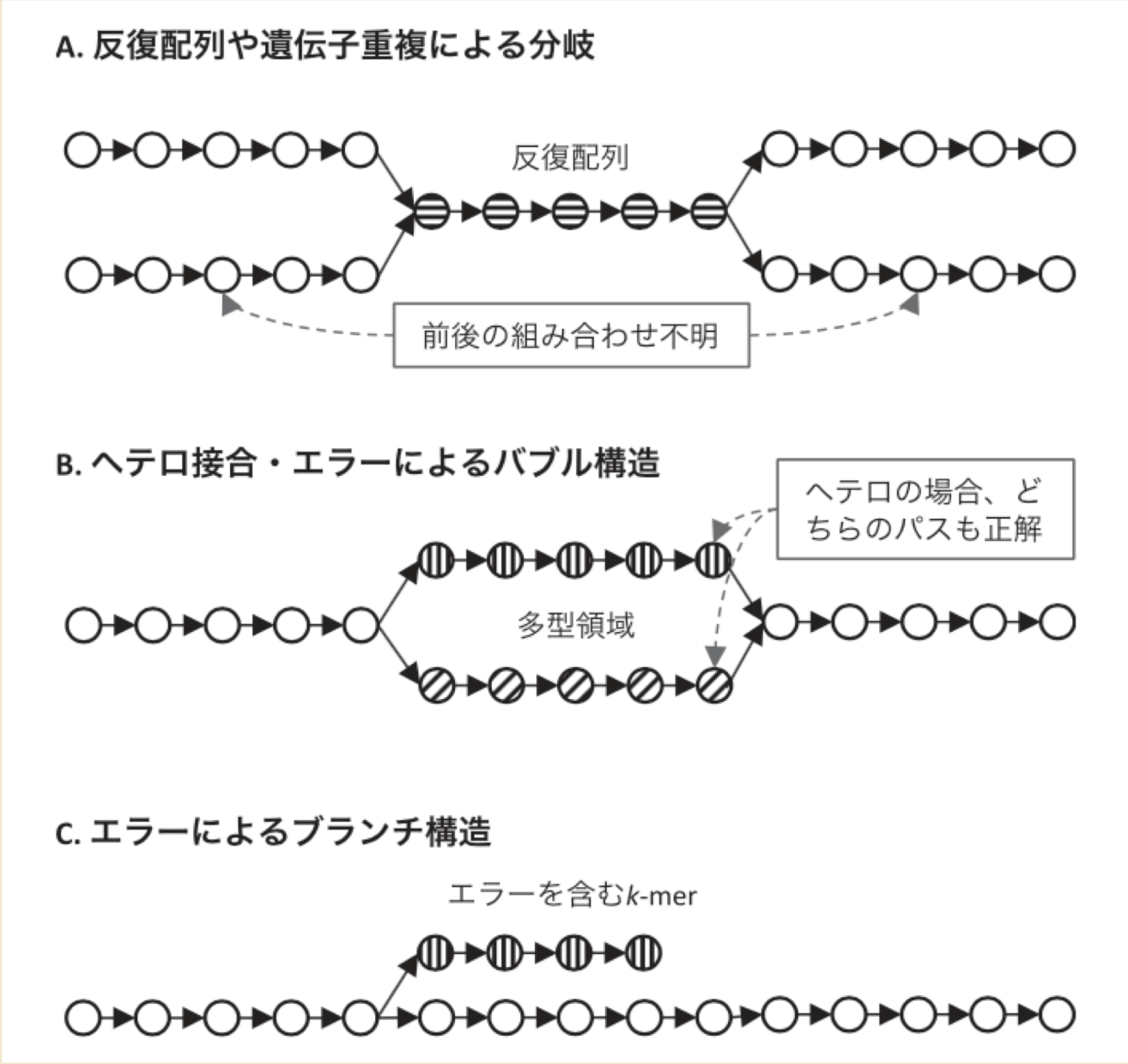



各*K-mer* を頂点

*K*-1のオーバラップを利用し、
*K-mer* 間のつながりを辺で表す

再構築されたコンティグ
GCCGTAGCTAC
TACCTTTAC
TACAATA

# グラフを複雑にさせる要因



# K-mer解析からゲノムアセンブルの難易度が予想できる

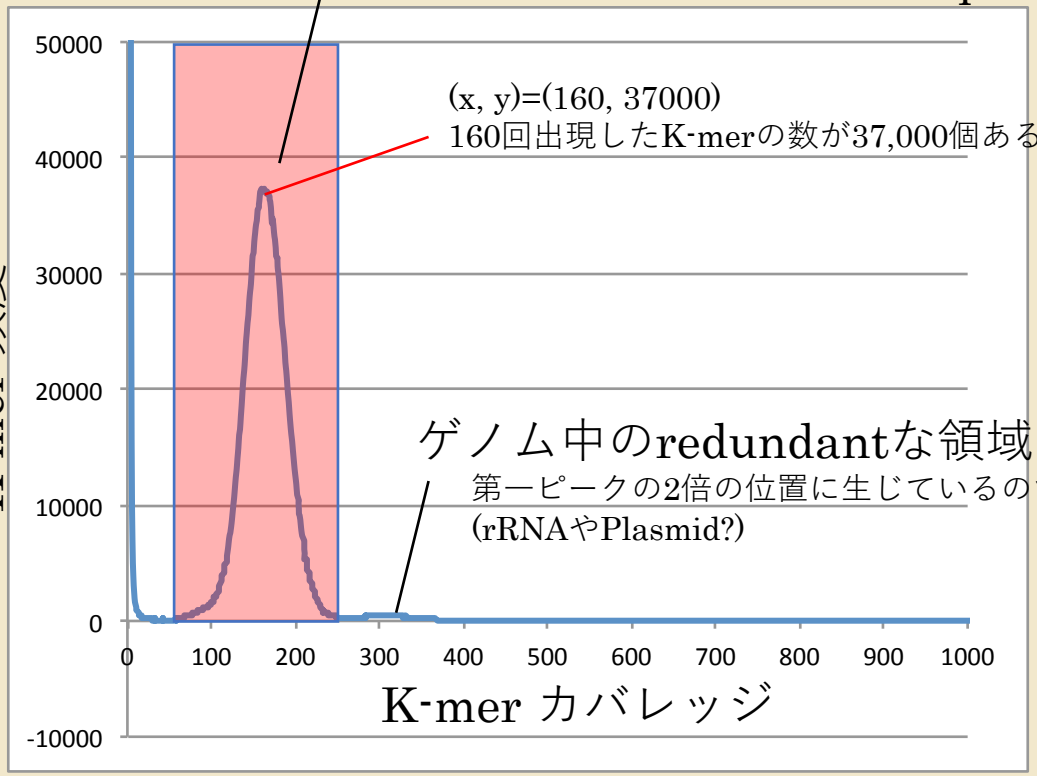
すべてのリードをK-merに分解し、出現回数 (K-merカバレッジ) を調べる

リード　CAACGAAACGAT

Kmerに分解
- CAACG
- AACGA
- ACGCA
- CGCAA
- .....

| K-mer | 出現回数 |
|---|---|
| CAACG | 20 |
| AACGA | 42 |
| ACGCA | 23 |
| CGCAA | 19 |
| ..... | ..... |

第一ピーク = ゲノム中のunique領域に由来すると考えられるK-mer




ゲノム中のredundantな領域に由来すると考えられるK-mer

第一ピークの2倍の位置に生じているので、おそらくゲノム中に2コピーある領域

(rRNAやPlasmid?)

# 出現回数が低いK-merを除くことでエラーの影響を減らせる

## エラーを含んだ K-mer の出現回数は少ない

正しいリード CAACGAAACGAT

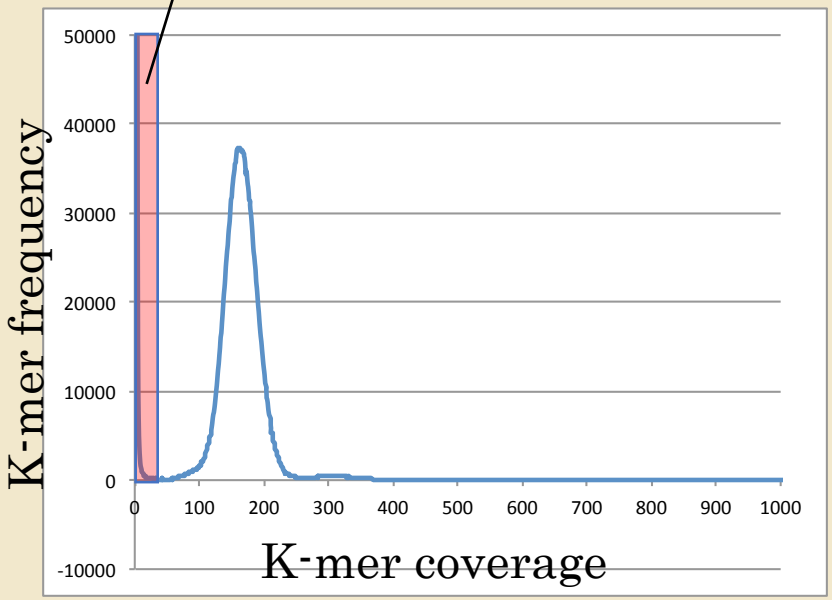
Kmerに分解
- CAACG
- AACGA
- ACGCA
- CGCAA
- .....

エラーを含んだリード CAACGTAACGAT

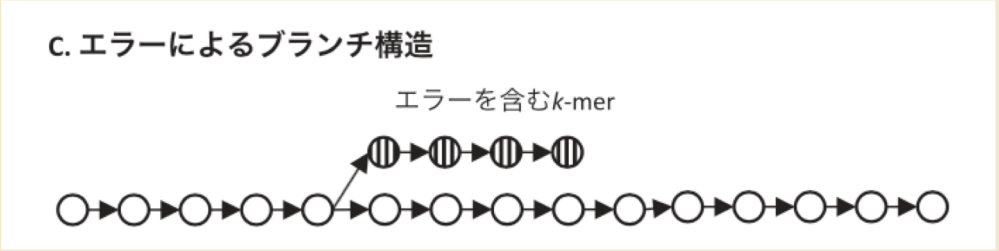
- CAACG
- AACGT
- ACGTA
- CGTAA
- .....

| K-mer | 出現回数 |
|---|---|
| CAACG | 20 |
| AACGA | 42 |
| ACGCA | 23 |
| ACGTA | 1 |
| ..... | ..... |

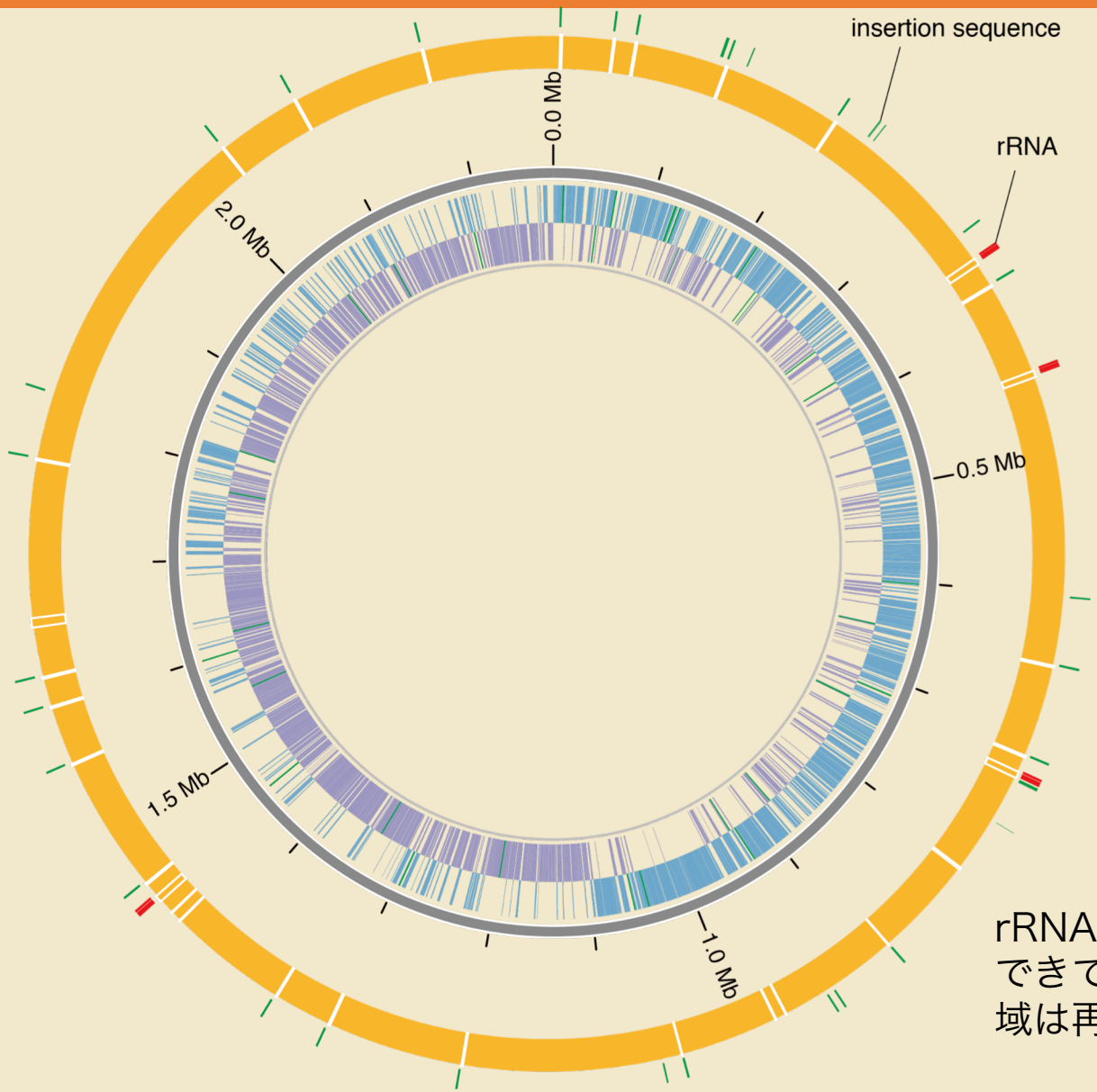
シークエンスエラー由来と考えられるK-mer




出現回数の少ない *K-mer* を取り除くことでブランチ構造を減らせる。



# ショートリードのアセンブルでは反復配列の部分で分断される




rRNAや挿入配列の部分はアセンブルできていないが、それ以外の遺伝子領域は再現できている。

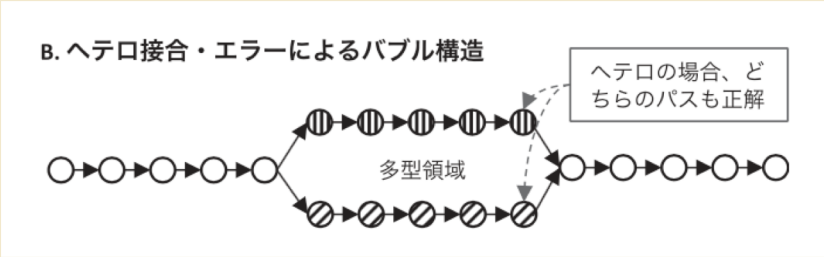
完全長のバクテリア環状染色体配列（内側）に、Illuminaショートリードから得られたコンティグ（外側）をマッピングした結果

# ヘテロ性が高いゲノムのアセンブリは難しい（真核生物の場合）

ホモ領域　　ヘテロ領域

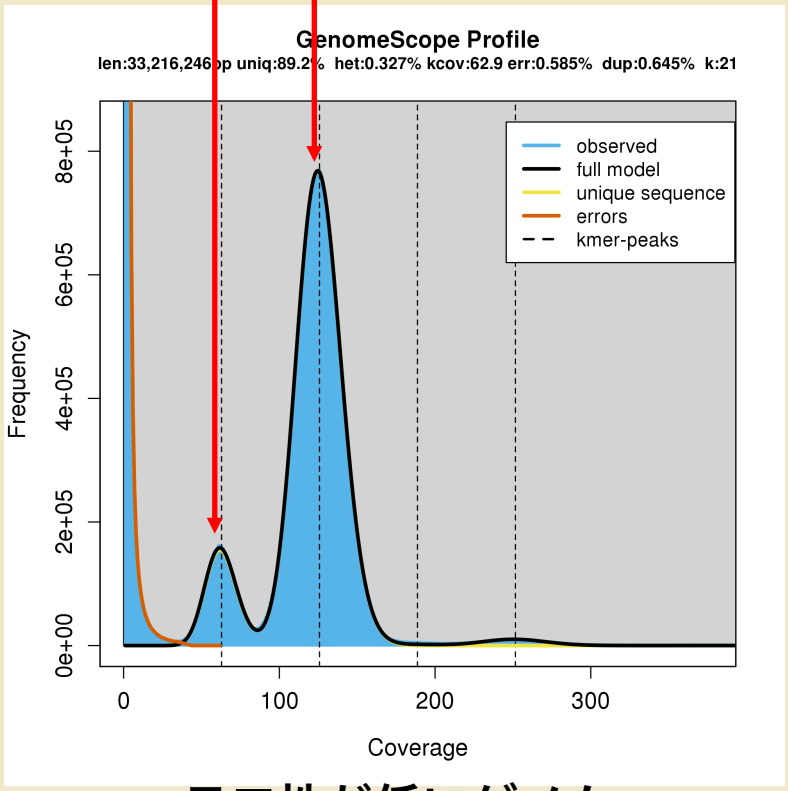
父方由来染色体 CAACGAAACGAT...GTTGCAGAAGCC

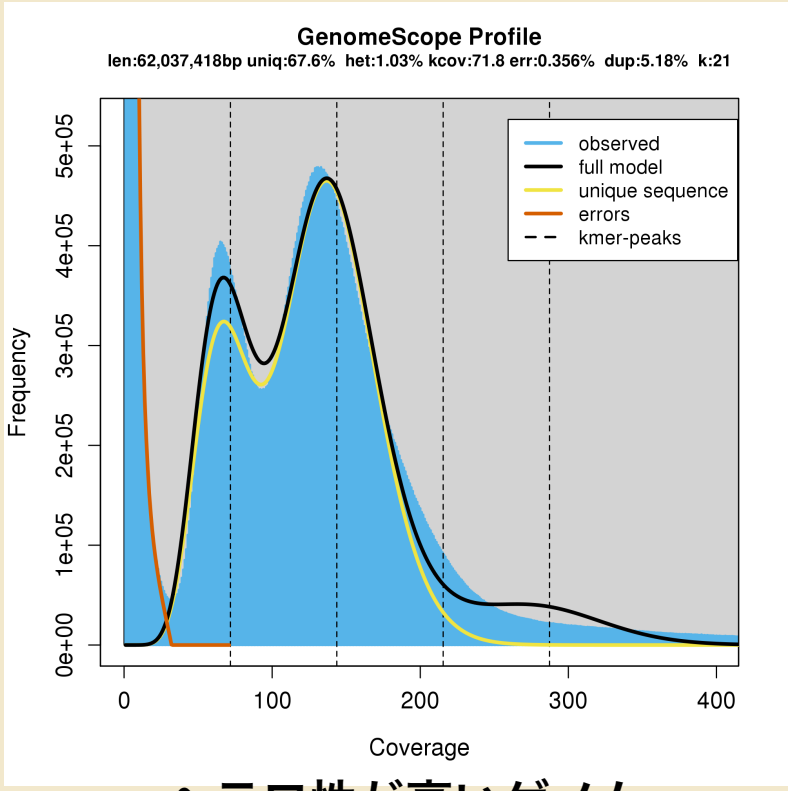
母方由来染色体 CAACGAAACGAT...GTAGCAGAAGGC




ヘテロ領域由来のK-merのピーク

ホモ領域由来のK-merのピーク




**ヘテロ性が低いゲノム**

推定ゲノムサイズ: 33Mbp

スキャフォールド数: 1,388本　N50: 96kbp

**ヘテロ性が高いゲノム**

推定ゲノムサイズ:  62Mbp

スキャフォールド数: 8,991本　N50: 16kbp

# N50：ドラフトゲノムの完成度を評価する指標

配列を長いものから順に足していったときに、全長の50%に達した時の配列の長さ
（配列長を考慮した加重平均）

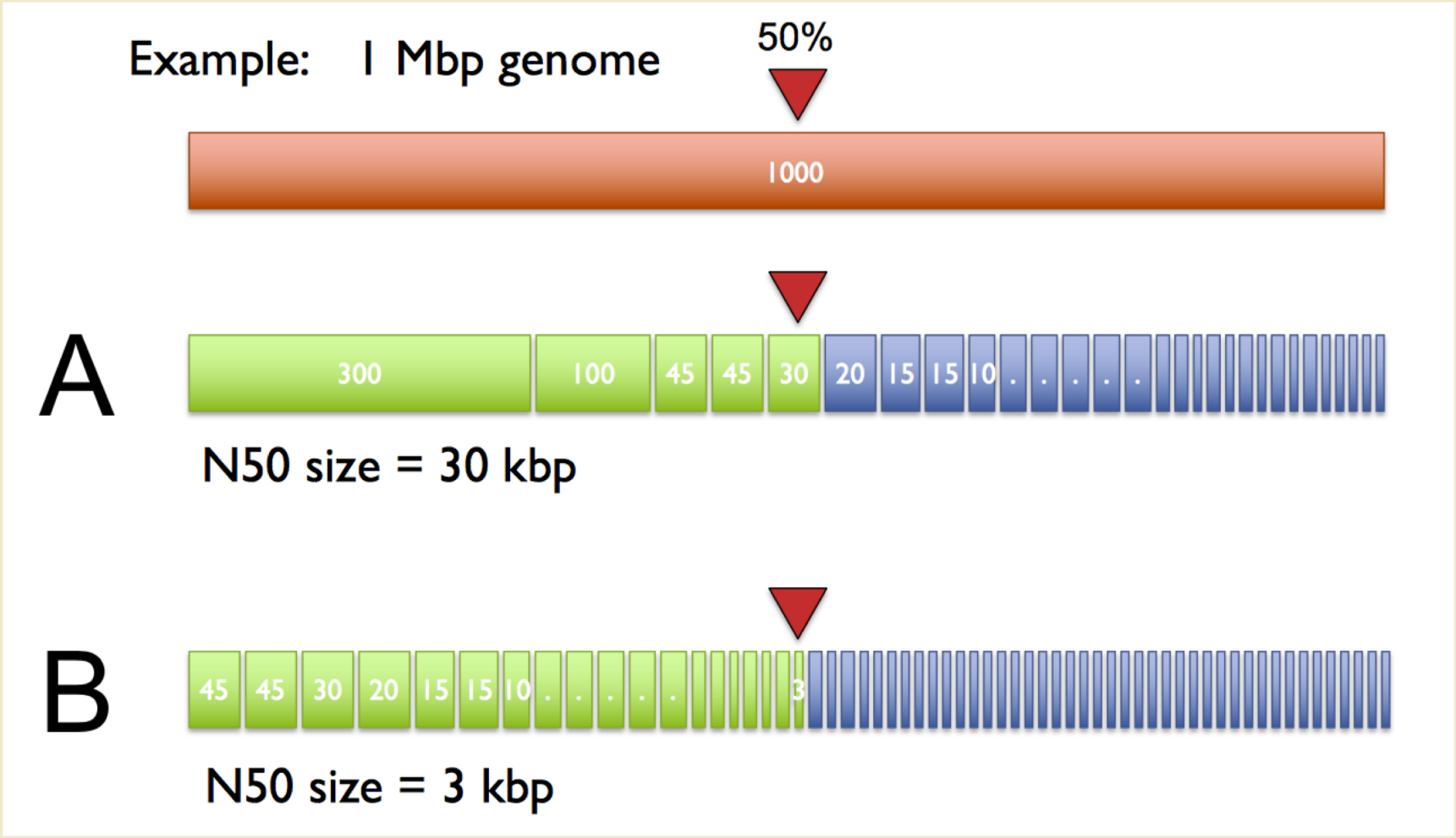




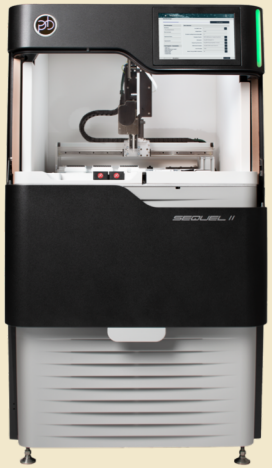

図. M. Schatz http://schatz-lab.org/appliedgenomics2018/lectures/03.GenomeAssembly.pdf

# ロングリードのアセンブルへの利用

ロングリードシークエンサー

PacBio

最大リード長 150kb以上
過半数は 50kb以上
エラー率 15%前後



Sequel II

Oxford Nanopore

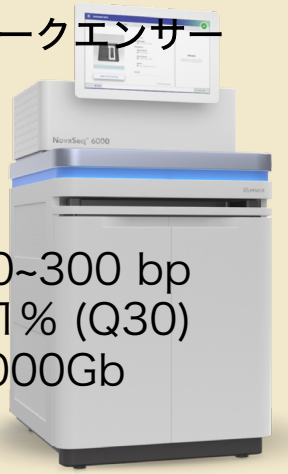
最大リード長 2Mb以上
(2018.5)
エラー率 15%前後

MinION

ショートリードシークエンサー

Illumina

リード長 100~300 bp
エラー率 ~0.1% (Q30)
データ量 ~6000Gb



NovaSeq 6000

rRNA領域の長さを十分に超えるリード長

エラー率を下げるための補正が必要

frameshiftが多すぎるという注釈付がされている

Nanoporeから得られたゲノムの例

**ASM386133v1**

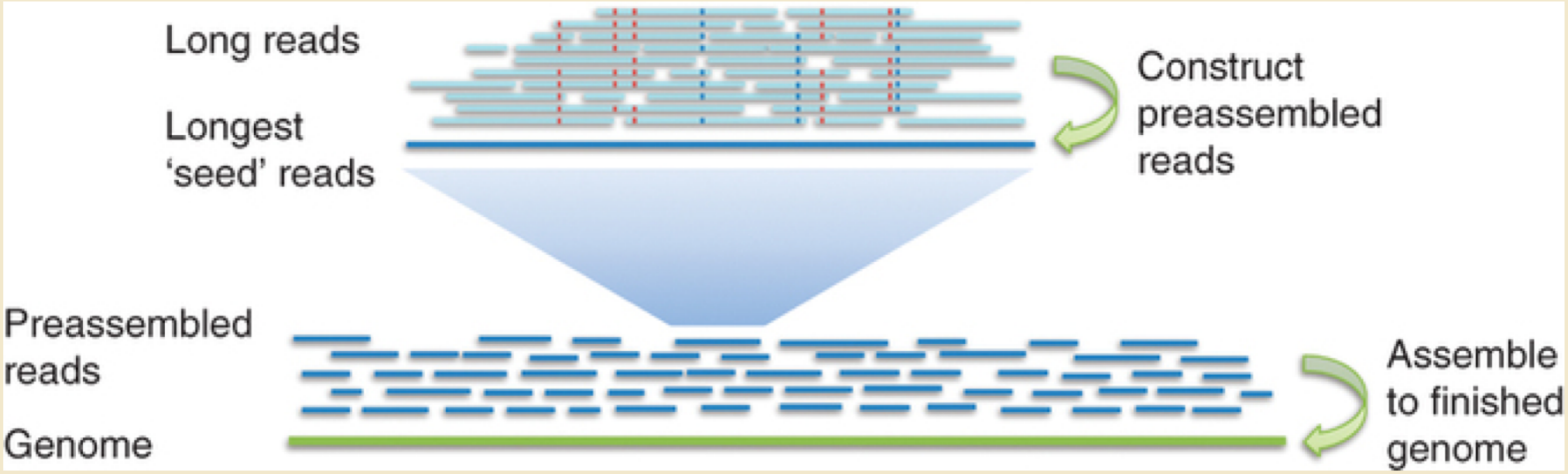
**Organism name:** Klebsiella pneumoniae (enterobacteria)
**Infraspecific name:** Strain: KLPN_9
**BioSample:** SAMN10241260
**BioProject:** PRJNA496461
**Submitter:** Johns Hopkins University
**Date:** 2018/12/04
**Assembly level:** Contig
**Genome representation:** full
**Excluded from RefSeq:**
- many frameshifted proteins

**GenBank assembly accession:** GCA_003861335.1 (latest)
**RefSeq assembly accession:** n/a
**RefSeq assembly and GenBank assembly identical:** n/a
**WGS Project:** RCZR01
**Assembly method:** Canu v. 1.6
**Expected final version:** yes
**Genome coverage:** 1153.962264x
**Sequencing technology:** Oxford Nanopore MinION

# ロングリードを利用したアセンブル方法

## HGAP法 (Hierarchical Genome-Assembly Process) による *de novo* アセンブリ

- Preassembly ：‘seed’ となるより長いリードに短いリードをマップし、エラーを補正したコンセンサス配列を得る
- Assembly：エラー補正済の’seed’ リードをアセンブルしコンティグを作成
- Consensus Polishing：得られたコンティグにリードを再マップし、さらにエラー補正




# Chin et al. *Nature Methods* (2013)

HGAP 法ではQS=60 (99.9999%の信頼性) のアセンブルが可能

この他にも Canu, Falcon, miniASM など多くのツールが利用可能になっている

# まとめ：ゲノムアセンブリ

**ショートリードによるアセンブリ**

de Bruijn graphを用いた手法が主流

rRNAや挿入配列のような反復配列の部分のアセンブリは苦手

→ バクテリアゲノムであれば、それ以外の遺伝子についてはほぼ網羅できている

**ロングリードによるアセンブリ**

Overlap layout consensus 法が主流

バクテリアゲノムであれば１本の染色体につなげることも難しくはない

エラー補正(polishing)による精度向上が望ましい

# ゲノムアノテーションおよびその表現方法

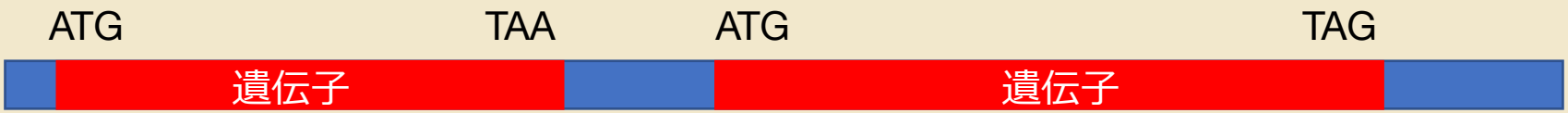
# ゲノムアノテーションとは

>Chromosome
TGGTAATATTACTGTTGATTCATCAACGAGTAGCCCCATAGGGGCAATGGCAAAAGCATACTCCCGTTAATTCGGATGT
ATAAATATTAAGTCGAATAAAAGGTATCTAGGAAAACTTGTGAGTACACGTGAAAAACGTCTGCTCTCCTTGCTCTTTT
TAAATGAAAAAGAGCCAAAGTCCATAAGGAGGTGTAACAGTTAATGGAACCAAAACGTTATGAAATTACGTACATCATT
CGTCCTGACATGGATGAAGCTGCTAAAACAGCGCTTGTTGAACGATTTGACAAGATTGTGTCAGATAATGGTGCTACGA
TCGTTGATTCGAAAGACTGGTCTACTCGTCGATTTGCTTATGAAATTGGTGATTACAACGAAGGTACTTACCATATCGT
TAATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGACGATATCTTGCGTCAC
ATGATTGTTAAGCGTGAAGCTTAATCTAATCAATTTAAAGTTAAGAAAGGAGTATTAGAATCAAACGTGCTCGGATTAT
GGGTCTGCTACCATTCGTTGCAGAAGACTAATTTGAAATTGTCCATATTGTATCTCTCGAGCCAATTAAATCAATTAGG
AAACTGCCAGAGGAGGGAAATTCAATGGCTCAACAAAGAAGAGGCGGACATCGTCGCCGTAAGGTTGACTTTATTGCCG
TTCACAGATTTAAGACACATACTTTTTGTTTTGTGTTCTTGTTTTATTAGTGCTATCGTGTTATAATTTTTGCTTACCG
AAAAACACGTTCACATCACATAGGCGTTAAAATAATACATCGATTACAAAGATACTGATTTACTAAAACGTTTTATTTC
TGAACGCGGTAAGATTTTACCACGTCGATTTAATGTAAATGTTTATTTAAATCCTAATTATGCCATGATTGTGGTGTGA
TTAGGTCTCGTCCCGTAAGGTAAGAACATTAACAATATCACCCACTATATGATTAATCGTACAATTCTTGTTGGACGCT
TAACTAGAGATCCTGAGTTGCGATACACAACTAGTGGAGCTGCTGTAGCAACGTTTACCGTTGCTGTCAATCGGCAGTT
TACCAATCAACAGGGTGAACGGGAAGCTGATTTTATTAGCTGCGTCATTTGGCGTAAAGCTGCTGAAAATTTTTCCAAT
TTCACTCATAAGGGTTCTTTGGTTGGGGTTGATGGCCGCATTCAAACGCGAAATTATGAAAATCAACAGGGTCAACGTG
TTTATGTAACGGAAGTAGTAGTTGAAAACTTCTCGTTACTAGAAACGAAAGCCCAAAGTCAAAACCATAATAATGGTGC
CCCAAGCTTTGACAATAATCAACAAGCCAATGCTCCTCAATCATCATCAGCAAATGATAATCCGTTTGGTAATGCTAAT
GACAATGCAAATGCGGGAAGTAGTAGTGCTAACAGCAATGCTAACGATCCATTCGCTAATAATGGCGAACCAATCGACA
TTTCAGATGACGATTTGCCGTTCTAACAAAGTTAGTGGAACAAGTGCTAAAAACCAGCGTCGTTTAACAATTGCAATCA
AACGTGCTCGGATTATGGGTCTGCTACCATTCGTTGCAGAAGACTAATTTGAAATTGTTTTAAT...

# ゲノムアノテーションとは

構造アノテーション (遺伝子領域の推定) ＋ 機能アノテーション (遺伝子機能の推定)

# 遺伝子領域の予測（バクテリアの場合。真核生物はもっと複雑）




- 開始コドン ATG で始まり、終始コドン (TAA, TAG, TGA) で終わる
- 長さは３の倍数

TGGTAATATTACTGTTGATTCATCAACGAGTAGCCCCATAGGGGCAATGGCAAAAGCATACTCCCGTTATTTCGGATGT
ATAAATATTAAGTCGAATAAAAGGTAACTAGGAAAACTTGTGAGTACACGTGAAAAACGTCTGCTCTCCTTGCTCTTTT
TAAATGAAAAAGAGCCAAAGTCCATAAGGAGGTGTAACAGTTAATGGAACCAAAACGTTATGAAATTACGTACATCATT
CGTCCTGACATGGATGAAGCTGCTAAAACAGCGCTTGTTGAACGATTTGACAAGATTGTGTCAGATAATGGTGCTACGA
TCGTTGATTCGAAAGACTGGTCTACTCGTCGATTTGCTTATGAAATTGGTGATTACAACGAAGGTACTTACCATATCGT
TAATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGACGATATCTTGCGTCAC
ATGATTGTTAAGCGTGAAGCTTAATCTAATCAATTTAAAGTTAAGAAAGGAGTATTAGAATCAAACGTGCTCGGATTAT
GGGTCTGCTACCATTCGTTGCAGAAGACTAATTTGAAATTGTCCATATTGTATCTCTCGAGCCAATTAAATCAATTAGG
AAACTGCCAGAGGAGGGAAATTCAATGGCT

正しい遺伝子とそうでないものを見分ける必要がある

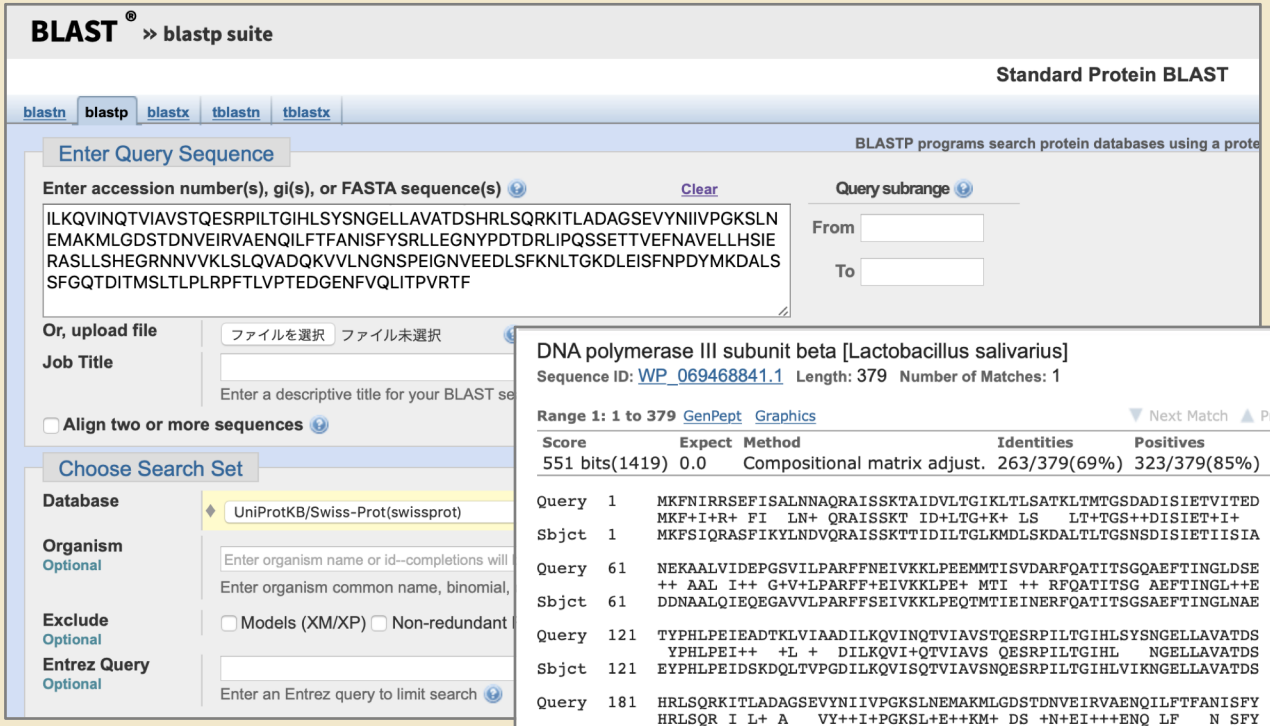
遺伝子らしさの推定
（既知の遺伝子領域の特徴を学習。学習結果から予測）

相同性検索

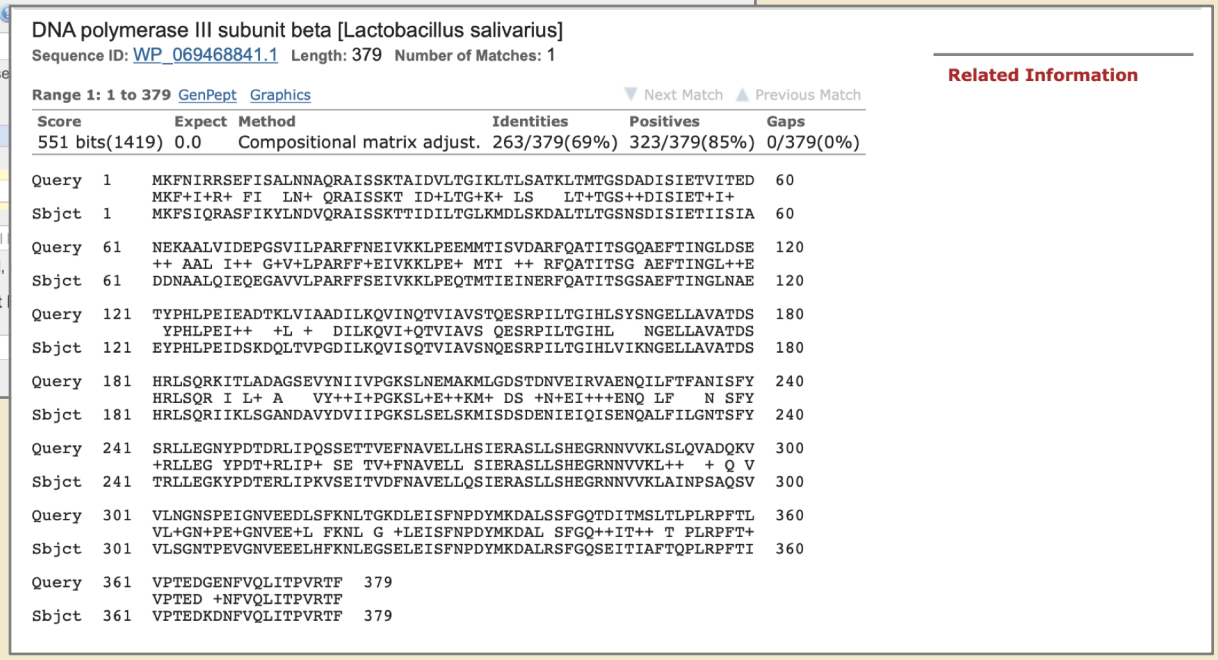
# 相同性検索で遺伝子機能を予測する

配列が類似 ≒ 機能が類似

塩基配列やタンパク質配列を入力として、配列データベースに対して検索する



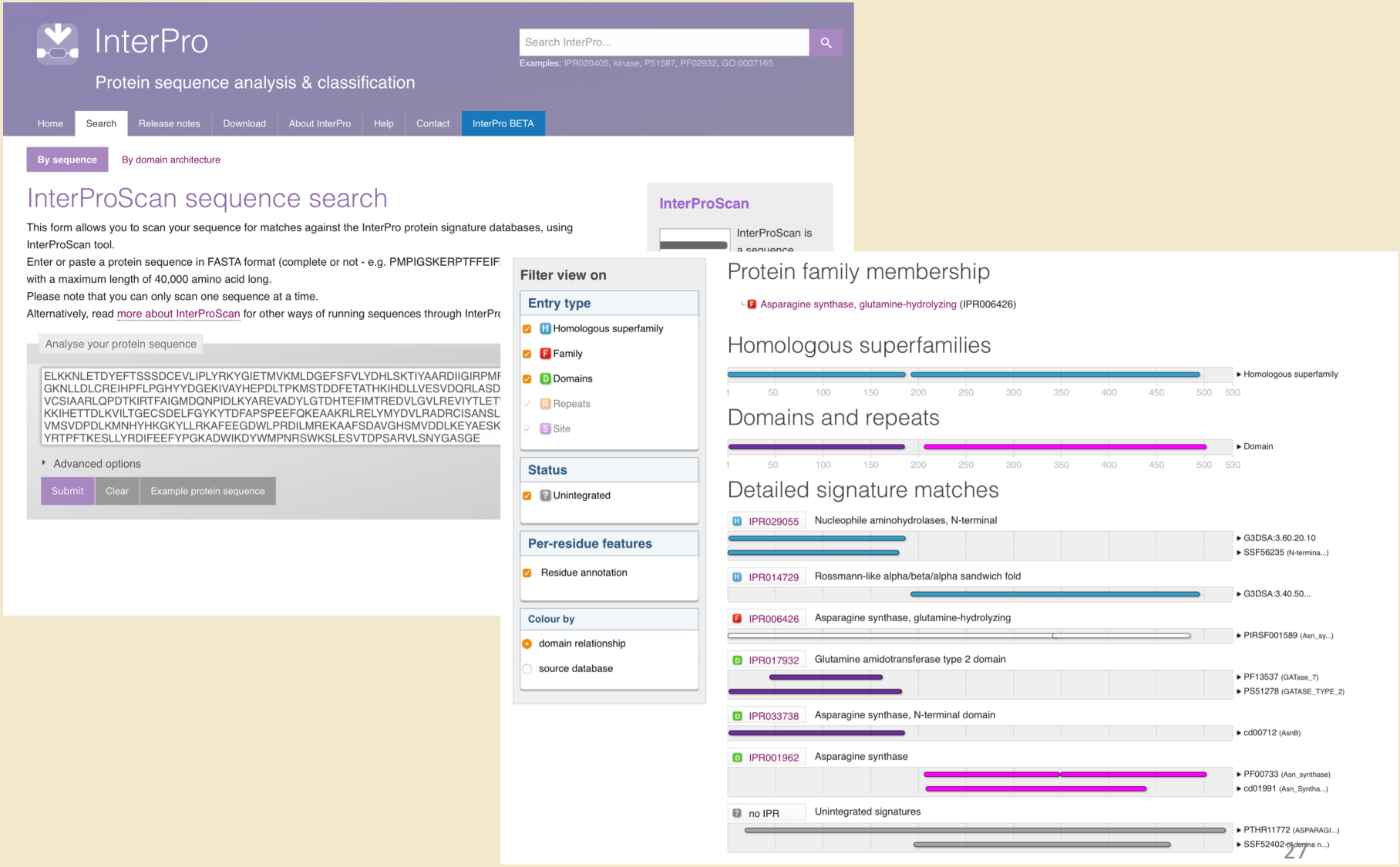

検索結果




信頼できる参照データベースを利用することが大事

# タンパク質の機能にかかわる部分配列（モチーフやドメイン）の検索

InterPro: さまざまなタンパク質機能探索のための統合データベース



# アノテーションされたゲノムの一般的な書式 (GenBank形式)

## GenBankに登録された大腸菌 K-12株染色体のエントリー

メタデータ
登録者情報、文献情報など

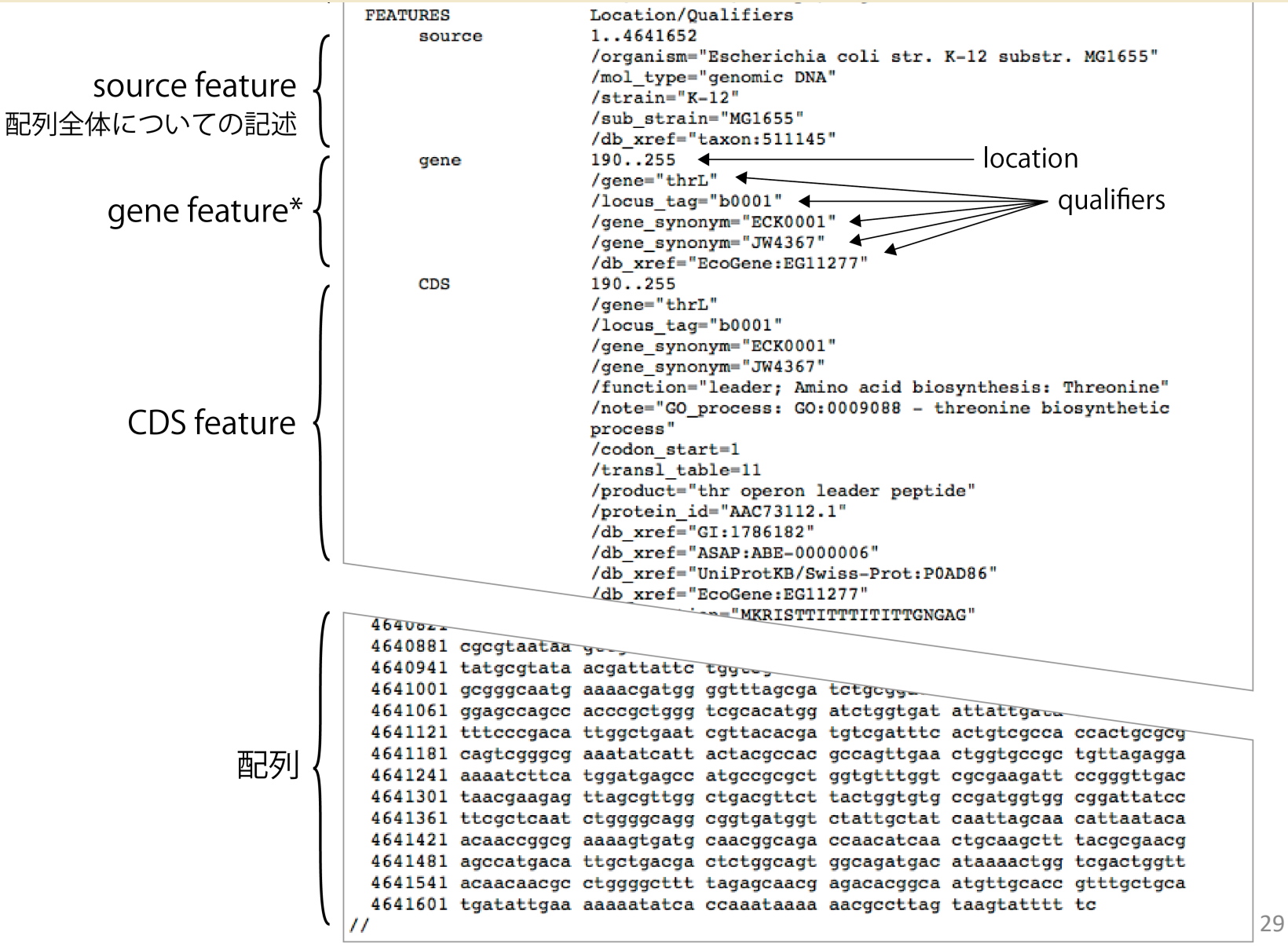
source feature
配列全体についての記述

```
LOCUS       U00096               4641652 bp    DNA     circular BCT 01-AUG-2014
DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION   U00096
VERSION     U00096.3
DBLINK      BioProject: PRJNA225
            BioSample: SAMN02604091
KEYWORDS    .
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4641652)
  AUTHORS   Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
            Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
            Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
            Mau,B. and Shao,Y.
  TITLE     The complete genome sequence of Escherichia coli K-12
  JOURNAL   Science 277 (5331), 1453-1462 (1997)
   PUBMED   9278503
REFERENCE   2  (bases 1 to 4641652)
  AUTHORS   Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
            Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
  TITLE     Highly accurate genome sequences of Escherichia coli K-12 strains
            MG1655 and W3110
            Syst. Biol. 2, 2006 (2006)
  JOURNAL
            Wisconsin,                                    Blattner,F.R.,
  REMARK    Protein update by submitte                                Kosuge,T.,
COMMENT     On Sep 26, 2013 this sequence version
            Current U00096 annotation updates are derived from
            http://ecogene.org. Suggestions for updates can be sent to Dr.
            Kenneth Rudd (krudd@miami.edu). These updates are being generated
            from a collaboration that also includes ASAP/ERIC, the Coli Genetic
            Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.
FEATURES             Location/Qualifiers
     source          1..4641652
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene            190..255
                     /gene="thrL"
```

location

# アノテーションされたゲノムの一般的な書式 (GenBank形式)

## GenBankに登録された大腸菌 K-12株染色体のエントリー（続き）



# GenBank形式は Entry – Feature – Qualifier の階層構造

| | | | |
|---|---|---|---|
| Genome | Entry | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |
| | | | Qualifier /key=value |
| | | | Qualifier /key=value |
| | | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |
| | | | Qualifier /key=value |
| | | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |
| | Entry | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |
| | | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |
| | | | Qualifier /key=value |
| | Entry | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |
| | | Feature | Location xxx…yyy |
| | | | Qualifier /key=value |

ドラフトゲノムなら１ゲノムは複数エントリーから構成される。コンプリートゲノムであれば１エントリーで１レプリコン（染色体 or プラスミド）の配列を構成する

# プログラミングによってGenBankファイルから情報を得る

Biopythonを使った例

```
from Bio import SeqIO

for entry in SeqIO.parse(file_name, "genbank"):
    for feature in entry.features:
        print(feature.type, feature.location)
        for key, value in feature.qualifiers.items():
            print("\t", key, value)
```

```
source [0:4641652](+)
        organism ['Escherichia coli str. K-12 substr. MG1655']
        mol_type ['genomic DNA']
        strain ['K-12']
        sub_strain ['MG1655']
        db_xref ['taxon:511145']
gene [189:255](+)
        gene ['thrL']
        locus_tag ['b0001']
        gene_synonym ['ECK0001']
        db_xref ['ASAP:ABE-0000006', 'ECOCYC:EG11277', 'EcoGene:EG11277']
CDS [189:255](+)
        gene ['thrL']
        locus_tag ['b0001']
        gene_synonym ['ECK0001']
        codon_start ['1']
        transl_table ['11']
        product ['thr operon leader peptide']
        protein_id ['AAC73112.1']
        db_xref ['UniProtKB/Swiss-Prot:P0AD86', 'ASAP:ABE-0000006', 'ECOCYC:EG11277', …
        translation ['MKRISTTITTTITITTGNGAG']
gene [336:2799](+)
        gene ['thrA']
        …
```

# DDBJ 登録フォーマットも Entry – Feature – Qualifier の階層構造

## DDBJ-MSS (大規模登録システム）用の登録ファイル

| Entry | Feature | Location | Qualifier Key | Qualifier Value |
|---|---|---|---|---|
| contig01 | source | 1..2038844 | organism | Lactobacillus delbrueckii subsp. lactis |
| | | | strain | DSM 20072 |
| | | | mol_type | genomic DNA |
| | | | submitter_seqid | @@[entry]@@ |
| | | | ff_definition | @@[organism]@@ @@[strain]@@ DNA, @@[submitter_seqid]@@ |
| | | | culture_collection | DSM:20072 |
| | CDS | complement(30..1622) | product | asparagine synthase |
| | | | transl_table | 11 |
| | | | codon_start | 1 |
| | | | locus_tag | LDL20072_00010 |
| | | | gene | asnB |
| | | | inference | similar to AA sequence:INSD:ADQ60884.1 |
| | tRNA | 2156..2230 | product | tRNA-Asn |
| | | | locus_tag | LDL20072_t00010 |
| | | | inference | COORDINATES:profile:Aragorn:1.2.38 |
| contig02 | CDS | 4470..5972 | product | ABC transporter substrate-binding protein |
| | | | transl_table | 11 |
| | | | codon_start | 1 |
| | | | locus_tag | LDL20072_00020 |
| | tRNA | 7221..7345 | product | tRNA-Asn |
| | | | locus_tag | LDL20072_t00100 |
| | | | inference | COORDINATES:profile:Aragorn:1.2.38 |
| contig03 | CDS | 1280..2900 | product | hypothetical protein |
| | | | transl_table | 11 |
| | | | codon_start | 1 |
| | | | locus_tag | LDL20072_00660 |

# GFF形式は複雑な遺伝子構造の記述に便利

ID – Parent を使って階層構造を表現

真核生物ゲノムの例

```
#gff-format3
chr1  feature  gene            825832  830976  .  -  .  ID=MP076;Name=MP076
chr1  feature  mRNA            825832  830976  .  -  .  ID=MP076.1;Name=MPALOM;Parent=MP076
chr1  feature  CDS             830806  830976  .  -  0  Parent=MP076.1
chr1  feature  CDS             826928  827188  .  -  0  Parent=MP076.1
chr1  feature  three_prime_UTR 826846  826927  .  -  .  Parent=MP076.1
chr1  feature  three_prime_UTR 825832  826334  .  -  .  Parent=MP076.1
chr1  feature  exon            830806  830976  .  -  .  Parent=MP076.1
chr1  feature  exon            826846  827188  .  -  .  Parent=MP076.1
chr1  feature  exon            825832  826334  .  -  .  Parent=MP076.1
##
chr1  feature  gene            833612  837015  .  -  .  ID=MP077;Name=MPkDE
chr1  feature  mRNA            833613  837015  .  -  .  ID=MP077.1;Name=MPkDE.1;Parent=MP077
chr1  feature  five_prime_UTR  835166  837015  .  -  .  Parent=MP077.1
chr1  feature  CDS             834852  835165  .  -  0  Parent=MP077.1
chr1  feature  CDS             834016  834118  .  -  0  Parent=MP077.1
chr1  feature  three_prime_UTR 833613  834015  .  -  .  Parent=MP077.1
chr1  feature  exon            834852  837015  .  -  .  Parent=MP077.1
chr1  feature  exon            833613  834118  .  -  .  Parent=MP077.1
chr1  feature  mRNA            833612  837015  .  -  .  ID=MP077.2;Name=MPkDE.2;Parent=MP077
chr1  feature  five_prime_UTR  835598  837015  .  -  .  Parent=MP077.2
chr1  feature  five_prime_UTR  835166  835318  .  -  .  Parent=MP077.2
chr1  feature  CDS             834785  835165  .  -  0  Parent=MP077.2
chr1  feature  three_prime_UTR 833612  834784  .  -  .  Parent=MP077.2
chr1  feature  exon            835598  837015  .  -  .  Parent=MP077.2
chr1  feature  exon            833612  835318  .  -  .  Parent=MP077.2
##
chr1  feature  gene            841118  841351  .  -  .  ID=MP078;Name=MP078
chr1  feature  mRNA            841118  841351  .  -  .  ID=MP078.1;1;Name=MP078.1;Parent=MP078
chr1  feature  CDS             841118  841351  .  -  0  Parent=MP078.1
chr1  feature  exon            841118  841351  .  -  .  Parent=MP078.1
```

# 原核生物のゲノムアノテーションはほぼ自働化されている

## Prokka

**Prokka: rapid prokaryotic genome annotation** FREE

Torsten Seemann

*Bioinformatics*, Volume 30, Issue 14, 15 July 2014, Pages 2068–2069,
https://doi.org/10.1093/bioinformatics/btu153
**Published:** 18 March 2014 **Article history** ▾

コマンドラインツール、軽量・高速

## PGAP

**NCBI prokaryotic genome annotation pipeline** FREE

Tatiana Tatusova, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky ✉, James Ostell

*Nucleic Acids Research*, Volume 44, Issue 14, 19 August 2016, Pages 6614–6624,
https://doi.org/10.1093/nar/gkw569
**Published:** 24 June 2016 **Article history** ▾

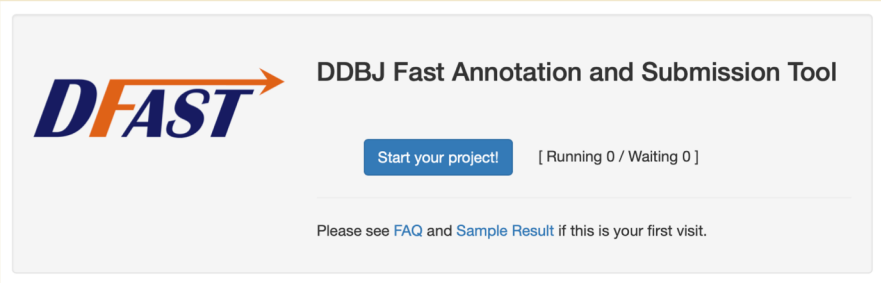
GenBankの登録システムに統合化

## RAST


http://rast.theseed.org

ウェブツール。SEEDというゲノムアノテーション・比較ゲノムのプラットフォームをベースにしている

## DFAST




https://dfast.nig.ac.jp

ウェブ/コマンドライン。DDBJの登録ファイルを自動生成可能

# 例外的な遺伝子アノテーションの例

## アミノ酸への翻訳ルールの例外
## (UGA -> セレノシステイン、UAG -> ピロリジン)

```
CDS             2026249..2029296
                /gene="fdnG"
                /note="codon on position 196 is selenocysteine opal codon"
                /transl_except=(pos:2026834..2026836,aa:Sec)
                /product="formate dehydrogenase-N, alpha subunit,
                nitrate-inducible"
                /translation="MDVSRRQFFKICAGGMAGTTVAALGFAPKQALAQARNYKLLRAK
                EIRNTCTYCSVGCGLLMYSLGDGAKNAREAIYHIEGDPDHPVSRGALCPKGA...
```

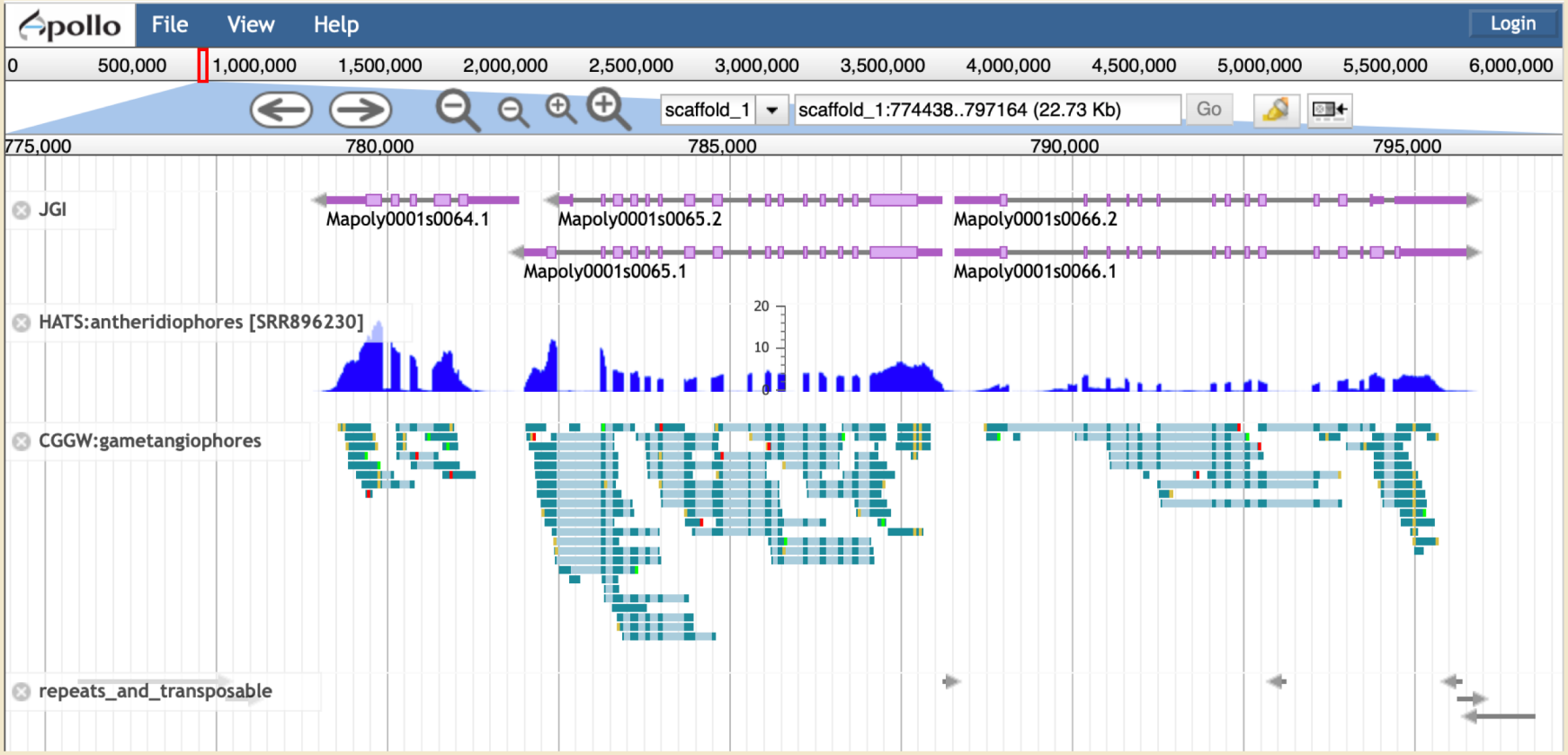
大腸菌 O26:H11 (accession no: GCF_000091005.1) の例

## Ribosomal slippage によるコドン読み枠のずれ

```
CDS             join(266..13468,13468..21555)
                /gene="ORF1ab"
                /ribosomal_slippage
                /note="pp1ab; translated by -1 ribosomal frameshift"
                /product="ORF1ab polyprotein"
                /translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
                HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELE...
```

SARS-Cov-2 reference genome (accession no: NC_045512.2) の例

# 真核生物の遺伝子予測は難しい

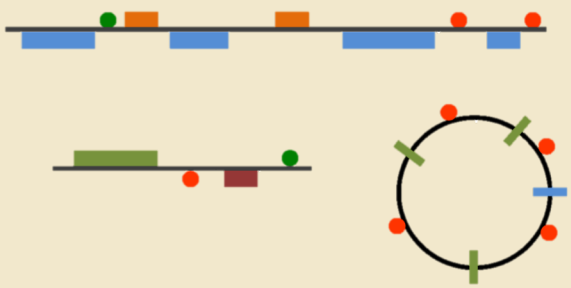



複雑なエクソン・イントロン構造
アイソフォーム
　（同一遺伝子座からスプライシングパターンの異なる複数の転写産物）

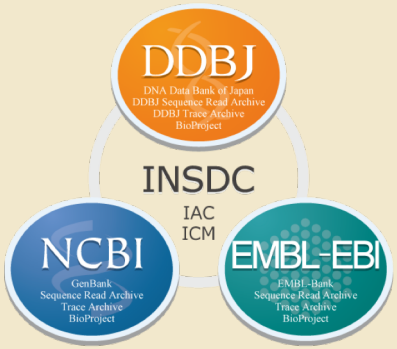
**RNA-seqなどの発現データを組み合わせて遺伝子領域の予測を行う**

# What's DFAST?

## DDBJ Fast Annotation and Submission Tool



DFAST



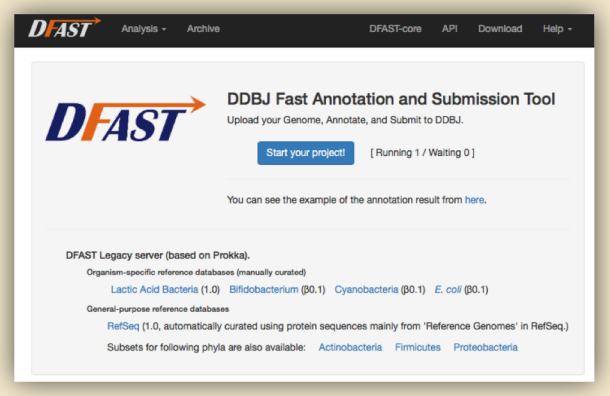

Prokaryotic genome annotation

Data submission to DDBJ

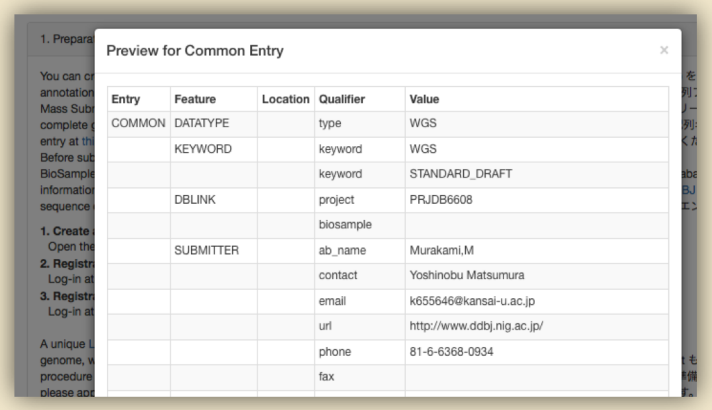
Fast, flexible, and powerful

# Friendly for both beginners and experts

## Graphical user interface for beginners

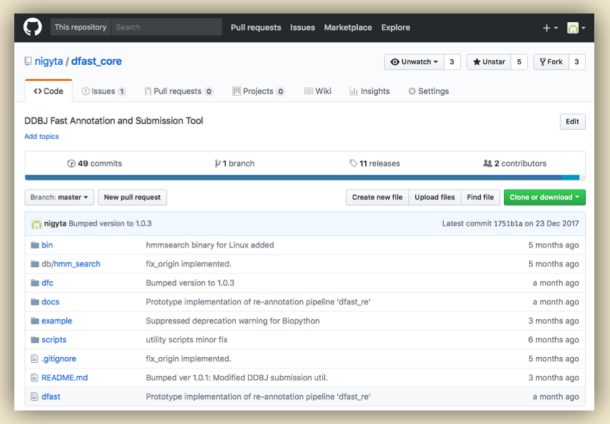



https://dfast.nig.ac.jp




Create DDBJ submission file using online editor

## Command operations for experts




Sample usage

```
dfast --genome your_genome.fna
      --config sample.cfg
```

Stand-alone version available for download

# Release history

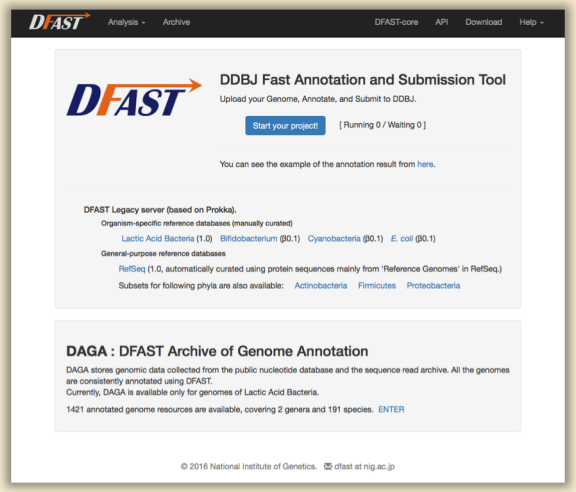
**V 0.1 (2016.1)**

**Web service to assist DDBJ submission**

Tanizawa et al., BMFH (2016)

- Based on the lightweight annotation pipeline *Prokka* (Seemann, 2014)
- GUI to edit annotation and metadata
- Curated reference database for specific organism groups

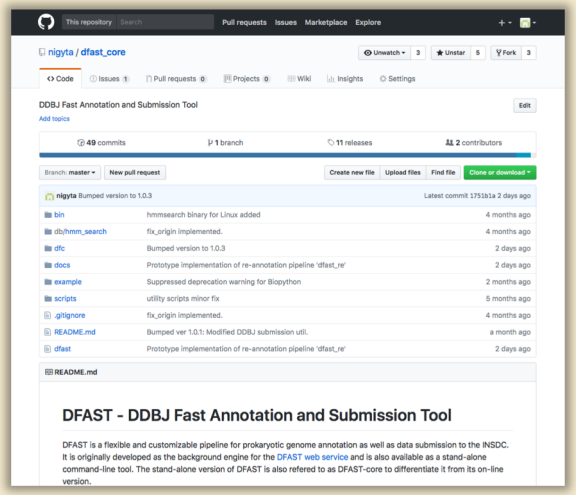
  Lactic acid bacteria, ecoli, cyanobacteria … etc..




Web version (https://dfast.nig.ac.jp)

**General-purpose reference database for all kinds of prokaryote**

**Original annotation engine *DFAST-core***

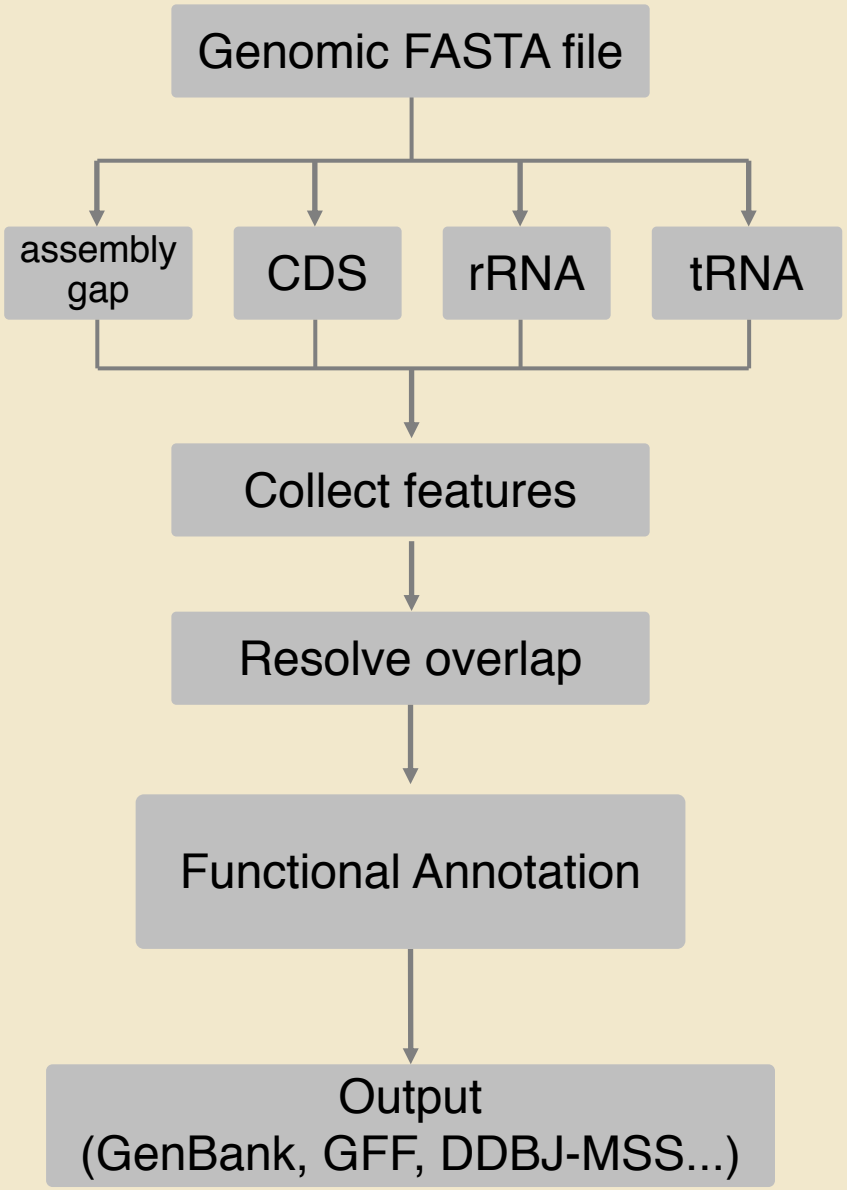



**V 1.0 (2017.8) First major release**

Tanizawa et al., Bioinformatics (2018)

Stand-alone version
Available at GitHub:nigyta/dfast_core

# How does it work?




## Structural annotation phase

*de facto* standard gene prediction tools

parallel processing

## Functional annotation phase

Ultrafast homology search using GHOSTX (Suzuki et al. 2014)

- 10 times faster

Small, but well-curated references

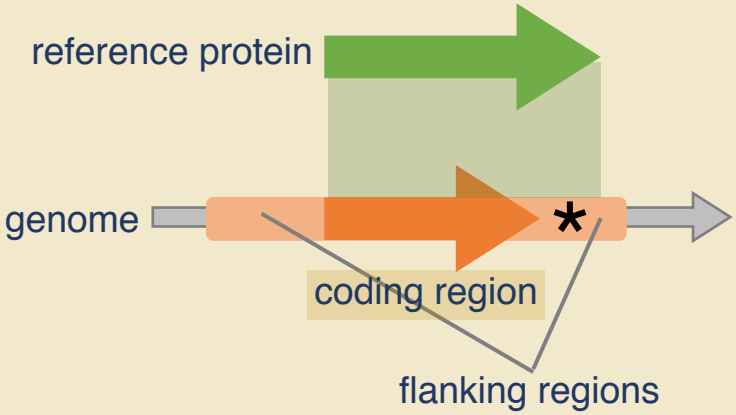
- Default database constructed from 120 representative genomes
- Optional organism-specific database

Pseudogene detection

## Flexible and customizable

# How to annotate pseudogenes

## Pseudogene: Premature stop codon / frameshift mutation




Coding region with its flanking regions

↓ Aligned using LAST (Frith et al. 2017)

Reference protein sequence

Example) *Lactobacillus parakefiri* JCM 8573, gluconate transporter

```
REF    AQKMVPDAFTGKPHLPLSSNKRQFKVSEAPGFGLSVLTALFPVILMTITTVY-ELVVDHGVTPKNPSTLDQII
QUERY  ARKFAPAAFERKGNLSSIGEVKQFTPEESPSFGLSVLTALFPVLLLSIATIY/QMTVNGGVDPKNPSVLDSII
```

coding region　　flanking region

**Description**

```
/note="Partial hit; WP_003643223.1 gluconate permease
(Lactobacillus plantarum WCFS1) [pid:71.3%, q_cov:100.0%,
s_cov:44.8%, Eval:6.5e-78]"
/note="frameshifted; deletion at around 14464"
/product="hypothetical protein"
```

# Translational exceptions

Under special conditions,

**UGA → selenocysteine (U), UAG → pyrrolysine (O)**

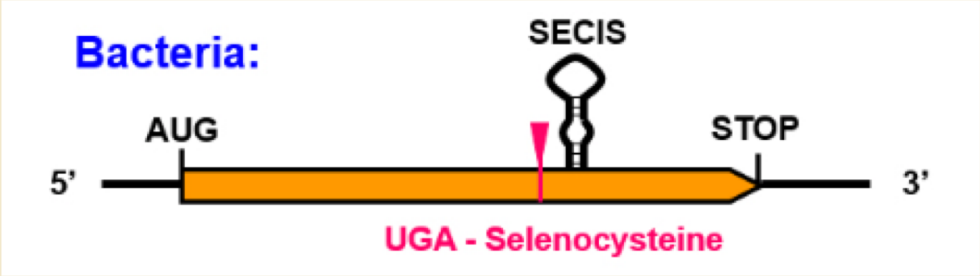



Figure from http://genomics.unl.edu/RBC_EDU/sp.html

Example) *Escherichia coli* O26, formate dehydrogenase alpha subunit

```
SBJCT   TGMLCASGASNETGMLTQKFARSLGMLAVDNQARVUHGPTVASLAPTFGRGAMTNHWVDIKNANVVMV
QUERY   TGMLCASGASNETGMLTQKFARSLGMLAVDNQARV*HGPTVASLAPTFGRGAMTNHWVDIKNANVVMV
```

**Description**

```
/inference="DESCRIPTION:similar to AA
sequence:RefSeq:NP_310105.1"
/transl_except=(pos:2026834..2026836,aa:Sec)
/note="codon on position 196 is selenocysteine opal codon."
/note="NP_310105.1 nitrate-inducible formate
dehydrogenase-N alpha subunit (Escherichia coli O157:H7
str. Sakai) [pid:99.8%, q_cov:100.0%, s_cov:100.0%, Eval:0.0e+00]"
```

# まとめ：ゲノムアノテーション

**構造アノテーション + 機能アノテーション**

**バクテリアゲノムのアノテーションはほぼ自動化されている**

DFAST をぜひご利用ください。https://dfast.nig.ac.jp

**真核生物の構造アノテーションは難しく、発現データとの組み合わせが必要**

**GenBankやGFF形式は標準的なゲノムアノテーションの書式**

Entry – Feature – Qualifier の階層構造

# 公共配列データベースの現状と諸問題

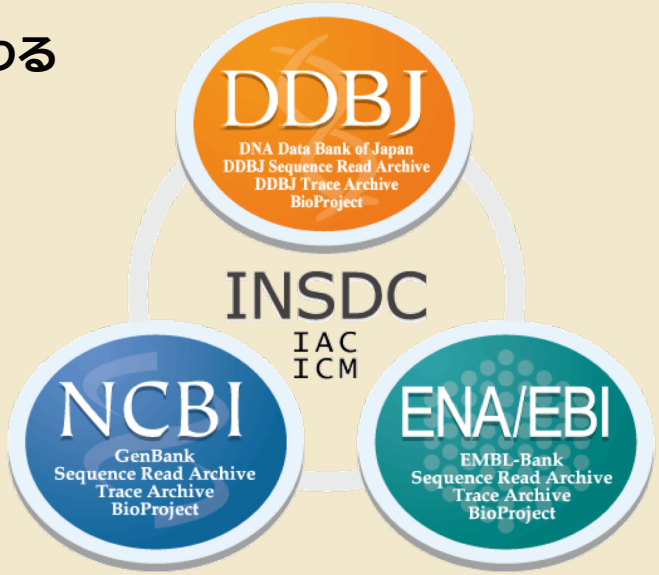
# INSDC: International Nucleotide Sequence Database Collaboration

**DDBJ は INSDC の一員として国際塩基配列DBの運用に携わる**

三極間で登録されたデータは同期・共有される

登録されたデータは目的や国籍に拘わらず、閲覧転用が可能

技術の進歩・時代の要請にあわせてデータベースの種類も増えている

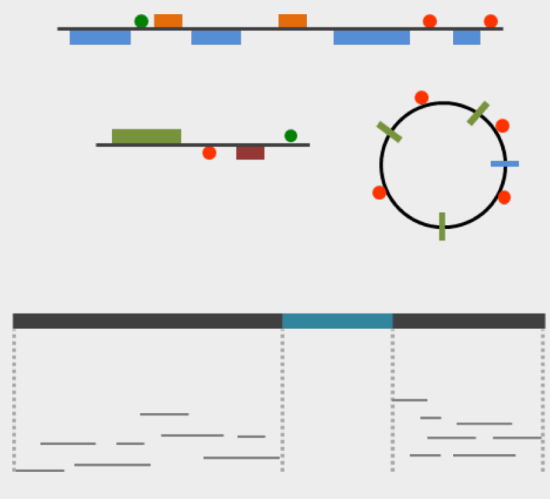



| Data type | DDBJ Center | EMBL-EBI | NCBI |
|---|---|---|---|
| **Next generation reads** | Sequence Read Archive | European Nucleotide Archive (ENA) | Sequence Read Archive |
| **Sequence Read Archive** | Trace Archive | | Trace Archive |
| **Annotated sequence** | DDBJ | | GenBank |
| **Samples** | BioSample | | BioSample |
| **Studies** | BioProject | | BioProject |



DDBJ児玉さんの資料を改変

# DDBJ が管理・運営する様々なデータベース



**DDBJ**
新規に決定（シーケンスまたはアセンブル)
された配列。アノテーションされた遺伝子
情報も含む。Traditional-DDBJとも呼ばれる

**BioProject**
**BioSample**

「なぜ」シーケンスを行なったのか
「何を」シーケンスしたのか

**DRA**
シーケンサーからの生データ または
参照配列にアラインメントしたデータ

**INSDC: オープンアクセスデータベース**

**アクセス制限データベース**

**JGA**
個人レベルの遺伝型と表現型

NBDC
ヒトデータ審査委員会

この他にGenomic Expression Archive (GEA)
遺伝子発現やエピゲノムなどの機能ゲノミクスデータを登録

DDBJ福田さんの資料を改変

# INSDCの歴史

- 1960-70年代　Margaret Dayhoff

  Atlas of Protein Sequence and Structure
  後の Protein Information Resource (1984, 最古のタンパク質配列DB)

- 1982年　(欧) EMBL data library 活動開始

- 1982年　(米) ロスアラモス国立研究所のWalter Goadら
  後のGenBank (NCBI, 1992) につながる塩基配列データベース設立

- 1984年　遺伝研にてDDBJ業務が部分的に始まる(丸山・五条堀)

- 1987年　DDBJ本格活動開始

参考)
DDBJ30年のあゆみ
https://www.ddbj.nig.ac.jp/ddbj30th/timeline.html

遺伝子バンク30年
三島・国立遺伝研

世界のＤＮＡデータ集約

生命の設計図を形作るDNA。A、T、G、Cの4文字で表されるDNA塩基配列のデータベースは日米欧の国際協力によって維持され、生命科学や医学研究の根幹を支えている。その国内拠点が、三島市の国立遺伝学研究所内にある「日本DNAデータバンク（DDBJ）」だ。

DDBJの心臓部は、1室にぎっしりと並んだスーパーコンピューター。大量の熱を発するコンピューターを安定稼働させるために、大きな音を立ててファンが回り、室温を一定に保っている。公共の知的財産であるDNAデータを常に最新の状態で無償公開し、世界中からの検索要求に応えるためには、このような大型のシステムが必要になる。研究の計画を立てたり、実験で得た塩基配列の意味を読み解くためには、データベースでの検索や解析が欠かせない。

世界中の大学や研究所から読み出されるDNAデータが現在のように集約されるまでの道はけっして平坦ではなかった。ヒトゲノム計画がまだ構想段階にあった1980年代、発足したばかりの欧米のデータベースでは、学術誌のページをめくってDNAデータを探し、1文字ずつ手入力していた。DDBJは遺伝研内外の研究者が利用するDDBJのスーパーコンピューター＝三島市の国立遺伝学研究所

発足当初から欧米と緊密に連携・データ交換し、2005年から三者は国際塩基配列データベース（INSD）として一体的に運営されている。

現在、DNAデータが含まれる論文には、INSDから発行されたIDを併記するよう求める学術誌が多い。つまりINSDに登録しないと研究が前に進まない。そして、集約されたINSDを検索することから次の研究が始まるのだ。

（伊東真知子・国立遺伝学研究所特任研究員）

◇

DDBJはことし30周年。5月27～29日に三島市で開く国際シンポジウムを前に、その生い立ちを紹介する。

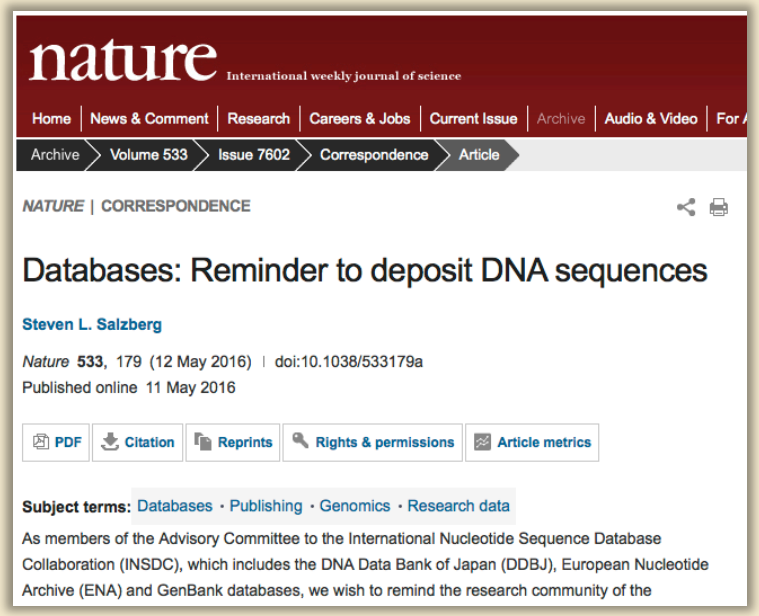
「遺伝子バンク30年」（遺伝研・伊藤、静岡新聞連載記事）
https://www.ddbj.nig.ac.jp/files/pdf/30th/shinbun_all.pdf

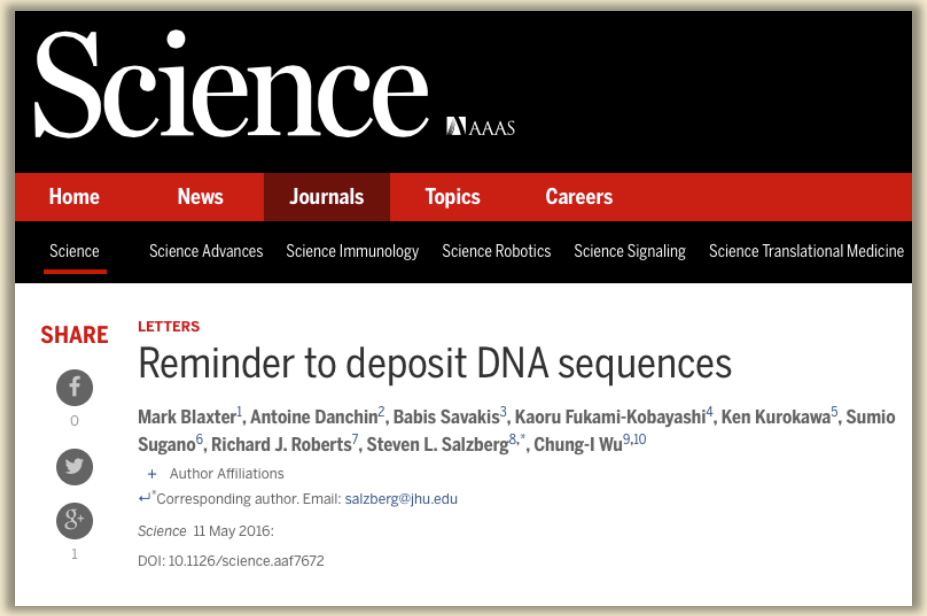
# データの登録は論文投稿の条件となっている

Goad の理念　　**研究によって得られた成果は人類共有の資産**

データベースへの配列登録は論文発表に伴う義務とする

無償でのデータ利用・再配布を認める → 論文結果の再現性の担保・再利用の促進




INSDC国際諮問委員会によるデータ登録への呼びかけ (2016)

# PrimaryデータベースとSecondaryデータベース

| | 一次データベース<br>Primary database | 二次データベース<br>Secondary database |
|---|---|---|
| 別の呼び方 | Archival database | Curated database; Knowledgebase |
| データソース | 研究者（登録者）が実験で得たデータを直接登録 | 一次データベースのデータや文献を解析、解釈、キュレーションした結果 |
| 例 | • DDBJ/ENA/GenBank<br>• DOR/ArrayExpress/GEO<br>• DRA/ERA/SRA<br>• EVA·DGVa/dbSNP·dbVar<br>• PDB | • RefSeq<br>• Ensembl<br>• Expression Atlas<br>• ChIP-Atlas<br>• UniProt |

2018.1 AJACS 講習会浜松 DDBJ児玉さんの資料より

RefSeq は NCBI によって管理・運営されている塩基配列二次データベース

GenBank に登録されたデータに対して、独自の品質チェック・再アノテーションを行なった上で再公開している。

# DDBJ (INSDC) が抱える問題点

**予算・人員の不足**

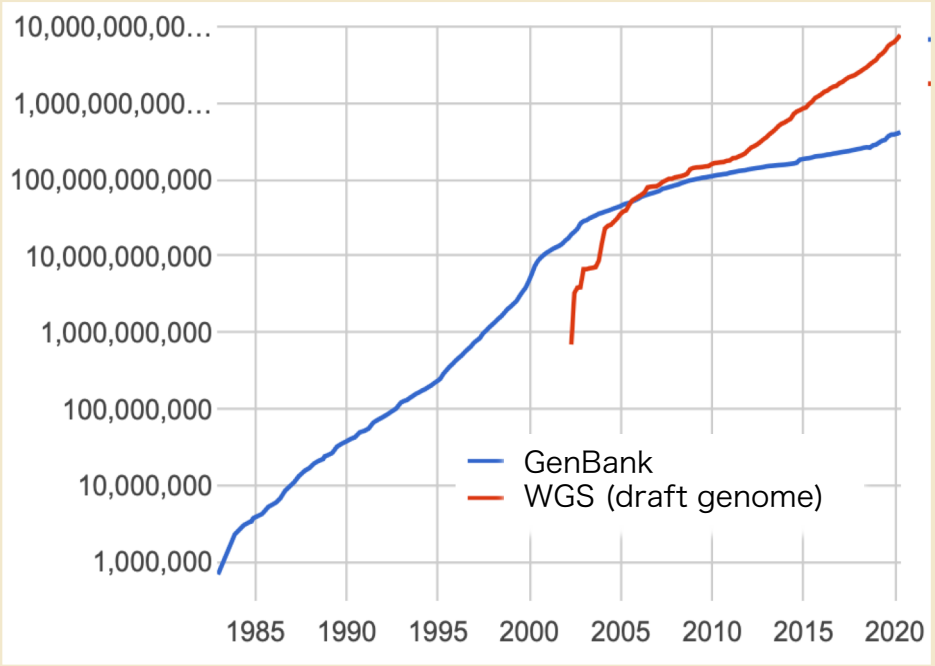
**データ量の増大**

**問題のある登録情報**

- 誤った注釈情報の伝播
- 生物種名の誤り
- 低品質な配列・コンタミネーション

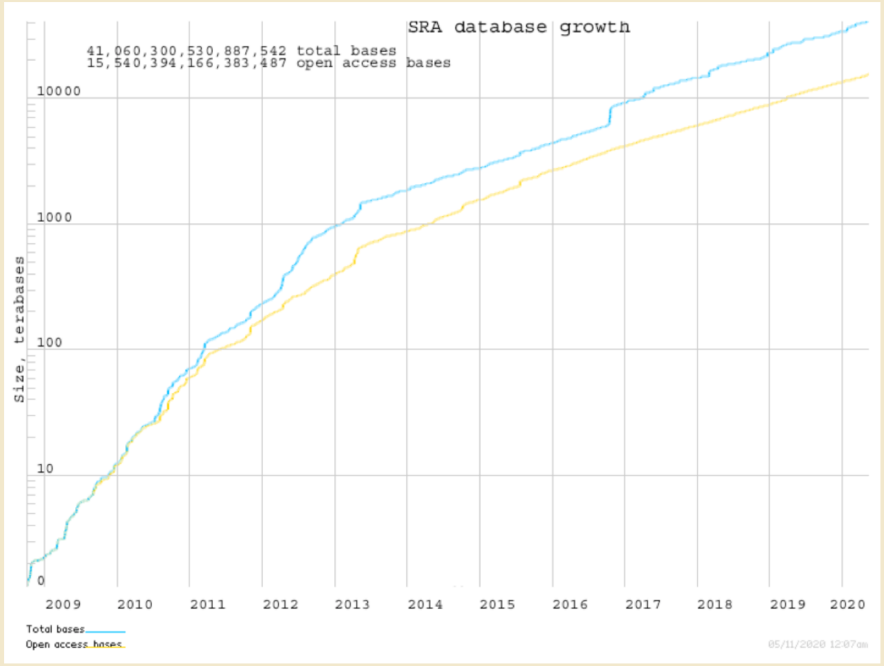
# 増大するデータ量

GenBankのデータ量（塩基数）




GenBank: 415,770,027,949塩基
WGS: 7,788,133,221,338塩基

Sequence read archive のデータ量（塩基数）




41,060,300,530,887,542塩基
(4京1060兆、41ペタ)
(2019年同時期、27ペタ)

GenBankデータベースは1982年以来、18ヶ月で約2倍のペース

2018.3 の DDBJ スパコンのリプレイスでは、
データベースストレージ 5.6PB → 30 PB に増強

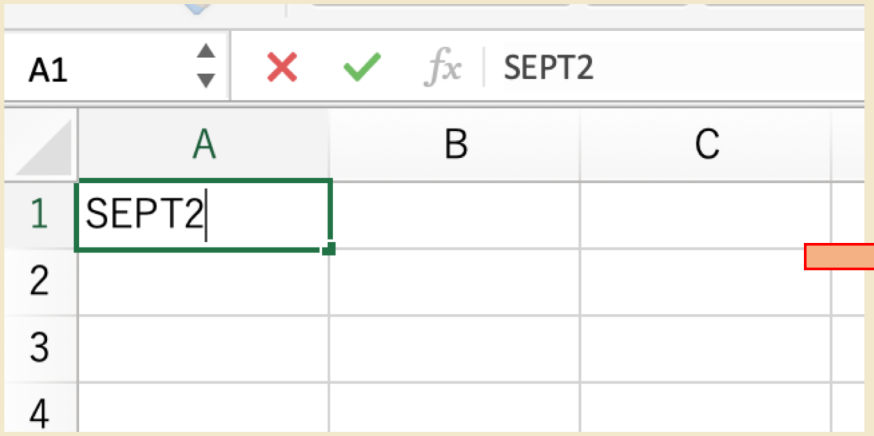
https://www.ncbi.nlm.nih.gov/genbank/statistics/ および https://trace.ncbi.nlm.nih.gov/Traces/sra/ より

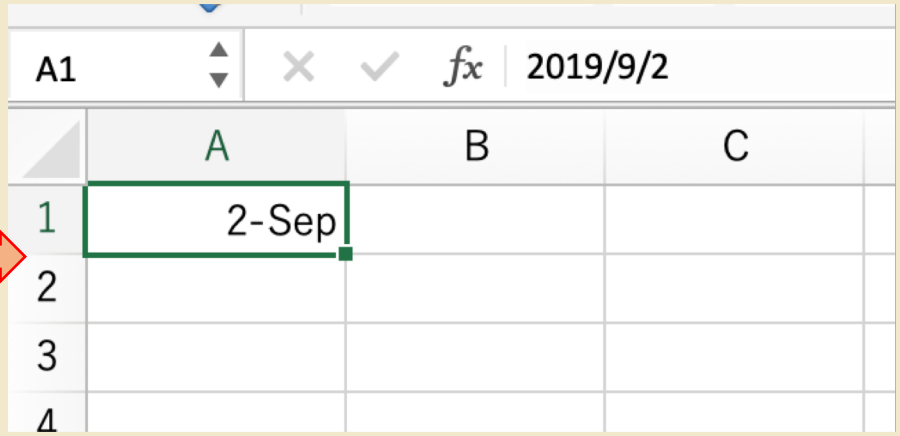
# 遺伝子名の伝言ゲーム

actin-related protein → similar to actin-related protein

→ similar to similar to actin-related protein

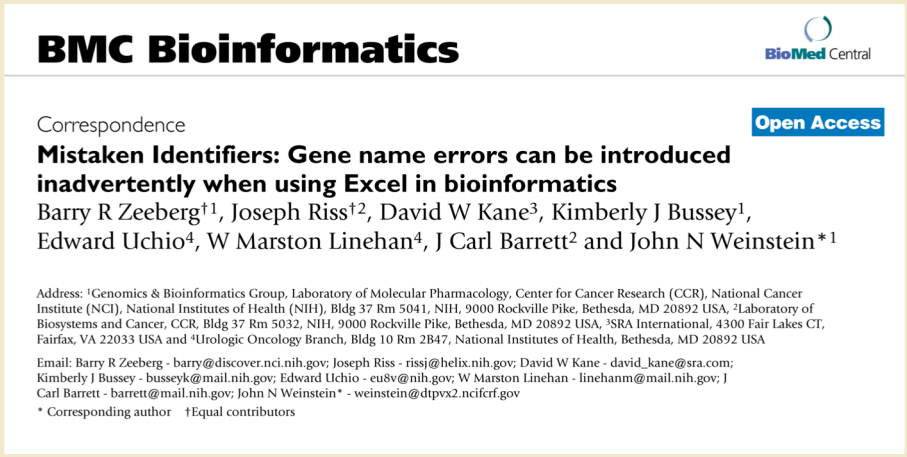
# 2004年から続くExcel 問題




日付データに変換されてしまう

2004年　Zeeberg et al.



**BMC Bioinformatics**

BioMed Central

Correspondence

Open Access

**Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics**

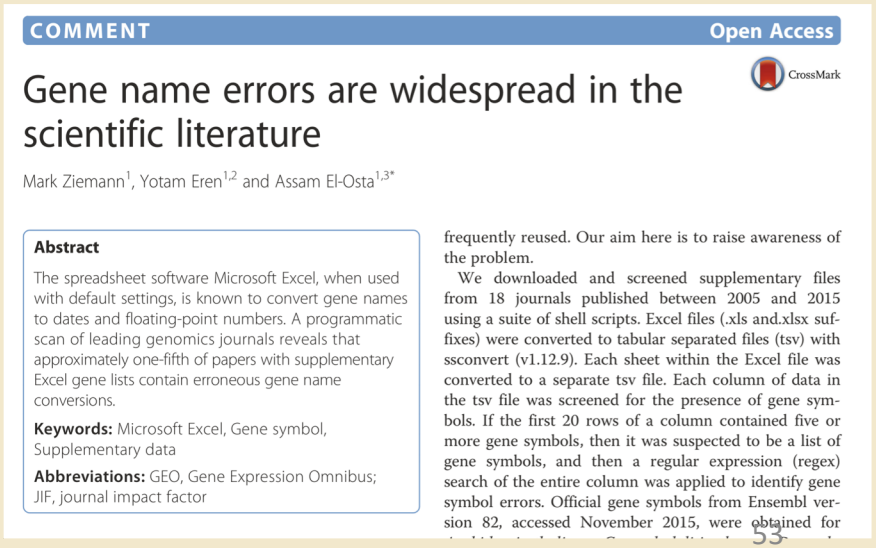
Barry R Zeeberg[†1], Joseph Riss[†2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein[*1]

Address: [1]Genomics & Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research (CCR), National Cancer Institute (NCI), National Institutes of Health (NIH), Bldg 37 Rm 5041, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, [2]Laboratory of Biosystems and Cancer, CCR, Bldg 37 Rm 5032, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, [3]SRA International, 4300 Fair Lakes CT, Fairfax, VA 22033 USA and [4]Urologic Oncology Branch, Bldg 10 Rm 2B47, National Institutes of Health, Bethesda, MD 20892 USA

Email: Barry R Zeeberg - barry@discover.nci.nih.gov; Joseph Riss - rissj@helix.nih.gov; David W Kane - david_kane@sra.com; Kimberly J Bussey - busseyk@mail.nih.gov; Edward Uchio - eu8v@nih.gov; W Marston Linehan - linehanm@mail.nih.gov; J Carl Barrett - barrett@mail.nih.gov; John N Weinstein* - weinstein@dtpvx2.ncifcrf.gov

* Corresponding author　†Equal contributors

2016年　Zeeberg et al.



COMMENT　Open Access

CrossMark

## Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

**Abstract**

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

**Keywords:** Microsoft Excel, Gene symbol, Supplementary data

**Abbreviations:** GEO, Gene Expression Omnibus; JIF, journal impact factor

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for

# 誤った登録情報が間違った結果につながった例

Zheng, et al. Appl Environ Microbiol. (2015)

**A Genomic View of Lactobacilli and Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology**

→ *Lactobacillus parakeri* は *L. kefiri* と同じ種である

Sun, et al. Nat Commun. (2015)

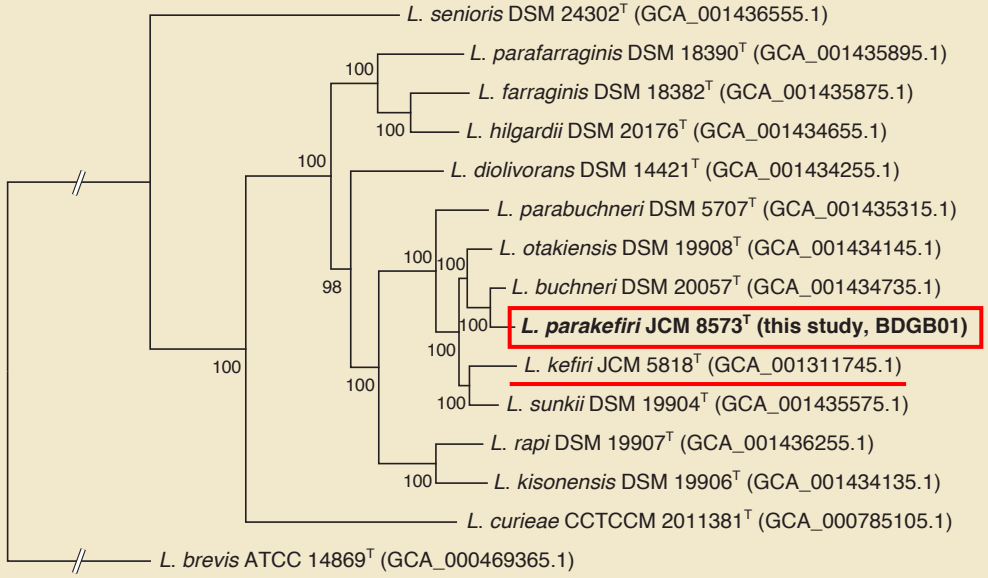
**Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera.**

→ *Lactobacillus parakeri* は *Lactobacillus* の中で最大のゲノムサイズを持つ

実際には解析に用いたゲノムにコンタミネーションが含まれていたことが原因

**Genomic characterization reconfirms the taxonomic status of *Lactobacillus parakefiri*.**

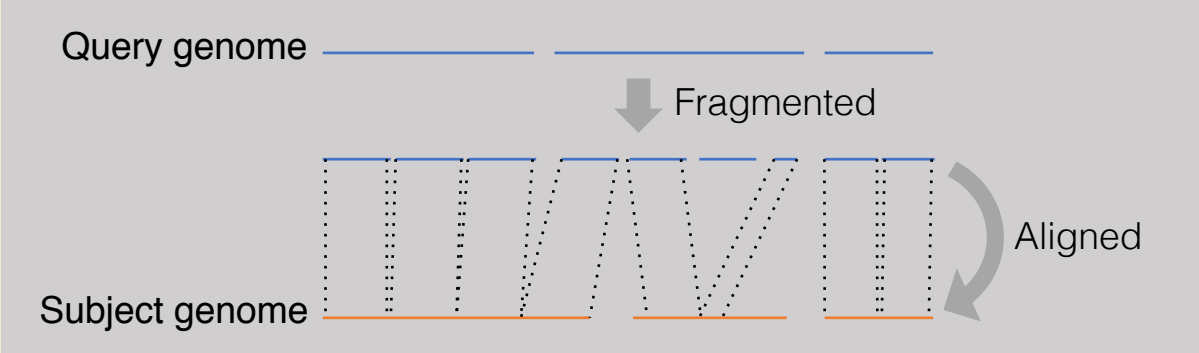
Tanizawa, et al. BMFH. (2017)



# 生物種やクオリティのチェック方法

**ANI (average nucleotide identity)**

(2005 Konstantinidis et al., 2007 Goris et al.)




ANI > 95% : 同種　　ANI < 95% : 別種

ゲノムデータに基づく系統分類の標準的な指標になっている

**single copy orthologous 遺伝子を用いたクオリティチェック**

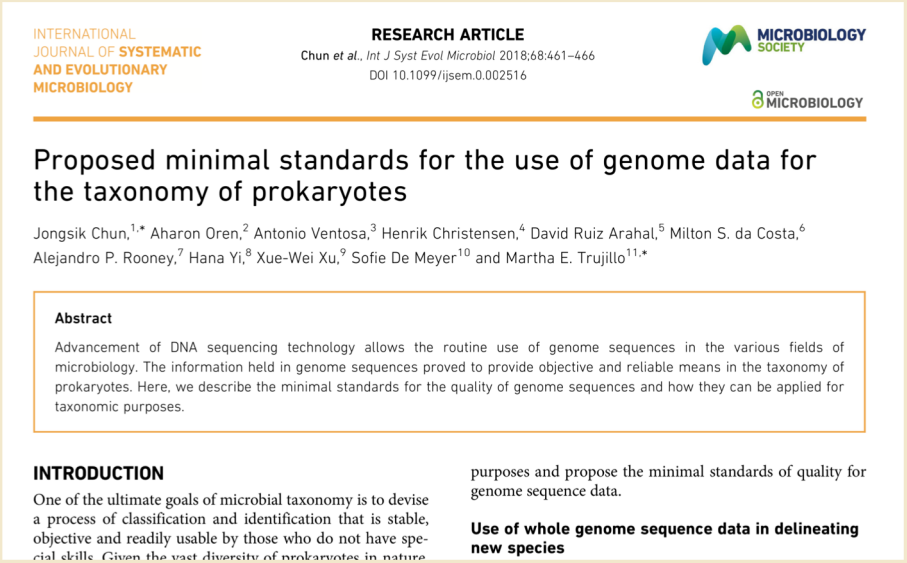
CheckM (Parks et al. 2015)

completeness, contamination を計算

BUSCO (Simão et al. 2015)

真核生物にも対象

# 系統分類学でもゲノムデータの利用が広まる



INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Chun *et al.*, *Int J Syst Evol Microbiol* 2018;68:461–466
DOI 10.1099/ijsem.0.002516

MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

**Abstract**

Advancement of DNA sequencing technology allows the routine use of genome sequences in the various fields of microbiology. The information held in genome sequences proved to provide objective and reliable means in the taxonomy of prokaryotes. Here, we describe the minimal standards for the quality of genome sequences and how they can be applied for taxonomic purposes.

**INTRODUCTION**

One of the ultimate goals of microbial taxonomy is to devise a process of classification and identification that is stable, objective and readily usable by those who do not have special skills. Given the vast diversity of prokaryotes in nature

purposes and propose the minimal standards of quality for genome sequence data.

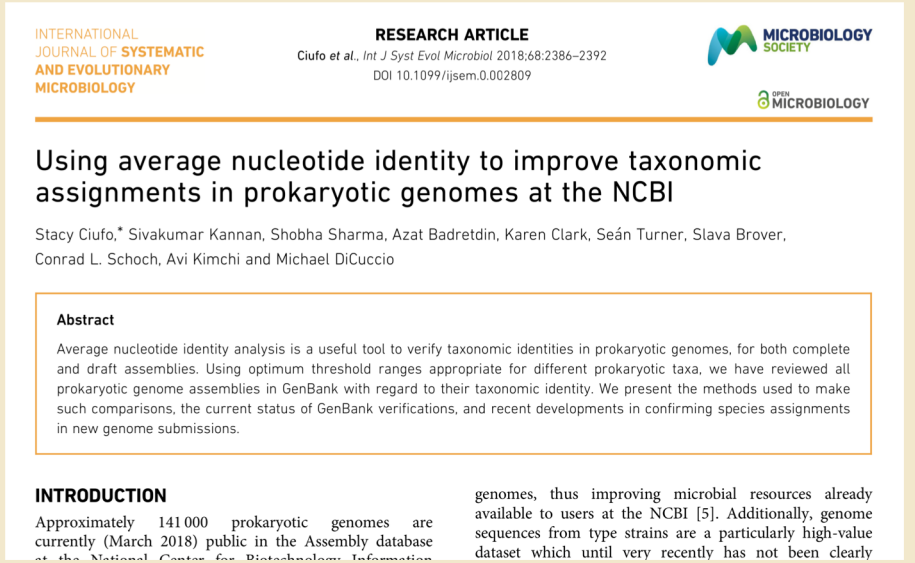
**Use of whole genome sequence data in delineating new species**

Chun et al. 2018 IJSEM

◀ 系統分類におけるゲノムデータの利用に関するガイドライン

ANI やクオリティチェックの利用が推奨されている。

NCBI でも ANI を利用した系統名のチェックが行われているようになる ▶



INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Ciufo *et al.*, *Int J Syst Evol Microbiol* 2018;68:2386–2392
DOI 10.1099/ijsem.0.002809

MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI

Stacy Ciufo,* Sivakumar Kannan, Shobha Sharma, Azat Badretdin, Karen Clark, Seán Turner, Slava Brover, Conrad L. Schoch, Avi Kimchi and Michael DiCuccio

**Abstract**

Average nucleotide identity analysis is a useful tool to verify taxonomic identities in prokaryotic genomes, for both complete and draft assemblies. Using optimum threshold ranges appropriate for different prokaryotic taxa, we have reviewed all prokaryotic genome assemblies in GenBank with regard to their taxonomic identity. We present the methods used to make such comparisons, the current status of GenBank verifications, and recent developments in confirming species assignments in new genome submissions.

**INTRODUCTION**

Approximately 141 000 prokaryotic genomes are currently (March 2018) public in the Assembly database

genomes, thus improving microbial resources already available to users at the NCBI [5]. Additionally, genome sequences from type strains are a particularly high-value dataset which until very recently has not been clearly

Ciufo et al. 2018 IJSEM

# NCBIでのクオリティチェックの例

*Lactobacillus parakefiri* のゲノムには ”Anomalous assembly” というラベルがつけられている

ASM143421v1

Organism: **Lactobacillus parakefiri** DSM 10551 (firmicutes)
Infraspecific name: Strain: DSM 10551
Submitter: Shanghai Majorbio
Date: 2015/11/06
Assembly level: Scaffold
Genome representation: full
Relation to type material: assembly from type material
GenBank assembly accession: GCA_001434215.1 (**latest**)
RefSeq assembly accession: GCF_001434215.1 (suppressed)
Assembly anomaly: contaminated
Excluded from RefSeq: contaminated, genome length too large

⚠ **Anomalous assembly.**

IDs: 578601 [UID] 2581688 [GenBank] 2686808 [RefSeq]

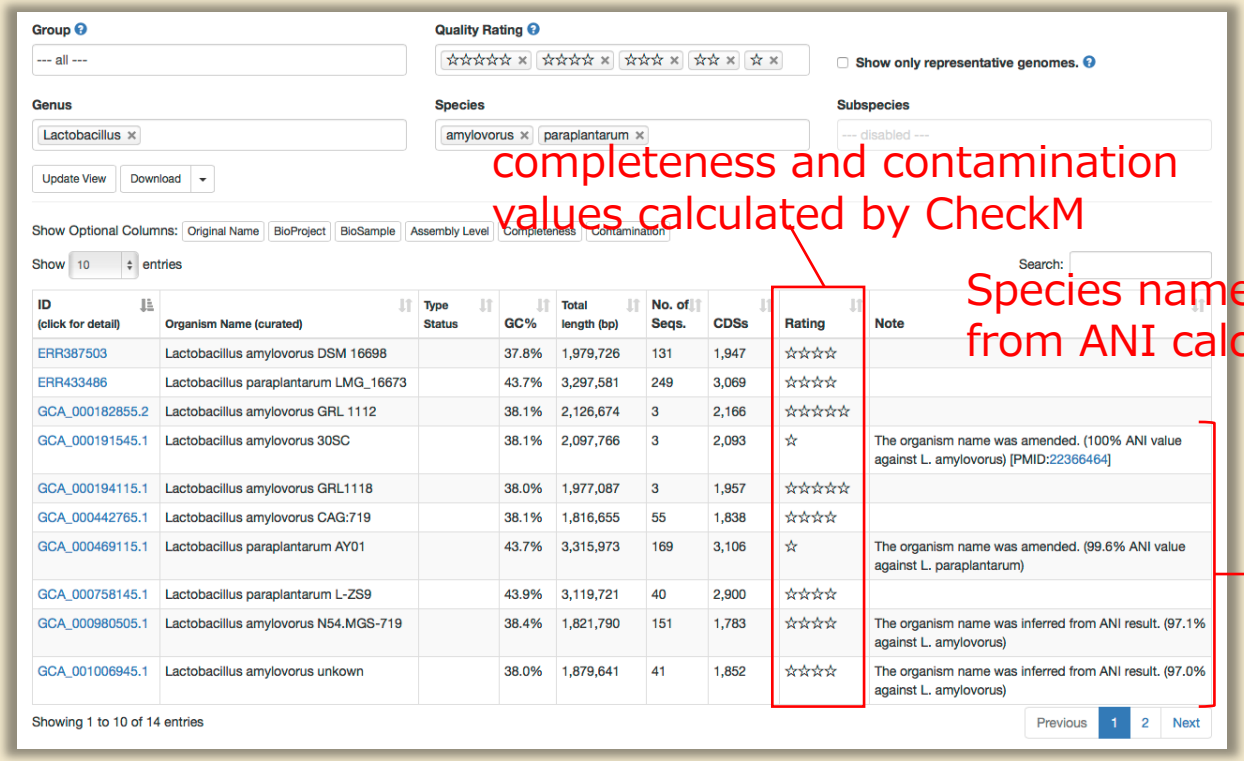
# 1,421件の乳酸菌ゲノムを独自にチェック (2016)

**Taxonomy and quality assessment**

-系統名の誤り　155 件 (ANI)

- 低品質・低クオリティのゲノム 38 件 (CheckM)

**DFAST Archive of Genome Annotation (DAGAとして公開)**



# *Lactobacillus casei* の多くは ‘*paracasei*’

約10年間、 ’*L. paracasei*’ の菌株が *L. casei* の基準株として用いられていた。

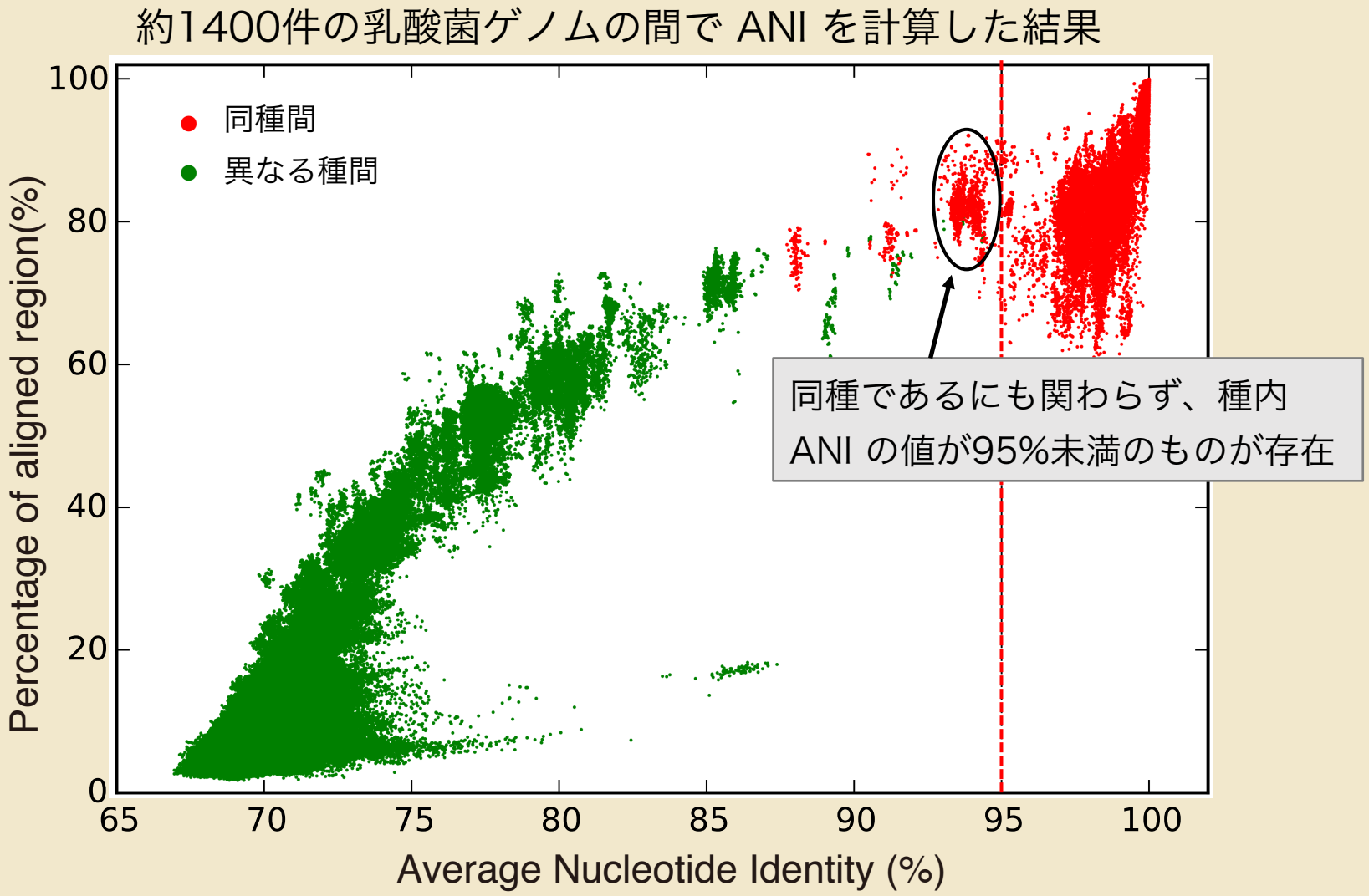
16S rRNAの配列では判定が難しいが、ANI を用いれば判定は容易

乳酸菌シロタ株も *paracasei*

*L.casei*とその近縁種の間でのANIの値

| | **L. casei** | **L. paracasei** subsp. paracasei | **L. paracasei** subsp. tolerans | L. rhamnosus |
|---|---|---|---|---|
| **L. casei** | | 78.2% | 78.4% | 79.2% |
| **L. paracasei** subsp. paracasei | 78.6% | | 98.2% | 77.3% |
| **L. paracasei** subsp. tolerans | 78.4% | 98.2% | | 77.3% |
| L. rhamnosus | 79.2% | 77.2% | 77.3% | |

# データベースから見つかった新種の乳酸菌



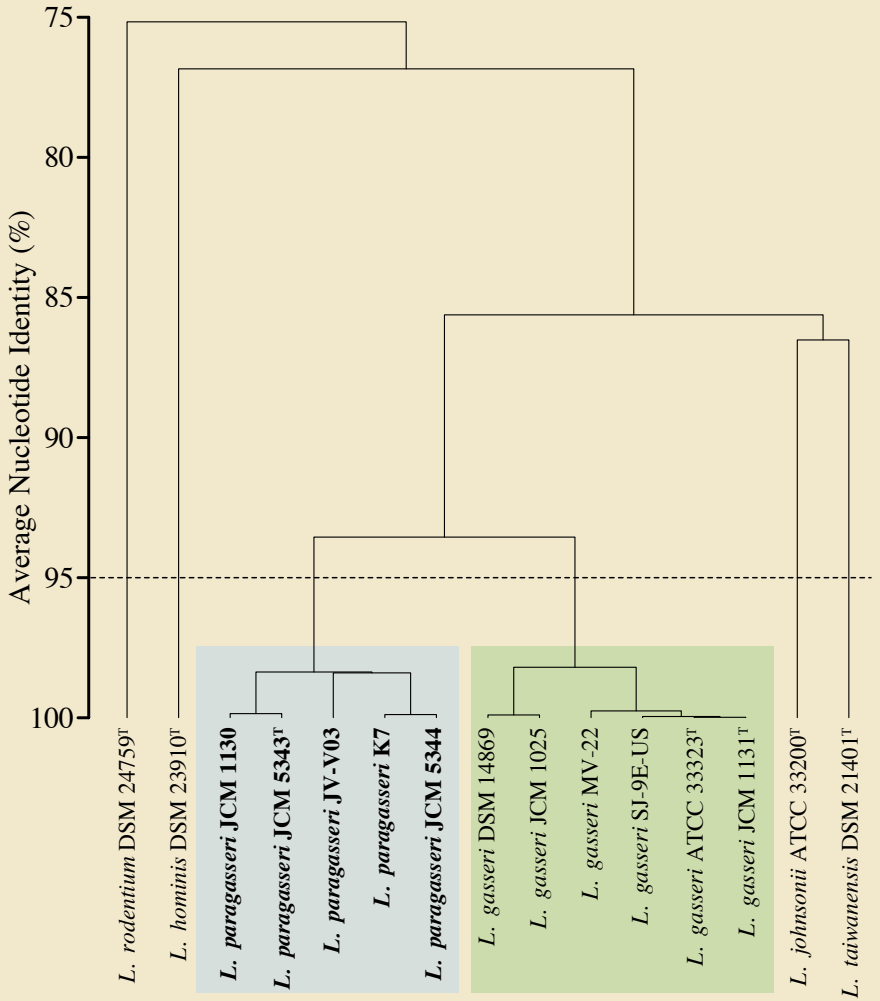
# 新種乳酸菌 *Lactobacillus paragasseri* の提唱

(Tanizawa et al. IJSEM, 2018)

*L. gasseri* はANIによって２つのグループに分かれる。

２グループ間でバクテリオシンの産生に関わる遺伝子の有無に違いがある

LG21 may become LP21?




UPGMA tree based on ANI values within *L. gasseri* and related species

# まとめ：配列データベース

**INSDC は新規決定された塩基配列を収集し、公開している**

**INSDC (DDBJ/GenBank/ENA・EBI) の間でデータは同期・共有されている**

**塩基配列を登録することは論文投稿の条件の１つとなっている**

**INSDCのデータは自由に利用・再配布が可能**

論文結果の再現性の担保、データの再利用

primary database と secondary database

**登録情報の正確さは、登録者の責任による**

データベース側でチェックを行う試みも広がっている