

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)

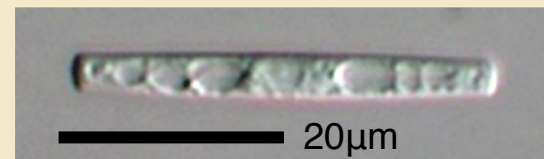


Figure (Ishii and Kamikawa, 2017, PBR)

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



MarpolBase
Genome Database for *Marchantia polymorpha*

2018.12.13 **Gene Nomenclature site updated**
We welcome your registration. The former site (wiki-based) is no longer accessible.

2017.10.5 **Genome paper published!**
The genome paper for *M. polymorpha* has been published from Cell.

2017.8.17 **Site renewal**
A new Blast server, utility tools, and a download site have become available!

Analytical Tools
BLAST similarity search against the *Marchantia* genome, transcripts and proteins
GMAP cDNA mapping/alignment to the *Marchantia* genome
Castfinder: sgRNA designing for the CRISPR/Cas9 system
GGGenome: Ultrafast genome sequence search
CRISPRdirect: Fast target search for the CRISPR/Cas9 system

Sequence retrieval utilities
Gene Fetcher by gene ID or keywords/FASTA format
Genome Slicer by genomic positions/FASTA format
SNA-Patrol by gene ID or sequence/GenBank format

Download *Marchantia* Genome Resources
Nuclear genome (2013.1.10.6)
Organelle genomes (Kikuchi & Kawasawa 2)

© 2018 *Marchantia* Working Group and Genome Informatics Laboratory, NIG

Available Tracks

- Gene models
- User-created Annotations
- Reference sequence
- Gene models
- Transcripts
- Full-length end sequences
- Ag. miRNA
- De novo assemblies of NGS reads
- De novo assemblies of longer reads
- FL transcript (DK1807; Isoseq)
- Expression
- XY plot
- HATC:sporelings [SR896224]
- HATC:archegonophores [SR896225]
- HAT1:TK1 thalli, heat-shock [SR896226]
- HATN:mixed [SR896227]
- HATD:TK1 thalli, cut-off [SR896228]
- HATP:TK1 thalli, cut-off [SR896229]
- HATS:antheridophores [SR896230]
- HATC:sporelings [SR896223]
- CGGW:gametangiohores
- CGCW:gametangiohores
- HATC:antheridophores [SR896224]
- HATC:sporelings [SR896225]
- HATC:antheridophores [SR896225]
- HAT1:TK1 thalli, heat-shock [SR896226]
- HATD:TK1 thalli, cut-off [SR896228]
- HATP:TK1 thalli, cut-off [SR896229]
- HATN:mixed [SR896227]
- HATN:TK1 thalli, cut-off [SR896228]
- HATP:TK1 thalli, cut-off [SR896229]

DDBJ DNA Data Bank of Japan

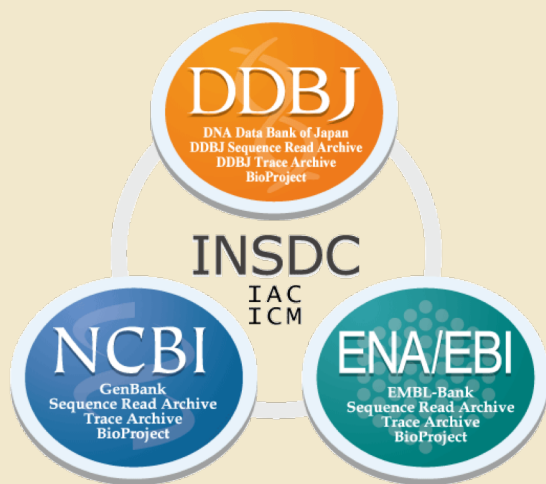


INSDC (International Nucleotide Sequence Database Collaboration) の一員として、塩基配列データを収集・公開。

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

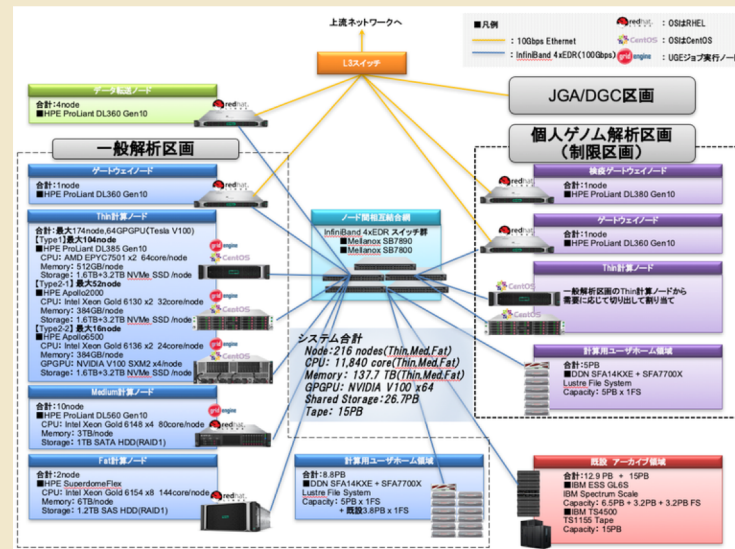
1台・144 CPUコア 12TBメモリ

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



微生物ゲノムの*de novo*アッセンブリ手法

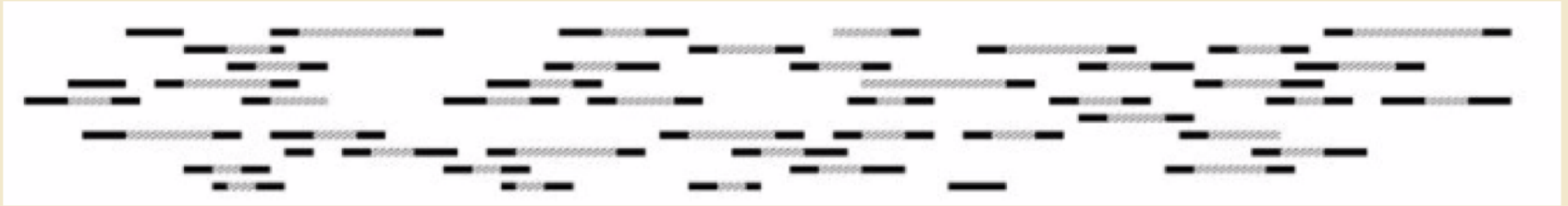
ゲノムアノテーションおよびその表現方法

公共配列データベースの現状と諸問題

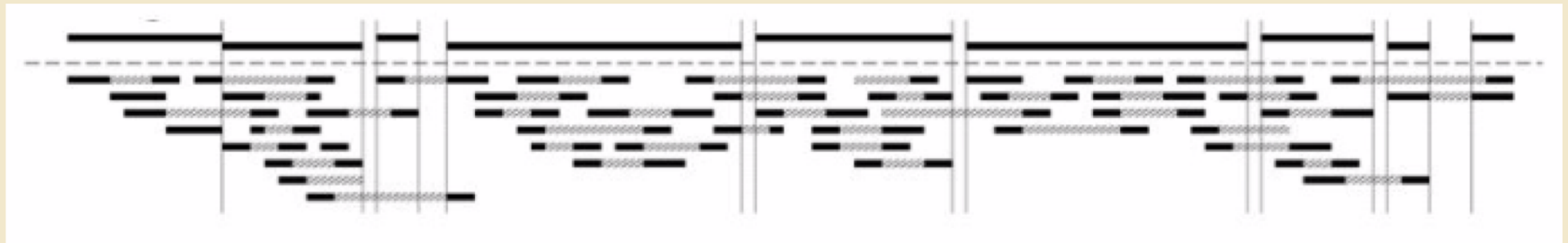
微生物ゲノムの *de novo* アセンブリ手法

ゲノムアッセムブリの一般的な流れ

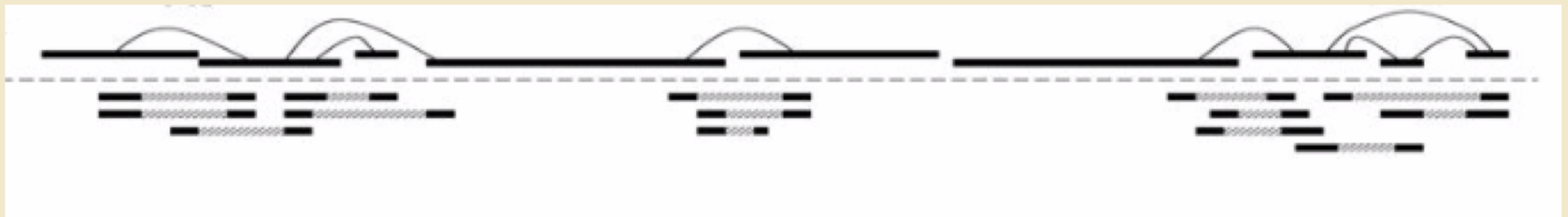
DNAを断片化し、断片の両末端を読み取る（リード）



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（スキヤフォールド）



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
```

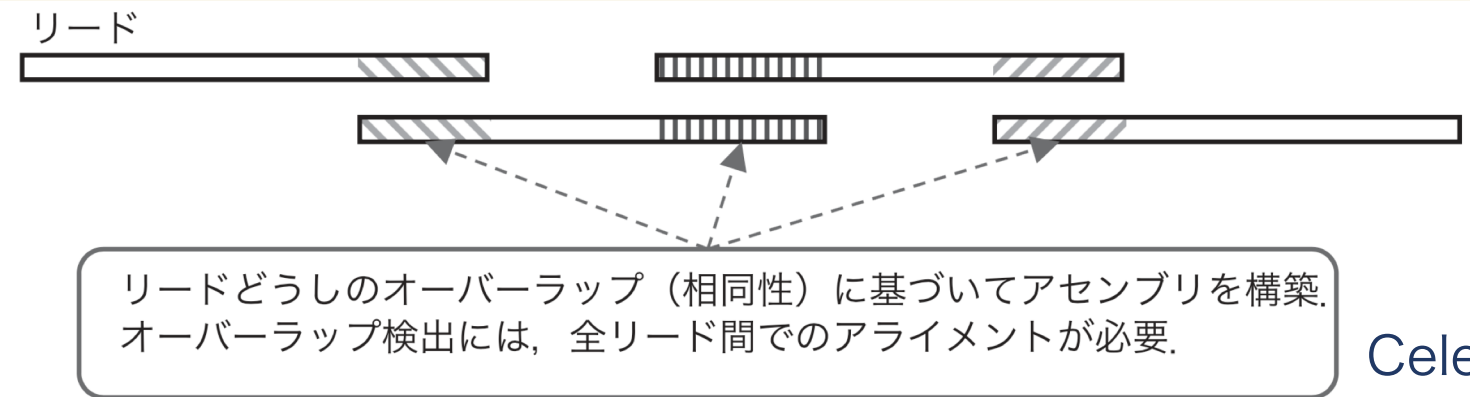
```
TGGTAATATTACTGTTGATTCATCAACGAGTAGCCCCATAGGGGCAATGGCAAAGCATAAAGGTATCTA  
TTAATTCGGATGTATAAATATTAAGTCGAATAAAAGGTATCTAGGAAAACCTTGTGAGTACGTGAAGCTTA  
AAAACGTCTGCTCGCGCTAAACGTCCTTGCTCTTTTTAAATGAAAAAGAGCCAAAGTCCATAAGGAGGTG  
TTAATGGAACCAAACGTGAAGCTTAGTTATGAAATTACGTACATCATTTCGTCCTGACATGGATGAAGCT  
ACAGCGCTTGTTGAACGATTTGACAAGATTGTGGCGCTAAACGTCAGATAATGGTGCTACGATCGTTGAT  
GACTGGTCTACTCGTCGATTTGCTTATGAAATTGTGAAGCTTAGGTGATTACAACGAAGGTACTTACCAT  
AATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGCGCTAAACG  
ATCTTGCGTCACATGATTGTTAAGCGTGAAGCTTAATCACATGATTGTTCTAATCAATTTAAAGTTAAGA  
TTAGAATCAAACGTGCTCGGATTATGGGTCTGCTACCATTTCGTTGCAGAAGACTAATTTGGAAGACTGAC  
TCCATATTGTATCTCTCGAGCCAATTAATCAATTTTTGTGTTCTTAGGAAACTGCCAGAGGAGGGAAAT  
CTCAACAAAGAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNTGTTTTGTGTTCTCACATGATTGTTTGTATTATTAGTGCTATCG  
AAACACGTTACATCACATCACATGATTGTTAGGCGTTAAAATAATACATCGATTACAAAGATACTGATT  
CGTTTTATTTCTGAACGCGGTAAGATTTTACCACGTCG...
```

```
>scaffold_02
```

```
AATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGCGCTAAACG  
GACTGGTCTACTCGTCGATTTGCTTATGAAATTGTGAAGCTTAGGTGATTACAACGAAGGTACTTACCAT  
GGATTATGGGTCTGCT...
```

de novo アセンブリアルゴリズムの二大潮流

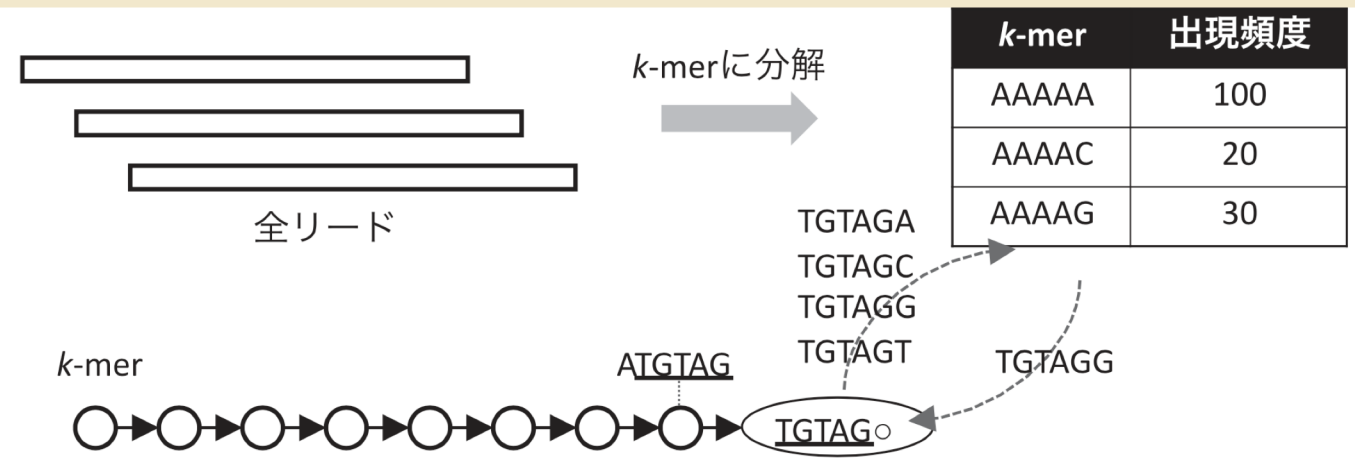
Overlap Layout Concensus (OLC)



(例)
CAP
PHRAP
Celera Assembler

ロングリードの時代になり再評価

de Bruijn graph (DBG)



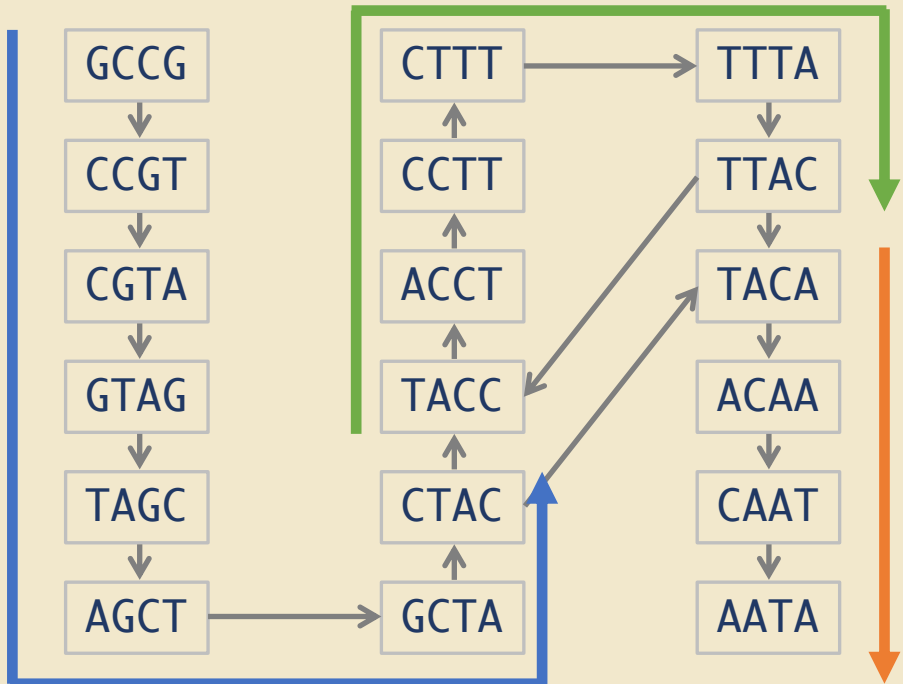
(例)
EULER
Velvet
Trinity
Platanus

ショートリードのアセンブリに最適

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	AGCT	GCTA	CCTT	CTTT	<i>K-mer</i> に分解 ($K=4$)
CCGT	GCTA	CTAC	CCTT	TTTA	
CGTA	CTAC	TACC	TTTA	TTAC	
GTAG	TACC	ACCT	TTAC	TACA	
TAGC		CCTT		ACAA	
AGCT		CTTT		CAAT	
				AATA	



各 *K-mer* を頂点
K-1 のオーバーラップを利用し、
K-mer 間のつながりを辺で表す

再構築されたコンティグ

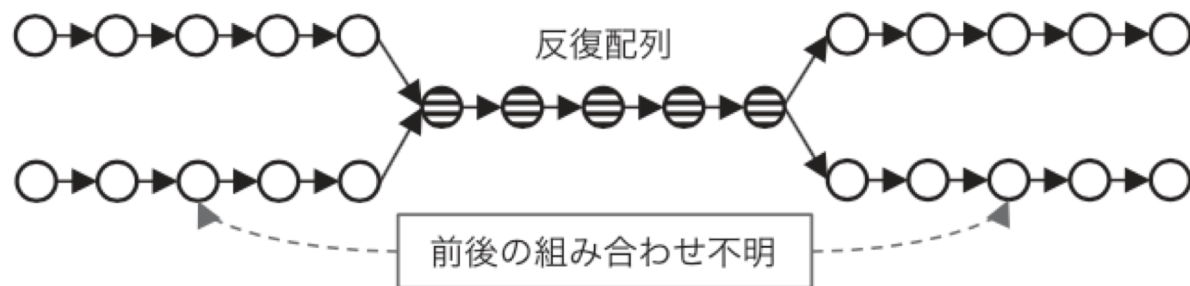
GCCGTAGCTAC

TACCTTTAC

TACAATA₁₂

グラフを複雑にさせる要因

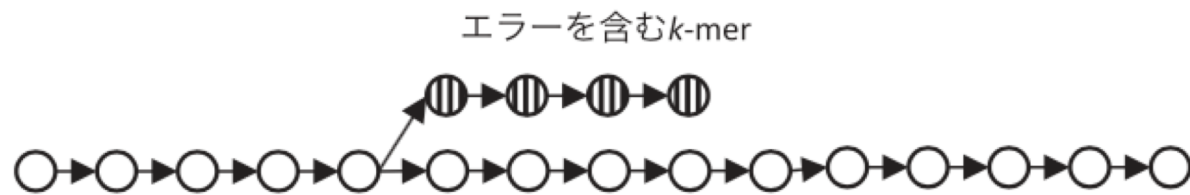
A. 反復配列や遺伝子重複による分岐



B. ヘテロ接合・エラーによるバブル構造



C. エラーによるブランチ構造



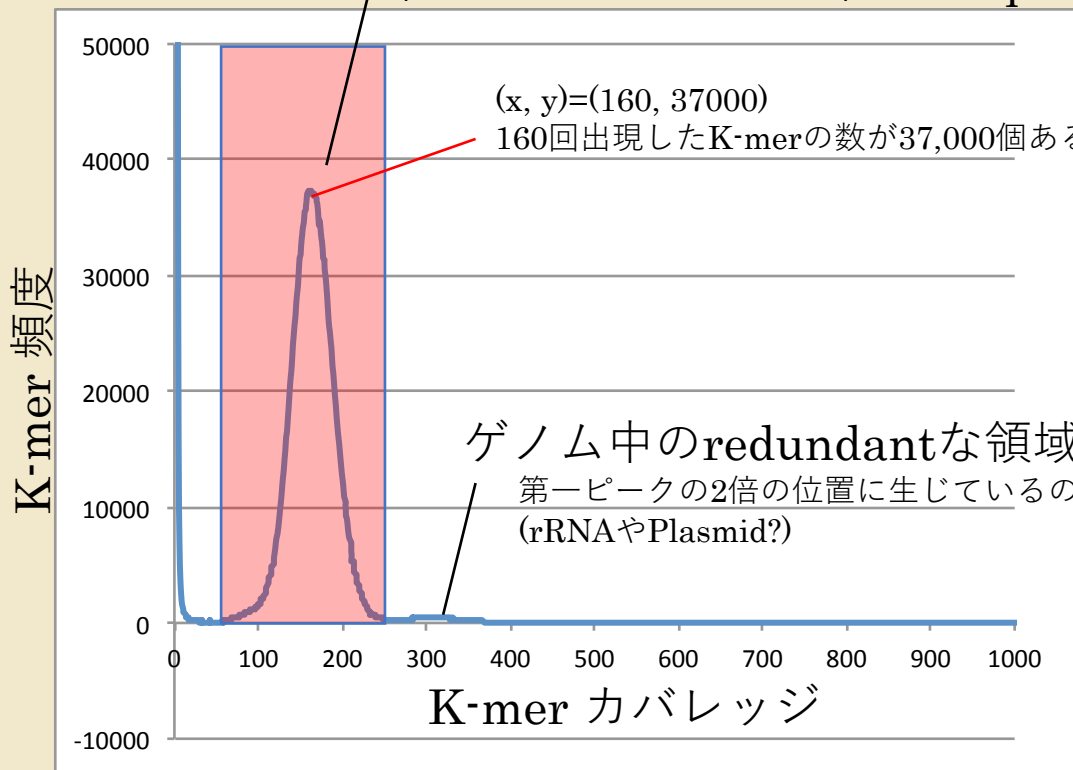
K-mer解析からゲノムアセンブルの難易度が予想できる

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



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

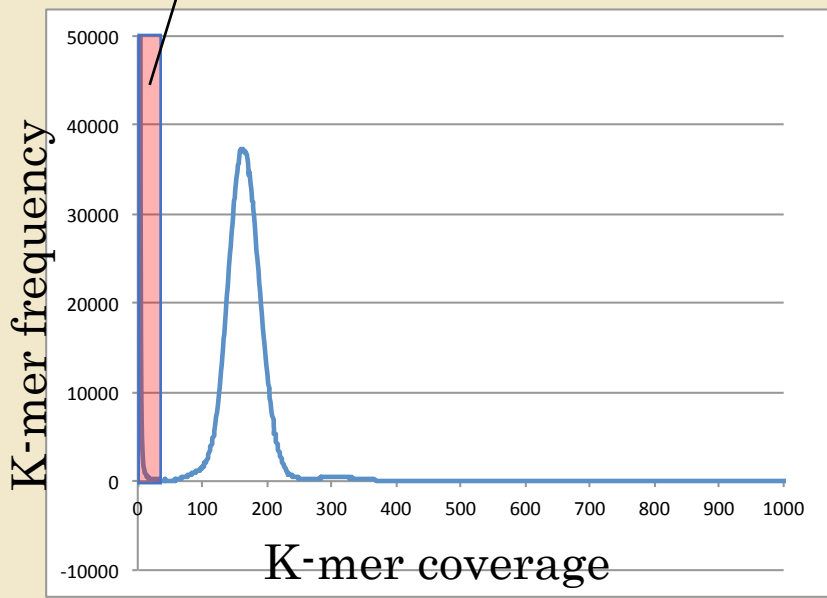
エラーを含んだ K-mer の出現回数は少ない

正しいリード CAACGAAACGAT
 Kmerに分解
 CAACG
 AACGA
 ACGCA
 CGCAA

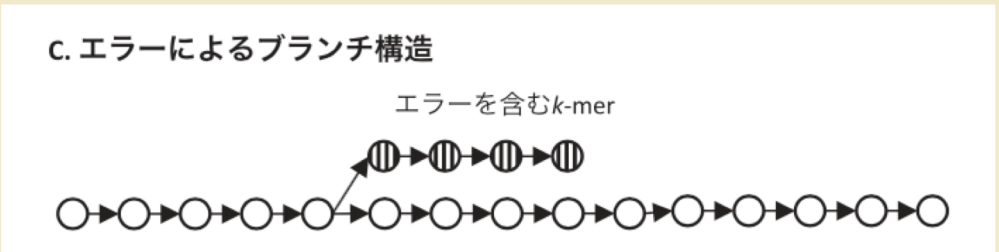
エラーを含んだリード CAACGTAACGAT
 CAACG
 AACGT
 ACGTA
 CGTAA

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

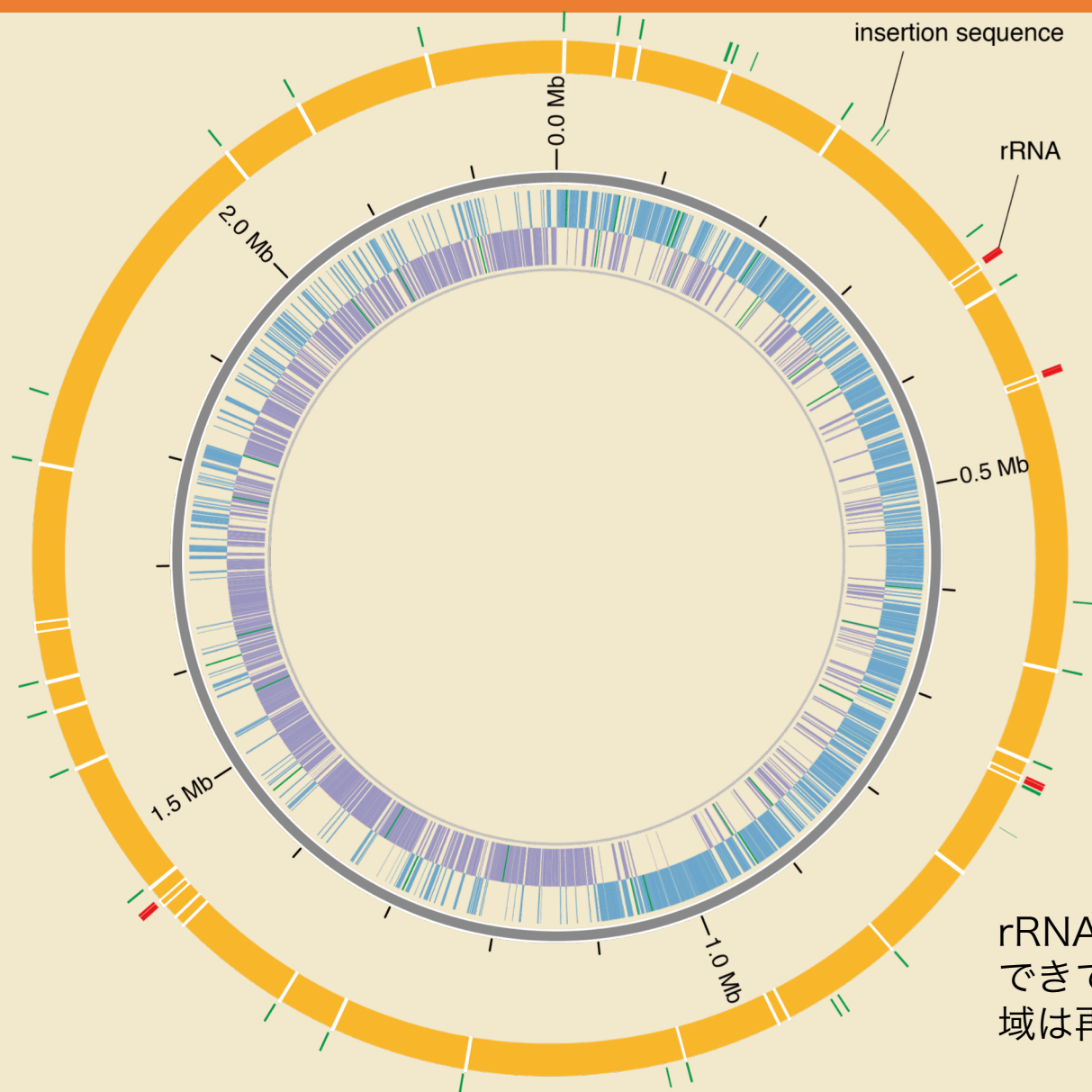
シーケンスエラー由来と考えられるK-mer



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

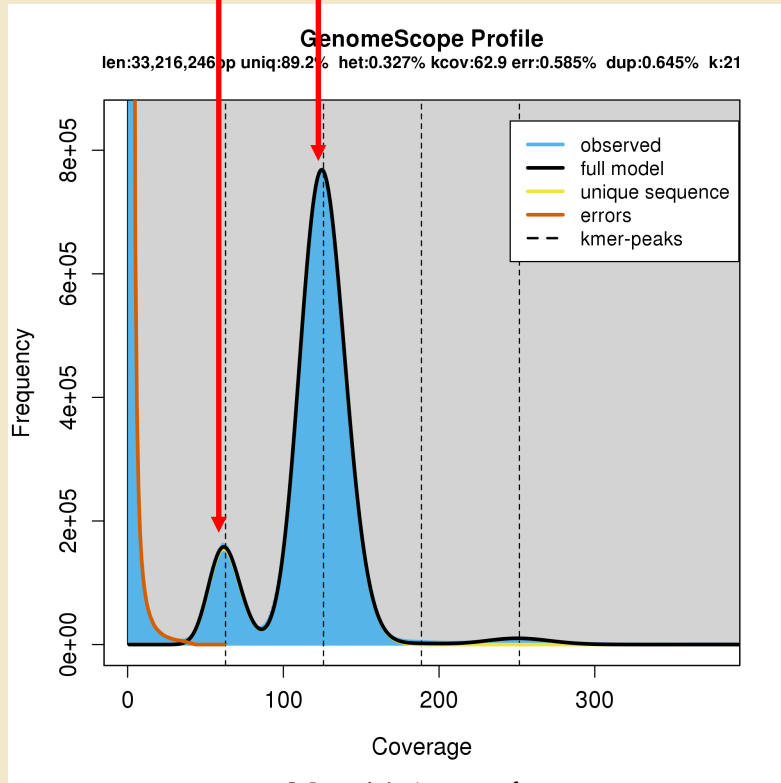
母方由来染色体 CAACGAAACGAT...GTAGCAGAAGGC

B. ヘテロ接合・エラーによるバブル構造



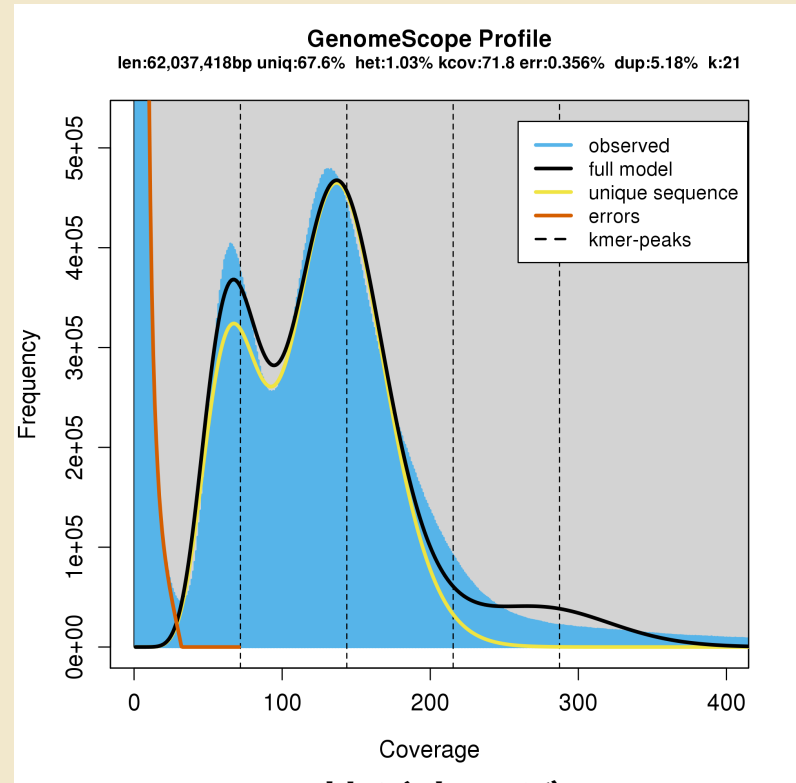
ヘテロ領域由来のK-merのピーク

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

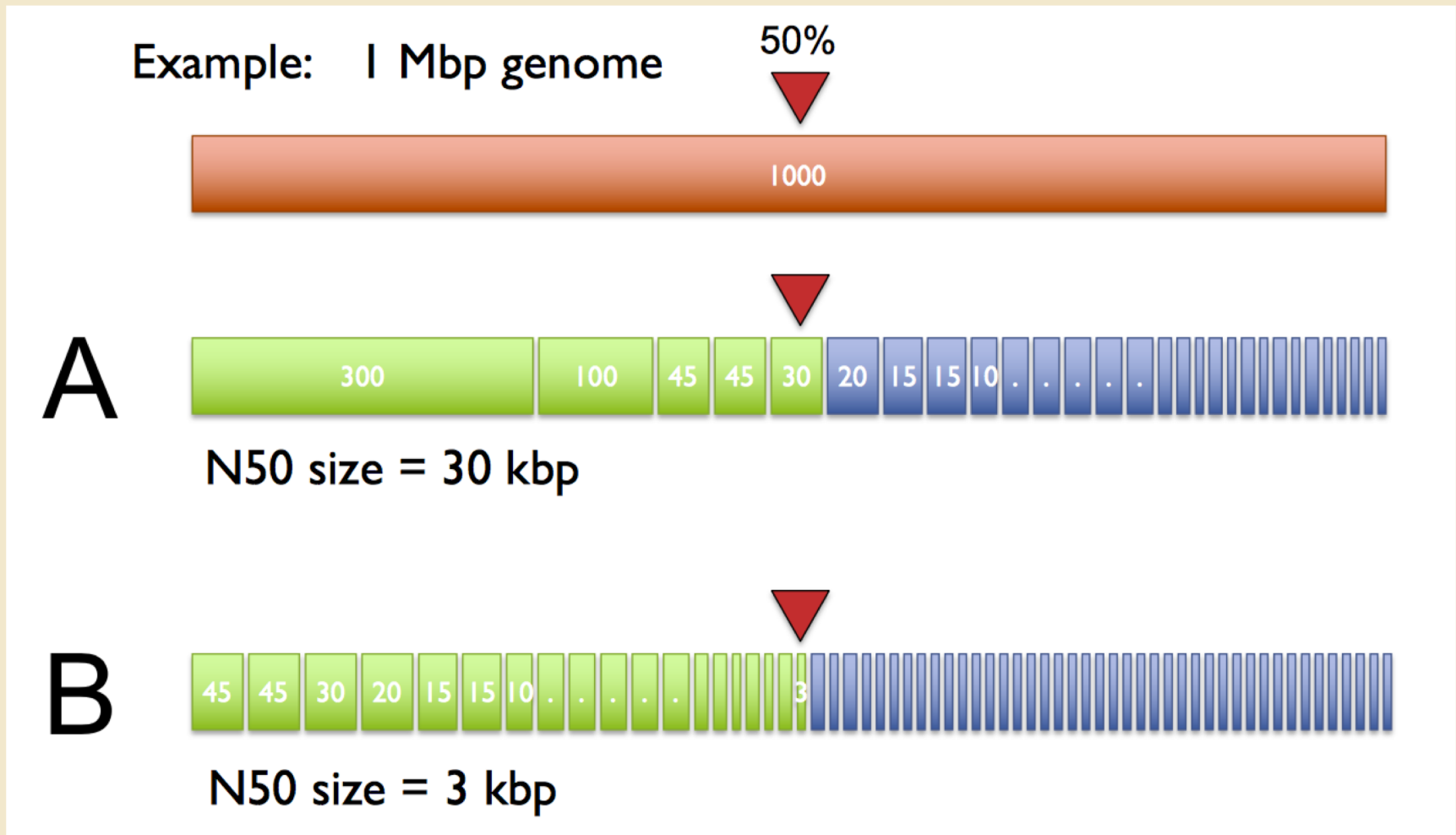


ヘテロ性が高いゲノム

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)



ロングリードのアセンブルへの利用

ロングリードシーケンサー

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

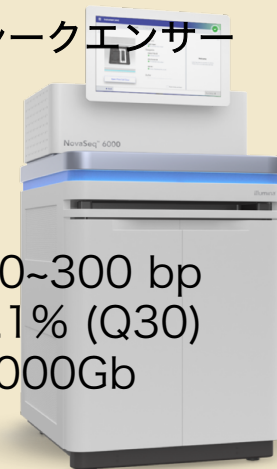


MinION

ショートリードシーケンサー

Illumina

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



NovaSeq 6000

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 MinION₉

rRNA領域の長さを十分に超えるリード長

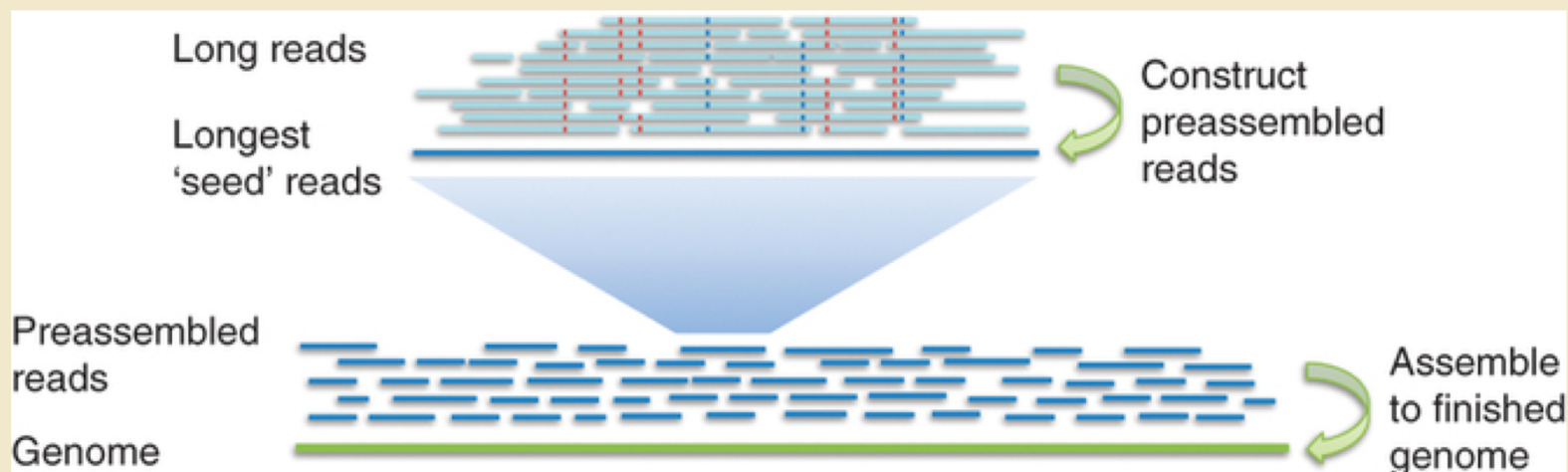
エラー率を下げるための補正が必要

frameshiftが多すぎるという注釈付
がされている

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

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

TGGTAATATTACTGTTGATTCATCAACGAGTAGCCCCATAGGGGCAATGGCAAAGCATACTCCCGTTAATTCGGATGT
ATAAATATTAAGTCGAATAAAAGGTATCTAGGAAAACCTTGTGAGTACACGTGAAAAACGTCTGCTCTCCTTGCTCTTTT
TAAATGAAAAAGAGCCAAAGTCCATAAGGAGGTGTAACAGTTAATGGAACCAAACGTTATGAAATTACGTACATCATT
CGTCCTGACATGGATGAAGCTGCTAAAACAGCGCTTGTTGAACGATTTGACAAGATTGTGTCAGATAATGGTGCTACGA
TCGTTGATTGAAAGACTGGTCTACTCGTCGATTTGCTTATGAAATTGGTGATTACAACGAAGGTACTIONTACCATATCGT
TAATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTAGTGACGATATCTTGCGTCAC
ATGATTGTTAAGCGTGAAGCTTAATCTAATCAATTTAAAGTTAAGAAAGGAGTATTAGAATCAAACGTGCTCGGATTAT
GGGTCTGCTACCATTCGTTGCAGAAGACTAATTTGAAATTGTCCATATTGTATCTCTCGAGCCAATTAATCAATTAGG
AAACTGCCAGAGGAGGGAAATTCAATGGCTCAACAAAGAAGAGGCGGACATCGTCGCCGTAAGGTTGACTTTATTGCCG
TTCACAGATTTAAGACACATACTTTTTGTTTTGTGTTCTTGTTTTATTAGTGCTATCGTGTTATAATTTTTGCTTACCG
AAAAACACGTTTACATCACATAGGCGTTAAAATAATACATCGATTACAAAGATACTGATTTACTAAAACGTTTTATTTT
TGAACGCGGTAAGATTTTACCACGTGATTTAATGTAAATGTTTATTTAAATCCTAATTATGCCATGATTGTGGTGTGA
TTAGGTCTCGTCCCGTAAGGTAAGAACATTAACAATATCACCCACTATATGATTAATCGTACAATTCTTGTTGGACGCT
TAACTAGAGATCCTGAGTTGCGATACACAAGTAGTGGAGCTGCTGTAGCAACGTTTACCGTTGCTGTCAATCGGCAGTT
TACCAATCAACAGGGTGAACGGGAAGCTGATTTTATTAGCTGCGTCATTTGGCGTAAAGCTGCTGAAAATTTTTCCAAT
TTCACTCATAAGGGTTCTTTGGTTGGGGTTGATGGCCGCATTCAAACGCGAAATTATGAAAATCAACAGGGTCAACGTG
TTTATGTAACGGAAGTAGTAGTTGAAAACCTTCTCGTTACTAGAAACGAAAGCCCAAAGTCAAACCATAATAATGGTGC
CCCAAGCTTTGACAATAATCAACAAGCCAATGCTCCTCAATCATCATCAGCAAATGATAATCCGTTTGGTAATGCTAAT
GACAATGCAAATGCGGGAAGTAGTAGTGCTAACAGCAATGCTAACGATCCATTCGCTAATAATGGCGAACCAATCGACA
TTTCAGATGACGATTTGCCGTTCTAACAAAGTTAGTGGAACAAGTGCTAAAAACCAGCGTCGTTTAAACAATTGCAATCA
AACGTGCTCGGATTATGGGTCTGCTACCATTCGTTGCAGAAGACTAATTTGAAATTGTTTTAAT...

ゲノムアノテーションとは

>Chromosome

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

TGGTAATATTACTGTTGAT TAATTCGGATGT
ATAAATATTAAGTCGAATAAAAGGTATCTAGGAAAACCTTGTGAGTACACGTGAAAAACGTCTGCTCTCCTTGCTCTTTT
TAAATGAAAAAGAGCCAAAGTCCATAAGGAGGTGTAACAGTTAATGGAACCAAACGTTATGAAATTACGTACATCATT
CGTCCTGACATGGATGAAGCTGCTAAACAGCGCTTGTTGAACGATTTGACAAGATTGTGTCAGATAATGGTGCTACGA
TCGTTGATTCGAAAGACTGGTCTACTCGTCGATTTGCTTATGAAATTGGTGATTACAACGAAGGTACTIONTACCATATCGT
TAATATCACAGCAAACGATGATGTAGCGCTAAACGAATTTGATCGTTTAGCTAAGTTTGTGACGATATCTTGCGTCAC
ATGATTGTTAAGCGTGAAGCTTAATCTAATCAATTTAAAGTTAAATGAAAGCAATGCAAGCTGCTCGGATTAT
GGGTCTGCTACCATTGTTGCAGAAGACTAATTTGAAATTGTCCAATCAATTAGG
AAACTGCCAGAGGAGGGAAATTCAATGGCTCAACAAAGAAGAGGCATTATTGCCG
TTCACAGATTTAAGACACATACTTTTTGTTTTGTGTTCTTGTTTTATTAGTGCTATCGTGTTATAATTTTTGCTTACCG
AAATACAAAGATACTGATTTACTAAAACGTTTTATTTC
TGAATTTAAATCCTAATTATGCCATGATTGTGGTGTGA
TTAGGTCCTCGTCCGTAAGGTAAGAACAATTAACAATATCACTACTATATGATTAATCGTACAATTCTTGTTGGACGCT
TAACTAGAGATCCTGAGTTGCGATACACAACACTAGTGGAGCTGCTGTAGCAACGTTTACCGTTGCTGTCAATCGGCAGTT
TACCAATCAACAGGGTGAACGGGAAGCTGATTTTATTAGCTGCGTCATTTGGCGTAAAGCTGCTGAAAATTTTTCCAAT
TTCACCTCATAAGGGTTCTTTGGTTGGGGTTGATGGCCGCATTCAAACGCGAAATTATGAAAATCAACAGGGTCAACGTG
TTTATGTAACGGAAGTAGTAGTTGAAAACCTTCTCGTTACTAGAAACGAAAGCCCAAAGTCAAACCATAATAATGGTGC
CCCAAGCTTTGACAATAATCAACAAGCCAATGCTCCTCAATCATCATCAGCAAATGATAATCCGTTTGGTAATGCTAAT
GACAATGCAAATGCGGGAAGTAGTAGTGCTAACAGCAATGCTAACGATCCATTCGCTAATAATGGCGAACCAATCGACA
TTTCAGATGACGATTTGCCGTTCTAACAAAGTTAGTGGAACAAGTGCTAAAAACCAGCGTCGTTTAAACAATTGCAATCA
AACGTGCTCGGATTATGGGTCTGCTACCATTGTTGCAGAAGACTAATTTGAAATTGTTTTAAT...

Gene XXX
Function for xxxxxx

Gene YYY with ZZZ domain
Similar to xxx of yyy (zz.z%)

相同性検索で遺伝子機能を予測する

配列が類似 \Rightarrow 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

BLAST \gg blastp suite

Standard Protein BLAST

blastn blastp **blastx** tblastn tblastx

Enter Query Sequence

BLASTP programs search protein databases using a prote

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#) [To](#)

ILKQVINQTVIAVSTQESRPILTGHIHLSYNGELLAVATDSHRLSQRKITLADAGSEVYNIIVPGKSLN
EMAKMLGDSTDNVEIRVAENQILFTFANISFYRLLLEGNYPTDRIPQSSETTVEFNAVALLHSIE
RASLLSHEGRNNAVVKLSLQVADQKVVLNGNSPEIGNVEEDLSFKNLTGKDLSEISFNPDYMKDAL
SFGQTDITMSLTLPLRPFTLVPTEDGENFVQLITPVRTF

Or, upload file

Job Title

Enter a descriptive title for your BLAST se

Align two or more sequences

Choose Search Set

Database [UniProtKB/Swiss-Prot \(swissprot\)](#)

Organism [Optional](#)

Enter organism name or id--completions will

Enter organism common name, binomial,

Exclude [Optional](#)

Models (XM/XP) Non-redundant

Entrez Query [Optional](#)

Enter an Entrez query to limit search

検索結果

DNA polymerase III subunit beta [Lactobacillus salivarius]
Sequence ID: [WP_069468841.1](#) Length: 379 Number of Matches: 1

Range 1: 1 to 379 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
551 bits(1419)	0.0	Compositional matrix adjust.	263/379(69%)	323/379(85%)	0/379(0%)
Query 1	MKFNIRREFISALNNAQRAISSKTAIDVLTGIKLTLTSATKLTMTGSDADISIETVITED	60			
Sbjct 1	MKF+I+R+ FI LN+ QRAISSKT ID+LTG+K+ LS LT+TGS++DISIET+I+ MKFSIQRASFIKYLNDVQRAISSKTTIDILTLGKMDLSKDALTLTGSNSDISIETIISIA	60			
Query 61	NEKAALVIDEPGSVILPARFFNEIVKKLPEEMMTISVDARFQATITSGQAEFTINGLDSE	120			
Sbjct 61	DDNAALQIEQEGAVLPLPARFFSEIVKKLPEQTMTEINERFQATITSGSAEFTINGLNAE	120			
Query 121	TYPHLPETIADTKLVIAADILKQVINQTVIAVSTQESRPILTGHIHLSYNGELLAVATDS	180			
Sbjct 121	YPHLPEI++ +L + DILKQVI+QTVIAVS QESRPILTGHIH NGELLAVATDS EYPHLPETIDSKQLTVPGDILKQVISQTVIAVSNQESRPILTGHIHLVKNGLAVATDS	180			
Query 181	HRLSQRKITLADAGSEVYNIIVPGKSLNEMAKMLGDSTDNVEIRVAENQILFTFANISFY	240			
Sbjct 181	HRLSQR I L+ A VY++I+PGKSL+E++KM+ DS +N+EI+++ENQ LF N SFY HRLSQR I I KLSGANDAVYDVI I PGKSLSELSKMSISDSDENIEIQISENQALFILGNTSFY	240			
Query 241	SRLLEGNYPTDRILIPQSSETTVEFNAVALLHSIERASLLSHEGRNNAVVKLSLQVADQKV	300			
Sbjct 241	+RLLG YPDT+RLIP+ SE TV+FNAVALL SIERASLLSHEGRNNAVVKL++ + Q V TRLLEGYPDTERLIPKVS EITVDFNAVALLQS IERASLLSHEGRNNAVVKLAINPSAQS	300			
Query 301	VLNGNSPEIGNVEEDLSFKNLTGKDLSEISFNPDYMKDALSSFGQTDITMSLTLPLRPFTL	360			
Sbjct 301	VL+GN+PE+GNVEE+L FKNL G +LEISFNPDYMKDAL SFGQ++IT++ T PLRPPT+ VLSGNTPEVGNVEEELHFKNLEGESELEISFNPDYMKDALRSFGQSEITIAFTQPLRPPTI	360			
Query 361	VPTEDGENFVQLITPVRTF 379				
Sbjct 361	VPTED +NFVQLITPVRTF 379 VPTEDKDNFVQLITPVRTF 379				

Related Information

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

The screenshot displays the InterPro website interface. At the top, the InterPro logo and navigation menu are visible. The main content area is titled "InterProScan sequence search" and includes instructions on how to use the tool. A search form is present with a text input field containing a protein sequence and a "Submit" button. To the right of the search form, there are filter options for "Entry type", "Status", "Per-residue features", and "Colour by". The search results are displayed in a list format, showing protein family membership, homologous superfamilies, domains and repeats, and detailed signature matches. The results include various InterPro IDs and their corresponding protein families, such as "Asparagine synthase, glutamine-hydrolyzing" and "Nucleophile aminohydrolases, N-terminal".

InterPro
Protein sequence analysis & classification

Search InterPro...
Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Home Search Release notes Download About InterPro Help Contact **InterPro BETA**

By sequence By domain architecture

InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool.

Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFEIF with a maximum length of 40,000 amino acid long.

Please note that you can only scan one sequence at a time.

Alternatively, read [more about InterProScan](#) for other ways of running sequences through InterPro

Analyse your protein sequence

```
ELKKNLETDYEFTSSSDCEVLIPLYRKYGIETMVKMLDGEFSVLYDHLSTKIYAARDIIGIRPMF
GKNLLDLCREIHPFLPGHYDGEKIVAYHEPDLTPKMSTDDFETATHKIHLLVESVDQRASD
VCSIAARLPDPTKIRTFAGMDQNPIDLKYAREVADYLGTDHTEFIMTREDVLGVLREVIYTLT
KKIHETDLDKVIITGECSDLEFGYKYTFAPSPPEEFQKEAAKRLRELYMVDVLRADRCISANSL
VMSVDPDLKMNHYHKGKYLRLKAFEEGDWLPDRILMREKAAFSDAVGHSMVDDLKEYAESK
YRTPFTKESLLYRDIFFEEFYPGKADWIKDYWMPNRSWKSLESVTDPSARVLSNYGASGE
```

Advanced options
Submit Clear Example protein sequence

InterProScan

InterProScan is a sequence

Filter view on

Entry type

- Homologous superfamily
- Family
- Domains
- Repeats
- Site

Status

- Unintegrated

Per-residue features

- Residue annotation

Colour by

- domain relationship
- source database

Protein family membership

Asparagine synthase, glutamine-hydrolyzing (IPR006426)

Homologous superfamilies

Homologous superfamily

Domains and repeats

Domain

Detailed signature matches

- IPR029055 Nucleophile aminohydrolases, N-terminal
 - G3DSA:3.60.20.10
 - SSF56235 (N-termina...)
- IPR014729 Rossmann-like alpha/beta/alpha sandwich fold
 - G3DSA:3.40.50...
 - PIRSF001589 (Asn_sy...)
- IPR017932 Glutamine amidotransferase type 2 domain
 - PF13537 (GATase_7)
 - PS51278 (GATASE_TYPE_2)
- IPR033738 Asparagine synthase, N-terminal domain
 - cd00712 (AsnB)
- IPR001962 Asparagine synthase
 - PF00733 (Asn_synthase)
 - cd01991 (Asn_Syntha...)
- no IPR Unintegrated signatures
 - PTHR11772 (ASPARAGI...)
 - SSF52402 (glutamine n...)

21

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

```
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
DEBLINK    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
           Mol. Syst. Biol. 2, 2006 (2006)
```

source feature
配列全体についての記述

```
JOURNAL    Wisconsin, ... Blattner,F.R., ... Kosuge,T., ...
REMARK     Protein update by submitter
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.

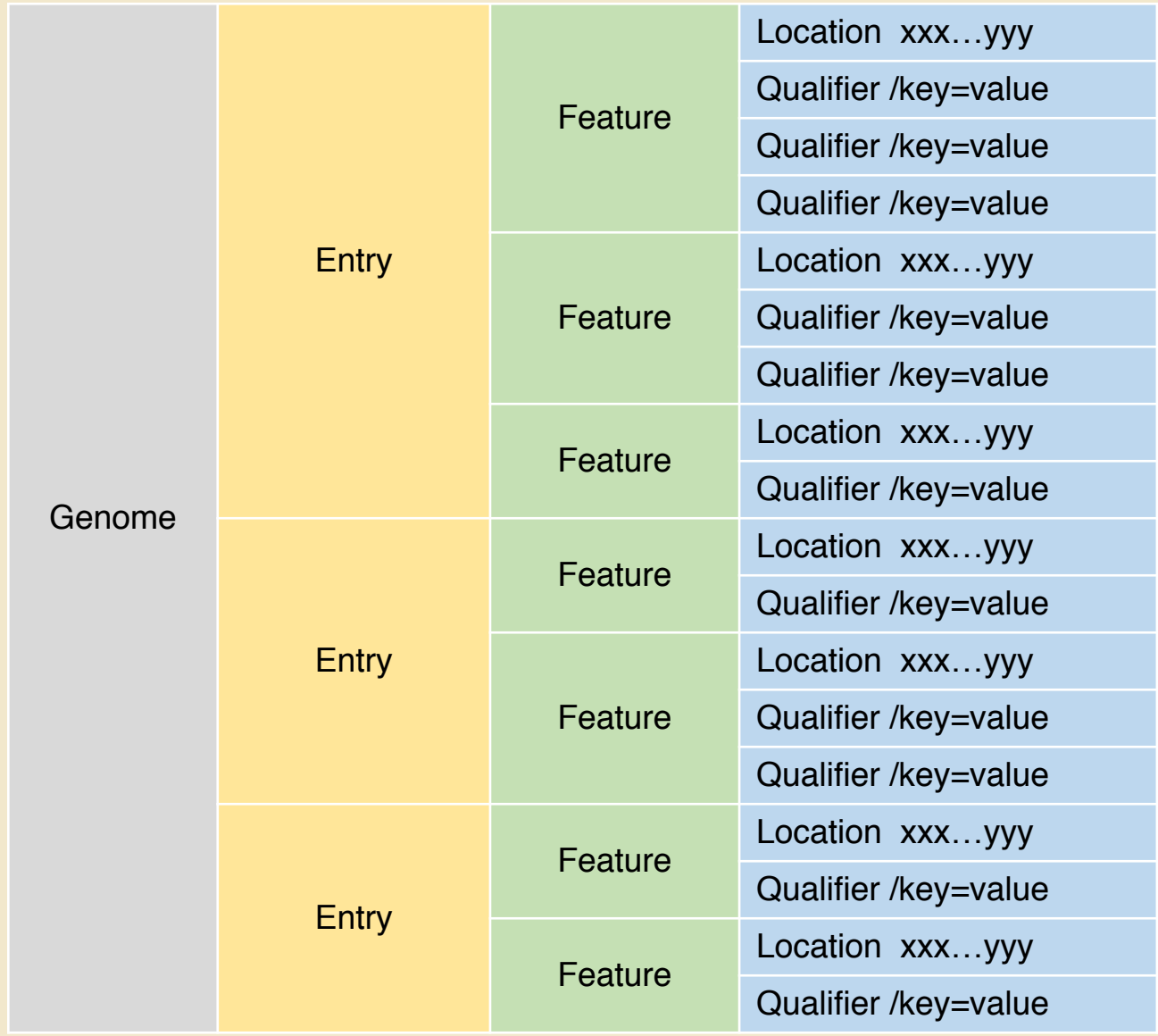
FEATURES   Location/Qualifiers
   source   1..4641652
            /organism="Escherichia coli str. K-12 substr. MG1655"
            /mol_type="genomic DNA"
            /strain="K-12"
            /sub_strain="MG1655"
            /db_xref="taxon:511145"
   gene     190..255 ← location
            /gene="thr." ←
```

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)

	FEATURES	Location/Qualifiers
source feature 配列全体についての記述	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 ← location /gene="thrL" ← qualifiers /locus_tag="b0001" ← qualifiers /gene_synonym="ECK0001" ← qualifiers /gene_synonym="JW4367" ← qualifiers /db_xref="EcoGene:EG11277"
gene feature*		
	CDS	190..255 /gene="thrL" /locus_tag="b0001" /gene_synonym="ECK0001" /gene_synonym="JW4367" /function="leader; Amino acid biosynthesis: Threonine" /note="GO_process: GO:0009088 - threonine biosynthetic process" /codon_start=1 /transl_table=11 /product="thr operon leader peptide" /protein_id="AAC73112.1" /db_xref="GI:1786182" /db_xref="ASAP:ABE-0000006" /db_xref="UniProtKB/Swiss-Prot:POAD86" /db_xref="EcoGene:EG11277" /db_xref="MCRISTITTTTITTTGNGAG"
CDS feature		
配列		4640881 4640881 cgcgtaataa g... 4640941 tatgcgtata acgattattc tgg... 4641001 gcgggcaatg aaaacgatgg ggtttagcga tctg... 4641061 ggagccagcc acccgctggg tcgcacatgg atctggtgat attattgata 4641121 tttcccgaca ttggctgaat cgttacacga tgtcgatttc actgtcgcca ccaactgcgcg 4641181 cagtccggcg aatatcatt actacgccac gccagttgaa ctggtgccgc tgtagagga 4641241 aaaatcttca tggatgagcc atgccgcgct ggtggttggc cgcgaagatt ccgggttgac 4641301 taacgaagag ttagcgttgg ctgacgttct tactggtgtg ccgatggtgg cggattatcc 4641361 ttcgctcaat ctggggcagg cggtgatggt ctattgctat caattagcaa cattaataca 4641421 acaaccggcg aaaagtgat caacggcaga ccaacatcaa ctgcaagctt tacgcgaacg 4641481 agccatgaca ttgctgacga ctctggcagt ggcagatgac ataaaactgg tcgactggtt 4641541 acaacaacgc ctggggcttt tagagcaacg agacacggca atggtgcacc gtttgcgca 4641601 tgatattgaa aaaaatatca ccaataaaa aacgccttag taagtatttt tc //

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



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

```
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-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			
contig01	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
```

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




RAST Rapid Annotation using
Subsystem Technology version 2.0

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

<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



DDBJ Fast Annotation and Submission Tool

[Start your project!](#) [Running 0 / Waiting 0]

Please see [FAQ](#) and [Sample Result](#) if this is your first visit.

<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..."
```

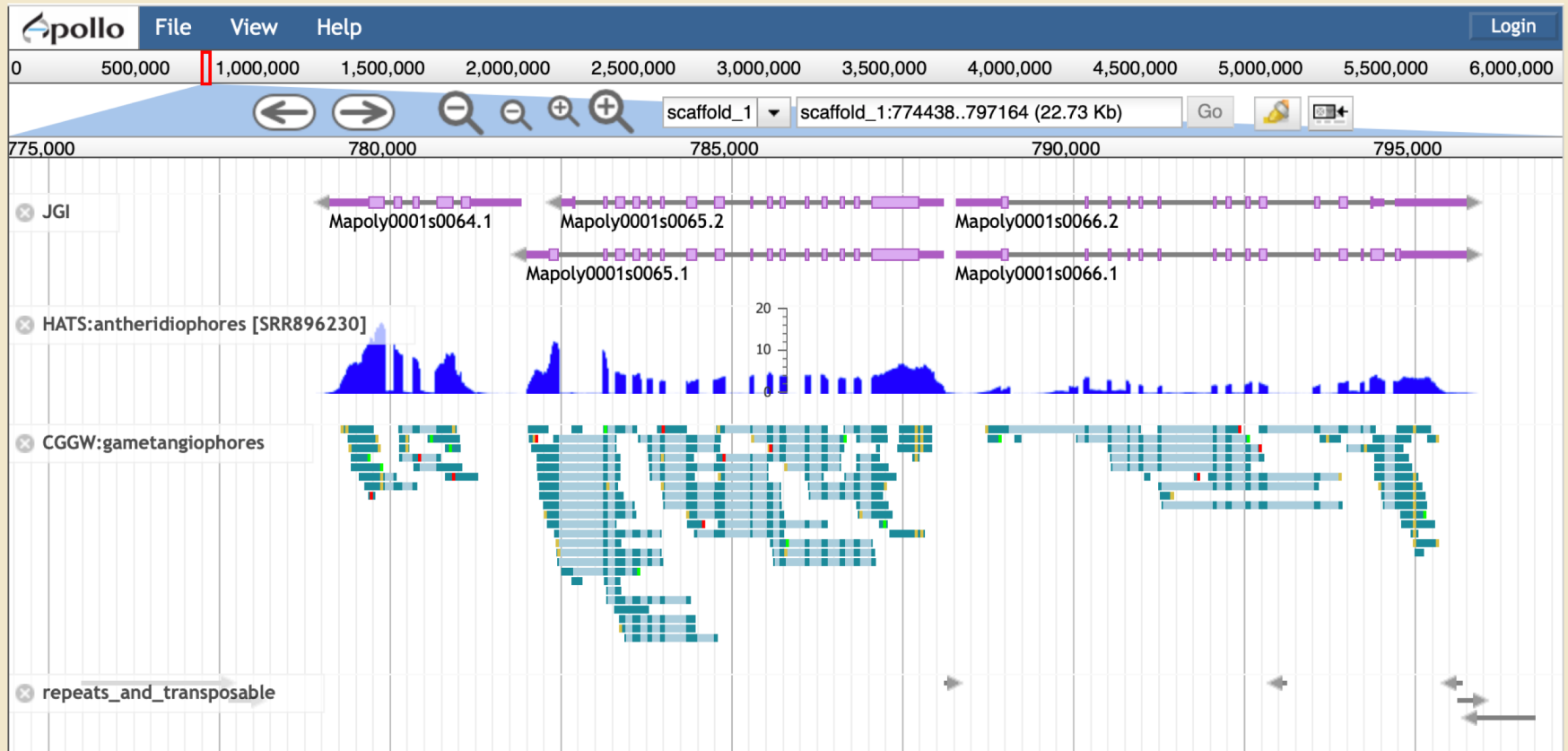
大腸菌 O26:H11 (accession no: GCF_000091005.1) の例

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
              HLKDGTCGLVEVEKGVLPQLEQPYVFIKRS DARTAPHGHVMVELVAELE..."
```

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



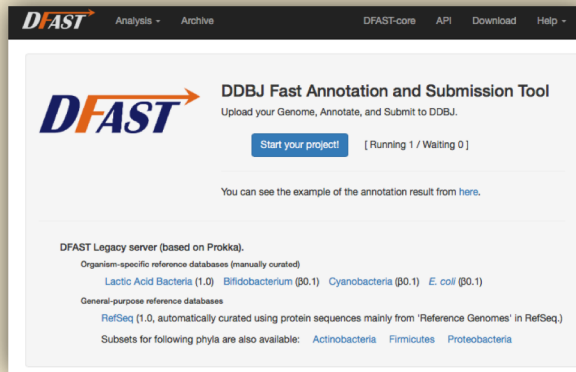
Prokaryotic genome annotation

Data submission to DDBJ

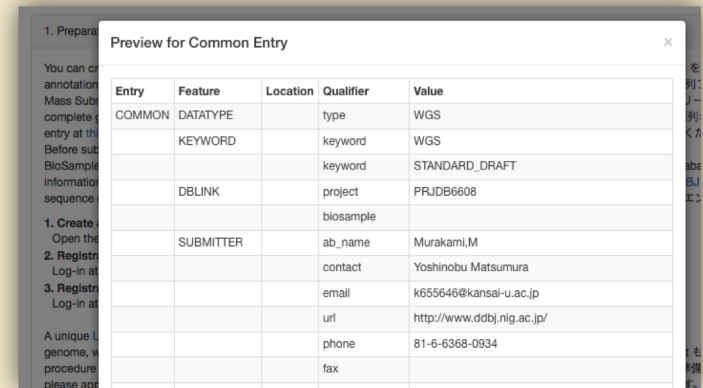
Fast, flexible, and powerful

Friendly for both beginners and experts

Graphical user interface for beginners

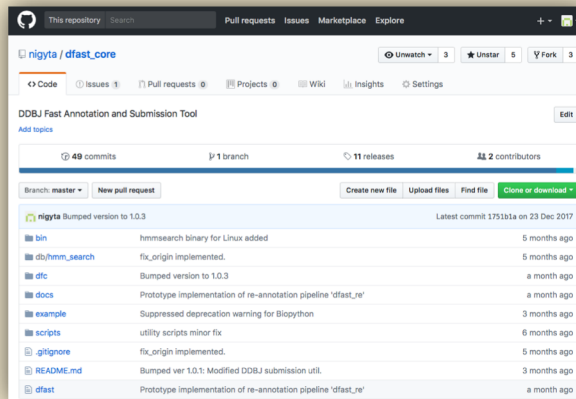


<https://dfast.nig.ac.jp>



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

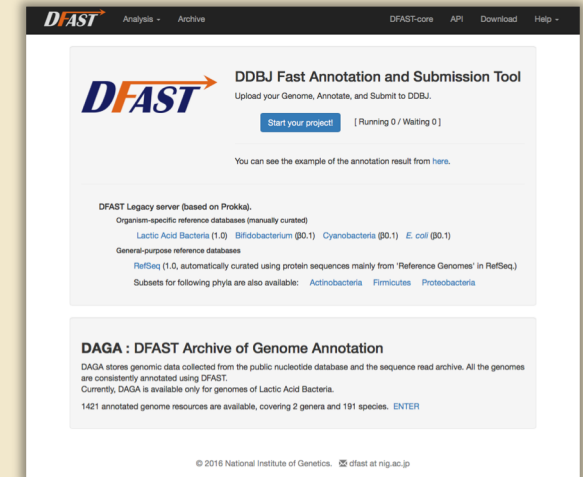
Release history

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..



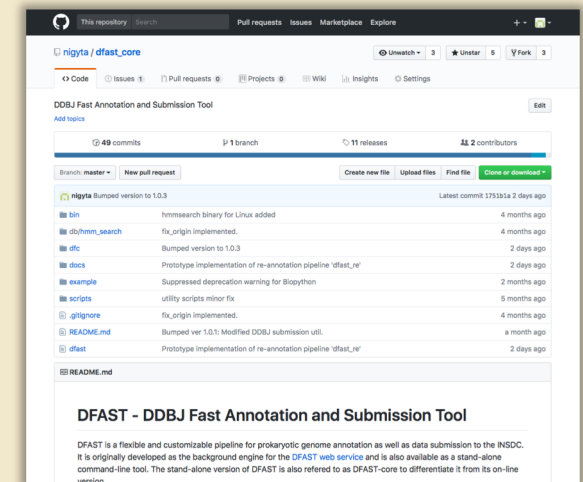
Web version (<https://dfast.nig.ac.jp>)

General-purpose reference database for all kinds of prokaryote

Original annotation engine *DFAST-core*

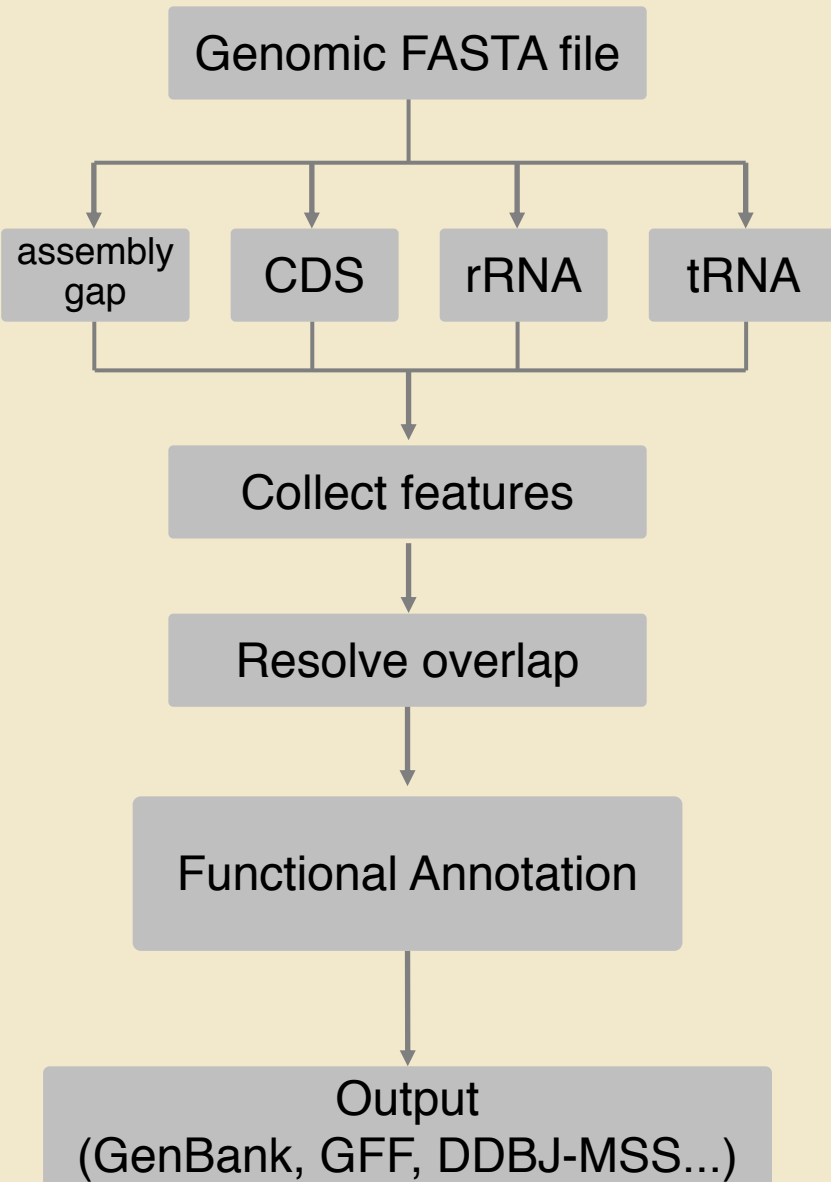
V 1.0 (2017.8) First major release

Tanizawa et al., Bioinformatics (2018)



Stand-alone version
Available at GitHub:[nigyta/dfast_core](https://github.com/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

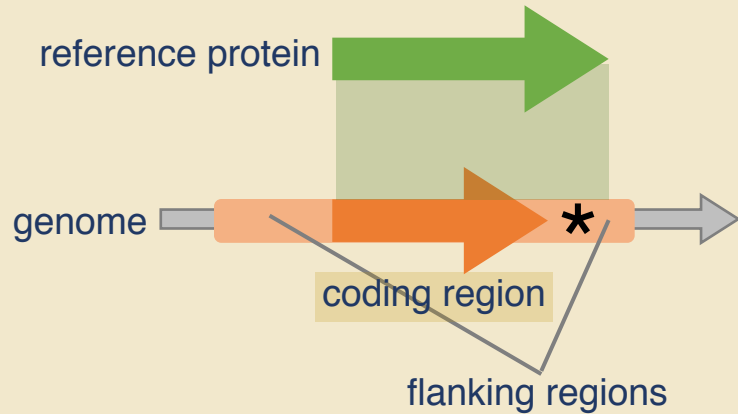
- Default database constructed from 120 representative genomes
- Optional organism-specific database

Pseudogene detection

Flexible and customizable ⁴⁰

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-E-VVDHGVTPKNPSTLDQII
QUERY ARKFAPAAFERKGNLSSIGEVKQFTPEESPSFGLSVLTALFPVLLLSIATIIY/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)

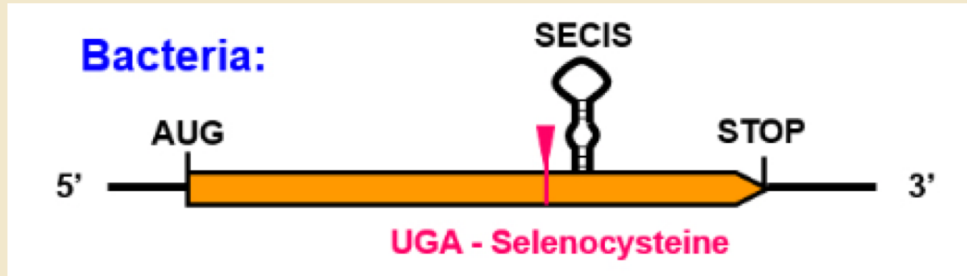


Figure from http://genomics.unl.edu/RBC_EDU/sp.html

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 0157:H7
str. Sakai) [pid:99.8%, q_cov:100.0%, s_cov:100.0%, Eval:0.0e+00]"
```

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

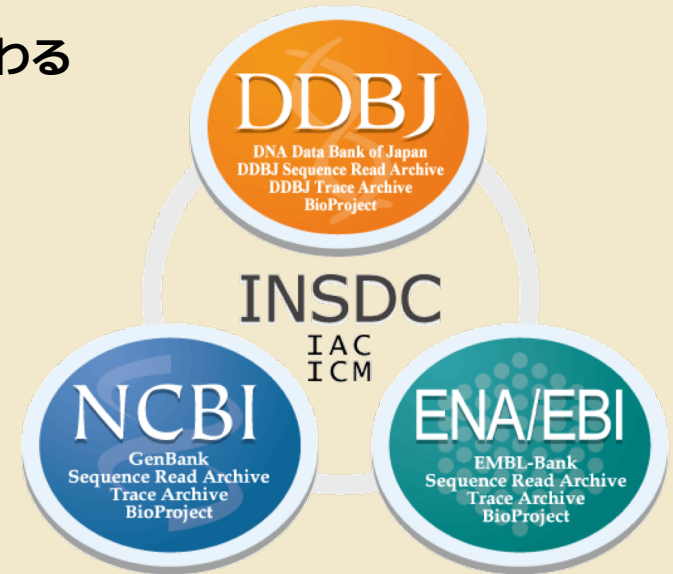
公共配列データベースの現状と諸問題

DDBJ は INSDC の一員として国際塩基配列DBの運用に携わる

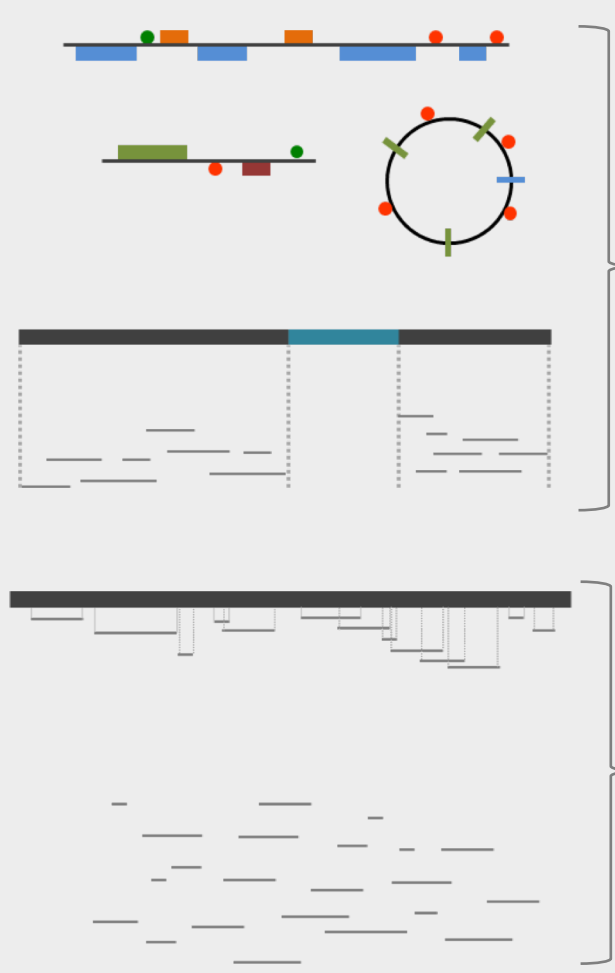
三極間で登録されたデータは同期・共有される

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		Sequence Read Archive
Sequence Read Archive	Trace Archive	European Nucleotide Archive (ENA)	Trace Archive
Annotated sequence	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



DDBJ

新規に決定（シーケンスまたはアSEMB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

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

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

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ムデータを登録

- 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」(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

The screenshot shows the top portion of a Nature journal article. The header includes the 'nature' logo and navigation links. The article title is 'Databases: Reminder to deposit DNA sequences' by Steven L. Salzberg. It includes publication details: 'Nature 533, 179 (12 May 2016) | doi:10.1038/533179a' and 'Published online 11 May 2016'. There are buttons for PDF, Citation, Reprints, Rights & permissions, and Article metrics. The subject terms are 'Databases', 'Publishing', 'Genomics', and 'Research data'. The main text begins with 'As members of the Advisory Committee to the International Nucleotide Sequence Database Collaboration (INSDC)...'.

The screenshot shows the top portion of a Science journal article. The header includes the 'Science' logo and AAAS logo. The article title is 'Reminder to deposit DNA sequences' under the 'LETTERS' section. The authors listed are Mark Blaxter¹, Antoine Danchin², Babis Savakis³, Kaoru Fukami-Kobayashi⁴, Ken Kurokawa⁵, Sumio Sugano⁶, Richard J. Roberts⁷, Steven L. Salzberg^{8,*}, and Chung-I Wu^{9,10}. It includes the publication date 'Science 11 May 2016' and the DOI '10.1126/science.aaf7672'. There are social media share icons for Facebook, Twitter, and Google+.

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 Primary database	二次データベース 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	<ul style="list-style-type: none">• DDBJ/ENA/GenBank• DOR/ArrayExpress/GEO• DRA/ERA/SRA• EVA·DGVa/dbSNP·dbVar• PDB	<ul style="list-style-type: none">• RefSeq• Ensembl• Expression Atlas• ChIP-Atlas• 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。

予算・人員の不足

データ量の増大

問題のある登録情報

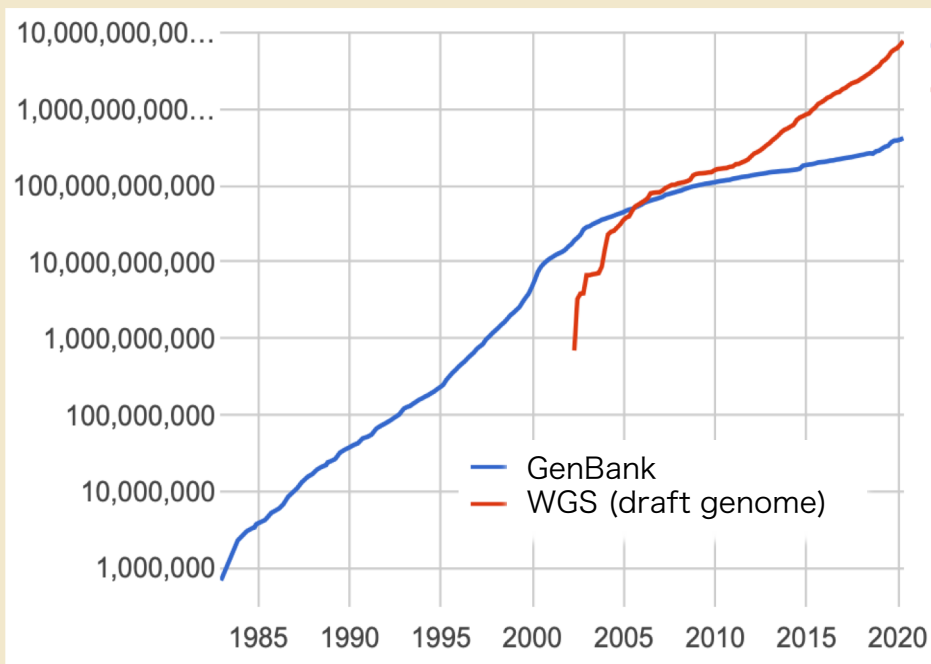
誤った注釈情報の伝播

生物種名の誤り

低品質な配列・コンタミネーション

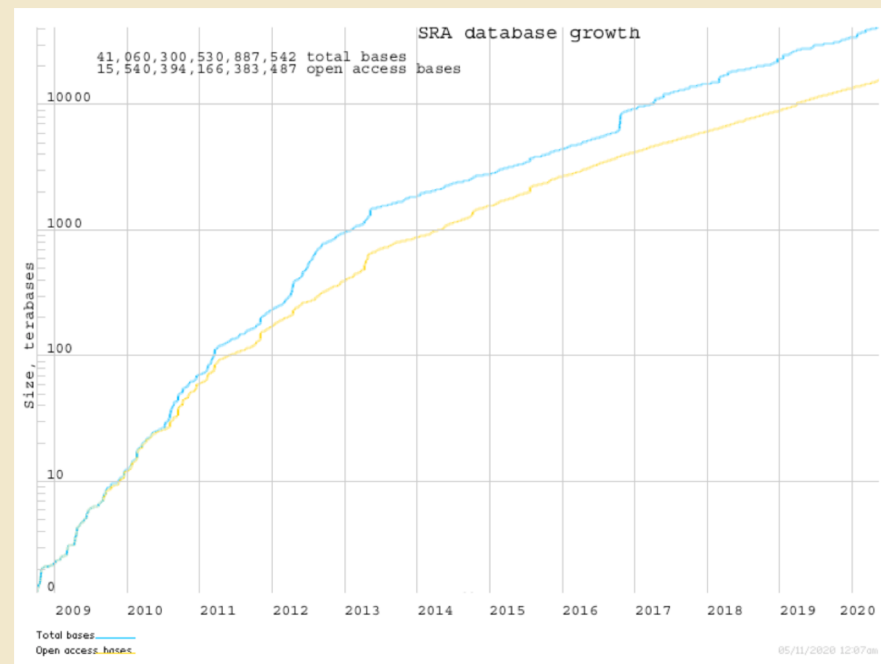
増大するデータ量

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

actin-related protein → similar to actin-related protein
→ similar to similar to actin-related protein

UniProtKB - G2Y5W9 (G2Y5W9_BOTF4)

Display

Entry

Publications

Feature viewer

None

Function

BLAST

Align

Format

Add to basket

History

Protein | Submitted name: **Similar to similar to actin-related protein R07**

Gene | **BofuT4_P112340.1**

Organism | *Botryotinia fuckeliana* (strain T4) (Noble rot fungus) (*Botrytis cinerea*)

Status | **Unreviewed** - Annotation score: 0.0000 - Protein inferred from homologyⁱ

Names & Taxonomy

2004年から続く Excel 問題

	A	B	C
1	SEPT2		
2			
3			
4			



	A	B	C
1	2-Sep		
2			
3			
4			

日付データに変換されてしまう

2004年 Zeeberg et al.

2016年 Zeeberg et al.

BMC Bioinformatics



Correspondence

Open Access

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

Barry R Zeeberg¹, Joseph Riss^{1,2}, David W Kane³, Kimberly J Bussey¹, Edward Uchio⁴, W Marston Linehan⁴, J Carl Barrett² and John N Weinstein^{*1}

Address: ¹Genomics & 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, ²Laboratory of Biosystems and Cancer, CCR, Bldg 37 Rm 5032, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, ³SRA International, 4300 Fair Lakes CT, Fairfax, VA 22033 USA and ⁴Urologic 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@dtpx2.ncicrf.gov

* Corresponding author †Equal contributors

COMMENT

Open Access

Gene name errors are widespread in the scientific literature



Mark Ziemann¹, 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. kefir* と同じ種である

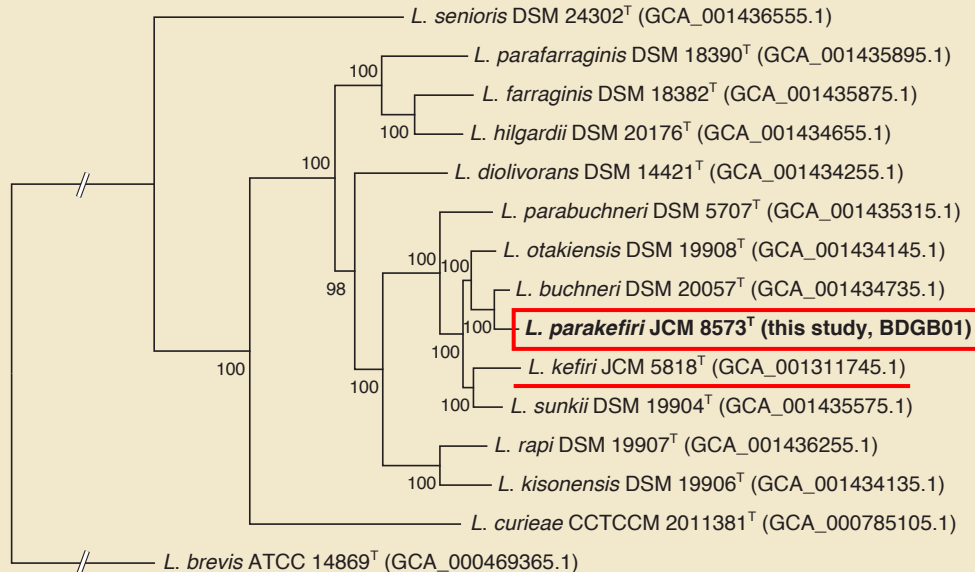
Sun, et al. Nat Commun. (2015)

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 parakefir*.

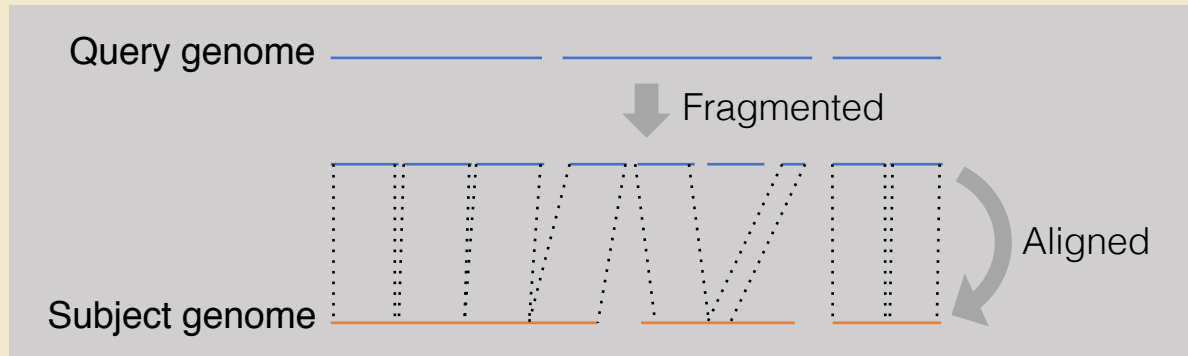


Tanizawa, et al. BMFH. (2017)

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



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 Meyer¹⁰ 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

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 ▶

◀ 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。

INTERNATIONAL
JOURNAL OF SYSTEMATIC
AND EVOLUTIONARY
MICROBIOLOGY

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



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

Stacy Ciufu,* 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 at the National Center for Biotechnology Information

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

Ciufu 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)

Intraspecific 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)

The screenshot shows the DAGA web interface with the following filters and options:

- Group: --- all ---
- Quality Rating: ☆☆☆☆ × ☆☆☆ × ☆☆☆ × ☆☆☆ × ☆ ×
- Show only representative genomes:
- Genus: Lactobacillus ×
- Species: amylovorus × paraplantarum ×
- Subspecies: --- disabled ---
- Show Optional Columns: Original Name, BioProject, BioSample, Assembly Level, Completeness, Contamination
- Show: 10 entries

The table below shows the first 10 entries, with the 'Rating' column highlighted in red. Red annotations point to the 'Rating' column and the 'Note' column.

ID (click for detail)	Organism Name (curated)	Type Status	GC%	Total length (bp)	No. of Seqs.	CDSs	Rating	Note
ERR387503	Lactobacillus amylovorus DSM 16698		37.8%	1,979,726	131	1,947	☆☆☆☆	
ERR433486	Lactobacillus paraplantarum LMG_16673		43.7%	3,297,581	249	3,069	☆☆☆☆	
GCA_000182855.2	Lactobacillus amylovorus GRL 1112		38.1%	2,126,674	3	2,166	☆☆☆☆☆	
GCA_000191545.1	Lactobacillus amylovorus 30SC		38.1%	2,097,766	3	2,093	☆	The organism name was amended. (100% ANI value against L. amylovorus) [PMID:22386464]
GCA_000194115.1	Lactobacillus amylovorus GRL1118		38.0%	1,977,087	3	1,957	☆☆☆☆☆	
GCA_000442765.1	Lactobacillus amylovorus CAG:719		38.1%	1,816,655	55	1,838	☆☆☆☆	
GCA_000469115.1	Lactobacillus paraplantarum AY01		43.7%	3,315,973	169	3,106	☆	The organism name was amended. (99.6% ANI value against L. paraplantarum)
GCA_000758145.1	Lactobacillus paraplantarum L-ZS9		43.9%	3,119,721	40	2,900	☆☆☆☆	
GCA_000980505.1	Lactobacillus amylovorus N54.MGS-719		38.4%	1,821,790	151	1,783	☆☆☆☆	The organism name was inferred from ANI result. (97.1% against L. amylovorus)
GCA_001006945.1	Lactobacillus amylovorus unknow		38.0%	1,879,641	41	1,852	☆☆☆☆	The organism name was inferred from ANI result. (97.0% against L. amylovorus)

completeness and contamination values calculated by CheckM

Species names inferred from ANI calculation

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

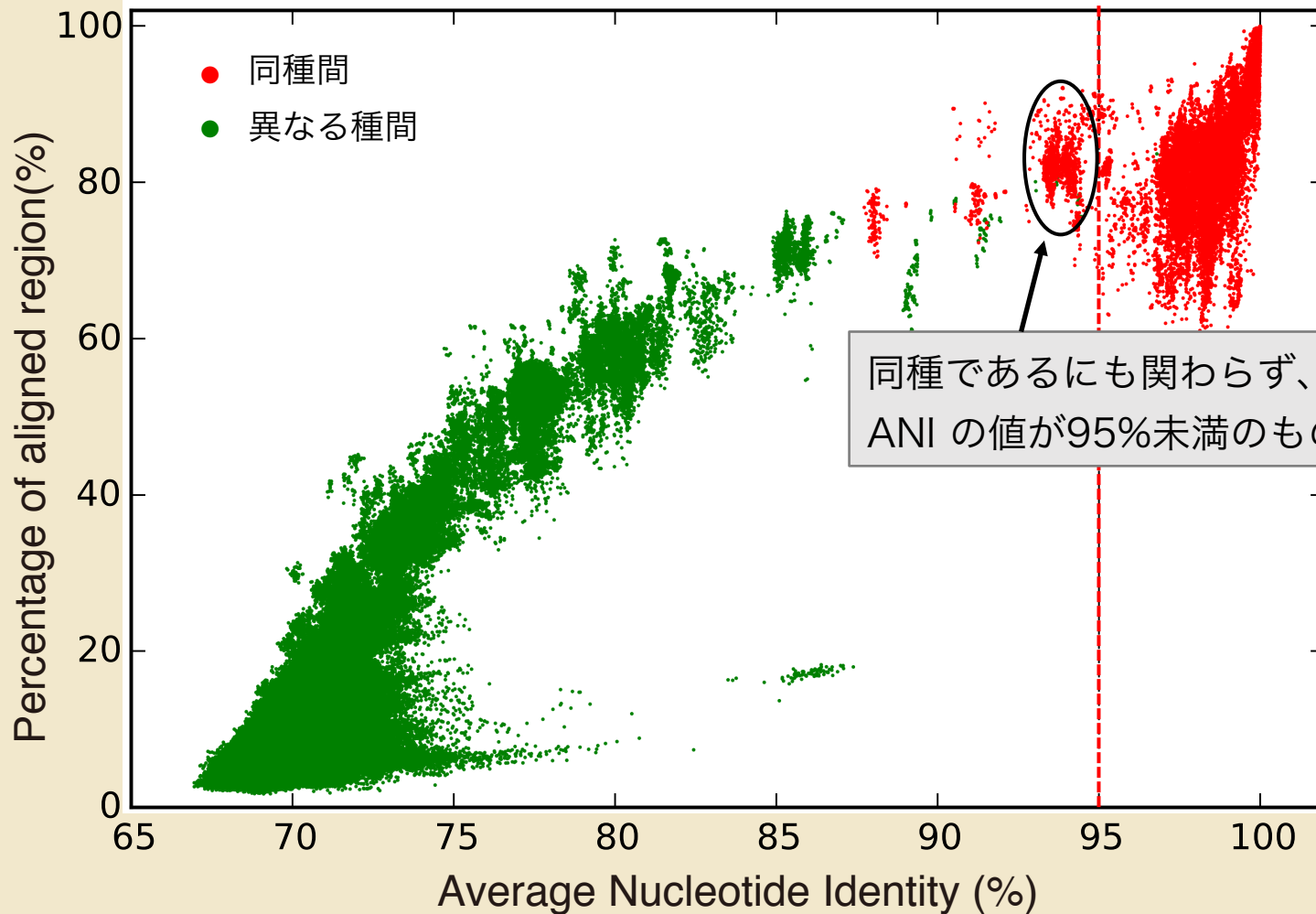
*L. casei*とその近縁種の間でのANIの値

乳酸菌シロタ株も *paracasei*

	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けた新種の乳酸菌

約1400件の乳酸菌ゲノムの中で ANI を計算した結果



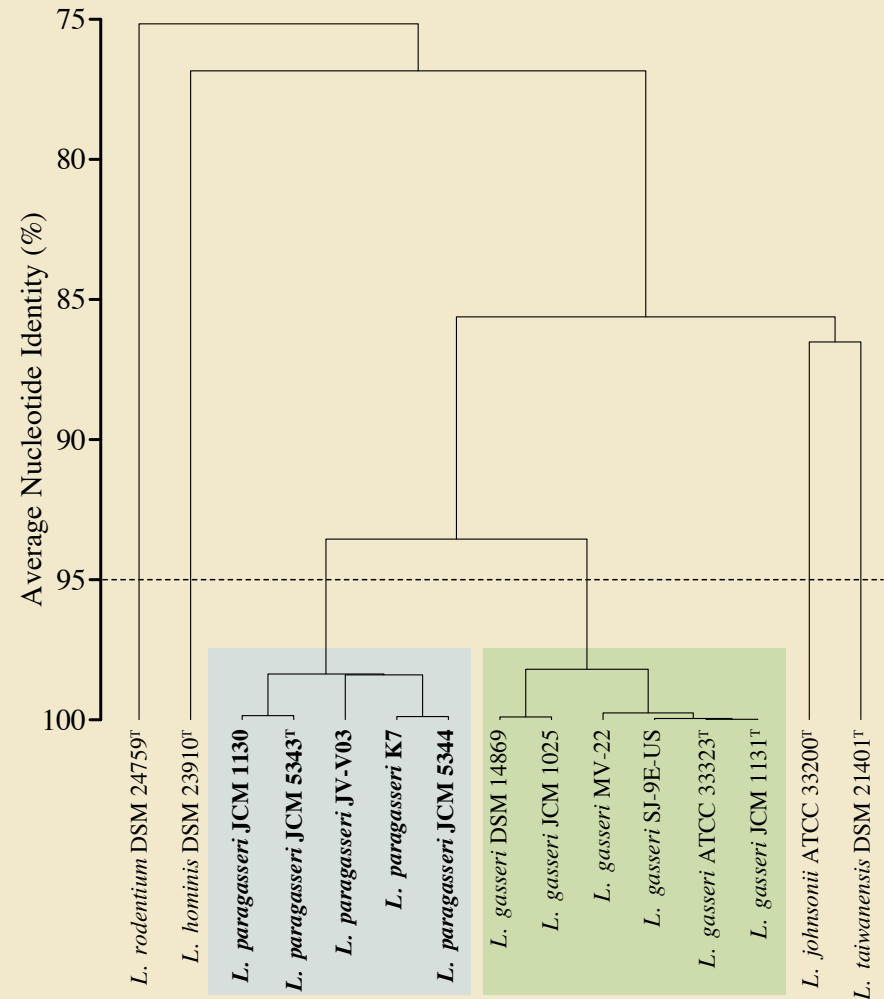
新種乳酸菌 *Lactobacillus paragasseri* の提唱

(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?



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