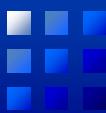


ゲノムデータベース ゲノムアノテーション

阿部 貴志

新潟大学大学院自然科学研究科
工学部工学科知能情報プログラム



自己紹介

- 専門は、バイオインフォマティクスです。特に、一括学習型自己組織化マップを始めとする機械学習を活用し、比較ゲノムやメタゲノム解析のための情報学的手法の開発とその応用研究を行なっています。
 - <http://bioinfo.ie.niigata-u.ac.jp>
- ゲノムアノテーションとしては、大腸菌ゲノムの再アノテーションプロジェクト(NAR, 2006)、微生物ゲノム再アノテーションワークフローの構築とDB化(DNA Res., 2006)、マニュアルキュレーションによるtRNA遺伝子DB構築・公開(tRNADB-CE)、タイレリア原虫ゲノムプロジェクト(mBio, 2012)、南極コケ坊主微生物プロジェクトなどに参加。

Nucleic Acids Research, 2006, Vol. 34, No. 1 1–9
doi:10.1093/nar/gkj405

Escherichia coli K-12: a cooperatively developed annotation snapshot—2005

Monica Riley*, Takashi Abe¹, Martha B. Arnaud², Mary K.B. Berlyn³, Frederick R. Blattner⁴,

Roy R. Chaud¹ DNA RESEARCH 13, 245–254 (2006)

Takehiko Kosi¹

Kenneth E. Ru⁵

David Wishart¹

Exploration and Grading of Possible Genes from 183 Bacterial Strains by a Common Protocol to Identification of New Genes: Gene Trek in Prokaryote Space (GTPS)

Takehiko KOSUGE,¹ Takashi ABE,¹ T_C
Yutaka MARUYAMA,^{1,2} Jun MASHIMA,¹
Satoshi FUKUCHI,¹ Satoru MIYAZAKI,¹
Hideaki SUGAWARA^{1,*}

Comparative Genome Analysis of Three Eukaryotic Parasites with Differing Abilities To Transform Leukocytes Reveals Key Mediators of *Theileria*-Induced Leukocyte Transformation

Kyoko Hayashida,^a Yuichiro Hara,^b Takashi Abe,^c Chisato Yamasaki,^d Atsushi Toyoda,^d Takehiko Kosuge,^e Yutaka Suzuki,^f Yoshiharu Sato,^g Shuichi Kawashima,^h Toshiaki Katayama,^h Hiroyuki Wakaguri,ⁱ Noboru Inoue,^j Keiichi Homma,^e Masahito Tada-Umezaki,^k Yukio Yagi,^k Yasuyuki Fujii,^l Takuuya Habara,^b Minoru Kanehisa,^m Hidemi Watanabe,ⁿ Kimihito Ito,^o Takashi Gojobori,^{p,e} Hideaki Sugawara,^e Tadashi Imanishi,^b William Weir,^q Malcolm Gardner,^q Arnab Pain,^r Brian Shiels,^p Masahira Hattori,^t Vishvanath Nene,^s and Chihiro Sugimoto^a

mBio, 3, 00204-12, 2012

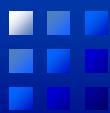
tRNADB-CE: tRNA gene database curated manually by experts

Takashi Abe¹, Toshimichi Ikemura¹, Junichi Sugahara², Akio Kana², Yasuo Ohara¹, Hiroshi Uehara¹, Makoto Kinouchi³,

Shigeohiko Kanaya⁴, Yuko Yamada¹, Akira Muto⁵, Hachiro Inokuchi¹

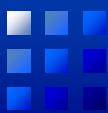
¹ Nagahama Inst. of Bio-Sci. and Tech., ² Keio Univ., ³ Yamagata Univ., ⁴ NAIST, ⁵ Hirosaki Univ.

<http://trna.nagahama-i-bio.ac.jp>; NAR, 2009, 2011



本日の講習会の内容

1. ゲノムデータベース
 - 微生物統合DB MicrobeDB.jp
 - 植物統合DB PGDBj
2. ゲノムアノテーションの概要
 - アノテーション手法の紹介
3. アノテーションパイプライン構築について
 - 実例を元に、構築の考え方について
4. 演習
 - 任意のゲノム配列に対し、実際にアノテーションしてみよう
- 資料について
 - 解析ツール等の紹介では、今後利用してもらいたく、論文ではなく、URLを記載しています。
 - データベースは、DBと省略しています。



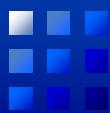
ゲノムデータベースの現状

- ゲノム分野では、塩基配列解読技術の急速な進歩によって解読されるゲノム種が急激に増加
- 國際論文誌Nucleic Acids Research誌で年1回DB特集号が公開されているが、年間100以上のDBが新規に公開。現在は、1500以上のDBが公開。
- なぜ、そんなにもDBは増えていくのか？
 - バイオデータの異種性
 - 分子種の違い
 - DNA, タンパク質、その他、低分子化合物など
 - 研究コミュニティの違い
 - 微生物/植物/動物
 - 生化学/生物物理/バイオインフォマティクス
 - 基礎生物学/臨床・医療
 - 共通表記法の欠如
 - DNA塩基配列データベース(1次DB)から、上記の対象ごとにDBを構築・公開(2次DB)しているが、公開するデータの内容、記載方法は開発者ごとに異なる

そもそもどのDBにアクセスすれば必要な情報を取得できるのか？



統合DBの必要性



ゲノムデータベースの現状

- 微生物統合データベース



<http://microbedb.jp>

国立遺伝学研究所: 黒川 顕, 中村保一, 神沼英里, 森 宙史, 藤澤貴智, 東 光一

基礎生物学研究所: 内山郁夫, 千葉啓和, 西出浩世

東京工業大学: 山田拓司

千葉大学: 高橋弘喜, 矢口貴志

- 植物統合データベース:



<http://pgdbj.jp>

かずさDNA研究所: 原田大士朗, Jeffrey A. Fawcett, 平川英樹,
磯部祥子, 田畠哲之

大阪大学: 市原寿子, 中谷明弘

謝辞:

本講義にあたり, 各開発グループの森先生と平川先生よりスライドデータ
をご提供頂きましたこと, 深く感謝申し上げます.

Microbe DB^{.JP} integrates lots of data related to microbes.

Especially, we integrates the microbial data that can be linked to **genomes**, since 2011



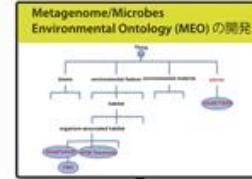
Microbe DB^{.JP}

<http://microbedb.jp/>

Microbe DB.jp

MicrobeDB.jp プロジェクトでは様々な微生物学上の知識を、ゲノム情報を軸として遺伝子、系統、環境の3つの軸に沿ってセマンティックウェブの技術を駆使して整理統合し、幅広い分野での微生物学の発展に貢献することの出来るデータベースの構築を目指しています。

Ontology



Gene

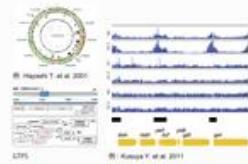
Taxon

Environment



Ortholog: MBGD

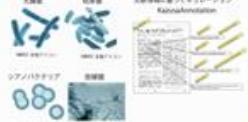
オーソログデータ



Genome: RefSeq

オミックスデータ

Annotation: TogoAnnotation



モデル微生物の高品質アノテーションデータ



Taxonomy:
NCBI Taxonomy

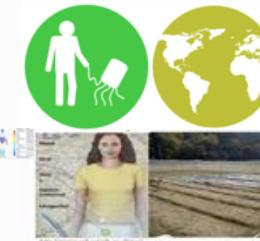
系統分類データ



Culture Collection:
NBRC/JCM

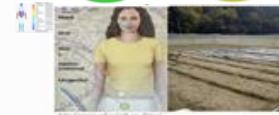


菌株データ
菌種保存情報(切歎条件含む)



Metadata:
INSDC DRA

環境のメタデータ



Metadata:
INSDC DRA

環境のメタデータ



Metagenome:
INSDC DRA

メタゲノムデータ

Red color indicates our collaborators

MicrobeDBの基本的な使い方

The screenshot shows the MicrobeDB search interface. A search bar at the top contains "Escherichia coli". Below it, a search result for "Escherichia coli" is displayed, including environment (hot spring), taxonomy (Enterococcus faecalis, Streptomyces avermitilis), gene (psbA), and ID (29). Navigation tabs for Text, Analysis, and Statistics are visible.



検索結果: 必要な情報を一覧表示可能

This screenshot shows the detailed search results for "Escherichia coli". It includes a Taxon Description section with scientific name (Escherichia coli), synonyms (Bacillus coli, Bacterium coli commune, Enterobacteriaceae, Bacterium coli), and a Taxonomic Hierarchy table. The table lists various taxonomic levels and their corresponding Taxonomy IDs, such as 131597 (Bacteria) and 562 (Escherichia coli). Below this is a Genome Information (High quality) section, which lists three genome entries: NC_001051 (Escherichia coli O157:H7 str. Sakai plasmid pO157), NC_005855 (Escherichia coli O157:H7 str. Sakai chromosome), and NC_001127 (Escherichia coli O157:H7 str. Sakai plasmid pOSAKI).

メタゲノムサンプルの検索、比較が可能

This screenshot shows the MicrobeDB analysis tools interface. It features sections for Analysis tools, Comparison between the metagenome samples (Metagenome samples), Comparison between the environments (Environment), Comparison between the taxa (Taxonomy), and Correlation between sample metadata and taxonomic/functional analysis (Metadata).

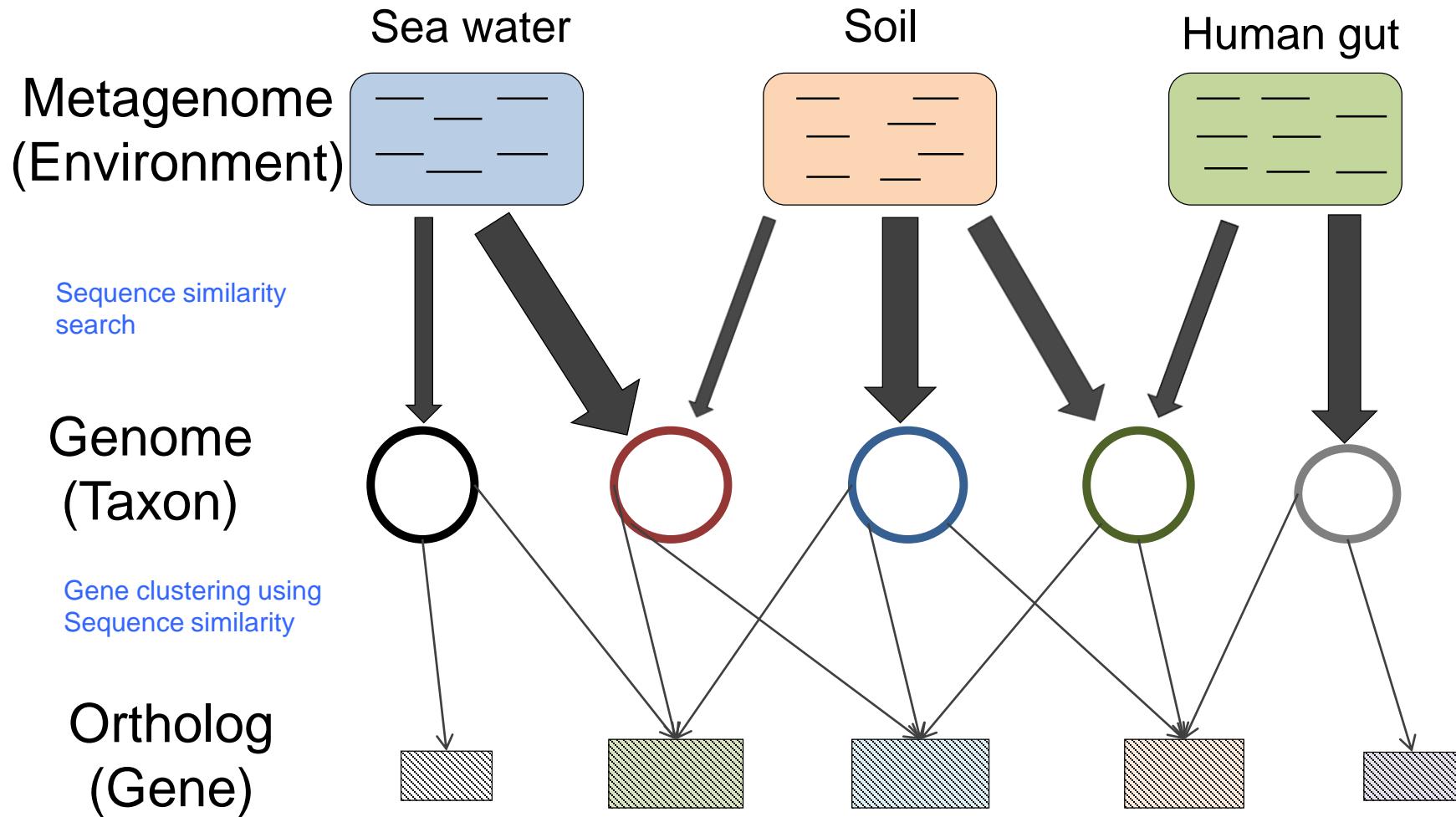
ゲノム解読済みの株情報

This screenshot shows the genome information section. It includes a "Draft genome" section with assembly details (ASM59970v1, S7380) and a "Strains" section listing various Escherichia coli strains with their names, isolation sources, and temperatures. The strains listed include JCM 1326, JCM 16274, JCM 16275, JCM 15425, JCM 39671, and JCM 39714.

株(単離、温度)情報の情報

MicrobeDBの基本的な使い方

メタゲノムデータの統合: 系統情報とオルソログ情報で実施

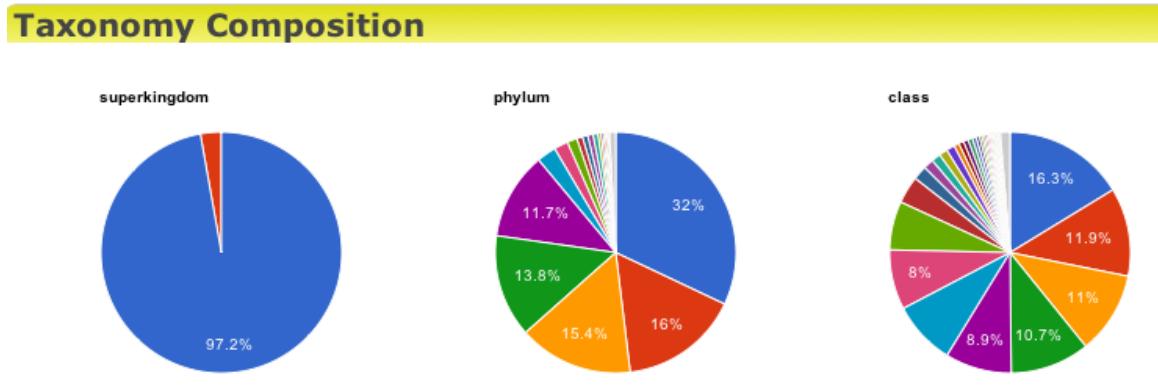




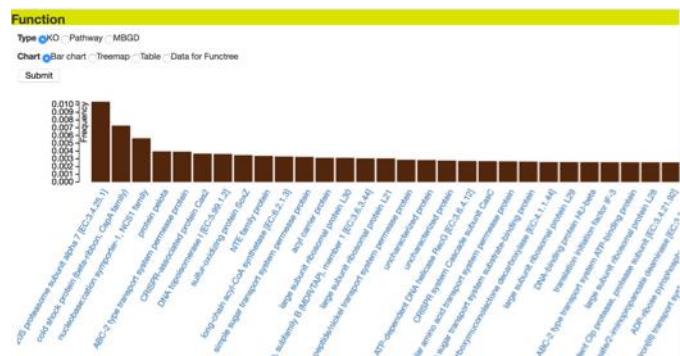
MicrobeDBの基本的な使い方

メタゲノムサンプルに対する検索結果

- Taxonomic composition
of a metagenome sample



- Functional composition of a metagenome sample



MicrobeDBの基本的な使い方

メタゲノムサンプルに対する検索結果

[top page](#) | [hmori](#) [Sign out](#)



Sample Comparison

Selected samples:

[150914164924C0](#) [150914165143C1](#) [150914165601C3](#) [150914170106C6](#) [150914170719C12](#) [150914171251C24](#) [ERS017997](#) [ERS033075](#)

Taxonomic rank: Genus

[Compare samples](#)

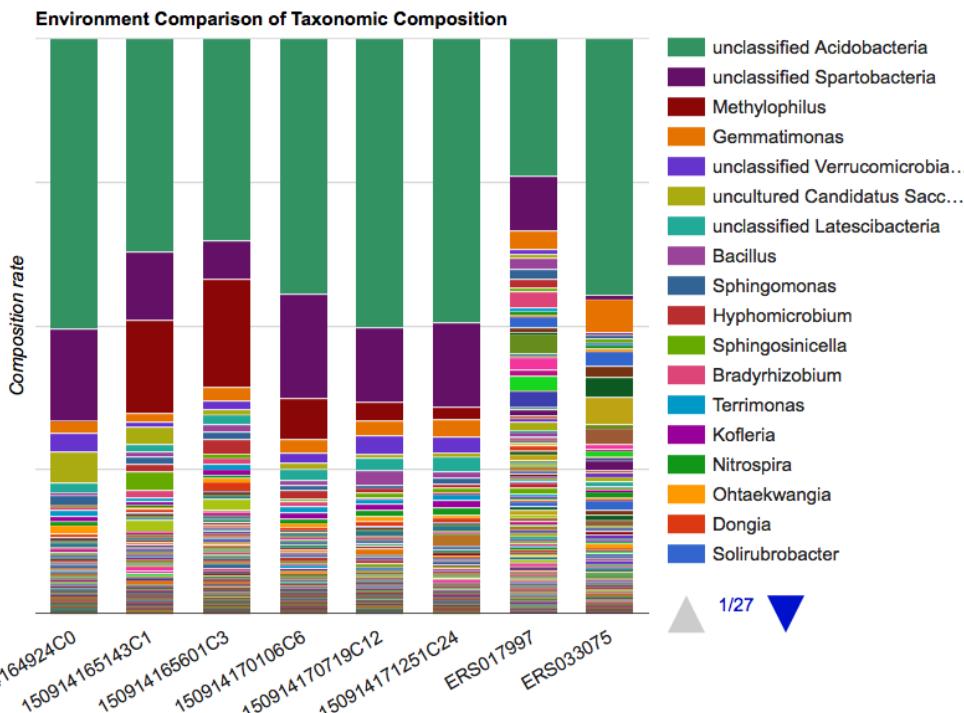
[Taxonomic composition \(bar\)](#)

[Taxonomic composition \(heatmap\)](#)

[Diversity index](#)

[Hierarchical clustering](#)

[PCoA](#)



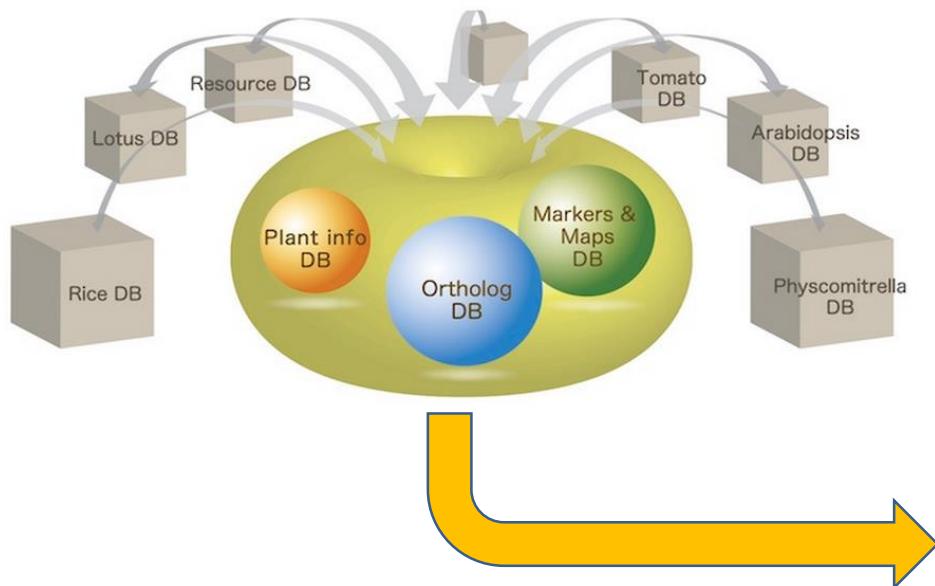


植物ゲノム関連データベースの統合化

植物ゲノム関連のデータベース：
1,000以上

Plant Genome DataBase Japan
(PGDBj)

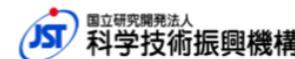
<http://pgdbj.jp>





PGDBj(Plant Genome DataBase Japan; <http://pgdbj.jp>)

植物ゲノム統合化データベース(統合化推進プログラム)



横断検索
→



リソースDB

緑色植物: 40種、約114万配列
ラン藻: 213種、約80万配列

API
→



15種
154万件



6種、85万2千件



約900個体のカンキツ在来種



DNAマーカー: 65種、約26万件
QTL: 45種、約1万4千件

横断検索
(メタボローム)
↓

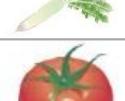
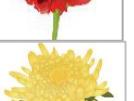
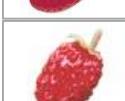
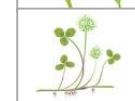
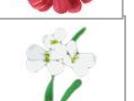


KNapSAcK MassBase



文献からのマニュアルキュレーション
・DNAマーカー(主にSSR、SNP)
・QTL

PGDBj: マーカー情報を公開している80の植物種

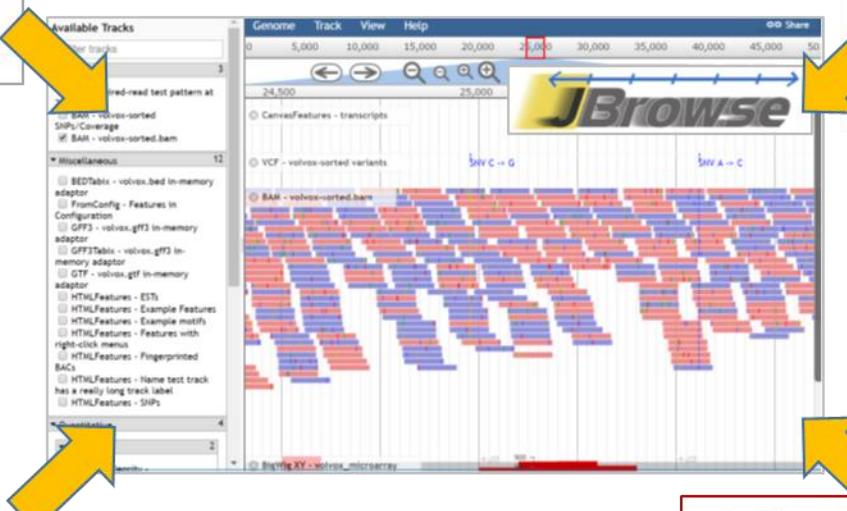
	タマネギ		セイヨウア布拉ナ		メロン		イネ		ミナトカモジグサ		ナツメヤシ
	ネギ		トウガラシ		リンゴ		コムギ		ルベラナズナ		ヒノキ
	テンサイ		キマメ		モモ		アワ		ドイツトウヒ		アンボレラ
	キャベツ		ヒヨコマメ		セイヨウナシ		ソルガム		カナダトウヒ		クリ
	ハクサイ		ダイズ		ナシ		オオムギ		テーダマツ		ウメ
	キュウリ		ラッカセイ		パパイア		トウモロコシ		ポプラ		ジャトロファ
	サツマイモ		タルウマゴヤシ		ブドウ		キャッサバ		コットンウッド		ワタ
	レタス		ミヤコグサ		バナナ		アサガオ		スギ		カカオ
	ダイコン		ソラマメ		キウイフルーツ		カーネーション		ギニアアブラヤシ		トウゴマ
	トマト		タバコ		オレンジ		キク		セキザイユーカリ		ゴマ
	ナス		オランダイチゴ		ウンシュウミカン		ヒヤクニチソウ		ユーカリ		モウソウチク
	ジャガイモ		エゾヘビイチゴ		アカクローバ		ミヤマハタザオ		イヌカタヒバ		スサビノリ
	ホウレンソウ		スイカ		シロクローバ		シロイヌナズナ		チヤ		ヒメツリガネゴケ
									アサ		クラミドモナス

PGDBj: ゲノムブラウザへの多型情報の集約

ゲノム配列(Lj3.0)
Pseudomolecule(5本)
総延長: 447.4 Mb
遺伝子モデル
48,106個



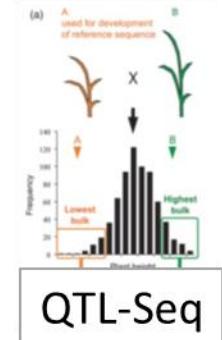
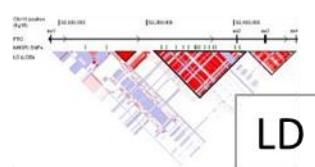
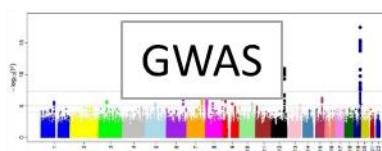
例) ミヤコグサ (*Lotus japonicus*)



SSR: 1,073
CAPS/dCAPS: 82

ゲノムワイド多型情報
(NCBI SRA)

文献からのキュレーション



PGDBj: JBrowseに集約させる植物種

かずさDNA研究所で解読した15種

Pseudomolecule



ミヤコグサ



サブクローバ



シバ

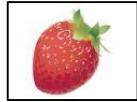


アウトウ

スキヤフォールド



ダイコン



イチゴ



バラ野生種



ナス



サツマイモ
野生種



カーネーション



ソバ



ユカリ



ジャトロファ



イチジク



キヌア

他研究機関で解読された27種

Pseudomolecule



キウイフルーツ



ミヤマハタザオ



セイヨウアブラナ



キャベツ



レタス



ルベラナズナ



スイカ



メロン



キュウリ



ヒヨコマメ



ダイズ



タルウマゴヤシ



アカクローバ



カカオ



バナナ



ゴマ



ミナトカモジグサ



アワ



エゾヘビイチゴ



リンゴ



モモ



オレンジ



ポプラ



トウガラシ



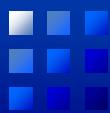
トマト



ジャガイモ



ブドウ



本日の講習会の内容

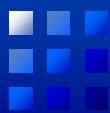
1. ゲノムデータベース
 - 微生物統合DB MicrobeDB.jp
 - 植物統合DB PGDBj
 2. ゲノムアノテーションの概要
 - アノテーション手法の紹介
 3. アノテーションパイプライン構築について
 - 実例を元に、構築の考え方について
 4. 演習
 - 任意のゲノム配列に対し、実際にアノテーションしてみよう
- 資料作成にあたって
 - 解析ツール等の紹介では、今後利用してもらいたく、論文ではなく、URLを記載しています。
 - 謝辞
 - アノテーションの資料作成にあたり、かずさDNA研究所の平川英樹先生に、スライドデータ、ならびに、NGSデータを用いたゲノムアノテーションについての情報をご提供頂きましたこと、深く感謝申し上げます。



アノテーション(Annotation)

- 一般にはある文章に対する注釈をいうが、生物学では**あるデータの属性を記述すること**を指す。データとしては、塩基配列（ゲノムやトランスクリプトーム）、アミノ酸配列、遺伝子、標本などさまざまなタイプがある。これらの付加情報を生物学的知識を駆使して作成する人をアノーター（Annotator）と呼ぶ。DNAや蛋白質などのデータベース構築は必須の作業である。

岩波生物学辞典第5版より



ゲノムアノテーション

aattcgataaatctctggtttattgtcagttatggttccaaatgcctttgctgtatatactcacagcataactgttatatacaccagggggcggaatgaaagcgtaacggccag gcaacaagaggtgtttgatctcatccgtatcacatcagccagacaggtatgccgcccac gcgtcgaaaaatcgcgcagcggtttgggttccgttccccaaacgcggctgaagaacatctaaggcgctggcacgcaaaggcgttattgaaattgtttccggcgcatcagcgggattcg tctgttgcatggaaagaggaagaagggttgcgcgttgttaggtcgtgtggctgccggtaacc acttctggcgcaacacagcatattgaaggtcattatcaggtcgatccttcatttcaagc gaatgctgatttcctgctgcgcgtcagcggatgtcgatgaaaagatatcggcattatgatggacttgctggcagtgcataaaactcaggatgtacgtaacggcaggtcggtgc acgtattgtacgaaagttaccgttaagcgcctgaaaaaaaaacaggcaataaagtcaact gttgccagaaaatagcgagttaaaccaattgtcggtgaccttcgtcagcagagcttcac cattgaaggctggcggttgggttattcgcaacggcactggctgtaacatctctgacccgcgtgccgcctggcggttgcgtttttcatctcttcattcaggcttgcgtcatggcatttcacttcatctgataaaag

塩基配列中にコードされている遺伝子領域やその制御領域情報を付与していく

赤:ORF(大腸菌lexA)

緑:lexAリプレッサーのコンセンサス結合部位

緑:-10と-35のプロモーター領域



ゲノムプロジェクトの流れ

1. シークエンシング

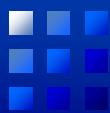
- ショットガンシークエンス
 - Sanger法、次世代シークエンサー
- アセンブル
- フィニッシング

2. アノテーション

- 遺伝子領域(rRNA, tRNA, mRNAなど)の同定
- 遺伝子領域・機能の予測
- などを行いゲノム配列へ生物学的知見を加えていく。

3. ゲノム機能の解明

- 比較ゲノム解析など

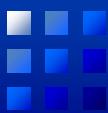


アノテーションの流れ

1. リピート配列の検出とマスク
2. 機能性RNA(rRNA, tRNA, ncRNA)の予測とマスク
3. 遺伝子領域予測
4. 遺伝子機能予測

主に計算機による処理

マニュアルアノテーション



1. リピート配列の検出とマスク

- 主に真核生物では、リピート配列があると正確な遺伝子領域予測が難しくなるために、リピート配列をマスクする必要がある。
- リピート配列の検出とマスク手法
 - RepeatMasker (<http://www.repeatmasker.org/>)
 - 既知リピート配列との相同性検索、ならびに、タンデムリピートの検出が可能。
 - 大抵の真核生物ゲノムプロジェクトで採用されている。
 - 出力結果として、リピート配列検出結果に加え、マスク済み配列も取得可能。
 - その他として、RepeatScout (<https://bix.ucsd.edu/repeatscout/>)も併用される場合がある



Services

- [RepeatMasking](#)
- [Protein-based RepeatMasking](#)
- [Pre-Masked Genomes Search](#)
- [Genome Analysis and Downloads](#)
- [Server Queue Status](#)
- [FEAST – Gene Prediction](#)

Welcome!

RepeatMasker is a program that screens DNA sequences for interspersed repeats and as well as a modified version of the query sequence in which all the annotated repeats have been masked. Sequence comparisons in RepeatMasker are performed by one of several popular

Latest News

If you would like to keep up with news and announcements relating to RepeatMasker, yo

2. 機能性RNAの予測とマスク

- 機能性RNA予測プログラムとして、主なものを以下に示す。
 - リボソームRNA(rRNA)
 - 原核生物の場合: 28S; ~2300塩基, 16S; ~1500塩基, 5S; ~100塩基が数セット
 - RNAmmer (<http://www.cbs.dtu.dk/services/RNAmmer/>)
 - Barrnap (<https://github.com/tseemann/barrnap>)
 - tRNA(76~120塩基)
 - tRNAscan-SE (<http://lowelab.ucsc.edu/tRNAscan-SE/>)
 - Aragorn (<http://mbio-serv2.mbioekol.lu.se/ARAGORN/>)
 - tmRNAも予測可能。一方で、tRNAscan-SEに比べて、過分に候補数を予測する傾向がある。
 - ncRNA
 - Infernal (<http://eddylab.org/infernal/>)
- これらは、配列も保存性が比較的高く、どの生物種でも共通のプログラムで予測可能。

rRNAやtRNA遺伝子領域は、遺伝子領域予測の際に、誤って予測されることもあり、遺伝子予測の際には予測した領域をマスクする。



3. 遺伝子領域の予測(1)

【目的】ゲノムDNAからタンパク質をコードしている可能性の最も高い領域を同定する。

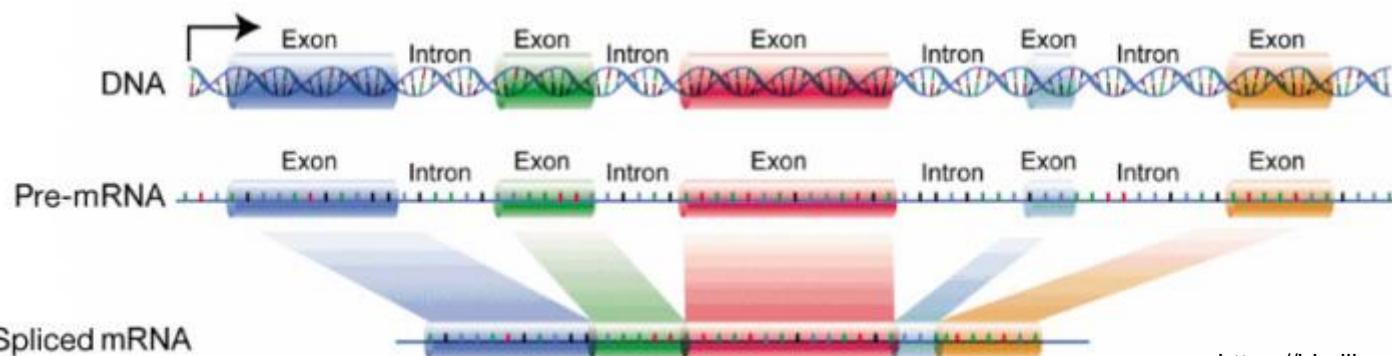
• 原核生物

プロモーター



ターミネーター

• 真核生物



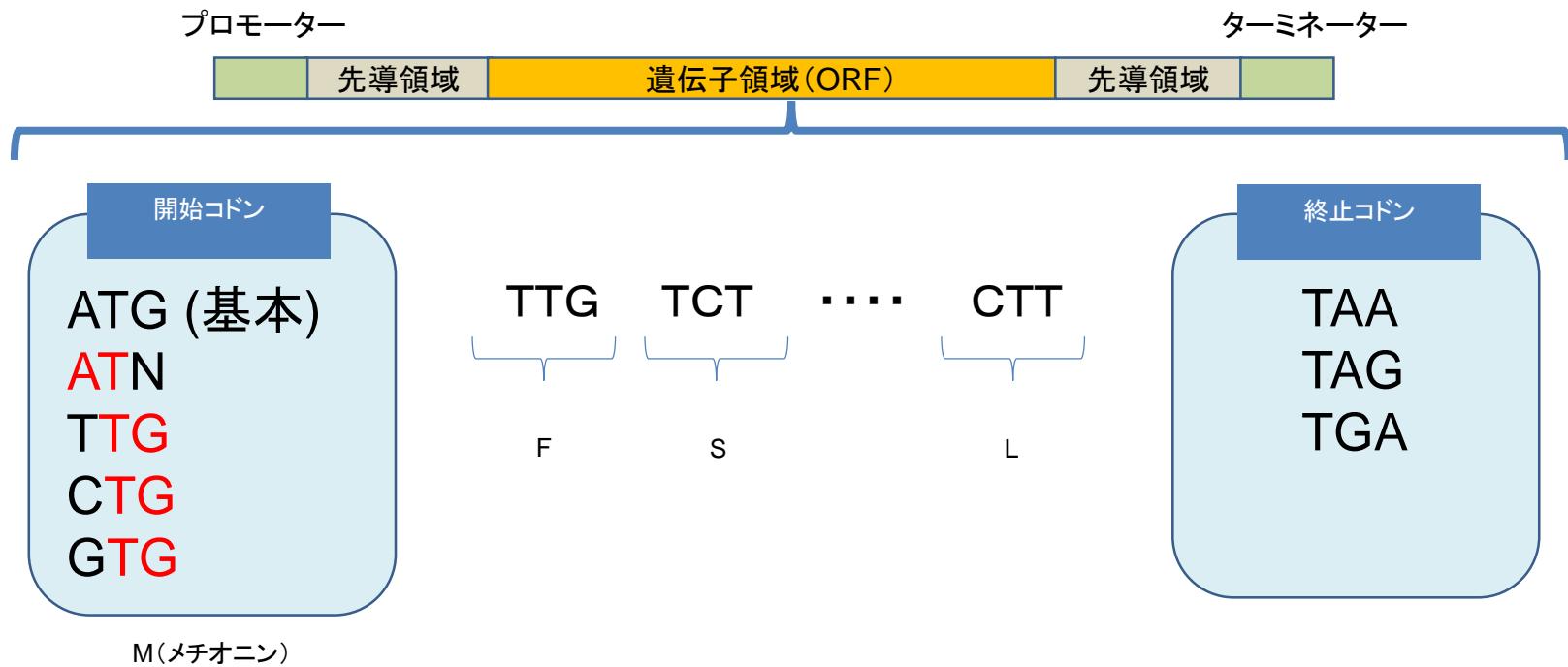
<https://bio.libretexts.org>

原核生物と真核生物では、遺伝子構造に大きな違いがあり、それぞれに特化した予測法の利用が必須



遺伝子領域の予測(2)：原核生物

• 原核生物の遺伝子構造



単純な遺伝子構造のため、遺伝子領域中とその周辺に存在する特徴的な配列パターンを識別できれば、予測可能。
特徴的な配列パターンとしては、主に以下が挙げられる。

- 遺伝子領域中のコドン使用(2連続コドン、8連続塩基など)パターン
- 開始コドン近傍の塩基配列パターン
- これらの特徴のある生物種の既知遺伝子セットから認識できるように訓練(学習)し、得られた学習データを用いて、類似パターンを持つ領域を予測する。
- 学習に用いられるアルゴリズムとしては、隠れマルコフモデル(HMM)、動的計画法、回帰分析などがある。

遺伝子領域の予測(3)：原核生物

遺伝子予測プログラムの種類と精度

Table 1. Comparison of Some Features for Gene Finders

Gene finders	CDS	stable RNA ^a	FA ^b	developed for :	Output files format
Prodigal	y	n	n	bacterial & archaeal	GBK, GFF or SCO
GeneMark.hmm	y	n	n	prokaryotes	algorithm-specific
GeneMark	y	n	n	prokaryotes	algorithm-specific
GenMarkS	y	n	y	prokaryotes	algorithm-specific
RAST	y	y	y	bacteria and archarea	GTF,GFF3,GenBank,EMBL
JCVI Annotation Service	y	y	y	prokaryotes	algorithm-specific
AMIGene	y	n	n	prokaryotes	EMBL, GenBank, GFF
Glimmer3	y	n	n	prokaryotes	algorithm-specific
EasyGene	y	n	n	prokaryotes	GFF2
Maker	y	n	y	small eukaryotes and prokaryotes	GFF3
Augustus	y	n	y	eukaryotes	GTF (similar to gff). GFF

^a stable RNA refers to rRNA, tRNA, tmRNA, RNA Component of RNaseP

^b FA stands for functional annotations i.e. mRNA, operons, promoters, terminators, protein-binding sites, DNA bends

Table 2. Results from Testing the Gene Finders on *P.a. LESB58*

Gene Finder	# Genes	# Genes on the + Strand	# Genes on the - Strand	#Correct Genes	% Correct Genes (compared to the Original)	% Correct Genes from (from all found genes)
Original	6061	2993	3067	6061	100,00%	100,00%
Prodigal	6055	3014	3041	5286	89,14%	87,30%
FGenesB	6197	3094	3103	5070	85,50%	81,81%
Glimmer3.0	6276	3100	3176	5043	85,04%	80,35%
GeneMarkS	6100	3043	3057	5006	84,42%	82,07%
JCVI	6270	3098	3172	5036	83,10%	80,32%
GeneMarkHMM	6129	3055	3074	4920	82,97%	80,27%
Rast	6297	3116	3181	4940	81,52%	78,45%
MED	7475	3708	3767	4747	80,05%	63,51%
Maker with model	6149	3065	3084	4588	75,71%	74,61%
Maker	5884	2904	2980	4370	72,11%	74,27%
Augustus	5268	2587	2681	3529	59,51%	66,99%
AMIGene	6154	3077	3077	2967	50,03%	48,21%
EasyGene	3150	0	3150	2570	43,34%	81,59%

P.a. LESB58: Pseudomonas aeruginosa LESB58

Correct Genesとは、5', 3'末端の両方一致を指す。

予測プログラム/パイプラインの出力結果は、まちまちで、プログラム固有な出力形式の場合、ファイル形式の変更も必要である。完全一致での予測精度は、概ね80%以上可能であるが、予測プログラムによっては、開始位置(5'末端)が異なる場合が多い。



遺伝子領域の予測(3-1)：原核生物

Prodigalの性能比較結果

nk Annotations

3'末端が一致/5'-3'が一致した遺伝子数

Organism	Genbank Genes with no Joins	Prodigal 1.20	Prodigal 1.20 +TiCo	Prodigal 1.20 +TriTisa	GenemarkHMM 2.6	Glimmer 3.02	EasyGene 1.2	MED 2.0
<i>Escherichia coli K12</i>	4268	4118/3823 (96.5%/ 89.6%)	4118/3779 (96.5%/88.5%)	4118/3778 (96.5%/88.5%)	4122/3685 (96.6%/86.3%)	4076/3563 (95.5%/ 83.5%)	3977/3565 (93.2%/ 83.5%)	4102/ 3711 (96.1%/ 86.9%)
<i>Halobacterium salinarum</i>	2110	2062/1857 (97.7%/ 88.0%)	2062/1809 (97.7%/85.7%)	2061/1790 (97.6%/84.8%)	2042/1676 (96.7%/79.4%)	2054/1609 (97.3%/ 76.2%)	2018/1692 (95.6%/ 80.2%)	2008/ 1469 (95.1%/ 69.6%)
<i>Natronomonas pharaonis</i>	2661	2630/2398 (98.8%/ 90.1%)	2630/2358 (98.8%/88.6%)	2630/2348 (98.8%/88.2%)	2624/2251 (98.6%/84.6%)	2622/2220 (98.5%/ 83.4%)	2548/2271 (95.7%/ 85.3%)	2586/ 1953 (97.2%/ 73.4%)
<i>Bacillus subtilis</i>	4174	4113/3705 (98.5%/ 88.8%)	4113/3678 (98.5%/88.1%)	4113/3679 (98.5%/88.1%)	4136/3713 (99.1%/89.0%)	4102/3569 (98.3%/ 85.5%)	3977/3578 (95.3%/ 85.7%)	4127/ 3596 (98.9%/ 86.2%)
<i>Aeropyrum pernix</i>	1699	1670/1430 (98.3%/ 84.2%)	1670/1363 (98.3%/80.2%)	1670/1353 (98.3%/79.6%)	1672/1364 (98.4%/80.3%)	1671/1317 (98.4%/ 77.5%)	1652/1389 (97.2%/ 81.8%)	1689/ 1309 (99.4%/ 77.1%)
<i>Synechocystis PCC6803</i>	3171	3146/2587 (99.2%/ 81.6%)	3146/2364 (99.2%/74.6%)	3146/2447 (99.2%/77.2%)	3124/2337 (98.5%/73.7%)	3123/2236 (98.5%/ 70.5%)	3053/2288 (96.3%/ 72.2%)	3126/ 2192 (98.6%/ 69.1%)
<i>Pseudomonas aeruginosa</i>	5565	5514/5038 (99.1%/ 90.5%)	5514/4885 (99.1%/87.8%)	5514/4821 (99.1%/86.6%)	5484/4698 (98.5%/84.4%)	5491/4705 (98.7%/ 84.5%)	5522/4761 (99.2%/ 85.5%)	5292/ 4539 (95.1%/ 81.6%)

遺伝子予測プログラムの精度

MetaGeneAnnotatorの性能比較結果

Noguchi et al., DNA Res., 15, 387-396, 2008.

MetaGeneAnnotator : <http://metagene.nig.ac.jp/>

Hyatt et al., BMC Bioinformatics, 2010, 11:119

Prodigal: <https://github.com/hyattpd/Prodigal>

Species	GC%	RBS%	MGA		GeneMarkS		Glimmer3	
			Sn (exact) (%)	Sp (%)	Sn (exact) (%)	Sp (%)	Sn (exact) (%)	Sp (%)
<i>S. marinus</i>	35.7	85.4	99.4 (87.8)	94.5	99.6 (87.2)	92.5	99.8 (87.6)	90.8
<i>C. acetobutylicum</i>	30.9	93.7	98.3 (92.1)	96.1	98.5 (74.1)	92.8	98.0 (90.9)	94.5
<i>F. nodosum</i>	35.0	90.2	99.6 (91.2)	94.8	99.8 (90.6)	92.8	99.7 (91.1)	94.0
<i>L. lactis</i>	35.3	81.1	98.5 (88.0)	95.1	98.9 (88.4)	92.7	98.2 (86.2)	93.2
<i>D. radiodurans</i>	67.0	47.9	97.8 (63.5)	93.6	96.3 (56.7)	93.1	96.5 (58.3)	92.1
<i>A. caulinodans</i>	67.3	64.8	99.2 (66.2)	95.4	98.8 (61.5)	95.8	98.6 (63.6)	93.6
Average			98.7 (80.2)	95.0	98.5 (74.3)	93.4	98.2 (78.0)	93.5

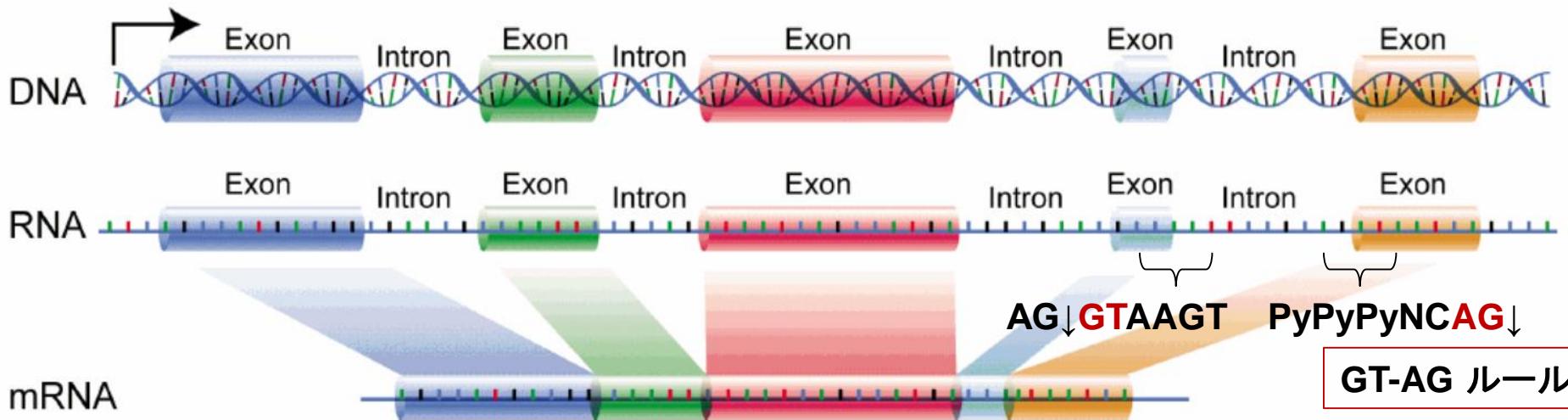
RBS%, the RBS ratio (the proportion of genes having representative RBSs); Sn, sensitivity to genes; (exact), sensitivity to start codons; Sp, specificity.

- 原核生物に限れば、どの予測プログラムも遺伝子領域の予測はかなりの精度(90%以上)で予測可能である。
- 現在、主に使用される予測ソフトとして、事前の学習セットを必要としない **prodigal**、**MetaGeneAnnotator** が挙げられる。
- 使用する予測プログラムでの予測結果の違いもあり、最終的には、人手をかけて、チェックを行うことも大事である。



遺伝子領域の予測(4): 真核生物

<https://bio.libretexts.org>



スプライスサイトの予測。原核生物よりも困難。

エキソンとイントロンの境界における特定の配列が存在するが、エキソンやイントロンの中にも存在し、例外も多く存在するため、正確な境界を特定することが困難。

原核生物で利用した隠れマルコフモデルなどの統計的手法とニューラルネットワークのような機械学習法を組み合わせた予測法などがあるが、予測精度はあまり高くない。上手くいっても60%~70%程度。

必要な情報(RNA-seq, EST, cDNA)や複数の予測プログラムを組み合わせた解析が必須

遺伝子領域の予測(5) : 真核生物

主な遺伝子予測プログラム: 対象生物種によって精度が大分異なるので注意が必要

- 近縁種のゲノムが**決定済み**の場合
 - 近縁種のタンパク質アミノ酸配列との**相同性**を元にした予測
 - GeneWise (<http://www.ebi.ac.uk/Tools/psa/genewise/>)
 - FGENESH++ (<http://linux1.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfs>)
 - genBlastG (<http://genome.sfu.ca/genblast/index.html>)
 - 近縁種のゲノムが**未決定**の場合
 - *Ab initio* による予測: 統計的な特徴(塩基配列の偏り)のみに基づいて遺伝子の位置を予測
 - FGENESH (<http://linux1.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfind>)
 - GlimmerHMM (GlimmerM) (<http://www.cs.jhu.edu/~genomics/GlimmerHMM/>)
 - Augustus (<http://bioinf.uni-greifswald.de/augustus/>)
 - GeneMark.hmm-E (<http://exon.gatech.edu/>)
 - GeneID (<http://genome.crg.es/software/geneid/>)
 - SNAP (<http://korflab.ucdavis.edu/software.html>)
 - 複数の予測結果を組み合わせる予測
 - Combiner (<http://www.cs.jhu.edu/~genomics/Combiner/>)
 - JIGSAW (<http://www.cbcn.umd.edu/software/jigsaw/>)
 - EVidenceModeler (<https://evidencemodele.github.io/>)

Software	Accuracy
HMM Gene	64.87%
CRITICA	67%
AUGUSTUS	71.58%
JIGSAW	72%
GENSCAN	75-80%
GLIMMER	87%
GAZE	85-90%
GeneBuilder	91%

Sharma R and Kaushik A. *Int. J. Eng. Res. Appl.*, 1, 1436-1440 (2014)



遺伝子の機能予測(1)

主に、配列相同性に基づき、機能予測を行う。

1. 既知タンパク質アミノ酸配列との相同性検索(BLAST)を実施

- 対象DBとしては、主に以下があげられる。
 1. 対象ゲノムの近縁種や系統グループに特化したゲノムDB
(公開されていれば)
 2. COG/KOG(原核生物)、BUSCOなどのオーソログDB
 3. Uniprot (Swissprot)
 4. NCBI RefSeq-NR (non-redundant protein), UniRef

BUSCO
(<https://busco.ezlab.org/>) (Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs)

2. タンパク質のドメイン・モチーフ検索を実施

- Interproscan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>)
 - 問い合わせた配列が属するタンパク質ファミリーやドメイン構成およびモチーフなどの特徴を効率的に推測することができるツール。プロファイル・正規表現DBとして、PROSITE、HMM DBとしてGene3D、PANTHER、PIRSF、Pfam、SMART、SUPERFAMILY及びTIGRFAMs、重み行列DBとしてPRINTSが統合されている。
 - WEBサイトには、これらのDBに対する一括検索機能が備わっており、入力配列とのアライメント結果・GOターム・生物種間の進化系統関係を得ることができる。
 - 【統合TV Interproscanを使ってアミノ酸配列からたんぱく質の機能を予測する】
(<http://tогotv.dbcls.jp/20150227.html#p01>)
- 個別のモチーフDBからの検索
 - 【統合TV Pfamを使ってたんぱく質のドメインを調べる2017】
(<http://tогotv.dbcls.jp/20171212.html>)

各種DBの特性を考慮し、
使用するDBの優先順位を
決める必要がある。



遺伝子の機能予測(2)

- 機能推定時の基準を明確化しておく必要がある。
 - 相同性検索時のしきい値など(E-value、配列一致率、ヒット配列とのヒットした領域とのカバー率)
- 遺伝子産物の名称(product名)の記載に注意しましょう。
 - 現在、国際塩基配列データベースに登録されているデータ中のアノテーションは、登録した研究者が各自に定義、記述しているため、同じ機能でも表記にゆらぎが生じている。
 - 【参考】DDBJの遺伝子命名に関する考え方 (<http://www.ddbj.nig.ac.jp/sub/cds-j.html#product>)
 - アノテーションの際に、「どの表記を採用すればよいか」、「本当に同じ機能なのか」などの混乱が生じている。
 - 機能に関する語彙を統一化するための用語辞書として、Gene Ontology (GO)がある。また、マニュアルで整理されているUniprotKB/Swiss-protや COG/KOG、BUSCOなどのオーソログ遺伝子DB中の機能アノテーションを参考にするのが一般的である。
 - GOについては、【統合TV Gene Ontologyを使って特定遺伝子の機能情報を検索する 2011】(<http://tогotv.dbcls.jp/20111028.html>)などを参考にして下さい。
 - COG/KOG (Cluster of Orthologous Group of proteins, <http://www.ncbi.nlm.nih.gov/COG/>)
 - COG : 主に、原核生物を対象にOrthologous Groupを作成
 - KOG : 真核生物を対象



本日の講習会の内容

1. ゲノムデータベース
 - 微生物統合DB MicrobeDB.jp
 - 植物統合DB PGDBj
 2. ゲノムアノテーションの概要
 - アノテーション手法の紹介
 3. アノテーションパイプライン構築について
 - 実例を元に、構築の考え方について
 4. 演習
 - 任意のゲノム配列に対し、実際にアノテーションしてみよう
- 資料作成にあたって
 - 解析ツール等の紹介では、今後利用してもらいたく、論文ではなく、URLを記載しています。
 - 謝辞
 - アノテーションの資料作成にあたり、かずさDNA研究所の平川英樹先生に、スライドデータ、ならびに、NGSデータを用いたゲノムアノテーションについての情報をご提供頂きましたこと、深く感謝申し上げます。

原核生物ゲノムにおける 共通自動アノテーションパイプラインの必要性

- 公開されているゲノムのアノテーション情報には様々な問題点がある。

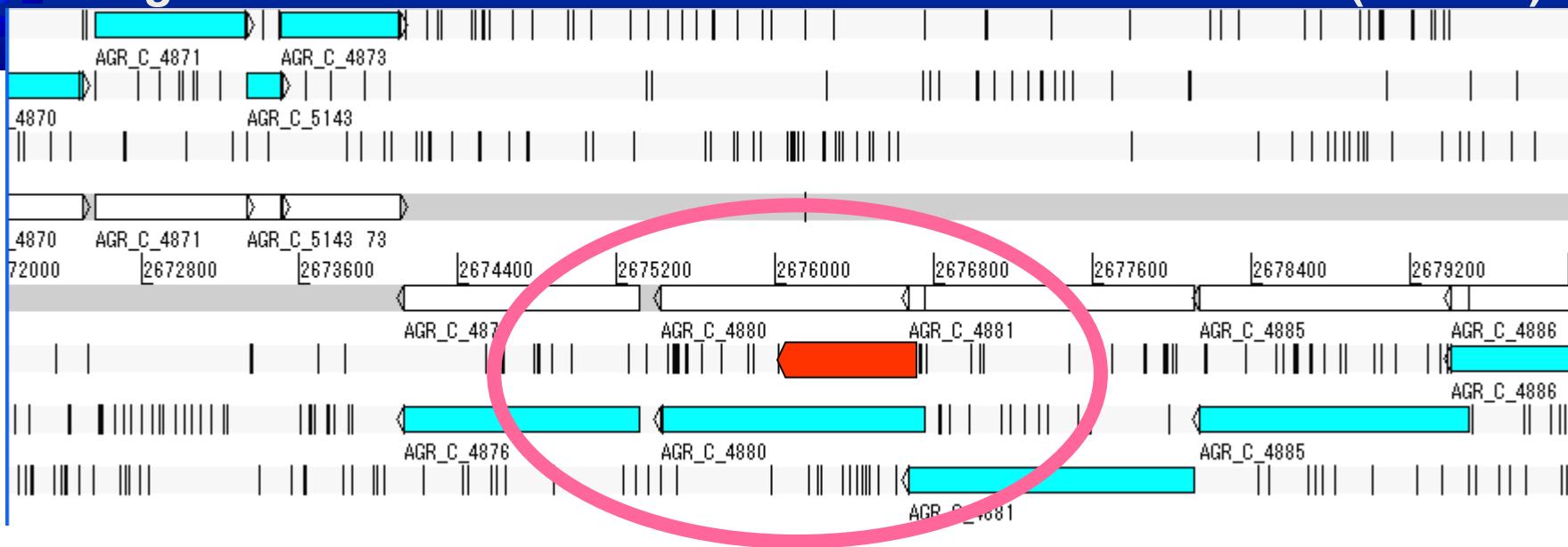
- 予測プログラムの違い
- 最短ORF長の設定が異なる
- 相同性検索のthreshold値の設定の違い
- 相同性検索・モチーフ解析のリファレンスデータベースのバージョンの違い
- プロダクト記載が不統一
- ORF決定の根拠(確かさ)が不明
- アップデートが不定期

ゼロから構築するのでは、各プログラムの選定、ファイル形式の変換など構築するだけでも多大な時間がかかる。

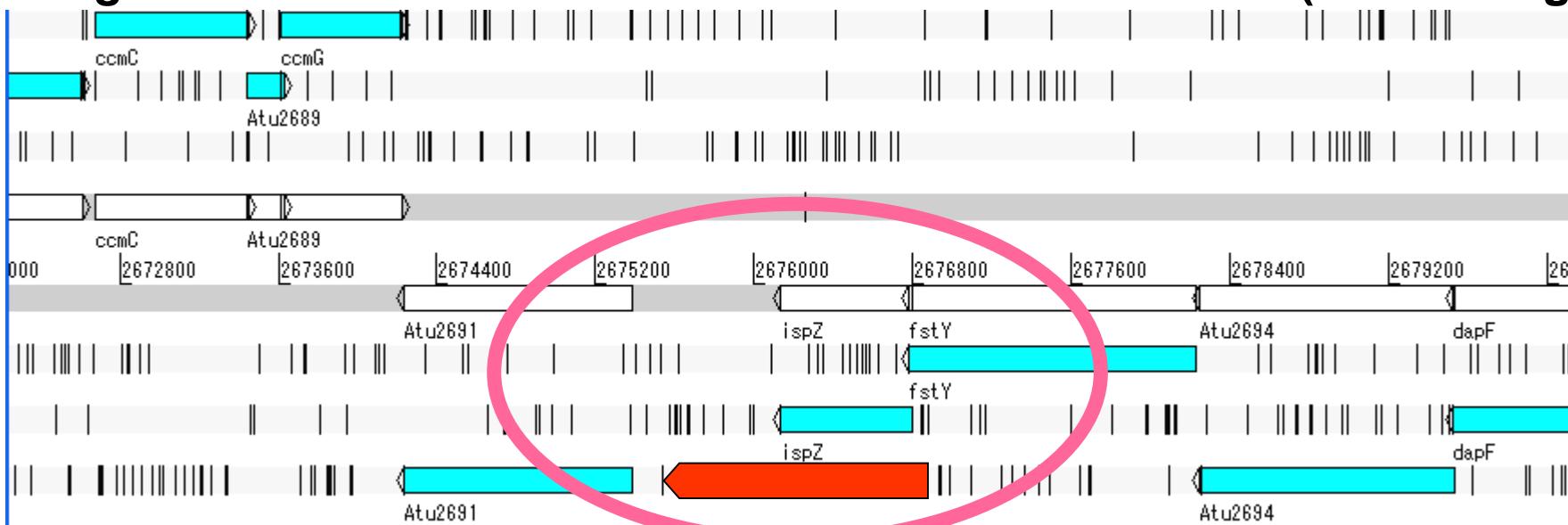


共通化された
パイプライン
があれば、こ
れらの問題は
解消される。

***Agrobacterium tumefaciens* C58 circular chromosome (Cereon)**



***Agrobacterium tumefaciens* C58 circular chromosome (U. Washington)**





実際のゲノムプロジェクトにおいて アノテーションを開始するには？

- 原核生物
 - 遺伝子領域(ORF)の予測精度が高く、高速な全自动パイプラインが開発され、公開されている。
 - 次世代シーケンサーの登場による微生物ゲノムの解読が容易となり、アノテーションにも迅速性が求められている。
 - 研究グループごとのアノテーションの違いを少なくでき、他生物種との比較ゲノム解析にも活用しやすい。
- 真核生物
 - エクソン－イントロン構造を持つため、遺伝子予測の精度も低く、ゲノム配列だけではなく、RNA-seq、ESTやfull-length cDNAなど追加実験のデータ、ならびに、専門家によるマニュアルアノテーションが必須となる。
 - 過去には、ヒト、マウス、イネなどでは専門家コミュニティによるアノテーション会議（アノテーションジャンボリー）が実施してきた。



原核生物ゲノムにおける 自動アノテーションシステム

- 代表的な自動アノテーションシステムとして、以下のシステムが挙げられる。
()内は解析手段と大まかな解析時間を示す。
 - DFAST@DDBJ (WWW/standalone, ~3 min)
 - <https://dfast.nig.ac.jp/>
 - Prokka: rapid prokaryotic genome annotation (standalone, ~3 min)
 - <https://github.com/tseemann/prokka>
 - NCBI Prokaryotic Genome Annotation Pipeline (WWW, >1 month*)
 - <http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html>
 - RAST: Rapid Annotation using Subsystem Technology (WWW, 12-24 hours*)
 - <http://rast.nmpdr.org/>
 - BG7 (standalone, >10 hours*)
 - <http://bg7.ohnosequences.com/>



自動アノテーションシステム DFAST

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

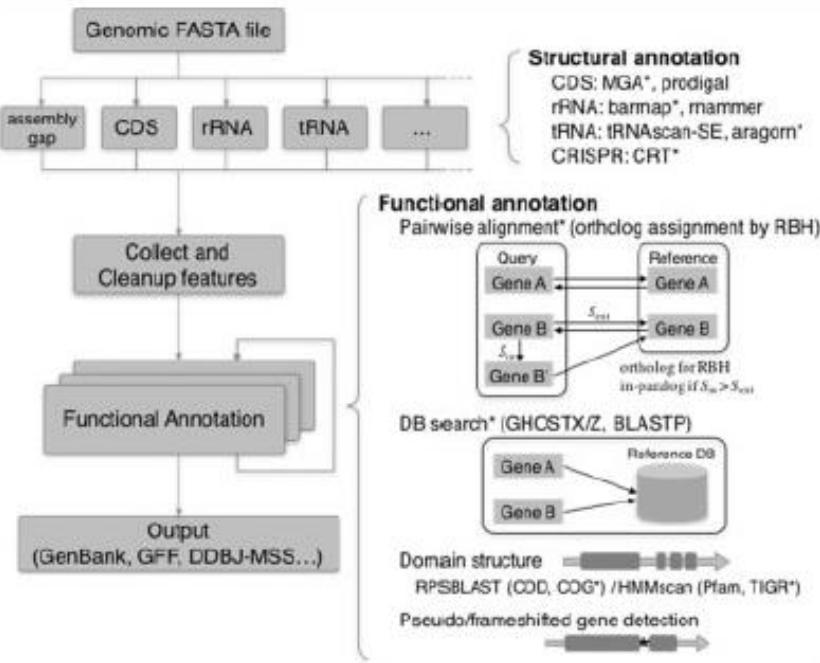


Fig. 1. DFAST annotation workflow. Items marked with asterisks are included in the default workflow

予測パイプラインで使用する予測プログラム

1. CDS ([MetaGeneAnnotator](#)),
2. rRNA (Barrnap)
3. tRNA/tmRNA (Aragorn)
4. **CRISPR (CRT)**

Partial and overlapping features will be cleaned up.

機能アノテーション

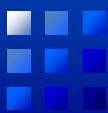
1. Homology search against an additional reference database using the **GHOSTX** aligner (optional)
2. Homology search against the DFAST default database using the **GHOSTX** aligner
3. PseudoGeneDetection (internal stop codons and frameshifts)
4. Profile HMM database search against TIGRFAM (optional, for CDSs without significant hits in upstream processes)
5. **RPSBLAST** search against COG database from NCBI Conserved Domain Database (optional)
6. **OrthoSearch** (standalone only, optional)

出力形式

- Standard annotation format (GFF, GenBank)
- Sequence files in a FASTA format (CDS, protein, RNA)
- Annotated features in a tabular format
- Genome and annotation statistics
- [Submission files for DDBJ Mass Submission System](#)

DDBJへの登録ファイルも作成してくれる。

赤字: Prokkaとの違い



自動アノテーションシステム DFAST

DFAST Analysis Archive DFAST-core API Download FAQ Help

[2018-Mar-15] DFAST paper has been published. *Bioinformatics*, 34(6) 1037–1039, 2018. DFAST-core, a stand-alone version, is available for download at GitHub. ×

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.

DFAST Legacy server (based on Prokka).
Organism-specific reference databases (manually curated)
[Lactic Acid Bacteria \(1.0\)](#) [Bifidobacterium \(β0.1\)](#) [Cyanobacteria \(β0.1\)](#) [E. coli \(β0.1\)](#)
General-purpose reference databases
[RefSeq \(1.0\)](#), automatically curated using protein sequences mainly from 'Reference Genomes' in RefSeq.
Subsets for following phyla are also available: [Actinobacteria](#) [Firmicutes](#) [Proteobacteria](#)

DAGA : DFAST Archive of Genome Annotation

DAGA stores genomic data collected from the public nucleotide database and the sequence read archive. All the genomes are consistently annotated using DFAST.
Currently, DAGA is available only for genomes of Lactic Acid Bacteria.
1421 annotated genome resources are available, covering 2 genera and 191 species. [ENTER](#)

© 2016-2018 National Institute of Genetics. dfast(at)nig.ac.jp

DFAST Analysis Archive DFAST-core

DFAST Prokaryotic genome annotation pipeline

Query File (Fasta format, up to 15Mbyte)

ファイルが選択されていません。 Run in demo mode

Job Title
(optional)

Mail Address
E-mail notification will be sent to this address when the job is completed. (optional)

--- options ---

Additional DB **Minimum sequence length**

Enable HMM scan against TIGRFAM Enable RPSBLAST against COG

Sort sequences by length (the longer comes first)

Fix the sequence origin (only for a finished genome with a circular chromosome)

Rotate/flip the chromosome so that the dnaA gene comes first

Offset from the start codon of the dnaA gene: bp ←ゲノム配列の開始位置の修正も行ってくれる。

Run



自動アノテーションシステム DFAST

大腸菌 K-12 MG1655 (U00096)での実行結果

DFAST Analysis Archive DFAST-core API Download FAQ Help

Remember the current URL to access this page. The result will be deleted 30 days after your last visit. Delete this job now. => [Delete] This procedure cannot be undone.

Title : aa5b38ec-c63c-47b9-9b16-6b6171e2272e
JobID : aa5b38ec-c63c-47b9-9b16-6b6171e2272e
Status : COMPLETE

[2018-06-14 14:49:14.527384] Job submitted.
[2018-06-14 14:49:14.554846] Job started.
[2018-06-14 14:50:27.269201] Job completed.

Result Features DDBJ Submission Log

Genome Statistics

Total Length (bp)	4,639,675
No. of Sequences	1
GC Content (%)	50.8%
N50	4,639,675
Gap Ratio (%)	0.0%
No. of CDSs	4,299
No. of rRNA	22
No. of tRNA	88
No. of CRISPRs	2
Coding Ratio (%)	87.2%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

Download Files

- Genbank Flat File : annotation.gbk
- GFF3-formatted File : annotation.gff
- Genome Fasta File : genome.fna
- Protein Fasta File : protein.faa
- CDS Fasta File : cds.fna
- RNA Fasta File : rna.fna
- Feature Table : features.tsv
- Genome Statistics : statistics.txt
- Zip Archive : annotation.zip

今回の場合、1分13秒で実行終了!!

Web版以外にも、コマンドユーザーインターフェース(CUI)版も提供
データベースや機能アノテーションがカスタマイズ可能

DDBJへの登録に必要なファイル作成のサポートもあり。

DFAST Analysis Archive DFAST-core API Download FAQ Help

Remember the current URL to access this page. The result will be deleted 30 days after your last visit. Delete this job now. => [Delete] This procedure cannot be undone.

Title : aa5b38ec-c63c-47b9-9b16-6b6171e2272e
JobID : aa5b38ec-c63c-47b9-9b16-6b6171e2272e
Status : COMPLETE

[2018-06-14 14:49:14.527384] Job submitted.
[2018-06-14 14:49:14.554846] Job started.
[2018-06-14 14:50:27.269201] Job completed.

Result Features DDBJ Submission Log

1. Preparation for Submit.

You can create DDBJ Submission Files (sequence file and annotation file) required to submit the genome through DDBJ Mass Submission System (MSS). If you want to submit a complete genome, you must provide a sequence name for each entry at [this page](#). Before submission, you need to register BioProject and BioSample. If necessary, raw sequence data should be deposited in SRA.

- Create a DDBJ submission account**
Open the submission portal page D-way, and create a new account.
- Registration to the BioProject Database**
Log-in at D-way, and create a new BioProject.
- Registration to BioSample Database**
Log-in at D-way, and create a new BioSample.

A unique Locus Tag Prefix is required for submitting an annotated genome, which can be registered during the submission procedure of BioProject or BioSample. To start using MSS, please apply from [MSS application form](#). Note that DFAST is not an official service of DDBJ. Please check the latest information at the [MSS guideline](#).

2. Input Metadata

Input metadata by filling the form to create the submission file. Please refer the [instruction](#) for more information for each item. If you want to add items not shown in the form, you can add them manually after downloading the annotation file.

You can "Preview" the provided metadata. You can also import metadata from

このページでは DDBJ Mass Submission System (MSS) を用いて拡基配列を登録するためには必要な 2 種類のファイル（配列ファイルとアノテーションファイル）を作成できます。コンピリートゲムを DDBJ に登録する場合には [こちらのページ](#)で配列名・配列種別（染色体/プラスミド）・直鎖/環状の指定を行ってください。

登録に先立ち、BioProject Database と BioSample Database への登録を次の手順に従って行います。必要に応じてシーケンスデータの SRA への登録も行います。

- 1. DDBJ登録アカウントの取得**
登録ポータル D-way でアカウント申請を行います。
- 2. BioProject Database の登録**
D-way にログインし、新規 BioProject を登録します。
- 3. BioSample Database の登録**
D-way にログインし、新規 BioSample を登録します。

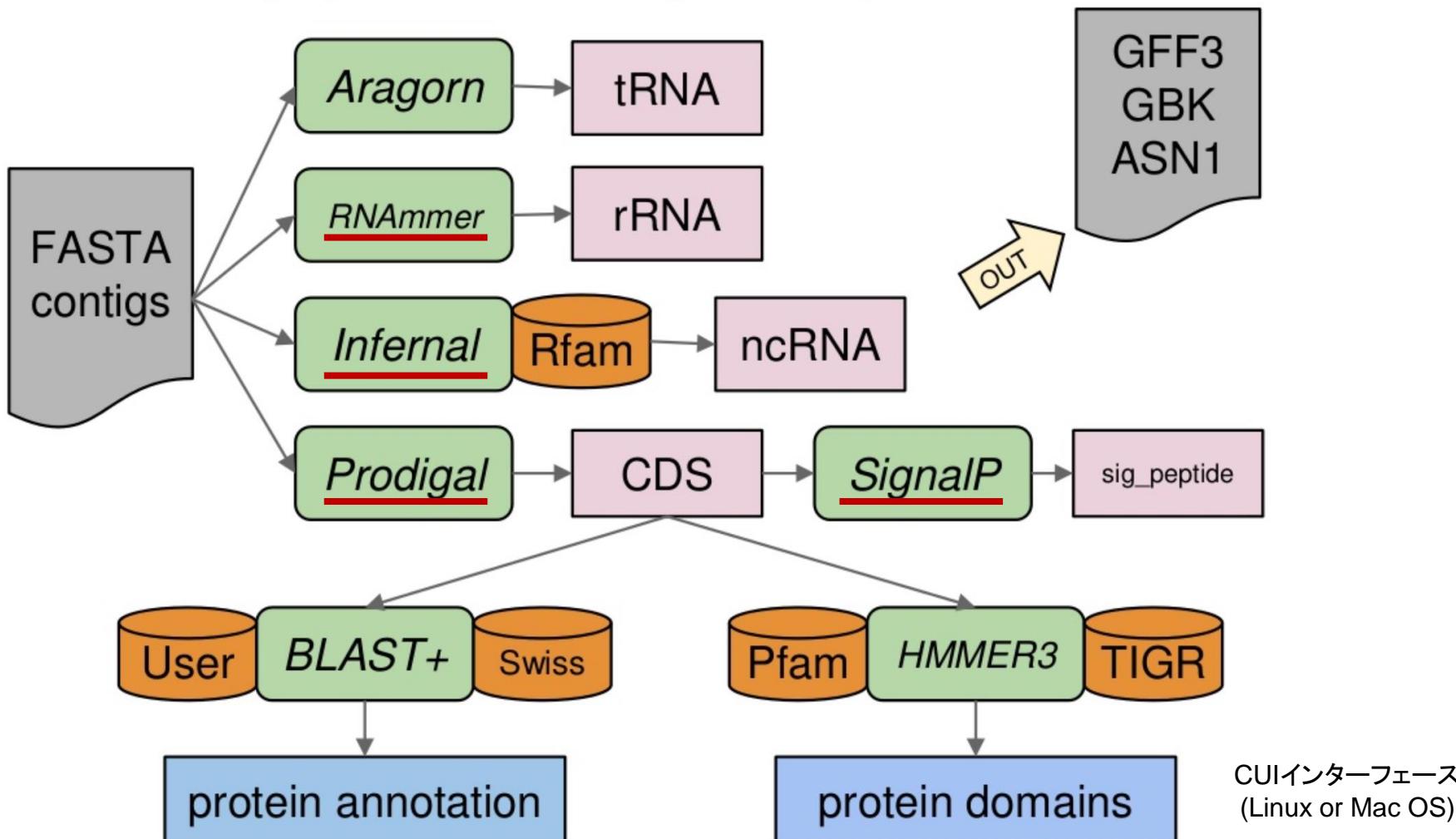
ファイル作成時に必要な Locus Tag Prefix は BioProject もしくは BioSample 登録時に取得することができます。以上の準備ができたら [申し込みフォーム](#) から MSS の利用申請を行います。MSS の詳細については DDBJ の [ガイドライン](#) を参照してください。(DFAST は DDBJ の公式サービスではありません。)



Prokka: rapid prokaryotic genome annotation

<https://github.com/tseemann/prokka>

Prokka pipeline (simplified)





DFAST、Prokkaのアノテーション結果比較

INSD	GC%	CDS	rRNA	tRNA	Pseudo gene
<i>Azorhizobium caulinodans</i> ORS 571	67.3	4717	9	53	0
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	43.5	4328	30	86	88
<i>Clostridium acetobutylicum</i> ATCC 824	30.9	3672	33	73	0
<i>Escherichia coli</i> str. K-12 substr. MG1655	50.8	4386	22	89	199
<i>Fervidobacterium nodosum</i> Rt17-B1	35	1750	6	50	38
<i>Halobacterium salinarum</i> NRC-1	67.9	2058	4	47	0
<i>Pseudomonas aeruginosa</i> LESB58	66.3	5965	13	67	34
<i>Pseudomonas aeruginosa</i> PAO1	66.6	5572	13	63	5

DFAST, Prokkaともにデフォルトで実行

予測遺伝子件数の差は、使用する予測プログラムやフィルタリング方法の違いによる。

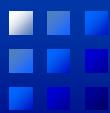
機能予測で使用するデフォルトのデータベースにより、COGIDを付与できた件数や検出されるhypothetical geneの割合が大分異なる。

なお、dfastでは、MGAからProdigalへの変更が可能であり、なるべく予測ソフトを併用することをおすすめする。

各パイプラインともデータベースのカスタマイズが可能。

Prokka	CDS#	exact	%	3' match	%	partial	%	partial	%	pseudo	COG	%	hypothetical	%
						match (5' and 3')*1		match (3' only)*1						
<i>Azorhizobium caulinodans</i> ORS 571	4802	3161	67.0	3938	83.5	3892	82.5	4304	91.2	38	2316	48.2	1533	31.9
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	4225	3759	86.9	3964	91.6	4053	93.6	4106	94.9	23	2200	52.1	1208	28.6
<i>Clostridium acetobutylicum</i> ATCC 824	3715	3373	91.9	3478	94.7	3531	96.2	3564	97.1	35	1732	46.6	1372	36.9
<i>Escherichia coli</i> str. K-12 substr. MG1655	4315	3830	87.3	4006	91.3	4040	92.1	4123	94.0	22	2942	68.2	756	17.5
<i>Fervidobacterium nodosum</i> Rt17-B1	1828	1605	91.7	1682	96.1	1699	97.1	1729	98.8	16	930	50.9	597	32.7
<i>Halobacterium salinarum</i> NRC-1	2104	1552	75.4	1765	85.8	1683	81.8	1841	89.5	10	745	35.4	1072	51.0
<i>Pseudomonas aeruginosa</i> LESB58	6048	5280	88.5	5553	93.1	5549	93.0	5698	95.5	40	2937	48.6	2008	33.2
<i>Pseudomonas aeruginosa</i> PAO1	5671	5030	90.3	5273	94.6	5256	94.3	5396	96.8	33	2893	51.0	1708	30.1
Average			84.9		91.3		91.3		94.7			50.1		32.7
DFAST														
<i>Azorhizobium caulinodans</i> ORS 571	4897	3132	66.4	3934	83.4	3836	81.3	4289	90.9	44	3215	65.7	1512	30.9
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	4242	3778	87.3	3975	91.8	4056	93.7	4129	95.4	16	3007	70.9	1118	26.4
<i>Clostridium acetobutylicum</i> ATCC 824	3746	3381	92.1	3501	95.3	3525	96.0	3575	97.4	32	2665	71.1	1006	26.9
<i>Escherichia coli</i> str. K-12 substr. MG1655	4306	3661	83.5	3911	89.2	3918	89.3	4056	92.5	86	3532	82.0	761	17.7
<i>Fervidobacterium nodosum</i> Rt17-B1	1831	1596	91.2	1669	95.4	1676	95.8	1717	98.1	22	1100	60.1	604	33.0
<i>Halobacterium salinarum</i> NRC-1	2104	1455	70.7	1711	83.1	1619	78.7	1801	87.5	2	1464	69.6	718	34.1
<i>Pseudomonas aeruginosa</i> LESB58	6122	5153	86.4	5403	90.6	5499	92.2	5707	95.7	56	4502	73.5	2232	36.5
<i>Pseudomonas aeruginosa</i> PAO1	5727	4863	87.3	5186	93.1	5203	93.4	5375	96.5	31	4504	78.6	1921	33.5
Average			83.1		90.2		90.0		94.2			71.4		29.9

Exact は遺伝子の開始位置と終了位置が完全に一致していた件数とその割合、3' matchは終了位置が完全一致していた件数とその割合、partial match (5' and 3')は予測遺伝子の開始位置と終了位置の±50bに、INSDの開始位置と終了位置が含まれていた件数とその割合、partial match (3' only)は予測遺伝子の終了位置±50bにINSDの開始位置と終了位置が含まれていた件数とその割合を示す。

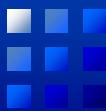


DFAST、Prokkaのアノテーション結果比較(2)

参考資料:各パイプラインでの予測遺伝子領域の一致割合の検証

	exact	3' match	DFAST	exact(%)	3' match (%)	Prokka	exact(%)	3' match (%)
<i>Azorhizobium caulinodans</i> ORS 571	4211	307	4897	86.0	92.3	4802	87.7	94.1
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	3867	133	4242	91.2	94.3	4225	91.5	94.7
<i>Clostridium acetobutylicum</i> ATCC 824	3480	96	3746	92.9	95.5	3715	93.7	96.3
<i>Escherichia coli</i> str. K-12 substr. MG1655	3842	179	4306	89.2	93.4	4315	89.0	93.2
<i>Fervidobacterium nodosum</i> Rt17-B1	1691	66	1831	92.4	96.0	1828	92.5	96.1
<i>Halobacterium salinarum</i> NRC-1	1781	138	2104	84.6	91.2	2104	84.6	91.2
<i>Pseudomonas aeruginosa</i> LESB58	5584	208	6122	91.2	94.6	6048	92.3	95.8
<i>Pseudomonas aeruginosa</i> PAO1	5225	187	5727	91.2	94.5	5671	92.1	95.4
Average				89.8	94.0		90.4	94.6
部分一致(±50b)	partial match (5' and 3')*1	partial match (3' only)*1	DFAST			Prokka		
<i>Azorhizobium caulinodans</i> ORS 571	4555	123	4897	93.0	95.5	4802	94.9	97.4
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	4066	50	4242	95.9	97.0	4225	96.2	97.4
<i>Clostridium acetobutylicum</i> ATCC 824	3596	43	3746	96.0	97.1	3715	96.8	98.0
<i>Escherichia coli</i> str. K-12 substr. MG1655	4029	91	4306	93.6	95.7	4315	93.4	95.5
<i>Fervidobacterium nodosum</i> Rt17-B1	1771	24	1831	96.7	98.0	1828	96.9	98.2
<i>Halobacterium salinarum</i> NRC-1	1974	51	2104	93.8	96.2	2104	93.8	96.2
<i>Pseudomonas aeruginosa</i> LESB58	5813	94	6122	95.0	96.5	6048	96.1	97.7
<i>Pseudomonas aeruginosa</i> PAO1	5461	86	5727	95.4	96.9	5671	96.3	97.8
Average				94.9	96.6		95.5	97.3

パイプライン間で95%近くが一致しており、残りをチェックするだけでも良い状況となりつつある。



【参考】原核生物でのアノテーション結果の整理

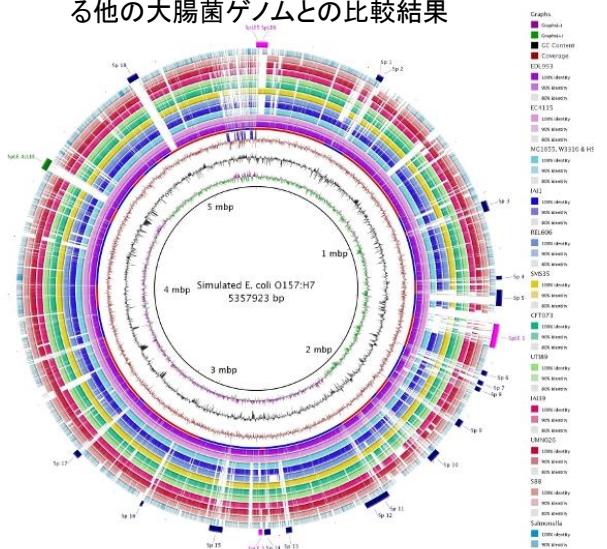
- ゲノム全体の生物学的な機能特徴
 - COGの機能能力テゴリを用いた遺伝子数の分布
 - COGとの相同性検索を実施
 - 保有する代謝経路をチェック: KEGG Pathway
(<http://www.genome.jp/kegg/pathway.html>)
- より詳細なアノテーションの実施
 - 水平伝播遺伝子の同定(IS含む)
 - リピート配列の同定など
- 比較ゲノム解析
 - オーソログ遺伝子グループ作成
 - ゲノム構造比較など

Functional Category	UW4
A RNA processing and modification	1
B Chromatin structure and dynamics	3
C Energy production and conversion	288
D Cell cycle control, cell division, chromosome partitioning	37
E Amino acid transport and metabolism	500
F Nucleotide transport and metabolism	90
G Carbohydrate transport and metabolism	232
H Coenzyme transport and metabolism	156
I Lipid transport and metabolism	255
J Translation, ribosomal structure and biogenesis	174
K Transcription	431
L Replication, recombination and repair	175
M Cell wall/membrane/envelope biogenesis	261
N Cell motility	114
O Posttranslational modification, protein turnover, chaperones	179
P Inorganic ion transport and metabolism	229
Q Secondary metabolites biosynthesis, transport and catabolism	102
R General function prediction only	441
S Function unknown	374
T Signal transduction mechanisms	231
U Intracellular trafficking, secretion, and vesicular transport	43
V Defense mechanisms	62
Total	4378

doi:10.1371/journal.pone.0058640.t002

Pseudomonas sp. UW4のCOGカテゴリ別の件数
Duan et al. PLoS ONE, 8(3), e58640

大腸菌O-157株(*E. coli* O-157)と公開されている他の大腸菌ゲノムとの比較結果





実際のゲノムプロジェクトにおいて アノテーションを開始するには？

- 原核生物
 - 遺伝子領域(ORF)の予測精度も高く、全自動化を目指したワークフローの開発され、公開されている。
 - 次世代シーケンサーの登場による微生物ゲノムの解読が容易となり、アノテーションにも迅速性が必要となる。
 - 研究グループによるアノテーションの違いを少なくでき、他生物種との比較ゲノム解析にも活用しやすい。
- 真核生物
 - エクソン－イントロン構造を持つため、遺伝子予測の精度も低く、ゲノム配列だけではなく、RNA-seq、ESTやfull-length cDNAなど追加実験のデータ、ならびに、専門家によるマニュアルアノテーションが必須となる。
 - 過去には、ヒト、マウス、イネなどでは専門家コミュニティによるアノテーション会議（アノテーションジャンボリー）が実施してきた。

実際のゲノムプロジェクトにおいて アノテーションを開始するには？

nature REVIEWS GENETICS

参考文献

Journal home > Archive > Review > Abstract

JOURNAL CONTENT

- Journal home
- Advance online publication
- Current issue
- Archive**
- Web Focuses
- Supplements
- Article Series
- Multimedia
- Posters

Journal information

- Guide to Nature Reviews Genetics
- Online submission
- Guidelines for referees
- About the journal
- Subscribe
- Feedback for editors

NPG services

- Help
- Authors and Referees
- Librarian gateway
- Advertising information
- work@npg
- Reprints

Review

Nature Reviews Genetics 13, 329-342 (May 2012) | doi:10.1038/nrg3174

ARTICLE SERIES: Study designs

A beginner's guide to eukaryotic genome annotation

Mark Yandell¹ & Daniel Ence¹ [About the authors](#)

[top ↑](#)

The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.

[View At a Glance](#)

[top ↑](#)

Author affiliations

1. Department of Human Genetics, Eccles Institute of Human Genetics, School of Medicine, University of Utah, Salt Lake City, Utah 84112-5330, USA.

Correspondence to: Mark Yandell¹ Email: myandell@genetics.utah.edu

Published online 18 April 2012

Search [Advanced search](#)

nature journals

Download on the App Store

This issue

- Table of contents
- Previous abstract
- Next abstract

Article tools

- Full text
- Download PDF
- Send to a friend
- CrossRef lists 22 articles citing this article
- Scopus lists 16 articles citing this article
- Export citation
- Rights and permissions
- Order commercial reprints

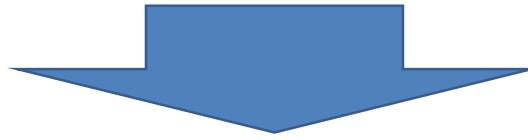
Article navigation

- Author affiliations
- At a glance
- Figures and tables

アノテーションの一連の流れに関する総説なので、ご参考下さい。

真核生物ゲノムへのアノテーション アノテーションパイプラインの構築(1)

- 原核生物と同様に、自動アノテーションシステムが公開されている。原核生物とは異なり、全ての真核生物に対し、精度よく遺伝子領域予測するのは難しく、対象ゲノムに特化したカスタマイズが必要。
 - 特に、遺伝子領域予測プログラムの選択や遺伝子予測用の学習セット作成については、対象ゲノムに特化させる必要がある。



- 対象ゲノムの近縁ゲノムが公開されている場合は、そのプロジェクトで利用されているアノテーション方法を参考にし、パイプラインを構築するのが良い。
- 今回は、原虫(*Theileria orientalis*)で構築したアノテーションパイプラインを中心に、検証事例も加えて、紹介していく。

真核生物ゲノムへのアノテーション アノテーションパイプラインの構築(2)

- 真核生物では、遺伝子予測の際に、ESTやcDNA配列を、ゲノム配列へマッピングし、遺伝子位置同定に活用する。

- 全長cDNA (complementary DNA)

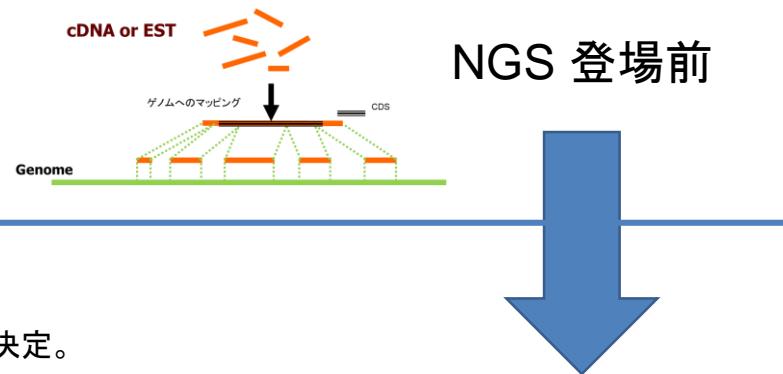
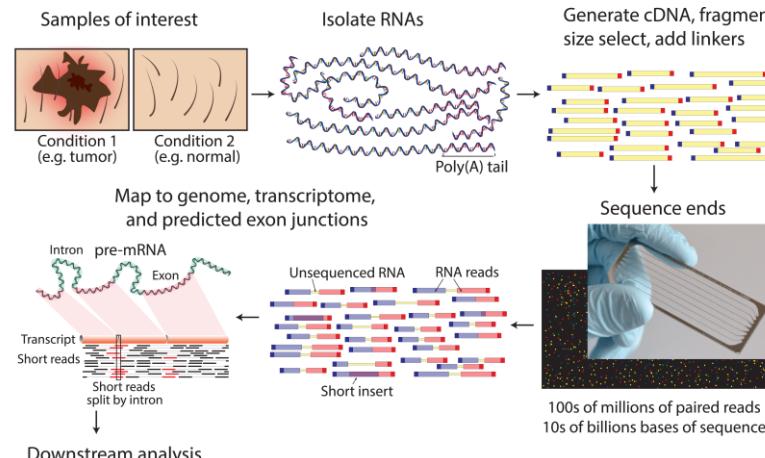
- EST (Expressed Sequence Tag)

- ## - RNA-seq

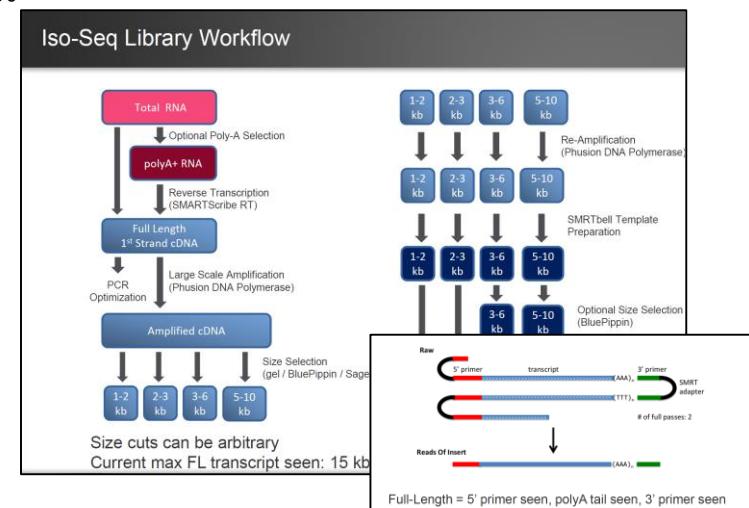
- NGSを用いて、ライブラリ中のRNAの断片配列決定。

- Iso-seq (Isoform sequencing)

- PacBioのロングリードを用いた完全長cDNA配列決定。



NGS 登場後





タイレリア原虫(*T. orientalis*)について

日本、韓国、中国(東北部)には牛小型ピロプラズマ症が分布し大きな経済的被害を与えており。原因是マダニ媒介性の *Theileria orientalis* であり、主に赤血球に寄生し、重度の貧血と黄疸を起こす。本原虫のゲノム解析プロジェクトを実施し、これまでに4本の染色体、全900万塩基のゲノムを完全解読した。“悪性”タイレリア種”とされる *T. parva*, *T. annulata* と “良性”タイレリア種 *T. orientalis*との比較ゲノム解析により、病原性関連遺伝子の解明を試みている。

*T. orientalis*と近縁種ゲノムのゲノム構造

Comparative Genome Analysis of Three Eukaryotic Parasites with Differing Abilities To Transform Leukocytes Reveals Key Mediators of *Theileria*-Induced Leukocyte Transformation

Kyoko Hayashida,^a Yuichiro Hara,^b Takashi Abe,^c Chisato Yamasaki,^b Atsushi Toyoda,^d Takehide Kosuge,^e Yutaka Suzuki,^f Yoshiharu Sato,^g Shuichi Kawashima,^h Toshiaki Katayama,^h Hiroyuki Wakaguri,ⁱ Noboru Inoue,^j Keiichi Homma,^e Masahito Tada-Umezaki,^j Yukio Yagi,^k Yasuyuki Fujii,^l Takuwa Habara,^b Minoru Kanehisa,^m Hidemitsu Watanabe,ⁿ Kimihito Ito,^o Takashi Gojobori,^{b,e} Hideaki Sugawara,^a Tadashi Imanishi,^p William Weir,^p Malcolm Gardner,^q Arnab Pain,^r Brian Shiels,^p Masahira Hattori,^s Vishwanath Nene,^t and Chihiro Sugimoto^{o,2}

Division of Collaboration and Education, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido, Japan;^a Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan;^b Information Engineering, Niigata University, Niigata, Japan;^c Comparative Genomics Laboratory, Center for Genetic Resource Information, National Institute of Genetics, Research Organization of Information and Systems, Mishima, Shizuoka, Japan;^d Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, Japan;^e Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan;^f Graduate School of Pharmaceutical Sciences, Chiba University, Chiba, Japan;^g Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan;^h National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido, Japan;ⁱ Institute of Natural Medicine, University of Toyama, Toyama, Japan;^j Hokkaido Research Station, National Institute of Animal Health, National Agricultural Research Organization, Sapporo, Hokkaido, Japan;^k Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan;^l Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan;^m Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan;ⁿ Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido, Japan;^o Institute of Comparative Medicine, Glasgow University Veterinary School, Glasgow, United Kingdom;^p Seattle Biomedical Research Institute, Seattle, Washington, USA;^q Pathogen Genomics, Computational Bioscience Research Center, Chemical Life Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia;^r and International Livestock Research Institute, Nairobi, Kenya²

Hayashida et al. mBio, 3, e00204-12, 2012.

謝辞：北海道大学人獣共通感染症リサーチセンター 杉本千尋先生と代表として、国立遺伝学研究所、産業技術総合研究所、新潟大等との共同研究成果を本講義に使わせていただきますこと、深く感謝申し上げます。

TABLE 1 Comparison of genome characteristics of *T. orientalis*, *T. parva*, *T. annulata*, and *B. bovis*

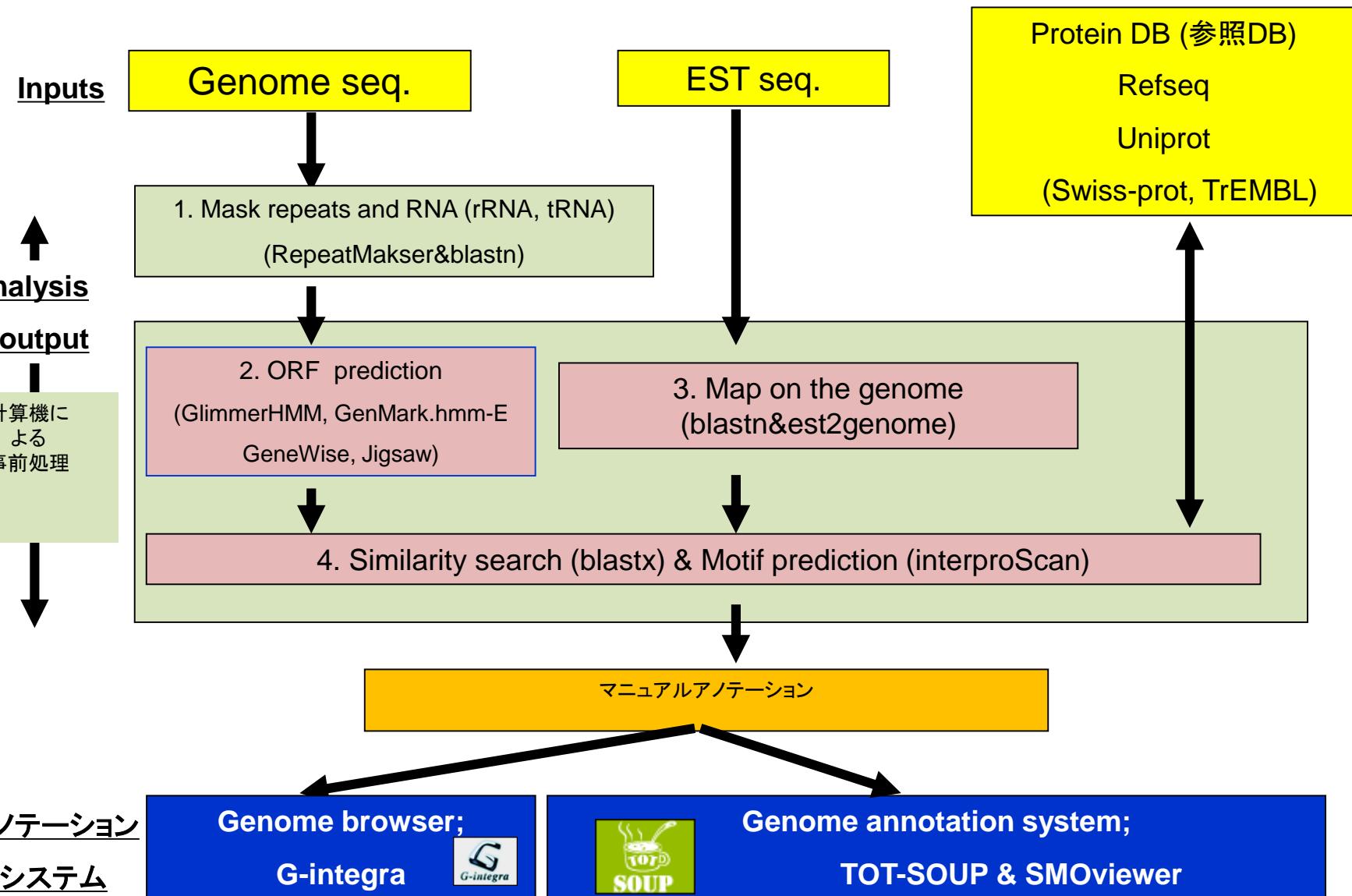
Nuclear genome feature	<i>T. orientalis</i>	<i>T. annulata</i>	<i>T. parva</i>	<i>B. bovis</i>
Size (Mbp)	9.0	8.4	8.3	8.2
No. of chromosomes	4	4	4	4
Total G + C content (%)	41.6	32.5	34.1	41.8
No. of protein-coding genes	3,995	3,792	4,035	3,641
% of genes with introns	78.3	70.6	73.6	61.5
Mean gene length (bp)	1,861	1,606	1,407	1,514
% Coding	68.6	72.8	68.4	70.2
Mean intergenic length (bp)	390	396	402	589
% G + C composition of exons	44.5	37.6	35.9	44.0
% G + C composition of intergenic regions	35.2	22.5	24.9	37.0
% G + C composition of introns	38.1	22.2	23.6	35.9
No. of tRNA genes	47	47	47	44
No. of 5S rRNA genes	3	3	3	NA ^b
No. of 5.8S, 18S, and 28S rRNA units	2	2	2	3
Mitochondrial genome size (kb)	2.5	6	6	6
Apicoplast genome size (kb)	26.5	NA	39.5	33
Gene density ^a	2,249	2,202	2,059	2,228

^a Genome size/number of protein-coding genes.

^b NA, not available.



*T. orientalis*ゲノムアノテーションパイプライン概要





*T. orientalis*ゲノムアノテーション

1. リピート配列の検出と機能性RNAの予測

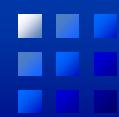
- リピート配列検出
 - RepeatMasker Ver. 3.1.6 を使用(解析当時. 現在は4.0.7)
- rRNA 予測
 - 探索方法
 - Theileria属の公開されている18S, 28S配列(完全長)を収集し、blastnを実行
 - 5.8S, 5S rRNA, nc-RNA : Rfamを対象に、blastnを実行。
 - 18S, 28Sについては、ヒット領域がほぼ全長だったものを抽出。
 - Rfamについては、一致率, カバー率が90%以上を抽出。
 - tRNA 予測
 - tRNAScan-SE 1.23 を使用 (option: -G のみ)



*T. orientalis*ゲノムアノテーション

2. 遺伝子領域予測(1)

- 遺伝子領域予測プログラムとして、以下を使用
 - 選定理由:既に公開されていた*T. parva*, *T. annulata*で利用されていて、フリーで学習セットの作成を自前で行えること。
 - 両末端EST配列からの遺伝子領域予測
 - 相同性に基づく方法
 - Genewise: *T. parva*, *T. annulata*の全ORFを対象にtblastnの実行結果を利用
 - *ab initio*法
 - GlimmerHMM
 - オプション: デフォルトを使用(最小長: 60bpで実施)。
 - 教師セット: *T. annulata*, *T. parva*のゲノムの遺伝子領域(400aa以上)とEST配列から予測した遺伝子領域データ。
 - GeneMark.hmm-E (WEB版)
 - 真核生物用自己学習型ORF finding予測プログラム
 - 複数の予測プログラム結果を組み合わせる方法
 - Jigsaw



2. 遺伝子領域予測(2)

- 遺伝子領域予測プログラム使用の際には、学習セット(既知の遺伝子モデル)をどう設定するかが重要となる。
 - 遺伝子モデルとしては、近縁種、もしくは、ゲノム内で予めアノテーションできているものを利用すると良い。
 - 近縁種の場合：配列長が比較的長い遺伝子領域を対象
 - 近縁種がない場合：自身のゲノムで、相同性などでアノテーションを行い、ある程度長く、確実な遺伝子領域を対象
 - 学習モデルによって、予測される件数が大きく異なる場合も多く、学習モデル作成に使用する遺伝子セットは幾つか検討を行った方が良い。



*T. orientalis*ゲノムアノテーション

2. 遺伝子領域予測(3)

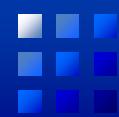
- 各予測プログラムの結果

	GlimmerHMM	GeneMark	GeneWise*	Jigsaw
chr1	1204	891	5854	817
chr2	962	720	6152	668
chr3	879	647	4886	607
chr4	889	629	4946	549
計	3934	2887	21838	2641

*GeneWiseはLow qualityも含む

GlimmerHMMとGeneMarkで予測領域が一致していた割合は、GlimmerHMMの60%程度であり、各予測プログラムで予測件数にばらつきが大きい。

各予測プログラム結果とEST配列のマッピング結果を用いて、遺伝子座の同定のためのクラスタリングを実施する。



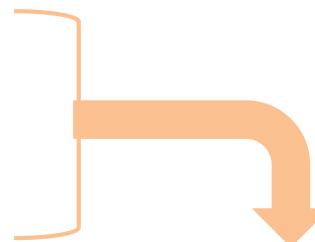
3. EST配列のゲノムへのマッピングについて

- EST配列: 27,478配列
- マッピング方法
 - Blastでゲノム上の大まかな位置を同定し、est2genomeで詳細な領域を同定。
 - Blastの条件は、【e値が1e-4以下】で上位10件をマッピング候補とする。
 - マッピング候補位置より±10kbの領域をゲノムより抜き出し、est2genomeを実行。
 - 一致率90%, coverage 50%以上の条件を満たすこと。



■実行対象:

- 1: JIGSAW
- 2: EST端読みペア
- 3: EST配列マッピング結果
- 4: GlimmerHMM
- 5: GeneWiseの結果



■実行結果: クラスター件数

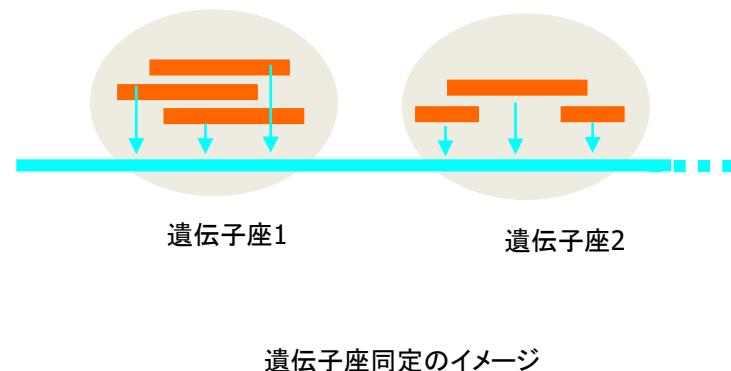
chr1 1185

chr2 924

chr3 865

chr4 859

計 3833



この結果を用いて、マニュアルアノテーションを実施



マニュアルアノテーション結果

- 遺伝子座クラスター**3833件**に対し、実施
 - 複数に分割するもの : 401
 - 除去すべきもの : 437
 - 2つの遺伝子座を融合させたいもの
 - 相同性が低く、決めかねるものなどのコメントがあり、アノテーションを修正し、最終的に、**3995遺伝子**となった。



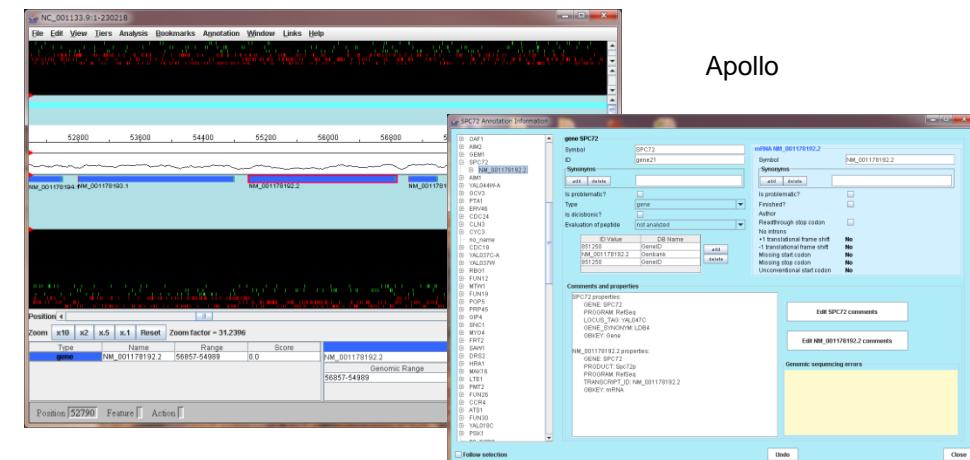
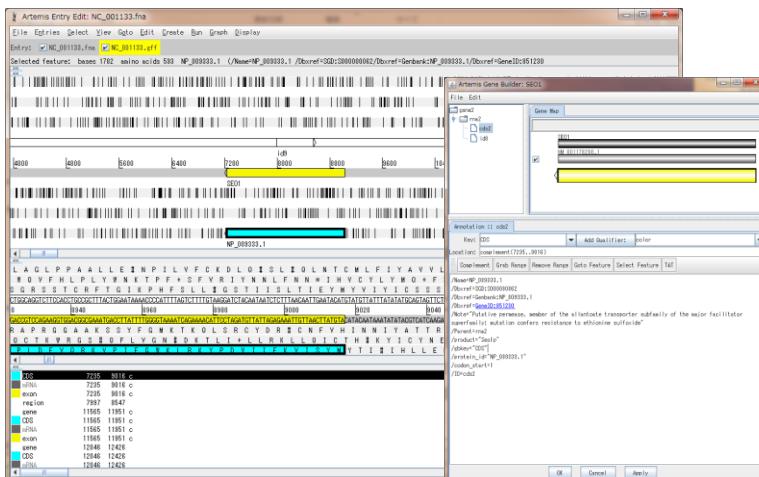
マニュアルアノテーションシステム

- ・ゲノムアノテーションを行う人数に応じて、システムを構築する必要があるが、アノテーションシステムとして、フリー版もある。
 - －ゲノムサイズが小さく、参加人数が少ない場合
 - ・ソフトウェア型で、windows PC上で実行可能
 - ・アノテーション結果を管理する担当者を決めておく必要がある。
 - －ゲノムサイズが大きく、参加人数が多い場合
 - ・サーバー・クライアント方式によるアノテーションシステム
 - ・サーバー構築やデータ管理を行う担当者が必要

マニュアルアノテーションシステム

- 少人数の場合: アノテーション結果をGFF(General Feature Format)等で管理することで、取り込み、アノテーションを実施できる。
 - Artemis (<http://www.sanger.ac.uk/resources/software/artemis/>)
 - 単一のグループでアノテーションを行う場合
 - Apollo (<http://apollo.berkeleybop.org/current/index.html>)
 - 複数のグループでアノテーションを行う場合

Artemis



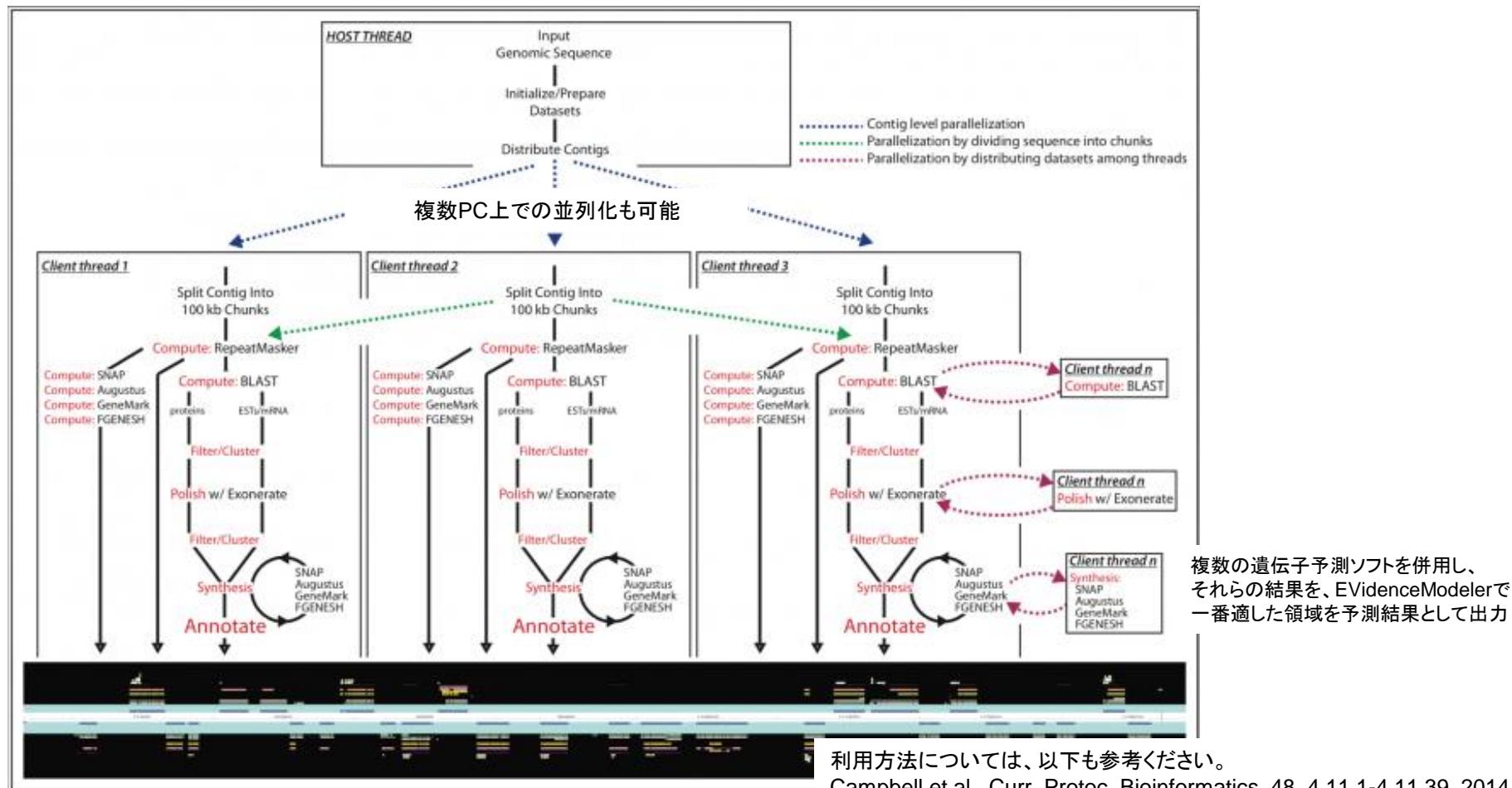
Apollo

真核生物ゲノムアノテーションシステム(1)

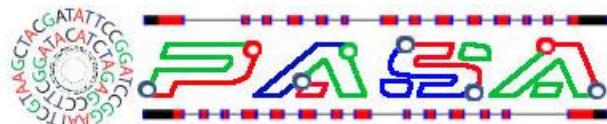


<http://www.yandell-lab.org/software/maker.html>

ゲノムアノテーションパイプライン用のプログラムシステム(植物で主に利用)
特に、真核生物(サイズの小さいゲノム)にも対応しており、Mac / Linuxで、
CUIで一連の解析を各プログラム間でのファイル形式の変更等なく実行することができる。
アノテーション結果は、GFF形式で出力でき、ArtemisやApolloなどへの取り込みも容易にできる。



真核生物ゲノムアノテーションシステム(2)



<https://github.com/PASApipeline/PASApipeline/wiki>

ゲノムアノテーションワークフロー用のプログラムシステム(主に、Fungiで利用)
アノテーション結果は、GFF形式で出力でき、ArtemisやApolloなどへの取り込みも可能

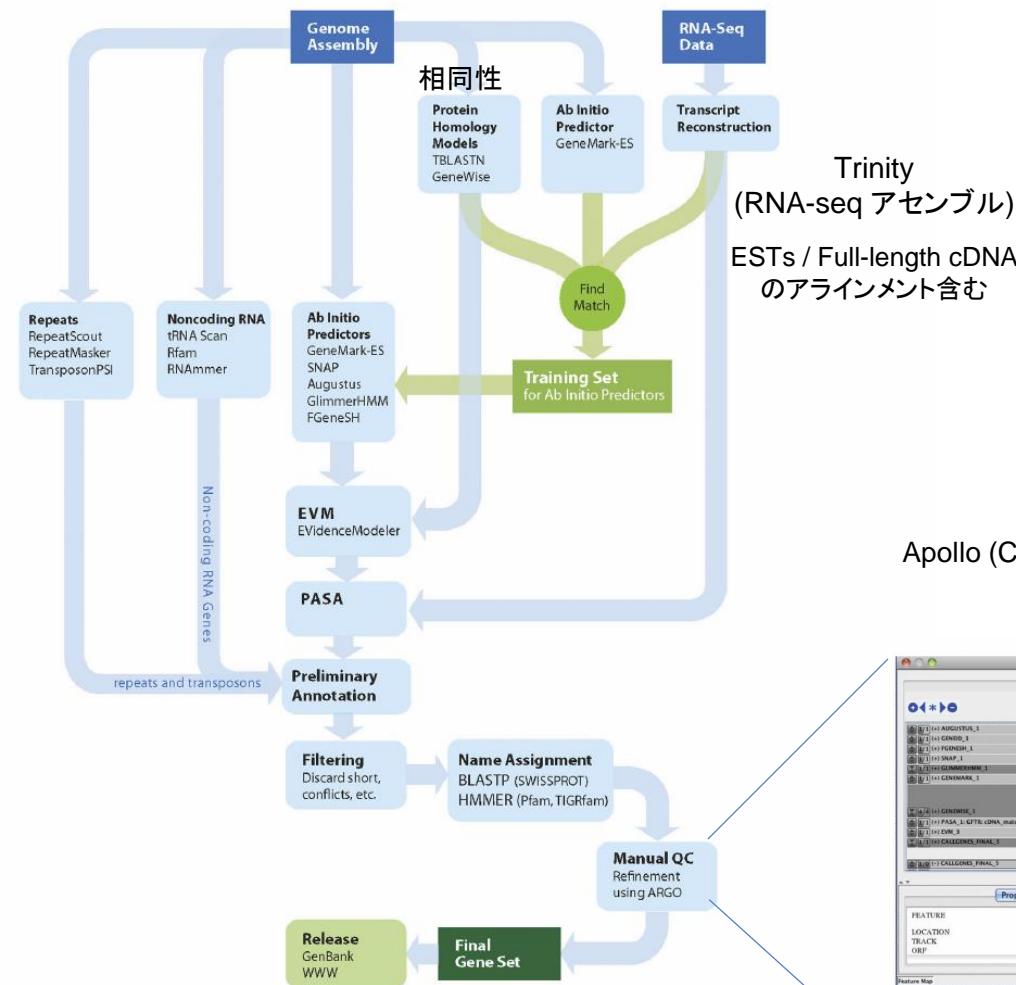


Figure 6. The Broad Institute Eukaryotic Genome Annotation Pipeline

Haas et al., Mycology. 2011 Oct 3;2(3):118-141.

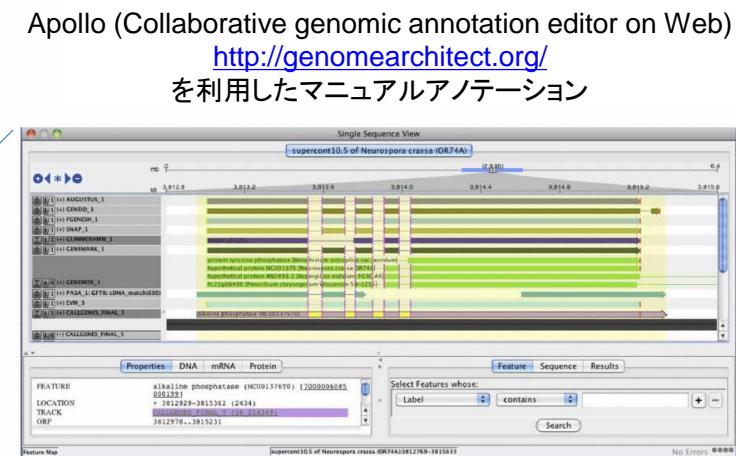


Figure 4. ARGO genome annotation editor display

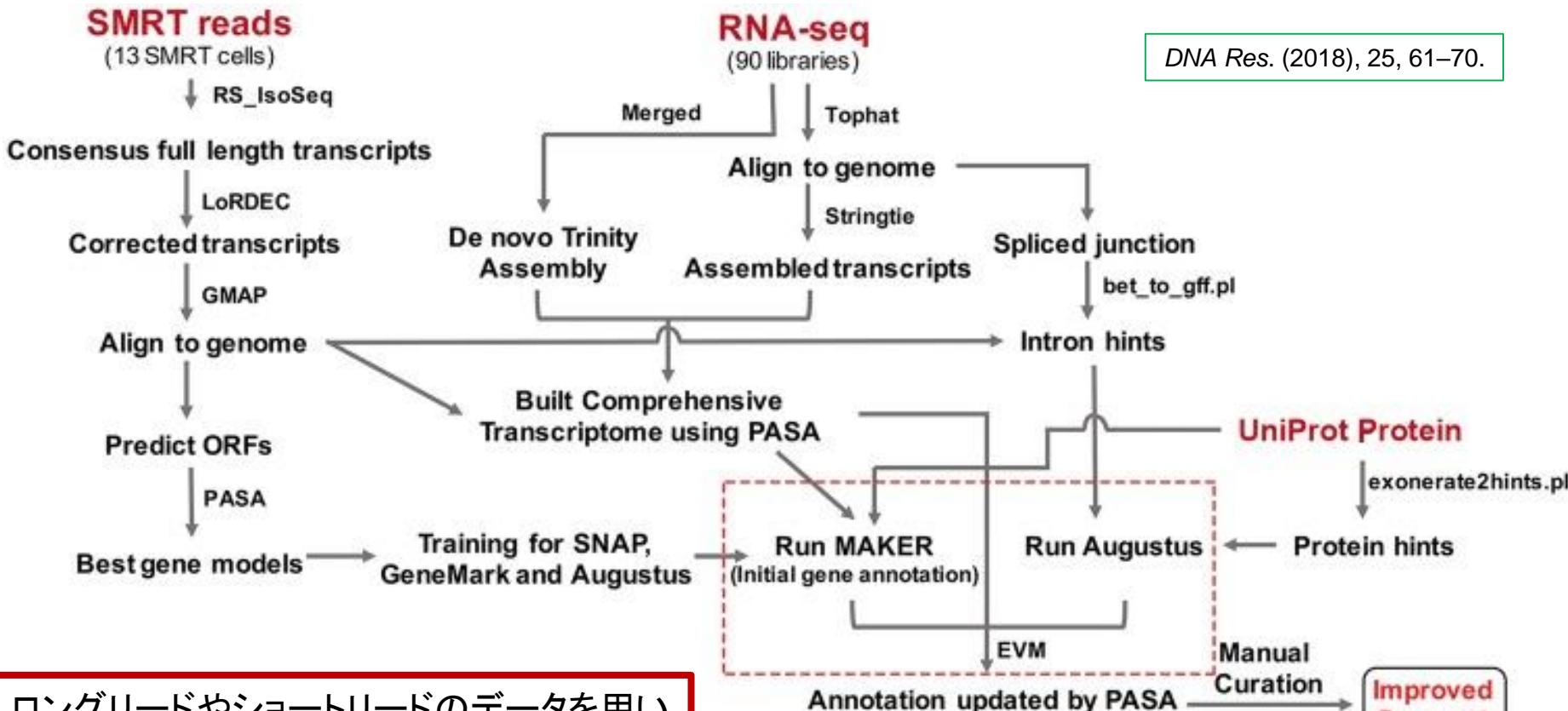
遺伝子の再アノテーション(*Fragaria vesca*(イチゴ野生種))

Iso-Seq(完全長cDNA)、RNA-Seqを用いた遺伝子予測

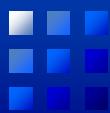


Evidence Modeler (<https://evidencemodeler.github.io>)

*ab initio*法とタンパク質・転写産物のアライメントにより重み付けされたコンセンサスな遺伝子構造を予測する



ロングリードやショートリードのデータを用い、複数のプログラムを用いて遺伝子予測の精度を向上させる



マニュアルアノテーションシステム

- ・ゲノムアノテーションを行う人数に応じて、システムを構築する必要があるが、アノテーションシステムとして、フリー版もある。
 - －ゲノムサイズが小さい、参加人数が少ない場合
 - ・ソフトウェア型で、windows PC上で実行可能
 - ・アノテーション結果を管理する担当者を決めておく必要がある。
 - －ゲノムサイズが大きい、参加人数が多い場合
 - ・サーバー・クライアント方式によるアノテーションシステム
 - ・サーバー構築やデータ管理を行う担当者が必要



H-Invitationalアノテーションプロジェクト

国際アノテーション会議(ジャンボリー)開催
専門家によるアノテーション(注釈付け)



アノテーションジャンボリー風景

統一基準でアノテーションを実施

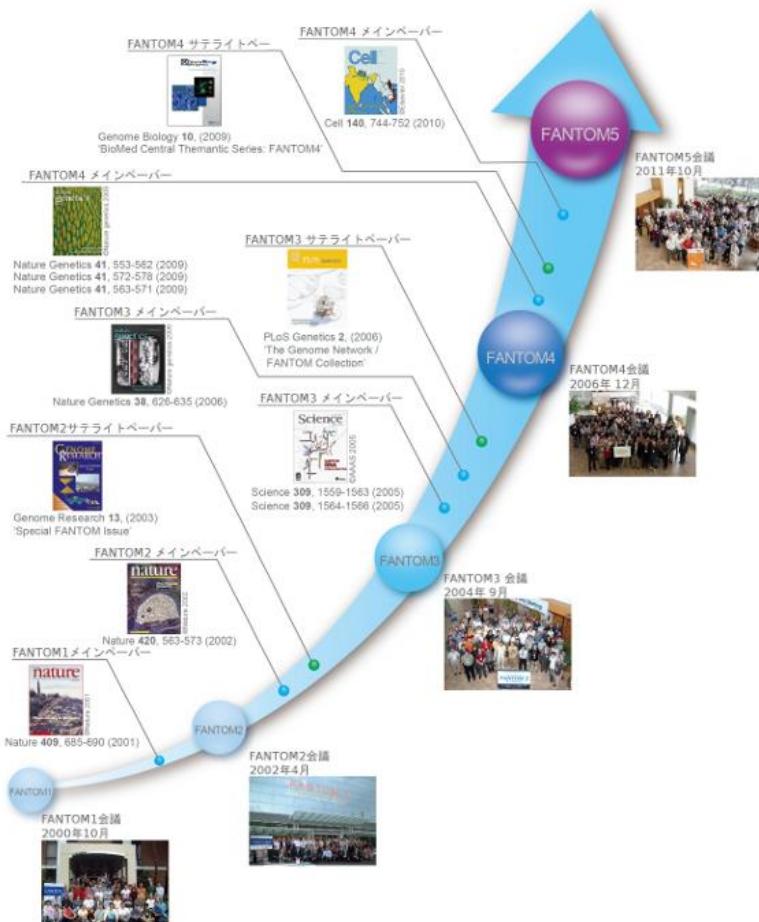


精査されたアノテーション情報を無償で公開

生物情報解析研究センター/現:バイオメディシナル情報研究
センター(JBIC、産総研)およびDDBJ(遺伝研)が主催

H-InvDB リリース1.0 2004/04/20公開(現在は、リリース8.3, 2013/03/26)

国際ゲノム会議は、日本でも精力的に開催されている。



イネアノテーションプロジェクトデータベース（RAP-DB）公開

イネの高精度アノテーション情報を収録したデータベース「RAP-DB」が(独)農業生物資源研究所・(独)産業技術総合研究所・国立遺伝学研究所により共同開発され、このたび公開されました。.

2004年に日本を中心とする国際イネゲノム解読プロジェクト (International Rice Genome Sequencing Project : IRGSP) によって *Oryza sativa* ssp. *japonica* cultivar Nipponbareゲノムの完全測定が達成され、イネ研究はポストゲノムの時代に突入しました (IRGSP, *Nature*, 2005 Aug 11;436 (7052):793 - 800)。

これを受けて世界中からイネ研究者・アーティスターが集合しマンパワーによりイネ全遺伝子を高精度に注釈付ける試み、イネアノテーションプロジェクト (Rice Annotation Project : RAP) が推進されています (論文集備申)。.

RAP-DB は、RAP によって得られたイネゲノムの注釈情報を収録・公開することによって、この貴重な情報をイネ研究のコミュニティに還元するためのデータベースです (Nucleic Acids Research 2006年データベース特集号, in press)。RAP-DB は各種類のアノテーションビューアからなり、キーワードや配列類似性による検索機能を備えています。また完全な cDNA や TIGI ミュータントベース等の有用情報へのリンクを持つことにより、イネゲノム学の「ハブ」となることを目指しています。

RAP-DBへは rapdb.lab.nia.go.jp (WIG) または rapdb.dna.affrc.go.jp (NIAS mirror) よりアクセスできます。

(参考)

- IRGSP (International Rice Genome Sequencing Project)
- 更新情報 [Bu14d4.0] (2005.8.31)
- URL その他の生命情報リンク：生物ごとのデータベース（イネ）

December 15, 2005.

rapwebmaster@ml.affrc.go.jp
Copyright (C) Rice Genome Research Program (RGP), 2000 All rights reserved



イネ@農業生物資源研究所など



演習内容(1)

- *Saccharomyces cerevisiae*のある染色体のゲノム配列に対し、アノテーションしてみましょう。
1. サンプル配列の入手
 - 以下のサイトよりダウンロードして下さい。
– http://bioinfo.ie.niigata-u.ac.jp/AJACS69/AJACS69_testseq.txt
 2. 遺伝子領域予測プログラムの実行
 - 今回は、酵母で最近よく利用されている以下の2種類を実行してみよう。
– Augustus
– FGENESH
 3. 遺伝子予測結果の検証と機能予測
 - 各出力結果を比較し、どちらが正しいのかを相同性検索で検証する。

演習内容(2)

- 遺伝子領域予測プログラムAugustusの実行(1)
 - <http://bioinf.uni-greifswald.de/augustus/submission>

1. 実行画面

なお、40名が一気に実行するとサーバーがダウンしてしまう可能性があります。

上手く実行できない場合は、以下より実行結果をダウンロードして下さい。

<http://bioinfo>

アミノ酸配列

2. 実行結果

3. 各種ファイル ダウンロードページ

Augustus [folder AUG-441208837]

graphical browsable results
input sequence
text results (gff)
predicted amino acid sequences
predicted coding sequences

Search for an amino acid pattern in the predicted amino acid sequences:

PROSITE pattern help example.

[submit another job](#)

【graphical browsable results】
をクリックすると、





演習内容(3)

- 遺伝子領域予測プログラムAugustusの実行(2)
 - <http://bioinf.uni-greifswald.de/augustus/submission>

結果表示について

この場合、+鎖上に、2exonで
322..454, 522..1654となる。
(stop codon含)

演習内容(4)

- 遺伝子予測プログラムFGENESHの実行(1)
 - <http://linux1.softberry.com/berry.phtml>
 - 民間企業のサイトとなるため、今回の演習では、実行せずに、実行結果のみを以下のURLよりダウンロード下さい。
 - http://bioinfo.ie.niigata-u.ac.jp/AJACS69/FGENESH_result.pdf

Softberry社トップページ

The screenshot shows the Softberry website with several key sections highlighted:

- FGENESHをクリック**: A button in the top navigation bar.
- CASE STUDIES:** A section listing various applications of FGENESH, including:
 - Annotating of animal genomes: genes, promoters, functional motifs, protein sub-cellular localization; softberry software, solutions and services.
 - Annotation of plant genomes: genes, promoters, functional motifs, protein sub-cellular localization: Softberry software, solutions and services.
 - Annotation of bacterial Genomes and Community Sequences: Genes, Operons, Promoters, Terminators, Protein Sub-Cellular Localization.
 - Annotation of Bacterial Genomes and Community Sequences: Genes, Operons, Promoters, Terminators, Protein Sub-Cellular Localization.
 - Analysis RNASeq Next Generation Sequencing Data:
 - Accurate alignment of high-throughput RNA-seq data to a reference genome (ReadsMap);
 - De novo transcriptome reads assembly into RNA transcripts (TransSeq);
 - Transomics - pipeline to map RNAseq data, assemble them into transcripts and quantify abundance of these transcripts in particular datasets.
 - Analytic Genome Next Generation Sequencing Data:
 - De novo reconstruction (assembling) of genomic sequence;
 - Reconstruction of sequences using reference genome;
 - Mutation profiling and SNP discovery (OligoZip Assembler);
 - Functional analysis of SNP (SNP-effect);
- For BIOTECH and PHARMA Companies**: A red button in the center of the page.
- TEST ON LINE**: A sidebar on the left containing links to various tools and databases.
- SEARCH FOR MOTIFS /promoters&functional**: A search bar for motif search.
- PROTEIN LOCATION /patterns/Epitops**: A section for protein location prediction.
- RNA STRUCTURE COMPUTING**: A section for RNA structure prediction.
- PROTEIN STRUCTURE**: A section for protein structure prediction.
- PROTEIN / DNA 3D-Visual Works**: A section for protein-DNA 3D visualization.
- SEQUOM**: A section for sequencing.
- MULTIPLE ALIGNMENTS**: A section for multiple alignments.
- ANALYSIS OF EXPRESSION DATA**: A section for expression data analysis.
- PLANT PROMOTERS DATABASE**: A section for plant promoter databases.
- REPEATS /rndMap repeats**: A section for repeats and RNDMap repeats.
- SNP Extracting known SNPs**: A section for SNP extraction.
- Proteomics**: A section for proteomics.
- Software Summary**: A summary of available software.
- Applied in hundreds of species**: A statement indicating the wide application of the tools.
- Sequence comparison**: A section for sequence comparison.
- Automatic genome annotation**: A section for automatic genome annotation.
- Eukaryotic, animal, plant, fungi**: A section for eukaryotic, animal, plant, and fungal genome analysis.
- Bacterial and bacteriophage**: A section for bacterial and bacteriophage genome analysis.
- Comparative genomics**: A section for comparative genomics.
- Visualization of Annotations**: A section for visualization of annotations.
- Genome/Sequence Explorer**: A section for genome and sequence exploration.
- Visualization of Bacterial genome comparison and annotation**: A section for bacterial genome comparison and annotation.
- Analysis of Gene Regulation**: A section for gene regulation analysis.
- Promoter prediction for animal and plant genes**: A section for promoter prediction.
- Search for functional motifs and conserved motifs**: A section for motif search.
- Regsite database of regulatory motifs**: A section for regulatory motif databases.
- Protein 3D-structure analysis and modeling**: A section for protein 3D-structure analysis.
- Assignment of secondary structure and accessibility**: A section for secondary structure assignment.
- Restoring native structures**: A section for restoring native protein structures.
- Restoring coordinates of side chains**: A section for side chain coordinate restoration.
- Prediction of secondary structures**: A section for secondary structure prediction.
- Fold recognition**: A section for fold recognition.
- Homology modeling**: A section for homology modeling.

On the right side of the page, there is a search interface for yeast:

- Paste nucleotide sequence here:** A text input field containing a yeast sequence: TATATCTGGGGGATCATAAACATGGTAAAGTGATTGTGATTGCAAGCTTAAGATTATAATCTATATAACACATAACAAAATTCGAOCCCAGTGGGTTATATAAGCAGGTTTGTGTTCTCTATTCATCTGTTAAAAGTCCTGSGGGTTATGCTCTCACAGAGTTATACCTT.
- Alternatively, load a local file with sequence in Fasta format:** A text input field with a placeholder "Local file name: [選択] ファイルが選択されていません。".
- Select organism specific gene-finding parameters:** A dropdown menu set to "Saccharomyces cerevisiae (baker's yeast)".
- Total 353 genome-specific parameters are available for genefinders of FGENESH suite**: A note indicating the number of available parameters.
- SEARCH / RESET**: Buttons for searching and resetting the parameters.
- Help [Show advanced options]**: A link to show advanced search options.
- Example: Homo sapiens genomic beta globin region (HBB@) on chromosome 11**: An example search query.
- Example: Search in-chain**: Another example search query.
- Return to page with other programs of group: Gene finding**: A link to return to the gene finding group page.
- Most gene finding parameters presented here were trained by Softberry for its own use and distribution, using proprietary and publicly available data. Some of the parameters were created for our academic customers, including Broad Institute/MIT, Washington University, University of Minnesota and The Institute for Genomic Research (TIGR).**: A note about the training data.
- Your use of Softberry programs signifies that you accept Terms of Use**: A note about accepting terms of use.
- Last modification date: 12 Dec 2013**: The last modification date of the parameters.

A large blue arrow points from the "FGENESHをクリック" button to the "For BIOTECH and PHARMA Companies" button. Another blue arrow points from the "Saccharomyces cerevisiae" dropdown menu to the text "サンプル配列を貼り付け".

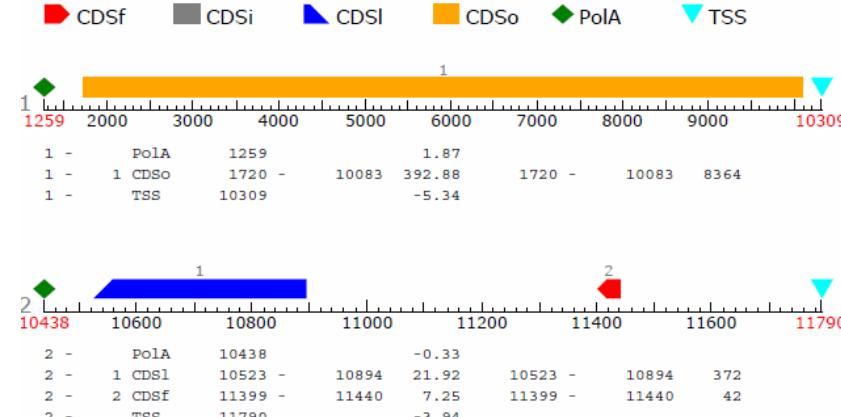


演習内容(5)

- 遺伝子予測プログラムFGENESHの実行(2)
 - <http://linux1.softberry.com/berry.phtml>
 - 民間企業のサイトとなるため、今回の演習では、実行せずに、実行結果のみを以下のURLよりダウンロード下さい。
 - テキストファイル：http://bioinfo.ie.niigata-u.ac.jp/AJACS69/FGENESH_result.txt
 - PDFファイル：http://bioinfo.ie.niigata-u.ac.jp/AJACS69/FGENESH_result.pdf

実行結果ファイル

FGENESH 2.6 Prediction of potential genes in *Saccharomyces* genomic DNA
Seq name: AJACS41_testseq
Length of sequence: 12001
Number of predicted genes 2: in +chain 0, in -chain 2.
Number of predicted exons 3: in +chain 0, in -chain 3.
Positions of predicted genes and exons: Variant 1 from 1, Score:408.34362



一 位置情報

— 配列情報

実行画面

テキストファイル

下方部に配列情報がある



演習内容(6)

- 遺伝子予測結果を比較してみよう
 - AugustusとFGENESHの実行結果ファイルから、予測された遺伝子領域を以下の表に書き出してみましょう。
 - http://bioinfo.ie.niigata-u.ac.jp/AJACS69/AJACS69_annotation_result.xlsxの「比較結果シート」

	Augustus			FGENESH		
	Position	#Exon	Strand	Position	#Exon	Strand
gene1	322..454, 522..1654	2	+			
gene2						
gene3						

結果が異なる場合には、どちらが正しいのかを
各プログラムで出力された各遺伝子のアミノ酸配列を用いて
相同性検索を実行して、確認してみましょう。



演習内容(7)

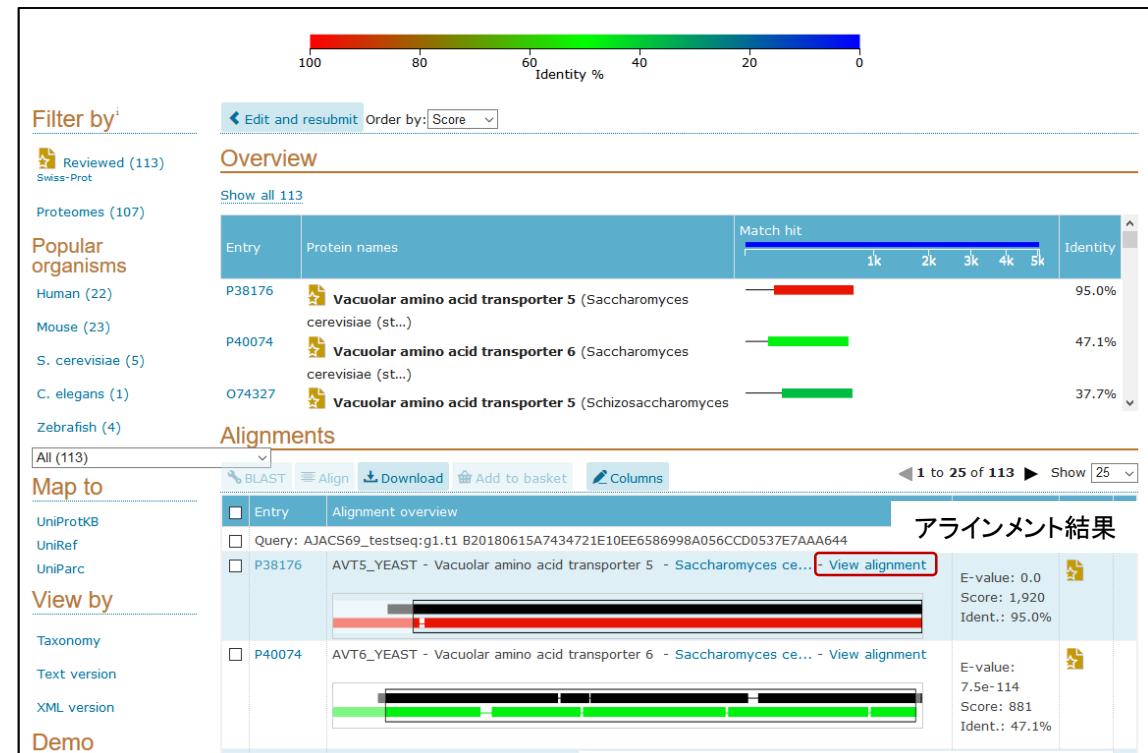
- 予測結果の検証と機能予測のため、相同性検索の実行
 - 今回は、UniprotKB/Swiss-Protをデータベースに検索を実行してみます。
 - <http://www.uniprot.org/>

実行結果画面

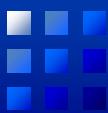
トップページで、ここをクリック

The screenshot shows the UniprotKB/Swiss-Prot BLAST search interface. At the top, there are links for BLAST, Align, Retrieve/ID mapping, Peptide search, Help, and Contact. Below this, a message states: "From June 29, 2018 all traffic will be automatically redirected to HTTPS. More information or view this page using https". The main section is titled "BLAST" and contains instructions: "How to use this tool", "The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.", and a "Protein sequence, Nucleotide sequence or Uniprot identifier" input field. Below the input field are dropdown menus for "Target database", "E-Threshold", "Matrix", "Filtration", "Gapped", and "Hits". A "Run BLAST" button is at the bottom. A blue arrow points to the "UniProtKB/Swiss-Prot" dropdown menu.

UniprotKB/Swiss-Protを選択



相同性が得られた領域情報が記載されているので、確認する。一致率やカバー率をみてみましょう。



演習内容(8)

- 最終遺伝子領域予測結果をまとめてみましょう。
 - AugustusとFGENESHの実行結果に対し、相同性検索等で比較した結果を元に以下の表にまとめてみる。
 - http://bioinfo.ie.niigata-u.ac.jp/AJACS69/AJACS69_annotation_result.xlsxの「最終結果シート」

	Position	#Exon	Strand	Function
gene1				
gene2				
gene3				

どちらの遺伝子領域予測プログラムが良かったか？
相同性検索結果から評価するときのポイントは何か？



演習(9)

- 遺伝子予測結果の検証(2)
 - 相同性検索でヒットしたアミノ酸配列を用いて、ゲノム上 のどの領域に位置するかを確認する。
 - 相同性に基づく遺伝子予測プログラムGeneWiseを利用してみよう。
 - <http://www.ebi.ac.uk/Tools/psa/genewise/>

GeneWise

Input form | Web services | Help & Documentation

Tools > Pairwise Sequence Alignment > GeneWise

Pairwise Sequence Alignment

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

STEP 1 - Enter your sequences

Enter or paste your protein sequence in any supported format:

sp|P38176|AVT5_YEAST Vacuolar amino acid transporter 5 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=AVT5 PE=3 SV=2

MPSNVRSGVL
VPKSENASFA
NMGSQEHH
VNRHQLERGC
PIFAVILAYFLY
HPCRSVSKNIIIFIERNFRKGKLYDNRASIPLDNFNSEDPQEAPQQNNNEPRLRSLESRL
HINIITGFSYLLAATISLAKVLAIVGATGSTSISFLPGLFGYKLIGSEFTGTNE

予測したいアミノ酸配列を貼り付ける。

Or, upload a file: 参照 ファイルが選択されていません。

AND

Enter or paste your DNA sequence in any supported format:

CGGGTATACTGATTCACTAGTCATCTTGTCGA
CATTAATGTTACCTTAAACAAAATCATTCT
GTACGTAGATGCACTCAGAAAATAAAAAAC
TTTATGAAAGTCGAATATACGTTGTTATCAT
CAATTCTCAAGTTTCAATGAACCTTATATT
TTCCTGCTGAAAAATCTCTGAAAGGCAAGCGAACCATAGTTGAAGTTTACTCAA
ATATTCCTTGGAAACATGTGTTCAAAATGTAAATTGCAAGCAGCTTGAAGAACCA
ACCTGTTTGGAAACATGTGTTCAAAATGTAAATTGCAAGCAGCTTGAAGAACCA

サンプル配列を貼り付ける。

Or, upload a file: 参照 ファイルが選択されていません。

STEP 2 - Set your options

SHOW PARAMETERS PRETTY ASCII

ON ON

GeneWise

Input form | Web services | Help & Documentation

Tools > Pairwise Sequence Alignment > GeneWise

Results for job genewise-l20130725-074630-0427-12847811-oy

Alignment Submission Details

View Alignment File

genewise \$Name: wise2-4-1 \$ (unreleased release)
This program is freely distributed under a GPL. See source directory
Copyright (c) GRU limited: portions of the code are from separate copyright

Query protein: sp|P38176|AVT5_YEAST
Comp Matrix: BLOSUM62.bla
Gap open: 12
Gap extension: 2
Start/End local
Target Sequence AJACS41_testseq
Strand: forward
Start/End (protein) local
Geno Parameter file: geneo.stat
Splice site model: GT/AG only
GT/AC bits penalty: -8.96
Codon Table: codon.table
Subs error: 1e-06
Indel error: 1e-06
Null model: syn
Algorithm: 623

genewise output
Score 1101.40 bits over entire alignment
Scores as bits over a synchronous coding model

Warning: The bits scores is not probabilistically correct for single seqs
See WWW help for more info

sp|P38176|AVT5_ 1 MPSNVRSGVLTTLHTACGAGVLAMPFAFPKPFGLMPGLITLTFCGICSLC

最下方に
予測結果の
位置情報が出力される。