MicrobeDB.jpポータル:
開発中のゲノム比較解析機能の紹介

beta.microbedb.jp

MicrobeDB.jp

Home Document Analysis e.g. hot spring, Enterococcus faecalis, psbA Search guest

Comparative analysis:
Metagenomic samples
Taxon vs Taxon
Genome vs Genome
Environments (MEO vs MEO)

ゲノム比較解析（未公開）

MicrobeDB.jp

Integrating and representing genome, metagenome, taxonomy resources and the analysis datasets with Semantic Web Technologies.

Database statistics

Total number of Metagenomic samples (SRA/SRS):	173,359 samples
- with taxonomic analysis results:	60,551 samples
- with functional analysis results:	4,048 samples
Total number of Assembled Genomes (RefSeq/Genbank):	16,983 taxa
Total number of Strains (JCM/NBRC):	16,671 strains
Total number of Environmental terms in ontology (MEO):	2,381 terms

Show graph

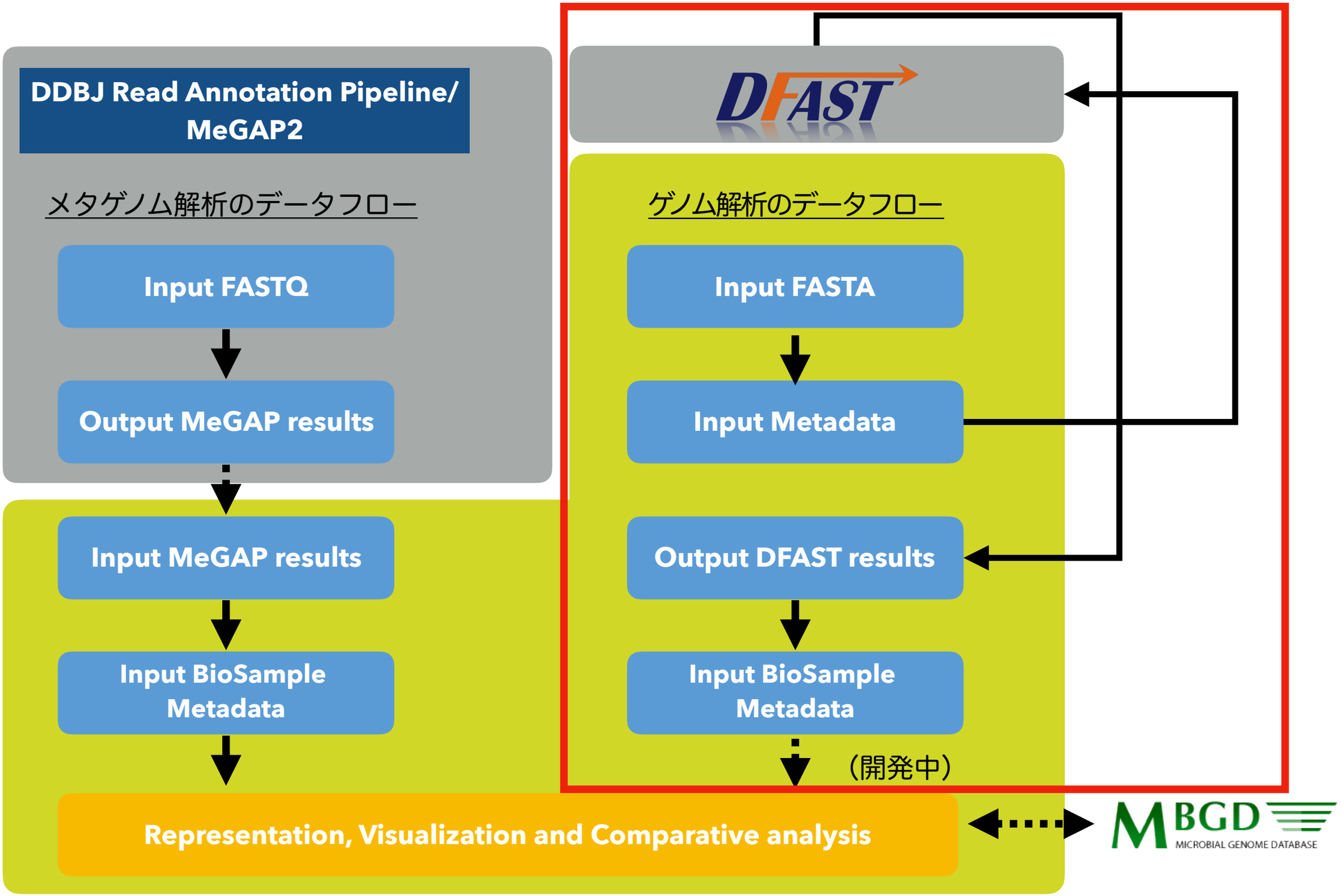
Features

Keyword Search

MicrobeDB.jp provides a keyword search function with a simple interface. The keyword search gives the user free-text access to the literal fields of all RDF/OWL resources on MicrobeDB.jp. See document#sesource for more information.

https://beta.microbedb.jp/#

MicrobeDB.jp解析パイプラインとの連携



MicrobeDB.jpポータル上からユーザ自身のメタゲノム・ゲノムデータ解析および比較解析を実現するため、それぞれメタゲノム解析パイプラインMeGAP、微生物ゲノムアノテーションパイプラインDFASTと連携させた。MeGAPについては非常に計算リソースを要求するため、国立遺伝学研究所スパコンで運用中のDDBJ Read Annotation Pipelineを介して実行後、解析結果が入力となる。

Bioinformatics, 34(6), 2018, 1037–1039
doi: 10.1093/bioinformatics/btx713
Advance Access Publication Date: 2 November 2017
Applications Note

OXFORD

Genome analysis

DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication

Yasuhiro Tanizawa, Takatomo Fujisawa and Yasukazu Nakamura*

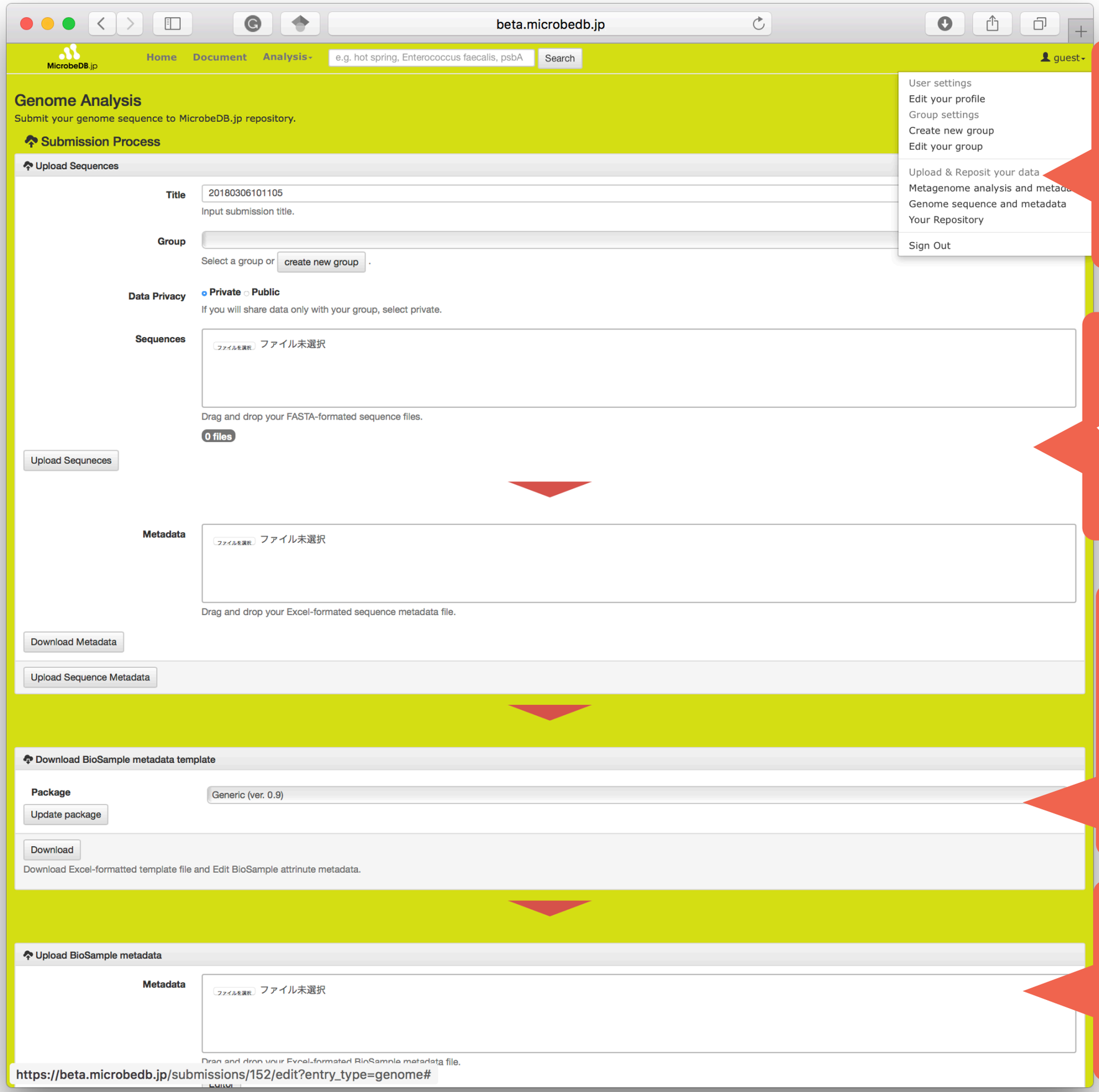
Center for Information Biology, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima 411-8540, Japan

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 15, 2017; revised on October 27, 2017; editorial decision on October 30, 2017; accepted on November 6, 2017

ゲノム比較解析のためのデータアップロード機能



サインイン後にグループ作成・管理、ユーザデータアップロードが可能

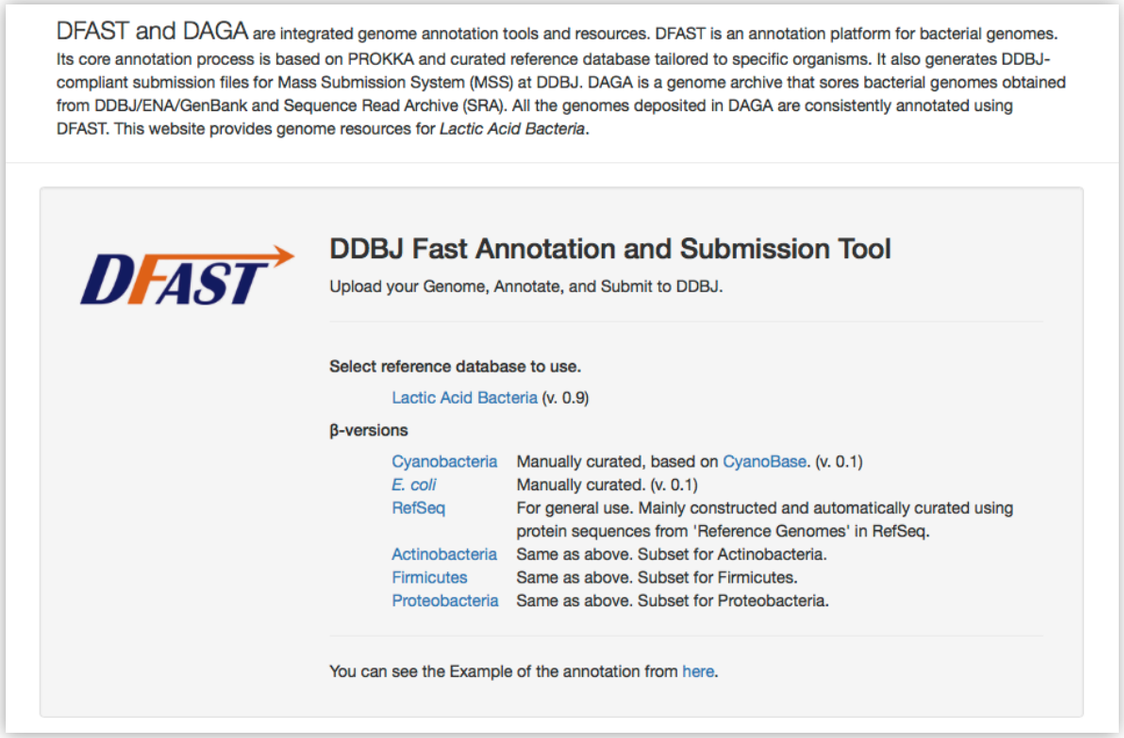
ゲノム配列およびDFASTメタデータ（APIパラメーター情報)のアップロード

DDBJ BioSample登録に準拠したBioSample Packageと必須なメタデータ入力テンプレートのダウンロード

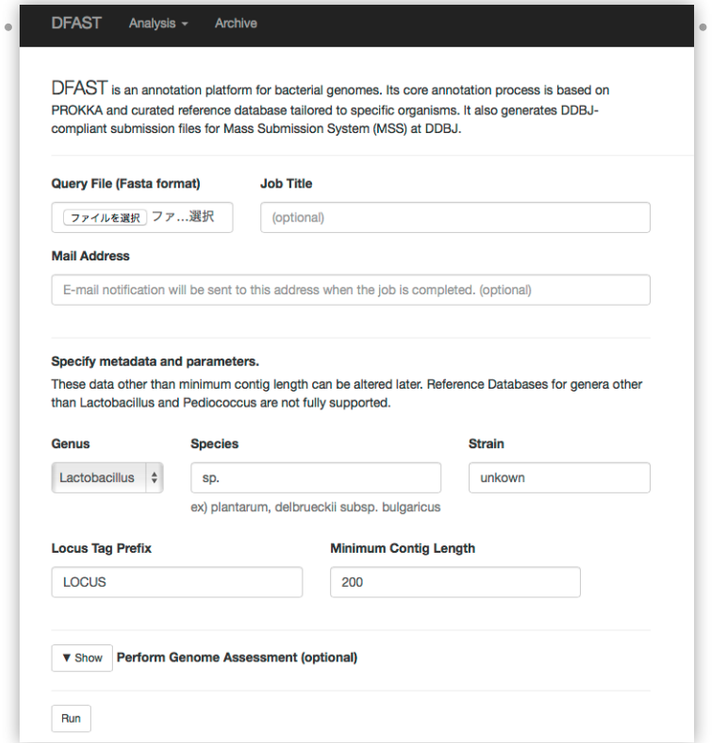
DDBJ BioSample登録に準拠したサンプルメタデータのアップロード

DFAST webサービス

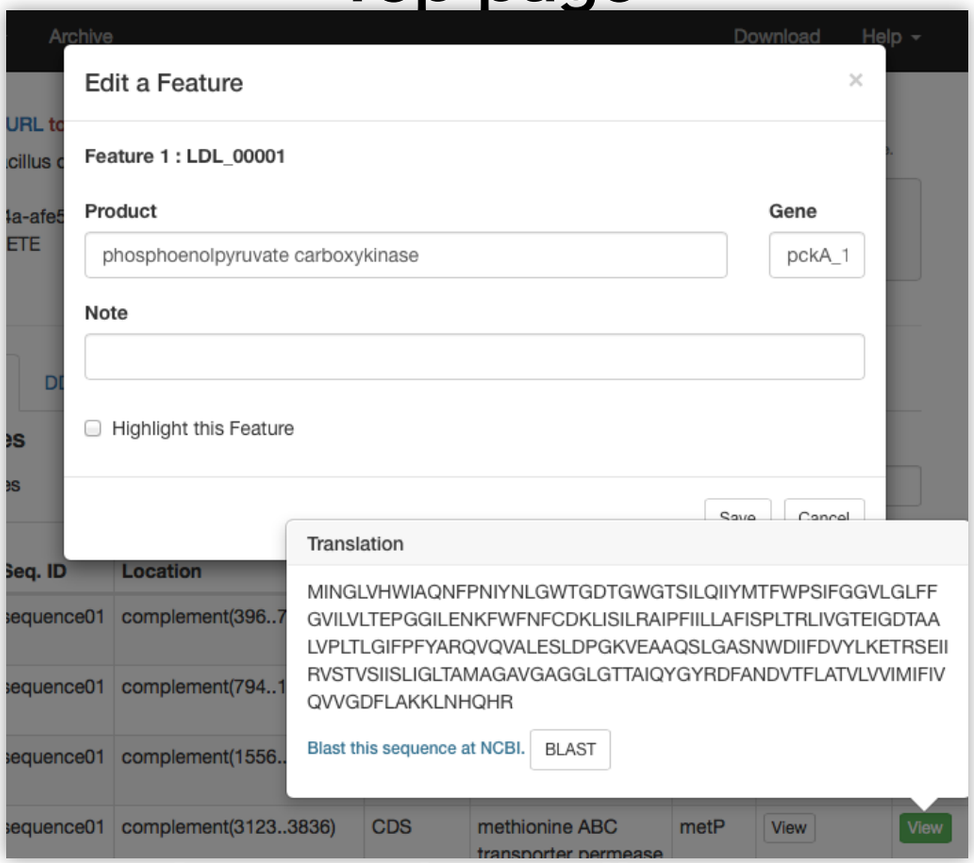
https://dfast.nig.ac.jp



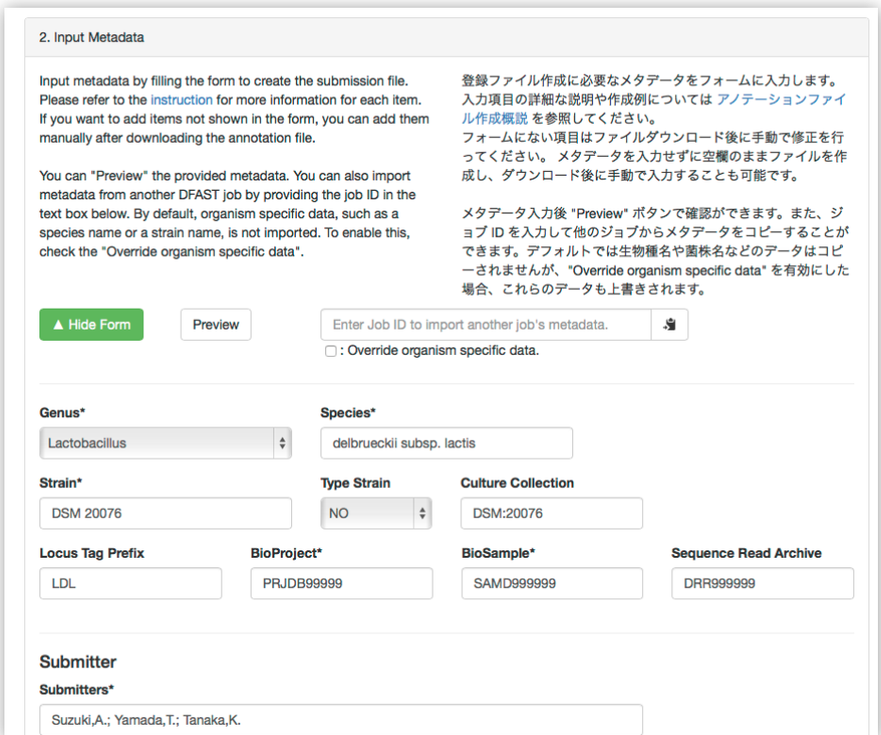
Top page



Job submission



Annotation editor



DDBJ submission file editor

DFAST: DDBJ Fast Annotation and Submission Tool

V 0.1 (2016.1) Tanizawa et al., BMFH (2016)

軽量アノテーションパイプライン Prokka をベース

生物種群ごとにキュレーションされたDBを用意

メタデータ・アノテーション結果を編集するGUI

→ DDBJ compliant な登録ファイルの生成

Web APIによるジョブ投入・管理機能の実装

汎用的な参照DBを用意

独自アノテーションエンジンDFAST-coreの新規開発

Stand-alone ツールとしても公開 (GitHub: nigyta/dfast_core)
Tanizawa et al., Bioinformatics (2017)

V 1.0 (2017.8)

DFAST-core: 新バックグラウンドエンジン

スタンドアローンツールとしても利用可能

https://github.com/nigyta/dfast_core

Linux/Mac, Python 2.7/3.4- で動作

外部バイナリー同梱

カスタマイズ可能なワークフロー

高速かつリッチなアノテーション

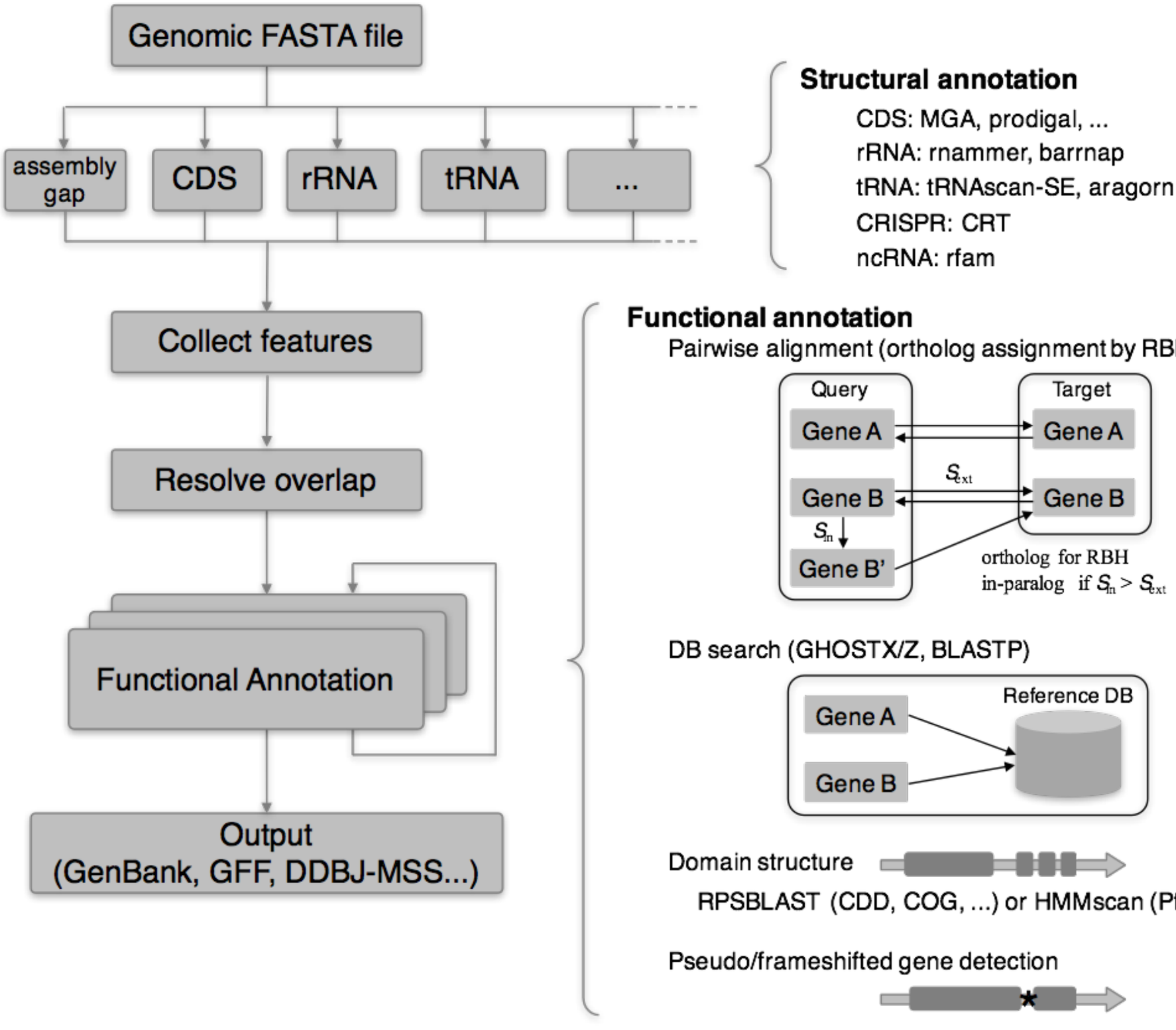
GHOSTX をデフォルトで利用

偽遺伝子・フレームシフトの検出

DDBJへの登録ファイル作成

DFAST-core アノテーションワークフロー

参照DB・外部ツール・パラメータ等を自由に設定可能



典型的なサイズのバクテリアゲノムを約５分でアノテーション

Prokka と同等の実行時間でより多くの遺伝子の機能を予測

参照DBのサイズはProkkaの20倍

BLASTPを使った場合の実行時間の1/10

セレノシステインを含む3遺伝子の検出に成功

Escherichia coli O26 アノテーション結果

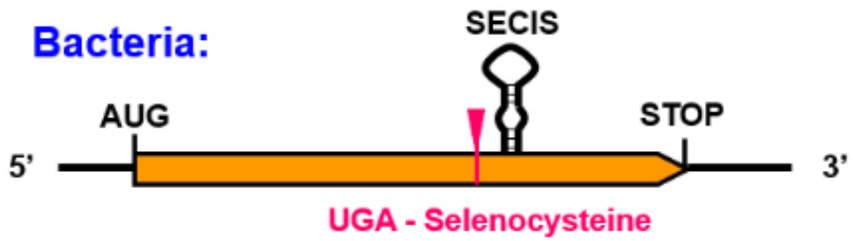
Data source / Annotation tool	INSDC	RefSeq (PGAP)	**DFAST**	Prokka	MiGAP
Total CDS	5795	6243	**5740**	5759	5721
Pseudogene frameshift or internal stop / partial	276	337 (250/87)	**344 (158/186)**	[30*]	-
Selenoprotein	3	1	**3**	-	-
with COG number	-	-	**3965**	-	4392
Unknown function	1203	1514	**1347**	2068	418
tRNA	101	101	**105**	105	100
rRNA	22	22	**22**	22	22
CRISPR arrays	-	2	**2**	2	-
Running time	-	-	**3m27s**	3m20s	4h43m

* Indicated as possible pseudo in the log file

(INSDC: GCA_000091005.1, RefSeq: GCF_000091005.1)

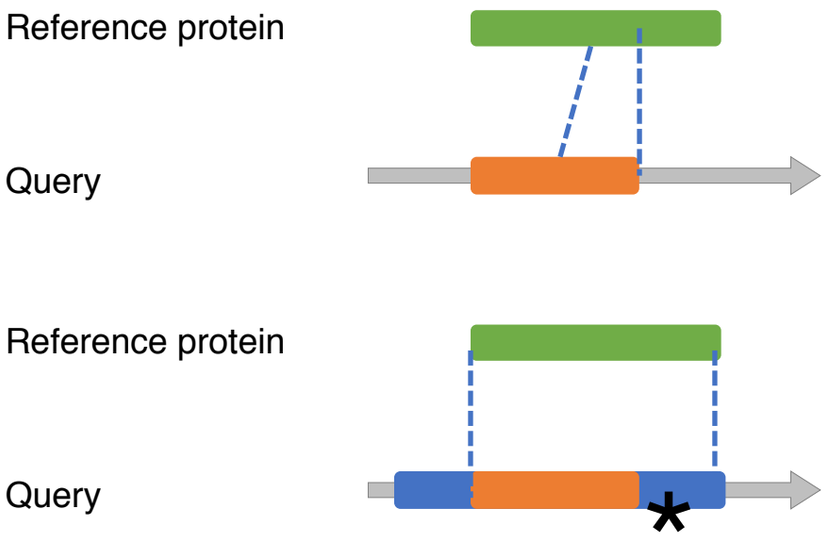
DFASTで対応しているアノテーション出力例

翻訳の例外



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

偽遺伝子・フレームシフトの検出



/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"

MBGD Cluster IDのアサイン（開発中）

/inference="similar to AA sequence:RefSeq:WP_003548611.1"
/inference="similar to AA sequence:MBGD:2015-01_default:16"
/codon_start=1
/product="AraC family transcriptional regulator"
/transl_table=11
/note="WP_003548611.1 AraC family transcriptional regulator (Lactobacillus acidophilus NCFM) [pid:36.0%, q_cov:99.7%, s_cov:99.4%, Eval:8.0e-54]"
/note="MBGD: {gene_id: 'crn:CAR_RS11760', cluster_id: '16', gene_description: 'transcriptional regulator', version: '2015-01', tabid: 'default', pid: 40.2%, q_cov: 31.1%, s_cov: 29.8%, Eval: 1.1e-15}"

ゲノム比較解析インターフェース（イメージ）

課題

- ゲノムにおいては、MBGDオーソログ解析、メタゲノムにおいては、MeGAPメタゲノム解析（taxonomy/function）に大規模な計算リソースと時間を必要とするため、最新の公開データに関して解析データを含めたデータベースからの提供は困難である

今後の対応

- ゲノム、メタゲノムサンプルのメタデータについてはの更新を行い、最新の公開データを対象とした比較解析のためのMeGAP, DFASTの実行はユーザ自身に委ねる
- MBGD、MeGAPによる解析結果の反映についてはメジャーリリース時に対応