08

# 微生物ゲノムアノテーションパイプラインDFAST

## DDBJの登録システムとの統合による登録査定業務の効率化を目指して

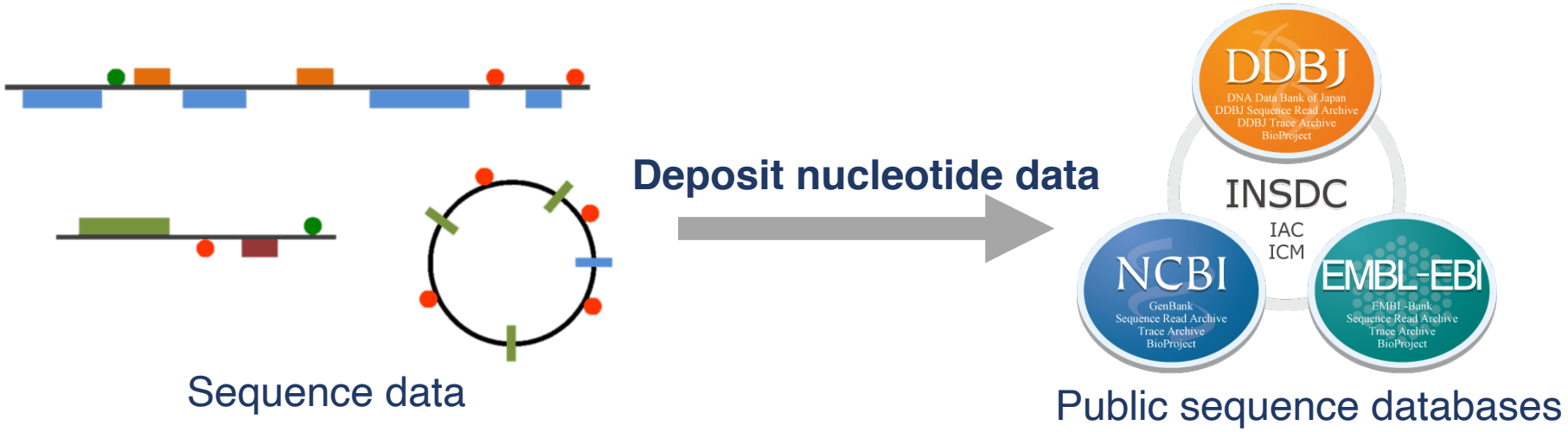


〇谷澤靖洋[1, 2]、藤澤　貴智[1, 2]、大城戸 利久[2]、青野 英雄[2]、李 慶範[2]、有田　正規[1, 2]、中村　保一[1, 2]

1) 情報・システム研究機構 国立遺伝学研究所 情報研究系
2) 情報・システム研究機構 国立遺伝学研究所 生命情報・DDBJセンター

# 1. 概要: 微生物ゲノムアノテーションパイプラインDFAST






**Genome annotation / data submission pipelines**

DFAST → DDBJ DNA Data Bank of Japan

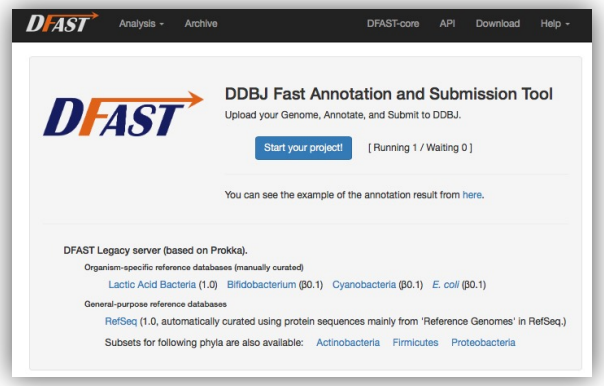
原核生物ゲノムの自動アノテーション

DDBJへのゲノム塩基配列の登録支援機能

高速なアノテーション (5Mbpのバクテリアゲノムで 3～5分)

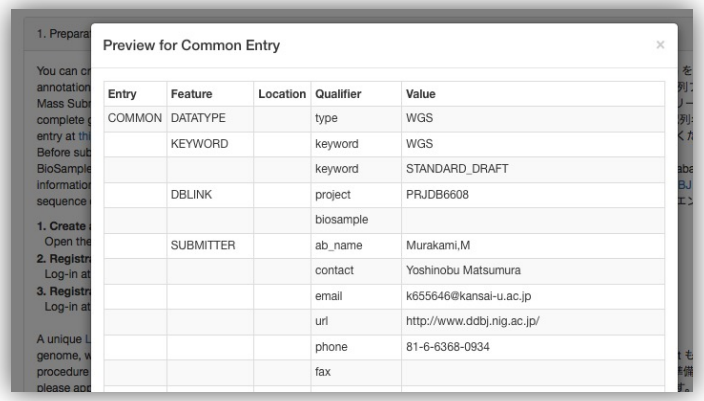
# 2. Web版とStand-alone版が利用可能

## Web version (Graphical user interface face)



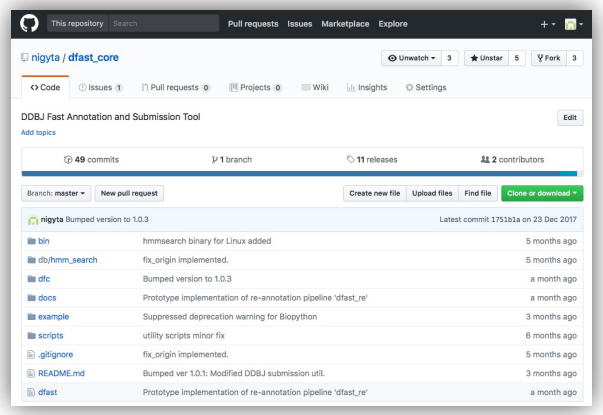

https://dfast.ddbj.nig.ac.jp

ファイルのアップロードのみで実行可能




DDBJ への登録用ファイルをオンラインで作成可能

## Stand-alone version (command-line user interface, Python3, Mac/Linux)




https://github.com/nigyta/dfast_core

シンプルなコマンド

パイプラインのカスタマイズ・拡張

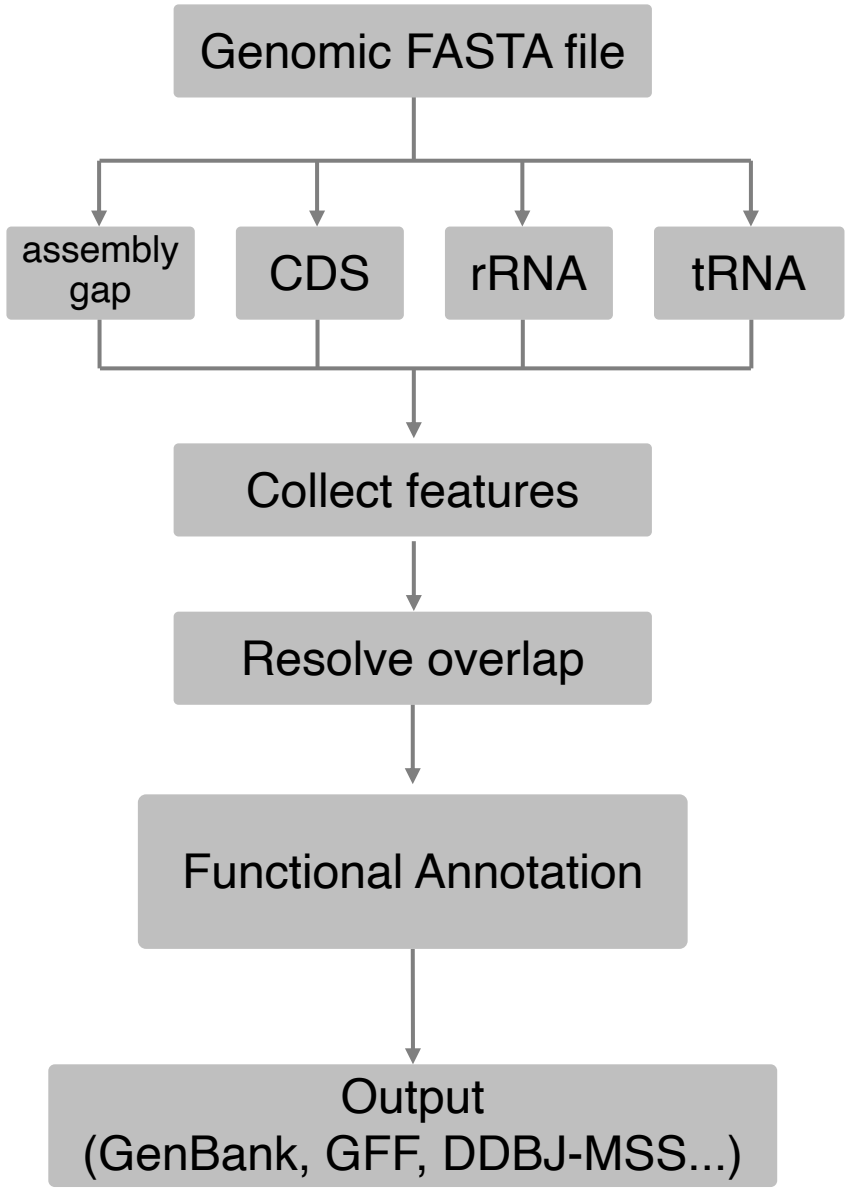
Sample usage

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

Bioconda からもインストール可能

```
conda install -c bioconda dfast
```

# 3. DFASTの処理ワークフロー




## 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

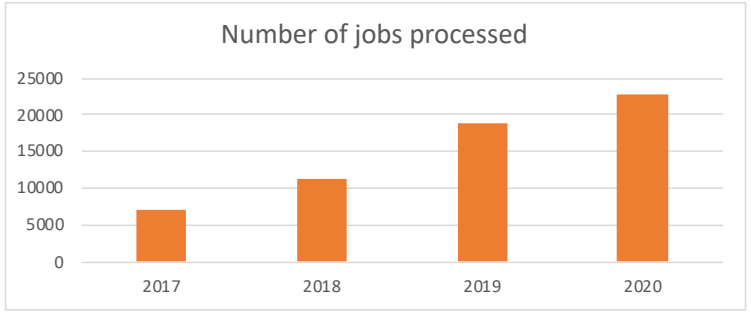
- Default database constructed from 120 representative genomes
- Optional organism-specific database

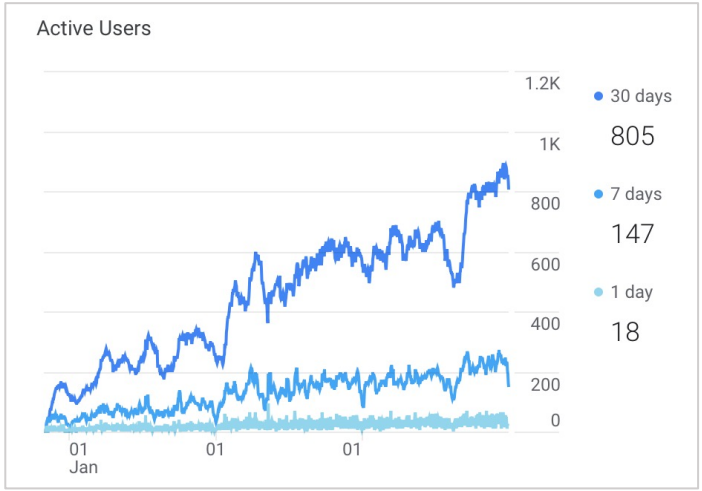
Pseudogene detection

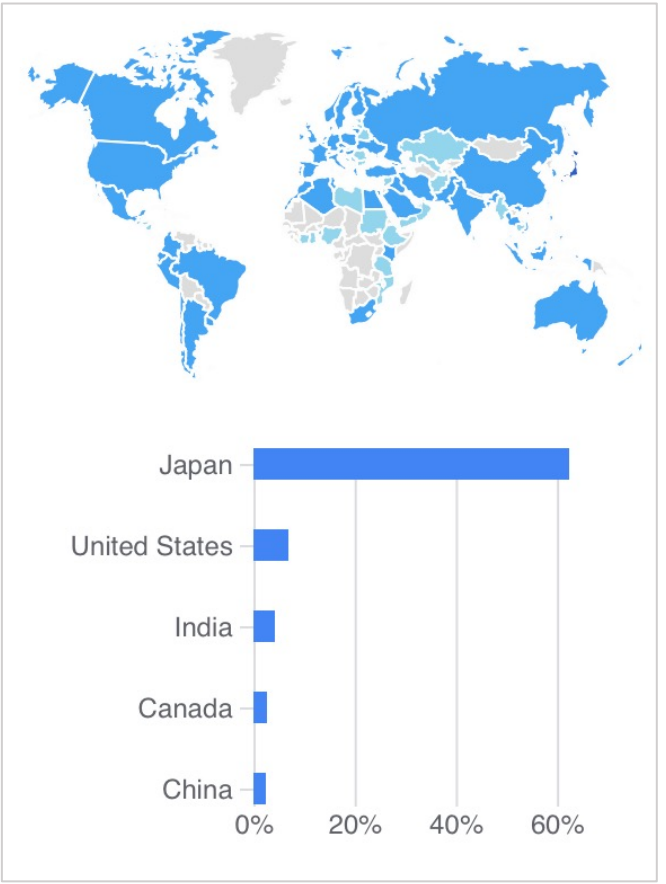
# 4. 利用実績

2019年にDDBJ公式サービスとして採用

2020年のWeb版での年間ジョブ投入数 22,895 件




DDBJから公開されたアノテーション付きバクテリアゲノムの
92% でDFASTが使用されている (1,716件中1,579件, 2020年)

WDCM* による GCM** 10K type strain sequencing project
の基準株ゲノム約600件が DFAST を使ってDDBJに登録された

http://gcm.wdcm.org/

* World Data Centre for Microorganisms
** Global Catalogue of Microorganisms

# 5. 遺伝研スパコンへのサーバー移転と今後の計画

## 2021.2.22　遺伝研スパコンへのサーバー移転

最大同時実行可能ジョブ数の増大（２ → ８）

ジョブあたりのCPU数・メモリの増強（１CPU → ３CPU）

実行時間の短縮（5Mbp 程度のバクテリアゲノムで 4 min → 2 min）

クオリティチェック機能 DFAST_QC の正式運用開始

## 今後の拡張計画 (Web版)

DDBJ ユーザーアカウントとの連携

→ ユーザーごとのジョブ履歴の管理を容易化

DDBJ 登録システムとの統合

→ DFAST の画面上でDDBJへの登録作業が可能に

# 6. 公共塩基配列DBにおける諸問題

## Quality (accuracy)

他の菌株・生物種のコンタミネーション

シークエンス・アッセンブリのエラー (indelエラーによるフレームシフト)

## Completeness

Partial / draft / complete

必須と思われる遺伝子が欠如している

## Authenticity

生物種の誤同定、生物種名の記載の間違い

細分類されて使用されなくなった生物種名

## Traceability

メタデータの不足や誤りによる検索性の低下・データ再利用の阻害

# 7. DFAST_QC の概要

参照マーカー遺伝子配列への相同性検索と
average nucleotide identity (ANI) を組み合わせた高速な種同定

CheckM (Parks, 2014) を用いたcompleteness と
contamination の評価

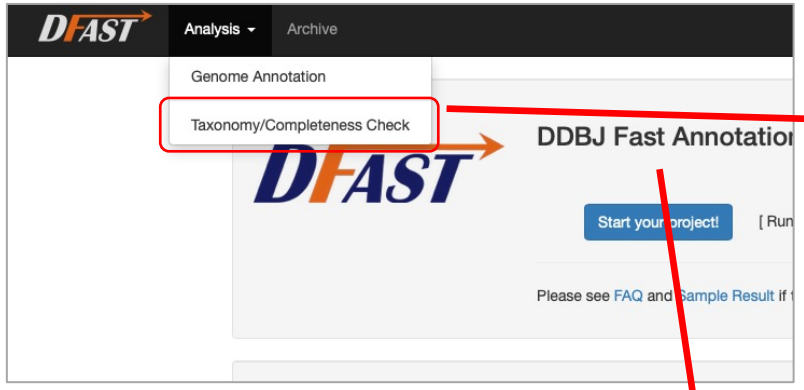
13,000 件以上の参照基準株ゲノムデータ

バクテリア・アーキアに対応

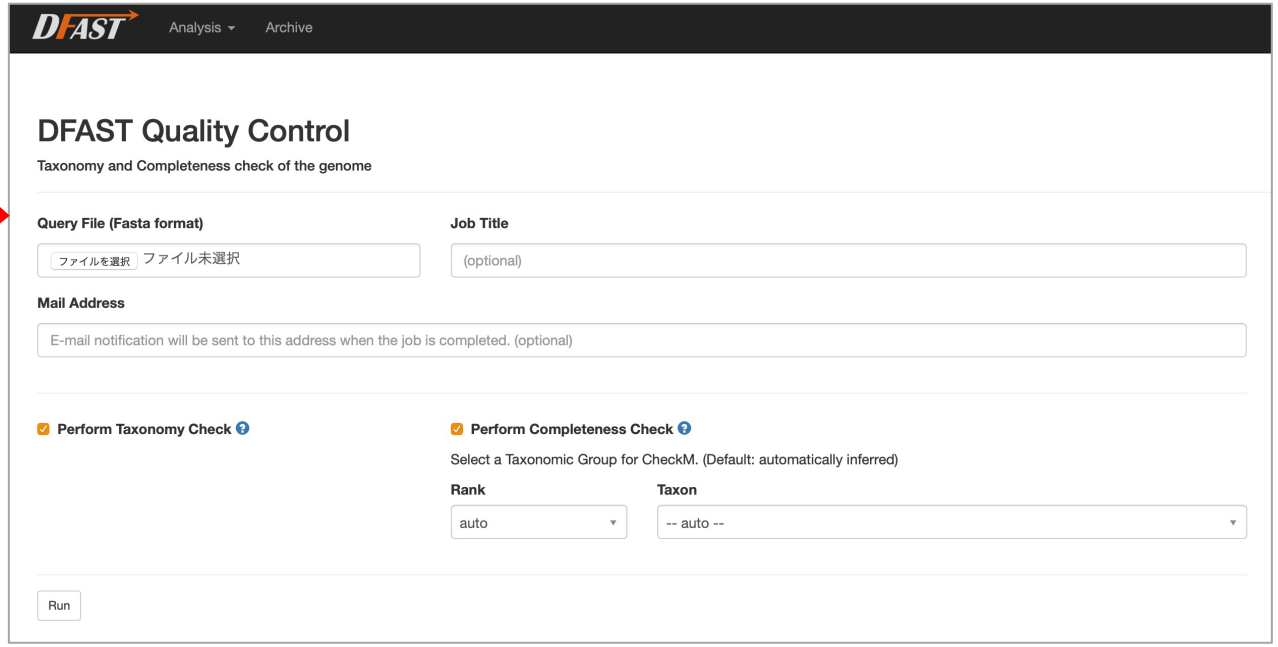
Web版と stand-alone 版が利用可能

# 8. ジョブの投入

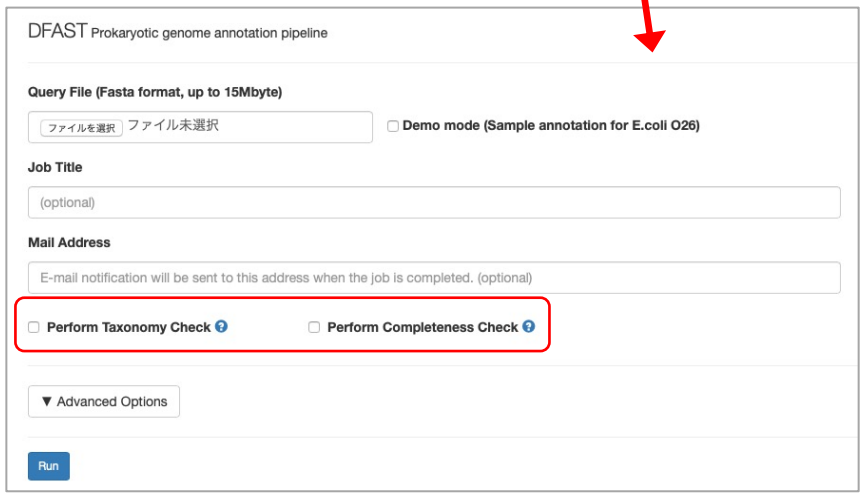
メニューバーから Analysis →
Taxonomy/Completeness Check を選択




https://dfast.ddbj.nig.ac.jp




FASTA形式のゲノムをアップロード




◀ DFAST のジョブ投入時に、
Perform Taxonomy/Completeness Check
を有効にするとアノテーションとともにクオリティ
チェックも実行される。

# 9. 結果画面

Query genome: *Corynebacterium stationis* strain ATCC 6872 (GCA_001561975.1)

**Taxonomy check**

The organism name inferred from ANI result is Corynebacterium stationis.

**Download Results**

| Organism name | strain | Accession | Taxonomy ID | Relation to type | Validated* | ANI (%) | Matched fragments | Total fragments |
|---|---|---|---|---|---|---|---|---|
| Corynebacterium stationis | strain=622=DSM 20302 | GCA_001941345.1 | 1705 | type | True | 98.2198 | 876 | 951 |
| Corynebacterium casei | strain=LMG S-19264 | GCA_000550785.1 | 160386 | type | True | 83.0316 | 652 | 951 |
| Corynebacterium ammoniagenes | strain=DSM 20306 = 9.6 | GCA_001941425.1 | 1697 | type | True | 82.7836 | 648 | 951 |
| Corynebacterium ammoniagenes | strain=DSM 20306 | GCA_000164115.1 | 1697 | type | True | 82.5756 | 650 | 951 |
| Corynebacterium camporealensis | strain=DSM 44610 | GCA_000980815.1 | 161896 | type | True | 78.4683 | 228 | 951 |
| Corynebacterium camporealensis | strain=CIP 105508 | GCA_000766885.2 | 161896 | type | True | 78.3596 | 226 | 951 |

* Based on NCBI Assembly Report and ANI report. See NCBI ANI report README.

**Completeness check**

**Completeness** 99.49%
**Contamination** 0.49%
**Download Results**

Please refer to the CheckM web site for the description of reported statistics.

## [Result]

Identified as *Corynebacterium stationis* by DFAST_QC (ANI 98.2%)

Completeness: 99.49%, Contamination: 0.49%

# 10. DFAST_QC パイプラインの概要 1

## 参照ゲノムデータの取得

NCBI Assembly DB で公開されている
ゲノム配列 (>700,000 genomes)

NCBI Assembly Report を参照

基準株ゲノムを抽出

NCBI ANI Report, NCBI Tax Dump を参照して検証

Reference data for Taxonomy check

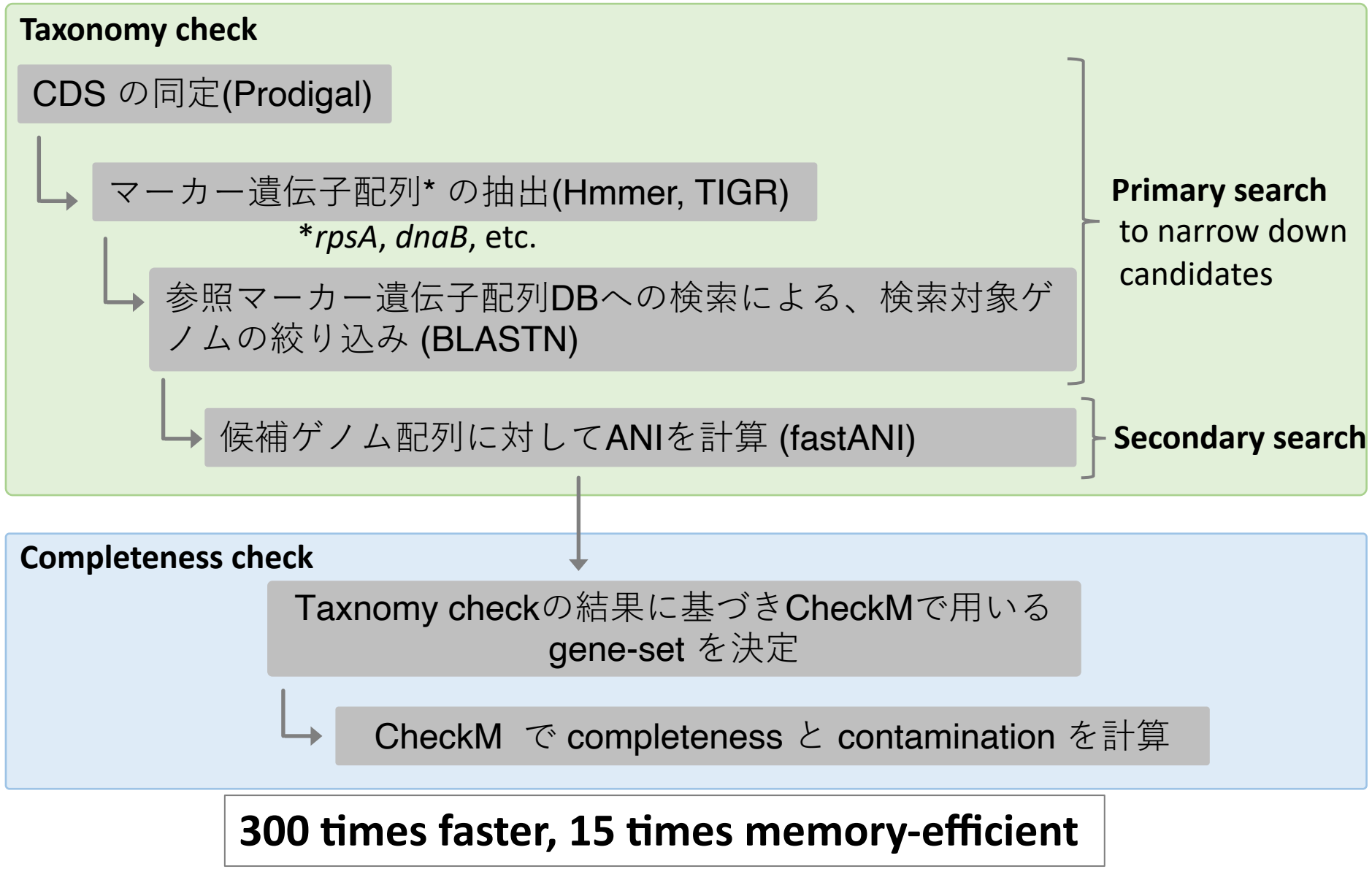
検証済みの基準株ゲノム配列
*10,574種　13,700 件
2020-09-15現在

Prodigal でCDS予測。*rpsA, dnaB, recA, gyrA, pheS, ksgA* を抽出

参照マーカー遺伝子塩基配列

- 最初のデータ取得に約 1 日
- 自動更新 (1ヶ月に1回程度を予定)

# 11. DFAST_QC パイプラインの概要 2




**300 times faster, 15 times memory-efficient**

# 12. ベンチマーク用データの準備

NCBI Assembly DB で公開されている
バクテリアゲノム配列取得 (*742,505 genomes)

* as of 2020-09-15

基準株ゲノムを除く

種名が同定されていないゲノム**を除外 (*637,824 genomes)

** Uncultured bacterium, *Clostridiales* bacterium,
*Spiroplasma* endosymbiont, *Lactobacillus* sp.　... etc,

**Dataset A**: ランダムに抽出した10,000件のゲノム配列

→ 病原菌等の比較的研究事例の多い菌種の結果を反映

**Dataset B**: 1菌種1ゲノムとなるようにランダムに抽出した 5,445件のゲノム配列

→ より多様な菌種の結果を反映

# 13. ベンチマーク結果 (Dataset A)

| classification | # of genomes (percentage) |
|---|---|
| 記載された生物種名と一致 | 9,669 (96.7%) |
| 〃　　不一致　*1 | 199 (2.0%) |
| Secondary search でのaccepted hit (ANI>95%) なし　*2 | 101 (1.0%) |
| Primary search で候補ゲノムがみつからなかった | 30 (0.3%) |
| マーカー遺伝子がみつからなかった*3 | 1 (0.0%) |

*1 *Shigella sonnei* (130 genomes, 参照ゲノムが存在しない)
*Yersinia pestis* (9 genome , 参照ゲノムが存在しない)
*Bacillus cereus* (8 genomes, 登録者の誤同定と推測される)

*2 93件では同じ属の別の菌種に対してもっとも高い ANI を示していた。
61件に関しては同じ菌種の参照ゲノムが存在していない。

*3 ゲノムサイズ が小さすぎるのが原因と考えられる

# 14. ベンチマーク結果 (Dataset B)

| classification | Reference genome | | Total (percentage) |
|---|---|---|---|
| | available | missing | |
| 記載された生物種名と一致 | 3,074 *1 | | 3,074 (56.5%) |
| 〃 不一致 | 121 | 328 | 449 (8.2%) *2 |
| Secondary search でのaccepted hit (ANI>95%) なし | 246 | 1,281 | 1,527 (28.0%)*2 |
| Primary search で候補ゲノムがみつからなかった | 77 | 315 | 392 (7.2%) *2 |
| Failedマーカー遺伝子がみつからなかった | 3 | | 3 (0.1%) |
| Total | 3,521 | 1,924 | 5,445 |

*1 同じ菌種の参照ゲノムが存在していれば、 87% のケースで一致 (3074/3521).

*2 43.4% のケースで、様々な理由により不一致 または 同定の失敗

登録者による誤同定の可能性
ゲノムのクオリティが低い
種分類の結果が反映されていない

参照ゲノムが存在しない
あいまいな種名、タイポ

# 15. DFAST_QC まとめ

原核生物ゲノムのクオリティチェックツール DFAST_QC を新規開発した

DFAST Webサービスから利用可能で、FASTAファイルのアップロードを
アップロードするだけで実行できる。

ベンチマークの結果、研究事例の多い菌種については96.7%のケースで記
載された生物種名と一致した。

DDBJ に登録されるゲノムの Quality, Completeness, Authenticity の向上
に役立つと期待される。

[今後の進展] DDBJの登録システムとの統合により、登録・公開処理の自動化実現、メタデータ
バリデーション機能の強化Traceabilityの向上を目指す