トーゴーの日シンポジウム 2020

ChIP-Atlasの二次解析による多彩な転写因子認識配列の解明および更なるデータベース化

2020 / 10 / 05

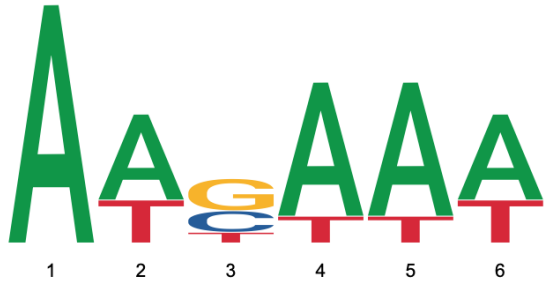
○田原 沙絵子[1], 土屋 貴穂[1,2], 尾崎 遼[1,2]

1. 筑波大学 医学医療系 生命医科学領域 バイオインフォマティクス研究室
2. 筑波大学 人工知能科学センター

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

Background

転写因子の認識配列は
位置特異的重み行列 (PWM) で表現される



PWMはゲノム上の位置間の相互依存性が損失する

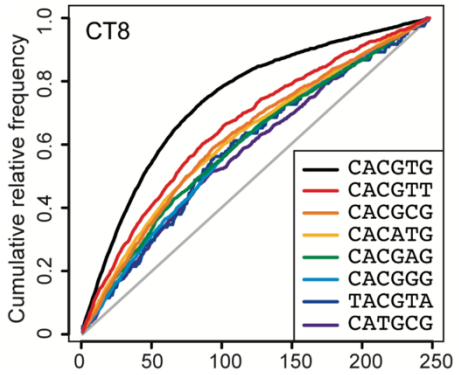
PWMは頻度の低い認識配列が見逃されている

ChIP-seqデータが蓄積され
統合データベースが開発されている

ex. ChIP-Atlas

[Shinya Oki et al. 2018. EMBO reports]

MOCCS は転写因子認識配列を k-mer として正確に抽出できる



k-mer	MOCCS2score
CACGTG	102.8
CACGTT	37.7
CACATG	58.1
CACGCG	42.4
CACGAG	28.1
CATGCG	21.6

*[**Haruka Ozaki** and Wataru Iwasaki. 2016. Comput. Biol. Chem.]*

MOCCSはゲノム上の位置間の相互依存性を損失しない

MOCCSは頻度の低い認識配列も正確に抽出できる

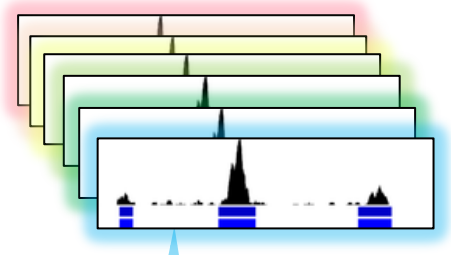
MOCCS と ChIP-Atlas を使えば、

多彩な転写因子認識配列が捉えられる！

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

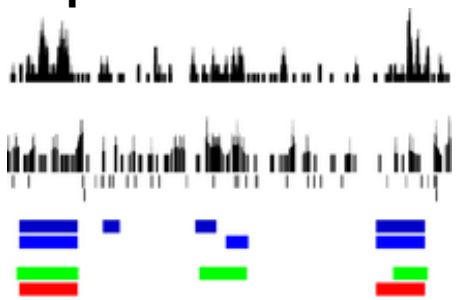
Method

ChIP-Atlas



BED file
~140,000

ChIP-seq peak call data



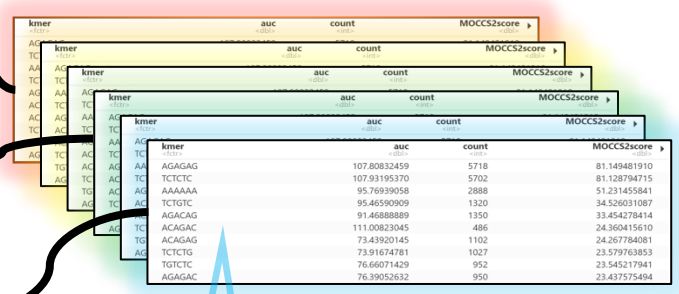
MOCCS

MOCCS output

Antigen
Cell type

Antigen
Cell type

Antigen
Cell type

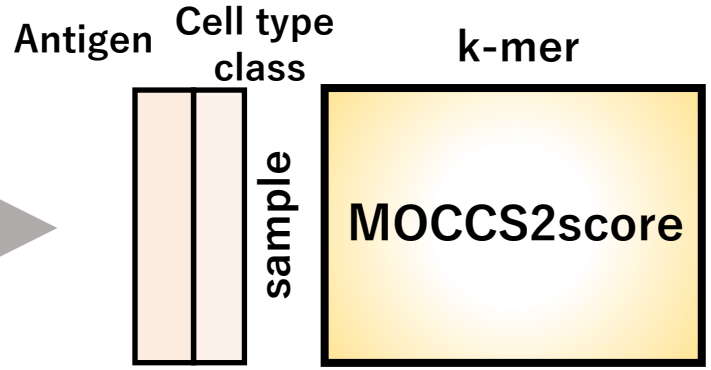


ID : SRX…
Antigen : XX , Cell type class : YY

Rank	Sequence	AUC	Appearance count	MOCCS2 score
D-box#1	TTATGCAA	94.9	537	30.9
D-box#2	TTATGTAA	60.4	390	16.7
D-box#3	TGATGTAA	71.8	189	13.9
D-box#4	TTATGTCA	71.8	119	11.0
D-box#5	TTGTGTAA	64.7	132	10.4
D-box#6	TTATACAA	60.5	75	7.3
D-box#7	TTAGGTAA	60.1	74	7.3

k-mer **MOCCS2score**

k-mer profile



Database development

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

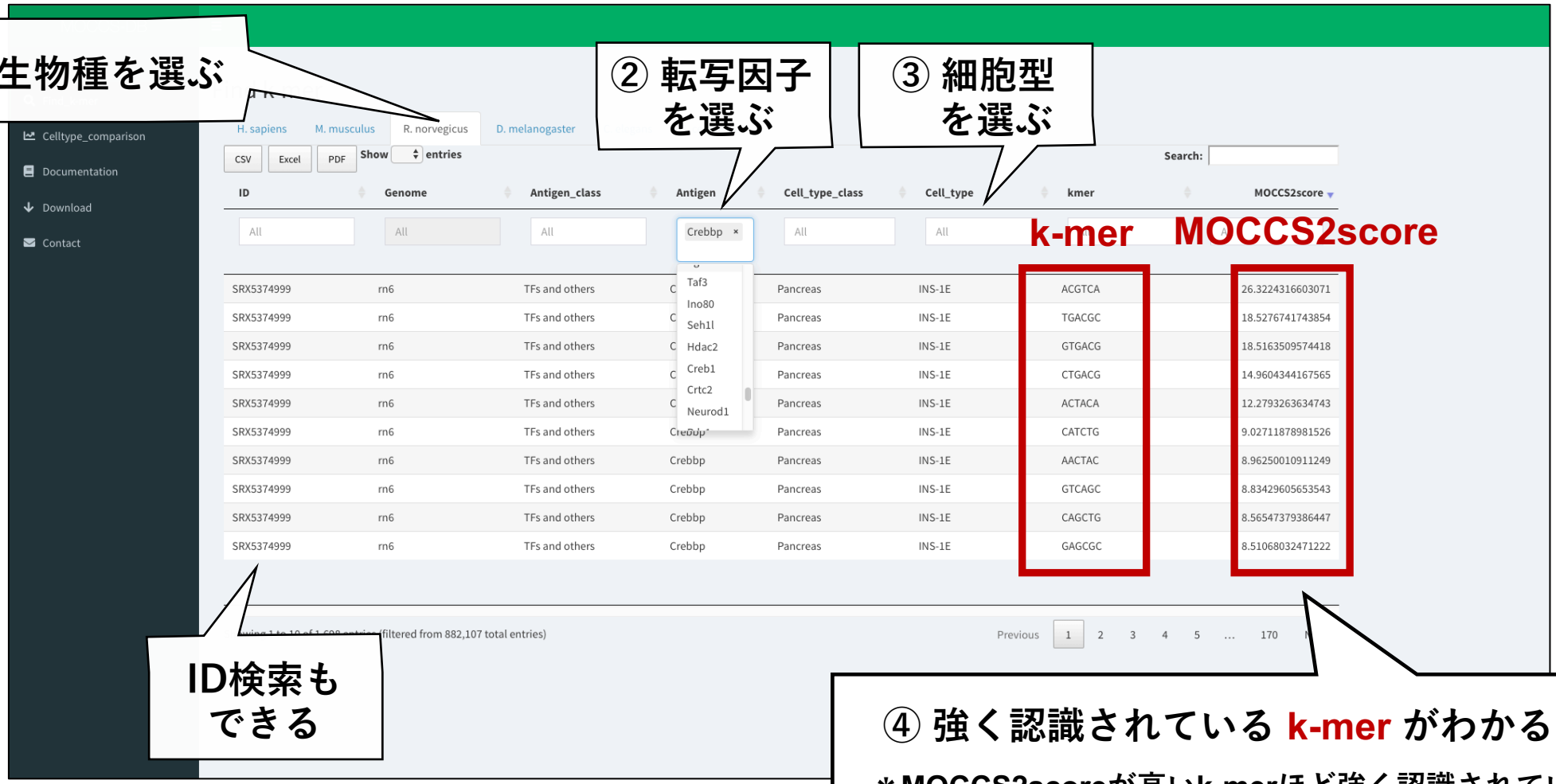
k-mer profile をデータベースとして公開予定

- **k-mer profileとは？**
  - ✓ **ChIP-seq データに MOCCS を適用して得られたデータセットです。**
  - ✓ **転写因子の認識配列が k-mer と MOCCS2score (結合特異性スコア)で表現されています。**
- **従来の PWM で表現されたデータベースと比べて優れている点は？**
  - ✓ **転写因子の認識配列を k-mer と MOCCS2score で得られます。**
  - ✓ **2つのサンプル間で転写因子認識配列の比較ができます。**
  - ✓ **細胞型間で異なる転写因子認識配列が捉えやすくなりました。**
- **データはどのくらいあるのですか？ 実装方法は何ですか？**
  - ✓ **6種の生物種, 10万以上の ChIP-seq データから得られた k-mer profile が入っています。**
  - ✓ **Shinyで実装しています。**

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

データベースの機能 - 1

興味のある転写因子の認識配列が k-mer としてわかる！



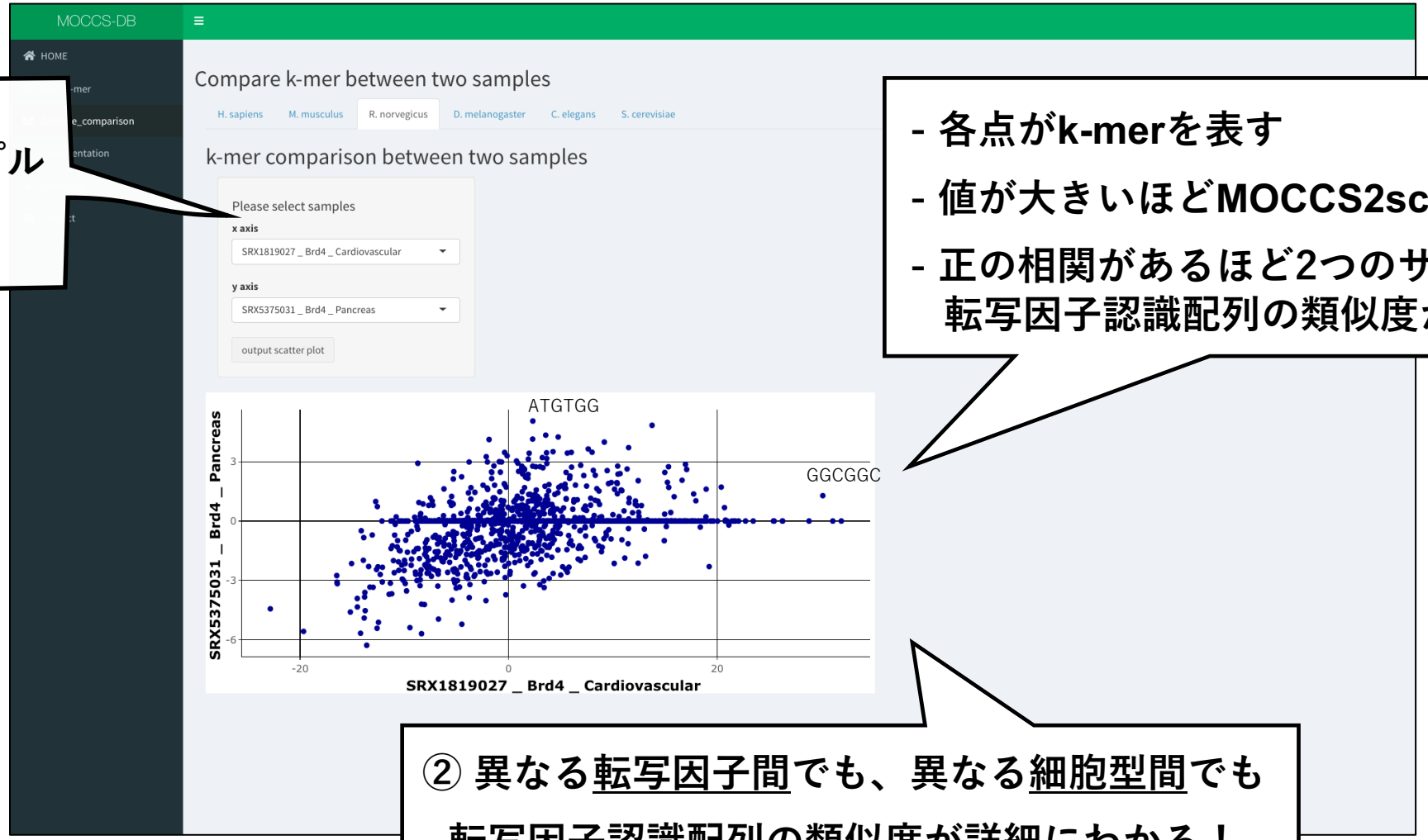
Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

Result

データベースの機能 - 2

2つのChIP-seqサンプル間で変動のある k-mer がわかる！

① 比べたいサンプルを選ぶ



- 各点がk-merを表す
- 値が大きいほどMOCCS2scoreが大きい
- 正の相関があるほど2つのサンプル間で転写因子認識配列の類似度が大きい

② 異なる<u>転写因子間</u>でも、異なる<u>細胞型間</u>でも転写因子認識配列の類似度が詳細にわかる！

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

細胞型間で異なるk-mer profileの例

転写因子認識配列は細胞型ごとに異なる

各sampleでMOCCS2score >= 7の
k-merを抽出

sample1

TTATGCAA
TTATGTAA
TGATGTAA
TTATGTCA
TTGTGTAA
TTATACAA

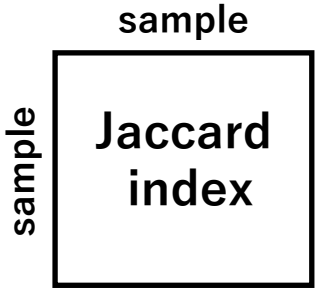
……..

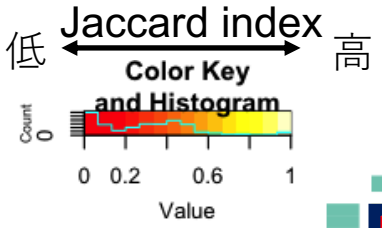
sampleX

TTAGGTAA
TGATGCAA
TTATCTAA
TTACGCAA
TGCTGTAA
TTATGAAA

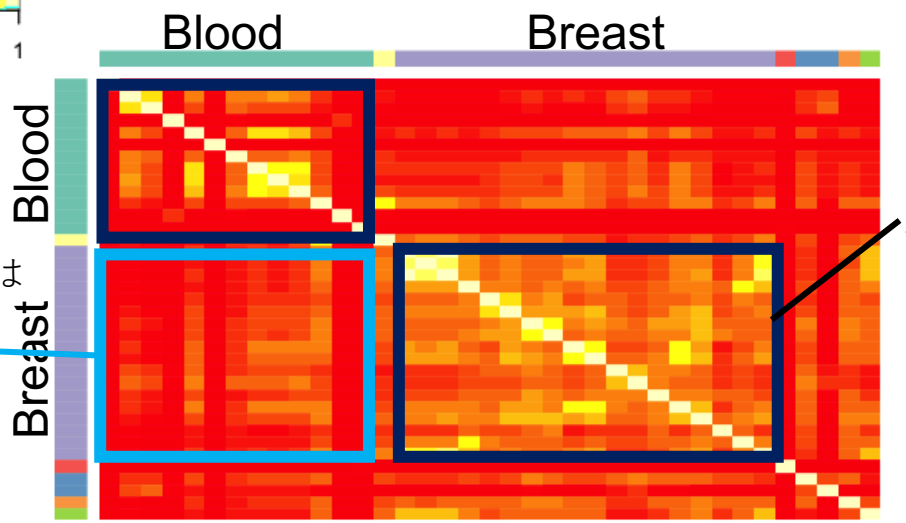
2つのsample間で
jaccard indexを計算

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



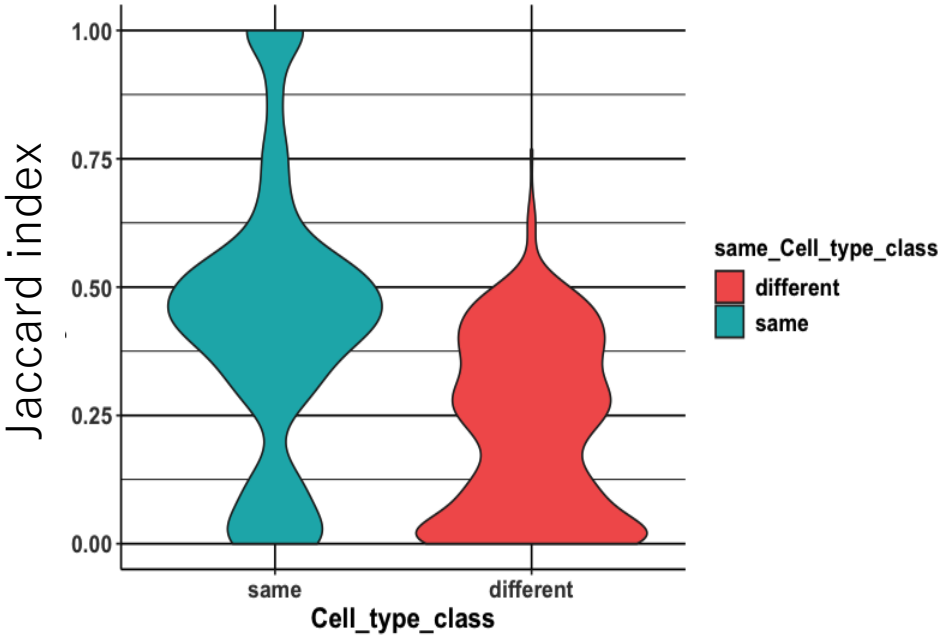


STAT3



違う細胞型同士では
認識配列に
差異がある

同じ細胞型同士は
認識配列が
似ている



＊Jaccard indexが高いほど2sample間の転写因子認識配列の類似度が大きいことを示す

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）

Discussion & Future application

- ChIP-seqデータのクオリティ管理
- 転写因子認識配列に重なるSNPの影響度評価
- 異なる生物種間の転写因子認識配列多様性の理解
- 転写因子同士のネットワーク理解

Acknowledgement

- 大学共同利用機関法人情報・システム研究機構
  データサイエンス共同利用基盤施設
  ライフサイエンス統合データベースセンター

  大田 達郎 博士

- 解析環境

  国立遺伝学研究所 スーパーコンピュータシステム

Licensed under a Creative Commons 表示4.0国際ライセンス
©2020（筑波大学医学医療系 バイオインフォマティクス研究室）