02 第2部

NCBI塩基配列データベースに蓄積するエラー解析

理化学研究所　バイオリソース研究センター　遺伝子材料開発室

三輪　佳宏、飯田　哲史、野﨑　晋五、木嶋　順子、岸川　昭太郎、中島　謙一、
笹沼　俊一、大波　純一、中村　宣篤、村田　武英

問題提起: データベース精度を向上する方法

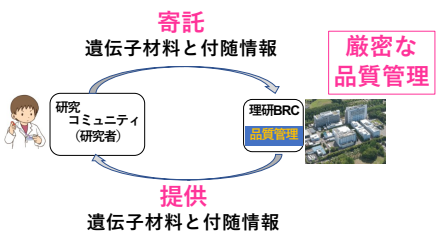
データベースにエラーが蓄積するには複数の原因がある

1）実験そのものの精度の低さ
2）データのコピー＆ペースト
3）人為的ミス
4）バリアント（エラーではない可能性）

原因に応じて、エラー防止のために求められる対策は全く異なる。
→実験研究者から情報研究者まで連携した対策づくりが必要

理化学研究所バイオリソース研究センター 遺伝子材料開発室

日本の遺伝子バンク

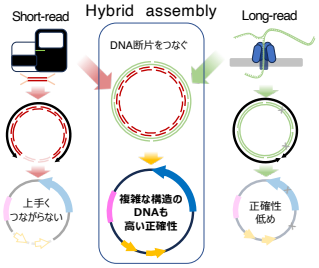


これまでの研究で作製された
品質が保証された豊富な遺伝子材料
→ 効果的、効率的な研究開発を促進

遺伝子材料開発室で遺伝子材料の品質管理
寄託者の意図する遺伝子材料が寄託されたか確認

塩基配列の確定法

HTSを用いたDNA配列検査



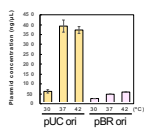
Short readシークエンサーだけでは、タンデムリピートなどの繰り返し配列を正しく解析できない。（左）

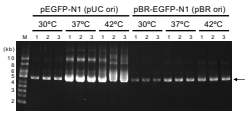
Long readシークエンサーだけでは、細かな塩基配列の正確性や連続塩基などが正しく解析できない。（右）

Short readとLong readを組み合わせて解析することで、正確な塩基配列を高速に決定することが可能になる。（中央）

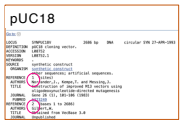
塩基配列エラー実例2　pUC-ori

```
..CGGCTACACTAGAAGAACAGTATTTGGTATCTGCGC..  正しいpUC oriの配列
  ||||||||||||||| ||||||||||||||||||||
..CGGCTACACTAGAAGGACAGTATTTGGTATCTGCGC..  データベース上の配列
```

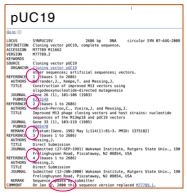




データベースと同じ配列の変異を導入（pBR）して大腸菌でのコピー数を解析した結果
1塩基の違いでコピー数は激減する



↓
エラーのまま未修正



MCS以外は同じ塩基配列の“兄弟ベクター”である
pUC18とpUC19のうち
pUC19のみが2000年に修正されたがpUC18はエラーのまま未修正

国際的塩基配列データベースのデータ検証

高度化した配列解析力＋遺伝子バンクとしてのDNA資料

正しい塩基配列を確定させて、データベース上の配列エラーを探索・修正することが可能

NIH National Library of Medicine
National Center for Biotechnology Information

検証結果の広報 「配列探偵」

1）生物系月刊誌への連載

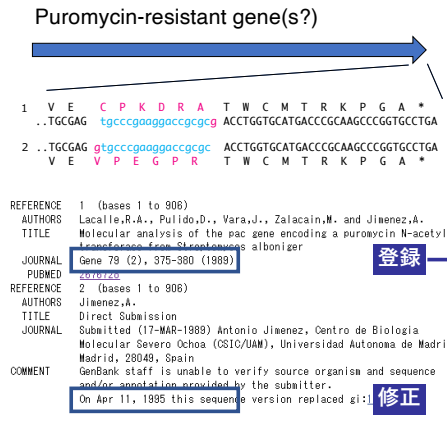
配列探偵 志久延子と弟子入り修行中の江良正、そして、配列探偵の後輩で理研BRCで研究員として働く穴理志津、の３人が、配列エラーに苦しむ研究者を救うためにデータベースの闇と戦う物語　紹介データは全てリアル！

2）RIKEN NEWS

NCBIの塩基配列データベースを検証し、蓄積したエラーを探し出す
リアル配列探偵達の紹介

RIKEN NEWS
「データベースに潜む塩基配列エラーに警戒せよ」 →

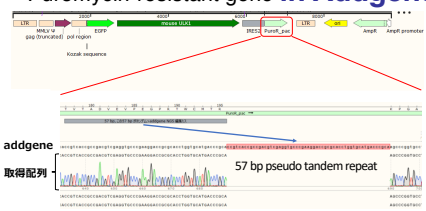
塩基配列エラー実例1　puromycin耐性遺伝子



品質管理検査
２の配列しか出てこない

2022年に新規登録

Puromycin-resistant gene **in Addgene**



Short readのみ
→Assembleミス

← 実在しない幻のタンデムリピート

塩基配列エラー実例3 EF-1α プロモーター

・4types of EF-1 alpha promoter

genome : NC_000006.12

		clones
Type 1. genomic sequence	RDB18336 : pAY5 (mAID-EGFP-NLS piggyBac)	1
Type 2. Dr. Nagata JBC paper	RDB18973 : pEF321-Flag-EX	11
Type 3. Dr. Nagata pEF-BOS plasmid	RDB07939 : pEF-BOS, pEF-BOS-EX	158
Type 4. lenti virus modified?	RDB17867 : CSII-EF-IRES2-Bsd	34

「EF-1α promoter」という１つの名前で異なる塩基配列が流通してしまっているが、そのことを認識している研究者は少ない