

# 微生物ゲノムシーケンシングの 状況とアノテーションの課題

日本微生物資源学会第回大会  
平成15年6月5-6日

平畠壮規<sup>1)2)</sup>, 田中尚人<sup>1)2)</sup>, 丸山穰<sup>1)2)</sup>, 宮崎智<sup>1)3)</sup>, 菅原秀明<sup>1)3)4)</sup>  
1)遺伝研生命情報・DDBJ、2)JST・BIRD、3)総研大・遺伝、4)情報研・生命情報

# DDBJは微生物ゲノムデータをGIBにより提供

## GIB

正式名：Genome Information Broker

- *E.coli* ゲノムプロジェクトのゲノムブラウザーとして開発
- 微生物ゲノム公開後 1 ~ 2 日以内に格納
- 複数のゲノムを比較することも可能

格納ゲノム数  
(5月23日現在)

Archaea	16
Bacteria	107
Eukaryota	6
Total	129

# インターフェースの改良

- GIBのトップページに分類階層を導入
  - 階層段階の調節も可能
  - 名前によりソートした目次もあり
- 個々のゲノムのトップページに機能分類情報などを表示
- ゲノムプロジェクトのページを新設

<http://gib.genes.nig.ac.jp/>

# 微生物ゲノムデータの再評価

GIBにおけるゲノムデータの比較から、  
アノテーションの統一性が欠けていることが  
明らかになってきた

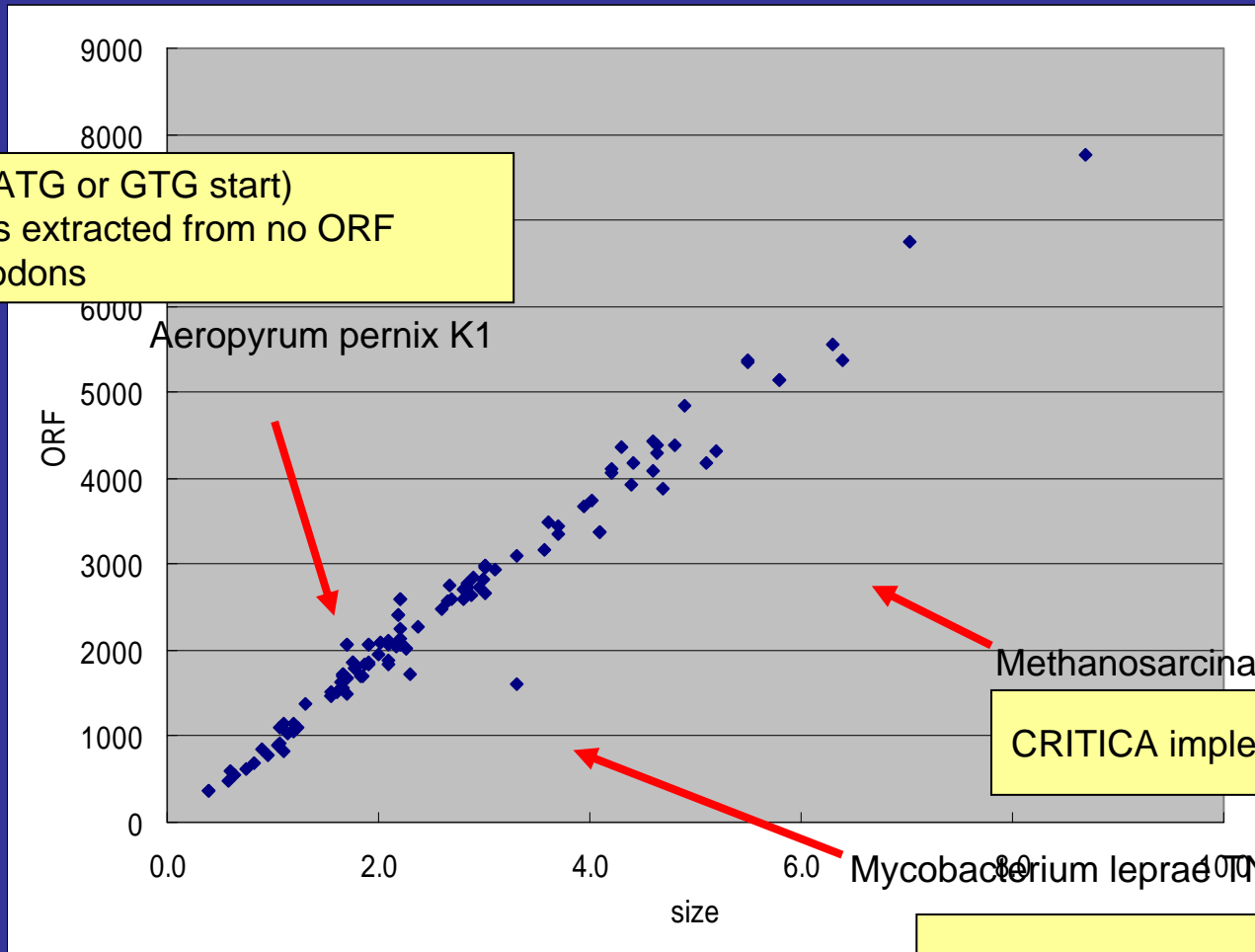
# アノテーションの非統一性の例

Escherichia coli O157:H7 RIMD 0509952	putative virulence protein
Escherichia coli O157:H7 EDL933	putative transposase TnpB of insertion sequence IS609
Escherichia coli K12 W3110	Hypothetical protein
Escherichia coli K12 MG1655	putative virulence protein
Escherichia coli CFT073	Peyer's patch-specific virulence factor GipA

# 微生物ゲノムデータ概観

## ゲノムサイズ対ORF数

>100 codons (ATG or GTG start)  
+ 50-99 codons extracted from no ORF  
having >100 codons



Methanosarcina mazei Goe1

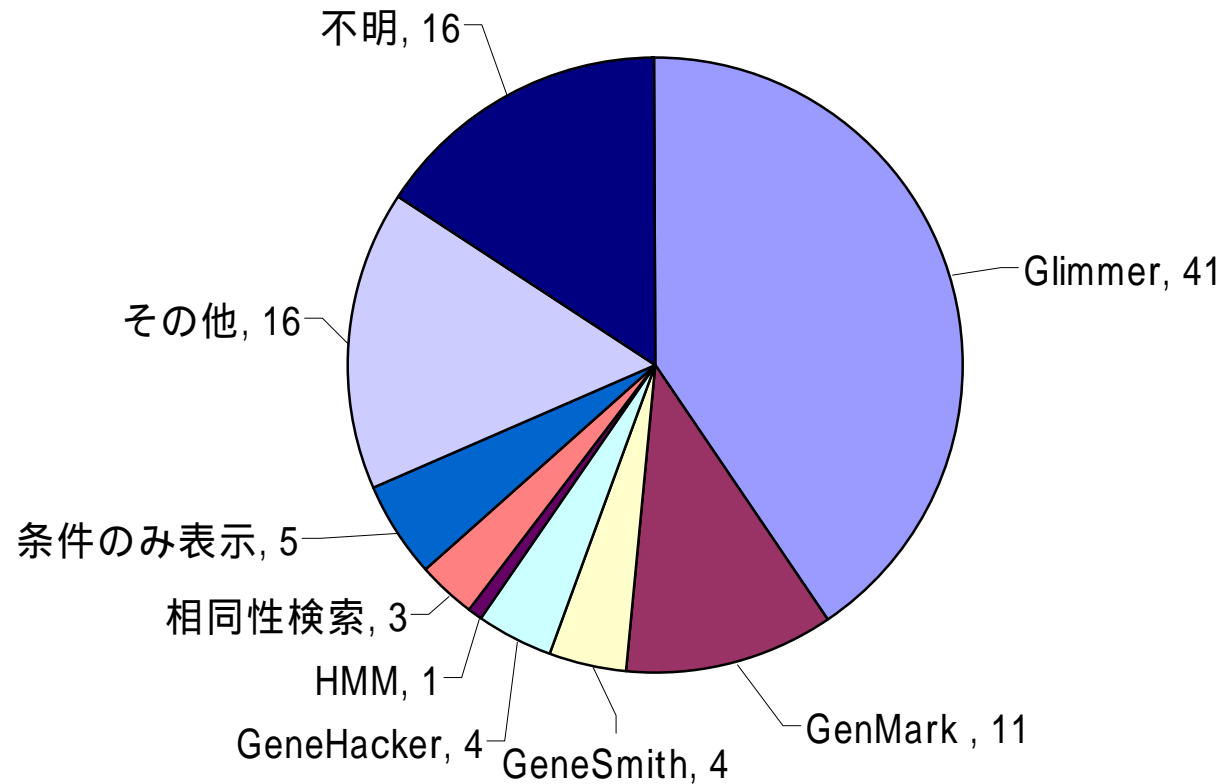
CRITICA implemented in ERGO software

Mycobacterium leprae TN

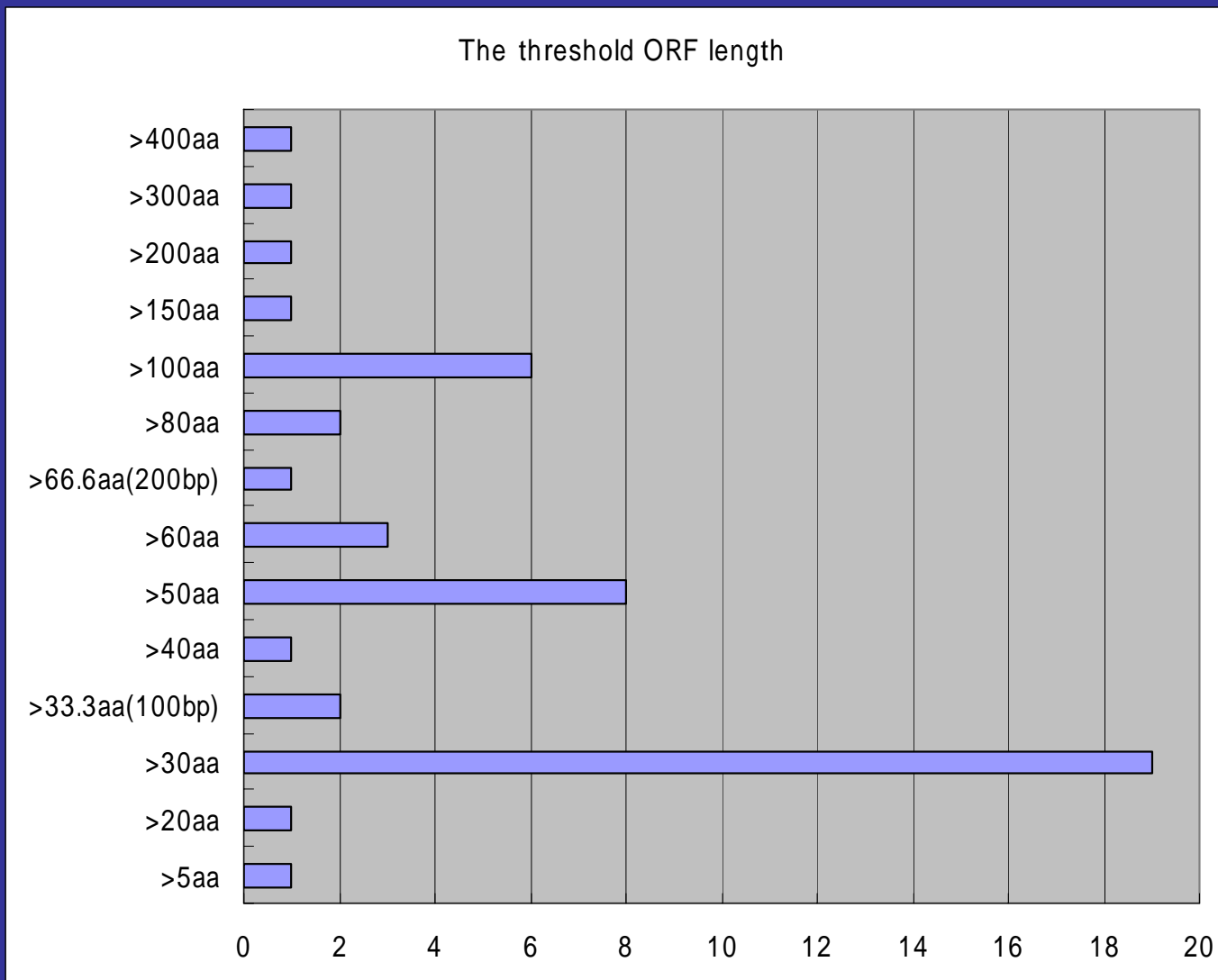
ORPHEUS and GLIMMER (both trained on an initial ORF set generated by ORPHEUS)

# ORF予測方法の多様性

ORF抽出方法



# 最短ORFに対する閾値の多様性



*Escherichia coli* K12 MG1655 Glimmer2 10272件 vs 既存CDS 4289件  
(Glimmer2 ORF最短長 45bp, ORF長以外はすべてデフォルト)

Glimmer2

5'

3'

既存CDS

Case 0

Case 1

Case 2

Case 3

Case 4

Case 5

Case 6

Case 7

Case 8

Case 9

Case 10

2765件

0件

694件

784件

0件

2件

3件

423件

665件

128件

4806件

Glimmerの結果が  
既存CDSとフ  
レームが合  
っている

4245件

Glimmerの結果がフ  
レームと合  
わない

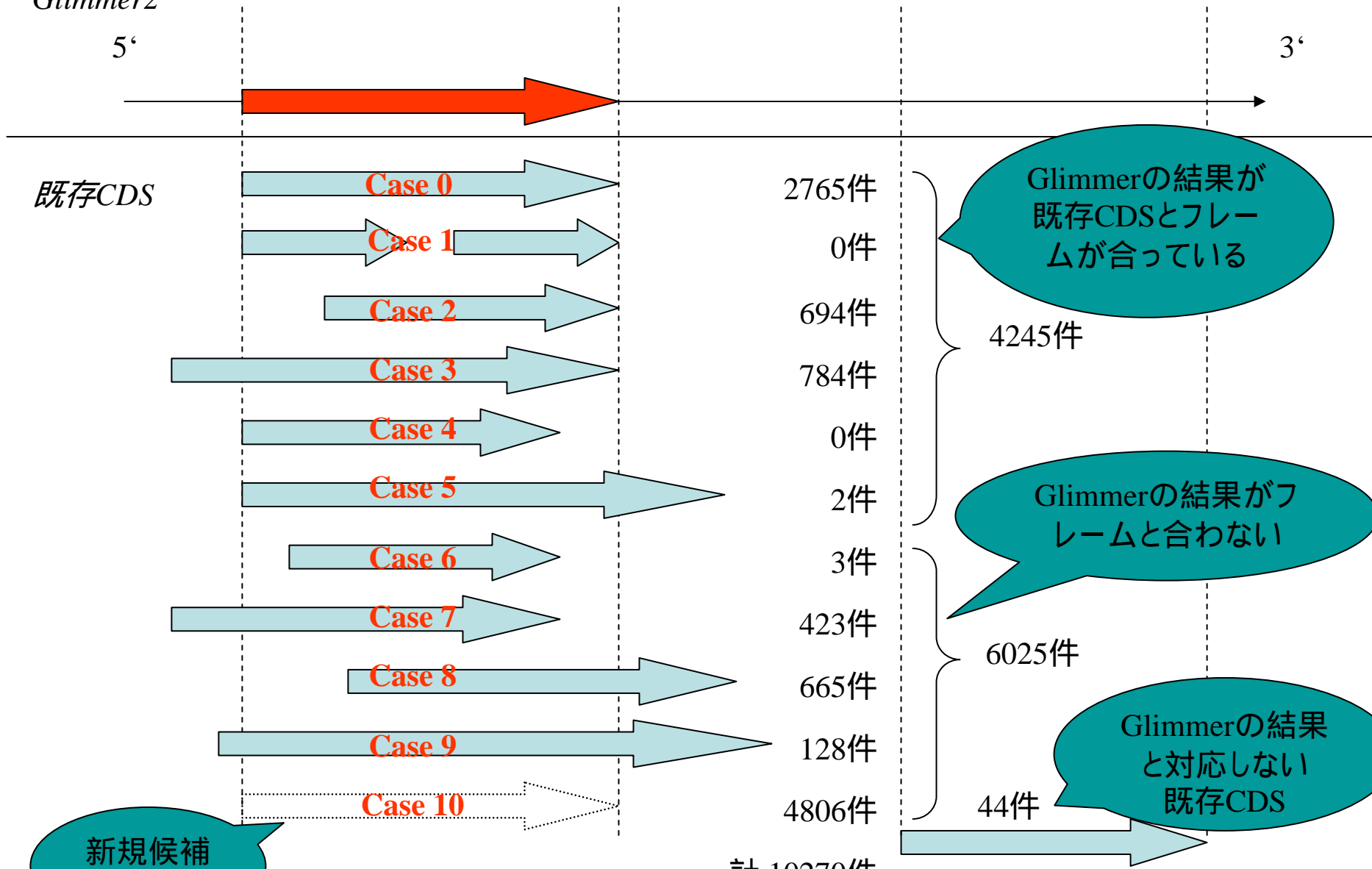
6025件

Glimmerの結果  
と対応しな  
い  
既存CDS

44件

新規候補  
件数

計 10270件



# 評価法検討のために対象とした 微生物ゲノム

<i>Bacillus subtilis</i>	実験データによるORF情報の蓄積が多い
<i>Escherichia coli</i> K12	実験データによるORF情報の蓄積が多い
<i>Streptomyces coelicolor</i> A3(2)	GC high, ゲノムサイズ大
<i>Nostoc sp.</i> PCC 7120	シアノバクテリア、ゲノムrearrangementが起こる
<i>Mycobacterium leprae</i>	pseudo geneが多い
<i>Yersinia pestis</i> KIM	病原菌、pseudo gene、frameshiftが多い
<i>Aeropyrum pernix</i> K1	古細菌 : Crenarchea
<i>Pyrococcus horikoshii</i> OT3	古細菌 : Euryarchea

カテゴリー	数	%
RBSあり、50bp以上補正、5'、3'end共に一致	1068	1.27
RBSあり、50bp未満補正、5'、3'end共に一致	2064	2.46
フレーム不一致、BLAST,motif Hitあり、GIB_ORFと重複なし、mRNA Hitあり	7	0.01
フレーム不一致、BLAST,motif Hitあり、GIB_ORFと重複あり、mRNA Hitあり	6	0.01
RBSあり、補正なし、5'、3'end共に一致	9955	11.86
RBSあり、補正なし、3'endのみ一致	3290	3.92
フレーム不一致、BLAST,motif Hitあり、GIB_ORFと重複なし、mRNA Hitあり	16	0.02
RBSなし、補正なし、5'、3'end共に一致	3458	4.12
フレーム不一致、BLAST,motif Hitあり、GIB_ORFと重複なし、mRNA Hitあり	15	0.02
RBSあり、50bp以上補正、3'endのみ一致	5862	6.98
RBSあり、50bp未満補正、3'endのみ一致	3730	4.44
RBSあり、補正あり、BLAST,motif Hitあり、GIB_ORFと重複なし、mRNA Hitなし	64	0.08
RBSあり、補正あり、BLAST,motif Hitあり、GIB_ORFと重複あり、mRNA Hitなし	233	0.28
RBSあり、補正なし、BLAST,motif Hitあり、GIB_ORFと重複なし、mRNA Hitなし	121	0.14
RBSなし、補正なし、3'endのみ一致	1772	2.11
BLAST,motif Hitあり、GIB_ORFと重複なし、mRNA Hitなし	63	0.08
対象外ORF	52205	62.20
総計	83929	100

# 予測ORF対GIB\_ORFのLocation比較

カテゴリー	数	%
5'と3'が一致	16,545	52.16%
3'のみ一致	14,654	46.19%
一致なし	525	0.02%
合計	31,724	100.00%
対象外	52,205	

# まとめ

- 登録ゲノム数の増加に堪えて、GIBの利用者インターフェースを改良した。
- GIBに集約した微生物ゲノムの比較から、微生物ゲノムデータの再評価が必要と考えた。
- ゲノムプロジェクトごとに、ゲノム解析の手法と適用条件が異なり、単純な比較は誤った結果をもたらす恐れがある。
- 共通の解析手法・条件設定で微生物ゲノムを再評価する価値があると思われる。

# 謝辞

インフォコム株式会社  
ライフサイエンス本部

富士通株式会社  
ライフサイエンスシステム事業部

三井情報開発株式会社  
バイオサイエンス本部研究開発部