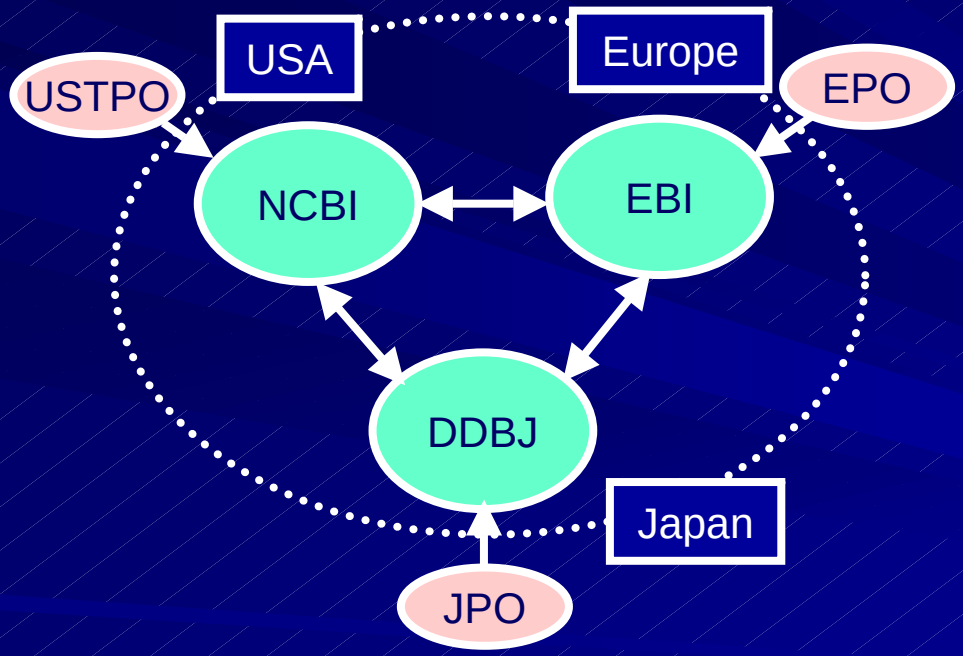
微生物ゲノムの共通プロトコルによる遺伝子配列情報の提供

国立遺伝学研究所
生命情報・DDBJ研究センター
阿部貴志

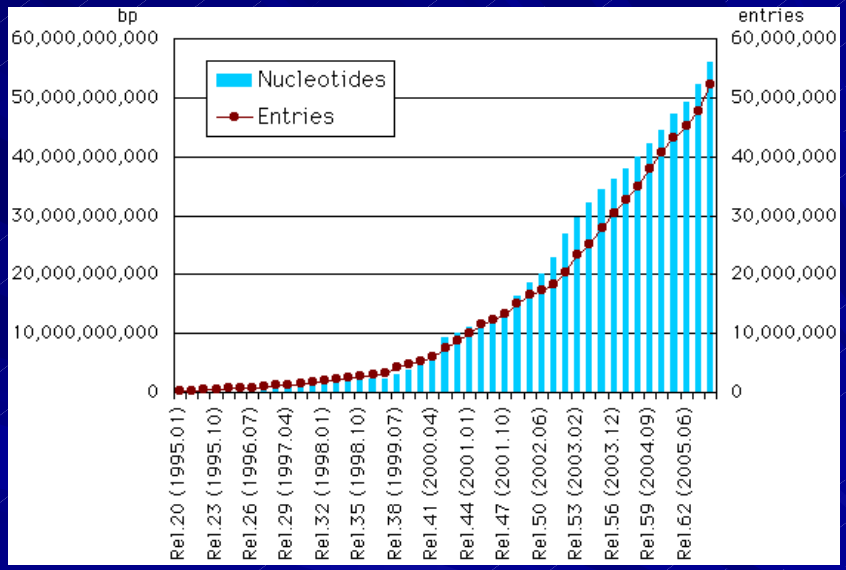
http://gtps.ddbj.nig.ac.jp/

DNA Data Bank of Japan (DDBJ)

International Nucleotide Sequence Database Collaboration (INSDC)



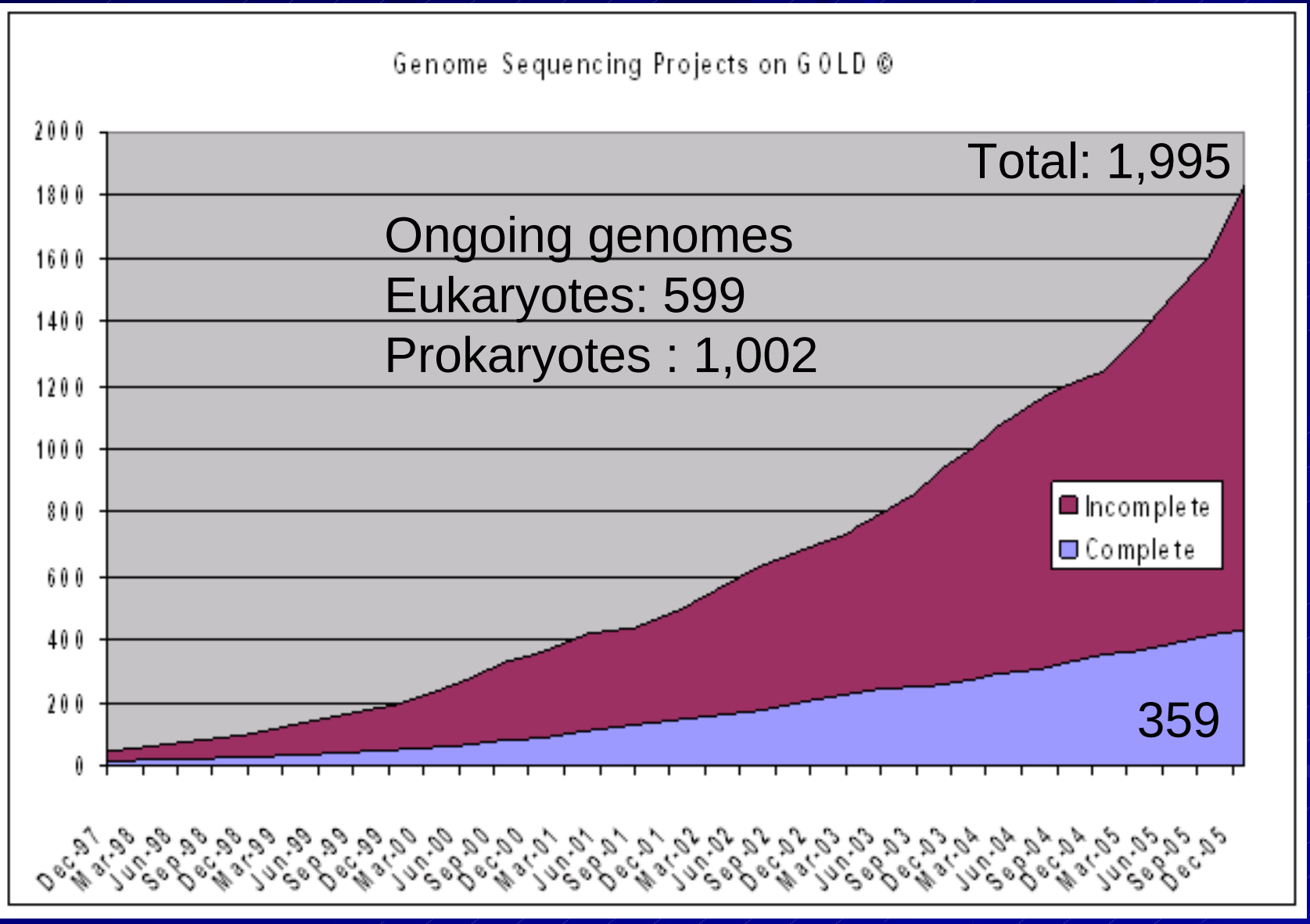
Growth of the International Nucleotide Sequence Database (INSD)



> 50 M rentries

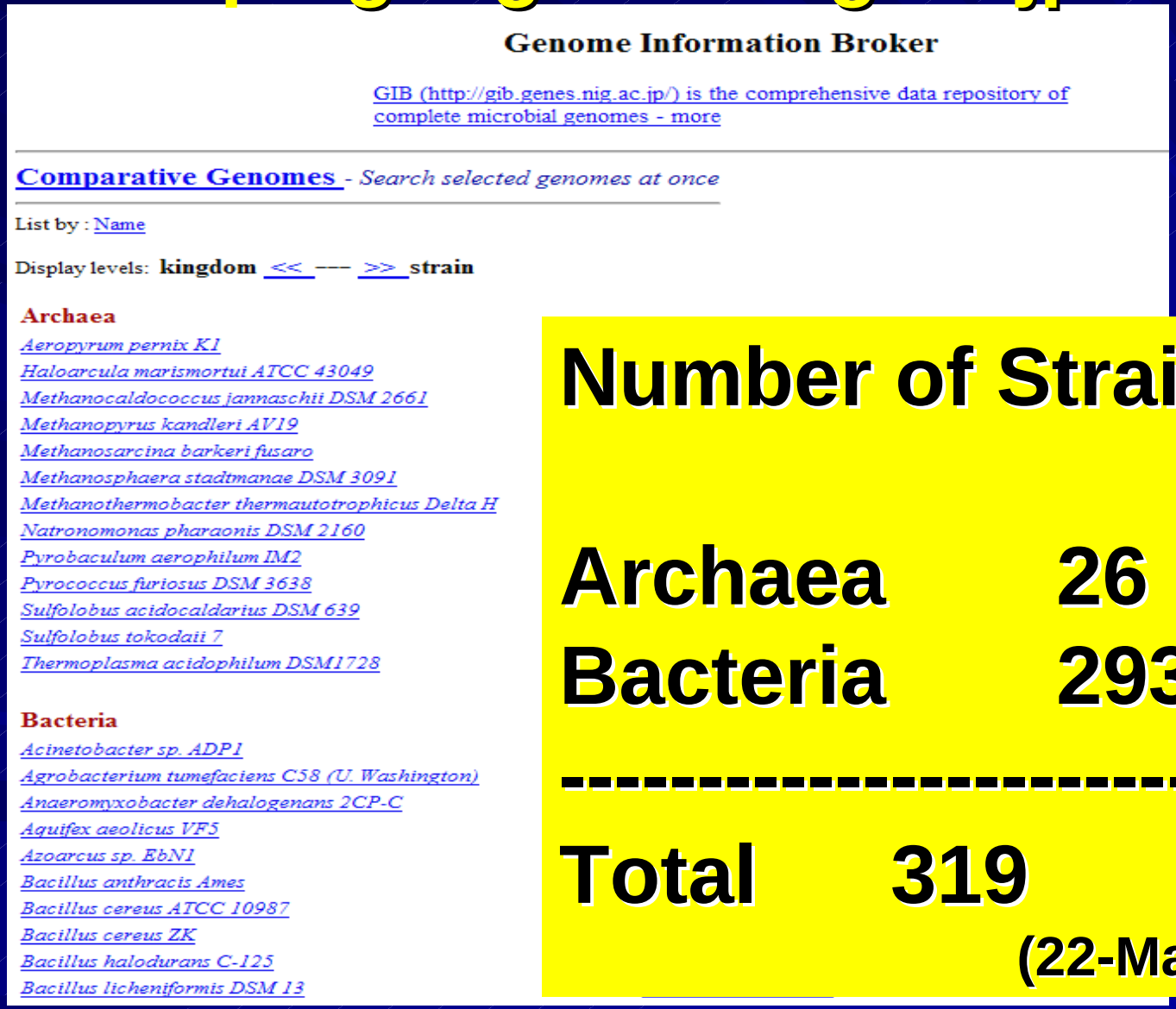
> 50 billion nucleotides

Genome Projectの動向



Genome OnLine Database (http://www.genomesonline.org/)

GIB: Genome Information Broker
http://gib.genes.nig.ac.jp/



Genome Information Broker

GIB (http://gib.genes.nig.ac.jp/) is the comprehensive data repository of complete microbial genomes - more

Comparative Genomes - *Search selected genomes at once*

List by : Name

Display levels: **kingdom** << --- >> **strain**

Archaea

Aeropyrum pernix K1
Haloarcula marismortui ATCC 43049
Methanocaldococcus jannaschii DSM 2661
Methanopyrus kandleri AV19
Methanosarcina barkeri fusaro
Methanosphaera stadtmanae DSM 3091
Methanothermobacter thermautotrophicus Delta H
Natronomonas pharaonis DSM 2160
Pyrobaculum aerophilum IM2
Pyrococcus furiosus DSM 3638
Sulfolobus acidocaldarius DSM 639
Sulfolobus tokodaii 7
Thermoplasma acidophilum DSM1728

Bacteria

Acinetobacter sp. ADP1
Agrobacterium tumefaciens C58 (U. Washington)
Anaeromyxobacter dehalogenans 2CP-C
Aquifex aeolicus VF5
Azoarcus sp. EbN1
Bacillus anthracis Ames
Bacillus cereus ATCC 10987
Bacillus cereus ZK
Bacillus halodurans C-125
Bacillus licheniformis DSM 13

Number of Strains

Archaea 26
Bacteria 293

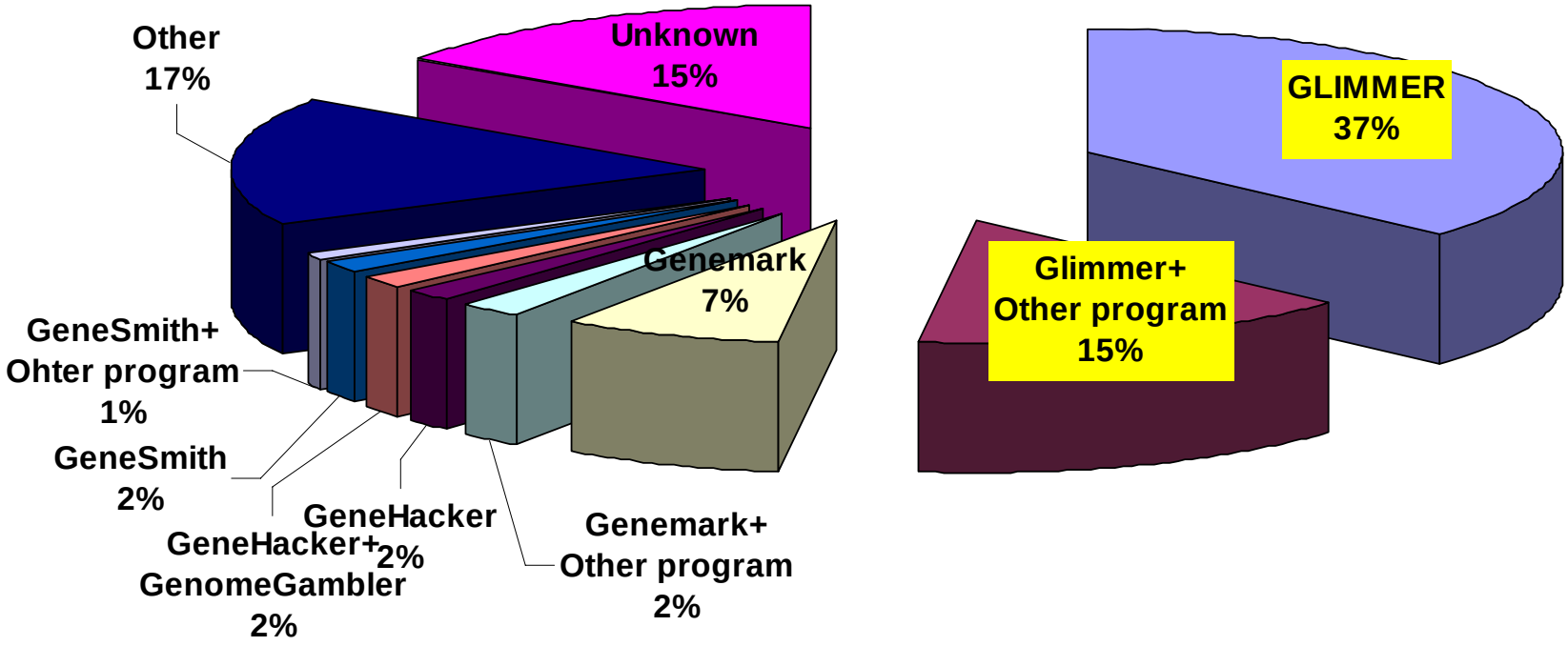
Total 319

(22-Mar-2006)

INSDから公開されているゲノムデータの問題点

- 遺伝子領域予測プログラムの違い
- 最短ORF長の設定が異なる
- 相同性検索のthreshold値の設定の違い
- 相同性検索・モチーフ解析のリファレンスデータベースの差異
- プロダクト記載が不統一
- ORF決定の根拠(確かさ)が不明
- アップデートが不確定

The diversity of ORF prediction programs



Diversity of the minimum length used in prediction program

length	number
＞20	1
＞(＝)30	25
>33.3aa （100bp)	3
>40aa	1
>50aa	7
>60aa	4
>66.6aa (200bp)	1
>80	2
>100aa	6
>150aa	1
>200aa	1
>300aa	1
>400aa	1

同じプロダクトでも記載内容が異なる

~ *Hahella chejuensis* KCTC 2396

```
     CDS             1023521..1024429
                     /gene="argB"
```

~ *Archaeoglobus fulgidus* DSM 4304

```
 CDS             complement(1141715..1142587)
                     /locus_tag="AF_1280"
```

~ *Agrobacterium tumefaciens* C58 circular chromosome

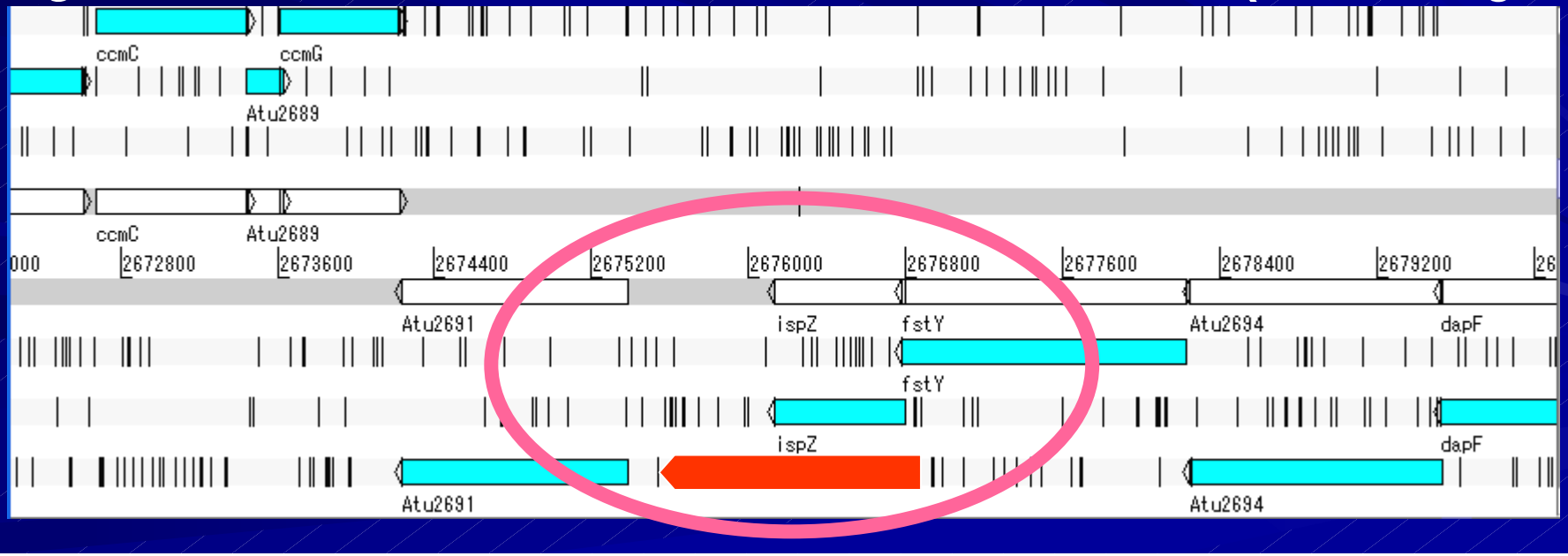
```
     CDS             complement(373582..374466)
                     /gene="AGR_C_666"
                     /note="acetylglutamate kinase PA5323 {imported} -
                     Pseudomonas aeruginosa (strain PA01)"
                     /codon_start=1
                     /transl_table=11
                     /product="AGR_C_666p"
                     /protein_id="AAK86197.1"
                     /db_xref="GI:15155294"
                     /translation="MTSSESEIQARLLAQALPFMQKYENKTIVVKYGGHAMGDSTLGK
                     AEDIALLKQSGINPIVVHGGGPQIGAMLSKMGIESKFEGGLRVTDAKTVEIVEMVL
                     SINKEIVALINQTGEWAIGLCGKDGNMVFAEKAKKTVIDPDSNIERVLDLGFVGEV
                     EVDRTLLDLLAKSEMIPVIAPVAPGRDGATYNINADTFAGAIAGALHATRLLFLTDV
                     PGVLDKNKELIKELTVSEARALIKDGTISGGMIPKVETCIDAIKAGVQGVVILNGKTP
                     HSVLLEIFTEGAGTLIVP"
```

/noteに
プロダクト名称

/productに
IDのような記載

Agrobacterium tumefaciens C58 circular chromosome (Cereon)

Agrobacterium tumefaciens C58 circular chromosome (U. Washington)

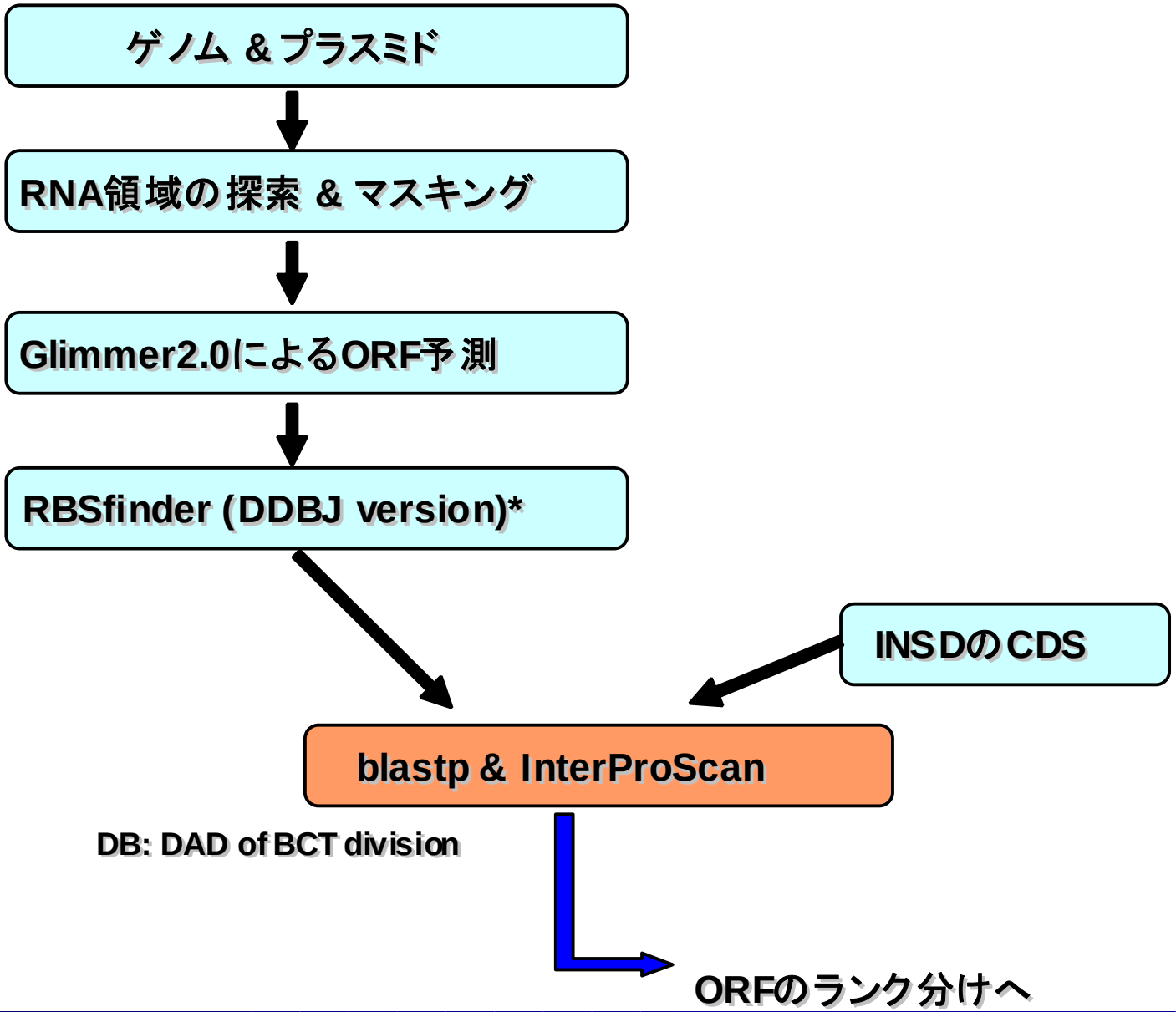


GTPS
~ Gene Trek in Procaryote Space ~

微生物ゲノムの一斉アノテーション
chromosome & plasmid

ver.	strains	Archaea	Bacteria
2003	123	14	109
2004	183	17	166
2005	> 300 strains (予定)		

GTPS overview



Grading of CDS (A and B)

Grade	blastp hit		InterProScan hit
	Coverage	Quality	Quality
AAAA / BBBB	≧70%(CDS Matched protein) & or	Valid protein	Significant motif
AAA / BBB			Unknown motif
AA / BB			No hit
A / B		Putative membrane or unknown proteins	Significant or unknown motif

Grading of CDS (subcategory)

1 = 完全一致

2 = 3'のみ一致

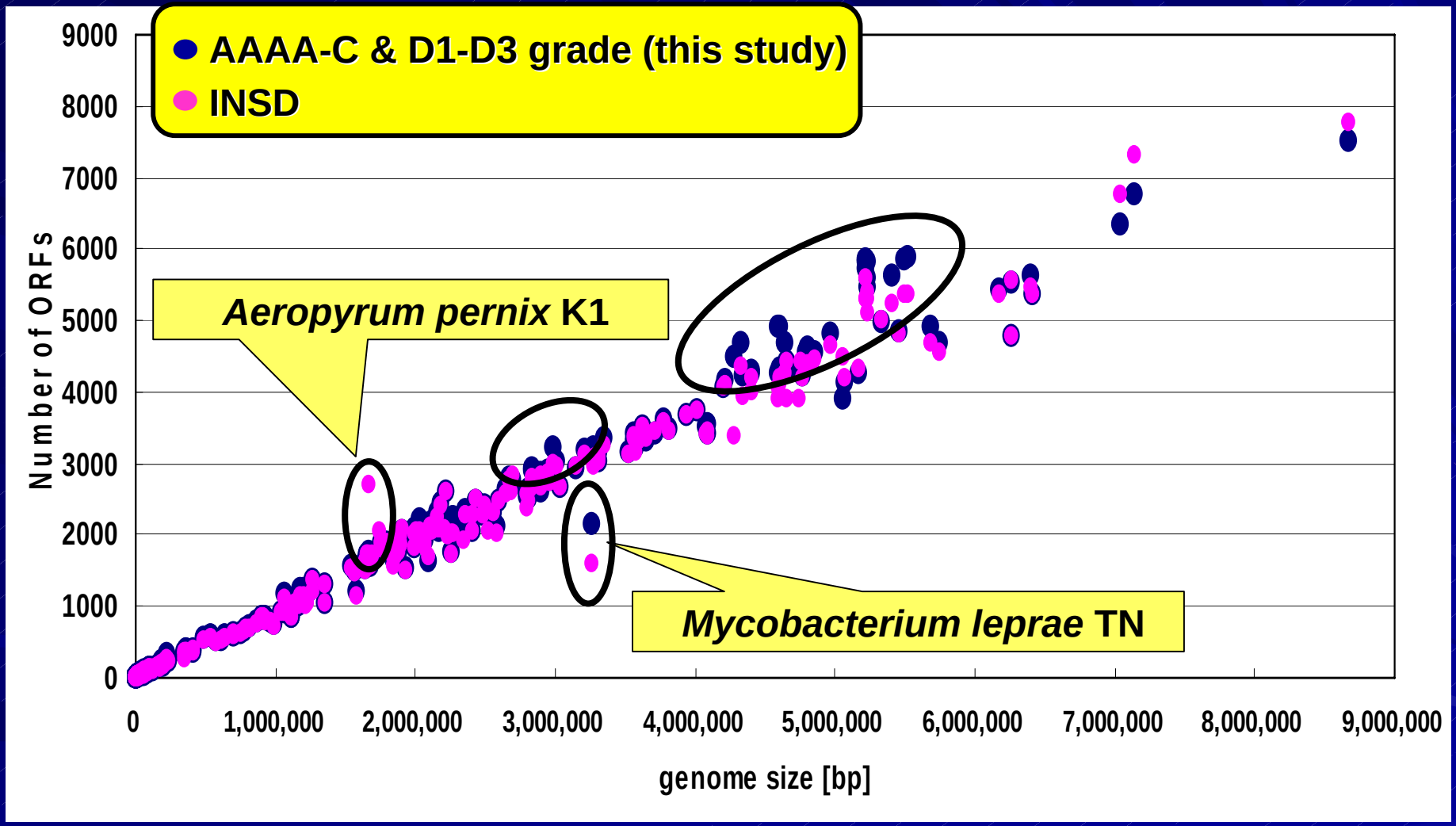
3 = INSDにない
(新規にみつかった)

4 = Glimmerで予測されない

Result

	ver. 2003	ver. 2004
AAAA-A	283,247	431,672
BBBB-B	7,208	10,250
C	4,680	7,511
D	79,779	107,382
E	6,788	10,225
X	466,681	687,110
	848,383	1,254,150

Correlation of number of ORFs and genome size

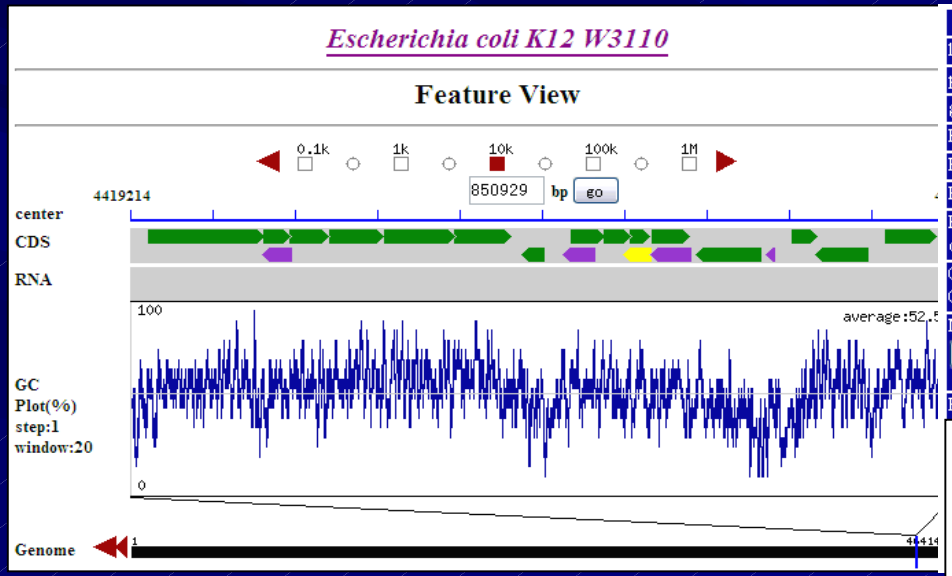


E. coli K12のアノテーション
(Acc#: U00096 & AP009048)に
加えられたORF

ECK4368:JW5891:b4568
/gene="ytjA"
/product="hypothetical protein"
/translation="MVKETLMFRWGIIFLVIA
LIAAALGFGGLAGTAAGAAKIVFVVGII
LFLVSLFMGRKRP"

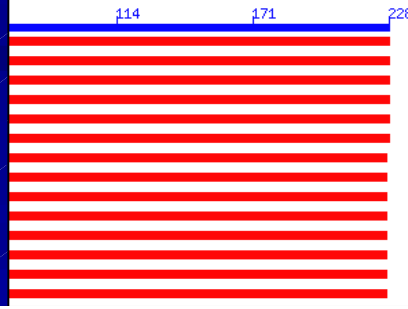
GTPS annotation data are freely available at

http://gtps.ddbj.nig.ac.jp/



CDS	
location	4544541..4547177
product	FimD protein
grade	AAAA1
Predicted by	Glimmer (option minimum length = 60 AA & 15 AA)
RBS	4544533..4544537 gagga
Relation of GIB ORF	identical
BLAST against BCT division	Sorted by score/Sorted by taxonomy class
Comparison of CGM ORFs	Homologues in CGM ORFs graded A-E
Motif Graphic View	IPR000015 Fimbrial biogenesis outer membrane usher protein
Protein Structure	GTOP

species	#of hits	
Bacteria;Escherichia		download multi Fasta
Escherichia coli K12	10	download multi Fasta
Escherichia coli CFT073	5	download multi Fasta
Escherichia coli	13	download multi Fasta
Escherichia coli O157:H7	3	download multi Fasta
Escherichia coli O157:H7 EDL933	3	download multi Fasta



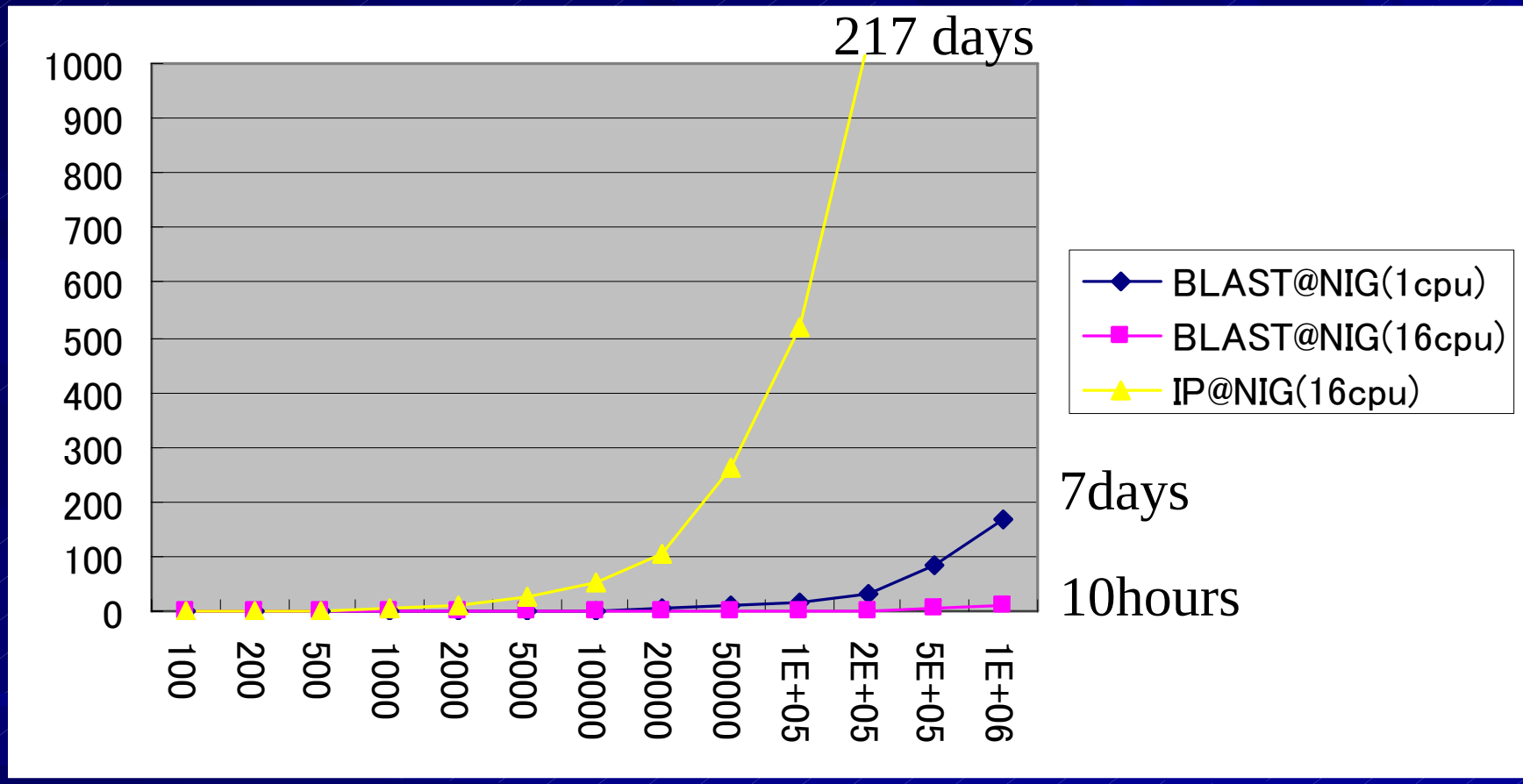
E-value	Score(bits)	ProteinID	Definition
1e-133	472	AAC77155.1	AE000491-10|AAC77155.
1e-133	472	AAA97094.1	U14003-110|AAA97094.1
1e-133	472	AAC77155.1	U00096-4055|AAC77155.
1e-131	466	AAG59394.1	AE005652-11|AAG59394.
1e-131	466	BAB38597.1	AP002568-167|BAB38597
1e-130	464	AAN83709.1	AE016771-220|AAN8370
9e-79	293	AAA23464.1	M15263-3|AAA23464.1|
3e-78	291	AAG54365.1	AE005183-2|AAG54365.1
3e-78	291	AAC73172.1	AE000116-7|AAC73172.1
3e-78	291	AAC73172.1	U00096-62|AAC73172.1|
3e-78	291	BAB96630.1	D10483-52|BAB96630.1|
3e-78	291	CAA39519.1	X56048-1|CAA39519.1|
3e-78	291	AAA24405.1	M62646-1|AAA24405.1|
3e-78	291	BAB33488.1	AP002550-65|BAB33488.

GTPSにかかる計算時間に関して

	プログラム名	プログラムの機能
1	Glimmer	遺伝子領域探索プログラム
2	BlastP	アミノ酸配列に基づく相同性領域の探索
3	InterPro	遺伝子領域内のモチーフ探索

この protocaol では、InterProの計算時間が他と比べ、圧倒的にかかる。そのため、PCクラスタ上での分散環境を構築し、実行を行っている。

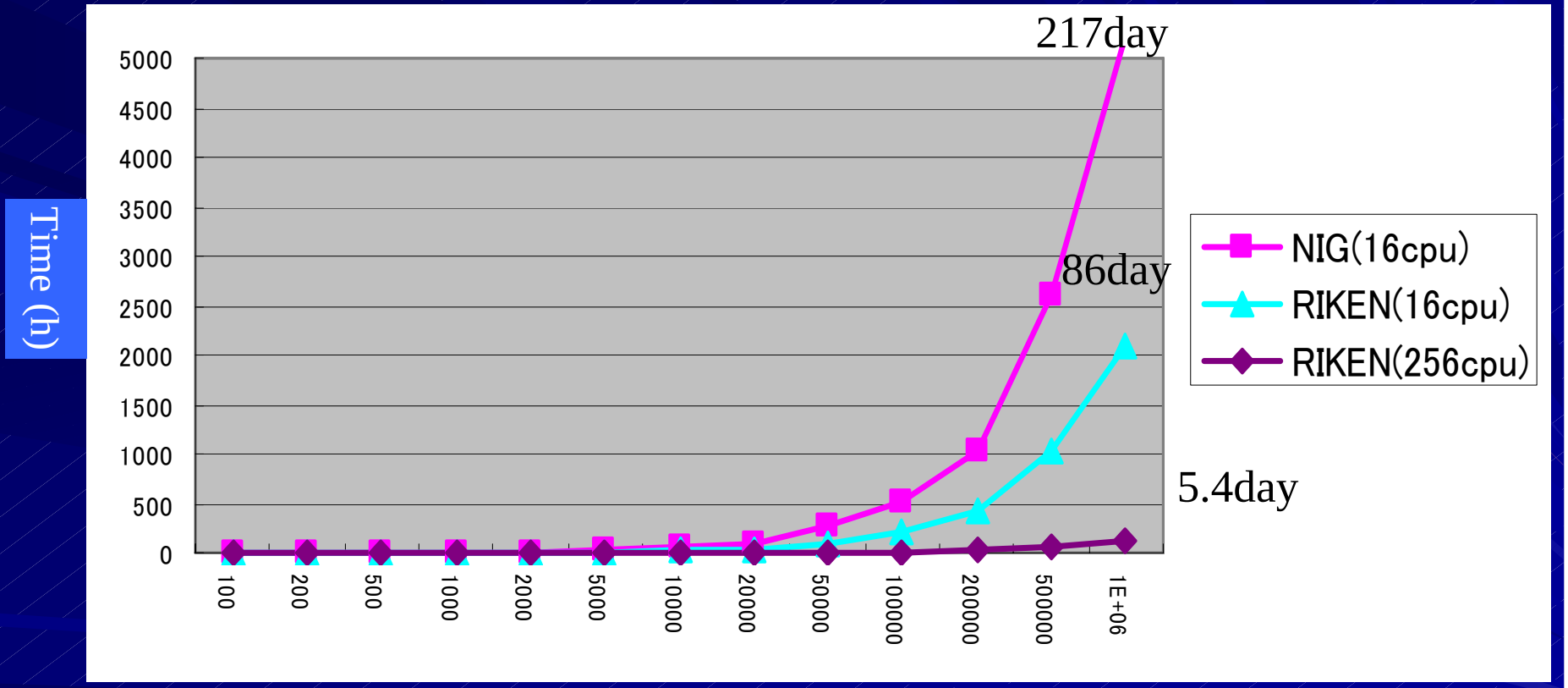
BLAST とInterProScan(IP)との計算時間の計測



・約400倍も計算時間がかかる。

BLAST/1CDS : 0.6s*
InterPro(IP)/1CDS : 5min*
*: Xeon 3.2GHz

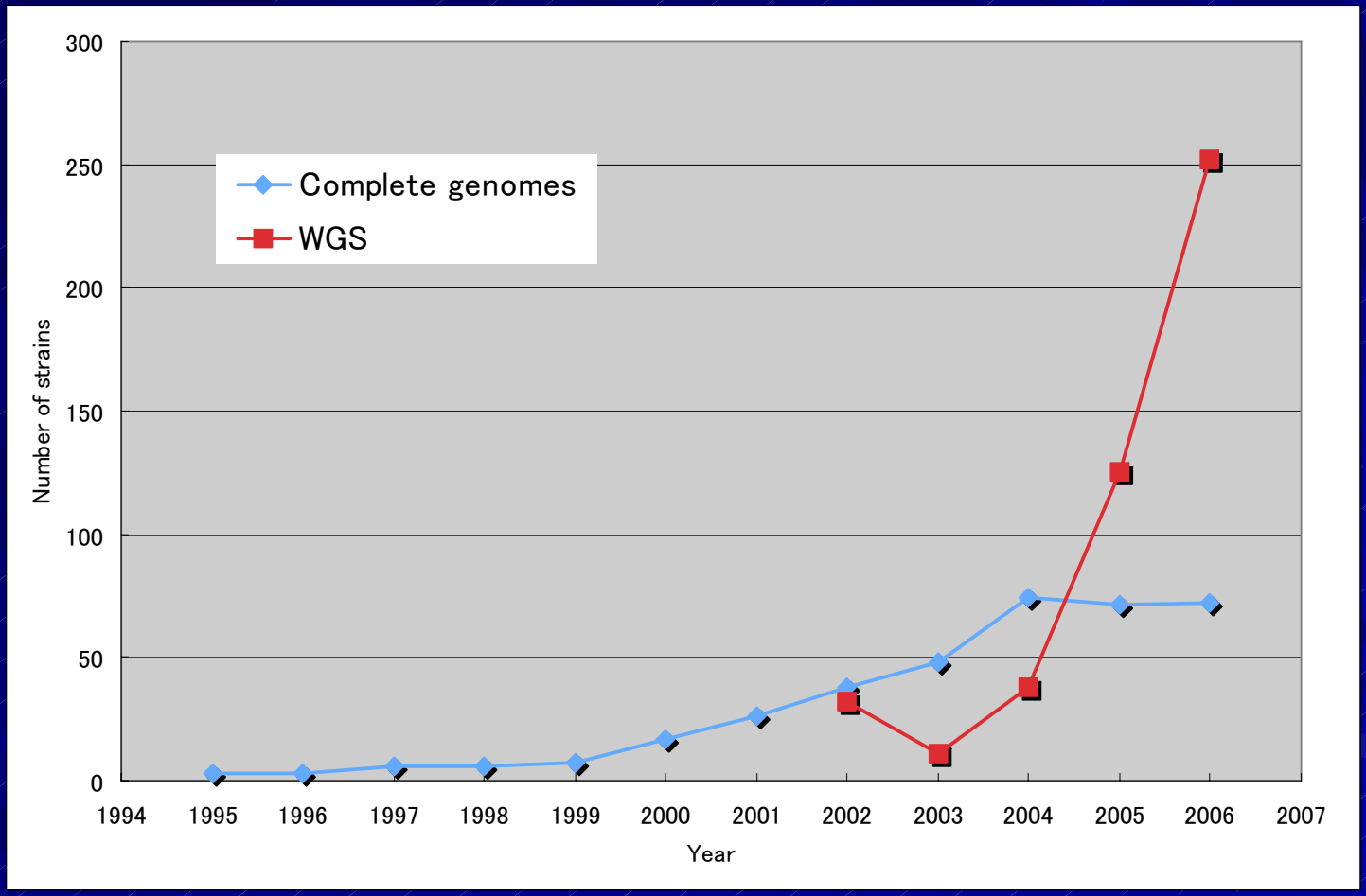
InterProScan(IP)との計算時間の計測



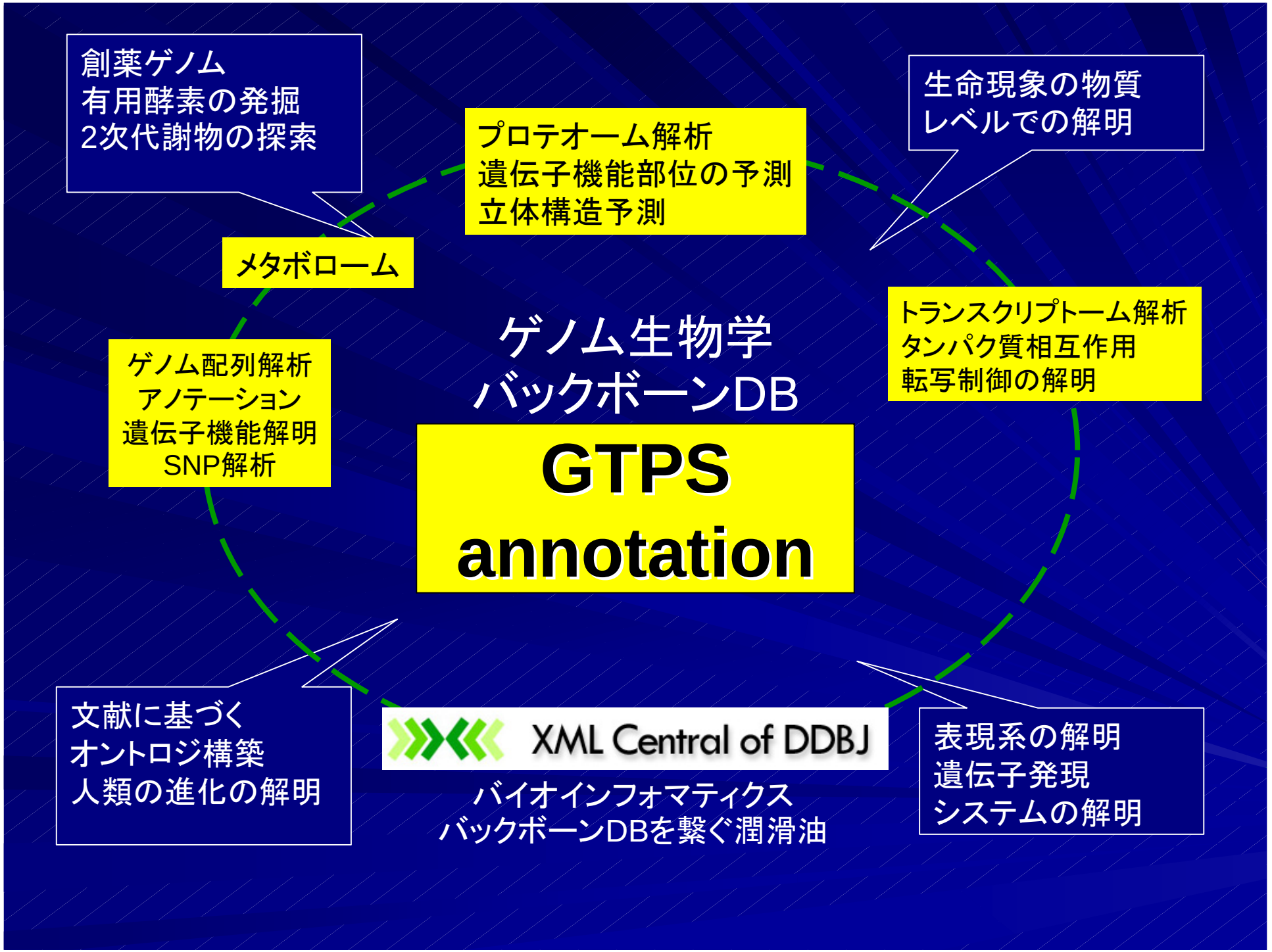
迅速な更新のためには、
高性能計算機資源が必要

NIG(1cpu)/1CDS : 5min*
RIKEN(1cpu)/1CDS : 2min
*: Xeon 3.2GHz

毎年INSDから公開される完全ゲノムの件数とWGSプロジェクト数の比較



WGS: ゲノムプロジェクト由来だが、完全に繋がっていない断片配列の集合が登録されている。


創薬ゲノム
有用酵素の発掘
2次代謝物の探索
プロテオーム解析
遺伝子機能部位の予測
立体構造予測
生命現象の物質
レベルでの解明
メタボローム
トランスクリプトーム解析
タンパク質相互作用
転写制御の解明
ゲノム配列解析
アノテーション
遺伝子機能解明
SNP解析
ゲノム生物学
バックボーンDB
GTPS
annotation
文献に基づく
オントロジ構築
人類の進化の解明
XML Central of DDBJ
バイオインフォマティクス
バックボーンDBを繋ぐ潤滑油
表現系の解明
遺伝子発現
システムの解明

GTPS
Gene Trek in Procaryote Space

§ プロトコル開発

国立遺伝学研究所	菅原秀明, 阿部貴志
東京理科大学薬学部	宮崎智
JST BIRD	田中尚人, 平畠壮規
DDBJアノテーター	小菅武英, 大城戸利久

§ データ解析・web製作

富士通株式会社	重元康昌
(株)東海ソフト開発	桑名良和

§ InterProScan 実行環境の提供

理化学研究所・情報基盤センター	姫野龍太郎, 黒川原佳

§ OASYS(アノテーションツール)

三井情報開発株式会社	菅原貴俊