# 生命科学データ共有のあり方
## --- DDBJを例に ---

国立遺伝学研究所　有田正規

Licensed under a Creative Commons 表示 4.0 国際ライセンス
(c)2018 有田正規（国立遺伝学研究所）

# 生命科学データの特徴

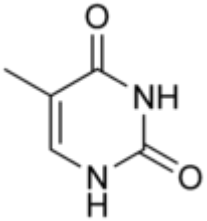
観察記録にすぎず、利用目的や適正さの判断が不可能

しかし、我々の健康・医療に直結

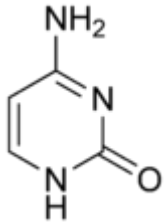
**例.　DNA塩基配列**

ATGCという4文字表現が最適？

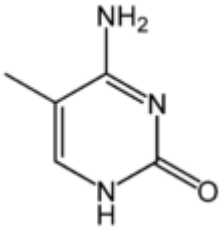
チミン




シトシン




メチルシトシン

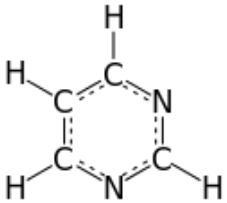



近年、DNAメチル化解析で
脚光を浴びている

**代謝物**

構造をどこまで特定？

ピリミジン



# DDBJ（INSDC） の歴史

1980　EMBL Data Library　開始

1982　EMBL & GenBank の国際連携発足、日本に参加要請

1983　DNAデータバンク運営委員会
内田久雄委員長、東大教授　→　利用者への提供係
大井龍夫　京大教授　→　文献からデータ取得係

1984　遺伝研の丸山毅夫がDDBJ受入を表明（木村資生は反対）
当時のマシン(FACOM M-150F) は　CPU 2M,　Disk 800M

1985　遺伝研に宮沢三造が着任、データ入力開始

1987　DDBJリリース1.0　（その後、丸山は急逝）

1988　ヒトゲノム国際機構設立

1991　遺伝研はヒトゲノムの研究センター設置を拒否（富澤所長）
３年間、文科省に干される

その後、論文投稿時に配列情報をバンクに投稿することが義務化される。

# データベースとフォーマットの変遷

30年前（手入力）

1993 **dbEST**

1995 **UniGene**

2000 **Gene Expression Omnibus (GEO)**

2007 **Genotype and Phenotype (dbGaP)**

2008 **Sequence Read Archive (SRA)**

2009 **human genome variants (dbVar)**

2011 **BioProject, BioSample**

2013 **Clinical Variation (clinVar)**

現在（一部RDF）

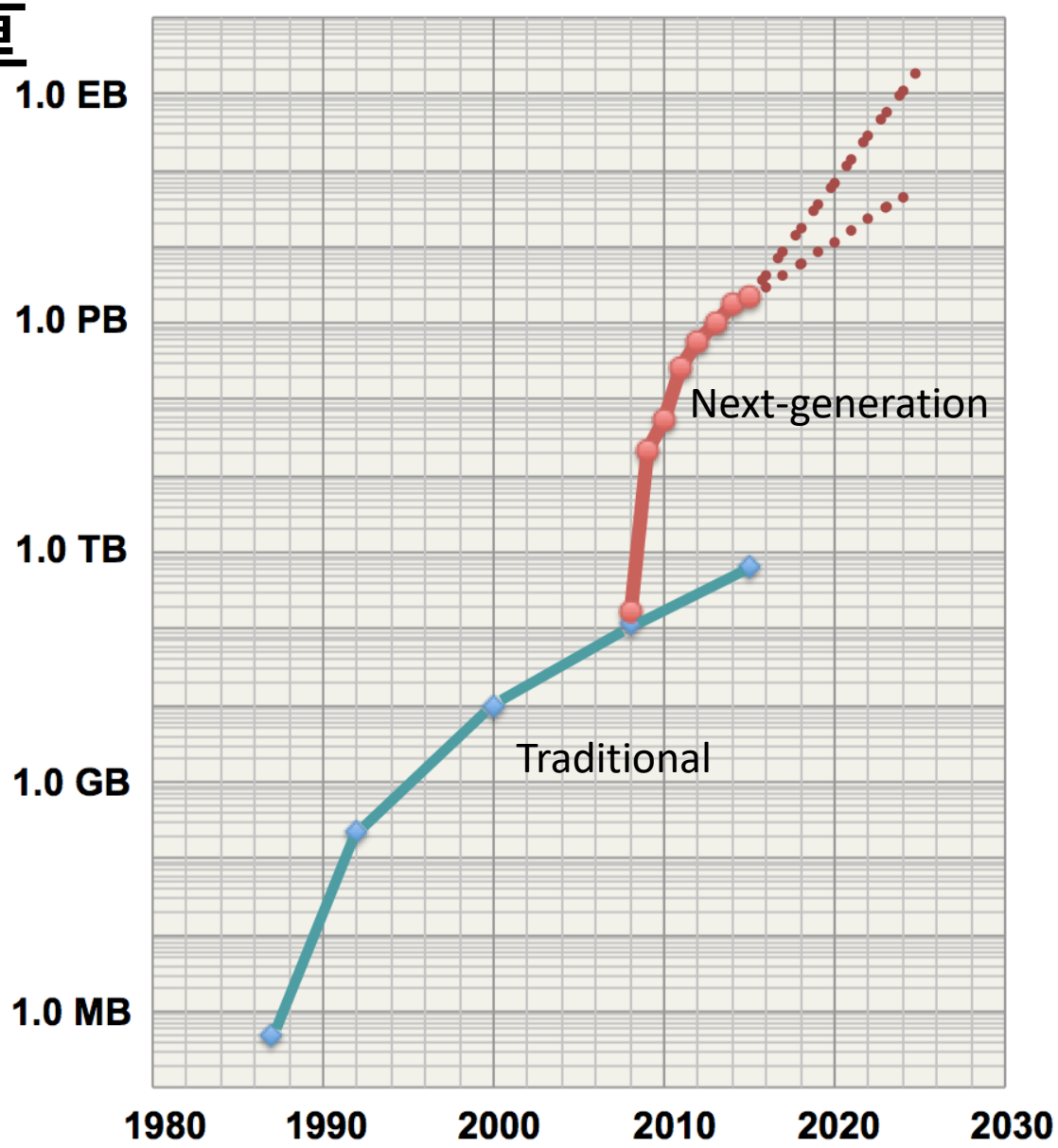
# 配列データ量

INSDCのサイズは、
データ交換できる限界
に近い。

10 Gbps = 100 TB / day

= 3 PB / month

ストレージのコストも大。
基本は磁気テープ。



# 多様化するDDBJ構成

配列の取得形態やタイプによって、様々なデータベースが必要になっている。

| | Annotated sequences | Capillary reads | NGS reads | Study | Sample | Assembly | Functional genomics | Variation | Genotype and phenotype |
|---|---|---|---|---|---|---|---|---|---|
| NCBI 米国 | GenBank | Trace Archive # SRA へ統合予定 | Sequence Read Archive | BioProject | BioSample | Assembly | GEO | #2017年9月 ヒト以外データ受付停止 dbSNP/dbVar | dbGaP |
| EBI 欧州 | European Nucleotide Archive (ENA) | | | | | | #2017年10月 取り込み停止 ArrayExpress | EVA/DGVa | EGA |
| DDBJ 日本 | DDBJ | Trace Archive | Sequence Read Archive | BioProject | BioSample | Assembly # 代理登録準備中 | DOR #2017年度末に公開 | ない | JGA |

20年前まで 1 遺伝子 1 論文だったが、今は 1微生物ゲノムでも論文にならない。
( Smith DR “**Goodbye genome paper, hello genome report**”
*Brief Funct Genomics* 16(3), 156, 2017)

# オープン化のコスト

DDBJ の年間予算は 10 億円弱 （DDBJ報告書より）

スパコン 6.5億、運用支援・開発 2億、キュレータ 1億、その他

これで 4 PB のINSDCデータと、全国120機関800ユーザを支援

今はざっくり 1 PB の維持管理に 1 億円

| NCBI (2015) | EBI (2015) | DDBJ (2015) |
|---|---|---|
| 287 人 | 513人 | 39人 |
| 181.3 億円 | 77.1 億円 | 14.6億円 |
| 30 PB? | 20 PB? | 当時 12 PB |

NBDC の 2016 統合推進事業 事業報告概要より

# コストをどう負担する？

1. **受益者負担、寄付など**
   プレプリントサーバ arXiv のコストは 1.5 億円　(140万論文)
   （コーネル大は17万ドル支出、上位機関$4500。
   　200超の利用上位機関に請求して、50万ドルが目標。）
   Wikipedia のコストは 60 億円ぐらい
   （財団のAnnual Plan Questions and Answersより。寄付キャンペーン）

⇒　１０億円規模だと、上位機関に最低１００万円は請求

2. **国民負担**　（＝運営費交付金）
   現在のINSDCモデル
   OAモデルもこちら

⇒　国民に科学・健康情報の運営費を認識して貰う必要

# OA出版とは国民負担

学術出版は税金ビジネス　　購読料300億円

“Springer Nature revenues grew 2.5 percent to €1.64 billion last year. Adjusted group profits fell slightly to €374 million, with operating profit margins of 22.9 percent.” (Handelsblatt, Apr 13, 2018)

*Nature Communications* の出版費用70万円超（税込）は、15万が投資家へ。

受益者負担の購読モデル→　まだ余裕のある国民負担へ

OAは学術出版に多大な利益をもたらしている

でも、日本はここで考えるべき

# 国民負担にするなら、日本語で

Human Genome Organization “human genome forms part of the common heritage of mankind.”

Francis Collins “The human genome is our shared inheritance.”

こういっても、普通の日本人には無意味

生命科学を身近に感じて貰う必要

映画
「Lorenzo’s Oil」

**Adrenoleukodystrophy, ALDを自力で予防する実話**

# 結論：我々がすべきこと

（個人の見解で、DDBJセンターの決定ではありません）

- OA唱道者は、長期的コストに目を向けよ
  独りよがりな研究者の意見は無意味
  基礎研究の場合、受益者負担が原則

- 生命科学データを医療・社会福祉に直結させよ
  日本語による発信を重視
  国際化の前に、国内貢献こそ重要

有田正規「科学はどこまでオープンにできるか」情報管理2015