

DDBJ Centerと NIG SuperComputer

DDBJセンター
国立遺伝学研究所
有田 正規

DDBJ (DNA Data Bank of Japan) センターの主要業務

- 国際塩基配列データベースの構築
- スーパーコンピュータ（スパコン）運用
- 解析ソフトの開発と提供
- 教育と広報

スタッフ構成： 教員 7、キュレータ 3 + 1 1
システム管理 2 3、 秘書 2

INSDC (International Nucleotide Sequence Database Collaboration)

塩基配列データベース事業



サイズは 4 PB

世界中のDNA配列情報を

- 査定して受け入れ
- データベースに蓄積し
- 無償公開して共有する

三極連携 (毎日交換)

- 米国 : GenBank (NCBI)
- 欧州 : ENA (EBI)
- 日本 : DDBJ (遺伝研)

生命科学論文では、データの公開が必須

DNA 配列やタンパク質立体構造は指定のリポジトリに登録。

DDBJはマイクロアレイ、NGS配列のリポジトリも提供。

Mandates for specific datasets *Nature*誌の投稿規定

For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank

データベースの構成

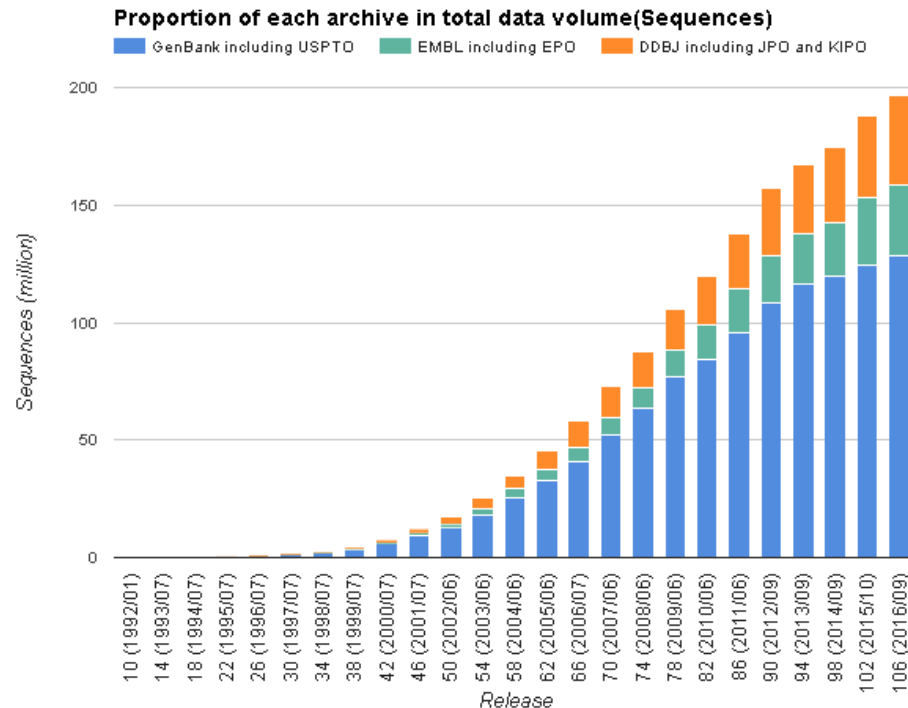
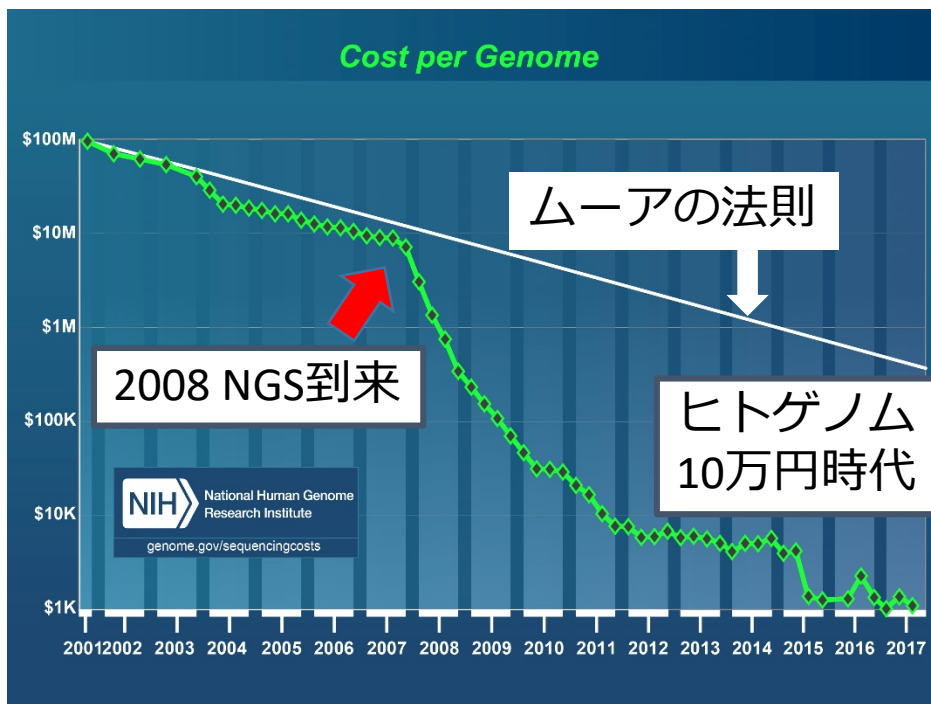
- **INSDC** (www.insdc.org)

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

- **GEA** (Genomic Expression Archive)
欧州のArrayExpress, 米国のGEO に相当
- **JGA** (Japanese Genotype-Phenotype Archive)
欧州のEGA, 米国の dbGaP に相当。制限公開

配列コスト減とデータ量の増加

大規模化と、制限共有化が急激に進んでいる



生物学は情報学になる

大規模化

- 1000 human genome
- 1001 *A. thaliana*
- 1011 *S. cerevisiae*
- 10K Bacteria
- 10K Vertebrates



ARTICLE

OPEN
<https://doi.org/10.1038/s41586-018-0030-5>

Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates

Jackson Peter^{1,6}, Matteo De Chiara^{2,4}, Anne Friedrich¹, Jia-Xing Yue², David Pflieger¹, Anders Bergström², Anastasie Sigwalt¹, Benjamin Barre², Kelle Freel¹, Agnès Llored², Corinne Cruaud³, Karine Labadie³, Jean-Marc Aury³, Benjamin Istace³, Kevin Lebrigand⁴, Pascal Barbry⁴, Stefan Engelen³, Arnaud Lemainque³, Patrick Wincker^{3,5,7}, Gianni Liti^{2,7*} & Joseph Schacherer^{1,7*}

ゲノムコホート

- UK BioBank (500K), UK10K (英国)
- 東北メガバンク三世代コホート (7万人)

ダウンロードすら難しい状況へ ← 計算リソース必要

遺伝研スパコン <http://sc.ddbj.nig.ac.jp/>

全国130機関、600超のユーザーが利用

ニュース

- [予告]【Phase2システム】 UGE month_ssd.q廃止のお知らせ
2018年08月10日
- (終了) 8月13日(月)10:00～12:00【LUSTRE4障害】 LUSTRE4緊急メンテナンスのお知らせ
2018年08月10日
- 2018年5月16日 Phase1 ディスク使用量削減のお願い
2018年05月16日
- (終了) 2018年4月30日 3:00～5月1日 10:45 Phase1システム LUSTRE1

遺伝研スーパーコンピュータシステム

国立遺伝学研究所は遺伝学を中心とした生命・医学系研究における情報処理のための大規模計算基盤として、大規模クラスター型計算機・大規模メモリ共有型計算機・大容量高速ディスク装置を備えた最新鋭のスーパーコンピュータシステムサービスを提供しています。



スパコン構成と利用法

30 TB までは無料。利用申請はウェブサイトから

ハードウェア構成 (年度末に入れ替え)

2018年度中

高速HDD (Lustre) 7 PB

554 x 64 GB memory, 10 x 2 TB, 1 x 10 TB

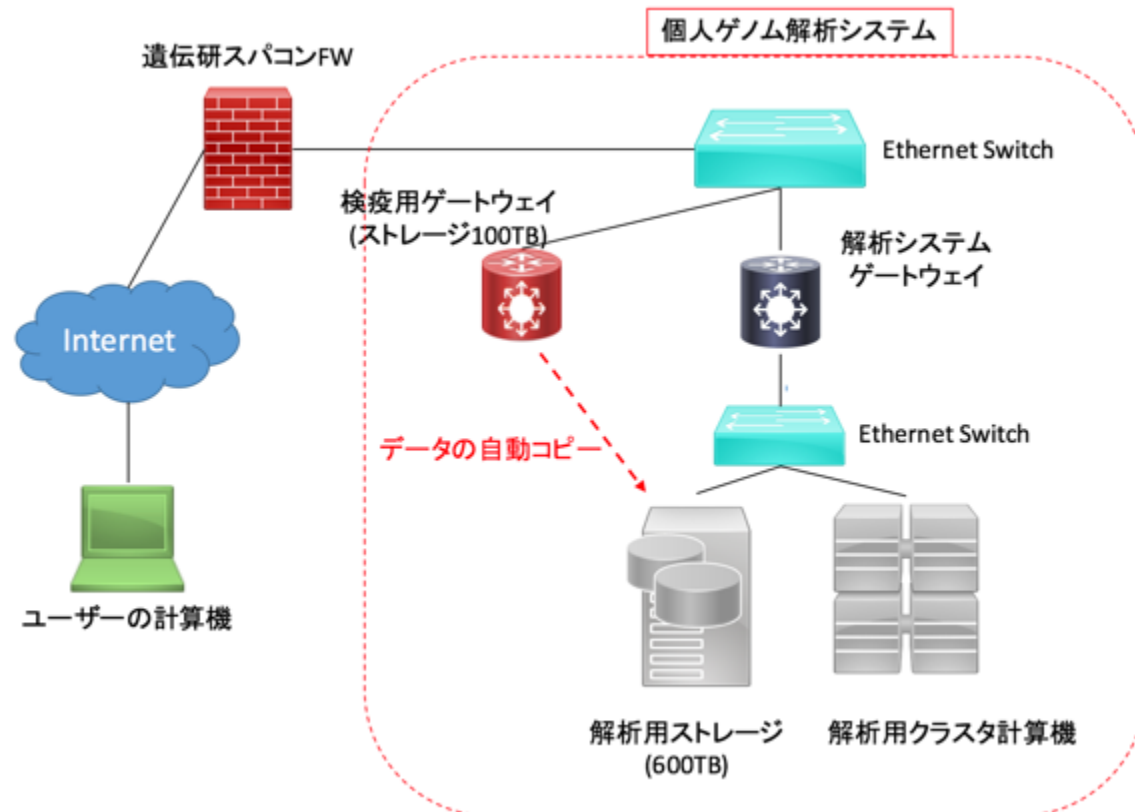
2019から(予定)

高速HDD (Lustre) 13 PB

200程度 x 512 GB memory, 10 x 4 TB, 1 x 12 TB

個人ゲノム解析システム

制限共有されているヒトゲノムデータを、NIG
スパコン上で解析できるサービス（有償）



データのアップ・ダウンロードは、ウィルスチェック用のゲートウェイを利用

利用してください

DDBJ は大学共同利用機関法人のため、皆さんの利用実績が必要です。

現在のサービスを維持するためにも、DDBJ およびスパコンの利用をお願いします。