国際DNA塩基配列データベース構築および遺伝研スパコンにおけるデータ解析の現状

国立遺伝学研究所 DDBJセンター
小笠原　理

国立遺伝学研究所

HOME > 遺伝研について

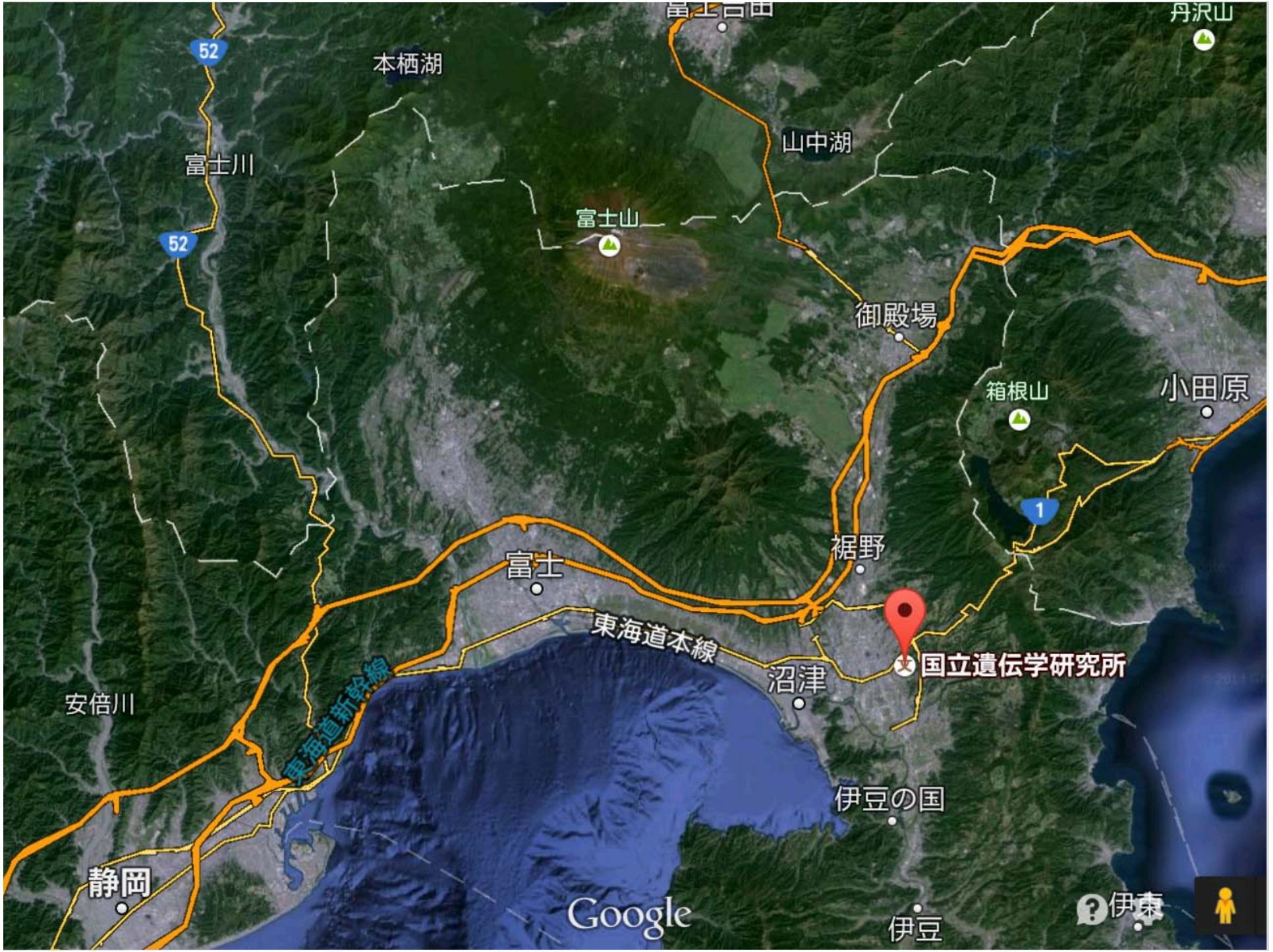
遺伝研について



大学共同利用機関法人　情報・システム研究機構

国立遺伝学研究所

住所：〒411-8540　静岡県三島市谷田1111
TEL番号：055-981-6707（管理部）
FAX番号：055-981-6715
最寄駅：JR三島駅
三島駅～国立遺伝学研究所
距離 約4km，バスまたはタクシーで約20分
アクセス方法や地図はこちら

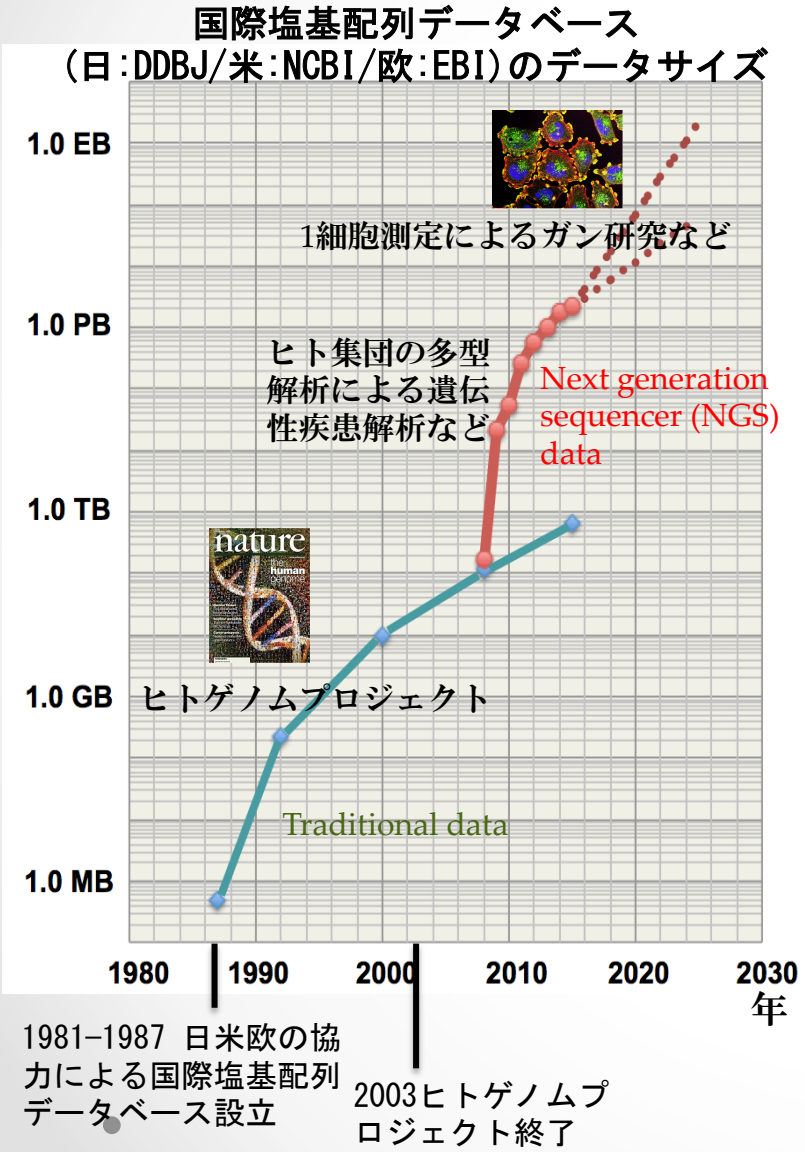

丹沢山
52
本栖湖
山中湖
富士川
富士山
52
御殿場
箱根山
小田原
1
裾野
富士
東海道本線
国立遺伝学研究所
沼津
安倍川
東海道新幹線
伊豆の国
静岡
Google
伊豆
伊東

遺伝研スパコン：目的

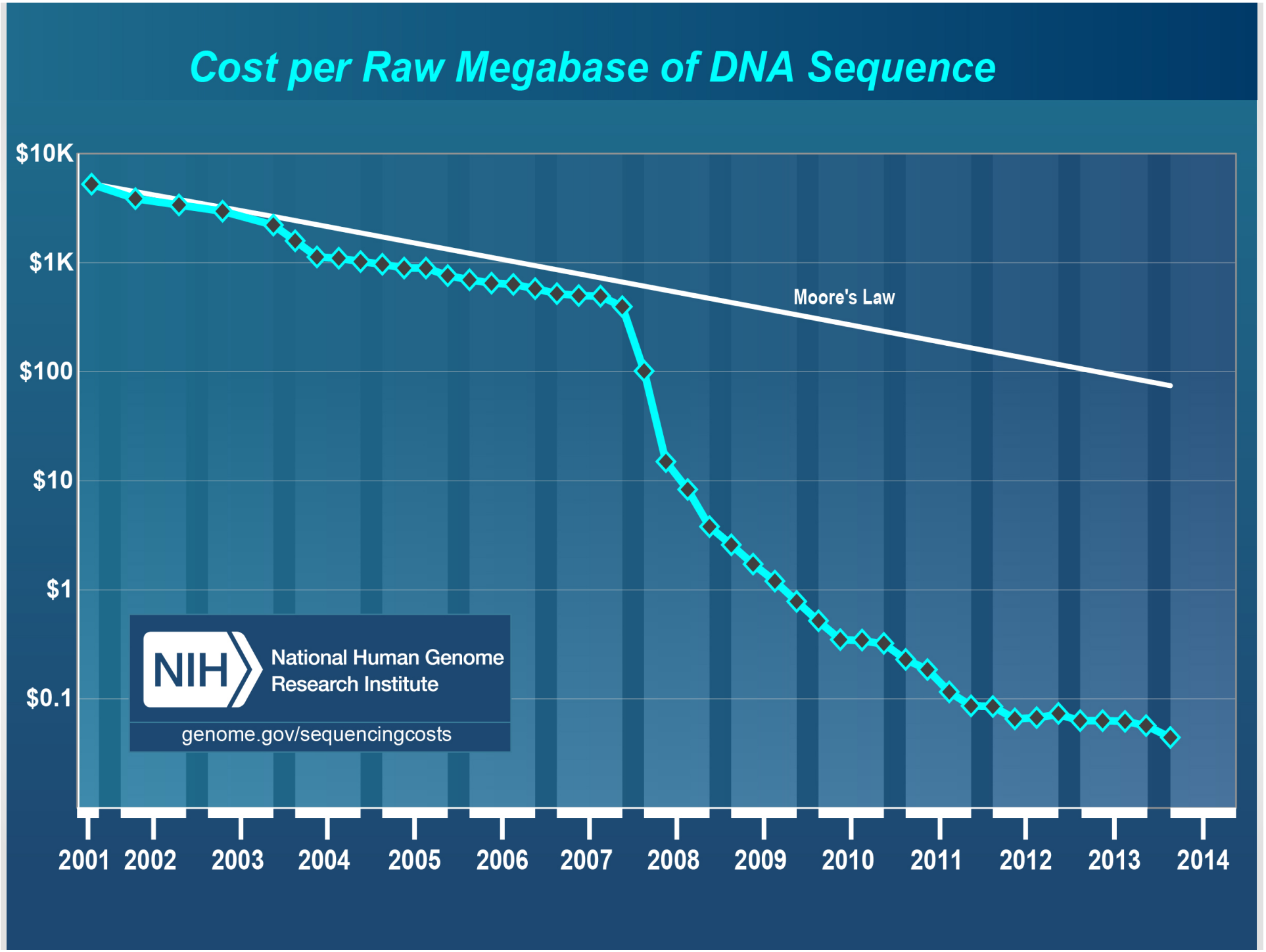
1. **国際塩基配列データベース** (International Nucleotide Sequence Database : INSD)**の構築**
2. **大学共同利用機関として、主に生物学医学系研究者への計算機資源の提供**



国際塩基配列データベースINSDCとDDBJの沿革



- 1977年 DNA塩基配列決定法 (Sanger法)
- 1981年 EMBL(欧) 1982年 GenBank（米）1984年 DDBJ 設立
- 1990 年頃 論文発表時にデータを登録することを著者に義務づける現在の運用形態が確立。
- 1993年 日米欧の特許庁との協力体制。6極会議。
- 2003年 ヒトゲノムプロジェクト終了
- **2008年 次世代シーケンサーのアーカイブを日米欧３極で開始**
- **2013年 個人ゲノム情報のアーカイブ(JGA/EGA/dbGaP)の運用を日米欧3極で開始。DDBJはJST傘下のNBDC と協同でJGAを運用。**


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

遺伝研スパコン

1. **分散メモリ型計算ノード** (Thin)
   - 64GB memory/node
   - 554**ノード**
   - GPGPU**搭載**128**ノード**
   - XeonPhi**搭載**32**ノード**
2. **共有メモリ型計算ノード** (Medium, Fat)
   - 2TB memory/node 10**ノード**
   - 10TB memory/node 1**ノード**
4. Lustre File system 7PB
5. **アーカイブ用ストレージ** 5.5PB
6. Infiniband QDR, FDR

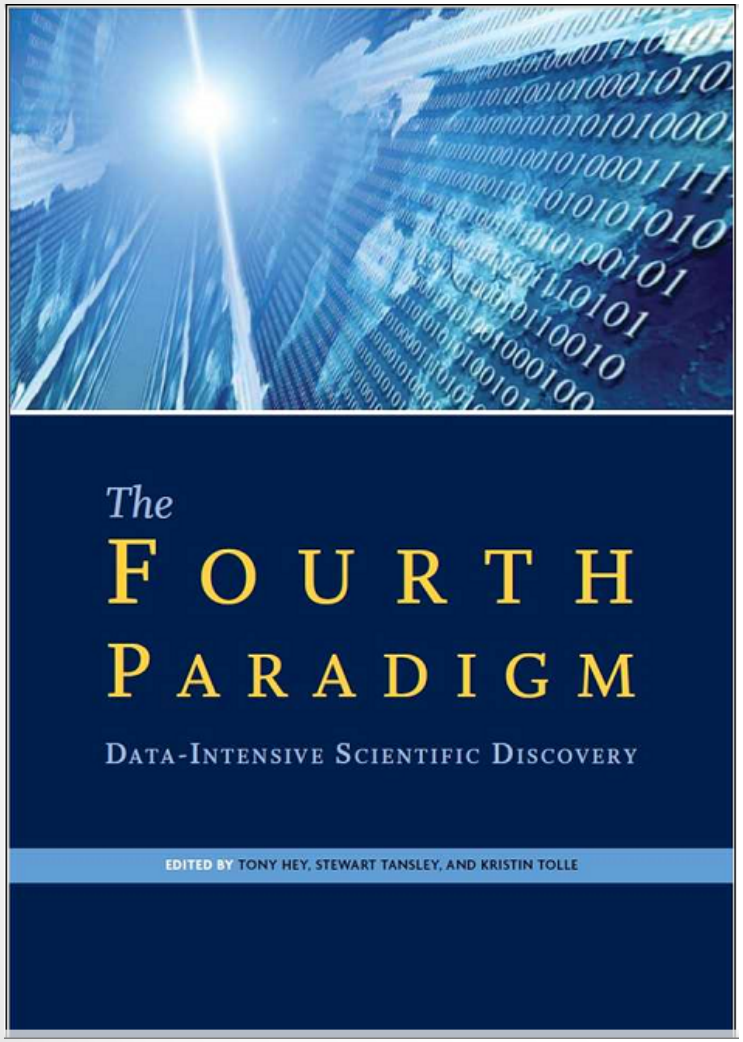


遺伝研スパコンは1995**年から。**
現スパコンは2012**年導入、**2014**年中間増強、**2017**年次期スパコンへリプレース**

国際塩基配列データベースコンソーシアム
INSDC = DDBJ（日）＋ NCBI(米）＋ EBI (欧)



毎日日米欧の拠点の間でデータを交換



"The fourth paradigm : Data-intensive scientific discovery" Microsoft (2009)

Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data

CHRISTOPHER Southan

Graham Cameron

EMBL-European Bioinformatics Institute

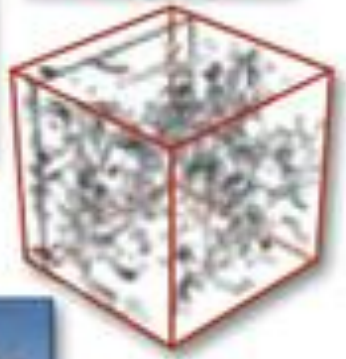
Custodianship of the data is held by the International Nucleotide Sequence Database Collaboration (INSDC), which consists of the DNA Data Bank of Japan (DDBJ), GenBank in the U.S., and EMBL-Bank in the UK. These three repositories exchange new data on a daily basis. As of May 2009, the totals stood at approximately 250 billion bases in 160 million entries.

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$





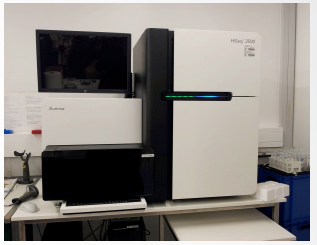


"The fourth paradigm : Data-intensive scientific discovery" Microsoft (2009)

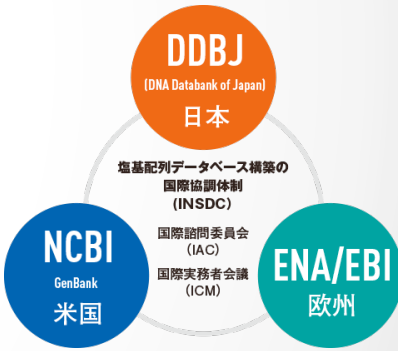
Components of data intensive science ("the fourth paradigm")

- Capture
  - The vast volume of data are captured by instruments.
- Curation
  - The data is stored in huge storages for long-term preservation.
  - The data is curated. (the data is made understandable and usable for the other researchers.)
  - The data is made publicly accessible.
- Analysis
-
  - The data is analyzed by many scientists.

NGS



INSDC



Supercomputer



INSDCにおける"Curation"

1. 研究者がDNA塩基配列データおよびメタデータを登録
   - メタデータ: (1)登録者情報 (2)生体サンプル情報
   - (1)シーケンサーの出力生データ (2)計算の中間データ (3)アノテーション付きの塩基配列データ
   - INSDCはこれらのデータのCurationを行いデータベースの品質を保持

2. 論文の投稿、査読
   - この時INSDCから発行されたaccession numberが必要

3. 論文公開。同時に登録されたDNA塩基配列データおよびメタデータも公開される。(キーワード検索・相同性検索・ダウンロード)

4. 公開後のデータ更新。データの差し替えへの対応

GenBank/DDBJ Flat File Format (1/3)

```
LOCUS       DSU64875                5805 bp    DNA     linear   INV 24-AUG-1996
DEFINITION  Drosophila simulans white gene, complete cds.
ACCESSION   U64875
VERSION     U64875.1
KEYWORDS    .
SOURCE      Drosophila simulans
  ORGANISM  Drosophila simulans
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1  (bases 1 to 5805)
  AUTHORS   Kirby,D.A. and Stephan,W.
  TITLE     Haplotype test reveals departure from neutrality in
            the white gene of Drosophila melanogaster
  JOURNAL   Genetics 141 (4), 1483-1490 (1995)
   PUBMED   8601488
REFERENCE   2  (bases 1 to 5805)
  AUTHORS   Kirby,D.A. and Stephan,W.
  TITLE     Multi-locus selection and the structure of variatior
            gene of Drosophila melanogaster
  JOURNAL   Genetics 144 (2), 635-645 (1996)
   PUBMED   8889526
REFERENCE   3  (bases 1 to 5805)
  AUTHORS   Kirby,D.A. and Stephan,W.
  TITLE     Direct Submission
  JOURNAL   Submitted (23-JUL-1996) Zoology, University of Maryl
            Park, MD 20742, USA
```



GenBank/DDBJ Flat File Format (2/3)

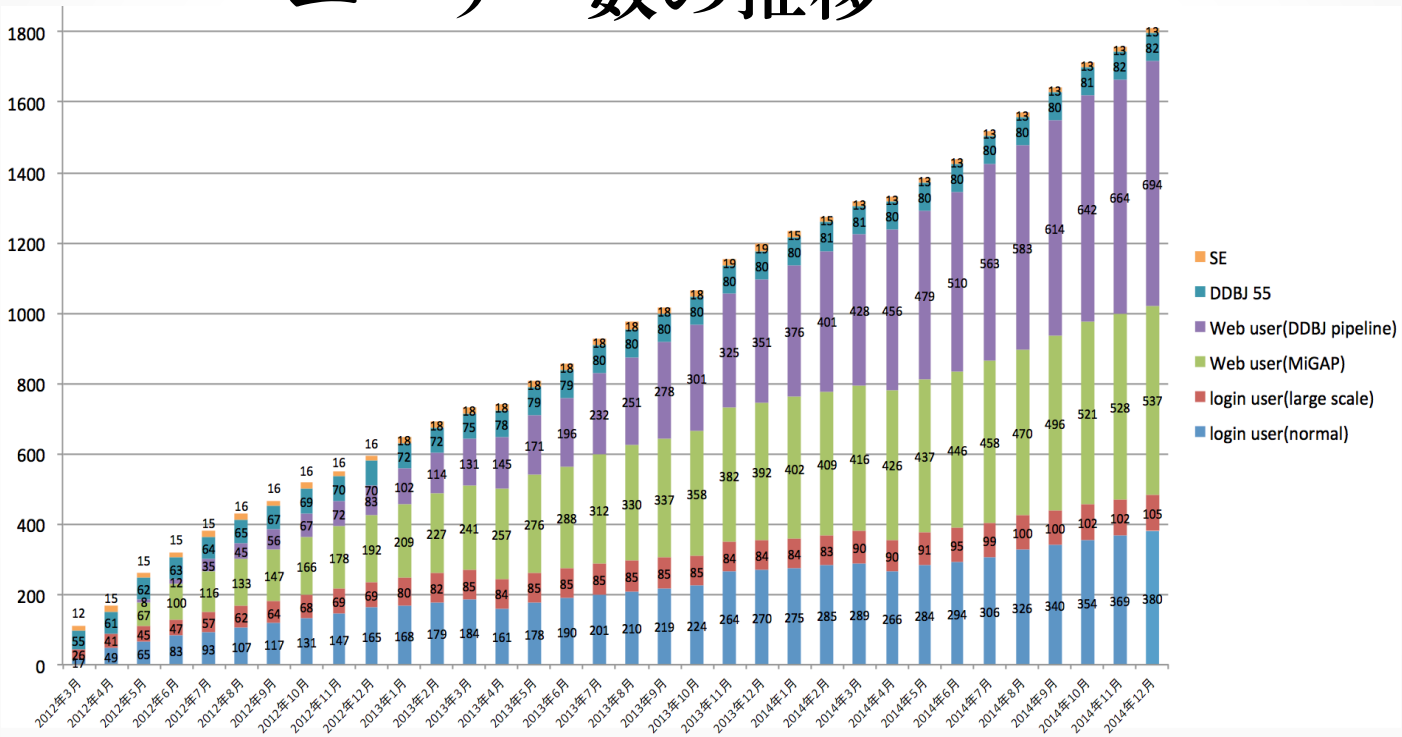
```
FEATURES             Location/Qualifiers
     source          1..5805
                     /organism="Drosophila simulans"
                     /mol_type="genomic DNA"
                     /strain="sim"
                     /db_xref="taxon:7240"
     gene            1..5805
                     /gene="white"
     mRNA            join(1..299,3287..3560,3631..4285,4342..4657,4860..4991,
                     5056..5805)
                     /gene="white"
     exon            1..299
                     /gene="white"
                     /number=1
     5'UTR           1..227
                     /gene="white"
     CDS             join(228..299,3287..3560,3631..4285,4342..4657,4860..4991,
                     5056..5670)
                     /gene="white"
                     /codon_start=1
                     /protein_id="AAB06578.1"
                     /db_xref="GI:1507738"
                     /translation="MGQEDQELLIRGGSKHPSAEHLNNGDSGAASQSCINQGFGQAKN
                     YGTLRPPSPPEDSGSGSGQVAENLTYAWHNMDIFGAVNQPGSGWRQLVNRTRGLFCNE
                     RHIPAPRKHLLKNVCGVAYPGELLAVMGSSGAGKTTLLNALAFRSPQGIQVSPSGMRL
                     LNGQPVDAKEMQARCAYVQQDDLFIGSLTAREHLIFQAMMRMPRHLTYRQRVARVDQV
                     IQELLLSKCQHTIIGVPGRAKGLSGGERKRLAFASEALTDPPLLICDEPTSGLDSFTA
                     HSVVQVLKKLSQKGKTVILTIHQPSSELFELFDKILLMAEGRVAFLGTPSEAVDFFSY
```

GenBank/DDBJ Flat File Format (3/3)

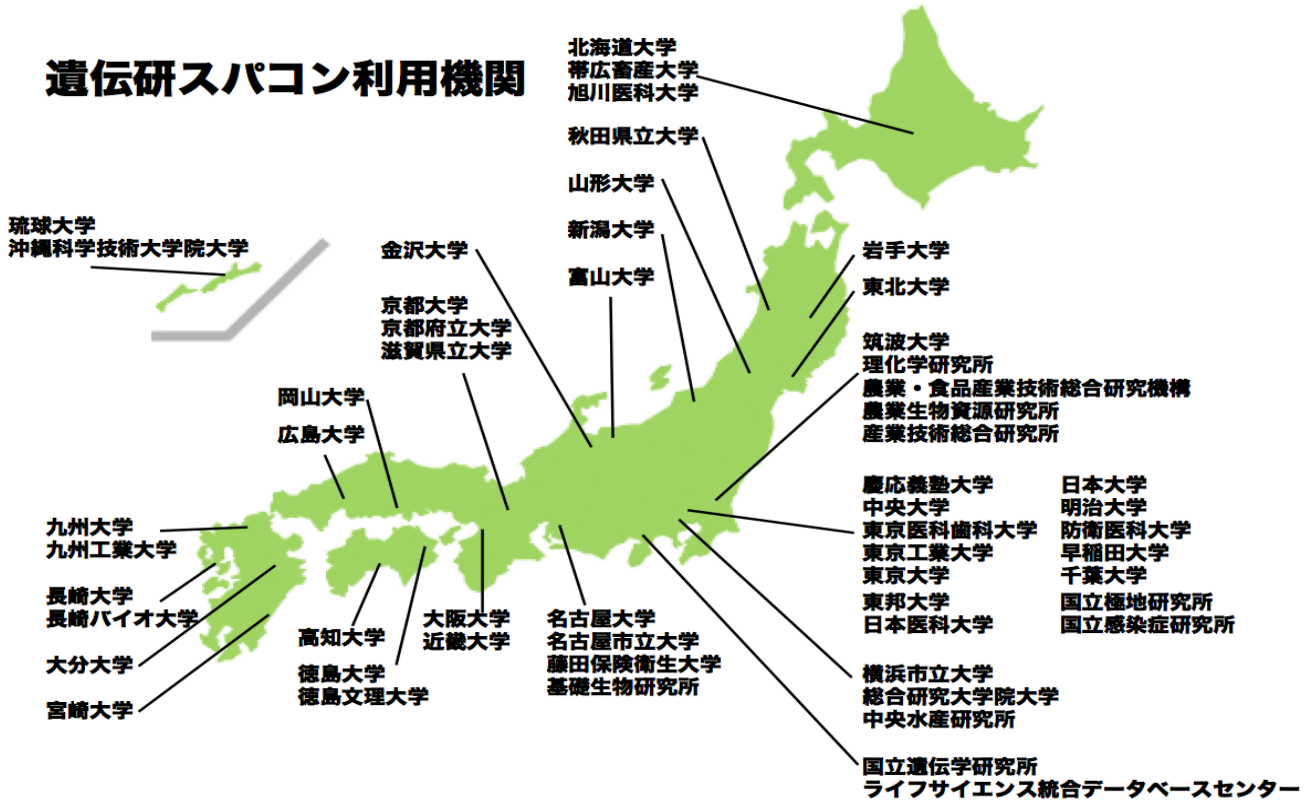
```
BASE COUNT         1489 a         1376 c         1362 g         1578 t
ORIGIN
        1 tgtctctttc gccaccgttt gtagcgttac ctagcgtcaa tgtccgcctt cagttgcatt
       61 ttgtcagcgg tttggtgacg aagctccaag cggtttacgc catcaattaa acacaaagtg
      121 ctgtgccaaa actcctctcg cttcttattt ttgtttgttt tttgagtgat tggggtggtg
      181 attggttttg ggtgggtaag caggggggaag tgtgaaaaat cccggcaatg ggccaagagg
      241 atcaggagct attaattcgc ggaggcagta aacacccatc tgccgagcat ctgaacaatg
      301 tgagtagtac ttgtgcatac gtcttaagat cagttgatct ataggaacag cgattgcaac
      361 atccaatggt ctgcggcgtg agaactgcga cccacaaaaa tcccaaaccg caatcgaaca
      421 aacaaatagt gacgcgaaac agattattct ggtagcttaa catgttataa gagacaacta
      481 ttaagatcat atcatgatca agacatcgaa aggcattcat tttctctaag aatatattct
      541 tttttataaa tataacgagc agatatttat ttggaagtag tagctcgtaa ttagggacgc
      601 ataaccagag ggtgtgtggg ggaggaggag tcggtttggt ttggtttatc ggcttaagtt
      661 ggccaacaac attgttgggt gtctataact cgggggttgc cagatactag acactagata
      721 ctgtatccgt atccatttct ggttgtgtac ttgcatctcc tacatgatct taatacctcg
      781 ctgtttgctc gtctcgctcg acgaaaaatg tacaatcttg tcttatctgg gtcattattt
      841 ggctagacga atgcttcagg ctcagcatct gatatcttgg catctgcgtg cgtatcttgc
      901 ttcaaattct tagcacctcg gcttgtataa caaaataaat aagtgagtac gatatgcata
      961 tctagcccgg gtctttgaaa caattttgaa aagtctcaaa aagttatacg agagataaga
     1021 actttaattc tttggggaaa gactaaatat gatatgtatg gaatatttta aaccatcttt
     1081 aattatatgt accttggact gtttacgctc tgatatttcc ctaataatat cgagtgatcg
     1141 tgagctttc cgctgaataa ttgcgcctcc ttcagtattt cgttctcatc aaagtgaatg
     1201 acttttgagt gcactcatta gccactcgca caatcgacaa aatgggaggg agaacaaatt
     1261 gagtcgcttc tttattcagt catttcaatt tgtcaagttc atcaaaaaaa aaaaaaagcg
     1321 atactccttc gcctgcctgc gcagcgccat ctatcgccca ctccgggagt ttccttcgca
     1381 tttcgtaagt aattcaaaaa tagcacgcaa agtaaaaagt aagccaaaaa attaaataaa
     1441 tatgaaatgt agcatcttac gctacttatt tcagcgaaat acgaaaatag ggaaaaggtg
     1501 ggagagtgaa ccagggacag ggagaagata agaaaaccca gtgggaaagg gacggcgaaa
     1561 acagaaagaa aagaaaactt tttaataaca tttttcggcc tcatttgttg tgaaaacaat
     1621 ttcgccaggt ccctctctct ctatctctgt ctctttcccc gccatatggc catctcgctc
```

大学共同利用機関としての計算機資源の提供

ユーザー数の推移



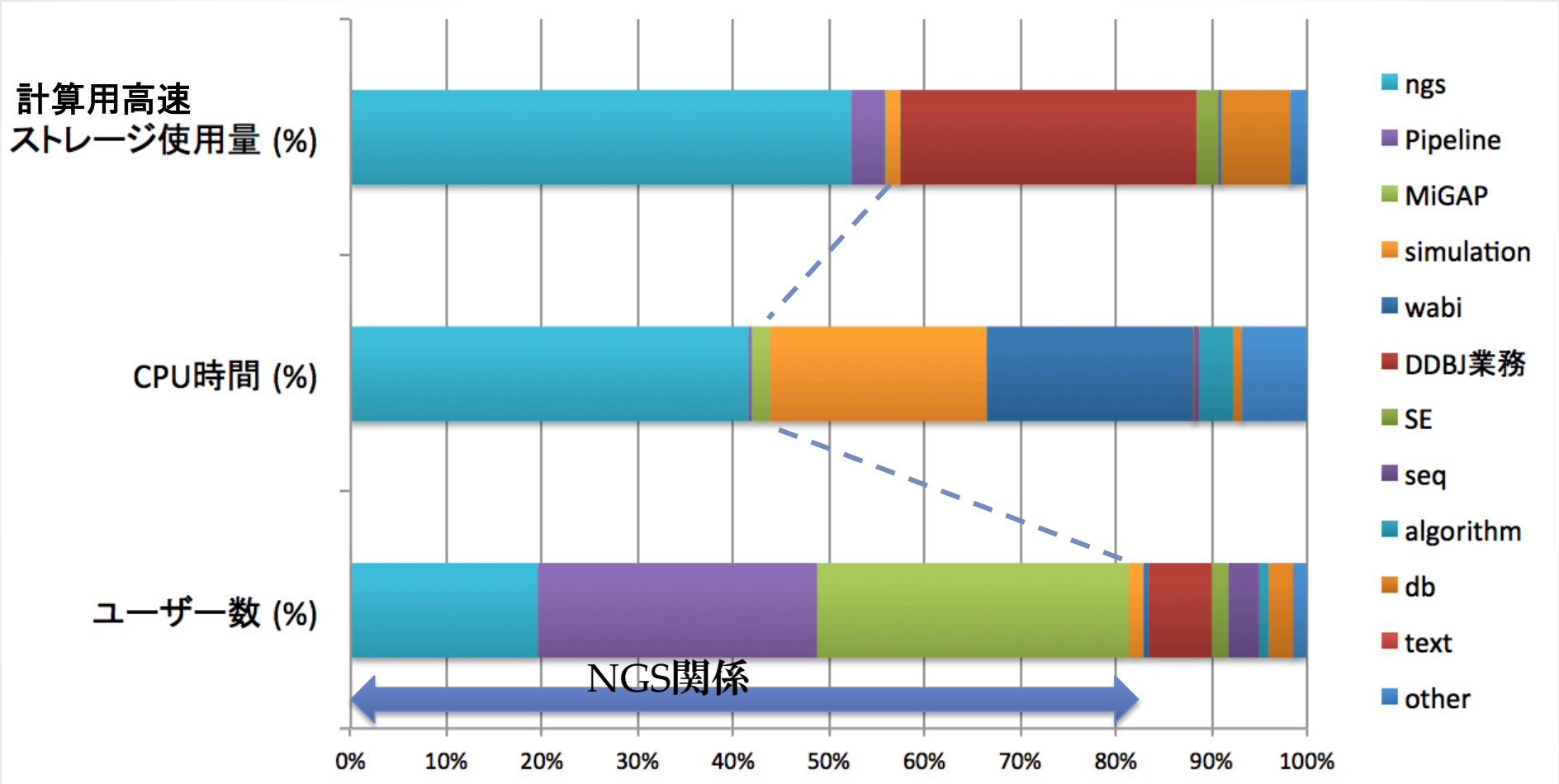
ユーザ分類	外部ユーザ数	遺伝研所属ユーザ数	総ユーザ数
一般研究ユーザ	322	36	358
一般研究ユーザ - 大規模 -	77	31	108
Webサービスユーザ	559	12	571
DDBJ pipelineユーザ	743	35	778
業務ユーザ	0	80	80
スパコン管理者	0	14	14
全ユーザ合計数	**1701**	**208**	**1909**



スパコン利用機関数

ユーザ区分	機関数	国内		海外	
		大学・研究所	企業	大学・研究所	企業
ログインユーザ	128	116	9	3	0
MiGAPユーザ	199	127	14	47	11
Pipelineユーザ	367	159	23	176	9
総数	520	247	39	215	19

利用内訳


計算用高速
ストレージ使用量 (%)
CPU時間 (%)
ユーザー数 (%)
NGS関係
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
ngs
Pipeline
MiGAP
simulation
wabi
DDBJ業務
SE
seq
algorithm
db
text
other

遺伝研スパコンの現状

- NGSデータに対応するため現スパコンではストレージサイズ、計算性能とも10～20倍の性能向上。
- DRA/ERA/SRAのアーカイブサイズの見通し
  - 増加率1.4倍/年 => 2017年　5PB, 2022年　25PB
  - 増加率2倍/年　=> 2017年 10PB, 2022年 300PB
- 大学共同利用機関としての計算機資源の提供
  - 月平均50人増加：現在2000人以上
  - 利用機関数：国内約300機関、全体で520機関
  - ユーザーの約8割はNGS解析関係
  - いつも2000ジョブ程度が実行されており、2000ジョブ程度が待ち状態にある。

ユーザーからの要望とその対応

1. スパコン用ソフトウェア開発研究の促進
2. データ解析の簡便化
3. ヒト遺伝性疾患やがんゲノムなど、個人ゲノム解析環境の提供
4. 各計算機センター間の連携
5. 種々のデータベースのミラーリング

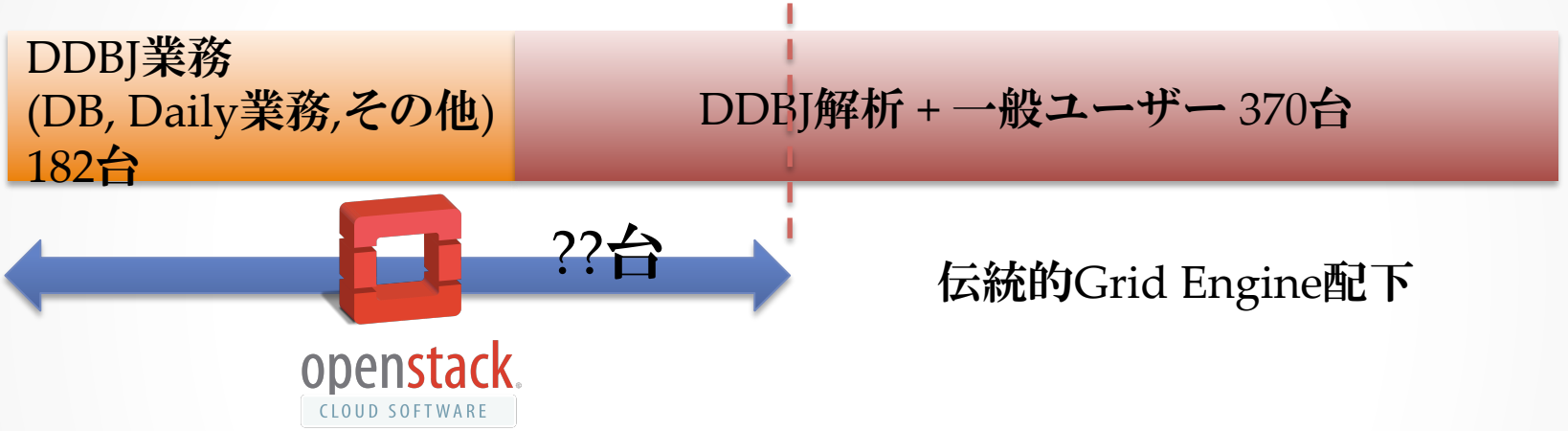


これらの要望に応えるために以下のような取り組みをしている。

(1) 遺伝研スパコンの仮想環境、分散計算環境への対応 (OpenStack, Hadoop,Spark,...)
(2) スパコン間連携
(3) 個人ゲノム対応(高度なセキュリティー)

仮想環境等への対応

現在は、グリッドエンジンを用いてDDBJと一般ユーザーで計算機をなるべく共用する（赤の範囲370台)ことで計算機の無駄な空き時間がなくなるよう工夫している。



- 個人ゲノム解析
- Webサーバー、DBサーバー
- 各研究者に最適化された環境

- 大規模処理、高速処理

OpenStackを用いた仮想環境の導入により、物理結線の変更なしに、設定の変更だけで台数を調整できる => 必要な計算機の台数を最適化

計算機センター間の連携システム
ソフトー式のインストールと実行制御

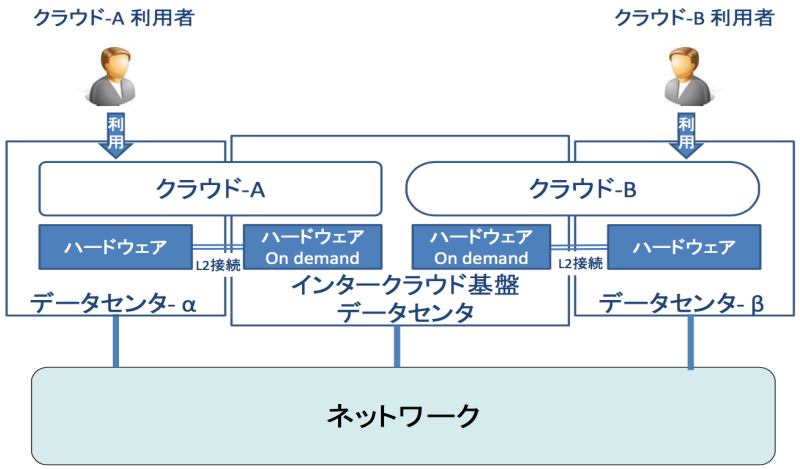
1. 関連技術の標準化、共通APIの開発

2. オンデマンドで環境を構築

3. Docker (Linux Container)



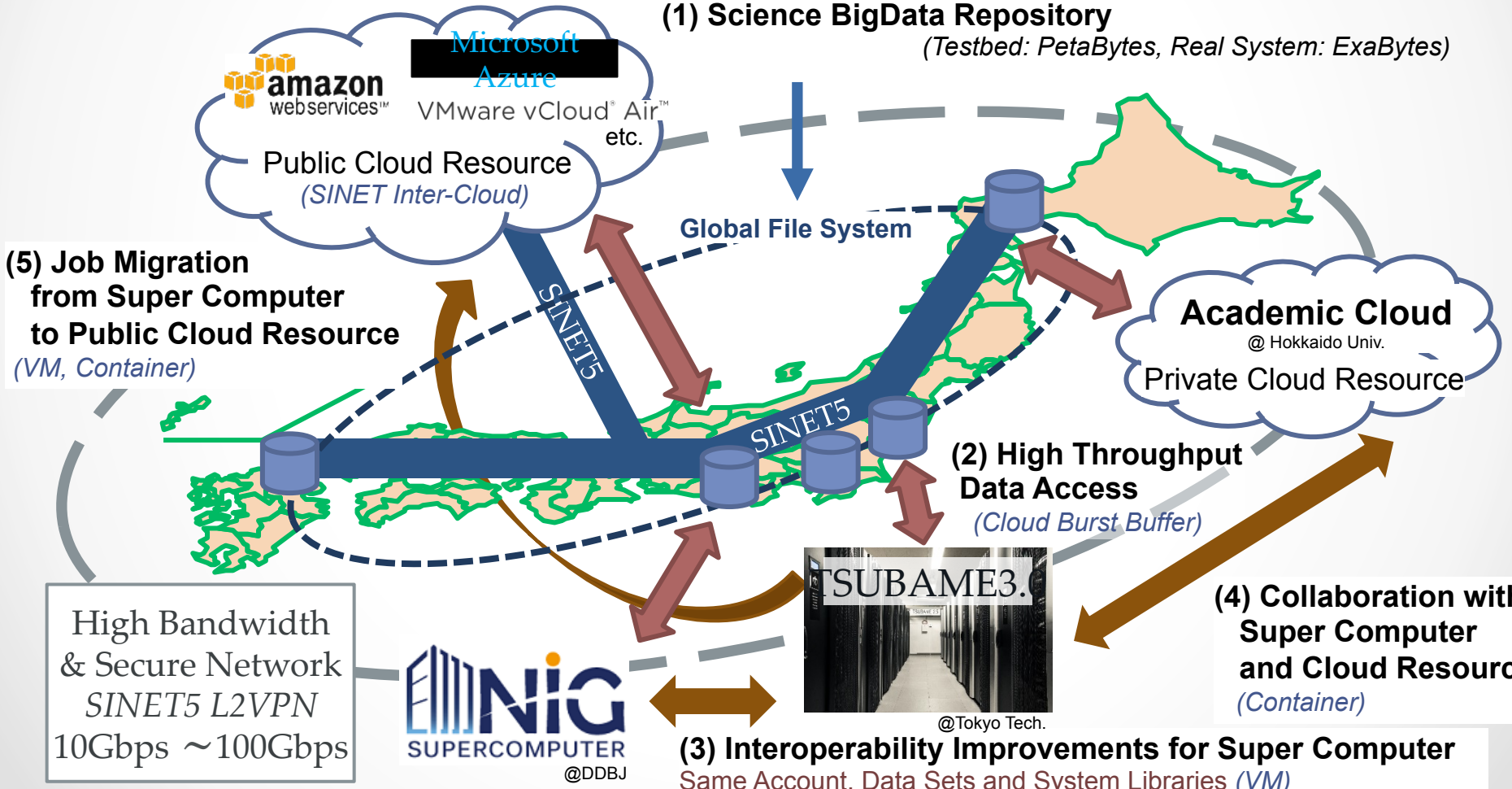
Helix Nebula Marketplace



情報研　横山先生のグループの研究

情報研,東工大などとの共同研究
各計算機センター間の連携・実証実験

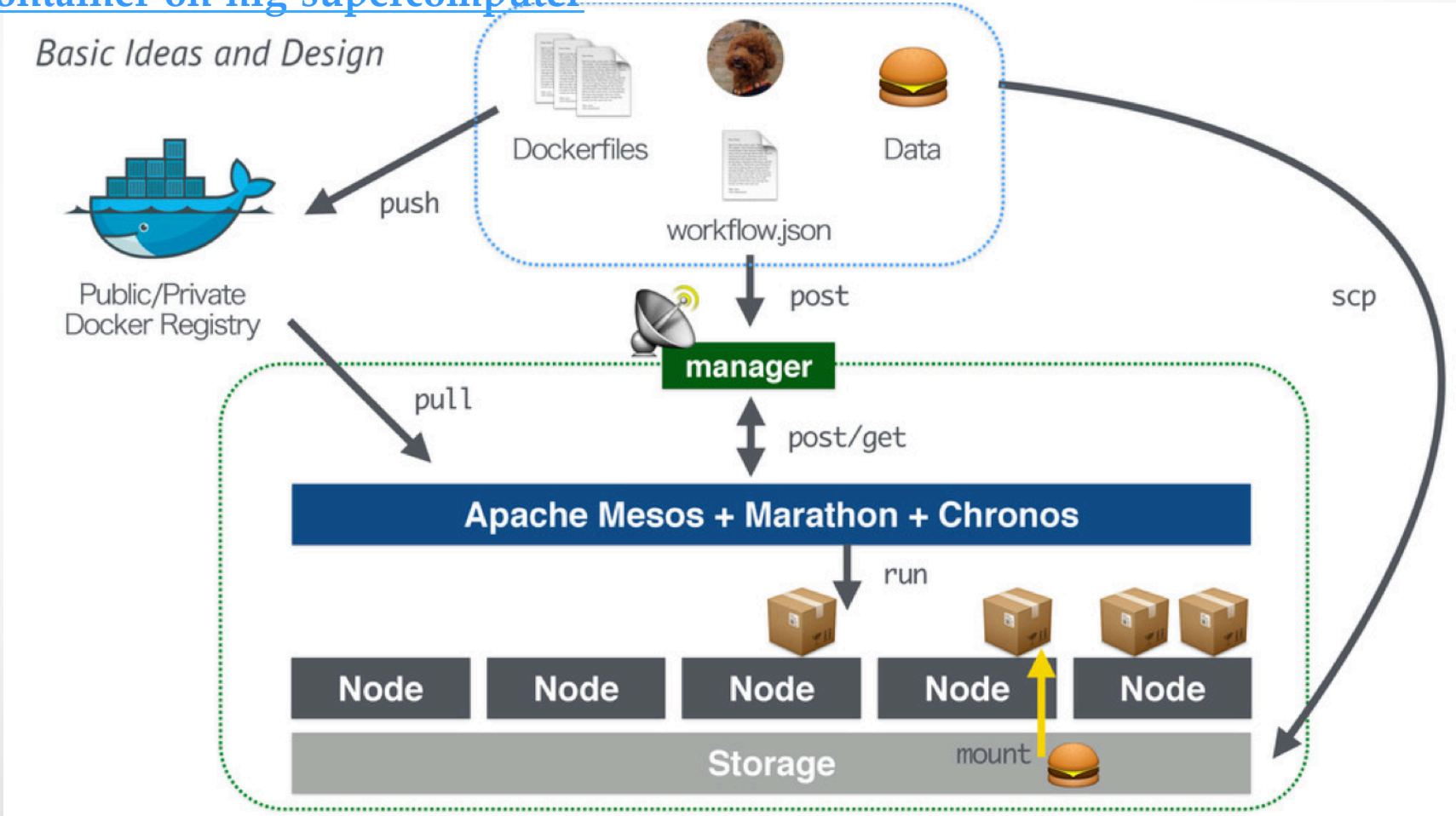
- **data store/access method** on inter-cloud
- **testbed operation**



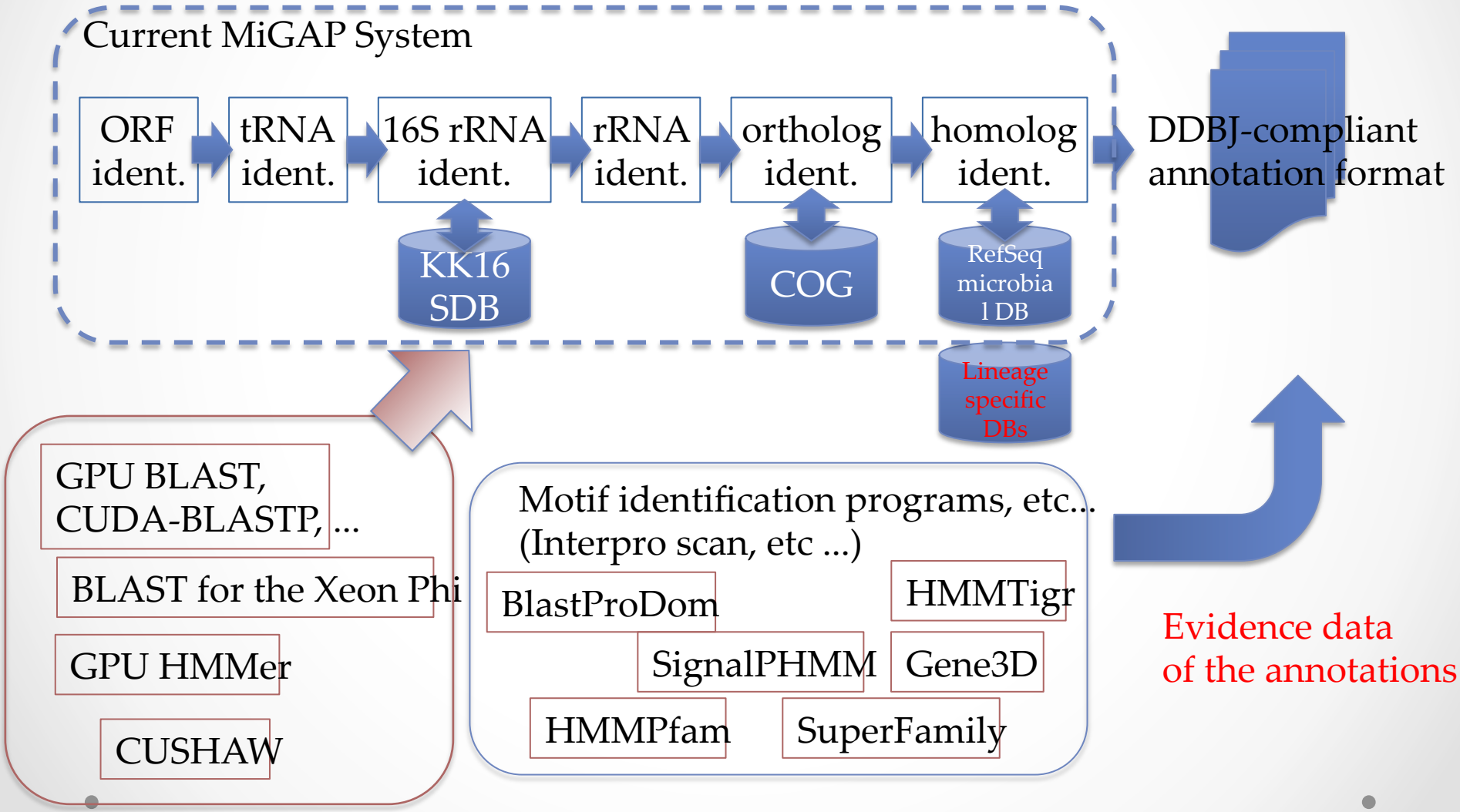
軽量仮想化環境上の大規模塩基配列データ解析フレームワーク開発およびスパコン間連携の実証実験 (DBCLS/NIG/NII)

RaaS: Reproducibility as a Service by docker container on NIG supercomputer by Tazro Ohta

https://speakerdeck.com/inutano/raas-reproducibility-as-a-service-by-docker-container-on-nig-supercomputer



Extension of MiGAP (Microbial Genome Annotation Pipeline)

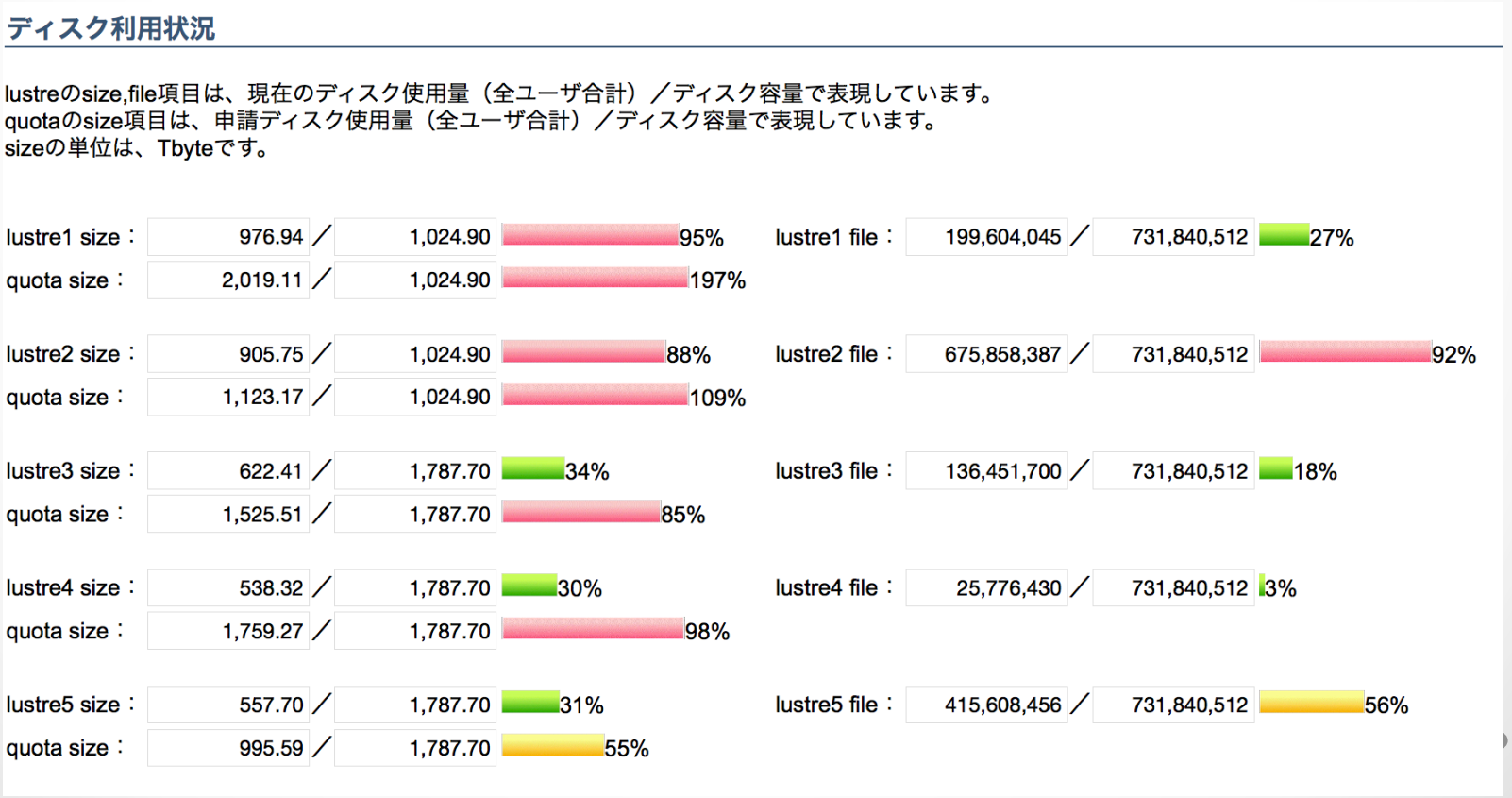


ストレージ(アーカイブ)

- アーカイブサイズの見通しは困難
  - 増加率1.4倍/年 => 2017年　5PB, 2022年　25～30PB
  - 増加率2倍/年　=> 2017年 10PB, 2022年 300PB
  - このほかに「個人ゲノムデータ」もある。
  - 電気代の問題。
  - スパコンリース終了時のデータ移行の方法？
- バックアップ
  - 遺伝研内におけるバックアップ？（軽微なディスクやテープの障害に対する対応）
  - ディザスタカバリー？遠隔地バックアップ
- データ利用の方法
  - ユーザーがアーカイブデータをダウンロードする方法
  - 全件アクセスは今後困難に => メタデータのデータベース管理

ストレージ（計算用）

- 伝統的Grid Engineの系 + OpenStack等の連携?
  - いわゆるSQL, NoSQLデータベースや全文検索エンジンの運用
  - その他ログの統計処理など



ディスク利用状況

lustreのsize,file項目は、現在のディスク使用量（全ユーザ合計）／ディスク容量で表現しています。
quotaのsize項目は、申請ディスク使用量（全ユーザ合計）／ディスク容量で表現しています。
sizeの単位は、Tbyteです。

lustre1 size：	976.94	1,024.90	95%	lustre1 file：	199,604,045	731,840,512	27%
quota size：	2,019.11	1,024.90	197%				
lustre2 size：	905.75	1,024.90	88%	lustre2 file：	675,858,387	731,840,512	92%
quota size：	1,123.17	1,024.90	109%				
lustre3 size：	622.41	1,787.70	34%	lustre3 file：	136,451,700	731,840,512	18%
quota size：	1,525.51	1,787.70	85%				
lustre4 size：	538.32	1,787.70	30%	lustre4 file：	25,776,430	731,840,512	3%
quota size：	1,759.27	1,787.70	98%				
lustre5 size：	557.70	1,787.70	31%	lustre5 file：	415,608,456	731,840,512	56%
quota size：	995.59	1,787.70	55%				