

トーゴーの日シンポジウム2011

「画期的な農畜産物作出のための
ゲノム情報データベース」の試み

(独)農業生物資源研究所
農業生物先端ゲノム研究センター
伊藤 剛

謝辞

本研究は農林水産省「画期的な農畜産物作出のためのゲノム情報データベースの整備」の委託を受けて行われています。

詳細については宮尾安藝雄のポスターも御覧下さい。

背景～これまでの状況

- 農業生物資源研究所は植物ゲノム研究、昆虫ゲノム研究、家畜ゲノム研究で成果を上げ、関連データベースを整備してきた。

Accessibility statement | Jump to main content | Jump to main navigation | Jump to search | Jump to login

NATURE.COM | NEWS@NATURE.COM | NATUREJOBS | NATUREEVENTS | ABOUT NPG

nature International weekly journal of science

Search journal | Advanced search

Journal home > Archive > Article > Abstract

Journal home
Advance online publication
Current issue
Archive
Supplements
Web focuses
About the journal
For authors and referees
Online submission
Reprints and Permissions

Gateways
Asia gateway
German gateway

NPG Subject areas
Biotechnology
Cancer
Chemistry
Clinical Practice & Research
Dentistry
Development
Drug Discovery
Earth Sciences

Article
#Nature 436, 793-800 (11 August 2005) | doi:10.1038/nature03895

The map-based sequence of the rice genome
International Rice Genome Sequencing Project*

Rice, one of the world's most important food plants, has important syntenic relationships with the other cereal species and is a model plant for the grasses. Here we present a map-based sequence of the 389 Mb genome, including virtually all of the euchromatin and 80% of the predicted genes. A total of 37,544 non-transposable-element-related protein-coding genes were identified, of which 71% had a putative homologue in *Arabidopsis*. In a reciprocal analysis, 90% of the predicted genes in *Arabidopsis* appear in clustered gene families. The number and classes of genes in the rice genome are consistent with the expansion of syntenic regions in the grasses. We find evidence for widespread and recurrent gene transfer from the organelles to the nuclear chromosomes. The map-based sequence has proven useful for the identification of agronomic traits. The additional single-nucleotide polymorphisms and simple sequence repeats identified in this study should accelerate improvements in rice production.

1. Affiliations for participants: National Institute of Agrobiological Sciences/Institute of the Society for Techno-Forestry and Fisheries, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan
2. The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
3. Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (CAS), 900 Caobao Road, Shanghai 200031, China
4. Centre National de Séquençage, INRA-URGV, and CNRS UMR-8030, 2, rue Gaston Crémieux, CP 5706, 91191 Evry-Courcouronnes, France
5. UMR PIA, Cirad-Amis, TA40-03 avenue Agropolis, 34396 Montpellier Cedex 05, France
6. Department of Plant Sciences, BIOS Institute, The University of Arizona, Tucson, Arizona 85721, USA
7. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11723, USA
8. Institute of Botany, Academia Sinica, 128, Sec. 2, Yen-Chiu-Yuan Rd, Nankang, Taipei 11529, Taiwan
9. National Cheng Kung University, No. 1, Ta-Hsueh Road, Tainan 701, Taiwan
10. National Yang-Ming University, 155, Sec. 2, Li-Nong St, Peltou, Taipei 112, Taiwan
11. Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi 110021, India
12. National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi 110018, India
13. Waksman Institute, Rutgers University, Piscataway, New Jersey 08854, USA

Rice genome sequencing (*Nature*, 436:793-800, 2005)

Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*

The Rice Annotation Project^{1,2}

We present here the annotation of the complete genome of rice (*Oryza sativa* ssp. *japonica*). The genome contains 37,544 protein-coding genes, 1,200 non-coding RNA genes, and 1,200 insertion elements. Functional annotations for proteins and non-protein-coding RNAs were identified or inferred in 19,969 (70%) of the predicted genes. A total of 1,200 antisense transcripts were found. Almost 5000 annotated protein-coding genes were found. Insertional mutant lines, which will accelerate future experimental determination of gene function, were identified. The genome is annotated based on these loci and an extrapolation suggested that the gene content of the rice genome is similar to that of other grass genomes. We conducted comparative analyses between rice and *Arabidopsis thaliana* genomes and found that rice genomes possessed several lineage-specific genes, which might account for the differences between the two species, while they had similar sets of predicted functional domains. The efficiency of gene prediction seems to be conserved across grasses. The efficiency of gene prediction of protein-coding genes was examined. Our analysis shows that the rice genome contains duplicated genes in both species, so that duplication is a common feature of the genome.

Complete list of authors

Takeshi Itoh,^{1,2} Tsuyoshi Tanaka,^{1,3} Roberto A. Barrero,³ Chisato Yamasaki,^{2,4} Yasuyuki Fujii,^{2,4} Phillip B. Hilton,^{2,4} Baltazar A. Antonito,¹ Hideo Aono,³ Rolf Apweiler,³ Richard Bruskiewich,⁵ Thomas Bureau,⁷ Frances Burr,⁸ Antonio Costa de Oliveira,⁹ Gallina Fuku,¹⁰ Takuya Habara,^{2,4} Georg Haberer,¹¹ Bin Han,¹² Ertim Hara,¹² Aiko T. Hiraki,⁴ Hirohiko Hirochika,¹ Douglas Hoen,⁷ Hiroki Hokari,⁴ Satomi Hosokawa,¹³ Yue-le Hsing,¹⁴ Hiroshi Ikawa,¹⁵ Kazuo Ikeno,³ Tadashi Imanishi,^{2,16} Yukiyo Ito,¹³ Pankaj Jaiswal,¹⁷ Masako Kanno,^{2,4} Yoshihiro Kawahara,^{2,18} Toshiyuki Kawamura,⁴ Hiroaki Kawashima,⁴ Jitendra P. Khurana,¹⁹ Shoshi Kikuchi,¹ Setsuko Komatsu,^{1,20} Kanako O. Koyanagi,¹⁶ Hitromi Kubooka,⁴ Damien Lieberherr,²¹ Yao-Cheng Lin,¹⁴ David Lonsdale,² Takashi Matsumoto,¹ Akihiro Matsuya,⁴ W. Richard McCombie,²² Joachim Messing,¹⁰ Akio Miyao,¹ Nicola Mulder,⁵ Yoshiaki Nagamura,³ Jongmin Nam,^{23,24} Nobukazu Namiki,¹³ Hisataka Numa,¹ Shin Nurimoto,⁴ Claire O'Donovan,⁵ Hajime Ohyanagi,^{3,15} Toshihisa Okido,² Satoshi Ohta,²⁵ Naoki Osato,² Lance E. Palmer,^{22,26} Francis Quetier,²⁷ Saurabh Raghuvanshi,¹⁹ Naomi Saichi,^{2,4} Hiroaki Sakai,^{1,4} Yasumichi Sakai,¹⁵ Katsumi Sakata,¹⁵ Tetsuya Sakurai,²⁸ Fumihiko Sato,⁴ Yoshiharu Sato,^{2,4} Heiko Schoof,^{11,29,30} Motoaki Seki,³¹ Michie Shibata,¹³ Yuji Shimizu,¹⁵ Kazuo Shinozaki,³² Yuji Shinso,⁴ Nagendra K. Singh,³³ Brian Smith-White,³⁴ Jun-ichi Takeda,^{2,4} Motohiko Tanino,^{2,4} Tatiana Tatusova,³⁴ Supat Thongjuea,³⁵ Fusano Todorokoro,⁴ Mika Tsugane,¹² Akhlesh K. Tyagi,¹⁹ Apichart Vanavichit,³⁵ Aihui Wang,³⁶ Rod A. Wing,³⁷ Kaori Yamaguchi,⁴ Mayu Yamamoto,¹³ Naoyuki Yamamoto,⁴ Yeisoo Yu,³⁷ Hao Zhang,⁴ Qiang Zhao,¹² Kenichi Higo,^{38,39} Benjamin Burr,⁴ Takashi Gojobori,^{2,3} and Takuji Sasaki²⁸

Rice genome annotation (*Genome Res*, 17:175-183, 2007)

背景～これまでの状況

- 農業生物資源研究所は植物ゲノム研究、昆虫ゲノム研究、家畜ゲノム研究で成果を上げ、関連データベースを整備してきた。

Contents lists available at ScienceDirect

 **Insect Biochemistry and Molecular Biology**

journal homepage: www.elsevier.com/locate/ibmb



The genome of a lepidopteran model insect, the silkworm *Bombyx mori*

The International Silkworm Genome Consortium ¹

ARTICLE INFO

Article history:
Received 27 November 2008
Received in revised form
28 November 2008
Accepted 28 November 2008

Keywords:
Bombyx mori
Silkworm
Genome
Transposable elements
Silk production
Gene duplication

Archibald *et al.* *BMC Genomics* 2010, **11**:438
<http://www.biomedcentral.com/1471-2164/11/438>



CORRESPONDENCE

Open Access

Pig genome sequence - analysis and publication strategy

Alan L Archibald^{1*}, Lars Bolund^{2,3}, Carol Churcher⁴, Merete Fredholm⁵, Martien AM Groenen⁶, Barbara Harlizius⁷, Kyung-Tai Lee⁸, Denis Milan⁹, Jane Rogers¹⁰, Max F Rothschild¹¹, Hirohide Uenishi¹², Jun Wang^{2,13}, Lawrence B Schook^{14*}, the Swine Genome Sequencing Consortium

背景～これまでの状況

- 農業生物資源研究所は植物ゲノム研究、昆虫ゲノム研究、家畜ゲノム研究で成果を上げ、関連データベースを整備してきた。
- さらに、種内や近縁種の比較解析を推進している。

Yamamoto et al. *BMC Genomics* 2010, **11**:267
<http://www.biomedcentral.com/1471-2164/11/267>

コシヒカリのゲノム配列決定



RESEARCH ARTICLE

Open Access

Fine definition of
closely related
genome-wide
polymorphisms

Toshio Yamamoto[†], Hideki
Masahiro Yano^{*}

the plant journal



The Plant Journal (2011) 66, 796–805 **アフリカイネのゲノム配列決定** 10.1111/j.1365-313X.2011.04539.x

Distinct evolutionary patterns of *Oryza glaberrima* deciphered by genome sequencing and comparative analysis

Hiroaki Sakai¹, Hiroshi Ikawa^{2,3}, Tsuyoshi Tanaka¹, Hisataka Numa¹, Hiroshi Minami^{2,3}, Masaki Fujisawa¹, Michie Shibata², Kanako Kurita², Ari Kikuta², Masao Hamada², Hiroyuki Kanamori², Nobukazu Namiki^{2,3}, Jianzhong Wu¹, Takeshi Itoh¹, Takashi Matsumoto^{1,*} and Takuji Sasaki¹

¹Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan,

²Research Division I, Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki 305-0854, Japan, and

³Tsukuba Division, Mitsubishi Space Software Co., Ltd., Tsukuba, Ibaraki 305-0032, Japan

背景～これまでの状況

- 農業生物資源研究所は植物ゲノム研究、昆虫ゲノム研究、家畜ゲノム研究で成果を上げ、関連データベースを整備してきた。
- さらに、種内や近縁種の比較解析を推進している。

The Plant Cell, Vol. 23: 1249–1263, April 2011, www.plantcell.org © 2011 American Society of Plant Biologists

LARGE-SCALE BIOLOGY ARTICLE

オオムギのゲノム配列決定

Unlocking the Barley Genome by Chromosomal and Comparative Genomics ^{[M][OA]}

Klaus F.X. Mayer,^{a,1} Mihaela Martis,^a Pete E. Hedley,^b Hana Šimková,^c Hui Liu,^b Jenny A. Morris,^b Burkhard Steuernagel,^d Stefan Taudien,^e Stephan Roessner,^a Heidrun Gundlach,^a Marie Kubaláková,^c Pavla Suchánková,^c Florent Murat,^f Marius Felder,^e Thomas Nussbaumer,^a Andreas Graner,^d Jerome Salse,^f Takashi Endo,^g Hiroaki Sakai,^h Tsuyoshi Tanaka,^h Takeshi Itoh,^h Kazuhiro Sato,ⁱ Matthias Platzer,^e Takashi Matsumoto,^h Uv

Comprehensive Sequence Analysis of 24,783 Barley Full-Length cDNAs Derived from 12 Clone Libraries ^{[W][OA]}

オオムギの完全長cDNA配列決定

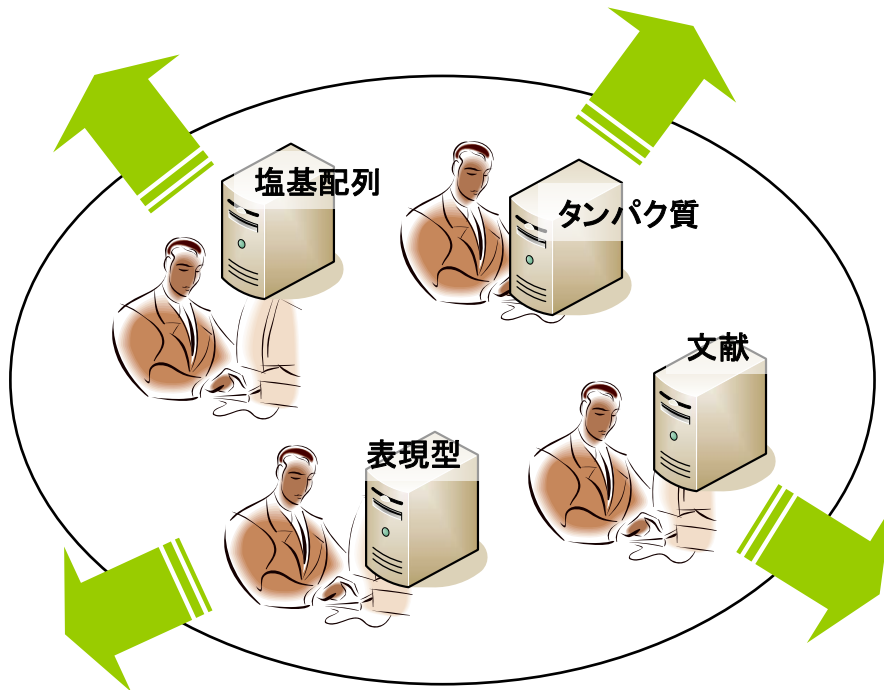
Takashi Matsumoto^{2*}, Tsuyoshi Tanaka², Hiroaki Sakai, Naoki Amano, Hiroyuki Kanamori, Kanako Kurita, Ari Kikuta, Kozue Kamiya, Mayu Yamamoto, Hiroshi Ikawa, Nobuyuki Fujii, Kiyosumi Hori, Takeshi Itoh, and Kazuhiro Sato

背景～これまでの状況

- 農業生物資源研究所は植物ゲノム研究、昆虫ゲノム研究、家畜ゲノム研究で成果を上げ、関連データベースを整備してきた。



- これらの成果を受け、農林水産生物ゲノム情報統合データベースのプロジェクトを推進してきた。
- 作物研では、イネの品種・特性データを、日本中の機関と共同で収集、提供。



農林水産生物ゲノム情報統合データベース

バラバラに作られ、発信されていたデータベース

本計画の全体像

H23～H27

ゲノム解読新時代

- 膨大な配列情報が蓄積(数百塩基の断片が数億本の単位で)
- 多様な生物種で全ゲノムが決まる

これまでと違い膨大な情報が生産される。
ゲノム競争に勝てる仕組みの樹立、その
ための実験研究者への支援が必要。

情報を最大限に活用するシステムの開発・確立・利用

①データベースの改良、運用

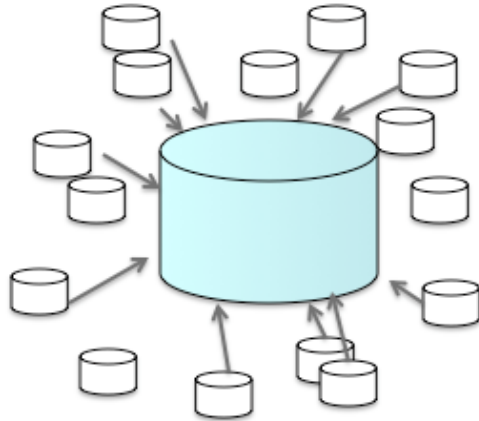
過去の資産を継承しつつ、大学や民間の幅広い研究者がデータベースを横断しながら効率的に情報を取得できるシステムを整備する。

②高次解析システムの開発

大量の断片配列を繋ぎ合わせて整理し、そこからマーカを整備したり、配列情報解析によって有用遺伝子を探索する。

研究者がストレス無く作業できる環境を提供する

H18～H22



有機的統合

利便性の
向上追求

農水統合DBポータル

イネ統合ブラウザ



カイコ統合ブラウザ

下記2機関で研究グループを組織する

【中核機関】



【プロジェクトリーダー】

- ・研究実施場所:
基盤研究領域(つくば市)
- ・研究課題:
データベースの改良、運用
高次解析システムの開発

ゲノム情報、リソース、植物、昆虫、動物の主要な研究者が参加

【共同研究機関】



- ・研究実施場所:
低コスト稲育種研究チーム
(つくば市)
- ・研究課題:
データベースの改良、運用
(イネの品種・特性情報)

イネの品種・特性データを広範に収集し、データ構築を行うための全日本的ネットワークがある

データベースの運用と管理

これまでに多数のサービスを公開、運用してきている。

- **ウェブポータル**(NIAS DNA Bankウェブサイト、ゲノムリソースセンターウェブサイト、anonymous-ftpによる各種データベースの提供、等)
- **植物系サーバ**(イネ完全長cDNAクラスタリング、イネミトコンドリアゲノム、イネタンパク構造データベース、44Kマイクロアレイ用BLAST検索、RGP EST 自動BLAST解析、イネいもち病菌 cDNA BLAST解析、Rice Gateway Library Database、オオムギcDNA、ダイズゲノム、フェノーム、RAP-DB、ミュータントパネル、RiceXPro、Q-Taro、イネ品種・特性データベース)
- **イネゲノムプロジェクトサーバ**(MDATAデータ提供用、MDATA BLASTサーバ)
- **イネ完全長cDNAサーバ**(KOME: Knowledge-based Oryza Molecular biological Encyclopedia、RMOS: Rice Microarray Opening Site、RED: Rice Expression Database、RED II: Rice Expression Database II、Rice Pipeline、RPD: Rice Proteome Database)
- **昆虫系サーバ**(遺伝子予測関連ツール、EST/完全長cDNA関連ツール、プロテオーム関連ツール、ミュータントパネル関連ツール、地図関連ツール、相同性検索関連ツール、ホームページ/ダウンロードサイト関連ツール)
- **動物系サーバ**(AGPウェブサイト、Pig Expression Data Explorer : Database of full-length cDNA clones and ESTs in pigs、自動BLAST解析サービス)
- **微生物系サーバ**(イネ白葉枯病菌、Genome Gambler、Rice HMM Viewer、イネいもち病菌EST)
- **解析ウェブサーバ**(PLACE: A Database of Plant Cis-acting Regulatory DNA Elementsの関連サービス、RiceGAAS、RiceBLAST、Rice-GD、RICEDB: RGP Genome MAP Information、GLocate: イネゲノム遺伝子予測プログラム、RiceGAAS解析処理、KAIKOGAAS解析処理)

データベースの運用と管理

- 切れ目の無いデータ提供
- 大型システムの安定運用



専門知識を持った
人員で対応

これまでに
大きなデータベースシステムを
長期間にわたって
運用してきている

- データベース運用管理
- WWW運用管理
- ユーザ管理・支援
- メール管理
- ファイル管理
- ログ管理
- ネットワーク運用管理
- セキュリティ運用管理

新規データやデータベースは次々と個別に作られる

- 開発はしたがプロジェクトが終了した
- 適当なハードが維持できない
- データはあるがどうしたらよいか分からない



受け皿となるべく、
個別の研究者に
対して打ち合わ
せの上、支援す
る。

利便性向上等のため各種のご意見を反映

生命科学データベースアーカイブ構築機能の開発

NBDC/DBCLSとの連携～その1



-あのデータベースが、丸ごとダウンロード可能に！-
生命科学系データベース アーカイブ

アーカイブ横断検索

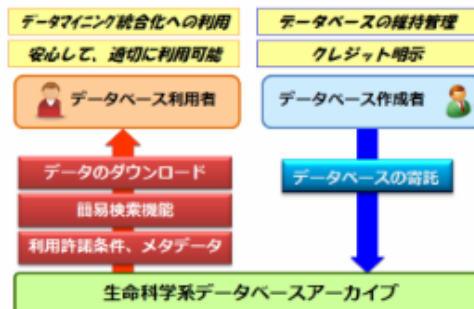
検索

ホーム アーカイブの説明 寄託応募要領 更新履歴 ヘルプ お問い合わせ

いくら良質なデータベースでも、説明が十分でない、利用条件が明確でない、ダウンロードできないなどの理由で十分に利用され、引用され、相応しい評価を受ける機会を逃していることがあります。

生命科学系データベースアーカイブは、国内のライフサイエンス研究者が生み出したデータセットをわが国の公共財としてまとめて長期間安定に維持保管し、データ説明(メタデータ)を統一して検索を容易にすると共に、利用許諾条件などの明示を行うことで、多くの人が容易にデータへアクセスしダウンロードを行えるようにするサービスです(詳細説明)。

データを長期にわたり保全し、データベース作成者のクレジットを明示する一方、公的機関や民間等様々なユーザが利用しやすい研究にすることで、それぞれの研究の生命科学へのいっそうの貢献を支援します。データベースの寄託を随時募集しています(寄託応募要領)。

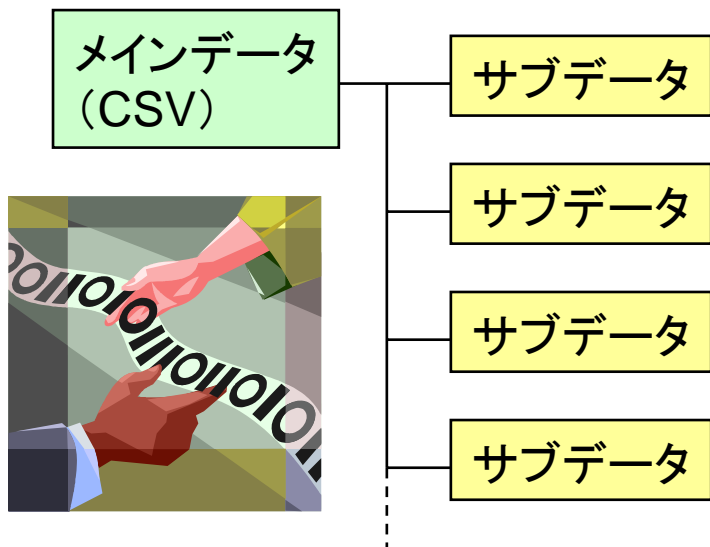


新アーカイブ情報

2011/08/29 Medaka Full-length cDNA Database (基礎生物学研究所 成瀬清准教授)を追加しました。
2011/08/25 「D-HaploDB」(九州大学 林健志名誉教授)を追加しました
2011/08/25 「RIKEN SSBC」(理化学研究所 横山茂之領域長)を追加しました

生命科学データベースアーカイブ構築機能の開発

NBDC/DBCLSとの連携～その1



- 1つのアーカイブに1つのメインデータ
- テキスト以外はサブデータとして持つ
- その他、ガイドラインに沿って作成

NBDC/DBCLSと連携し、アーカイブを作成する。

各データベースで、ガイドラインに沿ったデータの再編と加工。

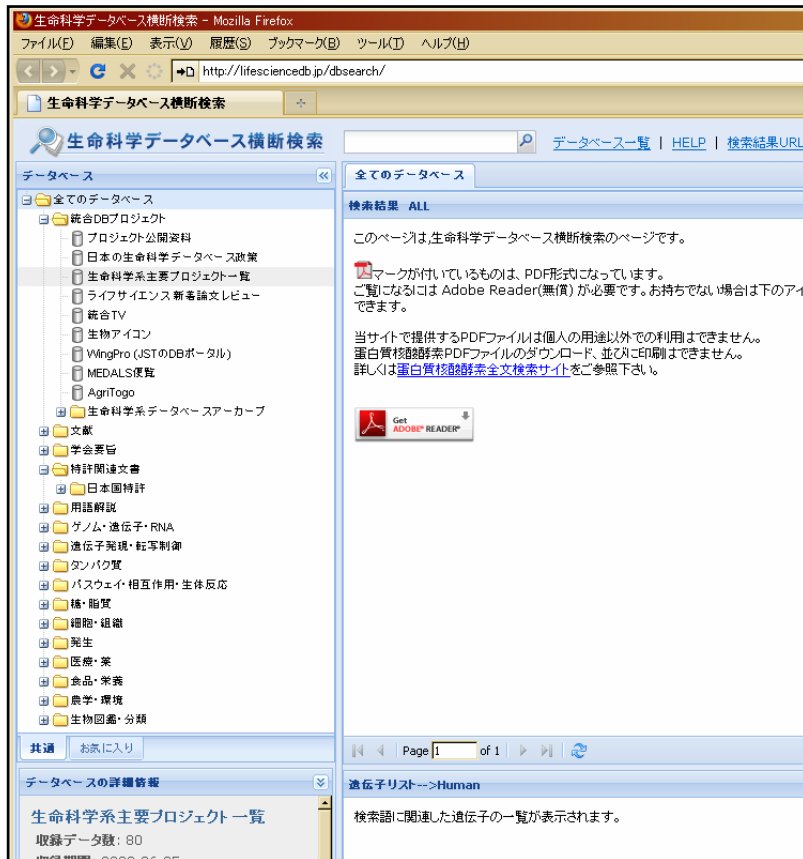
このため、各データベース内容の精査し、既に候補を幾つか選定。変換プログラム作成。

将来的には、定期的なアーカイブ化とデータ提供の実現。

定められた形式で確実にデータの受け渡しを行えるようにする

生命科学データベース横断検索機能の開発

NBDC/DBCLSとの連携～その2



NBDC/DBCLSと連携し、横断検索機能を開発する。

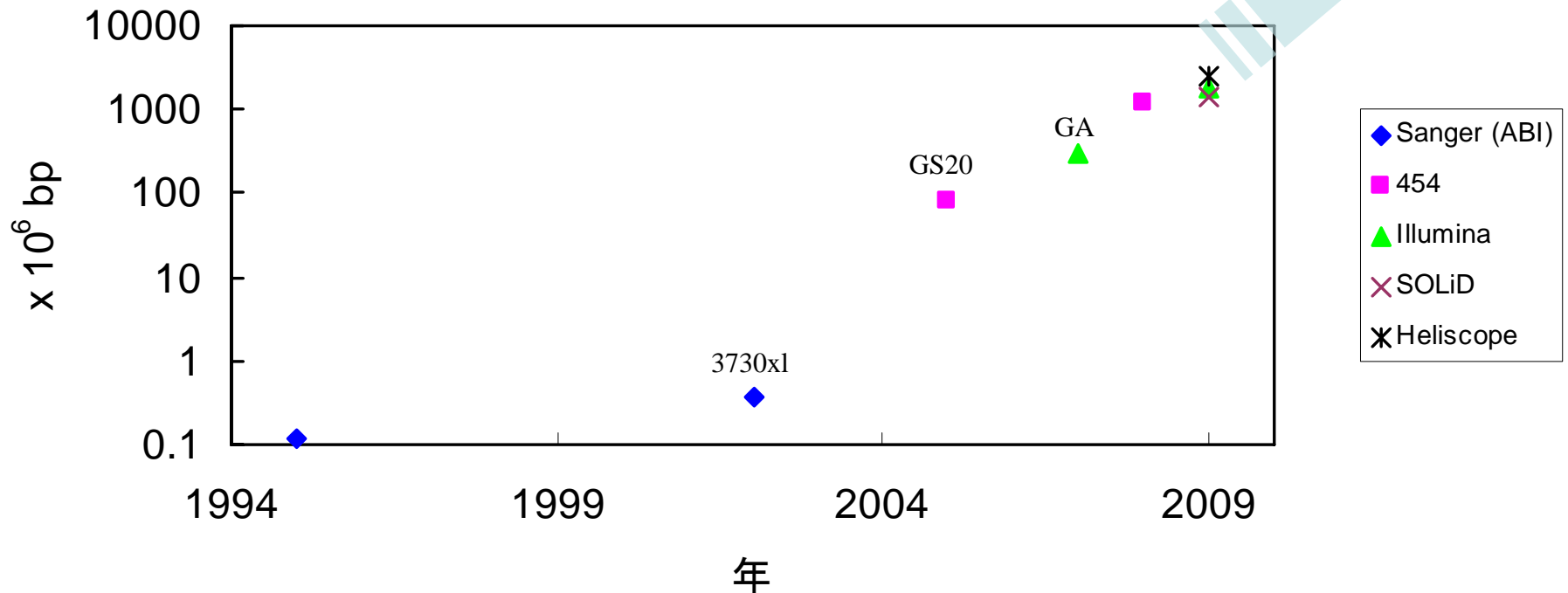
検索エンジンとして **Hyper Estraier** を使用する。

本提案の管理下にある **データベースのインデックスを作成** する。(定期的作業が必要)

超大量シーケンシング時代への対応

更に加速？

劇的に加速する「一日あたり」の配列決定量



現在は必要最小限のものだけ残しているが、それでも巨大なファイルになる。

解析を行うと数倍に膨れ上がる。

大型ファイルサーバーの高価なディスクは予算が追いつかない。

安価なディスクを積み上げて対処していると、今度は物理的スペースが...



現在、一回あたり数テラのデータを見込んで考えなければならない。

Nature Methods 2009年9月号の論説

Metagenomics versus Moore's law

Metagenomics sprang from advances in sequencing technology, and continued improvements are providing data in quantities unimaginable a few years ago. But without concerted efforts, the amount of data will quickly outpace the ability of scientists to analyze it.

As Craig Venter sails the oceans collecting seawater samples to profile microbial communities by high-throughput sequence analysis, microbiologists around the world are busy collecting their own samples. The diversity of locations—from Antarctic lakes to human armpits—highlights the reality that microscopic organisms represent a significant fraction of the Earth's ecosystem.

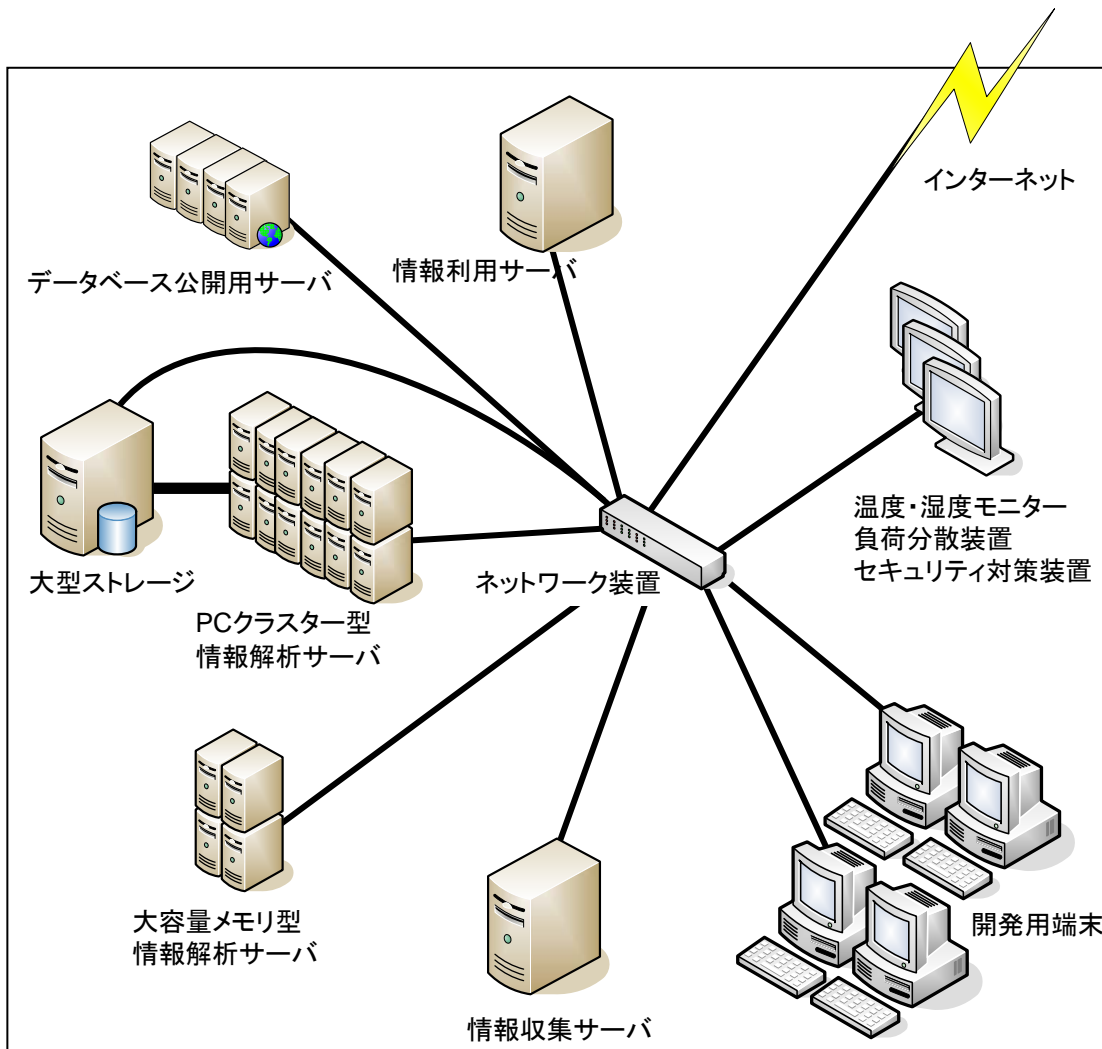
Any population this large is certain to have profound influences on its environment. Yet our knowledge of these communities and their functions is rudimentary, partly owing to our inability to culture the vast majority of

40 megabases. Today there are over 4,000 sequenced metagenomes, and their size and number are increasing. Each new pyrosequenced metagenome is 200–500 megabases, and those generated on Illumina platforms are 20–50 gigabases. To analyze these metagenomes using established pipelines would take tens of years on a single processor and weeks to months on machines with up to 1,000 processors. The rate of increase in sequence generation is far outpacing Moore's law, and the cost of analyzing the largest datasets already exceeds the cost of generating them.

Analysis of new metagenomes requires assembly. σ ene

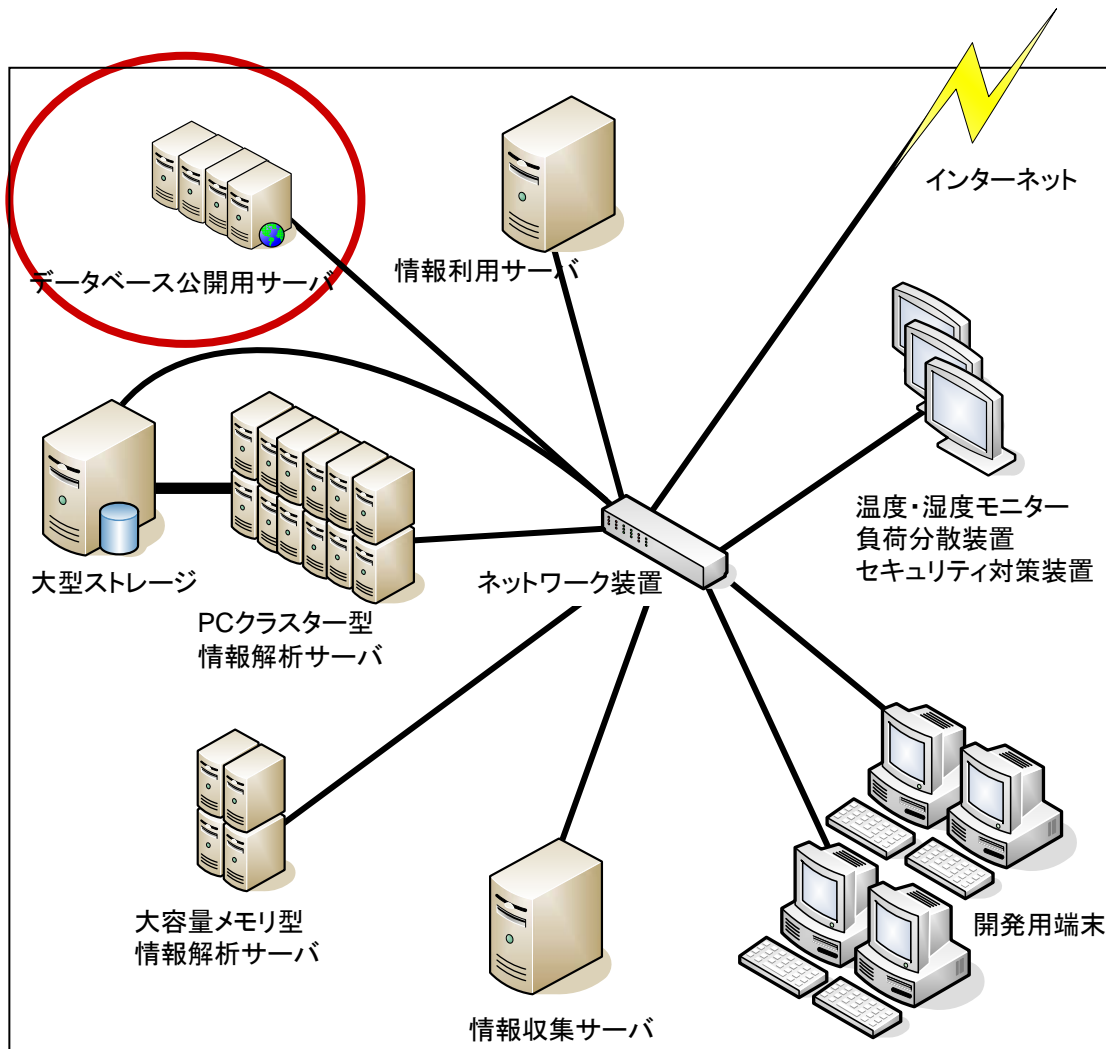
... (高速シーケンサーの) Illuminaで生産される塩基は20~50ギガにもなり... 配列生産の増加速度はムーアの法則をはるかに上回り、(現在)最大のデータセット解析のコストは、データ生産コストを既に超えている。

新しいシステムの導入 (今年度後半～)



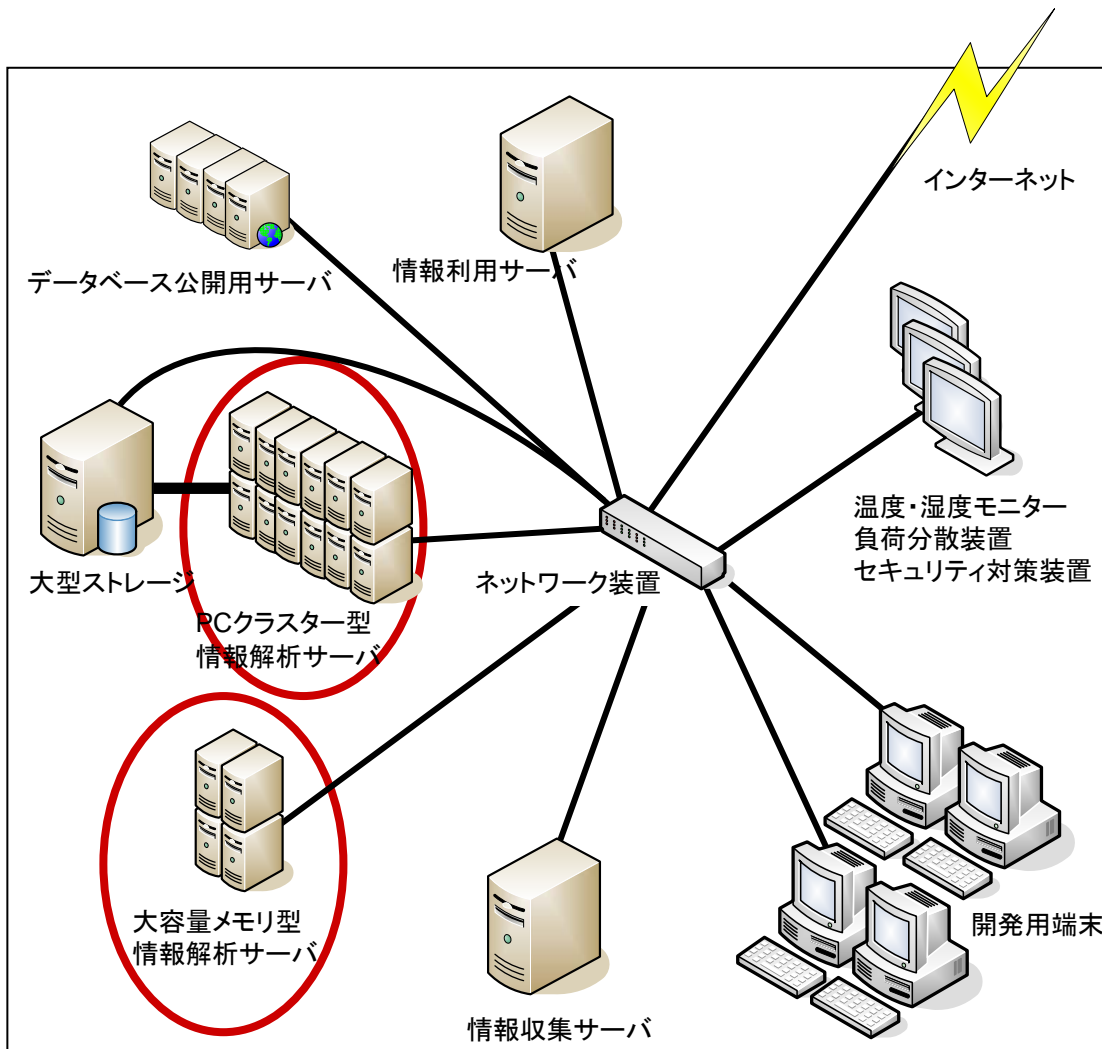
1. データベース公開用サーバ(年間30万訪問件数に対応)
2. PCクラスター型情報解析サーバ(100ノードを高速ネットワーク装置で接続、合計で1000コア以上。大規模遺伝子予測を遂行できる)
3. 大容量メモリ型情報解析サーバ(主記憶を2TB、2台で構成する。特にメモリを要求するゲノム断片の整列化機能等に対応)
4. 大型ストレージ(2PBが利用可能なディスク容量、高速バックアップに対応)
5. 情報収集サーバ(外部データベースからのデータ収集)
6. 情報利用サーバ(ユーザーがログインして収集データをオンライン解析で利用)
7. ネットワーク装置(全体を10Gbpsで構成、大量情報の転送に対応)

新しいシステムの導入 (今年度後半～)



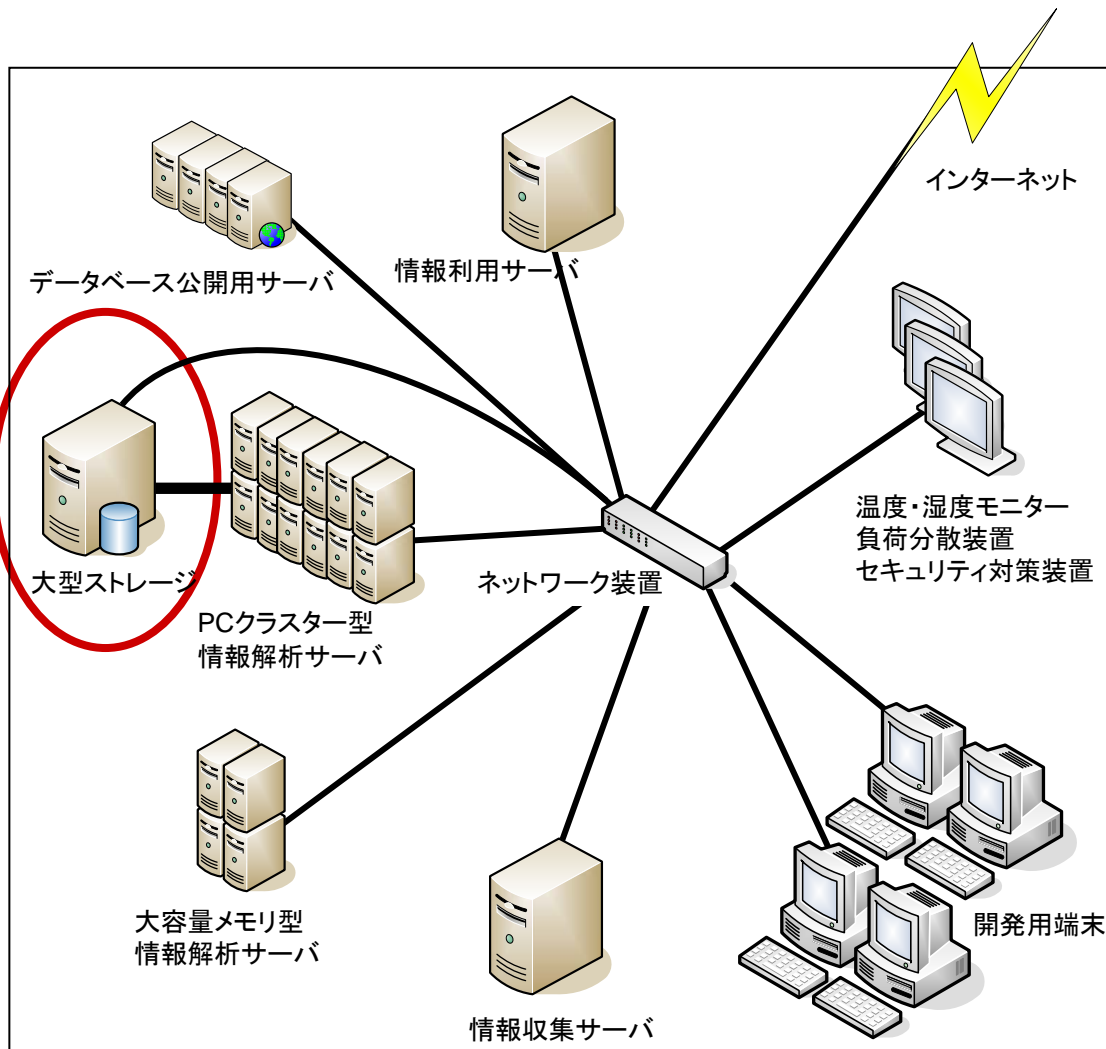
1. データベース公開用サーバ(年間30万訪問件数に対応)
2. PCクラスター型情報解析サーバ(100ノードを高速ネットワーク装置で接続、合計で1000コア以上。大規模遺伝子予測を遂行できる)
3. 大容量メモリ型情報解析サーバ(主記憶を2TB、2台で構成する。特にメモリを要求するゲノム断片の整列化機能等に対応)
4. 大型ストレージ(2PBが利用可能なディスク容量、高速バックアップに対応)
5. 情報収集サーバ(外部データベースからのデータ収集)
6. 情報利用サーバ(ユーザーがログインして収集データをオンライン解析で利用)
7. ネットワーク装置(全体を10Gbpsで構成、大量情報の転送に対応)

新しいシステムの導入 (今年度後半～)



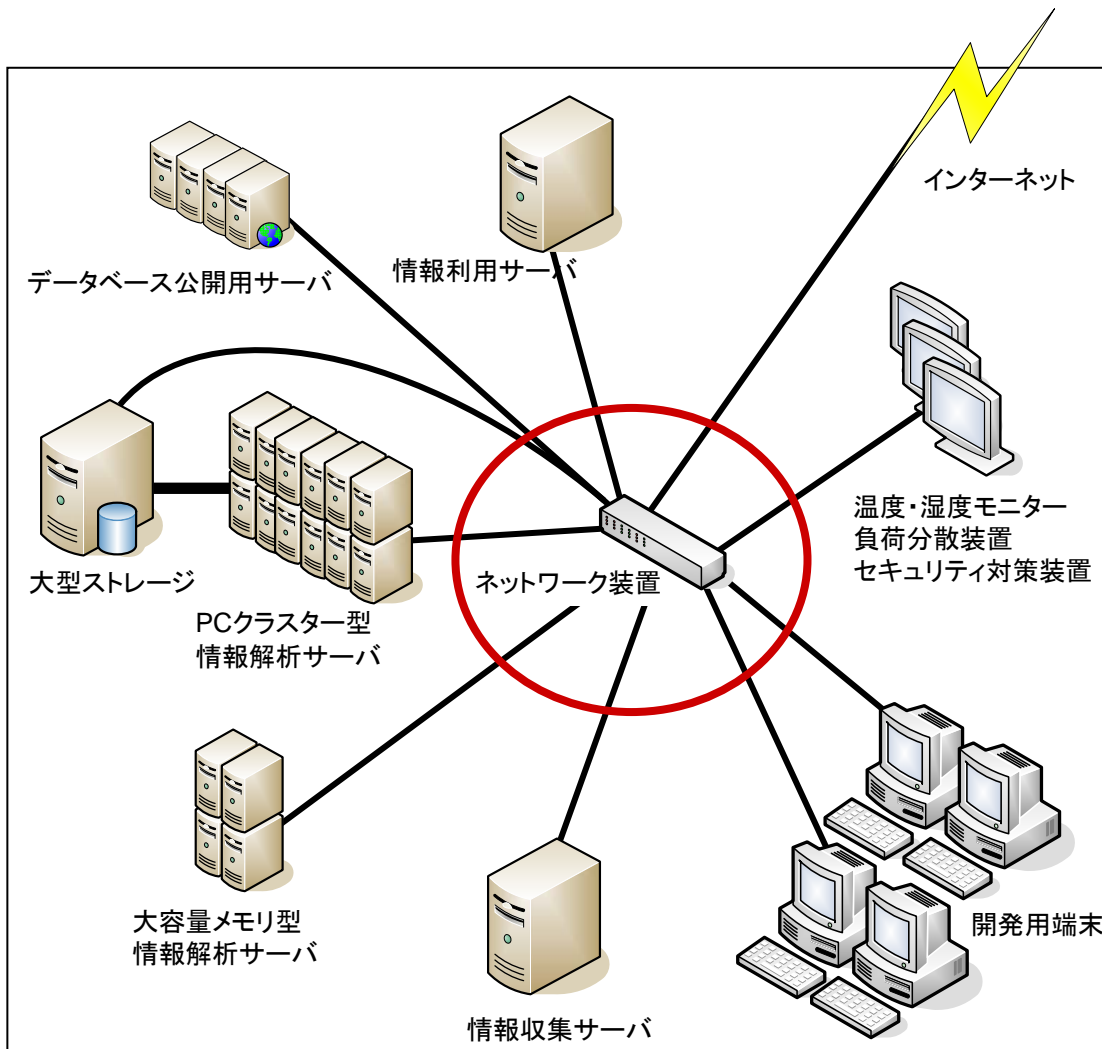
1. データベース公開用サーバ(年間30万訪問件数に対応)
2. PCクラスター型情報解析サーバ(100ノードを高速ネットワーク装置で接続、合計で1000コア以上。大規模遺伝子予測を遂行できる)
3. 大容量メモリ型情報解析サーバ(主記憶を2TB、2台で構成する。特にメモリを要求するゲノム断片の整列化機能等に対応)
4. 大型ストレージ(2PBが利用可能なディスク容量、高速バックアップに対応)
5. 情報収集サーバ(外部データベースからのデータ収集)
6. 情報利用サーバ(ユーザーがログインして収集データをオンライン解析で利用)
7. ネットワーク装置(全体を10Gbpsで構成、大量情報の転送に対応)

新しいシステムの導入 (今年度後半～)



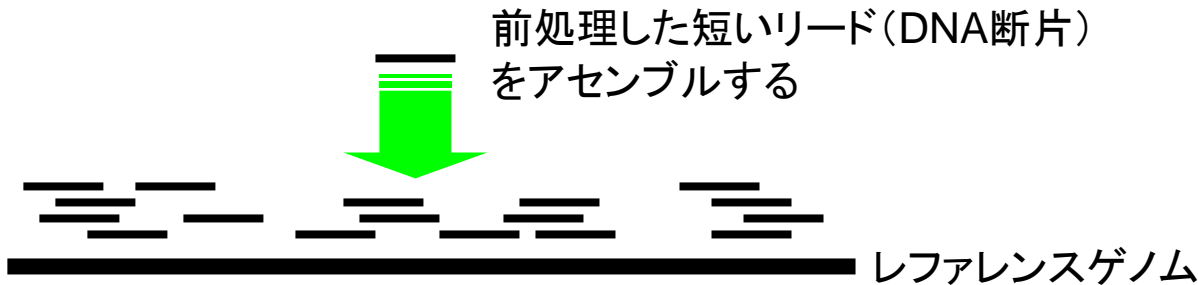
1. データベース公開用サーバ(年間30万訪問件数に対応)
2. PCクラスター型情報解析サーバ(100ノードを高速ネットワーク装置で接続、合計で1000コア以上。大規模遺伝子予測を遂行できる)
3. 大容量メモリ型情報解析サーバ(主記憶を2TB、2台で構成する。特にメモリを要求するゲノム断片の整列化機能等に対応)
4. 大型ストレージ(2PBが利用可能なディスク容量、高速バックアップに対応)
5. 情報収集サーバ(外部データベースからのデータ収集)
6. 情報利用サーバ(ユーザーがログインして収集データをオンライン解析で利用)
7. ネットワーク装置(全体を10Gbpsで構成、大量情報の転送に対応)

新しいシステムの導入 (今年度後半～)



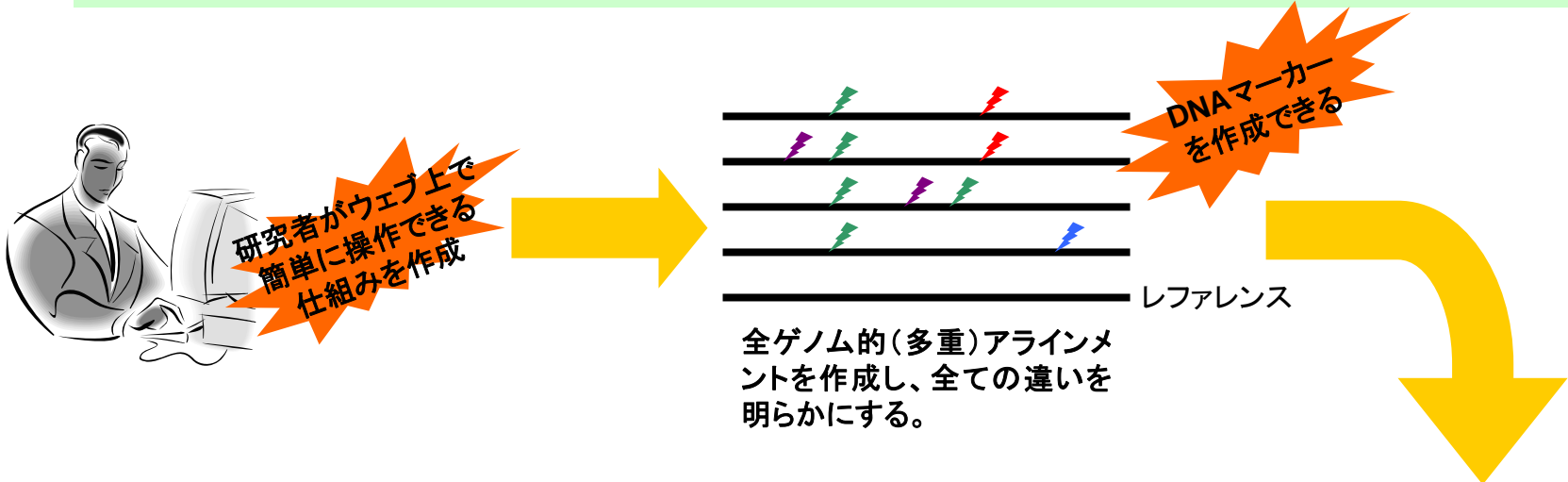
1. データベース公開用サーバ(年間30万訪問件数に対応)
2. PCクラスター型情報解析サーバ(100ノードを高速ネットワーク装置で接続、合計で1000コア以上。大規模遺伝子予測を遂行できる)
3. 大容量メモリ型情報解析サーバ(主記憶を2TB、2台で構成する。特にメモリを要求するゲノム断片の整列化機能等に対応)
4. 大型ストレージ(2PBが利用可能なディスク容量、高速バックアップに対応)
5. 情報収集サーバ(外部データベースからのデータ収集)
6. 情報利用サーバ(ユーザーがログインして収集データをオンライン解析で利用)
7. ネットワーク装置(全体を10Gbpsで構成、大量情報の転送に対応)

(1) 超高速シーケンサーに対応したゲノム断片の整列化機能



「大規模ゲノムリシーケンシング」への対応を一連のパイプライン化する

レファレンスとの比較(BWAなど)もしくは新規アセンブル(Velvetなど)により、長い一続きの配列(コンティグ)を形成する。



次の「新規遺伝子予測」のステップの基礎データになる。

- 既存プログラムの速い更新についていかなければならない
- 結果をBAM/SAMのような一般的なフォーマットで整理し、SNP等を抽出

```

+BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:1245/1
paaaaaaaaa_S_baaab_[E]UaaaUEZaaZEUEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:772/1
TTGCATCCGTCGCGCTCGTGANCGCCANCAACCNATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:772/1
paaaaaaaaa_aaaaaaaaaXE[Z]aUgEUaa[E]OEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:163/1
AGGAGAGGGGGCCGGTCAAGGNAGAGGNGGGANGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:163/1
paaa^aaaaUaaaaaZaaaaREMa[L]EXEUUEMXEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:1857/1
AACAGATCAAGTCAATCTCATNGCTCANCACACTNACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:1857/1
pbaaaabbbabaababaaaY[EMa^XOE]UaaXEZ[EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:125/1
TTAAAGCGTACTTTTCTGACTNCAACTNGGTGNATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:125/1
abaaaa_V^aabaabaabaZEXaaa[ERaPUe]UEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:2018/1
ATCTTCACCTTCTCGGTGCGTNTCGGTNTGGANATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+BRITNEYSPEARS_6_FC30F6TAAXX_RD1:1:4:16:2018/1

```

EMBL-EBI
Velvet
Sequence assembler

- [Current version: 1.1.0](#)
- [Manual and extensions](#)
- [Public Git URL: git clone](#)
- [For up-to-date info, visit](#)
- [For transcriptomic applications](#)

Bowtie
An ultrafast memory-efficient short aligner

Bowtie is an ultrafast, memory-efficient aligner for the human genome at a rate of over 2 million reads per second. It uses Burrows-Wheeler index to keep its memory footprint small (2.9 GB for paired-end).

Stable Releases

ABYSS 1.2.7 (Apr 15, 2011)
Support using bwa or bowtie to align reads to contigs. New parameter, d, to specify the acceptable error of a distance estimate.
[Read more...](#)

ABYSS 1.2.6 (Feb 07, 2011)
Sequence variants are popped if the two variants are at least 90% similar. Contigs that overlap by fewer than k-1 bp are found and may be merged.
[Read more...](#)

ABYSS 1.2.5 (Nov 15, 2010)
Fix a colour-space-specific bug and a bug causing the error Assertion `fstSol.size() == 1' failed.
[Read more...](#)

ABYSS 1.2.4 (Oct 14, 2010)
Replace gaps of Ns that span a region of ambiguous sequence with a consensus sequence of the possible sequences that fill the gap. The consensus sequence uses IUPAC-IUB ambiguity codes.
[Read more...](#)

ABYSS 1.2.3 (Sep 08, 2010)
Fix two bugs that caused the error messages: Assertion `m_comm.receiveEmpty()' failed. and error: unexpected ID
[Read more...](#)

ABYSS 1.2.2 (Aug 25, 2010)
Merge contigs after popping bubbles. Handle multi-line FASTA sequences. Report the amount of memory used.
[Read more...](#)



●既存の枠組みの活用

最近、Galaxyの利用が急激に増えている。一般向けサービスの一つとして導入することを検討している。



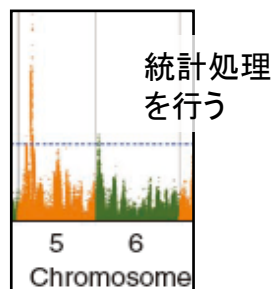
The screenshot displays the Galaxy web interface. On the left is a sidebar titled 'ツール' (Tools) with a dropdown menu 'オプション' (Options). The sidebar lists various tool categories such as 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analyses', 'Human Genome Variation', 'EMBOSS', 'NGS TOOLBOX BETA', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: SAM Tools', 'NGS: Indel Analysis', 'NGS: Peak Calling', and 'NGS: RNA Analysis'. The main content area features a dark navigation bar with links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. Below this is a large banner for 'Galaxy 2011 Community Conference' held on '25-26 May' in 'Lunteren, The Netherlands', with a 'Register now!' button. Underneath the banner is a 'Live Quickies' section with three cards: 'Basic fastQ manipulation: Galactic quickie # 13', 'Advanced fastQ manipulation: Galactic quickie # 14', and '454 Mapping: Single End: Galactic quickie # 15'. At the bottom, there is a text block about the Galaxy team and their affiliations, a 'Galaxy build: \$Rev 5353:d44244ed570a\$' string, and a Twitter link for 'galaxyproject'.

(2) ゲノム情報を活用した新規遺伝子予測機能

大きく分けて、GWASのような統計的手法で有用な遺伝子の位置を探す方法や、興味ある変異体の全ゲノムをシーケンシングによって決める方法が考えられる。

「ゲノム断片整列化」を受けて...

大量の近縁ゲノム配列
×
表現型情報



面白そうな変異体が...



ゲノムを全解読し、
レファレンスと比較

農業上重要な
遺伝子の発見
を加速する

大量データ解析
の支援

大量かつ複雑なデータを効率的に処理するシステムを研究者に提供する

今後の予定

平成23年度

平成24年度

平成25年度

平成26年度

平成27年度

データベース運用管理(維持管理、データバックアップ、故障対応、新規構築依頼の受付等)

旧システムの
データ移行

横断検索、アーカイブ機能
の設計と試用

実装、完成

効率よく情報検索できる
総合的データベースシステム
(一般研究者からの要望を反映)

ゲノム断片整列化の
調査と実装

ウェブサービス
構築

ウェブサービス提供
個別解析支援

多重アラインメント構成機能、
SNPとFNPの検出、統計処理機能の作成と実装

充実したバイオ
インフォマティクス
基盤を持って、

大量情報処理

データ発信

データ検索

を支援し、
重要遺伝子研究
を加速する
総合的システム
を構築する

今年度は...

- 新システム導入と旧システムのデータの移行
- ゲノム断片整列化機能等設計
- 作業に必要なシステムの導入

- 多様な農業上の重要種で大規模シーケンシングが想定されることから、解析手法を具体的に適用していく事が重要であると考えている。
- イネやダイズのように、既に大規模シーケンシングのデータが公開されているものもあるので、実データで検証しつつシステムを組み上げる。
- 作物、果樹、家畜、昆虫等でそれぞれ新規の配列情報増加が見込まれるので、対応していきたい。
- 省庁間連携によって、アーカイブ作成や横断検索を実現する。