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

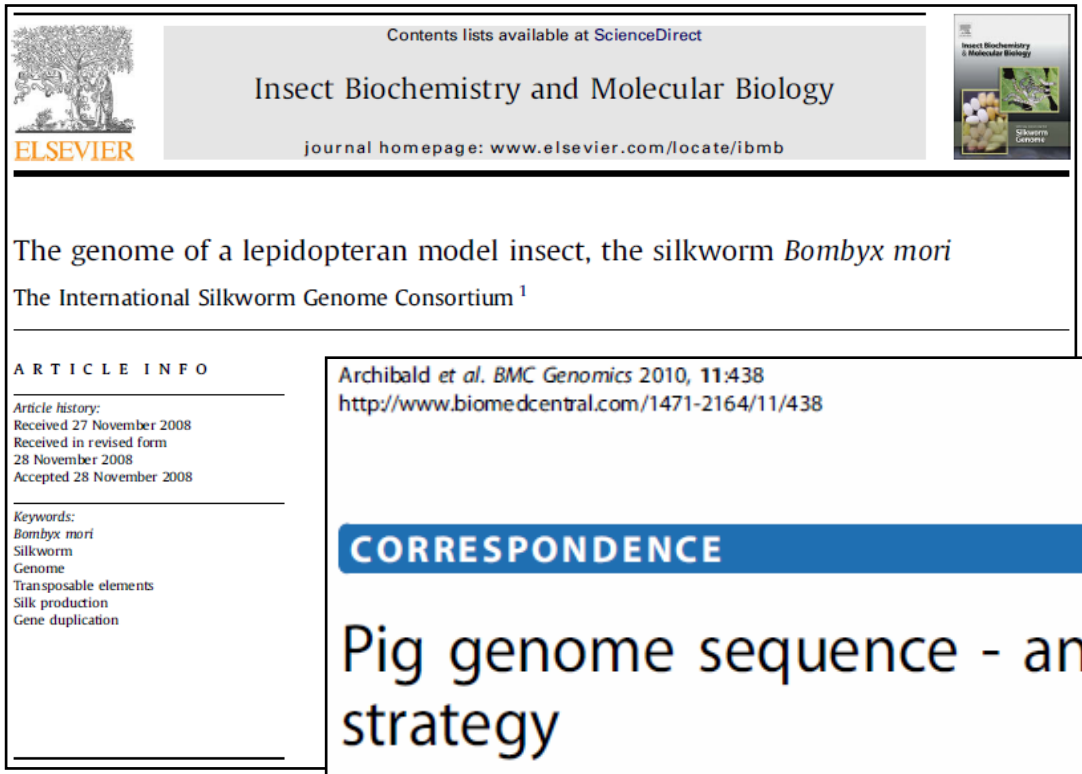
- 農業生物資源研究所は植物ゲノム研究、昆虫ゲノム研究、家畜ゲノム研究で成果を上げ、関連データベースを整備してきた。

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

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

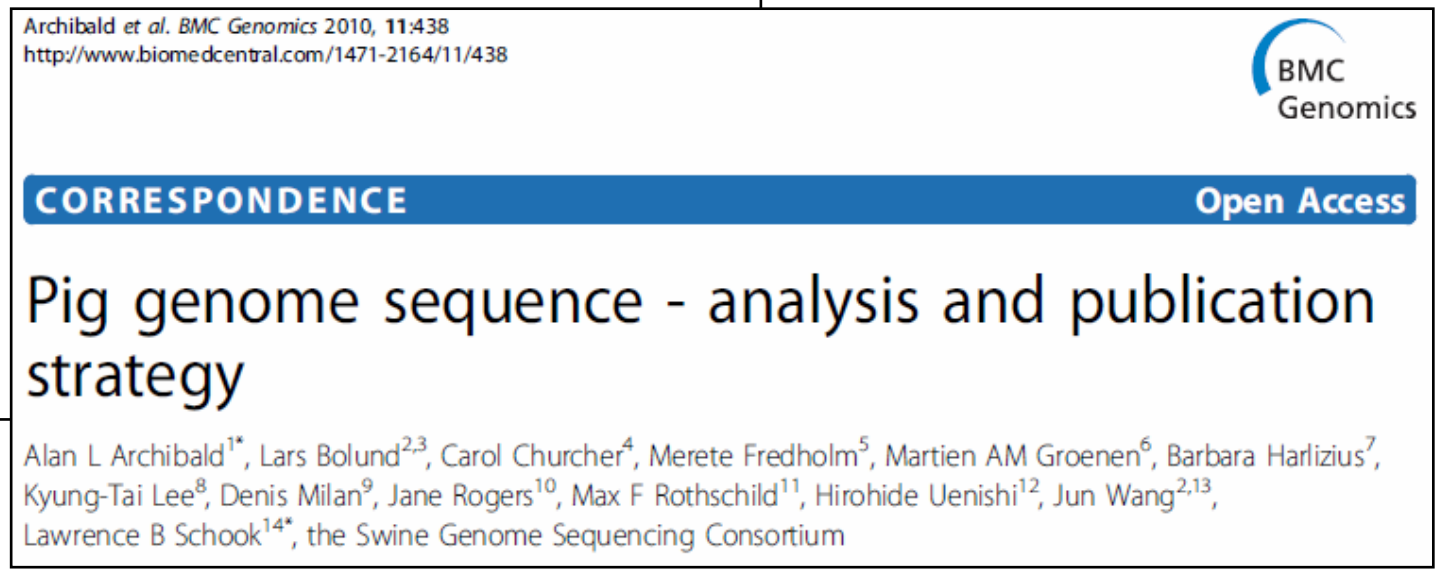
The genome of a lepidopteran model insect, the silkworm *Bombyx mori*

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

BMC Genomics

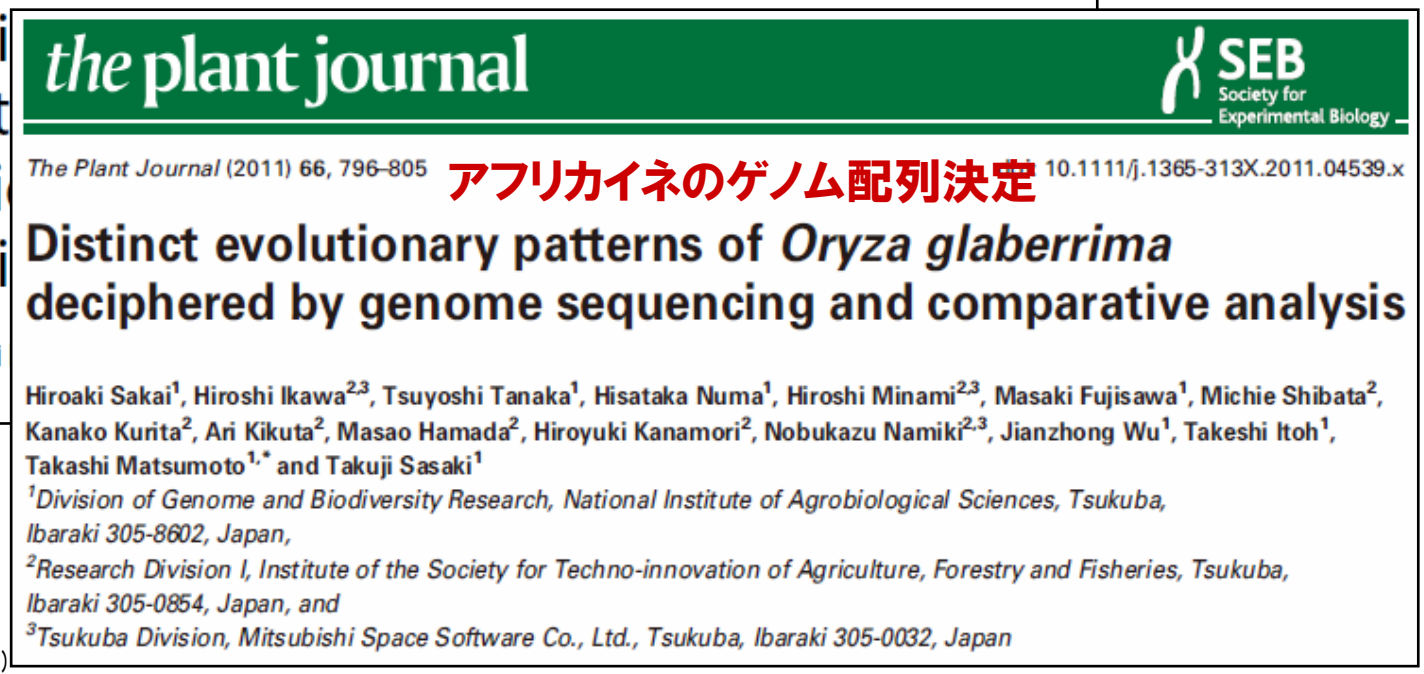
CORRESPONDENCE Open Access

Pig genome sequence - analysis and publication strategy

Alan L Archibald[1*], Lars Bolund[2,3], Carol Churcher[4], Merete Fredholm[5], Martien AM Groenen[6], Barbara Harlizius[7], Kyung-Tai Lee[8], Denis Milan[9], Jane Rogers[10], Max F Rothschild[11], Hirohide Uenishi[12], 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。



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 [W][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[1][W][OA]

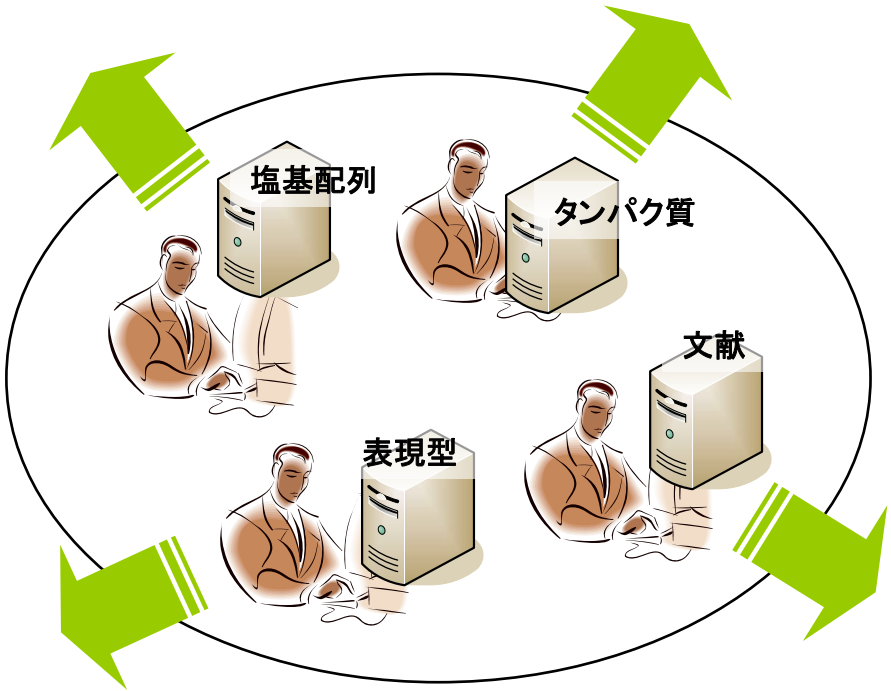
オオムギの完全長cDNA配列決定

Takashi Matsumoto[2,*], Tsuyoshi Tanaka[2], 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。

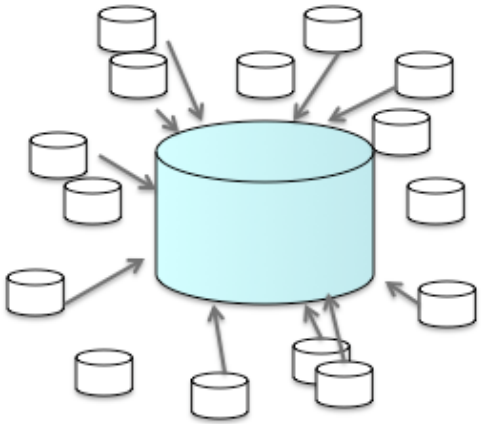


バラバラに作られ、発信されていたデータベース

農林水産生物ゲノム情報統合データベース

本計画の全体像

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

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、そこからマーカーを整備したり、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

下記２機関で研究グループを組織する

【中核機関】

独立行政法人
農業生物資源研究所

【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】

独立行政法人 農業・食品産業技術総合研究機構
作物研究所

・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

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

- 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

- 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

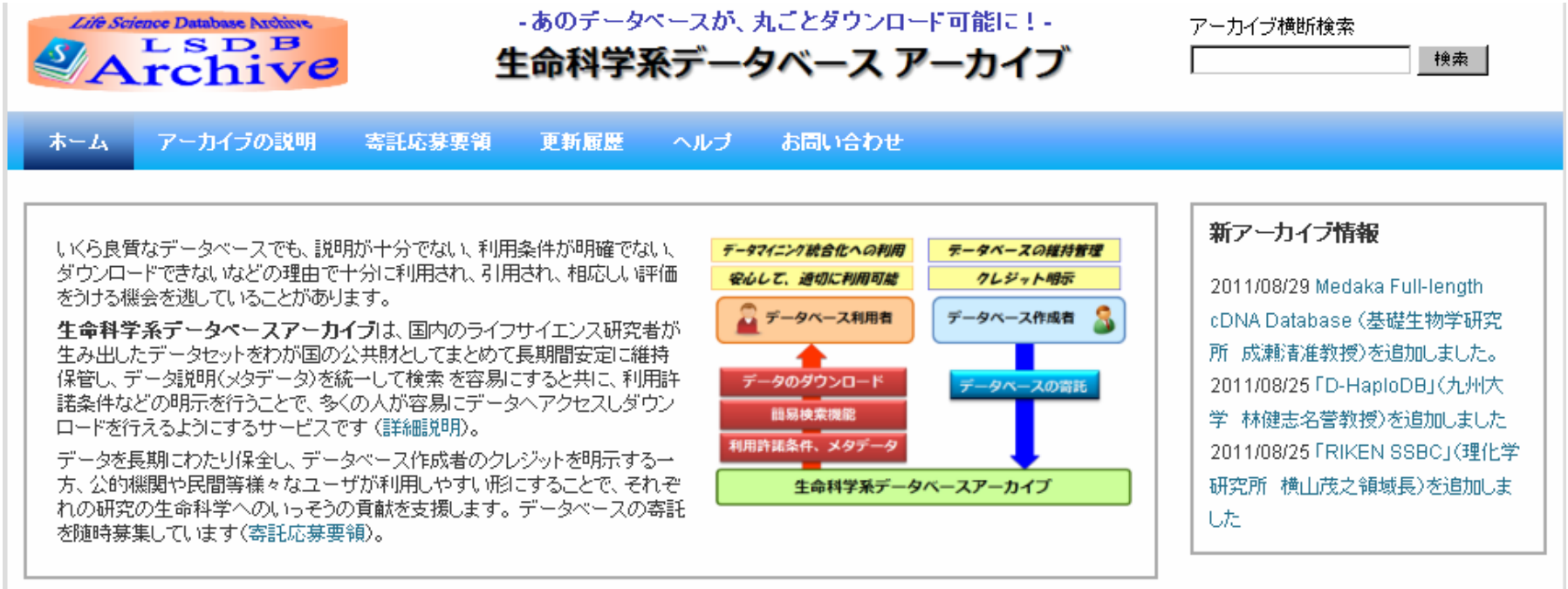
新規データやデータベースは次々と個別に作られる
- 開発はしたがプロジェクトが終了した
- 適当なハードが維持できない
- データはあるがどうしたらよいか分からない

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

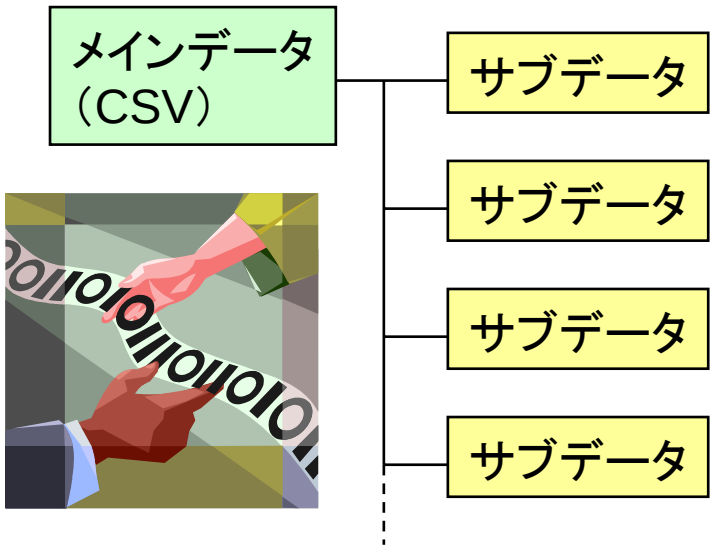
生命科学データベースアーカイブ構築機能の開発

NBDC/DBCLSとの連携～その1



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。

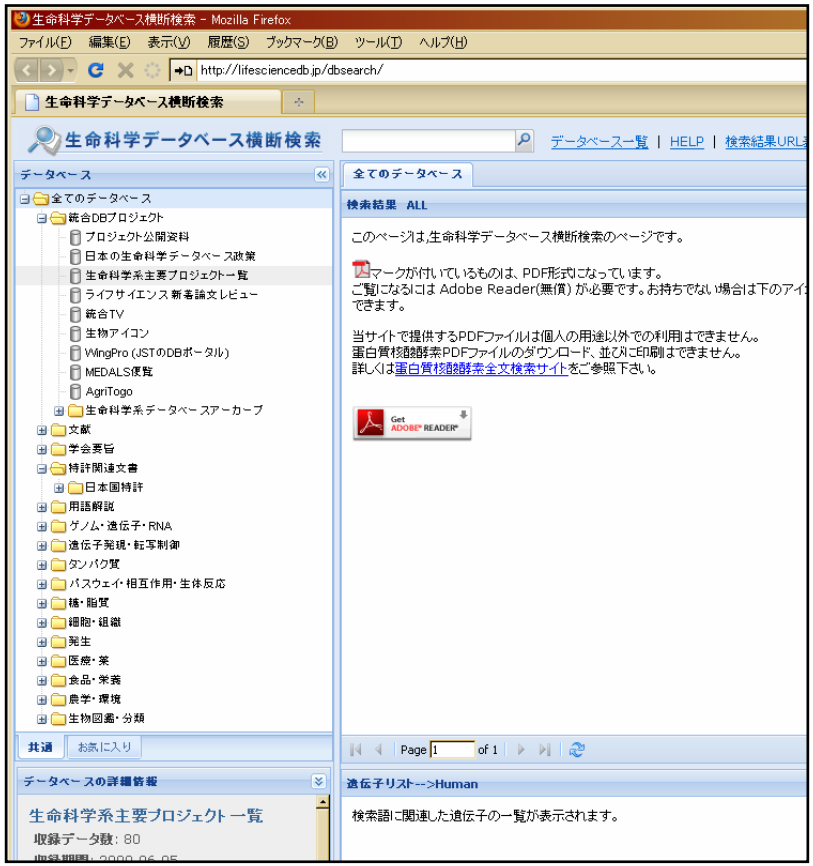
このため、各データベース内容の精査し、既に候補を幾つか選定。変換プログラム作成。

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



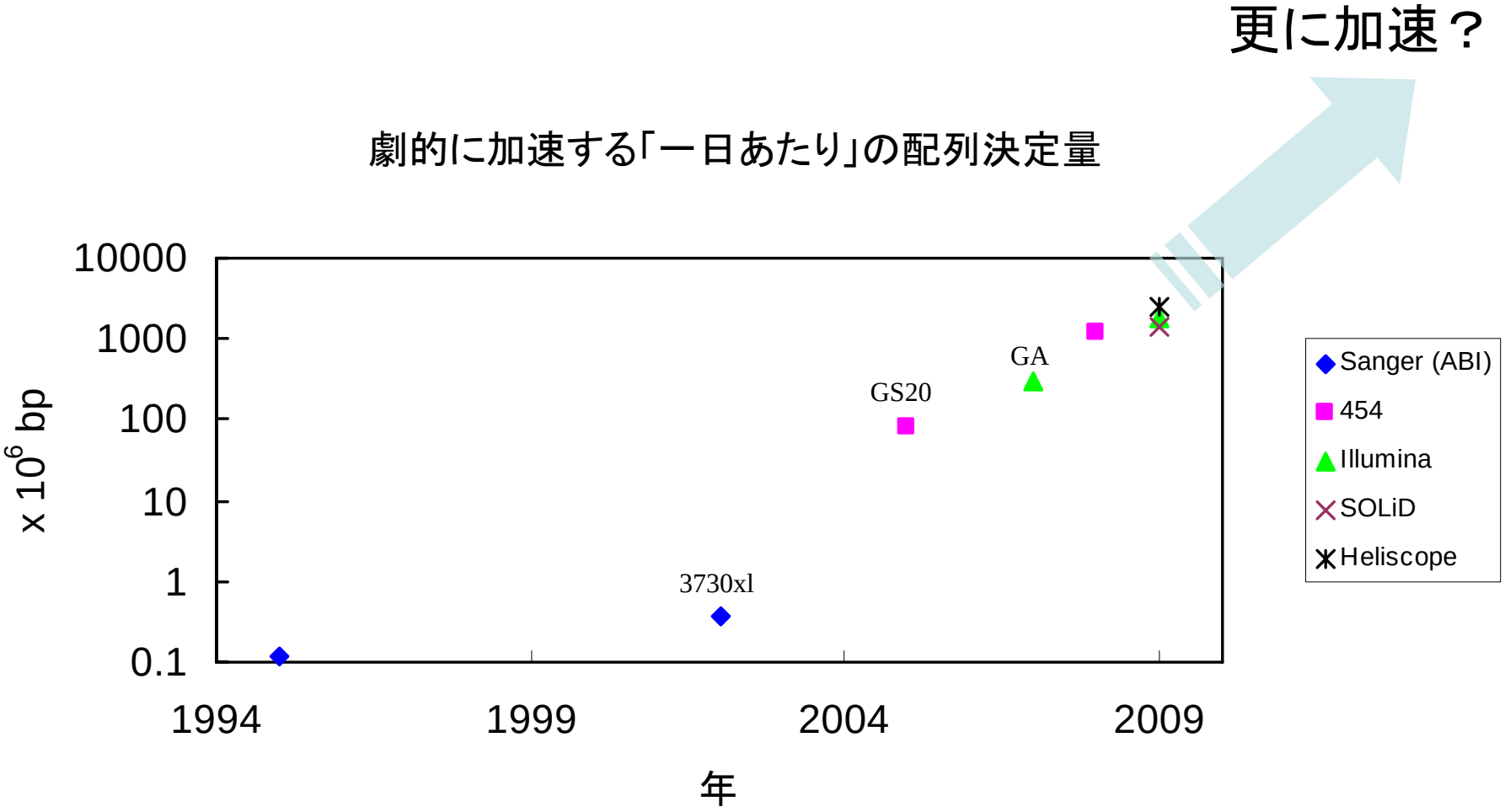
NBDC/DBCLSと連携し、横断検索機能を開発する。

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？



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.

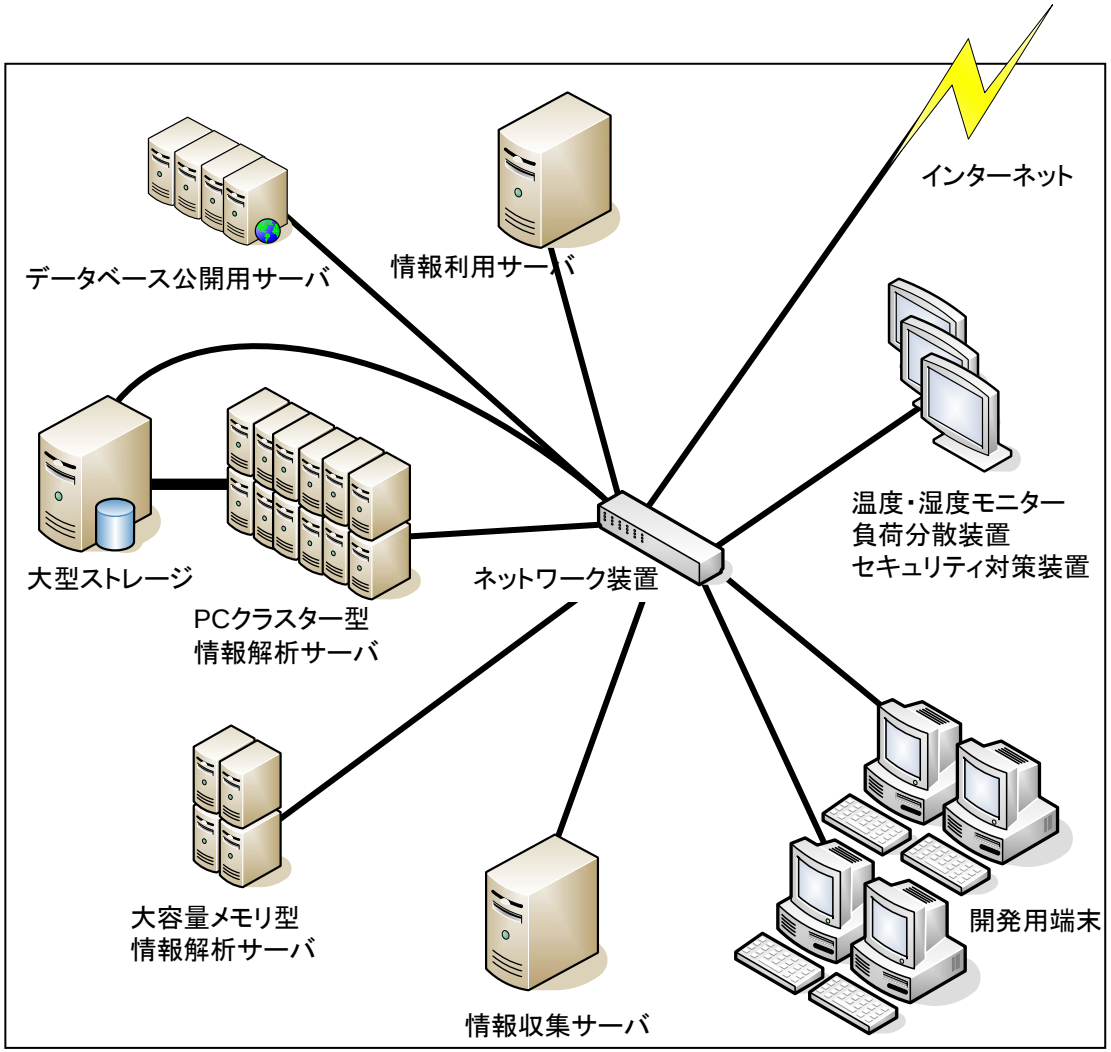
Any population this large is certain to have profound influences on its environment. Yet our knowledge of these communities and their functions is rudimentary, partly owing to our inability to culture the vast majority of

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, gene

…（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）

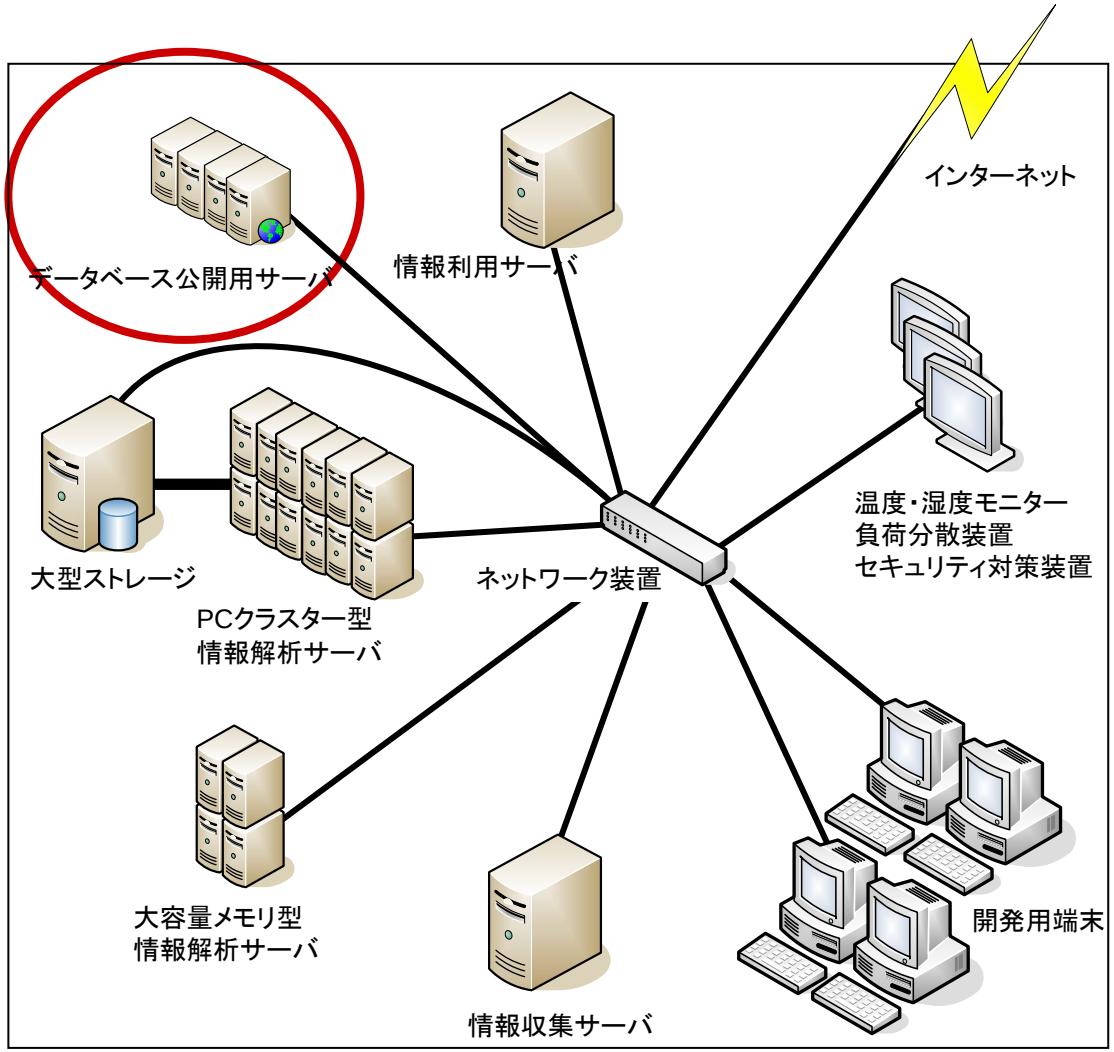
4．大型ストレージ（2PBが利用可能なディスク容量、高速バックアップに対応）

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）

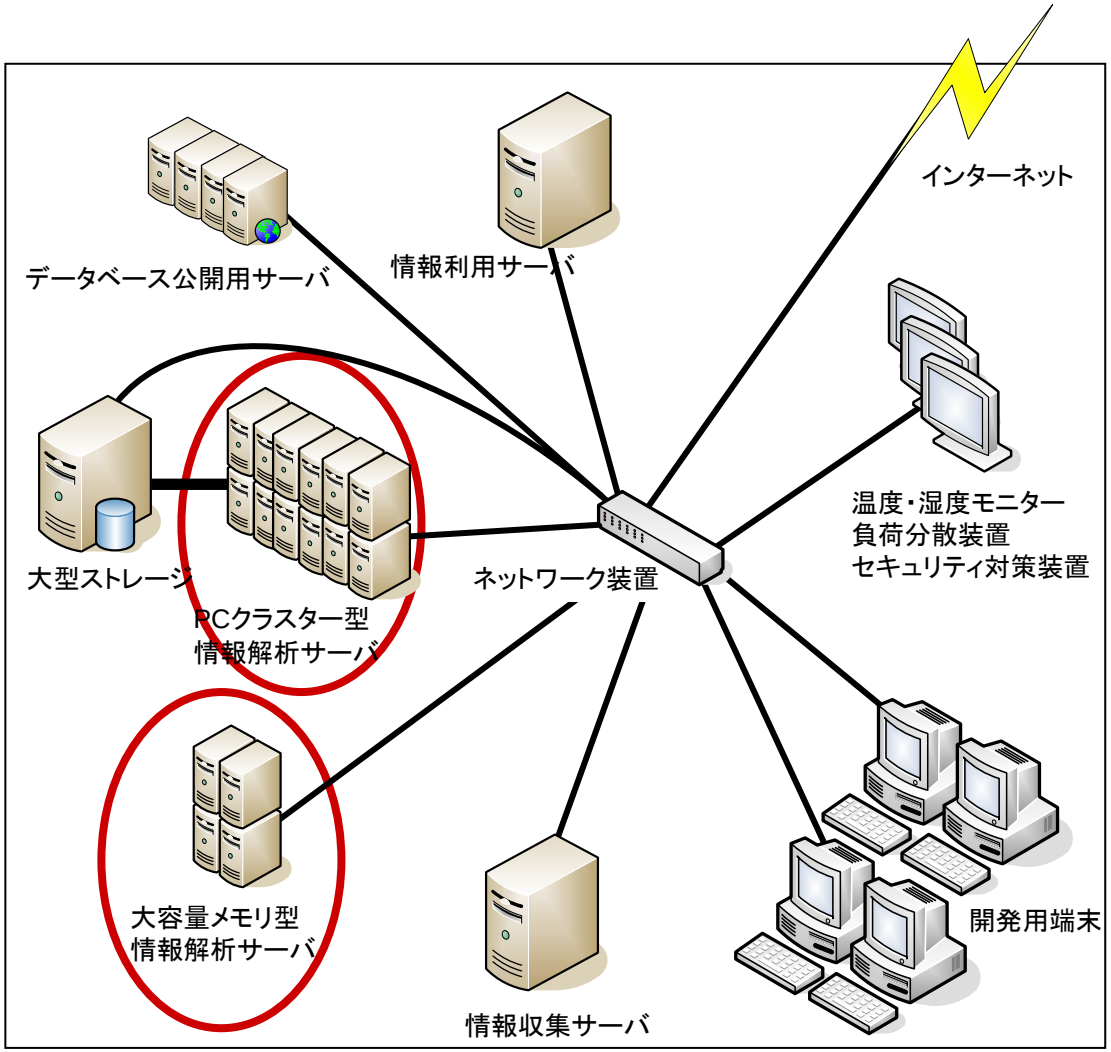
7．ネットワーク装置（全体を10Gbpsで構成、大量情報の転送に対応）

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）

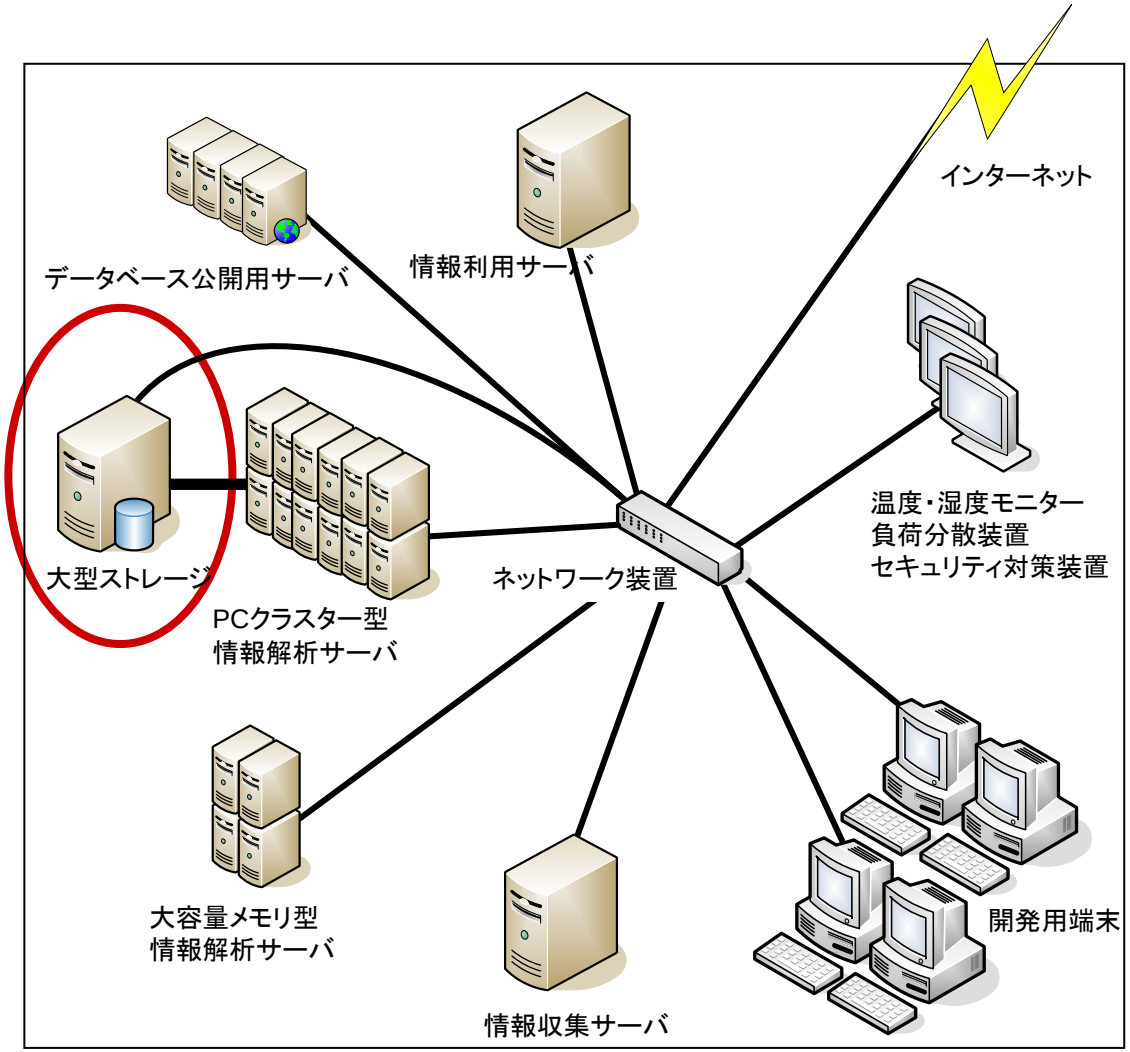
4．大型ストレージ（2PBが利用可能なディスク容量、高速バックアップに対応）

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）

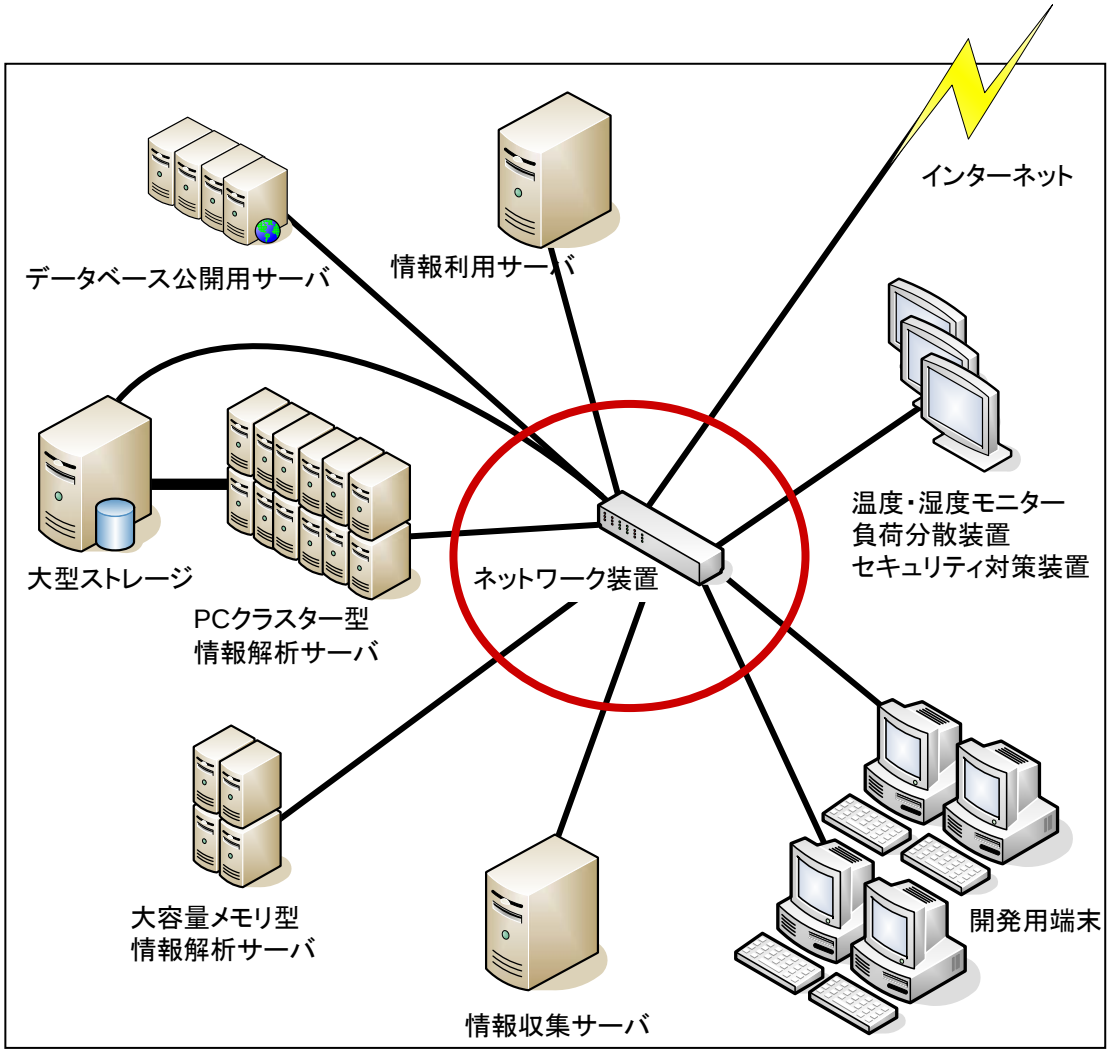
4．大型ストレージ（2PBが利用可能なディスク容量、高速バックアップに対応）

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）

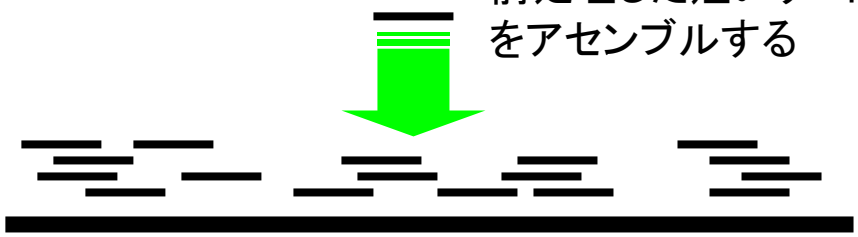
5．情報収集サーバ（外部データベースからのデータ収集）

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

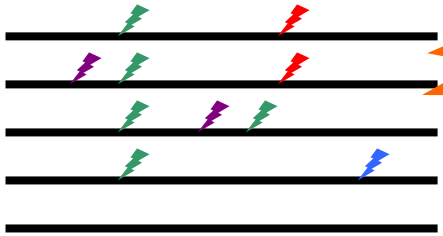
前処理した短いリード（DNA断片）
をアセンブルする



「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。

研究者がウェブ上で
簡単に操作できる
仕組みを作成

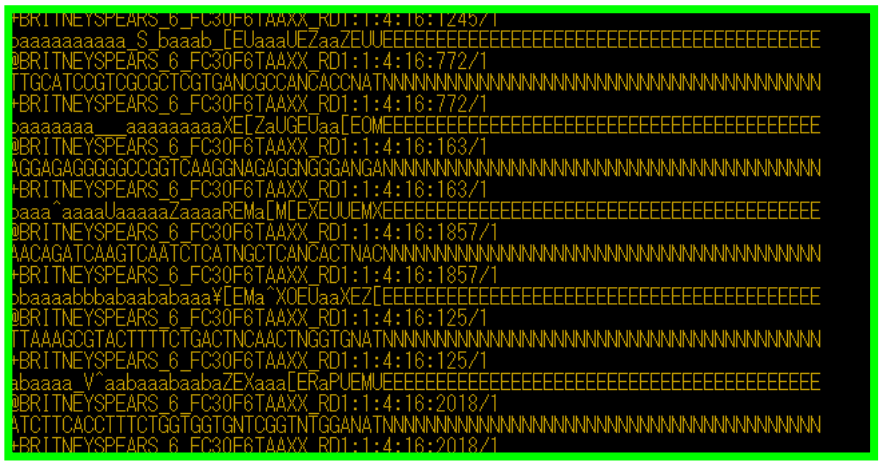


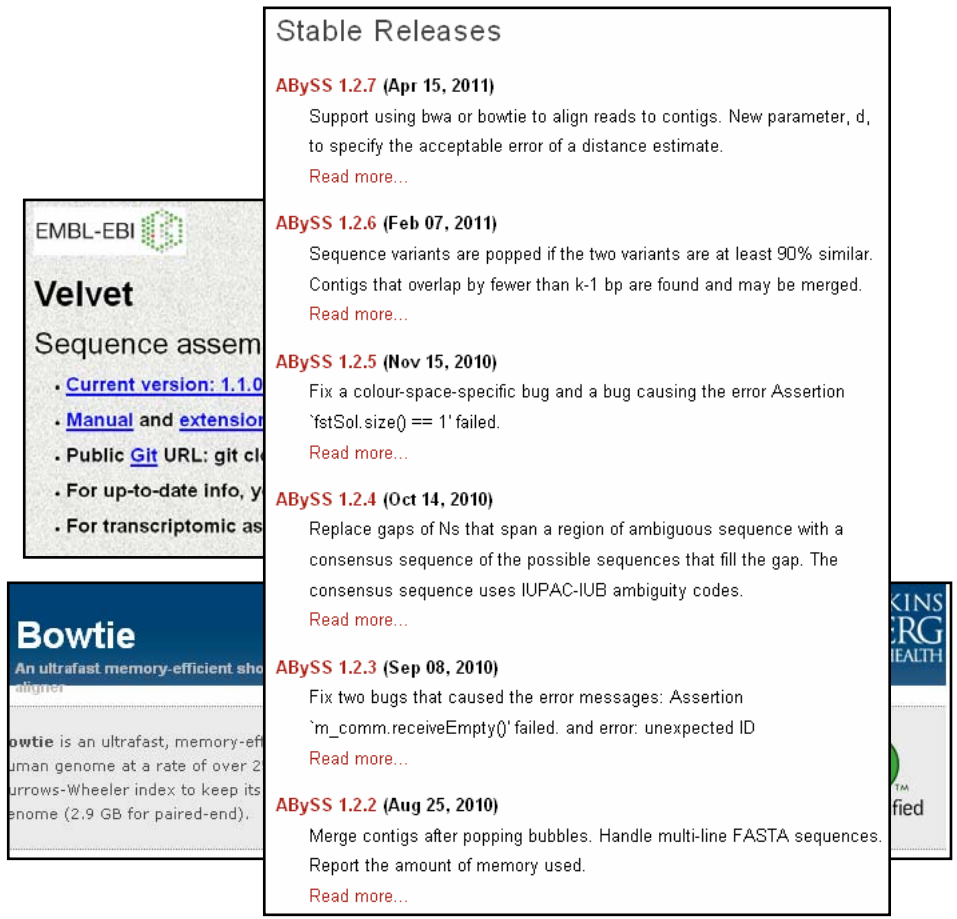
DNAマーカー
を作成できる

全ゲノム的（多重）アラインメントを作成し、全ての違いを明らかにする。

次の「新規遺伝子予測」のステップの基礎データになる。

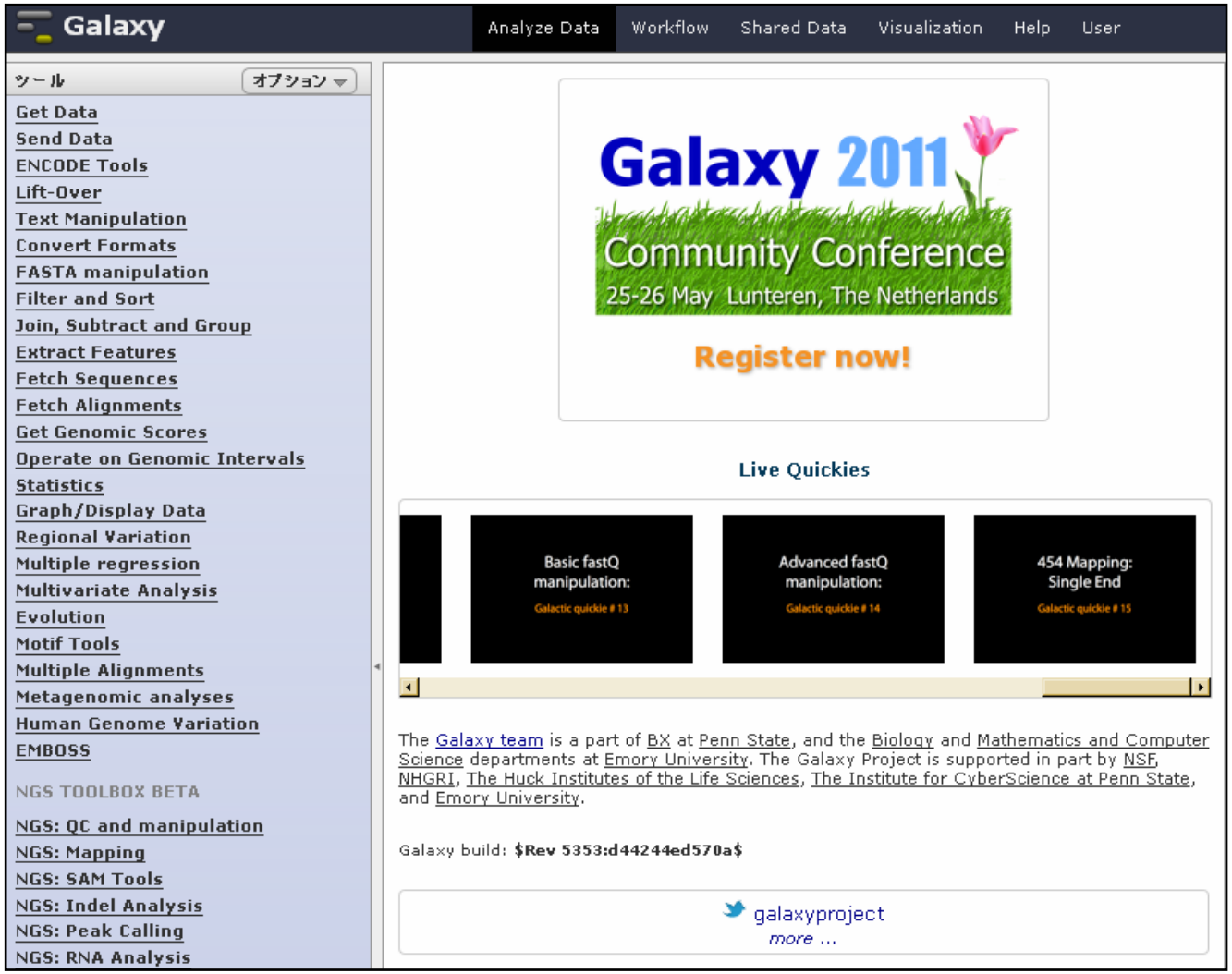
- 既存プログラムの速い更新についていかなければならない
- 結果をBAM/SAMのような一般的なフォーマットで整理し、SNP等を抽出





●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。

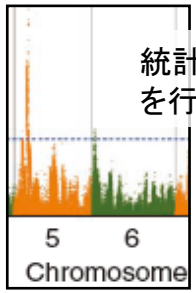


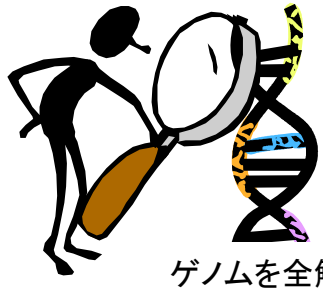
(2) ゲノム情報を活用した新規遺伝子予測機能

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



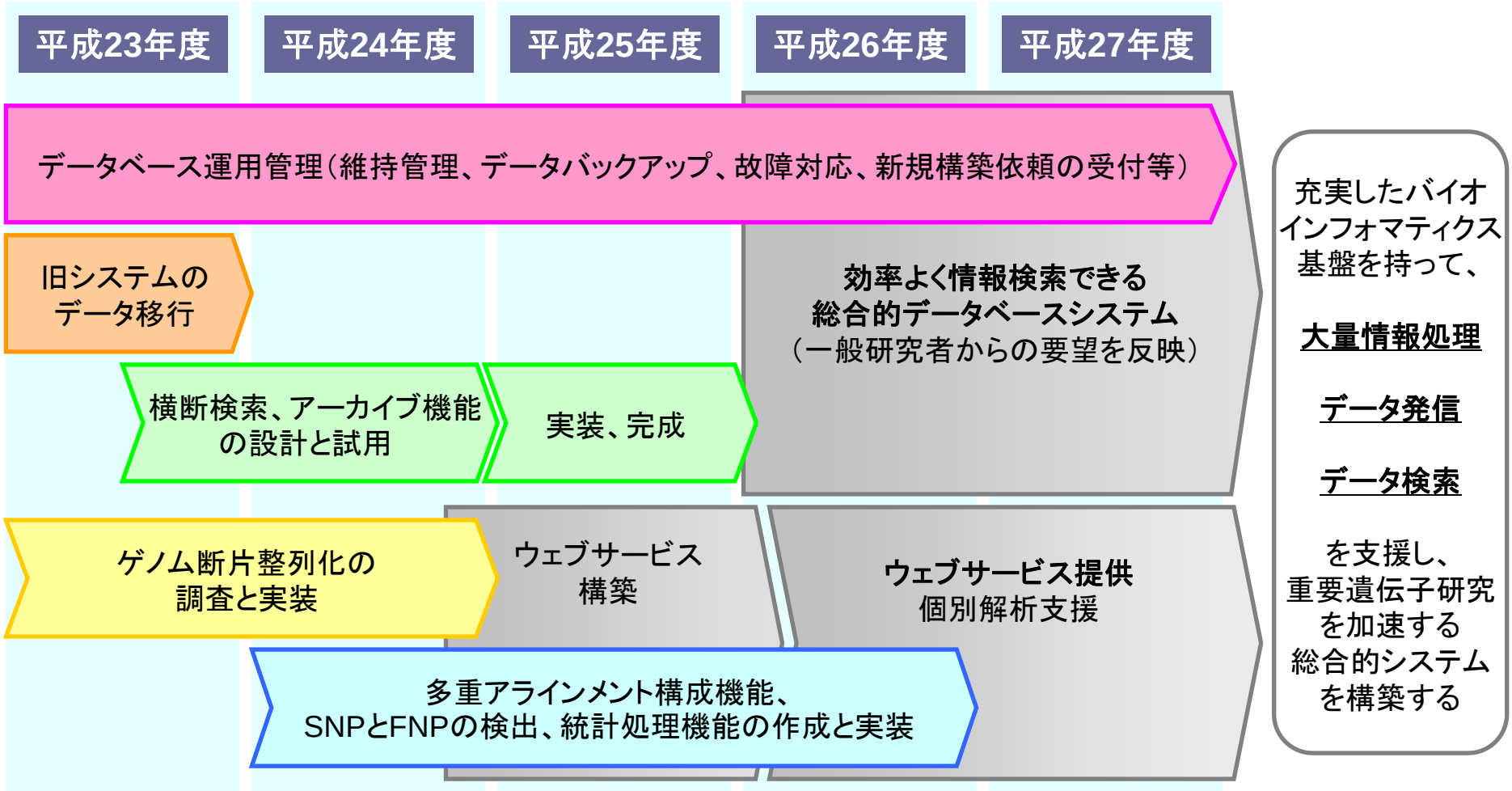


農業上重要な遺伝子の発見を加速する

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



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。
- イネやダイズの様に、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。