

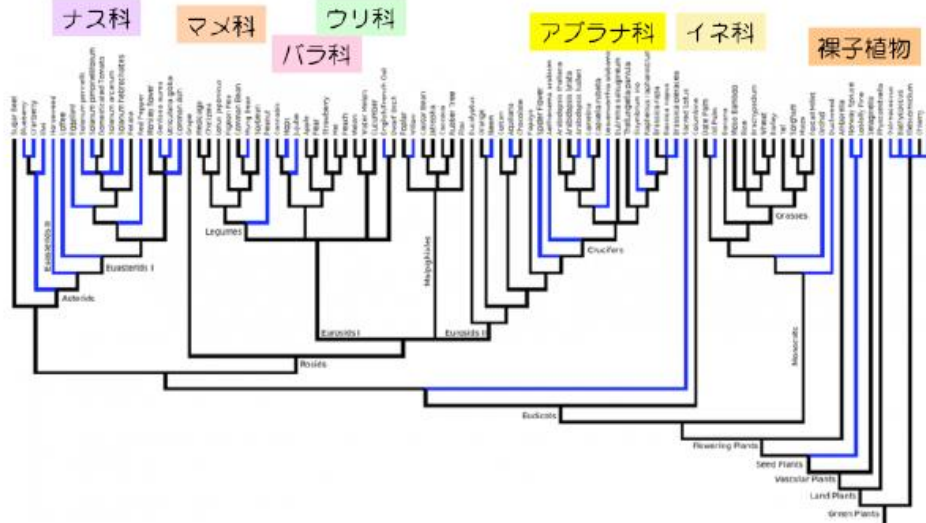
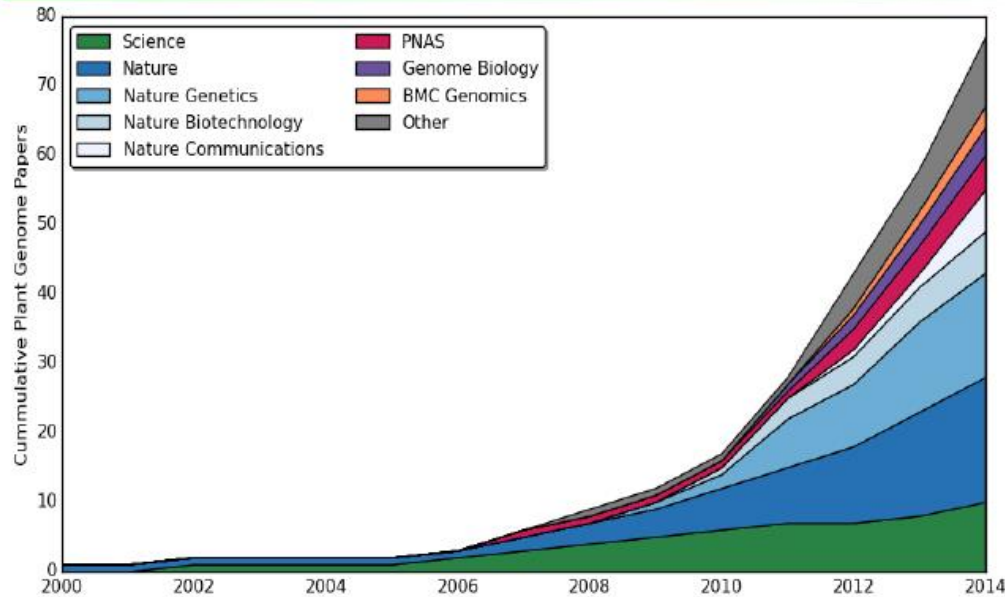


個体ゲノム時代に向けた植物ゲノム情報解析基盤の構築

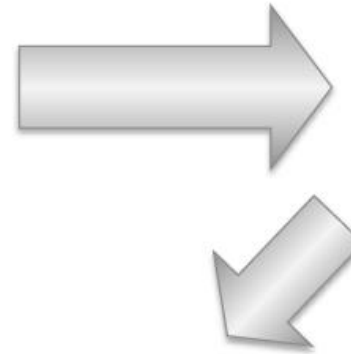
かずさDNA研究所
田畑哲之
大阪大学大学院医学系研究科
中谷明弘



研究開発の背景

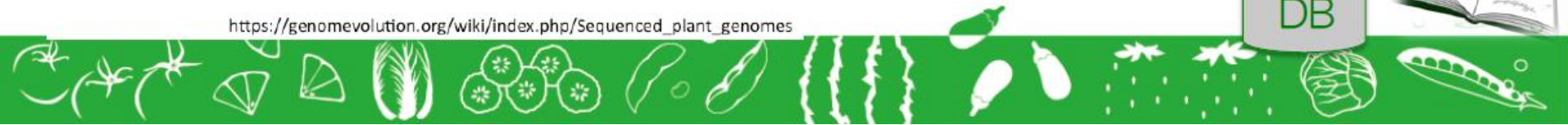


https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes



ゲノム構造
転写産物
代謝産物
変異体
DNAマーカー
連鎖地図

遺伝子機能
ネットワーク
有用遺伝子
QTL解析
品種改良



平成23年度から28年度の統合化推進プログラムにおいて統合化のハブとして構築したポータルサイト

Global Engageのサイトで
The Top 10 Plant Genome Databases
の一つとして紹介されている。

リソース

- 植物リソース横断検索システム
- NBRPの提供する植物リソース (cDNAクローン)を検索(横断検索)
- 追加されたリソース
- DB(カンキツリソース)
- 植物リソースについて

オルソログDB

- オルソログテーブル
遺伝子のアミノ酸配列の階層的な相互リンクテーブル
- キーワード検索
オルソログDBの関連データをキーワード検索
- BLAST検索
オルソログDBの関連データをアミノ酸配列で相同性検索
- データダウンロード
- オルソログDBについて

PGDB
Plant Genome DataBase Japan

例: "Pyruvic acid"

多種多様な
植物ゲノム関連情報

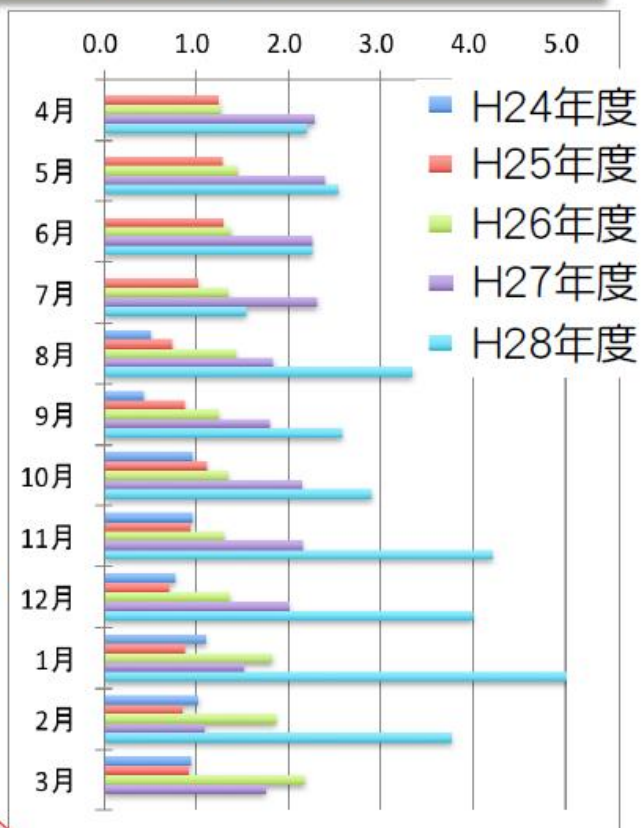
様々な目的をもった
幅広いユーザー

- 植物データベースリンク
- ゲノム解読状況
- メニュー
- ・ QTL ・ 植物種基本情報について

DNAマーカー他



育種向けDNAマーカー検索



訪問者数 (ユニークIP数)
($\times 10^3$ 件)

総数 > 94,500



本提案の背景

次世代型シーケンサー（NGS）の普及

新たな植物種で全ゲノム情報が次々と公開

品種間や自然変異がもつゲノムワイドな
多型情報を迅速安価に収集

異種間でのゲノム関連情報を比較

増え続けるゲノム関連情報の統合

品種、系統などの個体ゲノム解析
による多型・ハプロタイプ検出



(1) ゲノム横断的関連情報表示サイトの開発

JBrowseによるゲノムビューアをプラットフォームとして、個々の生物種のゲノム塩基配列に対し、互いのゲノム配列情報、遺伝子、転写産物、多型、DNAマーカー等さまざまな情報を合わせて閲覧、比較できるサイトを構築する。

(2) 種を超えたゲノム情報統合のためのデータリンク基盤の構築

上記サイトの基礎となる遺伝子間の配列類似性を大規模に探索するシステムを新たに開発する。

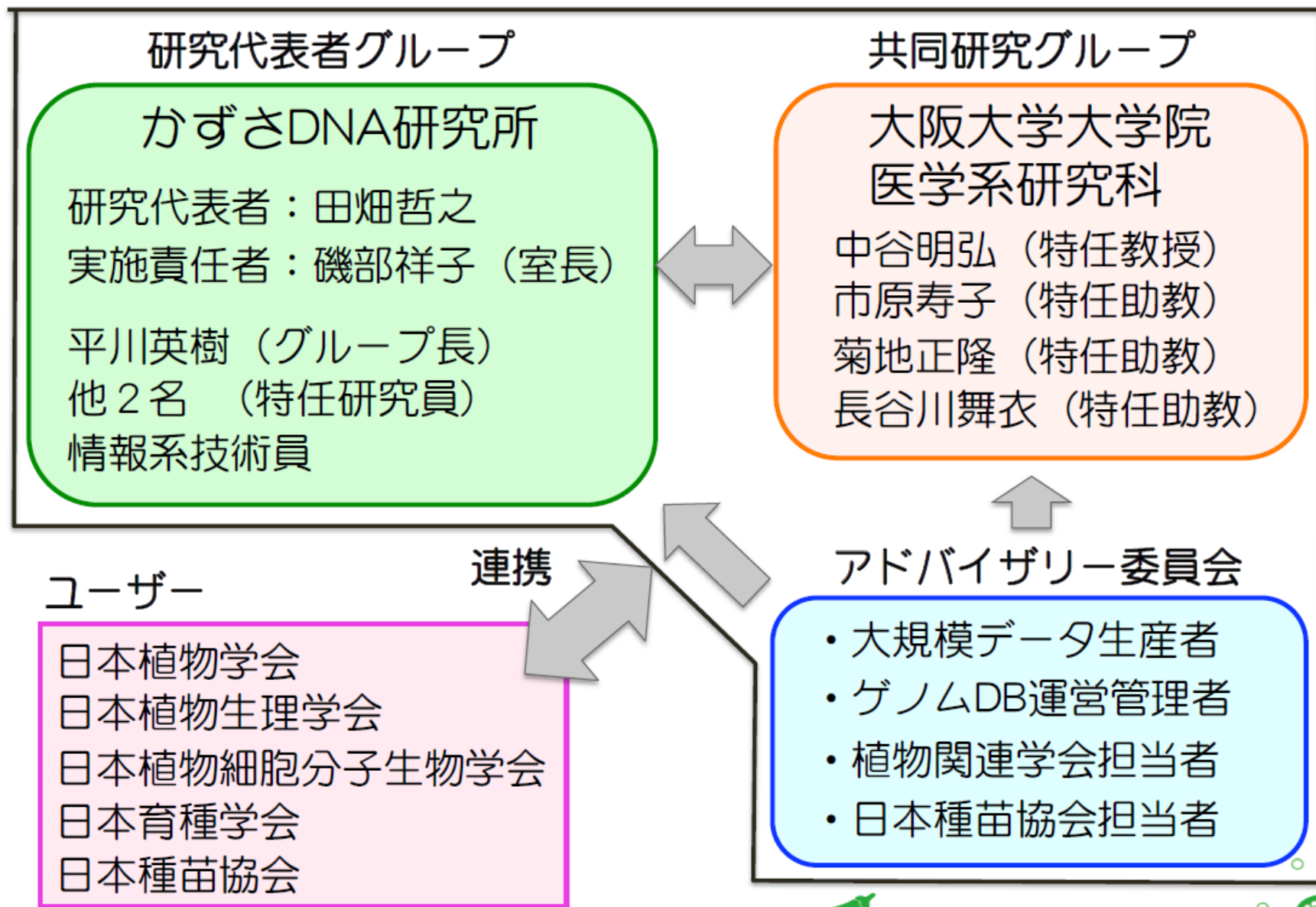
(3) カスタム型多型・ハプロタイプ検出システムの構築

ユーザが独自で取得したNGSデータをシステムに投入し、多型 (SNP、InDel、CNV、SV) やLD、ハプロタイプを検出するシステムを構築する。

(4) PGDBjコンテンツの更新

これまで対象とした植物種を中心にDNAマーカーやバイオリソース情報に加え、GWASやRAD-Seq、QTL-Seq等のゲノムワイドな多型に関するデータマニュアルキュレーションにより収集し、整理、公開する。





大林 武	東北大・院・情報科学研究科	<日本植物生理学会推薦>
川本祥子	国立遺伝研・系統情報研究室	
庄司 翼	奈良先端大・バイオサイエンス研究科	<日本植物細胞分子生物学会推薦>
鈴木隆夫	(株)サカタのタネ・掛川総合研究センター	<日本種苗協会推薦>
田中 剛	農研機構次世代作物開発研究センター・情報解析ユニット	
永野 淳	龍谷大・農学部	<日本植物学会推薦>
中村保一	国立遺伝研・大量遺伝情報研究室	
長谷部光泰	基生研・生物進化研究部門	
深海 薫	理研・バイオリソースセンター、「統合化推進プログラム」研究アドバイザー	
松井 南	理研・環境資源科学研究センター	
矢野健太郎	明治大・農学部・生命科学科	<日本育種学会推薦>
(敬称略)		

年2回開催

議題：事業計画や事業成果の報告、学会、DB管理者、データ生産者からの
要望聴取、技術的な情報提供、ユーザ獲得や現場での利用、その他
事業全般に関する情報交換や議論



- 日本植物学会、日本植物生理学会、日本細胞分子生物学会、日本育種学会との連携



年会等での発表、展示、講習会開催
会員の要望聴取

- 日本種苗協会との連携



DNAマーカー関連の講習会等で紹介
育種担当者の要望聴取

- アドバイザリー委員会



学協会、DB管理者、データ生産者への情報提供
要望聴取



各研究項目

今季からのPGDBjの構成について

かずさDNA研究所・磯部

(1) ゲノム横断的関連情報表示サイトの開発

かずさDNA研究所・平川

(2) 種を超えたゲノム情報統合のためのデータリンク基盤の構築

大阪大学・中谷

(3) カスタム型多型・ハプロタイプ検出システムの構築

かずさDNA研究所・磯部

(4) PGDBjコンテンツの更新

かずさDNA研究所・平川

今季からのPGDBjの構成について

かずさDNA研究所・磯部祥子

現在のコンテンツ

Version 1として当面運用

<http://pgdbj.jp/>



・ NBRPリソース情報

遺伝研, 理研BRC, 14 生物種 (約150万件)

・ イチゴ、カーネーション他

(かずさDNA研) 6 生物種 (約85万件)

・ カンキツDB (近畿大、果樹研)

約900種分 + INSDC他配列・形質情報
(約66万件)

・ 65植物種のゲノム関連情報

・ 文献約1,200報より収集した

DNAマーカー (約26万件)

QTL (約16,500件)

今季から構築するDB

Ver2.0として構造を変える



データを大幅追加
DNA研（DB構築）
阪大（類似性検索法）

各植物種のDBを前面に
だす

前期までのコンテンツ
カスタム型Variant検出シ
ステム

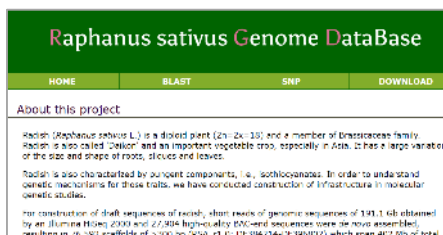
<育種向け> 分けなくても
使いやすいユニバーサ
ルデザイン
高度な検索は<Genome>
と統合

Resource
=物理的に配布するコ
ンテンツとする

<Omicsデータ> は
Genome/Geneへ

Genome/Gene/Variant

「(1)ゲノム横断的関連情報表示サイトの開発」



統合DB (PGDBj)

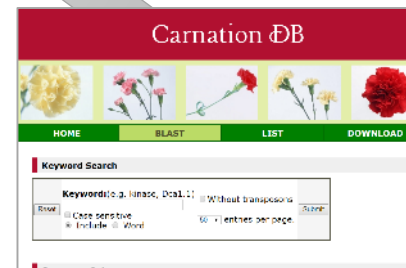
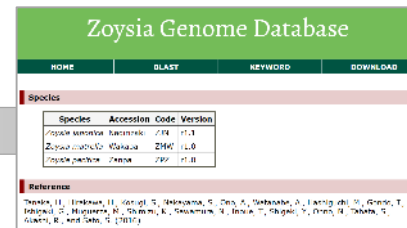
独立したDB間のデータをリンク

構造が類似しているDB
(かずさDNA研究所で構築)
で基本構造を検討

ゲノム配列、遺伝子、DNAマーカー
RNA-Seq
配列多型・ハプロタイプ
(NCBIのSRAから独自に検出)

リンクの整備

「(2)種を超えたゲノム情報統合のためのデータリンク基盤の構築」



Genome/Gene/Variant

統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみで
DBが構築されていないデータ

新規に公開されたデータ

統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみで
DBが構築されていないデータ

新規に公開されたデータ

統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみで
DBが構築されていないデータ

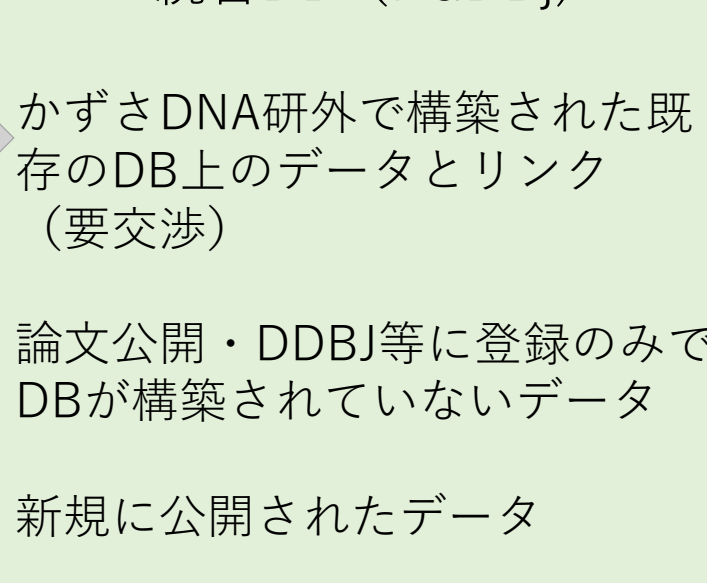
新規に公開されたデータ

統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみで
DBが構築されていないデータ

新規に公開されたデータ

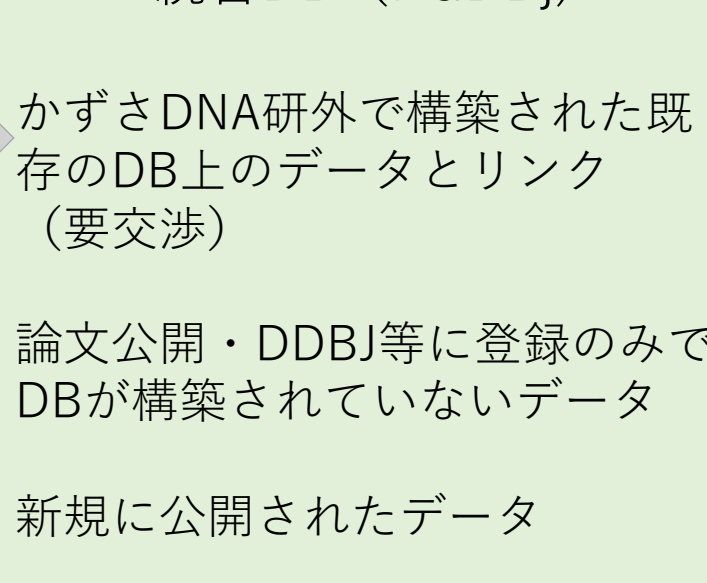


統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみでDBが構築されていないデータ

新規に公開されたデータ

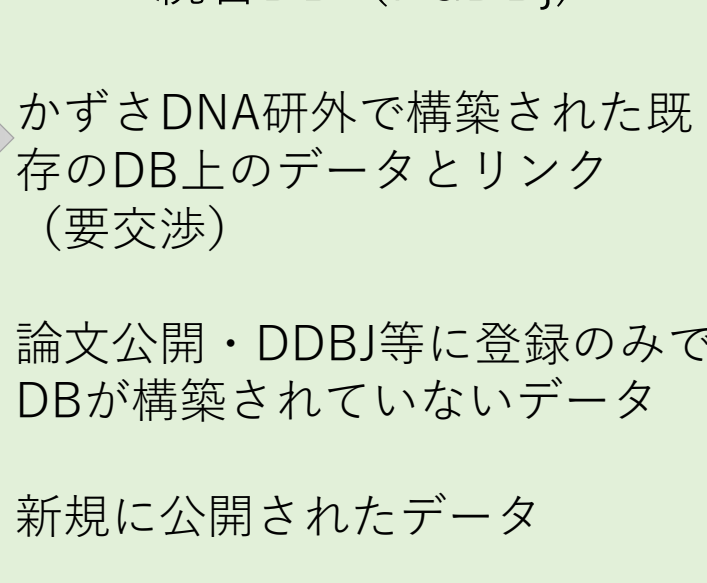


統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみでDBが構築されていないデータ

新規に公開されたデータ



統合DB (PGDBj)

かずさDNA研外で構築された既存のDB上のデータとリンク
(要交渉)

論文公開・DDBJ等に登録のみでDBが構築されていないデータ

新規に公開されたデータ

[illegible]



Sato Chives

@Chives (1934 ~ 1975)

既存のDB

STRAWBERRY GARDEN

Strawberry Genome And Variation Database (SNGVDB)

HOME

BLAST

SEQUENCING

MARKER

DOWNLOAD



Keyword Search

#AND #OR #NOT (use space as delimiter)

Database:

Reference sequences

Strawberry sequences

Strawberry annotations (JANU1.1)

[illegible]



miyakogusa.jp 3.0

Index

About Us

Contact us

Copyright

FAQ

Data download

Lotos japonicus News

Lotos japonicus genome assembly build 3.0

Lotos ERM Research Institute is pleased to release Lotos japonicus genome

- Release of build 3.0
- Release of build 3.0 (pre-release)

Lotos japonicus predicted gene keyword search

Carnation DB







[HOME](#)
[BLAST](#)
[LIST](#)
[DOWNLOAD](#)

Keyword Search

Keywords: Name, Deal:

☐ Case sensitive
 ☐ Include ☐ Word

 60 entries per page

(1) ゲノム横断的関連情報表示サイトの開発

かずさDNA研究所・平川英樹

(1) ゲノム横断的関連情報表示サイトの開発

植物関連情報へのアクセスを容易にして幅広いユーザの利便性を高める

PGDB j (Plant Genome DataBase Japan; <http://pgdbj.jp>)



これまでにキュレーションした内容

- ・ DNAマーカー (65種、約26万件)
- ・ QTL (45種、約1万4千件)

- ・ これまでにキュレーションしたDNAマーカー、QTL情報のゲノムブラウザ上への集約
- ・ 今期に収集する情報のゲノムブラウザ上への集約
ゲノムワイドな多型データ (GWAS、RAD-Seq、QTL-Seqなど)、LD、原因遺伝子
- ・ 新規植物や変異体の全ゲノム構造情報を表示するシステムの構築
- ・ SRAデータのキュレーション、多型検出を行い、ユーザが利用しやすい形で提供

対象植物種

かずさDNA研究所で解読した11植物種

科名	学術名	和名	ゲノム配列
アブラナ科	<i>Raphanus sativus</i>	ダイコン	Scaffold
マメ科	<i>Lotus japonicus</i>	ミヤコグサ	Pseudomolecule
	<i>Trifolium subterraneum</i>	サブリローバ	Pseudomolecule
イネ科	<i>Zoysia japonicus</i>	シバ	Pseudomolecule
バラ科	<i>Fragaria × ananassa</i>	イチゴ	Scaffold
ナス科	<i>Solanum melongena</i>	ナス	Scaffold
ヒルガオ科	<i>Ipomoea trifida</i>	サツマイモ野生種	Scaffold
ナデシコ科	<i>Dianthus caryophyllus</i>	カーネーション	Scaffold
タデ科	<i>Fagopyrum esculentum</i>	ソバ	Scaffold
フトモモ科	<i>Eucalyptus camaldulensis</i>	ユーカリ	Scaffold
トウダイグサ科	<i>Jatropha curcas</i>	ジャトロファ	Scaffold

植物種は適宜追加する

ニンジン
ペチュニア
ヒマワリ
アズキ
キヌア
アサガオ

イチジク
マスタード
レタス
オオムギ
チャ
ハウレンソウ

第二期までに対象とした植物種のうち
Pseudomoleculeとして解読された27植物種

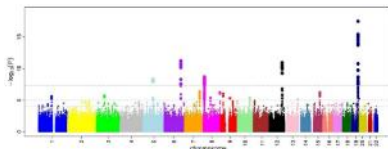
科名	学術名	和名
マタタビ科	<i>Actinidia chinensis</i>	キウイフルーツ
アブラナ科	<i>Arabidopsis lyrata</i>	ミヤマハタザオ
	<i>Brassica napus</i>	セイヨウアブラナ
	<i>Brassica oleracea</i>	キャベツ
	<i>Brassica rapa</i>	ハクサイ
	<i>Capsella rubella</i>	ルペラナズナ
ウリ科	<i>Citrullus lanatus</i>	スイカ
	<i>Cucumis melo</i>	メロン
	<i>Cucumis sativus</i>	キュウリ
マメ科	<i>Cicer arietinum</i>	ヒヨコマメ
	<i>Glycine max</i>	ダイズ
	<i>Medicago truncatula</i>	タルウマゴヤシ
	<i>Trifolium pratense</i>	アカクローバ
	<i>Theobroma cacao</i>	カカオ
バショウ科	<i>Musa acuminata</i>	バナナ
ゴマ科	<i>Sesamum indicum</i>	ゴマ
イネ科	<i>Brachypodium distachyon</i>	セイヨウヤマカモジ
	<i>Setaria italica</i>	アワ
バラ科	<i>Fragaria vesca</i>	エゾヘビイチゴ
	<i>Malus × domestica</i>	リンゴ
	<i>Prunus persica</i>	モモ
	<i>Citrus sinensis</i>	オレンジ
	<i>Populus trichocarpa</i>	ポプラ
ナス科	<i>Capsicum annuum</i>	トウガラシ
	<i>Solanum lycopersicum</i>	トマト
	<i>Solanum tuberosum</i>	ジャガイモ
ブドウ科	<i>Vitis vinifera</i>	ヨーロッパブドウ

ゲノムブラウザに掲載する情報

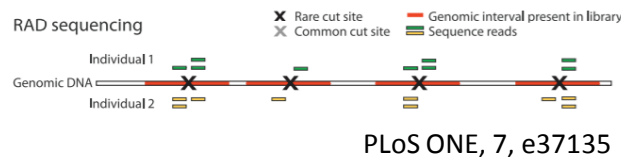
収集されたゲノムワイド多型情報の掲載

実験や文献などから収集した多型情報

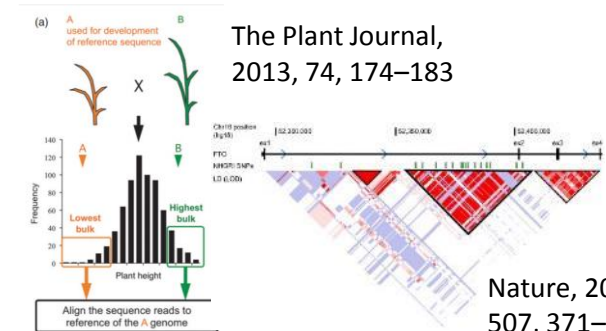
ゲノムワイド多型データ



GWAS



RAD-Seq



QTL-Seq

LD

原因遺伝子（Fruit ripening、Softening、Anthocyanin biosynthesis、Volatilesなど）

SRAデータから検出された多型情報

- 低ヘテロ接合性、高ヘテロ接合性、高次倍数体など多様
VCFとタブ区切りファイルの提供（ユーザがホモ/ヘテロ判定を実施）
- SNPの遺伝子機能への影響に関する情報を提供
（High impact：ナンセンス変異、非同義置換など、Moderate impact：変異）



ミヤコグサ



サブクローバ



キャベツ



スイカ



トマト

新規サーバの構築

* 新規サーバの構築

- PGDBj (plantdb: RHEL5)
- PGDBjテストサーバ (ライン社)
- togodb (RHEL6)
- togodb2 (マーカーキュレーションシステム: RHEL6)
- BLASTサーバ (RHEL6)

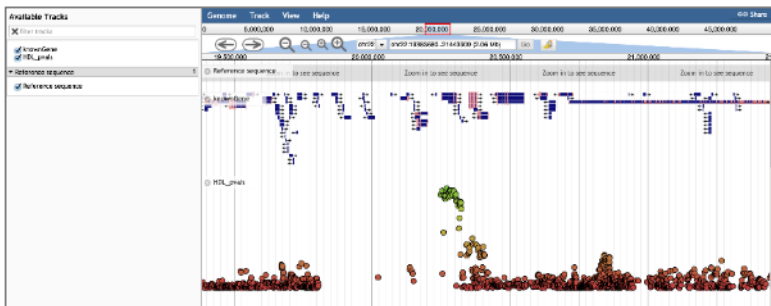


- PGDBj (新サーバ: RHEL7)
- PGDBjテストサーバ (新サーバかずさに設置)
- togodb (RHEL6)
- togodb2 (マーカーキュレーションシステム: RHEL6)
- BLASTサーバ (RHEL6)

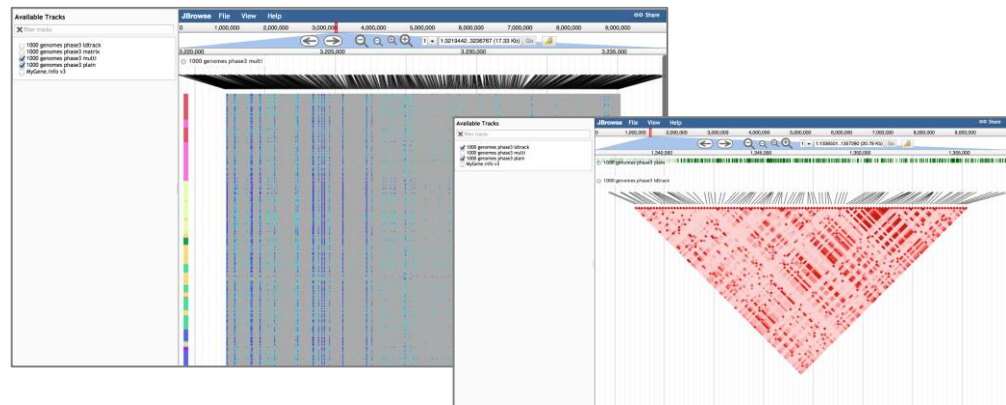
* JBrowseのテストサーバの構築 プラグインのインストール

gwasviewer、multivariantviewer、myvariantviewer、BlastView、
FeatureSequence、fastloader

gwasviewer



multivariantviewer



研究開発スケジュール

～第3年次末

収集した情報のゲノムブラウザへの集約

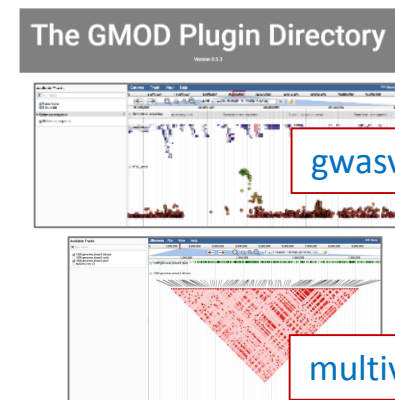
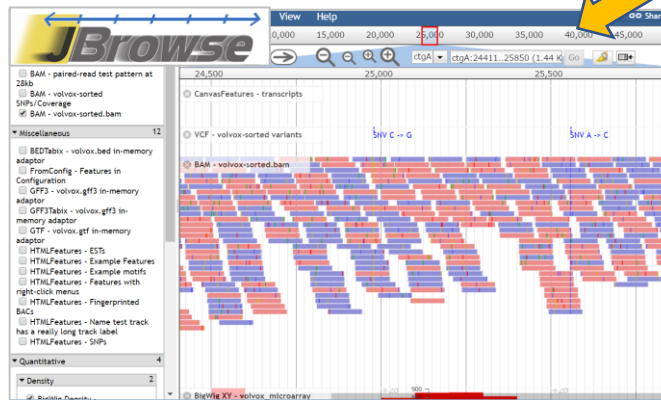
DNAマーカー、QTL情報

- ・第二期までに対象とした植物種
35種 (Pseudomolecule)
- ・かずさDNA研究所で解読した植物種
16種 (Pseudomolecule: 3種)
- ・遺伝子、転写産物 (RNA-Seq) など

今期から収集する情報

- ・ゲノムワイド多型データ
GWAS、RAD-Seq、QTL-Seq、LDなど
- ・原因遺伝子など
- ・SRAデータを用いた品種間多型

新規植物や変異体の全ゲノム構造情報を表示するシステムの構築



～第5年次末

- ・新規植物や変異体の全ゲノム構造情報を表示するシステムの実装
Phytozome、Ensembl Plants、PlantGDB、plaBi (Pseudomolecule)
- ・ゲノム比較ツールの開発 (ユーザが実行できる)
ゲノム配列全体の比較 (多型、転座)、品種の形質と遺伝子型、ゲノム構造との相関解析

(2) 種を超えたゲノム情報統合のための
データリンク基盤の構築

大阪大学・中谷明弘

(2) 種を超えたゲノム情報統合のためのデータリンク基盤の構築

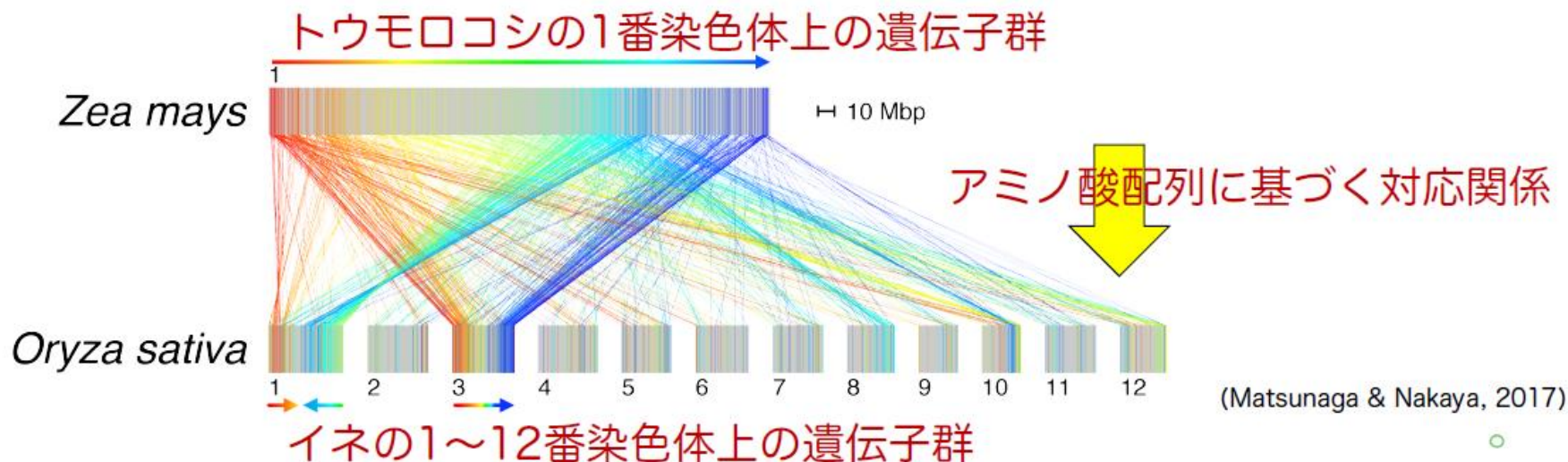
大阪大学 大学院医学系研究科 ゲノム情報学

◎中谷明弘（特任教授） ○市原寿子（特任助教）
菊地正隆（特任助教） 長谷川舞衣（特任助教）

遺伝子間の配列類似性を大規模に探索するシステムを開発

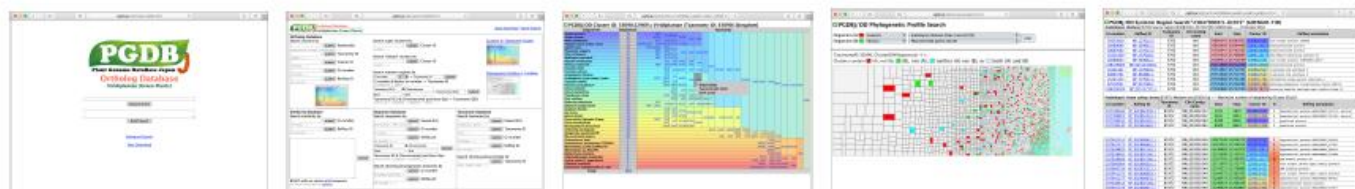
- 異なる生物種ゲノムにおける機能が類似する遺伝子候補の検索
- シンテニー様領域の検出と表示

例) トウモロコシとイネの遺伝子群の対応関係



第2期までの開発

- NCBI RefSeqのアミノ酸配列の類似クラスタのデータベースの構築
 - 自動化パイプラインの整備 (配列データ取得 → データ処理 → データベース格納)
- データおよび検索機能の公開
 - タブ区切りデータ → NBDC LSDBアーカイブ
 - RDF化 → NBDC RDFポータル
 - RDB化 → データベース可視化のウェブアプリ (<http://pgdbj.jp/od3/>)



(Nakaya, et al., 2017)

全ての検索結果がユニークなURL (検索用API) を持つように構成

検討事項

1. 植物種ゲノムデータベースのエントリーとの対応づけ
2. 総当たりの類似性検索の負荷軽減
3. 更新に際しての後方互換性の確保



(2) 種を超えたゲノム情報統合のためのデータリンク基盤の構築

1. “生物種毎の配列”との対応づけ

最初に対象とする25生物種（かずさDNA研究所が中心）を選定

科名	属名	種名	和名	配列数
バラ科	イチゴ属	Fragaria x ananassa (FANhybrid_r1.2)	オランダイチゴ(r1.2)	45,377
		Fragaria x ananassa (FAN_r1.1)	オランダイチゴ(r1.1)	230,838
		Fragaria iinumae	野生イチゴ(能郷莓)	76,760
		Fragaria nipponica	シロバナノヘビイチゴ(日本)	87,803
		Fragaria nubicola	野生イチゴ(ヒマラヤ)	85,062
		Fragaria vesca (v1.1.a2)	野生イチゴ(vesca)	33,439
		Fragaria orientalis	野生イチゴ(orientalis)	99,674
ヒルガオ科	サツマイモ属	Ipomoea trifida (0431-1)	野生サツマイモ(0431-1)	109,449
		Ipomoea trifida (Mx23Hm)	野生サツマイモ(Mx23Hm)	62,403
ナデシコ科	ナデシコ属	Dianthus caryophyllus L.	カーネーション	56,382
ナス科	ナス属	Solanum melongena L.	ナス	85,446
		Solanum lycopersicum	トマト	77,779
タデ科	ソバ属	Fagopyrum esculentum Moench	ソバ	286,768
フトモモ科	ユーカリ属	Eucalyptus camaldulensis	ユーカリ	118,501
トウダイグサ科	ナンヨウアブラギリ属	Jatropha curcas	ヤトロファ/ナンヨウアブラギリ	57,437
マメ科	シャジクソウ属	Trifolium subterraneum (Daliak)	サブクローバー(Daliak)	42,706
		Trifolium subterraneum (Woogenellup)	サブクローバー(Woogenellup)	62,251
	ミヤコグサ属	Lotus japonicus	ミヤコグサ	48,105
アブラナ科	ダイコン属	Raphanus sativus L.	ダイコン	80,521
	シロイヌナズナ属	Arabidopsis thaliana (Araport11)	シロイヌナズナ	48,359
ウリ科	スイカ属	Citrullus lanatus (v1)	スイカ	23,440
イネ科	イネ属	Oryza sativa (1.0)	イネ	42,208
	シバ属	Zoysia japonica	シバ	59,271
		Zoysia matrella	コウシュンシバ	95,079
		Zoysia pacifica	コウライシバ	65,252



2. 総当たりの類似性検索の負荷軽減

- 特定の系統群内での相同性検索の結果に基づいてコンセンサス配列を生成
- 系統群間の対応関係はコンセンサス配列間の対応関係を経由する

系統群① 複数生物種の類似配列

マルチプルアラインメント

M S A V K S A L G D M V I T F L W V I ...
M H A V R S A L G D M V I T F F W V I ...
M G V V K S A I G D M L M T F S W V V ...

コンセンサス配列を生成

M A V K S A L G D M V I T F W V I ...

系統群② 複数生物種の類似配列

マルチプルアラインメント

M S A V K S A L G D M V I T F L W V I ...
M H A V R S A L G D M V I T F F W V I ...
M G V V K S A I G D M L M T F S W V V ...
M G V V K S A I G D M L M T F S W V V ...

コンセンサス配列間の類似関係でリンク

M G V K S A G D M T F S W V ...

※コンセンサス配列は位置特異的スコア行列等で保持

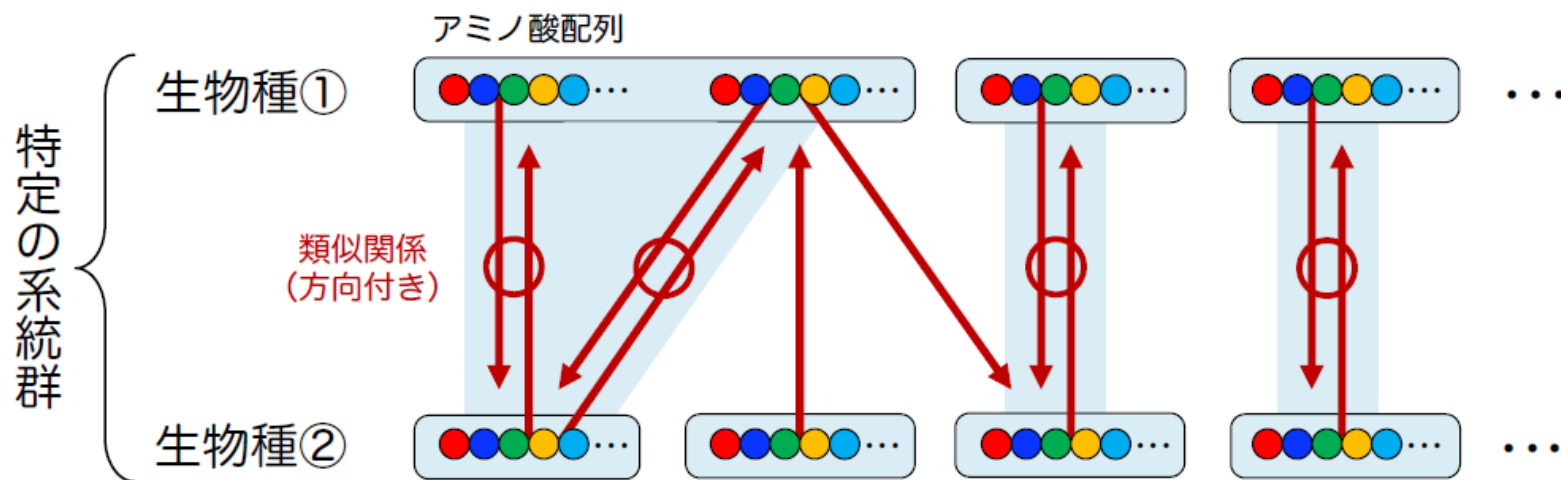


3. 更新に際しての後方互換性の確保

- 生物種の追加・置換・削除・選択を実現
 - 生物種の個数や順番に依らずに同じ結果
- 配列の類似関係に基づいたグラフ構造 (2部グラフ)
 - カバー率・一致度・E-valueによる閾値 (e.g., 80%・30%・ 10^{-4})
- 双方向で類似している配列群 (連結成分) を抽出

➡ マルチプルアラインメント
コンセンサス配列

➡ コンセンサス配列間の類似関係で系統群間をリンク



(3) カスタム型多型・ハプロタイプ検出 システムの構築

かずさDNA研究所・磯部祥子

植物も個体ゲノム時代に本格的に突入

SNP検出

- ・多検体処理（例：分離集団）に対応していない

ハプロタイプ検出

- ・GUI付きは利用できるリファレンスが限られている。
- ・植物ではあまり整備されていない（倍数性種への対応なし）

既存のGUIベースのパイプライン



ユーザーがリードデータを投げ込み
多型（SNP、Indel、CNV、SV）を検出するシステムの構築（～3年次）
ハプロタイプ検出システムの構築（～5年次）

達成目標

達成目標（～第3年次末）

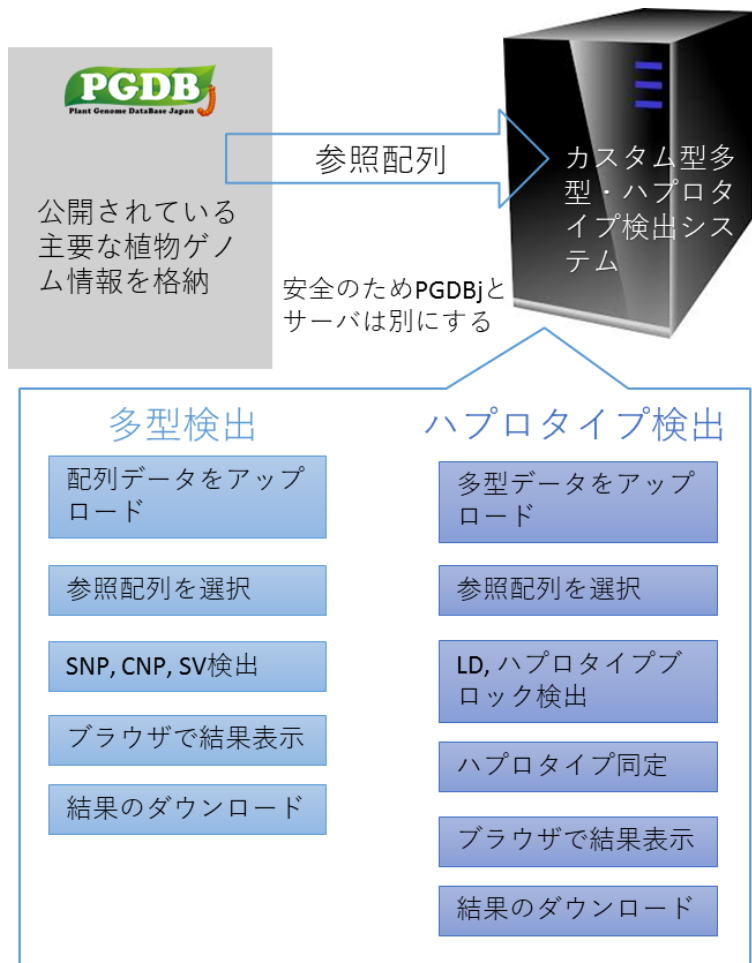
- ① カスタム型多型検出システムの構築
 - ・ ユーザが取得したNGSデータをアップロードすることで多型を検出
 - ・ 参照配列はPGDBjに格納されたものからユーザが選択
- ② ハプロタイプ検出システムの構築
 - ・ SNP表など多型データをアップロードすることでLDやハプロタイプブロックを検出、表示
 - ・ ハプロタイプ検出プログラムの評価、解析プログラムの開発も視野

達成目標（～第5年次末）

- ① システムで開発解析した結果をブラウザ上で確認する機能の開発
 - ・ 課題（1）で構築するゲノム横断的関連情報表示サイトとの連携
 - ・ 検出したLDやハプロタイプブロックからヘテロ系統に対してゲノム全体のハプロタイプを同定できるシステムの開発

ハードウェアの整備

* 新規回線の契約



* 新規サーバの構築

DELL PowerEdgeR430

Xeon E5-2698 v4 2.2 GHz, 50 cache, 20C/40T × 2

32 GB RDIMM

10 TB SATA × 4 (RAID01)

図 カスタム型多型・ハプロタイプ検出システムの概略

カスタム型多型検出システムの構築

① カスタム型多型検出システムの構築

- ・ ユーザが取得したNGSデータをアップロードすることで多型を検出
- ・ 参照配列はPGDBjに格納されたものからユーザが選択

かずさ内部で使用しているSNP解析パイプラインの改変

直列で処理するシステムの構築

対象植物種・Pseudomolecules/PGDBjに格納

SNP解析パイプライン（内部版）

入力ファイルを選択してください (必須)

メールアドレスを登録してください (必須)

処理内容

- FastQCによるクオリティチェックのみ
- FastQCによるクオリティチェック → PRINSEQによるトリミング → fastx_clipperによるアダプター除去 → ペア/シングルに仕分ける
- FastQCによるクオリティチェック → PRINSEQによるトリミング → fastx_clipperによるアダプター除去 → ペア/シングルに仕分ける → マッピング(bowtie2)
- FastQCによるクオリティチェック → PRINSEQによるトリミング → fastx_clipperによるアダプター除去 → ペア/シングルに仕分ける → マッピング(bowtie2) → SNP検出
- FastQCによるクオリティチェック → マッピング(bowtie2)
- FastQCによるクオリティチェック → マッピング(bowtie2) → SNP検出
- トリミングのみ: PRINSEQによるトリミング → fastx_clipperによるアダプター除去 → ペア/シングルに仕分ける
- トリミングとマッピング: PRINSEQによるトリミング → fastx_clipperによるアダプター除去 → ペア/シングルに仕分ける → マッピング(bowtie2)
- トリミングとマッピング: PRINSEQによるトリミング → fastx_clipperによるアダプター除去 → ペア/シングルに仕分ける → マッピング(bowtie2) → SNP検出
- マッピング(bowtie2)のみ
- マッピング(bowtie2) → SNP検出
- マッピング(tophat)のみ
- SNP検出のみ

クオリティタイプ	●Illumina1.8以降	●クオリティタイプ古い
prinseq を行う場合	指定されません	-phred64 が指定されます

○prinseq オプション

-trim_right: ●指定しない (3'末端を除去しない) ●指定する (除去する) → 塩基数 1

-trim_qual_right (3'末端クオリティス 1'ア下位): ●指定しない ●指定する 10

-trim_left: ●指定しない (5'末端を除去しない) ●指定する (除去する) → 塩基数 1

-min_len (残すリードの最長): ●指定しない ●指定する 100

-ns max n (許容するNの数): 0

-derep: ●重複リードを考慮しない ●考慮する ☒ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

○fastx_clipper オプション

-a (アダプター配列): ●AGATCGGAAGAGC (TruSeq) ●CTGTCTCTTATACACATCT (Nextera) ●任意

-l (残すリードの最長): ●指定しない ●指定する 100

-M (clippingに必要とするアライメント最長): ●指定しない ●指定する 10

-n: ●指定しない (Nを含むリードを残さない) / ●指定する (残す)

○マッピング・SNP検出

リファレンスFASTA:

○マッピング (bowtie2) オプション

PCクラスターでのCPU使用数: ●1 ●5 ●10

マッピングモード: ●--end-to-end / ●--local

入力リードタイプ: ●-q (FASTQ) / ●-f (FASTA, 「マッピングのみ」選択時のみ)

-I (pairのインリートサイズ min): 0

-X (pairのインリートサイズ max): 1000

ペアのタイプ: ●-fr / ●-rf

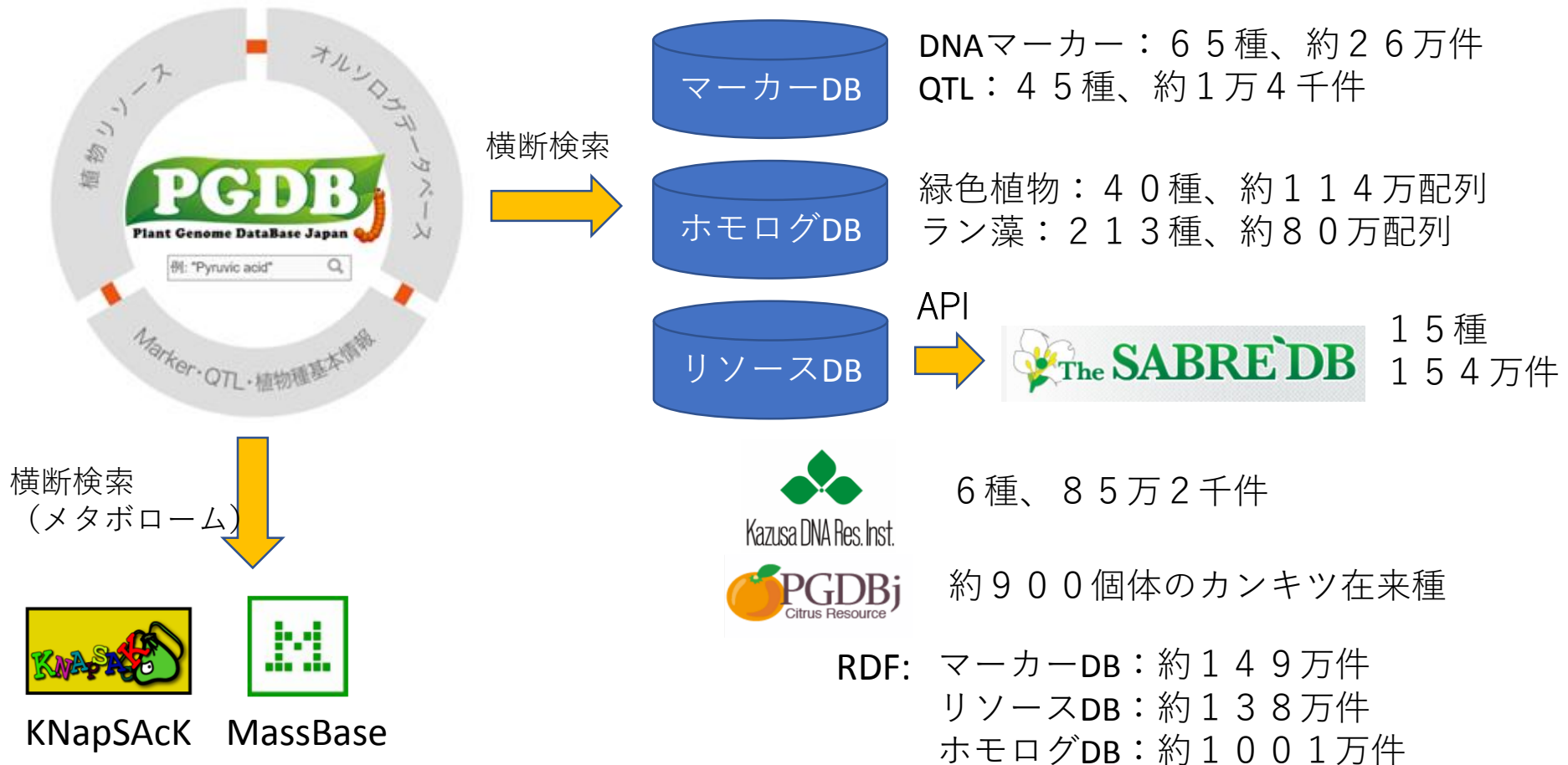
(4) PGDBjコンテンツの更新

かずさDNA研究所・平川英樹

(4) PGDBjコンテンツの更新

新しいデータを追加することで最新の情報を公開する

従来



データの更新・追加、植物種の追加

PGDBjコンテンツの更新の方針

従来のコンテンツ

- ・ DNAマーカー（6 5 種、約 2 6 万件）
- ・ QTL情報（4 5 種、約 1 万 4 千件）
- ・ リソースDB
- ・ ホモログDB

新たなコンテンツ

- ゲノムワイド多型に関するデータ
 - ・ GWAS、RAD-Seq、QTL-Seq、LD
 - ・ SRAデータ
- 新たな植物種（適宜実施）

- 
- ・ RDF化
 - ・ オントロジーの採用（GO、TO、POなど）



他課題との横断検索

- ・ DB構成の変更
- ・ ユーザーインターフェースの開発、利用マニュアルの整備
- ・ キュレーションの自動化
- ・ 更新の自動化

PGDBjコンテンツの更新（実施内容）

コンテンツの更新

- DNAマーカー、QTLについてマニュアルキュレーションを適宜に実施（2010年以降の文献を調査）
- ゲノムワイド多型に関する文献（GWAS、RAD-Seq、QTL-Seq、LDなど）
- リソース情報についても適宜に登録
- 植物種の追加（新たに解読された植物種（Pseudomolecule））

キュレーションの自動化

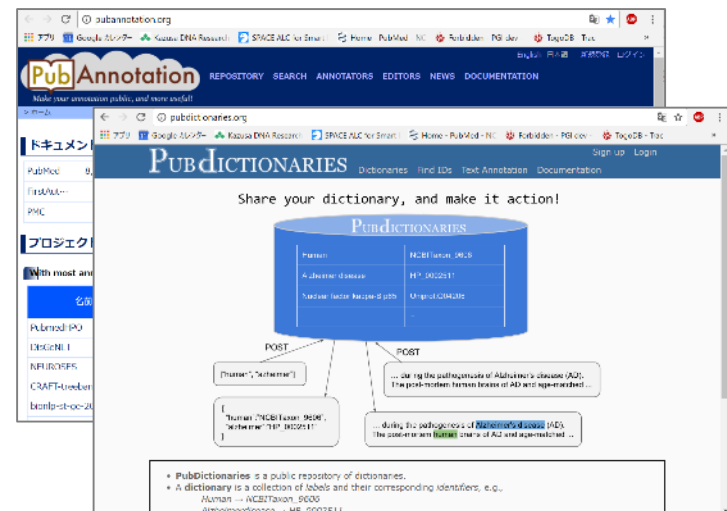
- センテンスキュレーション

PubTator



RESTful APIでコマンドラインで
文章を抽出することが可能

PubAnnotation/PubDictionaries



スクリプトの作成（辞書作成、オントロジー化）

研究開発スケジュール

達成目標（～第3年次末）

コンテンツの更新

- ・ DNAマーカー、バイオリソース関連についてマニュアルキュレーションを適宜に実施
- ・ ユーザーインターフェースの開発、利用マニュアルの整備

キュレーションの自動化

- ・ センテンスキュレーションをローカルで実施するためのツール開発
- ・ RDF化（PGDBjコンテンツ、新たなコンテンツ（植物種、ゲノムワイド解析など））

DB構造の変更

達成目標（～第5年次末）

すべての植物種についてのコンテンツの更新

- ・ キュレーション（DNAマーカー、QTL、ゲノムワイド多型）
- ・ RDF化
- ・ オントロジーの整備

更新の自動化

- ・ ゲノム情報の収集
- ・ DBの運用・管理