

ライフサイエンスデータベース統合推進事業
(統合化推進プログラム)

研究開発中間報告書

「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」
田畑 哲之 | かずさ DNA 研究所 所長



©2019 田畑 哲之(かずさDNA研究所) Licensed under CC BY 4.0

§1. 研究開発実施の概要

【目的】 本開発は、多種多様な植物種を対象とする基礎研究から実用技術開発に至る幅広い分野に対して、その基盤となるゲノム関連情報を提供するものである。前期までに開発したデータベース「PGDBj」で提供されている国内外の植物ゲノム情報に加え、開発期間中に公開される新たな情報を収集し、JBrowse を基軸としてこれらのデータを相互に比較することのできるデータベース(DB)の開発を目指す。本研究開発では、オルソログ検出手法を基軸に、今後一層の増加が見込まれるゲノム配列データを柔軟に取り入れ、ユーザーが必要な情報を抽出できる DB の開発を目指す。さらに個体ゲノムの時代が植物にも到来することを見据え、ユーザー自身が分析した NGS のショートリードデータを投入できるカスタム性の高い多型・ハプロタイプ検出システムを開発・実装する。

【方法】 PGDBj で格納されている植物の DNA マーカー、QTL 情報に加え、新たに多数の植物の全ゲノムアセンブル配列や遺伝子配列、機能予測やゲノムワイドな SNPs 等を格納する DB「Plant GARDEN」を開発する。また、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やユーザーが自らのデータと Plant GARDEN 内のデータを比較できるカスタム型多型検出システムの開発を行う。さらにより効率的な情報収集を目指して、遺伝子機能解析や QTL 情報に関するオートキュレーション方法の検討を行う。

【結果】 新たに構築した DB「Plant GARDEN」は研究開始直後から設計を開始し、2019 年 3 月に β 版(日本語)を公開した。また、2019 年 6 月には β 版として英語版も公開した。データの検索は植物種を指定して実施されることが基本であると考え、植物種毎のページを作成してデータを格納する構成とした。β 版開発当初は 9 種のみでの公開であったが、6 か月後の 2019 年 9 月現在で 42 種に種数を拡大している。

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では Plant GARDEN に格納した植物種のアミノ酸配列情報を整備するとともに配列プロファイリングに基づく類似度指標の算出と蓄積を行う手法を検討した。また、カスタム型多型検出システムの開発では、専用のサーバー上にユーザーがデータを Upload して SNP を検出できるパイプラインを構築した。現在は任意のサーバー上で解析を実行できるようにパイプラインのコンテナ化を行っている。

コンテンツの収集と更新に関しては必要情報のマニュアルキュレーションを行うとともに、センテンスキュレーション法の検討や QTLTable- Miner++ の試用とマニュアルキュレーションとの比較を実施した。また、アドバイザー委員会や学会展示等を通じてユーザーの意見の積極的な収集を行い、Plant GARDEN の開発にフィードバックした。

§2. 研究開発実施体制

1. 各グループの担当項目

(1) かずさ DNA 研究所グループ(研究代表者グループ)

開発項目のうち「ゲノム横断的関連情報表示サイトの開発」、「カスタム型多型・ハプロタイプ検出システムの構築」および「PGDBj コンテンツの更新」を担当する。

(2) 大阪大学グループ(主たる共同研究者グループ(1))

開発項目のうち「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」を担当する。

(3) 筑波大学グループ(主たる共同研究者グループ(2))

「ゲノム横断的関連情報表示サイトの開発」のうち、開発する DB 内のキーワード検索システムの構築を担当する。

2. 有識者会議等

(1) 会議概要

名称	ライフサイエンスデータベース統合推進事業統合化推進プログラム 「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」 アドバイザー委員会
目的	本開発の効率的な推進とユーザーの意見を開発に反映するため、国内データベース維持管理者およびユーザーとなる植物ゲノム研究者や植物関連団体関係者からなるアドバイザー委員会を年に2回程度開催する。
委員数	10 人

(2) 開催歴

年月日	場所	主な議題・指摘事項等
2017 年 8 月 3 日	東京ステーションコンファ レンス	平成29年度第1回アドバイザー委員会 プロジェクト開始にあたって、開発計画に 対する意見交換など
2018 年 2 月 8 日	東京ステーションコンファ レンス	平成29年度第2回アドバイザー委員会 新規に開発するデータベースに関しての 意見交換など
2018 年 9 月 3 日	東京ステーションコンファ レンス	平成 30 年度第 1 回アドバイザー委員 開発を実施したデータベース「Plant GAR DEN」に関しての意見交換など
2019 年 2 月 18 日	東京ステーションコンファ レンス	平成 30 年度第 2 回アドバイザー委員 Plant GARDENβ 版の公開および各開発 項目に関して意見交換など
2019 年 8 月 7 日	東京ステーションコンファ レンス	2019年度第 1 回アドバイザー委員会 Plant GARDENβ 版の開発内容および各

年月日	場所	主な議題・指摘事項等
		開発項目に関して意見交換など

§3. 研究開発の目的、実施内容及び成果

1. 研究開発対象のデータベース・ツール

(1) データベース

主要なもの

正式名称	略称	概要
Plant GARDEN	Plant GARDEN	様々な植物のゲノムやマーカー情報を集めたポータルサイト。全ゲノム配列、遺伝子配列、アノテーション、DNA マーカー、形質座、連鎖地図、SNPs 変異等の情報を格納している。

上記以外のもの

正式名称	略称	概要
Plant Genome Database Japan	PGDBj	植物のゲノム情報に関するポータルサイト。オルソログ、マーカー・QTL およびリソース情報を格納している。
PGDBj オルソログデータベース		PGDBj 内に格納しているアミノ酸配列の類似性に基づいた植物の遺伝子のオルソログ情報を蓄積したデータベース
PGDBj DNA マーカー・QTL データベース		PGDBj 内に格納している植物の DNA および QTL 情報を格納したデータベース
PGDBj カンキツリソースデータベース		PGDBj 内に格納しているカンキツの遺伝資源情報を格納したデータベース
ゲノム解読状況データベース		PGDBj 内に格納しているゲノムの解読状況を格納したデータベース

(2) ツール等

正式名称	略称	概要
PGDBj 横断検索システム		PGDBj 内に格納されたデータを横断検索するシステム
PGDBj 育種向け DNA マーカーページ		PGDBjに格納された情報のうち、育種関係者にもわかりやすくマーカー情報を表示するページ
Hayai-Annotation Plants		植物の遺伝子配列から機能アノテーションを高速に実施するプログラム。Plant GARDEN 内に格納する遺伝子の機能アノテーションも本ツールで再アノテーションし、情報の精度と内容を統一させた。
SNP Detection		Plant GARDEN に格納されている全ゲノム配列とユーザーが保有するゲノム配列情報とを比較して変異を検出するためのパイプライン。

2. 達成目標及び実施計画

(1) 当初の実施計画・達成目標

本開発は、多種多様な植物種を対象とする基礎研究から実用技術開発に至る幅広い分野に対して、その基盤となるゲノム関連情報を提供するものである。前期までに開発した DB「PGDBj」で提供されている国内外の植物ゲノム情報に加え、開発期間中に公開される新たな情報を収集し、JBrowse を基軸としてこれらのデータを相互に比較することのできる DB の開発を目指す。本研究開発では、オルソログ検出手法を基軸に、今後一層の増加が見込まれるゲノム配列データを柔軟に取り入れ、ユーザーが必要な情報を抽出できる DB の開発を目指す。さらに個体ゲノムの時代が植物にも到来することを見据え、ユーザー自身が分析した NGS のショートリードデータを投入できるカスタム性の高い多型・ハプロタイプ検出システムを開発・実装する。

以上の目標を実現するために、以下の4項目で開発を実施する。

(1) ゲノム横断的関連情報表示サイトの開発

Pseudomolecules が構築されている植物種 8 種以上のゲノム情報を相互に表示できるシステムを開発することを目標とする。本開発では新たな DB(Plant GARDEN)を構築することとし、DB のデザインとゲノム情報を相互に表示ができるシステム開発として、Pseudomolecules が構築されている植物種 8 種のゲノム配列および遺伝子情報の類似性検索と JBrowse 上での情報の集約が表示できるプロトタイプ A を作成する。

(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

対象植物種のアミノ酸配列の情報整備を行い、配列間の対応関係整理とデータストアを構築する。また、植物種の追加・更新システムの開発を行う。

(3) カスタム型多型・ハプロタイプ検出システムの構築

解析を実施するためのサーバーおよびデータ転送システムを構築する。また、SNP および CNV(Copy Number Variation) 検出のためのパイプラインを構築する。

(4) PGDBj コンテンツの更新

センテンスキュレーションのローカルでの実施を行う。また、新たな情報収集・キュレーションと RDF 化を実施する。

(2) 期間中に追加・削除・変更した実施計画・達成目標

当初計画においては Plant GARDEN の β 版を平成 32 年度までに開発し、最終年度(平成 33 年度)に正規版を完成させる予定だった。しかし、NGS の技術発展や Web ベースの解析ツールの開発などユーザーを取り巻く環境は急速に変化しており、Plant GARDEN が世界に幅広く認知される DB となるには一刻も早い公開が必要と考え、開発を前倒しして進めた結果、2019 年 3 月に β 版公開とすることとした。

一方、Plant GARDEN 開発過程におけるユーザーヒアリングから、Plant GARDEN が目指す「種を超えた植物ゲノム情報のポータルサイト」「異なる種間のデータ比較」「シンプルで使いやすい DB」といった事項に加え、「NGS データ処理や統計処理を行うツールの整備」「ユーザーが保有するデータを表示する Viewer の開発」が求められていることがわかった。ユーザーが保有するデータを Plant GARDEN のシステムに投げ込み、DB 内に格納

されているデータと比較して SNPs などの配列変異等を検出するシステム開発に関しては、研究課題(3)カスタム型多型・ハプロタイプ検出システムの構築において実施予定である。しかし当初計画では Plant GARDEN 内で設置したサーバーにユーザーがデータを投げ込むことで解析をするシステム開発のみにとどまっていた。本システム開発については SNP/Indel 検出についてすでにパイプラインを構築、Plant GARDEN の β 版公開とあわせて平成 31 年度にはユーザーにシステム開放する予定であるが、サーバー容量が限られているためユーザーの要求に応えきれない可能性が高い。より多くのユーザーを獲得し、高い評価を得る DB とするには、Plant GARDEN 内で管理するサーバー外で稼働するツールの開発とユーザーの解析結果を表示し、かつ Plant GARDEN とシームレスに連動する Viewer の開発が必要であると考えた。そこで研究課題(3)カスタム型多型・ハプロタイプ検出システムの構築のうち、利用者が比較的限られる CNV 検出パイプラインの構築を中止し、代わりに「NGS 配列解析ツールの開発」として、Plant GARDEN 内のデータと連動し、NGS のマッピングや配列変異の検出等を任意のサーバー上で実施する解析パイプラインを構築する。さらに、ユーザーデータを Plant GARDEN 内のデータと比較した結果をユーザーがローカルで表示できる Viewer 開発に着手する。

一方、Plant GARDEN が広く世界に認知されるためには、国内だけでなく国際ユーザーとの意見交換も重要であることから、国際学会でのユーザー意見の収集も積極的に行うこととした。また、Plant GARDEN 内のキーワード検索システム開発をより効率的に実施するため、2019 年度から筑波大学を研究グループに新たに加えた。

3. 実施内容

(1) 実施内容

① ゲノム横断的関連情報表示サイトの開発

本研究開発では種、属、科などさまざまな階層間のゲノム関連情報を容易に比較できる仕組みを整備し、特定の種で得られている知見を他の種で参照できる基盤を構築する。2017 年度にはスキーマの設計を行い、2018 年度に JBrowse によるゲノムビューアを軸に細部ページの設計を進めた。主な担当者は磯部祥子、平川英樹、Jeffrey Fawcett (2017 年度まで)、原田大士朗(2018 年度から)、小原光代(いずれもかずさ DNA 研究所)である。

新しく作成した基盤 DB の名称を「Plant GARDEN (Genome And Resource Database ENtry)」として、2019 年 3 月に β 版(日本語)を公開した。また、2019 年 6 月には β 版として英語版も公開した。Web ページデザインなど作業の一部は(株)バスキュールに委託し、策定したスキーマ(DB 構成)原案をもとに、ページ構成を決定した。トップページはユーザーがデータ検索時に行うべきアクションを直観的に理解できるよう、4 つのメニュー(植物種からさがす、他の方法でさがす、解析してみよう、データ一覧)から構成し、シンプルな画面構成となるようデザインした(図1)。また、データの検索は植物種を指定して実施されることが基本であると考え、植物種毎のページを作成してデータを格納する構成とした。各植物種のページには「この種について」「ゲノム配列をみる」「キーワード検索」「マーカーをさがす」「形質との関連をさがす」「その他の検索」「リンク」の 6 つの項目をメインにおき、各項目より目的とするデータを表示する。また、ゲノム配列は複数のバー

ジョンが存在することから、バージョン毎に情報を表示することとし、遺伝子情報もゲノムのバージョンに依存することから、ゲノムのバージョン別のページに格納することとした。一方、異なる種類のデータを横断的に表示するため、格納したデータをJBrowse上に表示できるよう整備し、JBrowseを介したリンクにより他植物種、もしくは他種類のデータを取得できるようにした。

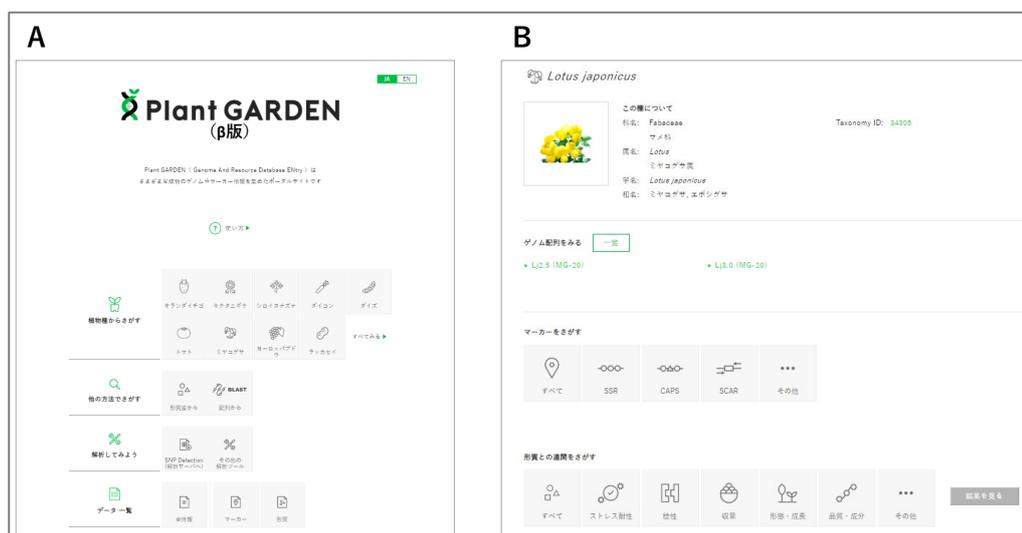


図 1. Plant GARDEN のページ例 A: トップページ B: 植物種毎のページ

Plant GARDEN の整備はデータが比較的整理されており、かつかずさ DNA 研究所でこれまでゲノム配列 DB を開発してきたミヤコグサ (*Lotus japonicus*) をテストケースとして、情報整備とデータ格納および Web ページの構築を進めた。β 版公開時の 2019 年 3 月では 9 種の公開であったが、2019 年 9 月末現在では 42 種を格納している。格納する種は、公開されているゲノムのうち、pseudomolecule レベルの高品質のゲノムから優先することとしているが、ユーザーの希望が高い場合は Scaffold レベルでも格納している。また、今後は DB 整備を自前で実施するのが困難なユーザーに対し、公開前(論文投稿準備時点)での情報受付も呼びかける予定である。また、現在公開中の β 版ではキーワード検索機能が実装されていない。そのため、筑波大グループが 2019 年度から新たに参加し、田村卓郎、杉原 英志が担当して開発を進めている。

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

本研究開発では、複数の生物種に跨ったゲノム情報の統合を目的として、遺伝子のアミノ酸配列の類似度情報に基づいたデータリンク基盤の構築を行う。担当は中谷明弘、市原寿子(大阪大学)である。ここでは植物種・系統・アセンブリバージョンのゲノム配列を追加してアミノ酸配列の相同性検索を実施し、類似度指標の情報を算出、蓄積する。また、配列間の対応関係を整理しクラスタリングするための方法を検討する。

これまで、新しい植物種と配列を追加してアミノ酸配列の相同性検索を実施し、類似度指標の情報を算出、蓄積した。また、二配列間の情報に基づいて三配列以上を含む配列間の対応関係を整理しクラスタリングするための方法を検討した。具体的には、近年、クラスタリングツールとして広く使用されている ProteinOrtho(<https://www.bioinf.uni-lei>

pzig.de/Software/proteinortho)と OrthoFinder(<http://www.stevkellylab.com/software/orthofinder>)を実行して、得られた結果を比較した。比較に際しては、国際標準 DB でもあり、また多種多様な植物種のアミノ酸配列情報を含む UniProt Reference Clusters(Uniref, <https://www.uniprot.org/uniref>)をクラスタリングの正解データセットとして参照することにした。デフォルトのパラメータで計算を実行した結果、形成されるクラスタの数と各クラスタに含まれるアミノ酸配列の数(クラスタサイズ)に大きな違いが見られた。本課題で扱った植物種データで二者を比較した場合、ProteinOrtho では配列同士が分割されて、サイズの小さいクラスタが形成される傾向を示したのに対し、OrthoFinder では多くの配列同士がまとめられ、サイズの大きいクラスタが形成される傾向を示した。

本課題では特に、多数の植物種間の配列関係を重視することから、多くの配列を一つのクラスタにまとめる OrthoFinder よりも、より類似した配列だけでクラスタを形成させる ProteinOrtho を採用することにした。また、クラスタデータセット構築に際しては、正解データとして参照する Uniref のデータセット(Uniref50, Uniref90, Uniref100)の構築基準(配列カバー率 80%, 配列一致度 50%, 90%, 100%)も参考に実施した。また、将来的な植物ゲノムデータの追加・更新作業を軽減するためのシステム構築を目的とし、クラスタリングの結果に基づいて生成する配列プロファイルを利用する方法を検討した。その結果、クラスタ情報に基づき生成した配列プロファイルの利用により、検索にかかる時間が 1/6~1/8 に短縮された。従来法で形成されたクラスタとの一致度は 80~84%だった(図 2)。

③ カスタム型多型・ハプロタイプ検出システムの構築

本項目ではユーザーが自身のデータを専用のサーバーに投げ込み、ゲノム横断的関連情報表示サイトに格納されている配列からリファレンスとなる配列を選択して、GUI 上でプログラムを操作して多型を検出するシステムを構築する。担当者は藤代継一(2017 年度)、山田学(2018 年度)、中村保一(2019 年度)および磯部祥子(いずれもかずさ DNA 研究所)である。2017~2018 年度は SNP 検出用のサーバーをかずさ DNA 研究所に設置して解析用パイプラインを組み込み、ユーザーがデータをアップロードして配列データの品質評価、トリミング、リファレンスへのマッピング、SNP の検出、フィルタリングを実施できるシステムを 2018 年 5 月に制限つきでテスト公開した。なお、外部からのデータは専用で設置した回線を通じて受け付ける。テスト公開ではユーザーサイトの IP アドレス管理方法などが原因で一部アクセスできない事例も発生したが、概ね順調に運用できたことから 2019 年 3 月末には β 版として制限なし公開とすることにした。また、パイプラインを実行する際の GUI を改良し、初心者でも分かりやすい表示に変えた(図3)。

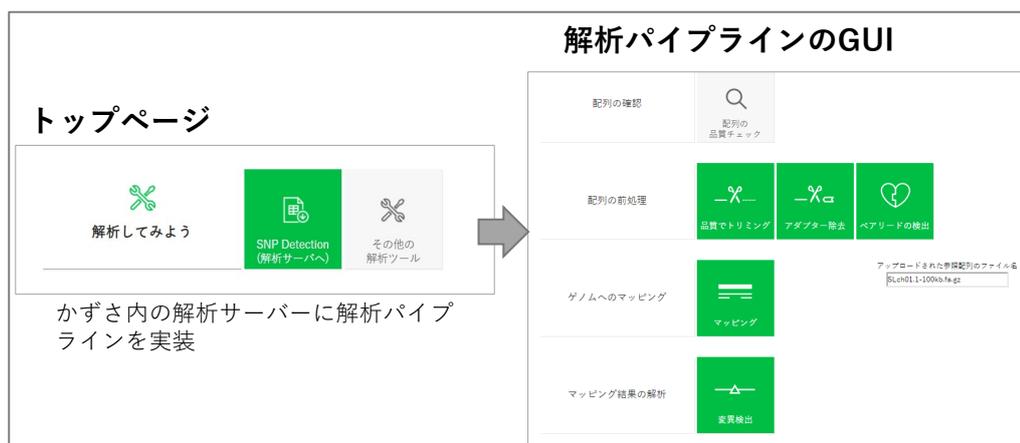


図 3. SNP 解析パイプラインの GUI

リードデータをサーバーに投げ込んで SNP 検出をするシステムはサーバーに負荷がかかるためアクセス制限を除くことが難しい。そこで 2019 年度からは任意のサーバー上で SNP 解析パイプラインを実行するために、パイプラインのコンテナ化に取り組んでいる。今後はコンテナ化したパイプラインと Plant GARDEN を連動させて、Plant GARDEN 内に格納した参照配列とユーザーが独自に解析した配列をより容易に比較できるシステム開発を実施する。

一方、ユーザーが独自に取得したゲノム情報から遺伝子予測やアノテーションを実施するケースが今後増えることを見越し、Ghelfi Andrea が担当者となって植物に特化した遺伝子アノテーションツール「Hayai-Annotation Plants」を開発した (Ghelfi et al. 2019、文献1)。「Hayai-Annotation Plant」は一般に普及している程度のスペックを有する PC で解析を実行できるプログラムであり R パッケージとして Git hub から公開している (<https://github.com/kdri-genomics/Hayai-Annotation-Plants>)。さらに Plant GARDEN に格納する遺伝子情報を再アノテーションするために探索する遺伝子 DB を全生物種に拡大した「Hayai-Annotation ZEN」を開発した。独自のアノテーションシステムを開発することで、従来のツールに比べて格段に解析を高速化させ、精度の高い結果を検出することができた (表1)。

表1 シロイヌナズナ (Araport11)の遺伝子配列を Hayai-Annotation と既存のツール (Blast2GO, TRAPID)で機能アノテーションした結果の比較.

	Hayai-Annotation Plants ^a (4 Best Hits)	Hayai-Annotation Plants ^a (20 Best Hits)	Blast2GO ^b (Default)	TRAPID ^c
TP	103,927	105,283	105,470	59,071
FP	13,589	14,379	34,841	632,636
FN	24,557	23,201	23,014	69,413
Specificity (%)	88.4	88.0	75.2	8.5
Sensitivity (%)	80.9	81.9	82.1	46.0
Accuracy (%)	84.7	85.0	78.6	27.3
p value	0.6330	0.6168	reference	6.2e-07
Running time (Total)	4 m 11 s	11 m 58 s	21 h 28 m*	6 h 11 m
Running time (Alignment)	3 m 57 s	11 m 07 s	17 h 35 m*	NS
Running time (Annotation)	14 s	51 s	3 m 53 s	NS

TP: True Positive, FP: False Positive, FN: False Negative, NS: non-specified.

^a Hayai-Annotation Plants の実行環境: Macintosh laptop (2.2 GHz, 2 cores, 8 GB RAM).

^b Blast2GO はクラウド上に設定した。*BLAST+は 2.40 GHz, 16 cores, 32 GB RAM のサーバーで実施し、FASTA ファイルを 10 分割して並列で計算した。

^c TRAPID は Web ベースシステムのため、計算速度は状況により異なる。

④ Plant GARDEN コンテンツの更新

PGDBj から Plant GARDEN へ DNA マーカーと QTL 情報に関するコンテンツを移行させるため、現 PGDBj に格納されているデータを整理しフォーマットや記述をより統一させた。また、PGDBj に格納していない 2010 年以降の文献を対象として新たなマーカー情報のキュレーションを実施した。さらに PGDBj に新たに格納するコンテンツとして全ゲノム配列および遺伝子配列等の情報を収集・精査した。主な担当は原田大士朗(かずさ DN

A 研究所)である。

一方、キュレーションを効率的に実施するため、平川英樹(かずさ DNA 研究所)を担当者として、センテンスキュレーション手法の検討を実施した。対象とする文献 DB は比較検討の結果、NCBI の PubMed とすることとし、入手した論文 PDF をテキストファイルに変換して Gene ID、オントロジー(GO, TO, PO, EO)、NCBI の各種 DB(EST, CDS, PEP)、NCBI の遺伝子名および Nuccore の遺伝子名からキーワードを抽出するスクリプトを作成した。抽出されたセンテンスの一部を目視で確認して精度を検討した結果、開発した手法によりキュレーション対象とする候補論文のリストが作成できること、DB への情報の格納においては目視による最終確認が必要であることが明らかとなった。さらに QTL 情報のキュレーションとして QTLTable- Miner++(QTM, Singh et al, 2018, BMC Bioinformatics)を試用した。ミヤコグサを対象として検索を実施したところ、植物種名のキーワード検索で検出された論文は 2,487 報であり、うち QTL 関連の表が掲載された論文は 87 報だった。また、その中で実際に QTL 情報が抽出できたのは 36 報であり、QTM の利用により QTL 情報のキュレーションの効率化を図ることができると考えられた。今後はセンテンスキュレーションとマニュアルキュレーションを組み合わせ、Plant GARDEN に格納するコンテンツの充実を図る。

SRA 配列の収集については、現在は1種あたり上限 100 データと決めて収集し、SNPs の検出と情報の格納を行っている。さらに格納したデータの RDF 化や API との連携をスムーズに行い、かつキーワード検索を効率的にすすめるため、Plant GARDEN の RESTful 化をすすめた(図4)



図 4. Plant GARDEN の RESTful 化

(2) 「対象とするデータベース条件」のうち採択時に未達成であった項目の対応状況

採択時に記載した「対象とするデータベース」は PGDBj であったが、採択時に全ての条件を満たしていた。また、新たに開発している Plant GARDEN はまだ開発途中であるため、g) 開発対象のデータベースについて、データ搭載対象とする主な研究分野において十分な数の研究者が利用しており、今後、さらにより多く多様な研究分野からの利用が期待できる。

のうち「十分な数の研究者が利用しており」はまだ未達成であるが、今後は達成できるものと見込んでいる。

(3) 統合化推進プログラムの他のチームや DBCLS との連携

微生物統合データベース(MicrobeDB.jp)とは植物-微生物感と統合検索を行うシステム開発を実施することで合意しており、Plant GARDEN に格納しているデータを MicrobeDB.jp に提供するなどしている。また、メタボロームデータとの連携を図るために、KNApSAcK core、Skwered KNApSAcK、MassBase および食品メタボロームDBとPlant GARDEN のデータをリンクさせた。DBCLS との連携では PGDBj や Plant GARDEN に格納している情報の RDF 化にあたって協力をいただいている。Plant GARDEN の正規版が開発されたのちには統合 TV への掲載などさらなる連携を図りたい。

§4. 主要なデータベースの利活用状況

1. アクセス数

(1) 実績

表 研究開発対象の主要なデータベースの利用状況(月間平均)

Plant GARDEN(2017年度から開発を開始し、2019年3月11日β版公開)

種別	2016年度	2017年度	2018年度	2019年度 (9月末時点)
ユニークIP数	未公開	未公開	未公開	473
ページ数	未公開	未公開	未公開	24,834

PGDBj(2017年度以降はPlant GARDEN開発のためコンテンツを拡充していない)

種別	2016年度	2017年度	2018年度	2019年度 (9月末時点)
ユニークIP数	3,204	6,145	4,885	2,496
ページ数	13,819	15,561	19,271	10,994

(2) 分析

Plant GARDENのβ版公開は2019年3月11日だったことから、Plant GARDENの利用状況は2019年度のみとした。まだ、β版であり格納しているコンテンツが限られていることから、ユニークIP数は473と少ない。一方でページ数は24,833とPGDBjに比べても非常に多いが、これは6月のページ数が133,201と多いことが原因であった。2019年6月は英語版を公開した月でPAG Asiaでの学会発表、展示を実施したことから特定のユーザーが多くのページにアクセスしたことがその要因と考える。PGDBjのアクセス数は2017年度以降、コンテンツの拡充を実施していないため、徐々に減少している。本年度中にPGDBjに格納しているマーカー情報等をPlant GARDENへ移殖することを予定しており、今後はPGDBjユーザーがPlant GARDENへ移行することが期待される。

2. データベースを利用して得られた研究成果・産業応用の例

1. PGDBjに格納されたイチゴのSSRマーカー情報を用いて、日本のイチゴ品種の遺伝子型解析とコアコレクションの作成を実施した。
関連:Wada, T., Noguchi, Y., Isobe, S., Kunihiisa, M., Sueyoshi, T., Shimomura, K. 2017, Development of a Core Collection of Strawberry Cultivars Based on SSR and CAPS Marker Polymorphisms. The Horticulture Journal doi: 10.2503/hortj.MI-142
2. 新たなイチゴ品種育成のために、PGDBjに格納されたイチゴのSSRマーカー情報を用いながらイチゴのMAGIC集団を作成した。

関連:16.Wada, T., Oku, K., Nagano, S., Isobe, S., et al. 2017, Development and characterization of a strawberry MAGIC population derived from crosses with six strawberry cultivars. *Breed Sci*, 67, 370-381 (DOI: 10.1270/jsbbs.17009)

3. その他

特になし

§5. 今後の展開

現在公開している Plant GARDEN は β 版であり、2019 年度末の正規版公開に向けてキーワードや遺伝子検索機能の追加、不具合の解消などを進めている。また、Plant GARDEN 内で用いている Viewer は JBrowse のみであるが、SNPs の表示に優れている TASUKE などの他の Viewer の導入も順次検討する。

コンテンツの拡充は DB 構築において最も重要な点であるが、まずは公開されている全ゲノムアセンブル配列および遺伝子配列情報をできるだけ速やかに収集し格納する体制を作る。また、オートキュレーションの助けをかりて、形質座と関連するマーカー情報の収集と格納も引き続き実施する。SRA 配列の収集については、現在は1種あたり上限 100 データと決めて収集し、SNPs の検出と情報の格納を行っている。RNA-Seq や ChIP-Seq 配列の格納などもユーザーの要望が高いが、限られた研究資源の中でどこを重視して開発するか、ということについては、ゲノム配列解析をとりまく状況の変化を鑑みながら検討していきたい。

Plant GARDEN の大きな特徴の1つは多数の植物種のゲノム・遺伝子配列を格納し、遺伝子配列の類似性により、種を超えた配列の比較ができることである。現在、異なる種間の遺伝子配列をユーザーがより自在に取り扱えるようなシステム開発を行うことも計画している。これにより、植物種に共通する新たな遺伝子機能の解明や、あるいは特定の種のみがもつユニークな遺伝子の発見が加速することを期待している。

個体ゲノムの時代においてはより容易に比較解析等を実施できる情報解析ツールを提供し、ユーザーが各自で Plant GARDEN 内の参照情報と自らのデータを比較できる環境を提供することが重要であると考え。本年度から追加計画として NGS 配列解析ツールの開発とローカルデータを表示する Viewer の開発を提案したが、これらを手始めとして、実行しているサーバーの場所をできるだけユーザーが意識しないで Plant GARDEN のデータとユーザーのデータをシームレスに比較、閲覧できるシステムの開発を行っていきたい。

Plant GARDEN の直接的な社会貢献分野である育種においては、育種データを管理するシェアウェア(例:PhenomeOne、<http://phnserver.phenome-networks.com/>)の普及や SIP 第2期プロジェクト(https://www8.cao.go.jp/cstp/gaiyo/sip/keikaku2/7_smartbio.pdf)等で開発が進んでいる。ゲノム情報だけでなく、このような育種、農業に関する個々のデータともつながることができるよう、これらのツール開発者とも協議して連携のとれたシステム開発にも取り組んで参りたい。

ゲノム解析は今後ますます普及し、より生活に身近な情報になると考えている。教育を通じてそのすそ野が広がることも予想されることから、Plant GARDEN は研究者だけでなく高等教育に携わる教育者や学生にも使ってもらいたいと考えている。そのため、日本語版と英語版の両方を開発し、特に日本語版では平易な表現を心掛けた。また、初心者でもわかりやすいシンプルなページデザインとなるように留意した。正規版が完成した後は、学会関係者だけでなく教育分野でも Plant GARDEN のデモンストレーションを実施して、ユーザーの拡大を目指す。

気候変動により、大きな自然災害が発生している一方で、人口増加に対応した食料増産技術の開発は全世界の喫緊の課題である。一方で、農業はデータを駆使した精密農業の時代へと変わろうとしている。農業におけるビッグデータとゲノム情報はまだ直接の繋がりはないが、遺伝子×環境の相互作用に関する研究は進んでおり、今後は農業データと植物ゲノム・遺伝子データの連携が進むと予想される。Plant GARDEN はゲノム・遺伝子情報だけでなく、SNPs や DNA マーカー、形質座など種内の変異に関わる情報を格納していることも特徴である。Plant GARDEN が育種や

農業データとの連携を図る上で中心的な役割を担えるように、様々な DB や開発関係者と連携する。

§6. 自己評価

開発項目については、いずれの目標も達成することができ、順調に進捗している。特に、当初計画ではβ版の構築を2021年9月に終了する予定であったが、ユーザーの意見をフィードバックするために、開発を前倒して2019年3月にβ版公開とした。さらに、計画当時からゲノム配列解析がさらに進化し、かつクラウドコンピューティングやAPIの普及、解析プログラムのコンテナ化などコンピュータ技術に関しても状況が変化している。このような変化を鑑みて、ユーザーが独自で取得したデータとPlant GARDEN内のデータを比較するツールの開発方針を修正するとともに、より開発に適した人材を参画者に取り入れた。

Plant GARDENはできるだけ多くの植物種のゲノム情報を格納する方針としており、今後は情報の論文公開を目指しているユーザーに対しても、論文公開と同時にDB上からも配列公開できるサービスを提供できるようにシステムを整備する予定である。そのため、「わが国の生命科学研究等によって産出された研究データを広く収集する」という統合化推進プログラムの主旨に十分に沿っていると考えている。また、データを格納するための体制の構築だけでなく、格納するデータの標準化を行うために遺伝子アノテーションを高速に実施する「Hayai-Annotation」の開発も行った。DBCLSと連携して、データのRDF化に関する準備も実施している。

研究者や産業コミュニティとの連携については、関連する有識者によるアドバイザー委員会を開催し、意見を収集して開発に反映させた。また、学会展示も積極的に実施し、普及やユーザー意見の収集にも努めている。現在はβ版のためまだ実施はしていないが、正規版公開後は講習会の開催などアウトリーチ活動をより活性化する計画である。

Plant GARDENの特徴の1つは多数の植物種のゲノム情報を格納し、種横断的に情報を比較できることである。この特徴を生かして、種を超えた有用遺伝子の探索を実施する情報基盤として植物研究に欠かせないDBとなるよう、引き続き開発に努めたい。

§7. 外部発表等

1. 原著論文発表

(1) 論文数概要

種別	国内外	件数
発行済論文	国内(和文)	0 件
	国際(欧文)	1 件
未発行論文 (accepted, in press 等)	国内(和文)	0 件
	国際(欧文)	0 件

(2) 論文詳細情報

1. Andrea Ghelfi, Kenta Shirasawa, Hideki Hirakawa and, Sachiko Isobe, “Hayai-An notation Plants: an ultra-fast and comprehensive functional gene annotation system in plants”, Bioinformatics, btz380, 2019 (DOI: 10.1093/bioinformatics/btz380).

2. その他の著作物(総説、書籍など)

該当なし

3. 国際学会発表及び主要な国内学会発表

(1) 概要

種別	国内外	件数
招待講演	国内	2 件
	国際	1 件
口頭発表	国内	8 件
	国際	10 件
ポスター発表	国内	8 件
	国際	3 件

(2) 招待講演

〈国内〉

1. 磯部祥子、植物ゲノムポータルサイト・PGDBj と Plant GARDEN の紹介、トーゴーの日シンポジウム、日本科学未来館、2019 年 10 月 5 日
2. 中村保一、植物ゲノムの進化研究と DDBJ と遺伝研スーパーコンピュータ活用法、木村資生進化学セミナー、御殿場、2019 年 8 月 4 日

〈国際〉

1. Hideki Hirakawa, Genome assembly of crops in Kazusa DNA Research Institute, 5th Asia-Pacific Agrobiological Genome Symposium、全州、2019 年 10 月 17 日

(3) 口頭講演

〈国内〉

1. Ghelfi A, Antezana E, Interaction between Genotype and Phenotype data, Functional Gene Annotation and Plant Breeding Ontology, 11th NBDC/DBCLS BioHackathon, Matsue, Shimane, 2018年12月14日
2. 原田大士朗、市原寿子、中谷明弘、ジェルフィアンドレア、藤代継一、小原光代、平川英樹、田畑哲之、磯部祥子、植物ゲノム情報ポータルサイト・PlantGARDENの開発と植物ゲノム解析の現状、日本育種学会第134回講演会、岡山大学(岡山)、2018年9月22-23日
3. 原田大士朗、市原寿子、中谷明弘、ジェルフィアンドレア、藤代継一、小原光代、平川英樹、田畑哲之、磯部祥子、植物ゲノム情報ポータルサイト PlantGARDENの開発、日本育種学会第135回講演会、千葉大学(千葉)、2019年3月16-17日
4. 平川英樹、原田大士朗、Ghelfi Andrea、Fawcett Jeffrey、白澤沙知子、市原寿子、中谷明弘、磯部祥子、田畑哲之、植物ゲノム情報統合ポータルサイト Plant GARDENの構築、第41回日本分子生物学会、横浜、2018年11月28-30日
5. 平川英樹、品種間ゲノムワイド多型情報の収集および Plant GARDENの構築、園芸学会平成30年度秋季大会イルミナランチョンセミナー、鹿児島、2018年9月22-24日
6. 平川英樹、育種への利用を目指したゲノムデータベースの開発、園芸学会平成31年度春季大会小集会、東京、2019年3月23-24日
7. 原田大士朗、市原寿子、中谷明弘、ジェルフィアンドレア、山田学、小原光代、平川英樹、田畑哲之、磯部祥子、植物ゲノム情報ポータルサイト Plant GARDENの拡張、日本育種学会第136回講演会、近畿大学、2019年9月6日
8. 中村保一、ゼニゴケ・高密度な実験情報の集約によるモデル植物化の加速、ワークショップ「データ駆動型植物科学最前線」第37回日本植物細胞分子生物学会大会、京都府立大学、2019年9月8日

〈国際〉

1. Ghelfi A, Hayai-Annotation: An Ultra-Fast and Comprehensive Gene Annotation System in Plants', PAG ASIA 2018, Seoul, 2018年5月31日
2. Ghelfi A, Hayai-Annotation: An Ultra-Fast and Comprehensive Gene Annotation System in Plants, Argonne National Laboratory, Chicago, 2018年7月11日

(4) ポスター発表

〈国内〉

1. 市原 寿子, 菊地 正隆, 長谷川 舞衣, 小原 光代, 平川 英樹, 磯部 祥子, 田畑 哲之, 中谷 明弘、種を超えたゲノム情報のためのデータリンク基盤の構築、トーゴーの日シンポジウム2017、東京大学弥生講堂、2017年10月4日
2. 市原 寿子, 菊地 正隆, 長谷川 舞衣, 小原 光代, 平川 英樹, 磯部 祥子, 田畑 哲之, 中谷 明弘、Construction of a data link system for integrating genomic information across species、第40回日本分子生物学会年会、神戸国際展示場、2017年12

月 8 日

3. 市原 寿子, 原田大士朗, Jeffrey Fawcett, 白澤沙知子, 小原 光代, 菊地 正隆, 長谷川 舞衣, 平川 英樹, 磯部 祥子, 田畑 哲之, 中谷 明弘、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、トーゴの日シンポジウム 2018、日本科学未来館、2018年10月5日
4. 市原 寿子, 原田大士朗, Jeffrey Fawcett, 白澤沙知子, 小原 光代, 菊地 正隆, 長谷川 舞衣, 平川 英樹, 磯部 祥子, 田畑 哲之, 中谷 明弘、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、第 41 回日本分子生物学会年会、パシフィコ横浜、2018年11月30日
5. Ghelfi A, Shirasawa K, Hirakawa H, Isobe S, Hayai-Annotation: five levels of high-accurate ultra-fast gene annotation in plants, 第 134 回講演会(岡山), 2018年9月22日
6. 原田大士朗、市原寿子、中谷明弘、ジェルフィアンドレア、藤代継一、小原光代、平川英樹、田畑哲之、磯部祥子、世界における植物ゲノム解析の現状と課題、トーゴの日シンポジウム 2018、日本科学未来館(東京)、2018年10月5日
7. 原田大士朗、市原寿子、中谷明弘、ジェルフィアンドレア、藤代継一、小原光代、平川英樹、田畑哲之、磯部祥子、植物ゲノム情報ポータルサイト・PlantGARDEN の開発にむけて、第 41 回日本分子生物学科年会、パシフィコ横浜(神奈川)、2018年11月28-30日
8. 平川英樹、原田大士朗、Andrea Ghelfi, Jeffrey Fawcett、白澤沙知子、市原寿子、中谷明弘、磯部祥子、田畑哲之、植物ゲノム統合ポータルサイト Plant GARDEN の構築、トーゴの日シンポジウム 2018、東京、2018年10月5日
9. 原田大士朗、市原寿子、中谷明弘、ジェルフィアンドレア、山田学、小原光代、平川英樹、田畑哲之、磯部祥子、植物ゲノム統合データベース Plant GARDEN の開発、トーゴの日シンポジウム 2019、東京、2019年10月5日
10. 市原寿子、磯部祥子、平川英樹、原田大士朗、Andrea Ghelfi、小原光代、山田学、白澤沙知子、中村保一、田村卓郎、杉原英志、田畑哲之、中谷明弘、種を超えた植物ゲノム情報統合のためのデータリンク基盤の構築、トーゴの日シンポジウム 2019、東京、2019年10月5日

〈国際〉

1. Fawcett JA, Hirakawa H, Kohara M, Ghelfi A, Fujishiro T, Ichihara H, Nakaya A, Isobe S, Tabata S, Developing a Plant Genome Database PGDBj in the NGS Era, Animal Genome XXVI Conference, P0022, San Diego, California, 2018年1月13-17日
2. Ghelfi A, Shirasawa K, Isobe S, Hosokawa, Development of a New Pipeline for Haplotype-Specific Expression: case-study in F1 reciprocal cross in pepper, International Society for Computational Biology, Chicago, 2018年7月9日
3. Harada D, Ichihara H, Nakaya A, Ghelfi A, Yamada M, Kohara M, Hirakawa H, Tabata S and Isobe S, Development of a Plant Genome Portal Site, Plant GARDEN, PAG Asia 2019, Shenzhen, China, 2019年6月6-8日

4. 知財出願

該当なし

5. 受賞・報道等

(1) 受賞

該当なし

(2) メディア報道

1. 日経バイオテック、植物ゲノム DB を統合した「Plant GARDEN」公開、2019 年1月時点で解析された植物ゲノムは 380 種超、オンライン 2019 年 3 月 9 日、雑誌掲載 2019 年 4 月、<https://bio.nikkeibp.co.jp/atcl/news/p1/19/03/18/05401/>

(3) その他

植物関連学会での出展による広報活動

〈国内〉

1. 第 35 回 日本植物細胞分子生物学会(埼玉)、2017 年 8 月 29～31 日
2. 日本育種学会 第 132 回講演会(盛岡)、2017 年 10 月 7～8 日
3. 2017 年生命科学系学会合同年次大会 (ConBio2017) (神戸)、2017 年 12 月 6～9 日
4. 日本育種学会 第 133 回講演会(福岡)、2018 年 3 月 25～26 日
5. 第 59 回 日本植物生理学会年会(北海道)、2018 年 3 月 28～30 日
6. 第 36 回 日本植物細胞分子生物学会(金沢)、2018 年 8 月 26～28 日
7. 日本育種学会秋季大会 第 134 回講演会(岡山)、2018 年 9 月 22～23 日
8. 園芸学会平成 30 年度秋季大会(鹿児島)、2018 年 9 月 22～23 日
9. 第 41 回 日本分子生物学会年会(横浜)、2018 年 11 月 28～30 日
10. 第 60 回 日本植物生理学会年会(名古屋)、2019 年 3 月 13～15 日
11. 日本育種学会春季大会 第 135 回講演会(千葉)、2019 年 3 月 16～17 日
12. 園芸学会平成 31 年度春季大会(川崎)、2019 年 3 月 23～24 日
13. 日本育種学会 第 136 回講演会(奈良)、2019 年 9 月 6～7 日
14. 第 37 回 日本植物細胞分子生物学会(京都)、2019 年 9 月 7～8 日
15. 園芸学会令和元年度秋季大会(島根)、2019 年 9 月 15～17 日

〈国際〉

1. Plant & Animal Genome XXVI (サンディエゴ 米国)、2018 年 1 月 13～17 日
2. PAG Asia 2019 (深圳、中国)、2019 年 6 月 6～8 日

§8. 研究開発期間中の活動

1. 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要
2017年 8月3日	平成29年度第一回統合 化推進プログラムアドバ イザリー委員会	東京ステーシ ョンコンファレ ンス	25人	プロジェクトの開始にあたって 10名の外部アドバイザーより計 画に対して意見をいただく会議
2017年 10月4日	担当者ミーティング	東京大学	7人	プロジェクトの進捗報告と協議 を行う担当者会議。トーゴの 日にあわせて実施した。
2018年 1月26日	担当者ミーティング	東京ステーシ ョンコンファレ ンス	7人	プロジェクトの進捗報告と協議 を行う担当者会議
2018年 2月8日	第2回アドバイザー委 員会	東京ステーシ ョンコンファレ ンス	22人	2017年度の進捗に対して外部 アドバイザーより意見をいただく 会議
2018年 5月17日	担当者ミーティング	TKP 品川カ ンファレンス センター	10人	研究進捗報告と今後の進め方 の協議のためのミーティング
2018年 9月3日	平成30年度第1回アド バイザリー委員	ステーション コンファレン ス東京	23人	外部アドバイザー委員と進捗 に対する意見交換を行うための ミーティング
2019年 2月18日	平成30年度第2回アド バイザリー委員会	ステーション コンファレン ス東京	25人	2017年度の進捗に対して外部 アドバイザーより意見をいただく 会議
2019年 4月15日	担当者ミーティング	TKP 品川カ ンファレンス センター	12人	研究進捗報告と今後の進め方 の協議のためのミーティング
2019年 8月7日	2019年度第1回アドバ イザリー委員会	ステーション コンファレン ス東京	25人	外部アドバイザー委員と進捗 に対する意見交換を行うための ミーティング

2. 主催したワークショップ、シンポジウム、アウトリーチ活動等

年月日	名称	場所	参加人数	目的・概要
2018年 3月9日	植物ゲノム情報活用に向 けたデータベース利用法	フクラシア東 京ステーショ ン会議室	85人	PGDBj ユーザーの獲得増加を 目指し、PGDBj、DDBJ の利用 法とゲノム配列・マーカー解析 法の講習会を実施した。

以上