



# 植物ゲノム情報ポータルサイト 「Plant GARDEN」の改訂 (2022年度・第2四半期版)

○市原 寿子<sup>1</sup>, 平川 英樹<sup>1</sup>, 山田 学<sup>1</sup>, 小原 光代<sup>1</sup>, 山下 サマツチャヤー<sup>1</sup>,  
白澤 沙知子<sup>1</sup>, 戸田 陽介<sup>1</sup>, 中村 保一<sup>1,2</sup>, 七夕 高也<sup>1</sup>, 田畑 哲之<sup>1</sup>, 磯部 祥子<sup>1</sup>

1. (公財) かずさDNA研究所、 2. 国立遺伝学研究所



# 要旨

近年、多数の植物種のゲノム配列が次々に決定され、また、一つの植物種でも品種やアセンブリバージョンが異なる多数のゲノムが公開されている。**Plant GARDEN** (Genome And Resource Database Entry ; <https://plantgarden.jp>) は、多数の研究機関から公開されている多種のゲノムデータを一元化し、目的の情報へ簡単にアクセスできるように開発された植物ゲノムポータルサイトである。

提供しているコンテンツは、ゲノム、遺伝子の配列情報、文献からキュレーションされたDNAマーカー、QTL、連鎖地図、遺伝子機能センテンスの情報、SRAデータから算出された塩基多型の情報、及び、これらを用いてユーザーが保有する情報を解析するためのツールである。

文献に由来するコンテンツの効率的な更新、拡充のため、キュレーション情報の共有に際し、担当者間の記述表現の揺らぎやミスを最小限に抑えるためのデータ入力システムを構築した。また、これまでのパソコン用に加え、タブレット端末からのアクセスに適したタブレット端末用インターフェースを開発し、2022年5月に公開した。本発表では、これらをはじめとした開発状況を紹介する。

# 格納されている植物種(他)の種類と数

	Order (目)/件	Species (種)/件
被子植物	33	303
裸子植物	≥3	11
シダ植物	2	2
ゼニゴケ植物	1	1
マゴケ植物	3	4
緑藻植物	5	8
紅藻植物	2	2
キノコ(担子菌門)	1	1

- 裸子植物のマツ綱に属する種に、目名候補が複数ある等不明なものがあった。
- Plant GARDENの中で裸子植物に属する綱は3つ→その下の目は3つ以上
- キノコは植物ではないが、Plant GARDENから公開

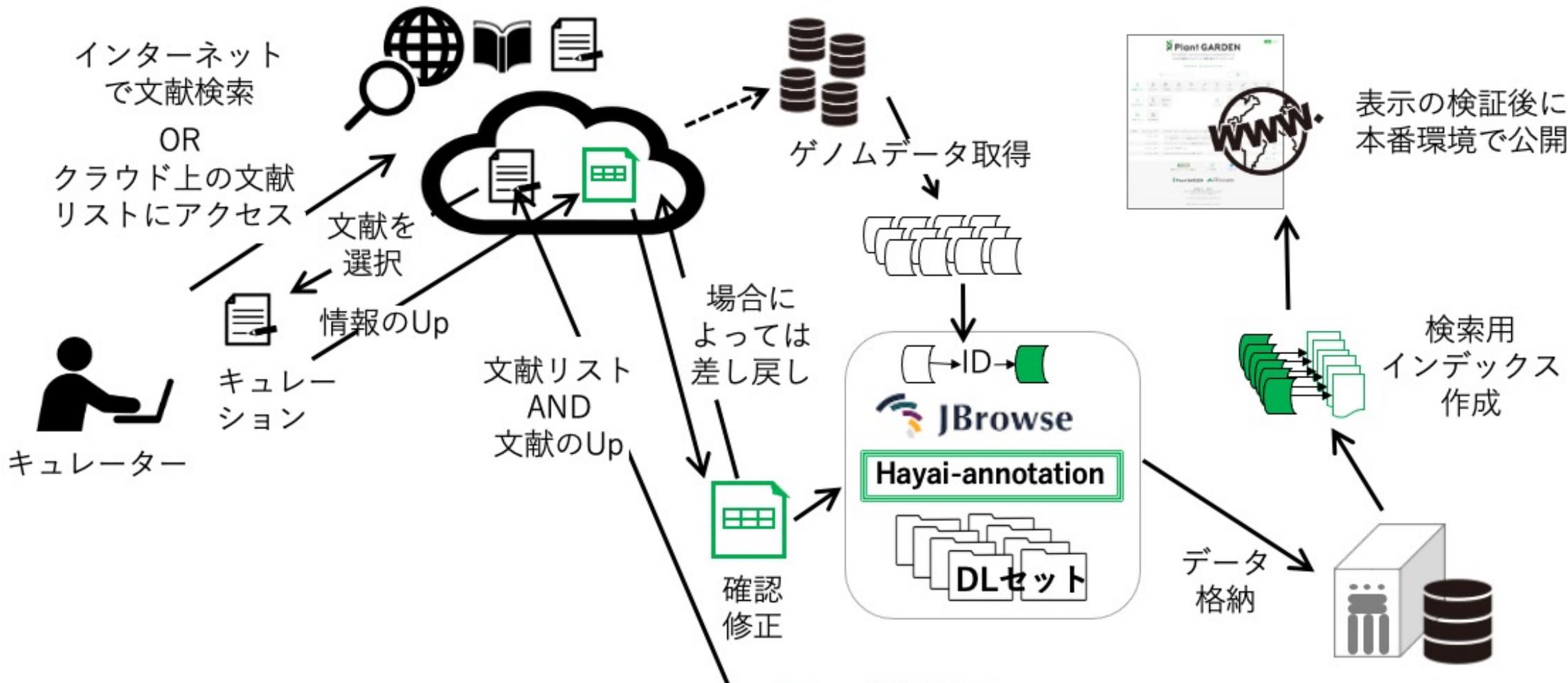


# 主要コンテンツの公開数（2022.9月現在）

データタイプ	生物種数	データ件数
ゲノム	184	236
遺伝子配列	160	10,948,303
VCF (Variant Call Format) ファイル	106	107
gVCF (Genomic Variant Call Format)ファイル	106	5,981
DNA マーカー	61	373,596
形質連鎖 DNA マーカー (QTL)	28	8,255

- かずさDNA研で解読されたもの
- ユーザーが解析し、掲載依頼のあったもの
- ユーザーから掲載要望のあったもの

# キュレーションコンテンツが公開されるまでの流れ



公益財団法人  
**かずさDNA研究所**

# キュレーション体制

## メンバー構成

かずさメンバー（4名）  
外部キュレーター（10名）

## 対象情報

DNAマーカー、QTL、遺伝子、連鎖地図  
ゲノム情報

## 現在までの対象種

3 3 2 生物種

## 着手済み論文数／収集済み論文数

1,120件／2,244件

# キュレーションデータDB登録時の問題点と対策

## データベース格納時のUNIX処理でトラブルの原因となる文字やコードの存在

- pdfファイルからのコピー&ペースト操作での文字化け
- pdfファイル由来のテーブルに含まれるセル内の改行文字
- 不適切な空白文字
- スプレッドシート→エクセル→テキスト変換の過程で挿入される文字コード

## データ入力時の表記ゆらぎ

- ルールを決めていても、間違えてしまうことがある



## キュレーションデータの入カシステムを開発

- Webブラウザを介して、指定したフォームへ情報を入力→クラウドDBへ格納
- 不正文字の一部を自動修正する。
- 問題のある入力に対して、受け付けずにエラーメッセージを出し、修正を促す。
- 途中のステップを減らし、不正文字が挿入される機会を下げる。
- 選択肢の設定により、表記揺らぎの頻度を下げる。

# キュレーションデータが反映されるページの例

Plant GARDEN

Lotus japonicus

この種について

科名: Fabaceae  
属名: Lotus  
学名: Lotus japonicus  
Taxonomy ID: 34305

グノム種別を見る

- Lj3.0Miyakojima MG-20
- Lj3.0Miyakojima MG-20

遺伝子を見る

- Lj3.0Miyakojima MG-20
- Lj3.0Miyakojima MG-20

全植物種リスト

学名 ↓	和名 ↓	科名 ↓
Acer truncatum	マンジュウイタヤ	ムクロジ科
Acer yangbiense	カエデ	ムクロジ科
Actinidia chinensis	キウイフルーツ (chinensis属)	マタタビ科
Actinidia eriantha	キウイフルーツ (eriantha属)	マタタビ科
Aegilops tauschii subsp. strangulata	タルホコムギ	イネ科
Amaranthus hypochondriacus	アマランサス	ヒユ科
Ananas comosus	パイナップル	パイナップル科
Ananas comosus var. F153	パイナップル	パイナップル科
Antirrhinum majus	キンギョソウ	オオバコ科
Apium graveolens	セロリ	セリ科
Arabidopsis thaliana	シロイヌナズナ	アブラナ科
Arabis alpina	ヤマハタザオ	アブラナ科
Arachis duranensis	ラッカセイ野生種 (duranensis属)	マメ科
Arachis hypogaea	ラッカセイ	マメ科
Arachis ipaensis	ラッカセイ野生種 (ipaensis属)	マメ科
Arachis monticola	ラッカセイ野生種 (monticola属)	マメ科
Asparagus kiusianus	ハマタマボウキ	クサカズラ科
Asparagus officinalis	アスパラガス	キジカクシ科
Bathycoccus prasinos	パチコッカス	パチコッカス科
Benincasa hispida	トウガン	ウリ科
Beta vulgaris subsp. vulgaris	テンサイ	ヒユ科

Lotus japonicus

配列名: Lj3.0

グノム配列の詳細

配列名	系統名	系統名	系統名
配列名	Lj3.0	系統名	Miyakojima MG-20
配列長 (bp)	447,416,816	染色体数	2n = 2x = 12
遺伝子の数	48,105	NSO長 (bp)	62,285,374
Hayai-annotationにより注釈づけられた遺伝子の数	48,105	遺伝子数由来ファイル	Lj3.0_pep.fna
シーケンシングの方法	Sanger, Illumina, 454	遺伝子数由来ファイル	zen_v2.0
アセンブリ方法	Paracel Genome Assembler	取得した配列量	35x
シーケンシングの方法のコメント		推定グノムサイズ (Mb)	465
論文 (DOIコード)	10.1093/dnares/dsn008	コメント	
データソース名	Lotus japonicus Genome Sequencing Project	責任者	Satoshi TABATA (Kazusa DNA Research Institute)
BUSCO バージョン	5.1.0	データソースURL	http://www.kazusa.or.jp/lotus/
BUSCO グノム	C:93.5%(S:86.9%,D:6.6%),F:3.7%,M:2.8%,n:1614	BUSCO データセット	embryophyta_odb10
		BUSCO 遺伝子 (pep/cds)	C:79.8%(S:59.5%,D:20.3%),F:11.9%,M:8.3%,n:1614

← → https://plantgarden.jp/ja/download/t34305/t34305.G002

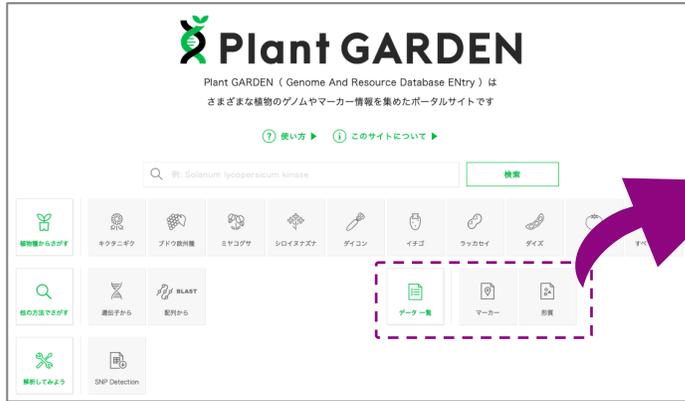
Index of /ja/download/t34305/t34305.G002

- Parent Directory
- Lj3.0\_cDNA.ffn.gz
- Lj3.0\_cds.ffn.gz
- Lj3.0\_gene\_models2.gff3.gz
- Lj3.0\_pep.fna.gz
- Lj3.0\_pseudomol.fna.gz
- Lotus japonicus t34305.G002\_zen\_v2.0.tar.gz
- SRA/



# Plant GARDENから取得できるデータの種類の見方

## トップページ



TOP > データ一覧

植物種名  頭文字に飛ぶ

学名 ↓	和名 ↓	科名 ↓	参照配列			
			ゲノム	遺伝子 (cds / transcript / gene)	アノテーション情報 (Hayai Annotation ZEN)	SRA (Sequence Read Archive) タンパク (pep / aa)
 <i>Musa balbisiana</i>	バナナ野生種 (balbisiana種), リュウキュウバショウ	バショウ科	●	●	●	●
 <i>Neopyropia yezoensis</i>	スサビノリ	ウシケノリ科	●			
 <i>Nicotiana attenuata</i>	タバコ野生種 (attenuata種), コヨーテタバコ	ナス科	●	●	●	●

2022年9月現在、*Neopyropia yezoensis* (スサビノリ)のゲノム情報は、アセンブリ配列のみが公開されており、アノテーション情報は未公開

- ゲノム列にのみ●
- cdsやpepは空欄

# アセンブリ配列とアノテーションの情報の公開パターン

## 1. アセンブリ配列 & アノテーション情報の両方を一つのDBで一緒に公開

"Genome assembly and annotation have been deposited in GenBank under BioProjects PRJNA680555 ('Oaxaca'), PRJNA680556 ('Pawnee'), PRJNA680557 ('Lakota'), and PRJNA680558 ('Elliott'). Genomes and annotations are also available through phytozome: Pawnee, Elliott, Lakota, and Oaxaca." (10.1038/s41467-021-24328-w) → NCBI has assembly and annotation data

## 2. アセンブリ配列とアノテーション情報を、別々のDBやウェブサイトから公開

"The coast redwood v2.2 assembly was deposited at NCBI as GenBank accession VDFB02000000, BioProject PRJNA542879. The genome assembly, annotation, and functional assessment are also available in the TreeGenes database (<https://treegenesdb.org/>)."  
→ NCBI has the assembly data and TreeGenes database has the annotation data.

## 3. アセンブリ配列のみを一つのDBから公開

"Genomic assembled sequences, ONT raw reads, and raw short reads have been deposited to NCBI database under Bioproject accession no. PRJNA781352." (10.1093/gbe/evac060) → NCBI has assembly data.

## 4. 論文に記述されているIDなどが見つからず、データを取得できない

"Genome sequences, assembly, and annotation data have been deposited at NCBI GenBank under BioProject/BioSample numbers PRJNA727440/SAMN19020793." (10.1016/j.xplc.2021.100247) → No corresponding data in NCBI

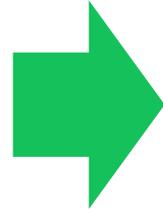
## 5. 論文に記述されているリンクが機能しておらず、データを取得できない

"The genome assembly file is available at NCBI. All the annotation tables containing results of the draft genome analysis are available at figshare" (doi.org/10.6084/m9.figshare.14790132) → NCBI has assembly data, but annotation data was inaccessible because the link was dead.

オリジナルで分かれて登録されているデータをPlant GARDENから一つのセットとして取得可

# タブレット端末用のウェブページを公開

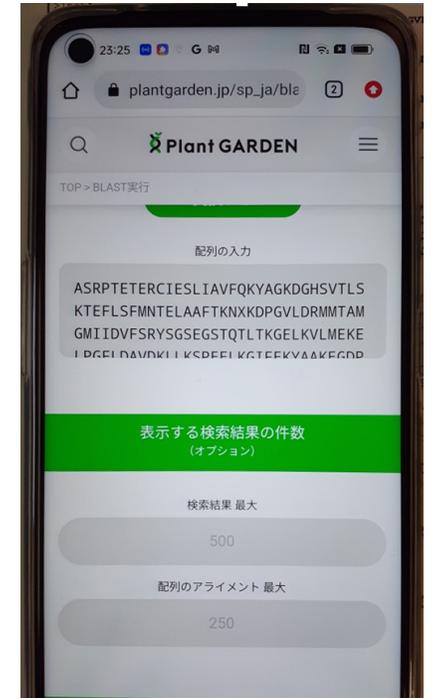
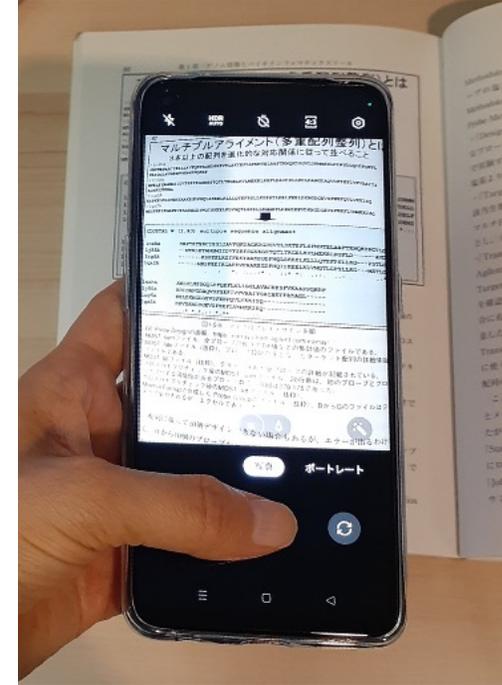
## パソコン画面表示



## スマートフォン画面表示



- 身近な情報収集のツールとして日常的に使用されるスマートフォンから、Plant GARDENにアクセスしやすくするインターフェースを開発した。
- 検索先として、**植物学の正確な情報の基盤**となることで、タブレット端末の**学習ツール**としての可能性を高めることが期待できる。

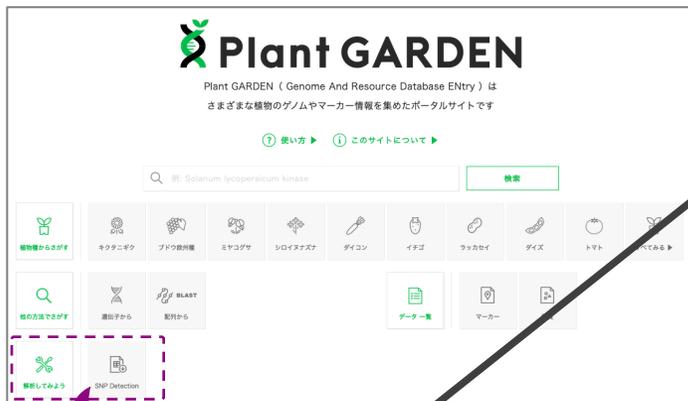


活用例：

✓タブレット端末のカメラ機能を使って紙媒体から配列を取得→配列相同性を検索

# 解析ツールを追加開発

トップページ



解析してみよう

SNP Detection

- ダウンロード版の変異検出のための解析パイプライン

Plant GARDENのデータを参照してユーザーの植物ゲノムデータをローカルで解析できる。

- コンテナ技術(Docker)を用いて、解析パッケージを一括インストールして使用。

解析ツール一覧のページへ

解析ツールを選ぶ

- お持ちのデータをアップロードしてブラウザ上で解析する
- 解析ツールをダウンロードしてローカル環境で解析する
- ANPLAT社 ANCATプラットフォームで解析する

SNP Detection	SNPやIn/Delを検出する解析パイプライン	
SNP Workflow (bwa)	Docker-compose版 SNP解析パイプライン	
RNA-Seq Workflow (hisat2)	Docker-compose版 RNA-seq解析パイプライン	
JBrowseContainer	Docker-compose版 JBrowseセットアップ/実行ツール	
Hayai-Annotation	遺伝子配列のアノテーション	
Hayai-Annotation Plants v2.0	植物種に特化した機能アノテーションシステム	
KusakiDB v1.0	タンパク質同族体群の検証と完全性のための新しいアプローチ	
merge-gvcf	複数個のvcf(variant call format)またはgVCF(genomic VCF)ファイルを1つに統合するツール	
gz2bgz	ドラッグドロップで、bcftools 等で用いるBgzip形式のファイルに変換するツール	

Licensed under a Creative Commons 表示4.0国際ライセンス  
©2022 市原 寿子 ( (公財) かずさDNA研究所)



# Plant GARDENデータとユーザーデータを比較解析する仕組み

## Plant GARDEN



- 多種多様な植物ゲノムを保有

アクセス



データの取得



- Plant GARDENのゲノム配列情報などとユーザーデータを比較し、SNP検出やRNA-Seq解析などを実施するプラットフォーム
- コンテナ化されたツールイメージからローカルに解析環境をセットアップして起動。
- セットアップおよびプログラム起動後は、ウェブブラウザ上で解析操作できる。

- 「Mi-GARDEN」をクラウド上で実施できる有料のシステム
- 面倒なセットアップが不要で、普段使いのPCでは難しい大きなデータも解析可能。
- (株) ANPLAT社が提供するサービスを介してシステムを公開。

# まとめ

1. ゲノムデータの追加を実施した。被子植物の分類体系である**APG IVに基づいて見た場合、半数以上の目をカバー**できた。カバーしきれていない目の中には、絶滅危惧種などの希少植物であるためにゲノム解読が困難であるなどの理由で、今後もカバーしきれない可能性の高いものも含まれる。
2. 複数の異なるデータベースに分けられて登録されているゲノムデータセットについて、可能なかぎり**Plant GARDENから一括で取得**できるようにした。
3. Plant GARDENに格納されているデータをリファレンスとして、ユーザーが保有するNGSデータを解析するための仕組みを新たに公開した。これまでは、かずさDNA研究所の解析サーバー上にデータをアップロードして解析する必要があった。今回追加されたのは、**ユーザー自身が解析環境をローカルマシン上でセットアップ**するものと、**セットアップ不要でクラウド上で解析できるサービス**である。
4. これまでのパソコン用に加え、タブレット端末からのアクセスに適した**タブレット端末用インターフェース**を開発し、公開した。

# 謝辞

## ■ キュレーター（敬称略, アルファベット順、50音順）

Koilkonda Padmalatha

Lau Nyok Sean

池崎 由佳

磐佐 まりな

岡田 眞銀

嶋本 育泰

嶋 羊子

畠山 剛臣

文屋 慧亮

三輪 晃敬

いつも大変丁寧なキュレーション業務を実施くださいます。  
誠に有難うございます。