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）
研究開発実施報告書　様式

研 究 開 発 実 施 報 告

□概要

研究開発課題名	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
開発対象データベースの名称（URL）	PGDBj (Plant Genome DataBase Japan) (http://pgdbj.jp/)
研究代表者氏名	田畑 哲之
所属・役職	かずさ DNA 研究所　　所長 (2019 年 3 月時点)

□目次

§1. 研究実施体制

グループ名	研究代表者または主たる共同研究者氏名	所属機関・役職名	研究題目
かずさDNA研究所グループ	田畑　哲之	かずさDNA研究所・所長	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、PGDBjコンテンツの更新
大阪大学グループ	中谷　明弘	大阪大学・特任教授	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

§2. 研究開発対象とするデータベース・ツール等

（1）データベース一覧

【主なデータベース】

No.	名称	別称・略称	URL
1	Plant GARDEN	Plant GARDEN	http://plantgarden.jp

【その他のデータベース】

No.	名称	別称・略称	URL
1	Plant Genome DataBase Japan	PGDBj	http://pgdbj.jp
2	PGDBj オルソログデータベース		http://pgdbj.jp/od3/
3	PGDBj DNA マーカー・QTL データベース		http://pgdbj.jp/markerdb/marker.html?sbj=m&ln=ja
4	PGDBj カンキツリソースデータベース		http://pgdbj.jp/estui/citrus/CR.html
5	ゲノム解読状況データベース		http://pgdbj.jp/plantdb/plantgenome.html
6	Strawberry GARDEN		http://strawberry-garden.kazusa.or.jp/
7	Sweetpotato GARDEN		http://sweetpotato-garden.kazusa.or.jp/
8	Carnation DB		http://carnation.kazusa.or.jp/
9	Zoysia Genome Database		http://zoysia.kazusa.or.jp/
10	Eggplant Genome DataBase		http://eggplant.kazusa.or.jp/
11	Raphanus sativus Genome DataBase		http://radish.kazusa.or.jp/
12	Buckwheat Genome DataBase (BGDB)		http://buckwheat.kazusa.or.jp/
13	Eucalyptus camaldulensis Genome Database		http://www.kazusa.or.jp/eucaly/
14	Jatropha Genome Database		http://www.kazusa.or.jp/jatropha/
15	CloverGarden		http://clovergarden.jp
16	Lotus japonicus Genome Sequencing Project		http://www.kazusa.or.jp/lotus/
17	Kazusa Marker DataBase		http://marker.kazusa.or.jp/
18	Tomato Functional SNP DataBase		http://plant1.kazusa.or.jp/tomato/

（2）ツール等一覧

No.	名称	別称・略称	URL
1	PGDBj 横断検索システム		http://pgdbj.jp
2	PGDBj 育種向け DNA マーカーページ		http://pgdbj.jp/pages/index.html?dir=ag&page=menu&ln=ja
3	Hayai-Annotation Plants		https://github.com/kdri-genomics/Hayai-Annotation-Plants
4	SNP Detection		https://pgdbjsnp.kazusa.or.jp/
5	ASE-pipeline		未公開

§3. 実施内容

（1）本年度の研究開発計画と達成目標

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。第２年次では、前年度に設計したスキーマと JBrowse によるゲノムビューアを軸に PGDBj の次バージョン（仮称・PGDBj 2）の細部ページの設計をすすめ、年度内の試験公開を目指す。前年度に動作や表示を検討した JBrowse のプラグインを DB に組み込み、まずはかずさ DNA 研究所でこれまで開発したゲノム DB のうち情報量が多く Pseudomolecules が整備されているミヤコグサ、トマトおよびイチゴを対象として、互いのゲノム配列情報、遺伝子、transcripts 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、比較できる基盤 DB を構築する。また、「（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」とも連携して Pseudomolecules が構築されている数種のゲノム情報を相互に表示するためのシステム開発を行う。

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、複数の生物種に跨が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。第２年次では前年度に引き続き、かずさ DNA 研究所グループが新たに選択した、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、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

本項目ではユーザ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。第２年次は前年度導入した SNP 検出用サーバに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制限つき公開する。外部からのデータは専用に設置した回線を通じて受け付ける。また、Copy number variation(CNV)や structure variation (SV)を検出するパイプラインの構築を開始する。一方、ユーザ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、遺伝子予測やアノテーションを高速で実施する解析パイプラインもあわせて開発する。開発するパイプラインは SNP 解析パイプラインと組み合わせて３年次以降に SNP アノテーションや RNA-Seq データを用いたアレル特異的発現解析を実施するツールへと発展させる予定である。

一方、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。そこでプロジェクト年度の後半にはリファレンス配列と解析ツールを PGDBj2(仮称)よりダウンロードし、ローカルで SNP 検出を実施、得られた vcf ファイルを Upload して他データと比較するシステムへと発展させることを計画している。そのための準備としてローカルで SNP 検出を実施するため計算量を抑えた SNP 検出プログラムの開発も着手する。

④ PGDBj コンテンツの更新

前年度までに設計したスキーマに合わせ、現 PGDBj に格納されているデータを整理し、コンテン

ツの更新を継続して実施する。また、前年度に引き続き 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行う。さらに全ゲノム配列の解読手法やゲノムワイドな多型情報、RNA-Seq 情報など新たな種類のコンテンツの収集を継続して実施すると共にキュレーションを効率的に実施するため、センテンスキュレーションの導入を進める。また、新たに解読された植物種のうち特に重要な作物や要望があった植物を掲載する。また、PGDBj を永続的に運用するため、前年度までは Line 社に PGDBj の管理を委託していたが、その手法を引き継ぐことで、かずさ DNA 研究所内で運用できる体制を整える。また、SNP 解析用サーバーや新たな PGDBj2 用サーバーの構築を引き続き実施し、テストサーバも構築することで PGDBj2 の試験公開を開始する。

（2）進捗状況

①　ゲノム横断的関連情報表示サイトの開発

新しく作成する基盤 DB（昨年度までの仮称・PGDBj2）の名称を「Plant GARDEN（Genome And Resource Database Entry）」として、2019 年度 3 月にβ版を公開することで準備を進めた。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、昨年度までに策定したスキーマ（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。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」「形質との連関をさがす」「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、バージョンごと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。

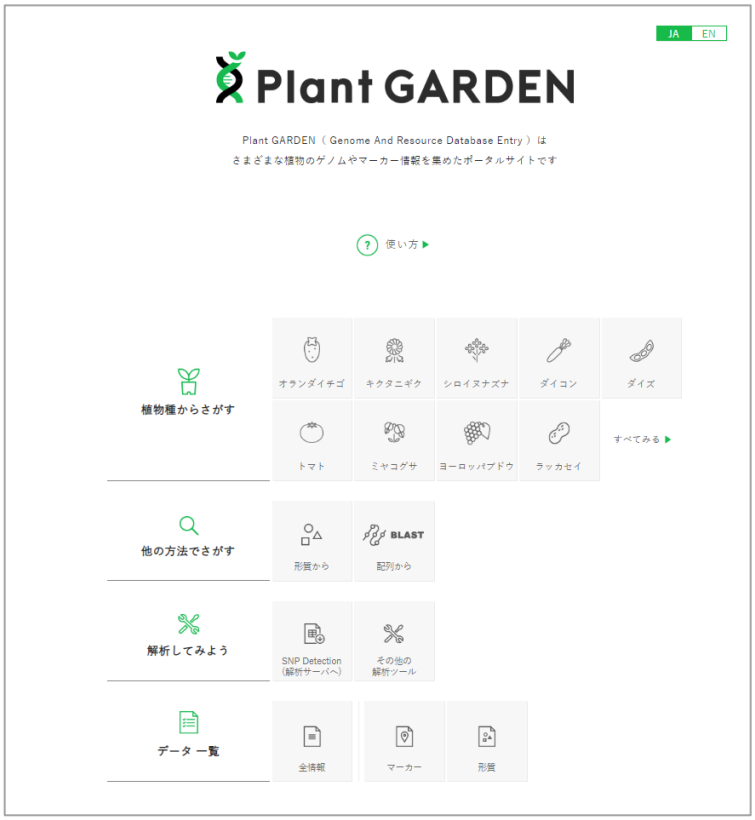

図　Plant GARDEN のトップ画面

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。ミヤコグサについては全情報、その他の 8 種（オランダイチゴ、キクタニギク、シロイヌナズナ、ダイコン、ダイズ、トマト、ヨーロッパブドウ、ラッカセイ）

については全ゲノム配列のみ格納したβ版を 2019 年 3 月 11 日に公開した。

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

第 1 年次に引き続き、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-leipzig.de/Software/proteinortho)と OrthoFinder(http://www.stevekellylab.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 Uni Prot 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。

③ カスタム型多型・ハプロタイプ検出システムの構築

前年度に購入した SNP 検出用サーバのセキュリティ対策を実施したのち、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制限つきで一部ユーザーに公開した。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。

さらに新たなキラーアプリケーションとして、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。これは 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の「Hayai-Annotation ZEN」とは異なり内部に実装する DB は植物遺伝子に特化している。また、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。「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 から公開した（https://github.com/kdri-genomics/Hayai-Annotation-Plants）。

④ PGDBj コンテンツの更新

PGDBj から Plant GARDEN へ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、前年度に引き続き 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。

一方、キュレーションを効率的に実施するため、センテンスキュレーションの検討を実施した。

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）、NBCI 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確認した結果、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。

§4. 成果発表等

（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）	0 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

該当なし

（2） その他の著作物（総説、書籍など）

1. 該当なし

（3） 国際学会および国内学会発表

① 概要

種別	国内外	件数
招待講演	国内	0 件
	国際	0 件
口頭発表	国内	6 件
	国際	2 件
ポスター発表	国内	6 件
	国際	1 件

② 招待講演

該当なし

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

〈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, 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、トーゴーの日シンポジウム 2018、日本科学未来館、10 月 5 日
2. 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、11 月 30 日
3. 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
4. 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、トーゴーの日シンポジウム 2018、日本科学未来館（東京）、10 月 5 日
5. 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
6. 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、トーゴーの日シンポジウム 2018、東京、2018 年 10 月 5 日

〈国際〉

1. 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

植物関連学会での出展による広報活動

〈国内〉

1. 第 36 回 日本植物細胞分子生物学会（金沢） 2018 年 8 月 26～28 日
2. 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
3. 園芸学会平成 30 年度秋季大会（鹿児島） 2018 年 9 月 22～23 日
4. 第 41 回 日本分子生物学会年会（横浜） 2018 年 11 月 28～30 日
5. 第 60 回 日本植物生理学会年会（名古屋） 2019 年 3 月 13～15 日
6. 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
7. 園芸学会平成 31 年度春季大会（川崎） 2019 年 3 月 23～24 日

（4）知的財産権の出願　出願件数

該当なし

（5）受賞・報道等

該当なし

§5. 研究開発期間中に主催した活動（ワークショップ等）

1. 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要
2018 年 5 月 17 日	担当者ミーティング（非公開）	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（非公開）	ステーションコンファレンス東京	23 人	外部アドバイザリ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（非公開）	ステーションコンファレンス東京	25 人	同上

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

該当なし

以上

別紙1　既公開のデータベース・ウェブツール等

No.	正式名称	別称・略称	概要	URL	公開日	状態	分類	関連論文
1	Plant Genome Database Japan	PGDBj	植物ゲノム関連情報を統合化するハブとして構築したポータルサイトである。進化情報、リソース情報、ゲノム上の位置や構造情報を軸に遺伝子機能等を検索できる。横断検索を用いることで植物に特化した多種類のDBへ効率的にアクセスできる。	http://pgdbj.jp	2012/8/20	維持・発展	データベース等	1. Asamizu E, Ichihara H, Nakaya A, Nakamura Y, Hirakawa H, Ishii T, Tamura T, Fukami-Kobayashi K, Nakajima Y, Tabata S., Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases.,Plant Cell Physiol. 55(1):e8 (2014) 2. Nakaya A, Ichihara H, Asamizu E, Shirasawa S, Nakamura Y, Tabata S, Hirakawa H., Plant Genome DataBase Japan (PGDBj)., Methods Mol Biol. 1533:45-77 (2017)
2	Plant GARDEN		植物ゲノム関連情報を格納したポータルサイトである。PGDBjでは格納されていなかった全ゲノム配列情報を基軸に、植物種毎に情報を閲覧できるページを基軸としている。他に遺伝子配列、アノテーション、PCRベースのDNAマーカー、SNPs、形質連関マーカー等の情報を格納し、ゲノムブラウザ（Jbrowse）上で横断的に情報を検索することも可能である。また、異なる植物種間で類似する遺伝子配列を検索することも可能である。現在はβ版として公開中である。	https://plantgarden.jp	2019/3/11	新規	データベース等	
3	Hayai-Annotation Plants		植物を対象に遺伝子機能アノテーションを実施するツール。ローカル環境で動作し、実行速度が極めて速く、正確かつ包括的なアノテーションが可能である。	https://github.com/kdri-genomics/Hayai-Annotation-Plants	2018/11/20	新規	ツール等	Ghelfi A, Shirasawa K, Hirakawa H, Isobe S, Hayai-Annotation Plants: an ultra-fast and comprehensive functional gene annotation system in plants, Bioinformatics, btz380, https://doi.org/10.1093/bioinformatics/btz380
4	SNP Detection		配列の精査、マッピングおよび変異検出を実施する解析パイプラインである。Plant GARDENのユーザーを対象にβ版として公開中。FTPを通じてユーザーがリード配列とリファレンス配列をかずさDNA研究所のサーバーにUploadして解析を実行する。実行にはユーザー登録が必要である。	https://pgdbjsnp.kazusa.or.jp/	2019/3/11	新規	ツール等	