# メタボローム統合データベースMetaboBank

○長崎英樹1)、荒武1)2)、福島敦史3)4)、高橋みき子4)、大澤祥子1)、小林紀郎5)、藤澤貴智6)、時松敏明6)、福田亜沙美6)、櫻井望6)、諏訪和大7)、金谷重彦8)、平川英樹1)、有田正規4)6)

1) かずさDNA研究所・植物DNA解析グループ
2) 京都大学・生存圏研究所
3) 京都府立大学・生命環境科学研究科
4) 理化学研究所・環境資源科学研究センター
5) 理化学研究所・情報統合本部・データ知識化開発ユニット
6) 国立遺伝学研究所・生命情報DDBJセンター
7) (株) リオレクト
8) 奈良先端科学技術大学院大学・情報科学領域・計算システムズ生物学

# 要旨

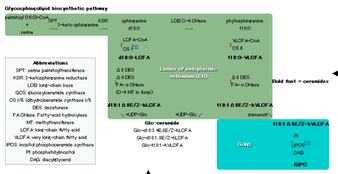
多様な代謝物をより多く同定することを目指すメタボローム解析から得られるデータは、分析装置による実験生データ、解析手法や設定を記載したメタデータなど多種多様である。メタボロームデータ統合化に向けてMetaboBankでは以下の開発を行っている。1)データレポジトリとしてBioProject、BioSample等DDBJメタデータとの連携、登録フォーマットMAGE-TABの整備、データ登録作業のシステム化。 2)メタボローム関連データベースKNApSAcK、MassBankのデータ移設、英国MetaboLightsのメタデータ追加、かずさDNA研、理研からの植物メタボロームデータのデータ移行と再解析。3) Resource Description Framework (RDF)によるデータアーカイブ化、2次データベースMetaboBank Wikiの構築。
2018年より開発が進めてられているMetaboBankは現在Ver2となり以下のサイトよりデータ公開と一般からのメタボロームデータの登録受付を開始している(https://mb2.ddbj.nig.ac.jp)。

# 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/物質循環を考慮したメタボロミクス情報基盤

## メタボローム 統合データベースMetaboBank

国立遺伝学研究所 (NIG)

代謝マップ



生物種-代謝物関係データベース
(http://www.knapsackfamily.com)

KNApSAcK

MassBank

標準マススペクトルデータベース
(http://www.massbank.jp)

メタデータ

解析で得られた代謝物の
データレポジトリ

かずさDNA研究所
(KDRI)

かずさDNA研究所

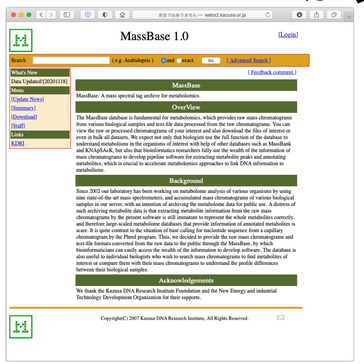
実験生データとメタデータ
を提供

RIKEN

理化学研究所

データ連携協定によりKDRIのデータセットは
理研にも提供される。


Metabolonote
http://metabolonote.kazusa.or.jp/




MassBase
http://webs2.kazusa.or.jp/massbase/


RIKEN Plant Metabolome MetaDatabase (RPMM)
http://metabobank.riken.jp/pmm/db/plantMetabolomics

# Resource Description Framework (RDF)のための自由書式の文章メタデータの意味づけ作業

## Analytical Method Details Information

| | |
|---|---|
| ID | MS1 |
| Title | LC-FT-ICR-MS ESI positive method 1 |
| Instrument | Agilent1100 HPLC (Agilent), LTQ-FT (Thermo Fisher Scientific) |
| Instrument Type | LC-FTICR-MS |
| Ionization | ESI |
| Ion Mode | Positive |
| Description<br>サンプル抽出 | Harvested sample is frozen by liquid N2 and resulting powder (100mg) are solved in 300uL 80% methanol solution. 20uL sample is injected into HPLC after 0.2um membrane filter treatment. |
| HPLC | HPLC conditions: Agilent 1100 series (Agilent), Column: TSKgel-100V (4.6 x 250 mm, 5 micrometer; TOSOH), Solvent: A; 0.1% formic acid aq. B; ACN (addition 0.1% formic acid fc.), Gradient: (B);3 to 30% (0.0 to 25.0 min), 30 to 90% (25.0 to 40.0 min), 90% (40.0 to 45.0 min), 95% (45.1 to 50.0 min), 3% (50.1 to 57.0 min), Column temp.: 30 degree C, Flow rate=0.5mL/min, PDA: 200-650 nm (2 nm step). |
| マス測定 | FT-ICR-MS conditions: Filter 1: FTMS + c norm !corona !pi res=50000 o(200.0-1500.0); 2: ITMS + c norm !corona !pi Dep MS/MS Most intense ion from (1).;3: ITMS + c norm o(200.0-1500.0)., Rejected mass=266.0000;294.0000;391.0000. |

| | |
|---|---|
| 植物のみ | 85 studies |
| その他生物 | 11 studies |
| 総数 | 96 studies |
| (食品 | 16 studies登録中) |
| 理研データ | 59 studies |

共通の述語(クラス)で関連づけられたRDF化用の50シートのエクセルファイルに決まった項目にデータを当てはめる意味づけ作業
理研は論文よりデータ抽出

| Chromatography | comment | description | temperature gradient | column type | column temperature | column pressure | column name |
|---|---|---|---|---|---|---|---|
| クロマトグラフィ | コメント | 説明 | 温度勾配 | カラムの種類 | カラム温度 | カラム圧力 | カラム名 |
| | rdfs:comment | dcterms:description | temperatureGradient | columnType | columnTemperature | columnPressure | columnName |
| Chromatography | rdf:langString | rdf:langString | xsd:string | xsd:string | Temperature | Pressure | xsd:string |
| pm_chromato:MN_SE2_HPLC | "Elute monitoring by PDA equipped with Agilent1100 HPLC in 2 nm step."@en | | | | [temp:30C] | | "TSKgel-100V (4.6 x 250 mm, 5 |

=> JAVAのプログラムでRDF化 (turtle ファイルフォーマット)

メタボローム解析は、手法、解析機種、その設定も多種多様でそれらを収めたメタデータも複雑になりがち。
一般のユーザーの登録でどこまでやってもらうか、という熟考を重ねつつ登録用フォーマットを整備。

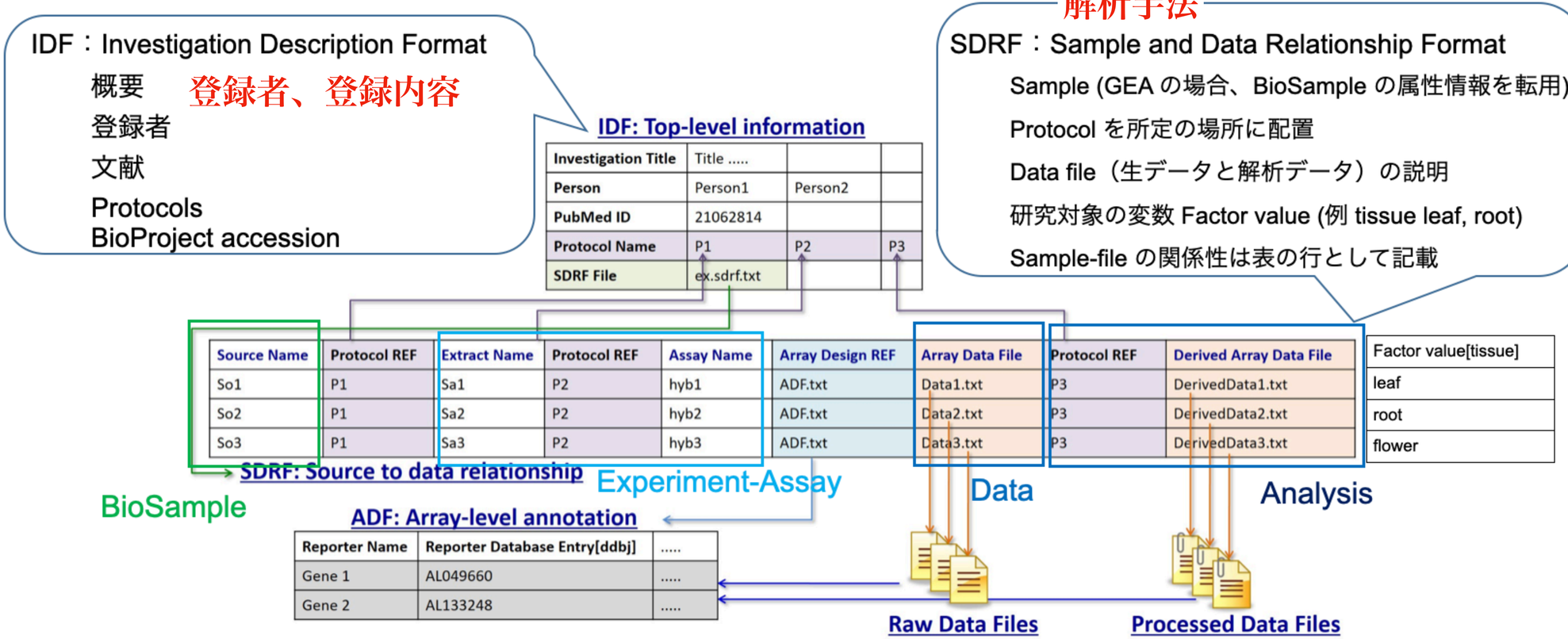
## MicroArray Gene Expression Tabular (MAGE-TAB)

機能ゲノミクスデータを構造化・標準化された方法で表現するための形式

IDF：Investigation Description Format
- 概要　登録者、登録内容
- 登録者
- 文献
- Protocols
- BioProject accession

解析手法

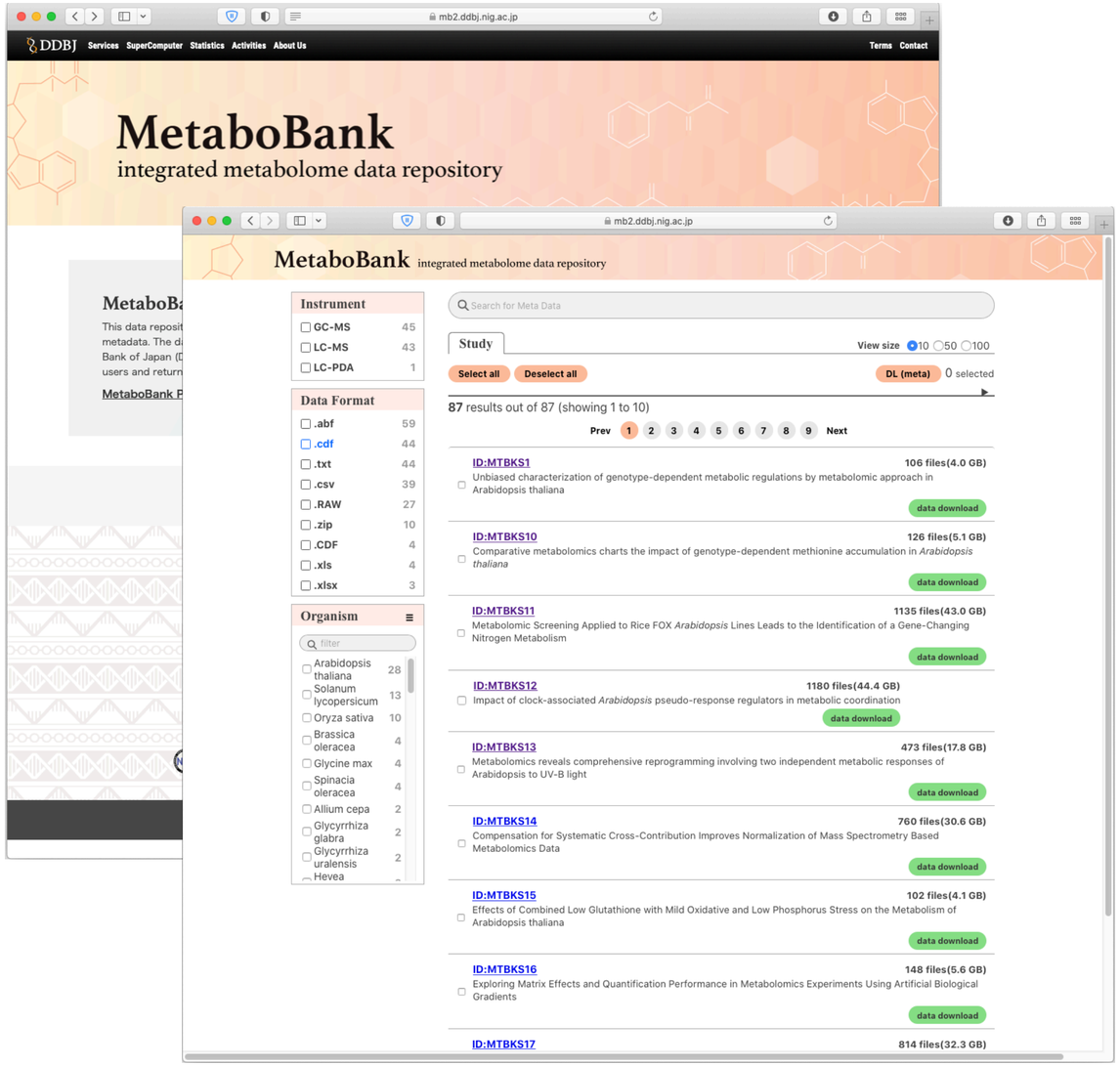
SDRF：Sample and Data Relationship Format
- Sample (GEA の場合、BioSample の属性情報を転用)
- Protocol を所定の場所に配置
- Data file（生データと解析データ）の説明
- 研究対象の変数 Factor value (例 tissue leaf, root)
- Sample-file の関係性は表の行として記載



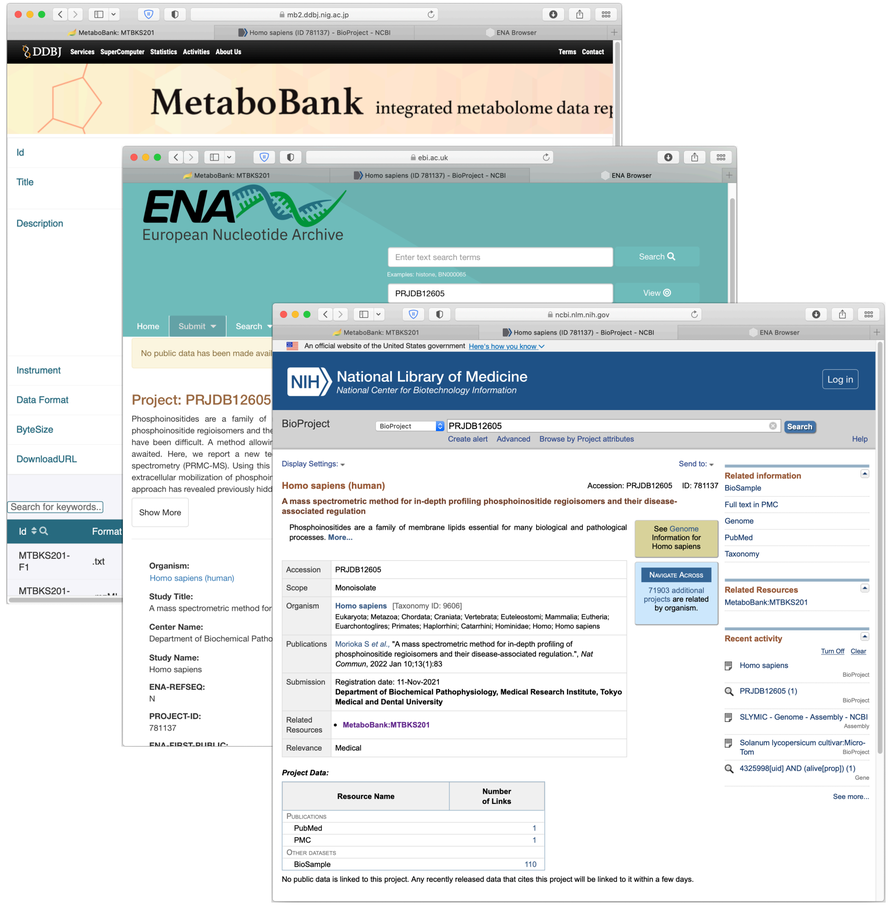

サンプル>解析手法>解析ファイルそれぞれをID化
サンプルから結果までの手順を追跡可能

EBIのProteomics IDEntications Database (PRIDE)やjPostといったプロテオミクス関連dbもMAGE-TABを参考にしたフォーマットの採用を検討中。

# RDFを見直したMetaboBank v2を公開中 (https://mb2.ddbj.nig.ac.jp)



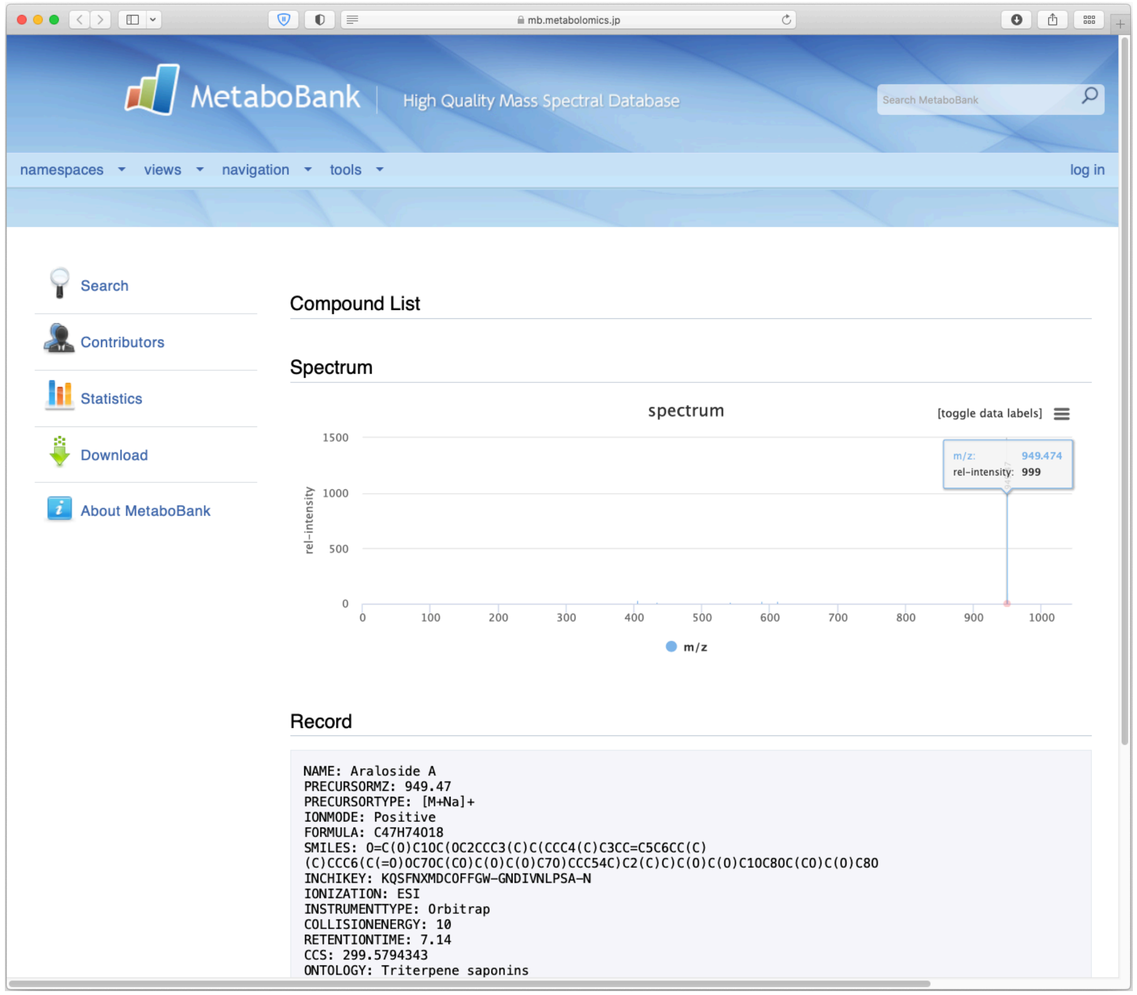

# 当プロジェクト以外の一般からの登録とデータ公開も始めている。




**https://mb2.ddbj.nig.ac.jp/study/MTBKS201.html**
**https://www.ebi.ac.uk/ena/browser/view/PRJDB12605**
**https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB12605**

現在 3箇所から7studyの登録申請。うち1study公開済

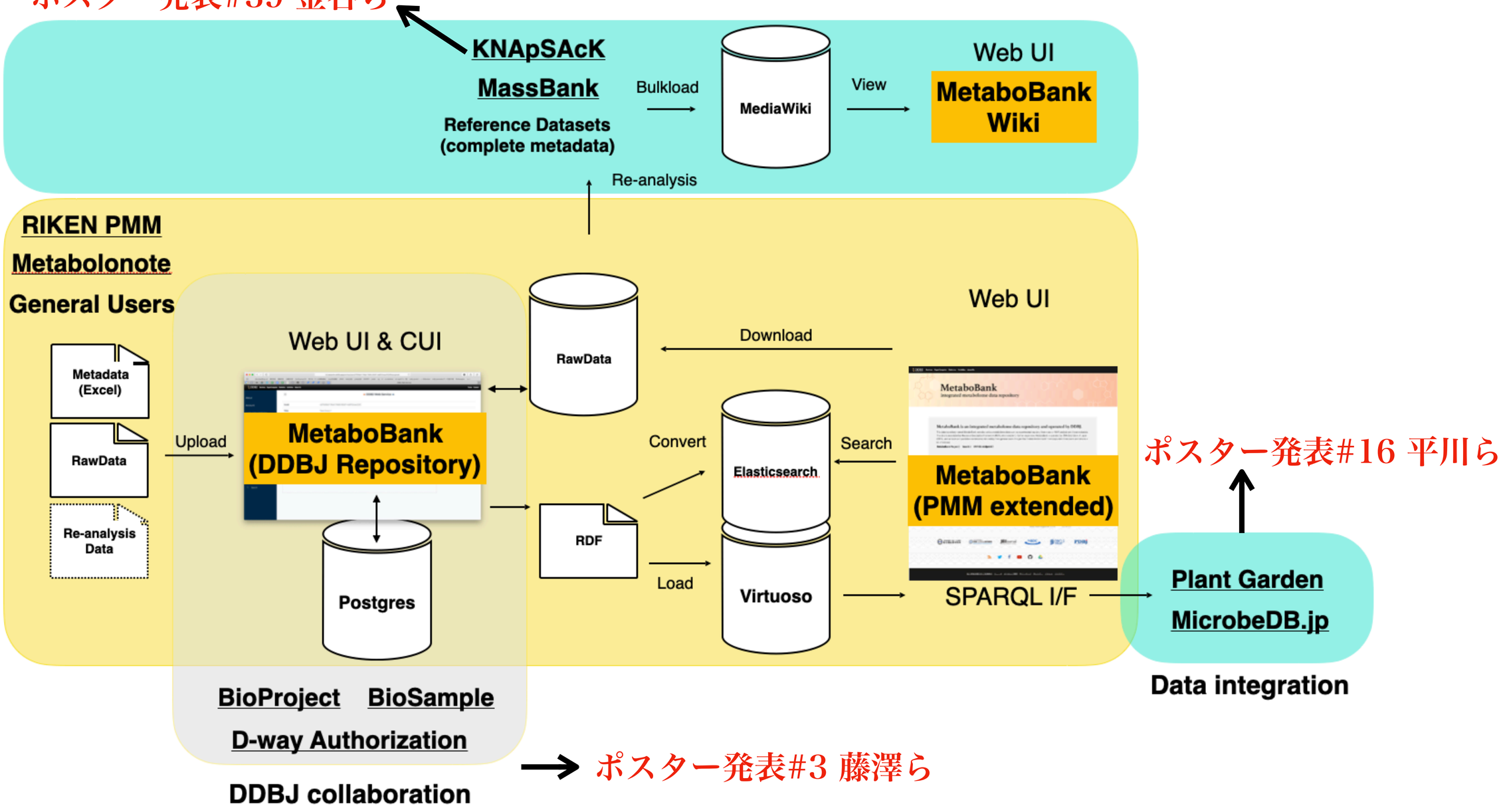
# 1次データレポジトリのMetaboBankに対して2次データベースMetaboBank Wikiを開発中




内包するデータ:
MassBank (http://www.massbank.jp)からのスペクトルデータ
天然化合物データベースKNApSAcK (http://www.knapsackfamily.com/)から化合物データ、代謝マップCobWeb
MetaboLights (https://www.ebi.ac.uk/metabolights/)のメタデータ
MetaboBankの登録データ

これらデータより物質循環を考慮したメタボロミクス情報基盤を構築する。

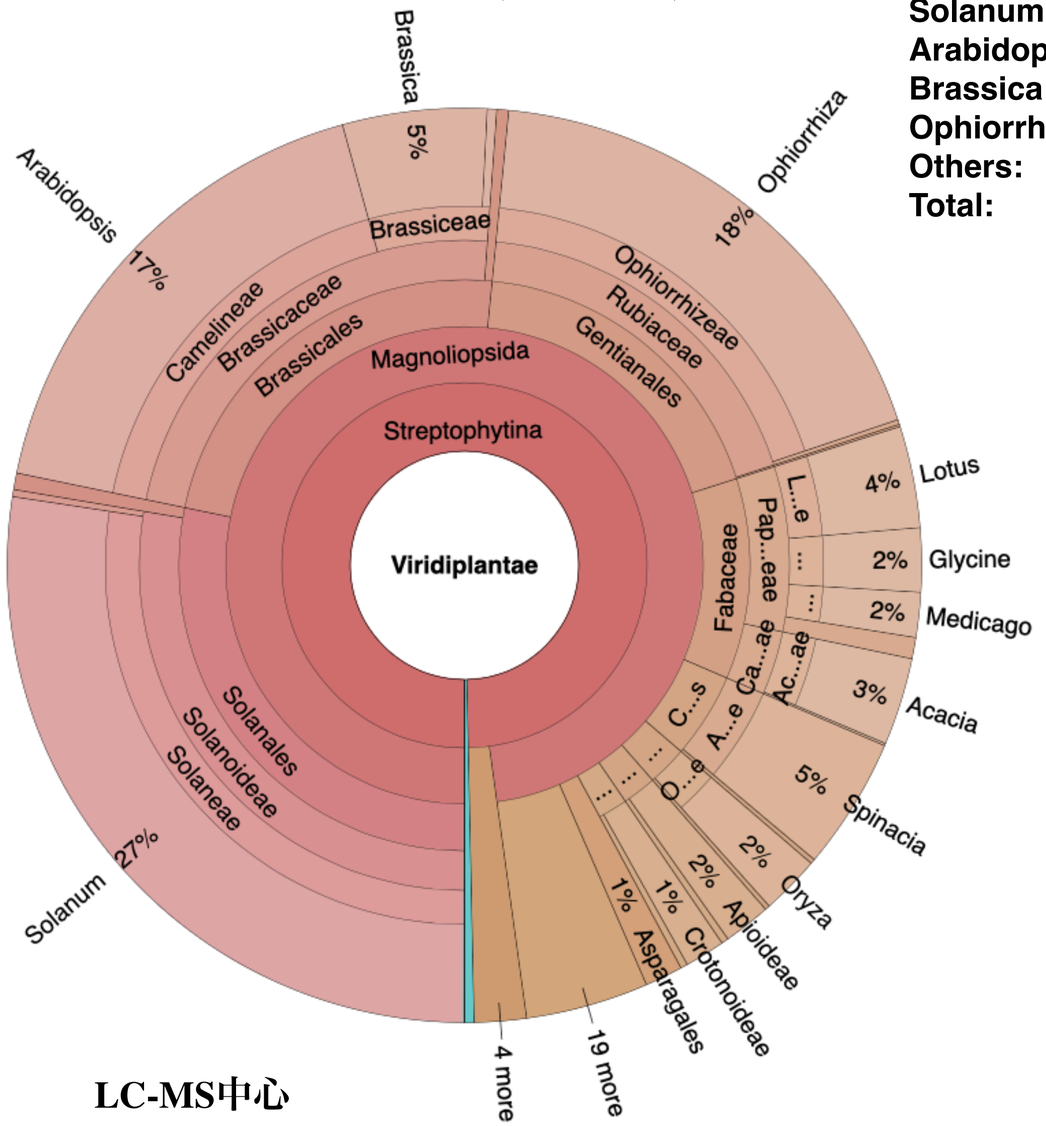
# MetaboBankの構造と本シンポジウムでの関連発表



# MetaboBank内KDRI、理研の植物サンプルの内訳

## Metabolonote (2021年8月)

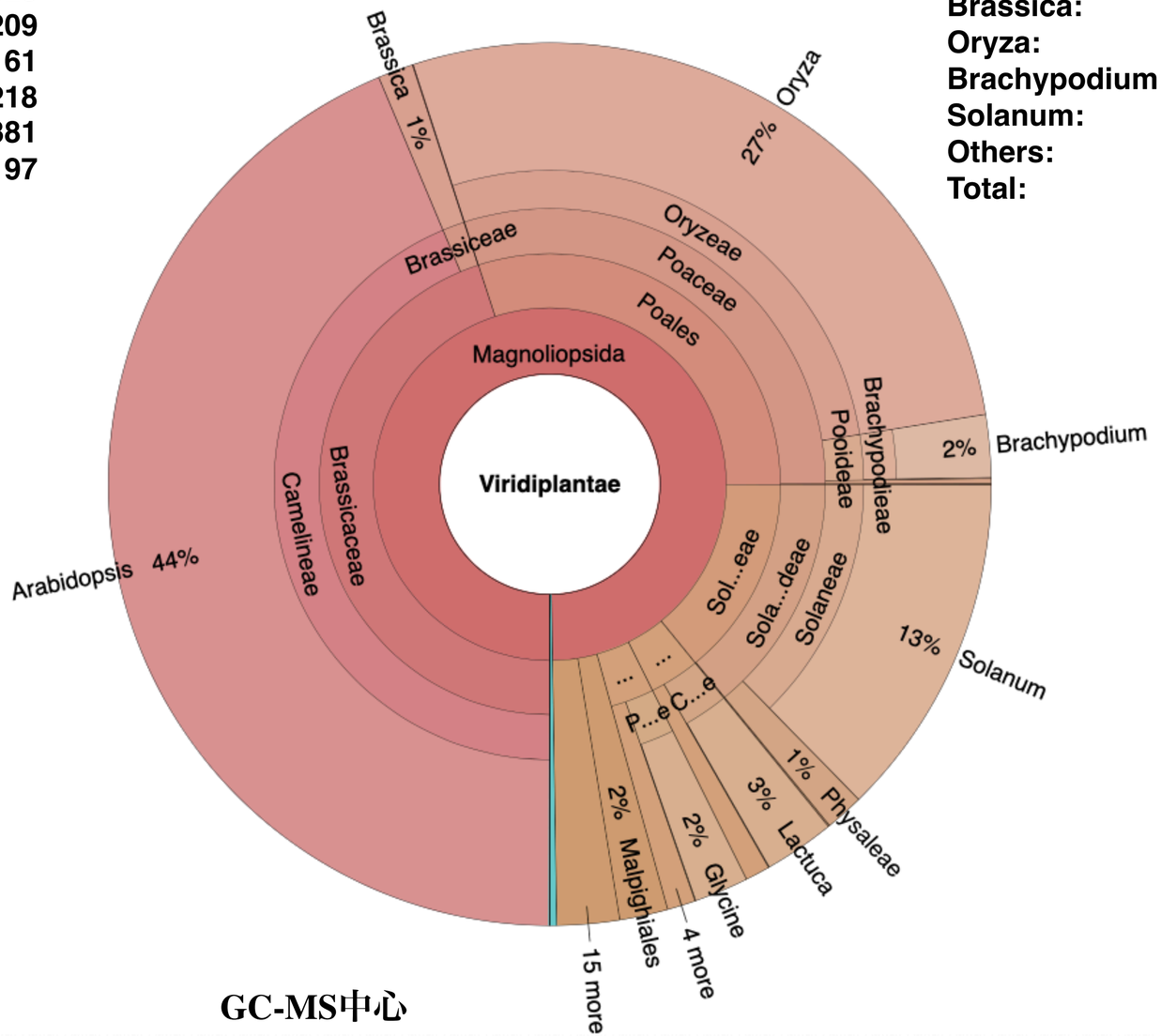
| | |
|---|---|
| Solanum: | 328 |
| Arabidopsis: | 209 |
| Brassica: | 61 |
| Ophiorrhiza: | 218 |
| Others: | 381 |
| Total: | 1,197 |




LC-MS中心

## RIKENの植物データ (2021年8月)

| | |
|---|---|
| Arabidopsis: | 3,337 |
| Brassica: | 96 |
| Oryza: | 1,935 |
| Brachypodium: | 174 |
| Solanum: | 964 |
| Others: | 1,136 |
| Total: | 7,642 |




GC-MS中心

理研データの再解析について
Fukushima et al. (2022) Plant Cell Physiol. 63:433–440.

# PowerGetBatchによるメタボローム データの再解析


PowerGetBatch (http://www.kazusa.or.jp/komics/software/PowerGetBatch)
液体クロマトグラフィー(LC)-高分解能質量分析(MS)のデータから、ピーク抽出、
サンプル間のアラインメントを行い、化合物データベース検索による一次アノテーション
を行う。
KDRIのスーパーコンピュータシステムで並列実行している。

# 質量分析計の分析方法ごとのパラメータの設定

Orbitrap　　シグナルピークの検出感度　高

フーリエ変換イオンサイクロトロン共鳴質量分析計
(FT-ICR)　　シグナルピークの検出感度　中

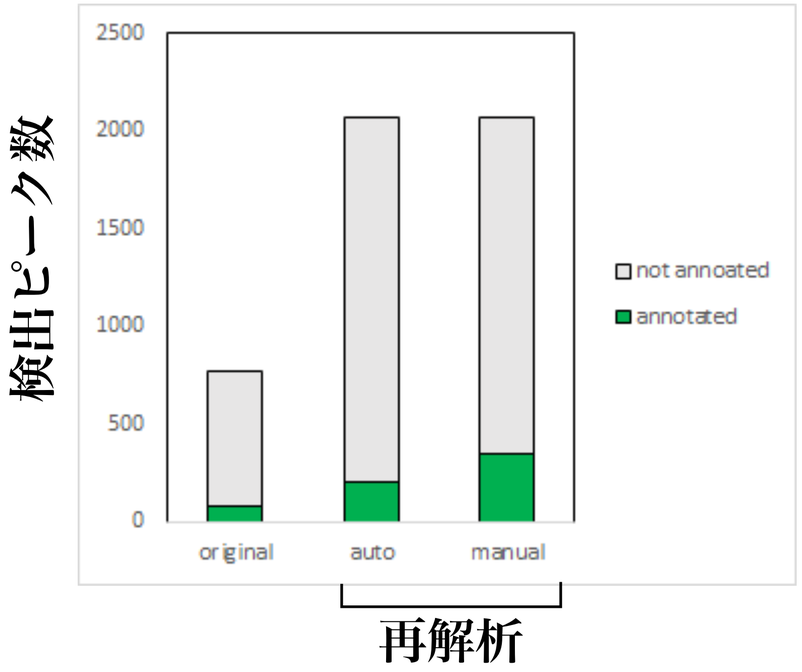
四重極飛行時間型質量分析計
(Q-TOF)　　シグナルピークの検出感度　低

装置のシグナル検出感度が高いものは、ノイズもそれだけ出るため、解析パラメータの感度を低くする、などピークの検出感度に関連するパラメータ4つを調整する。

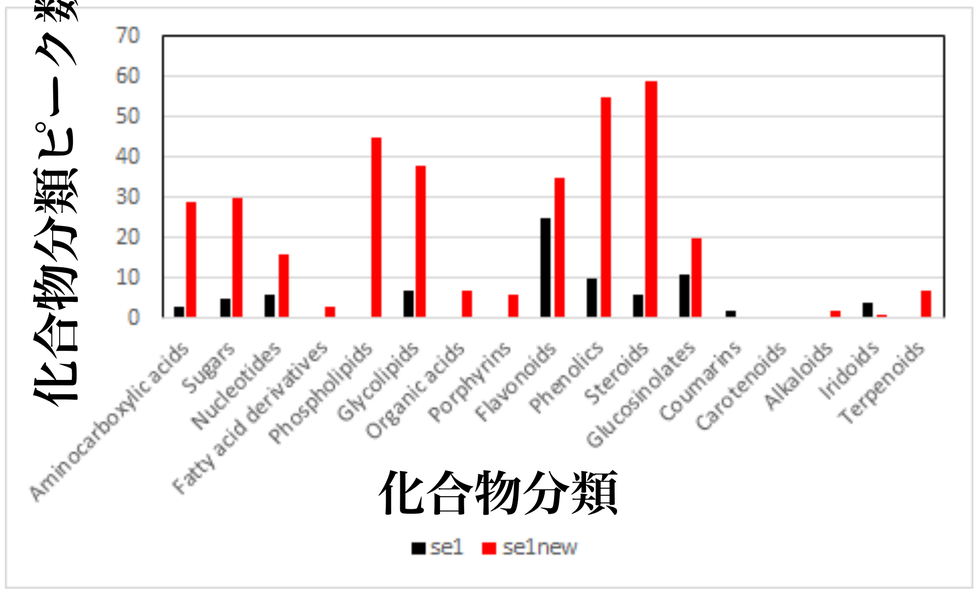
→マニュアルキュレーションした結果と比較を行う。

# メタボローム再解析でのアノテーションの改善
# (MTBKS102 シロイヌナズナの葉のFT-ICRの分析結果)

## 全体の結果




## 化合物分類毎の結果




化合物分類できたピーク数は自動解析で2倍、手動解析後は4倍に増加

アミノ酸、糖、脂質、フェノール性化合物、ステロイドなど多様な分類群でアノテーションできたピーク数が増加

# MTBKS52 (イネQTOFによる分析データ) の再解析

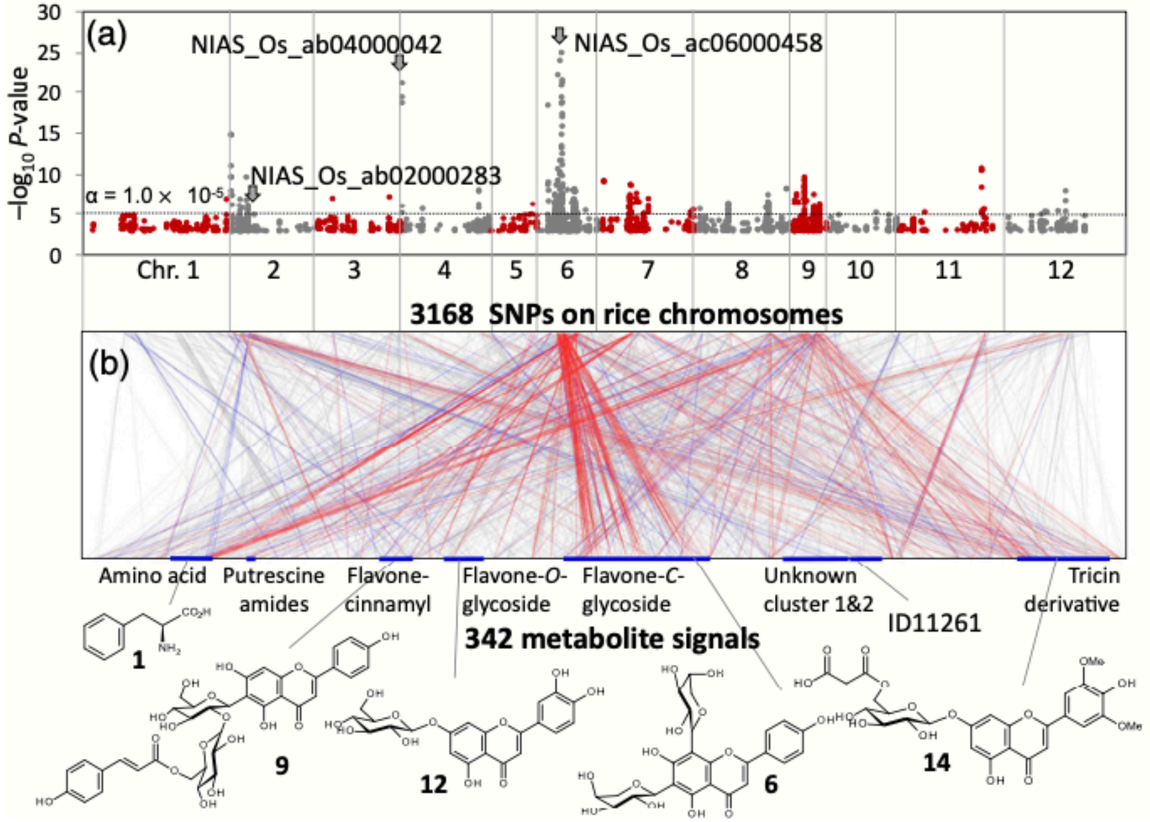



**Figure 3.** Genetic architecture of rice secondary metabolism.
(a) Manhattan plot for genome-wide association mapping of rice metabolic phenotypes. SNPs significantly associated with some metabolite levels were plotted on the rice genome ($\alpha = 1.0 \times 10^{-3}$).
(b) Associations between 3168 SNPs aligned on the upper boundary and 342 metabolites aligned on the lower boundary. Positions of SNPs correspond to the above panel. Red, blue, and gray lines represent significant associations between SNPs and metabolites with threshold levels of $\alpha = 1.0 \times 10^{-5}$, $5.0 \times 10^{-5}$, and $1.0 \times 10^{-3}$, respectively. Positions of metabolite clusters and representative metabolites are also represented (Table S4 for metabolite names).

Matsuda et al., The Plant Journal (2015) 81, 13–23

日本型イネ175品種で668run。=> 342 シグナルピーク検出=> 91個の代謝産物の構造を推定
GWASは、143のSNPと89の代謝物の間で323の関連を特定することに成功

188品種 752ファイルをPowerGetBatchによる再解析 => 7,000ピーク検出 => 300個の代謝産物の構造を推定。そのうちフラボノイド候補が62個確認。=> 今後はGWAS解析の予定。

# まとめ

メタボロームデータの1次レポジトリとしてMetaboBankはデータ公開と一般からの登録を受け付けている。

メタボローム2次データベースMetaboBank Wikiを開発中である。

新しいツール、データベースによる再解析はデータの充実化や利活用の観点から有効だと考えられる。
現在再解析データとKNApSAcKデータからフラボノイドのデータの収集を行っている。