

平成26年度ライフサイエンスデータベース統合推進事業

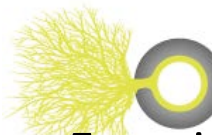
ゲノム・メタゲノム情報統合による微生物DBの超高度化推進

東京工業大学地球生命研究所

黒川 顕

微生物研究を取り巻く現状

- 微生物は地球上のいたる所に存在し環境と密接に関与している
- 微生物研究はバイオ分野のみならず、他の多くの分野と連携可能
- 既に多様なDBが多数存在する
- しかし、微生物と環境との関連性を記述しているDBは存在しない
- さらに、専門知識を持っていない**バイオ分野以外の人**は利用困難

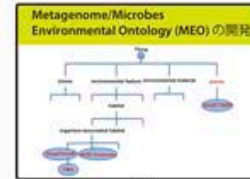
 **Microbe DB .JP** integrates lots of data related to microbes.
Especially, we integrate the microbial data that can be linked to **genomes**.

 **Microbe DB .JP**
<http://microbedb.jp/>

Microbe DB.jp
MicrobeDB.jp プロジェクトでは様々な微生物学上の知識を、ゲノム情報を核として遺伝子、系統、環境の3つの軸に沿ってセマンティックウェブの技術を利用して整理統合し、幅広い分野での微生物学の見解に資することの出来るデータベースの構築を目標としています。

Ontology

オントロジー: 検索タームの柔軟化&明確化



Gene

Taxon

Environment



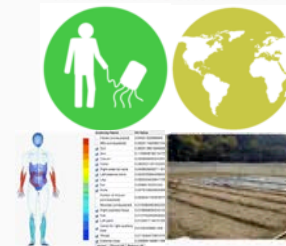
Ortholog: **MBGD**

オースログデータ



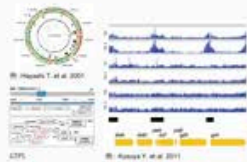
Taxonomy:
NCBI Taxonomy

系統分類データ



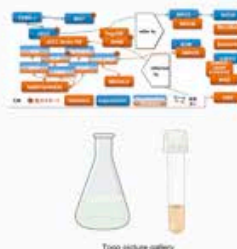
Metadata:
INSDC SRA

環境のメタデータ



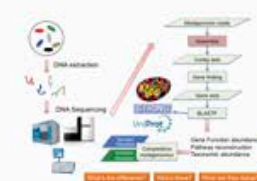
Genome: **GTPS/RefSeq**

オミックスデータ



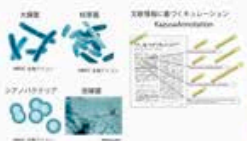
Culture Collection:
NBRC/JCM

菌株データ



Metagenome:
INSDC SRA

メタゲノムデータ



Annotation:

TogoAnnotation

モデル微生物の高品質
アノテーションデータ

Red color indicates our collaborators.

微生物における各種DBを統合化し、環境情報との連携を徹底的に記述した新規DB構築を実現

- 異分野データの統合化のため、セマンティックウェブの技術を徹底的に導入
- データ間をリンクするためのゲノム情報、オーソログ遺伝子情報、メタゲノム情報の整備
- 全データのRDF化、各データID間のリンク構築
- 各種オントロジーの開発、各データにマッピング
- アノテーション高度化システムの開発
- ユーザ認証システム基盤開発
- 検索結果可視化アプリケーション「Stanza」の開発
- ゲノム自動アノテーションシステムMiGAPとの連携

各種データのトリプル数

グラフ名	説明	作成元	トリプル数
refseq	RefSeq Prokaryoteゲノムデータ	DBCLS	550,273,744
mbgd	MBGD Orthologデータ	基生研	291,714,037
gtps	GTPSゲノムデータ	遺伝研	197,069,932
taxonomy	SPARQLthonで作成したNCBI Taxonomyオントロジー改良版	DBCLS, 遺伝研, 東工大	10,183,714
meta16S	各SRSメタ16Sの系統組成データ	東工大	9,831,600
gazetteer	地理オントロジー	外部機関	7,062,536
srs_metadata	SRSメタ16S・メタゲノムの様々なメタデータ	東工大	4,982,739
srs_ortholog	各SRSメタゲノムのMBGD Ortholog組成	東工大, 基生研	2,026,746
go	Geneオントロジー	外部機関	1,211,571
brc	JCM/NBRC菌株データ with NCBI Taxonomy ID	遺伝研, 東工大, DBCLS	903,319
gold	GOLDの個別ゲノムのMEO等へのオントロジーマッピングデータ	東工大, DBCLS	150,899
srs	SRSメタ16S・メタゲノムのMEO等へのオントロジーマッピングデータ	東工大	53,691
so	Sequenceオントロジー	外部機関	43,060
pdo	感染症オントロジー + 症状オントロジー + ゲノムへのオントロジーマッピングデータ	東工大	8,809
meo	微生物の生息環境オントロジー	東工大	4,975
msv	SRSメタ16S・メタゲノムのメタデータオントロジー	東工大	1,601
mpo	微生物フェノタイプオントロジー	DBCLS	734
mccv	菌株オントロジー	東工大, DBCLS	293
その他中間データ	いくつかのデータ集計系のSPARQLクエリは遅いため、MSSが集計結果のデータを作成		440,773
合計			1,075,964,773

開発したオントロジー


- FALDO (Feature Annotation Location Description Ontology)
 - ゲノム中の各featureの位置情報を記述するためのオントロジー (w/BioHackathon)
- INSDC Ontology
 - INSDCエントリのfeatureとqualifierのターム記述のためのオントロジー (w/DBCLS)
- MCCV (Microbial Culture Collection Vocabulary)
 - 菌株データを記述するためのオントロジー
- MEO (Metagenome/Microbe Environmental Ontology)
 - 細菌の生息環境を記述するためのオントロジー
- PDO/CSSO (Pathogenic Disease Ontology with Symptom)
 - 細菌が引き起こす感染症の情報および感染症の症状を連結したオントロジー
- GMO (Growth Media Ontology)
 - 細菌の培地情報を記述するためのオントロジー (w/DBCLS)

<http://microbedb.jp/>

MicrobeDB

microbedb.jp/MDB リーダー

[Sign In](#)



Search

Gene: psbA
Taxonomy: Streptococcus glycerinaceus
Mapping: Escherichia coli O157:H7 str. Sakai
Environment: hot spring
SRS: rumen
Strain: Bifidobacterium
Disease: Cholera
MiGap: GAF

MicrobeDB.jpの開発で実現したこと

1. 既存のゲノム中の各遺伝子の情報 (オーソログ、系統プロファイル、環境プロファイル)
2. 菌株保存機関に存在する菌株の情報 (生育培地、表現型情報、遺伝子機能組成)
3. 様々な環境中の細菌群集の情報 (系統組成、遺伝子機能組成)
4. 上記の情報をシームレスに統合



問合せ例:

高温環境に多く存在する遺伝子はどのような遺伝子か? その遺伝子は、どの系統が主に持っているのか?

本研究開発の目標・ねらい

MicrobeDB.jpを

- より広い微生物種を対象として拡張
- データ収集や更新自動化による持続可能なシステム
- 最先端解析プロトコルを実装した解析結果の可視化

研究者コミュニティだけでなく不特定多数の
イノベータを対象とした利用性の向上を徹底する

単なる統計量の羅列ではなく、大規模データから新規知識発見を容易に行う事が可能な、今までのDBを超えたDBシステムを構築する事を目標とする。

主たる共同研究者

東京工業大学

黒川 顕: 微生物DBにおける研究統括

山田拓司: ゲノム・メタゲノムDBの構築

森 宙史: 真菌類データの整備、DB自動更新システムの開発

山本 希: 解析Stanza & オントロジー開発

国立遺伝学研究所

中村保一: 藻類データの整備

菅原秀明: MiGAPとの連携強化

神沼英里: MeGAPとの連携強化

藤澤貴智: アクセスレベルの制限システムの開発

基礎生物学研究所

内山郁夫: 真核生物に対するオーソログ解析手法の開発

千葉啓和: ドラフトゲノムのオーソログ解析

西出浩世: オーソログを基軸とした各種データ統合の推進

具体的な研究開発項目（3年間）

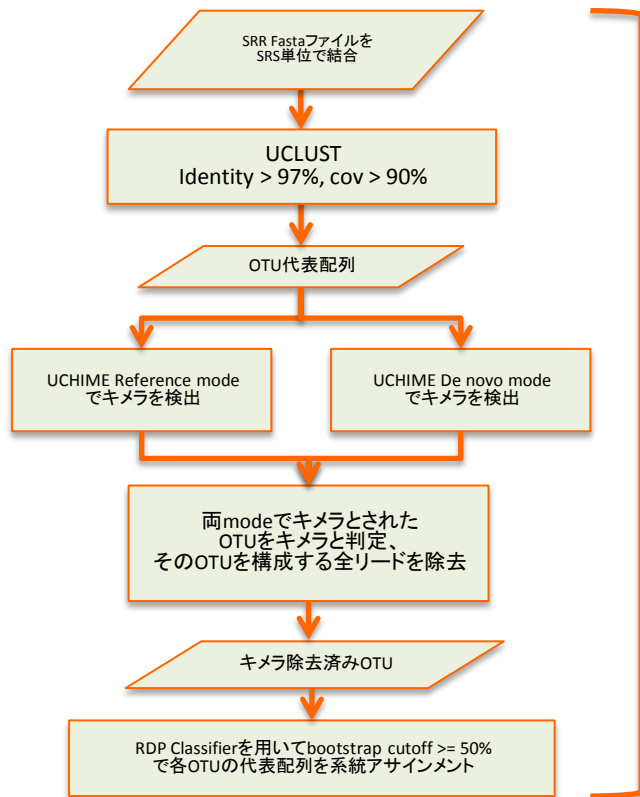
1. 各種オミックスデータへの対応
2. 真菌類および藻類を対象とした拡張
3. 各種オントロジー、ボキャブラリの開発
4. 解析プロトコルを実装した各種Stanzaの開発
5. データの収集およびクオリティコントロール、更新の自動化など持続可能なシステムの構築
6. データ共有、公開におけるアクセスレベルの制御システムの構築
7. 構築したシステムを幅広い分野の研究者に活用してもらうためのユーザビリティの向上

具体的な研究開発内容

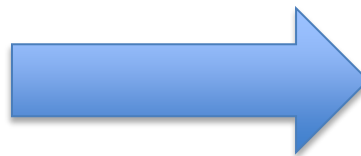
1. 各種オミックスデータへの対応
 - ドラフトゲノム、メタゲノム、RNA-seq等の各種オミックスデータを対象としたシステム整備とデータ収集およびそのセマンティックリソース化を実施
2. 真菌類(酵母・麹菌)および単細胞藻類を対象とした拡張
 - データの整理、各種オントロジーの開発、オーソログ遺伝子解析
 - 全データをRDF形式で記述しMicrobeDB.jpに統合
 - SGD、AspGDおよびFungiDBとの連携
3. 各種オントロジー・ボキャブラリの開発
 - 抗生物質や脂質などの生成物
 - RNA-seqなどにおける実験条件

4. 解析プロトコルを実装した 各種Stanzaの開発

多種多様な情報が混在しているゲノムやメタゲノム等の複雑なデータから知識発見をするために、比較ゲノム解析や比較メタゲノム解析など様々な解析Stanzaを開発する



解析Stanzaによる
結果の可視化

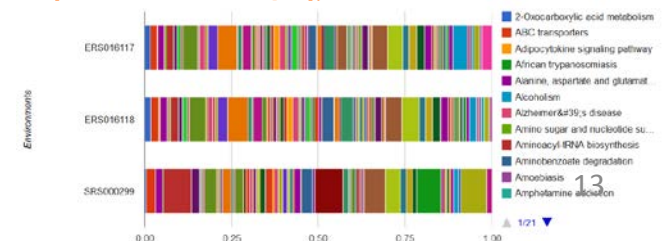


各系統と温度との相関係数リスト

メタデータ: temperature
表示種別: pathway
全メタデータ幅平均: 10.863
ダウンロード
件数: 224

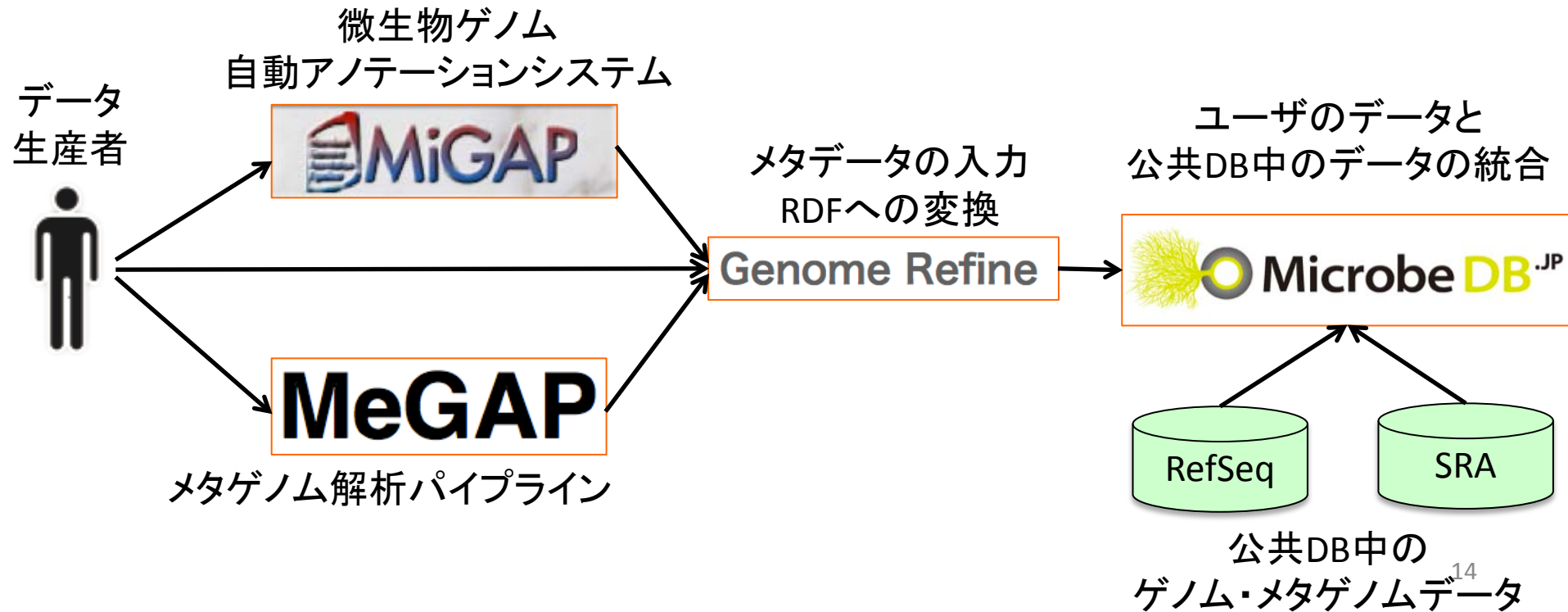
function ID	機能名	相関係数	サンプル数	メタデータ幅平均
http://www.genome.jp/dbget-bin/www_tget?i=000565	Ether lipid metabolism	0.9655330320905	3	8.5
http://www.genome.jp/dbget-bin/www_tget?i=000072	Synthesis and degradation of ketone bodies	0.61790144161164	25	11.3
http://www.genome.jp/dbget-bin/www_tget?i=000202	Two-component system	0.44564123374872	33	12.6
http://www.genome.jp/dbget-bin/www_tget?i=000910	Nitrogen metabolism	0.4390991977256	34	13.1
http://www.genome.jp/dbget-bin/www_tget?i=000643	Styrene degradation	0.3512843632747	17	10.2
http://www.genome.jp/dbget-bin/www_tget?i=000780	Biotin metabolism	0.34291921434483	26	11.3
http://www.genome.jp/dbget-bin/www_tget?i=003450	Non-homologous end-joining	0.27328119544247	7	12.5
http://www.genome.jp/dbget-bin/www_tget?i=000253	Tetracycline biosynthesis	0.26271210448501	26	11.3
http://www.genome.jp/dbget-bin/www_tget?i=000206	MicroRNAs in cancer	0.19776999332279	12	11.1
http://www.genome.jp/dbget-bin/www_tget?i=000560	Butanoate metabolism	0.18878105249684	26	11.3
http://www.genome.jp/dbget-bin/www_tget?i=004940	Type 1 diabetes mellitus	0.17906842287637	27	11.2
http://www.genome.jp/dbget-bin/www_tget?i=000340	Primary immunodeficiency	0.15863303443048	6	12.3
http://www.genome.jp/dbget-bin/www_tget?i=000152	Tuberculosis	0.15430685889307	29	11.4
http://www.genome.jp/dbget-bin/www_tget?i=000134	Legionellosis	0.1418962728288	28	11.2
http://www.genome.jp/dbget-bin/www_tget?i=000018	RNA degradation	0.08409246481617	29	11.4
http://www.genome.jp/dbget-bin/www_tget?i=000051	Fructose and mannose metabolism	0.07385682314052	27	11.1
http://www.genome.jp/dbget-bin/www_tget?i=000280	Valine, leucine and isoleucine degradation	0.05254845491534	26	11.3
http://www.genome.jp/dbget-bin/www_tget?i=000030	Pentose phosphate pathway	0.059475153688065	27	10.6
http://www.genome.jp/dbget-bin/www_tget?i=000190	Oxidative phosphorylation	0.03063403416373	28	11.4
http://www.genome.jp/dbget-bin/www_tget?i=000110	Carbon fixation in photosynthetic organisms	0.0186195171474291	28	10.7
http://www.genome.jp/dbget-bin/www_tget?i=000051	Fatty acid biosynthesis	0.0091971570069616	26	11.3
http://www.genome.jp/dbget-bin/www_tget?i=000945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	2	2	9.1
http://www.genome.jp/dbget-bin/www_tget?i=000504	Glycosphingolipid biosynthesis - ganglio series	2	2	10.4
http://www.genome.jp/dbget-bin/www_tget?i=000150	Staphylococcus aureus infection	1	1	10.7

環境ごとの系統組成のグラフ



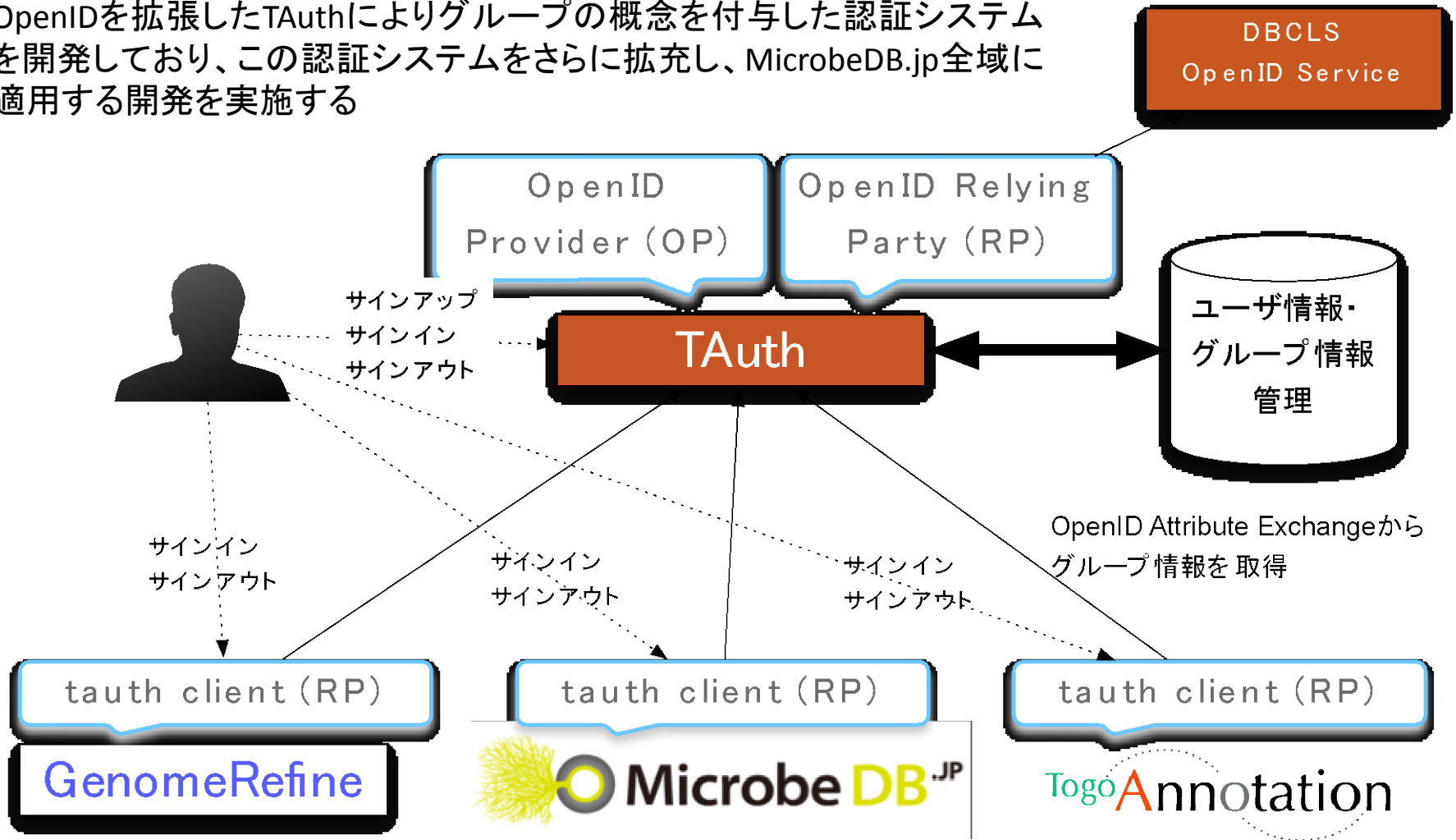
5. データの収集およびクオリティコントロール、更新の自動化など持続可能なシステムの構築

データ生産者から継続的にデータを受け付ける窓口のシステムとして微生物ゲノム自動アノテーションシステム「**MiGAP**」およびメタゲノム解析パイプライン「**MeGAP**」を利用し、MicrobeDB.jpと一体運用を実現する。また、これまで手作業で実施してきたDBの更新作業を可能な限り自動化し更新体制を強化する。



6. データ共有、公開における アクセスレベルの制御システムの構築

OpenIDを拡張したTAAuthによりグループの概念を付与した認証システムを開発しており、この認証システムをさらに拡充し、MicrobeDB.jp全域に適用する開発を実施する



グループ単位でのアクセスレベルコントロールを実現し、
配列解析者の利便性を増すと同時に、配列公開の促進を図る

7. 構築したシステムを幅広い分野の研究者に活用してもらうためのユーザビリティの向上

- Stanza間の関係性の記述 (Stanzaオントロジーの開発)
- キーワードとStanza間の関係性をDB化
 - 作成したRDFデータからキーワードを抽出し、どのような概念に関する語句なのかをトリプル中の述語やマッピングされているオントロジーの種類から判断し、適切なStanzaを関係付ける

検索語と各種Stanzaとの対応関係を明確にし、検索語によるStanzaの自動選択システムを開発する事で、検索システムを向上させる

研究開発プロジェクトの運営

- DBCLSと密に連携する
- チーム内MTGを定期的 to 実施
 - これまでに46回のチーム内MTGを実施
- BioHackathon、SPARQLthon、STANZAthonなどDBCLS主催の開発WSへの積極的な参加
- 各種学会などに出展しニーズの調査
- 企業などへのシーズの展開

本研究開発においては、これまで続けてきた上記活動をさらに強化し、メタボロームやフェノームDBとの連携も強める