平成29年度統合化推進プログラム研究開発課題

データサイエンスを加速させる微生物統合データベースの高度実用化開発

国立遺伝学研究所
黒川 顕

微生物・微生物叢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
- Metagenome、Microbiome等の研究が加速度的に進展している
- ヒト細菌叢研究は、文科省のH28年度研究戦略目標
- 環境、農業、海洋など関連研究がH28年度日本学術会議の大型研究計画マスタープランに含まれる

微生物統合DB「MicrobeDB.jp」

第1期および第2期統合化推進プログラムを通じて、微生物学の専門家のみならず非専門家も、微生物のゲノムやメタゲノム情報を容易に利活用できるDBを目指し開発。

- 微生物に関するデータを系統・遺伝子・環境の３つの軸に沿って整理・統合し、フルRDFのDBを構築
- 約90億トリプルから構成
- 12種類のオントロジー＆ボキャブラリの開発
- 公開済みの約17万サンプルのメタゲノムデータ、約1万7千株のゲノム・ドラフトゲノムデータを収録
- 195種類のStanzaの開発
- 解析プロトコルの標準化および解析パイプラインの開発
- 単細胞の真菌類(28種)・藻類(26種)のゲノムデータも整理・統合
- 自動更新技術の開発

微生物統合DB「MicrobeDB.jp」

Microbe DB.JP

hot spring x

Category colors: Environment Taxonomy Gene [hit column] (hit count) Phenotype Other category

Displaying related keywords . Please press for change to the new term instead of **"hot spring"**

Environment serpentine hot spring | calcite hot spring | alkaline hot spring | hot spring water | neutral hot spring | acid hot spring | artificial hot spring | acidic hot spring water

Taxonomy

Gene

Phenotype

This search term has exact match.
Now displaying stanzas in the category: **Environment** . Parameters are meo_id: **MEO_0000029**

Environment attributes

MEO ID	MEO_0000029
Title	hot spring
Synonyms	hot springs, hotspring, spring, thermal feature, thermal spring, thermal springs
Comment	A spring that is produced by the emergence of geothermally-heated groundwater from the Earth's crust.
MEO SuperClass ID	MEO_0000083
MEO SuperClass Title	spring

Environment Hierarchy

ID	Title
MEO_0000817	Environment for microbes

195種類のStanza群

Feature

[http://purl.obolibrary.org/obo/SO_0000704]

dbxref	http://www.ncbi.nlm.nih.gov/gene/897644
feature_gene	polC
feature_locus_tag	TM0576
location	605923..610026
isPartOf	http://genome.db/uuid/b4d48cd7-00ef-4e03-9adb-fda7de39e078
type	http://purl.obolibrary.org/obo/SO_0000704
label	TM0576

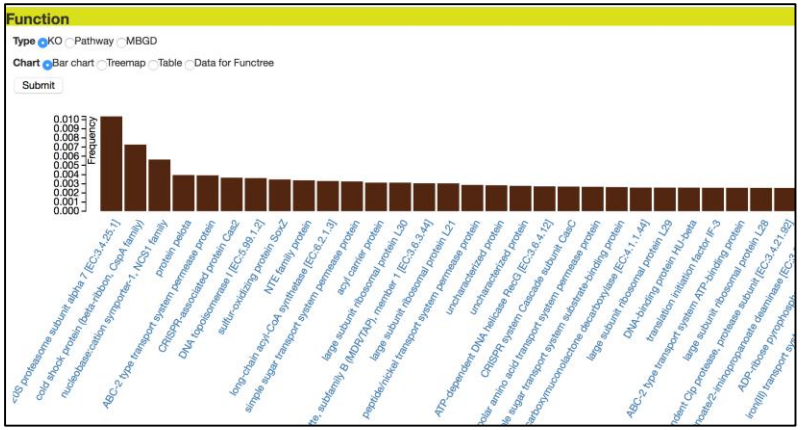
[http://purl.obolibrary.org/obo/SO_0000316]

dbxref	http://www.ncbi.nlm.nih.gov/gene/897644
dbxref	http://www.ncbi.nlm.nih.gov/nuccore/15643342
exons	nodeID://b71582

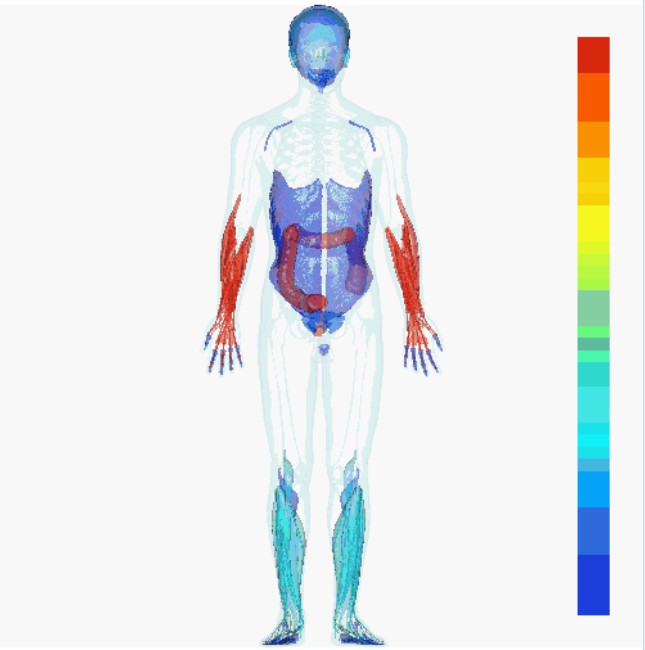
Genome

length	1860725
location	1..1860725
molecularType	genomic DNA
organism	Thermotoga maritima MSB8
sequence	http://genome.db/uuid/b4d48cd7-00ef-4e03-9adb-fda7de39e078.fasta
start	1
stop	1860725
strain	MSB8
version	NC_000853.1
modified	2012-02-13
type	http://purl.obolibrary.org/obo/SO_0000340
type	http://purl.obolibrary.org/obo/SO_0000988
comment	Thermotoga maritima MSB8 chromosome, complete genome.

Function



Body Mapping



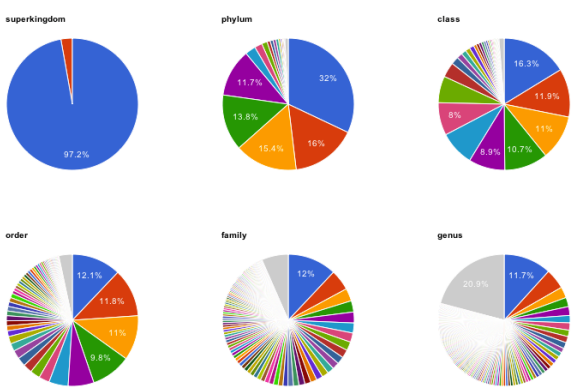
Anatomy Name	Hit Value
Feces (unvisualized)	0.84051325986689
Milk (unvisualized)	0.585317462682724
Gut	0.362513891694809
Arm	0.110064819613473
Cecum	0.093869656324387
Colon	0.090222145120303
Right external naris	0.045853955671191
Left external naris	0.043376094428822
Leg	0.032520323991776
Ear	0.02869155324353
Nose	0.027815662800968
Portion of mucus (unvisualized)	0.022045172407877
Mucosa (unvisualized)	0.015296295285225
Right popliteal fossa	0.013866629324515
Hair	0.013750255660852
Left palm	0.012587714675729
Canal for right auditory tube	0.01234339691499
Throat	0.011534477907475
External nose	0.009956159861758

reload | Check All | Release All

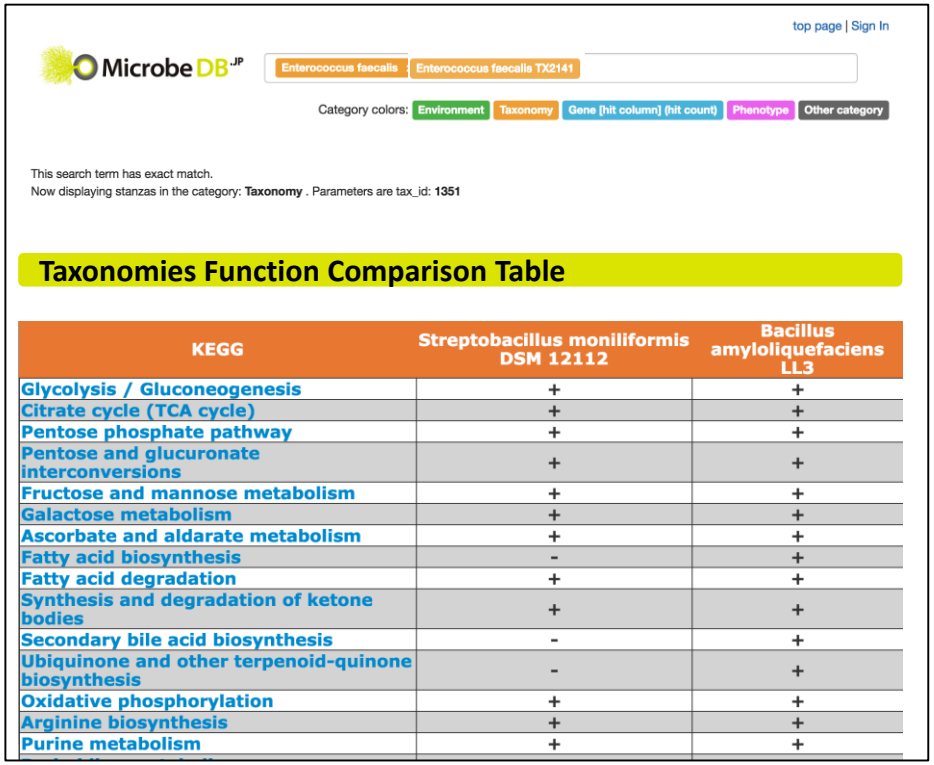
Ortholog

ID	Genome	Description	Protein	UniProt	GTPS	RefSeq
aac:AACI_1427	aac	DNA polymerase III subunit alpha	YP_003184842.1	C8WWI2	AACI_ACIDOCALDARIUSDSM446:ST2344	NC_013205.1
aar:ACEAR_1599	aar	DNA polymerase III catalytic subunit, PolC type	YP_003828170.1		AARA_DSM5501:ST105	
acl:ACL_0247	acl	DNA polymerase III subunit alpha	YP_001620249.1	A9NEU3	ALAI_PG8A:ST588	NC_010163.1
afl:AFLV_1700	afl	DNA polymerase III PolC	YP_002316046.1	B7GG80	AFLA_WK1:ST2505	NC_011567.1
afn:ACFER_1370	afn	DNA polymerase III subunit alpha	YP_003399045.1		AFER_DSM20731:ST1519	NC_013740.1
amt:AMET_2678	amt	DNA polymerase III subunit alpha	YP_001320489.1	A6TRL2	AMET_QYMF:ST2214	NC_009633.1

Taxonomy Composition

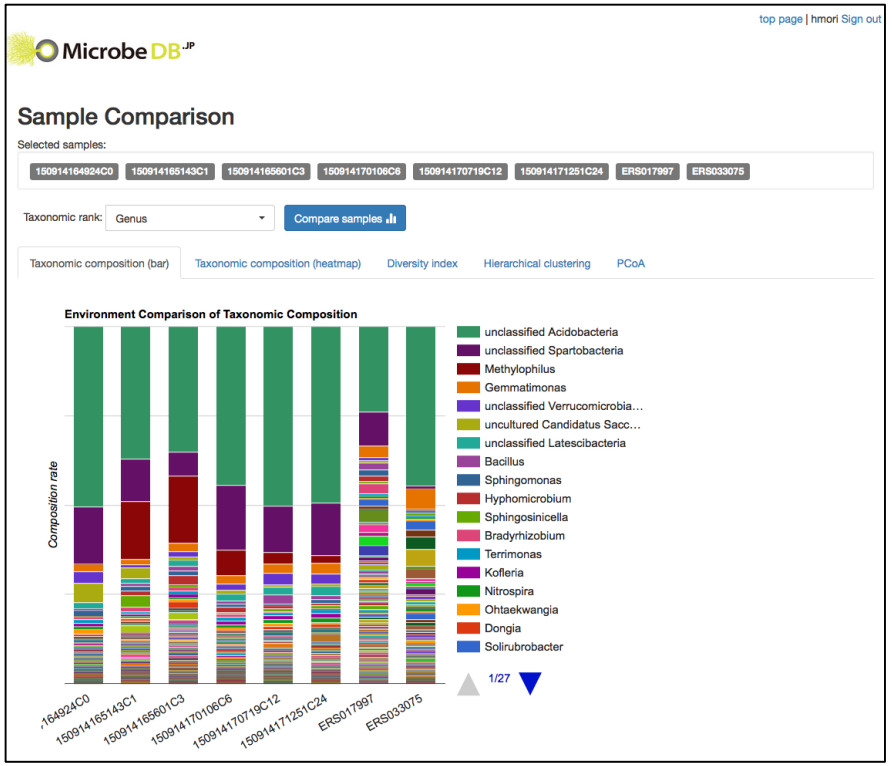


MiGAP&MeGAP連携による プライベートデータとの比較解析



KEGG	Streptobacillus moniliformis DSM 12112	Bacillus amyloliquefaciens LL3
Glycolysis / Gluconeogenesis	+	+
Citrate cycle (TCA cycle)	+	+
Pentose phosphate pathway	+	+
Pentose and glucuronate interconversions	+	+
Fructose and mannose metabolism	+	+
Galactose metabolism	+	+
Ascorbate and aldarate metabolism	+	+
Fatty acid biosynthesis	-	+
Fatty acid degradation	+	+
Synthesis and degradation of ketone bodies	+	+
Secondary bile acid biosynthesis	-	+
Ubiquinone and other terpenoid-quinone biosynthesis	-	+
Oxidative phosphorylation	+	+
Arginine biosynthesis	+	+
Purine metabolism	+	+

比較ゲノムStanza



比較メタゲノムStanza

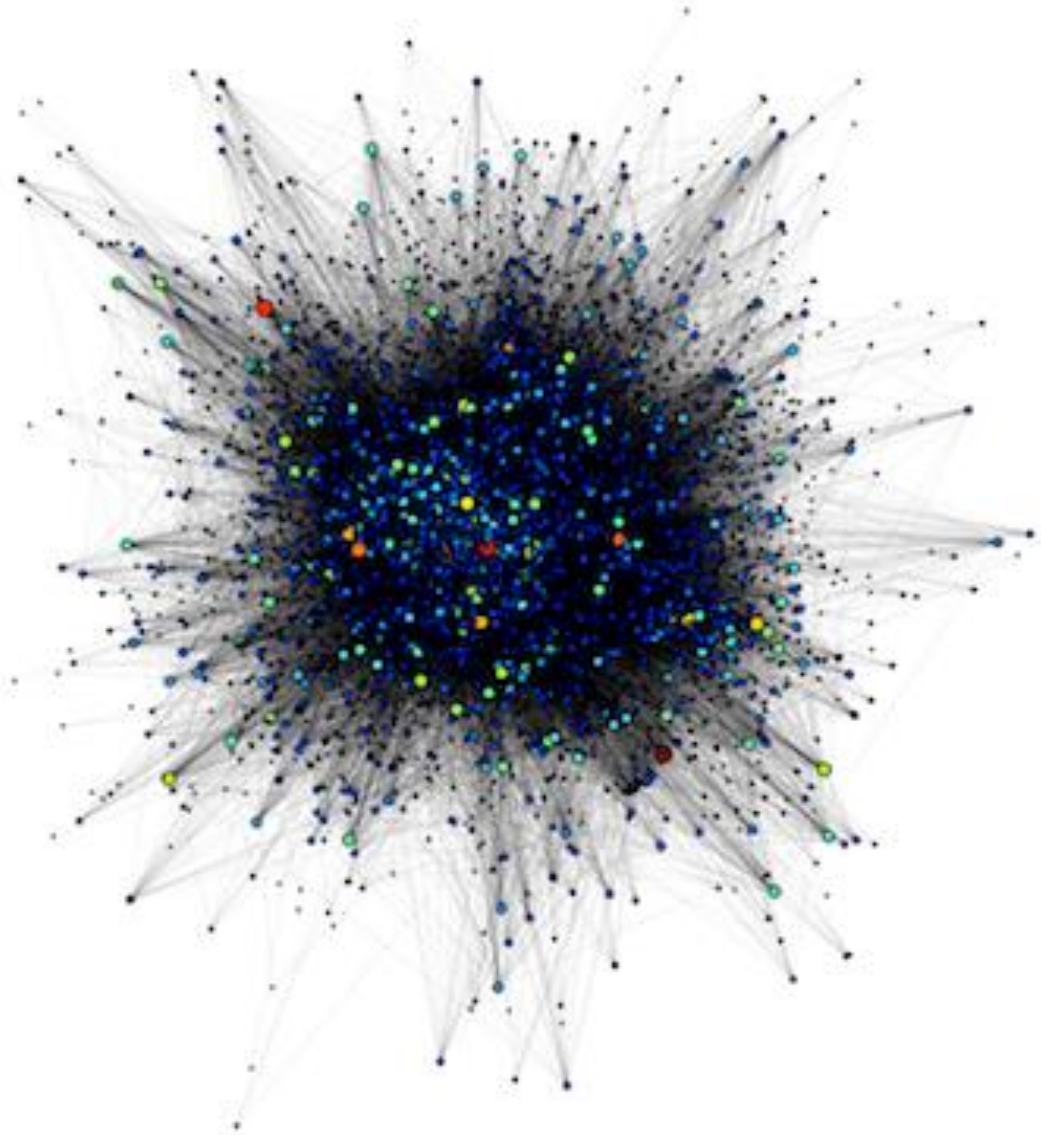
MicrobeDB.jpの課題

未だ十分な数の研究者がMicrobeDB.jpを利用しているとは言い難く、ユーザの獲得が喫緊の課題となっている。

MicrobeDB.jpでは非専門家もユーザとして想定しているため、第1期統合化推進プログラムから一貫して、google型UIを採用している。一方で専門家からは、「**何を入力して検索して良いかわからない**」との意見を多数頂いた。我々は当初、この原因が単純にgoogle型UIにあると考えていた。

しかし、ユーザの意見をさらに詳細に収集・分析したところ、①使い方がよくわからない、②意図しない検索結果が表示されており意味がわからない、③入力した検索語に対して、どのような結果が得られるのか予測できない、という意見に集約される事がわかった。

RDF-DBの全体像



- すべてのデータが連結された巨大なグラフとなっている。
- 統合DBを検索する際、利用者は巨大なグラフの全貌が不明であるため、「巨大グラフをどのように辿ればどのような答えが出てくるのか」を想定する事が困難となる。

全体像がわからない迷路で、出口に辿り着く方法を探すようなもの

統合DBに対しては、データ統合化のメリットを存分に活かすという点において、これまでのDBの利用形態を適用する事が本質的に困難

上図はイメージ：http://current.cs.ucsb.edu/rigel/

SPARQLおよびStanza

```
DEFINE sql:select-option "order"
PREFIX pdo: <http://purl.jp/bio/11/pdo/>

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX msv: <http://purl.jp/bio/11/msv/>
PREFIX mdbv: <http://purl.jp/bio/11/mdbv#>

SELECT
DISTINCT ?biosample ?title ?description ?environment ?meo_id ?me
o_label
FROM <http://microbedb.jp/pdo>
FROM <http://microbedb.jp/disease>
FROM <http://microbedb.jp/biosample>
FROM <http://microbedb.jp/meo>
WHERE
{
  {
    SELECT DISTINCT ?biosample
    {
      VALUES ?pdo_id { pdo:PDO_000193 }
      VALUES ?relate_tax_pdo_prop { pdo:hasInfectiousAgent
pdo:hasRelatedOrganism }
      ?pdo_id ?relate_tax_pdo_prop ?tax_blank .
      ?biosample skos:broader ?tax_blank ;
        a sio:SIO_010000 ;
        msv:sampleTitle/rdf:value ?title ;
        msv:sampleDescription/rdf:value ?description .
    } ORDER BY ?biosample LIMIT 100 OFFSET 0
  }
  ?biosample msv:sampleTitle/rdf:value ?title ;
    msv:sampleDescription/rdf:value ?description .
  OPTIONAL { ?biosample msv:environment/rdf:value ?environment }
  VALUES ?meo_mapping { mdbv:envMapping mdbv:stateMapping
mdbv:componentMapping mdbv:positionMapping }
  OPTIONAL
  {
    ?biosample ?meo_mapping ?meo_id .
    ?meo_id rdfs:label ?meo_label .
  }
}
```

肺炎に関与する菌種を介して、肺炎に関連するBioSampleの一覧を取得し、そのサンプルがどの環境由来であるかを取得する

- 統合DBを検索する際、発行するSPARQLは複雑であるため、SPARQLに不慣れなユーザに利用してもらうためには、SPARQLを生成し巨大な統合DBに対して検索を実行するとともに、得られた検索結果を可視化する機能を持つ「Stanza」を用意しておく必要がある。
- 入力された検索語に対して、検索語に関連するデータをどの階層まで検索し結果に含めるのか、などユーザに合わせたSPARQLの繊細なコントロールが必要となる。しかし、ユーザの検索意図は多様であり、システムはそれら意図を予め知り得ないため、このコントロールを誤ると意図しない結果を返すことになってしまう。

解決策

検索者の意図として、「ACDS 含量の高い hot spring 由来のメタゲノムサンプルを検索したい」（ACDS AND “hot sprint”)、である事がわかっている場合は簡単。
→ACDS が多く含まれるメタゲノムサンプルを取得して、そのサブグラフに対して hot spring を検索し一覧を返す SPARQLを発行すれば良い。

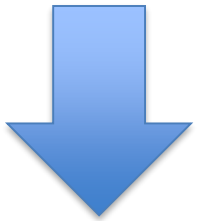
しかし、「ACDS含量の高い hot spring 由来のメタゲノムサンプルの平均的な群集構造を知りたい」場合は、全く異なるSPARQLを発行する必要がある。

SPARQLを記述するのはハードルが高く、そのような検索インターフェイスは非現実的。

当面は、検索者の意図（ユースケース）を反映させたStanzaを多数用意する事で対応可能→さらに我々は他の手法を検討中

微生物統合DBの高度実用化開発における目標・ねらい

「統合化されたデータをどのように渡り歩き、どのような新規知見を得るか」、という統合 DB の実用化に向けた、データサイエンスを加速する統合 DB の利活用方法の開発に重点を置き、MicrobeDB.jpの高度実用化を目指す。



統合 DB 活用による新たな科学的手法、
すなわちデータサイエンス研究手法を提案する

主たる共同研究者

国立遺伝学研究所

黒川　顕：微生物DBにおける研究統括
中村保一：DDBJ&植物統合DBの連携
神沼英里：解析パイプラインの高度化
森　宙史：統合DB構築、オントロジー&Stanza開発
藤澤貴智：セキュリティの高度化
東　光一：機械学習による統合データ解析技術開発

基礎生物学研究所

内山郁夫：オーソログデータの構築
千葉啓和：オーソログ情報のRDF化、Stanza開発 (現DBCLS)
西出浩世：オーソログデータの構築補助

東京工業大学

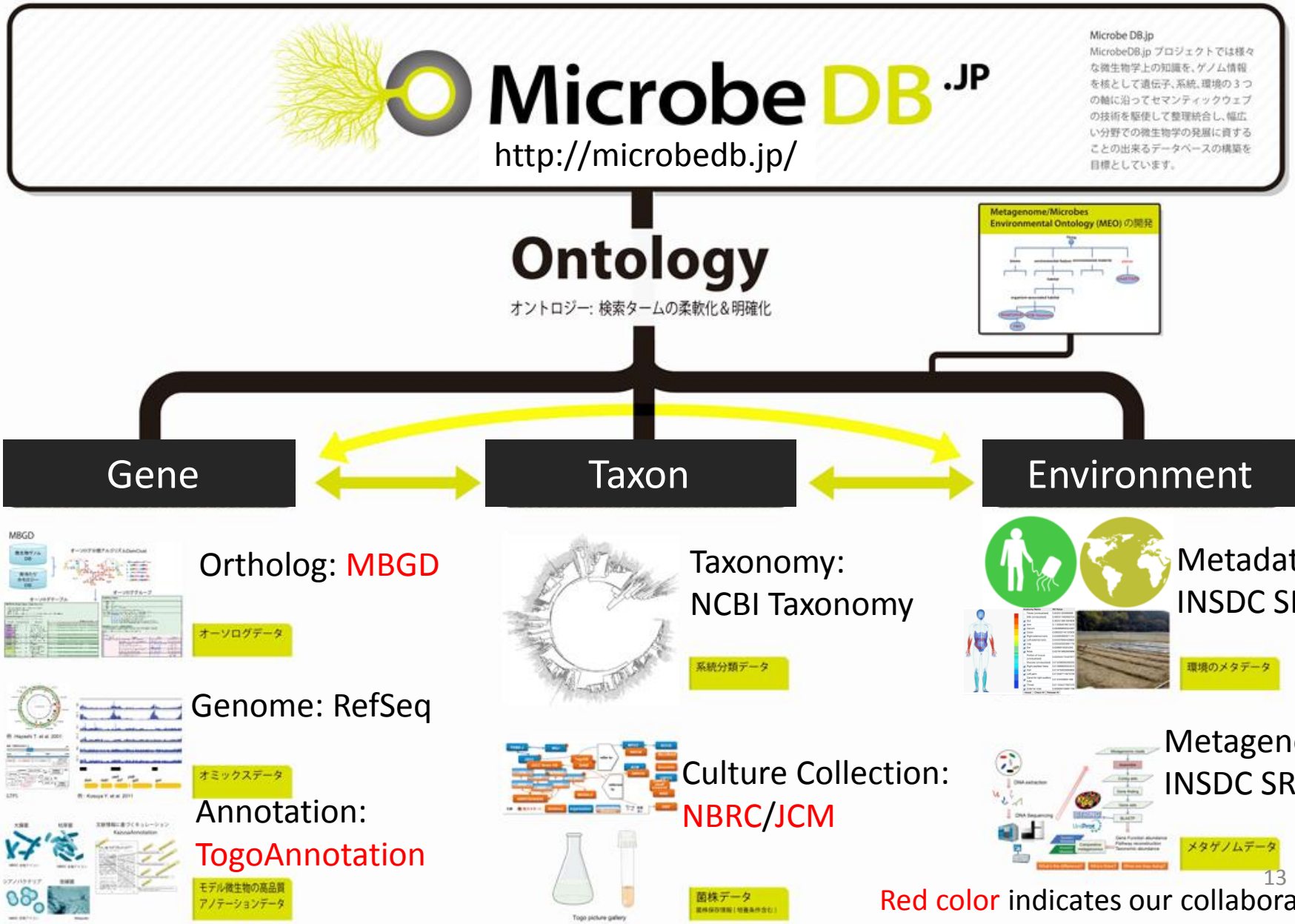
山田拓司：ヒトマイクロバイオームデータのメタデータ構築

千葉大学

高橋弘喜：真菌類ゲノム・菌株・オミックス情報の収集と高度化
矢口貴志：真菌類分類情報の整理

Microbe DB .JP integrates lots of data related to microbes.

Especially, we integrates the microbial data that can be linked to **genomes**.



Red color indicates our collaborators.

具体的な研究開発項目

1. 徹底したユーザビリティの向上
2. データ品質の向上
3. キラーアプリケーションの開発
4. さらなるデータの統合
5. 基盤データ解析技術の高度化
6. 効率的運用
7. ホロゲノム対応

具体的な研究開発内容

1. 徹底したユーザビリティの向上
   - ポータルサイト構築
   - プライベートデータアーカイブサービスの拡張
   - 多様なユースケースに対応できるStanzaの開発
2. データ品質の向上
   - 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
   - ヒトマイクロバイオーム関連のメタデータ整備
3. キラーアプリケーションの開発
   - VITCOMIC2（系統組成推定）
   - LEA（微生物GPS）

具体的な研究開発内容

4. さらなるデータの統合
   - 真核メタITS解析への対応
   - 真菌類ゲノム・菌株・オミックス情報の収集と統合
   - オミックスデータRDF拡張とアノテーション統合
5. 基盤データ解析技術の高度化
   - オーソログクラスタリングツールの開発（ビッグデータ対応）
   - 解析パイプラインの高速化、高精度化
6. 効率的運用
   - オントロジー自動アノテーションツールの精度向上
7. ホロゲノム対応
   - 植物ホロゲノム研究対応のStanza開発（植物統合DBとの連携）

達成目標（第1年次）

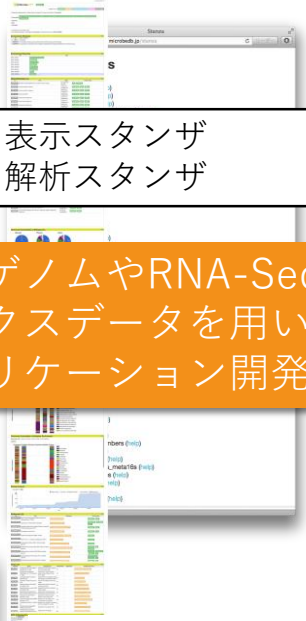
- MicrobeDB.jpトップページにポータルサイトを設置。
- MicrobeDB.jpユーザ会を設置。
- 公共DB中のヒトメタ16S・メタゲノムデータの詳細なメタデータを効率的に抽出する手法の開発。
- VITCOMIC2等のキラーアプリ候補にMicrobeDB.jpへの誘導のための機能を追加する。
- 真菌類のゲノムデータとRNA-Seqデータ、菌株メタデータの収集・整理を行う。
- MicrobeDB.jpで用いているメタ16S・メタゲノム解析パイプラインのアップデートを行う。
- MEOの自動アノテーションツールの精度向上のための、マニュアルアノテーションデータの追加。

各グループの
H29年度実施計画

遺伝研グループ: MicrobeDB.jp研究開発

検索

- 表示スタンザ
- 解析スタンザ



OWL

公開RDF

非公開RDF

⑥オントロジー自動アノテーションツールの精度向上

④オミックスデータRDF拡張・アノテーション統合

PGDB Plant Genome DataBase Japan

⑦藻類・植物ホロゲノム解析等を支援するサンプルメタデータおよび生物分類の再資源化

③メタゲノムやRNA-Seq等のマルチオミックスデータを用いたスタンザ・アプリケーション開発

①ポータルサイト構築

- 検索
- データ登録
- データ解析
- スタンザ
- ユースケース

TAuth認証サーバ

GenomeRefine

②ユーザアーカイブサービス拡張API開発

MeGAP2

⑤メタゲノム解析パイプラインの速度・精度向上、真核メタITS解析

MBGD MICROBIAL GENOME DATABASE

解析支援アプリケーション

DFAST

BioSample、JGA

DDBJ DNA Data Bank of Japan

ユーザ認証、登録UIなどユーザ利便性向上のための広域な連携

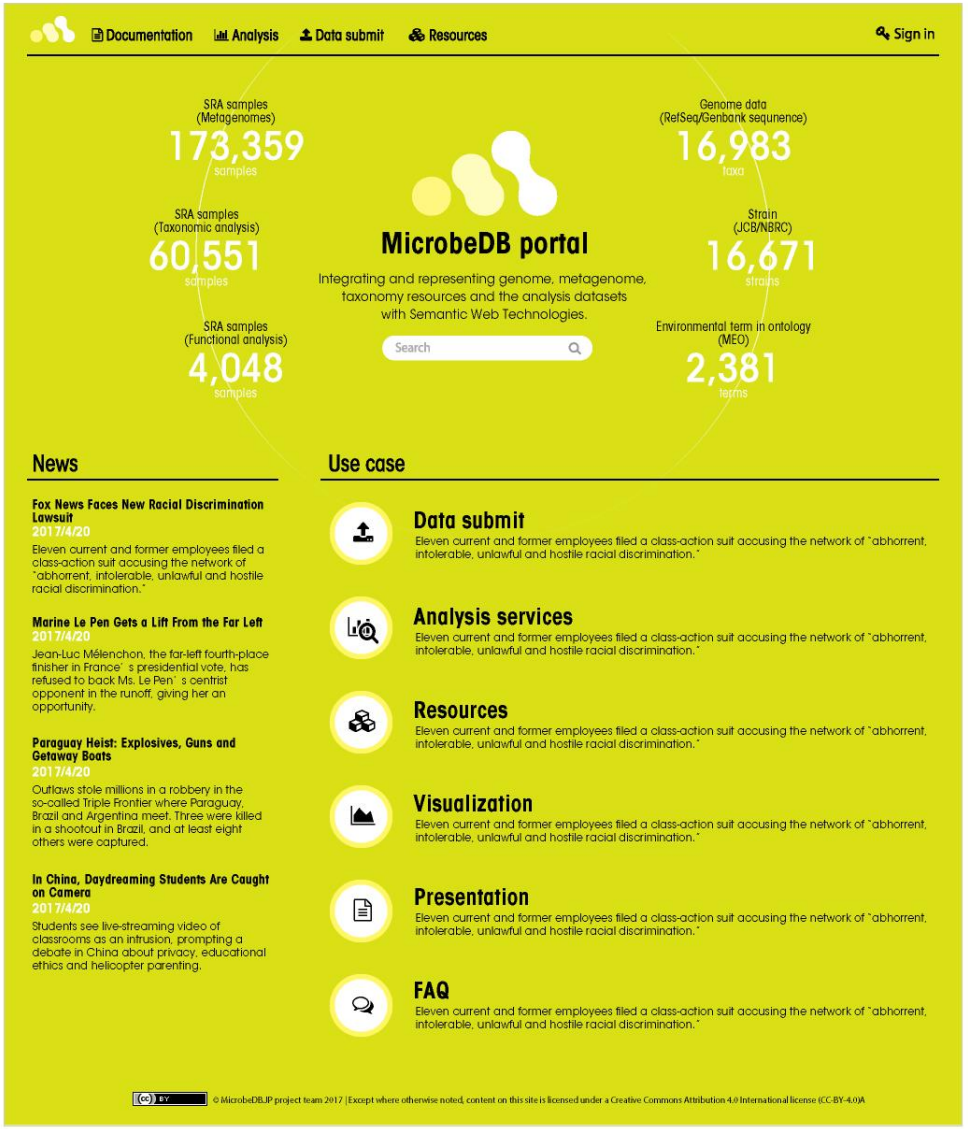
国際標準に則した生物分類情報および関連メタデータの再資源化

登録機能を統合するとともにユースケースの提供とユーザフィードバックに基づき継続的な発展

遺伝研グループ

① MicrobeDB.jpポータルサイト構築（2018年4月公開予定）
② DFAST, MeGAP解析パイプラインが提供するAPIと連携し、ポータルサイトからそれら機能を提供（2018年4月公開予定）
③ VITCOMIC2やLEA等のキラーアプリケーション候補とMicrobeDB.jpとの連携強化
④ オミックスデータRDF拡張とアノテーション統合については、配列IDならびに配列アライメント・マッピング情報を集積し、配列位置情報オントロジー FALDOに基づく変換系を構築
⑤ メタ16S・メタゲノム解析パイプラインの高速化・高精度化
⑥ MEOの自動アノテーションツールの精度向上
⑦ 藻類および植物ホロゲノム研究等を支援するサンプルメタデータおよび生物分類の再資源化、CyanoBase, RhizoBaseのユーザコミュニティと連携し、公共データベースに登録されている生物分類に関するメタデータについてRDF化
⑧ 2018年3月開催の日本ゲノム微生物学会において、第一回MicrobeDB.jp講習会・ユーザ会の開催を予定

MicrobeDB.jpポータルサイト構築

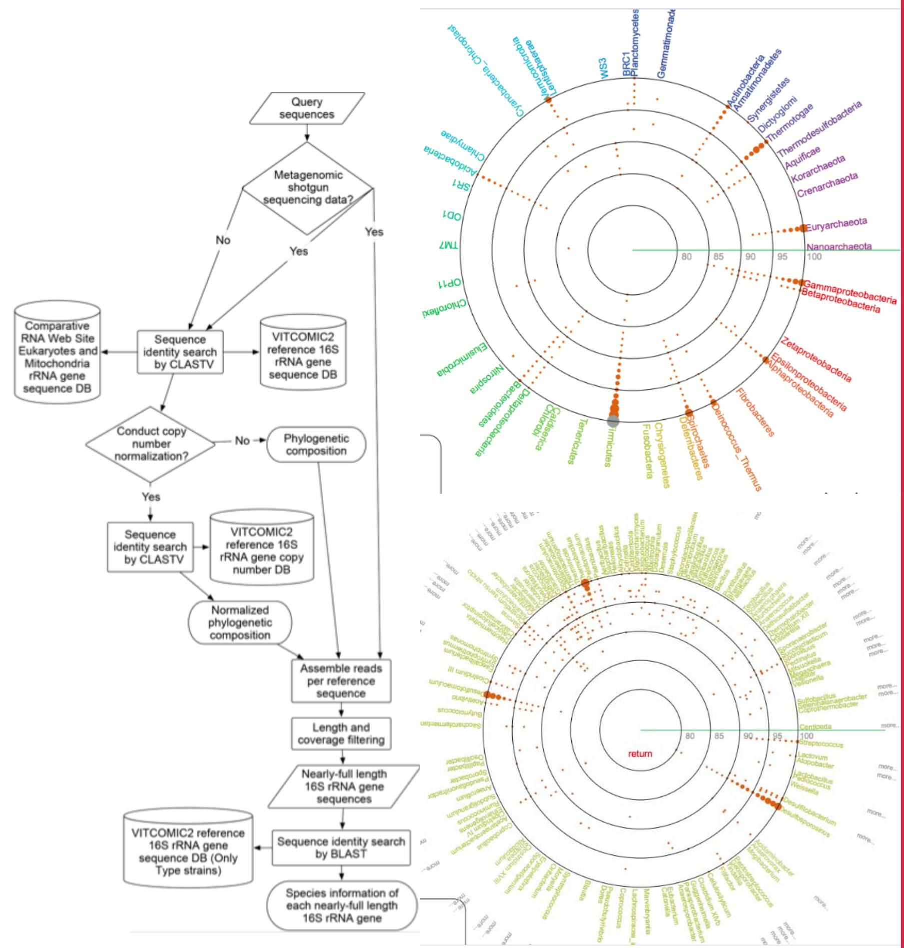


- 現行MicrobeDB.jpの機能・コンテンツの引き継ぎ
- ユーザレポジトリ機能（現行GenomeRefineの機能）の組み込み
- 公開以降も継続的にコンテンツ・ユースケースの追加

ユーザビリティの向上のための新しいポータルサイト構築（2018年4月公開予定）

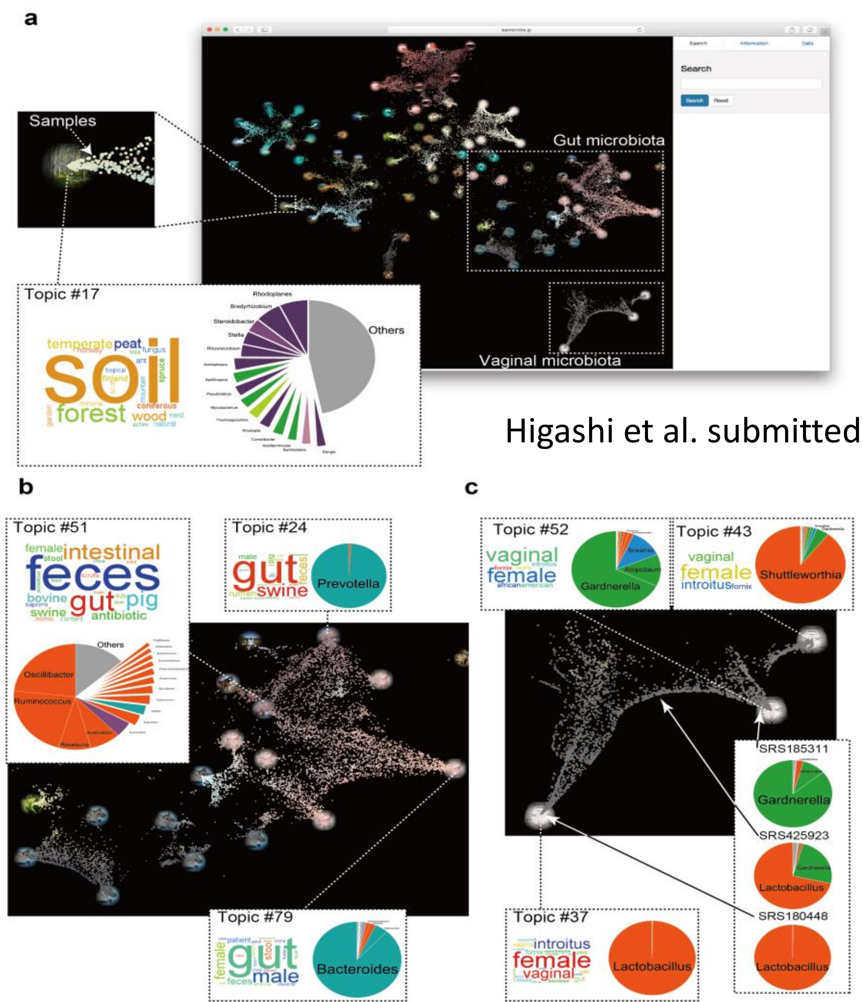
キラーアプリケーションとの連携

VITCOMIC2



Mori et al. submitted

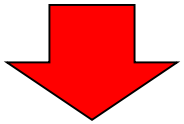
LEA



Higashi et al. submitted

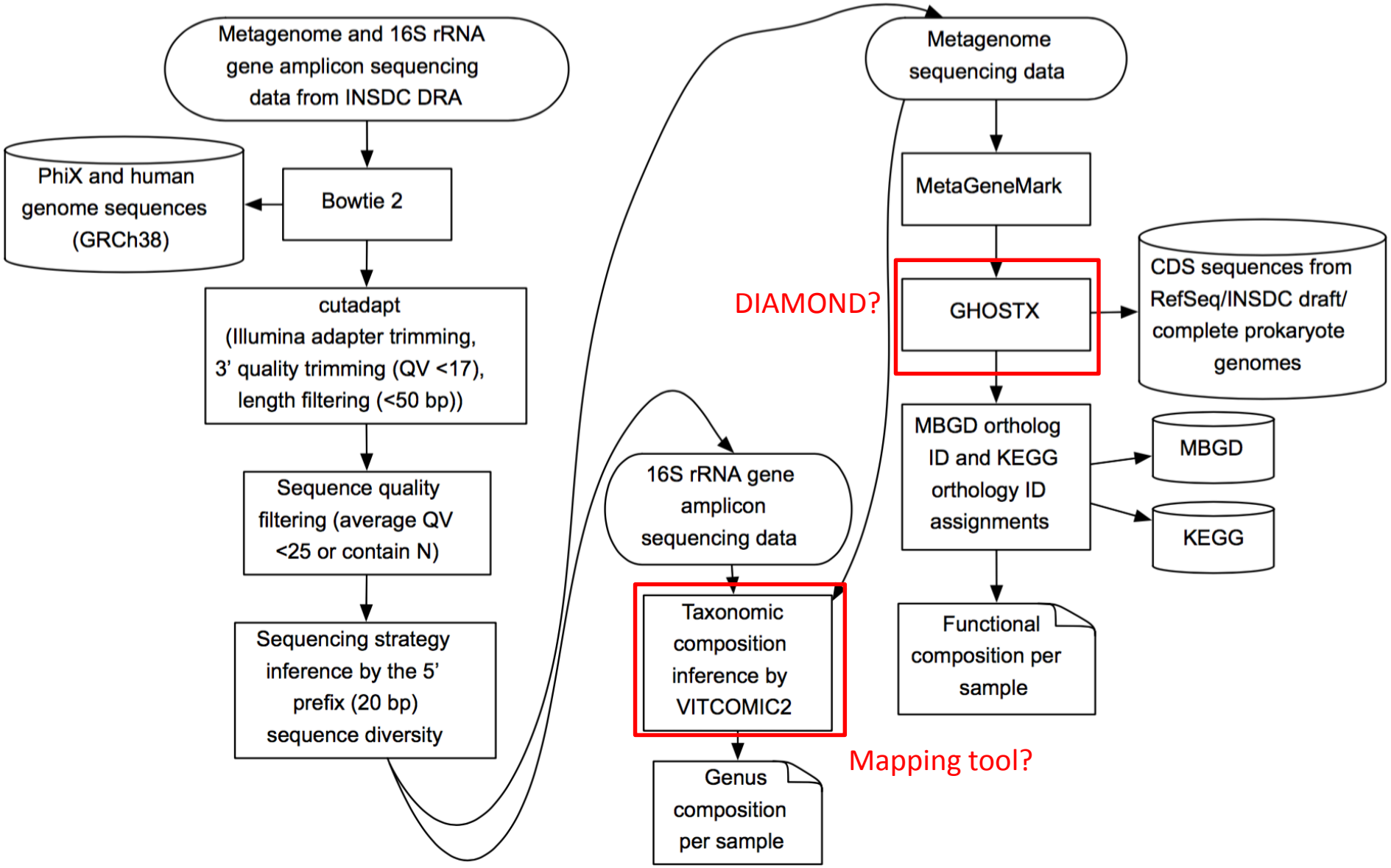
メタ16S・メタゲノムのデータ数の増加

- 2014年 Sample数
  - Metagenomes: 173,359
    - Meta 16S: 60,551
    - Shotgun metagenome: 4,048



- 2017年 Sample数
  - Metagenomes: 694,865
    - ecological metagenomes: 230,971
      - marine: 40,000
      - soil: 76,000
    - organismal metagenomes: 431,530
      - human metagenomes: 270,000
      - mouse metagenomes: 35,600

Metagenome解析パイプラインの改良
(MeGAP2 → MeGAP3)



MEOマニュアルアノテーションデータの追加

JCMの菌株に付随する分離源情報について、RIKEN側と協力してMEOのマニュアルアノテーションを行い毎月更新

→オートアノテーションにおける教師データとして利用

基生研グループ：超大規模データに対応したオーソログ解析プロトコルの確立および比較ゲノムアプリケーションの開発

① 階層的アプローチにより構築した新バージョンのオーソログデータのMBGD上での公開 (8月末ごろ)

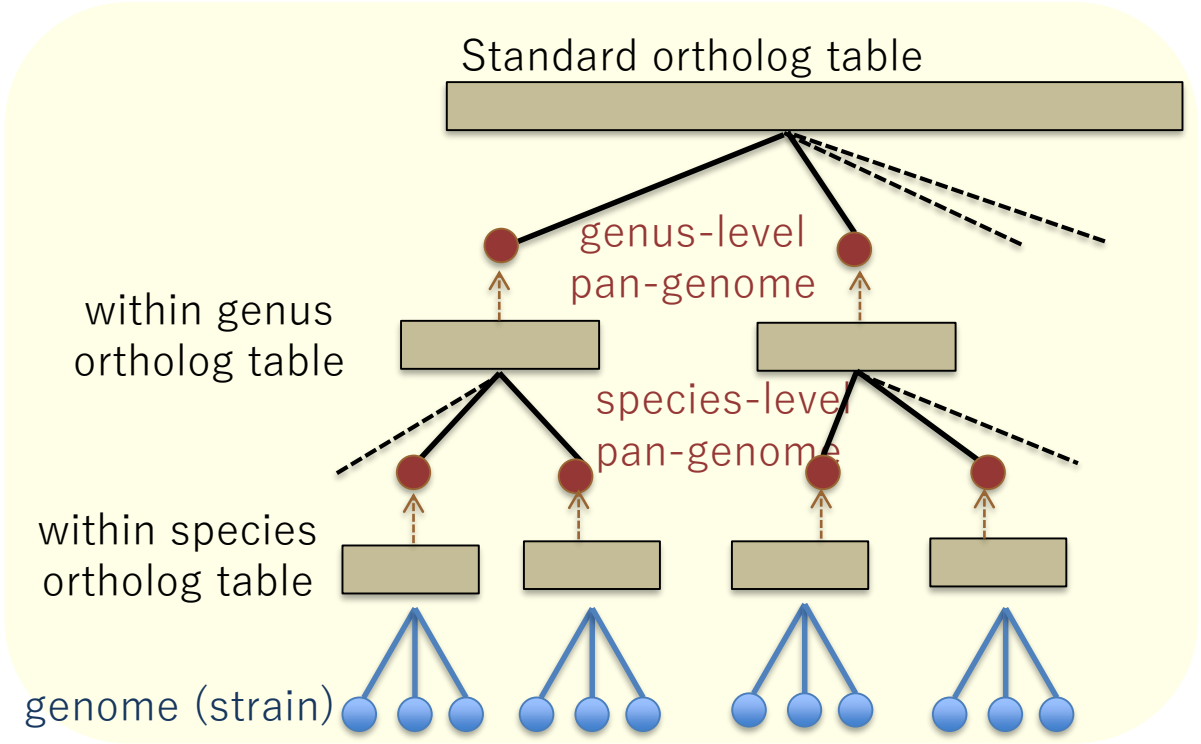
② 差分更新プロトコルの実装と、それを用いて最新データを取り込んだ次期バージョンの構築

③ MicrobeDB.jpに登録されたユーザゲノムデータをオーソログ解析する体制の構築

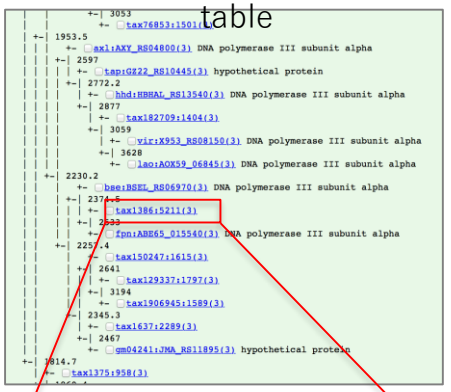
④ ユーザゲノムに対するオーソログ対応づけと、その結果を表示するアプリケーションの作成

基生研グループ課題１
超大規模データに対応したオーソログ解析プロトコルの確立

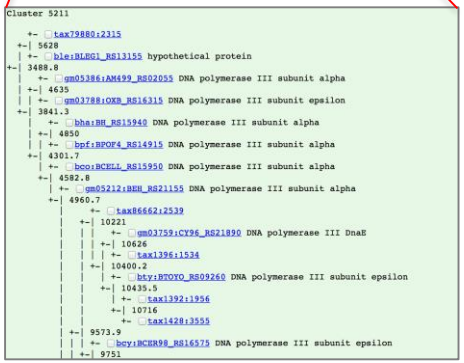
①階層的アプローチによる遺伝子レパートリー全体をカバーしたオーソログテーブルの構築



Standard ortholog table



Within genus ortholog table



②差分データ更新による効率的かつ安定的なデータ更新体制の構築

③解析対象とするゲノムを選択するための優先度指標の開発
（更新日付、ゲノムの完成度、コア遺伝子カバー率、リンク情報の数など）

千葉大グループ：真菌類ゲノム・菌株・オミックス情報の収集と高度化

真菌類のゲノム情報と生息環境や生理活性等のメタデータの収集および分類情報の整理を核として、MicrobeDB.jpの真菌に関するデータの基盤整備を実施する。

① 真菌類の完全長ゲノムデータ、ドラフトゲノムデータの収集および整備
② 真菌類のRNA-Seq等の各種オミックスデータの収集及び整備
③ 真菌類の菌種分類用配列データ（ITSなど）の収集およびそれを用いた真菌類分類情報の整備
　ITS配列（～760,000本）の収集とMicrobeDB.jpでの利用
④ 千葉大学真菌医学研究センターの菌株リソースのメタデータの整理・統合
　22,817の菌株情報の整理とMicrobeDB.jpへの格納

東工大グループ：ヒト常在菌細菌叢メタゲノムデータの価値最大化を目指したメタデータ構築

① ヒト共生細菌叢の関連論文リスト作成

- ヒト常在菌メタ16S・メタゲノムデータが報告されている約4,500報の論文をリスト化済み（ヒトメタゲノム関連のNCBI Taxonomy IDとPubMed Central論文を紐付けて抽出）
- オート・マニュアルキュレーションによる有用論文のスクリーニング
- 論文探索の基準作成

② 論文からのメタデータ抽出

- 上記論文に報告されているサンプルに紐付くメタデータを抽出しRDF化（INSDC BioSampleやSRA等の公共データベースにはこれらのメタデータは登録されていないことが多い）

例）

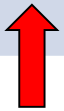
	身長	体重	性別	年齢	既往歴
サンプル_01	180	80	男	21	炎症性腸疾患
サンプル_02	161	55	女	33	なし
サンプル_03	162	60	男	44	胃がん

研究者コミュニティとの連携

- MicrobeDB.jp利用者講習会の開催
  - MicrobeDB.jp の利用方法を浸透させ、データサイエンス の裾野を拡大する。
  - 参加者は学術界、産業界を問わず募集。
  - 年1回 開催予定。
- MicrobeDB.jpユーザ会の設置
  - コアユーザからの意見や要望を収集するとともに、新規技術の開発や、解析 Stanza の開発などを技術共有しデータサイエンスの醸成を図る。
  - 各関連学会会員、企業コンソーシアム、共同研究者等で構成。
  - 年 2 回 開催予定。
- アンケート調査
  - 利用者講習会およびユーザ会に参加した研究者を名簿管理し、毎年度末にアンケート調査を実施し、MicrobeDB.jp への要望、利活用状況、論文発表状況に関する情報を収集する。
- 関連学会や展示会への積極的な出展

MicrobeDB.jpのアクセス数

	2013年度	2014年度	2015年度	2016年度	2017年度予想	2018年度予想	2019年度予想	2020年度予想	2021年度予想
訪問者数	5,712	6,673	7,489	10,048	12,000 (計画書では9,000)	13,000	18,000	26,000	28,000
訪問数	15,867	24,957	16,741	14,248	25,000 (計画書では20,000)	30,000	40,000	60,000	80,000
ページ数	313,272	561,522	729,093	250,891	80万	150万	200万	280万	300万



サーバ移転 & version 2へ更新

ページ遷移が多いversion 1からページ遷移が少ないversion 2へ更新したことと、サーバ移転時にサーバをしばらく停止していたことにより、訪問数とページ数が減少

達成目標（～第3年次）

- MicrobeDB.jpトップページにポータルサイトを設置。ユースケースを示すため、10種以上のStanzaからなるショーケースを設置。
- 真核メタITSデータに対する検索＆比較のための解析パイプライン開発。
- MEO等のオントロジーによるメタデータのアノテーションにかかる時間を、オントロジー自動アノテーションツールの精度向上により60%低減。
- オーソログデータの差分更新技術の開発、公表されたゲノムデータ全体を取り込んだオーソログDBの構築。
- ゲノムデータ爆発に対応したオーソログデータベース構築の汎用的な仕組みの開発。
- これら開発を通して、MicrobeDB.jpの年間訪問者数を13,000人以上にする。

達成目標（～第5年次）

- ホロゲノム研究に活用できるよう、植物統合DBと密に連携する。
- MEO以外のオントロジーアノテーションの自動化。
- オントロジー自動アノテーションの高効率化を受け、MicrobeDB.jpのローコストかつ定期的なバージョンアップを実施する体制を整備する。
- より安定的な運用を目指して、DDBJとの連携を強める。
- キラーアプリケーションの開発。
- 3年次までに開発した、選択的にデータを取り込むことによりオーソログデータを更新する仕組みを実装し、ゲノムデータ爆発時にも最善のオーソログデータを作成できるような更新体制を構築する。
- これら開発を通して、MicrobeDB.jpの年間訪問者数を26,000人以上にする。

将来展望：安定した運用体制の構築

- 微生物が関与する新たなデータの統合化を検討する。
- DDBJと連携を深め、MicrobeDB.jpの持続の安定性を図るとともに、より多くのユーザの利便性を向上させていきたい。
- ユーザ会を中心として、統合DBを駆使したデータサイエンスの推進に貢献したい。
- 営利団体ユーザを対象とした課金制度を導入したい。

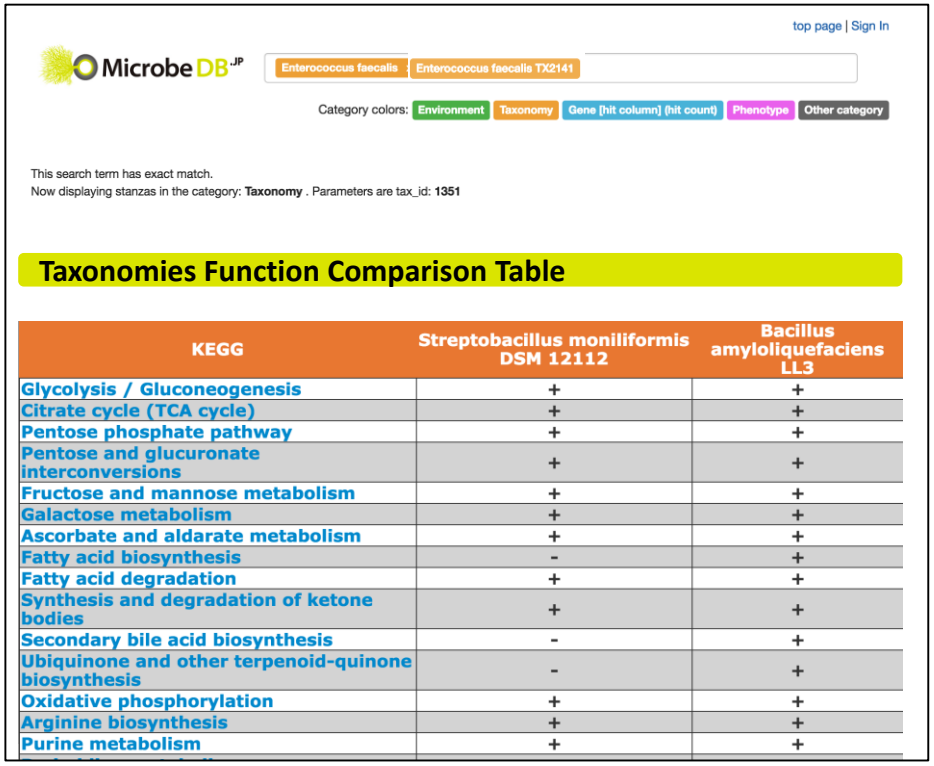
課金制度の導入の検討

DBおよび格納しているデータやStanzaに関しては、CC-BYでの利用を前提としている。MicrobeDB.jpでは、学術界、産業界問わず以下のサービスを提供している。

1. プライベートなゲノム・メタゲノムデータを受け入れ、統一した解析プロトコルで解析し、解析結果をMicrobeDB.jpのプライベートDBに格納する。
2. プライベートDBに格納された解析結果は、直ちにMicrobeDB.jpのデータへのリンクが得られるとともに、パブリックデータとの比較解析が可能となる。
3. プライベートDBへのアクセスはグループIDで管理されており、グループ内での共有が可能である。

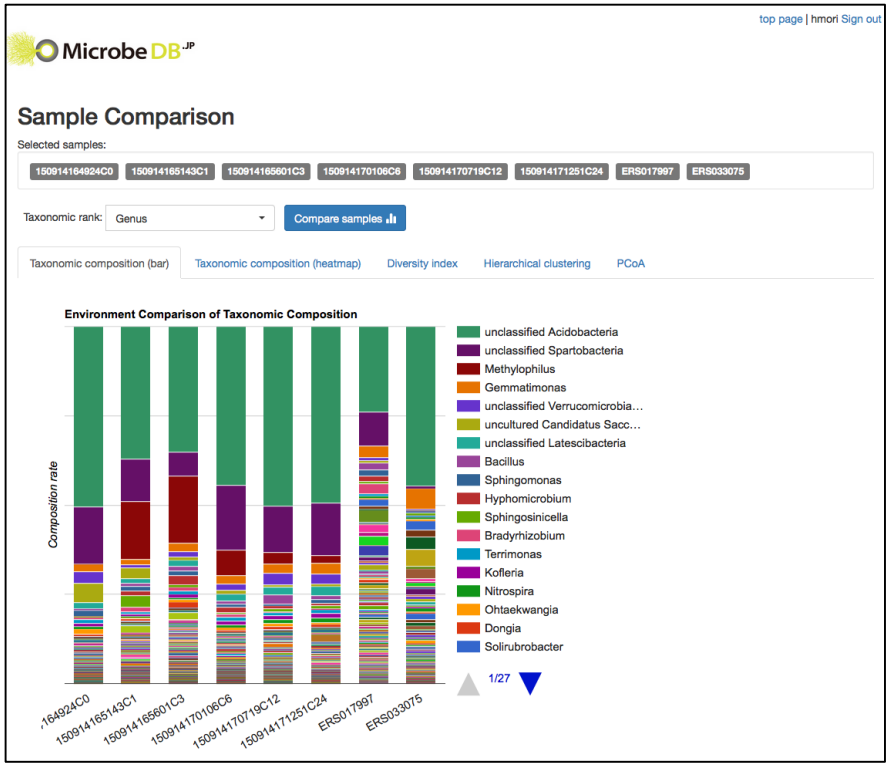
営利企業に対しては、プライベートデータの受け入れ、統一プロトコルによるそれらデータの解析、MicrobeDB.jpへのリンク、比較解析などに対して課金する事を検討している。

MiGAP&MeGAP連携による プライベートデータとの比較解析



KEGG	Streptobacillus moniliformis DSM 12112	Bacillus amyloliquefaciens LL3
Glycolysis / Gluconeogenesis	+	+
Citrate cycle (TCA cycle)	+	+
Pentose phosphate pathway	+	+
Pentose and glucuronate interconversions	+	+
Fructose and mannose metabolism	+	+
Galactose metabolism	+	+
Ascorbate and aldarate metabolism	+	+
Fatty acid biosynthesis	-	+
Fatty acid degradation	+	+
Synthesis and degradation of ketone bodies	+	+
Secondary bile acid biosynthesis	-	+
Ubiquinone and other terpenoid-quinone biosynthesis	-	+
Oxidative phosphorylation	+	+
Arginine biosynthesis	+	+
Purine metabolism	+	+

比較ゲノムStanza



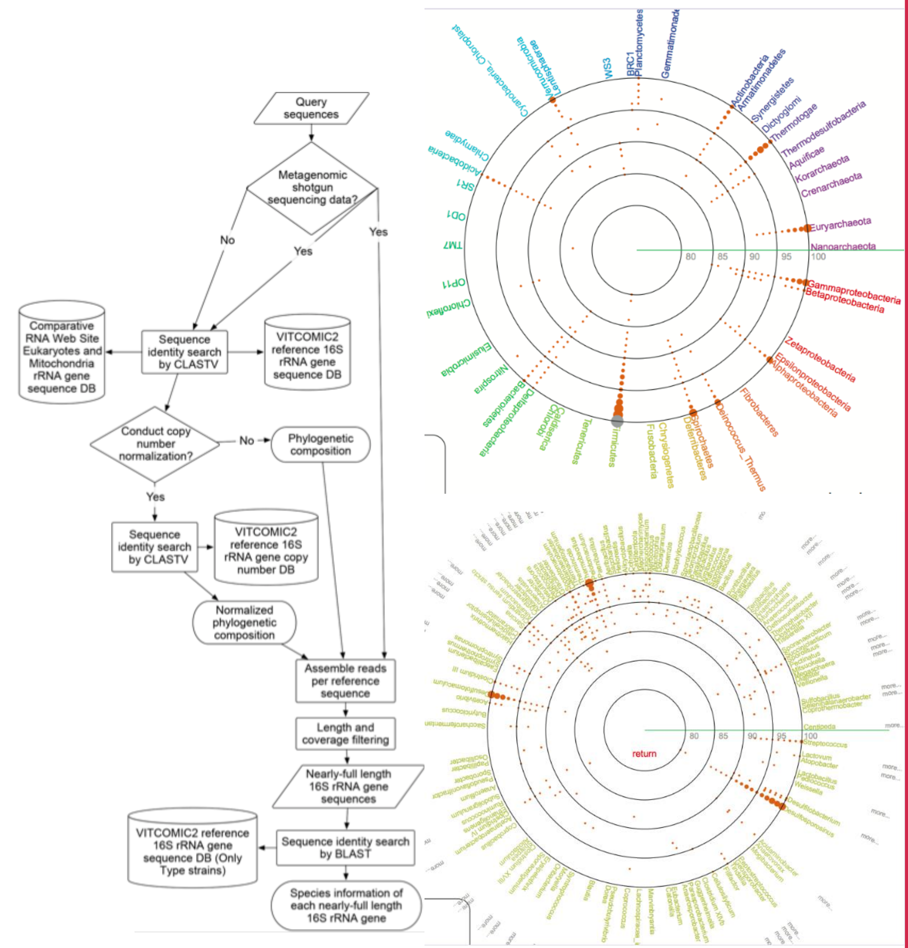
比較メタゲノムStanza

産業コミュニティからの利用促進方策

- 直感的なUIの実現、文章による検索の実現（非専門家に対する対応）
- ユーザ会の立ち上げ（第一回ユーザ会は2018年3月の日本ゲノム微生物学会年会で行う予定）
- 企業コンソーシアムとの連携
- 受託解析サービス会社との提携
- 学会、展示会での出展および講習会の開催（第一回講習会は2018年3月の日本ゲノム微生物学会年会で行う予定）
- キラーアプリケーションの開発およびライセンス契約

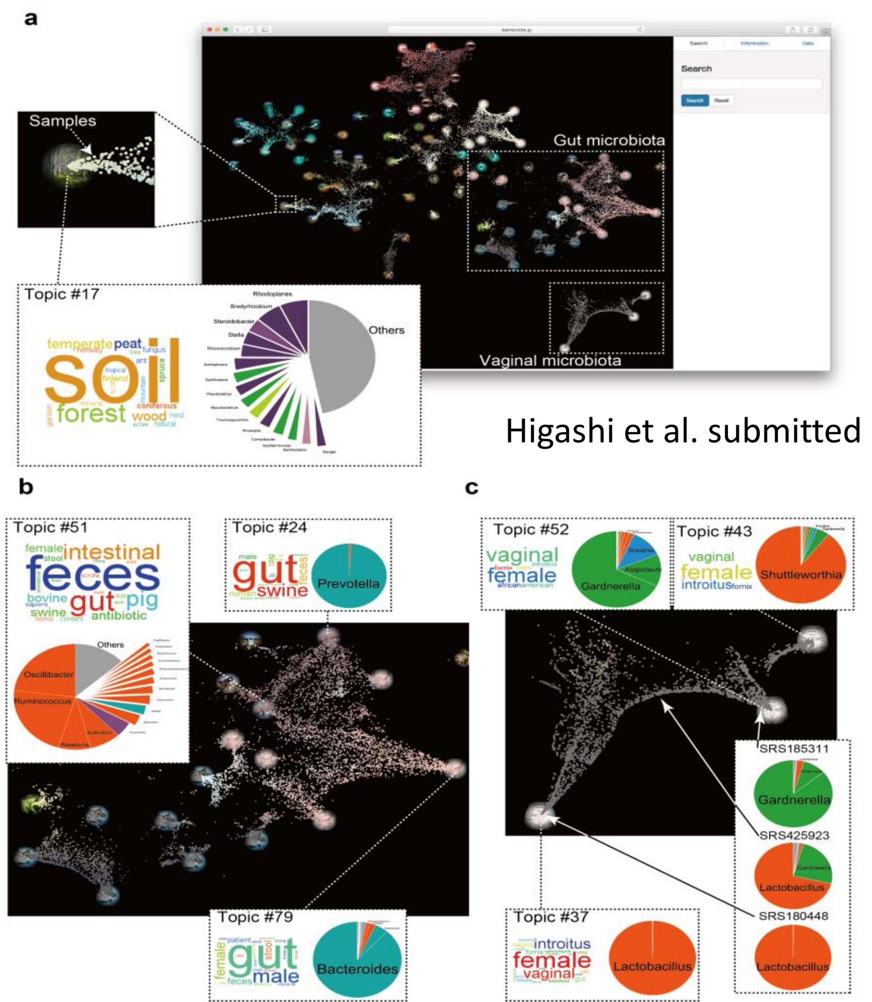
キラーアプリケーションとの連携

VITCOMIC2



Mori et al. submitted

LEA



Higashi et al. submitted

キラーアプリケーションとの連携

- データサイエンスで重要な事は、研究者の想像、発想、感性など…。
- それらphysicalな要素を具体化する作業が研究である。
- 一方で、統合DBはlogicの集合体でしかない。
- 統合DBをデータサイエンスで利活用するためには、logicalな集合体にphysicalな問いかけをする必要がある。

不動産検索アプリ「SUUMO」

今年度更新によるデータ数の推移予想

	MicrobeDB.jp version 2	MicrobeDB.jp version 3
prokaryote genome	16,983	53,500 (Max)*1
fungal genome	28	40?
metagenome	173,359	695,000
JCM & NBRC strains	16,671	17,500?
Chiba University Fungi strains	-	22,817

*1 INSDC Genome中のProkaryoteのDraft genome(Scaffold or Chromosome)とComplete genomeの合計数であり、実際はここからMicrobeDB.jpの基準によりクオリティフィルタリングされるので、あくまで最大値