

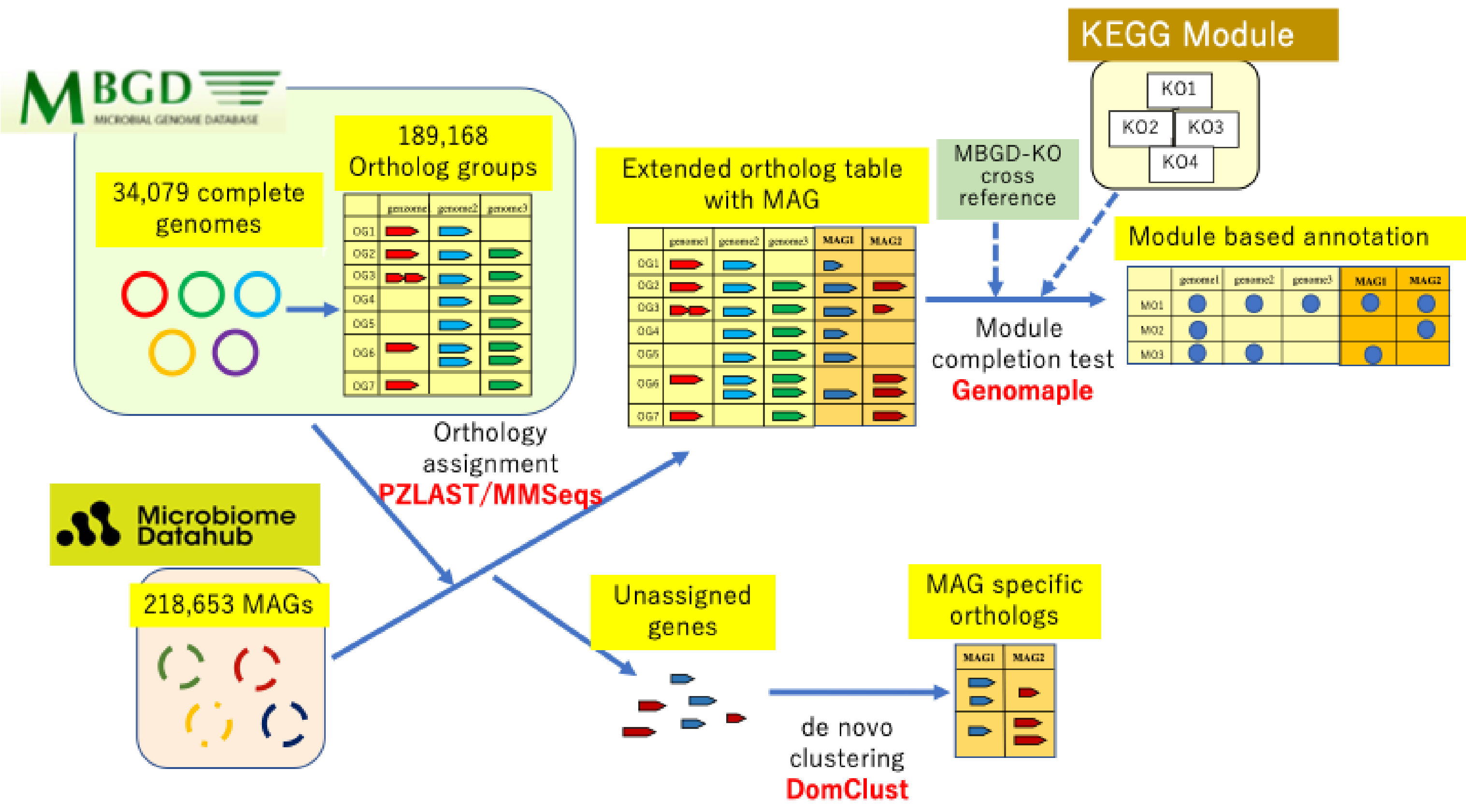


微生物比較ゲノムデータベースMBGDの大規模ゲノム・メタゲノムデータ解析への適用

内山郁夫、西出浩世、河合幹彦（基生研） 三原基広（ダイナコム）
千葉啓和（DBCLS） 高柳正彦（ウェブブレイン） 森宙史(遺伝研)

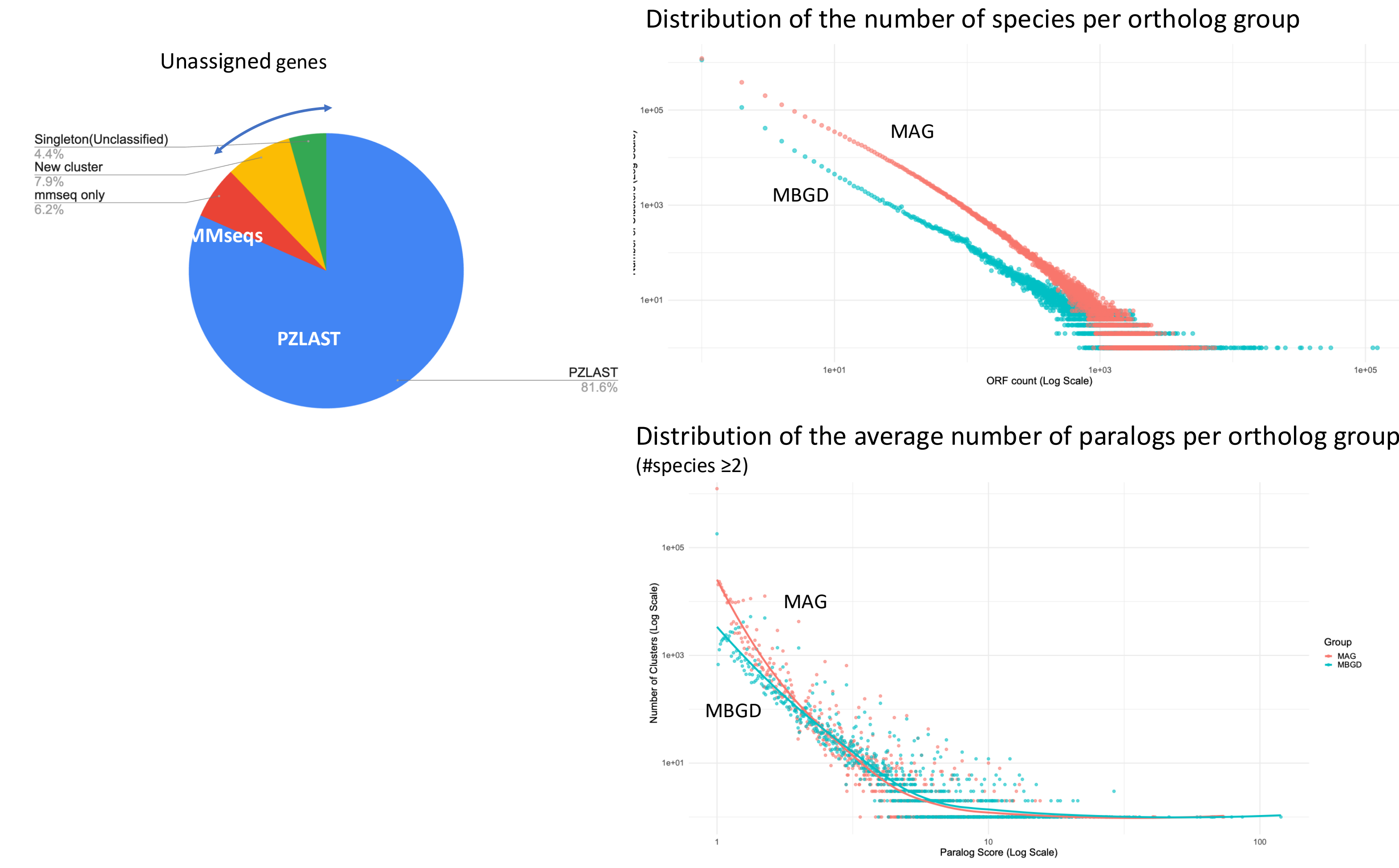
オーソログ解析に基づくMAGアノテーション

MBGDはオーソログ解析に基づいて公開された微生物のゲノム情報を整理したデータベースである。これまでに、MBGDのオーソログ情報を新規ゲノム解析に応用する目的で MyMBGD の開発を進めてきた。現在、この解析パイプラインを用いて、Microbiome Datahubに登録された21万MAGデータの解析を進めている。

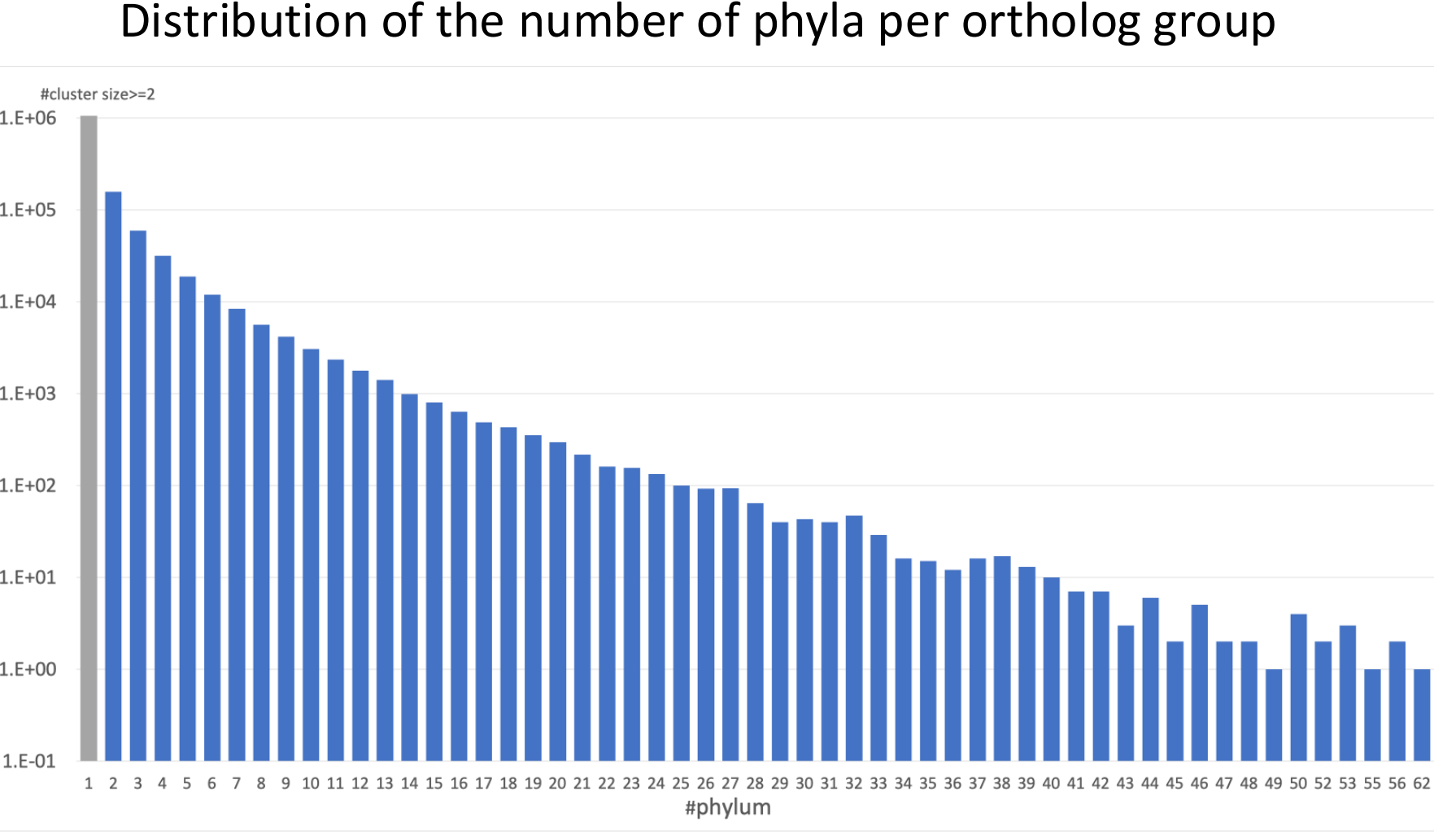


MBGDオーソログとヒットしないMAGの新規クラスタリング

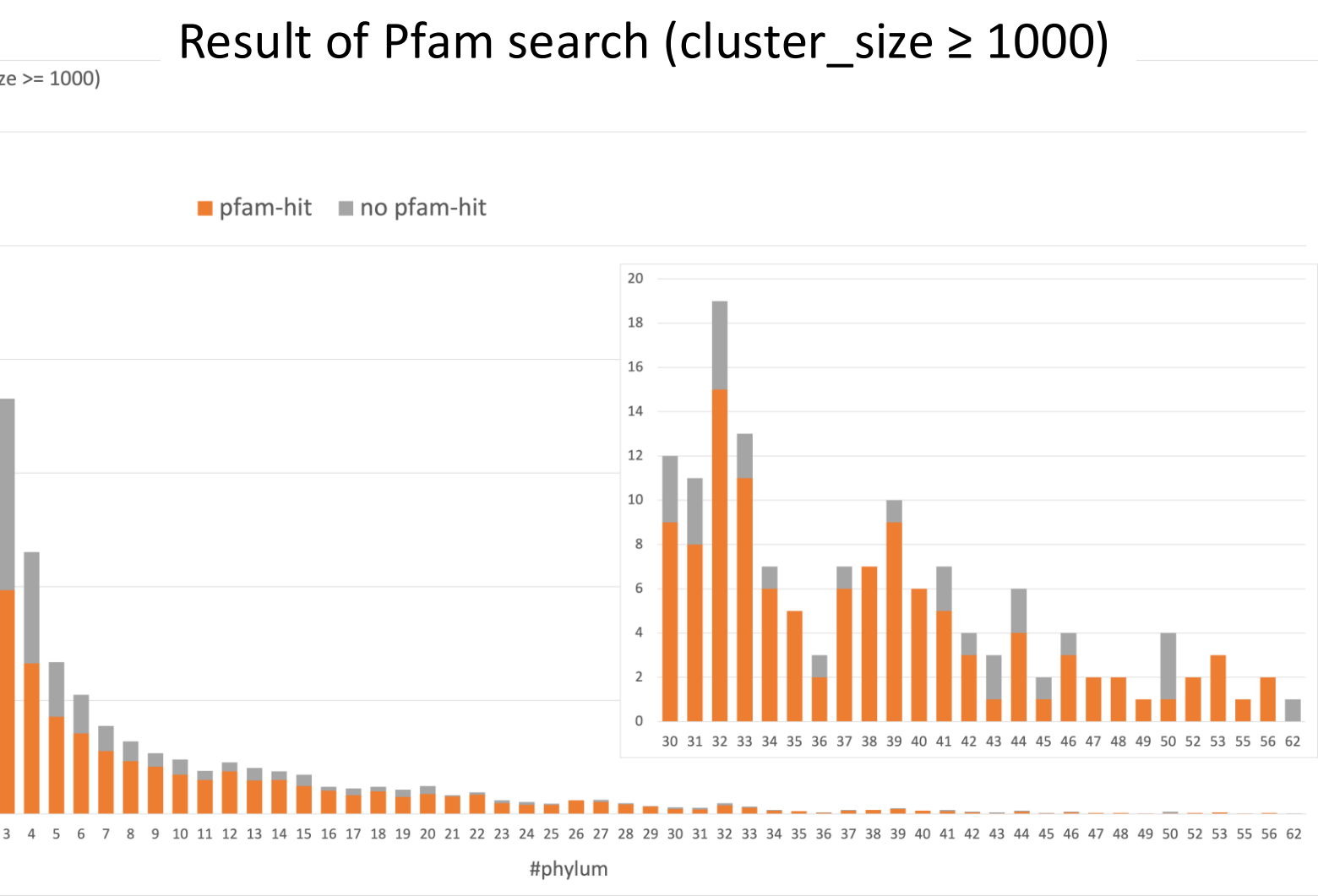
MBGDオーソロググループにヒットしなかった遺伝子は、DomClustを用いて新規のオーソログ分類を行った。5300万あまりのunassigned遺伝子が、Singleton含めて280万弱の新規オーソロググループに分類された。最大のオーソロググループは10000弱のMAGの遺伝子を含んでいた。



多様性の大きなオーソロググループを調べるため、オーソロググループ当たりの門(phylum)数の分布を調べたところ、数十門にまたがって出現するものがいくつかあった。

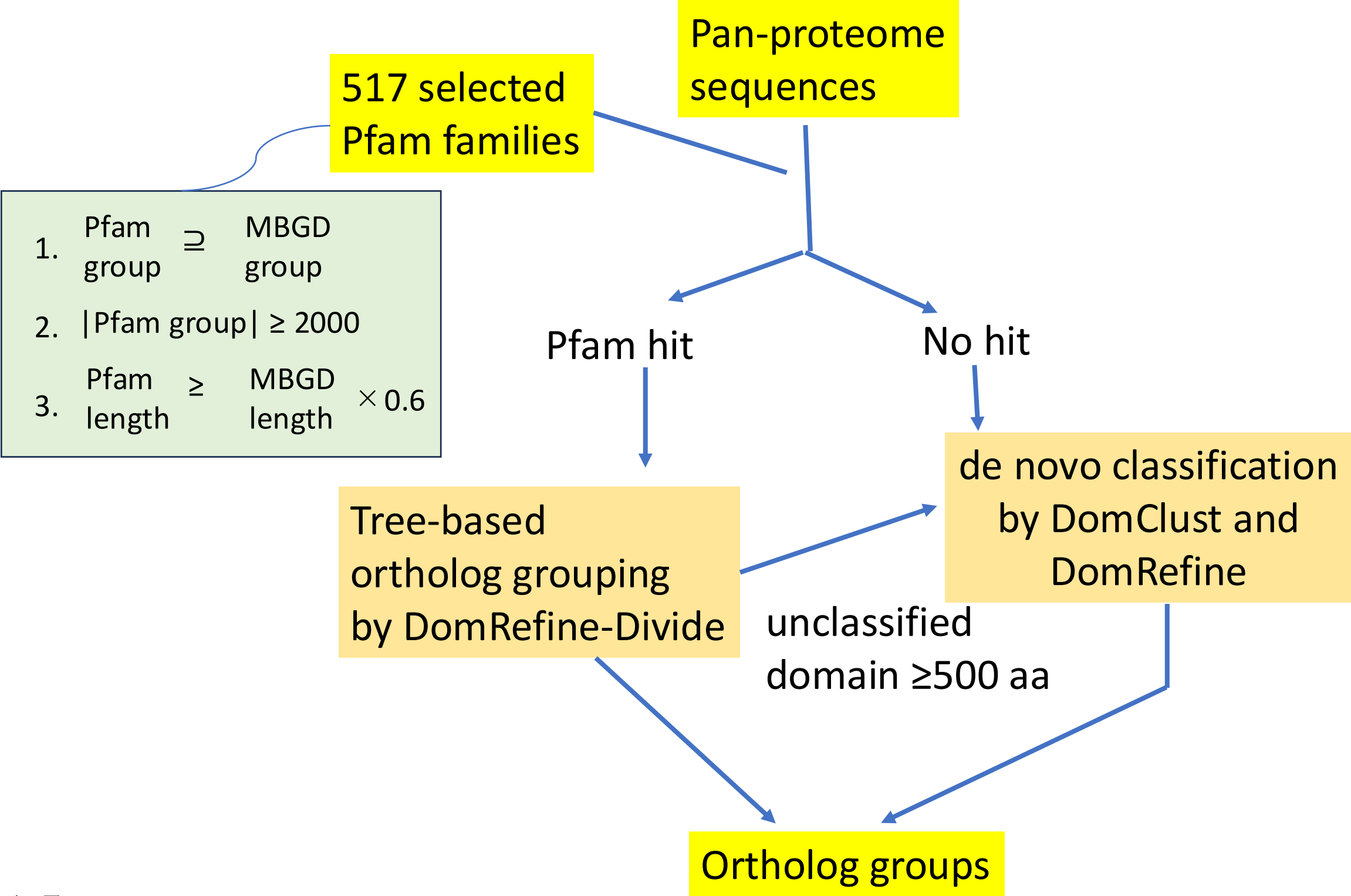
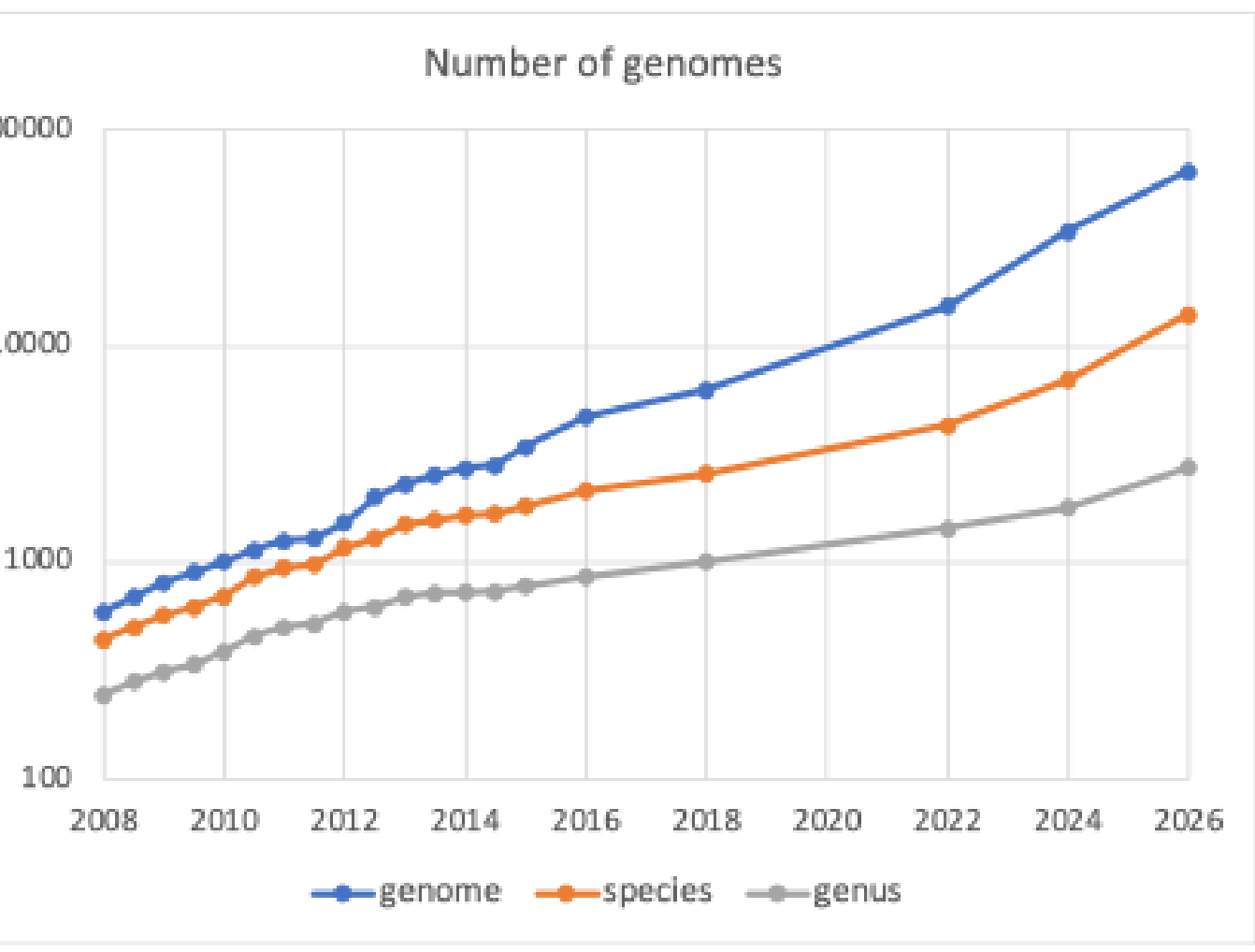


メンバー数が1000以上の新規オーソロググループについてPfam検索したところ、6割程度がPfamにヒットした。その中で多くの新規オーソロググループに出現するPfamエントリーを列挙した。



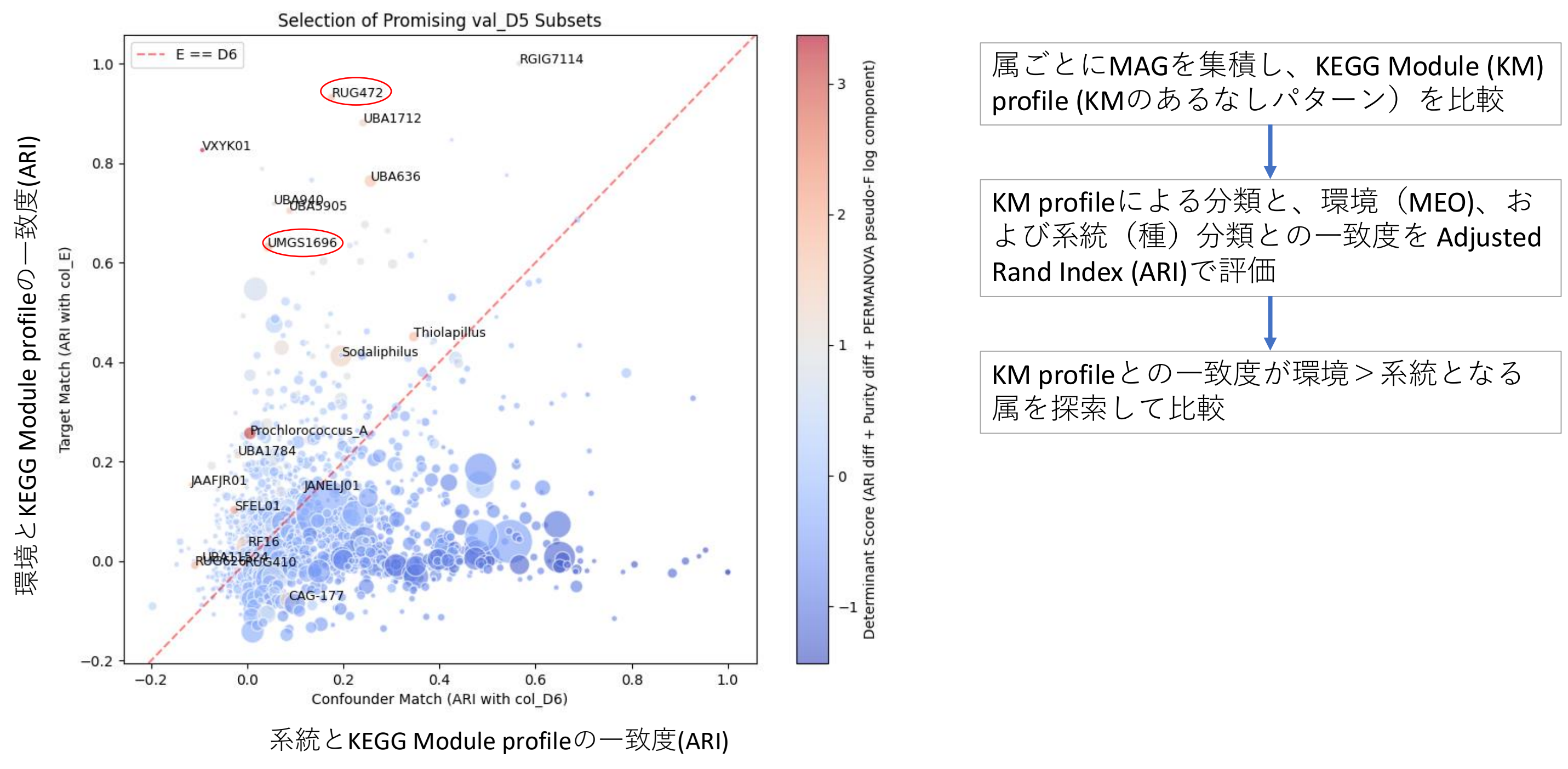
オーソロググループの差分更新

公開されたゲノムデータは指数関数的に増大しており、オーソログ解析に必要な計算資源の増大が切実な問題になってきている。ドメイン単位の分類というMBGDの特徴を生かしつつ、この問題に対処するため、安定的なホモロジードメインのデータベースであるPfamをベースにした差分更新の体制を検討している。

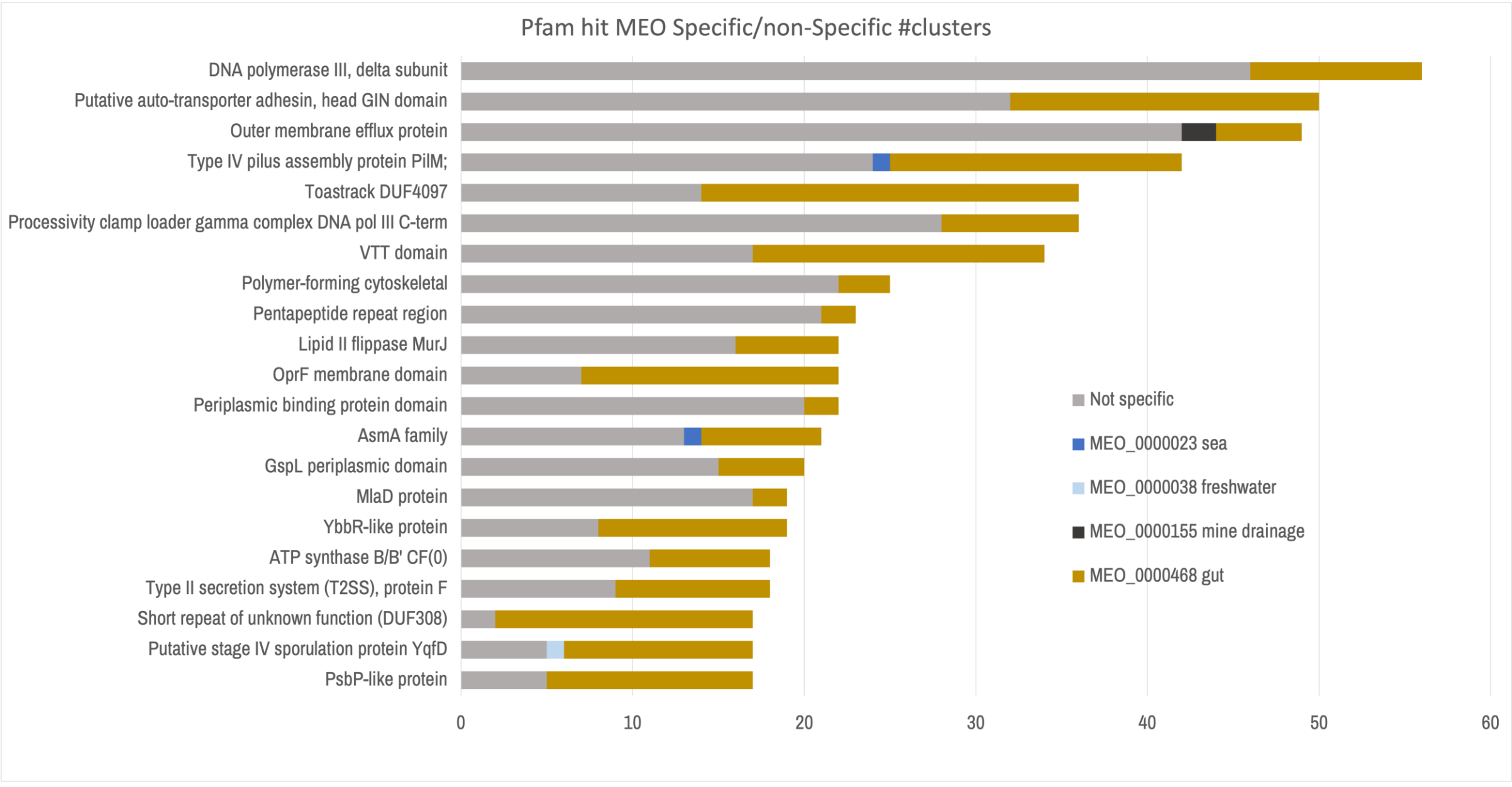
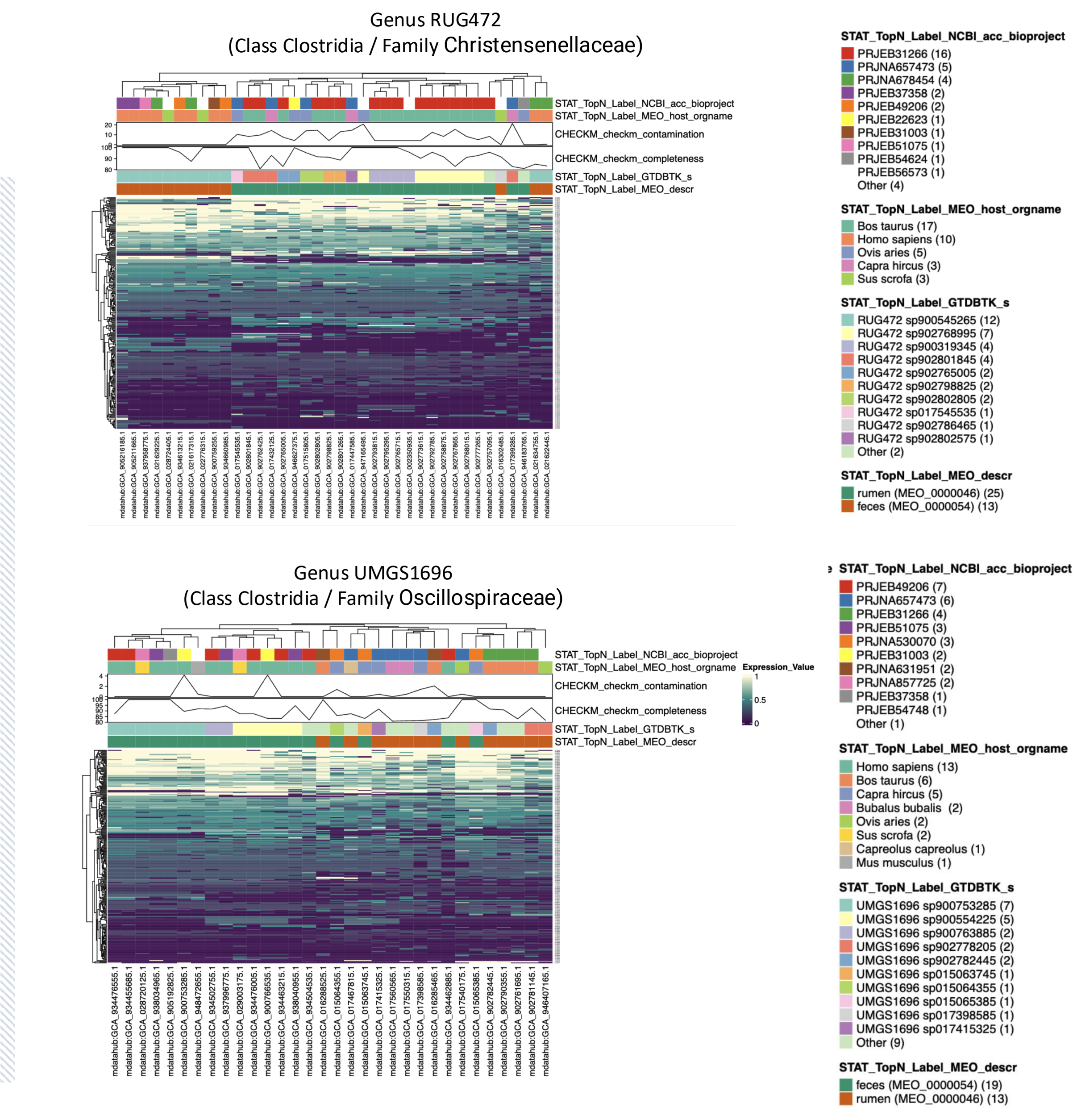


MAGオーソログ解析に基づくゲノム機能-環境マッピング

MBGDオーソロググループにヒットした遺伝子は、MBGD-KOマッピングを通じてKEGG Moduleによるゲノム機能推定を行った。さらにMAGにアサインされたMEOによる環境情報と結びつけることで、ゲノム機能-環境マッピングを行った。この情報を用いて、近縁種間で、系統よりは環境によってゲノム機能が支配されているケースの探索を行っている。



属ごとにMAGを集積し、KEGG Module (KM) profile (KMのあるなしパターン) を比較
↓
KM profileによる分類と、環境 (MEO)、および系統 (種) 分類との一致度を Adjusted Rand Index (ARI) で評価
↓
KM profileとの一致度が環境 > 系統となる属を探索して比較



DomClust 実行効率の変化

	MBGD従来版	517 Pfamを除去	割合
ゲノム数	1,812	1,812	100%
配列数	9,214,351	7,712,104	83.7%
アミノ酸数	2,954,621,833	2,365,830,863	80.1%
ホモロジーデータサイズ	606G	231G	38.1%
クラスター数	1,451,649	1,380,407	95.1%
最大サイズ	19,570	16,909	86.4%
平均サイズ	10.88	8.78	80.7%
計算時間	22:38:00	6:08:00	27.1%
使用メモリ	2TB	759GB	38.0%