55 バイオサイエンスデータベースの俯瞰調査

NBDC事業推進部

科学技術振興機構 NBDC事業推進部（NBDC）

サマリー

バイオサイエンスデータベース（以下、バイオDB）整備政策立案の基礎資料とするため、世界の動向と日本の立ち位置を明らかにすることを目的に、世界のバイオDBの整備状況を調査・分析した。

NAR、FAIRsharing、Integbio、Database Commons、Web of Scienceを用いてDBリストを作成し、調査時点でアクセス可能だった3,881件について、Species、Data Type、Number of Data elements、Data Domain Classification、Type of Database、Access Type、Functional Classification、Development Year、Last Known Updated Date、Funding Source、Country of Originなどの情報を取得。これらをカテゴリごと、期間ごとにまとめるとともに、その現状と動向を分析した。

2000年以降、バイオDBの開発数は年々増加してきたが、ここ3年は若干鈍化傾向が見られた。新規に開発されたDB数で日本は世界2位だったが、被引用数では世界トップ131 DB中、日本のDBは1DBに留まった。日本のトップ30 DBでは、かずさDNA研、京大、理研、農研機構、東大の上位5機関が7割を占めた。モデル生物種から非モデル生物種、単一生物種から複合生物種といったバイオDBの多様化・多機能化が進むとともに、生物種やデータ種を超えたデータ統合が進んでいることが明らかとなった。分野別ではヒト、健康医療分野のDBが増加していた。海外と比較すると、日本は植物、昆虫、無脊椎動物に強みがあると思われた。

調査方法

① DBリストの作成

•以下のDBレポジトリを用いてDBリストを作成。

- ✓**NAR** (1)
- ✓**FAIRsharing** (2)
- ✓**Integbio** (3)
- ✓**Database Commons** (4)

(1) https://www.oxfordjournals.org/nar/database/c
(2) https://fairsharing.org/
(3) https://integbio.jp/dbcatalog/?lang=en
(4) https://bigd.big.ac.cn/databasecommons/

•**Web of Science** (WoS)を用いて新規DB関連の論文を特定し、それらを読み込み調査することで近年になって構築されたDBを特定しリストに加える。

5,986件のDBリスト

アクセス可否の調査

3,881件のDBリスト

② 各DBの情報取得

•各DBにアクセスし、調査時点でアクセス可能だった3,881件の、生物種, データタイプ, 収載データ数, データドメイン分類, データベースタイプ, アクセスタイプ, 機能別分類, 開発年次, 最終更新日, 開発資金提供元, 主な開発国などの情報を取得。

③各DBに関連する文献数と引用数の取得

•DBリストの各DBの名称・URL等を使用してWoSで検索し、それぞれの文献ヒット数を調査。

•DBの開発論文を引用している文献数（被引用数）、その中のTop 10%論文数、国際共著数を調査。

•2019~2021年など3年ごとの区間にわけ、各区間の文献数をそれぞれ取得。

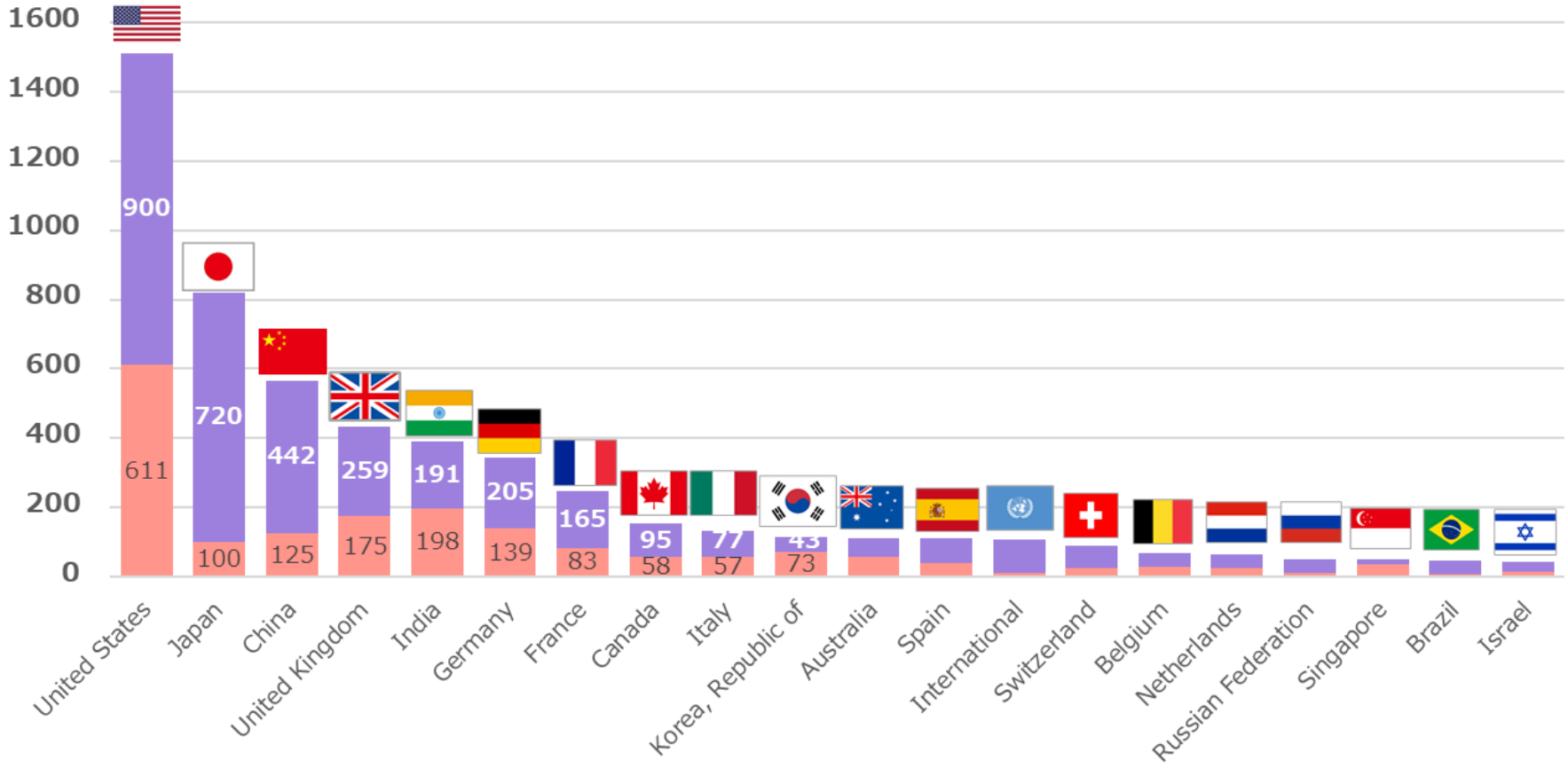
④ 分析

•取得した各DBの情報・分類、文献ヒット数等のデータをパラメーターとして、国別、機能別、論文の被引用数での国別ランキング、直近3年間に開発されたDBの国別ランキングやDBのカテゴリーごとのランキングなどをまとめた。

•さらに、対象とする生物種やータ種などの再分類して年代別にカウントし、新規に開発されたバイオDBのトレンドを調べるとともに、地域別の特徴を分析した。

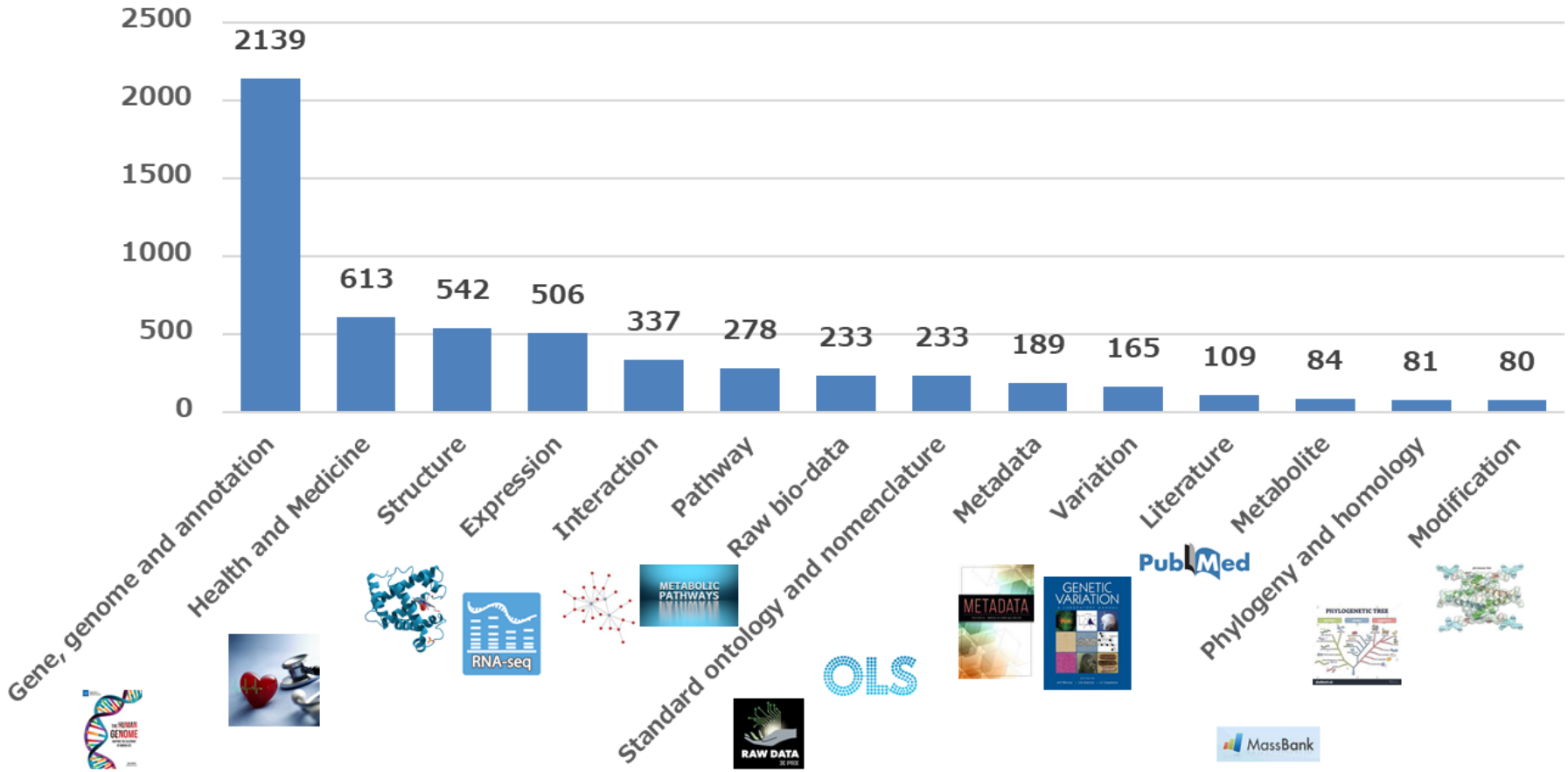
世界のDBの構築状況（国別DB数）

✓USが最も多く1,511件。
✓次いで日本、中国、UK、インドの順。
✓件数Top10の国でアクセス可能なDBの割合が最も多かったのが日本（88%）。
✓一方で、韓国では37%と低く、USも60%に留まっていた。



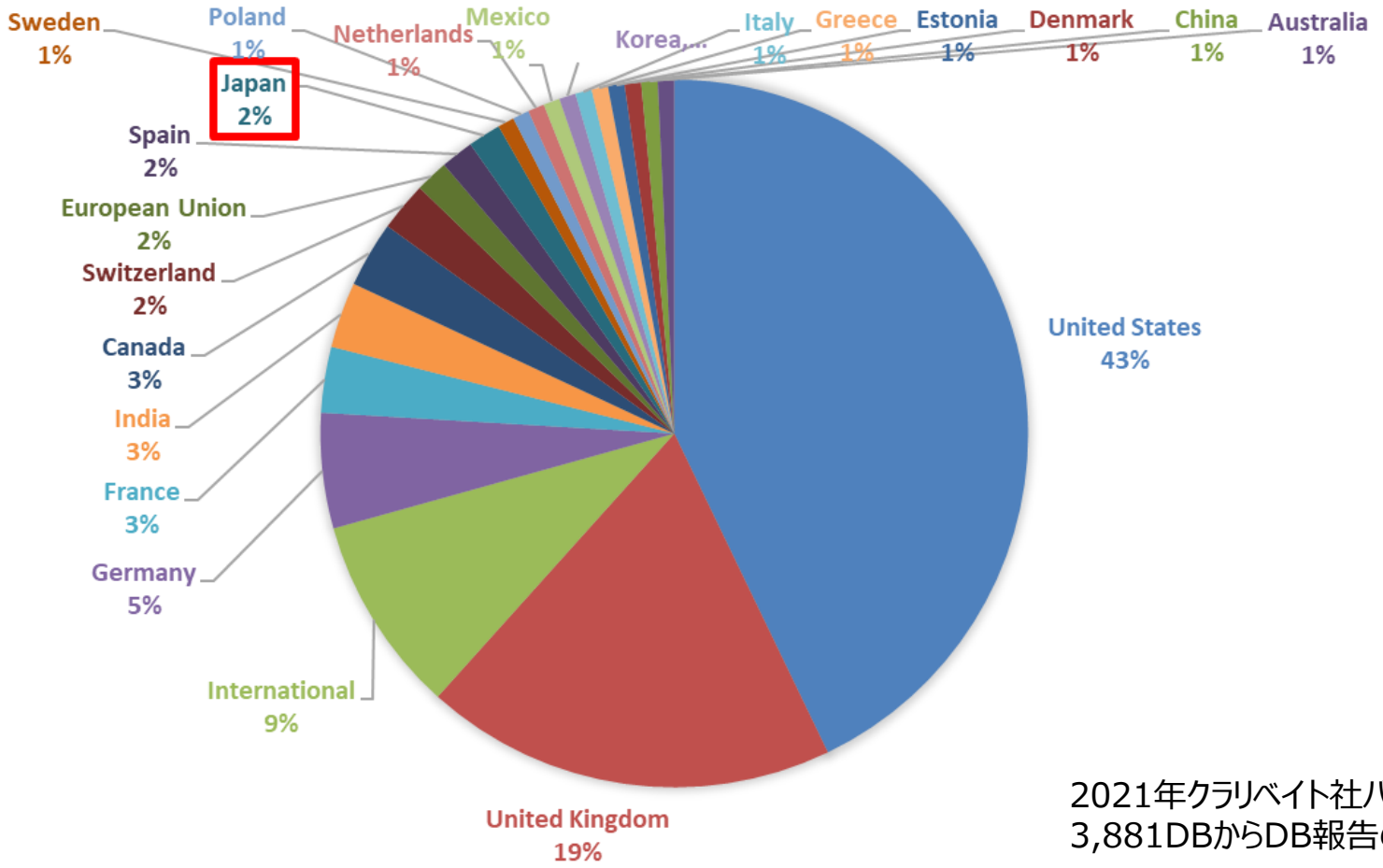
機能別のDB構築状況（Functional classification）

✓「Gene genome and annotation」が最も多く2,139件。
✓次いで「Health and Medicine」,「Structure」,「Expression」,「Interaction」,「Pathway」,「Raw bio-data」と続いた。



被引用回数1000回以上のトップ133バイオDB（国別）

✓USが最も多くトップ133の43%を占めた。
✓次いでUK 19%、上位2ヶ国でトップ2/3弱を占めた。
✓日本は2%にとどまった。



2021年クラリベイト社バイオDB俯瞰調査でアクセス可能な
3,881DBからDB報告の文献の引用回数が高いDBを抽出

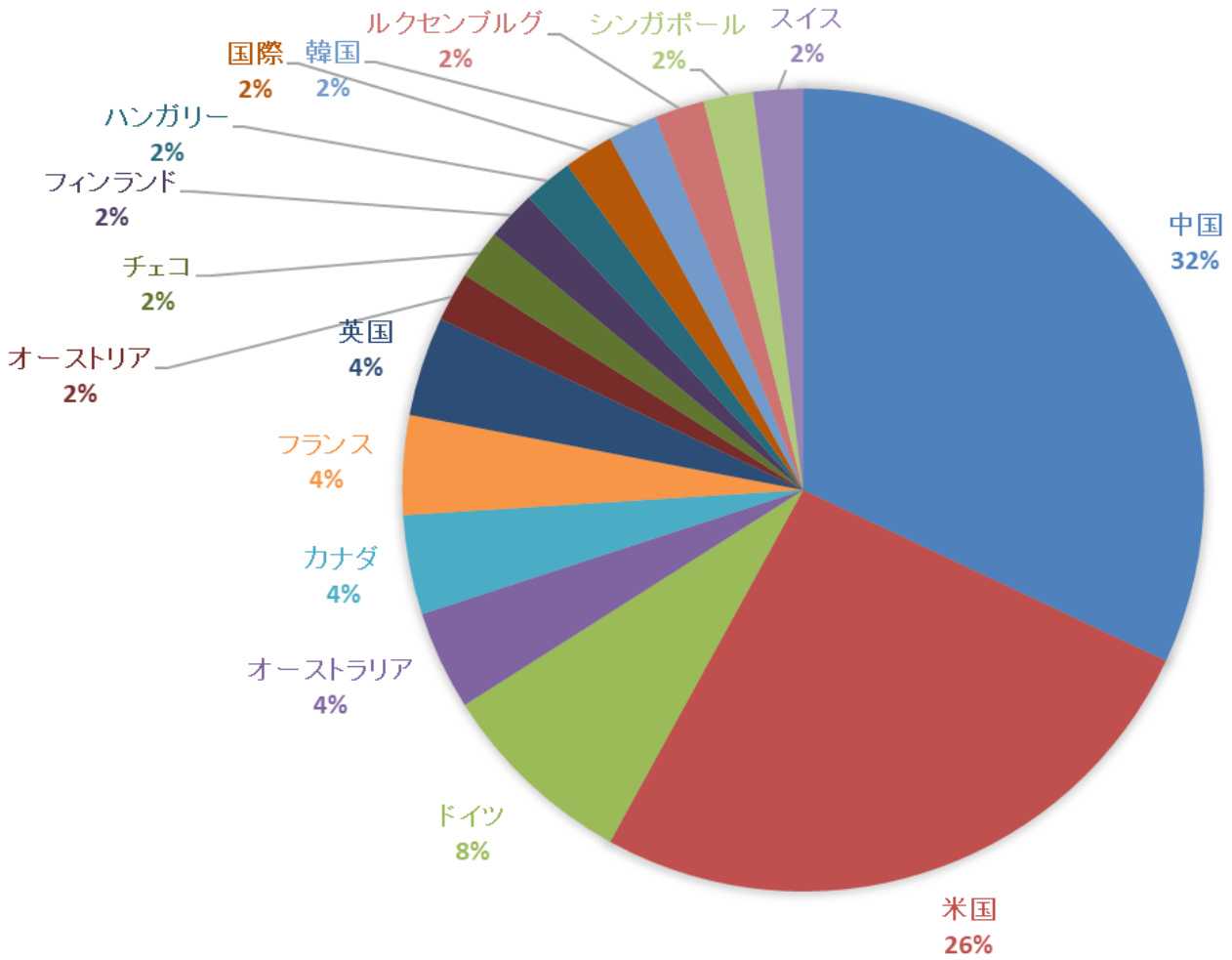
国内バイオDBランキング（被引用回数150回以上）

✓かずさDNA研５件、京大５件、理研４件、農研機構４件、東大３件で、上位5機関で日本のトップ30 DBの7割を占めた。

	589DS001 - DATABASE ANNOTATION - *Revised - 20-Dec-21*			
	Full Name	Short Name	機関	被引用数
1	Saccharomyces Genome Database (SGD)	SGD	スタンフォード大	2035
2	UMIN Clinical Trials Registry	UMIN-CTR	東京大学	1170
3	Codon Usage Database	Codon Usage Database	かずさDNA	867
4	DNA Data Bank of Japan	DDBJ	遺伝学研究所	842
5	Food Metabolome Repository	Food Metabolome Repository	かずさDNA	766
6	Arabidopsis thaliana trans-factor and cis-element prediction database	ATTED-II	東京工業大学	676
7	DBGET Search	DBGET Search	京都大学	629
8	Functional Annotation of th Mammalian Genome	FANTOM	理化学研究所	608
9	The Rice Annotation Project	RAP-DB	農研機構	573
10	Mitochondrial Genome Database of Fish	Mitofish	東京大学	482
11	Hepatitis Virus Database	Hepatitis Virus Database	名古屋市立大学	457
12	PRIMe	PRIMe	理化学研究所	442
13	DataBase of Transcriptional Start Sites.	DBTSS	東京大学	428
14	Protein Data Bank Japan	PDBj	大阪大学	425
15	Rice Expression Profile Database	RiceXPro	農研機構	425
16	Database of Transcriptional Regulation in Bacillus subtilis	DBTBS	東京大学	417
17	Multiple alignment program for amino acid or nucleotide sequences	MAFFT-DASH	農研機構	415
18	H-Invitational Database	H-Invitational Database, an integrated database of human genes and transcripts	東海大学	403
19	Functional Annotation of the Mammalian Genome	FANTOM5	理化学研究所	374
20	Functional RNA Analysis Database	fRNAdb	理化学研究所	306
21	Microbial Genome Database	MBGD	遺伝学研究所	299
22	KNApSAcK family databases	KNApSAcK family databases	奈良先端大学	289
23	Coexpression Database	COXPRESdb	東北大学	257
24	Jatropha Genome Database	Jatropha Genome Database	かずさDNA	255
25	Q-TARO	Q-TARO	農研機構	248
26	LIGAND	LIGAND	京都大学	228
27	Human Unidentified Gene-Encoded Large Protein Database	HUGE	かずさDNA	219
28	Kyoto Encyclopedia of Genes and Genomes- GLYCAN	KEGG GLYCAN	京都大学	191
29	KEGG EXPRESSION	KEGG EXPRESSION	京都大学	166
30	Japan ProteOme STandard Repository	jPOSTrepo	京都大学	160
31	Kazusa Marker DataBase	Kazusa Marker DataBase	かずさDNA	151

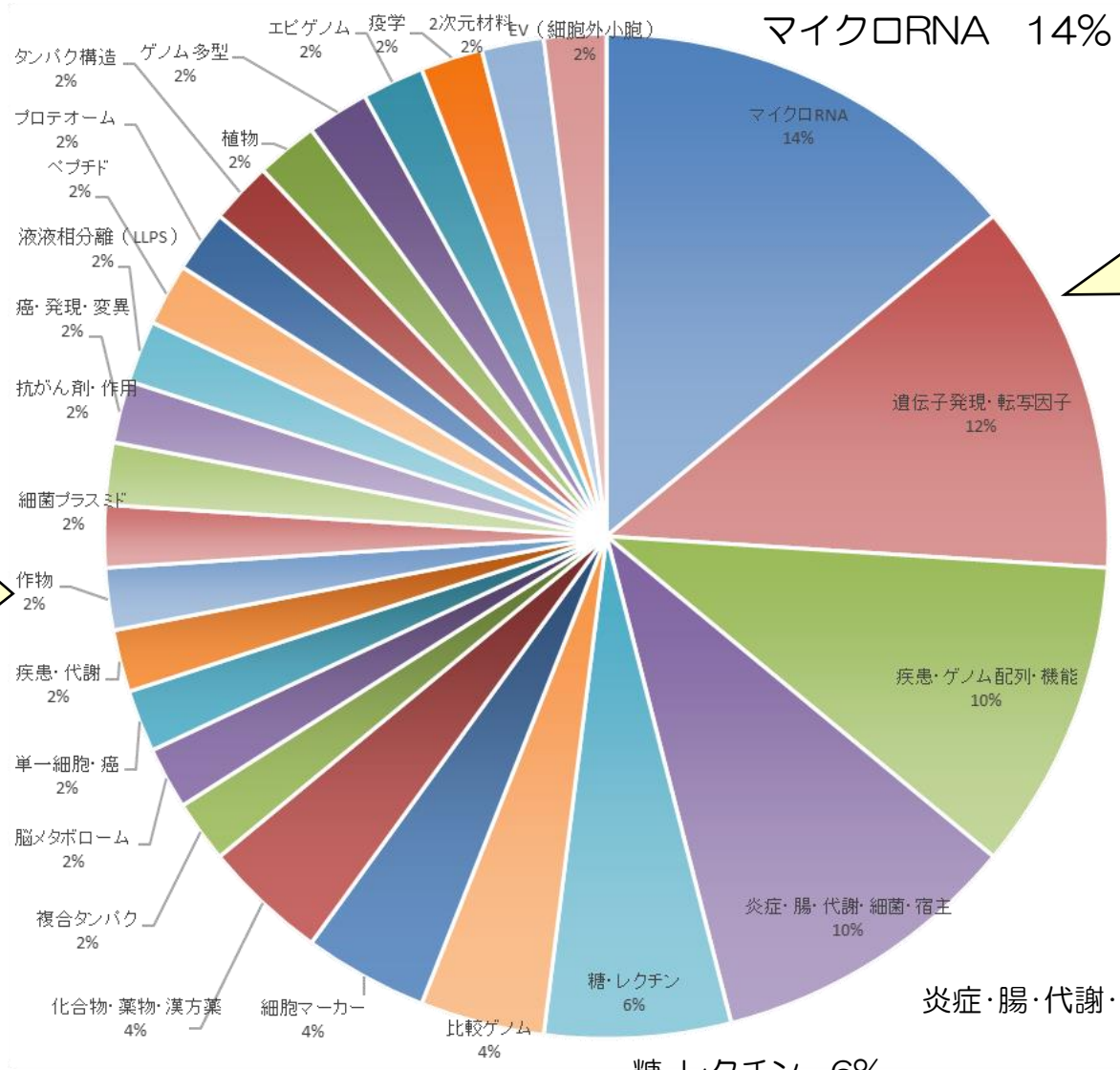
2019年以降にリリースされたDBのランキング TOP50（国別）

✓中国とUSで58%を占めた。
✓日本はTOP50に1DBもランクインせず。



クラリベイト社バイオDB俯瞰調査2021-2022

2019年以降にリリースされたDBランキング TOP50（対象別）



マイクロRNA　14%

✓ マイクロRNA、遺伝子発現、疾患ゲノム、腸内細菌など近年研究が盛んな対象に関するDBが全体の約50%を占めた。

遺伝子発現･転写因子　12%

疾患･ゲノム配列･機能　10%

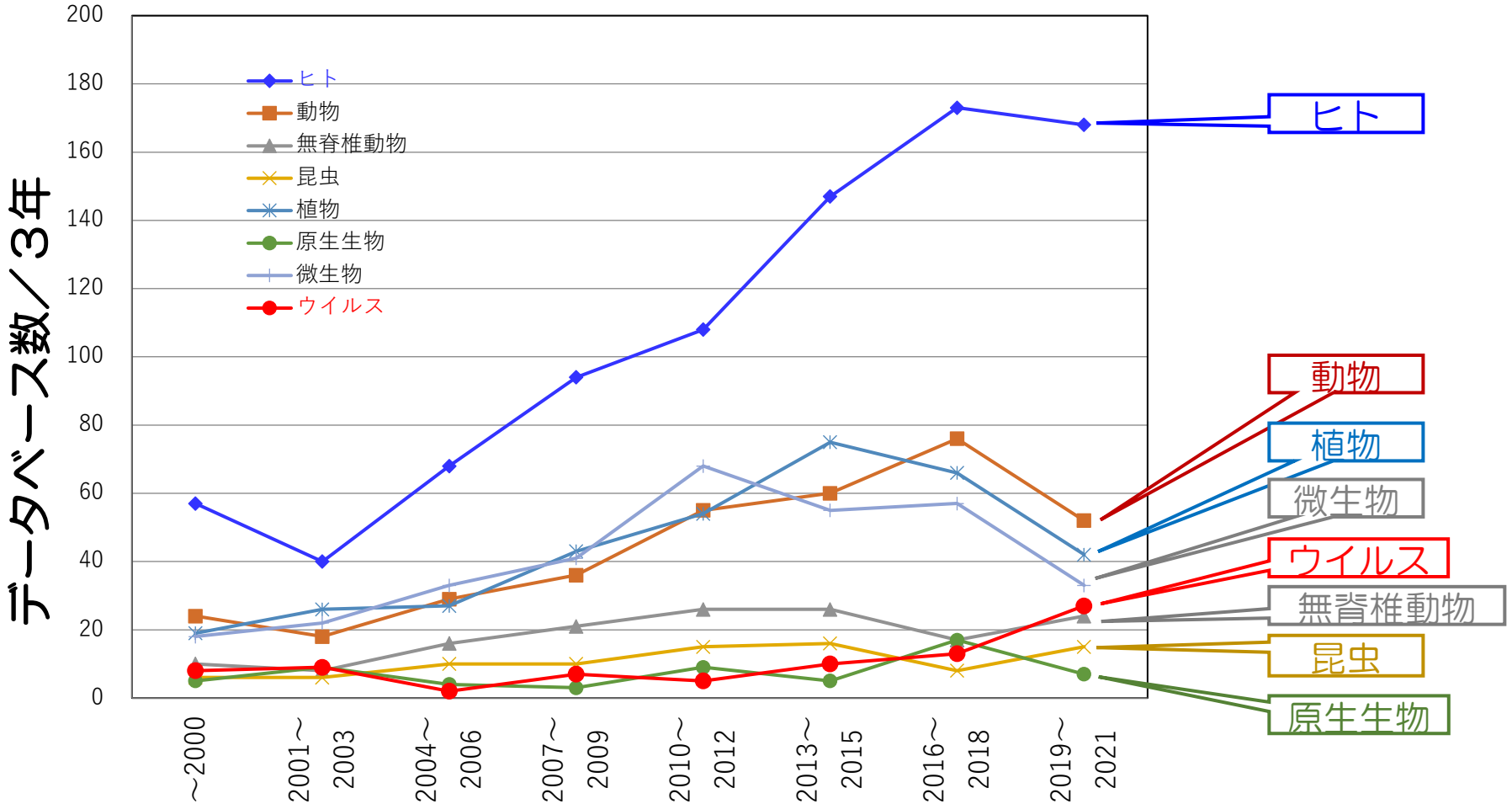
炎症･腸･代謝･細菌･宿主　10%

糖･レクチン　6%

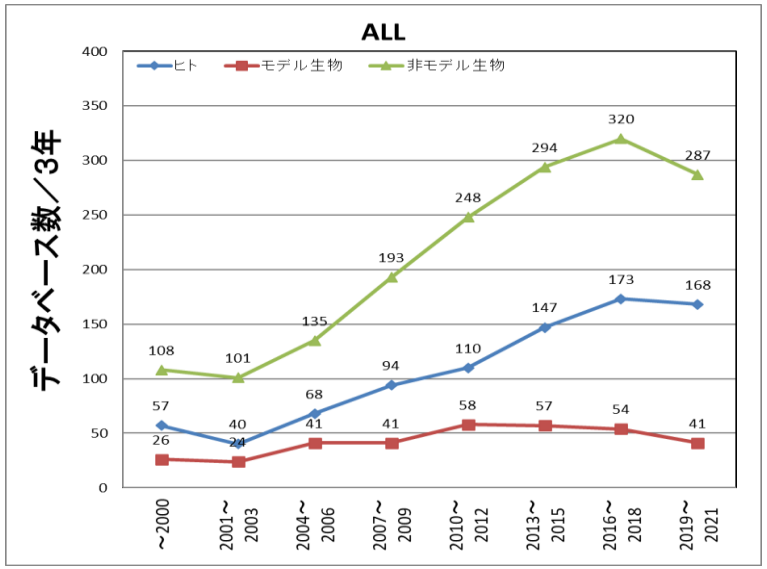
✓ 一細胞解析、細菌、がん変異、LLPSなどの先端技術がランクインした。

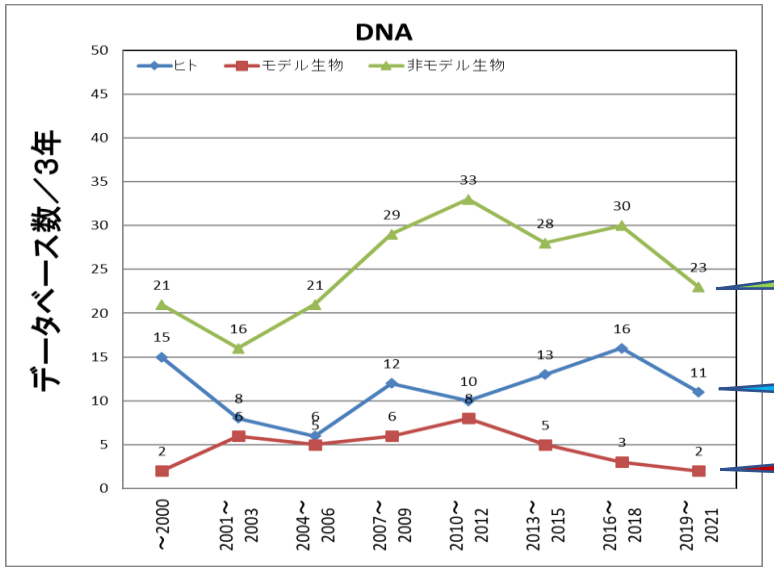
新規に公開されたDB数の推移（生物種別）

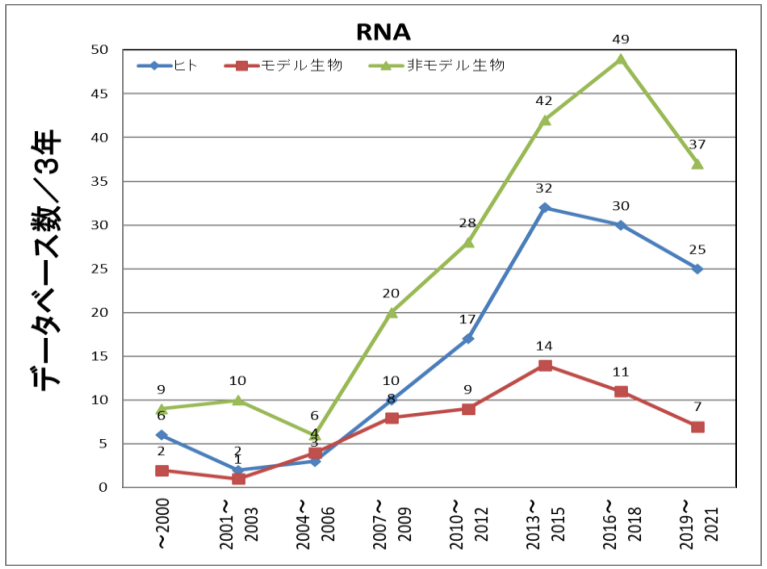
✓2000年以降年々増加してきたが、ここ3年は多くの生物種で鈍化傾向が見られた。
✓直近3年間では「ウイルス」が急増した（新規感染症拡大の影響か？）。

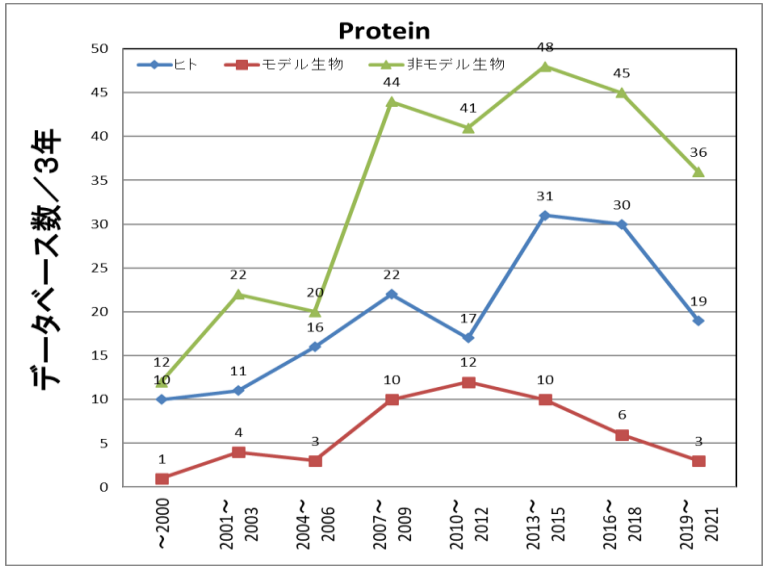


新規に公開されたDB数の推移（モデル生物/非モデル生物）









非モデル生物

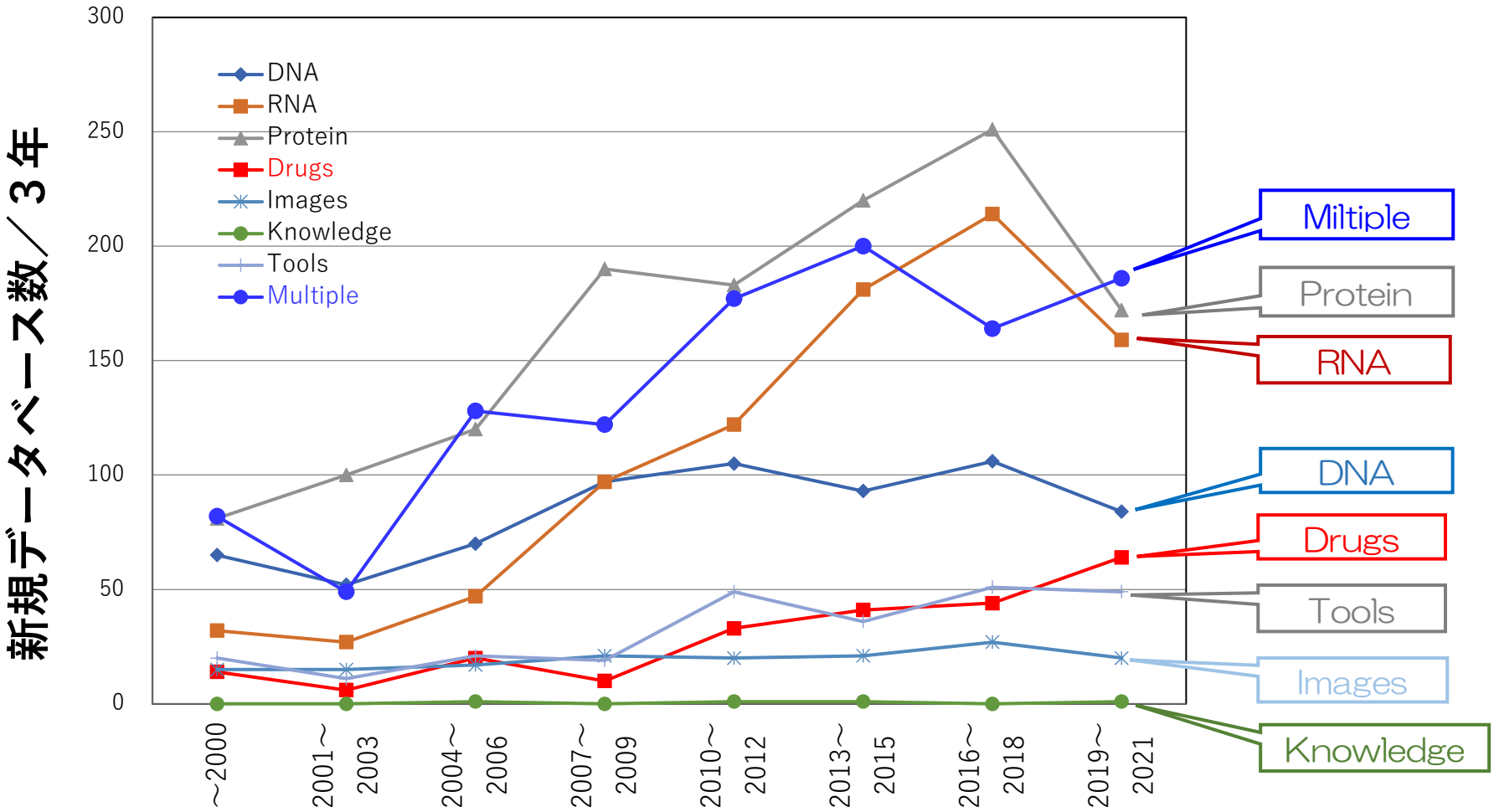
ヒト

モデル生物

2004年以降、
特にRNAとProteinで、
ヒトと非モデル生物のDBが
増加した。

新規に公開されたDB数の推移（データ種）

✓「Multiple」の増加からデータ種を超えた統合が進んでいることが読み取れた。
✓直近3年間では「Drug」が増加した。



新規に公開されたDB数の推移（生物種×データタイプ）

✓世界でヒト･動物･植物･微生物のDNA･RNA･タンパク質･その他が多い。
✓日本は、生物種では無脊椎動物･昆虫･植物のDBの割合が高い。
✓日本は、データ種ではDNAとMultipleの割合が高い。

世界	真核生物	原核生物	その他	マルチ	ヒト	動物	無脊椎動	昆虫	植物	原生生物	微生物	ウイルス	情報なし	合計
DNA	184	26	3	86	97	63	20	17	53	14	34	6	119	722
RNA	210	25	0	153	131	61	27	14	44	9	30	7	216	927
Protein	225	61	0	222	162	48	14	8	34	9	80	27	497	1,387
Drugs	70	5	0	20	66	5	1	0	3	0	5	3	69	247
Images	43	3	0	28	22	7	8	7	9	4	5	0	49	185
Knowledge	0	1	0	1	0	0	0	0	0	0	1	0	5	8
Tools	57	5	0	43	26	11	8	3	16	5	7	3	87	271
Multiple	271	45	1	176	151	59	35	19	75	6	57	15	306	1,216
Other	469	85	8	268	252	119	51	29	138	16	128	26	680	2,269
情報なし	0	0	0	0	0	0	0	0	0	0	0	0	0	0
合計	1,529	256	12	997	907	373	164	97	372	63	347	87	2,028	7,232

日本	真核生物	原核生物	その他	マルチ	ヒト	動物	無脊椎動	昆虫	植物	原生生物	微生物	ウイルス	情報なし	合計
DNA	76	4	1	25	22	20	15	12	32	9	5	1	28	250
RNA	24	0	0	17	12	5	4	3	8	2	1	0	19	95
Protein	25	3	0	29	20	5	0	0	5	1	3	0	57	148
Drugs	11	0	0	3	10	0	0	0	1	0	0	0	12	37
Images	21	2	0	17	7	6	5	4	5	4	2	0	20	93
Knowledge	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tools	5	1	0	1	1	0	0	0	3	1	1	0	5	18
Multiple	78	4	0	57	37	18	17	6	24	1	7	2	80	331
Other	0	0	0	0	0	0	0	0	0	0	0	0	129	129
情報なし	0	0	0	0	0	0	0	0	0	0	0	0	0	0
合計	240	14	1	149	109	54	41	25	78	18	19	3	350	1,101

調査した5,986データベース中、アクセス可能だった3,881データベースを分析し、DBの延べ数を集計（重複あり）。
データタイプは、クラリベイトの調査による。生物種分類は、NBDCにてキュレーションしたもの。

まとめ（地域別のトレンドの概要）

地域	生物種	データ種	機能
世界	•ヒトが増加 •ウイルスが急増 •非モデル生物が増加 •微生物、植物、動物は減少	•ツールが増加 •医薬品が増加 •タンパク質、RNAが増加 •Multipleが増加	•Annotationが増加 •Health & Medicineが増加
日本	•無脊椎動物・昆虫、植物の割合が高い	•DNAが多い •Multipleが多い	•Gene, genome & annotaionが多い
米国 ＋カナダ	•ヒト、微生物、ウイルスの割合が高い	•ペプチド（タンパク質）が多い •Multipleが多い	•Gene, genome & annotaionが多い •Annotationが多い
西欧地域	•ヒト、微生物、ウイルスの割合が高い	•ペプチド（タンパク質）が多い •Multipleが多い	•Gene, genome & annotaionが多い •Annotationが多い
中国 ＋台湾	•ヒト、動物、植物の割合が高い	•Multipleが多い •RNAが多い	•Gene, genome & annotaionが多い •Expressionが多い •Health & Medicineが多い

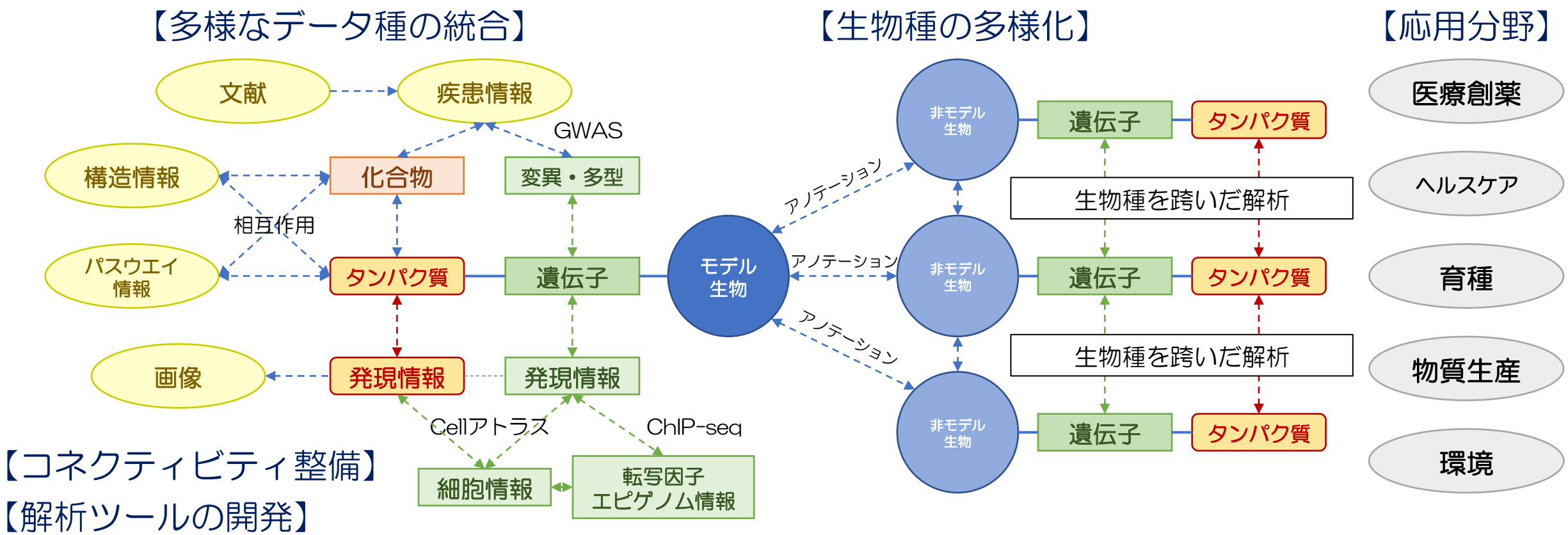
☞ **世界的なトレンド**　ヒト・疾患、医薬品関連のDB数が伸びる。特にウイルス関連のDBが急増。
複数生物種を含むDBや多機能なDBが増加。

☞ **日本の特徴**　他の地域に比べると、ヒト・疾患関連のDBの割合が低い。
無脊椎動物・昆虫・植物などの割合が高い。

☞ **米国・西欧地域の特徴**　ヒト・疾患関連のDBの割合が高い。ペプチド（タンパク質）の割合が高い。
米国と西欧地域は類似した傾向。

☞ **中国の特徴**　近年急速にDB数が増加しており、生物種も多様。
他の地域よりもヒト・疾患関連の割合が10ポイント以上高い。

バイオDBのトレンド：データ統合化と知識抽出



- ☞ **生物種の多様化** モデル生物から非モデル生物へ、病原生物や実用生物への展開が進む
- ☞ **データ種の多様化** データ種を超えた統合化やコネクティビティの整備が進む
- ☞ **ツールの増加・多機能化** 生物種やデータ種を超えた多様な情報からさまざまな知識抽出が進む

バイオDB俯瞰調査　今後の課題

【調査】

✓さらなる網羅性の担保。
✓初期調査におけるカテゴライズ、キュレーションの甘さの改善。
✓バイオDBの開発報告論文調査の徹底（WoS解析に調査漏れあり）。
✓調査の継続。

【分析】

✓国内外の主要なバイオDBの開発経緯の把握。
✓バイオDBの国内外のトレンドのさらなる詳細分析。
✓萌芽的バイオDBのトレンド把握。
✓わが国のバイオDB基盤の強み・弱みの把握。
✓バイオDB政策への反映。