

採択課題：データベース統合に関わる基盤技術開発

データベース統合の実現に向けて



基盤技術開発 (II)

岡本 忍 so@dbcls.rois.ac.jp

情報・システム研究機構
ライフサイエンス統合データベースセンター





データベースの統合はなぜ必要か

データ爆発

超高速ゲノムシーケンサー：10,000倍

ゲノムプロジェクト：10,000以上

計算機の処理能力：ムーアの法則

イメージデータの増加

データベース爆発

世界：10,000以上

日本：1,000以上

解析ツール：2,000以上

知識爆発

論文出版：2,000万報

論文オープンアクセス化



データ / データベース統合の重要性

大規模プロジェクトの成果物としてのデータベース

多額の研究費がデータとして産出される

大規模データ（ビッグデータ）の出現

データ駆動型サイエンスの台頭

仮説駆動型 \leftrightarrow データ駆動型

データの囲い込み \rightarrow データの分断、断片化

オープンイノベーションには知識の共有・統合化が必要

ブラウズ、検索や一部のデータの再利用では不十分

データ生産者以外の不特定多数のイノベーター

再利用、転用が自由にできる必要がある



データ/データベース統合を巡る諸問題

- データ量の爆発
- データの囲い込み
- データ再利用性が低い
- データベースの乱立
- データベースが維持されない



統合にむけたDBCCLSの取り組み

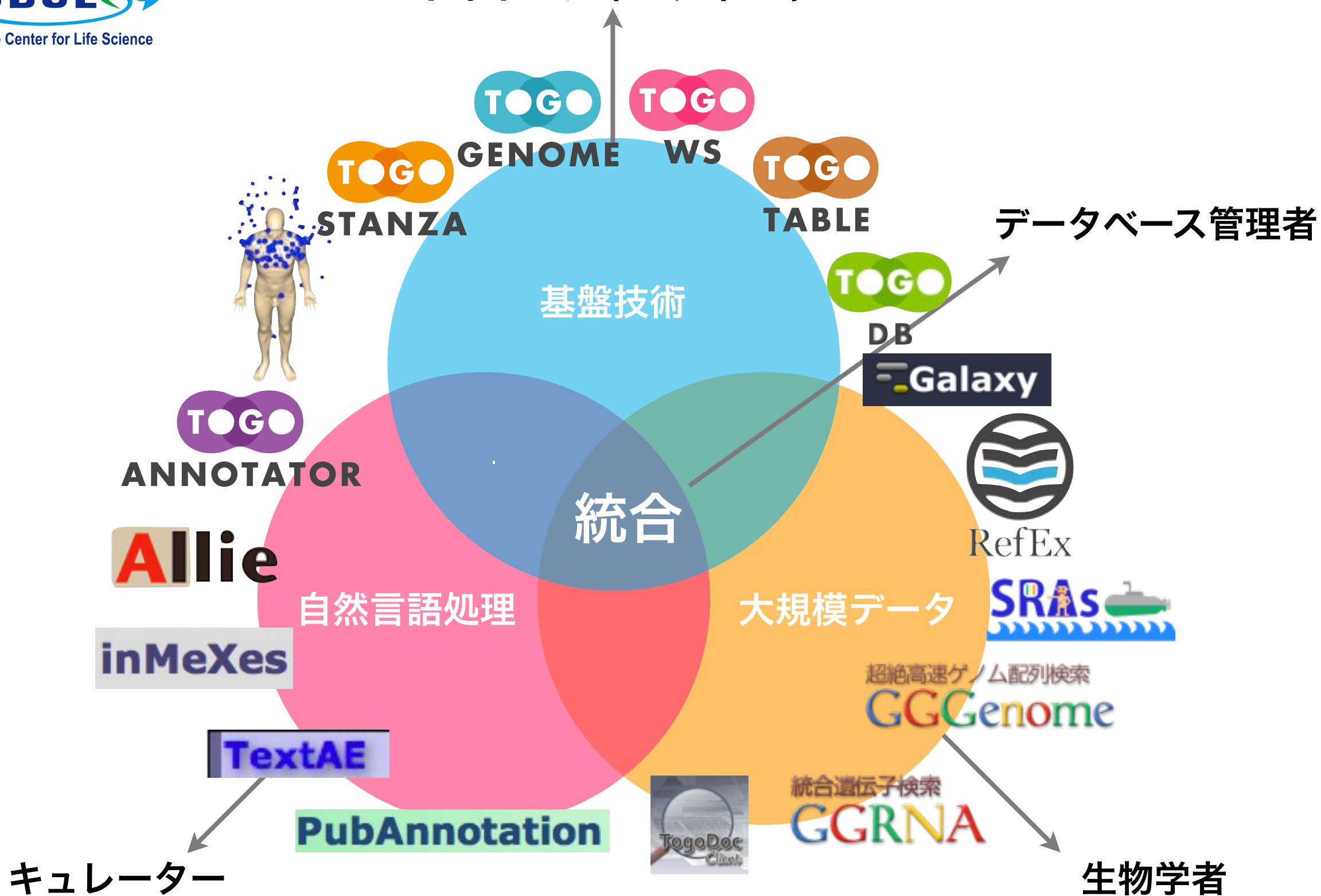
- 専門領域ごとのニーズの掘り起こし
- オープンなデータの収集と整理
- 語彙、辞書、データ、技術の標準化
- 可視化、検索手法の開発
- 労働集約型作業の自動化手法の開発
- 講習会、チュートリアル動画、レビュー



統合のためにDBCCLSが提供するサービス



バイオインフォマティクス





統合のためにDBCCLSが提供するサービス

大規模データ

サービス名	URL	サービスタイプ
RefEX	http://refex.dbcls.jp/	整理、検索、閲覧、解析
DBCLS SRA	http://sra.dbcls.jp/	整理、検索
GGRNA	http://ggrna.dbcls.jp/	検索
GGGenome	http://gggenome.dbcls.jp/	検索
TogoDoc/Client	http://tdc.cb.k.u-tokyo.ac.jp/	集積、整理、自動化

基盤技術

サービス名	URL	サービスタイプ
TogoDB	http://semantic.togodb.dbcls.jp/	標準化、集積、検索、整理、共有
TogoWS	http://togows.dbcls.jp/	標準化、共有
TogoTable	http://togotable.dbcls.jp/	標準化、集積、検索、整理、共有
TogoGenome/TogoStanza	http://togogenome.org/	標準化、集積、検索、整理、共有、可視化
DBCLS Galaxy	http://galaxy.dbcls.jp/	自動化、共有
BodyParts 3D	http://lifesciencedb.jp/bp3d/	標準化、整理、共有、可視化

自然言語処理

サービス名	URL	サービスタイプ
PubAnnotation	http://pubannotation.dbcls.jp/	標準化、集積、整理、共有
TextAE		標準化、整理、自動化、共有
Allie	http://allie.dbcls.jp/ja	検索
inMeXes	http://docman.dbcls.jp/im/index.html.ja	検索

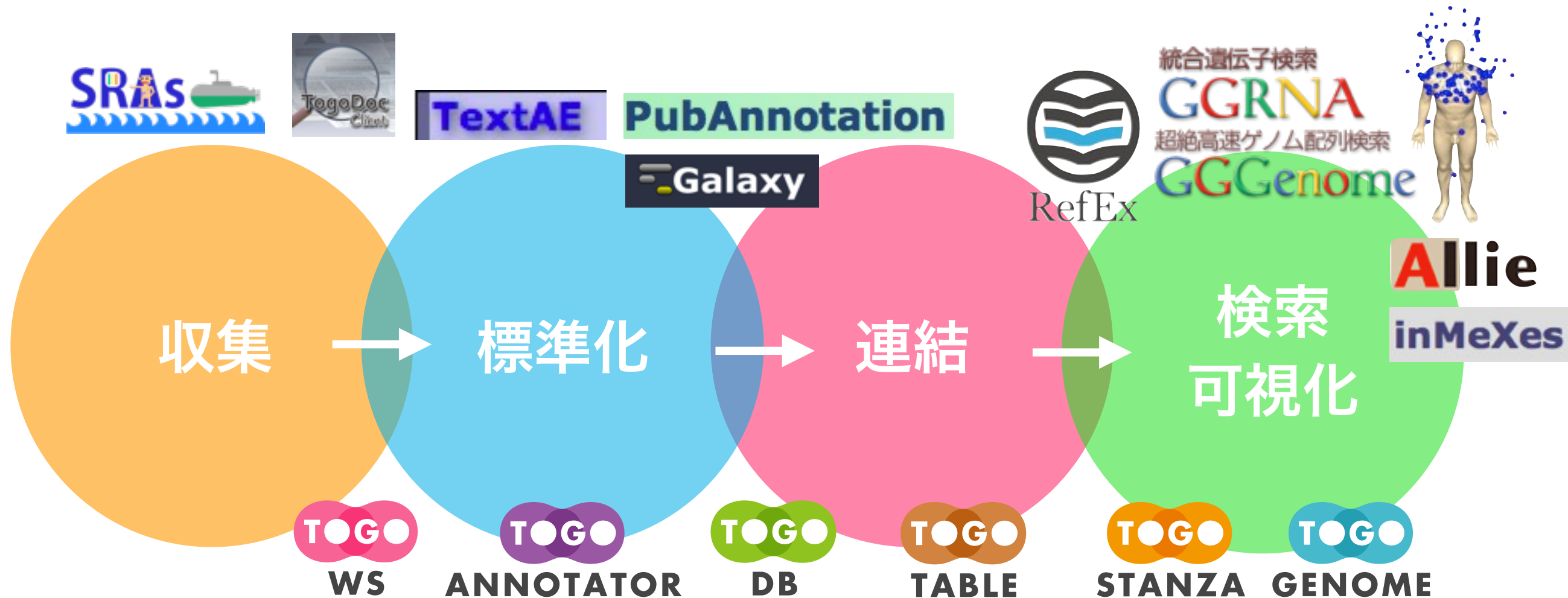
教育

サービス名	URL	サービスタイプ
TogoTV	http://togotv.dbcls.jp/ja/	教育、動画
AJACS	http://motdb.dbcls.jp/	教育、講習会、コンテンツ
First Author, Leading Author	http://first.lifesciencedb.jp/	教育、コンテンツ、オープンアクセス



統合のためにDBCCLSが提供するサービス

統合（トーゴー）



生命知識の統合と発見



基盤となるゲノム情報を集積しDB化



特徴

- 多種多様なデータをユニークなID (URI) とゲノム座標に基づき集積
- 世界標準なオントロジーやデータ開発
- さまざまな検索機能でデータを絞り込む
 - ファセット検索、配列検索、ID検索
- 外部から再利用可能な可視化パーツ
 - TogoStanza
- 閲覧したい視点ごとにTogoStanzaを組み合わせてレポートページを作成可能



TogoGenome

http://togogenome.org/mappings/convert?utf8=✓&from_database=togogenome&identifiers=103690%3Aall0004%0D%0A103690%3Aalr0022%0D%0A103690%3Aall0374

TOGO GENOME

Facets | Sequence | ID Mapping | Hide

TogoGenome

UniProt

RefSeq

Pfam

NCBI Gene

NCBI Protein

103690:all0004
103690:alr0022
103690:all0374

Identifiers

TogoGenome

UniProt

RefSeq

Pfam

NCBI Gene

NCBI Protein

From Database To Database

[Map](#)

From	TogoGenome URL	To	Pfam URL
103690:all0374	http://togogenome.org/gene/103690:all0374	PF00155	http://pfam.sanger.ac.uk/family/PF00155
103690:all0004	http://togogenome.org/gene/103690:all0004	PF00231	http://pfam.sanger.ac.uk/family/PF00231
103690:all0374	http://togogenome.org/gene/103690:all0374	PF00266	http://pfam.sanger.ac.uk/family/PF00266
103690:all0374	http://togogenome.org/gene/103690:all0374	PF01041	http://pfam.sanger.ac.uk/family/PF01041
103690:all0374	http://togogenome.org/gene/103690:all0374	PF01053	http://pfam.sanger.ac.uk/family/PF01053
103690:alr0022	http://togogenome.org/gene/103690:alr0022	PF00502	http://pfam.sanger.ac.uk/family/PF00502
103690:all0004	http://togogenome.org/gene/103690:all0004	PF08967	http://pfam.sanger.ac.uk/family/PF08967

生育環境
6 Stanza

生物種
12 Stanza

遺伝子
13 Stanza

Environment attributes

Environment: MEO_000029

Inhabitants statistics

GOLD	26
JCM	97
NBIC	77

Inhabitants

Original source	Organism name	Taxonomy ID	Selection
GD0272			
GD0304	Thermotoga gammatolerans EJ		
GD1508	Calditerrivibrio modestus DL	522118	
GD0354	Subdoligranulum turgidum DSM 1848	543362	
GD4433			
GD1747	Thermotoga sibirica DSM 14796	747361	
GD0231	Methanothermobacter thermautotrophicus DSM 3952	509580	
GD2331	Alicyclobacillus acidocaldarius LAJ3	540803	
GD1509	Calditerrivibrio modestus DSM 1847	543360	
GD1540	Calditerrivibrio modestus DSM 1848	543362	
GD1884			
GT0578	Sulfolobus	2281	

Geographical map

Taxonomic composition

Organism: 1351

Organism name

Scientific name

Synonym

- Dimerotoga faecalis
- Dimerotoga profectans
- Dimerotoga
- Dimerotoga maffei
- Dimerotoga sibirica
- Streptotoga faecalis
- Streptotoga profectans
- Streptotoga maffei

Authority

- "Dimerotoga" Throckmoller 1952
- "Dimerotoga maffei" Throckmoller 1952
- "Streptotoga profectans" Throckmoller 1952
- "Streptotoga faecalis" Throckmoller 1952
- Streptotoga faecalis (Throckmoller and Throckmoller 1952) (Throckmoller and Throckmoller 1952)
- "Dimerotoga profectans" Throckmoller and Throckmoller 1952
- "Dimerotoga maffei" Throckmoller and Throckmoller 1952
- "Dimerotoga sibirica" Throckmoller and Throckmoller 1952
- "Streptotoga faecalis" Throckmoller and Throckmoller 1952
- "Streptotoga profectans" Throckmoller and Throckmoller 1952
- "Dimerotoga sibirica" Throckmoller and Throckmoller 1952

Genome information

No data

Ortholog profile

Taxonomic information

Rank	Lineage
Super-Kingdom	cellular organisms
Kingdom	Bacteria
Phylum	Firmicutes
Class	Bacilli
Order	Lactobacillales
Family	Dimerotogaceae
Genus	Dimerotoga
Species	Dimerotoga faecalis

Gene: 1111708:slr1311

Protein names

Protein names	Recommended Name	EC number	Phosphoprotein [DB:PF00134]
Phosphoprotein 2			0.3.1.24

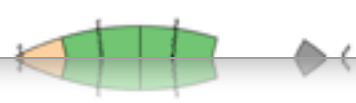
Gene names

Gene names	Name	Synonym	Ordered Locus Name
Phosphoprotein 2	phoA2		phoA2
Phosphoprotein 1	phoA1		phoA1

Genomic context

Gene attributes

Attribute	Value
Gene symbol	phoA2
Local tag	phoA1
Gene type	gene
Organism	Streptococcus pneumoniae
Position	NC_009611.3
Strand	7228..8211
Strand	Positive strand
Length	1083





生育環境レポートページ



TogoGenome

http://togogenome.org/environment/MEO_0000038

TOGO GENOME

Environment attributes Inhabitants statistics Inhabitants Geographical map Taxonomic composition Environmental ontology (MEO)

Environment: MEO_0000038

Environment attributes

Environment	fresh water
Description	Fresh water is generally characterized by having low concentrations of dissolved salts and other total dissolved solids.
MEO	MEO_0000038
Exact synonyms	

Inhabitants statistics

GOLD	303
JCM	25
NBRC	34

Inhabitants

Original source	Organism name	Taxonomy ID	Isolation	Environments
Gc01999	Thermoproteus tenax Kra 1	768679		fresh water
Gi12071	Hydrogenobaculum acidophilum	34092		fresh water
Gc00944	Thermotoga neapolitana DSM 4359	309803		fresh water



生育環境レポートページ



TogoGenome

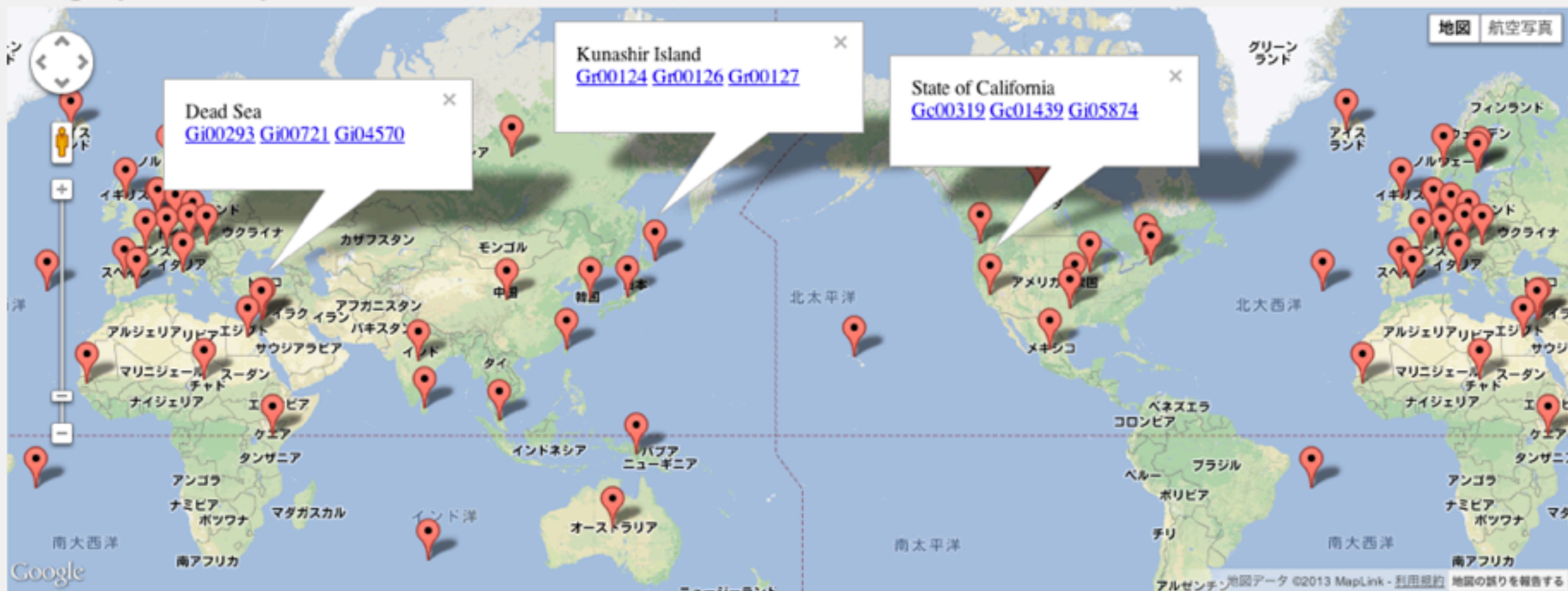
http://togogenome.org/environment/MEO_0000038#environment_geographical_map

tfac operon Import to Mendeley cga1 bookmarklet DBCLSgs DBCLSwiki Google Bookmarklet AI NI Wikipedia BI UI Diigo! CBSrc MyHB Gmail SGA:Trac Res



Environment attributes Inhabitants statistics Inhabitants Geographical map Taxonomic composition Environmental ontology (MEO)

Geographical map

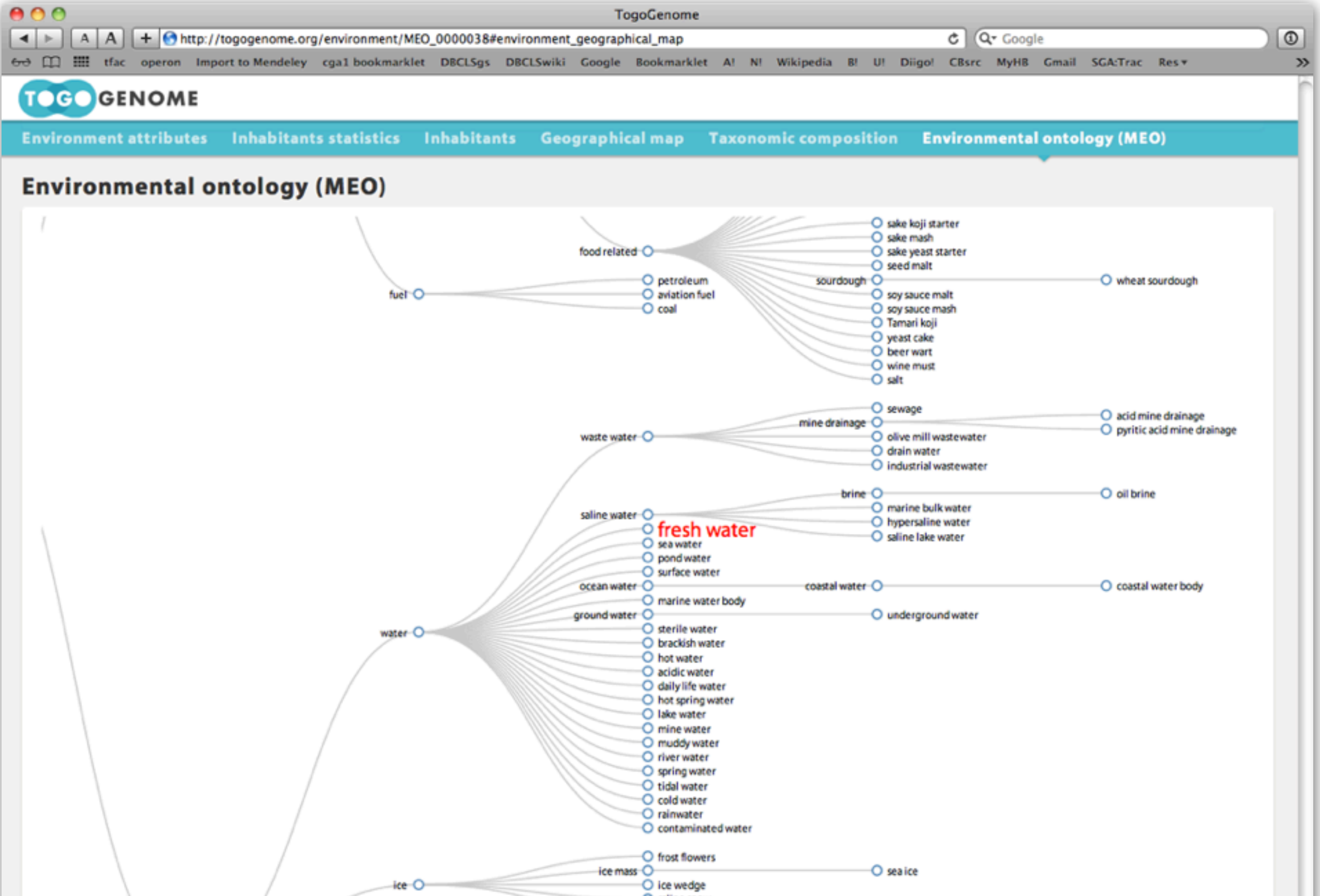


Taxonomic composition





生育環境レポートページ





生物種レポートページ

TogoGenome

http://togogenome.org/organism/103690#organism_phenotype

TOGO GENOME

Organism name Genome information Ortholog profile Taxonomic information Culture collections Medium information Phenotype

Organism: 103690

Organism name

Scientific name	<ul style="list-style-type: none">Nostoc sp. PCC 7120
Synonym	<ul style="list-style-type: none">Anabaena variabilis UTCC 387Anabaena sp. DCC D0672Nostoc sp. ATCC 72893Anabaena sp. (ATCC 27893)Nostoc sp. AKM24Anabaena sp. SAG 25.82Nostoc sp. Ind43Anabaena sp. IRRl 'Ab 47 XX'Nostoc muscorum ISUAnabaena sp. (PCC 7120)Nostoc sp. ATCC 27347Anabaena sp. UTEX 'B 2576'
Genbank synonym	<ul style="list-style-type: none">Anabaena sp. PCC 7120
Misspelling	<ul style="list-style-type: none">Anabaena sp. AKM8Nostoc PCC7120Anabaena PCC7120Nostoc sp. strain PCC 7120Anabaena sp. (strain PCC 7120)Anabaena sp. UTEX 2576Anabaena 7120Anabaena sp. PCC7120



生物種レポートページ

TogoGenome

http://togogenome.org/organism/1351#organism_medium_information

TOGO GENOME

ation Culture collections Medium information Phenotype information Genomic plot Pathogen information Organism cross references

Medium information

Medium ID	NBRC_M310
Medium name	MRS Medium
Medium type	Undefined medium
Defined ingredient	Polysorbate 80, Potassium phosphate dibasic, Magnesium sulfate heptahydrate, Sodium acetate, Distilled water, Manganese sulfate n-hydrate, Diammonium hydrogen citrate, Glucose
Undefined ingredient	Peptone, Yeast extract, Meat extract
Support medium	Agar
Water	Distilled water

Phenotype information

No data

Genomic plot



生物種レポートページ

TogoGenome

http://togogenome.org/organism/103690#organism_phenotype

TOGO GENOME

og profile Taxonomic information Culture collections Medium information Phenotype information Genomic plot Pathogen informatio

Phenotype information

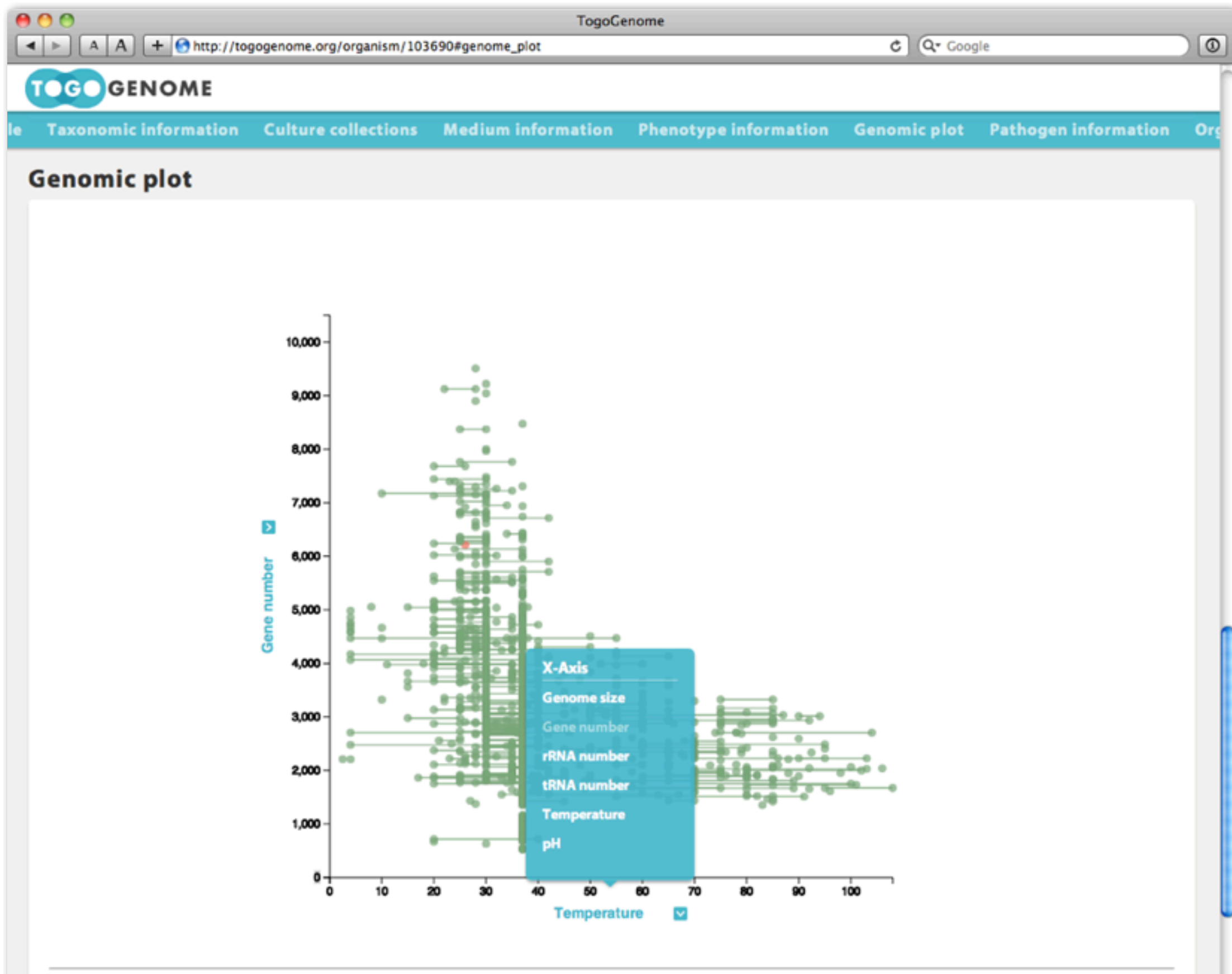
Cell shape	Filament
Oxygen requirement	Aerobe
Temperature range	Mesophile
Optimal growth pH	7.1
Optimal growth temperature	26°C
Motility	Motile

Genomic plot

10,000
9,000
8,000
7,000
0,000



生物種レポートページ





遺伝子レポートページ

TogoGenome

http://togogenome.org/gene/103690:all2699

TOGO GENOME

Protein names Genomic context Gene attributes Nucleotide sequence Protein attributes Protein sequence Protein general annotat

Gene: 103690:all2699

Protein names

Protein names	
Genes names	Ordered Locus Names all2699
Organism	Nostoc sp. (strain PCC 7120 / UTEX 2576)
Taxonomic identifier	103690
Taxonomic lineage	<ul style="list-style-type: none">Bacteriacellular organismsNostocNostocaceaeNostocalesCyanobacteriaNostoc sp. (strain PCC 7120 / UTEX 2576)

Genomic context

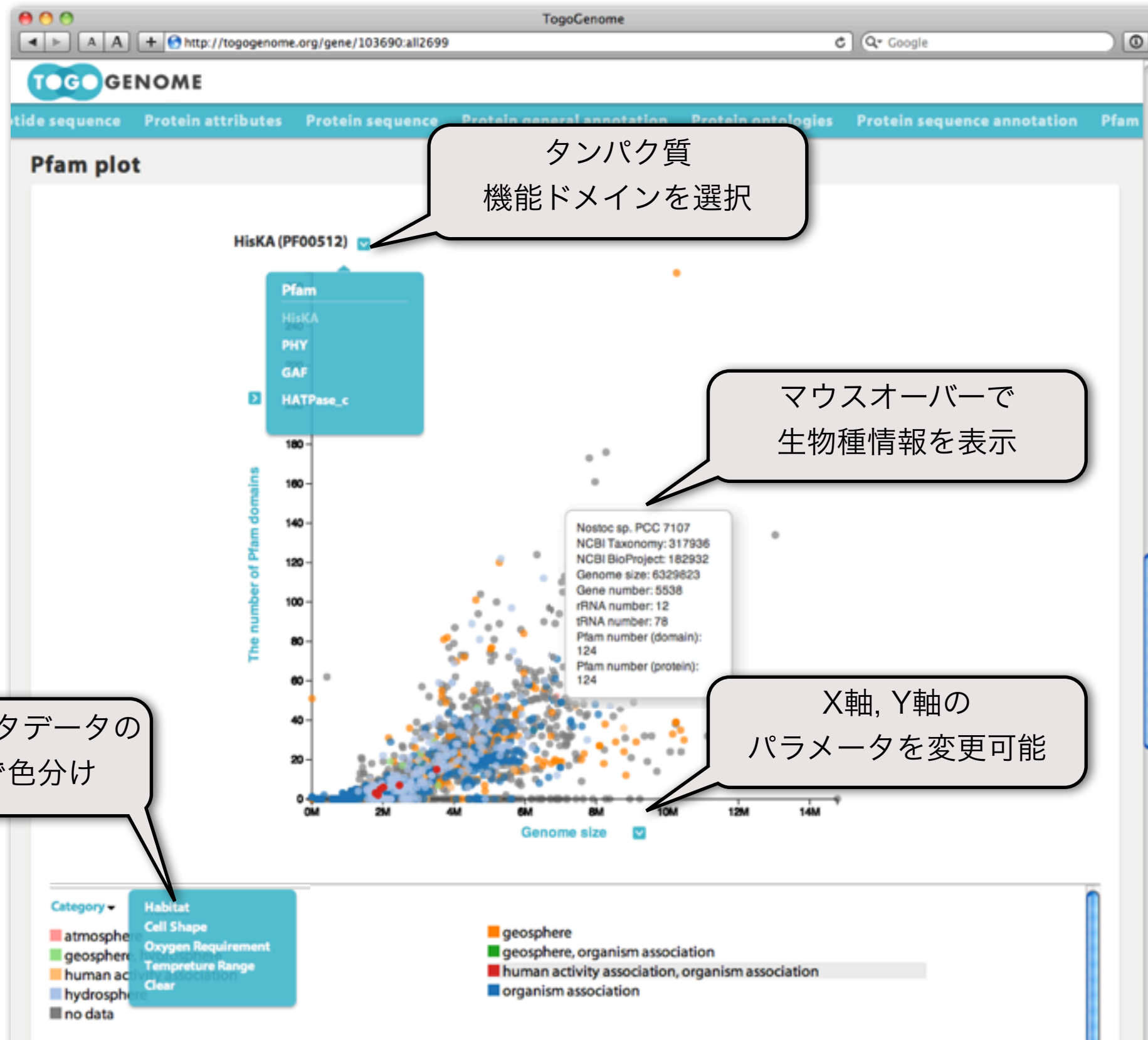
3288210 protein coding gene 3291652

alr2698 all2699 all2700



遺伝子レポートページ

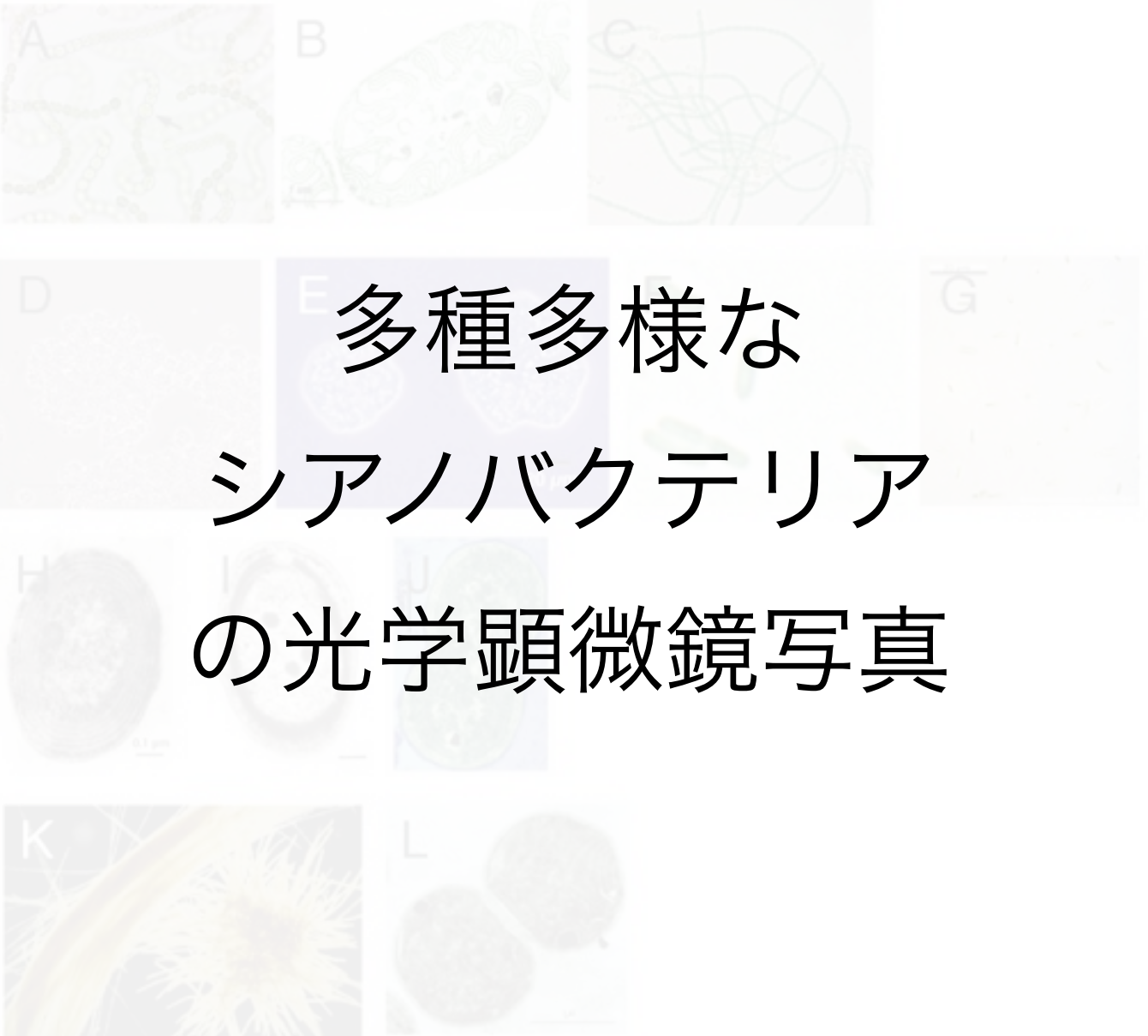
データを探索的に眺めることが出来る機能





ユースケース：比較ゲノム解析

光合成細菌（シアノバクテリア）



多種多様な
シアノバクテリア
の光学顕微鏡写真

- モデル生物
- 酸素発生型光合成
- **生物学的な特徴が多様**
 - 形態
 - 生育場所
 - 窒素固定
 - 細胞分化
 - 運動



ユースケース：比較ゲノム解析

The screenshot shows the TOGO Genome web interface. The browser address bar is <http://togogenome.org/>. The main navigation bar includes 'Facets', 'Sequence', and 'ID Mapping'. The 'Facets' section is active, showing filters for Environment, Taxonomy, GO: BiologicalProcess, GO: MolecularFunction, and GO: CellularComponent. The 'Environment' facet is expanded to 'fresh water', with a breadcrumb trail: 'All > hydrosphere > water'. A dropdown menu is open, listing various water types: acidic water, brackish water, cold water, contaminated water, daily life water, fresh water (highlighted), ground water, hot spring water, hot water, lake water, marine water body, mine water, muddy water, ocean water, pond water, rainwater, river water, saline water, sea water, spring water, sterile water, surface water, tidal water, and waste water. A 'Clear' button is visible next to the dropdown. Below the facets, it says 'Showing 1 to 25 of 12,853 entries (filtered from 1,292,481 total entries)'. A 'Download CSV' button is present. The main table displays search results with columns: Description, Gene, UniProt ID, Gene ontologies, Organism, and Environments. The first three rows are visible, all for the organism *Arthrospira platensis str. Paraca*.

Description	Gene	UniProt ID	Gene ontologies	Organism	Environments
NAD(P)H-quinone oxidoreductase subunit M	• APPUASWS_00373	K6DUG3	<ul style="list-style-type: none">transportoxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptorquinone bindingthylakoid membrane	Arthrospira platensis str. Paraca	<ul style="list-style-type: none">fresh waterfreshwater habitat
S05 ribosomal protein L27	• APPUASWS_00598	K6E668	<ul style="list-style-type: none">translationstructural constituent of ribosomeribosome	Arthrospira platensis str. Paraca	<ul style="list-style-type: none">fresh waterfreshwater habitat
Cytochrome b6-f complex subunit 8	• APPUASWS_00834	K6EQW5	<ul style="list-style-type: none">photosynthesiscytochrome complex assemblyelectron transport chainelectron transporter, transferring electrons within	Arthrospira platensis str. Paraca	<ul style="list-style-type: none">fresh waterfreshwater habitat



ユースケース：比較ゲノム解析

TogoGenome

http://togogenome.org/mappings/convert?utf8=✓&from_database=togogenome&identifiers=103690%3Aall0004%0D%0A103690%3Aalr0022%0D%0A103690%3Aall0374

TOGO GENOME

Facets | Sequence | ID Mapping | Hide

From Database

- TogoGenome
- UniProt
- RefSeq
- Pfam
- NCBI Gene
- NCBI Protein

Identifiers

103690:all0004
103690:alr0022
103690:all0374

To Database

- TogoGenome
- UniProt
- RefSeq
- Pfam
- NCBI Gene
- NCBI Protein

Map

From	TogoGenome URL	To	Pfam URL
103690:all0374	http://togogenome.org/gene/103690:all0374	PF00155	http://pfam.sanger.ac.uk/family/PF00155
103690:all0004	http://togogenome.org/gene/103690:all0004	PF00231	http://pfam.sanger.ac.uk/family/PF00231
103690:all0374	http://togogenome.org/gene/103690:all0374	PF00266	http://pfam.sanger.ac.uk/family/PF00266
103690:all0374	http://togogenome.org/gene/103690:all0374	PF01041	http://pfam.sanger.ac.uk/family/PF01041
103690:all0374	http://togogenome.org/gene/103690:all0374	PF01053	http://pfam.sanger.ac.uk/family/PF01053
103690:alr0022	http://togogenome.org/gene/103690:alr0022	PF00502	http://pfam.sanger.ac.uk/family/PF00502
103690:all0004	http://togogenome.org/gene/103690:all0004	PF08967	http://pfam.sanger.ac.uk/family/PF08967



Address: 2-11-16 Yayoi, Bunkyo-ku, Tokyo

Phone: +81 (3) 5841 6754



ユースケース：比較ゲノム解析

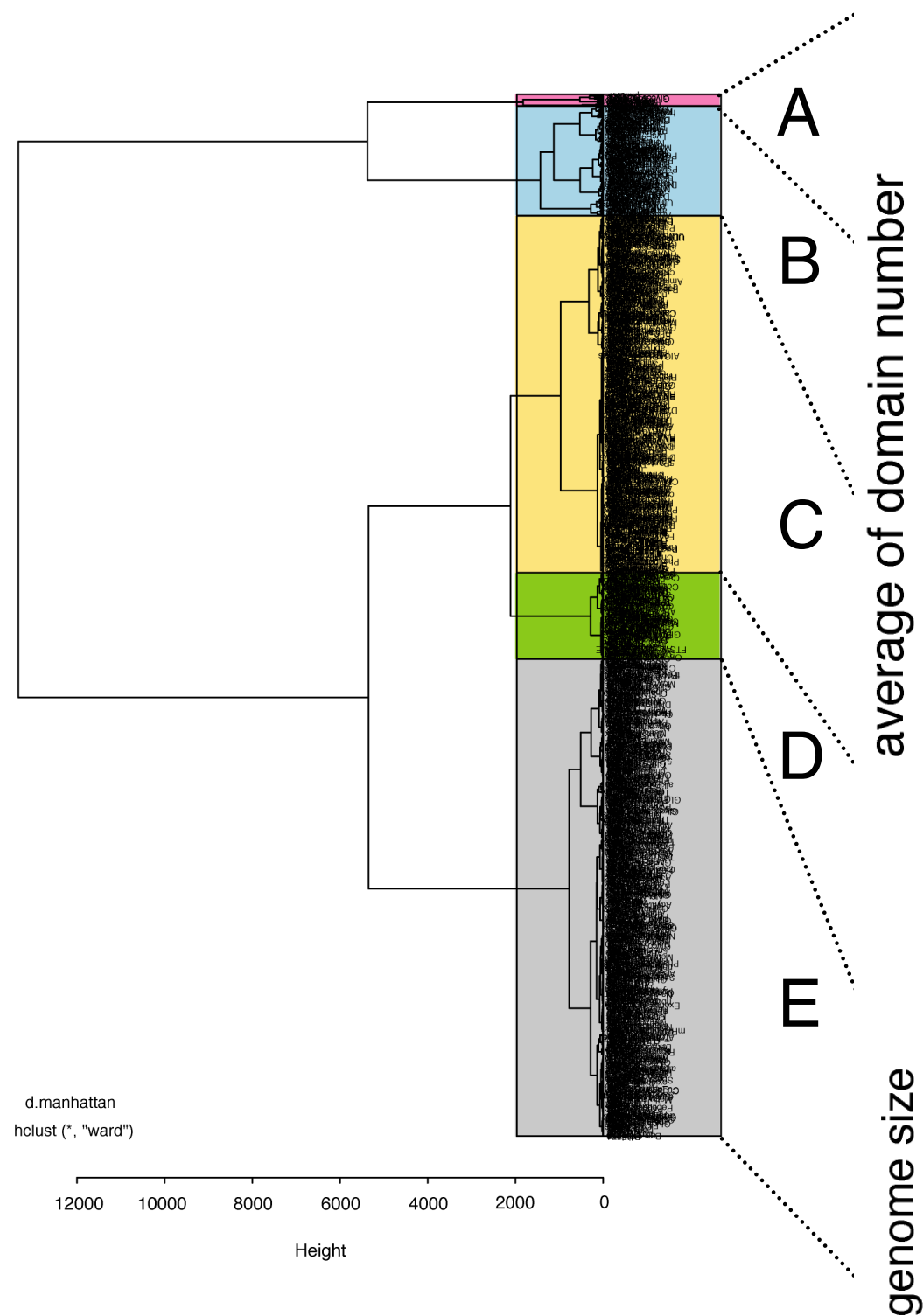
データセットの取得（選択された遺伝子と機能ドメイン）

生物種	生育場所	遺伝子	Pfam
<i>Nostoc punctiforme</i> ATCC291330	淡水	7323	5997
<i>Anabaena variabilis</i>	淡水	5754	5083
<i>Anabaena</i> sp. PCC 7120	淡水	6131	5052
<i>Gloeobacter violaeus</i> PCC 7421	淡水	4430	3833
<i>Thermosynechococcus elongatus</i> BP-1	淡水	2475	2452
<i>Synechocystis</i> sp. PCC 6803	淡水	3264	3142
<i>Synechococcus</i> sp. PCC 6301	淡水	2525	2475
<i>Synechococcus elongatus</i> PCC 7942	淡水	2653	2516
<i>Prochlorococcus marinus</i> MIT 9313	海水	2265	1933
<i>Prochlorococcus marinus</i> MED 4	海水	1712	1560
<i>Prochlorococcus marinus</i> SS 120	海水	1882	1609
<i>Synechococcus</i> sp. WH 8102	海水	2517	2066
<i>Crocospaera watsonii</i> WH 8501	海水	6751	3759
<i>Trichodesmium erythraeum</i>	海水	7713	3462

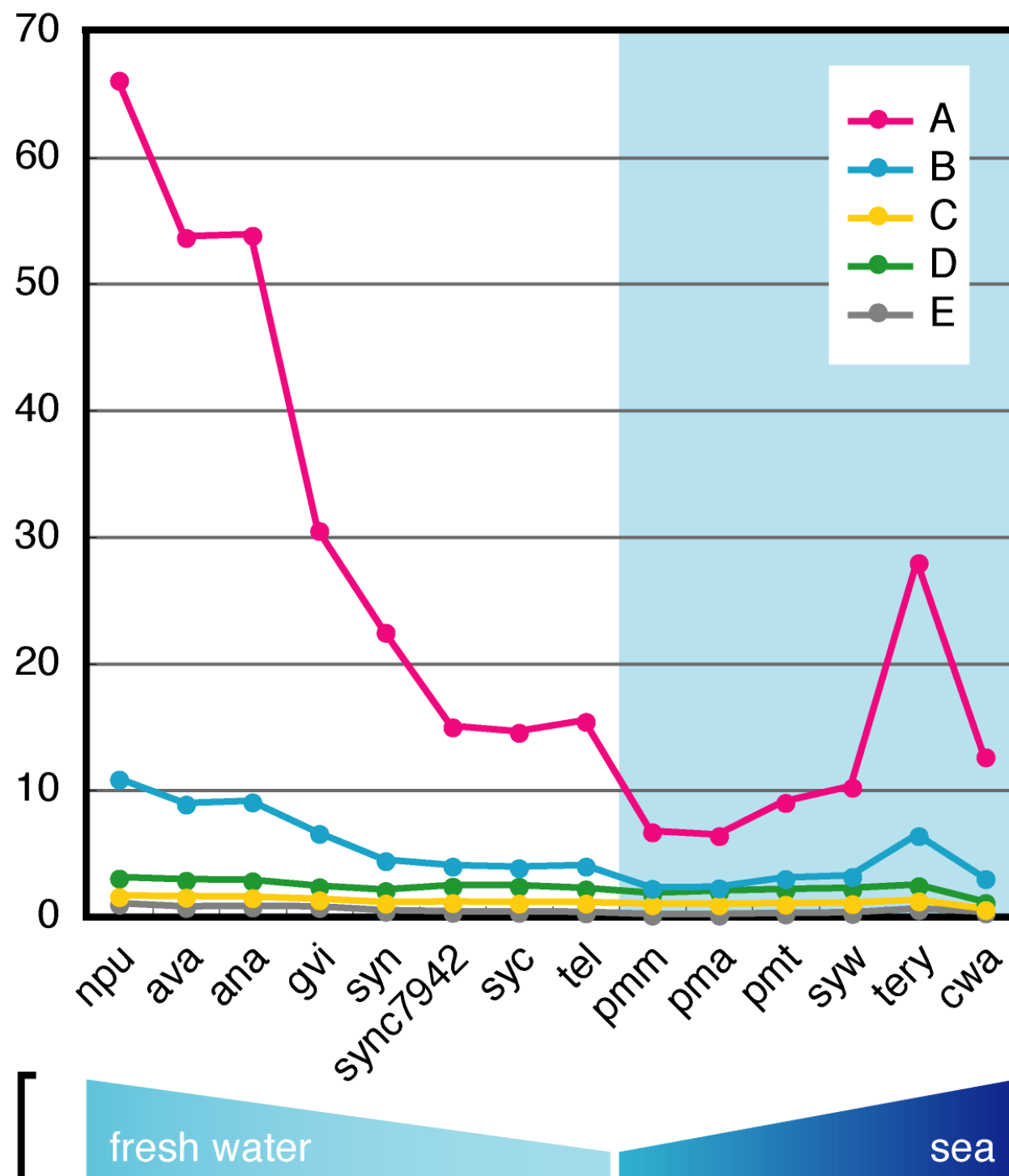


ユースケース：比較ゲノム解析

結果



機能ドメインの生物種別プロファイル

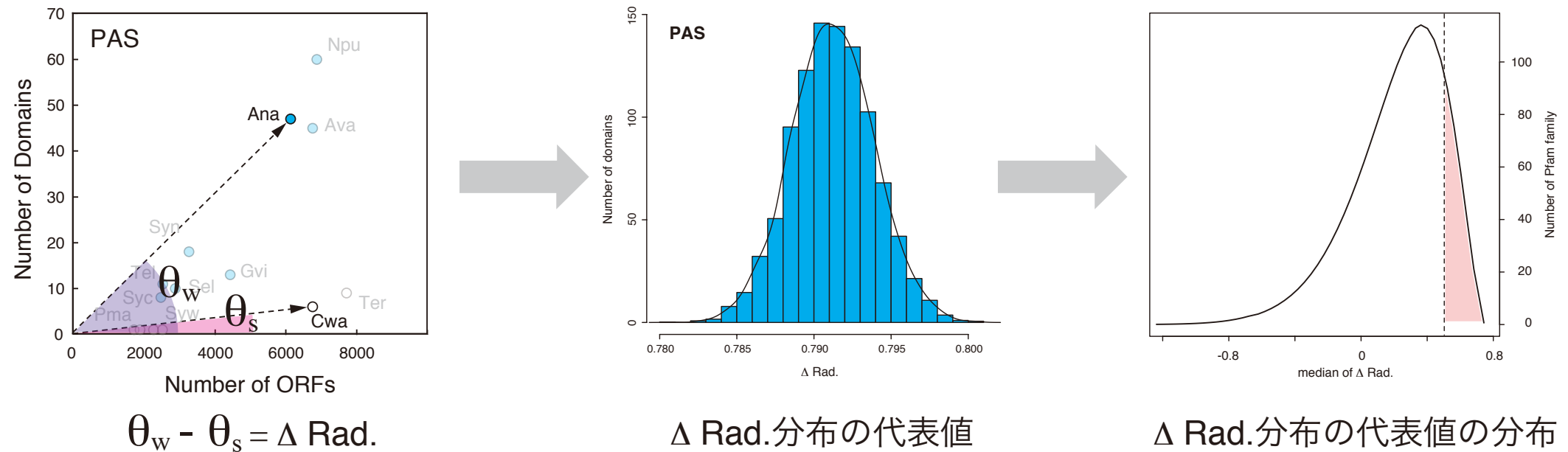




ユースケース：比較ゲノム解析

解析手法と結果

各機能ドメインごとに海産性と淡水性のドメイン重複の差を判定



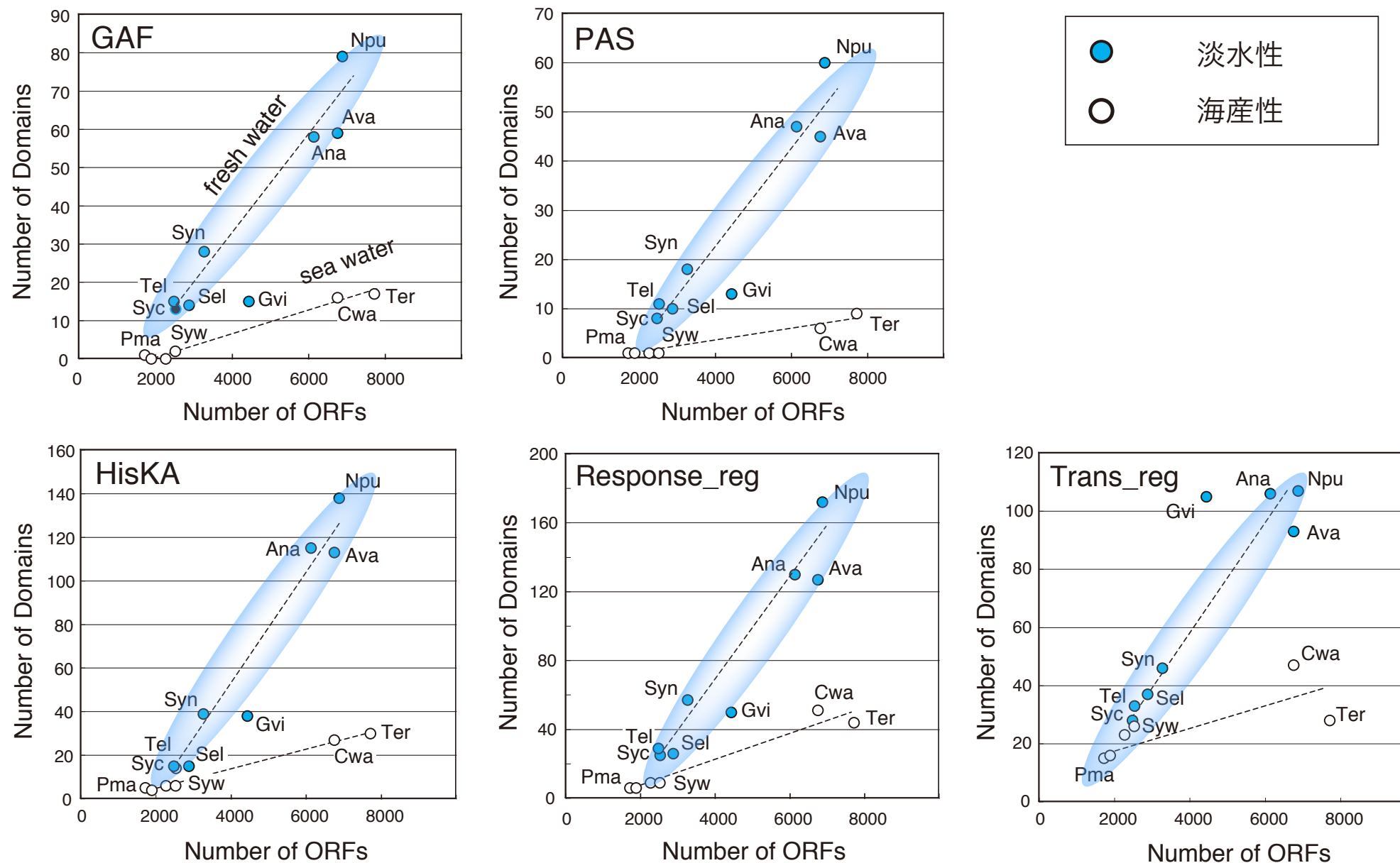
海産性で顕著な重複が起きている機能ドメイン

Functional class	Number of Pfam domains
Transporter, Outer membrane	8
Signal transduction	10
Metabolism	13
Others	3



ユースケース：比較ゲノム解析

結果 リン酸化による信号伝達に関連するタンパク質機能ドメイン

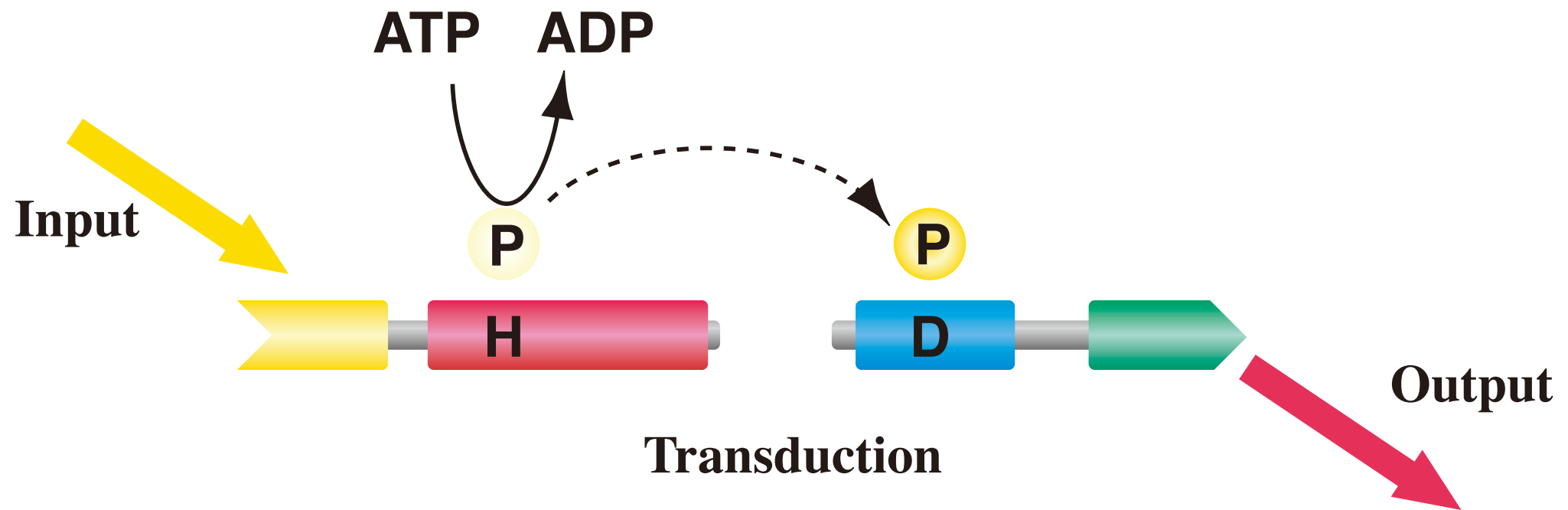


さまざまな生物学的特徴により生物種やタンパク質機能ドメインを絞り込むことで、表現型とタンパク質機能の関連を見いだすことができる



ユースケース：比較ゲノム解析

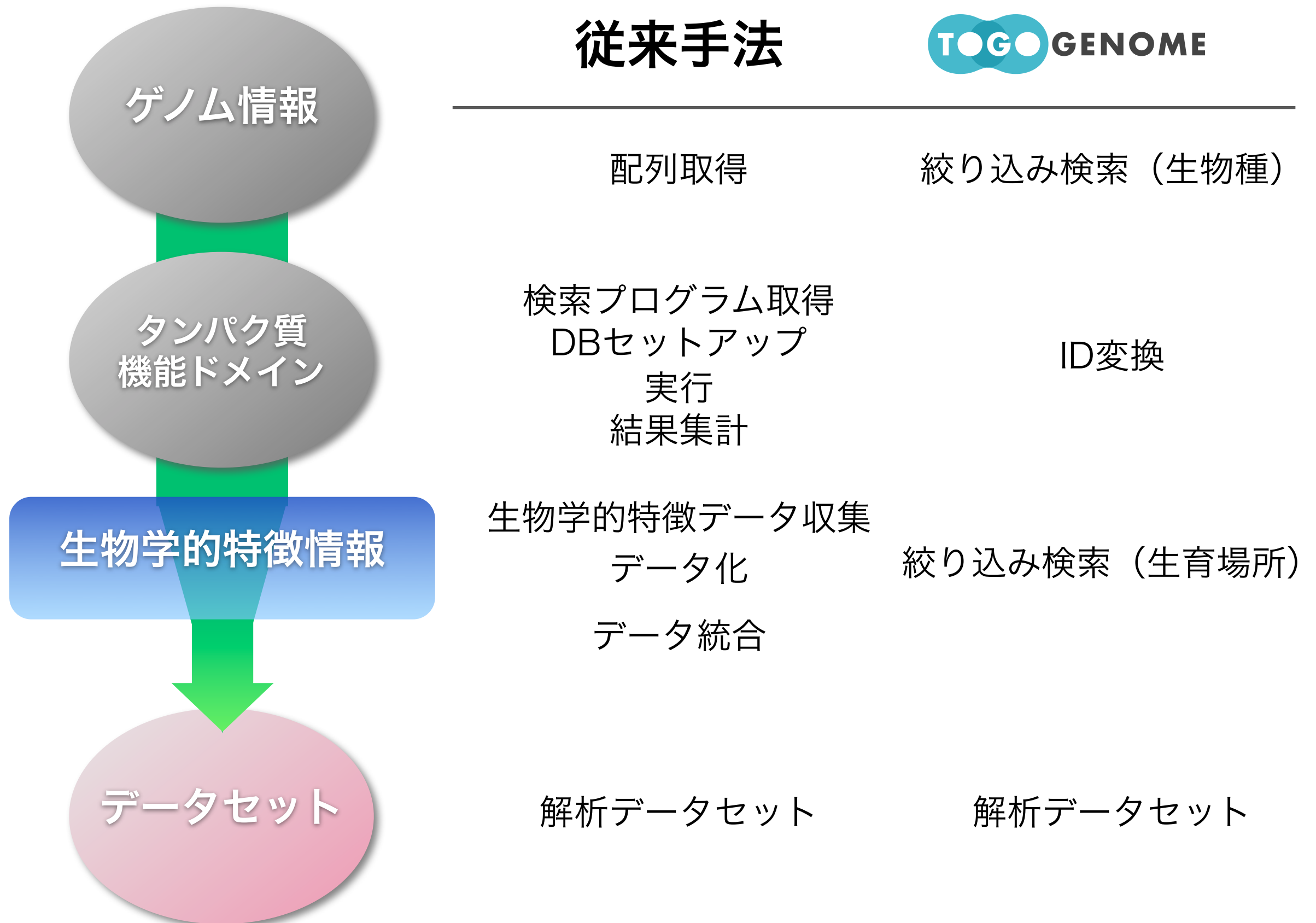
結果 リン酸化による信号伝達に関連するタンパク質機能ドメイン



sensor	transmitter	receiver	transcriptional regulation
GAF PAS	HisKA	Response_reg	GerE HTH_8 Trans_reg_C etc...



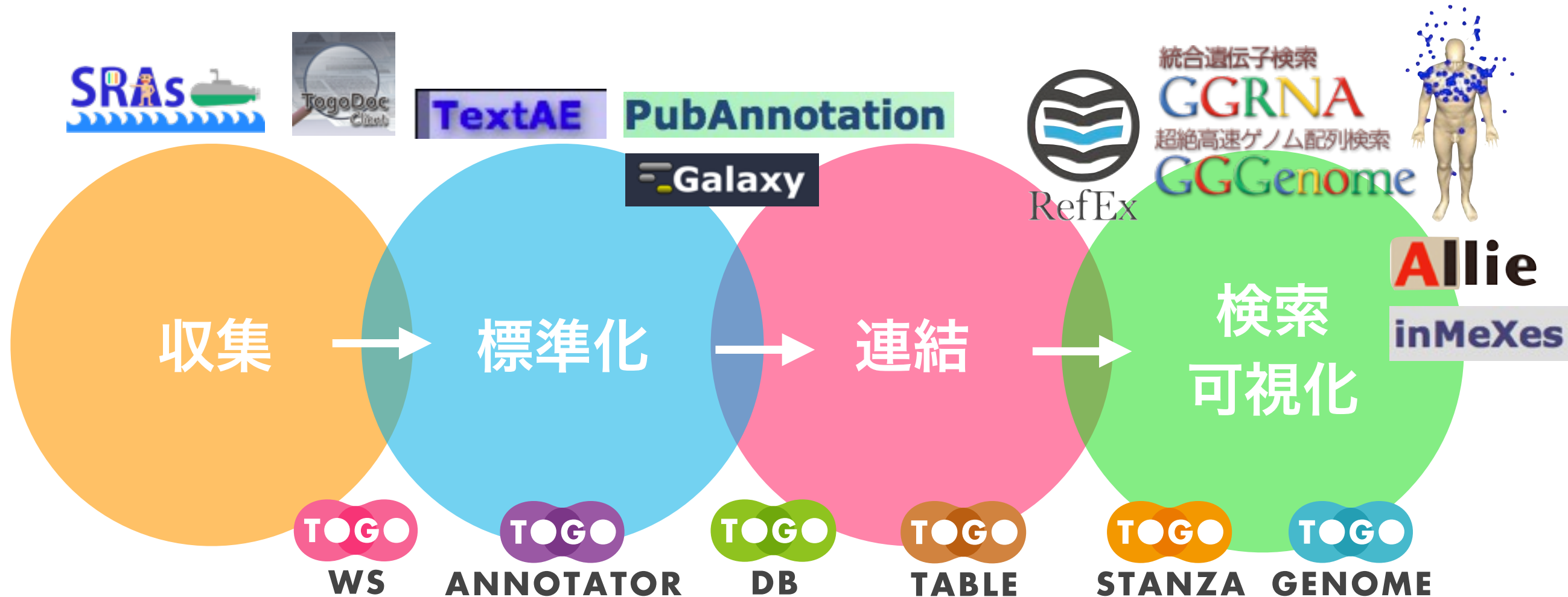
ユースケース：比較ゲノム解析





統合のためにDBCCLSが提供するサービス

統合（トーゴー）



生命知識の統合と発見