

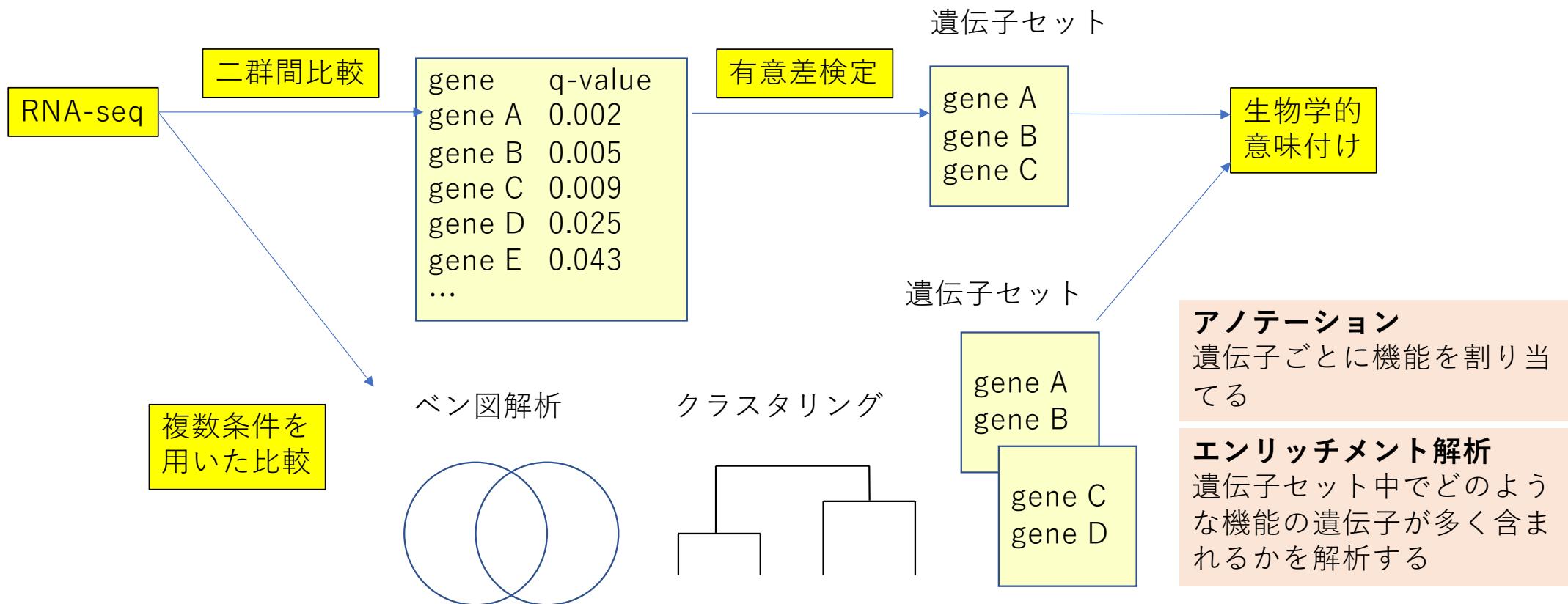
機能アノテーションと GO解析

基礎生物学研究所

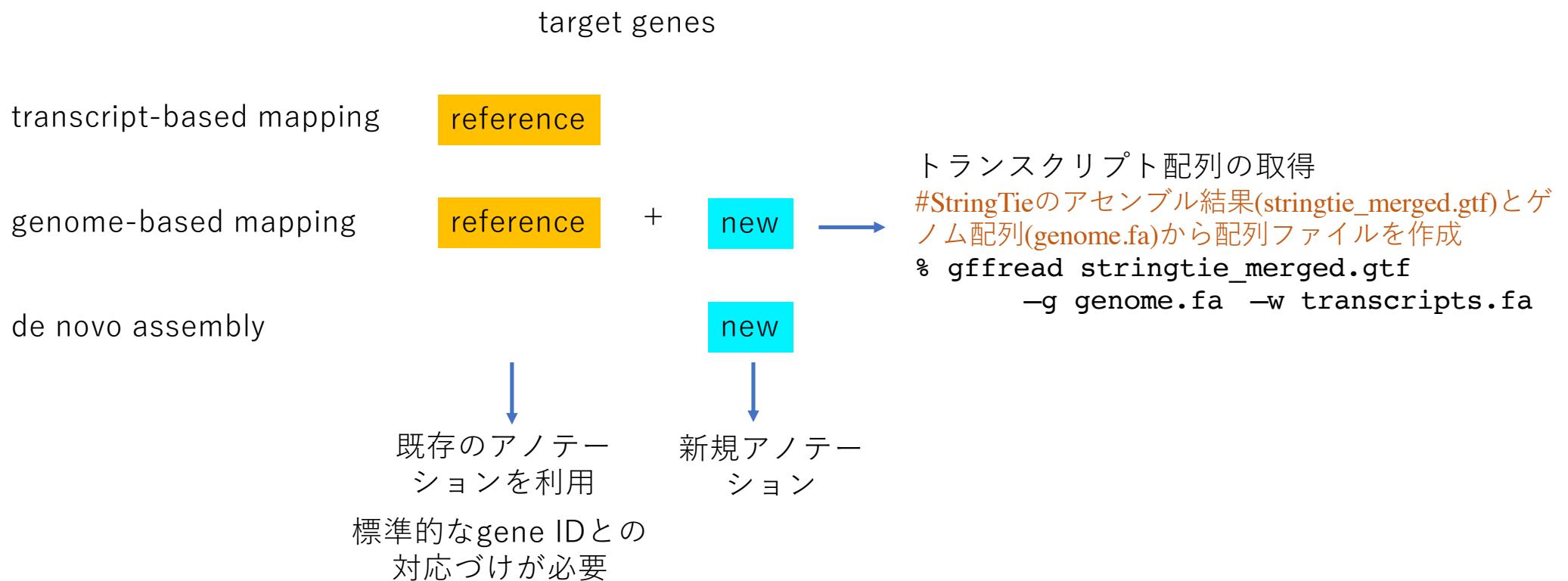
情報管理解析室

内山 郁夫

RNA-seq解析結果の解釈



遺伝子アノテーション 基本戦略



遺伝子アノテーション

- 遺伝子構造の予測

- 通常の遺伝子構造予測では、イントロンーエクソン構造の予測が必要だが、RNA-seq解析では、すでにmRNA配列が得られているので、これは不要。ただし、配列中のタンパク質をコードする領域(ORF)を予測する必要はある。
- blastxなどのツールを用いると、ORF予測をせずに、全読枠を翻訳してホモロジー検索を実行することも可能。ただし時間がかかる。

- 遺伝子機能の予測

- ホモロジー検索、もしくはモチーフ検索を行い、ヒットした類似配列もしくはモチーフの機能に基づいて機能を推定する。

TransDecoder

- 転写配列中のコード領域を予測するツール。もともとTrinityに付属のツールとして開発されたが、現在は独立のツールとして公開されている。

TransDecoderの実行

クエリ配列： seqfile.fa

- 長いORFを抽出する(≥ 100 aa)

% TransDecoder.LongOrfs -t seqfile.fa

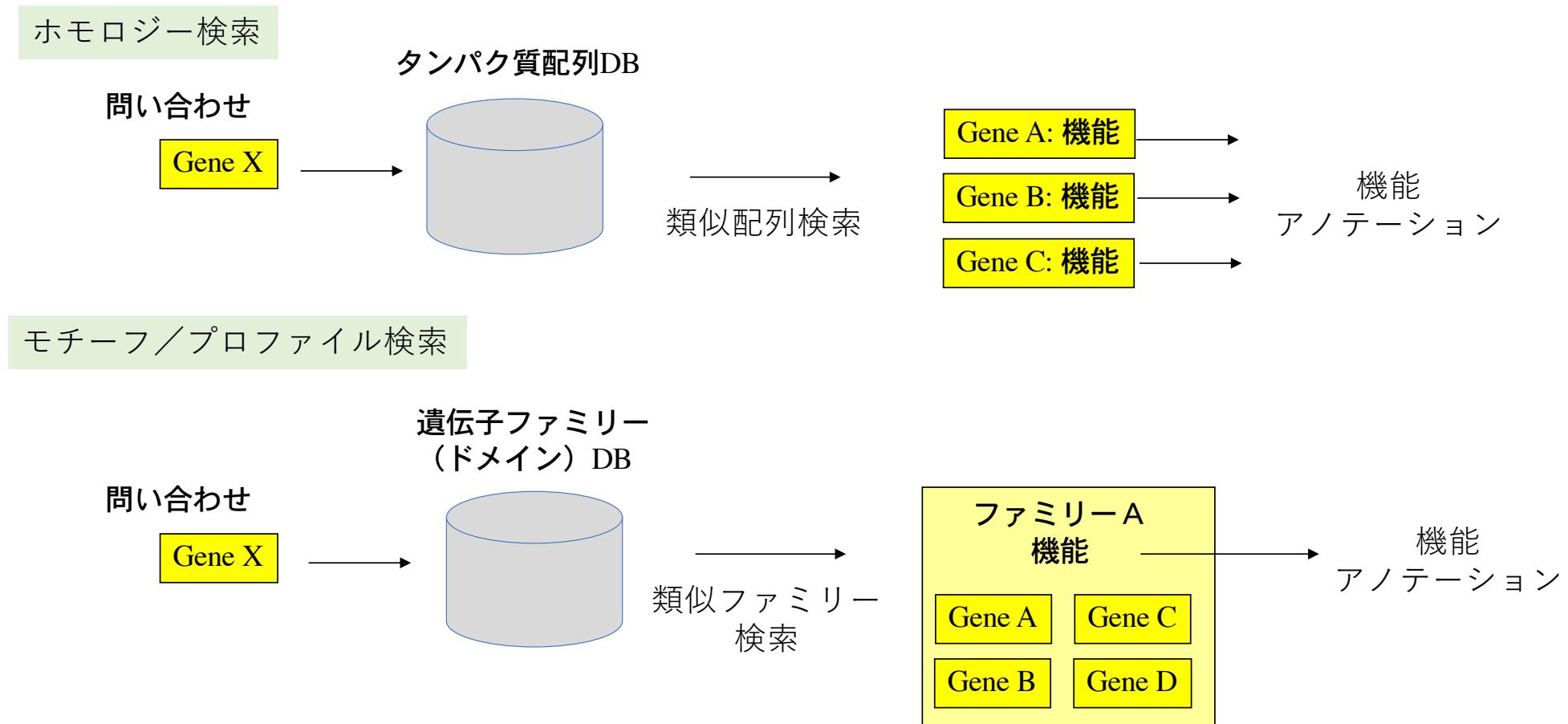
- 長さ上位500のORFを使って、コード領域の確率モデルパラメータを推定し、それを用いてコード領域を予測する

% TransDecoder.Predicts -t seqfile.fa

出力ファイル： seqfile.fa.transdecoder.???

pep: アミノ酸配列、 cds: 塩基配列、 gff3: GFF形式の遺伝子座標

ホモロジーに基づく機能推定



ホモロジー検索とアノテーション

ヘモグロビン
β鎖
||
オーソログ

他の
グロビン族
(パラログ)

非相同
蛋白質

	Score	E
	(bits)	Val
Sequences producing significant alignments:		
ue ベストヒット		
sp:HBB_HUMAN HEMOGLOBIN BETA CHAIN.	306	1e-83
sp:HBB_GORGO HEMOGLOBIN BETA CHAIN.	305	3e-83
sp:HBB2_PANLE HEMOGLOBIN BETA-2 CHAIN.	302	2e-82
sp:HBB_HYLLA HEMOGLOBIN BETA CHAIN.	300	6e-82
sp:HBB_PREEN HEMOGLOBIN BETA CHAIN.	298	4e-81
sp:HBB_COLPO HEMOGLOBIN BETA CHAIN.	295	2e-80
sp:HBB_CERAE HEMOGLOBIN BETA CHAIN.	295	3e-80
sp:HBB_MACFU HEMOGLOBIN BETA CHAIN.	293	1e-79
sp:HBB_COLBA HEMOGLOBIN BETA CHAIN.	293	1e-79
sp:MYG_BALAC MYOGLOBIN.	49	5e-06
sp:MYG_MEGNO MYOGLOBIN.	48	8e-06
sp:MYG_ESCGI MYOGLOBIN.	48	1e-05
sp:MYG_BALPH MYOGLOBIN.	47	2e-05
sp:MYG_ZIPCA MYOGLOBIN.	46	4e-05
sp:GLB1_ARTSX GLOBIN E1, EXTRACELLULAR.	45	9e-05
sp:GLP2_GLYDI GLOBIN, POLYMERIC COMPONENT P2.	42	6e-04
sp:GLP1_GLYDI GLOBIN, MAJOR POLYMERIC COMPONENT P1.	41	8e-04
sp:HBAZ_MACEU HEMOGLOBIN ZETA CHAIN (FRAGMENTS).	39	0.005
sp:GLP3_GLYDI GLOBIN, POLYMERIC COMPONENT P3.	38	0.009
sp:LGB2_PEA LEGHEMOGLOBIN II.	36	0.035
sp:LGB1_PEA LEGHEMOGLOBIN I.	35	0.079
sp:LGB2_SESRO LEGHEMOGLOBIN 2.	34	0.18
sp:HBP_CANLI LEGHEMOGLOBIN.	32	0.40
sp:LGB1_VICFA LEGHEMOGLOBIN I.	32	0.53
sp:LACG_LACCA 6-PHOSPHO-BETA-GALACTOSIDASE (EC 3.2.1.85) (BETA-...)	32	0.53
sp:LGB3_SESRO LEGHEMOGLOBIN 3.	31	0.90
sp:LGBA_PHAVU LEGHEMOGLOBIN A.	31	0.90
sp:HMPA_BACSU FLAVOHEMOPROTEIN (HAEMOGLOBIN-LIKE PROTEIN) (FLAV...)	31	1.2

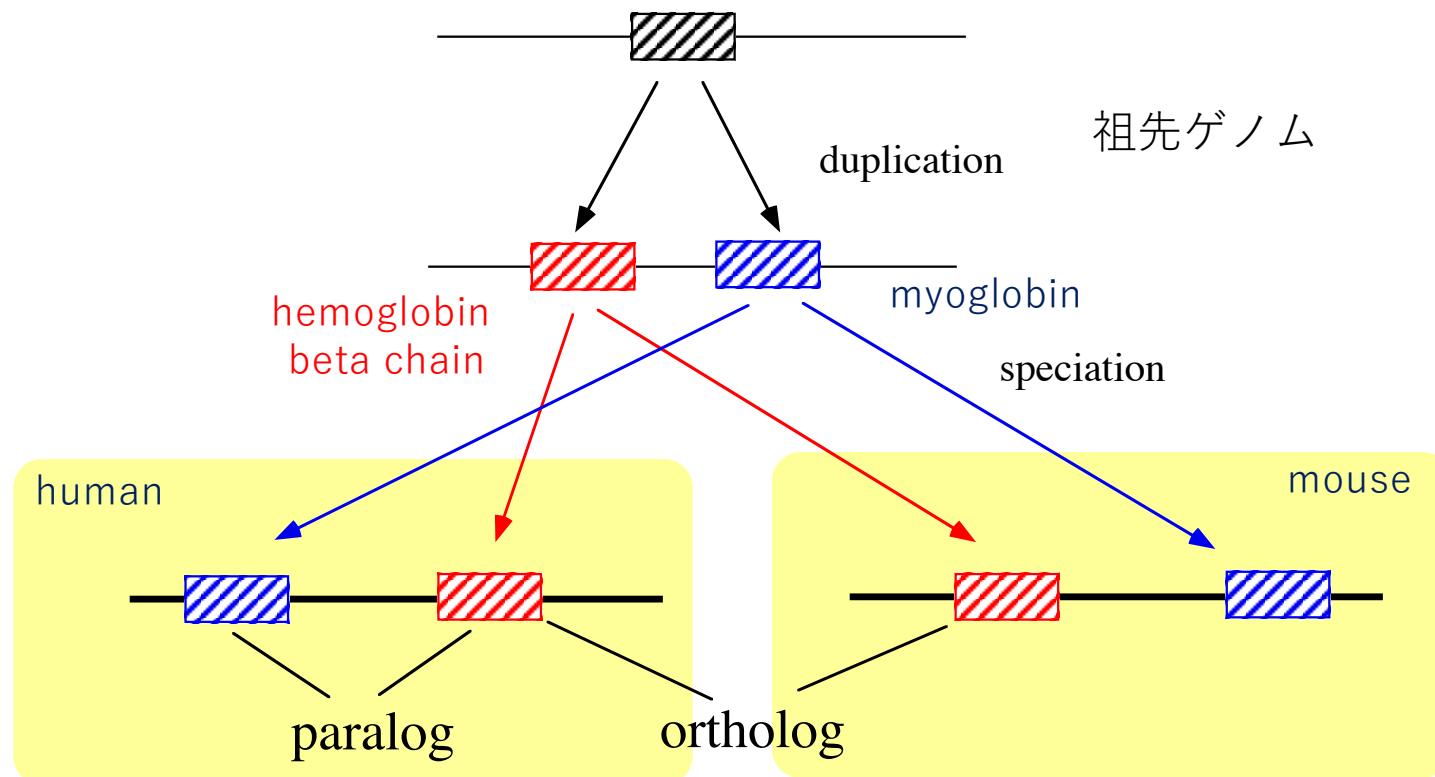
E-value
非相同配列を
間違って拾って
しまう個数の
期待値

統計的に
有意な類似性

微妙な類似性

統計的に
有意でない

オーソログとパラログ



BLAST 検索の実行

```
# 配列ファイルdb.faに検索用のインデックスを作成。データベース名 dbnameで出力。
% makeblastdb -in db.fa -dbtype prot -parse_seqids -out dbname
# 作成したデータベースに対して、query.faをクエリとした検索の実行。
# タブ区切り形式 (outfmt 6)で、標準の形式にタイトル行を附加して、上位10ヒットを出力。
% blastp -query query.fa -db dbname -evaluate 0.001
          -outfmt "6 std stitle" -max_target_seqs 10 > blastout.tab
```

1. query	2. subject (database)	3. %identity	5. mismatch 6. gap_open 9. s_start												12. bit-score	13. title
			4. align-len	7. q_start		8. q_end		10. s_end	11. evaluate							
spo:NP_001018179.1	sce:NP_010076.1	40.373	322	181	4	6	322	5	320	2.05e-81	248	hydroxymethylbilane synthase [Saccharomyces cerevisiae CEN.PK2-1	248	2.05e-81	hydroxymethylbilane synthase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018181.1	sce:NP_014526.1	41.772	316	179	4	213	524	149	463	2.04e-80	265	non-canonical poly(A) polymerase [Saccharomyces cerevisiae CEN.PK2-1	265	2.04e-80	non-canonical poly(A) polymerase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018181.1	sce:NP_014100.2	34.541	414	239	9	138	545	89	476	8.12e-72	243	non-canonical poly(A) polymerase [Saccharomyces cerevisiae CEN.PK2-1	243	8.12e-72	non-canonical poly(A) polymerase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018187.2	sce:NP_013081.1	47.176	301	132	6	345	628	294	584	1.65e-88	290	serine/threonine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	290	1.65e-88	serine/threonine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018187.2	sce:NP_013081.1	53.125	32	15	0	655	686	693	724	5.58e-04	41.2	serine/threonine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	41.2	5.58e-04	serine/threonine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018187.2	sce:NP_012394.1	27.249	378	222	6	343	684	346	706	1.86e-41	159	serine/threonine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	159	1.86e-41	serine/threonine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018187.2	sce:NP_013214.1	28.614	332	182	10	357	681	18	301	3.31e-31	125	mitogen-activated protein kinase [Saccharomyces cerevisiae CEN.PK2-1	125	3.31e-31	mitogen-activated protein kinase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018187.2	sce:NP_014092.1	28.571	364	191	16	341	684	15	329	2.97e-30	120	serine/threonine/tyrosine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	120	2.97e-30	serine/threonine/tyrosine protein kinase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018187.2	sce:NP_009537.1	28.319	339	183	10	362	682	13	309	2.04e-27	112	mitogen-activated serine/threonine kinase [Saccharomyces cerevisiae CEN.PK2-1	112	2.04e-27	mitogen-activated serine/threonine kinase [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018191.1	sce:NP_011482.3	27.803	223	135	7	1	208	1	212	1.41e-12	62.4	Hop2p [Saccharomyces cerevisiae CEN.PK2-1	62.4	1.41e-12	Hop2p [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018193.1	sce:NP_012148.1	64.126	223	53	3	7	204	10	230	6.88e-98	281	Rho family GTPase RHO3 [Saccharomyces cerevisiae CEN.PK2-1	281	6.88e-98	Rho family GTPase RHO3 [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018193.1	sce:NP_015491.1	51.934	181	83	2	12	188	9	189	1.09e-61	189	Rho family GTPase RHO1 [Saccharomyces cerevisiae CEN.PK2-1	189	1.09e-61	Rho family GTPase RHO1 [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018193.1	sce:NP_014309.3	49.432	176	88	1	9	183	3	178	8.31e-54	168	Rho family GTPase RHO2 [Saccharomyces cerevisiae CEN.PK2-1	168	8.31e-54	Rho family GTPase RHO2 [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018193.1	sce:NP_012981.3	45.000	180	93	3	7	184	65	240	5.61e-48	156	Rho family GTPase RHO4 [Saccharomyces cerevisiae CEN.PK2-1	156	5.61e-48	Rho family GTPase RHO4 [Saccharomyces cerevisiae CEN.PK2-1	
spo:NP_001018193.1	sce:NP_013330.1	44.509	173	92	1	15	183	5	177	1.07e-46	150	Rho family GTPase CDC42 [Saccharomyces cerevisiae CEN.PK2-1	150	1.07e-46	Rho family GTPase CDC42 [Saccharomyces cerevisiae CEN.PK2-1	

BLASTタブ区切り出力からのベストヒット抽出

ベストヒットの出力

```
% sort -k 1,1 -u blastout.tab > blast_top.tab
```

(クエリ配列名をキーとしてソートし、キーが重複した場合は最初の行のみを出力。元の並びがE-value(スコア)の順になっており、その順でチェックされるため、ベストヒット1つのみが出力される)

双方向ベストヒット（オーソログの推定）の出力

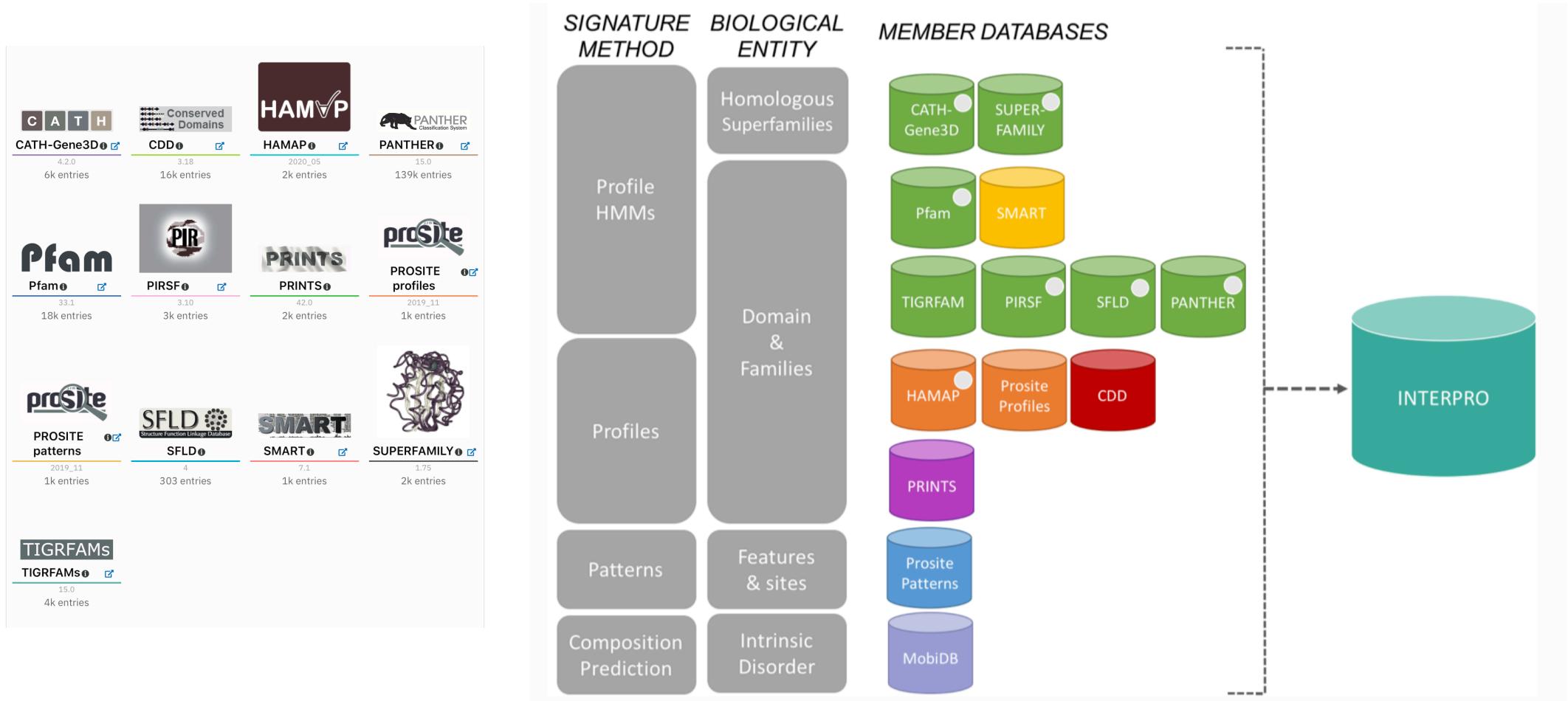
```
% sort -k 2,2 -k 11,11g -k 12,12nr blast_top.tab  
| sort -k 2,2 -u > blast_bbh.tab
```

(上記の結果を、サブジェクト配列ごとのE-valueとスコア順に並べ直した上で、上記と同様のコマンドで逆方向のベストヒットを抽出する)

DIAMOND 超高速ホモジー検索

```
# インデックスの作成
% diamond makedb --in db.fa --db db
# 検索の実行
% diamond blastp --query query.fa --db db
  --eval 0.001 --max-target-seqs 10
  --outfmt 6 qseqid sseqid pident eval bitscore stitle
  --out diamondout.tab
```

InterProScan モチーフ／ドメイン検索

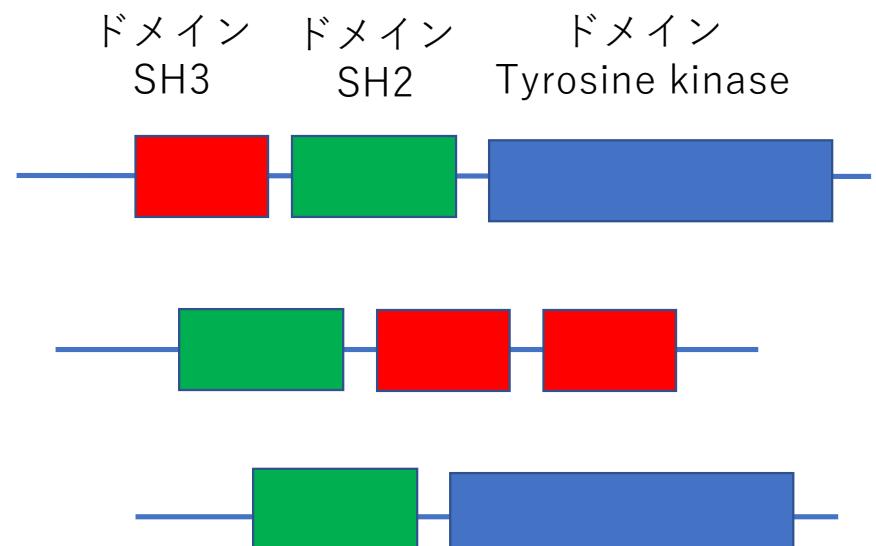
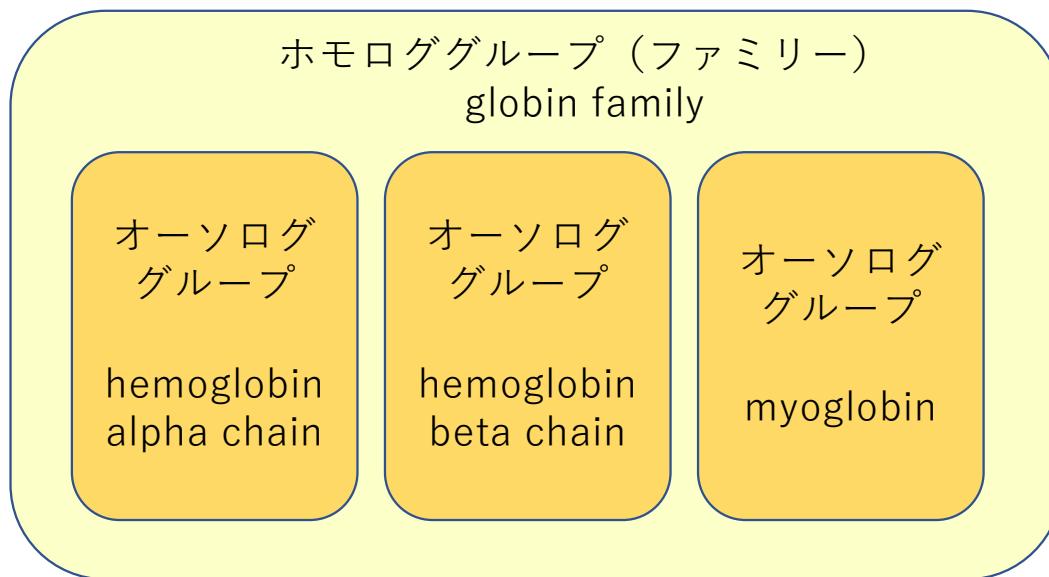


InterProScan の実行

query.faをクエリとして、InterProのデータベース全てを対象として検索。検索結果にはGO termも含める。

```
% interproscan.sh -i query.fa -goterms
```

ファミリー／ドメインアノテーションの複雑さ



検索結果は各配列に一つではなく、ドメイン単位でつけられるもの、ドメインはオーバーラップしているが、ファミリーの大きさが異なるものなどが含まれる。

ホモロジーに基づくアノテーションの戦略

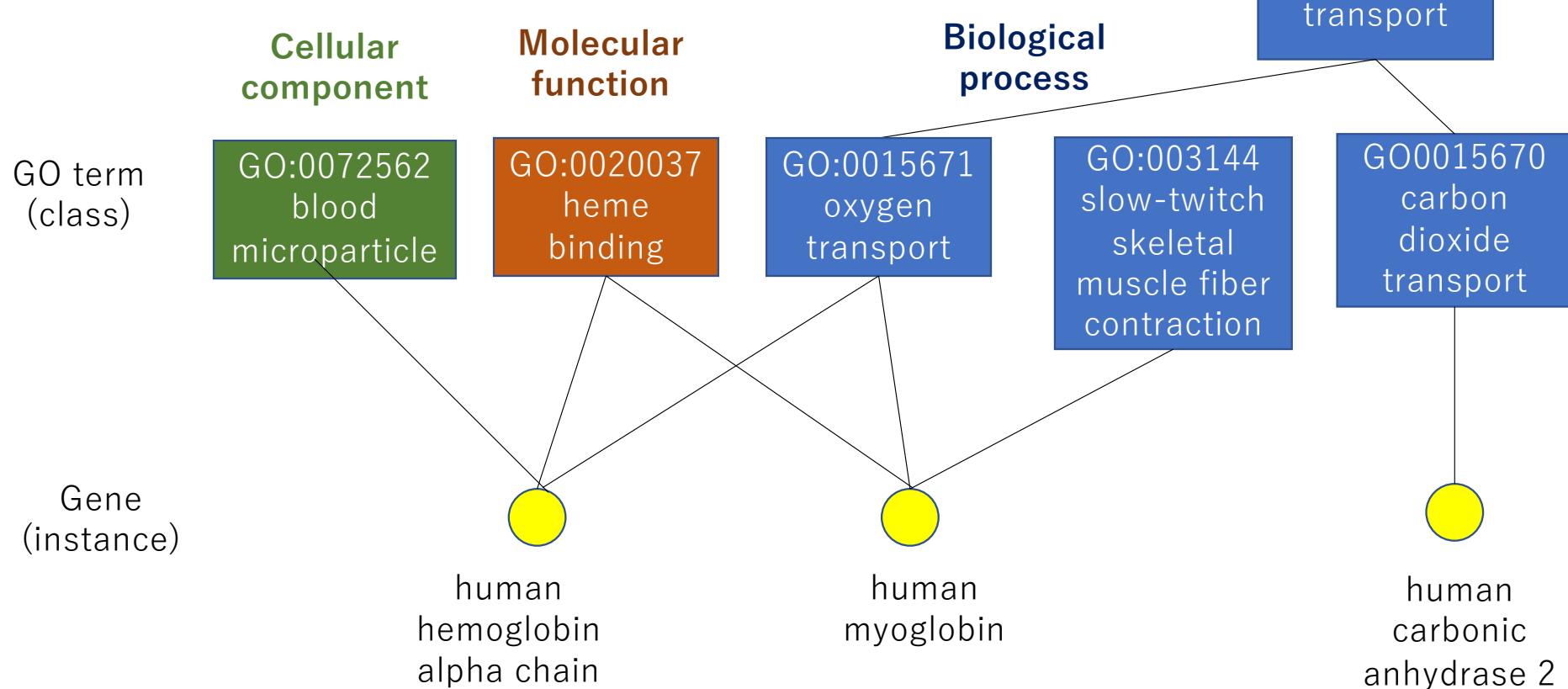
- 近縁種で、高品質なアノテーションがついたモデル生物ゲノムが利用可能な場合——そのゲノムに対するオーソログを同定し、アノテーションをコピーする。
- 特にターゲットとする生物種を絞らず、幅広い生物種から情報を集めたい場合——nrなどの網羅的なデータベースを検索して上位のヒットのアノテーションをコピーする。
- ホモロジー検索だけでは類似性が低い配列しかヒットせず、信頼性が低い場合——InterProなどのモチーフデータベースの検索を併用する。

テキスト記述によるアノテーションの問題点

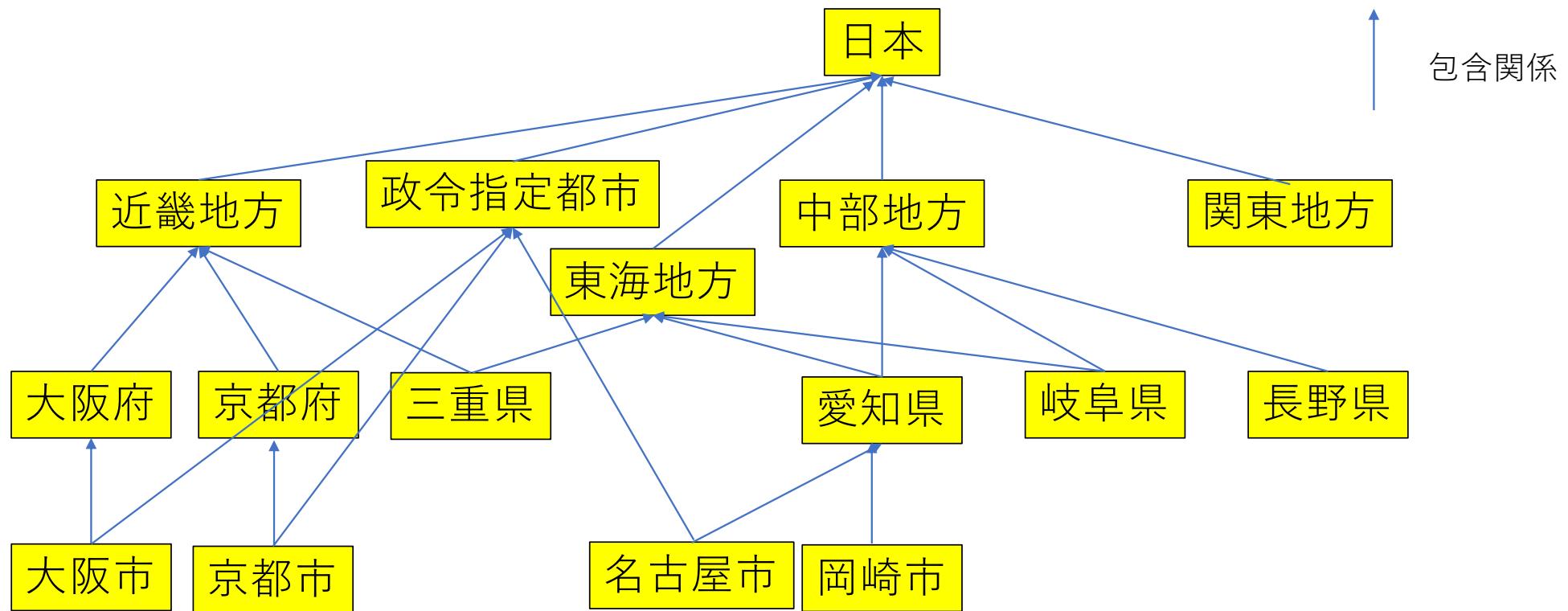
- 基本的に遺伝子（タンパク質）の名前を記載しただけで、具体的な機能について記載しているわけではない。
- 生物学的な解釈を考えるには、各遺伝子の機能に関する知識が別途必要になる。
- 大規模なRNA-seq解析結果を解釈するには、この部分についても計算機のサポートが必要。

Gene Ontology (GO)

機械可読な遺伝子機能の表現

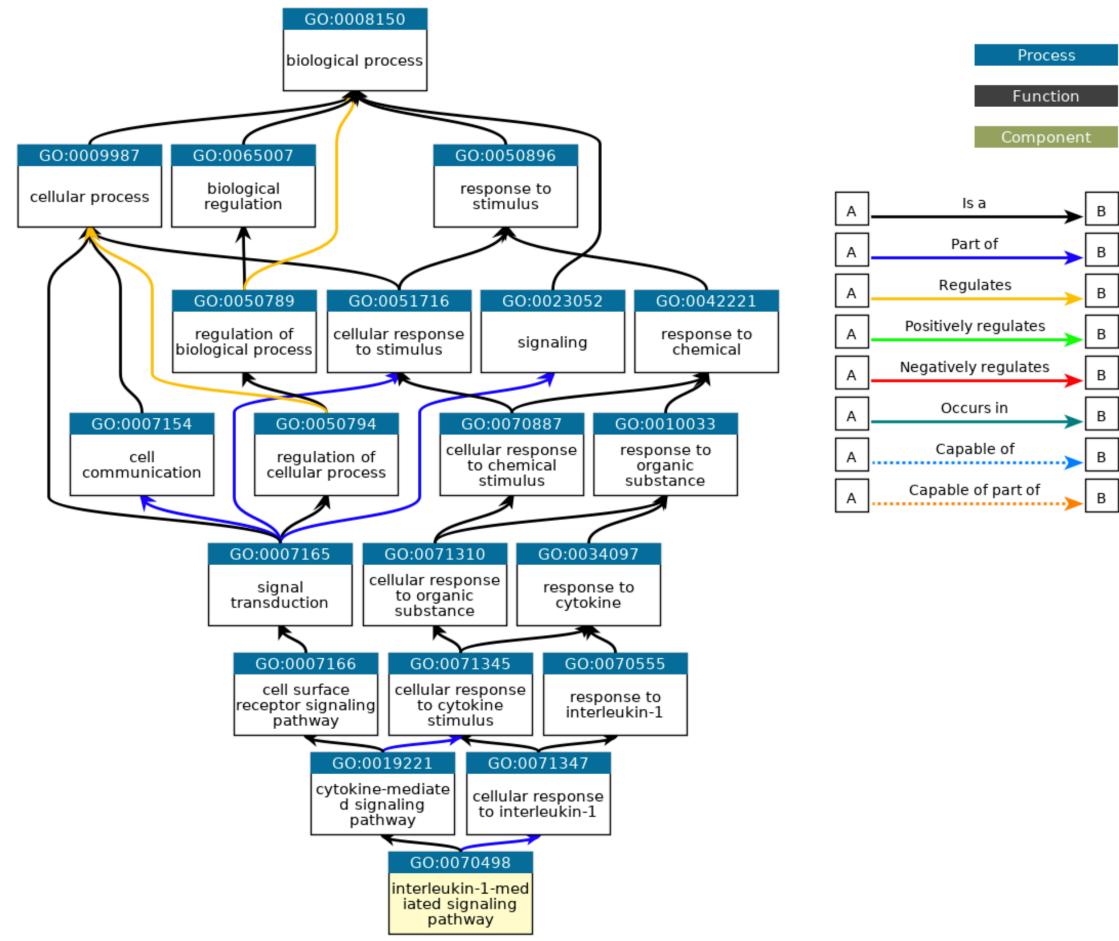


包含関係を表すグラフ：有向非巡回グラフ Directed Acyclic Graph (DAG)



GO階層のグラフ

- 各ノードはGO termを表す
 - 最上位ノードは以下の3種類
 - biological process
 - molecular function
 - cellular component
- 各矢印（エッジ）はGO term間の関係を表す
 - 包含関係を表す矢印は2種類
 - is_a AはBの一種である
 - part_of AはBの一部である
 - その他の関係を表す矢印
 - regulates
 - occurs_in など



QuickGO - <https://www.ebi.ac.uk/QuickGO>

GO annotation

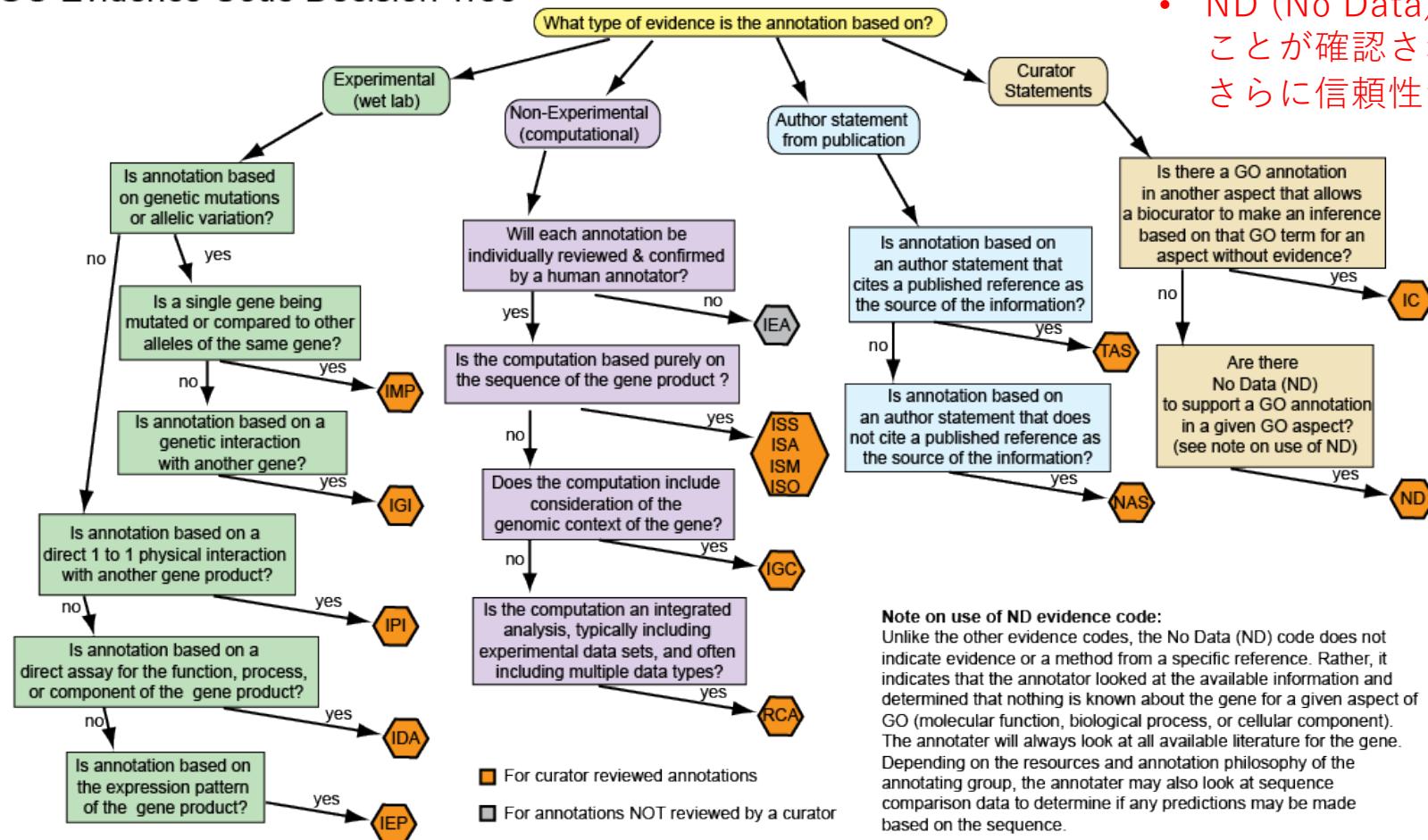
既知遺伝子へのGO termの割り当て

- GOアノテーションデータベース
 - モデル生物ゲノムデータベース
 - MGI (マウス)
 - FlyBase (ショウジョウバエ)
 - TAIR (シロイヌナズナ)
 - SGD (酵母) など
 - 網羅的なデータベース
 - タンパク質配列データベースUniProt
 - モチーフドメインデータベース InterPro
 - パスウェイデータベースReactome など
- アノテーションの根拠を Evidence Codeで示す

Human IKBKB gene に対するGO アノテーション													
Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Type	Isoform	Reference	Date
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		stimulatory C-type lectin receptor signaling pathway		Reactome	Homo sapiens	TAS		ikb kinase pthr22969	protein		Reactome:R-HSA-5621481	20181121
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent		Reactome	Homo sapiens	TAS		ikb kinase pthr22969	protein		Reactome:R-HSA-1236974	20180419
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		MyD88-independent toll-like receptor signaling pathway		Reactome	Homo sapiens	TAS		ikb kinase pthr22969	protein		Reactome:R-HSA-168927	20171201
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein kinase activity		UniProt	Homo sapiens	IDA		ikb kinase pthr22969	protein		PMID:20434986	20100804
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein serine/threonine kinase activity		UniProt	Homo sapiens	IDA		ikb kinase pthr22969	protein		PMID:15084260	20151001
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein serine/threonine kinase activity		Reactome	Homo sapiens	EXP		ikb kinase pthr22969	protein		PMID:18692471	20170505
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein serine/threonine kinase activity		Reactome	Homo sapiens	EXP		ikb kinase pthr22969	protein		PMID:23613522	20170505
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein serine/threonine kinase activity		UniProt	Homo sapiens	IDA		ikb kinase pthr22969	protein		PMID:25326418	20200702
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein serine/threonine kinase activity		CACAO	Homo sapiens	IDA		ikb kinase pthr22969	protein		PMID:25636800	20151001
IKBKB	Inhibitor of nuclear factor kappa-B kinase subunit beta		protein serine/threonine kinase activity		Reactome	Homo sapiens	TAS		ikb kinase pthr22969	protein		Reactome:R-HSA-168140	20170811
IKBKB	Inhibitor of nuclear factor kappa-B		protein serine/threonine kinase activity		Reactome	Homo sapiens	TAS		ikb kinase pthr22969	protein		Reactome:R-HSA-202541	20150207

Evidence Code

GO Evidence Code Decision Tree



- IEA (Inferred from Electronic Annotation) は計算機による予測のみなので要注意！
- ND (No Data) は、証拠がないことが確認されているので、さらに信頼性が低い

Note on use of ND evidence code:

Unlike the other evidence codes, the No Data (ND) code does not indicate evidence or a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene for a given aspect of GO (molecular function, biological process, or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

ホモロジーに基づくアノテーションの戦略（再掲）

- 近縁種で、高品質なアノテーションがついたモデル生物ゲノムが利用可能な場合——そのゲノムに対するオーソログを同定し、アノテーションをコピーする。
- 特にターゲットとする生物種を絞らず、幅広い生物種から情報を集めたい場合——nrなどの網羅的なデータベースを検索して上位のヒットのアノテーションをコピーする。 (**→要注意！**)
- ホモロジー検索だけでは類似性が低い配列しかヒットせず、信頼性が低い場合——InterProなどのモチーフデータベースの検索を併用する。

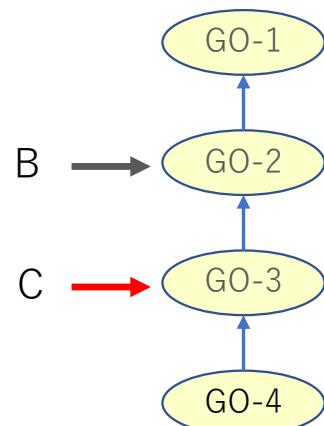
GOアノテーションのついたデータベースを検索して、その結果に基づいてGOアノテーションをコピーすればよい。ただし、アノテーションのクオリティを考慮すると、2番めのケースは特に注意が必要。

ホモロジーに基づくGOアノテーション

Hit	Score	GO	Evidence
Gene-A	90	GO-2	IEA
Gene-B	85	GO-2	IDA
Gene-C	80	GO-3	IDA
Gene-D	60	GO-4	IEA

- 類似性検索でトップヒットのGOを採用すれば良いとは限らない。

→スコアが同程度なら、エビデンスコードを考慮して、より信頼性の高いアノテーションを採用した方がよい。



- あるGO termがアサインされる場合、その上位のGO termも必ずアサインされる。

→アノテーションとしては、その遺伝子に当てはまる最も下位のGO termを記載する。

BLAST2GO

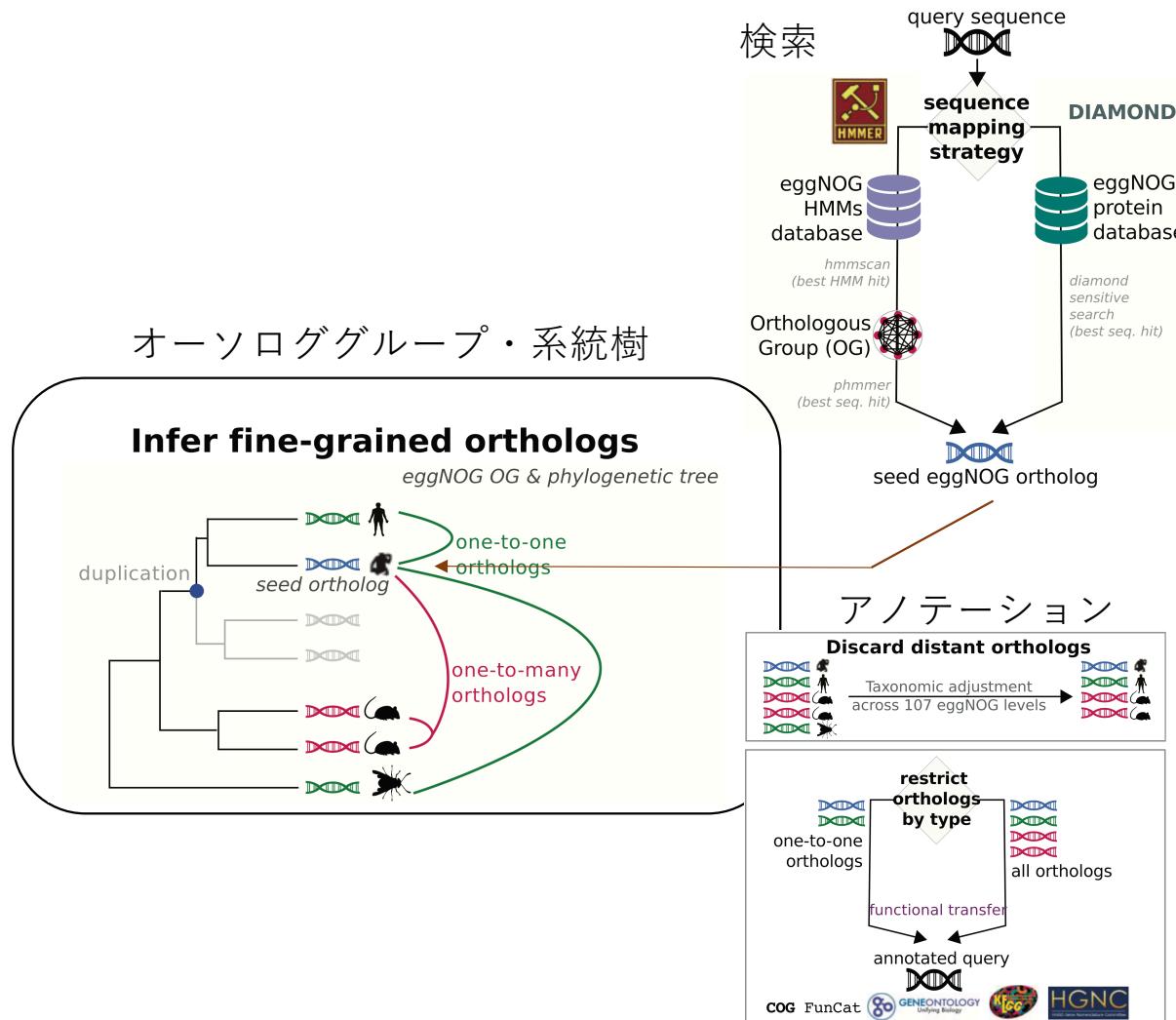
- Annotation Score: $AS = DT + AT$
- $DT = \max(\text{similarity} \times ECw)$
 ECw : Evidence Code weight
- $AT = (\#GO - 1) \times GOw$
 $\#GO$: number of child GOs assigned
 GOw : GO weight
- $AS \geq \text{threshold}$ を満たす最下層のGOをアサインする

Evidence code weight

EC	Description	Default
IDA	Inferred from direct assay	1
IMP	Inferred from mutant phenotype	1
IGI	Inferred from genetic interaction	1
IPI	Inferred from physical interaction	1
IEP	Inferred from expression pattern	1
TAS	Traceable author statement	0.9
NAS	Non-traceable author statement	0.9
IC	Inferred by curator	0.9
ISS	Inferred from sequence or structural similarity	0.9
RCA	Inferred from reviewed computational analysis	0.9
IEA	Inferred from electronic annotation	0.7
ND	No biological data available	0.5
NR	Not recorded	0.5

→OmixBoxという有償ソフトへ統合化

EggNOG Mapper オーソログベースのアノテーション



- あらかじめデータベース中の配列をオーソロググループに分類し、各グループの系統樹を作成。
- クエリ配列に対するホモジー検索によりseed ortholog配列を同定し、系統樹を用いて他生物種におけるオーソログを同定する。
- seed orthologに対する類縁関係を考慮し、近縁種のオーソログに付けられた機能アノテーションをコピーする。

EggNOG mapper の実行

DIAMONDを用いて、EggNOG mapperの実行

```
% emapper.py -i query.fa -o outname -m diamond
```

アノテーション結果から、遺伝子名とテキスト記述を抽出

```
% cut -f1,6,22 outname.emapper.annotations
```

```
#query_name Preferred_name eggNOG free text desc.  
HMPREF9499_RS03205      Glucosyl transferase GtrII  
HMPREF9499_RS03210      rpll  Forms part of the ribosomal stalk which helps the ribosome interact with GTP-bound translation ...  
HMPREF9499_RS03215      rplJ   Forms part of the ribosomal stalk, playing a central role in the interaction of the ribosome ...  
HMPREF9499_RS03220      rpla  Binds directly to 23S rRNA. The L1 stalk is quite mobile in the ribosome, and is involved in ...  
HMPREF9499_RS03225      rplK   Forms part of the ribosomal stalk which helps the ribosome interact with GTP-bound translation ...
```

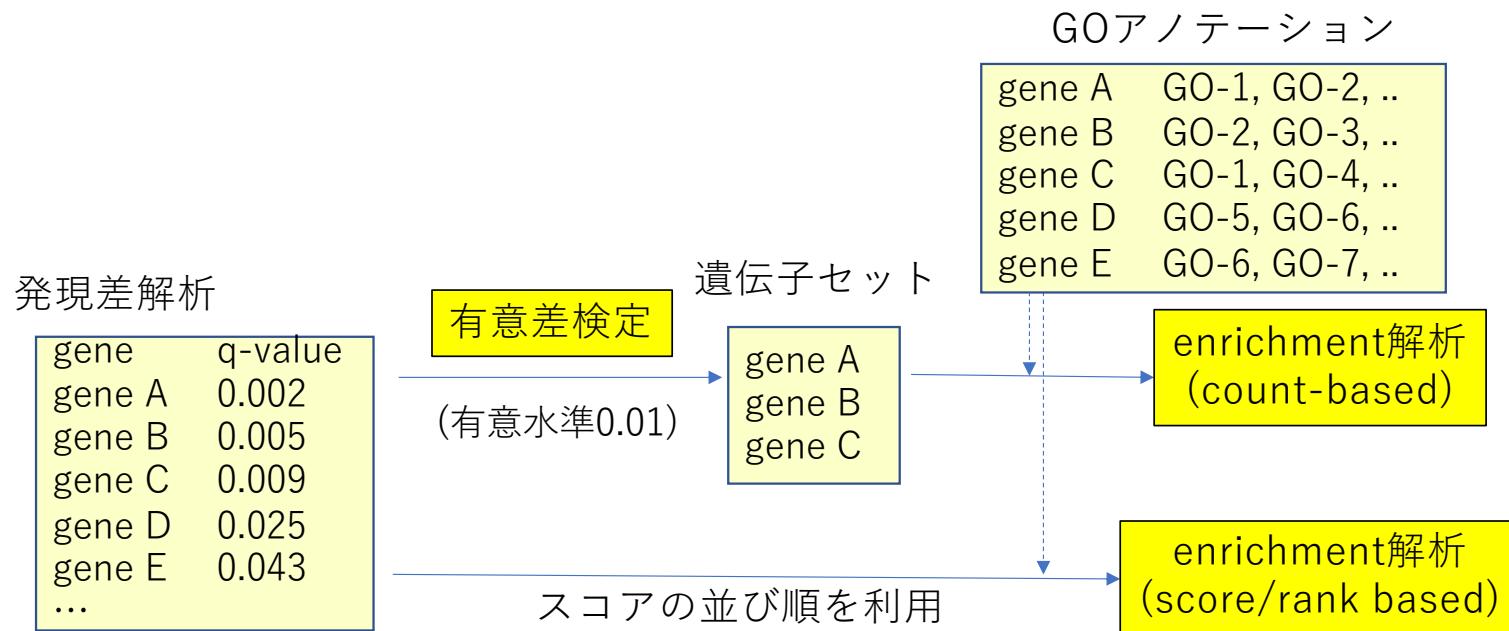
アノテーション結果から、アサインされたGOのリストを抽出

```
% cut -f1,7 outname.emapper.annotations
```

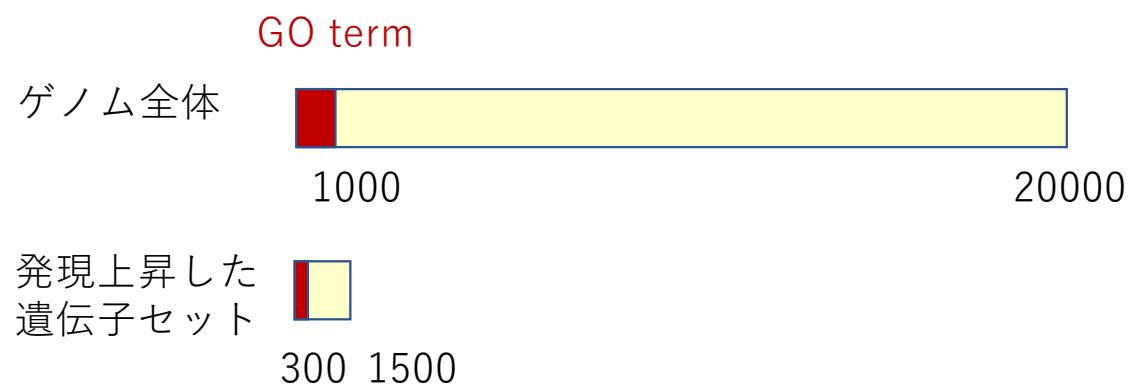
```
#query_name      GOs  
HMPREF9499_RS03205  
HMPREF9499_RS03215      GO:0003674,GO:0003735,GO:0005198,GO:0005488,GO:0005575,GO:0005618,GO:0005622,GO:0005623,...  
HMPREF9499_RS03220      GO:0000027,GO:0000470,GO:0003674,GO:0003676,GO:0003723,GO:0003735,GO:0005198,GO:0005488,...  
HMPREF9499_RS03225      GO:0000027,GO:0003674,GO:0003676,GO:0003723,GO:0003735,GO:0005198,GO:0005488,GO:0005575,...
```

GO enrichment 解析

- ・発現量が増加／減少した遺伝子群において、より多く出現する（エンリッチしている）機能(GO term)を抽出する。
- ・まず設定した閾値（有意水準）によって遺伝子セットを抽出し、その中でエンリッチメント解析を行うアプローチと、スコアの並び順を用いてエンリッチメント解析を行うアプローチがある。



Enrichment analysis by Fisher's exact test



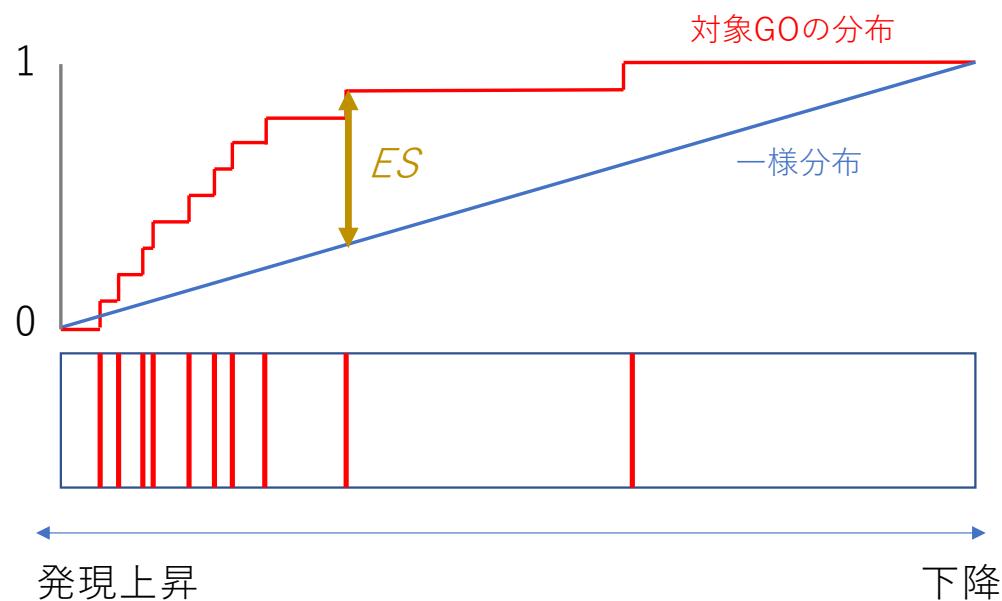
分割表			
	機能を持つ	機能を持たない	計
発現上昇あり	300	1200	1500
発現上昇なし	700	17800	18500
計	1000	19000	20000

そのGOは、発現上昇した遺伝子セット中でエンリッチしていると言えるか？

赤玉1000個、白玉19000個入った壺から、ランダムに玉を1500個抽出するとき、
その中に赤玉が300個以上含まれる確率を求める。

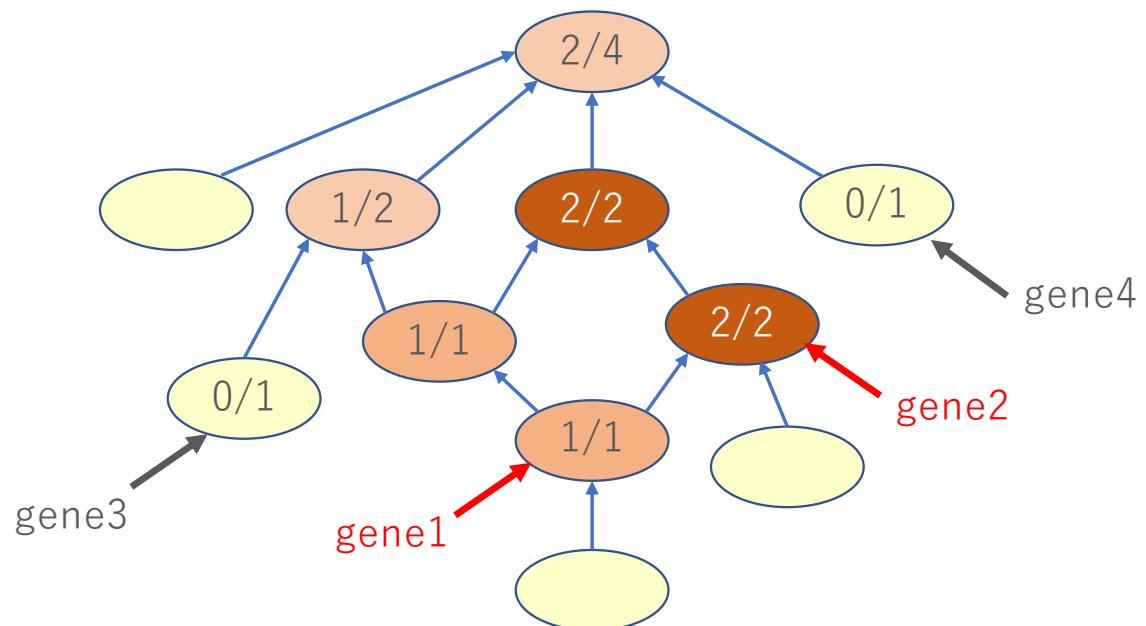
超幾何分布(hypergeometric distribution)を使って正確に計算できる
→ Fisherの正確確率検定 (Fisher's exact test)

Gene Set Enrichment Analysis (GSEA)



- 全遺伝子を発現解析の結果に基づいてソートし、その並び順の中で、対象とする遺伝子セットの出現が、偏りのない分布（一様分布）からずれているかを判定する（Kolmogorov-Smirnov検定）。
- カウントを、ソートに用いた指標で重み付けすることにより、上位のヒットにより大きな重みを与えることもできる。

GO enrichment 解析とGO階層

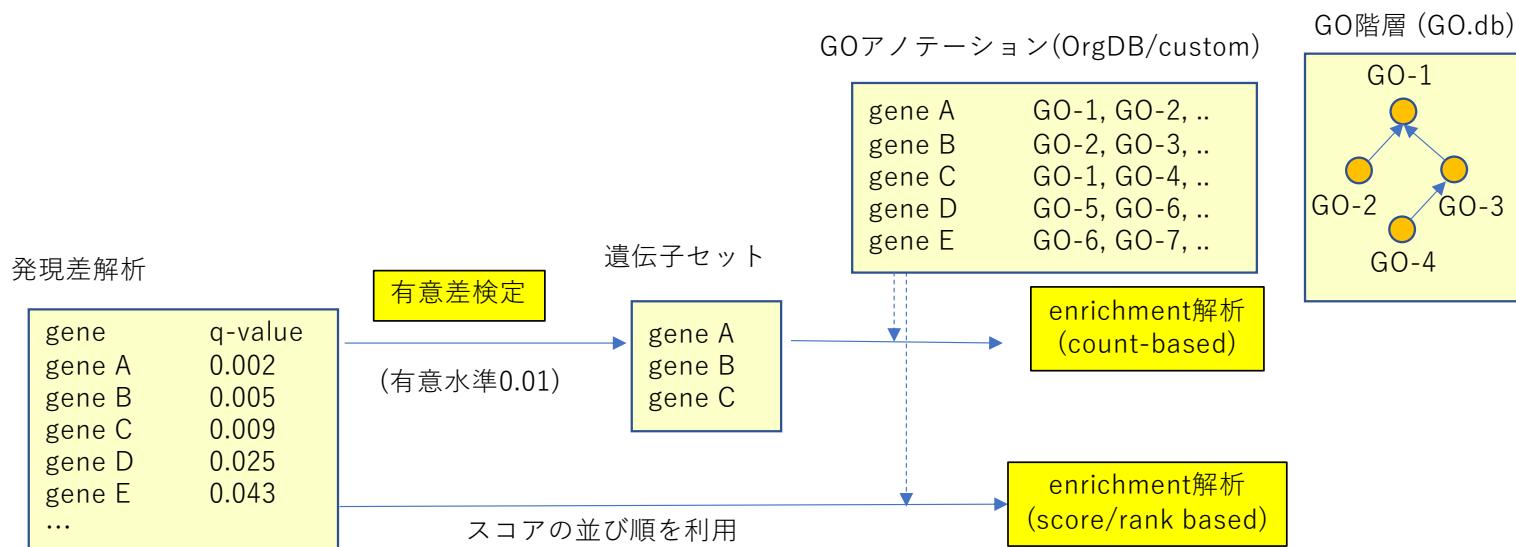


- 下位ノードにアサインされた遺伝子は上位ノードにも自動的にアサインされる。
- 上位ノードは分母が大きくなるため、「濃縮度」は低くなる。
- 有意なGO termは、GO階層上近傍に集まって出現することが多い。

Rを用いたGO enrichment 解析

- clusterProfiler
 - GOのほか、KEGGやDisease Ontology(DO)など、様々なデータベースに対応
 - 結果を可視化するツールが充実している

- TopGO
 - 隣接するGO階層間で生じる冗長性を排除するアルゴリズムを実装
 - 様々な統計手法とアルゴリズムを組み合わせてenrichment解析を行う汎用的な枠組みを提供



入力データの準備

```
# edgeRのデータオブジェクト(DGEList)から解析結果全体をテーブルとして抜き出す
etab <- TopTags(DGEList, n=999999)$table

## count-based 解析用の遺伝子リスト作成
# 有意水準0.001として有意に発現上昇した遺伝子を抽出
upreg <- subset(etab, logFC > 0 & etab$FDR < 0.001)
# 発現上昇した遺伝子と全体の遺伝子名リストを作成
upregGenes <- rownames(upreg)
allGenes <- rownames(etab)

## score-based 解析用のスコアリスト作成
# -log(FDR)にlogFCの符号をかけてup/downを区別した指標を算出
expScore <- sign(etab$logFC) * -log(etab$FDR)
# expScoreの各値にnameとして遺伝子名を対応づける
names(expScore) <- rownames(etab)
```

GOアノテーションの準備

公開されたアノテーションデータベースを利用する場合

OrgDb

- 代表的モデル生物種の遺伝子データベースでGOのアノテーションを含む。
- org.<Sp>.<id>.db という名称のパッケージとして公開。ヒトorg.Hs.eg.db、マウスorg.Mm.eg.dbなど、<id>=eg (Entrez Gene)のものを中心に現在約20種類。
- enrichment解析に用いるには、発現解析に用いた遺伝子IDが、NCBI GeneIDなど、このデータベースに登録されたIDのいずれかと一致する必要がある。

```
> library(org.Hs.eg.db)
# ロードすると、org.Hs.eg.db という名前のオブジェクトを介してデータベースにアクセスできる
# 利用可能な属性の一覧を表示
> columns(org.Hs.eg.db)
"ACCCNUM" "ALIAS" "ENSEMBL" "ENSEMBLPROT" "ENSEMBLTRANS" "ENTREZID" "ENZYME"
"EVIDENCE" "EVIDENCEALL" "GENENAME" "GO" "GOALL" "IPI" "MAP"
"OMIM" "ONTOLOGY" "ONTOLOGYALL" "PATH" "PFAM" "PMID"
"PROSITE" "REFSEQ" "SYMBOL" "UCSCKG" "UNIGENE" "UNIPROT"
# SYMBOL="KRAS" 遺伝子のGOを表示
> select(org.Hs.eg.db, keys="KRAS", keytype="SYMBOL", columns="GO")
SYMBOL GO EVIDENCE ONTOLOGY
1 KRAS GO:0000165 TAS BP
2 KRAS GO:0001889 IEA BP
3 KRAS GO:0001934 IMP BP
...
```

GOアノテーションの準備

公開されたアノテーションデータベースを利用する場合

AnnotationHub

- より多くの生物種のアノテーションを蓄積。必要なものをダウンロードして使う。"OrgDb"として登録されたものは前項のOrgDbデータベースと同様に使える。

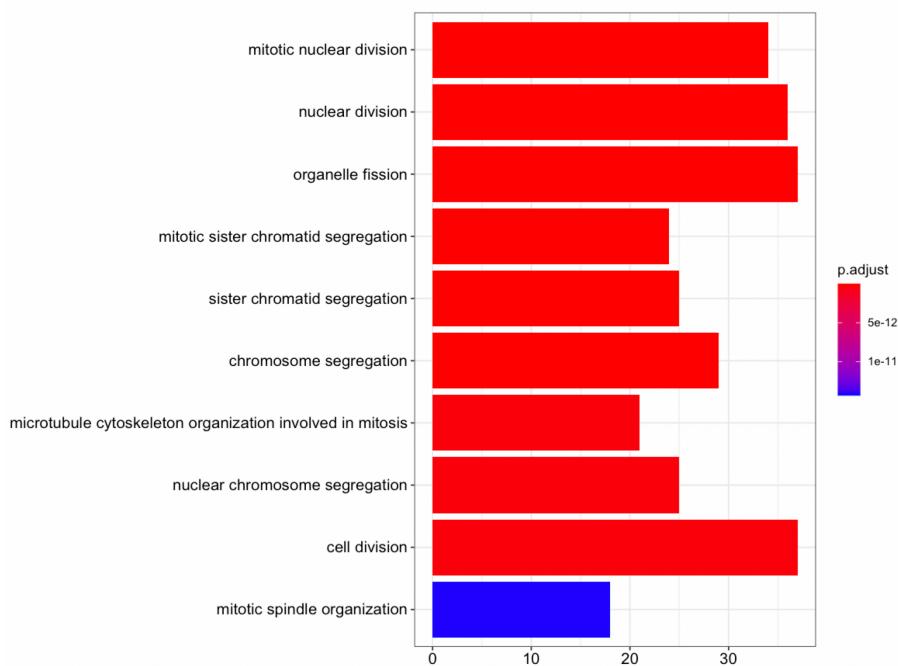
```
> library(AnnotationHub)
# AnnotationHubへのアクセスインターフェイスを変数ahに格納
> ah <- AnnotationHub()
# イネ(Oryza sativa)を検索。Queryコマンドの2番目の引数は、任意の数のキーワードをベクトルとして与える。
> query(ah, c("OrgDb", "oryza sativa"))
AnnotationHub with 3 records # <- ヒットが3つ見つかっている
# snapshotDate(): 2020-10-27
...
# retrieve records with, e.g., 'object[[ "AH85565" ]]' # <- データへのアクセス方法。objectは変数名ahに置き換える
      title          # 以下、ヒットしたデータの一覧
AH85565 | org.Oryza_sativa_(japonica_cultivar-group).eg.sqlite
AH85566 | org.Oryza_sativa_Japonica_Group.eg.sqlite
AH85567 | org.Oryza_sativa_subsp._japonica.eg.sqlite
# データをダウンロードし、org.os.dbという変数に格納
> org.os.db <- ah[["AH85565"]]
> columns(org.os.db)
"ACCCNUM"    "ALIAS"    "CHR"      "ENTREZID"   "EVIDENCE"   "EVIDENCEALL"  "GENENAME"   "GID"       "GO"
"GOALL"      "ONTOLOGY"  "ONTOLOGYALL" "PMID"        "REFSEQ"     "SYMBOL"
```

clusterProfilerによる解析(OrgDbを用いる場合)

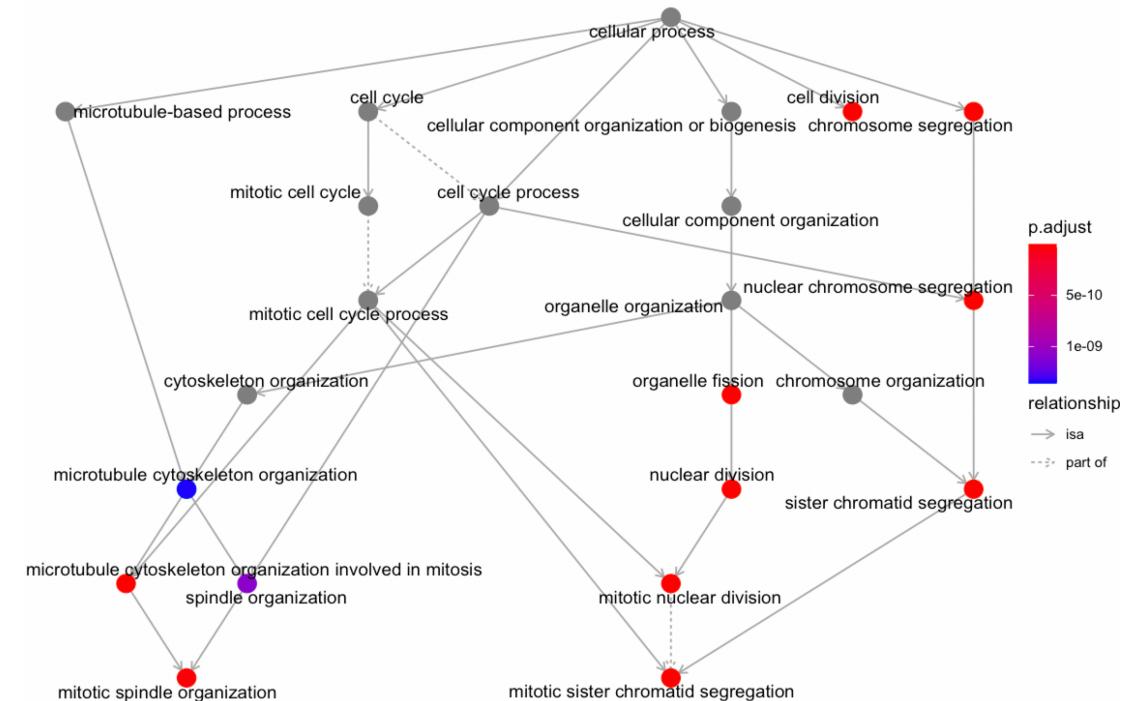
```
# Fisher's exact testの実施
> cprof.fisher <- enrichGO(gene=upregGenes, universe=allGenes,
+                             OrgDb=org.Hs.eg.db, ont="BP", qvalueCutoff=0.05)
# Fisher's exact testの実施
> head(cprof.fisher, 20)
# Gene Set Enrichment Analysisの実施
> cprof.gsea <- gseGO(geneList=expScore, OrgDb=org.Hs.eg.db, ont="BP",
+                        qvalueCutoff=0.05)
```

解析結果の可視化

```
> barplot(cprof.fisher, showCategory=10)
```



```
> goplot(cprof.fisher, showCategory=10)
```



GOアノテーションの準備

自分で作成したアノテーションを取り込む場合

- アプリケーションで指定された形式で遺伝子とGO ID の対応表を準備する。
- clusterProfilerでは、以下のいずれかの形式で準備する。TERM2GENE引数で与える。

GMT 形式

GO-1	GOname1	gene1	gene2	gene3
GO-2	GOname2	gene4	gene5	

`read.gmt(gmfile)` で読み込む。
GOnameを読み飛ばして、右の形式と同じ
データフレームを生成する。

または、以下の形式

GO-1	gene1
GO-1	gene2
GO-1	gene3
GO-2	gene4
GO-2	gene5

`read.delim(file)` で読み込む

- clusterProfilerでは、GO ID と名前の対応表(TERM2NAME)も用意する必要がある。

GO-1	GOname1
GO-2	GOname2

GOアノテーションの準備

遺伝子とトランスクリプトの対応表の作成

- 実際の解析はトランスクリプトではなく、遺伝子単位で行うことが多い
- StringTieの出力GTFファイルのtranscript行9カラム目から、gffreadを用いて、gene ID(@geneid)、transcript ID (@id)、トランスクリプト長の情報を抜き出す。
% gffread -table @geneid,@id,@covlen stringtie_merged.gtf
 > gene2transcript.txt
- 遺伝子ごとに最長のトランスクリプトを取る
% sort -k 1,1 -k 3,3nr gene2transcript.txt | sort -k 1,1 -u

clusterProfilerによる解析(対応表を読み込む)

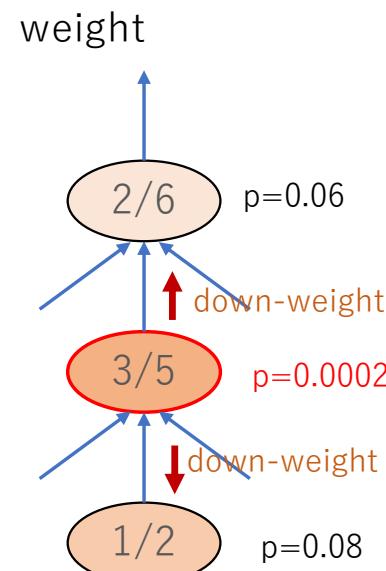
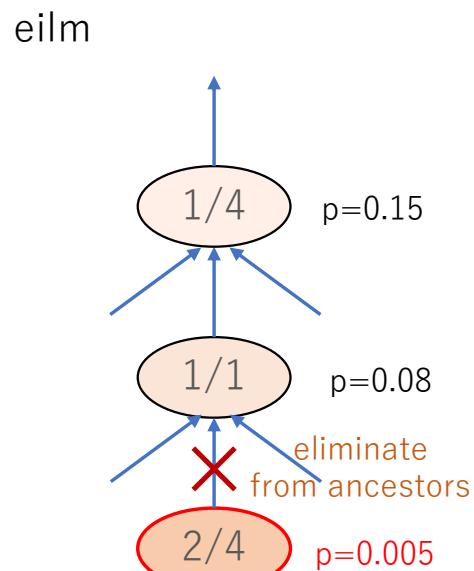
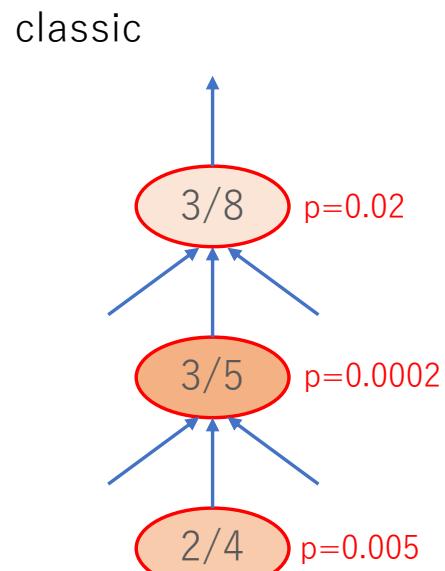
```
# 遺伝子とGOID対応表の読み込み
> go2gene <- read.table("go2gene.txt")
# 祖先ノードに遡ってGOをアサインする
> go2gene.expand <- biuldGOmap(go2gene)
# TERM2NAMEの対応表をGO.dbから作成する
> gonames <- select(GO.db, keys=keys(GO.db), columns="TERM")
# Fisher's exact testの実施
> cprof.fisher2 <- enricher(gene=upregGenes, TERM2GENE=go2gene.expand,
    TERM2NAME=gonames)
# Gene Set Enrichment Analysisの実施
> cprof.gsea2 <- GSEA(geneList=expScore, TERM2GENE=go2gene.expand
    TERM2NAME=gonames)
```

TopGO

- ・多様なGOエンリッチメント解析に対応したRパッケージ。
- ・様々な「統計手法」と、それをGO階層に適用する際の「アルゴリズム」の組み合わせを選択できる。

algorithms	statistics				
	fisher	ks	t	globaltest	sum
classic	✓	✓	✓	✓	✓
elim	✓	✓	✓	✓	✓
weight	✓	—	—	—	—
weight01	✓	✓	✓	✓	✓
lea	✓	✓	✓	✓	✓
parentchild	✓	—	—	—	—

Algorithms in TopGO



何もしない

子ノードを優先

p-valueが低い
ノードを優先

- ボトムアップに検定を行い、p-valueを計算。
- 有意性が確認されたノード中の遺伝子の影響が隣接するノードに広がらないように、親ノードや子ノードから遺伝子を除いたり、重みを下げたりする。

