

第12回バイオインフォマティクス研究会

GOtermに基づくPathway解析

contents

- GO term、pathway databaseについて
- GO解析、Pathway解析の有意性判定（超幾何分布）
- 実例

Runx2 K/Oマウス

発現変動遺伝子 → GO解析 → 機能推定

Gene Ontology

□ Gene Ontology Consortium

<http://www.geneontology.org>

生物学的概念を記述するための共通の語彙を策定するプロジェクト

GO term

- biological process (生物学的プロセス)
- cellular component (細胞の構成要素)
- molecular function (分子機能)

階層構造をもつ。上層 → 一般的機能表現、下層 → 特異的機能表現

非循環有向グラフで記述できる。

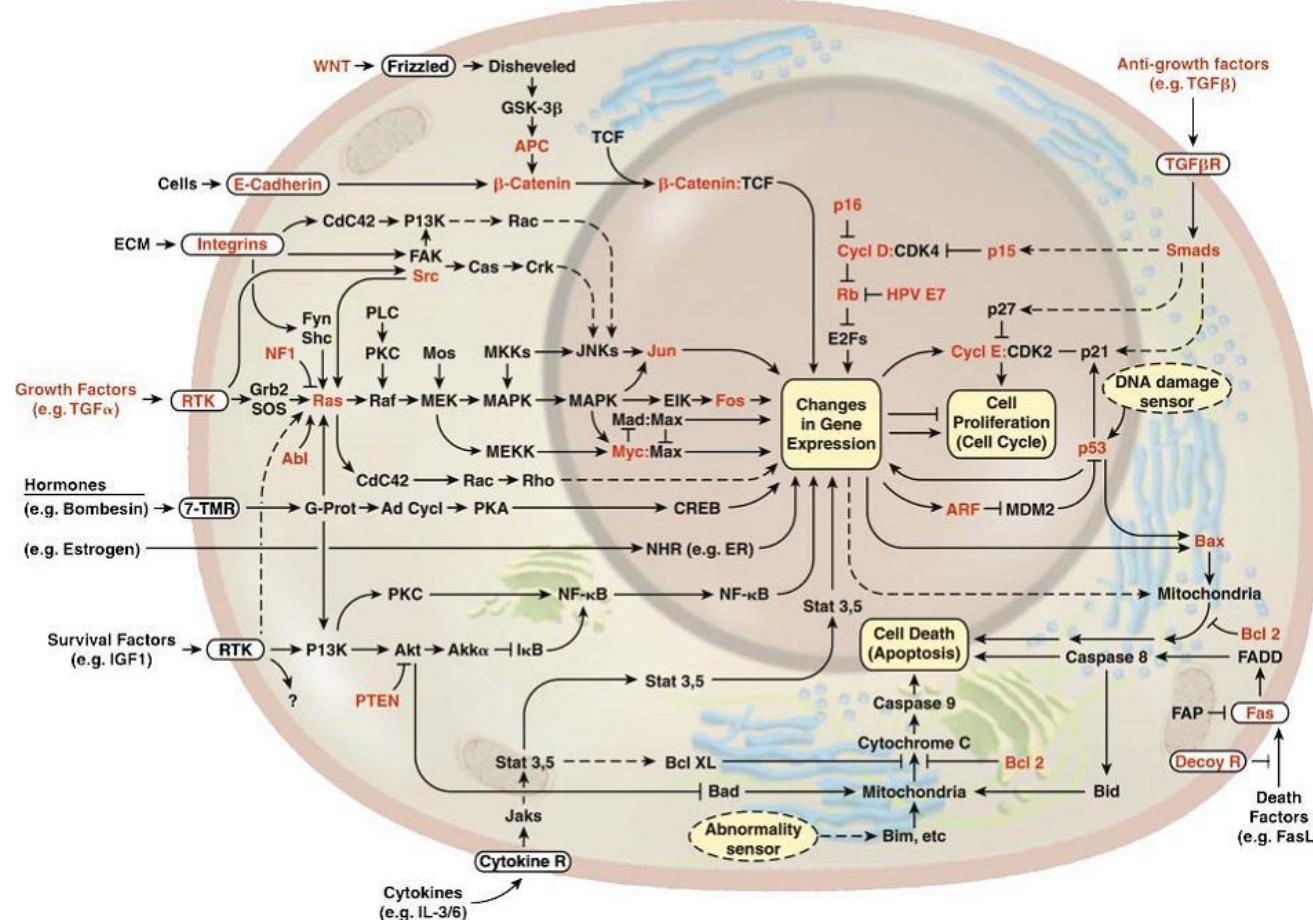
Biological Pathway

- 細胞内で生起する一連の分子間相互作用。特定の産物を產生したり、細胞に変化を引き起こしたりする。

Signal transduction pathway

Metabolic pathway

signal transduction pathway



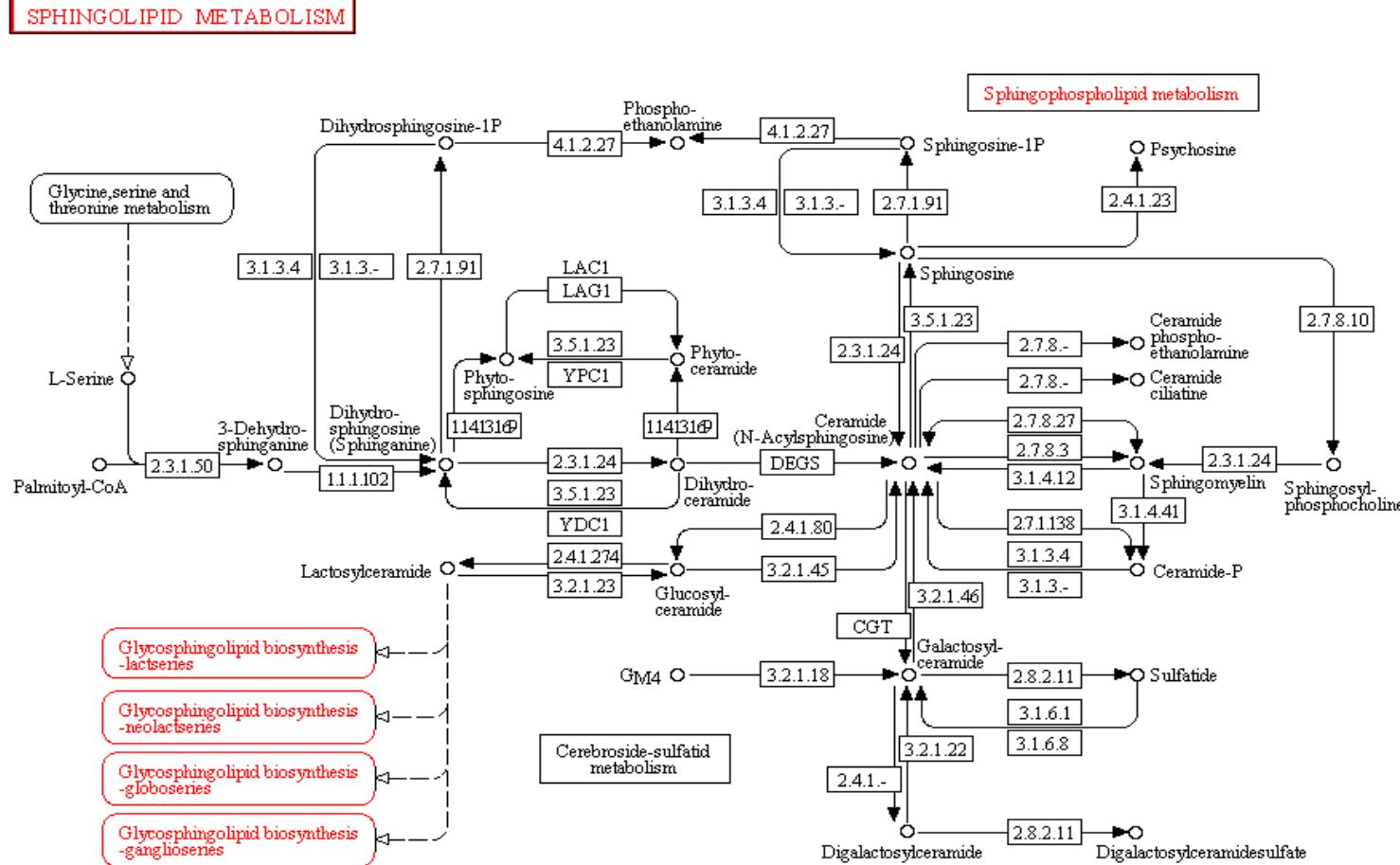
pathway database

■ 例 Kyoto Encyclopedia of Genes and Genomes (KEGG)

<http://www.genome.jp/kegg/pathway.html>

代謝やシグナル伝達など、分子間相互作用のネットワークに関する情報を統合したデータベース

Sphingolipid metabolism



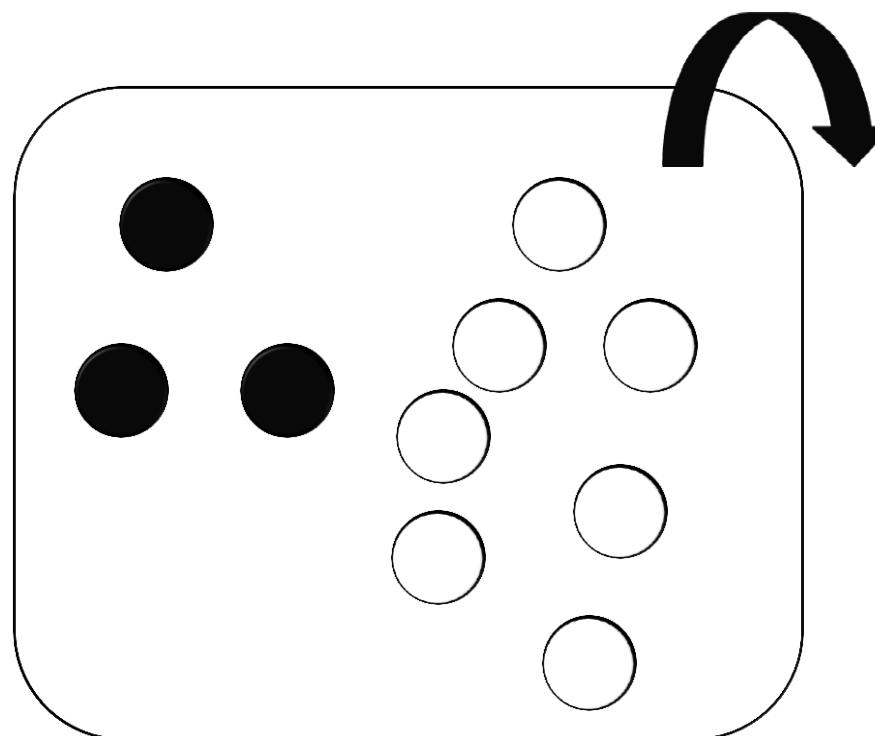
超幾何分布

$$P(X=k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$$
$$= \frac{{}_M C_k \cdot {}_{N-M} C_{n-k}}{{}_N C_n}$$

属性Aを持つM個と属性Bを持つN-M個からなる総数N個の集団から、n個のサンプルを抽出したとき、属性Aを持つものがk個存在する確率

超幾何分布

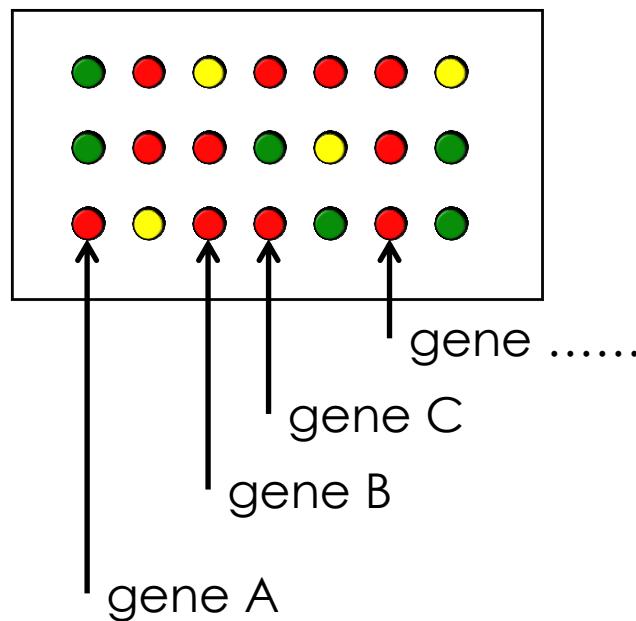
碁石100個
白70個
黒30個



10個抽出
4個黒の確率

$$\frac{\binom{30}{4} \binom{100-30}{10-4}}{\binom{100}{10}} = \frac{_{30}C_4 \cdot {}_{70}C_6}{{}_{100}C_{10}} = 0.2075777$$

GO解析



gene	GO term
gene A	GO term 1
gene B	GO term 2
gene C	GO term 1
...	...

total gene :N
annotated to specific GO term :M
interested gene :n
annotated specific GO term in list :k

実例

■ GEO database

GSE4911

Runx2 K/O microarray

WT / Runx2 KO

WT で発現の高い遺伝子上位100をSAM法で抽出

GO finder

microarray data

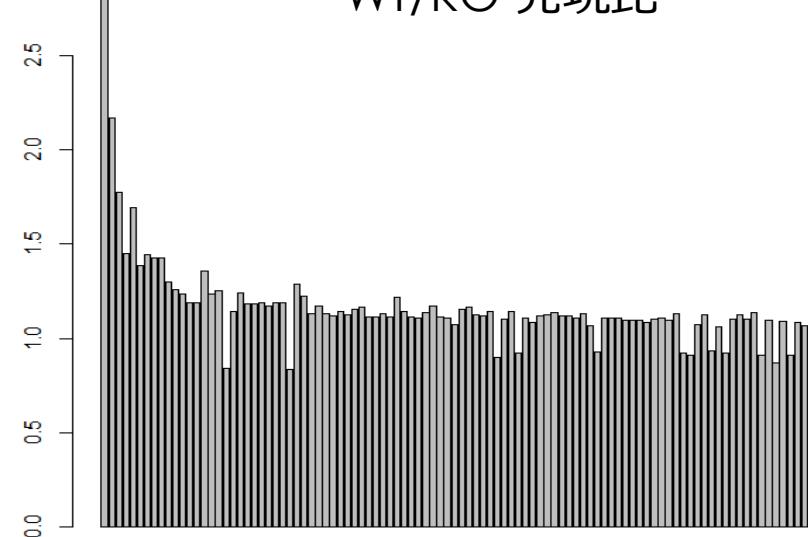
probe ID	WT sample1	WT sample2	KO sample1	KO sample2
100476_at	11.308503	12.065447	4.120034	4.222905
100484_at	9.683591	10.557372	4.668391	4.657318
...

WT / Runx2 KO



SAM法で上位100遺伝子を抽出

WT/KO 発現比



gene list

Affy mg u74a probeset	Gene Name
102925_at	Dusp9
99903_at	Dmp1
100476_at	Ibsp
97519_at	Spp1
100484_at	Mmp13
92596_at	Cacybp
98850_at	Tll1
102644_at	Mdfic
93369_at	Omd
...	...

GOTermFinder

<http://go.princeton.edu/cgi-bin/GOTermFinder>

Required Basic Input Options [Help](#)

1. Enter a list of genes, one per line. [SGD sample gene list](#)

list
gene name

OR

Upload a file containing lists of genes: no file selected

(10 MB max, plain text, one per line, gz ok)

This version of the Generic GO Term Finder has a batch mode to process multiple gene lists in parallel. To use this feature, upload an archive in tar, tar.gz, tgz, or zip format. Please refer to the [help document](#) for instructions.

Batch processing is done on one of our grid systems. Processing time will depend on the grid's load.

For long running jobs, it is highly recommended that an email address be provided so that notification may be sent when results are ready. If you do not do this, you will have to keep your browser window open to monitor the progress.
A notification email address is required for batch jobs.

Please enter a notification email address:
And please enter it again for confirmation: mail address

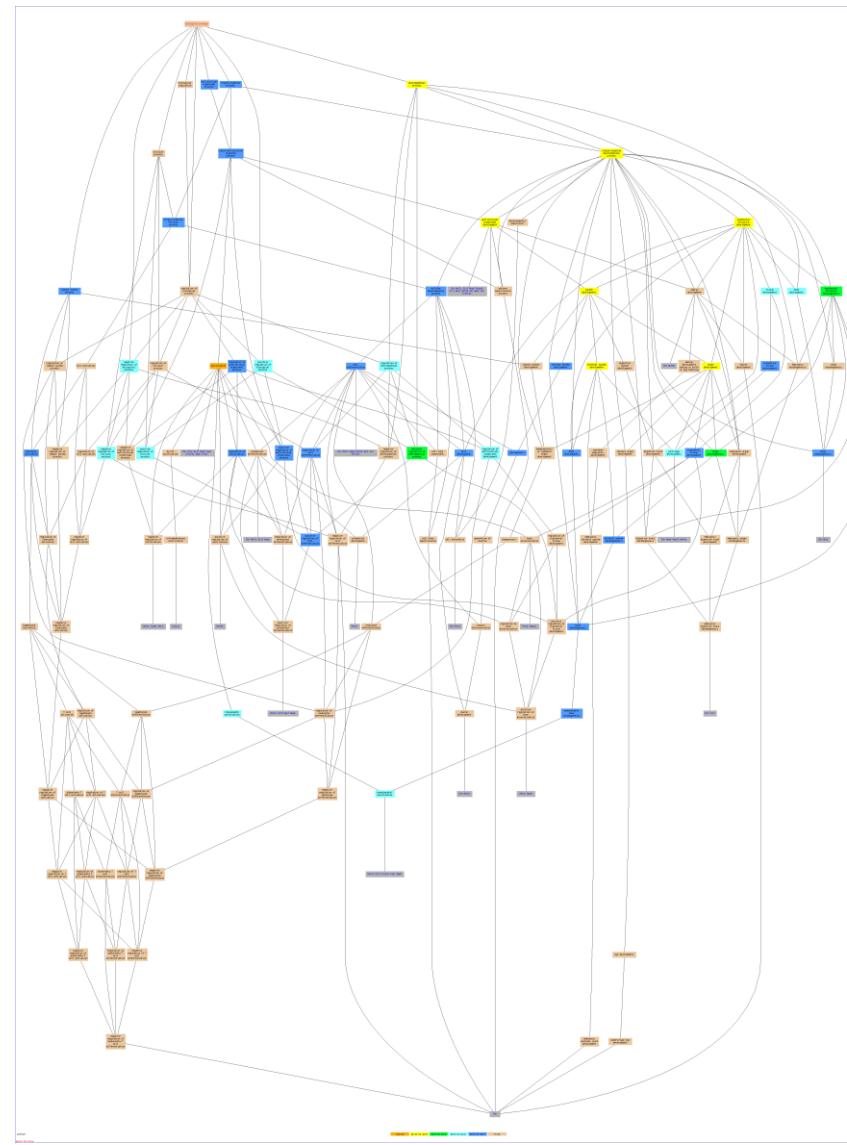
2. Choose 1 of the 3 ontology aspects: Process Function Component

3. Choose annotation: (or upload your own in the advanced options)
Example identifiers: P2ry12 MGE:1918089 P2Y12
[Longer sample list](#)

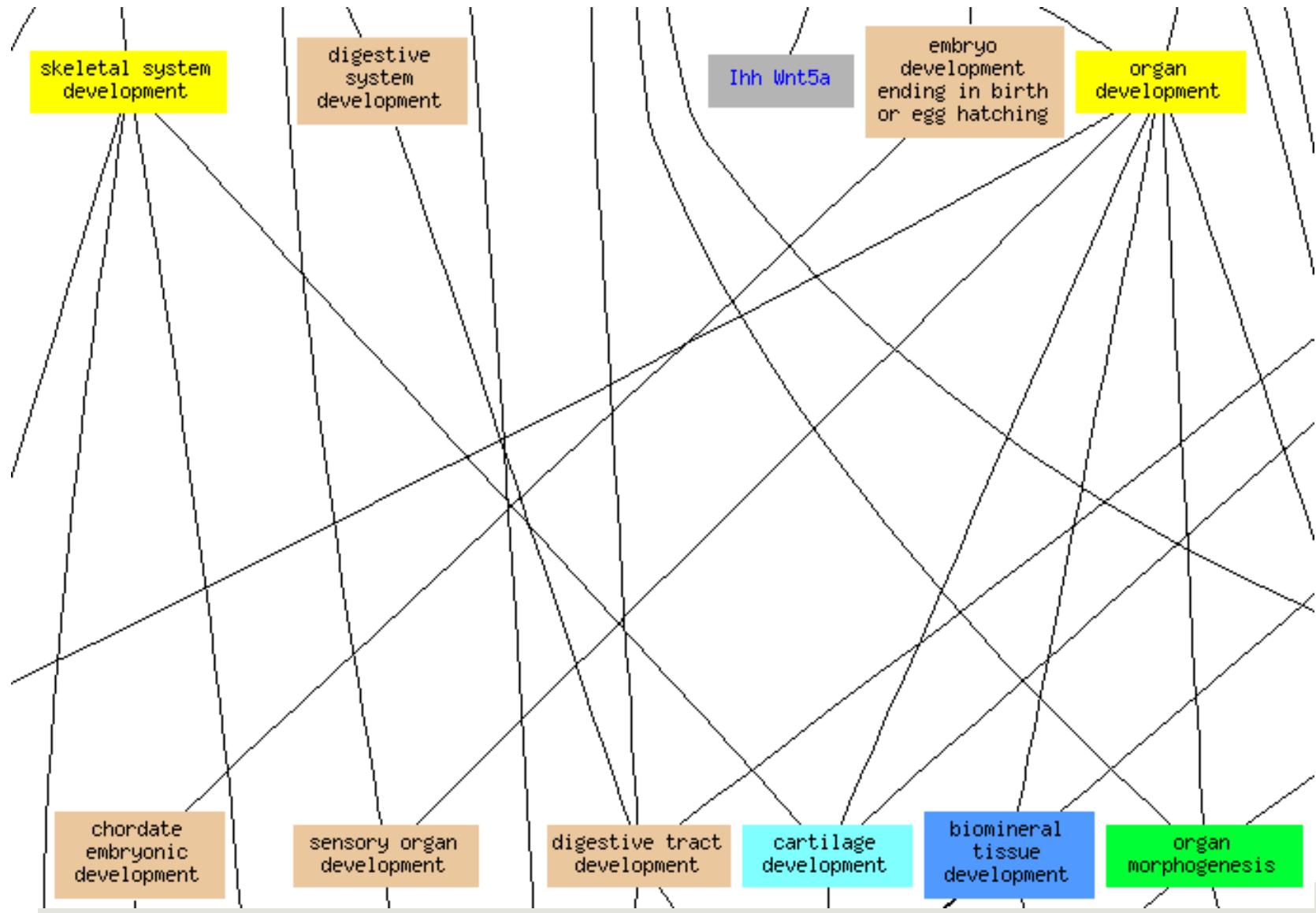
species
(mouse)

4. Plain text will be produced. Choose additional output format(s): HTML table GO tree view images

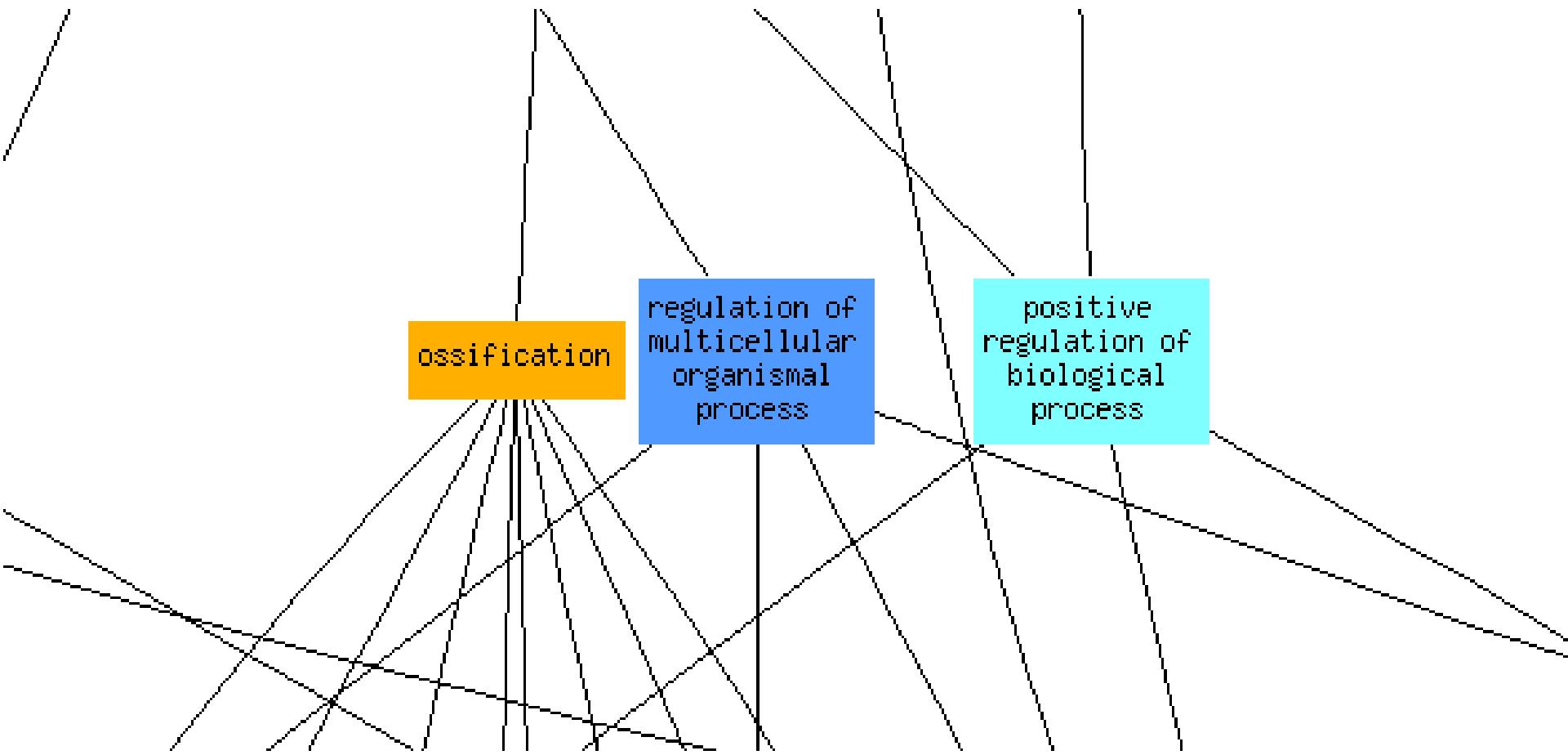
Directed Acyclic Graph



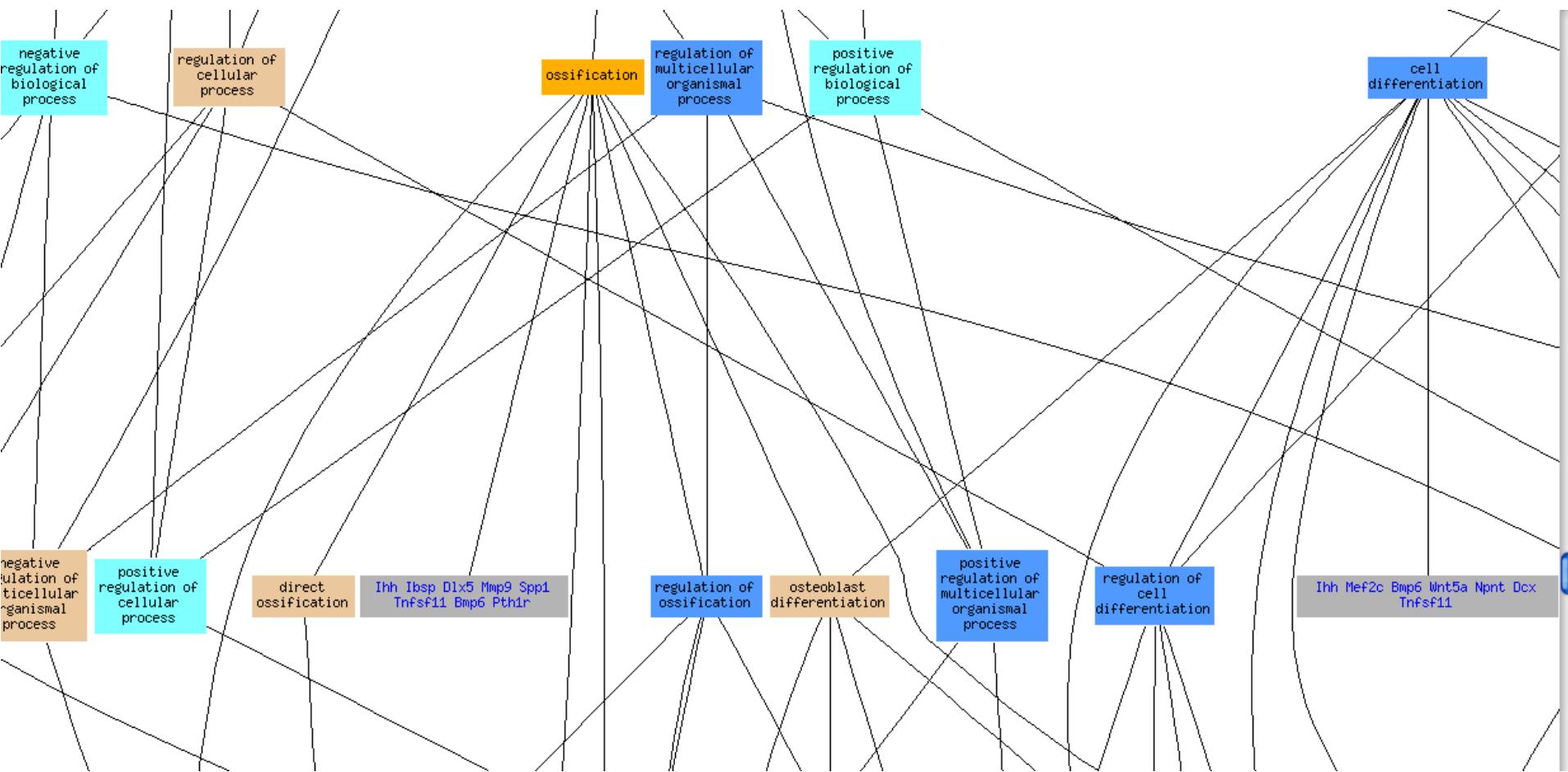
Directed Acyclic Graph



Directed Acyclic Graph



Directed Acyclic Graph



KEGG Pathway Search

KEGG Mapper – Search&Color Pathway

http://www.genome.jp/kegg/tool/map_pathway2.html

KEGG

Biomart バイオインフォマティクス ポータル query.png 7627x10203 pixels DBGET Search Result: GENES dcx KEGG Mapper – Search&Color Pa...

KEGG Mapper – Search&Color Pathway

About KEGG Mapper

Search Pathway
Search&Color Pathway
Color Pathway
Color Pathway WebGL

Search Brite
Search&Color Brite
Join Brite

Search Module
Search&Color Module

Reconstruct Pathway
Reconstruct Brite
Reconstruct Module
Map Taxonomy

KEGG Atlas
KEGG

Search against: ko Enter: map, ko, ec, rn, hsadd, or org

Primary ID: KEGG identifiers (Outside IDs for organism-specific pathways only)

Enter objects one per line followed by bgcolor, fgcolor:

16147
15891
13395
17395
211660
21943
12161
19228
114249
67880

Examples:
(Reference pathway (KO))
K01803 red,blue
C00118 pink

(Homo sapiens pathway)
7167 red,blue
C00118 pink

Alternatively, enter the file name containing the data:

Choose File no file selected

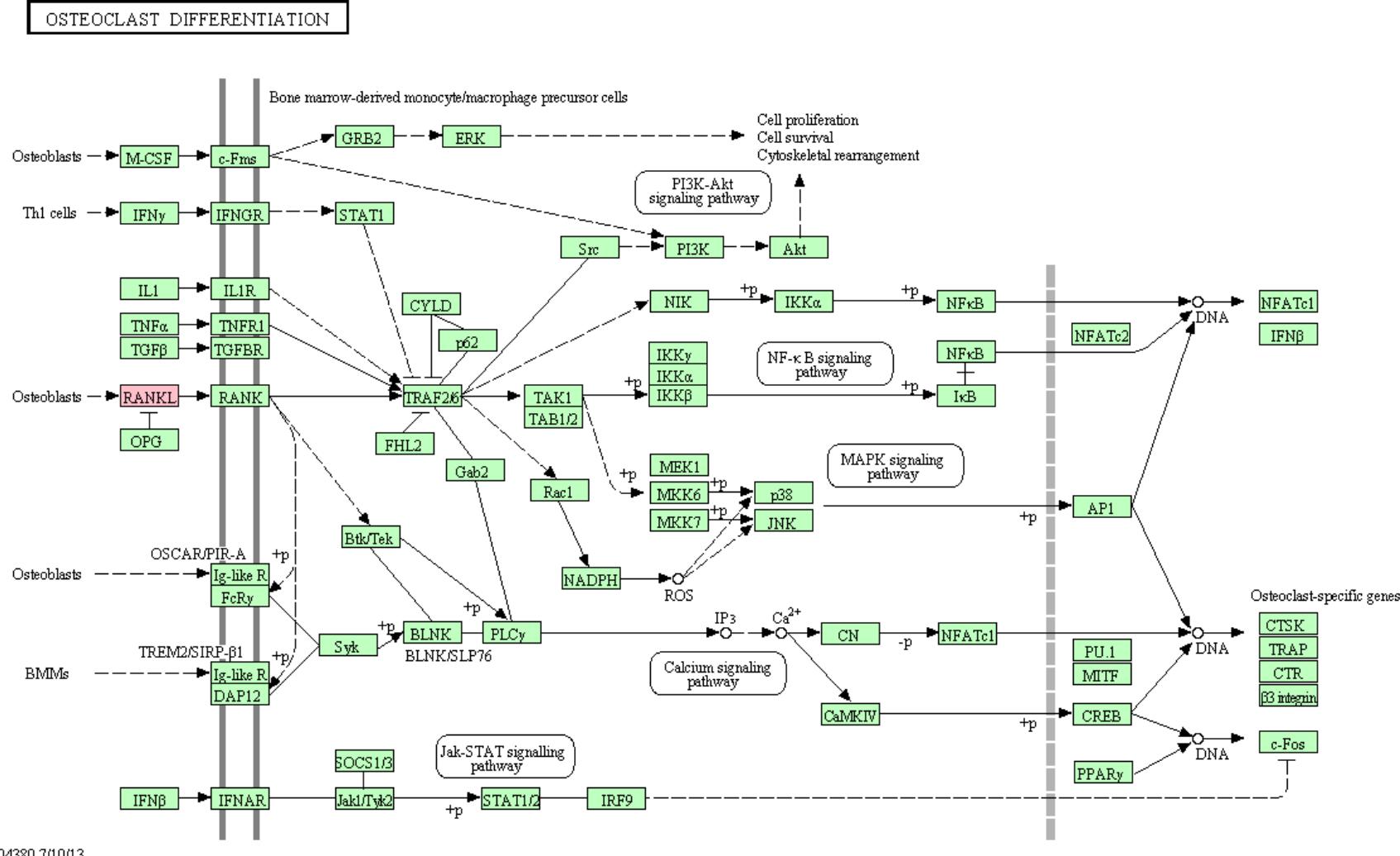
If necessary, change default bgcolor: pink

Include aliases
 Use uncolored diagrams
 Display objects not found in the search
 Search pathways containing all the objects (AND search)

Exec Clear

Search&Color Pathway is an advanced version of the KEGG pathway mapping tool, where...

KEGG pathway search



Runx2 - Osteoblast

Heparin-binding EGF-like growth factor and miR-1192 exert opposite effect on Runx2-induced osteogenic differentiation.

Yu S, Geng Q, Ma J, Sun F, Yu Y, Pan Q, Hong A.
Cell Death Dis. 2013 Oct 17;4:e868.

Histone demethylase Jmjd3 regulates osteoblast differentiation via transcription factors Runx2 and osterix.

Yang D, Okamura H, Nakashima Y, Haneji T.
J Biol Chem. 2013 Nov 22;288(47):33530-41.

Runx2 induces bone osteolysis by transcriptional suppression of TSSC1.

Wang DC, Wang HF, Yuan ZN.

Biochem Biophys Res Commun. 2013 Sep 6;438(4):635-9.

まとめ

- WT / Runx KO 発現変動遺伝子について、GO解析を行った。
- 骨形成に関するGO termが有意にエンリッチされていた。
- Runx2 の機能をGO termから推測することができた。

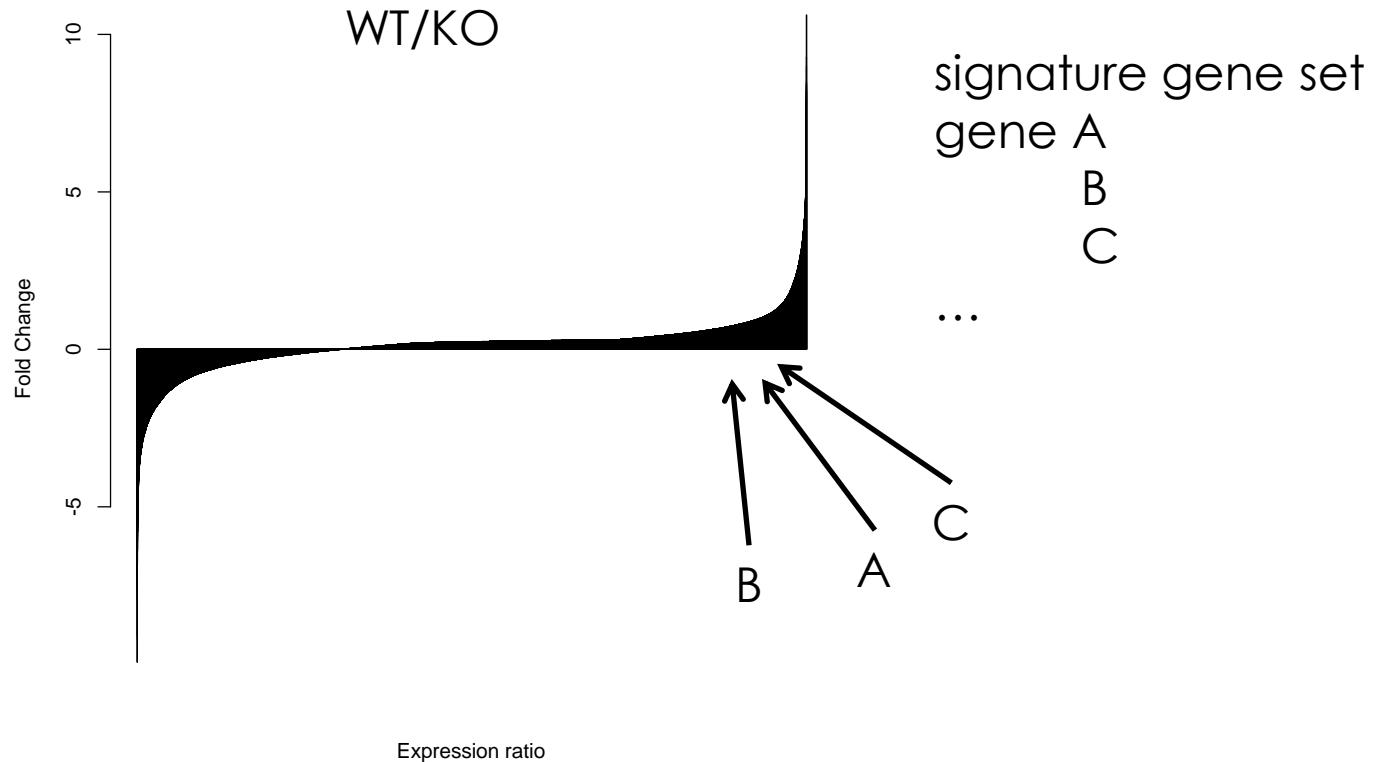
問題点

- 一つの遺伝子に複数のGOタームが定義されている
- GOタームを持たない遺伝子も存在する
- 下層のタームは必ず上層のタームと併せて検出される
- GO termは人間が定義している（主観的）

発展

□ 発現量の情報で重み付け

GSEA (Gene Set Enrichment Analysis)



Bioinformatics portal site

□ <http://genome.med.yokohama-cu.ac.jp>

bioinformatics解析サーバ

12 CPU

190GB メモリ

40TB ハードディスク

学内LANに接続された端末からアクセスが可能

- 1) ウェブブラウザ上から操作できる解析ツール
- 2) ssh ログインして使用する解析用ソフトウェア
- 3) ウェブベースのツールへのリンク

Bioinformatics portal site

<http://genome.med.yokohama-cu.ac.jp>

The screenshot shows a web browser window for the "バイオインフォマティクス ポータル" (Bioinformatics Portal) at <http://genome.med.yokohama-cu.ac.jp/biport/>. The page features a header with the university logo and navigation links. A central search bar is present, along with a banner image of laboratory equipment and a building.

Biomart search results are displayed, showing a query image (query.png 7627x10203 pixels) and a "GO解析 (知識)" (GO analysis (knowledge)) link.

バイオインフォマティクス ポータル (Bioinformatics Portal) is highlighted in red. Below it, a section titled "バイオインフォマティクス解析サーバー" (Bioinformatics Analysis Server) provides information about the server's capabilities and contact details.

メニュー (Menu) is shown on the right side.

バイオインフォマティクス 解析サーバー (Bioinformatics Analysis Server) includes links to "ウェブツール" (Web Tools) and "コマンドラインツール" (Command Line Tools).

その他 (Other) includes links to "解析ツールリンク集" (List of analysis tools) and "バイオインフォマティクス研究会" (Bioinformatics Research Society).

ウェブツール (Web Tools) section describes the server as a high-performance computer used for processing large-scale sequencing data and lists the available tools.

UCSC Genome Browser section features a screenshot of the browser interface showing genomic tracks and a brief description of its capabilities.

統合TV : 「UCSC Genome Browserの使い方 ～基本編～」 (Integrated TV: 'How to use UCSC Genome Browser - Basic Edition') is mentioned at the bottom.

Web tools

- 1) UCSC genome Browser
- 2) NIA Array Analysis
- 3) Galaxy
- 4) Blat

ウェブツール

ウェブブラウザ上で動作するバイオインフォマティクス解析ツールです。学内LANに接続された端末から、Internet ExplorerやSafariなどのウェブブラウザで聞くだけで使用することができます、かつ高機能な解析が可能です。

UCSC Genome Browser

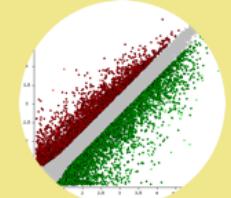
ヒト、マウスをはじめとする各動物種のゲノムデータベース。一塩基多型(SNP) やマイクロRNA(miRNA)など様々なデータベースと統合されており、それらのゲノム情報の閲覧や配列の取得に使用します。また、RNA-seqやChIP-seqなど次世代シーケンサーで解析したデータを表示することができます。本家のゲノムブラウザと異なり、アップロードした解析データを長期に保存しておくことができます。

- 統合TV：「UCSC Genome Browserの使い方～基本編」
- 各教室のアカウント

NIA Array Analysis

マイクロアレイデータの解析ツール。スキャッターブラウザの作成や二群間比較などの基本的な解析により、サンプル間で変動する遺伝子群を見つけることができます。

- General Description of NIA Array Analysis tool
- 日本語マニュアル



Galaxy

ソフトウェア & ウェブツールへのリンク

— こじて使い回す —

コマンドラインツール

バイオインフォマティクス解析サーバには、ウェブベースのツールの他にコマンドベースのソフトウェアが使用できます。

Perl、Javaといった基本的なプログラミング環境と、以下のようなバイオインフォマティクス解析用ソフトウェアがインストールされています。学内LANに接続した端末からSSH接続して、これらのソフトウェアを使用することができます。アカウントの作成を希望する方はご連絡ください。

◆ R/Bioconductor

統計解析用ソフトウェア。[Bioconductor](#)に収録されたパッケージを用いることで、マイクロアレイデータの解析やChIP-seqデータの解析など、幅広い種類の実験データを取り扱うことが可能です。

◆ Bowtie

シーケンスデータのマッピング用ソフトウェア。シーケンスされた配列をゲノム上にマッピングし、SAM形式のファイルとして出力します。出力されたSAMファイルはその後の解析に使用します。

◆ HOMER

ChIP-seqデータ解析用ソフトウェア。ピーク同定やアノテーションなどChIP-seqデータの解析に使用します。付属のツール群を使ってピーク情報の可視化なども出来ます。

◆ SICER

ChIP-seqデータ解析ソフトウェア。ピーク同定などChIP-seqデータ解析に使用します。ユーザによりパラメータを細かく設定することが可能です。

その他

◆ バイオインフォマティクス解析ツール リンク集

連絡先

nakabaya@yokohama-cu.ac.jp

バイオインフォマティクス解析室

中林潤