
遺伝子発現データの知識化

～公開レポジトリデータの活用～

2021年8月19日



Japan Science and Technology Agency

バイオサイエンスデータベースセンター(NBDC)

自己紹介

太田紀夫 (おおた としお) toshio.ota@jst.go.jp

・製薬会社で35年間、研究業務に従事。退職後、NBDCに再就職

'84～'93 抗生物質生産菌育種研究、組換え酵素の大量生産系開発

'93～'96 放線菌ゲノム解析・ゲノム育種

'96～'03 ヒト完全長cDNA解析／経産省 NEDO ヒト完全長cDNAプロジェクト

'00～'08 DNAマイクロアレイを活用した創薬研究

'08～'19 生命科学情報を活用した創薬研究・開発支援

'19～ NBDC企画運営室 (現職)

【会社員時代にやってきたこと】

- ✓ 創薬標的探索、バイオマーカー探索（データマイニング）
- ✓ 病態機序、薬効機序、毒性機序（プロファイリング・機序解析）
- ✓ 予後予測、薬効予測、毒性予測（判別分析）

【担当していた業務】

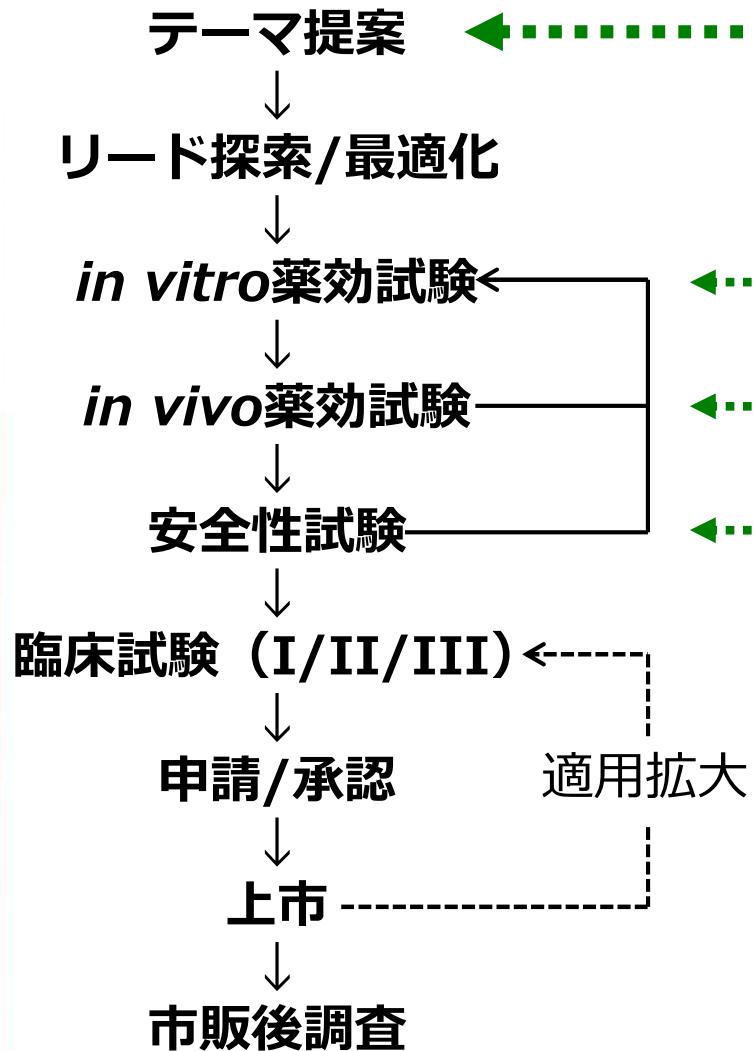
がん、免疫・アレルギー、中枢神経、腎臓、安全性、新規テーマ探索、開発支援

【使っていたデータベース・ツール等】

NCBI Gene, Blast, GEO, SRA, dbEST, UniGene, HomoloGene, PubMed, OMIM, ClinVar, GRASP, GSEA, Connectivity Map, Cell Encyclopedia, NCI 60, caArray, UniProt, Ensembl, COSMIC, ArrayExpress, BioMart, RefEx, GGRNA, GGGenome, Allie, Chip-Atlas, FANTOM, DBTSS, *GeneSpring GX*, *SpotFire*, *IPA*, *MetaCore/KPA*, *NextBio (Correlation Engine)*ほか

くすりの開発とデータ活用

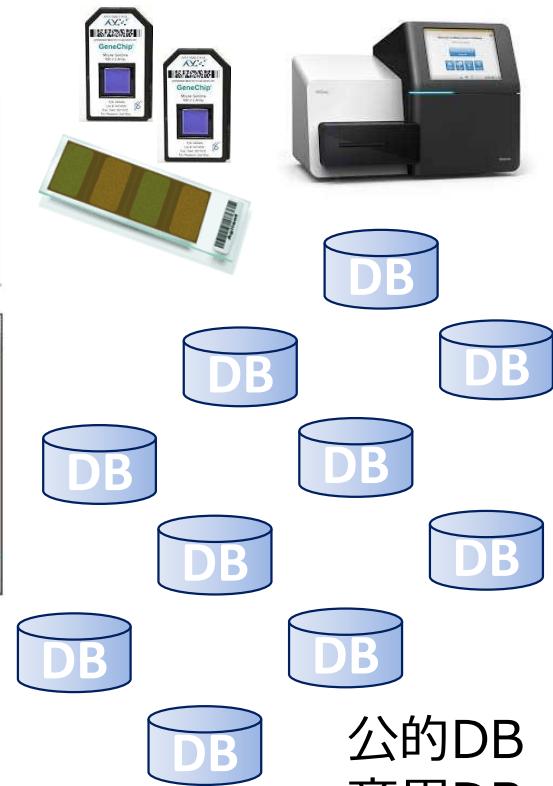
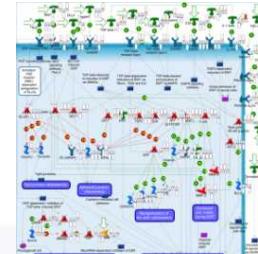
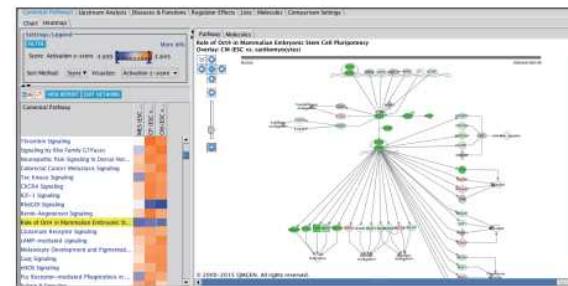
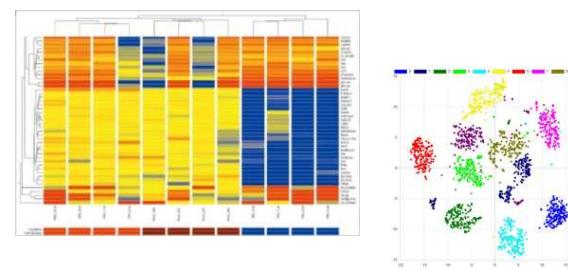
【創薬研究 – 開発の流れ】



情報調査・解析

- データマイニング
- プロファイリング
- 判別分析

自社・外注分析



バイオインフォマティシャン？

"No"

or

データサイエンティスト？

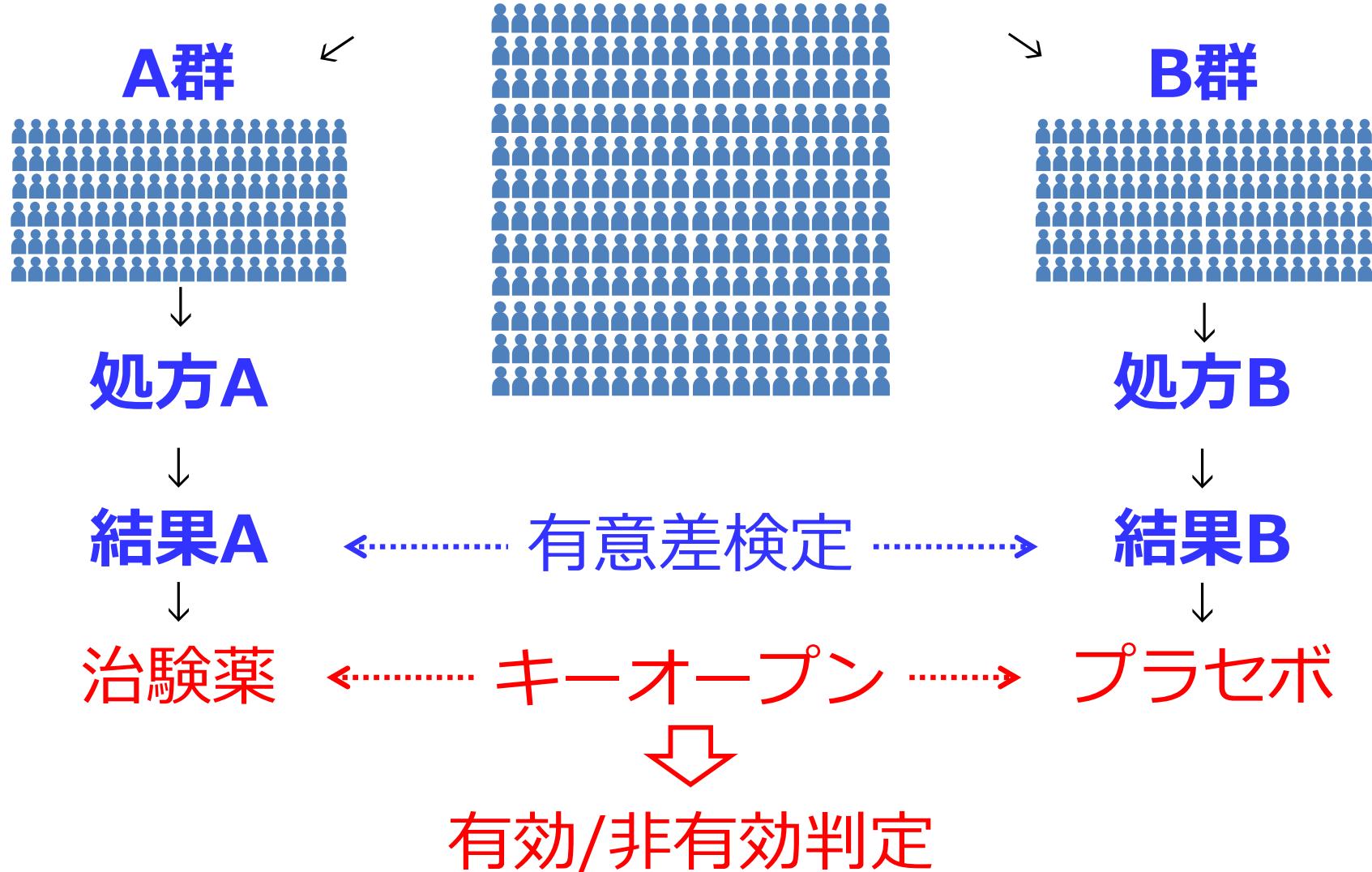
"Yes"

(たぶん…)

「統計学的有意差」

臨床試験(二重盲検比較試験)

被験者エントリー



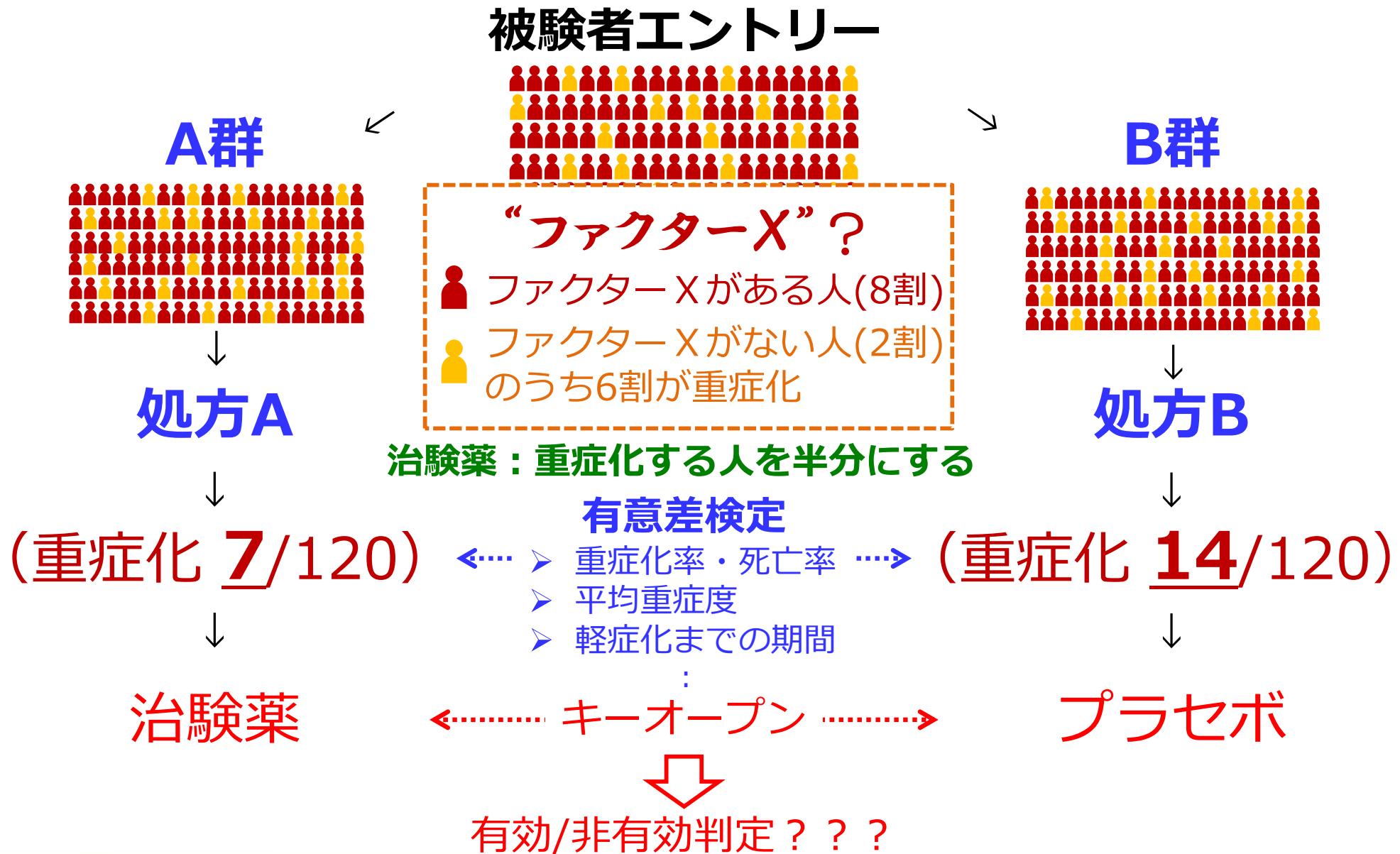
バイオインフォマティクスにおける 統計学

「統計学」はとても重要です。
だから、よく勉強して理解しましょう。



“ファクターX”の呪い

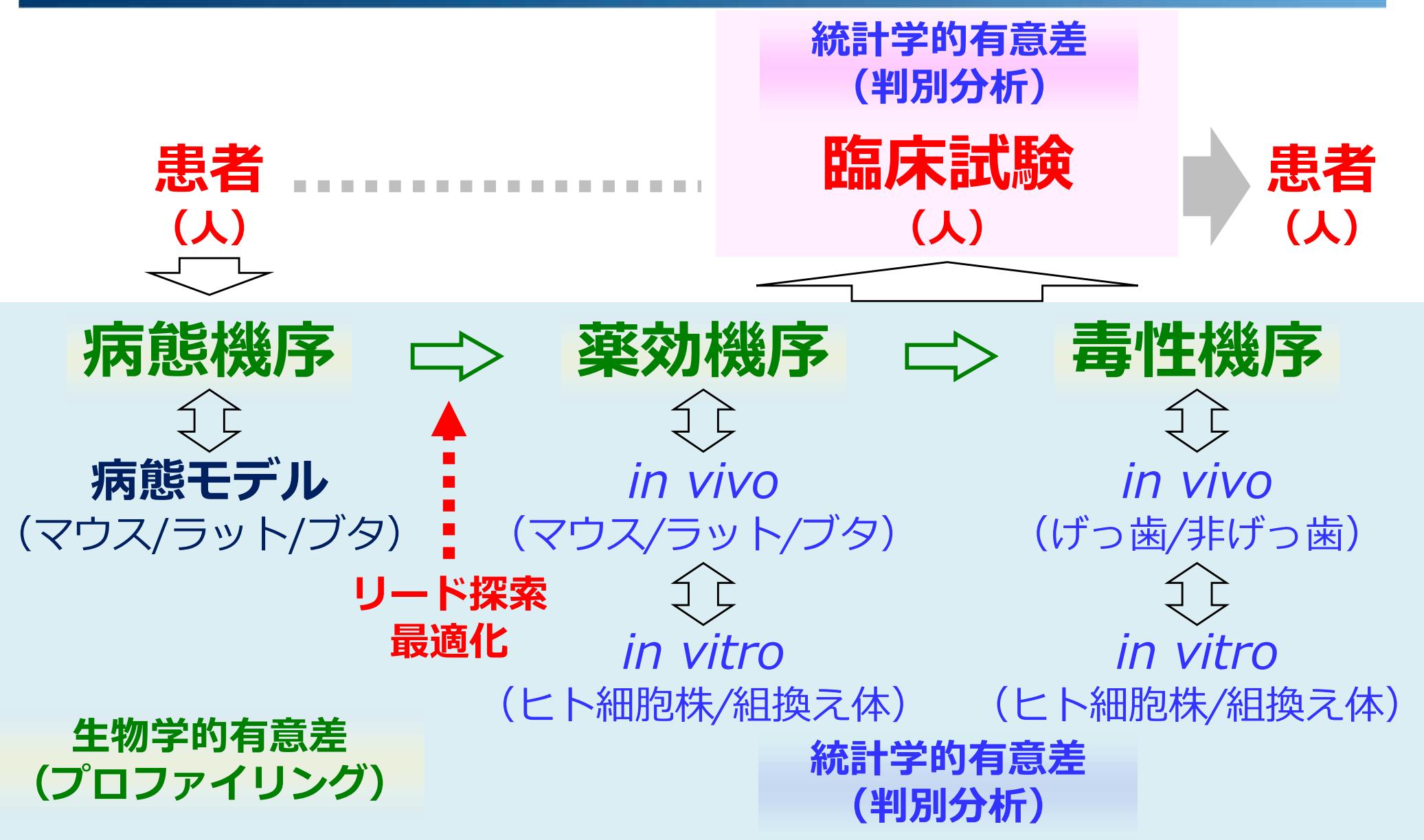
不均一な標本での有意差検定の問題



層別化マーカーを使った臨床試験



統計学的有意差と生物学的有意差



”統計学的“有意差

&

”生物学的“有意差

どちらも大切！

公共レポジトリの遺伝子発現データ

福田繁雄 「潮風公園島の日曜日の午後」



公共レポジトリの遺伝子発現データ

【気になる点】

- 衆目に晒されている使い古されたデータ。
- どうせデータを取った人がしゃぶりつくしている。
- サンプル調製や実験などの条件がまちまち。
- サンプルに関する情報が十分でないデータも多い。
- 記載されている情報が本当かどうかわからない。
- ひと昔前の技術で取られた古いデータ。
- どこの馬の骨かもわからないデータ。

⋮
⋮

所詮、「他人が取ったデータ」…

遺伝子発現データの公共レポジトリ

NCBI Gene Expression Omnibus (GEO)

(<https://www.ncbi.nlm.nih.gov/geo/>)



EBI Array Express

(<http://www.ebi.ac.uk/arrayexpress/>)



Genomic Expression Archive (GEA)

(<https://www.ddbj.nig.ac.jp/gea/index-e.html>)



【URL】 <https://www.ncbi.nlm.nih.gov/geo/>

NCBI GEOの公開データ数 (2021.08.08現在)

DataSets (GDS): 4,348 NCBIがまとめた解析単位

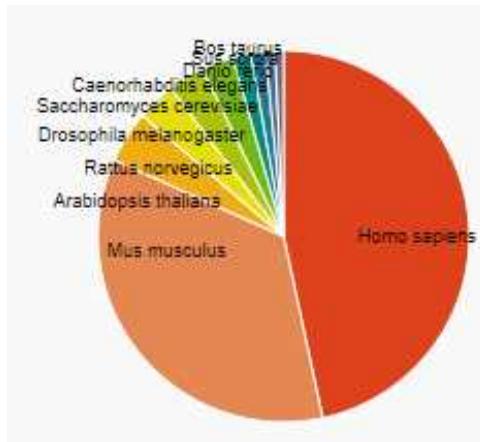
Series (GSE): 158,136 登録された実験セット

Platforms (GPL): 22,456 測定プラットフォーム

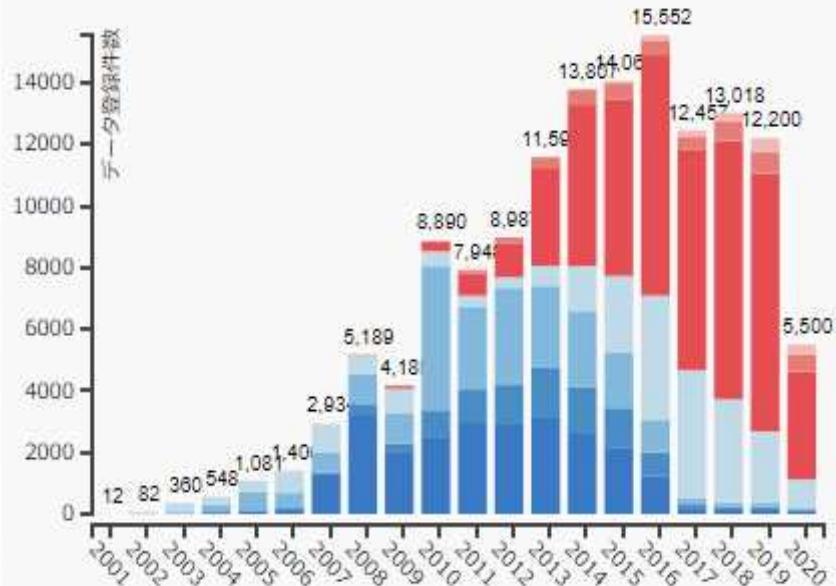
Samples (GSM): 4,560,461 サンプルごとのデータ



NCBI BioSampleの遺伝子発現データ登録数



1	<i>Homo sapiens</i>	52,665	
2	<i>Mus musculus</i>	39,674	
3	<i>Arabidopsis thaliana</i>	5,088	
4	<i>Rattus norvegicus</i>	3,808	



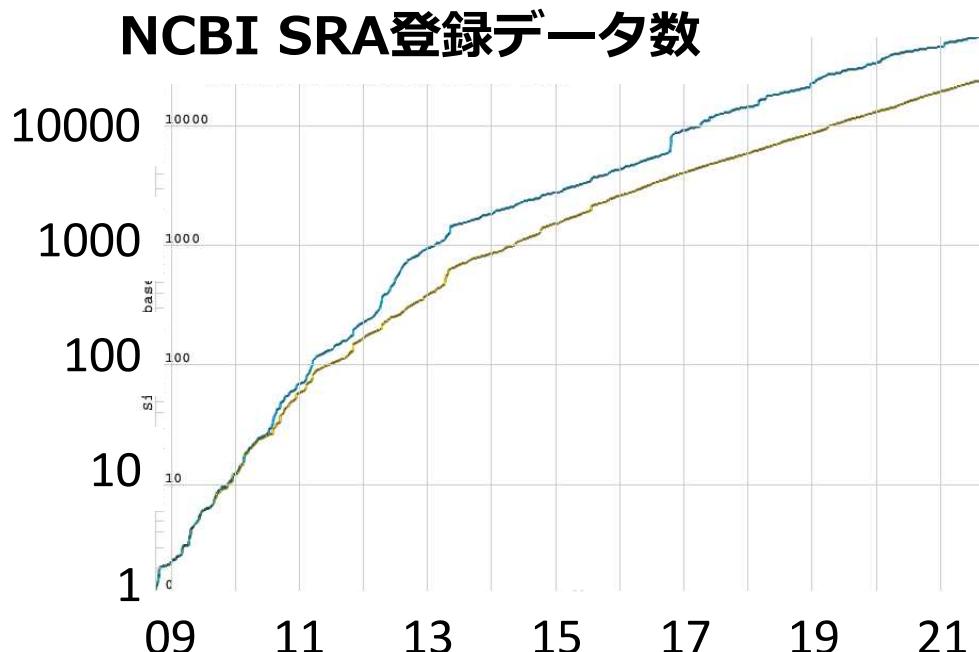
1	Illumina	51,589
2	Affymetrix	25,040
3	Unknown_Array	24,626
4	Others_Array	22,959
5	Agilent	9,429
6	Others_NGS	4,470
7	Unknown_NGS	1,702

ヒト
マウス
NGS
アレイ
半分弱
1 / 3
4割
6割

【URL】 <https://trace.ncbi.nlm.nih.gov/Traces/sra/>
NGSデータの一次レポジトリ

NCBI SRAの公開データ数 (2021.08.08現在)

Study: 324,672 登録された実験セット
Samples: 11,056,478 シーケンスデータ



ゲノム、遺伝子発現、エピゲノム解析など様々なシーケンスデータを含む。SRAはあくまで一次レポジトリなので、このサイト内でSRAに登録されたRNA-seqデータから得られた遺伝子発現情報を見るといったことはできない。一部のRNA-seqデータについては、GEOに遺伝子発現情報が登録されていることもあるが、すべてではない。



EBI Expression Atlas

【URL】 <https://www.ebi.ac.uk/gxa/home>

RNA-seqデータのビューワ。65生物種, 139,128データを公開

(2021.07.19現在)

Homo sapiens	Mus musculus	Rattus norvegicus	Drosophila melanogaster	Gallus gallus	Caenorhabditis elegans
1518 experiments	1185 experiments	152 experiments	142 experiments	39 experiments	30 experiments
Baseline: 79	Baseline: 49	Baseline: 3	Baseline: 4	Baseline: 4	Baseline: 1
Differential: 1439	Differential: 1136	Differential: 149	Differential: 138	Differential: 35	Differential: 29

【統合TV】 TO GO TV

「Expression Atlasで様々な生物種の組織や疾患などにおける遺伝子発現の情報を調べる」 (11分30秒) [2018-07-25]

<https://togotv.dbcls.jp/20180725.html>

Expression Atlas は様々な生物種における遺伝子発現情報を提供するEMBL-EBI のウェブサイト。組織や細胞種、発生段階、疾患の有無などの条件別に、遺伝子の発現情報をまとめている。データは GTEx、Cancer Cell Line Encyclopedia (CCLE)、ENCODE、FANTOM5 などのプロジェクトに由来する。全てのデータは専門家がキュレーションしている。

遺伝子名で検索し、発現している実験条件を検索する方法をはじめ、遺伝子の発現に差が見られる実験条件の検索、生物学的条件から発現している遺伝子を検索する方法、さらに、一細胞 RNA-seq の実験データを検索できるSingle Cell Expression Atlas の使い方を紹介します。



EBI Single Cell Expression Atlas

【URL】 <https://www.ebi.ac.uk/gxa/sc/home>

EBIに登録されているssRNA-seqデータのビューワー。

Single Cell Expression Atlasのデータ数 (2021.08.08現在)

Species :

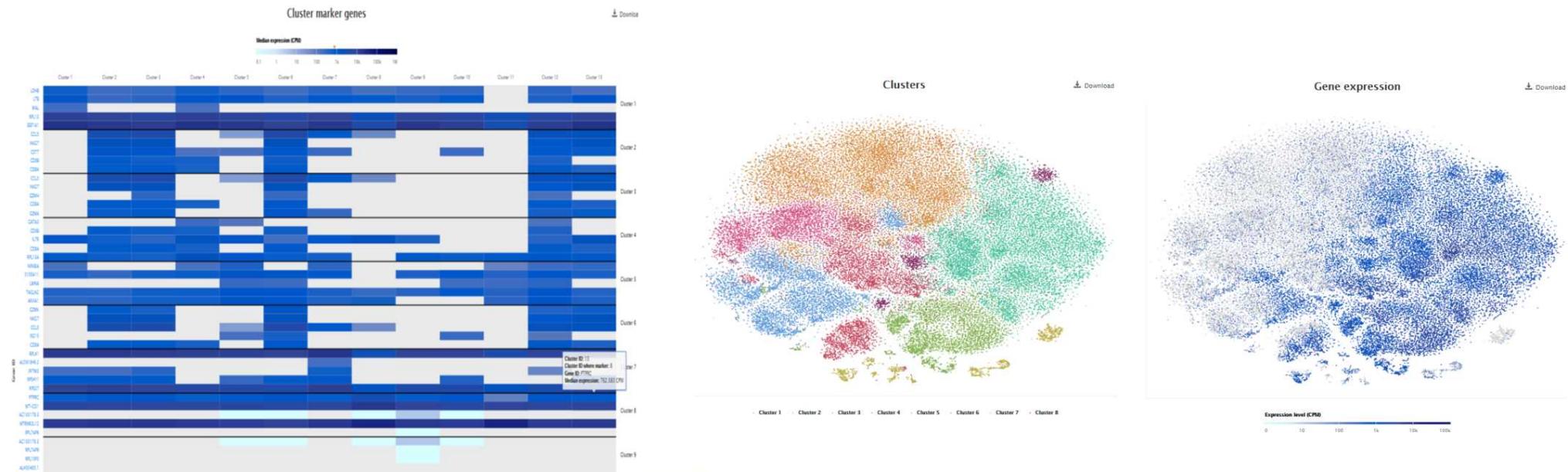
18 生物種

Study:

217 登録された実験セット

Cells:

5,312,183 細胞数

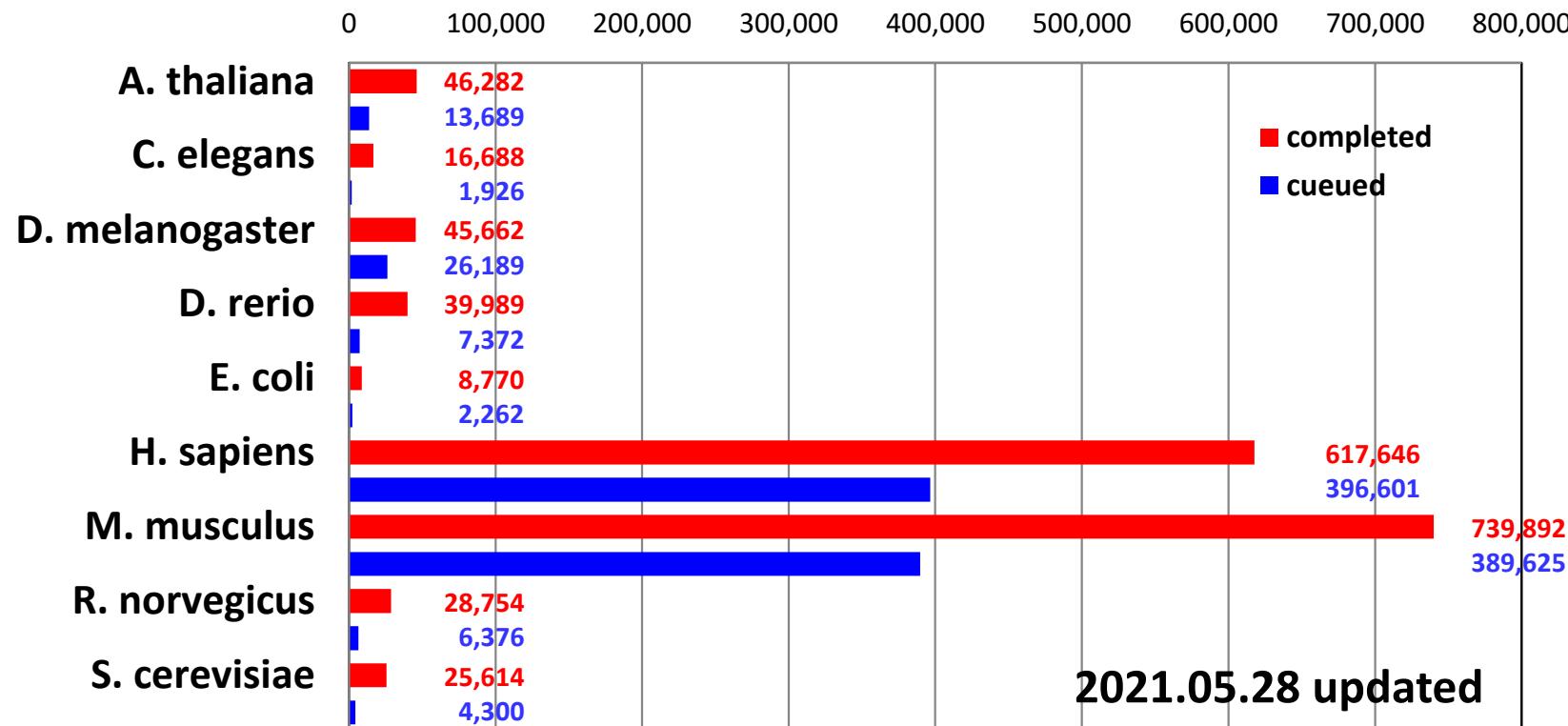


Digital Expression Explorer 2 (DEE2)

【URL】 <http://dee2.io>

NCBI SRAのRNA-seqデータを統一プロトコールで再解析したデータのレポジトリ。データごとのQCデータも提供。

9生物種, 1,569,297データを公開。 (2021.05.28現在)



複数のSeriesを組み合わせた第三者解析

Series GSE119087

Public on Aug 29, 2018

URL

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119087>

Title

Human Boolean Implication Network

Third-party reanalysis

Summary

Numerous gene expression datasets from diverse human tissue samples have been already deposited in the public domain. There have been several attempts to do large scale meta-analyses of all of these datasets. Most of these analyses summarize pairwise gene expression relationships using correlation, or identify differentially expressed genes in two conditions. We propose here a new large scale meta-analysis of all of the publicly available human datasets to identify Boolean logical relationships between genes. Boolean logic is a branch of mathematics that deals with two possible values. In the context of gene expression datasets we use qualitative high and low expression values. A strong logical relationship between genes emerges if at least one of the quadrants is sparsely populated.

Overall design

25,955 published human microarray samples assayed on the GPL570 were re-analyzed. RMA was used to normalize the RAW CEL files all together.

Contributor(s)

Sahoo D

Citation(s)

Dabydeen SA, et al., Unbiased Boolean analysis of public gene expression data for **cell cycle gene identification**. Mol Biol Cell 2019 Jul 1;30(14):1770-1779. PMID: 31091168

Dang D, et al., Computational Approach to Identifying **Universal Macrophage Biomarkers**. Front Physiol 2020;11:275. PMID: 32322218

Organization

UCSD

公共レポジトリの遺伝子発現データの特徴

【良い点】

- 膨大な数のデータがある。
- 多様な種類のサンプルのデータがある。
- 簡単には入手できないサンプルのデータもある。
- さまざまな測定手法で取られたデータがある。
- 異なる施設やラボで取られたデータがある。
- すぐに解析できる（迅速性）。
- 無償で利用できる（タダ！ \$）

：

実は、**誰でも使える「宝の山」！**

公共レポジトリの遺伝子発現データを取り扱う際の留意点

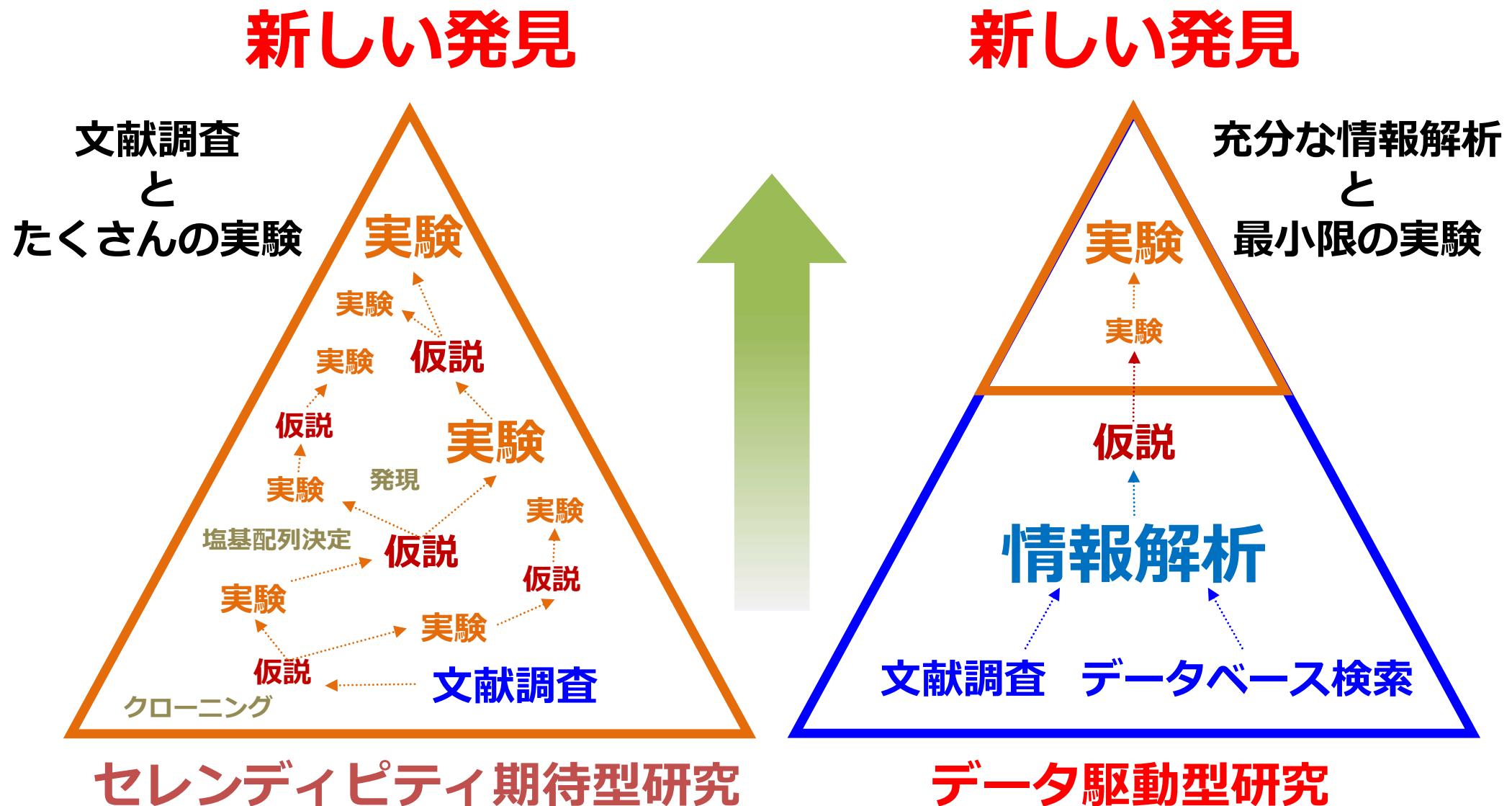
【取り扱うデータについて理解を深める】

- ✓ 対応論文があればよく読みこむ (Supplementaryも含めて)。
- ✓ **サンプルの由来・調製法**などに注意を払う (**サンプルのQC**)。
- ✓ **測定手法・計測方法**にも注意を払う。
- ✓ **定量方法や正規化方法**にも注意を払う。

【データの利用にあたって】

1. 必要に応じ、適切なプロトコールで**再正規化**を行う。
2. **信頼できるデータ区間**を見極める (**データのQC**)。
3. **適切な閾値**を設定する (**データのQC**)。
4. **最適な解析手法**を選ぶ。
5. できる限り、複数のデータやほかの論文でも**裏を取る**。

「データ駆動型研究」



[URL] <https://togotv.dbcls.jp/>



DBCLS 小野浩雅
特任助教

The screenshot shows the homepage of TOGO TV, featuring a search bar and a navigation menu. A large yellow arrow points from the text below to a specific video thumbnail titled "公共の遺伝子発現データの検索や解析を行う" (Search and analyze public gene expression data).

「公共の遺伝子発現データの検索や解析を行う」

- NCBI GEOの使い方1 ~マイクロアレイデータの検索・取得~2017
- NCBI GEO の使い方2 ~遺伝子プロファイルの検索・処理済みデータの取得~ 2018
- NCBI GEOのデータセットブラウザを使って公共データの遺伝子発現解析を行う 2019
- AOEを使って遺伝子発現データベースの統計を見ながら検索する 2018
- RefExの使い方
- OmicsDI を使って様々なデータベースからオミクスデータを縦断的に検索する
- Expression Atlas で 様々な生物種の組織や疾患などにおける遺伝子発現の情報を調べる
- IMOTA を使って、組織別に分類された様々なマルチオミクスデータ (miRNA/mRNA/protein)を調べる
- Spotfireを用いた公共マイクロアレイデータとローカルなデータの統合 2019
- Bgee を使って複数の生物種の正常組織における遺伝子発現データを検索、比較、取得する
- MGIを使ってマウスの遺伝子発現情報を調べる 2019
- MGIを使ってノックアウトマウスの情報を調べる 2018
- GTEx Portalを使ってヒトの各組織での遺伝子発現量や影響するeQTLを調べる
- UCSC VisiGeneを使って生体内におけるmRNAの局在を調べる 2017
- The Human Protein Atlasでヒトのタンパク質の発現情報をRNA-seqデータ、画像データとともに調べる 2017
- Allen Brain Atlasを使い倒す ~基本編~ 2017
- Arabidopsis eFP Browser でシロイヌナズナの遺伝子発現情報を見る
- Human Brain Transcriptomeを使ってヒトの脳の発達に関する時空間トランскriptomを見る
- NCBI GEOのGEO2Rを使って公開されているマイクロアレイデータを解析する
- 疾患に関連するバリエントや遺伝子発現の情報を調べる
- ⋮

同じデータでも視点を変えると違うものが見えてくる!!



福田繁雄 「潮風公園島の日曜日の午後」

福田繁雄
「潮風公園島の日曜日の午後」

スーラ
「ラ・グランド・ジャッド島の日曜日の午後」

遺伝子発現情報の 「知識化」 について考える

本日のデモデータ [1]

GSE21422

非侵襲性乳管がん (DCIS) と侵襲性乳管がん (IDC)

URL : <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21422>

非侵襲性乳管がん (DCIS)・侵襲性乳管がん (IDC)

Series GSE21422

URL	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21422
Title	Expression profiling of human DCIS and invasive ductal breast carcinoma
Organism	Homo sapiens
Overall design	Expression profiling by array
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570
BioProject	PRJNA126373 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA126373

【関連論文】

Cell Oncol. (2011) 34:419–434

Latent transforming growth factor binding protein 4 (LTBP4) is downregulated in mouse and human DCIS and mammary carcinomas

Celine Kretschmer *et al.* (Charité Berlin, CVK, Med. Klinik Hepatologie & Gastroenterologie)

【解析サンプル】

Healthy breast	5 サンプル
Ductal carcinoma in situ (DCIS)	9 サンプル
Invasive ductal carcinoma (IDC)	5 サンプル

【要旨】

TGF- β は上皮細胞の増殖を阻害し、乳腺腫瘍の発がんに関与する。また、3種類のLTBPがTGF- β の機能の調節に関与していることが知られている。

ヒトとマウスの非浸潤性乳管がん (DCIS) におけるLTBP4とそのアイソフォームLTBP1、LTBP3、そしてTGF β 1、TGF β 2、TGF β 3、SMAD2、SMAD3、SMAD4の発現を調べ、浸潤性乳管がん (IDC) と比較した。さらに乳がん細胞株 (MCF7、Hs578T、MDA-MB361) と良性乳腺細胞株 (Hs578BsT) の発現を調べた。マイクロアレイ、q-PCR、イムノプロット、免疫組織化学、免疫蛍光抗体法を用いた。

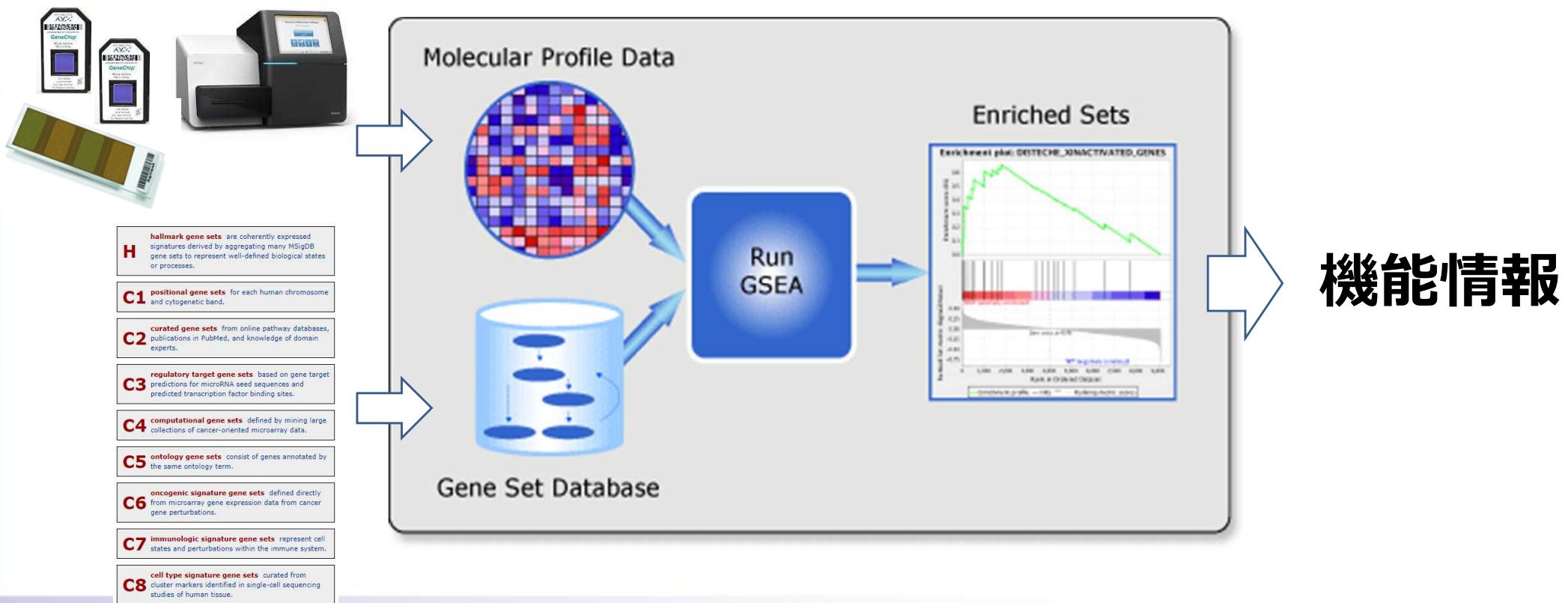
非がん組織 ($n=5$) と比較し、調べた全てのWAP-TNP8マウスとヒトのDCIS ($n=9$)、およびIDC ($n=5$)において、**LTBP4の発現低下**とともに、**BMP4の発現が低下**し、またその阻害因子である**GREM1の発現が増加**していることを見出した。さらに、乳がん細胞株 (Hs578T) を組換えTGF β 1で処理したところ、BMP4は上昇しGREM1は低下した。

悪性乳腺腫瘍組織では、LTBP4を介したターゲッティングが欠如することにより、TGF β 1とBMPのバイオアベイラビリティと機能の変化が引き起こされている可能性が考えられた。

GSEA

(Gene Set Enrichment Analysis)

<https://www.gsea-msigdb.org/gsea/index.jsp>



【クエリ】アレイデータ/RNA-seqデータ (gct/clsファイル)

【得られる結果】どのような機能が亢進/減衰しているか

MSigDB

- H** **hallmark gene sets** (biological states or processes).
- C1** **positional gene sets** (human chromosome and cytogenetic band).
- C2** **curated gene sets** (pathway databases in PubMed and knowledge of domain experts).
- C3** **regulatory target gene sets** (microRNA target seed sequences and predicted transcription factor binding sites).
- C4** **computational gene sets** (large collections of cancer-oriented microarray data).
- C5** **ontology gene sets** (annotated by the same ontology term).
- C6** **oncogenic signature gene sets** (cancer gene perturbations).
- C7** **immunologic signature gene sets** (cell states and perturbations within the immune system).
- C8** **cell type signature gene sets** (cluster markers identified in single-cell studies of human tissue).

GSEAを使った解析事例 (GSE21422 乳管がん)

C2: "curated gene sets"

Gene sets enriched in phenotype DCIS (9 samples), top 10

Rank	GS follow link to MSigDB	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME CONDENSATION OF PROPHASE CHROMOSOMES	58	0.80	1.83	0.000	0.230	0.125	2945	tags=52%, list=14%, signal=60%
2	REACTOME PRC2 METHYLATES HISTONES AND DNA	57	0.80	1.82	0.000	0.147	0.152	2390	tags=47%, list=11%, signal=53%
3	WP_FBXL10_ENHANCEMENT_OF_MAPERK_SIGNALING_IN_DIFFUSE_LARGE_BCELL_LYMPHOMA	28	0.71	1.81	0.000	0.133	0.190	1650	tags=36%, list=8%, signal=39%
4	REACTOME AMYLOID FIBER FORMATION	94	0.58	1.81	0.021	0.101	0.191	2390	tags=30%, list=11%, signal=33%
5	REACTOME MEIOTIC RECOMBINATION	71	0.73	1.80	0.000	0.093	0.214	2902	tags=39%, list=14%, signal=46%
6	REACTOME SIRT1 NEGATIVELY REGULATES_RRNA_EXPRESSION	52	0.78	1.79	0.006	0.101	0.278	3040	tags=44%, list=15%, signal=52%
7	REACTOME PROCESSING_OF_DNA_DOUBLE_STRAAND_BREAK_ENDS	90	0.68	1.78	0.000	0.092	0.293	2902	tags=46%, list=14%, signal=53%
8	REACTOME ACTIVATED_PKN1 STIMULATES TRANSCRIPTION_OF_AR_AND_AR_GEN RECEPTOR REGULATED GENES_KLK2 AND_KLK3	51	0.76	1.78	0.010	0.082	0.298	2390	tags=39%, list=11%, signal=44%
9	REACTOME DNA METHYLATION	49	0.84	1.78	0.004	0.083	0.325	2390	tags=49%, list=11%, signal=55%
10	REACTOME OXIDATIVE_STRESS_INDUCED_SENESCENCE	108	0.56	1.77	0.000	0.089	0.367	2390	tags=32%, list=11%, signal=36%

Gene sets enriched in phenotype IDC (5 samples), top 10

Rank	GS follow link to MSigDB	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	PID_AURORA_B_PATHWAY	39	0.68	1.83	0.000	0.161	0.122	1162	tags=46%, list=6%, signal=49%
2	JEON_SMAD6_TARGETS_DN	19	0.73	1.79	0.000	0.189	0.204	1022	tags=42%, list=5%, signal=44%
3	NAKAYAMA_SOFT_TISSUE_TUMORS_PCA2_UP	89	0.75	1.74	0.000	0.342	0.439	1093	tags=49%, list=5%, signal=52%
4	SU_TESTIS	80	0.63	1.74	0.000	0.264	0.443	2301	tags=31%, list=11%, signal=35%
5	LEE_EARLY_T_LYMPHOCYTE_UP	109	0.77	1.73	0.000	0.242	0.455	2268	tags=61%, list=11%, signal=69%
6	REACTOME tRNA MODIFICATION IN THE NUCLEUS AND CYTOSOL	41	0.63	1.73	0.000	0.204	0.455	2900	tags=44%, list=14%, signal=51%
7	CHEMNITZ_RESPONSE_TO_PROSTAGLANDIN_E2_UP	147	0.60	1.73	0.000	0.179	0.459	1422	tags=42%, list=7%, signal=45%
8	REACTOME PROTEIN ubiquitination	77	0.59	1.72	0.010	0.171	0.481	3210	tags=45%, list=15%, signal=54%
9	FERREIRA_EWINGS_SARCOMA_UNSTABLE_VS_STABLE_UP	161	0.66	1.72	0.000	0.161	0.500	2211	tags=48%, list=11%, signal=54%
10	ODONNELL_TFRC_TARGETS_DN	141	0.67	1.71	0.000	0.187	0.580	2739	tags=47%, list=13%, signal=54%

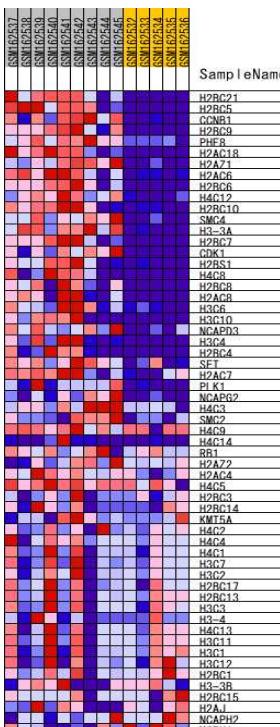
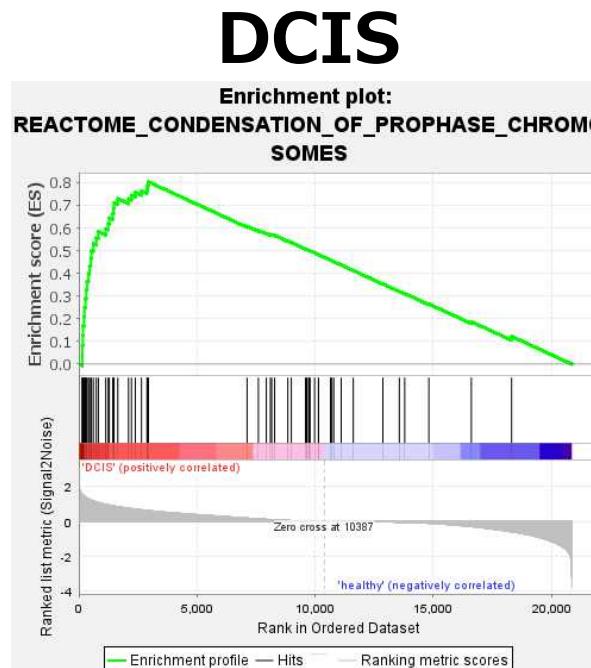
GSEAを使った解析事例（GSE21422 乳管がん）

C2: "curated gene sets"

【DCISで亢進していた機能】

"REACTOME_CONDENSATION_OF_PROPHASE_CHROMOSOMES"

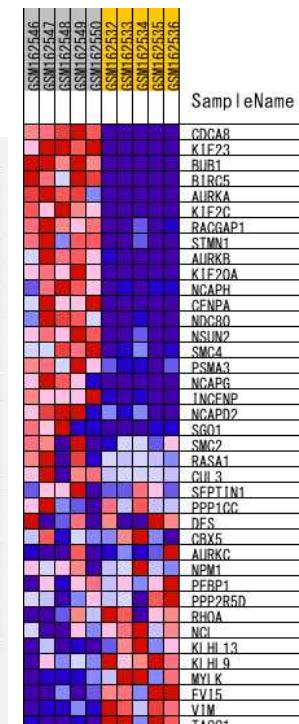
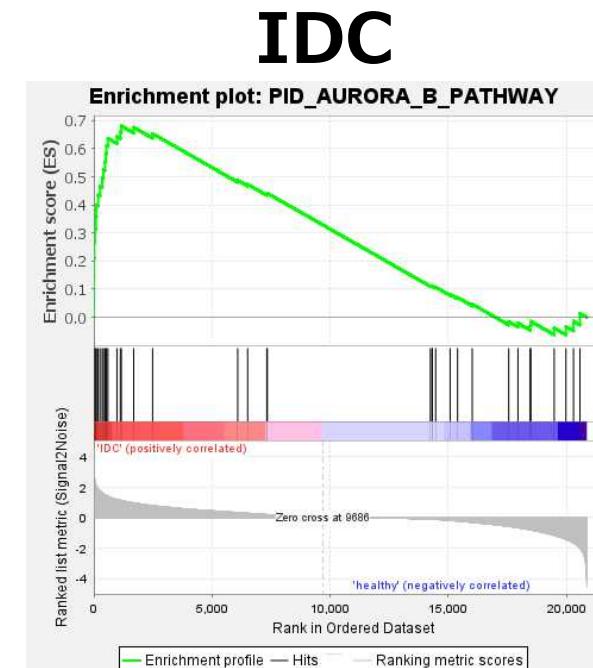
Group	Rank	ES	NES	NON P
DCIS	1	0.80	1.83	0.000
IDC	329	0.72	1.47	0.058



【IDCで亢進していた機能】

"PID_AURORA_B_PATHWAY"

Group	Rank	ES	NES	NON P
DCIS	377	0.57	1.39	0.154
IDC	1	0.68	1.83	0.000



【統合TV】GSEAの使い方はこちらへ

1) GSEA softwareの使い方 基本編 (9分56秒) [2010-07-23]

<https://togotv.dbcls.jp/20100723.html>

GSEA softwareは米国Broad Instituteによって提供されているGSEAを行うソフトウェアです。Gene Set Enrichment Analysis (GSEA) は、予め用意した遺伝子セットが異なる条件下でどう振舞うかを調べる手法です。これを利用し発現プロファイルを解析することができます。詳しいアルゴリズムは、[GeneSet Enrichment Analysis PNAS paper \(pdf\)](#)を参照してください。今回は、GSEA softwareのGUI版の導入と簡単な解析を行い、GSEA softwareで何ができるかを示します。

2) GSEA softwareの使い方 発展編 (8分57秒) [2010-08-30]

<https://togotv.dbcls.jp/20100830.html>

NCBI GEOより取得した公共の遺伝子発現データ (GSE1657:Adipocyte Differentiation Homo sapiens]のSeries Matrix Files) を表計算ソフトを使い加工し、GSEA softwareに読み込ませ、解析を行う手順を解説します。

3) GSEA software を使って遺伝子リストのエンリッチメント解析を行う

(8分52秒) [2019-04-01]

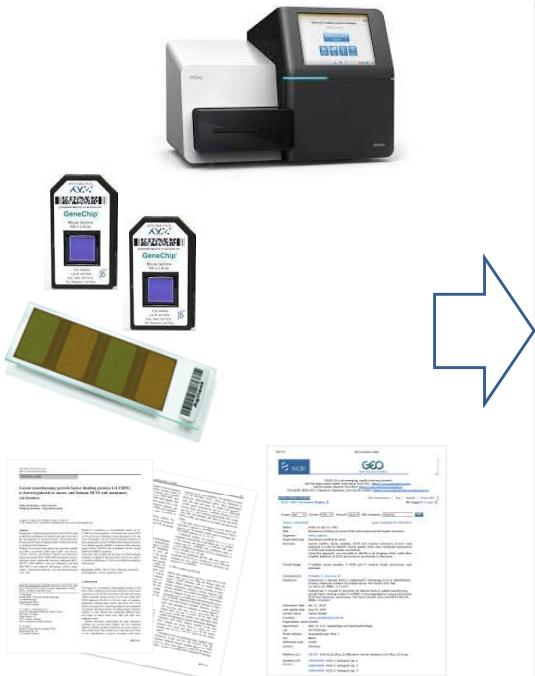
<https://togotv.dbcls.jp/20190401.html>

GSEA softwareのGUI版の導入と設定方法の解説とともにデモデータを用いた簡単な解析を行いながら、GSEA softwareで何ができるかを示します。



DAVID

(The Database for Annotation, Visualization and Integrated Discovery)
[\(https://david.ncifcrf.gov/home.jsp/\)](https://david.ncifcrf.gov/home.jsp/)



A screenshot of the DAVID Bioinformatics Resources 6.8 homepage. The header reads "DAVID Bioinformatics Resources 6.8" and "Laboratory of Human Retrovirology and Immunoinformatics (LHRI)". The menu includes Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, About DAVID, and About LHRI.

Overview

The Database for Annotation, Visualization and Integrated Discovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
 - List interacting proteins
 - Explore gene names in batch
 - Link gene-disease associations
 - Highlight protein functional domains and motifs
 - Redirect to related literatures
 - Convert gene identifiers from one type to another.
 - And more

Hot Links

Postdoctoral Fellow position available in LHRI

The Laboratory of Human Retrovirology and Immunoinformatics (LHRI) has collaborated with the National Institute of Allergy and Infectious Diseases (NIAID) and supported NIAID clinical trials for patients infected with HIV mutants resisting anti-retroviral therapy. LHRI has isolated the multiple-class drug-resistant (MDR) variants from patients and characterized each variant's drug sensitivity and infectivity. The study aims to define salvage therapy and develop novel therapy (chemotherapy and immunotherapy). During the investigation, LHRI has characterized the emergence of novel mutations on drug susceptibility and viral replication. LHRI is a pioneer in researching the anti-viral cytokine, Interleukin-27, DNA-repair protein (Ku70)-mediated innate immune response against HIV and other virus co-infection, and novel subsets of immune cells. LHRI maintains the Database for Annotation, Visualization and Integrated Discovery (**DAVID**). **Postdoctoral Fellow position** available to perform Microbiology/Cellular Immunology research in our **Basic Research Section**.

Call for papers

Submit papers for a Special Issue of **Frontiers in Immunology: "IL-27 in Health and Disease"**



機能情報

Upload List Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -
 Homo sapiens(2090)
 Unknown(104)

[Select Species](#)

[List Manager](#) [Help](#)

IDC_UpDown 4 of 6 (2)

Select List to:

[Use](#) [Rename](#)
[Remove](#) [Combine](#)
[Show Gene List](#)

[View Unmapped Ids](#)

Annotation Summary Results

Current Gene List: IDC_UpDown 4 of 6 (2194) 1537 DAVID IDs
 Current Background: Homo sapiens Check Defaults Clear All

[Help and Tool Manual](#)

Disease (1 selected)			
<input type="checkbox"/> GAD_DISEASE	75.5%	1160	Chart
<input type="checkbox"/> GAD_DISEASE_CLASS	75.5%	1160	Chart
<input checked="" type="checkbox"/> OMIM_DISEASE	24.6%	378	Chart

Functional Categories (3 selected)			
<input checked="" type="checkbox"/> COG_ONTOLOGY	9.2%	142	Chart
<input type="checkbox"/> PIR_SEQ_FEATURE	14.2%	218	Chart
<input type="checkbox"/> SP_COMMENT_TYPE	93.8%	1442	Chart
<input checked="" type="checkbox"/> UP_KEYWORDS	96.0%	1475	Chart
<input checked="" type="checkbox"/> UP_SEQ_FEATURE	95.1%	1462	Chart

Gene Ontology (3 selected)			
<input type="checkbox"/> GOTERM_BP_1	88.1%	1354	Chart
<input type="checkbox"/> GOTERM_BP_2	87.8%	1350	Chart
<input type="checkbox"/> GOTERM_BP_3	87.2%	1341	Chart
<input type="checkbox"/> GOTERM_BP_4	86.1%	1323	Chart
<input type="checkbox"/> GOTERM_BP_5	84.8%	1303	Chart
<input type="checkbox"/> GOTERM_BP_ALL	88.1%	1354	Chart
<input checked="" type="checkbox"/> GOTERM_BP_DIRECT	88.1%	1354	Chart
<input type="checkbox"/> GOTERM_BP_FAT	87.2%	1341	Chart
<input type="checkbox"/> GOTERM_CC_1	92.6%	1424	Chart
<input type="checkbox"/> GOTERM_CC_2	91.5%	1407	Chart
<input type="checkbox"/> GOTERM_CC_3	91.5%	1406	Chart
<input type="checkbox"/> GOTERM_CC_4	89.9%	1382	Chart
<input type="checkbox"/> GOTERM_CC_5	84.8%	1303	Chart
<input type="checkbox"/> GOTERM_CC_ALL	92.6%	1424	Chart

DAVIDを使った解析事例 (GSE21422 乳管がん)

【IDCで亢進していた機能】

Annotation Cluster 1		Enrichment Score: 10.4		
		Count	P_Value	Benjamini
<input type="checkbox"/>	UP_SEQ_FEATURE	signal peptide	RT	366 1.2E-17 4.5E-14
<input type="checkbox"/>	UP_KEYWORDS	Secreted	RT	228 8.7E-14 1.3E-11
<input type="checkbox"/>	UP_KEYWORDS	Disulfide bond	RT	340 5.7E-11 4.2E-9
<input type="checkbox"/>	UP_KEYWORDS	Signal	RT	398 6.6E-11 4.3E-9
<input type="checkbox"/>	UP_KEYWORDS	Glycoprotein	RT	420 2.1E-9 1.1E-7
<input type="checkbox"/>	UP_SEQ_FEATURE	disulfide bond	RT	292 3.7E-9 7.1E-6
<input type="checkbox"/>	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT	377 5.7E-6 5.5E-3
Annotation Cluster 2		Enrichment Score: 5.35		
<input type="checkbox"/>	UP_KEYWORDS	Cell division	RT	58 1.7E-7 5.3E-6
<input type="checkbox"/>	UP_KEYWORDS	Cell cycle	RT	82 6.9E-7 1.9E-5
<input type="checkbox"/>	UP_KEYWORDS	Mitosis	RT	43 7.1E-7 1.9E-5
<input type="checkbox"/>	GOTERM_BP_DIRECT	cell division	RT	51 5.1E-5 2.0E-2
<input type="checkbox"/>	GOTERM_BP_DIRECT	mitotic nuclear division	RT	37 4.0E-4 5.8E-2
Annotation Cluster 3		Enrichment Score: 5.26		
<input type="checkbox"/>	GOTERM_CC_DIRECT	bicellular tight junction	RT	29 2.9E-8 2.4E-6
<input type="checkbox"/>	UP_KEYWORDS	Tight junction	RT	21 2.1E-6 4.6E-5
<input type="checkbox"/>	KEGG_PATHWAY	Tight junction	RT	18 2.7E-3 6.0E-2
Annotation Cluster 4		Enrichment Score: 4.34		
<input type="checkbox"/>	UP_KEYWORDS	Oxidoreductase	RT	74 1.9E-6 4.6E-5
<input type="checkbox"/>	GOTERM_MF_DIRECT	oxidoreductase activity	RT	32 1.8E-4 2.5E-2
<input type="checkbox"/>	GOTERM_BP_DIRECT	oxidation-reduction process	RT	73 2.7E-4 5.6E-2
Annotation Cluster 5		Enrichment Score: 3.97		
<input type="checkbox"/>	KEGG_PATHWAY	Focal adhesion	RT	41 5.8E-6 7.9E-4
<input type="checkbox"/>	KEGG_PATHWAY	ECM-receptor interaction	RT	22 3.9E-5 3.5E-3
<input type="checkbox"/>	KEGG_PATHWAY	PI3K-Akt signaling pathway	RT	48 5.5E-3 7.4E-2
Annotation Cluster 6		Enrichment Score: 3.65		
<input type="checkbox"/>	INTERPRO	Pleckstrin homology-like domain	RT	58 1.7E-5 8.4E-3
<input type="checkbox"/>	INTERPRO	Pleckstrin homology domain	RT	39 1.5E-4 3.7E-2
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:PH	RT	33 9.4E-4 1.9E-1
<input type="checkbox"/>	SMART	PH	RT	39 1.1E-3 1.3E-1
Annotation Cluster 7		Enrichment Score: 3.2		
<input type="checkbox"/>	INTERPRO	Insulin-like growth factor binding protein, N-terminal	RT	31 1.2E-7 2.1E-4
<input type="checkbox"/>	INTERPRO	Epidermal growth factor-like domain	RT	42 2.1E-7 2.1E-4
<input type="checkbox"/>	SMART	EGF	RT	38 1.3E-6 5.0E-4
<input type="checkbox"/>	INTERPRO	EGF-like conserved site	RT	35 5.4E-6 3.6E-3
<input type="checkbox"/>	UP_KEYWORDS	EGF-like domain	RT	37 1.6E-5 2.8E-4
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 4; calcium-binding	RT	12 1.1E-5 1.3E-2
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 1	RT	23 1.1E-5 1.7E-2
<input type="checkbox"/>	INTERPRO	EGF-like calcium-binding	RT	24 1.1E-5 2.3E-2
<input type="checkbox"/>	INTERPRO	Complement C1-like EGF domain	RT	10 1.5E-4 3.7E-2
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 2; calcium-binding	RT	14 2.2E-4 7.7E-2
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 3	RT	16 3.3E-4 1.1E-1
<input type="checkbox"/>	SMART	EGF_CA	RT	17 3.1E-4 1.0E-2
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 3; calcium-binding	RT	18 1.3E-3 1.8E-1
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 5; calcium-binding	RT	19 1.2E-3 1.2E-1
<input type="checkbox"/>	INTERPRO	EGF-like calcium-binding conserved site	RT	18 1.3E-3 1.8E-1
<input type="checkbox"/>	INTERPRO	EGF-type aspartate/asparagine hydroxylation site	RT	15 4.9E-3 4.8E-1
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 2	RT	7 9.6E-3 5.2E-1
<input type="checkbox"/>	UP_SEQ_FEATURE	domain:EGF-like 6; calcium-binding	RT	

Annotation Cluster 1	
<input type="checkbox"/>	UP_SEQ_FEATURE
<input type="checkbox"/>	UP_KEYWORDS
<input type="checkbox"/>	UP_KEYWORDS
<input type="checkbox"/>	UP_KEYWORDS
<input type="checkbox"/>	UP_SEQ_FEATURE
<input type="checkbox"/>	UP_SEQ_FEATURE

Annotation Cluster 1	
signal peptide	Enrichment Score: 10.4
Secreted	糖タンパク質
Disulfide bond	
Signal	
Glycoprotein	
disulfide bond	
glycosylation site:N-linked (GlcNAc...)	

Annotation Cluster 2	
<input type="checkbox"/>	UP_KEYWORDS
<input type="checkbox"/>	UP_KEYWORDS
<input type="checkbox"/>	UP_KEYWORDS
<input type="checkbox"/>	GOTERM_BP_DIRECT
<input type="checkbox"/>	cell division
<input type="checkbox"/>	mitotic nuclear division

Annotation Cluster 2	
Cell division	Enrichment Score: 5.35
Cell cycle	
Mitosis	
cell division	
mitotic nuclear division	

Annotation Cluster 3	
<input type="checkbox"/>	GOTERM_CC_DIRECT
<input type="checkbox"/>	bicellular tight junction
<input type="checkbox"/>	Tight junction
<input type="checkbox"/>	Tight junction

Annotation Cluster 3	
bicellular tight junction	Enrichment Score: 5.26
Tight junction	
Tight junction	

タイトジャンクション

ヒットした機能情報を
クラスター化して結果表示

【統合TV】DAVIDの使い方はこちらへ

1) DAVIDを使ってマイクロアレイデータを解析する (7分40秒) [2009-09-25]

<https://togotv.dbcls.jp/20090925.html>

DAVIDはマイクロアレイ実験から得られたデータを解析するツールです。このツールを使うことで発現変動のあった遺伝子群の特徴を可視化し、直感的に分析することができます。DAVIDという名前はThe Database for Annotation, Visualization and Integrated Discoveryの頭文字に由来しています。ムービーではサンプルデータとして、NCBI GEOに登録されている公共の遺伝子発現データ (GSE17913:Effects of Cigarette Smoke on the Human Oral Mucosal Transcriptome) からGEO2Rを用いて取得した、喫煙者女性と非喫煙者女性の口腔内で発現に差のある250個の遺伝子群のリストを使って説明しています。

2) DAVIDを使ってマイクロアレイデータを解析する2012 (6分56秒) [2012-09-27]

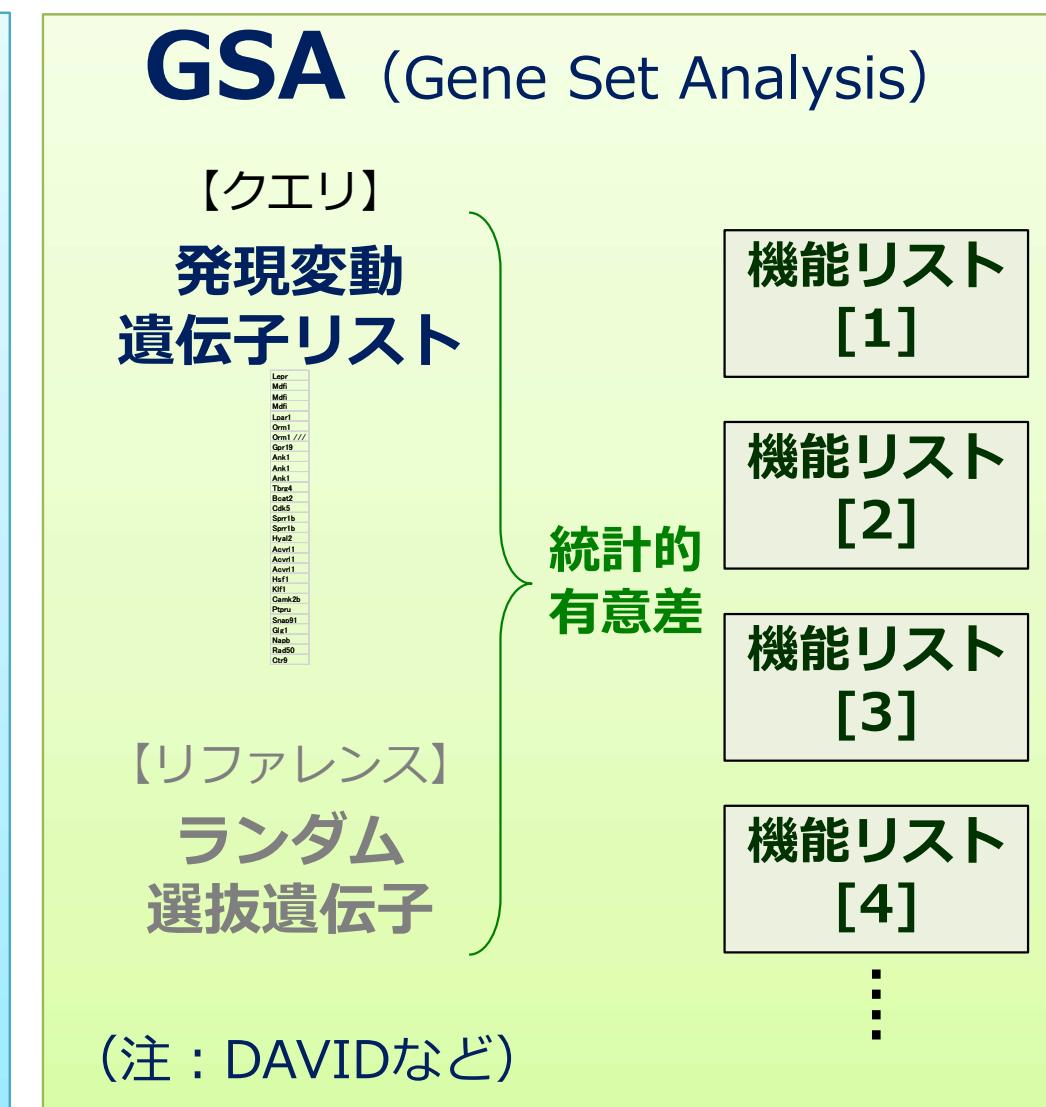
<https://togotv.dbcls.jp/20120927.html>

3) DAVIDの使い方 実践編 (5分42秒) [2013-05-28]

<https://togotv.dbcls.jp/20090925.html>

「DAVID の使い方 実践編」と題して、DAVID による GO 解析、KEGG パスウェイ解析と、Gene ID Conversion Tool の解説をしています。動画内では、NCBI GEO に登録されているデータ(GSE18121: Gene expression regulation in response to heat stress in different yeast strains)から得た遺伝子リストを用いています。酵母に熱ショックを 30分間与えた後で発現が上昇した遺伝子約 2000 個を、GEO2R を用いて取得したものです。

GSEAとGSAの比較

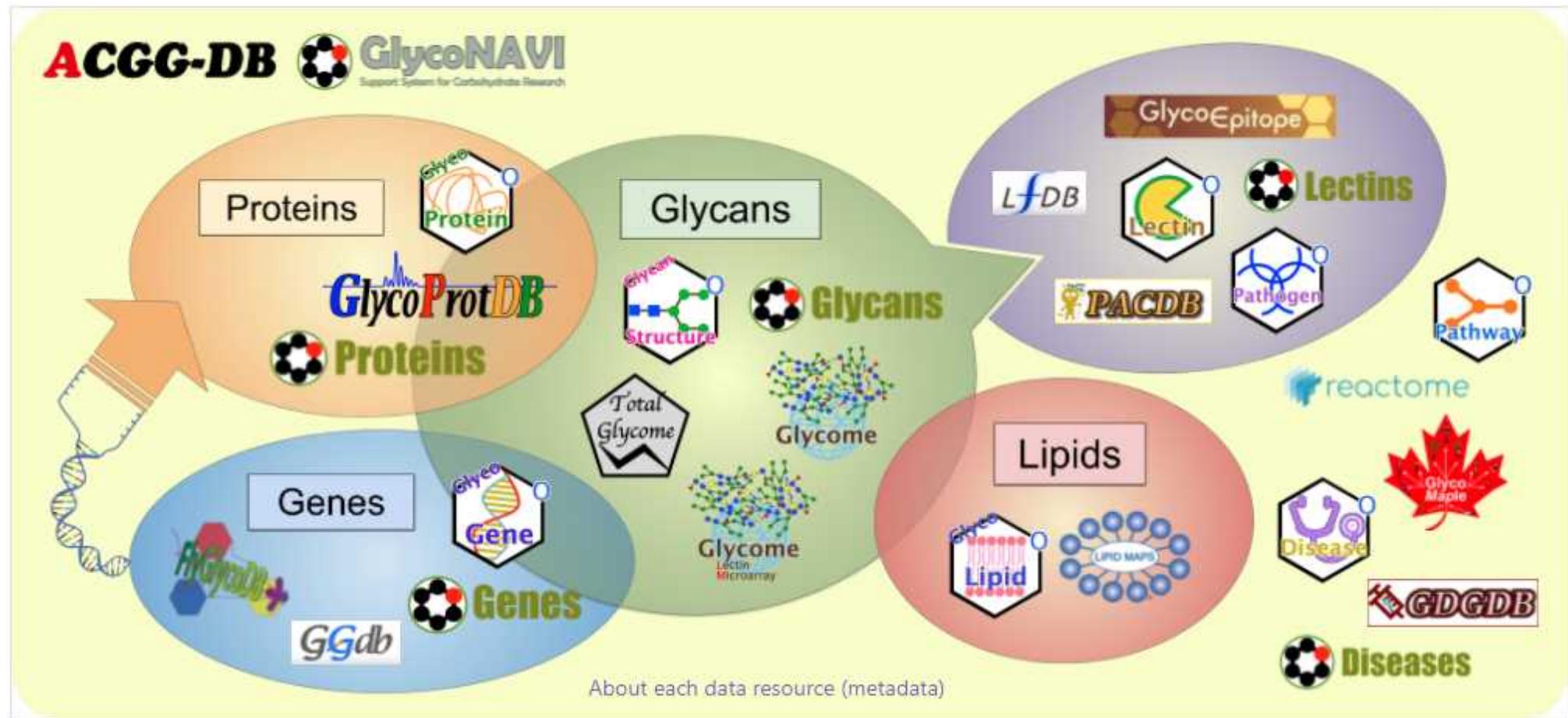




糖鎖科学ポータル

GlyCosmos

(<https://glycosmos.org>)





GlycoMaple (<https://glycosmos.org/glycomaple/index>)

【クエリ】 遺伝子発現データ (マイクロアレイ/RNA-seq)

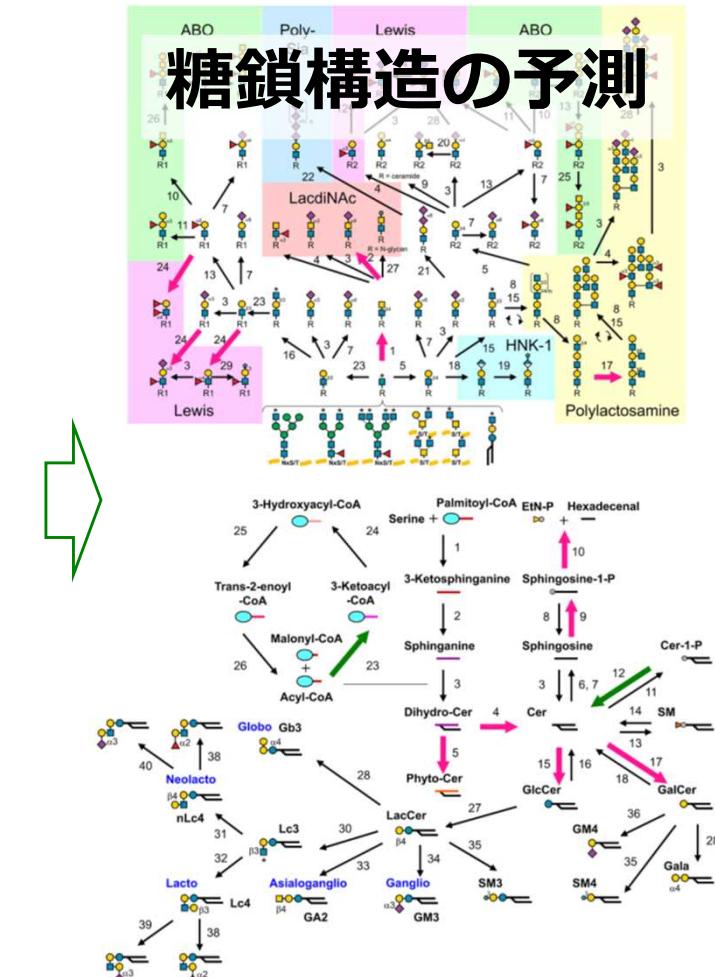
【得られる結果】 どのような糖鎖が合成されているかを予測

糖鎖関連遺伝子
発現データ
マイクロアレイ
RNA-seq



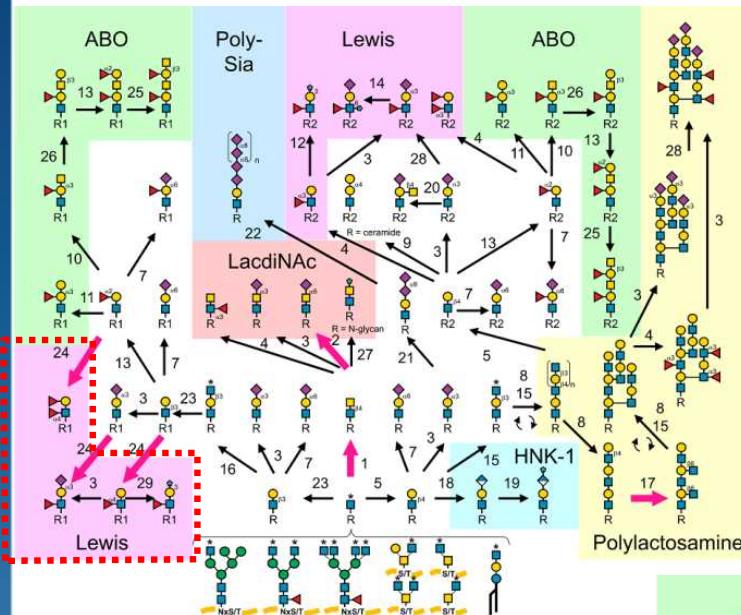
GlycoMaple

The screenshot shows the Glycosmos Portal interface. At the top, there are dropdown menus for 'Database Name' (set to 'glycomaple') and 'Last Update' (set to 'April 1, 2020'). Below these are several search and filter options, including 'Search' and 'Complex' dropdowns. A central panel displays a diagram of a cell membrane with various glycan structures attached to proteins and lipids. The bottom of the interface has a 'Download Image' button and a 'Legend' section.





GlycoMapleを使った解析事例（GSE21422 乳管がん）



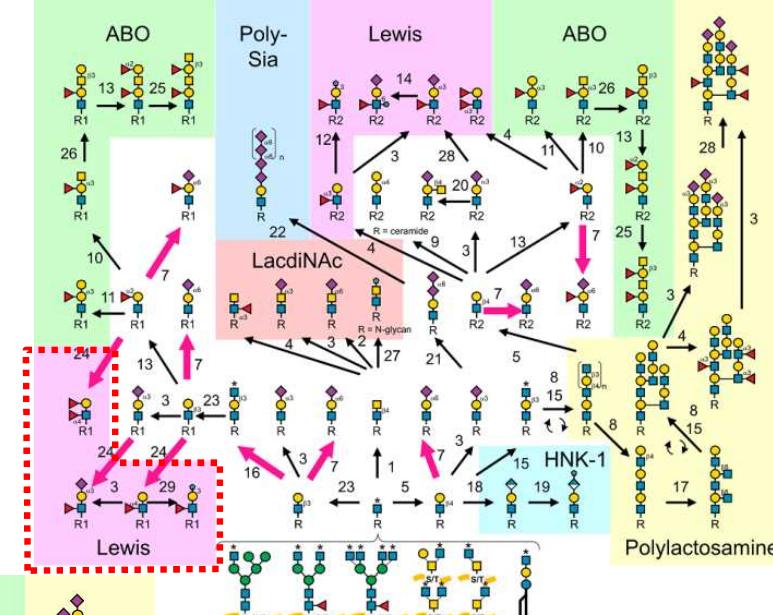
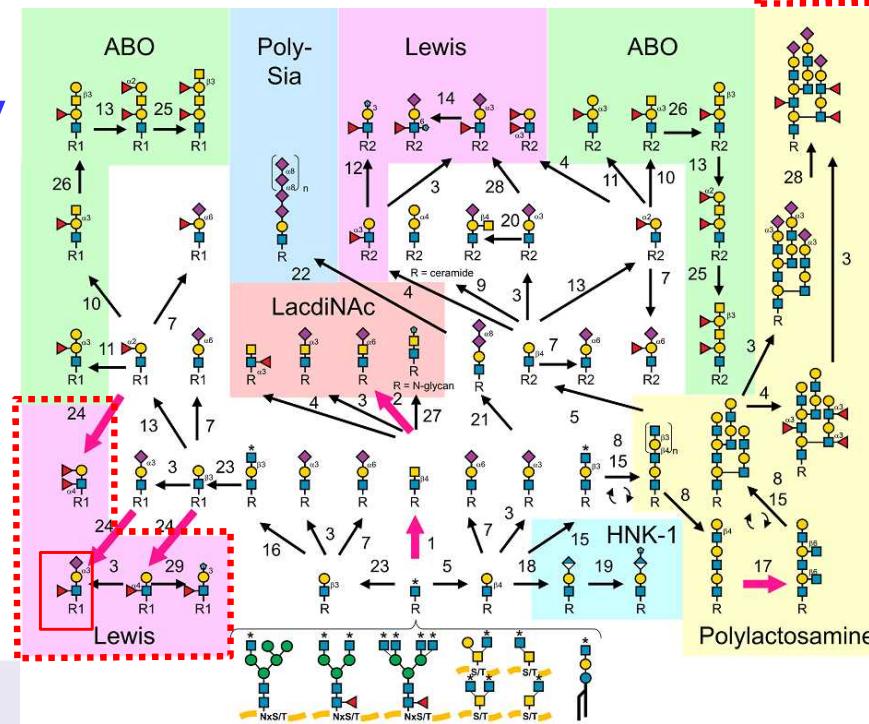
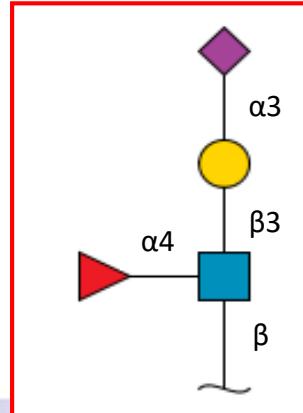
乳管がんで増加する
糖鎖構造を予測

Healthy < DCIS < IDC

IDC vs. DCIS

DCIS vs. Healthy

CA19-9



IDC vs. Healthy

【統合TV】GlyCosmosの使い方はこちらへ

1) GlyCosmosを使って糖鎖の情報を検索する (16分32秒) [2020-11-24]

<https://togotv.dbcls.jp/20201124.html>

GlyCosmos Portalは、日本糖質学会の公式ポータルとして認定されている、糖鎖科学とライフサイエンスの融合を目指したWebポータルです。リポジトリ（Submissions）からは、国際糖鎖構造リポジトリであるGlyTouCanと、グライコミクス・グライコプロテオミクスのための質量分析データリポジトリであるGlycoPOSTにアクセスすることができます。そして2020年8月より、質量分析から同定された糖鎖構造データのリポジトリUniCarb-DRがスウェーデンから移行され新たに追加されました。データセット（Resources）では、遺伝子、タンパク質、脂質、糖鎖、複合糖質、グライコーム、パスウェイ、疾患についてのデータがまとめられており、網羅的に検索することができます。

2) Find information related to glycans using GlyCosmos

(16分33秒) [2021-01-23]

<https://togotv.dbcls.jp/20210123.html>

3) 3分ちょっとで分かる！「世界初 糖鎖ポータルサイト開発」

(3分42秒) (創価大作成動画)

<https://www.youtube.com/watch?v=5UgGG4cXoIQ>

本日のデモデータ [2]

GSE7032

マウス 褐色脂肪細胞・白色脂肪細胞 初代培養細胞

URL : <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7032>

GSE7032:褐色脂肪細胞と白色脂肪細胞

URL	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7032
Organism	Mus musculus
Overall design	Comparisons of white and brown pre- and mature-adiposites
Platforms (1)	GPL81 [MG_U74Av2] Affymetrix Murine Genome U74A Ver. 2 Array https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL81
BioProject	PRJNA98391 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA98391

【関連論文】

PNAS March 13, 2007 vol. 104 no. 11 4401–4406

Myogenic gene expression signature establishes that brown and white adipocytes originate from distinct cell lineages

James A. Timmons et al. (Stockholm University)

【解析サンプル】

NMRI雄マウス(3-4週齢)

- ↓ 肩甲骨間・頸部・腋窩褐色脂肪組織、精巣上体白色脂肪組織
- ↓ DMEM with 10% 新生仔牛血清/2.4 nM インスリン/25 µg/ml アスコルビン酸Na/10 mM Hepes, pH 7.4 /4 mM グルタミン/50 U/ml ペニシリン/50 ug/ml ストレプトマイシン.
- ↓ 表現型をUcp1の発現で確認

Treated with 0.1 µM ノルエピネフリン処理 (4時間)

- ↓ 培養

4日後 前駆褐色脂肪細胞・白色脂肪細胞 (未分化)

7日後 褐色脂肪細胞・白色脂肪細胞 (分化)

【要旨】

- ✓ 褐色脂肪前駆細胞の初代培養は、筋原性のシグネチャを示した。
- ✓ 分化した褐色脂肪細胞では、筋原性シグネチャをサイレンシングする **Sirt11** 関連の転写シグネチャを示した。
- ✓ 白色脂肪前駆細胞では、筋形成に関わる核内受容体を抑制する転写因子 **Tcf21** が発現していた。
- ✓ 褐色脂肪前駆細胞と白色脂肪前駆細胞では、肥満に関する多くの分化関連遺伝子が発現していた。
- ✓ 褐色脂肪前駆細胞と筋細胞の類似性から褐色脂肪組織と白色脂肪組織は起源が異なることが示唆された。
- ✓ 褐色脂肪細胞が骨格筋組織のような酸化的な脂質異化作用に特化した機能を有する理由を説明するものと考えられた。



ChIP-seq統合データベース

ChIP-Atlas

(<https://chip-atlas.org/>)

➤ **Peak Browser**

IGVゲノムブラウザで転写因子が結合するゲノム領域を調べる

➤ **Target Genes**

転写因子が結合するターゲット遺伝子を調べる

➤ **Colocalization**

近傍に一緒に結合する転写因子の組み合わせを調べる

➤ **Enrichment Analysis**

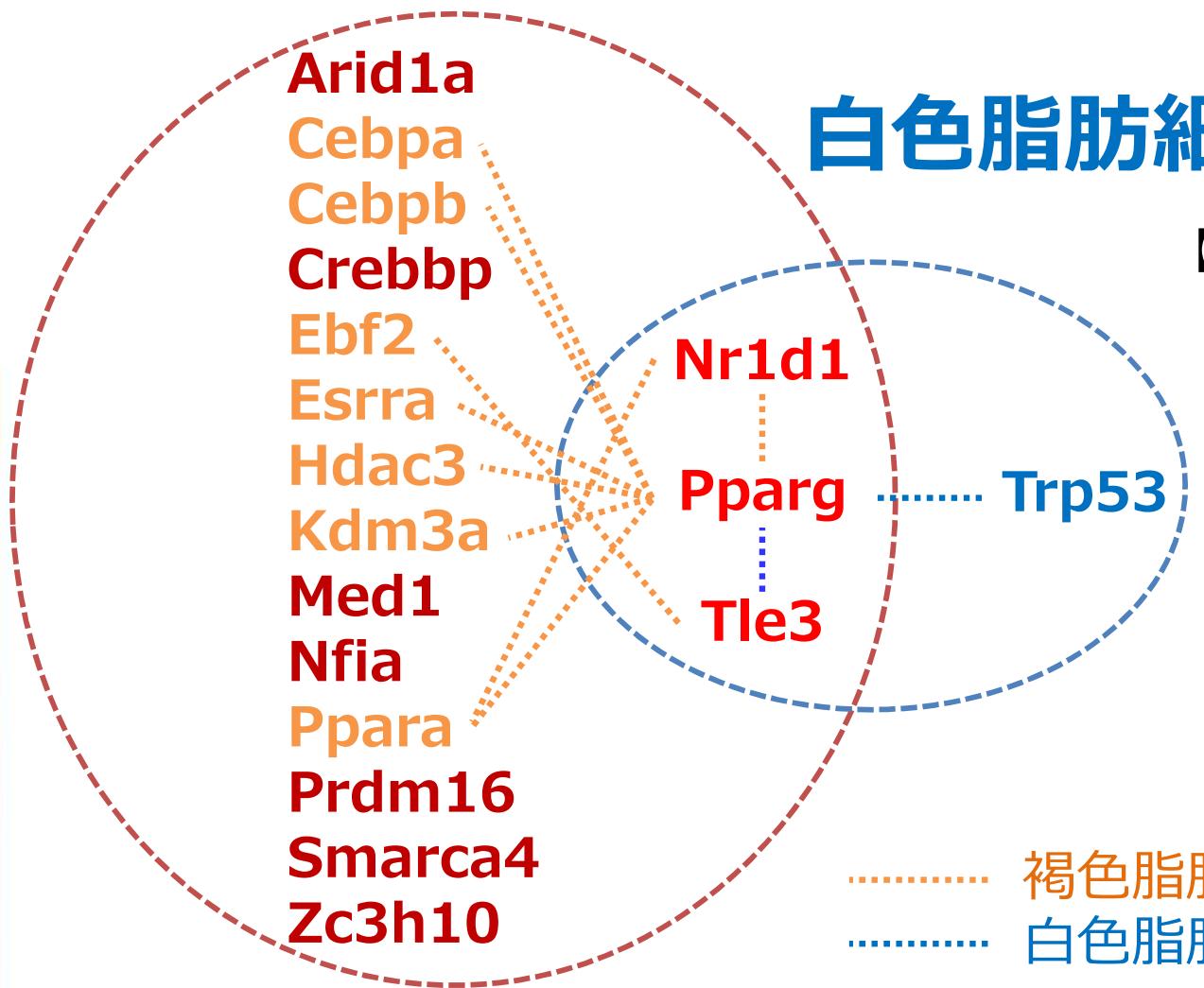
与えた遺伝子群の発現制御に関わる転写因子を予測する

褐色脂肪細胞と白色脂肪細胞の特徴的転写因子

褐色脂肪細胞

白色脂肪細胞

(使用データ : GSE7032)



【解析手順】

褐色脂肪細胞と白色脂肪細胞で発現差があった遺伝子群を **Enrichment Analysis**にかけ、褐色脂肪細胞と白色脂肪細胞に共通で挙がった3つの転写因子 Nr1d1、Pparg、Tle3に関して褐色脂肪細胞と白色脂肪細胞で共局在する転写因子パートナーを **Colocalization**で調べた。

Ppargの褐色/白色脂肪細胞におけるターゲット遺伝子

褐色脂肪細胞でのターゲット

Ech1	Ate1	Art3
Pdgfra	Acot8	Smyd5
Acaa2	Gpr137b	Acsf3
Etfdh	Mcm7	Entpd5
Slc39a1	Myl4	Ghitm
Pcx	Myl1	Fli1
Hcls1	Pdk4	Angpt1
Dnajc15	Nfyc	Ndufa9
Gabpa	Hccs	Tomm6
Chp1	Cox7a2	Gipc1
Coa3	Pvr	Tspan12
Aco2	Arl8b	Txlng
Nos3	Fubp1	Rad51ap1
Gapdh	Rangap1	Uqcc1
Pdhb	Idh3a	:

白色脂肪細胞でのターゲット

Mgst1	Ggt5	C3
Lpl	Mkln1	Entpd2
Cidec	Ncoa4	Tnfaip3
Serpine1	Atp1a2	Bdnf
Cdkn1a	Arid1a	Zranb1
Me1	C1qtnf1	Atp6v0a1
Cast	Nupr1	Tcf4
Acp5	Taf15	Klf9
Slc27a1	Amy1	Abca1
Rom1	Ttc14	Selenbp1
Clcn2	Otud5	Notch2
Nr2f2	Repin1	Dicer1
S100a13	Cmip	Ptprs
Rbm39	Fmr1	Apbb2
Rbck1	Orm2	:

【統合TV】ChIP-Atlasの使い方はこちらへ

1) ChIP-Atlasを使って既報のChIP-seqデータをまとめて閲覧する

～Peak Browserの使い方～ (9分40秒) [2018-01-23]

<https://togotv.dbcls.jp/20180123.html>

ChIP-Atlas は、(NCBI, EMBL-EBI, DDBJなどに登録されたほぼ全ての ChIP-seq データが解析できます。既報のChIP-seqデータをまとめて閲覧する「Peak Browser」の使い方。Peak Browserでは、Integrative Genomics Viewer (IGV)を用いてスムーズなブラウジングができ、遺伝子のシス調節領域を予測したり、その遺伝子の発現を制御する転写因子を予測します。

2) ChIP-Atlasを使って興味のある転写因子を選択しその標的遺伝子候補を検索する

～Target Genesの使い方～ (7分21秒) [2018-01-24]

<https://togotv.dbcls.jp/20180124.html>

興味ある転写因子を選択し、転写開始点と検索範囲を指定すると、その転写因子が制御する標的遺伝子を予測する「Target Genes」の使い方を紹介します。

3) ChIP-Atlasを使って共局在タンパク質を探す

～Colocalizationの使い方～ (6分11秒) [2018-01-28]

<https://togotv.dbcls.jp/20180128.html>

興味ある転写因子を選択し、それとゲノム上で共局在する転写因子候補を検索、つまり結合部位が近接する転写因子ペアを検索する「Colocalization」の使い方を紹介します。

4) ChIP-Atlasを使って興味ある遺伝子リストを制御する可能性の高い転写因子を調べる

～Enrichment Analysisの使い方～ [2019-01-05]

<https://togotv.dbcls.jp/20190105.html>

利用者の遺伝子リストと既存のChIP-seqデータを比較解析する「Enrichment Analysis」の使い方を紹介します。入力した遺伝子群の発現をまとめて制御する転写因子を予測するほか、BED形式ファイルやシーケンスマチーフを入力すると、それらにenrichmentする転写因子群を調べることができます。



プロテオーム統合データベース jPOST (<https://jpostdb.org/>)

jPOST
Repository/Database

Japan Proteome Standard Repository/Database

Recent posts

- other**
jPOST member's co-authored paper about USI has been published.
⌚ 2021-06-30 pjost
 The Universal Spectrum Identifier (USI) is an essential mechanism for the wide use of proteomics data provided by the ProteomeXchange (PX) repositories. It has been discussed for a long time in the HUPO Proteomics Standards Initiative, to which jPOST member Prof. Kawano has contributed greatly. The USI facilitates access to the huge amount of spectral data registered in the PX repositories, and is expected to make proteomics data more findable, accessible, interoperable, and reusable. We expect that the USI leads to the use of jPOST data and contributes to research in the field of life science, including proteome research. Universal Spectrum Identifier for mass spectra. Eric W. Deutsch, Yasset [...]
- event**
JCompMS 6th workshop
⌚ 2021-05-26 pjost
 Workshop details in Japanese
- other**
jPOST database and repository will be stopped. (on Nov. 27-30)
⌚ 2020-11-19 pjost
 jPOST database and repository server will be unavailable during the time shown below due to a scheduled system maintenance. We thank you for your understanding and cooperation in this matter.
 Date and Time: Friday, November 27, 17:00 – Monday, November 30, 10:00 (UTC+9)

Post raw data **Views**

To repository To database

Submission form data dashboard Slices

Register Faceted search Aggregate

Repository Cube Globe

Reprocessing Pack into database Results

jPOST REPOSITORY A member of ProteomeXchange

Repository Submit Help ⌚ Sign in 👤 Sign up

About jPOSTrepo
 jPOSTrepo (Japan Proteome Standard Repository) is a new data repository of sharing MS raw/processed data. It consists of a newly-developed, high-speed file upload process, flexible file management system and easy-to-use interfaces. Users can release their "raw/processed" data via this site with a unique identifier number for the paper publication. Users also can suspend (or "embargo") their data until their paper is published. The file transfer from users' computer to our repository server is very fast (roughly ten times faster than usual file transfer) and uses only web browsers – it does not require installing any additional software.

jPOST is a certified member of ProteomeXchange Consortium and jPOSTrepo provides official ProteomeXchange Identifiers to projects stored in our repository.

Reference

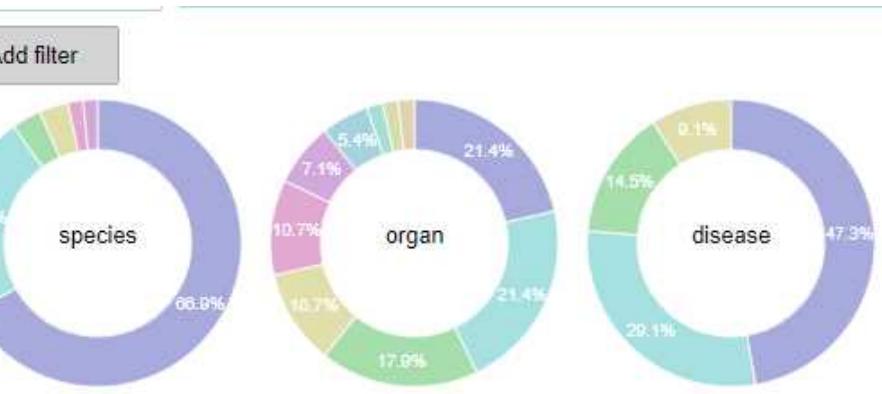
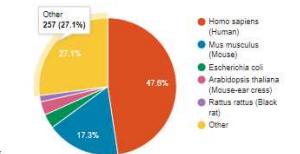
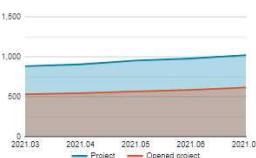
Please cite the following article when using jPOSTrepo:
 Okuda, S. et al. jPOSTrepo: an international standard data repository for proteomes. *Nucl. Acids Res.* 45 (D1): D1107-D1111 (2017). doi: 10.1093/nar/gkw1080 [pubmed]

Statistics

1021 projects are registered. 615 are opened.

77931 files amount to 33.5 TB

164 species.



プロテオームデータからの転写因子予測

jPOST (<https://jpostdb.org/>)

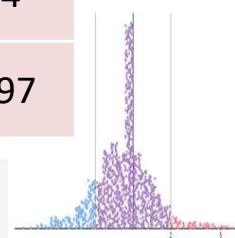
Slice1: White adipocyte

Dataset ID	Project ID	#proteins	#spectra
DS791_1	JPST000791	1,385	4,047
DS791_2	JPST000791	1,182	3,527
DS791_3	JPST000791	1,615	5,404
DS791_4	JPST000791	1,519	5,089

Slice2: Brown Adipocyte

Dataset ID	Project ID	#proteins	#spectra
DS793_1	JPST000793	1,223	4,421
DS793_2	JPST000793	1,100	3,499
DS793_3	JPST000793	1,321	4,513
DS793_4	JPST000793	786	2,408

	White adipocyte		share	Brown Adipocyte	
	total	unique		total	unique
# proteins	2,000	716	1,284	1,688	404
# peptides	3,771	1,644	2,127	3,524	1,397



ChIP-Atlas (<https://chip-atlas.org/>)

ChIP-Atlas: Enrichment Analysis

Predict proteins bound to given genomic loci and genes

H. sapiens (hg19) H. sapiens (hg38) M. musculus (mm10) M. musculus (mm10) R. norvegicus (rn5) D. melanogaster (dm3) D. melanogaster (dm3)

C. elegans (ce10) C. elegans (ce11) S. cerevisiae (SacCer2)

1. Antigen Class: Human antigens (3337)
Disease (942)
Hepatitis (11)
Hippo (248)
TFs and others (1310)
Translating (10)
Unclassified (1935)
Not described (324)

2. Cell type Class: All cell types (3337)
Adipocytes (63)
Blood (146)
Brain (10)
Breast (92)
Colon (278)
Digestive tract (539)
Endocrine (10)
Epithelial (1000)

3. Threshold for Significance: 100, 200, 300, 500

4. Enter dataset A: Gene list (Gene symbols) Hapt1, Mid2, Pcbp1, Acsl1, Tubn, Eif4e, Tmem52b, Sfha
5. Enter dataset B: Ribaoq coding genes (excluding dataset A) Gene list (Gene symbols) Ahsgf1, Fln1, Plau2, Alaa2, Crm1, Bst1
6. Analysis description: Analysis title: White vs Brown
Dataset A title: White adipocyte
Dataset B title: Brown adipocyte
Distance range from TSS: -5000 bp & TSS & +5000 bp

submit Estimated run time: 5 mins

Brown adipocytes

Cebpa
Cebpb
Ebf2
Esrra
Hdac3
Med1

Nfia
Nr1d1
Ppara
Pparg
Prdm16
Tle3
Zc3h10

White adipocytes

Pparg
Tle3
Nr1d1
Trp53

プロテオームデータからも、遺伝子発現データとほぼ同じ転写因子群の関与が予測された。

「信頼できるデータ区間と 変動遺伝子の選抜方法」 について考える

発現量で見るか、発現変動で見るか

発現量

(シグナル値)

発現変動

(対照群に対する変動比)

シグナル値 \neq 発現量

発現変動 \neq 機能変化

\neq 機能の重要性

\neq 生物学的機能の重み

➤ 測定系の検出感度とノイズ

➤ 低発現遺伝子ほど変動幅が
振れやすい

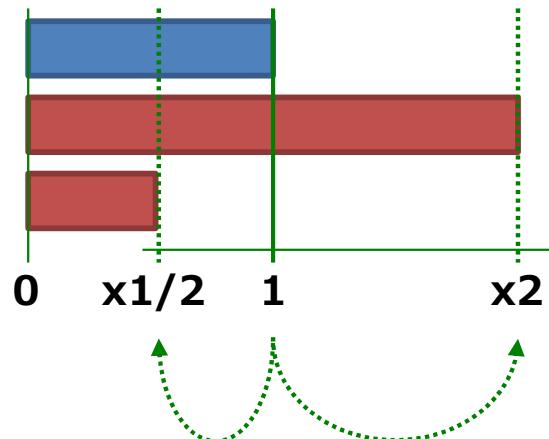
変動値は「対数値」で扱う

線形数値だと発現上昇遺伝子と発現低下遺伝子が均等に比較できない

対照群

x_2 発現上昇

$x_{1/2}$ 発現低下



変動幅は「A（処置）/B（対照）」

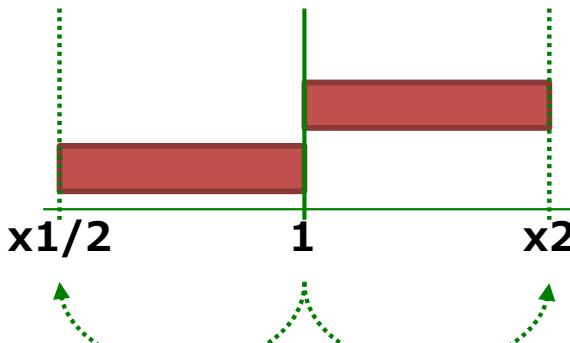
等距離にならない

対数化することで発現上昇遺伝子と発現低下遺伝子が均等に比較できる

対照群

x_2 発現上昇

$x_{1/2}$ 発現低下



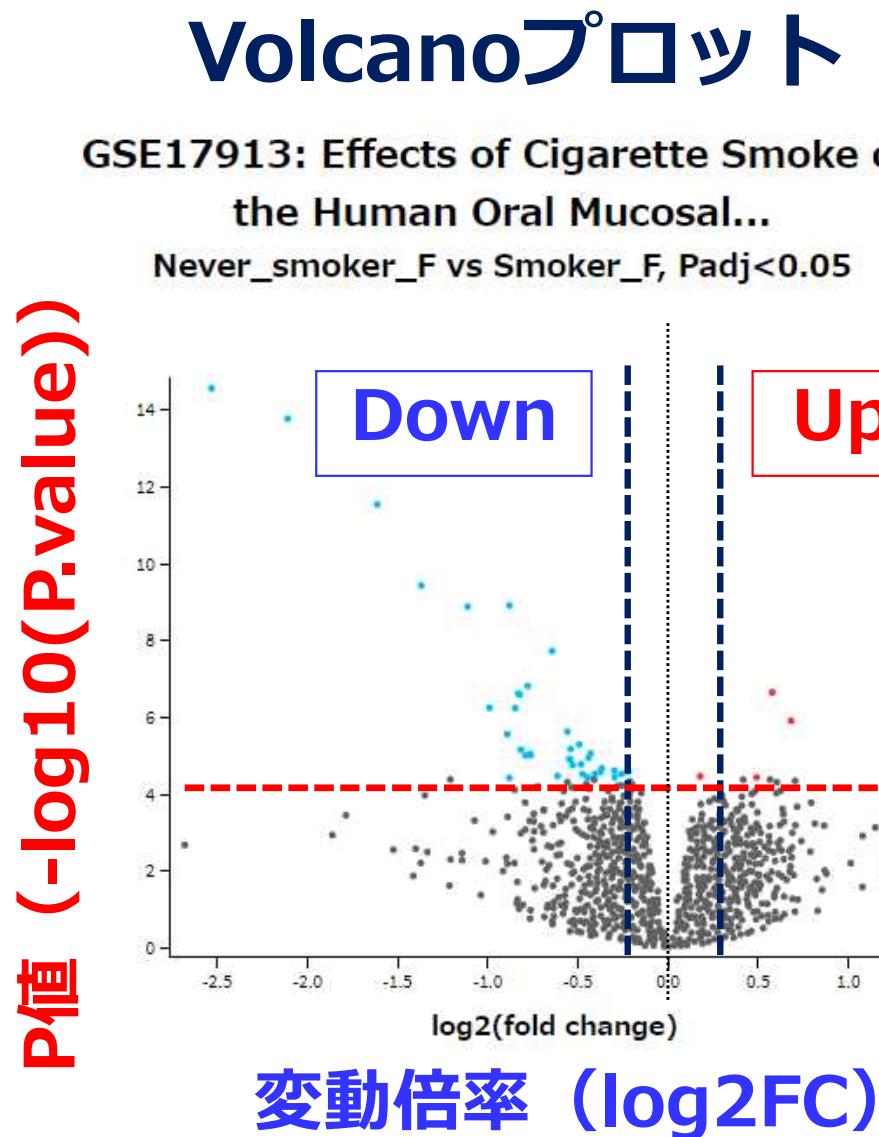
変動幅は「A（処置）-B（対照）」

等距離になる

※ 慣習的に、対数の底は「2」を使うことが多い。

※ 対数以外にも+/-で示したFold Changeを使うケースもある。

① 変動倍率とP値による選抜(Volcanoプロット)



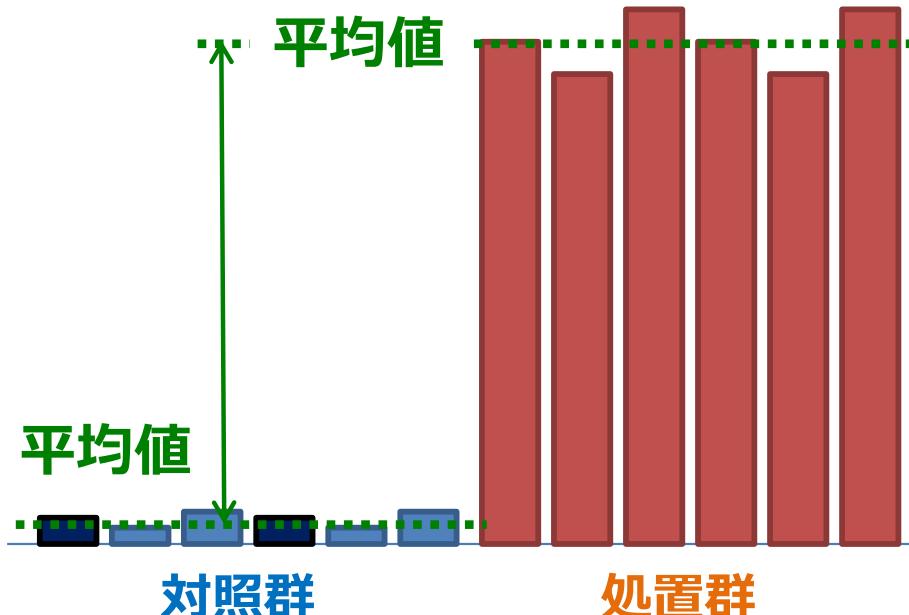
比較する2群間のデータを比較し、『統計学的に有意な差』がある遺伝子を選抜する手法。多くの論文で採用され、一般的で、かつ最もよく使われている手法。

- 【前提（暗黙の条件）】
- ✓ 2つのサンプル群内のデータが『質的に揃っている』こと。
 - ✓ 群間差が有意かどうかが判定できる『十分な数』のデータがあること。

- 【問題点】
- ✓ 臨床サンプルデータでは、この二つの条件を満たすことがむずかしい。

「群代表値」の選択肢

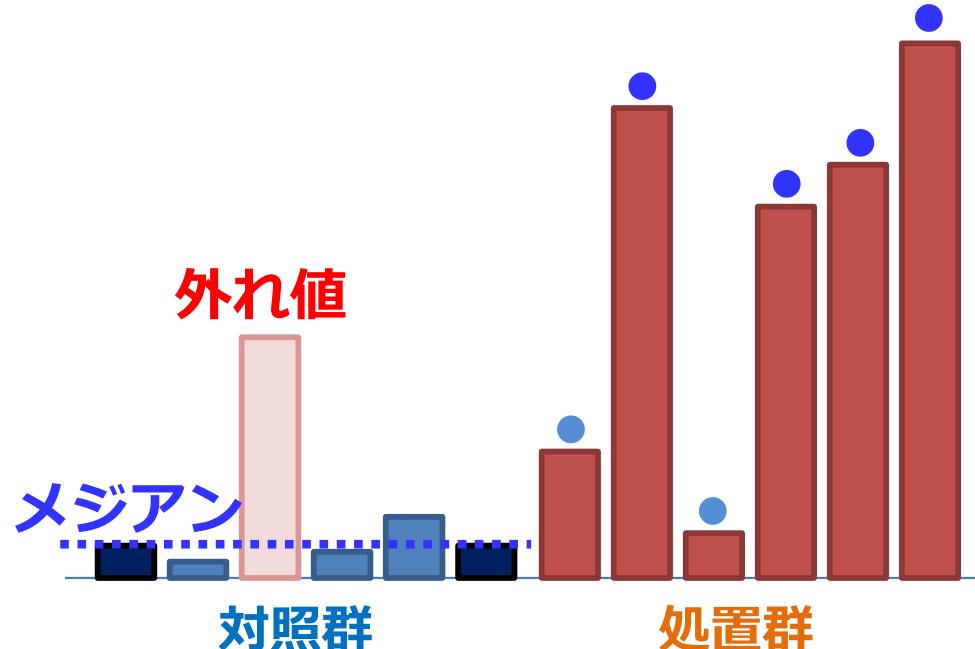
A) 群内のはらつきが小さい場合



ばらつきが小さくN数がある場合は、
T検定(2群)やANOVA(3群以上)などの
検定手法で、**平均値**の比とP値から変動
遺伝子を抽出すると良い。

※ T検定やANOVAで変動遺伝子を抽出する場合は、群代表値はメジアンではなく平均値を使う。

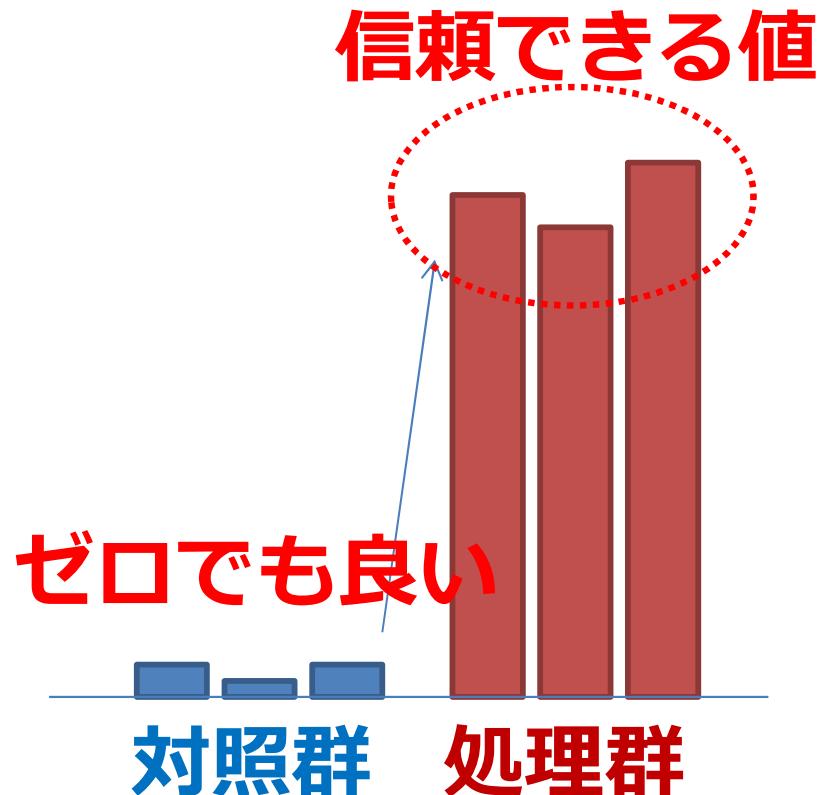
B) 群内のはらつきが大きい場合



ばらつきが大きい場合やN数が少ない
場合は、**メジアン**を対照群の代表値と
して個別サンプル毎の変動を算出し、
3/6以上とか4/6以上の**頻度条件**で
変動遺伝子を抽出すると良い。

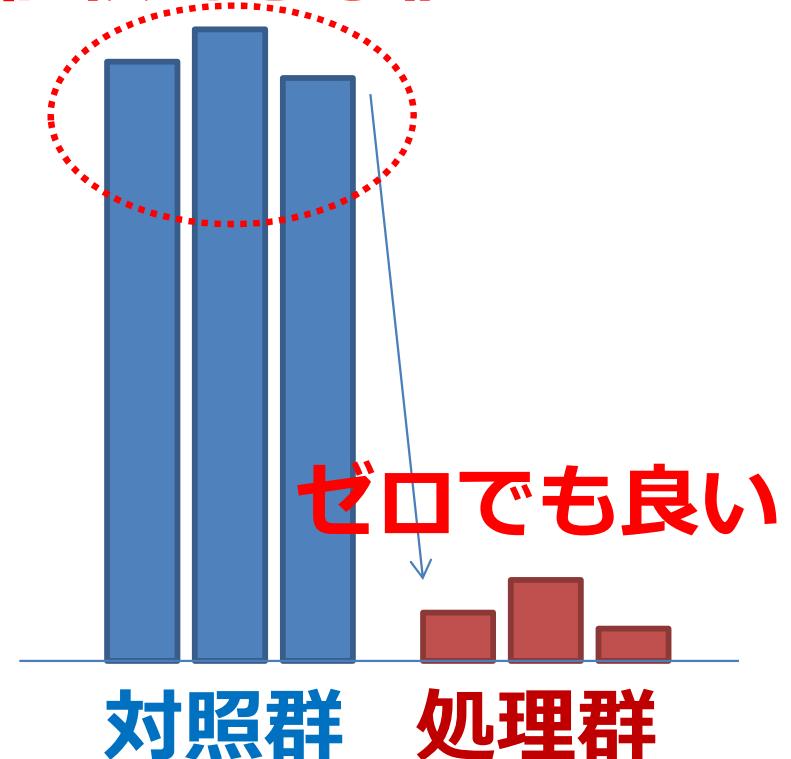
「信頼できる発現変動」の考え方

発現上昇遺伝子のQC



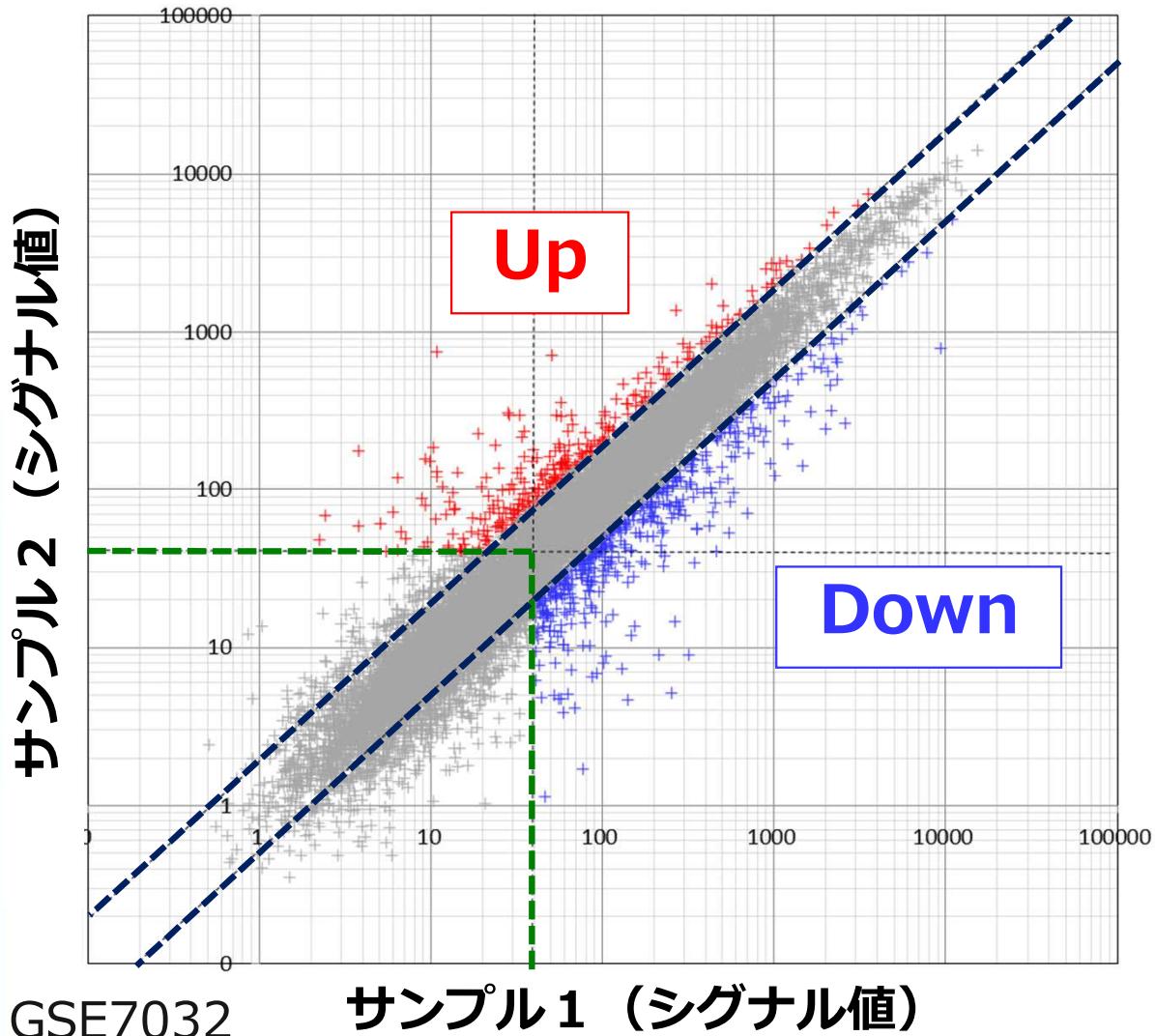
発現低下遺伝子のQC

信頼できる値



② シグナル閾値と変動倍率を使った選抜

Brown vs. White, adipocytes (Expression and FC)



『2つのサンプル』を比較し、
『適切な閾値』を設定した上で、
遺伝子を選抜する。

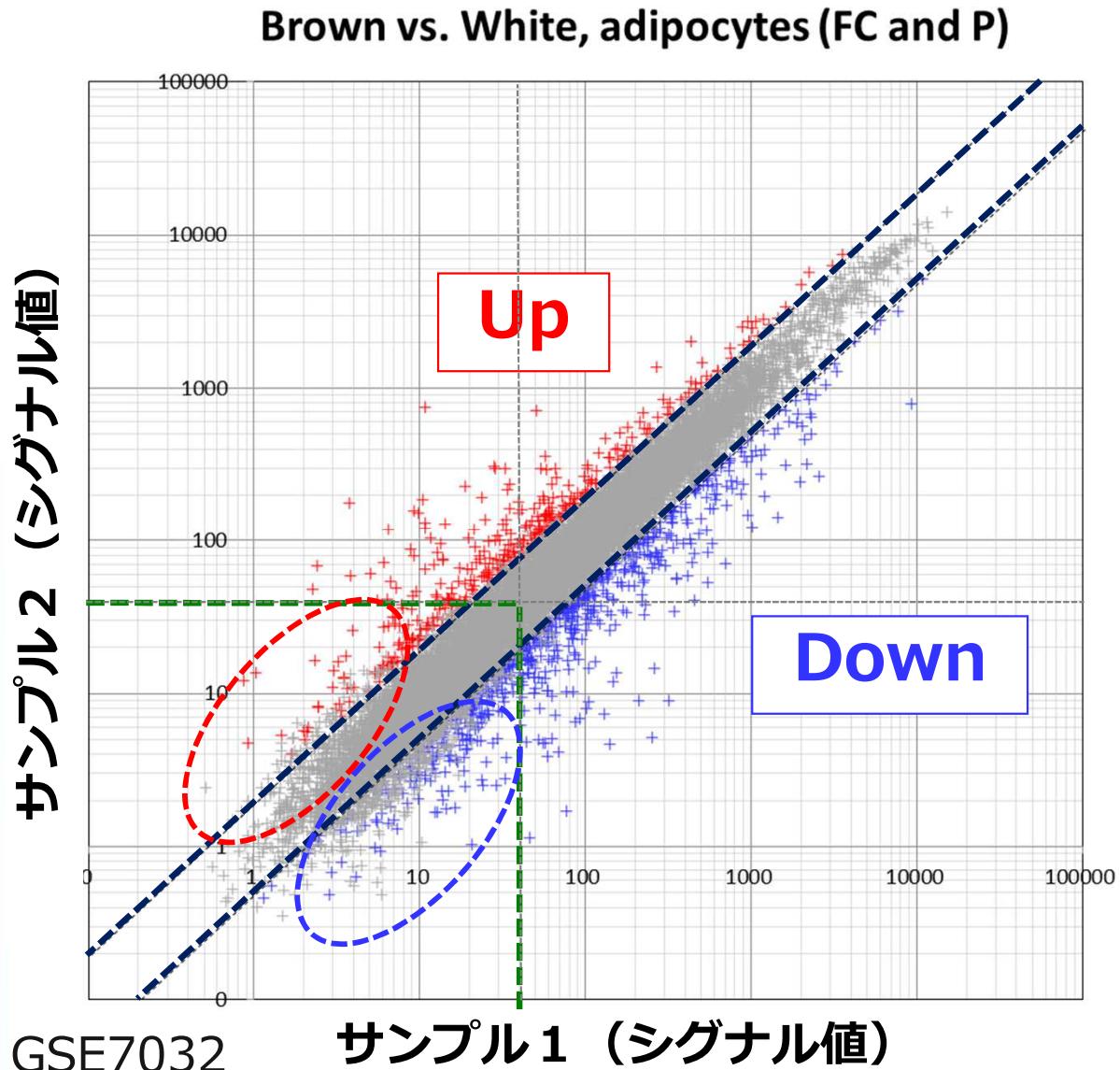
【利点】

- ✓ サンプル数に依存しない。
- ✓ 質的に揃っていないデータに対して
も対応可能。

【課題】

- ✓ 『適切な閾値』はどのように決める
か。

変動倍率とP値による選抜の問題点



変動遺伝子選抜（検定手法）では、有意差さえつけば、検出限界以下の発現レベルが低い遺伝子についても『有意な発現変動遺伝子』と判定してしまう。

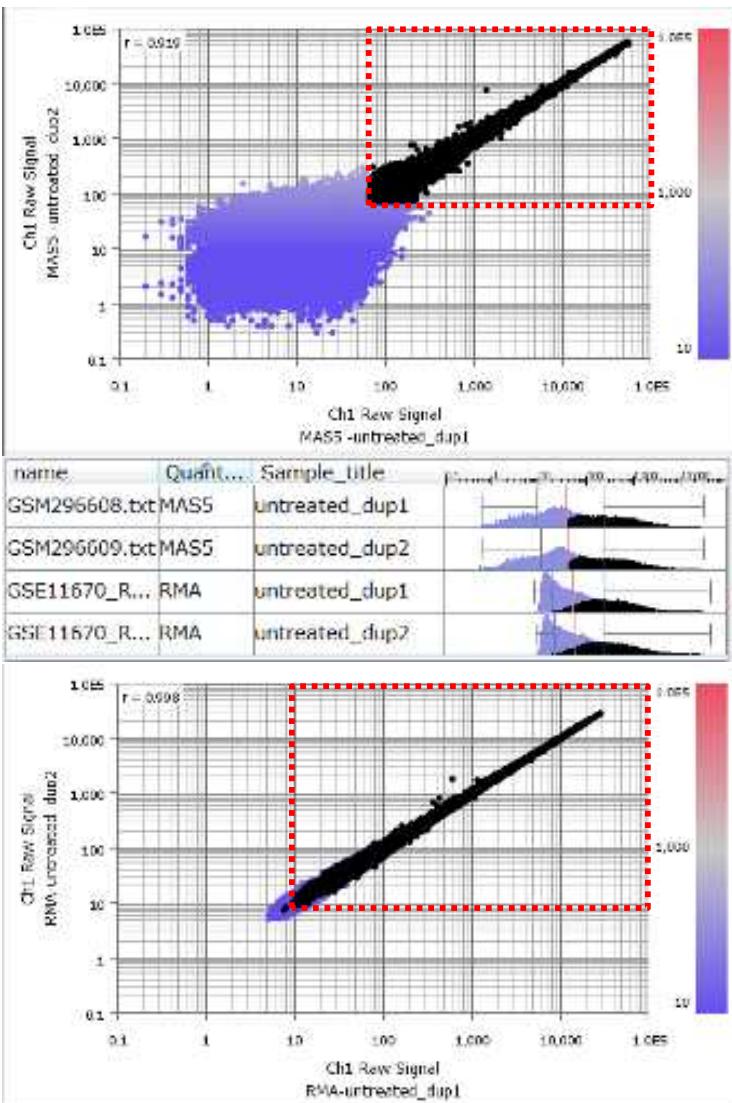
発現レベルの低い遺伝子は実験誤差やサンプル間のばらつきに起因する『見かけ上の変動倍率』が大きく出る傾向があり、選抜した発現変動遺伝子を『変動倍率』で評価すると、意味のない誤った解釈につながってしまう危険性があるので注意する必要がある。

この過ちを犯し、結論をミスリードしている論文も少なくない。統計学的には正しくても生物学的に正しいとは必ずしも言えない点に注意。

「測定技術と検出限界」 について考える

Affymetrix 3'-IVT GeneChipデータの信頼区間

MAS5



Affymetrix 3'-IVT GeneChip

低温ハイブリで測定系は16ビット取り込み。

RMA

【数値化方法による違い】

MAS5では、CM/MMプローブの検定でプローブセットごとに評価フラグ (P/M/A) を出す。シグナル値<100の信頼性は高くないが、評価フラグを使ってノイズではない値を拾える。

信頼区間のダイナミックレンジは 10^3 程度、約25,000プローブ(約12,000遺伝子)を検出。

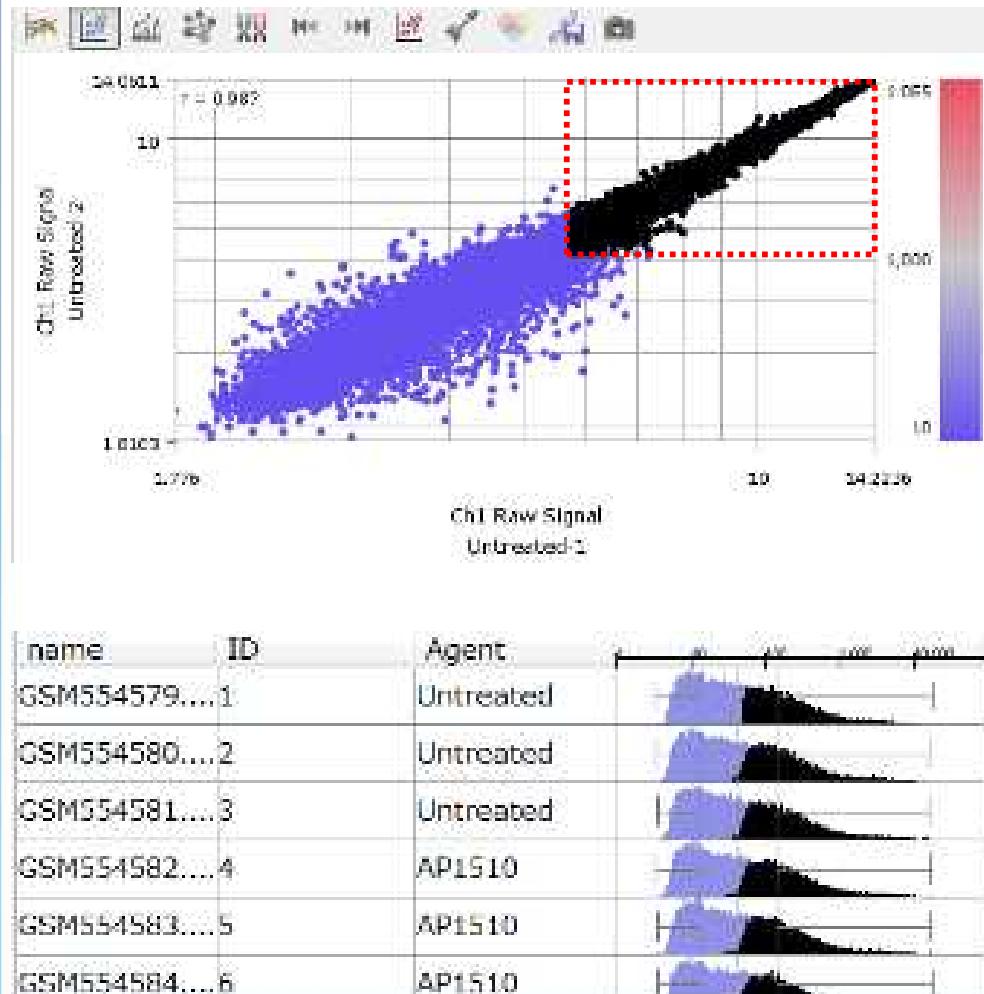
RMAでは、データセット内で数値分布を正規化した上で数値化する。フラグは出さない。検出感度と再現性が高くなつたように見えるが、正規化でノイズが隠れただけなので、信頼区間とノイズの見きわめがむずかしくなっている。

[GSE11670 \(K562 コントロール実験duplicate\)](#)

出展: Subio (<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-microarrays>)

Affymetrix Gene ST Arrayデータの信頼区間

Affymetrix HuGene-1.0-st



[GSE22288](#) (MCF10A-HER2-FKBP-HAコントロールduplicate)

出展: Subio (<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-microarrays>)

Affymetrix Gene ST Array

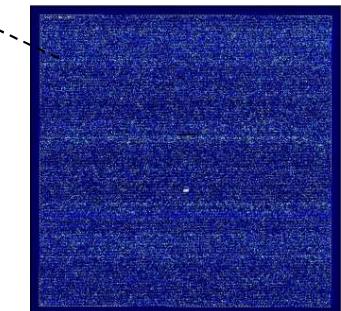
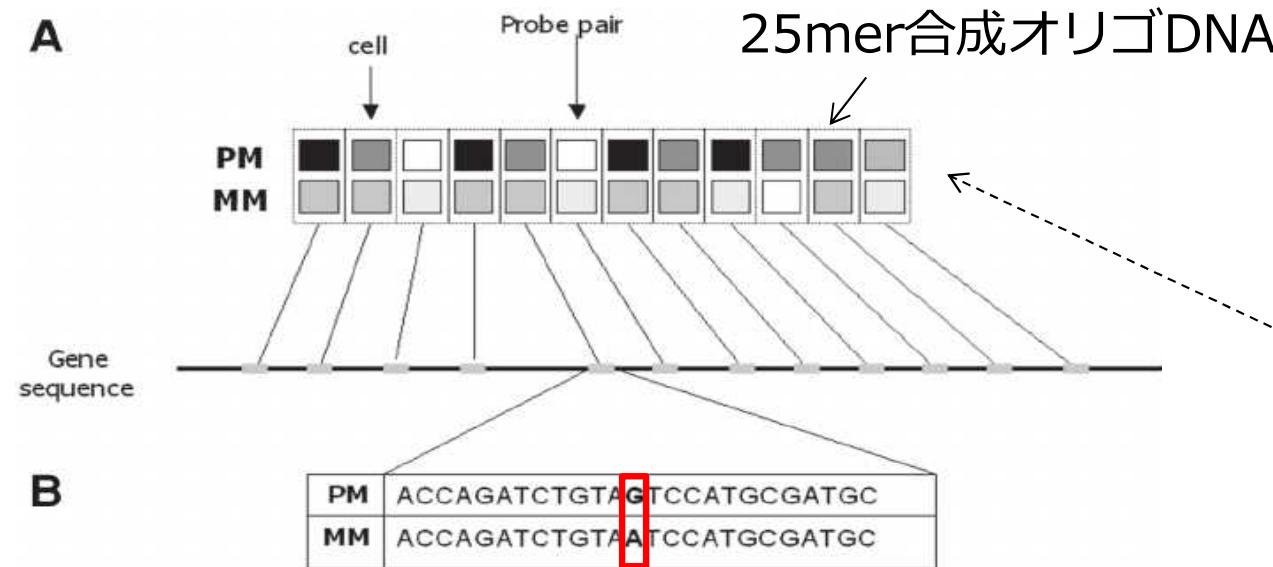
3'-IVT GeneChipの後継技術として出された。3'-IVT GeneChipは3'-UTRにプローブが設計されていたが、Gene ST Arrayではエクソン領域にプローブが設計されている。低温ハイブリのでもともとノイズが高いが、ランダムプライマー標識でさらにS/N比が悪くなつた。

数値化ソフトでバイアスとノイズを隠しているため、低シグナル域ほどばらつきが小さくなるという**不自然なシグナル分布**をする。

信頼区間の**ダイナミックレンジは10³弱**、およそ12,000～15,000プローブ(**9,000～11,000遺伝子**)の発現が測定できていると思われるが、信頼区間とノイズの境界が分かりにくく、解析しづらい。

Affymetrix GeneChip

Probeset design



Affymetrix GeneChip (3'IVT Expression)

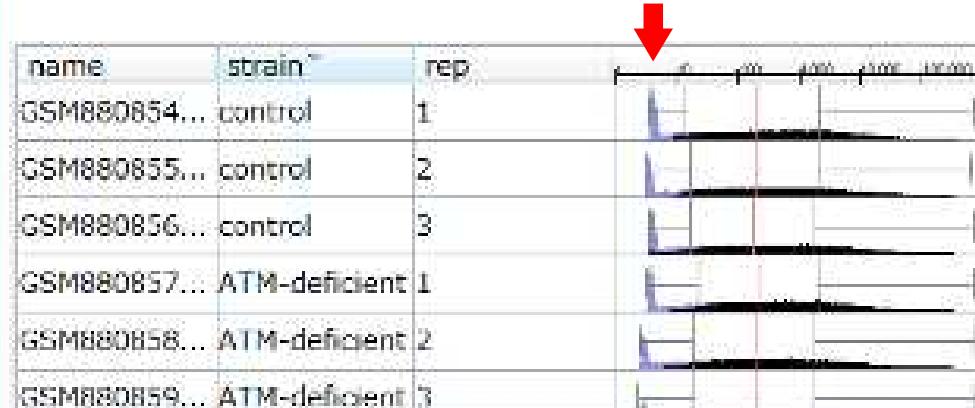
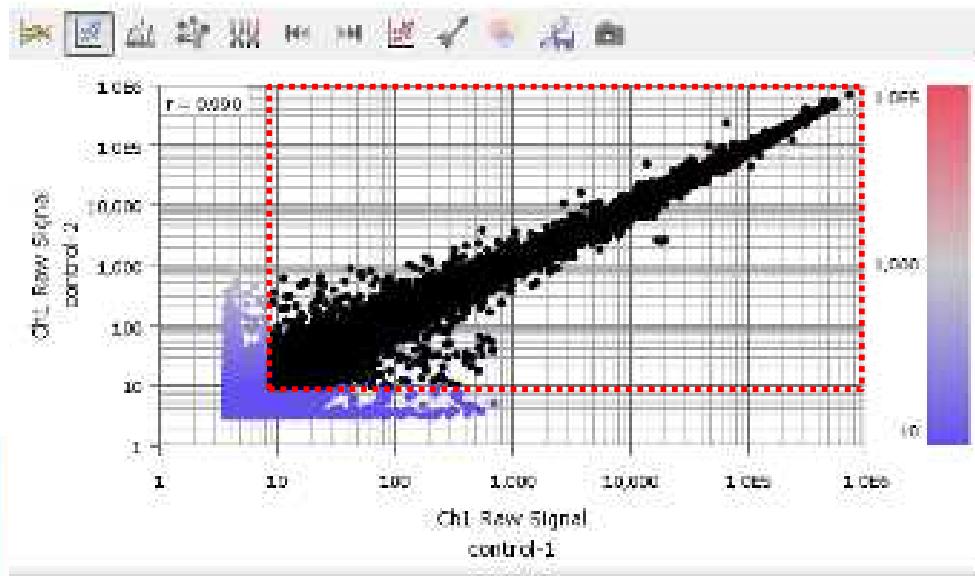
遺伝子の3'末端側に1ないし20プローブセットをデザイン
Perfect Match probe (PM) および Mismatch probe (MM)

Affymetrix Gene ST Array

1エクソンあたり約4プローブ、1遺伝子あたり約40プローブ
Perfect Match probe (PM) のみ

Agilent Microarrayデータの信頼区間

Agilent Whole Genome 44K

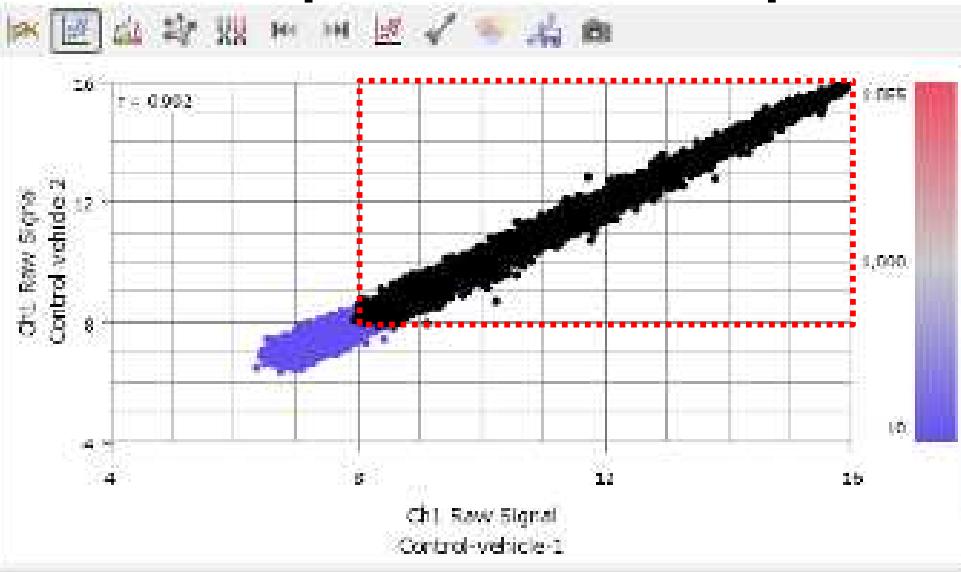


[GSE36082 \(HME-CCコントロールduplicate\)](#)

出展: Subio (<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-microarrays>)

Illumina BeadChipデータの信頼区間

Illumina HumanHT-12 v4.0 expression beadchip



GSE25315 (MCF-7コントロールduplicate)

出展: Subio (<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-microarrays>)

Illumina Expression Beadchip

高温ハイブリだがデータ取り込みは16ビット。ビーズアレイのため同一プローブを複数回測定し信頼性を高めている。散布図ではばらつきが小さく見えるが、RMAでの数値化によるもの。

信頼区間のダイナミックレンジは**10²~10³弱**とAffymetrix 3'-IVT GeneChipよりも狭い。約15,000プローブ(**約11,000遺伝子**)を検出。

ビーズスキャナーが専用装置のため、RMA以外の数値化手法が選択できない制約がある。

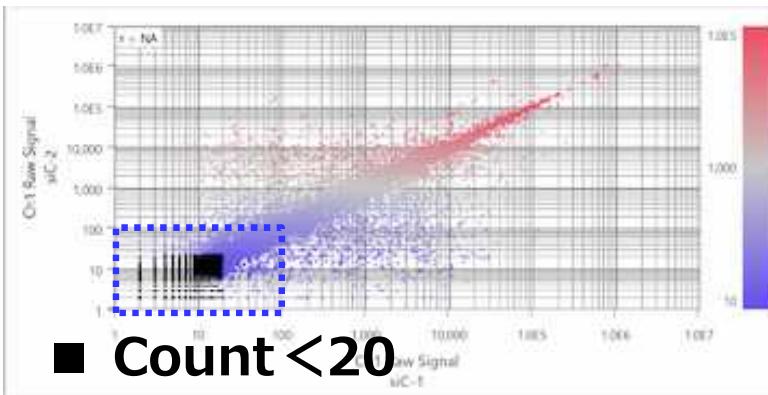
=====

【参考】TIFFファイルの形式と理論階調

- 16ビット : 65,536階調 ($\sim 10^4$)
- 20ビット : 1,048,576階調 ($\sim 10^6$)

RNA-seqデータの信頼区間

Count



Count (Read)

ゲノム上にマッピングされた各転写物ごとのリード数。

RPM (read per million read mapped)

各サンプル総リード数を100万あたりに換算して正規化。

TPM (tag per million read mapped)

Countを遺伝子長1000 bpあたりに補正後、各サンプルの総リード数を100万あたりに換算して正規化。

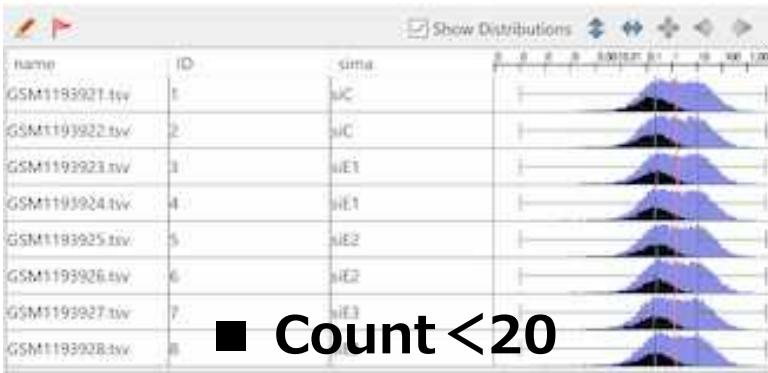
FPKM/RPKM (fragments per kilobase of exon per million read mapped)

RPMを各転写産物ごと遺伝子長が1000 bpあたりに換算して補正。

RNA-seqのダイナミックレンジ

- Countで判断しないと信頼区間が特定できない。
- 総リード数によって信頼区間が変わる。
- 総リード数、数1000万で上位1万遺伝子くらい(?)
- **2000~5000万リード ≈ Affymetrix GeneChip**
- **1億リード ≈ Agilent Microarray**

TPM



出展: Subio (<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-rna-seq>)

NCBI SRA RNA-seqデータの品質(ヒト)

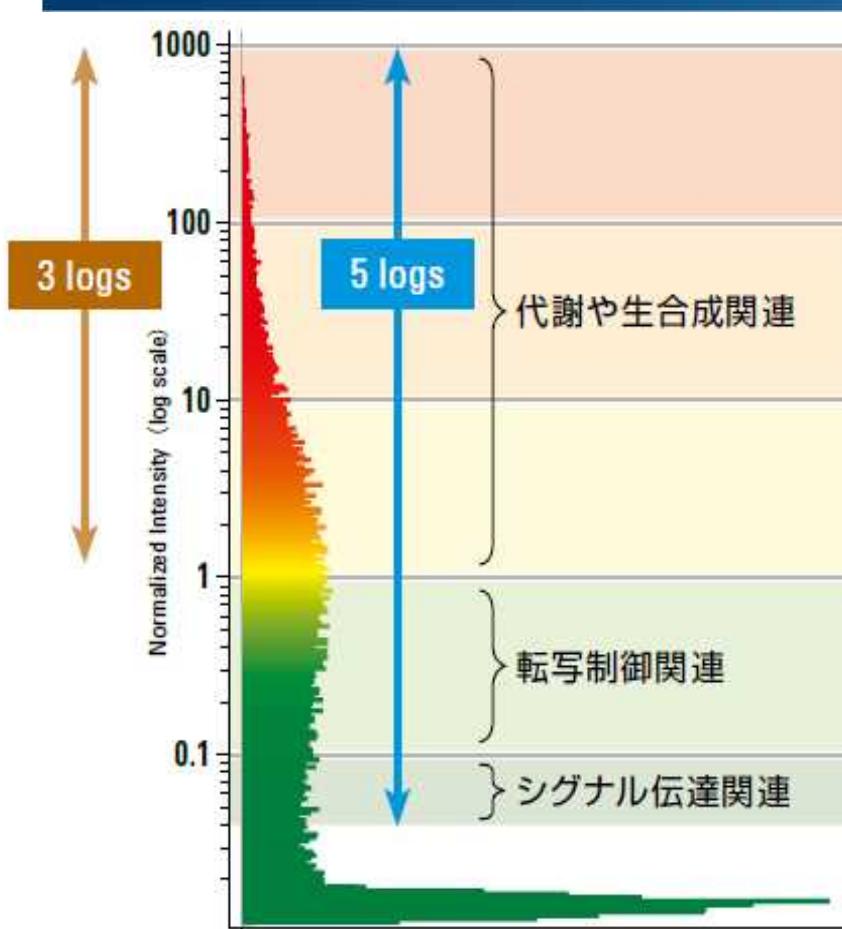
QC Metric	Num	Ratio
PASS	2,275	21.8%
WARN	2,800	26.8%
FAIL	5,358	51.4%
Total	10,433	100.0%

← 十分な品質が確保できている
RNA-seqデータはわずか2割。
多くはマップされたリード数の
不足が原因。

Code	Metric	Warn threshold	WARN		Fail threshold	FAIL	
1	NumReadsQcPass	< 500 reads per gene	2,322	22.3%	< 50 reads per gene	639	6.1%
2	QcPassRate	< 80%	343	3.3%	< 60%	592	5.7%
3	STAR_UniqMapRate	< 70%	463	4.4%	< 50%	2,771	26.6%
4	STAR_AssignRate	< 60%	1,045	10.0%	< 40%	3,318	31.8%
5	STAR_AssignedReads	< 500 reads per gene	2,566	24.6%	< 50 reads per gene	2,941	28.2%
6	Kallisto_MapRate	< 60%	184	1.8%	< 40%	3,291	31.5%
7	Kallisto_MappedReads	< 500 reads per gene	2,520	24.2%	< 50 reads per gene	1,954	18.7%
8	DatasetCorrel	< 0.5	0	0.0%	-	0	0.0%

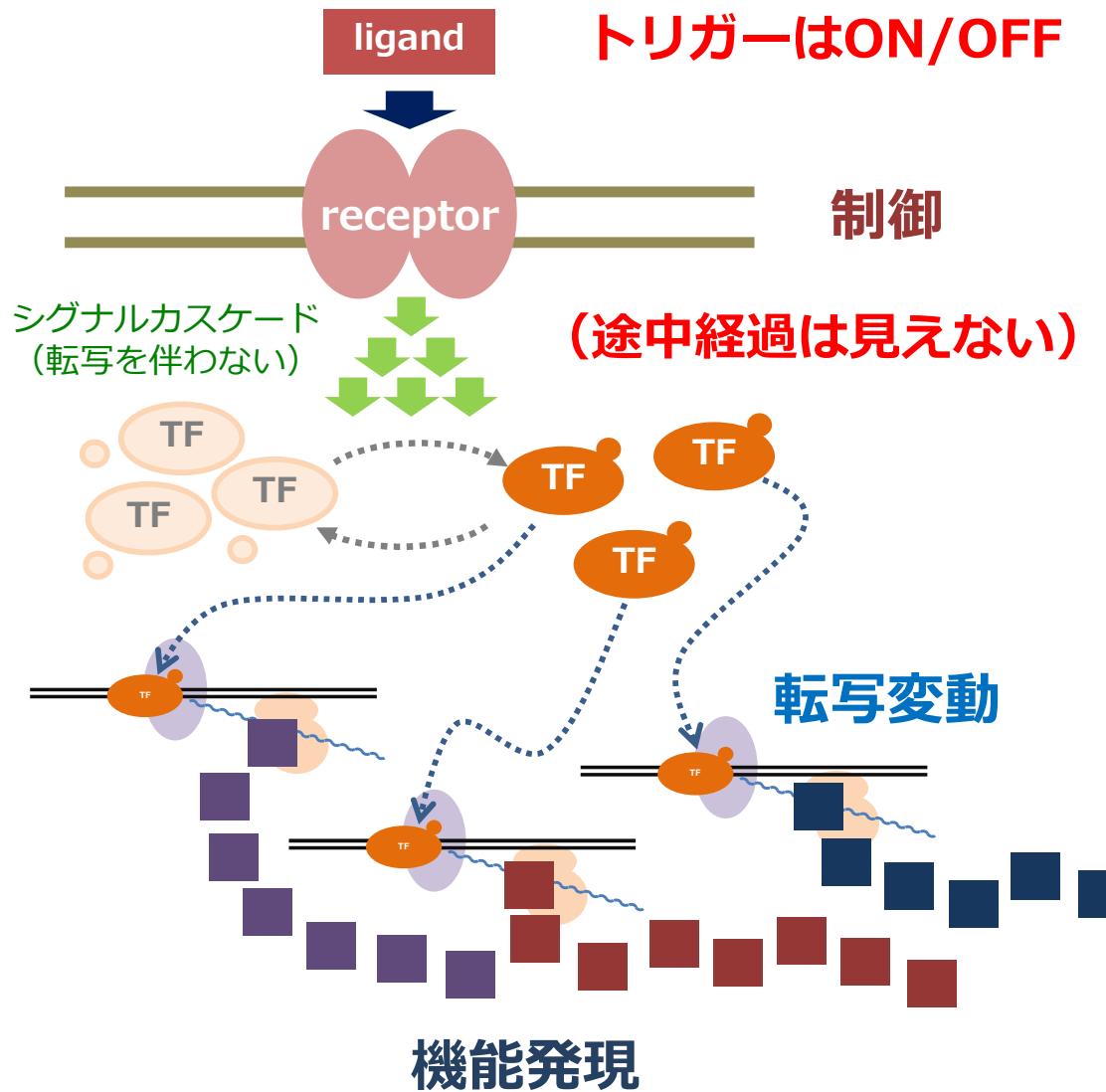
DEE2 (<http://dee2.io/index.html>) での評価を集計 (2021.5.28現在)

計測値の感度と機能の変化の関係



乳がんセルライン (MCF7) のヒストグラム

(Agilent Technologies社資料より)

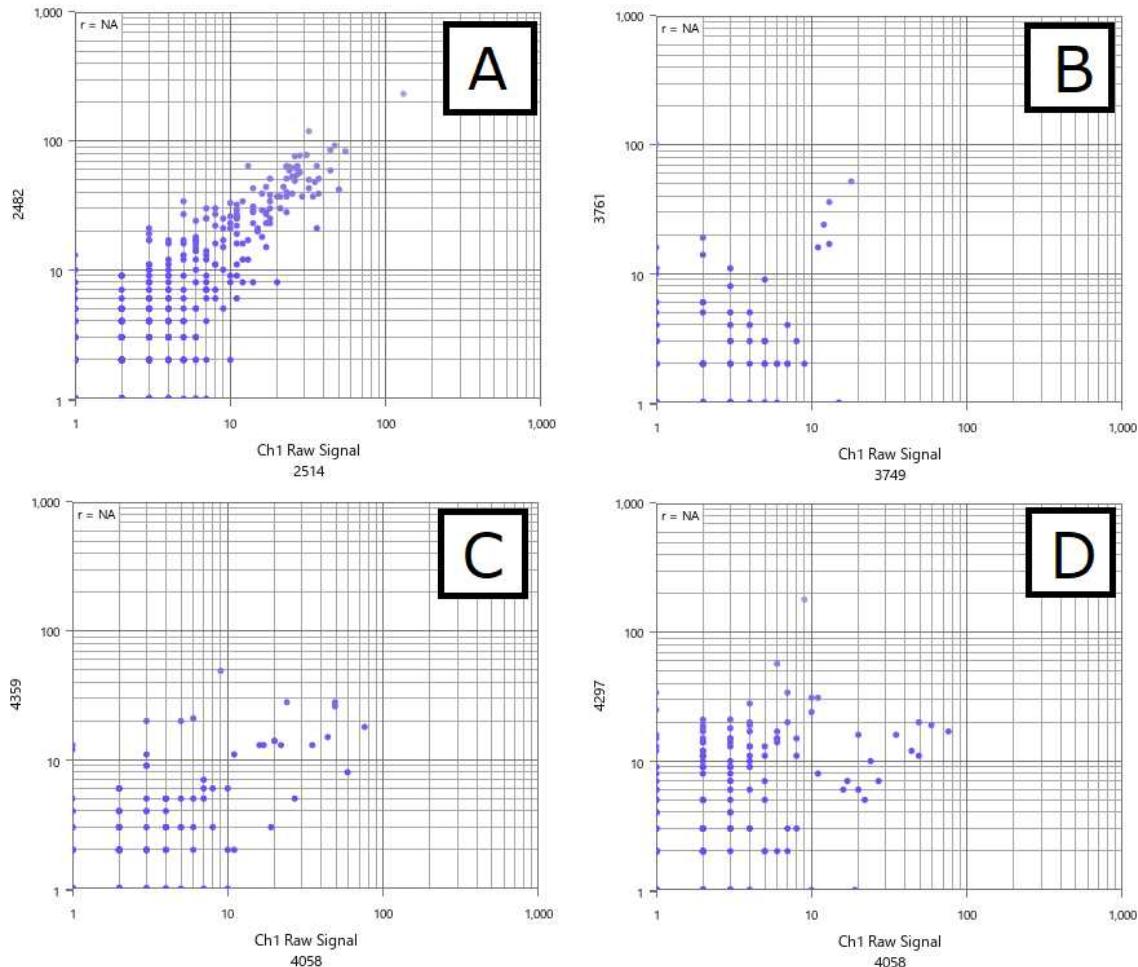


発現変動のダイナミックレンジは 10^6

下流ほど遺伝子数も多く変動幅も大きい

ssRNA-seqデータの信頼区間

ssRNA-seqデータの分散図の例(GSE134520)



GSE134520 (Gastric cancer)

ssRNA-seqのCount

13細胞サンプル中少なくとも1細胞でCount値が1以上だったのは15,245遺伝子。少なくとも1細胞でCount値が5以上だったのは1,854遺伝子。さらに、少なくとも2細胞でCount値が10以上だったのはたった301遺伝子だった。半分の細胞でCount値が10を超えた6遺伝子のうち、4遺伝子はミトコンドリアの遺伝子だった。

このデータでは何らかの生物学的解析に資すると言える遺伝子は200~300個しかなく、それですらごく一部の細胞でしか検出されていなかった。

ssRNA-seqでは十分な深さのデータを取得することが、技術的にもコスト的にも困難なことから、その限界（何がわかり、何がわからないのか）をよく理解して使うことが重要。

出展: Subio (https://www.subioplatform.com/ja/info_casestudy/340/how-is-the-quality-of-single-cell-rna-seq-actually)

「群分け」 について考える

「がん悪液質マーカーの探索」

Genome Med. 2010; 2(1): 1.

Using transcriptomics to identify and validate novel biomarkers of human skeletal muscle cancer cachexia

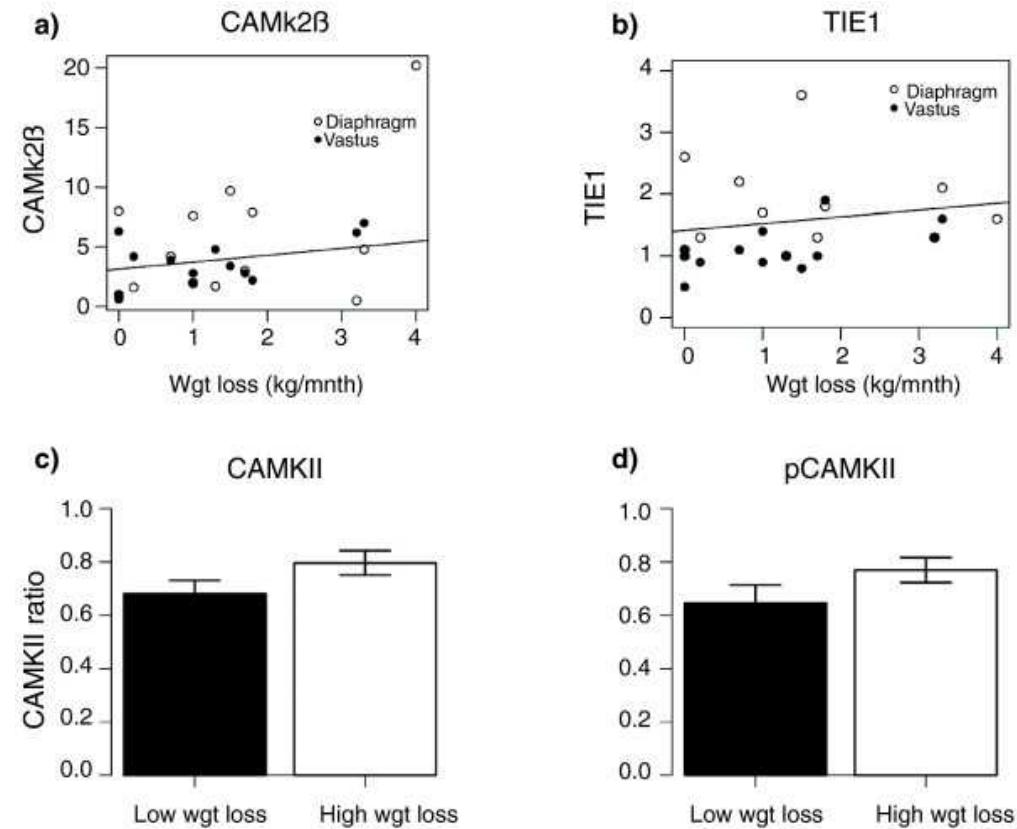
Nathan A Stephens et al. (University of Edinburgh)

Published online 2010 Jan 15. doi: 10.1186/gm122, PMCID: PMC2829926, PMID: 20193046

GeneChipを用いて上部消化器がん患者
腹直筋サンプルの遺伝子発現を調べ、

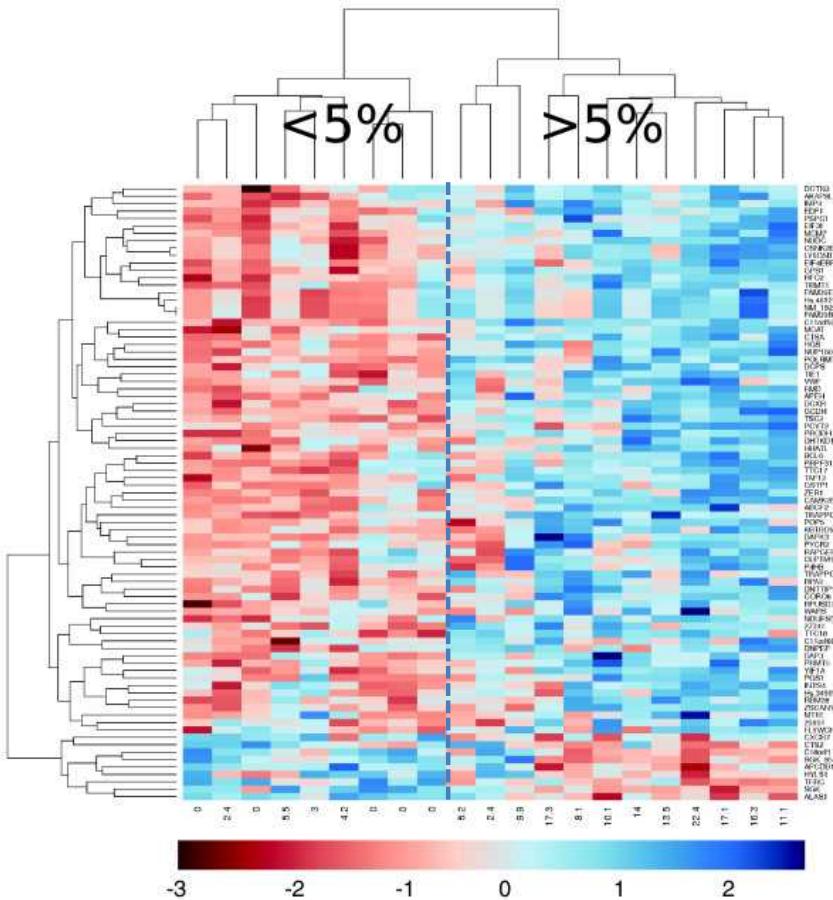
Student's two tailed t-test, one way
ANOVA, Mann-Whitneyで統計学的に
有意な発現差のある遺伝子を抽出。

- 悪液質患者で**CaMKII β** と**TIE1**が
発現上昇。
- 動物モデルの解析で悪液質との関係
が報告された**FOXO**や**ユビキチンE3
リガーゼ (MURF1, MAFbx)**は、
体重減少とは無関係だった。

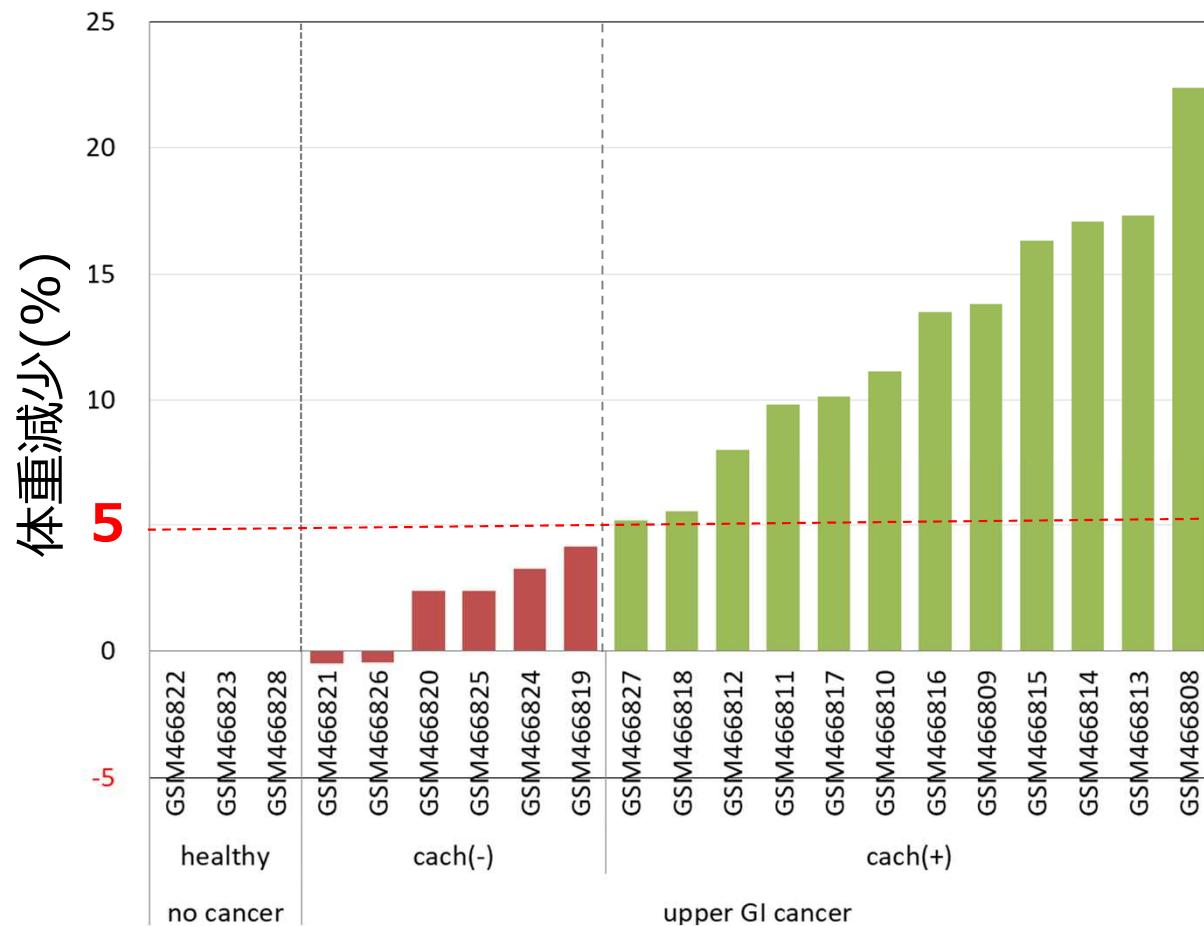


この論文の「群設定」は正しいでしょうか？

Significance Analysis of Microarrays
(SAM) で選択した83遺伝子を用いた
クラスタリング



- 悪液質あり（体重減少5%以上） 12検体
- 悪液質なし（体重減少5%未満） 6検体
- 健常人 3検体



「群の不均一性」と「数の力」 について考える

サンプルの多様性(例:「乳がん」)

データセット	乳がん	正常				合計
		乳房	乳腺	間質	乳首	
GSE2109	335	0	0	0	0	335
GSE8977	7	0	0	14	0	21
GSE5460	127	0	0	0	0	127
GSE7307	0	2	3	0	4	9
ca1015897591112000	22	0	0	0	0	22
合計	491	2	3	14	4	514

乳がんサンプルの多様性(1) : ER/PR/Her2

ER	nega	nega	nega	nega	nega	nega	pos	pos	pos	pos	pos	pos	pos	pos	pos	low	unk	unk	NT	合計	
PR	nega	nega	nega	pos	unk	unk	nega	nega	nega	pos	pos	pos	pos	pos	unk	unk	unk	unk			
Her2	nega	pos	unk	nega	nega	pos	nega	pos	unk	nega	pos	unk	nega	pos	unk	unk	nega				
1 Adenocarcinoma, Metastatic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	
2 Adenoid Cystic Carcinoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
3 Carcinoma, Metastatic	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	
4 Colloid Adenocarcinoma	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
5 Cribiform Carcinoma	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	
6 Ductal and Lobular Carcinoma	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	3	
7 Ductal Carcinoma	36	16	7	0	39	15	10	7	3	54	22	6	60	15	0	0	2	97	0	389	
8 Ductal Carcinoma, Metastatic	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	4	
9 Inflammatory Carcinoma	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	
10 Intracystic Carcinoma	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	
11 Intraductal Carcinoma	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	
12 Lobular Carcinoma	0	0	1	1	0	0	6	0	1	11	1	1	0	0	0	0	1	12	0	35	
13 Lobular Carcinoma, Metastatic	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	
14 Medullary Carcinoma	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
15 Metaplastic Carcinoma	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	
16 Metaplastic Squamous Carcinoma	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	
17 Mucinous Carcinoma	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	2	0	6	
18 Papillary Carcinoma	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	2	
19 Pleomorphic Liposarcoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
20 Stroma Tumor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	7	
21 Histology unknown	0	0	9	0	0	0	0	0	1	0	0	10	0	0	1	1	0	0	0	22	
0 Normal	0	0	0	0	0	0	0	0	0	75	25	17	60	15	1	1	3	129	23	23	
合計		42	18	18	1	39	15	19	8	5	75	25	17	60	15	1	1	3	129	23	514

乳がんサンプルの多様性 (2) : ステージ

STAGE	1[P]	2	2a[C]	2a[P]	2a[R]	2b[P]	2b[R]	3	3a[P]	3a[R]	3b[P]	3b[R]	3c[P]	4	4[C]	4[P]	4[R]	unk	NT	合計
1 Adenocarcinoma, Metastatic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
2 Adenoid Cystic Carcinoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
3 Carcinoma, Metastatic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
4 Colloid Adenocarcinoma	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5 Cribriform Carcinoma	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2
6 Ductal and Lobular Carcinoma	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	3
7 Ductal Carcinoma	21	0	2	58	1	31	2	0	22	2	3	0	14	0	1	4	0	228	0	389
8 Ductal Carcinoma, Metastatic	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	4
9 Inflammatory Carcinoma	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	2
10 Intracystic Carcinoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
11 Intraductal Carcinoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4
12 Lobular Carcinoma	4	0	0	6	0	5	0	0	4	0	0	0	3	0	0	0	0	13	0	35
13 Lobular Carcinoma, Metastatic	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	2
14 Medullary Carcinoma	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
15 Metaplastic Carcinoma	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	3
16 Metaplastic Squamous Carcinoma	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2
17 Mucinous Carcinoma	2	0	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	6
18 Papillary Carcinoma	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
19 Pleomorphic Liposarcoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
20 Stroma Tumor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	7
21 Histology unknown	0	8	0	0	0	0	0	6	0	0	0	0	0	8	0	0	0	0	0	22
0 Normal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	23	514
合計	30	8	2	68	2	39	3	6	29	3	5	1	19	8	1	6	3	258	23	514

「乳がん」サンプルの多様性 (3) : クラス

CLASS	1	1[C]	1[P]	2	2[P]	3[C]	3[P]	3[R]	unk	NT	合計
1 Adenocarcinoma, Metastatic	0	0	0	0	0	0	0	1	1	0	2
2 Adenoid Cystic Carcinoma	0	0	1	0	0	0	0	0	0	0	1
3 Carcinoma, Metastatic	0	0	0	0	0	0	0	0	1	0	1
4 Colloid Adenocarcinoma	0	0	1	0	0	0	0	0	0	0	1
5 Cribriform Carcinoma	0	0	0	0	1	0	1	0	0	0	2
6 Ductal and Lobular Carcinoma	0	0	0	0	1	0	1	0	1	0	3
7 Ductal Carcinoma	31	0	17	69	66	3	105	5	93	0	389
8 Ductal Carcinoma, Metastatic	0	0	0	0	1	0	1	2	0	0	4
9 Inflammatory Carcinoma	0	0	0	0	0	0	2	0	0	0	2
10 Intracystic Carcinoma	0	0	0	0	1	0	0	0	0	0	1
11 Intraductal Carcinoma	0	0	0	0	1	0	0	0	3	0	4
12 Lobular Carcinoma	0	1	4	0	20	0	1	0	9	0	35
13 Lobular Carcinoma, Metastatic	0	0	0	0	0	0	0	0	2	0	2
14 Medullary Carcinoma	0	0	0	0	0	0	1	0	0	0	1
15 Metaplastic Carcinoma	0	0	0	0	0	0	1	1	1	0	3
16 Metaplastic Squamous Carcinoma	0	0	0	0	0	0	2	0	0	0	2
17 Mucinous Carcinoma	0	0	3	0	3	0	0	0	0	0	6
18 Papillary Carcinoma	0	0	1	0	1	0	0	0	0	0	2
19 Pleomorphic Liposarcoma	0	0	0	0	0	0	0	0	1	0	1
20 Stroma Tumor	0	0	0	0	0	0	0	0	7	0	7
21 Histology unknown	0	0	0	0	0	0	0	0	22	0	22
0 Normal	0	0	0	0	0	0	0	0	0	23	23
合計	31	1	27	69	95	3	115	9	141	23	514

乳がんサンプルの多様性 (5)

他にも・・・

- ・採取部位
- ・発症からの時間
- ・リンパ節転移の有無
- ・家族歴
- ・年齢
- ・性別

⋮

- ・“X”
- ・“Y”
- ・“Z”

⋮

人為的な要因

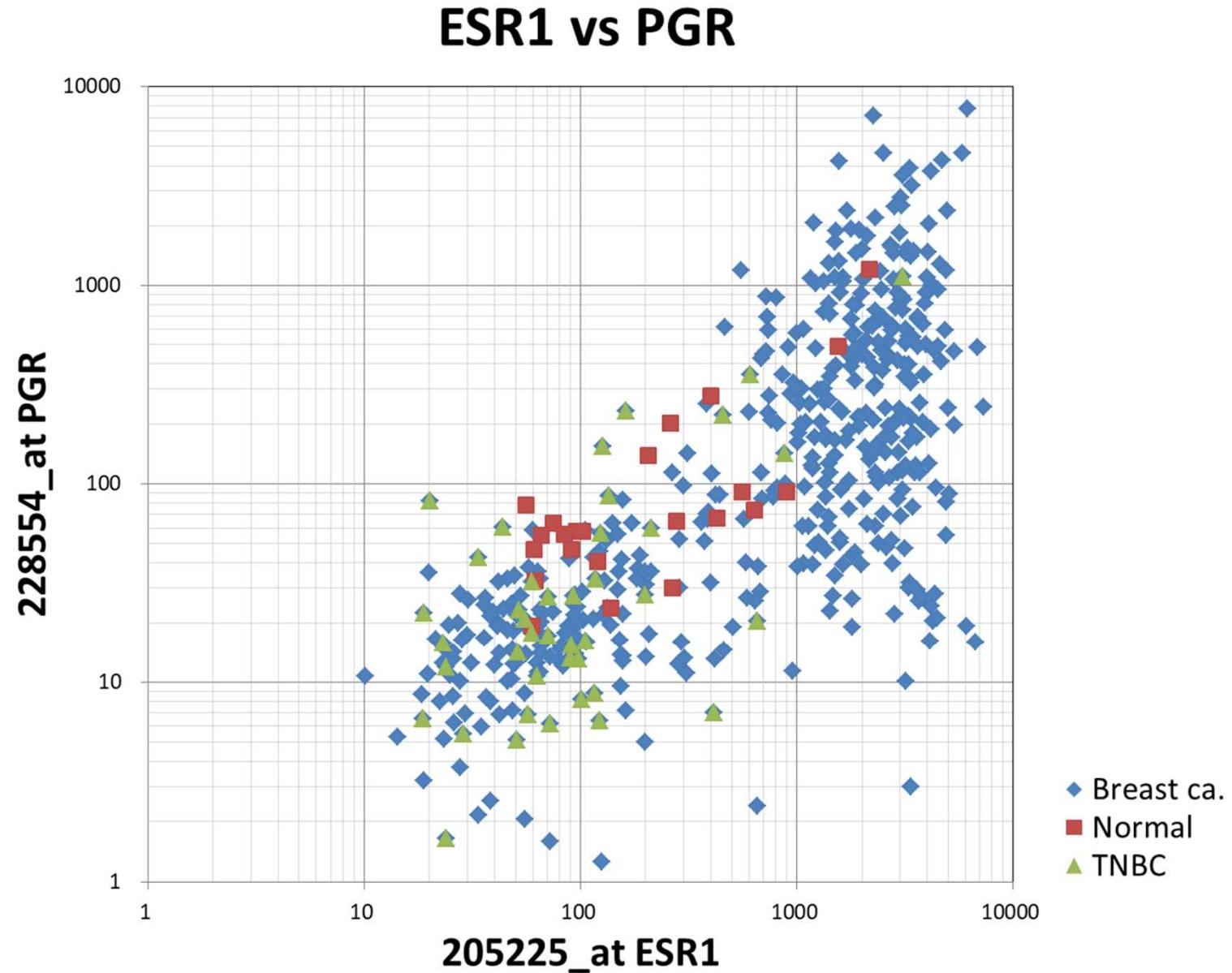
- ・周辺組織の混入度合い
- ・血球細胞の混入度合い
- ・摘出から凍結までの時間
- ・RNAの品質
- ・RNAの抽出方法
- ・サイト・実験者の違い

⋮

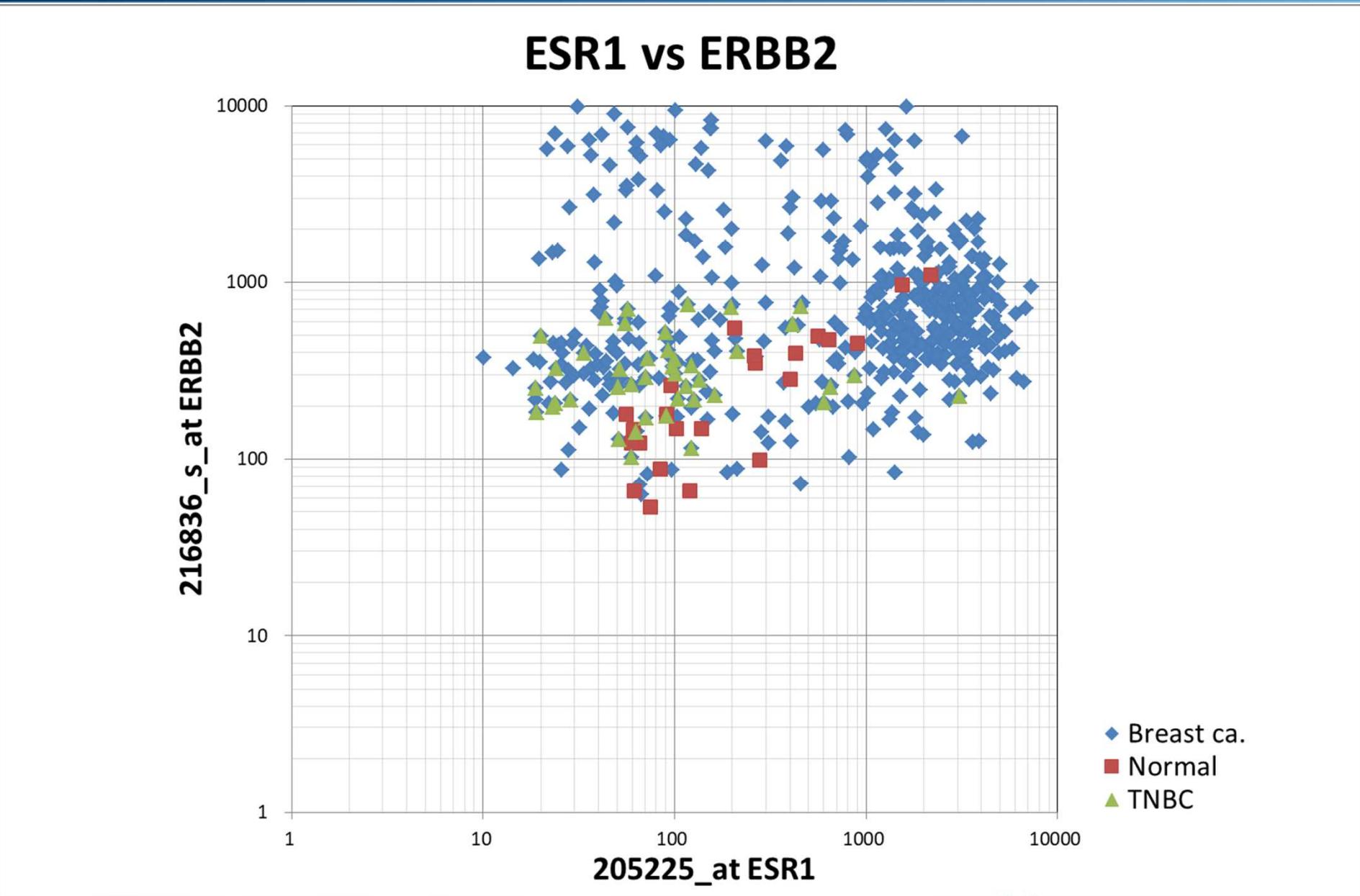
そもそも、臨床サンプルでは、
解析対象を「均質な集団」にすることは
きわめて困難。

→ 「不均一な集団」であることを
前提に、どのように扱うべきかを
考える。

ESR1とPGRの発現相関



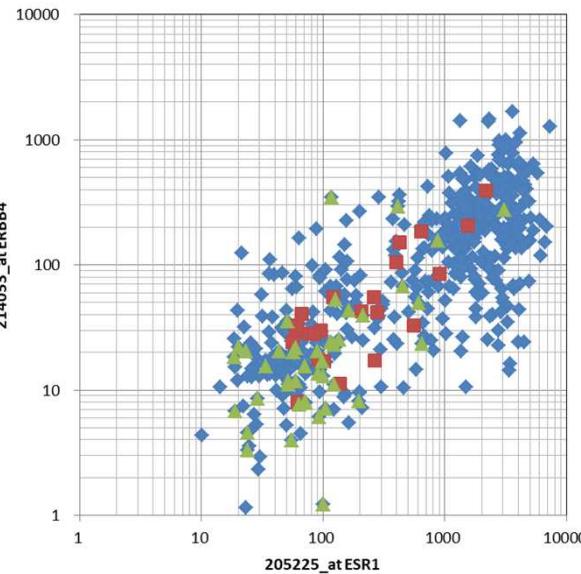
ESR1とHer2(ERBB2)の発現



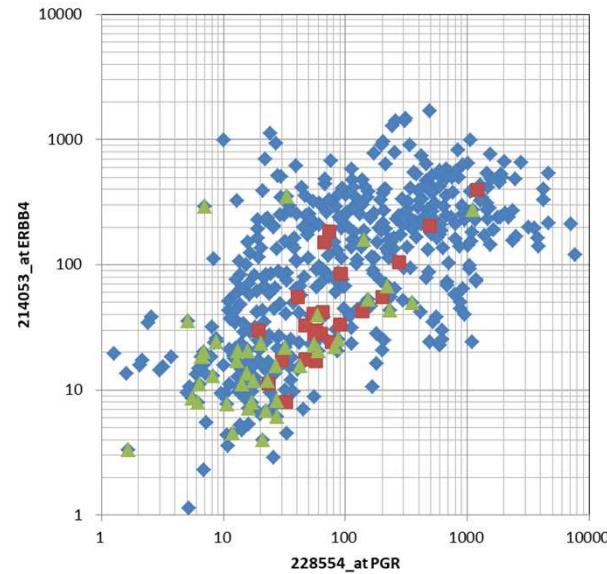
ERBB4の発現

ESR1やPGRとERBB4の発現が相関？

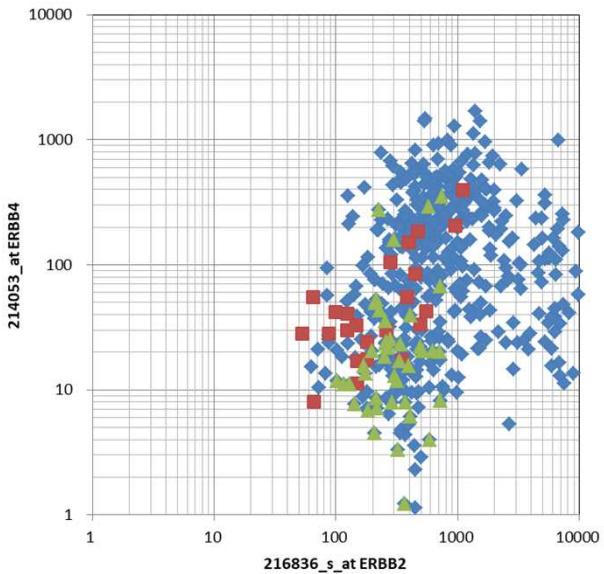
ESR1 vs ERBB4



PGR vs ERBB4



ERBB2 vs ERBB4



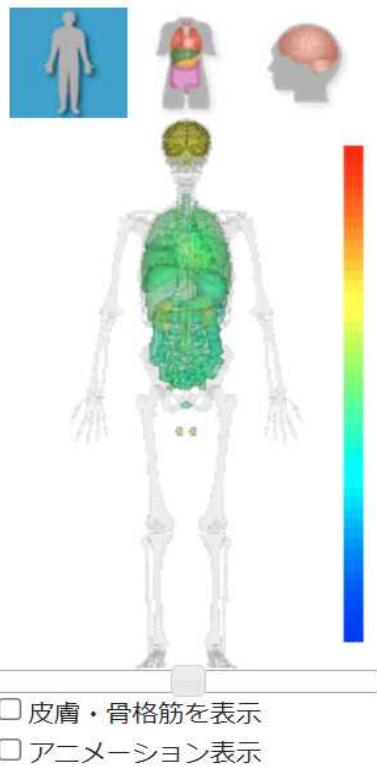
多様性があるからこそ見えること
データがたくさんあるからこそ見えること

ERBB4が発現している組織と機能

RefEx (<https://refex.dbcls.jp/>)

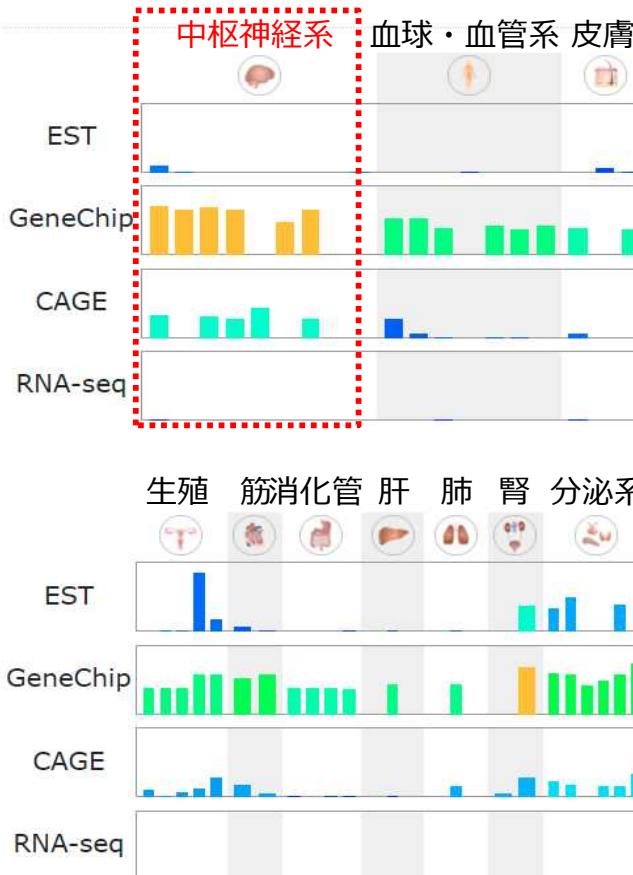
発現マップ on BodyParts3D

相対発現量を、人体 3D 画像にマップしたものです。
Genechip 組織40分類 の発現パターンを使用しています。



組織40分類別データ

[Download](#)



ERBB4

- Neuregulinや他の成長因子により活性化され、分化や細胞増殖を誘導
- 発生、中枢神経系、様々な組織・器官の分化・増殖に関与
- 複数の構造異性体が存在
- 統合失調症・双極性障害の発症に関与（リスク因子）
- 乳がんに関連する変異

【統合TV】RefExの使い方はこちらへ

1) はじめてのRefEx(Reference Expression dataset) (3分2秒) [2021-01-23]

<https://togotv.dbcls.jp/20100618.html>

ヒトやマウス遺伝子の解剖学的な発現パターンデータの統合サイトで、4種類の発現データ (EST, GeneChip, iAFLP, CAGE)に対し、NCBIのRefSeqでデータを整理し、遺伝子発現データ解析のリファレンス（参照）データセットを収載しています。対象生物種はヒトとマウスのみですが、複数の手法による客観的な遺伝子発現データの比較が可能となっているのが特徴です。「検索」ボタンから「Descriptionから遺伝子ファミリー(Interpro ID)を検索」する機能を用いて、Wntスーパーファミリーを付与されたInterProのアノテーションの中から検索し、Wntファミリーに属する遺伝子群をその発現パターンと共に表示し可視化する方法を説明しています。次に、囊胞性線維症の責任領域として知られている7q31.2領域のエントリを発現の高い順に表示する方法を紹介します。さらに、その中のNM_000245 met proto-oncogeneに関して遺伝子の情報の詳細（絶対発現量、相対発現量、3D人体マップ上の発現量のヒートマップ表示）をブラウズする方法を説明しています。

2) RefExの使い方 (7分57秒) [2014-2-22]

<https://togotv.dbcls.jp/20140222.html>

RefEx (Reference Expression dataset) は、4つの異なる実験手法 (EST, GeneChip, CAGE, RNA-seq) によって得られた40種類の正常組織における遺伝子発現データを統合し並列に表現することで、手法間の比較とともに各遺伝子の発現量を直感的に比較することが可能なリファレンス(参照)データセットです。論文を読んでいて見かけた馴染みのない遺伝子がどの組織・臓器で発現しているのか、どんな特徴があるのかを論文の記述ではなく、バイアスの少ない実際の測定データから研究者自身の目で簡単に確認することができます。また、RefExではさまざまな実験における比較対照などに用いられる『組織特異的遺伝子』を測定データから独自に算出し、組織ごとに一覧することができます。さらに、リスト機能を用いて任意で選んだ最大3つの遺伝子について発現データを含む全ての詳細データを並列に比較することができ、遺伝子発現解析などで見出された不詳な遺伝子群の関係性を知るためのツールとしても有用です。

バイオインフォマティクスにおける **生物学**

「統計学はとても重要」です。
だから、よく勉強して理解してください。

しかし、

「統計学だけでは生物は理解できない」
ということも忘れないでください。

だから、

「生物学」

も勉強してください。

NCBI GEOから 遺伝子発現データを取ってくる

NCBI GEOでの検索

(<https://www.ncbi.nlm.nih.gov/geo/>)

生物種、疾患名、組織、細胞名などで探すことが出来ます。

① 検索クエリ窓

② 検索結果

③ クリック!

④ 検索結果リスト

【略号の意味】

GSE : Series (実験セット)

GPL : Platform (測定プラットフォーム)

GSM: Data (サンプルデータ)

GDS: Data Set (NCBIがまとめた解析単位)

NCBI Resources How To Sign in to NCBI

GEO Home Documentation Query & Help

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started Tools Browse Content

diabetes human kidney Search

There are 1772 results for "diabetes human kidney" in the GEO DataSets Database.

There are 40084 results for "diabetes human kidney" in the GEO Profiles Database.

DataSets: 4348 Series: 79636

Information for Submitters Login to Submit Submission Guidelines MIAME Standards

Sign in to NCBI

GEO DataSets GEO DataSets diabetes human kidney Create alert Advanced

Search results Items: 1 to 20 of 1772

Dietary fatty acid composition effect on *vastus lateralis muscle*

Analysis of skeletal muscle (SM) from insulin resistant men 4 hrs after consumption of meals high in saturated FA (SFA), monounsaturated FA (MUFA), or polyunsaturated FA (PUFA). Results provide insight into mechanisms underlying effects of FA composition on SM FA handling and insulin sensitivity.

Organism: Homo sapiens Type: Expression profiling by array, transformed count, 10 individual, 3 protocol, 2 series, 1 sample

Platform: GPL1281 Series: GSE1901 56 Samples

Download data: GEO (CEL) Database: Accession: GDS4412 ID: 4412 PubMed: Similar studies: GEO Profiles Analyze DataSet

Diabetic OVE26 glomerulus

Analysis of glomeruli isolated from kidneys of 8 week old, diabetic OVE26 males. The OVE26 type 1 mouse is a model of progressive glomerulosclerosis and decline of renal function. Results provide insight into the pathogenic mechanisms linked to diabetic nephropathy in the OVE26 model.

Organism: Mus musculus Type: Expression profiling by array, transformed count, 2 disease state sets Platform: GPL1281 Series: GSE20944 7 Samples

Download data: GEO (CEL) Database: Accession: GDS3992 ID: 3992 PubMed: Full text in PMC: Similar studies: GEO Profiles Analyze DataSet

Pancreatic beta cells and panel of primary tissues

Analysis of adult pancreatic beta cell-enriched fractions and duct cell-enriched fractions and other human primary tissues. Results provide insight into conserved beta cell markers for assessment of beta cell phenotypic.

Organism: Homo sapiens Type: Expression profiling by array, count, 17 tissue sets Platform: GPL08 Series: GSE30903 33 Samples

Download data: GEO (CEL) Database: Accession: GDS3996 ID: 3996 PubMed: Full text in PMC: Similar studies: GEO Profiles Analyze DataSet

Resveratrol (resVida) effect on obese men: *vastus lateralis muscle* biopsies

Analysis of vastus lateralis muscle from obese men after 30 days of resveratrol supplementation. Resveratrol, a caloric restriction mimetic, significantly reduced sleepiness and resting metabolic rate. Results provide insight into molecular mechanisms underlying the metabolic effects of resveratrol.

Organism: Homo sapiens Type: Expression profiling by array, transformed count, 2 agent sets Platform: GPL1152 Series: GSE32357 20 Samples

JST Japan Science and Technology Agency 89

NCBI Resources How To

GEO DataSets diabetes human kidney Help

Entry type
DataSets (0)
Series (60)
Samples (1,063)
Platforms (18)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling array
Customize ...

Author
Customize

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ Filters: Manage Filters

Items: 1 to 20 of 1772 Page 1 of 89 First <Prev Next > Last >>

Dietary fatty acid composition effect on vastus lateralis muscle

Analysis of skeletal muscle (SM) from insulin resistant men 4 hrs after consumption of meal high in saturated FA (SFA), monounsaturated FA

Mus musculus (1474)
Homo sapiens (292)
Rattus norvegicus (6)

⑤ Seriesをクリック

⑥ 実験セット（GSE）のリスト

データの種類で絞り込み

strain (1,467)
Customize ...
Publication dates
30 days
1 year
Custom range...

✓ Series (60)
Samples (0)
Platforms (0)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

絞り込み検索

Japan Science and Tech

ここから再検索も出来ます

生物種での絞り込み

⑦ 選択した実験セット（GSE）をクリック
(「右クリック>新しいタブで開く」がお薦め)

⑧ 実験セット（GSE）の情報

NCBI Resources How To

GEO DataSets diabetes human kidney Help

Find items [MeSH Terms]

NCBI GEO Accession Display

Series GSE53119 Status Public on Dec 05, 2016

Title A novel HbA1c-lowering traditional Chinese medicinal formula, identified by translational medicine study

Organism Mus musculus

Experiment type Expression profiling by array

Summary CYSKT affected the expressions of genes associated with insulin signaling pathway, increased the amount of phosphorylated insulin receptor in cells and tissues, and stimulated the translocation of glucose transporter 4 to the cell membrane. Moreover, CYSKT affected the expressions of genes related to diabetic complications, improved the levels of renal function indexes, and increased the survival rate of diabetic mice.

Recent activity

diabetes hu [Filter] (60)

diabetes hu [Filter] (60)

Platforms (1) GPL6845 Phalanx Mouse OneArray

Series (6) GSM1282668 Mock_1

al More... GSM1282669 Mock_2

GSM1282670 Mock_3

GSE7032 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7032>)

Series GSE7032 Query DataSets for GSE7032

Status Public on May 31, 2007
Title Brown and white adipocyte differentiation
Organism *Mus musculus*
Experiment type Expression profiling by array
Summary Attainment of a brown adipocyte cell phenotype in white adipocytes, with their abundant mitochondria and increased energy expenditure potential, is a legitimate strategy for combating obesity. The unique transcriptional regulators of the primary brown adipocyte genotype are...

Overall design Comparisons of white and brown pre- and mature-adiposites

Contributor(s) Larsson O, Timmons JA
Citation(s) Timmons JA, Wennmalm K, Larsson O, Walden TB et al. Myogenic gene expression signature establishes that brown and white adipocytes originate from distinct cell lineages. *Proc Natl Acad Sci U S A* 2007 Mar 13;104(11):4401-6. PMID: 17360536

Submission date Feb 14, 2007
Last update date Feb 18, 2018
Contact name Ola Larsson

Platforms (1) [GPL81 \[M_3_U74Av2\]](#) Affymetrix Murine Genome U74A Version 2 Array

Samples (24)
More...
GSM162532 Primary brown adipocytes 4 days in culture B_y_05
GSM162533 Primary brown adipocytes 4 days in culture B_y_06
GSM162534 Primary brown adipocytes 4 days in culture B_y_09

Relations
BioProject PRJNA98391

Analyze with GEO2R

Download family
SOFT formatted family file(s)
MINiML formatted family file(s)
Series Matrix File(s)

Format
SOFT
MINiML
TXT

Supplementary file
GSE7032_RAW.tar
Raw data provided as supplementary file

Size 64.3 Mb
Download (<http://>(custom))

File type/resource TAR (of CEL)

→ 原著論文へのリンク

① プラットフォーム (GPL81) のアノテーション
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL81>

→ BioProjectへのリンク
<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA98391>

GEO2Rでの解析（変動遺伝子抽出）

⇒ 統合TV (<https://togotv.dbcls.jp/20210213.html>) 参照

② マトリクス形式データ

③ 生データ（数値化から再解析する場合）



① Series matrix file

実験セット情報

..... サンプル

サンプル情報

正規化データ

←… ピローピ… (4~6万行)

⇒ 実験セット（GSE）に関する情報

- 登録日、公開日
- 研究者の所属、連絡先
- 関連論文、等

⇒ サンプル（GSM）に関する情報
生物種、組織、細胞、実験条件等
RNA抽出方法、プラットフォーム、
実験条件、使用機器、正規化法、
研究者所属、連絡先、等

⇒ 正規化されたデータ
MAS5で正規化されているが、
このケースもフラグ情報はない

ダウンロードしたgzファイルを解凍した後、
Tクセルで開いたもの

② GPL81の情報

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	#ID = Affymetrix Probe Set ID																								
2	#GB_ACC = GenBank Accession Number																								
3	#SPOT_ID = identifies controls																								
4	#Species Scientific Name = The genus and species of the organism represented by the probe set.																								
5	#Annotation Date = The date that the annotations for this probe array were last updated. It will generally be earlier than the																								
6	#Sequence Type =																								
7	#Sequence Source = The database from which the sequence used to design this probe set was taken.																								
8	#Target Description =																								
9	#Representative Public ID = The accession number of a representative sequence. Note that for consensus-based probe sets																								
10	#Gene Title = Title of Gene represented by the probe set.																								
11	#Gene Symbol = A gene symbol, when one is available (from UniGene).																								
12	#ENTREZ_GENE_ID = Entrez Gene Database UID																								
13	#RefSeq Transcript ID = References to multiple sequences in RefSeq. The field contains the ID and Description for each ent																								
14	#Gene Ontology Biological Process = Gene Ontology Consortium Biological Process derived from LocusLink. Each annota																								
15	#Gene Ontology Cellular Component = Gene Ontology Consortium Cellular Component derived from LocusLink. Each annota																								
16	#Gene Ontology Molecular Function = Gene Ontology Consortium Molecular Function derived from LocusLink. Each annota																								
17	ID	GB_ACC	SPOT_ID	Species	S	Annotation	Sequence	Sequence	Target	Det	Represent.	Gene	Title	Gene	Sym	Ent									
18	100001_at	M18228		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc M18228	CD3 antigen	Cd3g															
19	100002_at	X70893		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X70893	inter-alpha	Ith3															
20	100003_at	D38216		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc D38216	ryanodine	Ryr1															
21	100004_at	AW120890		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AW120890	integrator	Jnts7															
22	100005_at	X92346		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X92346	TNF recep	Traf4															
23	100006_at	D21253		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc D21253	cadherin 1	Cdh11															
24	100007_at	AI837573		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI837573	interferon	Irf2bp1															
25	100009_r_a	X94127		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X94127	SRY (sex	Sox2															
26	100010_at	U36340		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc U36340	Kruppel-like	Klf13															
27	100011_at	AI851658		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI851658	Kruppel-like	Klf3															
28	100012_at	U29539		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc U29539	lysosomal	Laptm5															
29	100013_at	AW121732		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AW121732	interferon-	Ifi35															
30	100014_at	AI845038		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI845038	tousled-like	Tlk2															
31	100015_at	X67677		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X67677	Yamaguchi	Yes1															
32	100016_at	Z12604		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc Z12604	matrix met	Mmp11															
33	100017_at	U68267		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc U68267	myosin	bir Mybph															
34	100018_at	X71327		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X71327	metal resp	Mtf1															
35	100019_at	D45889		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc D45889	versican	Vcan															
36	100020_at	J04036		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc J04036	solute car	Slc4a2															
37	100021_at	M17640		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc M17640	cholinergic	Chrna1															
38	100022_at	D89613		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc D89613	cytokine	Ir Cish															
39	100023_at	X70472		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X70472	myeloblast	Mybl2															
40	100024_at	AI641895		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI641895	shroom	fai Shroom3															
41	100026_at	U42443		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc U42443	branched	c Bcat1															
42	100027_s_a	X81897		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X81897	peroxisom	Pex14															
43	100028_r_a	AI881987		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI881987	peroxisom	Pex14															
44	100029_at	AW208667		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AW208667	peroxisom	Pex14															
45	100030_at	D44464		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc D44464	uridine	ph Upp1															
46	100032_at	X60136		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X60136	trans-actin	Sp1															
47	100033_at	X81143		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc X81143	mutS hom	Msh2															
48	100034_at	U54705		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc U54705	serine	(or Serpinb5															
49	100035_at	L31932		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc L31932	natriuretic	Npr1															
50	100037_at	AW213225		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AW213225	DEAD	(As Ddx18															
51	100039_at	AW125880		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AW125880	canopy	2 Cnpy2															
52	100040_at	AI843081		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI843081	mitochondri	Mpr17															
53	100041_at	AW124183		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AW124183	solute car	Slc25a39															
54	100042_at	AI837921		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI837921	hydroxyac	Hagh															
55	100043_f_a	AI173973		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc AI173973	prosaposin	Psap1															
56	100044_at	U19582		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc U19582	claudin	11 Cdln11															
57	100046_at	J04627		Mus musci	6-Oct-14	Consensus	GenBank	Cluster Inc J04627	methylene	Mthfd2															

- ID
- GB_ACC
- SPOT_ID
- Species Scientific Name
- Annotation Date
- Sequence Type
- Sequence Source
- Target Description
- Representative Public ID
- Gene Title
- Gene Symbol
- ENTREZ_GENE_ID
- RefSeq Transcript ID
- Gene Ontology Biological Process
- Gene Ontology Cellular Component
- Gene Ontology Molecular Function
- (※ 付加情報はGPLにより異なる)

③ Supplementary file (Raw data)

生データ

- **Affymetrix GeneChip™の.CELファイル**
 - **Agilent MicroarrayのFeature Extractionの.txtファイル、等**
- 論文とは違う正規化法で再解析したり、同じプラットフォームで取られた別のGSEデータと組み合せた解析が可能。
- Agilent MicroarrayやIllumina BeadArrayのデータはテキストファイルから数値が取れる。
- Affymetrix GeneChip™データの正規化には専用の数値化ソフトが必要。
 - **Transcriptome Analysis Console (TAC)Software
(GeneChip™ Expression Console Softwareを含む)**
Thermo Fisherサイトから無償で入手可能（要登録）
<https://www.thermofisher.com/jp/ja/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software.html>
 - BioconductorのAffymetrixデータ処理用パッケージ、等
- **Supplementary fileが公開されていないデータもある。**

【URL】 <https://www.thermofisher.com/jp/ja/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/affymetrix-expression-console-software.html>

Expression Console Software is now part of
Transcriptome Analysis Console (TAC) Software.

TAC Software (version 4.0 and subsequent releases) now includes
**the normalization, probe summarization,
and data quality control functions**
of Expression Console Software.

Manuals

[Expression Console Software](#)

[Sample Attribute Editor](#)

White papers

[Microarray normalization using SST and GCCN](#)

Additional support

[Expression Console Software 1.4.1 release notes](#)

エクセルTIPS

エクセルの準備(設定変更)

【おすすめの設定変更】

ファイル>オプション>詳細設定>編集オプション

- ✓ 「セルを直接編集する」のチェックを外す
(クリックで参照先セル指定するため)
- ✓ 「パーセンテージを自動入力する」のチェックを外す
- ✓ 「オートコンプリートを使用する」のチェックを外す

ファイル>オプション>クイックアクセスバー

- ✓ 「参照元のトレース」、「参照元トレース矢印の削除」 を追加 
- ✓ 「参照先のトレース」、「参照先トレース矢印の削除」 を追加 

クイックアクセスバー

- ✓ (自動保存は解除しておく) 注: マニュアルで適時保存する!

【お好みで】

ファイル>オプション>数式

- ✓ 計算方法の設定: 自動 → 手動 (計算式を含む大きなテーブルを扱う場合)

エクセルTIPS

【操作】

1. 「**Ctrl+C**」（コピー）と「**Ctrl+V**」（ペースト）
2. 「**Ctrl+Z**」（直前の操作の取り消し）と「**Ctrl+Y**」（直前の操作の繰り返し）
3. 「\$」（位置固定符号）と**F4キー**（A1 → \$A\$1 → A\$1 → \$A1 → A1）
4. 連続セル間でのジャンプ移動と選択（「**Ctrl+矢印**」、「**Ctrl+Shift+矢印**」）
5. ワイルドカード（「*」、「?」）、エスケープ（「"」）

【関数】

1. 参照関数（**VLOOKUP**、**HLOOKUP**）
2. 論理関数（**IF**、**AND**、**OR**、**NOT**）
3. 文字列関数（**TRIM**、**LEN**、**LEFT**、**RIGHT**、**SUBSTITUTE**、**&**）
4. 数値関数（**ROUND**、**ROUNDDOWN**、**ROUNDUP**）
5. 数学関数（**SUM**、**PRODUCT**、**QUOTIENT**、**MOD**、**ABS**）
6. 統計関数（**MAX**、**MIN**、**AVERAGE**、**MEDIAN**、**STDEV**、**SMALL**、**LARGE**、**RANK**、**PERCENTILE**、**QUARTILE**、**TTEST**、**PEARSON**、**CORREL**）
7. 数を数える関数（**COUNT**、**COUNTA**、**COUNTIF**、**COUNTIFS**、**SUMIF**）
8. データベース関数（**DAVERAGE**、**DCOUNT**、**DCOUNTA**、**DMAX**、**DMIN**、**DPRODUCT**、**DSTDEV**、**DSTDEVP**、**DSUM**、**DVAR**、**DVARP**、**DGET**）
9. その他（**IFERROR**）

エクセル操作

デモデータ [2] : GSE7032 (脂肪細胞初代培養)

GSE7032(Fc-P).xlsx

変動倍率 (Fc) とt検定のP値 (P) で変動遺伝子を選抜

The screenshot displays the GSE7032(Fc-P).xlsx Excel spreadsheet with several distinct sections highlighted:

- 数値分布**: A green-highlighted section in the upper left containing numerical data.
- 詳細フィルターの条件**: A blue-highlighted section in the upper right showing filter settings for multiple columns.
- GPLシートから参照したプロープ情報**: A green-highlighted section in the lower left containing probe information.
- series matrix のデータ部分**: A red-highlighted section in the lower center containing data from the series matrix.
- 各群平均値**: A blue-highlighted section in the lower middle containing average values for each group.
- 比較群間での平均値変動倍率 t検定のP値 選抜遺伝子**: A large blue-highlighted section on the right containing results of differential expression analysis.
- 複合条件での選抜遺伝子**: A blue-highlighted section on the far right showing selected genes under complex conditions.

【注意】 GSE7032(Exp-Fc) (練習用) .xlsxには計算式を入れています (青色セル)。変動遺伝子選抜用のワーキングファイルでは、計算式を消さないと動作が重くなります。

【練習①】 遺伝子情報のLOOKUP

参照列
(数字を参照させる)

「\$」を上手に使いましょう！

	1	2	3	4	5	6
検索キー	ID FEF	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology	Gene Ontology
#	#A	#B	#C	#D	#E	#F
1	100001 at	Cd3g				
2	100002 at					
3	100003 at					
4	100004 at					
5	100005 at					
6	100006 at					
7	100007	=VLOOKUP(\$B43,'GPL81'!\$A\$20:\$P\$12508,C\$40, FALSE)&"'				
8	100009					
9	100010					
10	100011 at					
11	100012 at					
12	100013 at					
13	100014 at					
14	100015 at					
15	100016 at					
16	100017 at					
17	100018 at					
18	100019 at					
19	100020 at					
20	100021 at					
21	100022 at					

後ろに「&'''」を付けると
空白セルが「#N/A」にならない

検索キー
(列を固定)

参照先
(行と列を固定)

参照列
(行を固定)

- 1) C43セルに「=VLOOKUP(\$B43,'GPL81'!\$A\$20:\$P\$12508,C\$40, FALSE)&"'''」と入力。
- 2) C43セルを選択し、右下の■をG43セルまで引っ張る。
- 3) C43～G43セルを選択し、右下の■をダブルクリックする。
- 4) C43～G12530セルを値に変換（領域が選択されている状態で「Ctrl+C」→「値でペースト」）

【練習②】 数値分布表の作成



- I7セルに「=COUNTA(I\$43:I\$12530)」と入力。
(「(」の後ろで「I43」をクリック、「Ctrl+Shift+↓」でI43～I12530セルを選択し、F4キーを1回押す)
- I7セルを選択し、右下の■をダブルクリックする。
- I8セル（最大値）：「=COUNTA(I\$43:I\$12530)」→「=MAX(I\$43:I\$12530)」
- リボンメニュー（ホーム>数値）で、I8セルの桁数を小数点以下2桁表示にする。
- I8セルを選択し、右下の■をダブルクリックする。
- I9セル：「=COUNTA(I\$43:I\$12530)」→「=PERCENTILE(I\$43:I\$12530,\$H9)」
(「\$H9」の入力は「,」の後ろにカーソルを置いて、H9セルをクリックし、F4キーを3回押す)
- I9セルを選択し、右下の■をつまんでI27セルまで引っ張る。
- I28セル（最小値）：「=COUNTA(I\$43:I\$12530)」→「=MIN(I\$43:I\$12530)」
- I29セル（平均値）：「=COUNTA(I\$43:I\$12530)」→「=AVERAGE(I\$43:I\$12530)」
- I30セル（中央値）：「=COUNTA(I\$43:I\$12530)」→「=MEDIAN(I\$43:I\$12530)」
- I7～I30セルを選択し、右下の■をつまんでAF30まで引っ張る。
- I7～AF30セルを値に変換（領域が選択されている状態で、「Ctrl+C」→「値でペースト」）

【練習③】シグナル閾値と変動倍率を使った選抜

各群の平均値

群平均値				選択された変動遺伝子				0.05			
ID_REF	PreBr(Av)	BD(Ave)	White	ID_REF	#G	#H	PreBr vs. PreWh	PreBr vs. PreWh FCave >=log	PreBr Up 0	PreBr Down 0	ot_selected 0
100001_at	2.72	2.87	3.18	4.50	100001_at	-0.226	0.634				
100002_at					100002_at						
100003_at					100003_at						
100004_at					100004_at						
100005_at					100005_at						
100006_at					100006_at						
100007_at					100007_at						
100009_r_at					100009_r_at						
100010_at					100010_at						
100011_at					100011_at						
100012_at					100012_at						
100013_at					100013_at						
100014_at					100014_at						
100015_at											

1)-a Non differentiated., brown vs. white

FC(Ave.log2)	1.000
TTest P-val	0.050
Up	0
Down	0
Selected	0
Total	12,488

未分化(Pre-)群での選抜

「2」:両側検定
「2」:等分散の2標本

=TTEST(\$I43:\$M43,\$S43:\$Z43,2,2)

=LOG(\$AH43/\$AJ43,2)

=AVERAGE(\$AA43:\$AF43)

=AVERAGE(\$S43:\$Z43)

=AVERAGE(\$N43:\$R43)

=AVERAGE(\$I43:\$M43)

Pre-群とDiff-群平均値の変動比
(Brown群)／(White群)
「2」を底とするLOG値

【練習④】 詳細フィルターによる変動遺伝子の選抜

Brown > White
のフィルター

詳細フィルター設定		
PreBr vs. PreWh		
#G	#H	
PreBr_Up	≥ 1	< 0.05
	#G	#H
PreBr_Down	≤ -1	< 0.05

#G: 「=">="&AN\$32」
#H: 「="<"&AN\$33」

「">="& (参照) 」
で条件を設定

Brown < White
のフィルター

PreBr vs. PreWh		
FC(Ave.log2) TTest P-value		
#G	#H	
PreBr_Up	≥ 1	< 0.05
PreBr_Down	≤ -1	< 0.05

#G: 「=">="&AN\$32」
#H: 「="<"&AN\$33」

フィルター条件は
□の範囲を選択

<選択遺伝子参照用テンプレート>

	PreBr Up	PreBr Down	Not_selected
PreBr_Ave	0.00	0.00	
PreWh_Ave			0.00
Not_selected			0.00

閾値

1)-a Non differentiated, brown vs. white	
FC(Ave.log2)	1.000
TTest P-val	0.050
Up	0
Down	0
Selected	0
Total	12,488

詳細フィルタで条件を満たすプローブを選択後、各群平均値を参照するためのコピペ用テンプレート

選択された変動遺伝子

PreBr vs. PreWh		PreBr vs. PreWh FCave >=log1.0 & P<0.05							
ID_REF		#G	#H	PreBr_Av	PreWh_Av	PreBr_Av	PreWh_Av	PreBr_Av	PreWh_Av
100001_at		-0.226	0.634						
100002_at									
100003_at									
100004_at									
100005_at									

リスト範囲は「\$A\$42:\$BW\$12530」
(#G, #H列を含む範囲)

【練習⑤】 詳細フィルターによる複合条件による選抜

#I: 「=">="&BZ\$32」
#J: 「="<"&BZ\$33」

3)複合的な条件による変動遺伝子選択

例)「分化した
変動条件: 1)~bの条件
「GO Biological Proce
「GO Cellular Compon

文字列は「="* (文字列) *"」で指定
(文字列を「*」で挟む)

#詳細フィルター設定

DifBr vs. DifWh						
FC(Ave.log2)	TTest P-val	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology	Gene Ontology
#I	#J	#B	#C	#D	#E	#F
DifBr_Up	>=1	<0.05		*regulation	*nucleus*	
DifBr_Down	<=-1	<0.05		*regulation	*nucleus*	

詳細フィルタで条件を満たすプローブを選択後、各群平均値を参照するためのコピペ用テンプレート

行がAND条件、列がOR条件
UpとDownの条件を別々に指定する
フィルター条件は□の範囲を選択

変動倍率とP値

1)~b Differentiated., brown vs. white						
	FC(Ave.log2)	TTest P-val	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology
Up	1.000	0.050				
Down	0					
Selected	0					
Total	12,488					

<選択遺伝子参照用テンプレート>						
	FC(Ave.log2)	TTest P-val	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology
0.00	0.000	0	0	0	0	0

1)~b Differentiated., brown vs. white						
	FC(Ave.log2)	TTest P-val	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology
Up	1.000	0.050				
Down	0					
Selected	0					
Total	12,488					

閾値

選択された変動遺伝子						
ID_REF	FC(Ave.log2)	TTest P-val	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology
100001_at	-0.666	0.382				
100002_at	-0.668	0.317				
100003_at	-0.644	0.269				
100004_at	0.082	0.619				
100005_at	0.512	0.00				
100006_at	-0.154	0.512				
100007_at	0.422	0.022				
100009_r_at	2.062	0.005				
100010_at	0.390	0.111				

リスト範囲は「\$A\$42:\$BW\$12530」
(#D, #E, #G, #H列を含む範囲)

遺伝子情報

検索キー	1	2	3	4	5	6
ID_FEF	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology	Gene Ontology	
#A	#B	#C	#D	#E	#F	
100001_at	Cd3g	CD3 antigen, g	0007163 // est	0016020 // me	0004888 // tra	
100002_at	Itih3	inter-alpha try	0010466 // ne	0005576 // ex	0004867 // ser	
100003_at	Ryr1	ryanodine rece	0001666 // res	0005622 // int	0002020 // pro	
100004_at	Ints7	integrator com	0000077 // DN	0005634 // nu	0005488 // bin	
100005_at	Traf4	TNF receptor a	0006915 // ap	0005634 // nu	0004842 // ubi	
100006_at	Cdh11	cadherin 11	0007155 // cel	0005737 // cyt	0005509 // cal	
100007_at	Irf2bp1	interferon regul	0000122 // ne	0005634 // nu	0003714 // tra	
100009_r_at	Sox2	SRY (sex deter	0000122 // ne	0005634 // nu	0000976 // tra	
100010_at	Klf3	Kruppel-like fa	0006351 // tra	0005634 // nu	0003676 // nu	

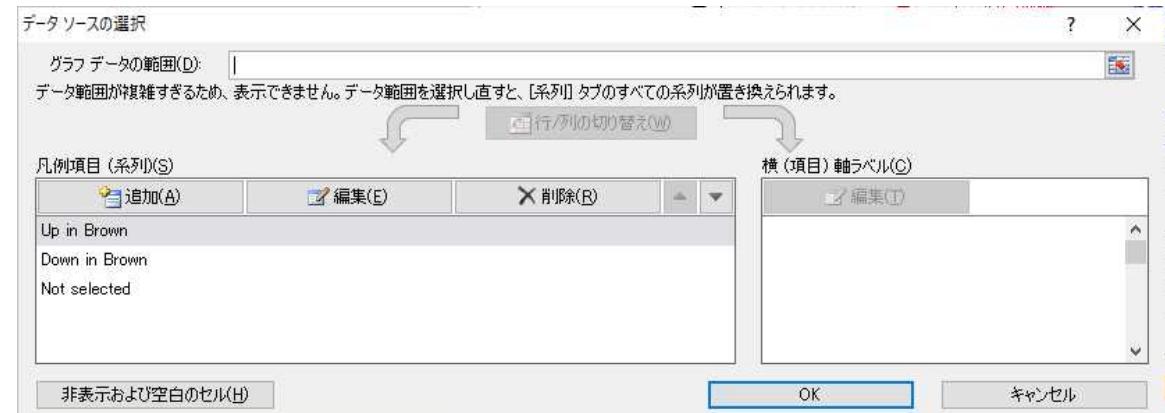
分散グラフの作成(GSE7032)

ID_REF	DifBr vs. DifWh		DifBr vs. DifWh FCave >=log1.0 & P<0.05			
	#I	#J	DifBr Up 364	DifBr Down 649	Not selected 11475	
100001_at	-0.666	0.382			2.87	4.56
100002_at	-0.668	0.317			15.12	24.03
100003_at	-0.644	0.269			11.56	18.07
100004_at	0.082	0.619			47.27	44.65
100005_at	0.512	0.001			319.25	223.90
100006_at	-0.154	0.512			25.62	28.51
100007_at	0.422	0.022			205.06	153.04
100009_r_at	2.062	0.005	15.38	3.68		
100010_at	0.390	0.111			50.64	38.65
100011_at	-1.210	0.002		41.65	96.37	
100012_at	-1.010	0.140			17.69	35.63
100013_at	0.344	0.160			210.04	165.49
100014_at	-0.203	0.317			39.98	46.01
100015_at	0.195	0.203			73.58	64.26
100016_at	-0.411	0.176			310.85	413.36
100017_at	0.929	0.002			62.97	33.06
100018_at	0.661	0.033			33.22	21.02
100019_at	1.208	0.071			70.23	30.40
100020_at	-0.518	0.005			280.22	401.25
100021_at	1.763	0.083			11.79	3.47
100022_at	0.099	0.589			94.15	87.88
100023_at	0.907	0.164			32.64	17.41
100024_at	-0.328	0.109			72.80	91.39
100026_at	0.011	0.940			236.24	234.44
100027_s_at	0.173	0.181			469.34	416.27
100028_r_at	1.361	0.001	102.31	39.83		
100029_at	0.545	0.004			712.61	488.32
100030_at	3.561	0.000	71.29	6.04		
100032_at	-1.687	0.000		20.88	67.21	
100033_at	0.805	0.001			82.96	47.48
100034_at	-0.214	0.257			1.45	1.68
100035_at	-0.0				6.88	6.99
100037_at	0.8				87.58	47.24
100039_at	-0.1				542.73	592.84
100040_at	-0.0				222.93	234.60
100041_at	0.7				1345.64	809.65
100042_at	0.4				239.46	178.24
100043_f_at	0.142	0.376			89.51	81.10
100044_at	-0.778	0.151			0.00	13.73
100046_at	0.612	0.002			85.89	
100047_at	-1.137	0.287			5.31	
100048_at	0.110	0.665			28.92	
100049_at	-1.634	0.089			4.43	
100050_at	0.833	0.000			14.36	
100051_at	-0.114	0.465			39.43	
100052_at	0.082	0.748			3.79	3.58
100054_s_at	0.418	0.121				

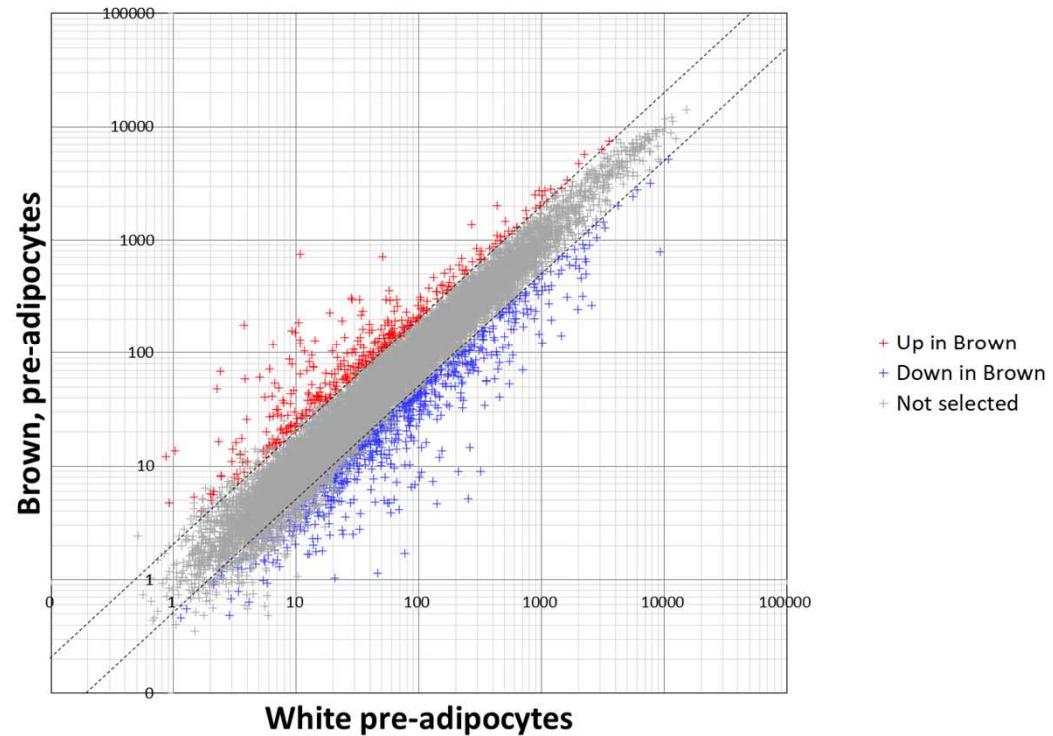
Up in Brown
のデータ範囲

Down in Brown
のデータ範囲

Not selected
のデータ範囲



Brown vs. White, adipocytes (FC and P)



個別グラフの作成(GSE7032)

検索キー

1	ID_REF	100009_r_at
2	Gene Symbol	Sox2
3	Gene Title	SRY (sex determining region Y)-box 2
4	Gene Ontology Biological Process	0000122 // negative regulation of transcription from RNA polymerase II promoter // inferred from genetic interaction // GO process
5	Gene Ontology Cellular Component	0005634 // nucleus // inferred by curator // GO:0005634 // nucleus // inferred from direct assay // GO:0005634
6	Gene Ontology Molecular Function	0000976 // transcription regulatory region sequence-specific DNA binding // not recorded // GO:000981 // sequence-specific DNA binding
7	Cell type	Differentiation
8	Brown adipocytes	Non-differentiated
9		GSM162532
10		8.68
11		14.64
12		5.93
13	Differentiated	GSM162536
14		15.01
15		GSM162537
16		24.89
17		6.71
18	White adipocytes	Non-differentiated
19		GSM162542
20		11.92
21		9.62
22		GSM162545
23	Differentiated	3.38
24		GSM162546
25		10.42
26		GSM162547
27		8.99
28	Non-differentiated	GSM162549
29		8.34
30		GSM162550
31		1.12
32		0.97
33	Differentiated	6.03
34		7.56
35		0.65
36		5.77
37		5.77

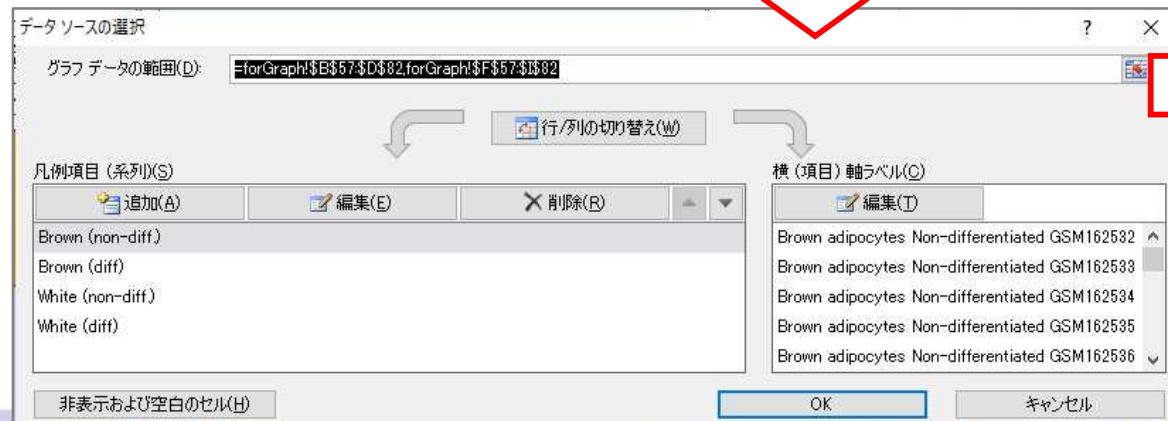
GSE7032(Fc-P).xlsx

データシートから
VLOOKUPで参照

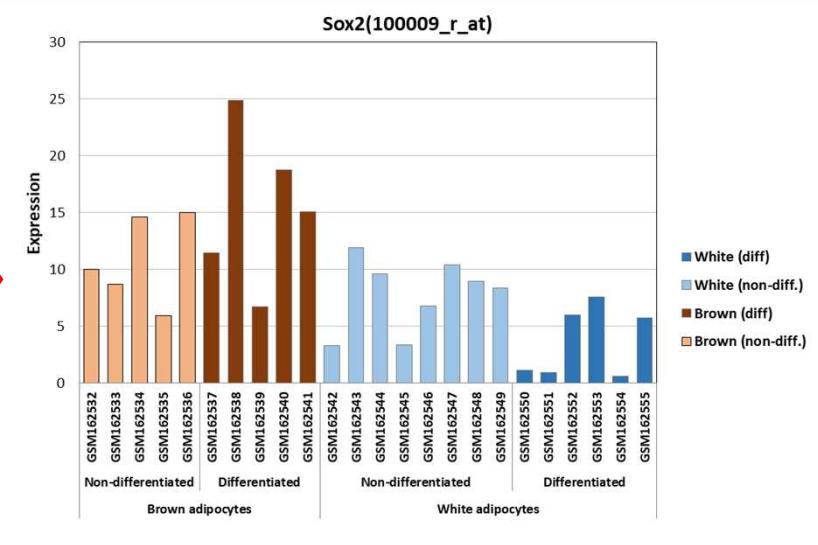
データ範囲

セル参照

データシートから
VLOOKUPで参照



積み重ね棒グラフ



データ [1] : GSE21422 (乳管がん患者)

GSE21422(Exp-Fc).xlsx

シグナル値 (Exp) と変動倍率 (Fc) の閾値を超えたサンプル数で変動遺伝子を選択

The screenshot shows a complex Excel spreadsheet with multiple tabs and data sections. Several sections are highlighted with colored boxes:

- 数値分布**: A green box highlighting a section of the main data area.
- 分布**: A green box highlighting a section of the main data area.
- フィルター条件**: A blue box highlighting a section of the main data area.
- GPLシートから
参照した
プローブ情報**: A green box highlighting a section of the main data area.
- series matrix
のデータ部分**: A red box highlighting a section of the main data area.
- 各群
median値**: A blue box highlighting a section of the main data area.
- Healthy群median値
に対する変動 (log2)**: A blue box highlighting a section of the main data area.
- 選抜結果**: A blue box highlighting a section of the main data area.

【注意】 GSE21422(Exp-Fc) (練習用) .xlsxには計算式を入れています (青色セル)。変動遺伝子選択用のワーキングファイルでは、計算式を消さないと動作が重くなります。

変動倍率とP値の閾値を超えたサンプル数による選抜

=COUNTIFS(
 \$O51:\$W51,">="&BD\$20,
 \$AM51:\$AU51,">="&BD\$21)

「">="&(参照)」で条件を設定

各サンプルの
シグナル値

シグナル

	GSM53561	GSM53560	GSM53560
healthy	breast	1, bi	DCIS 2, bi
healthy	DCIS	DCIS	
breast	tumor	tumor	
healthy	DCIS	DCIS	
GSM53561	GSM53560	GSM53560	

各群のメアン値

変動倍率
(log2)

healthyメアン値に
対するサンプルごとの
変動倍率 (log2)

	disease state	healthy	DCIS	IDC
probeset	Median	Median	Median	
531.6	1230.6	1871.3	2,659.0	
103.0	157.6	123.1		
45.1	130.9	42.4		
54.9	45.3	60.4		
4.9	4.9	4.8		
159.1	128.4	97.1		
12.3	10.5	10.2		
7.3	6.4	6.3		
328.4	590.8	285.8		
6.9	7.1	6.8		
1431_at				

シグナルと変動倍率の閾値

	閾値とサンプル数分布	DCIS	IDC	IDC	
	healthy_U healthy_D DCIS_Up DCIS_Down IDC Up(2) IDC Down(2)	DCIS_Down	IDC Up(2)	IDC Down(2)	
Signal	100.0 100.0 100.0 100.0 100.0 100.0	1.585	-1.585	1.585	-1.585
log2(FC)	1.585 -1.585 1.585 -1.585 1.585 -1.585	90	462		
9					
8			254	667	
7			444	887	
6			656	1,107	
5			887	1,305	338 1,038
4			1,242	1,560 707 1,487	
3			1,784	1,875 1,261 1,977	
2		100	268	2,710 2,292 2,424 2,647	
1		1,034	1,365	5,272 3,338 5,404 3,977	
0	54,675	54,675	54,675	54,675 54,675 54,675 54,675	

	選抜用フィルタ条件	DCIS_Down	IDC Up(2)	IDC Down(2)
6	healthy_U healthy_D DCIS_Up DCIS_Down IDC Up(2) IDC Down(2)	>=6	>=6	
4	healthy_U healthy_D DCIS_Up DCIS_Down IDC Up(2) IDC Down(2)	>=4	>=4	

	選抜用フィルタ条件	DCIS_Down	IDC Up(2)	IDC Down(2)
4	healthy_U healthy_D DCIS_Up DCIS_Down IDC Up(2) IDC Down(2)	>=4	>=4	
2				

条件を満たした
サンプルの数

	条件を満たすサンプル数	DCIS_Down	IDC Up(2)	IDC Down(2)
1007_s_at	0	4	0	0
1053_at	0	0	0	2
117_at	0	0	0	0
121_at	0	0	0	0
1255_g_at	0	0	0	0
1294_at	0	0	0	0
1316_at	0	0	0	0
1320_at	0	0	0	0
1405_i_at	1	0	2	2
1431_at	0	0	0	0

必要なものは…

「統計学」

と

「生物学」

と

ほんの少しの

「情報処理スキル」

エクセルでもOK！

もし興味がもてたら、R、Python、Perlなどが出来ると便利ですが…

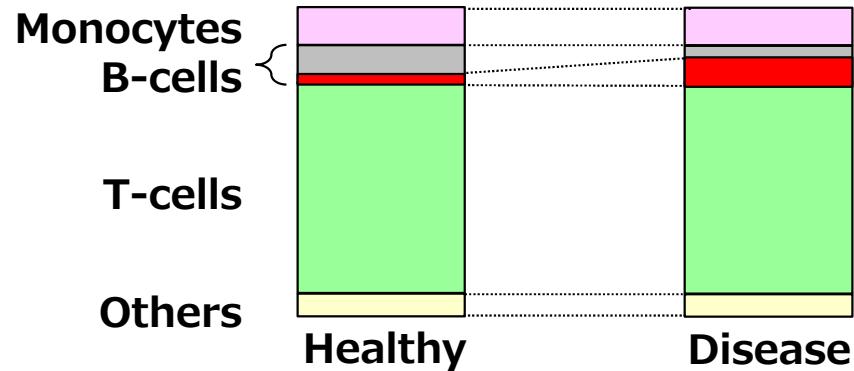
一番大切なものは…

「知りたい」という気持ち
好奇心

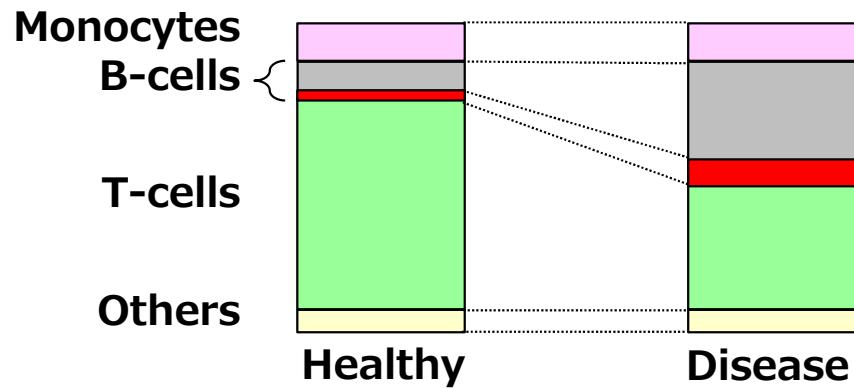
【参考】

組織・混合サンプルのデータを見る際の注意点

1) サブポピュレーション内の発現量が変化する場合

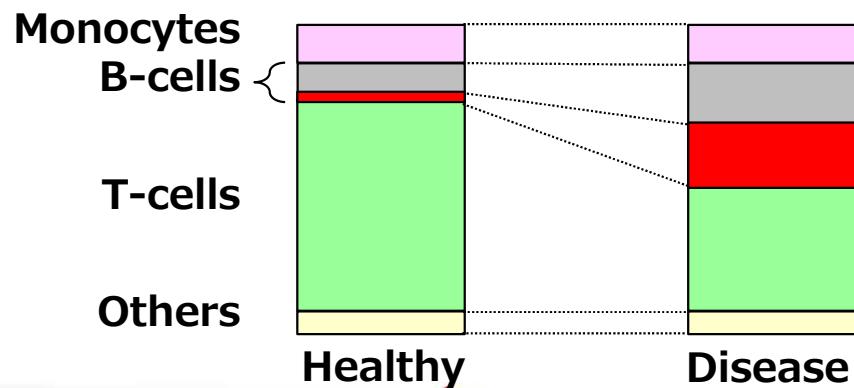


2) サンプル内のポピュレーションが変化する場合



3) 1)と2)の組み合わせ

遺伝子発現データだけから判断することはむずかしい…



【参考】

NCBI DataSet Browserの使い方

DataSet Browser (GDSビューワ)

データセット検索結果

The screenshot shows the NCBI GEO DataSet search results for "diabetes human kidney". It lists several datasets, each with a brief description, organism, platform, and sample count. A red box highlights the first dataset entry, which is about diabetic OVE26 glomerulus.

Data Set Browser画面

The screenshot shows the Data Set Browser for DataSet Record GDS3992. It displays the title, summary, organism, platform, citation, reference series, sample count, and value type. Below this, there are sections for "Data Analysis Tools" and "Find genes". A red box highlights the "Find genes" section, which includes options for comparing samples, creating cluster heatmaps, and viewing experiment designs and value distributions.

- Data Set Browserで出来る解析
- ① 特定の遺伝子のプロファイル表示
 - ② 2群間の変動遺伝子抽出
 - ③ クラスターヒートマップ表示
 - ④ 実験デザインと数値分布表示

Data Set BrowserではNCBIでGDS化したデータセットのみ解析可能。GEOにはGDS化されていないデータも多いが、NCBIのGDSプロジェクトが終了しているため、今後、これ以上に増えることはない。

1) 特定の遺伝子の発現を調べる

NCBI DATASET BROWSER CURATED BROWSE

Search for GDS3992[ACCN] Search Clear Show All Advanced Search

DataSet Record GDS3992: Expression Profiles Data Analysis Tools Sample Subsets

Title: Diabetic OVE26 glomerulus

Summary: Analysis of glomeruli isolated from kidneys of 8 week old, diabetic OVE26 males. The OVE26 type 1 mouse is a model of progressive glomerulosclerosis and decline of renal function. Results provide insight into the pathogenic mechanisms linked to diabetic nephropathy in the OVE26 model.

Organism: Mus musculus

Platform: GPL1261: [Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array

Citation: Reiniger N, Lau K, McCalla D, Eby B et al. Deletion of the receptor for advanced glycation end products reduces glomerulosclerosis and preserves renal function in the diabetic OVE26 mouse. *Diabetes* 2010 Aug;59(8):2043-54. PMID: 20627935

Cluster Analysis Download DataSet full SOFT file DataSet SOFT file Series family SOFT file Series family MINI ML file

Summary ▾

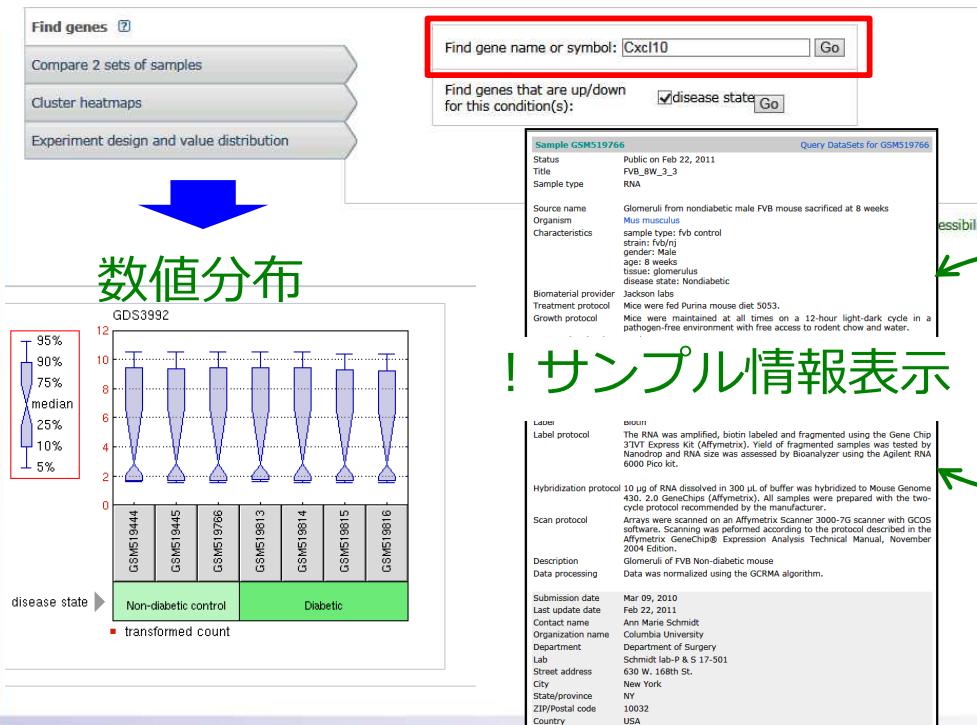
Cxcl10 - Diabetic OVE26 glomerulus

Annotation: Cxcl10, chemokine (C-X-C motif) ligand 10
Organism: Mus musculus
Reporter: GPL1261, 1418930_at (ID_REF), GDS3992, 15945 (Gene ID), NM_021274
DataSet type: Expression profiling by array, transformed count, 7 samples
ID: 75305661
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors

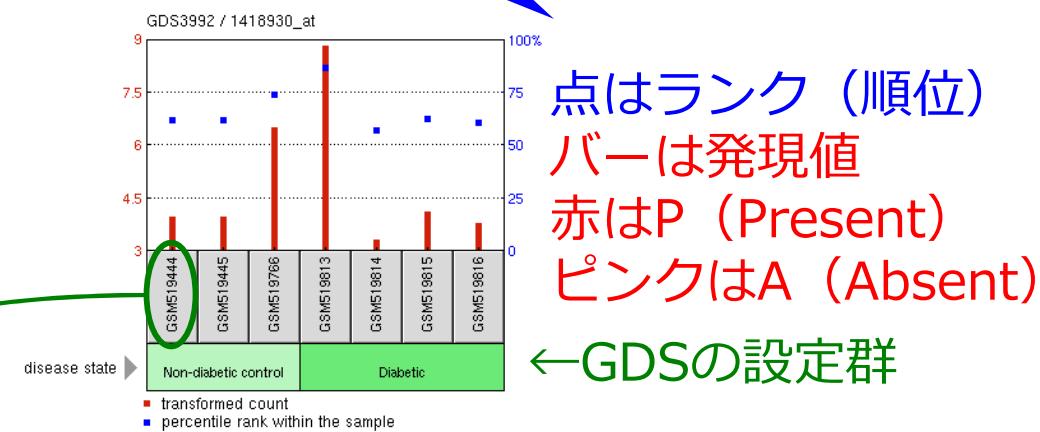
Send to: ▾

Profile GDS3992 / 1418930_at
Title Diabetic OVE26 glomerulus
Organism Mus musculus

① 遺伝子名・プローブ名等を入力し[Go]



② グラフをクリック



2) 2群間の発現を比較する

The screenshot shows the GEO Dataset Browser interface. On the left, the dataset record for GDS3992 is displayed, including details like Title (Diabetic OVE26 glomerulus), Summary (Analysis of glomeruli isolated from kidneys of 8 week old, diabetic OVE26 males.), Organism (Mus musculus), Platform (GPL1261: [Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array), and Reference Series (GSE20844). The Data Analysis Tools section is highlighted with a red box, showing Step 1 (Select test and significance level: Two-tailed t-test (A vs B), Significance level: 0.100), Step 2 (Select which Samples to put in Group A and Group B: Group A: GSM519444, GSM519445, GSM519766; Group B: GSM519813, GSM519814, GSM519815), and Step 3 (Query Group A vs. B).

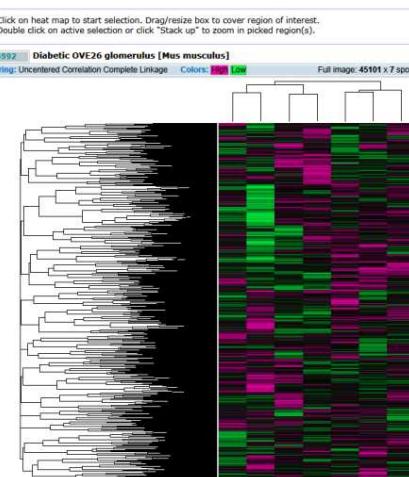
A central table lists samples grouped by disease state: Non-diabetic control (GSM519444, GSM519445, GSM519766) and Diabetic (GSM519813, GSM519814, GSM519815). A blue arrow points from this table to the 'Selected items' section below.

The 'Selected items' section shows a list of 5883 items, starting with Mrpl27 - Diabetic OVE26 glomerulus. To the right, there are two bar charts: one for Mrpl27 showing multiple red bars above a green bar, and another for GDS3992 / 1415690_at showing a single red bar above a green bar. A blue arrow points from the first chart to the second.

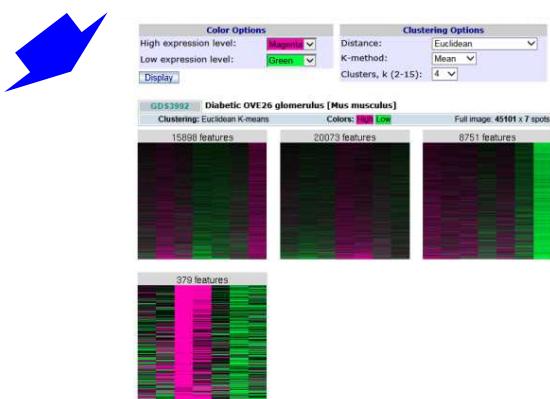
On the right side of the interface, there are sections for Profile pathways, Find related data, Recent activity, and Important Links.

- ① Step 1で変動抽出方法と閾値を選択
- ② Step 2 でサンプル群を設定
- ③ Step 3をクリック
- ④ 結果リストから表示する遺伝子のグラフをクリック

3) クラスタリング解析



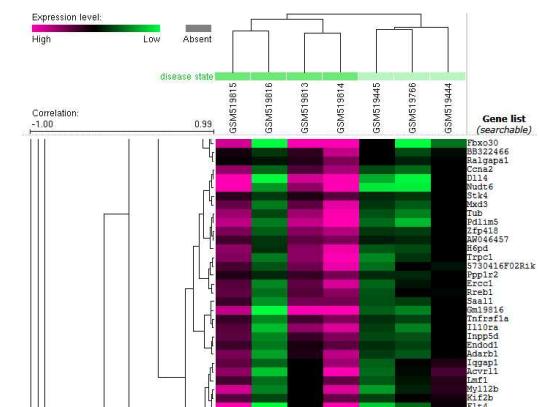
② Partitional (K-means/K-medians)で条件を設定し、[Display]をクリック



③ Location by chromosomeをクリック



領域選択して拡大、遺伝子名表示



領域選択して拡大、遺伝子名表示



【統合TV】DataSet Browserの使い方はこちらへ

1) NCBI GEO データセットブラウザをさらに使い倒す (6分26秒) [2009-03-7]

<https://togotv.dbcls.jp/20090307.html>

データセットブラウザからダウンロードできるファイルの説明や、発現に有意な変化のある遺伝子を探すツール、各サンプルの発現量の分布を見るツールなどの説明をしています。

2) NCBI GEO データセットブラウザの使い方1～2012 (7分59秒) [2012-1-28]

<https://togotv.dbcls.jp/20120128.html>

データセットブラウザを利用して、一つの実験データセットにおける様々な遺伝子発現を詳細に調べる方法を紹介しています。データセットの検索、データセット内の遺伝子の検索、及びヒートマップツールの使い方などについて説明しています。

3) NCBI GEO データセットブラウザの使い方2～2012 (6分26秒) [2009-03-7]

<https://togotv.dbcls.jp/20120227.html>

データセットブラウザからダウンロードできるファイルの説明や、発現に有意な変化のある遺伝子を探すツール、各サンプルの発現量の分布を見るツールなどの説明をしています。

4) NCBI GEOのデータセットブラウザを使って公共データの遺伝子発現解析を行ふ～2019 (6分26秒) [2019-4-3]

<https://togotv.dbcls.jp/20190403.html>

データセットの検索からデータセット内の遺伝子の検索、生データや処理済みデータなどNCBI GEOからダウンロード可能なファイルの説明、遺伝子発現パターンに有意な変化がある遺伝子を探す機能、クラスタリングとヒートマップ作成機能、各サンプルの発現量の分布を見るツールなどの説明をしています。

【参考】

NCBI GEO2Rの使い方



【統合TV】 <https://togotv.dbcls.jp/20210213.html>

(手順1) GEO2Rを起動する

GSE17913

喫煙による口腔内粘膜での
遺伝子発現変動を調べた
マイクロアレイデータ

- ① [Analyze with GEO2R] をクリック

The screenshot shows the NCBI GEO Accession Display page for study GSE17913. The page header includes the NCBI logo and the GEO logo. The main content area displays study details: Status (Public on Feb 15, 2010), Title (Effects of Cigarette Smoke on the Human Oral Mucosal Transcriptome), Organism (Homo sapiens), Experiment type (Expression profiling by array), and Summary (description of the study involving 40 smokers and 40 non-smokers). Below this, there is an 'Overall design' section. At the bottom left, a blue button labeled 'Analyze with GEO2R' is highlighted with a red box. A large red arrow points downwards from this button towards the 'Supplementary file' table at the bottom of the page.

Scope: self Format: HTML Amount: Quick GEO accession: GSE17913 GO

Series GSE17913 Query DataSets for GSE17913

Status Public on Feb 15, 2010

Title Effects of Cigarette Smoke on the Human Oral Mucosal Transcriptome

Organism Homo sapiens

Experiment type Expression profiling by array

Summary 40 current smokers and 40 age- and gender- matched never smokers underwent buccal biopsies.The study had four objectives: (a) to define the effects of smoking on the transcriptome of oral epithelial cells; (b) to determine if any of the effects of tobacco smoke on the transcriptome are gender-dependent; (c) to compare the effects of tobacco smoke exposure on the transcriptome in oral v. bronchial epithelium and (d) to identify agents with the potential to suppress the effects of tobacco smoke on the transcriptome.
We used microarrays to provide new insights into the carcinogenic effects of tobacco smoke and offer insights that may prove useful in developing preventive strategies.

Overall design 40 never smokers / 40 smokers

Analyze with GEO2R

Download family Format

SOFT formatted family file(s) SOFT

MINiML formatted family file(s) MINiML

Series Matrix File(s) TXT

Supplementary file	Size	Download	File type/resource
GSE17913_RAW.tar	373.1 Mb	(http)(custom)	TAR (of CEL)

(手順2) グループ設定とメンバーの選択

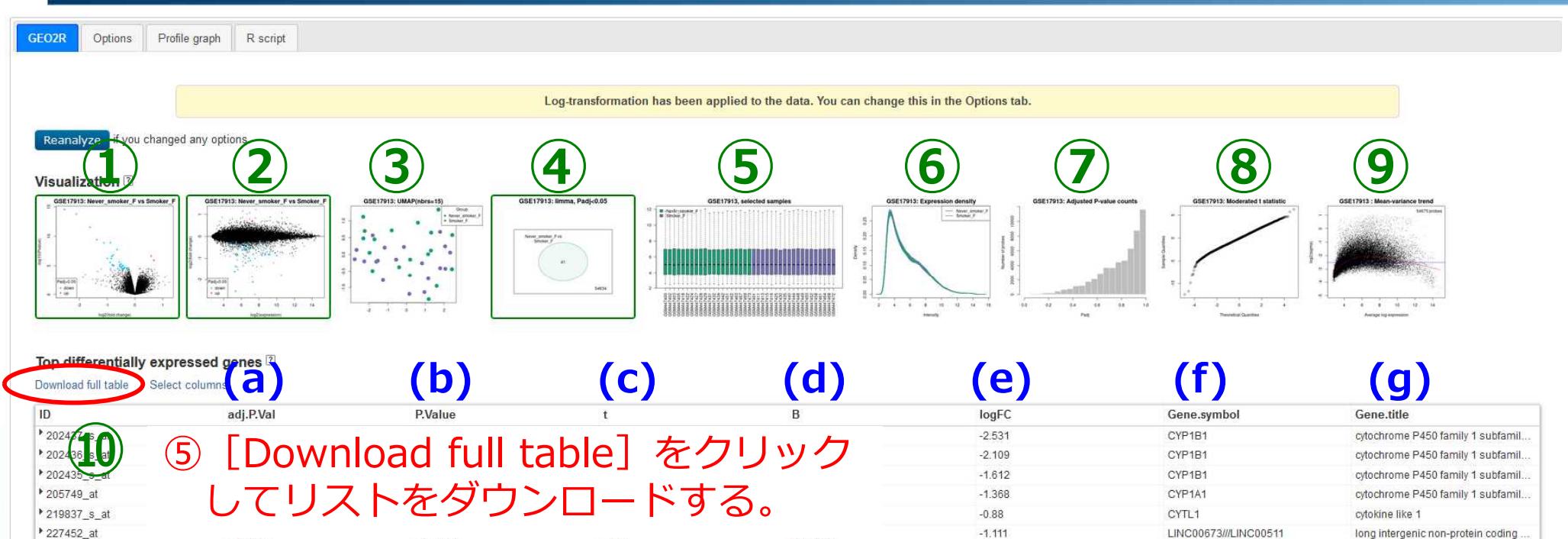
② グループの設定
[Define groups] をクリックし、
・ "Never_smoker_F"と入力して [リターン]
・ "Smoker_F"と入力して [リターン]

③ グループメンバーの選択
サンプルを選択して、
・ "Never_smoker_F"
・ "Smoker_F"
のメンバーを設定
(グループ名をクリック)。

④ [Analyze] をクリック

The screenshot shows the GEO2R interface for GSE17913. In the top left, there's a navigation bar with NCBI, GEO Publications, FAQ, MIAME, Email GEO, and Login. The main area has two panels. The left panel shows a table of samples with columns for Group and Accession. A red box highlights the 'Smoker_F' group. The right panel shows a 'Define groups' dialog where 'Never_smoker_F' and 'Smoker_F' have been entered. Red boxes highlight these entries. A blue arrow points from the 'Never_smoker_F' entry to the second panel. The second panel shows a list of samples with their details. Red boxes highlight the 'Never_smoker_F' and 'Smoker_F' groups in the list. A blue arrow points from the 'Smoker_F' entry in the first panel to the second panel. A large red downward-pointing arrow is at the bottom left. The bottom right contains a 'How to use' section with an 'Analyze' button and a note about differential expression analysis.

解析結果



- ① Volcano plot (FC vs. Pvalue)
- ② Mean-difference plot (Expression vs. FC)
- ③ UMAP plot
- ④ Venn diagram
- ⑤ Box plot (シグナル値分布)
- ⑥ Expression density plot (Intensity vs. Density)
- ⑦ P-value histogram
- ⑧ t-static quantile-quantile plot
- ⑨ Mean-variance trend plot
- ⑩ 発現有意差があった遺伝子のリスト (t-検定で P 値が小さい順に250件のリスト)

- ⑪ ⑩のリストのカラムヘッダー
 - (a) adj. P val (多重検定の補正後のP値)
 - (b) P value (元のP値)
 - (c) t (普通のtの標準偏差を全遺伝子の標準偏差で調整したmoderated-t)
 - (d) B (グループ間で有意に変動しているかのオッズ比の対数値)
 - (e) logFC (2を底とする発現変動対数値)
 - (f) Gene symbol
 - (g) Gene title

【統合TV】GEO2Rの使い方はこちらへ

1) NCBI GEOの使い方5～GEO2Rを使う～

(9分33秒) [2012-05-24]

<https://togotv.dbcls.jp/20120524.html>

NCBI GEO (Gene Expression Omnibus)はNCBIが提供・維持管理している遺伝子発現情報のデータベースです。GEOに登録されているマイクロアレイ実験のデータを、フリーのデータ解析環境 R をベースに解析できるツール GEO2Rの使い方を紹介します。GEOに登録されているデータから、それぞれのサンプルを発現量の差を調べたいグループに分け、検定の結果発現量に差が大きいとされた上位の遺伝子を表示するまでの流れを例にあげています。

2) GEO2Rを使ってマイクロアレイデータを解析する～ 2018

(7分24秒) [2018-04-05]

<https://togotv.dbcls.jp/20180405.html>

GEOに登録されているデータから、それぞれのサンプルを発現量の差を調べたいグループに分け、検定の結果発現量に差が大きいとされた上位の遺伝子を表示するまでの流れを例にあげています。

3) GEO2Rを使って公開されているマイクロアレイデータを解析する

(10分59秒) [2021-02-13]

<https://togotv.dbcls.jp/20210213.html>

GEO2Rを使って、2つ以上のグループの実験データを比較し、実験条件によって発現が異なる遺伝子を特定することができます。今回は、喫煙による口腔粘膜の遺伝子発現変化を調べた実験 (Effects of Cigarette Smoke on the Human Oral Mucosal Transcriptome: GSE17913を例に、それぞれのサンプルを発現量の差を調べたいグループに分け、検定の結果発現量に差が大きいとされた上位の遺伝子を表示したり、各統計量について解釈する方法について紹介します。

NCBIサイトで出来ること出来ないこと

<出来る事>

- ・ 単一データセットでの変動遺伝子抽出 (**GEO2R**)。
- ・ 単一データセット内でのプロファイル表示 (**GEO2R**)。
- ・ 単一のGDS化されたデータセットのクラスタリング解析 (**DataSet Browser**)。

= GEO2R =



<出来ない事>

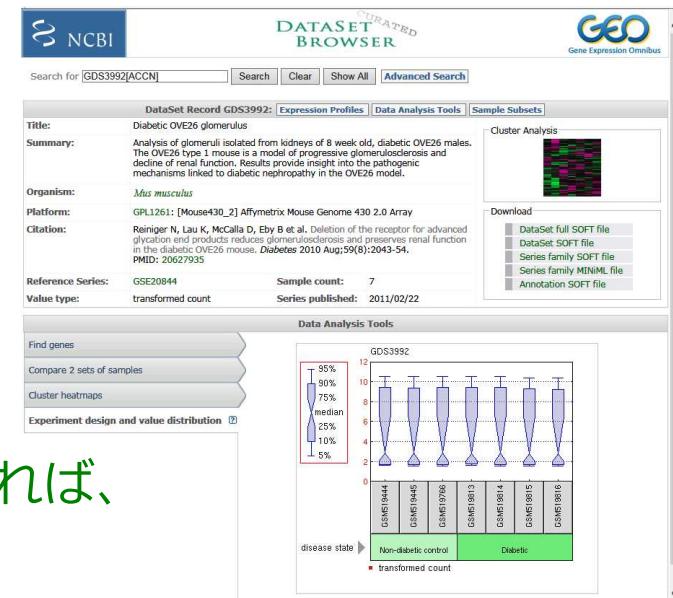
- ・ GDS化されていないデータのクラスタリング解析。
- ・ 複数のシリーズデータを組み合わせた解析。
- ・ 機能解析、パスウェイ解析、ネットワーク解析、上流制御因子予測、判別分析等、より複雑で多様な解析。

✓ NCBI GEOにはGDS化されていない実験セット(GSE)

がたくさん登録されている。

✓ 同じプラットフォームのデータでもRaw Dataが入手出来れば、異なる実験セットを組み合わせた解析も可能
(組合せ解析をする上で注意点は諸々あります)

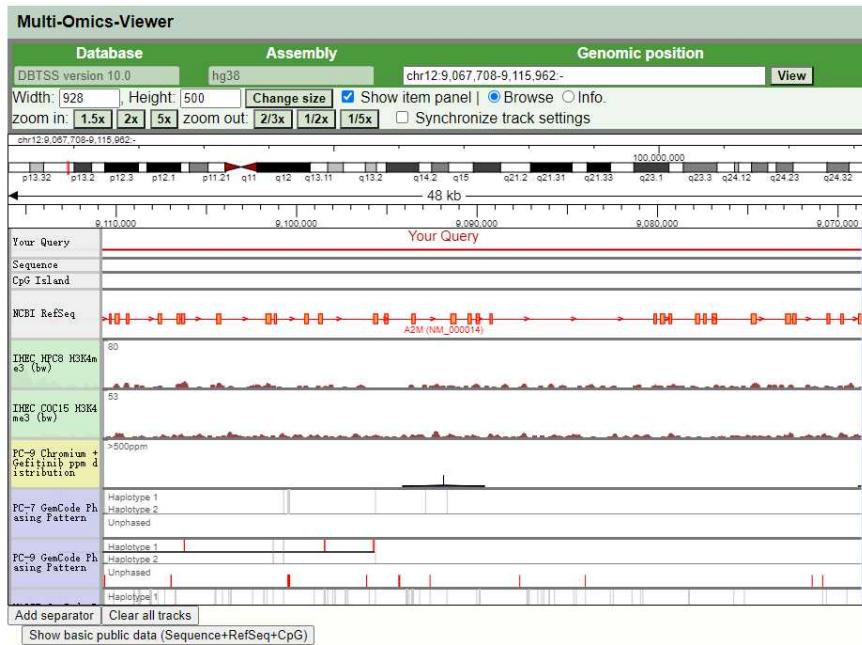
= DataSet Browser =



【参考】

そのほかの便利なサイト

【URL】 <https://kero.hgc.jp/Multi-Omics-Viewer>



TSS Seq Viewer



Multi-Omics-Viewerでは、遺伝子のエクソン-インtron構造、変異・多型情報、発現情報、エピゲノム情報（IHEC）、ChIP-seq情報（ChIP-Atlas）等を並列表現。
 TSS Seq Viewerでは、転写開始点情報を表示。

NCBI HomoloGene

【URL】 <https://www.ncbi.nlm.nih.gov/homologene/>

1: HomoloGene:3273. Gene conserved in Euteleostomi

Genes
Genes identified as putative homologs of one another during the construction of HomoloGene.

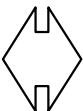
- ERBB2, *H.sapiens*
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
- ERBB2, *C.lupus*
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
- Erbb2, *M.musculus*
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
- Erbb2, *R.norvegicus*
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
- ERBB2, *G.gallus*
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
- erbb2, *D.rerio*
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog

Proteins
Proteins used in sequence comparisons and their conserved domain architectures.

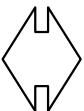
Species	Protein ID	Length	Domain Architecture
ヒト	NP_004439.2	1255 aa	[Schematic]
イヌ	NP_001003217.1	1259 aa	[Schematic]
マウス	NP_001003817.1	1256 aa	[Schematic]
ラット	NP_058699.2	1259 aa	[Schematic]
ニワトリ	NP_001038126.1	1235 aa	[Schematic]
ゼブラフィッシュ	NP_956413.2	1275 aa	[Schematic]

同じ名前の遺伝子がオルソログとは限らない！

HG: 117693
CXCL1 (Hs)
Cxcl3 (Mm)
Cxcl3 (Rn)



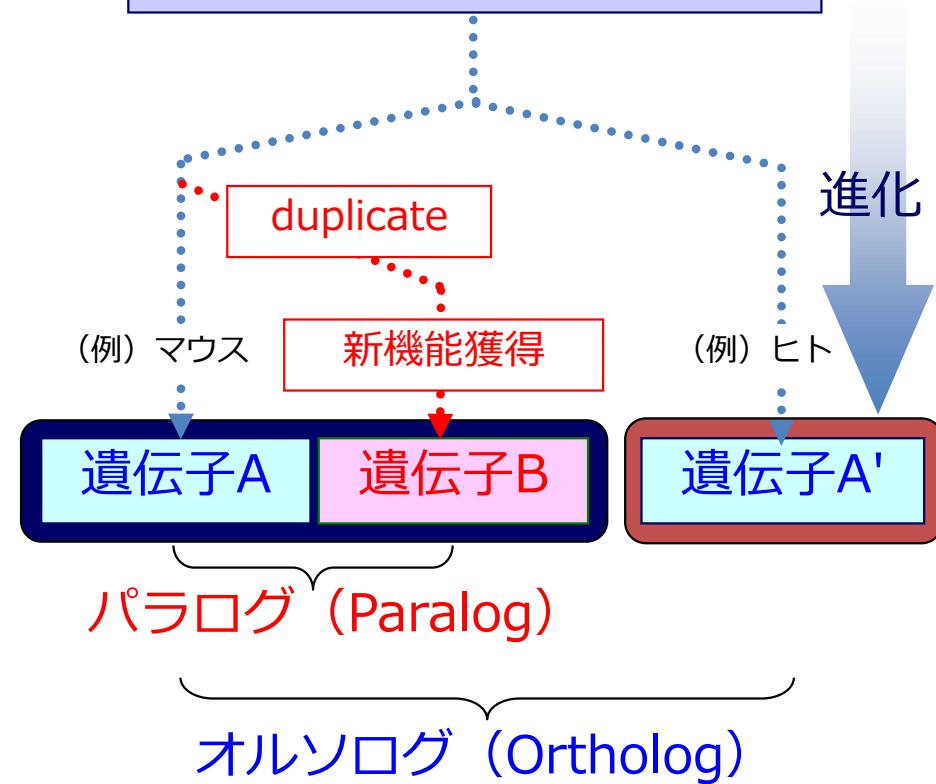
HG: 105490
CXCL2 (Hs)
Cxcl1 (Mm)
Cxcl1 (Rn)



HG: 117695
CXCL3 (Hs)
Cxcl2 (Mm)
Cxcl2 (Rn)

パラログとオルソログ

祖先遺伝子 (Ancestor gene)



Ortholog : 共通祖先に由来する異種間の同機能相同遺伝子
Paralog : ゲノム内で複製され新機能獲得した別機能相同遺伝子

【URL】 <http://www.biomart.org/>

各種ID変換（遺伝子、タンパク質、プローブ、等々）

異種生物のオルソログ対応付け、等

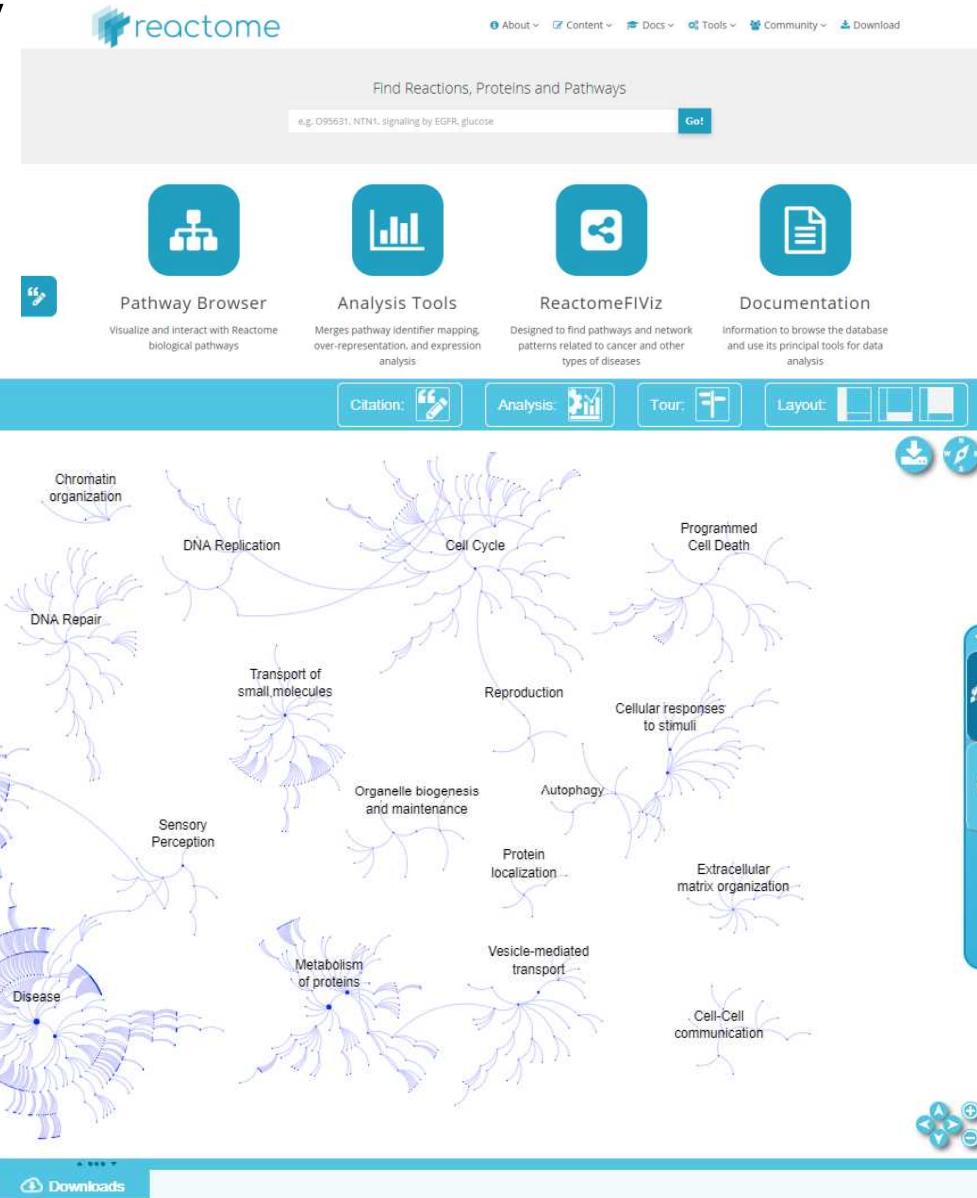
【統合TV】 biomartの使い方はこちらへ

- 「[BiomartでIDの対応表を作成する](#)」 (2分53秒) [2009-03-04]
- 「[BioMartを使い倒す～比較ゲノミクス編～](#)」 (6分48秒) [2010-3-29]
- 「[BioMart を用いてAffymetrixとAgilentのマイクロアレイのプローブID対応表を作成する 2011](#)」 (6分9秒) [2011-02-24]
- 「[Biomartを使い倒す～遺伝子の上流配列を取得する～2011](#)」 (3分41秒) [2011-05-27]
- 「[BioMartを使ってさまざまなIDの変換対応表を作成する](#)」 (6分26秒) [2011-09-27]
- 「[Biomart v0.8を使ってIDから遺伝子情報を取得する](#)」 (6分18秒) [2012-01-27]
- 「[BioMartを使って二つの生物種の対応するデータを取得する](#)」 (7分15秒) [2012-06-28]
- 「[BioMartを使い倒す～遺伝子の上流配列を取得する～2012](#)」 (4分34秒) [2012-07-20]
- 「[BioMart v0.9を使い倒す～遺伝子情報や配列を取得する～](#)」 (8分10秒) [2014-8-30]
- 「[BioMart v0.9を使い倒す～Enrichment analysis編～](#)」 (4分33秒) [2014-10-10]

:

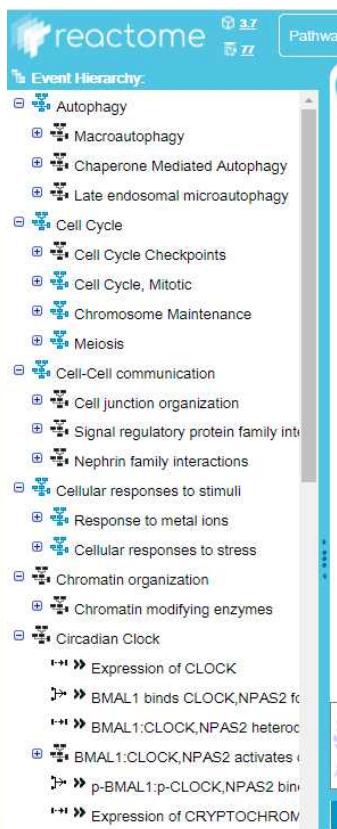
Reactome

【URL】 <https://reactome.org/>
 パスウェイデータベース



 Pathway Browser
 Analysis Tools
 ReactomeFIViz
 Documentation

Visualize and interact with Reactome biological pathways
Merges pathway identifier mapping, over-representation, and expression analysis
Designed to find pathways and network patterns related to cancer and other types of diseases
Information to browse the database and use its principal tools for data analysis



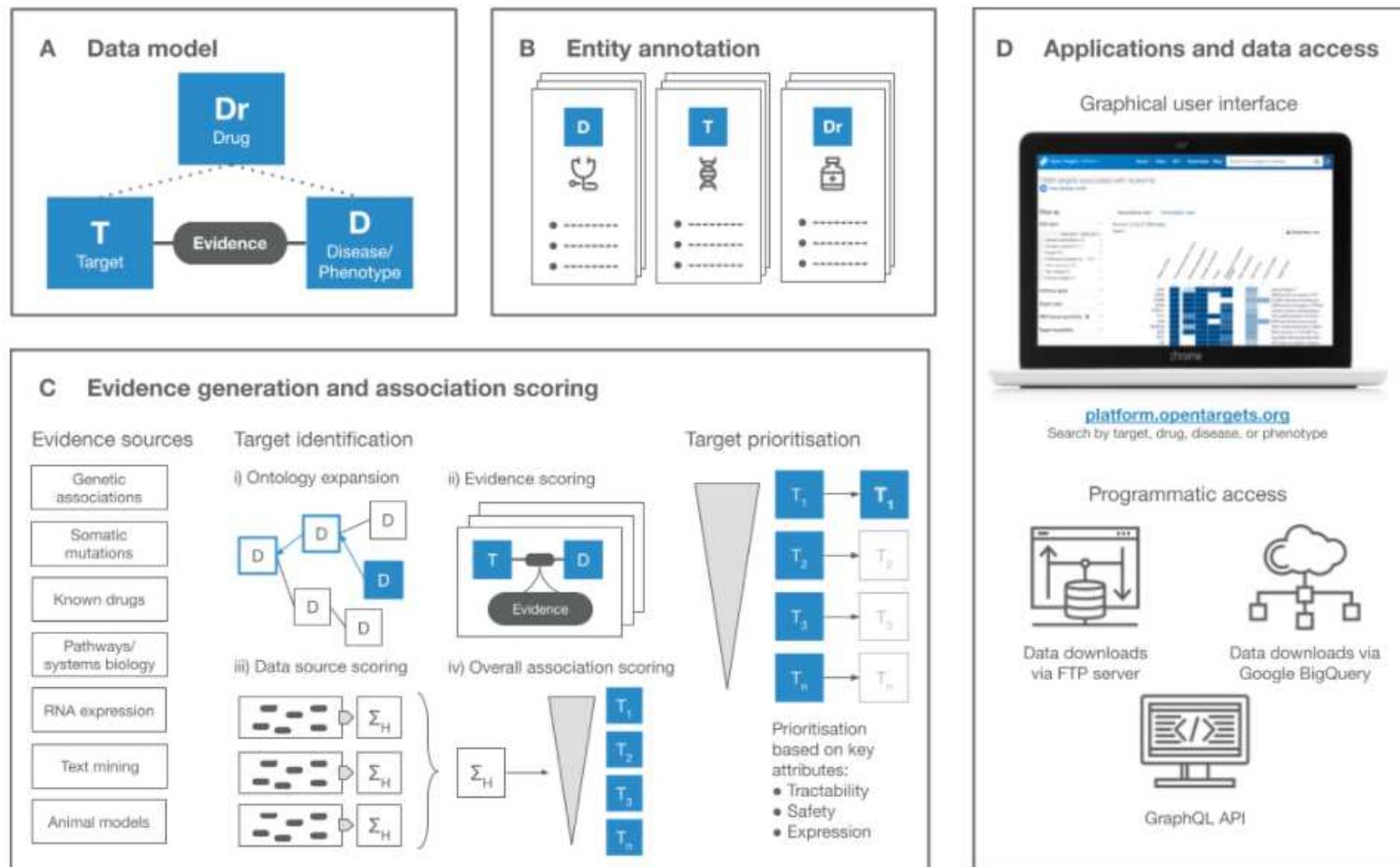
Event Hierarchy:

- Autophagy
 - Macroautophagy
 - Chaperone Mediated Autophagy
 - Late endosomal microautophagy
- Cell Cycle
 - Cell Cycle Checkpoints
 - Cell Cycle, Mitotic
 - Chromosome Maintenance
 - Meiosis
- Cell-Cell communication
 - Cell junction organization
 - Signal regulatory protein family interactions
 - Nephrin family interactions
- Cellular responses to stimuli
 - Response to metal ions
 - Cellular responses to stress
- Chromatin organization
 - Chromatin modifying enzymes
- Circadian Clock
 - Expression of CLOCK
 - BMAL1 binds CLOCK,NPAS2 to DNA
 - BMAL1:CLOCK,NPAS2 heterodimer binds DNA
 - BMAL1:CLOCK,NPAS2 activates transcription
 - p-BMAL1:p-CLOCK,NPAS2 binds DNA
 - Expression of CRYPTOCHROM

OpenTargets Platform

【URL】 <https://platform.opentargets.org>

疾患や表現型と薬剤標的分子との関係をデータに基づいて理解するための情報やツールが提供されている解析サイト。

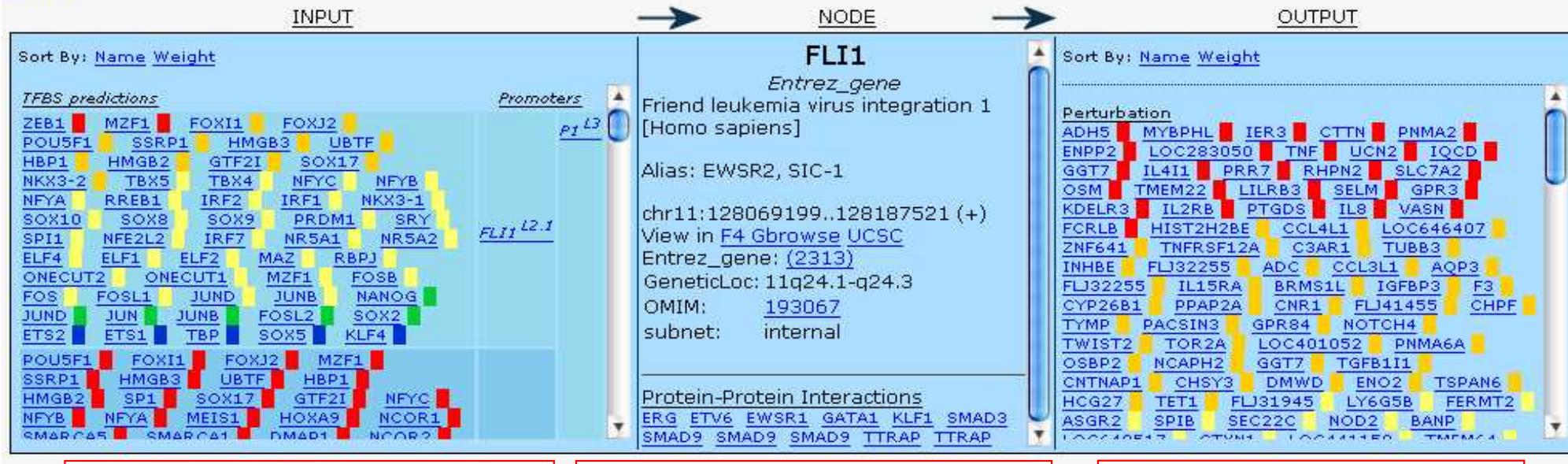




Search: FLI1

FLM FLII

<http://fantom.gsc.riken.jp/4/edgeexpress/>

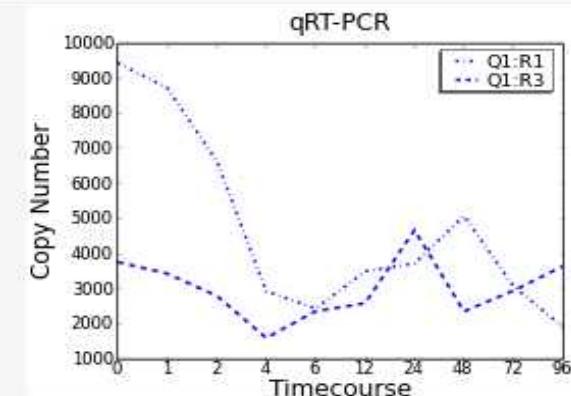
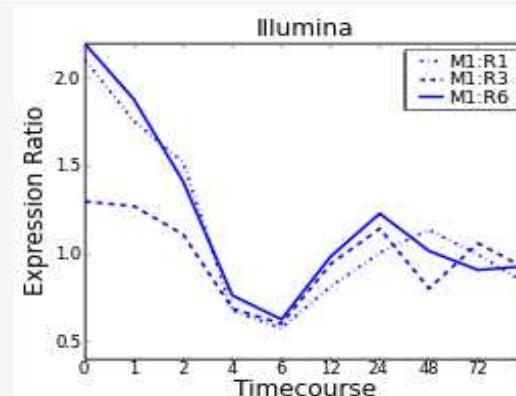
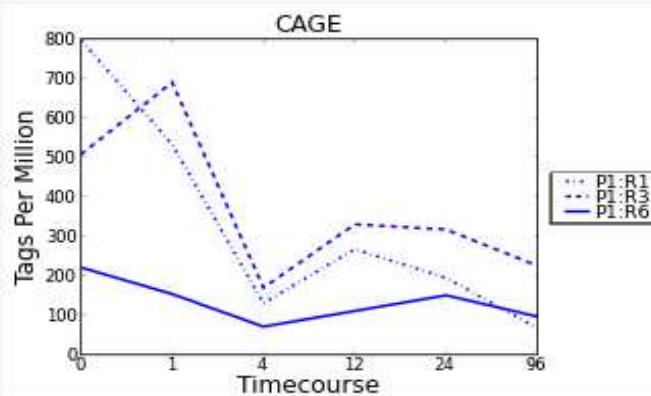


Save

① 制御因子予測(ChIP)

② 相互作用TF (M2H)

③ TF結合サイト予測



④ 転写変動 (CAGE, Illumina, qRT-PCR)

EdgeExpressDB SubnetView

FANTOM

HOME GENOME BROWSER EDGE EXPRESS DB DATA PUBLICATIONS ABOUT

Center view Subnet view About

Search: FLI1
(add all 2) FLI1 FLI1

Clear load genelist cookie
demos: net1 net2 net3 net4 net5

FLI1
ERG ERG ERG ERG ETV6 ETV6 ETV6
ETV6 EWSR1 EWSR1 EWSR1
GATA1 GATA1 GATA1 GATA1
GATA1 KLF1 KLF1 KLF1 KLF1
SMAD3 SMAD9 SMAD9 SMAD9
SMAD9 SMAD9 SMAD9 SMAD9
TTRAP TTRAP

primary edge types
 TFBS pred miRNA target pred
 Published protein-DNA PPI
 ChIP perturbation siRNA/miRNA

AND secondary edge types
 TFBS pred miRNA target pred
 Published protein-DNA PPI
 ChIP perturbation siRNA/miRNA

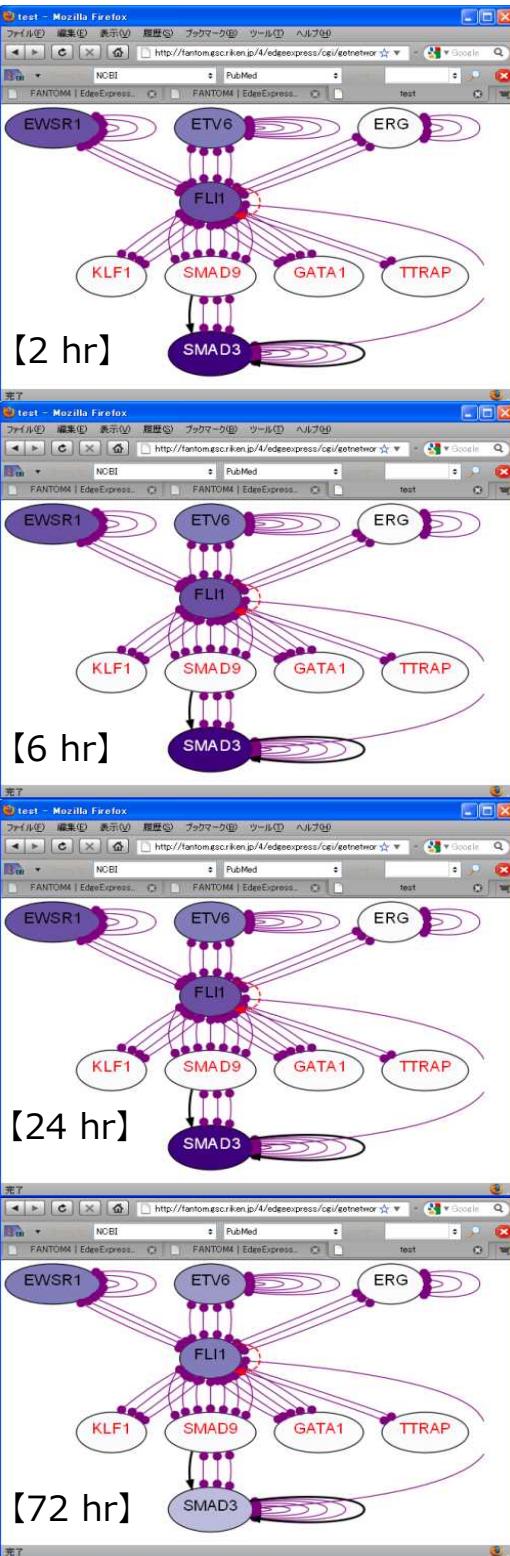
hide singletons hide leaves
expression timepoint: 0hr
SVG
generate subnet SVG preview
IE requires plugin for SVG
please download [Adobe SVG plugin](#)

【0 hr】 Monoblast

【96 hr】 Monocyte

相互作用と転写変動 (M2H, qRT-PCR)
[PMA刺激によるTHP-1細胞の分化過程]

EEDB subnet LEGEND info



【謝辞】

本講習会資料の「測定技術と検出限界について考える」の内容の多くは、株式会社Subio社長 田部暁郎氏の承諾を得て、SubioのHPで紹介されている資料をもとに作成させていただきました。心より感謝申し上げます。

- マイクロアレイの正規化手法とデータの解釈について
(<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-microarrays>)
- マイクロアレイのダイナミックレンジの比較
(<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-microarrays>)
- RNA-Seqのダイナミックレンジ
(<https://www.subioplatform.com/ja/products/subioplatform/the-dynamic-range-of-rna-seq>)



Akio Tanabe

【参考】

解析ツール

【URL】 <https://www.genepattern.org>

様々なゲノムデータを解析したり閲覧したりするための
解析ツールのレポジトリ。

Powerful genomics tools in a user-friendly interface



GenePattern provides hundreds of analytical tools for the analysis of gene expression (RNA-seq and microarray), sequence variation and copy number, proteomic, flow cytometry, and network analysis. These tools are all available through a Web interface with no programming experience required.

GenePattern Notebook



The GenePattern Notebook environment extends the Jupyter Notebook system, allowing researchers to create documents that interleave formatted text, graphics and other multimedia, executable code, and GenePattern analyses, creating a single "research narrative" that puts scientific discussion and analyses in the same place.

Analysis Pipelines



GenePattern pipelines allow you to capture, automate, and share the complex series of steps required to analyze genomic data. By providing a way to create and distribute an entire computational analysis methodology in a single executable script, pipelines enable a form of *in silico* reproducible research.

Reproducible Research



Published research, particularly *in silico* research, should contain sufficient information to completely reproduce the research results. By capturing the analysis methods, parameters, and data used to produce the research results, GenePattern pipelines enable reproducible research. By versioning every pipeline and its methods, GenePattern ensures that each version of a pipeline (and its results) remain static, even as your research and the pipeline continue to evolve.

Programming Environment



GenePattern provides a simple application interface that gives users access to computational analysis methods and tools, regardless of their computational experience. GenePattern also provides a programmatic interface that makes those analysis modules available to computational biologists and developers from Java, MATLAB, and R.



biostatistics

【URL】 <https://stats.biopapyrus.jp/>

生物統計学に関する情報サイト。

【目次】

1. プログラミング言語 (R、Python)
2. 確率分布 (確率、確率変数、正規分布、二項分布、ポアソン分布、t分布、カイ二乗分布、F分布)
3. 基礎統計 (標本分散と不偏分散、t検定、F検定、分散分析、主成分分析 Benjamini-Hochberg補正)
4. ベイズ統計学
5. 一般化線型モデル GLM (単回帰 重回帰 誤差構造・リンク関数・線形予測子 最尤推定 尤度比検定 スコア検定 Wald 検定 対数線形モデル ロジスティック回帰)
6. スパース推定 (正則化、LASSO、Ridge、Elastic Net、Fused LASSO Group LASSO、スパース主成分分析、グラフィカル LASSO)
7. 時系列解析
8. GxE 解析