

# 計算生命科学の基礎II

## 1.4 到来する大規模生命情報の解析に備えて

---

土井 淳

atsushi\_doi@cell-innovator.com

株式会社セルイノベーター  
研究開発部

〒812-8582 福岡市東区馬出3-1-1

九州大学 ウェストウィング8階 806 システム生命科学府 遺伝子制御学分野内

<http://www.cell-innovator.com>

# お伝えしたいこと

---

- 公開データの利用方法
  - BioGPS
  - Connectivity Map
  - The Cancer Genome Atlas (TCGA); cBioPortal
- データを表示する方法と、その見方
  - ヒートマップとクラスタリング
  - 機能解析 (GO解析、DAVID, GSEA)
  - パスウェイ解析
  - ネットワーク解析 (WCGNA)

# 公開されている大量の遺伝子発現データ

---

- BioGPS
  - <http://biogps.org/>
  - ある遺伝子は、どの組織で発現しているか？
- GEO
  - <http://www.ncbi.nlm.nih.gov/geo/>
  - マイクロアレイ、NGSのデータの公開データベース
- Connectivity Map
  - <https://www.broadinstitute.org/cmap/>
  - 薬剤を加えた時に変動する遺伝子はどれか？
- The Cancer Genome Atlas
  - <http://cancergenome.nih.gov>
  - がんに関するデータのデータベース (mRNA, SNP, CNV, etc.)

# BioGPS

BioGPS - your Gene Portal System

Sign Up or Login BioGPS account username OpenID account (what's that?) Google YAHOO! more

**BIOGPS** A free extensible and customizable gene annotation portal, a complete resource for learning about gene and protein function.

Plugins Datasets

**Simple to use**

- 1 Search for your gene of interest
- 2 View the gene annotation report
- 3 Browse the gene report layouts
- 4 Build your own gene report

**Search genes by Symbol or Accession**

ANXAS

Example Searches (click to try these samples)

- Gene Symbol(s)
- Wildcard queries
- Gene Ontology
- Affymetrix IDs
- Interpro

Press Ctrl-Enter or click Search by Symbol or Accession

Advanced Search (keyword, interval, etc.)

**User Love**

"I don't know how I didn't know Thank you for your fantastic website. It is more importantly FREE..." Jose University of Edinburgh

**News and Musings**

Data Chart Plugin Beta by Max Nanis  
How shall we find the concord of this discord? —William Shakespeare, A Midsummer Night's Dream & #160; Big news coming out [...] [view more]

New BioGPS and MyGene.info Paper published  
Introducing the Dataset Library  
ISMB 2012 recap  
Data chart plugin version 2.0

About BioGPS Blog Help FAQ Downloads API iPhone App Email updates Terms of Use © 2013 The Scripps Research Institute (ver 2.1.225-8acf773e8316)

ANXAS (annexin A5) | Gene Report | BioGPS

My Stuff Plugins Datasets Login here or Sign up Quick gene search

Current Gene List View Undo Save All (2) ANXAS anxas5

**ANXAS (annexin A5)** current layout: Default layout options Add a Plugin

Gene expression/activity chart Species: Hs

Dataset: GeneAtlas UI33A, gcma change  
Probeset: 200782\_at

Summary: The tissue-specific pattern of mRNA expression can indicate important clues about gene function. Hig... more

Interactive Image Static Image Correlation Downloads

Search: Zoom:

Gene Identifiers Species: Hs

Symbol: ANXAS  
Description: annexin A5  
Accessions: 308 (NCBI Gene), ENSG00000164111 (Ensembl), P08758 (UniProt), 131230 (OMIM), 20312 (HomoloGene)  
Aliases: ANX5, ENX2, PP4, RPRGL3  
Genome Location: chr4:122589110-122618268 (hg19)  
Molecular Function: phospholipase inhibitor activity (GO:0004859), calcium ion binding (GO:0005509), phospholipid binding (GO:0005543), calcium-dependent phospholipid binding (GO:0005544), receptor tyrosine kinase binding (GO:0030971), eukaryotic cell surface binding (GO:0043499)  
Biological Process: signal transduction (GO:0007165), blood coagulation (GO:0007596), response to organic substance (GO:0010033), positive regulation of apoptotic process (GO:0043065), negative regulation of apoptotic process (GO:0043066), negative regulation of coagulation (GO:0050819), protein homooligomerization (GO:0051260)  
Cellular Component: intracellular (GO:0005622), cytoplasm (GO:0005737), external side of plasma membrane (GO:0009897), intercalated disc (GO:0014704), sarcolemma (GO:0042383), cell projection (GO:0042995), extracellular vesicular exosome (GO:0070062)

Gene Wiki Species: Hs

About BioGPS Blog Help FAQ Downloads API iPhone App Email updates Terms of Use © 2013 The Scripps Research Institute (ver 2.1.225-8acf773e8316)

- 遺伝子名で検索すると、対象の遺伝子について、複数の組織や細胞における発現レベルが棒グラフで表示される。

# BioGPS の特徴

---

- 対象生物種は、ヒト、マウス、ラット、ブタ。
- 逆引きはできない。
  - 肝臓で発現している遺伝子を全てを取得したい。
  - 膵臓だけで得意的に発現している遺伝子はあるのか？
- GEO から、BioGPS で使用しているデータはダウンロードできる。
  - ヒト (GSE1133)、マウス (GSE10246) など。
  - ヒートマップを書けば、上記のような遺伝子も一目瞭然。

# 参考 : Gene Expression Omnibus (GEO)

The screenshot shows the main page of the Gene Expression Omnibus (GEO) website. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus. Below this, there are links for 'GEO Home', 'Documentation', 'Query & Browse', and 'Email GEO'. The main heading is 'Gene Expression Omnibus'. A paragraph describes GEO as a public functional genomics data repository supporting MIAME-compliant data submissions. Below the text, there are three columns of links: 'Getting Started' (Overview, FAQ, About GEO DataSets), 'Tools' (Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation), and 'Browse Content' (Repository Browser, DataSets, Series).

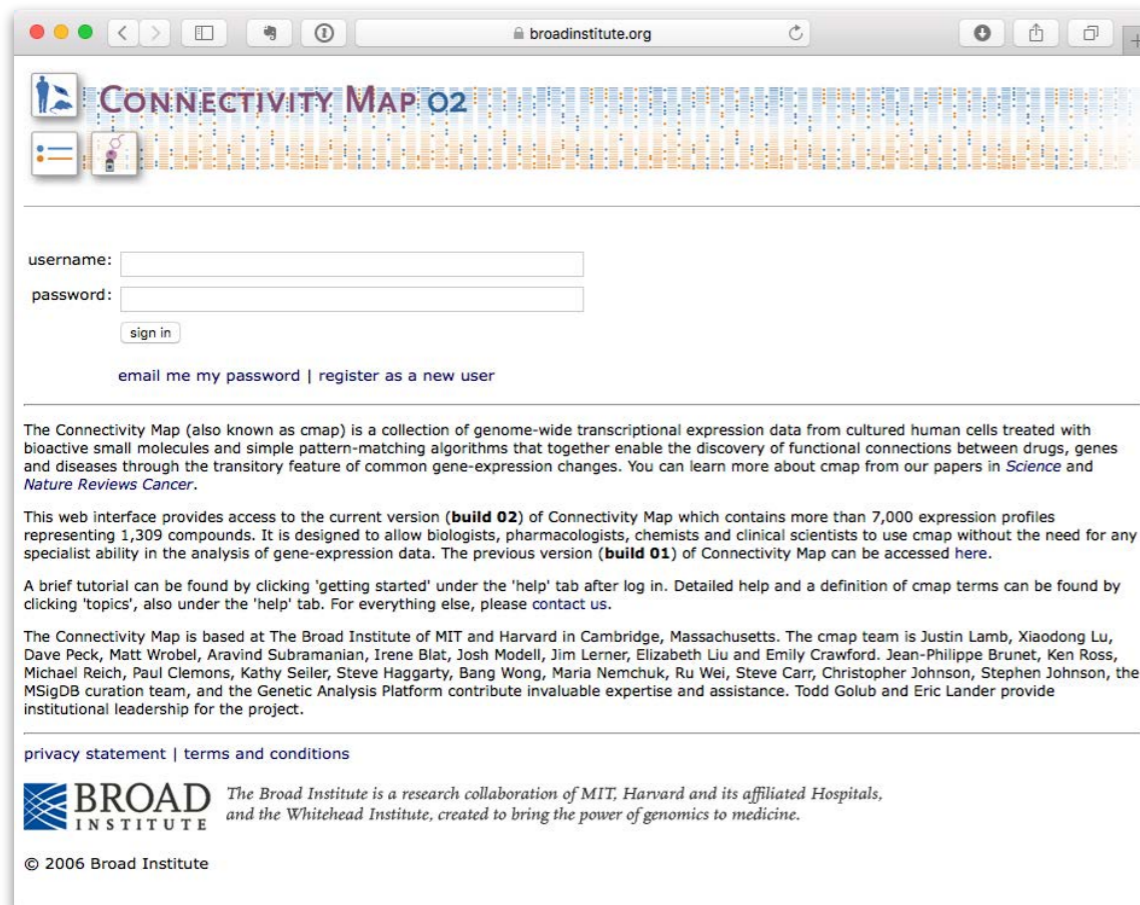
The screenshot shows a detailed view of a specific GEO accession, GSE53614. The page includes a search bar at the top with 'Scope: Self', 'Format: HTML', 'Amount: Quick', and 'GEO accession: GSE53614'. The main content area is titled 'Series GSE53614' and provides the following information:

- Status: Public on Dec 24, 2013
- Title: Gene expression signatures of 293T cell lines transfected with mutant TRPV3
- Organism: *Homo sapiens*
- Experiment type: Expression profiling by array
- Summary: We identified a number of affected pathways through transcriptome analysis on the skin biopsy samples of the FPPK patients. Our findings suggest that TRPV3 dysfunction may increase apoptotic activity, inhibit keratinocyte differentiation and disturb the intricate balance between proliferation and differentiation state of keratinocytes in the skin.
- Overall design: To understand the effect of TRPV3 mutation, transcriptome of 293T cell lines transfected with mutant TRPV3 were profiled in time-course manner (16, 24 and 40hr).
- Contributor(s): He Y, Wang K
- Citation missing: Has this study been published? Please login to update or notify GEO.
- Submission date: Dec 23, 2013
- Last update date: Dec 25, 2013
- Contact name: Ji-Hoon Cho
- E-mail: jcho@systemsbiology.org
- Organization name: Institute for Systems Biology
- Street address: 401 Terry Avenue North
- City: Seattle
- State/province: WA
- ZIP/Postal code: 98109
- Country: USA
- Platforms (1): GPL17077 Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381 (Probe Name version)
- Samples (6): GSM1296890 293T\_16hr\_Control, GSM1296891 293T\_24hr\_Control, GSM1296892 293T\_40hr\_Control, GSM1296893 293T\_16hr\_muTRPV3, GSM1296894 293T\_24hr\_muTRPV3, GSM1296895 293T\_40hr\_muTRPV3

At the bottom, there is a description of the series: 'Gene expression signatures in the skin of focal palmoplantar keratoderma (FPPK) patients, and cell lines transfected with mutant TRPV3'.

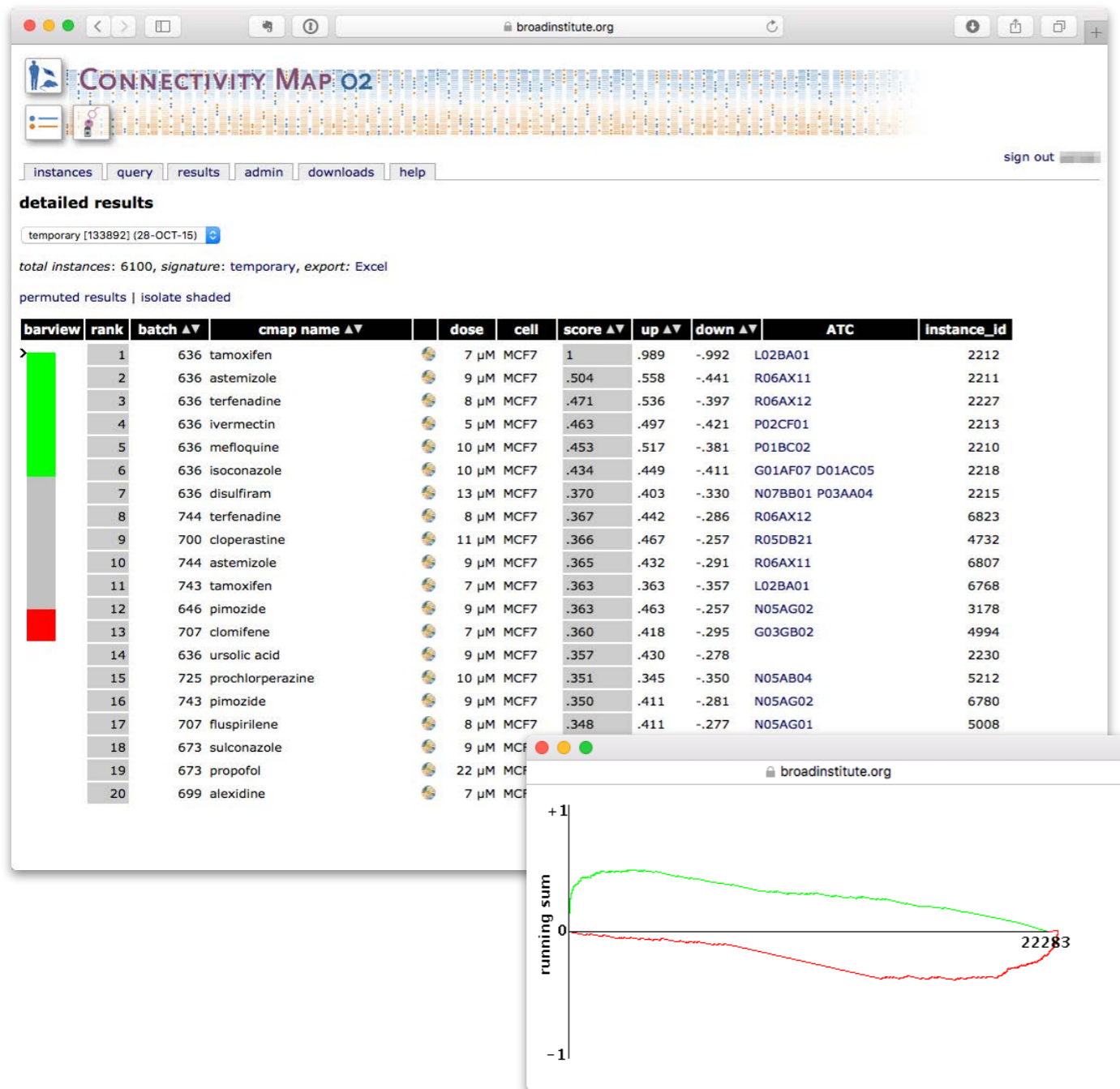
- <http://www.ncbi.nlm.nih.gov/geo/>
- NCBI のデータベースで、論文で使用されたマイクロアレイやシーケンスのデータが登録され、公開されている。
- 近年の論文では、投稿時に登録を求められることも多い。
- 同様のデータベースとして、ほかに ArrayExpress がある。

# Connectivity Map



- メールアドレスを登録して利用。
- up と down に分けて入力するのがポイント。
- 入力は、ProbeSetID (Affymetrix GeneChip Human Genome U133A Array) のみを受け付ける。(他のアレイの場合、BioMartなどで変換しておく)。

# どの薬剤と似ているか？



- results から details を確認すると、変動している遺伝子が似ている薬剤が表示される。

- ランク（全体のうち何番目に上がっていたか、下がっていたか）を用いて評価する。（fold-change そのものではない。）



# データの取得方法



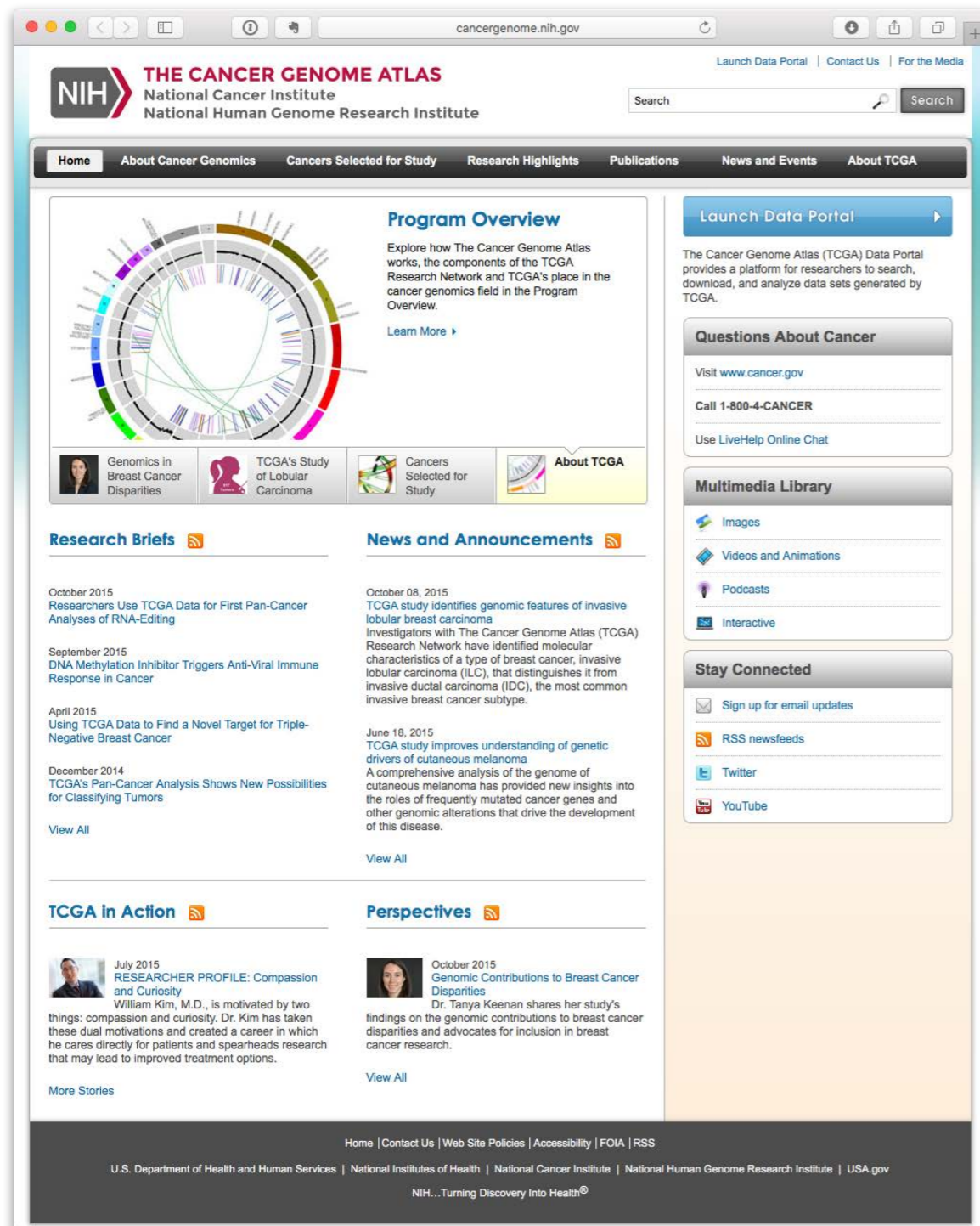
- 直接、CELファイル（何も処理していないデータ）をダウンロード可能。
- 処理済みの data matrix は、ランク形式のデータ。
- 6100インスタンス（=サンプル）。

# 備考：Connectivity Map のデータ

---

- 数は多い。
- 使用しているマイクロアレイは古い。
  - 搭載された遺伝子数が少ない。
  - よく使われたプラットフォームのため、比較対象が多いのは利点。
- 1つの薬剤につき、濃度の異なる複数のデータがある。
  - 薬剤は、1300種類。
  - 使用されている細胞は限定的（HL60, MCF7, PC3, SKMEL5, ssMCF7）。
- 後述の TCGA のデータが比較的新しく、様々な癌の種類があるので、組み合わせて利用すると良いかもしれない。

# The Cancer Genome Atlas (TCGA)



The screenshot shows the homepage of the Cancer Genome Atlas (TCGA) website. The header includes the NIH logo and the text "THE CANCER GENOME ATLAS National Cancer Institute National Human Genome Research Institute". A search bar is located in the top right. The main navigation menu includes "Home", "About Cancer Genomics", "Cancers Selected for Study", "Research Highlights", "Publications", "News and Events", and "About TCGA". The main content area is divided into several sections: "Program Overview" with a circular diagram, "Launch Data Portal" button, "Questions About Cancer" with contact information, "Multimedia Library" with links to Images, Videos and Animations, Podcasts, and Interactive content, and "Stay Connected" with links to email updates, RSS newsfeeds, Twitter, and YouTube. There are also sections for "Research Briefs" and "News and Announcements" with recent articles, and "TCGA in Action" and "Perspectives" with featured stories.

- クリニカル情報も含め、がんの研究データが公開されている。（\*データによっては、公開時期などの制限がある。）
- 制限なく公開されている mRNA, SNP, CNV などのデータは、後述の **cBioPortal** から閲覧すると簡単。
- がんの種類（Acute Myeloid Leukemia [LAML], Adrenocortical carcinoma [ACC], Bladder Urothelial Carcinoma [BLCA], Brain Lower Grade Glioma [LGG] など、34種類）

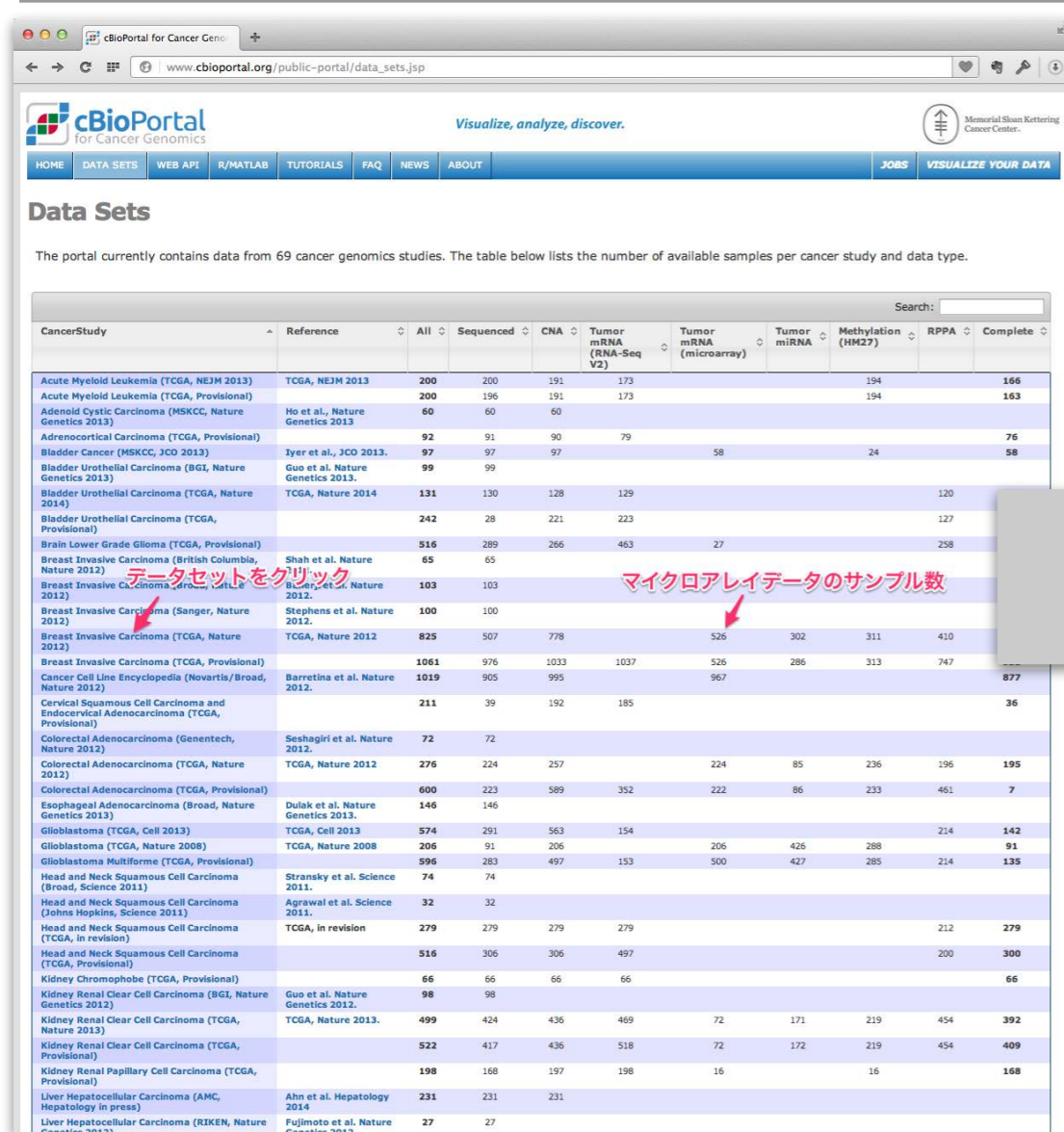
# cBioPortal 経由で TCGA のデータを閲覧

The screenshot shows the cBioPortal homepage. The header includes the logo and navigation links: HOME, DATA SETS, WEB API, R/MATLAB, TUTORIALS, FAQ, NEWS, ABOUT, JOBS, and VISUALIZE YOUR DATA. The main content area features a 'What's New' section with hiring announcements and a 'Data Sets' section stating that the portal contains data for 17,584 tumor samples from 69 cancer studies. A search interface is visible on the left, including a 'Select Cancer Study' dropdown, 'Select Data Type Priority' radio buttons, and an 'Enter Gene Set' field with a 'Submit' button.

The screenshot shows the cBioPortal study page for Breast Invasive Carcinoma (TCGA, Nature 2012). The page displays a 'Study Summary' tab with several data visualizations: 'Overall Survival' (Kaplan-Meier plot), 'Disease Free Survival' (Kaplan-Meier plot), 'Mutation Count vs CNA' (scatter plot), 'Overall Survival Status' (pie chart), 'Subtype' (pie chart), 'Disease Free Status' (pie chart), 'Overall Survival (Months)' (histogram), 'Disease Free (Months)' (histogram), 'Mutation Count' (histogram), and 'Copy Number Alterations' (histogram). The footer includes contact information for cBioPortal, MSKCC, and TCGA.

- <http://www.cbioportal.org>
- データをダウンロードせずに、その場で表示して確認もできる。

# cBioPortal: 変異のある遺伝子を表示する (1)



CancerStudy	Reference	All	Sequenced	CNA	Tumor mRNA (RNA-Seq V2)	Tumor mRNA (microarray)	Tumor miRNA	Methylation (HM27)	RPPA	Complete
Acute Myeloid Leukemia (TCGA, NEJM 2013)	TCGA, NEJM 2013	200	200	191	173			194		166
Acute Myeloid Leukemia (TCGA, Provisional)		200	196	191	173			194		163
Adenoid Cystic Carcinoma (MSKCC, Nature Genetics 2013)	Ho et al., Nature Genetics 2013	60	60	60						
Adrenocortical Carcinoma (TCGA, Provisional)		92	91	90	79					76
Bladder Cancer (MSKCC, JCO 2013)	Iyer et al., JCO 2013	97	97	97		58		24		58
Bladder Urothelial Carcinoma (BGI, Nature Genetics 2013)	Guo et al. Nature Genetics 2013	99	99							
Bladder Urothelial Carcinoma (TCGA, Nature 2014)	TCGA, Nature 2014	131	130	128	129					120
Bladder Urothelial Carcinoma (TCGA, Provisional)		242	28	221	223					127
Brain Lower Grade Glioma (TCGA, Provisional)		516	289	266	463	27				258
Breast Invasive Carcinoma (British Columbia, Nature 2012)	Shah et al. Nature 2012	65	65							
Breast Invasive Carcinoma (Sanger, Nature 2012)	Bauer et al. Nature 2012	103	103							
Breast Invasive Carcinoma (Sanger, Nature 2012)	Stephens et al. Nature 2012	100	100							
Breast Invasive Carcinoma (TCGA, Nature 2012)	TCGA, Nature 2012	825	507	778	526	302	311	410		
Breast Invasive Carcinoma (TCGA, Provisional)		1061	976	1033	1037	526	286	313	747	
Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012)	Barretina et al. Nature 2012	1019	905	995	967					877
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA, Provisional)		211	39	192	185					36
Colorectal Adenocarcinoma (Genentech, Nature 2012)	Seshagiri et al. Nature 2012	72	72							
Colorectal Adenocarcinoma (TCGA, Nature 2012)	TCGA, Nature 2012	276	224	257		224	85	236	196	195
Colorectal Adenocarcinoma (TCGA, Provisional)		600	223	589	352	222	86	233	461	7
Esophageal Adenocarcinoma (Broad, Nature Genetics 2013)	Dutak et al. Nature Genetics 2013	146	146							
Glioblastoma (TCGA, Cell 2013)	TCGA, Cell 2013	574	291	563	154				214	142
Glioblastoma (TCGA, Nature 2008)	TCGA, Nature 2008	206	91	206	206	426	288			91
Glioblastoma Multiforme (TCGA, Provisional)		596	283	497	153	500	427	285	214	135
Head and Neck Squamous Cell Carcinoma (Broad, Science 2011)	Stransky et al. Science 2011	74	74							
Head and Neck Squamous Cell Carcinoma (Johns Hopkins, Science 2011)	Agrawal et al. Science 2011	32	32							
Head and Neck Squamous Cell Carcinoma (TCGA, in revision)	TCGA, in revision	279	279	279	279				212	279
Head and Neck Squamous Cell Carcinoma (TCGA, Provisional)		516	306	306	497				200	300
Kidney Chromophobe (TCGA, Provisional)		66	66	66	66					66
Kidney Renal Clear Cell Carcinoma (BGI, Nature Genetics 2012)	Guo et al. Nature Genetics 2012	98	98							
Kidney Renal Clear Cell Carcinoma (TCGA, Nature 2013)	TCGA, Nature 2013	499	424	436	469	72	171	219	454	392
Kidney Renal Clear Cell Carcinoma (TCGA, Provisional)		522	417	436	518	72	172	219	454	409
Kidney Renal Papillary Cell Carcinoma (TCGA, Provisional)		198	168	197	198	16		16		168
Liver Hepatocellular Carcinoma (AMC, Hepatology in press)	Ahn et al. Hepatology 2014	231	231	231						
Liver Hepatocellular Carcinoma (RIKEN, Nature Genetics 2012)	Fujimoto et al. Nature Genetics 2012	27	27							



- データセットをクリックすると、サマリーが表示される。
- サマリーから、Mutated Genes のタブを選択する。

# cBioPortal: 変異のある遺伝子を表示する (2)

Breast Invasive Carcinoma (TCGA, Nature 2012)

1076 mutated genes

Gene	Cytoband	Gene size (Nucleotides)	# Mutations	# Mutations / Nucleotide	MutSig Q-value
TP53	7p13.1	3924	191	0.0487	8.64e-12
PIK3CA	3q26.32	9411	194	0.0206	8.64e-12
GATA3	10p15	3253	58	0.0178	5.25e-11
MAP3K1	5q11.2	6428	6	9.32e-4	1.13e-1
KMT2C	7q36.1	22650	38	1.678e-3	5.25e-11
CDH1	16q22.1	6148	33	5.368e-3	5.43e-11
MAP2K4	17p12	5714	21	3.675e-3	1.13e-1
TBX3	12q24.21	5950	13	2.185e-3	1.13e-1
PTEN	10q23.3	10048	19	1.891e-3	1.13e-1
PIK3R1	5q13.1	10791	13	1.205e-3	1.13e-1
RUNX1	21q22.3	15347	17	1.108e-3	1.13e-1
AKT1	14q32.32	10838	12	1.107e-3	1.14e-8
CTCF	16q21-q22.3	4301	13	3.023e-3	1.08e-7
NCOR1	17p11.2	15393	18	1.169e-3	1.010e-6
RPRG	Xp21.1	7143	10	1.400e-3	9.900e-5
RB1	13q14.2	6169	10	1.621e-3	2.800e-4
CBFB	16q22.1	4129	8	1.938e-3	3.700e-4
CDKN1B	12p13.1-p12	2811	5	1.779e-3	3.900e-4
ZFP36L1	14q22-q24	6676	7	1.049e-3	3.900e-4
FOXA1	14q12-q13	4111	8	1.946e-3	4.100e-4
PRRX1	1q24	8246	5	6.064e-4	3.400e-3
MUC4	3q29	21325	20	9.379e-4	3.800e-3
AFF2	Xq28	14006	13	9.282e-4	6.000e-3
DSPP	4q21.3	4373	8	1.829e-3	6.700e-3

Breast Invasive Carcinoma において変異のある遺伝子のリスト

遺伝子名をクリックすると詳細が見れます。

Gene Set / Pathway is altered in 22.7% of all cases.

Breast Invasive Carcinoma (TCGA, Nature 2012) / All Tumors: (825) / User-defined List/1gene

TP53

TP53: [Germline Mutation Rate: 0.5%, Somatic Mutation Rate: 22.2%]

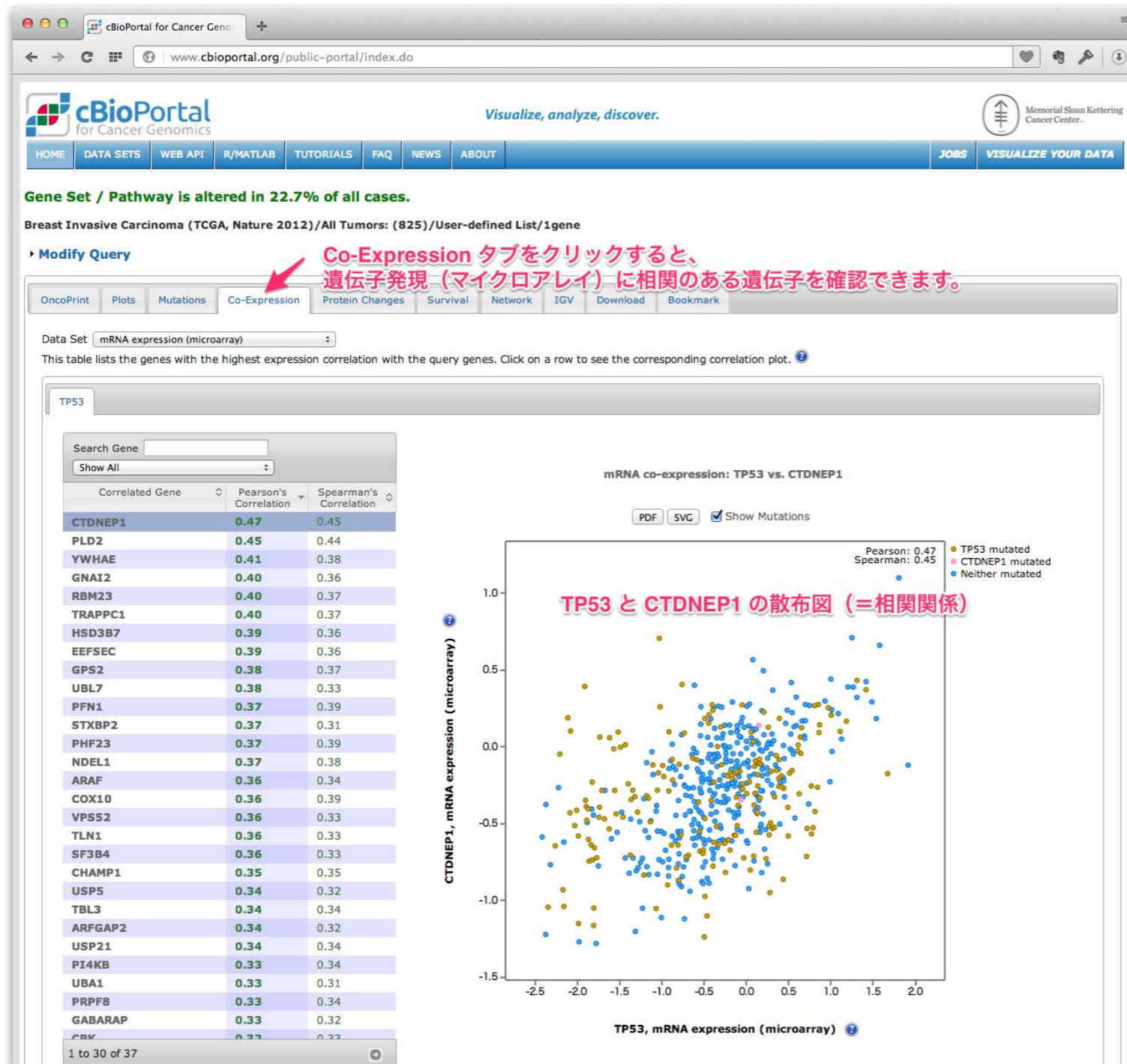
変異の位置が示されています。マウスでポイントすることで、どの検体に見られる変異が確認できます。

Case ID	AA change	Type	Copy #	COSMIC	MS	Mutation Assessor	# Mut in Sample
TCGA-A7-A0CJ	R175H	3D Missense	hetloss	878	S	Medium	39
TCGA-C8-A12L	C238F	3D Missense	hetloss	134	S	Medium	67
TCGA-B6-A0IQ	R213*	3D Nonsense	diploid	243	S	Medium	51
TCGA-BH-A0C0	R110fs	3D FS del	diploid	11	S	Medium	65
TCGA-AN-A0AJ	R110fs	3D FS del	hetloss	7	S	Medium	62
TCGA-AN-A0FJ	V157fs	3D FS ins	hetloss	14	S	Medium	67
TCGA-BH-A1EV	S215R	3D Missense	hetloss	62	S	Medium	75
TCGA-BH-A18H	Q331*	3D Nonsense	diploid	31	S	Medium	25
TCGA-DB-A147	C176F	3D Missense	hetloss	234	S	Medium	108
TCGA-B6-A0WX	R248W	3D Missense	gain	1190	S	Medium	28
TCGA-AR-A0TS	E198*	3D Nonsense	diploid	28	S	Medium	34
TCGA-BH-A18V	E56*	3D Nonsense	hetloss	8	S	Medium	149
TCGA-AN-A0G0	R342*	3D Nonsense	NA	83	S	Medium	31
TCGA-C8-A12K	I195T	3D Missense	diploid	119	S	Medium	81
TCGA-AO-A03N	A161T	3D Missense	diploid	72	S	Medium	87
TCGA-BH-A18U	E285fs	3D FS del	hetloss	25	S	Medium	63
TCGA-AB-A07W	H179R	3D Missense	hetloss	289	S	Medium	69
TCGA-BH-A0AW	F270S	3D Missense	hetloss	69	S	Medium	109
TCGA-AN-A0FY	V173M	3D Missense	diploid	131	S	Medium	47
TCGA-E2-A1B6	R342*	3D Nonsense	diploid	83	S	Medium	12
TCGA-BH-A0RX	R342*	3D Nonsense	diploid	83	G	Medium	18
TCGA-E2-A150	C176*	3D Nonsense	hetloss	22	S	Medium	36
TCGA-AR-A0TV	V216M	3D Missense	diploid	80	S	Medium	59
TCGA-BH-A0DZ	E258Q	3D Missense	hetloss	85	S	Medium	180
TCGA-BH-A0BC	R248W	3D Missense	hetloss	1190	S	Medium	23
TCGA-A2-A0D1	Q331*	3D Nonsense	hetloss	31	S	Medium	35
TCGA-AO-A03R	L265R	3D Missense	hetloss	23	S	Medium	27
TCGA-AR-A0U2	C176F	3D Missense	diploid	234	S	Medium	42

変異のある検体のリスト

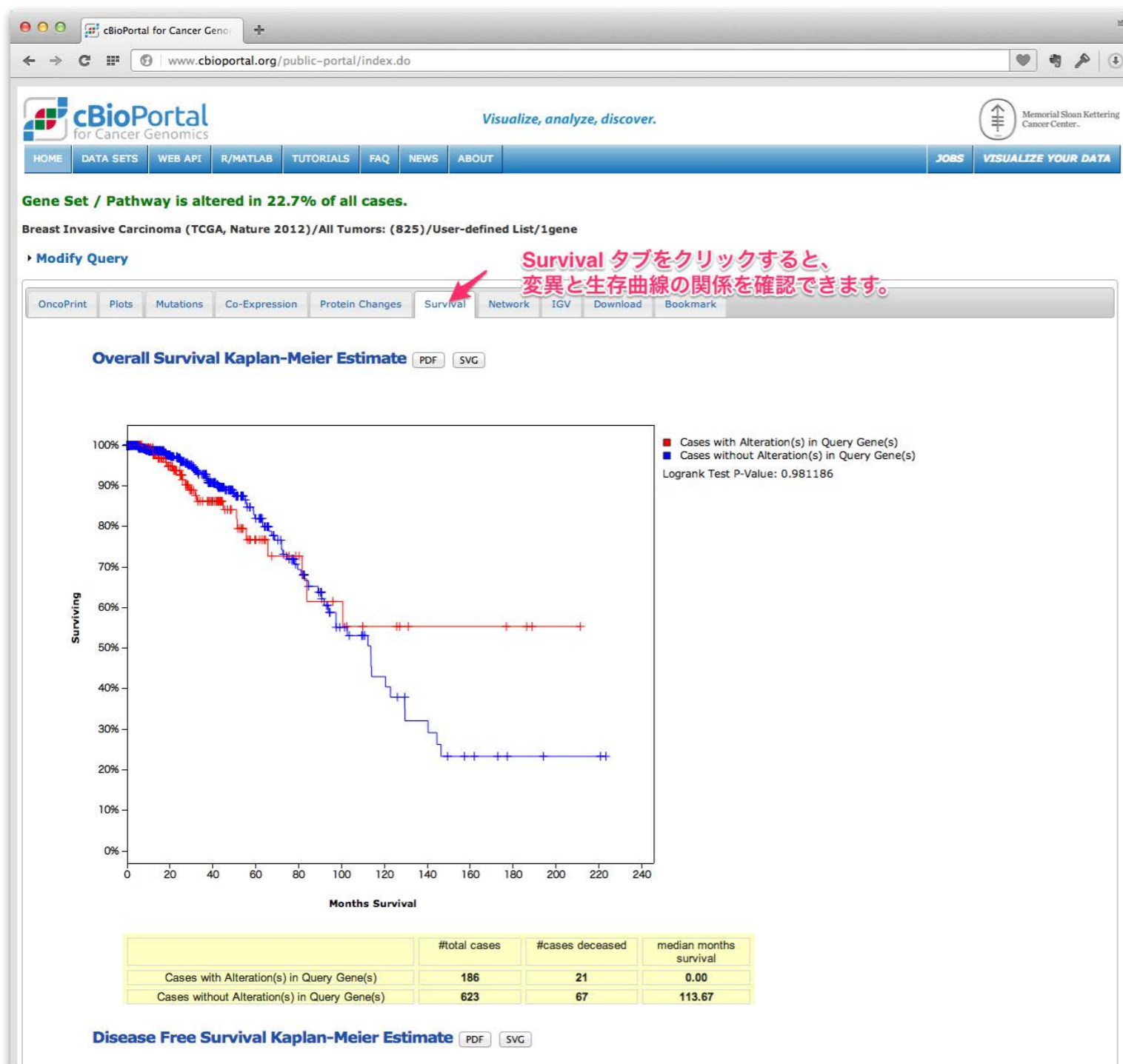
- 変異のある遺伝子の一覧が表示される。
- 遺伝子名をクリックすると、変異のポイントなどの詳細が表示される。

# cBioPortal: Co-Expression の表示



- Co-Expression のタブをクリックすると、共に発現している遺伝子（共発現遺伝子）の関係を表示できる。
- 相関関係があれば、散布図の表示が、左下から右上（または、左上から右下に）点が集まって見える。

# cBioPortal: 生存曲線の表示



- **Survival** のタブをクリックすると、変異と生存曲線の関係を確認できる。
- 遺伝子群を指定して、データセットを表示していれば、指定された遺伝子群について生存曲線を確認できる。



# cBioPortal: ネットワークの表示

Gene Set / Pathway is altered in 22.7% of all cases.  
Breast Invasive Carcinoma (TCGA, Nature 2012)/All Tumors: (825)/User-defined List/1gene

Modify Query

OncoPrint Plots Mutations Co-Expression Protein Changes Survival **Network** IGV Download Bookmark

The network below contains 51 nodes, including your 1 query gene and the 50 most frequently altered neighbor genes (out of a total of 390).  
Download the complete network in GraphML or SIF for import into Cytoscape (GraphMLReader plugin is required for importing GraphML).

File Topology View Layout Legends Double-click nodes/edges for details

Genes Interactions Details Help

Type:

- In Same Component 19.5%
- Reacts with 20.3%
- State Change 8.9%
- Targeted by Drug 0.4%
- Other 50.8%

Source:

- NCI\_NATURE
- HPRD
- REACTOME
- CELLMAP
- DrugBank
- Unknown

Update

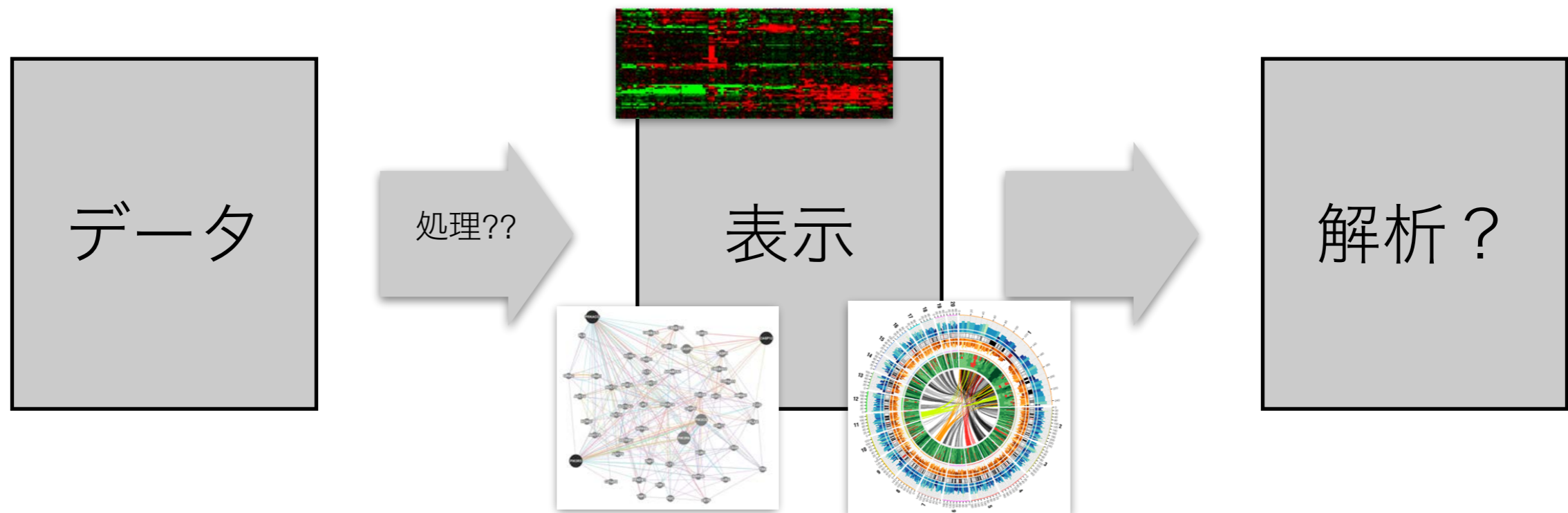
cBioPortal | MSKCC | TCGA  
Questions and feedback: [cbioportal@googlegroups.com](mailto:cbioportal@googlegroups.com) | [User discussion group](#)

- **Network** のタブをクリックすると、選択した遺伝子に関連した遺伝子がネットワークで表示される。
- 関係性の情報は、NCI\_Nature (Pathway Interaction Database), HPRD, REACTOME, DrugBank などのデータベースの情報が使用される。

データを表示する方法と、その見方

# データを表示して解析というパラダイム

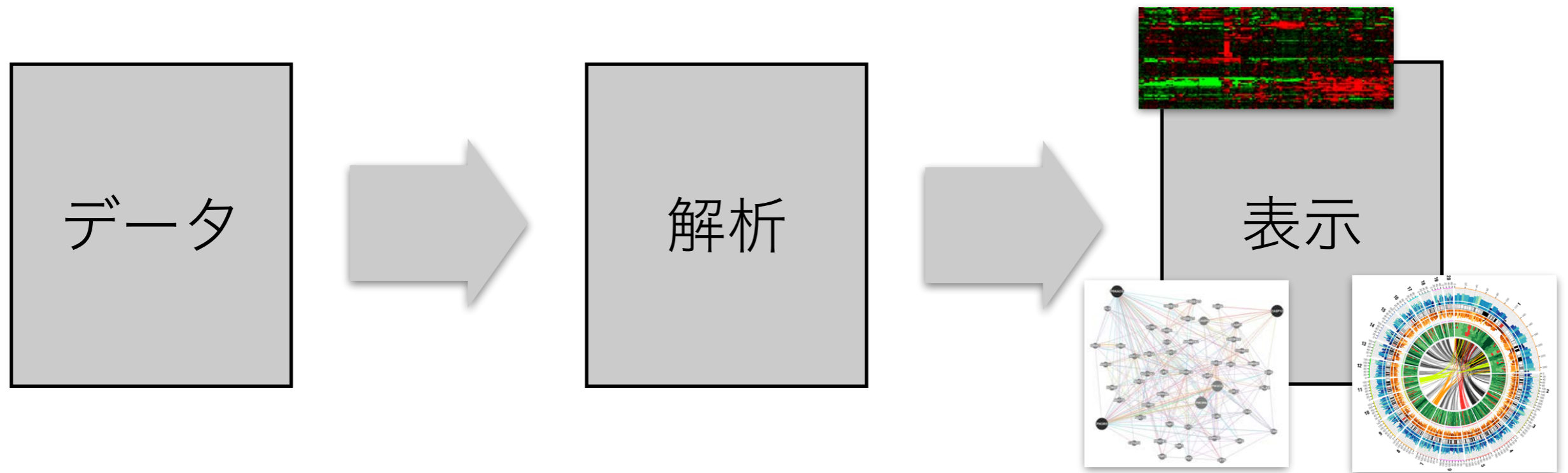
---



- 念のため、のお話。
- データをプログラムで処理して、表示してから、解析、と思っていないですか？

# データを解析して表示というパラダイム

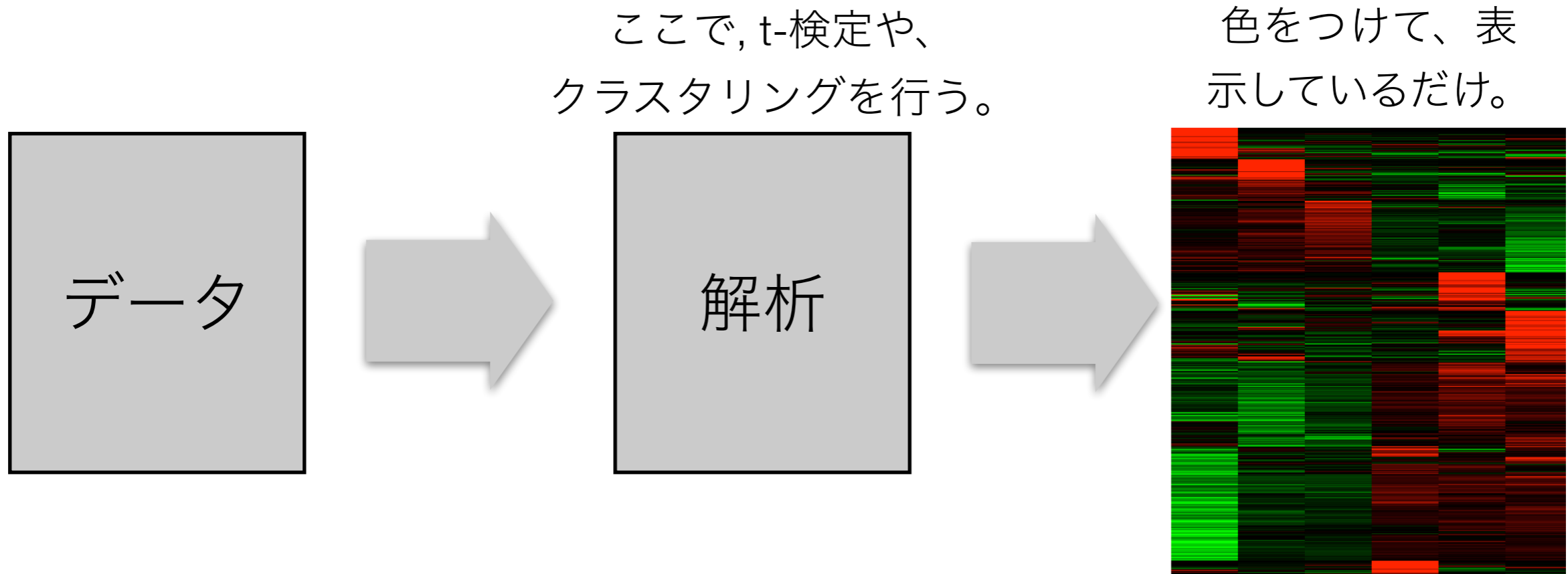
---



- 多くの場合、ヒートマップ、ネットワーク、環状ダイアグラム (circos, chord diagram) などは、**解析した後**、結果を表示している。
- (もちろん、その先に、さらに解析が続く場合もあるが。)

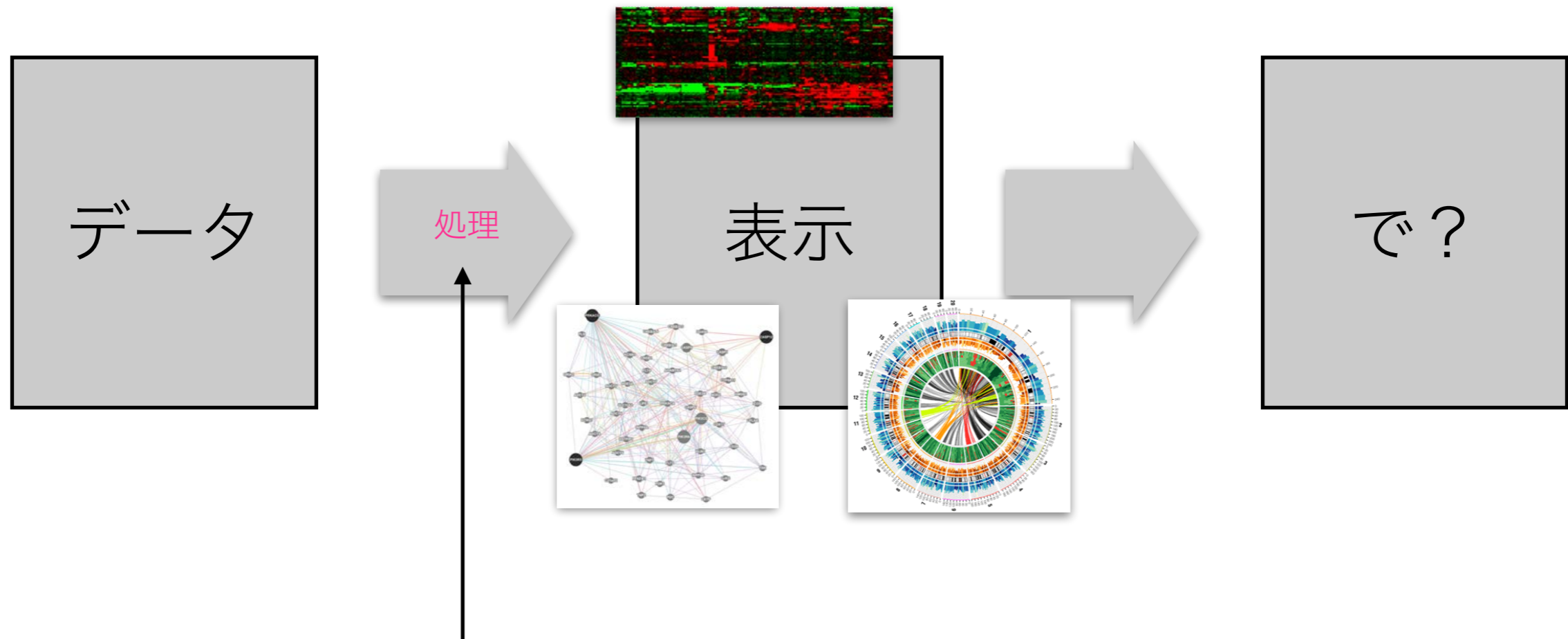
# 例えば、ヒートマップ

---



- ハッキリ分かれたヒートマップでは、検定した結果、**有意差のある遺伝子だけを取り出して**書いている場合に注意。（良い悪いという話ではなく。。。）
- ヒートマップを書いてから、分かれている部分（クラスター）を探しているわけではない。

# データを表示して解析というパラダイム

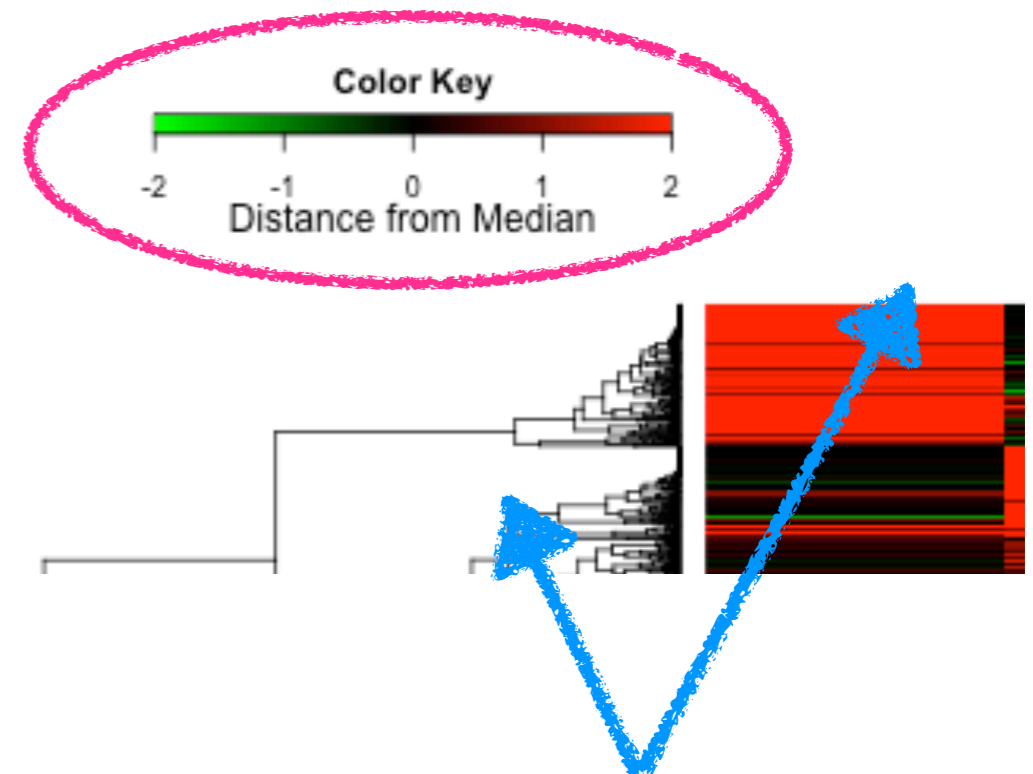


- この流れでも良いけれど。。
- 「処理の部分で何をしているか」「何を表示しているか」を理解しないと、表示を解釈する時に誤解を招く。

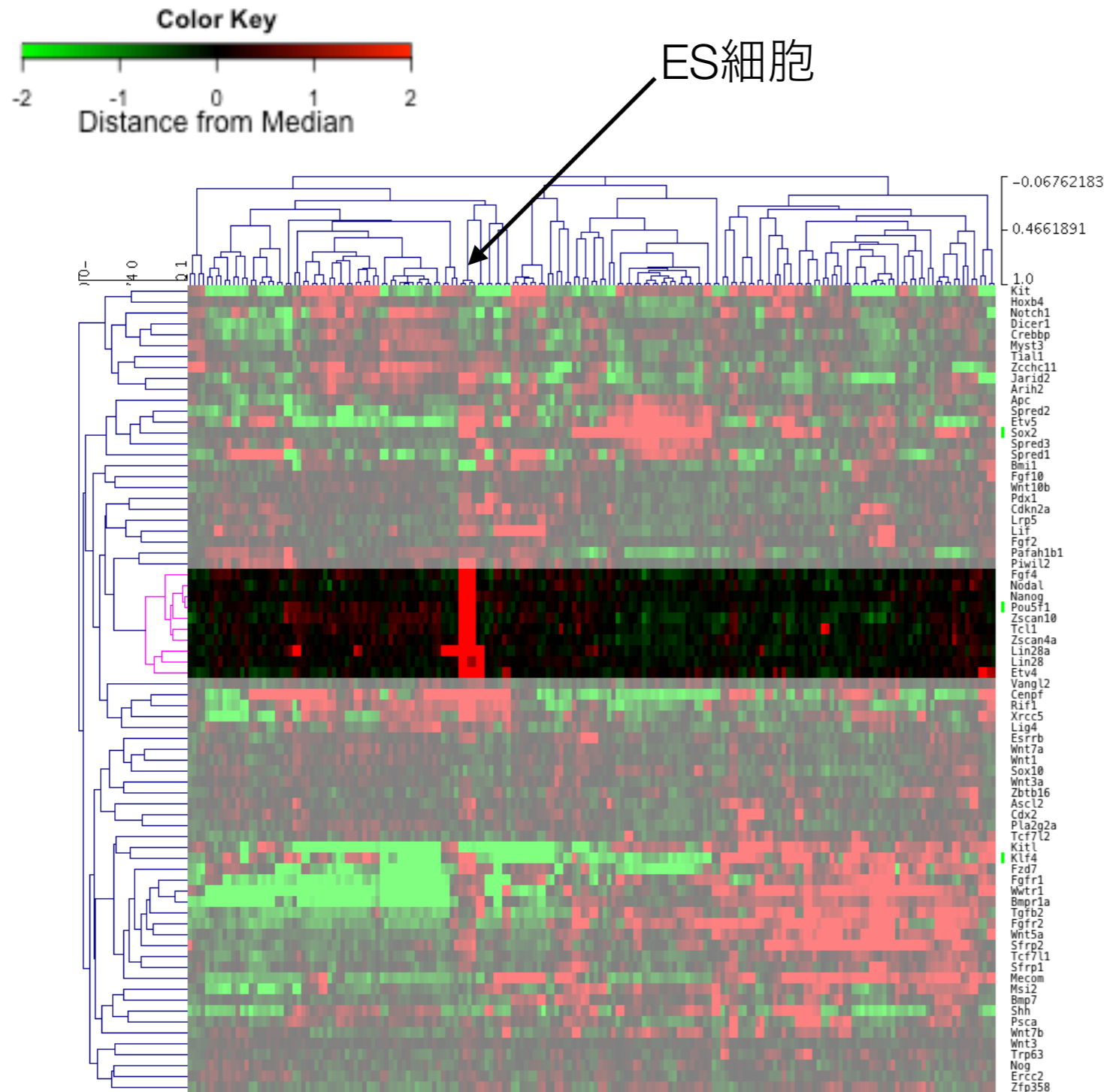
# ヒートマップ（クラスタリング図）を見るときのポイント

- 「カラーキー（スケール）」 = 「何を、何色にしているか」を確認。
  - 真ん中の色は？
  - 中央値 = 黒であれば、黒は発現していないという意味でない。
- 対象データは？
  - 一部の遺伝子だけ抽出していないか？
- クラスタリングしてある？
  - クラスタリング = ツリーの表示の有無。
  - 時系列データの場合は、時間で並んでいることも。

\*どれをどの濃さにするかは任意！



# stem cell 関連遺伝子のヒートマップ



- BioGPSより取得したマウスのマイクロアレイデータから、stem cell 関連遺伝子を抽出して作成されたヒートマップ。
- 横方向に、ES細胞をはじめ、liver, heart など 180サンプル並んでいる。
- \*GOのアノテーションに“stem cell”を含むものと定義。
- 何が読み取れるか？



# 機能解析 (functional analysis): GO, DAVID, GSEA

---

- 発現している遺伝子、変動している遺伝子の集団があった時、
  - 「生物学的に見て、どの機能 (biological function) を持った遺伝子が多いのか」を見る手法。
  - 広い意味で、「**機能解析**」と呼ばれる。
- 解析のために、様々なアルゴリズムがある。
  - 一番基礎的なものが、「**GO解析**」
  - データベースを利用したものとして、下記の2つが有名。**エンリッチメントアナリシス**という言い方もある。
    - The Database for Annotation, Visualization and Integrated Discovery (**DAVID**)
    - Gene Set Enrichment Analysis (**GSEA**)

# そもそも GO とは？

---

- Gene Ontology (ジーンオントロジー)
  - もともと、オントロジーとは、用語 (ターム) を構造化して整理する手法。
  - 遺伝子用のオントロジーなので、 Gene Ontology (GO)。
- 簡単にいうと、構造化された用語集。
- その用語集を、遺伝子ごとに割りあてる作業がアノテーションの1種。
- 例えば、Myc 遺伝子の持つGOは、
  - DNA binding, E-box binding, RNA polymerase II core promoter proximal region sequence-specific DNA binding, core promoter proximal region sequence-specific DNA binding, double-stranded DNA binding, protein binding, protein complex binding, protein dimerization activity, protein heterodimerization activity, etc.

# 構造化されているとは？

- ▶ GO:0005575 cellular\_component [445431 gene products]
- ▶ GO:0005623 cell [277375 gene products]
- ▶ GO:0044464 cell\_part [277326 gene products]
- ▶ GO:0005622 intracellular [249494 gene products]
  - ▶ GO:0044424 intracellular\_part [235566 gene products]
  - ▶ GO:0043226 organelle [192890 gene products]
    - ▶ GO:0043229 intracellular\_organelle [192707 gene products]
    - ▶ GO:0043227 membrane-bounded\_organelle [163999 gene products]
      - ▶ GO:0043231 intracellular\_membrane-bounded\_organelle [163826 gene products]
      - ▼ GO:0005634 nucleus [75407 gene products]
        - ▶ GO:0000794 condensed\_nuclear\_chromosome [1031 gene products]
        - ▶ GO:0000780 condensed\_nuclear\_chromosome\_centromeric\_region [382 gene products]
        - ▶ GO:0048555 generative\_cell\_nucleus [1 gene products]
        - ▶ GO:0043073 germ\_cell\_nucleus [160 gene products]
        - ▶ GO:0071686 horsetail\_nucleus [1 gene products]
        - ▶ GO:0031039 macronucleus [1 gene products]
        - ▶ GO:0043076 megasporocyte\_nucleus [6 gene products]
        - ▶ GO:0031040 micronucleus [0 gene products]
        - ▶ GO:0048556 microsporocyte\_nucleus [0 gene products]
        - ▶ GO:0031618 nuclear\_centromeric\_heterochromatin [22 gene products]
        - ▶ GO:0000790 nuclear\_chromatin [1908 gene products]
        - ▶ GO:0000228 nuclear\_chromosome [3974 gene products]
        - ▶ GO:0000784 nuclear\_chromosome\_telomeric\_region [296 gene products]
        - ▶ GO:0000798 nuclear\_cohesin\_complex [32 gene products]

- 細胞の中に、細胞膜と細胞質があって、細胞質の中に核があって、、、
- 細胞内における機能に関する用語も階層化されている。
- ただし、「親は複数あってもよい」という特殊な階層関係にある。  
(DAG)
- GO コンソーシアムで決定される。
- <http://geneontology.org>

# GO解析

---

- 変動した遺伝子の中に、
  - 転写因子はどれくらい含まれるのか？
  - 膜タンパク質が多いのか、少ないのか？
  - アポトーシスに影響があるのかどうか？
- GO解析とは、変動した遺伝子のアノテーションを集計して、どのタームが何個あるか、何割くらいかなどを調べること。
  - 実際には単純な数や割合で評価することはできない。
  - アノテーションに“kinase”を持つものは、もともとたくさんあるので、ある一定の割合で見つからないと、偶然でないとは言えない。-> 検定

# GOの問題点

---

- そもそも、GO解析のためにアノテーションを設計しているわけではないので、用語に気をつける必要がある。
  - apoptosis ではなく、GO では、apoptotic process (“apoptosis”で検索しても見つからない。)
  - 転写関連なら何でもという場合は、GO:0001071 nucleic acid binding transcription factor activity より GO:0006351 transcription, DNA-templated に多くの遺伝子が存在。
- 決められた用語しかない。
  - アノテーションから、転移に関連した遺伝子を調べたくても、“metastasis” というタームはない。

# GO解析の注意点

- 検定の結果、「inflammatory response」を持つ遺伝子が有意であった」 = 「炎症が亢進ではない」
- GO の inflammatory response には、negative regulation と positive regulation の両方が含まれている。
- 遺伝子によっては、negative, positive 両方のアノテーションが付いていることもある。

```
all : all [25485 gene products]
├── GO:0008150 : biological_process [24800 gene products]
│   ├── GO:0050896 : response to stimulus [7175 gene products]
│   │   ├── GO:0006950 : response to stress [2213 gene products]
│   │   │   ├── GO:0006952 : defense response [857 gene products]
│   │   │   │   └── GO:0006954 : inflammatory response [403 gene products]
│   │   │   │       ├── GO:0002526 : acute inflammatory response [102 gene products]
│   │   │   │       ├── GO:0002544 : chronic inflammatory response [13 gene products]
│   │   │   │       ├── GO:0002437 : inflammatory response to antigenic stimulus [44 gene products]
│   │   │   │       ├── GO:0002269 : leukocyte activation involved in inflammatory response [0 gene products]
│   │   │   │       ├── GO:0002523 : leukocyte migration involved in inflammatory response [9 gene products]
│   │   │   │       ├── GO:0050728 : negative regulation of inflammatory response [73 gene products]
│   │   │   │       └── GO:0050729 : positive regulation of inflammatory response [67 gene products]
│   │   │   │           └── GO:0002532 : production of molecular mediator involved in inflammatory response [30 gene products]
│   │   │   │               ├── GO:0050727 : regulation of inflammatory response [186 gene products]
│   │   │   │               └── negativeとpositive 2つのregulationが含まれる。
│   │   │   └── GO:0009611 : response to wounding [636 gene products]
│   │   │       ├── GO:0006954 : inflammatory response [403 gene products]
│   │   │       ├── GO:0002526 : acute inflammatory response [102 gene products]
│   │   │       ├── GO:0002544 : chronic inflammatory response [13 gene products]
│   │   │       ├── GO:0002437 : inflammatory response to antigenic stimulus [44 gene products]
│   │   │       ├── GO:0002269 : leukocyte activation involved in inflammatory response [0 gene products]
│   │   │       ├── GO:0002523 : leukocyte migration involved in inflammatory response [9 gene products]
│   │   │       ├── GO:0050728 : negative regulation of inflammatory response [73 gene products]
│   │   │       ├── GO:0050729 : positive regulation of inflammatory response [67 gene products]
│   │   │       ├── GO:0002532 : production of molecular mediator involved in inflammatory response [30 gene products]
│   │   │       ├── GO:0050727 : regulation of inflammatory response [186 gene products]
│   │   │       └── GO:0002246 : wound healing involved in inflammatory response [5 gene products]
```

# The Database for Annotation, Visualization and Integrated Discovery (**DAVID**)

- アノテーションのデータベースである DAVID を使うと、GO解析は簡単に行える。
- 使い方は、変動していた遺伝子のリスト（遺伝子名またはID）をアップロードするだけ。
- <https://david.ncifcrf.gov>

DAVID Functional Annotation | david.ncifcrf.gov

DAVID Bioinformatics Resources 6.7  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

Shortcut to DAVID Tools

- Functional Annotation
- Gene Functional Classification
- Gene ID Conversion
- Gene Name Batch Viewer

Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2015

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an 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

What's Important in DAVID?

- [Current \(v 6.7\) release note](#)
- [New requirement to cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)
- [Novel Classification Algorithms](#)
- [Pre-built Affymetrix and Illumina backgrounds](#)
- [User's customized gene background](#)
- [Enhanced calculating speed](#)

Statistics of DAVID

DAVID Bioinformatic Resources Citations

Year	Citations
2004	~100
2005	~200
2006	~300
2007	~400
2008	~500
2009	~600
2010	~700
2011	~800
2012	~900
2013	~1000
2014	~1100

- [> 17,000 Citations](#)
- Daily Usage: ~1200 gene lists/sublists from ~400 unique

# DAVIDの解析結果

149 Cluster(s) [Download File](#)

Annotation Cluster	Enrichment Score	Count	P-Value	Benjamini
<b>Annotation Cluster 1</b> Enrichment Score: 2.53				
GOTERM_BP_FAT	cell activation	18	1.1E-3	8.8E-1
GOTERM_BP_FAT	T cell activation	11	1.4E-3	7.4E-1
GOTERM_BP_FAT	lymphocyte activation	13	5.0E-3	7.9E-1
GOTERM_BP_FAT	leukocyte activation	14	9.3E-3	7.7E-1
<b>Annotation Cluster 2</b> Enrichment Score: 1.76				
SP_PIR_KEYWORDS	signal	113	2.4E-3	6.2E-1
UP_SEQ_FEATURE	signal peptide	113	3.0E-3	9.9E-1
GOTERM_CC_FAT	extracellular region	71	3.9E-3	6.9E-1
SP_PIR_KEYWORDS	Secreted	62	1.1E-2	7.6E-1
GOTERM_CC_FAT	extracellular region part	37	1.4E-2	6.5E-1
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	132	2.1E-2	1.0E0
SP_PIR_KEYWORDS	glycoprotein	136	2.5E-2	8.7E-1
GOTERM_CC_FAT	extracellular space	26	4.9E-2	8.8E-1
UP_SEQ_FEATURE	disulfide bond	86	1.5E-1	1.0E0
SP_PIR_KEYWORDS	disulfide bond	88	1.5E-1	9.3E-1
<b>Annotation Cluster 3</b> Enrichment Score: 1.67				
UP_SEQ_FEATURE	topological domain:Extracellular	97	3.1E-3	9.2E-1
GOTERM_CC_FAT	integral to membrane	158	1.0E-2	7.8E-1
GOTERM_CC_FAT	intrinsic to membrane	162	1.3E-2	7.3E-1
UP_SEQ_FEATURE	topological domain:Cytoplasmic	111	1.8E-2	1.0E0
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	132	2.1E-2	1.0E0
SP_PIR_KEYWORDS	glycoprotein	136	2.5E-2	8.7E-1
SP_PIR_KEYWORDS	receptor	56	2.9E-2	8.6E-1
GOTERM_CC_FAT	plasma membrane	114	2.9E-2	8.3E-1
SP_PIR_KEYWORDS	transmembrane	153	3.4E-2	8.2E-1
UP_SEQ_FEATURE	transmembrane region	152	3.5E-2	1.0E0
SP_PIR_KEYWORDS	membrane	188	3.8E-2	8.3E-1
SP_PIR_KEYWORDS	cell membrane	72	5.6E-2	8.3E-1

- 1つ1つのGOタームの p-value に加えて、似ているタームをまとめて、アノテーションクラスターとして表示。
- アノテーションクラスターごとに Enrichment Score が産出される。(Enrichment Score > 1.3 で有意差あり)
- 遺伝子の名前だけをアップロードするので、増加したか、減少したかは結果に影響しない。



# Gene Set Enrichment Analysis (GSEA)

**Overview**

**Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- ▶ **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- ▶ **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- ▶ **View documentation** describing GSEA and MSigDB.

**What's New**

14-Dec-2012: Version 2.0.9 of the GSEA desktop application is now available. This version makes GSEA compatible with both Java 6 and Java 7.

15-Oct-2012: Version 3.1 of the Molecular Signatures Database (MSigDB) is now available. Highlights include:

1. more than 1,000 new gene sets curated from publications,
2. a new collection of gene sets representing oncogenic pathway activation modules,
3. two new sources of gene sets representing canonical pathways, and
4. an improved mapping to common gene identifiers for all gene sets.

See the [MSigDB 3.1 Release Notes](#) for details. A minor update of the GSEA desktop application has also been released. See the [GSEA 2.0.8 Release Notes](#) for details.

01-Oct-2012: We recently submitted a manuscript to Statistical Methods in Medical Research which provides a systematic comparison of the GSEA method with other methods employing a "simpler" t-test assessment of enrichment.

01-Oct-2012: We now have a CiteULike page where users can browse a list of all citations of the GSEA algorithm.

**Registration**

Please [register](#) to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

**Contributors**

GSEA and MSigDB are maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

**Citing GSEA**

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Broad Home | Cancer Genomics

MSigDB database v3.1 updated Sep 27, 2012  
GSEA/MSigDB web site v3.86 released Dec 13, 2012

- GSEA は、DAVID と似ているが、増加したか、減少したかも考慮した上で Enrichment Score を算出。
- 背景に MSigDB というデータベースがあり、「遺伝子セット」という概念がある。
- <http://www.broadinstitute.org/gsea/index.jsp>

# Gene Set (遺伝子セット)

---

- GO には、定義された用語でないと使えないという制限があった。
  - では、誰かの論文で、metastasis が起きた時、変動のあった遺伝子をリストにしておけば？
  - 肝臓癌で発現が増加した遺伝子のリスト
  - EMT で増加（減少）した時の遺伝子のリスト、など。
- MSigDB には、キュレーターが選択した論文から拾い出した遺伝子のリストがデータベース化されている。
- それらの「遺伝子リスト」について、検定を行う。
- 特定のGOのタームに載っている遺伝子も、同様に「遺伝子リスト」として扱うことで、GO解析も可能。

ABBUD_LIF_SIGNALING_2_DN	HOLLEMAN_PREDNISOLONE_RESISTANCE_B_ALL_DN	PENG_LEUCINE_DEPRIVATION_DN
ABBUD_LIF_SIGNALING_2_UP	HOLLEMAN_PREDNISOLONE_RESISTANCE_B_ALL_UP	PENG_LEUCINE_DEPRIVATION_UP
ABDELMOHSEN_ELAVL4_TARGETS	HOLLEMAN_VINCRIStINE_RESISTANCE_ALL_DN	PENG_RAPAMYCIN_RESPONSE_DN
ABDULRAHMAN_KIDNEY_CANCER_VHL_DN	HOLLEMAN_VINCRIStINE_RESISTANCE_ALL_UP	PENG_RAPAMYCIN_RESPONSE_UP
ABDULRAHMAN_KIDNEY_CANCER_VHL_UP	HOLLEMAN_VINCRIStINE_RESISTANCE_B_A_LL_DN	PEPPER_CHRONIC_LYMPHOCYtic_LEUKEMIA_DN
ABE_INNER_EAR	HOLLEMAN_VINCRIStINE_RESISTANCE_B_A_LL_UP	PEPPER_CHRONIC_LYMPHOCYtic_LEUKEMIA_UP
ABE_VEGFA_TARGETS	HOLLMANN_APOPTOSIS_VIA_CD40_DN	PEREZ_TP53_AND_TP63_TARGETS
ABE_VEGFA_TARGETS_2HR	HOLLMANN_APOPTOSIS_VIA_CD40_UP	PEREZ_TP53_TARGETS
ABE_VEGFA_TARGETS_30MIN	HONMA_DOCETAXEL_RESISTANCE	PEREZ_TP63_TARGETS
ABRAHAM_ALPC_VS_MULTIPLE_MYELOMA_DN	HONRADO_BREAST_CANCER_BRCA1_VS_BRCA2	PETRETTO_BLOOD_PRESSURE_DN
ABRAHAM_ALPC_VS_MULTIPLE_MYELOMA_UP	HOOI_ST7_TARGETS_DN	PETRETTO_BLOOD_PRESSURE_UP
ABRAMSON_INTERACT_WITH_AIRE	HOOI_ST7_TARGETS_UP	PETRETTO_CARDIAC_HYPERTROPHY
ACEVEDO_FGFR1_TARGETS_IN_PROSTATE_CANCER_MODEL_DN	HOQUE_METHYLATED_IN_CANCER	PETRETTO_HEART_MASS_QTL_CIS_DN
ACEVEDO_FGFR1_TARGETS_IN_PROSTATE_CANCER_MODEL_UP	HORIUCHI_WTAP_TARGETS_DN	PETRETTO_HEART_MASS_QTL_CIS_UP
ACEVEDO_LIVER_CANCER_DN	HORIUCHI_WTAP_TARGETS_UP	PETRETTO_LEFT_VENTRICLE_MASS_QTL_CIS_DN
ACEVEDO_LIVER_CANCER_UP	HORTON_SREBF_TARGETS	PETRETTO_LEFT_VENTRICLE_MASS_QTL_CIS_UP
ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_DN	HOSHIDA_LIVER_CANCER_LATE_RECURRENCE_DN	PETROVA_ENDOTHELium_LYMPHATIC_VS_BLOOD_DN
ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_UP	HOSHIDA_LIVER_CANCER_LATE_RECURRENCE_UP	PETROVA_ENDOTHELium_LYMPHATIC_VS_BLOOD_UP
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_DN	HOSHIDA_LIVER_CANCER_SUBCLASS_S1	PETROVA_PROX1_TARGETS_DN
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_UP	HOSHIDA_LIVER_CANCER_SUBCLASS_S2	PETROVA_PROX1_TARGETS_UP
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_DN	HOSHIDA_LIVER_CANCER_SUBCLASS_S3	PHESSE_TARGETS_OF_APC_AND_MBD2_DN
ACEVEDO_LIVER_CANCER_WITH_H3K9ME3_UP	HOSHIDA_LIVER_CANCER_SURVIVAL_DN	PHESSE_TARGETS_OF_APC_AND_MBD2_UP
ACEVEDO_LIVER_TUMOR_VS_NORMAL_ADJACENT_TISSUE_DN	HOSHIDA_LIVER_CANCER_SURVIVAL_UP	PHONG_TNF_RESPONSE_NOT_VIA_P38
ACEVEDO_LIVER_TUMOR_VS_NORMAL_ADJACENT_TISSUE_UP	HOUSTIS_ROS	PHONG_TNF_RESPONSE_VIA_P38_COMPLETE
ACEVEDO_METHYLATED_IN_LIVER_CANCER_DN	HOWLIN_CITED1_TARGETS_1_DN	PHONG_TNF_RESPONSE_VIA_P38_PARTIAL
ACEVEDO_NORMAL_TISSUE_ADJACENT_TO_LIVER_TUMOR_DN	HOWLIN_CITED1_TARGETS_1_UP	PHONG_TNF_TARGETS_DN
ACEVEDO_NORMAL_TISSUE_ADJACENT_TO_LIVER_TUMOR_UP	HOWLIN_CITED1_TARGETS_2_DN	PHONG_TNF_TARGETS_UP
ACOSTA_PROLIFERATION_INDEPENDENT_MYC_TARGETS_DN	HOWLIN_CITED1_TARGETS_2_UP	PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_DN
ACOSTA_PROLIFERATION_INDEPENDENT_MYC_TARGETS_UP	HOWLIN_PUBERTAL_MAMMARY_GLAND	PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_UP
ADDYA_ERYTHROID_DIFFERENTIATION_BY_HEMIN	HSIAO_HOUSEKEEPING_GENES	PIEPOLI_LGI1_TARGETS_DN
AFFAR_YY1_TARGETS_DN	HSIAO_LIVER_SPECIFIC_GENES	PIEPOLI_LGI1_TARGETS_UP
AFFAR_YY1_TARGETS_UP	HU_ANGIOGENESIS_DN	PILON_KLF1_TARGETS_DN
AGARWAL_AKT_PATHWAY_TARGETS	HU_ANGIOGENESIS_UP	PILON_KLF1_TARGETS_UP
AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_DN	HU_GENOTOXIC_DAMAGE_24HR	PIONTEK_PKD1_TARGETS_DN
AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_UP	HU_GENOTOXIC_DAMAGE_4HR	PIONTEK_PKD1_TARGETS_UP
AIGNER_ZEB1_TARGETS	HU_GENOTOXIN_ACTION_DIRECT_VS_INDIRECT_24HR	PLASARI_NFIC_TARGETS_BASAL_DN
AIYAR_COBRA1_TARGETS_DN	HU_GENOTOXIN_ACTION_DIRECT_VS_INDIRECT_4HR	PLASARI_NFIC_TARGETS_BASAL_UP
AIYAR_COBRA1_TARGETS_UP	HUANG_DASATINIB_RESISTANCE_DN	PLASARI_TGFB1_SIGNALING_VIA_NFIC_10HR_DN
AKL_HTLV1_INFECTIOn_DN	HUANG_DASATINIB_RESISTANCE_UP	PLASARI_TGFB1_SIGNALING_VIA_NFIC_10HR_UP
AKL_HTLV1_INFECTIOn_UP	HUANG_FOXA2_TARGETS_DN	PLASARI_TGFB1_SIGNALING_VIA_NFIC_1HR_DN
ALCALA_APOPTOSIS	HUANG_FOXA2_TARGETS_UP	PLASARI_TGFB1_SIGNALING_VIA_NFIC_1HR_UP
ALCALAY_AML_BY_NPM1_LOCALIZATION_DN	HUANG_GATA2_TARGETS_DN	PLASARI_TGFB1_TARGETS_10HR_DN
ALCALAY_AML_BY_NPM1_LOCALIZATION_UP	HUANG_GATA2_TARGETS_UP	PLASARI_TGFB1_TARGETS_10HR_UP
ALFANO_MYC_TARGETS	HUI_MAPK14_TARGETS_UP	PLASARI_TGFB1_TARGETS_1HR_DN
ALONSO_METASTASIS_DN	HUMMEL_BURKITTs_LYMPHOMA_DN	PLASARI_TGFB1_TARGETS_1HR_UP
ALONSO_METASTASIS_EMT_DN	HUMMEL_BURKITTs_LYMPHOMA_UP	PODAR_RESPONSE_TO_ADAPHOSTIN_DN
ALONSO_METASTASIS_EMT_UP	HUMMERICH_BENIGN_SKIN_TUMOR_DN	PODAR_RESPONSE_TO_ADAPHOSTIN_UP
ALONSO_METASTASIS_NEURAL_UP	HUMMERICH_BENIGN_SKIN_TUMOR_UP	POMEROY_MEDULLOBLASTOMA_DESMOPLASIC_VS_CLASSIC_DN
ALONSO_METASTASIS_UP	HUMMERICH_MALIGNANT_SKIN_TUMOR_DN	POMEROY_MEDULLOBLASTOMA_DESMOPLASIC_VS_CLASSIC_UP
ALTEMEIER_RESPONSE_TO_LPS_WITH_MECHANICAL_VENTILATION	HUMMERICH_MALIGNANT_SKIN_TUMOR_UP	POMEROY_MEDULLOBLASTOMA_PROGNOSIS_DN
	HUMMERICH_SKIN_CANCER_PROGRESSION_DN	
	HUMMERICH_SKIN_CANCER_PROGRESSION_UP	

\*MSigDB に登録された  
「遺伝子セット」の一部

すべての論文のリストがあるわけではない。分野によっては偏りがある。

# GSEAの注意点

- Enrichment Score の判定には、ランクを用いる。
- n数が少ないと、シグナル値が低く、ノイズの可能性もある遺伝子が一律に扱われる。
- DAVID と違い、up, down を評価する。
  - 論文から取られた遺伝子リストは up, down に意味がある。
  - GO のタームは、 up か down か分からない。

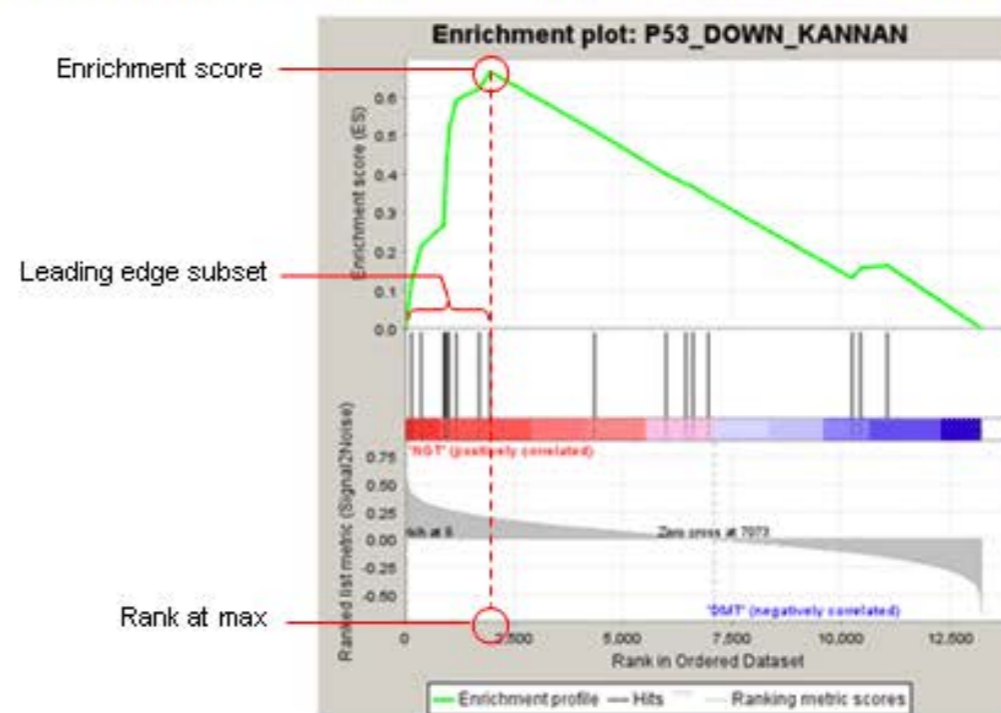
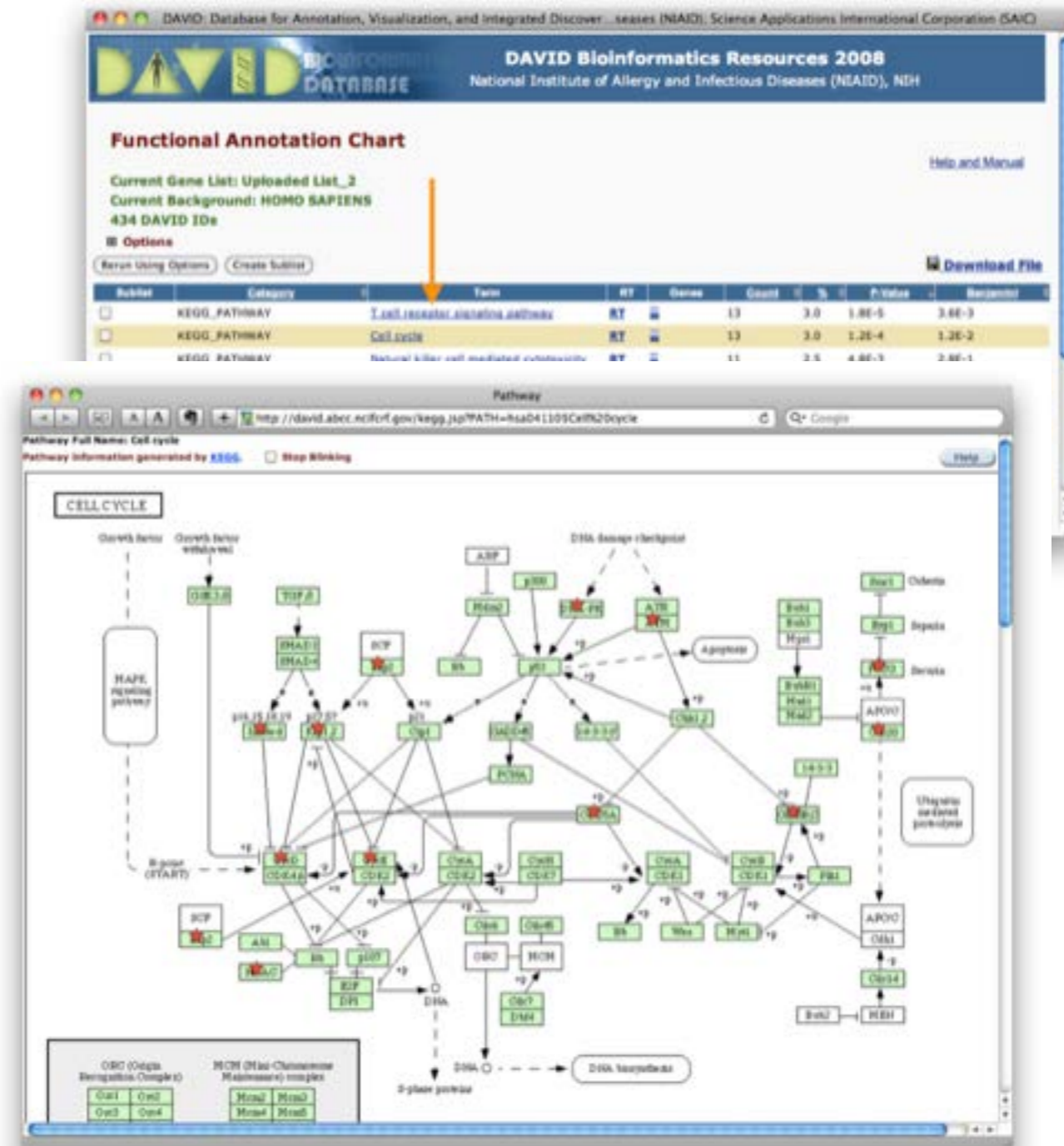


Fig 1: Enrichment plot: P53\_DOWN\_KANNAN  
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

\*GSEAガイドより

# パスウェイ解析

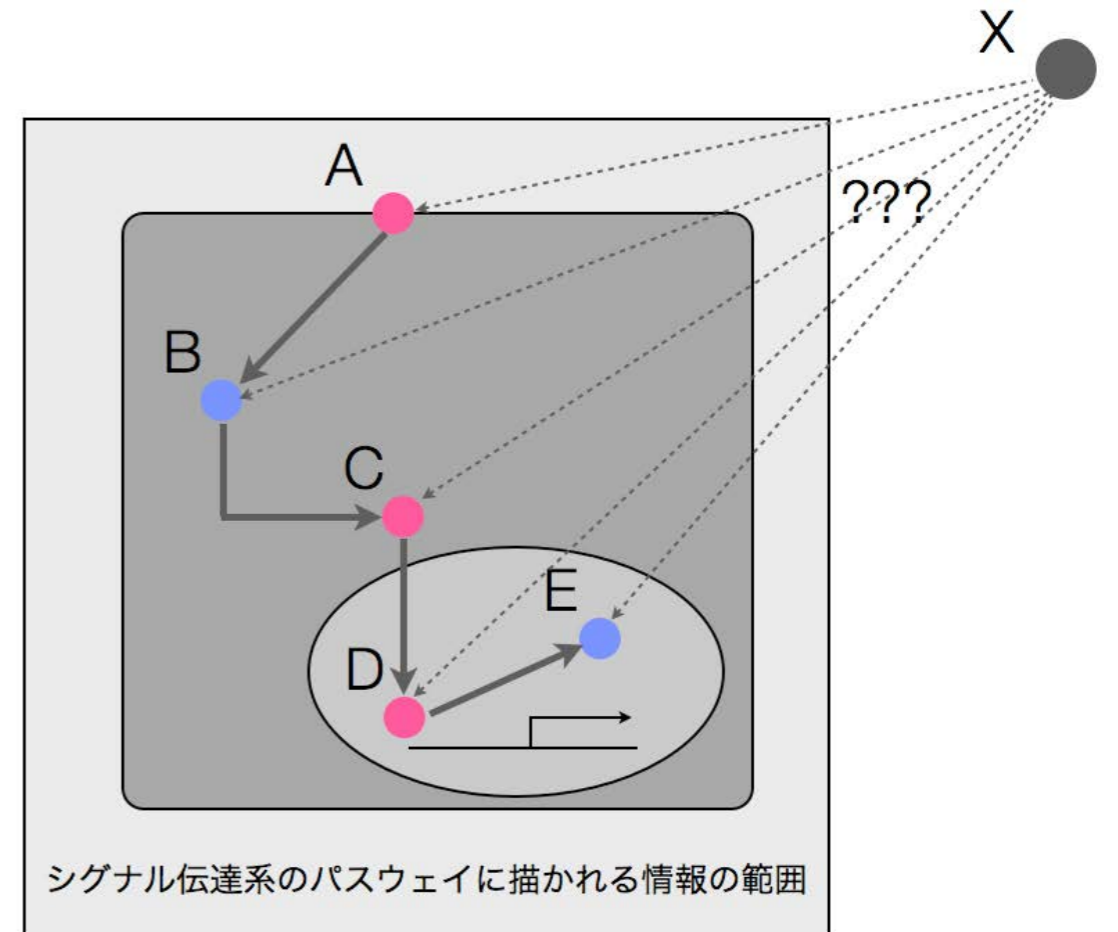
- いわゆるパスウェイ解析は、変動していた遺伝子が、特定のパスウェイに含まれているかどうかを検定。
- KEGG に登録されたパスウェイ中に含まれている遺伝子群を「遺伝子リスト」と捉え、GO解析の延長で対応されている。
- DAVID や GSEA でも解析が可能。
- KEGG パスウェイデータベース  
<http://www.genome.jp/kegg/pathway.html>



cell innovator

# パスウェイ解析の限界

- パスウェイ解析も、GO解析と同様の問題がある。
  - 有意になったパスウェイが、活性化されたか、抑制されたかは、厳密には評価できない。
  - GSEA のようにデータを up, down を分けたとしても、パスウェイの遺伝子セットが up, down に分かれていない。
- そもそも、パスウェイの外にある遺伝子は対象外。



# ネットワーク解析の考え方（データドリブン）

---

- パスウェイ解析で、パスウェイが活性化されたかどうか分からない。
  - → 活性化された遺伝子だけでパスウェイを作成したらどうか？
- パスウェイを制御する遺伝子が、パスウェイに載っていない。
  - → パスウェイに含まれる遺伝子に関係ありそうな遺伝子をデータから探せないか？
- 遺伝子発現データ（マイクロアレイやNGSのデータ）から、動きだけから、ネットワークを動的に作成する。（データドリブン）
  - 相関関係のある遺伝子でネットワークを作成 → Weighted correlation network analysis (WGCNA) など。

# 解析例：WCGNA, DAVID

**Table 1 Weighted gene-coexpression network analyses revealed significantly altered gene modules in c9ALS cerebellum and frontal cortex**

Module name	P value	Gene count	Notable GO terms	Enrichment P value
<b>c9ALS cerebellum</b>				
MEpink	0.0137	148	Neuron development, protein localization, transcription	$3.40 \times 10^{-3}$
MEgray60	0.0350	93	Vesicle transport, protein transport	$1.90 \times 10^{-2}$
MEdarkred	0.0359	68	rRNA processing, lysosome organization, RNA processing	$3.30 \times 10^{-2}$
MEgreen	0.0493	210	Intracellular transport	$1.60 \times 10^{-4}$
MEblue	0.0497	299	Chromatin modification	$4.60 \times 10^{-9}$
<b>c9ALS frontal cortex</b>				
MEsalmon	0.0217	43	Response to unfolded protein	$1.40 \times 10^{-8}$
MEturquoise	0.0413	1,678	Protein localization	$8.10 \times 10^{-25}$
MEblack	0.0428	66	Oxidative phosphorylation	$7.50 \times 10^{-8}$
MEgreen	0.0457	98	rRNA processing	$8.20 \times 10^{-3}$

WGCNA modules were done with data from c9ALS cerebellum and frontal cortex (versus sALS and controls;  $P < 0.05$ ). All significant modules are shown. The most notable GO terms and their enrichment  $P$  values are included for each module. Note that MEpink is shown in blue in Figure 2a, and MEsalmon is shown in red in Figure 2b.

the frontal cortex was enriched in UPR-related genes (Table 1 and Fig. 2b). Of note, genes involved in the UPR pathway were among the top DE genes in both cerebellum and frontal cortex in c9ALS subjects as determined by EdgeR analyses, as mentioned above, and validated by quantitative real-time PCR (qRT-PCR) (Supplementary Fig. 4). In sALS, the top modules identified included genes involved in calcium transport and synaptic transmission in cerebellum and in oxidative phosphorylation in frontal cortex (Supplementary Table 6). These studies revealed marked differences in gene-expression patterns between c9ALS and sALS subjects, with substantially more changes observed in individuals with c9ALS, especially in the cerebellum.

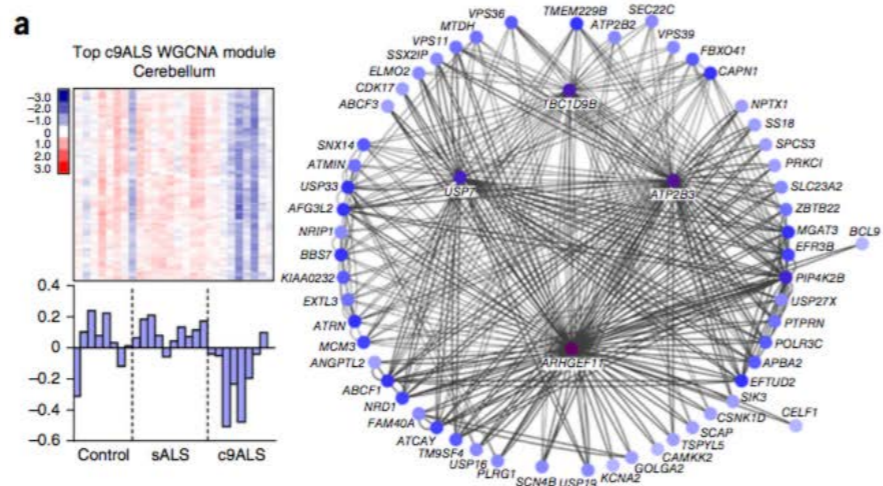
**Extensive misregulation of AS in ALS brain**  
To evaluate AS changes in ALS, we used OLEGO software<sup>30</sup> to align RNA-Seq reads to the hg19 genome assembly and splice junctions. The total number of AS events (false discovery rate (FDR) < 0.05) was more than three times higher in c9ALS subjects than in sALS subjects in both cerebellum (8,224 in c9ALS versus 2,229 in sALS) and frontal cortex (920 in c9ALS versus 282 in sALS) (Fig. 3a and Supplementary Tables 7 and 8). Also, the number of AS events was

approximately eight to nine times higher in the cerebellum than in the frontal cortex in c9ALS and sALS samples (Fig. 3a). Although 1,172 and 106 of the AS events that occurred in the cerebellum and frontal cortex, respectively, were shared between c9ALS and sALS subjects, they represented a relatively small percentage (~11%–15%) of the total AS changes seen in c9ALS subjects.

Among the different types of AS changes noted in ALS, cassette exon (CE) events were the most common, and intron-retention events represented a significant proportion of total changes (FDR < 0.05; Fig. 3b). A total of 918 intron-retention events occurred in c9ALS cerebellum, whereas 286 were found in c9ALS frontal cortex. In sALS subjects, there were 378 alternative intron-retention events in the cerebellum but only 7 in the frontal cortex (Fig. 3b and Supplementary Fig. 5). In c9ALS subjects, approximately 12 times more AS CE events (FDR < 0.05) were found in the cerebellum (4,419) than in the frontal cortex (369) (Fig. 3b and Fig. 4a,b). In sALS subjects, there were 949 AS CE events in the cerebellum, more than four times the number of events in the frontal cortex (203) (Fig. 3b and Fig. 4a,b).

The majority of CE events in c9ALS were the result of exon skipping, whereas similar proportions of CE exclusions and inclusions

- Prudencio et al., Nat. Neurosci. 2015 Aug;18(8): 1175-82.
- RNA-Seq のデータから、ネットワークを作成して、クラスターの機能を DAVID で解析





# まとめ

---

- その他のデータの見方について、下記のサイトで解説しています。
- <http://array.cell-innovator.com>

# 参考URL

---

- BioGPS - <http://biogps.org/>
- GEO - <http://www.ncbi.nlm.nih.gov/geo/>
- Connectivity Map - <https://www.broadinstitute.org/cmap/>
- TCGA - <https://www.broadinstitute.org/cmap/>
- cBioPortal - <http://www.cbioportal.org>
- DAVID - <https://david.ncifcrf.gov>
- GSEA - <http://www.broadinstitute.org/gsea/index.jsp>
- KEGG - <http://www.genome.jp/kegg/pathway.html>
- WCGNA - <http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/>