

COLUMN7 付録

Excelを使った遺伝子発現データ解析法

太田紀夫 科学技術振興機構情報基盤事業部 NBDC事業推進室

NCBI Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) の実験セットを用いた解析例、ケース1とケース2を紹介する。最後にExcelファイルの作り方や応用例についても解説する。COLUMN7も参照のこと。

■ ケース1

実験セットGSE7032（マウスの白色脂肪細胞と褐色脂肪細胞の分化実験）を用いた解析

GSE7032については <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7032> から詳細が見られる。

ここでは、変動比とP値で変動遺伝子を選抜することを行う。これは、Volcano plotとも呼ばれ、NCBI GEO2Rでも採用している手法である。大規模コホートの臨床サンプルなど、N数が大きいデータに向いている手法である。細胞株実験などの実験的なばらつきが小さいデータにも有効だが、その場合は統計学的解析に足るサンプル数がないデータに検定手法を使うので、結果は慎重に解釈することが必要になる。

1. Excel ファイル

この付録では、以下の3つのExcelファイルを用意した。

- (a) GSE7032_Fc-P.xlsx
- (b) GSE7032_Fc-P_practice.xlsx
- (c) GSE7032_Fc-P_formula.xlsx

(a)は解析用で、動作を軽くするために固定値セルの計算式を値に変換している。(b)は解析用ワークシート作成の練習用ファイル。各エリアの左上のセルに計算式を入れてあり、そのセルをコピペすれば(c)の解析用ワークシートができる。(c)の固定値セルの計算式を値に変換したもののが(a)の解析用ワークシートになる。

各Excelファイルには複数のシートが含まれているので、以下に構成を説明する。

(a) GSE7032_Fc-P.xlsxの構成

- ① GSE7032_samples : サンプル情報
- ② GSE7032_data : 解析用ワークシート
- ③ 1_forGraph : 選択プローブのシグナル値グラフ
- ④ 2_NonDiff(FC-P)Scattered : 分散グラフ
- ⑤ 3_NonDiff(FC-P)Volcano : Volcano plot
- ⑥ DiffBrvsWh_FC1.0-P0.05Up : 褐色脂肪細胞で発現上昇したプローブ
- ⑦ DiffBrvsWh_FC1.0-P0.05Down : 褐色脂肪細胞で発現低下したプローブ

(b) GSE7032_Fc-P_practice.xlsxの構成

- ① GPL81 : プローブ情報
- ② GSE7032_data : 解析用ワークシート ((2)は解説付き)
- ③ GSE7032_series matrix : GSE7032のseries matrix
- ④ GSE7032_samples : GSE7032のサンプル情報

- ⑤ 1_forGraph : 選択プローブのシグナル値グラフ
- ⑥ 2_NonDiff(FC-P)Scattered : 分散グラフ
- ⑦ 3_NonDiff(FC-P)Volcano : Volcano plot

(c) GSE7032_Fc-P_formula.xlsxの構成

- ① GPL81 : プローブ情報
- ② GSE7032_samples : サンプル情報
- ③ GSE7032_data : 解析用ワークシート
- ④ 1_forGraph : 選択プローブのシグナル値グラフ

2. GSE7032_Fc-P.xlsxの使い方

(1) 解析用ワークシート (GSE7032_data) の構成について (付録図1)

変動倍率 (Fc) と t 検定の P 値 (P) で変動遺伝子を選抜するための解析シートである (付録図1)。赤枠部分が series matrix のシグナルマトリクス。その左側にプローブ情報、上に数値分布、右側に変動プローブを選抜するためのテーブルを付けた。群平均値と P 値を計算し、詳細フィルターで変動プローブを選択する。群平均値と P 値のテーブルの上に、詳細フィルターの検索条件範囲で参照するセルを付けた。

なお、Excelファイルの GSE7032_Fc-P_formula.xlsx には計算式が残されている。

付録図1 GSE7032_Fc-P.xlsx のシート GSE7032_data

(2) 特定の遺伝子の発現情報を調べる

特定遺伝子の発現を調べる際は、遺伝子名で Gene Symbol (C列) にフィルターをかける。ワイルドカード (「?」や「*」) を使うと、一連のファミリー遺伝子をまとめて検索できる (例: 「IL*」)。

(3) 変動比とP値で発現変動プローブを選抜する (AL~AT列)

詳細フィルターで褐色脂肪前駆細胞 vs 白色脂肪前駆細胞で変動比 2 倍以上かつ P 値が 0.05 未満のプローブを選抜する。

● 発現上昇プローブの選抜 (付録図2)

[GSE7032_data!\$A\$42:\$BW\$12530]にフィルターをかける。

データ>詳細設定 から

- ・抽出先：選択範囲内
- ・リスト範囲：[\$A\$42:\$BW\$12530]
- ・検索条件範囲：[GSE7032_data!\$AM\$13:\$AN\$14]

<OK>を押すと196プローブが選択される。

各群のシグナル平均値、FC値・-logP値を参照するセル ([AO27:AZ27]) を[AO:AZ]列の選択されたセルにコピペする。

フィルターをクリアする。

付録図2 フィルターをかける

● 発現低下プローブの選抜

データ>詳細設定 から、

- ・リスト範囲：[\$A\$42:\$BW\$12530]
- ・検索条件範囲：[GSE7032_data!\$AM\$16:\$AN\$17]

<OK>を押すと571プローブが選択される。

各群のシグナル平均値、FC値・-logP値を参照するセル ([AO28:AZ28]) を[AO:AZ]列の選択されたセルにコピペする。

フィルターをクリアする。

● 上記以外のプローブの選抜 (付録図3)

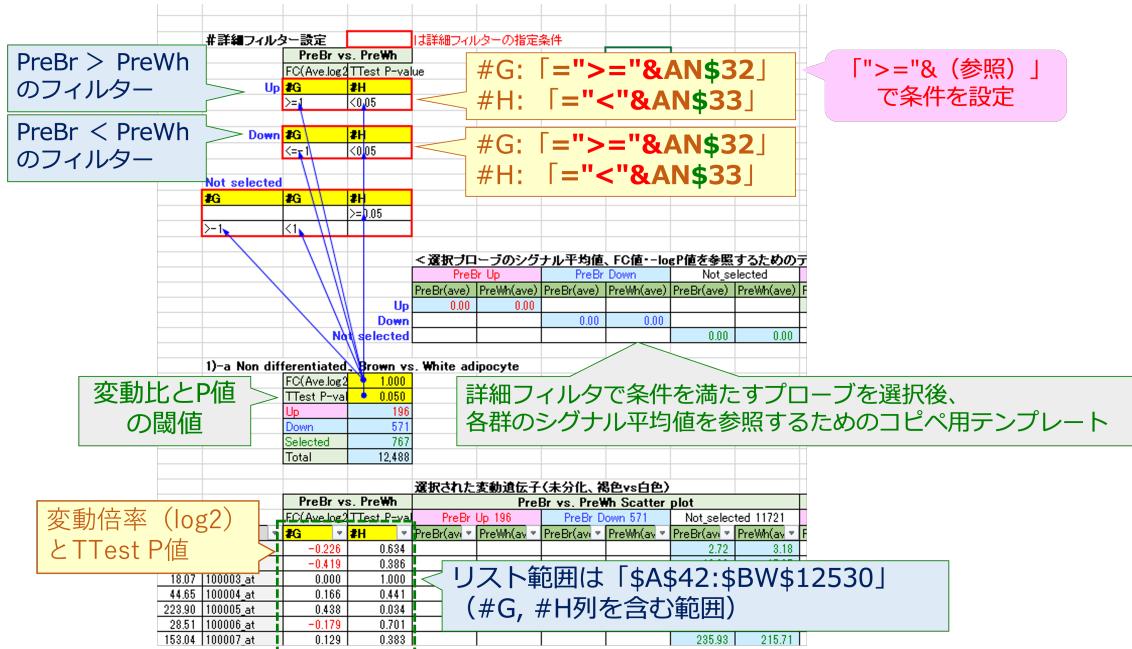
データ>詳細設定 から、

- ・リスト範囲：[\$A\$42:\$BW\$12530]
- ・検索条件範囲：[GSE7032_data!\$AL\$20:\$AN\$22]

<OK>を押すと11721プローブが選択される。

各群のシグナル平均値、FC値・-logP値を参照するセル ([AO29:AZ29]) を[AO:AZ]列の選択されたセルにコピペする。

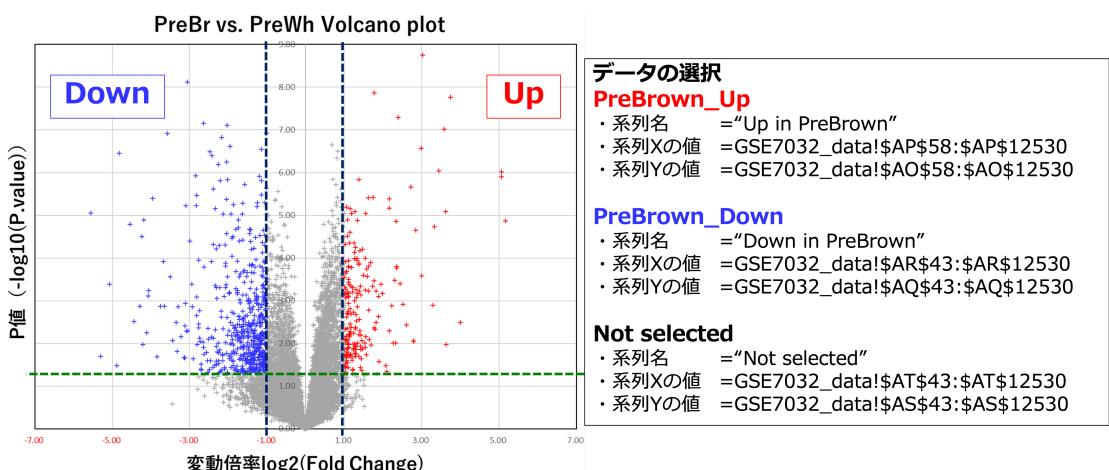
フィルターをクリアする。



付録図3 変動比とP値に基づく変動遺伝子の選抜

● 発現変動プローブの可視化

2_NonDiff(FC-P)Scattered シートに分散グラフを、3_NonDiff(FC-P)Volcanoシートには Volcano plot (付録図4左) を表示した。発現上昇プローブは赤、低下プローブは青、それ以外は灰色で表示されている。GSE7032_dataシートでフィルターをかけると、絞り込んだプローブだけが表示される。



付録図4 Volcano plot。右は各系列でプロットしているデータ範囲（「データの選択」で表示）。

(4) 特定の機能に関連する発現変動遺伝子を選抜する (BP~BW列)

● 複合条件での選抜 (付録図5)

詳細フィルターで褐色脂肪細胞 vs 白色脂肪細胞で変動比2倍以上かつP値が0.05未満で、かつGO Biological Process (E列) に"regulation of transcription"または"DNA-templated"、GO Cellular Componentに"nucleus"を含むプローブを選択する。GOの検索条件は検索対象の文字列を「""」で囲み、前後に「*」を付けて「=*"文字列"*」で表す（例：="*regulation of transcription, DNA-templated*")。詳細フィルターの検索条件範囲は行方向が「AND条件」、列方向が「OR条件」なので、発現上昇プローブと低下プローブの数値条件と文字列条件をそれぞれ別の行で「AND条件」で指定し、発現上昇プローブと低下プローブの「OR条件」で検索をする。

[GSE7032 data!\$A\$42:\$BW\$12530]にフィルターをかける。

データ> 詳細設定 から

- ・抽出先：選択範囲内
 - ・リスト範囲：[\$A\$42:\$BW\$12530]
 - ・検索条件範囲：[G\$F7032_data!\$B\$20:\$BW\$22]

<OK>を押すと158プローブが選択される。

各群のFC値・P値と遺伝子情報を参照するセル ([BQ27:BW27]) を[BQ: BW]列の選択されたセルにコピペする

フィルターをクリアする

数値条件は文字列関数を使って
「”（記号）“&参照セル”」で指定

文字列は「=* (文字列) *」で指定

DiffBr Upの検索条件範囲

#I: = ">=&\$BR\$32、 #J : = "<&\$BR\$33
#D: = "regulation of transcription, DNA-temp
#E: = "nucleus"

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

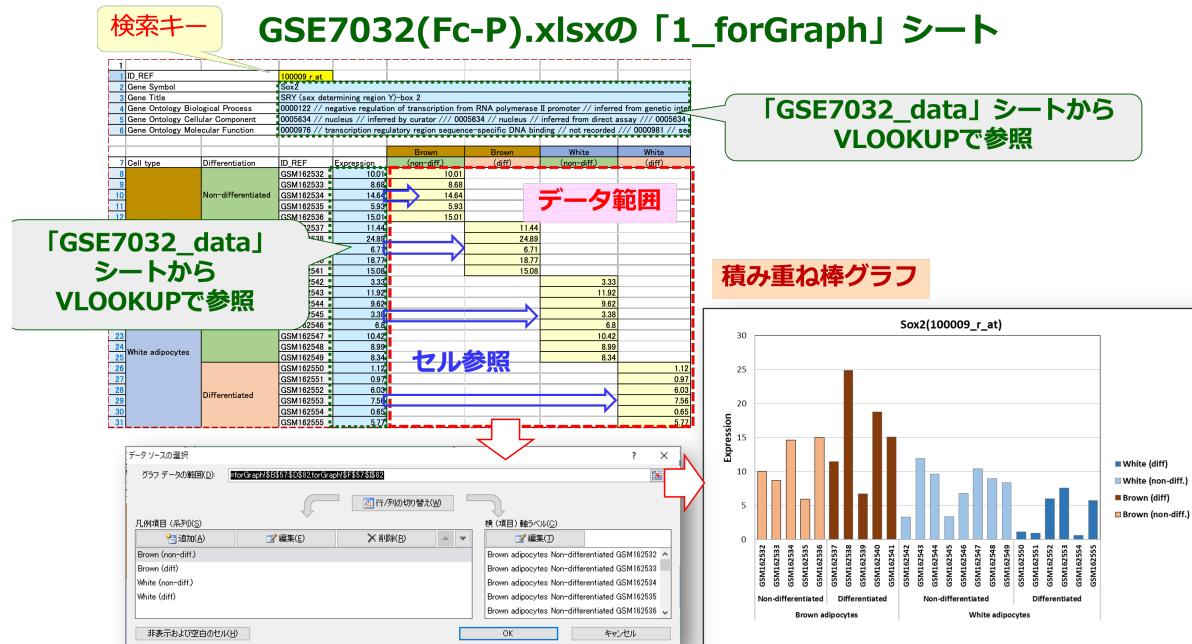
変動倍率とP値

	1	2	3	4	5	6
Eckart-#	11	10	14	15	16	
ID	FEF	Gene Symbol	Gene Title	Gene Ontology	Gene Ontology	Gene Ontology
#A	#B	#C	#D	#E	#F	
000001	Cd3g	CD3 antigen, polypeptide chain gamma	0001163 / es_0016020	0004888 / tra_0004888		
000002	Itih3	inter-alpha tryptase inhibitor, heparan-binding, member 3	0010468 / es_0055768	0004867 / tra_0004867		
000003	Ryr1	ryanodine receptor 1	0001666 / es_005622	0002020 / tra_0002020		
000004	In7s	integrator complex subunit 7	0000077 / es_005634	0004588 / tra_0004588		
000005	Traf4	TNF receptor type I, apoprotein B-100	0001995 / ap_0005634	0004842 / tra_0004842		
000006	Cdh11	cadherin 11	0001159 / es_0017537	0005009 / tra_0005009		
000007	Irfp2bp1	interferon regulatory factor 2 binding protein 1	0001222 / es_005634	0003714 / tra_0003714		
000009	Sox2	SRY (sex determining region Y)-box 2	0001223 / es_005634	0000978 / tra_0000978		
00010	Kif3	Kruppel-like factor 3	0000351 / es_005634	0003876 / tra_0003876		

付録図5 複合条件による選抜

(5) 選抜プローブのシグナル値をグラフ表示する (1_forGraph)

1_forGraphのシートを見てみよう。プローブIDをキーとしてVLOOKUPで表示するプローブの情報とシグナル値をGSE7032_dataシートから参照し、グラフ表示する。群ごとにシグナル値の表示列をずらし（F～I列），全体をグラフデータの範囲（例：[F59:I82]）に指定して積み重ねグラフにすると簡単に色分けグラフになる（[付録図6](#)）。



付録図6 個別グラフの作成

■ ケース2

実験セットGSE7032（非侵襲性乳管がん（DCIS）と侵襲性乳管がん（IDC）のプロファイル）を用いた解析

GSE7032については<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21422>から詳細が見られる。

ここでは、シグナル値と変動比に閾値を設定し、群ごとにn/N以上で条件を満たす変動遺伝子を選抜する。検体数が少ない臨床検体や動物実験で作製した病態モデルなど、ばらつきが大きいデータにも有効な選抜方法である

1. Excelファイル

この付録では、以下の3つのExcelファイルを用意した。

- (a) GSE21422_Exp-Fc.xlsx
- (b) GSE21422_Exp-Fc_practice.xlsx
- (c) GSE21422_Exp-Fc_formula.xlsx

(a) GSE21422_Exp-Fc.xlsxの構成

- ① GPL570：プローブ情報
- ② GSE21422_samples：サンプル情報
- ③ GSE21422_data：解析用ワークシート
- ④ DCISUpDown_Scattered：分散グラフ
- ⑤ DCISUpDown 6 of 9 (1763)：DCIS群の6/9サンプル以上でhealthy群のメジアン値より3倍以上に変動した1763プローブ
- ⑥ IDCUpDown 3 of 5 (3238)：IDC群の3/5サンプル以上でhealthy群のメジアン値より3倍以上に変動した3238プローブ
- ⑦ GlycoGenes：GSE21422の糖鎖関連遺伝子リスト
- ⑧ forGlycoMaple：GlycoMapleケエリ用の糖鎖関連遺伝子リスト

(b) GSE21422_Exp-Fc_practice.xlsxの構成

- ① GPL570：プローブ情報
- ② GSE21422_samples：サンプル情報
- ③ GSE21422_data：解析用ワークシート

(c) GSE21422_Exp-Fc_formula.xlsxの構成

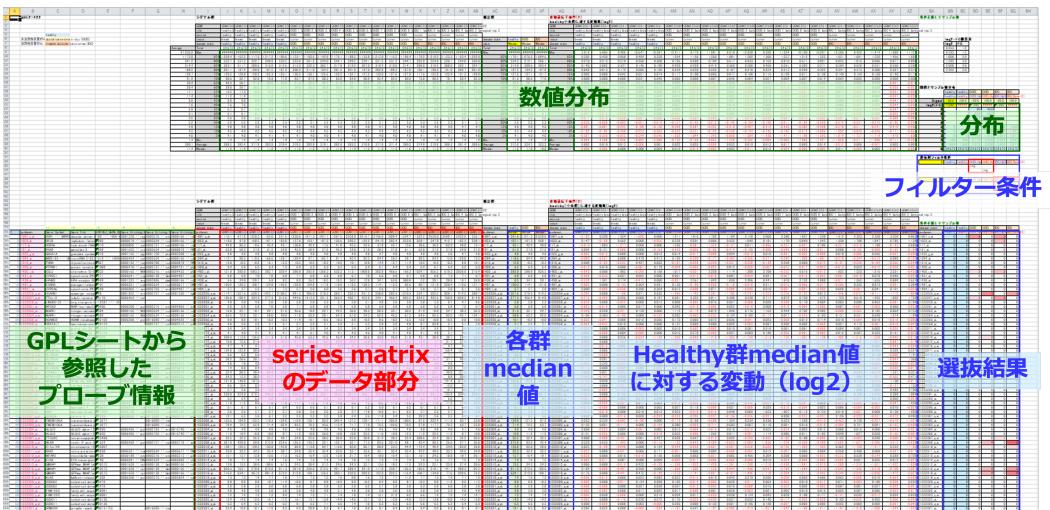
- ① GPL570：プローブ情報
- ② GSE21422_samples：サンプル情報
- ③ GSE21422_data：解析用ワークシート

2. GSE21422_Exp-Fc.xlsxの使い方

(1) 解析用ワークシート（GSE21422_data）（付録図7）

プローブ情報、データマトリクス、数値分布表の構成は「ケース1」と同じ。シグナルマトリクスの右側に、healthy群のメジアン値に対するサンプルごとの変動比をMEDIANで計算した。

シグナル値と変動比に閾値を設定し、群ごとに条件を満たすプローブを数えてn/Nサンプル以上で変動プローブを選抜できるようにした。発現上昇プローブは疾患群の各サンプルの、低下プローブはhealthy群のメジアン値のシグナルでクオリティコントロールを行った。



付録図7 GSE21422(Exp-Fc).xlsxのシートGSE21422_data

(2) 各群で発現が変動したプローブを選抜する

シグナル値とhealthy群のメジアン値に対する変動比の閾値 ([BB27:BG29], [BB6]と[BB7]から参照) を参考し、各群で条件を満たすサンプルをCOUNTIFSで数える ([BB51:BG54725]) (付録図8)。詳細フィルターでDCIS群またはIDC群で指定した数 ([BA34]または[BA38]) 以上に変動したプローブを選抜する。

● DCIS群6/9サンプル以上で上昇したプローブの選抜

サンプル数の閾値 ([BA21]) に「6」を入力する。

[GSE21422_data! \$A\$50:\$BG\$54725]にフィルターをかけ、

データ>詳細設定 から、

- ・ 抽出先 : 選択範囲内
- ・ リスト範囲 : [\$A\$50:\$BT\$54725]
- ・ 検索条件範囲 : [GSE21422_data!\$BB\$21:\$BG\$22]

<OK>を押すと656プローブが選択される。

各群のシグナルメジアン値を参照するセル ([B136:BN36]) を[B1:BN]列の選択されたセルにコピペする。

フィルターをクリアする。

● DCIS群6/9サンプル以上で低下したプローブの選抜

データ>詳細設定 から、

- ・ リスト範囲 : [\$A\$50:\$BT\$54725]
- ・ 検索条件範囲 : [GSE21422_data!\$BB\$24:\$BG\$25]

<OK>を押すと1107プローブが選択される。

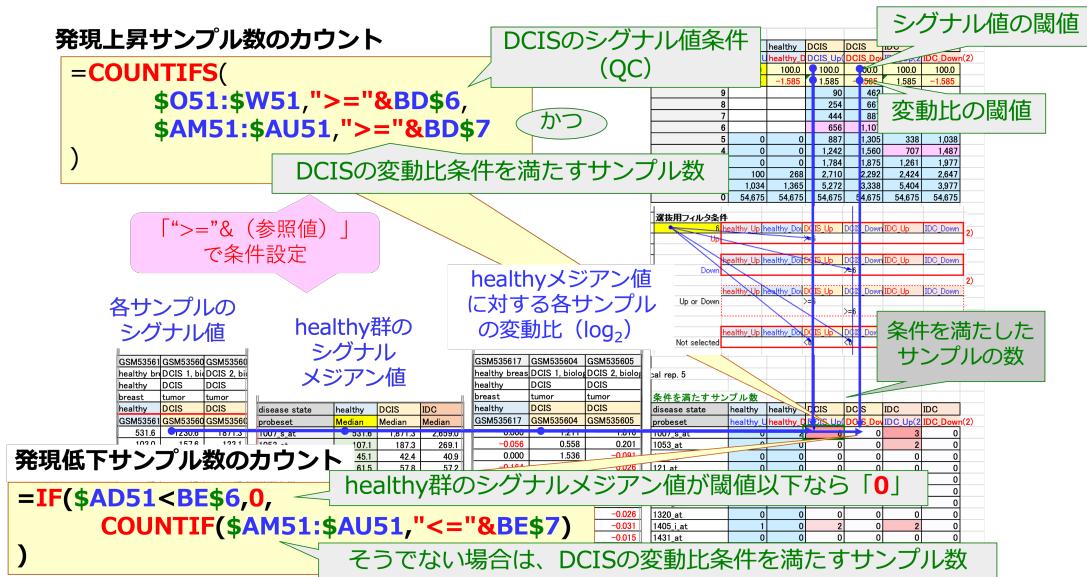
各群のシグナルメジアン値を参照するセル ([B137:BN37]) を[B1:BN]列の選択されたセルにコピペする。

フィルターをクリアする。

● 上記以外のプローブの選抜

データ>詳細設定 から、

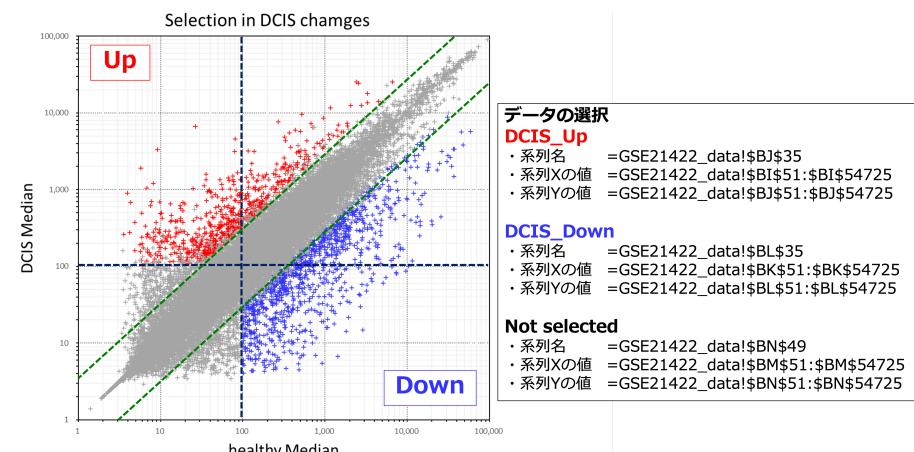
- リスト範囲 : [\$A\$50:\$BT\$54725]
 - 検索条件範囲 : [GSE21422_data!\$BB\$31:\$BG\$32]
- <OK>を押すと52912プローブが選択される。
- 各群のシグナルメジアン値を参照するセル ([BI38:BN38]) を[BI:BN]列の選択されたセルにコピペする。
- フィルターをクリアする。



付録図8 シグナル値と変動比の条件を満たすサンプル数による選択

● 発現変動プローブの可視化

DCIS_UpDown_Scatteredシートには、DCISで発現が上昇したプローブを赤、低下したプローブを青、それ以外を灰色でプロットした分散グラフを作成した（付録図9左）。GSE21422_dataシートでフィルターをかけると、絞り込んだプローブだけが表示される。



付録図9 Scattered plot。右は各系列でプロットしているデータ範囲（「データの選択」で表示）。

■ Excelファイルの作り方

1. NCBI GEOからデータファイルをダウンロードする。

NCBI GEO (<https://www.ncbi.nlm.nih.gov/geo/>) でのデータ検索は、PubMedと同様、キーワードなどで検索する。生物種やデータタイプ、実験の種類等で絞り込むこともできる。GSE番号がわかっている場合は、GSE番号で直接検索する（付録図10と11）。

① 検索クエリ窓
Keyword or GEO Accession Search

② 検索結果
diabetes human kidney Search

③ クリック！
The are 1772 results for "diabetes human kidney" in the GEO DataSets Database.

④ 検索結果リスト
GEO DataSets diabetes human kidney Create alert Advanced

[データタイプ]
GSE : 実験セット (シリーズ)
GPL : 測定プラットフォーム
GSM: サンプルデータ
GDS: NCBIがまとめた解析単位

付録図10 NCBI GEOでの検索 (1)

⑤ Seriesをクリック
diabetes human kidney

⑥ 実験セット (GSE) のリスト
Items: 1 to 2 of 1772

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
GEO DataSets diabetes human kidney Create alert Advanced

データタイプや実験タイプによる絞り込み検索

生物種での絞り込み

付録図11 NCBI GEOでの検索 (2)

以下、GSE7032の例を示す。

GSE7032: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7032>
 から、① GPL (GPL81-57556.txt) と② series matrix (GSE7032_series_matrix.txt) をダウンロードする。Affymetrix GeneChipを生データ (CELファイル) の数値化からやり直す場合は、③のsupplementary file (GSE7032_RAW.tar) も必要になる (付録図12)。

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

① プラットフォームの情報 (GPL81)
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL81>

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)

付録図12 GSE7032

series matrixにはサンプル情報の上部には実験セットの情報と各サンプルの実験条件や数値化条件などが記載されており、その下にシグナルマトリクスがある。これらを加工して解析用ワークシートを作る (付録図13)。

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

シグナルマトリクス
 ここではMAS5で数値化されているが、
 このケースではフラグ情報はない

付録図13 Series matrixファイル

series matrixにプローブIDが示されているが、そのプローブの遺伝子情報はGPLファイルにあるので、シグナルマトリクスとGPLの情報を合わせる。GPLによって異なるが、GPL81では遺伝子IDや遺伝子名のほか、遺伝子機能や遺伝子産物の細胞内局在を示すGene Ontology (GO) の情報もあった（[付録図14](#)）。

GPL81の情報

A	B	C	D	E	F	G	H	I	J	K	L
1	GB	_Attribute	Probe_Set_ID								
2	3	GB_ACC	=	GmBank	Accession Number						
4	SPOT_ID	=	identifies	controls							
5	Species	Scientific Name	=	the genus and species of the organism represented by the probe set.							
6	Annotation Date	=	The date that the annotations for this probe array were last updated. It will generally be earlier than the date the probe set was created.								
7	Sequence Source	=	The database from which the sequence used to design this probe set was taken.								
8	#Target Description	=									
9	#Representative Public ID	=	ID: The accession number of a representative sequence. Note that for consensus-based probe sets, this is the same as the Gene ID.								
10	#Gene Title	=	Title of Gene represented by the probe set.								
11	#Gene Symbol	=	A common symbol which is available from UniGene.								
12	ENTREZ_GENE_ID	=	Entrez Gene Database ID								
13	#RefSeq Transcript ID	=	References to multiple sequences in RefSeq. The field contains the ID and Description for each entry.								
14	Gene Ontology Biological Process	=	Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation has one GO term.								
15	Gene Ontology Cellular Component	=	Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation has one GO term.								
16	Gene Ontology Molecular Function	=	Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation has one GO term.								
17	GB_SPOT_ID	=	SPOT_ID	Species	S Affected	Sample Type	Sample ID	Sample Set ID	Sample Set Name	Sample Set Entrez ID	
18	100001_1st	=	M12328	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	M12328_C09_wtCell	1
19	100002_1st	=	X70398	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X70398_1st	1
20	100003_1st	=	D92616	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	D92616_1	1
21	100004_1st	=	AW120390	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AW120390_interprts_1	7
22	100005_1st	=	Z12654	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	Z12654_1	1
23	100006_1st	=	D21253	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	D21253_1	1
24	100007_1st	=	AB337573	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB337573_interprts_1	27
25	100008_1st	=	X94127	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X94127_SRY (sex: Sox2)	2
26	100009_1st	=	U83404	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U83404_kruppel-lk_KR9	1
27	100010_1st	=	AB91583	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB91583_kruppel-lk_KR9	1
28	100011_1st	=	AB91584	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB91584_kruppel-lk_KR9	1
29	100012_1st	=	AW121732	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AW121732_interprts_B05	7
30	100013_1st	=	AB45038	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB45038_toussled-Tk2	2
31	100014_1st	=	X67677	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X67677_Yamauchi_Yes1	2
32	100015_1st	=	Z12654	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	Z12654_mitochondrial_mimp1	1
33	100016_1st	=	Z12657	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	Z12657_mitochondrial_mimp1	1
34	100017_1st	=	X71237	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X71237_metal_loph	1
35	100018_1st	=	AB45898	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB45898_verسان_vcan	1
36	100020_1st	=	JO4038	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	JO4038_solute_carrier2_2	2
37	100021_1st	=	M17640	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	M17640_cholinergic_ChRNA1	1
38	100022_1st	=	Z12656	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	Z12656_hypothal_MyH2	1
39	100023_1st	=	X70472	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X70472_hypothal_MyH2	1
40	100024_1st	=	AB41945	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB41945_shroom_faShroom3	1
41	100025_1st	=	U24443	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U24443_branchied_cBcl1	1
42	100027_1st	=	AB81187	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB81187_peroxisom_Pex14	5
43	100028_1st	=	AB20867	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB20867_peroxisom_Pex14	5
44	100029_1st	=	AB20867	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB20867_peroxisom_Pex14	5
45	100030_1st	=	D44444	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	D44444_uridine_pkLsp1	2
46	100032_1st	=	X60138	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X60138_trans-actin-Spi1	2
47	100033_1st	=	X70143	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	X70143_tmstt5_homo_Mash2	1
48	100034_1st	=	U54705	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U54705_serine_Leucine	1
49	100035_1st	=	L10182	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	L10182_nucleotide_binding	1
50	100037_1st	=	M12325	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	M12325_DAD_As_Dx18	1
51	100038_1st	=	AW12250	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AW12250_2_Cryp2	2
52	100040_1st	=	AD43081	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AD43081_mitochondr_Mpl17	27
53	100041_1st	=	Z124138	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	Z124138_cytochrome_c_beta	1
54	100042_1st	=	AB337573	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	AB337573_prosurp_Pase1	1
55	100043_1st	=	Z124138	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	Z124138_cytochrome_c_beta	1
56	100044_1st	=	U19582	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U19582_cldlin_11_Odin11	1
57	100045_1st	=	J04527	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	J04527_myelene_Mhd	1
58	100046_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
59	100047_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
60	100048_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
61	100049_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
62	100050_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
63	100051_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
64	100052_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
65	100053_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
66	100054_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
67	100055_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
68	100056_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
69	100057_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
70	100058_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
71	100059_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
72	100060_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
73	100061_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
74	100062_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
75	100063_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
76	100064_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
77	100065_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
78	100066_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
79	100067_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
80	100068_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
81	100069_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
82	100070_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
83	100071_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
84	100072_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
85	100073_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
86	100074_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
87	100075_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
88	100076_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
89	100077_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
90	100078_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
91	100079_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
92	100080_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
93	100081_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
94	100082_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
95	100083_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
96	100084_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
97	100085_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
98	100086_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
99	100087_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
100	100088_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
101	100089_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
102	100090_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
103	100091_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
104	100092_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
105	100093_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
106	100094_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
107	100095_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
108	100096_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
109	100097_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
110	100098_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
111	100099_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
112	100100_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
113	100101_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
114	100102_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
115	100103_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
116	100104_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
117	100105_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
118	100106_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
119	100107_1st	=	U00018	Mus musculus	6	Oct-14	Consensus	GenBank	Cluster	U00018_nucleotide_binding	1
120	100108_1st	=	U00018	Mus musculus	6						

付録図14 GPLファイル

2. 解析用エクセルファイルを作る (GSE7032)

(1) 準備

GSE7032_series_matrix.txtとGPL81-57556.txtを1つのエクセルに読み込みそれぞれワークシートにして「.xlsx」形式で別名保存する。

(2) サンプル情報の整理

新たにシートを作り、GSE7032_series matrixシートから縦横を入れ替えて必要なサンプル情報のテーブルを作る（付録図15）。解析目的に合わせて並べ替える場合は、サンプル情報とシグナルマトリクスの順番に齟齬がないように注意する。加工前のseries matrix上で列単位で並べ替えた後に、サンプル情報とシグナルマトリクスのシートを作ると取り違えない。

(3) 解析用ワークシート (GSE7032 data) の作成

新しくシートを作り、GSE7032_series matrixのシグナルマトリクスをコピペする。その左側に空の列を挿入し、GPL81シート（[付録図16](#)）からプローブ情報をVLOOKUPで参照する（C～

G列)。series matrixとGPLファイルのプローブの順番は同じとは限らないので、必ずプローブID (B列) をキーにしてVLOOKUPで参照する。VLOOKUPで参照する列番号は計算式に直接数字を入力するのではなく、別のセル ([C40:G40]) から参照すると良い（付録図18）。GPL81シートのカラムヘッダーの上 ([A19:P19]) に数字を振り、ヘッダーと一緒にGSE7032_dataシートにコピペしてくると間違えない。VLOOKUPの参照先の英語と数字に適切に絶対参照記号「\$」を付けると、左上セルに計算式を入力後、そのセルをコピペすれば簡単に表が完成する。正しく参照できたら「コピー>値をペースト」で計算式を消す。

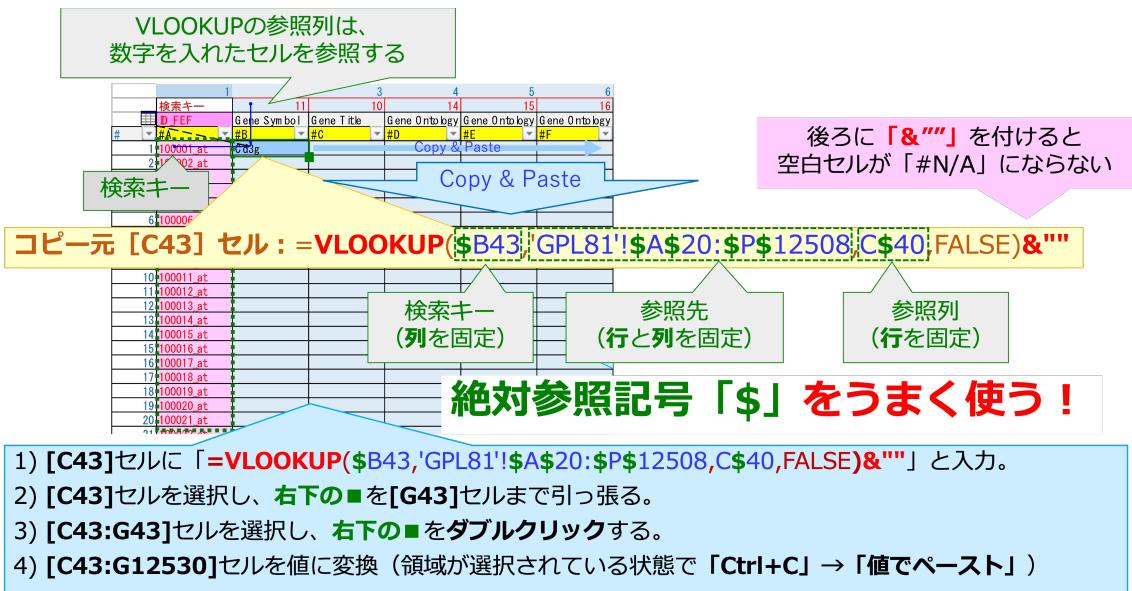
① series matrixのGSM情報を縦横を入れ替えて表にし、
解析に必要な情報を整理しておく。

② 左端をGSM番号にし、ここを起点(1)として番号を振っておくと、
VLOOKUPの列順として参照できる。

付録図15 サンプル情報シート

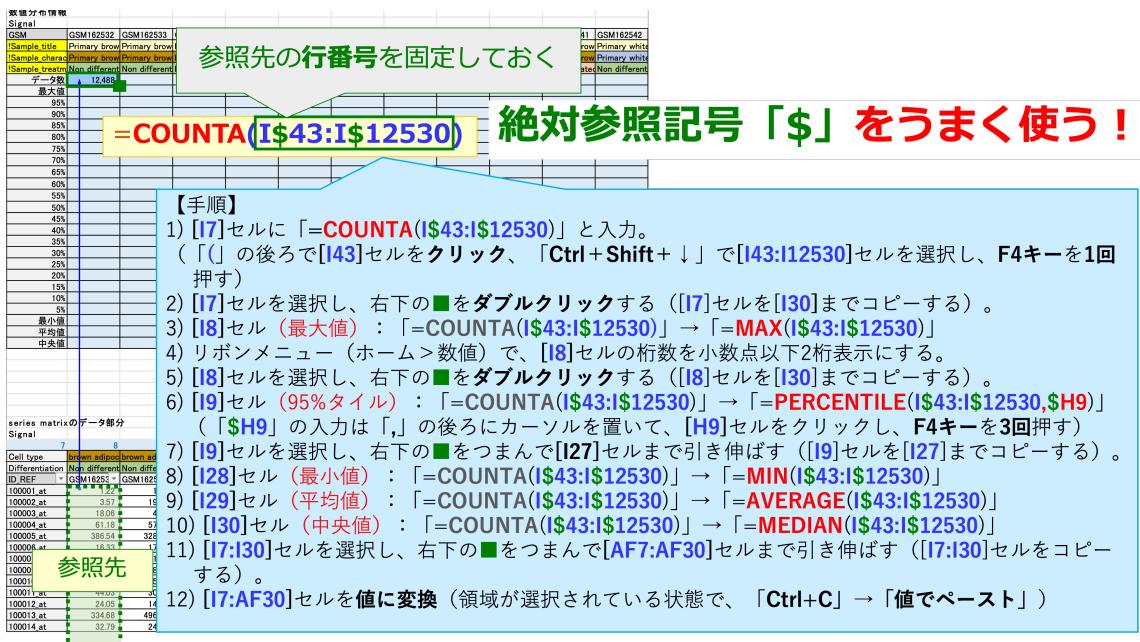
① データマトリクスに必要な情報をLOOKUPするため、19行目に
1行挿入し、IDを「1」にして番号を振った。
② 解析用シートから必要な情報を参照する際、ヘッダーと一緒に
解析用シートにコピーすれば、LOOKUPから参照できるので、
いちいち列番号を数えなくて済む。

付録図16 GPLシート



付録図17 遺伝子情報テーブル

シグナルマトリクスの上に空白行を挿入し、COUNTA, MAX, PERCENTILE, MIN, AVERAGE, MEDIANを使って数値分布表を作る（3～30行目）（付録図18）。信頼データ区間を判断する上で数値分布の把握は必須。ここでも絶対参照記号「\$」を使うとコピペで表が完成する。正しく表ができたら「コピー>値をペースト」で計算式を消す。



付録図18 数値分布表の作成

シグナルマトリクスの右側に変動プローブを選択するテーブルを作成する。ここでは、群平均値 ([AH:AK]) をAVERAGEで計算し、変動比とP値 ([AM:AN], [BB:BC]) をLOGとTTESTで計算した。

この上に詳細フィルターの検索条件範囲参照セル ([AM13:AN14], [AM16:AN17] , [AL20:AN22]と, [BB13:BC14], [BB16:BC17], [BA20:BC22]と, [BQ13:BW15]) を用意した。

この右側にはそれぞれ選択したプローブのシグナル平均値、FC値・-logP値などを参照・計算するテンプレート ([AO:AZ]と[BD:BO]と[BQ:BW]) を付けた。

選択条件の閾値は計算式に直接数値を入力せず、閾値を入れた表 ([AN32:AN33]と[BC32:BC33]) から参照すれば、条件を変えた変動遺伝子リストを簡単に作れる。また、選択した変動プローブのシグナル値を参照後、プローブ数をCOUNTで数えた表 ([AM34:AN37]と[BB34:BC37]) を作っておくと、選択条件を検討する際、参考になる。

3. 解析用エクセルファイルを作る (GSE21422)

(1) 解析用ワークシート (GSE21422_data) の作成

シグナルマトリクス ([I45:AB54725]) , プローブ情報 ([B49:H54725]) の付加、シグナルの数値分布表 ([I3:AB31]) の作成までは「ケース1」とほぼ同様なので省略する。

シグナルマトリクスの右側に変動遺伝子を選択するテーブルを作成する。ここでは、各群のメジアン値を計算し、シグナル値とhealthy群のメジアン値に対するサンプルごとの変動比を計算するテーブル ([AG45:AZ54725]) を、さらに、群ごとに条件を満たすサンプル数をCOUNTIFSで数えるテーブル ([BA49:BG54725]) を作成した。発現上昇プローブは個別サンプルのシグナル値で、発現低下プローブはhealthy群のメジアン値でクオリティコントロールの閾値を設定している点に注意する。

この上には、詳細フィルターの検索条件範囲参照セル ([BB21:BG22], [BB24:BG25], [BB31:BG32]など) を用意した。ここでも閾値は計算式に数値を直接入力するのではなく、閾値を入れたセル ([BA21]) から参照すれば、条件の違う変動プローブリストを簡単に作れる。条件設定の参考にするため、COUNTを使って条件を満たすプローブ数の集計テーブル ([BA4:BG17]) を作成した。

■ どんな場面（研究/実験 /解析 /操作）で利用すると便利か

● RNA-seqデータの解析

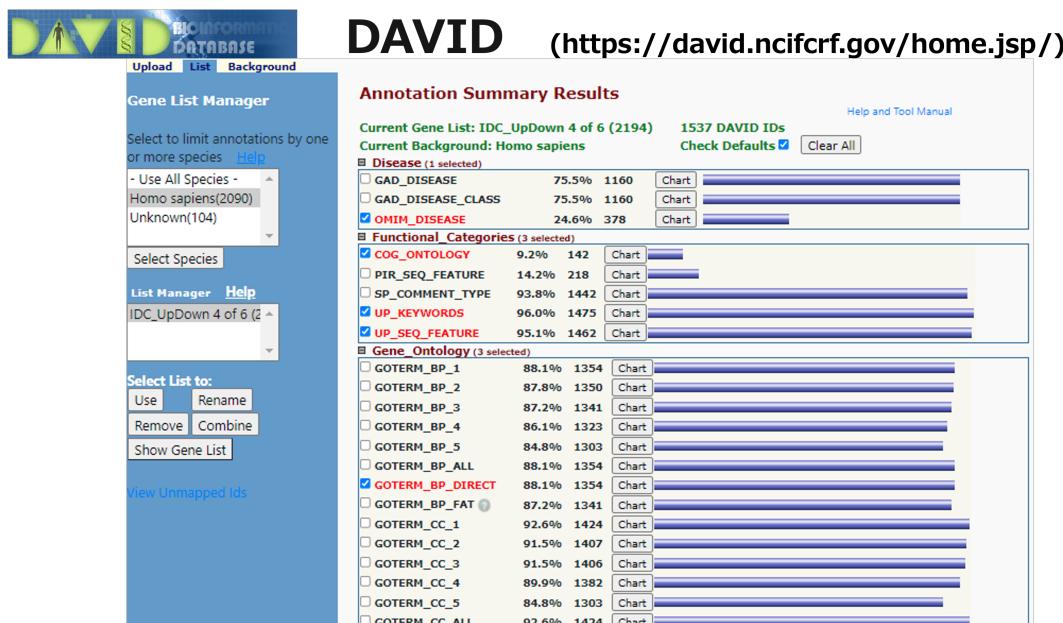
ここではDNAマイクロアレイデータの解析例を紹介したが、RNA-seqデータでも同様に解析できる。GEOでは計算された定量値が公開されていないデータセットも多いが、Digital Expression Explorer 2 (DEE2) (<https://dee2.io/>) ではGEOの多くのRNA-seqデータを定量化し公開している。

● 変動遺伝子リストのDAVIDでの利用

DAVID (<https://david.ncifcrf.gov/>) では、遺伝子リストをクエリに、それらの遺伝子リストがどのような機能に関係しているかを調べることができる。

Start Analysisから、左ウインドウの「Upload」タブに変動遺伝子のプローブリスト（例：「ケース2」のGSE21422_Exp-Fc.xlsxのUpDown 6 of 9 (1763)シートの赤枠部分 [B51:1813]）をコピペして<Gene List>として<Submit List>し、「List」タブで<Homo sapiens (2087)>を、「Background」タブで<Human Genome U133 Plus2>を指定して、右ウインドウのAnalysis Wizard の<Functional Annotation Tool>をクリックすると解析結果が得られる（付録図19）。

DAVIDでは変動比やシグナル値の閾値などの条件で結果が影響を受けるため、選択条件を変えた変動遺伝子リストをいくつか作り、結果を比較することをお勧めする。



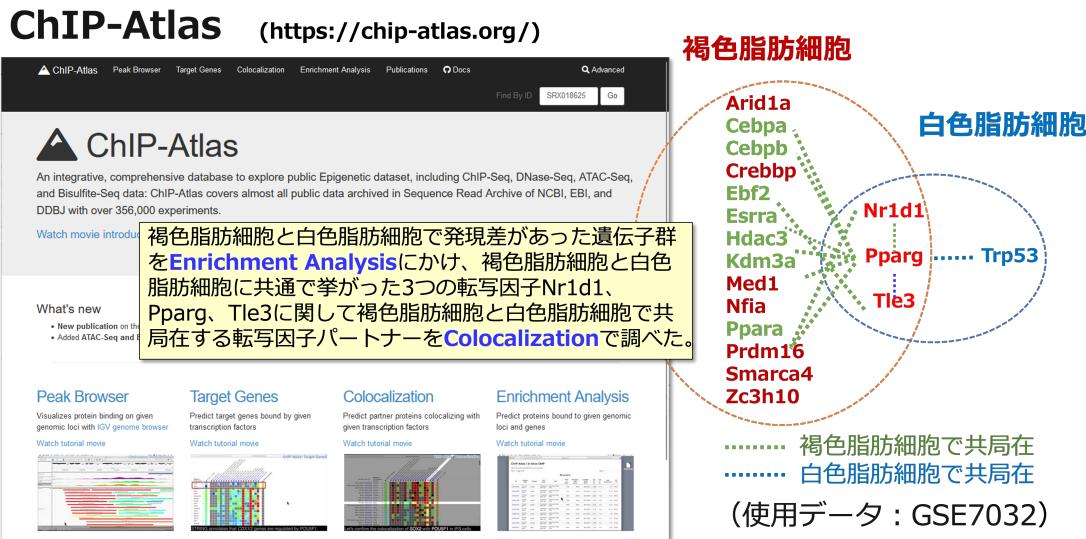
付録図19 DAVID

● 変動遺伝子リストのChIP-Atlasでの利用

ChIP-Atlas (<https://chip-atlas.org/>) はChIP-seqやATAC-seqなどのデータを集めたデータベース。Enrichment Analysisでは、遺伝子リストをクエリにし、細胞や組織を指定して、それらの発現制御に関する転写因子やエピゲノム状態を予測できる。Enrichment Analysisを選択し、「1. Experiment type」, 「2. Cell type Class」, 「3. Threshold for Significance」を設定後、「4. Enter dataset A」で<Gene list (Gene symbols)>を選択して

変動遺伝子のプローブリスト（例：「ケース 1」のGSE7032_Fc-P.xlsのDiff BrvsWh_FC1.0-P0.05Upシートの赤枠部分[C43:C406]）をコピペし、「5. Enter dataset B」で<Refseq coding genes (excluding dataset A)>を指定して、「6. Analysis descriptionで<submit>をクリックすると解析が始まる。指定されたURLにアクセスすると結果が得られる（付録図20）。

ChIP-Atlasでも、変動比やシグナル値の閾値で結果が影響を受けるため、選択条件を変えた変動遺伝子リストをいくつか作り、結果を比較することをお勧めする。

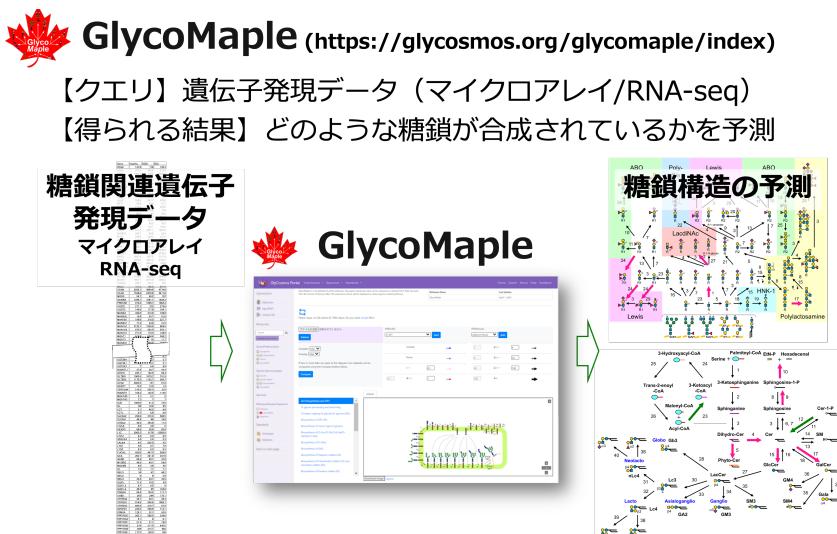


付録図20 ChIP-Atlas

● 変動遺伝子リストのGlycoMapleでの利用

糖鎖統合データベースGlyCosmosのGlycoMaple (<https://glycosmos.org/glycomaple>) では、糖鎖関連遺伝子の発見データからその細胞の糖鎖構造の変化を予測できる（付録図21）。

「ケース 2」のGlycoMaple用クエリサンプルをGSE21422_Exp-Fc.xlsxのforGlycoMapleシートを「.csv」形式で別名保存し、クエリとして<Submit>する。



付録図21 GlycoMaple

■ Excelを便利に使いこなす工夫

● Excelで大きなマトリクスファイルを扱うときに設定変更しておくとよい機能

自動保存(自動回復用データの保存)をしない

意図しないタイミングで自動保存が始まり作業が中断されることを避けるため、チェックを外す。ただしこまめに手動で保存する。作業のまとまりごとに、バージョンを付けて別名保存すると、ミスした際に少し前の状態からやり直すことができる。

「セルを直接編集する/セル内で編集する」を解除する

関数入力の際、参照先のセルのアドレスをキーボードから入力するのではなく、マウスで参照先をクリックすると間違いが減る。そのときには「セルの直接編集」を解除しておかないと、入力ウインドウが邪魔になり、すぐ右側のセルをクリックできなくなる。

オートコンプリートを使用しない

意図せず誤った文字列をオートコンプリートで入力してしまわないように解除する。

パーセンテージを自動入力しない

クイックアクセスバーを設定変更

「参照元のトレース」および「参照元トレース矢印の削除」を追加

「参照先のトレース」および「参照先トレース矢印の削除」を追加

入力した計算式が正しいセルを参照しているかを確認する際に便利。

Windowsではファイル>オプション>クイックバーアクセスから設定変更できる。

macでは環境設定>リボンとツールバー>クイックアクセスツールバー>コマンドの選択で

「すべてのコマンド」を選択から追加できる。

● エクセル操作のtips

便利なショートカットキー

- ・ 「Ctrl+C」（コピー）と「Ctrl+V」（ペースト）
- ・ 「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）
- ・ 「\$」（絶対参照記号）とF4キー（A1→\$A\$1→A\$1→\$A1→A1）
- ・ 「Ctrl+矢印」（連続セル間でのジャンプ移動）と「Ctrl+Shift+矢印」（連続セルの選択）
- ・ 「*」（任意の文字列）、「?」（任意の一文字）
- ・ 「"」（特殊文字のエスケープ、例えば、「"」（ダブルコーテーション）を""で囲って文字列として扱う際は「""""」となる（2番目の「"」がエスケープ記号）
- ・ 関数の「引数」は、計算式に直接数字や文字列を入力するのではなく、別のセルに入力した値を参照するようにすると扱いやすい。

よく使うエクセル関数

- ・ 参照関数：VLOOKUP, HLOOKUP
- ・ 論理関数：IF, AND, OR, NOT
- ・ 文字列関数：TRIM, LEN, LEFT, RIGHT, SUBSTITUTE, &
- ・ 数値関数：ROUND, ROUNDDOWN, ROUNDUP
- ・ 数学関数：SUM, PRODUCT, QUOTIENT, MOD, ABS
- ・ 統計関数：MAX, MIN, AVERAGE, MEDIAN, STDEV, SMALL, LARGE, RANK, PERCENTILE, QUARTILE, TTEST, PEARSON, CORREL
- ・ 数を数える関数：COUNT, COUNTA, COUNTIF, COUNTIFS, SUMIF
- ・ データベース関数：DAVERAGE, DCOUNT, DCOUNTA, DMAX, DMIN,

DPRODUCT, DSTDEV, DSTDEVP, DSUM, DVAR, DVARP, DGET

- ・その他：IFERROR

詳細フィルター

- ・ 詳細フィルターのリスト範囲のヘッダーは重複を避け、検索条件範囲で一義的に指定できるようとする。
- ・ 詳細フィルターの検索条件範囲は、等号や不等号を「'''」でくくり、文字列連結関数「&」でつないで閾値を参照すると、簡単に条件を変えてフィルターをかけ直せる。
- ・ 詳細フィルターの検索条件範囲は、行方向がAND条件、列方向がOR条件。
- ・ エクセルの詳細フィルターの検索条件範囲や検索範囲にエリア名（「DATA」，「criteria_1」，「criteria_2」・・・など）を付けておくと、フィルターをかけなおす際の手間が省ける。エリア名の編集・削除は「数式>名前ボックス」で行う。

大きなエクセルファイルの動作を軽くする工夫

- ・ 操作のたびに再計算になると時間がかかるため、固定値で扱う数値は計算式を消して値に変換しておく。検証や再利用のためには、計算式を残したファイルを別名ファイルで保存しておくと良い。
- ・ 条件を変えて詳細フィルターをかけ直す際、詳細フィルター設定画面で検索条件範囲を変えてできるが、再検索の結果表示に時間がかかるため、一旦フィルターをクリアして詳細フィルターをかけ直した方が早い。
- ・ 表示フォントをMSPゴシックやMSゴシックにすると表示が速くなる。

● Affymetrix GeneChipデータをCELファイルの数値化から行うときの注意！

Affymetrix GeneChipのCELファイルの数値化は、Thermo Fisher社のTranscriptome Analysis Console (TAC) Software等を使って行う。異なる方法で数値化したデータは直接比較できないため、複数のGSEを統合して解析する場合は、CELファイルを取ってきて、同じ方法で数値化し直す必要がある。

詳細はThermo Fisherのサイトを参照されたい（<https://www.thermofisher.com/jp/ja/home/technical-resources/technical-reference-library/microarray-analysis-support-center.html>）。