

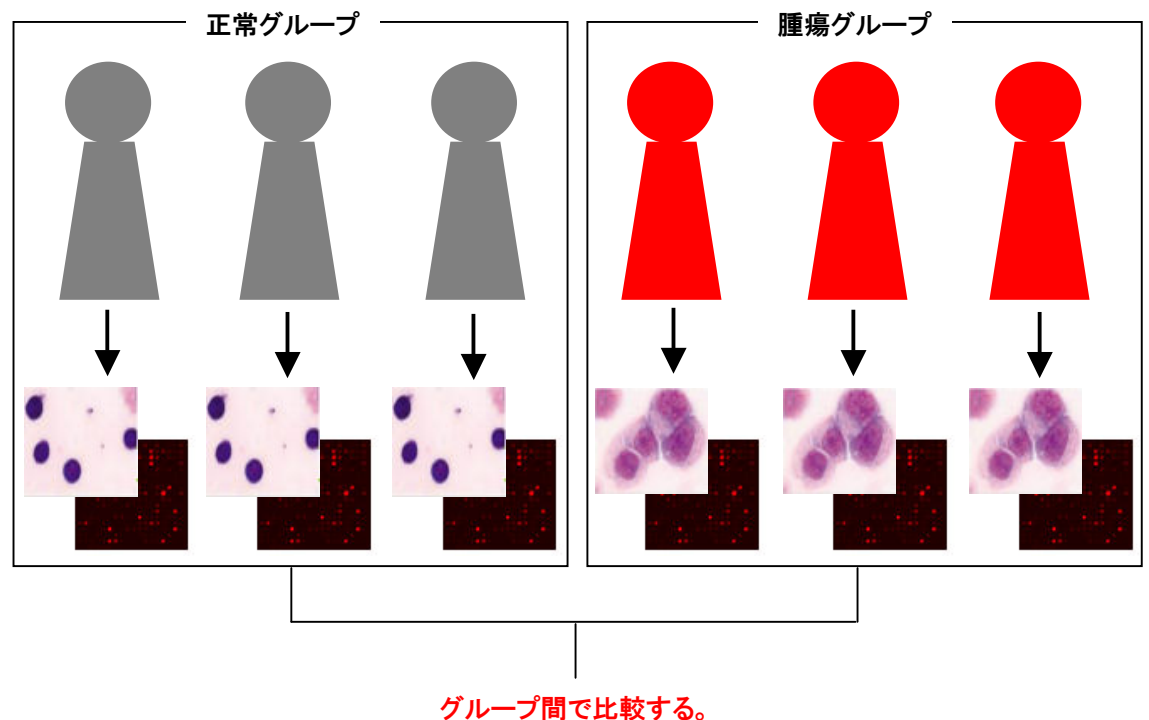
Microarray Data Analysis Tool データ解析の進め方 Vol.5

－ 2グループ間の比較解析 －

疾患グループ・正常グループなどのように、繰り返し実験を含む多数のサンプルを、2つのグループに分け、統計検定も使用して、グループ間の発現量を比較する場合の解析例を、Signal比較解析を使用して説明します。

解析テーマは

「正常グループ-腫瘍グループ間の、遺伝子発現プロファイルの探索」です。



上記のテーマについて解析を行うには、まず各サンプルを用いて、単色のマイクロアレイ実験を行い、その後、本ソフトウェアのSignal比較解析機能を用いて、下記のような処理を行う必要があります。

- ①全実験データの正規化(ノーマライズ)
- ②各サンプルデータのグループ分け
- ③統計検定機能を用いて、グループ間の発現量の差の有意確率を計算
- ④グループ間で発現が変動している遺伝子を抽出

Vol.4で説明したように、Signal比較解析では、全実験データの正規化およびグループ分けを行うことができます。そしてグループ情報に従い、グループ間発現量差の有意確率を計算し、それらの値を用いて、有意に発現が変動している遺伝子を抽出します。

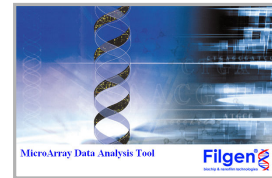
1. ソフトウェアの起動と実験データの読み込み

1-1. ソフトウェアの準備

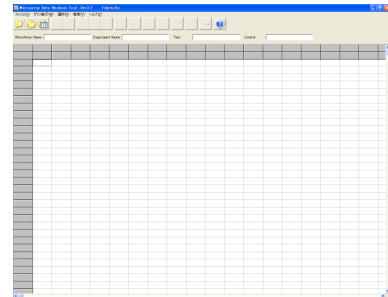
CDに入っている「Microarray Data Analysis Tool」のフォルダを、自分のコンピュータにコピーしてください。

1-2. ソフトウェアの起動


コピーしたフォルダを開き、「MAtool.exe」のアイコンをダブルクリックしてください。右図のタイトル画面が表示され、ソフトウェアが起動します。

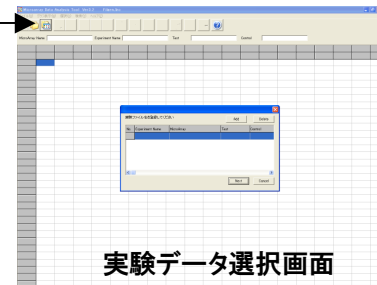


ソフトウェアが起動後は右図のメイン画面が表示されます。

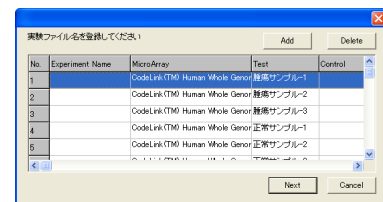


1-3. 実験データ読み込み

 アイコンをクリックすると、右の画面が表示されます。「Add」ボタンをクリックし、対象となる実験データの選択・読み込みを行います。なおSignal比較解析用の実験データファイルは、1サンプル分の実験データしかもっていません。そのためファイル名も「SampleA vs SampleB」ではなく、「SampleA」、「SampleB」のように1サンプルずつに分かれており、サンプル名の後ろに「Signal比較解析用」と入力されています。



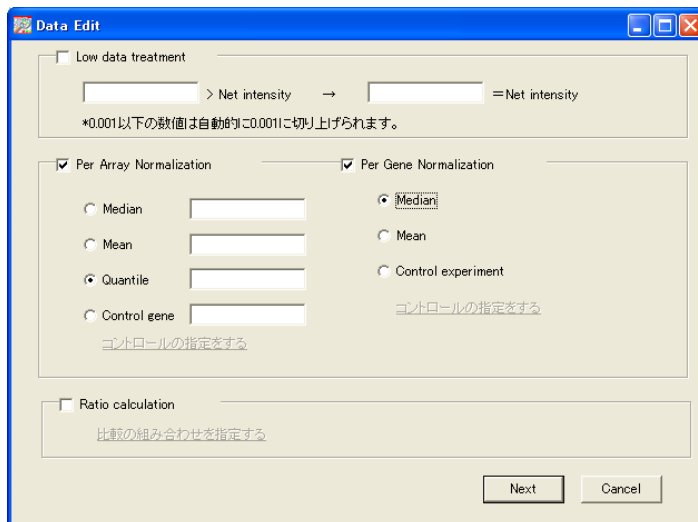
実験データを読み込むと、右図のように選択したファイルの使用マイクロアレイ名、サンプル名等が表示されます。ここでは「腫瘍サンプル1~3」と、「正常サンプル1~3」の、合計6つの実験データを読み込みます。表示の確認が終わったら、「Next」ボタンをクリックして次へ進みます。



2. 実験データの正規化とグループ分け

2-1. 正規化手法の指定

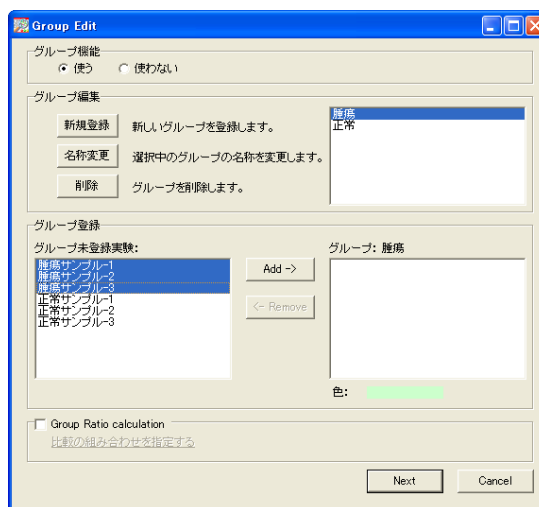
「Next」ボタンをクリックすると、「Data Edit」(下図)の画面が現れます。この画面では正規化の手法や、比較の組み合わせの選択を行います。各項目の詳細については、本ソフトウェアのユーザーマニュアル、またはVol.4を参照してください。



ここでは「Per Array Normalization」に「Quantile」を選択し、「Per Gene Normalization」に「Median」を選択しました。本解析のように、クラスタリングや発現量の統計検定を行う場合は、必ずPer Gene Normalizationの設定を行ってください。また本解析のように、サンプル同士の直接の比較を行わず、グループ間で比較を行う場合は、「Ratio calculation」の設定は必要ありません。

2-2. グループ化

Data Edit画面から進むと、「Group Edit」画面が現れます。ここで実験データをグループ分けすることで、グループ単位の検索や統計検定が実行可能になります。

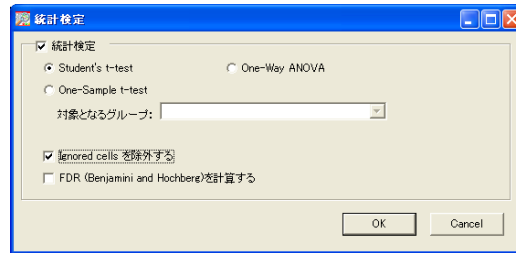
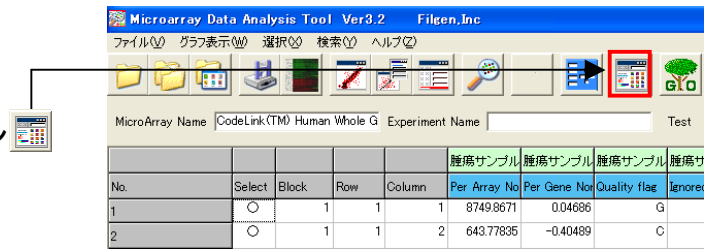


ここでは2つのグループを作成し、一方に「腫瘍サンプル-1~3」を、もう一方に「正常サンプル-1~3」を設定しました。

3. 統計検定の実行

3-1. 検定手法の選択

メイン画面の統計検定アイコンをクリックしてください。



すると上図のように、実行する統計検定手法とオプション項目の選択画面が表示されます。

使用する検定手法は、解析を行うデータのプラットフォームやグループ数によって、以下のよう使い分けます。

- Student's t-test : 2グループ間比較
- One-Sample t-test : 1グループ比較(2色法の場合)
- One-Way ANOVA : 3グループ以上の比較

また2つのオプション項目は、それぞれのチェックボックスにチェックを入れることによって、以下の機能が使用できるようになります。

- Ignored cellsを除外する : P-value計算時に、解析除外スポット(Ignored cellsが1のスポット)のデータを使用しない。
- FDR (Benjamini & Hochberg)を計算する : P-valueと同時に、FDR (False Discovery Rate)も計算する。


なお、本解析では2グループの比較解析になるので、使用する検定手法として「Student's t-test」を、オプション項目として「Ignored cellsを除外する」を選択しました。

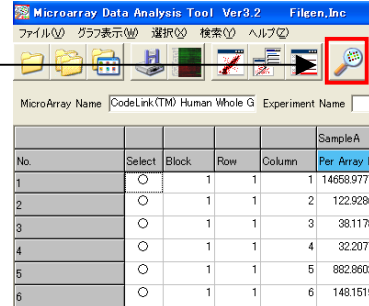
上記のように設定後、「OK」ボタンをクリックすると、メイン画面の「P-value」列に、各遺伝子のグループ間発現量の差を表す、P-valueが表示されます(下図)。

No.	正常サンプル Per Array No	正常サンプル Per Gene No	正常サンプル Quality flag	正常サンプル Ignored cells	腫瘍 vs 正常 ratio	腫瘍 vs 正常 log2ratio	Student's t-test P-value	Student's t-test FDR
1	8572.50823	0.01731	G	0	1.07964	0.11055	0.242953311756289	
2	916.77918	0.10512	G	0	1.20128	0.26457	0.202342912596338	
3	131.35457	0.25545	G	0	1.25724	0.33026	0.421228263974263	
4	140.12423	0.4119	G	0	1.80163	0.8493	0.332858890768491	
5	96.21778	0.20617	L	0	1.77299	0.82619	0.21260760352305	
6	157.45542	0.41573	G	0	1.49626	0.58136	0.258609568543353	
7	169.07073	0.50231	G	0	1.29721	0.37541	0.495648652537351	
8	8943.11052	0.09587	G	0	0.9643	-0.05244	0.597084885732903	
9	193.04647	-0.07338	G	0	1.57124	0.6519	0.145579454953325	
10	434.67242	-0.03688	G	0	1.30659	0.38581	0.104383004885151	

4. 発現変動遺伝子の抽出

4-1. Filter Options検索画面の表示

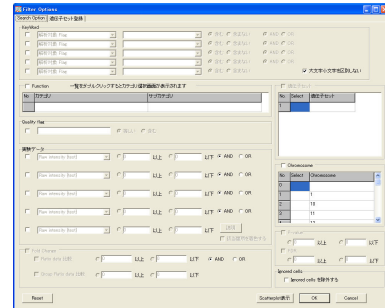
メイン画面のFilter Optionsアイコンをクリックし、Filter Options検索画面を開きます。



No.	Select	Block	Row	Column	Sample A
1	<input type="radio"/>	1	1	1	14858.9777
2	<input type="radio"/>	1	1	2	122.9206
3	<input type="radio"/>	1	1	3	38.1178
4	<input type="radio"/>	1	1	4	32.2077
5	<input type="radio"/>	1	1	5	882.8602
6	<input type="radio"/>	1	1	6	148.1519

Filter Options検索画面では各種検索条件でデータの抽出を行うことができます。

keyword
機能カテゴリー
実験データ
染色体番号
遺伝子セット



4-2. 検索条件を設定

ここでは下記を目的を達するための条件設定を行います。

- ①コントロール遺伝子などを除去し、解析対象遺伝子のみにする。
- ②グループ間で発現差のあるデータを抽出する。
- ③ノイズレベルの信頼性の低いデータを除外する。(CutOff)

条件設定の方法

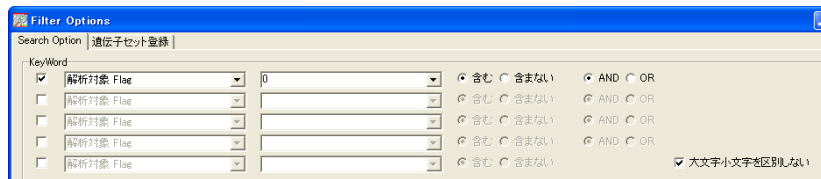
- ①コントロール遺伝子などを除去し、解析対象遺伝子のみにする。

ここでは、keyword検索を使用します。事前にアレイに搭載されているプローブは「0、2」という数字で識別をしています。

「0」は解析対象遺伝子

「2」は解析対象外遺伝子(Negative Control, Blank Spot, マーカーコントロールなど)

まずは、左のチェックボックスをクリックし、該当項目で「解析対象 Flag」と、「0」を選択してください。これで、解析対象遺伝子のみを抽出する条件設定ができました。



②グループ間で発現差のあるデータを抽出する。

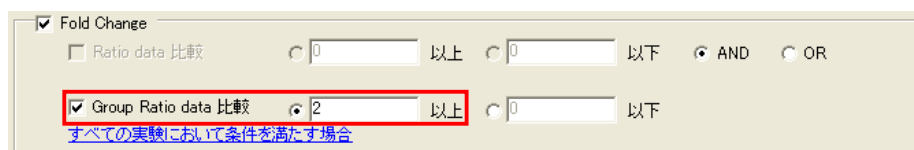
グループ間で発現差のあるデータを抽出する場合、グループ間のFold Changeと、統計検定のP-value(or FDR)で抽出する場合の2種類があり、この2つは同時に使用することができます。以下、それぞれの設定方法を紹介します。

②-1 Fold Changeで検索する。

4ページで指定したグループ間の発現比の検索には、「Fold Change」検索を使用します。「Ratio data 比較」ではサンプル間の発現比、「Group Ratio data 比較」では実験グループ間の発現比の検索条件の指定を行います。本解析ではサンプル間発現比の計算を行っていませんので、「Ratio data 比較」の項目は指定できないようになっています。

左のチェックボックスをクリックし、検索条件を入力します。(下図は「Group Ratio data 比較」を2以上にした例です。)

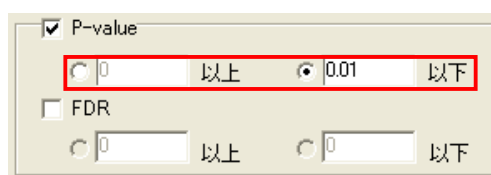
続いて、「複数データの検索条件を指定してください」の項目を設定してください。本解析ではGroup Ratioとして、「腫瘍グループ vs 正常グループ」だけしか計算していませんが、このFold Change検索では、Ratioデータが1つだけしかなくても、複数データの検索条件の指定を求められるようになっているので、必ず指定を行ってください。



The screenshot shows a search settings window titled "Fold Change". It contains two main sections. The first section, "Ratio data 比較", has an unchecked checkbox, a text input field with "0", and a radio button for "以上" (above). The second section, "Group Ratio data 比較", has a checked checkbox, a text input field with "2", and a radio button for "以上". Below this section is a blue link that says "すべての実験において条件を満たす場合". At the bottom, there are radio buttons for "AND" (selected) and "OR".

②-2 P-value(FDR)で検索する。

6ページで計算したグループ間発現量差のP-value(or FDR)の検索には、「P-value」「FDR」検索を使用します。それぞれの入力欄に数値を入力し、検索を実行すると、メイン画面上の「P-value」または「FDR」のデータに対して、指定の条件で抽出を行うことができます。なお、本解析ではP-valueのみ計算を行っているため、ここでは「P-value」が「0.01」以下と設定します。



The screenshot shows a search settings window titled "P-value". It contains two main sections. The first section, "P-value", has a checked checkbox, a text input field with "0", and a radio button for "以上" (above). The second section, "FDR", has an unchecked checkbox, a text input field with "0", and a radio button for "以上" (above). Below this section is a radio button for "以下" (below).

なお本解析では、Vol.2~4では設定を行った、Filter Optionsの「Ignored cellsを除外する」機能を使用しません。この機能を使用すると、全実験データ(本解析では6つ)のうち1つでも解析除外スポットがある遺伝子は、検索時に除去されてしまうからです。またGroup Ratioの計算時には、自動的に解析除外スポットのデータは、計算に使用しないようになっているうえ、統計検定のP-value(or FDR)においては、6ページで説明したように、計算時に解析除外スポットを除去することができます。そのためグループ間比較において、「Group Ratio」や「P-value(or FDR)」で検索を行う場合は、Filter Optionsの「Ignored cellsを除外する」に、チェックを入れないことを推奨しています。

③ノイズレベルの信頼性の低いデータを除外する。(CutOff)

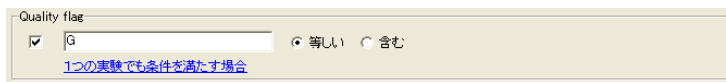
ここでは1比較解析のときと同様、1. Quality flagを用いる場合 2. Intensityを用いる場合の、2通りの解析手順を紹介します。

1・複数比較では、「G,G」のように2サンプル分のflagが、カンマで区切って記入されていますが、Signal比較解析用の実験データでは、「G」のように1サンプル分のflagしか記入されていません。そのためSignal比較解析では、検索文字列としてカンマを使用しないで下さい。

③-1 Quality flagを用いてCutOffを行う。

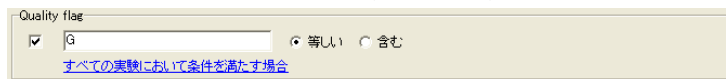
例1. 少なくとも1つの実験データで、Quality flagが「G」の遺伝子を抽出する場合

項目「Quality flag」で「Gと等しい」と指定し、複数データの検索条件の選択において、「1つの実験でも条件を満たす場合」を選択します。



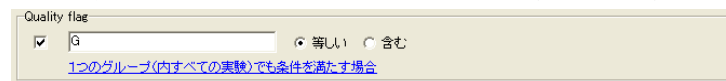
例2. すべての実験データで、Quality flagが「G」の遺伝子を抽出する場合

項目「Quality flag」で「Gと等しい」と指定し、複数データの検索条件の選択において、「すべての実験において条件を満たす場合」を選択します。



例3. 1つのグループでも、Quality flagが「G」の遺伝子を抽出する場合

項目「Quality flag」で「Gと等しい」と指定し、複数データの検索条件の選択において、グループ単位(「内1つの実験でも条件を満たすかどうか」or「内すべての実験で条件を満たすかどうか」)の、「1つのグループでも条件を満たす場合」を選択します。



***この検索条件は、実験データをグループ化していないと、使用できません。**


上記のようにCutOffを設定し、さらに②の発現変動遺伝子抽出の設定を同時に行うことで、発現変動を示し、なおかつノイズレベルのデータを除去した遺伝子のみを抽出することができます。

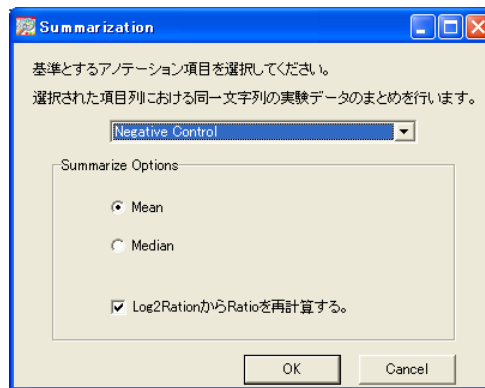
6サンプル全ての実験データで、Quality flagが「G」のデータのみを解析に使用すれば、データの信頼性は上がりますが、場合によっては条件が厳しすぎて、抽出される遺伝子が少なくなってしまうことがあります。例1～3を参考に、必要に応じて抽出条件の変更を行ってください。

③-2 intensityを用いてCutOffを行う。


Signal比較解析でintensityを用いてCutOffを行う場合は、複数比較の場合と同様に、サンプルごとのネガティブコントロールの値が異なるため、実験データ共通のCutOff値を決めなければいけません。以下にその手順を示します。

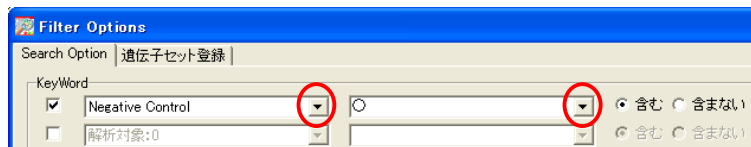
まずSummarizationを行い、データ内のネガティブコントロールスポット蛍光強度の平均値を求めます。

この処理を行うには、メイン画面上のSummarizationアイコンをクリックし、対象とするアノテーション項目に、「Negative Control」を選択してください(下図)。この状態で「OK」ボタンをクリックすると、計算が開始され、処理が終了するとメイン画面が表示されます。なおSummarizationを実行すると、統計検定で計算したP-value(or FDR)のデータは一度リセットされます。その場合は、Summarization実行後に、再び統計検定を実行し、P-value(or FDR)を計算してください。



次にメイン画面のデータの中から、Summarizationを行ったネガティブコントロールのデータを抽出します。

メイン画面上のFilter Optionsアイコンをクリックし、Key Word検索で左側の入力欄には「Negative Control」、右側の入力欄には「0」を指定します。指定を行うには、▼(下図)の部分をクリックし、表示される項目の中から目的の項目名をクリックしてください。項目の指定が終わったら、「OK」ボタンをクリックし、抽出を実行します。



それぞれのサンプルデータについて、ネガティブコントロールの値が表示されます。通常CutOffは、正規化を行った後のデータに対して行うことが多いので、基準とするネガティブコントロール値も、各サンプルのPer Array Normalized intensityの値を用います(下図)。また本解析では、6サンプルのデータがあるので、この中から最も小さい値をCutOffに使用します。

	腫瘍サンプル	腫瘍サンプル	腫瘍サンプル	腫瘍サンプル	腫瘍サンプル	腫瘍サンプル	腫瘍サンプル	腫瘍サンプル	腫瘍サ		
No.	Per Array No	Per Gene No	Quality flag	Ignored cells	Per Array No	Per Gene No	Quality flag	Ignored cells	Per Array No	Per Gene No	Quality
216	48.0613	-0.37204	L L L L L L L	0	56.68036	-0.23356	L L L G G L G	0	57.65342	-0.1545	G L L G
Mean value	-	-	-	-	-	-	-	-	-	-	-
Standard deviation	-	-	-	-	-	-	-	-	-	-	-

最後に、前ページで計算したネガティブコントロールの値を、CutOff値として設定を行います。

Filter Optionsの「実験データ」項目で、左のチェックボックスをクリックし、項目「Per Array Normalized intensity」でCutOff値を入力し、かつ複数データの検索条件の選択において、「すべての実験において条件を満たす場合」、または「1つの実験でも条件を満たす場合」などを選択します。

下図は、ノイズ除去の設定例です。

前ページの例では、Per Array Normalized intensityにおける6サンプルでの最小値が「48」だったので、このように設定を行いました。

実験データ

Per Array Normalized intensity 48 以上 0 以下 AND OR
[すべての実験において条件を満たす場合](#)

Per Array Normalized intensity 0 以上 0 以下 AND OR

Per Array Normalized intensity 0 以上 0 以下 AND OR

Per Array Normalized intensity 0 以上 0 以下 AND OR

Per Array Normalized intensity 0 以上 0 以下 説明

該当箇所を着色する

以上が検索条件の設定例です。ただし、「③ノイズレベルの信頼性の低いデータを除外する」で、③-2 intensityを用いてCutOffを行う場合は、Summarizationによるネガティブコントロール値の計算、および統計検定のやり直しを、変動遺伝子の抽出の前に行ってください。

最後にFilter Options検索画面右下の「OK」ボタンをクリックすると、データの抽出が実行され、メイン画面に抽出されたデータが表示されます。

No.	Select	Block	Row	Column	腫瘍サンプル		腫瘍サンプル		腫瘍サンプル		腫瘍サンプル		腫瘍サンプル		腫瘍サンプル		腫瘍サンプル		正常サンプル	
					Per Array No	Per Gene No	Quality flag	Ignored cells	Per Array No	Per Gene No	Quality flag	Ignored cells	Per Array No	Per Gene No	Quality flag	Ignored cells	Per Array No	Per Gene No	Quality flag	Ignored cells
1843	<input checked="" type="radio"/>	1	17	110	12857327	0.49504	G	0	12453997	0.44905	G	0	12541695	0.46919	G	0	3902905			
3388	<input type="radio"/>	1	32	9	13958617	0.32457	G	0	23414433	1.07061	G	0	17233038	0.62859	G	0	8149628			
5517	<input type="radio"/>	2	52	1	14872007	0.3998	G	0	16432782	0.54377	G	0	18408897	0.7076	G	0	679838			
6617	<input type="radio"/>	2	62	5	11909652	0.25884	G	0	18432188	0.88894	G	0	15276202	0.618	G	0	5618747			
7164	<input type="radio"/>	2	67	8	13240803	0.32163	G	0	17972337	0.76242	G	0	15843388	0.58052	G	0	794879			
12136	<input type="radio"/>	3	113	21	7490452	0.3505	L	0	10224778	0.79955	G	0	72562	0.30477	L	0	445266			
12409	<input type="radio"/>	3	115	74	8144463	0.48161	L	0	7190395	0.30187	G	0	9928627	0.76739	G	0	2792905			
12592	<input type="radio"/>	3	117	38	6568058	0.50678	L	0	78932	0.77193	G	0	5993608	0.37474	L	0	2500357			
14044	<input type="radio"/>	4	131	11	1139543	0.80005	G	0	8677975	0.40704	G	0	860731	0.39524	G	0	4482013			
16692	<input type="radio"/>	4	155	46	6178762	0.41221	L	0	7454855	0.66817	G	0	9695542	1.06338	G	0	2141023			
16723	<input type="radio"/>	4	155	77	9879498	0.47877	L	0	10336262	0.54121	G	0	9405145	0.40516	G	0	4795533			
17967	<input type="radio"/>	4	167	9	14391552	0.40267	G	0	18021417	0.72716	G	0	1292032	0.24709	G	0	57237			
17968	<input type="radio"/>	4	167	10	13850893	0.6193	G	0	11297778	0.32536	G	0	150355	0.73769	G	0	5863602			
18802	<input type="radio"/>	5	175	12	12049913	0.93545	G	0	1233933	0.97005	G	0	8034667	0.35074	G	0	4566632			
19459	<input type="radio"/>	5	181	21	11254332	0.71331	G	0	1433388	1.06226	G	0	892333	0.37848	G	0	2477723			
20328	<input type="radio"/>	5	189	7	7986913	0.42252	L	0	17354955	1.54218	G	0	12054987	1.01646	G	0	3662378			
24415	<input type="radio"/>	6	227	19	147334	0.86542	G	0	9752383	0.27016	G	0	14194395	0.91165	G	0	6421548			
24591	<input type="radio"/>	6	228	84	8901737	0.45834	L	0	1033014	0.67304	G	0	8060592	0.31514	G	0	4897168			
24732	<input type="radio"/>	6	230	3	5975398	0.66202	L	0	6924678	0.87473	G	0	4571645	0.2757	L	0	2848035			
25923	<input type="radio"/>	6	241	6	813992	0.39326	G	0	9794042	0.66015	G	0	952101	0.61936	G	0	3125012			
26298	<input type="radio"/>	6	244	56	7969163	0.50137	L	0	8081488	0.52157	G	0	8563712	0.60518	G	0	3290208			
27725	<input type="radio"/>	7	257	94	7420445	0.82395	L	0	5214155	0.31487	G	0	9257495	1.14306	G	0	174895			
28893	<input type="radio"/>	7	268	71	6331387	0.59304	L	0	5256365	0.32458	G	0	6128697	0.54421	L	0	3138368			
32538	<input type="radio"/>	8	302	54	511113	0.50661	L	0	6110112	0.76416	G	0	7312642	1.02336	L	0	1142293			
33181	<input type="radio"/>	8	308	49	6286625	0.32045	L	0	7316783	0.53933	G	0	944175	0.90717	G	0	3782395			
33593	<input type="radio"/>	8	312	8	1024035	0.3933	G	0	11979508	0.62401	G	0	12095462	0.63876	G	0	4141433			
34352	<input type="radio"/>	8	319	8	12813925	0.35379	G	0	18240957	0.89357	G	0	12592305	0.32961	G	0	7462293			
35616	<input type="radio"/>	8	330	77	529164	0.30403	L	0	9284683	1.11516	G	0	8039012	0.90733	G	0	32807			
35938	<input type="radio"/>	8	333	33	5336292	0.48114	L	0	8586387	1.16735	G	0	5933462	0.63417	L	0	1090085			
38498	<input type="radio"/>	9	357	69	6283662	0.36931	L	0	6295456	0.37203	G	0	701126	0.52738	L	0	2708188			
40974	<input type="radio"/>	10	380	112	8324098	0.44917	L	0	7627437	0.32307	G	0	1032305	0.75967	G	0	4566798			
41326	<input type="radio"/>	10	384	36	8289187	0.24586	L	0	120768	0.78879	G	0	1044418	0.57925	G	0	3596813			

本解析では、遺伝子の抽出に以下のような条件を設定しました。

- コントロール遺伝子などの除去: 解析対象 Flagが「0」
- 変動遺伝子抽出: 腫瘍グループ vs 正常グループで発現比が2倍以上かつ、P-valueが0.01以下
- ノイズデータ除去: 少なくとも1つの実験データで、Quality flagが「G」

この結果、43個の遺伝子が抽出されました。結論として、この43遺伝子は腫瘍グループの細胞で、発現が増加している遺伝子だといえます。また、この後Gene Ontology解析やPathway解析を行うことで、これらの遺伝子がどのような機能やパスウェイと関連があるかを調べることができます。

なお、腫瘍グループで発現が減少している遺伝子を抽出する場合は、その他の設定は変更せずに、Group Ratioの設定だけを「0.5以下」などに行えます。