RNA-Seqデータ解析と視覚化

フィルジェン株式会社　バイオインフォマティクス部
(biosupport@filgen.jp)

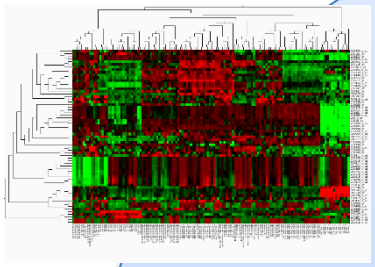
RNA-Seqデータ解析

Filgen® biosciences & nanosciences

Data import example (unsaved) 1.1

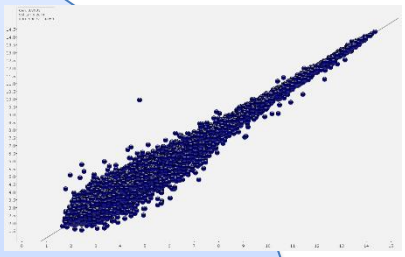
5305	5306	5307	5308	5309
0.025761	0.079122	0.17611	0.022006	-0.028059
0.74627	-0.40401	-1.7265	-0.31155	-0.88203
-0.59056	0.39811	0.2189	0.23436	0.016658
2.2421	2.6953	2.3714	1.9648	3.4437
-0.31883	-0.14073	0.46935	0.23085	-0.34026
0.14719	-0.33961	-0.3464	0.082997	-0.011759
0.31026	0.4793	-0.73888	0.55981	-0.39493
-0.53521	0.24693	-0.2871	-0.017066	0.13111

➢ データの正規化

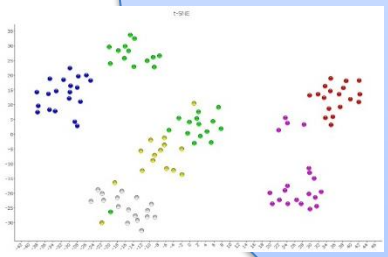


➢ クラスタリング

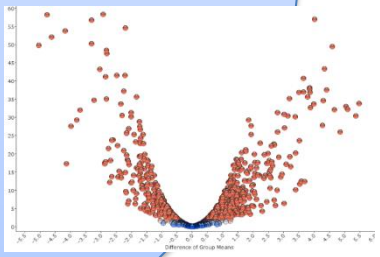
Qlucore Omics Explorer



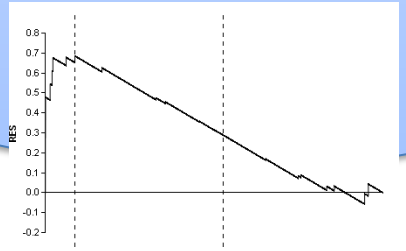
➢ クオリティコントロール



➢ 次元削減



➢ 発現差解析

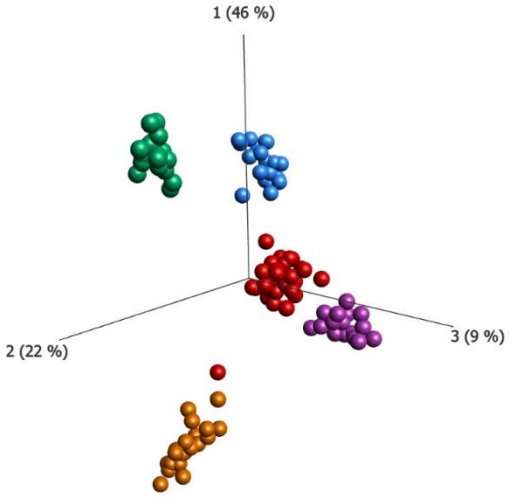


➢ 遺伝子セット解析

Qlucore Omics Explorer

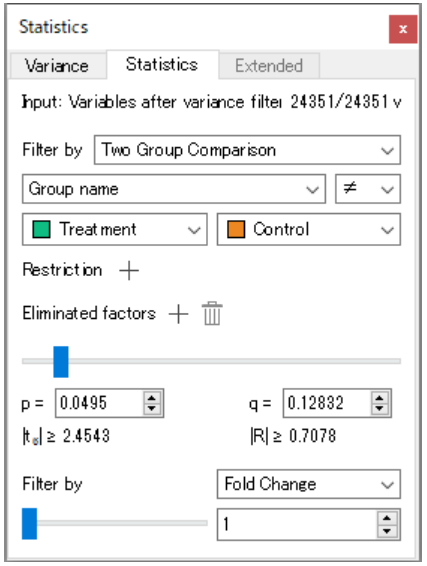
Visualize and Explore

- データのクオリティチェック
- データ構造の可視化
- 新たな仮説立て



Statistical Analysis

- t-test, ANOVA, 回帰分析
- Open API による統計メソッドのインテグレート
- 各種プロットやデータの作成と出力



Achieve Biological Insight

- アノテーションの探索
- GO Browser
- GSEA – Gene Set Enrichment Analysis

Machine learning

- データ分類モデルの構築
- kNN, SVM, RT, XG Boost
- 構築したモデルの新しいデータへの適用

データの正規化

非正規化データ
（リードカウントデータ）

BAM

quant.sf

TXT

正規化の実行

- TMM
- TPM
- FPKM

正規化済みデータ

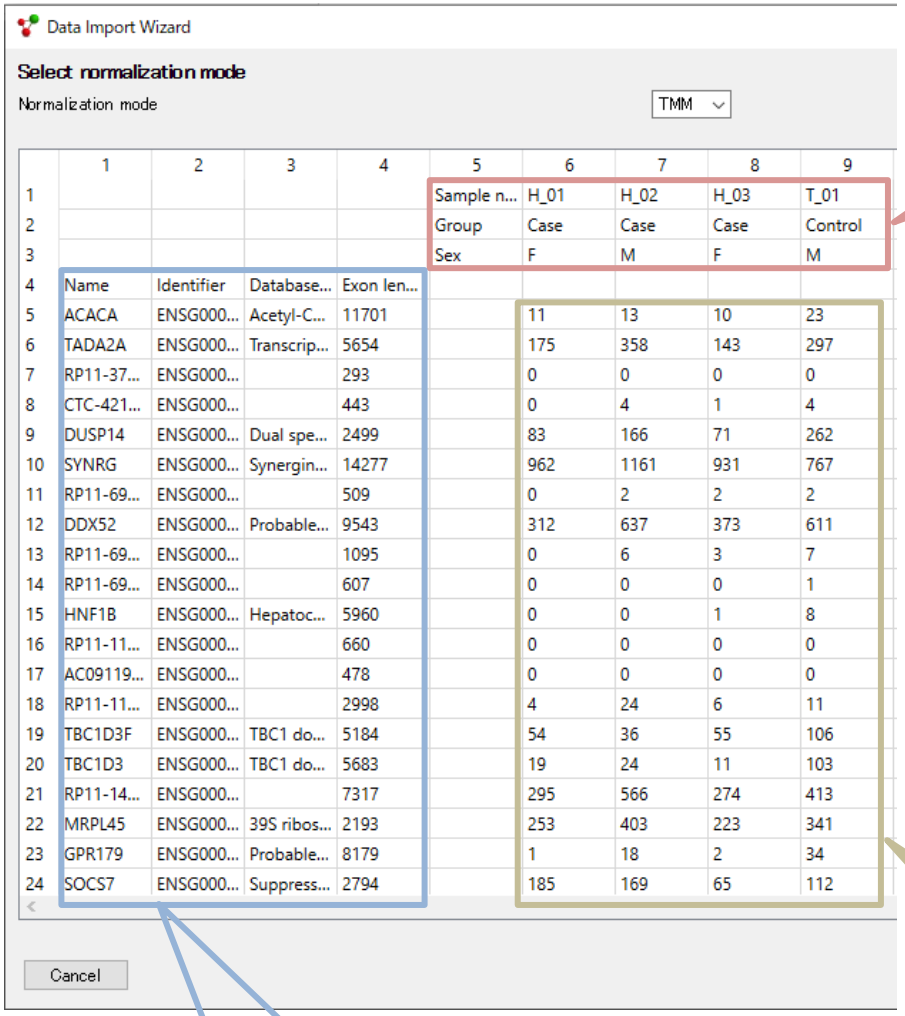
正規化済みデータ

10x

- barcodes.tsv.gz
- features.tsv.gz
- matrix.mtx.gz

TXT

リードカウントテキストの場合



	1	2	3	4	5	6	7	8	9
1					Sample n...	H_01	H_02	H_03	T_01
2					Group	Case	Case	Case	Control
3					Sex	F	M	F	M
4	Name	Identifier	Database...	Exon len...					
5	ACACA	ENSG000...	Acetyl-C...	11701		11	13	10	23
6	TADA2A	ENSG000...	Transcrip...	5654		175	358	143	297
7	RP11-37...	ENSG000...		293		0	0	0	0
8	CTC-421...	ENSG000...		443		0	4	1	4
9	DUSP14	ENSG000...	Dual spe...	2499		83	166	71	262
10	SYNRG	ENSG000...	Synergin...	14277		962	1161	931	767
11	RP11-69...	ENSG000...		509		0	2	2	2
12	DDX52	ENSG000...	Probable...	9543		312	637	373	611
13	RP11-69...	ENSG000...		1095		0	6	3	7
14	RP11-69...	ENSG000...		607		0	0	0	1
15	HNF1B	ENSG000...	Hepatoc...	5960		0	0	1	8
16	RP11-11...	ENSG000...		660		0	0	0	0
17	AC09119...	ENSG000...		478		0	0	0	0
18	RP11-11...	ENSG000...		2998		4	24	6	11
19	TBC1D3F	ENSG000...	TBC1 do...	5184		54	36	55	106
20	TBC1D3	ENSG000...	TBC1 do...	5683		19	24	11	103
21	RP11-14...	ENSG000...		7317		295	566	274	413
22	MRPL45	ENSG000...	39S ribos...	2193		253	403	223	341
23	GPR179	ENSG000...	Probable...	8179		1	18	2	34
24	SOCS7	ENSG000...	Suppress...	2794		185	169	65	112

サンプルアノテーション

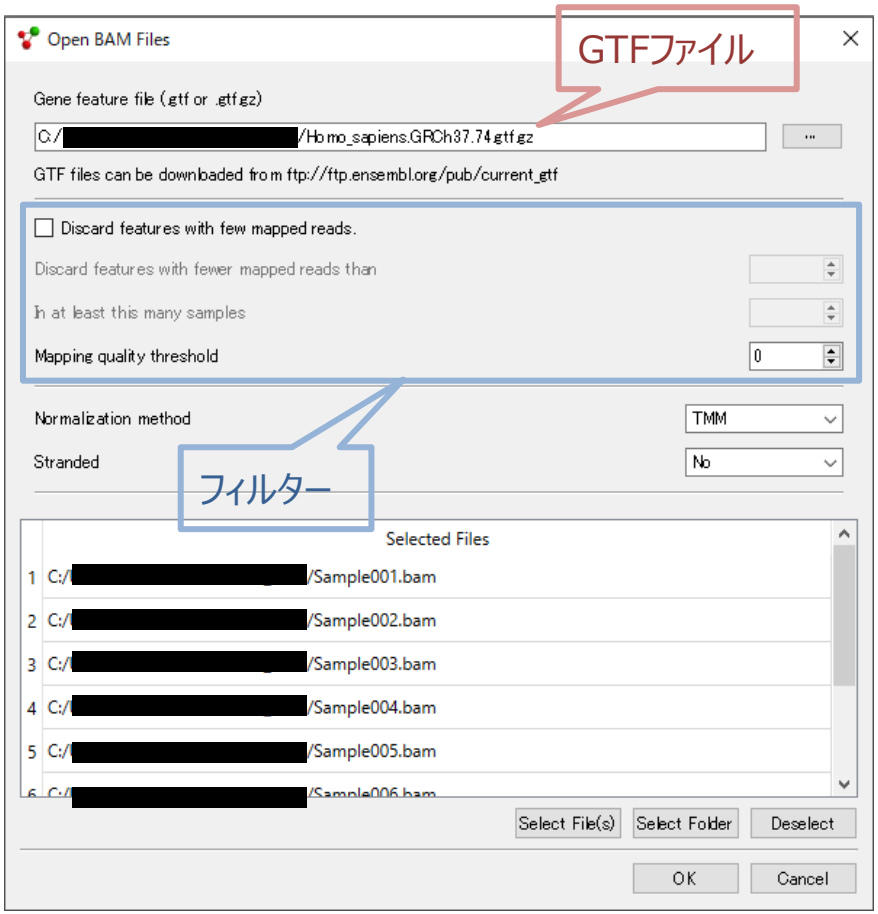
遺伝子アノテーション

リードカウントデータ

■ **リードカウントデータの正規化**

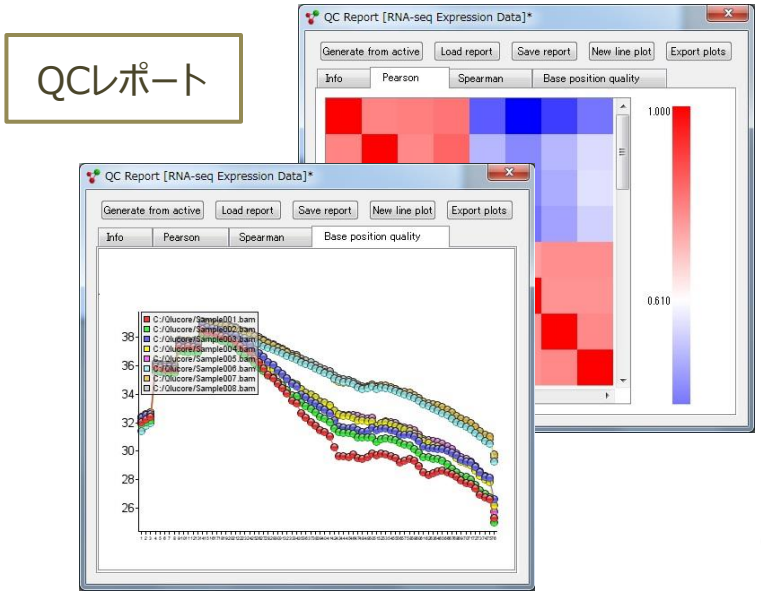
- 遺伝子ごとのリードカウント数を、全サンプル分まとめたマトリックスデータファイルを作成
- リードカウント数に加え、遺伝子アノテーションとサンプルアノテーションも、ファイルに書き込みが可能
- 遺伝子アノテーションとして、正規化の計算に必要な、遺伝子長（またはエクソン長）のデータが必要

BAMファイルの場合



■ BAMファイルの正規化

- サンプル別のBAMファイルと、リファレンスゲノムのGTFファイルが必要
- BAMファイル読み込み時に、リードや遺伝子のクオリティーなどで、フィルターをかけることが可能
- 読み込み完了時に、正規化済みデータに加え、サンプル間の相関やリードクオリティーのQCグラフを出力
- アノテーションは、別ファイルでインポートが必要



発現データテーブル

Example_count_data (unsaved) 1.1

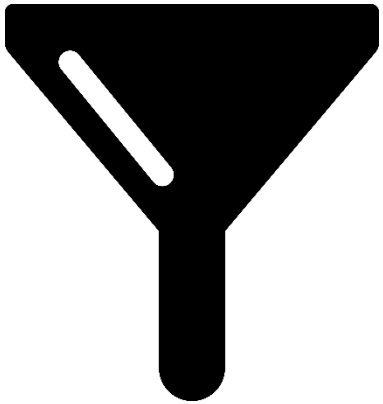
	p-values	q-values	SampleA	SampleB	SampleC	SampleD	SampleE	SampleF	SampleG	SampleH	SampleI	SampleJ
AC003958.			6.5964	11.271	11.506	10.851	10.972	11	10.767	6.5629	11.234	11.112
AC003958.			7.994	14.093	14.498	14.738	13.825	14.579	13.784	9.4724	14.35	13.609
AC004231.			14.487	16.029	15.667	15.29	15.166	15.073	14.996	14.675	15.993	15.653
AC006449.			6.611	7.7559	7.7807	8.7107	7.6518	8.5915	9.0316	6.8671	7.8375	7.9281
AC019349.			6.2319	7.3768	7.4016	8.3316	7.2727	8.2124	8.6525	6.488	7.4584	7.549
AC087491.			5.8599	7.0048	7.0297	7.9596	6.9008	7.8404	8.2805	6.116	7.0864	7.177
AC090283.			5.7691	6.9141	6.9389	7.8689	6.81	7.7496	8.1897	6.0252	6.9956	7.0863
AC091199.			13.483	13.846	13.839	12.573	13.495	13.064	13.039	13.168	13.827	13.986
AC124789.			10.396	10.379	10.259	11.333	10.637	10.729	10.432	10.98	10.823	11.288
ACACA			3.1973	4.3422	4.3671	6.034	4.2382	5.9148	5.6179	6.5123	6.2983	5.7368
ARHGAP23			2.1584	3.3033	3.3281	4.2581	3.1992	4.1389	4.579	2.4145	3.3849	3.4755
ARL5C			4.3511	5.496	5.5208	6.4508	5.3919	6.3316	6.7717	4.6072	5.5776	5.6682
C17orf96			3.4707	4.6156	4.6405	5.5704	4.5116	5.4512	5.8913	3.7268	4.6972	4.7879
C17orf98			14.456	13.076	13.259	12.781	13.434	12.447	12.887	13.906	12.777	13.206
CACNB1			2.3343	3.4792	3.5041	4.434	3.3752	4.3148	4.7549	2.5904	3.5608	3.6514
CASC3			11.507	11.21	10.857	11.157	11.567	11.587	11.963	11.729	10.455	11.153
CCR7			3.8344	4.9793	5.0041	5.9341	4.8752	5.8149	6.255	4.0904	5.0609	5.1515
CDC6			9.7429	10.276	9.9303	9.4096	10.059	9.084	8.1456	10.135	10.607	10.651
CDK12			1.8011	2.946	2.9709	3.9008	2.842	3.7816	4.2217	2.0572	3.0276	3.1182
CISD3			3.7479	4.8928	4.9176	5.8476	4.7887	5.7283	7.7534	4.0039	4.9743	5.065
CSF3			6.3323	9.5476	10.853	6.1101	8.7516	8.3128	10.338	5.8514	8.0442	7.6494

- 正規化された遺伝子発現データは、テーブル形式で表示
- 必要に応じて、タブ区切りテキストでファイル出力

クオリティコントロール

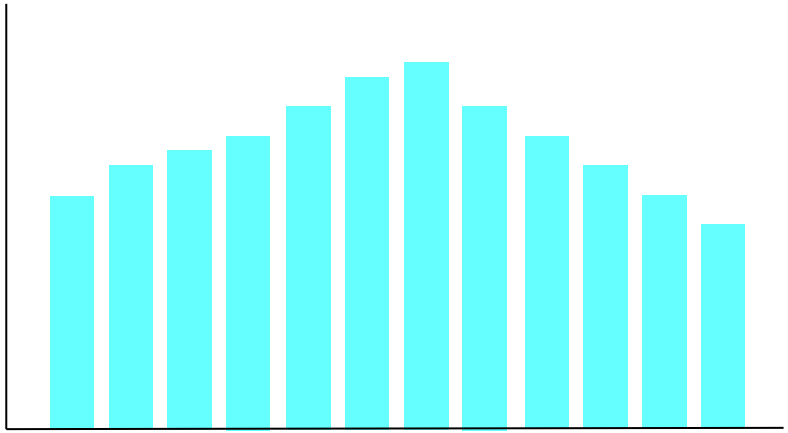
全データセット

ノイズフィルタリング



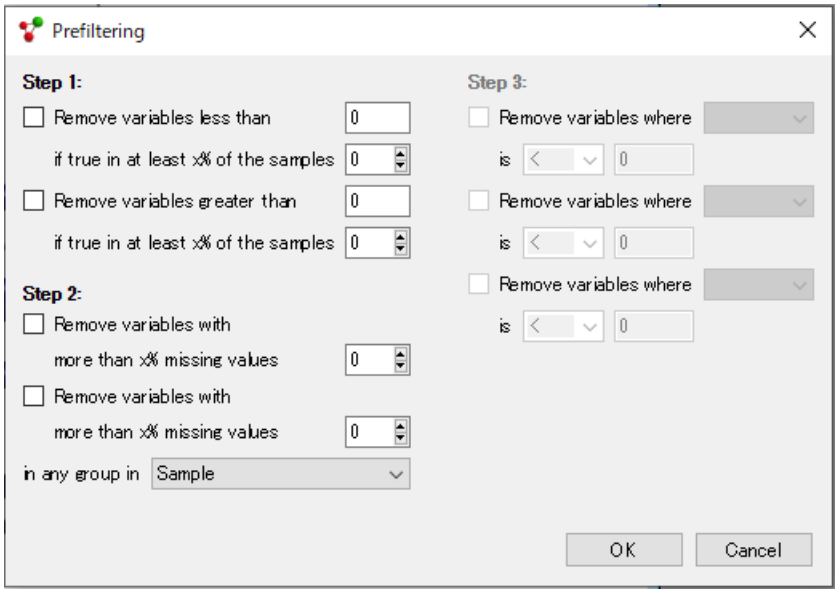
- ✓ 低発現遺伝子
- ✓ サンプル間発現変動の小さい遺伝子　…など

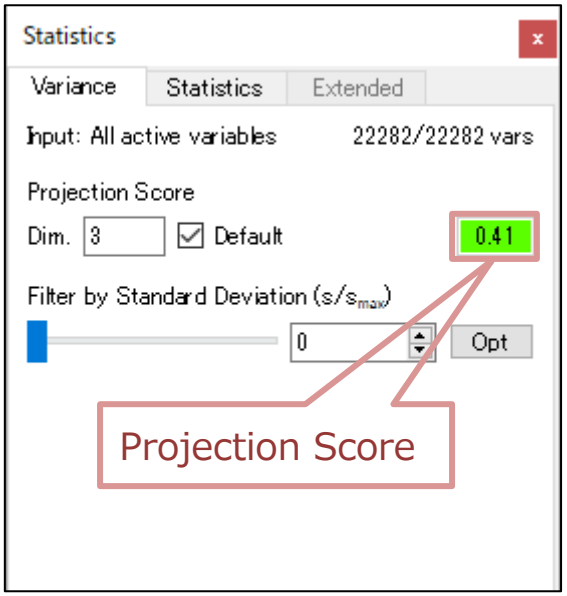
データ分布の確認



- ✓ ヒストグラム
- ✓ スキャッタープロット

ノイズフィルタリング





■ プレフィルター

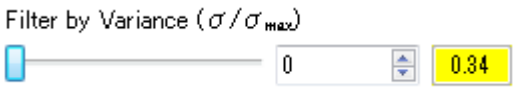
- 低発現遺伝子や、欠損値をもつ遺伝子などをフィルタリング
- 遺伝子アノテーションのデータに基づいたフィルタリングも可能

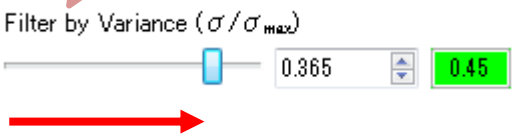
■ 分散フィルター

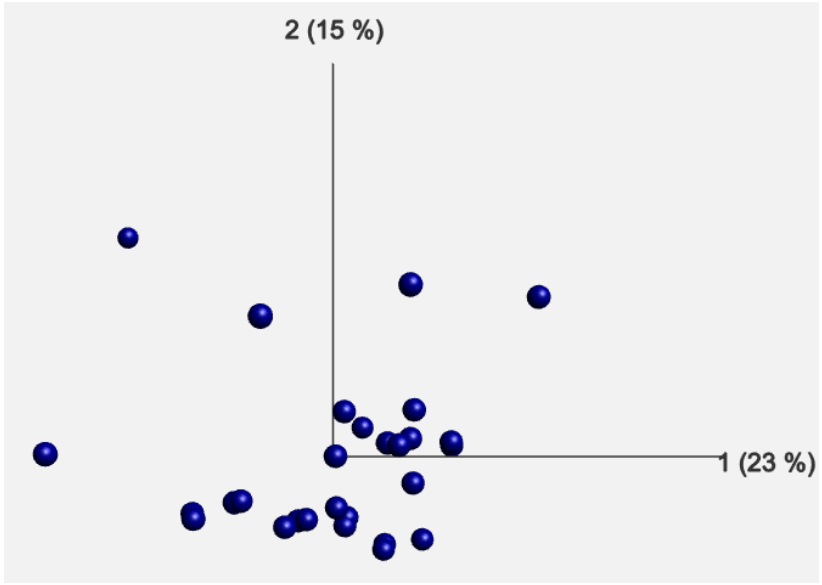
- サンプル間の発現変動のばらつきが小さい遺伝子をフィルタリング
- Projection Scoreを基準にすることで、フィルタリングの閾値を決定

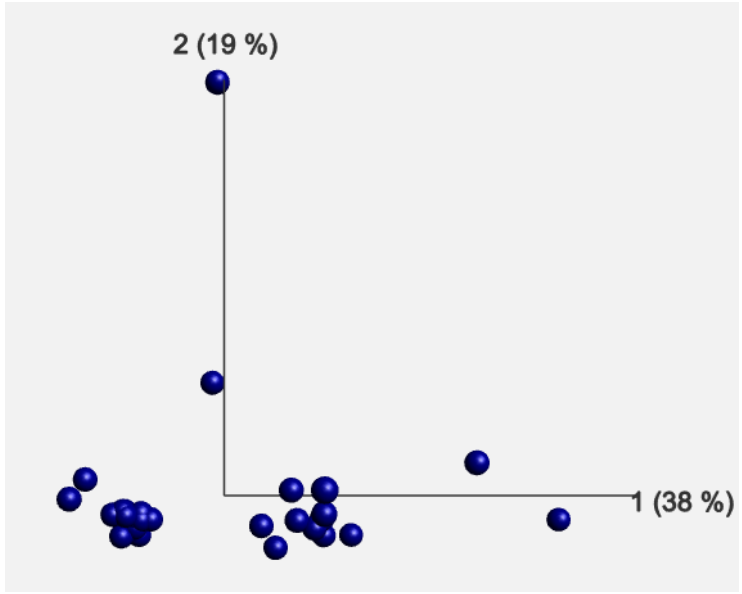
プロットへの反映

分散フィルターのスライダーバーを
横にスライド



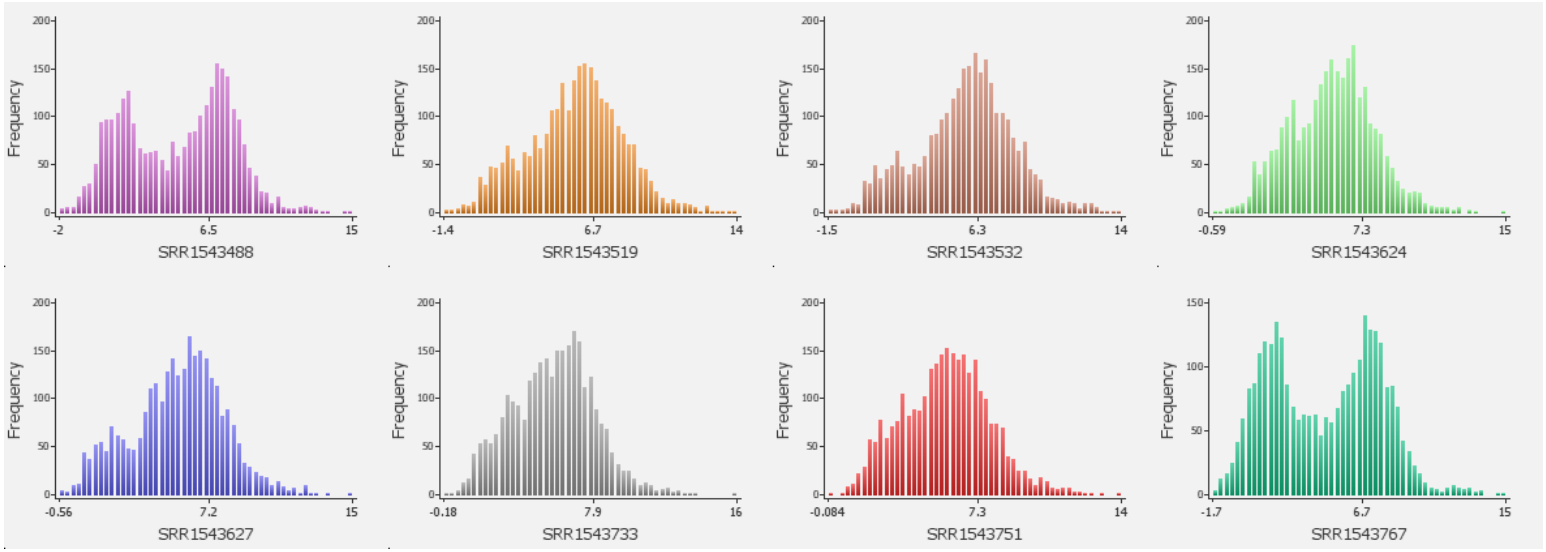




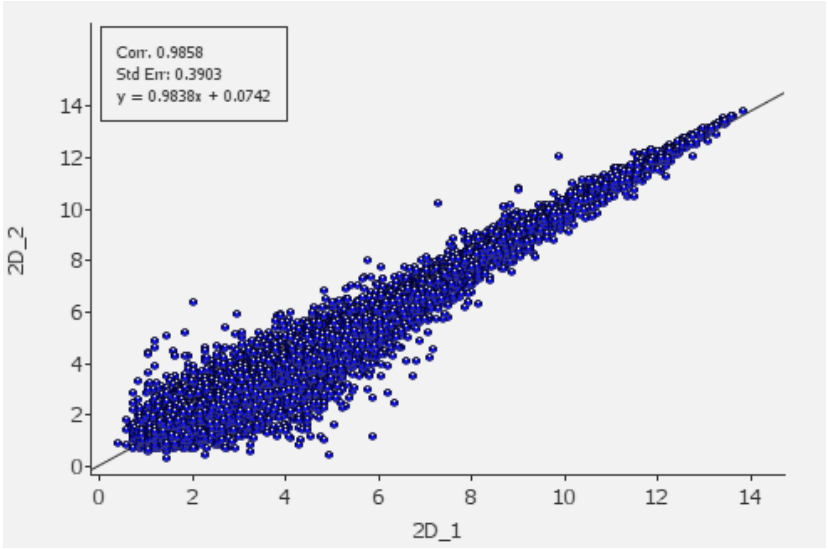


- フィルタリングを適用すると、データのプロットグラフもリアルタイムで変化する

データ分布の確認



➢ ヒストグラム



➢ スキャッタープロット

■ グラフプロットによるデータの確認

- ヒストグラムでサンプル内、スキャッタープロットでサンプル間の、それぞれデータ分布を確認
- 他のサンプルと比べて、データ分布が極端に異なるものがあれば、解析から除外するか再実験を検討する

発現差解析

Filgen® biosciences & nanosciences

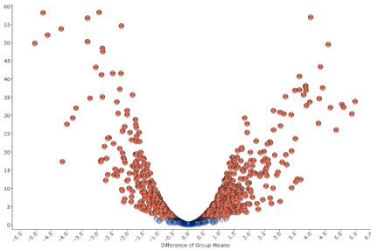
サンプルアノテーション

- **順位のないカテゴリー**（ 薬剤処理･未処理 など ）
- **順位のあるカテゴリー**（ 高用量･中用量･低用量 など ）
- **連続値**（ 年齢 など ）

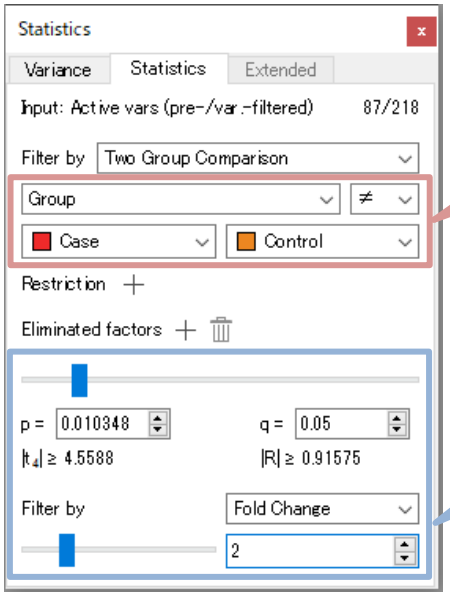
統計解析

- ✓ t-test
- ✓ Paired t-test
- ✓ ANOVA
- ✓ Two way ANOVA
- ✓ Rank Regression
- ✓ Linear Regression

可視化



設定と解析結果



■ 解析の実行とフィルタリング

- 統計解析実行時には、手法の選択に加え、サンプルのグループ分類に用いるサンプルアノテーションを選択する
- フィルター条件に閾値を入力、またはスライダーバーをスライドさせることで、フィルタリングも可能

■ 解析結果の確認

- P値などの統計解析結果データは、遺伝子アノテーションとともに、遺伝子リストとして表示
- 必要に応じて、遺伝子リストをファイル出力

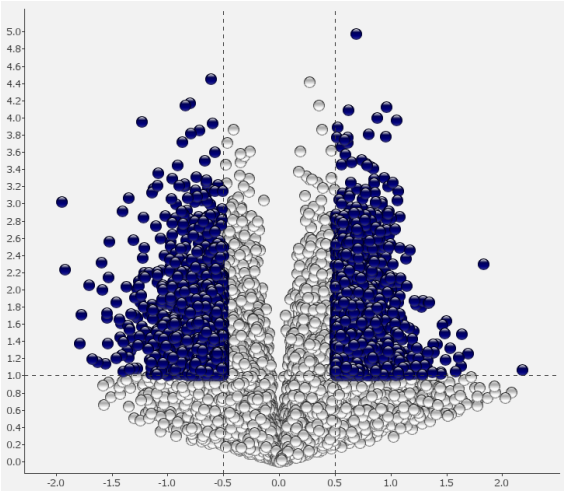
IDs

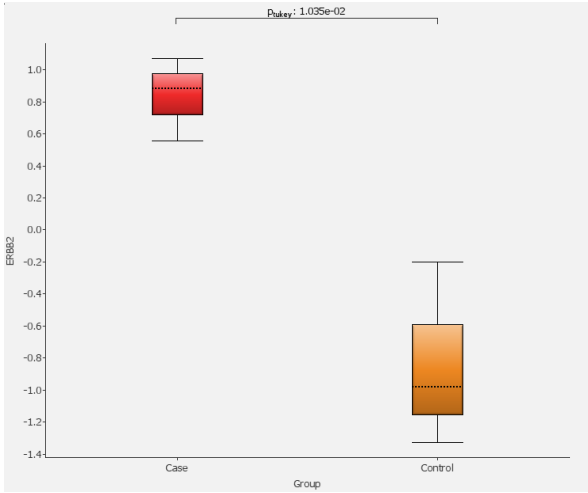
	Name	p-value	q-value	Difference	Fold change
1	AC003958.2	0.00675517	0.0381376	-4.96792	0.0319527
2	AC124789.1	0.000484734	0.0140573	-3.61779	0.0814586
3	ERBB2	0.0103481	0.0428707	3.98524	15.8372
4	GJD3	0.00913235	0.0397257	-3.71282	0.0762658
5	HNF1B	0.00855457	0.039527	-4.14099	0.0566809
6	KRT12	0.00146431	0.0254789	-2.33915	0.197627

1 / 1: 86

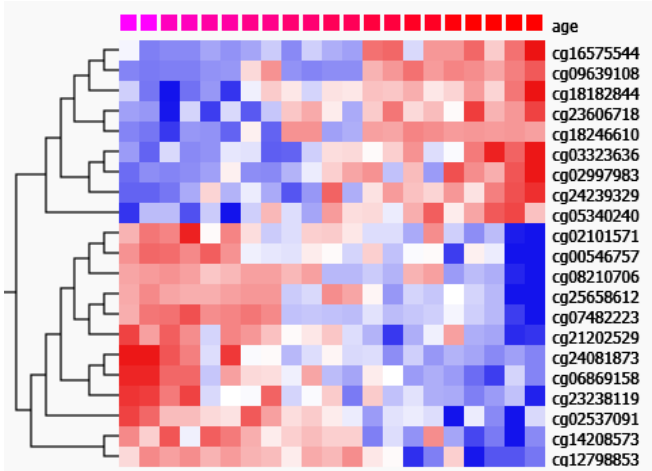
Database object name	Receptor tyrosine-protein kinase erbB-2
Exon length	10321
Identifier	ENSG00000141736
Name	ERBB2

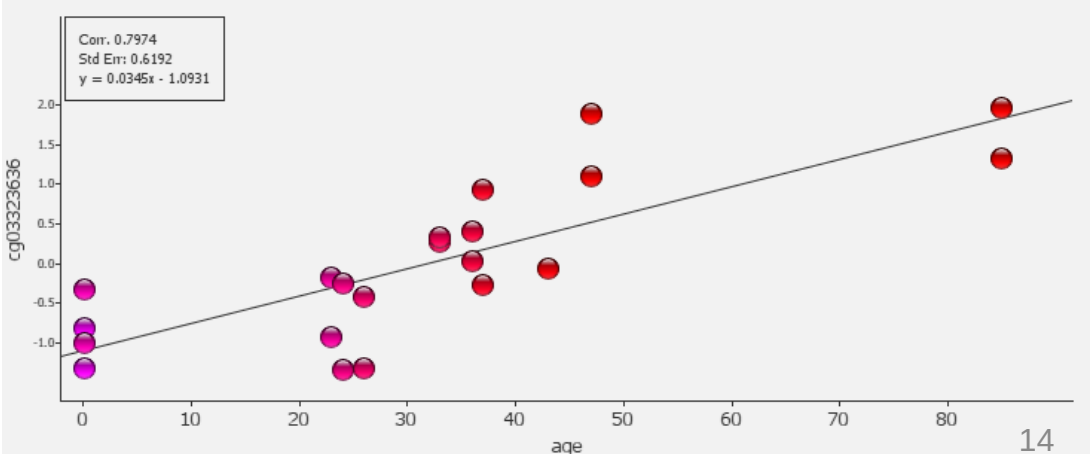
解析結果のグラフ表示



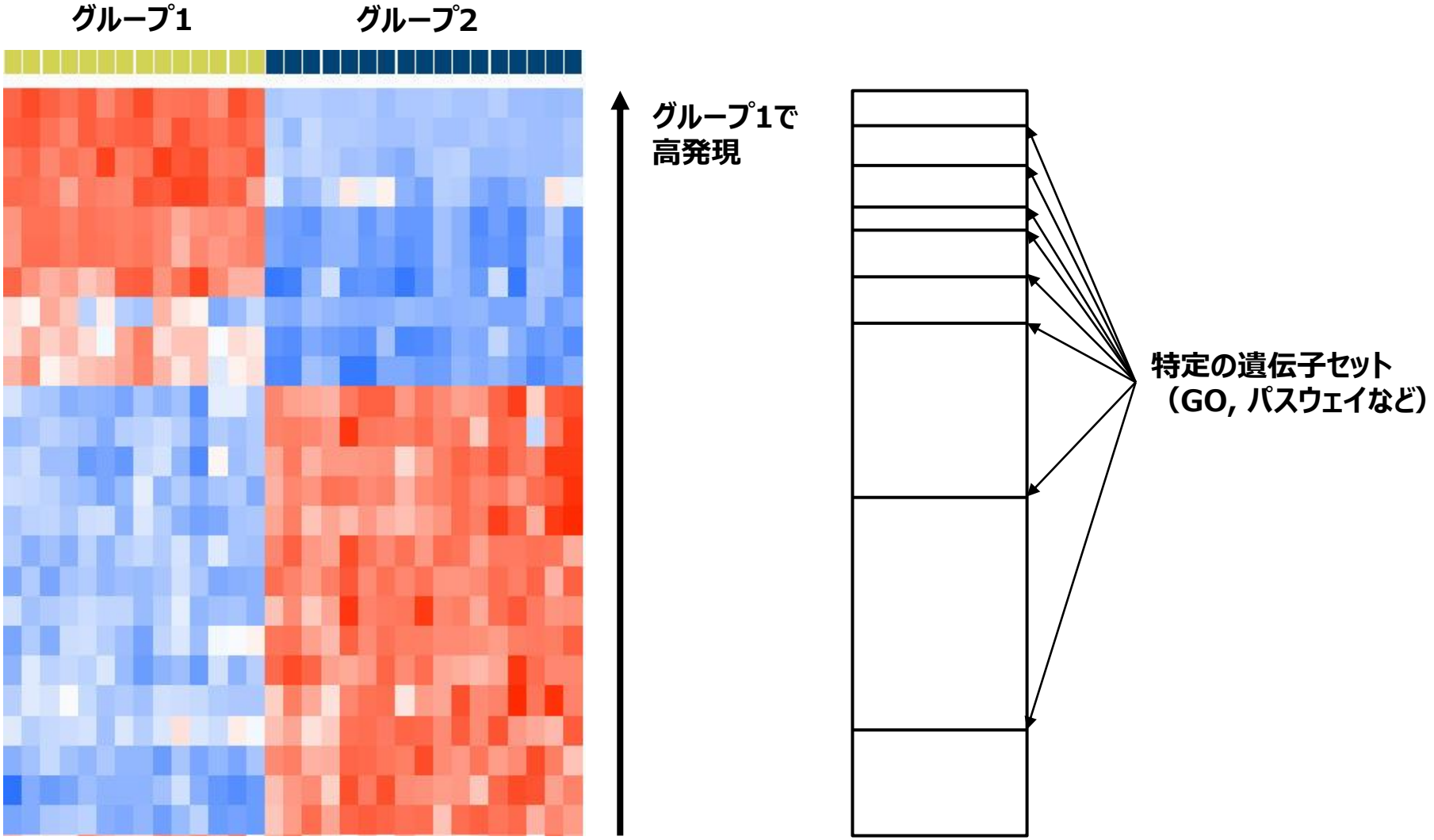


- ボルケーノプロットやボックスプロットで、カテゴリー間で有意差をもつ遺伝子をグラフ表示
- サンプルアノテーションが連続値や順位のあるカテゴリーの場合は、ヒートマップやスキャッタープロットなどで、データの傾向などを可視化





遺伝子セット解析



- 遺伝子セット解析（GSEA）では、発現差解析の結果に基づきデータセット内の全遺伝子をソートし、特定の遺伝子セットが、高ランクの遺伝子に濃縮されているかどうかの検定を行う

GSEAの実行

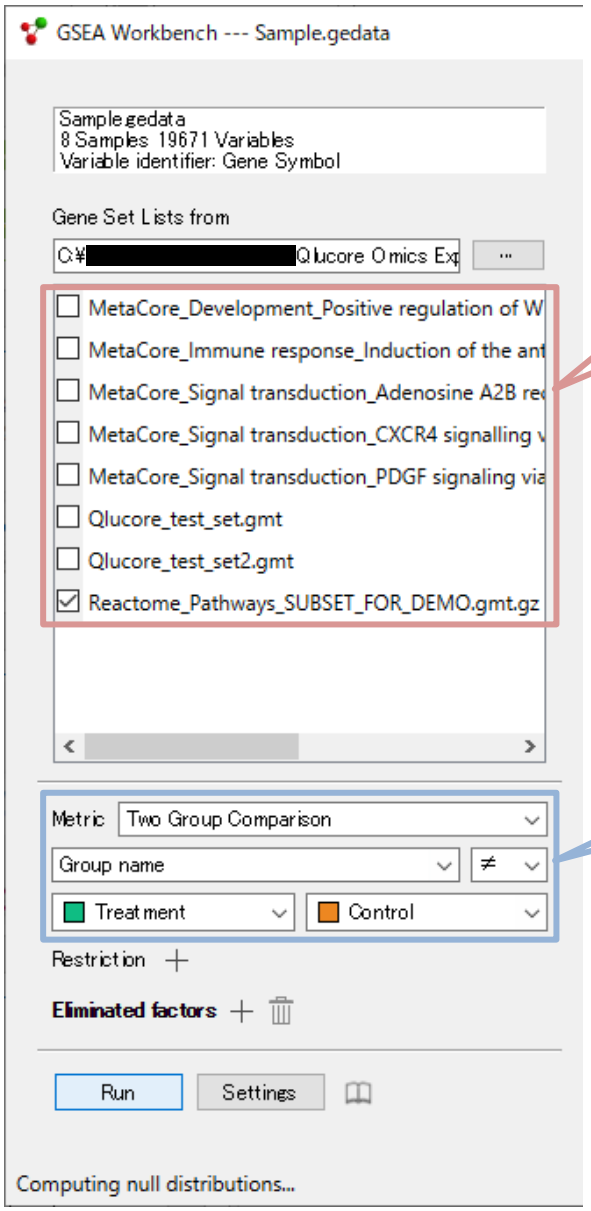
■ **遺伝子セットの選択**

- GSEA実行の際は、遺伝子セットデータベースが必要（GSEA Webサイトなどからダウンロード可能）

  ✓ パスウェイ
  ✓ Gene Ontology
  ✓ がん関連遺伝子
  ✓ 免疫関連遺伝子　など

■ **遺伝子のランク付け**

- 有意差検定のP-valueなど、遺伝子のランク付け条件を指定

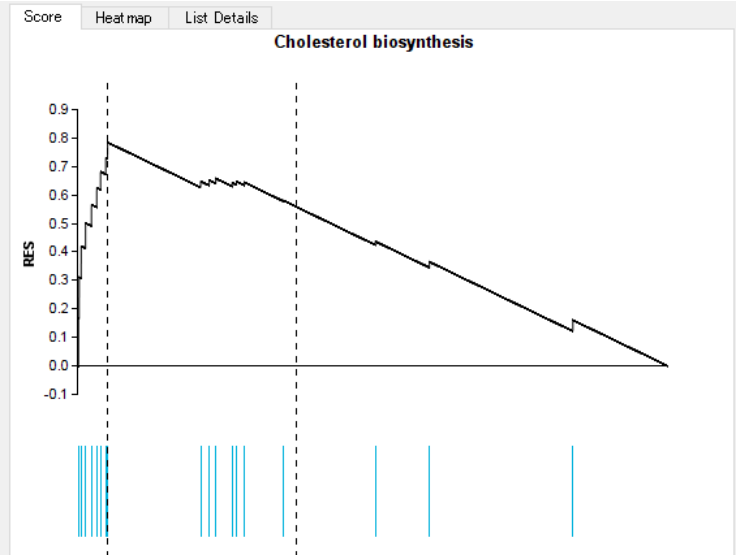


GSEA実行結果

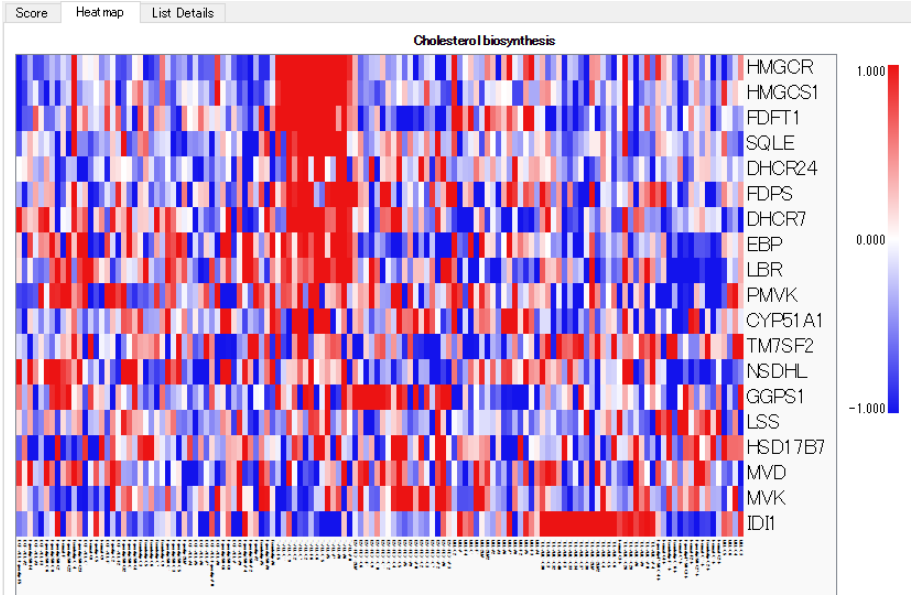
	Name	Size	Matches	ES	NES	p	q
1	Cholesterol biosynthesis	22	19	0.78	2.02	0	0.095513
2	Activation of Gene Expression by ...	41	38	0.56	1.86	0.0037951	0.35581
3	Regulation of Cholesterol ...	54	51	0.57	1.86	0.003937	0.23861
4	Effects of PIP2 hydrolysis	25	22	0.57	1.84	0.002004	0.22566
5	TCR signaling	63	52	0.49	1.68	0.021401	1
6	Formation of incision complex in G...	20	19	0.67	1.68	0.019881	0.86576
7	Dual incision reaction in GG-NER	20	19	0.67	1.68	0.019881	0.86576
8	Respiratory electron transport	76	55	0.78	1.67	0.0057915	0.78391
9	Respiratory electron transport, ATP ...	94	71	0.75	1.67	0.0096712	0.71466
10	Biosynthesis of the N-glycan ...	31	22	0.57	1.66	0.02	0.71047
11	The citric acid (TCA) cycle and ...	131	101	0.67	1.63	0.027397	0.78357
12	Generation of second messenger ...	36	29	0.56	1.63	0.031936	0.78177
13	mRNA Processing	156	134	0.59	1.62	0.056202	0.77492
14	Processing of Capped Intron-...	137	118	0.62	1.61	0.051923	0.74446

➢ **遺伝子セットリスト**

- 解析結果として、有意な遺伝子セットのリストと、各遺伝子セットのエンリッチメントスコアやヒートマップを出力

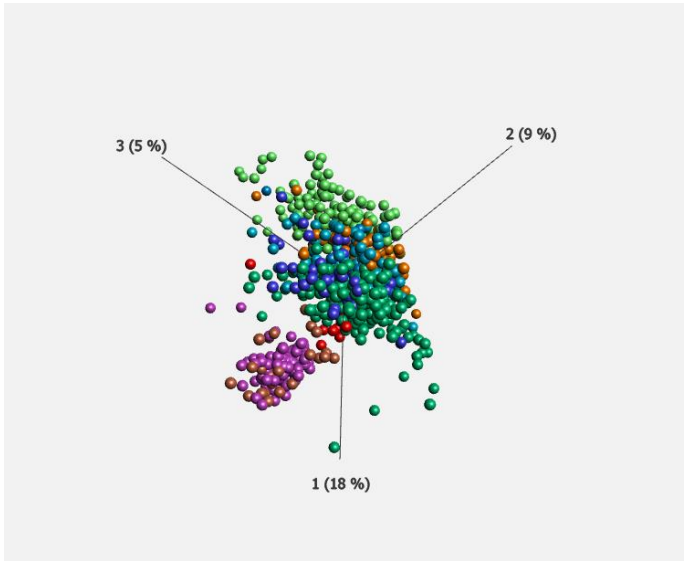


➢ **エンリッチメントスコア**

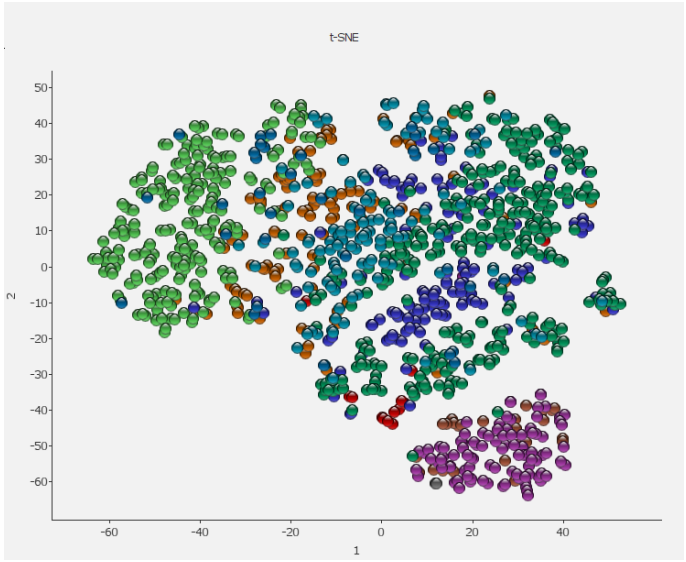


➢ **ヒートマップ**

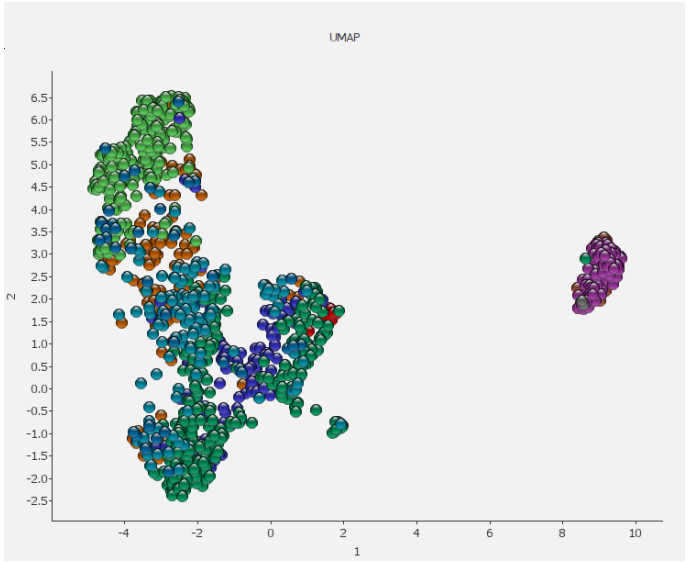
次元削減



➢ PCA（主成分分析）

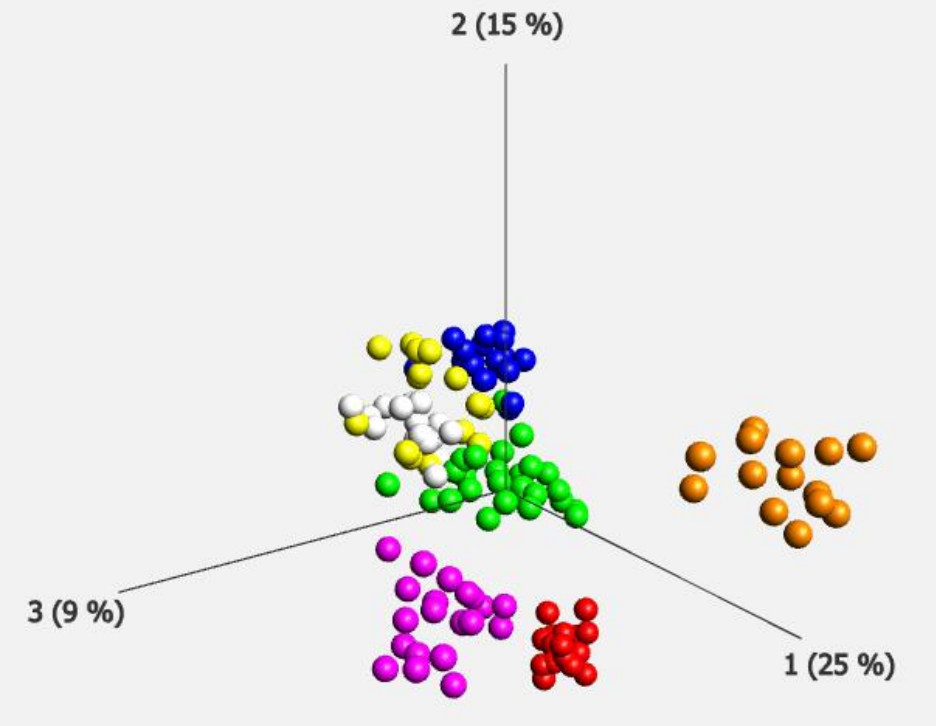


➢ t-SNE

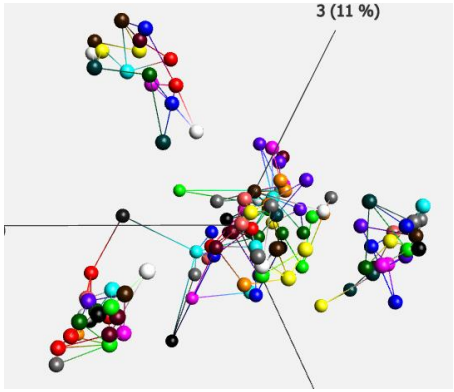


➢ UMAP

PCA



- 多次元データを、データの特徴を保持したまま、2または3次元に削減してプロット
- 計算が高速なため、遺伝子フィルタリング結果がリアルタイムでプロットに反映される
- プロットに対して、サンプルと遺伝子の切り替えや、データ間の相関を示す近傍グラフ表示などの調整が可能
- 各主成分に対する、サンプルまたは遺伝子ごとの座標や負荷量（寄与率）をファイル出力可能

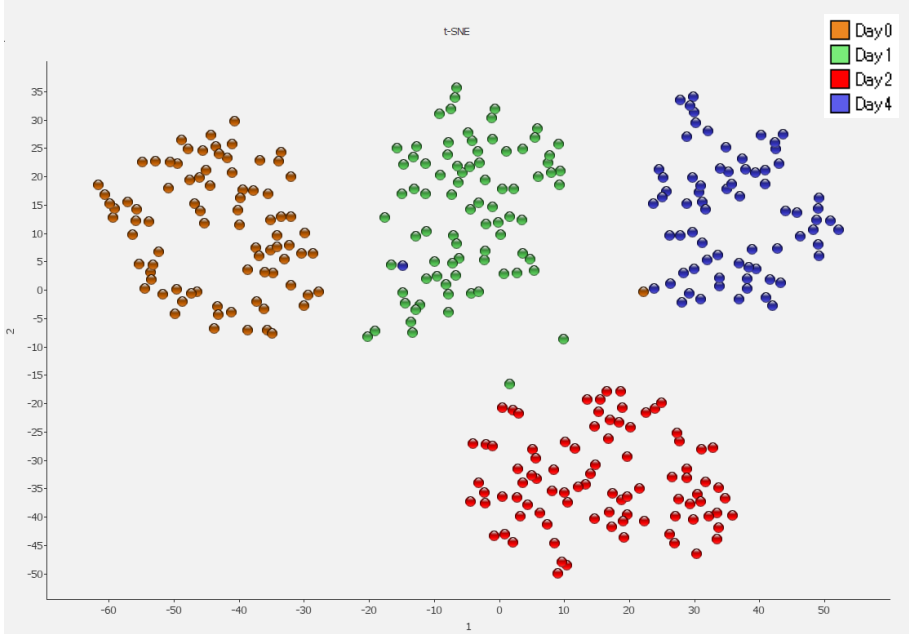


➢ **近傍グラフ表示**

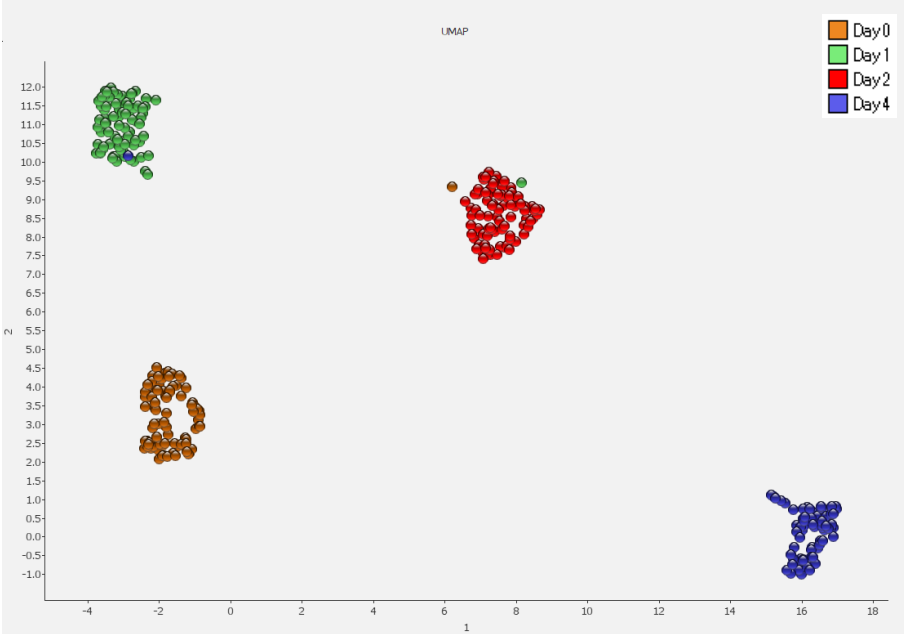
Variable Loadings									
ID	MTATP6P1	AURKAIP1	RPL22	RPL11	SH3BGRL3	ATPIF1	NDUFS5	YBX1	RPS8
PC1	−0.02699	−0.0083	0.099984	0.097311	−0.04408	−0.00138	0.030611	0.06949	0.10438
PC2	0.060679	0.090874	0.011818	0.045463	0.024263	0.099782	0.049087	0.02327	0.025292
PC3	0.051128	−0.02064	−0.02759	0.028116	−7.59E−04	−0.04239	−0.06956	−0.06762	0.002499
PC4	−7.71E−04	−0.02001	−0.02541	0.006545	0.10319	0.001674	0.009379	−0.02056	−0.05719
PC5	−0.16914	0.04707	−0.03074	0.017845	0.025878	0.053424	0.025652	−0.00418	0.024576
Sample Coordinates									
ID	1_AGTCTT	1_ATGTGT	1_CGTTCT	1_GGACAG	1_GGCGAC	1_GGGAGA	1_GTAACT	1_GTAGGC	1_GTCCTC
PC1	−11.77	−17.294	−13.259	−20.774	−9.8751	−12.789	−13.901	−14.473	−16.363
PC2	1.689	−2.5748	0.77786	−10.802	2.7939	2.652	−0.8279	0.23825	−4.9049
PC3	5.5775	3.9924	6.9636	2.5374	3.0656	4.777	−0.34342	4.4677	6.6199
PC4	3.3844	4.0225	5.342	0.43697	5.5632	2.1499	5.1986	−1.3531	2.4186
PC5	1.1099	1.9496	1.618	1.4684	−0.53881	3.1469	0.52476	0.14125	0.050511

➢ **座標・負荷量のファイル出力**

t-SNE, UMAP



➢ **t-SNE**

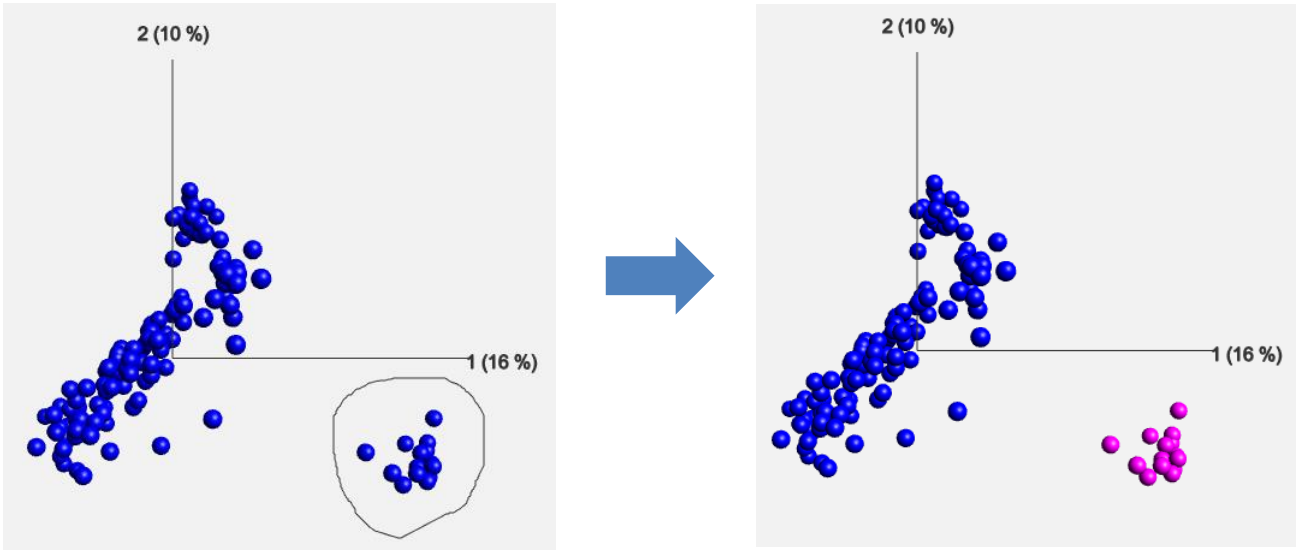


➢ **UMAP**

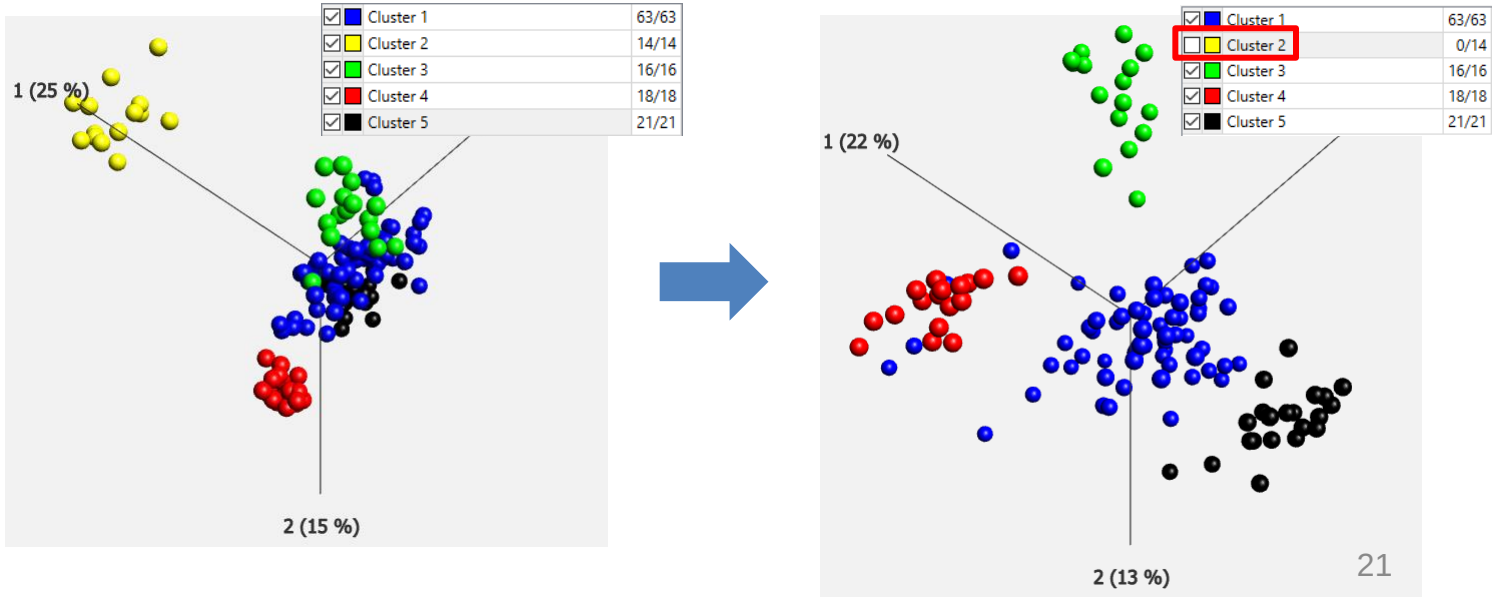
- t-SNE, UMAPともに、非線形的な次元削減の手法で、2次元プロットのみ可能
- フィルタリング結果のプロットへのリアルタイムの反映は不可
- 両手法とも、PCAに比べてデータの局所的な関係をつかみやすい

クラスターのアノテーション

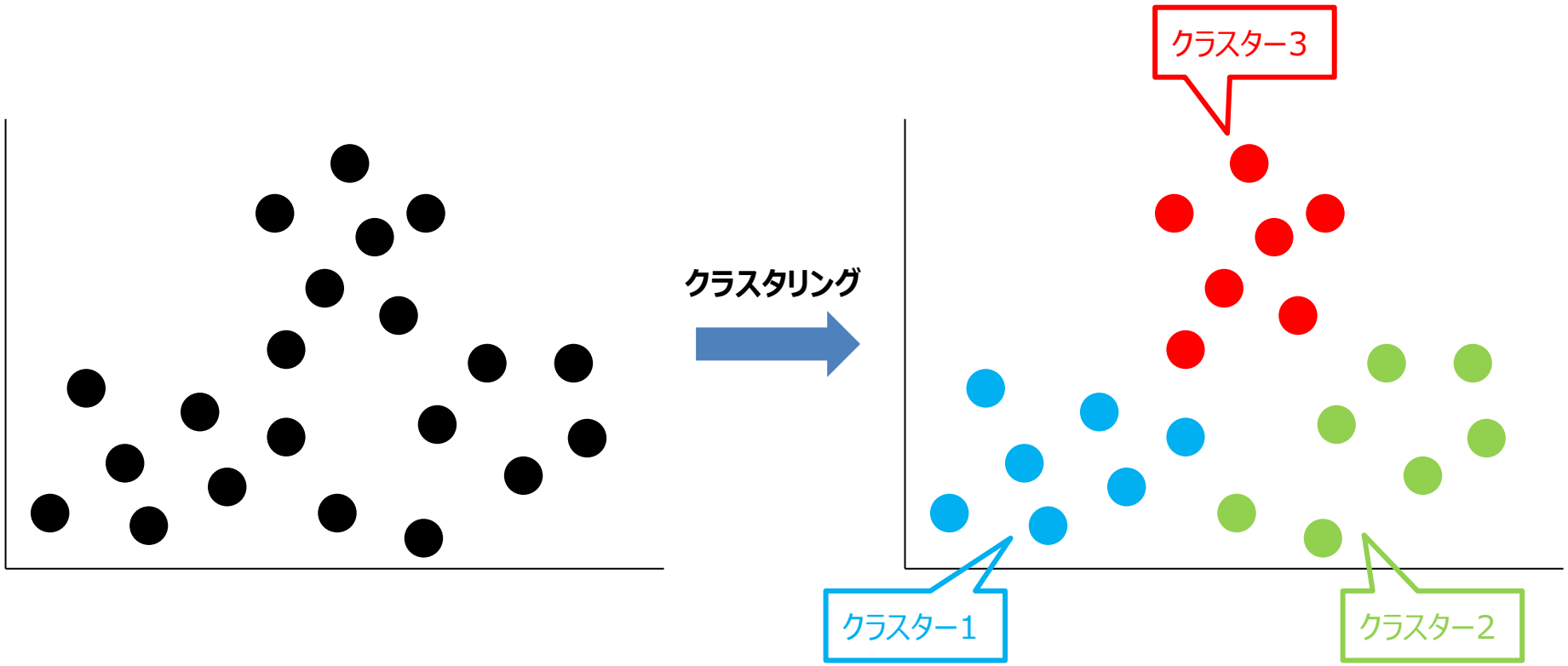
- クラスターの手動アノテーション



- アノテーション除去と連動したプロットの変化

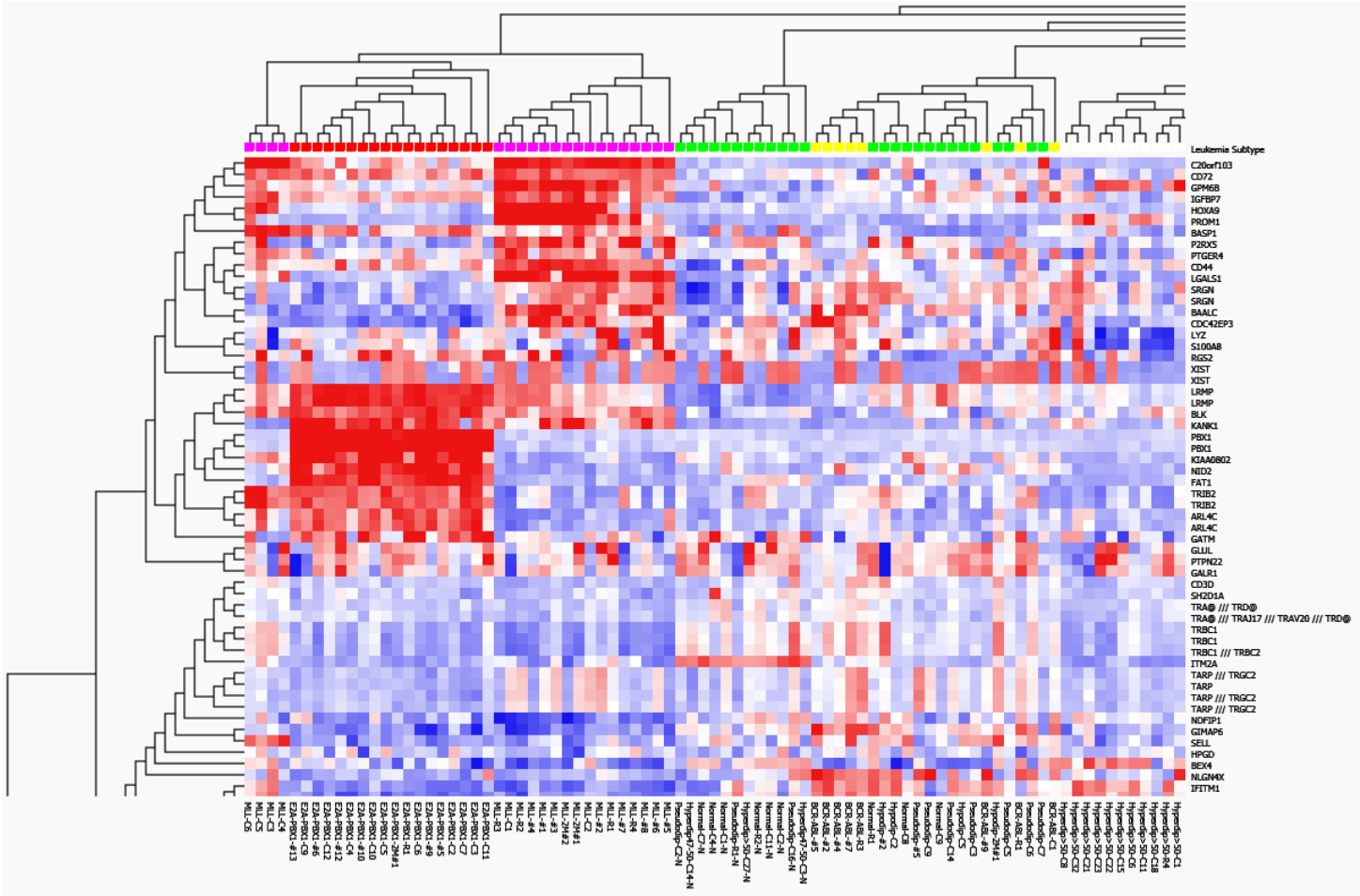


クラスタリング



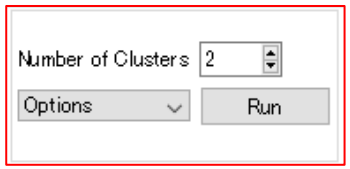
- 発現プロファイルデータより、各サンプルまたは遺伝子を、自動またはユーザー指定のパラメータに基づきクラスターに分類
- Qlucore Omics Explorerでは、2種類のクラスタリング手法が利用可能
  - 階層型クラスタリング
  - K-meansクラスタリング

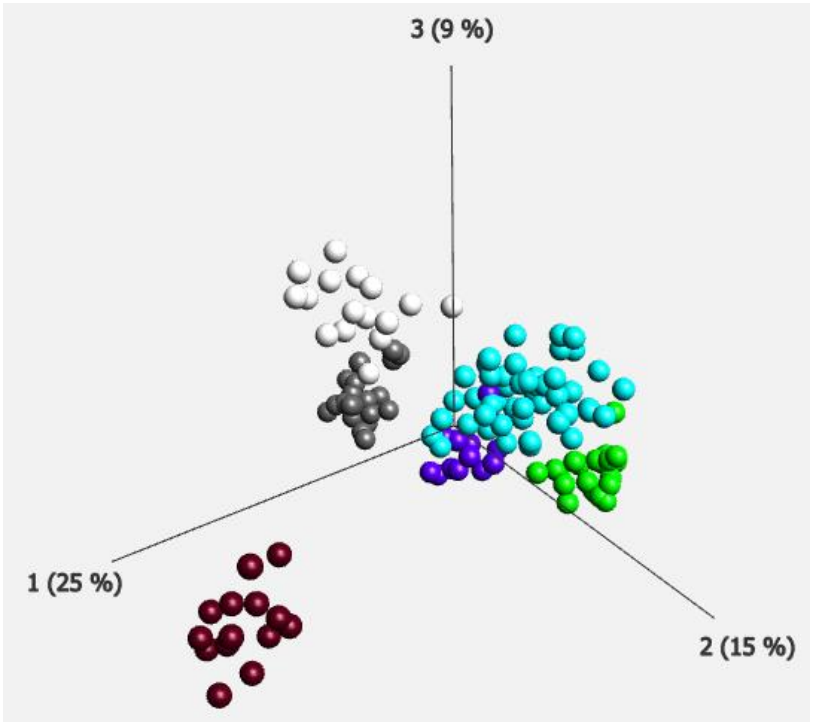
階層型クラスタリング

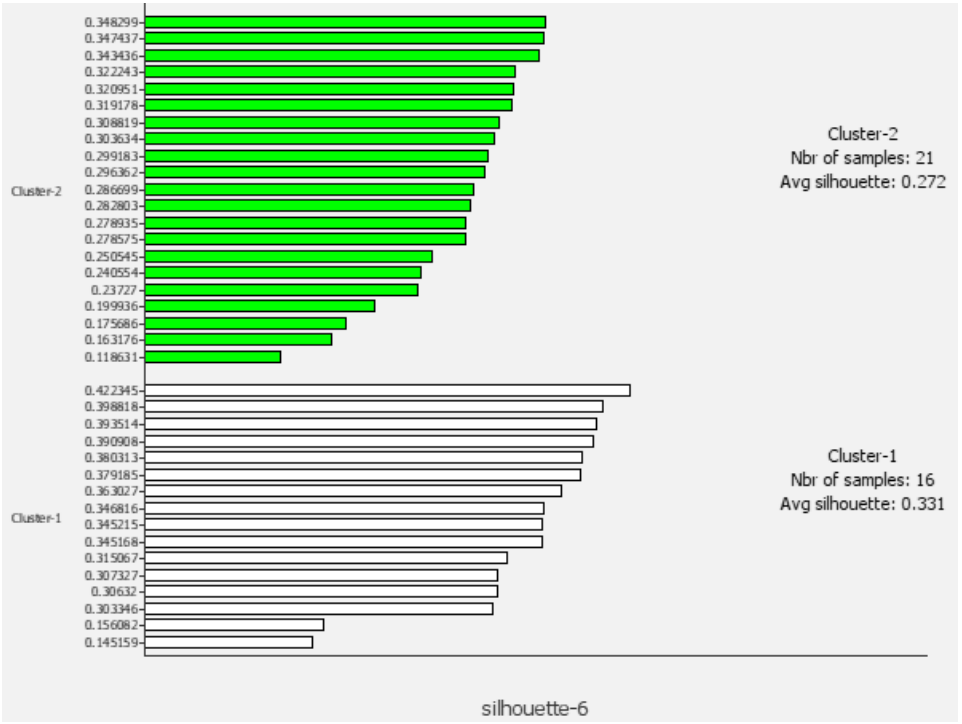


- 遺伝子発現プロファイルのヒートマップに対し、サンプル間・遺伝子間それぞれで距離が近いデータ同士を樹形図で表示
- 任意の距離で樹形図を区切ることで、クラスターを識別

K-meansクラスタリング







- クラスター数をユーザーが指定すると、サンプルを自動的に分類を行うクラスタリング手法
- クラスタリング結果はサンプルアノテーションとして保存され、主成分分析プロットなどに反映が可能
- クラスター数の妥当性の評価に用いる、シルエットグラフも出力

お問い合わせ先：フィルジェン株式会社

TEL: 052-624-4388 (9:00～18：00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp