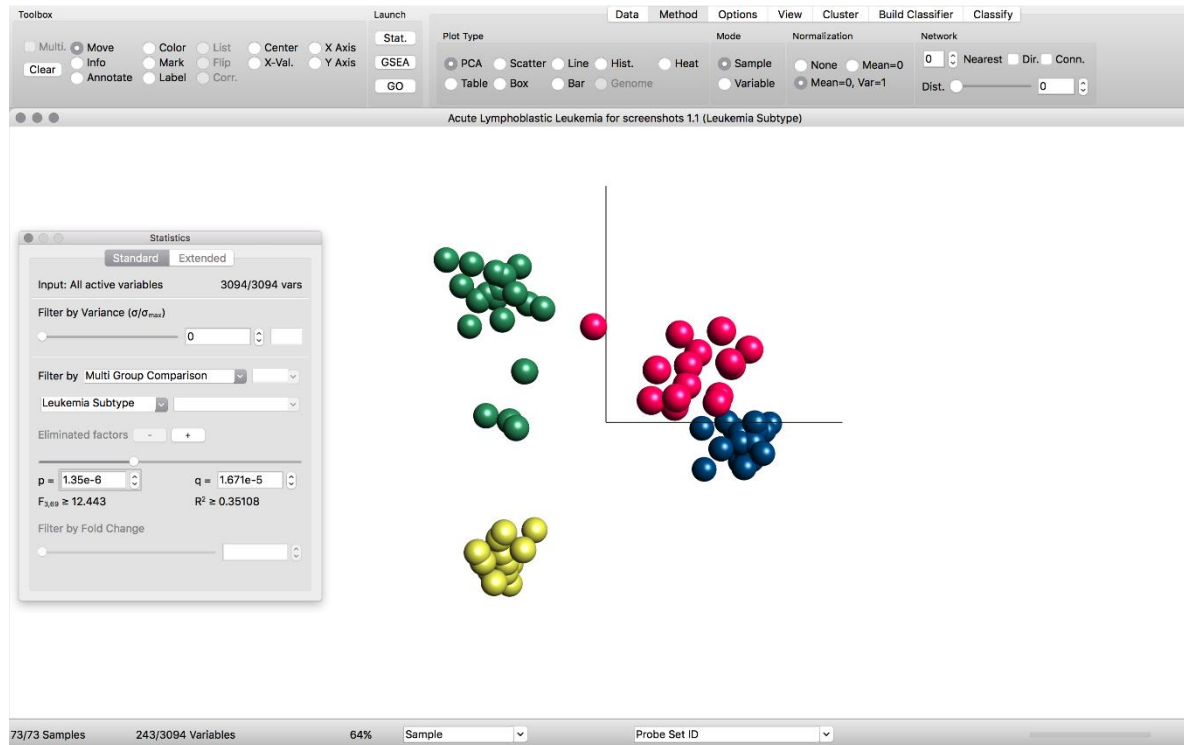


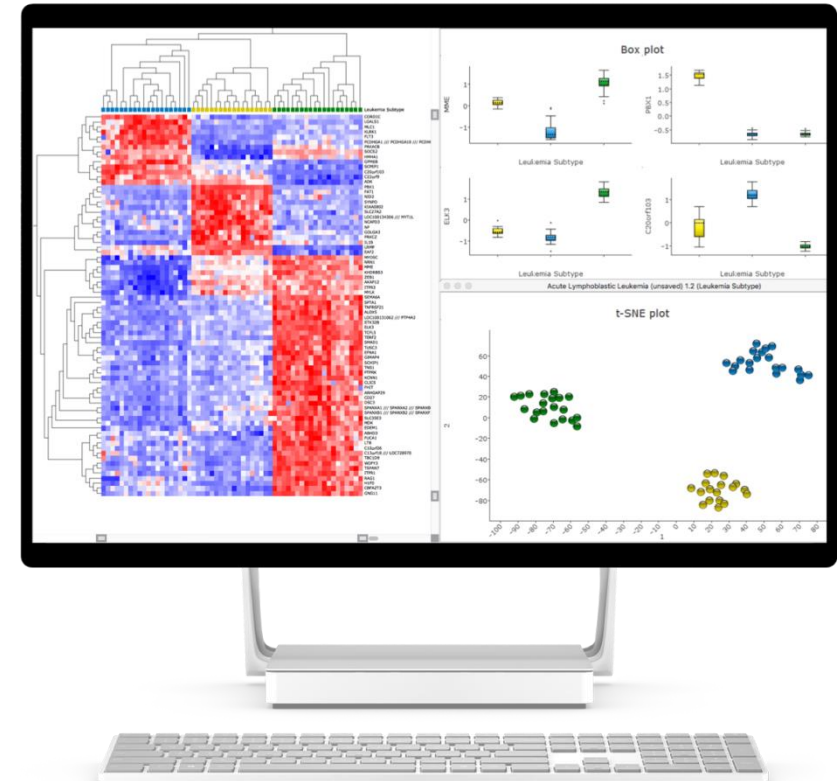
オミックスデータのクラスター解析

フィルジェン株式会社 バイオサイエンス部
(biosupport@filgen.jp)

- トランスクリプトームやプロテオームといったオミックスデータの解析において、多数の異なる性質のサンプルを同時に測定する場合、サンプルの分類や識別を行うために、クラスター解析が用いられる
- クラスター解析の実行には、専用の統計アルゴリズムに加え、データの品質チェックやビジュアライゼーションまでカバーした統合的なプラットフォームが必要となる
- Qlucore Omics Explorerは、これらクラスター解析に必要な機能を備え、また直感的な操作で、クラスター解析を効率的に実行することができる



- データを3Dプレゼンテーションとしてフルリアルタイムで操作が可能
- マウスクリックでフィルターやパラメーターの設定条件を簡単に変更し、自動でグラフに反映
- Affymetrix, Agilentマイクロアレイデータ、RNA-Seqデータ (BAMファイル & カウントデータファイル) のノーマライゼーションが可能
- 階層クラスタリングとヒートマップ、主成分解析 (PCA)、ボルケーノプロットやベン図、さらにシングルセルRNA-Seq解析用にt-SNEプロットをサポート
- 生物学的解釈を得るために、GSEA (Gene Set Enrichment Analysis) を利用可能
- サンプルの分類モデルの構築と、新サンプルへの適用を行うためのツールも搭載



対応アプリケーションエリア



Gene expression and miRNA
(Microarrays and RNA-Seq)



Proteomics



NGS



DNA Methylation



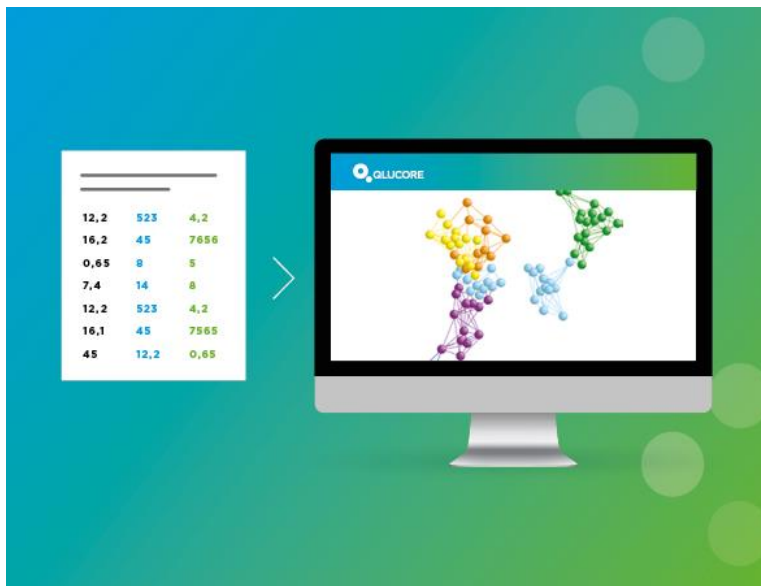
Metabolomics



Other

対応データファイル

- RNA-Seq
→ .bamファイル .txtファイル (カウントデータファイル)
- Affymetrix GeneChip WT, 3' and Clariom S Array
→ .celファイル .chpファイル
- Affymetrix GeneChip miRNA Array
→ .txtファイル
- Agilent Gene Expression Array
→ .txtファイル (Feature Extractionソフトウェア出力ファイル)
- GEO Data Set
→ .softファイル
- その他カスタムフォーマットファイル
→ .txtファイル

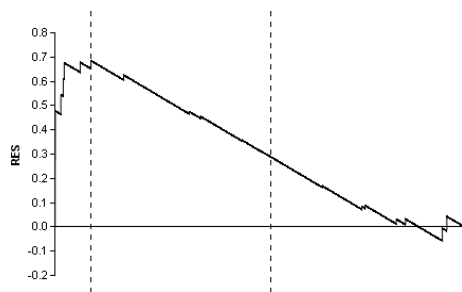
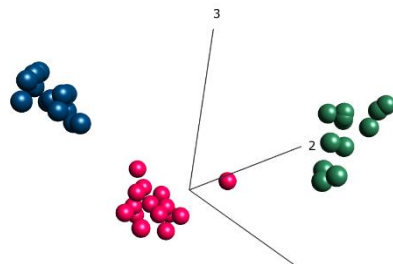


Array ID	5301	5302	5303	5304
Age	34	22	47	41
Sex	Female	Male	Male	Female

VarID	Symbol
3140	MR1
622	BDH1
7551	ZNF3

2.28	-1.23		0.45
1.04	0	-0.03	0
-0.67	3.14	2.18	0.53

Filter by Standard Deviation (s/s_{max})



- 各種オミックスデータのインポート

- ✓ マイクロアレイデータ
- ✓ RNA-Seqデータ
- ✓ プロテオミクスデータ ...など



- クオリティコントロール

- ✓ 分散フィルター
- ✓ プレフィルター
- ✓ サブサンプリング



- クラスタ解析

- ✓ 主成分分析 (PCA)
- ✓ t-SNE
- ✓ ヒートマップ・階層型クラスタリング
- ✓ K-meansクラスタリング



- 遺伝子セット解析

- ✓ GSEA

Data Import Wizard

Sample annotations

Click on sample annotation header cell(s) or click on the leftmost annotation cell(s). (Above the data matrix.)
Click again to deselect.

	1	2	3	4	5	6	7
1			Array ID	5301	5302	5303	5304
2			Age	34	22	47	41
3			Sex	Female	Male	Male	Male
4	varID	Symbol					
5	3140	MR1		2.28	-1.23	0.45	
6	622	BDH1		1.04	0	-0.03	0
7	7551	ZNF3		-0.67	3.14	2.18	0.53
8	1537	CVC1		-1.34	2.34	-0.3	2.73
9	961	CD47		2.73	1.07	0.83	-1.52

Cancel Step 8 of 8 Back Finish

Data Import Wizard

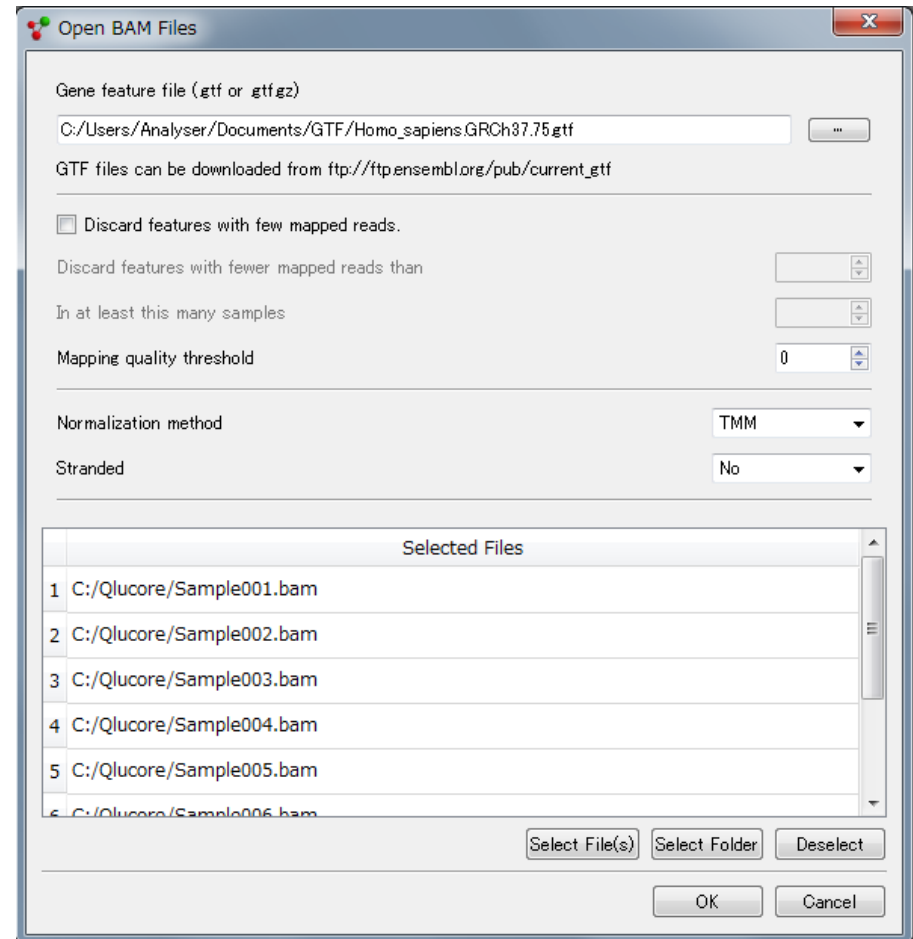
Is this a raw count matrix?

Yes No

TMM
TMM
FPKM
TPM

- ✓ タブ区切りテキストで作成した発現量データマトリックスをインポート
- ✓ RNA-Seqのリードカウントデータの場合は、データのノーマライゼーションを実行

- ✓ RNA-Seqデータの場合は、マッピングに用いたGTFファイルを用いて、BAMファイルによるインポートが可能
- ✓ リードカウントのノーマライゼーションに加え、マッピングのクオリティーやストランド設定などにも対応



Template Browser

The following templates are available in your template folders. To interact with a template, select it below.

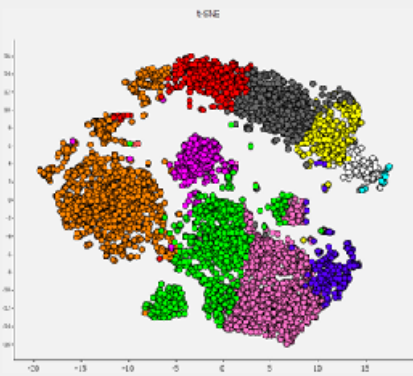
10X Genomics Assistant	Load single cell data from 10X Genomics
FTP download, load, and preprocess	Download RNA abundance data from ENCODE
Heatmap sample clustering	Plots a heatmap and orders samples by p-value
Multigroup anova	Performs an anova on a grouping annotation
t-test	Performs a t-test and displays a heatmap
t-test & fold-change	Filter by fold-change and p-value before performing a t-test
t-test (paired)	Performs a paired t-test and displays a heatmap
TCGA RSEM	Download RNA expression data from TCGA
Three plots	Displays a sample PCA plot, a heat map and a sample distance matrix

Loads data processed by 10X Genomic's CellRanger pipelines 'count' or 'aggr'.

Welcome to the loader for 10x Genomics data. Please type the directory where the data is found and click the button labeled, "OK". The directory must contain at least 3 files named, 'barcodes.tsv.gz', 'features.tsv.gz', and 'matrix.mtx.gz'. These files will be made if either the 'count' or 'aggr' pipelines was used from 10X Genomics CellRanger software. This is a beta-version of the assistant. Please consider reporting any errors of this template so that we can provide a fix.

Path

Execute Close

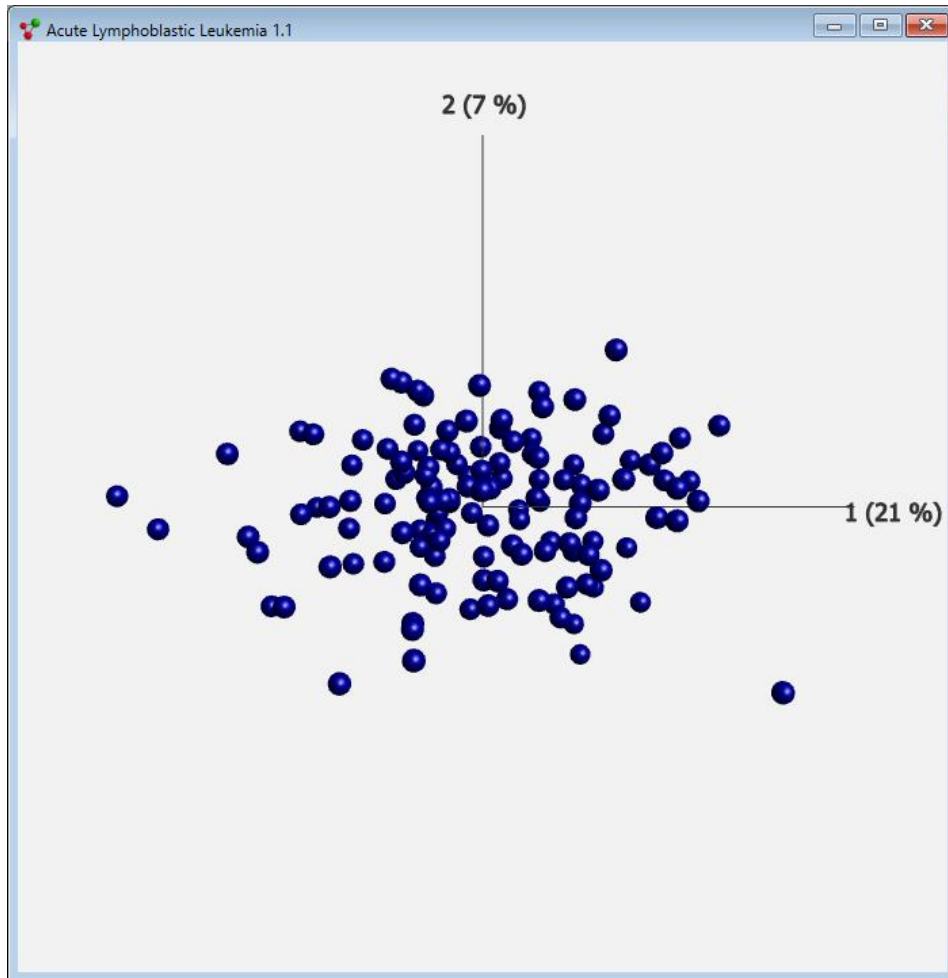


必要ファイル

- ✓ matrix.mtx.gz
- ✓ features.tsv.gz
- ✓ barcode.tsv.gz

- ✓ シングルセルRNA-Seqデータをインポートする場合は、10X Genomics Cell Rangerの解析パイプラインで作成された、ノーマライゼーション済みのファイルを指定する

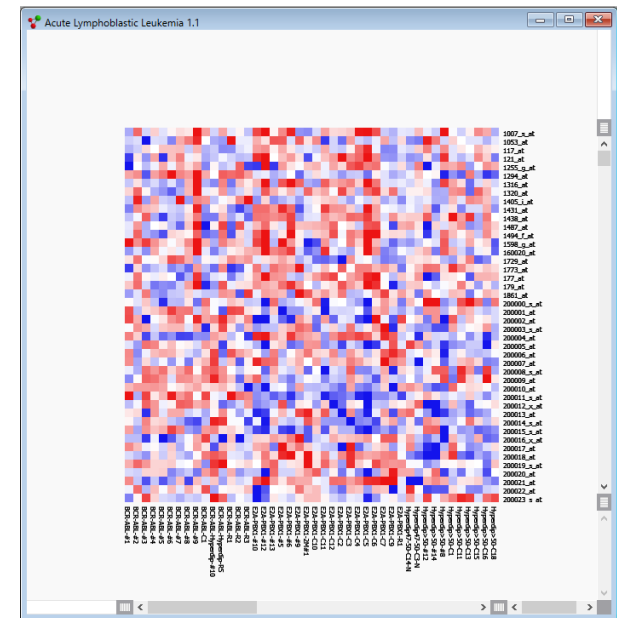
インポートされたデータの表示



Plot

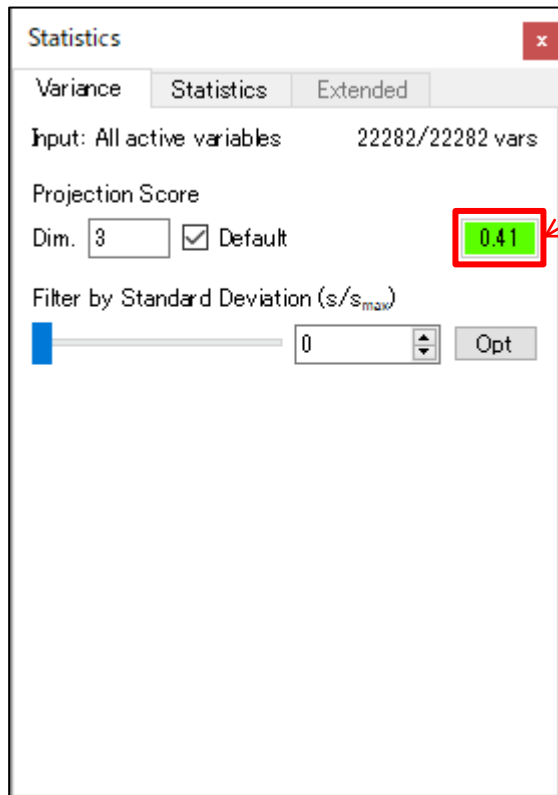
PCA Line Heat Genome

Scatter Bar Table Special



- ✓ インポートが完了すると、PCAプロット（サンプル数が多い場合はテーブルプロット）が表示される
- ✓ インポート後は、プロット形態を自由に切り替えが可能

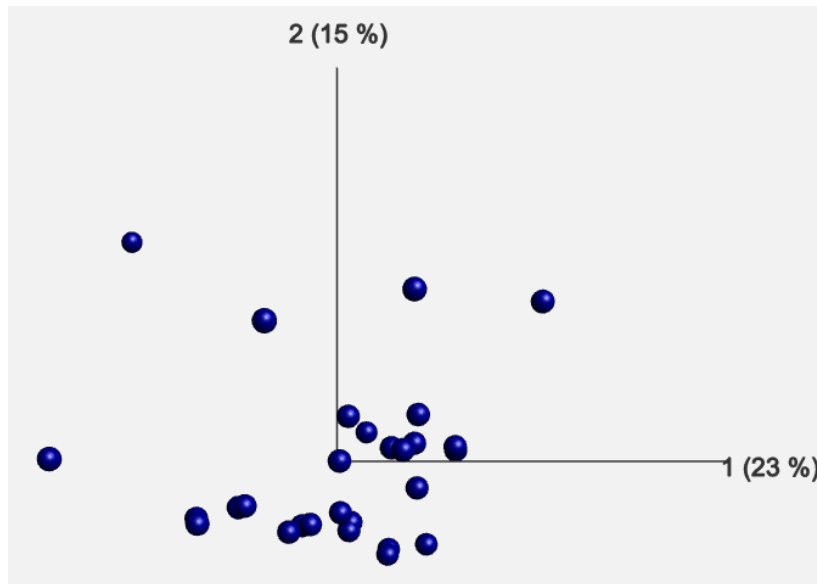
- ✓ Variance filterでは、サンプル間のばらつきが小さいデータのフィルタリングが可能
- ✓ Projection Scoreが高いフィルタリング閾値を採用することによって、閾値の選択の根拠を客観的に示すことができる



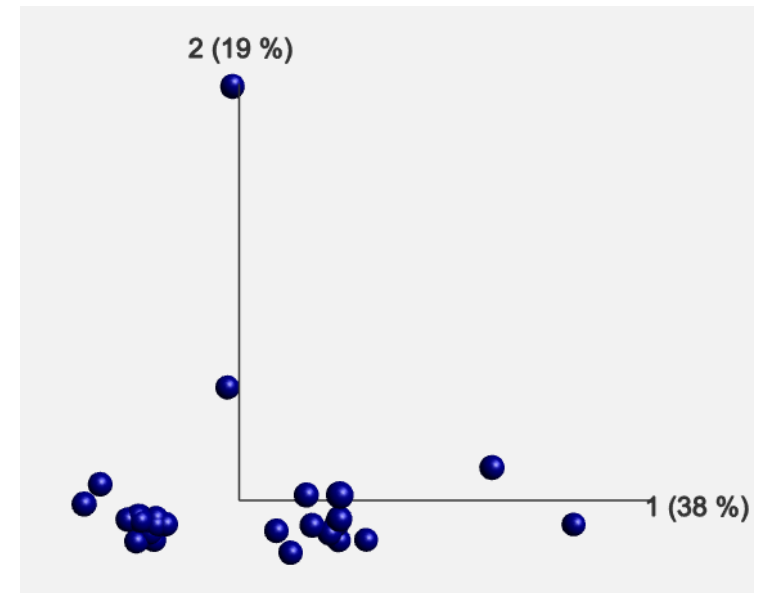
Projection Score

- ✓ PCA Plotにおいて、Variance filterでフィルタリングを実施する際に計算されるスコア
- ✓ フィルタリングによりデータの次元を減らしていった際の、データの情報性を数値化したものであり、この値が最大になるように調整することで、フィルタリングの閾値を決定できる
- ✓ スコアが高い場合は、緑色で表示される

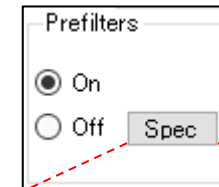
Filter by Variance ($\sigma/\sigma_{\text{max}}$)



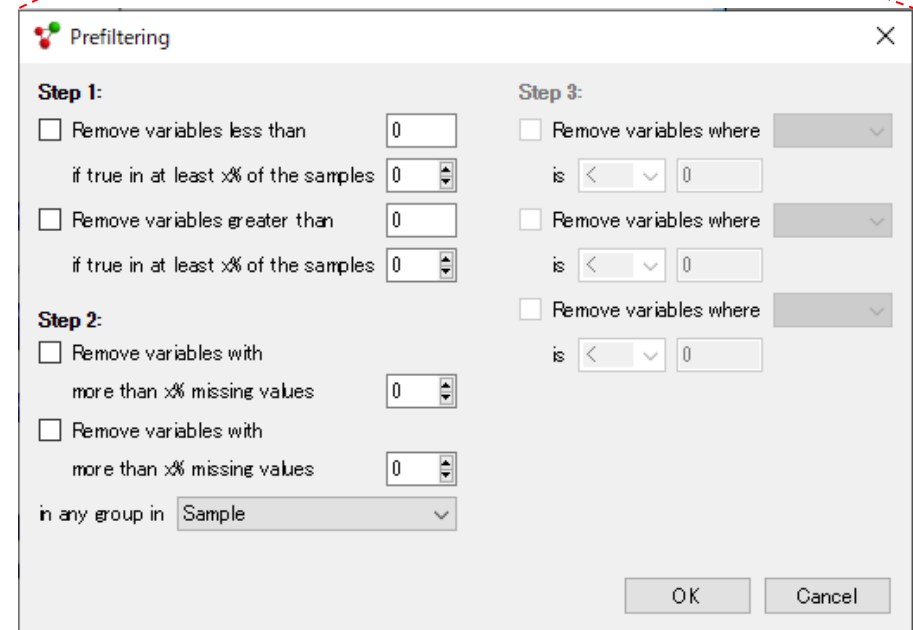
Filter by Variance ($\sigma/\sigma_{\text{max}}$)



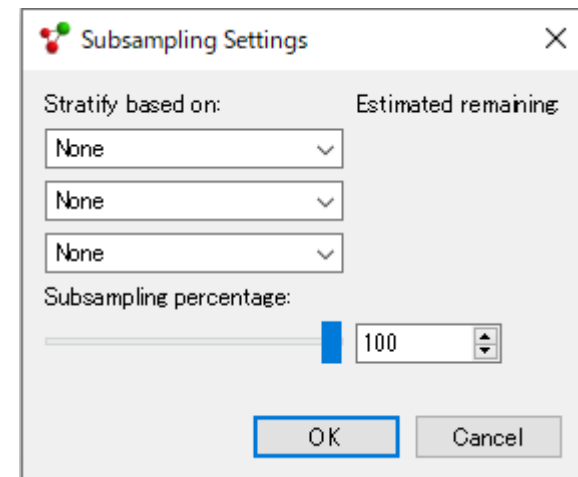
- ✓ スライダーバーを動かし、フィルタリングの閾値を変更すると、データのプロットグラフもリアルタイムで変化する

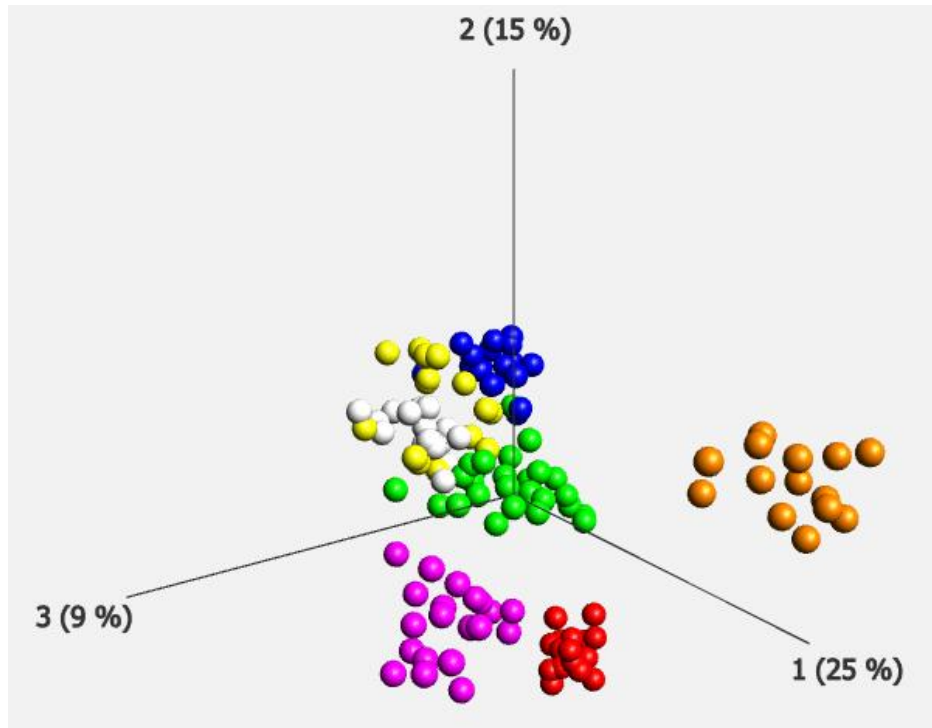


- ✓ 低クオリティデータ（ノイズデータ、空データなど）除去用のフィルター機能
- ✓ 多くのサンプルで、発現量が小さい遺伝子や、RNA-Seqでリードがカウントされなかった遺伝子などの一括除去に使用



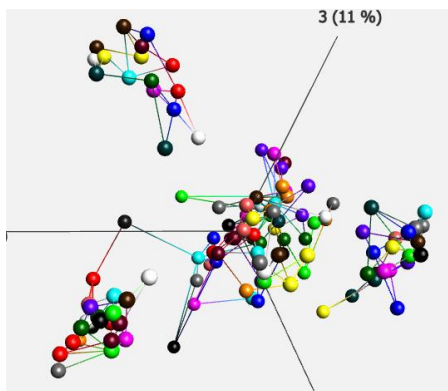
- ✓ サンプル数削減用のフィルター機能
- ✓ サンプルアノテーション（所属するグループ、クラスターなど）に基づき、アノテーションを保持したまま、代表するサンプルだけに削減したデータセットを作成が可能
- ✓ おもにシングルセルRNA-Seqなど、サンプル数が大きくコンピュータの負荷が大きい場合の、小サイズのデータセット作成に用いる



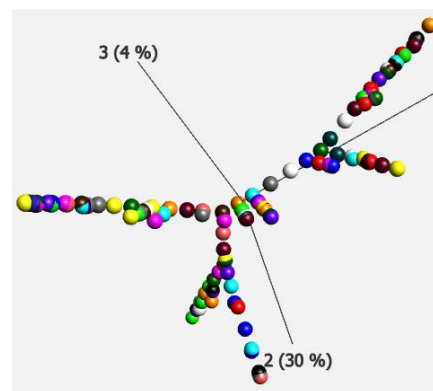


- ✓ 代表的なクラスター解析手法で、網羅的遺伝子解析データなどの多次元データを、データの特徴を保持したまま2または3次元に削減してプロット
- ✓ 分散フィルターと組み合わせて使用し、特徴的なクラスターの検出や、手動でのアノテーション付けに利用
- ✓ データプロットは、サンプルと変数（遺伝子など）の切り替えが可能

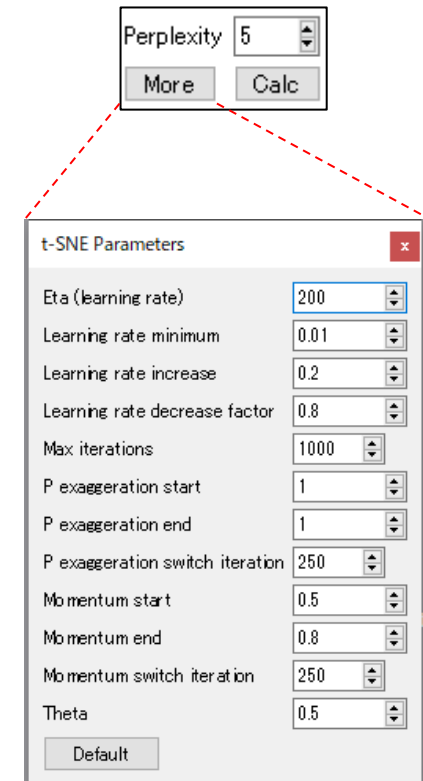
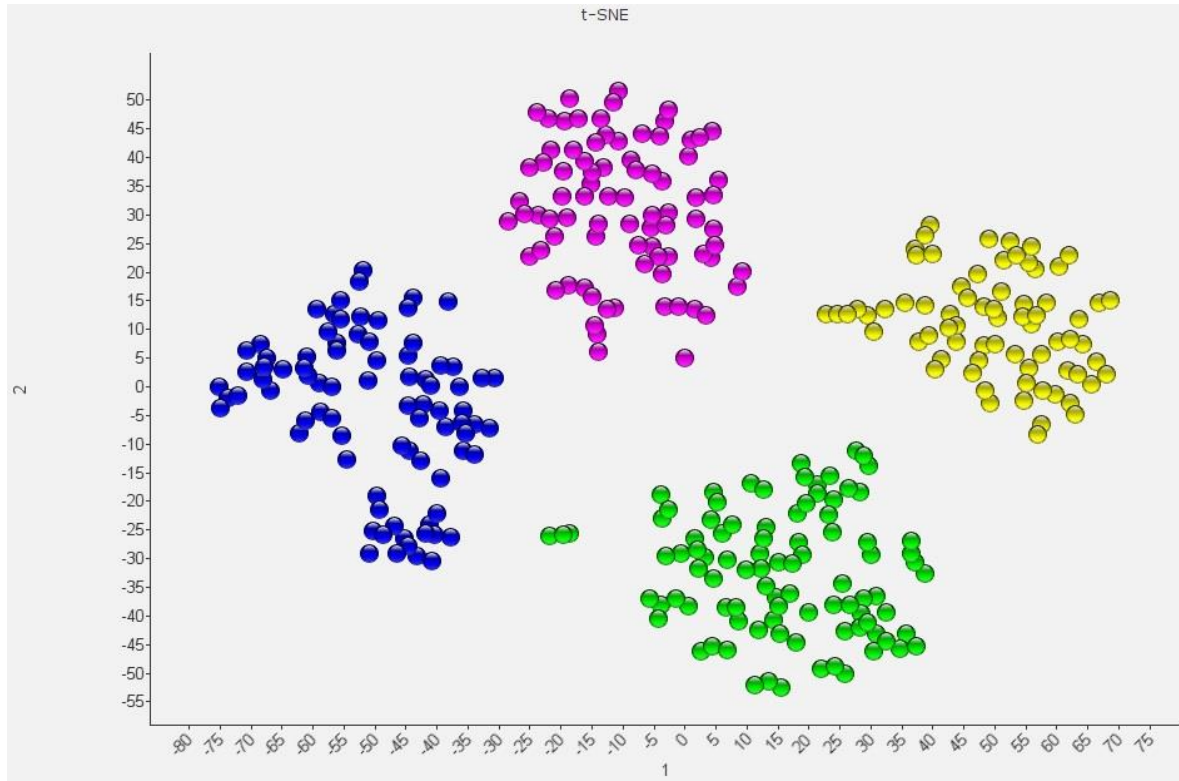
ネットワーク



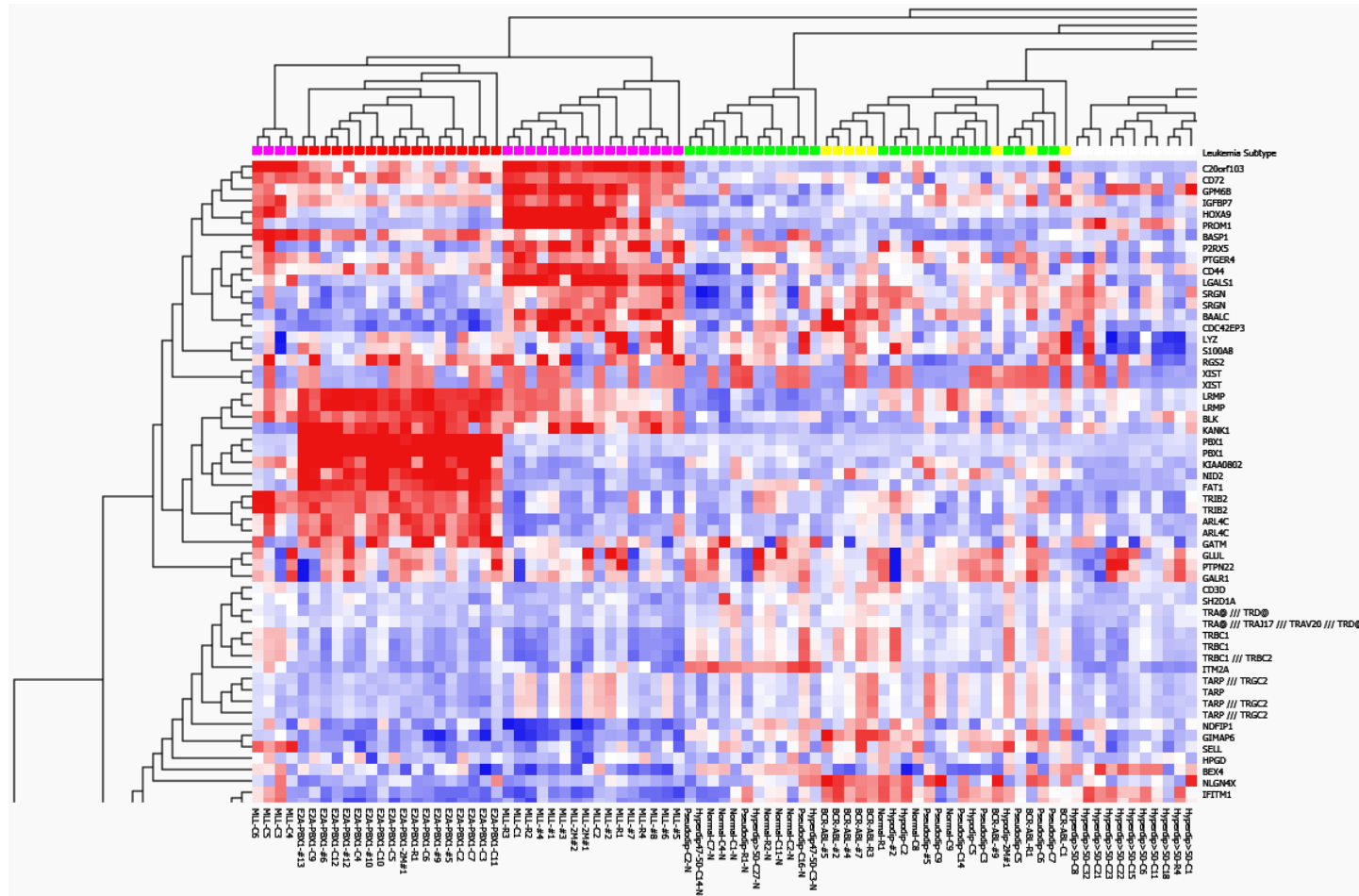
Isomap



- ✓ データ間の相関を示すネットワークや Isomapも同時に表示が可能



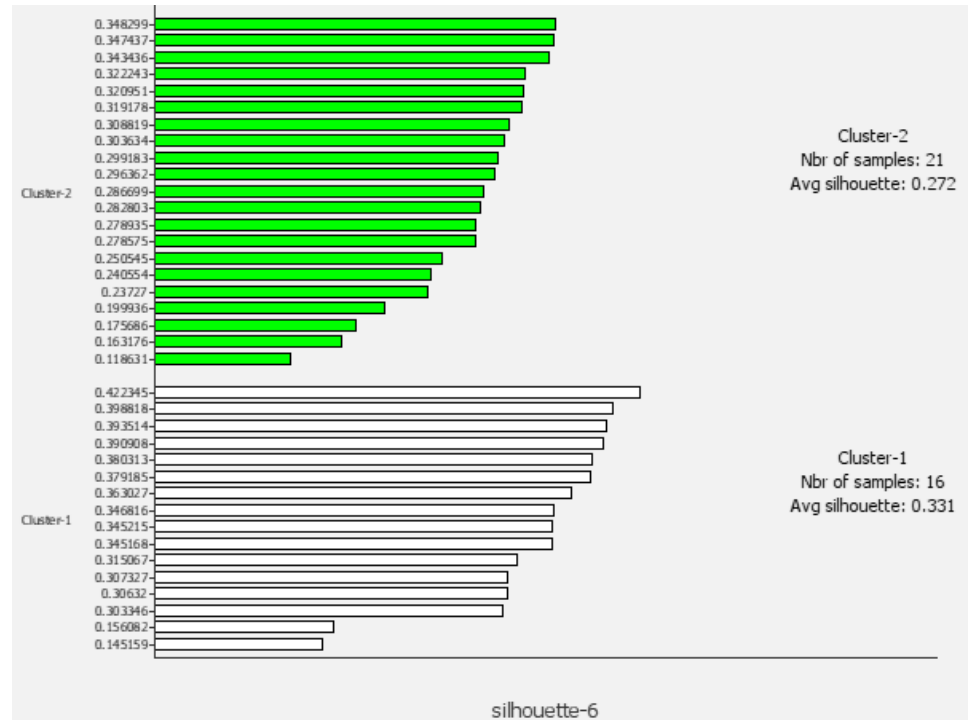
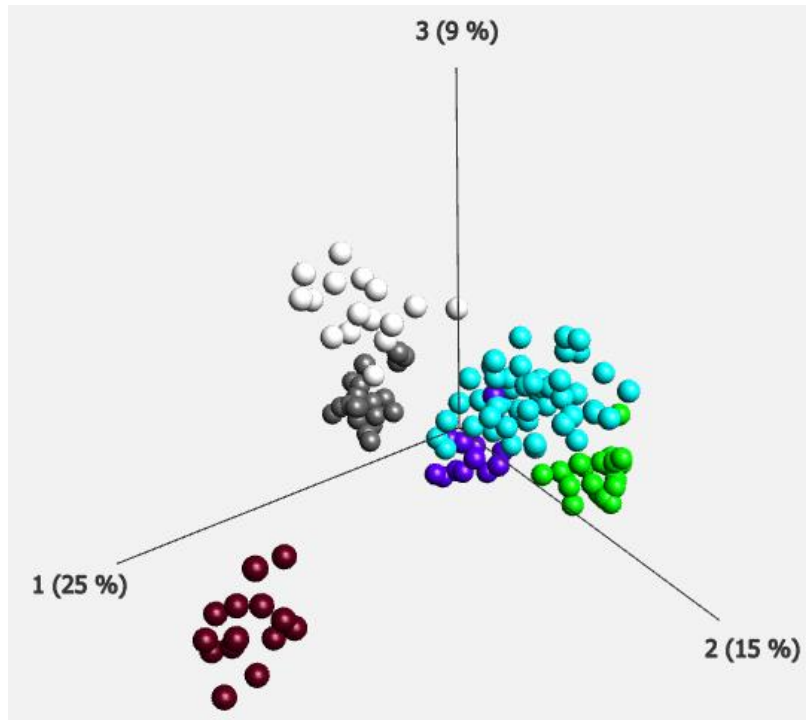
- ✓ 主成分分析と同じく次元削減の手法で、主成分分析と比べて多次元データの関係性を捉えやすい
- ✓ 主成分分析と同じく、クラスターの手動でのアノテーション付けが可能
- ✓ 計算実行時に、Perplexityなどのパラメータ調整が可能



- ✓ サンプルごとに遺伝子発現量で色付けを行ったヒートマップに対し、サンプル間・遺伝子間のそれぞれで類似したデータを樹形図で表示
- ✓ サンプルや遺伝子アノテーションに基づいたラベル表示や、遺伝子名なども同時に表示が可能で、クラスターのマーカー遺伝子の探索などに利用

Number of Clusters

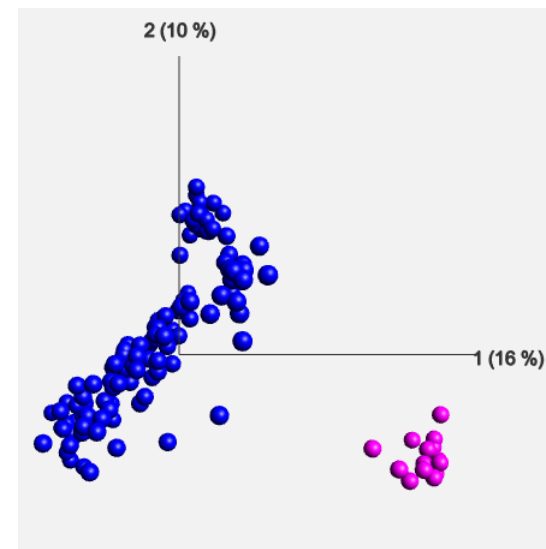
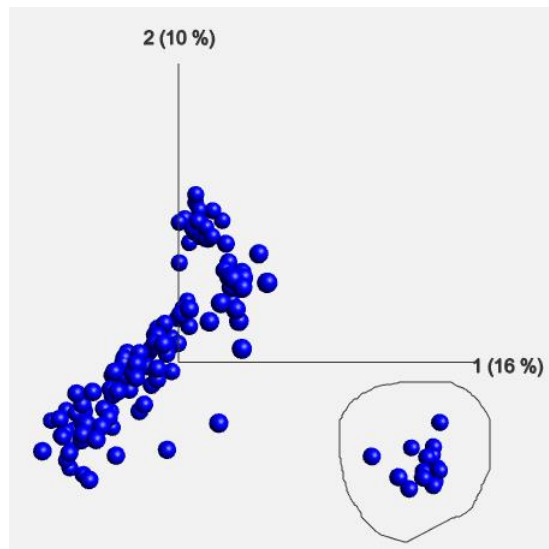
Options



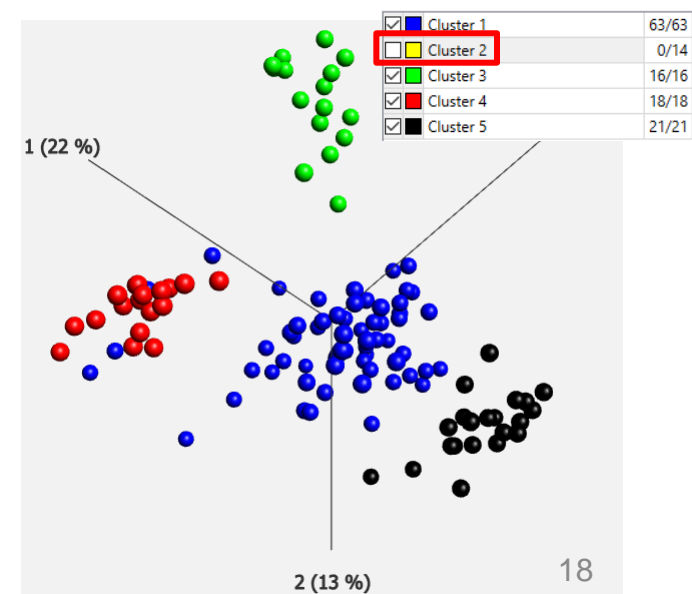
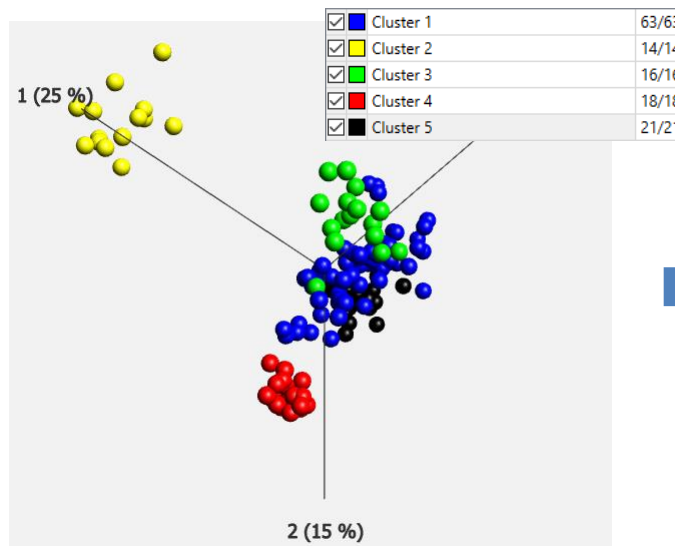
- ✓ クラスタ数ユーザーが指定すると、サンプルを自動的に分類を行うクラスタリング手法
- ✓ クラスタリング結果はサンプルアノテーションとして保存され、主成分分析プロットなどに反映が可能
- ✓ クラスタ数の妥当性の評価に用いる、シルエットグラフも出力

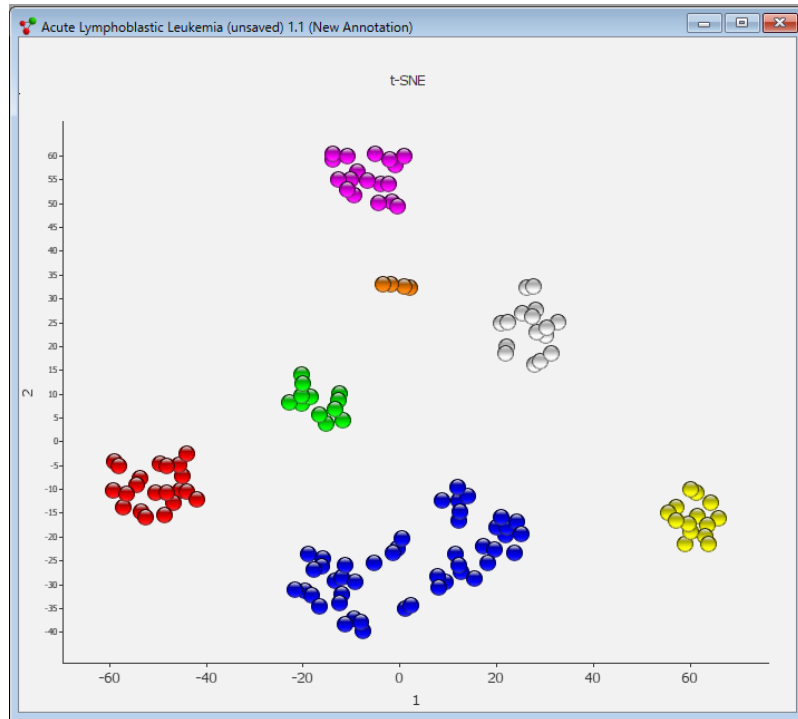
クラスターの検出

- ✓ クラスターの手動アノテーション

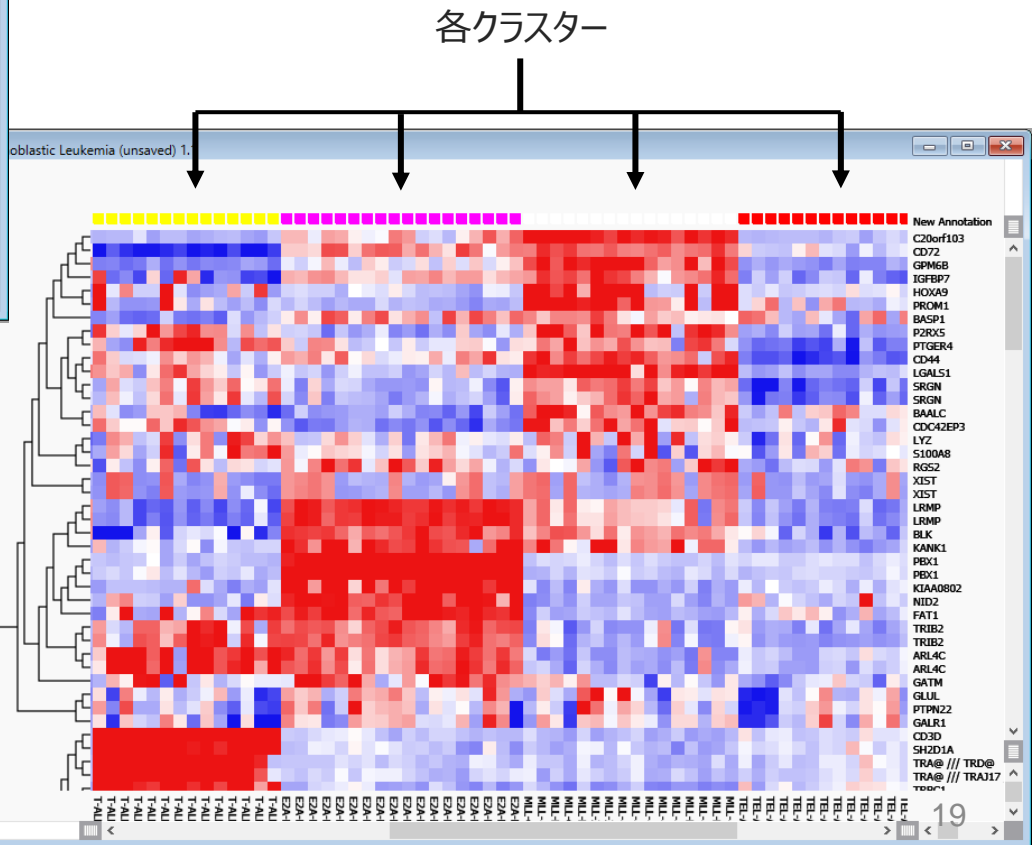


- ✓ アノテーション除去と連動したプロットの変化





- ✓ 主成分分析プロットやt-SNEプロットなどでクラスターのアノテーション付け後に、ヒートマップ表示に切り替え、クラスターに特徴的な遺伝子を確認



Statistics

Variance Statistics Extended

Input: Variables after variance filter: 207/22282 va

Filter by: Two Group Comparison \neq

New Annotation: v Value (2)

Eliminated factors: - +

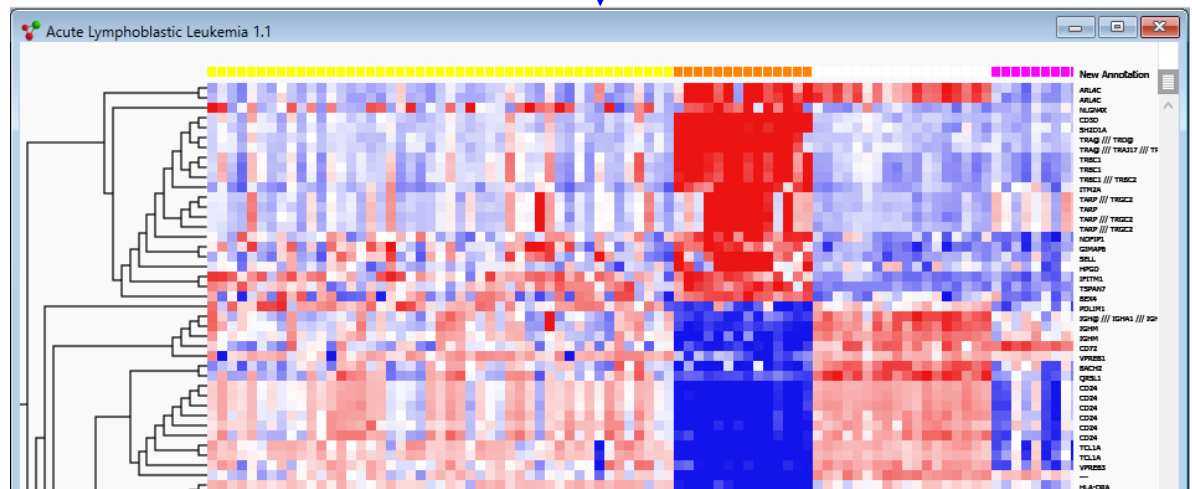
$p = 5.4e-6$ $q = 1.1557e-5$

$|t_{100}| \geq 4.7455$ $|R| \geq 0.38425$

Filter by: Fold Change

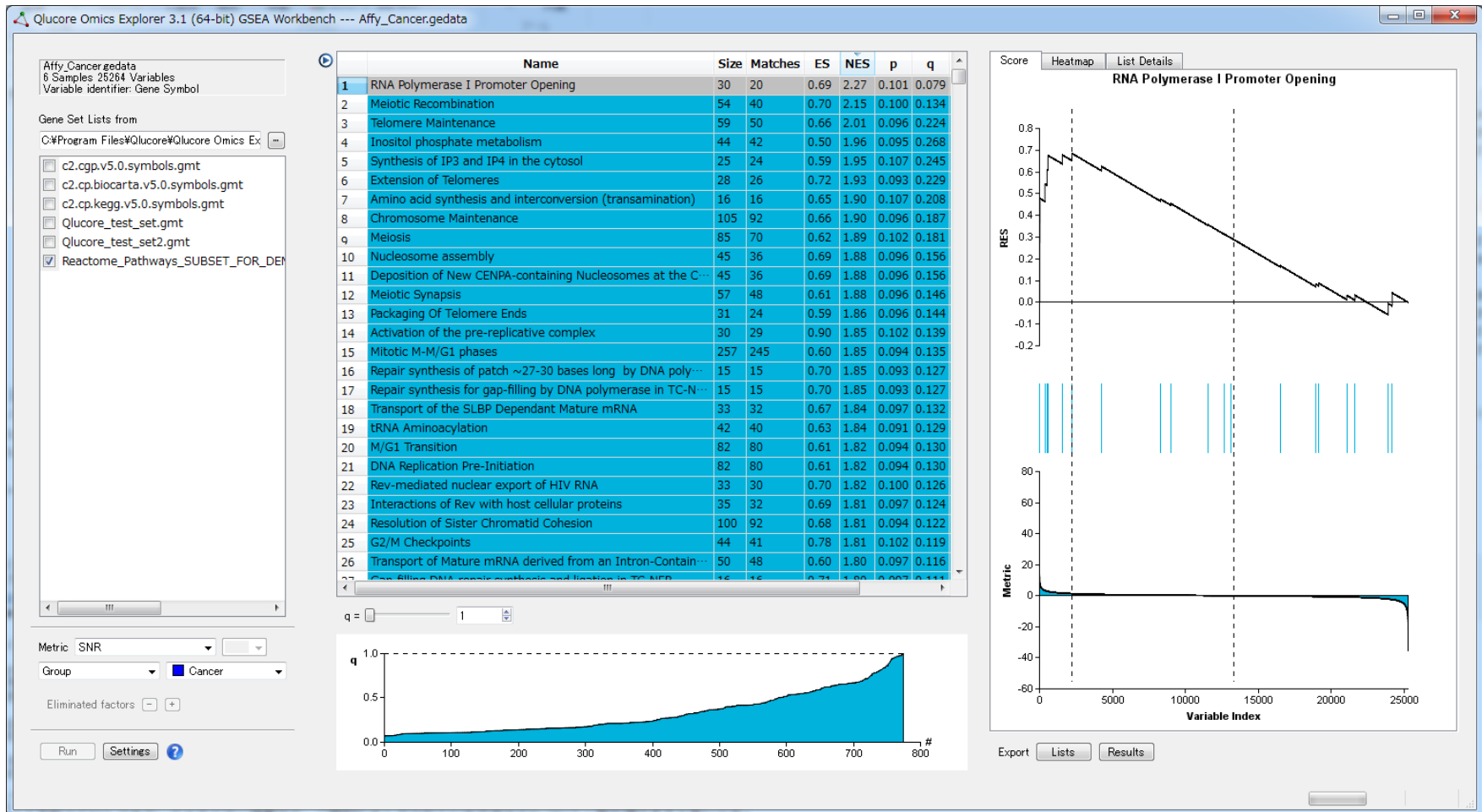
2

	p-values	q-values	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#	BCR-ABL-#
201005_at	2.096e-12	7.232e-12	5.3569	10.773	7.6947	10.848	9.3365	7.1835	10.216	10.118	9.6027	11.043	10.482		
201137_s_s	2.18e-33	2.257e-32	11.189	10.574	11.284	10.448	11.075	10.357	10.052	10.604	10.956	10.167	10.505		
202206_at	2.495e-14	9.934e-14	6.9966	5.3971	4.2979	5.2629	4.8614	4.5577	5.8244	4.1983	5.1734	5.4414	4.673		
202207_at	2.049e-14	8.318e-14	7.7239	6.2377	5.0618	6.2811	5.6602	5.403	6.7949	4.9388	5.1479	6.6057	5.3061		
202242_at	5.39e-10	1.691e-09	6.8589	7.8751	8.6288	7.8781	7.0958	8.2654	7.6067	8.0694	7.9298	5.0131	8.4911		
202746_at	6.211e-18	3.136e-17	5.5843	5.4445	2.985	4.674	4.4641	6.1527	5.8745	3.1528	4.5978	5.2681	3.7312		
203066_at	2.361e-17	1.137e-16	6.5014	8.5312	9.9243	9.0105	8.9287	9.3676	8.3766	9.0063	8.188	10.21	10.153		
203372_s_s	9.116e-11	3.044e-10	9.2274	10.297	9.7628	10.147	10.105	10.365	10.036	9.7453	9.6273	9.9406	10.027		
203373_at	1.087e-10	3.57e-10	9.9676	10.773	10.247	10.54	10.528	10.902	10.475	10.062	10.28	10.671	10.676		
203434_s_s	1.616e-06	4.289e-06	6.3363	9.6197	8.2477	9.7004	9.7824	10.558	9.7804	8.5474	10.115	9.6917	9.5521		
203435_s_s	1.815e-06	4.757e-06	7.1953	10.329	8.6064	10.368	10.444	10.877	10.33	8.8949	10.77	10.273	10.497		
203611_at	3.08e-06	7.97e-06	6.5156	10.492	9.005	10.721	10.143	8.7597	9.3802	9.7426	10.525	7.2123	10.773		
203753_at	2.62e-14	1.023e-13	6.7418	5.2151	7.1101	5.9468	6.8807	6.0462	5.7689	5.8944	4.8689	6.2794	6.9774		
203913_s_s	6.849e-13	2.403e-12	2.8237	2.5721	0.66544	2.6628	1.5027	1.9133	1.6187	3.2018	1.0257	2.8813	0.86808		
204563_at	2.074e-10	6.708e-10	10.72	6.8398	5.2376	6.9178	8.3462	8.5857	9.1828	5.7661	8.0291	7.3316	6.3889		
204670_x_s	1.065e-37	1.836e-36	11.562	11.1	11.347	10.974	10.578	10.903	11.115	10.734	9.597	9.9501	10.875		



- Two Group Comparison (t-test)
- Multi Group Comparison (ANOVA)

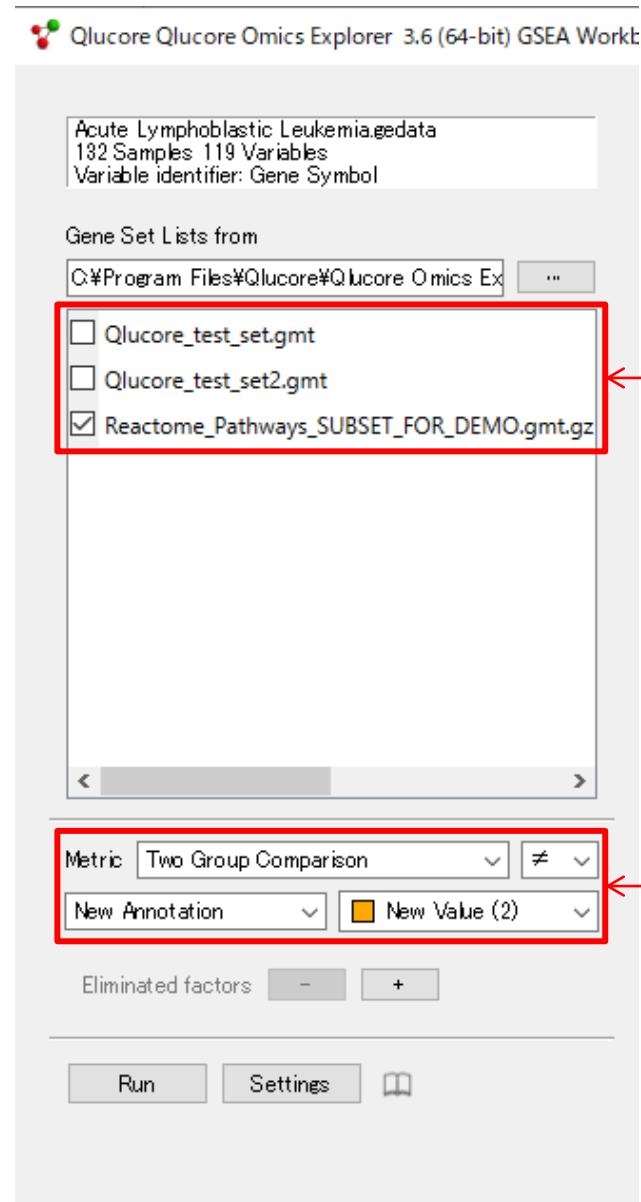
✓ 統計フィルターを用いた有意差検定により、クラスターに特徴的な遺伝子の検定も可能



- ✓ 遺伝子発現データに対し、発現変動の大きい順に遺伝子をランク付けし、ランクの高い遺伝子が、調べたい遺伝子セットにどれだけ多く含まれているか検定を行う
- ✓ 遺伝子セットには、Gene Ontologyやパスウェイデータベースなどを使用

- ✓ 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

- ✓ 有意差検定のP-valueなど、遺伝子のランク付け条件を指定



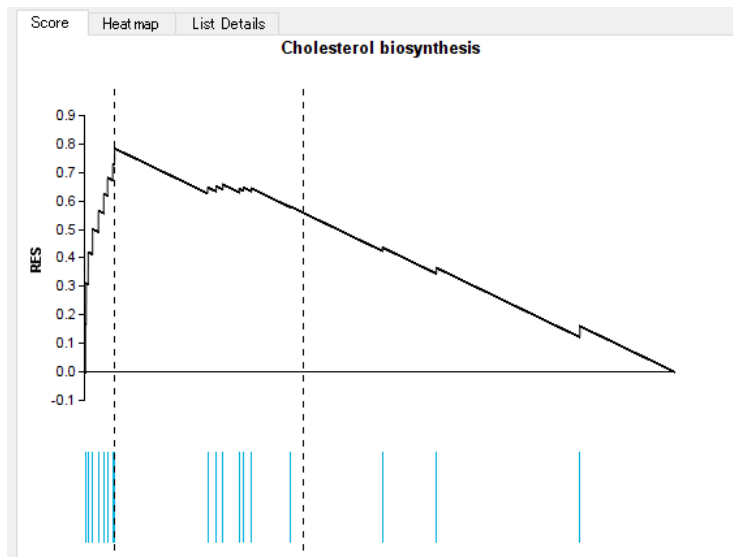
遺伝子セットデータベース
(.gmtファイル)

遺伝子のランク付け条件

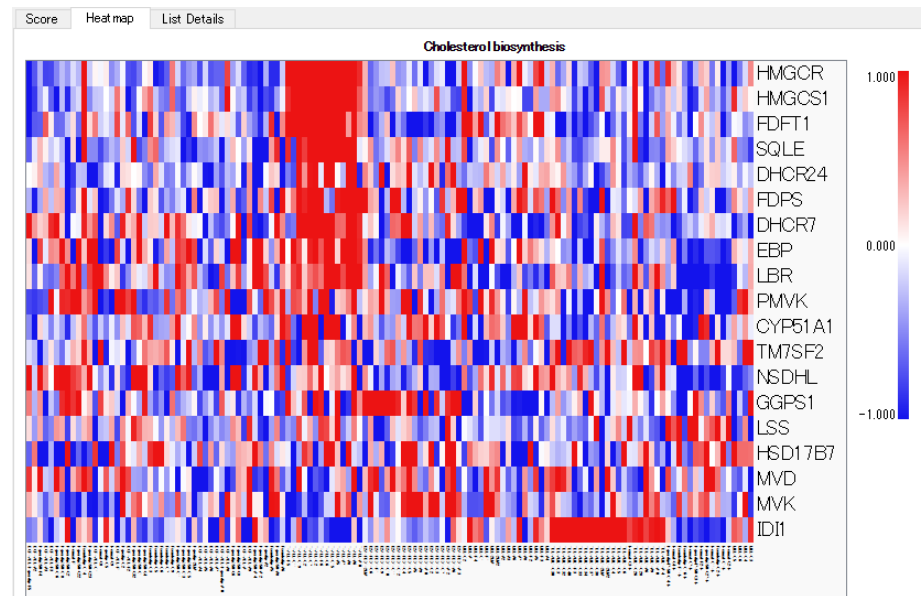
	Name	Size	Matche:	ES	NES	p	q
1	Cholesterol biosynthesis	22	19	0.78	2.02	0	0.095513
2	Activation of Gene Expression by ...	41	38	0.56	1.86	0.0037951	0.35581
3	Regulation of Cholesterol ...	54	51	0.57	1.86	0.003937	0.23861
4	Effects of PIP2 hydrolysis	25	22	0.57	1.84	0.002004	0.22566
5	TCR signaling	63	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

- ✓ 実行結果として、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

遺伝子セットリスト



エンリッチメントスコア



ヒートマップ

お問い合わせ先：フィルジエン株式会社

TEL: 052-624-4388 (9:00～18 : 00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp