

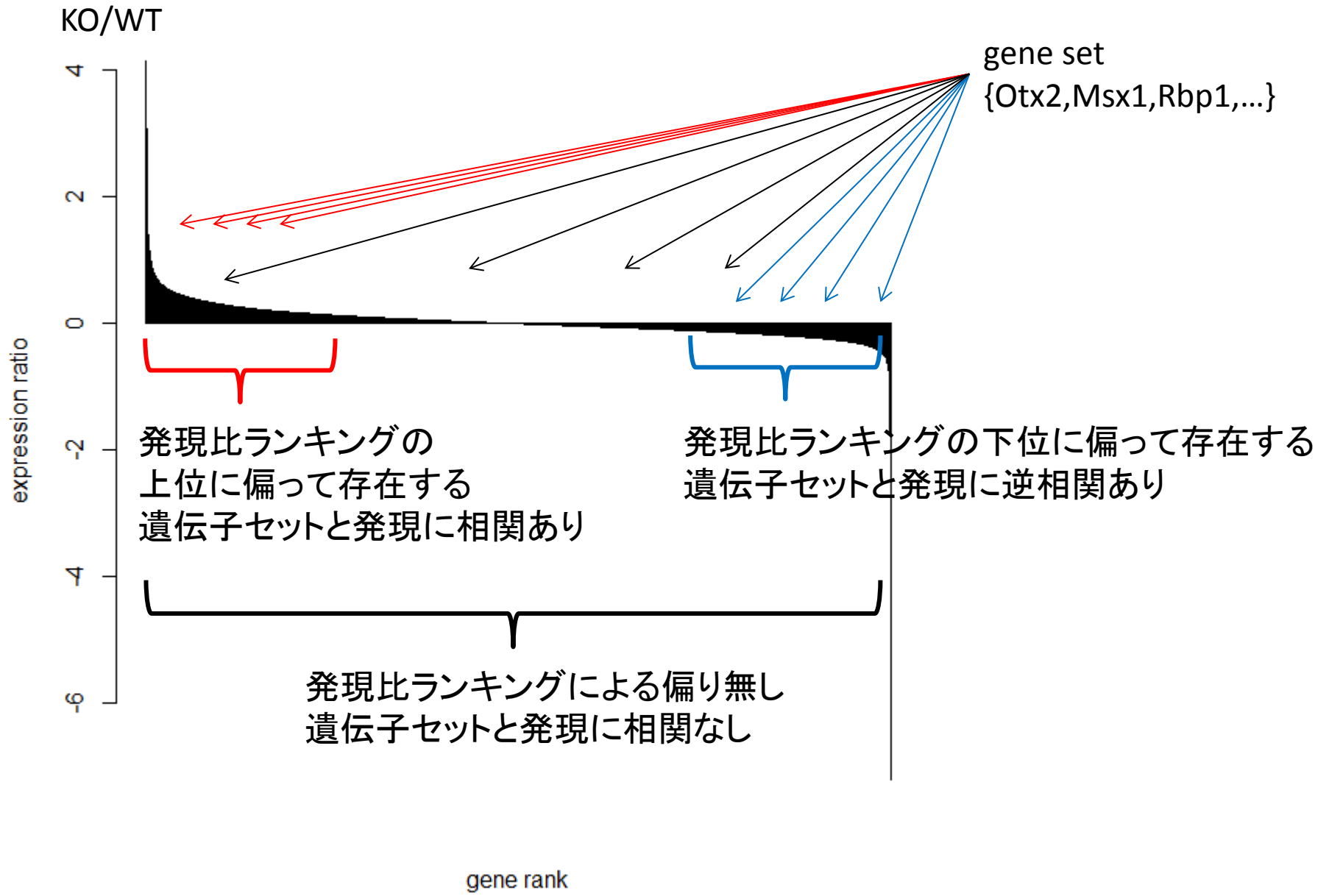


第3回バイオインフォマティクス実習コース
横浜市大 先端医科学研究センター
バイオインフォマティクス研究室

室長 田村智彦
准教授 中林潤
免疫学 藩龍馬

- Gene Set Enrichment Analysis

- Gene Set Enrichment Analysis (GSEA)
- 特定の遺伝子セットと発現比の間に相関があるか調べる



Enrichment Score

$$P_{hit}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{\sum_{g_j \in S} |r_j|^p}$$

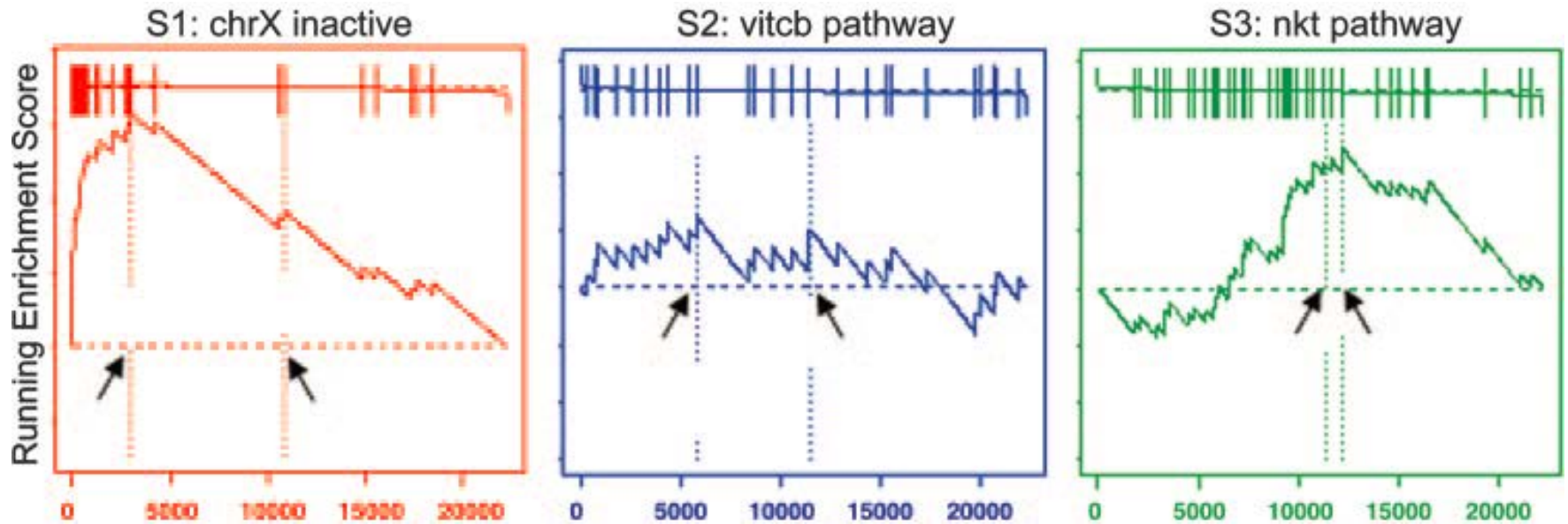
r_j 発現比
 p 重みづけ

$$P_{miss}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)}$$

発現比ランクの高い順から遺伝子を調べ、遺伝子リスト中に該当する遺伝子が存在したらenrichment scoreを加算、なければ減算する。

正の相関

相関なし



ENRICHMENT SCOREの最大値と位置から相関の有り無しを判定する

Subramanian A *et al.* PNAS 2005

http://www.broadinstitute.org/gsea/index.jsp

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

23-Jan-2014: Version 2.0.14 of the GSEA desktop application is now available, which contains a number of upgrades and bug fixes. See the [GSEA v2.0.14 Release Notes](#) for details.

05-Jun-2013: Version 4.0 of the Molecular Signatures Database (MSigDB) is now available, which includes a new gene set collection (C7) of 1,910 immunologic signatures generated as part of the Human Immunology Project Consortium. We also released a newer version (2.0.13) of the GSEA desktop application. There were no changes to 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).

Diagram: Molecular Profile Data and Gene Set Database feed into Run GSEA, which produces Enriched Sets.

DownloadセクションからGSEAを取得 Javaプログラム(OSに依存しない) メールアドレスを登録する必要あり

The screenshot shows the GSEA website interface. At the top, there's a navigation bar with links for GSEA Home, Downloads, Molecular Signatures Database, Documentation, and Contact. The 'Downloads' link is highlighted. Below the navigation bar, there's an 'Overview' section with a sub-section 'What's New' containing two entries: one from 2014 and one from 2013. To the right, a diagram illustrates the workflow: 'Molecular Profile Data' and 'Gene Set Database' feed into a 'Run GSEA' box, which outputs 'Enriched Sets' (a graph). Below the diagram, there are sections for 'Registration' (requiring email registration) and 'Contributors' (listing funding agencies). At the bottom, there's a 'Citing GSEA' section with a reference to Subramanian et al. (2005) and Mootha et al. (2003). The browser's taskbar at the bottom shows various application icons and the system clock indicating 23:21 on 2015/01/25.

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

23-Jan-2014: Version 2.0.14 of the GSEA desktop application is now available, which contains a number of upgrades and bug fixes. See the [GSEA v2.0.14 Release Notes](#) for details.

05-Jun-2013: Version 4.0 of the Molecular Signatures Database (MSigDB) is now available, which includes a new gene set collection (C7) of 1,910 immunologic signatures generated as part of the Human Immunology Project Consortium. We also released a newer version (2.0.13) of the GSEA desktop application. There were no changes to 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).

Download

The screenshot shows the GSEA Downloads page. The browser address bar is www.broadinstitute.org/gsea/downloads.jsp. The page title is "Downloads". Below the title, there is a paragraph: "The GSEA software and source code and the Molecular Signatures Database (MSigDB) are freely available to individuals in both academia and industry for internal research purposes. Please see the [GSEA/MSigDB license](#) for more details."

Under the "Software" section, there are four options:

Software Option	Description	Download/Action
javaGSEA Desktop Application	<ul style="list-style-type: none">Easy-to-use graphical user interfaceRuns on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java 6 or 7Produces richly annotated reports of enrichment resultsIntegrated gene sets browser to view gene set annotations, search for gene sets and map gene sets between platforms	Launch with 1GB (for 32 or 64-bit Java) memory: Launch
javaGSEA Java Jar file	<ul style="list-style-type: none">Command line usageRuns on any platform that supports Java 6 or 7We recommend using the 'Launch' buttons above instead of this mode for most users	download gsea2-2.1.0.jar
GSEA Java Source Code Java source files	<ul style="list-style-type: none">100% Java implementation of GSEAIncorporate GSEA into your own data analysis pipelineProgrammatically call the open source GSEA java API	download gsea2_distrib-2.1.0.zip
R-GSEA R Script	<ul style="list-style-type: none">Usage from within the R programming environmentEasily inspect, learn and tweak the algorithmIncorporate GSEA into your own data analysis pipelineProgrammatically call the open source GSEA R APIClick here to learn more about the R-GSEA script	download GSEA-P-R.1.0.zip

Annotations on the screenshot:

- An arrow points from the text "JNLPファイル その都度プログラムをダウンロードして実行する" to the "Launch" button in the "javaGSEA Desktop Application" row.
- An arrow points from the text "Java実行ファイル" to the "download gsea2-2.1.0.jar" link in the "javaGSEA Java Jar file" row.

JNLPファイル
その都度プログラムを
ダウンロードして実行
する

Java実行ファイル

- 課題配布フォルダからgsea2-2.1.0を各自のデスクトップにコピー
- gsea2-2.1.0をダブルクリック

GSEA

GSEA v2.1.0 (Gene set enrichment analysis -- Broad Institute)

File Options Downloads Tools Help

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis
- Enrichment Map Visualization

Gene set tools

- Chip2Chip mapping
- Browse MSigDB

Analysis history

GSEA reports



Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Home

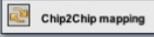
Steps in GSEA

- 1. What you need for GSEA**
 - Expression data set
 - Phenotype annotation
 - Gene sets – use MSigDB or your own gene sets
- 2. Run GSEA**
 - Start with default parameters
 - If you want to collapse probes to genes, specify chip platform
- 3. View results**
- 4. Leading edge analysis**
 - Leading edge finds genes driving enrichment results

Gene Set Tools

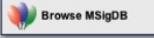
Chip2Chip mapping

- Convert gene sets between platforms



Explore MSigDB gene sets

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets



See also


- MSigDB online tools at: www.broadinstitute.org/msigdb

Getting Help

GSEA web site:
www.broadinstitute.org/gsea

GSEA documentation:
www.broadinstitute.org/gsea/wiki

Email the GSEA team:
gsea@broadinstitute.org



23:29:55

30M of 44M

23:29
2015/01/25

データファイルをload

GSEA v2.1.0 (Gene set enrichment analysis -- Broad Institute)

File Options Downloads Tools Help

Steps in GSEA analysis

Load data

Run GSEA

Leading edge analysis

Enrichment Map Visualization

Gene set tools

Chip2Chip mapping

Browse MSigDB

Analysis history

GSEA reports



Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Home

Steps in GSEA

- 1. What you need for GSEA**
 - Expression data set
 - Phenotype annotation
 - Gene sets – use MSigDB or your own gene sets
- 2. Run GSEA**
 - Start with default parameters
 - If you want to collapse probes to genes, specify chip platform
- 3. View results**
- 4. Leading edge analysis**
 - Leading edge finds genes driving enrichment results

Gene Set Tools

Chip2Chip mapping

- Convert gene sets between platforms

Chip2Chip mapping

Explore MSigDB gene sets

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

Browse MSigDB

See also


- MSigDB online tools at: www.broadinstitute.org/msigdb

Getting Help

GSEA web site:
www.broadinstitute.org/gsea

GSEA documentation:
www.broadinstitute.org/gsea/wiki

Email the GSEA team:
gsea@broadinstitute.org



23:29:55

30M of 44M

23:29
2015/01/25

データファイルのload

- 必要なファイルは3つ
- 発現プロファイル gctファイル
- 遺伝子セット grpファイル
- カテゴリー clsファイル

gctファイル

常に必要

遺伝子数

サンプル数

#1.2

21530	4				
NAME	Description	KO1	KO2	WT1	WT2
Ctss	NA	1730.1	1681.1	10.2	10.5
Ahnak	NA	1650.3	1510.1	11.3	14.2
...

常に必要

遺伝子名
大文字、小文字の区別に注意

ファイル名の拡張子はgct

grpファイル

#gene symbol
Evi1
Myct1
...

遺伝子名の羅列

gctファイルと大文字、小文字を一致させる
ファイル名の拡張子はgrp

clsファイル

サンプル数
クラス数
常に必要

4 2 1
#KO WT
KO KO WT WT

clsファイルはスペース区切りのテキストファイル
拡張子はcls

- 課題配布フォルダから
- GSE40493_Normalized_GSEA_UPPER.gct
- geneset_Bcl6.grp,geneset_BRAIN.grp
- GSE40493_Class.cls
- 各ファイルを各自のデスクトップフォルダへコピー

Load Data

Browse for filesをクリックしてファイルを選択

The screenshot shows the GSEA v2.1.0 application window. The 'Load data' window is active, displaying three methods for loading data. Method 1 includes a 'Browse for files ...' button, which is highlighted by an arrow from the Japanese text above. Method 2 has a 'Load last dataset used' button. Method 3 is a large empty area for drag-and-drop. The right side of the window lists supported file formats: Dataset (res or gct, pcl, txt), Phenotype labels (cls), and Gene sets (gmx or gmt). Below the main window, there are sections for 'Recently used files' and 'Object cache'. The 'Recently used files' list contains various files with icons indicating their type (e.g., class files, gct files, grp files). The 'Object cache' section is currently empty. The bottom of the window shows a taskbar with various application icons and a system tray with the time 23:55 and date 2015/01/25.

Run

Run GSEAをクリックして実行

The screenshot shows the GSEA v2.1.0 web application interface. The title bar reads "GSEA v2.1.0 (Gene set enrichment analysis -- Broad Institute)". The menu bar includes "File", "Options", "Downloads", "Tools", and "Help".

Steps in GSEA analysis:

- Load data
- Run GSEA** (highlighted with a red arrow)
- Leading edge analysis
- Enrichment Map Visualization

Gene set tools:

- Chip2Chip mapping
- Browse MSigDB
- Analysis history

GSEA reports:

Processes: click 'status' field for results

Name	Status

Show results folder

Main Content Area:

- Steps in GSEA**
 - 1. What you need for GSEA**
 - Expression data set
 - Phenotype annotation
 - Gene sets – use MSigDB or your own gene sets
 - 2. Run GSEA**
 - Start with default parameters
 - If you want to collapse probes to genes, specify chip platform
 - 3. View results**
 - 4. Leading edge analysis**
 - Leading edge finds genes driving enrichment results
- Gene Set Tools**
 - Chip2Chip mapping**
 - Convert gene sets between platforms
 - Chip2Chip mapping
 - Explore MSigDB gene sets**
 - Search the database of thousands of gene sets
 - Browse the gene sets by name
 - Find overlapping gene sets
 - Export gene sets
 - Browse MSigDB
 - See also**
 - MSigDB online tools at: www.broadinstitute.org/msigdb
- Getting Help**
 - GSEA web site:** www.broadinstitute.org/gsea
 - GSEA documentation:** www.broadinstitute.org/gsea/wiki
 - Email the GSEA team:** gsea@broadinstitute.org

BROAD INSTITUTE

System tray: 23:29:55, 30M of 44M, 23:29 2015/01/25

Run

The screenshot shows the GSEA v2.1.0 application window. The main panel is titled "Gsea: Set parameters and run enrichment tests". It contains several sections:

- Required fields:**
 - Expression dataset: GSE40493_Normalized_GSEA_UPPER [21535x8 (ann: 21535,8,chip na)]
 - Gene sets database: Users\Jun\Dropbox\BioinformaticsStudy\JishuCourse\geneset_Bcl6.grp
 - Number of permutations: 1000
 - Phenotype labels: %BioinformaticsStudy%JishuCourse%GSE40493_Class.cls#KO_versus_WT
 - Collapse dataset to gene symbols: false
 - Permutation type: gene_set
 - Chip platform(s):
- Basic fields:** (with a "Show" button)
- Advanced fields:** (with a "Show" button)

Annotations with arrows point to specific elements:

- gctファイルを選択 (Select gct file) - points to the Expression dataset dropdown.
- grpファイルを選択 (Select grp file) - points to the Gene sets database dropdown.
- 発現比の方向 WT/KO KO/WT (Direction of expression ratio WT/KO KO/WT) - points to the Phenotype labels dropdown.
- false - points to the Collapse dataset to gene symbols dropdown.
- gene_set - points to the Permutation type dropdown.
- runをクリックして実行 (Click run to execute) - points to the Run button at the bottom right.

In the bottom left, the "GSEA reports" section shows a table with the following data:

	Name	Status
1	Gsea	Success 5

An annotation "ステータスが表示 Successと表示されたらクリック 結果を確認" (Status is displayed, click when Success is displayed, check results) points to the "Success 5" cell in the table.

At the bottom of the window, the status bar shows: 23:58:31 | 4535 [INFO] Parsed from unigene / gene symbol: 38870 | 99M of 247M

ブラウザ上で結果を表示

GSEA Report for Dataset GSE40493_Normalized_GSEA_UPPER

Enrichment in phenotype: KO (4 samples)

- None of the gene sets are enriched in phenotype KO
- [Guide to interpret results](#)

Enrichment in phenotype: WT (4 samples)

- 1 / 1 gene sets are upregulated in phenotype WT
- 1 gene sets are significantly enriched at FDR < 25%
- 1 gene sets are significantly enriched at nominal pvalue < 1%
- 1 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to interpret results](#)

← enrichment result in htmlをクリック

Dataset details

- The dataset has 21535 features (genes)
- No probe set => gene symbol collapsing was requested, so all 21535 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 0 / 1 gene sets
- The remaining 1 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

Gene markers for the KO versus WT comparison

- The dataset has 21535 features (genes)
- # of markers for phenotype KO: 12807 (59.5%) with correlation area 51.8%
- # of markers for phenotype WT: 8728 (40.5%) with correlation area 48.2%
- Detailed [rank ordered gene list](#) for all features in the dataset
- [Heat map and gene list correlation](#) profile for all features in the dataset

Global statistics and plots

Windows taskbar: 0:02 2015/01/26

ブラウザ上で結果を表示

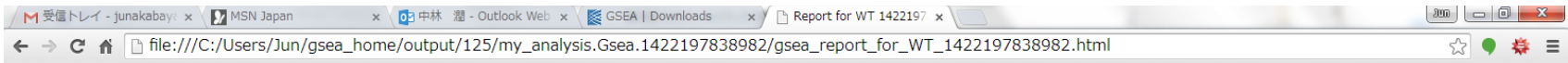


Table: Gene sets enriched in phenotype WT (4 samples) [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	geneset_Bcl6.gp	Details...	261	-0.48	-2.37	0.000	0.000	0.000	4412	tags=43%, list=20%, signal=53%

↑
detailsをクリック



ブラウザ上で結果を表示

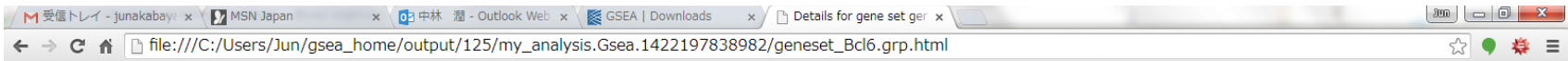
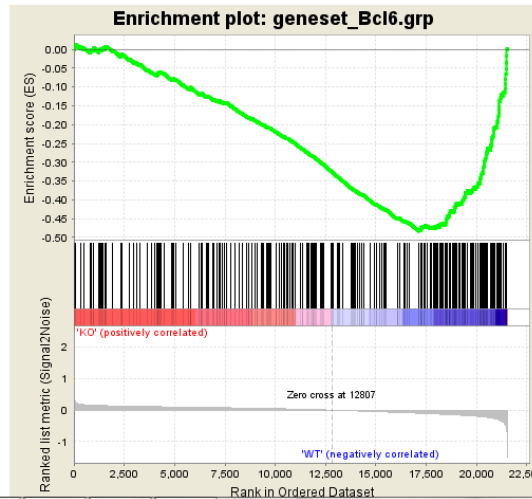


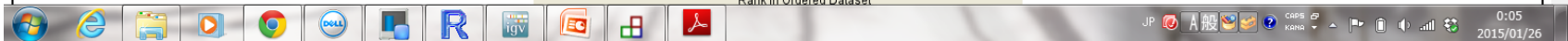
Table: GSEA Results Summary

Dataset	GSE40493_Normalized_GSEA_UPPER.GSE40493_Class.cls#KO_versus_WT
Phenotype	GSE40493_Class.cls#KO_versus_WT
Upregulated in class	WT
GeneSet	geneset_Bcl6.grp
Enrichment Score (ES)	-0.48353454
Normalized Enrichment Score (NES)	-2.373729
Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

統計量



enrichment score



結果のファイル

- 結果はgsea_homeフォルダに自動的に保存されます。

gene set database

The screenshot displays the GSEA v2.1.0 software interface. The main window is titled "GSEA v2.1.0 (Gene set enrichment analysis -- Broad Institute)". The "Run GSEA" dialog box is open, showing the "Gene sets database" field. A smaller dialog box, "Select one or more gene sets(s)", is overlaid on top. This dialog has a "Text entry" field and four tabs: "Gene matrix (from website)", "Gene sets (grp)", "Gene matrix (local gmx/gmt)", and "Subsets". The "Gene matrix (from website)" tab is selected. Below the tabs, there is a text area containing the message: "Error listing Broad website", "Read timed out", and "Choose gene sets from other tabs". An arrow points from the text "gene matrix (from website)" at the bottom right of the image to the selected tab in the dialog box. The main window also shows various settings like "Expression dataset" (GSE40493_Normalized_GSEA_UPPER [21535x8 (ann: 21535,8,chip na)]), "Number of permutations" (1000), and "Phenotype labels". The status bar at the bottom shows "14:19:23", "0908 [INFO] Loading ... 1 files:GSE40493_Class.clsFiles loaded successfully: 1 / 1There were NO errors", and "47M of 205M".

gene matrix (from website)

第4回

日時：平成27年2月23日（月）17:00～

「マイクロアレイデータ機能解析（続き）」

「RNA-seq データ解析

パブリックデータベースからデータ取得、

読み取り、正規化、保存」