

# User Manual

## ExDEGA v.5.1.1 & Data Analysis

## < 목 차 >

1.	Differentially Expressed Genes Analysis (ExDEGA)	3
2.	Functional Annotation Analysis (DAVID, ExDEGA GraphicPlus)	22
3.	Clustering Heatmap Analysis (Clustering, ExDEGA GraphicPlus)	36
4.	Principal Component Analysis (PCA, ExDEGA GraphicPlus)	41
5.	String Network Analysis (Network, ExDEGA GraphicPlus)	50
6.	Correlation Analysis (ExDEGA GraphicPlus)	54
7.	Pathway Analysis (KEGG mapper)	57
8.	Gene Set Enrichment Analysis (GSEA)	60
9.	Protein-Protein Network Analysis (Cytoscape STRING)	65

# 1. Differentially Expressed Genes Analysis (ExDEGA)

㈜이바이오젠은 RNA-Seq (Quant-Seq, mRNA-Seq, Total RNA-seq)과 Microarray data 를 엑셀 기반에서 쉽게 분석할 수 있도록 분석 결과 보고 시 ExDEGA (Excel based Differentially Expressed Gene Analysis) tool 과 ExDEGA GraphicPlus 를 함께 제공한다. ExDEGA 분석 툴은 ㈜이바이오젠이 연구자들이 RNA-Seq 및 Microarray 데이터를 보다 쉽게 다루고 원하는 데이터를 쉽게 얻을 수 있도록 사용자 편의를 최대한 반영한 분석 툴이며 엑셀 프로그램 안에서 다양한 분석을 직관적으로 수행할 수 있도록 개발되었다. ExDEGA 분석 툴은 사용자들의 요구사항을 지속적으로 반영하여 데이터 분석과 엑셀 사용에 익숙하지 못한 연구자들도 쉽게 사용이 가능하도록 계속 업데이트 될 예정이다.

이바이오젠에서 제공하는 Microarray data 와 RNA-Seq data (엑셀 데이터)를 열기 전에 함께 제공한 ExDEGA\_v(버전)\_Installer.zip 파일을 다운로드 폴더에서 압축을 풀고, setup.exe 를 실행하면 분석 툴이 설치된다(그림 1-1A). 만약 다운로드 폴더에서 압축을 풀지 않았을 경우, 별도로 압축된 파일에 있는 ExDEGAGraphicPlus.exe 를 컴퓨터의 로컬 C 드라이브 아래로 복사/붙여넣기 하면 ExDEGA GraphicPlus 프로그램이 설치 완료된다(그림 1-1B).

설치가 완료되고 ExDEGA format 의 엑셀 데이터를 열면 자동으로 ExDEGA 분석 툴이 구동된다. 참고로 ExDEGA 설치 전에 실행 중인 엑셀 파일이 있으면 종료시킨 후 다시 실행해야 ExDEGA 를 사용할 수 있다. ExDEGA 설치 및 구동에 오류가 있으면 ExDEGA 오류 해결 메뉴얼 ([Download link](#))을 확인한다.

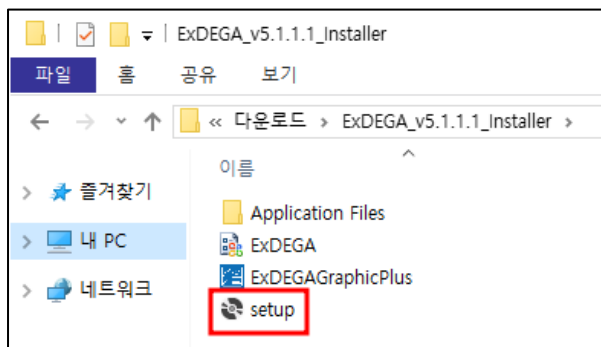


그림 1-1A. ExDEGA setup

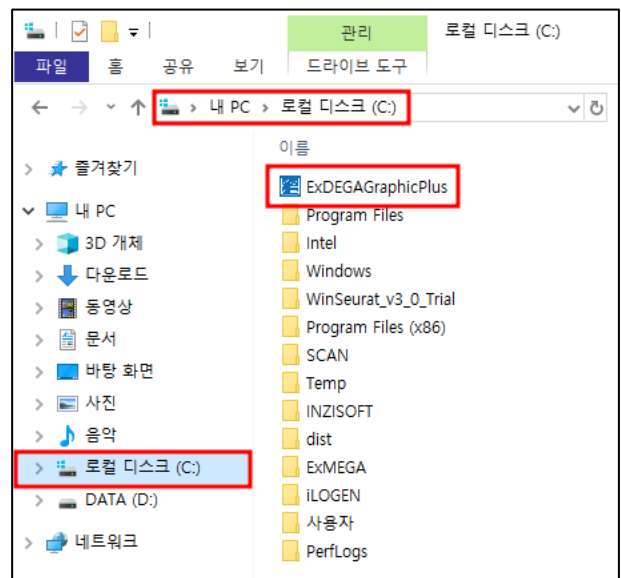


그림 1-1B. ExDEGA GraphicPlus Installation

ExDEGA format 의 엑셀 파일을 열면, 왼쪽에 Gene Category 창과 가운데에 gene expression data, 오른쪽에 DEG Analysis 창이 실행된다(그림 1-2). Gene Category 분석 창에서는 기본 설정된

Gene ontology (GO)가 있고 사용자가 원하는 대로 gene category 를 구성하여 분석할 수 있다. Gene category 창과 DEG Analysis 창은 함께 연동하여 데이터를 쉽게 얻을 수 있다. DEG Analysis 창에서는 Fold change, Normalized Data (log2), p-value 등을 선택하여 DEG 선별을 쉽게 할 수 있고 DEGs 를 gene category 별로 그래프를 작성할 수 있다. 뿐만 아니라, DEG 분석 창에서 Scatter Plot, Volcano Plot, Venn Diagram 을 직접 그릴 수 있고 선별된 유전자들을 대상으로 Clustering heatmap, KEGG 분석, DAVID 분석을 수행하기 위한 input file 을 자동으로 만들 수 있다. Gene expression graph, Gene search 기능도 이용할 수 있어 연구자가 RNA-Seq, microarray data 를 쉽게 활용할 수 있다.

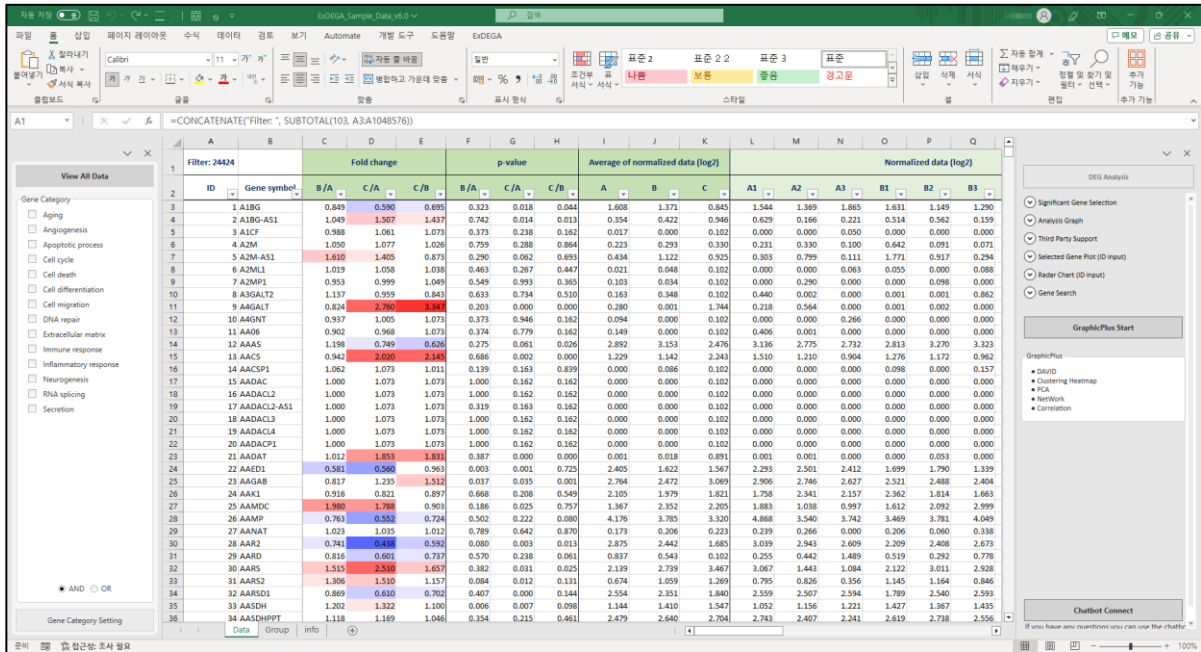


그림 1-2. RNA-seq or Microarray data in ExDEGA format

### 1-1. Gene Category 사용 방법

RNA-seq 또는 Microarray data 는 수 만개의 유전자를 포함하기 때문에 유전자를 한 개씩 분석하기 보다 기능별로 그룹을 지어 분석을 하는 것이 용이하다. 이를 위해 많은 연구자들이 Gene Ontology(GO)를 활용한다. GO 는 비슷한 기능의 유전자들을 묶어 놓은 그룹이라고 생각하면 이해하기 쉽다.

Gene Category 창은 분석된 종에 대한 GO 가 임의로 추가되어 있으며, 관련 유전자들을 필터링할 수 있다. 예를 들어, Aging 관련 유전자만 분석을 원할 경우, Gene Category 창에서 Aging 을 선택하면 해당 유전자 리스트만 필터링 된다. 그리고 Gene Category 의 여러 GO 들을 선택하여 동시에 해당하는 유전자 (교집합)를 필터링할 수 있는 "AND" 기능과 한 GO 만이라도 해당하는 유전자 (합집합)를 필터링할 수 있는 "OR" 기능을 갖추고 있다(그림 1-3)

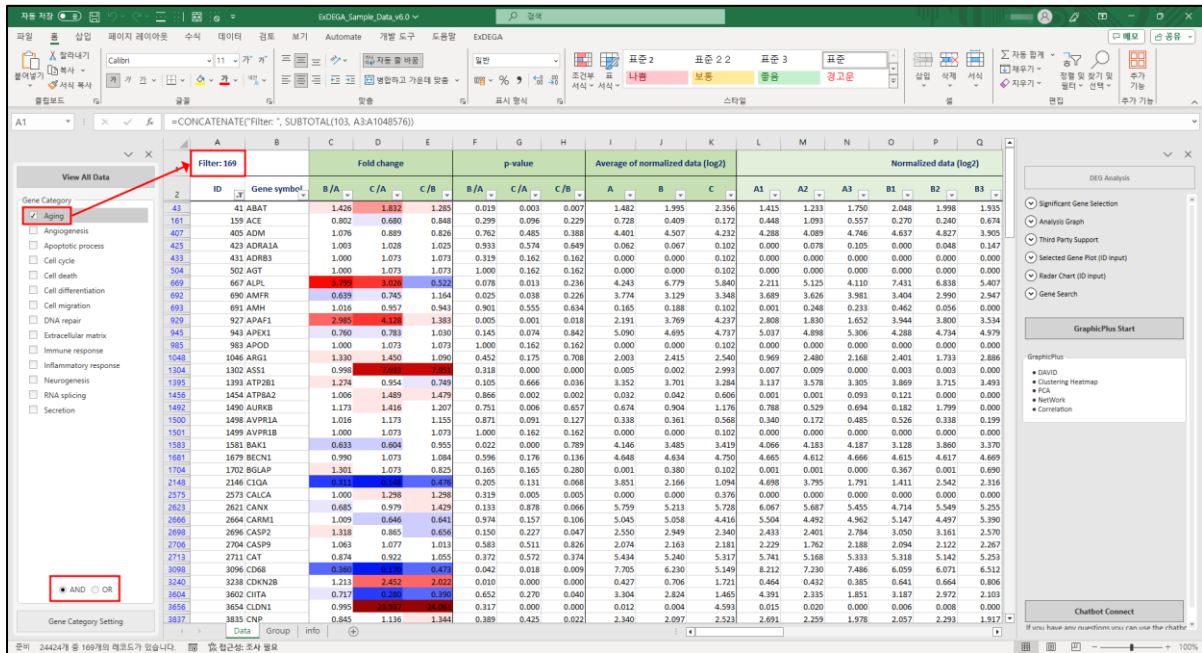


그림 1-3. Gene ontology selection

가장 왼쪽 상단에 'View All Data' 버튼을 누르면 필터를 모두 해제되어 다시 전체 결과를 볼 수 있고 기존 GO 중 관심있는 GO 가 없다면 'Gene Category Settings' 버튼을 이용하여 Quick GO site 에서 다른 GO 를 추가할 수 있다. '?' 버튼을 누르면 Quick GO site 를 이용한 Gene Category 를 추가하는 방법이 자세히 설명되어 있다(그림 1-4).

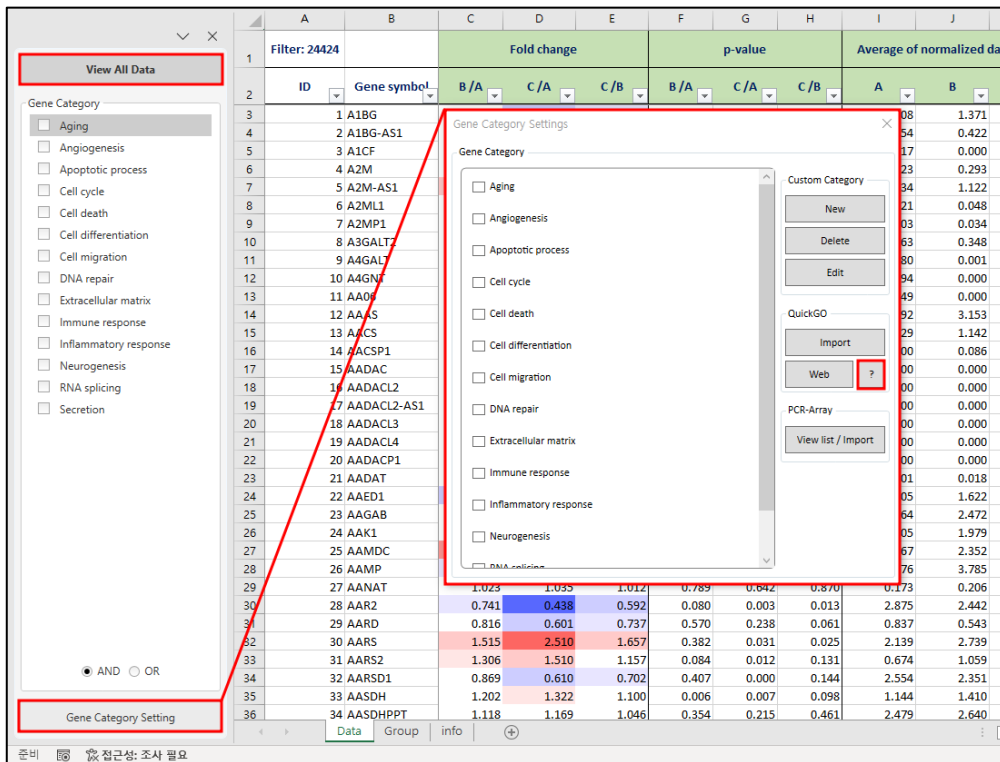


그림 1-4. Gene category settings

만약 원하는 유전자 그룹 목록을 알고 있다면, 직접 입력하여 새로운 Gene Category 를 추가할 수도 있다. Gene Category Settings 버튼을 누른 후 New 를 선택하고 원하는 Gene symbol list 입력(or 복사&붙여넣기) 한 뒤, Gene category 이름 설정 후 Setting 창을 닫아주면 새로운 Gene category 를 확인할 수 있다(그림 1-5A, B).

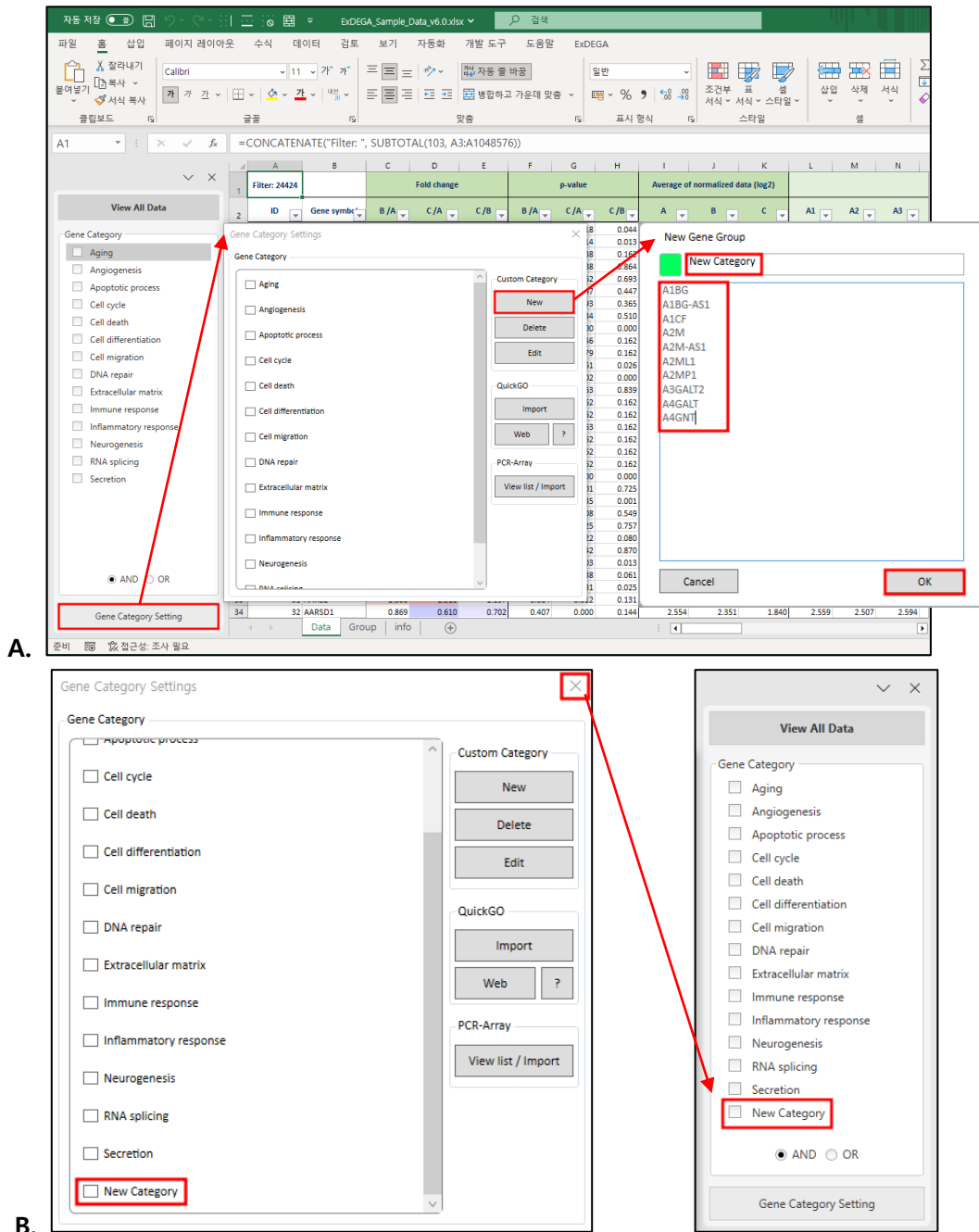


그림 1-5. Adding Genes to make a new gene category

생성된 Category 를 수정 혹은 삭제하고 싶다면 Category 선택 후 Delete 를 선택하면 삭제를, Edit 을 클릭하면 수정을 할 수 있다(그림 1-6).

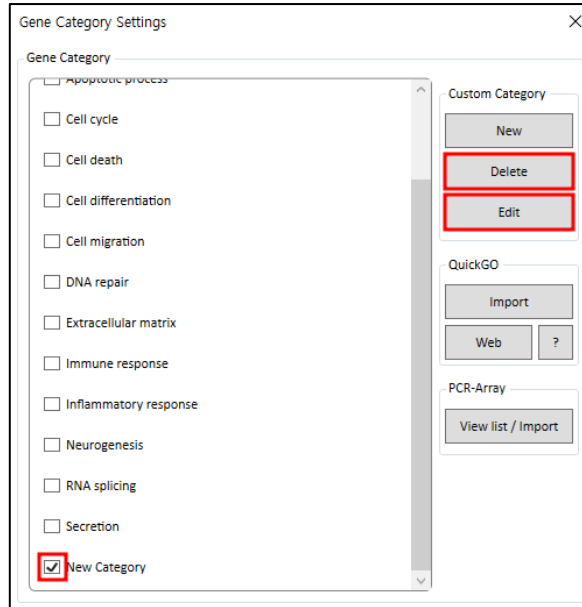


그림 1-6. Edit or Delete gene category

PCR-Array 항목의 View list / Import 버튼을 이용하여 Pathway 별 Gene list 를 추가 할 수 있다. Gene Category Settings 버튼을 누른 후 View list / Import 버튼을 누른다. Sub Window 창에서 species 를 선택하고 Keyword 에 추가하고자 하는 Pathway 이름이나 유전자 이름을 검색하고 Check box 에 체크한 뒤 Import 버튼을 누르면 자동으로 추가된다(그림 1-7).

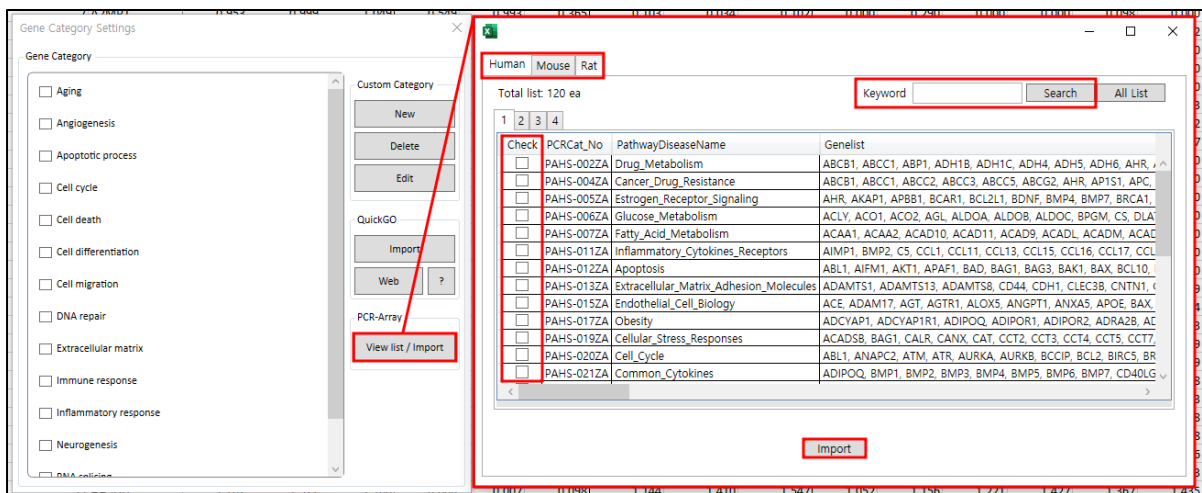


그림 1-7. PCR-Array Pathway settings

## 1-2. Significant Gene Selection 사용 방법

오른편의 DEG Analysis 부분에서 "Significant Gene Selection" 창은 전체 결과 중 대조군과 실험군을 비교한 결과에서 유의하게 발현 차이가 나는 유전자를 필터링 할 수 있도록 만들어 놓은 것이다. 예를 들어, B/A 비교조합을 선택하고 [Fold change : 2, Normalized Data (log2) : 4, p-value : 0.05] 를 선택하면, A 대비 B 에서 2 배 이상 발현이 증가 또는 감소하고, Normalized Data (log2)값이 4 이상이고, p-value 값이 0.05 이하인 유전자가 필터링 된다(그림 1-8). p-value 는 반복 실험한 데이터(N>=2)의 경우만 제공된다. 비교조합은 다중 선택할 수 있으며 "AND"나 "OR"를 기능을 이용하면 선택한 비교조합들에서 공통적인 DEGs (교집합) 또는 하나의 비교조합 이상 DEGs (합집합)을 선별할 수 있다.

유전자 선별 기준(Fold change, p-value 및 Normalized data (log2))은 연구자의 데이터에 맞게 조정하여 사용할 수 있다.

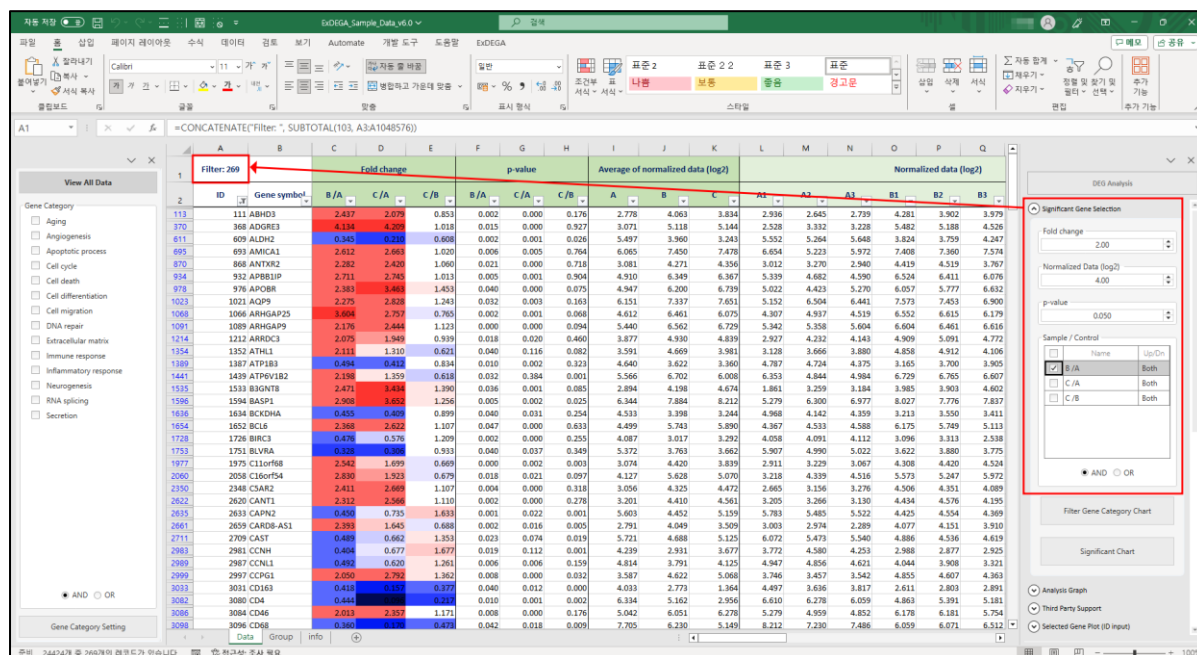


그림 1-8. Significant gene selection

Significant gene selection 에서 증가 또는 감소한 유전자를 각각 보고 싶다면 Up/Dn 의 selection box 에서 선택할 수 있다. Both 는 증가, 감소 유전자가 모두 필터링 되고 Up 은 증가한 유전자만 Dn 은 감소한 유전자만 따로 필터링 할 수 있다(그림 1-8).



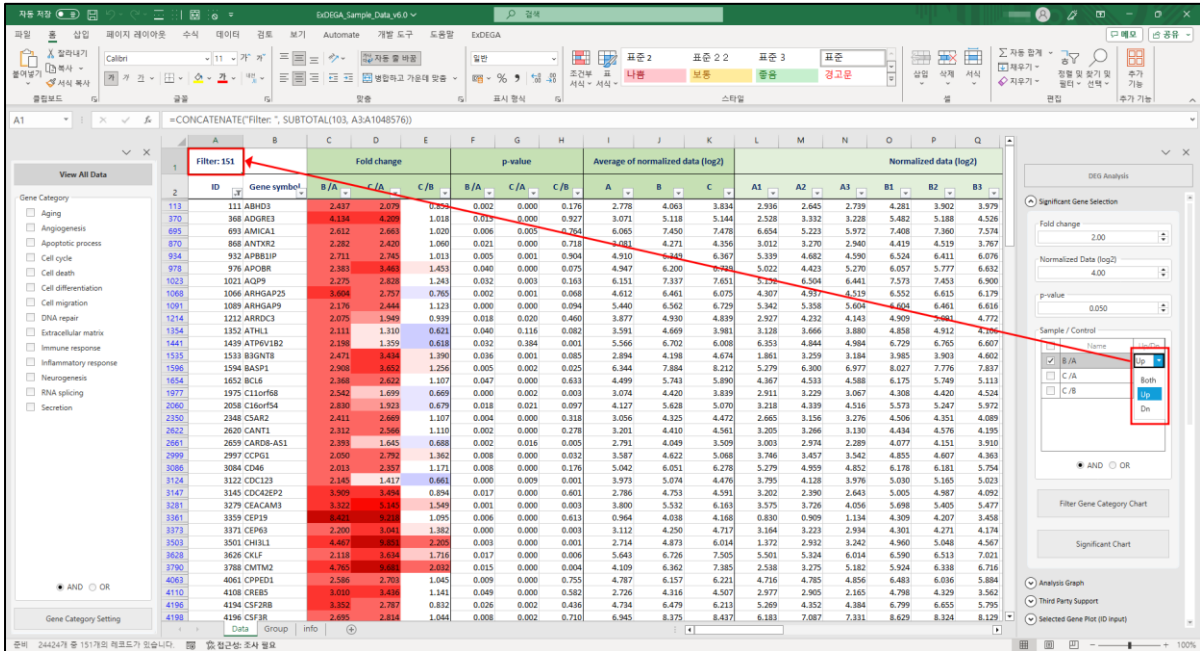


그림 1-8. Significant genes (separately Up and Dn)

Gene Category와 Significant gene selection은 연동이 가능하다. 그림 1-9 에서처럼 Significant Gene Selection에서 선별조건을 설정하고 Gene Category의 GO를 선택하면 모든 조건이 적용된 유전자가 선별된다. 선별된 유전자는 예시 데이터에서 선택한 GO와 관련된 유전자들 중 선별조건이 적용된 유전자를 의미한다.

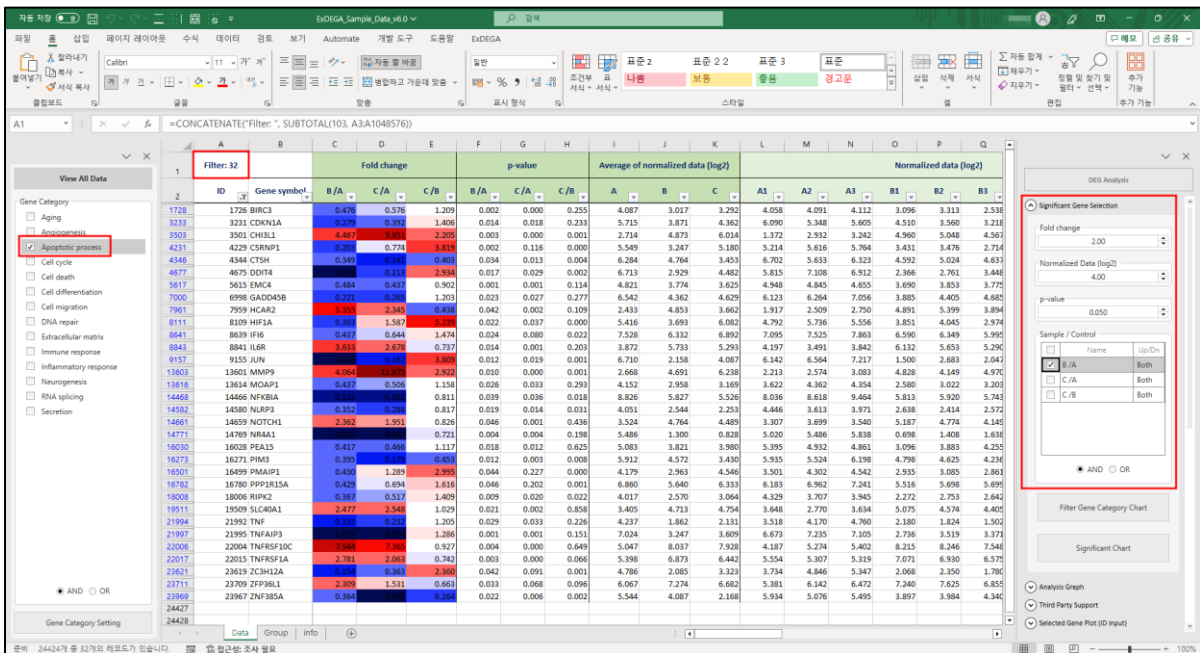


그림 1-9. Significant genes related to Apoptotic process



### 1-3. Analysis Graph 사용 방법

DEG Analysis 부분에서 "Analysis Graph" 창을 펼치면 그림 1-12 와 같이 Scatter Plot, Volcano Plot, Venn Diagram 을 그릴 수 있다.

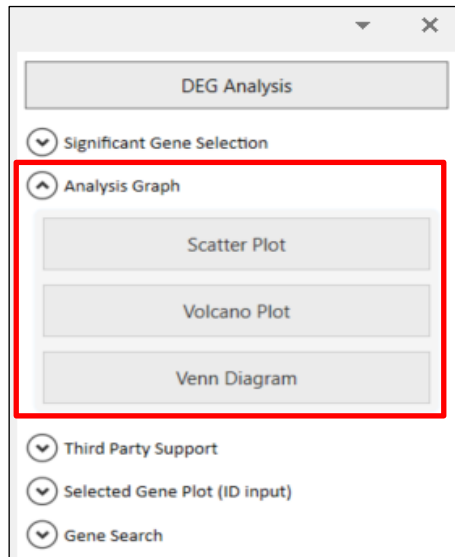


그림 1-12. Analysis Graph Tool

#### 1-3-1. Scatter Plot

Scatter Plot 은 대조군과 실험군의 발현양상을 확인할 수 있는 이미지이다. 오른쪽에 샘플 비교 그룹과 Fold threshold line (예시: 2fold)을 선택하고 "Graph View"를 클릭하면 왼쪽에 선택한 비교 그룹을 대상으로 Scatter Plot 이 자동 생성된다. 생성된 그래프는 마우스 스크롤로 크기 조절할 수 있다. x 축은 대조군의 Normalized data (log2), y 축은 실험군의 Normalized data (log2)이다. 초록색 사선 아래는 2fold 이상 감소한 유전자들, 빨간색 사선 위 2fold 이상 증가한 유전자들이다. 유의한 유전자를 식별하기 위한 색은 빨강, 파랑, 초록 중 사용자가 선택할 수 있다. Plot 에서 특정 spot 을 클릭하면 해당 유전자명이 표시되고 마우스 오른쪽을 클릭하여 지울 수도 있다. 표시된 유전자명은 마우스로 위치 조절이 가능하고 "Font Size"에서 표시된 유전자명의 글씨 크기를 조절할 수 있으며 "Hide"를 클릭하면 그래프의 grid line 을 숨길 수 있다. 그리고 여러 개의 유전자를 동시에 표시하고 싶다면 "Gene Select (ID Input)" 창에 해당 유전자의 ID 를 복사하여 입력하고 "Add"를 클릭하면 Gene Symbol 이 자동 생성된다(그림 1-13).

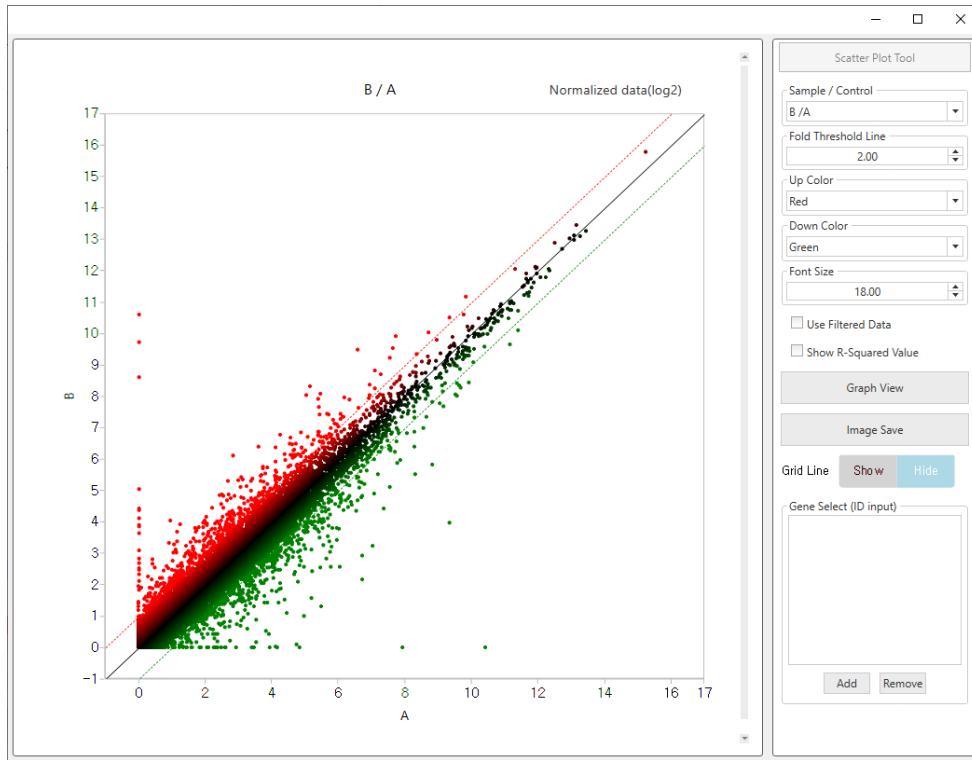


그림 1-13. Analysis Graph Tool – Scatter Plot

생성된 Scatter Plot은 전체 유전자를 기반으로 생성된 그래프이며 원하는 GO category에 대한 Scatter Plot을 그릴 수 있다. GO category 선택 후, “Use Filtered Data”를 체크하고 “Graph View”를 클릭하여 그래프를 그리면 선택한 GO Category를 기반으로 한 Scatter Plot을 생성할 수 있다(그림 1-14).

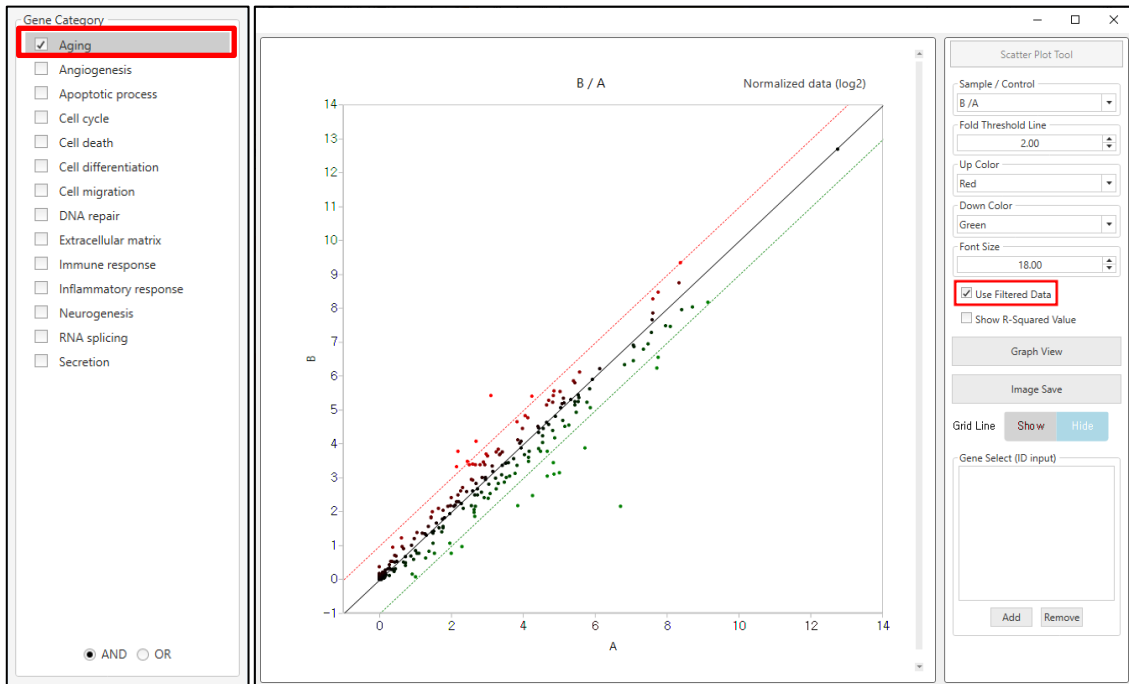


그림 1-14. Analysis Graph Tool – Use Filtered Data

### 1-3-2. Volcano plot

Volcano Plot 은 반복 실험( $N \geq 2$ )이 된 경우에만 분석 가능하다. Volcano Plot 은 Scatter Plot 의 기능과 거의 동일한데 오른쪽에 샘플 비교 그룹과 Fold threshold line (예시: 2fold), p-value (예시: p-value 0.05)를 선택하고 "Graph View"를 클릭하면 왼쪽에 선택한 비교 그룹을 대상으로 Plot 이 자동 생성된다. 생성된 그래프는 마우스 스크롤로 크기 조절할 수 있다. 초록색 세로선 왼쪽은 2fold 이상 감소한 유전자들, 빨간색 세로선 오른쪽은 2fold 이상 증가한 유전자들, 검은색 가로선 위는 p-value 0.05 이하인 유전자들이다. 유의한 유전자를 식별하기 위한 색은 빨강, 파랑, 초록 중 사용자가 선택할 수 있다. Plot 에서 특정 spot 을 클릭하면 해당 유전자명이 표시되고 마우스 오른쪽을 클릭하여 표시를 지울 수도 있다. 표시된 유전자명은 마우스로 위치 조절이 가능하고 "Font Size"에서 표시된 유전자명의 글씨 크기를 조절할 수 있으며 "Hide"를 클릭하면 그래프의 grid line 을 숨길 수 있다. 그리고 여러 개의 유전자를 동시에 표시하고 싶다면 "Gene Select(ID Input)" 창에 해당 유전자의 ID 를 복사하여 입력하고 "Add"를 클릭하면 Gene Symbol 이 자동 생성된다(그림 1-15).

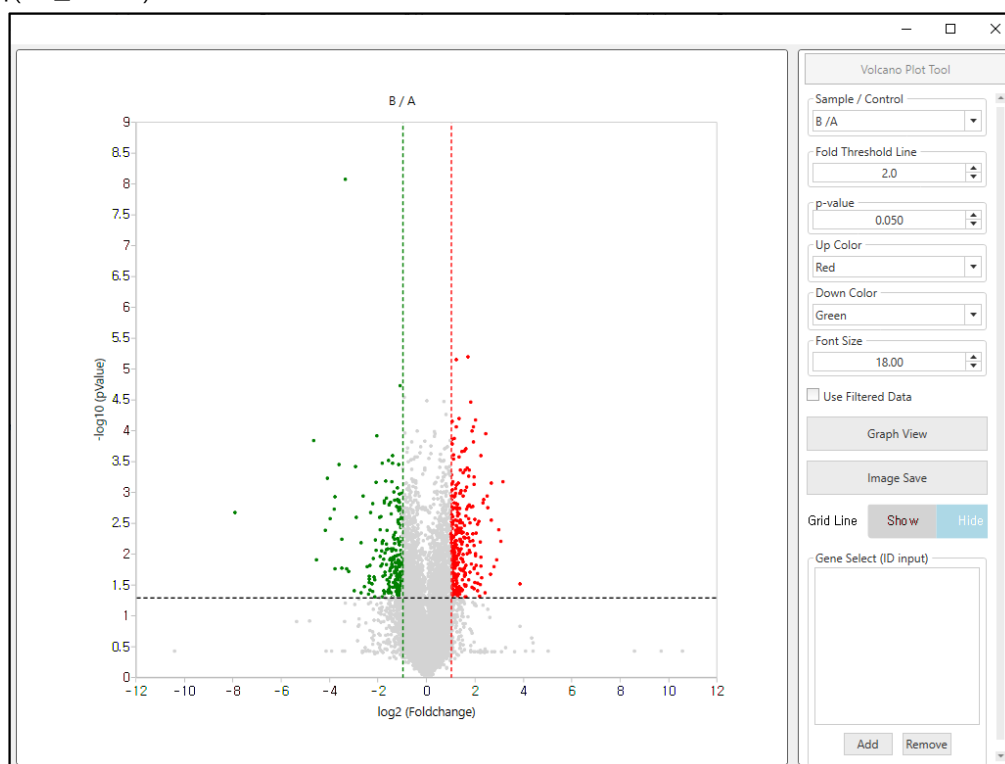


그림 1-15. Analysis Graph Tool – Volcano Plot

생성된 Volcano Plot은 전체 유전자를 기반으로 생성된 그래프이며 원하는 GO category에 대한 Volcano Plot을 그릴 수 있다. GO category 선택 후, "Use Filtered Data"를 체크하고 "Graph View"를 클릭하여 그래프를 그리면 선택한 GO Category를 기반으로 한 Volcano Plot을 생성할 수 있다(그림 1-14).

### 1-3-3. Venn diagram

Venn Diagram 을 통해 4 개 이하의 비교조합을 대상으로 Venn Diagram 을 작성할 수 있다. Venn Diagram 을 그릴 Fold Change 와 Normalized data (log2), p-value 을 선택 후, Diagram View 를 클릭하면 결과를 확인할 수 있으며 비교조합은 최대 4 개까지 선택 가능하다. 그리고 "theme"을 통해서 Venn Diagram 의 색상 테마 선택이 가능하다. 아래의 그림은 B/A, C/A, C/B 결과를 비교하여 2fold, Normalized data (log2) 4 이상, p-value 0.05 이하인 유전자 list 를 가지고 Venn Diagram 을 작성한 결과이다(그림 1-16).

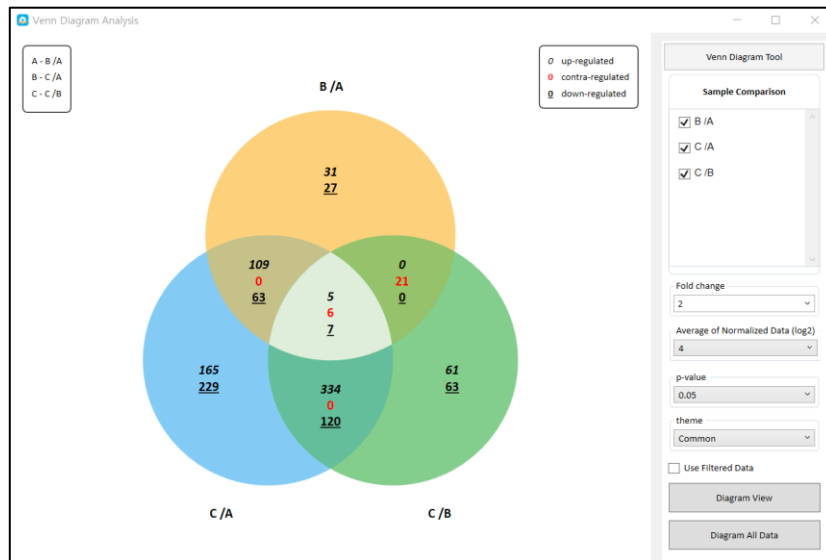


그림 1-16. Analysis Graph Tool – Venn Diagram

생성된 Venn Diagram 은 전체 유전자를 기반으로 생성된 그래프이며 원하는 GO Category 에 대한 Venn Diagram 을 그릴 수 있다. GO category 선택 후, "Use Filtered Data"를 체크하고 Diagram View 를 클릭하여 그래프를 그리면 선택한 GO Category 를 기반으로 한 Venn Diagram 을 생성할 수 있다(그림 1-17).

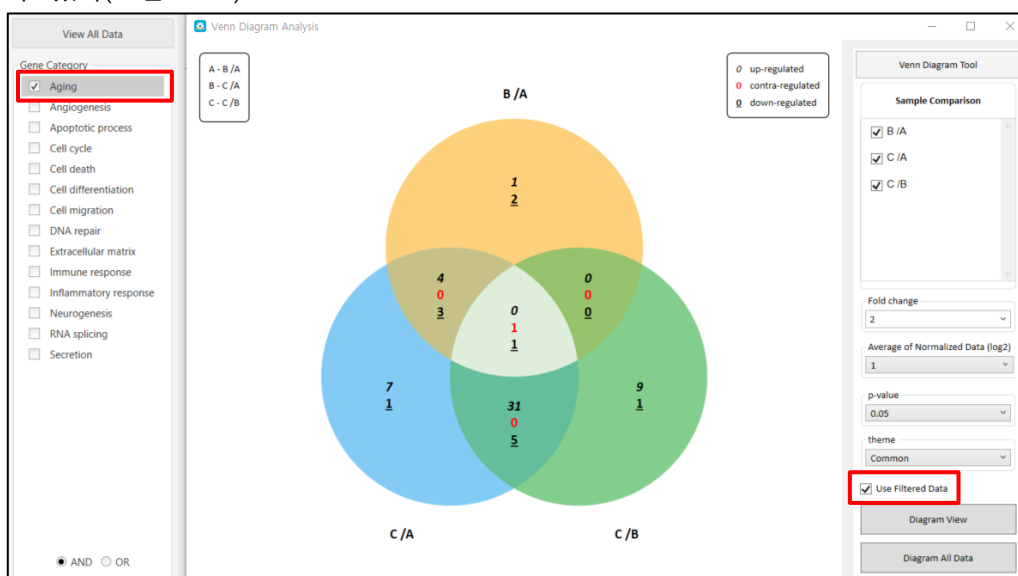


그림 1-17. Use Filtered Data

Venn Diagram 결과에서 표시되는 형식은 다음과 같다(그림 1-18).

- **기울어진 숫자** : up-regulated 된 gene 수
- **빨간색 숫자** : regulation 이 대조되는 gene 수  
(예: B/A 에서는 up 되고 C/A 에서는 down 되는 gene 수)
- **밑줄 친 숫자** : down-regulated 된 gene 수

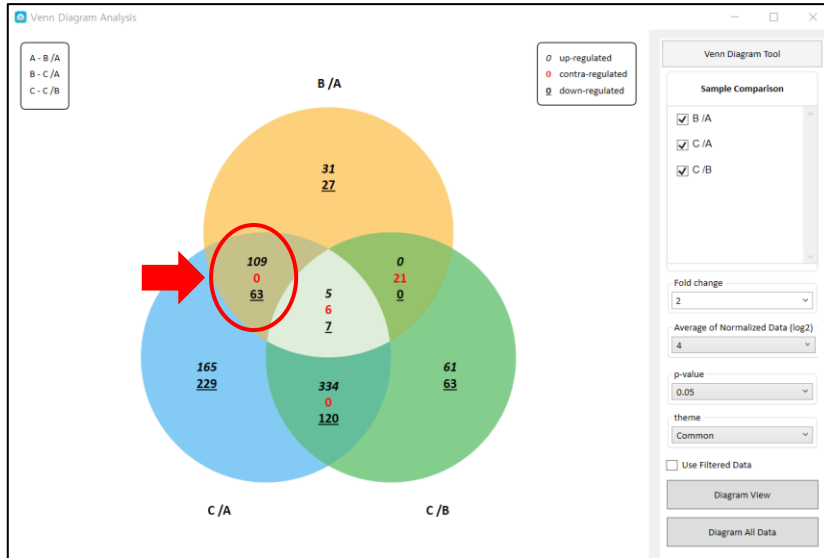


그림 1-18. For example of up, down, contra-regulated in Venn Diagram

Venn Diagram 각 영역에 어떤 유전자들이 있는지 확인할 수도 있다. 예를 들어, B/A 에서만 up 이 되는 유전자를 보고 싶으면, Venn Diagram 에서 B/A 에서만 해당되는 영역을 찾아 마우스 오른쪽 클릭하고 up-regulated 를 선택하면 증가된 유전자 list 가 엑셀 data sheet 에 filter 된다 (그림 1-19).

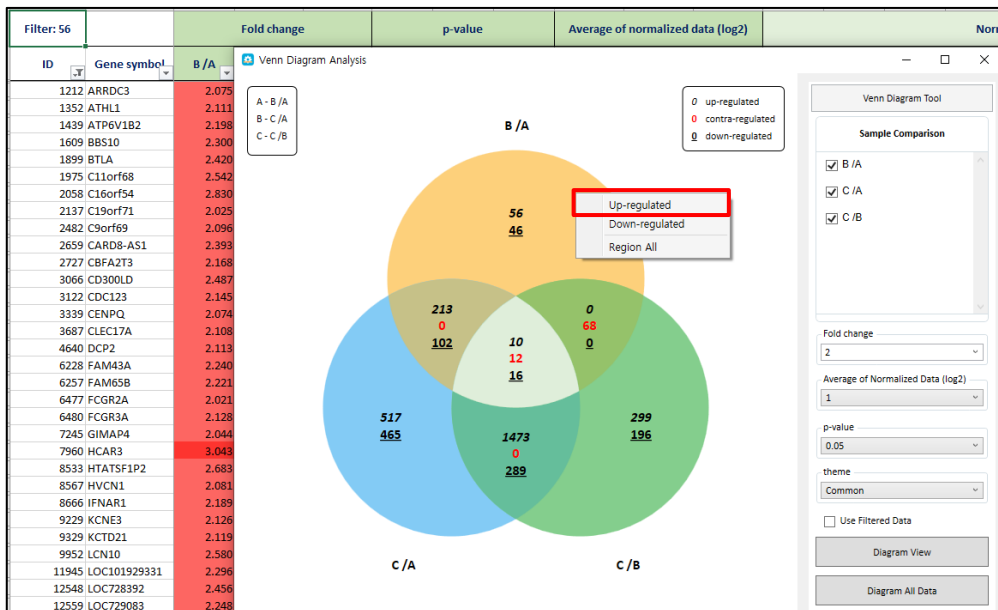


그림 1-19. Filtering 2fold up-regulated gene list in Venn Diagram

ExDEGA 에서 제공되는 모든 이미지는 오른쪽마우스를 눌러 'Save image' 버튼을 통해 저장이 가능하다(그림 1-20).

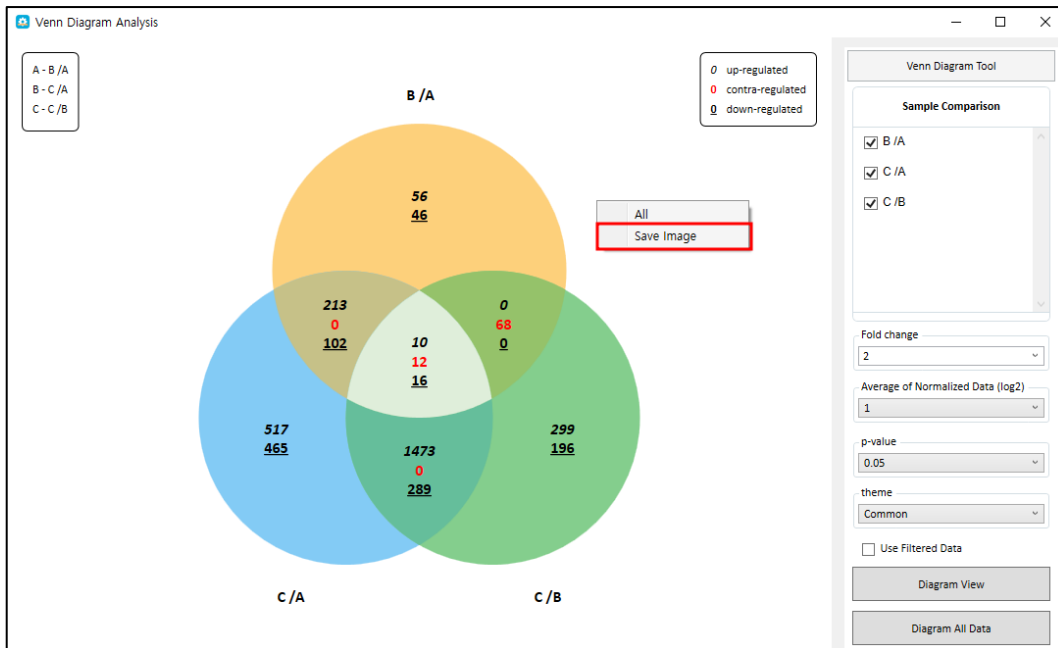


그림 1-20. Save image

#### 1-4. Third Party Support 사용 방법

Third Party Support는 연구자가 선별한 유전자를 기반으로 Clustering heatmap 과 DAVID 및 KEGG 그리고 GSEA분석을 진행하기 위한 입력 데이터를 제공한다(단, GSEA의 경우 선별하지 않고 전체 유전자 리스트를 기반으로 진행하는 것을 권장한다.)(그림 1-21). 먼저, Input File 제작에 앞서 유전자를 선별하는 것이 필요하다. 유전자 선별은 유의성 있는 DEG 분석, Gene Ontology 분석 등으로 선별할 수 있다.

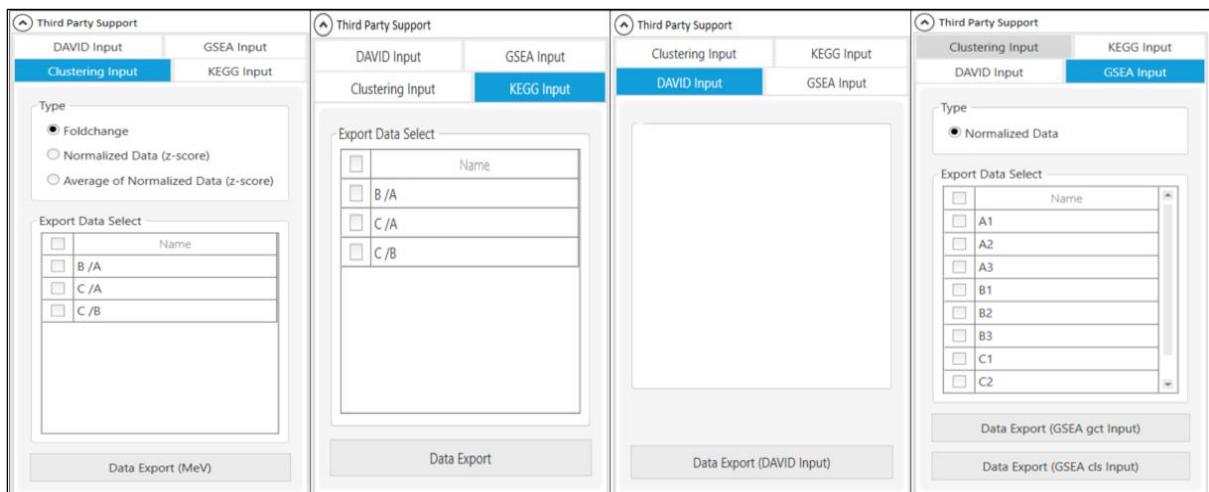


그림 1-21. Third Party Support

(왼쪽부터 Clustering Input, KEGG Input, DAVID Input, GSEA Input)



### 1-4-1. Clustering Input

필터링 된 유전자 리스트를 대상으로 Clustering Heatmap Input 은 크게 세 종류의 데이터를 이용할 수 있다. Input 파일 저장 시, 파일명에는 띄어쓰기가 들어가지 않도록 주의해야 하며 저장된 Clustering Input 파일은 GraphicPlus 혹은 MeV 프로그램을 이용하여 heatmap 을 작성할 수 있다.

Clustering Input 파일을 제작하는 방법은 다음과 같다.

첫 번째, Fold change 값을 이용할 시 Type 부분에 Fold change 를 체크하고 Export Data Select 에서 Heatmap 에 표현할 비교조합을 체크한다. "Data Export"를 클릭한 후 "(input 명).txt"로 저장한다. 저장된 파일은 Fold change 가 log2 변환된 값으로 추출된다.

두 번째, 개별 샘플의 발현 값인 Normalized Data 값을 이용할 시 Type 부분에 Normalized data (z-score)를 체크하고 이용하고자 하는 샘플을 체크한다. "Data Export"를 클릭한 후 "(input 명).txt"로 저장한다. 단, Z-score 를 이용할 때는 표준편차가 고려되기 때문에 반드시 샘플을 3 개 이상을 선택해야 한다.

세 번째, 그룹의 발현 값인 Average of Normalized Data 값을 이용할 시 Type 부분에 Average of Normalized Data (z-score)를 체크하고 이용하고자 하는 그룹명을 체크한다. "Data Export"를 클릭한 후 "(input 명).txt" 로 저장한다. 단, Z-score 를 이용할 때는 표준편차가 고려되기 때문에 반드시 그룹을 3 개 이상을 선택해야 한다.

\* Z-score 는 일반적으로 평균으로부터 얼마만큼 떨어져 있는지를 판단하는 지표이다.

계산방식은 Normalized data 를 log10 으로 변환 후 평균값을 뺀 후 표준편차로 나누어 계산한다.

$$Z\text{-score} = \frac{\{\text{Normalized data} (\log_{10}) - \text{average of Normalized data} (\log_{10})\}}{\text{standard deviation of Normalized data}(\log_{10})}$$

만약 이용하고자 하는 샘플 혹은 그룹의 수가 2 개 이하인 경우, z-score 변환이 불가능하므로 아래의 방법으로 input 파일을 직접 작성한다. 우선 유의성 있는 유전자를 선별한 후, 새로운 excel 에 선별된 Gene list 와 이용하고자 하는 Average of normalized data (log2) 혹은 Normalized data (log2)를 복사하여 Input 파일을 만든다(그림 1-22). Input 파일 저장 시, 파일 형식은 "텍스트(탭으로 분리)(\*txt)"을 선택한다. 저장된 Input 파일은 MeV 프로그램을 이용하여 Heatmap 을 작성할 수 있다.

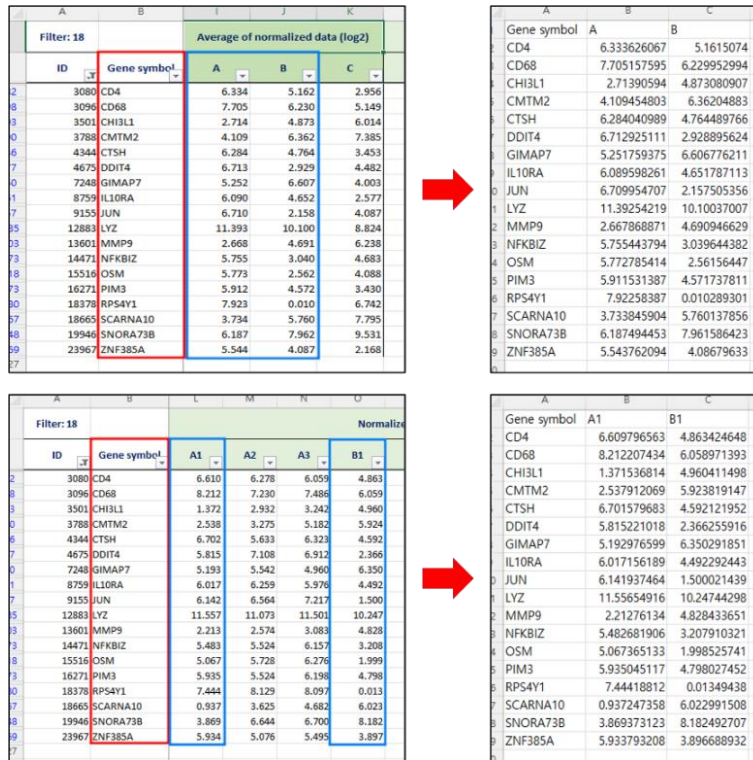


그림 1-22. Heatmap input of no more than 2 samples or groups

Clustering Input file 은 Gene Symbol 과 Fold change (log<sub>2</sub>) 또는 z-score 로 구성된다(그림 1-23). MeV 프로그램을 사용하여 Clustering heatmap 을 작성하는 방법은 MeV manual ([Download Link](#))에서 확인할 수 있다.

Gene Symbol	B / A	C / A	C / B
ABHD3	1.28484278864215	1.05564963600944	-0.229193276278856
ADGRE3	2.04741030637742	2.07337604787126	0.0259657102702727
ALDH2	-1.53704893440065		-2.25385030388101
AMICA1	1.38502619744398	1.41309305964294	0.0280669503482263
ANTXR2	1.19047263791619	1.27486038592147	0.0843881542963257
APBB1IP	1.43866864864493	1.45671679922983	0.0180472923397484
APOBR	1.2529266643143	1.79222275950347	0.539295814093037
AQP9	1.18587922828891	1.49957447547321	0.313695136333861

그림 1-23. Third Party Support - Clustering Input

### 1-4-2. KEGG Input

KEGG Input 은 KEGG Mapper 분석에 이용하기 위한 input 파일을 제공한다. KEGG Input 에서는 하나의 비교 조합만 선택 가능하며, 비교 조합 선택 후 Data Export 를 선택하면 그림 1-24 와 같이 유전자의 Entrez ID와 비교 조합, 발현 수준에 따른 색 코드로 구성된 input파일이 저장된다. KEGG Input 파일을 이용한 KEGG mapper 분석 방법은 7. Pathway Analysis (KEGG mapper)에 자세히 설명되어 있다.

```

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
Entrez ID B /A
18      #FFE4c3,black
100     #E0FFFF,black
133     .
148     .
177     #FFA07A,black
183     .
207     .
218     .
239     .
267     #B0E0E6,black
268     .
317     #FF6347,black

```

그림 1-24. Third Party Support - KEGG input

### 1-4-3. DAVID Input

DAVID Input 은 DAVID 분석에 이용하기 위한 input 파일을 제공한다. 분석하고자 하는 유전자를 선별하고 Data Export 를 클릭하면 그림 1-25 와 같이 Gene Symbol 로 구성된 input 파일이 저장된다. DAVID Input 파일을 이용한 DAVID 분석 방법은 **2. Functional Annotation Analysis (DAVID, ExDEGA GraphicPlus)**에 자세히 설명되어 있다.

```

DAVID_input.txt - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말
ABHD3
ADGRE3
ALDH2
AMICA1
ANTXR2
APBB1IP
APOBR
AQP9
ARHGAP25
ARHGAP9
ARRDC3
ATHL1
ATP1B3
ATP6V1B2
B3GNT8

```

그림 1-25. Third Party Support - DAVID input

### 1-4-4. GSEA Input

GSEA Input 은 GSEA 프로그램에 사용되는 input 파일을 제공한다. GSEA 는 그룹 데이터이면서 그룹당 3 반복 이상인 결과만 분석할 수 있다. GSEA Input 은 별도의 유전자 선별 없이 전체 데이터를 이용하는 것을 권장하며 유전자 발현값 정보가 포함되어 있는 gct 파일과 샘플 정보가 포함되어 있는 cls 파일이 필요하다. gct 및 cls 파일은 각각 그림 1-26, 1-27 와 같은 형식으로 되어있으며, 파일을 저장할 때는 파일명 뒤에 ".gct" 혹은 ".cls"을 반드시 붙이고 파일 형식은 "텍스트 (탭으로 분리) 파일"로 저장한다. GSEA Input 파일을 이용한 GSEA 분석 방법은 **8. Gene Set Enrichment Analysis (GSEA)**에 자세히 설명되어 있다.

	A	B	C	D	E	F	G	H	I	J	K
1	#1.2										
2	24424	9									
3	Gene syml	Gene Title	A1	A2	A3	B1	B2	B3	C1	C2	C3
4	A1BG	NA	1.544002	1.369196	1.864898	1.630973	1.149183	1.289642	0.842837	0.742837	0.942837
5	A1BG-AS1	NA	0.629423	0.166494	0.220651	0.514093	0.562111	0.158501	0.943667	0.843667	1.043667
6	A1CF	NA	4.39E-05	5.74E-05	0.050181	1.67E-05	2.14E-05	0	0.1	0	0.2
7	A2M	NA	0.230622	0.330027	0.100307	0.641994	0.090967	0.071063	0.328091	0.228091	0.428091
8	A2M-AS1	NA	0.303073	0.798804	0.111377	1.771126	0.91748	0.29448	0.922872	0.822872	1.022872

파일 이름(N): gsea\_input.gct  
 파일 형식(T): 텍스트 (탭으로 분리)

그림 1-26. gct file

	A	B	C
1	9 3 1		
2	#A B C		
3	A A A B B B C C C		
4			

파일 이름(N): gsea\_input.cls  
 파일 형식(T): 텍스트 (탭으로 분리)

그림 1-27. cls file

### 1-5. Selected Gene Plot 사용 방법

ExDEGA의 기능 중에 선별한 유전자 또는 연구자가 관심있는 유전자들을 대상으로 발현 패턴을 그래프로 표현하고자 할 때는 "Selected Gene Plot" 기능을 사용할 수 있다. 선별한 유전자의 ID를 복사하여 Selected Gene Plot 창에 붙여 넣고 "Expression Plot View"를 누르면 Normalized data (log2) 값, Fold change (log2) 값으로 line graph가 그려진다(그림 1-28).

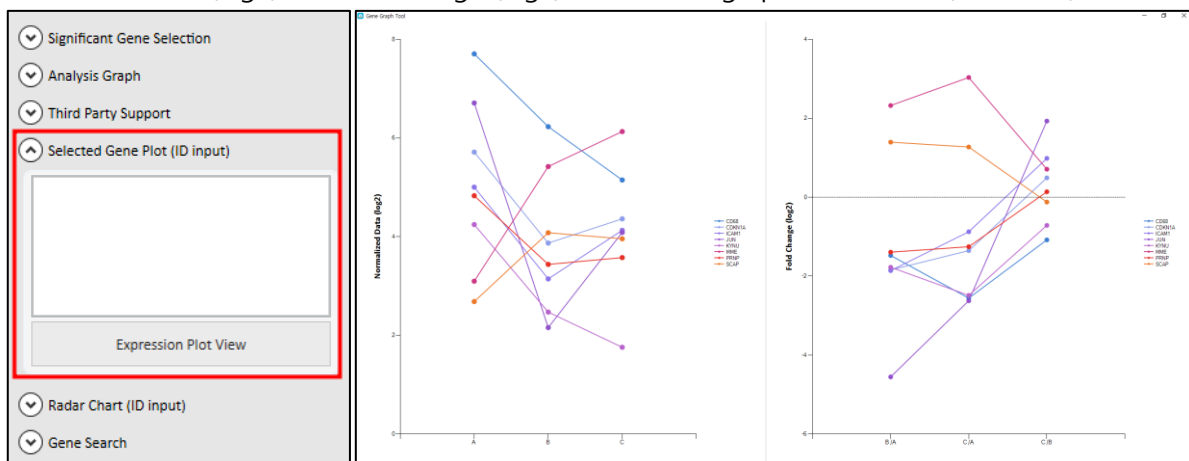


그림 1-28. Selected Gene Plot

### 1-6. Rader Chart 사용 방법

ExDEGA의 기능 중에 선별한 유전자나 연구자가 관심있는 유전자들을 대상으로 발현 패턴을 그래프로 표현하고자 할 때는 "Radar chart" 형식의 이미지로도 표현할 수 있다. 사용방법은 선별한 유전자의 ID를 복사하여 Radar chart 창에 붙여 넣고 옵션을 선택한다. 옵션은 Radar chart의 꼭지점으로 활용할 항목(유전자 or 샘플명)을 선택하고, 표현하고자 하는 발현값(평균 발현값 or 개별 샘플 발현값)을 선택한 뒤 "Radar chart View"를 누르면 Normalized data (log2) 값, Fold change (log2) 값으로 이미지를 그릴 수 있다(그림 1-29). 단, 꼭지점에 있는 항목이 3 가지 이상일 경우 제작 가능하다.

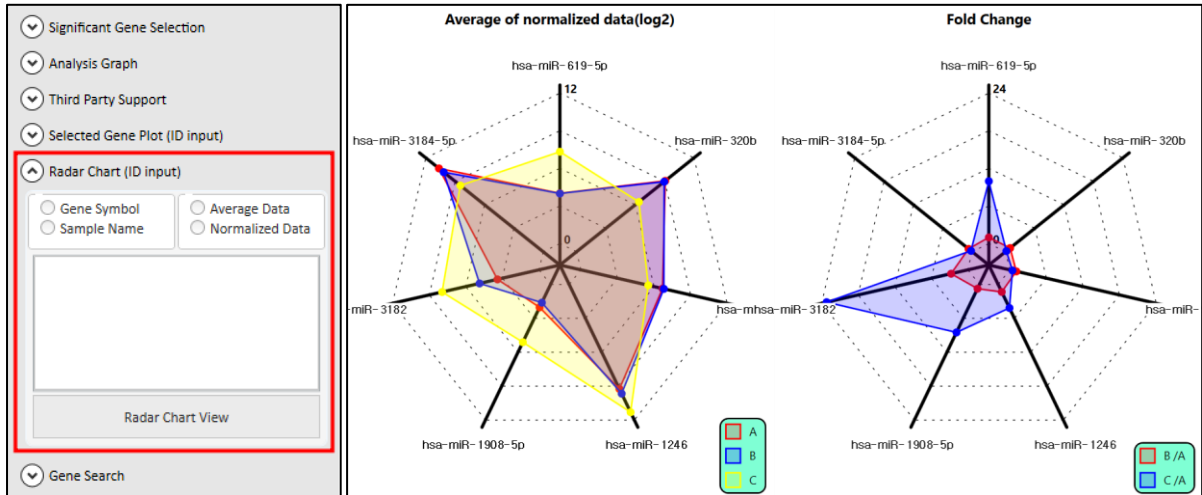


그림 1-29. Radar chart

### 1-7. Gene Search 사용 방법

특정 keyword 관련 유전자를 검색하고 싶을 때는 Gene search 창을 이용하면 된다. 예를 들어 'insulin'을 검색하면 엑셀 Data Sheet에 'insulin' keyword를 포함하는 행만 필터링 하여 확인할 수 있다(그림 1-30).

ID	Gene symbol	B/A	C/A	C/B	p-value	Average of normalized data (log2)	Normalized data (log2)
8611	8609 IDE	0.790	1.015	1.284	0.143	0.887	0.054
8704	8702 IGF1	1.306	1.016	0.778	0.390	0.765	0.414
8705	8703 IGF1R	3.781	2.238	1.740	0.005	0.008	0.444
8706	8704 IGF2	0.853	1.217	1.414	0.488	0.001	0.000
8708	8706 IGF2BP1	0.992	1.036	1.045	0.910	0.667	0.356
8709	8707 IGF2BP2	0.929	0.917	0.980	0.829	0.688	0.961
8711	8709 IGF2BP3	0.709	2.566	3.628	0.507	0.029	0.001
8712	8710 IGF2R	2.639	1.131	1.263	0.006	0.000	0.105
8713	8711 IGFALS	0.914	0.957	1.047	0.330	0.635	0.383
8714	8712 IGFBP1	1.000	1.185	1.185	0.319	0.017	0.017
8715	8714 IGFBP2	0.640	1.147	1.792	0.512	0.000	0.000
8716	8714 IGFBP3	0.685	1.011	1.476	0.140	0.000	0.000
8717	8715 IGFBP4	0.675	1.411	2.104	0.298	0.000	0.000
8718	8716 IGFBP5	0.972	4.267	4.440	0.281	0.000	0.000
8719	8717 IGFBP6	1.059	4.809	4.541	0.755	0.000	0.000
8720	8718 IGFBP7	0.519	1.119	2.155	0.122	0.000	0.000
8722	8720 IGFBP11	1.000	0.985	0.985	0.997	0.857	0.784
8911	8909 INS	1.000	1.073	1.073	1.000	0.162	0.162
8914	8912 INSIG1	0.677	0.692	1.450	0.059	0.199	0.005
8915	8913 INSIG2	1.379	1.263	0.916	0.138	0.207	0.419
8916	8914 INS1	0.949	0.809	0.852	0.860	0.292	0.494
8917	8915 INS1L4	1.000	1.073	1.073	0.319	0.162	0.162
8918	8916 INS1L5	1.000	1.073	1.073	1.000	0.162	0.162
8919	8917 INS1L6	1.000	1.073	1.073	0.319	0.163	0.162
8920	8918 INS1M1	1.000	1.073	1.073	1.000	0.162	0.162
8921	8919 INS1M2	1.000	1.073	1.073	0.319	0.162	0.162
8922	8920 INS1R	0.712	0.642	0.902	0.031	0.001	0.329
8923	8921 INS1RR	1.000	1.073	1.073	0.319	0.162	0.162
9007	9005 IRS1	0.920	2.838	3.078	0.030	0.001	0.000
9008	9006 IRS2	0.628	1.367	2.178	0.023	0.067	0.000
9009	9007 IRS4	1.000	1.073	1.073	1.000	0.162	0.162
18316	18314 RFXP1	1.000	1.073	1.073	1.000	0.162	0.162
18317	18315 RFXP2	0.989	1.052	1.061	0.653	0.206	0.218
18318	18316 RFXP3	1.000	1.073	1.073	1.000	0.162	0.162

그림 1-30. Genes related to insulin

## 2. Functional Annotation Analysis (DAVID, ExDEGA GraphicPlus)

### 2-1. DAVID 분석 틀을 이용한 Functional Annotation 분석

DAVID는 다양한 데이터 베이스를 기반으로 유전자의 상관관계를 통계적으로 분석하여 유전자의 주요 기능을 예측하는 analysis tool이다. 분석 과정은 그림 2-1 과 같다.

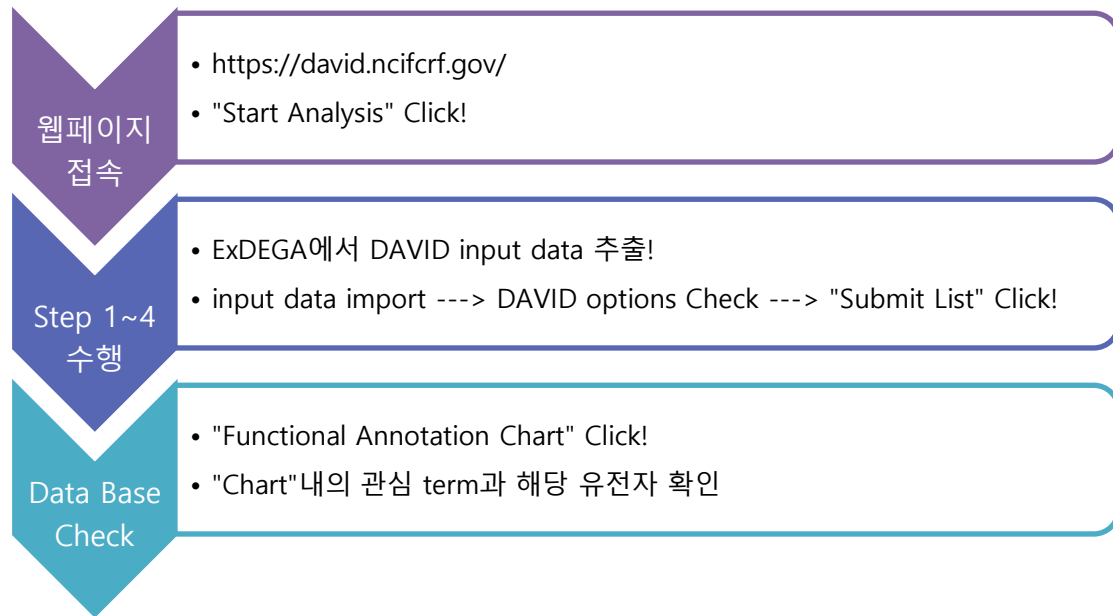


그림 2-1. DAVID tool analysis process

RNA-Seq 결과에서 significant gene을 선별하여 DAVID 분석을 한다. Significant gene selection 에서 Fold change, Normalized Data(log2), p-value (반복실험의 경우) 값을 지정하고, 확인하고자 하는 Fold change 조합을 선택하여 필터를 적용한다. 필터를 적용하여 선별된 유전자를 대상으로 Third Party Support를 통해 DAVID input을 추출하여, DAVID 분석에 사용한다. 반드시 하나의 비교 조합만 진행해야 한다(그림 2-2).

Filter: 269		Fold change			p-value			Average of normalized data (log2)			Normalized data (log2)					
ID	Gene symbol	B/A	C/A	C/B	B/A	C/A	C/B	A	B	C	A1	A2	A3	B1	B2	B3
368	ADGRE3	4.134	4.205	1.018	0.015	0.000	0.927	3.071	5.118	5.144	2.528	3.332	3.228	5.482	5.188	4.526
609	ALDH2	0.945	0.910	0.608	0.002	0.001	0.026	5.497	3.960	3.243	5.552	5.264	5.648	3.824	3.729	4.247
493	AMAC1	2.812	2.683	1.020	0.006	0.005	0.794	6.065	7.450	7.478	6.884	5.223	5.972	7.408	7.360	7.574
868	ANTXR2	2.282	2.420	1.060	0.021	0.000	0.718	3.081	4.271	4.356	3.012	3.270	2.940	4.419	4.519	3.767
932	APBB1P	2.711	2.745	1.013	0.005	0.001	0.904	4.910	6.349	6.367	5.339	4.682	4.590	6.524	6.411	6.076
976	APC8R	2.383	3.461	1.453	0.040	0.000	0.075	4.947	6.200	6.739	5.022	4.423	5.270	6.057	5.777	6.832
1021	ADP9	2.275	2.828	1.243	0.032	0.003	0.183	6.151	7.337	7.651	5.152	6.504	6.441	7.573	7.453	6.900
1066	ARHGAP25	3.908	2.757	0.780	0.002	0.001	0.088	4.612	6.461	6.075	4.307	4.937	4.519	6.552	6.615	6.179
1089	ARHGAP9	2.176	2.448	1.123	0.000	0.000	0.094	5.840	6.562	6.729	5.342	5.358	5.604	6.608	6.461	6.816
1112	ARRDC3	2.075	1.948	0.935	0.018	0.020	0.460	3.877	4.930	4.838	2.927	4.232	4.143	4.909	5.091	4.772
1352	ATH1	2.111	1.310	0.621	0.040	0.118	0.082	3.591	4.669	3.981	3.128	3.666	3.880	4.858	4.912	4.106
1387	ATP1B3	0.494	0.412	0.834	0.010	0.002	0.323	4.640	3.622	3.360	4.787	4.724	4.375	3.165	3.700	3.905
1439	ATP9V1B2	2.198	1.355	0.618	0.032	0.384	0.021	5.566	6.702	6.008	6.353	4.844	4.984	6.729	6.765	6.607
1533	BGNNT8	2.471	3.434	1.390	0.036	0.001	0.085	2.894	4.198	4.674	1.861	3.259	3.184	3.985	3.903	4.602
1594	BASP1	2.908	3.852	1.296	0.005	0.002	0.025	6.344	7.884	8.212	5.279	6.300	6.977	8.027	7.776	7.837
1634	BCKDHA	0.455	0.499	0.899	0.040	0.011	0.254	4.533	3.398	3.244	4.988	4.142	4.359	3.213	3.550	3.411
1652	BCLB	2.368	2.622	1.107	0.047	0.000	0.633	4.299	5.783	5.890	4.367	4.533	4.588	6.175	5.749	5.313
1726	BIRC3	0.476	0.578	1.205	0.002	0.000	0.255	4.087	3.017	3.292	4.058	4.091	4.112	3.096	3.313	2.538
1751	BLVRA	0.933	0.940	0.937	0.949	5.372	3.783	3.662	5.907	4.990	5.022	3.622	3.880	3.880	3.775	
1975	C11orf68	2.542	1.699	0.669	0.000	0.002	0.003	3.074	4.420	3.839	2.911	3.229	3.067	4.308	4.420	4.524
2058	C16orf54	2.830	1.923	0.679	0.018	0.021	0.097	4.127	5.628	5.070	3.218	4.339	4.516	5.573	5.247	5.972
2348	CSAR2	2.411	2.669	1.107	0.004	0.000	0.318	3.056	4.325	4.472	2.665	3.156	3.276	4.506	4.351	4.089
2620	CANT1	2.312	2.566	1.110	0.002	0.000	0.278	3.201	4.410	4.561	3.205	3.266	3.130	4.434	4.576	4.155
2633	CAFN2	0.450	0.735	1.633	0.001	0.022	0.003	5.603	4.452	5.159	5.783	5.489	5.322	4.425	4.554	4.369
2659	CARD6-AS1	2.393	1.645	0.688	0.002	0.018	0.005	2.791	4.049	3.509	3.003	2.974	2.289	4.077	4.151	3.910
2709	CAS7	0.488	0.662	1.353	0.023	0.074	0.019	5.721	4.688	5.125	6.072	5.473	5.540	4.886	4.536	4.619
2981	CCNH	0.490	0.677	1.677	0.019	0.112	0.001	4.239	2.931	3.677	3.772	4.580	4.253	2.988	2.877	2.925
2987	CENL1	0.492	0.620	1.261	0.006	0.006	0.159	4.814	3.791	4.125	4.947	4.856	4.621	4.044	3.908	3.321
2997	CCPG1	2.030	2.792	1.367	0.008	0.000	0.032	3.587	4.622	5.068	3.746	3.547	3.542	4.855	4.607	4.363
3031	CD183	0.418	0.566	0.778	0.040	0.022	0.000	4.033	2.773	1.864	4.487	3.656	3.817	2.611	2.803	2.851
3080	CD4	0.444	0.566	0.711	0.010	0.001	0.002	6.334	5.162	2.956	6.610	6.278	6.059	5.391	5.181	
3084	CD46	2.013	2.352	1.171	0.008	0.000	0.176	5.042	6.051	6.278	5.279	4.959	4.852	6.178	6.181	5.754

그림 2-2. DAVID input file generation process

DAVID 홈페이지는 그림 2-2의 "DAVID Web"을 클릭하여 접속하거나 웹사이트 (<https://david.ncicrf.gov/home.jsp>)를 검색하여 직접 접속한다.

\* 주의사항 : internet explorer 를 이용할 경우 다른 이름으로 저장 버튼이 보이지 않기 때문에, Chrome 을 이용하여 분석하기를 권장한다.

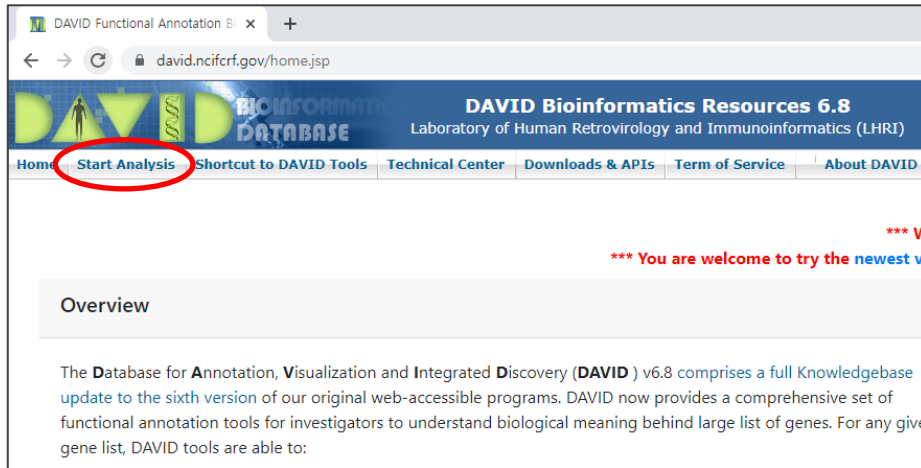


그림 2-3. DAVID tool webpage

"Upload" 탭에서 Step 1 에서 Step 4 까지 수행한다(그림 2-4). Step 1 에서 ExDEGA 에서 제작한 DAVID input 파일을 선택한다. Step 2 에서 "OFFICIAL\_GENE\_SYMBOL"를 선택한다. 만약 step 1 에서 Gene Bank No.를 넣었다면 "GENEBANK\_ACCESSION"을 선택한다. Step 2a 에서 분석하는 종의 학명을 입력한다. Step 3 에서 "Gene List"를 체크하고 Step 4 에서 "Submit List"를 누른다.

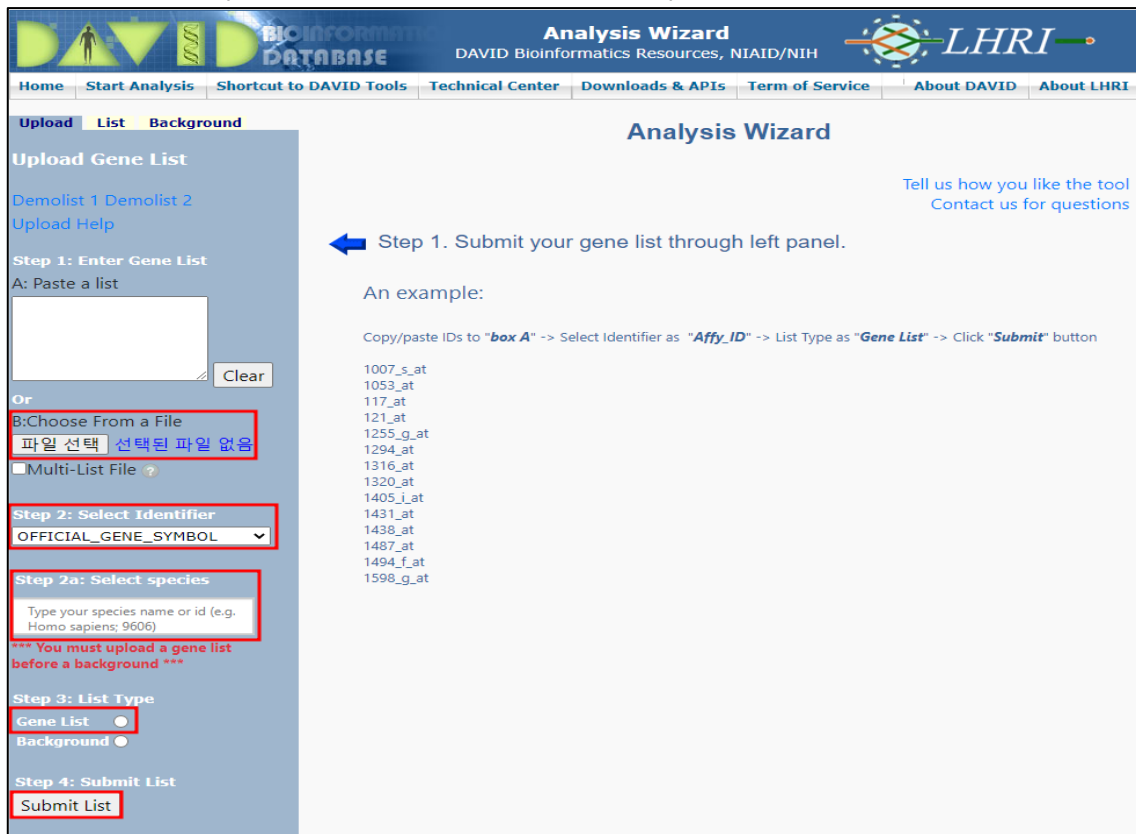


그림 2-4. DAVID tool : Step 1 ~ Step 4

List sheet 에서 분석하고자 하는 종을 선택한다(그림 2-5A). "List" Sheet 에서 해당 종(숫자)로 표기되어 있고 가로 안의 숫자가 분석에 적용된 유전자의 개수이다. 예시에서는 269 개의 유전자 리스트를 넣었고 데이터베이스에서 기능이 밝혀진 267 개만이 Functional Annotation 분석에 이용되었다는 의미이다.

만약 Current Background 에 분석하고자 하는 종이 아닌 다른 종이 나왔다면 좌측 "Background" Sheet 에서 알맞은 종을 선택하여 "Use"를 클릭한다(그림 2-5B).



그림 2-5. DAVID tool : Select Species

확인 후, 화면에서 Functional Annotation Tool 을 클릭하면 결과가 나온다(그림 2-6).



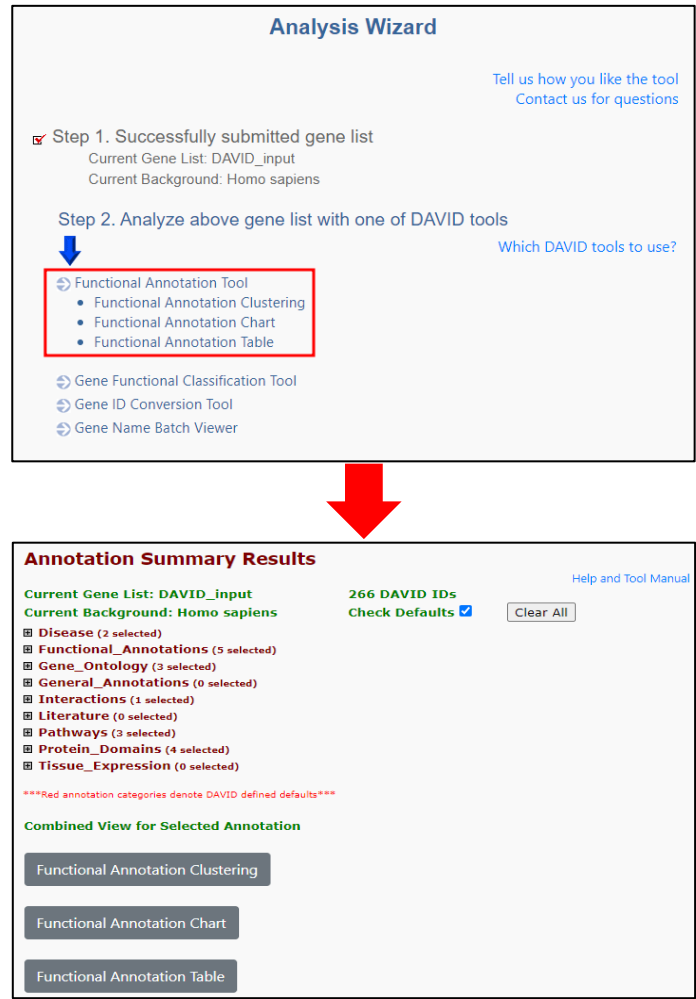


그림 2-6. DAVID results

DAVID 분석 결과 중 Gene Ontology Biological Process 결과를 확인하려면 "Gene\_Ontology"의 "+" 표시를 클릭하여 결과 창을 열고 "GOTERM\_BP\_Direct"의 "Chart"를 누른다(그림 2-7). Input 한 유전자들이 유의하게 관여하는 GO list 가 나온다. 관심 GO 를 클릭하면 QuickGO 데이터베이스로 연결되어 각 GO 의 정보를 확인할 수 있다. GO 의 Gene 막대를 클릭하면 해당 GO 관련 유전자들을 확인할 수 있다.

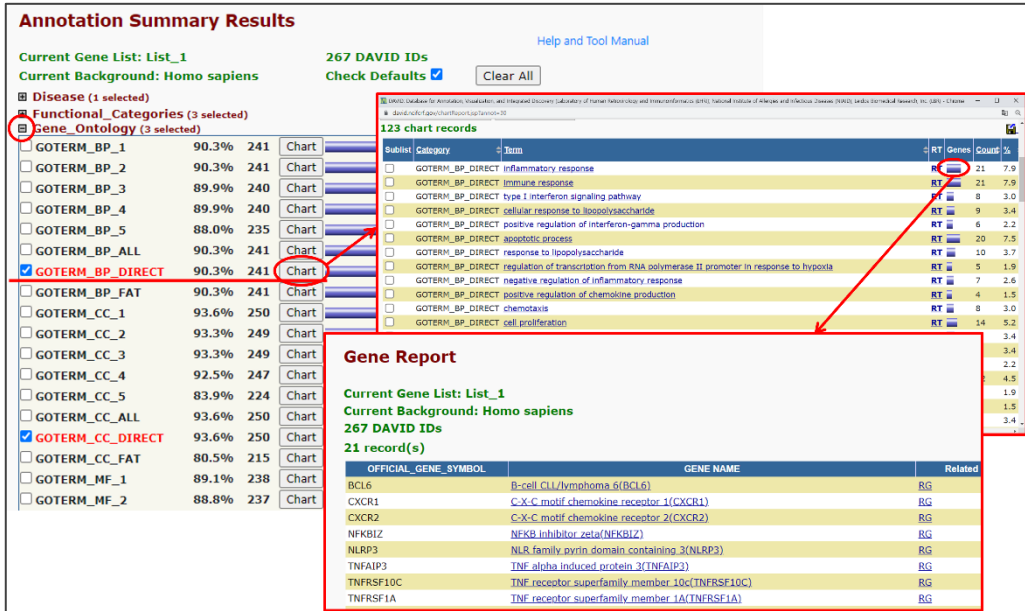


그림 2-7. DAVID tool : exploring Gene Ontology analysis result

이와 같은 방법으로 Pathway 결과를 확인해 보면 KEGG\_PATHWAY database 에서 주요 Pathway 가 나온다(그림 2-8). 각 pathway 를 누르면 pathway 그림을 확인할 수 있다. pathway 그림에서 별 표시가 되어 있는 유전자가 input 유전자 중 해당 pathway 에 관여하는 유전자이다. 유전자를 클릭하면 유전자 정보도 자세히 알 수 있다.

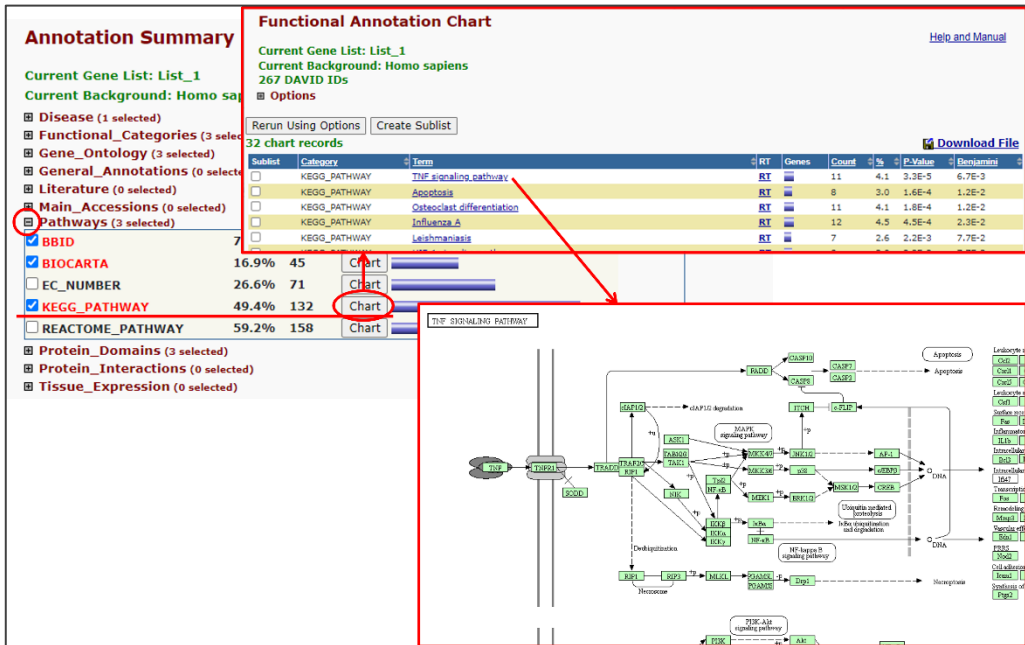


그림 2-8. DAVID tool : exploring Pathway analysis result

DAVID 분석은 input 한 유전자들이 유의하게 관련되는 GO, pathway 등을 분석하기에 유용한 tool 이다. 즉, input 한 유전자에서 많은 유전자들이 관련되는 GO, pathway 만 결과로 나오기 때문에 input 유전자 중 적은 수가 관련되는 GO, pathway 는 결과에 나오지 않는다. 또한 input 유전자의 수가 적으면 분석 결과가 없을 수도 있다.

위와 같이 GO, KEGG 등 각각의 chart 를 클릭하여 결과 데이터를 확인할 수 있고, ExDEGA GraphicPlus 용으로 "Functional Annotaiton Chart"를 클릭하여 전체 DAVID 결과를 확인할 수 있다(그림 2-9).

DAVID 분석 결과를 내 컴퓨터에 저장하려면, GO, KEGG 등 각각의 chart 를 클릭해서 나온 결과창(그림 2-7, 2-8) 또는 "Functional Annotaiton Chart"를 클릭해서 나온 결과창에서 Download File 링크를 마우스 오른쪽 버튼 클릭한 후 다른 이름으로 저장을 선택하면 DAVID Reults 파일을 다운로드 받을 수 있다(그림 2-10). 다운로드 받은 DAVID 결과 파일로 그래프 작성하는 방법은 '2-2. ExDEGA GraphicPlus 를 이용한 DAVID 결과 그래프 작성'에 설명되어 있다.

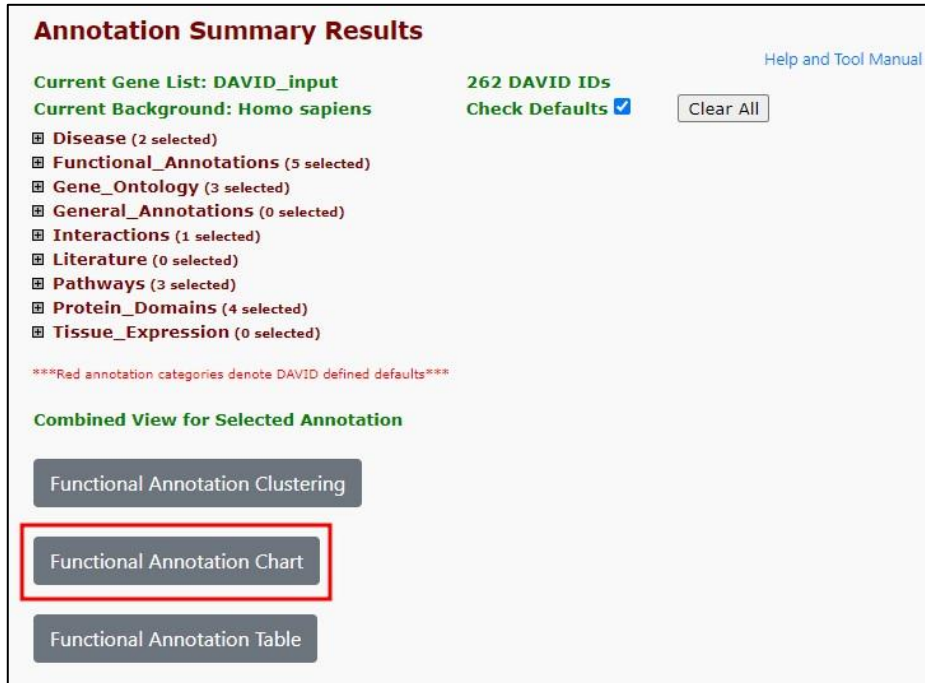


그림 2-9. DAVID All data Result

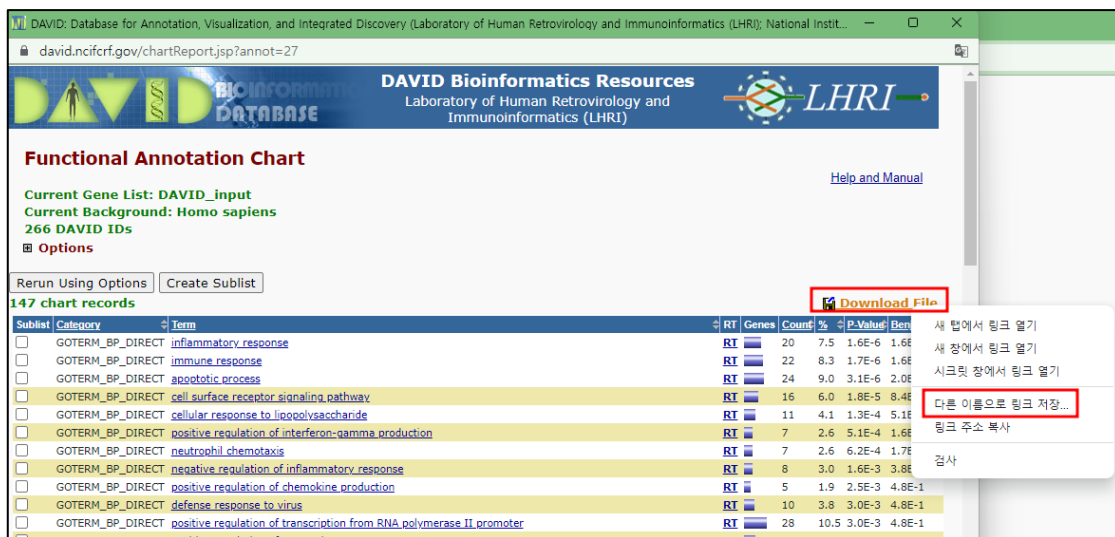


그림 2-10. DAVID data download

DAVID에서는 유전자 2 개 이상, EASE score 0.1 이하를 default 로 분석하여 이 기준에 적합한 결과를 보여준다. option 에서 이 기준을 조정하여 리스트를 더 볼 수 있다. DAVID 분석 결과의 각 항목은 DAVID 홈페이지의 Help and Tool Manual 에 자세히 설명되어 있다(그림 2-11).

### Annotation Summary Results

**Current Gene List:** List\_1

**Current Background:** Homo sapiens

- Disease (1 selected)
- Functional\_Categories (3 selected)
- Gene\_Ontology (3 selected)
- General Annotations (0 selected)
- Literature (0 selected)

**94 DAVID IDs**

Check Defaults  Clear All

[Help and Tool Manual](#)

### Functional Annotation Chart

Current Gene List: demolist1  
Current Background: Homo sapiens  
171 DAVID IDs

**Options**

Count Threshold  EASE Threshold  # of Records Displayed

Rerun Using Options Create Sublist [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">signal</a>	RT	<div style="width: 47%; height: 10px; background-color: #4a7ebb;"></div>	>47	27.5%	>3.0E-10
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">glycoprotein</a>	RT	<div style="width: 51%; height: 10px; background-color: #4a7ebb;"></div>	51	29.8%	4.9E-8
<input type="checkbox"/>	GOTERM_CC_ALL	<a href="#">extracellular region</a>	RT	<div style="width: 32%; height: 10px; background-color: #4a7ebb;"></div>	32	18.7%	1.1E-7
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">alternative splicing</a>	RT	<div style="width: 49%; height: 10px; background-color: #4a7ebb;"></div>	49	28.7%	6.4E-6
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">chromoprotein</a>	RT	<div style="width: 7%; height: 10px; background-color: #4a7ebb;"></div>	7	4.1%	1.1E-5
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">direct protein sequencing</a>	RT	<div style="width: 33%; height: 10px; background-color: #4a7ebb;"></div>	33	19.3%	1.2E-5
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">phosphorylation</a>	RT	<div style="width: 31%; height: 10px; background-color: #4a7ebb;"></div>	31	18.1%	1.6E-5
<input type="checkbox"/>	UP_SEQ_FEATURE	<a href="#">signal peptide</a>	RT	<div style="width: 47%; height: 10px; background-color: #4a7ebb;"></div>	47	27.5%	3.7E-5
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">metalloprotein</a>	RT	<div style="width: 8%; height: 10px; background-color: #4a7ebb;"></div>	8	4.7%	4.7E-5
<input type="checkbox"/>	GOTERM_BP_ALL	<a href="#">response to chemical stimulus</a>	RT	<div style="width: 14%; height: 10px; background-color: #4a7ebb;"></div>	14	8.2%	6.1E-5

그림 2-11. DAVID Help and Tool Manual

## 2-2. ExDEGA GraphicPlus 를 이용한 DAVID 결과 그래프 작성

ExDEGA 레포트의 ExDEGA Graphic Plus Start 버튼을 클릭하여 프로그램을 실행한다 (그림2-12).

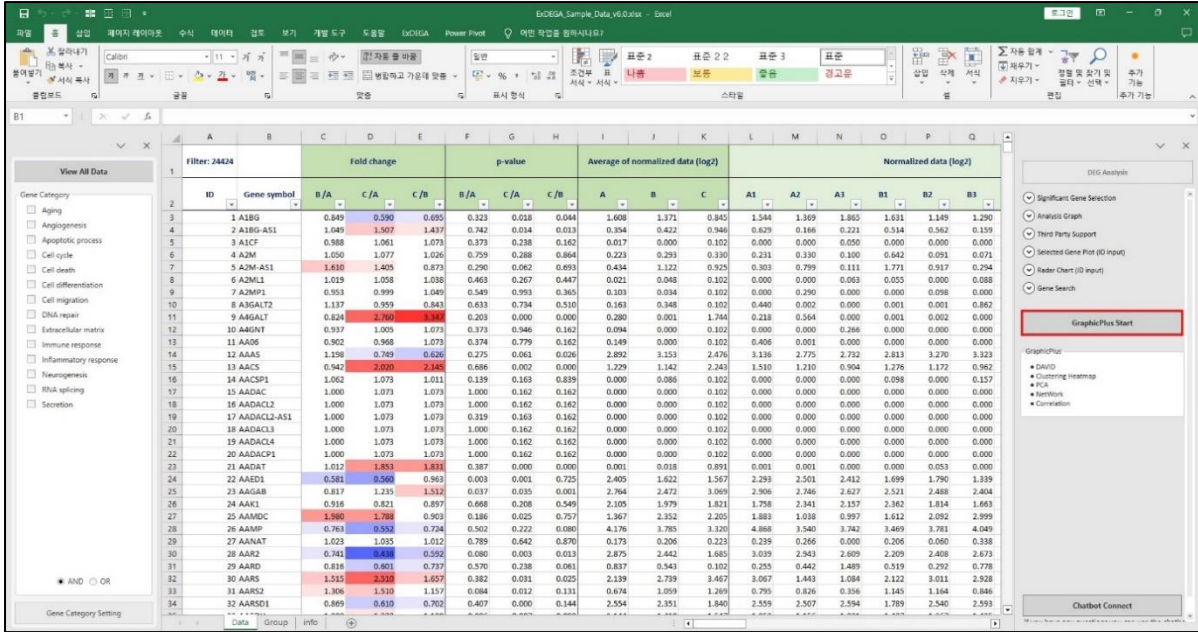


그림 2-12. Execute ExDEGA GraphicPlus

메인 화면의 4 개 탭 중 'DAVID' 탭에서 DAVID Graphic 분석을 수행할 수 있다. DAVID Graphic 분석 창은 그림 2-13 과 같다.

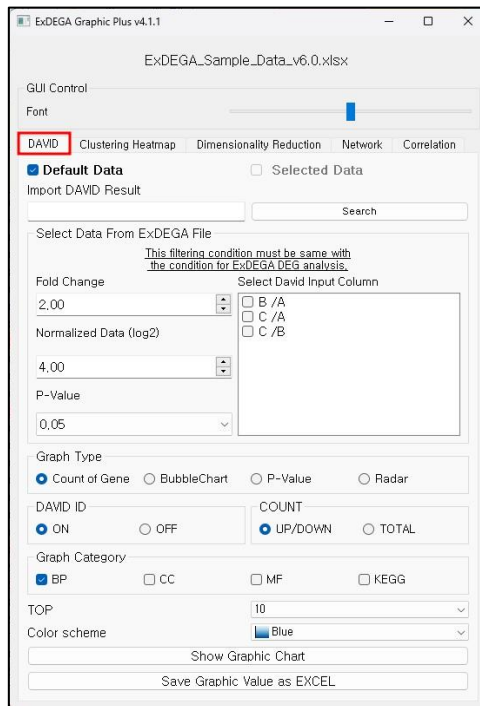


그림 2-13. Select DAVID tab to execute the DAVID Graphic Analysis

DAVID 분석 결과를 그래프로 제작하려면 DAVID result 파일과 DAVID input 이 필요하다. DAVID result 은 '2-1. DAVID 분석 틀을 이용한 Functional Annotation 분석'에서 저장한 DAVID

결과 파일이다. DAVID result 파일 안에 있는 내용은 그림 2-14 과 같다. 이 파일에 있는 Term, Count, P-value, Fold Enrichment 항목을 이용하여 그래프가 제작된다.

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_DIRECT	GO:0055114~oxidation-reduction process	36	16.43836	2.70E-14	D3YZE4, P5	203	676	18082	4.743580027	3.05E-11	3.05E-11	4.33E-11
GOTERM_BP_DIRECT	GO:0006631~fatty acid metabolic process	19	8.675799	1.50E-13	P04117, P5	203	156	18082	10.84874321	1.70E-10	8.48E-11	2.40E-10
GOTERM_BP_DIRECT	GO:0006635~fatty acid beta-oxidation	12	5.479452	1.43E-12	Q9DC50, C	203	44	18082	24.29287953	1.62E-09	5.40E-10	2.30E-09
GOTERM_BP_DIRECT	GO:0008152~metabolic process	28	12.78539	2.10E-12	P19157, P5	203	463	18082	5.386758025	2.37E-09	5.93E-10	3.36E-09
GOTERM_BP_DIRECT	GO:0006629~lipid metabolic process	24	10.9589	1.96E-09	P51174, Q5	203	459	18082	4.657458386	2.22E-06	4.43E-07	3.14E-06
GOTERM_BP_DIRECT	GO:0006810~transport	48	21.91781	3.28E-08	P04117, P2	203	1822	18082	2.346622831	3.71E-05	6.18E-06	5.26E-05
GOTERM_BP_DIRECT	GO:0006637~acyl-CoA metabolic process	7	3.196347	1.05E-06	Q8VCT4, Q	203	31	18082	20.1134594	0.001191	1.70E-04	0.001689
GOTERM_BP_DIRECT	GO:0070527~platelet aggregation	7	3.196347	3.70E-06	Q9Z1Q5, P	203	38	18082	16.40834846	0.004178	5.23E-04	0.005935
GOTERM_BP_DIRECT	GO:0006754~ATP biosynthetic process	5	2.283105	1.13E-04	Q03265, D	203	23	18082	19.36388948	0.120415	0.014155	0.181743
GOTERM_BP_DIRECT	GO:0015671~oxygen transport	4	1.826484	1.56E-04	P02089, P0	203	10	18082	35.62955665	0.161483	0.017458	0.249389
GOTERM_BP_DIRECT	GO:0006749~glutathione metabolic process	6	2.739726	2.12E-04	P48774, P1	203	49	18082	10.90700714	0.213205	0.021563	0.339389
GOTERM_BP_DIRECT	GO:0051791~medium-chain fatty acid metabol	3	1.369863	3.70E-04	Q9DC50, C	203	3	18082	89.07389163	0.342104	0.034291	0.591879

그림 2-14. DAVID output file

우선, Import DAVID Result 에서 DAVID Result 파일을 선택한다. 그리고 아래 선별조건에서 DAVID 분석 시 사용된 input 파일을 만들 때 이용했던 조건과 동일하도록 ExDEGA 레포트에 적용한 Fold Change 값, Normalized Data (log2)값, P-Value 값 및 비교조합 선택을 세팅해야 한다(그림 2-15). Input 파일과 마찬가지로 비교조합은 1 개만 선택한다. 사용자가 별도의 기준을 적용하지 않을 경우, Fold Change 2.00, Normalized Data (log2) 4.00, P-Value 0.05 로 자동 적용된다.

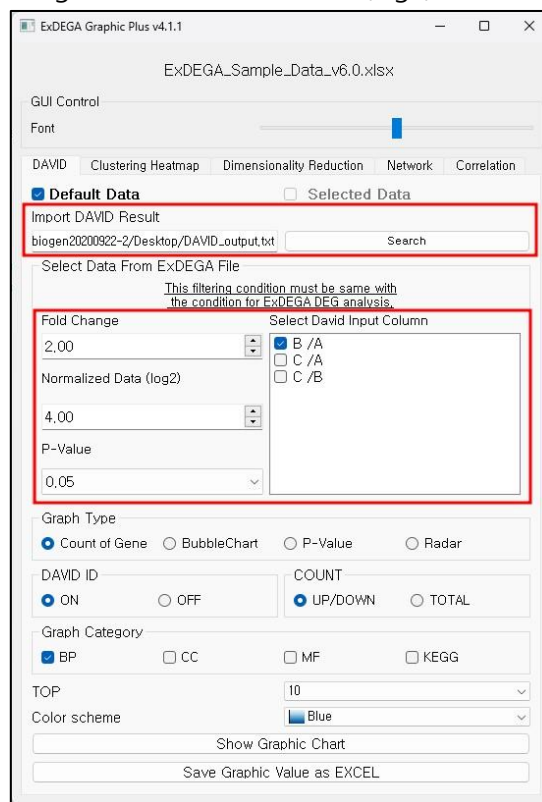


그림 2-15. Steps for Create DAVID Analysis Graph

위의 과정을 거친 뒤, Graph Type 에서 4 종류의 그래프를 제작할 수 있다(그림 2-16).

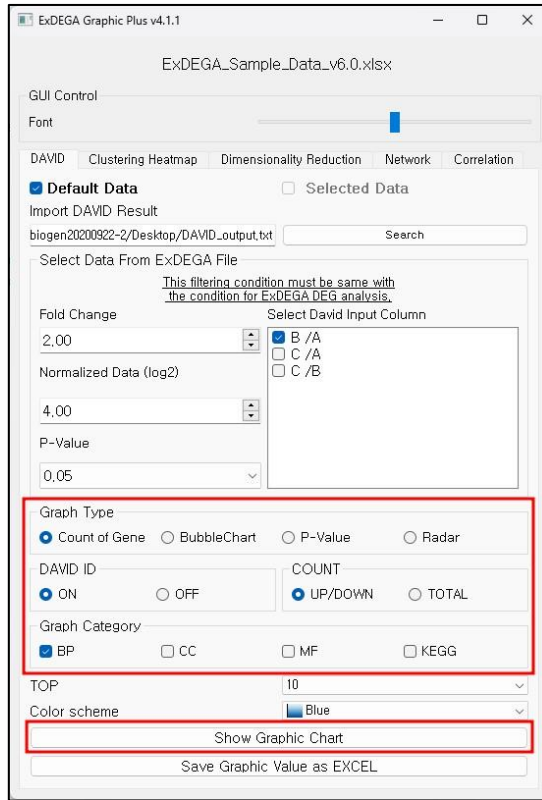
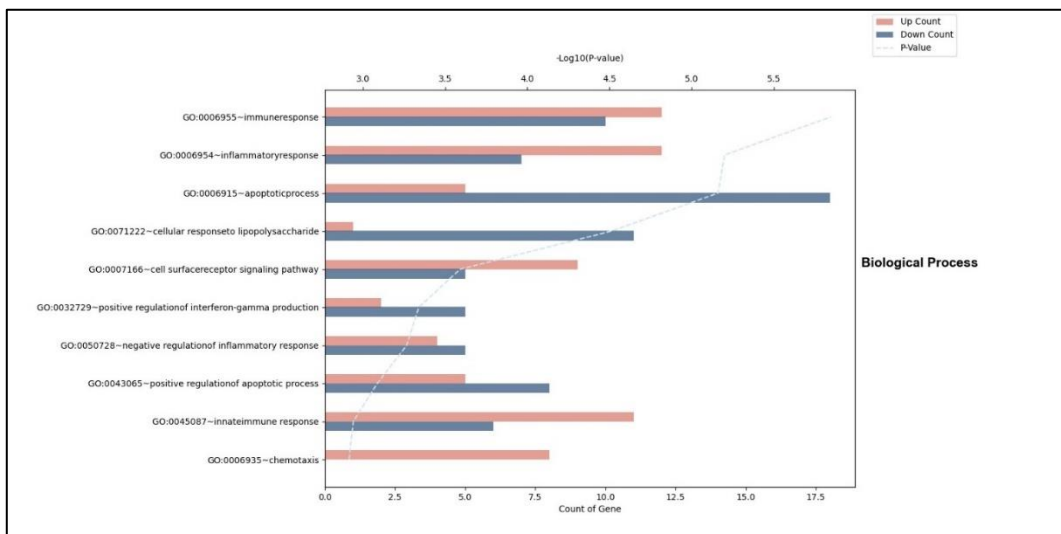


그림 2-16. Result of Graphic Chart

Count of Gene 을 선택하고 Show Graphic Chart 를 누르면 각 GO (or pathway)에서 발현이 증가하는 유전자, 감소하는 유전자 수가 그래프로 작성된다(그림 2-17). 그림 2-16 의 COUNT 항목은 해당 그래프에서 증가, 감소하는 유전자 개수를 나누어 표현할지, 구분없이 합쳐서 표현할지 선택할 수 있다.



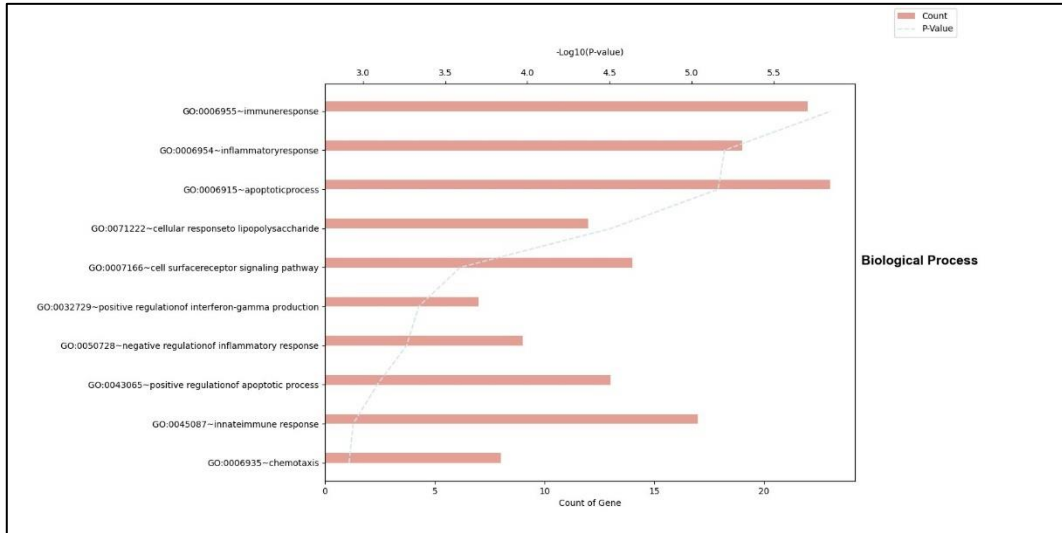


그림 2-17. Result of Graphic Chart (Count of Gene)

Bubble chart를 선택하고 Show Graphic Chart를 누르면 각 GO (or pathway)의 p-value, Fold enrichment 값으로 그래프가 작성된다 (그림2-18).

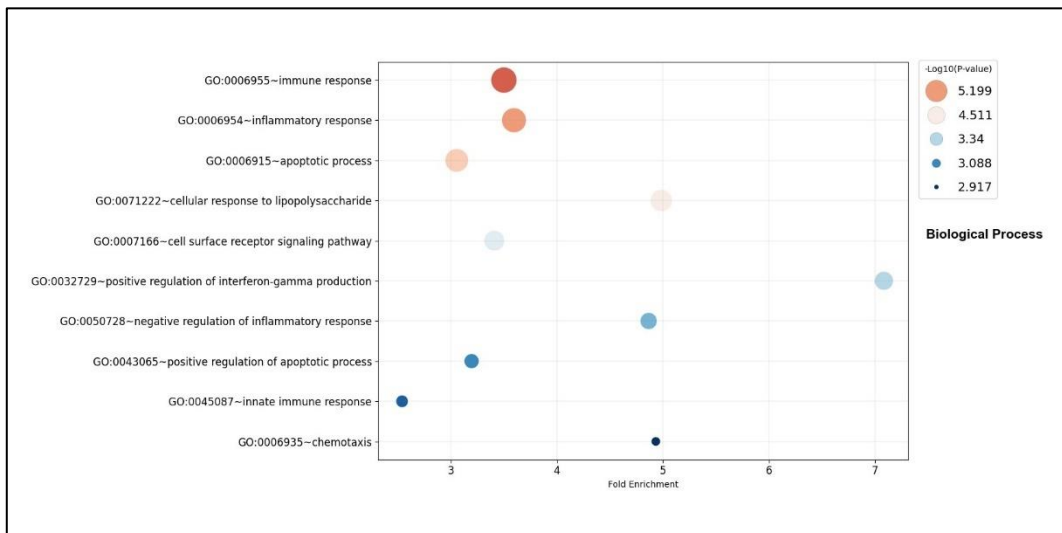


그림 2-18. Result of Graphic Chart (Fold Enrichment)



P-value 를 선택하고 Show Graphic chart 를 클릭하면 각 Term 마다 해당하는 p-value 값으로 막대 그래프가 작성된다(그림 2-19).

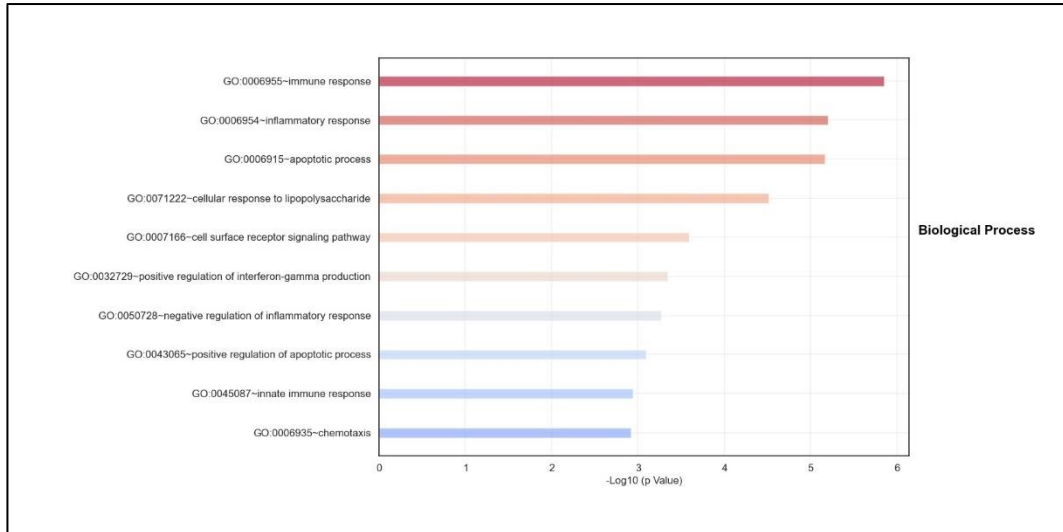


그림 2-19. Result of Graphic Chart (p-value)

Radar를 선택하고 Show Graphic Chart를 누르면 각 GO (or pathway)의 Fold enrichment 값으로 그래프가 작성된다(그림2-20).

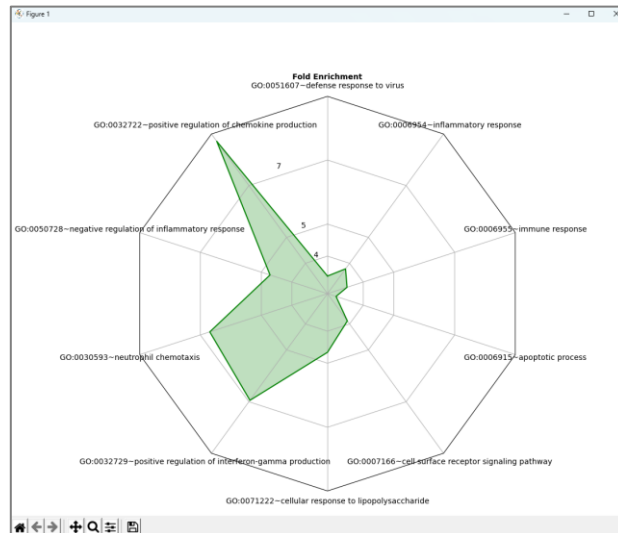


그림 2-20. Result of Graphic Chart (Radar)

DAVID ID에서 ON을 선택하면 DAVID의 고유번호가 앞에 붙고, OFF를 선택하면 고유번호 없이 Term만 출력된다. Graph Category는 input한 DAVID 분석 결과 데이터 내에서 해당하는 Category로 제작되며, 중복으로 선택할 수 있다(그림 2-16).

TOP 에서는 그래프 제작에 사용될 결과 리스트의 수를 선택할 수 있다. 이는 선택한 비교조합이 DAVID result 파일의 결과값과 함께 계산되어 DAVID 분석 그래프를 그릴 때 DAVID result 파일에서 상위 n 개의 리스트를 대상으로 그래프를 제작하는 옵션이다. 2 개에서 40 개

까지의 옵션을 선택 가능하며, 사용자가 별도의 기준을 적용하지 않을 경우 이 값은 자동으로 10 으로 설정된다. 그리고 Color scheme 을 이용하여 다른 색상을 적용시킬 수 있다.

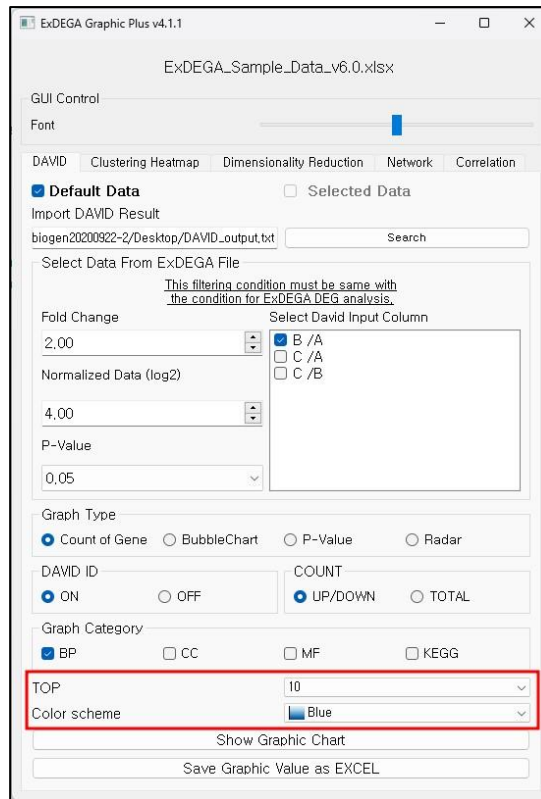


그림 2-21. Result of Graphic Chart (TOP, Color scheme)

예시로 Type 은 Count of Gene, Graph Category 의 항목은 전부 선택하고, TOP 을 40 으로 설정했을 때는 아래 그림과 같이 제작된다(그림 2-22).

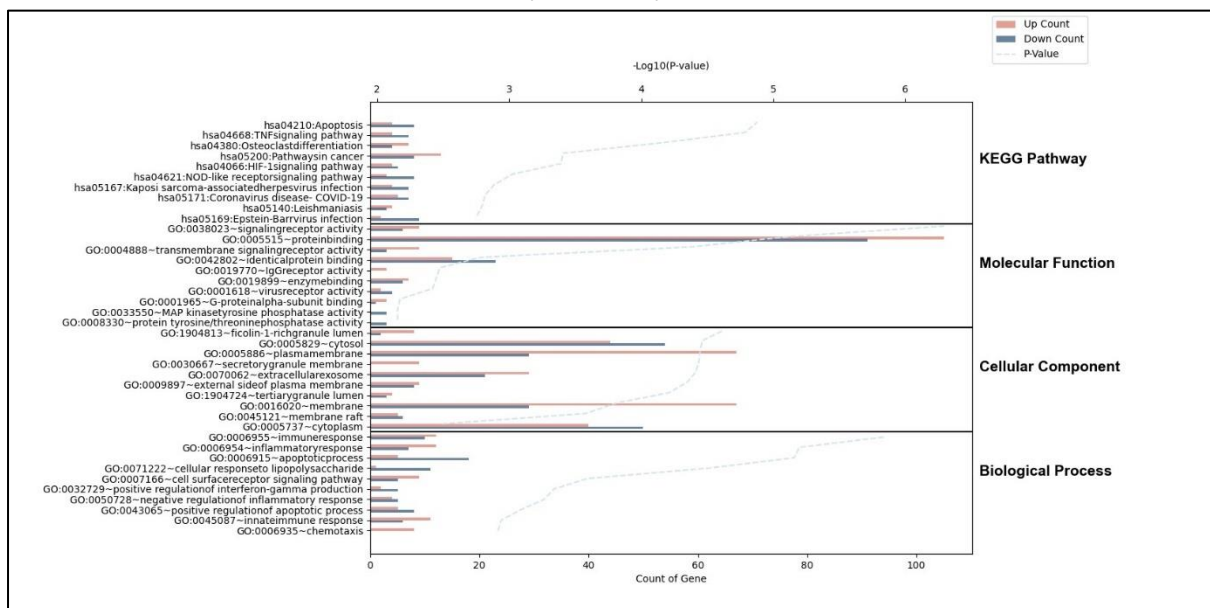


그림 2-22. Result of Graphic Chart



### 3. Clustering heatmap analysis (ExDEGA GraphicPlus)

Hierarchical Clustering Heatmap은 연구자가 선택한 유전자의 발현 유사성을 기반으로 Sample 간의 유사성, 유전자 간의 유사성을 판단할 때 사용한다. 본 챕터에서는 Heatmap (개별 색상의 직사각형 데이터 행렬)과 Dendrogram (계층적 클러스터링)을 합쳐 Hierarchical Clustering Heatmap을 그리는 방법을 설명한다.

Hierarchical Clustering Heatmap은 ExDEGA GraphicPlus를 이용하여 분석할 수 있다. ExDEGA GraphicPlus를 이용하기 위해서 먼저 해당 프로그램을 연다. ExDEGA 레포트의 ExDEGA GraphicPlus Start 버튼을 클릭하여 프로그램을 동작시킨 뒤, Clustering Heatmap 탭을 클릭하여 준비한다(그림 3-1).

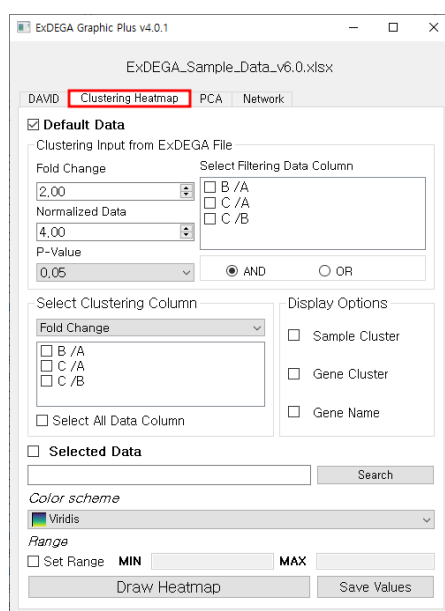


그림 3-1. Select Clustering Heatmap tab

Clustering Heatmap을 제작을 위해 input Data는 크게 두가지 형식으로 이용할 수 있다. GraphicPlus가 인식한 ExDEGA report의 전체 data를 이용하는 Default Data와 Third Party Support에서 export한 Selected Data가 있다.

Default Data를 이용하는 경우, 1~7의 과정을 거쳐 Clustering Heatmap을 그리게 된다.(그림 3-2) 1에서는 인식된 ExDEGA 레포트를 바탕으로, DEG 분석 기준 및 DEG 분석 기준을 적용할 Fold Change 샘플 그룹을 지정하게 된다. 사용자가 별도의 기준을 적용하지 않을 경우, Fold Change 2.00, Normalized Data 4.00, P-Value 0.05로 자동 적용되며 Fold Change 샘플 그룹은 AND/OR 기능을 통해 여러 개의 샘플에 기준을 복수 적용할 수 있다.

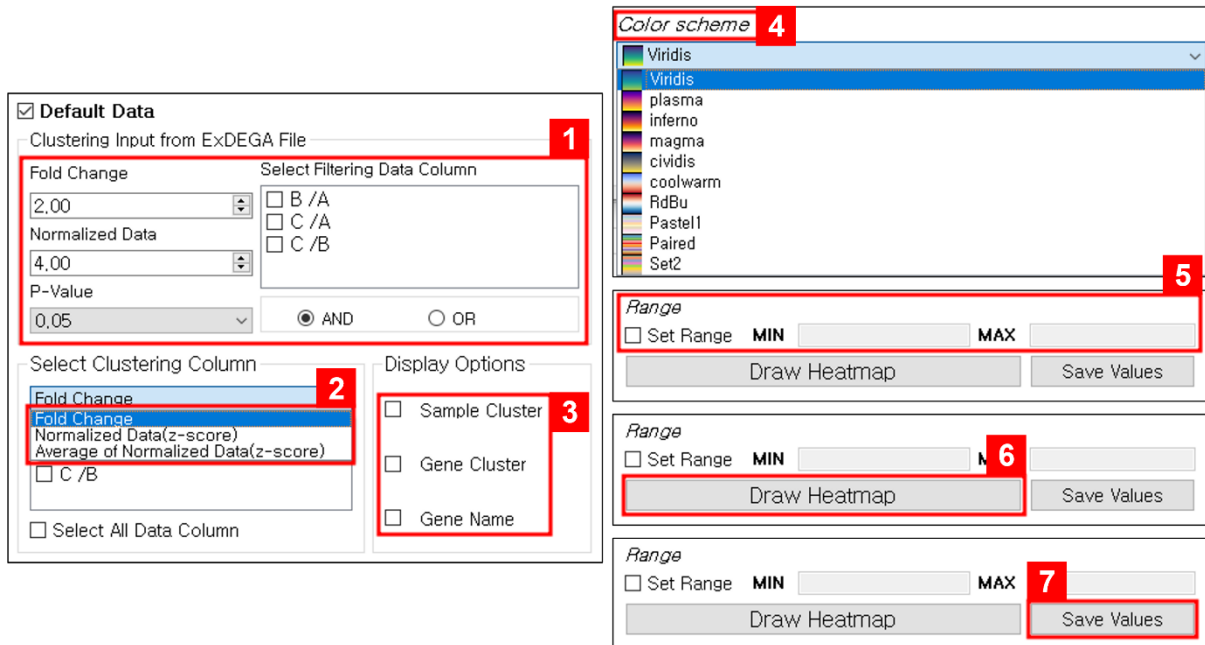


그림 3-2. Steps for Create Clustering Heatmap

2에서는 1에서 적용한 기준을 바탕으로, 실제 Clustering Heatmap을 그리고자 하는 데이터 컬럼을 선택한다. ExDEGA 레포트 유형(Group, Single)에 따라 Group 레포트에서는 Fold Change, Normalized Data(z-score), Average of Normalized Data(z-score) 타입이, Single 레포트에서는 Fold Change, Normalized Data(z-score) 타입이 제공된다. 반영하고자 하는 타입을 선택한 후, 그림 3-3와 같이 타입에 해당하는 샘플 그룹을 선정할 수 있다. 이 때, 하단의 'Select All Data Column'을 선택하면 해당하는 타입에 존재하는 모든 샘플 그룹을 동시에 선택/선택 해제할 수 있다.

단, 사용자가 Normalized Data(z-score), Average of Normalized Data(z-score) 옵션을 선택하더라도, 3에서 고른 샘플 그룹의 수가 2개 이하인 경우 z-score가 적용되지 않고 ExDEGA 레포트에 표현된 데이터 값을 그대로 적용하여 Clustering heatmap이 작성되니 이 점을 유의해야 한다.

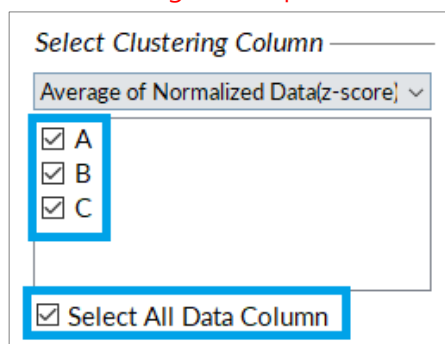


그림 3-3. Select specific sample group for composition of clustering heatmap

3에서는 Clustering Heatmap 옵션에 적용할 Display option을 선택할 수 있다(그림 3-4). Sample cluster를 선택하면 발현이 유사한 비교조합 또는 샘플 간의 dendrogram이 작성된다. Gene cluster를 선택하면 발현이 유전자 간의 dendrogram이 작성된다. Gene name을 표시하면 Raw에 해당하는 gene symbol이 표시된다. 단, 입력한 유전자가 80개 이상일 경우에는 gene symbol을 표시할 수 없다.

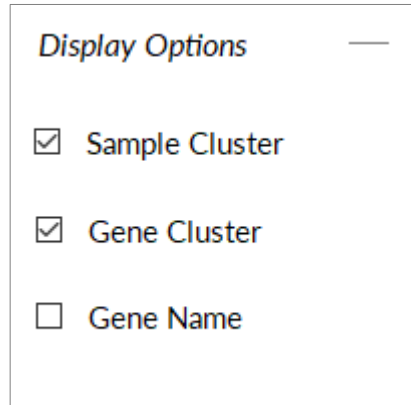


그림 3-4. Set a Display option for clustering heatmap

4에서는 Color scheme는 Heatmap의 색상을 설정할 수 있다. 10가지의 기본 옵션을 제공한다. 사용자가 별도로 옵션을 선택하지 않을 경우, 'Viridis' 옵션이 기본으로 설정된다.

5의 Set Range는 데이터 표현 범위를 설정하는 것으로 선택하지 않으면 input file의 최소값과 최대값으로 자동 설정된다. Set Range 앞의 체크박스를 선택하면 MIN(최소값), MAX(최대값) 입력 박스가 활성화되며, 사용자가 원하는 최소값/최대값을 설정할 수 있다(그림 3-5).



그림 3-5. Set a min-max option for clustering heatmap

6의 Draw Heatmap 버튼을 누르면 Hierarchical Clustering Heatmap 결과 창이 생성된다(그림3-6). 각 heatmap의 위쪽에 표시된 dendrogram은 비교조합 또는 샘플 간의 발현 유사성(sample cluster)을 표시한 결과이다. 왼쪽에 표시된 dendrogram은 유전자 간의 발현 유사성(gene cluster)을 표시한 결과이다. 가깝게 묶일수록 발현이 유사한 것이다.

또한 그림 3-6. a heatmap은 정제한 전체 발현값을 대상으로 제작된 heatmap이며, 그림 3-6. b heatmap은 min-max 값을 적용하여 제작된 heatmap이다. (각 그래프 왼쪽 상단의 legend 값에서 차이점을 확인할 수 있다.)

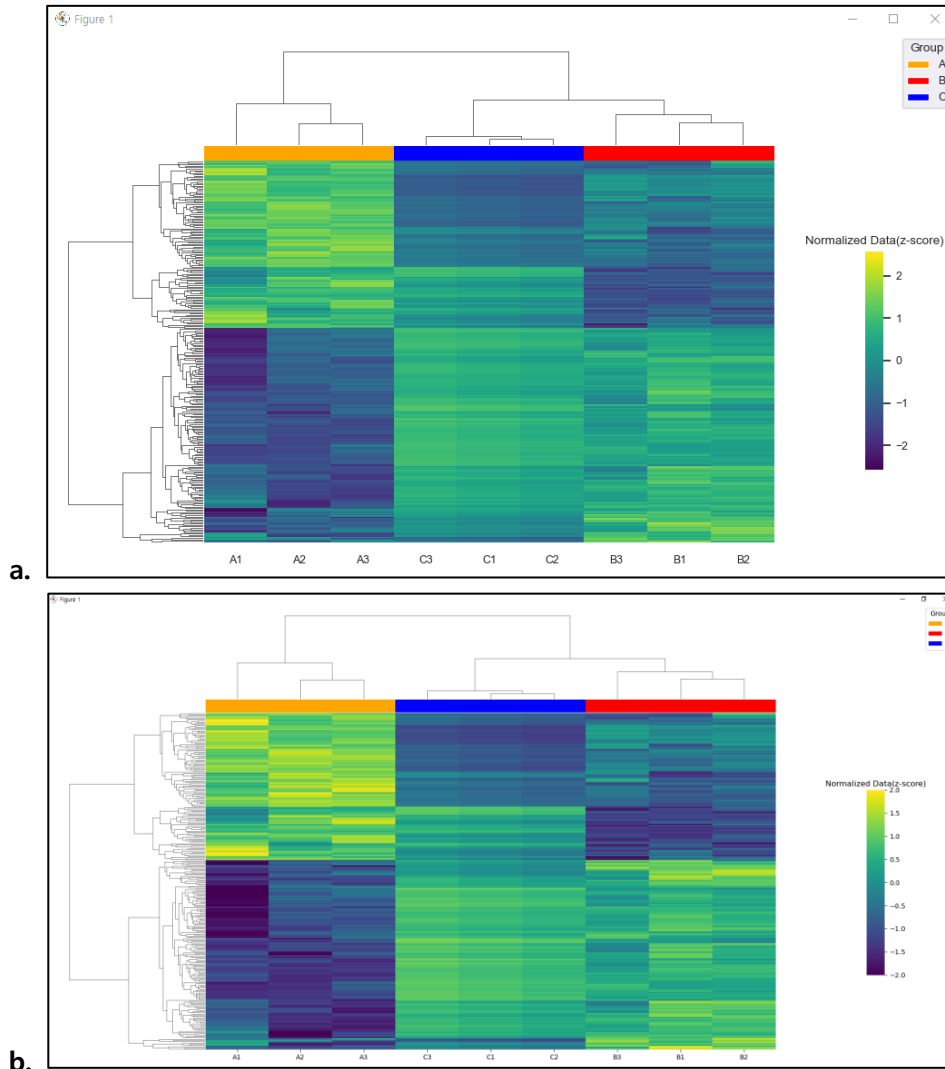


그림 3-6. Clustering heatmap result

7의 Save Values 버튼을 누르면 Clustering Heatmap을 만들 때 사용한 value값이 엑셀로 저장된다. 이 때, Save Values는 Clustering Heatmap을 제작하지 않았더라도 필터링 기준, 필터링 기준을 적용할 샘플그룹, Clustering Heatmap을 적용할 샘플 그룹들을 선택하면 결과값을 받아볼 수 있다. 각각의 값을 지정하고 Save Values 버튼을 누르는 경우, 조건을 확인하기 위한 결과창이 뜨며(그림 3-7) 이 결과창에서 'OK'를 누를 때 값이 사용자가 정한 위치에 .xlsx 형식으로 저장된다.

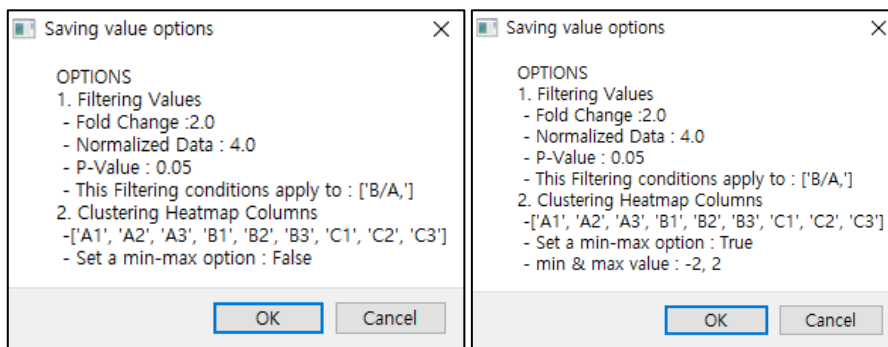


그림 3-7. Confirm all conditions for saving values used to create clustering heatmap

저장된 파일을 확인하면, 사용자가 지정한 Clustering Heatmap을 적용할 샘플 그룹에 대해서 Clustering Heatmap을 그릴 때 사용한 값이 저장되어 있다. 또한, 사용자가 MIN, MAX 옵션을 사용한 경우 최소값은 MIN 값으로, 최대값은 MAX값으로 대체되어 저장되므로 유의해야 한다.(그림 3-8, 그림에 노란색으로 표시된 부분은 최소값/최대값이 사용자가 지정한 값으로 대체되어 저장된 것을 나타내었음.)

	A	B	C	D	E
1		Gene Symbol	B/A	C/A	C/B
2	0	ADGRE3	2.04741024	2.073376243	0.025966003
3	1	ALDH2	-1.537049041	-2.253850295	-0.716801254
4	2	AMICA1	1.385026346	1.413092963	0.028066616
5	3	ANTXR2	1.190472629	1.27486025	0.084387621
6	4	APBB1IP	1.438668675	1.456716586	0.018047911
7	5	APOBR	1.252926787	1.792222798	0.539296011
8	6	AQP9	1.185878944	1.4995746	0.313695656
9	7	ARHGAP25	1.849565549	1.463314624	-0.386250925
10	8	ARHGAP9	1.121902679	1.289058964	0.167156285
11	9	ATP1B3	-1.017592617	-1.279867725	-0.262275108
12	10	B3GNT8	1.304915258	1.78003738	0.475122122
13	11	BASP1	1.540191899	1.86853178	0.328339881
14	12	BCKDHA	-1.135843533	-1.289436933	-0.1535934
15	13	BCL6	1.243655421	1.390589701	0.14693428
16	14	BLVRA	-1.609398178	-1.710154094	-0.100755915
17	15	C5AR2	1.269735395	1.416122499	0.146387104
18	16	CANT1	1.208913771	1.359269542	0.150355772
19	17	CCPG1	1.035597227	1.481048011	0.445450784
20	18	CD163	-1.259950238	-2.668497509	-1.408547271
21	19	CD4	-1.172118667	-3.37724913	-2.205606246
22	20	CD46	1.009085765	1.236811337	0.227725572
23	21	CD68	-1.475204601	-2.556510409	-1.081305808
24	22	CD69	-3.489771037	-4.496600244	-1.006829207
			-1.077319481	-2.915110714	-1.837791232
			1.966743553	1.804878193	-0.16186536
			-1.843939211	-1.352696295	0.491242916

	A	B	C	D	E
1		Gene Symbol	B/A	C/A	C/B
2	0	ADGRE3	2	2	0.025966003
3	1	ALDH2	-1.537049041	-2	-0.716801254
4	2	AMICA1	1.385026346	1.413092963	0.028066616
5	3	ANTXR2	1.190472629	1.27486025	0.084387621
6	4	APBB1IP	1.438668675	1.456716586	0.018047911
7	5	APOBR	1.252926787	1.792222798	0.539296011
8	6	AQP9	1.185878944	1.4995746	0.313695656
9	7	ARHGAP25	1.849565549	1.463314624	-0.386250925
10	8	ARHGAP9	1.121902679	1.289058964	0.167156285
11	9	ATP1B3	-1.017592617	-1.279867725	-0.262275108
12	10	B3GNT8	1.304915258	1.78003738	0.475122122
13	11	BASP1	1.540191899	1.86853178	0.328339881
14	12	BCKDHA	-1.135843533	-1.289436933	-0.1535934
15	13	BCL6	1.243655421	1.390589701	0.14693428
16	14	BLVRA	-1.609398178	-1.710154094	-0.100755915
17	15	C5AR2	1.269735395	1.416122499	0.146387104
18	16	CANT1	1.208913771	1.359269542	0.150355772
19	17	CCPG1	1.035597227	1.481048011	0.445450784
20	18	CD163	-1.259950238	-2	-1.408547271
21	19	CD4	-1.172118667	-2	-2
22	20	CD46	1.009085765	1.236811337	0.227725572
23	21	CD68	-1.475204601	-2	-1.081305808
24	22	CD69	-2	-2	-1.006829207
			-1.077319481	-2	-1.837791232
			1.966743553	1.804878193	-0.16186536
			-1.843939211	-1.352696295	0.491242916

그림 3-8. Examples of saved value file for clustering heatmap

(이 그림은 사용자의 이해를 돕기 위해 임의로 수정되었으며, 실제 저장된 파일에는 노란색 음영이 보이지 않는 것이 정상입니다.)

Selected Data를 이용하는 경우, ExDEGA report에서 Third Party Support를 통해 Export한 input 파일을 GraphicPlus에 Upload 한다(그림 3-9). 3~7의 과정을 거쳐 Clustering Heatmap을 그리게 된다(그림 3-2).

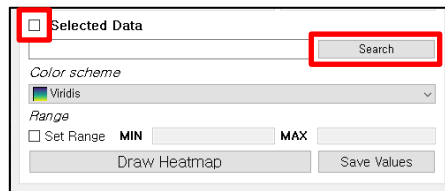


그림 3-9. Selected Data – input upload

Clustering heatmap은 MeV라는 프로그램을 이용하여 분석할 수도 있다. MeV 프로그램을 사용하여 Clustering heatmap을 작성하는 방법은 MeV manual ([Download Link](#))에서 확인할 수 있다.



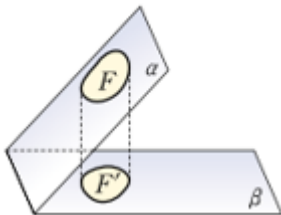
## 4. Principal component analysis (ExDEGA GraphicPlus)

본 챗터에서는 PCA 에 대한 이론과 PCA 2D/3D 를 그리는 방법에 대해 설명한다. PCA 는 Sample 간의 발현 유사성을 확인하기 위한 목적으로 Clustering Heatmap 과는 다르게 Sample 내의 유전자 전체 발현값을 기반으로 분석된다.

### 4-1. PCA (Princial Component Analysis) 이론

#### PCA Essential Description.

PCA 는 주성분 분석의 준말로 고차원 데이터를 정사영(구조 유지, 차원 감소) 시켜 저차원 데이터로 차원을 축소하는 알고리즘이다.



$F'$ 은  $F$ 의  $\beta$  위로의 정사영(=  $F'$ 은  $F$ 를  $\lambda$  만큼 표현한다)

PCA 는 앞서 설명한 바와 같이 Sample 에 속해 있는 전체 유전자 발현을 대상으로 Sample 간의 유사성을 확인하려는 목적으로 분석을 진행한다. Human RNA-seq 기준으로 각 Sample 에는 약 24,000 개 이상의 유전자가 포함된다. 24,000 개의 변수가 생긴다는 말과 동일하기 때문에 Sample 간의 유사성을 파악하기 힘들다. PCA 알고리즘을 통해 방향 변화 없이 variance 최대(선형 변환)가 되는  $\lambda$  값(eigenvalue)을 계산한다.  $\lambda$  값은 sample 의 개수만큼 나오게 되고 주성분 비율은  $\lambda$  값의 전체 합에서 해당  $\lambda$  값이 차지하는 비율이다.

예시 : i. PC1( 60%), PC2(30%) => 주성분 1, 2 로 데이터의 90% 표현

ii. PC1(50%), PC2(30%), PC3(10%) => 주성분 1,2,3 으로 데이터의 90% 표현

**PCA additional Info.**

다음은 공분산을 이용하여 PCA 를 계산하는 방법을 기술한다.

2 개의 Sample(열)에서 5 개의 유전자(행)에 발현을 관찰하여 행렬(D)로 만든다.

$$D = \begin{bmatrix} 100 & 111 \\ 152 & 45 \\ 19 & 33 \\ 22 & 31 \\ 27 & 10 \end{bmatrix} \in \mathbb{R}^{5 \times 2}$$

유전자 각각의 변동을 알기 위해 행렬(D)에서 평균 값(기준점)을 뺀으로써 행렬(X) 구할 수 있다. (평균 값을 원점으로 데이터가 확장된 정도)

$$X = D - m_{\text{ean}}(D) = \begin{bmatrix} 100 & 111 \\ 152 & 45 \\ 19 & 33 \\ 22 & 30 \\ 27 & 10 \end{bmatrix} - \begin{bmatrix} 64 & 46 \\ 64 & 46 \\ 64 & 46 \\ 64 & 46 \\ 64 & 46 \end{bmatrix} = \begin{bmatrix} 36 & 65 \\ 88 & -1 \\ -45 & -13 \\ -42 & -16 \\ -37 & -36 \end{bmatrix}$$

Transpose X 와 X 를 곱함으로써 데이터의 내적 값을 구할 수 있고 (유전자의 개수-1)을 나눔으로 데이터 공분산 행렬을 구할 수 있다.

$$X^T X = \begin{bmatrix} 36 & 88 & -45 & -42 & -37 \\ 65 & -1 & -13 & -16 & -36 \end{bmatrix} \begin{bmatrix} 36 & 65 \\ 88 & -1 \\ -45 & -13 \\ -42 & -16 \\ -37 & -36 \end{bmatrix} = \begin{bmatrix} 14198 & 4841 \\ 4841 & 5947 \end{bmatrix}$$

$$\frac{X^T X}{n-1} = \begin{bmatrix} 14198 & 4841 \\ 4841 & 5947 \end{bmatrix} / 4 = \begin{bmatrix} 3549.5 & 1210.25 \\ 1210.25 & 1486.74 \end{bmatrix}$$

공분산은 공통되게 움직이는 정도를 표현한 것으로 이제 방향(Eigenvalue)를 계산해야 한다. Eigenvalue(  $\lambda$  )는 nonzero solution vector K (Eigenvector) 가 존재해야 한다는 조건이 있다. 따라서,  $AK = \lambda K$  를 만족하여야 하며  $K(A - \lambda I) = 0$  과 같다. 이때  $K(A - \lambda I)$  이 역 행렬이 존재 한다면  $K = 0$  이기 때문에 조건에 모순이 된다.

$$\therefore \det(A - \lambda I) = 0$$

위의 예제를 공식에 따라 대입해보면 다음과 같다.

$$\det \left( \begin{bmatrix} 3549.5 - \lambda & 1210.25 \\ 1210.25 & 1486.74 - \lambda \end{bmatrix} \right) = 0 \rightarrow (3549.5 - \lambda)(1486.74 - \lambda) - 1210.25^2 = 0$$

$$\therefore \lambda_1 = 275.101, \lambda_2 = -5311.341$$

#### 4-2. ExDEGA GraphicPlus 를 이용한 PCA 분석 방법

ExDEGA GraphicPlus를 이용하려면, ExDEGA 레포트 우측 하단의 ExDEGA Graphic Plus Start 버튼을 클릭하여 ExDEGA Graphic Plus를 활성화시킨다.

메인 화면의 3 개 탭 중 'PCA' 탭에서 PCA 분석을 수행할 수 있다. PCA 분석 창은 그림 4-1 과 같다.

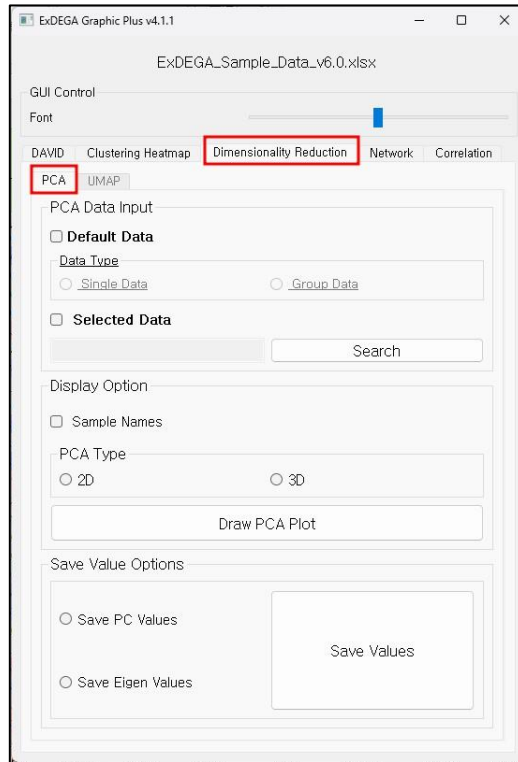


그림 4-1. Select PCA tab

1~3 의 과정을 거쳐 PCA 분석을 수행한다(그림 4-2).

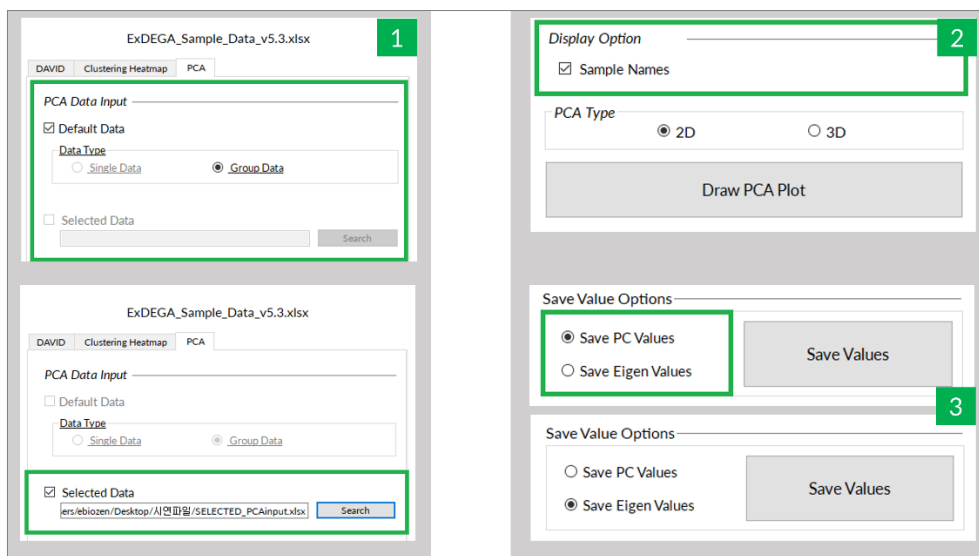


그림 4-2. Steps for Create PCA Graph

먼저, 1 에서 PCA 그래프 제작에 사용될 Data input 타입을 선택해야 한다. 현재 ExDEGA Graphic Plus 에 인식된 ExDEGA 레포트를 그대로 사용하는 경우 Default Data, 별도의 값을 이용하여 PCA 그래프를 제작할 경우 Selected Data 를 선택한다. 이 때, Default Data 를 선택하면 인식된 ExDEGA 레포트의 단일/반복 실험 유형에 따라 자동으로 Single/Group 으로 인식된다. Selected Data 의 경우, 반드시 정해진 양식의 .xlsx 형식의 파일만 입력 가능하니 주의가 필요하다. \* Array Data 의 경우 Default Data 옵션을 이용할 수 없다. Selected Data 옵션을 체크하고, input 파일을 별도로 만들어 이용해야 한다.

Default Data 이면서 Group Data 인 경우, 반복실험 결과( $N \geq 2$ )로 PCA 분석을 하는 경우에 해당한다.

Default Data 이면서 Single Data 인 경우, 반복실험 하지 않은 실험 결과( $N=1$ )로 PCA 분석을 하는 경우에 해당한다.

Selected Data 의 PCA input 파일 만드는 방법은 본 매뉴얼의 '**4-3. PCA Plot input 파일 작성방법**'에 설명되어 있다.

2 에서는 PCA 그래프의 표현 옵션을 제공한다. Display Option 에서 Sample Names 를 체크하면, PCA plot 에 Sample 의 인덱스(샘플 순서대로 부여하는 숫자)를 표시한다. 이 Sample Names 는 그래프 내의 범례로도 표현된다.

마지막으로, 3 에서 2D 를 누르면 PCA 2D (2 차원 평면) 분석 결과가 나오고 3D 를 누르면 PCA 3D (3 차원 공간) 분석 결과가 나온다(그림 4-3). PCA 2D 는 x 축이 PC1, y 축이 PC2 로 작성된 결과이다. PCA 3D 는 x 축이 PC1, y 축이 PC2, z 축이 PC3 로 작성된 결과이다.

각각 좌측은 Display Option 에서 Sample Names 옵션을 선택한 경우, 우측은 선택하지 Sample Names 옵션을 선택하지 않은 경우에 해당한다.

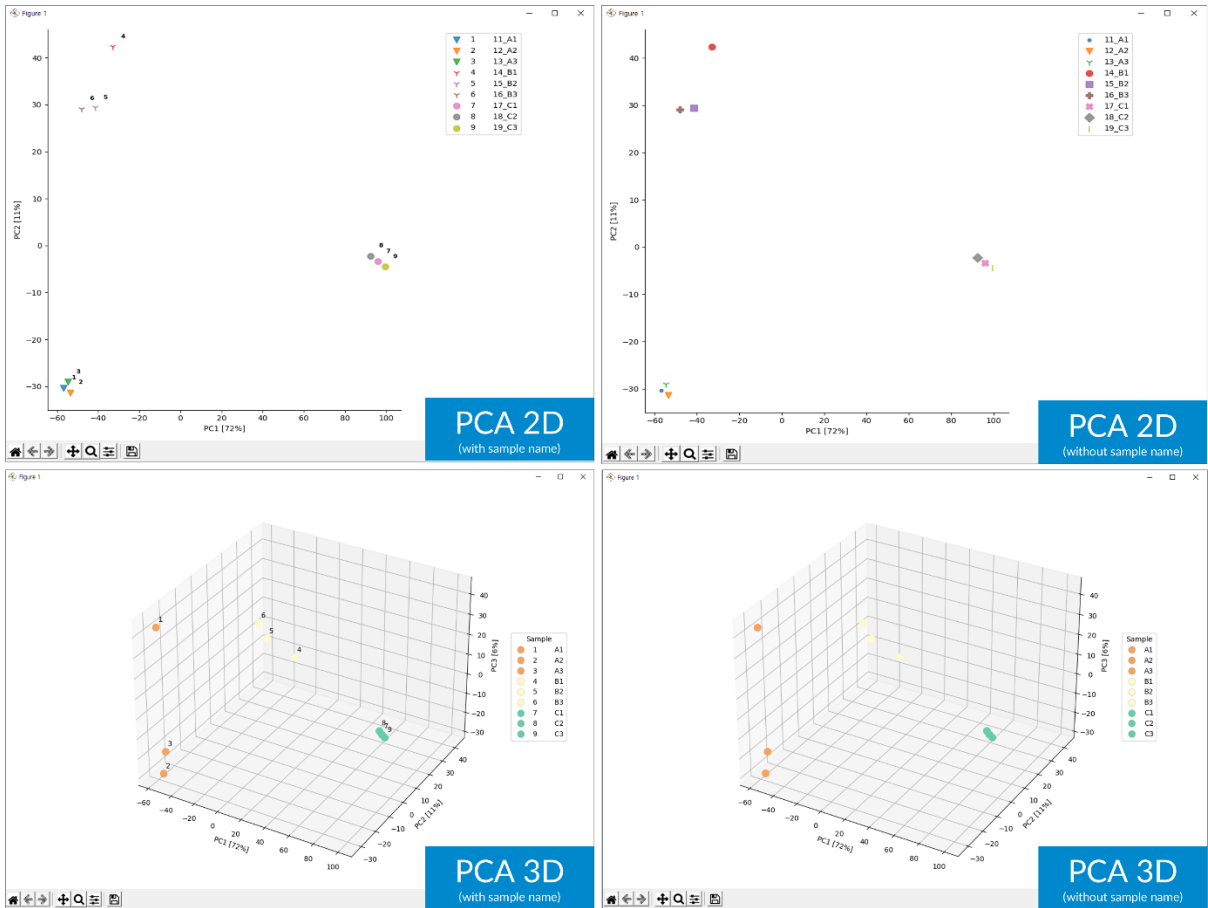


그림 4-3. PCA results 2D, 3D

3 의 Save Value Options 에서 Save PC Values 를 체크한 후 Save Values 버튼을 누르면 각 주성분(PC)이 전체 데이터를 얼마만큼 설명할 수 있는 지 분산 비율로 계산되어 Excel 파일 형식으로 저장된다(그림 4-4). 이 때, 저장하기 전 PC 값들을 그래프로 먼저 확인할 수 있다(그림 4-5).

	A	B
1	Principal Components	Variance Explained
2	PC1	0.716810026
3	PC2	0.10768151
4	PC3	0.056442647
5	PC4	0.047627244
6	PC5	0.038690301
7	PC6	0.024854022
8	PC7	0.007894249
9	PC8	4.87531E-30
10	PC9	3.9363E-31
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		

그림 4-4. Save PC Values as Excel

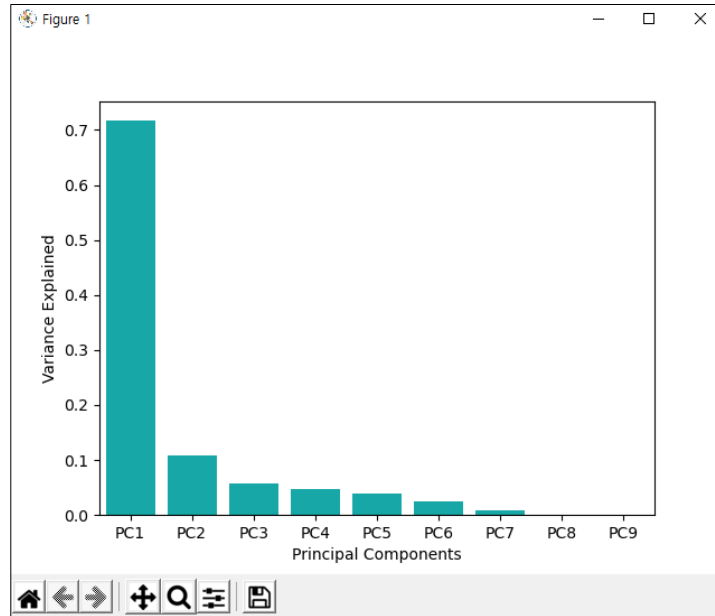


그림 4-5. Check pc values before saving the data as excel

Save Eigen Values 를 체크한 후 Save Values 버튼을 누르면 계산된 Eigenvalue 의 결과를 Excel 파일 형식으로 저장한다(그림 4-6). Eigen values 는 PCA plot 에서 각 샘플의 좌표이다.

	A	B	C	D	E	F	G	H	I	J
1	Sample	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
2	A1	-56.8864	-30.3457	43.08954	-13.297	7.524544	-4.4288	0.033604	1.73E-13	1.17E-14
3	A2	-53.6851	-31.329	-28.1938	-13.4698	-22.2992	-14.1345	0.252672	1.73E-13	1.17E-14
4	A3	-54.6259	-29.076	-18.6736	21.70531	22.02369	16.76101	-0.15929	1.73E-13	1.17E-14
5	B1	-32.9178	42.36758	-12.7574	-19.6184	26.06745	-11.3873	-0.07946	1.73E-13	1.17E-14
6	B2	-41.5794	29.39304	3.757813	-14.5291	-19.7461	25.75895	-0.5247	1.73E-13	1.17E-14
7	B3	-48.145	29.10471	10.89999	36.80833	-12.3007	-12.5653	-0.29896	1.73E-13	1.17E-14
8	C1	95.94653	-3.37155	0.625813	0.8002	-0.42324	-0.00138	0.258714	7.22E-14	1.32E-13
9	C2	92.43661	-2.23767	0.83776	1.080566	-0.51381	0.500988	15.43273	2.24E-13	-4.9E-14
10	C3	99.45645	-4.50543	0.413867	0.519833	-0.33267	-0.50374	-14.9153	2.24E-13	-4.9E-14
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										

그림 4-6. Save Eigen Values as Excel

이 값으로 엑셀에서 분산형 차트를 그려 직접 PCA 2D 를 작성할 수도 있다(그림 4-7).

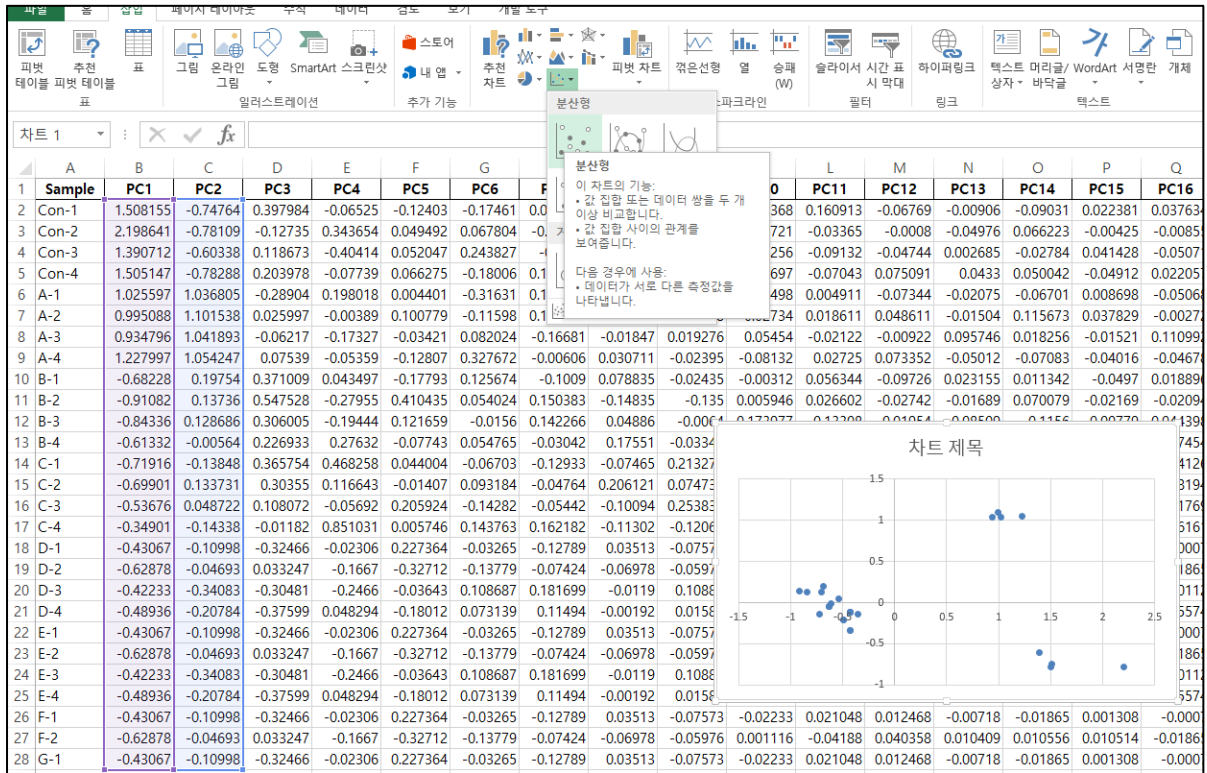


그림 4-7. Create PCA 2D Graph using saved eigen values

### 4-3. PCA Plot input 파일 작성방법

#### Selected Data – Group Data 의 경우

같은 Group 의 Sample 들을 묶어 같은 색상으로 표시하여 PCA 를 표현할 수 있다. Input File 은 Excel 에서 작성한다.

<작성 순서>

1) Normalized data(log2)항목을 복사한다(그림 4-9).

Normalized data (log2)									
	A1	A2	A3	B1	B2	B3	C1	C2	C3
5	1.544	1.369	1.865	1.631	1.149	1.290	0.843	0.743	0.943
6	0.629	0.166	0.221	0.514	0.562	0.159	0.944	0.844	1.044
2	0.000	0.000	0.050	0.000	0.000	0.000	0.100	0.000	0.200
0	0.231	0.330	0.100	0.642	0.091	0.071	0.328	0.228	0.428
5	0.303	0.799	0.111	1.771	0.917	0.294	0.923	0.823	1.023
2	0.000	0.000	0.063	0.055	0.000	0.088	0.100	0.000	0.200
2	0.000	0.290	0.000	0.000	0.098	0.000	0.100	0.000	0.200
2	0.440	0.002	0.000	0.001	0.001	0.862	0.100	0.000	0.200
4	0.218	0.564	0.000	0.001	0.002	0.000	1.742	1.642	1.842
2	0.000	0.000	0.266	0.000	0.000	0.000	0.100	0.000	0.200
2	0.406	0.001	0.000	0.000	0.000	0.000	0.100	0.000	0.200
6	3.136	2.775	2.732	2.813	3.270	3.323	2.474	2.374	2.574
3	1.510	1.210	0.904	1.276	1.172	0.962	2.241	2.141	2.341
2	0.000	0.000	0.000	0.098	0.000	0.157	0.100	0.000	0.200
2	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.200

그림 4-8. Copy normalized data

참고) 전체 데이터 선택 방법은 A1(초록색 네모) 클릭 후 Shift 를 누른 상태에서 C3(파란색 네모)를 클릭한다. A1~ C3 까지 선택 후에 Ctrl + Shift + 아래 방향키(↓)를 입력하면 전체 데이터가 한 번에 선택된다.

2) 새 엑셀 파일(기존 파일에 sheet 추가가 아닌 엑셀 새로 만들기)을 열어서 **2 번째** 열에 붙여넣기 한다(그림 4-10).

	A	B	C	D	E	F	G	H	I
1									
2	A1	A2	A3	B1	B2	B3	C1	C2	C3
3	1.544	1.369	1.865	1.631	1.149	1.290	0.843	0.743	0.943
4	0.629	0.166	0.221	0.514	0.562	0.159	0.944	0.844	1.044
5	0.000	0.000	0.050	0.000	0.000	0.000	0.100	0.000	0.200
6	0.231	0.330	0.100	0.642	0.091	0.071	0.328	0.228	0.428
7	0.303	0.799	0.111	1.771	0.917	0.294	0.923	0.823	1.023
8	0.000	0.000	0.063	0.055	0.000	0.088	0.100	0.000	0.200
9	0.000	0.290	0.000	0.000	0.098	0.000	0.100	0.000	0.200
10	0.440	0.002	0.000	0.001	0.001	0.862	0.100	0.000	0.200
11	0.218	0.564	0.000	0.001	0.002	0.000	1.742	1.642	1.842
12	0.000	0.000	0.266	0.000	0.000	0.000	0.100	0.000	0.200
13	0.406	0.001	0.000	0.000	0.000	0.000	0.100	0.000	0.200
14	3.136	2.775	2.732	2.813	3.270	3.323	2.474	2.374	2.574
15	1.510	1.210	0.904	1.276	1.172	0.962	2.241	2.141	2.341
16	0.000	0.000	0.000	0.098	0.000	0.157	0.100	0.000	0.200
17	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.200
18	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.200
19	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.200

그림 4-9. Paste normalized data in new excel file



3) 1 번째 열에는 그룹명 입력 후 병합한다(그림 4-11).

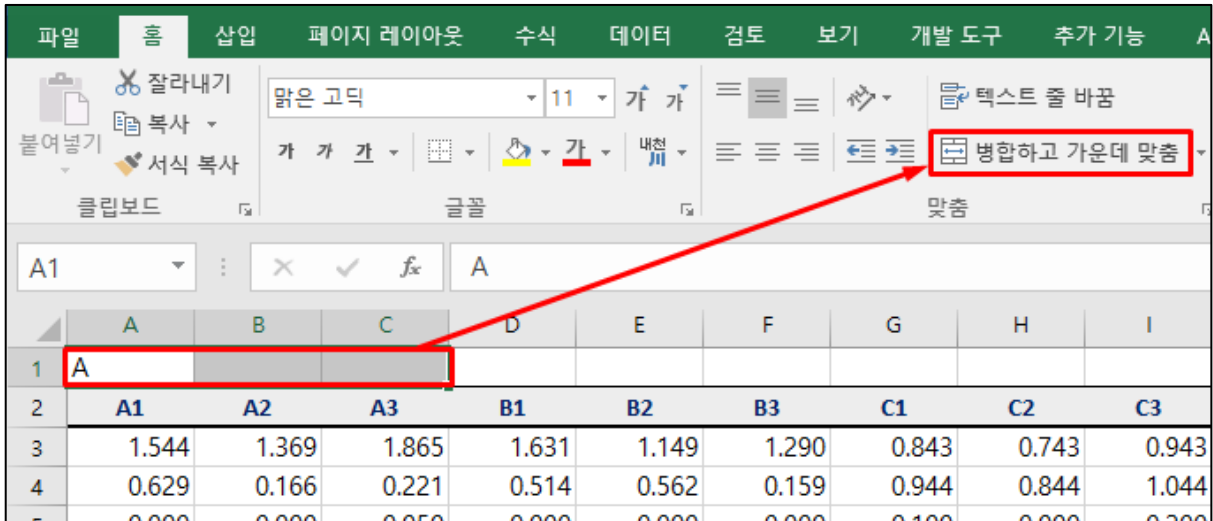


그림 4-10. Group information

4) 완성 후 파일형식은 엑셀로 저장한다(그림 4-12). Input file 명에는 띄어쓰기가 들어가지 않도록 주의한다.

	A	B	C	D	E	F	G	H	I
1	A			B			C		
2	A1	A2	A3	B1	B2	B3	C1	C2	C3
3	1.544	1.369	1.865	1.631	1.149	1.290	0.843	0.743	0.943
4	0.629	0.166	0.221	0.514	0.562	0.159	0.944	0.844	1.044
5	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.200
6	0.231	0.330	0.000	0.000	0.000	0.000	0.328	0.228	0.428
7	0.303	0.799	0.111	1.771	0.917	0.294	0.923	0.823	1.023
8	0.000	0.000	0.063	0.055	0.000	0.088	0.100	0.000	0.200
9	0.000	0.290	0.000	0.000	0.098	0.000	0.100	0.000	0.200
10	0.440	0.002	0.000	0.001	0.001	0.862	0.100	0.000	0.200
11	0.218	0.564	0.000	0.001	0.002	0.000	1.742	1.642	1.842
12	0.000	0.000	0.266	0.000	0.000	0.000	0.100	0.000	0.200

그림 4-11. Save PCA input (group data) file

### Selected Data – Single Data 의 경우

- 1) Group Data 의 (1) ~ (2) 과정은 동일하게 진행한다.
- 2) 2 열의 샘플명을 1 열에 동일하게 복사/붙여넣기 한다.

## 5. String Network Analysis (ExDEGA GraphicPlus)

STRING tool 은 Protein-Protein Interaction 데이터 베이스를 기반으로 유전자의 상관관계를 통계적으로 분석하여 Network 을 작성해주는 분석 툴이다. Graphic plus 를 이용하면 관심있는 유전자 리스트를 넣고 바로 STRING network 이미지를 얻을 수 있다. 이미지 편집이 필요하다면 매뉴얼 뒤쪽의 Cytoscape-STRING 분석 툴을 활용하여 분석이 가능하다.

분석 과정은 그림 5-1 과 같다.

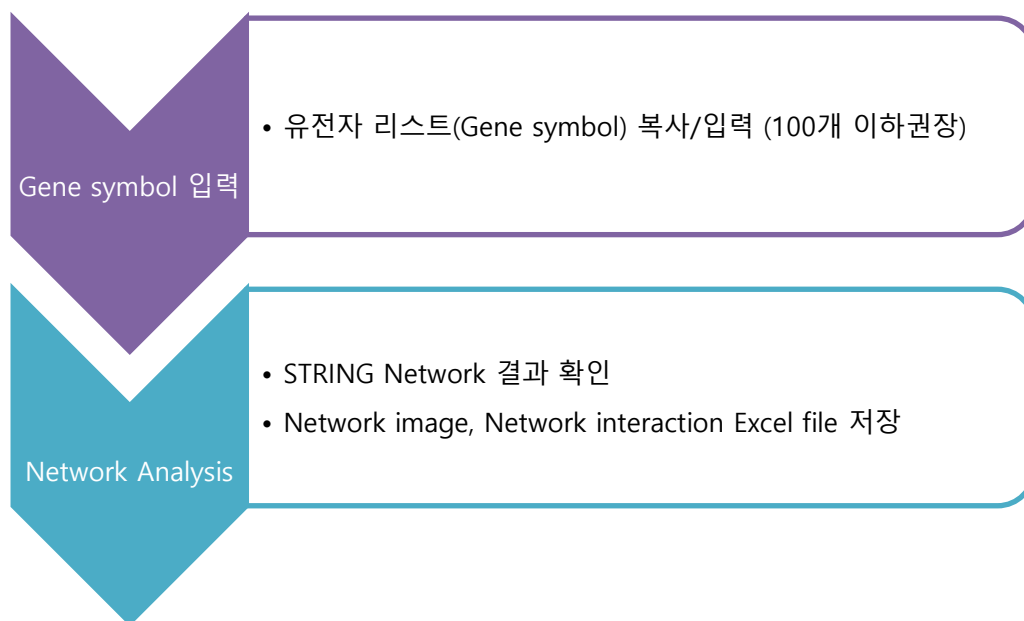


그림 5-1. STRING analysis process

분석을 진행하기 위해서는 먼저 관심있는 유전자를 선별해야 하는데, 이때 유전자 개수는 100 개 이내로 선별하는 것을 권장한다. 유전자 개수가 너무 적으면 network가 잘 형성이 되지 않고, 개수가 너무 많으면 복잡도가 높아져서 이미지를 직관적으로 보기 어렵기 때문이다.

유전자를 선별한 뒤 그림 5-2와 같이 빨간색 상자에 유전자 리스트를 복사 붙여넣기 해준다. 만약 유전자 리스트를 엑셀데이터 새창으로 저장했을 경우, 상단의 Import 버튼을 통해서도 불러올 수 있다.

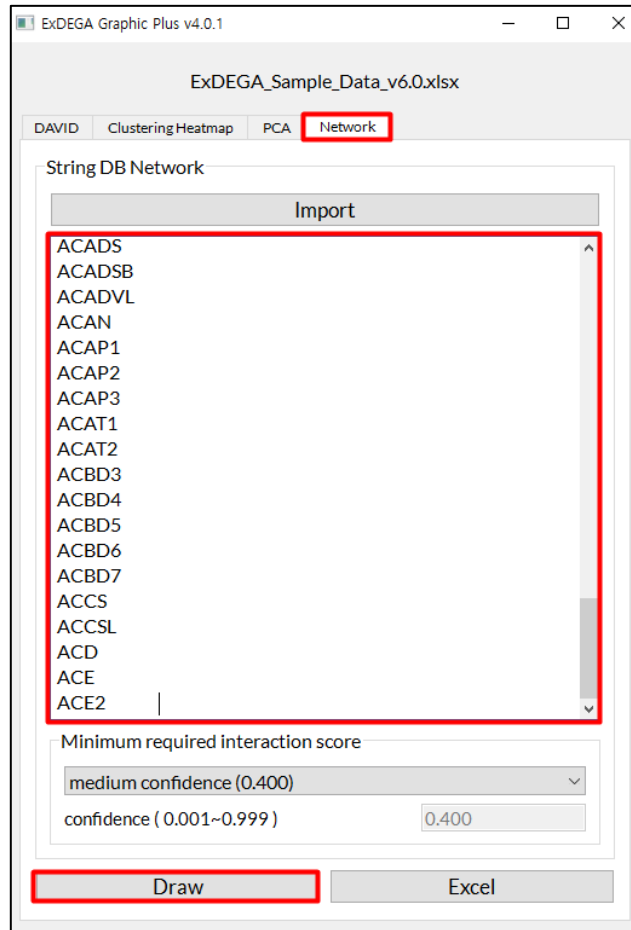


그림 5-2. Graphic plus STRING network 분석 창

그리고 Draw 버튼을 클릭하면 바로 STRING network 분석이 진행되어 이미지를 저장할 수 있는 창이 뜬다. 이미지는 벡터형식인 .svg 형식으로 추출되며, 저장하고자 하는 경로를 설정한 뒤 저장버튼을 클릭하면 이미지 저장이 완료된다. (그림5-3)

오른편의 Excel 버튼을 클릭하면 유전자 간의 interaction score 값을 엑셀형식으로 저장하여 확인할 수 있다. 결과자료는 그림 5-4와 같이 3개의 열로 구성이 되는데, 앞쪽 두개의 열(node1, node2)은 서로 연관 있는 유전자명을 나타내며 세번째 열은 STRING 분석을 통해 계산된 두 유전자 간의 interaction score 값을 나타낸다.

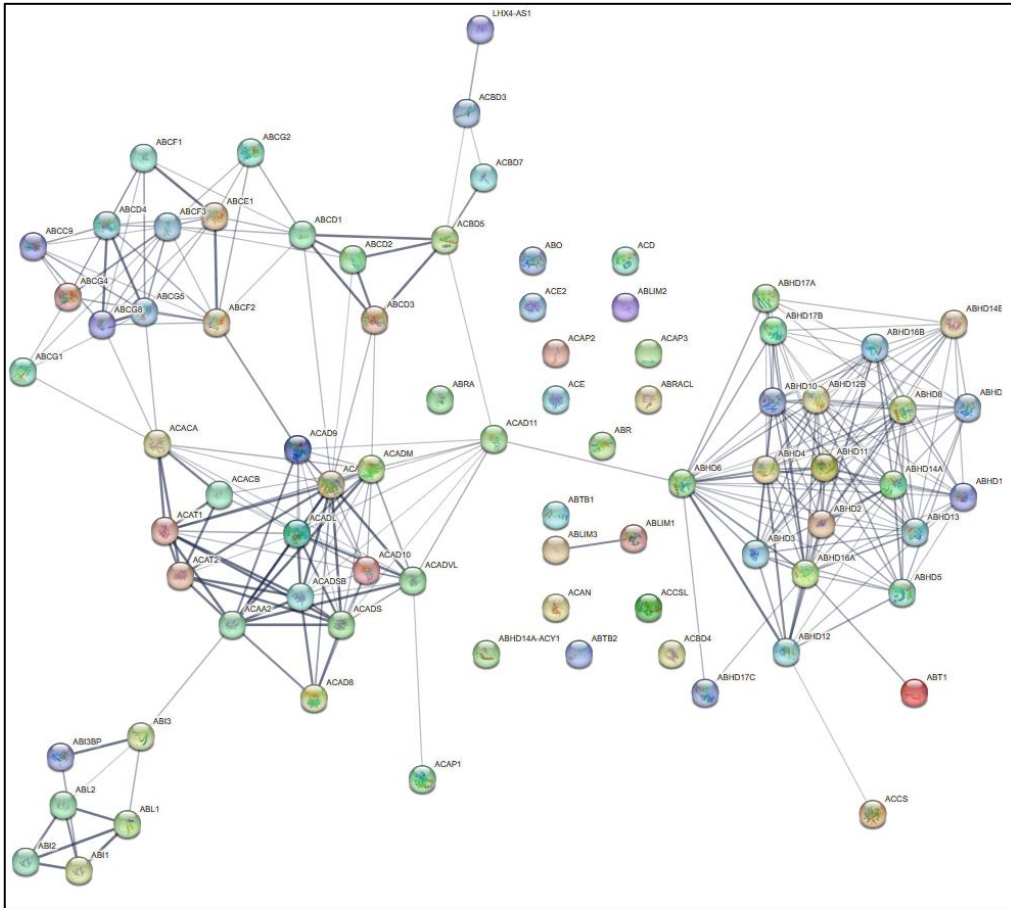


그림 5-3. STRING network 이미지

	A	B	C
1	node1	node2	score
2	ACAP1	ACADVL	0.453
3	ABCD1	ABCF2	0.451
4	ABCD1	ABCF1	0.46
5	ABCD1	ACAA1	0.499
6	ABCD1	ABCE1	0.535
7	ABCD1	ABCF3	0.539
8	ABCD1	ABCG2	0.587
9	ABCD1	ABCD2	0.902
10	ABCD1	ACBD5	0.947
11	ABCD1	ABCD3	0.959
12	ABCF2	ABCC9	0.462
13	ABCF2	ABCF3	0.487
14	ABCF2	ABCG8	0.529
15	ABCF2	ABCG2	0.59
16	ABCF2	ABCD4	0.604
17	ABCF2	ABCG5	0.609
18	ABCF2	ABCG4	0.616
19	ABCF2	ACAD9	0.725
20	ABCF2	ABCE1	0.964

그림 5-4. STRING network interaction score 값

STRING network 옵션으로 그림 5-5와 같이 interaction score 임계점을 설정할 수 있는데, 기본 값은 0.4 로 분석이 진행되며 0~1 사이 값을 설정할 수 있다. Score 값이 1에 가까워질수록 연관도가 높은 유전자 간의 edge(선)이 형성되며, 0에 가까워질수록 연관도가 적은 유전자까지도 edge가 형성된다. 만약 gene list가 너무 많아서 network가 복잡할 경우에는 score 값을 올리고, gene list가 적어서 단순할 경우에는 score값을 내리는 방식으로 이미지를 제작할 수 있다.

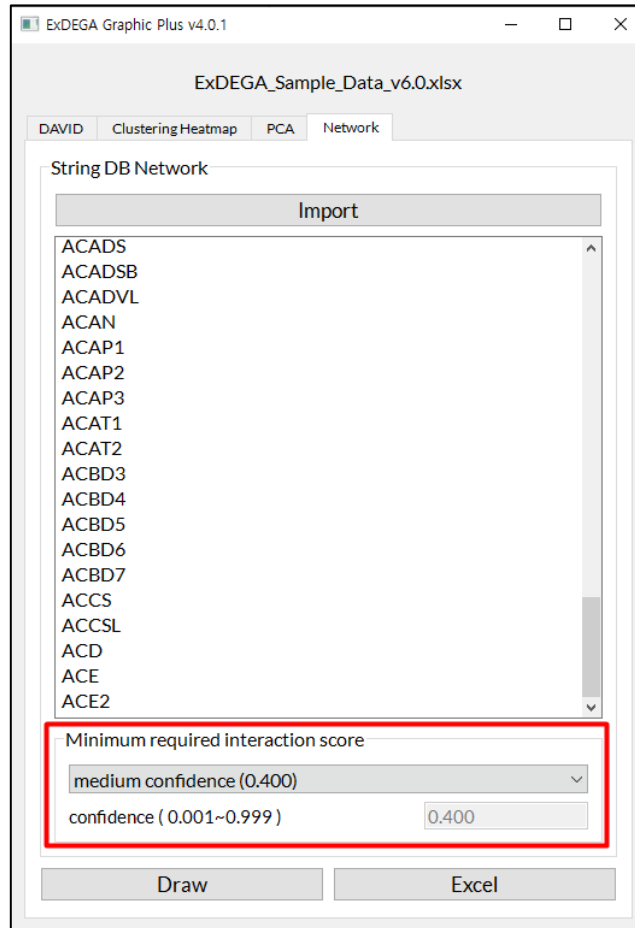


그림 5-5. STRING network score 옵션 설정

## 6. Correlation Analysis (ExDEGA GraphicPlus)

Correlation은 샘플들 간에 어떤 상관관계를 가지는지 상관계수를 이용하여 통계적으로 측정하는 분석이다.

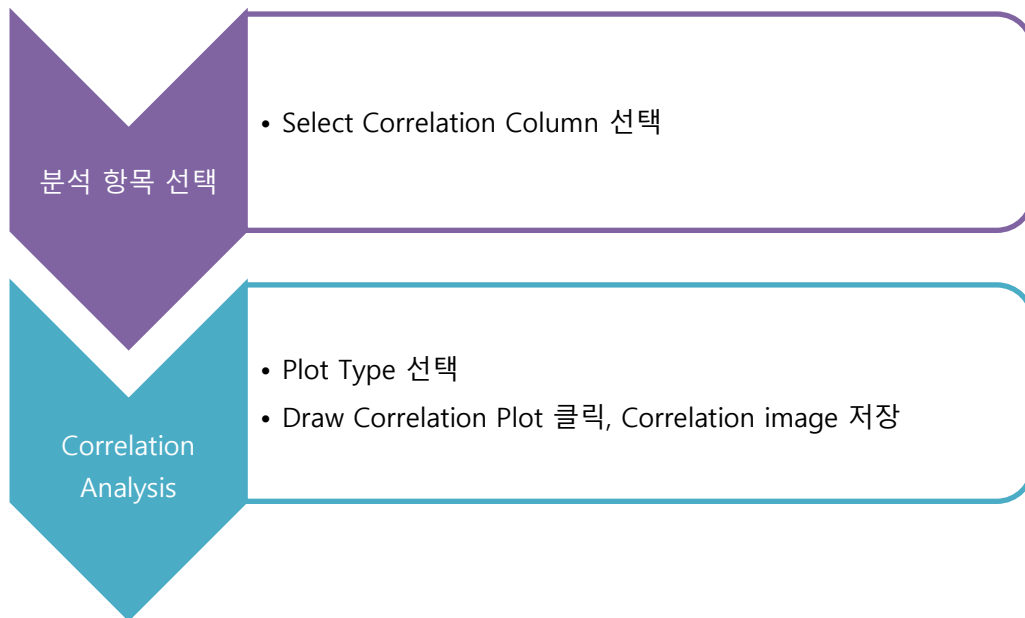


그림 6-1. Correlation analysis process

우선 Select Correlation Column 에서 원하는 항목을 선택한 후에 아래 칸에서 원하는 항목을 체크한다. 전체를 체크하고 싶은 경우, "Select All Data Column"을 클릭한다. 그리고 Plot Type 에서는 이미지의 형식을 선택하고 "Draw Correlation Plot"을 클릭하면 이미지가 제작되며, 저장하고자 하는 경로를 설정한 뒤 저장버튼을 클릭하면 이미지가 저장이 완료된다(그림 6-2).

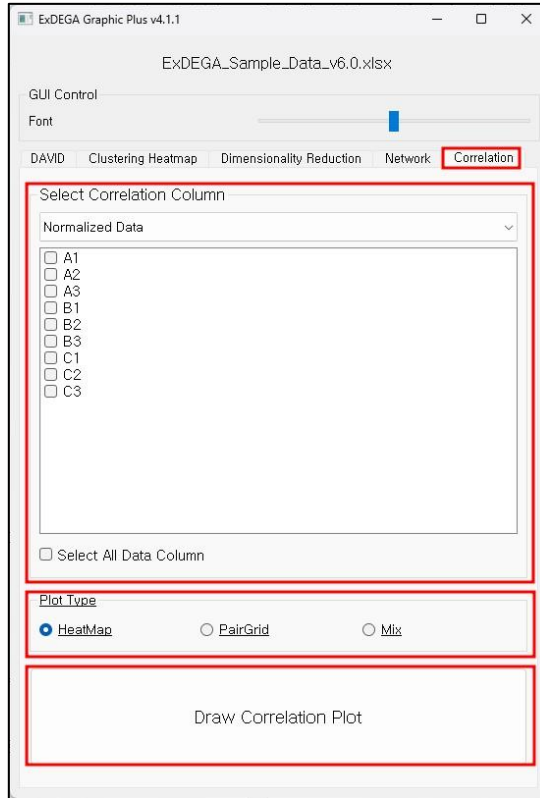


그림 6-2. Correlation analysis 옵션 설정

이미지는 총 3 개의 Type으로 선택할 수 있으며, HeatMap을 선택하면 아래 그림과 같이 제작되며, 1에 가까울수록 높은 상관관계를 갖는다고 볼 수 있다(그림 6-3).



그림 6-3. Correlation plot (HeatMap)

PairGrid를 선택하면 아래 그림과 같이 제작되며 증가하는 경향의 모양은 양의 상관관계, 감소하는 경향의 모양은 음의 상관관계로 볼 수 있다. 또한, 1에 가까울수록 직선에 가까운 분포도 모양이 나온다(그림 6-4).

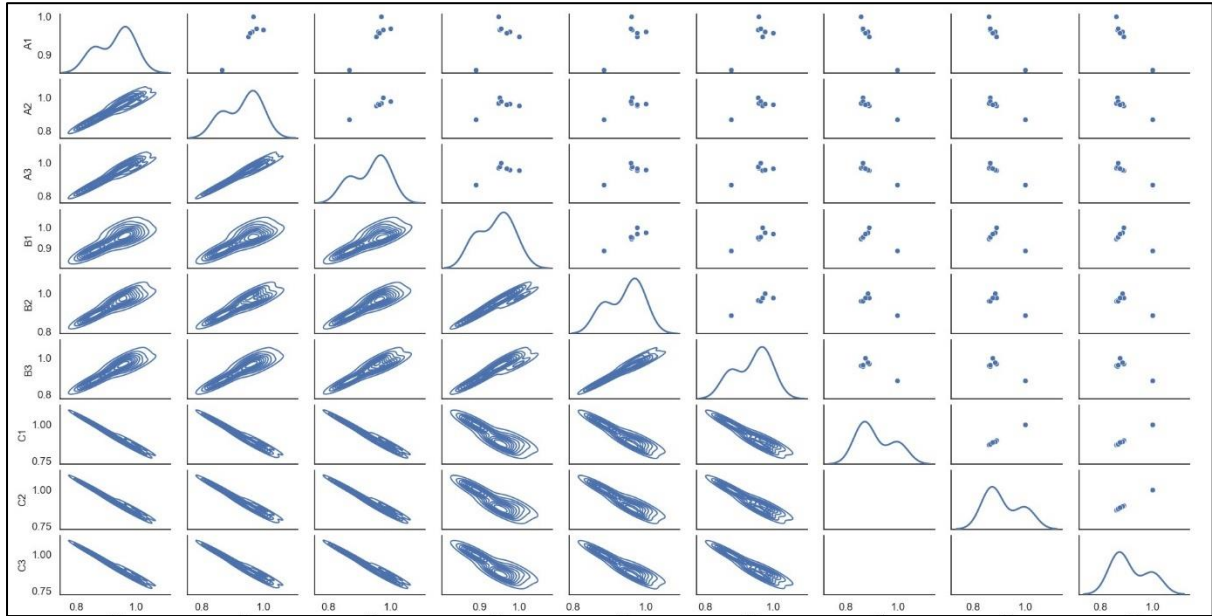


그림 6-3. Correlation plot (PairGrid)

Mix를 선택하면 이전의 HeatMap과 PairGrid가 반씩 섞여서 제작이 된다(그림 6-4).

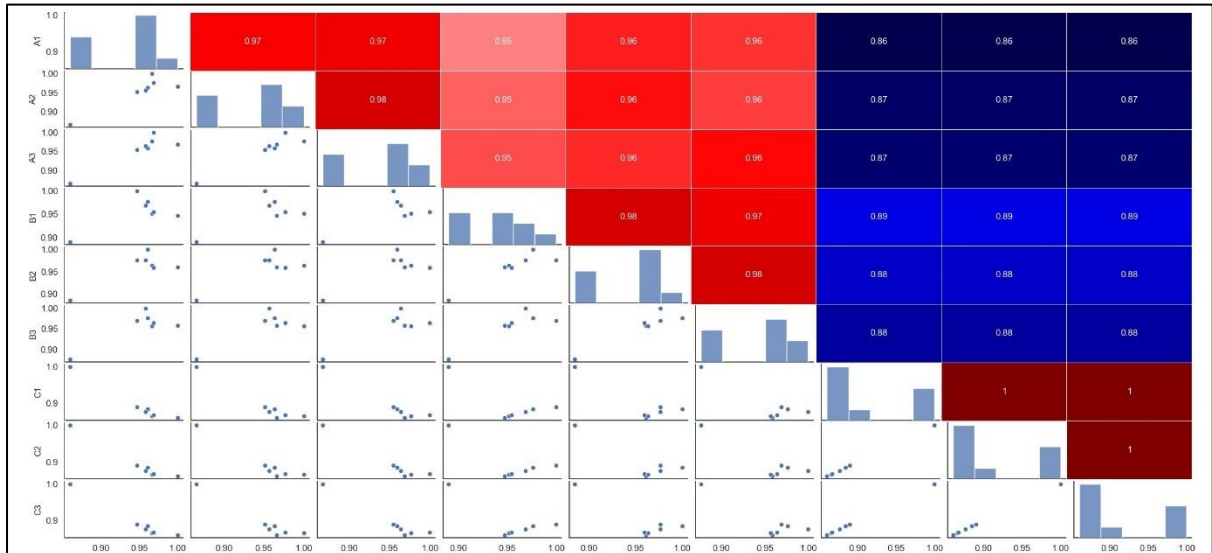


그림 6-4. Correlation plot (Mix)



## 7. Pathway analysis (KEGG mapper)

RNA-Seq 분석 결과에서 up/down-regulated genes들이 어떤 Pathway에 속하는지 확인하고자 한다면 KEGG에서 제공하는 KEGG Mapper를 이용한다. 사용방법은 그림 7-1과 같은 순서로 진행된다.

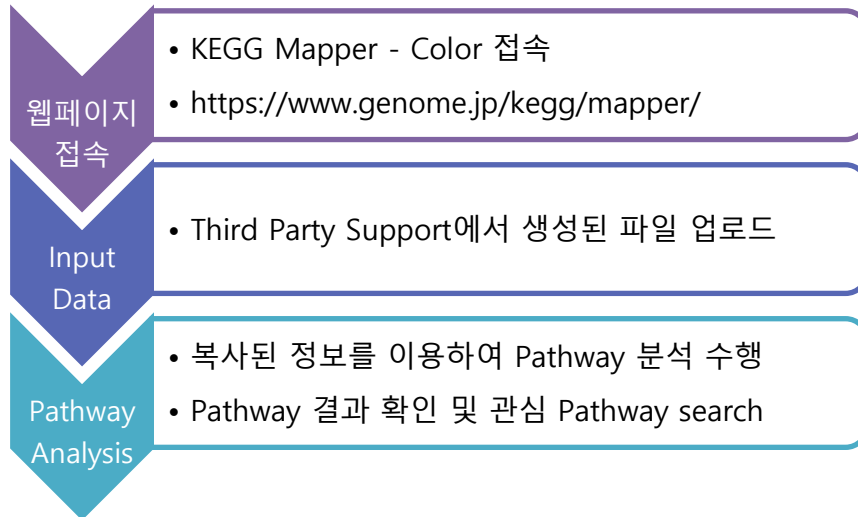


그림 7-1. KEGG Mapper tool analysis process

그림 7-2는 mRNA-Seq report에서 Fold change 2, normalized data(log2) 4, p-value<0.05을 기준으로 선별한 유전자를 KEGG 분석하는 과정이다. Significant gene selection에서 Fold change, Normalized Data(log2), p-value (반복실험의 경우) 값을 지정하고, 확인하고자 하는 Fold change 조합을 선택하여 필터를 적용한다. 필터를 적용하여 선별된 유전자를 대상으로 Third Party Support를 통해 KEGG input을 추출하여, KEGG 분석에 사용한다. 반드시 하나의 비교 조합만 진행해야 한다.

ID	Gene symbol	Fold change			p-value			Average of r
		B/A	C/A	C/B	B/A	C/A	C/B	
113	ABHD3	2.437	2.079	0.853	0.002	0.000	0.176	2.778
370	ADGRE3	4.134	4.259	1.018	0.015	0.000	0.827	3.071
611	609 ALDH2	0.345	0.210	0.608	0.002	0.001	0.026	5.497
695	693 ANICA1	2.612	2.663	1.020	0.006	0.005	0.764	6.065
870	868 ANXR2	2.282	2.420	1.060	0.021	0.000	0.718	3.081
934	932 APBB1P	2.711	2.745	1.013	0.005	0.001	0.904	4.910
978	976 APBR	2.383	3.468	1.453	0.040	0.000	0.075	4.947
1023	1021 AQP9	2.275	2.528	1.243	0.032	0.003	0.163	6.151
1068	1066 ARHGAP25	3.604	2.757	0.765	0.002	0.001	0.068	4.612
1091	1089 ARHGAP9	2.176	2.444	1.123	0.000	0.000	0.094	5.440
1214	1212 ARDC3	2.075	1.949	0.939	0.018	0.020	0.460	3.877
1354	1352 ATHL1	2.113	1.310	0.621	0.040	0.116	0.082	3.591
1389	1387 ATP1B3	0.494	0.448	0.904	0.010	0.002	0.328	4.640
1441	1439 ATP9V1B2	2.198	1.359	0.618	0.032	0.384	0.001	5.566
1535	1533 B3GNT8	2.471	3.434	1.390	0.036	0.001	0.085	2.894
1596	1594 BASP1	2.968	3.652	1.256	0.005	0.002	0.025	6.344
1636	1634 BCKDHA	0.455	0.409	0.899	0.040	0.031	0.254	4.533
1654	1652 BCL6	2.988	2.622	1.107	0.047	0.000	0.833	4.499
1728	1726 BIRC3	0.476	0.576	1.209	0.002	0.000	0.255	4.087
1753	1751 BLVRA	0.322	0.308	0.933	0.040	0.037	0.349	5.372
1977	1975 C11orf68	2.542	1.699	0.669	0.000	0.002	0.003	3.074
2060	2058 C16orf54	2.830	1.923	0.679	0.018	0.021	0.097	4.127
2350	2348 CSAR2	2.411	2.669	1.107	0.004	0.000	0.318	3.056
2622	2620 CANT1	2.312	2.998	1.110	0.002	0.000	0.278	3.201
2635	2633 CAPN2	0.458	0.735	1.638	0.001	0.022	0.001	5.603
2661	2659 CARD8-AS1	2.393	1.645	0.688	0.002	0.016	0.005	2.791
2711	2709 CAST	0.489	0.662	1.353	0.023	0.074	0.019	5.721
2883	2881 CCNH	0.404	0.677	1.677	0.019	0.112	0.001	4.239
2989	2987 CCN1L1	0.492	0.620	1.261	0.006	0.006	0.159	4.814
2999	2997 CCPG1	2.050	2.792	1.362	0.008	0.000	0.032	3.587
3033	3031 CD163	0.418	0.197	0.377	0.040	0.012	0.000	4.033
3082	3080 CD4	0.444	0.577	0.217	0.010	0.001	0.002	6.334
3088	3084 CD46	2.013	2.357	1.171	0.008	0.000	0.176	5.042

그림 7-2. KEGG Mapper input file generation process

그림 7-3과 같이 KEGG Mapper 웹페이지(<https://www.genome.jp/kegg/mapper/>)에 접속하고 Color 항목에 들어가면 아래와 같은 화면을 볼 수 있다.

- (1) 분석하고자 하는 유전자의 species를 선택 (Human이면 hsa 선택)
- (2) Or upload file의 파일 선택 버튼을 클릭하고 ExDEGA에서 생성된 입력 데이터 선택
- (3) "Use uncolored diagram"과 "Include aliases" 항목에 체크를 한 후
- (4) Exec 버튼을 누른다.

**KEGG Mapper – Color**

KEGG2 About Reconstruct Search Color Join Convert ID Assign KO Taxonomy

### Color tool

The Color tool searches various KEGG objects, including genes, KOs, EC numbers, metabolites and drugs, against KEGG pathway maps. Found objects may be marked in any combination of background and foreground colors.  
See new article: [KEGG mapping tools for uncovering hidden features in biological data](#)

**Search mode:**  Reference  hsa  other org

**Enter KEGG identifiers followed by color specification**

Examples:

Or upload file:  Kegg input.txt

Default bgcolor:

Use uncolored diagrams

Include aliases (for hsa and other org modes)

그림 7-3. KEGG Mapper tool analysis process

분석결과, 입력한 유전자들이 관여하는 pathway list 가 나온다(그림 7-4). pathway 이름 옆에 있는 괄호 안 숫자는 입력한 유전자 중 각 pathway 에 관여하는 유전자의 수이다. 괄호 안 숫자를 클릭하면 해당 유전자 목록을 볼 수 있다. pathway 이름을 클릭하면 해당 pathway chart 가 열리고 입력한 유전자의 발현 up/down (red/blue)이 색으로 표시되어 있다. Pathway 이미지는 “다른 이름으로 저장”이 가능하고 “html”으로 저장하면 이미지에 링크된 항목을 그대로 유지해서 저장이 가능하다.

\*참고사항 : 만약 오른쪽마우스 버튼을 클릭했을 때 다른 이름으로 저장이 보이지 않을 경우, Internet explorer 대신 chrome 창을 이용하면 확인할 수 있다.

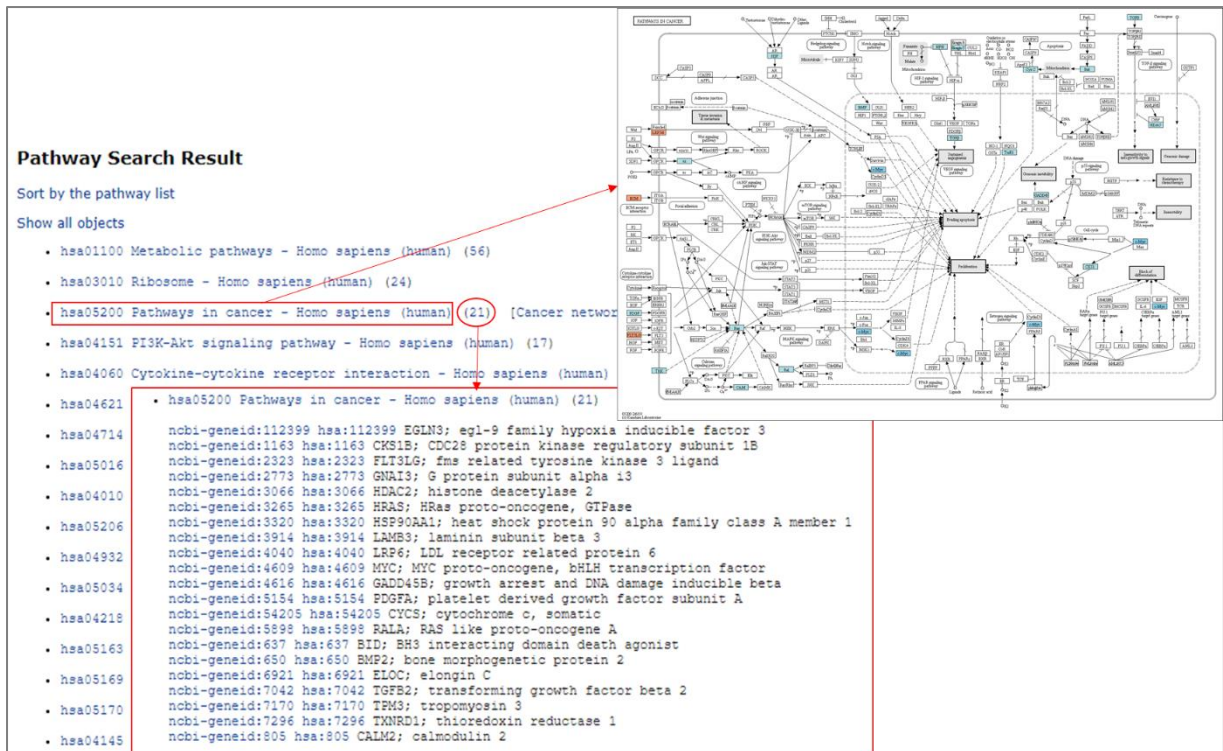


그림 7-4. KEGG Mapper tool analysis result

## 8. Gene set enrichment analysis (GSEA)

Gene set enrichment analysis(GSEA)는 Microarray 또는 RNA-seq data 를 넣어 대조군, 실험군 에서 유의한 gene set 을 분석하는 프로그램이다. GSEA 는 Human, Mouse, Rat 만 지원하고 또한 그룹 비교 (반복 실험) 데이터만 분석이 가능하며, 그룹 당 최소 3 반복 이상이어야 진행이 가능하다.

MSigDB 에 있는 gene set (GO, pathway 등)을 기반으로 분석한다. 분석 과정은 그림 8-1 과 같은 순서로 진행된다.

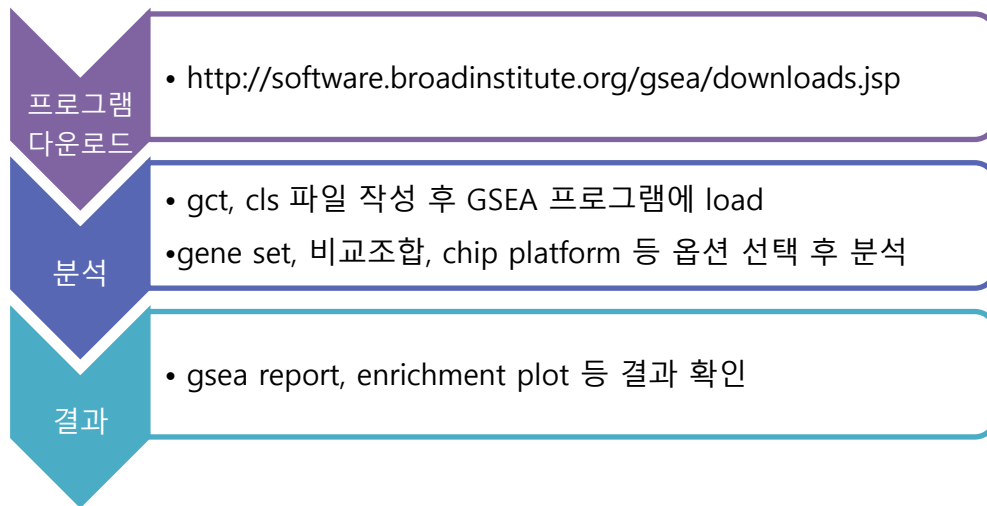


그림 8-1. GSEA tool analysis process

GSEA 홈페이지(<http://software.broadinstitute.org/gsea/downloads.jsp>)에 들어가 회원가입 후 로그인 하여 GSEA 프로그램을 다운로드 받는다 (그림 8-2).

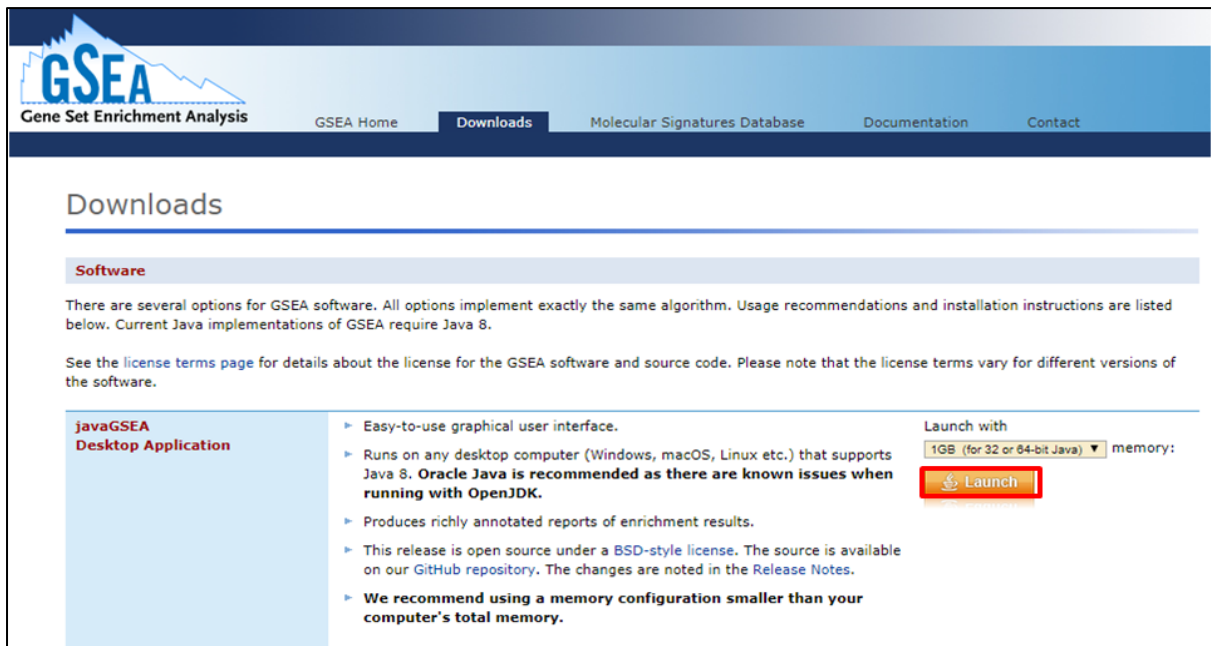


그림 8-2. GSEA program download

ExDEGA Report의 Third Party Support에서 추출한 input을 이용하여 GSEA 분석을 수행한다.

GSEA 프로그램을 열어 Load data 버튼을 누르고 Browse for files 버튼을 누른 후 ExDEGA Report 의 Third Party Support 에서 추출한 gct, cls 파일을 연다(그림 8-6). gct, cls 파일은 파일의 경로가 길면 input 파일을 잘 인식하지 못하므로 되도록 바탕화면에 두고 진행한다. 문제없이 완료가 되면 NO errors 라는 메시지창을 확인할 수 있다(그림 8-7). 메시지창 확인 후, Run GSEA 버튼을 누른다.

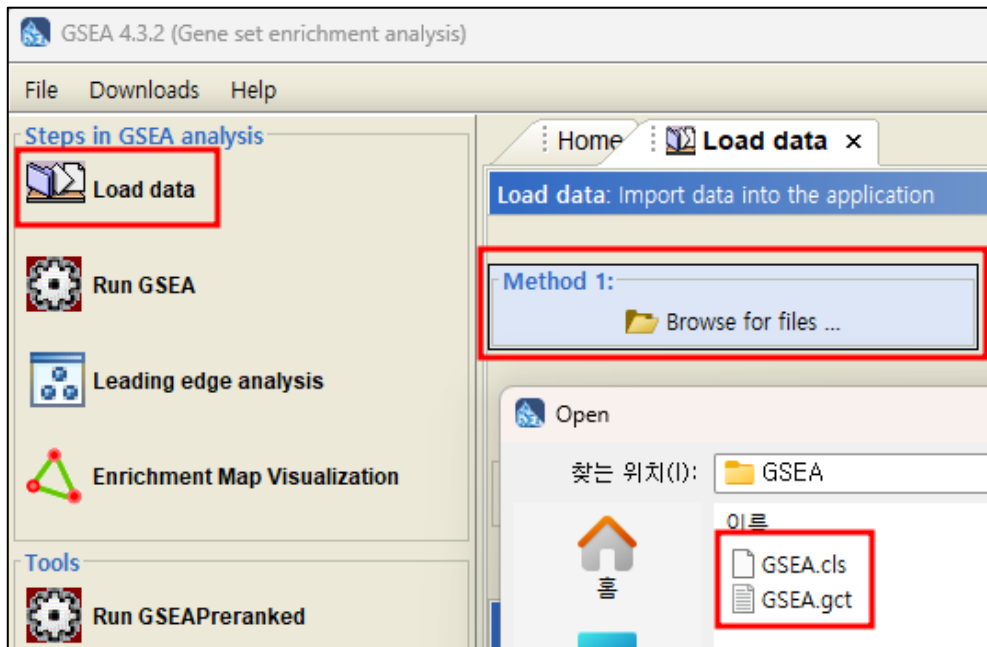


그림 8-6. Load data in GSEA program

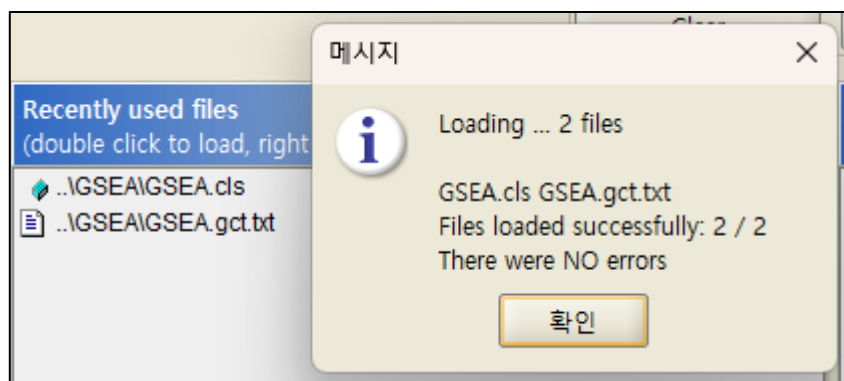


그림 8-7. Load data in GSEA program

Run GSEA 를 누르고 Expression dataset 는 gct 파일명을 선택, gene sets database 는 분석하고자 하는 gene set 을 선택한다(그림 8-8). pathway 분석을 하고자 하면 c2 에서 선택, gene ontology 분석을 하고자 하면 c5 에서 선택한다. Gene set 에 대한 자세한 설명은 GSEA 홈페이지 (<http://software.broadinstitute.org/gsea/msigdb/collections.jsp>)에 있다.

gene sets database 는 현재 Human, Mouse 탭만 있으며, 해당하는 종으로 선택하여 진행하면 되는데, Rat 의 경우, 아직 탭이 존재하지 않아 Human, 또는 Mouse 탭으로 진행한 후, 다음 chip 선택 과정에서 Orthologs 로 진행해야 한다.

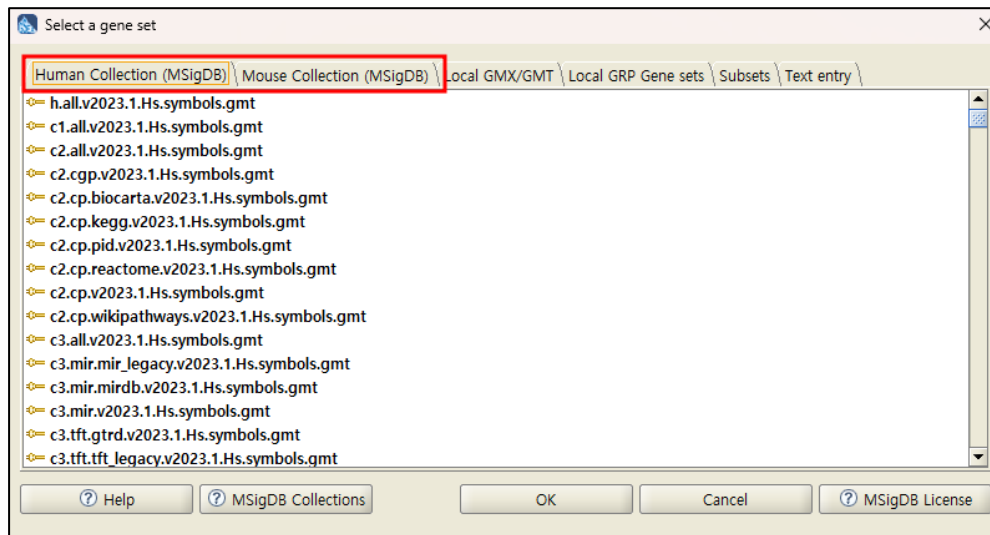


그림 8-8. Select options in GSEA program

Number of permutations은 기본값인 1000으로 기입하고, Phenotype labels은 분석하고자 하는 비교조합을 선택한다. 비교조합 선택 시, 순서는 Test (실험군) versus Control (대조군)인 점을 유의해야 한다.

Collapse/Remap to gene symbols 은 기본값인 Collapse 을 선택하고, permutation type 은 그룹 당 샘플 수가 7 개 미만이면 gene\_set 을 선택, 7 개 이상이면 phenotype 을 선택한다.

Chip platform 은 RNA-seq 의 경우 분석하려는 종의 탭을 선택하여 Human (or Mouse)\_Symbol\_with\_Remapping\_MSigDB~.chip 을 선택한다.

Rat 은 해당하는 탭이 없으므로 Human 또는 Mouse 탭에서 Rat\_Gene\_Symbol\_Remapping\_Human (or Mouse)\_Orthologs\_MSigDB~.chip 을 선택하여 진행한다 (그림 8-9). 여기서 주의할 점은, 이전 gene set 선택 과정(그림 8-8)에서 선택한 탭과 같은 탭으로 진행해야 한다.

Microarray 의 경우에는 실험한 chip 을 선택하여 진행한다.

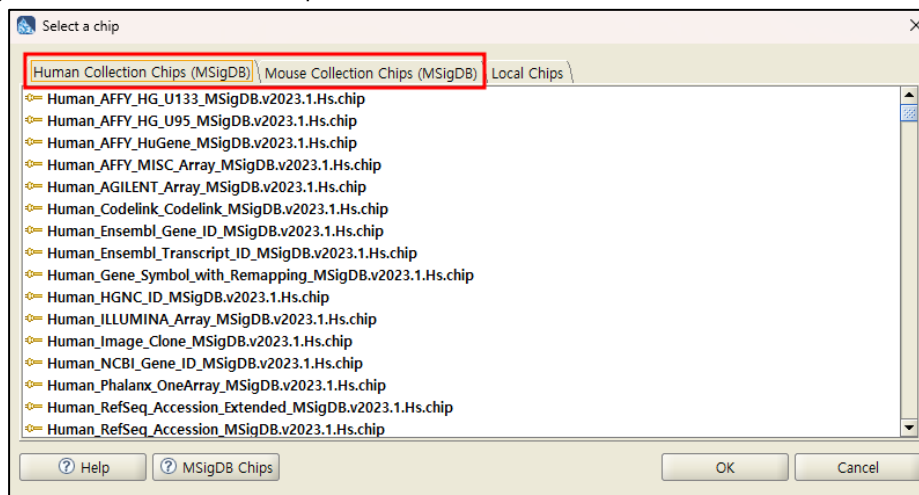


그림 8-9. Select options in GSEA program

모든 옵션이 선택이 완료되고 Run 버튼을 누르면 분석이 시작된다(그림 8-10).

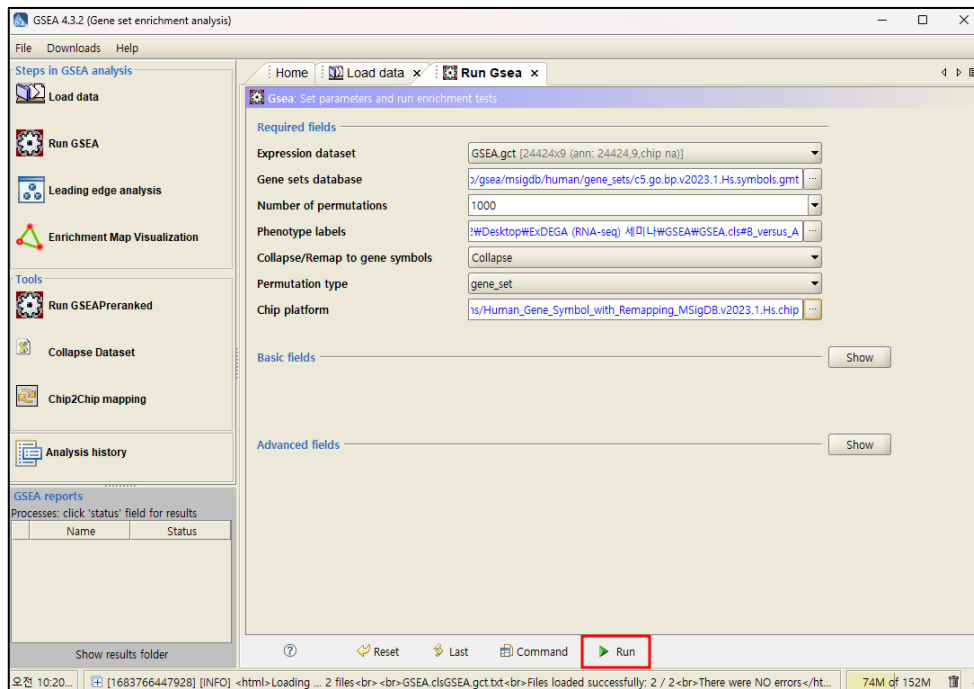


그림 8-10. Run GSEA program

분석이 완료되면 GSEA 왼쪽 아래 GSEA reports 창에 status가 Success로 바뀐다. Show results folder를 누르면 GSEA 분석 결과 창이 열린다(그림 8-11).

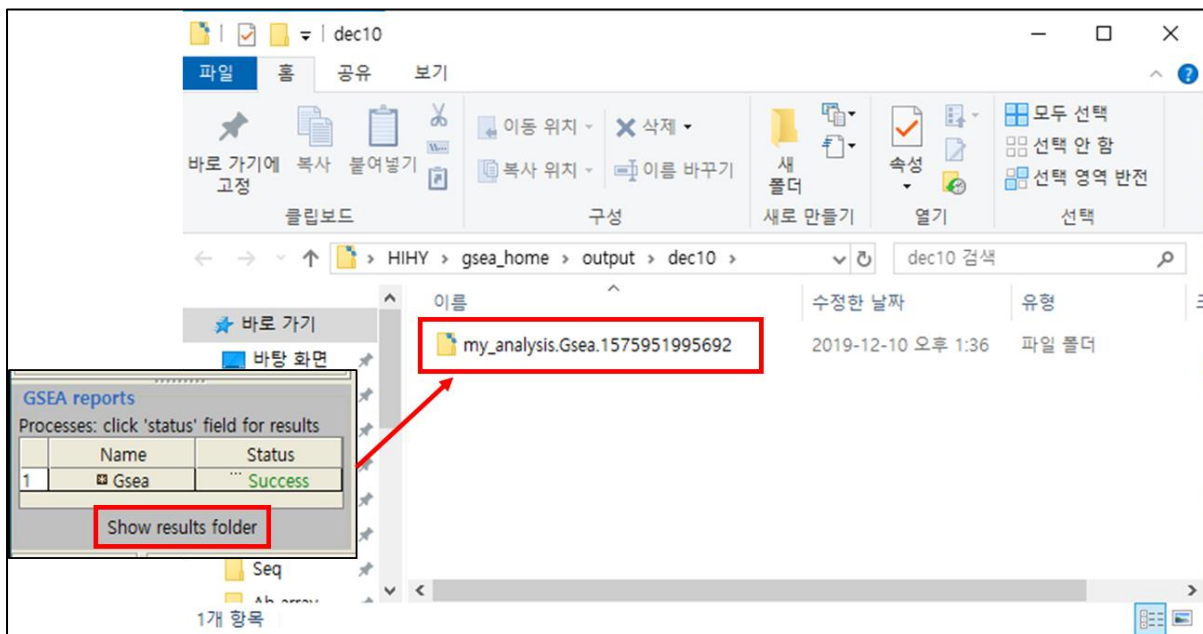


그림 8-11. GSEA results folder

GSEA 결과 중 중요 파일은 'gsea\_report\_for'로 시작하는 엑셀 파일이다. \_for 대조군 파일은 대조군에서 유의한 gene set, \_for 실험군 파일은 실험군에서 유의한 gene set이다(그림 8-12).

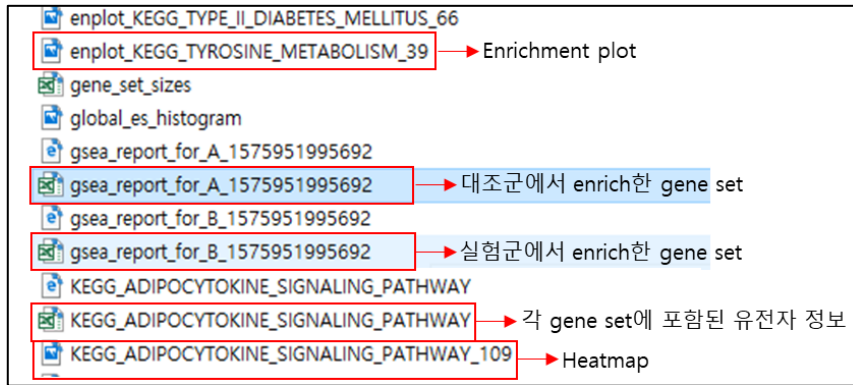
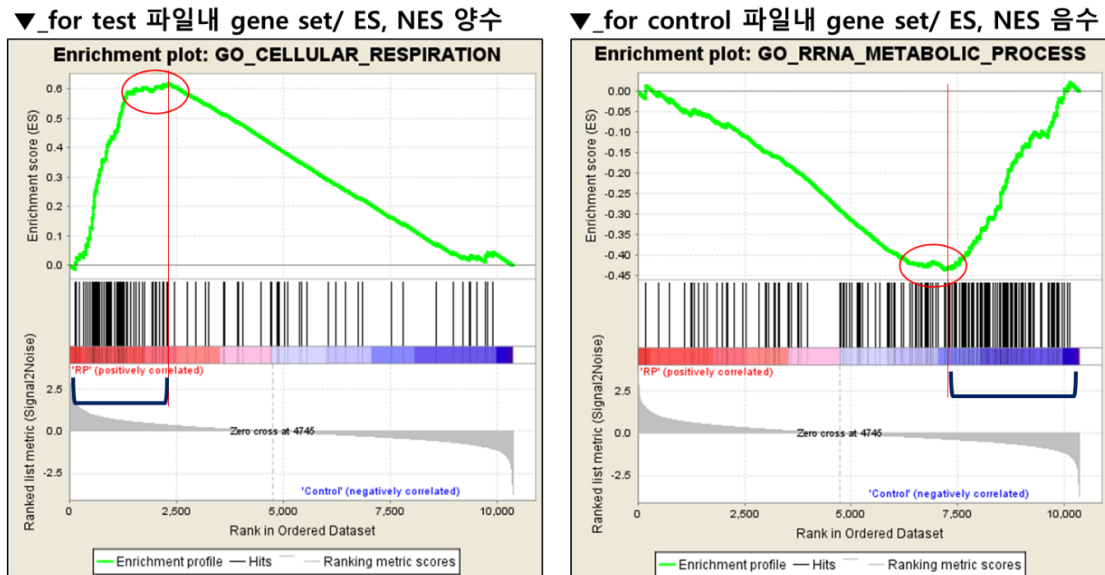


그림 8-12. GSEA result files

\_for 대조군 파일에는 enrichment score (ES)와 Normalized enrichment score (NES)가 음수, \_for 실험군 파일에는 ES와 NES는 양수다. 음수 양수와 관계없이 NES의 절대값이 큰 순서로 ranking 되어 있다. 음수는 DOWN (ranking 하위)에서 core gene의 밀집도가 있다는 것을, 양수는 UP (ranking 상위)에서 core gene의 밀집도가 있다는 것을 의미한다. NES 절대값이 높을수록 유의한 gene set이다. 상위 20개 gene set은 enrichment plot, heatmap, 각 gene set에 포함된 유전자들의 정보가 담긴 excel file이 있다. GSEA 분석 결과 중 Enrichment plot이 논문에 많이 실린다. Enrichment plot 이미지에서 세로 선이 해당 gene set에 포함된 유전자들이며 fold change 순으로 나열된다(그림 7-13). Peak가 왼쪽에 생기면 대조군 대비 실험군에서 up된 유전자들이 많다는 의미이고, peak가 오른쪽에 생기면 down된 유전자가 많다는 의미이다.



core enrichment(=core gene) 영역, 관련된 유전자영역이 밀집되어 있는 곳

그림 8-13. GSEA enrichment plot

GSEA 분석과정 및 결과에 대한 카테고리의 자세한 의미는 GSEA user guide (<https://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html>)에서 확인할 수 있다.



## 9. Protein-Protein Network Analysis (Cytoscape STRING)

STRING tool 은 Protein-Protein Interaction 데이터 베이스를 기반으로 유전자의 상관관계를 통계적으로 분석하여 Network 을 작성해주는 분석 툴이다. 분석 과정은 그림 9-1 과 같다.

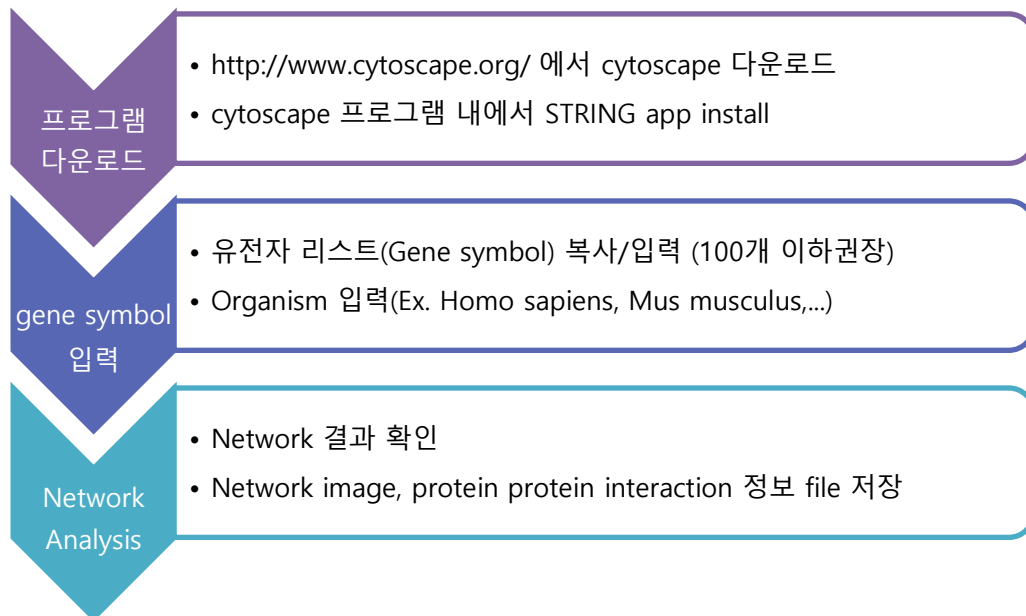


그림 9-1. STRING analysis process

Cytoscape 홈페이지 (<http://www.cytoscape.org/>)에서 cytoscape 프로그램을 다운로드 받아 설치한다(그림 9-2)(그림의 버전과 상관없이 최신 버전으로 다운받으면 된다.).



그림 9-2. Cytoscape download

Cytoscape 프로그램을 열어 상위에 있는 메뉴 중 [Apps] > [App Store] > [Show App Store]로 들어간다. 빈 칸에 String 을 검색하고 연동되어 검색된 웹사이트에서 StringApp 을 선택 후 Install 버튼을 누른다 (그림 9-3).

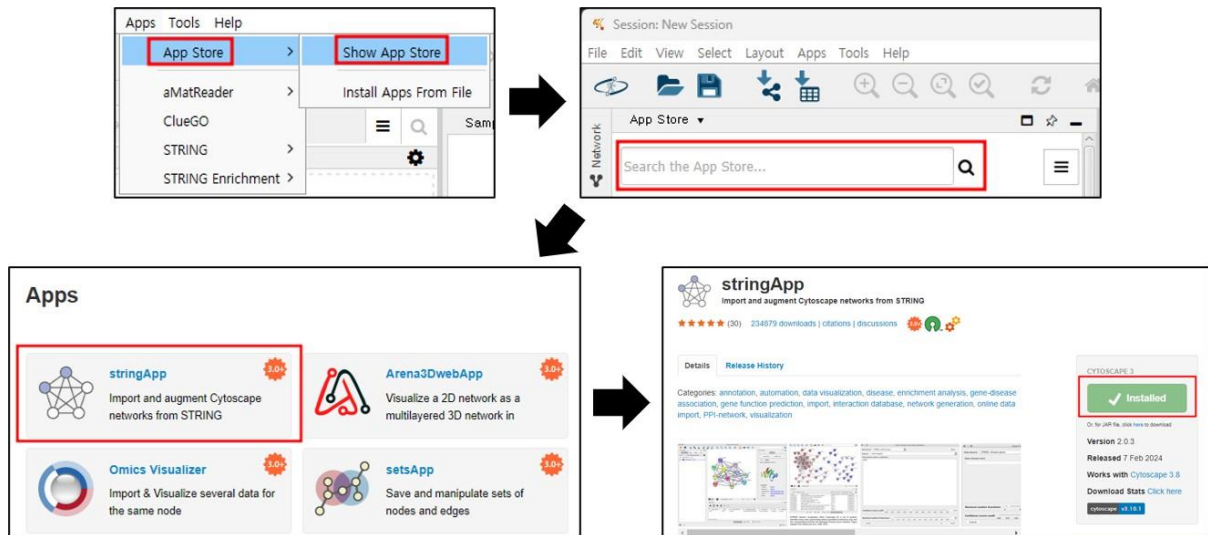


그림 9-3. STRING app installation in Cytoscape

Cytoscape 상위 메뉴 중 [File] > [Import] > [Network from Public Databases]로 들어간다(그림 9-4).

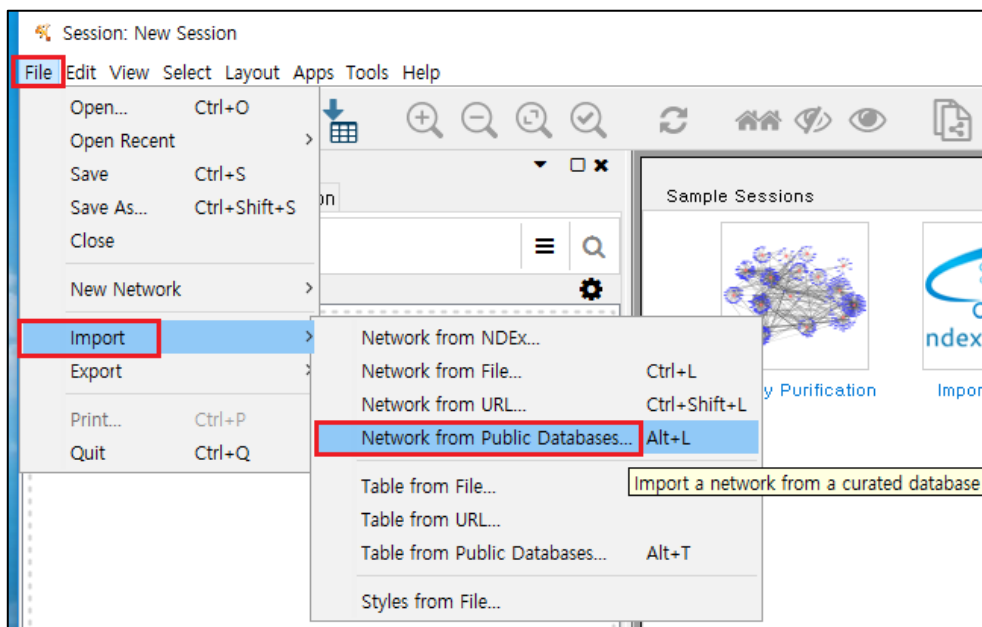


그림 9-4. STRING analysis process 1

Data Source 를 “STRING : protein query” 선택하고 Species 를 선택한다(그림 9-5). 분석하고자 하는 유전자들의 gene symbol 을 입력한다. Confidence (score)는 Protein-Protein Interaction 강도를 뜻하는 것으로 0 부터 1 까지이고, 1 로 갈수록 Interaction 이 강함을 의미한다. Maximum additional interactors 를 0 으로 하면 input 한 유전자 안에서만 network 이 그려지고 숫자를 높이면 input 하지 않은 neighborhood protein 까지 network 이 그려진다.

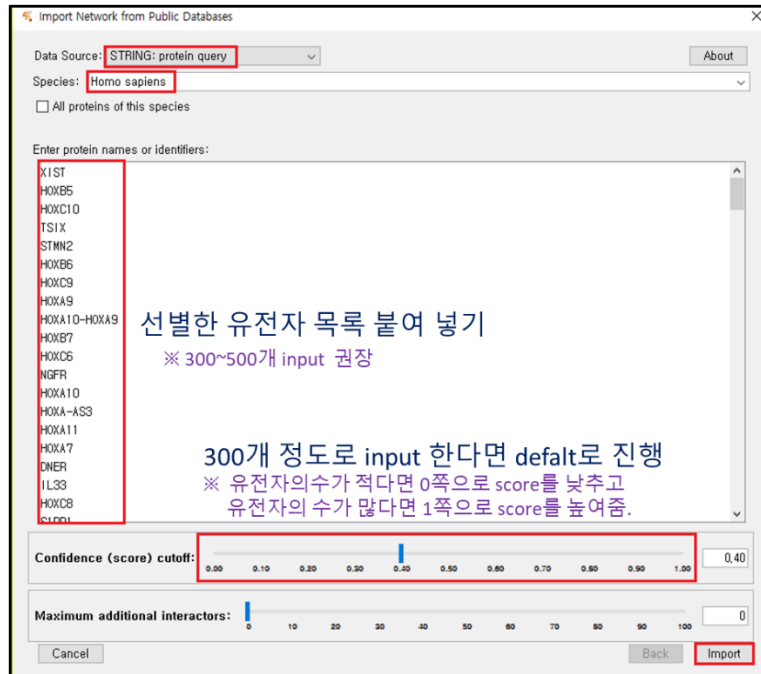


그림 9-5. STRING analysis process 2

Input 한 gene symbol 과 match 가 되지 않는 protein 이 있으면 그림 9-6 과 같은 화면이 나온다. 두 개 이상의 protein 이름이 나타나는 경우는 유사 protein 을 확인하라고 한다. 연구자의 선택에 따라 모두 check 또는 해지한다. Import 버튼을 누르면 분석이 진행된다.

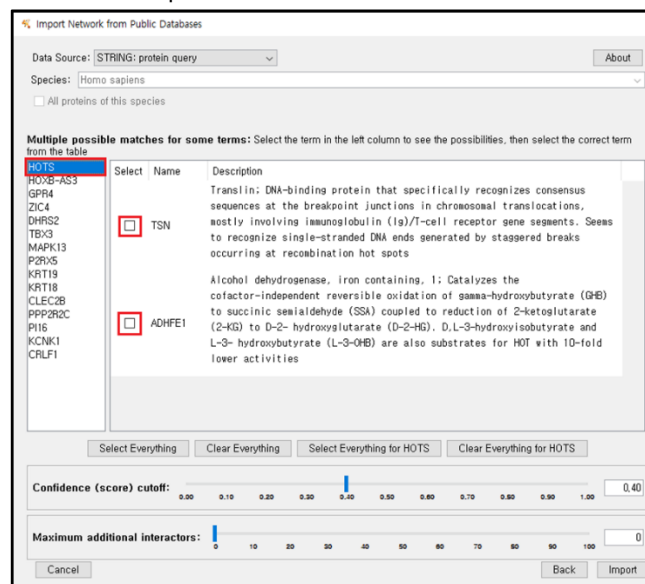


그림 9-6. Not matched proteins in STRING

분석이 완료되면 network image 가 나온다(그림 9-7).

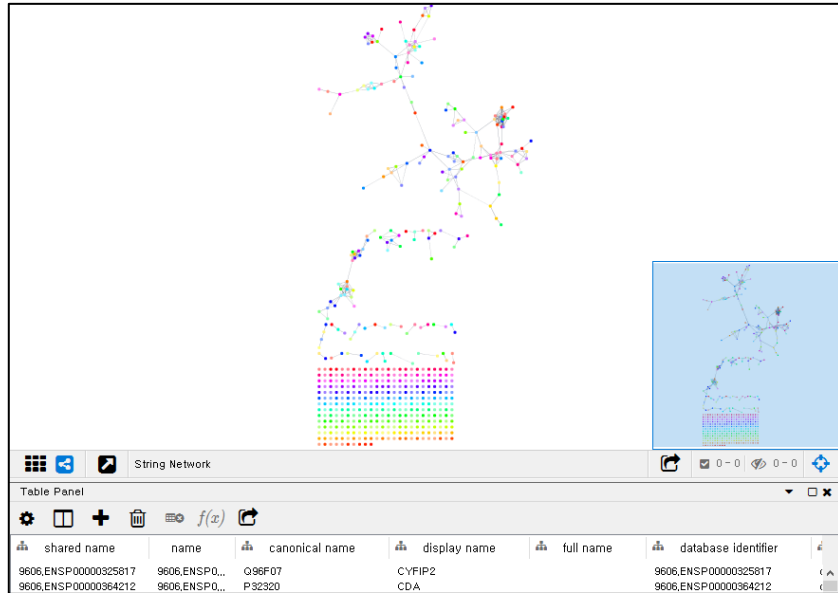


그림 9-7. Network result

[File] > [Export] > [Network to Image]를 눌러 이미지를 저장한다(그림 9-8). PDF 파일형식으로 저장하는 것을 권장한다. Pdf 파일로 저장하면 확대를 하여도 이미지가 깨지지 않는다.

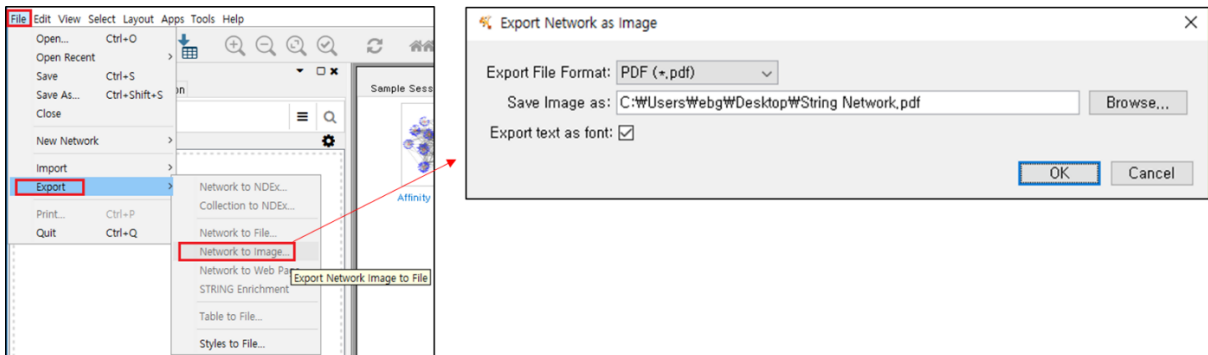


그림 9-8. Save network image

어떤 유전자들이 protein-protein interaction 을 하는지 정보를 저장하고 싶으면 [File] > [Export] > [Table to File...]로 들어가 String Network default edge 파일을 저장한다(그림 9-9).

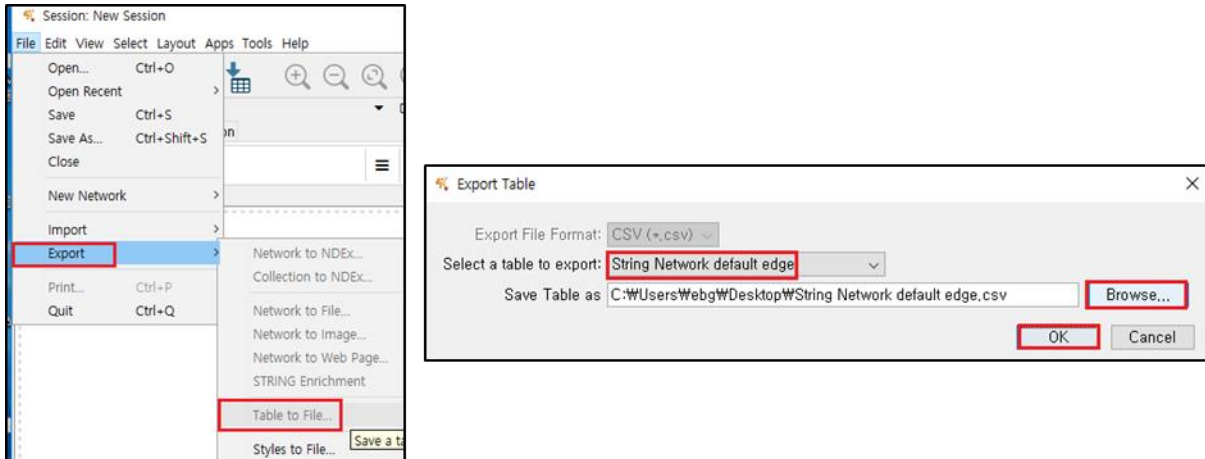


그림 9-9. Save edge table

String Network default edge 파일에서 name 에 interaction 정보, score 에 confidence score 가 나와있다(그림 9-10). Name 에 A (pp) B 라고 적혀있으면 A 유전자와 B 유전자가 Protein-Protein Interaction 한다는 것이고 score 값이 1 에 가까울수록 interaction 이 강한 것이다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	SUID	coexpri	cooccur	database	experiri	fusion	interac	intersp	name	neighb	score	selecte	shared	shared	textmi
2	701	0.397		0.9	0.961		pp		EFNB2 (pp) EPHA4		1.000	FALSE	pp	EFNB2 (pp)	0.926
3	702	0.397		0.9	0.817		pp		EFNB2 (pp) EPHB1		0.999	FALSE	pp	EFNB2 (pp)	0.887
4	961	0.929		0.9			pp		IFIT1 (pp) MX1	0.064	0.998	FALSE	pp	IFIT1 (pp)	0.722
5	700	0.397		0.9	0.76		pp		EFNB2 (pp) EPHA5		0.997	FALSE	pp	EFNB2 (pp)	0.768
6	680	0.878		0.9	0.292		pp		OAS3 (pp) IFIT1		0.996	FALSE	pp	OAS3 (pp)	0.519
7	683	0.888		0.9	0.132		pp		OAS3 (pp) MX1		0.995	FALSE	pp	OAS3 (pp)	0.506
8	935	0.892		0.9			pp		IFI6 (pp) MX1		0.995	FALSE	pp	IFI6 (pp) N	0.503

그림 9-10. Interaction information in edge table

Network image 에서 색이나 모양을 변경하고 싶은 경우에는 STRING Manual ([Download link](#))에서 image 수정 방법을 확인할 수 있다.