ebiogen

User Manual

ExDEGA v.5.0.0
& Data Analysis

Tel. 02-3141-0791 Fax. 02-3141-0792 E-mail: service@e-biogen.com

<목 차>

1. Differentially Expressed Genes Analysis (ExDEGA) ---------- 3

2. Functional Annotation Analysis (DAVID, ExDEGA GraphicPlus) ------- 22

3. Clustering Heatmap Analysis (Clustering, ExDEGA GraphicPlus) ------ 34

4. Principal Component Analysis (PCA, ExDEGA GraphicPlus) ---------- 39

5. String Network Analysis (Network, ExDEGA GraphicPlus) ---------- 48

6. Pathway Analysis (KEGG mapper) ---------- 52

7. Gene Set Enrichment Analysis (GSEA) ---------- 55

8. Protein-Protein Network Analysis (Cytoscape STRING) ---------- 63

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 Graphic Plus 를 함께 제공한다. ExDEGA 분석 툴은 ㈜이바이오젠이 연구자들이 Microarray 및 RNA-Seq 데이터를 보다 쉽게 다루고 원하는 데이터를 쉽게 얻을 수 있도록 사용자 편의를 최대한 반영한 분석 툴이고 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 1-1 A). 만약 다운로드 폴더에서 압축을 풀지 않았을 경우, 별도로 압축 푼 파일에 있는 ExDEGAGraphicPlus.exe 를 컴퓨터의 로컬 C 드라이브 아래로 복사+붙여넣기 하면 ExDEGA Graphic Plus 프로그램이 설치 완료된다(그림 1-1 B).

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.

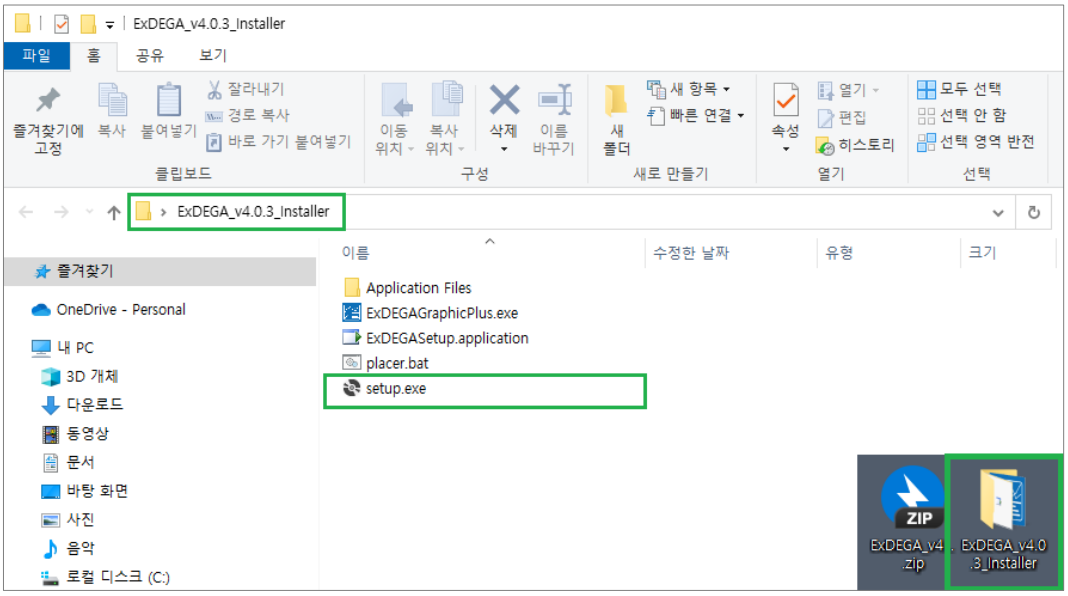


그림 1-1 A. ExDEGA set up

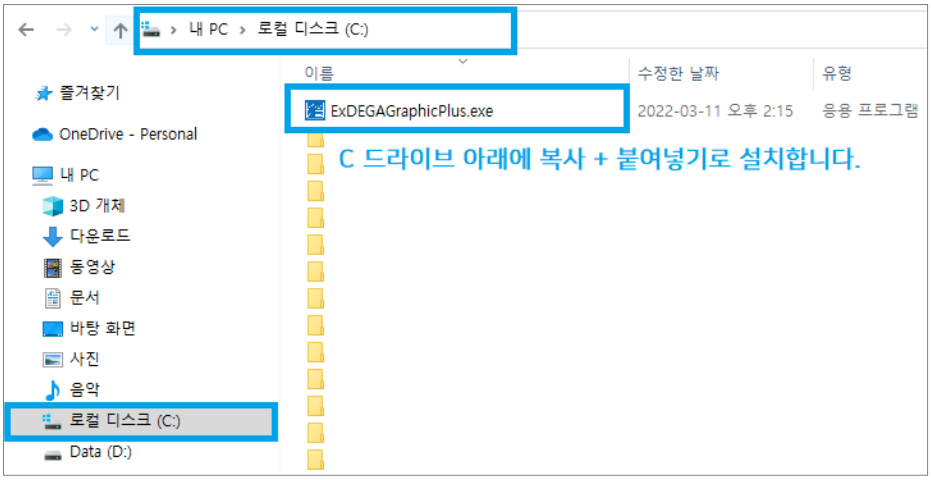


그림 1-1 B. ExDEGA Graphic Plus installation

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 창 에서는 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.

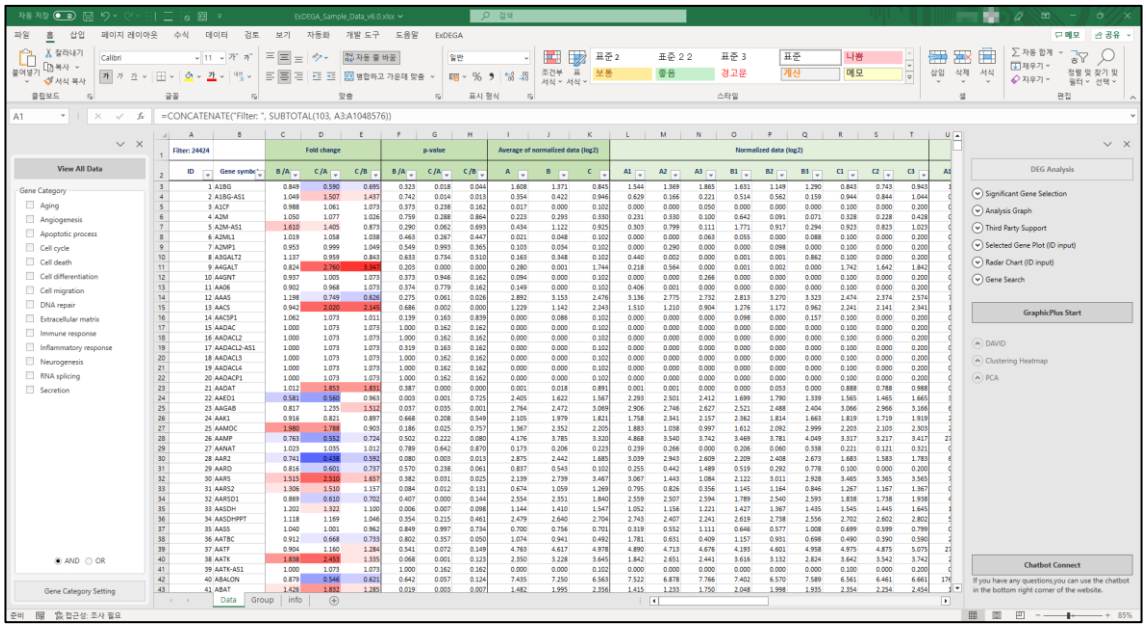


그림 1-2. RNA-seq or Microarray data in ExDEGA format

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)

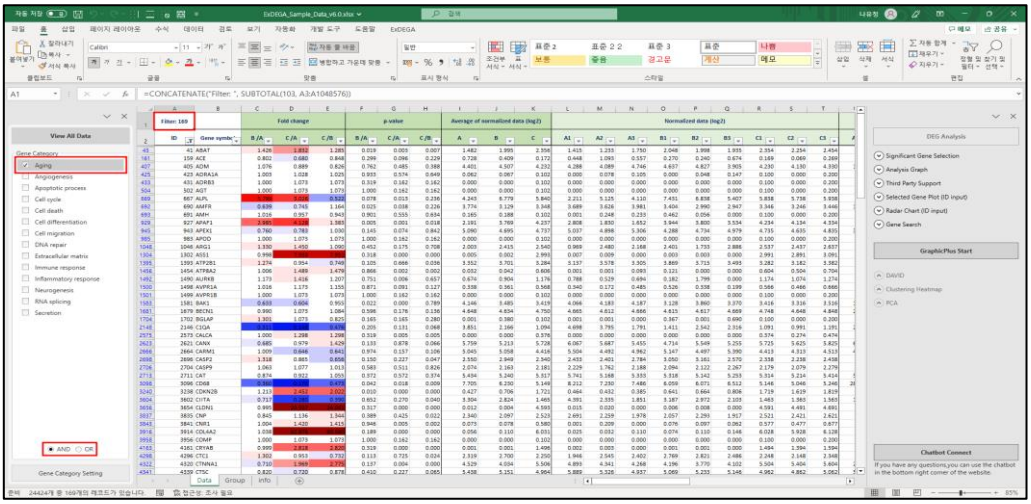


그림 1-3. Gene ontology selection

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(그림 1-4). '?' 버튼을 누르면 GO 추가하는 방법이 자세히 설명되어 있다.

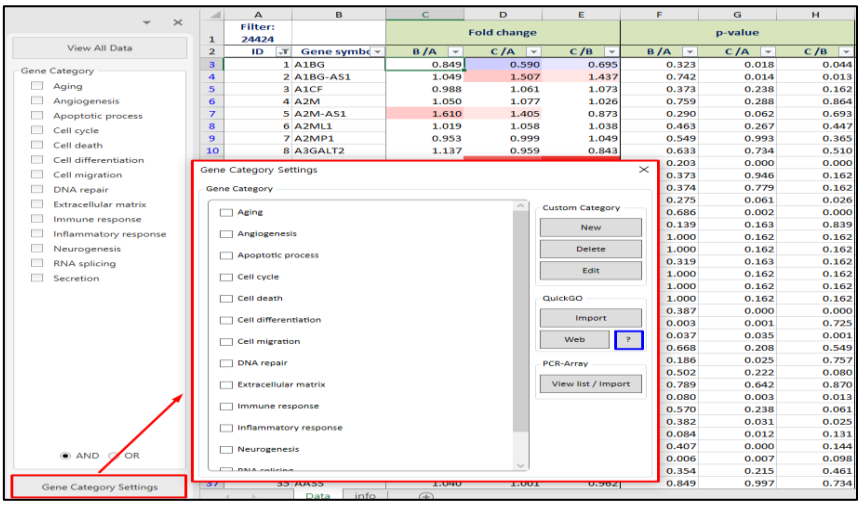
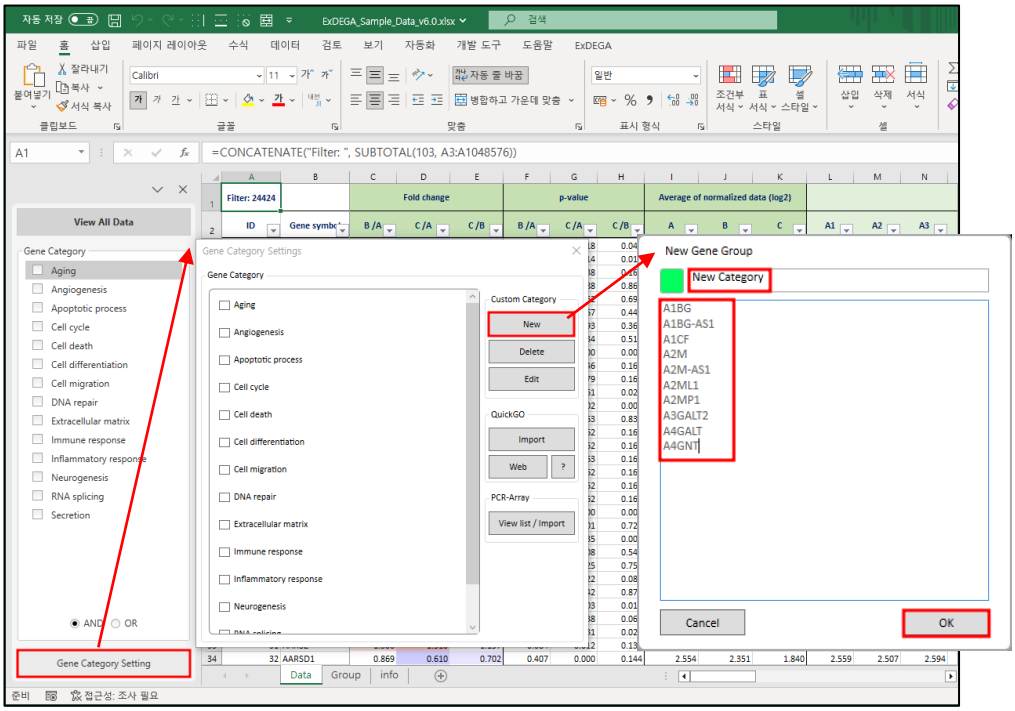


그림 1-4. Gene category settings

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 1-5. a, b).

a.



b.

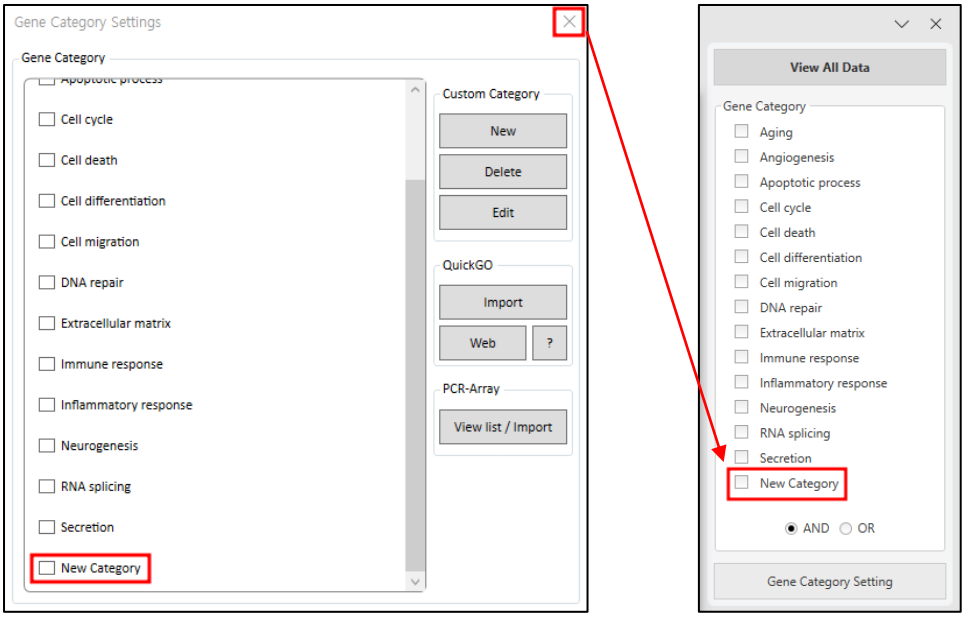


그림 1-5. Adding Genes to make a new gene category

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

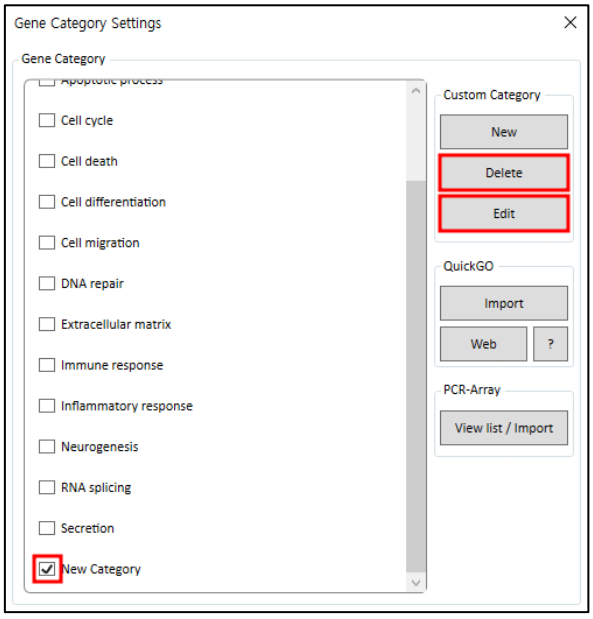


그림 1-6. Edit or Delete gene category

PCR-Array 항목의 View list / Import 를 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).

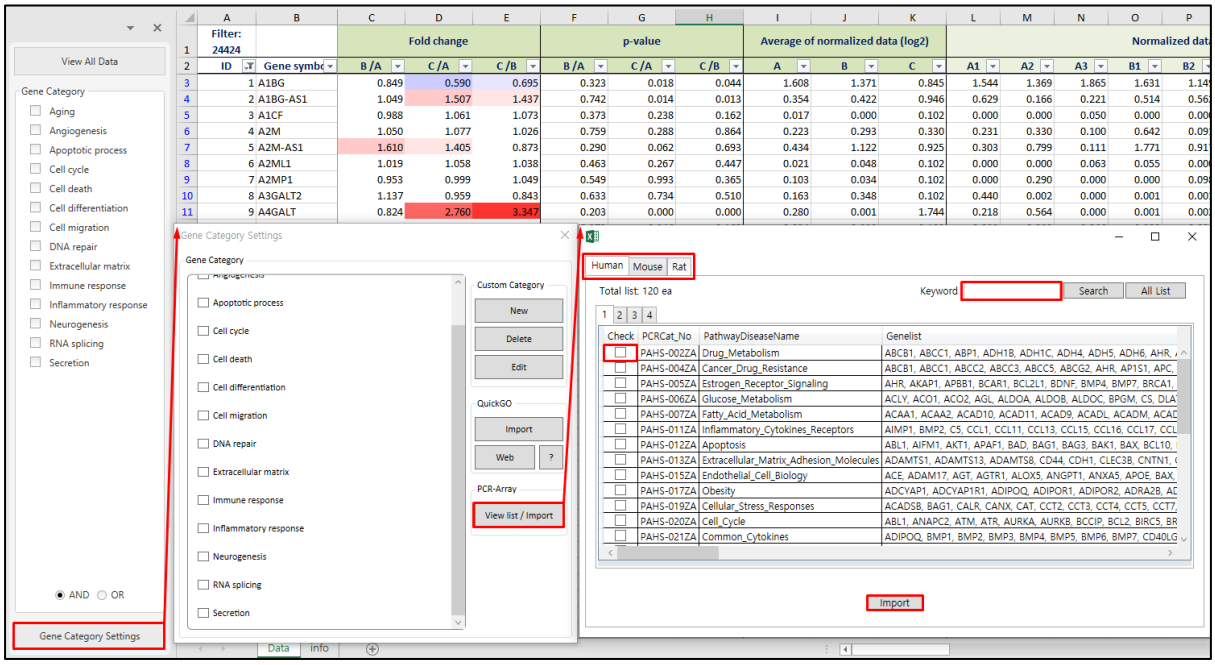


그림 1-7. PCR-Array Pathway settings

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 기준은 연구자의 데이터에 맞게 조정하여 사용할 수 있다.

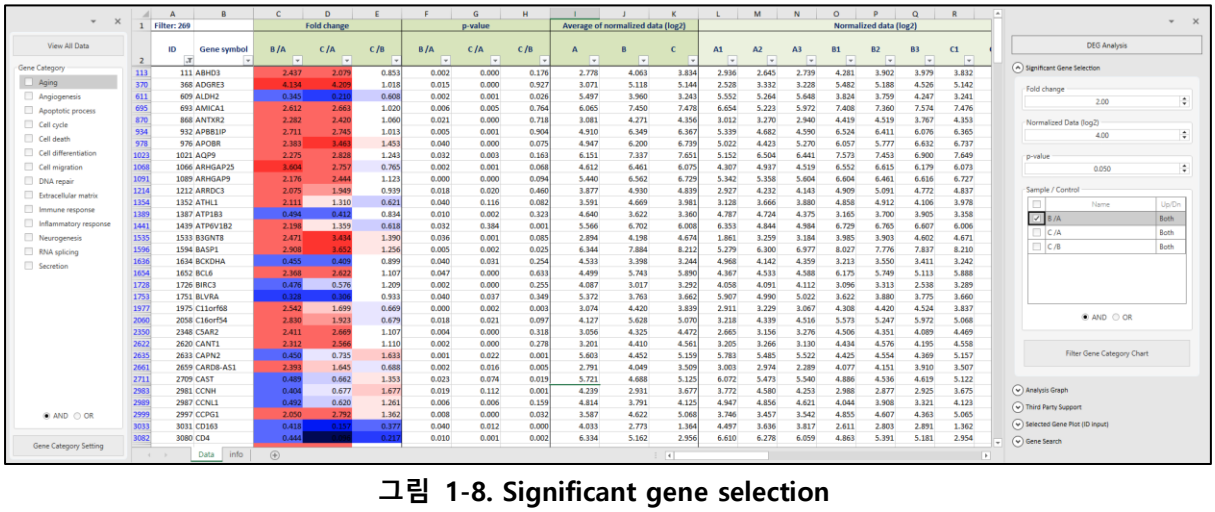


그림 1-8. Significant gene selection

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).

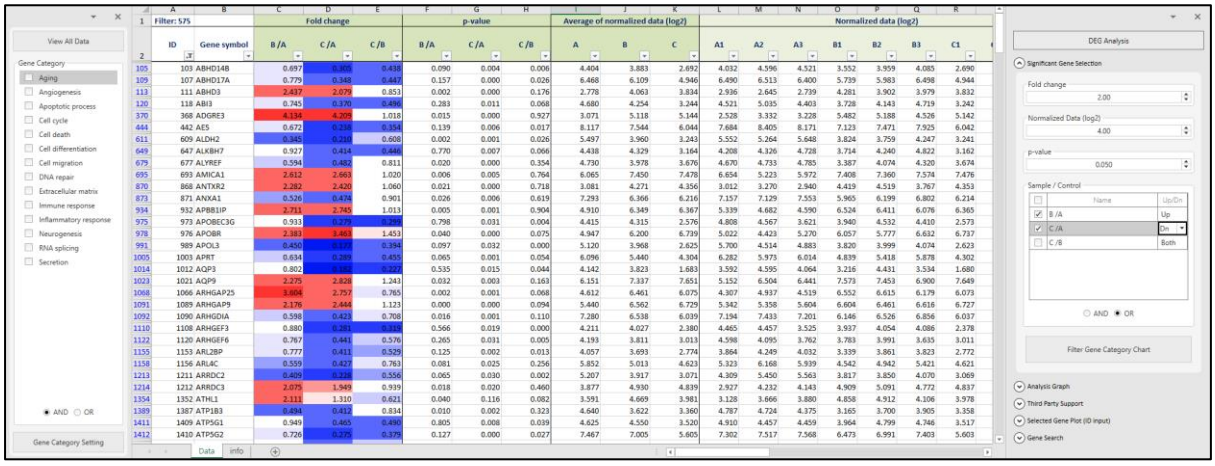


그림 1-8. Significant genes (separately up and down)

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

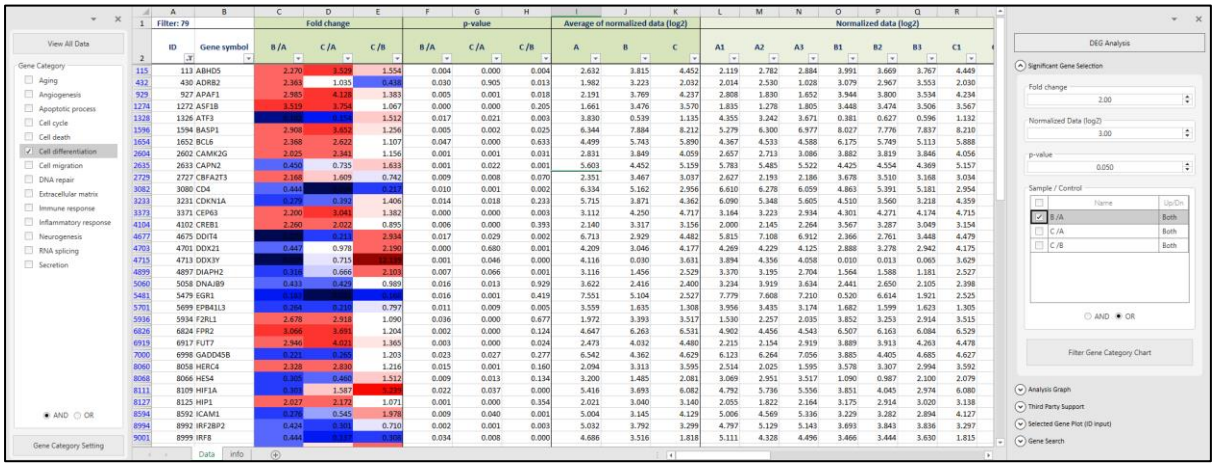


그림 1-9. Significant genes related to Cell differentiation

Gene Category Chart 는 각 GO 관련 유전자 중 발현이 유의하게 차이 나는 유전자의 %와 수를 나타낸 그래프이다. 본 분석을 통해 어떤 GO 의 유전자들이 상대적으로 많은 발현 변화가 있었는지를 확인할 수 있다. 전체 데이터 상태에서 Significant Gene Selection 의 비교 그룹을 선택하고 "View Gene Category Chart"를 클릭하면 증가/감소한 유전자들 대상으로 GO Chart 가 생성된다. 그래프의 각 영역을 클릭하면 해당 유전자들이 필터링 된다. 예를 들어 왼쪽의 Pie chart 의 특정영역을 클릭하면 해당 GO 의 증가/감소된 유전자가 함께 필터링 된다. 오른쪽의 bar chart 에서 bar 상단의 숫자는 증가/감소에 해당하는 유전자 수이며, bar 는 각각 Up significant(%)와 Down significant(%)를 의미한다. bar 를 클릭했을 때 해당 유전자가 필터링 된다(그림 1-10).

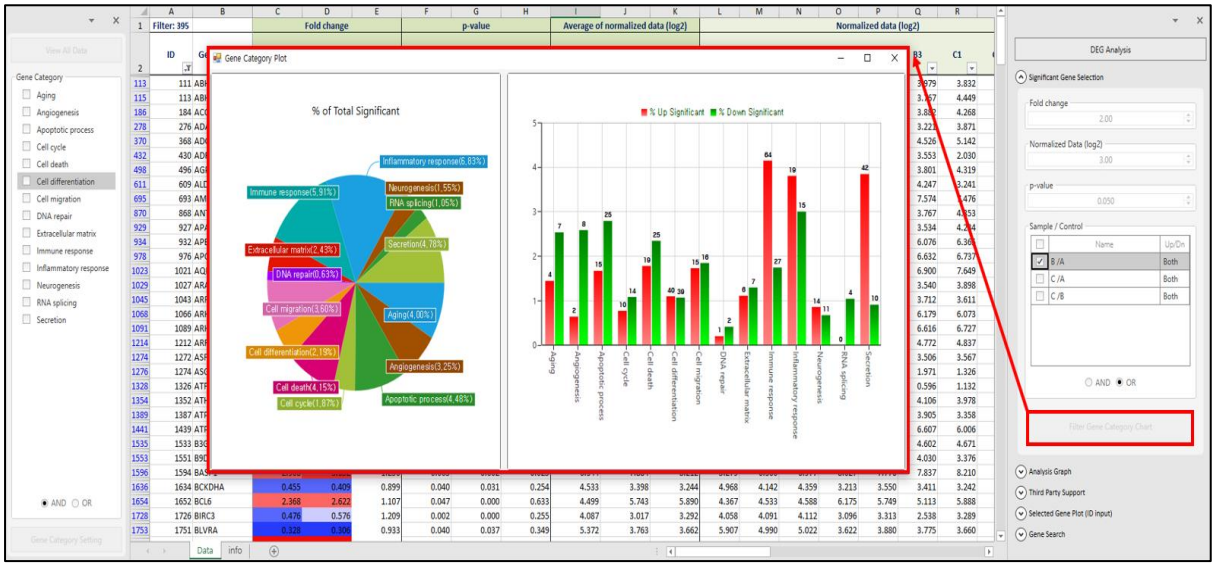


그림 1-10. View Gene Category Chart

significant chart 는 선택한 비교조합에 따른 유의한 유전자(최대 30 개)의 발현 값을 그룹 별로 확인할 수 있다. 유의한 유전자 필터링 기준은 선택한 비교조합의 p-value 순으로 표시된다. 차트의 x 축은 유전자 이름이고 y 축은 Normalized data (log2) 로 구성되어 있다. 또한 각 점(dot)은 샘플 하나를 의미하고 그룹에 따라 색이 구분된다. 마우스 커서를 각 점(dot)위에 올리면 해당되는 샘플 명도 알 수 있다(그림 1-11).

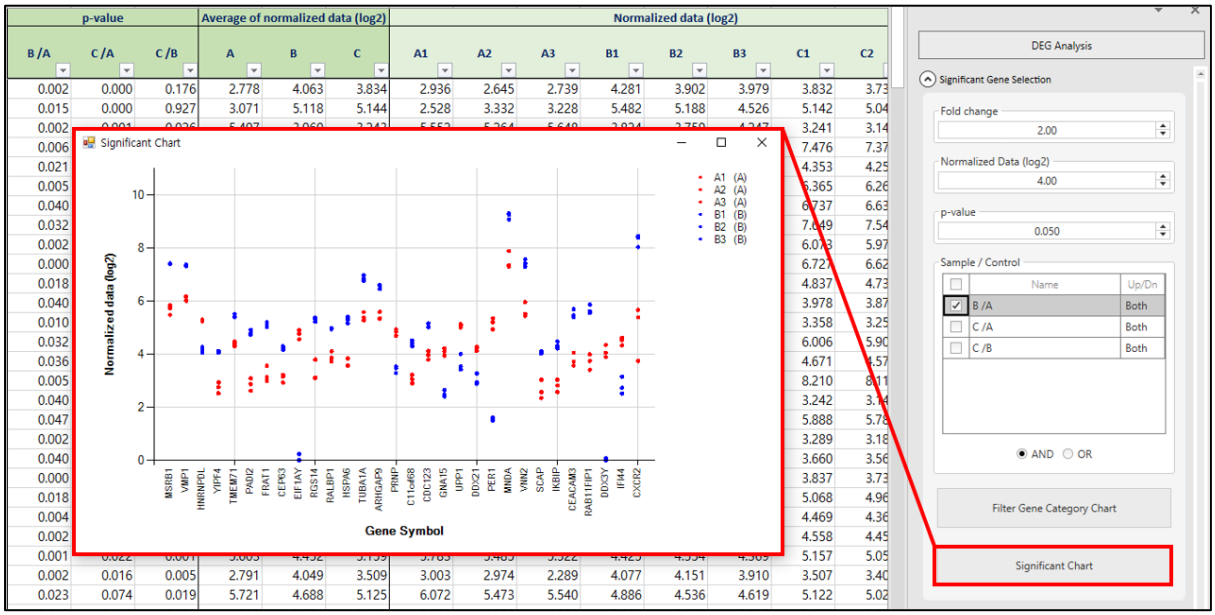


그림 1-11. Significant Chart

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 쉽게 그릴 수 있다.

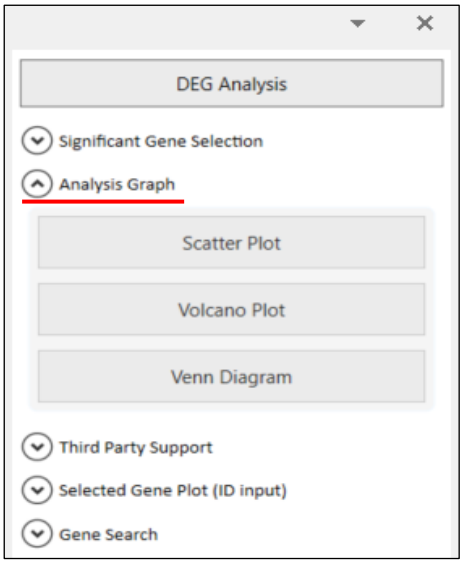


그림 1-12. Analysis Graph Tool

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 유전자 ID 를 복사하여 입력하고 "Add"를 클릭하면 Gene Symbol 이 자동 생성된다(그림 1-13).

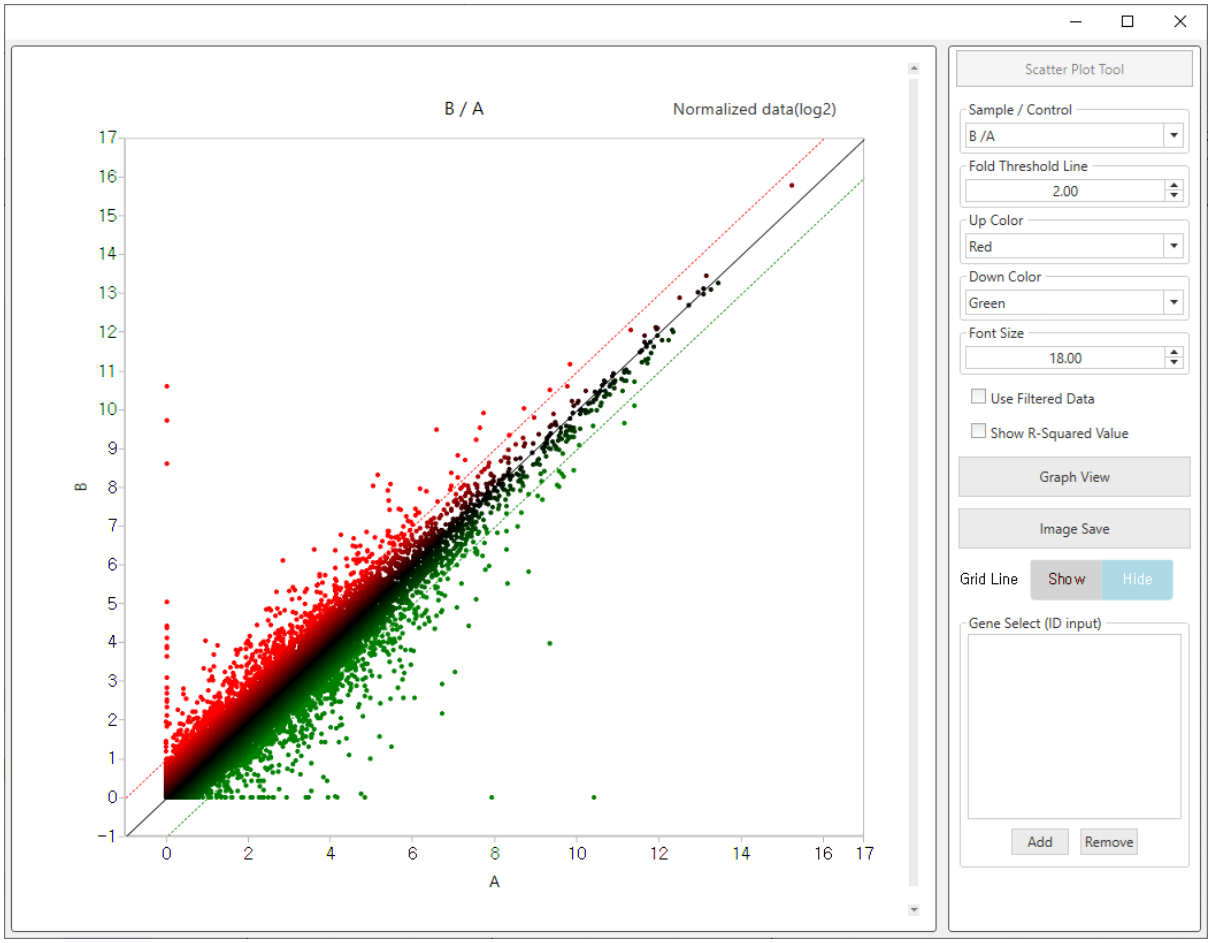


그림 1-13. Analysis Graph Tool – Scatter Plot

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).

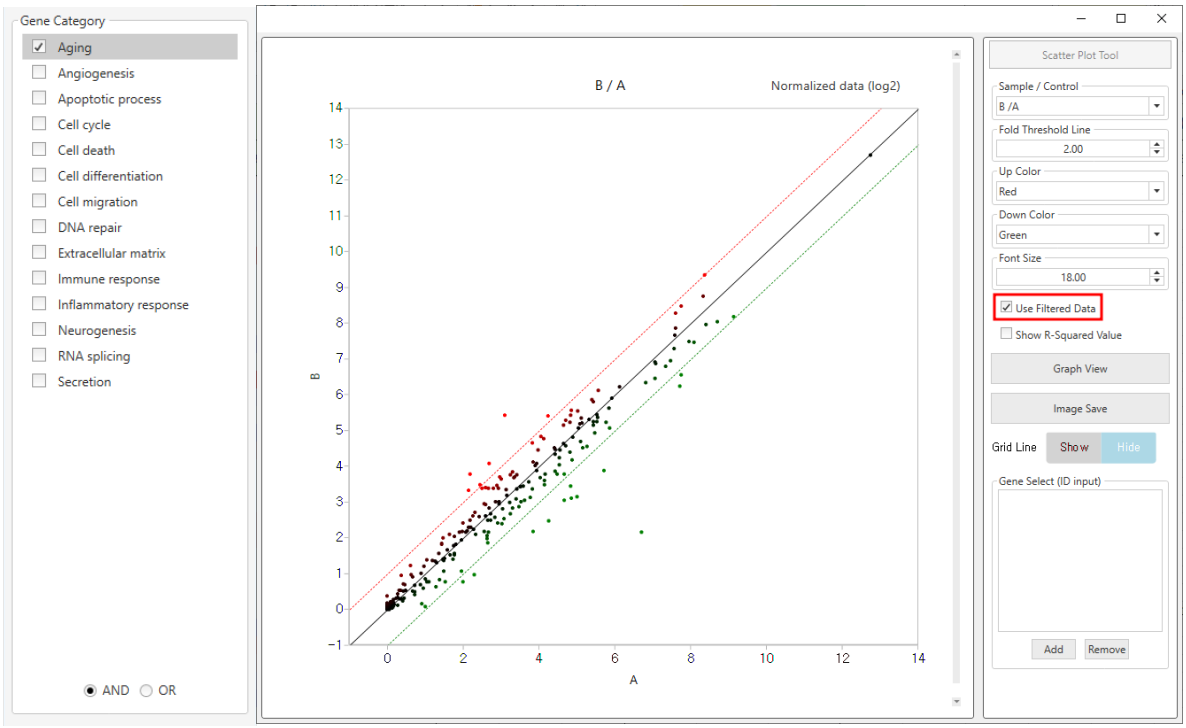


그림 1-14. Analysis Graph Tool – Use Filtered Data

1-3-2. Volcano plot

Volcano Plot 은 반복 실험(N>=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 유전자 ID 를 복사하여 입력하고 "Add"를 클릭하면 Gene Symbol 이 자동 생성된다(그림 1-15).

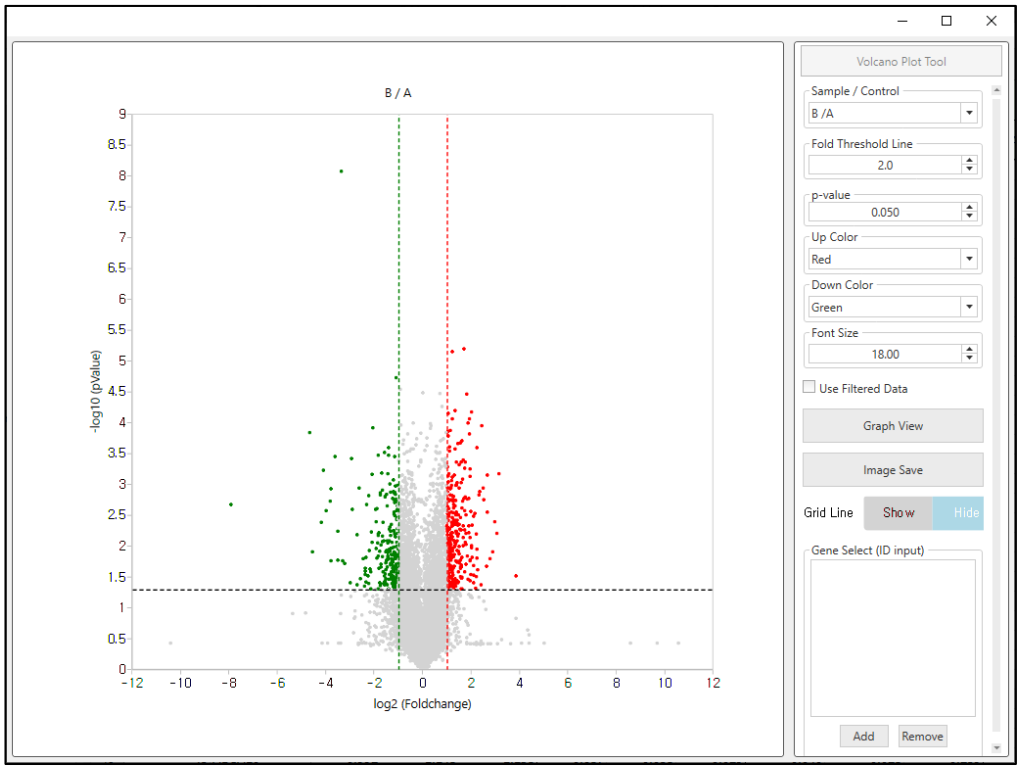


그림 1-15. Analysis Graph Tool – Volcano Plot

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 결과 중, 2fold, Normalized data (log2) 4 이상, p-value 0.05 이하인 유전자 list 를 가지고 Venn Diagram 을 작성한 결과이다(그림 1-16).

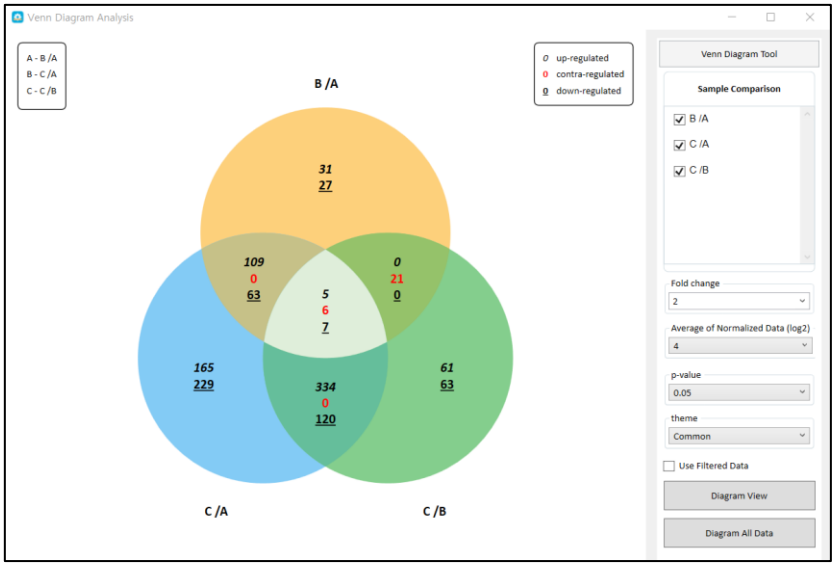


그림 1-16. Analysis Graph Tool – Venn Diagram

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).

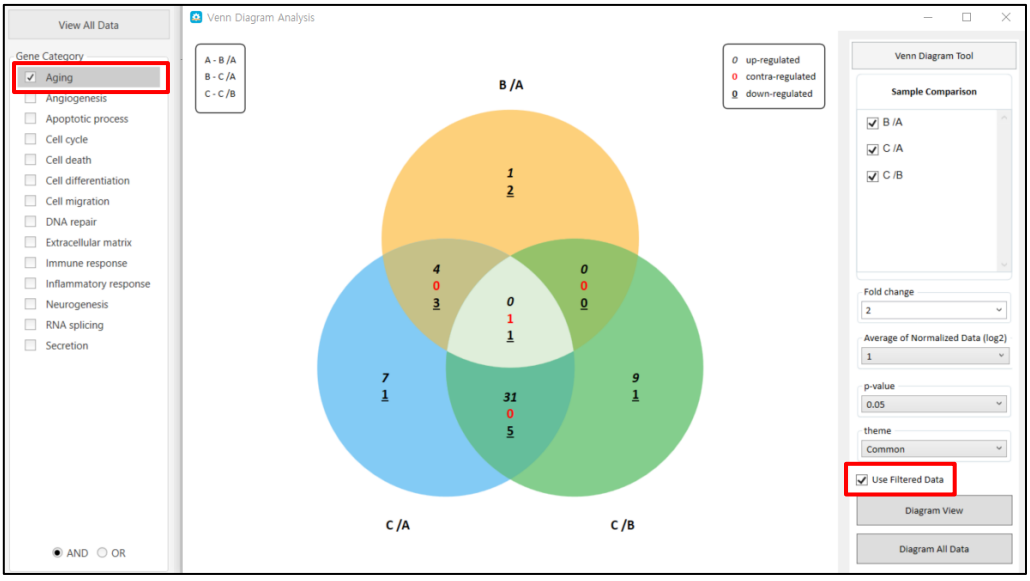


그림 1-17. Use Filtered Data

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

1. ***기울어진 숫자*** : up-regulated 된 gene 수
2. **빨간색 숫자** : regulation 이 대조되는 gene 수

(예: B/A 에서는 up 되고 C/A 에서는 down 되는 gene 수)

3. **밑줄 친 숫자** : down-regulated 된 gene 수

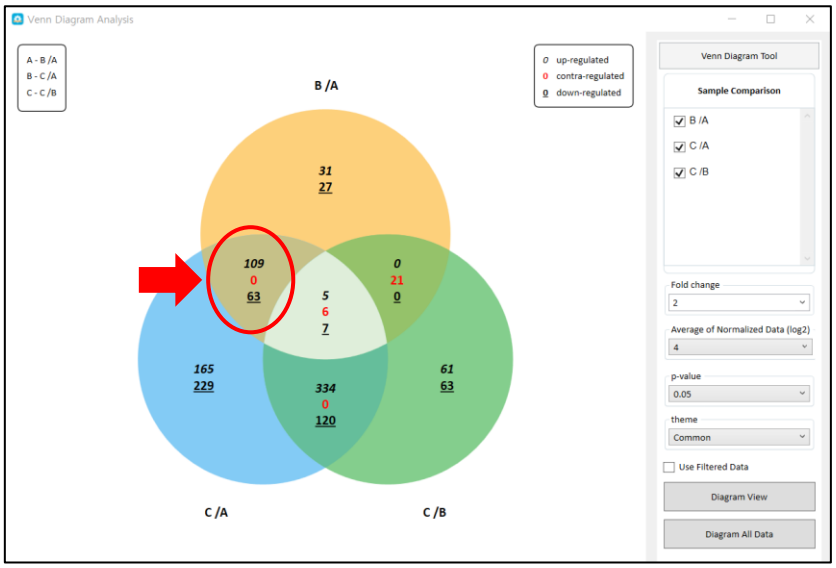


그림 1-18. For example of up, down, contra-regulated in Venn Diagram

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).

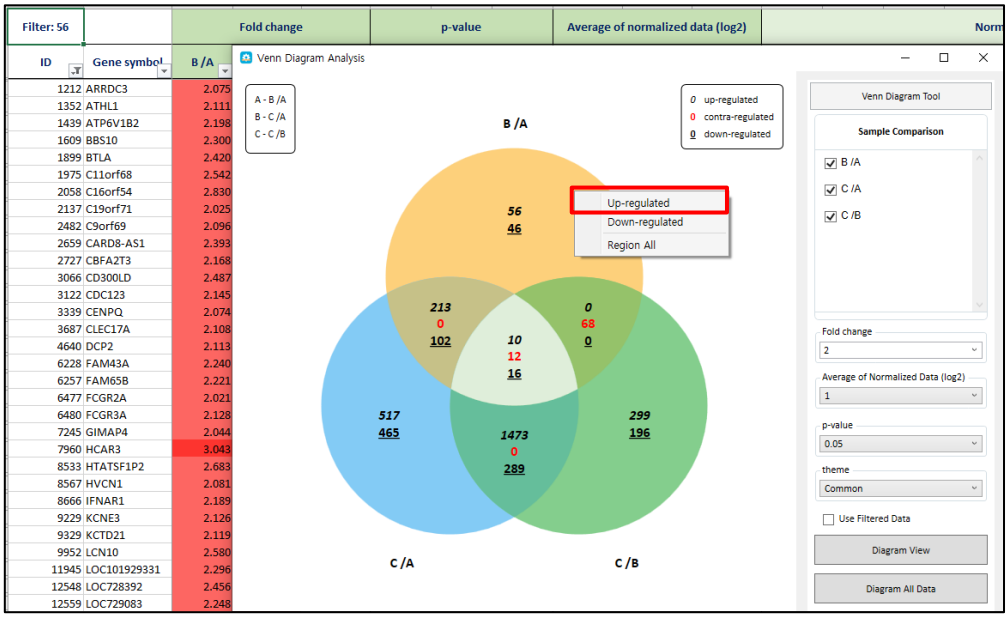


그림 1-19. Filtering 2fold up-regulated gene list in Venn Diagram

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

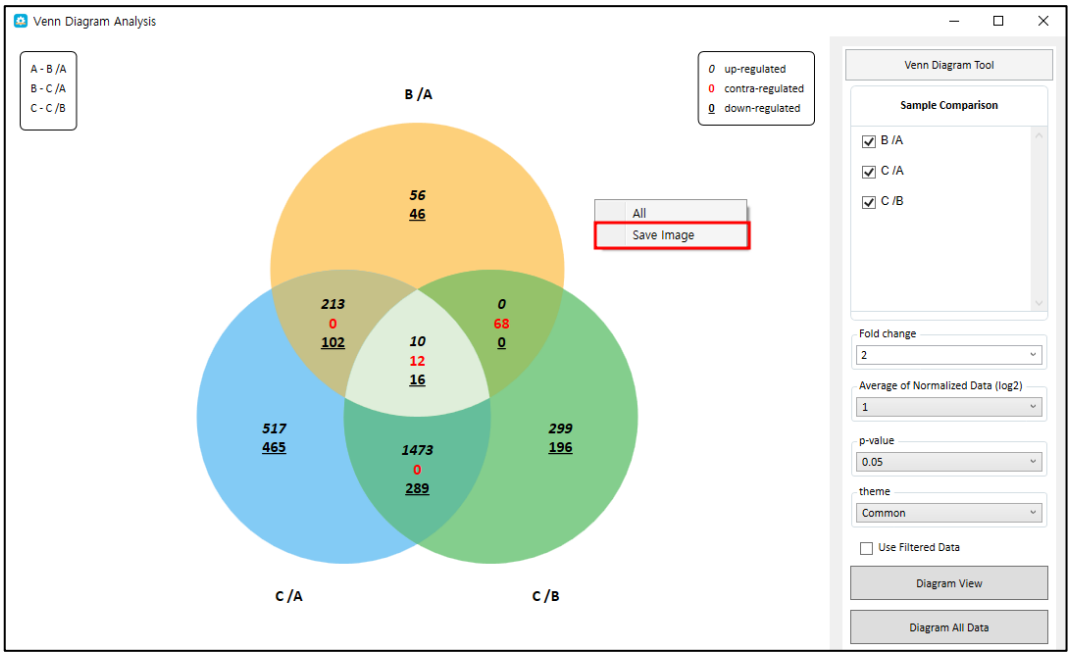


그림 1-20. Save image

1-4. Third Party Support 사용 방법

Third Party Support 는 연구자가 선별한 유전자를 기반으로 Clustering heatmap 과 DAVID 및 KEGG 분석을 수행하기 위한 입력 데이터를 제공한다. (단, GSEA의 경우 선별하지 않고 전체 유전자 리스트를 기반으로 진행하는 것을 권장한다.) 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).

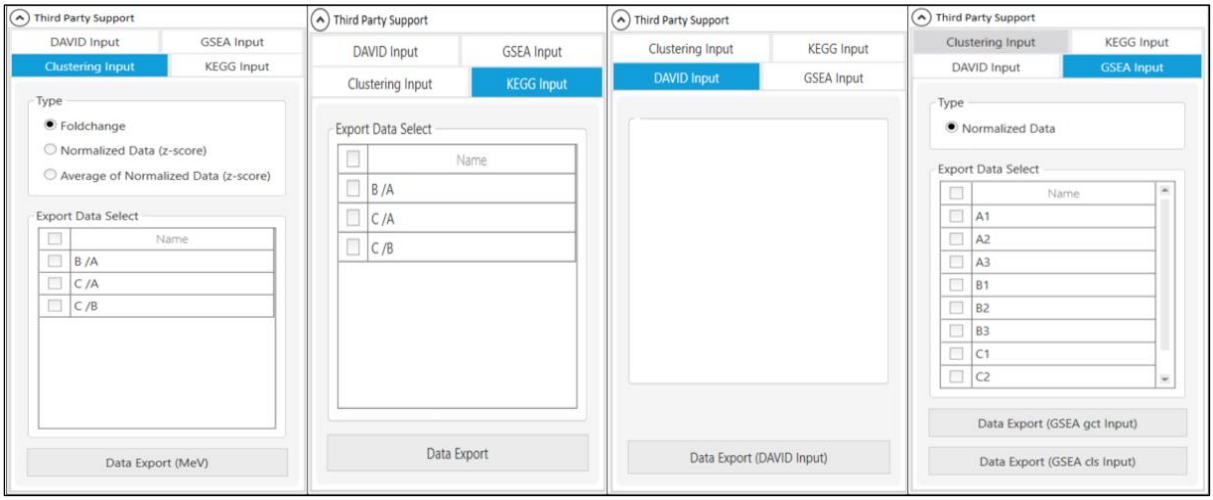


그림 1-21. Third Party Support

(왼편부터 Clustering Input, KEGG Input, DAVID Input, GSEA 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 에서 heatmap 을 작성하거나 MeV 프로그램을 이용하여 heatmap 을 작성할 수 있다.

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 로 그릴 때는 표준편차가 고려되기 때문에 샘플 3 개 이상에서만 가능하다

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

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 exel 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.

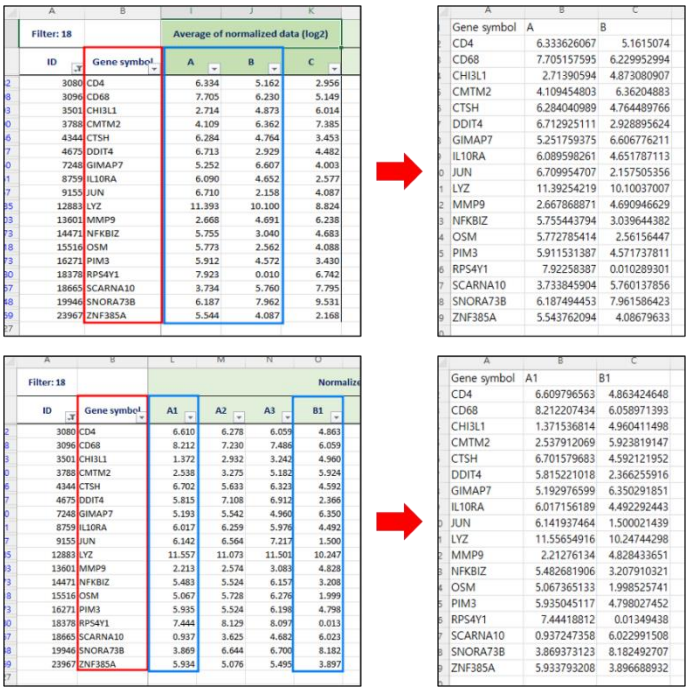


그림 1-22. Heatmap input of no more than 2 samples or groups

* 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-score = {Normalized data (log10) – average of Normalized data (log10)}
/standard deviation of Normalized data(log10)

Clustering heatmap input file 은 Gene Symbol 과 fold change (log2) 또는 z-score 로 구성된다(그림 1-23). MeV 프로그램을 사용하여 Clustering heatmap 을 작성하는 방법은 MeV manual (Download Link)에서 확인할 수 있다.

Heatmap(Allcompare_FC).txt - 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말

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	-0.716801237557984
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

KEGG input 은 분석 결과에서 Up-/Down-regulated 된 유전자들이 어떤 Pathway 에 속하는지 확인하고자 할 때 **KEGG Mapper** 를 이용하기 위한 입력 데이터를 제공한다. KEGG Input 에서는 하나의 비교 조합만 선택 가능하다. 비교 조합 선택 후 Data Export 를 선택하면 그림 1-24 과 같이 유전자의 Entrez ID 와 비교 조합, 발현 수준에 따른 색 코드로 구성된다. KEGG Input 파일은 **KEGG Mapper** 의 입력 데이터로 사용하여 Pathway 상에 속하는 유전자와 이들의 발현 수준을 확인할 수 있다.

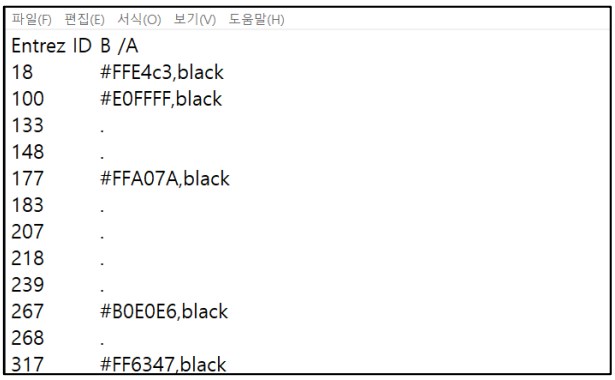


그림 1-24. Third Party Support - KEGG input

DAVID 는 다양한 데이터베이스를 기반으로 유전자의 상관관계를 통계적으로 분석하여 유전자의 주요 기능을 예측하는 Analysis tool 이다. DAVID 는 3 천 개 이상의 유전자는 분석할 수 없으므로 3 천 개 이하로 유전자를 선별해야 한다. Data Export (DAVID Input)는 DAVID 에서 유전자 이름을 입력하는 부분에 사용되는 파일이다. 이 파일은 유전자 이름으로 구성된다(그림 1-25). 유전자 이름을 DAVID 의 입력 데이터로 사용하여 분석하려는 GO 분석 결과 데이터를 다운로드 받는다.

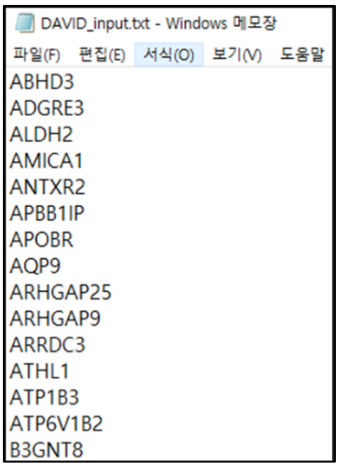


그림 1-25. Third Party Support - DAVID input

1-5. Selected Gene Plot 사용 방법

ExDEGA 의 기능 중에 선별한 유전자 또는 연구자가 관심있는 유전자들을 대상으로 발현 패턴을 그래프로 표현하고자 할 때는 "Selected Gene Plot" 기능을 사용할 수 있다. 선별한 유전자의 ID 을 복사하여 Selected Gene Plot 창에 붙여 넣고 "Expression Plot View"를 누르면 Normalized data (log2) 값, Fold change (log2) 값으로 line graph 가 그려진다(그림 1-26).

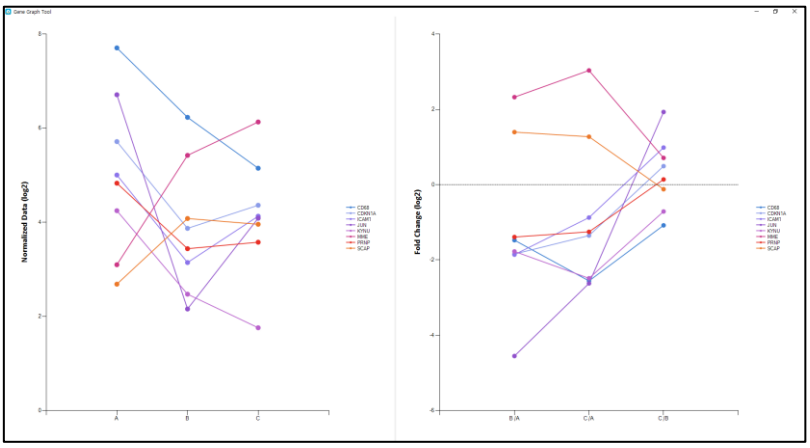


그림 1-26. Gene graph

1-6. Rader Chart 사용 방법

ExDEGA 의 기능 중에 선별한 유전자나 연구자가 관심있는 유전자들을 대상으로 발현 패턴을 그래프로 표현하고자 할 때는 "Radar chart" 형식의 이미지로도 표현할 수 있다. 사용방법은 선별한 유전자의 ID 을 복사하여 Radar chart 창에 붙여 넣고 옵션을 선택한다. 옵션은 Radar chart 의 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 1-27). 단, 꼭지점에 있는 항목이 3 가지 이상일 경우 제작 가능하다.

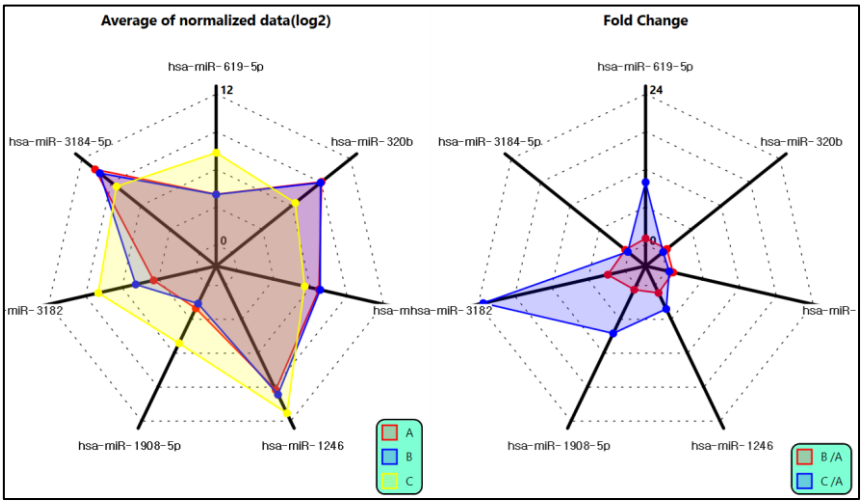


그림 1-27. Radar chart

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 Sheet 에 'insulin' keyword 을 포함하는 행만 필터링 하여 확인할 수 있다(그림 1-28).

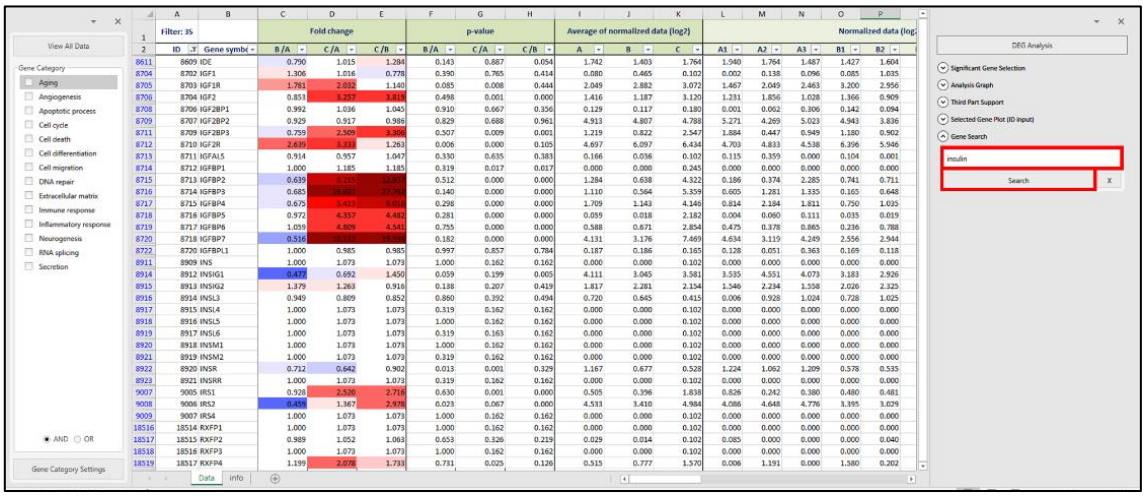


그림 1-28. Genes related to insulin

2. Functional Annotation Analysis (DAVID, ExDEGA GraphicPlus)

2-1. DAVID 분석 툴을 이용한 Functional Annotation 분석

DAVID 는 다양한 데이터 베이스를 기반으로 유전자의 상관관계를 통계적으로 분석하여 유전자의 주요 기능을 예측하는 analysis tool 이다. 분석 과정은 그림 2-1 과 같다.

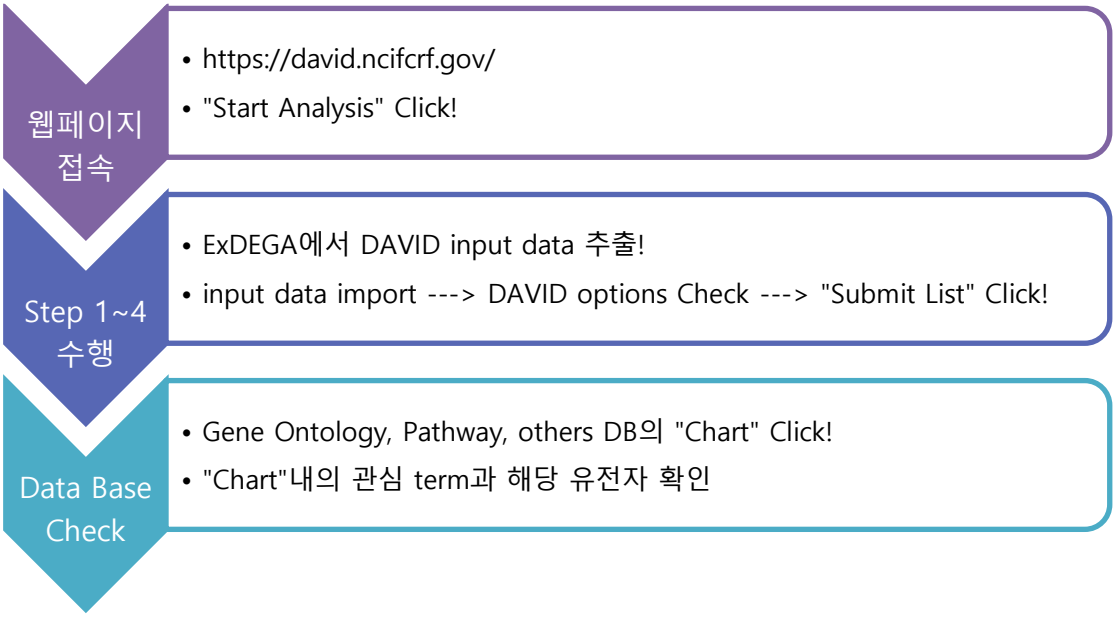


그림 2-1. DAVID tool analysis process

DAVID에서는 3천 개 이상의 유전자는 분석할 수 없으므로 3천 개 이하로 유전자를 선별해야 한다. RNA-Seq 결과에서 significant gene을 선별하여 DAVID 분석을 한다. 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 DAVID input을 추출하여, DAVID 분석에 사용한다. 반드시 하나의 비교 조합만 진행해야 한다(그림 2-2).

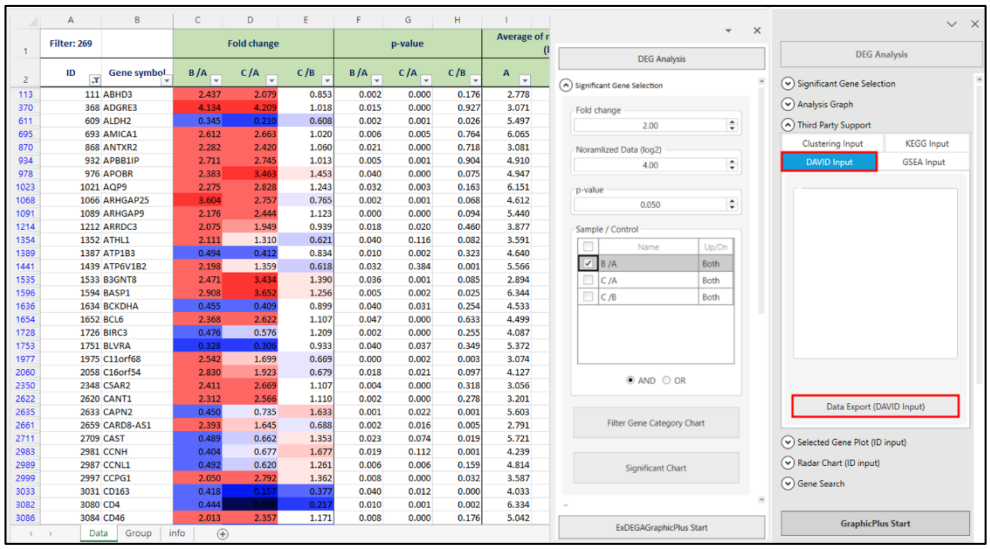


그림 2-2. DAVID input file generation process

DAVID 홈페이지 (https://david.ncifcrf.gov/home.jsp)에 접속하여 "Start Analysis"을 클릭한다(그림 2-3).

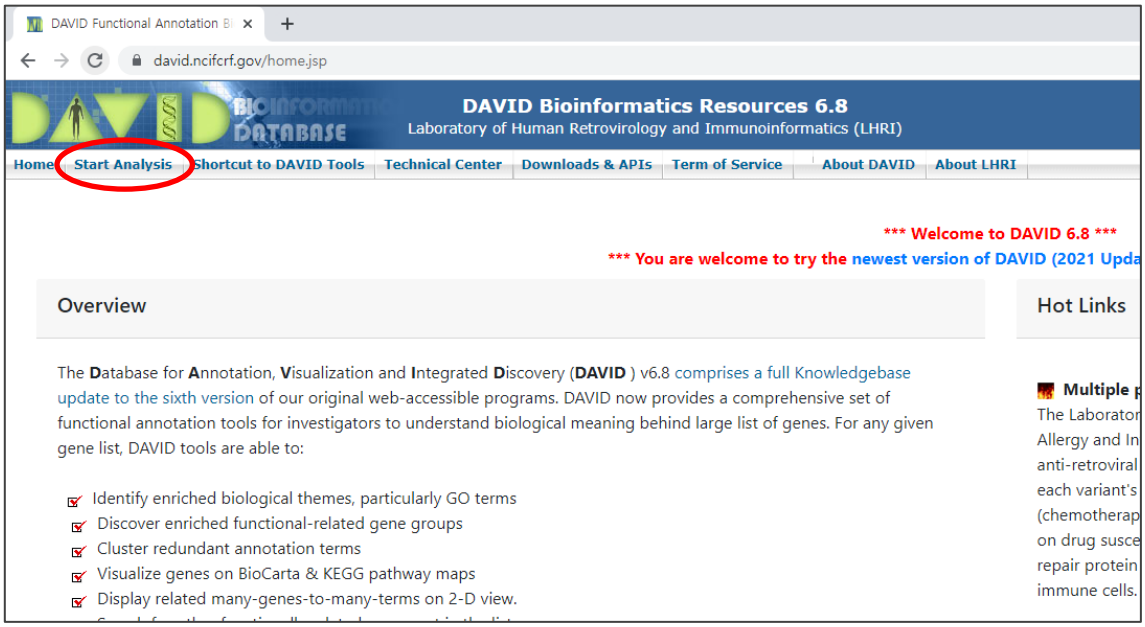


그림 2-3. DAVID tool webpage

"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.

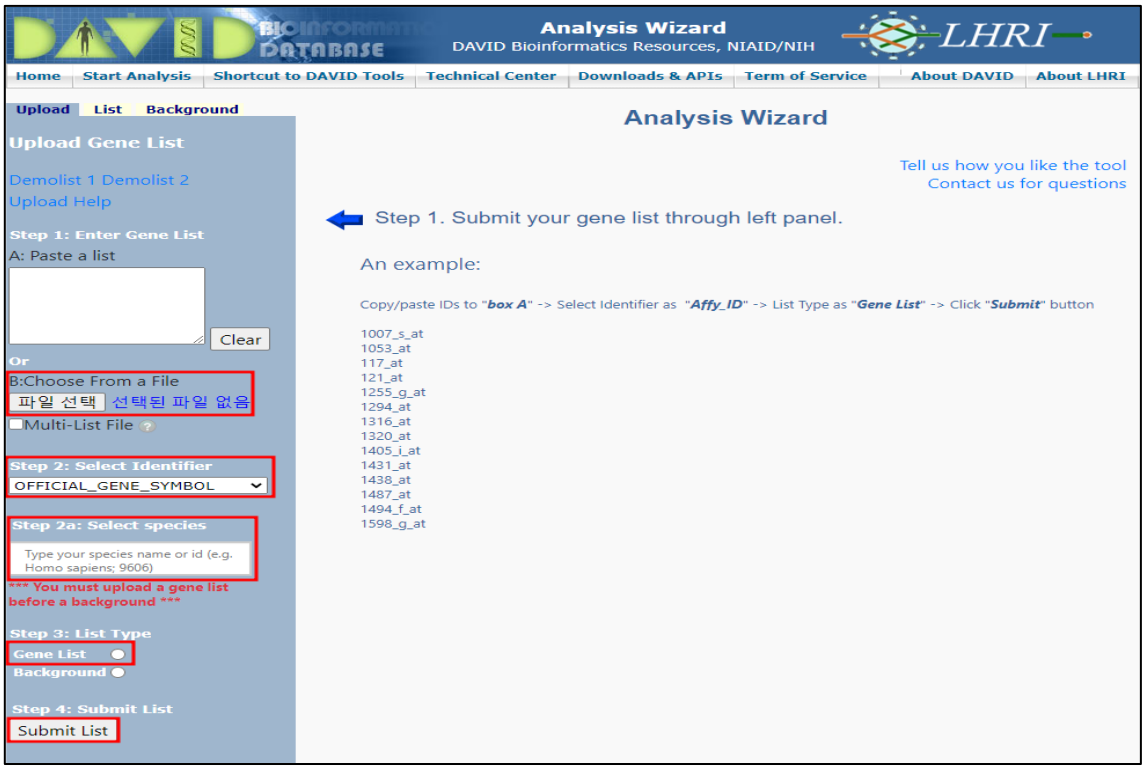
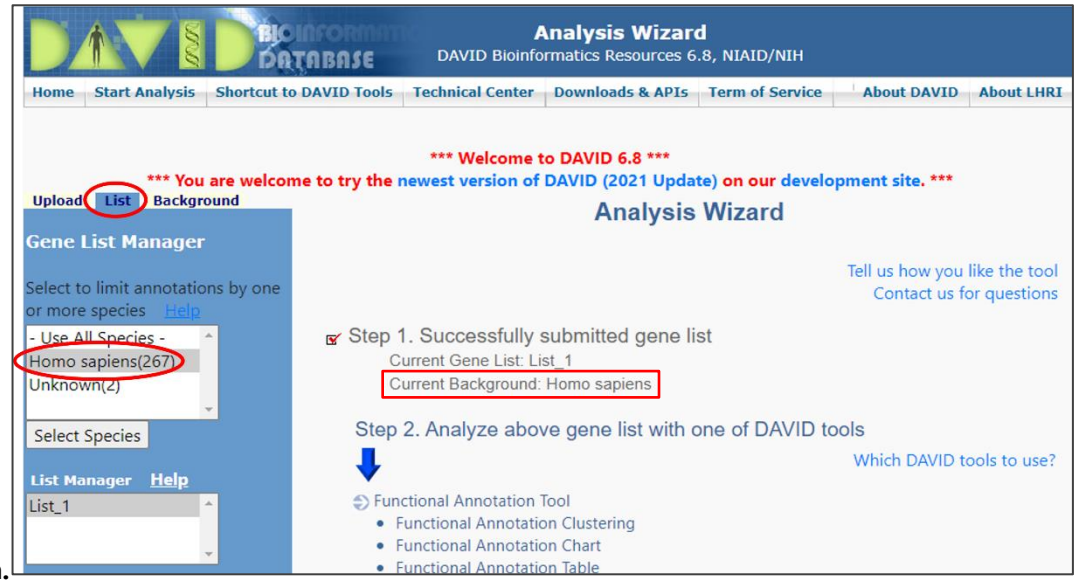


그림 2-4. DAVID tool : Step 1 ~ Step 4

List sheet 에서 분석하고자 하는 종을 선택한다(그림 2-5. a). "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 2-5. b).

a.



b.

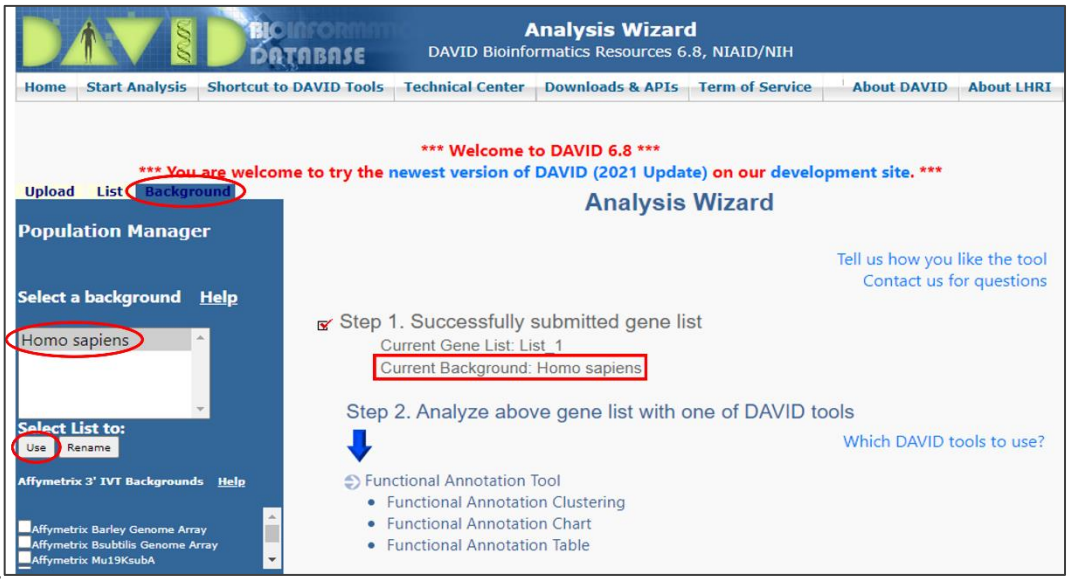


그림 2-5. DAVID tool : Select Species

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

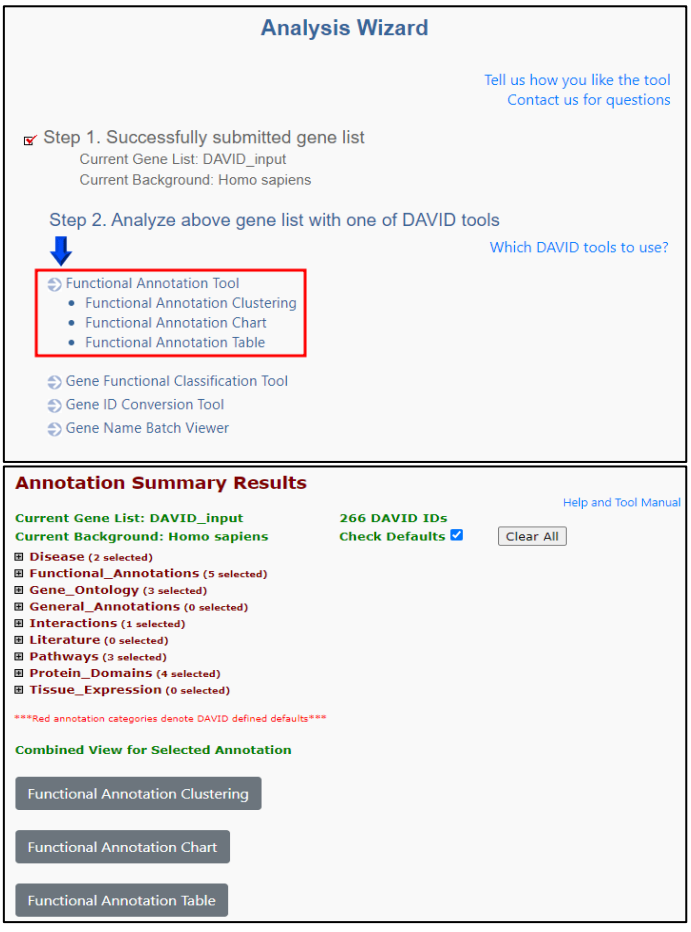


그림 2-6. DAVID results

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.

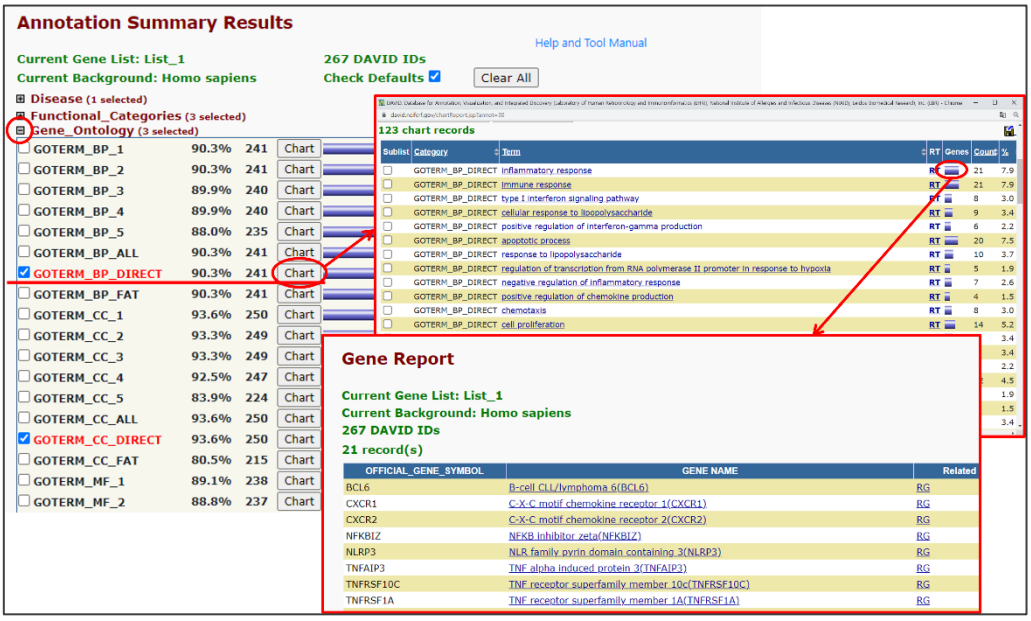


그림 2-7. DAVID tool : exploring Gene Ontology analysis result

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.

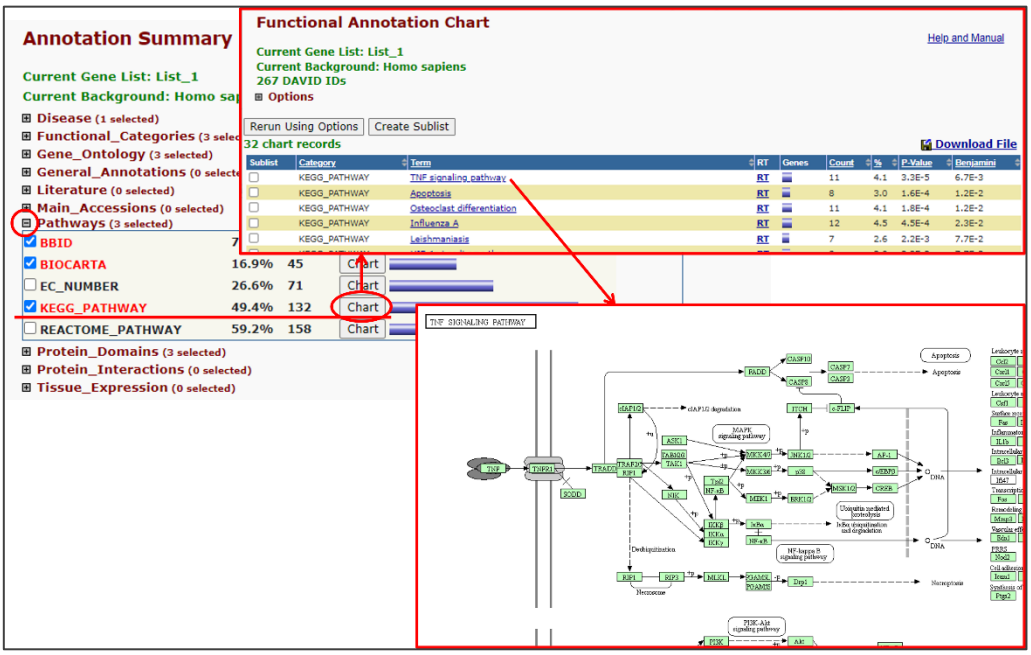


그림 2-8. DAVID tool : exploring Pathway analysis result

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.

DAVID 분석 결과를 내 컴퓨터에 저장하려면, chart 를 클릭해서 나온 결과창에서 Download File 링크를 마우스 오른쪽 버튼을 클릭한 후 다른 이름으로 저장을 선택하면 DAVID reults 파일을 다운로드 받을 수 있다 (그림 2-9. 다운로드 받은 DAVID 결과 파일로 그래프 작성하는 방법은 '**2-2. ExDEGA GraphicPlus 를 이용한 DAVID 결과 그래프 작성**'에 설명되어 있다.

* 주의사항

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

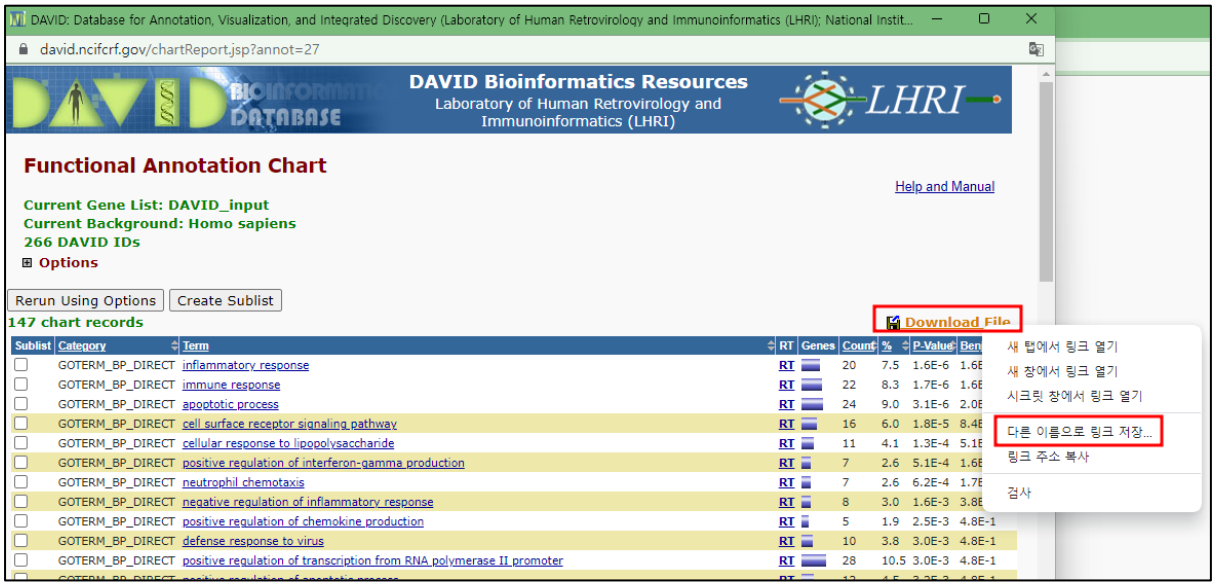


그림 2-9. DAVID data download

DAVID 에서는 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 2-10)

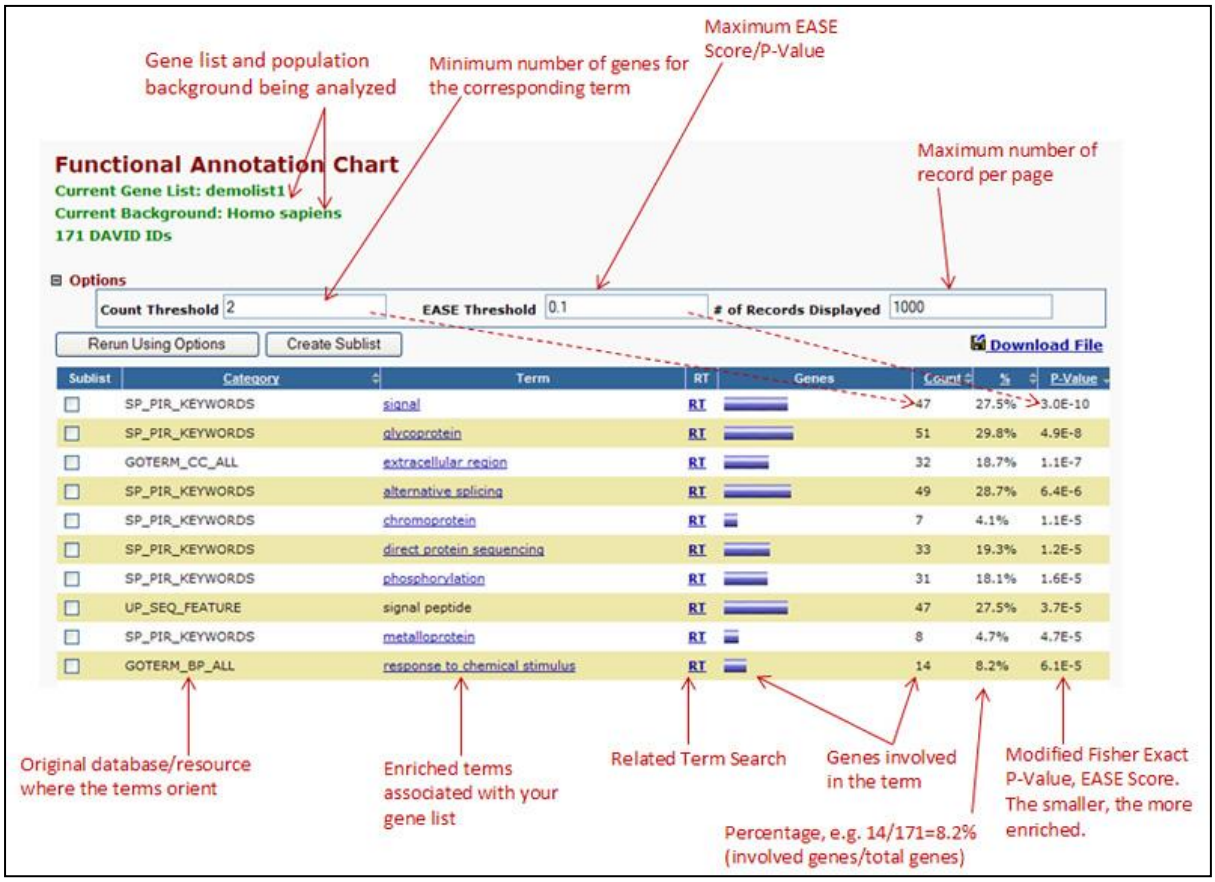


그림 2-10. DAVID Help and Tool Manual

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

ExDEGA 레포트의 ExDEGA Graphic Plus Start를 클릭하여 프로그램을 실행한다 (그림2-11).

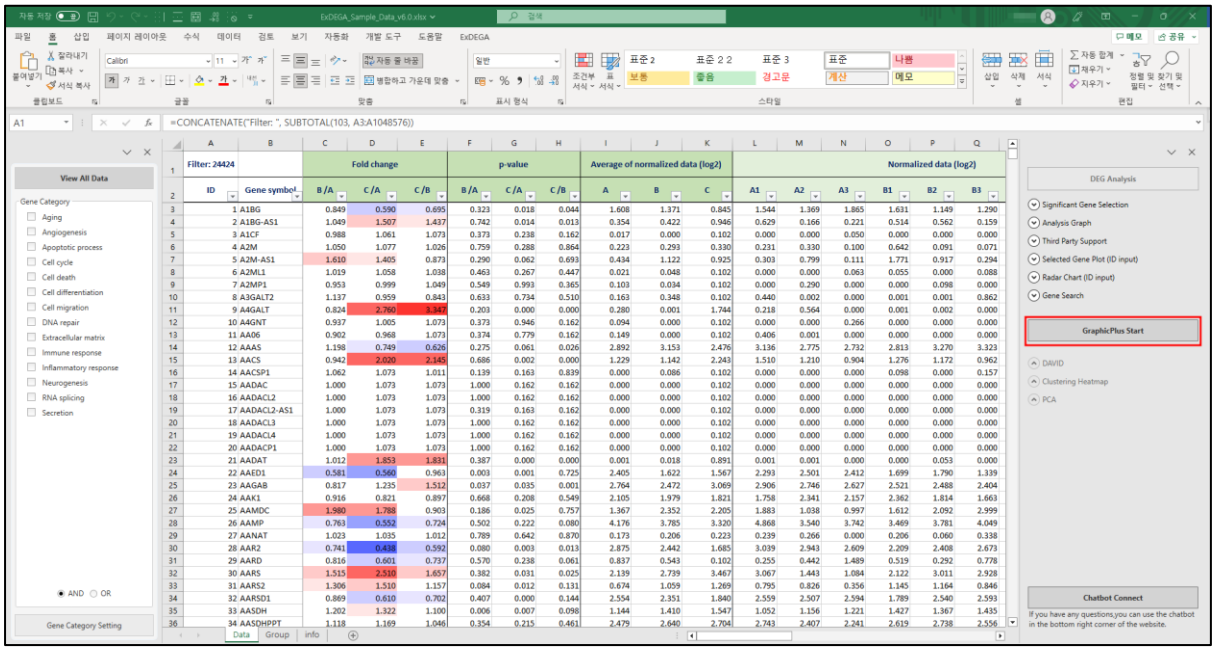


그림 2-11. Execute ExDEGA GraphicPlus

메인 화면의 4 개 탭 중 'DAVID' 탭에서 DAVID Graphic 분석을 수행할 수 있다. DAVID Graphic 분석 창은 그림 2-12 과 같다.

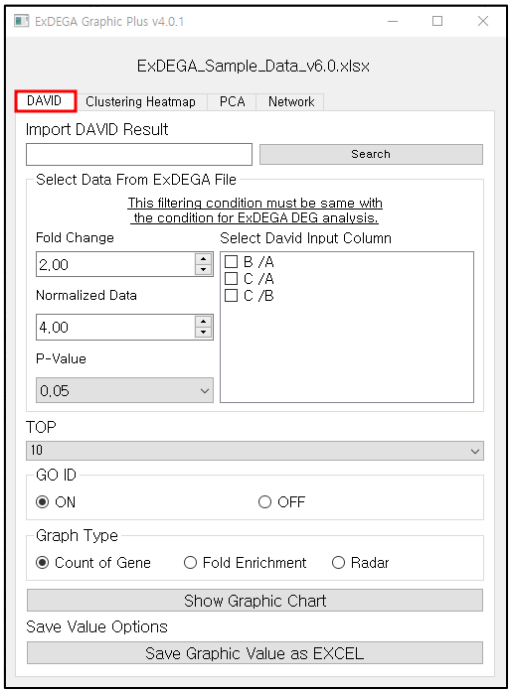


그림 2-12. Select DAVID tab to execute the DAVID Graphic Analysis

DAVID 분석 결과를 그래프로 제작하려면 DAVID result 파일과 DAVID input 이 필요하다. DAVID result 은 '2-1. DAVID 분석 툴을 이용한 Functional Annotation 분석'에서 저장한 DAVID 결과 파일이다. DAVID result 파일 안에 있는 내용은 그림 2-13 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	Bonferron	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, Q9	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 2-13. DAVID output file

그래프를 제작하기 위해서는 이전 DAVID input 파일을 만들 시, 적용했던 선별조건을 그대로 설정해야 한다 (그림 2-14).

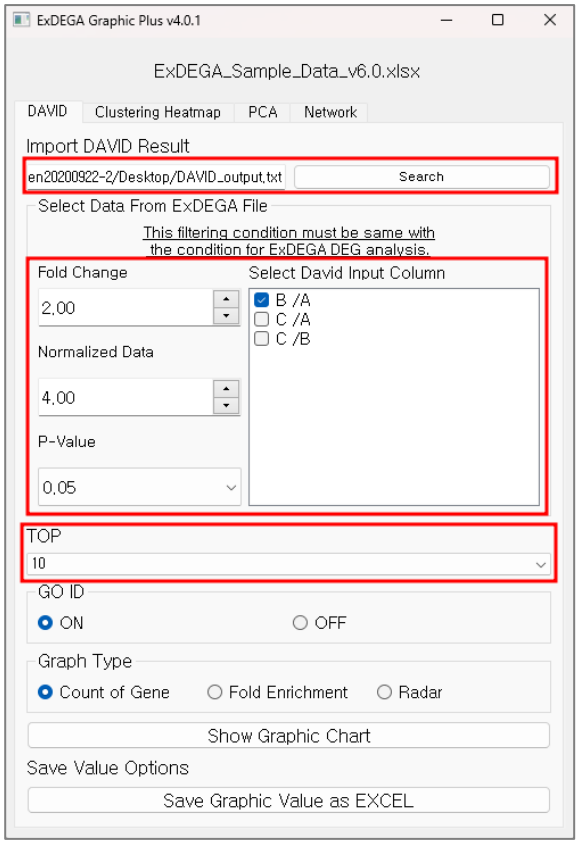


그림 2-14. Steps for Create DAVID Analysis Graph

우선, import DAVID Result 에서 DAVID result 파일을 선택한다. 그리고 아래 선별조건에서 DAVID 분석 시 사용된 input 파일과 동일하도록 ExDEGA 레포트에 적용한 Fold Change 값, Normalized Data (log2)값, P-Value 값 및 비교조합 선택을 세팅해야 한다. 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. TOP 에서는 그래프 제작에 사용될

결과 리스트의 수이다. 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 35 개 까지의 옵션을 선택 가능하며, 사용자가 별도의 기준을 적용하지 않을 경우 이 값은 자동으로 10 으로 설정된다.

위의 과정을 거친 뒤, Graph Type 에서 3 종류의 그래프를 제작할 수 있다 (그림 2-15).

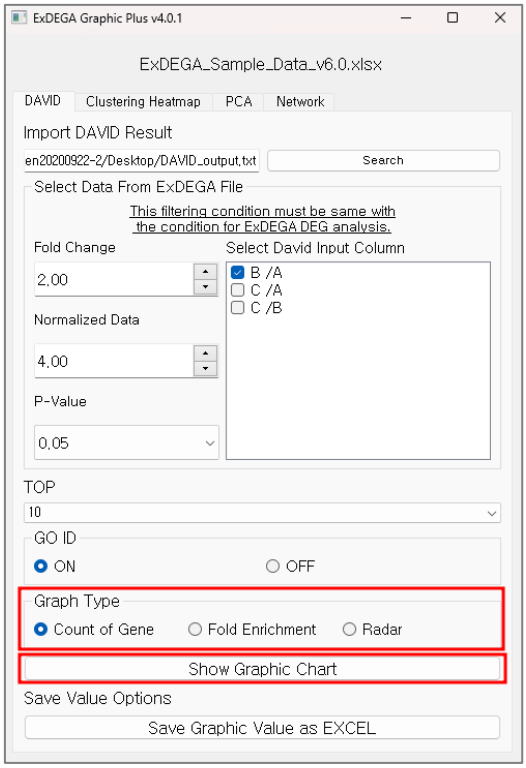


그림 2-15. Result of Graphic Chart

Count of Gene 을 선택하고 Show Graphic Chart 를 누르면 각 GO (or pathway)에서 발현이 증가하는 유전자, 감소하는 유전자 수가 그래프로 작성된다(그림 2-16).

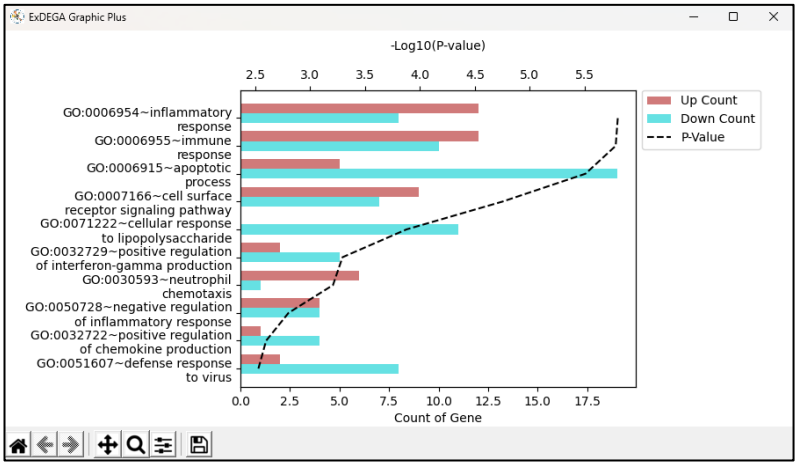


그림 2-16. Result of Graphic Chart (Count of Gene)

Fold Enrichment를 선택하고 Show Graphic Chart를 누르면 각 GO (or pathway)의 p-value, Fold enrichment 값으로 그래프가 작성된다 (그림2-17).

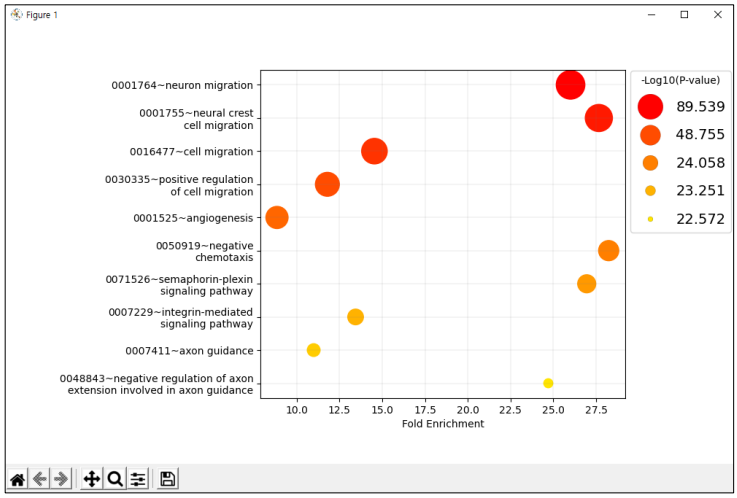


그림 2-17. Result of Graphic Chart (Fold Enrichment)

Radar를 선택하고 Show Graphic Chart를 누르면 각 GO (or pathway)의 Fold enrichment 값으로 그래프가 작성된다 (그림2-18).

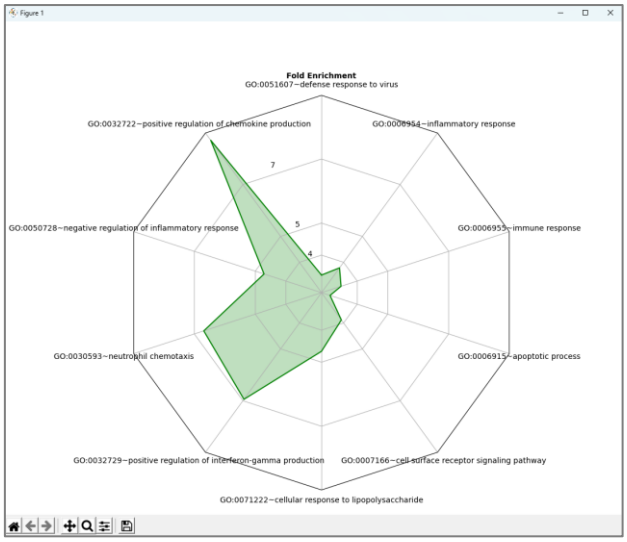


그림 2-18. Result of Graphic Chart (Radar)

GO ID에서 ON을 선택하면 DAVID의 고유번호가 앞에 붙고, OFF를 선택하면 고유번호 없이 Term만 출력된다 (그림 2-19).

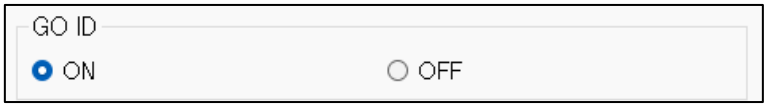


그림 2-19. Result of Graphic Chart (GO ID)

마지막으로, Save Graphic Value를 누르면 그래프에 사용된 값들이 Excel형식으로 저장된다 (그림 2-20).

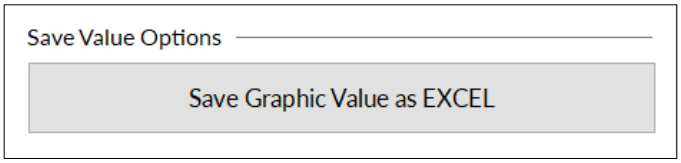


그림 2-20. Save DAVID Graphic Values as Excel

이 결과 파일은 DAVID Result (그림 2-21), Gene List & Fold Change (그림 2-22)의 2개 탭으로 이루어져 있다. DAVID Result 탭에서는 그래프 제작에 이용한 DAVID Graphic 분석에 대한 전체적인 결과값이, Gene List & Fold Change 탭에서는 설정한 Top Count에서 유의미한 결과값을 보인 Gene List와 각 gene의 fold change 결과값이 표시된다.

	Total	Number of Molecules	Up Molecules	Down Molecules	Fold Enrichment	Log10(P-value)
0001764~neuron migration	84	70	31	39	26.00451467268623	89.53874064847625
0001755~neural crest cell migration	44	39	20	19	27.659347424584445	50.73585351965816
0016477~cell migration	118	55	29	26	14.54489803726518	48.75523653693835
0030335~positive regulation of cell migration	127	48	24	24	11.794173583832496	37.14452140747271
0001525~angiogenesis	134	38	20	18	8.849297530406657	24.057940473511444
0050919~negative chemotaxis	21	19	8	11	28.23347307320219	23.97814634517612
0071526~semaphorin-plexin signaling pathway	22	19	11	8	26.950133388805664	23.251333489938407
0007229~integrin-mediated signaling pathway	65	28	11	17	13.442333738496266	23.055541647085267
0007411~axon guidance	88	31	16	15	10.992817566180998	22.571872514053652
0048843~negative regulation of axon extension involved in axon guidance	24	19	11	8	24.70428893905192	22.01361832716083

David Result | Gene List & Fold Change

그림 2-21. Result of Saved Graphic value (Tab 1 : DAVID Result)

	0001764~neuron mig	0001764_FC	neural crest cell	0001755_FC	477~cell migra	0016477_FC	ive regulation o	0030335_FC	1525~angiogen	0001525_FC	9~negative che	0050919_FC	phorin-plexin si	0071526_FC	rin-mediated sig	0007229_FC	411~axon guid	0007411_FC	of axon extens	0048843_FC
2	0	MEF2C	0.859782263	NRP2	0.96702913	EPHA1	1.083013992	CCL3	0.142713802	NRP2	0.96702913	NRP2	0.96702913	FLNA	0.650619717	ITGB4	1.201288745	WNT5A	1.014431086	WNT5A	1.014431086
3	1	TUBB2B	0.940041104	EDN3	0.955311516	CSPG4	1.005792733	DRD1	1.022817663	FGFR1	0.81416452	SLIT1	1.027894843	SEMA5A	0.975121774	ITGB5	0.913818968	CDH5R1	0.825573826	NRP1	0.846690016
4	2	PAX6	0.950597863	NRTN	1.029285219	SDC4	0.521640541	PDGFB	1.352585508	CAV1	1.003423852	SLIT3	0.996503173	SEMA6A	1.098783042	ITGB2	0.738046765	EPNA1	1.083013992	SEMA5A	0.975121774
5	3	PEX5	0.991673245	ERBB4	1.008209316	RAB1A	0.767316638	F2RL1	2.677534126	ACVRL1	0.835068183	APOA1	0.977553588	SEMA6C	1.178436305	ITGB3	0.630547126	PAX6	0.950597863	SEMA6A	1.098783042
6	4	GJA1	0.968055913	KITLG	0.99093864	SDC2	1.143314051	EDN1	0.96739145	EFNA1	1.083013992	SEMA4G	1.145906979	SEMA6D	1.014010591	ITGB1	0.689743825	CXCL12	0.864461	SEMA6C	1.178436305
7	5	FGF13	1.045138752	GDNF	1.006143849	TGFB1	0.641669032	FER	1.118601907	WASF2	1.150090002	SEMA4F	1.149019447	SEMA4G	1.145906979	ITGAM	0.95945261	GLI3	0.935657218	SEMA6D	1.014010591
8	6	PRKG1	0.926776102	SOX8	0.922659924	TGFB2	0.957342529	CXCL12	0.864461	EPHB3	1.105765553	SEMA3G	1.015347482	SEMA4F	1.149019447	SRC	0.430301835	SHH	0.997912483	SEMA4G	1.145906979
9	7	CTNNB1	0.872765289	SHH	0.997912483	CUL3	1.303301053	TGFB1	0.641669032	PTEN	1.564172914	ITGAV	0.88173207	SEMA3G	1.015347482	CUL3	1.303301053	TGFB2	0.957342529	SEMA4F	1.149019447
10	8	PEX7	0.743490745	SEMA5A	0.975121774	FAM83D	0.980343251	SEMA5A	0.975121774	PRKX	0.677769185	SEMA3F	0.98296323	SEMA3F	0.98296323	PTK2	0.986780604	LAMB2	0.868051993	SEMA3G	1.015347482
11	9	NDE1	1.536416938	EDNRB	0.933948792	PDPK1	1.07454371	PTK2	0.986780604	EPHB1	1.799663601	SEMA3E	0.9810739	RAC1	1.115886798	APOA1	0.977553588	ROBO1	1.105908686	SEMA3F	0.98296323
12	10	POMGNT2	0.702009099	SEMA3G	1.015347482	NDE1	1.536416938	SEMA3G	1.015347482	NRCAM	1.218190273	SEMA3D	1.027867246	SEMA3E	0.9810739	PTK2B	1.222727732	GATA3	0.839542686	SEMA3E	0.9810739
13	11	GATA3	0.839542686	SEMA7A	0.74064196	S1PR1	1.207926194	SEMA7A	0.74064196	PTK2	0.986780604	SEMA4C	0.910711274	SEMA3D	1.027867246	ITGB8	1.060729002	RAC1	1.115886798	SEMA3D	1.027867246
14	12	RAPGEF2	1.013904084	SEMA3F	0.98296323	TIAM1	0.958857526	SEMA3F	0.98296323	S1PR1	1.207926194	RHOA	0.806837702	SEMA4C	0.910711274	SEMA7A	0.74064196	GBX2	1.01494718	SEMA4C	0.910711274
15	13	NR2F2	0.99123908	SEMA3E	0.9810739	RHOA	0.806837702	ITGAV	0.88173207	ANG	0.593802654	SEMA4B	2.409734569	SEMA4B	2.409734569	ITGAV	0.88173207	SEMA3C	0.95972267	SEMA4B	2.409734569
16	14	ASPM	1.160999121	GBX2	1.01494718	PTPRK	1.20316042	SEMA3E	0.9810739	PTK2B	1.222727732	SEMA3C	0.95972267	SEMA3C	0.95972267	ITGB7	0.951101264	NKX2-1	1.023241608	SEMA3C	0.95972267
17	15	PITX2	1.033092698	SEMA3D	1.027867246	ARC	0.913979888	SEMA3D	1.027867246	ITGAV	0.88173207	SEMA3B	0.960108825	SEMA3B	0.960108825	ILK	0.944779977	USP33	1.073246406	SEMA3B	0.960108825
18	16	KIRREL3	1.029189125	SEMA3C	0.95972267	CCDC88A	0.678932544	SEMA3C	0.95972267	TEK	1.060102409	SEMA4D	1.569044066	SEMA4D	1.569044066	FCER1G	0.844611124	FLRT3	0.999499242	SEMA4D	1.569044066
19	17	TWIST1	0.942645058	SEMA3B	0.960108825	PTPRF	1.070568168	SEMA3B	0.960108825	PIK3CA	0.91191364	SEMA3A	0.926246444	SEMA3A	0.926246444	SYK	1.273158465	MATN2	0.980952064	SEMA3A	0.926246444
20	18	MATN2	0.980952064	SEMA3A	0.926246444	LYN	1.335519894	LAMB1	1.036702746	RHOB	0.466743059	SEMA4A	1.205043398	SEMA4A	1.205043398	ITGA1	1.068664471	EFNB1	1.090589862	SEMA4A	1.205043398
21	19	PHOX2B	1.008467428	SOX10	1.099415304	EFNB2	1.028375988	PIK3R1	0.307389252	SHC1	0.970948849	_	None	_	None	ITGA2	1.049418578	EFNB2	1.028375988	_	None
22	20	ELP3	1.189387588	RET	0.947689347	MMP14	0.686333138	EGFR	0.997992529	ANGPT1	1.366862918	_	None	_	None	ITGA3	0.9771769	VAX1	0.959406691	_	None
23	21	SATB2	0.955934816	PHACTR4	0.720957676	PALLD	0.855445416	ELP3	1.189387588	SOX17	0.999225997	_	None	_	None	ITGA4	0.638463652	PTPRO	1.256592246	_	None
24	22	VAX1	0.959406691	EFNB1	1.090589862	CDK5	0.971637616	RET	0.947689347	FGF2	1.039398378	_	None	_	None	VAV1	1.574332609	FEZF2	0.959375574	_	None
25	23	PCM1	0.920866647	LEF1	1.483036866	JUP	0.639666974	FLT1	1.130235841	ANGPT2	1.131631889	_	None	_	None	PTPN11	0.708425411	SEMA6A	1.098783042	_	None
26	24	CDK5	0.971637616	ISL1	0.979247508	HES1	0.497540011	LYN	1.335519894	FN1	0.847631869	_	None	_	None	ITGA6	1.082516384	EPHA8	0.984002982	_	None
27	25	DDIT4	0.072592812	CORO1C	1.090326572	NCK2	0.718684245	TGFBR1	1.589468185	TGFBR1	1.589468185	_	None	_	None	ITGA5	1.157167212	SEMA4F	1.149019447	_	None
28	26	ASCL1	1.010453192	SEMA6A	1.098783042	SDC1	1.074769832	PODXL	1.138712563	ACKR3	1.196680026	_	None	_	None	LAMA5	1.173156303	CNTN2	1.028627295	_	None
29	27	NDEL1	1.79520061	SEMA6C	1.178436305	NDEL1	1.79520061	PIK3CD	2.067214458	MYH9	1.043745286	_	None	_	None	ITGB1BP1	0.883151231	RELN	1.007248142	_	None
30	28	PSEN1	1.255145722	HIF1A	0.302931554	SH3KBP1	1.055996222	SUN2	0.881667121	LEP	1.07092241	_	None	_	None	_	None	APBB2	0.920877479	_	None
31	29	CCR4	0.59350018	SEMA4G	1.145906979	LAMC1	1.069478363	LEF1	1.483036866	VEGFC	0.936504698	_	None	_	None	_	None	CHL1	1.015587263	_	None
32	30	RELN	1.007248142	SEMA6D	1.014010591	IL12B	1.003428763	HGF	1.192310908	HIF1A	0.302931554	_	None	_	None	_	None	MYH10	1.448797351	_	None
33	31	CHL1	1.015587263	LAMA5	1.173156303	APC2	1.178523996	MMP14	0.686333138	VEGFA	0.367138658	_	None	_	None	_	None	_	None	_	None
34	32	BBS4	0.937120933	SEMA4F	1.149019447	USP9X	1.160889157	SNAI1	1.160889157	SNAI1	0.253511999	PECAM1	1.806444781	_	None	_	None	_	None	_	None
35	33	CDK5R1	0.825873826	SEMA4C	0.910711274	PEAK1	0.949040201	NOTCH1	2.361637016	AAMP	0.762977827	_	None	_	None	_	None	_	None	_	None
36	34	SDCCAG8	0.861983043	SEMA4B	2.409734569	PTK7	1.096392148	SEMA6C	1.178436305	SETD2	0.816334374	_	None	_	None	_	None	_	None	_	None
37	35	CDK5R2	0.977606945	SEMA4D	1.569044066	CDH2	1.42385724	SYNE2	0.928851683	JAM3	1.013778745	_	None	_	None	_	None	_	None	_	None
38	36	POMK	1.21623053	HTR2B	1.068028765	EPHB3	1.105765553	SEMA4G	1.145906979	WNT7A	0.513240858	_	None	_	None	_	None	_	None	_	None
39	37	NDN	0.963329027	SEMA4A	1.205043398	SRC	0.430301835	SEMA6D	1.014010591	SEMA4A	1.205043398	_	None	_	None	_	None	_	None	_	None
40	38	SOX1	0.972419536	ACVR1	0.804901866	SCRIB	1.114455363	SEMA4F	1.149019447	_	None	_	None	_	None	_	None	_	None	_	None
41	39	USP9X	1.160889157	_	None	DOCK1	0.878882533	SEMA4C	0.910711274	_	None	_	None	_	None	_	None	_	None	_	None
42	40	ASTN1	1.006495908	_	None	HOXA5	0.797884616	PDGFRA	0.962596518	_	None	_	None	_	None	_	None	_	None	_	None
43	41	SPOCK1	0.863511401	_	None	ITGAV	0.88173207	ROR2	1.169865317	_	None	_	None	_	None	_	None	_	None	_	None
44	42	SRF	1.158458563	_	None	PTK6	1.073194674	SEMA4B	2.409734569	_	None	_	None	_	None	_	None	_	None	_	None
45	43	PTK2	0.986780604	_	None	LIMD1	1.007805098	WNT11	0.891487669	_	None	_	None	_	None	_	None	_	None	_	None
46	44	GPM6A	1.152123224	_	None	CD24	1.464045018	SEMA4D	1.569044066	_	None	_	None	_	None	_	None	_	None	_	None
47	45	NKX2-1	1.023241608	_	None	NFATC2	0.79447628	COL11A1	0.857709502	_	None	_	None	_	None	_	None	_	None	_	None
48	46	PEX13	1.005176321	_	None	USP33	1.073246406	SEMA4A	1.205043398	_	None	_	None	_	None	_	None	_	None	_	None
49	47	PAFAH1B1	1.135320335	_	None	APC	1.364696101	ACVR1	0.804901866	_	None	_	None	_	None	_	None	_	None	_	None
50	48	UNK	1.036403012	_	None	FSCN2	1.04709182	_	None	_	None	_	None	_	None	_	None	_	None	_	None

David Result | Gene List & Fold Change

그림 2-22. Result of Saved Graphic value (Tab 2 : Gene List & Fold Change)

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

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 테이터 행렬)과 Dendrogram (계층적 클러스터링)을 합쳐 Hierarchical Clustering Heatmap을 그리는 방법을 설명한다.

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

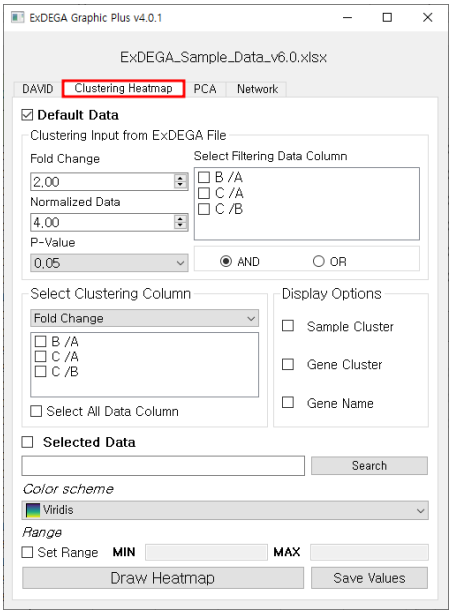


그림 3-1. Select Clustering Heatmap tab

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 0.05 로 자동 적용되며 Fold Change 샘플 그룹은 AND/OR 기능을 통해 여러 개의 샘플에 기준을 복수 적용할 수 있다.

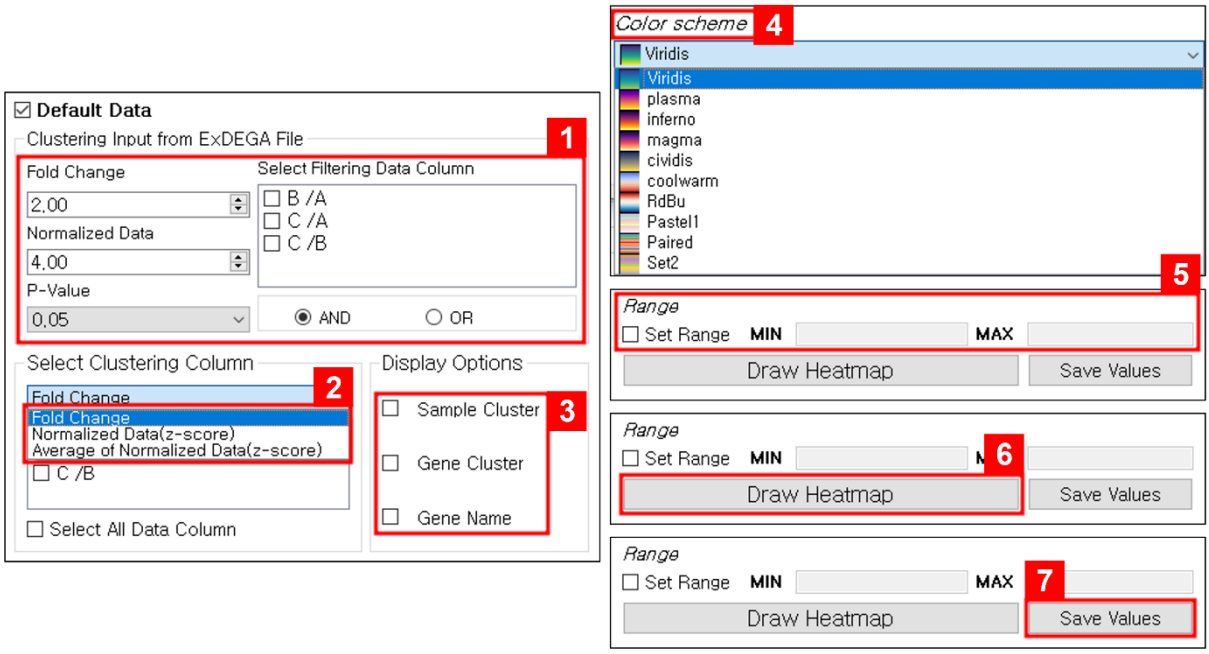


그림 3-2. Steps for Create Clustering Heatmap

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.

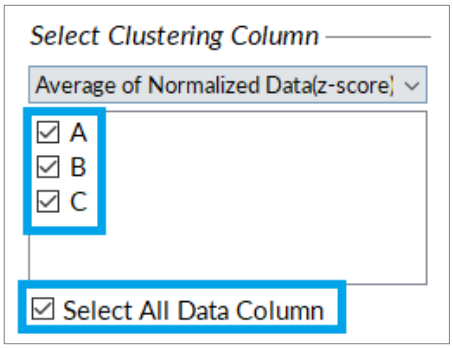


그림 3-3. Select specific sample group for composition of clustering heatmap

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.

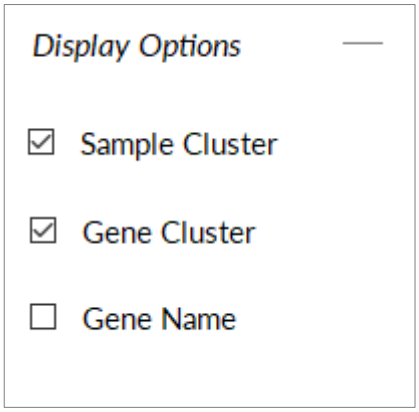


그림 3-4. Set a Display option for clustering heatmap

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).

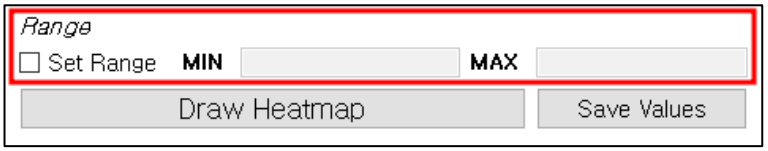
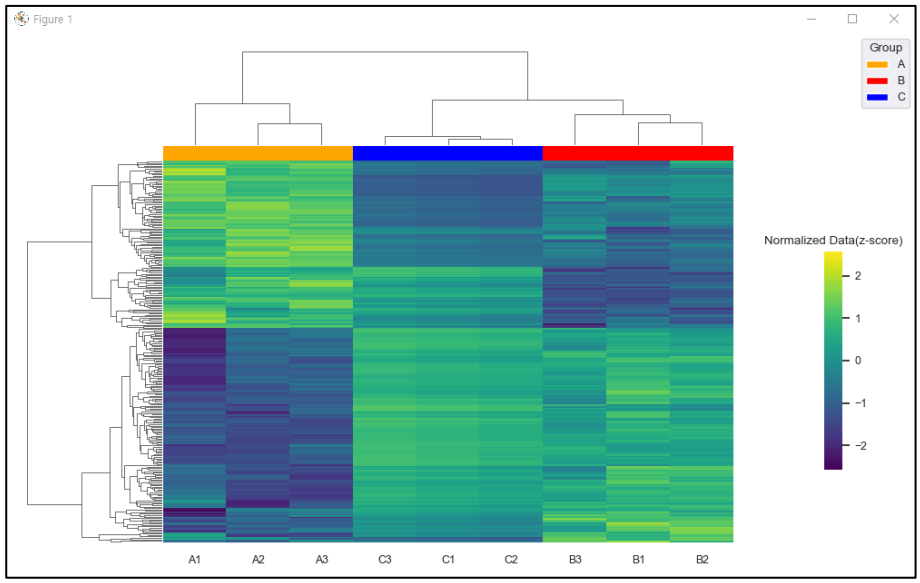


그림 3-5. Set a min-max option for clustering heatmap

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.)

a.



b.

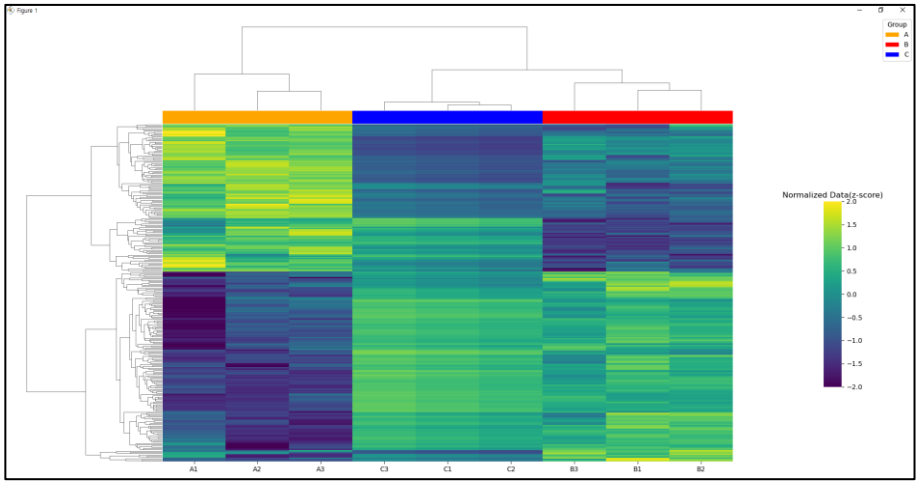


그림 3-6. Clustering heatmap result

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.

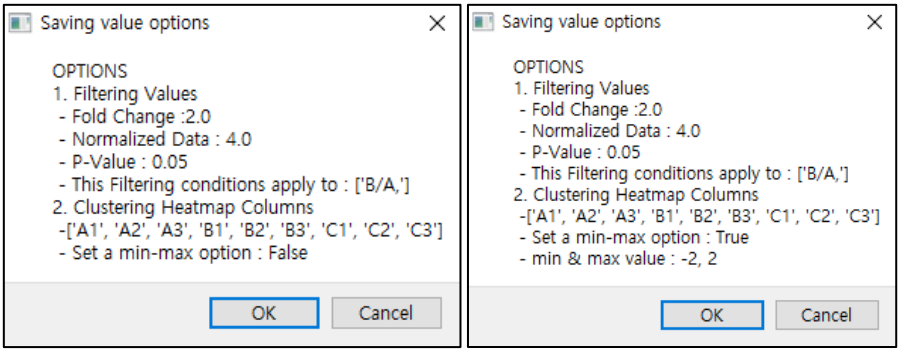


그림 3-7. Confirm all conditions for saving values used to create clustering heatmap

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.)

Save without min-max

	Gene Symbol	B/A	C/A	C/B
0	ADGRE3	2.04741024	2.073376243	0.025966003
1	ALDH2	-1.537049041	-2.253850295	-0.716801254
2	AMICA1	1.385026346	1.413092963	0.028066616
3	ANTXR2	1.190472629	1.27486025	0.084387621
4	APBB1IP	1.438668675	1.456716586	0.018047911
5	APOBR	1.252926787	1.792222798	0.539296011
6	AQP9	1.185878944	1.4995746	0.313695656
7	ARHGAP25	1.849565549	1.463314624	-0.386250925
8	ARHGAP9	1.121902679	1.289058964	0.167156285
9	ATP1B3	-1.017592617	-1.279867725	-0.262275108
10	B3GNT8	1.304915258	1.78003738	0.475122122
11	BASP1	1.540191899	1.86853178	0.328339881
12	BCKDHA	-1.135843533	-1.289436933	-0.1535934
13	BCL6	1.243655421	1.390589701	0.14693428
14	BLVRA	-1.609398178	-1.710154094	-0.100755915
15	C5AR2	1.269735395	1.416122499	0.146387104
16	CANT1	1.208913771	1.359269542	0.150355772
17	CCPG1	1.035597227	1.481048011	0.445450784
18	CD163	-1.259950238	-2.668497509	-1.408547271
19	CD4	-1.172118667	-3.377724913	-2.205606246
20	CD46	1.009085765	1.236811337	0.227725572
21	CD68	-1.475204601	-2.556510409	-1.081305808
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

Save with min-max

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

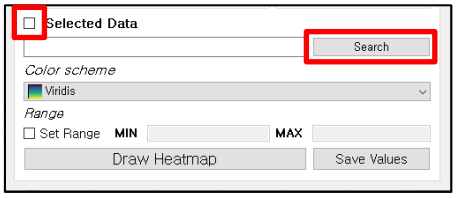


그림 3-9. Selected Data – input upload

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

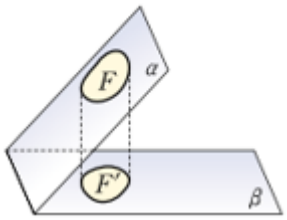
4. Principal component analysis (ExDEGA GraphicPlus)

본 챕터에서는 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 β 위로의 정사영(= F'은 F 를 λ 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 λ 값(eigenvalue)을 계산한다. λ 값은 sample 의 개수만큼 나오게 되고 주성분 비율은 λ 값의 전체 합에서 해당 λ 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.

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

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

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

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(λ)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.

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

$\therefore \boldsymbol{\lambda_1 = 275.101,\ \lambda_2 = -5311.341}$

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 4-2 과 같다.

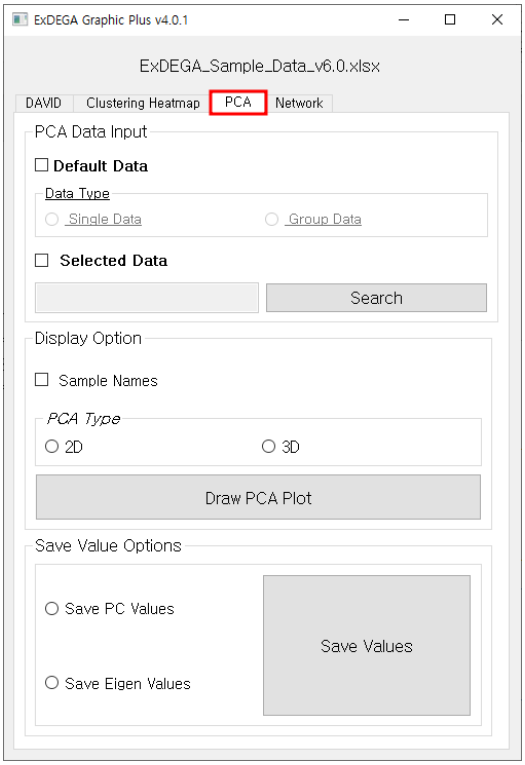


그림 4-2. Select PCA tab

1~3 의 과정을 거쳐 PCA 분석을 수행한다(그림 4-3).

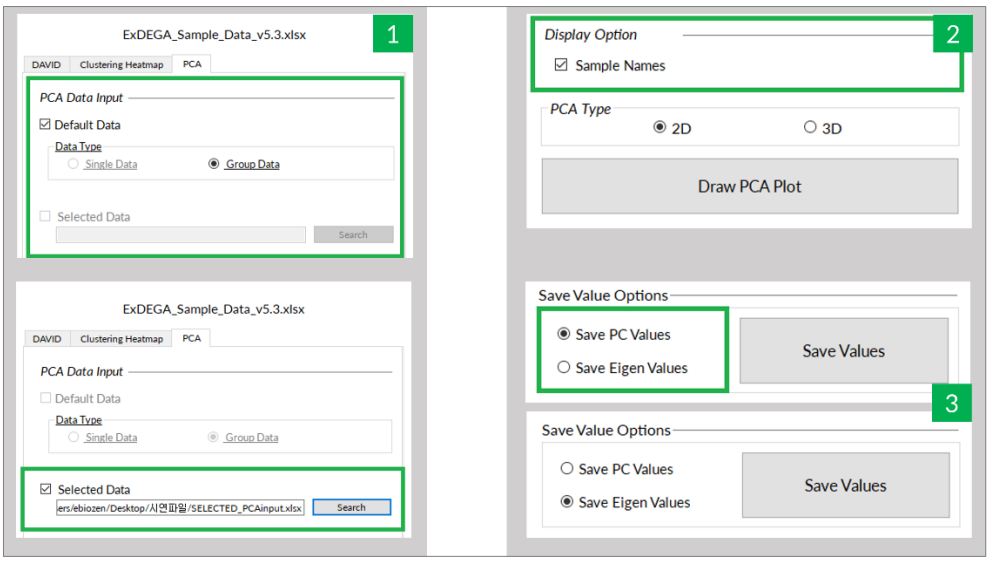


그림 4-3. Steps for Create PCA Graph

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>=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 4-4). 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.

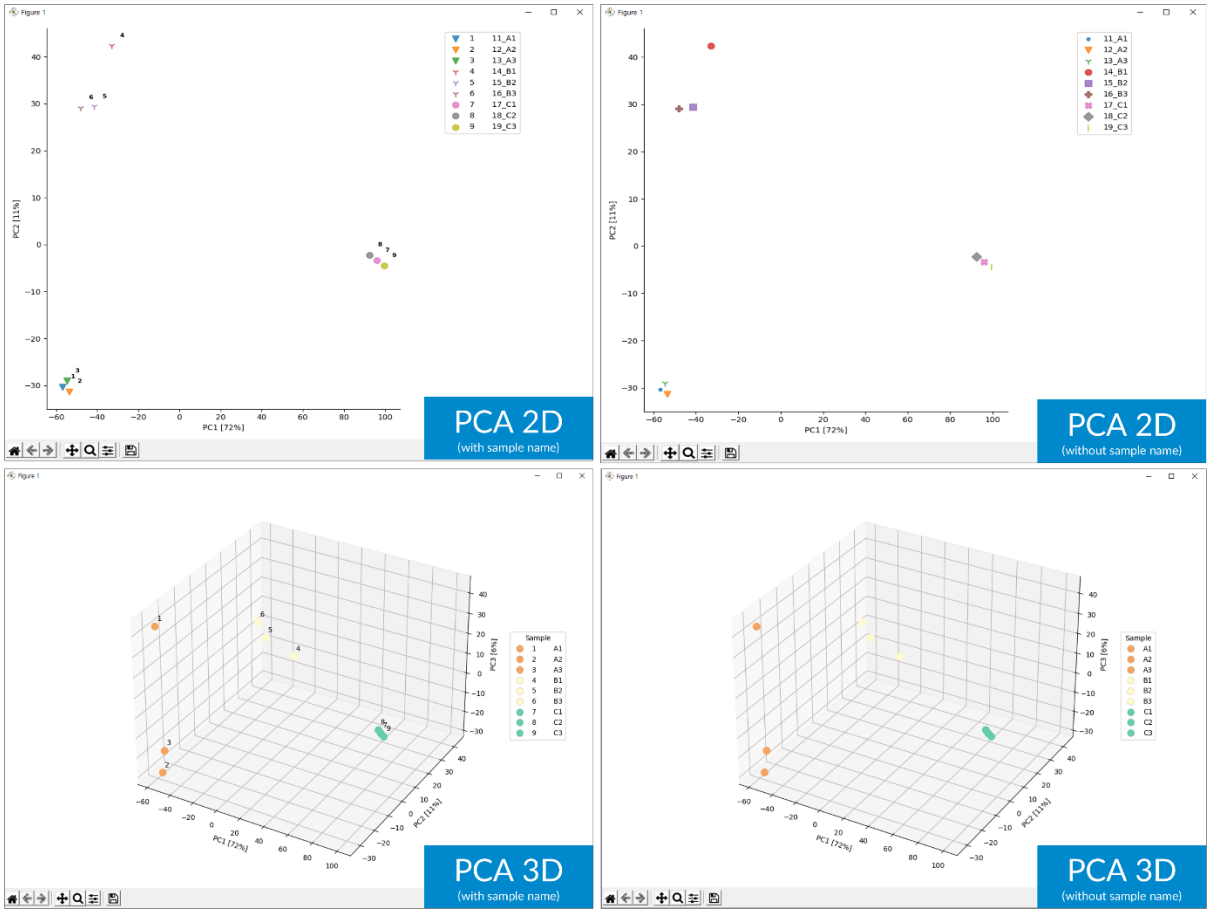


그림 4-4. PCA results 2D, 3D

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 4-5). 이 때, 저장하기 전 PC 값들을 그래프로 먼저 확인할 수 있다(그림 4-6).

	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		

PCA_Values

그림 4-5. Save PC Values as Excel

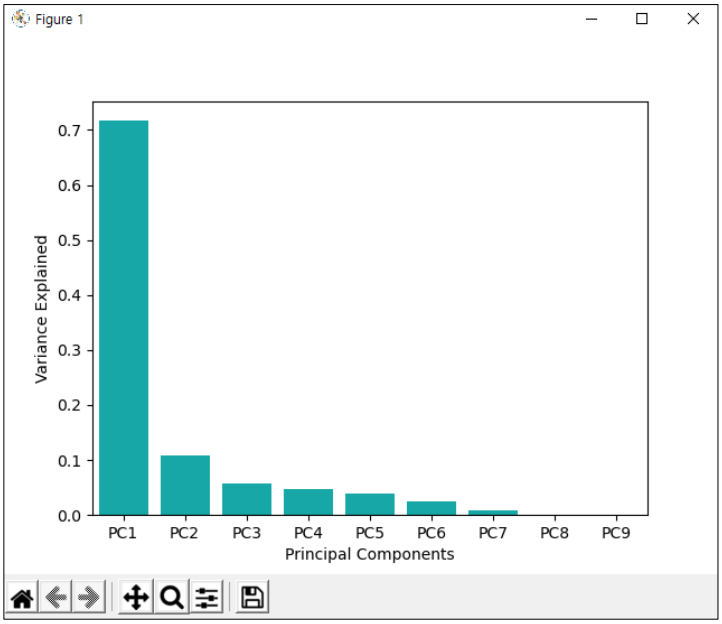


그림 4-6. Check pc values before saving the data as excel

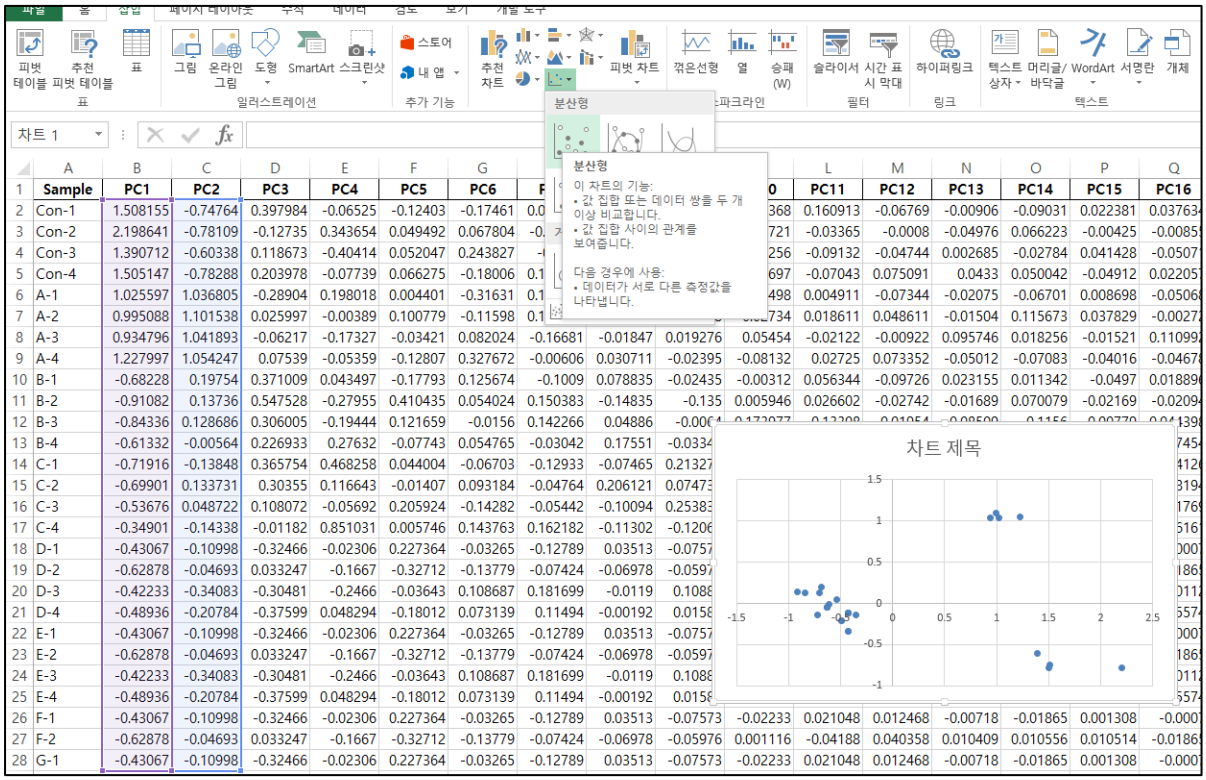
Save Eigen Values 를 체크한 후 Save Values 버튼을 누르면 계산된 Eigenvalue 의 결과를 Excel 파일 형식으로 저장한다(그림 4-7). Eigen values 는 PCA plot 에서 각 샘플의 좌표이다.

	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										

Eigen Values

그림 4-7. Save Eigen Values as Excel

이 값으로 엑셀에서 분산형 차트를 그려 직접 PCA 2D 를 작성할 수도 있다(그림 4-8).



	A	B	C	D	E	F	G	L	M	N	O	P	Q
1	Sample	PC1	PC2	PC3	PC4	PC5	PC6	PC11	PC12	PC13	PC14	PC15	PC16
2	Con-1	1.508155	-0.74764	0.397984	-0.06525	-0.12403	-0.17461	0.160913	-0.06769	-0.00906	-0.09031	0.022381	0.03763
3	Con-2	2.198641	-0.78109	-0.12735	0.343654	0.049492	0.067804	-0.03365	-0.0008	-0.04976	0.066223	-0.00425	-0.0085
4	Con-3	1.390712	-0.60338	0.118673	-0.40414	0.052047	0.243827	-0.09132	-0.04744	0.002685	-0.02784	0.041428	-0.0507
5	Con-4	1.505147	-0.78288	0.203978	-0.07739	0.066275	-0.18006	-0.07043	0.075091	0.0433	0.050042	-0.04912	0.02205
6	A-1	1.025597	1.036805	-0.28904	0.198018	0.004401	-0.31631	0.004911	-0.07344	-0.02075	-0.06701	0.008698	-0.0506
7	A-2	0.995088	1.101538	0.025997	-0.00389	0.100779	-0.11598	0.018611	0.048611	-0.01504	0.115673	0.037829	-0.0027
8	A-3	0.934796	1.041893	-0.06217	-0.17327	-0.03421	0.082024	-0.16681	-0.01847	0.019276	0.05454	-0.02122	-0.00922	0.095746	0.018256	-0.01521	0.11099
9	A-4	1.227997	1.054247	0.07539	-0.05359	-0.12807	0.327672	-0.00606	0.030711	-0.02395	-0.08132	0.02725	0.073352	-0.05012	-0.07083	-0.04016	-0.0467
10	B-1	-0.68228	0.19754	0.371009	0.043497	-0.17793	0.125674	-0.1009	0.078835	-0.02435	-0.00312	0.056344	-0.09726	0.023155	0.011342	-0.0497	0.01889
11	B-2	-0.91082	0.13736	0.547528	-0.27955	0.410435	0.054024	0.150383	-0.14835	-0.135	0.005946	0.026602	-0.02742	-0.01689	0.070079	-0.02169	-0.0209
12	B-3	-0.84336	0.128686	0.306005	-0.19444	0.121659	-0.0156	0.142266	0.04886	-0.0004							
13	B-4	-0.61332	-0.00564	0.226933	0.27632	-0.07743	0.054765	-0.03042	0.17551	-0.0334							745
14	C-1	-0.71916	-0.13848	0.365754	0.468258	0.044004	-0.06703	-0.12933	-0.07465	0.21327							412
15	C-2	-0.69901	0.133731	0.30355	0.116643	-0.01407	0.093184	-0.04764	0.206121	0.07473							319
16	C-3	-0.53676	0.048722	0.108072	-0.05692	0.205924	-0.14282	-0.05442	-0.10094	0.25383							176
17	C-4	-0.34901	-0.14338	-0.01182	0.851031	0.005746	0.143763	0.162182	-0.11302	-0.1206							516
18	D-1	-0.43067	-0.10998	-0.32466	-0.02306	0.227364	-0.03265	-0.12789	0.03513	-0.0757							001
19	D-2	-0.62878	-0.04693	0.033247	-0.1667	-0.32712	-0.13779	-0.07424	-0.06978	-0.0597							186
20	D-3	-0.42233	-0.34083	-0.30481	-0.2466	-0.03643	0.108687	0.181699	-0.0119	0.1088							011
21	D-4	-0.48936	-0.20784	-0.37599	0.048294	-0.18012	0.073139	0.11494	-0.00192	0.0158							557
22	E-1	-0.43067	-0.10998	-0.32466	-0.02306	0.227364	-0.03265	-0.12789	0.03513	-0.0757							001
23	E-2	-0.62878	-0.04693	0.033247	-0.1667	-0.32712	-0.13779	-0.07424	-0.06978	-0.0597							186
24	E-3	-0.42233	-0.34083	-0.30481	-0.2466	-0.03643	0.108687	0.181699	-0.0119	0.1088							011
25	E-4	-0.48936	-0.20784	-0.37599	0.048294	-0.18012	0.073139	0.11494	-0.00192	0.0158							557
26	F-1	-0.43067	-0.10998	-0.32466	-0.02306	0.227364	-0.03265	-0.12789	0.03513	-0.07573	-0.02233	0.021048	0.012468	-0.00718	-0.01865	0.001308	-0.000
27	F-2	-0.62878	-0.04693	0.033247	-0.1667	-0.32712	-0.13779	-0.07424	-0.06978	-0.05976	0.001116	-0.04188	0.040358	0.010409	0.010556	0.010514	-0.0186
28	G-1	-0.43067	-0.10998	-0.32466	-0.02306	0.227364	-0.03265	-0.12789	0.03513	-0.07573	-0.02233	0.021048	0.012468	-0.00718	-0.01865	0.001308	-0.000

그림 4-8. 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 4-9. Copy normalized data

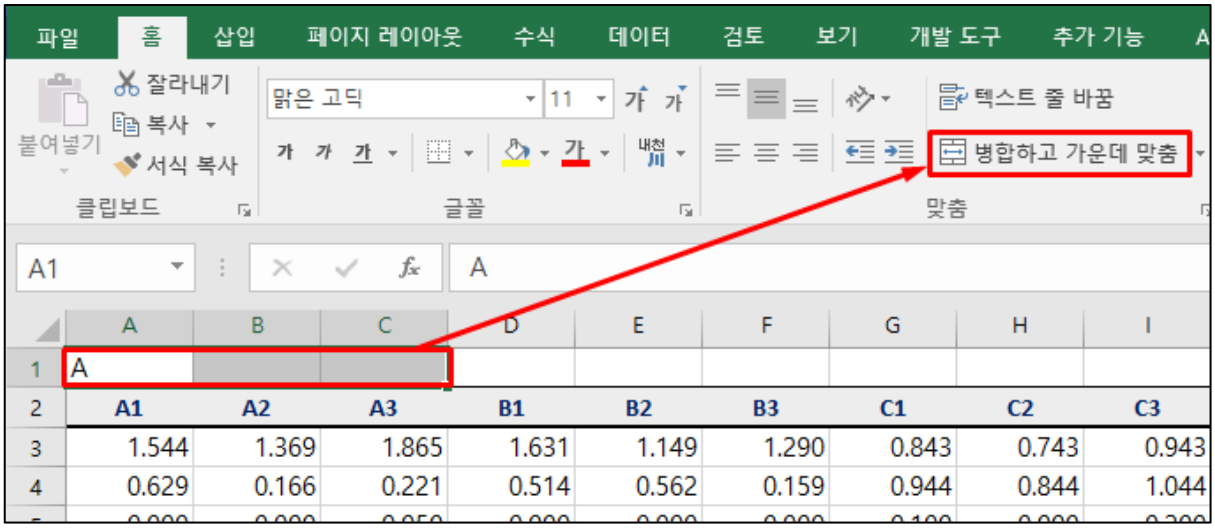
참고) 전체 데이터 선택 방법은 A1(초록색 네모) 클릭 후 Shift 를 누른 상태에서 C3(파란색 네모)를 클릭한다. A1~ C3 까지 선택 후에 Ctrl + Shift + 아래 방향키(↓)를 입력하면 전체 데이터가 한번에 선택된다.

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 4-10. Paste normalized data in new excel file

3) 1 번째 열에는 그룹명 입력 후 병합한다(그림 4-11).



	A	B	C	D	E	F	G	H	I
1	A								
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

그림 4-11. 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	0.071	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

그림 4-12. Save PCA input (group data) file

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.

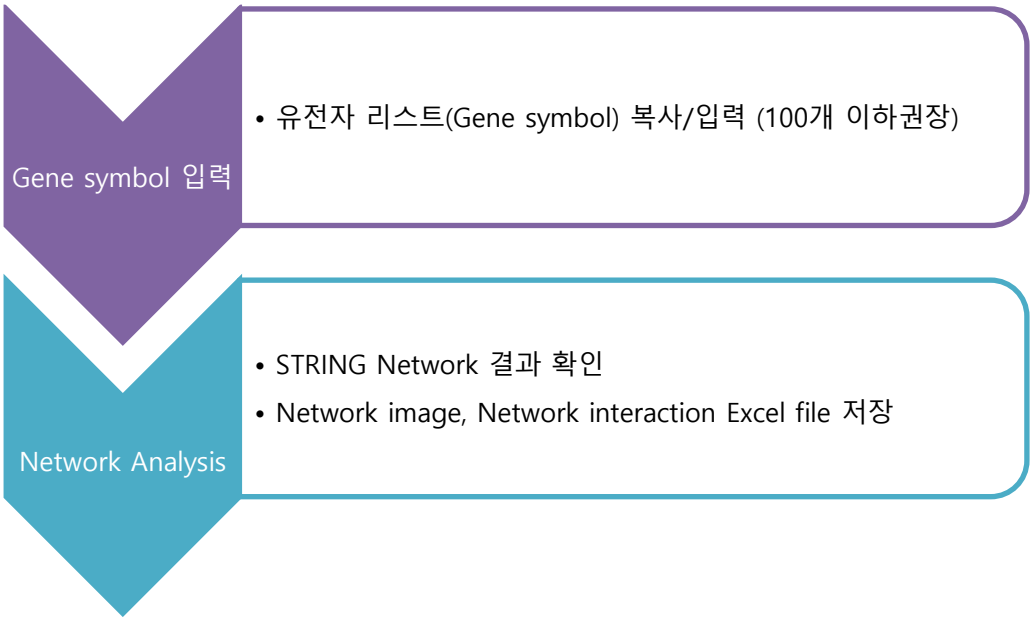


그림 5-1. STRING analysis process

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.

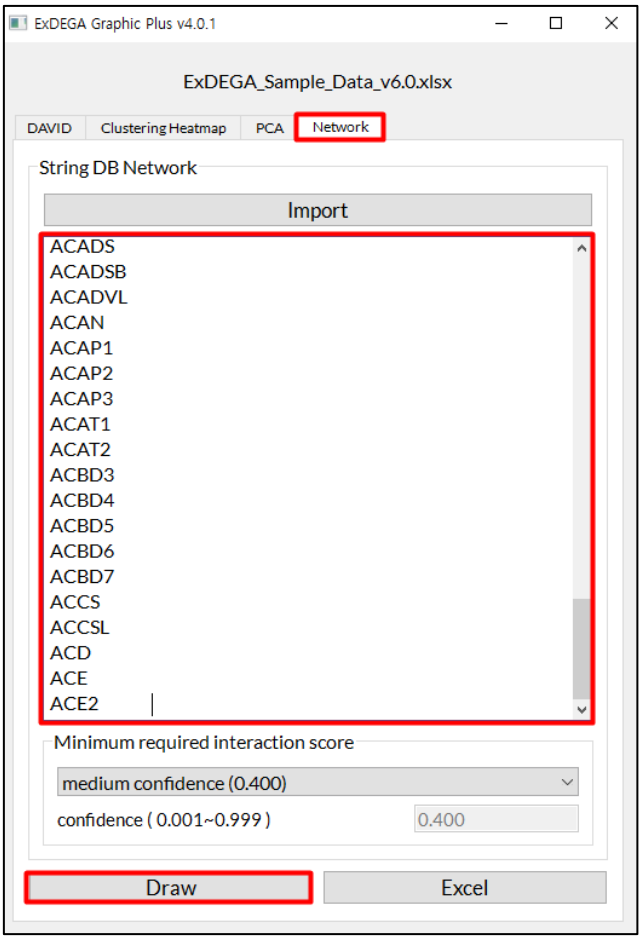


그림 5-2. Graphic plus STRING network 분석 창

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.

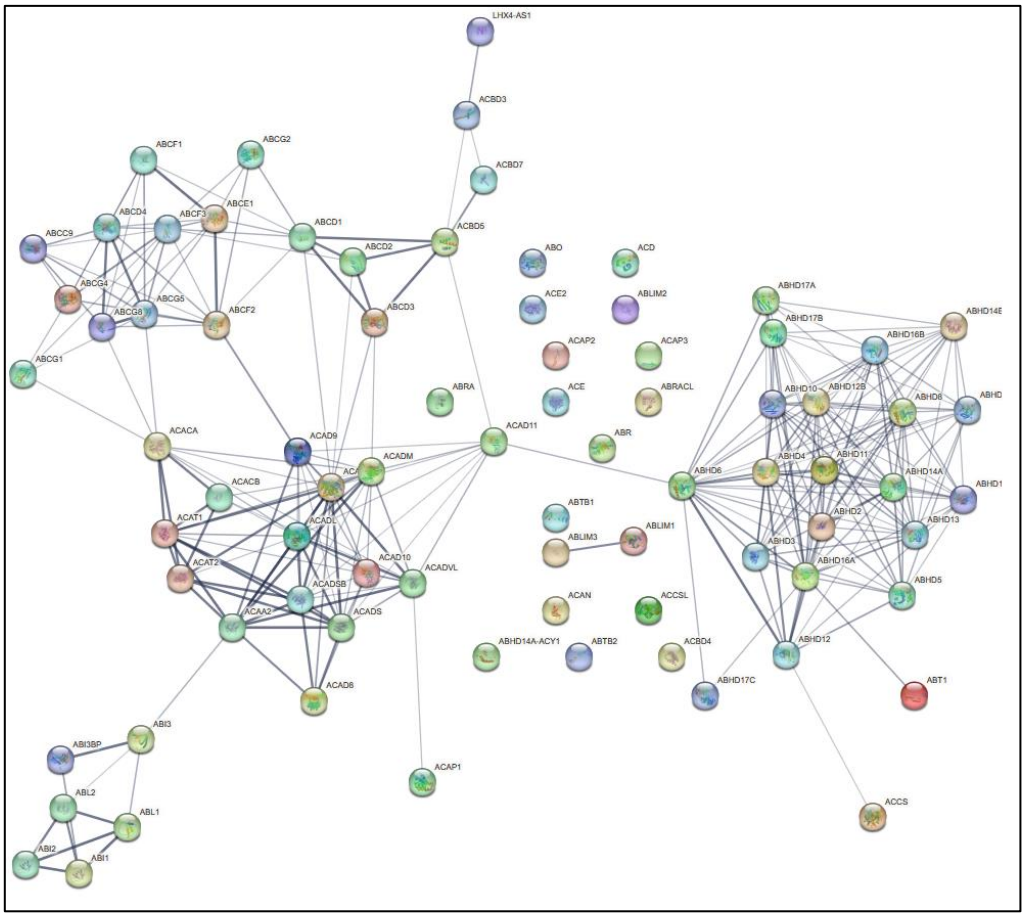


그림 5-3. STRING network 이미지

	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

STIRNG 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.

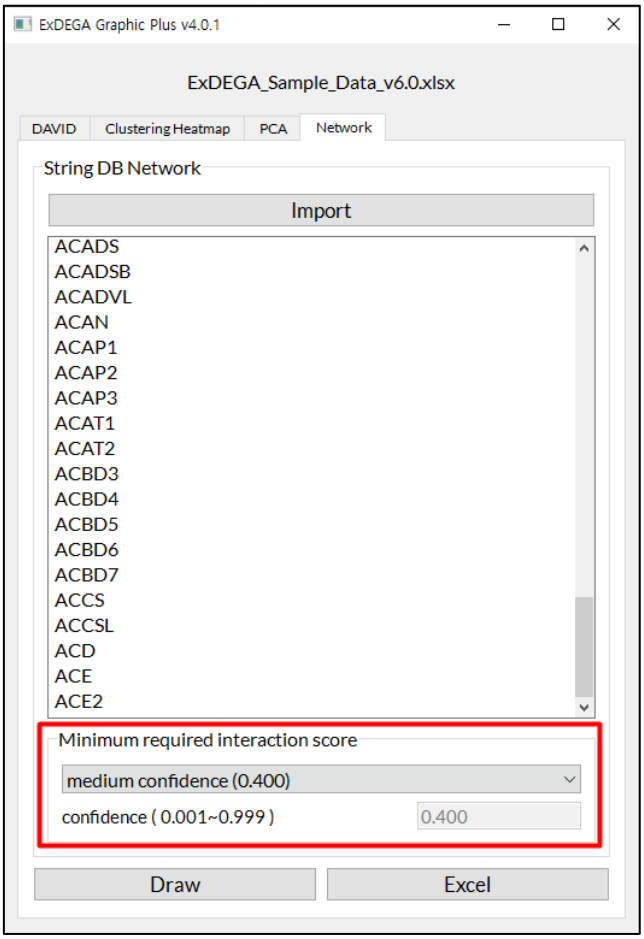


그림 5-5. STRING network score 옵션 설정

6. Pathway analysis (KEGG mapper)

RNA-Seq 분석 결과에서 up/down-regulated genes들이 어떤 Pathway에 속하는지 확인하고자 한다면 KEGG에서 제공하는 KEGG Mapper를 이용한다. 사용방법은 그림 6-1과 같은 순서로 진행된다.

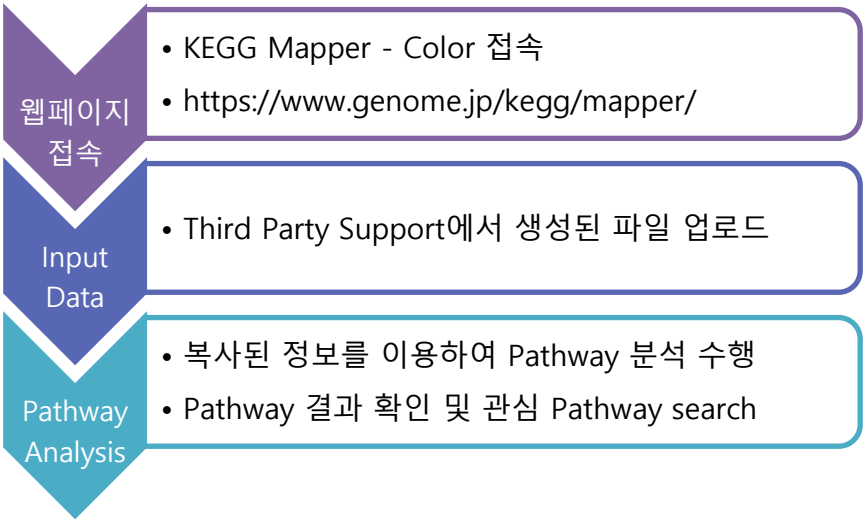


그림 6-1. KEGG Mapper tool analysis process

그림 6-2는 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.

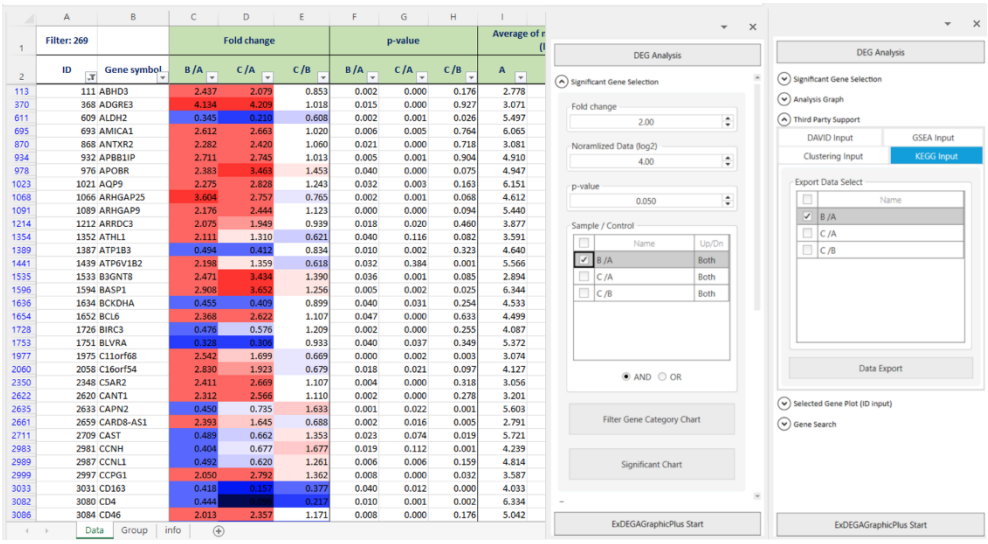


그림 6-2. KEGG Mapper input file generation process

그림 6-3과 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.

그림 6-3. KEGG Mapper tool analysis process

분석결과, 입력한 유전자들이 관여하는 pathway list 가 나온다(그림 6-4). 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.

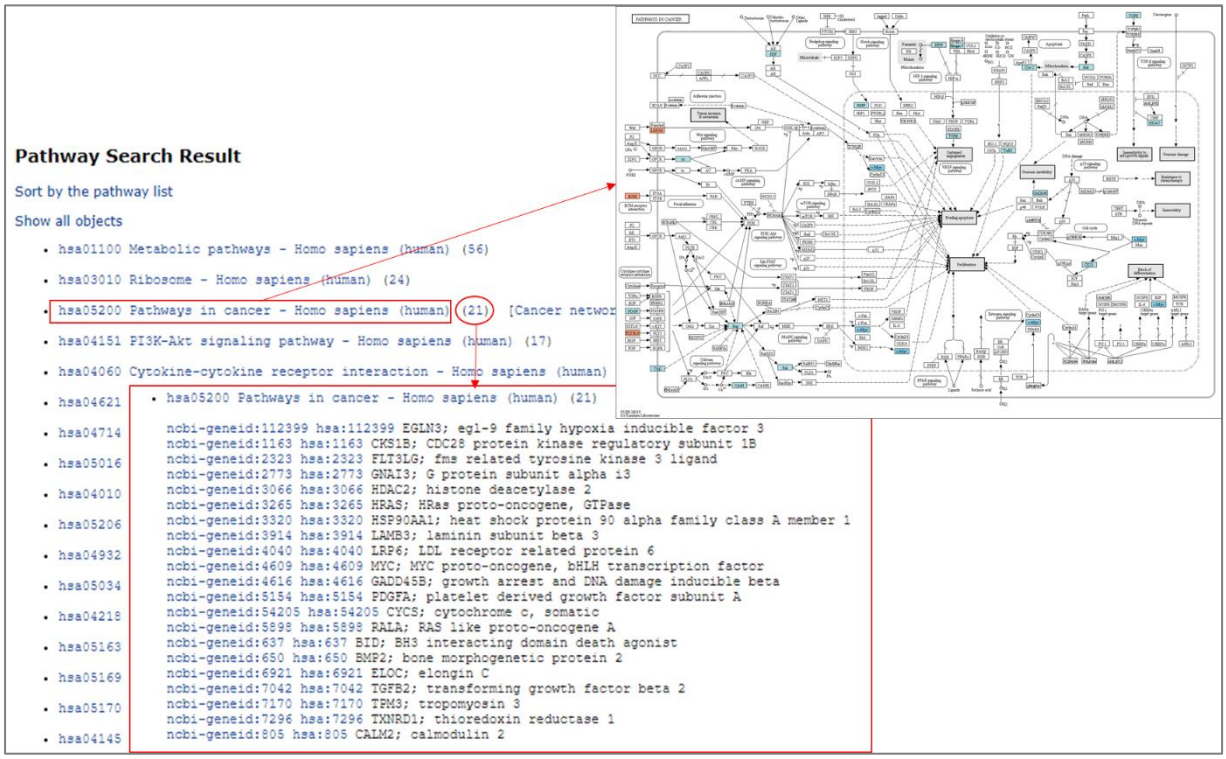


그림 6-4. KEGG Mapper tool analysis result

7. Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA)는 Microarray 또는 RNA-seq data 를 넣어 대조군, 실험군에서 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 이상이여야 진행이 가능하다.

MSigDB 에 있는 gene set (GO, pathway 등)을 기반으로 분석한다. 분석 과정은 그림 6-1 과 같은 순서로 진행된다.

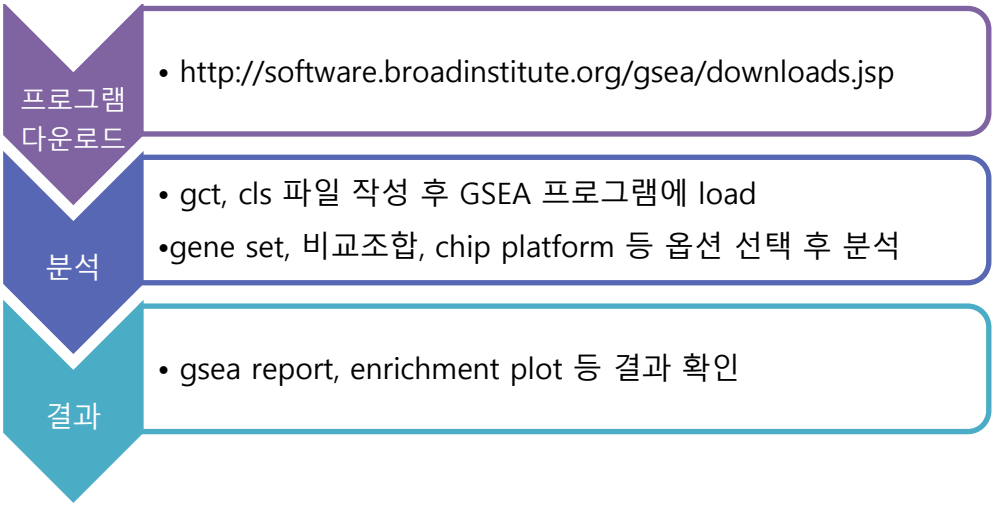


그림 7-1. GSEA tool analysis process

GSEA 홈페이지(http://software.broadinstitute.org/gsea/downloads.jsp)에 들어가 회원가입 후 로그인 하여 GSEA 프로그램을 다운로드 받는다 (그림 7-2).

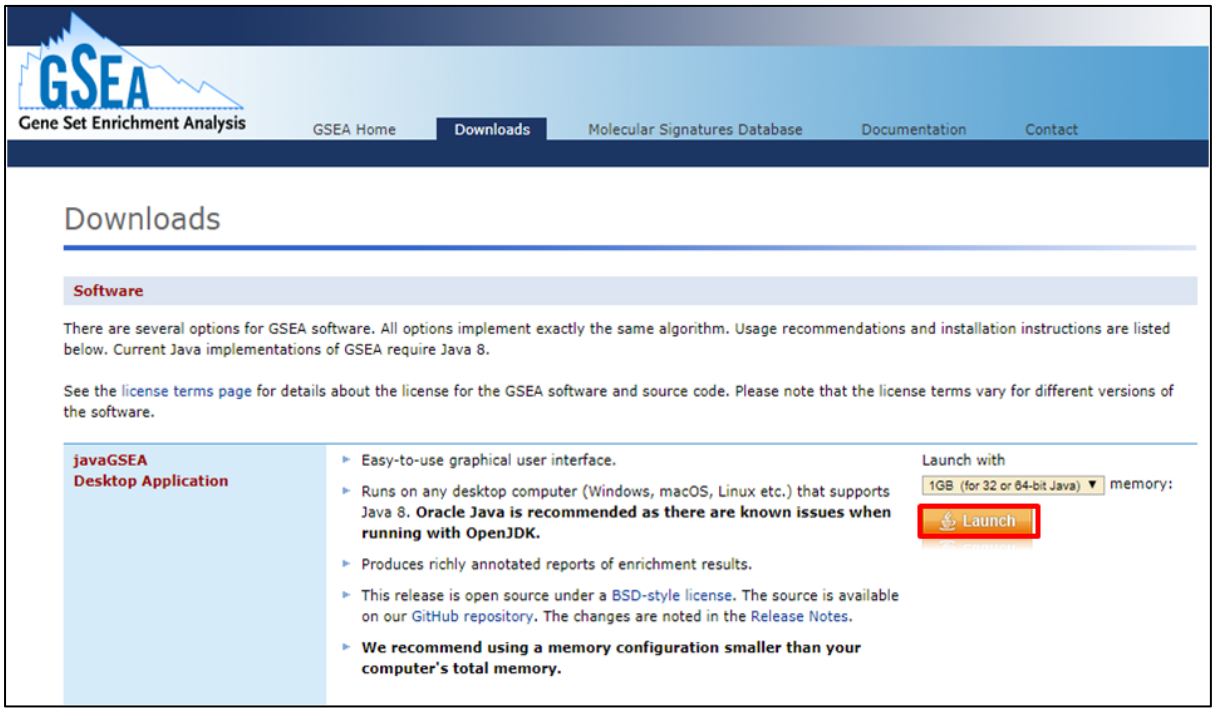


그림 7-2. GSEA program download

GSEA 분석을 위해서는 유전자 발현값 정보가 포함되어 있는 gct 파일과 샘플 정보가 포함되어 있는 cls 파일이 필요하다. Input data는 Third party support에서 추출하여 사용할 수 있다 (그림 7-3).

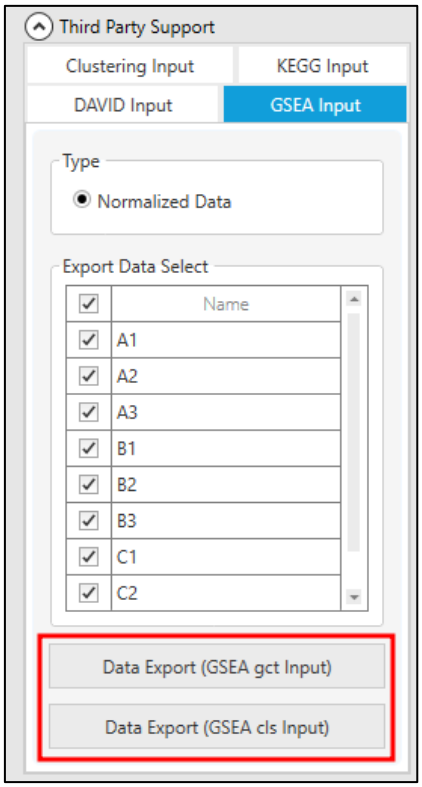
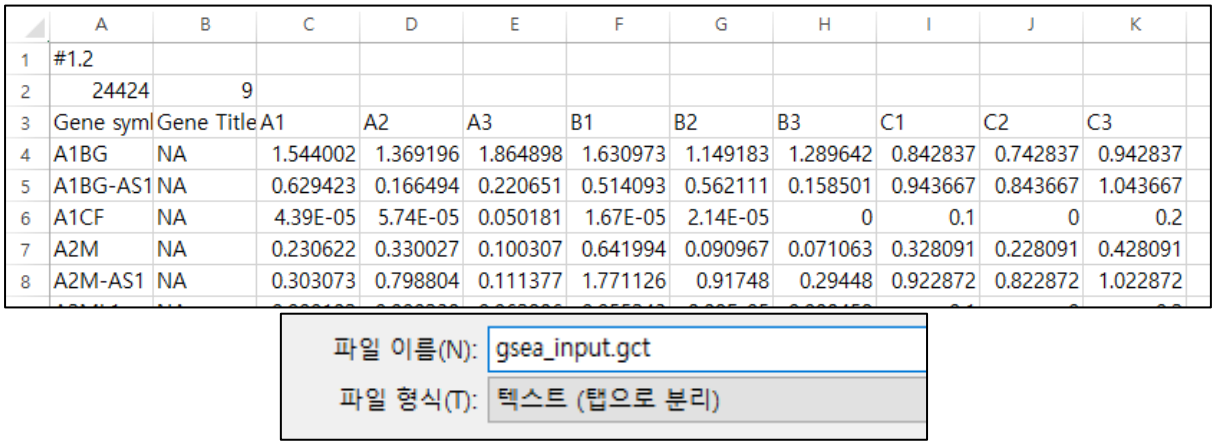


그림 7-3. GSEA Input data

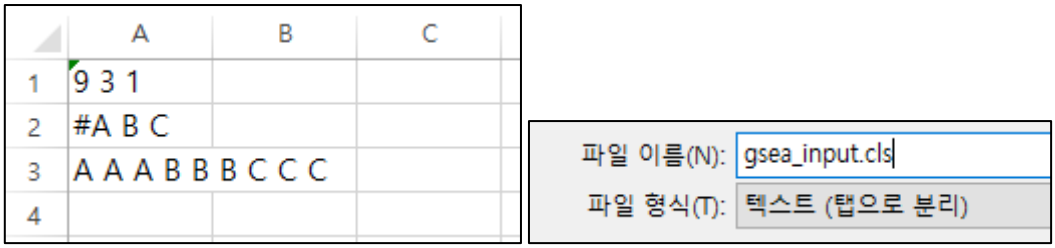
gct 및 cls 파일은 각각 그림 7-4, 7-5와 같은 형식으로 되어있으며, 전체 샘플 및 전체 유전자 리스트에 대해 분석하는 것을 권장한다.

파일 저장할 때는 파일명 뒤에 ".gct" 또는 ".cls"를 반드시 붙이고 파일 형식은 "텍스트 (탭으로 분리) 파일"로 저장한다.



	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

그림 7-4. gct file



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

그림 7-5. cls file

Microarray data의 경우, gct 파일은 해당 파일을 열어 A열에 probe ID, B열에 Gene symbol로 수정이 필요하다.

GSEA 프로그램을 열어 Load data 버튼을 누르고 Browse for files 버튼을 누른 후 gct, cls 파일을 연다(그림 7-6). 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 7-7). 메시지창 확인 후, Run GSEA 버튼을 누른다.

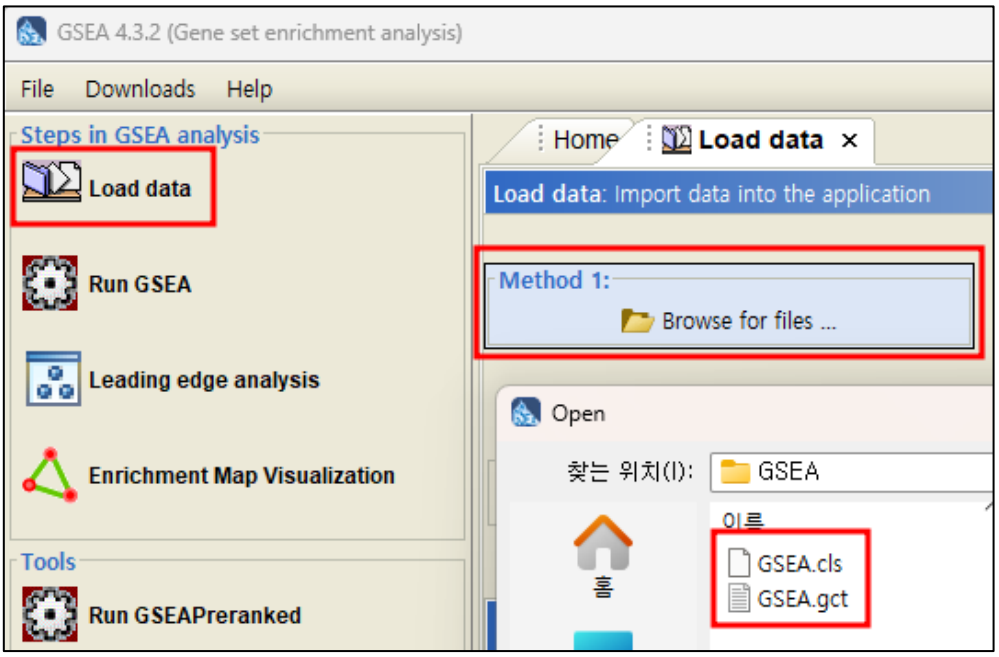


그림 7-6. Load data in GSEA program

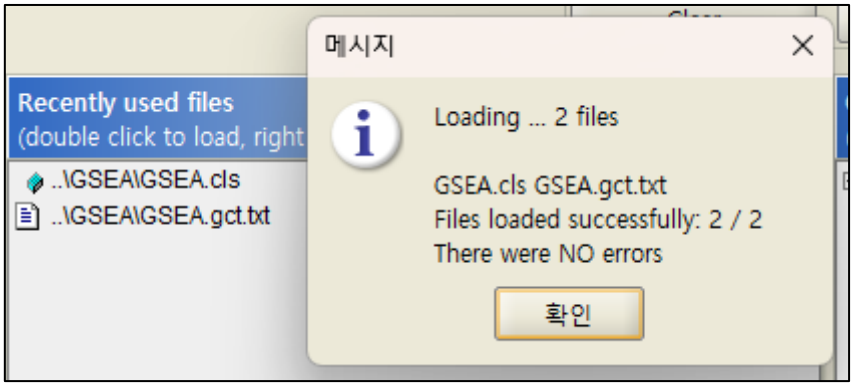


그림 7-7. Load data in GSEA program

Run GSEA 를 누르고 Expression dataset 는 gct 파일명을 선택, gene sets database 는 분석하고자 하는 gene set 을 선택한다(그림 7-8). 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.

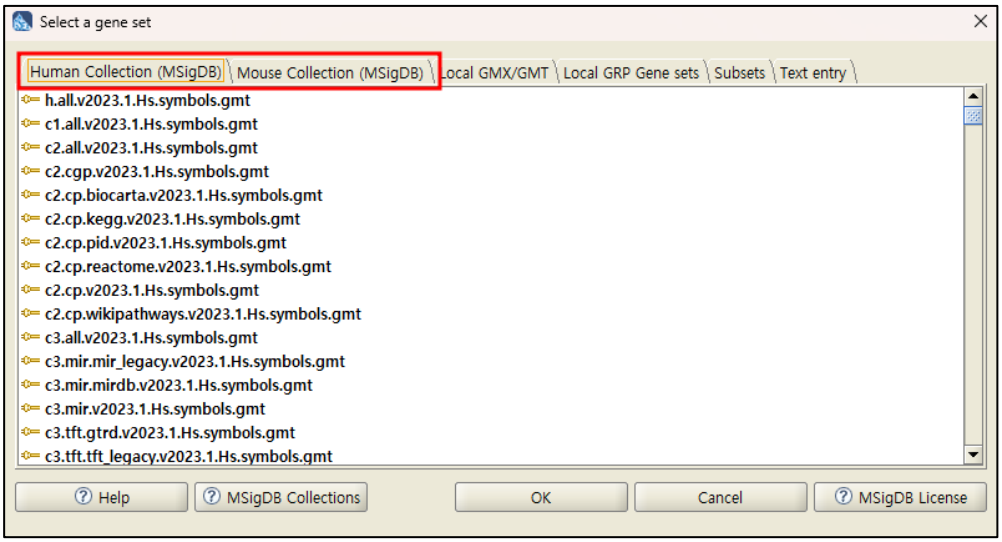


그림 7-8. Select options in GSEA program

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 gene_set 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 7-9). 여기서 주의할 점은, 이전 gene set 선택 과정 (그림 7-8)에서 선택한 탭과 같은 탭으로 진행해야 한다. Microarray 의 경우에는 실험한 chip 을 선택하여 진행한다.

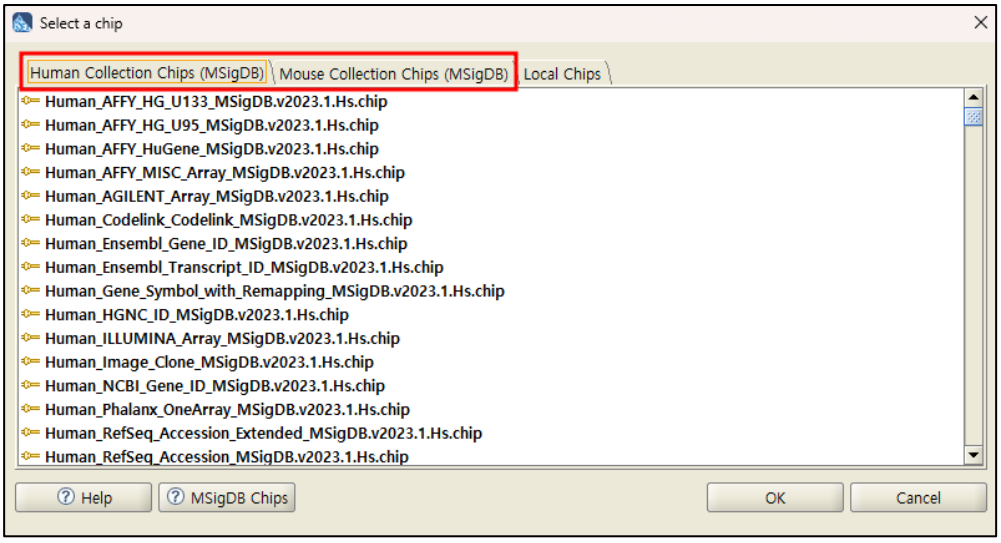


그림 7-9. Select options in GSEA program

모든 옵션이 선택이 완료되고 Run 버튼을 누르면 분석이 시작된다 (그림 7-10).

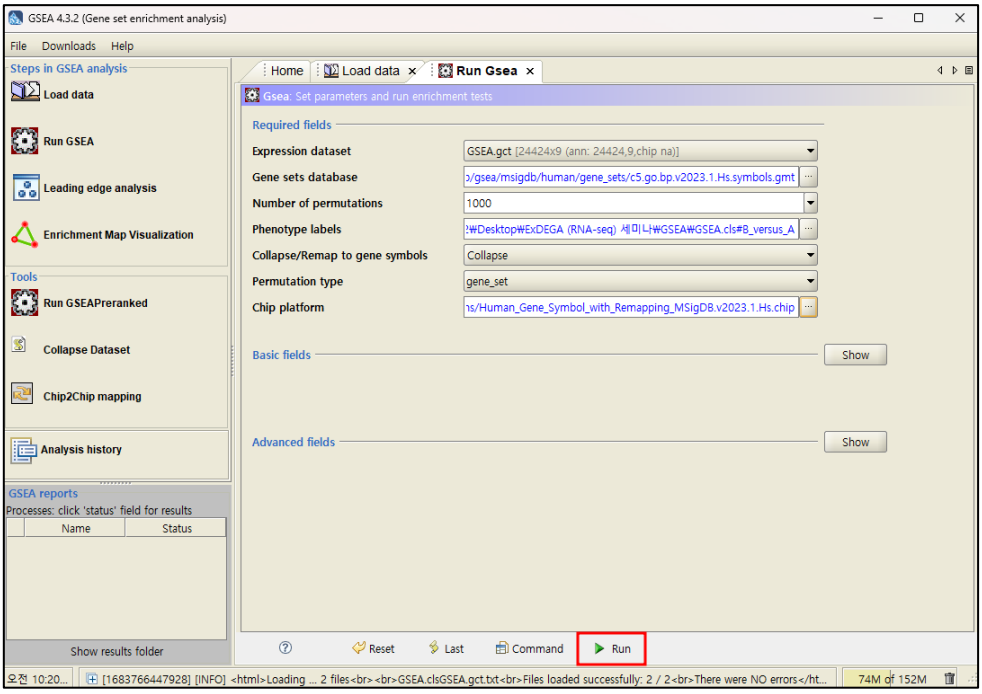


그림 7-10. Run GSEA program

분석이 완료되면 GSEA 왼쪽 아래 GSEA reports 창에 status 가 Success 로 바뀐다. Show results folder 를 누르면 GSEA 분석 결과 창이 열린다(그림 7-11).

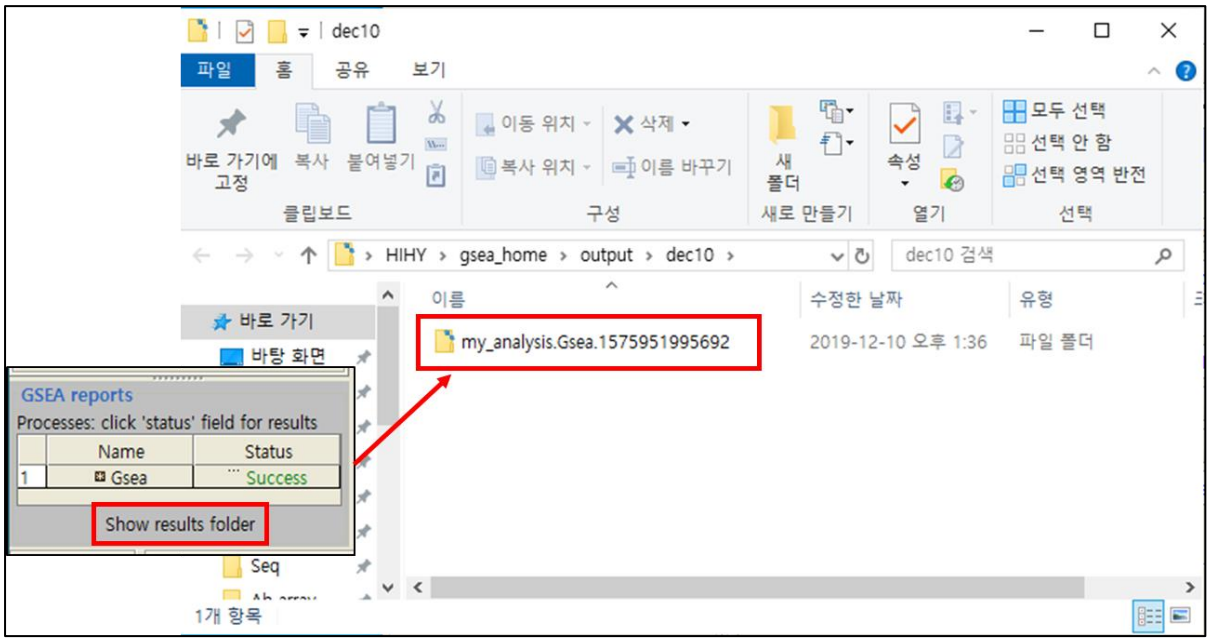


그림 7-11. GSEA results folder

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 set 이다(그림 7-12).

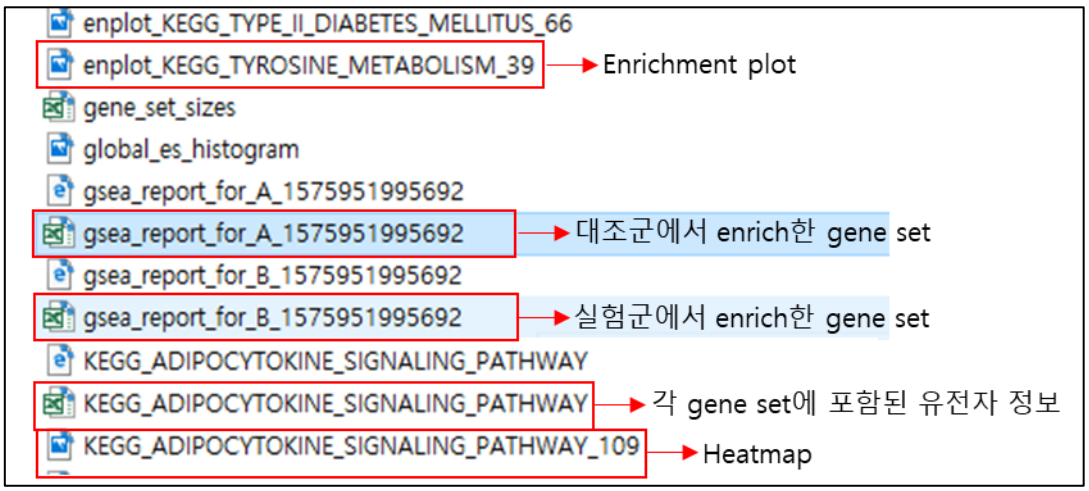
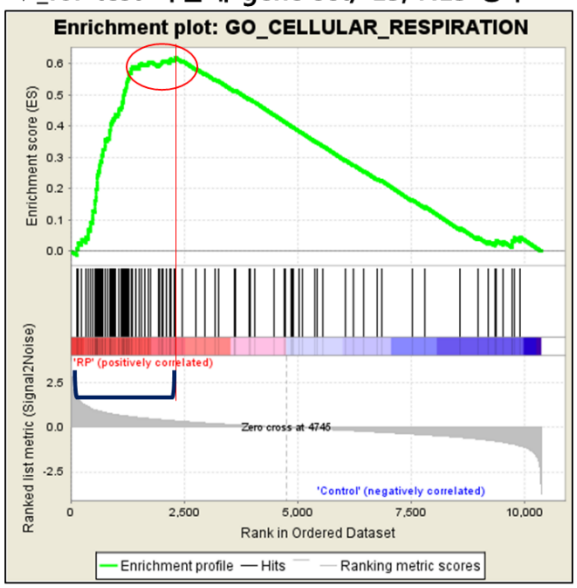


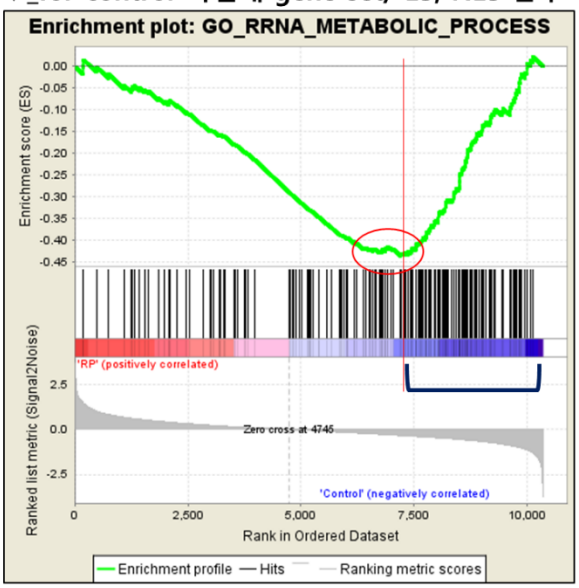
그림 7-12. GSEA result files

_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.

▼_for test 파일내 gene set/ ES, NES 양수



▼_for control 파일내 gene set/ ES, NES 음수



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

그림 7-13. 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.

8. 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 8-1 과 같다.

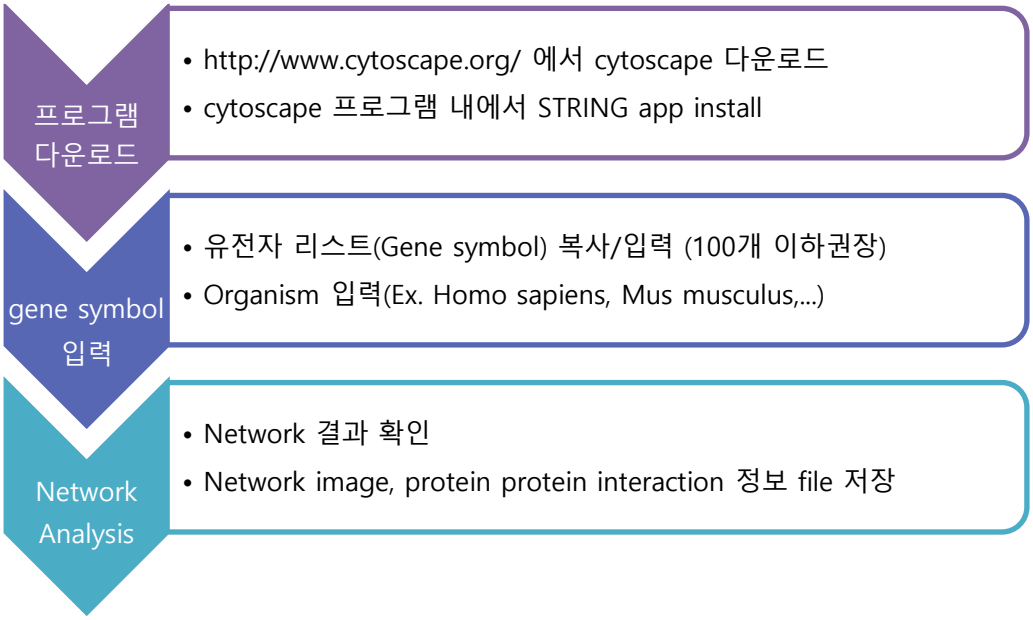


그림 8-1. STRING analysis process

Cytoscape 홈페이지 (http://www.cytoscape.org/)에서 cytoscape 프로그램을 다운로드 받아 설치한다(그림 8-2).

그림 8-2. Cytoscape download

Cytoscape 프로그램을 열어 상위에 있는 메뉴 중 [Apps] > [App Manager]로 들어간다(그림 8-3). StringApp 을 선택 후 Install 버튼을 누른다.

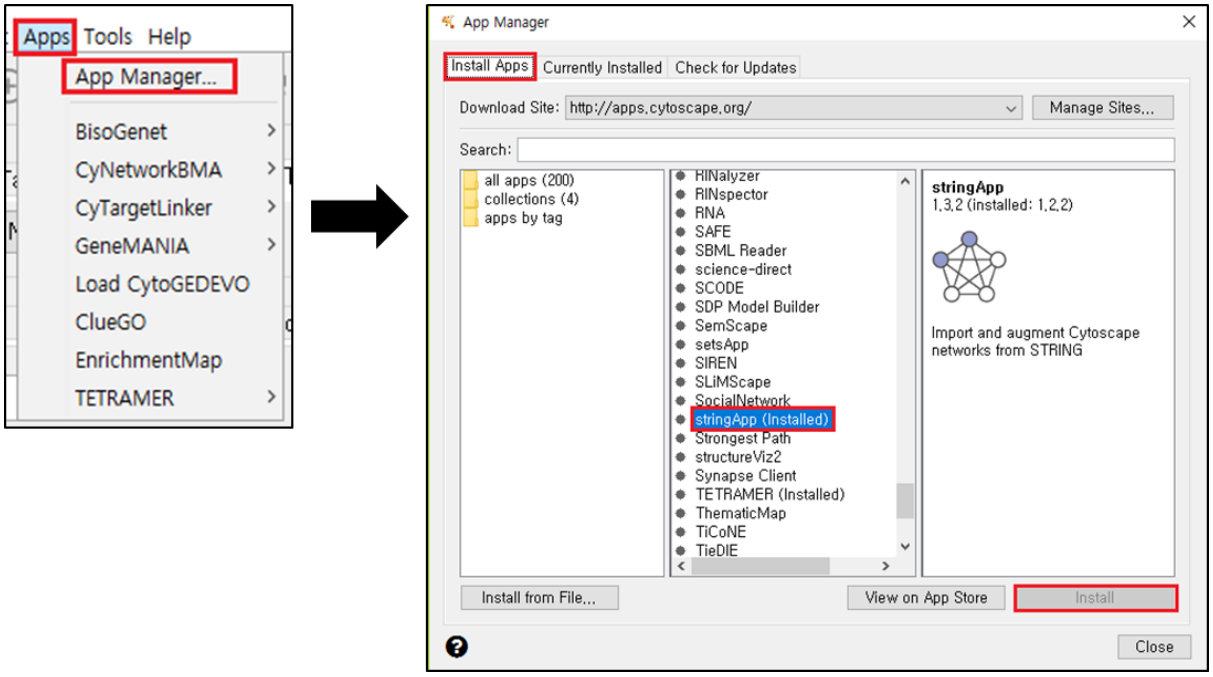


그림 8-3. STRING app installation in Cytoscape

Cytoscape 상위 메뉴 중 [File] > [Import] > [Network from Public Databases]로 들어간다(그림 8-4).

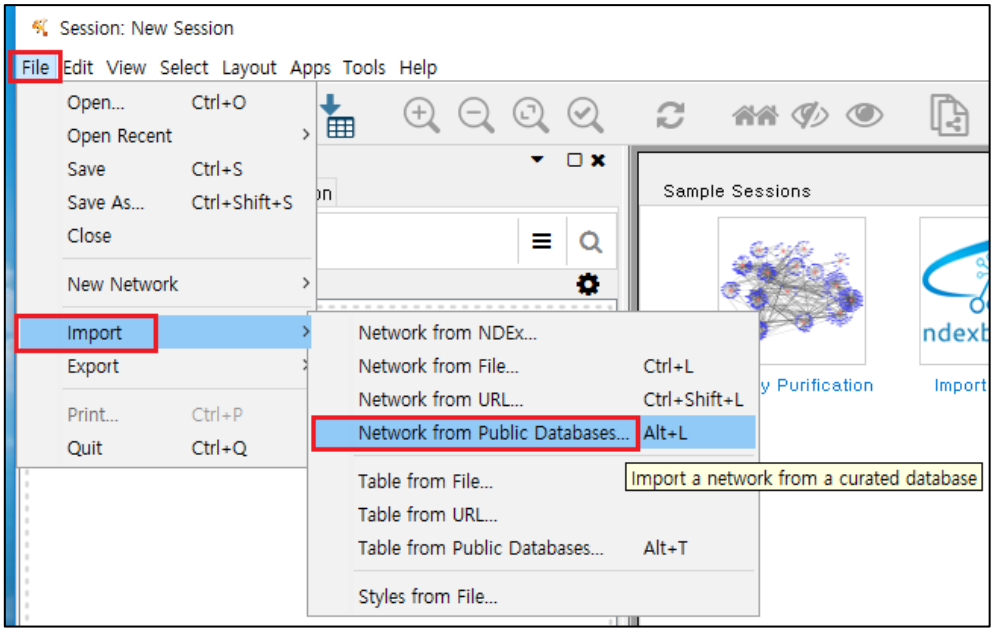


그림 8-4. STRING analysis process 1

Data Source 를 "STRING : protein query" 선택하고 Species 를 선택한다(그림 8-5). 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.

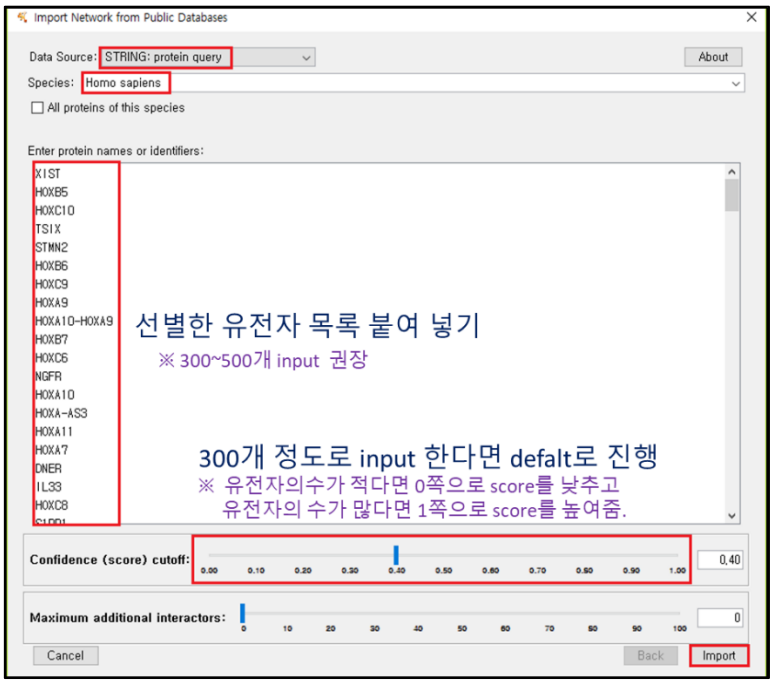


그림 8-5. STRING analysis process 2

Input 한 gene symbol 과 match 가 되지 않는 protein 이 있으면 그림 8-6 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.

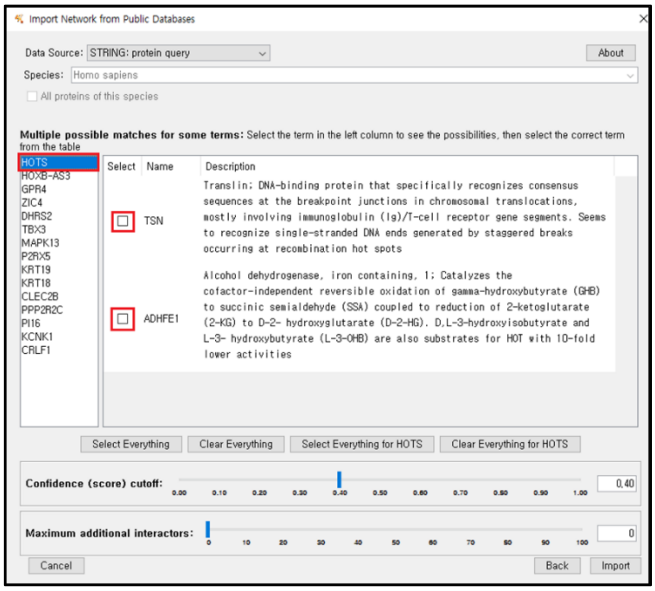


그림 8-6. Not matched proteins in STRING

분석이 완료되면 network image 가 나온다(그림 8-7).

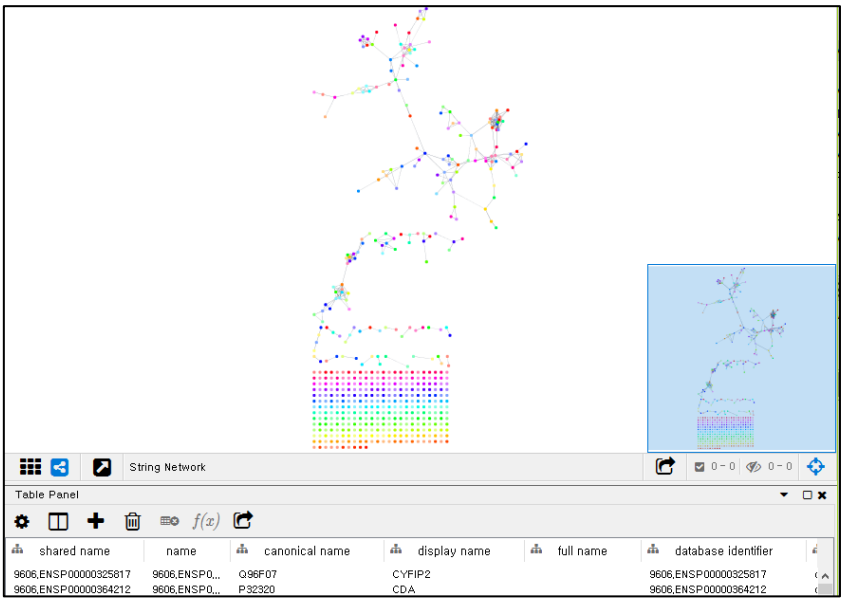


그림 8-7. Network result

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

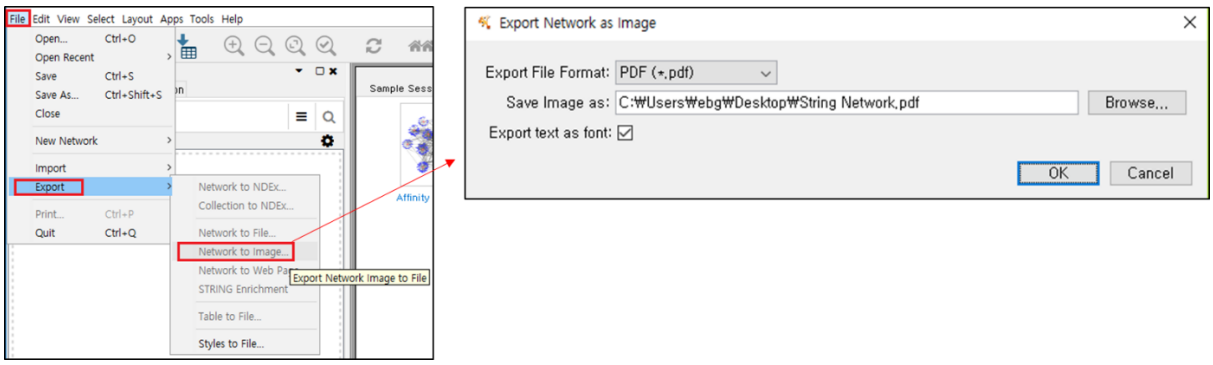


그림 8-8. Save network image

어떤 유전자들이 protein-protein interaction 을 하는지 정보를 저장하고 싶으면 [File] > [Export] > [Table to File...]로 들어가 String Network default edge 파일을 저장한다(그림 8-9).

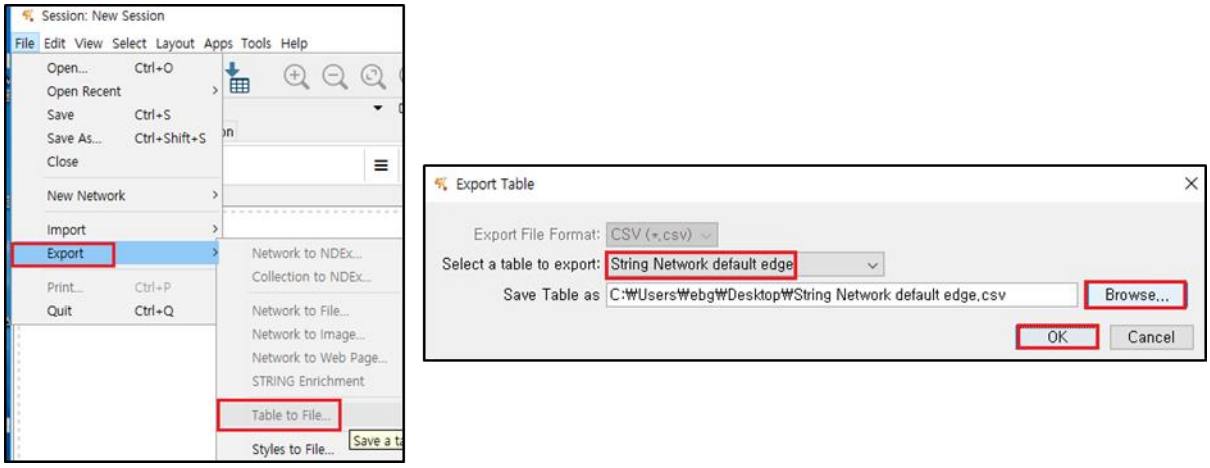


그림 8-9. Save edge table

String Network default edge 파일에서 name 에 interaction 정보, score 에 confidence score 가 나와있다(그림 8-10). 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	coexpr	cooccu	databa	experir	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 8-10. Interaction information in edge table

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