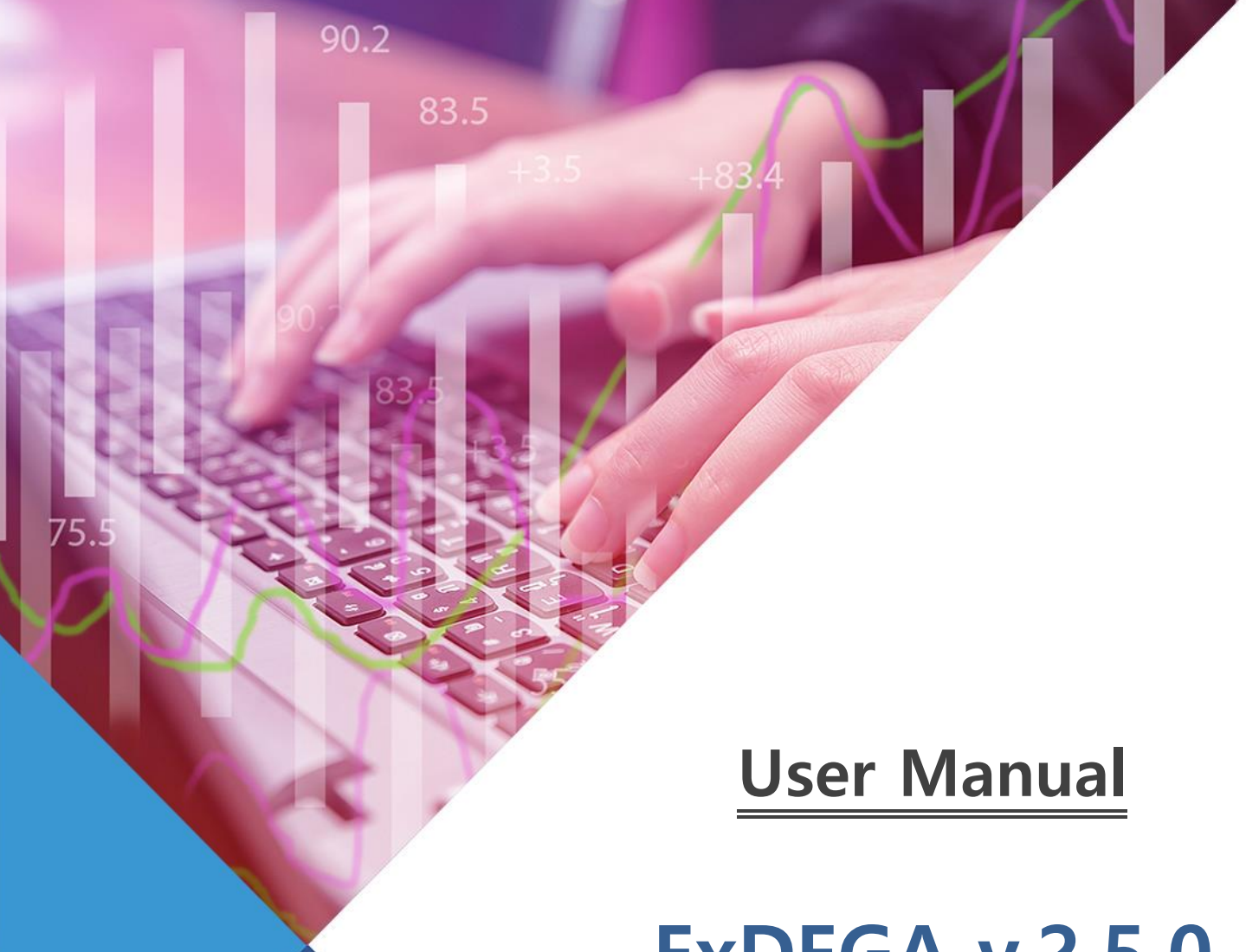
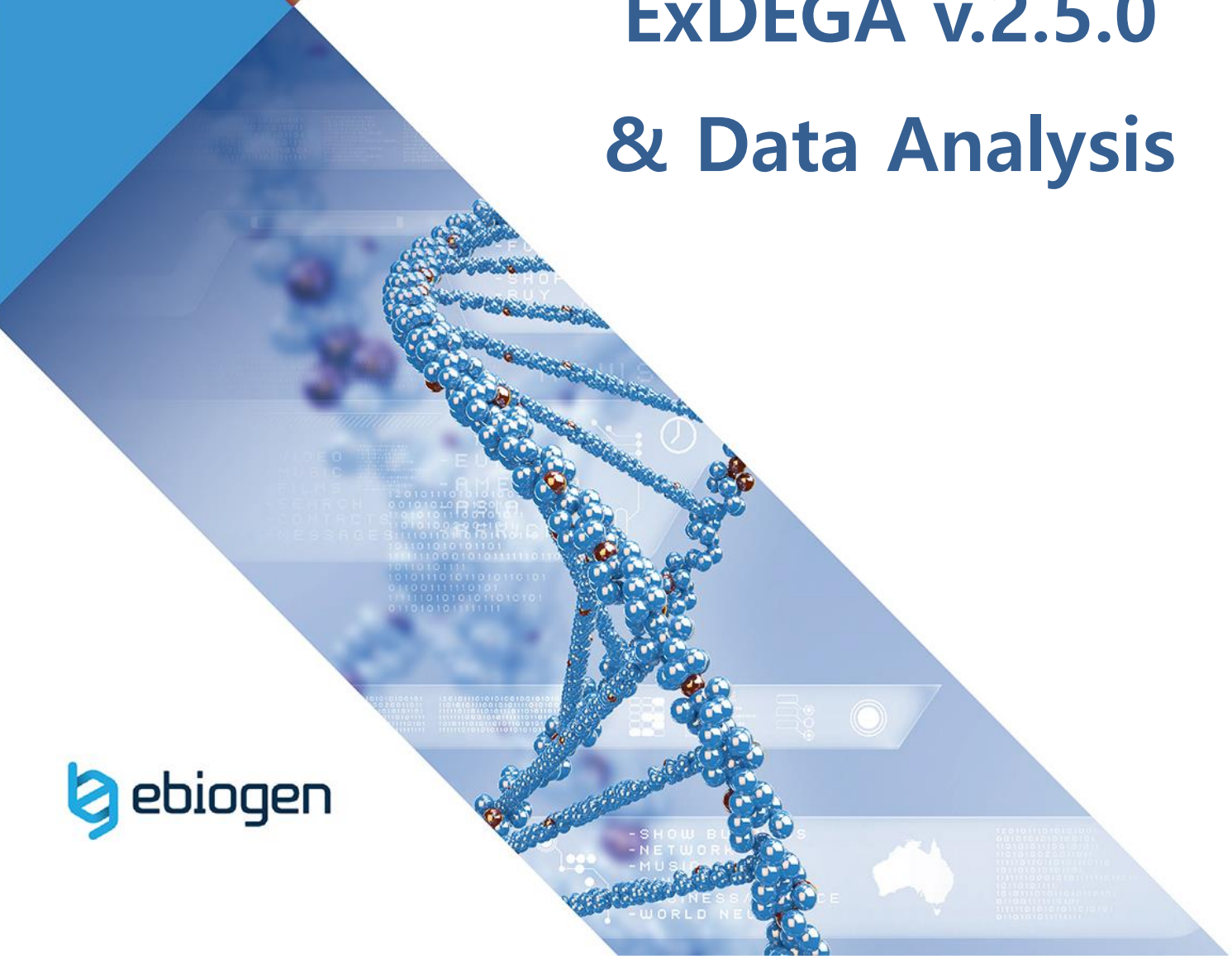


User Manual

ExDEGA v.2.5.0 & Data Analysis



ebiogen

<목 차>

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 을 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(버전).zip 파일의 압축을 풀고 setup 을 실행하면 분석 툴이 설치된다(그림 1-1).

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 매뉴얼 (Download link)을 확인한다.

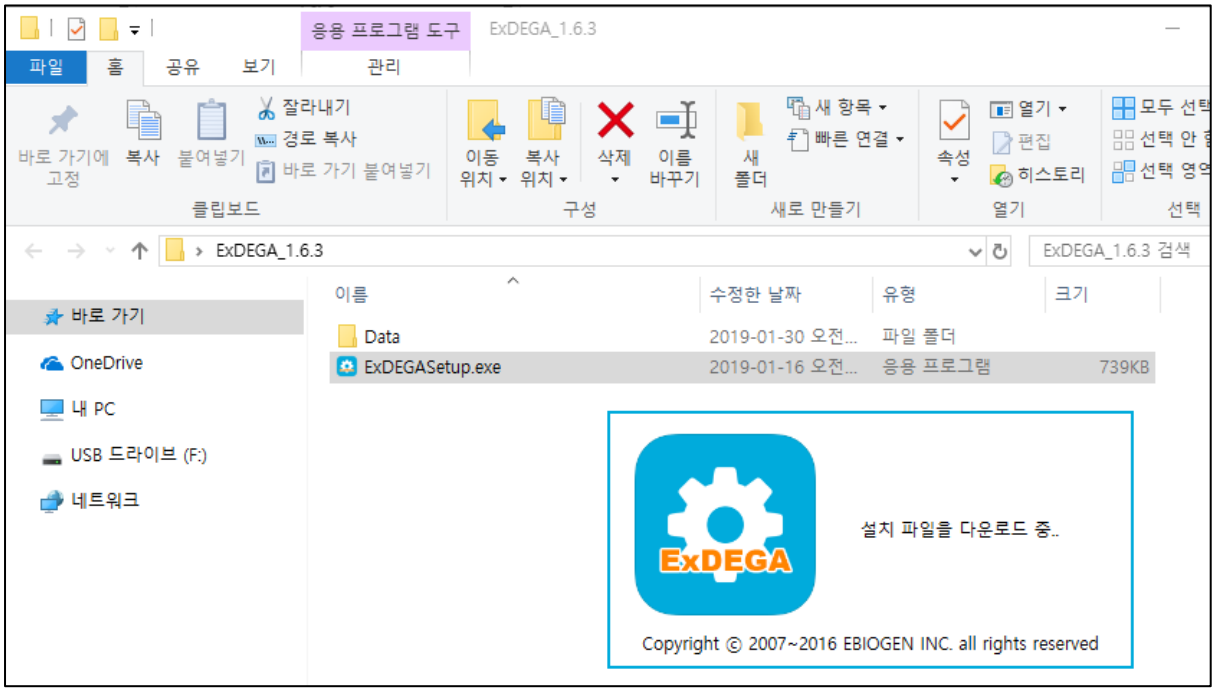


그림 1-1. ExDEGA set up

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 그래프 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.

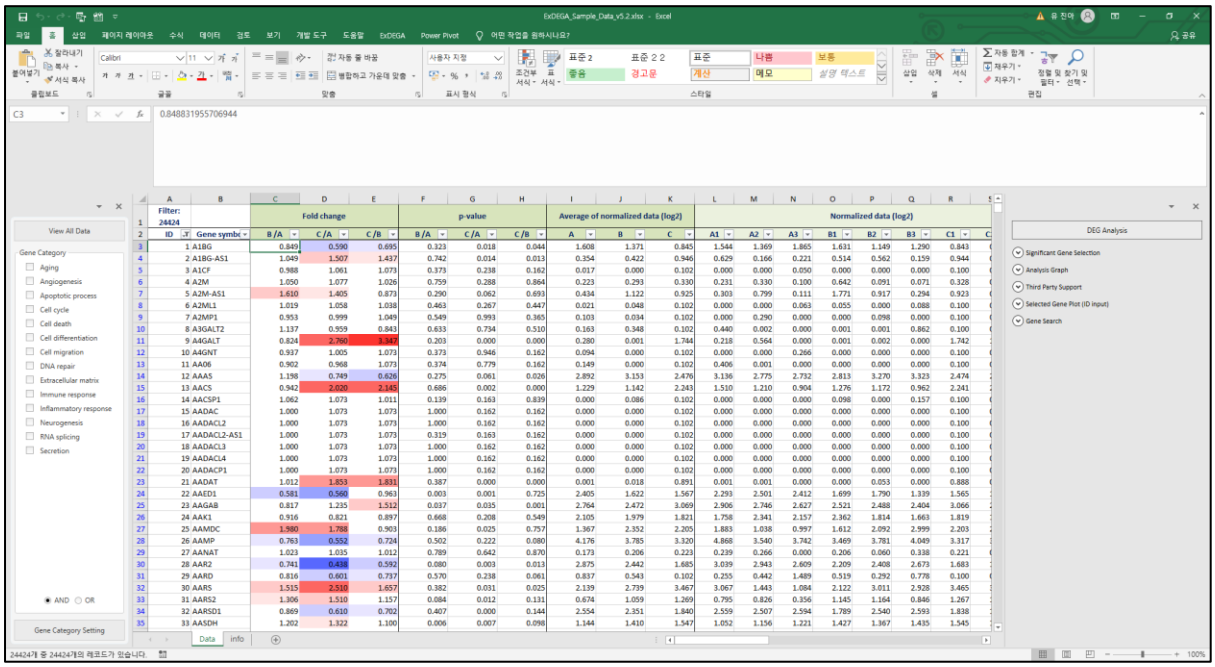


그림 1-2. RNA-seq or Microarray data in ExDEGA format

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 수많은 GO 중 임의로 15 개를 선택하여 관련 유전자를 필터링 할 수 있도록 만들어 놓은 것이다. 예를 들어, Aging 관련 유전자만 분석을 원할 경우, Gene Category 창에서 Aging 을 선택하면 해당 유전자 리스트만 필터링 된다(그림 1-3). 그리고 Gene Category 의 여러 항목들을 동시에 만족하는 유전자를 필터링할 수 있고 적어도 한 항목만이라도 포함하는 유전자를 보고자 하는 경우도 필터링이 가능하도록 "AND"와 "OR" 기능을 갖추고 있다.

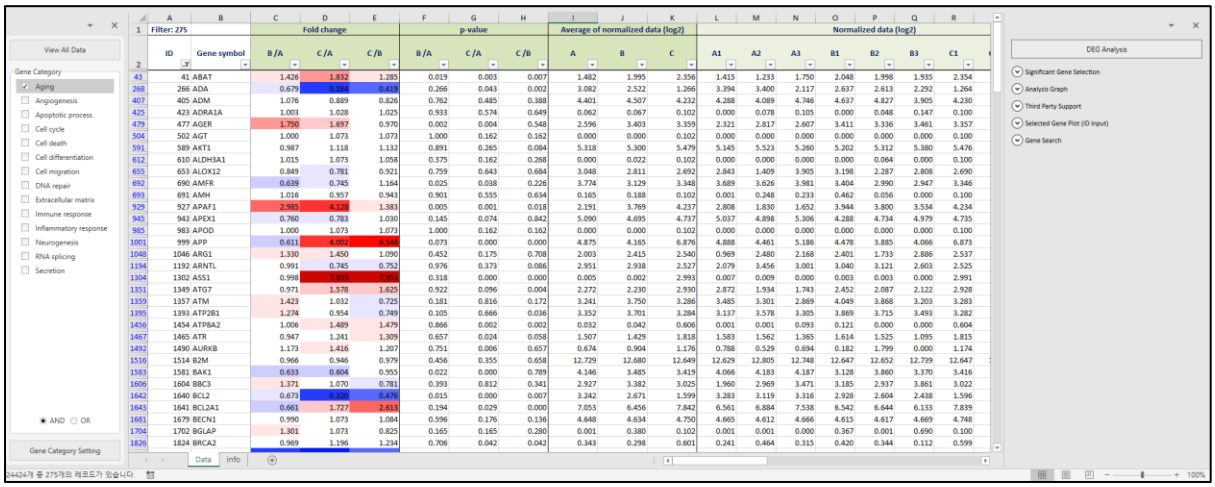


그림 1-3. Gene ontology (Aging) selection

가장 왼쪽 상단에 'View All Data' 버튼을 누르면 필터를 모두 해제되어 다시 전체 결과를 볼 수 있고 15 개의 GO 중 관심 기능이 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.

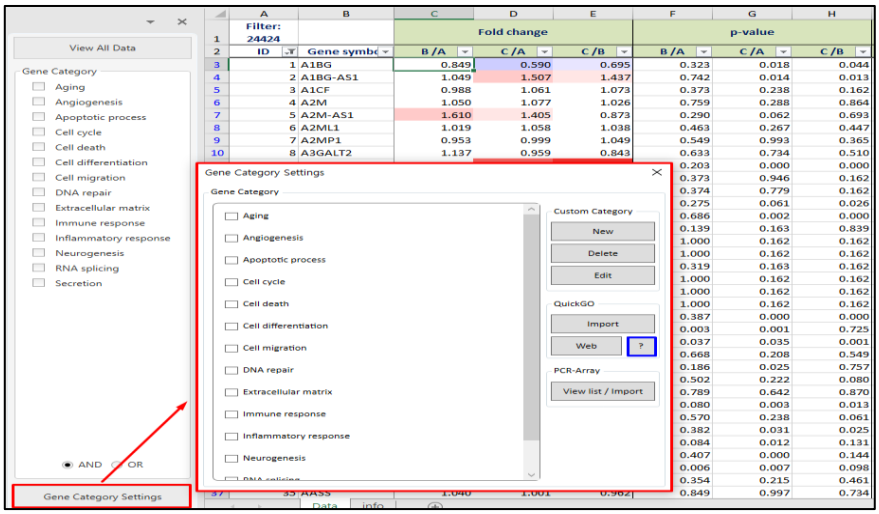


그림 1-4. Gene category settings

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 저장하면 새로운 Gene category 를 확인 할 수 있다(그림 1-5. a, b).

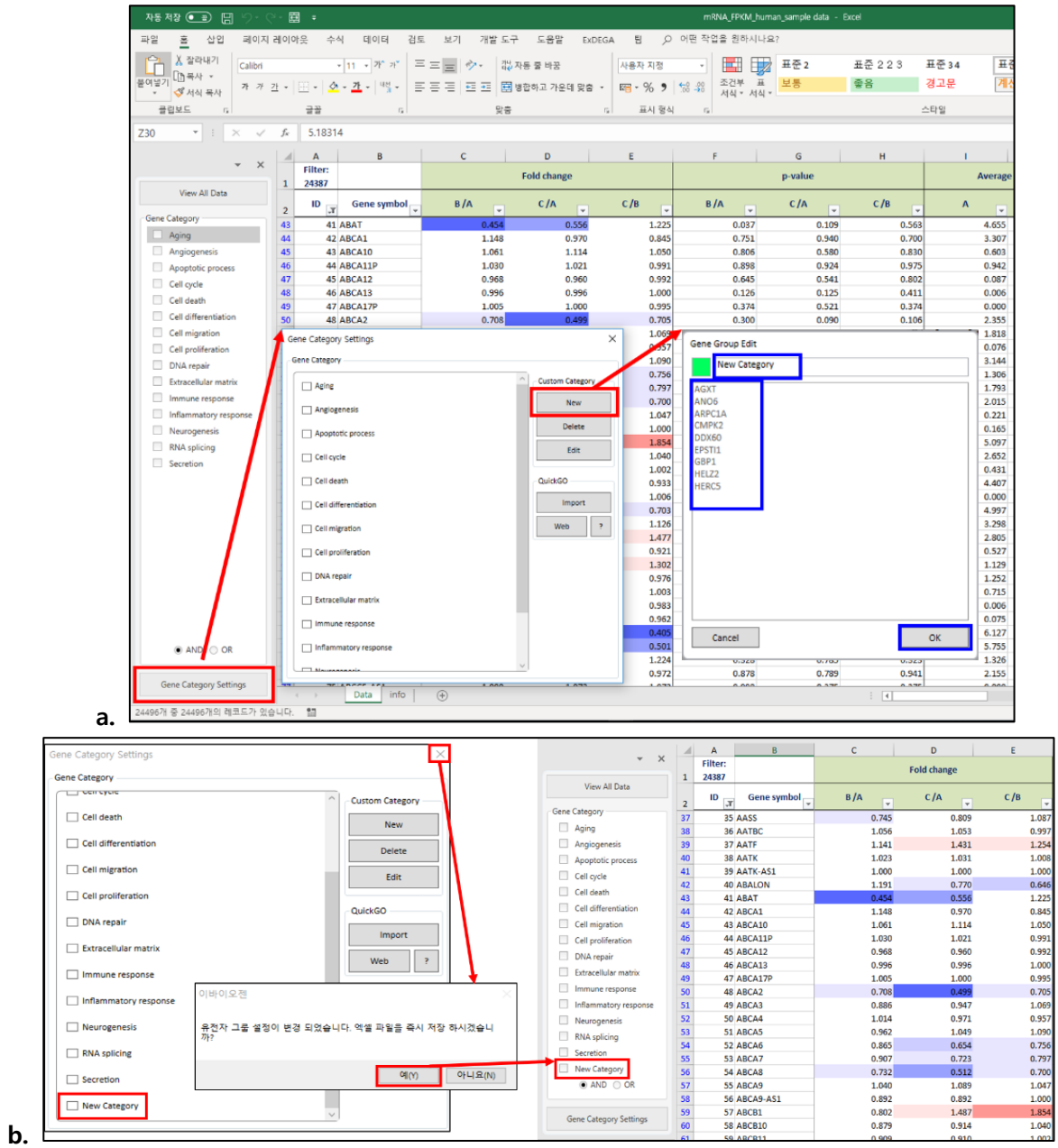


그림 1-5. Adding Genes to make a new gene category

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 1-6).

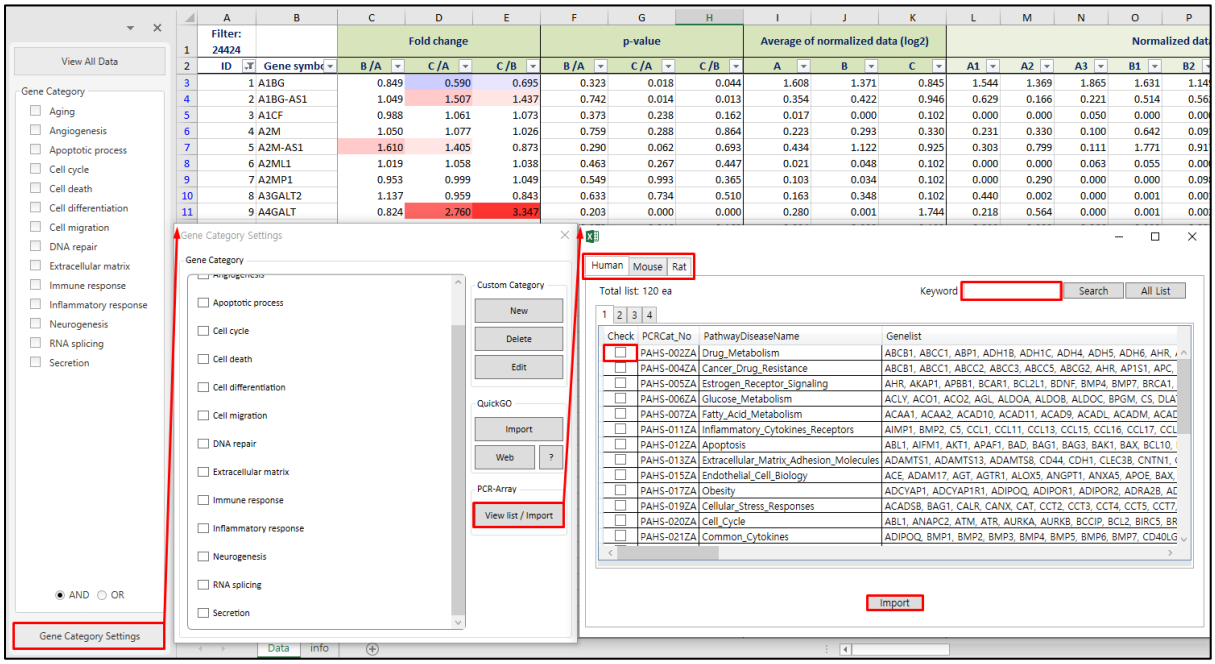


그림 1-6. PCR-Array Pathway settings

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 비교조건을 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 1-7). P-value 는 반복 실험한 데이터(N>=2)의 경우만 제공된다. 비교그룹을 다중 선택할 수 있다. "AND"나 "OR"를 기능을 이용하면 선택한 비교그룹들에서 공통적인 DEGs (교집합) 또는 하나의 비교그룹 이상 DEGs (합집합)을 선별할 수 있다.

참고로 유전자 선별 시 보통 fold change 2 이상, p-value 0.05 이하를 기준으로 선별한다. Normalized data (log2)는 정해져 있는 기준은 없으나 FPKM 의 경우엔 1 이상, RC 의 경우엔 4~6 이상을 기준으로 선별하면 발현값이 낮은 유전자들을 제외할 수 있다. 위 기준을 꼭 따라야 하는 것은 아니며, 유전자 선별 기준은 연구자의 데이터에 맞게 조정하여 사용할 수 있다.

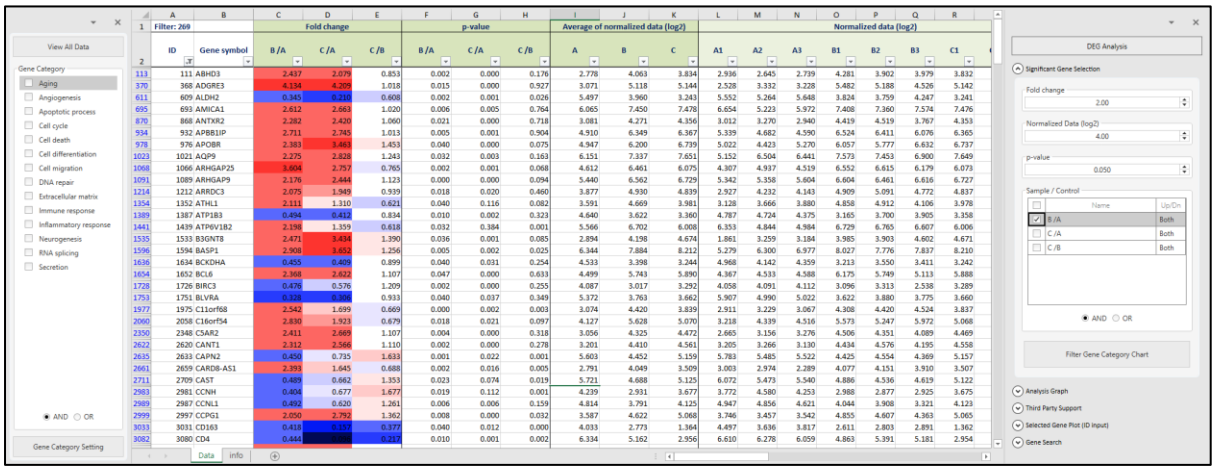


그림 1-7. Significant gene selection

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)

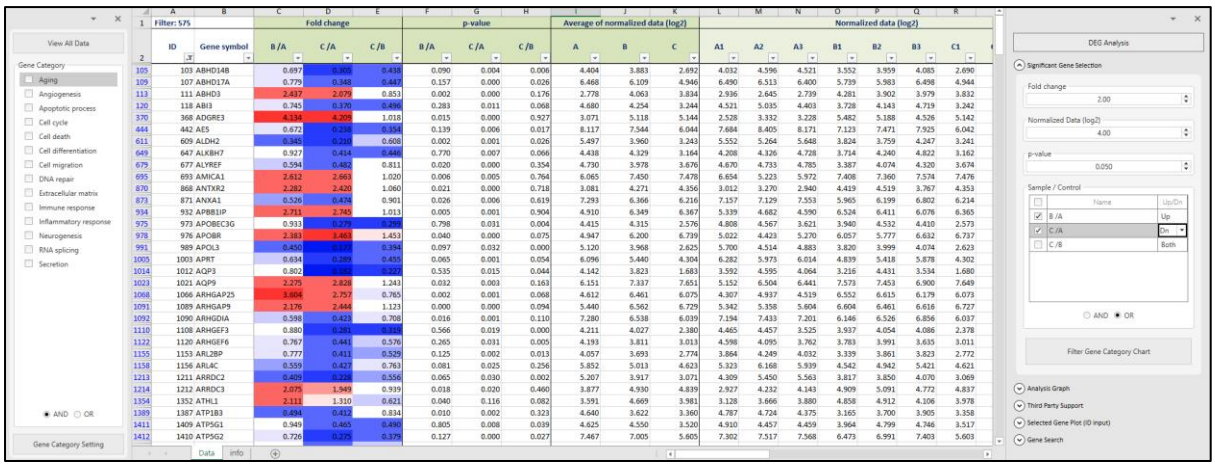


그림 1-8. Significant genes (separately up and down)

Gene Category와 Significant gene selection은 연동 가능하다. 그림 1-8에서처럼 Significant Gene Selection에서 DEGs를 선택하고 Gene Category의 Cell differentiation을 선택하면 5개의 유전자가 필터링 된다(그림 1-9). 5개의 유전자는 본 데이터에서 Cell differentiation 관련 유전자들 중 B/A 비교그룹에서 유의하게 발현이 증가 또는 감소한 유전자를 의미한다.

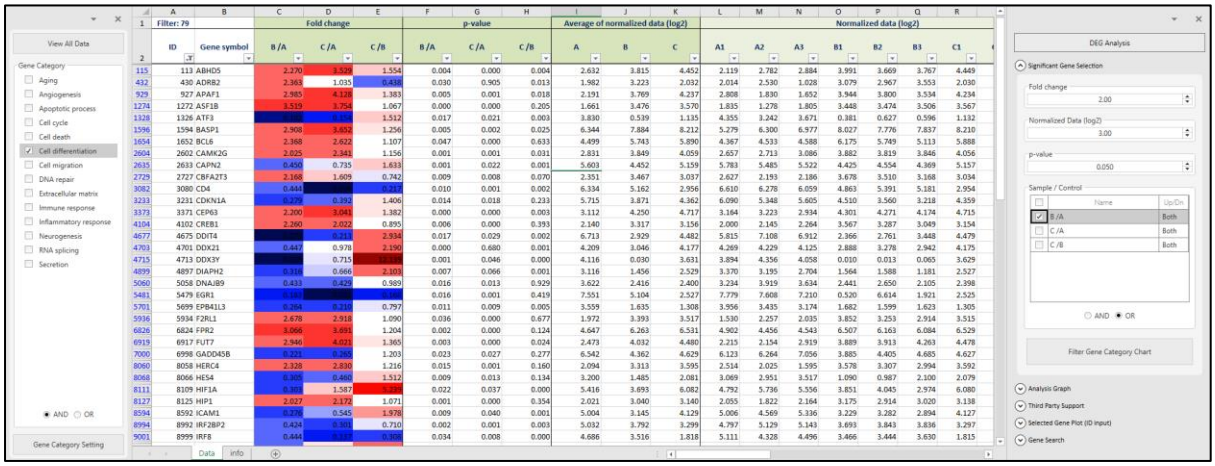


그림 1-9. Significant genes related to Cell differentiation

Gene Category Chart는 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 증가/감소된 bar chart 에서 bar 상단의 숫자는 해당 유전자 수이며 bar 를 클릭했을 때 해당 유전자가 필터링 된다(그림 1-10).

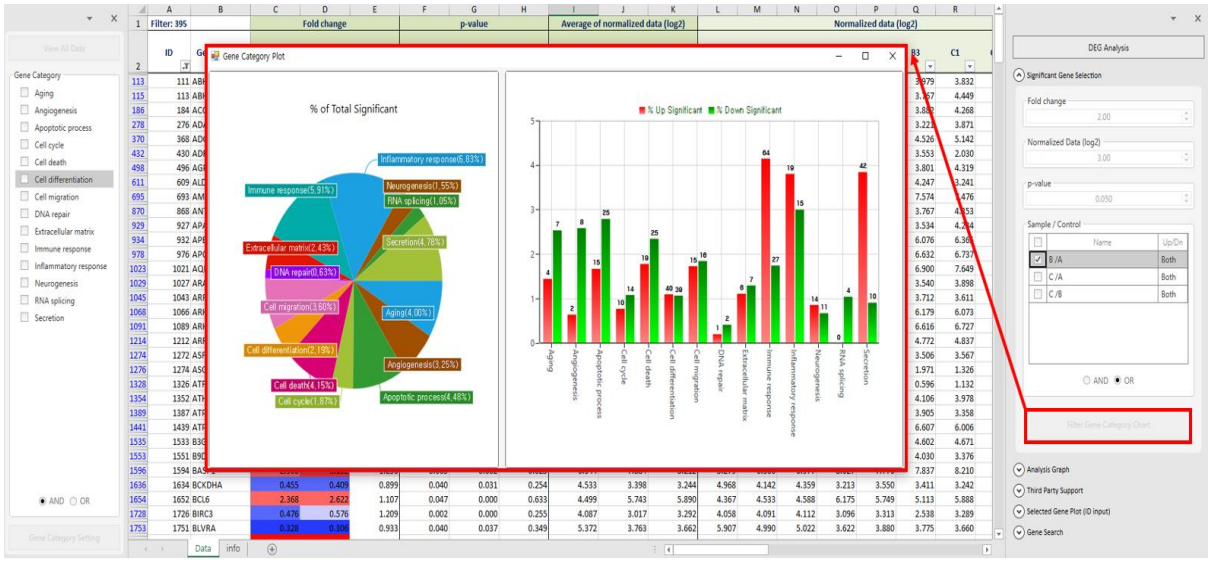


그림 1-10. View Gene Category Chart

1-3. Analysis Graph 사용 방법

DEG Analysis 부분에서 "Analysis Graph" 창을 펼치면 그림 1-11 와 같이 Scatter Plot, Volcano Plot, Venn Diagram 을 엑셀에서 쉽게 그릴 수 있다.

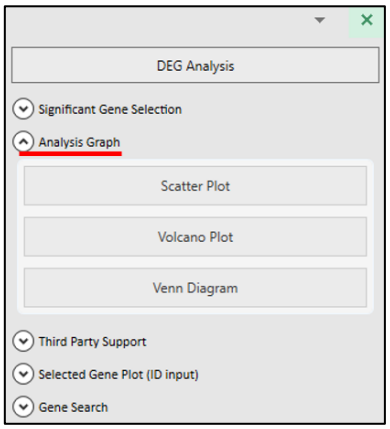


그림 1-11. Analysis Graph Tool

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. 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. 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 1-12).

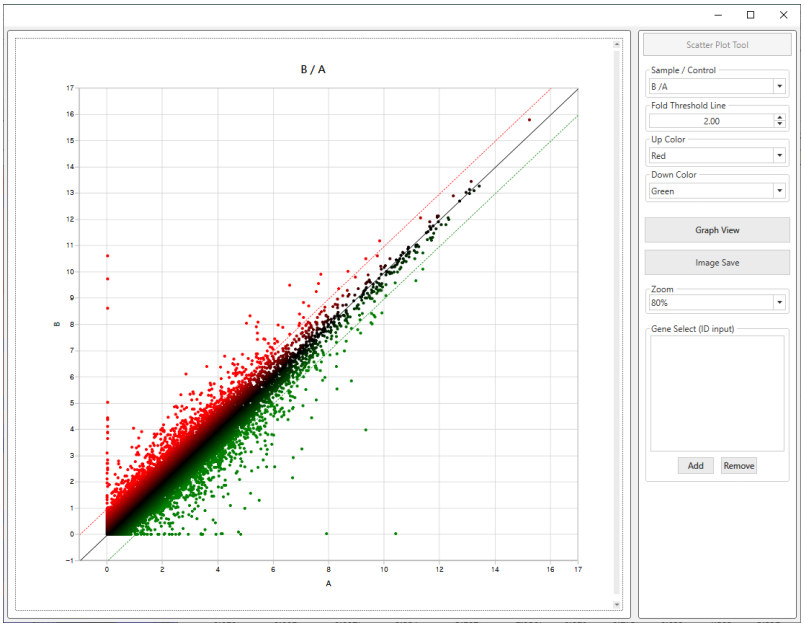


그림 1-12. Analysis Graph Tool – Scatter Plot

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. 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. 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).

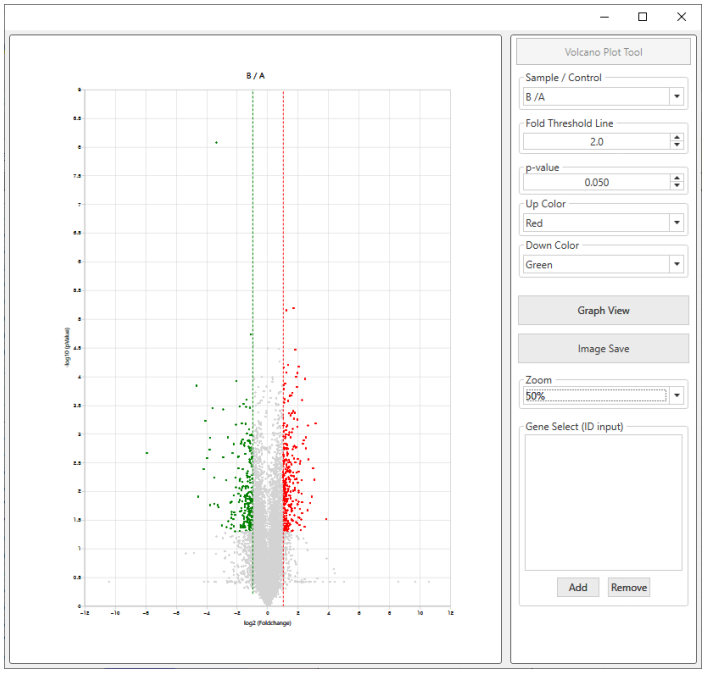


그림 1-13. Analysis Graph Tool – Volcano Plot

1-3-3. Venn diagram

Venn Diagram 을 통해 4 개 이하의 비교그룹을 대상으로 Venn Diagram 을 작성할 수 있다. Venn Diagram 을 그릴 샘플 비교그룹과 Fold Change, Normalized data (log2), p-value 을 선택 후, Diagram View 를 클릭하면 결과를 확인할 수 있으며 그룹은 최대 4 그룹까지 선택 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 1-14).

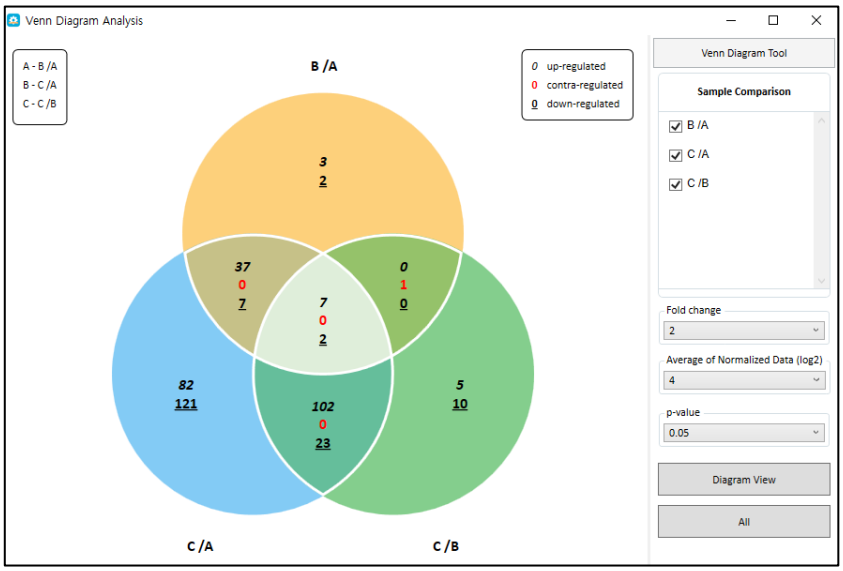


그림 1-14. Analysis Graph Tool – Venn Diagram

Venn Diagram 결과에서 표시되는 형식은 다음과 같다(그림 1-15).

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

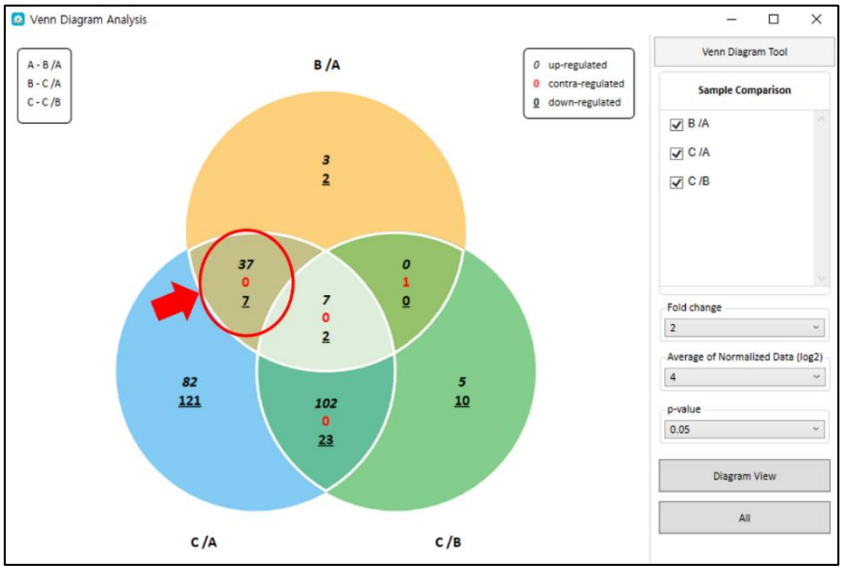


그림 1-15. For example of up, down, contra-regulated in Venn Diagram

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 유전자 list 3 개가 엑셀 data sheet 에 filter 된다(그림 1-16).

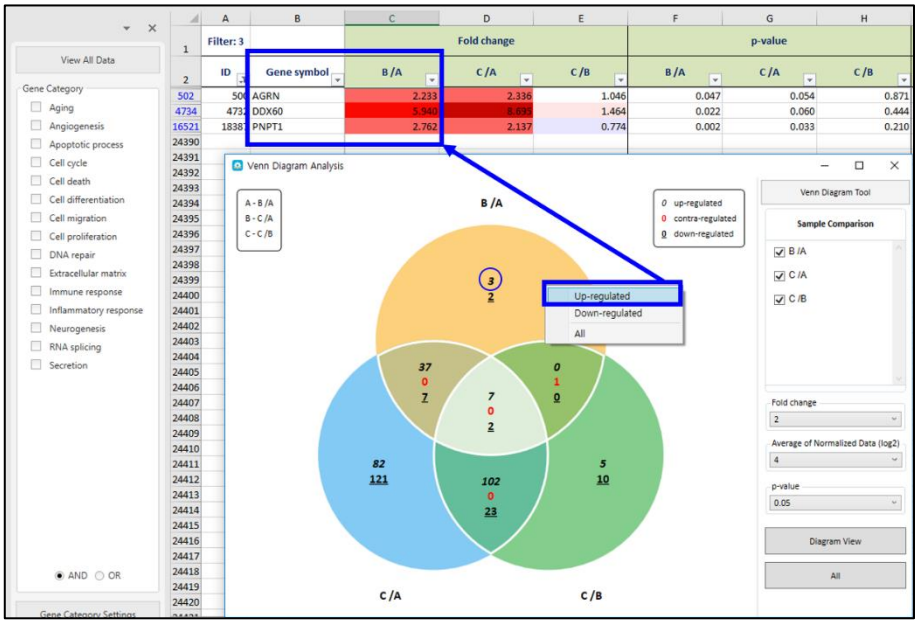


그림 1-16. Filtering 2fold up-regulated gene list in Venn Diagram

ExDEGA 에서 제공되는 모든 이미지는 오른쪽마우스를 눌러 'Save image' 버튼을 통해 저장이 가능하다(그림 1-17).

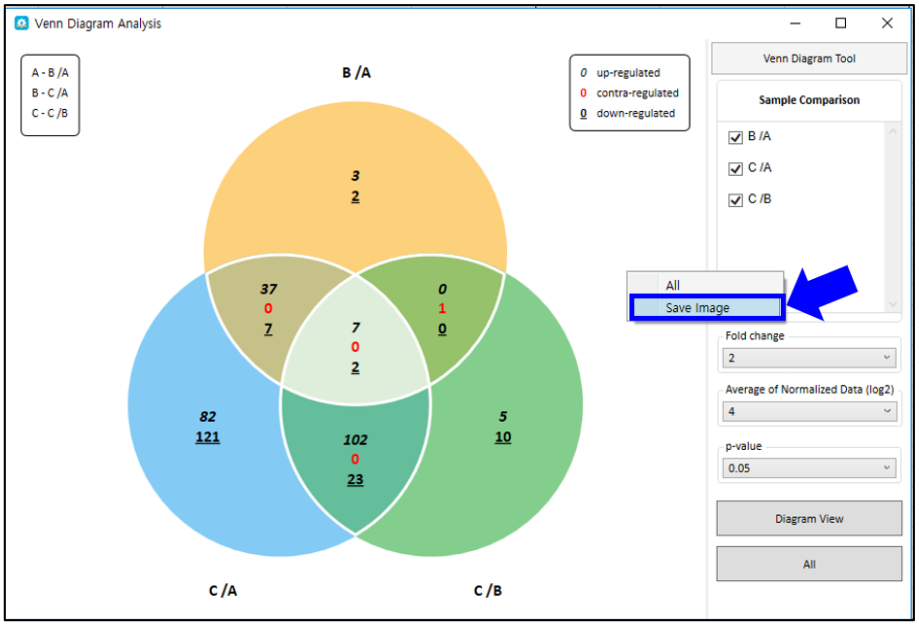


그림 1-17. Save image

1-4. Third Party Support 사용 방법

Third Party Support 는 연구자가 선택한 유전자를 기반으로 Clustering heatmap 과 KEGG 분석, DAVID 분석을 수행하기 위한 입력 데이터를 제공한다. 먼저, Input File 제작에 앞서 유전자를 선별하는 것이 필요하다. 유전자 선별은 유의성 있는 DEG 분석, Gene Ontology 분석 등으로 선별할 수 있다 (그림 1-18).

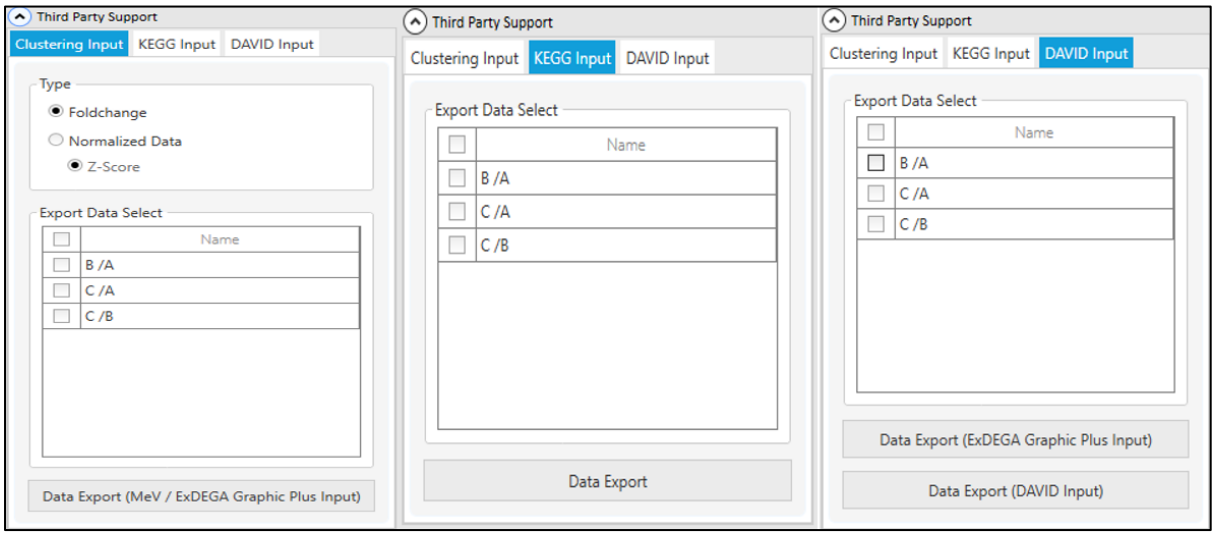


그림 1-18. Third Party Support (왼편부터 Clustering Input, KEGG Input, DAVID Input)

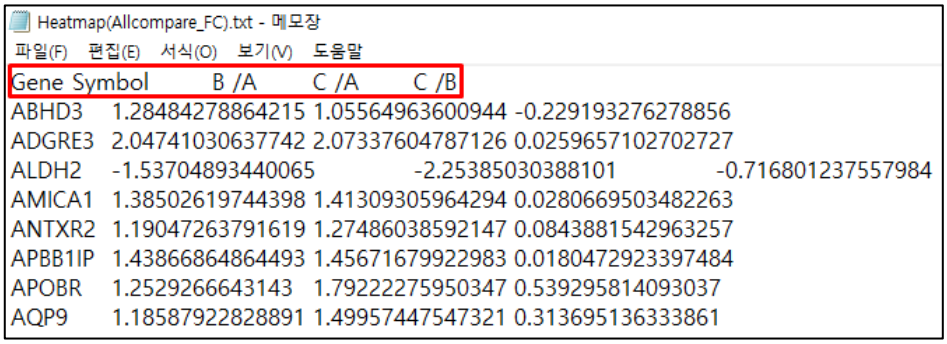
필터링 된 유전자 리스트를 대상으로 Clustering Heatmap 을 작성하려면 크게 두 종류의 데이터를 이용할 수 있다. Clustering Input 은 비교 조합의 Fold change 를 이용하여 제작된다. 원하는 비교 조합을 선택한 후 Data Export 버튼을 누르면 파일을 저장할 수 있다. 파일명에는 띄어쓰기가 들어가지 않도록 주의한다. Clustering Input 파일은 유전자 이름과 선택한 비교 조합, Z-score 로 구성된다. Clustering Input 파일은 **MeV** 또는 **ExDEGA Graphic Plus** 프로그램을 이용하여 Clustering 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 비교그룹을 체크한다. "Data Export"를 클릭한 후 "(input 명).txt"로 저장한다. Input file 명에는 띄어쓰기가 들어가지 않도록 주의한다.

두 번째, 개별 샘플의 발현 값인 Normalized Data 로 표현하고자 할 때 Z-Score 를 체크하고 확인하고자 하는 샘플을 체크한다. "Data Export"를 클릭한 후 "(input 명).txt"로 저장한다. 단, Z-score 로 그릴 때는 샘플 3 개 이상에서만 가능하다.

* Z-score 는 일반적으로 평균으로부터 얼마만큼 떨어져 있느냐를 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)

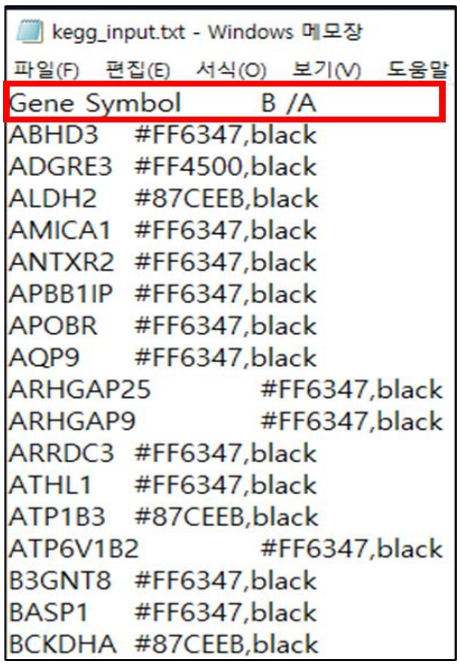
Clustering heatmap input file 은 Gene Symbol 과 fold change (log2) 또는 z-score 로 구성된다(그림 1-19). Clustering heatmap input file 을 ExDEGA Graphic plus tool 에 넣어 clustering heatmap 을 작성할 수 있다. Clustering heatmap 작성법은 본 매뉴얼의 3. Clustering heatmap analysis 부분에 설명되어 있다.


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 1-19. Clustering heatmap input file

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 1-20 과 같이 유전자 이름과 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.


kegg_input.txt - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말

Gene Symbol	B /A
ABHD3	#FF6347,black
ADGRE3	#FF4500,black
ALDH2	#87CEEB,black
AMICA1	#FF6347,black
ANTXR2	#FF6347,black
APBB1IP	#FF6347,black
APOBR	#FF6347,black
AQP9	#FF6347,black
ARHGAP25	#FF6347,black
ARHGAP9	#FF6347,black
ARRDC3	#FF6347,black
ATHL1	#FF6347,black
ATP1B3	#87CEEB,black
ATP6V1B2	#FF6347,black
B3GNT8	#FF6347,black
BASP1	#FF6347,black
BCKDHA	#87CEEB,black

그림 1-20. KEGG input file

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. DAVID input 은 크게 Data Export (ExDEGA Graphic Plus Input)와 Data Export (DAVID Input)로 구성된다. Data Export (DAVID Input)는 DAVID 에서 유전자 이름을 입력하는 부분에 사용되는 파일이다. 이 파일은 유전자 이름으로 구성된다 (그림 1-21). 유전자 이름을 DAVID 의 입력 데이터로 사용하여 분석하려는 GO 또는 Pathway 에 대한 데이터를 다운로드 받는다. Data Export (ExDEGA Graphic Plus Input)는 **ExDEGA Graphic Plus** 프로그램에서 DAVID 분석을 수행하기 위하여 사용되는 파일을 구축한다. 이 파일은 유전자 이름과 비교 조합, Fold change 로 구성된다 (그림 1-22).

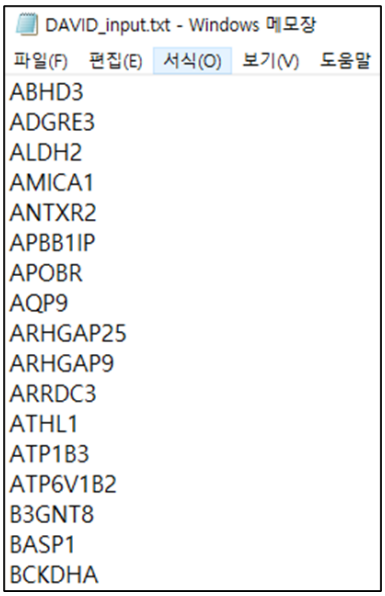


그림 1-21. DAVID input file

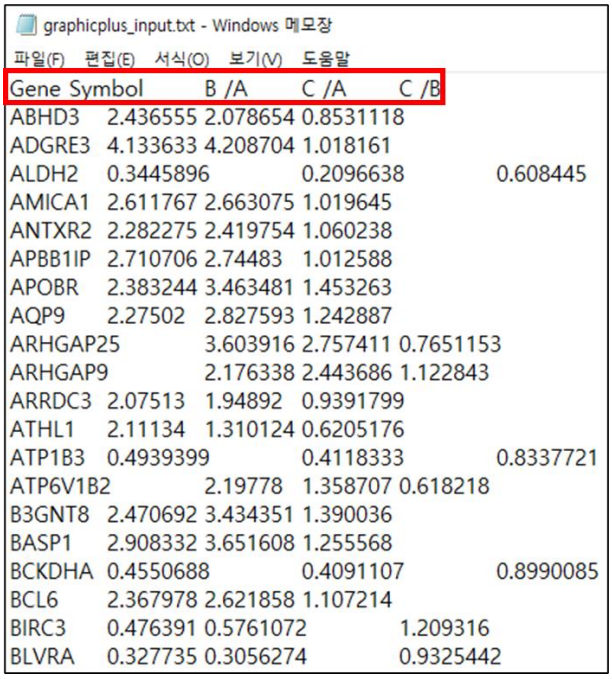


그림 1-22. Graphic Plus input file

1-5. Selected Gene Plot & Gene Search 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 1-23).

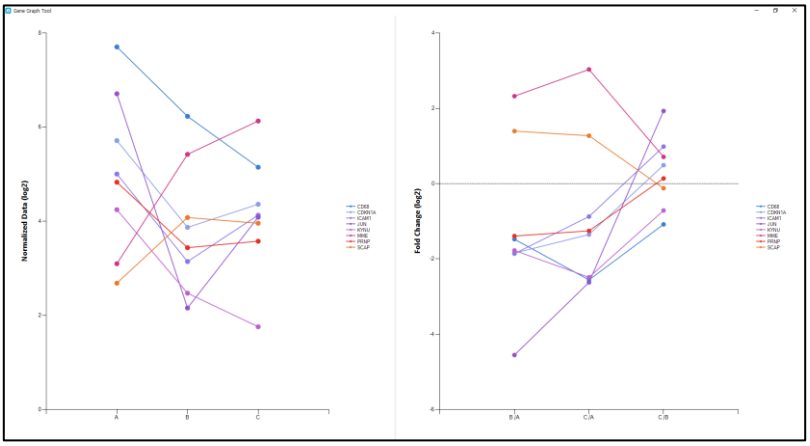


그림 1-23. Gene graph

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 1-24).

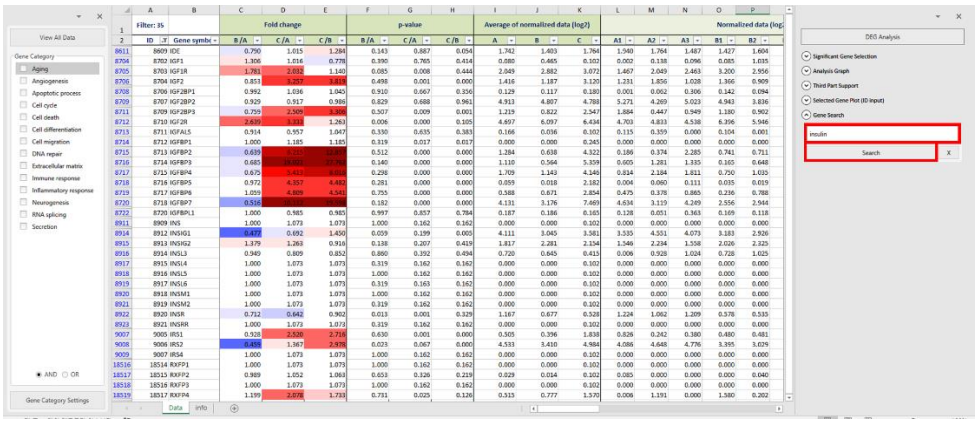


그림 1-24. Genes related to insulin

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.

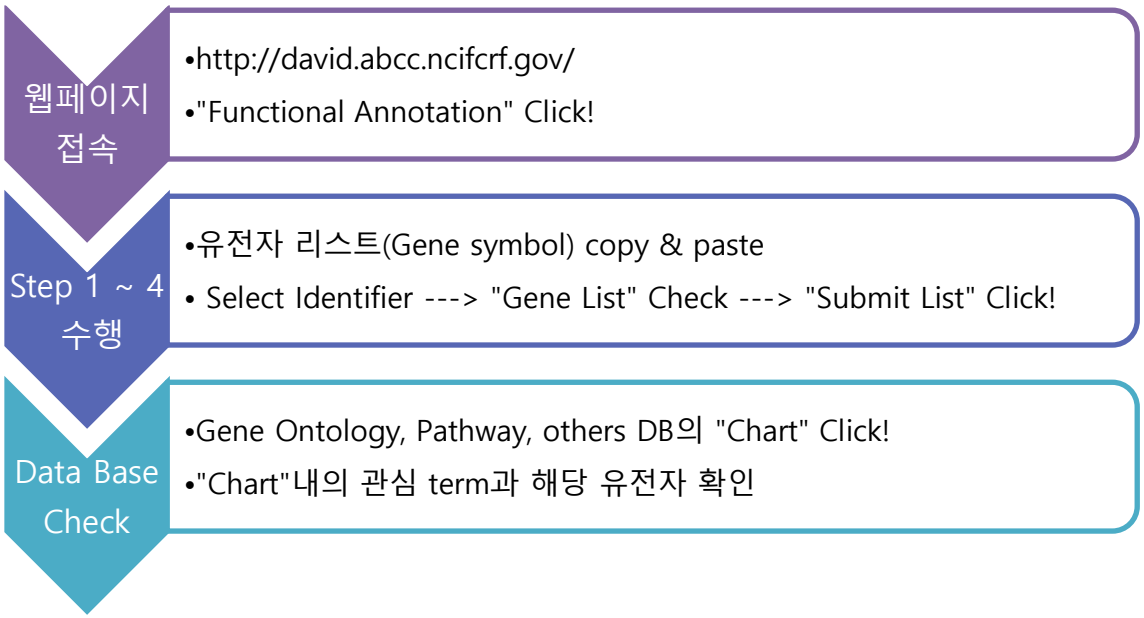


그림 2-1. DAVID tool analysis process

DAVID 에서는 3 천 개 이상의 유전자는 분석할 수 없으므로 3 천 개 이하로 유전자를 선별해야 한다. RNA-Seq 결과에서 significant gene 을 선별하여 DAVID 분석을 한다. DAVID 홈페이지 (http://david.abcc.ncifcrf.gov/)에 접속하여 "Functional Annotation"을 클릭한다(그림 2-2).

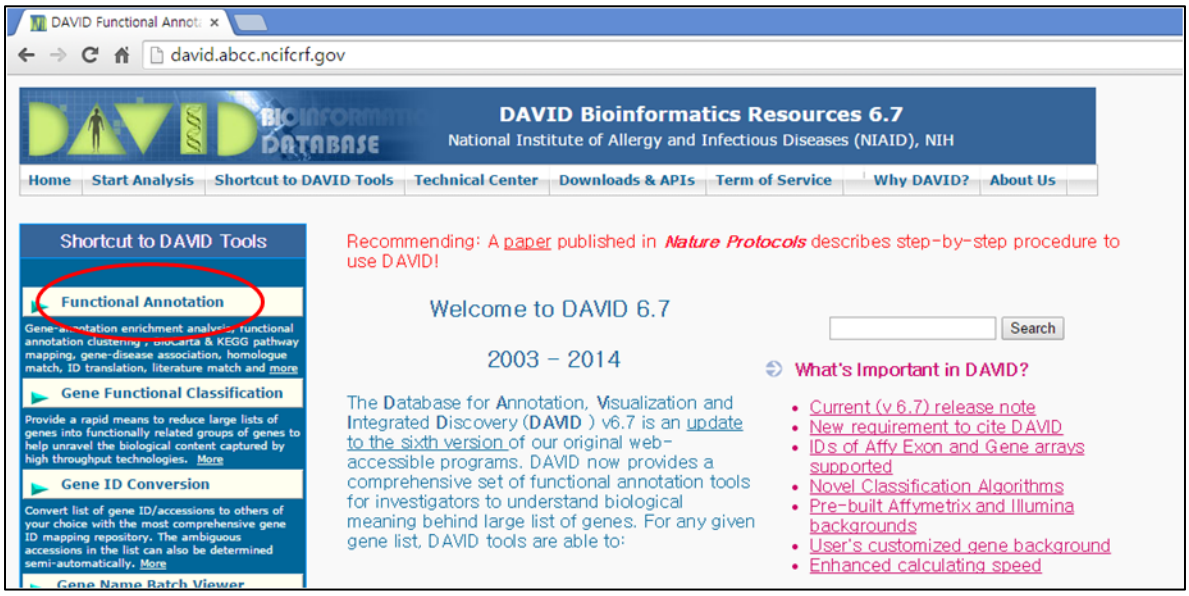


그림 2-2. DAVID tool webpage

"Upload" 탭에서 Step 1 에서 Step 4 까지 수행한다(그림 2-3). Step 1 에서 선별한 유전자의 Gene Symbol 을 복사하고 "A: Paste a list" 창에 붙여 넣는다. Step 2 에서"OFFICIAL_GENE_SYMBOL"를 선택한다. 만약 step 1 에서 Gene Bank No.를 넣었다면 "GENEBANK_ACCESSION" 을 선택한다. Step 3 에서 "Gene List"를 체크하고 Step 4 에서 "Submit List"를 누른다. Gene Symbol 을 넣은 경우, "multiple species have been detected in your gene list"라는 창이 뜨면"확인"을 누른다.

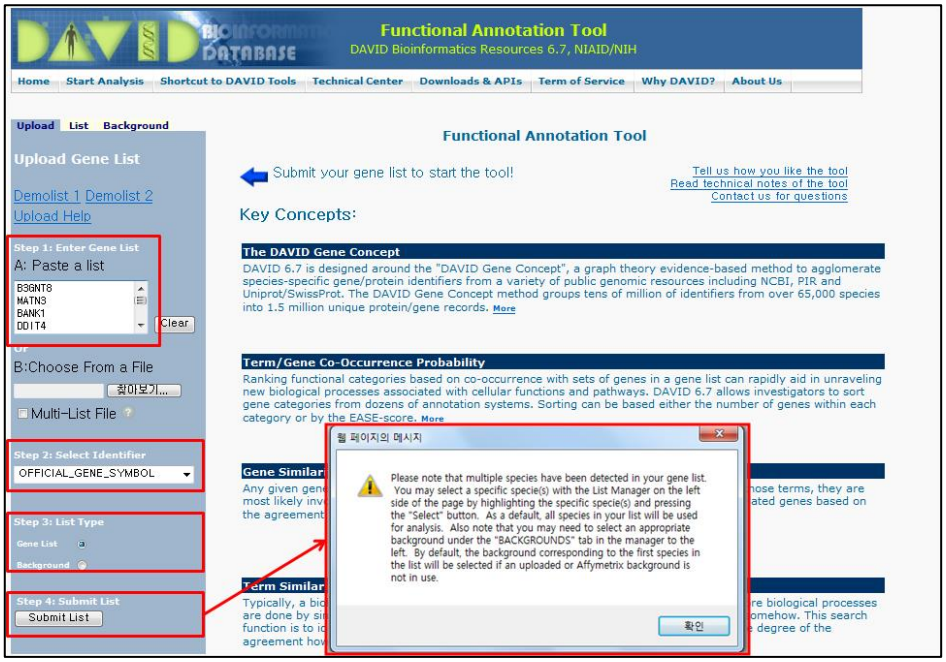


그림 2-3. DAVID tool : Step 1 ~ Step 4

List sheet 에서 분석하고자 하는 종을 선택한다(그림 2-4. a). "List" Sheet 에서 해당 종(숫자)로 표기되어 있고 가로 안의 숫자가 분석에 적용된 유전자의 개수이다. 예시에서는 59 개의 유전자 리스트를 넣었고 데이터베이스에서 기능이 밝혀진 48 개만이 Functional Annotation 분석에 이용되었다는 의미이다. Current Background 에 분석하고자 하는 종이 아닌 다른 종이 나왔다면 좌측 "Background" Sheet 에서 알맞은 종을 선택하여 "Use"를 클릭한다(그림 2-4. b).

a.

b.

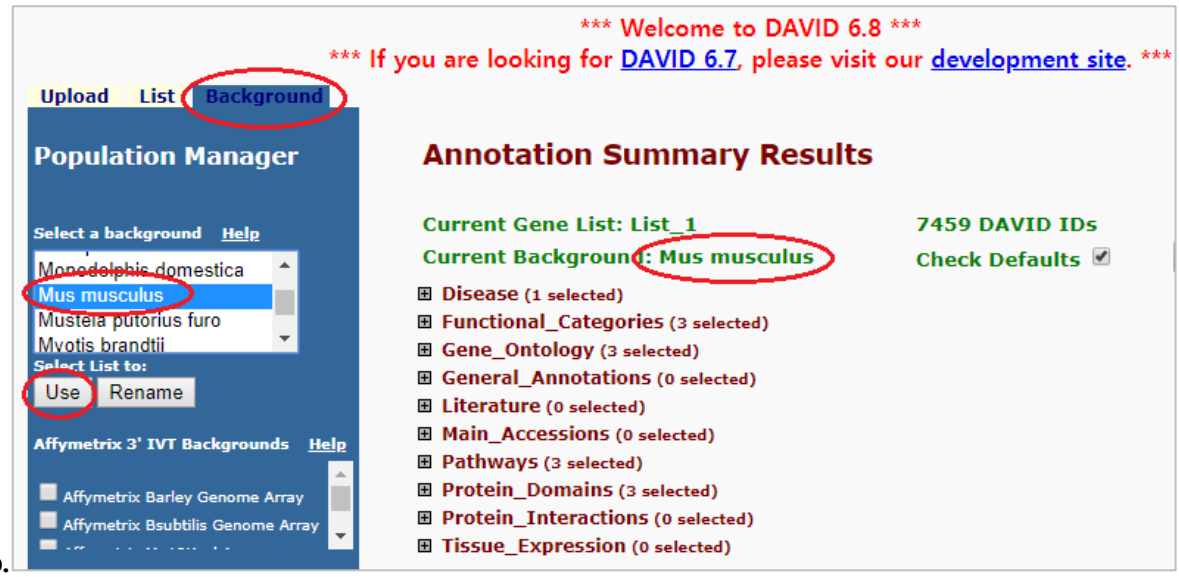


그림 2-4. DAVID tool : Select Species

DAVID 분석 결과 중 Gene Ontology Biological Process 결과를 확인하려면 "Gene_Ontology"의 "+" 표시를 클릭하여 결과 창을 열고 "GOTERM_BP_FAT"의 "Chart"를 누른다(그림 2-5). 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.

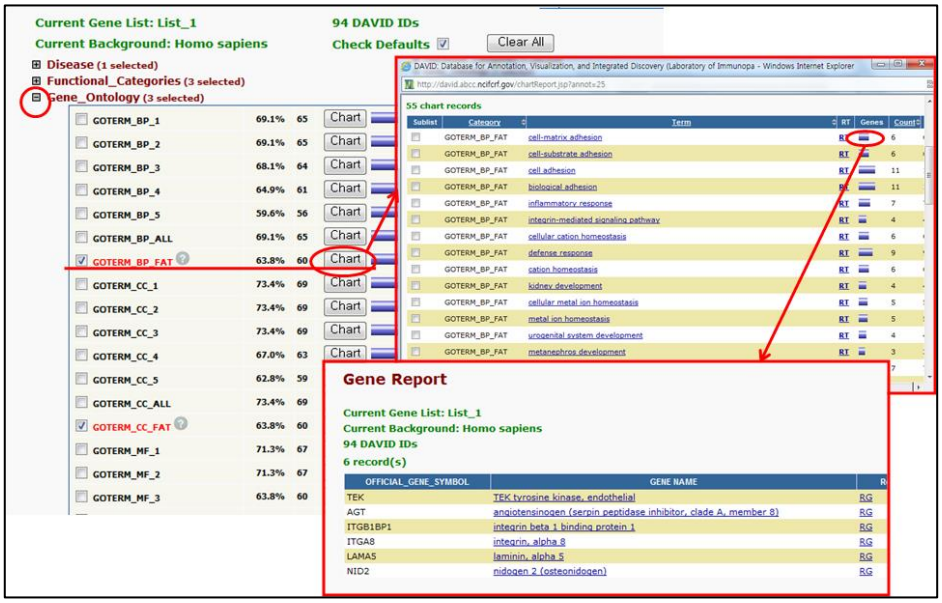


그림 2-5. DAVID tool : exploring Gene Ontology analysis result

이와 같은 방법으로 Pathway 결과를 확인해 보면 KEGG_PATHWAY database 에서 주요 Pathway 가 나온다(그림 2-6). 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.

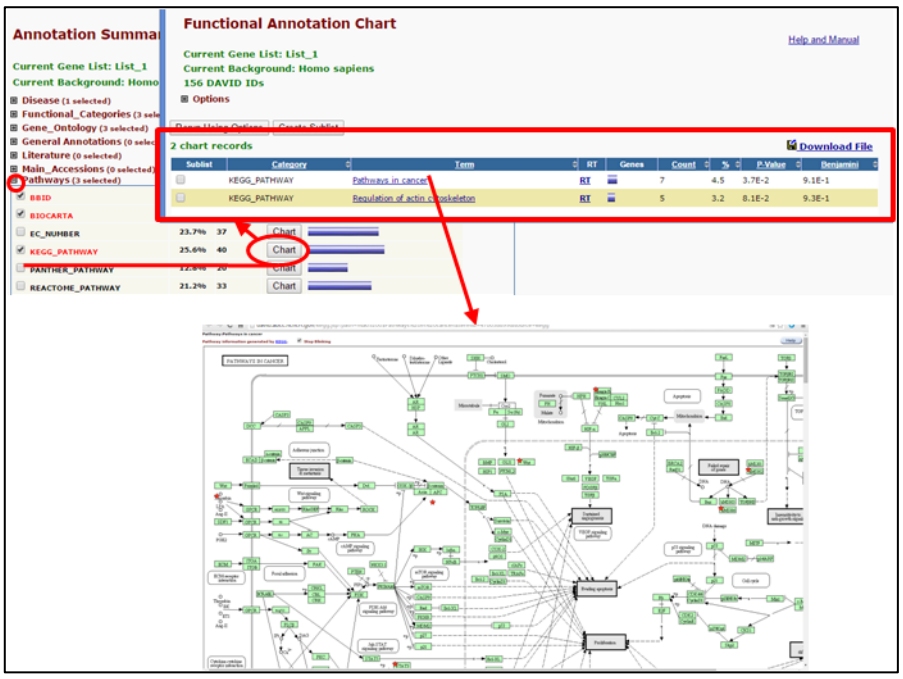


그림 2-6. DAVID tool : exploring Pathway analysis result

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 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 2-7).

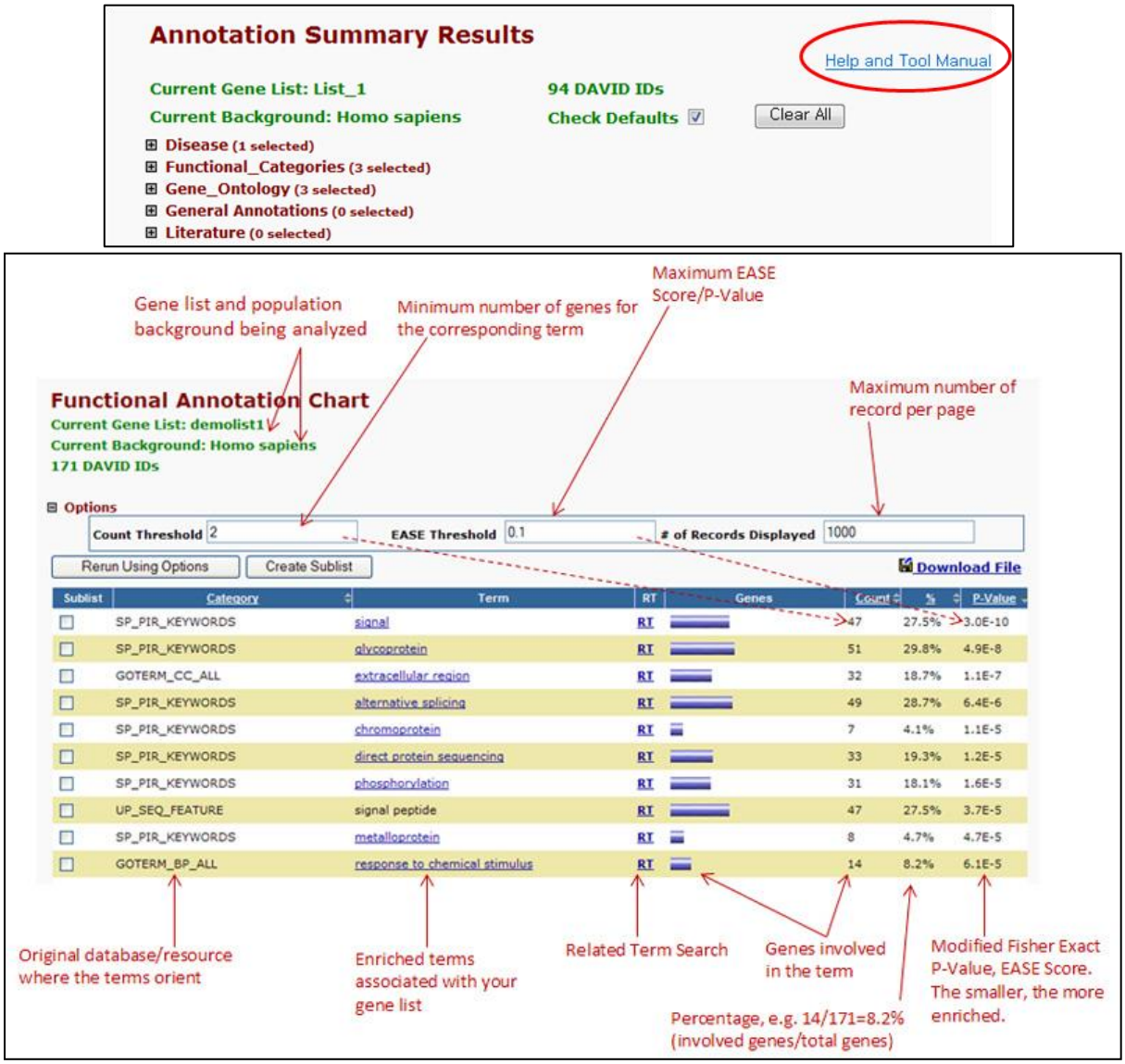


그림 2-7. DAVID Help and Tool Manual

DAVID 분석 결과를 내 컴퓨터에 저장하려면, 초록색으로 표시된 Download File 링크를 마우스 오른쪽 버튼을 클릭한 후 다른 이름으로 저장을 선택하면 파일을 다운로드 받을 수 있다(그림 2-8). 다운로드 받은 DAVID 결과 파일로 그래프 작성하는 방법은 '2-2. ExDEGA Graphic+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.

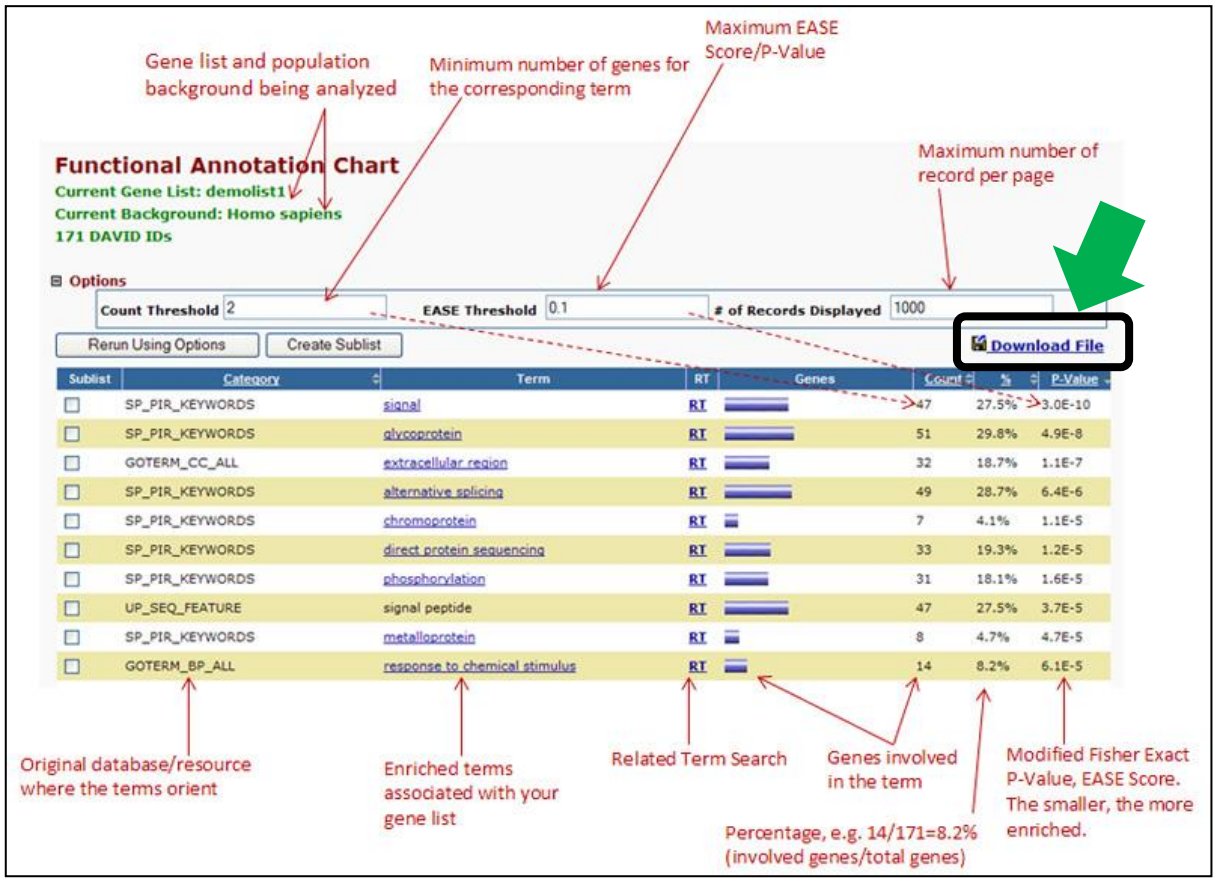


그림 2-8. DAVID data download

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

ExDEGA graphic plus을 클릭하여 프로그램을 실행한다(그림2-9).

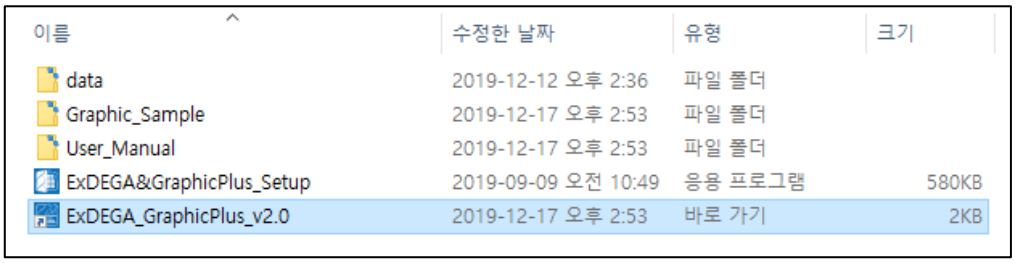


그림 2-9. ExDEGA GraphicPlus

먼저 ExDEGA import 항목에 분석하고자 하는 ExDEGA format의 엑셀 파일(RNA-seq report 또는 microarray report excel file)을 import 시킨다(그림2-10). Import가 완료되면 Graphic Tools에 해당하는 3개의 버튼이 활성화가 되는데, DAVID 분석을 위해 첫 번째 DAVID Graphic 버튼을 클릭한다.

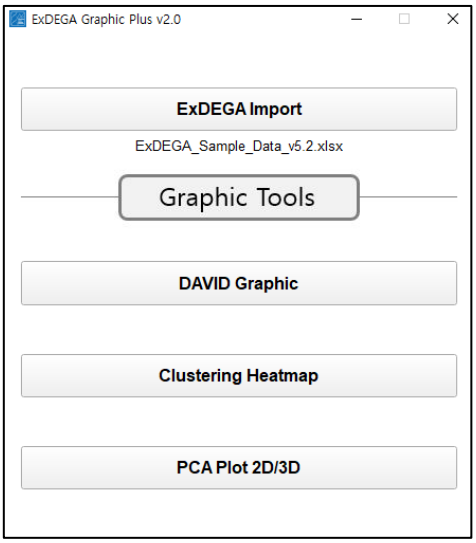


그림 2-10. Import ExDEGA Report to ExDEGA GraphicPlus

DAVID_Graphic 분석 창은 그림 2-11 과 같다.

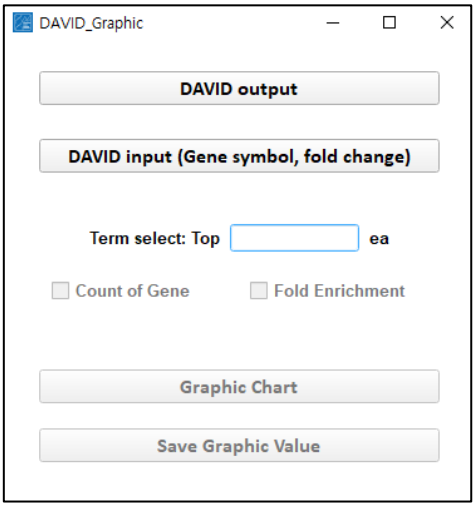


그림 2-11. ExDEGA GraphicPlus - DAVID Graphic

DAVID 분석 결과를 그래프로 작성하려면 2 가지 파일(DAVID output, input)이 필요하다.

DAVID output 은 '2-1. DAVID 분석 툴을 이용한 Functional Annotation 분석'에서 저장한 DAVID 결과 파일이다. DAVID 결과 파일 안에 있는 내용은 그림 2-12 과 같다. 이 파일에 있는 Term, Count, P-value, Fold Enrichment 항목을 이용하여 그래프가 작성된다.

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, 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 2-12. DAVID output file

DAVID input 은 DAVID 분석 시 입력한 유전자들의 Gene Symbol 과 Fold Change 값이 들어있는 파일을 말한다. DAVID input 파일은 그림 2-13 과 같이 만든다. RNA-seq report 파일에서 DAVID 분석한 유전자의 gene symbol 과 fold change 값을 복사하여 새 엑셀 파일에 붙여넣기 한다. 기존 파일에 sheet 를 추가하는 것이 아니고 엑셀을 새로 만들기 하여 새 파일을 만드는 것이다. DAVID input 파일을 저장할 때는 꼭 '텍스트 (탭으로 분리)' 형식으로 저장해야 한다. 파일 이름에는 띄어쓰기가 들어가지 않도록 한다.

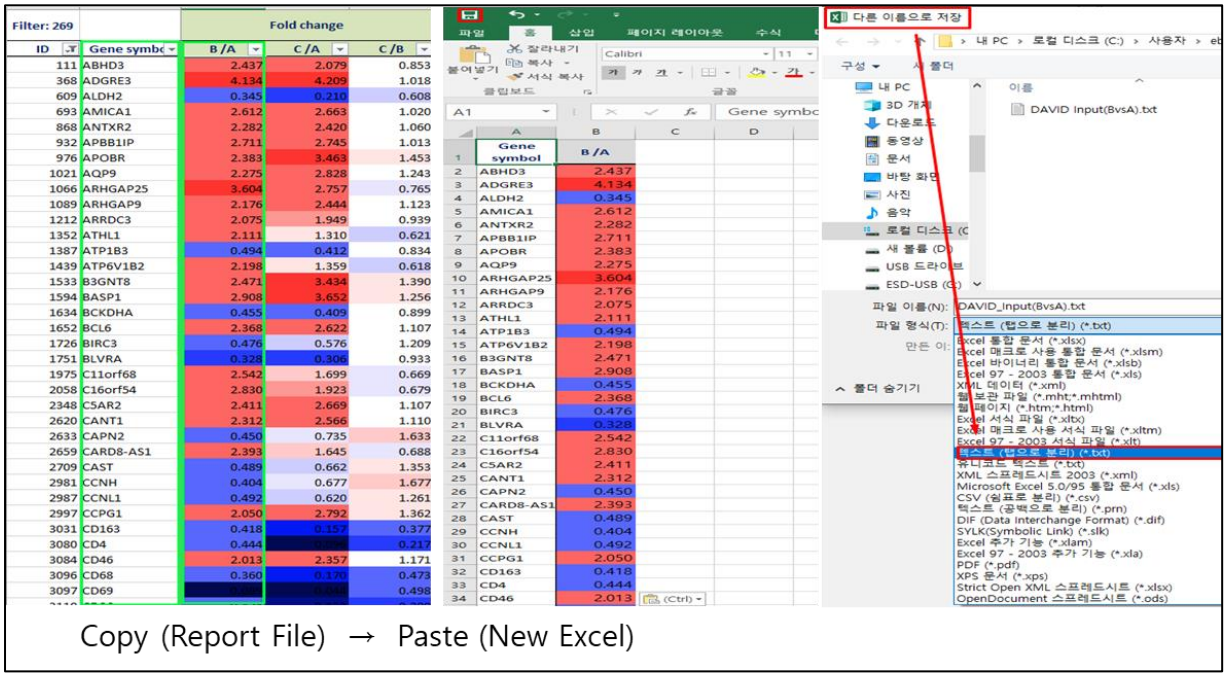


그림 2-13. DAVID input file

DAVID output 버튼을 눌러 DAVID 결과 파일을 열고, DAVID input (Gene symbol, fold change) 버튼을 눌러 DAVID input 파일을 연다(그림 2-14).

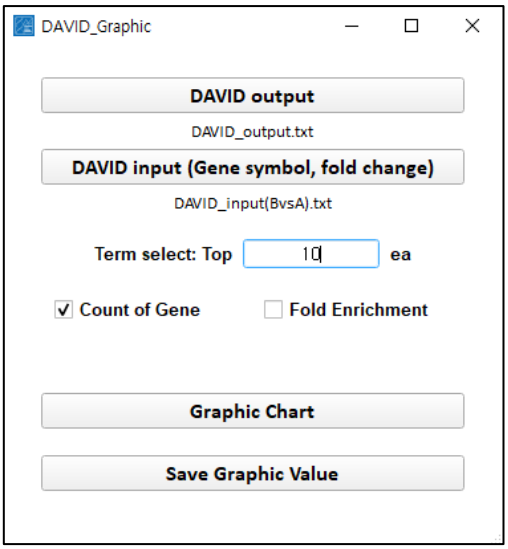


그림 2-14. DAVID Graphic options

Term select는 DAVID 분석 결과에서 p-value 순으로 상위 몇 개를 그래프로 작성할지에 대한 옵션이다. Count of Gene을 선택하고 Graphic Chart를 누르면 각 GO에서 발현이 증가하는 유전자, 감소하는 유전자 수가 그래프로 작성된다(그림 2-15).

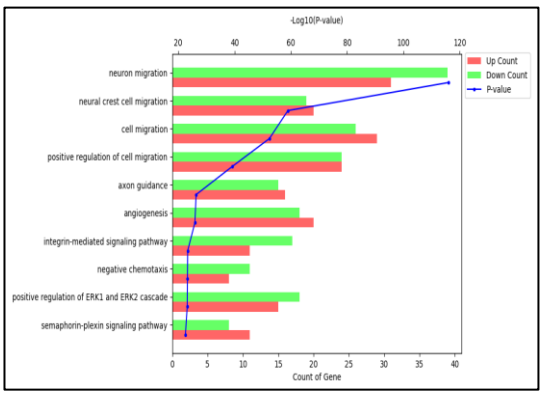


그림 2-15. Result of Graphic Chart (Count of Gene)

Fold Enrichment를 선택하고 Graphic Chart를 누르면 각 GO의 p-value, fold enrichment 값으로 그래프가 작성된다(그림2-16).

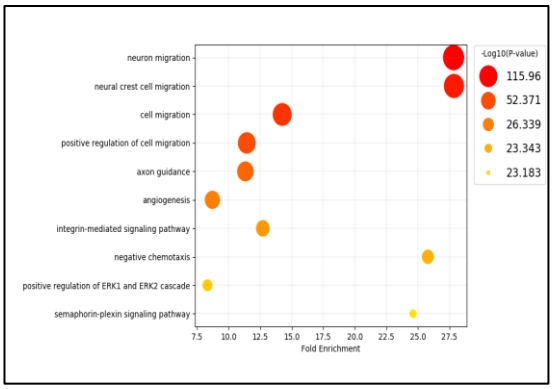


그림 2-16. Result of Graphic Chart (Fold Enrichment)

DAVID_Graphic 창에서 Save Graphic Value를 누르면 그래프에 사용된 값들이 Excel로 저장된다.

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 을 이용하여 분석할 수 있다. ExDEGA Graphic plus 를 이용하기 위해서는 먼저 해당 프로그램을 열고, ExDEGA import 항목에 분석하고자 하는 ExDEGA 엑셀 레포트 파일을 import 시킨다(그림 2-10). Import 가 완료되면 Graphic Tools 에 해당하는 3 개의 버튼이 활성화가 되는데, 2 번째 Clustering heatmap 버튼을 클릭한다.

Clustering Heatmap 창에서 Hierarchical Clustering Input 버튼을 눌러 clustering heatmap input file을 연다(그림3-1). clustering heatmap input file을 만드는 방법은 본 매뉴얼의 '**1-4. Clustering Heatmap Support 사용 방법**'에 설명되어 있다.

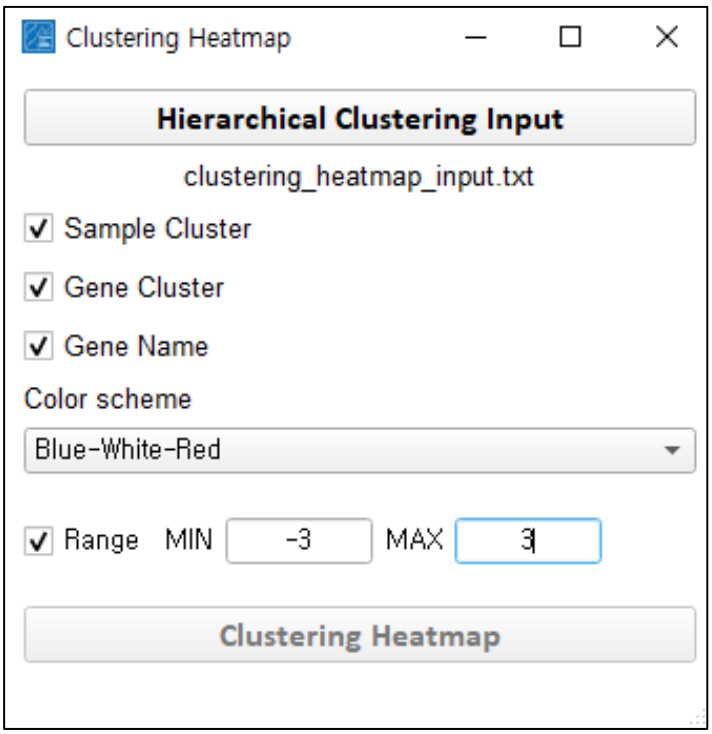


그림 3-1. Data export from Clustering heatmap support

Sample cluster를 선택하면 발현이 유사한 비교그룹 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. Color scheme는 Heatmap의 색상을 설정하는 옵션이다. Range는 데이터 표현 범위를 설정하는 것으로 선택하지 않으면 input file의 최소값과 최대값으로 자동 설정된다. 선택하면 원하는 최소값, 최대값을 설정할 수 있다.

Clustering Heatmap 버튼을 누르면 Hierarchical Clustering Heatmap 결과 창이 생성된다(그림3-2). 위쪽에 표시된 dendrogram은 비교조합 또는 샘플 간의 발현 유사성을 표시한 결과이다. 왼쪽에 표시된 dendrogram은 유전자 간의 발현 유사성을 표시한 결과이다. 가깝게 묶일수록 발현이 유사한 것이다.

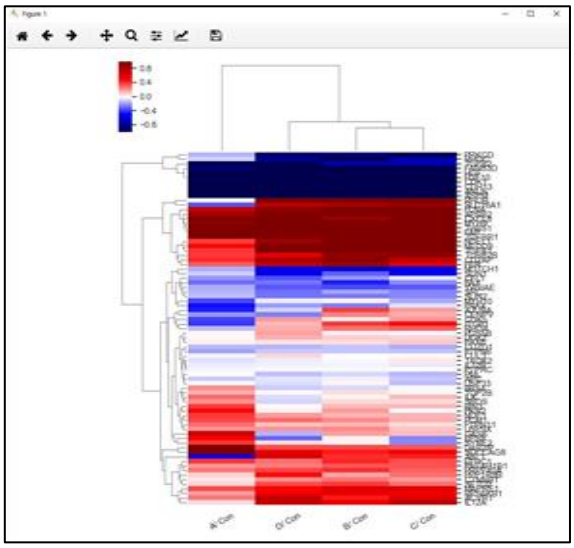


그림 3-2. Clustering heatmap result

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

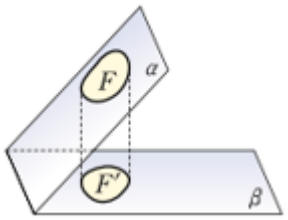
4. Principal component analysis (ExDEGA GraphicPlus)

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

먼저 ExDEGA GraphicPlus를 이용하려면, ExDEGA GraphicPlus의 ExDEGA import 항목에 분석하고자 하는 ExDEGA format의 엑셀 파일(RNA-seq report 또는 microarray report excel file)을 import 시킨다(그림2-10). Import가 완료되면 Graphic Tools에 해당하는 3개의 버튼이 활성화가 되는데, PCA 분석을 위해 세 번째 PCA Plot 2D/3D 버튼을 클릭한다.

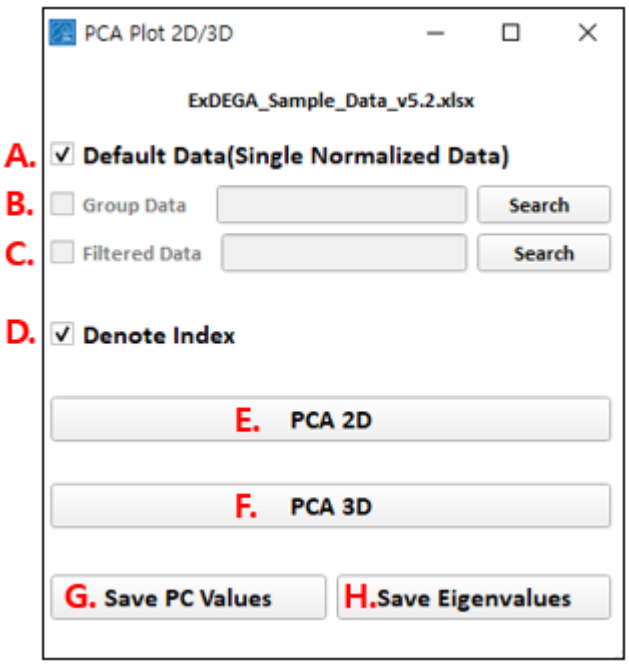


그림 4-1. PCA plot 2D/3D in ExDEGA graphic plus

PCA 분석 창은 그림 4-1 과 같다. PCA 분석 시 3 개 data (A, B, C) 중 하나를 선택하여 분석할 수 있다.

Default Data(Single Normalized Data) (A)를 선택하면 초기화면에서 입력한 Report 파일의 Normalized Data 가 자동으로 입력된다. * Array Data 의 경우 해당 옵션을 이용하실 수 없다. B 혹은 C 와 같이 input 파일을 별도로 만들어 이용해야 한다.

Group Data (B)는 반복실험 결과(N>=2)로 PCA 분석을 할 때 사용한다. Search 버튼을 눌러 반복실험 결과로 만든 PCA input 파일을 load 하여 해당 파일로 분석한다. B. Group Data 로 PCA plot 을 작성하면 샘플들의 색상을 Group 별로 구분하여 PCA 를 그릴 수 있다.

Filtered Data (C)는 반복실험 하지 않은 실험 결과(N=1)로 PCA 분석을 할 때 사용한다. Search 버튼을 눌러 반복실험 하지 않은 실험 결과로 만든 PCA input 파일을 load 하여 해당 파일로 분석한다.

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

Denote Index (D)를 체크하면 PCA plot 에 Sample 의 인덱스(샘플 순서대로 부여하는 숫자)를 표시한다.

PCA 2D 버튼(E)을 누르면 PCA 2D (2 차원 평면) 분석 결과가 나오고 PCA 3D 버튼(F)을 누르면 PCA 3D (3 차원 공간) 분석 결과가 나온다(그림 4-2). 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.

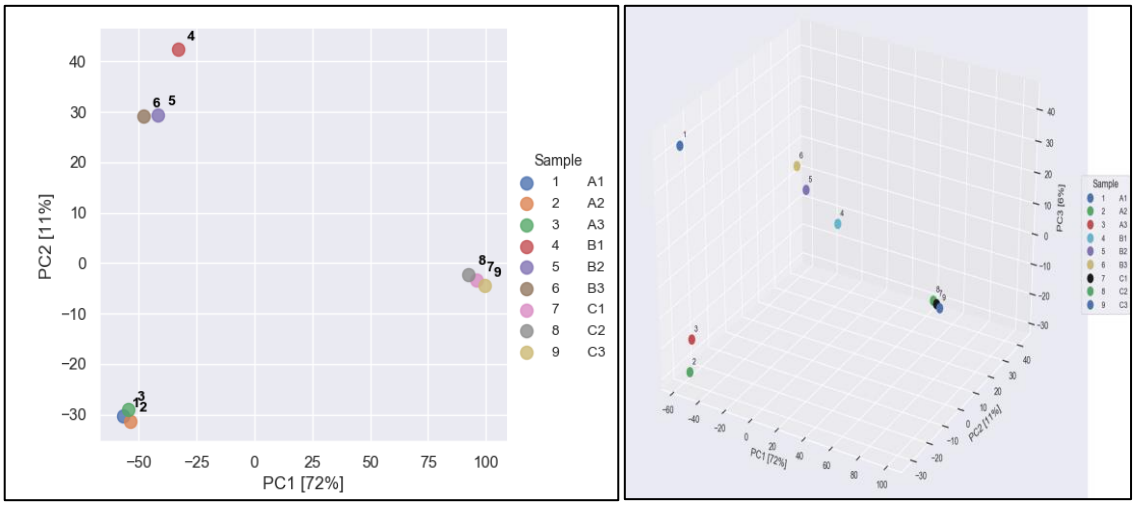


그림 4-2. PCA results 2D, 3D

Save PC Values 버튼(G)을 누르면 각 주성분(PC)이 전체 데이터를 얼마만큼 설명할 수 있는 지 분산 비율로 계산되어 Excel 파일 형식으로 저장된다.

Save Eigenvalues 버튼(H)를 누르면 계산된 Eigenvalue 의 결과를 Excel 파일 형식으로 저장한다. Eigen values 가 PCA plot 에서 각 샘플의 좌표이다. 이 값으로 엑셀에서 분산형 차트를 그려 직접 PCA 2D 를 작성할 수도 있다(그림 4-3).

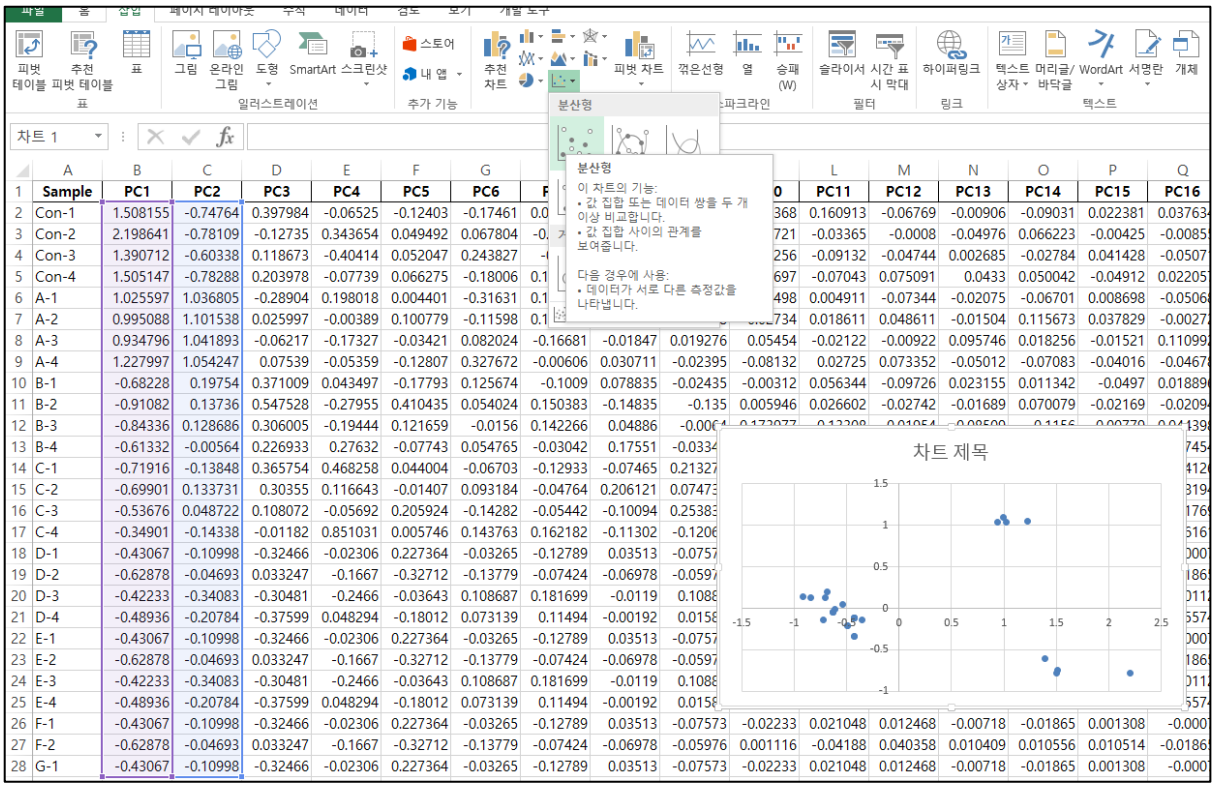


그림 4-3. PCA plot in Excel

4-3. PCA Plot input 파일 작성방법

Group Data

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 4-4).

	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 4-4. Copy normalized data

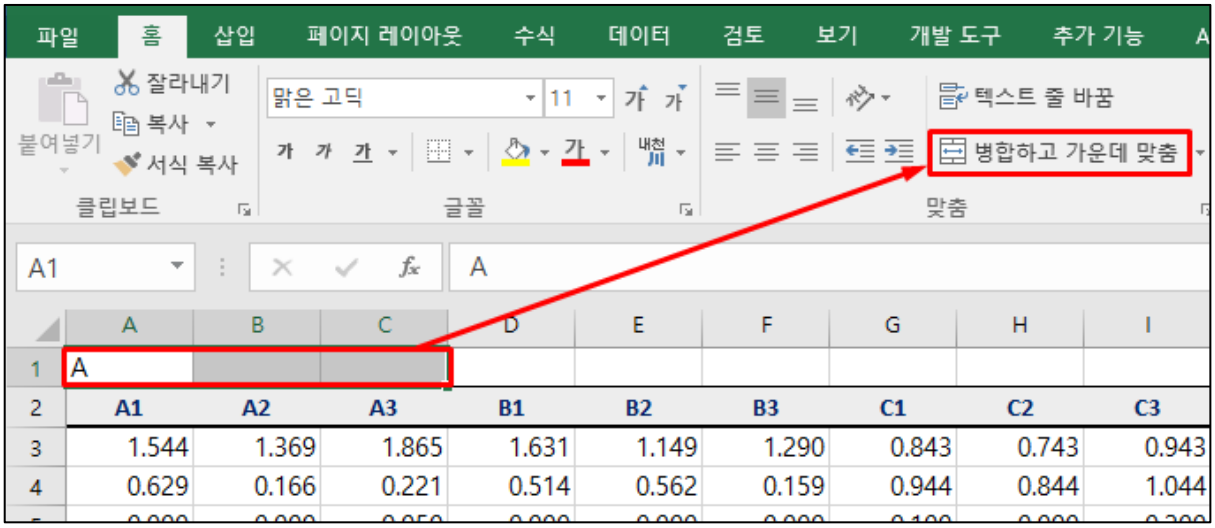
참고) 전체 데이터 선택 방법은 A1(초록색 네모) 클릭 후 Shift 를 누른 상태에서 C3(파란색 네모)를 클릭한다. A1~ C3 까지 선택 후에 Ctrl + Shift + 아래 방향키(↓)를 입력하면 전체 데이터가 한번에 선택된다.

2) 새 엑셀 파일(기존 파일에 sheet 추가가 아닌 엑셀 새로 만들기)을 열어서 **2 번째** 열에 붙여넣기한다(그림 4-5).

	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 4-5. Paste normalized data in new excel file

3) 1 번째 열에는 그룹명 입력 후 병합한다(그림 4-6).



	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 4-6. Group information

4) 완성 후 파일형식은 엑셀로 저장한다(그림 4-7). 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 4-7. Save PCA input (group data) file

Filtered Data

<작성 순서>

1) Normalized data(log2)항목에서 데이터 복사한다(그림 4-8).

Normalized data (log2)					
A	B	C	D	E	F
1.080	1.128	1.000	1.000	1.074	1.132
11.432	16.763	11.978	14.201	11.178	13.520
36.997	19.871	26.910	47.822	33.745	37.776
7.510	5.007	3.489	6.181	6.950	6.165
12.164	9.640	10.508	9.975	12.335	12.184
1.221	1.599	1.353	1.989	1.383	1.449
12.262	6.081	7.191	8.172	9.814	11.692
22.889	33.752	35.366	52.206	34.648	28.727
17.898	17.651	17.612	27.822	21.294	22.199
22.061	17.520	27.063	40.521	29.062	25.247
9.127	14.030	10.192	5.724	7.576	8.077
1.000	1.000	1.000	1.101	1.000	1.000

그림 4-8. Copy normalized data

2) 새 엑셀 파일(기존 파일에 sheet 추가가 아닌 엑셀 새로 만들기)을 열어서 **2 번째** 열에 붙여넣기 한다(그림 4-9). 완성 후 파일형식은 엑셀로 저장한다. Input file 명에는 띄어쓰기가 들어가지 않도록 주의한다.

	A	B	C	D	E	F	G	H	I
1	A1	A2	A3	B1	B2	B3	C1	C2	C3
2	1.544	1.369	1.865	1.631	1.149	1.290	0.843	0.743	0.943
3	0.629	0.166	0.221	0.514	0.562	0.159	0.944	0.844	1.044
4	0.000	0.000	0.050	0.000	0.000	0.000	0.100	0.000	0.200
5	0.231	0.330	0.100	0.642	0.091	0.071	0.328	0.228	0.428
6	0.303	0.799	0.111	1.771	0.917	0.294	0.923	0.823	1.023
7	0.000	0.000	0.063	0.055	0.000	0.088	0.100	0.000	0.200
8	0.000	0.290	0.000	0.000	0.098	0.000	0.100	0.000	0.200
9	0.440	0.002	0.000	0.001	0.001	0.862	0.100	0.000	0.200
10	0.218	0.564	0.000	0.001	0.002	0.000	1.742	1.642	1.842
11	0.000	0.000	0.266	0.000	0.000	0.000	0.100	0.000	0.200
12	0.406	0.001	0.000	0.000	0.000	0.000	0.100	0.000	0.200
13	3.136	2.775	2.732	2.813	3.270	3.323	2.474	2.374	2.574
14	1.510	1.210	0.904	1.276	1.172	0.962	2.241	2.141	2.341
15	0.000	0.000	0.000	0.098	0.000	0.157	0.100	0.000	0.200
16	0.000	0.000	0.000	0.000	0.000	0.000	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
20	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

5. Pathway analysis (KEGG mapper)

RNA-Seq 분석 결과에서 up/down-regulated genes들이 어떤 Pathway에 속하는지 확인하고자 한다면 KEGG에서 제공하는 KEGG Mapper를 이용한다. 사용방법은 그림 5-1과 같은 순서로 진행된다.

그림 5-1. KEGG Mapper tool analysis process

그림 5-2는 mRNA-Seq report에서 2fold, normalized data(log2)>4, p-value<0.05을 기준으로 선별한 유전자를 KEGG 분석하는 과정이다.

* KEGG input 값은 excel 파일의 Annotation 항목 앞에 제작되어 있다.

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 유전자의 KEGG input [Entrez ID, FC Color(#숫자,black)] 영역을 함께 복사하여, KEGG 분석에 사용할 것이다.

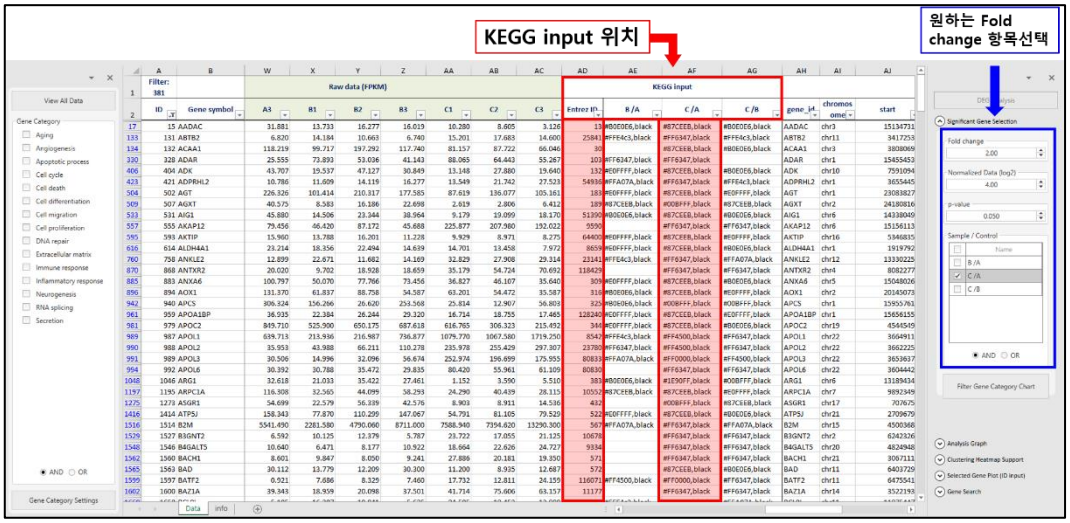


그림 5-2. KEGG Mapper tool analysis process

그림 5-3과 같이 KEGG Mapper 웹페이지(http://www.genome.jp/kegg/tool/map_pathway2.html)에 접속하고 Search & Color pathway 링크에 들어가면 아래와 같은 화면이 보여진다.

(1) 분석하고자 하는 유전자의 species를 선택

(2) 'Optional use of outside ID:'는 NCBI-GeneID로 선택

(3) 'Enter objects one per line followed bgcolor, fgcolor:' 창에 엑셀에서 준비해 놓은 Entrez ID, Color 항목을 복사-붙여넣기를 한다.

(4) "Include aliases"와 "Use uncolored diagram" 항목에 체크를 한 후

(5) Exec 버튼을 누른다.

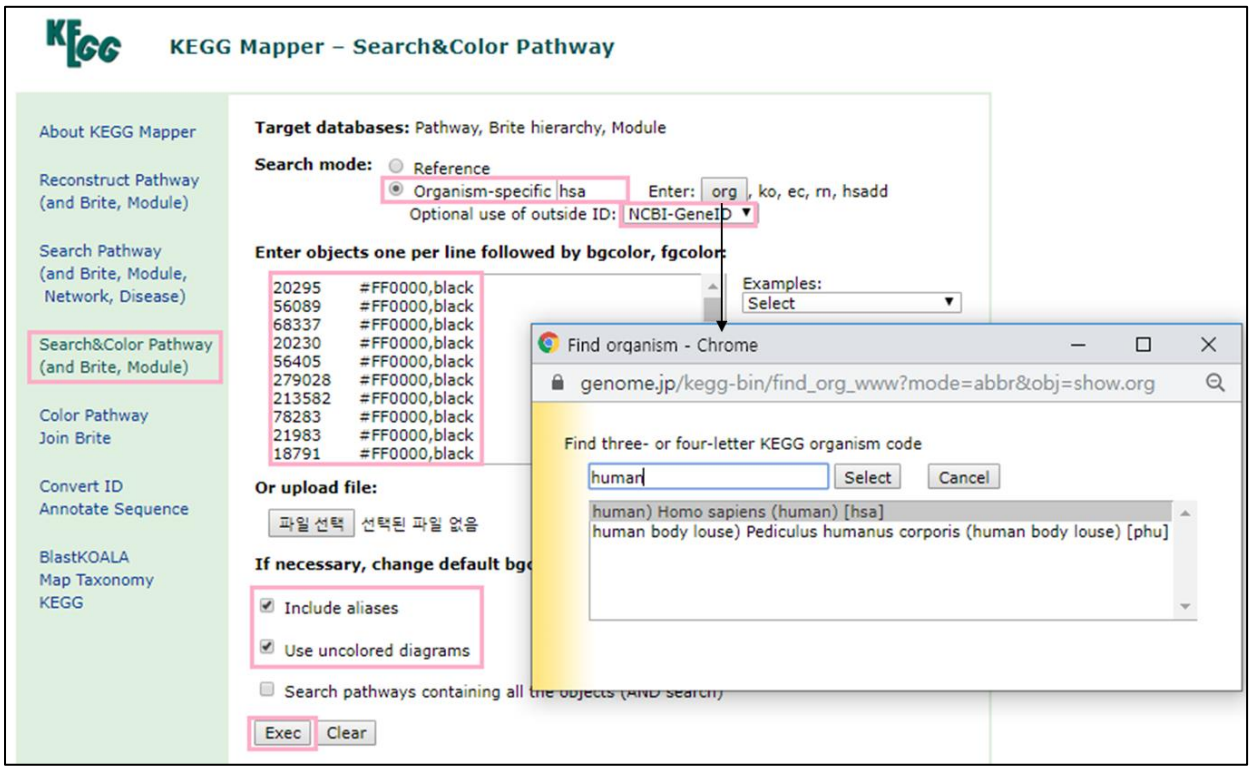


그림 5-3. KEGG Mapper tool analysis process

분석결과, 입력한 유전자들이 관여하는 pathway list가 나온다(그림 5-4). pathway 이름 옆에 있는 괄호 안 숫자는 입력한 유전자 중 각 pathway에 관여하는 유전자의 수이다. 괄호 안 숫자를 클릭하면 해당 유전자 목록을 볼 수 있다. pathway 이름을 클릭하면 해당 pathway chart가 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.

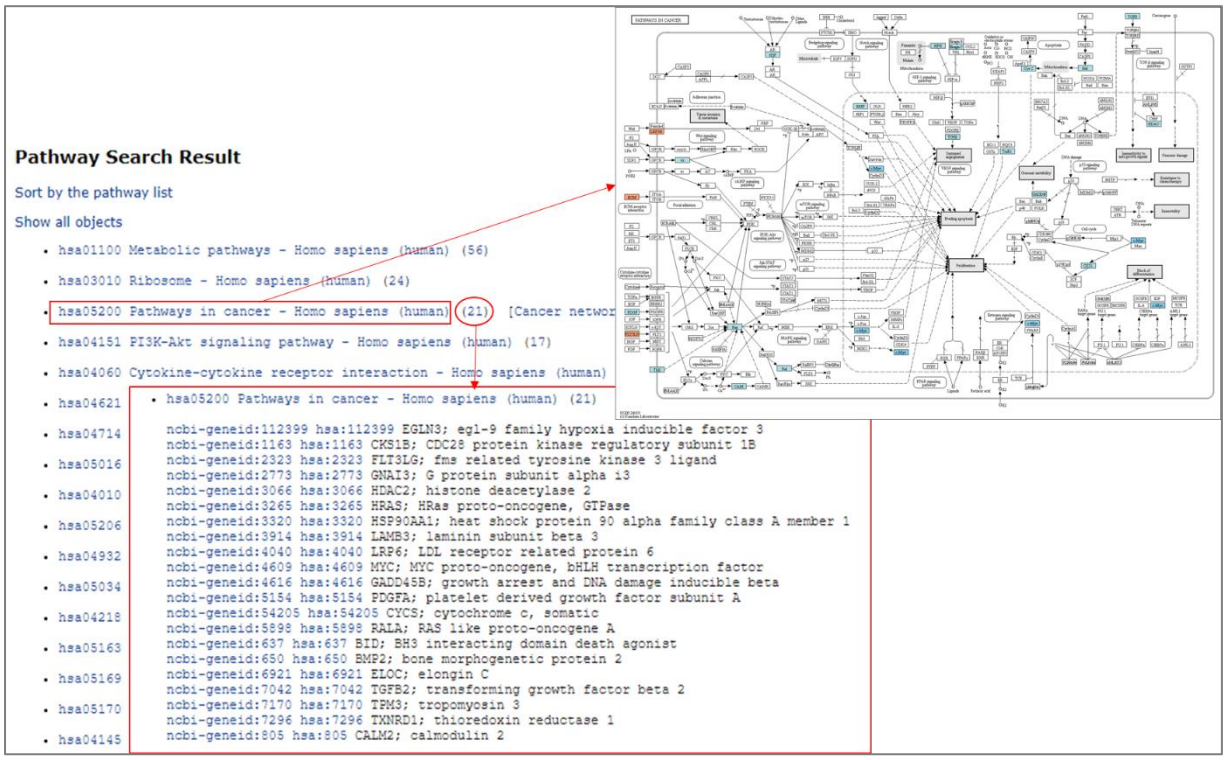


그림 5-4. KEGG Mapper tool analysis result

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

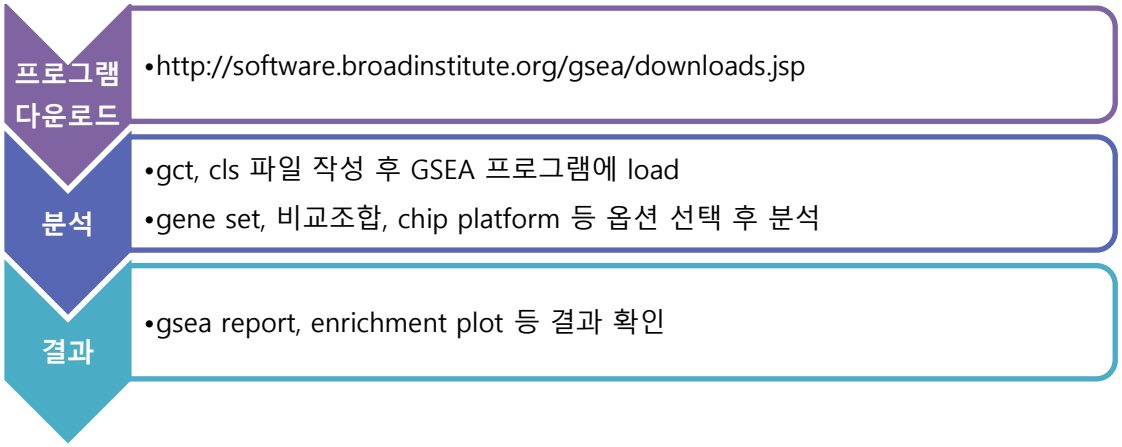


그림 6-1. GSEA tool analysis process

GSEA 홈페이지(http://software.broadinstitute.org/gsea/downloads.jsp)에 들어가 회원가입 후 로그인 하여 GSEA 프로그램을 다운로드 받는다 (그림 6-2).

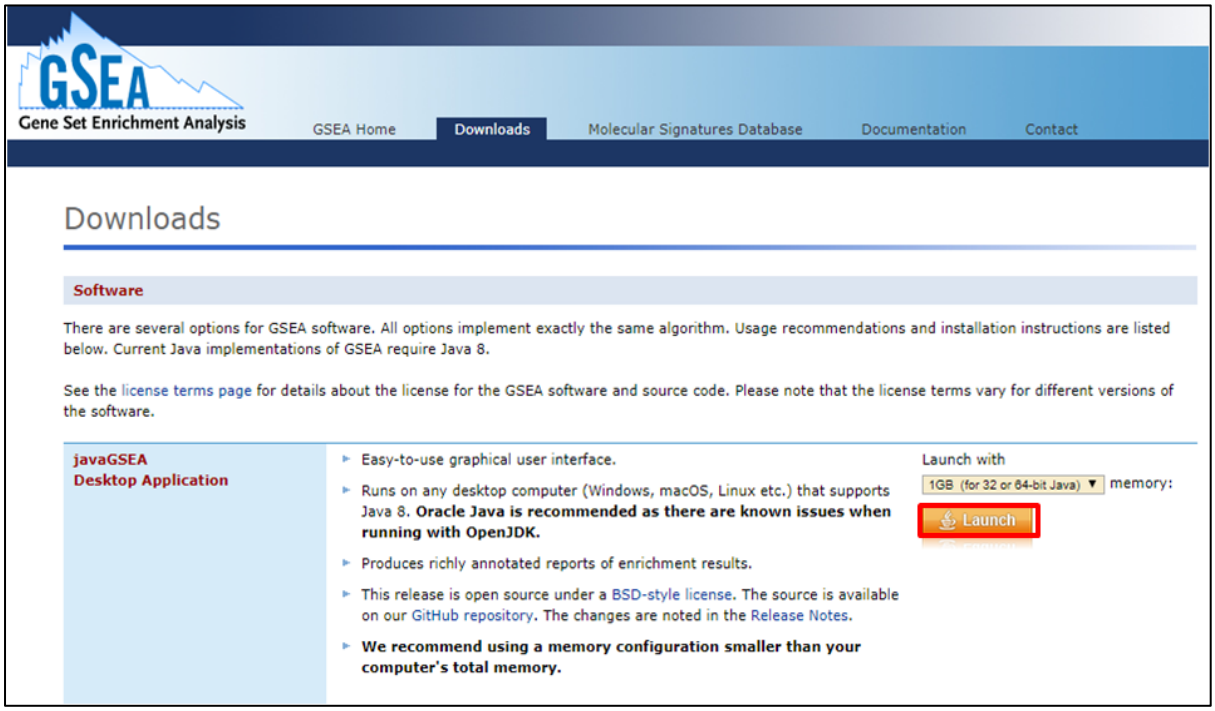
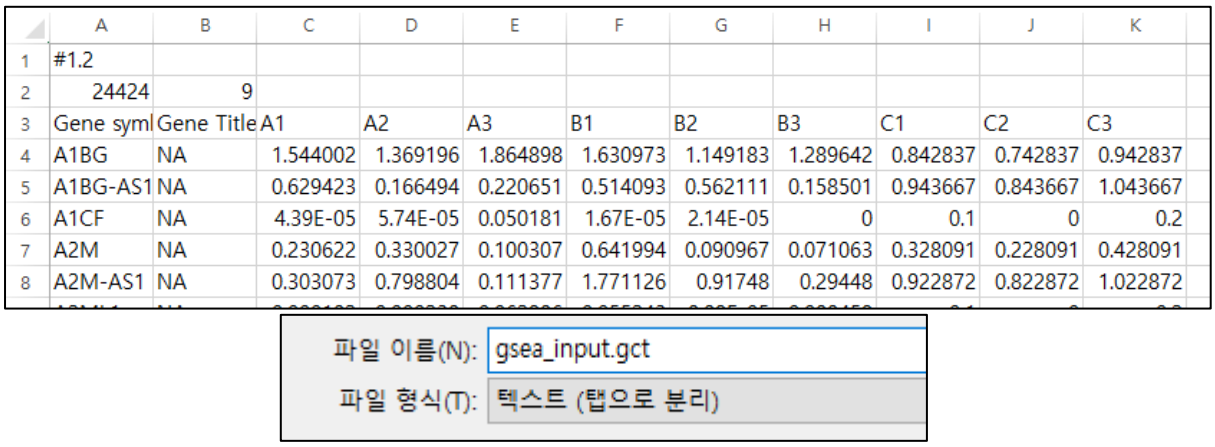


그림 6-2. GSEA program download

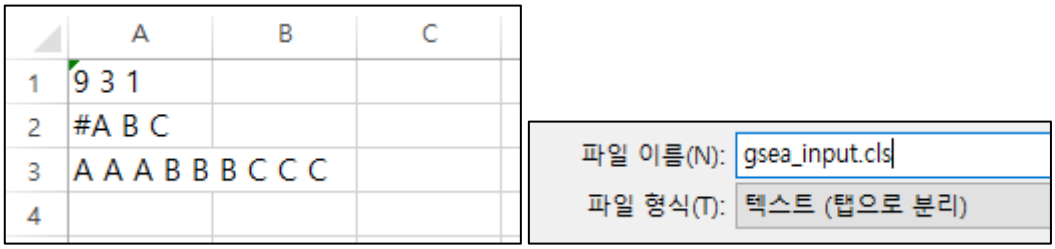
GSEA 분석을 위해서는 유전자 발현값 정보가 포함되어 있는 gct 파일과 샘플 정보가 포함되어 있는 cls 파일이 필요하다. Gct 파일은 그림 6-3과 같은 형식으로 만든다. A1칸에는 항상 "#1.2", A2칸에 유전자 수, B2칸에 샘플 수를 기입한다. RNA-seq data는 A열에 Gene symbol, B열에 Gene title (gene description), C열부터 각 샘플의 normalized data (log2 변환하지 않은 값)을 기입한다. Microarray data는 A열에 probe ID, B열에 Gene symbol, C열부터 각 샘플의 normalized data (log2 변환하지 않은 값)을 기입한다. 파일 저장할 때는 파일명 뒤에 ".gct"를 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 sym	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 6-3. gct file

cls 파일은 그림 6-4와 같은 형식으로 만든다. A1칸에는 "샘플수(띄어쓰기)그룹수(띄어쓰기)1", A2칸에는 "#그룹이름", A3칸에는 gct파일에 기입한 샘플의 순서대로 각 샘플이 어떤 그룹에 속하는지 그룹이름을 기입한다. A2, A3칸에서 띄어쓰기로 그룹을 구분한다. 파일 저장할 때는 파일명 뒤에 ".cls"를 붙이고 파일형식은 "텍스트 (탭으로 분리) 파일"로 저장한다.



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

그림 6-4. cls file

GSEA 프로그램을 열어 Load data 버튼을 누르고 Browse for files 버튼을 누른 후 gct, cls 파일을 연다(그림 6-5). gct, cls 파일은 파일의 경로가 길면 input 파일을 잘 인식하지 못하여 되도록 바탕화면에 두고 수행한다.

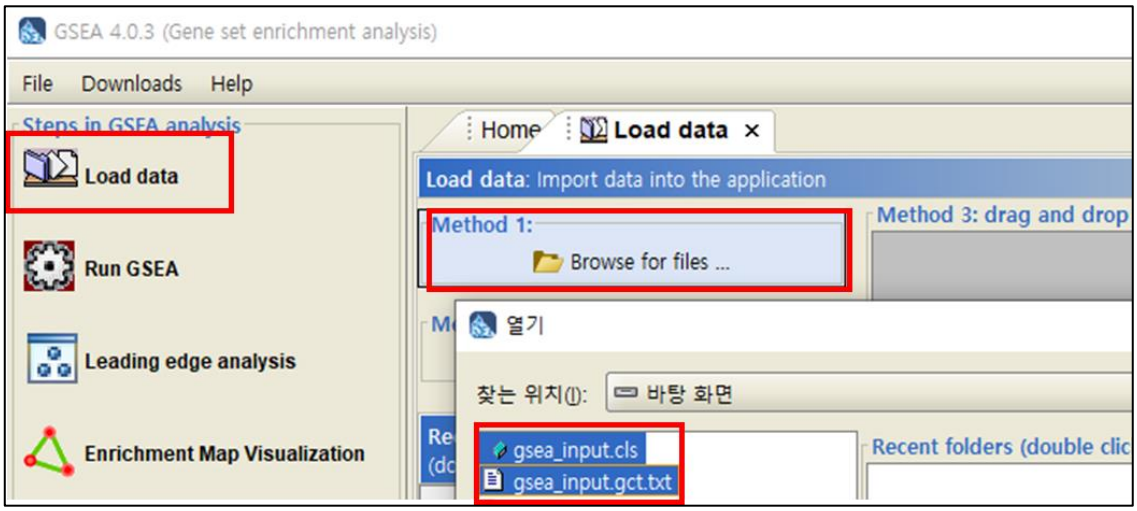


그림 6-5. Load data in GSEA program

Run GSEA 를 누르고 Expression dataset 는 gct 파일명을 선택, gene sets database 는 분석하고자 하는 gene set 을 선택한다(그림 6-6). pathway 분석을 하고자 하면 c2 를 선택, gene ontology 분석을 하고자 하면 c5 를 선택한다. Gene set 에 대한 자세한 설명은 GSEA 홈페이지 (http://software.broadinstitute.org/gsea/msigdb/collections.jsp)에 있다. Number of permutations 은 1000 으로 기입하고, Phenotype labels 은 분석하고자 하는 비교조합(test versus control)을 선택한다. Collapse/remap to gene symbols 은 Collapse 을 선택하고, permutation type 은 gene_set 을 선택한다. Chip platform 은 RNA-seq 의 경우엔 Human(or Mouse or Rat)_Symbol_with_Remapping_MSigDB.v7.0.chip 을 선택한다. Microarray 의 경우엔 실험한 chip 을 선택한다. Run 버튼을 누르면 분석이 시작된다.

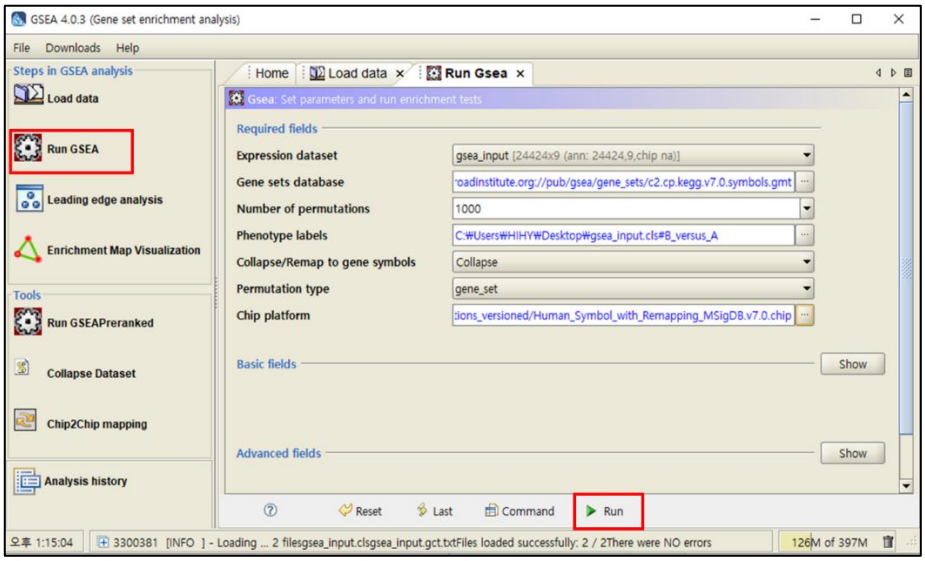


그림 6-6. Run GSEA

분석이 완료되면 GSEA 왼쪽 아래 GSEA reports 창에 status 가 Success 로 바뀐다. Show results folder 를 누르면 GSEA 분석 결과 창이 열린다(그림 6-7).

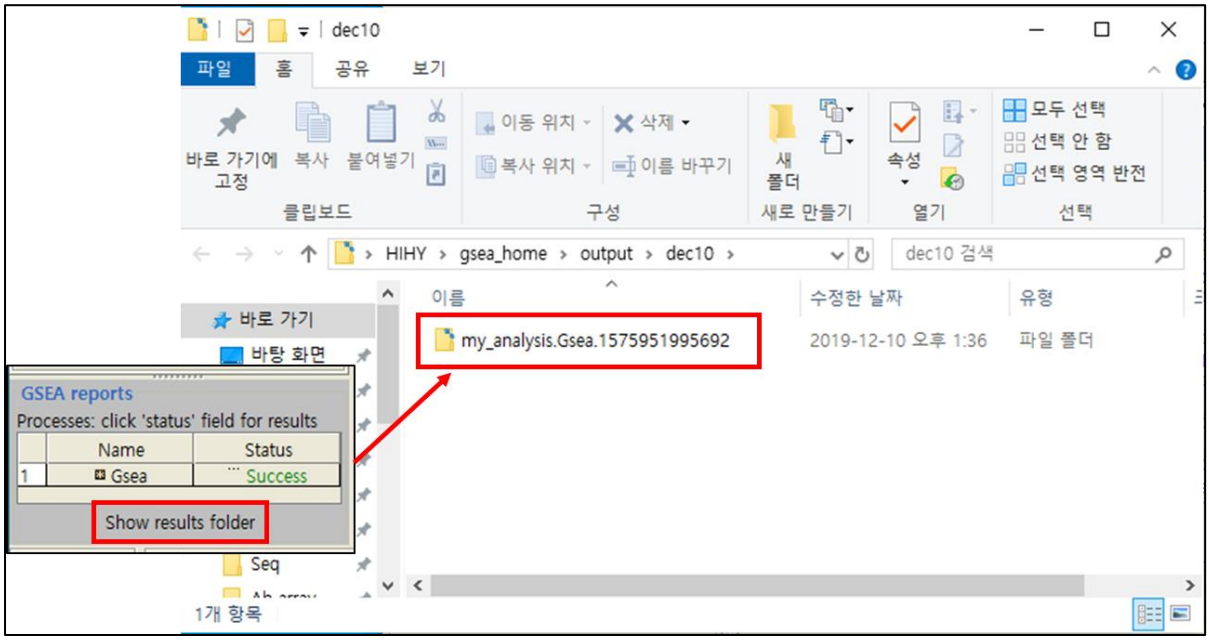


그림 6-7. GSEA results folder

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 6-8).

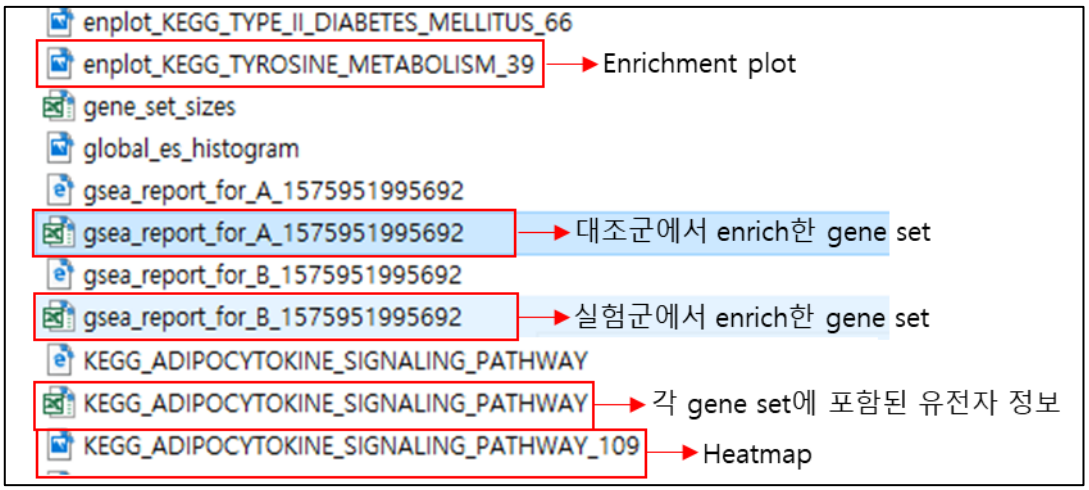


그림 6-8. GSEA result files

_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 6-9). Peak가 왼쪽에 생기면 대조군 대비 실험군에서 up된 유전자들이 많다는 의미이고, peak가 오른쪽에 생기면 down된 유전자가 많다는 의미이다.

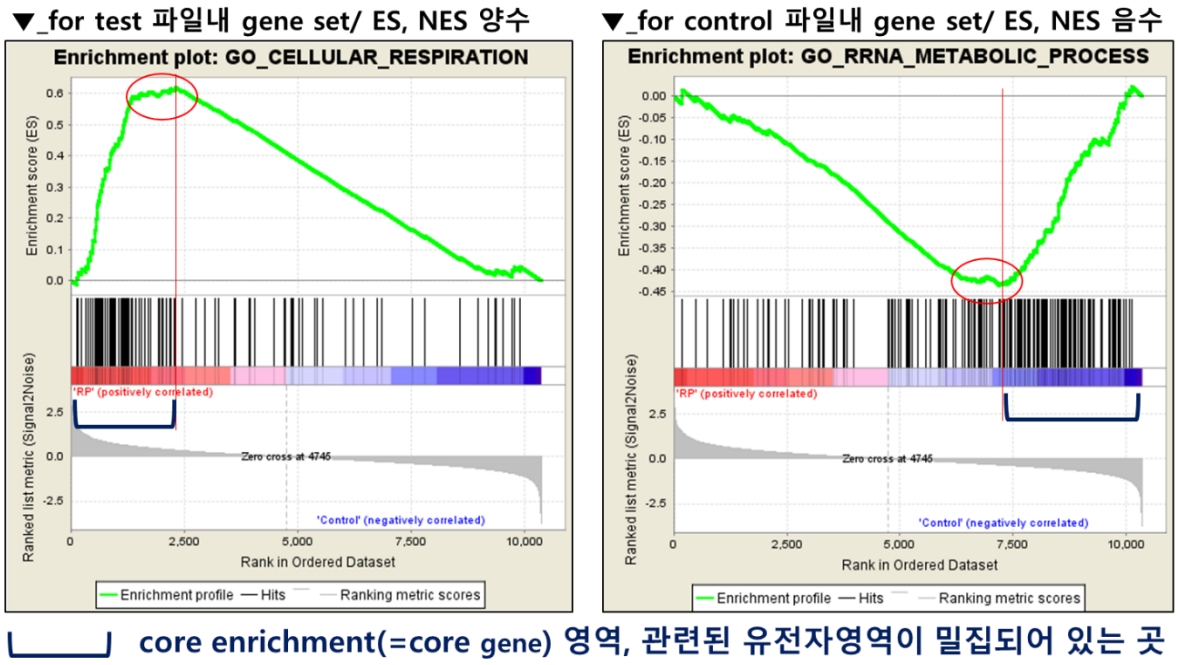


그림 6-9. GSEA enrichment plot

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

7. 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 7-1 과 같다.

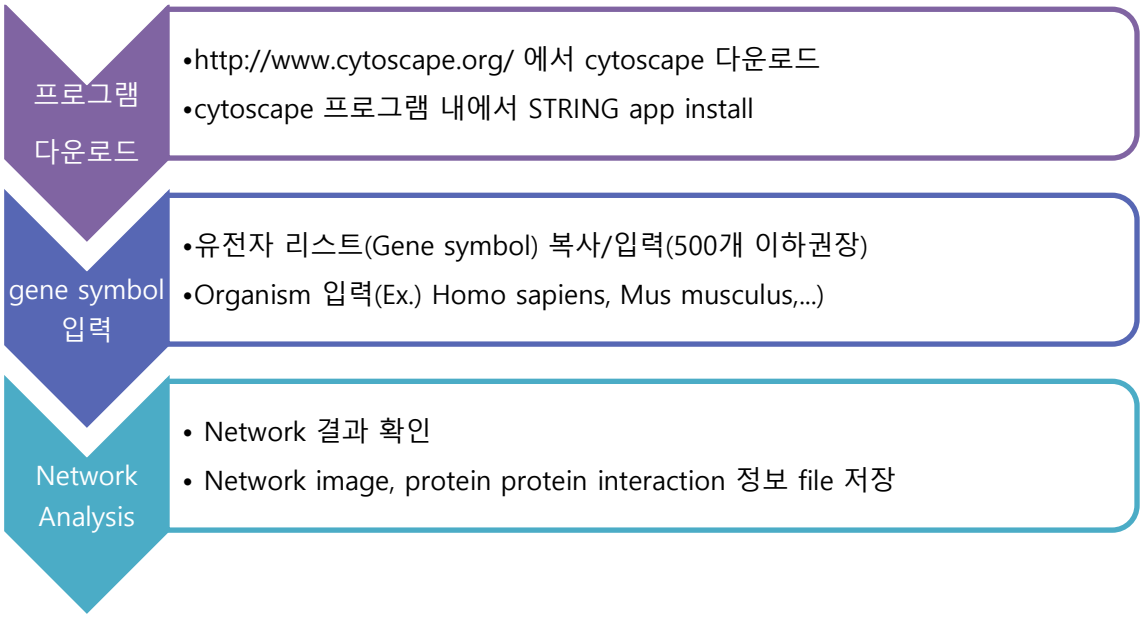


그림 7-1. STRING analysis process

Cytoscape 홈페이지 (http://www.cytoscape.org/)에서 cytoscape 프로그램을 다운로드 받아 설치한다(그림 7-2).

그림 7-2. Cytoscape download

Cytoscape 프로그램을 열어 상위에 있는 메뉴 중 [Apps] > [App Manager]로 들어간다(그림 7-3). StringApp 을 선택 후 Install 버튼을 누른다.

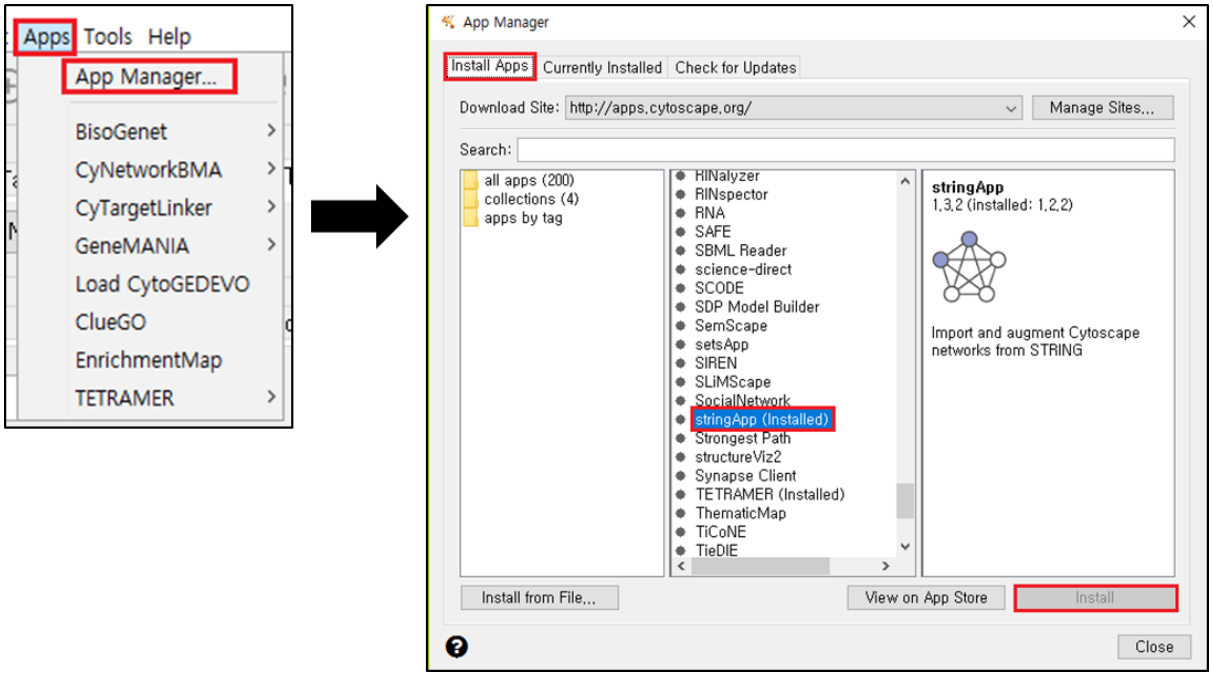


그림 7-3. STRING app installation in Cytoscape

Cytoscape 상위 메뉴 중 [File] > [Import] > [Network from Public Databases]로 들어간다(그림 7-4).

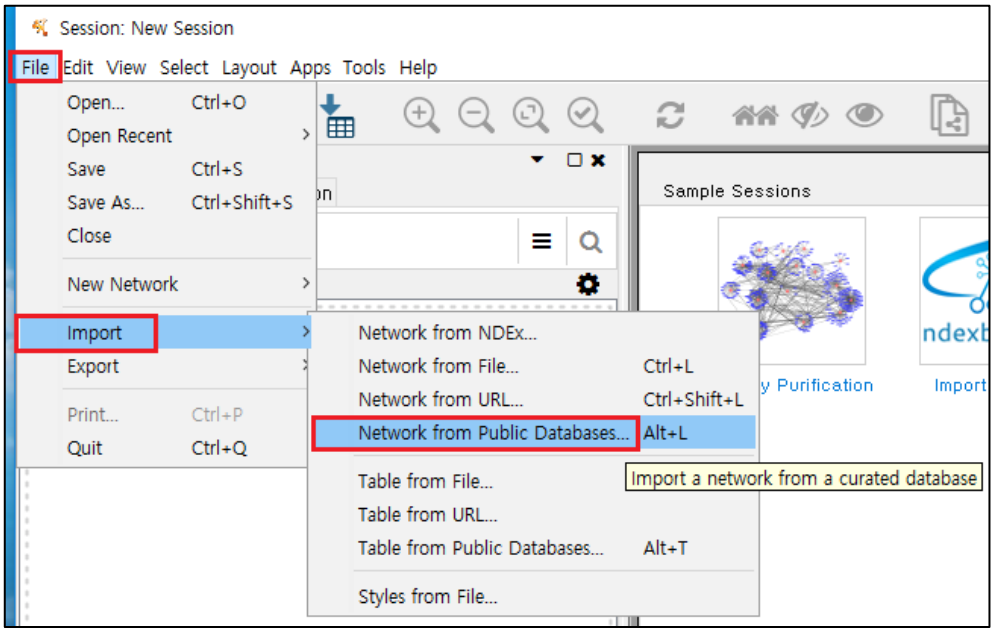


그림 7-4. STRING analysis process 1

Data Source 를 "STRING : protein query" 선택하고 Species 를 선택한다(그림 7-5). 분석하고자 하는 유전자들의 gene symbol 을 입력한다. Confidence (score)는 Protein-Protein Interaction 강도를 뜻하는 것으로 0 부터 1 까지 이고, 1 로 갈 수록 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.

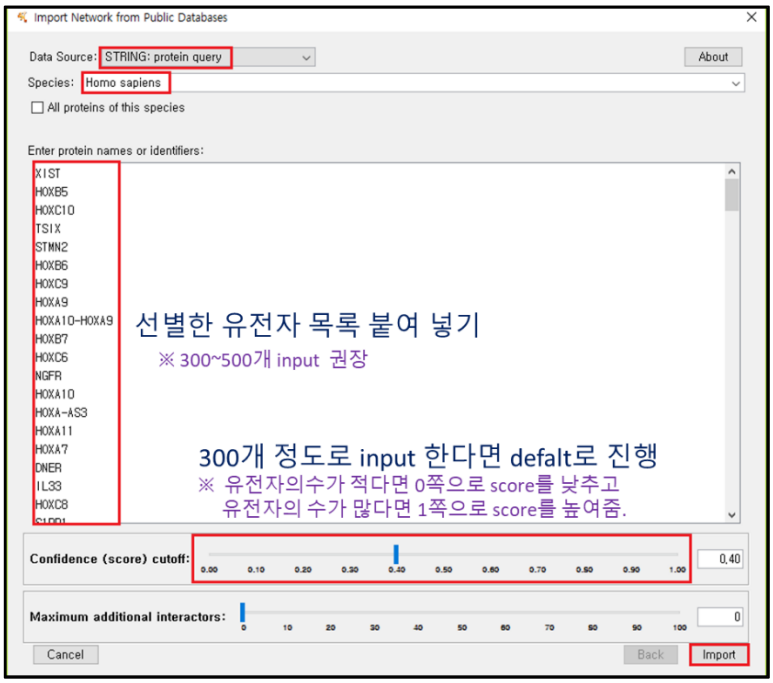


그림 7-5. STRING analysis process 2

Input 한 gene symbol 과 match 가 되지 않는 protein 이 있으면 그림 7-6 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.

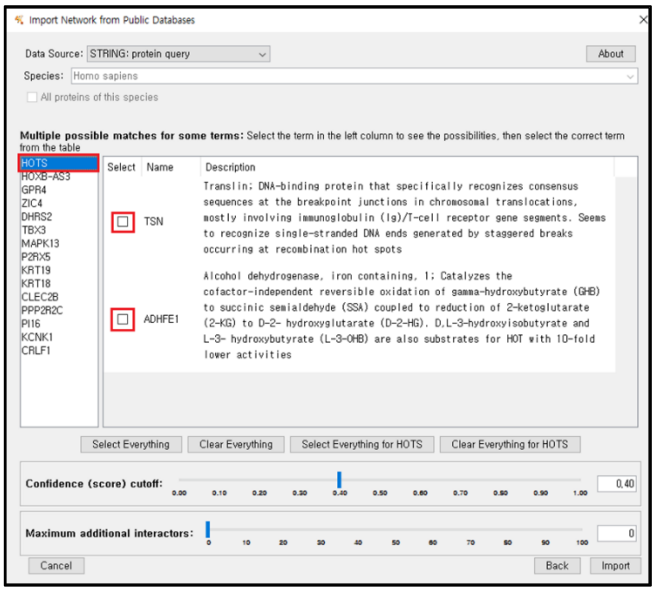


그림 7-6. Not matched proteins in STRING

분석이 완료되면 network image 가 나온다(그림 7-7).

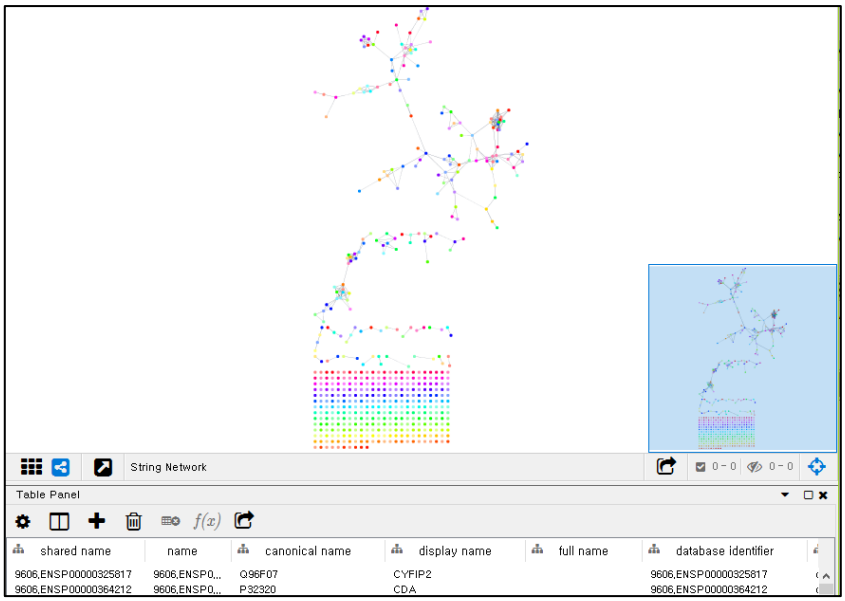


그림 7-7. Network result

[File] > [Export]> [Network to Image]를 눌러 이미지를 저장한다(그림 7-8). PDF 파일형식으로 저장하는 것을 권장한다. Pdf 파일로 저장하면 확대를 하여도 이미지가 깨지지 않는다.

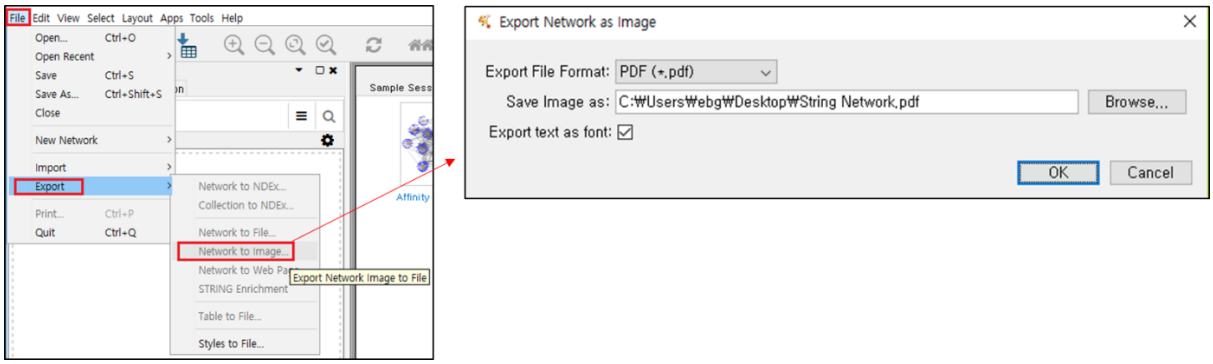


그림 7-8. Save network image

어떤 유전자들이 protein-protein interaction 을 하는지 정보를 저장하고 싶으면 [File] > [Export] > [Table to File...]로 들어가 String Network default edge 파일을 저장한다(그림 7-9).

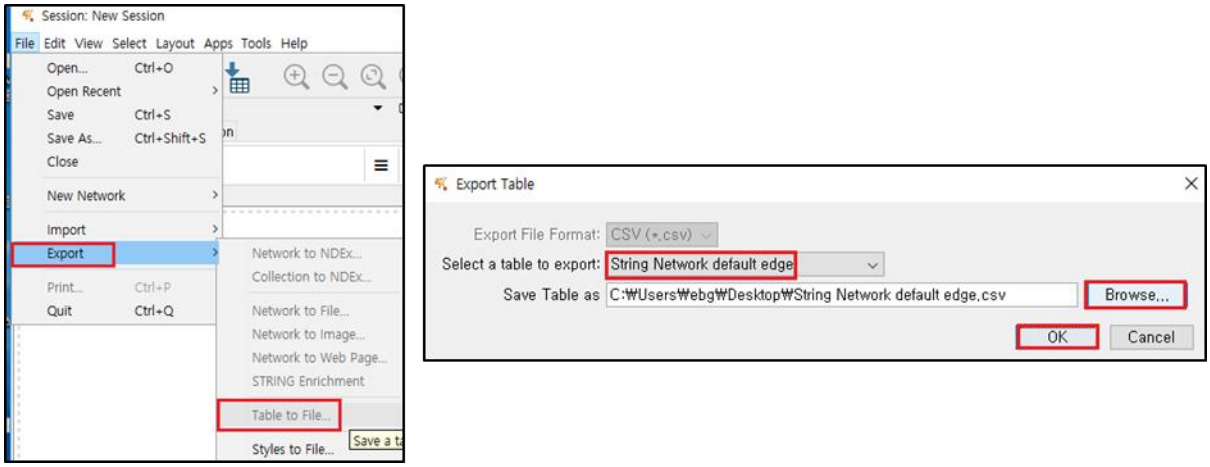


그림 7-9. Save edge table

String Network default edge 파일에서 name 에 interaction 정보, score 에 confidence score 가 나와있다(그림 7-10). 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	experi	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) M	0.503

그림 7-10. Interaction information in edge table

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