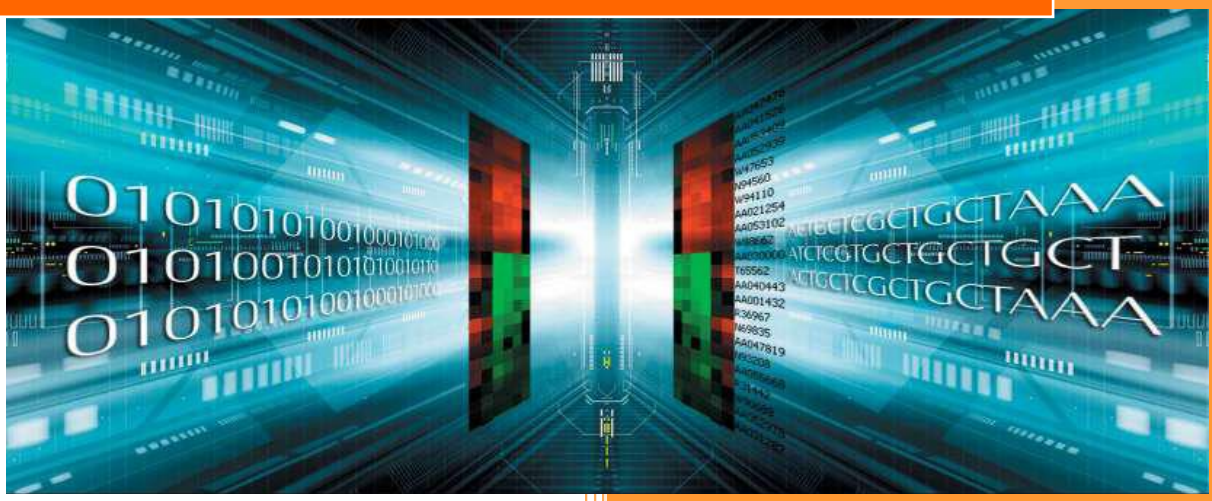
Version 3.0

RNA-Seq/Microarray DEG Analysis



㈜이바이오젠
서울특별시 영등포구 선유로13길 25
(문래동6가), 에이스하이테크시티2, 305호
Tel. 02-3141-0791
service@e-biogen.com
http://www.e-biogen.com

<목 차>

1. 엑셀기반 DEG 분석 (ExDEGA v.1.6.0)

㈜이바이오젠은 QuanSeq, mRNA-Seq, Total RNA-Seq 과 Micorarray data 를 엑셀 기반에서 DEG 를 쉽게 분석할 수 있도록 분석보고 시 ExDEGA (Excel based Differentially Expressed Gene Analysis) tool 을 함께 제공한다. ExDEGA 분석툴은 ㈜이바이오젠이 연구자들이 Microarray 및 RNA-Seq 데이터를 보다 쉽게 다루고 원하는 데이터를 쉽게 얻을 수 있도록 사용자 편의를 최대한 반영한 분석툴이고 엑셀 프로그램 안에서 다양한 분석을 직관적으로 수행할 수 있도록 개발되었다. ExDEGA 분석툴은 사용자들의 요구사항을 지속적으로 반영하여 데이터분석과 엑셀사용에 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 분석툴이 설치된다(그림 1-1).

설치가 완료되면 보고된 엑셀데이터를 열면 자동으로 ExDEGA 분석툴이 엑셀에 반영된 것을 확인할 수 있다. 참고로 ExDEGA 설치 전에 실행 중인 엑셀 파일이 있으면 종료시킨 후 다시 실행해야 ExDEGA 를 사용할 수 있다.

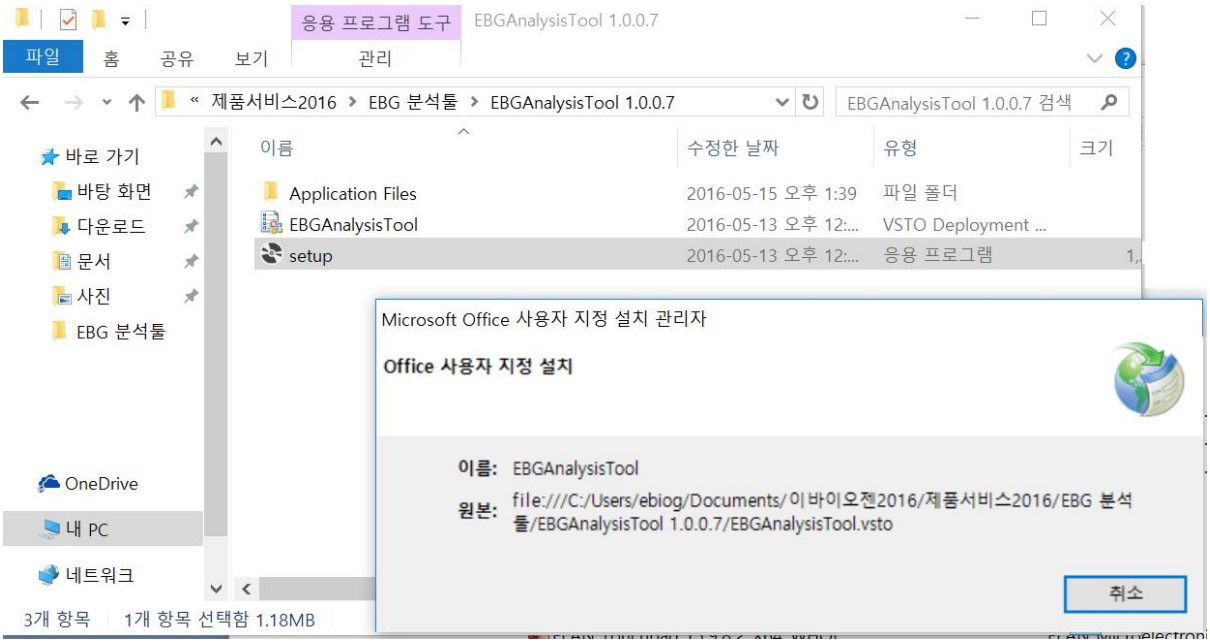


그림 1-1. ExDEGA set up

??? ExDEGA Report.xls 파일을 열면 왼쪽에 Gene Ontology (GO) 분석 창과 가운데에 mRNA expression data, 오른쪽에 DEG 분석 창이 나온다(그림 1-2).

GO 분석 창에서는 기본 설정된 GO 와 사용자가 원하는 대로 GO 를 구성하여 분석할 수 있고 DEG 분석과 함께 연동하여 데이터를 쉽게 얻을 수 있다. DEG 분석 창에서는 Fold change, Normalized RC, p-value 등을 선택하여 원하는 데이터를 쉽게 얻을 수 있고 GO graph 를 통해 전체적인 발현패턴을 확인할 수 있다. 뿐만 아니라, DEG 분석 창에서 Scatter Plot, Volcano Plot, Venn Diagram 을 직접 그릴 수 있고 필터링된 유전자들을 대상으로 Clustering heatmap 을 작성하기 위한 MeV 프로그램 input file 을 자동으로 만들 수 있고 Gene expression graph, Gene search 기능도 이용할 수 있어 연구자가 RNA-Seq data 를 쉽게 활용할 수 있다.

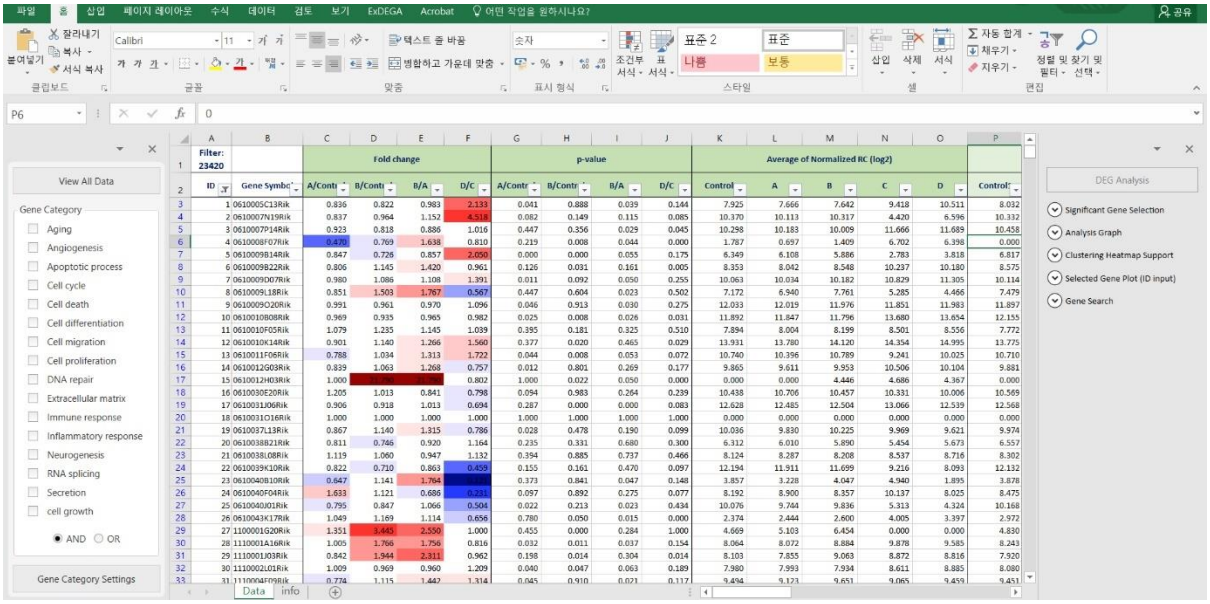


그림 1-2. mRNA expression data format made in E-Biogen

1-1. Gene Category 사용 방법

mRNA expression 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.

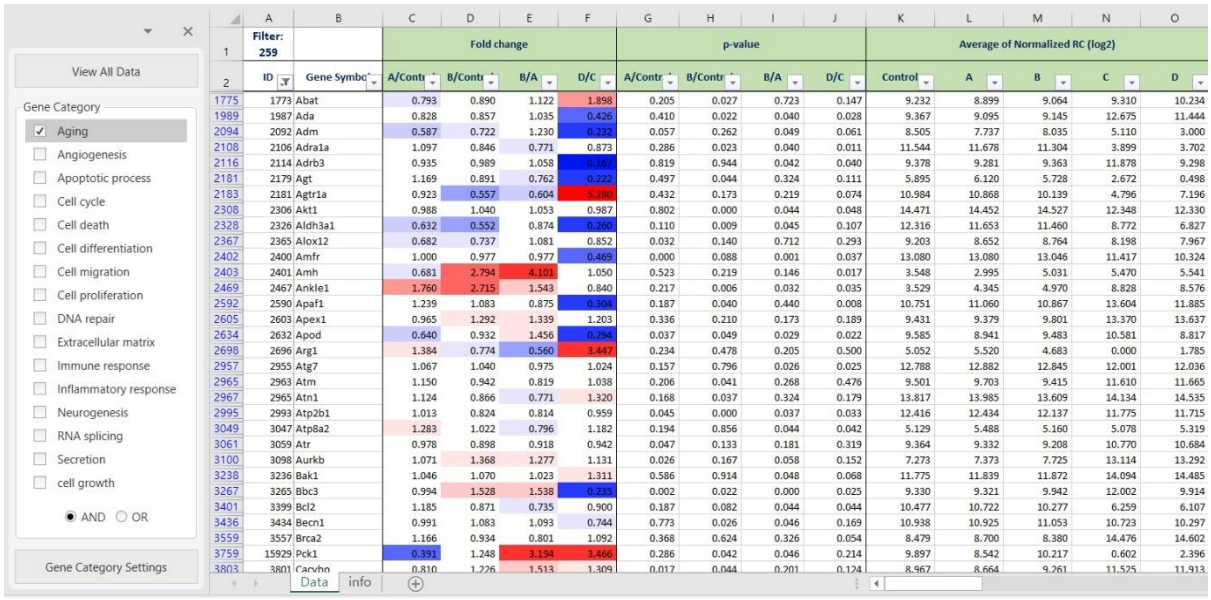


그림 1-3. Gene ontology (Aging) selection

'View All Data' 버튼을 누르면 필터를 해제하여 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.

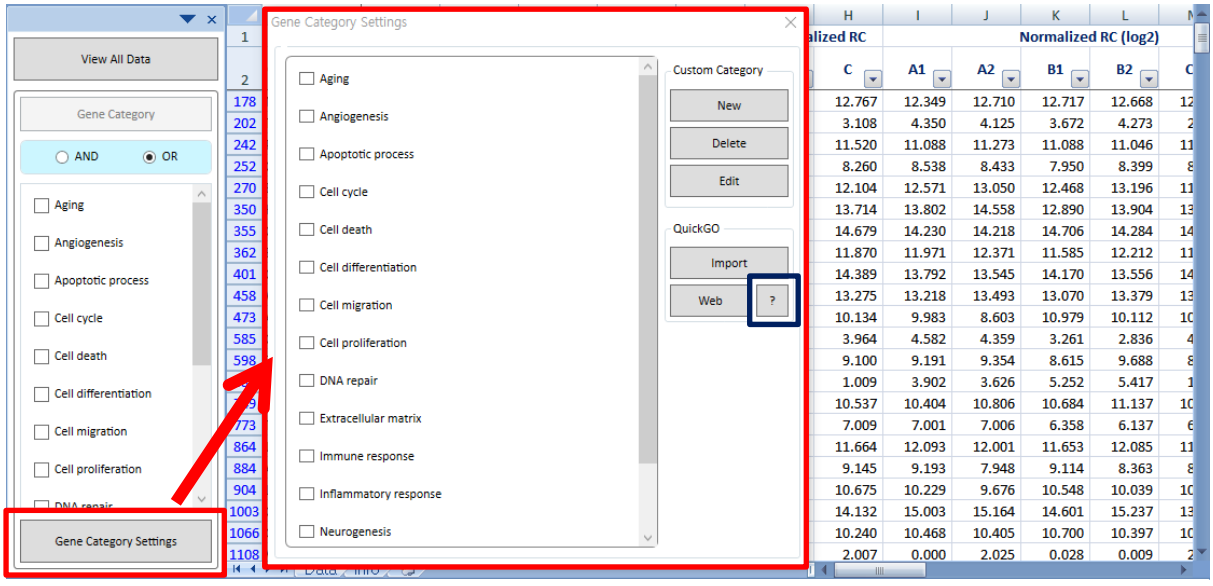


그림 1-4. Gene category settings

만약 원하는 유전자 그룹 목록이 있다면, 직접 입력하여 새로운 Gene Category 를 추가할 수도 있다. Gene Category Settings 버튼을 누른 후 New 를 선택하고 원하는 gene list 입력(or 복사-붙여넣기) 한 뒤, Gene category 이름 설정 후 저장하면 새로운 GO category 를 확인 할 수 있다(그림 1-5-a,b).

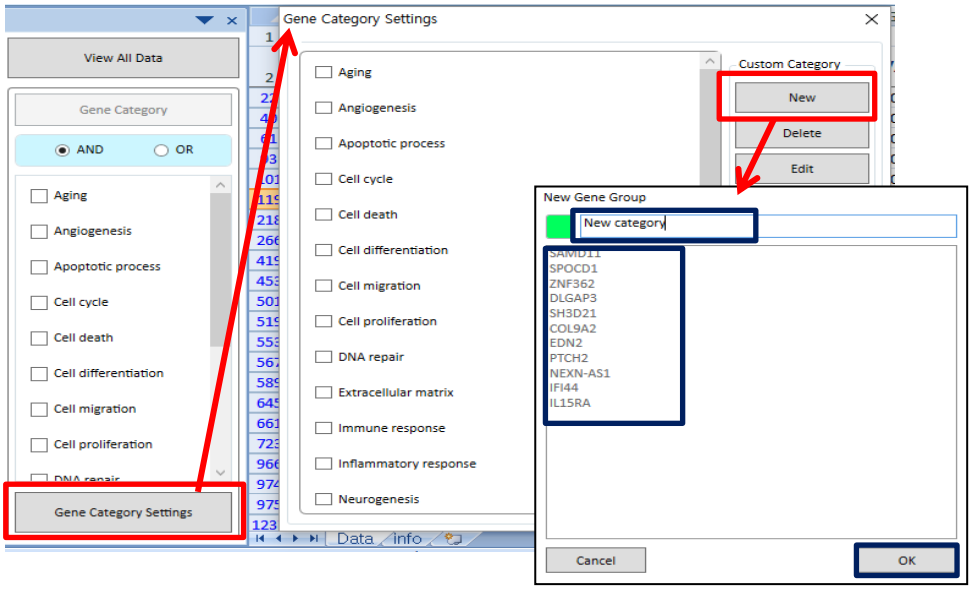


그림 1-5-a. Adding Genes to make a new gene category

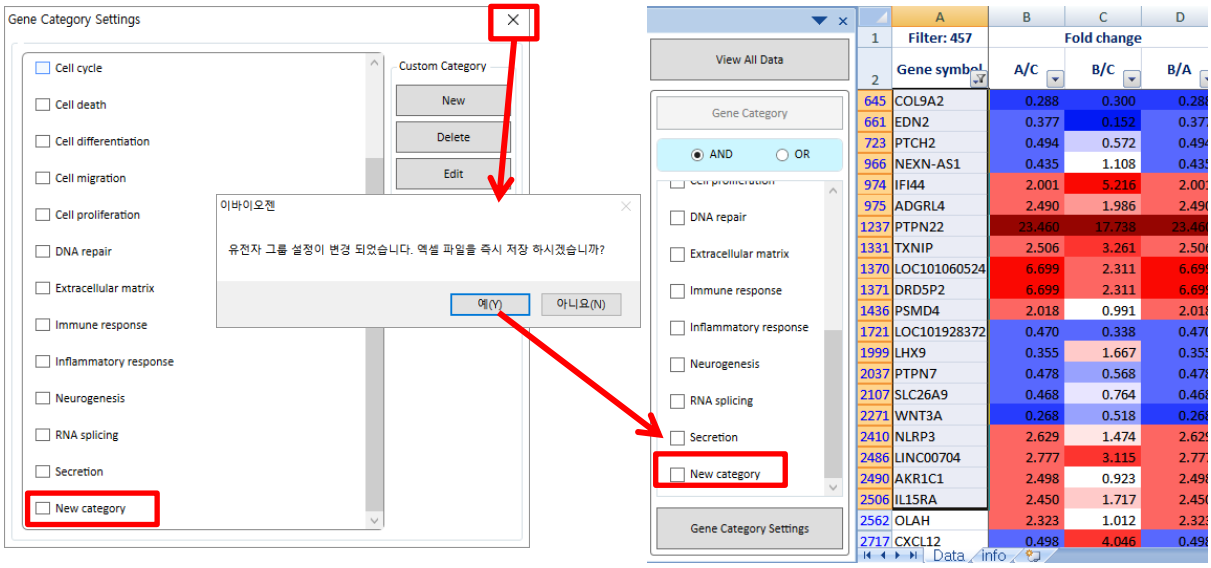


그림 1-5-b. Adding Genes to make a new gene category

1-2. Significant Gene Selection 사용 방법

오른편의 DEG Analysis 부분에서 "Significant Gene Selection" 창은 전체 결과 중 control 과 test 를 비교한 결과에서 유의하게 발현 차이가 나는 유전자를 필터링 할 수 있도록 만들어 놓은 것이다. 예를 들어, control 기준으로 A 에서 발현이 2 배 이상 증가 또는 감소하고, normalized RC(log)값이 4 이상이고, t-test 결과 p-value 값이 0.05 이하인 유전자(반복 실험한 데이터의 경우)를 선택하면 95 개의 유전자가 필터링 된다(그림 1-6).

그리고 여러 개의 비교그룹에서 동시에 Significant gene 을 선별하고자 할 경우와 적어도 한 비교그룹에서 Significant gene 을 선별하고자 할 경우에는 "AND"와 "OR" 기능을 사용하면 된다.

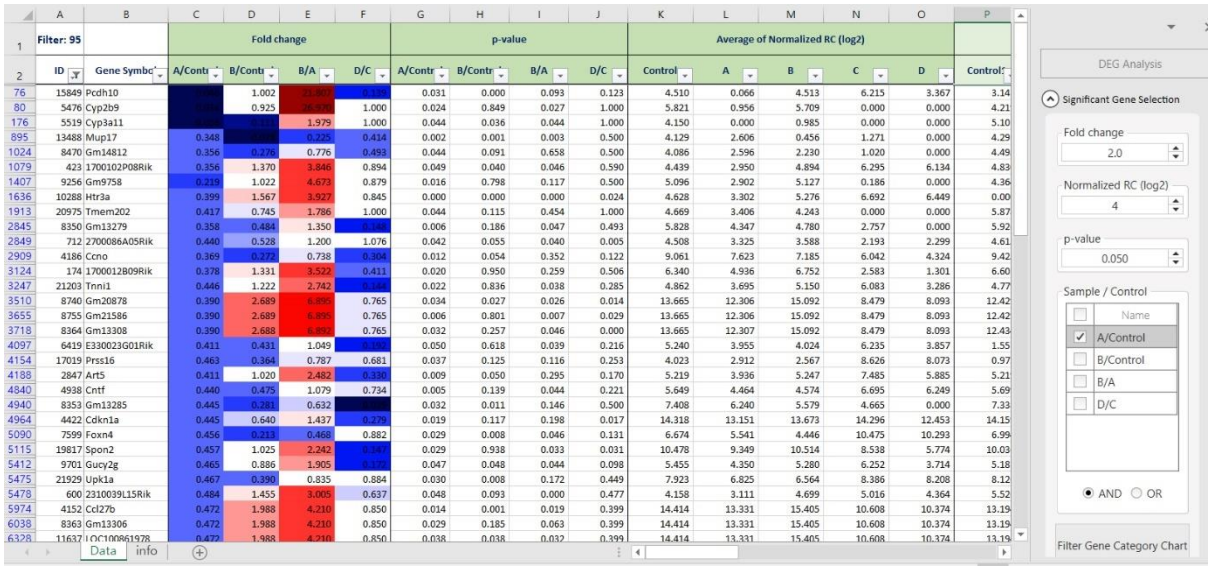


그림 1-6. Significant gene selection

Gene Category 와 Significant gene selection 은 연동 가능하다. 그림 1-7 에서 처럼 Gene Category 의 Cell differentiation 을 선택하면 10 개의 유전자가 필터링 된다(그림 1-7). 10 개의 유전자는 본 데이터에서 Cell differentiation 관련 유전자들 중 A/Control 비교그룹에서 유의하게 발현이 증가 또는 감소한 유전자를 의미한다.

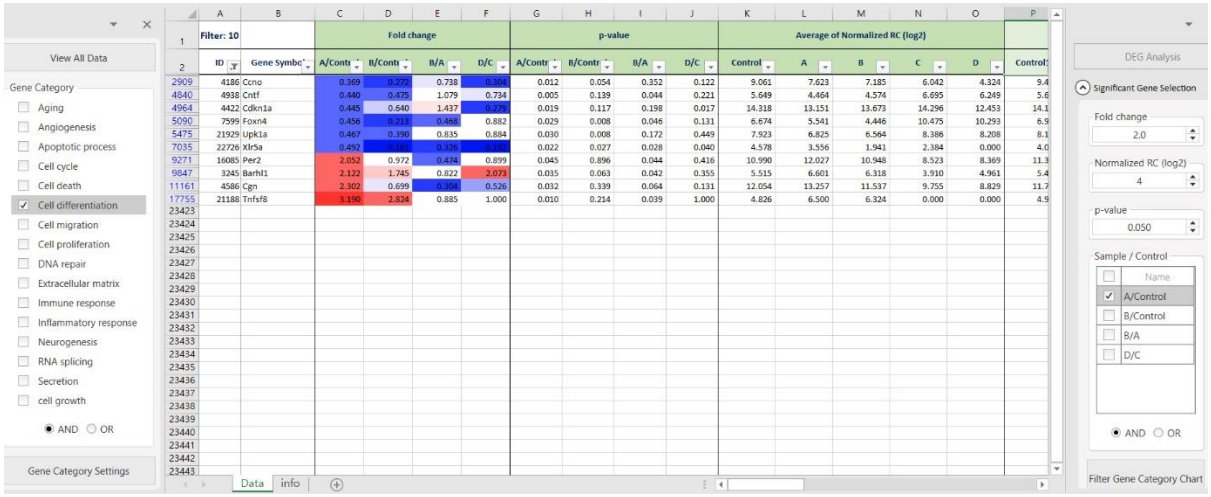


그림 1-7. Significant genes related to Cell cycle

실험 결과에 따라 발현 변화값 (fold change), p-value, normalized RC(log2) 기준을 조정할 수 있고 반복 실험인 경우만 p-value 를 선택할 수 있다.

"View Gene Category Chart" 버튼을 누르면 각 GO 관련 유전자 중 발현이 유의하게 차이 나는 유전자의 %와 수가 그래프로 그려진다. 본 분석을 통해 어떤 GO 의 유전자들이 상대적으로 많은 발현 변화가 있었는지를 확인할 수 있다. 전체 데이터 상태에서 Significant Gene Selection 의 비교그룹을 선택하고 "View Gene Category Chart"를 클릭하면 증가/감소한 유전자 들 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 되고 오른쪽의 증가/감소된 bar chart 에서 bar 상단의 숫자는 해당 유전자 수이고 bar 를 클릭하면 해당 유전자가 필터링 된다(그림 1-8).

그림 1-8. View Gene Category Chart

1-3. Analysis Graph 사용 방법

DEG Analysis 부분에서 "Analysis Graph" 창을 펼치면 아래 그림 1-9 와 같이 Scatter Plot, Volcano Plot, Venn Diagram 을 엑셀에서 쉽게 그릴 수 있다.

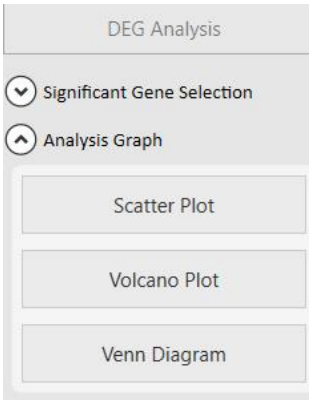


그림 1-9. Analysis Graph Tool

첫번째 Scatter Plot 은 오른쪽에 샘플 비교그룹을 선택하고 Fold threshold line 을 선택하고 "Graph View"를 클릭하면 왼쪽에 선택한 비교그룹을 대상으로 Scatter Plot 이 자동 생성된다. Plot 에서 특정 spot 을 클릭하면 해당 유전자가 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 1-10).

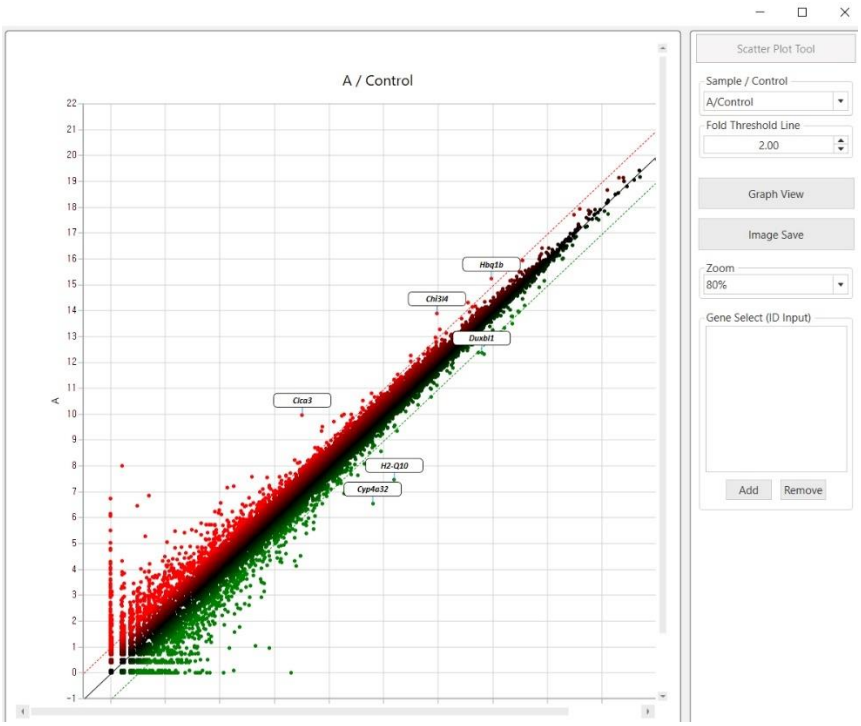


그림 1-10. Analysis Graph Tool – Scatter Plot

두번째 Volcano Plot 은 Scatter Plot 의 기능과 거의 동일한데 오른쪽에 샘플 비교그룹을 선택하고 Fold threshold line 과 p-value 를 선택하고 "Graph View"를 클릭하면 왼쪽에 선택한 비교그룹을 대상으로 Scatter Plot 이 자동 생성된다. Plot 에서 특정 spot 을 클릭하면 해당 유전자가 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 1-10).

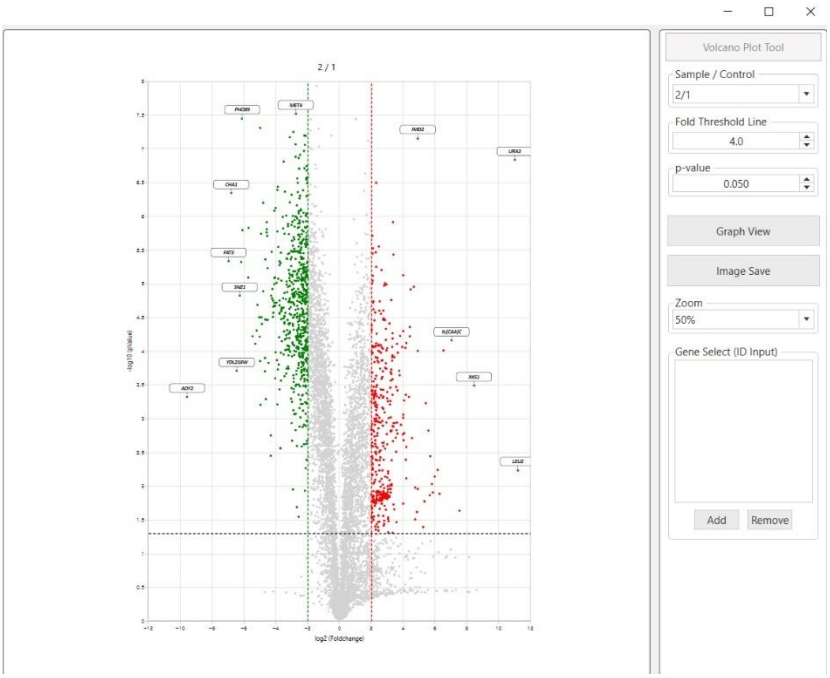


그림 1-11. Analysis Graph Tool – Volcano Plot

세번째 Venn Diagram 을 통해 2 개, 3 개 또는 4 개 까지의 비교그룹을 대상으로 Venn Diagram 을 작성할 수 있다. Venn Diagram 을 그릴 샘플 비교그룹과 Fold Change, p-value(반복실험시)을 선택 후, Diagram View 를 클릭하면 결과를 확인할 수 있으며 그룹은 최대 4 그룹까지 선택 가능하다. 아래의 그림은 A/C 와 B/C, B/A 결과 중, 2fc 이상 up, down 된 list 를 가지고 Venn Diagram 을 작성한 결과이다(그림 1-12).

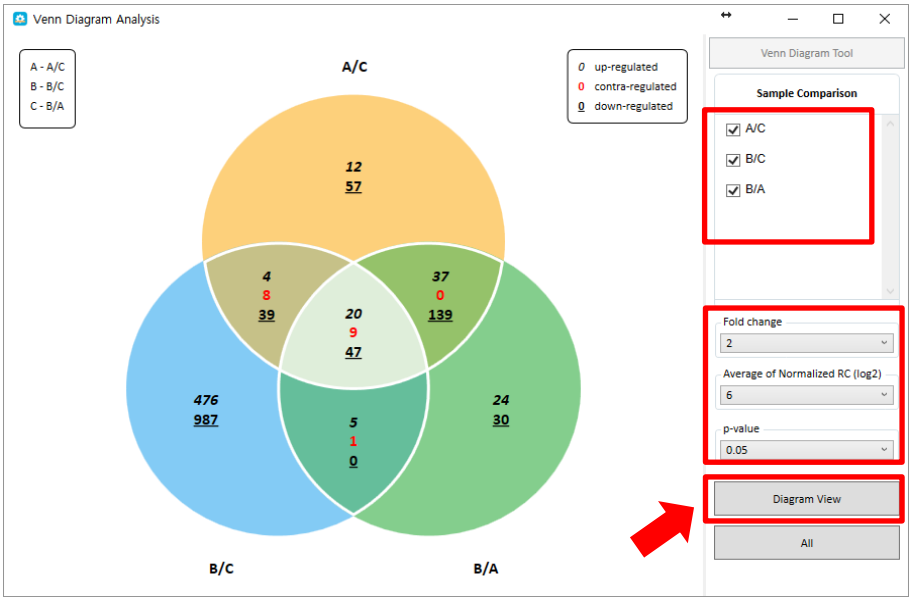


그림 1-12. Analysis Graph Tool – Venn Diagram

Venn Diagram 결과에서 표시되는 형식은 다음과 같다(그림 1-13).

1. ***기울어진 숫자*** : 2fold 이상 up-regulated 된 gene 수
2. **빨간색 숫자** : regulation 이 대조되는 gene 수
3. **밑줄 친 숫자** : 2fold 이상 down-regulated 된 gene 수

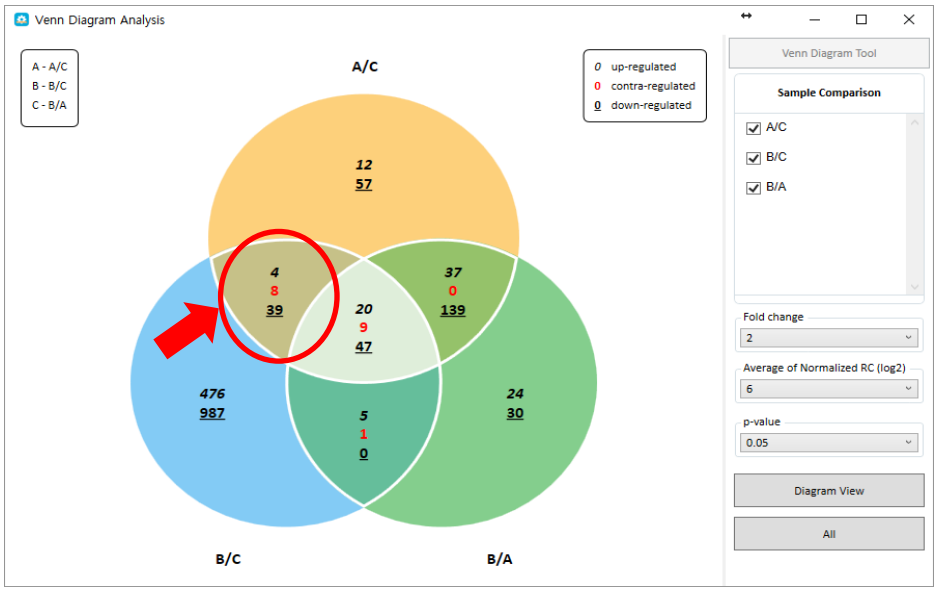


그림 1-13. For example of up ,down, contra-regulated in Venn Diagram

Venn Diagram 이미지를 오른쪽 클릭하면 Venn Diagram 각 영역에 어떤 유전자들이 있는지 확인할 수 있다. 예를 들어, A/C 에서만 2fold up 이 되는 유전자를 보고 싶으면, Venn Diagram 에서 A/C 에서만 해당되는 영역을 찾아 마우스 오른쪽 클릭 하면 2fold up 된 유전자 list 4 개가 엑셀 sheet 에 filter 된다(그림 1-14).

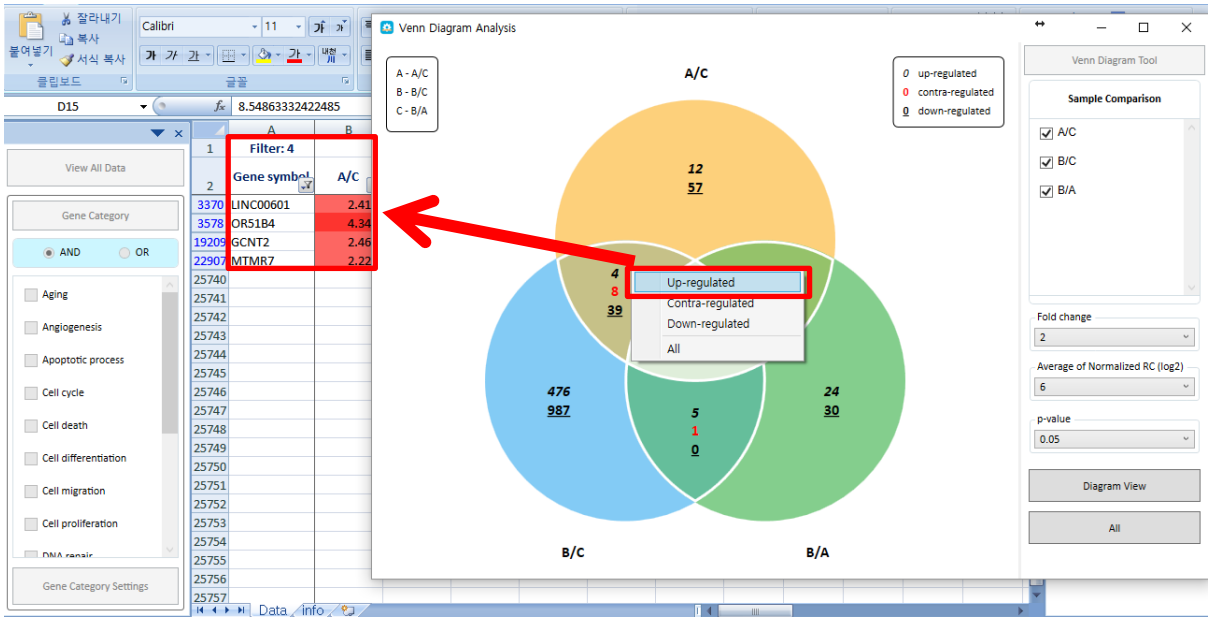


그림 1-14. Filtering 2fold up-regulated gene list in Venn Diagram

ExDEGA 에서 제공되는 모든 이미지는 오른쪽마우스를 눌러 'Save image' 버튼을 통해 저장이 가능하다(그림 1-15).

그림 1-15. Save image

1-4. Clustering Heatmap Support 사용 방법

ExDEGA 의 DEG Analysis 에서는 Significant Gene Selection 또는 Venn Diagram 등을 통해 Data Mining 을 수행한 후 정리된 유전자 리스트를 대상으로 Clustering Heatmap 을 쉽게 작성할 수 있도록 지원한다.

당사에서 추천하는 Clustering Heatmap 프로그램은 MeV 인데 ExDEGA 에서 MeV 용 Input file 을 자동 생성해 주고 MeV 에서 해당 파일을 불러오면 된다. 이후의 Clustering 방법 및 이미지 가공 및 저장 방법은 본 매뉴얼 "4. MeV Software 이용 Clustering Heatmap 작성" 부분을 참고하면 된다.

그림 1-16 에서 필터링된 유전자 리스트를 대상으로 Clustering Heatmap 을 작성하려면 크게 2 종류의 데이터를 이용할 수 있는데 첫번째는 Fold change 값을 이용할 시 Type 부분에 Fold change 를 체크하고 Export Data Select 에서 Heatmap 에 표현할 비교그룹을 체크하여 "Data Export"를 클릭한 후 "???.txt"로 저장하면 된다. 두번째는 발현값(Raw Data(RC))으로 표현하고자 할 때 Raw Data 를 체크하고 샘플이 3 개 이상이면 z-score 를 체크하고 샘플이 2 개면 median 을 체크하고 Export Data Select 에서 Heatmap 에 표현할 비교그룹을 체크하여 "Data Export"를 클릭한 후 "???.txt"로 저장하면 된다

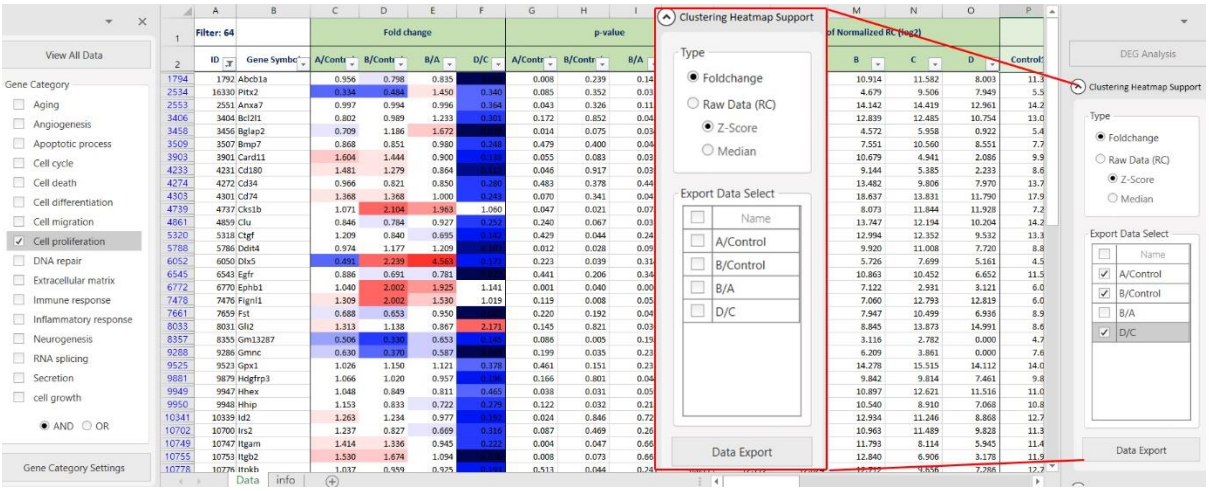


그림 1-16. Clustering Heatmap Support

1-5. Selected Gene Plot & Gene Search 사용 방법

ExDEGA 의 기능 중에 선별한 유전자 또는 연구자가 관심있는 유전자들을 대상으로 발현패턴을 그래프로 표현하고자 할 때는 "Selected Gene Plot" 기능을 사용하면 된다.

선별한 유전자의 gene symbol 을 복사하여 Selected Gene Plot 창에 붙여 넣고 "Expression Plot View"를 누르면 normalized RC(log2) 값, fold change 값으로 line graph 가 그려진다(그림 1-17).

그리고 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 모든 유전자가 검색되어 필터링 된다(그림 1-18).

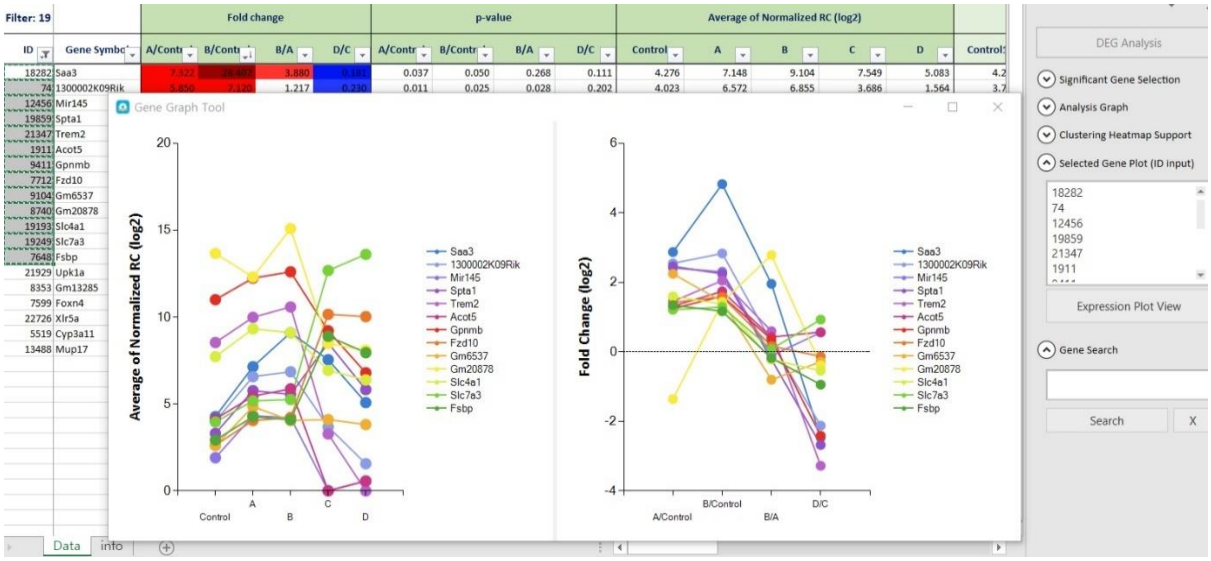


그림 1-17. Gene graph

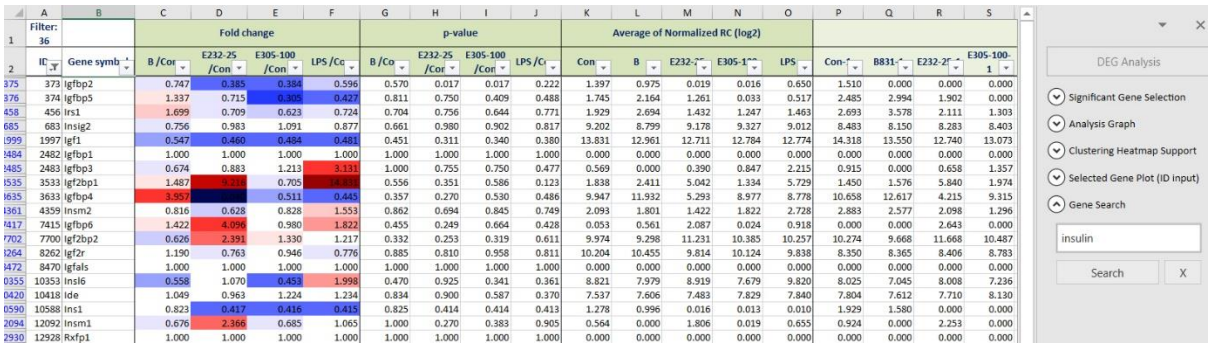


그림 1-18. Genes related to insulin

2. Web 기반 Gene Set Enrichment 분석

2-1. DAVID tool을 이용한 Functional Annotation 분석

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

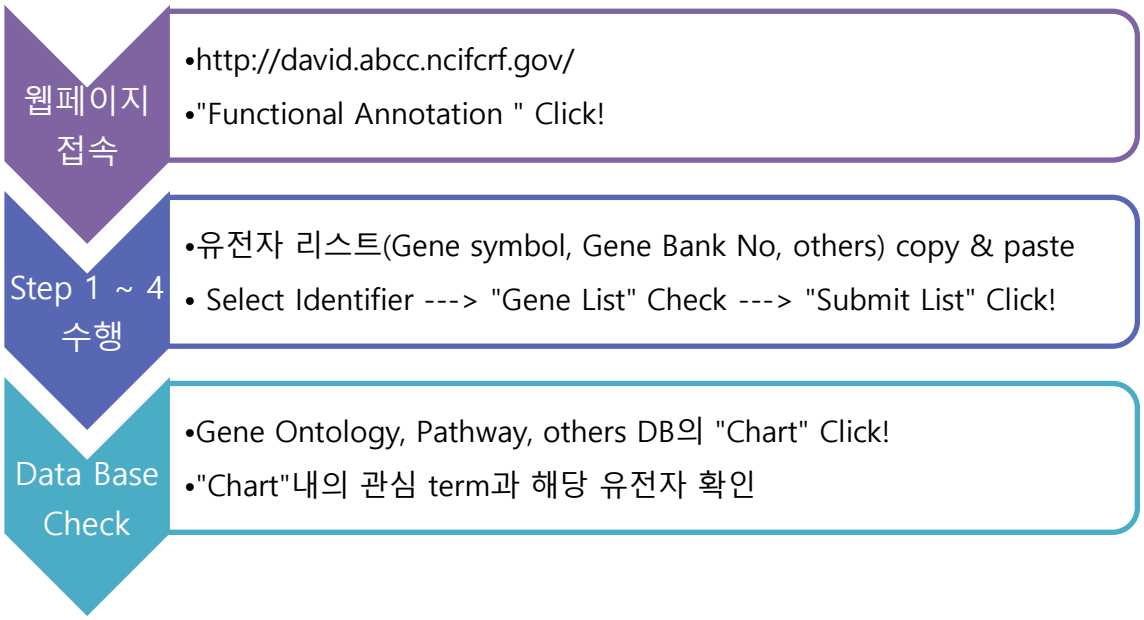


그림 2-1. DAVID tool analysis process

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

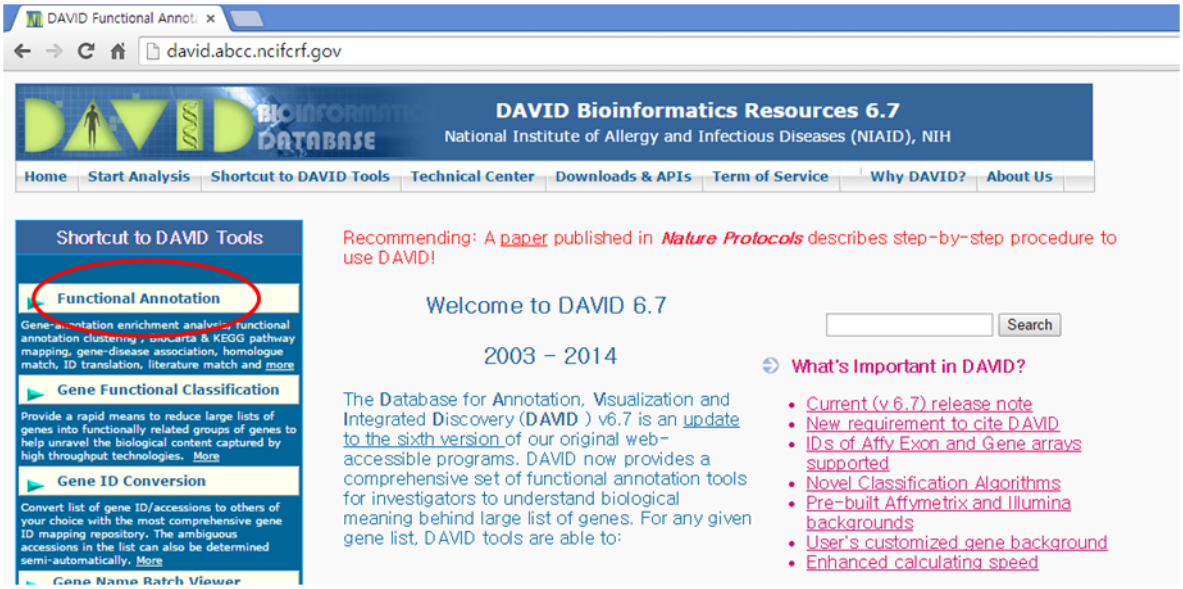


그림 2-2. DAVID tool webpage

"Upload" 탭에서 Step 1 에서 Step 4 까지 수행한다(그림 2-3). Step 1 에서선별한 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.

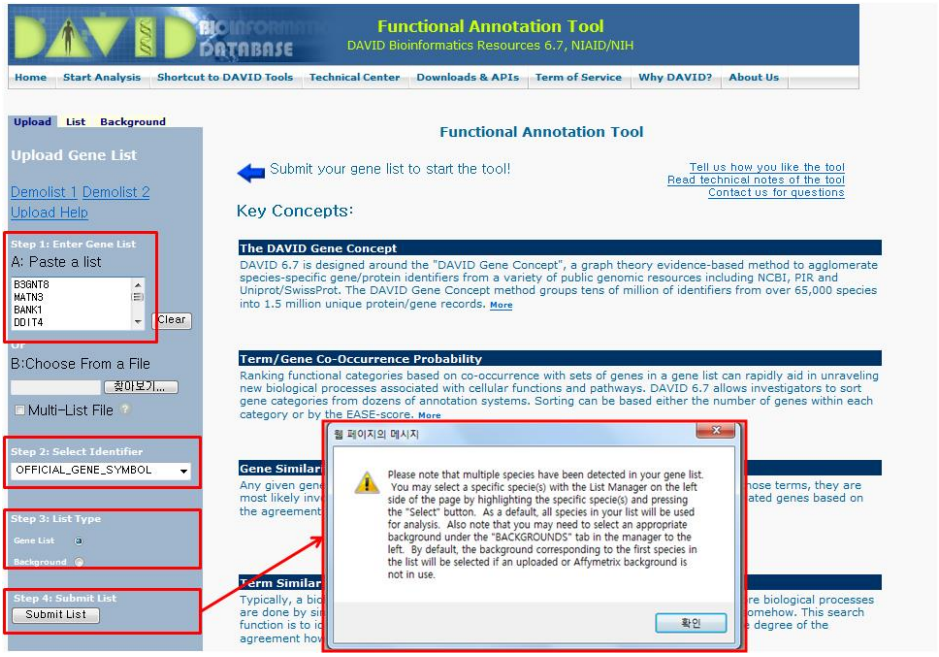


그림 2-3. DAVID tool : Step 1 ~ Step 4

실험한 종을 선택하고 "Select Species"를 누르면 해당 종의 유전자를 대상으로 다시 분석된다. 예시에서는 160개의 유전자 리스트를 넣었지만 데이터베이스에서 기능이 밝혀진94개이기에 최종 94 개 유전자를 대상으로 Functional Annotation 분석이 완료되었다(그림 2-4).

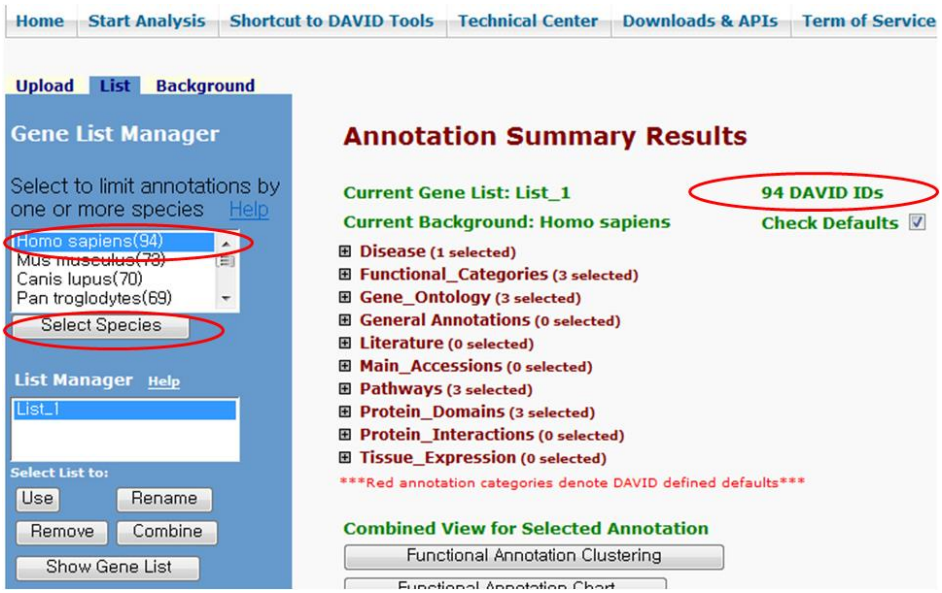


그림 2-4. DAVID tool : Select Species

분석 결과를 확인하기 위해 예로 Gene Ontology 중 Biological Process 를 확인한다."Gene_Ontology"의 "+" 표시를 클릭하여 결과 창을 열고 "GOTERM_BP_FAT"의 "Chart"를 누르면 94 개 유전자들이 관여하는 Biological Process 에 속하는 GO 를 확인할 수 있다(그림 2-5). 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.

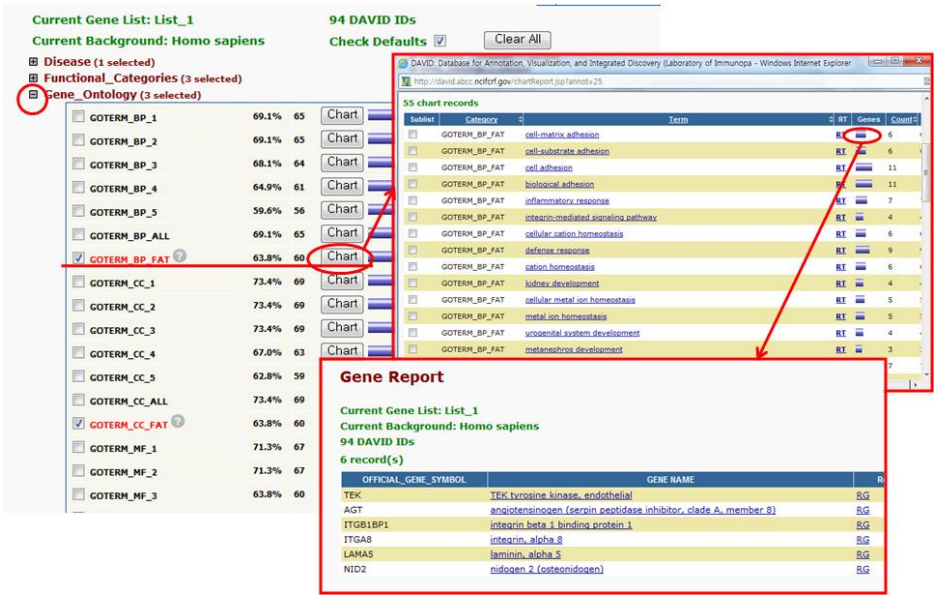


그림 2-5. DAVID tool : exploring Gene Ontology analysis result

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(160 개) 중 해당 pathway 에 관여하는 유전자이다. 유전자를 클릭하면 유전자 정보를 자세히 알 수 있다.

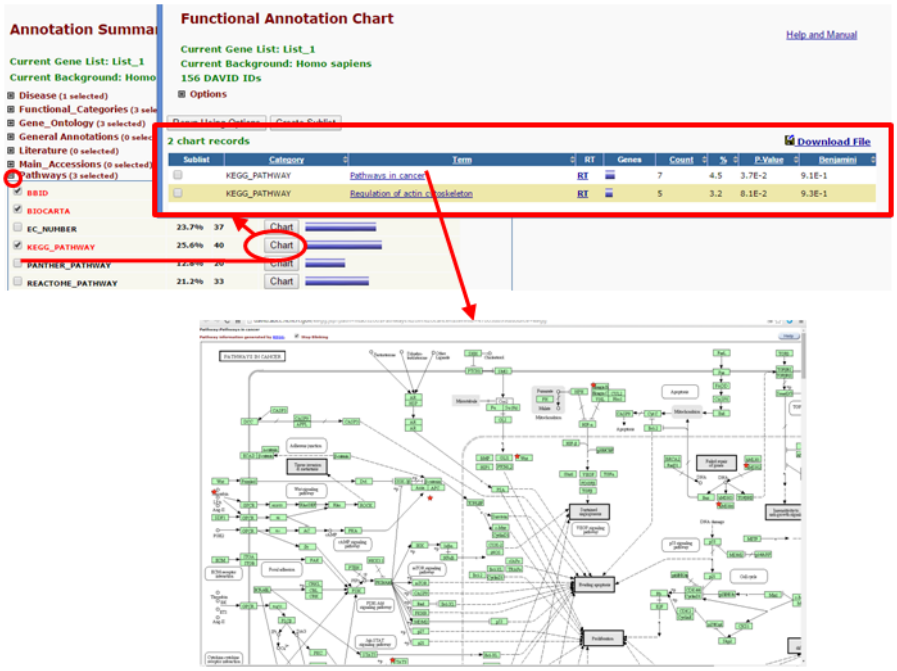


그림 2-6. DAVID tool : exploring Pathway analysis result

DAVID 분석은 input 한 유전자들이 유의하게 관련되는 GO, pathway 등을 분석하는 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 조정할 수 있다. David 분석 결과의 각 항목은 DAVID 홈페이지의 Help and Tool Manual 에 자세히 설명되어 있다(그림 2-7).

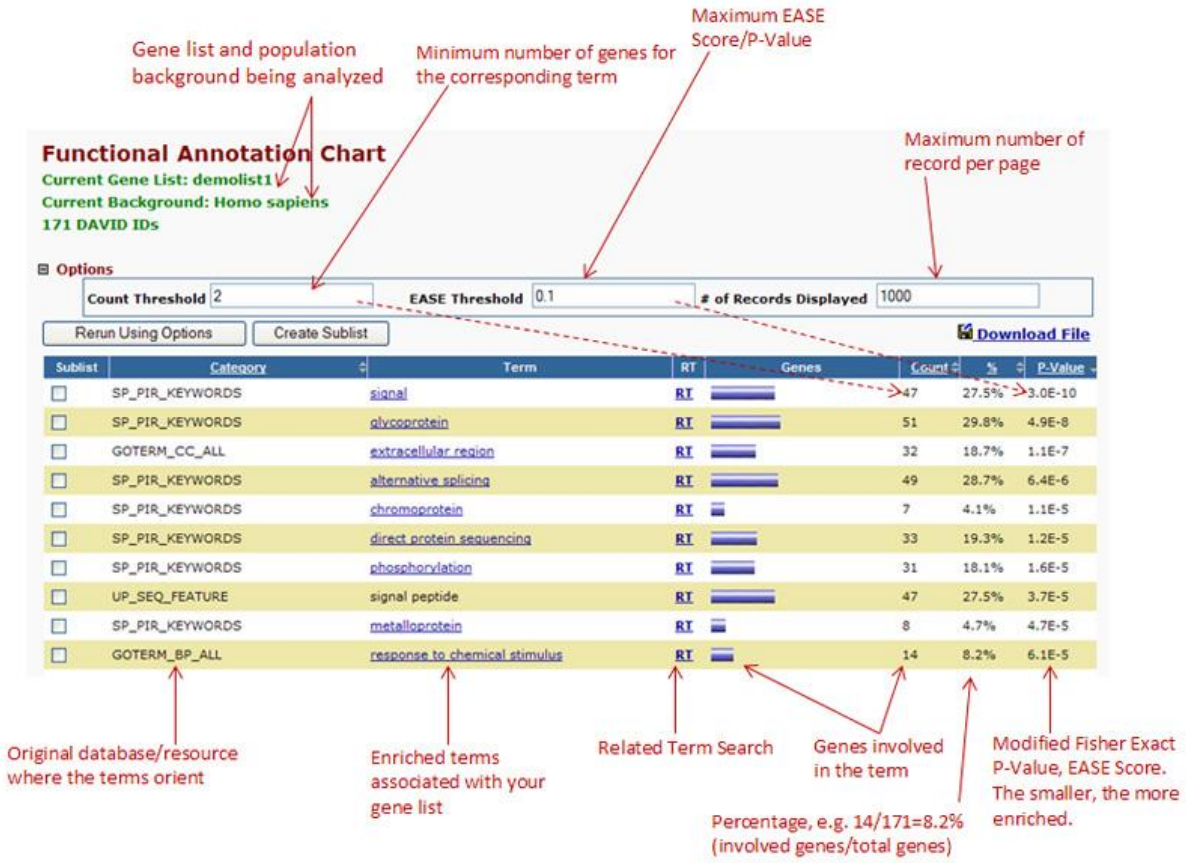


그림 2-7. DAVID Help and Tool Manual

2-2. String-db tool을 이용한 gene set분석

String-db tool 은 Protein-Protein Interaction 데이터 베이스를 기반으로 유전자의 상관관계를 통계적으로 분석하여 유전자의 주요 기능을 예측하고 Network 을 build 해 주는 분석툴이다. 분석과정은 그림 2-2-1 과 같다.

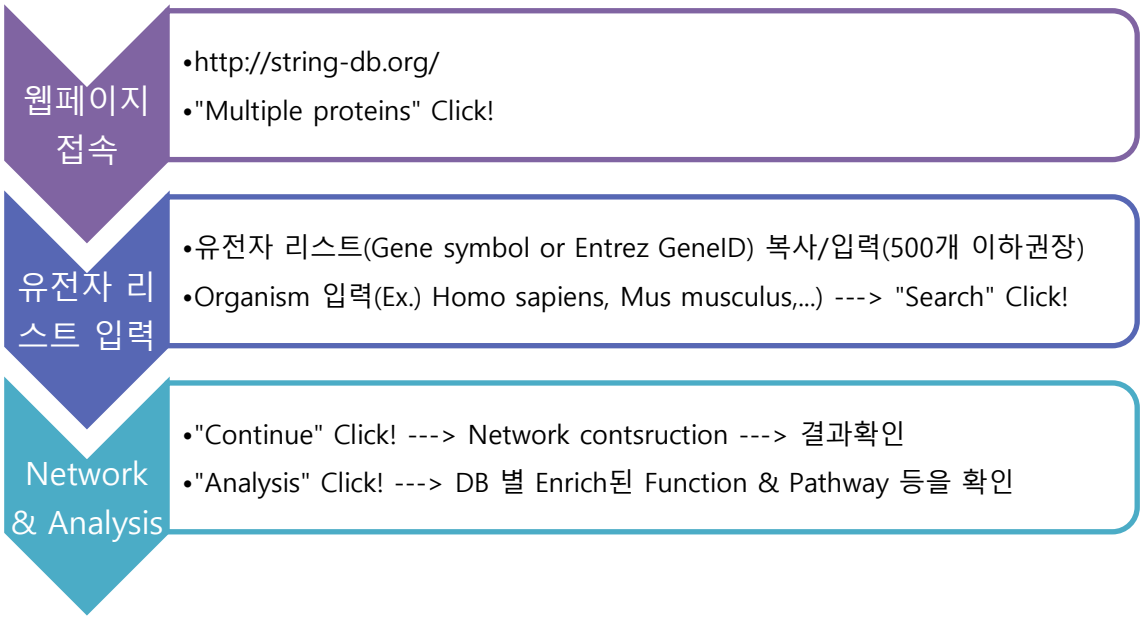


그림 2-2-1. String-db tool analysis process

String-db 에서는 500 개 이하의 유전자를 input 하는 것을 권장하고 있고 여러 public ID 중 EntrezGeneID 사용이 좀더 편리하다. mRNA-Seq 결과에서 significant gene 을 선별하고 String-db 홈페이지 (http://string-db.org/)에 접속하여 "Multiple proteins"을 클릭하고 "List of names" 입력창에 유전자 리스트를 복사한다.그리고 "Organism" 입력창에 해당 species 학명을 입력하고 "Search"를 클릭한다(그림 2-2-2).

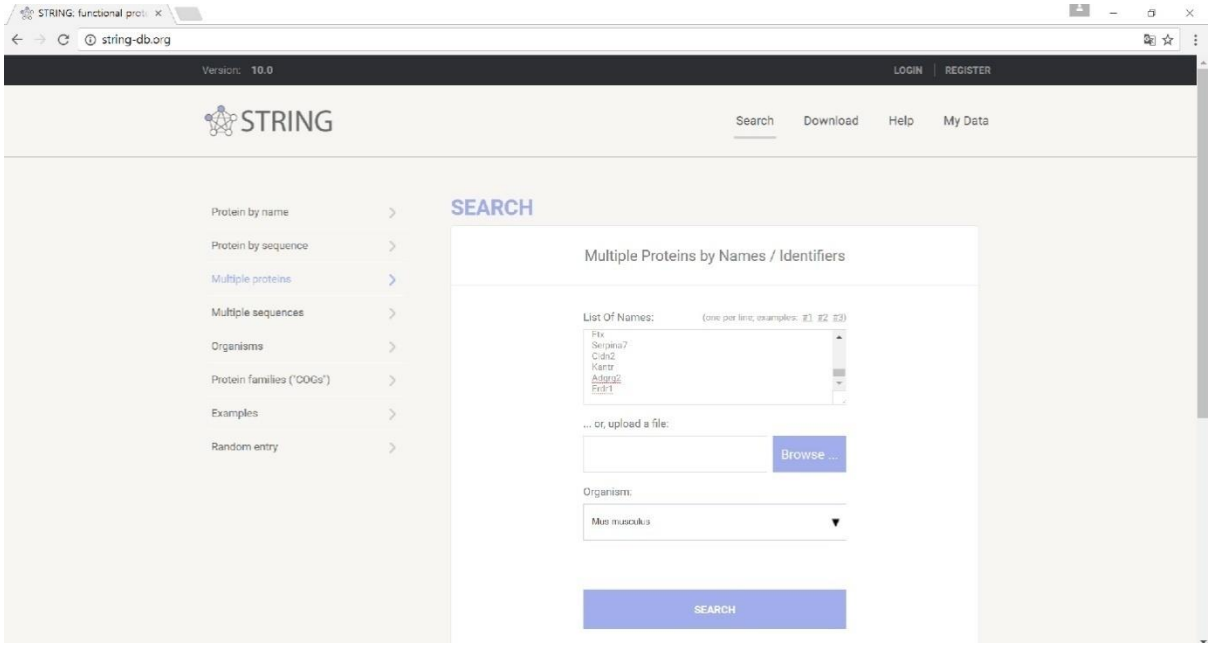


그림 2-2-2. Multiple proteins search

"Search" 결과 중간에 아래 그림과 같은 유전자 확인 단계가 있고 별 이상이 없으면 "continue"를 클릭하여 계속 진행한다(그림 2-2-3).

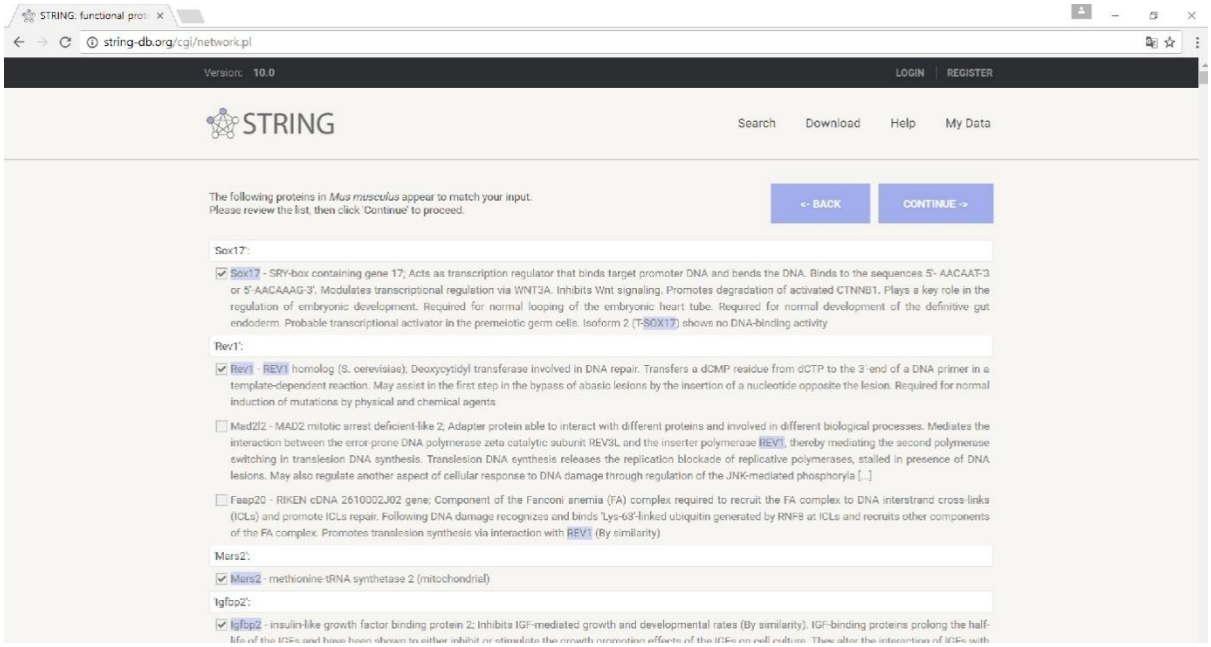


그림 2-2-3. Gene confirmation step

분석이 완료되면 그림 2-2-4와 같이 String DB 기반 Network 결과를 확인할 수 있고 "Analysis" 탭을 클릭하면 "Functional enrichments in your network" 결과를 확인할 수 있다(그림 2-2-5).
각 Functional DB 결과의 오른쪽 하단에 "more"를 클릭하면 FDR<0.05 이하에 해당하는 항목을 모두 볼 수 있다.

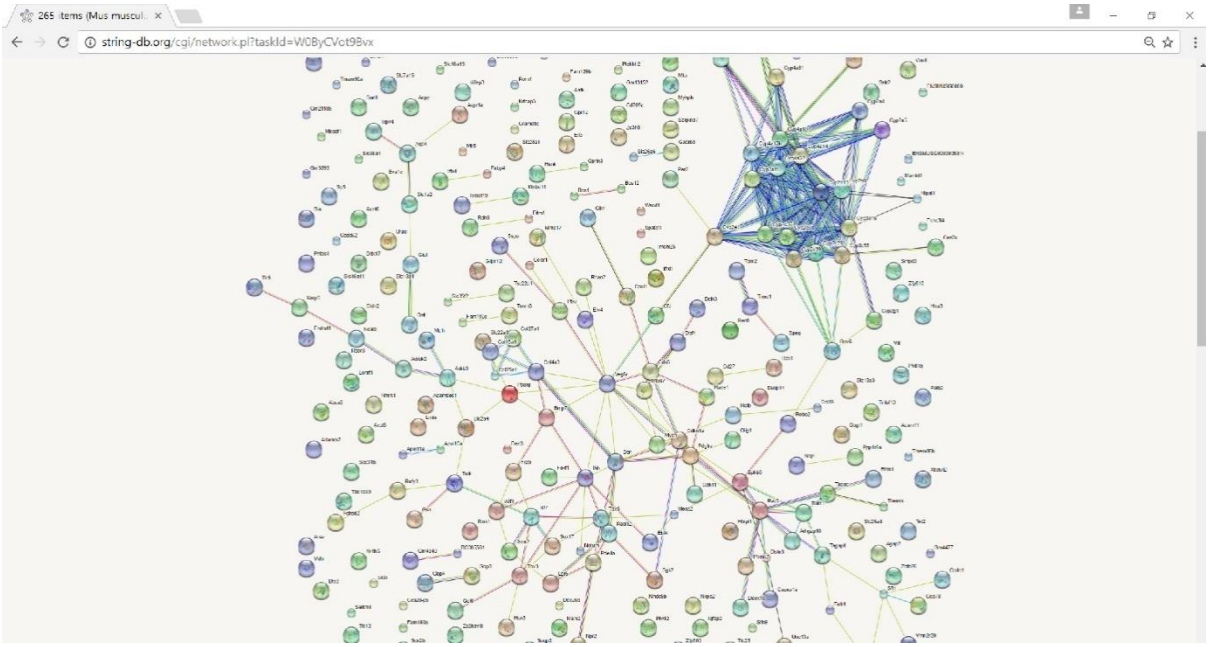


그림 2-2-4. String network result

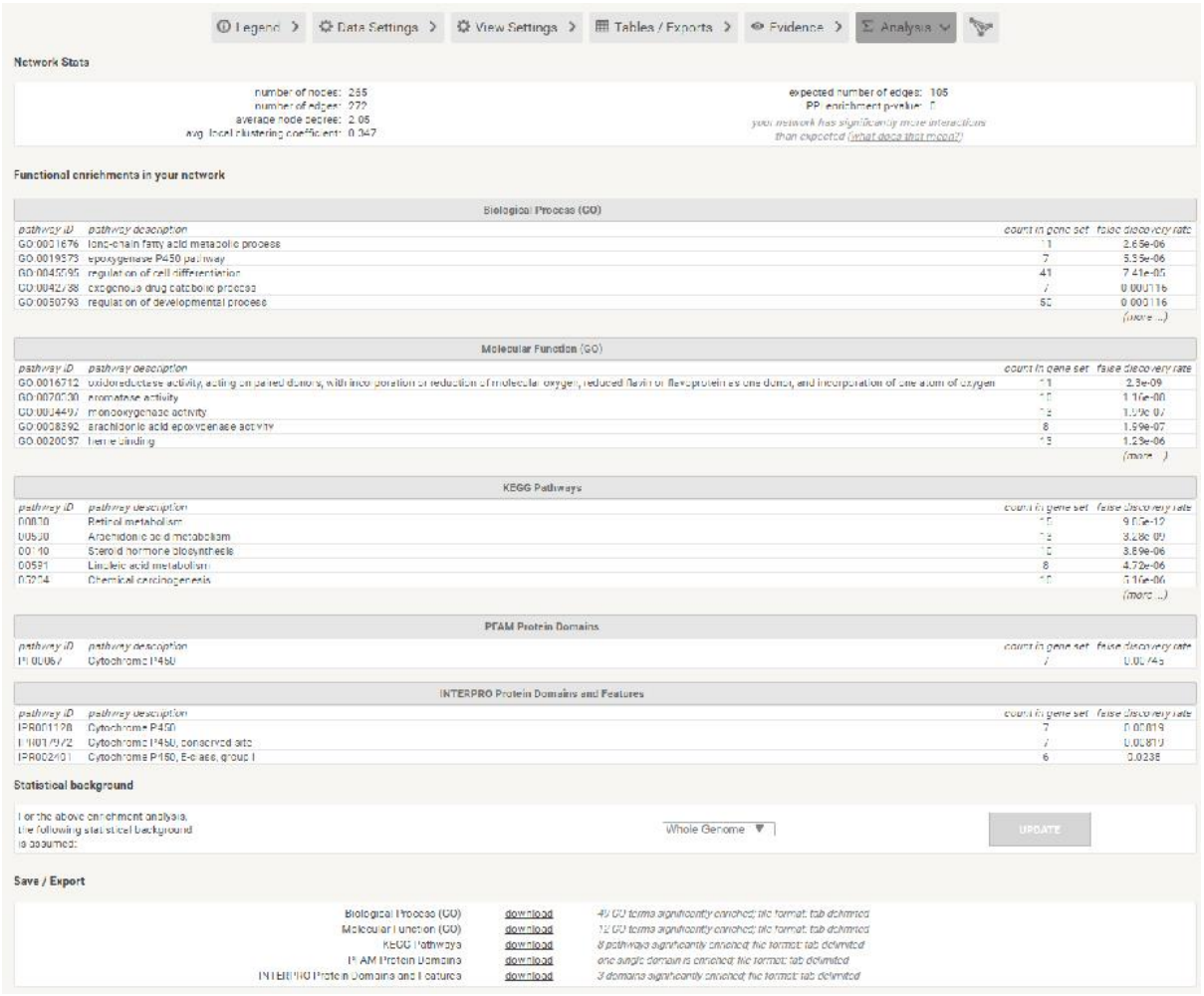


그림 2-2-5. Functional enrichments result

관심 있거나 중요한 Function을 클릭하면 Network상에서 해당 유전자들이 붉은색으로 표시되고(그림 2-2-6) 관심 있는 유전자를 클릭하면 해당 유전자의 자세한 정보를 추가로 얻을 수 있다(그림 2-2-7).

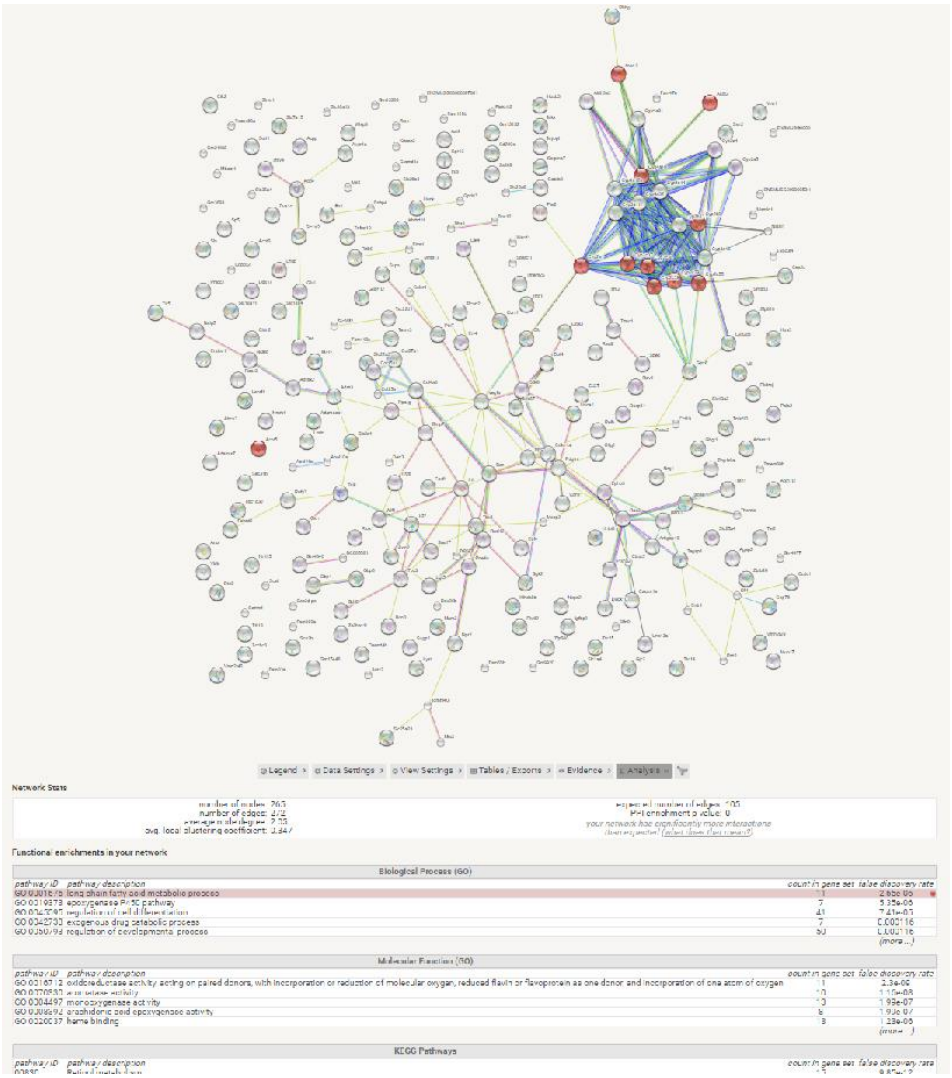


그림 2-2-6. Function selection on your network

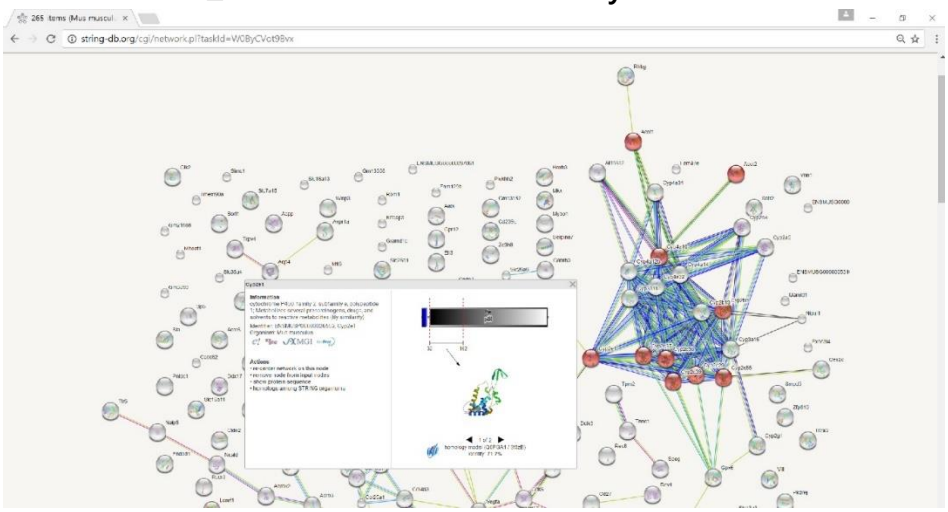


그림 2-2-7. Gene selection on your network

"Legend" 탭에서는 Node, Edge, Input 유전자의 설명을 자세히 볼 수 있고(그림 2-2-8)"Tables/Exports" 탭에서는 Network와 유전자 정보를 파일로 저장할 수 있다.(그림 2-2-9)

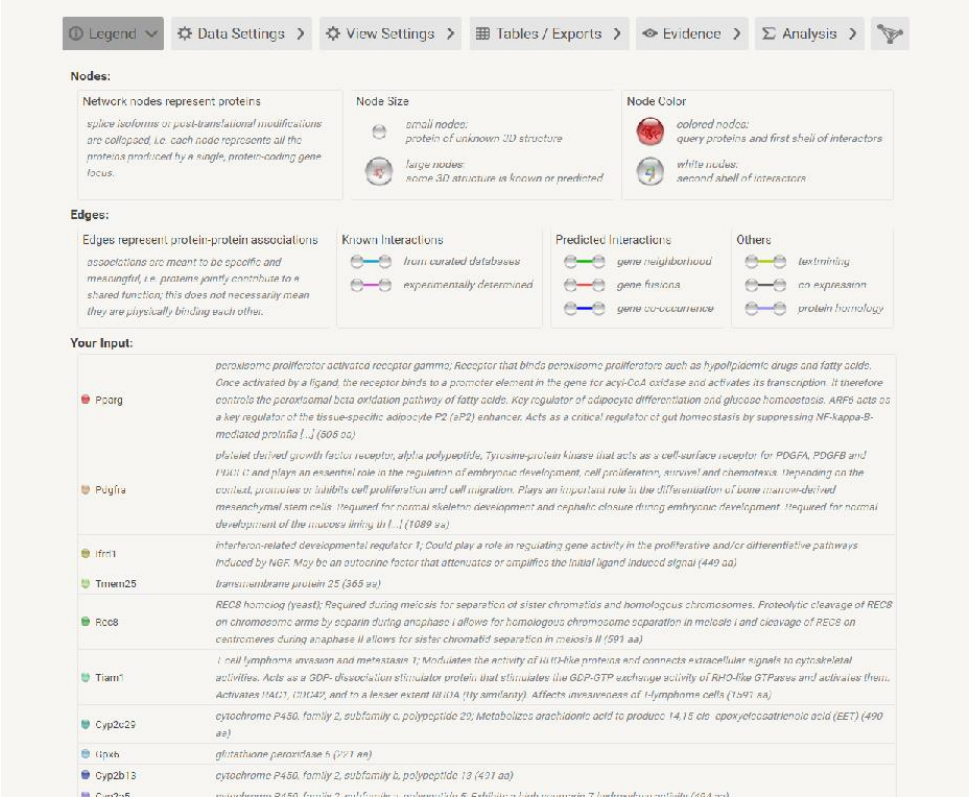


그림 2-2-7. Legend of your network

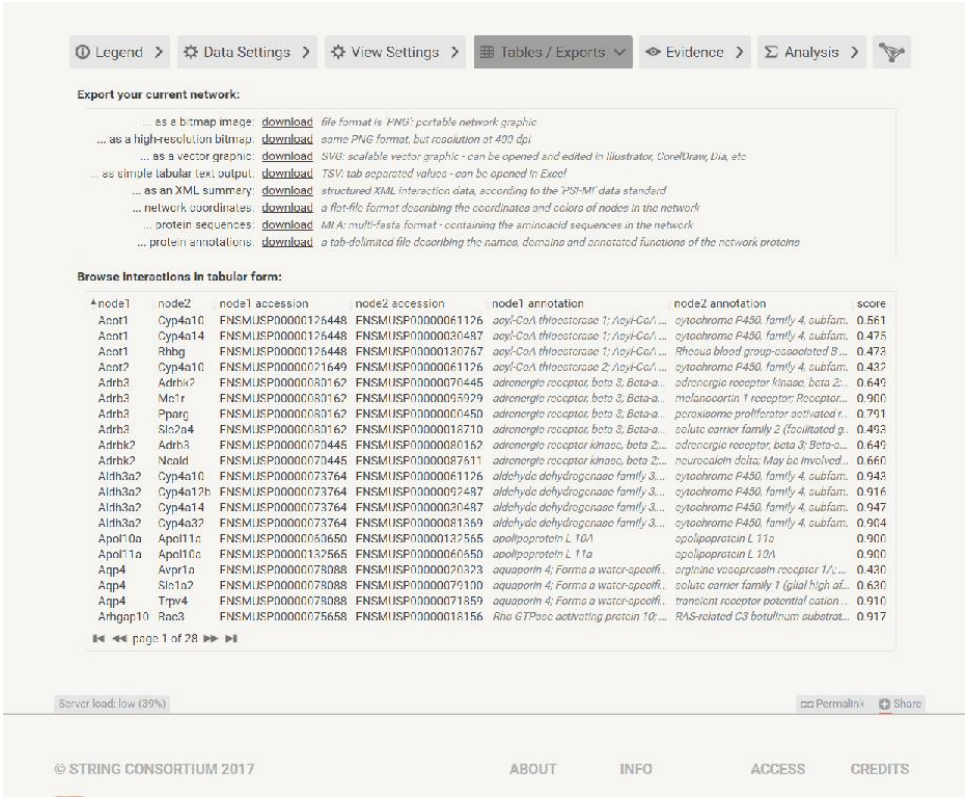


그림 2-2-8. Tables/Exports of your network

2-3. MSigDB기반 GSEA 분석

GSEA 분석은 MSigDB 기반으로 유전자의 상관관계를 통계적으로 분석하여 입력한 유전자 셋의 주요 기능을 예측하고 각 유전자가 어떤 기능들에 포함되는지 overlap 분석을 제공해 준다. 분석과정은 그림 2-3-1 과 같다.

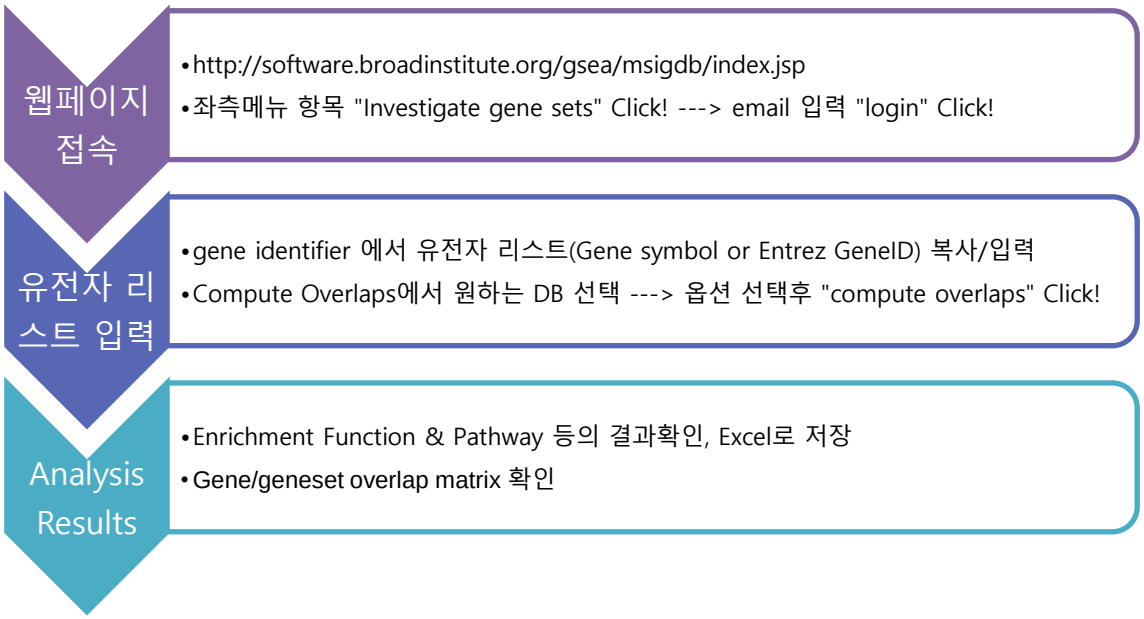


그림 2-3-1. Web based GSEA tool analysis process

MSigDB 에 접속하여 "Investigate gene sets"을 클릭하고 등록한 이메일을 입력하여 로그인을 수행한다.(그림 2-3-2).만약 등록이 필요할 시 "Click here"을 클릭하여 등록을 진행하면 된다.(그림 2-3-3).

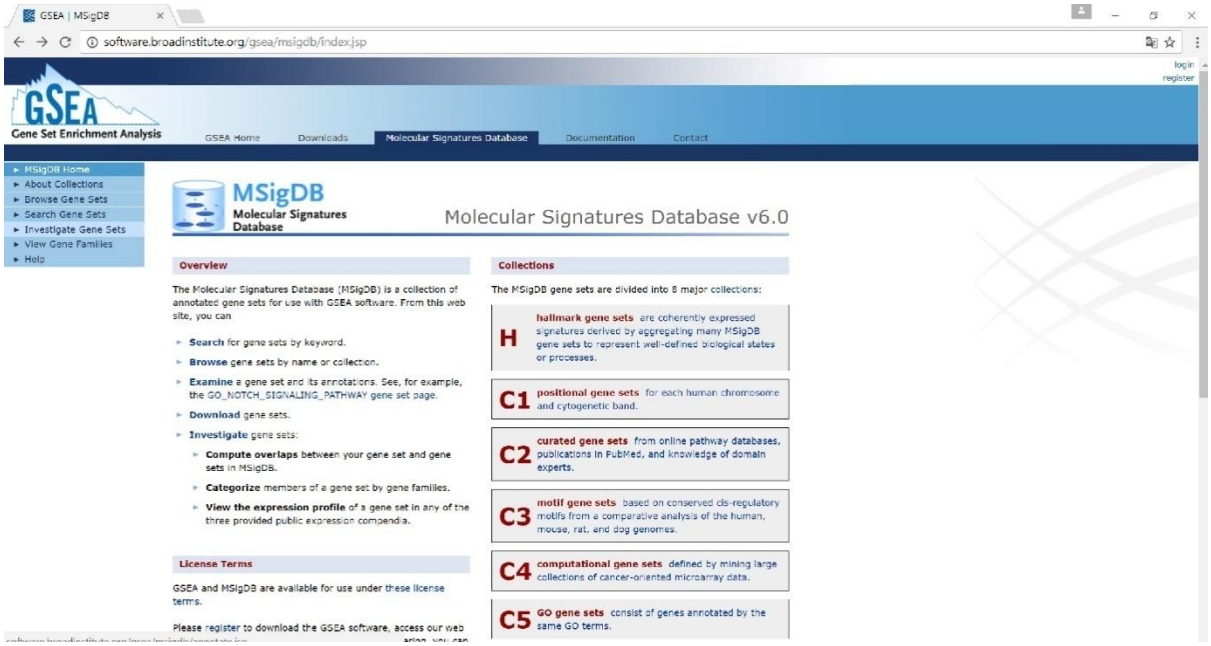


그림 2-3-2. GSEAmain page

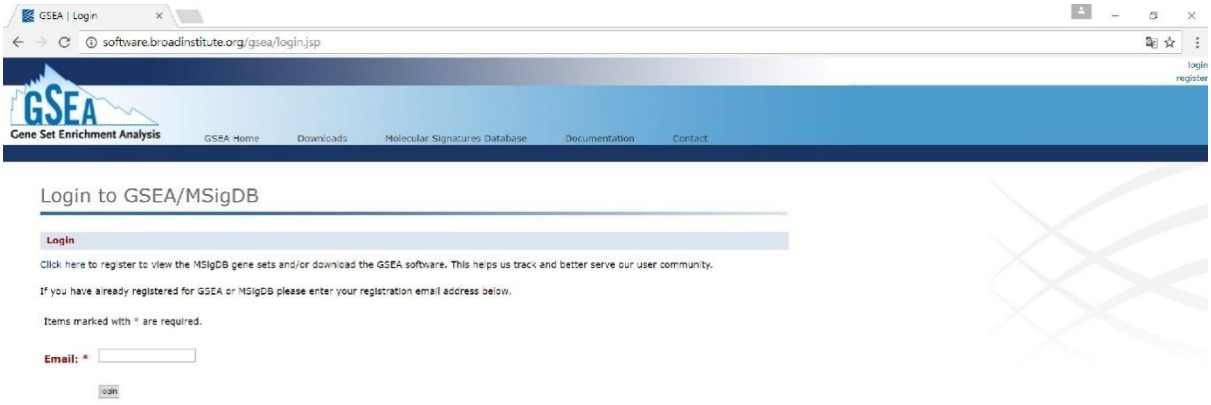


그림 2-3-3. GSEALogin page

"Gene Identifiers"입력창에 유전자 리스트(Gene Symbol, EntrezGeneID 또는 public ID)를 입력하고 "Compute Overlaps"에 원하는 DB 를 클릭한 후 맨 아래 "compute overlaps" 버튼을 클릭한다.(그림 2-3-4).DB 선택시 DB 명 앞의 파란색 글자를 누르면 해당 DB 정보를 확인할 수 있다.

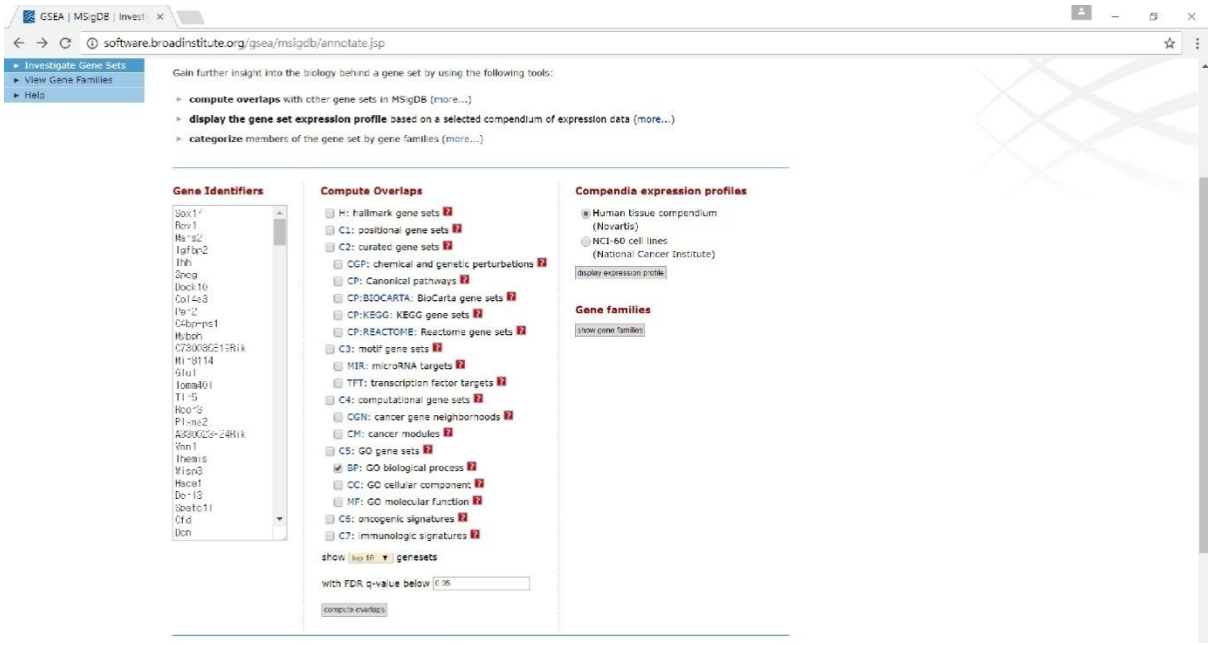


그림 2-3-4. GSEAAnalysis

분석이 완료되면 그림 2-3-5 와 그림 2-3-6 과 같이 통계적으로 유의한 Gene Set List 와 Gene/Gene-set Overlap Matrix 결과를 확인할 수 있다.

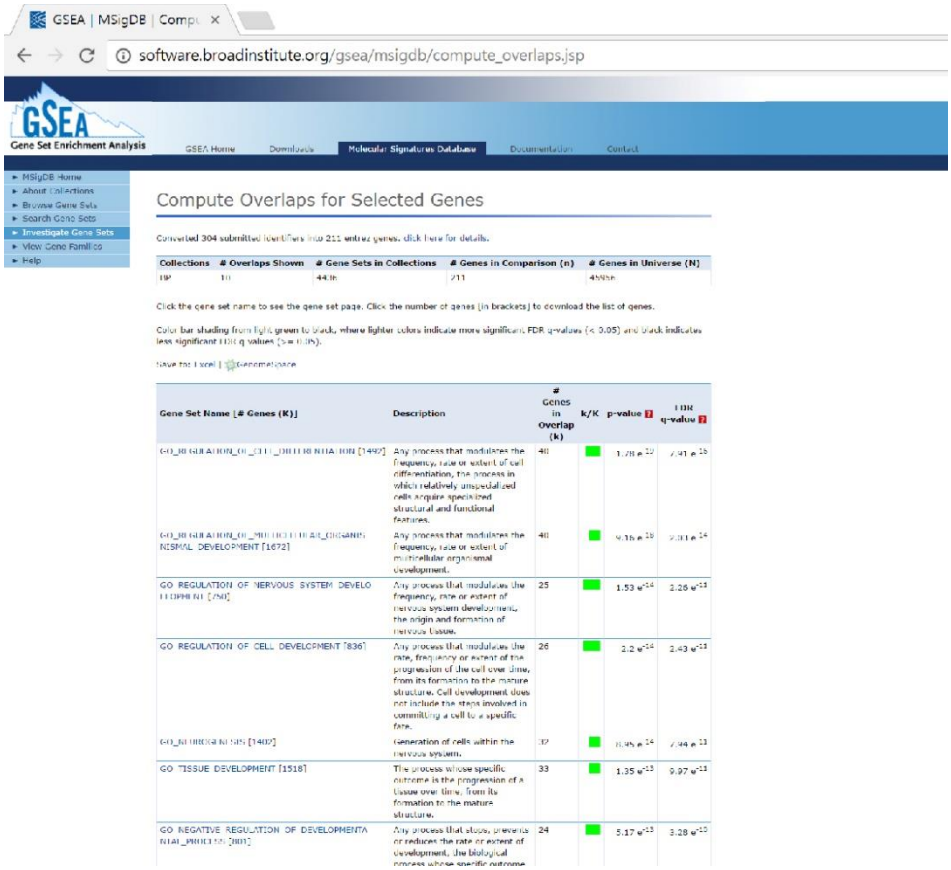


그림 2-3-5. GSEAAnalysisResult(Gene Set)

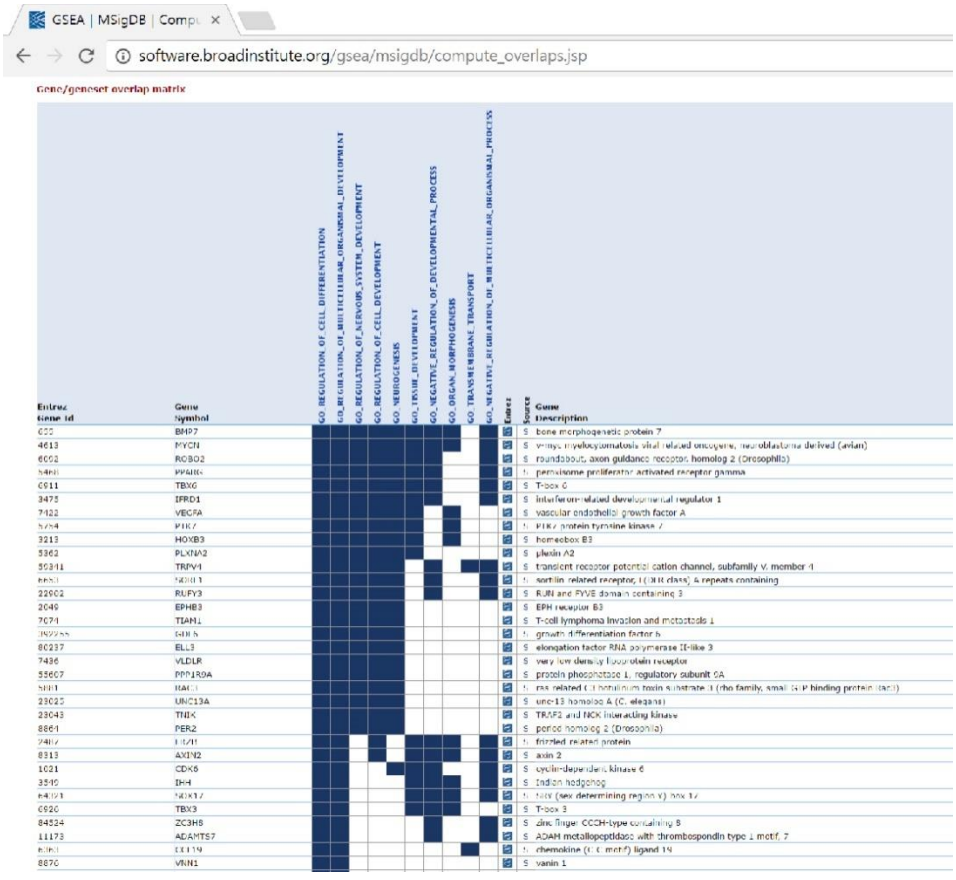


그림 2-3-6. GSEAAnalysisResult(Gene/Gene-set Overlap Matrix)

3. KEGG DB 기반 Pathway 분석

mRNA-Seq 분석 결과에서 up/down-regulated 유전자들이 어떤 Pathway에 속하는지 확인하고자 한다면 KEGG에서 제공하는 KEGG Mapper를 이용하면 된다. 사용방법은 그림 3-1과 같은 순서로 진행된다.

그림 3-1. KEGG Mapper tool analysis process

그림 3-2는 mRNA-Seq report에서 2fold, normalized RC(log2)>6을 기준으로 선별한 유전자를 KEGG 분석하는 과정이다.

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

오른쪽 필터에서 Fold change와 Normalized RC (반복실험의 경우 p-value) 값을 지정하고, 확인하고자 하는 Fold change 조합을 선택하여 필터를 적용 한다.

필터를 적용하여 선별 된 유전자의 KEGG input [Entrez ID, FC Color(#숫자,black)] cell을 복사하여, KEGG 분석에 사용할 것이다.

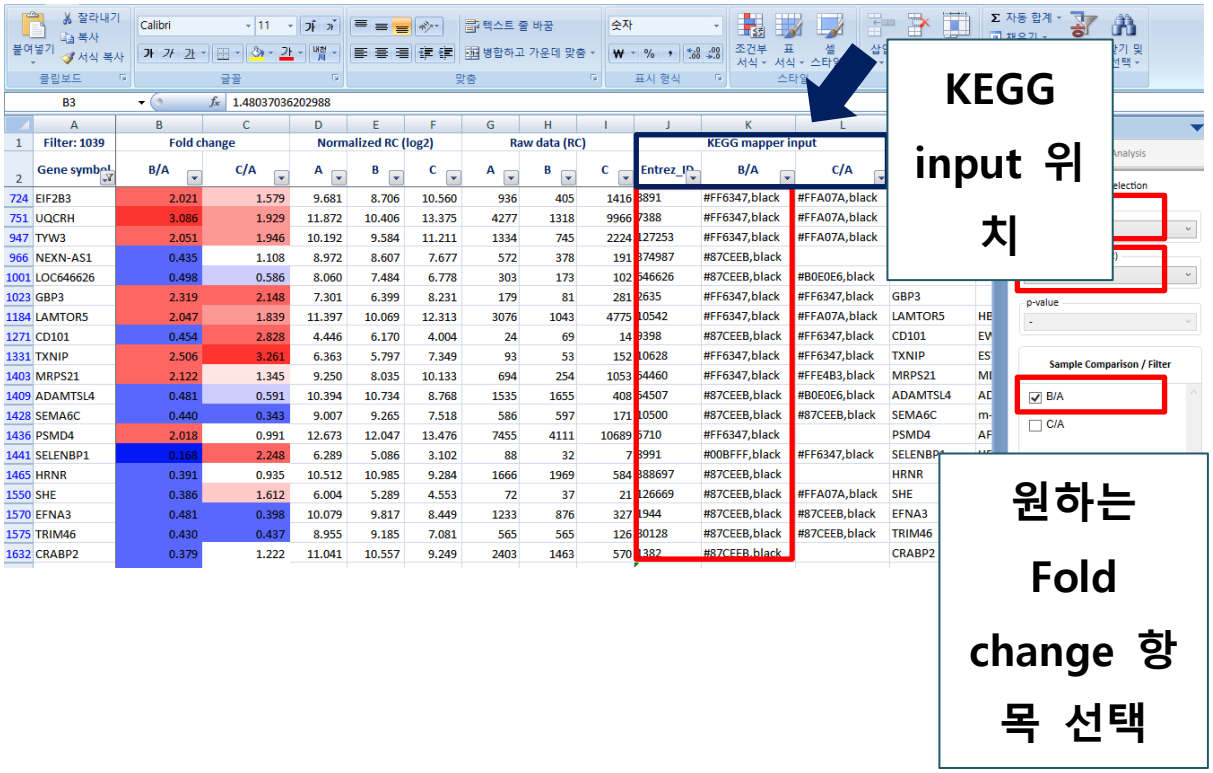


그림 3-2. KEGG Mapper tool analysis process

그림 3-3과 같이 KEGG Mapper 웹페이지(http://www.genome.jp/kegg/tool/map_pathway2.html)에 접속하고 Search & Color pathway 링크에 들어가면 아래와 같은 화면이 보여진다. 분석하고자 하는 유전자의 species를 선택하고, 'primary ID'는 KEGG identifiers로 선택한 뒤 'Enter objects one per line followed bgcolor, fgcolor' 창에 엑셀에서 준비해 놓은 Entrez ID, Color 항목을 복사-붙여넣기를 한다. 마지막으로 "Include aliases"와 "Use uncolored diagram" 항목에 체크를 한 후 Exec 버튼을 누른다.

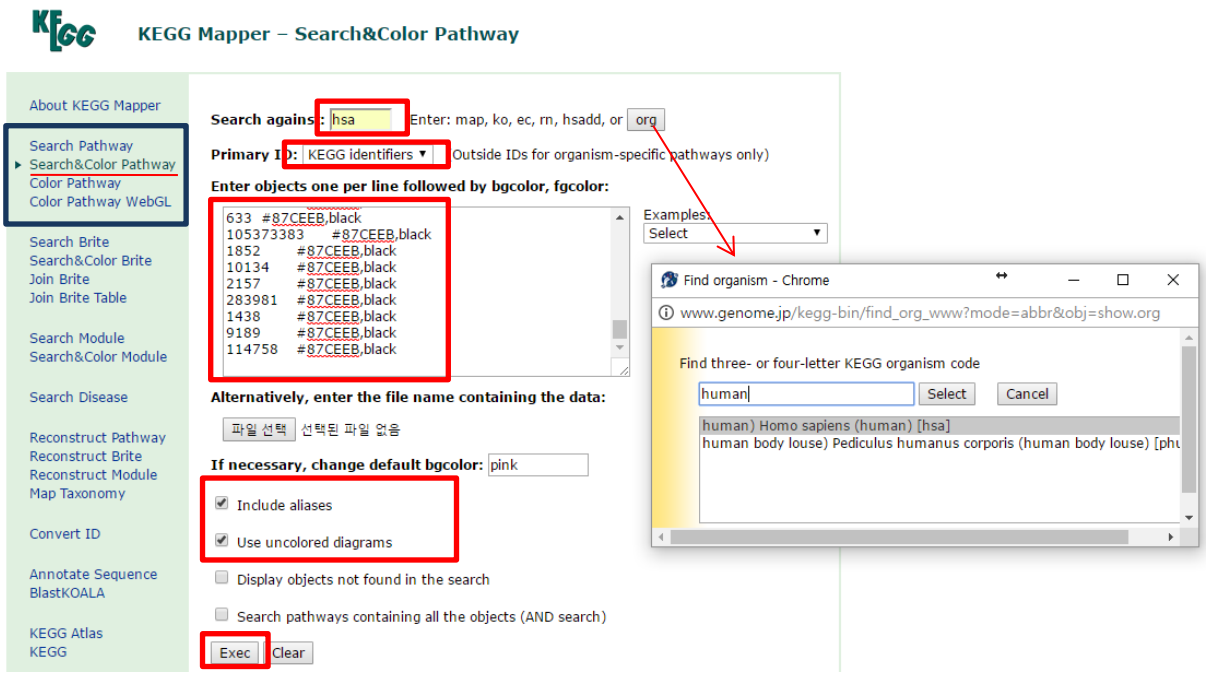


그림 3-3. KEGG Mapper tool analysis process

분석결과, 입력한 유전자들이 관여하는 pathway list가 나온다(그림 3-4). 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/green)이 색으로 표시되어 있다. Pathway 이미지는 "다른 이름으로 저장"이 가능하고 "html"으로 저장하면 이미지에 링크된 항목을 그대로 유지해서 저장이 가능하다.

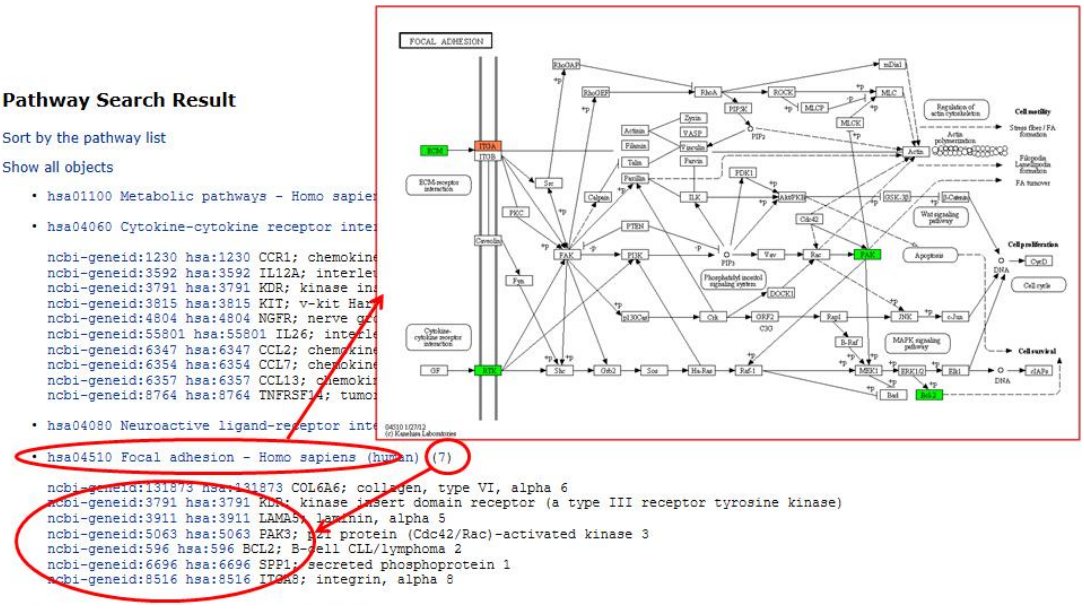


그림 3-4. KEGG Mapper tool analysis result

4. MeV Software 이용 Clustering Heatmap 작성

MeV 소프트웨어는 미국의 Dana-Farber Cancer Institute에서 개발한 Microarray, mRNA-Seq 전용 분석 프로그램으로 연구자들에게 무료로 공급하고 있다. 주로 clustering 분석과 통계분석(K-means clustering, Hierarchical clustering, t-test, Significance Analysis of mRNA-Seqs, Gene Set Enrichment Analysis, EASE)을 할 수 있는 프로그램이다. 아래 웹페이지에 접속하면 최신의 업데이트된 프로그램과 매뉴얼을 다운받을 수 있다.

http://www.tm4.org >> 오른쪽 Browse 항목내 "TM4 MeV Stand-Alone Client" 클릭

프로그램을 다운받아 압축을 풀고, MeV 또는 TMEV를 클릭해서 프로그램을 실행시킨다(그림4-1).MEV프로그램을 실행시키면 세 개의 창이 나타난다(그림4-2). 분석창은 프로그램창의 메뉴에서 file->New multiple array viewer를 통해 여러개를 생성할 수 있고 데이터 분석은 분석창을 통해 진행한다.

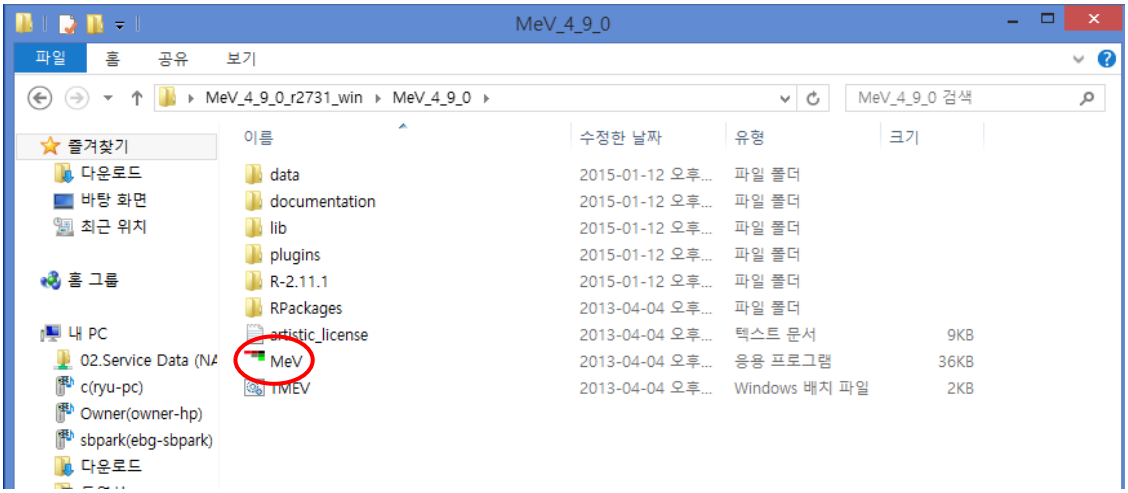


그림 4-1. MeV program folder and files

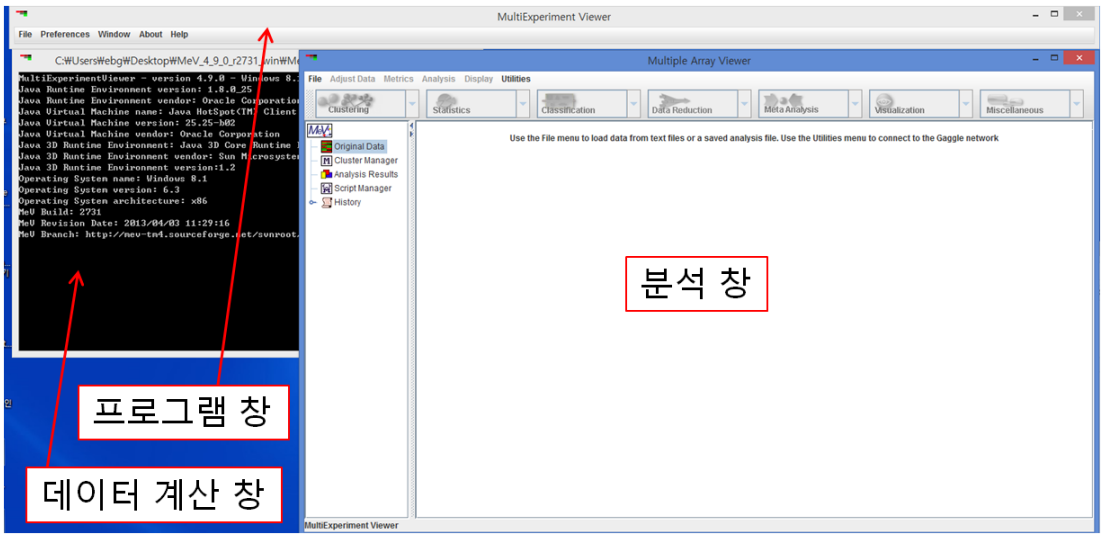
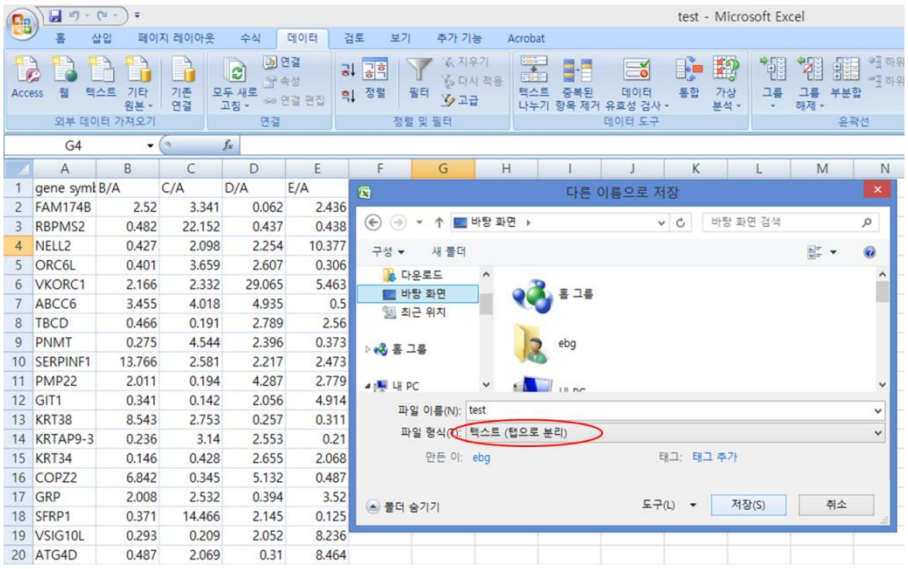


그림 4-2. MeV program windows

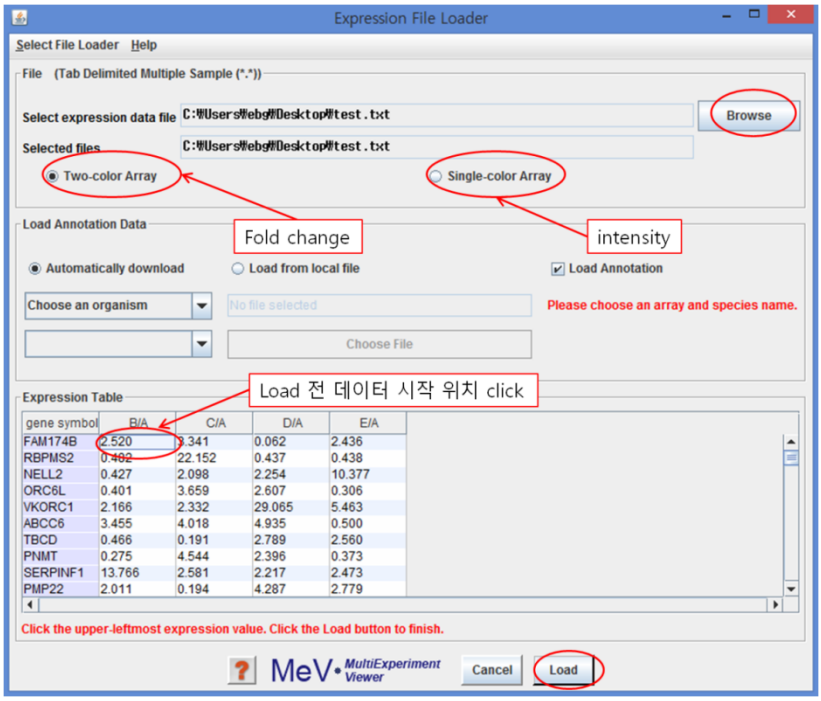
본 자료에서는 MeV 프로그램을 이용하여 Clustering 분석 방법을 설명한다. 우선 MeV 프로그램에 input할 데이터를 엑셀에서 파일 양식에 맞춰 저장해야 한다. 엑셀에 clustering 하고자 하는 유전자 이름과 fold change 또는 발현값(intensity)를 정리한다(그림 4-3). 그리고 '텍스트 (탭으로 분리)'파일 형식으로 저장해야 MeV에 upload 할 수 있다. MeV에서는 2만 개 이상의 유전자는 clustering 분석을 할 수 없으므로 2만 개 이하로 유전자를 선별해야 한다.



	gene symbol	B/A	C/A	D/A	E/A
2	FAM174B	2.52	3.341	0.062	2.436
3	RBPMS2	0.482	22.152	0.437	0.438
4	NELL2	0.427	2.098	2.254	10.377
5	ORC6L	0.401	3.659	2.607	0.306
6	VKORC1	2.166	2.332	29.065	5.463
7	ABCC6	3.455	4.018	4.935	0.5
8	TBCD	0.466	0.191	2.789	2.56
9	PNMT	0.275	4.544	2.396	0.373
10	SERPINF1	13.766	2.581	2.217	2.473
11	PMP22	2.011	0.194	4.287	2.779
12	GIT1	0.341	0.142	2.056	4.914
13	KRT38	8.543	2.753	0.257	0.311
14	KRTAP9-3	0.236	3.14	2.553	0.21
15	KRT34	0.146	0.428	2.655	2.068
16	COPZ2	6.842	0.345	5.132	0.487
17	GRP	2.008	2.532	0.394	3.52
18	SFRP1	0.371	14.466	2.145	0.125
19	VSIG10L	0.293	0.209	2.052	8.236
20	ATG4D	0.487	2.069	0.31	8.464

그림 4-3. Data format example

input 데이터 저장이 완료되면 MeV 프로그램의 분석창에서 file -> load data를 실행한다(그림 4-4). Browse를 클릭하여 input 데이터를 선택한다. 데이터가 fold change인 경우 "Two-color Array"로 체크하고 데이터가 intensity인 경우는 "Single-color Array"에 체크한다. 마우스로 데이터가 시작되는 부위를 클릭한 후 load를 누른다.



gene symbol	B/A	C/A	D/A	E/A
FAM174B	2.520	3.341	0.062	2.436
RBPMS2	0.482	22.152	0.437	0.438
NELL2	0.427	2.098	2.254	10.377
ORC6L	0.401	3.659	2.607	0.306
VKORC1	2.166	2.332	29.065	5.463
ABCC6	3.455	4.018	4.935	0.500
TBCD	0.466	0.191	2.789	2.560
PNMT	0.275	4.544	2.396	0.373
SERPINF1	13.766	2.581	2.217	2.473
PMP22	2.011	0.194	4.287	2.779

그림 4-4. Data uploading method

데이터가 열리면 Adjust Data -> Log Transformation -> Log2 Transform을 선택하여 fold change는 log2(fold change)로, intensity는 log2(intensity)로 바꿔준다(그림 4-5). 왼쪽 메뉴의 Original Data -> Expression image를 보면 log2 값으로 바뀌어 색이 변한 것을 확인할 수 있다.

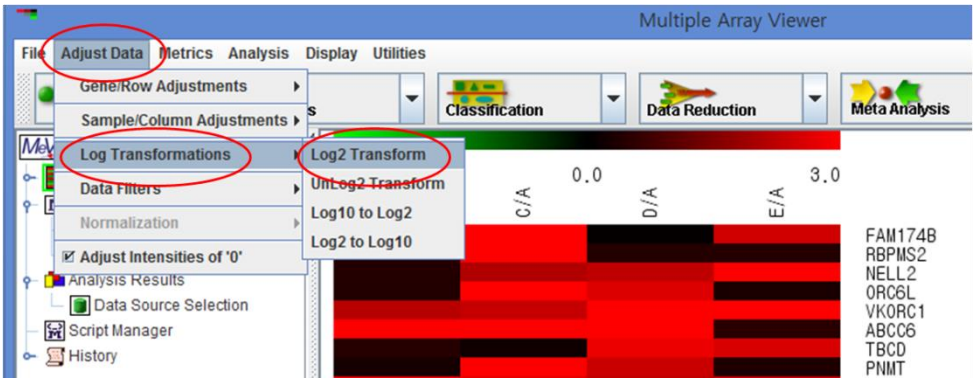


그림 4-5. Log2 transformation

Analysis-> Clustering-> HCL을 선택하여 Clustering 분석을 시작한다(그림 4-6).

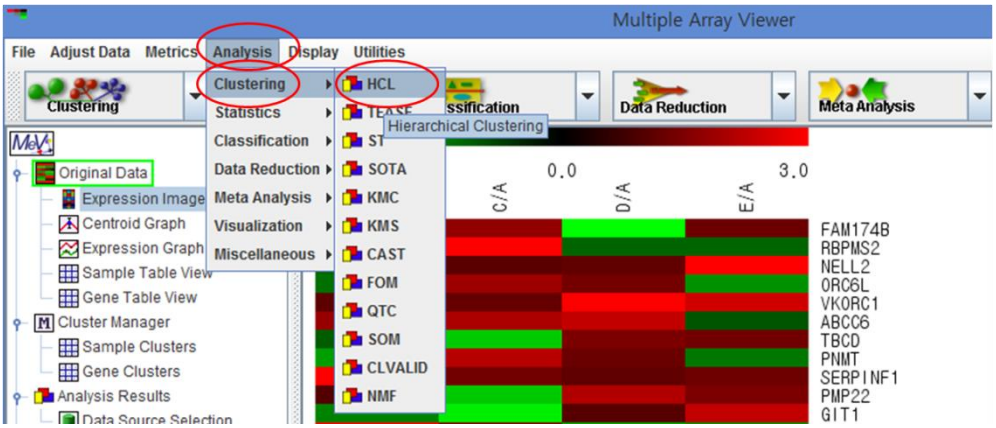


그림 4-6. Hierarchical Clustering Selection

Clustering 분석 시 다양한 옵션을 선택할 수 있다(그림 4-7). Gene tree를 선택하면 fold change 또는 intensity가 유사한 유전자끼리 clustering한 결과가 나온다. Sample tree를 선택하면 발현이 유사한 샘플끼리 clustering한 결과가 나온다.당사에서 clustering 분석을 할 때 Distance Metric는 Euclidean Distance로 Linkage Method Selection은 Average linkage clustering으로 설정한다. 다른 옵션을 선택해도 된다. 옵션을 선택하고 OK를 누른다.

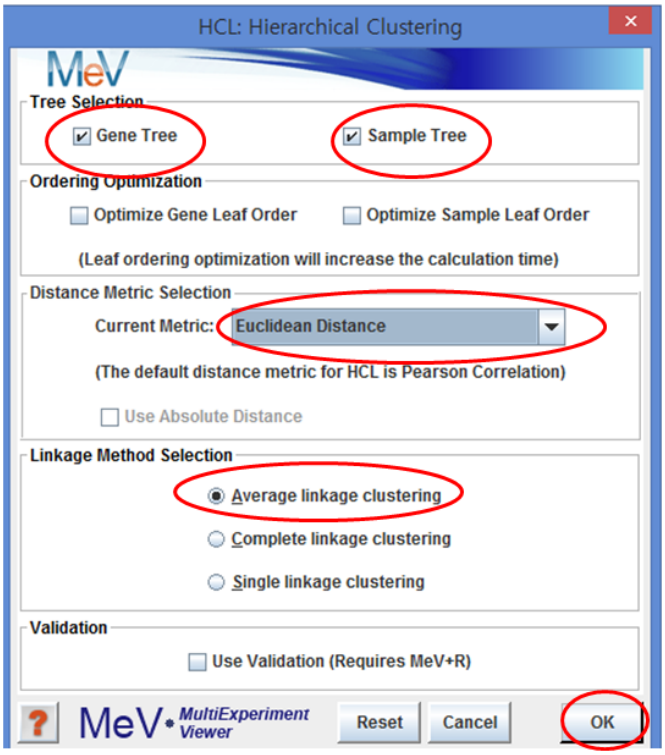


그림 4-7. Hierarchical Clustering Method

clustering이 완료되면 왼쪽 메뉴에 Analysis Results에 HCL 결과가 생긴다. HCL -> HCL tree를 클릭하면 clustering 결과가 화면에 나온다(그림 4-8). 위의 tree는 sample clustering 결과이고 왼쪽 tree는 gene clustering 결과이다.각 tree에는 distance scale bar가 있어서 tree의 길이를 가늠할 수 있다. tree의 길이는 distance이며, distance가 짧을수록 유전자 간 또는 샘플 간의 발현이 비슷한 것, 길수록 발현이 다른 것이다.

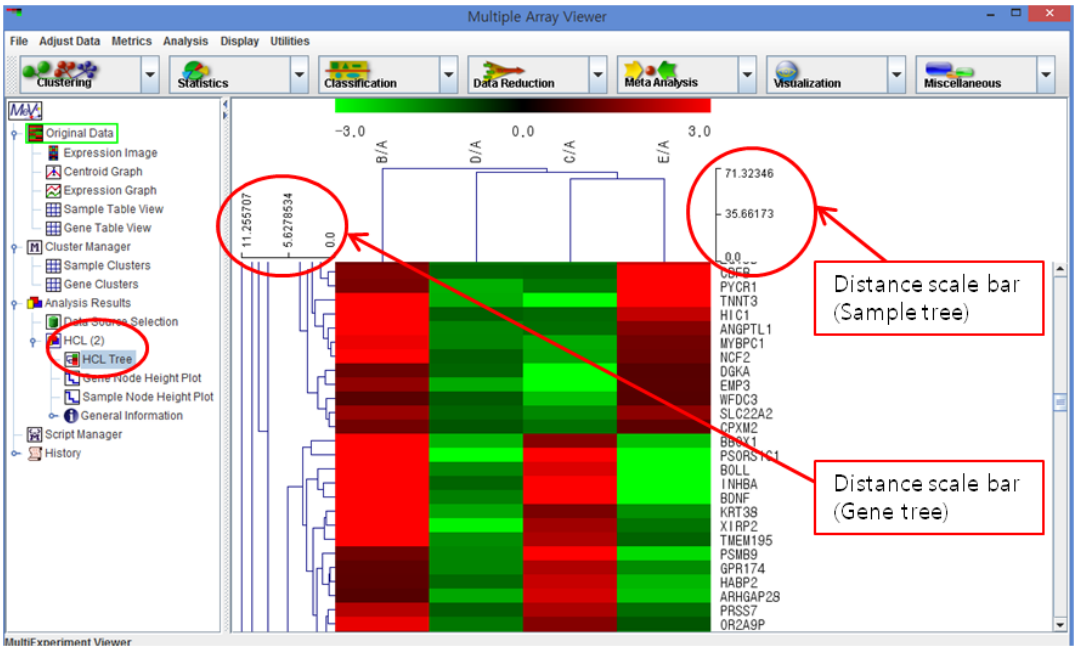


그림 4-8. Hierarchical Clustering Result

clustering 결과는 이미지의 크기와 색상을 조절하여 원하는 형태의 이미지를 만들 수 있다(그림 4-9, 4-10)

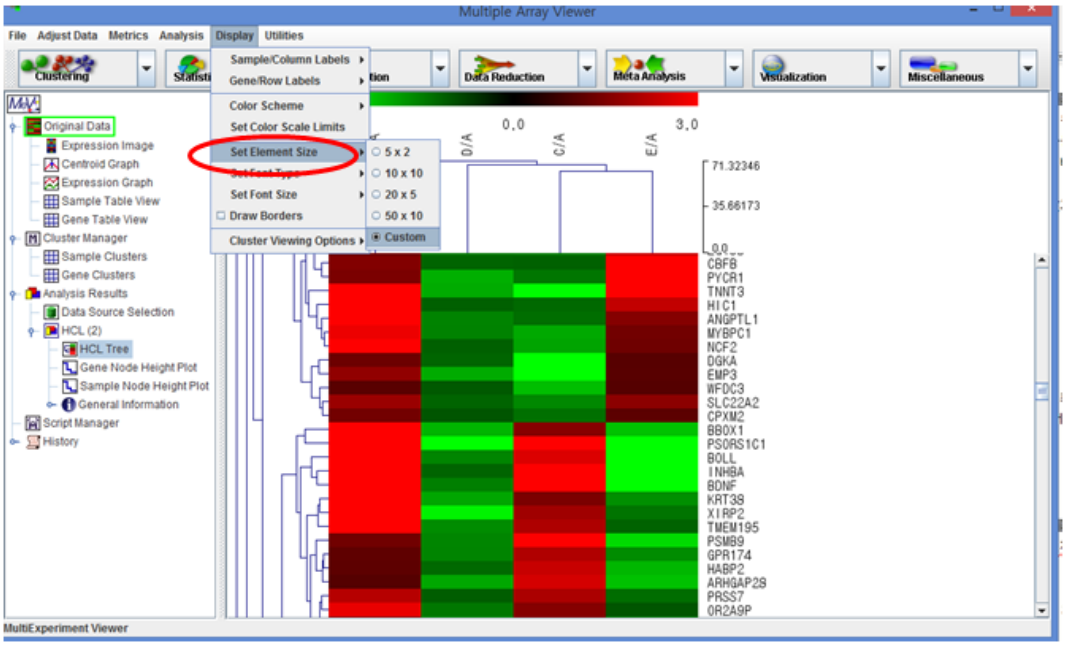


그림 4-9. Clustering image size control

Display -> Set Color Scale Limits을 누르면 color scale bar의 최소값, 중간값, 최대값을 설정할 수 있다. 보통 log2(fold change)는 최소값과 최대값은 같은 크기에 부등호만 바꿔주고(예: min:-3, max:3) 중간값은 0으로 설정해 준다(그림 4-10). 이렇게 하면 up-regulated genes은 red, down-regulated genes은 green으로 나타나게 된다.

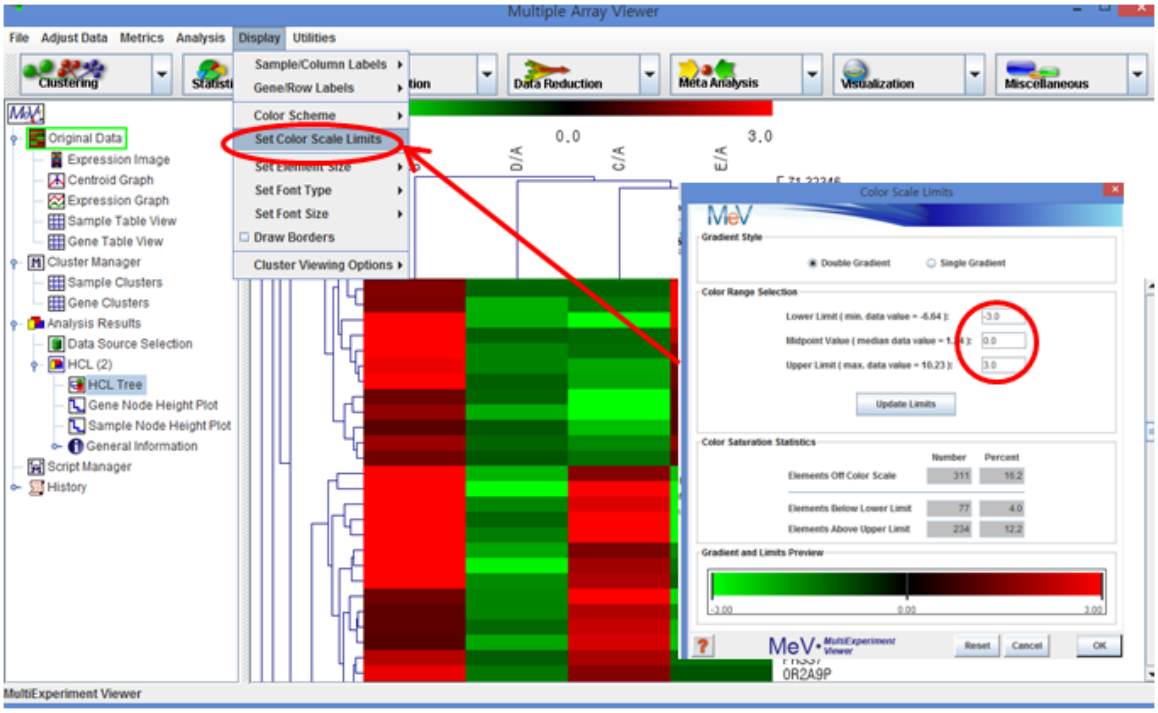


그림 4-10. Clustering image color setting

원하는 이미지 조절이 완료되면 File -> Save image를 눌러 이미지를 저장한다. 이때 파일 이름에 파일 확장자명(예: .jpg)을 꼭 기입하여야 이미지 파일로 저장이 된다(그림 4-11).

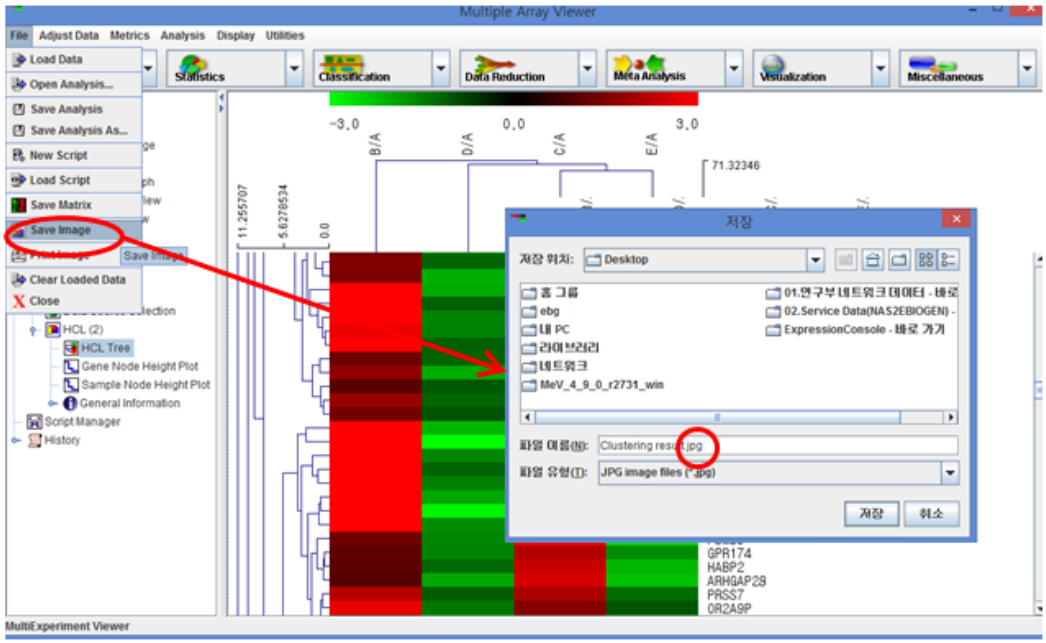


그림 4-11. Clustering image save