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 1-1 과 같다.

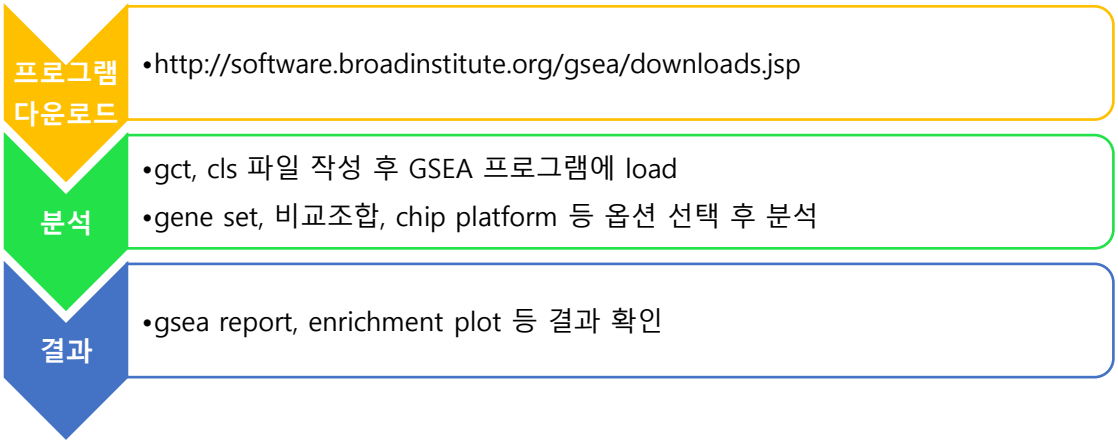


그림 1-1. GSEA tool analysis process

GSEA 홈페이지(http://software.broadinstitute.org/gsea/downloads.jsp)에 들어가 회원가입 후 로그인 하여 GSEA 프로그램을 다운로드 받는다 (그림 1-2).

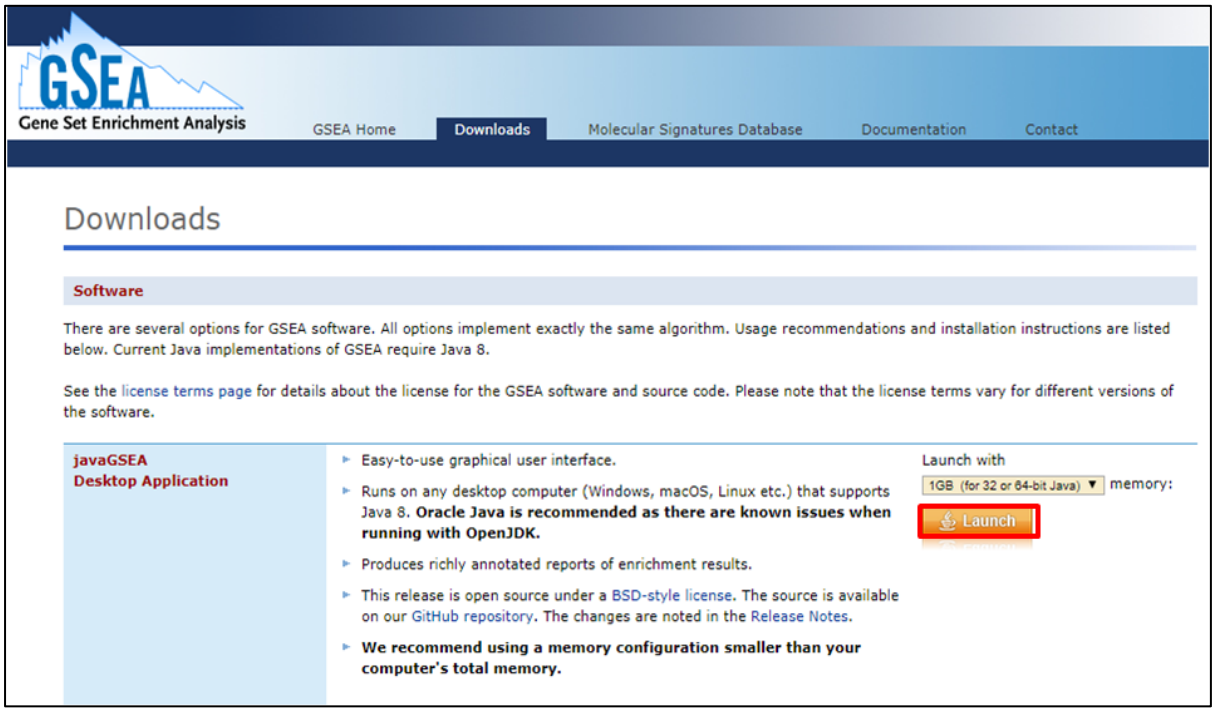
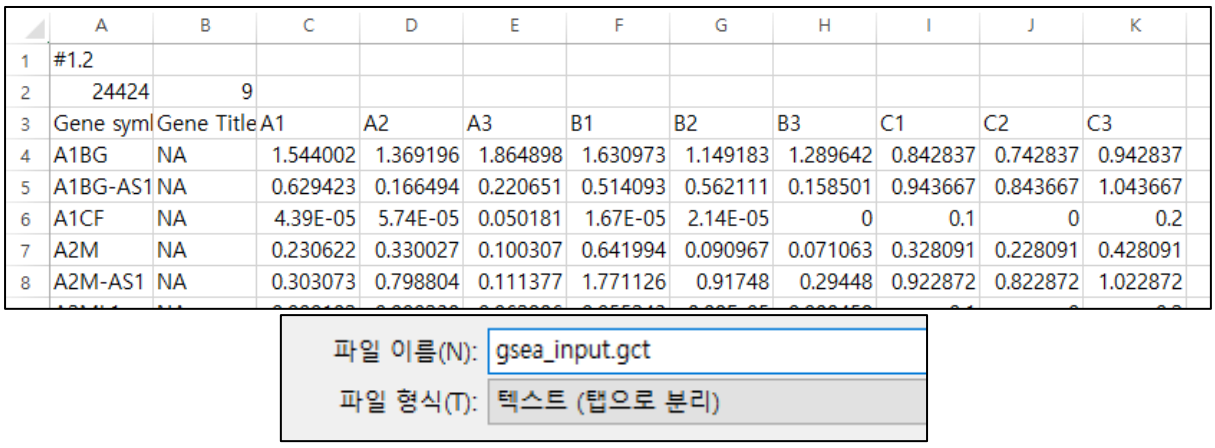


그림 1-2. GSEA program download

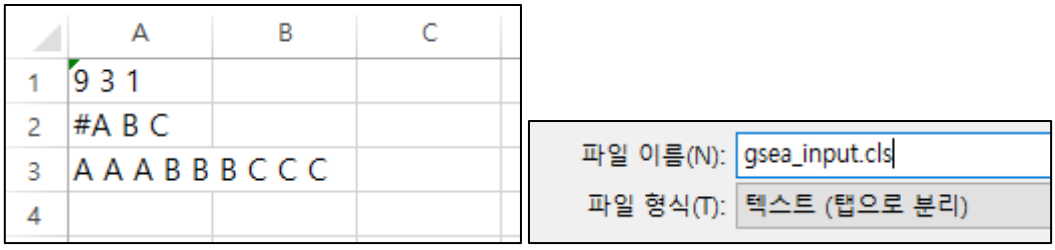
GSEA 분석을 위해서는 유전자 발현값 정보가 포함되어 있는 gct 파일과 샘플 정보가 포함되어 있는 cls 파일이 필요하다. Gct 파일은 그림 1-3과 같은 형식으로 만든다. A1칸에는 항상 "#1.2", A2칸에 유전자 수, B2칸에 샘플 수를 기입한다. RNA-seq data는 A열에 Gene symbol, B열에 Gene title (gene description), C열부터 각 샘플의 normalized data (log2 변환하지 않은 값)을 기입한다. Microarray data는 A열에 probe ID, B열에 Gene 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 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 1-3. gct file

cls 파일은 그림 1-4와 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 1-4. cls file

GSEA 프로그램을 열어 Load data 버튼을 누르고 Browse for files 버튼을 누른 후 gct, cls 파일을 연다(그림 1-5). gct, cls 파일은 파일의 경로가 길면 input 파일을 잘 인식하지 못하여 되도록 바탕화면에 두고 수행한다.

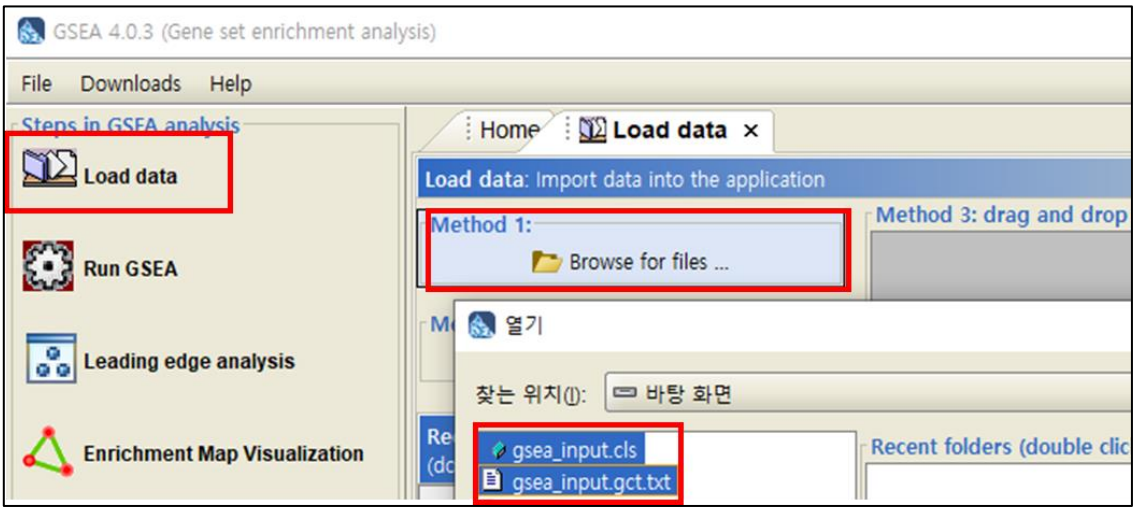


그림 1-5. Load data in GSEA program

Run GSEA 를 누르고 Expression dataset 는 gct 파일명을 선택, gene sets database 는 분석하고자 하는 gene set 을 선택한다(그림 1-6). 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.

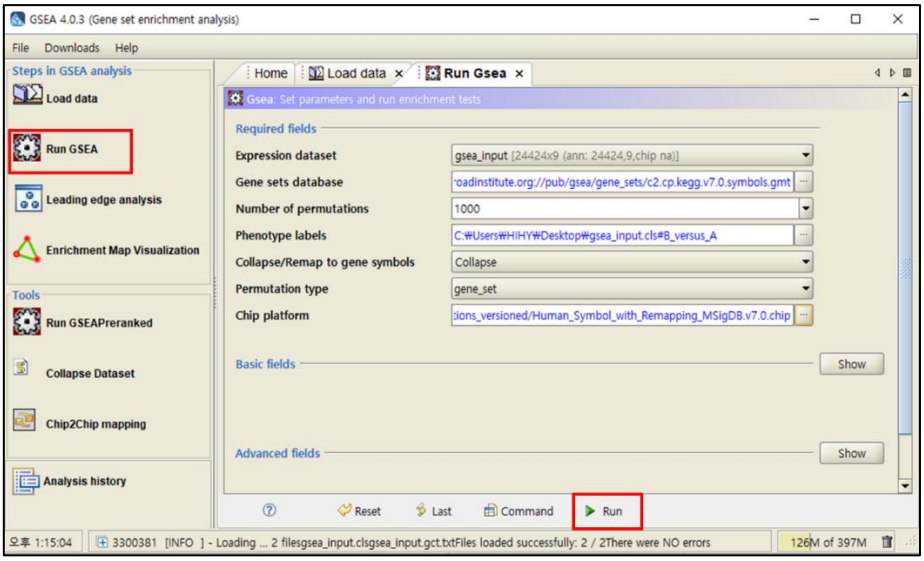


그림 1-6. Run GSEA

분석이 완료되면 GSEA 왼쪽 아래 GSEA reports 창에 status 가 Success 로 바뀐다. Show results folder 를 누르면 GSEA 분석 결과 창이 열린다(그림 1-7).

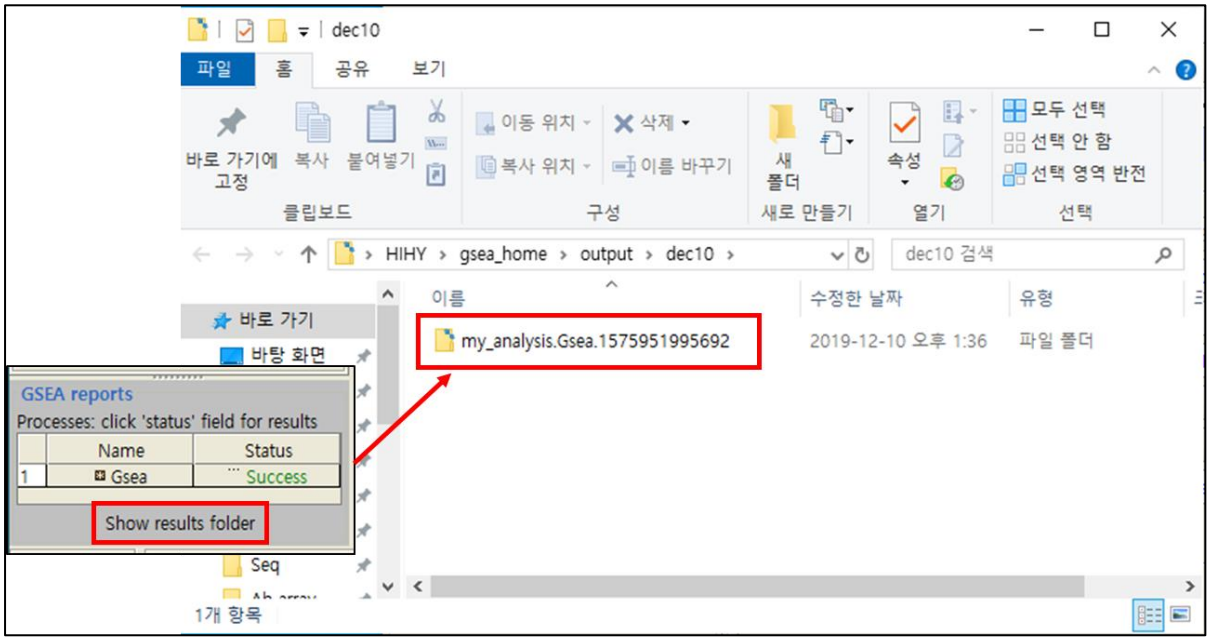


그림 1-7. GSEA results folder

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 1-8).

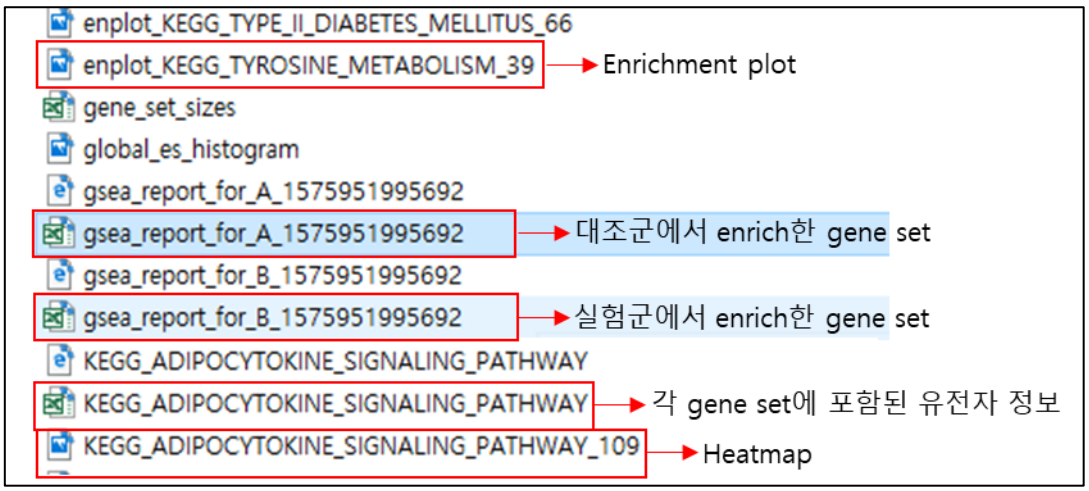


그림 1-8. GSEA result files

_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 1-9). Peak가 왼쪽에 생기면 대조군 대비 실험군에서 up된 유전자들이 많다는 의미이고, peak가 오른쪽에 생기면 down된 유전자가 많다는 의미이다.

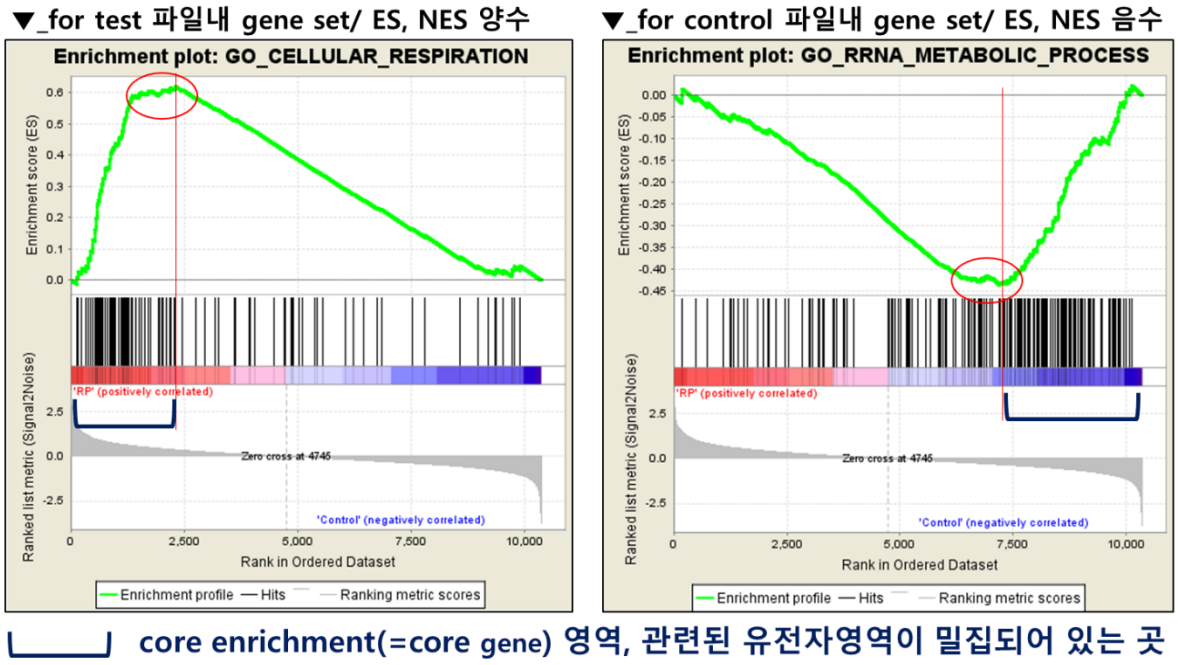


그림 1-9. GSEA enrichment plot

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