약학 박사 학위 논문

유전자 세트 분석을 이용한 유방암 종양개시 세포의 메타분석

Meta-Analysis of Tumor Stem-Like Breast Cancer Cells Using Gene Set Analysis

2016년 8월

서울대학교 대학원
약학과 약품분석학전공
이 원 준

약학 박사 학위 논문

유전자 세트 분석을 이용한 유방암 종양개시 세포의 메타분석

Meta-Analysis of Tumor Stem-Like Breast Cancer Cells Using Gene Set Analysis

지도교수 박 정 일

이 논문을 약학박사 학위논문으로 제출함

2016년 6월

서울대학교 대학원

약학과 약품분석학 전공

이 원 준

이원준의 박사학위논문을 인준함

2016년 6월

위 원 장 권성원 (인)

부 위 원 장 임요한 (인)

위 원 김유선 (인)

위 원 홍순선 (인)

위 원 박정일 (인)

국문 초록

일반적으로 종양개시세포는 epithelial-to-mesenchymal-transition 성질을 가지며 증식속도가 빨라 암의 전이를 일으키는데 중요한 역할을 한다. 그러나 아직까지 종양개시세포에 관하여 확실한 마커는 알려지지 않았으며 종양개시세포 관련기전과 마커의 연구를 위해 가장 많이 사용되는 방법은 종양개시세포의 성질을 나타낸다고 알려진 sphere cells과 그 대조군인 adherent cells의 비교방법이다. 본 연구는 서로 다른 유방암 종양개시세포 연구에서 얻은 sphere cells과 adherent cells의 유전자 발현 데이터를 이용하여 메타분석을 수행함으로써 유방암 종양개시세포의 기전을 밝히고 새로운 마커를 검색하였다. 이를 위해 Gene Expression Omnibus에서 출처가 다른 유방암 종양개시세포 연구에서 3개의 유전자 발현 데이터를 얻고 이를 ComBat 알고리즘을 이용하여 하나의 데이터로 통합하였으며, 여기에 본 연구진과 공동으로 연구를 수행한 아주대학교에서 유전자 발현 데이터를 제공하여 본 연구에 추가하였다. 메타분석을 진행하기 위해 위의 결과로 얻은 두 개의 데이터에 각각 gene set analysis를 적용하여 유의성 있는 gene set을 얻은 후 두 데이터 모두에서 공통적으로 유의성을 보인 4개의 gene set을 얻었다. 유방암 종양개시세포에 관여하는 유전자 마커는 유의성을 보인 4개의 gene set이 포함하는 유전자 중에서, 두 데이터 모두에서 p-value < 0.05를 만족하고 발현이 증가했던 CXCR4와 CXCL1을 포함하는 6개의 유전자로 선택하였다. 실험적 검증을 위하여 아주대학교에서 최종적으로 얻어진 6개의 유전자에 대해 quantitative reverse transcription-polymerase chain reaction을 수행하였고 6개의 유전자 중에서 CXCR4, CXCL1, HMGCS1이 MCF-7의 sphere cells에서 adherent cells과 비교했을 때 발현이 증가하였다. 본 연구는 gene set analysis를 이용하여 메타분석을 수행하였으며 이를 통해 유방암 종양개시세포 기전에 관여하는 4개의 gene set과 유전자 마커인 CXCR4, CXCL1, HMGCS1을 제시하였다. 최종적으로 메타분석에 gene set

analysis를 도입함으로써 통계적인 유의성을 가질 뿐만 아니라 gene set 개념을 바탕으로 생물학적 기전 정보를 고려한 유전자 마커를 제시하였다.

주요어

: 메타분석, 유방암 종양개시세포, 유전자 세트 분석, 유전자 마커, cytokine-cytokine receptor interaction gene set, valine, leucine and isoleucine degradation gene set

학 번 : 2012-30467

목차

List of Tables

List of Figures

List of Abbreviations

EMT: epithelial-to-mesenchymal-transition
ALDH1: aldehyde dehydrogenase activity
SP: ABC transporter dependent Hoechst side population
GEO: Gene Expression Omnibus
PAM: prediction analysis for microarrays
RT-PCR: quantitative reverse transcription-polymerase chain reaction
DAVID: Database for Annotation, Visualization, and Integrated Discovery
GAGE: Generally Applicable Gene-set Enrichment
KEGG: Kyoto Encyclopedia of Genes and Genomes
FDR: false discovery rate
TP: true positive
FP: false positive
TN: true negative
FN: false negative

Ⅰ. 서론

종양개시세포는 빠른 종양형성과 암의 재발에 관여한다고 알려져 있다 [1]. 유방암, 뇌암, 췌장암, 자궁암 등 다양한 암세포에서 종양개시세포는 항암치료에 강한 내성을 갖고 있으며 빠른 증식과 epithelial-to-mesenchymal-transition (EMT) 성질을 가지는 것으로 관찰되고 있다 [2]. 따라서 이러한 종양개시세포의 연구는 암 치료에 매우 중요하며 이를 위해 종양개시세포의 성질을 가지는 sphere cells을 대조군인 adherent cells과 비교하는 방법이 많이 사용되고 있다 [2]. 지금까지 알려진 종양개시세포 관련 마커는 CD24-/CD44+, aldehyde dehydrogenase activity (ALDH1), ABC transporter dependent Hoechst side population (SP) 등이 있지만 확실한 상관관계가 확인되지 않아 항암치료에 임상적으로 사용될 수 있는 종양개시세포 마커에 대한 연구는 아직까지 진행 중이다 [1, 2].

본 연구는 유방암 종양개시세포에 관여하는 새로운 유전자 마커의 검색을 위해 Gene Expression Omnibus (GEO)[1]와 같은 open database에서 얻은 데이터를 통합하는 메타분석[2]을 시도하였다. GEO는 미국에서 운영되는 database로 유전체 정보관련 세계 최대 규모의 database로 알려져 있다. 메타분석은 연구주제와 관련된 이미 나와 있는 결과를 체계적인 문헌고찰을 통해 수집하고 이를 통계적 방법을 사용하여 보다 일반적이고 신뢰성 있는 결론을 도출하기 때문에 경제적 비용과 시간이 절약되는 장점이 있다 [3, 4]. 일반적으로 어떤 한 분야에서 개별 연구의 결과가 누적이 되게 되면 그 누적된 연구결과를 통합하여 새로운 결론을 내려는 필요성이 생기게 되는데 메타분석은 이러한 필요성에서 시작이 되었다. 메타분석을 위해 본 연구진은 서로 다른 유방암 종양개시세포 연구에서 sphere cells과 adherent cells의 microarray 데이터를 GEO에서 얻었으며 데이터 개수는 총 3개로 모두 Affymetrix[3] Gene Chip Arrays를 사용한 유전자 발현 데이터이다. GEO에서 얻은 3개의 데이터는 ComBat[4] 알고리즘을 사용하여 batch effect를 제거하고 하나의 데이터로 통

합하였다 [5]. Batch effect는 메타분석에서 서로 다른 출처의 연구 결과를 하나로 합하는 경우 서로 다른 출처라는 변수가 데이터 값에 주는 영향을 의미하며 본 연구에서 cancer stem cell 관련 마커의 검색을 위해 adherent cell과 sphere cell이라는 변수를 고려하여 분석을 진행하기 위해서는 반드시 batch effect를 제거해주어야 한다. 데이터 추가를 위해 본 연구진과 공동 연구를 수행한 아주대학교는 MCF-7 cell line에서 유래된 adherent cells과 sphere cells을 자체적으로 배양하고 각각에서 Illumina[5] Gene Chip Arrays를 이용하여 얻은 유전자 발현 데이터를 제공하였다.

현재까지 제시된 메타분석은 크게 4가지 방법으로 vote counting, combing ranks, combining p-values, combining effect sizes가 있으나 이 4가지 방법은 모두 산술적 통계방법에 의존하고 있으며 실제 생물학적 기전 정보는 메타분석 과정에서 고려되지 않고 있다 [4, 6]. 본 연구는 Affymetrix와 Illumina platform에서 얻은 sphere cells과 adherent cells의 유전자 발현 데이터에 gene set analysis[6]를 적용하여 비교함으로써 개별 유전자의 통계적 유의성만을 고려할 뿐만 아니라 gene set 개념을 도입하여 생물학적으로 서로 연관성 있는 유전자 정보까지 고려할 수 있는 메타분석의 새로운 방법을 제시하였다. 기존의 유전자 발현 분석에서는 개별 유전자의 p-value를 계산하여 유의성을 보이는 개별 유전자를 선별하였지만 gene set analysis는 gene set, 즉 pathway에 해당하는 p-value를 계산하여 유의성을 보이는 pathway를 선별하게 된다. 본 연구에서는 기존 유전자 발현 분석과 같이 개별 유전자 선별에 바로 들어가지 않고 gene set analysis를 이용하여 유의성을 보이는 gene set을 선별한 후에 선별된 gene set 내에서 개별 유전자를 선별하는 방법을 제시하였다. Gene set의 경우 어떤 생물학적 현상이나 특정 질병에 대한 정보를 포함하기 때문에 gene set analysis를 이용하여 분석을 진행하면 분석 결과의 생물학적 해석이 용이하게 된다는 장점이 있다. 메타분석은 Affymetrix와 Illumina 각각에서 gene set analysis를 실행하여 p-value < 0.001을 만족하는 유의성 있는 gene set을 얻고 이 중에서 두 platform 모두에서 유의성을 나타내는 4개의 gene set

을 선택하여 진행하였다. 선택된 4개 gene set의 통계적 검증을 위하여 Affymetrix와 Illumina 데이터 각각에서 prediction analysis for microarrays (PAM)[7]을 이용하여 leave-one-out cross-validation을 진행하였고 4개의 gene set에 대한 accuracy를 얻었다. 최종 유전자 마커를 선별하기 위해 4개의 gene set을 구성하는 개별 유전자에 대한 p-value와 발현증감에 대한 정보를 R package인 Globaltest[8]를 이용하여 얻고 그 두 가지 정보를 고려하여 최종 유전자 마커를 선택하였다.

본 연구에서는 gene set analysis를 메타분석에 적용하여 유방암 종양개시세포의 기전을 설명할 수 있는 개별 유전자 마커를 검색 및 선별함으로써 기존 방법과는 차별화된 메타분석 방법을 제시하였으며 선택된 개별 유전자 마커는 아주대학교에서 quantitative reverse transcription-polymerase chain reaction (RT-PCR)를 이용하여 추가적 실험검증을 마쳤다. 본 연구는 단순한 산술적 데이터 통합에 의한 메타분석과 비교하여 생물학 정보에 기반 한 메타분석을 제시함으로써 유방암 종양개시세포연구 뿐만 아니라 다른 분야에도 적용 가능한 메타분석의 새로운 방법을 제시하였다.

Ⅱ. 재료 및 방법

1. 데이터 수집

유방암 종양개시세포의 유전자 발현 데이터 수집을 위해 GEO (www.ncbi.nlm.nih.gov/geo/)와 ArrayExpress[9] (www.ebi.ac.uk/arrayexpress)에서 검색어 "cancer stem", "breast", "sphere", "mammosphere", "tumor stem-like"를 이용하여 데이터 검색을 하였다. 검색 결과 총 49개 데이터를 찾았으며 동일한 데이터의 전처리를 위해 이 중 Affymetrix Human Genome U133 Plus 2.0 Array의 platform을 이용한 17개 데이터를 선택하였다. 17개 데이터 중에서 다음 2가지 기준을 이용하여 최종 3개 데이터를 선택하였다: (1) 데이터 발현값의 적절성과 (2) 종양개시세포 연구를 위한 sphere cell과 adherent cell 데이터의 포함여부. Affymetrix 데이터 이외에 본 연구진은 아주대학교에서 Illumina human HT12-v4 Beadchip을 이용하여 sphere cell과 adherent cell에 대해 얻은 유전자 발현 데이터를 분석 데이터에 추가하였다. 최종적으로 본 메타분석에서는 3개의 Affymetrix 데이터와 1개의 Illumina 데이터를 얻어서 분석을 진행하였다 (Figure 1).

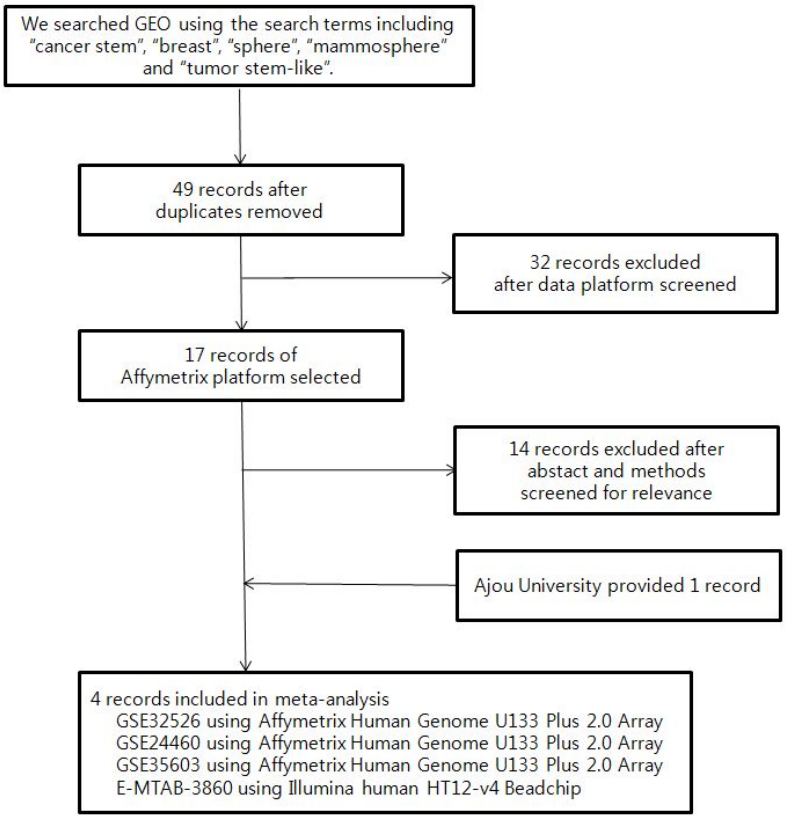


Figure 1. 메타분석을 위한 데이터 수집 과정.

2. 세포배양 및 유전자 발현 profiling

본 연구진과 공동 연구를 수행한 아주대학교는 MCF-7 세포주를 American Type Culture Collection (ATCC, Manassas, VA)에서 구입하여 10% fetal bovine serum의 DMEN medium에서 배양하고 유전자 발현 profiling을 수행하였다. MCF-7의 single cell suspensions은 1 x B27 supplement (Life Technologies, Carlsbad, CA), 20 ng/mL basic fibroblast growth factor (R&D Systems, Minneapolis, MN), 20 ng/mL recombinant epidermal growth factor (Life Technologies, Carlsbad, CA), 100 U/mL penicillin, 100 μ

g/mL streptomycin을 포함하는 DMEM/F12에 배양되었으며 ultralow adherence dish (Corning, Corning, NY)에서 길러졌다. 계대배양은 2주에 두 번 진행하였으며 sphere cell 배양을 위해 single cells을 mammosphere medium 환경으로 옮겨서 배양하였다 [7]. RNA 추출은 mirVana™ RNA isolation kit (Ambion, Inc., Carlsbad, CA)을 이용하여 제조사 프로토콜에 따라 진행하였으며 RNA의 총 양은 Illumina TotalPrep RNA amplification kit (Ambion, Inc., Carlsbad, CA)을 사용하여 cRNA 합성에 사용되었다. cRNA는 제조사의 프로토콜에 따라 Illumina human HT12-v4 Beadchip gene expression array의 hybridization에 사용되었고 hybridization된 array의 fluorescence signals은 Illumina Bead Array Reader (Illumina, San Diego, CA)를 사용하여 얻었다.

3. 데이터 전처리

수집된 데이터 중 Affymetrix platform의 데이터는 GSE32526, GSE24460, GSE35603을 포함했으며 이 데이터의 발현값을 표준화하기 위해 robust multi-array analysis (RMA) 방법을 R package인 affy[10]를 이용하여 실행하였다 [8]. 표준화 이후 3개 데이터의 발현값에서 존재하는 batch effect를 제거하고 하나의 데이터로 통합하기 위해 ComBat 알고리즘을 R package인 sva[4]를 이용하여 적용하였다 [5]. ComBat은 생물학 분야의 high-throughput 데이터에서 발생되는 발현값의 batch effect나 치우침 현상을 empirical Bayesian framework 방법을 통해 batch effect가 제거된 보정된 값으로 산출하여준다. Illumina platform의 데이터 발현값 표준화는 log2 변환과 quantile normalization 방법을 통해 아주대학교에서 진행하였다. 데이터 발현값의 표준화 이후 총 4개 데이터의 gene label을 Database for Annotation, Visualization, and Integrated Discovery (DAVID)[11]를 이용하여 Entrez IDs[12]로 변환하였다 [9].

4. Gene set analysis

Gene set analysis는 Generally Applicable Gene-set Enrichment (GAGE) 방법을 R package인 gage[6]를 이용하여 진행하였다. GAGE는 이전의 gene set analysis 방법인 GSEA나 PAGE가 가진 서로 다른 sample크기와 실험 디자인을 처리하는데서 오는 한계를 개선한 방법으로 이전의 방법과 비교하여 재현성이 좋은 결과를 보여 왔다 [10]. 본 연구진은 sphere cells과 adherent cells을 비교하여 유의성을 보이는 gene set을 선별하기 위해 gage를 Affymetrix와 Illumina platform 각각에서 적용하였다. gage는 Kyoto Encyclopedia of Genes and Genomes (KEGG)에서 정의한 각각의 gene set에 대해 실험군과 대조군의 발현값을 이용하여 p-value 값을 제공해준다. Affymetrix와 Illumina 각각에서 gene set에 대한 p-value 값을 얻은 후 p-value < 0.001을 cut-off로 하여 유의성 있게 차이를 보인 gene set을 선별하고 두 platform 모두에서 p-value < 0.001을 만족하는 4개 gene set을 최종 선택하였다.

5. 통계적 검증과 개별 유전자 마커 선택

최종 선택된 4개 gene set의 통계적 검증을 위하여 leave-one-out cross-validation[13] 방법을 사용하였다. Leave-one-out cross-validation은 Affymetrix와 Illumina platform 각각에서 실행하였으며 진행방법은 sphere cell과 adherent cell을 포함하는 전체 데이터에서 하나의 데이터를 제외시키고 나머지 데이터를 이용하여 prediction model을 만든 후, 이 모델을 이용하여 제외시켰던 데이터가 sphere cell인지 adherent cell 인지 예측하는 방법으로 하였다. Prediction model은 prediction analysis for microarrays (PAM)을 사용하여 만

들었으며 PAM은 nearest shrunken centroid 방법을 사용하여 데이터의 class를 예측하는 알고리즘으로 여러 연구에서 microarray 데이터의 통계적 검증에 사용되어왔다. 통계검증 결과 4개의 gene set 각각에서 Affymetrix와 Illumina platform에 해당하는 accuracy를 얻었다.

최종 개별 유전자 마커의 선별을 위해 선택된 4개의 gene set을 구성하는 개별 유전자에 대한 p-value를 R package "Globaltest"의 component test를 이용하여 얻었다 [11, 12]. Globaltest는 대조군과 실험군 사이의 p-value와 함께 positive association 또는 negative association에 관한 정보를 제공하며 본 연구에서 positive association은 sphere cell에서 해당 유전자가 발현이 증가했다는 의미이며 negative association은 발현이 감소했다는 의미이다. 개별 유전자 마커는 Globaltest 결과 Affymetrix와 Illumina 2개의 platform에서 p-value < 0.05를 만족하고 발현의 증감이 일치하는 유전자로 선택하였다. 4개의 gene set 각각에서 개별 유전자의 유의성을 시각화하기 위해 Globaltest를 이용하여 gene plot을 생성하였으며 gene plot의 bar는 개별 유전자의 p-value에 해당한다. Adherent cells과 비교하여 sphere cells에서 발현 증감을 표시하기 위해 bar의 색깔을 녹색 또는 붉은색으로 할당하였으며 붉은색은 해당 유전자가 sphere cells에서 발현이 증가했음을 의미하고 녹색은 발현이 감소했다는 것을 의미한다.

6. Reverse transcription-PCR

본 연구진과 공동연구를 수행한 아주대학교에서는 RT-PCR을 이용하여 선별된 개별 유전자의 발현을 실험적으로 확인하기 위해 각 sample에서 총 1 μg의 RNA를 추출하였고 cDNA 합성을 위해 reverse transcriptase kit (Promega)을 사용하였으며 Taq DNA polymerase (Promega)로 만든 cDNA를 PCR에 사용하였다. PCR 증폭은 최적화된 풀림온도에서 진행되었으며 PCR cycles 횟수는 17 또는 30으로 하였다.

Ⅲ. 결과

1. 수집된 데이터의 성격

데이터 수집결과 GEO에서 2010년과 2012년에 발표된 GSE32526, GSE24460, GSE35603을 포함한 3개의 Affymetrix Gene Chip Array 유전자 발현 데이터를 메타분석에 사용하였다. GSE32526은 55세, 85세 여성 유방암 환자의 유전자 발현 데이터이며 이 중 높은 종양 형성을 보였던 sample 데이터 S2N을 선택하였고 이 데이터는 각각 3개의 sphere cell과 adherent cell의 유전자 발현 데이터를 포함하였다 [2]. GSE32526은 유방암 환자의 외과 수술에서 얻은 sample의 유전자 발현 데이터이며 이 과정은 University of Palermo on human experimentation의 기관 위원회 윤리규범을 따라 진행하였다 [2]. GSE24460에서는 parental MCF-7과 MCF/ADR cells을 사용하였다. Parental MCF-7 cells은 wild-type이며 estrogen receptor positive luminal subtypes으로 분류 된다 [13, 14]. MCF-7/ADR

cells은 빠른 증식속도를 보이는 sphere cells이며 높은 농도의 doxorubicin이 첨가된 배지에서 배양되었다 [13]. Parental MCF-7 cells과 MCF-7/ADR cells은 각각 2개의 유전자 데이터를 포함하였다. GSE35603은 3개의 parental MCF-7 cells과 parental MCF-7 cells에서 유래된 2개의 sphere cell 유전자 발현 데이터를 포함하였다. GSE35603의 parental MCF-7은 wild-type이며 estrogen receptor-positive로 분류 된다 [15, 16]. Illumina platform의 데이터 추가를 위해, parental MCF-7과 parental MCF-7에서 유래된 sphere cells 각각에 해당하는 2개의 유전자 발현 데이터를 아주대학교로부터 얻었다. 아주대학교에서 사용한 parental MCF-7 cells은 luminal A subtype이며 estrogen receptor-positive로 분류 된다. 아주대학교에서 제공한 유전자 발현 데이터는 ArrayExpress에 업로드 되었으며 데이터는 E-MTAB-3860으로 accession number가 할당되었다.

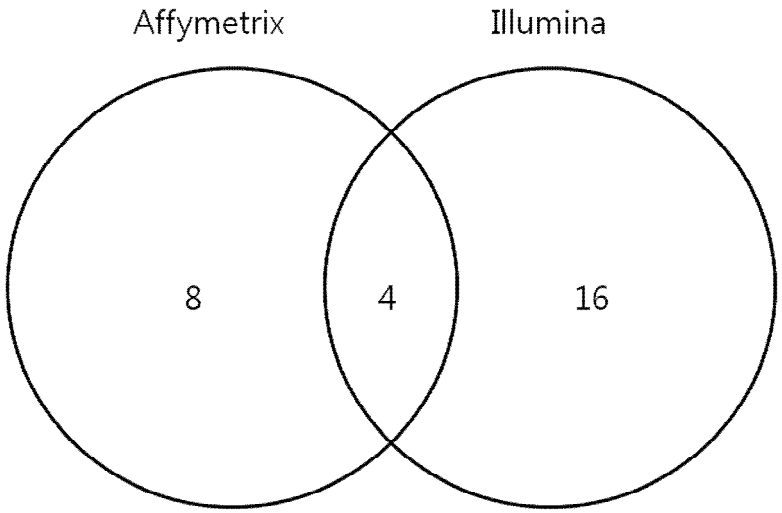


Figure 2. Affymetrix와 Illumina platform에서 얻은 Gene set analysis 결과를 보여주는 벤다이어그램.

2. Gene set analysis와 통계적 검증

Gene set analysis 결과, p-value < 0.001을 cut-off로 사용하여 Affymetrix와 Illumina platform 각각에서 유의성 있게 발현차이가 나는 12, 20개의 gene set을 얻었다. Figure 2는 gene set analysis 결과에서 유의성을 보인 gene set을 벤다이어그램을 이용하여 나타낸 것이며 이 중 Affymetrix와 Illumina platform 모두에서 유의성을 보인 4개의 gene set인 cytokine-cytokine receptor interaction, valine, leucine and isoleucine degradation, systemic lupus erythematosus, DNA replication을 선택하였다. 선택된 4개의 gene set은 Affymetrix와 Illumina에서 모두 false discovery rate (FDR) <0.05를 만족하였으며 DNA replication은 sphere과 adherent cells 사이에서 가장 높은 유의성을 보였다 (Table 1).

Gene Sets	Affymetrix				Illumina		
	p-value	FDR	accuracy (%)	AUC	p-value	FDR	accuracy (%)
DNA replication	9.81E-05	0.008	87	**0.939**	6.62E-12	**1.17E-09**	100
Valine, leucine and isoleucine degradation	0.000729	0.016	73	**0.816**	0.000251	**0.007**	100
Cytokine-cytokine receptor interaction	0.000852	0.041	87	**0.841**	0.000575	**0.010**	100
Systemic lupus erythematosus	0.000978	0.041	93	**0.982**	0.000553	**0.010**	100

Table 1. Gene set analysis 결과, Affymetrix와 Illumina platform 모두에서 p-value < 0.001을 만족하는 4개 gene set의 p-value, FDR, accuracy 값.

선택된 4개의 gene set에 대한 통계적 검증을 위해 leave-one-out cross validation을 이용하여 accuracy를 계산하였다. Leave-one-out cross validation 결과에서 positive와 negative는 각각 sphere cell과 adherent cell을 의미하며 accuracy의 정의에 따라 true positive (TP)는 실제 sphere cell이 prediction model에 의해 sphere cell로 예측된 경우의 수이며 false positive (FP)는 실제 adherent cell이 sphere cell로 예측된 경우의 수이다. 같은 방법으로 true negative (TN)은 실제 adherent cell이 adherent cell로 예측된 경우의 수이며 false negative (FN)은 실제 sphere cell이 adherent cell로 예측된 경우의 수이다 [17]. Accuracy는 정의에 따라 TP, FP, TN, FN 값을 이용하여 계산되었다. Table 2는 leave-one-out cross validation의 결과를 나타낸 표이다. Illumina platform에서는 모든 sample이 TP 또는 TN의 결과를 보였으며 Affymetrix platform에서는 몇 개의 sample이 FP 또는 FN의 결과를 보였다. Cytokine-cytokine receptor interaction의 gene set에서는 sample 4와 15가 각각 FP와 FN의 결과를 나타내었으며 valine, leucine and isoleucine degradation의 gene set에서는 sample 7과 8이 FP, sample 10과 15가 FN의 결과를 보였다. Systemic lupus erythematosus에서는 sample 15가 FN의 결과를 보였고 DNA replication에서는 sample 4와 10이 각각 FP와 FN의 결과를 보였다. Leave-one-out cross validation 결과를 바탕으로 선별된 4개 gene set의 accuracy를 구하였으며 4개의 gene set이 Affymetrix platform에서 70% 이상의 accuracy를 보였고 Illumina platform에서 100%의 accuracy를 보였다 (Table 2).

Gene Sets	Samples of Affymetrix															Samples of Illumina			
	Adherent cells								Sphere cells							Adherent cells		Sphere cells	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4
Cytokine-cytokine receptor interaction	T	T	T	F	T	T	T	T	T	T	T	T	T	T	F	T	T	T	T
Valine, leucine and isoleucine degradation	T	T	T	T	T	T	F	F	T	F	T	T	T	T	F	T	T	T	T
Systemic lupus erythematosus	T	T	T	T	T	T	T	T	T	T	T	T	T	T	F	T	T	T	T
DNA replication	T	T	T	F	T	T	T	T	T	F	T	T	T	T	T	T	T	T	T

Table 2. Leave-one-out cross validation 결과. T는 adherent 또는 sphere cell이 각각 adherent 또는 sphere cell로 예측된 경우이며 F는 adherent 또는 sphere cell이 각각 sphere 또는 adherent cell로 예측된 경우임.

3. 개별 유전자 마커 선택

개별 유전자 마커는 선별된 4개의 gene set을 구성하는 유전자들 내에서 선택하였으며 선택 기준은 Globaltest 결과인 p-value와 2개의 platform에서 발현증감의 일치 여부로 하였다.

Table 3은 4개의 gene set을 구성하는 유전자 중에 Affymetrix와 Illumina platform 모두에서 p-value < 0.05를 만족하는 유전자 리스트이다. Cytokine-cytokine receptor interaction에서는 IL12RB2, CXCL1, CXCR4의 발현이 2개의 platform 모두에서 증가했지만

CXCL10, CXCL6, TNFRSF11B은 Illumina에서만 발현이 감소하였다. Valine, leucine and isoleucine degradation에서는 ACADM, BCKDHB, HMGCS1의 발현이 2개의 platform 모두에서 증가했지만 PCCB, AOX1은 2개의 platform 어느 한 개에서만 발현이 감소하였다. Systemic lupus erythematosus에서는 HLA-DMA만이 2개의 platform에서 발현이 증가하였고 DNA replication에서는 발현증감이 일치하는 유전자가 없었다. 4개의 gene set 중에서는 cytokine-cytokine receptor interaction과 valine, leucine and isoleucine degradation이 발현증감이 일치하는 유전자를 많이 포함하였다.

Gene Sets	Genes	Affymetrix		Illumina	
		p-value	direction	p-value	direction
Cytokine-cytokine receptor interaction	CXCL10	0.004	down	0.038	up
	CXCL6	0.009	down	0.028	up
	IL12RB2	0.013	up	0.030	up
	CXCL1	0.014	up	0.011	up
	TNFRSF11B	0.022	down	0.009	up
	CXCR4	0.029	up	0.001	up
Valine, leucine and isoleucine degradation	PCCB	0.000	up	0.025	down
	ACADM	0.004	up	0.013	up
	BCKDHB	0.015	up	0.037	up
	AOX1	0.036	down	0.023	up
	HMGCS1	0.038	up	0.005	up
Systemic lupus erythematosus	HIST1H2BD	0.000	down	0.037	up
	HLA-DMA	0.001	up	0.049	up
	HIST2H3A	0.002	down	0.004	up
	HIST1H2BC	0.006	down	0.018	up
	H2AFJ	0.013	down	0.008	up
	SSB	0.026	up	0.025	down
	HIST2H2BE	0.045	down	0.000	up
DNA replication	RfC4	0.000	up	0.030	down
	RPA1	0.002	up	0.001	down
	MCM4	0.005	up	0.001	down
	MCM5	0.007	up	0.014	down
	MCM2	0.009	up	0.009	down
	FEN1	0.022	up	0.039	down
	RFC5	0.029	up	0.031	down

Table 3. 4개의 gene set을 구성하는 유전자 중 p-value <0.05를 만족하는 유전자의 p-value와 발현증감 정보.

A

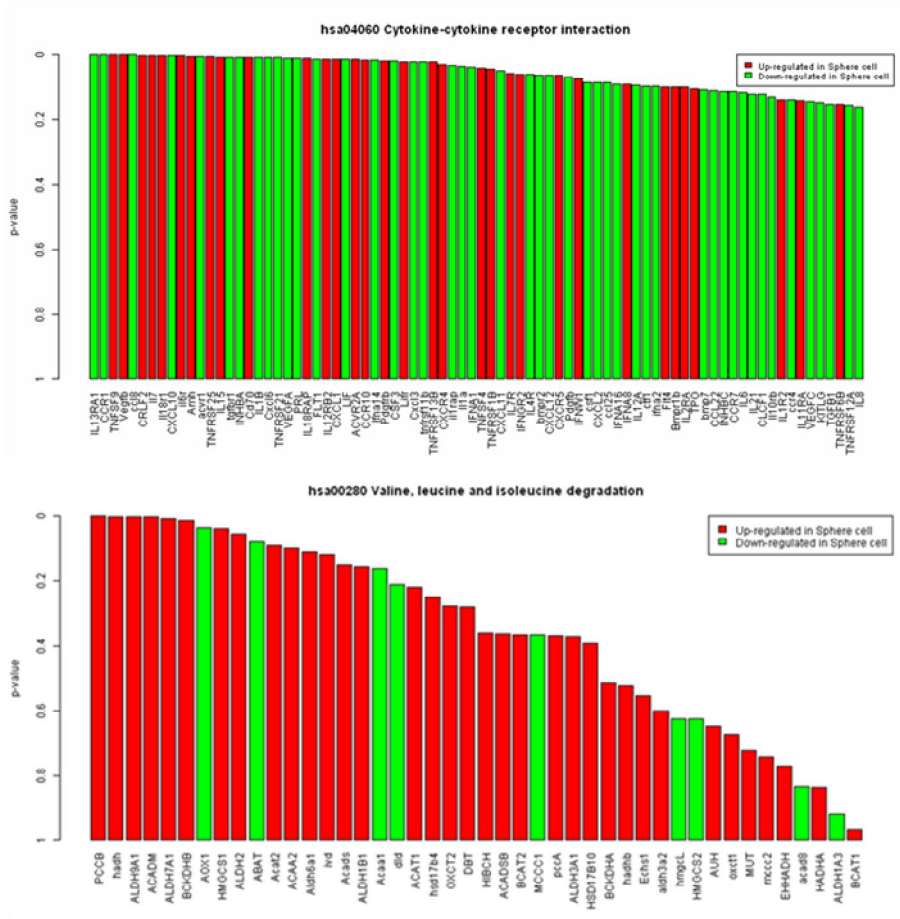


Figure 3A. Globaltest 결과, Affymetrix platform 데이터에서 cytokine-cytokine receptor interactionr과 valine, leucine and isoleucine degradation이 구성하는 유전자들의 p-value값과 발현증감 정보.

B

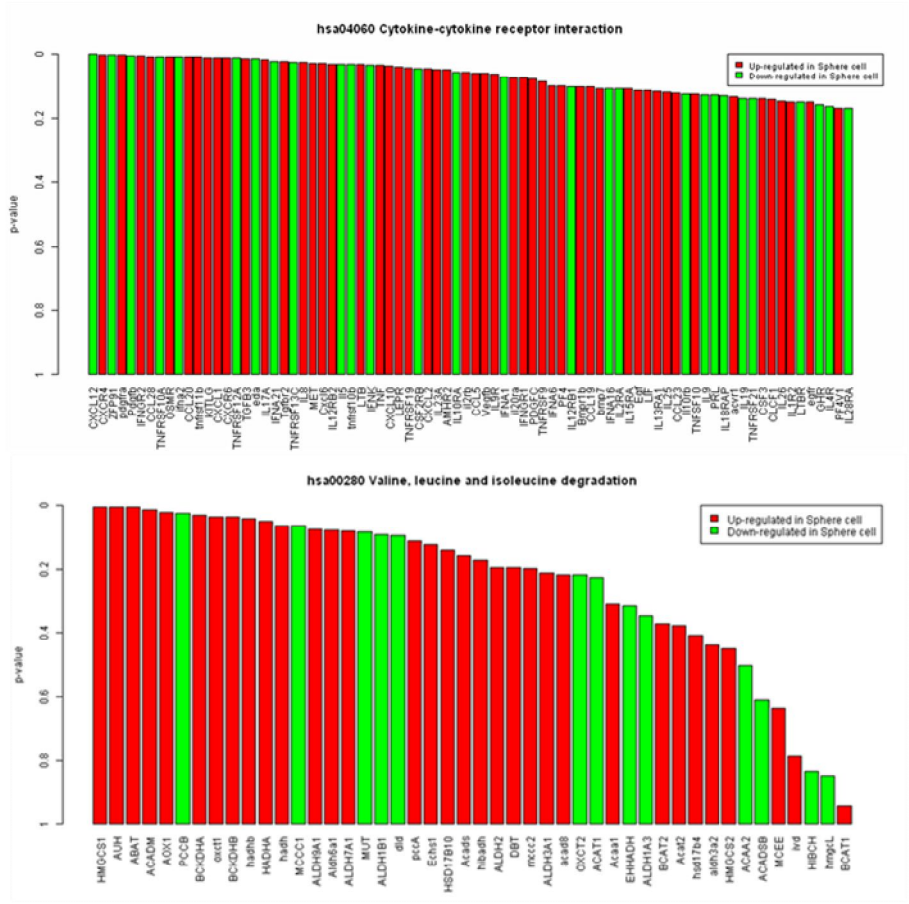


Figure 3B. Globaltest 결과, Illumina platform 데이터에서 cytokine-cytokine receptor interactionr과 valine, leucine and isoleucine degradation이 구성하는 유전자들의 p-value값과 발현증감 정보.

Figure 3은 Affymetrix와 Illumnian platform에서 cytokine-cytokine receptor interactionr과 valine, leucine and isoleucine degradation의 gene plot을 보여준다. Affymetrix에서 cytokine-cytokine receptor interaction은 TNFSF9, VEGFB, CRLF2, IL7, IL18R1을 포함한 21개의 유전자가 sphere cell에서 유의성 있게 발현이 증가하였고 IL13RA1, CCR1, CCL8, CXCL10, ACVR1을 포함한 22개의 유전자가 유의성 있게 발현이 감소하였다. Illumina에서 cytokine-cytokine receptor interaction은 CXCR4, PDGFRA, IFNGR2, CCL28, OSMR을 포함한 25개의 유전자가 sphere cell에서 유의성 있게 발현이 증가하였고 CXCL12, ZFP91, PDFGB, TNFRSF10A, IFNA2을 포함한 13개의 유전자가 유의성 있게 발현이 감소하였다. Affymetrix에서 valine, leucine and isoleucine degradation은 PCCB, HADH, ALDH9A1, ACADM, ALDH7A1을 포함한 7개의 유전자가 sphere cell에서 유의성 있게 발현이 증가하였고 AOX1이 유의성 있게 발현이 감소하였다. Illumina에서 valine, leucine and isoleucine degradation은 HMGCS1, AUH, ABAT, ACADM, AOX1을 포함한 10개의 유전자가 sphere cell에서 유의성 있게 발현이 증가하였고 PCCB가 유의성 있게 발현이 감소하였다. Affymetrix와 비교했을 때, Illumina platform에서 cytokine-cytokine receptor interaction와 valine, leucine and isoleucine degradation의 발현이 증가된 유전자 수가 더 많았다. 최종적으로 cytokine-cytokine receptor interaction과 valine, leucine and isoleucine degradation을 구성하는 유전자 중에 Affymetrix와 Illumina platform 모두에서 발현이 증가된 IL12RB2, CXCL1, CXCR4, ACADM, BCKDHB, HMGCS1을 포함한 6개의 유전자를 개별 유전자 마커로 선택하였다.

4. Reverse transcription-PCR

선별된 6개 유전자 마커의 발현을 실험적으로 확인하기 위해, 본 연구진과 공동연구를 수행한 아주대학교는 RT-PCR을 이용하여 6개 유전자 마커 (IL12RB2, CXCL1, CXCR4, ACADM, BCKDHB, HMGCS1)의 mRNA 발현을 MCF-7과 MCF-7에서 유래된 sphere cells에서 확인하였다. RT-PCR을 진행 시, SNAI와 ACTIN을 reference 유전자로 사용하였다. SNAI는 종양개시세포 기전에 관여하는 유전자로, 종양개시세포에서 발현이 증가한다고 알려져 있으며 ACTIN은 control 유전자로 사용되었다 [16, 18, 19]. RT-PCR 결과 CXCL1, CXCR4, HMGCS1의 mRNA 발현이 MCF-7과 비교했을 때 MCF-7에서 유래된 sphere cells에서 증가되었음을 확인하였고 IL12RB2, ACADM, BCKDHB는 유의성 있는 발현의 차이를 보이지 않았다 (Figure 4).

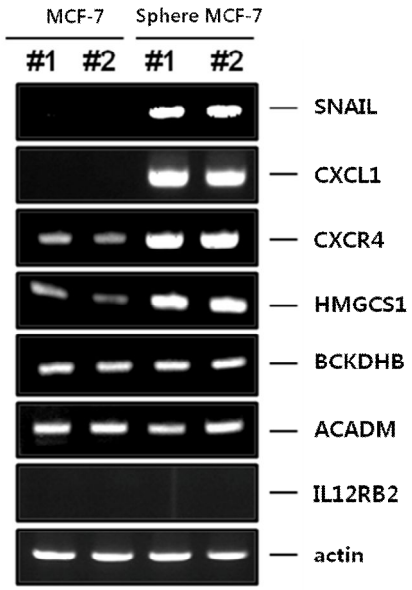


Figure 4. 메타분석을 통해 선별한 6개 유전자의 RT-PCR 결과 (아주대학교 제공)

Ⅳ. 결론

통계 알고리즘의 발달과 함께 경제적 비용이 들지 않고 신뢰성 있는 결과를 보여주는 장점으로 인해 최근 여러 연구에서 메타분석이 사용되고 있다. 특히 microarray 분야의 몇몇 연구에서 새로운 메타분석 방법이 제시되고 있으나 아직까지 포괄적이고 체계적인 분석 방법은 확립되지 않았다. 본 연구는 gene set analysis를 적용하여 여러 유방암 종양개시세포 연구로부터 얻은 유전자 발현 데이터를 이용한 메타분석을 수행하였고 그 결과로 유방암 종양개시세포에 관여하는 기전과 유전자 마커를 제시하였다. Gene set analysis 결과 총 4개의 gene set인 cytokine-cytokine receptor interaction, valine, leucine and isoleucine degradation, systemic lupus erythematosus, DNA replication을 cut-off인 p-value < 0.001를 이용하여 선택하였으며 이 4개의 gene set을 구성하는 유전자 중 Affymetrix와 Illumina platform 모두에서 p-value < 0.05와 발현의 증감이 일치한 IL12RB2, CXCL1, CXCR4, ACADM, BCKDHB, HMGCS1을 유전자 마커로 선별하였다. 이전 연구에서 종양개시세포 기전에 관여한다고 대표적으로 알려진 SNAIL의 경우 gene set analysis에서 선별된 4개의 gene set이 포함하는 구성 유전자에 없었기 때문에 본 연구에서 제시된 6개의 유전자 마커에는 누락이 되었지만 3개의 Affymetrix 데이터(GSE32526, GSE24460, GSE35603)가 통합된 Affymetrix platform과 Illumina platform 모두에서 adherent cell과 비교하여 sphere cell에서 발현이 증가되었음을 관찰하였다. Affymetrix platform에서 ComBat method로 데이터를 통합하기 전에 SNAIL의 발현은 GSE32526, GSE24460, GSE35603 모두에서 adherent cell과 비교하여 sphere cell에서 발현이 증가하였으며 3개 중 GSE32526은 SNAIL의 p-value가 6.69E-05로 높은 유의성을 보였다. 이를 통해 본 연구에서 사용된 데이터가 기존의 연구 결과인 SNAIL의 발현을 반영함을 확인하였다.

메타분석에서 선별된 유전자 마커의 발현을 실험적으로 확인하기

위해 본 연구진과 공동연구를 수행한 아주대학교에서 RT-PCR을 진행하였으며 그 결과 CXCL1, CXCR4, HMGCS1의 mRNA 발현이 MCF-7 cells과 비교하여 MCF-7에서 유래된 sphere cells에서 증가하였음을 관찰하였다. 증가된 유전자 중에서 CXCR4[14]는 CXC chemokine 수용체 중 하나로 stromal derived factor-1 (SDF-1)을 리간드로 가지며 건강한 사람의 조직에서는 CXCR4 유전자 발현이 낮거나 없는 것으로 알려져 있으나 유방암, 자궁암, 악성 흑색종, 전립선암 등 23가지 이상의 암에서는 발현이 증가하는 것으로 알려져 있다. 특히 CXCR4 발현이 증가한 암세포는 CXCL12[15]의 발현이 높은 조직인 폐, 간, 골수로 전이되는 경향이 있는 것으로 알려져 있다. 또한 많은 연구에서 유방암의 전이에 중요한 역할을 하는 인자로 보고하고 있으며 CXCR4의 발현이 증가된 cancer stem cells의 개체군은 종양의 전이에 결정적 역할을 한다는 보고가 있다 [20-23]. CXCL1[16]의 경우, 유방암, 폐암, 췌장암, 대장암, 전립선암 등 다양한 암에서 발현의 증가가 보고되었으며 유방암에서 이 유전자가 발현이 증가할 경우 폐로 전이 될 수 있는 위험이 높은 것으로 알려져 있다 [24-28]. 본 연구의 메타분석이 제시한 3개의 유전자 마커 중 2개가 이미 cancer stem cells 기전에 관여하는 중요한 마커로 밝혀졌으므로 본 연구에서 제시된 메타분석 방법은 유방암 종양개시세포에 관여하는 유전자 마커를 선별하는데 유용한 접근법이라고 할 수 있으며 나머지 1개인 HMGCS1[17] 또한 cancer stem cells 기전에 관여하는 마커로 제시하였다.

본 연구의 한계는 GEO와 같은 open database에 공개된 데이터 중 유방암 종양개시세포 관련 데이터 수가 많지 않아 메타분석을 위해 오직 3개의 Affymetrix platform의 데이터를 사용한 점이다. 또한 사용된 데이터의 sample이 종류가 모두 일치하지 않았으며 RT-PCR 결과도 낮은 sensitivity를 보였다. 데이터 분석에 사용된 유방암 cell line 중 GSE24460, GSE35603, E-MTAB-3860의 molecular subtype은 모두 estrogen receptor-positive luminal MCF-7 cell line이며 실험적 검증을 위한 RT-PCR에서는 estrogen receptor-positive luminal A subtype인 MCF-7 cell line을 사용하였

다. MCF-7 cell line은 cancer stem cells 연구에 많이 사용되어 왔으며 특히, 3-dimensional (3D) 배지에서 배양되었을 때 *in vivo*에서도 cancer stem cells의 성질이 많이 발현된다고 알려져 있다[29-32]. 데이터 수집을 위해 본 연구진은 Illumina platform의 유전자 발현 데이터를 아주대학교로부터 얻어 추가하였으며 Affymetrix의 데이터와 비교하여 Illumina 데이터는 sphere과 adherent cells 사이에서 보다 명확한 발현의 차이를 보여주었다. 따라서 본 연구진이 추가한 Illumina 데이터는 유전자 마커를 선택하는데 있어서 보다 명확한 발현 패턴을 제공하여 유전자 선별에 기여할 수 있었다. 또한, 데이터 분석에 있어서 Affymetrix와 Illumina 데이터를 모두 사용함으로써 두 platform의 특성을 모두 고려하였다.

결론적으로 본 연구진은 gene set analysis를 적용하여 새로운 메타분석 방법을 제시함으로써 산술적 정보에만 근거해서 마커를 선별하는 기존 방법에 생물학적 기전 정보를 추가로 고려하였으며 이를 통해 제시된 유전자 마커는 해당 gene set이 설명하는 특정 pathway 기전에 관여되기 때문에 종양개시세포 기전을 효과적으로 해석할 수 있다는 데 의미가 있다. 유의성과 발현정보를 고려하여 유방암 종양개시세포 기전에 관여하는 유전자 마커로서 CXCR4, CXCL1, HMGCS1을 선별하였으며 이를 RT-PCR을 이용하여 발현의 실험적 검증을 확인하였다.

Ⅴ. 참고 문헌

1. Wang H, Zhang Y, Du Y. Ovarian and breast cancer spheres are similar in transcriptomic features and sensitive to fenretinide. BioMed research international. 2013; 2013:510905. doi: 10.1155/2013/510905 PMID: 24222909; PubMed Central PMCID: PMC3816214.

2. Lehmann C, Jobs G, Thomas M, Burtscher H, Kubbies M. Established breast cancer stem cell markers do not correlate with in vivo tumorigenicity of tumor-initiating cells. Int J Oncol. 2012; 41(6):1932.42. doi: 10.3892/ijo.2012.1654 PMID: 23042145; PubMed Central PMCID: PMC3583871.
3. Goonesekere NC, Wang X, Ludwig L, Guda C. A meta analysis of pancreatic microarray datasets yields new targets as cancer genes and biomarkers. PLoS One. 2014; 9(4):e93046. doi: 10.1371/journal.pone.0093046 PMID: 24740004; PubMed Central PMCID: PMC3989178.
4. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 2008; 5(9):e184. doi: 10.1371/journal.pmed.0050184 PMID: 18767902; PubMed Central PMCID: PMC2528050.
5. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8(1):118.27. doi: 10.1093/biostatistics/kxj037 PMID: 16632515.
6. Oztemur Y, Bekmez T, Aydos A, Yulug IG, Bozkurt B, Dedeoglu BG. A ranking-based meta-analysis reveals let-7 family as a meta-signature for grade classification in breast cancer. PLoS One. 2015; 10(5):e0126837. doi: 10.1371/journal.pone.0126837 PMID: 25978727; PubMed Central PMCID: PMC4433233.
7. Kim YJ, Koo GB, Lee JY, Moon HS, Kim DG, Lee DG, et al. A microchip filter device incorporating slit arrays and 3-D flow for detection of circulating tumor cells using CAV1-EpCAM conjugated microbeads. Biomaterials. 2014; 35(26):7501.10. doi: 10.1016/j.biomaterials.2014.05.039 PMID: 24917030.
8. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4(2):249.64. doi: 10.1093/biostatistics/4.2.249 PMID: 12925520.
9. Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated

Discovery. Genome Biol. 2003; 4(5):P3. PMID: 12734009.
10. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics. 2009; 10:161. doi: 10.1186/1471-2105-10-161 PMID: 19473525; PubMed Central PMCID: PMC2696452.
11. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004; 20(1):93.9. PMID: 14693814.
12. Hulsegge I, Kommadath A, Smits MA. Globaltest and GOEAST: two different approaches for Gene Ontology analysis. BMC Proc. 2009; 3 Suppl 4:S10. doi: 10.1186/1753-6561-3-S4-S10 PMID: 19615110; PubMed Central PMCID: PMC2712740.
13. Calcagno AM, Salcido CD, Gillet JP, Wu CP, Fostel JM, Mumau MD, et al. Prolonged drug selection of breast cancer cells and enrichment of cancer stem cell characteristics. J Natl Cancer Inst. 2010; 102 (21):1637.52. doi: 10.1093/jnci/djq361 PMID: 20935265; PubMed Central PMCID: PMC2970576.
14. Mehta K. High levels of transglutaminase expression in doxorubicin-resistant human breast carcinoma cells. Int J Cancer. 1994; 58(3):400.6. PMID: 7914183.
15. Yu YH, Chiou GY, Huang PI, Lo WL, Wang CY, Lu KH, et al. Network Biology of Tumor Stem-like Cells Identified a Regulatory Role of CBX5 in Lung Cancer. Sci Rep. 2012; 2:584. Epub 2012/08/18. doi: 10. 1038/srep00584 PMID: 22900142; PubMed Central PMCID: PMC3419921.
16. Lien HC, Hsiao YH, Lin YS, Yao YT, Juan HF, Kuo WH, et al. Molecular signatures of metaplastic carcinoma of the breast by large-scale transcriptional profiling: identification of genes potentially related to epithelial-mesenchymal transition. Oncogene. 2007; 26(57):7859.71. doi: 10.1038/sj.onc.1210593 PMID: 17603561.
17. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240(4857):1285.93. PMID:3287615.
18. Yang J, Mani SA, Donaher JL, Ramaswamy S, Itzykson RA, Come C, et al. Twist, a master regulator of morphogenesis, plays an

essential role in tumor metastasis. Cell. 2004; 117(7):927.39. doi: 10.1016/j. cell.2004.06.006 PMID: 15210113.

19. Shipitsin M, Campbell LL, Argani P, Weremowicz S, Bloushtain-Qimron N, Yao J, et al. Molecular definition of breast tumor heterogeneity. Cancer Cell. 2007; 11(3):259.73. doi: 10.1016/j.ccr.2007.01.013 PMID: 17349583.
20. Hermann PC, Huber SL, Herrler T, Aicher A, Ellwart JW, Guba M, et al. Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. Cell Stem Cell. 2007; 1(3):313.23. doi: 10.1016/j.stem.2007.06.002 PMID: 18371365.
21. Gil M, Seshadri M, Komorowski MP, Abrams SI, Kozbor D. Targeting CXCL12/CXCR4 signaling with oncolytic virotherapy disrupts tumor vasculature and inhibits breast cancer metastases. Proc Natl Acad Sci U S A. 2013; 110(14):E1291.300. doi: 10.1073/pnas.1220580110 PMID: 23509246; PubMed Central PMCID: PMC3619300.
22. Rhodes LV, Short SP, Neel NF, Salvo VA, Zhu Y, Elliott S, et al. Cytokine receptor CXCR4 mediates estrogen-independent tumorigenesis, metastasis, and resistance to endocrine therapy in human breast cancer. Cancer Res. 2011; 71(2):603.13. doi: 10.1158/0008-5472.CAN-10-3185 PMID: 21123450; PubMed Central PMCID: PMC3140407.
23. Muller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME, et al. Involvement of chemokine receptors in breast cancer metastasis. Nature. 2001; 410(6824):50.6. doi: 10.1038/35065016 PMID: 11242036.

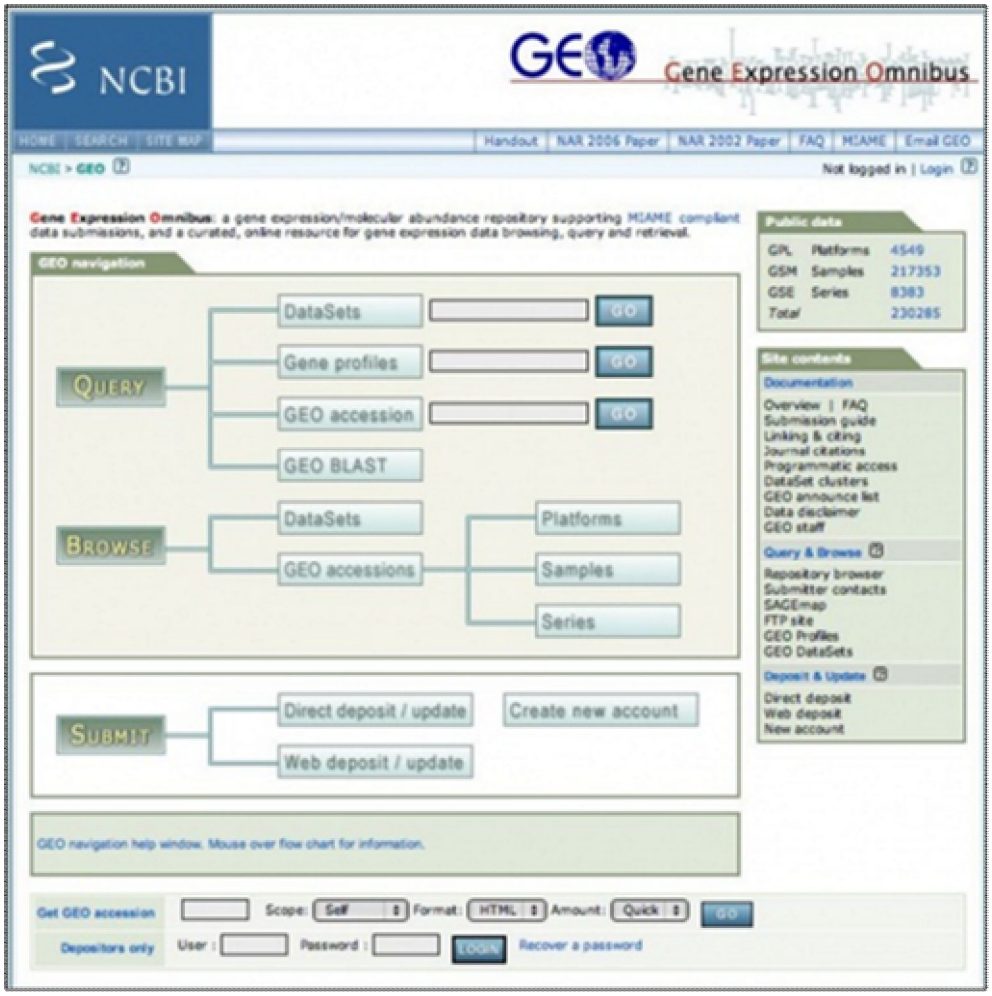
24. Kang Y, Siegel PM, Shu W, Drobnjak M, Kakonen SM, Cordon-Cardo C, et al. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell. 2003; 3(6):537.49. PMID: 12842083.
25. Kluger HM, Chelouche Lev D, Kluger Y, McCarthy MM, Kiriakova G, Camp RL, et al. Using a xenograft model of human breast cancer metastasis to find genes associated with clinically aggressive disease. Cancer Res. 2005; 65(13):5578.87. doi: 10.1158/0008-5472.CAN-05-0108 PMID: 15994930.

26. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, et al. Genes that mediate breast cancer metastasis to lung. Nature. 2005; 436(7050):518.24. doi: 10.1038/nature03799 PMID: 16049480; PubMed Central PMCID: PMC1283098.
27. Carbone C, Moccia T, Zhu C, Paradiso G, Budillon A, Chiao PJ, et al. Anti-VEGF treatment-resistant pancreatic cancers secrete proinflammatory factors that contribute to malignant progression by inducing an EMT cell phenotype. Clin Cancer Res. 2011; 17(17):5822.32. doi: 10.1158/1078-0432.CCR-11-1185 PMID: 21737511; PubMed Central PMCID: PMC3178272.
28. Kuo PL, Shen KH, Hung SH, Hsu YL. CXCL1/GROalpha increases cell migration and invasion of prostate cancer by decreasing fibulin-1 expression through NF-kappaB/HDAC1 epigenetic regulation. Carcinogenesis. 2012; 33(12):2477.87. doi: 10.1093/carcin/bgs299 PMID: 23027620.
29. Chen L, Xiao Z, Meng Y, Zhao Y, Han J, Su G, et al. The enhancement of cancer stem cell properties of MCF-7 cells in 3D collagen scaffolds for modeling of cancer and anti-cancer drugs. Biomaterials. 2012; 33(5):1437.44. doi: 10.1016/j.biomaterials.2011.10.056 PMID: 22078807.
30. Ponti D, Costa A, Zaffaroni N, Pratesi G, Petrangolini G, Coradini D, et al. Isolation and in vitro propagation of tumorigenic breast cancer cells with stem/progenitor cell properties. Cancer Res. 2005; 65 (13):5506.11. doi: 10.1158/0008-5472.CAN-05-0626 PMID: 15994920.
31. Hwang-Verslues WW, Kuo WH, Chang PH, Pan CC, Wang HH, Tsai ST, et al. Multiple lineages of human breast cancer stem/progenitor cells identified by profiling with stem cell markers. PLoS One. 2009; 4(12):e8377. doi: 10.1371/journal.pone.0008377 PMID: 20027313; PubMed Central PMCID: PMC2793431.
32. Wang R, Lv Q, Meng W, Tan Q, Zhang S, Mo X, et al. Comparison of mammosphere formation from breast cancer cell lines and primary breast tumors. Journal of thoracic disease. 2014; 6(6):829.37. doi: 10.3978/j.issn.2072-1439.2014.03.38 PMID: 24977009; PubMed Central PMCID: PMC4073404.

Ⅵ. 부록

1. The Gene Expression Omnibus (GEO)

본 연구에서 사용한 Genomics 관련 데이터베이스로 미국에서 운영되고 있으며 세계 최대 Genomics 데이터베이스 중 하나이다. 아래는 GEO에 관한 설명이다.



(GEO 웹사이트 홈페이지, 출처 : http://www.ncbi.nlm.nih.gov/geo/)

The Gene Expression Omnibus (GEO) is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data sets (1). Approximately 90% of the data in GEO are gene expression studies that investigate a broad range of biological themes including disease, development, evolution, immunity, ecology, toxicology, metabolism, and more. The non-expression data in GEO represent other categories of functional genomic and epigenomic studies including those that examine genome methylation, chromatin structure, genome copy number variations, and genome - protein interactions. A breakdown of GEO data types and technologies is provided on the repository Summary page.

Data in GEO represent original research submitted by the scientific community in compliance with grant or journal provisos that require data to be made available in a public repository, the objective being to facilitate independent evaluation of results, reanalysis, and full access to all parts of the study. The resource supports archiving of all parts of a study including raw data files, processed data, and descriptive metadata, which are indexed, cross-linked, and searchable. While the principal role of GEO is to serve as a primary data archive, the resource also offers several tools and features that allow users to explore, analyze, and visualize expression data from both gene-centric and study-centric perspectives.

To summarize, the main goals of GEO are to:

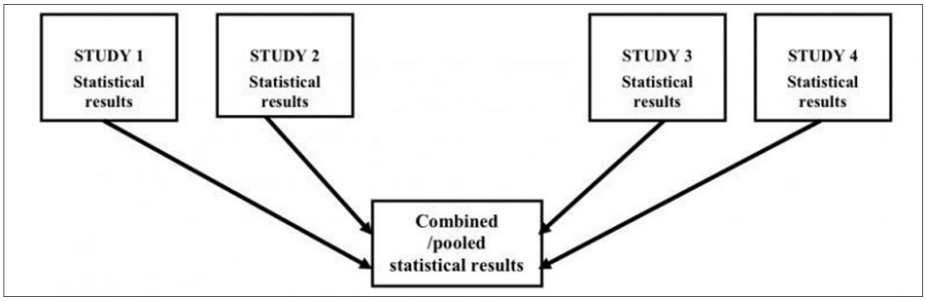
• Provide a robust, versatile primary data archive database in which to efficiently store a wide variety of high-throughput functional genomic data sets.

• Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community.
• Provide user-friendly mechanisms that allow users to locate, review, and download studies and gene expression profiles of interest.

(출처 : http://www.ncbi.nlm.nih.gov/books/NBK159736/)

2. 메타분석 (Meta-analysis)

본 연구에서 사용한 연구방법이다. 메타분석은 하나의 주제에 대해 출처가 다른 여러 연구에서 관련 데이터를 수집한 후 통계 알고리즘을 사용하여 데이터를 하나로 통합하고 통합된 데이터로부터 새로운 결론을 이끌어내는 방법이다. 메타분석은 실제 실험을 수행하지 않고 주요 과정인 데이터 처리가 컴퓨터에서 이뤄지므로 비용이 적게 들고 속도가 빠른 장점이 있다. 아래는 메타분석에 관한 설명이다.



(메타분석의 개요도,
출처 : https://hsl.lib.umn.edu/biomed/help/understanding-research-study-designs)

2-1. 메타분석의 배경

• 학문의 다양한 분야에서 쏟아져 나오는 정보의 홍수 속에서 단편적 연구의 발견점들을 통합하여 보다 객관적이며, 신뢰할 수 있는 강력한 결론을 도출하기 위한 노력은 1930년대부터 시작되었다고 볼 수 있음

• Hedges와 Olkin(1983)의 조사에 의하면 연구결과들을 통합하는 방법이 1930년대 농업분야의 실험결과들을 통합하기 위해 쓰여 지기 시작하였으며 그 이후 연구 문헌들을 종합, 정리하려는 노력은 초기의 연구 결과의 통계적 유의도를 분석하는 수준으로부터 차차 보다 세련된 측정 기법과 분석 기법이 적용되고 발달되었으며 Glass(1976)에 의해 메타분석 또는 통합적 분석으로 개념화된 이후에 본격적으로 활성화 됨

• Glass(1976)는 메타분석을 기존 연구에서 발견된 사항들의 통합을 목적으로 가지고 일련의 개별 연구들로부터 수립된 다양한 연구결과들을 통계적 분석하는 방법으로, “분석의 분석(analysis of analysis)”이라고 밝힘

• Hattie와 Hansford(1982)는 메타분석이란 많은 개별 연구 결과들의 결과를 하나로 종합하는 수량적인 접근 방법이라고 함

• 황정규(1988)는 메타분석이란 연구 결과들을 통합할 목적으로 많은 수의 개별적 연구 결과들을 통계적으로 분석하는 이론 및 방법이라고 정의함

2-2. 메타분석의 정의

메타분석 (meta-analysis) 이라는 것은.

1) 정의: 특정한 연구주제에 대해 행해진 여러 독립적인 연구의 결과를 합리적이고 체계적으로 종합하는 통계적 분석방법. 즉, 어떤 주제에 대해서 출판된 여러 논문의 결과들을 모으고 합쳐서 분석하는 방법임. Original article에서 환자 한명 한명을 엄격한 inclusion, exclusion criteria를 가지고 모아서 결과를 내듯이, meta-analysis에서는 논문 하나, 하나를 엄격한 inclusion, exclusion criteria를 가지고 모아서 결론을 도출하게 됨

2) 의의: 각각의 작은 연구들은 limitation이 있으므로, 같은 목적을 가지고 시행된 여러 연구들의 결과를 합리적이고 체계적으로 종합한다는데 의미

3) Meta-analysis가 적합한 자료의 조건 (1) 충분한 선행연구 결과물을 수집할 수 있어야 하며 (2) Effect size에 해당하는 값(mean, SD, OR, HR 등) 과 N수 p-value가 있어야 하고 (3) 연구논문의 질 판단이 가능해야 하며 (4) 논문 선정의 범위 문제가 명확해야 됨. (즉, publish된 논문만 할지, gray literature도 포함시킬지, 미발표된 논문들도 포함시킬지, 효과크기가 극단값인 연구도 넣을지 등에 대한 합리적인 기준을 마련할 수 있어야 함)

cf> publication bias란? positive result인 것은 잘 출판되는 경향이 있기 때문에, 출판된 논문들만 모아서 meta분석을 하면 결과가 positive하게 나오게 됨.

(출처 : http://openwiki.kr/med/meta-analysis#메타분석의_개념)

3. Affymetrix

본 연구에서는 메타분석을 수행하기 위하여 Gene Expression Omnibus (GEO)에서 총 3가지 microarray 데이터를 얻었으며 이 데이터는 모두 Affymetrix 회사의 microarray chip에서 유래된 데이터이다. Affymetrix는 microarray를 제조하는 미국 회사로 아래는 Affymetrix 회사에 대한 설명이다.

Affymetrix, Inc. is an American company that manufactures DNA microarrays; it is based in Santa Clara, California, United States. The company was founded by Dr. Stephen Fodor in 1992. It began as a unit in Affymax N.V. in 1991 by Fodor's group, which had in the late 1980s developed methods for fabricating DNA microarrays, called "GeneChip" according to the Affymetrix trademark, using semiconductor manufacturing techniques. The company's first product, an HIV genotyping GeneChip, was introduced in 1994 and the company went public in 1996. As a result of its pioneering work and the ensuing popularity of microarray products, Affymetrix derives significant benefit from its patent portfolio in this area.
Affymetrix has taken over Genetic MicroSystems for slide-based microarrays and scanners, Neomorphic for bioinformatics, ParAllele Bioscience for custom SNP genotyping, USB/Anatrace for biochemical reagents, eBioscience for flow cytometry, and Panomics and True Materials to expand its offering of low to mid-plex applications. In 2000, Perlegen Sciences spun out from Affymetrix to focus on wafer-scale genomics for massive data

(Affymetrix에서 생산하는 microarray chip,
출처 : http://www.genehk.com/news/news_20140205_5.php)

creation and collection required for characterizing population variance of genomic markers and expression for the drug discovery process.

Thermo Fisher Scientific announced on January 8, 2016 the acquisition of Affymetrix for approximately $1.3 billion. The acquisition is anticipated to close in the second quarter of 2016. The acquisition will enhance their diagnostic equipment offerings and lead to 10s of millions of dollars in savings in the first few years after completion of the deal.

(출처 : https://en.wikipedia.org/wiki/Affymetrix)

4. ComBat method (R 설치 및 R package 설치)

본 연구에서는 Gene Expression Omnibus (GEO)에서 얻은 3개의 데이터를 하나의 데이터로 통합하기 위해 ComBat method를 사용하여 batch effect를 제거하였다. 여기서 batch effect는 메타분석에서 사용된 용어로 일반적으로 출처가 다른 여러 데이터를 하나로 합할 때 연구자가 고려하고자 하는 변수 (연구자가 실험설계를 하면서 디

자인한 실험군, 대조군에 해당되는 변수) 이외에 불필요한 변수 (예를 들어 데이터를 생성한 날짜, 사용된 서로 다른 기기 등)에 의해 데이터에 나타나는 영향을 의미한다. 예를 들어 그림 1A는 서로 다른 출처 3개의 데이터를 하나로 합하여 clustering을 그린 결과이고 그림 1A에서처럼 데이터의 분포가 연구자가 고려하는 변수 Time, Temperature에 의한 영향 보다는 batch에 의한 영향이 더 큰 것을 볼 수 있다. 만약 연구자가 이 데이터를 추후 분석에 사용하게 될 경우 분석결과는 연구자가 고려하고자 했던 Time, Temperature와 상관없는 batch에 의한 분석결과가 나오게 되므로 연구자는 반드시 메타분석에서 데이터를 합한 후 batch effect를 제거해주는 과정을 거쳐야 한다. 그림 1B는 batch effect를 통계 알고리즘을 사용하여 제거해 준 결과이다. 그림 1A에서 같은 batch의 데이터가 cluster를 이루는 것과 비교해서 그림 2B에서는 이러한 batch effect는 제거되고 Time과 Treatment에 의해 데이터가 분포됨을 알 수 있다.

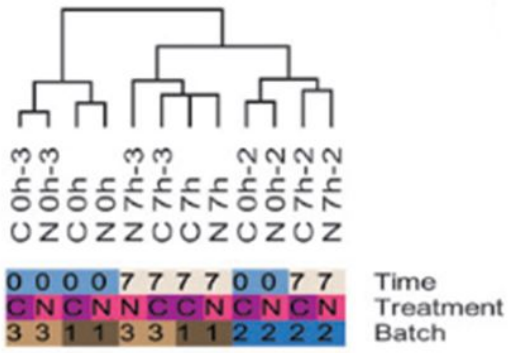


그림 1A. Clustering에 존재하는 batch effects (출처 : Leek et. al 2010)

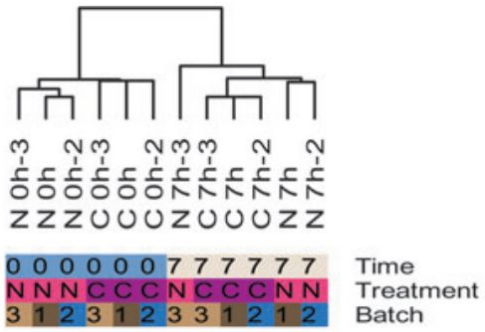


그림 1B. Clustering에 존재하는 batch effects (출처 : Leek et. al 2010)

본 연구에서는 batch effect를 제거하기 위해 Combat method라는

통계 알고리즘을 사용하였으며 이 알고리즘은 R package인 SVA package에 의해 실행되었다. R package는 R 프로그램(https://www.r-project.org/)에서 지원하는 package 형식의 software로 R 프로그램은 통계계산과 그래픽을 위한 프로그램이며 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있다. R 프로그램은 open source 기반이며 프로그래밍 언어로 실행이 된다. 사용자는 자신이 필요한 통계 알고리즘을 R 프로그램에서 package 형태로 실행이 가능하며 2016년 1월 기준으로 7,801개의 package가 발표되었다.

SVA R package 사용법을 간단하게 소개하면 우선 package는 R 프로그램 환경에서 실행이 되므로 R 프로그램을 설치해야 한다. R 프로그램 설치파일을 다운 받기 위해 https://cloud.r-project.org/에 접속하여 그림 2에서 Download R for Window를 선택한다.

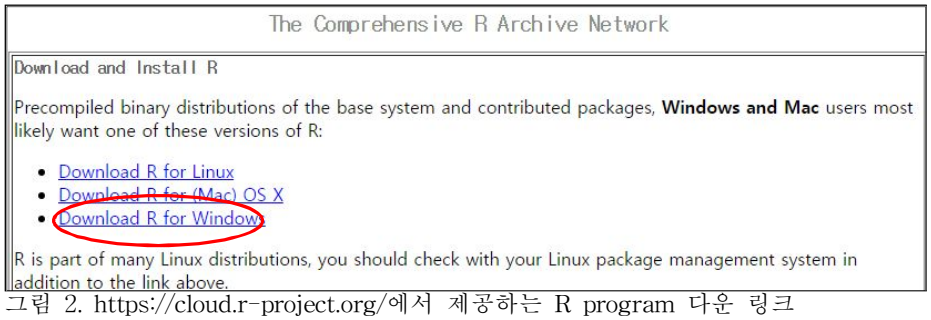


그림 2. https://cloud.r-project.org/에서 제공하는 R program 다운 링크

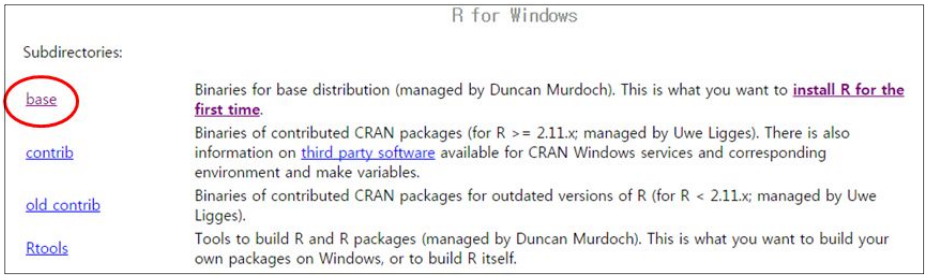


그림 3. R program 다운로드 링크

Download R for Window를 선택하게 되면 그림 3과 같은 화면이 나오고 여기서 다시 base를 선택한다.

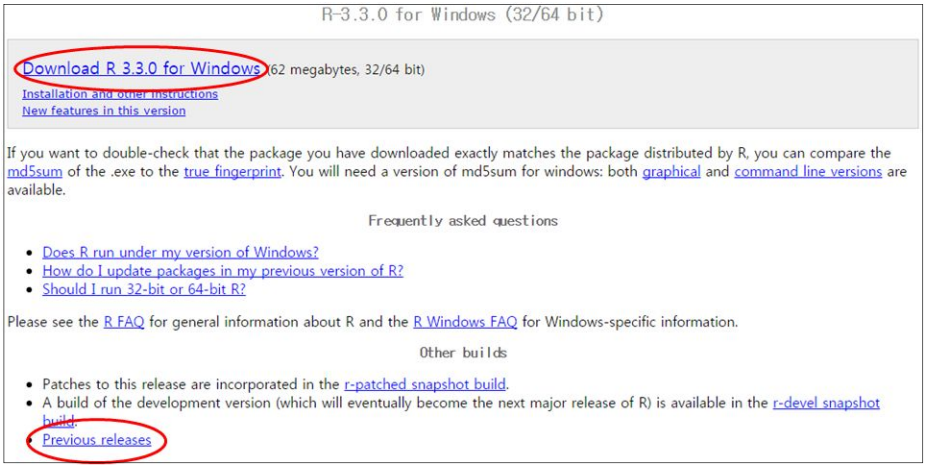


그림 4. R 버전에 따른 다운로드 링크

그림 3에서 base 링크를 선택하면 그림 4와 같이 상단에 최신 버전의 R 프로그램 다운로드 링크와 하단에 이전 버전의 R 프로그램 다운로드 링크가 보이게 된다. 여기서는 이전 버전의 R 프로그램을 사용해서 Combat method를 실행할 예정이므로 Previous release 링크를 선택한다. Previous release를 선택하면 그림 5와 같이 이전 버전에 해당하는 리스트가 나타나고 여기서는 R.2.15.3을 다운받아 설치할 예정이므로 R.2.15.3을 선택 한다 (R 버전에 따라 사용할 수 있는 package의 종류가 다르며 실행하는 코드도 달라질 경우가 발생하므로 자신이 사용하고자 하는 package에 맞게 R 프로그램 버전을 선택하여 설치해야한다).

R.2.15.3을 선택하게 되면 해당 설치 파일의 다운로드가 시작되고 다운로드가 완료된 후 실행하면 그림 6과 같은 화면이 나타난다. 설치 시 '다음' 버튼을 계속 누르게 되면 사용자 컴퓨터 사양에 맞게 설치가 완료된다.

그림 5. R 버전에 따른 다운로드 링크

R 프로그램의 설치가 완료되고 프로그램을 실행하게 되면 그림 7과 같이 command 실행 창이 나타나게 된다.

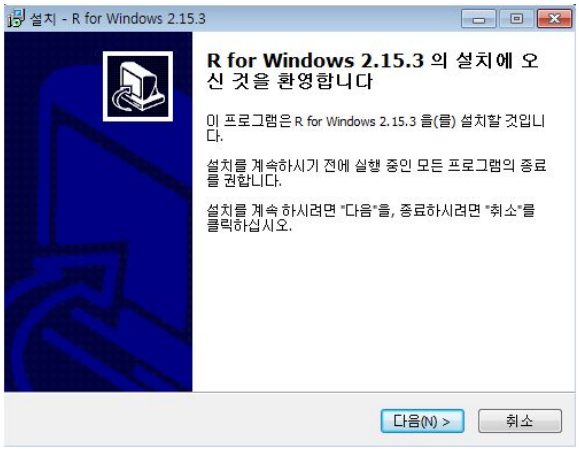


그림 6. R 프로그램 설치 파일 실행

R 프로그램 설치 및 실행이 완료되면 Combat method를 실행하기 위해 SVA R package를 설치해야 한다. SVA R package는 Bioconductor (https://bioconductor.org/)라는 웹 사이트에서 무료로 제공하며 이 웹사이트는 생물정보학 분야에서 사용하는 통계 알고리즘을 R package 형태로 제공하는 open source 기반의 웹사이트이다. Bioconductor의 R package들을 R program에 쉽게 설치할 수 있도록 R program에서는 간단한 설치 코드를 아래와 같이 제공하고 있다.

source("http://bioconductor.org/biocLite.R")
biocLite("sva")

이 코드는 R package 중에서 SVA package를 설치하는 코드로 이 코드를 복사해서 R 프로그램에 붙여넣기를 하면 자동으로 SVA package가 R 프로그램에 설치가 된다. SVA package에 대한 정보는 Bioconductor (https://bioconductor.org/packages/release/bioc/html/sva.html)에서 제공하고 있으며 SVA package에 사용되는 통계적 배경, work flow, 실제 예제와 같은 내용이 pdf 파일로 제공되고 있다. 본 연구에서는 이 정보를 바탕으로 분석하고자 했던 데이터에 맞게 코드를 작성하여 Combat method를 실행하였다.

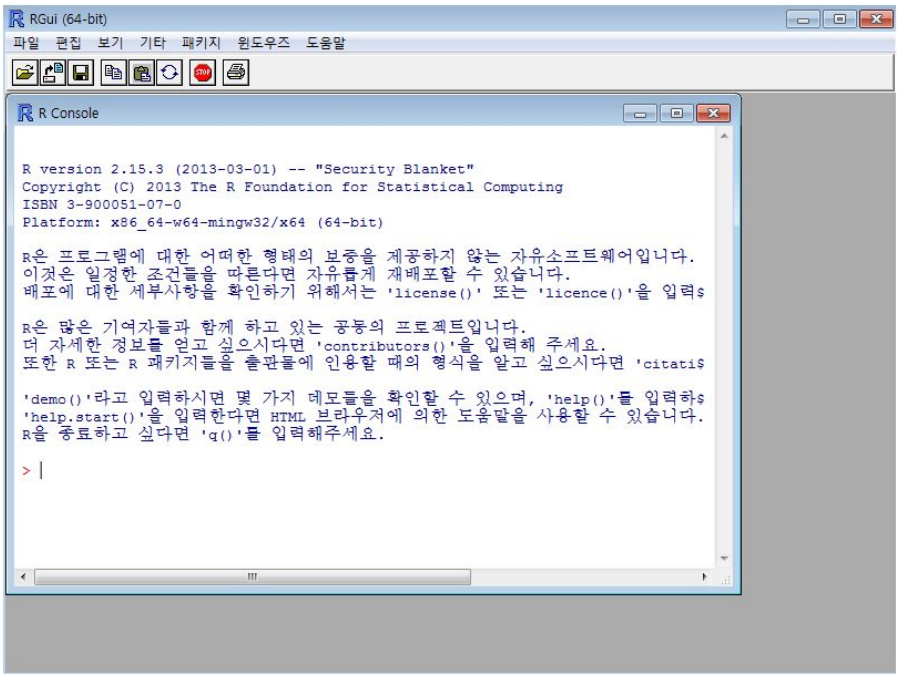


그림 7. R 프로그램 실행 시 나타나는 command 창

5. Illumina

본 연구의 메타분석에서는 아주대학교에서 제공한 Illumina platform의 데이터를 사용하였다. Illumina는 Affymetrix와 마찬가지로 microarray를 제조하는 미국의 회사이며 microarray 이외에 DNA Sequencing 기술 관련 제조 회사로도 알려져 있다. 아래는 Illumina 회사에 대한 설명이다.

Illumina, Inc. is an American company incorporated in April 1998 that develops, manufactures and markets integrated systems for the analysis of genetic variation and biological function. In 2014, Illumina was named the world's smartest company by MIT Technology Review. Using its technologies, the company provides a line of products and services that serve the sequencing, genotyping and gene expression markets. This technology had purportedly by 2013 reduced the cost of sequencing a human genome to US$4,000, down from a price of US$1 million in 2007. Customers include genomic research centers, pharmaceutical companies, academic institutions, clinical research organizations and biotechnology companies. Its tools provide researchers with the capability to perform genetic tests needed to extract medical infor

(Illumina에서 생산되는Arrays
출처: : https://www.eurofinsgenomics.eu/en/genotyping-gene-expression/service-platforms/illumina-array-platforms.aspx)

mation from advances in genomics and proteomics. Its headquarte rs are located in San Diego, California.

5-1. History

Illumina was founded in April 1998 by David Walt, Larry Bock, John Stuelpnagel, Anthony Czarnik, and Mark Chee. While working with CW Group, a venture capital firm, Bock and Stuelpnagel uncovered what would become Illumina's BeadArray technology at Tufts University and negotiated an exclusive license to that technology. Illumina completed its initial public offering in July 2000.

Illumina began offering single nucleotide polymorphism (SNP) genotyping services in 2001 and launched its first system, the Illumina BeadLab, in 2002, using GoldenGate Genotyping technology. Illumina currently offers microarray-based products and services for an expanding range of genetic analysis sequencing, including SNP genotyping, gene expression, and protein analysis. Illumina's technologies are used by a broad range of academic, government, pharmaceutical, biotechnology, and other leading institutions around the globe.

On January 26, 2007, the Company completed the acquisition of Hayward based Solexa, Inc. Solexa Ltd,based in Cambridge UK was founded in June 1998 by Shankar Balasubramanian, and David Klenerman to develop and commercialize genome sequencing technology invented by the founders in the University of Cambridge. Solexa INC was formed 2005 when Solexa Ltd reversed into Lynx Therapeutics of Hayward. technology uses also the DNA colony sequencing technology, invented in 1997 by

Pascal Mayer and Laurent Farinelli and which was acquired by Solexa in 2004 from the company Manteia Predictive Medicine. It is being used to perform a range of analyses, including whole genome resequencing, gene expression analysis and small ribonucleic acid (RNA) analysis.

In June 2009, Illumina announced the launch of their own Personal Full Genome Sequencing Service at a depth of 30X for $48,000 per genome, and a year later dropped the price to $19,500. This is still too expensive for true commercialization but the price will most likely decrease substantially over the next few years as they realize economies of scale and given the competition with other companies such as Complete Genomics and Knome. As of May 2011, Illumina reduced the price to $4,000. Illumina acquired Epicentre Biotechnologies, based in Madison, Wisconsin, on January 11, 2011. On January 25, 2012, Hoffmann-La Roche made an unsolicited bid to buy Illumina for $44.50 per share or about $5.7 billion. Roche tried other tactics, including raising its offer (to $51.00, for about $6.8 billion). Illumina rejected the offer, and Roche abandoned the offer in April. As of April 2013, the company's chief executive officer was Jay Flatley. In 2014, the company announced a multimillion-dollar product, HiSeq X Ten, that it forecast would provide large-scale whole-genome sequencing for $1,000/genome. The company claimed that forty such machines would be able to sequence more genomes in one year than had been produced by all other sequencers to date. In January 2014, Illumina already held 70 percent of the market for genome-sequencing machines. Illumina machines accounted for more than 90 percent of all DNA data produced.

5-2. Products

1) Golden Gate Methylation

The GoldenGate Methylation Cancer Panel allows the user to probe 1,505 CpG loci selected from 807 genes across a large sample size. The array based method allows 96 samples to be probed simultaneously on one array matrix.

2) Infinium methylation

Utilizing Illumina's HumanMethylation27 DNA Analysis BeadChip and the Infinium technology, this method allows the user to map single methylation resolution for 27,578 CpG sites across over 14,000 genes. This Chip has been replaced by the 450K Methylation Chip and later by the EPIC Array, covering about 850k sites.

3) DNA sequencing

Illumina sells a number of very high-throughput DNA sequencing systems, also known as DNA sequencers, based on technology developed by Solexa. The technology features bridge amplification to generate clusters and reversible terminators for sequence determination. The technology behind these sequencing systems involves ligation of fragmented DNA to a chip, followed by primer addition and sequential fluorescent dNTP incorporation and detection.

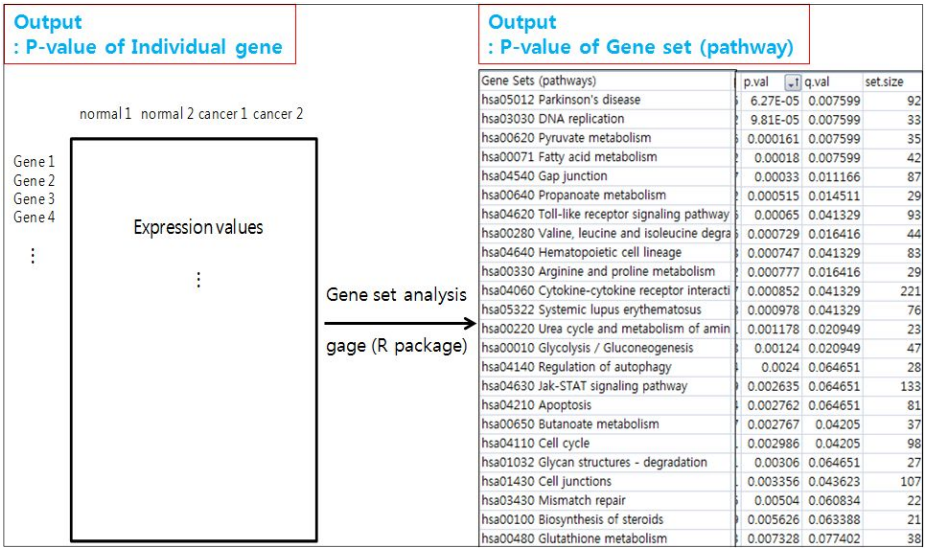
(출처 : https://en.wikipedia.org/wiki/Illumina_(company))

6. Gene set analysis

일반적으로 유전자 발현 분석 (gene expression analysis)에서는 개별 유전자에 해당하는 p-value를 sample들의 발현 정보를 이용하여 계산하고 이를 통해 유의성을 보이는 '개별 유전자'들을 마커로서 제시하게 된다. 하지만 gene set analysis는 개별 유전자들에 대한 p-value를 계산하는 대신에 개별 유전자들의 묶음인 gene set을 기준으로 p-value를 계산하는 알고리즘이다. 본 연구에서 사용한 gene set 개념은 pathway 개념과 같으며 이 pathway는 KEGG (Kyoto Encyclopedia of Genes and Genomes)에서 정한 기준을 사용하였다. Gene set, 즉 pathway는 특정 생물학적 현상이나 질병의 메커니즘에 해당하는 정보를 포함하고 있기 때문에 gene set analysis를 이용하여 분석결과를 얻게 되면 기존의 유전자 발현 분석에서 사용되었던 산술적인 유의성 뿐 만 아니라 생물학적 정보까지 고려된 결과를 얻을 수 있으며 최종적으로 분석결과의 생물학적 해석이 용이하게 된다. 아래는 gene set analysis에 관한 설명이다.

Gene set enrichment (also functional enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. Microarray and proteomics results often identify thousands of genes which are used for the analysis. Researchers performing high-throughput experiments that yield sets of genes (for example, genes that are differentially expressed under different conditions) often want to retrieve a functional profile of that gene set, in order to better understand the underlying biological processes.

(출처 : https://en.wikipedia.org/wiki/Gene_set_enrichment)



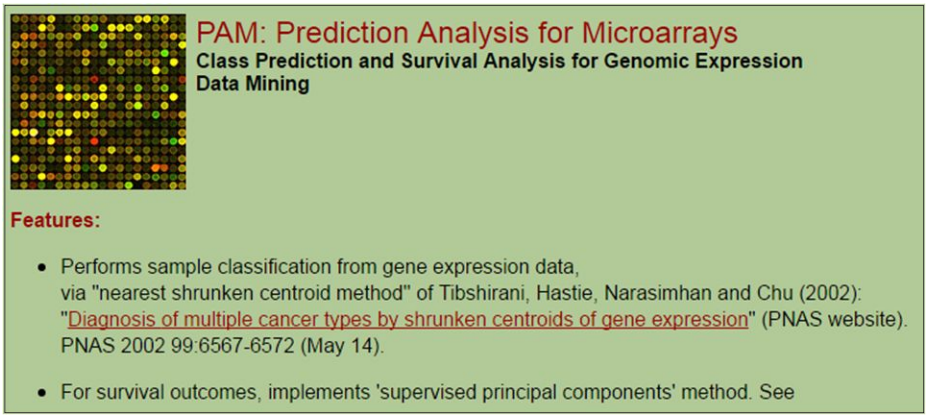
Gene Sets (pathways)	p.val	q.val	set.size
hsa05012 Parkinson's disease	6.27E-05	0.007599	92
hsa03030 DNA replication	9.81E-05	0.007599	33
hsa00620 Pyruvate metabolism	0.000161	0.007599	35
hsa00071 Fatty acid metabolism	0.00018	0.007599	42
hsa04540 Gap junction	0.00033	0.011166	87
hsa00640 Propanoate metabolism	0.000515	0.014511	29
hsa04620 Toll-like receptor signaling pathway	0.00065	0.041329	93
hsa00280 Valine, leucine and isoleucine degra	0.000729	0.016416	44
hsa04640 Hematopoietic cell lineage	0.000747	0.041329	83
hsa00330 Arginine and proline metabolism	0.000777	0.016416	29
hsa04060 Cytokine-cytokine receptor interacti	0.000852	0.041329	221
hsa05322 Systemic lupus erythematosus	0.000978	0.041329	76
hsa00220 Urea cycle and metabolism of amin	0.001178	0.020949	23
hsa00010 Glycolysis / Gluconeogenesis	0.00124	0.020949	47
hsa04140 Regulation of autophagy	0.0024	0.064651	28
hsa04630 Jak-STAT signaling pathway	0.002635	0.064651	133
hsa04210 Apoptosis	0.002762	0.064651	81
hsa00650 Butanoate metabolism	0.002767	0.04205	37
hsa04110 Cell cycle	0.002986	0.04205	98
hsa01032 Glycan structures - degradation	0.00306	0.064651	27
hsa01430 Cell junctions	0.003356	0.043623	107
hsa03430 Mismatch repair	0.00504	0.060834	22
hsa00100 Biosynthesis of steroids	0.005626	0.063388	21
hsa00480 Glutathione metabolism	0.007328	0.077402	38

그림 8. 왼쪽 그림은 gene set analysis에서 사용되는 input 데이터의 format이며 이 input 데이터를 일반적인 유전자 발현 분석에 사용할 경우 개별 유전자에 해당하는 p-value를 얻지만 gene set analysis를 적용할 경우 오른쪽과 같이 gene set (pathway)에 해당하는 p-value를 얻을 수 있다.

본 연구에서는 gene set analysis를 실행하기 위해 R package ‘gage ’를 사용하였다. gage R package는 Bioconductor (http://bioconductor.org/packages/release/bioc/html/gage.html)에서 무료로 다운받아 사용이 가능하며 이곳에서 package에 대한 통계적 배경, work flow, 실제 사용되는 코드가 포함된 guideline을 제공하고 있다. 본 연구는 Bioconductor가 제공하는 gage R package의 guideline을 참고하여 사용했던 데이터에 맞게 코드를 작성하고 gene set analysis를 진행하였다.

7. Prediction analysis for microarrays (PAM)

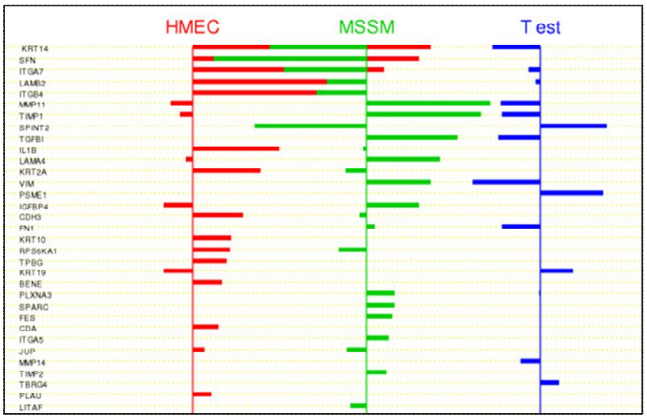
PAM은 본 연구에서 gene set analysis를 통해 얻은 significant gene set의 통계적 검증을 위해 사용한 소프트웨어이다. PAM은 유전자 발현 데이터의 class 예측을 할 때 사용되는 프로그램으로 nearest shrunken centroid method라는 통계 알고리즘을 사용하여 prediction model을 만들고 이 prediction model을 이용하여 test sample의 class를 예측하는 방법으로 진행된다. Nearest shrunken



(Prediction Analysis for Microarrays 웹사이트 홈페이지,
출처 : http://statweb.stanford.edu/~tibs/PAM/)

centroid는 class 예측방법의 하나로 각 class에 대해 가장 큰 특징을 가지는 변수들을 구별하며 이 방법은 고차원 분류의 문제들에서도 사용 가능하다. Nearest shrunken centroid 분류의 계산 방법은 훈련자료에서 전체 표본에 대한 변수들의 발현값에 대해 각 중심값(변수의 평균 발현값)을 뺀 후, nearest centroid 분류를 적용해서 테스트 표본의 변수 발현값과 각 클래스의 중심값의 거리를 계산하여 거리가 가장 짧은 클래스에 분류하는 방법이다. Nearest shrunken centroid 분류와 같이 고차원적인 특성들을 지닌 클래스에 대한 예측은 중요한 문제이며 최근에는 cDNA 마이크로어레이(Microarray)에서 유전자 발현 profile을 기초로 표본(sample)들의 진단에 도움이

되는 범주를 분류하고 예측하는데 사용되고 있다.



Plot of the shrunkencentroids for the specified threshold with the ranked list of significant genes that are used for the predictions

(출처 : http://dspace.inha.ac.kr/pdfupload/15715.pdf, https://www.researchgate.net/figure/7153056_fig1_PAM-Prediction-Analysis-of-Microarrays-This-is-a-statistical-technique-for-class)

PAM은 open source 기반의 무료로 제공되는 프로그램으로 누구나 다운받아서 사용가능하며 본 연구에서는 PAM을 R package인 'pamr'을 사용하여 실행하였다. pamr을 R에 설치하는 파일은 https://cran.r-project.org/web/packages/pamr/index.html에서 다운받을 수 있으며 다운로드가 끝나면 그림 9와 같이 메뉴창에서 '로컬에 있는 zip 파일들로부터 패키지(를) 설치'를 선택하여 pamr을 설치하면 된다.

pamr 이외의 무료로 배포되는 R package는 대부분 Google을 통해 검색 및 다운로드가 가능하다. pamr을 다운받았던 url 주소는 pamr에 해당하는 매뉴얼, work flow, 실제 예제 코드에 관한 파일을 제공하고 있으며 본 연구에서는 이 정보를 바탕으로 분석하고자 했던 데이터에 맞게 R 코드를 작성하여 데이터 분석을 진행하였다.

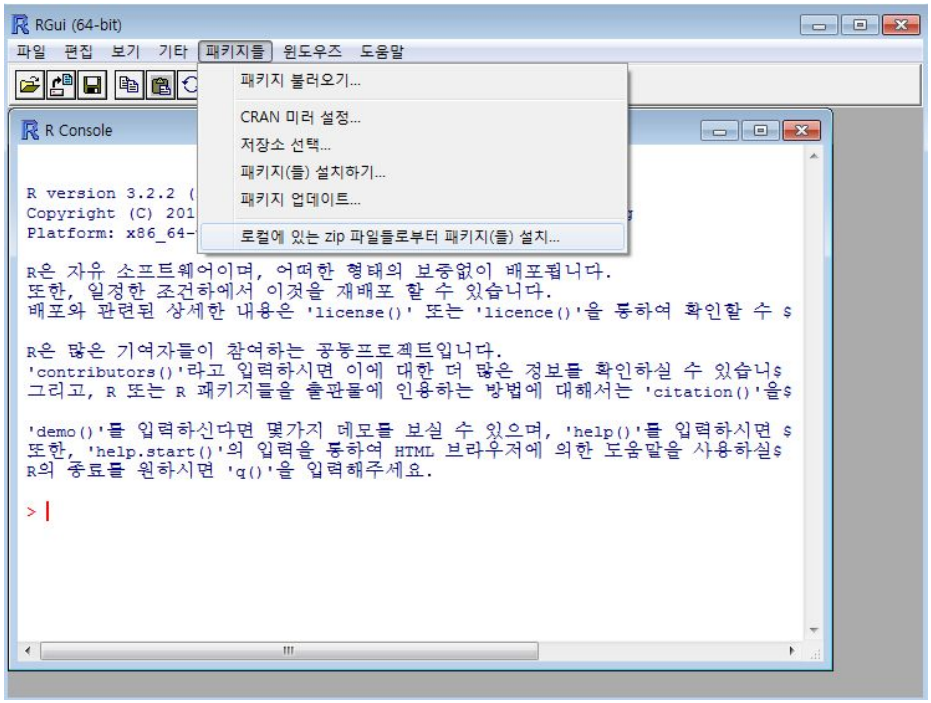


그림 9. Open source에서 다운받은 R package를 설치하는 메뉴

8. Globaltest

R package인 Globaltest는 본 연구에서 gene set analysis 결과를 통해 얻은 4개의 significant gene set에서 개별 유전자 candidate를 선별할 때 사용한 소프트웨어로 Bioconductor 웹사이트 (http://bioconductor.org/packages/release/bioc/html/globaltest.html)에서 무료로 제공한다. Package 사용에 관한 매뉴얼, work flow, 실제 예제 코드도 같이 첨부하여 제공하고 있으며 본 연구에서는 이 정보를 바탕으로 분석하고자 했던 데이터에 맞게 코드를 작성하여 Globaltest를 실행하였다. Globaltest는 개별 유전자에 대한 p, q-value를 sample의 expression value를 이용하여 결과값으로 계산하고 control과 비교하여 treatment에서 발현의 증감 정보를 제공한

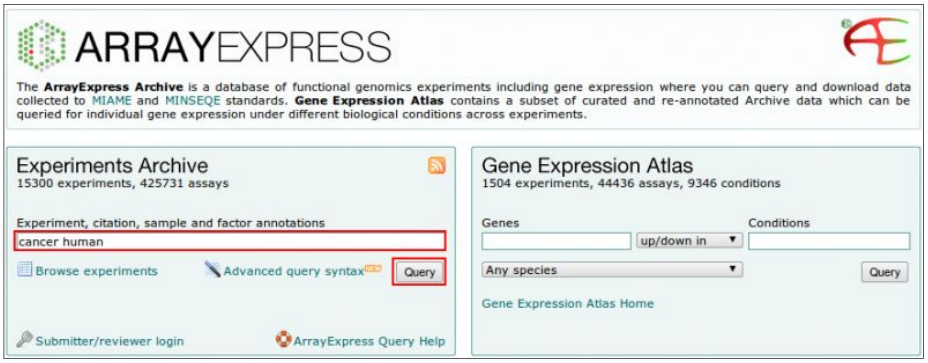
다. 본 연구에서는 Globaltest에서 제공하는 p-value와 발현 증감정보를 이용하여 gene plot을 얻었으며 gene plot 역시 Globaltest가 제공하는 guideline을 참고하여 생성하였다.

9. ArrayExpress

ArrayExpress는 본 연구의 메타분석에서 데이터 수집을 위해 Gene Expression Omnibus(GEO)와 함께 사용했던 Genomics 데이터베이스이다. 아래는 ArrayExpress에 대한 설명이다.

ArrayExpress is a new public database of microarray gene expression data at the EBI, which is a generic gene expression database designed to hold data from all microarray platforms. ArrayExpress uses the annotation standard Minimum Information About a Microarray Experiment (MIAME) and the associated XML data exchange format Microarray Gene Expression Markup Language (MAGE-ML) and it is designed to store well annotated data in a structured way. The ArrayExpress infrastructure consists of the database itself, data submissions in MAGE-ML format or via an online submission tool MIAMExpress, online database query interface, and the Expression Profiler online analysis tool. ArrayExpress accepts three types of submission, arrays, experiments and protocols, each of these is assigned an accession number. Help on data submission and annotation is provided by the curation team. The database can be queried on parameters such as author, laboratory, organism, experiment or array types. With an increasing number of organisations adopting MAGE-ML standard, the volume of submissions to ArrayExpress is increasing rapidly. The database can be accessed at

http://www.ebi.ac.uk/arrayexpress.
(출처 : http://www.ncbi.nlm.nih.gov/pubmed/12519949)



(ArrayExpress 웹사이트 홈페이지,
출처 : http://bioinfo.cipf.es/babelomicstutorial/data_downloading)

10. affy

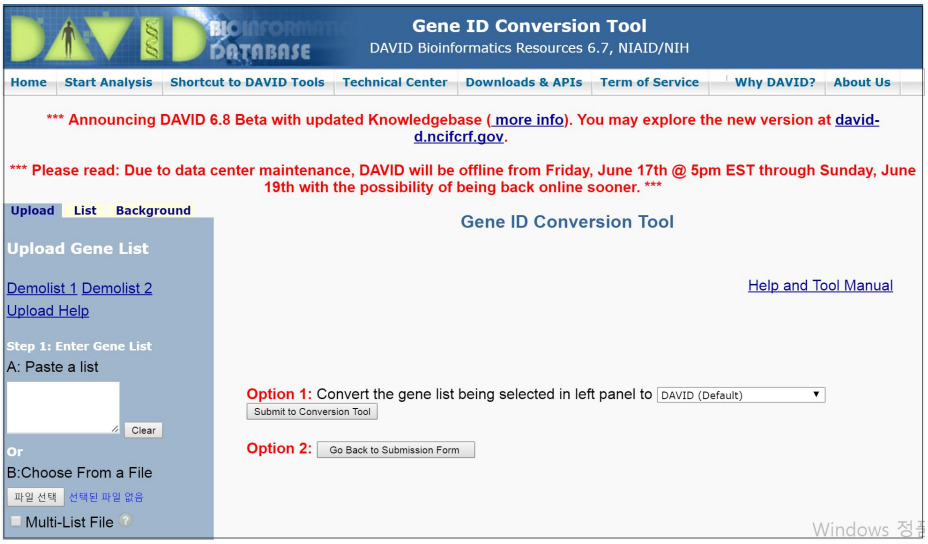
본 연구에서 사용된 Affymetrix data의 표준화를 위해 사용한 R package이다. Affymetrix에서 제공하는 파일의 확장자가 '.CEL'인 유전자 발현 데이터는 R 프로그램에서 affy라는 package를 이용하여 importing 및 normalization을 실행할 수 있다. 이 package는 Biocondoctor (http://bioconductor.org/packages/release/bioc/html/affy.html) 에서 무료로 R에 설치하여 사용가능하며 package에 대한 매뉴얼, work flow, 실제 예제 코드를 제공하고 있다. 본 연구에서는 이 매뉴얼을 참고하여 분석하고자 하는 데이터에 맞게 코드를 작성하고 분석을 진행하였다.

11. DAVID

Database for Annotation, Visualization, and Integrated Discovery (DAVID)는 본 연구에서 수집된 유전자 발현 데이터의 유전자 ID mapping에 사용했던 웹사이트이다. 본 연구에서는 두 가지 platform, Affymetrix와 Illumina 데이터를 사용하였으며 두 platform은 서로 다른 유전자 ID를 포함하기 때문에 DAVID를 이용하여 Entrez ID로 유전자 ID를 통일시켰다. 메타분석에서는 서로 다른 출처의 데이터를 사용하기 때문에 유전자 ID 통일 과정은 반드시 거쳐야하는 단계이다. 아래는 DAVID에 관한 설명이다.

DAVID (the Database for Annotation, Visualization and Integrated Discovery) is a free online bioinformatics resource developed by the Laboratory of Immunopathogenesis and Bioinformatics. All tools in the DAVID Bioinformatics Resources aim to provide functional interpretation of large lists of genes derived from genomic studies, e.g. microarray and proteomics studies. DAVID can be found at http://david.niaid.nih.gov or http://david.abcc.ncifcrf.gov

The DAVID Bioinformatics Resources consists of the DAVID Knowledgebase and five integrated, web-based functional annotation tool suites: the DAVID Gene Functional Classification Tool, the DAVID Functional Annotation Tool, the DAVID Gene ID Conversion Tool, the DAVID Gene Name Viewer and the DAVID NIAID Pathogen Genome Browser. The expanded DAVID Knowledgebase now integrates almost all major and well-known public bioinformatics resources centralized by the DAVID Gene



(DAVID 웹사이트,
출처 : https://david.ncifcrf.gov/conversion.jsp)

Concept, a single-linkage method to agglomerate tens of millions of diverse gene/protein identifiers and annotation terms from a variety of public bioinformatics databases. For any uploaded gene list, the DAVID Resources now provides not only the typical gene-term enrichment analysis, but also new tools and functions that allow users to condense large gene lists into gene functional groups, convert between gene/protein identifiers, visualize many-genes-to-many-terms relationships, cluster redundant and heterogeneous terms into groups, search for interesting and related genes or terms, dynamically view genes from their lists on bio-pathways and more.
(출처 : https://en.wikipedia.org/wiki/DAVID)

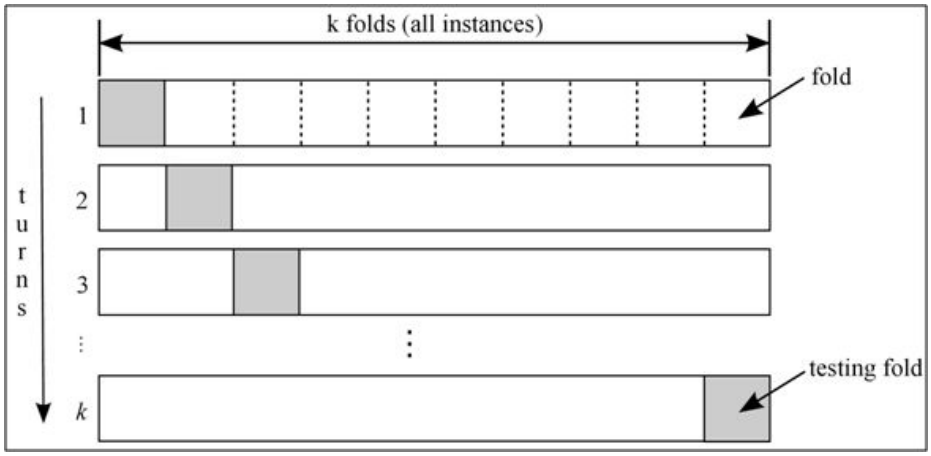
12. Entrez ID

일반적으로 microarray 데이터와 같은 유전자 발현 데이터는 그 데이터를 생산한 회사에 따라 서로 다른 유전자 ID로 구성되게 된다. Enrez ID는 Entrez Global Query Cross-Database Search System에서 제공하는 유전자 ID 중 하나로 본 연구에서는 서로 다른 출처의 유전자 발현 데이터를 통합할 때 유전자 ID를 하나로 통합하기 위해 사용하였다. 본 연구의 gene set analysis를 위해 사용한 gage R package는 Kyoto Encyclopedia of Genes and Genomes (KEGG)가 분류해놓은 pathway에 대한 대조군과 실험군 사이의 p-value를 계산하게 된다. gage 알고리즘에 input으로 들어가는 유전자 발현 데이터는 유전자 ID를 Entrez ID로 변환해주어야 gage 알고리즘이 유전자를 pathway의 구성원으로 인식하고 해당 유전자의 발현값을 pathway에 대한 p-value값을 계산하는데 사용할 수 있게 된다. GEO database에서 Affymetrix와 Illumina의 raw data를 다운받아 바로 열어보게 되면 유전자 ID는 각각 Affymetrix와 Illumina platform에서 제공하는 포맷의 유전자 ID로 구성되게 된다. 본 연구에서는 gage R package를 이용하여 gene set analysis를 실행하기 위해 DAVID를 이용하여 raw data의 유전자 ID를 모두 Enrez ID로 변환해주었다.

13. Leave-one-out cross validation

본 연구에서는 gene set analysis에서 선별된 significant gene set을 통계적으로 검증하기 위해 Leave-one-out cross validation을 사용하였다. 이 방법은 선별된 significant gene set의 유의성을 평가하기 위한 방법으로 본 연구에서는 prediction analysis for microarrays (PAM)을 이용하여 만든 prediction model을 가지고 Leave-one-out

cross validation을 진행하였다. 진행방법은 k개의 sample이 있다고 가정할 때 그 중 1개의 sample을 제외시키고 나머지 k-1개의 sample을 이용하여 prediction model을 만들고 이 prediction model을 이용하여 제외시켰던 sample의 class를 진행하였다. 이때 prediction model의 modeling은 유의성을 평가하고자 하는 gene set의 발현값을 이용하며 예측하고자 하는 sample (제외시켰던 sample)의 발현값 역시 유의성을 평가하고자 하는 gene set의 발현값으로 구성되어야 한다. 이 과정을 k개의 모든 샘플에서 총 k번 진행하고 이 결과를 이용하여 유의성을 평가하고자 하는 각각의 gene set에서 accuracy값을 얻었다.



(Leave-one-out cross validation 개요도,
출처 : http://www.intechopen.com/books/advances-in-data-mining-knowledge-discovery-and-applications/selecting-representative-data-sets)

14. C-X-C chemokine receptor type 4 (CXCR4)

C-X-C chemokine receptor type 4 (CXCR-4) also known as fusin or CD184 (cluster of differentiation 184) is a protein that in humans is encoded by the CXCR4 gene.

14-1. Function

CXCR-4 is an alpha-chemokine receptor specific for stromal-derived-factor-1 (SDF-1 also called CXCL12), a molecule endowed with potent chemotactic activity for lymphocytes. CXCR4 is one of several chemokine receptors that HIV can use to infect CD4+ T cells. HIV isolates that use CXCR4 are traditionally known as T-cell tropic isolates. Typically, these viruses are found late in infection. It is unclear as to whether the emergence of CXCR4-using HIV is a consequence or a cause of immunodeficiency.

CXCR4 is upregulated during the implantation window in natural and hormone replacement therapy cycles in the endometrium, producing, in presence of a human blastocyst, a surface polarization of the CXCR4 receptors suggesting that this receptor is implicated in the adhesion phase of human implantation.

CXCR4's ligand SDF-1 is known to be important in hematopoietic stem cell homing to the bone marrow and in hematopoietic stem cell quiescence. Until recently, SDF-1 and CXCR4 were believed to be a relatively monogamous ligand-receptor pair (other chemokines are promiscuous, tending

to use several different chemokine receptors). Recent evidence demonstrates ubiquitin is also a natural ligand of CXCR4 Ubiquitin is a small (76-amino acid) protein highly conserved among eukaryotic cells. It is best known for its intracellular role in targeting ubiquitylated proteins for degradation via the ubiquitin proteasome system. Evidence in numerous animal models suggests ubiquitin is anti-inflammatory immune modulator and endogenous opponent of proinflammatory damage associated molecular pattern molecules. It is speculated this interaction may be through CXCR4 mediated signalling pathways. MIF is an additional ligand of CXCR4.

CXCR4 is present in newly generated neurons during embryogenesis and adult life where it plays a role in neuronal guidance. The levels of the receptor decrease as neurons mature. CXCR4 mutant mice have aberrant neuronal distribution. This has been implicated in disorders such as epilepsy.

14-2. Clinical significance

Drugs that block the CXCR4 receptor appear to be capable of "mobilizing" hematopoietic stem cells into the bloodstream as peripheral blood stem cells. Peripheral blood stem cell mobilization is very important in hematopoietic stem cell transplantation (as a recent alternative to transplantation of surgically harvested bone marrow) and is currently performed using drugs such as G-CSF. G-CSF is a growth factor for neutrophils (a common type of white blood cells), and may act by increasing the activity of neutrophil-derived proteases such as neutrophil elastase in the

bone marrow leading to proteolytic degradation of SDF-1. Plerixafor (AMD3100) is a drug, recently approved for routine clinical use, which directly blocks the CXCR4 receptor. It is a very efficient inducer of hematopoietic stem cell mobilization in animal and human studies.

It has been associated with WHIM syndrome. WHIM like mutations in CXCR4 were recently identified in patients with Waldenstrom's macroglobulinemia, a B-cell malignancy.

While CXCR4's expression is low or absent in many healthy tissues, it was demonstrated to be expressed in over 23 types of cancer, including breast cancer, ovarian cancer, melanoma, and prostate cancer. Expression of this receptor in cancer cells has been linked to metastasis to tissues containing a high concentration of CXCL12, such as lungs, liver and bone marrow. However, in breast cancer where SDF1/CXCL12 is also expressed by the cancer cells themselves along with CXCR4, CXCL12 expression is positively correlated with disease free (metastasis free) survival. CXCL12 (over-)expressing cancers might not sense the CXCL12 gradient released from the metastasis target tissues since the receptor, CXCR4, is saturated with the ligand produced in an autocrine manner. Another explanation of this observation is provided by a study that shows the ability of CXCL12 (and CCL2) producing tumors to entrain neutrophils that inhibit seeding of tumor cells in the lung.
(출처 : https://en.wikipedia.org/wiki/CXCR4)

15. C-X-C motif chemokine 12 (CXCL12)

The stromal cell-derived factor 1 (SDF1) also known as C-X-C motif chemokine 12 (CXCL12) is a chemokine protein that in humans is encoded by the CXCL12 gene.
Stromal cell-derived factors 1-alpha and 1-beta are small cytokines that belong to the chemokine family, members of which activate leukocytes and are often induced by proinflammatory stimuli such as lipopolysaccharide, TNF, or IL1. The chemokines are characterized by the presence of 4 conserved cysteines that form 2 disulfide bonds. They can be classified into 2 subfamilies. In the CC subfamily, the cysteine residues are adjacent to each other. In the CXC subfamily, they are separated by an intervening amino acid. The SDF1 proteins belong to the latter group.
(출처 : https://en.wikipedia.org/wiki/Stromal_cell-derived_factor_1)

16. chemokine (C-X-C motif) ligand 1 (CXCL1)

The chemokine (C-X-C motif) ligand 1 (CXCL1) is a small cyto kine belonging to the CXC chemokine family that was previously called GRO1 oncogene, GROα, KC, neutrophil-activating protein 3 (NAP-3) and melanoma growth stimulating activity, alpha (MS GA-α). In humans, this protein is encoded by the CXCL1 gene.
CXCL1 is secreted by human melanoma cells, has mitogenic properties and is implicated in melanoma pathogenesis. CXCL1 is expressed by macrophages, neutrophils and epithelial cells, and ha s neutrophil chemoattractant activity. CXCL1 plays a role in spin

al cord development by inhibiting the migration of oligodendrocyte precursors and is involved in the processes of angiogenesis, arteriogenesis, inflammation, wound healing, and tumorigenesis. This chemokine elicits its effects by signaling through the chemokine receptor CXCR2. The gene for CXCL1 is located on human chromosome 4 amongst genes for other CXC chemokines. An initial study in mice showed evidence that CXCL1 decreased the severity of multiple sclerosis and may offer a neuro-protective function.
(출처 : https://en.wikipedia.org/wiki/CXCL1)

17. Hydroxymethylglutaryl-CoA synthase (HMGCS1)

In molecular biology, Hydroxymethylglutaryl-CoA synthase or HMG-CoA synthase is an enzyme which catalyzes the reaction in which Acetyl-CoA condenses with acetoacetyl-CoA to form 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA). It is the second reaction in the mevalonate-dependent isoprenoid biosynthesis pathway. HMG-CoA is an intermediate in both cholesterol synthesis and ketogenesis. This reaction is over-activated in patients with diabetes mellitus type 1 if left untreated, due to prolonged insulin deficiency and the exhaustion of substrates for gluconeogenesis and the TCA cycle, notably oxaloacetate. This results in shunting of excess acetyl-CoA into the ketone synthesis pathway via HMG-CoA, leading to the development of diabetic ketoacidosis.
(출처
: https://en.wikipedia.org/wiki/Hydroxymethylglutaryl-CoA_synthase)

Abstract

The intratumoral heterogeneity of solid tumors in vivo has been more complicated by the advent of a tumor-initiating cell, also known as cancer stem cells. Cancer stem cells have epithelial-to-mesenchymal-transition characteristics and more aggressive properties, which may cause metastasis. However, there is no confident identification of cancer stem cells-related markers. For investigating the mechanism of cancer stem cells, the comparative method between adherent cells and sphere cells is widely used because sphere cells have been known to maintain the capacity of cancer stem cells. Here, using gene set, we conducted meta-analysis that combined gene expression profiles from several studies which utilized tumorsphere technology for investigating tumor-stem like breast cancer cells. To collect gene expression profiles, we brought our own gene expression profiles from Ajou University and three different gene expression profiles from the Gene Expression Omnibus (GEO) together. Gene expression profiles of GEO were combined by ComBat method and gene set analysis was conducted using our dataset and the combined dataset. In gene set analysis, we found four gene sets including cytokine-cytokine receptor interaction and valine, leucine and isoleucine degradation, which commonly demonstrated significance. Among the genes of four significant gene sets, we selected CXCR4, CXCL1, IL12RB2, ACADM, BCKDHB and HMGCS1 that satisfied p-value < 0.05 and consistently up-regulated in both datasets. Finally, Ajou University confirmed the expression of candidates using quantitative reverse transcription-polymerase chain reaction and CXCR4, CXCL1 and HMGCS1 showed significant up-regulation in

MCF-7 derived sphere. In this study, we demonstrated a framework of meta-analysis by applying gene set analysis and d etected CXCR4, CXCL1 and HMGCS1 as candidates that involved in tumor-stem like breast cancer cells.

Keywords:
tumor-stem-like breast cancer cells, meta-analysis, gene set analysis, cytokine-cytokine receptor interaction gene set, valine, leucine and isoleucine degradation gene set

Student Number: 2012-30467