1. TCGA(The Cancer Genomic Atlas) – 암 유전체학 지도(…)

National Cancer Institute + National Human Genome Research Institute(NHGRI)

2. Biospecimens

- 암 진단 및 분석에 사용될 수 있는 인체 유래 조직 및 체액 샘플.

- 암 세포의 생물학적 특성을 식별 할 수 있는 cell, gene, protein등 매우 큰 양의 생물학적 정보를 포함.

(1) TCGA는 각 종양 유형별로 최대 500개의 샘플을 검사. 각 암에 대한 포괄적인 게놈 프로필 생성 🡪 약물 개발을 위한 최상의 목표를 식별.

(2) 모든 TCGA 연구 팀은 동일한 견본을 연구. 모든 샘플은 모든 플랫폼으로 분석되므로 암 게놈에 대한 완전하고 상세한 분석 결과를 얻을 수 있음.

(3) 연구된 모든 TCGA 암 표본에 대해 암 조직과 정상 조직의 샘플을 수집. 연구된 암의 발달에 역할을 할 수 있는 genomic change를 식별할 수 있음.

\*’종양 조직’과 이에 대응하는 ‘정상 조직’의 quality와 quantity의 이용 가능성은 TCGA에서 연구를 위해 선택된 암의 주된 요소.

3. Precision Medicine & Cancer Genomics

- 정밀 의학(Precision Medicine)은 사람의 질병에 대한 유전 정보가 질병을 진단, 치료하는 데 사용됨.

- 암은 게놈의 질병, 각 종양에는 고유한 유전적 변화가 있음. 암세포의 유전적 변화를 이해

🡪 각 환자마다 암의 유전적 특성에 맞는 효과적인 치료 전략 사용.

- 암 유전체학은 DNA 시퀀싱 및 환자 종양 분석을 통해 맞춤형 의학을 발전, 특정 암과 관련된 새로운 유전적 변이를 찾는 것을 목표함.

- major type & subtype의 암에서 key genomic change에 대한 포괄적인 카탈로그를 제공

🡪 암 진단, 치료 및 예방을 좀 더 효과적으로.

(1) imatinib (Gleevec) : 만성 골수성 백혈병에서 발견되는 두 가지 유전자가 융합해 생성된 변형 효소를 억제.

(2) trastuzumab (Herceptin) : 유방암 치료제. HER-2 양성이라는 특정 유전적 특징의 환자에게만 작용. 종양이 EGFR 돌연변이에 양성인 폐암 환자는 gefitinib (Iressa) 와 erlotinib (Tarceva) 이 돌연변이를 표적으로 삼는다. 반면에 KRAS라는 종양에 돌연변이가 있는 결장암 환자의 경우 cetuximab ( Erbitux )과 panitumumab ( Vectibix )의 효과는 거의 없음.

- TCGA 및 기타 암 게놈 프로젝트에 의해 생성된 게놈 정보 유형은 주어진 게놈 변화에 대해 가장 효과적인 치료 전략을 개발하기 위해 연구.

4. 관련 그룹 및 “GDC(Genomic Data Commons)”

[Biospecimen Core Resource (BCR)](https://cancergenome.nih.gov/abouttcga/overview/howitworks/bcr) – 다른 TCGA 프로젝트 사이의 인터페이스, 표준 규격 준수 여부 판별 등.

[게놈 특성화 센터 (Genome Characterization Centres, GCC)](https://cancergenome.nih.gov/abouttcga/overview/howitworks/characterizationcenters) - 게놈 특성화 센터는 유전자 발현 수준 및 게놈의 구조적 재 배열을 포함하여 암과 관련된 게놈 변화를 분석하기 위해 몇 가지 기술을 사용했습니다.

[게놈 시퀀싱 센터 (GSC)](https://cancergenome.nih.gov/abouttcga/overview/howitworks/sequencingcenters) - 높은 처리량의 게놈 시퀀싱 센터는 특정 유형의 암과 관련된 DNA 서열의 변화를 확인했습니다.

[Proteome Characterization Centers (PCCs)](https://cancergenome.nih.gov/abouttcga/overview/howitworks/proteomecharacterization) - NCI의 [임상 Proteomic 종양 분석 컨소시엄](http://proteomics.cancer.gov/) 의 한 구성 요소. TCGA 샘플의 전체 proteomic contents를 분석.

[데이터 조정 센터 (DCC) 및 암 유전체학 허브 (CGHub)](http://cancergenome.nih.gov/abouttcga/overview/howitworks/datasharingmanagement) - TCGA에 의해 생성된 정보는 DCC에서 중앙 관리, TCGA 데이터 포털 및 Cancer Genomics Hub에 입력.

데이터는 이제 NCI [***Genomic Data Commons****에*](https://gdc.cancer.gov/) 의해 저장되고 배포.

[GDAC (Genome Data Analysis Centres)](https://cancergenome.nih.gov/abouttcga/overview/howitworks/dataanalysiscenters) - GDAC은 수천 개의 샘플에 걸쳐 배열 및 시퀀싱 기술에서 엄청난 양의 데이터를 통합. 이 센터는 TCGA 데이터의 광범위한 사용을 촉진하기 위해 전체 연구 커뮤니티에 새로운 정보 과학 도구를 제공.

[분석 실무 그룹 (AWG)](http://cancergenome.nih.gov/abouttcga/overview/howitworks/analysis-working-groups) - AWG는 각 TCGA 종양 유형에 대한 포괄적이고 통합적인 분석을 수행하는 학제 간 국제 그룹의 과학자. 모든 AWG는 모든 TCGA 플랫폼을 사용하여 특정 종양 유형을 연구하고 암 연구 및 임상 공동체에 도움이되는 동료 검토 저널에 연구 결과 분석을 게시.

\*GDC?

- precision medicin의 홍보를 위한 data 공유 platform.

- The Cancer Genome Atlas (TCGA) & Therapeutically Applicable Research to Generate Effective Therapies (TARGET) 포함. clinical + genomic data

- common set of bioinformatics 사용 🡪 바로 비교할 수 있다!

5. 본격 유전자명 mapping

**1) 데이터 다운로드**

GDC data portal – repository – gene expression quantification & open access

\*data\_search.py 참조.

**2) expression data의 gene code를 gene name(hvnc symbol)으로 mapping**

Ensemble biomaRt(R package) 사용.

설치) <https://bioconductor.org/packages/release/bioc/html/biomaRt.html>

source("https://bioconductor.org/biocLite.R")

biocLite("biomaRt")

gene exp data 예시) counts, txt 파일.

ENSG00000000003.13\t829\n

\* 60483개 or 60488개 element 가지고 있음. 후자의 경우 5개는 이름 없는 gene.

**(1) 프록시 서버 접속 & data set preparation**

Sys.setenv("http\_proxy" = "http://www.ensembl.org:80")

혹은

options(RCurlOptions = list(proxy="http://www.ensembl.org:80",proxyuserpwd=NULL))

로 접속.

ensembl = useMart("ensembl",dataset="hsapiens\_gene\_ensembl", host = "www.ensembl.org", ensemblRedirect = FALSE)

# 인간 유전자 data set 사용.

**(2) getIBM으로 검색 – hgnc symbol**

id\_with\_hgnc <- getBM(attributes = c("ensembl\_gene\_id\_version","hgnc\_symbol"),

filters = "ensembl\_gene\_id\_version",

values = raw\_ref$ids, mart = ensembl)

# id\_with\_hgnc에 gene id와 hgnc symbol로 이뤄진 data frame을 저장.

\*총 19287개 hgnc symbol 검색.

**(3) hgnc symbol을 총 유전자 길이의 hgnc\_syms에 mapping**

hgnc\_syms <- rep("NA", length(raw\_ref[,1]))

# hgnc\_syms를 NA로 초기화. 총 유전자 길이만큼 선언.

for(n in 1:length(id\_with\_hgnc)){

if(id\_with\_hgnc$hgnc\_symbol[n]!=""){

index = which(raw\_ref$ids == id\_with\_hgnc$ensembl\_gene\_id\_version[n])

hgnc\_syms[index] = id\_with\_hgnc$hgnc\_symbol[n]

}

}

# id\_with\_hgnc 길이만큼 반복. id와 hgnc symbol이 묶여진 파일에서, 그 id에 해당하는 gene symbol이 존재한다면 raw\_ref에서 어떤 행에 위치하는지 index로 가져옴. index에 해당하는 hgnc\_syms 값에 해당 gene symbol을 넣어줌.

\*하나의 gene id가 2개 이상의 hgnc symbol을 마킹하는 경우 있다. 총 31개의 error 발생. 어떻게 처리할까?

ref <- data.frame("ids" = raw\_ref$ids, "hgnc\_syms" = hgnc\_syms)

ref\_shorts <- ref[1:60483,]

번외) gene expression data의 gene name을 GEO data와 비교.

**3) gene expression data의 tumor 여부 판별**

file name 🡪 TCGA barcode 🡪 판별 🡪 file name & Tumor/Normal 여부 묶기(txt file)

<https://www.biostars.org/p/215175/>

ref file preprocessing

filename-TCGA barcode mapping file

1. 헤더 추가

2. barcode의 가장 뒤 3자리(sample id)를 추출한 후 앞의 두 자리만 남김.

3. filename에서 “.gz” 있는 경우 제거

barcode$sample\_id[barcode$file\_name == table$file\_name]

file name으로 ref 파일에서 Tumor 여부 찾아서 TCGA table에 추가