



مقدمه ای بر بیوانفورماتیک

دانشگاه صنعتی شریف

پاییز 1401

اساتید: دکتر کوهی – دکتر شریفی

اعضای گروه:

ارسلان مسعودی فرد-99105718

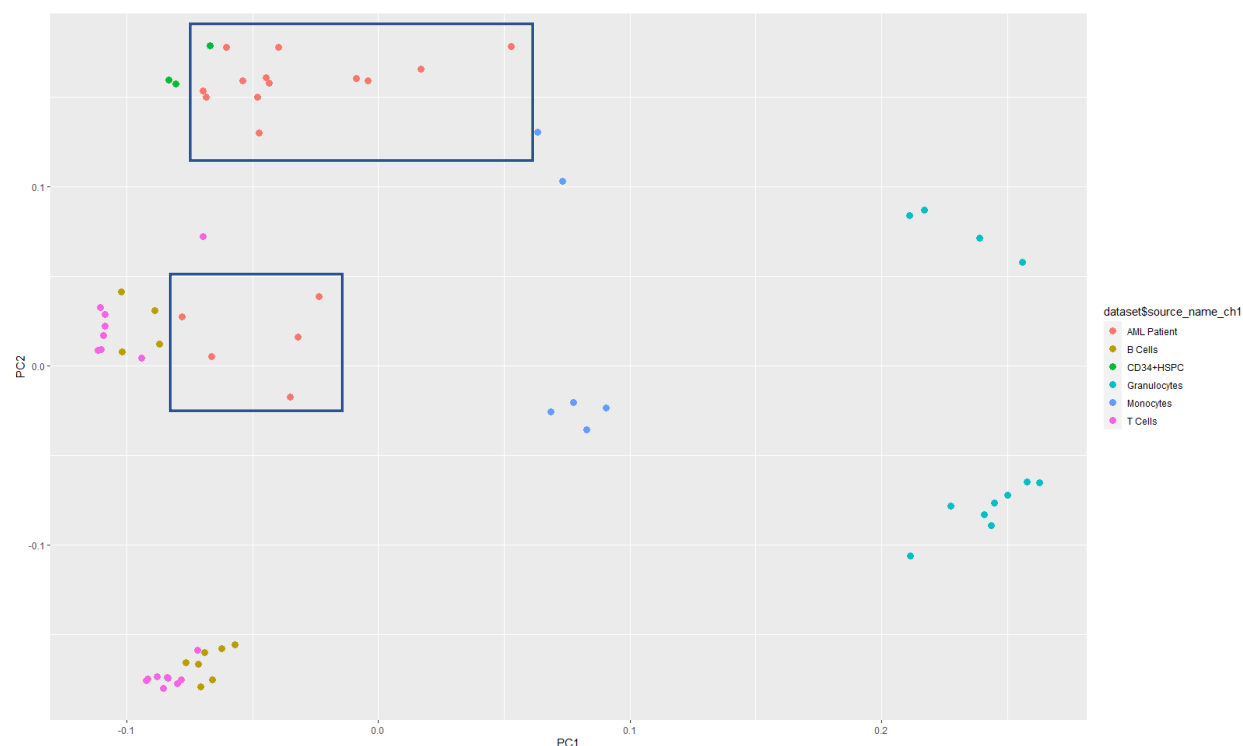
امیرحسین محمودی-98108779

آرش ملکپور-98108821

بخش اول

با توجه به نمودار PCA که در فاز قبل بررسی شد دو دسته بودن AML ها مشهود است و با قطعه کد زیر آنها را از یکدیگر جدا می کنیم. بدین صورت که نمونه های Test یا AML که شرط $PC2 > 0.1$ را دارا باشند به نمونه CD34+HSPC نزدیک ترند، پس جدا می شوند.

```
grouped_near_AML <- divided_group
grouped_near_AML[which((pcar$PC2 > 0.1 & pcar$group == "Test"))] <- "AML_CD34"
```



(*)اصلاحیه: در فاز یک هر دو نمودار PCA یک شکل بودند. در صورتی که قطعه کد دوم(86-91) شکل بالا را می دهد.)

حال با استفاده از تابع فاکتور فرمت گروه ها را تغییر داده و مشاهده می شود که باز همان 6 گروه را داریم ولی AML های نزدیک به CD34 باقی مانده اند. سپس متغیر design را تعریف می کنیم که یک ماتریس است و برای هر گروه یک سطر دارد. سپس با تابع lmFit به دیتا مدل خطی داده و در contrast به مقایسه نمونه های AML نزدیک با CD34 می پردازیم.

```
grouped_near_AML <- factor(grouped_near_AML)
dataset$description <- grouped_near_AML
design <- model.matrix(~ description + 0, dataset)
colnames(design) <- levels(grouped_near_AML)

fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(AML_CD34 - Normal_CD34, levels = design)
```

در ادامه متغیر fit2 را تعریف می‌کنیم که شیب بر اساس تفاوت AML_CD34 و Normal_CD34 تعیین شده است و با یک مدل Bayesian آن را fit می‌کنیم.

```
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
```

پس از آنکه posterior توزیع نهایی بدست آمد در یک جدول موارد مورد نیاز مثل P_value و logFC را نگاه می‌داریم.

```
table <- topTable(fit2, adjust = "fdr", sort.by = "B", number = Inf)
table1 <- subset(table, select = c("ID", "Gene.symbol", "Gene.title", "P.value", "adj.P.val", "logFC"))
write.table(table1, file = "table1.txt", sep = "\t")
```

از اینجا به بعد کافیسست تاثیرگذارترین ژن‌هایی که p-value آنها از 0.05 کمتر بوده را بدست آورده و سپس آنهایی که logFC آنها بیشتر از 1 است را به عنوان جدول ژن‌هایی با بیان بیشتر و کمتر از 1 را به عنوان بیان کمتر خروجی دهیم.

```
aml.high <- subset(table1, logFC > 1 & adj.P.val < 0.05)
aml.high.genes <- unique(as.character(strsplit2((aml.high$Gene.symbol), "///")))
write.table(aml.high.genes, "aml_cd34_higher.txt", quote=FALSE,
            row.names = FALSE, col.names = FALSE)

aml.low <- subset(table1, logFC < -1 & adj.P.val < 0.05)
aml.low.genes <- unique(as.character(strsplit2((aml.low$Gene.symbol), "///")))
write.table(aml.low.genes, "aml_cd34_lower.txt", quote=FALSE,
            row.names = FALSE, col.names = FALSE)
```

بخش دوم

* بررسی pathwath برای ژن‌های با بیان بیشتر

در این بخش با کمک جدولی (فایل تکست) که در بخش قبلی به دست آوردیم یک سری از ژن‌ها را وارد سایت Enrichr می‌کنیم. فقط باید دقت کرد که برخی ژن‌ها چند اسم دارند و باید در سطرهای جداگانه‌ای قرار بگیرند. که برای حل این مشکل نیز عبارت زیر را به قطعه کد صفحه قبل اضافه کردیم.

```
as.character(strsplit2((am1.low$Gene.symbol), "///"))
```




حال تعدادی یا همه‌ی اسامی جدول تهیه شده را وارد سایت می‌کنیم و در بخش pathway که برای تعیین مسیر زیستی به کار می‌رود از دو مجموعه‌ی Reactome 2022 و WikiPathways 2021 Human استفاده می‌کنیم. چهار مورد زیر adjuster p-value های مناسبی داشتند.

Reactome 2022					
Bar Graph Table Clustergram Appyter ⚙️ ⓘ					
Hover each row to see the overlapping genes.					
10 ▾	entries per page		Search: <input type="text"/>		
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Immune System R-HSA-168256	3.000e-48	3.879e-45	3.56	389.32
2	Innate Immune System R-HSA-168249	4.292e-37	2.775e-34	3.96	331.79




WikiPathway 2021 Human					
Bar Graph Table Clustergram Appyter ⚙️ ⓘ					
Hover each row to see the overlapping genes.					
10 ▾	entries per page		Search: <input type="text"/>		
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	TYROBP causal network in microglia WP3945	1.799e-11	9.011e-9	9.47	234.32
2	Complement system WP2806	6.960e-11	1.744e-8	6.52	152.60

* بررسی pathwath برای ژن‌های با بیان کمتر

مراحل ذکر شده در قسمت قبل را برای جدول ژن‌ها با بیان کمتر هم پیاده‌سازی می‌کنیم. اما در دو مجموعه ذکر شده در قسمت قبل adjusted p-value های بسیار بالاتری داریم (نزدیک یا برابر 1). پس به سراغ مجموعه ARCHS4 Kinases Coexp می‌رویم.

ARCHS4 Kinases Coexp					
Bar Graph Table Clustergram Appyter  					
Hover each row to see the overlapping genes.					
10  entries per page		Search: <input type="text"/>			
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	PRKG2 human kinase ARCHS4 coexpression	2.639e-10	1.269e-7	3.45	76.14
2	PAK5 human kinase ARCHS4 coexpression	0.00001701	0.004092	2.48	27.26

با اینکه AVP نمونه بالا خوب بود اما اطلاعات کمتری در رابطه با آن وجود دارد. پس مجموعه زیر هم در نظر می‌گیریم.

HumanCyc 2016					
Bar Graph Table Clustergram Appyter  					
Hover each row to see the overlapping genes.					
10  entries per page		Search: <input type="text"/>			
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	cholesterol biosynthesis III (via desmosterol) Homo sapiens PWY66-4	0.0001809	0.002081	13.21	113.83

* بررسی ontology برای ژن‌های با بیان بیشتر

این حوزه را در دو مجموعه‌ی biological process و molecular function مورد بررسی قرار می‌دهیم.

GO Biological Process 2021 Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

Hover each row to see the overlapping genes.

10 ▾ entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	regulation of microglial cell mediated cytotoxicity (GO:1904149)	2.128e-7	0.00002590	95360.00	1464985.63
2	neutrophil activation involved in immune response (GO:0002283)	1.231e-37	4.844e-34	5.86	498.20
3	neutrophil degranulation (GO:0043312)	3.422e-37	6.734e-34	5.84	490.60

با اینکه مورد اول AVP بیشتری نسبت به بقیه دارد اما odds ratio بسیار بالای آن که از fisher test حاصل شده آن را قابل توجه می‌کند.

GO Molecular Function 2021 Bar Graph **Table** Clustergram Appyter ⚙️ ⓘ

Hover each row to see the overlapping genes.

10 ▾ entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	hemoglobin alpha binding (GO:0031721)	0.00002218	0.001935	82.56	884.72
2	aryl sulfotransferase activity (GO:0004062)	0.00002293	0.001935	25.82	275.87

* بررسی ontology برای ژن‌های با بیان کمتر

بار دیگر در دو مجموعه بالا AVP مناسبی نداریم پس سراغ مجموعه‌های دیگر می‌رویم.

GO Cellular Component 2021

Bar Graph

Table

Clustergram

Appyter



Hover each row to see the overlapping genes.

10

entries per page

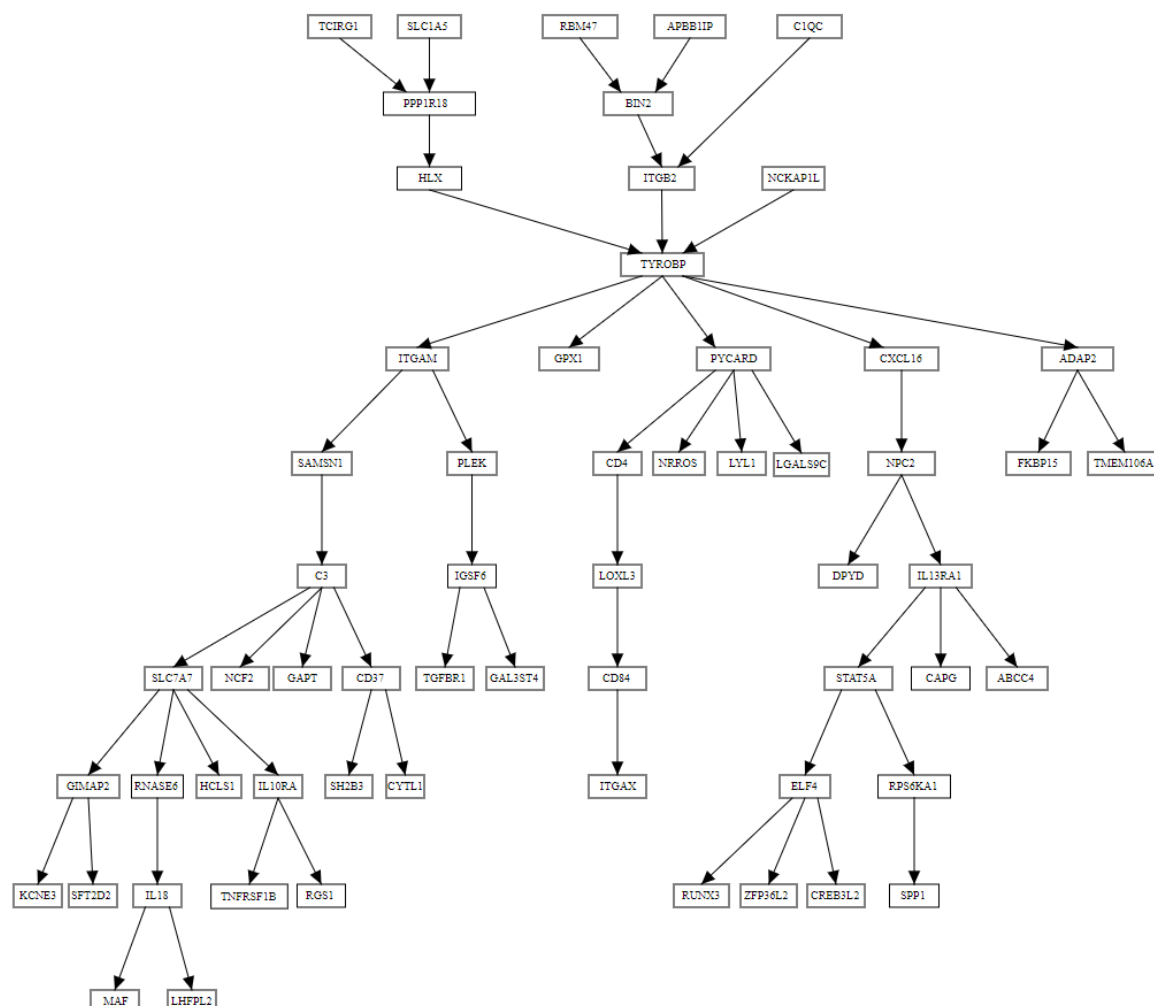
Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	integral component of luminal side of endoplasmic reticulum membrane (GO:0071556)	0.0001989	0.02834	7.06	60.13
2	luminal side of endoplasmic reticulum membrane (GO:0098553)	0.0001989	0.02834	7.06	60.13

بخش سوم

(الف)

مورد اول: TYROBP causal network in microglia WP394mn. در مجموعه‌ی WikiPathway 2021Human که در سایت مربوط به خود این مجموعه pathway آن آورده شده و در رابطه با بیماری AML و این مقالات زیادی وجود دارد که به نام برخی از آنها اشاره کرده‌ایم.



<https://www.wikipathways.org/index.php/Pathway:WP3945>

ncbi.nlm.nih.gov/pmc/articles/PMC6606333/

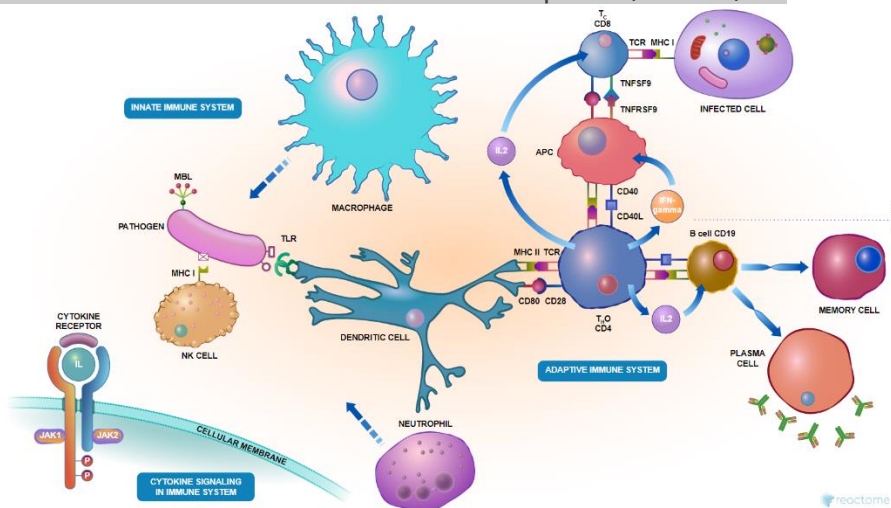
proteinatlas.org/ENSG00000011600-TYROBP/pathology

pubmed.ncbi.nlm.nih.gov/25928846/

که همگی ژن TYROBP را عامل موثری بر AML می‌دانند. چراکه این ژن مسئول تولید پروتئین TYRO protein tyrosine kinase binding بوده که در تشکیل سلول‌های ایمنی بدن نقش دارد و از آنجایی که در AML سلول‌های سفید خونی فعالیت نرمالی ندارند این ارتباط توجیه می‌شود.

مورد دوم: در مجموعه‌ی Reactome دو نمونه دیگر با adjuster p-value های بسیار پایین دیدیم که مشخصاً به سیستم ایمنی مربوط می‌شوند. دو جدول و عکس زیر مربوط به pathway های خود سایت این منبع بوده و در ادامه به مقاله‌هایی که به این ارتباط پرداخته‌اند اشاره شده.

PathwayBlob	R-HSA-168256
Source	Reactome
Taxonomic Scope	organism_specific
Taxonomy	Homo sapiens (human)



PathwayBlob

R-HSA-168249

Source

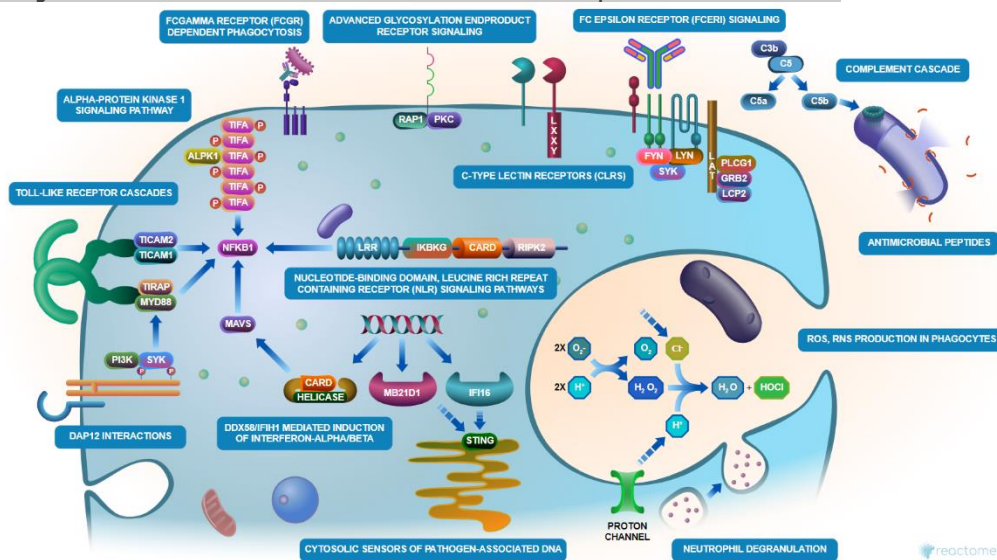
Reactome

Taxonomic Scope

organism_specific

Taxonomy

Homo sapiens (human)



ncbi.nlm.nih.gov/pmc/articles/PMC8461066/

در اینجا با مسیری مشابه با کاری که در این تحقیق کردیم بر روی دیتاستی دیگر از GSE13591 و GSE13591 در نهایت نتیجه‌ای مشابه گرفته شده.

ncbi.nlm.nih.gov/pmc/articles/PMC7201587

در مورد بالا هم با اینکه با کار انجام شده توسط این تحقیق متفاوت است اما در بخش 3.3 آن نتیجه‌گیری مشابهی شده است.