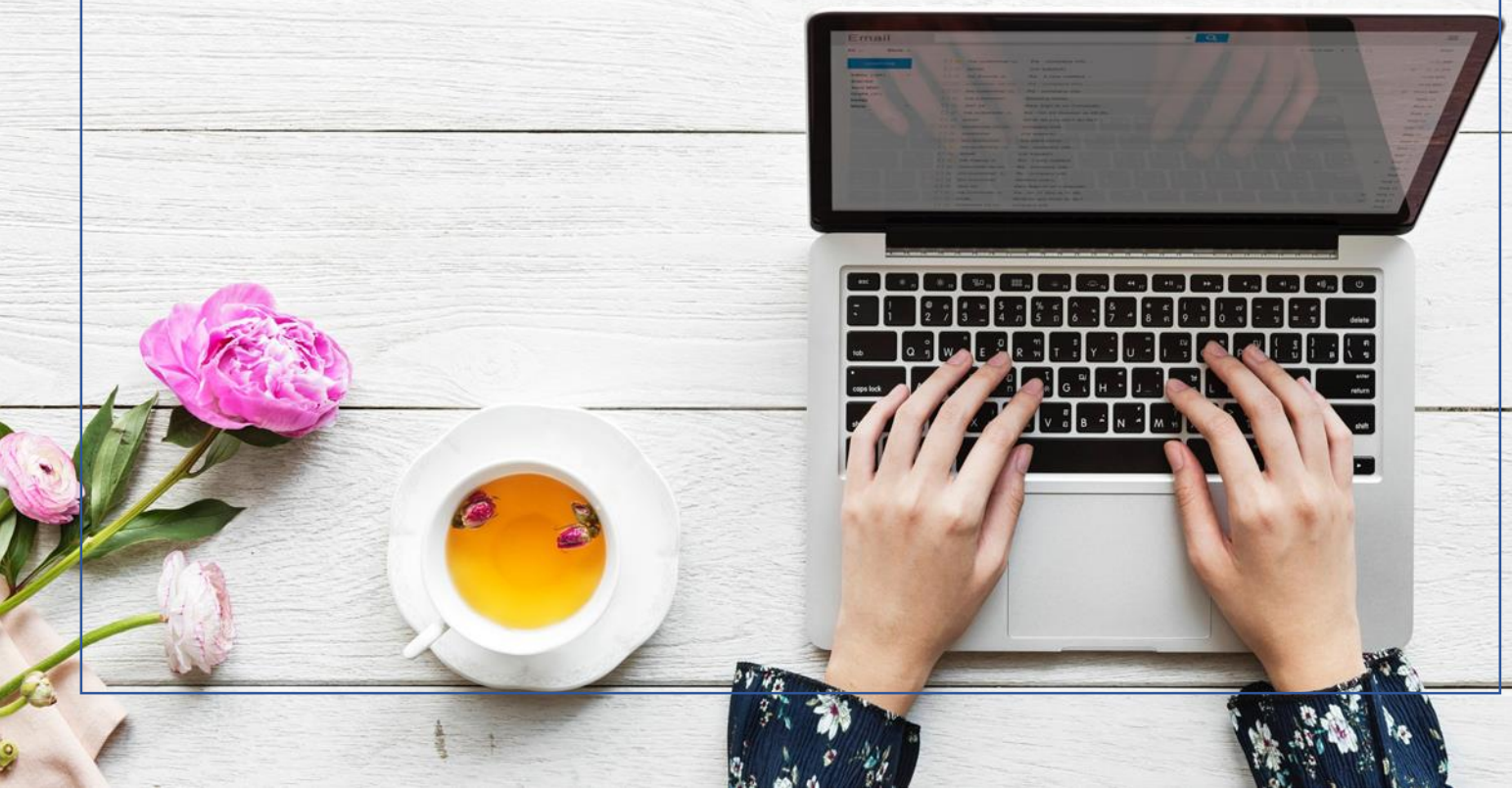
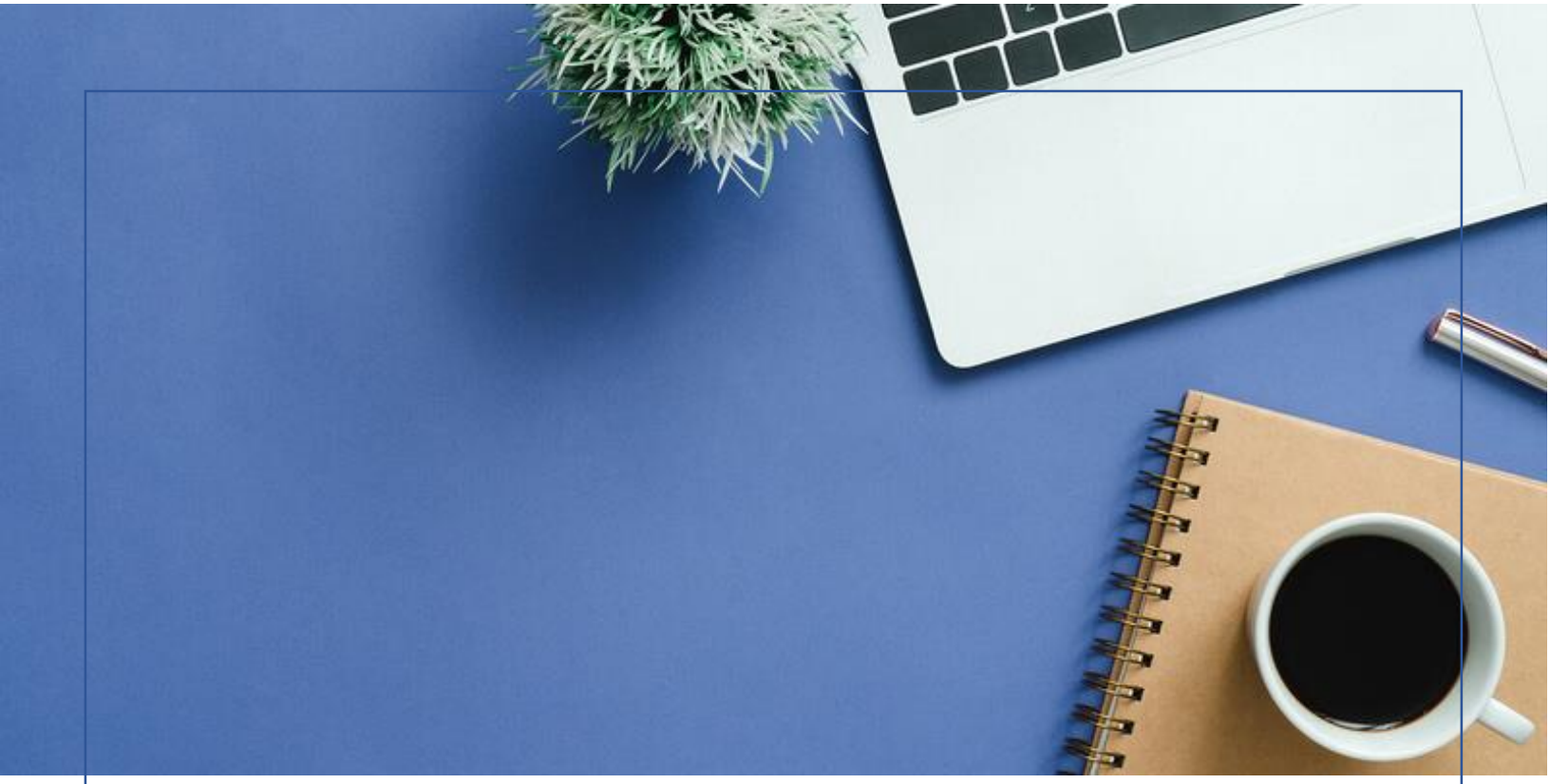
پروژه بیوانفورماتیک

محمدرضا غمخوار

95106494



فهرست



مقدمه

لوکمیا یا سرطان خونِ، یکی از انواع سرطان است ک معمولا از مغز استخوان آغاز می شود و تعداد زیادی سلول خونی غیرعادی و نابالغ تولید می کند .این سلول های خونی به طور کامل تکامل نیافته اند به آن ها بلاست(blast) و یا سلول لوکمی گفته می شود .علت بروز این سرطان هنوز ناشناخته است اما ترکیبی از عوامل ژنتیکی و محیطی(غیر ارثی) به عنوان عوامل موثر درنظر گرفت می شوند.

لوکمیا دارای انواع مختلفی است ک یکی از آن ها لوکمی حاد مغز استخوان یا به اختصار AML است. AML دومین سرطان خونِ شایع در کودکان است .معمولا در اثر بروز جهشی در دی ان ای سلول های پیش ساز خون که از تمایز کامل این سلول ها جلوگیری می کندو جهشی دیگر که موجب تقسیم و تکثیر غیرقابل کنترل سلول ها می شود، رخ می دهد.

در سال های اخیر از آنالیز داده های میکرواری برای تشخیص این بیماری و بیماری های مشابه که در اثر بروز جهش و تغییر در بیان ژن ها به وجود می آیند استفاده می شود.

در این پروژه قصد داریم با تحلیل داده های GSE48558 [1] که شامل تعدادی نمونه ی سرطانی و تعدادی نمونه ی سالم است، ژن هایی که در بروز AML موثرند را شناسایی کنیم.

به طور کلی برای نایل آمدن به این هدف مراحل زیر را طی می کنیم:

1. کنترل کیفیت داده: در ابتدا مطمئن می شویم که داده ی ما کیفیت کافی را برای ادامه تحلیل داشته باشد؛ زیرا که ممکن است در مراحل نمونه گیری خطاهایی رخ داده باشد و مارا به سمت نتیجه گیری های اشتباه سوق دهد.

2. کاهش ابعاد داده: داده ما از probe 5494570 به ازای هریک از 170 نمونه تشکیل شده است و در نتیجه از تعداد زیادی نقطه در فضایی با تعداد زیادی بعد تشکیل شده است؛ در نتیجه برای سهولت کار نیاز داریم که ابعاد داده ها را به طور موثری کاهش دهیم.
3. بررسی همبستگی بین نمونه ها: سپس بعد از کاهش ابعاد داده ها موقعیت هرکدام از 170 نمونه را در فضای جدید بدست می آوریم و با توجه به گروه نمونه(بیمار، سالم یا لوکمیای نوع دیگر) موقعیت آن ها را نسبت به هم می سنجیم؛ انتظار می رود که نمونه های هر گروه به یکدیگر نزدیک باشند.
4. بررسی تمایز در بیان ژن های نمونه ها: بعد از اطمینان از کیفیت داده ها ژن ها را از منظر تفاوت میزان بیان بین گروه سالم و بیمار بررسی می کنیم و تاثیر گذار ترین ژن ها را شناسایی می کنیم.
5. آنالیز gene anthology و pathway ها: در نهایت با داشتن متفاوت ترین ژن ها سعی می کنیم عامل های مشترک بین این ژن ها و دلایل متفاوت شدن این ژن ها را شناسایی کنیم و از این طریق می توانیم امیدوار باشیم که تشخیص AML ممکن و راحت تر شده و همچنین قدمی در راه شناخت بیشتر این بیماری و عوامل ایجاد کننده آن و درمان آن برداشته باشیم.
6. مقایسه نتایج بدست آمده با سایر مقالات زیستی: در نهایت نتایج بدست آمده در مرحله قبل را با مقالات به روز این حوزه مقایسه می کنیم.
7. بررسی تفاوت زیر گروه های داده: ما همه افراد سالم را یک گروه و همه افراد دارای سرطان به جز نوع AML را نیز در یک گروه بررسی کردیم. در این مرحله تفاوت داده ها را جزیی تر بررسی می کنیم.
8. مباحثات آینده: در این بخش سعی می کنیم چالش های پیش رو را شناخته و راهکار های احتمالی را معرفی کرده و جهت حرکت این شاخه از علم را پیش بینی کنیم.

در این گزارش سعی شده به جزئیات پیاده سازی پرداخته نشود و بیشتر نتایج و توضیحات ارائه شود. برای بررسی دقیق تر می توانید به کد پروژه مراجعه کنید.

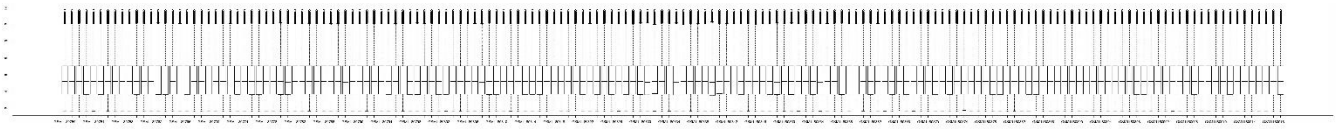
راه اندازی اولیه:

از خط 1 تا 28 کد کتابخانه های مورد نیاز را بارگذاری کرده و دیتاست مدنظر را دانلود می کنیم و آن ها را به سه دسته افراد سالم (normal) ، بیمار (AML) و سایر (leukemia) تقسیم می کنیم.

در انتها گزارش این دسته بندی جزیی تر نیز می شود.

کنترل کیفیت داده:

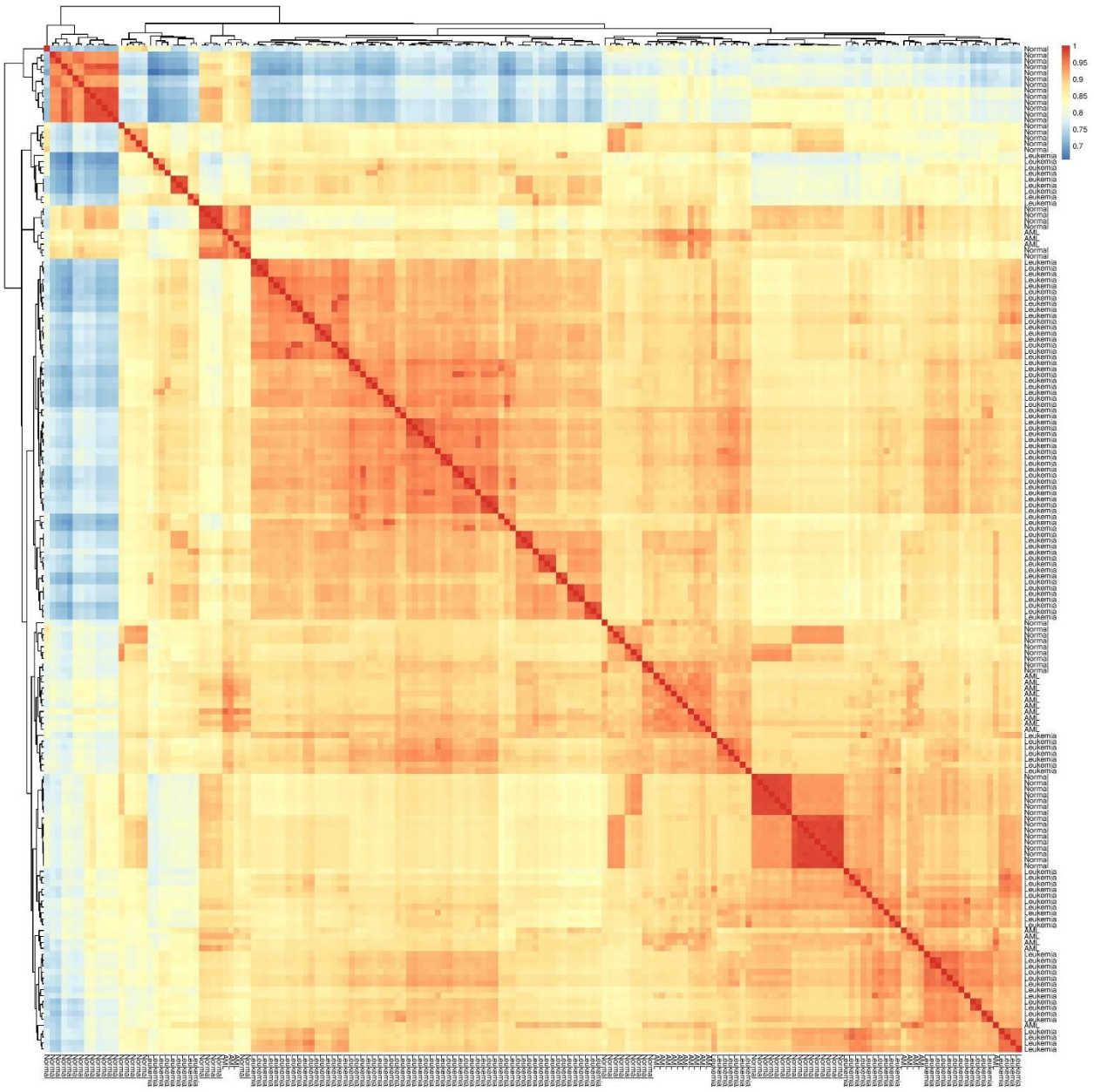
حال میزان بیان هر ژن را بدست می آوریم و نتیجه را روی نمودار می بریم. نمودار زیر حاصل می شود:



نمودار 1. توزیع داده ها – فایل boxplot.pdf

همانطور که مشاهده می کنید مقدار ماکسیمم و مینیمم برای هر ژن به ترتیب کمتر از 14 و بیشتر از 2 است در نتیجه مقیاس ما لگاریتمی است و نیازی به log2 گرفتن از داده ها نداریم. همچنین چارک های ژن های مختلف نزدیک به هم هستند پس داده ها از قبل normalize شدن و نیازی به نرمال کردن داده ها نداریم.

تا اینجا شکل کلی داده ها را برای ادامه کار مناسب کردیم حال باید ببینیم که آیا محتوای داده ها درست هستند و در نتیجه تفاوت بین دسته های مختلف و شباهت بین دسته های یکسان را نشان می دهند یا این که دچار خطا و اشتباه زیادی هستند. برای این کار در قدم اول یک heatmap از correlation بین ژن های مختلف هر گروه می کشیم:



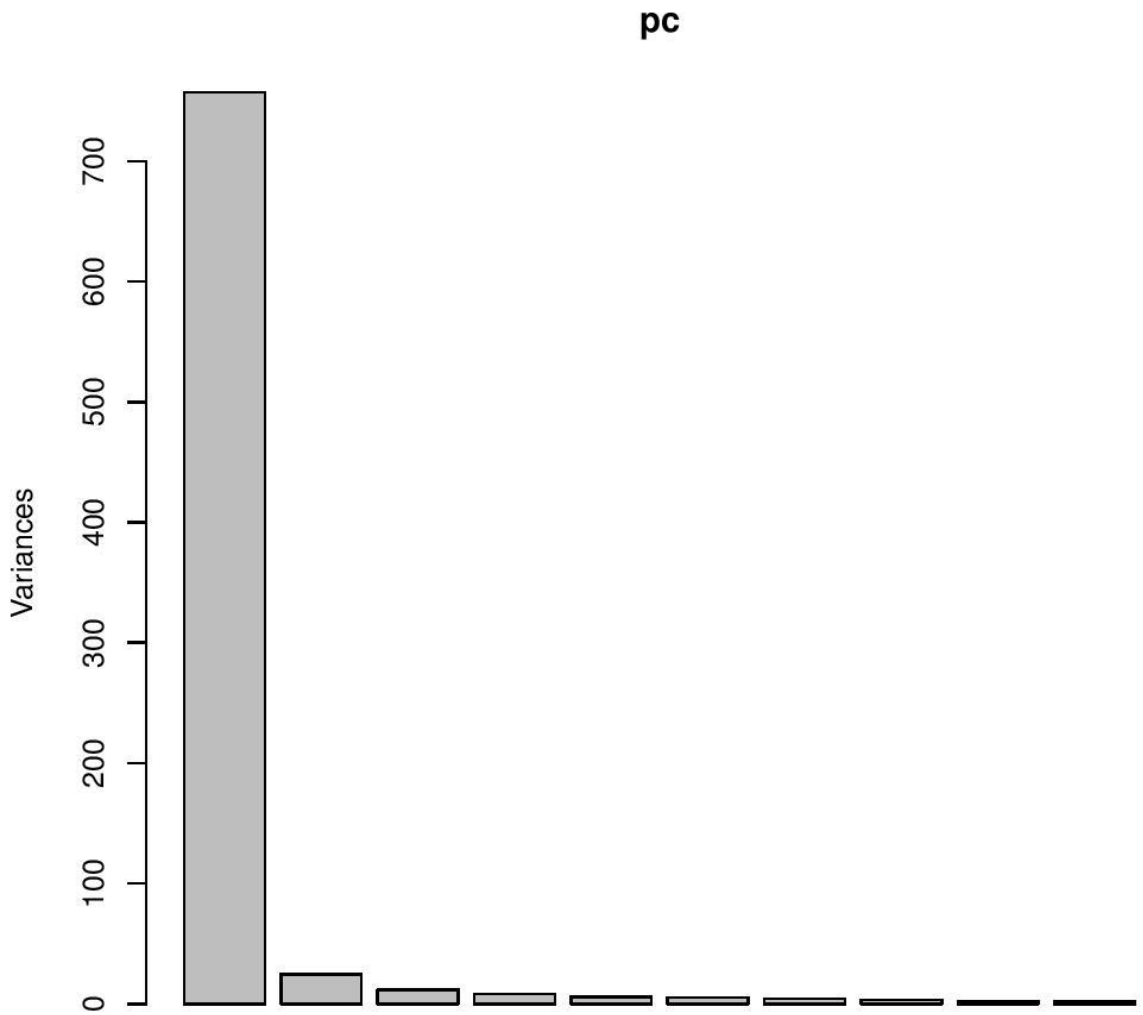
نمودار 2 شباهت داده ها – فایل corHeatmap.pdf

همانطور که مشاهده می کنید گروه های یکسان به یکدیگر شباهت دارند و از گروه های دیگر متفاوت اند؛ این بیانگر کیفیت نسبتا خوب داده هاست. در نتیجه می توانیم به تحلیلمان ادامه دهیم.

در قدم بعدی کیفیت را به صورت دقیق تر و با استفاده از روش PCA یا principal component analysis می سنجیم(در بخش کاهش ابعاد داده).

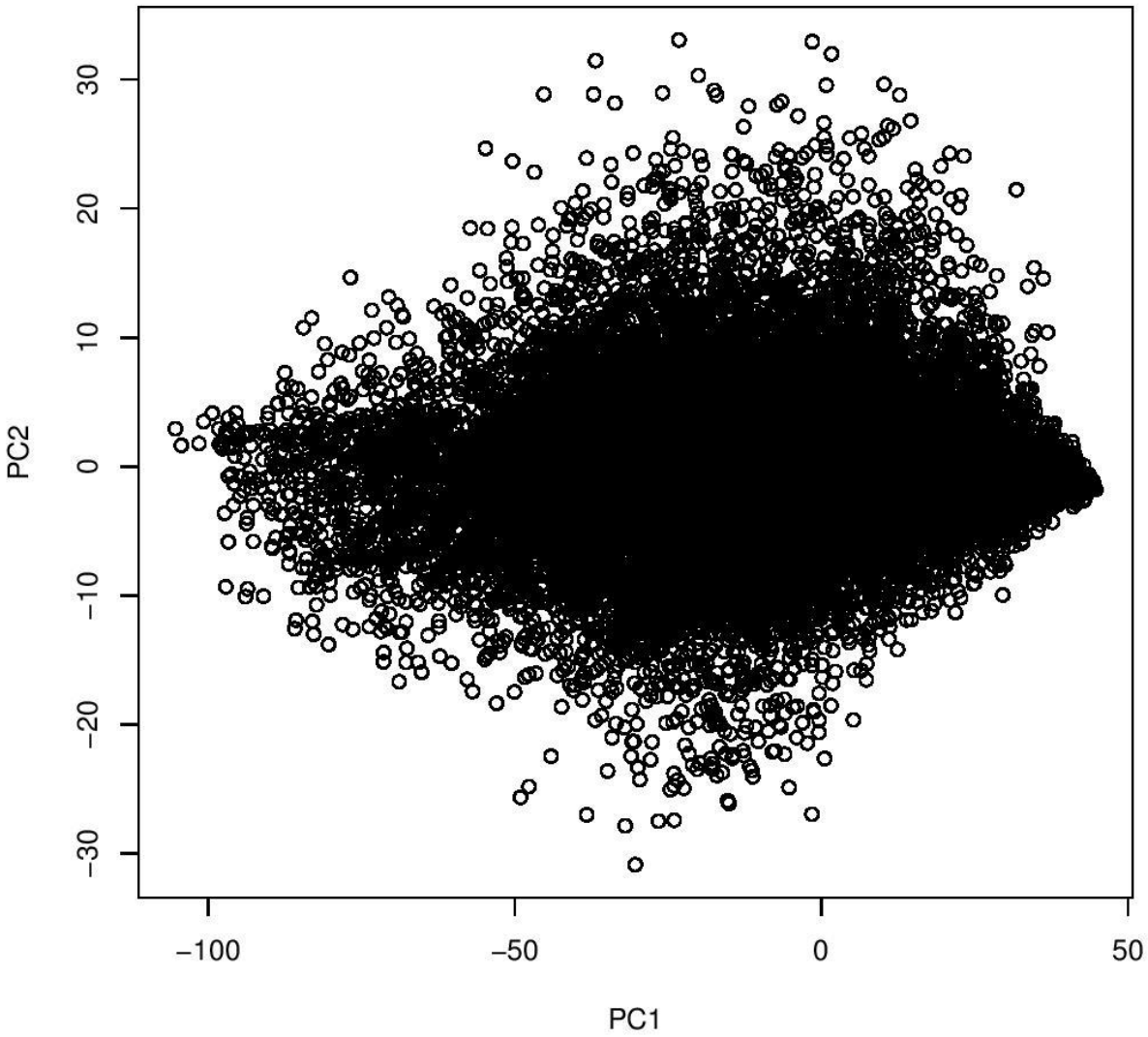
کاهش ابعاد داده:

در این مرحله اولین راه کاری که به ذهنمان می رسد بدست آوردن principal component یا pc بر اساس همین داده ها و به همین شکل است(با استفاده از تابع ()prcomp). نتیجه به این شکل می شود:



نمودار 3 میزان تفاوت داده ها براساس pc ها - فایل PC.pdf

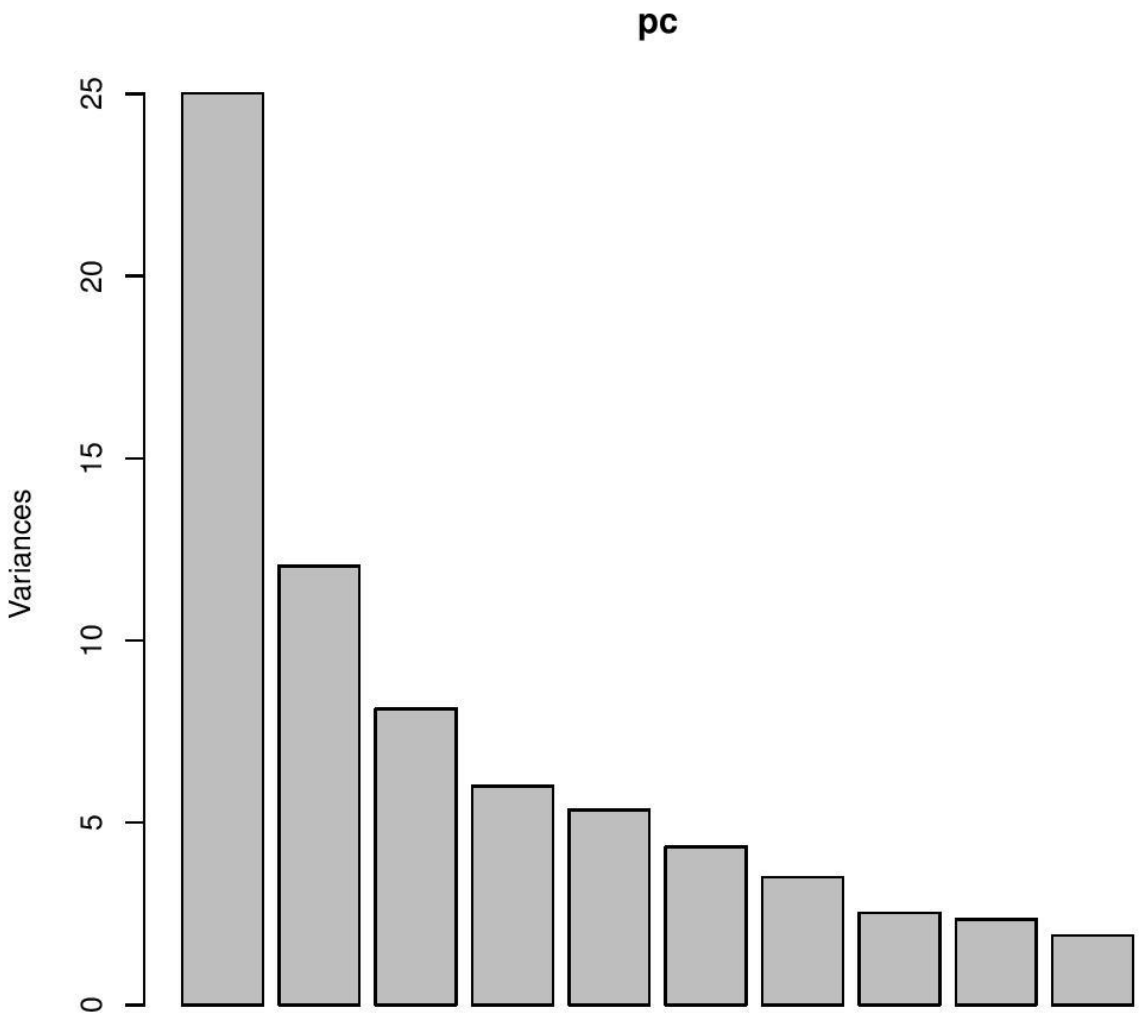
حال می توانیم از pc2، pc3 و ... چشم پوشی کنیم و ابعاد داده را فقط به pc1 و pc2 که تاثیر گذار ترین (متفاوت کننده ترین) محور هستند کاهش دهیم. نمایش داده ها در این دو بعد این چنین می شود:



نمودار 4 نمایش داده ها در فضای کاهش ابعاد داده شده اولیه - فایل PC.pdf

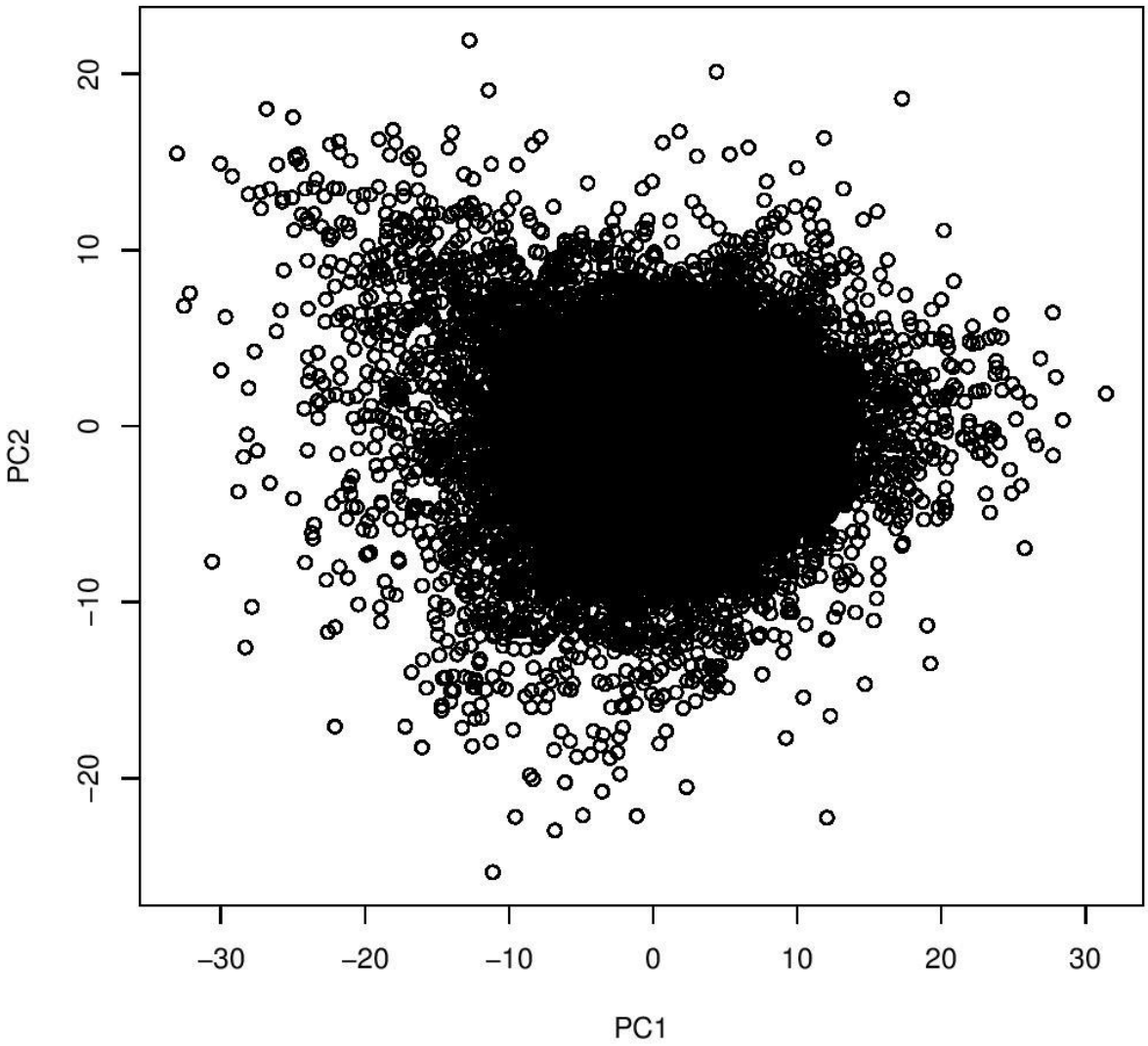
اما با کمی دقت متوجه می شویم که pc1 که به مراتب تاثیر بیشتری از بقیه pc ها دارد بیانگر میزان بیان ژن هاست که یک معیار خوب نیست زیرا که ممکن است ژن هایی در بخش های مختلف زیادی بیان شوند و از طرفی ژن هایی دیگر بسیار کم در بخش های مختلف بیان شوند و این پدیده ارتباطی به سالم یا بیمار بودن یک فرد ندارد و بدرد ما نمی خورد.

پس حال داده ها را به این صورت تغییر می دهیم که به ازای هر نمونه، میانگین بیان یک ژن در بین همه نمونه ها را از میزان بیانش در آن ژن خاص کم می کنیم و جایگزین مقدار قبلی می کنیم. حال مراحل بالا را تکرار می کنیم:



نمودار 5 میزان تفاوت داده های جرید به ازای pc ها - فایل PC_scaled.pdf

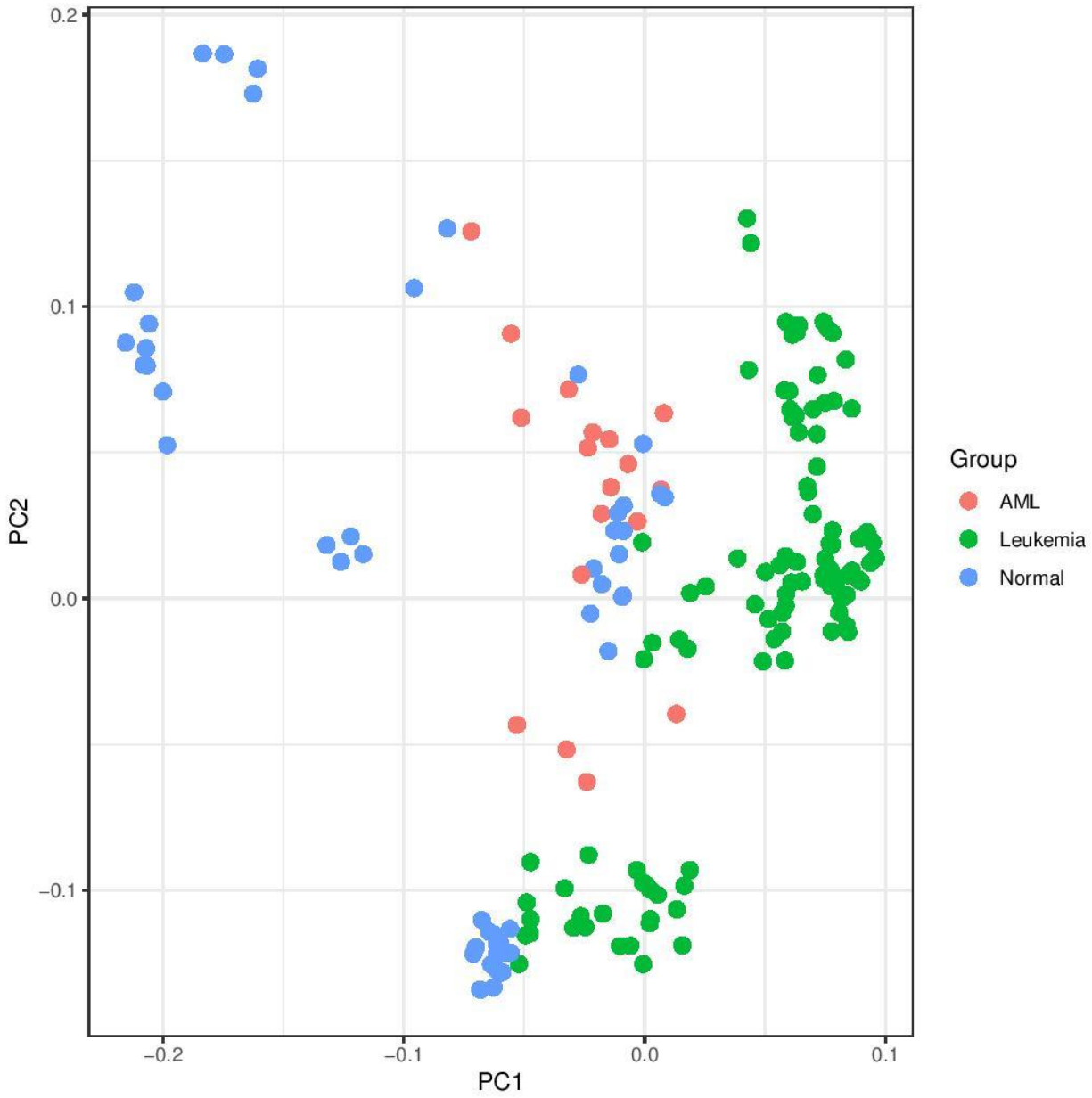
حال دیگر تفاوت pc1 با بقیه pc ها بسیار زیاد نیست. نمایش داده ها در دو بعد pc1 و pc2:



نمودار 6 نمایش داده های جدید در فضای کاهش ابعاد داده شده - فایل PC_scaled.pdf

بررسی همبستگی بین نمونه ها:

حال که ابعاد داده ها را کاهش دادیم در این مرحله سعی می کنیم تفاوت داده های هر گروه با گروه دیگر و نزدیک بودنشان به دیگر اعضای گروه خود را در فضای دو بعدی pc1 و pc2 به تصویر بکشیم یا به عبارتی دیگر همبستگی نمونه های گروه های مختلف را بررسی کنیم. نتیجه به این شکل می شود:



نمودار 7 نمایش نمونه ها در فضای pc1 و pc2 براساس گروهشان - فایل PCA_samples.pdf

همانطور که مشاهده می کنید نمونه ها تقریبا براساس گروهشان از هم جدا شده اند از طرفی می توان حدس زد که هر گروه از زیر گروه های متفاوتی تشکیل شده است که آن ها را در بخش "بررسی تفاوت زیر گروه های داده" بررسی می کنیم.

بررسی تمایز در بیان ژن های نمونه ها:

حال که از کیفیت مناسب داده ها مطمئن شدیم سراغ differential expression analysis یا بررسی تمایز بیان ژن می رویم. در ابتدا یک مدل خطی برروی ماتریس داده ها و گروه هایشان fit می کنیم و با استفاده از این مدل و تفاوت های ایجاد شده بین گروه های بیمار (AML) و سالم (Normal) با استفاده از یک مدل بیزین (eBayes()) ژن ها را به همراه p-value و logFC و دیگر اطلاعاتشان بدست می آوریم. در نهایت داده ها را براساس نمره "B" آن ها که بدست آمده از مقدار Adjusted p-value و logFC است و با استفاده از روش بنجامین هاچبرگ (fdr) مرتب می کنیم. نتایج در فایل AML_Normal.txt قابل مشاهده است:

Gene.symbol	Gene.ID	adj.P.Val	logFC
KIAA0101	9768	3.65E-31	4.559135
DTL	51514	5.92E-30	3.679218
TYMS	7298	2.83E-29	3.670352
MAMDC2	256691	7.45E-27	4.03101
MYB///MYB	4602///4602	1.50E-26	3.598965
CBX7	23492	1.58E-26	-2.24008
PRC1	9055	7.03E-26	3.080097
TPX2	22974	3.06E-22	3.156415
ZBP1	81030	5.42E-22	-2.24401
SCCPDH	51097	1.05E-20	2.833054
TOP2A	7153	1.05E-20	3.298104
CHEK1	1111	1.35E-20	2.946827
MKI67	4288	1.68E-19	2.77093
STMN1	3925	2.10E-19	2.788174
BUB1B	701	2.33E-19	2.756554

جدول 1 چند ژن اول تمایز ایجاد کننده بین افراد سالم و سرطانی - فایل AML_Normal.txt

این فایل همه ژن ها چه آن ها که به طور معنی دار بین گروه سالم و بیمار تفاوت دارند و چه آن هایی که به یکدیگر شبیه اند را نشان می دهد. حال می خواهیم تنها آن هایی که تفاوت معنی دار دارند را جدا کنیم برای این کار آن ژن هایی که adjusted p-value شان کمتر از 0.05 باشد را انتخاب می کنیم. به جز این آن هایی را که در نمونه سرطانی نسبت به سالم بیشتر بیان شده اند را از آن هایی که در نمونه سالم کمتر بیان شده اند را جدا می کنیم(به وسیله logFC). نتایج را در فایل های AML_Normal_Down و AML_Normal_Up ذخیره کرده ایم:

AML_Normal_Down	AML_Normal_Up
CBX7	KIAA0101
ZBP1	DTL
NUAK2	TYMS

AKTIP	MAMDC2
TTC9	MYB
LINC00324	PRC1
ZNF211	TPX2
SNTB2	SCCPDH
MINK1	TOP2A
DNAJB9	CHEK1

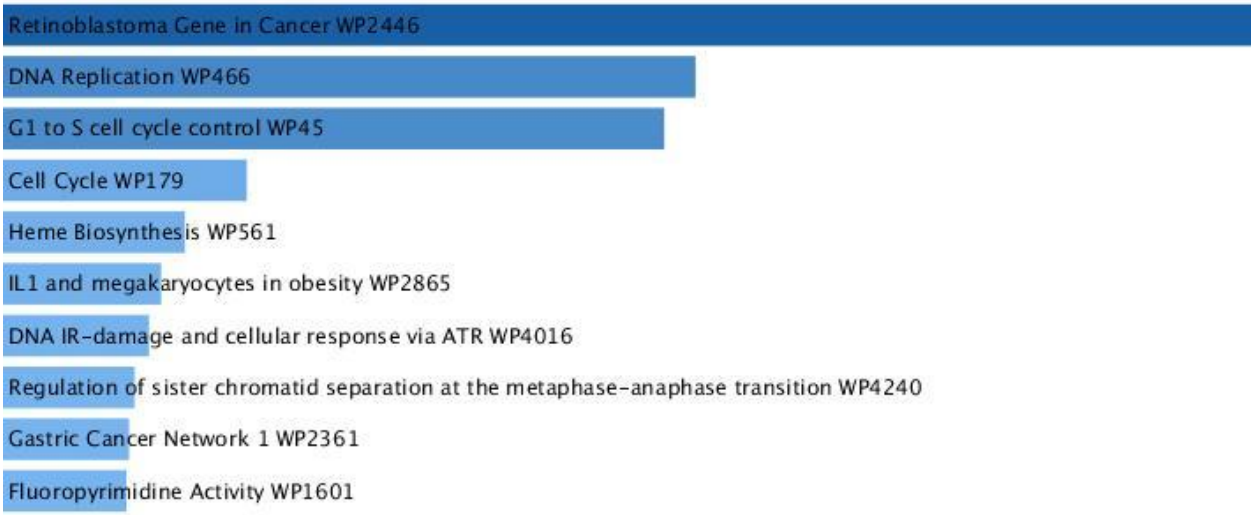
جدول 2 چند ژن اول AML_Normal_Up و AML_Normal_Down

آنالیز gene anthology و pathway ها:

در مرحله نهایی براساس ژن هایی که در قسمت قبل شناسایی کردیم pathway های مرتبط با این ژن ها را شناسایی می کنیم و در نهایت سراغ gene antology می رویم. برای این عمل از پلتفورم Enrichr [2]استفاده می کنیم ولی از آن جایی که این سایت فعلا دچار مشکل است به جای آن از OxEnrichr [3] استفاده می کنیم.

آنالیز pathway ها:

در ابتدا با استفاده از ژن هایی که در نمونه های بیمار بروز بیشتری داشتند (AML_Normal_Up.txt) و با استفاده از سایت enrichr این کار را انجام می دهیم:



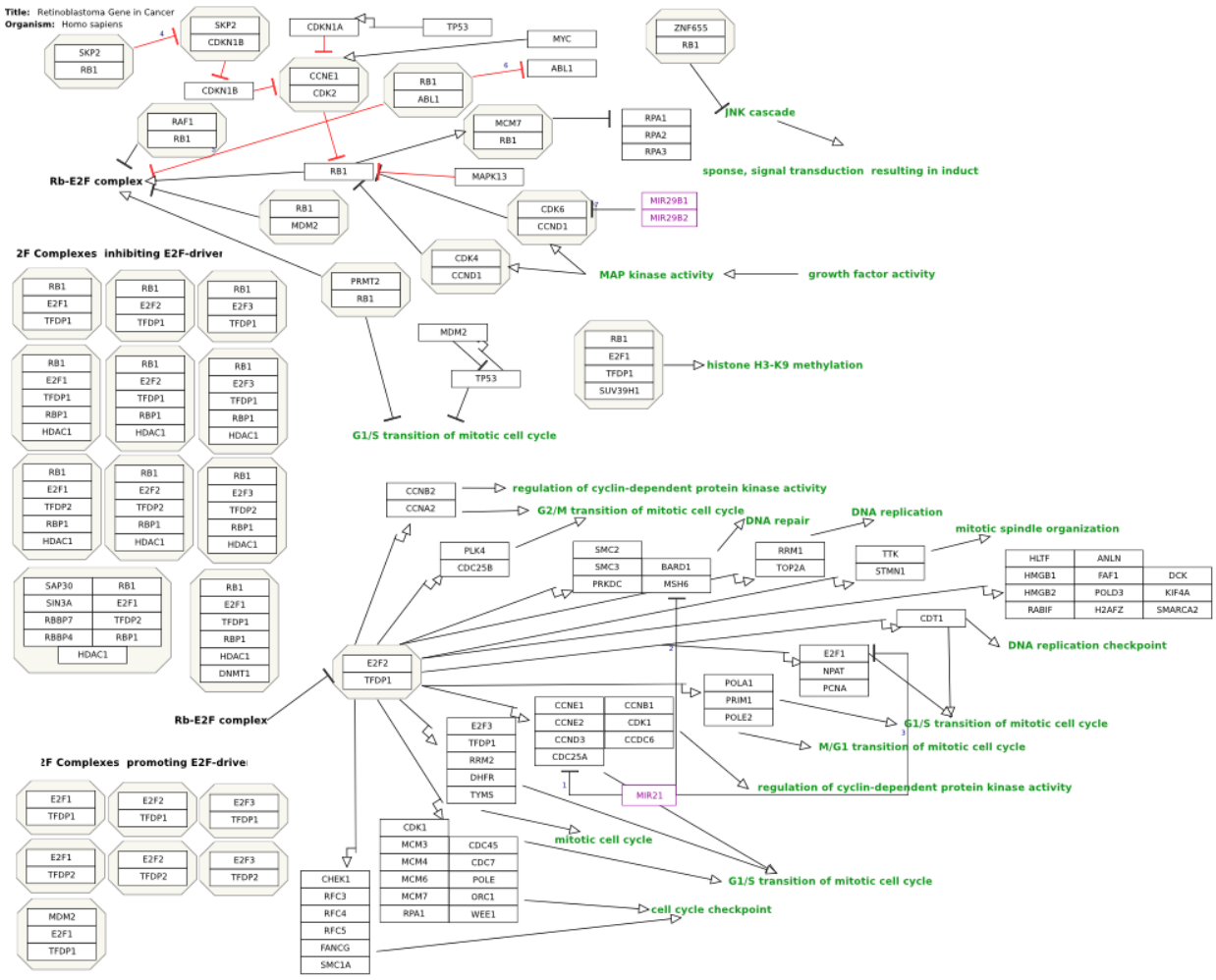
نمودار 8 pathway براساس wikipathways

Index	Name	P-value	Adjusted p-value	Z-score	Combined score
1	Retinoblastoma Gene in Cancer WP2446	1.05E-23	4.14E-21	-1.75	92.76

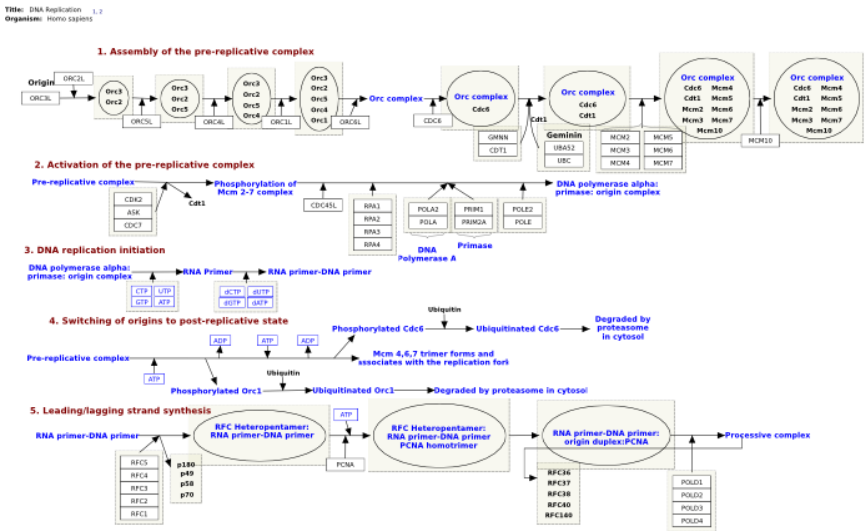
2	DNA Replication WP466	6.81E-13	7.14E-11	-2.1	58.9
3	G1 to S cell cycle control WP45	2.48E-13	4.88E-11	-1.96	56.99
4	Cell Cycle WP179	7.25E-13	7.14E-11	-1.13	31.55
5	Heme Biosynthesis WP561	0.0001585	0.003468	-3.17	27.78
6	IL1 and megakaryocytes in obesity WP2865	0.0000182	0.0007171	-2.41	26.32
7	DNA IR-damage and cellular response via ATR WP4016	2.71E-09	2.14E-07	-1.3	25.58

جدول 3 pathway براساس wikipathways

طبق پایگاه داده wikipathways [4] بروز این ژن ها معلول pathway هایی مانند retinoblasomia gene in cancer و DNA replication است. در نتیجه می توان اینگونه برداشت کرد که AML بر همانندسازی DNA تاثیر دارد و همچنین همانند ژن های مربوط به سرطان retinoblasomia است. این pathway ها را که ژن هایشان شباهت بسیاری به ژن های aml.up.genes دارند را در زیر مشاهده می کنید:



نمودار 9 *Retinoblastoma Gene in Cancer pathway*



نمودار 10 *DNA Replication pathway*

شاید wikipathways خیلی پایگاه داده قابل اطمینانی نباشد پس نتایج براساس پایگاه داده Reactom [5]را نیز گزارش می کنیم:

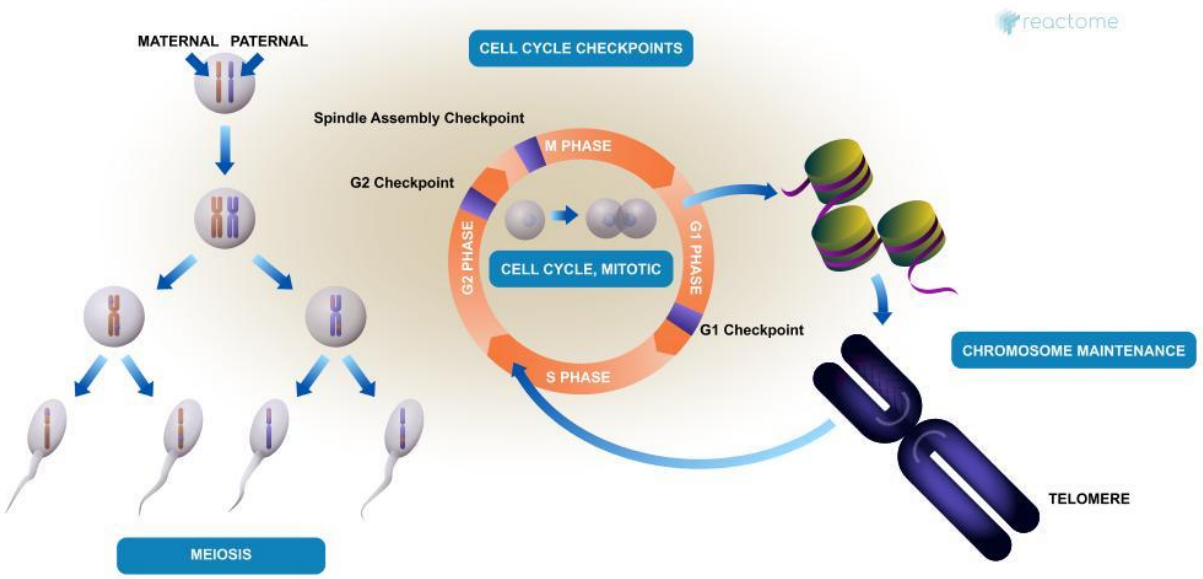
نمودار 11 pathway براساس reactome

Index	Name	P-value	Adjusted p-value	Z-score	Combined score
1	Cell Cycle_Homo sapiens_R-HSA-1640170	2.30E-50	2.61E-47	-2.46	281.3
2	Cell Cycle, Mitotic_Homo sapiens_R-HSA-69278	1.72E-45	9.73E-43	-2.47	254.39
3	M Phase_Homo sapiens_R-HSA-68886	4.21E-22	1.59E-19	-2.43	119.7
4	Cell Cycle Checkpoints_Homo sapiens_R-HSA-69620	2.04E-21	5.77E-19	-2.34	111.63
5	G2/M Checkpoints_Homo sapiens_R-HSA-69481	5.04E-20	1.14E-17	-2.32	102.92
6	Mitotic Prometaphase_Homo sapiens_R-HSA-68877	1.11E-19	2.10E-17	-2.01	87.61

7	Mitotic G1-G1/S phases_Homo sapiens_R-HSA-453279	7.36E-18	1.13E-15	-2.08	82.11

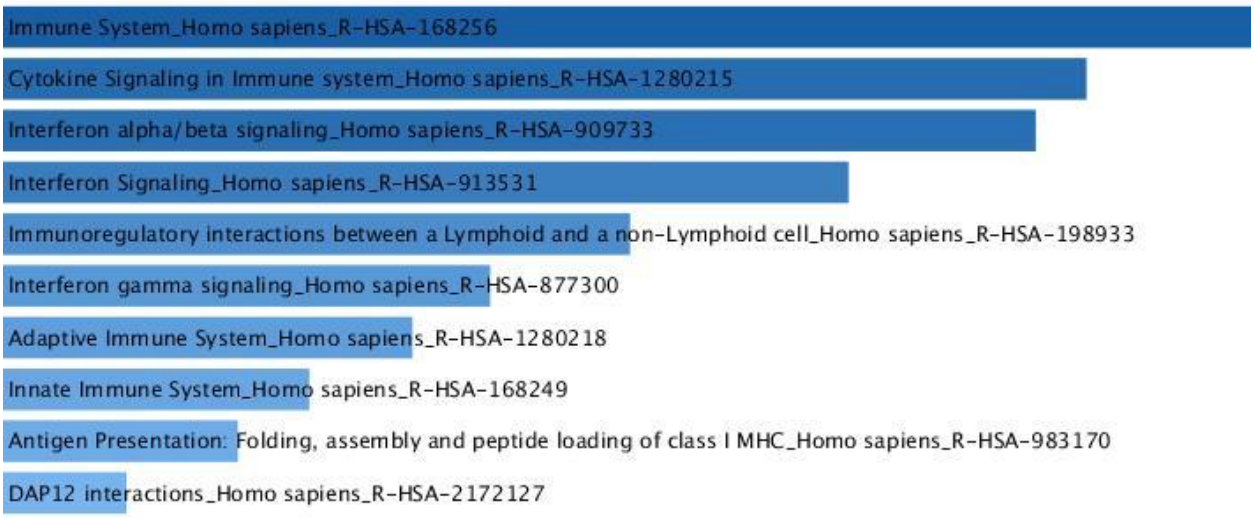
جدول 4 pathway براساس reactome

طبق این پایگاه داده pathway ما مربوط به چرخه سلولی است:



نمودار 12 cell cycle pathway

حال این کار را دوباره برای AML.Down.genes انجام می دهیم نتایج این چنین می شوند:

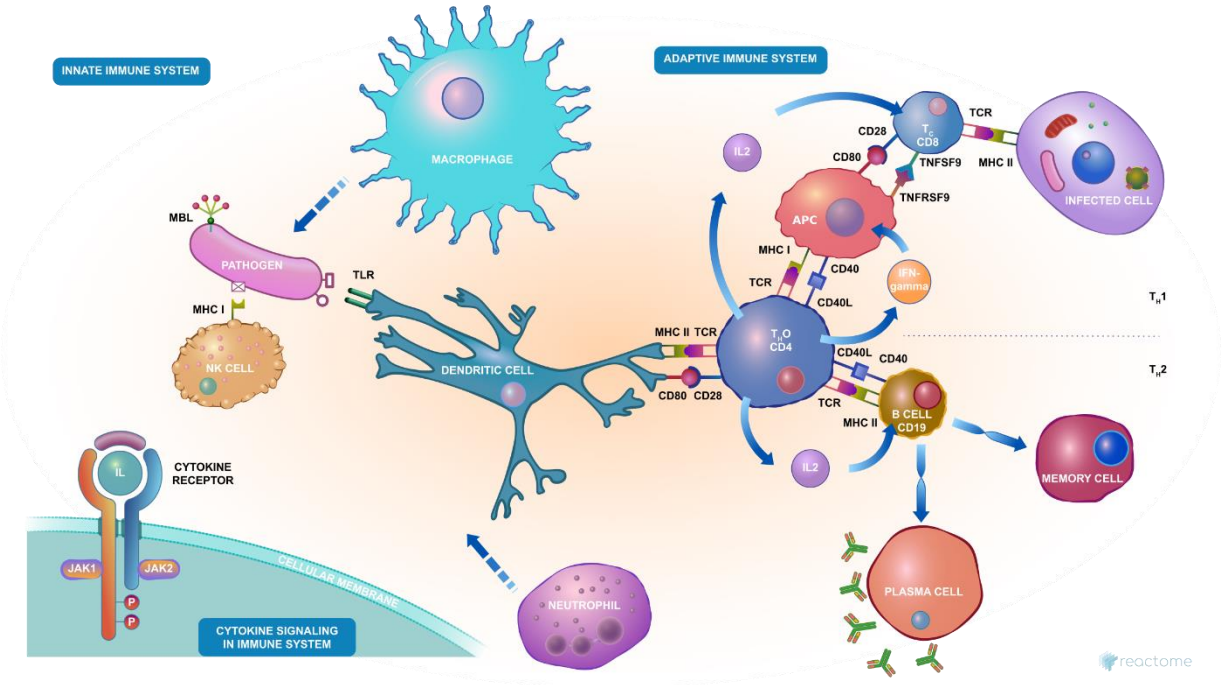


نمودار 13 pathway براساس reactome

Index	Name	P-value	Adjusted p-value	Z-score	Combined score
1	Immune System_Homo sapiens_R-HSA-168256	3.08E-14	1.40E-11	-2.22	69.19
2	Cytokine Signaling in Immune system_Homo sapiens_R-HSA-1280215	5.00E-12	1.52E-09	-2.39	62.11
3	Interferon alpha/beta signaling_Homo sapiens_R-HSA-909733	1.10E-14	1.00E-11	-1.87	59.94
4	Interferon Signaling_Homo sapiens_R-HSA-913531	1.43E-11	3.25E-09	-2.08	51.89
5	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell_Homo sapiens_R-HSA-198933	4.65E-10	8.46E-08	-1.98	42.5

جدول 5 pathway براساس reactome

براساس این نتایج می توان به این پی برد که AML مربوط به سیستم ایمنی بدن است. Pathway سیستم ایمنی:



نمودار 14 Immune system pathway

آنالیز gene antology:

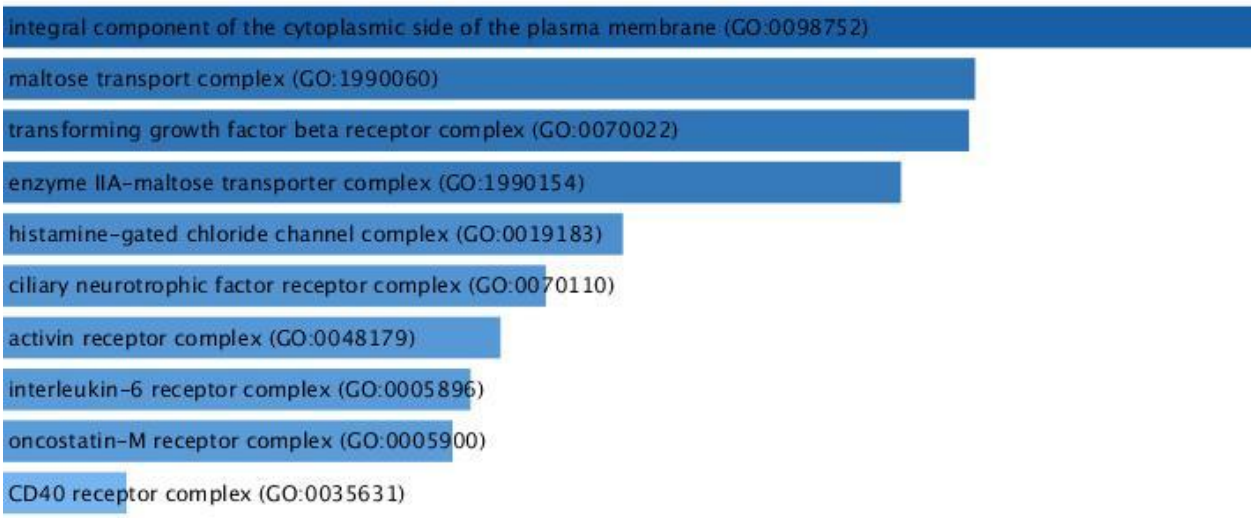
همچنان از سایت enrichr برای این کار استفاده می کنیم. برای این قسمت نتایج را از سه دید متفاوت برسی کنیم:

1. جز سلولی یا Cellular Component
2. فرآیند زیستی یا Biological Process
3. عملکرد مولکولی یا Molecular function

جز سلولی:

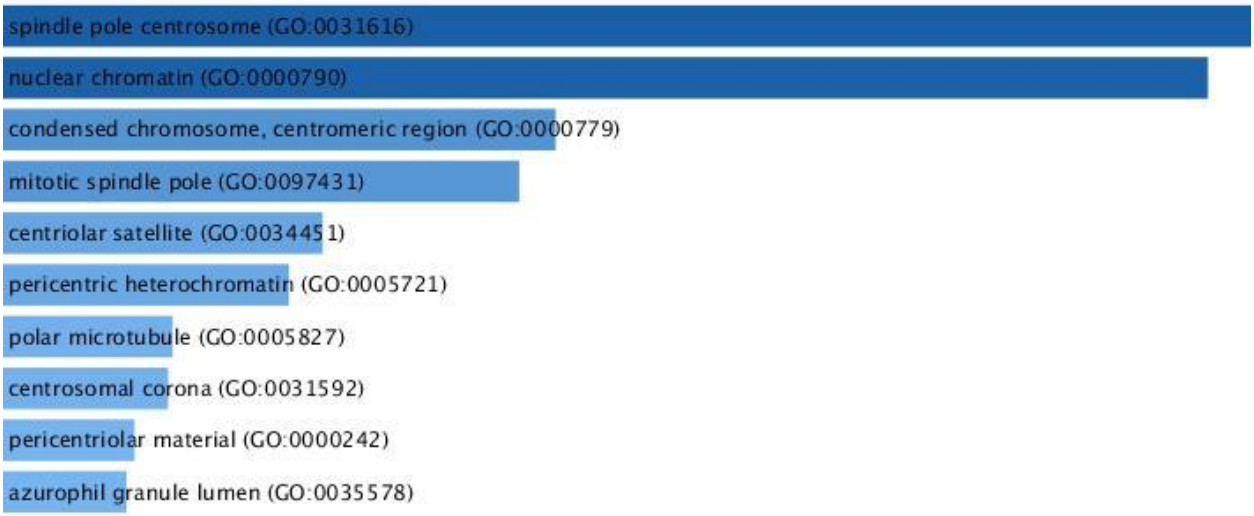
در این قسمت سعی می کنیم که براساس ژن هایی که تفاوت بیان معنی دار دارند بتوانیم سلول هایی که سرطان تحت تاثیر قرار می دهد را شناسایی کنیم.

براساس ژن های موجود در aml.down.genes این سرطان اکثرا ژن های موجود در بخش داخلی سمت سیتوپلاسمای غشای سیتوپلاسم را تحت تاثیر خود قرار می دهد:



نمودار15 جز سلولی ژن های aml.down.genes

و براساس ژن های aml.up.genes نتایج این چنین می شود:

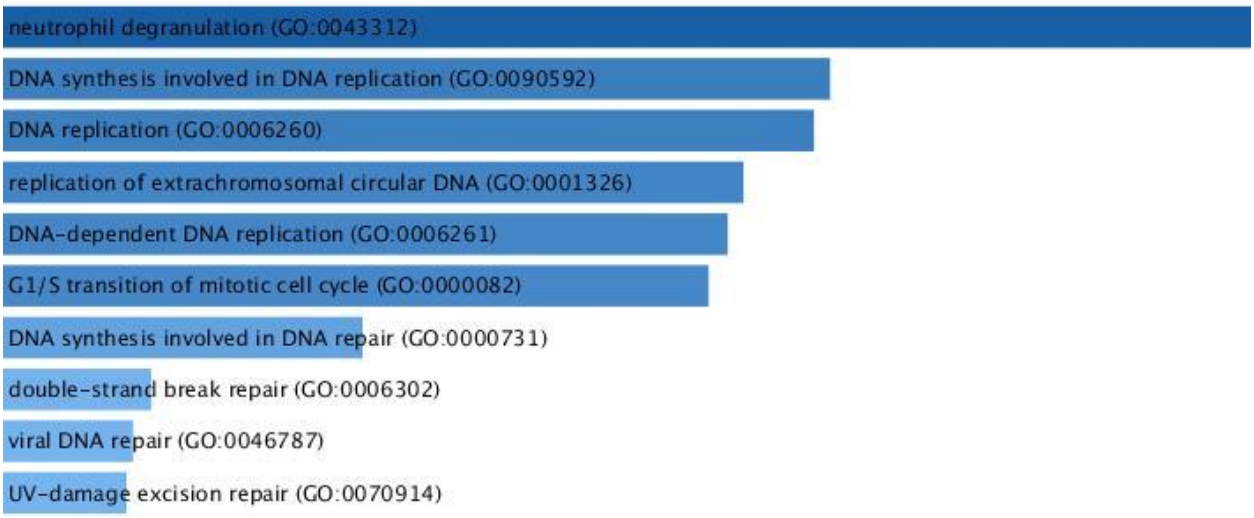


نمودار 16 جز سلولی aml.up.genes

فرآیند زیستی:

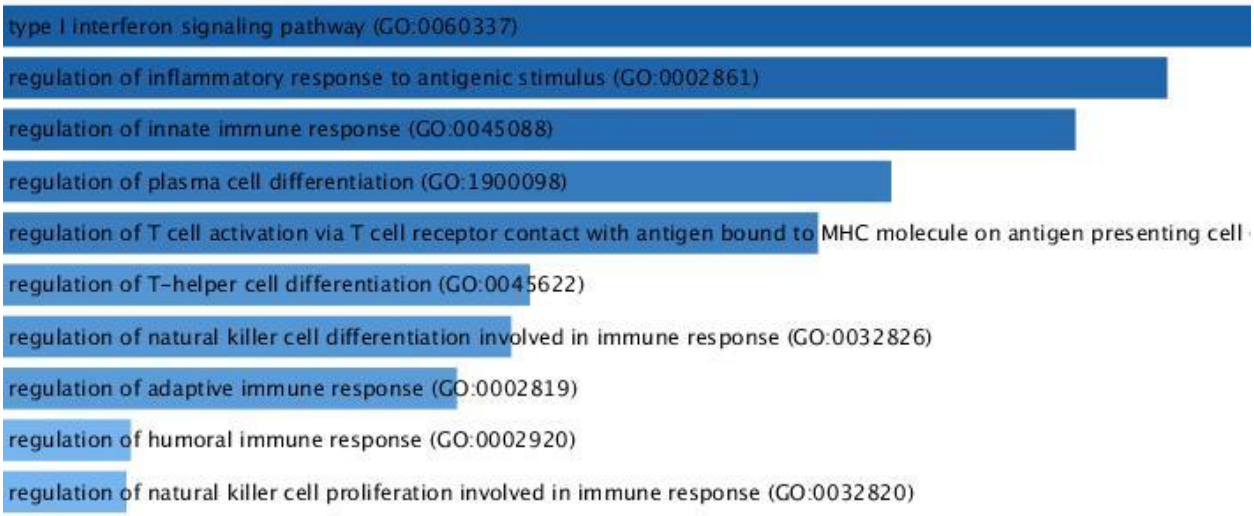
در این بخش سعی می کنیم فرآیند های زیستی ای که به AML مربوط هستند را تشخیص دهیم.

در ابتدا براساس aml.up.genes بررسی می کنیم و متوجه ارتباط تنگاتنگ این نوع سرطان با فرآیند انقراض نوتروفیل ها (neutrophil degranulation) می شویم:

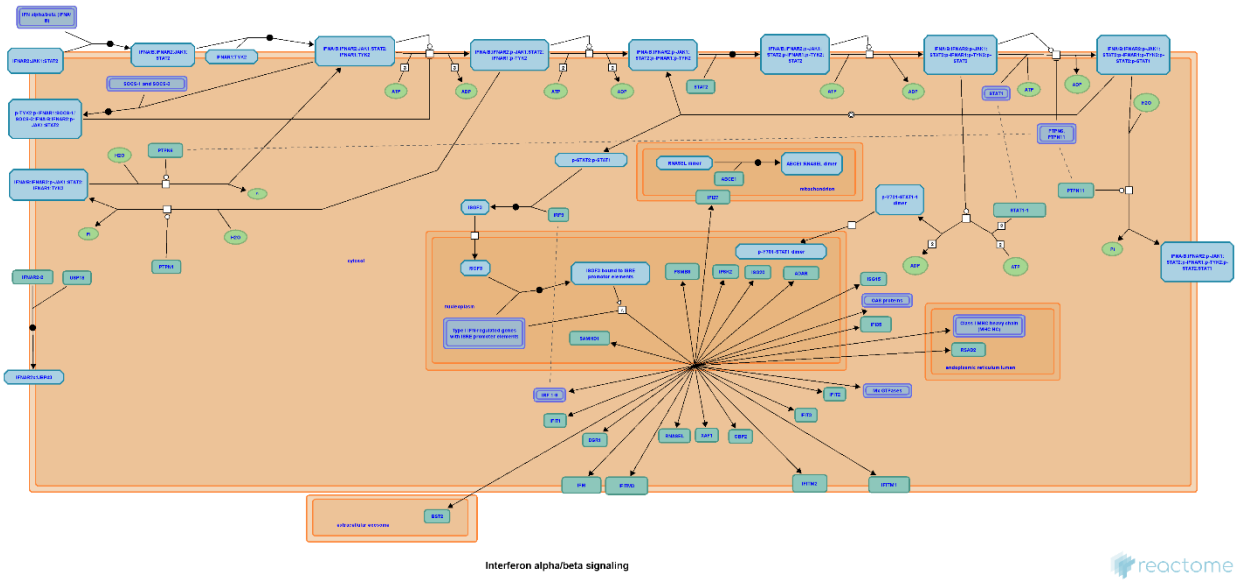


نمودار 17 فرآیند سلولی براساس aml.up.genes

براساس ژن های موجود در aml.down.genes نیز این فرآیند ها را شناسایی کردیم:



نمودار 18 فرآیند سلولی براساس Aml.down.genes



نمودار 19 *type I Interferon signalling pathway*

عملکرد مولکولی:

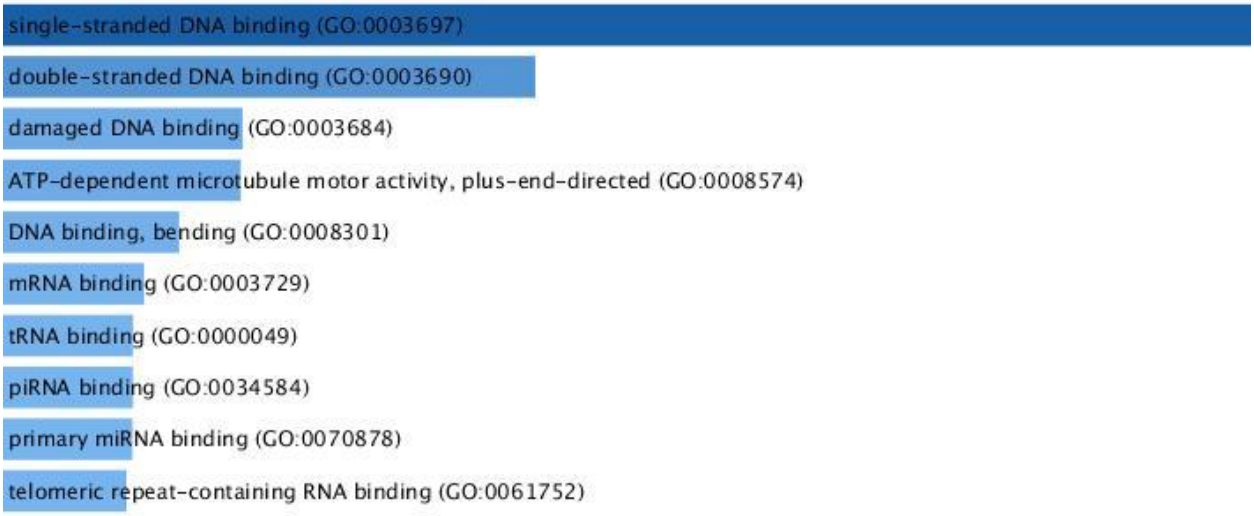
در این مرحله آنزیم ها و پروتیین ها یا به طور کلی عملکرد مولکولی تاثیرگذار یا تاثیرپذیر از AML را شناسایی می کنیم.

براساس aml.down.genes فعال کننده آنزیم GTPase یکی از نشانه های AML است:

GTPase activator activity (GO:0005096)
transmembrane receptor protein serine/threonine kinase activity (GO:0004675)
calcium-dependent protein serine/threonine kinase activity (GO:0009931)
calmodulin-dependent protein kinase activity (GO:0004683)
ribosomal protein S6 kinase activity (GO:0004711)
cyclic nucleotide-dependent protein kinase activity (GO:0004690)
Fas-activated serine/threonine kinase activity (GO:0033867)
G-protein coupled receptor kinase activity (GO:0004703)
GTP-dependent protein kinase activity (GO:0034211)
Rho-dependent protein serine/threonine kinase activity (GO:0072518)

نمودار 20 عملکرد مولکولی براساس *aml.down.genes*

و با استفاده از aml.up.genes به نتایج زیر می رسیم:



نمودار 21 عملکرد مولکولی براساس aml.up.genes

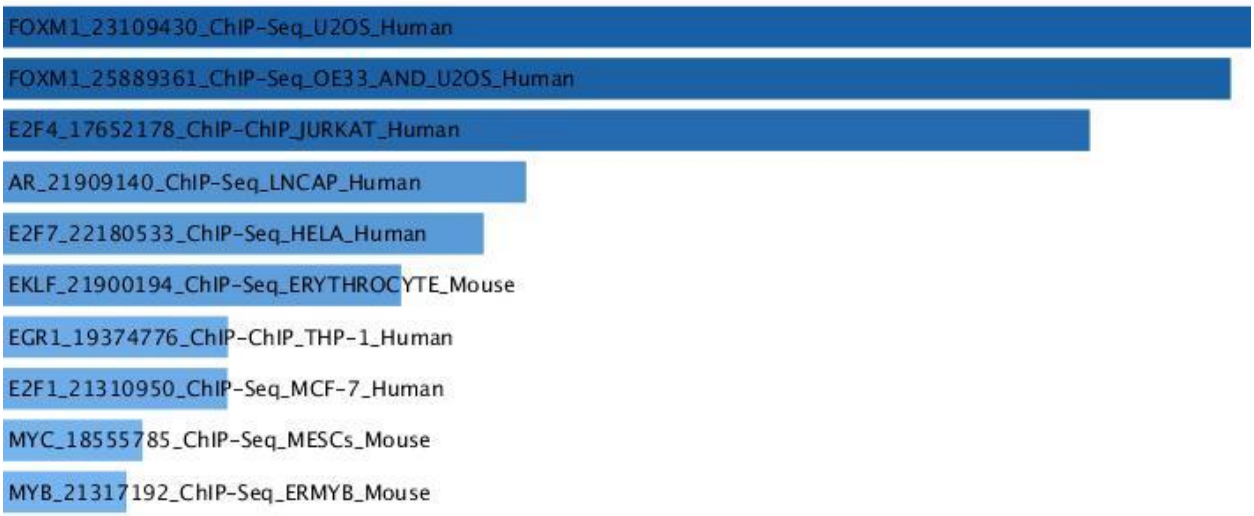
مقایسه نتایج بدست آمده با سایر مقالات زیستی:

برای نمونه سه مورد از نتایج بدست آمده در مرحله قبل را با مقالات روز حوزه پزشکی مقایسه می کنیم:

در قسمت Gene Anthology ما فرآیند زیستی type I interferon signaling pathway را به عنوان فرآیندی موثر و مربوط به AML شناسایی کردیم. مقاله Inflammatory Signaling Pathways in Preleukemic and Leukemic Stem Cells نوشته خانوم Shayda Hemmati و Tamanna Haque و Kira Gritsman [6]به بررسی این ارتباط پرداخته و در نتیجه صحت نتایج بدست آمده در این پروژه را تصدیق می کند.

فرآیند دیگری که در مرحله Gene Anthology شناسایی کردیم انقراض نوتروفیل ها (neutrophil degranulation) بود که در سایت cancer.net برای بیان توضیح کلی و مقدمه ای برای سرطان AML دقیقا متاثر از انقراض نوتروفیل ها می داند. [7]

به عنوان یک transcription factor اساسی در AML ما FOX1 را شناسایی کردیم:



نمودار 22 Transciptions factors in AML Accordimg to ChEA 2016

که می توان این transcription factor و ارتباط آن با AML را به وضوح در مقالات این حوزه مشاهده کرد مثلا مقالات [8] و [9] به اثر FOX1 در تکثیر AML می پردازند.

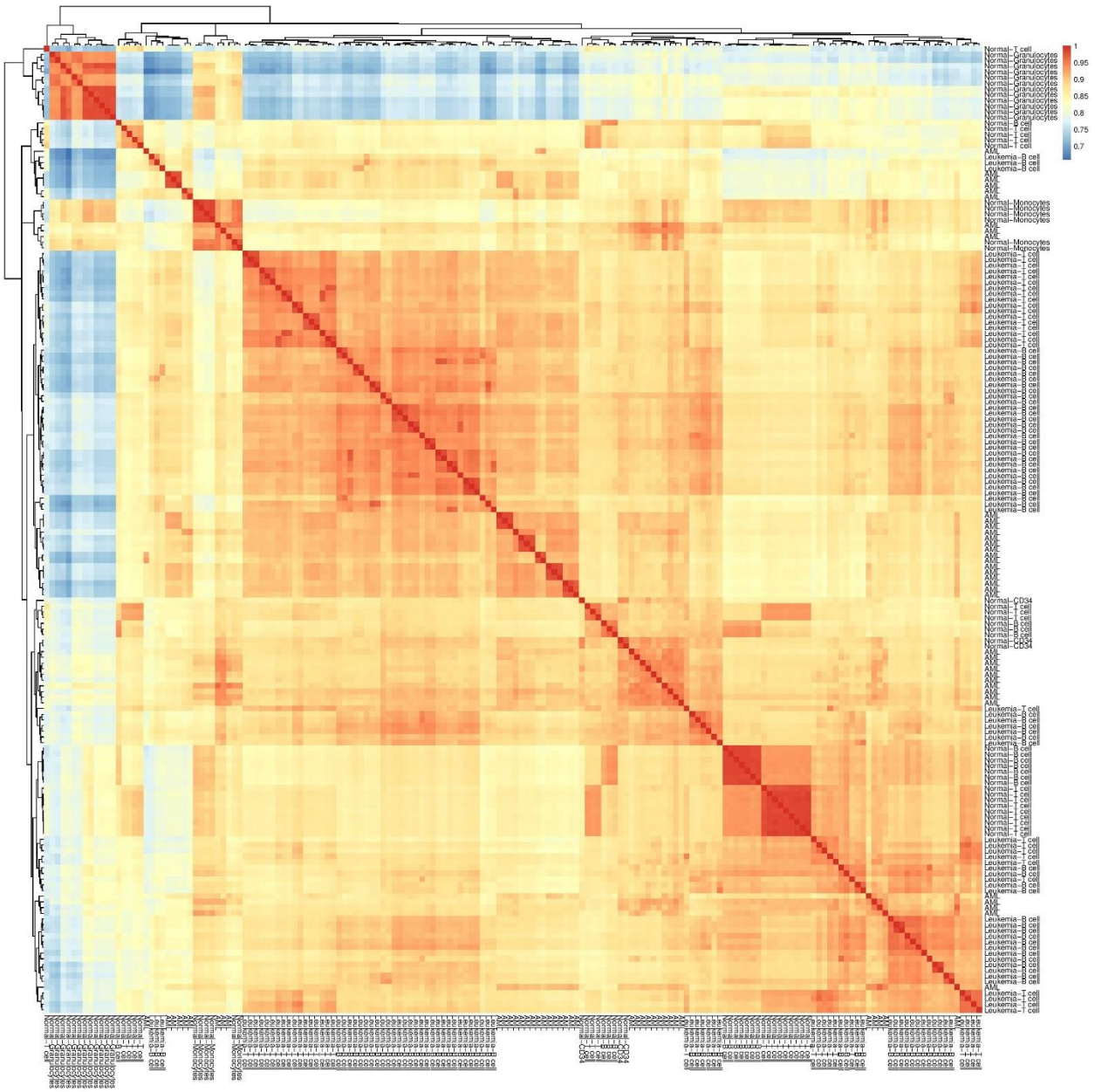
بررسی تفاوت زیر گروه های داده:

روال همانند قبل است با این تفاوت که به جای تنها سه گروه Leukemia ، AML و Normal دسته بندی را جزیی تر کرده و براساس:

```
AML, Leukemia-B cell, Leukemia-T cell, Normal-B cell, N
ormal-CD34, Normal-Granulocytes, Normal-Monocytes, Norm
al-T cell
```

دسته بندی کرده ایم. برای اجرا این بخش فایل subgroupsOfData.R را اجرا کنید.

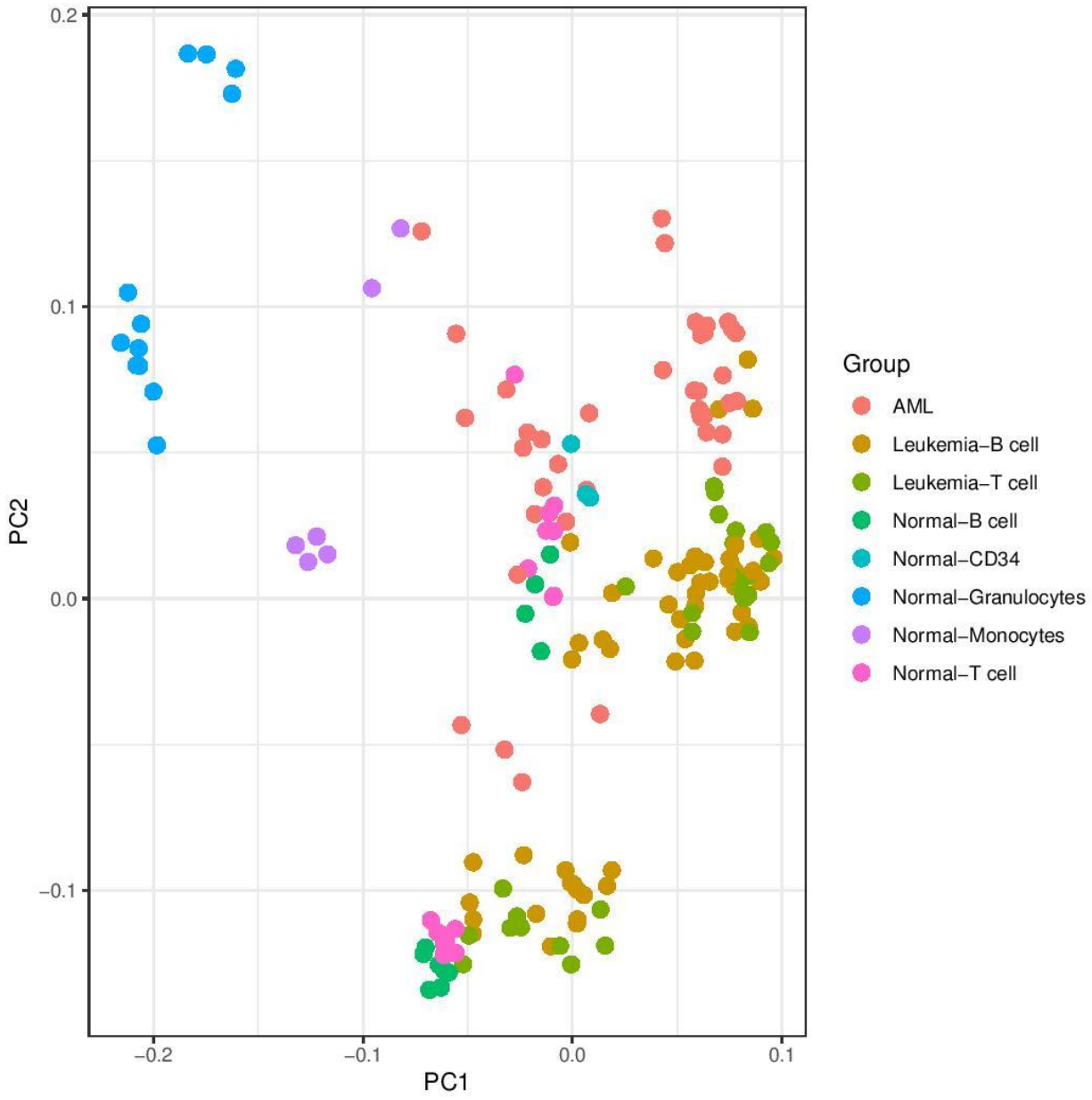
نمودار heatmap ما به این شکل می شود:



نمودار 23 heatmap با در نظر گرفتن زیرگروه های داده ها - فایل CorHeatmap_subgroups.pdf

حال می بینیم که داده های یک نوع correlation بسیار مشابه ای دارد پس کیفیت داده ما احتمالا عالی است.

در مرحله آخر بررسی کیفیت، به همبستگی بین نمونه ها می پردازیم. نتیجه چنین می شود:



نمودار 24 نمایش همبستگی بین نمونه با در نظر گرفتن زیر گروه ها در فضای PC1,PC2 - فایل PCA_samples_subgroups.pdf

مشاهده می شود که گروه های یکسان کاملا کنار یکدیگر قرار می گیرند.

برای آنالیز pathway و gene anthology هم باید دو به دو AML را با Normal B cell و Normal T cell و ... مقایسه کنیم.

مباحثات آینده:

با توجه به ورود توانایی های پردازشی کامپیوتر ها به علم پزشکی و پیشرفت علم بیوانفورماتیک از طرفی به دست آوردن داده های بیشتر از افراد مبتلا به AML، در اختیار قرار گرفتن اطلاعات بیشتر درباره ژن ها، Transcription factor ها و pathway ها و همچنین ظهور روش هایی همچون crispr gene editing [10] ، انتظار می رود که عوامل موثر در AML بیشتر و بیشتر شناخته شوند و همچنین عوامل موثر بر این عوامل هم شناخته شوند؛ در نتیجه می توان از بروز آن جلوگیری به عمل آورد و یا داروهایی موثر تر و به تناسب این عوامل برای آن ساخت. در نهایت امید می رود که با پیشرفت روش های مهندسی ژنتیک از جمله crispr gene editing بتوان به طور مستقیم ژن های سرطانی شناخته شده را اصلاح کرد و خطر سرطان AML را به طور کامل از بین برد.

منابع:

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48558

[2] https://amp.pharm.mssm.edu/Enrichr/

[3] https://amp.pharm.mssm.edu/OxEnrichr/

[4] https://www.wikipathways.org/index.php/WikiPathways

[5] https://reactome.org/

[6] https://www.frontiersin.org/articles/10.3389/fonc.2017.00265/full

[7] https://www.cancer.net/cancer-types/leukemia-acute-myeloid-aml/introduction

[8] https://www.ncbi.nlm.nih.gov/pubmed/20823107

[9] http://www.bloodjournal.org/content/early/2019/04/25/blood.2018893982?sso-checked=true

[10] https://en.wikipedia.org/wiki/CRISPR_gene_editing