



دانشگاه صنعتی شریف

گزارش پژوهه درس مقدمه‌ای بر بیوانفورماتیک

تحلیل داده‌های میکرواری لوکمیا

علیرضا اکبری ۹۵۱۰۵۳۷۹

توجه: به علت حجم زیاد یکی از فایل‌های txt. در فایل گذاشته نشد. با ران کردن کد، ایجاد می‌شود

۱ مقدمه ، دریافت داده و دسته‌بندی

بیماری AML یکی از انواع سرطان خون است. در این بررسی قصد داریم با استفاده از داده‌های GSE۴۸۵۵۸ ژن‌هایی که تفاوت بیان معنی‌داری میان دو دسته نرمال و سرطانی دارند را پیدا کنیم و با استفاده از پایگاه داده‌های آنلاین به تحلیل و پیدا کردن pathway مربوطه پردازیم. داده‌ها به صورت کلی بدین صورت دسته‌بندی شده‌اند که براساس phenotype داده‌های نرمال در یک دسته قرار گرفته‌اند و نمونه‌های نرمال و سالم را تشکیل داده‌اند. همچنین براساس source name داده‌های AML Patient نیز همگی در یک دسته قرار گرفته‌اند و دسته سرطانی را تشکیل می‌دهند. باقی سمپل‌ها بر حسب source name دسته‌بندی شده‌اند. همچنین هرجایی که لازم بوده است بررسی دقیق‌تر شود ، تمام داده‌ها به غیر از دو دسته نرمال و AML Patient حذف شده‌اند و تنها بین دو گروه مقایسه انجام داده‌ایم. در این موقعیت سمپل‌های نرمال را بر حسب source name شان هم دسته‌بندی کردہ‌ایم تا شهود بهتری نسبت به سمپل‌ها پیدا کنیم. در نهایت هم بررسی تمایز بیان ژن‌ها میان دو گروه AML Patient و نرمال (بر حسب phenotype) انجام شده است. در ادامه به کنترل کیفیت داده و سپس به تحلیل می‌پردازیم.

۲ کنترل کیفیت داده

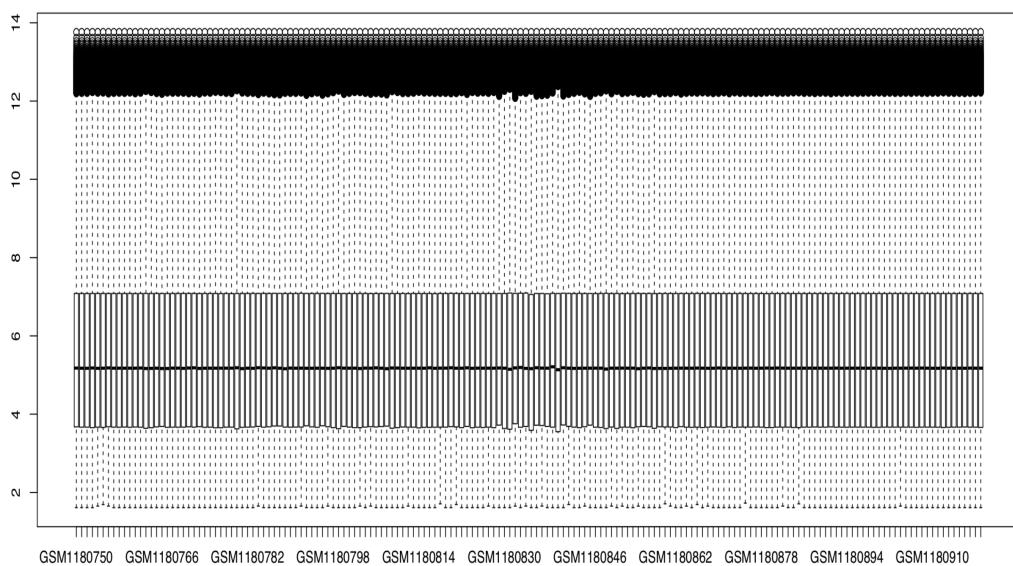
در قسمت‌ها کنترل کیفیت (همچنین در دو بخش بعدی)، همانند کاری که در ویدیوها انجام شد، تحلیل کیفیت را روی کل داده‌ها انجام می‌دهیم و هرجا که لازم شد داده‌هایی به غیر از AMLP و AML را حذف می‌کنیم و کیفیت این دو دسته را به صورت جداگانه بررسی می‌کنیم. Normal بعد از دریافت داده‌های GSE۴۸۵۵۸ و جداسازی ماتریس بیان ژن‌ها، می‌خواهیم در این بخش کیفیت داده را کنترل کنیم.

۱.۲ بررسی مقیاس لگاریتمی

ابتدا می‌بایست بررسی کنیم که آیا اعداد موجود در ماتریس بیان ژن‌ها به مقیاس لگاریتمی تبدیل شده است یا نه. برای این کار بزرگترین عدد موجود در ماتریس را بررسی می‌کنیم که ۱۳,۷۶۱۵۴ است. نتیجه می‌گیریم که ماتریس بیان در مقایس لگاریتمی است.

۲.۲ بررسی نرمال بودن داده‌ها

برای آنکه بتوانیم داده‌ها را با یکدیگر مقایسه کنیم، می‌بایست همگی نرمال شده باشند. برای بررسی نرمال بودن داده‌ها نمودار جعبه‌ای داده‌ها را در شکل ۱ رسم می‌کنیم.

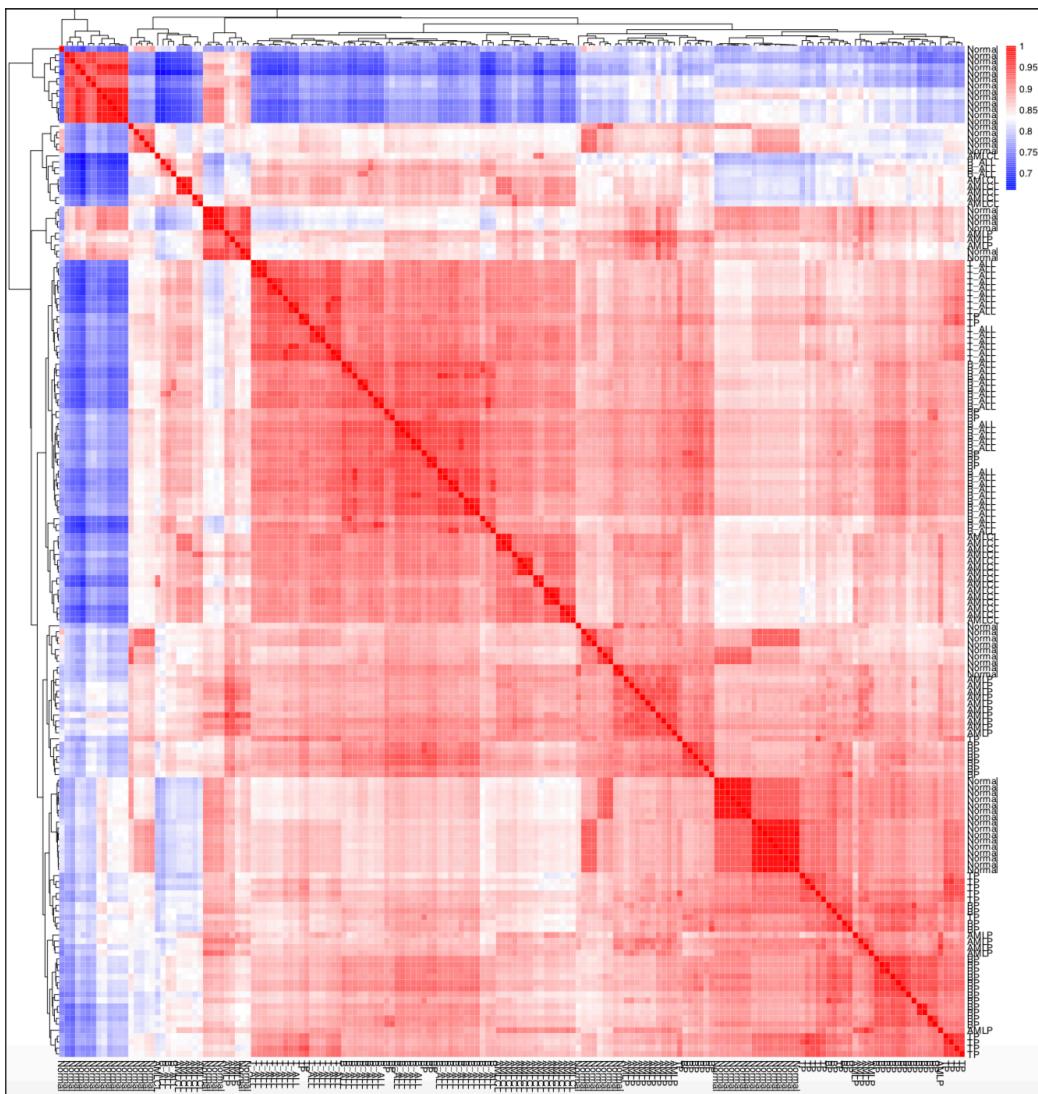


شکل ۱ : Boxplot of Expression Matrix

همان‌طور که مشاهده می‌شود میانه هر نمونه تقریبا همگی روی یک خط قرار دارند پس می‌توان نتیجه گرفت که داده‌ها نرمال نیز شده بودند. همچنین از این شکل دارای مقیاس لگاریتمی بودن نیز قابل فهم است.

۳ بررسی همبستگی بین نمونه‌ها

این بخش و بخش بعدی (کاهش ابعاد داده) همچنان جز مراحل کنترل کیفیت داده‌ها می‌باشد. اما چون در صورت پروژه جدا شده است اینجا نیز به صورت جداگانه بررسی می‌کنیم. در این بخش می‌خواهیم کوریلیشن میان سمپل‌ها را بررسی کنیم. نحوه دسته‌بندی سمپل‌ها را توضیح دادیم. برای بررسی کوریلیشن میان سمپل‌ها از Heatmap استفاده می‌کنیم. ابتدا Heatmap را بر روی کل سمپل‌ها با دسته‌بندی گفته شده رسم می‌کنیم. ۲ در ابتدا برداشتی که می‌توان از شکل ۲ کرد این است که سمپل‌های نرمال به چند دسته تقسیم



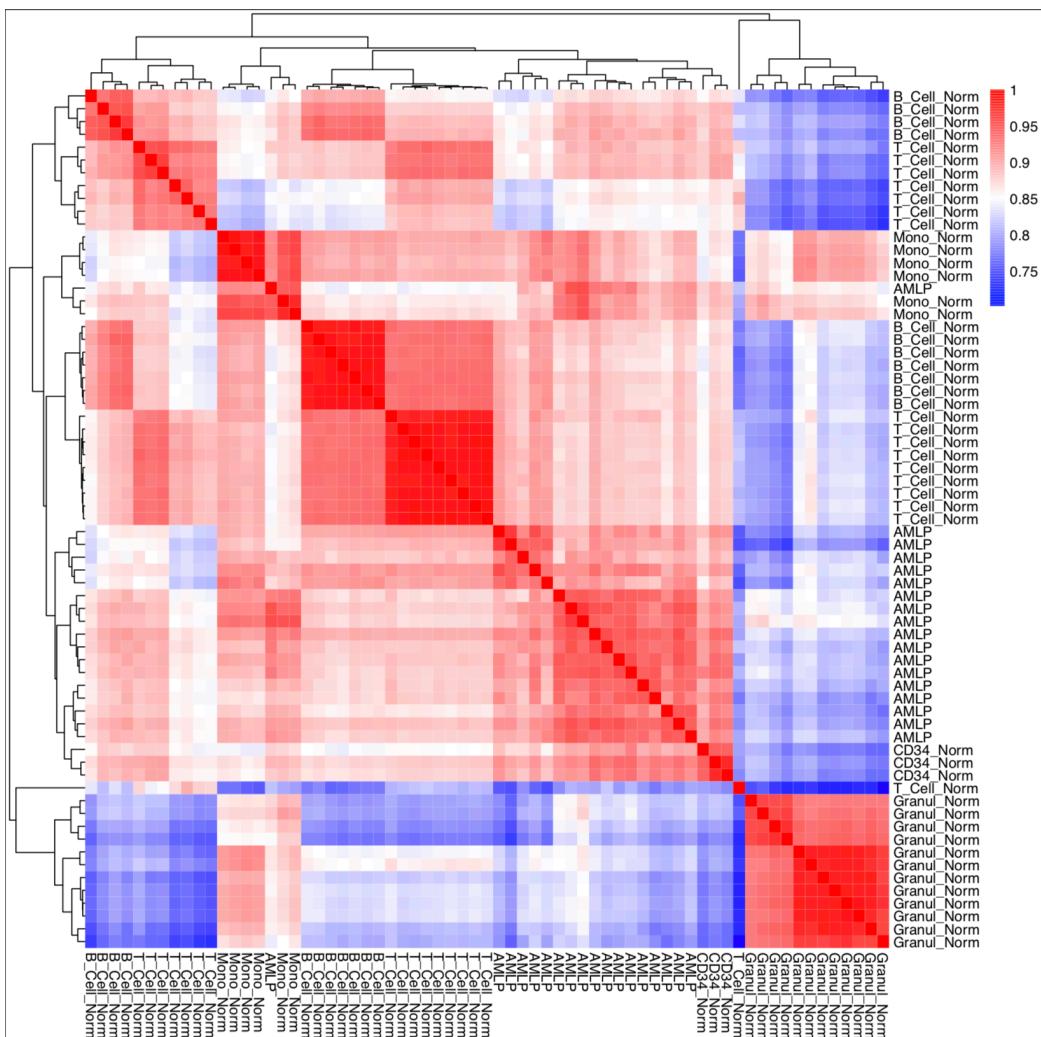
شكل ٢ : Correlation Heatmap on whole Dataset

شده‌آند که هر دسته با دسته خود correlation بالایی دارد. یعنی دسته نرمال خود به چند -
sub group تقسیم شده است . که البته قابل انتظار نیز بود چرا که ما داده‌های نرمال را تنها بر حسب
phenotype دسته‌بندی کرده بودیم در حالی که طبق داده‌های GSE۴۸۵۵۸ داده‌های نرمال بر
حسب source name خود به چند دسته تقسیم شده‌اند. از نمودار مشخص است که هر subgroup
با خودش کوریلیشن بالایی دارد که البته قابل انتظار هم بود . همچنین مشاهده می‌شود که بعضی از
زیرگروه‌های normal نیز انگار خود دوباره به تعدادی subgroup تقسیم شده است، جلوتر
به بررسی دقیق‌تر می‌پردازیم. سمپل‌های AML Patient نیز تشکیل چند دسته داده‌اند که البته طبق
انتظار کوریلیشن هر دسته با خودش لزوماً زیاد نیست که این ناشی از ماهیت سرطان است.
حال برای بررسی دقیق‌تر، سمپل‌هایی به غیر از دو دسته Normal و AML Patient را حذف
می‌کنیم. همچنین برای شناسایی بهتر انواع سمپل‌های نرمال، سمپل‌های نرمال را هم براسن
name تفکیک می‌کنیم. و مجدداً heatmap را حال برای این دسته‌بندی جدید که از ابتدا دنبال
بررسی تفاوت میان آن‌ها بودیم رسم می‌کنیم. ۳

نتایج جالبی در این نمودار دیده می‌شود و می‌توانیم تحلیل دقیق‌تری داشته باشیم. نمونه‌های نرمال
T_Cell به سه subgroup تقسیم شده‌اند که یک دسته تنها دارای یک سمپل است . نمونه‌های
نرمال B_Cell نیز به دو subgroup تقسیم شده‌اند. نکته جالبی که دیده می‌شود این است که هر
یک از subgroup های B_Cell و T_Cell با یکدیگر تشکیل کلاستر داده‌اند.
سمپل‌های سرطانی طبق انتظار یک دسته تشکیل داده‌اند به غیر از یک نمونه. نمونه‌های نرمال
Monocytes نیز به دو subgroup تقسیم شده‌اند که یکی از زیرگروه‌ها با یک عدد سمپل سرطانی
کوریلیشن نسبتاً قابل توجهی دارد . در ادامه باید تصمیم بگیریم که آیا هیچ یک از این داده‌ها
را می‌توانیم outlier در نظر بگیریم یا خیر. این موضوع را در بخش PCA بیشتر بررسی خواهیم
کرد.

مشاهده می‌شود که سمپل‌های نرمال Granulocytes با یکدیگر تشکیل دسته داده‌اند(که خود به
دو subpopulation تقسیم شده است) و کوریلیشن بالایی با هم دارند اما با بقیه نمونه‌ها هیچ
ارتباطی ندارد و شاید گزینه مناسبی برای مقایسه نباشد. این بحث در در بخش PCA بیشتر بررسی
خواهیم کرد.

همچنین سه سمپل نرمال CD۳۴ با سمپل‌های سرطانی تشکیل یک کلاستر داده‌اند. درباره
سمپل‌های CD۳۴ نیز با در بخش PCA با بررسی دقیق‌تر سعی می‌کنیم ارتباطشان را با سمپل‌های
سرطانی پیدا کنیم و بینیم واقعاً شباهت قابل توجهی با یکدیگر دارند یا نه .
در مجموع به غیر از دو مورد نتایج ماتریس کوریلیشن طبق انتظارمان بود و باعث شد شناخت
بیشتری نسبت به داده‌ها و کیفیت آن‌ها پیدا کنیم.



شكل ٣ Correlation Heatmap on Normal and AML Patient :

۴ کاهش ابعاد داده

در ادامه کنترل کیفیت داده، می‌خواهیم با روشی نمونه‌ها را در یک فضای برداری دو بعدی یا سه بعدی رسم کنیم تا پراکندگی آن‌ها را بهتر ببینیم و بررسی کنیم که آیا ناظاطمان تشکیل دسته‌های بخصوصی می‌دهند یا نه.

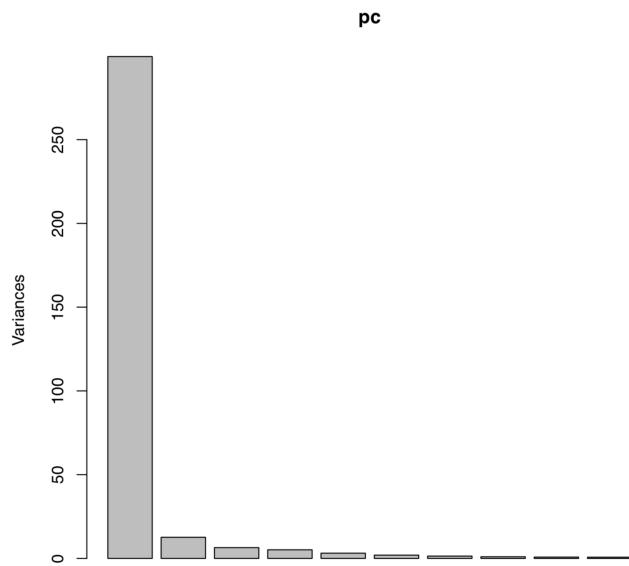
برای این مفهوم از روش Principal Component Analysis استفاده می‌کنیم. این روش بدین صورت عمل می‌کند که پایه‌هایی برای فضا پیدا می‌کند که آن پایه‌ها بیشترین پراکندگی نقاط را شامل شوند. و از دو یا سه پایه با بیشترین پراکندگی (واریانس) برای ترسیم نقاط استفاده می‌کنیم. حال این روش را روی ماتریس بیان ژن‌ها (کل داده‌ها) اتخاذ می‌کنیم. واریانس نقاط روی PC ها به صورت شکل ۴ در می‌آید. همچنین پراکندگی ژن‌ها روی دو PC اول به صورت شکل ۵ در می‌آید. همچنان که از روی شکل ۵ مشخص است، ما انگار راستایی را پیدا کردیم که نشان‌دهنده میانگین بیان ژن‌ها از کم به زیاد است. انگار ژن‌ها را بر اساس میانگین بیان در سمپل‌های مختلف سورت کرده‌ایم که این چیزی نبود که دنبال آن بودیم. این امر از روی شکل ۴ نیز قابل استنباط بود. چرا که واریانس PC اول بسیار زیاد و واریانس باقی بسیار ناچیز است.

برای رفع این مشکل، در ماتریس بیان، عدد بیان هر ژن در یک سمپل را از میانگین بیان آن ژن در تمام سمپل‌ها کم می‌کنیم. در واقع با این کار هر عدد نشان‌دهنده تغییرات بیان آن ژن خاص در آن نمونه است. حال دوباره دو نمودار واریانس PC ها (۶) و پراکندگی ژن‌ها (۷) را برای ماتریس جدید رسم می‌کنیم.

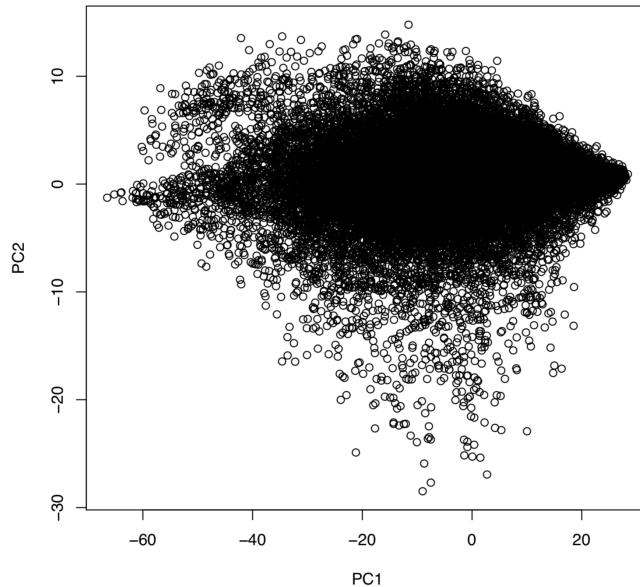
با توجه به نمودارهایی بدست آمده، نتیجه می‌گیریم که به هدف خود رسیده‌ایم و پایه‌هایی پیدا کرده‌ایم که پراکندگی ژن‌ها را بهتر به تصویر کشانده‌اند. همچنین برخلاف حالت قبلی PC دوم و PC سوم نیز واریانس نسبتاً بالایی پیدا کرده‌اند.

تا بدینجا پراکندگی ژن‌ها را ترسیم کردیم. حال می‌خواهیم پراکندگی سمپل‌ها را رسم کنیم. اگر این کار را روی دو PC اول رسم کنیم به نمودار ۸ می‌رسیم. همان طور که در بخش همبستگی نیز دیدیم، مطابق انتظار سمپل‌های نرمال به چند زیردسته تقسیم شده‌اند. و نمونه‌های سرطانی طبق انتظار نسبتاً ناهمگون هستند. همچنین از روی شکل به نظر می‌آید که سمپل‌های AML Patient و یک سری از سمپل‌های نرمال به خوبی از یکدیگر تفکیک نشده‌اند. برای آنکه دقیق‌تر بررسی کنیم، سمپل‌ها را در سه بعد PC₁ و PC₂ و PC₃ بررسی می‌کنیم چرا که همان‌طور که دیده شد واریانس PC₃ نیز زیاد است. همچنین سمپل‌هایی به غیر از AML Patient و Normal را از سمپل‌ها حذف می‌کنیم تا تفاوت‌های این دو گروه را جزئی‌تر بررسی کنیم. سمپل‌های نرمال را براساس آن‌ها دسته‌بندی کرده‌ایم. رسم سمپل‌ها در سه بعد به نمودار ۹ می‌انجامد. (در فایل source name html کنار فایل گزارش می‌توان نمودار سه بعدی را به صورت تعاملی بررسی کرد)

نتیجه مشاهده شده در این نمودار مصدق نتایج مشاهده شده در قسمت کوریلیشن است. همچنین مشاهده می‌شود که هر نوع دسته‌بندی در subpopulation خود کنار یکدیگر قرار گرفته‌اند که نشان از کیفیت داده دارد.

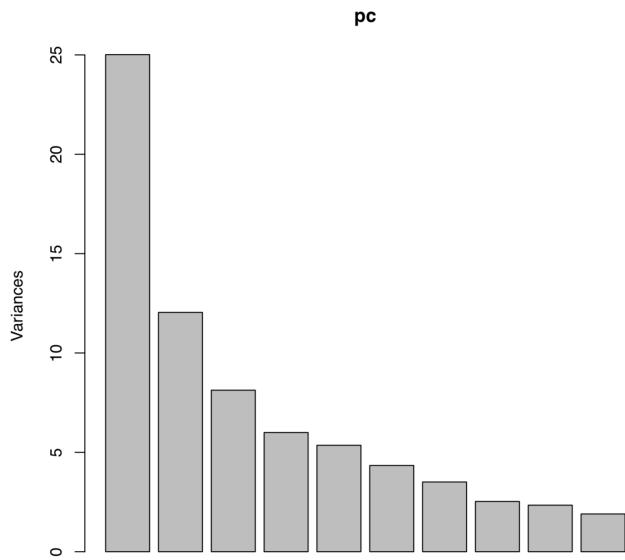


Variances on Principal Components - not scaled : ٤ شکل

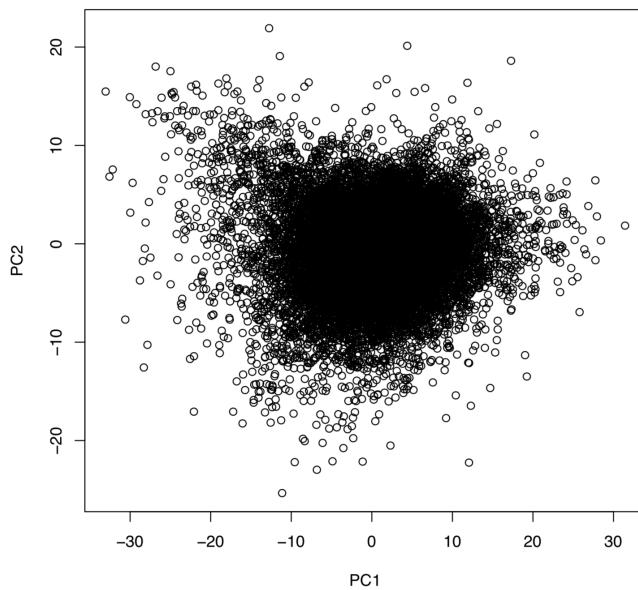


Genes on PC1 and PC2 - not scaled : ٥ شکل

v

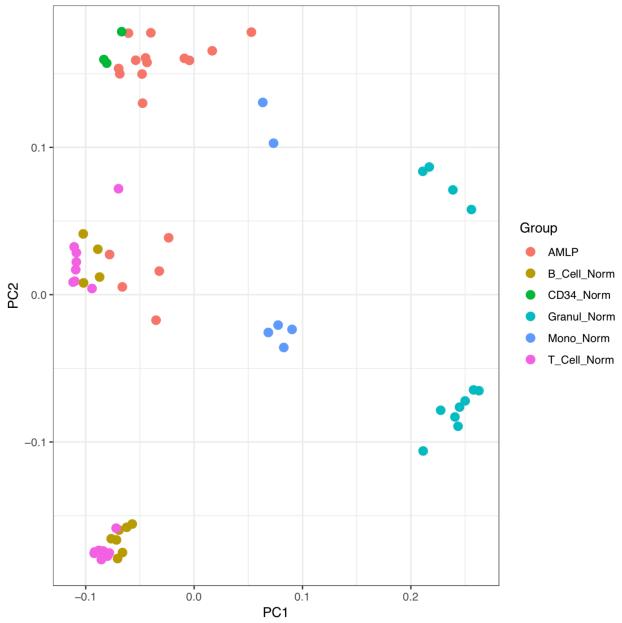


Variances on Principal Components - scaled : ٦ شکل



Genes on PC1 and PC2 - scaled : ٧ شکل

٨



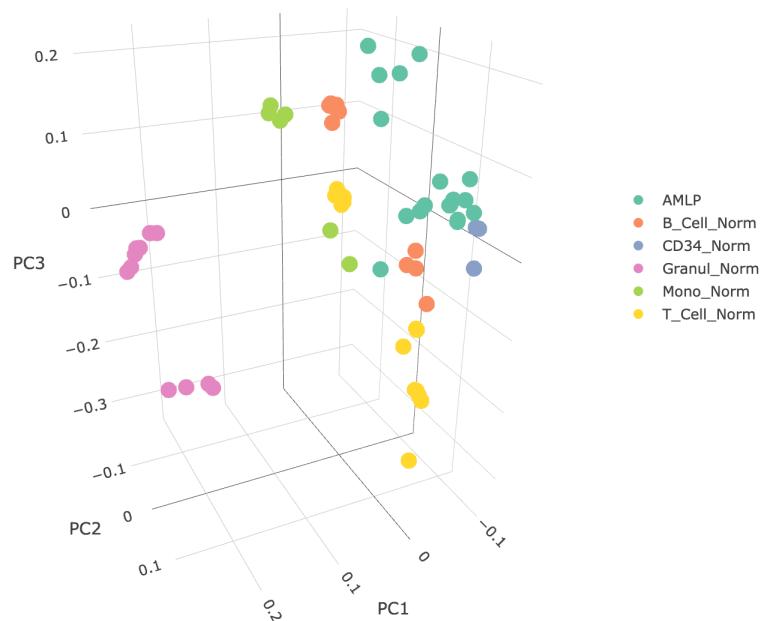
شکل : ۸ Samples :

متوجه می شویم که به غیر از دو مورد، تمام سمپل های سرطانی به خوبی از سمپل های نرمال جدا شده اند. این دو مورد همان دو موردی است که در بخش بروسی کوریلیشن هم به آن ها برخور迪م. سمپل های CD34 به یکی از کلاستر های سرطانی بسیار نزدیک است. و یک سمپل سرطانی که به نظر جز هیچ یک از دو کلاستر سمپل های سرطانی نیست به یکی از زیر گروه های سمپل نرمال monocyte نزدیک است.

همچنین همان طور که در بخش قبل دیدیم، سمپل های نرمال Granulocytes دو - subpopula tion دارند که در نتیجه دو دسته تشکیل داده اند. همان طور که قابل انتظار بود، این دو دسته از باقی سمپل ها بسیار متفاوت است. پس نهایتا می توان گفت که مجموع داده ها به خوبی پخش و دسته شده اند و می توان به داده ها اطمینان کرد.

برای تحلیل در ابتدا هیچ داده ای را به عنوان داده پرت در نظر نمی گیریم و همان طور که در صورت پژوهه گفته شده است، نتایج را میان دو گروه AML Patient و Normal Bedst می اوریم و تحلیل می کنیم و در صورت الزام به داده ها باز می گردیم.

AML Patient and Normal samples on PC1, PC2, PC3



Samples : ٩ شكل

۵ بررسی تمایز در بیان ژن‌های نمونه‌ها

در این بخش می‌خواهیم لیستی از ژن‌هایی را پیدا کنیم که تفاوت معنی‌داری در بیان آن‌ها میان دو دسته نرمال و AML Patient وجود دارد. همچنین در ابتدا هیچ داده‌ای را به عنوان outlier در نظر نمی‌گیریم. در واقع می‌خواهیم جدولی درست کنیم که نشان دهد برای هر ژن چه مقدار تغییر (logFC) داشته‌ایم و این تغییر چقدر معنادار بوده است. این کار را همانند کاری که در ویدیوها انجام شد با استفاده از قطعه کد موجود در صفحه مربوط به داده‌ها انجام می‌دهیم. در نهایت به جدولی مانند جدول ۱ می‌رسیم

جدول ۱ : چند نمونه از تمایز بیان ژن‌ها میان دو دسته نرمال و سرطانی

| Gene.Symbol | Gene.ID | adj.PVal | logFC |
|-------------|-------------|----------------|-----------|
| KIAA0101 | 9768 | 4.097322e - 35 | 4.308426 |
| TYMS | 7298 | 1.030528e - 28 | 3.353002 |
| DTL | 51514 | 1.749621e - 28 | 3.412998 |
| CBX7 | 23492 | 3.201488e - 27 | -1.992279 |
| MYB///MYB | 4602///4602 | 2.500051e - 25 | 3.609204 |

این جدول به طور کامل در بخش Results به نام AMLP_Normal.txt وجود دارد. در واقع این جدول برای آنالیز pathway و ontology می‌کند که در بخش بعدی نقش آن را خواهیم دید.

۶ آنالیز pathway و gene ontology ها

در این بخش ابتدا می‌خواهیم با توجه به جدولی که در قسمت قبلی بدست آورdim، دو لیست از ژن‌ها ایجاد می‌کنیم. یک لیست حاوی ژن‌هایی که بیان آن‌ها به طرز معناداری افزایش یافته است و لیست دیگر حاوی ژن‌هایی که بیان آن‌ها به طرز معناداری کاهش یافته است. حال از این دو لیست استفاده می‌کنیم و با استفاده از سایت enrichR به تحلیل می‌پردازیم. اینجا با توجه به خواسته خود صورت پروژه تنها به تحلیل بخش‌های Ontlogy و Pathway پرداخته‌ایم. البته در بخش Up به یک نتیجه جالب از Transcription نیز اشاره کردہ‌ایم.

۱.۶ ژن‌هایی که بیان آن‌ها به طرز معناداری کاهاش یافته است

Pathway Analysis ۱.۱.۶

در واقع در آنالیز pathway می‌خواهیم پیدا کنیم که کدام pathway با لیست ژن‌هایی که پیدا کرده‌ایم بیشترین اشتراک را دارد. یعنی ژن‌های درگیر در کدام pathway همانند ژن‌های در لیست ما هستند که افزایش بیان یا کاهاش بیان داشته‌اند. در واقع یک pathway عبارت است از زنجیره اعمال علت و معلولی که منجر به یک عملیات خاص در بدن می‌شود.

Reactome •

نتایجی که از این پایگاه داده مشاهده می‌کنیم طبق انتظار ما از بیماری AML است. مثلا در مورد اول طبق ۱۰ مشاهده می‌شود که pathway سیستم ایمنی مربوط به انسان بسیار درگیر شده است که با توجه به بیماری سرطان خون، نتیجه درستی است چرا که این بیماری مربوط به از بین رفتن سیستم ایمنی می‌باشد.
در نمونه دوم، سایتوکاین، پروتئینی است که عملکرد تنظیمی در سیستم ایمنی بدن دارد و در برابر تحريك ایمنی واکنش نشان می‌دهد. در همین لینک نوشته می‌شود که این pathway خود مربوط به چندین دیگر می‌باشد که همگی در سیستم ایمنی بدن نقش دارند.

| Reactome 2016 | | | | | |
|--|--|-------------|------------------|---------|----------------|
| Hover each row to see the overlapping genes. | | | | | |
| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
| 1 | Immune System_Homo sapiens_R-HSA-168256 | 3.352e-17 | 3.077e-14 | -2.23 | 84.68 |
| 2 | Cytokine Signaling in Immune system_Homo sapiens_R-HSA-1280215 | 1.308e-12 | 4.002e-10 | -2.39 | 65.31 |
| 3 | Interferon alpha/beta signaling_Homo sapiens_R-HSA-909733 | 2.461e-14 | 1.130e-11 | -1.86 | 58.14 |
| 4 | Interferon Signaling_Homo sapiens_R-HSA-913531 | 8.747e-12 | 1.686e-9 | -2.08 | 52.91 |
| 5 | Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell_Homo sapiens_R-HSA-198933 | 9.185e-12 | 1.686e-9 | -1.98 | 50.26 |
| 6 | Adaptive Immune System_Homo sapiens_R-HSA-1280218 | 2.703e-9 | 3.545e-7 | -2.23 | 44.04 |
| 7 | Interferon gamma signaling_Homo sapiens_R-HSA-877300 | 2.514e-10 | 3.846e-8 | -1.74 | 38.54 |
| 8 | Innate Immune System_Homo sapiens_R-HSA-168249 | 2.923e-7 | 0.00002981 | -2.36 | 35.49 |
| 9 | Generation of second messenger molecules_Homo sapiens_R-HSA-202433 | 3.398e-8 | 0.000003899 | -1.94 | 33.35 |
| 10 | Costimulation by the CD28 family_Homo sapiens_R-HSA-388841 | 0.000001037 | 0.00009458 | -1.89 | 26.06 |

شكل ۱۰ : Reactome on down gene list

KEGG •

مورد اول از ۱۱ درباره تمایز سلولی سلول‌های نوع Th است که با توجه به توضیح اینجا

به سیستم ایمنی بدن مرتبط است. همچنین نتیجه مرتبط دیگر با سیستم ایمنی ، مورد دهم است که با توجه به این توضیح این pathway نیز در سیستم ایمنی نقش دارد.

| KEGG 2019 Human | | Bar Graph | Table | Clustergram | | |
|--|--|-------------|------------------|-------------|----------------|--|
| Hover each row to see the overlapping genes. | | | | | | |
| 10 | entries per page | | Search: | | | |
| Index | Name | P-value | Adjusted p-value | Z-score | Combined score | |
| 1 | Th17 cell differentiation | 4.294e-9 | 3.492e-7 | -68.54 | 1320.46 | |
| 2 | Glycosphingolipid biosynthesis | 0.3564 | 0.7764 | -685.31 | 707.06 | |
| 3 | Glycosaminoglycan degradation | 0.2246 | 0.5957 | -460.38 | 687.53 | |
| 4 | Autoimmune thyroid disease | 0.0001577 | 0.001673 | -74.35 | 650.90 | |
| 5 | Graft-versus-host disease | 0.00009852 | 0.001145 | -64.89 | 598.67 | |
| 6 | Natural killer cell mediated cytotoxicity | 0.000003269 | 0.00007252 | -45.34 | 572.75 | |
| 7 | NF-kappa B signaling pathway | 7.740e-8 | 0.000003147 | -34.08 | 558.09 | |
| 8 | Allograft rejection | 0.00005205 | 0.0007055 | -51.50 | 507.96 | |
| 9 | T cell receptor signaling pathway | 1.360e-9 | 1.659e-7 | -22.56 | 460.55 | |
| 10 | Intestinal immune network for IgA production | 0.006850 | 0.04180 | -86.45 | 430.82 | |

شکل ۱۱ : Kegg on down gene list

Gene Ontology Analysis ۲.۱.۶

GO Biological Process

در این بخش نیز نتایج مربوط به سیستم ایمنی است. اینترفرون نوعی از پروتئین است که موجب تحریک سیستم ایمنی و افزایش مقاومت بدن می‌شوند. [۱] در نتیجه باز هم نتایج مربوط به بیماری aml می‌باشد.

| GO Biological Process 2018 | | Bar Graph | Table | Clustergram | | |
|--|--|-------------|------------------|-------------|----------------|--|
| Hover each row to see the overlapping genes. | | | | | | |
| 10 | entries per page | | Search: | | | |
| Index | Name | P-value | Adjusted p-value | Z-score | Combined score | |
| 1 | type I interferon signaling pathway (GO:0060337) | 1.170e-14 | 1.652e-11 | -2.33 | 74.60 | |
| 2 | cellular response to type I interferon (GO:0071357) | 1.170e-14 | 1.652e-11 | -1.40 | 44.75 | |
| 3 | cytokine-mediated signaling pathway (GO:0019221) | 1.961e-10 | 1.474e-7 | -1.34 | 30.04 | |
| 4 | regulation of immune response (GO:0050776) | 2.089e-10 | 1.474e-7 | -1.29 | 28.69 | |
| 5 | interferon-gamma-mediated signaling pathway (GO:0060333) | 3.240e-9 | 0.000001829 | -1.31 | 25.56 | |
| 6 | T cell receptor signaling pathway (GO:0050852) | 5.293e-8 | 0.00002490 | -1.21 | 20.25 | |
| 7 | T cell differentiation (GO:0030217) | 1.337e-7 | 0.00005005 | -1.93 | 30.58 | |
| 8 | T cell activation (GO:0042110) | 1.418e-7 | 0.00005005 | -1.33 | 20.95 | |
| 9 | positive regulation of cytokine production (GO:0001819) | 1.838e-7 | 0.00005764 | -1.95 | 30.24 | |
| 10 | cellular defense response (GO:0006968) | 0.000001838 | 0.0004901 | -1.92 | 25.37 | |

Biological Process on down gene list : ۱۲

GO Molecular Process •

GO Molecular Function 2018

Bar Graph **Table** Clustergram

Hover each row to see the overlapping genes.

| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
|-------|--|-------------|------------------|---------|----------------|
| 1 | GTPase activator activity (GO:0005096) | 0.000002582 | 0.0007063 | -2.04 | 26.26 |
| 2 | GTPase regulator activity (GO:0030695) | 0.000002554 | 0.0007063 | -1.68 | 21.60 |
| 3 | chemokine receptor activity (GO:0004950) | 0.000004165 | 0.0007595 | -2.19 | 27.16 |
| 4 | Ras GTPase binding (GO:0017016) | 0.000006738 | 0.0009215 | -1.77 | 21.07 |
| 5 | cytokine receptor activity (GO:0004896) | 0.00003745 | 0.004097 | -1.82 | 18.52 |
| 6 | G-protein coupled chemoattractant receptor activity (GO:0001637) | 0.00005726 | 0.005221 | -2.54 | 24.86 |
| 7 | SH3/SH2 adaptor activity (GO:0005070) | 0.0001335 | 0.009464 | -1.96 | 17.45 |
| 8 | Rab GTPase binding (GO:0017137) | 0.0001430 | 0.009464 | -1.33 | 11.78 |
| 9 | protein serine/threonine kinase activity (GO:0004674) | 0.0001557 | 0.009464 | -1.23 | 10.81 |
| 10 | C-C chemokine binding (GO:0019957) | 0.0005094 | 0.02786 | -3.37 | 25.59 |

شكل ۱۳: Molecular Process on down gene list

GO Cellular Process •

GO Cellular Component 2018

Bar Graph **Table** Clustergram

Hover each row to see the overlapping genes.

| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
|-------|---|-----------|------------------|---------|----------------|
| 1 | T cell receptor complex (GO:0042101) | 1.898e-8 | 0.000004649 | -2.71 | 48.25 |
| 2 | MHC protein complex (GO:0042611) | 0.0001718 | 0.01403 | -2.09 | 18.15 |
| 3 | early endosome membrane (GO:0031901) | 0.0001644 | 0.01403 | -1.60 | 13.97 |
| 4 | integral component of luminal side of endoplasmic reticulum membrane (GO:0071556) | 0.0003948 | 0.02338 | -1.93 | 15.12 |
| 5 | integral component of plasma membrane (GO:0005887) | 0.0005725 | 0.02338 | -1.62 | 12.10 |
| 6 | phagocytic vesicle (GO:0045335) | 0.0005502 | 0.02338 | -1.46 | 10.94 |
| 7 | specific granule membrane (GO:0035579) | 0.001253 | 0.04387 | -1.80 | 12.03 |
| 8 | phagocytic vesicle membrane (GO:0030670) | 0.002063 | 0.06318 | -1.54 | 9.52 |
| 9 | membrane raft (GO:0045121) | 0.004300 | 0.1054 | -1.31 | 7.16 |
| 10 | Golgi subcompartment (GO:0098791) | 0.004165 | 0.1054 | -1.28 | 7.03 |

شكل ۱۴: Cellular Process on down gene list

۲.۶ ژنهایی که بیان آنها به طرز معناداری افزایش یافته است

Pathway Analysis ۱.۲.۶

نتایج بدست آمده بدین صورت هستند:

Kegg •

با بررسی نتایج نمودار ۱۵ و بررسی ارتباط آنها با بیماری AML در می‌باییم که مثلا نتایج دوم و سوم و چهارم با بیماری AML مرتبطند.

| KEGG 2019 Human | | | | | |
|---|--|-------------|------------------|---------|----------------|
| Bar Graph Table Clustergram ⚙️ ⓘ | | | | | |
| Hover each row to see the overlapping genes. | | | | | |
| 10 | entries per page | | | | |
| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
| 1 | Systemic lupus erythematosus | 2.388e-21 | 7.091e-19 | -12.52 | 594.55 |
| 2 | Cell cycle | 2.072e-14 | 3.077e-12 | -5.05 | 159.10 |
| 3 | Alcoholism | 1.993e-13 | 1.974e-11 | -1.28 | 37.35 |
| 4 | Transcriptional misregulation in cancer | 6.398e-10 | 4.751e-8 | -3.77 | 79.73 |
| 5 | Viral carcinogenesis | 6.719e-8 | 0.000003991 | -1.02 | 16.87 |
| 6 | DNA replication | 5.136e-7 | 0.00002542 | -48.80 | 706.67 |
| 7 | Propanoate metabolism | 0.000006792 | 0.0002882 | -12.02 | 143.02 |
| 8 | p53 signaling pathway | 0.000008071 | 0.0002997 | -7.15 | 83.82 |
| 9 | Valine, leucine and isoleucine degradation | 0.00001929 | 0.0006367 | -24.46 | 265.58 |
| 10 | Malaria | 0.00002458 | 0.0007299 | -19.87 | 210.86 |

شکل ۱۵ : Kegg on up Gene list

Reactome •

نتایج نمودار ۱۶ نیز قابل پیش‌بینی از بیماری سرطان است چرا که روی چرخه سلولی و تقسیم سلولی تاثیر گذاشته است

| Reactome 2016 | | | | | |
|---|--|-----------|------------------|---------|----------------|
| Bar Graph Table Clustergram ⚙️ ⓘ | | | | | |
| Hover each row to see the overlapping genes. | | | | | |
| 10 | entries per page | | | | |
| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
| 1 | Cell Cycle_Homo sapiens_R-HSA-1640170 | 4.782e-50 | 5.461e-47 | -2.46 | 279.49 |
| 2 | Cell Cycle, Mitotic_Homo sapiens_R-HSA-69278 | 2.417e-45 | 1.380e-42 | -2.47 | 253.55 |
| 3 | M Phase_Homo sapiens_R-HSA-68886 | 1.191e-21 | 4.532e-19 | -2.43 | 117.17 |
| 4 | Cell Cycle Checkpoints_Homo sapiens_R-HSA-69620 | 4.851e-21 | 1.385e-18 | -2.34 | 109.60 |
| 5 | G2/M Checkpoints_Homo sapiens_R-HSA-69481 | 1.086e-19 | 2.480e-17 | -2.32 | 101.15 |
| 6 | Mitotic Prometaphase_Homo sapiens_R-HSA-68877 | 2.153e-19 | 4.097e-17 | -2.01 | 86.28 |
| 7 | Mitotic G1-G1/S phases_Homo sapiens_R-HSA-453279 | 2.271e-18 | 3.705e-16 | -2.08 | 84.55 |
| 8 | Chromosome Maintenance_Homo sapiens_R-HSA-73886 | 1.406e-17 | 2.006e-15 | -2.02 | 78.23 |
| 9 | Resolution of Sister Chromatid Cohesion_Homo sapiens_R-HSA-2500257 | 3.877e-17 | 4.920e-15 | -2.02 | 76.38 |
| 10 | G1/S Transition_Homo sapiens_R-HSA-69206 | 7.735e-17 | 8.788e-15 | -2.07 | 76.65 |

شکل ۱۶ : Reactome on up Gene list

Gene Ontology Analysis ۲.۲.۶

GO Biological Process •

همان طور که در شکل ۱۷ مشخص است، فرآیندهایی راجع به سیستم ایمنی بدست آمده است. همچنین عملکردهای پایه‌ای سلول مثل Metabolic Process و Replication درگیر شده‌اند.

GO Biological Process 2018

Bar Graph [Table](#) Clustergram

Hover each row to see the overlapping genes.

50 entries per page

| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
|-------|--|-----------|------------------|---------|----------------|
| 1 | DNA metabolic process (GO:0006259) | 9.732e-19 | 3.624e-15 | -1.36 | 56.56 |
| 2 | G1/S transition of mitotic cell cycle (GO:0000082) | 6.776e-17 | 1.262e-13 | -1.19 | 44.45 |
| 3 | mitotic cell cycle phase transition (GO:0044772) | 8.475e-15 | 1.052e-11 | -1.17 | 37.89 |
| 4 | DNA replication (GO:0006260) | 5.140e-14 | 4.786e-11 | -1.55 | 47.55 |
| 5 | cell cycle G1/S phase transition (GO:0044843) | 9.376e-14 | 6.983e-11 | -1.20 | 36.09 |
| 6 | neutrophil activation involved in immune response (GO:0002283) | 3.508e-11 | 2.177e-8 | -1.25 | 30.02 |
| 7 | neutrophil degranulation (GO:0043312) | 6.009e-11 | 3.197e-8 | -2.01 | 47.25 |
| 8 | neutrophil mediated immunity (GO:0002446) | 1.324e-10 | 5.478e-8 | -1.93 | 43.99 |
| 9 | DNA repair (GO:0006281) | 1.195e-10 | 5.478e-8 | -1.70 | 38.91 |
| 10 | mitotic sister chromatid segregation (GO:0000070) | 2.739e-10 | 1.020e-7 | -1.36 | 30.03 |
| 11 | microtubule cytoskeleton organization involved in mitosis (GO:1902850) | 4.720e-10 | 1.598e-7 | -1.64 | 35.25 |
| 12 | regulation of transcription involved in G1/S transition of mitotic cell cycle (GO:0000083) | 6.067e-10 | 1.883e-7 | -1.91 | 40.48 |
| 13 | centromere complex assembly (GO:0034508) | 1.101e-9 | 3.154e-7 | -1.70 | 35.03 |

شکل ۱۷ Biological Process on up gene list :

GO Molecular Process •

GO Molecular Function 2018

Bar Graph Table Clustergram ⚙ ⓘ

Hover each row to see the overlapping genes.

10 entries per page Search:

| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
|-------|---|-------------|------------------|---------|----------------|
| 1 | ATPase activity (GO:0016887) | 3.229e-8 | 0.00001308 | -2.34 | 40.44 |
| 2 | DNA binding (GO:0003677) | 2.370e-8 | 0.00001308 | -1.20 | 21.09 |
| 3 | single-stranded DNA binding (GO:0003697) | 5.954e-7 | 0.0001608 | -1.94 | 27.77 |
| 4 | DNA-dependent ATPase activity (GO:0008094) | 0.000002259 | 0.0003850 | -1.53 | 19.86 |
| 5 | RNA binding (GO:0003723) | 0.000002377 | 0.0003850 | -1.35 | 17.51 |
| 6 | nucleoside-triphosphatase activity (GO:0017111) | 0.000004698 | 0.0006342 | -1.16 | 14.29 |
| 7 | purine ribonucleoside triphosphate binding (GO:0035639) | 0.000005729 | 0.0006629 | -1.75 | 21.08 |
| 8 | DNA helicase activity (GO:0003678) | 0.00001700 | 0.001721 | -2.63 | 28.92 |
| 9 | double-stranded DNA binding (GO:0003690) | 0.00001922 | 0.001730 | -1.25 | 13.53 |

شکل ۱۸ Molecular Process on up gene list :

GO Cellular Process •

GO Cellular Component 2018

Bar Graph Table Clustergram ⚙ ⓘ

Hover each row to see the overlapping genes.

10 entries per page Search:

| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
|-------|---|------------|------------------|---------|----------------|
| 1 | MutLalpha complex (GO:0032389) | 0.00002017 | 0.0002023 | -4.97 | 53.74 |
| 2 | chromosome, centromeric region (GO:0000775) | 5.099e-11 | 5.456e-9 | -2.25 | 53.30 |
| 3 | nuclear chromosome part (GO:0044454) | 1.854e-14 | 5.951e-12 | -1.25 | 39.52 |
| 4 | Golgi cis cisterna (GO:0000137) | 2.473e-7 | 0.000008822 | -2.49 | 37.88 |
| 5 | chromatin (GO:0000785) | 9.943e-10 | 7.979e-8 | -1.66 | 34.34 |
| 6 | spindle (GO:0005819) | 3.449e-12 | 5.535e-10 | -1.18 | 31.05 |
| 7 | Golgi cisterna membrane (GO:0032580) | 2.473e-7 | 0.000008822 | -2.00 | 30.48 |
| 8 | spindle microtubule (GO:0005876) | 1.748e-7 | 0.000008822 | -1.95 | 30.28 |
| 9 | nuclear chromosome (GO:0000228) | 4.303e-7 | 0.00001151 | -2.06 | 30.26 |
| 10 | condensed chromosome, centromeric region (GO:0000779) | 9.540e-7 | 0.00002187 | -1.98 | 27.49 |

شکل ۱۹ Cellular Process on up gene list :

Transcription ۳.۲.۶

یک نتیجه جالبی که در بخش Transcription مشاهده می شود این است که طبق نمودار ۲۰ با توجه به مقدار اصلاح شده p-value برای ژن E2F4 نمی توانیم آنرا قبول کنیم. با بررسی بیشتر بدست امد که در یک آزمایش توانسته اند نشان بدهند که این ژن اصولاً در بیماری AML نقشی ندارد. [۲]

TRANSFAC and JASPAR PWMs

Bar Graph Table Grid Network Clustergram ⚙ ⓘ

Hover each row to see the overlapping genes.

10 entries per page

Search:

| Index | Name | P-value | Adjusted p-value | Z-score | Combined score |
|-------|---------------|-----------|------------------|---------|----------------|
| 1 | E2F4 (human) | 0.0003699 | 0.1161 | -1.97 | 15.55 |
| 2 | FOXA3 (human) | 0.006058 | 0.9511 | -1.99 | 10.15 |
| 3 | E2F1 (mouse) | 0.01516 | 1.000 | -1.57 | 6.56 |
| 4 | NFYB (human) | 0.05695 | 1.000 | -2.26 | 6.48 |
| 5 | CBEPB (human) | 0.01780 | 1.000 | -1.61 | 6.48 |
| 6 | RARA (human) | 0.05253 | 1.000 | -1.73 | 5.10 |
| 7 | STAT1 (human) | 0.07250 | 1.000 | -1.62 | 4.25 |
| 8 | BCL6 (human) | 0.07167 | 1.000 | -1.61 | 4.24 |
| 9 | PGR (human) | 0.08054 | 1.000 | -1.63 | 4.10 |
| 10 | TBP (human) | 0.06894 | 1.000 | -1.53 | 4.09 |

شکل :۲۰ Transcription :

۷ مباحثات آینده

در این پژوهه با بررسی و تحلیل داده‌های AML ، تفاوت‌های میان GSE۴۸۵۵۸ مربوط به بیماری AML ، دو دسته نرمال و AML Patient را بررسی کردیم و به pathway ها و Gene Ontology های آینده می‌تواند صرف مجموعه این pathway ها و جالبی دست یافتیم. بررسی‌ها و پژوهش‌های آینده می‌تواند صرف مجموعه این pathway ها و جالبی دست یافتیم. بررسی‌ها و پژوهش‌های آینده می‌توانند صرف مجموعه این pathway ها و جالبی دست یافتیم. بررسی‌ها و با بررسی دقیق آنها شناخت دقیق‌تری نسبت به بیماری پیدا شود. همچنین در سمپل‌ها دیدیم که سمپل‌های نرمال CD۳۴ به یک subpopulation از سمپل‌های Gene Ontology شباخت دارند. اما مشکل اینجا بود که تعداد آنها زیاد نبود و تنها دو سمپل CD۳۴ موجود داشتیم. در آزمایش‌های آینده برای حل این مشکل می‌توان با نمونه‌گیری بیشتر از این دسته، دریافت که آیا واقعاً شباهت معناداری میان سمپل‌های CD۳۴ و سمپل‌های سرطانی وجود دارد یا خیر. همچنین در نمودار سه‌بعدی دیدیم که نمونه‌های سرطانی انگار به دو subpopulation ناشناخته تقسیم شده بودند. چالشی که وجود دارد این است که با توجه به ماهیت سلول سرطانی، شناخت و تحلیل دقیق‌تر اینگونه اتفاقات سخت است. به عنوان راه حل، می‌توان چندین مرتبه دیگر نمونه‌گیری را تکرار کرد و مشاهده که آیا دوباره subpopulation ایجاد می‌شود یا خیر. که البته چندین مرتبه نمونه‌گیری نیز خودش به نوعی یک چالش است.

مراجع

- [1] Parkin J, Cohen B An overview of the immune system. *Lancet*. 357 (9270): 1777–89., PMID 11403834, June 2001
- [2] Komatsu, N., Takeuchi, S., Ikezoe, T., Tasaka, T., Hatta, Y., Machida, H., Williamson, I. K., Bartram, C. R., Koeffler, H. P., & Taguchi, H Mutations of

the E2F4 gene in hematological malignancies having microsatellite instability
Blood, 95(4), 1509-1510., (2000)