به نام خدا

محمدرضا منعمیان

شماره دانشجویی : ۴۰۲۱۰۶۶۰۴

Mohammad Reza Monemian

bioinformatic

بخش اول :

۱. با استفاده از دستور dim ابعاد ماتریس را بررسی می‌کنیم :

```
dim(data_matrix)
```

```
28889 · 36
```

عدد اول تعداد سطر ها است که برابر با تعداد کل ژن هایی است که در آزمایش RNA-seq اندازه گیری شده اند که حدود ۲۸ هزار تا است.

عدد دوم تعداد ستون های ماتریس است که تعداد کل نمونه های بیولوژیک مورد مطالعه است که شامل جمعا ۳۶ نمونه بیمار و کنترل است.

---

۲. سطر 5s-rRNA را بررسی کردیم همانطور که مشاهده می‌کنیم در اکثر نمونه ها مقدار بیان آن ژن ها صفر یا بسیار کم است. مقدار آن به صورت زیر است :

```
         CCI031 CCI041 CCI042 CCI050 CCI064 CCI069 CCI077 CCI088 CCI12  CCI136
5S_rRNA       0      0      0      0      0 1.0116      0      0     0 1.15624
         CCI138 CCI58 CCI93 HP1193 HP1194 HP1301 HP851 S013_11 S014_11 S016_11
5S_rRNA       0     0     0      0      0      0     0       0       0       0
         S020_11 S023_11 S024_11 S025_11  S026_11 S027_11 S029_11  S030_11
5S_rRNA 1.25403       0       0       0 0.857024       0       1 0.116938
         S032_11 S033_11 S034_11 S035_11 S038_11 S039_11 S041_11 S042_11
5S_rRNA       0       0       0       1       0       0       0       0
0.177662
```

در آزمایش‌های RNA-seq هدف اصلی مطالعه‌ی **mRNA**(ژن‌های کدکننده پروتئین) است. اما حدود ۸۰ تا ۹۰ درصد RNA کل سلول راRNA های ریبوزومی (rRNA) تشکیل می‌دهند. برای اینکه بودجه توالی‌یابی هدر نرفته و بتوانیم ژن‌های اصلی را ببینیم، دانشمندان در آزمایشگاه با روشی به نام **rRNA Depletion** این ژن‌ها را حذف می‌کنند. مقدار کمی که می‌بینی، باقی‌مانده‌های ناچیزی هستند که در فرآیند حذف باقی مانده‌اند.

۳.

برای بررسی اینکه آیا در ماتریس مقدار NAN و منفی وجود دارد یا نه با استفاده از any این موضوع رو صحت سنجی می‌کنیم :

```
any(is.na(data_matrix))

any(data_matrix < 0)

range(data_matrix)
```

```
FALSE
FALSE
0 · 95726.724
```

همانطور که مشاهده می‌کنیم خروجی ها FALSE است و این یعنی نه مقدار منفی و نه مقدار NAN در ماتریس وجود ندارد.
همانطور که می‌بینیم رنج ماتریس از صفر تا حدود ۹۵ هزار تا است.

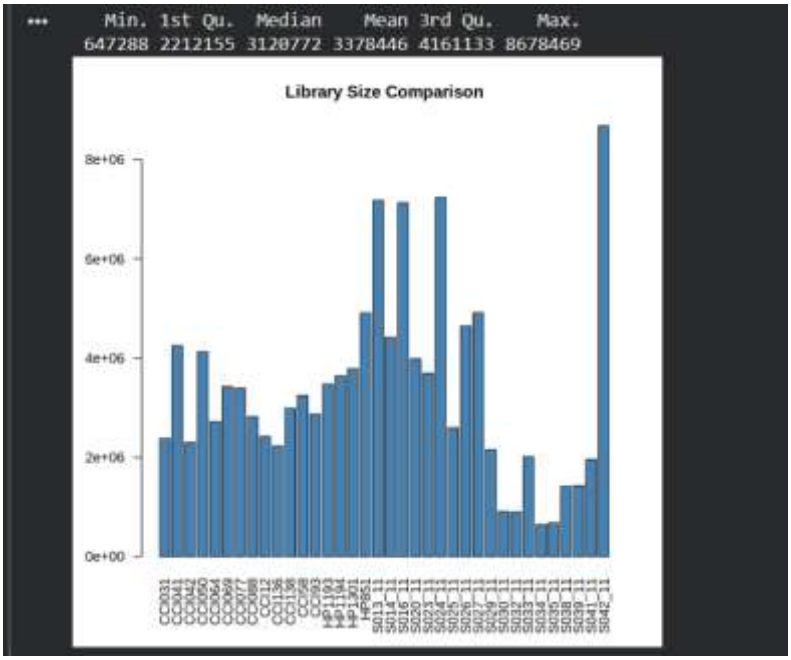
---

۴. اگر داده ها row-count بودند داده ها لزوما باید اعداد صحیح می‌بود اما مشاهده می‌کنیم مقادیر اعشاری نیز در ماتریس وجود دارد پس داده ها row-count نیستند.
وجود اعداد اعشاری و همچنین بازه ای که برای مقادیر ماتریس مشاهده می‌کنیم دلیلی بر این است که داده های ماتریس نرمال سازی شده اند این نرمال سازی برای مشهود بودن تفاوت در عمق library size است. اما داده ها log-transform نشدند به این خاطر که اگر لگاریتم بر آنها اعمال می‌شد مقدار آنها بین ۰ تا حدودا ۲۰ می‌رفت اما همانطور که مشاهده می‌کنیم مقدار داده ها در ماتریس تا حدود ۹۵ هزار تا نیز رفته است و این امکان انجام log-transform را از بین می‌برد.

انتخاب روش تحلیل به این دلیل حیاتی است که ابزارهای بیوانفورماتیکی بر پایه فرض‌های آماری متفاوتی بنا شده‌اند؛ ابزارهایی مانند DESeq2 و edgeR منحصرا به شمارش‌های خام (Raw Counts) یا همان اعداد صحیح نیاز دارند تا بتوانند با استفاده از توزیع «دوجمله‌ای منفی»، نویزهای آماری را مدل‌سازی کنند، در حالی که روش‌هایی مانند limma برای کار با داده‌های نرمال‌شده و لگاریتمی (Log-transformed) طراحی شده‌اند تا تفاوت شدت بیان را در یک مدل خطی بسنجند.

---

۵. با رسم نمودار library size می‌توانیم سوالات پرسیده شده را بررسی کنیم:



همانطور که مشاهده می‌کنیم نمونه S042_11 با اختلاف فاحشی نسبت به بقیه بیشترین مقدار خوانش را در حدود ۸.۶ ملیون دارد در صورتی که ژن هایی مانندS034_11 , S035_11 مقدار خوانشی در حد ۶۰۰ هزار تا دارند. می‌بینیم که اختلاف بین library size ها تا حد زیادی قابل توجه است بطوریکه تفاوت بین کمترین و بیشترین مقدار ۱۳ برابر است.

میانگین خوانش ها حدود ۳ ملیون است اما واریانس بسیار زیادی دارد و این یک عامل مخدوش کننده است و باید نرمال سازی رخ دهد مثلا ژن S042_11 به طور کاذب بیش بیان شده است اما فقط به این خاطر که دستگاه توالی یاب برای این نمونه بیشتر کار کرده است نه به دلیل بیماری. در اصل اگر نمونه‌های گروه Long COVID به طور میانگین Library Size بالاتری نسبت به گروه کنترل داشته باشند، تمام ژن‌های آن‌ها به غلط "بالاتر" نشان داده می‌شوند.

در اصل بررسی library-size قبل از تحلیل باعث می‌شود از خطاهای مثبت کاذب جلوگیری شود. تفاوت در Library Size معمولا ریشه در مسائل فنی آزمایشگاه مانند غلظت اولیه RNA یا کارایی دستگاه توالی‌یاب دارد و نه تفاوت بیولوژیک بین بیمار و سالم. برای مثال، در داده‌های ما نمونه S042_11 بیش از ۸.۶ میلیون خوانش دارد در حالی که نمونه‌های دیگر زیر ۱ میلیون هستند. اگر این تفاوت اصلاح نشود، تمام ژن‌های نمونه S042_11 به اشتباه بیش‌بیان (Up-regulated) به نظر می‌رسند.

---

۶. نخست، وجود تعداد قابل توجهی ژن با بیان **صفر مطلق** در تمامی ۳۶ نمونه یا ژن‌هایی با بیان بسیار ناچیز (مانند وضعیت ژن 5S-Rrna در اکثر نمونه‌ها) که فاقد ارزش اطلاعاتی برای تحلیل تفاضلی هستند. دوم، ناهماهنگی شدید در Library Size نمونه هاست که در آن نمونه‌ای مانند S042_11 با حدود ۸.۶ میلیون خوانش، بیش از ۱۳ برابر بزرگتر از نمونه‌های کم‌عمق (زیر ۱ میلیون) است و می‌تواند به عنوان یک **Outlier** پتانسیل، نتایج را به نفع خود منحرف کند. در نهایت، تضاد میان اعشاری بودن داده‌ها (که نشانه نرمال‌سازی است) و بازه گسترده مقادیر (از ۰ تا بیش از ۹۵ هزار) نشان می‌دهد که نرمال‌سازی انجام شده احتمالاً از نوع درون‌نمونه‌ای) بوده و نتوانسته است نوسانات شدید فنی بین نمونه‌ها را که در نمودار مشهود است خنثی کند.

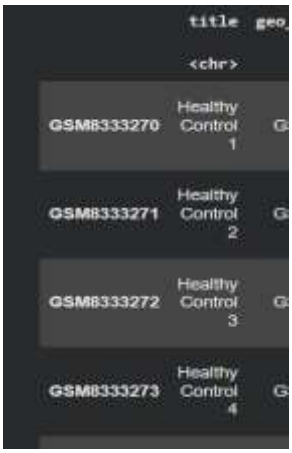
پاسخ به سوالات بخش اول :

سوال اول : در پایگاه داده GEO دو نوع فایل series-matrix , LC_count وجود دارد. فایل series-matrix حاوی مقادیر نهایی بیان ژن ها است که توسط محقق اصلی پژوهش شده است. اما مشکل اینجاست که ممکن است این مقادیر از قبل نرمالایز یا ترنسفورم شده باشند به روشی که با استاندارد های فعلی ما قابل تطابق نباشد اما در فایل LC_count حاوی داده های اولیه تر است حتی در حالت های ایده آل شامل داده های خام row-count است در اصل ما برای یک تحلیل آماری نیاز داریم تا کنترل کاملی بر روی روش نرمال سازی و فیلترینگ داشته باشیم.

سوالات ۲ تا ۴ پاسخ به پرسش ها در سوالات ۱ تا ۵ قسمت قبلی به تفضیل پاسخ داده شده اند پس به توضیح مجدد آن نمی‌پردازیم.

بخش دوم :

۱. وقتی با استفاده از تابع PData جدول اطلاعات نمونه ها را استخراج کردیم و به مشاهده ستون های آن پرداختیم مشاهده کردیم در ستون title که نوع داد های آن به صورت کاراکتر بود داده ها یا Healthy-control و یا Long-covid هستند و این موضوع دقیقا مربوط به دسته بندی نمونه ها به کنترل و یا بیمار است .



۲. با استفاده از table(group) تعداد نمونه ها در هر گروه را به دست آوردیم که ۱۷ تا از نمونه ها برای گروه کنترل و ۱۹ تا مربوط به گروه بیمار است. تعادل بین نمونه ها نیز نسبتا مناسب و حدود ۵۰ ۵۰ است که باعث می‌شود قدرت آماری آزمون‌های شما برای مقایسه دو گروه در بالاترین حد خود باشد و نتایج به نفع یک گروه خاص سنگینی نکند.

```
# Display the total count of samples in each group to verify the split
table(sample_info$group)
```

```
Control  ME_CFS
     17      19
```

۳. مطابقت دقیق نام ستون‌های ماتریس بیان (counts) با شناسه‌های موجود در metadata از این جهت حیاتی است که ابزارهای آماری مانند DESeq2 یا limma فرض می‌کنند ترتیب نمونه‌ها در هر دو فایل کاملاً یکسان است. اگر این همخوانی وجود نداشته باشد یا ترتیب آن‌ها به‌هم‌ریخته باشد، اطلاعات فنوتیپی (مثلاً وضعیت سالم یا بیمار) به نمونه‌های اشتباهی نسبت داده می‌شود؛ در نتیجه، تحلیل تفاضلی بیان ژن مقایسه‌ای بی‌معنی بین نمونه‌های نامرتبط انجام داده و خروجی‌هایی کاملاً غلط و فاقد اعتبار علمی تولید می‌کند که می‌تواند منجر به شناسایی اشتباه ژن‌های شاخص بیماری شود.

۴. تطبیق دقیق نام ستون‌های ماتریس بیان ژن با شناسه‌های موجود در متادیتا از آن جهت حیاتی است که ابزارهای آماری مانند DESeq2 یا limma فرض می‌کنند ترتیب نمونه‌ها در هر دو فایل کاملاً یکسان است. اگر این همخوانی به‌هم بخورد یا نمونه‌های بیمار به‌اشتباه در گروه Control قرار گیرند، واریانس درون‌گروهی به‌شدت افزایش یافته و توان آماری برای شناسایی تغییرات واقعی کاهش می‌یابد که این امر منجر به تولید نتایج مثبت و منفی کاذب (False Positives/Negatives) می‌شود. در نهایت، کل خروجی تحلیل تفاضلی بیان ژن فاقد اعتبار علمی شده و ژن‌های معرفی‌شده به‌عنوان شاخص بیماری، در واقع ناشی از خطای دسته‌بندی خواهند بود نه بیولوژی واقعی بیماری.

۵. انتخاب گروه کنترل به عنوان سطح مرجع (Reference Level) از آن جهت اهمیت دارد که مقایسه آماری و تفسیر مقادیر logFC بر پایه آن انجام می‌شود؛ در واقع، تمام تغییرات بیان ژن نسبت به وضعیت پایه یا «سالم» سنجیده می‌شوند. با این انتخاب، اگر مقدار logFC برای یک ژن **مثبت** باشد، به این معناست که بیان آن ژن در بیماران نسبت به افراد سالم افزایش (Up-regulated) یافته و اگر **منفی** باشد، نشان‌دهنده **کاهش بیان** (Down-regulated) در بیماران است. این کار باعث استانداردسازی گزارش‌ها شده و از سردرگمی در تفسیر جهت تغییرات بیولوژیکی جلوگیری می‌کند، چرا که بدون تعیین مرجع، جهت تغییرات می‌تواند به صورت معکوس گزارش شود و نتیجه‌گیری‌های بالینی را تحت تأثیر قرار دهد.

اگر سطح مرجع را بر روی گروه بیمار قرار دهید، فرمول ریاضی تغییر نمی‌کند اما تفسیر نتایج دقیقاً معکوس می‌شود؛ یعنی ژن‌هایی که در واقعیت در بیماران افزایش بیان دارند، با عدد logFC منفی نمایش داده می‌شوند. این کار باعث سردرگمی در تحلیل مسیرهای زیستی می‌شود، زیرا در گزارش‌های استاندارد علمی، همیشه تغییرات «بیمار» نسبت به حالت «سالم» سنجیده می‌شود تا اعداد مثبت به معنای افزایش بیان در بیماری باشد.

بخش سوم :

۱. در زیر خلاصه ای از نتایج به دست آمده از میانگین میانه و ... را مشاهده می‌کنیم:

```
================================================
   GENE EXPRESSION MATRIX REPORT
================================================

--- Summary Statistics ---
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    0.000     0.000     4.622   116.946    60.687 95726.724

--- Data Integrity Checks ---
Any Missing Values (NAs)?:   FALSE
Any Negative Values?:        FALSE
Value Range (Min to Max):    0  to  95726.72

--- Conclusion on Data Type ---
Result: Data appears to be NORMALIZED or LOG-TRANSFORMED.
================================================
```

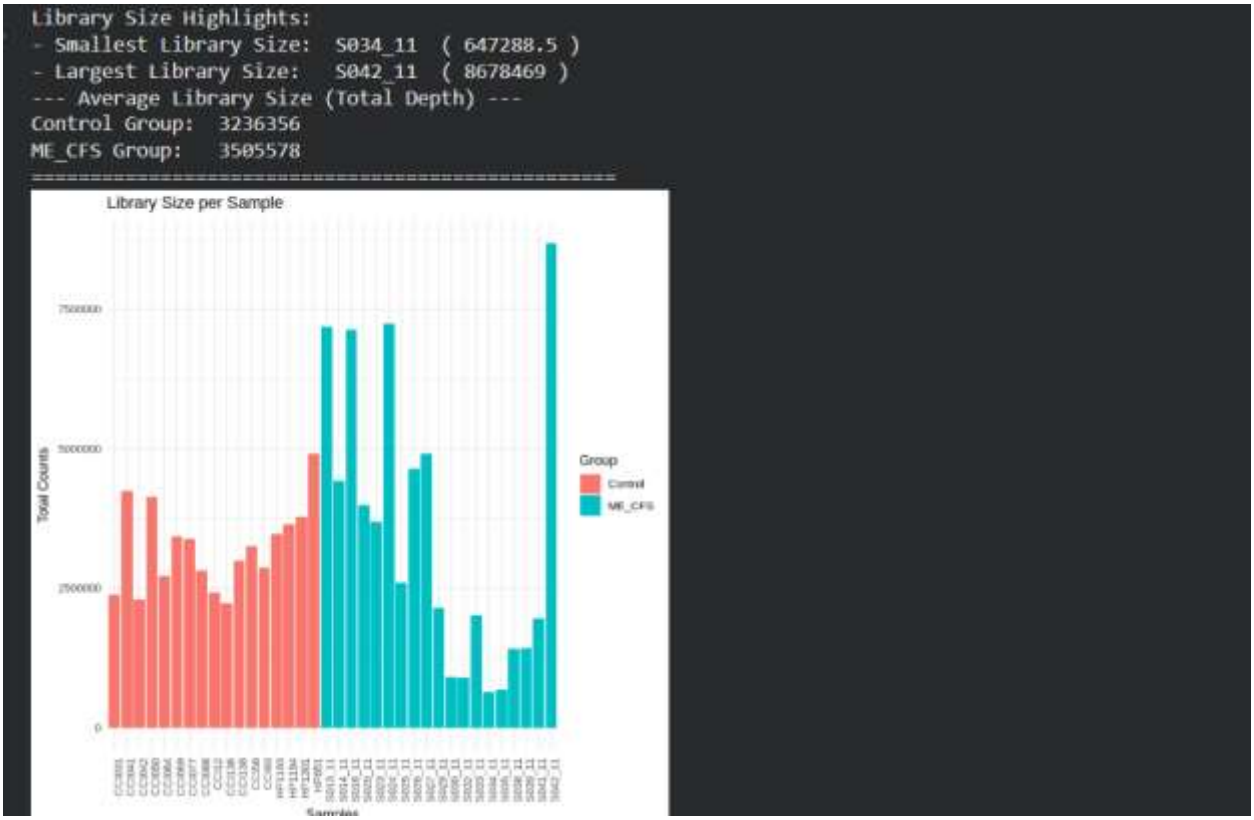
نبودن مقادیر منفی و همچنین وجود اعداد اعشاری نشان می‌دهد که داد های ما خام نیستند بلکه نرمالایز شدند البته از آنها لگاریتم گرفته نشده است به این خاطر که مقدار librarySize ها تا چندین ملیون نیز می‌رود. اما اینکه داده های ما فقط عدد صحیح نیستند نشان از این می‌دهد که داده ها فقط شمارش خام نیستند.

۲. در خروجی کدی که زدیم مشاهده می‌کنیم :

کمترین بیان متعلق به ژن S034-11 با حدودا مقدار ۶۴۷ هزار تا و بیشترین بیان ژن مربوط به S042-11 با حدود ۸ ملیون و ۶۰۰ هزار تا بیان است.

همانطور که در نمودار مشاهده می‌کنیم به جز نمونه آخر که مربوط به گروه بیمار است و بیان بسیار زیادی در حد ۸.۶ ملیون دارد در بقیه موارد تفاوت چشمگیری بین بیان ژن در گروه کنترل و بیمار نیست.

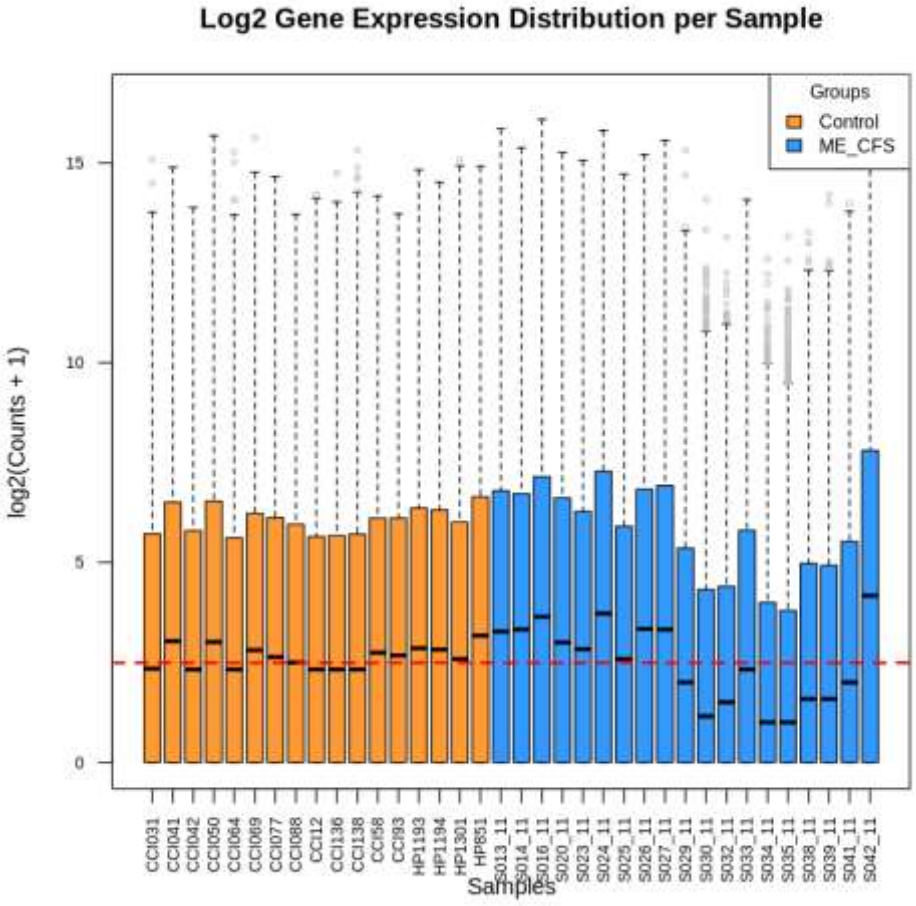
اختلاف زیاد باعث می‌شود ژن‌های یک نمونه فقط به دلیل اینکه بیشتر توالی‌یابی شده‌اند، "پر بیان" به نظر برسند، نه به دلیل بیولوژی واقعی. این موضوع باعث ایجاد نویز و نتایج غلط در تحلیل تفاضلی بیان می‌شود و حتماً باید قبل از تحلیل با روش‌های نرمال‌سازی (Normalization) اصلاح شود.



۳. همان‌طور که در نمودار مشخص است، نمونه‌های گروه Control توزیع نسبتاً یکنواخت و میانه‌های نزدیک به هم دارند، اما در گروه ME_CFS، پراکندگی بسیار زیاد است. برخی نمونه‌ها مانند S030_11، S034_11 و S035_11 بسیار پایین‌تر از بقیه قرار گرفته‌اند و میانه‌های آن‌ها حتی به عدد ۱ نزدیک شده

است بنابراین می‌توان گفت شکل کلی توزیع ها به هم شباهت زیادی ندارد و برخی نمونه ها گسترده متفاوتی دارند..

دلیل اصلی این تفاوت‌ها در این مرحله، تفاوت‌های بیولوژیکی نیست، بلکه عمدتاً ناشی از عوامل فنی است . هم‌ترین دلیل این است که برخی نمونه‌ها (به‌ویژه در سمت راست نمودار) عمق توالی‌یابی بسیار کمتری داشته‌اند. وقتی مجموع خوانش‌های یک نمونه کم باشد، کل توزیع بیان ژن‌های آن به سمت پایین کشیده می‌شود نمونه‌هایی که میانه‌های بسیار پایینی دارند، احتمالاً در مرحله استخراج RNA یا آماده‌سازی کتابخانه دچار افت کیفیت شده‌اند و خوانش‌های کمتری از آن‌ها به دست آمده است.



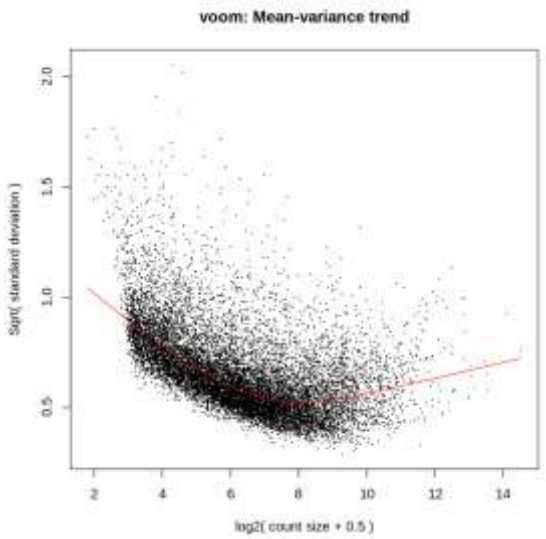
یک نکته را توجه کنیم که دایره هایی که خیلی با جعبه ها فاصله دارند داده های پرت هستند. این نقاط نمایانگر ژن‌هایی با سطح بیان بسیار بالا یا بسیار پایین در مقایسه با سایر ژن‌های همان نمونه می‌باشند که خارج از محدوده چارک‌ها قرار گرفته‌اند.

۴. نمودار voom در اصل چه چیزی را محاسبه می‌کند ابتدا کل داد های برای هر ژن را برای هر نمونه جمع آوری کرده و همچنین آن داده هایی که مقدار بیانشان بسیار کم است را حذف می‌کند. حالا برای هر ژن مقدار لگاریتم شده را میانگین و همچنین واریانس اش را محاسبه می‌کند و در یک نمودار نقطه ای به این صورت که محور افقی میانگین بیان لگاریتم ژن و محور عمودی واریانس آن است رسم می‌کند.

در مورد منحنی قرمزی که رسم شده است نیز باید گفت این منحنی بر اساس داده ها به دست آمده است یعنی بین ژن هایی که میانگین نزدیک بهم دارند می‌بیند که میانگین واریانس آنها چقدر است در اصل روی نقلط سیاه یک رگرسیون لاسو انجام می‌دهد. حالا اگر نقطه ای پایین این منحنی قرمز باشد یعنی واریانس آن ژن از چیزی که ما امتظار داشتیم کمتر است و این چیز خوبی است و اگر بالاتر باشد یعنی واریانسش از چیزی که ما انتظار داشتیم بسیار بالاتر است.

ما با استفاده از چیزی که voom به ما می‌دهد به ژن ها وزن دهی می‌کنیم. در اصل مشاهده نمودار **voom** برای limma به این دلیل حیاتی است که مدل‌های خطی limma فرض می‌کنند واریانس در تمامی داده‌ها مستقل از میانگین است، در حالی که در RNA-seq ، ژن‌های کم‌بیان همیشه نویز (واریانس) بیشتری نسبت به ژن‌های پربیان دارند. voom با محاسبه این رابطه و تبدیل آن به وزن‌های آماری(Precision Weights) ، به limmaمی‌گوید که به کدام ژن‌ها بیشتر و به کدام‌ها کمتر اعتماد کند؛ این کار باعث می‌شود که اثر نویزهای ناشی از قدرت بیان پایین حذف شده و توان آماری مدل برای شناسایی ژن‌های با بیان متفاوت (DEGs) به شکل چشم‌گیری افزایش یابد.

نمودار voom خروجی به صورت زیر است :

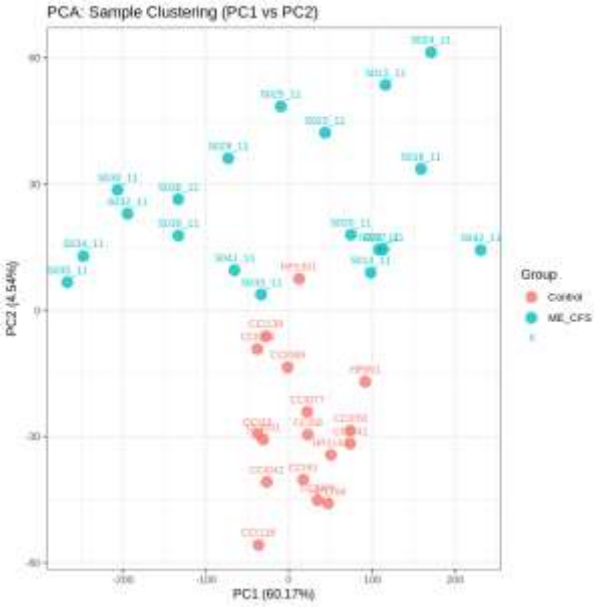


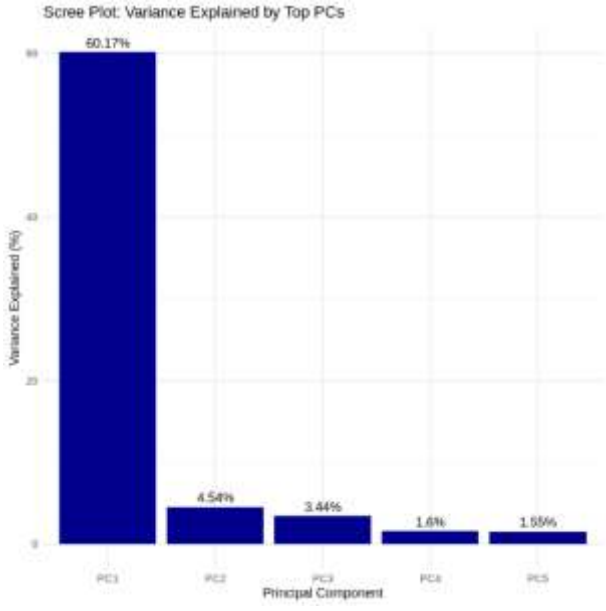
۵.ابزار DESeq2 بر پایه ی توزیع دو جمله ای منفی طراحی شده اند. این توزیع مخصوص داده های شمارشی است که حتما اعدادی صحیح و غیر اعشاری هستند. اما مشاهده می‌کنیم که داده های اولیه ما نرمالایز شدند و اعشاری هستند. برخلاف DESeq2 که از مدل‌های پیچیده غیرخطی استفاده می‌کند و برای آن داده های صحیح و غیر اعشاری نیاز هست مدل limma از مدل‌های خطی (Linear Models) بهره می‌برد. مدل‌های خطی هیچ مشکلی با اعداد اعشاری ندارند و فرض می‌کنند توزیع داده‌ها (پس از وزن‌دهی) به توزیع نرمال نزدیک است. همچنین voom با استفاده از منحنی قرمزی که رسم شد به ژن ها وزندهی می‌کند تا ناهماهنگی بایاس ورایانس را جبران کند.

بخش چهارم :

۱. همانطور که مشاهده می‌کنیم مولفه اول یا pc1 حدود ۶۰ درصد از واریانس کل داده را توضیح می‌دهد و این واضحا به معنای روند غالب بر داده هاست چراکه بیش از نیمی از کل تفاوت‌های موجود در بیان ژنِ تمامی نمونه‌ها تنها توسط همین یک محور توجیه می‌شود.

در زیر نتیجه الگوریتم pca را به صورت بصری مشاهده می‌کنیم :





همچنین نتایج مقدار واریانس هر مولفه به صورت زیر است:

```
--- PCA Variance Summary ---
  Component Variance_Percentage
1       PC1               60.17
2       PC2                4.54
3       PC3                3.44
4       PC4                1.60
5       PC5                1.55
----------------------------
PC1 explains: 60.17%
PC2 explains: 4.54%
PC3 explains: 3.44%

--- Preparing PCA Plot (PC1 vs PC2) ---
Plotting PC1 (60.17%) against PC2 (4.54%) ...

--- PCA Explained Variance (Scree Plot Data) ---
PC1: 60.17% of total variance explained
PC2: 4.54% of total variance explained
PC3: 3.44% of total variance explained
PC4: 1.6% of total variance explained
PC5: 1.55% of total variance explained
```

PCA: Sample Clustering (PC1 vs PC2)

۲.همانطور که در نمودار دو تایی pc2 , pc1 مشاهده می‌کنیم دو گروه کنترل و بیمار بر اساس مولفه دوم یعنی pc2 به وضوح از هم جدا شده اند. این موضوع از نظر زیستی به این معناست که یک **تفاوت پروفایل بیان ژن معنادار و سیستماتیک** بین افراد بیمار و سالم وجود دارد؛ به طوری که می‌توان بر اساس داده‌هایRNA-seq ، این دو گروه را با دقت بالایی از هم تفکیک کرد.

وقتی می‌بینید ۶۰٪ تنوع (PC1) باعث جدا شدن بیماری نشده و فقط نمونه‌ها را در عرض نمودار پخش کرده است، یعنی:

- یک عامل غیر از بیماری (مثل سن، جنسیت، تفاوت‌های ژنتیکی فردی یا حتی نویزهای تکنیکال در روز انجام آزمایش) وجود دارد که تنوع بسیار بیشتری نسبت به خودِ بیماری ایجاد کرده است.
- اما چون نمونه‌های سالم و بیمار هر دو در طول PC1 پخش شده‌اند، این عامل بزرگ (۶۰٪) تأثیری در توانایی ما برای تشخیص بیماری ندارد، چون بیماری خودش را به طور مستقل روی PC2 نشان داده است.

۲. همانطور که مشاهده می‌کنیم در گروه کنترل نمونه هایCCL136, HP1301 به صورت outlier قرار دارند و خارج از محل متمرکز شدن بقیه نمونه ها هستند. همچنین در نمونه های گروه بیمار نیز داده های S034_11,S035_11 , S030_11 از یک طرف و از طرف دیگر نمونه S042_11 خارج از محدوده بیان بقیه نمونه ها هستند. این موضوع به دلایل مختلفی است:

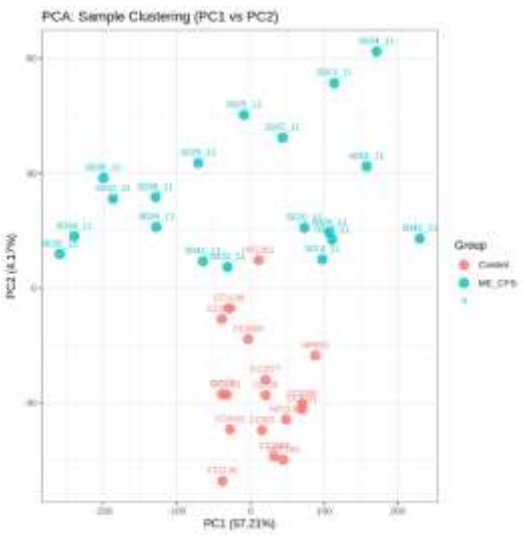
- ❖ **تفاوت بیولوژیک واقعی:** شدت بیماری یا پاسخ فیزیولوژیک در این افراد خاص بسیار شدیدتر یا متفاوت از سایر هم‌گروهی‌ها است.
- ❖ **خطای تکنیکی** : بروز مشکل در مراحل آماده‌سازی کتابخانه (Library Prep) یا توالی‌یابی آن نمونه خاص.
- ❖ **کیفیت پایین نمونه :** پایین بودن یکپارچگی RNA (RIN score) در آن نمونه که باعث ایجاد نویز در داده‌های نهایی شده است.

۳.واریانس کم و نزدیک به صفر یعنی بین همه نمونه ها آن ژن مقدار تقریبا مشابهی داشته است. حذف این ژن ها می‌تواند به دلایل زیر مناسب باشد :

- ❖ **کاهش نویز Noise Reduction** : ژن‌هایی که تغییر بیان ندارند، اطلاعات مفیدی برای تفکیک گروه‌ها ارائه نمی‌دهند و فقط باعث اضافه شدن نویز محاسباتی به مدل می‌شوند.
- ❖ **تمرکز بر متغیرهای کلیدی PCA :** به دنبال یافتن جهت‌هایی است که بیشترین تفاوت (Variance) را در داده‌ها نشان می‌دهند. با حذف ژن‌های ایستا، الگوریتم بر روی ژن‌هایی تمرکز می‌کند که واقعاً بین نمونه‌ها متفاوت هستند (مثلاً تحت تأثیر بیماری تغییر کرده‌اند).
- ❖ **بهبود کارایی محاسباتی :**کاهش تعداد متغیرها (ژن‌ها) سرعت اجرای الگوریتم `prcomp` را افزایش داده و از پیچیدگی بی‌مورد مدل جلوگیری می‌کند.

---

۵. مقدار گفته شده را برابر با true گذاشتیم نتیجه به صورت زیر در آمد :



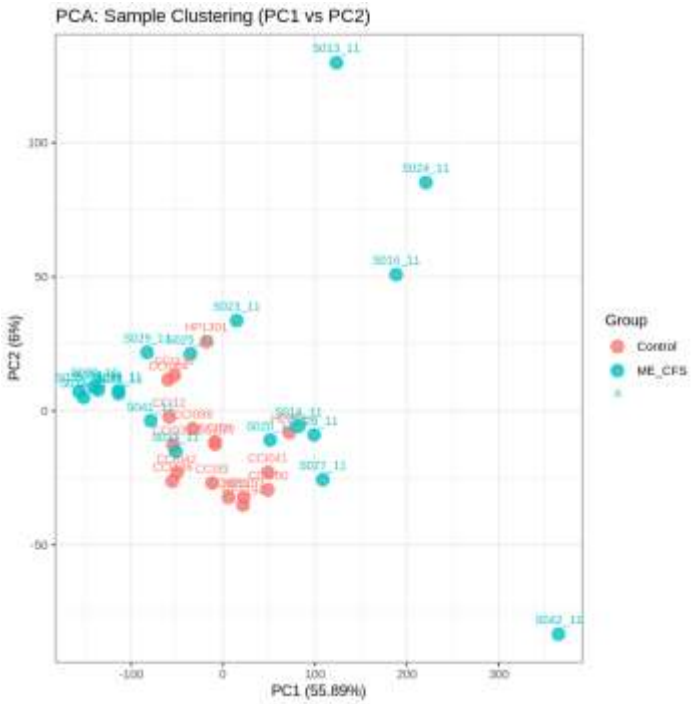
همانطور که مشاهده می‌کنیم مقداری از مولفه اول کم شد. اما دلیل این موضوع چیست ؟

ژن‌هایی که بیان بسیار کمی دارند و تغییرات آن‌ها صرفاً ناشی از نویز تکنیکال است، وزنی برابر با ژن‌های حیاتی پیدا می‌کنند. این کار باعث می‌شود نویز در نمودار PCA غالب شود که این باعث کاهش جدایش گروه ها می‌شود.در اصل استفاده از scale = false به دلایل زیر ترجیح داده می‌شود :

حفظ اهمیت بیولوژیکی :ژن‌هایی که واریانس بالاتری دارند، معمولاً همان ژن‌هایی هستند که تحت تأثیر شرایط آزمایش بیماری در مقابل کنترل تغییر بیان داده‌اند. اگر ما داده‌ها را استاندارد (scale=TRUE) کنیم، واریانس همه ژن‌ها را برابر با ۱ قرار می‌دهیم. این کار باعث می‌شود ژن‌های کلیدی که تفاوت اصلی را ایجاد می‌کنند، هم‌سطح با ژن‌های کم‌اثر و نویزها قرار بگیرند و سیگنال بیولوژیکی ضعیف شود.

ماهیت داده‌های لگاریتمی :تبدیل log_2 تا حد زیادی مشکل ناهمگنی واریانس را حل کرده و داده‌ها را به توزیع نرمال نزدیک می‌کند. در این حالت، تفاوت در مقیاس بیان ژن‌ها یک "ویژگی" است، نه یک "خطا" که نیاز به حذف داشته باشد.

۶.ما این کار را کردیم و با استفاده از داده های خام الگوریتم را ران کردیم و به نتیجه زیر رسیدیم :



مشاهده می‌کنیم اولا چند ژن که بیان بسیار بالایی دارند(Outliers) ، تمام واریانس را به خود اختصاص دادند. در نتیجه، PCAفقط تفاوت‌های آن چند ژن را نشان داده و تغییرات هزاران ژن دیگر که برای تشخیص بیماری ME_CFS مهم هستند، نادیده گرفته شده است. همچنین اکثر نمونه‌ها در یک گوشه نمودار به صورت فشرده جمع شدند و تفکیک زیبایی که اکنون بین گروه‌های Control و ME_CFSمی‌بینیم, کاملاً از بین رفته است . به طور خلاصه به دلیل ماهیت داده‌های شمارشی، واریانس با میانگین رابطه مستقیم دارد. بدون لگاریتم، نمودار تحت تأثیر شدید نویزِ ژن‌های پربیان قرار می‌گرفت و تفسیر بیولوژیکی غیرممکن می‌شود.

بخش پنجم :

۱.تابع filterbyexpr بر اساس فرمولی که در ادامه تعریف می‌کنیم ژن هایی که بیان کمی دارند را حذف می‌کند:

ابتدا می‌آید و به محاسبه CMP بر اساس فرمول زیر می‌پردازد:

$$cpm_i = \frac{c_i}{L_i}$$

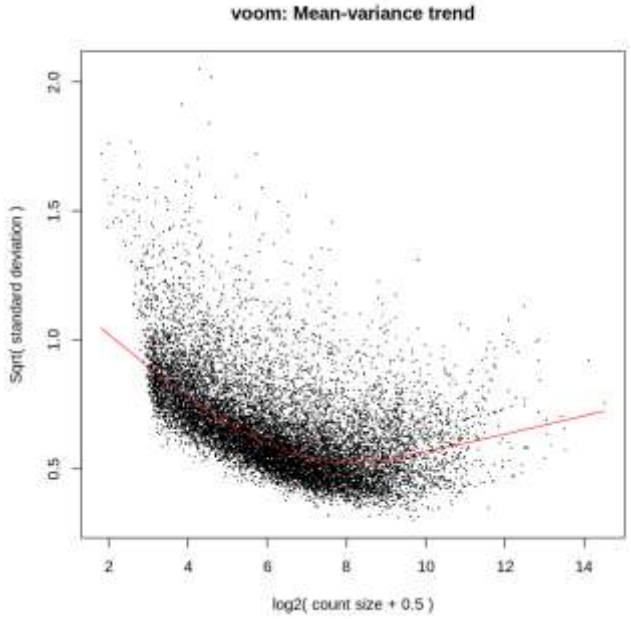
که در آن صورت تعداد خوانش ژن در یک نمونه و مخرج تعداد کل خوانش های آن ژن در کل نمونه ها است. در ادامه یک حد آستانه ای محاسبه می‌کنیم که نمونه های با اندازه متوسط در حد ۱۰ تا ۱۵ خوانش داشته باشند حال اگر یک ژن CMP بیشتری از آن مقدار آستانه در حداقل n نمونه (که این بر اساس تعداد نمونه ها در کوچکترین گروه آزمایشی است) باشد آن ژن نگه داشته می‌شود و در غیر اینصورت حذف خواهد شد.

اما این موضوع چه اهمیتی برای تحلیل های DE دارد؟ توجه داریم اگر یک ژن در کل بیان کمی داشته باشد در اصل به جای اضافه کردن داده نویز اضافه می‌کند و در اصل اطلاعات بیولوژیکی مفیدی به مدل اضافه نمی‌کنند. همچنین حذف این ژن‌ها باعث می‌شود که آزمون‌های فرض مکرر کمتری انجام شود که در نتیجه آن، دقت adj.P.Val برای ژن‌های مهم افزایش می‌یابد.

۲.به نظر در سوال ۴ قسمت ۳ مفصلا به این موضوع پرداختیم اما باز بیان می‌کنیم :

نمودار voom با هدف مدیریت نویزهای آماری در داده‌هایRNA-seq ، ابتدا داده‌های نرمال‌شده را به مقیاس log-CPM تبدیل کرده و رابطه میان میانگین بیان ژن‌ها (محور افقی) و واریانس آن‌ها (محور عمودی) را مدل‌سازی می‌کند. از آنجایی که در توالی‌یابی نسل جدید، ژن‌های کم‌بیان به طور طبیعی واریانس و

نویز بیشتری نشان می‌دهند، voom یک منحنی رگرسیون (Lowess) روی نقاط رسم کرده تا واریانس انتظاری را در هر سطح از بیان مشخص کند. در نهایت، این ابزار بر اساس فاصله هر ژن از منحنی، «وزن‌های آماری Precision Weights محاسبه می‌کند تا به مدل‌های خطی limma(که به طور پیش‌فرض واریانس را ثابت فرض می‌کنند) بفهماند که به ژن‌های دقیق‌تر، وزن بیشتر و به ژن‌های نویزی، وزن کمتری اختصاص دهند؛ این فرآیند باعث حذف اثر نویز فنی و افزایش چشم‌گیر توان آماری برای شناسایی دقیق ژن‌های با بیان متفاوت (DEGs) می‌شود.



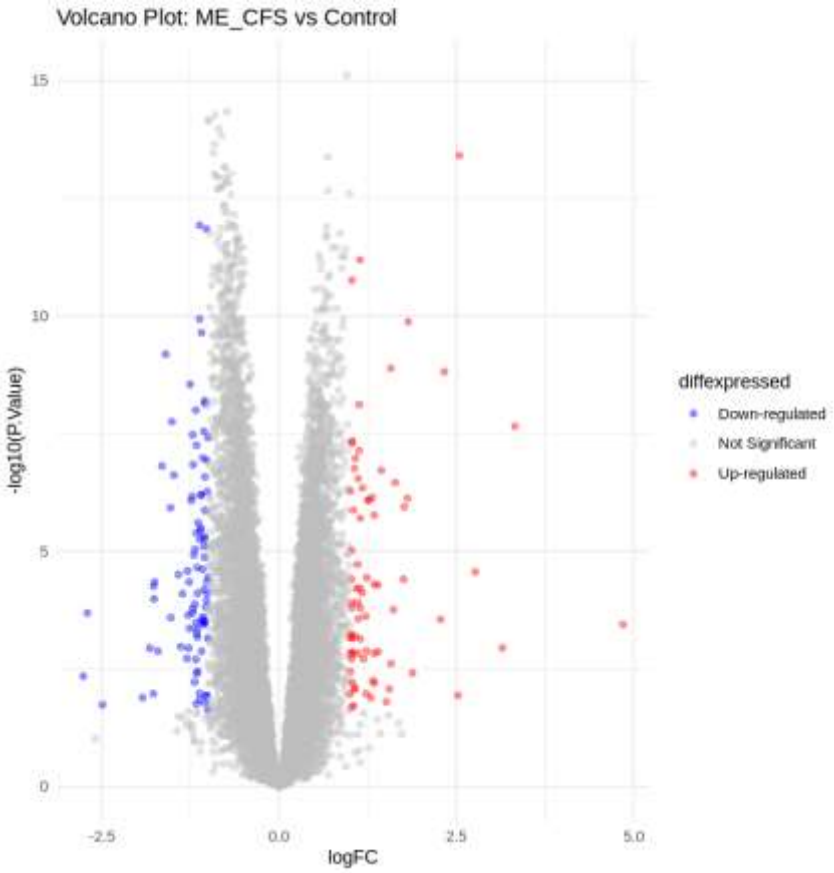
۳. توجه داریم که نمونه های کنترل را رفرنس گرفتیم بنابراین اگر log2fc به دست آمده برای یک ژن بیشتر باشد این بدان معناست که بیان آن ژن در نمونه بیمار بیشتر بوده است و اگر منفی باشد به معنای بیان کمتر ژن مورد نظر در نمونه بیمار است. همچنین توجه داریم هر چقدر مقدار p-value کمتر باشد بهتر است. با توجه به این موضوعات داریم :

```
---  Top 3 DEGs ---
1. Gene: TAF1D | logFC: 0.958 | adj.P.Val: 1.02e-11 | Higher expression in: ME_CFS (Up-regulated)
2. Gene: ZBTB7A | logFC: -0.733 | adj.P.Val: 1.87e-11 | Higher expression in: Control (Down-regulated)
3. Gene: MAP1S | logFC: -0.902 | adj.P.Val: 1.87e-11 | Higher expression in: Control (Down-regulated)
```

مشاهده می‌کنیم ژن TAF1D در نمونه های بیمار بیان بیشتری دارد و دو ژن بعدی از نظر رتبه بندی p-value یعنی ZBTB7A و MAP1S در نمونه های بیمار بیان کمتری دارد.

۴.در نمودار volcano در نظر گرفتیم ژن هایی که adj-p-value کمتر از مقدار ۰.۰۵ با log2fc بیشتر از ۱ دارا باشد را بیش بیانی و اگر log2fc کمتر از -۱ داشته باشد را کم بیانی در نظر بگیرد. همانطور که مشاهده می‌کنیم ژن هایی که در سمت راست نمودار با رنگ آبی مشخص شدند بیش بیانی و آنهایی که در سمت چپ نمودار با رنگ قرمز مشخص شدند را کم بیانی در نظر گرفتیم.

مشاهده می‌شود که تفاوت بیان به صورت متقارن توزیع شده است، اما با نگاه دقیق به تراکم نقاط، تعداد ژن‌های کاهش‌یافته (آبی) در این مطالعه کمی بیشتر از ژن‌های افزایش‌یافته به نظر می‌رسد.



همچنین به طور دقیق تر داریم :

```
Down-regulated Not Significant    Up-regulated
            93           12984              77
```

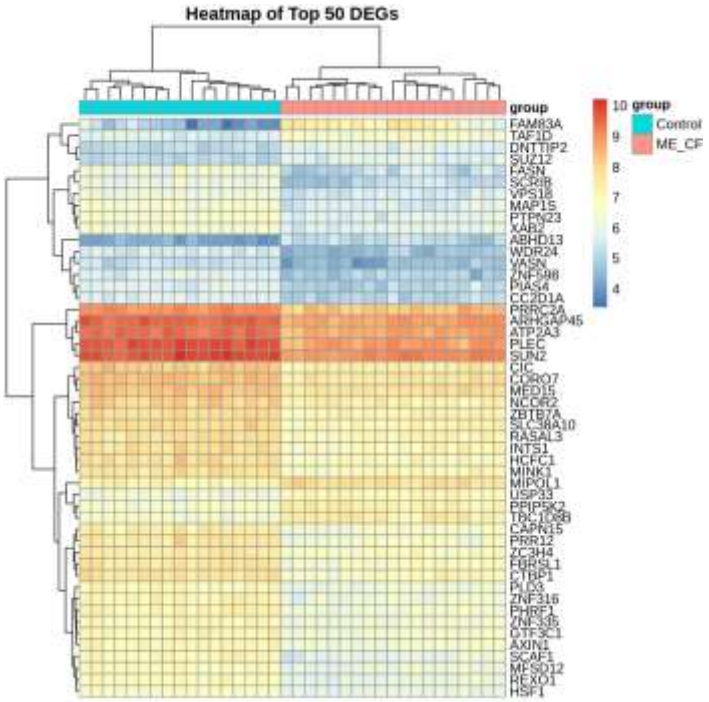
۵.اولا توجه داریم این دو عدد نشان دهنده یک اختلاف بیولوژیکی قوی در یک ژن بین گروه کنترل و بیمار است. چون هم اطمینان زیادی با توجه adj-p-value داریم هم تفاوت زیادی در بیان بین دو گروه است.

با توجه به عکسی هم که در بالا مشاهده کردیم در ۹۳ ژن کم بیانی و در ۷۷ ژن بیش بیانی مشاهده می‌شود که یعنی در کل در ۱۷۰ ژن اعداد گفته شده در سوال صادق است.

۶.بله در دو خوشه قرار گرفته است به این خاطر که در بعضی از ژن ها مشاهده می‌کنیم دقیقا برای همان نمونه هایی که مربوط به کنترل است رنگ در حد قرمز یعنی بیان زیاد است اما برای نمونه های بیمار رنگ در حد زرد مایل به آبی و بیان کم است. همچنین برای بعضی از ژن ها نیز به صورت برعکس این موضوع وجود دارد. در اصل جداسازی بین نمونه ها به صورت درست انجام شده است.

همچنین در نمودار درختی که برای خو.شه بندی انجام شده است درستی خوشه بندی قابل مشاهده است چرا که اگر درست انجام نمی‌شد یک نمونه از گروه کنترل با یک نمونه از گروه بیمار جفت می‌شد که در این حالت نیفتاده است یا رنگ هایی برای نمونه های در یک گروه بود با همدیگر متفاوت می‌شد و آنها که در یک گروه نیستند رنگ مشابهی داشته باشند.

نمودار گرمایی به صورت زیر است:



۷.بیان کردیم که چون رفرنس را گروه کنترل گذاشتیم log2fc مثبت به معنای بیش بیانی یک ژن در گروه بیمار و مقدار منفی آن به معنی کم بیانی یک ژن در گروه بیمار نسبت به کنترل است.