به نام خدا

دانشکده مهندسی کامپیوتر

مقدمه‌ای بر بیوانفورماتیک
نیم‌سال اوّل ۰۵-۰۴
دکتر شریفی - کوهی

گزارش تمرین عملی سوم **فراز دعاگوی تهرانی ۴۰۲۱۰۵۹۹۸**

سوالات بخش اول

۱. زیرا که داده های ماتریس سریال پردازش شده و یا خلاصه است و مناسب نیست.

۲. تعداد سطر ها تعداد ژن ها و ستون ها تعداد نمونه هاست که به ترتیب ۲۸۸۸۹ و ۳۶ تا اند.

۳. به دلیل اینکه rRNA ها معمولا حذف میشوند و به خوبی توالی یابی نمیشوند.

۴. به دلیل گسترده بودن دامنه و غیرصحیح بودن آن، normalized هستند.

۵. توجه نکردن به آن، میتواند باعث بایاس شدن خروجی شود و به یک نمونه بیشتر توجه کرد.

سوالات بخش دوم

۱. ستون tile بیشترین کمک را کرد چرا که در آن مشخص بود که نوع health control بوده یا برای بیماری.

۲. از گروه control ۱۷ تا و از دیگری ۱۹ تا داریم که تقریبا برابرند و تعادل خوبی را نشان میدهد.

۳. اگر یکی نباشند تطابق و انتصاب نمونه ها به گروه ها درست انجام نمیشود و ادامه تحلیل دچار خطا میشود.

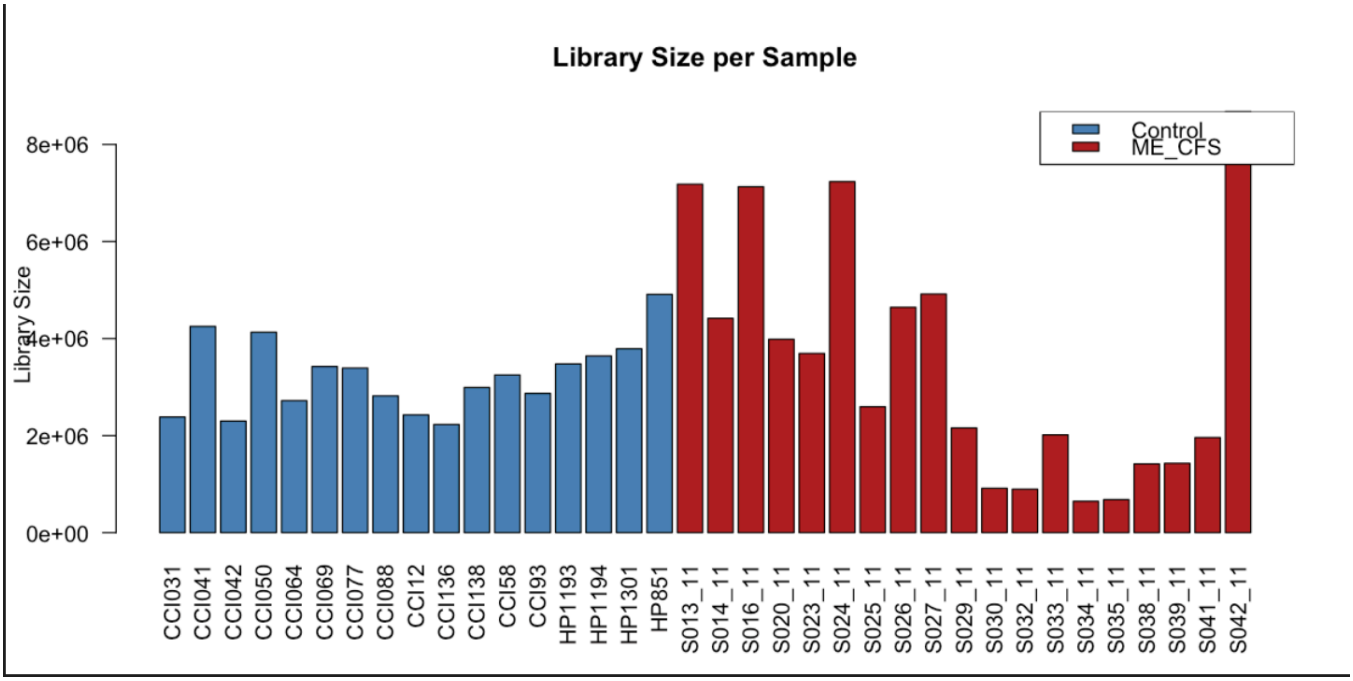
۴. برخی نتایج برعکس میشوند و پیش بینی ها را خراب میکنند.

۵. معمولاً سطح مرجع Control انتخاب می‌شود تا logFC بیان تغییرات بین بیمار و سالم محاسبه شود. انتخاب Con-trol به عنوان مرجع باعث می‌شود مقدار مثبت logFC نشان‌دهنده افزایش بیان در بیمار نسبت به سالم باشد و تفسیر ساده و مستقیم داشته باشد.

سوالات بخش سوم

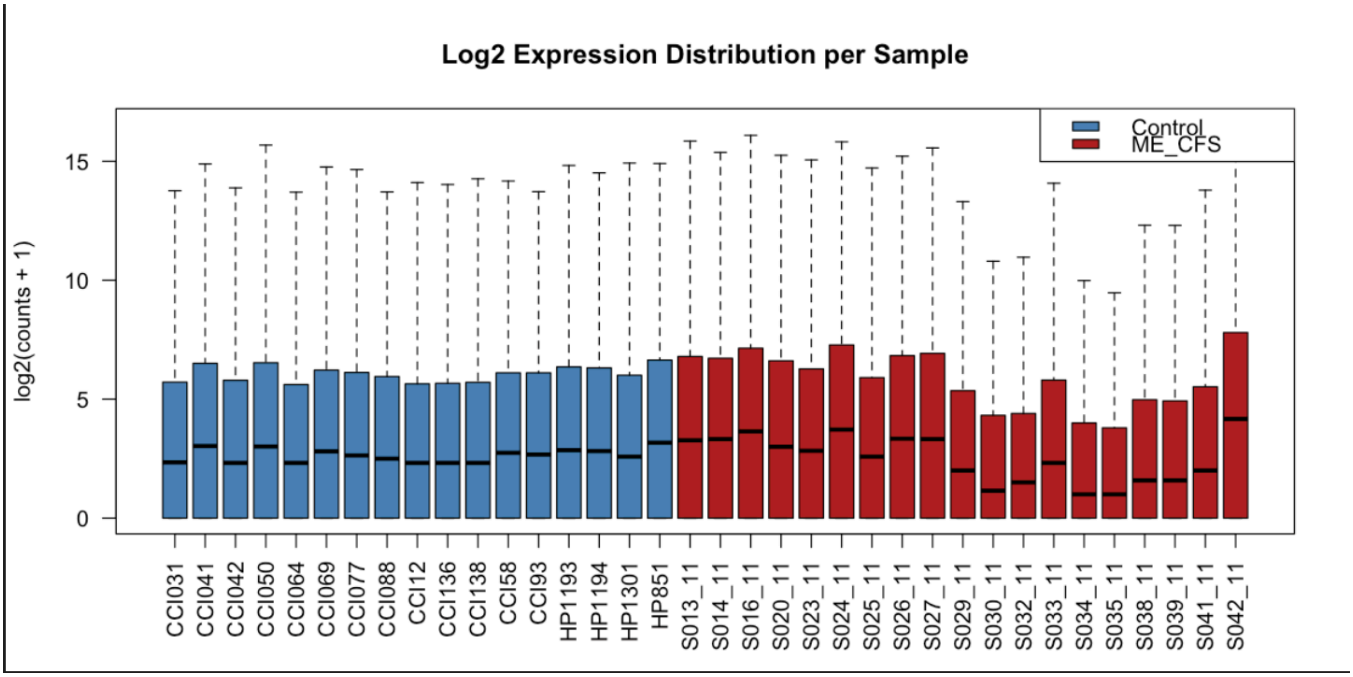
۱. با توجه به خروجی اولین سلول این بخش، کمترین مقدار برابر۰، و بشترین مقدار برابر ۷۲.۹۵۷۲۶ میباشد. همچنین میانه و میانگین به ترتیب برابر ۶۲.۴ و ۹۵.۱۱۶ میباشند. نبود اعداد منفی و وجود اعشار نشان میدهد که داده ها شمارشی خام نیستند و نرمال شده هستند.

۲. با توجه به نمودار و خروجی بخش های قبلی، کمترین library size مربوط به S034_11 با مقدار ۶۴۷۲۸۸/۵ و بیشترین مقدار مربوط به S042_11 با مقدار ۸۶۷۸۴۶۹ میباشد.
با توجه به خروجی های سلول دوم این بخش، تفاوت چشم گیری بین میانگین ها وجود ندارد اما مقدار ماکسیمم و مینیمم ها با هم فاصله زیادی دارد.
در صورت اختلاف زیاد، به سمت یکی بایاس پیدا میکنیم و در ادامه پیش‌بینی هایمان درست نمیشود.



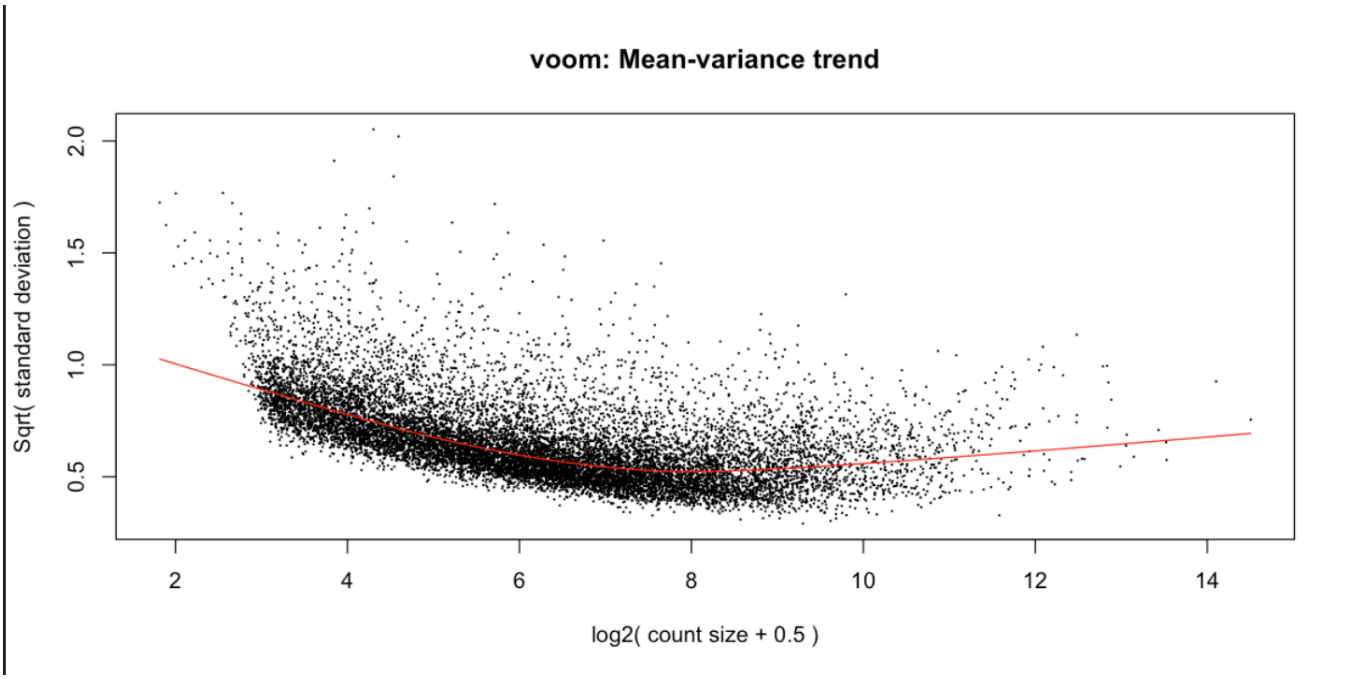
شکل ۱: نمودار library size ها به تفکیک گروه

۳. در کل، همانطور که در نمودار پایین میبینید، تقریبا همه مقادیر میانه و کلیت آنها با هم برابرند هرچند که برخی نمونه ها مثل راست ترین نمونه بزرگتر از بقیه و برخی دیگر نیز کوچکترند.
دلیل این تفاوت میتواند در کیفیت بد، داده پرت یا library prepration متفاوت باشد.



شکل ۲: بیان لگاریتمی به تفکیک گروه

۴. با کاهش میانگین بیان، واریانس زیاد میشود و با افزایش آن، پایدار تر میشود. در این مدل واریانس های بیشتر وزن کمتری میگیرند تا پایداری حفظ شود.



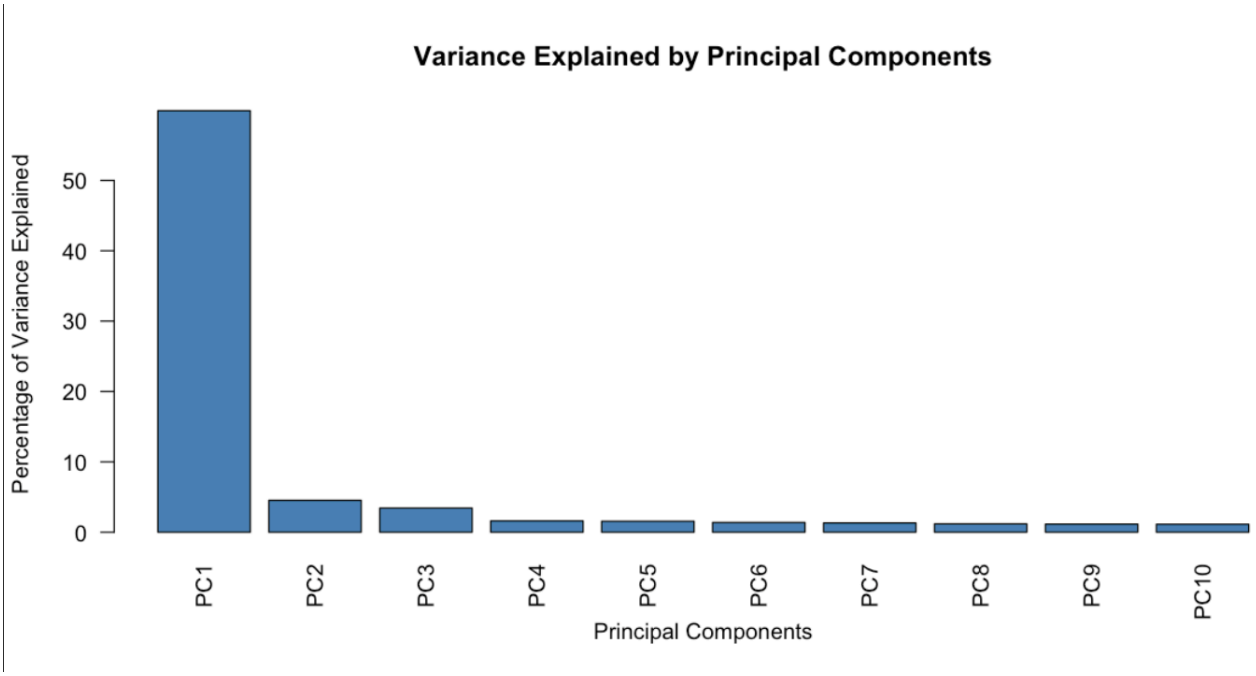
شکل ۳: نمودار voom

۵. در روش DESeq2 فرض میشود که داده ها صحیح اند اما در limma-voom نیازی به این فرض نیست و بر داده های

تقریبا نرمال هم درست عمل میکند.

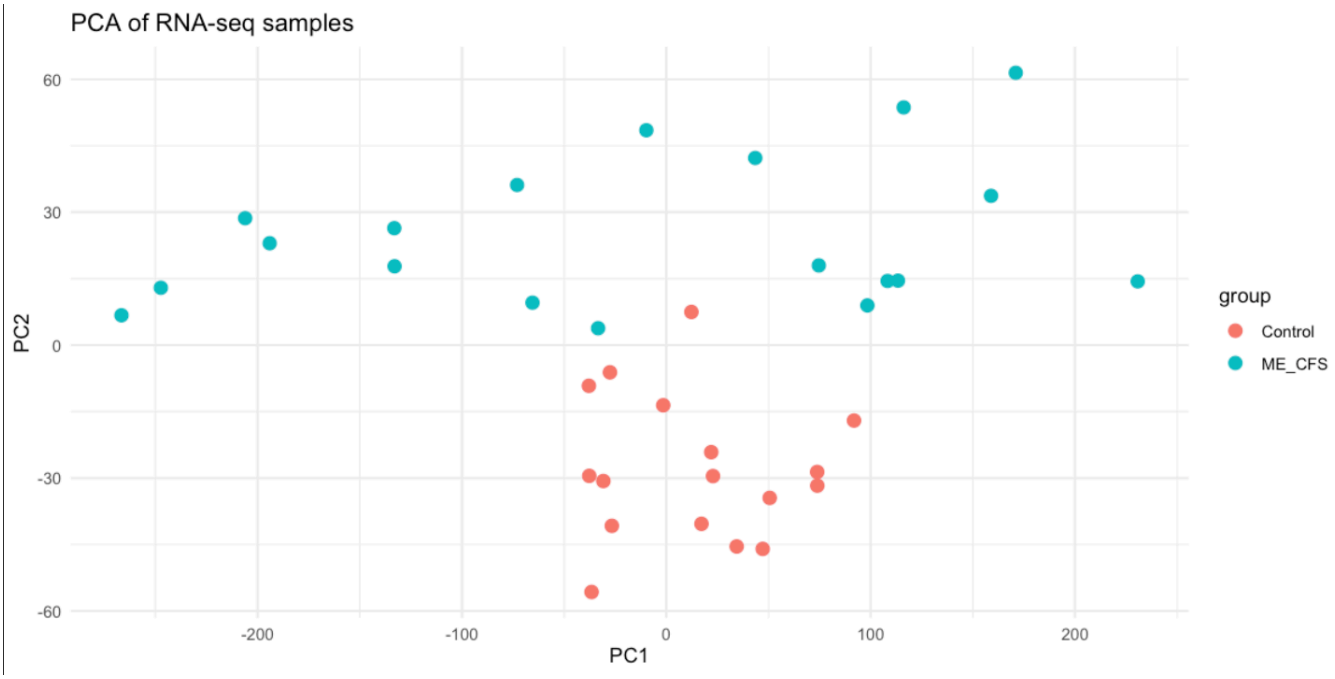
سوالات بخش چهارم

۱. بنا به خروجی سلول سوم این بخش، حدود ۶۰ درصد واریانس از PC۱ که این مسئله، نشان دهنده غلبه این بعد است و آن را توضیح میدهد.



شکل ۴: درصد واریانس توضیح داده شده توسط اولین ۱۰ PC

۲. در این نمودار، داده ها تقریبا به صورت کامل با یک خط جدا میشوند که این مسئله بیانگر این است که داده های این دو گروه تفاوت های معنا داری دارند و در فضای کم‌بعد، قابل تفکیک اند.



شکل ۵: داده ها در فضای PC1-PC2

۳. در نمودار بالا، نقاطی که فاصله بسیار زیادی به دسته خود دارند، مانند یک نقطه ای که قرمز است ولی در بخش آبی قرار گرفته است، به عنوان داده پرت درنظر گرفته میشود.
دلیل این تفاوت میتواند موارد گوناگونی از جمله کیفیت پایین نمونه، خطا در sequencing، آلودگی نمونه، وضعیت بالینی خاص و یا دیگر موارد باشد.

۴. زیرا این نمونه ها، اطلاعات خاصی را به ما اضافه نمیکنند و در هر کجا مقدار نسبتا ثابتی دارند و وجود آنها تنها باعث نویز میشود.

۵. اگر از scale. = TRUE استفاده میکردیم، ژن های کم‌واریانس باقی مانده که بیشتر آنها نویز هستند، وزن بیشتری میگرفتند و موثر تر واقع میشدند که باعث ایراد در تصمیم گیری ما میشد.

۶. ژن ها با بیان بالاتر کاملا غالب میشدند و کل ساختار با به خود بایاس میکردند. با استفاده از log این داده ها متعادل تر میشوند و استفاده از آنها منطقی تر میشود.

سوالات بخش پنجم

۱. ژن هایی که بیان کمی دارند و اطلاعات کافی ندارند را حذف میکند. چرا که به پیشبرد و دسته بندی کمکی نمیکنند و تنها باعث نویز میشوند.

۲. این نمودار که در بخش ۳ به آن پرداختیم، نشان میدهد که با افزایش بیان، واریانس کاهش میابد.
وزن دهی بر همین اساس در این روش انجام میگیرد و ضرورت آن به دلیل برقرار ماندن فرض همسانی واریانس میباشد.

۳. با توجه به خروجی سلول سوم این بخش که در پایین میتوانید آن را ببینید، این سه ژن بیشترین معنا را دارند. همچنین، هرچقدر logFC بیان بیشتری در گروه ME_CFS نسبت به کنترل داشته باشد، مثبت تر میشود و هر چقدر بیان در دیگری بیشتر باشد، منفی تر میشود.

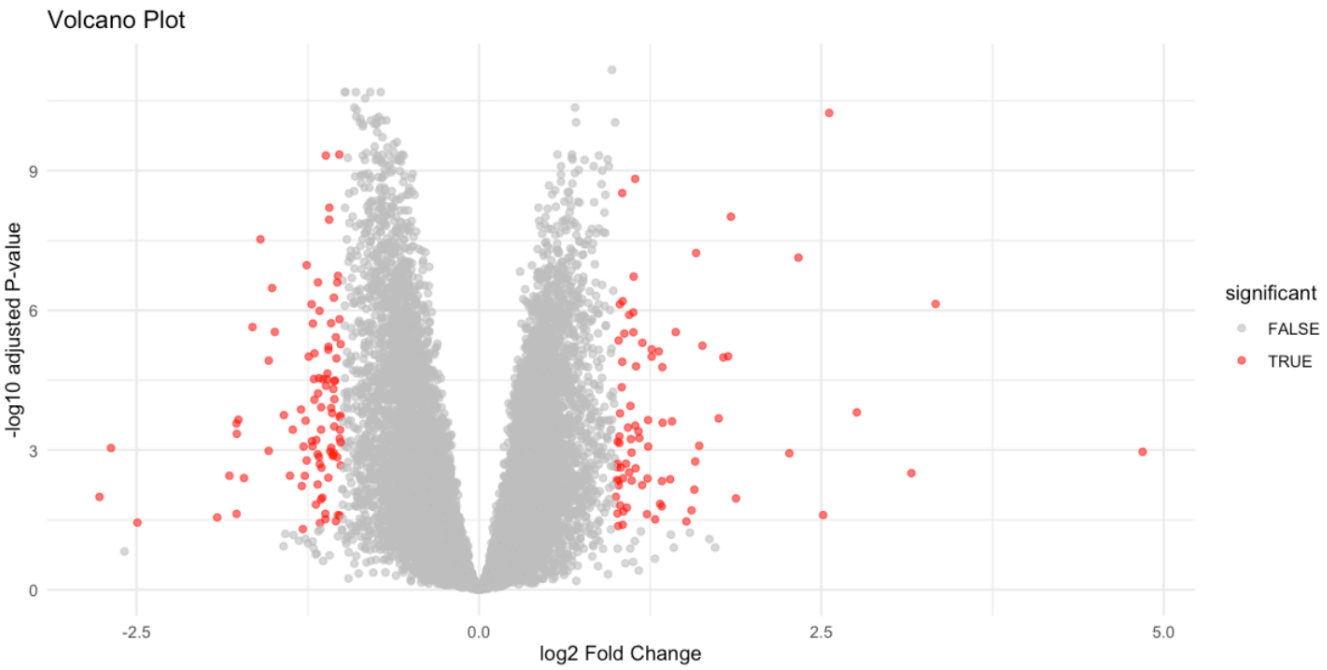
در نتیجه در خروجی پایین اولی در ME_CFS بیان بیشتری دارد و دوتای دیگر در Control.

A data.frame: 3 × 2

	logFC	adj.P.Val
	<dbl>	<dbl>
TAF1D	0.9711999	6.822068e-12
SCAF1	-0.9807698	2.057448e-11
MAP1S	-0.8988431	2.057448e-11

شکل ۶: سه ژن برتر بر اساس adj.P.Val

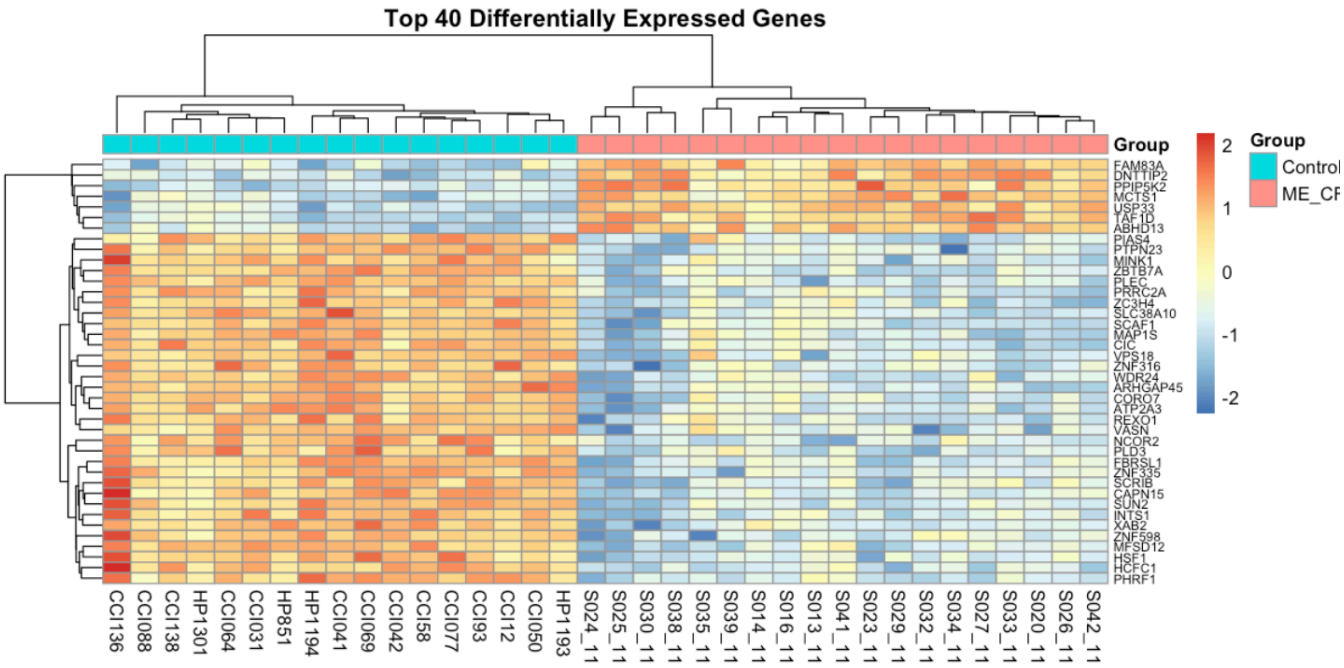
۴. بر اساس این نمودار، بخش های سمت چپ محور عمودی که logFC منفی دارند به گروه Control و مثبت ها به ME_CFS تعلق دارند. هرچقدر به محور عمودی نزدیک تر باشند، از معنای کمتری برخوردارند. هر جقدر مقدار |logFC| بیشتر میشود و مقدار adj.P.Val کمتر میشود، نقاط از اهمیت بیشتری برخوردارند و در نمودار زیر با رنگ قرمز مشخص شده اند.
همچنین با توجه به خروجی سلول ۵ این بخش، ۳۳۰۸ تا از ژن ها افزایش بیان و ۳۲۹۱ تا کاهش بیان دارند. در نتیجه تقریبا برابرند.



شکل ۷: نمودار volcano

۵. با توجه به خروجی سلول چهارم این بخش، ۱۷۴ ژن این ویژگی را دارند و آن‌ها ژن‌هایی هستند که هم مقدار تغییر زیاد دارند و هم معناداری آماری کافی دارند.

۶. بله همانطور که در شکل زیر میتوانید ببینید نمونه های دو دسته در خوشه های صحیح قرار گرفتند.
اگر نمونه ای اشتباه میشد، احتمالا ناشی از خطای بیولوژیکی یا خطای تکنیکی و یا داده پرت بود.



شکل ۸: نمودار volcano

۷. همانطور که بالاتر گفته شد، همچنین، هرچقدر logFC بیان بیشتری در گروه ME_CFS نسبت به کنترل داشته باشد، مثبت تر میشود و هر چقدر بیان در دیگری بیشتر باشد، منفی تر میشود. در این مدل تفسیر ME_CFS vs. Control اختلاف آنها میشود.