

باسمه تعالی



پروژه درس بیوانفورماتیک

فاز یک

نیم سال اول سال تحصیلی ۱۴۰۲-۱۴۰۱

اعضای گروه: علی نظری، پرهام باطنی، سیدمحمدیوسف نجفی

شماره های دانشجویی:

علی نظری: ۹۹۱۰۲۴۰۱

پرهام باطنی: ۹۹۱۰۵۲۹۴

سیدمحمدیوسف نجفی: ۹۹۱۰۲۳۶۱

ایمیل ها:

[ali.nazari.8102@gmail.com](mailto:ali.nazari.8102@gmail.com)

[mp.bateni@gmail.com](mailto:mp.bateni@gmail.com)

[najafim2002@gmail.com](mailto:najafim2002@gmail.com)

۱. درمورد microarray، روش کار آن و فرمت داده های خروجی آن به طور مختصر توضیح دهید. (۱۰)

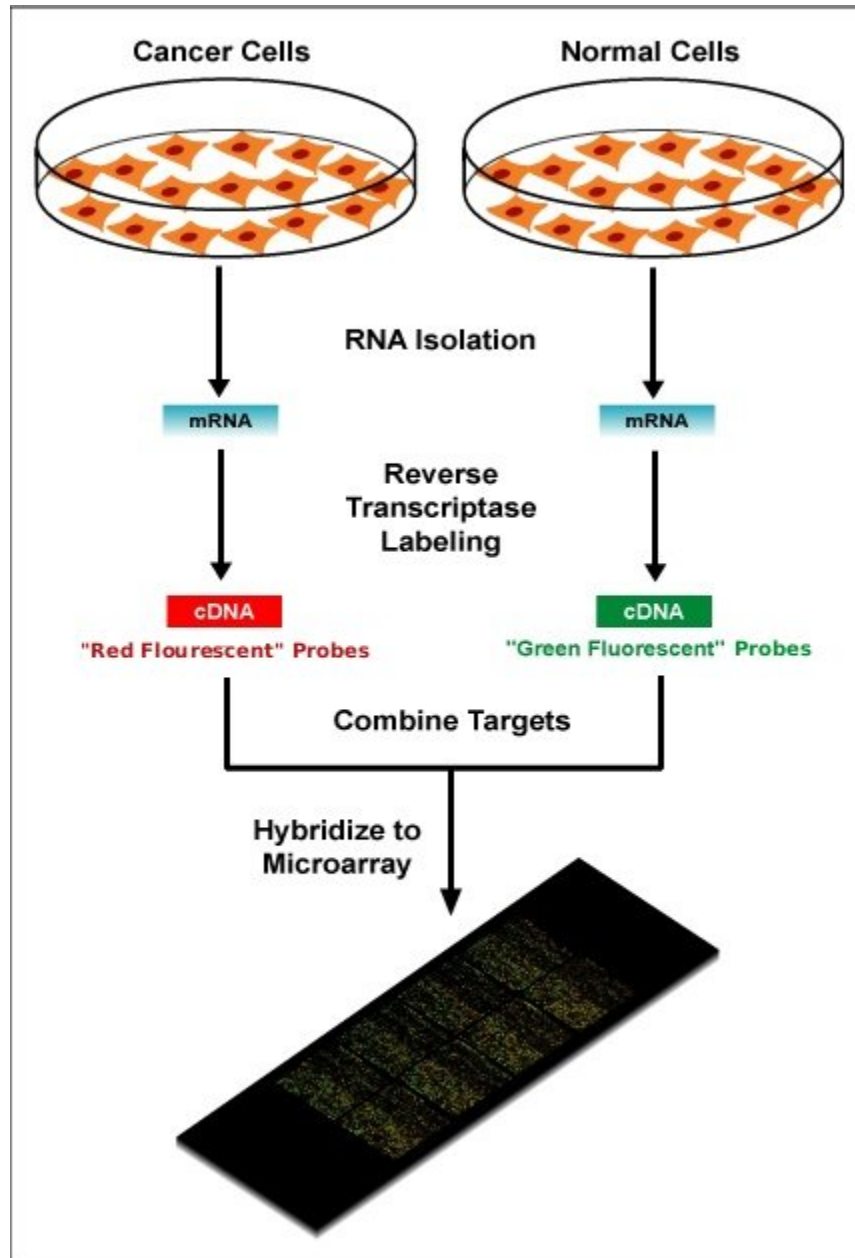
روش microarray یک روش جدید آزمایشگاهی برای شناسایی میزان بیان ژن چند ژن یا شناسایی جهش موجود در چند ژن به صورت همزمان است. روش کار در به طور خاص DNA microarray به این صورت است که یک مجموعه شناخته شده شامل هزاران یا حتی میلیون ها قطعه nucleic acid بر روی یک قطعه سیلیکونی در ابعاد ۱.۵ در ۱.۵ سانتی متر قرار داده می شوند یا به اصطلاح bind می شوند. این قطعه سیلیکونی شامل تعداد زیادی pixel است که هر کدام شامل یک تکه کوتاه از DNA یکسان به نام oligonucleotide است که این قطعات کوچک DNA طولی در حدود ۲۵ base pair دارند و این قطعات DNA مشخص هر کدام بیانگر مکمل قسمتی از یک ژن متمایز است به نوعی که در ادامه با اتصال هر sample به آن می توان آن sample را به یک ژن مشخص ارتباط داد. در ادامه نیاز است که قسمتی از DNA که در هر sample ما express می شود را به یک روشی extract کنیم. منطقاً این کار از طریق RNA امکان پذیر است و به این منظور از هر کدام از سلول های sample ها RNA آن ها را با استفاده از آنزیم هایی که هر کدام به طور مجزا DNA، پروتئین و قند و چربی ها را می شوید و می برد، استخراج کنیم. سپس با استفاده از یک آنزیم خاص به نام Reverse transcriptase، بر روی این رشته RNA فرایند Reverse transcript را پیاده سازی کنیم تا به cDNA یا complementary DNA معادل رشته RNA برسیم. طبیعتاً در هر سلول هر ژنی که میزان بیان بیشتری داشته باشد نمونه های cDNA متناظر با قسمتی از آن ژن ها، نیز در آن بیشتر است. پس از استخراج cDNA، آن را با استفاده از یک رنگ Fluorescent رنگ می کنیم و سپس نمونه های خود را روی microarray می ریزیم و در دستگاهی قرار می دهیم که به خوبی مخلوط شود و پس از گذشت زمان کافی بسته به بیان شدن برخی ژن ها در نمونه خود شاهد این خواهیم بود که بعضی خانه ها رنگی می شوند و برخی رنگی نمی گیرند. هرچه رنگ خانه های بیشتر باشند این به معنا است که آن ژن متناظر با آن pixel بیشتر بیان شده و لذا میزان بیان آن بالاتر است. توجه داشته باشید که بین هر cDNA و DNA آن pixel متناظر خود در صورت امکان پیوند هیدروژنی برقرار می شود همچنین ممکن است گاهی بین یک cDNA و یک probe در یک pixel یک پیوند ضعیف ایجاد شود که به این منظور فرایند wash انجام می شود تا پیوندهای ضعیف از بین بروند و فرصت برای پیوندهای قوی ایجاد شود. در نهایت خروجی مورد نظر یک آرایه از خانه هایی با یک رنگ مشخص خواهند بود که هر کدام بیانگر میزان بیان یک ژن خاص هستند.

همچنین گفتنی است ممکن است این سوال پیش بیاید که از کجا می توان مطمئن بود که احتمال ایجاد پیوند هیدروژنی بین هر cDNA و probe متناظرش یکسان است و از کجا معلوم که دلیل کمتر بیان شدن یک ژن کمتر بودن احتمال ایجاد پیوند هیدروژنی نباشد؟ در جواب باید گفت با استفاده از مقوله ای به نام GC content استفاده می کنیم که تعداد حرف های G و C در یک probe است و با تنظیم probe ها به گونه ای که همه GC content یکسان داشته باشند می توان احتمال ها را تقریباً برابر کرد چرا که در صورت برابر بودن GC content تعداد پیوندهای هیدروژنی ۳ تایی و ۲ تایی در probe ها یکسان می شود و به سبب آن احتمال چسبیدن cDNA برای همه probe ها یکسان می شود.

همچنین گفتنی است در تحلیل داده های microarray معمولاً از یک تکنیک دیگر به نام Background correction نیز استفاده می کنند که به این شکل عمل می کند که یک threshold برای میزان رنگ probe ها در نظر می گیرند که در ادامه نحوه محاسبه threshold توضیح داده می شود و سپس از میزان رنگ همه probe ها این میزان threshold کم می شود چرا که این میزان صرفاً یک noise بوده است. روش تشخیص threshold نیز به این صورت است که با استفاده از یک سری مجموعه probe که در آزمایش ما حکم داده control دارند استفاده می کنیم به این صورت که مطمئنیم به این probe هیچ cDNA از نمونه آزمایش ما نخواهد چسبید و در نهایت میزان رنگی که این probe می گیرد بیانگر میزان noise موجود در آزمایش است.

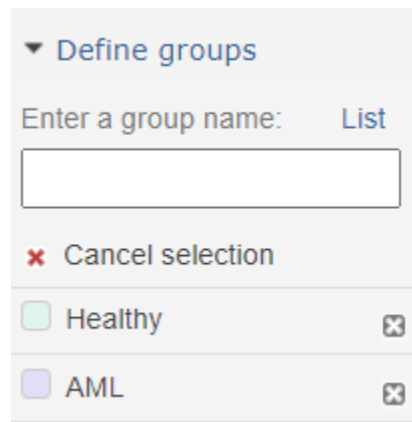
برای درک بیشتر مسئله به مثال زیر نیز توجه کنید.

به عنوان مثال دو sample یکی از یک بیمار سرطانی و دیگری از یک فرد سالم در اختیار داریم و قصد داریم ببینیم که این دو فرد در چه ژن‌هایی دچار تغییر شده‌اند تا بتوانیم ژن‌هایی که در بیماری سرطان نقش دارند را بیابیم. به این منظور تمامی مراحل آماده سازی sample ها که در قسمت قبل گفته شده را انجام می‌دهیم و cDNA های فرد بیمار را قرمز و برای فرد سالم را سبز می‌کنیم. سپس این دو را با هم مخلوط می‌کنیم و حاصل را روی microarray می‌ریزیم. نتیجه نهایی بسیار هیجان انگیز است چرا که pixel هایی که رنگ آن‌ها بیشتر به سمت قرمز باشد بیانگر ژن‌هایی خواهند بود که بیشتر در سلول‌های سرطانی بیان شده‌اند و در سلول‌های فرد سالم کمتر بیان شده‌اند و pixel هایی که بیشتر به رنگ سبز باشند بیانگر ژن‌هایی هستند که بیشتر در سلول‌های سالم بیان شده‌اند و در سلول‌های بیمار سرطانی کمتر بیان شده‌اند. به همین راحتی می‌توان تاثیر up/down regulation برخی ژن‌ها روی به وجود آمدن بیماری سرطان را با استفاده از این روش بررسی کرد.



۲. داده هایی که phenotype آنها normal است را به عنوان داده های گروه سالم و داده هایی که source name آن ها AML patient است را به عنوان گروه داده های بیمار در نظر بگیرید. داده های اولیه ممکن است برای تحلیل های بعدی آماده نباشند. کیفیت داده ها را از جنبه هایی که به نظرتان میرسد بررسی کنید و در صورت لزوم تغییرات لازم را روی آن ها اعمال کنید. (راهنمایی: برای مثال نرمال سازی داده ها). برای هر ویژگی ای که کیفیتش را کنترل میکنید ذکر کنید که این کنترل چه لزومی دارد، مراحل بررسی و کنترل خود را گزارش کنید (برای مثال نمودارها و...) و اگر لازم بود تغییری در داده ها ایجاد کنیم، تاثیر تغییرات را گزارش کنید. (۳۰)

ابتدا داده ها را از سایت گفته شده دانلود می کنیم و سپس package های لازم برای کار با داده مثل BiocManager را دانلود می کنیم. در ادامه قبل از شروع به کار با داده ها با استفاده از ابزار آماده سایت رفرنس کمی به بررسی داده ها می پردازیم. به این منظور ابتدا داده های سالم و سرطانی را به دو گروه جدا می کنیم.



▼ Define groups

Enter a group name: [List](#)

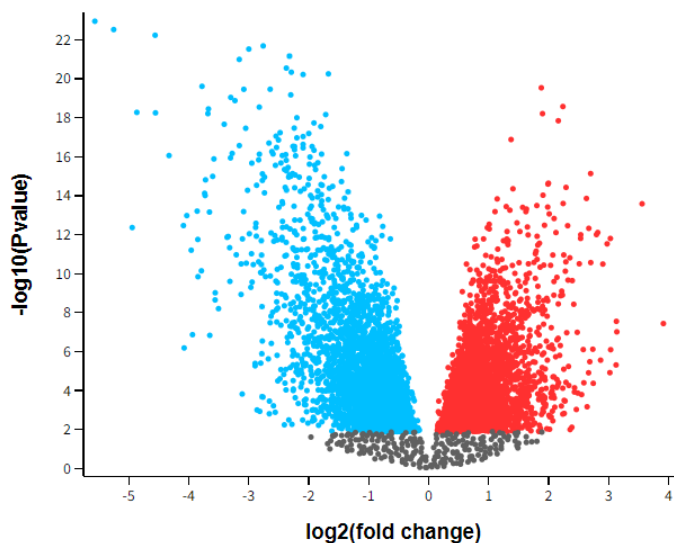
✖ Cancel selection

☐ Healthy ✖

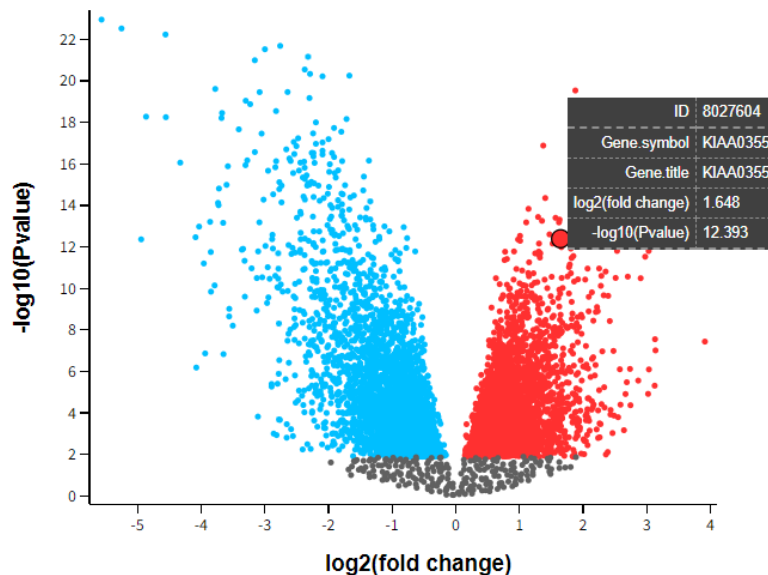
☐ AML ✖

پس از این تقسیم بندی ۴۹ sample سالم و ۱۸ sample سرطانی خواهیم داشت که روی این داده ها آنالیز را انجام می دهیم. در ادامه خروجی آنالیز انجام شده و عکس های خروجی را توضیح می دهیم.

**Volcano plot**  
**GSE48558: Expression data from normal and**  
**Malignant hematopoietic...**  
 Healthy vs AML, Padj<0.05



**Volcano plot**  
**GSE48558: Expression data from normal and**  
**Malignant hematopoietic...**  
 Normal vs Patient, Padj<0.05



نمودار فوق بیانگر نسبت statistical significance به میزان تغییر مقدار برای هر probe در مقیاس log را نشان می‌دهد که به آن volcano plot نیز می‌گویند. به ازای هر probe اگر مقدار داخل sample ها را در هر گروه بیمار سرطانی و سالم به صورت جدا میانگین بگیریم و از هم کم کنیم  $\log(\text{fold change})$  را خواهیم داشت.

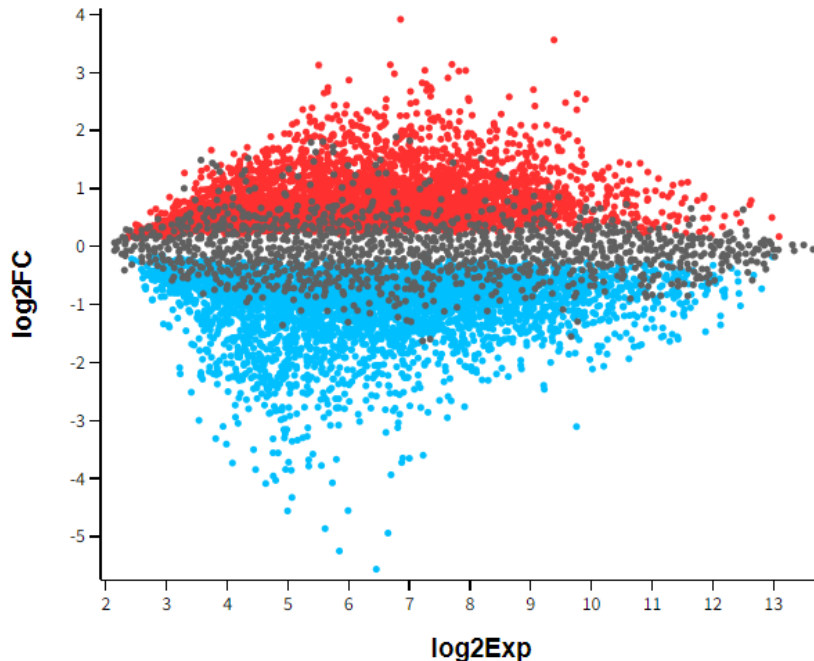
$\log(\text{fold change for probe } x \text{ (gene } x)) =$

$\text{mean}(\text{value each probe } x \text{ in patients samples}) - \text{mean}(\text{value each probe } x \text{ in normal samples})$

همچنین محور عمودی نیز بیانگر این است که آیا  $\log \text{ fold change}$  برای هر probe و متناظر آن هر gene در دو نمونه سالم و سرطانی اختلاف واضحی دارند یا خیر و نمونه‌های قرمز شده بیانگر ژن‌هایی خواهند بود که در نمونه‌های سرطانی به صورت محسوسی up regulate شده‌اند و ژن‌های آبی بیانگر ژن‌هایی خواهند بود که در نمونه‌های سرطانی به صورت محسوسی down regulate شده‌اند. نمونه‌های سیاه نیز از میزان adj Pvalue کافی برای اطمینان از محسوس بودن این اختلاف برخوردار نیستند.

در یکی از قسمت‌های بعدی به توضیح بیشتر درباره adj p value و دلیل استفاده از آن خواهیم پرداخت.

**Meandiff plot**  
**GSE48558: Expression data from normal and**  
**Malignant hematopoietic...**  
 Healthy vs AML, Padj<0.05



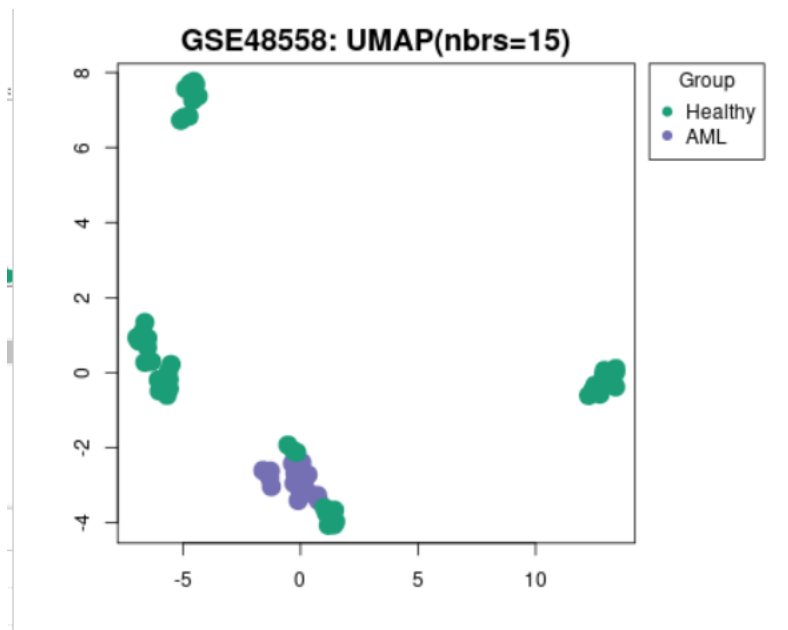
مشابه نمودار قبلی تا حدودی در این نمودار که به نمودار Mean difference plot مشهور است قصد داریم ژن‌های با میزان بیان متفاوت را به تصویر بکشانیم. به این منظور برای هر ژن از فرمول‌های زیر برای بدست آوردن logFC و logExp استفاده می‌کنیم.

$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

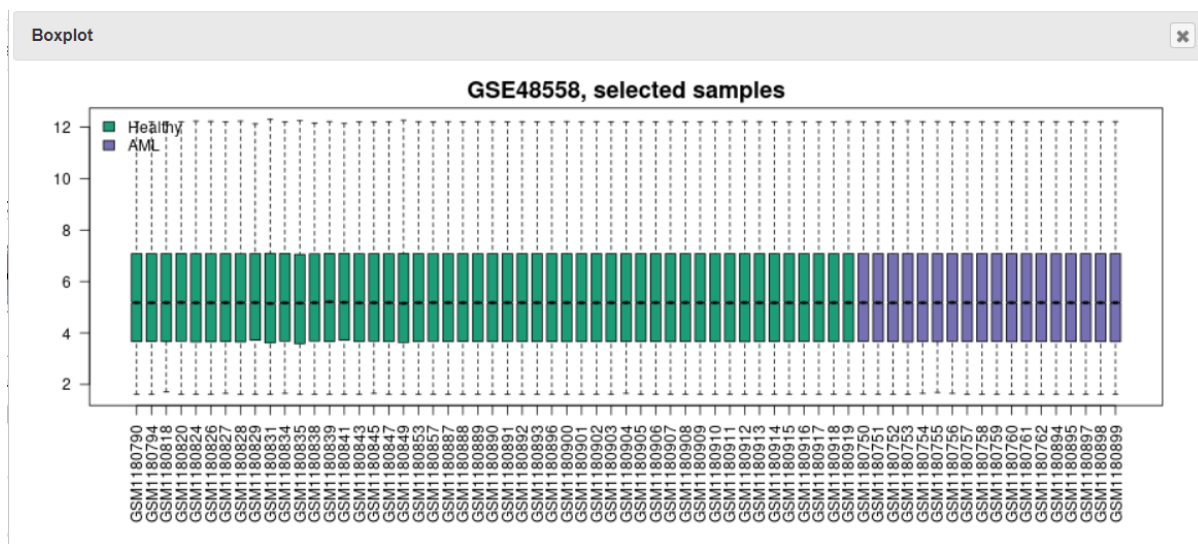
$$A = \frac{1}{2} \log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$

در فرمول‌های بالا R یک ژن بیانگر میزان میانگین Intensity مربوط به probe ژن موردنظر در نمونه‌های سرطانی و G مربوط به نمونه‌های سالم است و M برابر logFC خواهد بود و A بیانگر logExpression خواهد بود که میانگین Intensity مربوط به probe ژن موردنظر در بین دو دسته سرطانی و سالم خواهد بود.

به این ترتیب انتظار داریم هرچه قدر مطلق logFC بیشتر باشد و logExpression کمتر باشد به احتمال بیشتری ژن‌های متناظر تفاوت محسوسی در دو نمونه سرطانی و سالم داشته باشند. نحوه رنگ شدن بعضی ژن‌ها نیز مشابه قسمت قبل بر اساس up/down regulate شدن ژن و میزان محسوس بودن آن بر حسب adj p value خواهد بود.



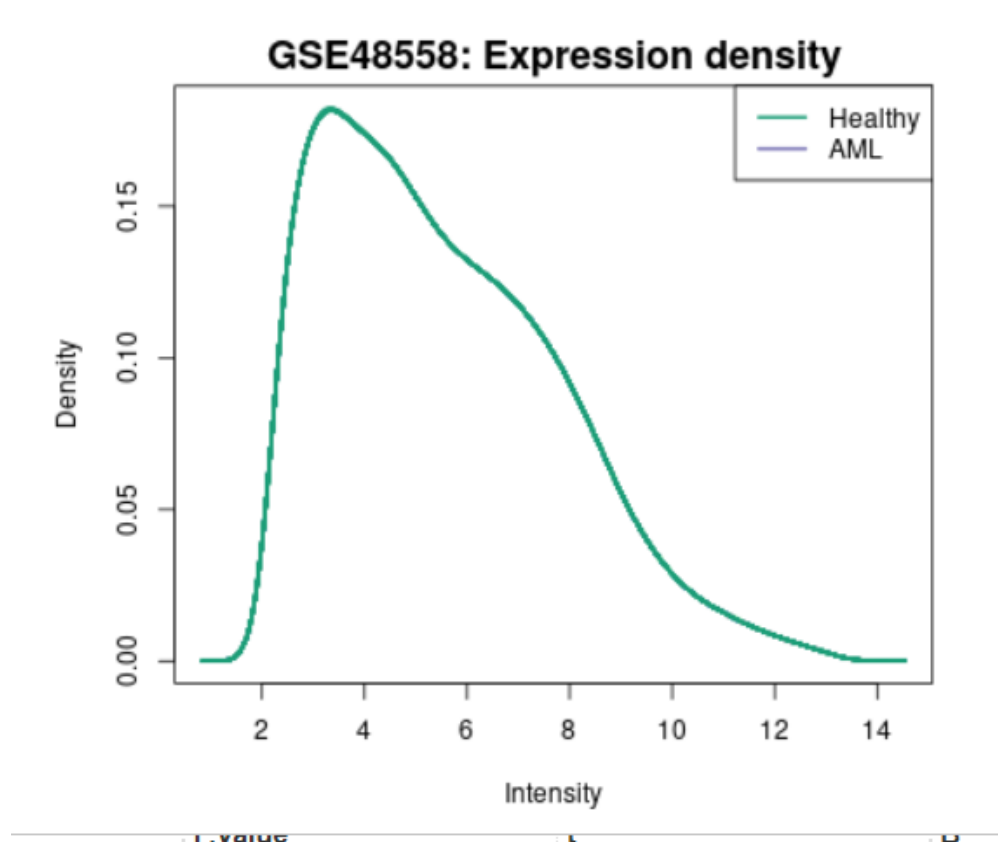
نمودار بالا خروجی یکی از روش‌های dimension reduction به نام Uniform manifold approximation and projection را نشان می‌دهد.



نمودار box plot فوق، توزیع مقادیر هر sample را نشان می‌دهد که به صورت box plot مشخص شده است و مقادیر داخل هر sample مقادیر intensity حدود ۳۰ هزار probe در array ما هستند که از قضا log transform نیز شده‌اند. مطابق شکل بالا از آنجا که range داده‌های هر sample بسیار نزدیک به هم است و همه تقریباً میانگین یکسانی دارند و چارک‌های نمونه‌ها اختلاف فاحشی با هم ندارند (برای مثال میانه یکی پایین‌تر از چارک اول دیگری و یا بالاتر از چارک سوم دیگری نیست)، می‌توان فهمید که نیاز به normalize کردن مقادیر sample ها نیست. توجه داشته باشید دلیل استفاده از log transformation این است که در این مسئله مقدار اختلاف بین بیان دو ژن مد نظر نیست بلکه نسبت بیان دو ژن چیزی است که به دنبال آن هستیم و مهم این است که برای مثال مقدار بیان یک ژن در نمونه سرطانی چند برابر بیان آن در نمونه سالم است. لذا از log استفاده می‌کنیم که اختلاف هر دو مقدار آن معیاری از این نسبت خواهد بود و کمک به normalize کردن داده می‌کند.

$$\log(\text{expression1}) - \log(\text{expression2}) = \log(\text{expression1} / \text{expression2})$$

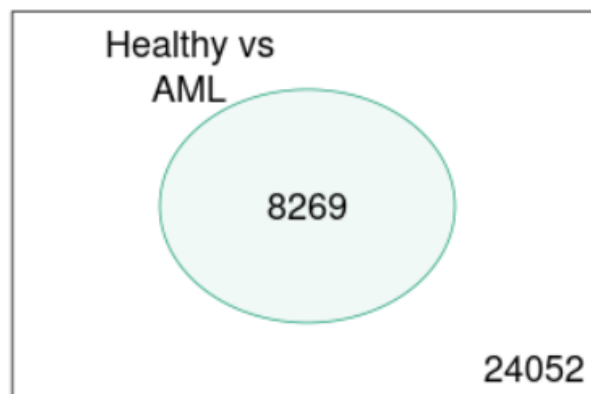




در نمودار بالا توزیع مقادیر هر sample سرطانی و سالم را می‌توانید مشاهده کنید که همانطور که از شکل واضح است چیزی نزدیک به ۲۰ درصد مقادیر همه نمونه‌ها Intensity نزدیک به ۴ داشته‌اند.

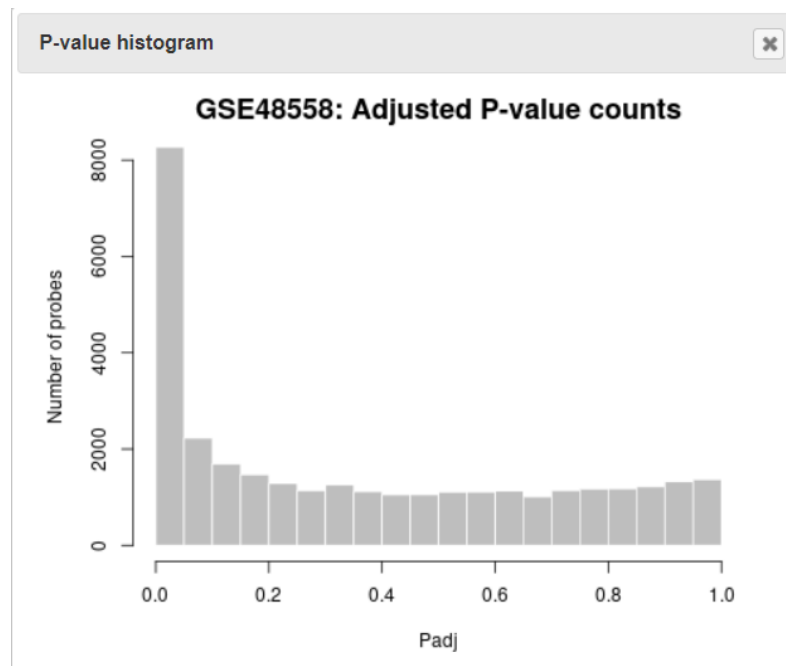
### Venn Diagram

GSE48558: limma,  $P_{adj} < 0.05$

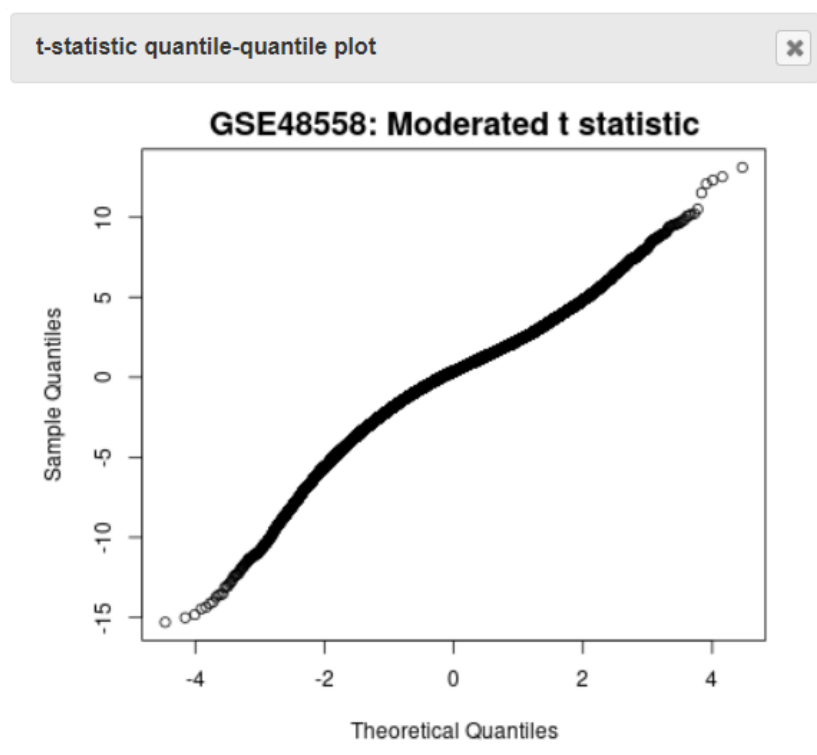


نمودار بالا نیز نمودار ون نمونه‌ها را نشان می‌دهد که از ۳۲۳۲۱ ژن موجود ۸۲۶۹ ژن به طور محسوس (significant) در دو نمونه سرطانی و سالم اختلاف بیان دارند.



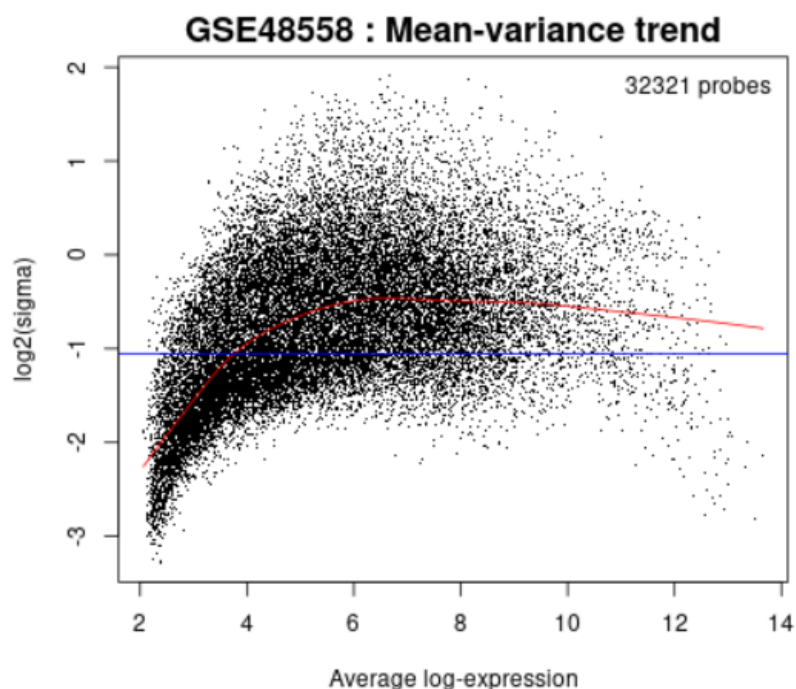


نمودار بالا توزیع Adjusted P-value های هر ژن را نشان می‌دهد به این شکل که ژن‌هایی که P-value های آن‌ها کمتر از threshold موردنظر است به عنوان ژن‌های موثر در بیماری سرطان معرفی خواهند شد. همانطور که در شکل مشخص است و در نمودار ون نیز عنوان کردیم چیزی نزدیک به ۸۲۶۹ ژن P-value کمتر از threshold دارند و این ژن‌ها، ژن‌های محسوس (significant) هستند.



نمودار بالا نمودار مقایسه چارک‌ها (Quantile ها) مربوط به داده‌ها و چارک‌های توزیع t-distribution را نشان می‌دهد. نمودار quantile-quantile plot در حالت کلی برای مقایسه دو توزیع احتمالاتی استفاده می‌شود و این مقایسه توسط مقایسه quantile های هر توزیع با توزیع دیگری انجام می‌شود. در این‌جا نیز حاصل این نمودار تقریباً یک خط است که می‌توان متوجه شد توزیع داده‌های اصلی به صورت خطی با توزیع student-t distribution در ارتباط است.

Mean-variance trend plot



نمودار فوق رابطه میانگین و واریانس مقادیر Intensity برای هر ژن را نشان می‌دهد. هر نقطه یک ژن است که میانگین و واریانس مقدار Intensity آن در کل نمونه‌ها محاسبه می‌شود و حاصل نمودار بالا می‌شود. خط قرمز تقریبی از روند کلی trend میانگین و واریانس ژن‌ها را نشان می‌دهد و خط افقی آبی تقریب کلی لگاریتم واریانس مقدار Intensity همه ژن‌ها است. همچنین برای هر ژن هم جدول بسیار مفیدی قرار دارد که به شکل زیر است:

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
8016932	3.62e-19	1.12e-23	-15.31	43.2	-5.564	MPO	myeloperoxidase
7970737	4.84e-19	2.99e-23	-15.03	42.3	-5.25	FLT3	fms related tyrosine kinase 3
7989647	6.31e-19	5.86e-23	-14.83	41.6	-4.559	KIAA0101	KIAA0101
7982663	1.66e-18	2.06e-22	-14.48	40.4	-2.757	BUB1B	BUB1 mitotic checkpoint seri...
8083422	1.94e-18	3.00e-22	-14.37	40.1	-2.997	SUCNR1	succinate receptor 1
7926259	3.71e-18	6.89e-22	-14.14	39.3	-2.319	MCM10	minichromosome maintenanc...
8061579	4.70e-18	1.02e-21	-14.03	38.9	-3.156	TPX2	TPX2, microtubule nucleation...
7966878	1.15e-17	2.84e-21	-13.75	37.9	-2.371	CIT	citron rho-interacting serineit...
8071212	1.66e-17	4.62e-21	-13.61	37.4	-2.288	CDC45	cell division cycle 45
7921033	1.78e-17	5.60e-21	-13.56	37.2	-1.67	IQGAP3	IQ motif containing GTPase ...
8089875	1.78e-17	6.04e-21	-13.54	37.2	-2.092	POLQ	DNA polymerase theta
8064539	6.59e-17	2.45e-20	-13.16	35.8	-3.777	CPXM1	carboxypeptidase X, M14 fa...
8126018	7.23e-17	2.91e-20	13.12	35.6	1.88	STK38	serine/threonine kinase 38
8132318	7.50e-17	3.45e-20	-13.07	35.5	-2.641	ANLN	anillin actin binding protein
7991406	7.50e-17	3.48e-20	-13.07	35.5	-3.08	PRC1	protein regulator of cytokinesi...
8155214	1.35e-16	6.70e-20	-12.89	34.8	-2.297	MELK	maternal embryonic leucine z...
8014974	1.73e-16	9.11e-20	-12.81	34.5	-3.298	TOP2A	topoisomerase (DNA) II alpha
8094278	2.38e-16	1.32e-19	-12.71	34.2	-3.228	NCAPG	non-SMC condensin I compl...
8076185	4.53e-16	2.66e-19	12.53	33.5	2.24	CBX7	chromobox 7
7984540	4.59e-16	2.84e-19	-12.51	33.4	-2.822	KIF23	kinesin family member 23
8019842	5.46e-16	3.55e-19	-12.45	33.2	-3.67	TYMS	thymidylate synthetase
8095110	7.79e-16	6.99e-19	-12.35	32.8	-4.864	KIT	KIT proto-oncogene receptor

اگر بر روی هر ژن کلیک کنیم، جدول مربوطه برای هر sample را می‌آورد. ولی قبل از بررسی یک نمونه از جدول هر ژن به توضیح ستون‌های جدول بالا می‌پردازیم.

ID شماره خاصی است که به یک probe در array داده شده است و این ID به نوعی به یک ژن اختصاص دارد. توجه داشته باشید که لزوماً از همه ID ها نمی‌توان ژن آن را حدس زد.

P.value نیز مقدار نهایی آزمون تست برای هر ژن را نشان می‌دهد. این تست برای بررسی این است که تفاوت بیان این ژن در sample های مختلف، معنا دار بوده است یا نه. به عبارتی دیگر هر سطر جدول یک probe است و برای اینکه متوجه شویم آیا اختلاف بیان ژن‌های متناظر هر probe در دو sample سرطانی و سالم معنا دار بوده است یا خیر از تست آماری استفاده می‌کنیم و p-value را حساب می‌کنیم. همچنین می‌توان از تست‌های آماری دیگر نظیر t-test نیز استفاده کرد.

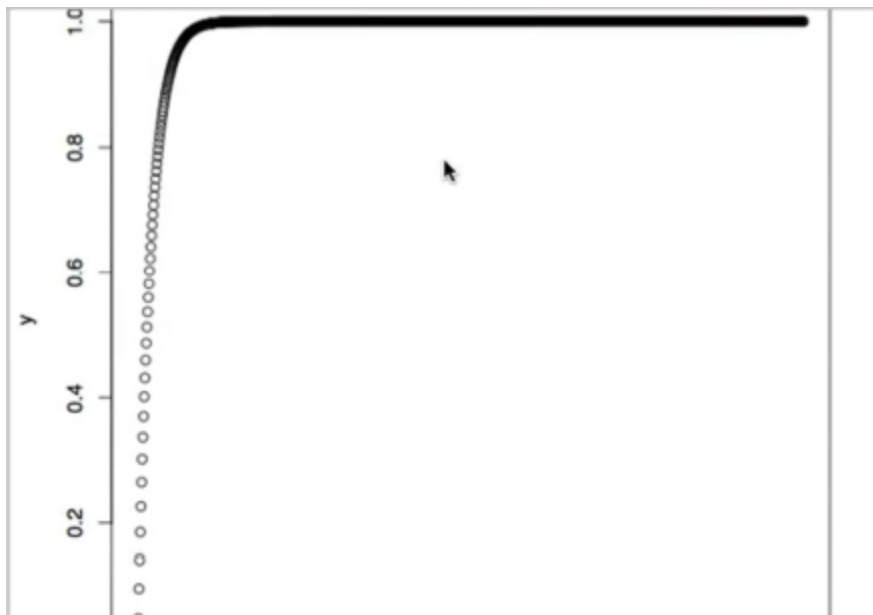
اکنون لازم است در مورد adj p-value توضیحاتی بدهیم. فرض کنید که چندین hypothesis را می‌خواهیم چک کنیم. این همان multiple test hypothesis است. یعنی روی چندین تست، فرض صفر های مختلف داریم و با آن فرض صفر، تست را انجام می‌دهیم. به عنوان مثال فرض صفر اول ما این است که در دو نمونه در یک جفت probe های خاص، این دو تفاوت معناداری ندارند. حال فرض صفر دوم ما، این است که ژن دوم در دو گروه تفاوت معناداری ندارد. پس به تعداد probe ها ما فرض صفر داریم. لذا به تعداد probe ها p.value داریم و در یک ستون این مقادیر مشخص شده‌اند. حال فرض کنید فرض صفر هر کدام از probe ها را اگر p.value متناظر آن‌ها کمتر از ۰.۰۵ بود، رد کنیم. چقدر احتمال دارد ژن‌هایی که تفاوت بیان معناداری ندارند را به عنوان اینکه تفاوت معنی دار دارند، گزارش کنیم؟ در مورد ژن اول، این موضوع، احتمال 0.05 دارد. هر ژن به تنهایی به این مقدار احتمال دارد که معنادار نباشند، ولی ما معنادار گزارش کنیم. پس به عبارتی دیگر دقت ما در تشخیص درست معنادار بودن تفاوت بیان هر ژن ۹۵ درصد است. حال اگر ژن ۱ و ۲ را در نظر بگیریم، چطور؟ این احتمالات در هم ضرب می‌شوند و میزان دقت به ازای تعداد زیادی ژن بسیار کاهش می‌یابد و لذا خطا بسیار بالا می‌رود که این موضوع اصلاً مطلوب نیست و به جای p-value از مفهوم دیگری به نام adj p.value استفاده می‌کنیم. قبل از توضیح دقیق‌تر این مفهوم به عکس زیر توجه کنید:

		Condition (as determined by " <u>Gold standard</u> ")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	False Negative (Type II error)	True Negative	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

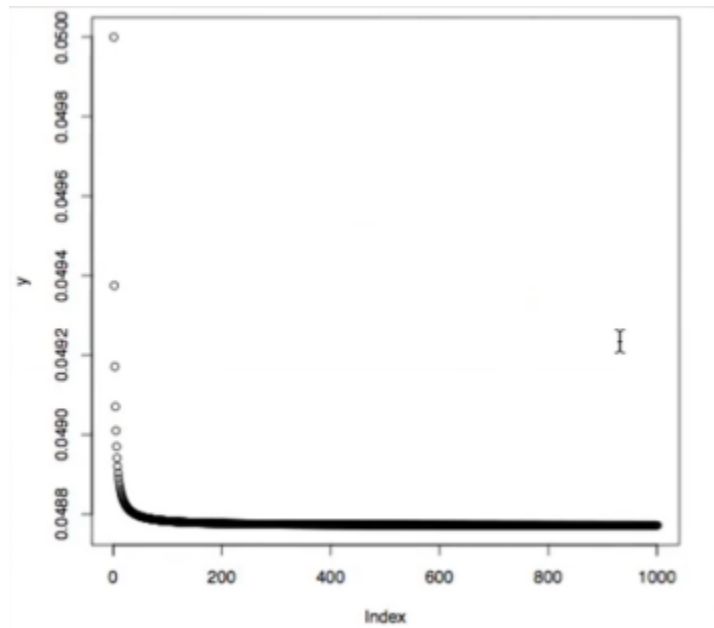
مطابق توضیحات گفته شده برای ۲ ژن احتمال خطا داشتن نتایج گزارش به شکل زیر است:

$$1 - (1 - 0.05)^2$$

همچنین برای تعداد نمونه‌های خیلی زیاد می‌توان با R نمودار آن را کشید و میزان خطا را مشاهده کرد.



همانطور که در تصویر بالا مشاهده می‌کنید از یک جایی به بعد خطا به یک میل می‌کند که این مسئله اصلا مطلوب نیست. لذا نیاز است از روش دیگری مثل adjacent p-value استفاده کنیم. برای adjacent کردن p-value روش های زیادی موجود است

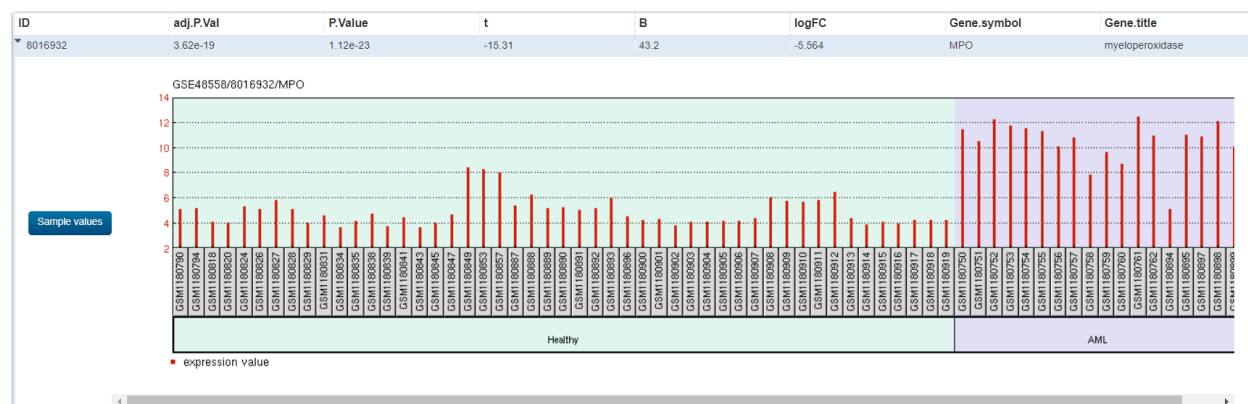


که یکی از معروف ترین های آن ها، روش بون فرونی است. در این روش، threshold ای که داریم را بر تعداد نمونه ها تقسیم می کنیم. در واقع هر کدام از p.value ها را در تعداد نمونه ها ضرب می کنیم. اگر بعد از این ضرب، باز هم تفاوت معنادار بود، در آن صورت تفاوت میزان بیان ژن را معنادار گزارش می کنیم. در این حالت نمودار به شکل زیر می شود:

پس می بینیم که وضعیت خیلی بهتر شده است و عدد adj p.value که در جدول قبلی وجود دارد، با چنین روشی به دست آمده است. البته نکته این است که روش بون فرونی می گوید اگر هر hypothesis مستقل از دیگری باشد، از این روش می توان استفاده کرد.

در توضیح بقیه ستون ها، ستون های B و t مقدار statistics test ها هستند.

همچنین یک ستون LogFC هم داریم که در قسمت‌های قبل این پارامتر را توضیح دادیم و عنوان کردیم که این مقدار مشخص می‌کند که مقدار بیان یک ژن در نمونه سرطانی چند برابر نمونه سالم آن است. برای مثال اگر بیان یک ژن در سمپل های سرطانی ۸ برابر سلول های سالم باشد، FC آن ۸ است و LogFC آن هم ۳ است. البته در این جا چون که داده ها در حالت لگاریتمی بای دیفالت هستند، دیگر تقسیم نمی کنیم و منها می کنیم. در کل ترکیبی از استفاده adj p-value و LogFC مناسب است. برای مثال خیلی وقت ها برای LogFC یک threshold می گذاریم که یا از یک بالاتر باشند و یا اینکه از -۱ کمتر باشد تا حداقل یکی دو برابر دیگری بیان شده باشد. ولی فقط به LogFC نمی توان نگاه کرد چون ممکن است واریانس خیلی زیاد باشد و همین باعث می شود که داده ها معنادار نباشند. پس ترکیب آن با adj p-value برای هر ژن مطلوب است. در انتها جدول مقادیر مربوط به یک probe را می‌توانید مشاهده کنید که میزان Intensity هر نمونه با تفکیک نوع sample به سالم یا سرطانی را نشان می‌دهد.



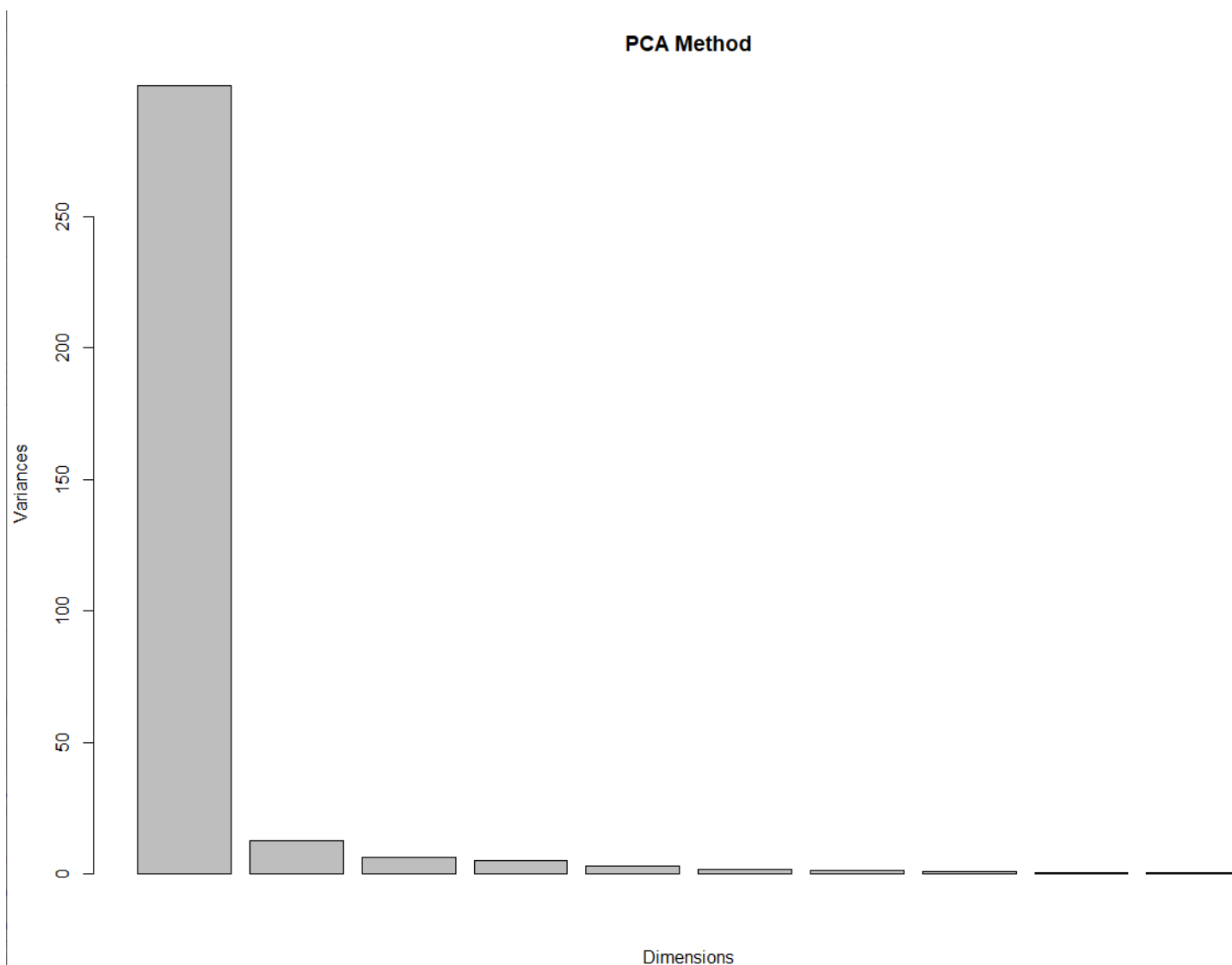
این جدول ها را می توان دانلود کرد و در R لود کرد و بسیاری تحلیل‌های دیگری با آن ها انجام داد. همچنین تحلیل‌های نموداری در کد R که تحویل داده می شود، موجود است. در نهایت بعد از تحلیل این نمودارها با R متوجه می‌شویم که داده ها به Log2 برده شده اند و نیاز به این کار نیست و نیاز به normalization هم نداریم. همچنین در کشیدن heatmap روی correlation داده ها هم به تفکیک خوبی رسیده‌ایم که این موضوع مطلوب ما است. لذا برای رسیدن به نمودارهای مطلوب بالا کافی است داده‌ها را در کد R به log2 ببریم.

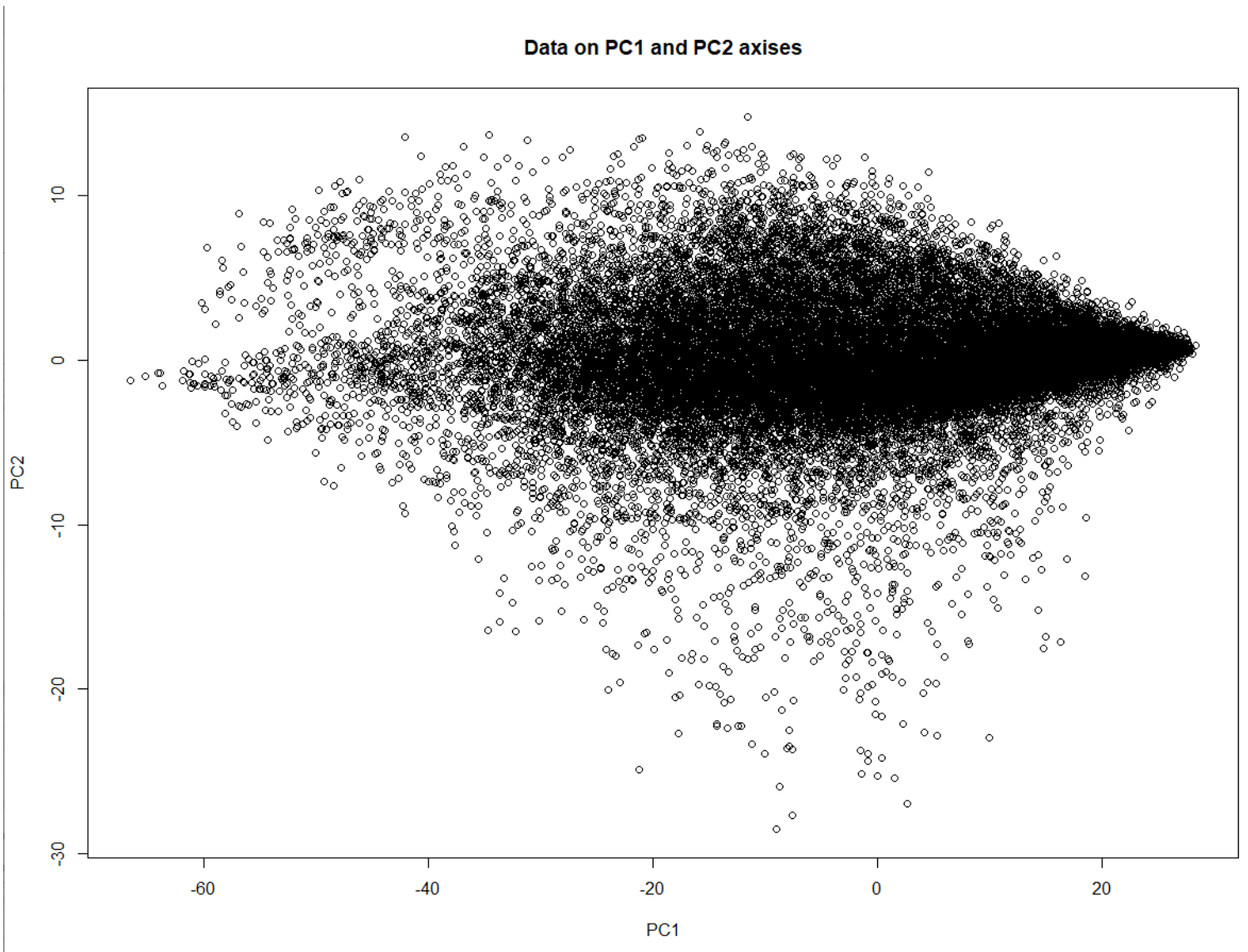
۳. لزوم کاهش ابعاد داده ها چیست؟ سه روش مختلف برای کاهش ابعاد را انتخاب کرده و نتایج حاصل از هر سه روش را گزارش کنید. سپس با مقایسه این نتایج، روشی که بهترین خروجی را نتیجه داده است، انتخاب کنید. دلایل انتخاب روش بهتر را ذکر کنید.

راهنمایی: برای مثال می توانید کاهش ابعاد را با سه روش PCA، MDS و tSNE انجام دهید. (۳۰)

لزوم کاهش ابعاد این است که در چنین مسائلی با تعداد زیادی sample هرکدام با تعداد زیادی ژن (۳۲۰۰۰) یا feature سروکار داریم و اگر بخواهیم بدون کاهش ابعاد از داده ها استفاده کنیم با پیچیدگی زمانی و حافظه ای زیادی مواجه خواهیم شد که مطلوب ما نیست. برای مثال تصور کنید بخواهیم مدل learning بر روی داده های خود fit کنیم و واضح است که اگر با داده های کاهش ابعاد نیافته بخواهیم کار کنیم چقدر فرایند یادگیری طولانی می شود. لذا استفاده از یک روش کاهش ابعاد الزامی است و در ادامه نتیجه سه روش کاهش ابعاد را بررسی کرده و بهترین روش را انتخاب می کنیم. (کدهای هر نمودار در فایل R موجود است.)

برای PCA خروجی کاهش ابعاد یافته داده ها به صورت زیر است:

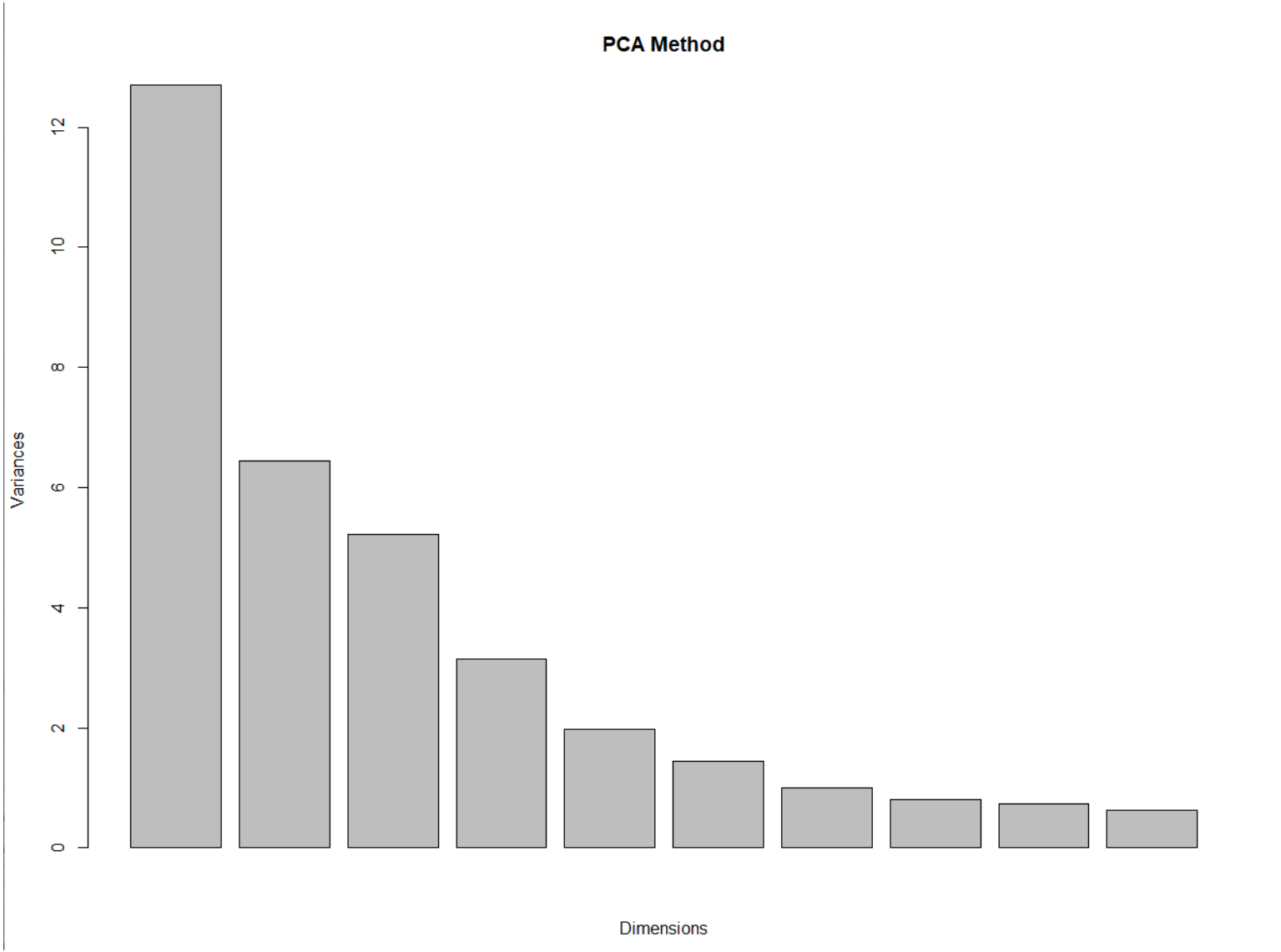


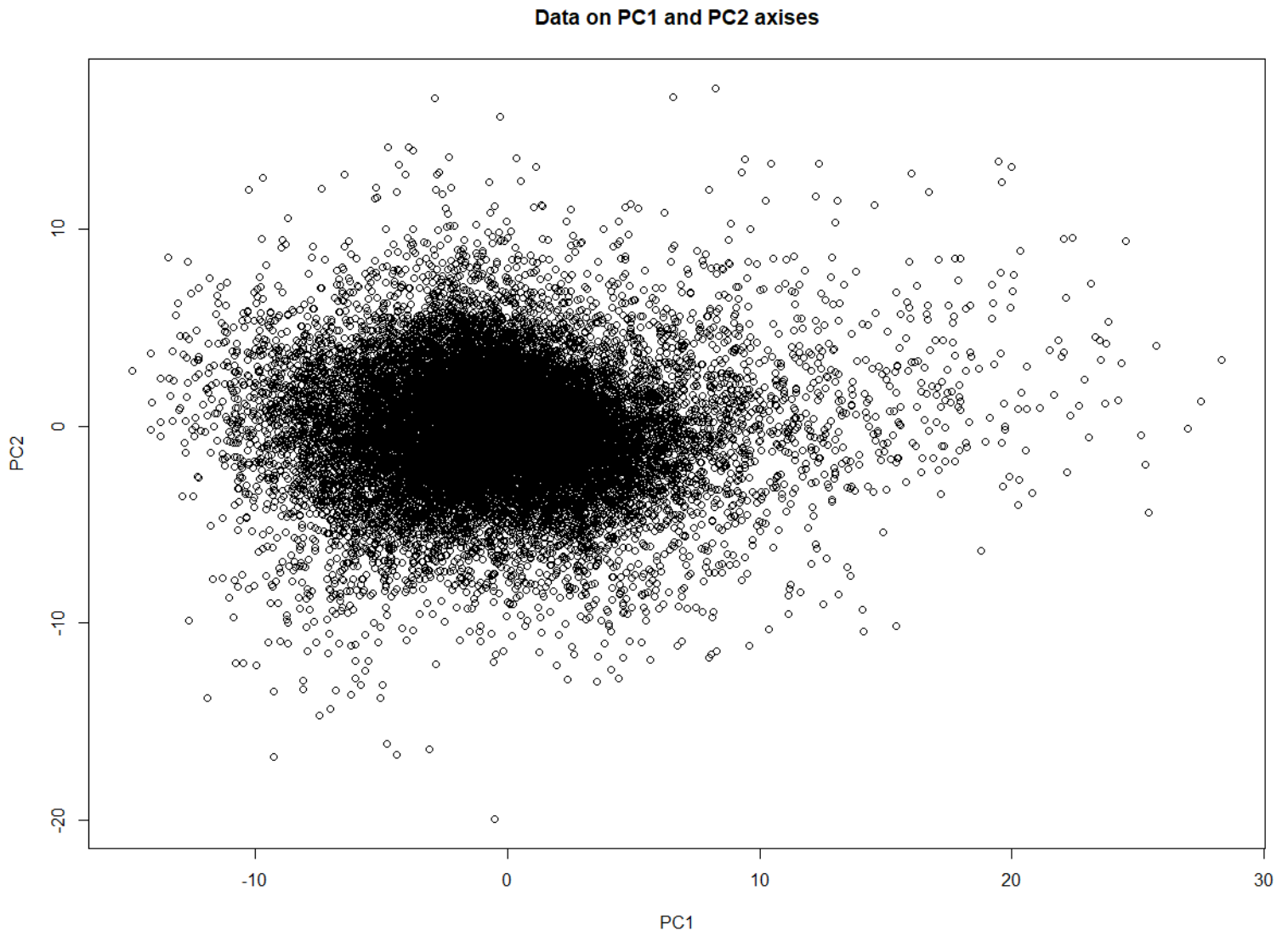


ولی باید داده‌ها را center کنیم تا تفاوت آن‌ها بهتر مشخص و جدا شود.



PCA Method

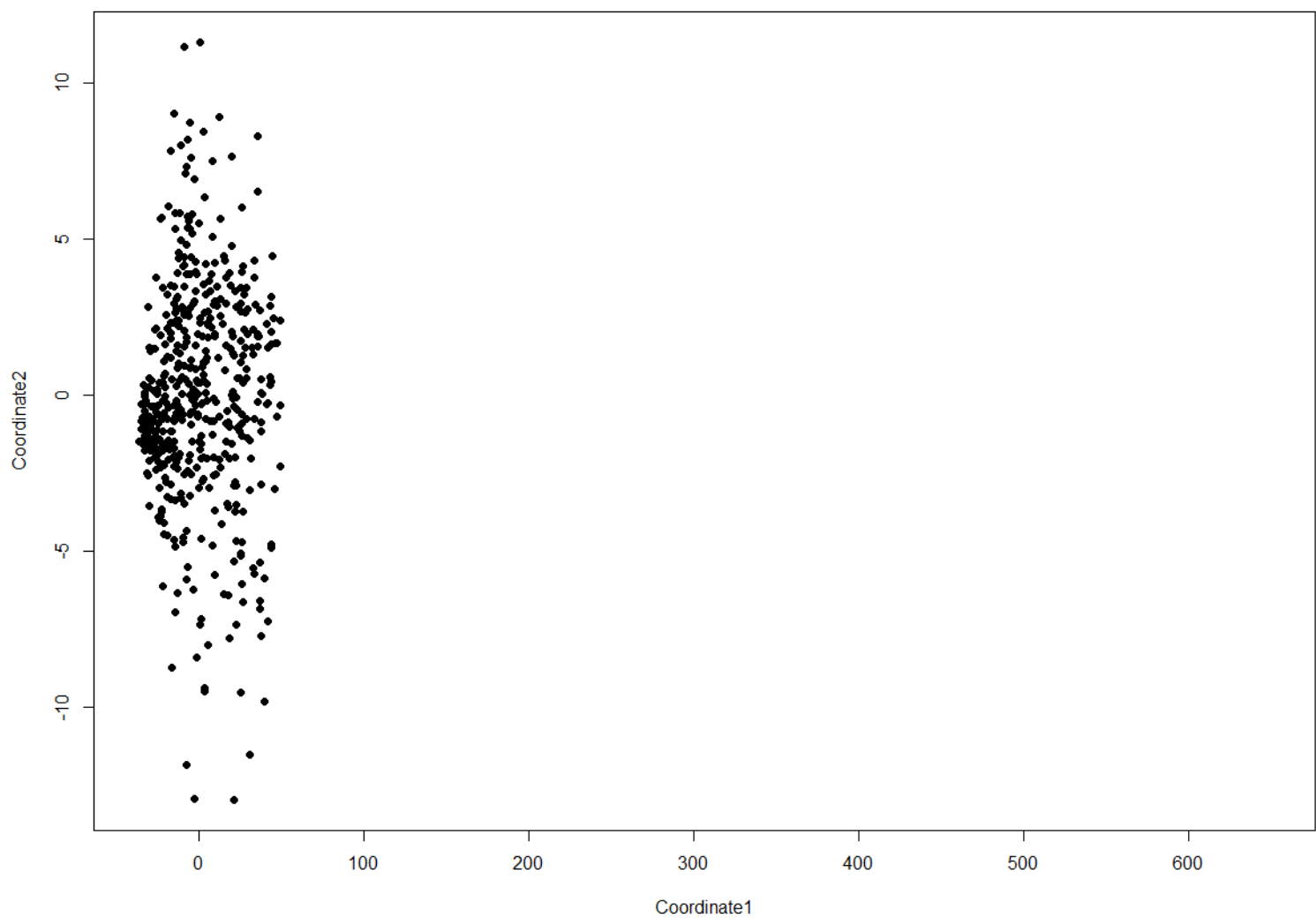




با این کار درصد اهمیت PC1 کمتر شده است و بهتر داده ها پخش شده اند.

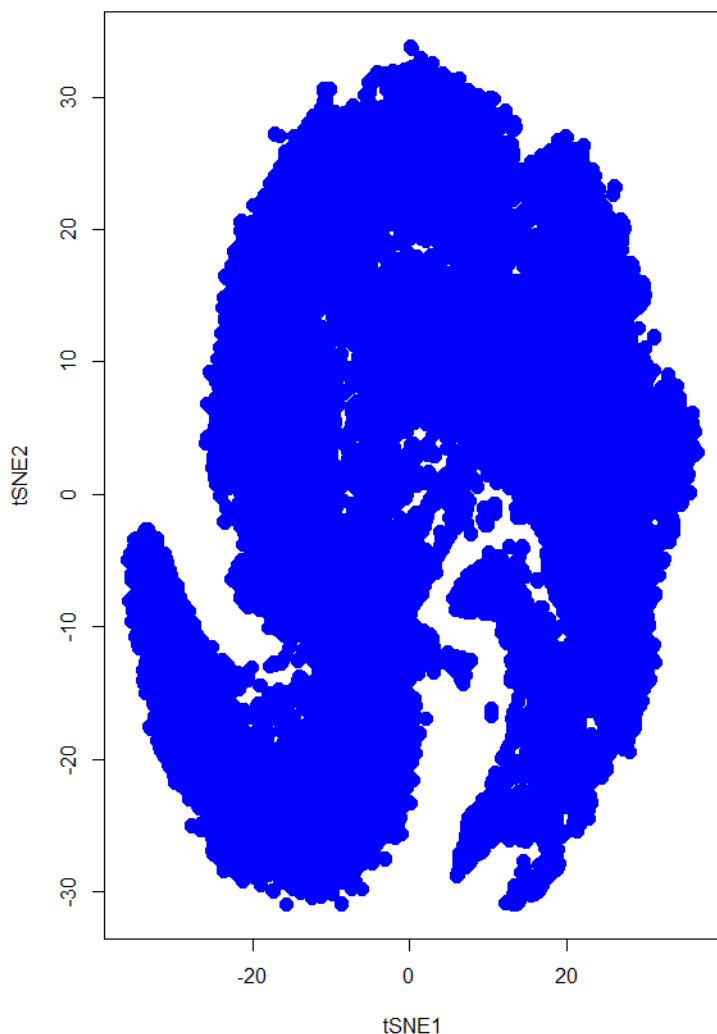
برای روش MDS خروجی کاهش ابعاد یافته داده ها به صورت زیر است:

MDS Method

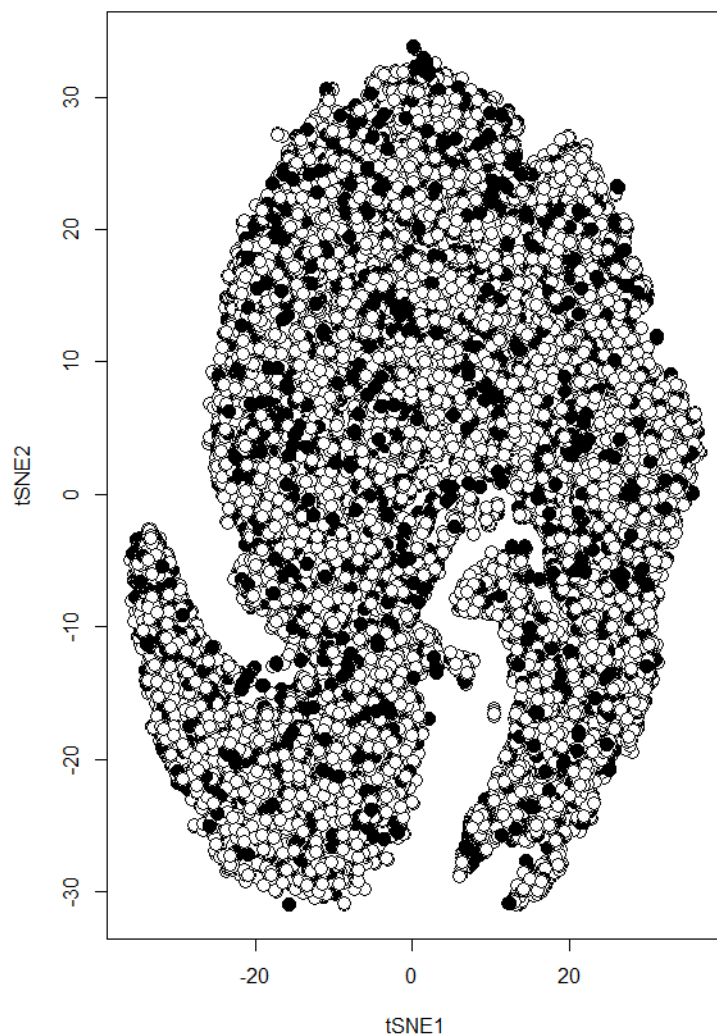


و در نهایت خروجی بخش tSNE به شکل زیر است:

tSNE Method



tSNE Method



در کل خروجی هر سه متد داده‌های کاهش ابعاد یافته هستند ولی هر کدام از یک روش خاص برای کاهش ابعاد استفاده می‌کنند. روش PCA سعی می‌کند که variance داده‌ها را در بعد کاهش یافته حفظ کند و ابعاد را به گونه‌ای کاهش دهد که این واریانس بیشترین حالت ممکن باقی بماند. از آن طرف روش MDS سعی می‌کند فاصله بین جفت نقطه‌ها را در فضای کاهش یافته حفظ کند. اما روش tSNE کاملاً هدف متفاوتی دارد و سعی می‌کند که همسایگی داده‌ها را حفظ کند. یعنی آن نقطه‌هایی که در فضای اصلی نزدیک به هم بودند، در فضای کاهش بعد یافته هم سعی می‌شود نزدیک به هم بمانند. لذا از tSNE برای expose کردن clustering ها و از MDS برای زمانی که رابطه global مورد بحث باشد و از PCA برای کاهش ابعاد با حفظ فاصله‌ها و واریانس و حذف نویزها استفاده می‌شود.

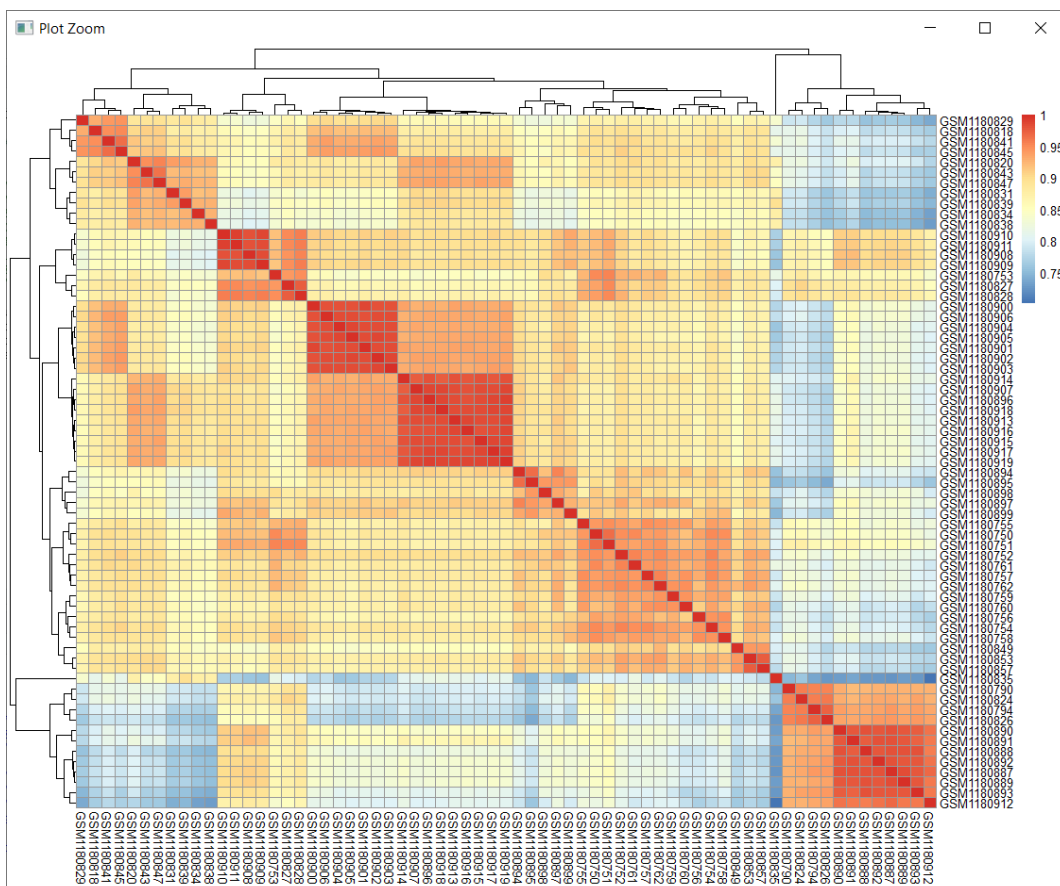
در نهایت با توجه به داده‌های ما و نتیجه‌گیری از نمودارها، و هدفی که داریم، همان PCA مناسب تر از بقیه روش‌های کاهش ابعاد است. همچنین این روش به کاهش نویز داده هم کمک می‌کند و دلیل دیگر استفاده از این روش این موضوع است. این روش به feature selection در مراحل بعدی هم کمک می‌کند و باعث بالا رفتن دقت می‌شود و همچنین feature هایی که تولید می‌شوند مستقل و uncorrelated هستند. با این روش کاهش ابعاد، مقداری از دیتا از بین می‌رود ولی با توجه به مزیت‌های آن تصمیم گرفتیم که از این روش استفاده کنیم.

۴. اگر دقت کنید source name نمونه های نرمال مختلف با یکدیگر تفاوت دارند. این فیلد بیانگر چیست؟

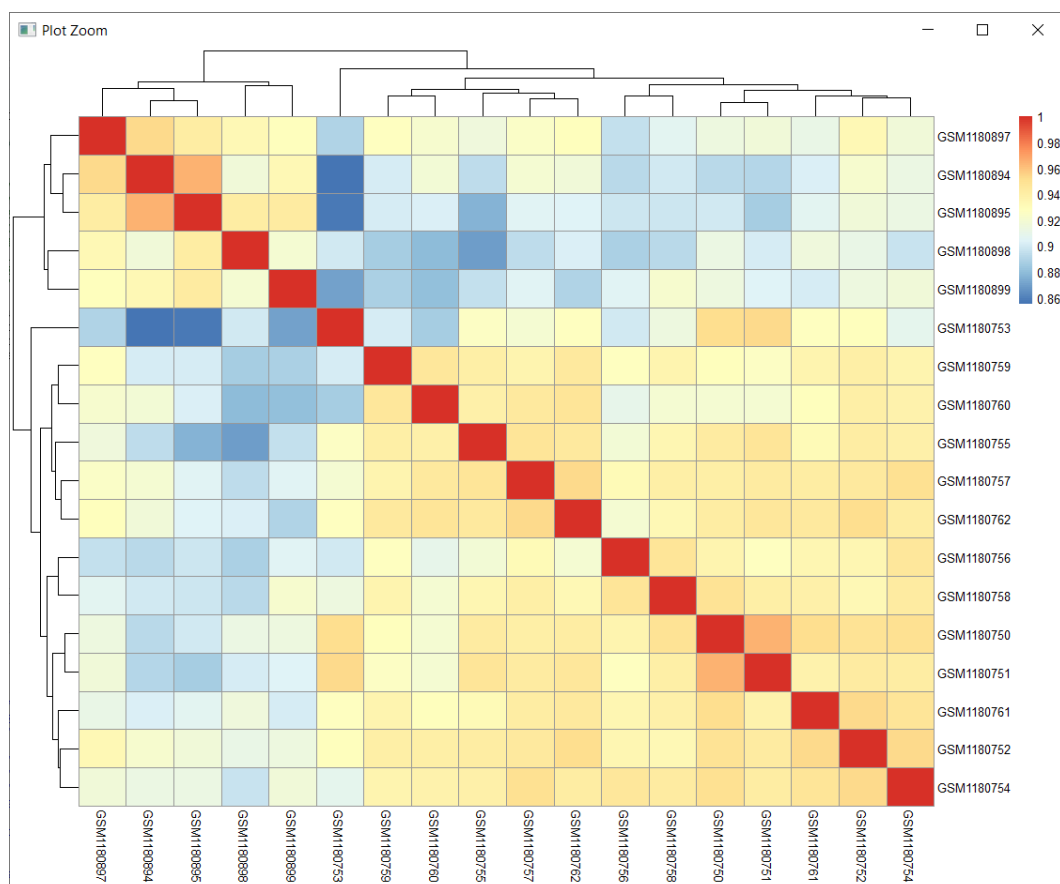
اگر داده ها را بر اساس source name گروه بندی کنیم (همه ی نمونه های بیمار در یک گروه هستند و هر نمونه ی سالم در گروه source name متناظر با خود است) همبستگی بین گروه ها با هم را بررسی کنید و به صورت یک نمودار نمایش دهید. گروهی از داده های سالم که بیشترین همبستگی با نمونه های گروه بیمار دارند را از روی نمودار مشخص کنید تا در مراحل بعد از این گروه برای تحلیل های بعدی استفاده شود. به نظر شما لزوم انجام این مرحله چیست؟ (۳۰)

این فیلد نشان دهنده منشأ داده است و بیانگر این است که sample از کجا گرفته شده است. برای مثال در بین source name های موجود برای نمونه های نرمال عبارت های T cell و B cell و Monocytes را می توان در جدول sample ها مشاهده کرد که هرکدام یک نوع خاص از سلول خونی سفید هستند. یکی دیگر از عبارت هایی که می توان مشاهده کرد CD34+HSPC است که یکی از کلیدی ترین سلول ها در تشخیص سرطان است چرا که یک سلول بنیادی است و در مغز استخوان قرار دارند و قابلیت تبدیل شدن و mature شدن به سلول های خونی را دارند. در مورد روش جداسازی این سلول های خاص باید گفت که به sample های خونی فلورسنت می زنند و با استفاده از روش FACS یا Fluorescence Activated Cell Sorting این سلول های خاص را جدا می کنند. این سلول های جدا شده سلول های خاصی هستند چرا که بیشترین شباهت را به سلول های سرطانی دارند و معمولاً در مغز استخوان موجود هستند. پس Source name در واقع source آن sample را مشخص می کند که از چه سلولی استفاده شده است. همانطور که مشاهده می شود، نمونه های CD34 خالص تر هستند و range بیان آن ها هم بسیار نزدیک تر است و به همین دلیل، کیفیت تحلیل بالاتر می رود. در مورد بقیه منشأها هم هر چه سلول ها متفاوت تر باشند، واریانس بیشتر می شود و در نتیجه احتمال بالاتر رفتن و غیر معنادارتر شدن p.value هم بیشتر می شود. تنها موردی که در مورد CD34 مطرح است، این است که تعداد نمونه های آن کمتر است و همین موضوع درجه آزادی را کاهش می دهد. در ادامه همبستگی را بین گروه های مختلف به شکل heatmap نمایش می دهیم تا مبنای تصمیم گیری خود را بیان کنیم.

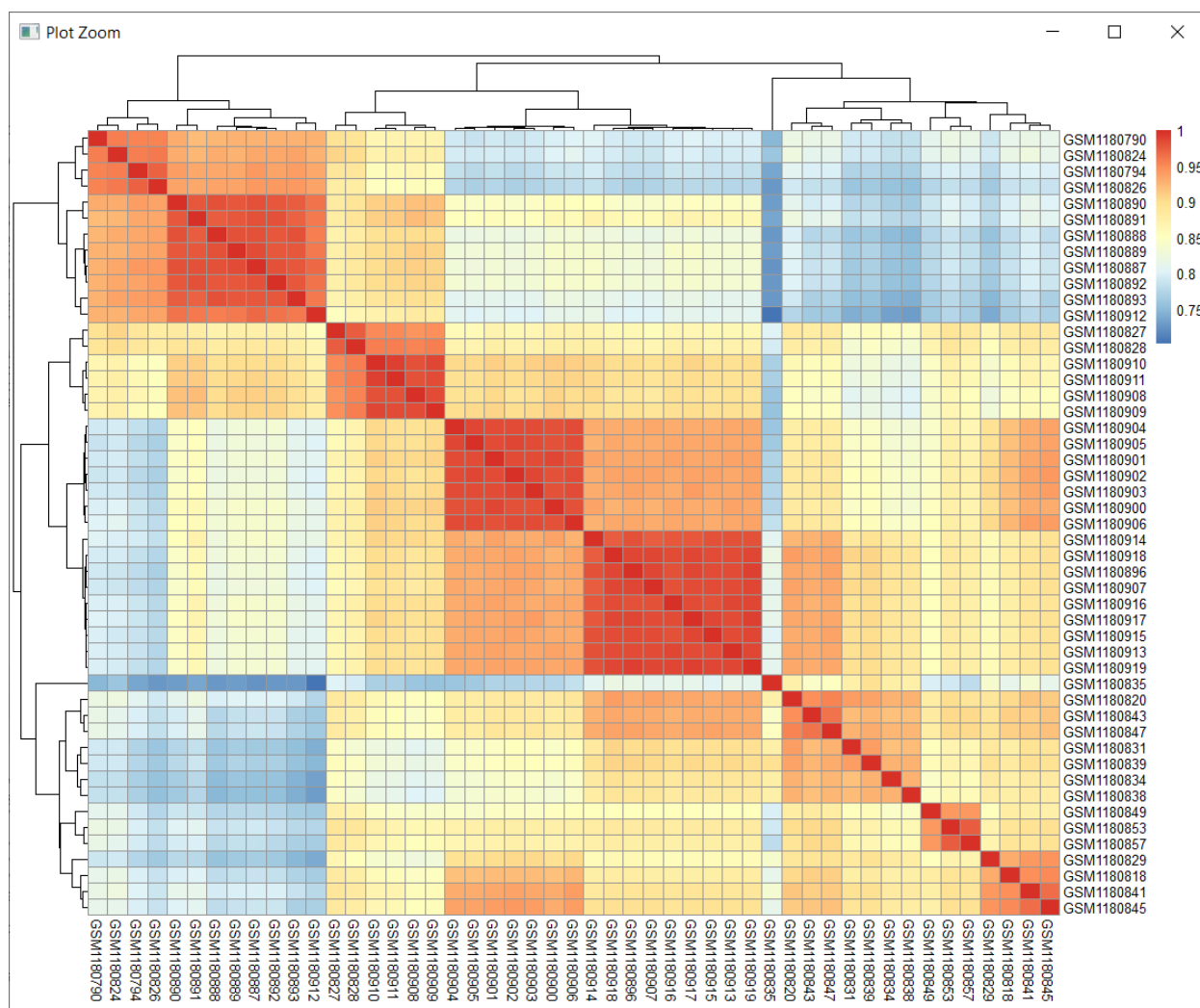
بین کلیه‌ی نمونه‌ها:



فقط بین داده های سرطانی:

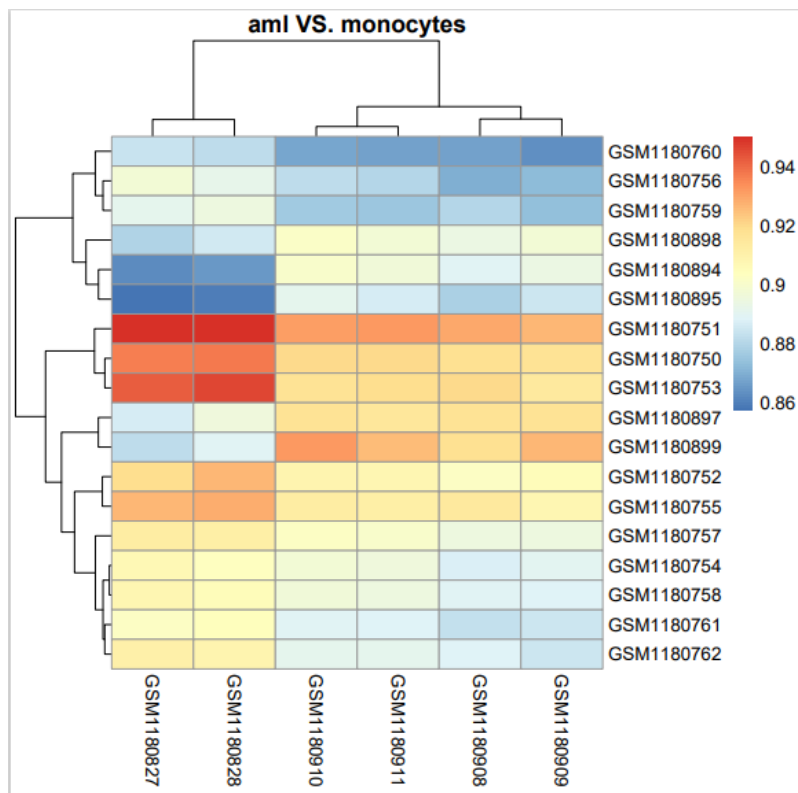
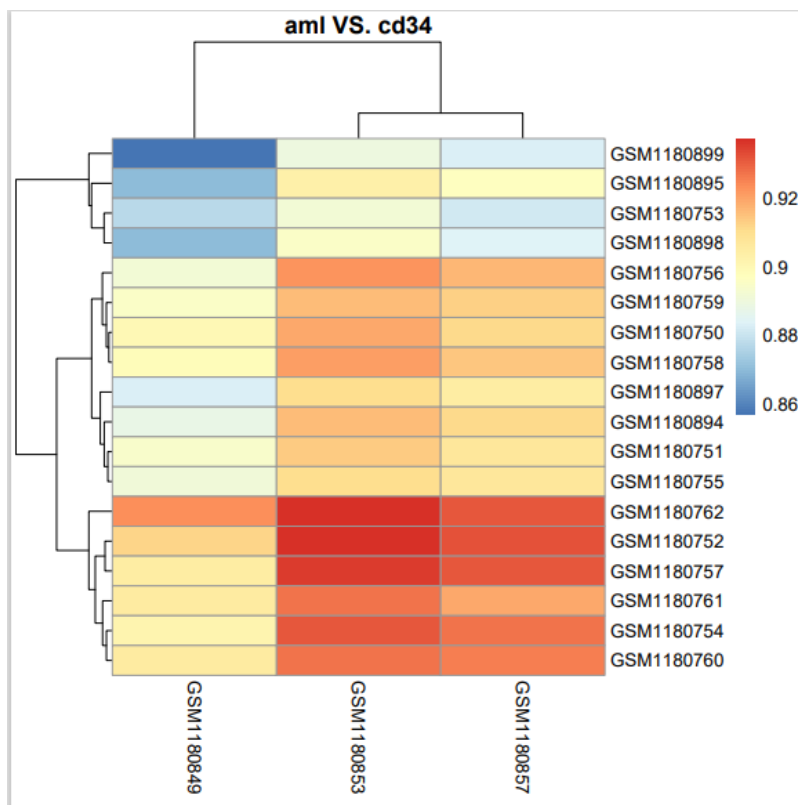


بین همه سالم ها هم به شکل زیر است:



خروجی مقایسه 2 به 2 گروهها را در فایل PDF ای تولید می‌کنیم و با اجرا کردن بخش آخر کد می‌توانید به آن‌ها دست پیدا کنید. ولی دو مورد از آن‌ها را اینجا قرار می‌دهیم تا توضیحات را روی آن‌ها بیان کنیم.





در سوال خواسته شده است که گروهی که بیشترین شباهت را با داده‌های AML دارد بیابیم تا در ادامه تحلیل‌ها را بر مبنای آن‌ها بگذاریم. به عنوان مثال با مقایسه دو نمونه بالا متوجه می‌شویم که میزان همبستگی بین CD34 با AML بسیار بیشتر از همبستگی Monocytes با AML است، زیرا میزان رنگ قرمز و زرد و نارنجی که بیانگر همبستگی بالاتر است، در Heatmap مربوط به CD34 بیشتر دیده می‌شود و به همین ترتیب می‌توان همه گروه‌ها را مقایسه کرد و نتیجه نهایی که به نظرمان آمد، این بود که گروه CD34 بیشترین همبستگی را دارد.