فاز اول

پروژه مقدمه‌ای بر بیوانفورماتیک

دکتر علی شریفی‌زارچی و دکتر سمیه کوهی

دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

نیم‌سال اول ۰۱-۰۲

امیرحسین باقری - ۹۸۱۰۵۶۲۱
مهدی مستانی - ۹۷۱۰۰۵۱۳
محمدرضا مفیضی - ۹۸۱۰۶۰۵۹

فهرست مطالب

۱ ریزآرایه چیست؟

ریزآرایه [1]، ابزاری آزمایشگاهی است که برای تشخیص بیان هزاران ژن به طور همزمان استفاده می‌شود. ریزآرایه‌های DNA لام‌های میکروسکوپی هستند که با هزاران نقطه کوچک در موقعیت‌های مشخص چاپ می‌شوند و هر نقطه حاوی یک توالی DNA یا ژن شناخته‌شده است.

روش کار

برای انجام تحلیل ریزآرایه، مولکول‌های mRNA معمولاً از هر دو نمونه آزمایشی و نمونه مرجع جمع‌آوری می‌شوند. به عنوان مثال، نمونه مرجع را می‌توان از یک فرد سالم، و نمونه آزمایشی را می‌توان از یک فرد مبتلا به بیماری مانند سرطان جمع‌آوری کرد. سپس دو نمونه mRNA به DNA مکمل (cDNA) تبدیل می‌شوند و هر نمونه با یک ترکیب فلورسنت [2] با رنگ متفاوت برچسب‌گذاری می‌شود. مثلا، نمونه آزمایشی cDNA ممکن است با رنگ فلورسنت قرمز برچسب‌گذاری شود، در حالی که cDNA مرجع با رنگ فلورسنت سبز برچسب‌گذاری می‌شود.

سپس دو نمونه با هم مخلوط شده و اجازه داده می‌شود تا به لام ریزآرایه متصل شوند. فرآیندی که در آن مولکول‌های cDNA به ترکیب‌های DNA روی لام متصل می‌شوند، هیبریداسیون [3] نامیده می‌شود. پس از هیبریداسیون، ریزآرایه برای اندازه‌گیری میزان بیان هر ژن چاپ‌شده روی لام اسکن می‌شود. اگر بیان یک ژن خاص در نمونه آزمایشی بیشتر از نمونه مرجع باشد، نقطه مربوطه روی ریزآرایه قرمز به نظر می‌رسد. از طرفی، اگر بیان در نمونه آزمایشی کمتر از نمونه مرجع باشد، آن نقطه سبز به نظر می‌رسد. در نهایت، اگر میزان بیان در دو نمونه یکسان باشد، نقطه زرد خواهد بود. داده‌های جمع‌آوری‌شده از طریق ریزآرایه‌ها را می‌توان برای ایجاد پروفایل‌های بیان ژن، که تغییرات همزمان در بیان بسیاری از ژن‌ها در پاسخ به یک بیماری یا درمان خاص را نشان می‌دهد، استفاده کرد. [۱]

فرمت داده‌های خروجی

مجموعه داده‌های ریزآرایه معمولاً بسیار بزرگ هستند و فرمت داده‌های خروجی به صورت یک فایل خام (Raw Matrix) در قالب یک متن tab-seperated حاوی داده‌های بیش از یک سنجش ترکیبی (ترکیب‌ها در سطر و نتایج آزمایش‌ها در ستون) است. در تصویر ۱ میزان بیان هر ژن به صورت heatmap نمایش داده شده است.

همچنین داده‌های ریزآرایه در دیتاست با توجه به شکل ۲ قابل دسترسی خواهد بود.

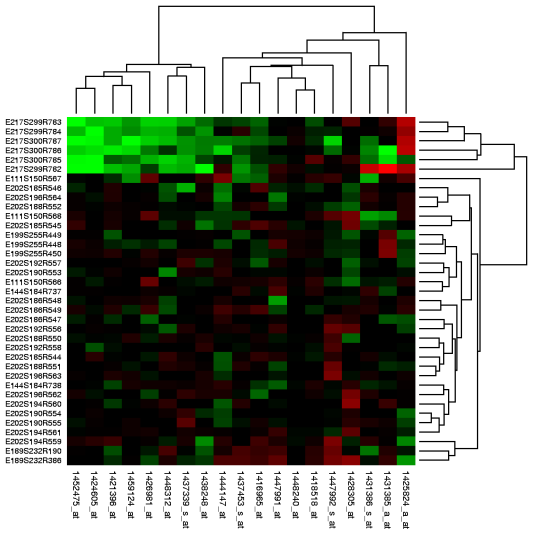
۲ کیفیت داده‌ها

در ابتدا بررسی می‌کنیم که بیشترین و کمترین مقدار داده‌ها در ماتریس بیان چقدر است. مشاهده می‌شود که بیشترین، عددی برابر با 13.76154 و کمترین برابر با 1.611473 است که نتیجه گیری این است که داده‌ها نرمال است(اگر که بیشترین مقدار عددی بیشتر از ۱۰۰ بود از لگاریتم داده‌ها استفاده می‌کردیم تا نتیجه‌گیری‌های بهتری داشته باشیم). در ادامه نیز کد لازم در صورتی که داده‌ها نیاز به تبدیل به مقیاس لگاریتمی داشته باشیم، قرار داده‌شده‌است تا صورت نیاز، این کار انجام شود (برای مثال اگر اختلاف ابتدا و انتهای بازه داده‌ها بیشتر از ۵۰ باشد یا ۱ درصد انتهای داده‌های عددی بیشتر از ۱۰۰ باشد). هم‌چنین داده‌های بدون مقدار و داده‌های تکراری را حذف می‌کنیم. در انتها فقط بخشی از ستون‌های داده که اطلاعات بهتری به می‌دهند یعنی Accession و SourceName و Phenotype را نگه می‌داریم.

در ادامه به بررسی برخی از تحلیل‌های انجام‌شده روی داده‌ها می‌پردازیم.

برای بررسی‌های بیشتر در رابطه با آنالیز کیفیت داده‌ها، نمودار Adjusted P-value را بررسی می‌کنیم. همان‌طور که در نمودار ۳ مشاهده می‌شود، توزیع Adjusted P-value ها بدین صورت است که بخش قابل توجهی از آن‌ها در بازه 0 تا 0.05 قرار دارند که بیان‌گر این موضوع است که در بخش قابل از توجهی از داده‌ها تفاوت معنی‌داری وجود دارد که یعنی نمونه‌های انتخاب شده این ویژگی را دارند که تفاوت‌های لازم را نمایش دهند. از طرفی نیز برای مقادیر بزرگتر از 0.05 نیز توزیع تقریبا یکنواختی را مشاهده می‌کنیم که این موضوع کیفیت خوب داده‌ها را تایید می‌کند؛ زیرا که بسیاری از ژن‌ها تفاوت بیان قابل توجهی در نمونه‌های مریض و سالم ندارند و در هر دو حالت بیان تقریبا یکسانی دارند.

[1] microarray
[2] fluorescent
[3] hybridization



شکل ۱: میزان بیان ژن در ریزآرایه [۲]

سپس در نمودار ۴ چندک‌های [۴] نمونه با چندک‌های تئوری که از توزیع t-Student حاصل می‌شود را رسم می‌کنیم. مشاهده می‌شود که این نمودار تقریبا یک خط را توصیف می‌کند که یک حالت بسیار مناسب است که بیان‌گر این موضوع است که مقادیر مورد استفاده در نمونه از توزیع نظری پیش‌بینی شده تقریبا پیروی می‌کند.

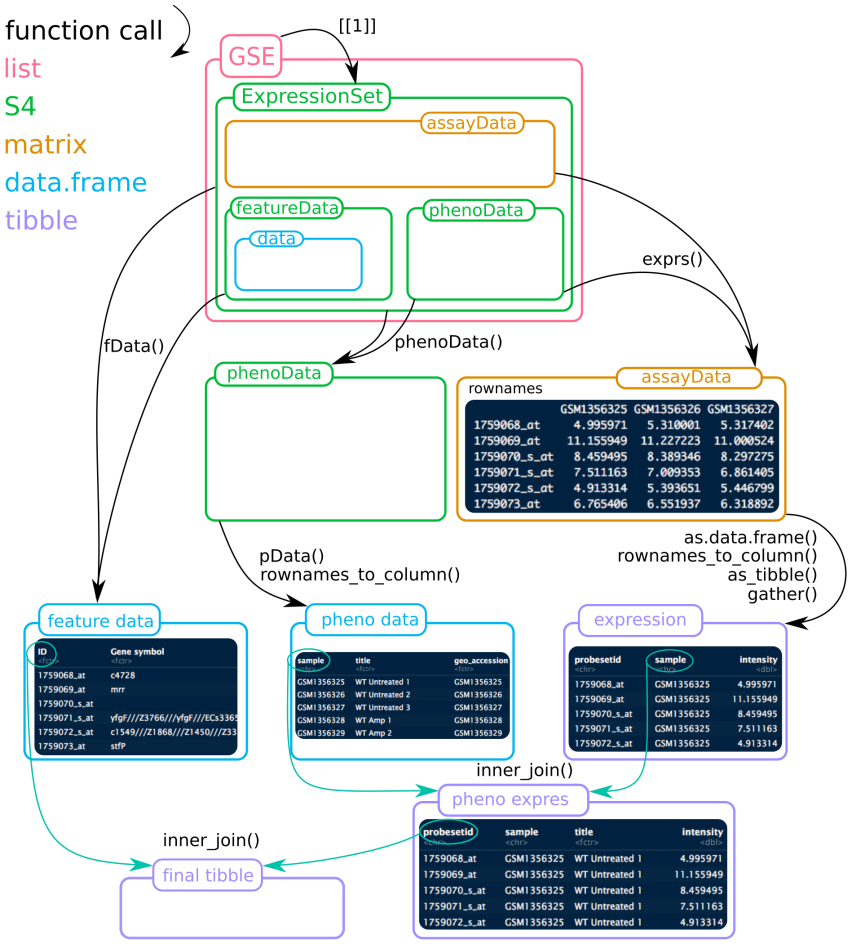
در مرحله بعد، نمودار آتشفشانی [۵] را بررسی می‌کنیم که این نمودار (شکل ۵) بدین صورت است که میزان اهمیت آماره (که در این آزمایش p-value است که البته برای نمایش بهتر آن از log p-value- استفاده می‌شود) در مقابل میزان تغییر دیتا (که در این قسمت نیز برای نمایش بهتر از lg fold change بهره می‌گیریم) مورد بررسی قرار می‌گیرد.

در این نمودار، داده‌هایی که p-value آن‌ها مناسب است (کمتر 0.05) ، آبی شده‌اند و از طرفی ژن‌های آبی‌رنگ در سمت راست نمودار، میزان بیان آن‌ها در نمونه سالم بیشتر از بیمار است و ژن‌های آبی‌رنگ در سمت چپ نمودار، میزان بیان آن‌ها در نمونه بیمار بیشتر از سالم است و هر چه قدر به سمت راست و چپ نمودار حرکت کنیم، این تفاوت بسیار بیشتر و قابل توجه‌تر می‌شود . این نمودار بیان می‌کند که بخشی از ژن‌ها شرط p-value را برقرار نمی‌کنند و برخی از دیتا ها که این شرط را برقرار می‌کند، fold change آن‌ها چندان قابل توجه نیست ولی تعدادی از داده‌ها که مقادیر p-value آن‌ها بسیار پایین است و نیز fold change آن‌ها مقادیر مثبت نسبتا زیاد دارند (نقاط بالا و سمت راست نمودار) و یا مقادیر منفی نسبتا زیاد دارند (نقاط بالا و سمت چپ نمودار) را می‌توان به‌طور ویژه بررسی نمود و در مجموع نمودار بیان می‌کند که کیفیت داده‌ها در مرحله قابل قبولی قرار دارد.

در قسمت بعدی نمودار اختلاف میانگین [۶] را بررسی می‌کنیم. این نمودار (شکل ۶) تا حدودی مشابه نمودار آتشفشانی عمل می‌کند با این تفاوت که $\log_2$ میانگین بیان ژن‌ها را در آن در نظر می‌گیریم و با lg fold change مقایسه می‌کنیم. در این نمودار مقادیری که p-value آن‌ها مناسب است (کمتر 0.05) است و نیز $\log FC$ آن‌ها مثبت است را قرمز و آن‌هایی که منفی است را آبی کرده‌ایم.

در مرحله بعد، نمودار جعبه‌ای را مورد بررسی قرار می‌دهیم. نمودار حاصل (شکل ۷) تقریبا نشان می‌دهد که میانه‌ی نمونه‌ها با یکدیگر برابر است که نشان‌گر این است که داده‌ها نرمال و قابل مقایسه هستند و از طرفی طول جعبه‌ها نیز این ویژگی را دارد . چارک‌ها و کمترین و بیشترین مقدار در هر یک از نمونه‌ها نیز تا حدودی با یکدیگر برابر است و از طرفی نیز بیشترین مقدار نمونه‌ها، عددی حداکثر ۱۵ است که یعنی داده‌ها در مقیاس لگاریتمی می‌باشد و نیازی به تبدیل آن‌ها نیست.

[4] quantiles
[5] volcano plot
[6] Mean Difference (MD) plot



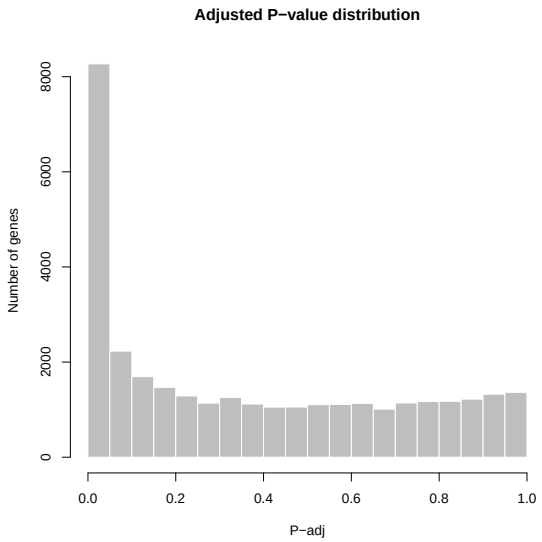
شکل ۲: ساختار داده‌ها در دیتاست [۳]

در بخش بعدی، نمودار توزیع چگالی بیان نمونه‌ها (شکل ۸) را بررسی می‌کنیم که در آن توزیع همه گروه‌های نمونه‌ها را با رنگ‌های متفاوتی در یک نمودار رسم می‌کنیم. مشاهده می‌کنیم که خم تمام گروه‌ها تقریبا بر یکدیگر منطبق است که بیان‌گر این است که نمونه‌ها نرمال و قابل مقایسه با هم هستند.

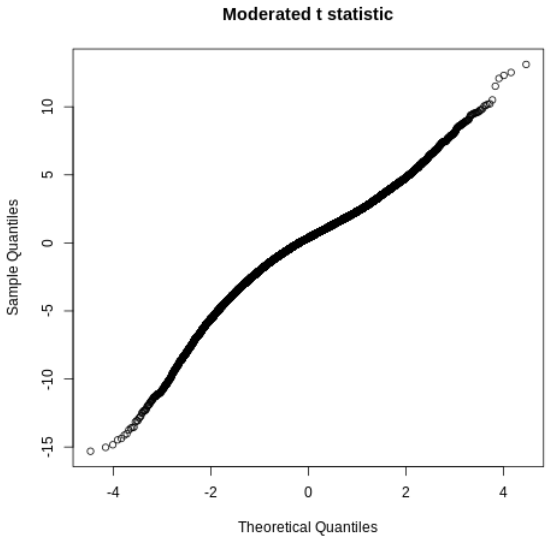
همچنین نمودار Mean-Variance (شکل ۹) برای بررسی نسبت واریانس به میانگین میزان بیان ژن رسم شده که می‌تواند نشان دهد داده‌ها چقدر از هم پراکنده هستند.

در انتها نیز از UMAP [۷] استفاده می‌کنیم که یک روش کاهش بعد داده است که برای نحوه ارتباط داده‌ها با یکدیگر استفاده می‌شود. در نمودار ۱۰ مشخص است که نمونه‌های سالم به چند دسته تقسیم شده‌اند که البته در هر دسته نمونه‌ها بسیار شبیه به یکدیگرند (تقسیم آن‌ها به چند دسته نیز به این علت است که در نمونه‌های سالم نیز چند دسته‌ی متفاوت داشتیم) و از طرفی نمونه‌های بیمار نیز در یک دسته قرار دارند که موکد این است که نمونه‌ها از کیفیت مناسبی برخوردار هستند. البته اگر که توجه نماییم، برخی از نمونه‌های سالم در دسته‌ی بیمار قرار گرفته‌اند که در واقع بیان‌گر این است که این نمونه‌ها شباهت زیادی به نمونه‌های بیمار دارند که این نکته می‌تواند یکی از روش‌های یافتن نمونه‌های شبیه به بیمار

[7] Uniform Manifold Approximation and Projection

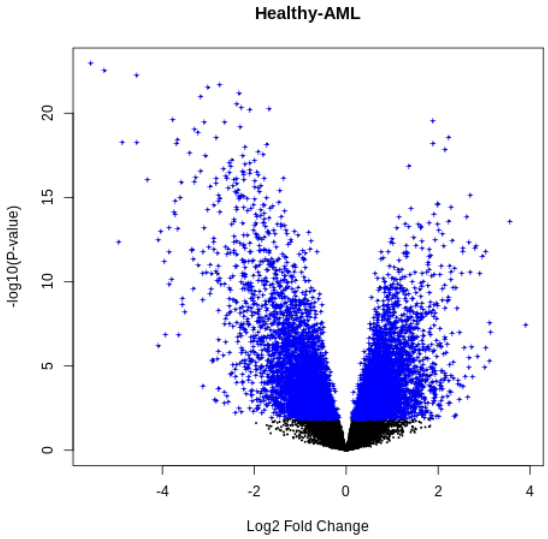


شکل ۳: نمودار Adjusted P-value

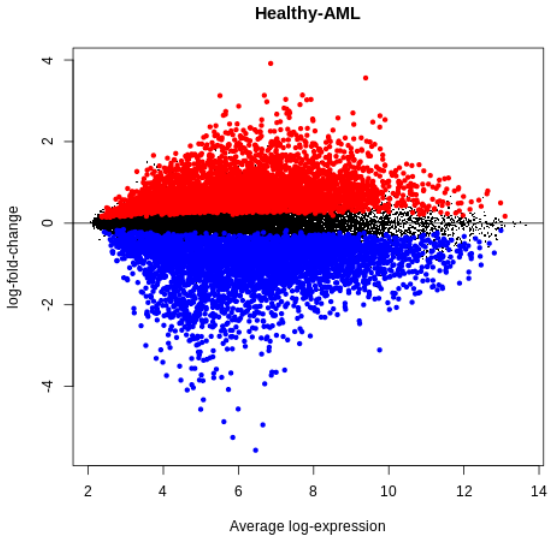


شکل ۴: نمودار چندک‌های نمونه

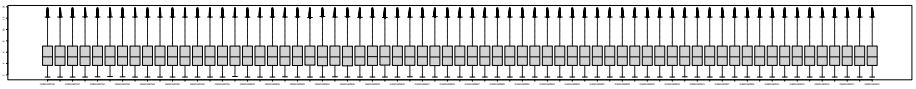
در قسمت ۴ام باشد.



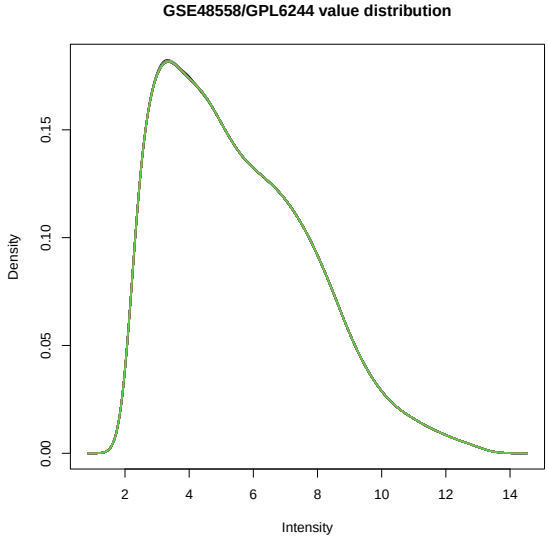
شکل ۵: نمودار آتشفشانی



شکل ۶: نمودار اختلاف میانگین



شکل ۷: نمودار حعبه‌ای



شکل ۸: نمودار توزیع چگالی بیان نمونه‌ها

۳ کاهش ابعاد داده‌ها

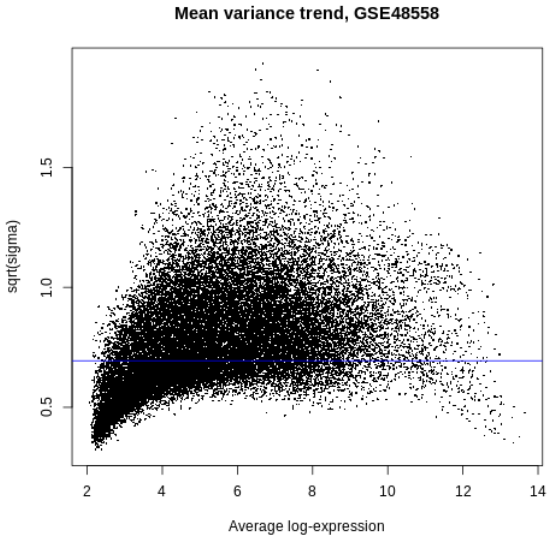
۱.۳ دلیل کاهش بعد داده‌ها

کاهش ابعاد یک بعد کلیدی در مطالعات بایواینفورماتیک است. که به ما قابلیت نمایش دیتا های پیچیده و همچنین امکان بررسی آماری آنها را می‌دهد. مطابق آنچه در درس دکتر شریفی گفته شد این کار برای ۲ مهم صورت می‌گیرد. زمانی که نتایج یک آزمایش در اختیار ما قرار می‌گیرد و ما می‌خواهیم یک مساله را بررسی کنیم. ممکن است که اشتباهی در فرایند جمع آوری داده رخ داده باشد ممکن است یک باکتری در محیط رشد کرده باشد و بناکردن نتایج تحقیقات بر روی یک داده ناصحیح ممکن است نتایج زیان باری داشته باشد بنابراین باید مطمعن شیم که داده های ما صحت دارند. برای این کار از کاهش بعد کمک می‌گیریم تا مطمعن شیم به طور مثال که آیا واقعا رابطه ای بین مساله و داده ها وجود ندارد یا اینکه داده ها صحت دارند و می‌توان کار را ادامه داد.

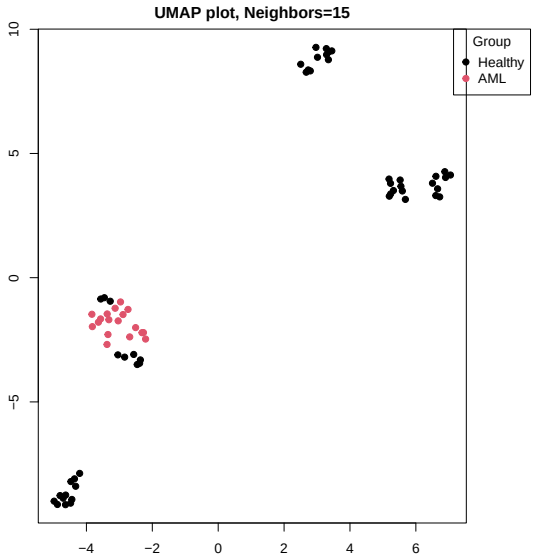
داده های بایوانفرماتیک عموما از پیچیدگی زیادی برخورداند بدین صورت که اکثر آنها تعداد نمونه ها از تعداد ویژگی ها بیشتر است با کاهش بعد داده ها می توان داده ها را به متغیر های جدیدی تقلیل داد به طوری که تفاوت داده ها (تفاوت می تواند با متریک های متفاوتی اندازه گیری شود) حفظ شود. تا بتوان داده هارا بررسی کرد.

PCA ۲.۳

یک کاهش بعد خطی است که در نهایت داده هارا در یک ماتریس دوران ضرب می‌شود. این کاهش بعد سعی می‌کند که در عین کاهش بعد بیشترین واریانس را حفظ کند. به دلیل اینکه به این پروژه مربوط نیست وارد توضیحات ریاضی آن نمی‌شویم و به همین توضیح بسنده می‌کنیم.

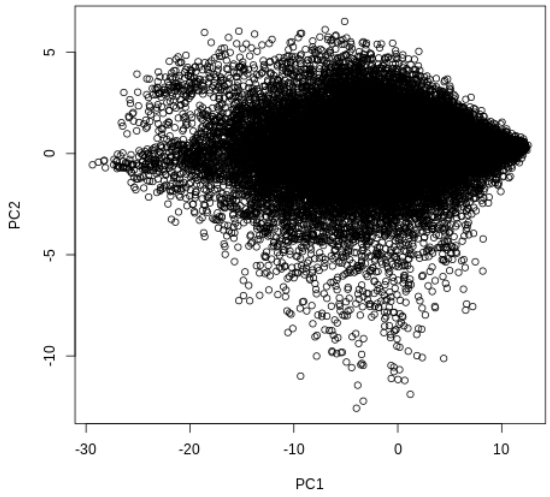


شکل ۹: نمودار Mean-Variance

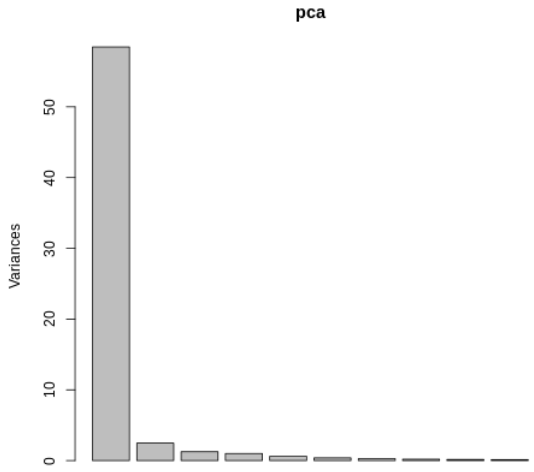


شکل ۱۰: نمودار UMAP

ابتدا داده هایی که scale نشده اند را بررسی می‌کنیم. همانطور که می‌بینید در این داده ها به دلیل اینکه اختلاف بیان ژن ها مورد بررسی نیست و خود ژن ها مورد بررسی هستند. در نمودار آن داده ها بیان ژن ها به صورت بیضوی است به معنای آنکه بیان بعضی ژن ها زیاد و بیان بعضی کم است و این چیزی نیست که ما دنبال آن می‌گردیم. (۱۱)
همچنین در این حالت میزان واریانسی که در هر component وجود دارد میزان زیادی از آنها در بعد اول جمع شده(که همین عامل بیضوی شدن شکل شده است.) (شکل ۱۲) حال داده ها را که در ماتریس چرخش ضرب شده اند. در

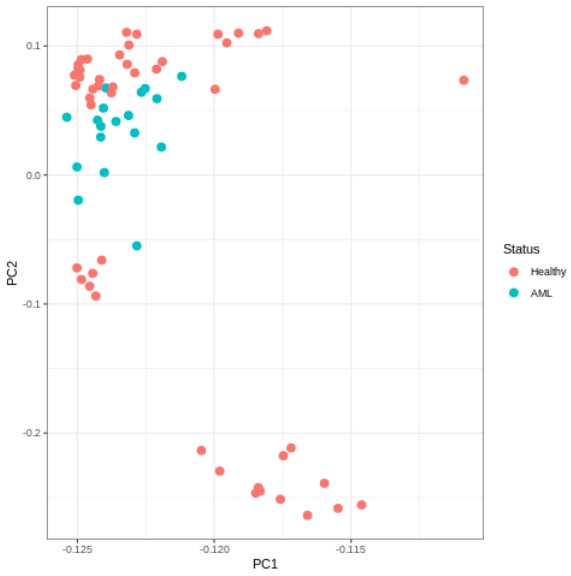


شکل ۱۱: میزان بیان هر ژن پس از کاهش بعد



شکل ۱۲: واریانس هر بعد

شکل ۱۳ مشاهده می‌کنید که جدایی پذیری کمی دارند.
حال داده های را scale می‌کنیم بدان معنا که میانگین هر سطور را از هر می‌کاهیم. بدین شکل تفاوت بیان ژن ها مشخص می‌شود.
همانطور که مشاهده می‌کنید واریانس موجود در هر component اکنون قابل مقایسه شده است. همچنین ژن ها در ۲ بعد از حالت بیضوی در آمده اند. ۱۴



شکل ۱۳: داده های کاهش بعد یافته

حال کاهش بعد یافته داده هارا پس از اعمال اسکیل بر روی آنان در شکل ۱۵ مشاهده می‌کنید. هماطور که مشاهده می‌شود در این حالت کاهش بعد بهتر عمل کرده و جدایی پذیری بیشتری بین گونه ها توانسته تشخیص دهد.

TSNE ۳.۳

(t-distributed Stochastic Neighbor Embedding) یک روش غیر خطی و احتمالاتی است که در سال ۲۰۰۸ معرفی گشته است. این روش قادر است که داده هارا به طوری کاهش بعد دهد که جدایی پذیری آنان بسیار عالی شود. به عنوان مثال فرض کنید که داده ها بر روی یک خم (manifold) قرار دارند. روش pca قادر به کاهش بعد آنان به صورتی که داده ها جدا سازی خوبی شده باشند ندارد اما TSNE می‌تواند این کار را انجام دهد.

به عنوان مثال داده های زیر در یک فضای ۳ بعدی بر روی ۲ خم قرار دارند. (شکل ۱۶)

که کاهش شکل یافته آنان در فضای ۲ بعدی با استفاده از tsne به صورت شکل ۱۷ در می‌آید.

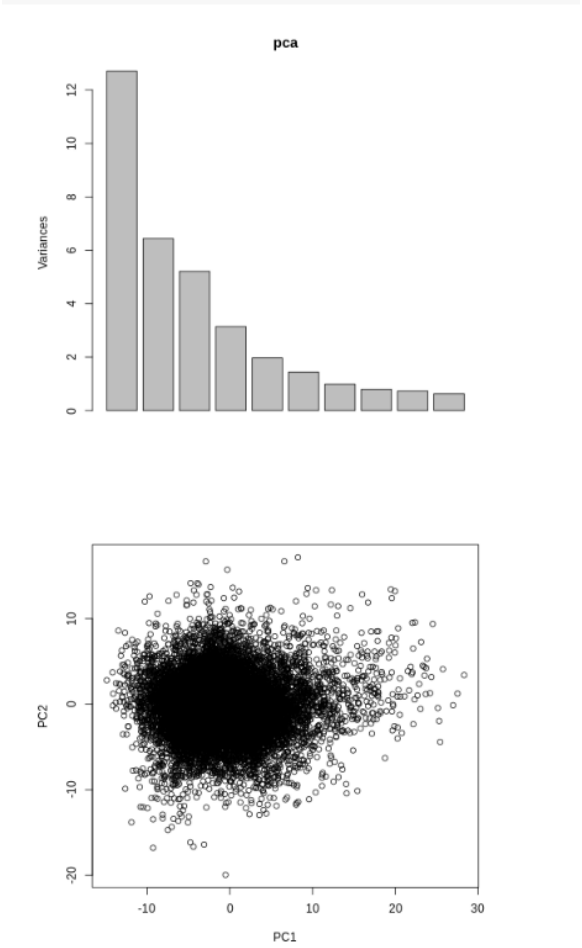
این الگوریتم یک معیار perplexity نیز دارد که به معنای زیر است. این مقدار برابر ۲ به توان آنتروپی شنون است. در واقع این متغیر برابر تعداد همسایه های موثر در الگوریتم است. در واقع می‌توان آن را به نوعی به تعداد همسایه های نزدیک در خم فضا ارجاع داد. هر چه دیتا پر چگال تر باشد اندازه این متغیر باید بیشتر باشد تا به نتایج بهتری دست یابیم. درواقع این معیار متناسب با اندازه دیتا ست خواهد شد.

در این بخش حال به کاهش بعد داده های اصلی می‌پردازیم. در شکل ۱۸ مشاهده می‌کنید که مقدار ۵ از همه بهتر عمل می‌کند. در نهایت با اندکی تغیر دادن فضای سرچ در پارانتر prep بهترین نتیجه با prep = 8 حاصل می‌شود. که در شکل ۱۹ می‌توانید مشاهده کنید.

MDS ۴.۳

این کاهش بعد سعی می‌کند که فاصله نسبی بین داده ها در بعد کمتر حفظ شود. دقت کنید که در pca هدف حفظ واریانس داده ها بود. در اینجا هدف حفظ فاصله نسبی دو به دو داده ها خواهد بود.

برای این کار ما ۳ روش از روش های موجود در R را انجام می‌دهیم.



شکل ۱۴: نتایج داده های اسکیل شده

۱.۴.۳ Principal Coordinates Analysis

نتایج در شکل ۲۰

۲.۴.۳ Stress Minimization: SMACOF

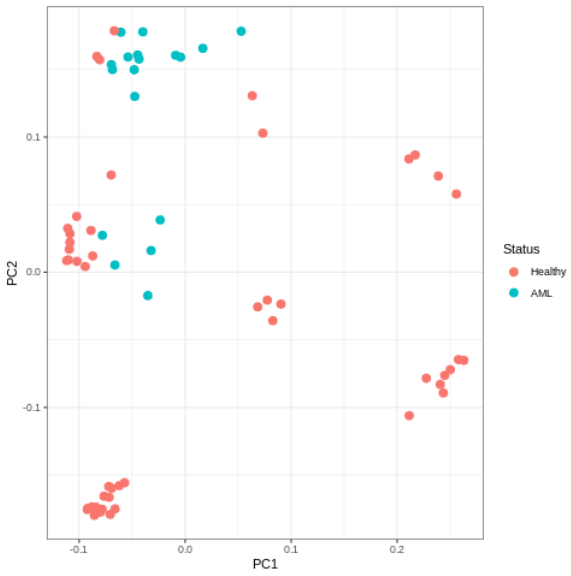
نتایج در شکل ۲۱

۳.۴.۳ Sammon

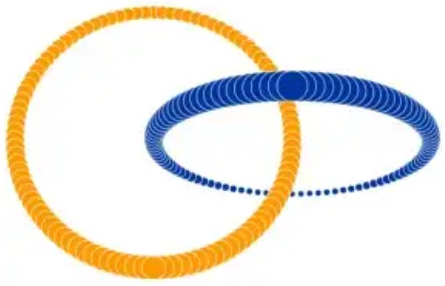
نتایج در شکل ۲۲

۵.۳ انتخاب بهترین کاهش بعد

از آنجا که هدف ما کاهش بعد داده ها به صورتی است که تفاوت معناداری بین سمپل های بیمار و سالم حاصل شود. در واقع اگر بخواهیم یک مدل لرنینگ یا هر معیار تصمیم گیری دیگری قرار دهیم تا بتواند به ما کمک کند بهتر است



شکل ۱۵: کاهش بعد یافته داده های اسکیل شده



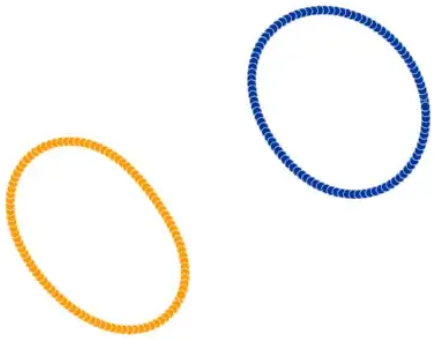
شکل ۱۶: داده ها در فضای ۳ بعدی

که تفاوت یا margin بین داده ها بیشتر شود که در حالت TSNE with prep = 5 از بقیه کاهش بعد ها بهتر عمل می‌کند. زیرا هم داده ها تفاوت معنا داری دارند. هم داده های AML در این حالت رفتار یک توزیع پارامتری دارند.

۴ همبستگی بین گروه‌ها

هرکدام از انواع Source Name در واقع نشان‌دهنده نوع خاصی از سلول‌ها را نشان می‌دهد که از نمونه مورد نظر گرفته شده است.

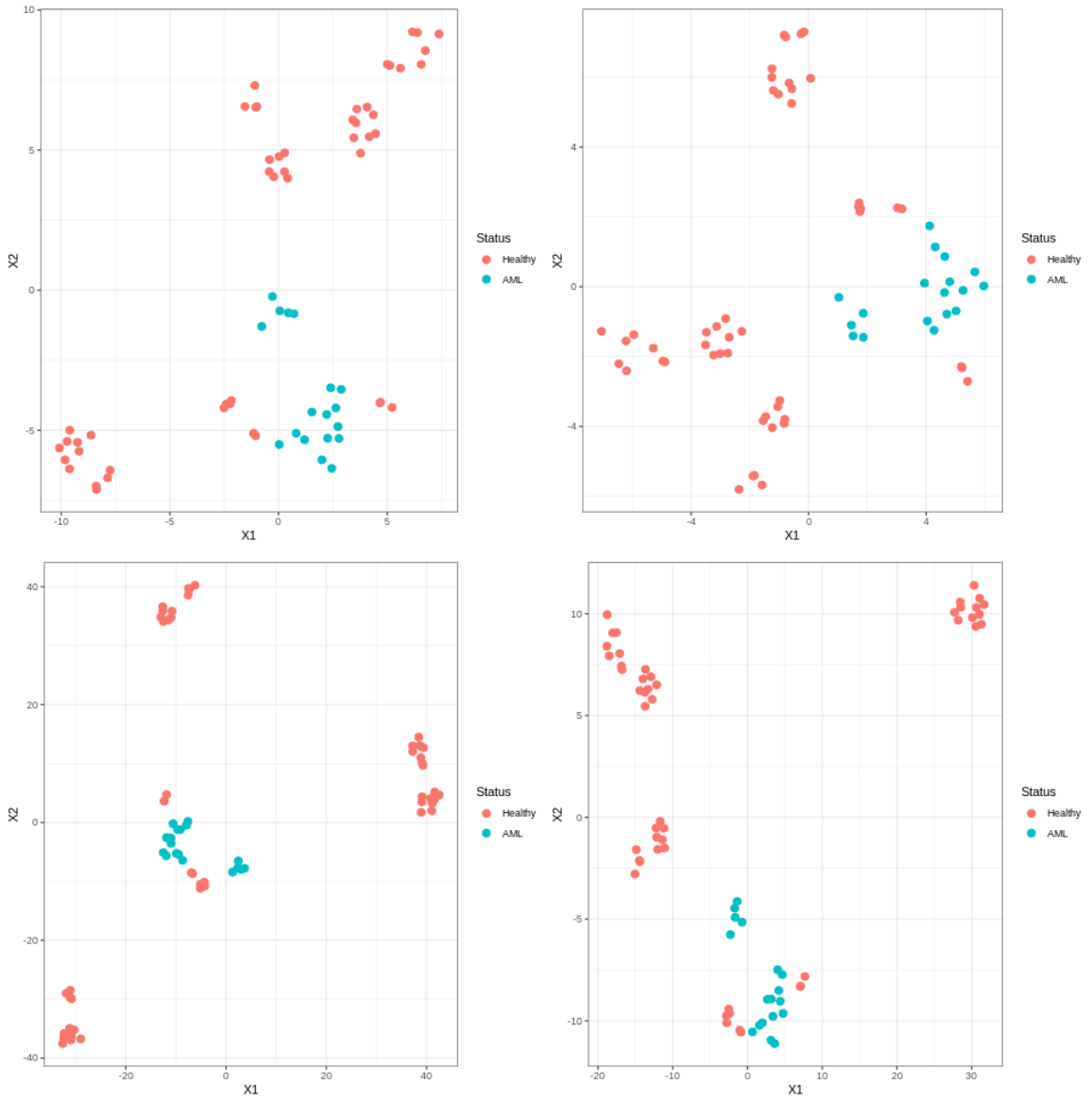
نمودار هم‌بستگی بین گروه‌های مختلف در شکل ۲۳ رسم شده است. با توجه به نمودار هم‌بستگی، گروه Granulocytes



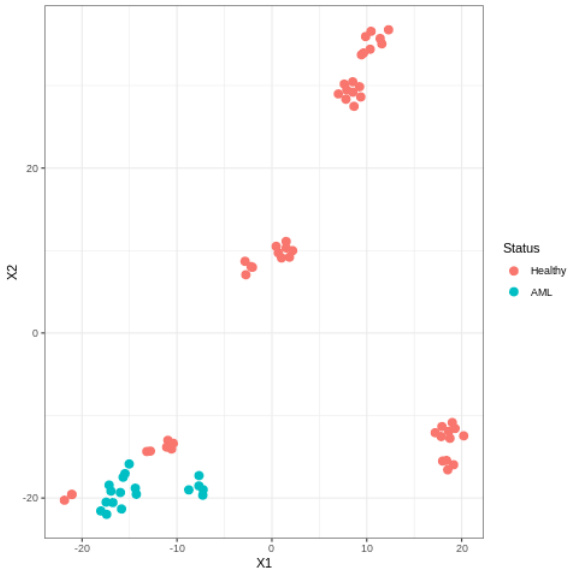
شکل ۱۷: کاهش بعد یافته آنان در فضای ۲ بعدی به کمک tsne

هم‌بستگی نسبتا پایینی با گروه نمونه‌های بیمار دارد. و از طرفی نمونه‌های T Cells و B Cells هم‌بستگی قابل توجهی با نمونه‌های بیمار دارند.

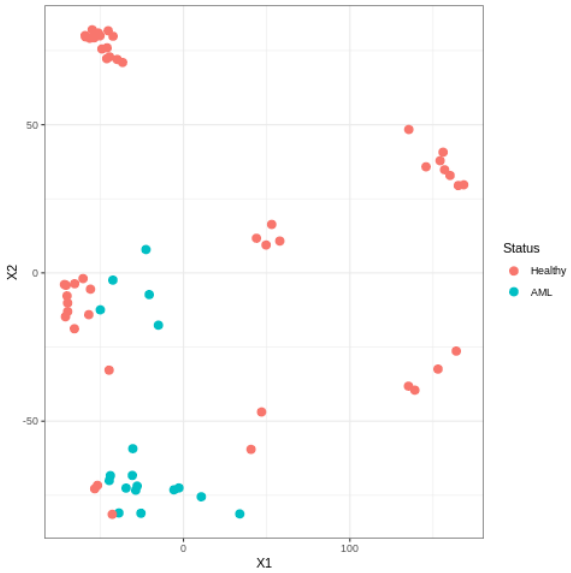
لزوم انجام این مرحله شناسایی گروهی از نمونه‌های سالم است که شباهت بیشتری با گروه بیمار دارند. شناسایی بیماران از این گروه نمونه‌های سالم و بررسی آن‌ها ما را به نتایج ارزشمندتری می‌رساند (زیرا گروه دیگر نمونه‌های سالم تفاوت فاحشی با گروه بیمار دارند که تصمیم‌گیری براساس آن‌ها لزوما نتیجه خوبی نخواهد داشت). هم‌چنین می‌دانیم که هرچه سلول‌ها در هر گروه به هم شبیه‌تر باشند، تحلیل بهتر خواهد بود.



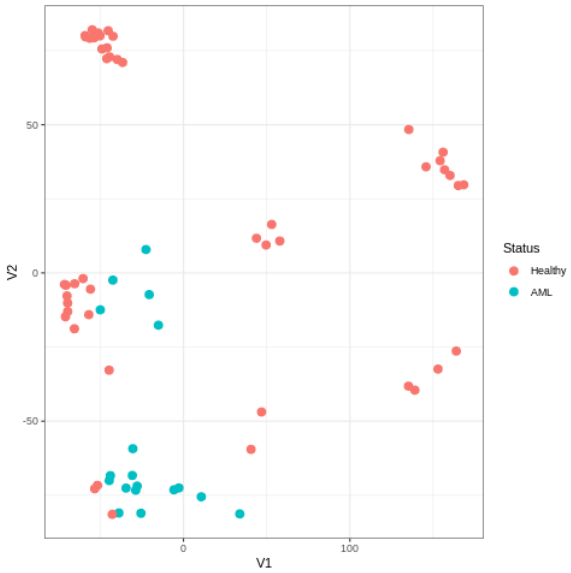
شکل ۱۸: کاهش بعد داده ها به ترتیب با مقادیر perplexity ۵ ۱۰ ۱۵ ۲۰



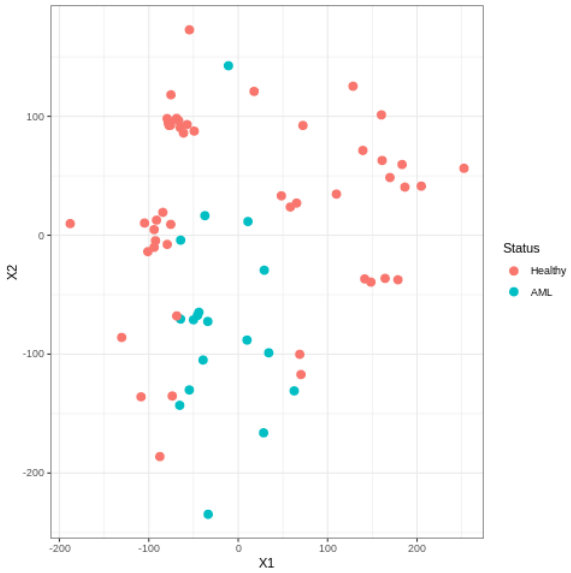
شکل ۱۹: کاهش بعد با prep8=8



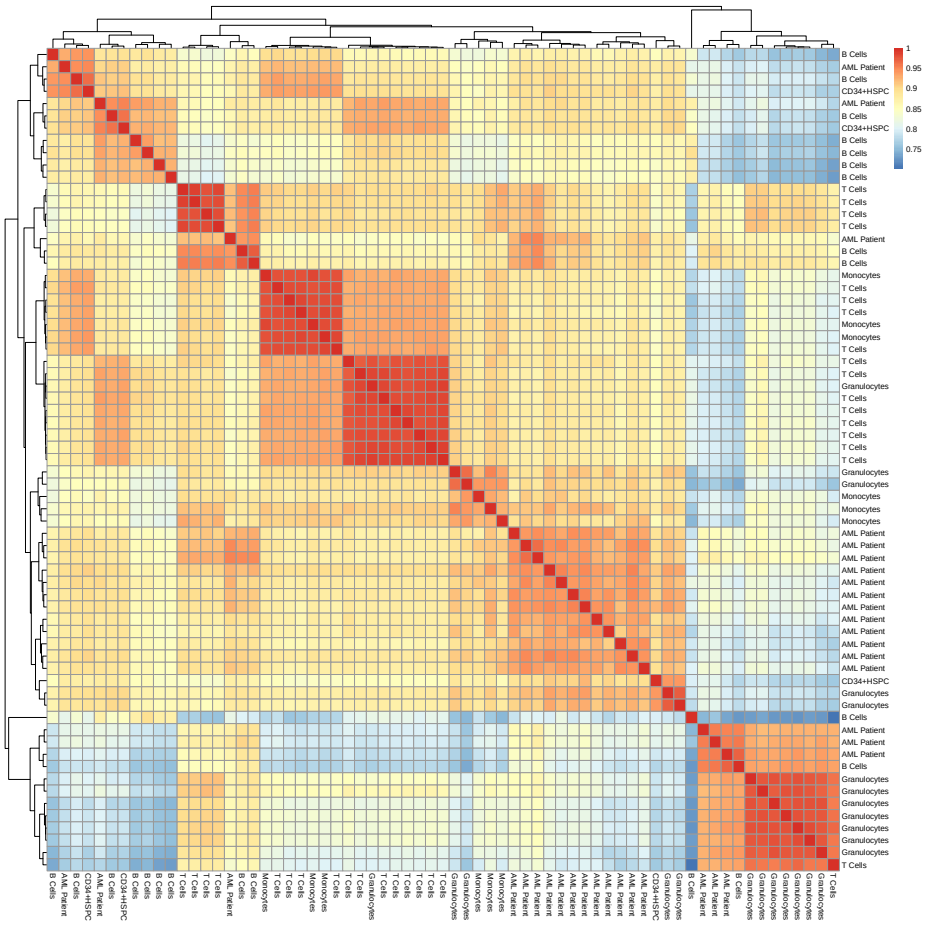
شکل ۲۰: کاهش بعد با pco



شکل ۲۱: کاهش بعد با SMACOF



شکل ۲۲: کاهش بعد با sammon



شکل ۲۳: میزان هم‌بستگی نمونه‌ها با یک‌دیگر

مراجع

[1] Nature Defenition Microarray (2014), Nature Education, https://www.nature.com/scitable/definition/microarray-202

[2] Microarray Hitmap (2006), Wikipedia, https://commons.wikimedia.org/wiki/File:Heatmap.png#/media/File:Heatmap.png

[3] GEO study, R tidyverse, https://lsru.github.io/tv_course/TD_project_solution.html