به نام خدا

گزارش پروژه

بیوانفورماتیک

فاز اول

علی ماهانی

۹۷۱۱۰۸۶۳

علی قبله

۹۹۱۰۹۹۷۱

مقدمه

سرطان خون یا لوسمی[1] یا لوکمیا بیماری پیشرونده و بدخیم اعضای خون ساز بدن است. این بیماری در اثر تکثیر و تکامل ناقص گلبول‌های سفید خون و پیش سازهای آن در خون و مغز استخوان ایجاد می‌شود. لوسمی یکی از سرطان‌های شایع در میان کودکان است. در لوسمی مغز استخوان به صورت غیر عادی، مقدار بسیار زیادی سلول خونی تولید می‌کند. این سلول‌ها با سلول‌های خون نرمال و عادی متفاوت هستند و درست عمل نمی‌کنند. در نتیجه، تولید گلبول‌های سفید خون طبیعی را متوقف کرده و توانایی فرد را در مقابله با بیماری‌ها از بین می‌برند. سلول‌های لوکمی همچنین بر تولید سایر انواع سلول‌های خونی که توسط مغز استخوان ساخته می‌شود از جمله گلبول‌های قرمز خون که اکسیژن را به بافت‌های بدن می‌رسانند، و پلاکت‌های خونی که از لخته شدن خون جلوگیری می‌کنند، اثر منفی می‌گذارند لذا در انواع لوسمی ضعف ایمنی، کم خونی و اختلال انعقاد خون داریم.

لوسمی حاد میلوئیدی[2] یا به اختصار ***AML*** یکی از انواع سرطان خون است. این نوع لوکمی سلول‌های مغز استخوان یا میلوسیت[3]ها را تحت تأثیر قرار می‌دهد و روندی حاد دارد. در این بیماری مغز استخوان، میلو بلاست‌ها (نوعی گلبول سفید)، گلبول‌های قرمز یا پلاکت‌های غیرطبیعی می‌سازد.

در انتهای پرسش های این بخش، با توجه به نتایج می‌بینیم که لوکمی میلوسیت‌ها را تحت تاثیر قرار می‌دهد و داده های ما با دانش پیشیین درست کار می‌کنند.

روند انجام کار[4]

در ابتدا لازم به ذکر است که تمام برنامه با زبان ***R*** نوشته شده است. در ابتدا کتابخانه های مورد نیاز را بارگزاری می‌کنیم. سپس لازم است که ماتریس ***Annotation*** را استخراج کنیم. داده‌ها را از طریق لینک داده شده، وارد می‌کنیم و گروه‌بندی میان ۲ دسته سالم و بیمار را انجام می‌دهیم. برای صحت سنجی بایستی دو بخش را بررسی کنیم، ابتدا باکس پلات را رسم می‌کنیم. طبق داده هایی که داریم، باکس پلات به شکل خوبی نمایش داده شده است و نیاز به نرمال کردن ندارد. در بخش دوم، نمودار ***Heat map*** را ترسیم می‌کنیم. نرمال کردن داده‌ها به طور کلی باعث می‌شود که توزیع آن‌ها مانند هم شود تا در مراحل بعدی بررسی‌ها دقیق‌تر انجام شود اما در داده های ما این مسئله نیاز نبود چرا که از پیش داده‌ها نرمال شده بودند و لازم نبود که آن‌ها را دوباره نرمال کنیم. بخش دوم نیز برای آن است که بتوانیم ارتباط میان گروه‌های متفاوت را بررسی کنیم تا بتوانیم نسبت ارتباط میان گروه های بیمار و سالم را نشان دهیم و در صورت وجود کورلیشن های اشتباه بتوانیم آن را تغییر دهیم. تغییر داده در این بخش باعث رخداد یک ***Heat map*** از ***Correlation*** بهتر میان گروه ها می‌شود و میتوان جداسازی را بهتر انجام داد.

---

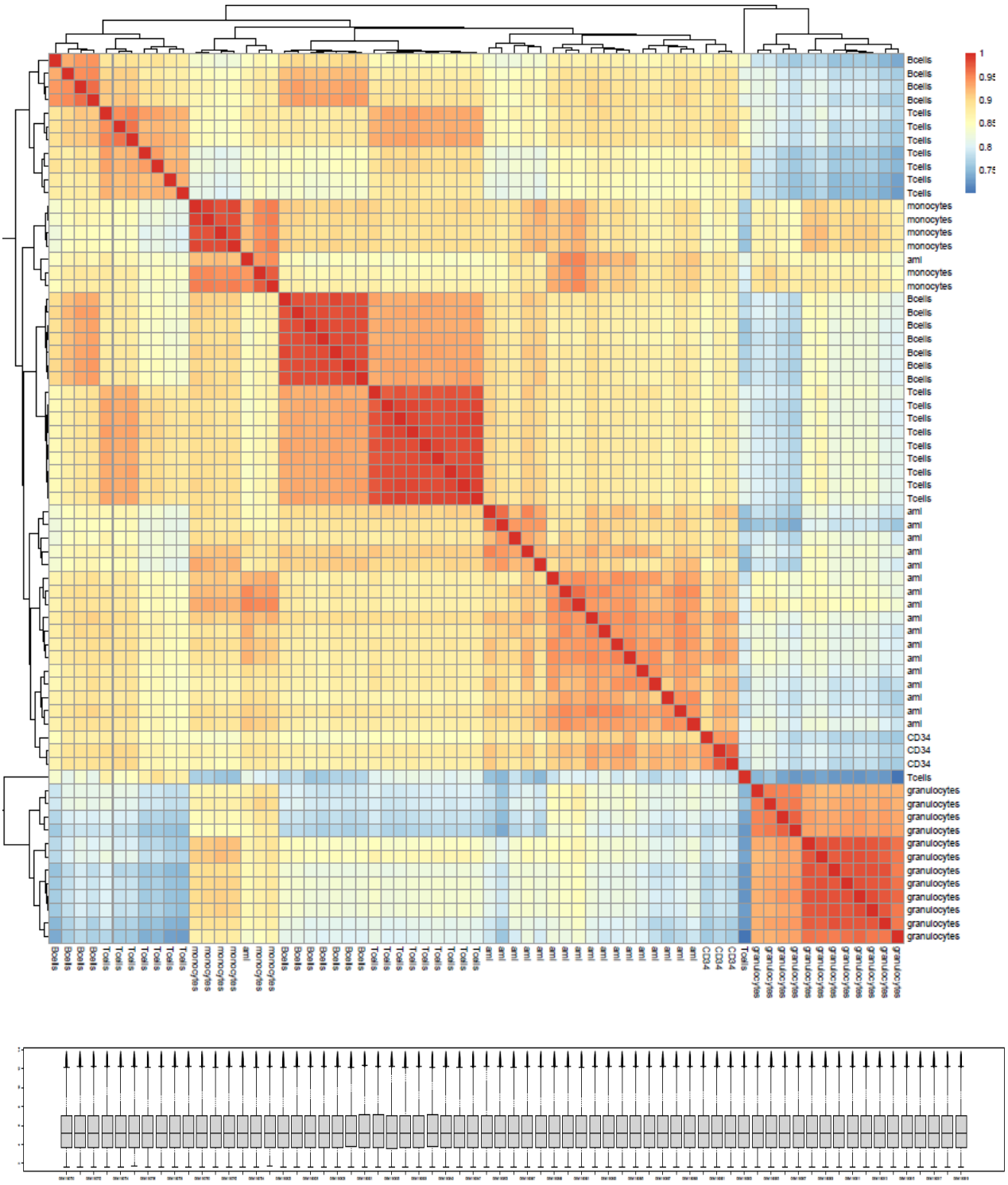
[1] Leukemia
[2] Acute Myeloid Leukemia
[3] Myelocyte
[4] با توجه به ویدئو های ۹ تا ۱۲

پس از مشخص شدن داده‌ها و صحت سنجی آن‌ها، لازم است ابعاد آن‌ها را کاهش دهیم. با استفاده از سه روش *PCA*، *MDS* و *tSNE* ابعاد داده‌ها را کاهش می‌دهیم. نتایج را در ادامه می‌توان مشاهده کرد. لازم به ذکر است که الگوریتم های *PCA* و *MDS* با استفاده از *Pattern* های خطی اجزای اصلی را جداسازی می‌کنند. برای مثال روش *PCA*، با گرفتن ماتریس و قطری کردن آن و سپس چینش ویژه مقادیر به صورت نزولی، می‌تواند ماتریس انتقالی را بدست آورد شامل ویژه بردارها است که هر ویژه بردار مربوط به همان ویژه مقدار ستون می‌باشد. با استفاده از همین شیوه، دو گروه مورد بررسی را جداسازی می‌کند. روش *tSNE* بر *Pattern* های غیر خطی متکی است و در درون خود الگوریتم رندوم را دارا است برای همین پس از هر بار اجرای برنامه نمودار های متفاوتی به ما می‌دهد اما نتیجه و میزان جداسازی آن در کل یکی است و تنها شکل نمودار متفاوت است. استفاده از این روش در مقابل دو روش دیگر در داده‌هایی که *Pattern* غیر خطی دارند، نتایج بهتری را شامل می‌شود.

سوالات

۱. *Microarray* تکنولوژی است که از یک چیپ حاوی تعداد زیادی پیکسل تشکیل شده و در هر پیکسل یک توالی تک رشته ای از *DNA* متصل شده بر روی یک سطح سیلیکونی. این تکنولوژی به ما این امکان را می‌دهد که از دو نمونه‌ی مختلف *RNA* استخراج کنیم، از آن قطعه های کوچک *cDNA* بسازیم و آنها را روی این چیپ ها قرار می‌دهیم. سپس این چیپ ها را زیر دستگاه قرار می‌دهیم و مشاهده می‌کنیم که هر چیپ چقدر رنگ فلوروسانت دارد. دستگاه اسکنر پس از اسکن کردن چیپ، به ما یک تصویر می‌دهد و آن را پیش پردازش می‌کند. خروجی پیش پردازش به صورت یک ماتریس از اعداد است که به تعداد سمپل ها ستون و به تعداد پروب ها سطر دارد. (ممکن است بعضی ژن ها در یک *microarray* خاص پروب نداشته باشند یا بیش از یک پروب به ازای آیزومورف های مختلف داشته باشند.)

۲. با توجه به توضیحات داخل متن گزارش، ۲ صحت سنجی صورت گرفته است که نتایج به صورت زیر است و تغییری در داده‌ها صورت نگرفته است. لزوم این کنترل‌ها در پاراگراف اول بخش روند کار، توضیح داده شده است. تصویر اول، *Heat map* و تصویر دوم *Box Plot* می‌باشد.

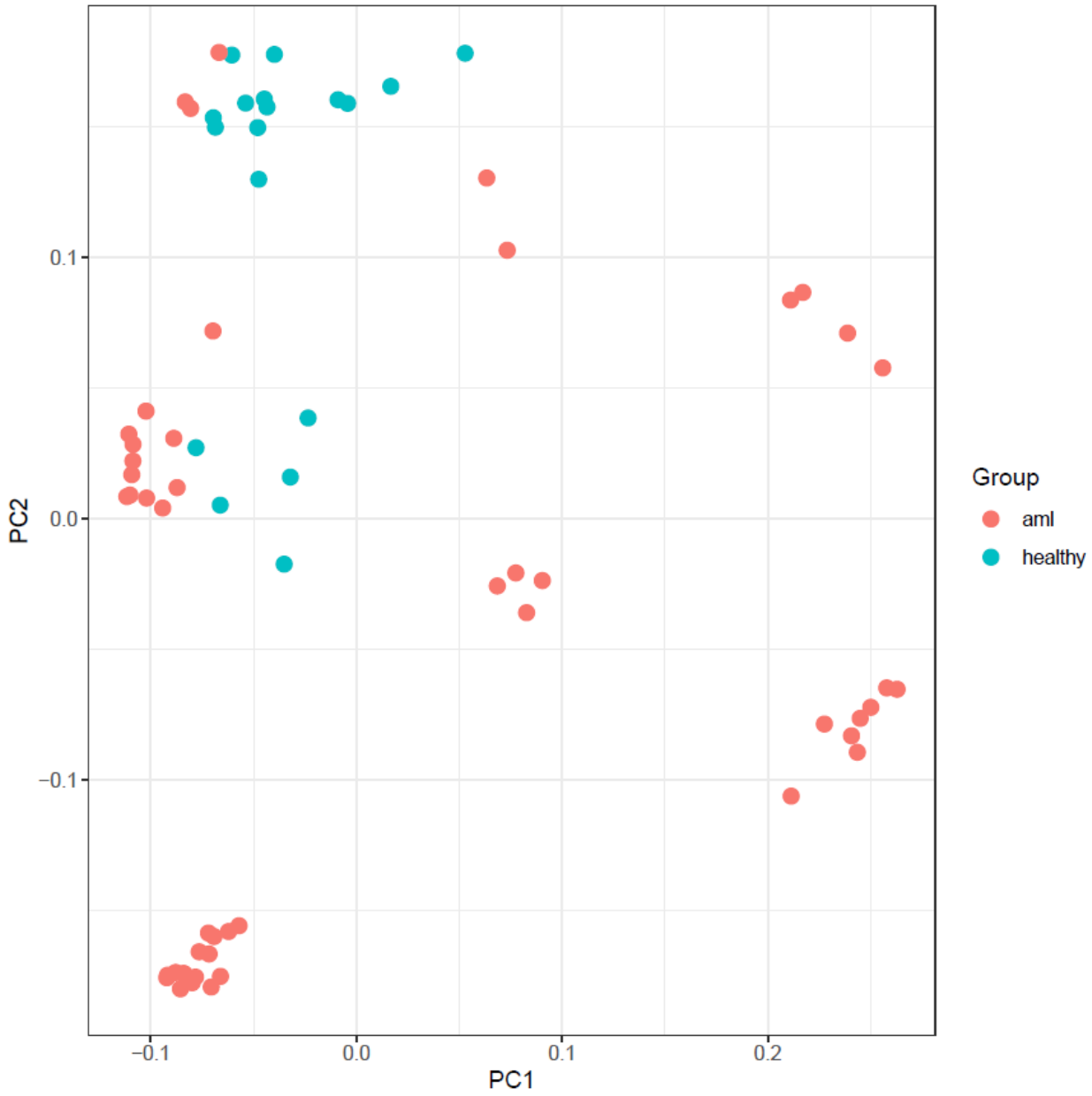


تصویر ۲

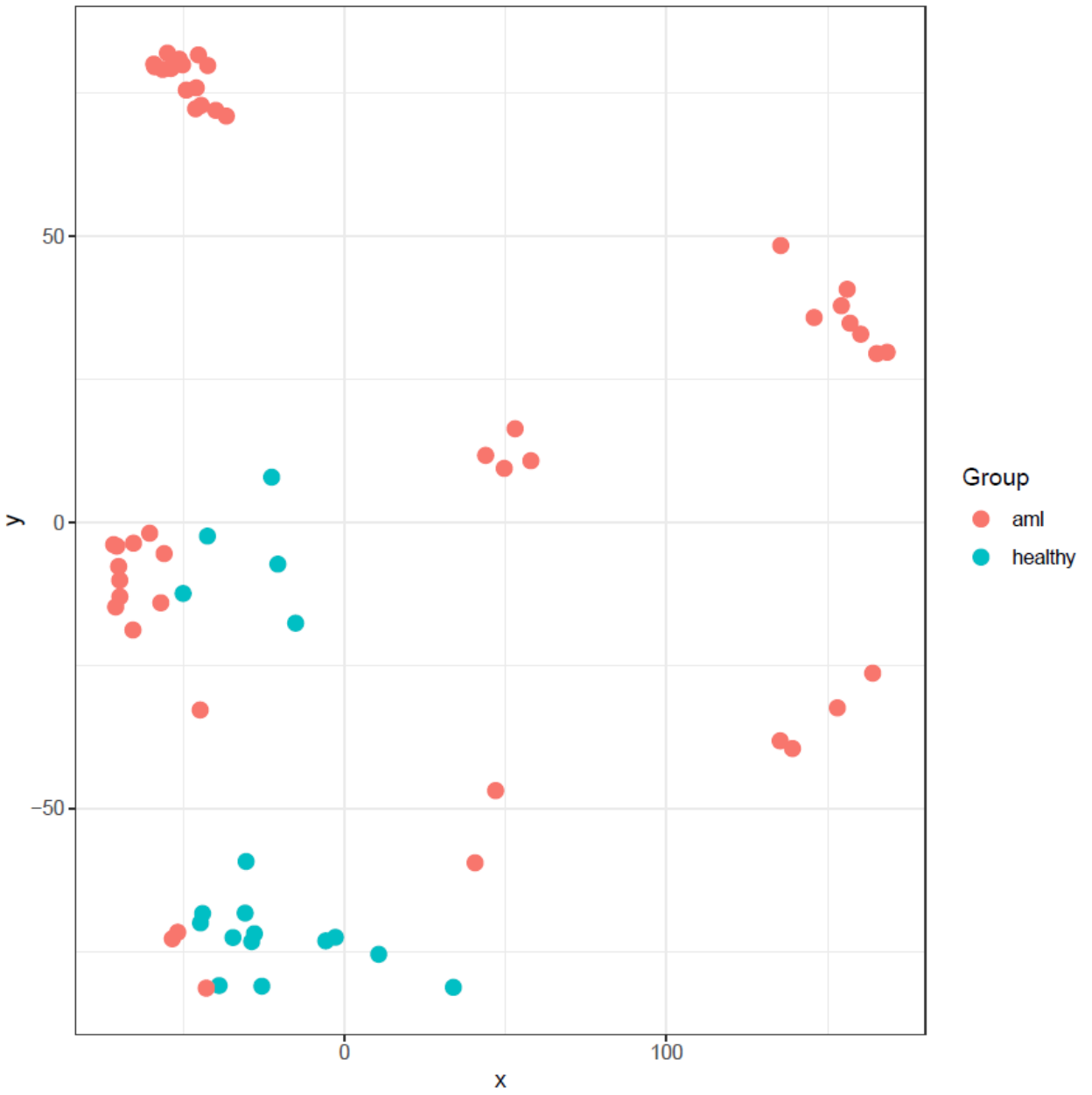
۳. لزوم کاهش ابعاد آنجا مشخص می‌شود که تعداد زیادی داده وجود داشته باشد و تحلیل این داده‌ها بر روی نمودار سخت شود. حال نیاز است از کاهش ابعاد استفاده کنیم تا با ترکیب فیچرها باعث تولید فیچرهای جدید بشود که بتوان در ادامه مراحل این تغییرات را مشاهده کرد و تحلیل ساده‌تر صورت گیرد. روش tSNE که در زیر می‌بینیم به خوبی توانسته است که گروه‌ها را از هم تشخیص دهد و گروه های سالم و بیمار را از یکدیگر جدا کند اما دو روش

دیگر نتوانسته اند که این مهم را انجام دهند و در بخش های زیادی دو گروه داده های بسیار نزدیک و غیر قابل جداسازی دارند.

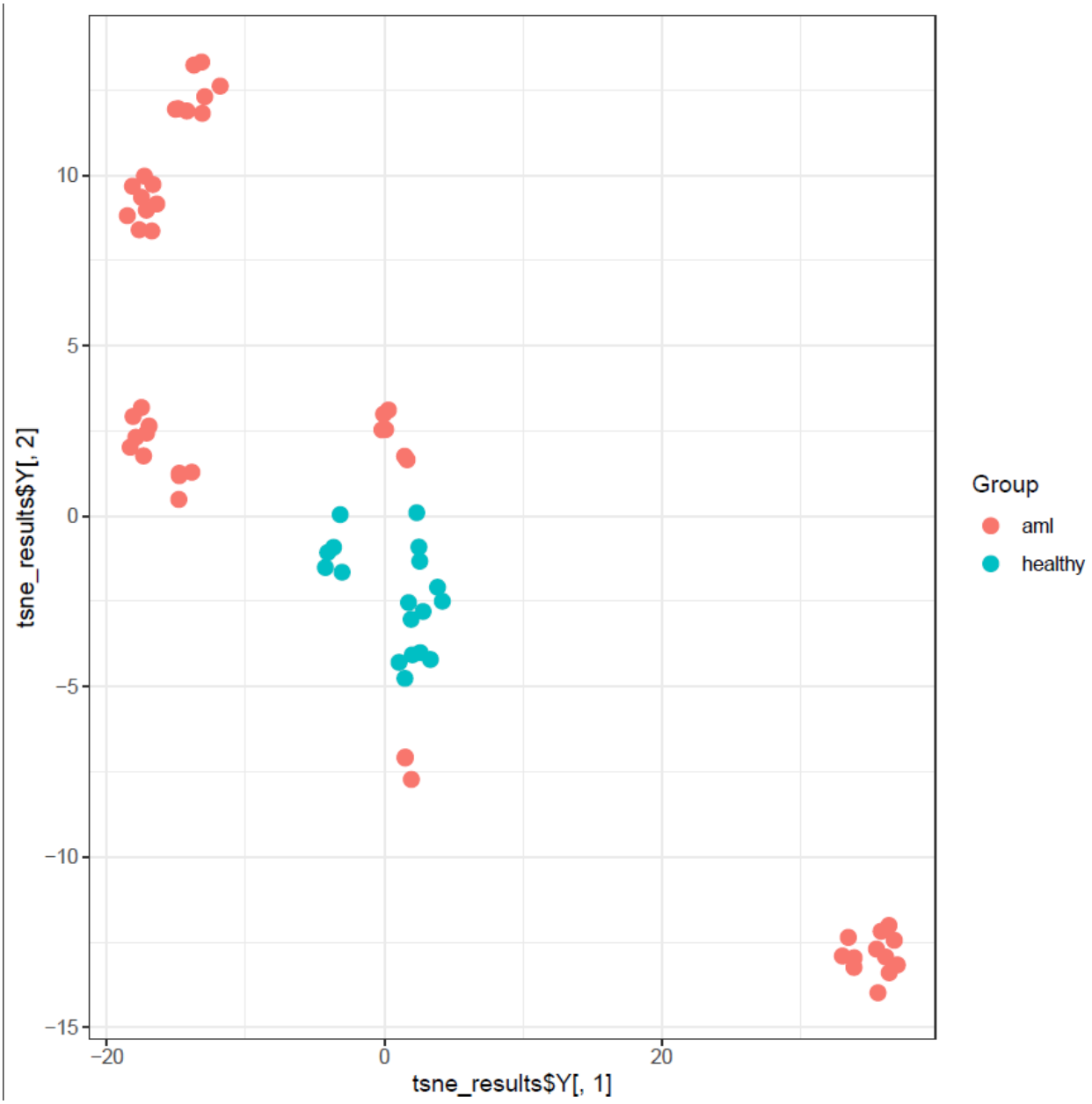
روش PCA:



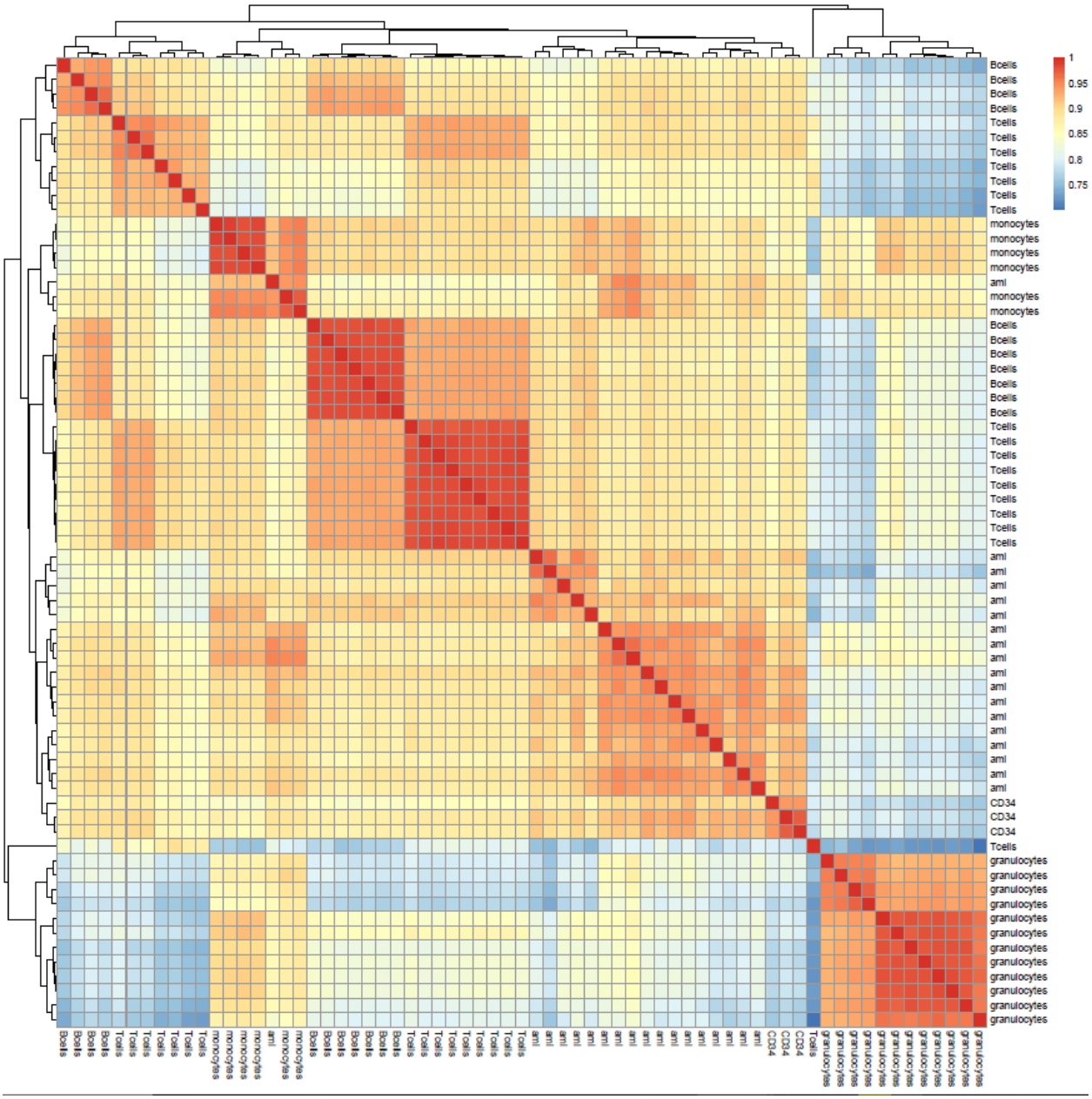
روش MDS:



روش tSNE:



۴. نشان می‌دهد که سلول از کجا نمونه برداری شده است و به کمک نمودار زیر می‌توان تشخیص داد که کدام سلول برای تشخیص سرطان مناسب هستند. با بررسی می‌توان به نتیجه ای که در مقدمه بیان شد برسیم.


1
0.95
0.9
0.85
0.8
0.75
Bcells
Tcells
monocytes
aml
CD34
granulocytes