بخش ۱.

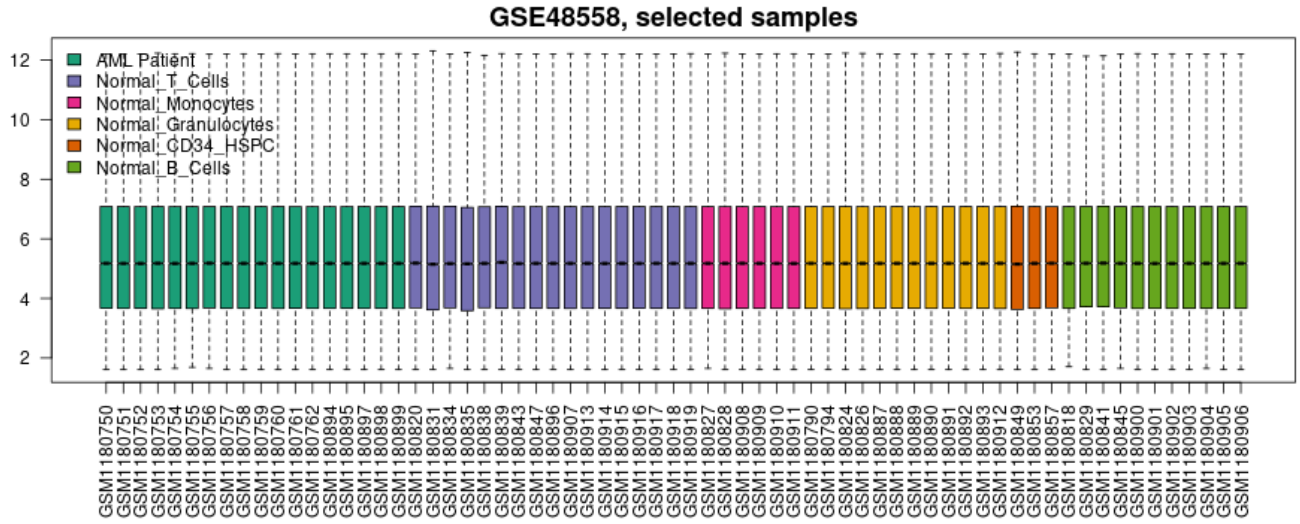
کنترل کیفیت داده

گام ۱ – Importing Libraries

ابتدا کتاب‌خانه‌های لازم را import می‌کنیم. این قسمت کپی شده‌ی خود سایت ncbi است که یک کد R به صورت ابتدایی قرار داده است.

گام ۲ – Analysis

همانطور که در صورت پروژه گفته شده می‌بایست داده‌های Normal را به عنوان پایه و AML Patient را تست در نظر بگیریم. در این بخش، با کمک برچسب‌گذاری انجام شده در دیتاست، این دو گروه را جدا می‌کنیم. به کمک خود سایت از قسمت داده‌های خود سایت به خوبی نرمالایز شده‌اند. با توجه به نمودار زیر می‌توان دید که میانگین‌ها به اندازه‌ی کافی نزدیک به هم هستند. البته برای آنکه نرمال‌های متفاوت را داشته باشیم، بر اساس جفت Normal و Source name آن‌ها را جدا می‌کنیم.

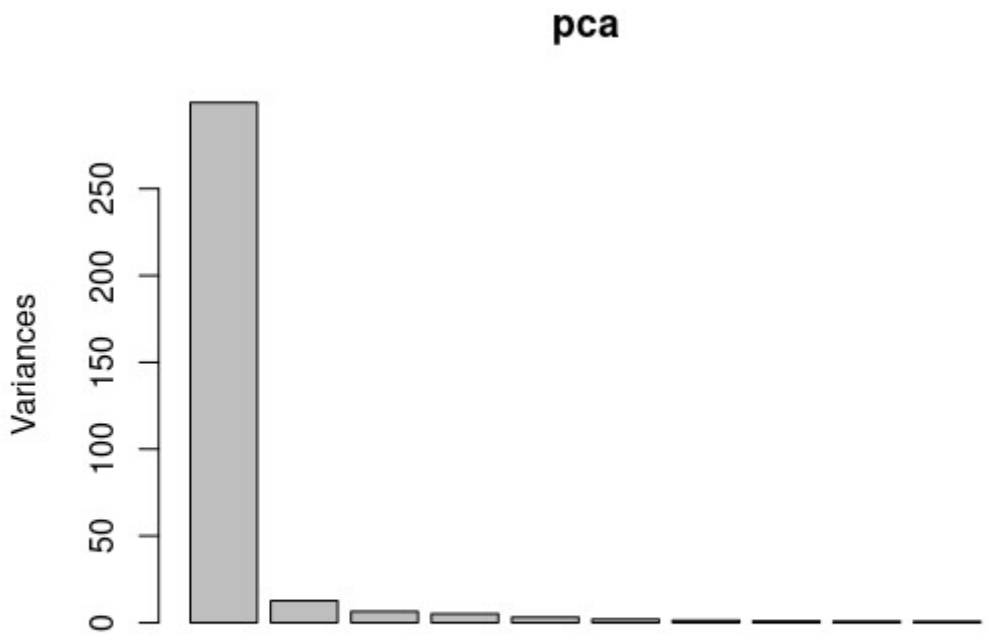


بخش ۲.

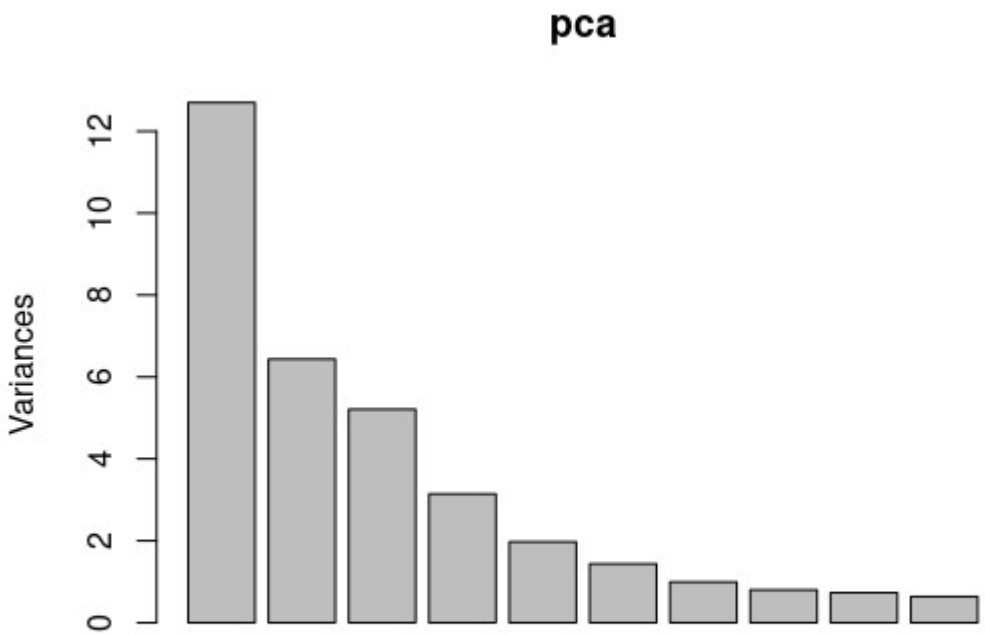
کاهش ابعاد داده

گام ۳ – Dimenstion reduction

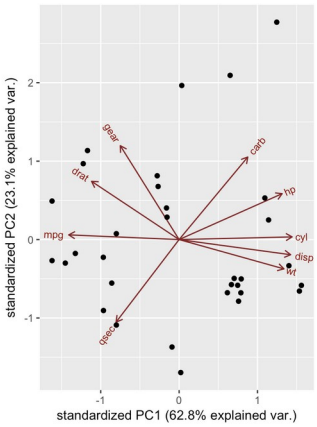
برای کاهش بعد دو راه PCA و t-SNE وجود دارد. در اینجا من PCA را استفاده کرده‌ام.


pca
Variances
0
50
100
150
200
250

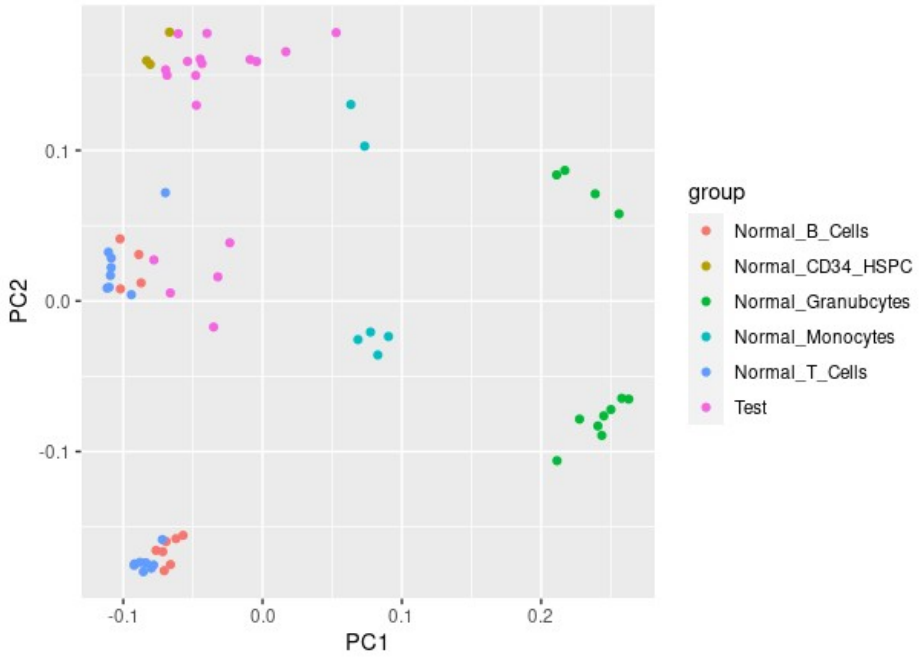
حالتی که PCA را بررسی می‌کنیم می‌بینیم که یک سری از ژن‌ها بیان بالایی دارند و باقی خیر. به صورت کلی، ژن‌هایی که مسئول بقای سلول هستند به صورت کلی بیانشان بالاست. یعنی به هر حال این ژن‌ها متمایزکننده‌ی خوبی نیستند. یک راه این است که داده‌ها را مقیاس کنیم، سپس میانگین را صفر کنیم.



از طرفی ترسیم خروجی PCA اساس چیزی شبیه به این است:



در خود سایت از umap استفاده شده ولی ما اینجا با کمک زاویه و طول بردار می‌توان PCA را نمایان کرده‌ایم..

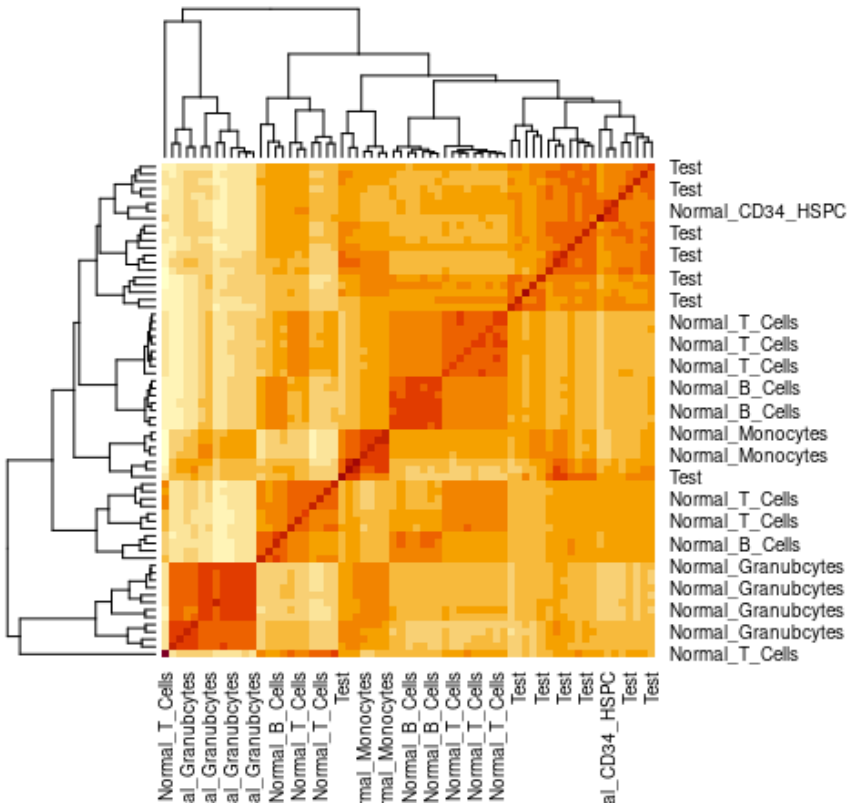


همانگونه که می‌بینید، داده‌های تست به Normal_CD34_HSPC بسیار نزدیک هستند (گوشه بالا سمت چپ). این یعنی اینکه در بخش چهارم، این گروه از داده‌های تست را جدا می کنیم.

بخش ۳.

خروجی Heatmap

گام ۴ – Correlation



بخش ۴.

بررسی تمایز ژن‌ها

گام ۵ – جداسازی.

همانطور که در انتهای بخش ۲ گفته شد، گروهی از داده‌های تست که PC2 بیشتر ۰.۱۳ و PC1 کمتر از ۰/۰۲۵- دارند نزدیک به CD34 هستند. همچنین در نمودار همبستگی نیز نواحی تست با این گروه همبستگی بالاتری دارند.

گام ۵ – مرتب‌سازی.

اکنون (مانند کلاس آقای سلیمی) این گروه Near را به گروه‌های gset اضافه می‌کنیم و نهایتاً آن را به صورت one-hot در می‌آوریم. این برای این است که از حالت categorical خارج شود.

گام ۶ – اعمال مدل‌ها.

در این قسمت با کمک کتاب‌خانه‌ی Limma از یک مدل خطی استفاده می‌کنیم. پس از fit کردن آن، می‌توان اختلاف بین بیانگرها را برای حالت Test و CD34_HSPC به دست آورد. سپس با فرض اینکه آستانه‌ی تفاوت ۰/۰۵ است، توزیع پیدا می‌شود. مشابه خود سایت، logFC و adj.P_value را جدولی نگه می‌داریم.

گام ۶ – نتایج.

با توجه به شرط کمتر ۰.۰۵ بودن adj.P_value و همچنین logFC (بسته به اینکه بالا را بخواهیم یا پایین را) ژن‌ها را جدا می‌کنی.