RayBiotech
Empowering your proteomics

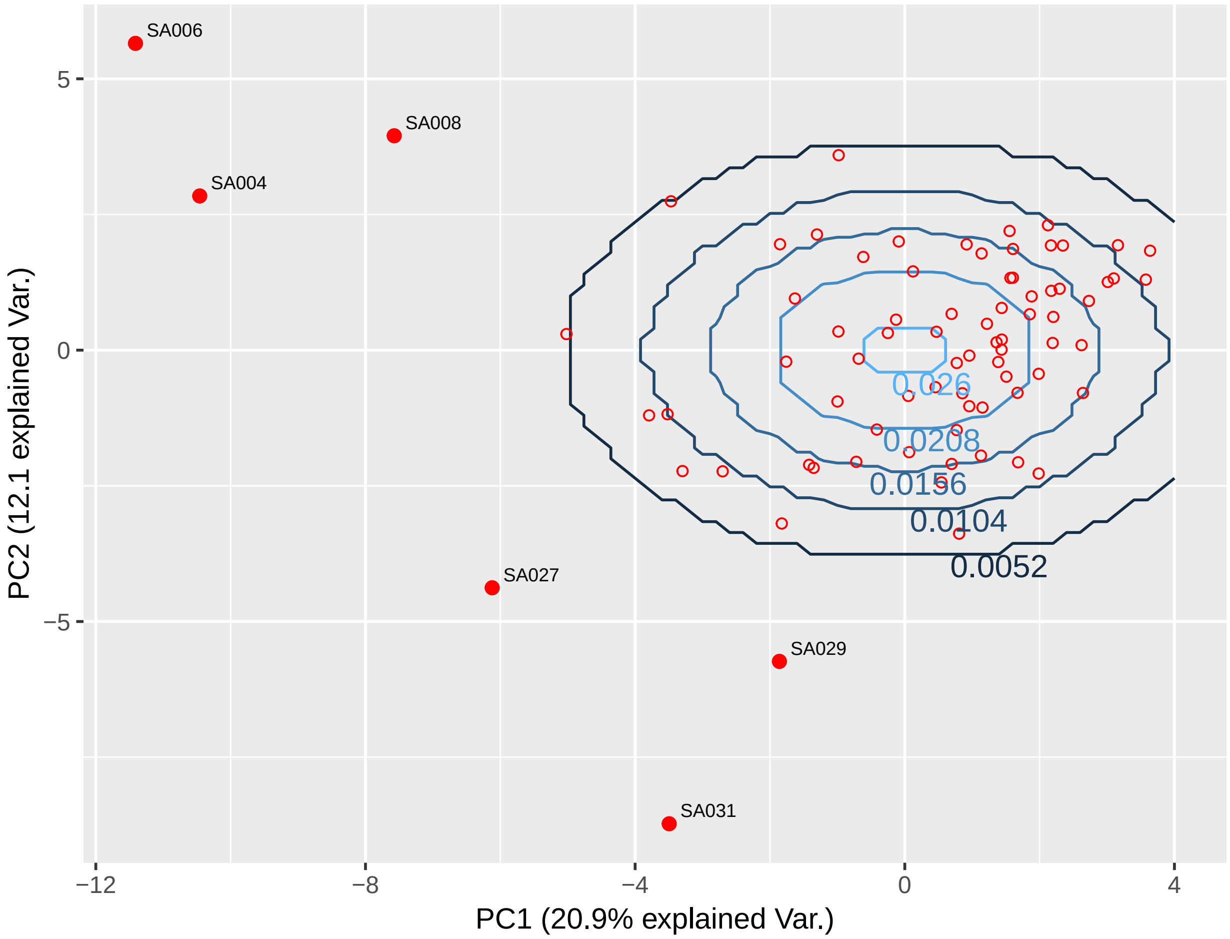
3607 Parkway Ln, Suite 200
Norcross GA 30092

1-888-494-8555
www.raybiotech.com

# EXAMPLE REPORT

## Biostatistics & Bioinfomatics Service

## "Data Clean-up" Service



Bioinformatics Team, RayBiotech
December 06, 2018

# Contents

# 1 Introduction

The “Data clean-up” service aims to filter data points with missing value, standardize data by scaling and centering, and identify outliers among samples.

> *Need help understanding how the statistical analyses were performed in layman’s terms? Please visit our* website.

# 2 Method

## 2.1 Data filtration

Samples with missing data will be identified, and excluded from the analysis. Biomarkers that show no variation across all the subjects (i.e., zero-variance) will be excluded from the analysis.

## 2.2 Data transformation

The original biomarker values will be scaled/centered to remove potential influences of differently-scaled biomarkers.

The scaled data will be transformed with Principal Component Analysis (PCA) to obtain principal components (PC) that are mutually orthogonal to each other. Each PC is a linear combination of products of original biomarker values and dedicated weights/coefficients. With the data set of $n$ subjects and $p$ biomarkers, for each $i^{th}$ of $n$ subjects with $p$ biomarkers $x_{ij}$, there is

$$\begin{aligned}
PC_1 &= w_{11} * x_{i1} + w_{12} * x_{i2} + ... + w_{1p} * x_{ip} \\
PC_2 &= w_{21} * x_{i1} + w_{22} * x_{i2} + ... + w_{2p} * x_{ip} \\
&... \\
PC_{min(n,p)} &= w_{min(n,p)1} * x_{i1} + w_{min(n,p)2} * x_{i2} + ... + w_{min(n,p)p} * x_{ip}
\end{aligned}$$

The transformed data, weights of all the biomarkers in different PCs, variations explained by each PC, and contributions of the biomarkers will be listed as separate files.

## 2.3 Outlier identification

All of the subjects will be plotted as points in a 2-dimensional figure by the first two PCs where the explained data variation are the highest compared to the other PCs. The distribution of the points will be observed, and those lying outside the swarm of most of the subjects will be considered as outliers.

### 2.4 Software

All the analysis will be conducted using R programming language V 3.5.1 (R Core Team 2017).

## 3 Results

### 3.1 Data filtration

#### 3.1.1 Subjects with missing values

Subject SA001 has missing values, thus was excluded from the analysis.

#### 3.1.2 Biomarkers with zero-variance

There is no biomarker with zero-variance across 79 subjects, thus no biomarker was excluded from the analysis based on zero-variance. Biomarkers missing values were not considered in this filtration step.

### 3.2 Data transformation

#### 3.2.1 Data scaling

The data scaling process will modify the scale of data, but will leave the shape of the distribution unchanged (Figure 1).

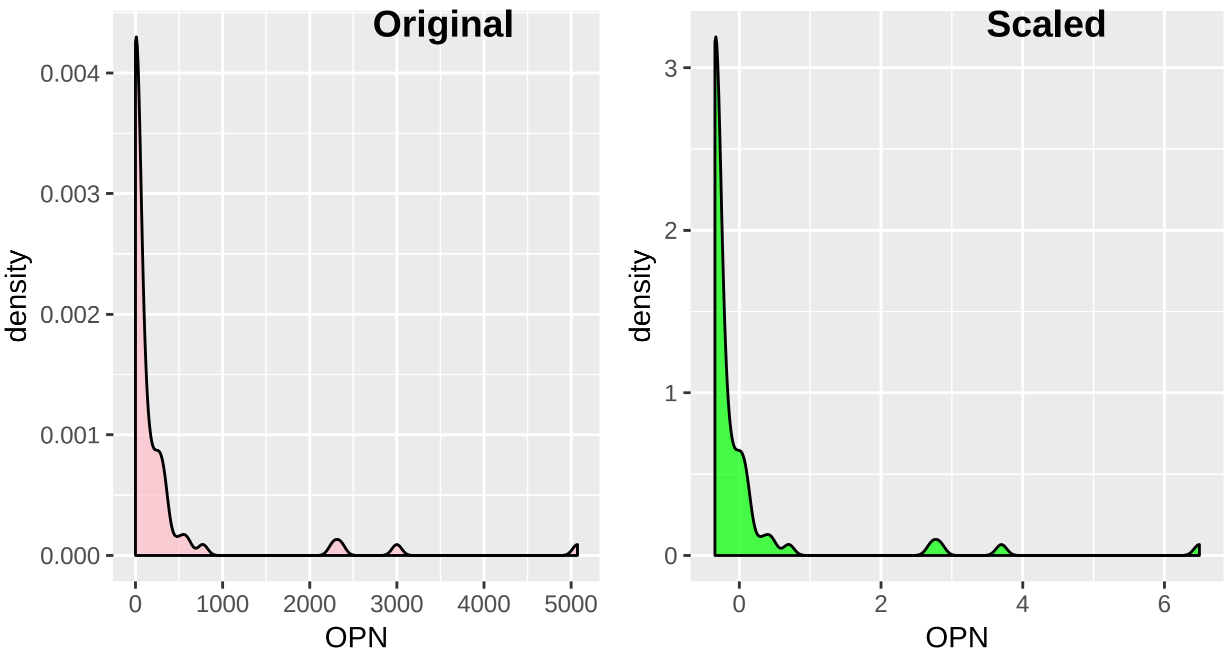


Figure 1: Density plot of biomarker OPN before and after scaling

#### 3.2.2 Principal component analysis

The PCA transforms the original dataset (79 X 38) into another data matrix of 79 samples and 38 PCs, considering the number of biomarkers is less than subjects.

Table 1 and Figure 2 demonstrate the percentage of variance explained by each PC.

Table 1: Variance explained by each Principal Component

| PC | Variance Explained % | PC | Variance Explained % | PC | Variance Explained % | PC | Variance Explained % |
|---|---:|---|---:|---|---:|---|---:|
| PC1 | 20.93 | PC11 | 2.99 | PC21 | 0.93 | PC31 | 0.27 |
| PC2 | 12.13 | PC12 | 2.63 | PC22 | 0.78 | PC32 | 0.20 |
| PC3 | 8.05 | PC13 | 2.22 | PC23 | 0.72 | PC33 | 0.11 |
| PC4 | 7.29 | PC14 | 2.13 | PC24 | 0.59 | PC34 | 0.09 |
| PC5 | 5.84 | PC15 | 1.81 | PC25 | 0.53 | PC35 | 0.05 |
| PC6 | 5.52 | PC16 | 1.49 | PC26 | 0.48 | PC36 | 0.05 |
| PC7 | 4.60 | PC17 | 1.36 | PC27 | 0.42 | PC37 | 0.03 |
| PC8 | 4.25 | PC18 | 1.21 | PC28 | 0.38 | PC38 | 0.01 |
| PC9 | 3.98 | PC19 | 1.14 | PC29 | 0.33 | | |
| PC10 | 3.15 | PC20 | 1.04 | PC30 | 0.28 | | |

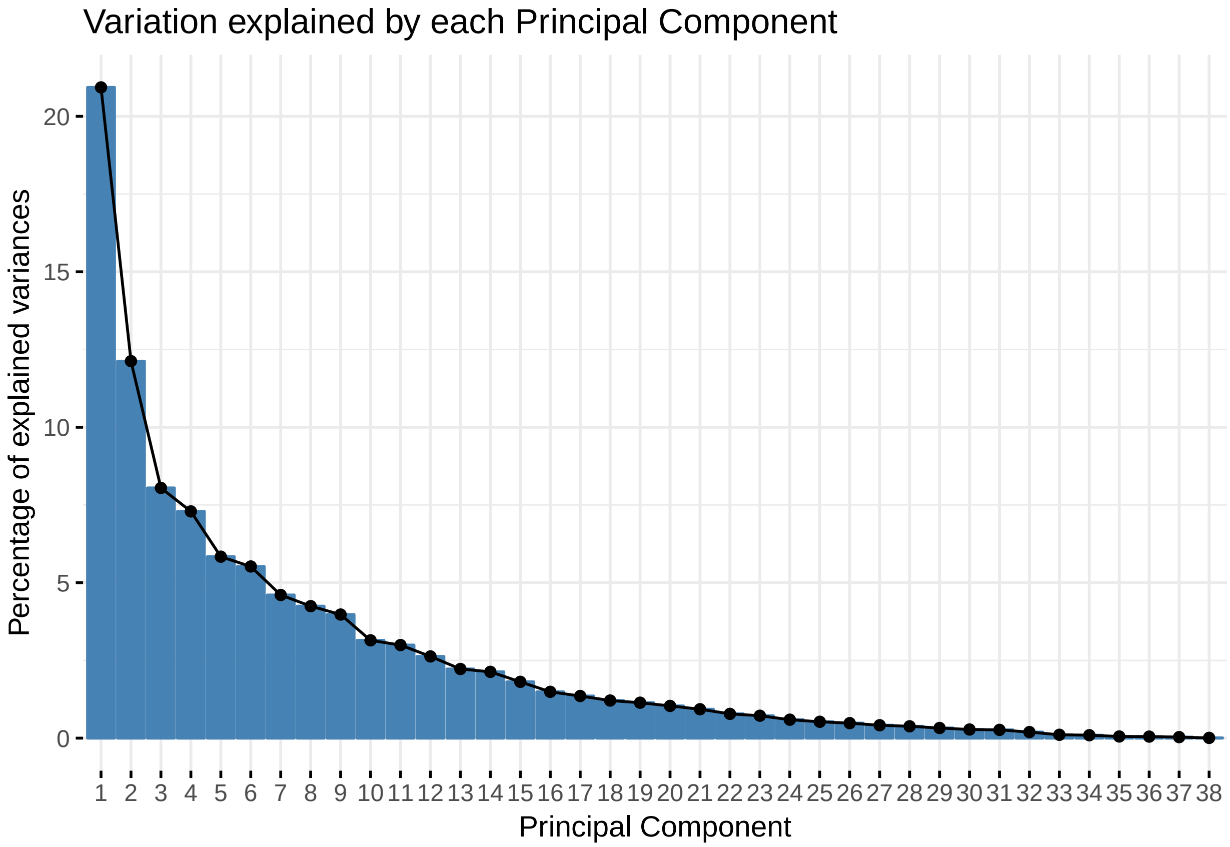


Figure 2: Variation explained by each principal component

Table 2 lists the weights of 38 biomarkers in the first 10 PCs.

Table 2: Weights of biomarkers in the first 10 Principal Components

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 |
|---|---|---|---|---|---|---|---|---|---|---|
| AFP | -0.0675 | -0.2515 | 0.0393 | -0.2425 | 0.0555 | -0.0510 | 0.3407 | -0.0356 | 0.1071 | -0.1005 |
| AgRP | -0.2567 | 0.0225 | 0.1875 | -0.0988 | 0.1575 | -0.0089 | 0.0614 | 0.0910 | -0.1272 | 0.1381 |
| BDNF | -0.1420 | -0.3081 | 0.0000 | -0.0047 | 0.0694 | -0.0372 | 0.0332 | 0.1812 | -0.1998 | -0.0558 |
| CA125 | -0.1463 | -0.2308 | 0.1049 | 0.0132 | 0.0234 | 0.0251 | -0.3393 | -0.1655 | -0.2274 | -0.1267 |
| CA15-3 | -0.2714 | 0.0019 | 0.0921 | -0.0422 | 0.1591 | 0.1303 | -0.0479 | 0.0574 | -0.0752 | 0.1381 |
| CEA | -0.1048 | -0.2739 | -0.0210 | 0.0375 | -0.1200 | 0.2718 | -0.1876 | -0.1303 | -0.0525 | 0.2355 |
| CXCL16 | -0.1357 | -0.2784 | 0.0347 | -0.0520 | 0.0137 | 0.2539 | 0.1023 | 0.2060 | 0.1369 | 0.0278 |
| EGF | -0.1972 | -0.2250 | -0.0919 | 0.1030 | -0.0045 | 0.0867 | 0.0010 | 0.0175 | -0.1263 | -0.1769 |
| EGF R | -0.2154 | -0.1693 | -0.1545 | -0.0254 | 0.0173 | 0.2141 | 0.0533 | 0.1401 | 0.2577 | 0.0504 |
| GROa | -0.1631 | 0.1173 | -0.0731 | -0.0518 | 0.0102 | -0.3457 | -0.3077 | 0.1253 | -0.0883 | 0.0064 |
| IFNa | -0.1622 | 0.0779 | -0.0621 | 0.0520 | 0.0339 | 0.2976 | -0.1108 | 0.0727 | 0.4166 | -0.2880 |
| IGFBP-4 | -0.2365 | 0.0224 | 0.2559 | -0.0878 | 0.1088 | -0.1551 | 0.1358 | -0.1607 | 0.0746 | 0.1042 |
| IL-1 R6 | -0.2743 | 0.1180 | 0.2291 | -0.0667 | 0.1129 | -0.0796 | -0.0196 | -0.1092 | 0.0479 | 0.0887 |
| IL-2 Ra | -0.2844 | 0.1327 | 0.2269 | -0.0401 | -0.0035 | -0.0382 | -0.0357 | -0.0106 | 0.0052 | 0.1346 |
| IL-6 | -0.2319 | 0.1564 | -0.0832 | 0.0368 | 0.0759 | -0.1536 | 0.0519 | -0.0358 | -0.0403 | -0.0647 |
| IL-6 sR | -0.1443 | -0.2326 | -0.1636 | -0.0442 | -0.0035 | 0.0535 | 0.1708 | 0.0686 | -0.1297 | 0.1446 |
| IL-8 | -0.2199 | 0.1212 | -0.2063 | 0.0150 | 0.0145 | -0.2069 | -0.1656 | 0.1206 | -0.1109 | -0.1171 |
| Leptin | -0.1024 | -0.2522 | 0.0773 | 0.0141 | -0.0123 | -0.0593 | -0.2240 | -0.2973 | -0.2257 | -0.1668 |
| MCSF | -0.1578 | 0.1389 | 0.1301 | 0.1200 | -0.4862 | 0.0400 | 0.0700 | 0.0841 | -0.0015 | -0.0437 |
| Mesothelin | -0.1263 | -0.0530 | -0.3198 | 0.1854 | -0.0837 | -0.0032 | 0.0221 | -0.1344 | 0.0740 | 0.3778 |
| MIF | -0.2556 | 0.1238 | 0.1847 | 0.0539 | -0.2755 | -0.0478 | 0.1048 | -0.0671 | -0.0107 | 0.0449 |
| MSPa | -0.0959 | 0.1099 | -0.0414 | 0.0167 | 0.1923 | -0.1591 | 0.1500 | -0.1041 | -0.0010 | 0.1716 |
| OPN | -0.2381 | 0.1651 | -0.0946 | 0.0026 | 0.0319 | 0.0400 | -0.2560 | 0.1529 | 0.2519 | -0.1912 |
| PDGF Ra | -0.2234 | -0.0159 | -0.0826 | 0.1547 | -0.2226 | 0.1234 | 0.1021 | -0.0369 | -0.1098 | -0.2309 |
| PDGF Rb | -0.0645 | 0.0599 | 0.1698 | 0.1060 | -0.5164 | -0.0119 | 0.1929 | 0.0941 | -0.0273 | 0.0243 |
| PDGF-AA | -0.1336 | 0.0926 | -0.3949 | 0.0834 | 0.1334 | 0.0188 | 0.1755 | 0.0019 | -0.0815 | 0.1398 |
| Prolactin | -0.0412 | 0.0431 | -0.1415 | -0.4974 | -0.1755 | 0.0671 | 0.0556 | -0.2303 | -0.0552 | -0.0211 |
| Prostasin | -0.0372 | 0.0500 | -0.2887 | -0.3669 | -0.1716 | -0.1126 | -0.1307 | 0.1455 | -0.0781 | -0.0592 |
| TIMP-4 | -0.1919 | 0.1996 | -0.2112 | 0.0439 | 0.1543 | 0.1005 | 0.0461 | -0.0744 | 0.1166 | -0.0174 |
| VEGF | -0.0310 | 0.0517 | -0.1747 | -0.4917 | -0.1927 | 0.0689 | -0.0183 | -0.1762 | -0.0721 | -0.0093 |
| B2M | -0.0125 | -0.2433 | -0.0136 | -0.1301 | -0.0956 | -0.2716 | -0.0958 | 0.3532 | 0.1331 | -0.0351 |
| HE4 | 0.0394 | -0.1853 | -0.0039 | -0.0205 | -0.1494 | -0.3791 | -0.0544 | 0.0915 | 0.3741 | 0.0703 |
| TIMP-2 | 0.0012 | -0.0786 | -0.2380 | 0.2119 | -0.1162 | -0.1469 | 0.1362 | -0.2862 | -0.0278 | -0.0629 |
| ICAM-1 | -0.0028 | -0.1680 | -0.0274 | 0.0486 | -0.1248 | -0.0893 | -0.2656 | 0.0115 | 0.1444 | 0.5191 |
| IGFBP-3 | -0.0904 | -0.1959 | -0.0954 | 0.2495 | 0.0421 | -0.3035 | 0.1549 | 0.0426 | -0.0919 | -0.2327 |
| ApoA1 | -0.0647 | -0.1967 | 0.1498 | -0.1809 | 0.1167 | -0.1642 | 0.1988 | -0.1815 | 0.2736 | -0.1448 |
| Adiponectin/ACRP30 | -0.0380 | -0.0713 | -0.0681 | 0.1090 | -0.0743 | -0.0883 | -0.1679 | -0.4944 | 0.3481 | -0.0909 |
| transferrin | 0.0305 | -0.0382 | 0.1455 | 0.0033 | 0.0240 | 0.1469 | -0.2799 | -0.0252 | 0.0011 | -0.0522 |

Figure 3 shows the weights of biomarkers in the first 2 PCs.

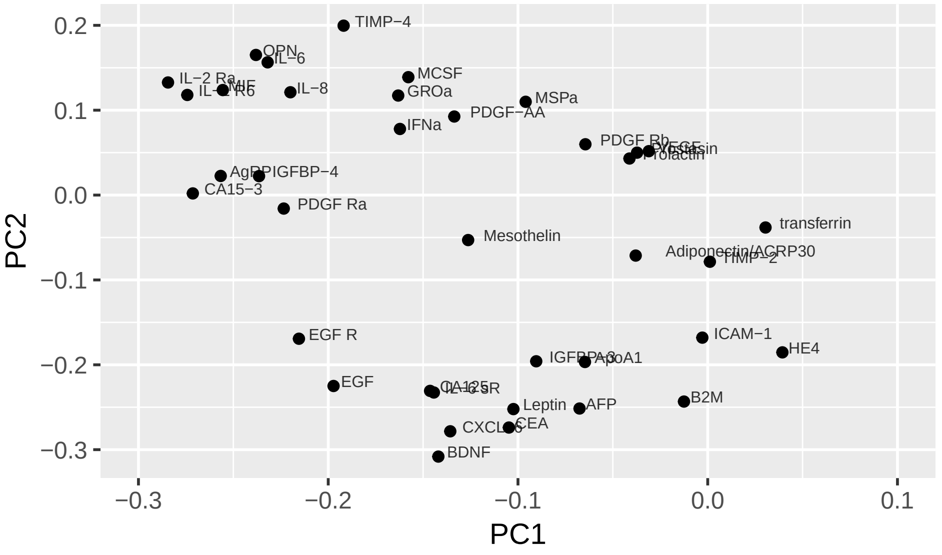


Figure 3: Weights of biomarker in the first 2 PCs

Table 3 lists 38 PCs for the first 10 samples.

Table 3: PCs of the first 10 samples

| | SA002 | SA003 | SA004 | SA005 | SA006 | SA007 | SA008 | SA009 | SA010 | SA011 |
|---|---|---|---|---|---|---|---|---|---|---|
| PC1 | -0.981 | 0.917 | -10.459 | -1.851 | -11.412 | -1.629 | -7.574 | 2.297 | 1.672 | 0.694 |
| PC2 | 3.592 | 1.949 | 2.841 | 1.952 | 5.653 | 0.952 | 3.950 | 1.132 | -0.785 | 0.669 |
| PC3 | -1.756 | -0.425 | -1.998 | -2.780 | 2.703 | -1.496 | -0.427 | -0.258 | -0.725 | -1.249 |
| PC4 | 0.202 | 0.317 | 1.325 | 1.780 | -1.349 | 0.967 | 0.913 | 0.729 | 0.722 | 1.335 |
| PC5 | 3.368 | 0.442 | 0.677 | 0.597 | 3.189 | 0.160 | -3.307 | 0.017 | -0.168 | 0.100 |
| PC6 | -1.283 | -0.294 | 5.623 | 0.987 | -2.047 | 2.329 | 0.343 | 0.444 | 0.832 | -1.122 |
| PC7 | 2.158 | 0.964 | -1.630 | 1.828 | -2.640 | 0.799 | -1.798 | 0.259 | -0.106 | -0.212 |
| PC8 | -0.431 | -0.268 | 1.054 | -0.976 | -0.640 | 0.232 | -0.568 | -1.448 | -0.444 | -1.890 |
| PC9 | -0.599 | -1.338 | 5.606 | -1.152 | -1.100 | -1.158 | -1.232 | 0.257 | 0.025 | -0.154 |
| PC10 | 1.250 | -0.762 | -3.062 | -0.259 | 1.489 | 0.768 | -0.966 | -0.225 | 0.216 | -0.982 |
| PC11 | 6.263 | -0.034 | 0.506 | -0.015 | -1.243 | 0.466 | -1.189 | -0.479 | -0.783 | -0.366 |
| PC12 | -0.402 | -0.381 | -2.572 | 1.045 | -0.160 | 0.735 | 2.378 | -0.421 | -0.554 | 0.285 |
| PC13 | -3.027 | -0.842 | -0.218 | -0.093 | 0.606 | -0.018 | -1.299 | 0.120 | -0.621 | 0.314 |
| PC14 | -1.063 | 1.463 | 0.212 | -1.384 | 0.234 | 0.364 | 0.651 | 0.304 | -0.276 | 0.589 |
| PC15 | -0.081 | -0.554 | -1.033 | -0.242 | 0.710 | 0.328 | 1.229 | 0.428 | -0.012 | 0.614 |
| PC16 | -0.268 | -0.915 | 0.060 | 0.574 | -1.795 | 0.984 | 0.906 | 0.658 | 0.100 | -2.281 |
| PC17 | -1.951 | 0.407 | 0.213 | 1.383 | -0.549 | 1.386 | -0.263 | 0.071 | -0.753 | 0.947 |
| PC18 | 0.122 | 0.009 | -0.612 | -0.694 | 1.371 | 0.085 | 0.897 | -0.024 | 0.151 | 0.973 |
| PC19 | -0.499 | 0.319 | -0.365 | 0.547 | -0.256 | 0.208 | 1.199 | -0.628 | -0.046 | -0.039 |
| PC20 | -1.140 | -0.386 | -0.268 | 0.966 | 0.974 | 0.224 | -1.877 | 0.085 | 0.135 | 0.170 |
| PC21 | 0.373 | -0.526 | -0.047 | -1.532 | -1.341 | -0.332 | 1.707 | -0.149 | -0.017 | 0.856 |
| PC22 | -0.023 | 0.501 | -0.588 | -0.345 | 0.388 | 0.402 | 0.343 | 0.290 | -0.254 | -0.840 |
| PC23 | -0.072 | 0.167 | 0.026 | 0.120 | 0.683 | 0.237 | -0.445 | 0.289 | 0.280 | 0.012 |
| PC24 | -0.211 | 0.231 | 0.102 | 0.519 | -0.229 | -0.427 | 0.444 | -0.032 | -0.773 | -1.486 |
| PC25 | 0.119 | -0.377 | -0.437 | 0.645 | 0.067 | 0.508 | 0.583 | 0.172 | -0.231 | -0.128 |
| PC26 | -0.165 | 0.634 | 0.292 | -0.527 | -0.424 | 0.439 | -0.208 | -0.131 | -0.977 | 0.853 |
| PC27 | -0.115 | -0.083 | 0.192 | -0.436 | -0.428 | -0.947 | -0.226 | -0.030 | -0.262 | -0.309 |
| PC28 | 0.194 | 0.461 | 0.017 | 0.636 | -0.167 | -0.432 | -0.196 | 0.023 | 0.060 | 0.156 |
| PC29 | 0.317 | -0.228 | 0.162 | -0.382 | -0.110 | 0.164 | -0.604 | 0.375 | -0.109 | 0.643 |
| PC30 | -0.208 | 0.074 | 0.039 | 0.242 | -0.148 | 0.029 | -0.633 | -0.250 | 0.356 | 0.357 |
| PC31 | 0.024 | -0.084 | 0.043 | 0.424 | -0.268 | -0.097 | 0.257 | -0.336 | -0.584 | -0.181 |
| PC32 | -0.017 | 0.062 | 0.027 | -0.008 | -0.121 | -0.174 | -0.242 | -0.096 | -0.643 | 0.296 |
| PC33 | 0.013 | -0.074 | -0.007 | 0.224 | 0.174 | 0.117 | -0.279 | 0.035 | -0.169 | -0.272 |
| PC34 | 0.050 | 0.155 | 0.040 | 0.106 | -0.070 | -0.058 | 0.017 | 0.108 | -0.345 | 0.205 |
| PC35 | -0.014 | -0.030 | -0.024 | -0.047 | 0.278 | -0.143 | -0.104 | 0.067 | -0.169 | -0.175 |
| PC36 | 0.025 | -0.046 | -0.033 | -0.046 | -0.095 | 0.276 | 0.110 | 0.002 | 0.129 | -0.222 |
| PC37 | -0.011 | 0.131 | -0.050 | 0.030 | 0.013 | 0.064 | 0.114 | 0.163 | -0.118 | 0.136 |
| PC38 | -0.022 | -0.066 | -0.004 | -0.063 | 0.025 | 0.097 | -0.011 | -0.007 | -0.005 | -0.033 |

## 3.3 Outlier identification based on the first 2 PCs

The first 2 PCs calculated from the data set account for 33.06% of total variance. Figure 4 plots the first 2 PCs of 79 subjects, which demonstrates the cluster of healthy subjects, and provides a way to identify those outside of the majority (i.e., the outliers).

We calculated the joint probability density of PC1 and PC2, assuming both PCs follow a normal distribution. The joint probability density, plotted as contour lines, indicates the boundary of the majority of the subjects. From the plot, we can tell that the samples SA004, SA006, SA008, SA027, SA029 and SA031 are probably outliers (Figure 4), and should be excluded from normal range estimation.

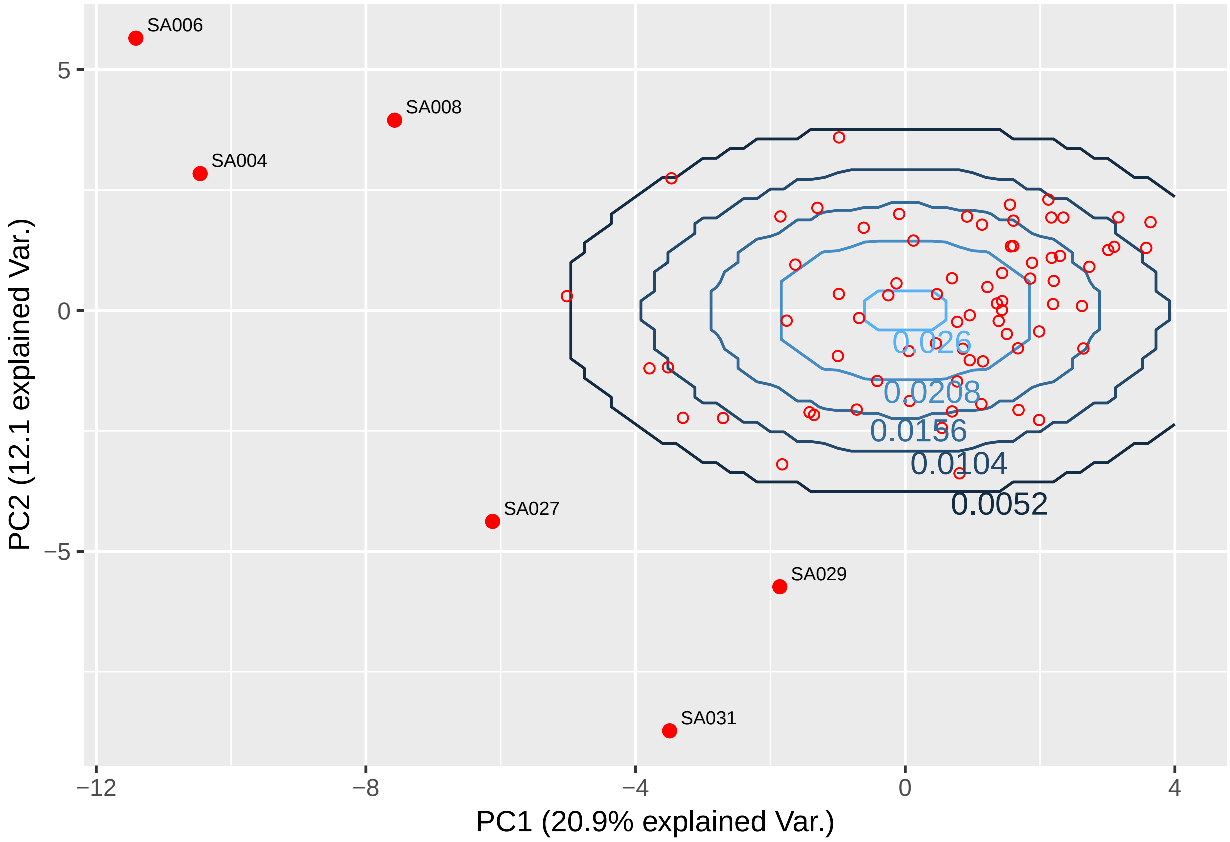


Figure 4: Plot of PC1 and PC2 values of 79 subjects

## 4 Summary

In this analysis we conducted data filtration, scaling and transformation, and the outlier identification based on PCA to prepare appropriate data set for normal range estimation of 38 biomarkers in 80 healthy subjects. One subject with missing data, and 6 probable outliers were identified. Please note that our strategy was quite conservative because this project aimed to provide a robust estimation of the normal range of expressed protein. The criteria could be less stringent for projects with different objectives.

## Reference

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.