Dimension reduction for genomics

Jeff Leek

Dependencies

This document depends on the following packages:

```
  library(devtools)
  library(Biobase)
```

To install these packages you can use the code (or if you are compiling the document, remove the `eval=FALSE` from the chunk.)

```
install.packages(c("devtools"))
source("http://www.bioconductor.org/biocLite.R")
biocLite(c("Biobase"))
```

General principles

- Can we find patterns in matrices of data?

Load some data

We will use this expression set that combines two studies Transcriptome genetics using second generation sequencing in a Caucasian population. (http://www.ncbi.nlm.nih.gov/pubmed?term=20220756%5Buid%5D) and Understanding mechanisms underlying human gene expression variation with RNA sequencing. (http://www.ncbi.nlm.nih.gov/pubmed?term=20220758). These studies are different populations but we counted the same genes for both. Then we'll explore the differences.

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
ls()
```

```
## [1] "con"           "edata"         "fdata"         "montpick.eset"
## [5] "mp"            "pdata"         "tropical"
```

Calculate the singular vectors

Here we calculate the singular vectors:

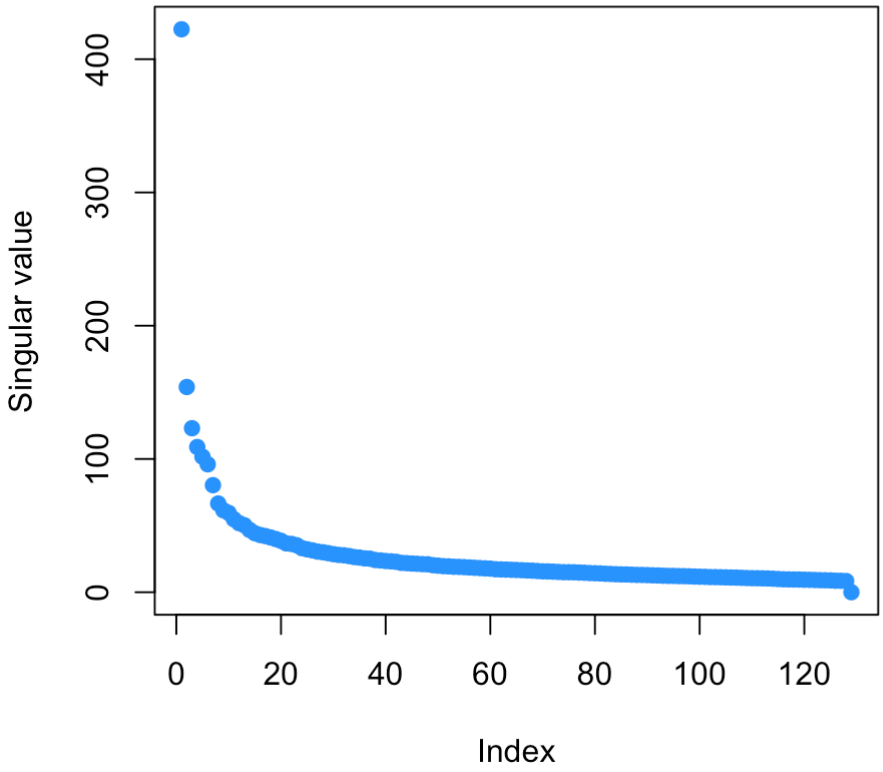
```
edata = edata[rowMeans(edata) > 100, ]
edata = log2(edata + 1)
edata_centered = edata - rowMeans(edata)
svd1 = svd(edata_centered)
names(svd1)
```

```
## [1] "d" "u" "v"
```

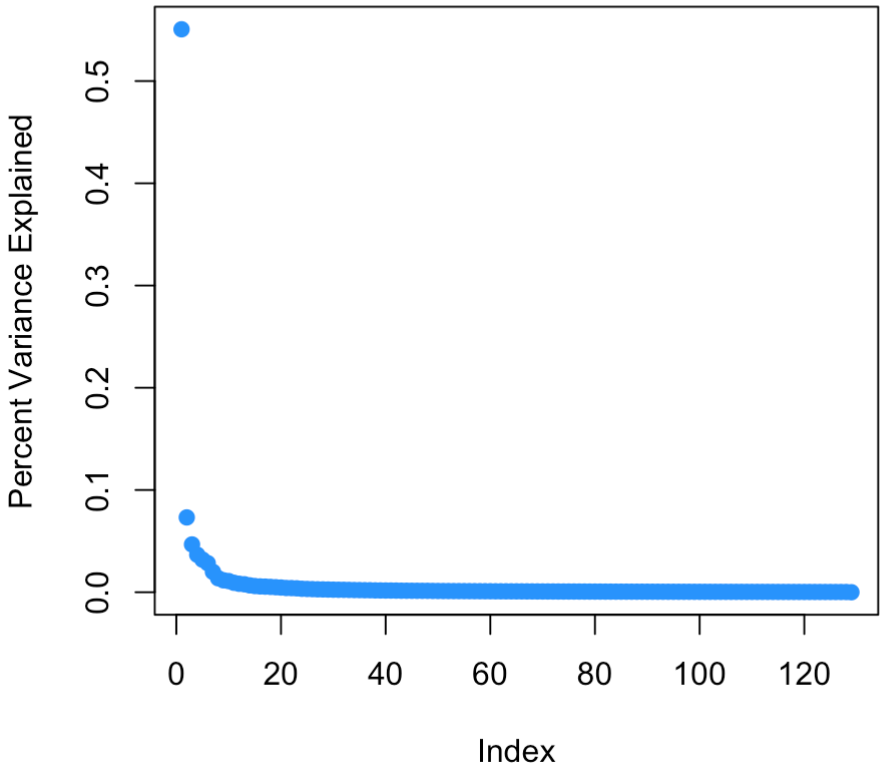
Look at the percent variance explained

The percent of variance explained is given by $\frac{d_{ii}}{\sum_j d_{jj}^2}$

```
plot(svd1$d,ylab="Singular value",col=2)
```

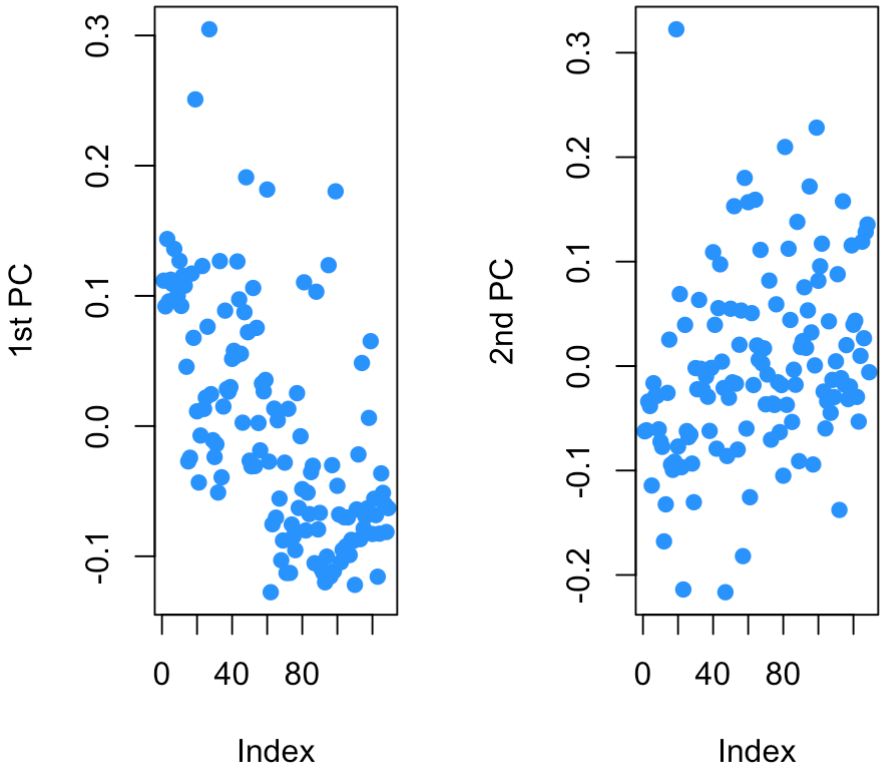



```
plot(svd1$d^2/sum(svd1$d^2),ylab="Percent Variance Explained",col=2)
```



Plot top two principal components

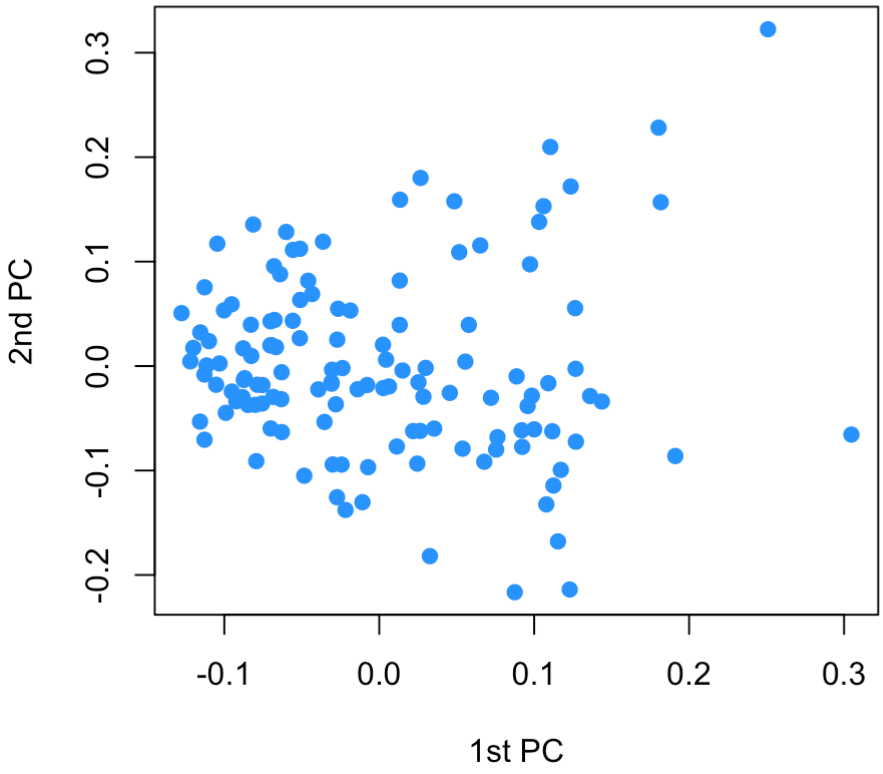
```
par(mfrow=c(1,2))
plot(svd1$v[,1],col=2,ylab="1st PC")
plot(svd1$v[,2],col=2,ylab="2nd PC")
```



Plot PC1 vs. PC2

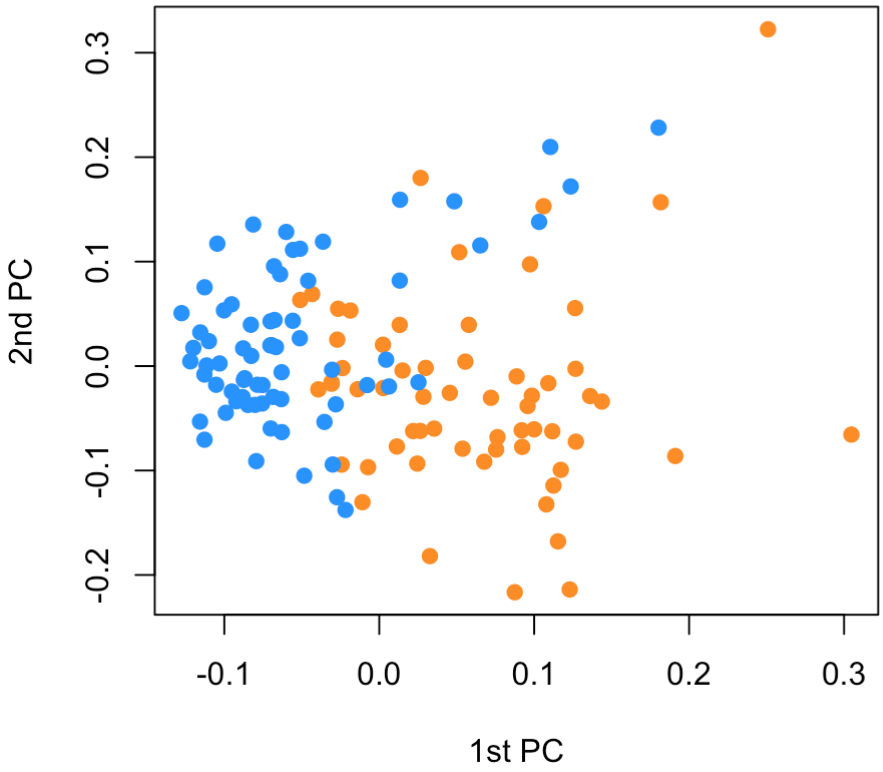
A very common plot is to plot PC1 versus PC2 to see if you can see any “clusters” or “groups”.

```
plot(svd1$v[,1],svd1$v[,2],col=2,ylab="2nd PC",xlab="1st PC")
```



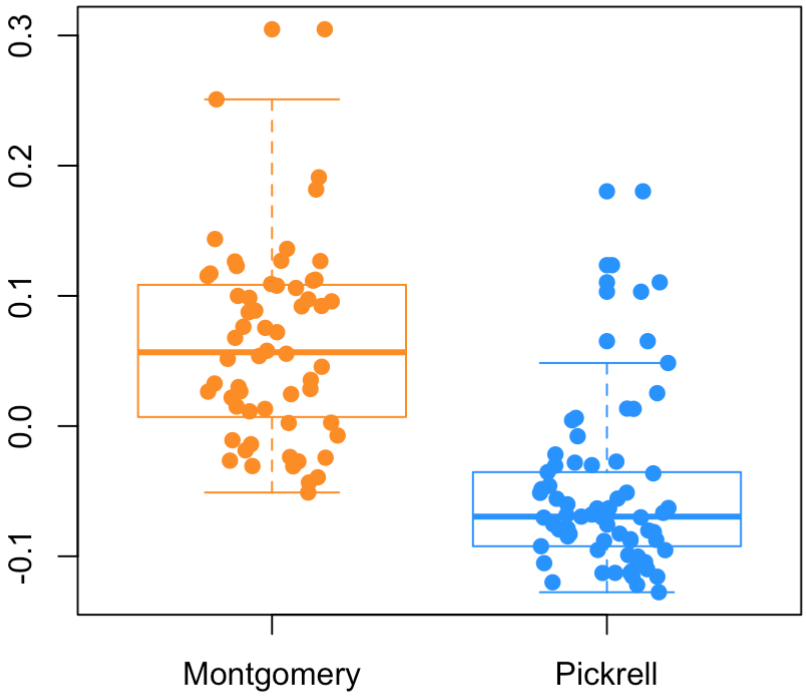
One thing you can do is color them by different variables to see if clusters stand out.

```
plot(svd1$v[,1],svd1$v[,2],ylab="2nd PC",
     xlab="1st PC",col=as.numeric(pdata$study))
```



Another common plot is to make boxplots comparing the PC for different levels of known covariates (don't forget to show the actual data!).

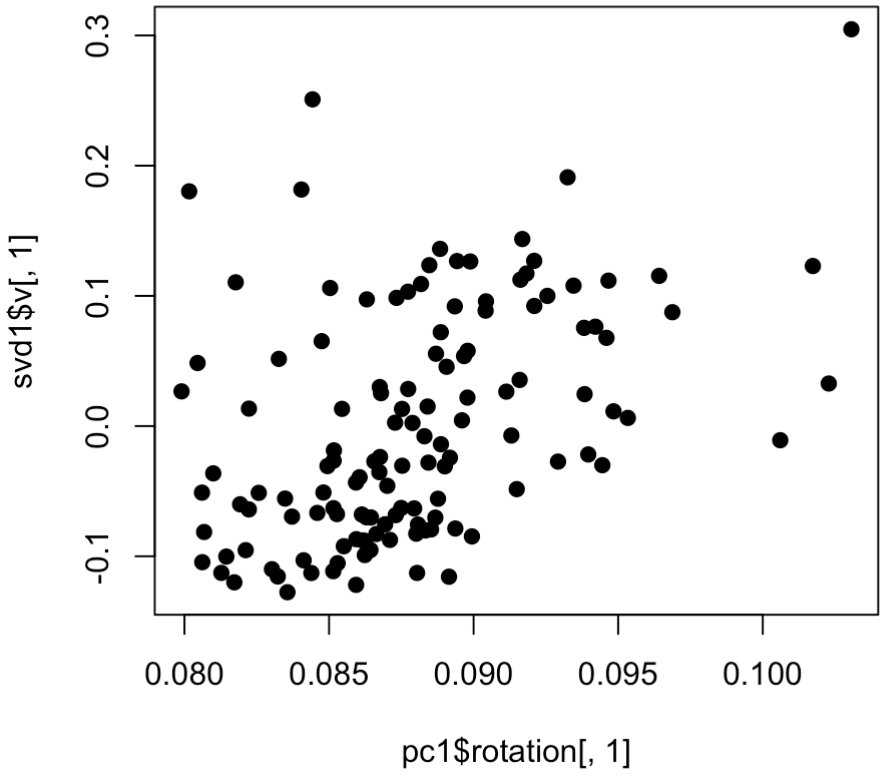
```
boxplot(svd1$v[,1] ~ pdata$study,border=c(1,2))
points(svd1$v[,1] ~ jitter(as.numeric(pdata$study)),col=as.numeric(pdata$study))
```



PCs versus SVs

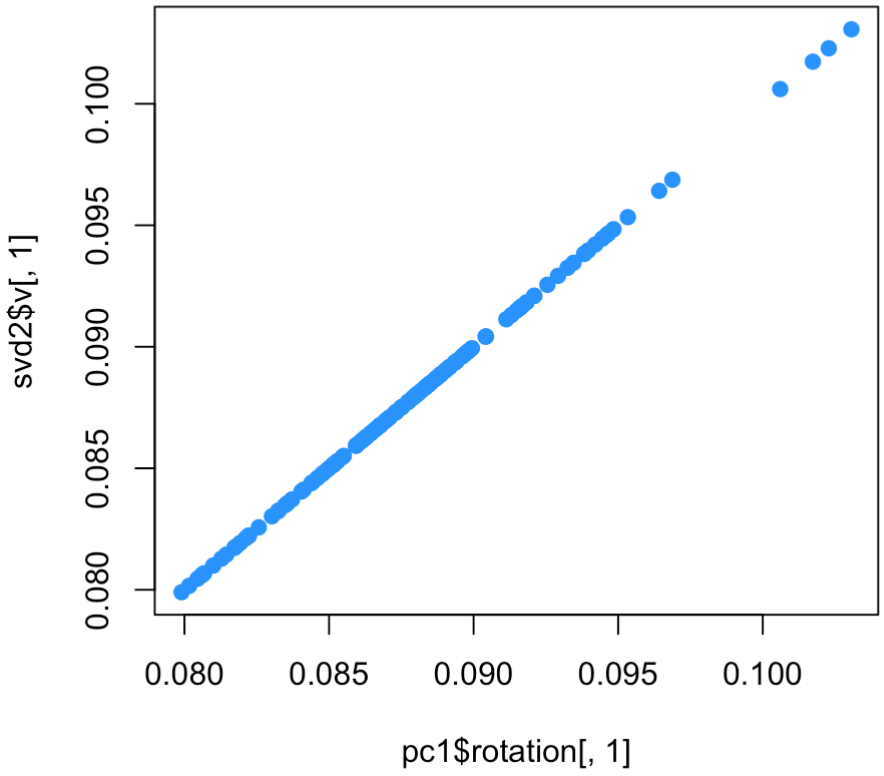
What we have been plotting is not exactly the principal components.

```
pc1 = prcomp(edata)
plot(pc1$rotation[,1],svd1$v[,1])
```



To get the actual PCs you have to subtract the column means rather than the row means when normalizing.

```
edata_centered2 = t(t(edata) - colMeans(edata))
svd2 = svd(edata_centered2)
plot(pc1$rotation[,1],svd2$v[,1],col=2)
```

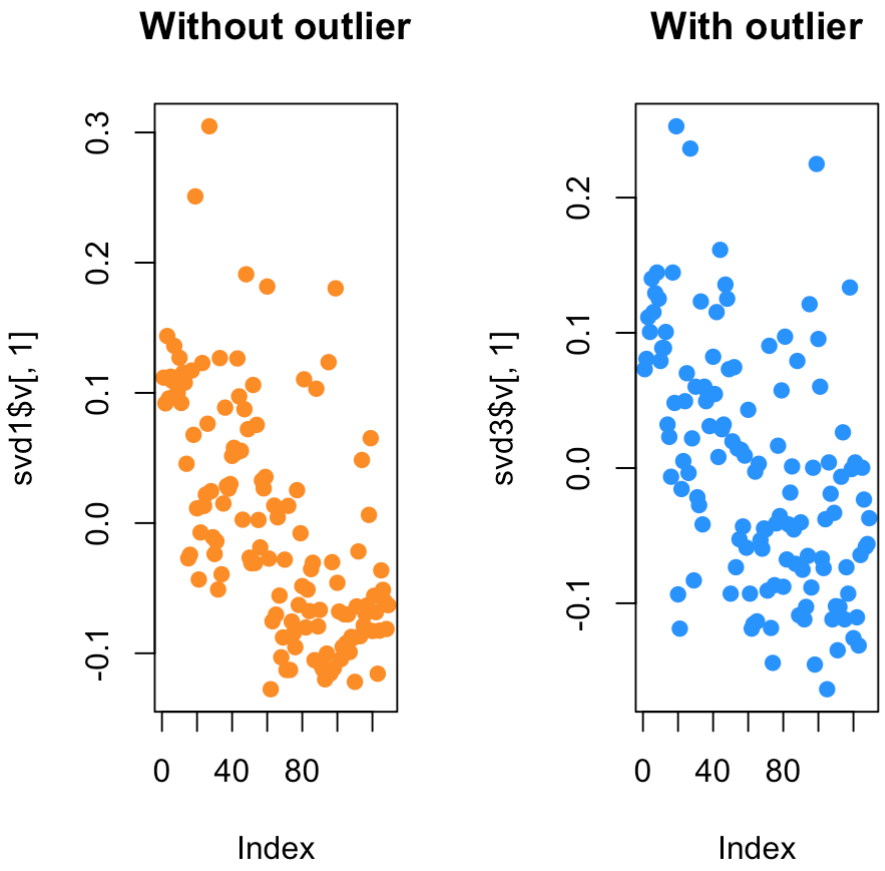


Despite this, it is most common for people to perform row-centering and then plot the singular vectors (sometimes labeling them PCs like I have done in this document)

Outliers

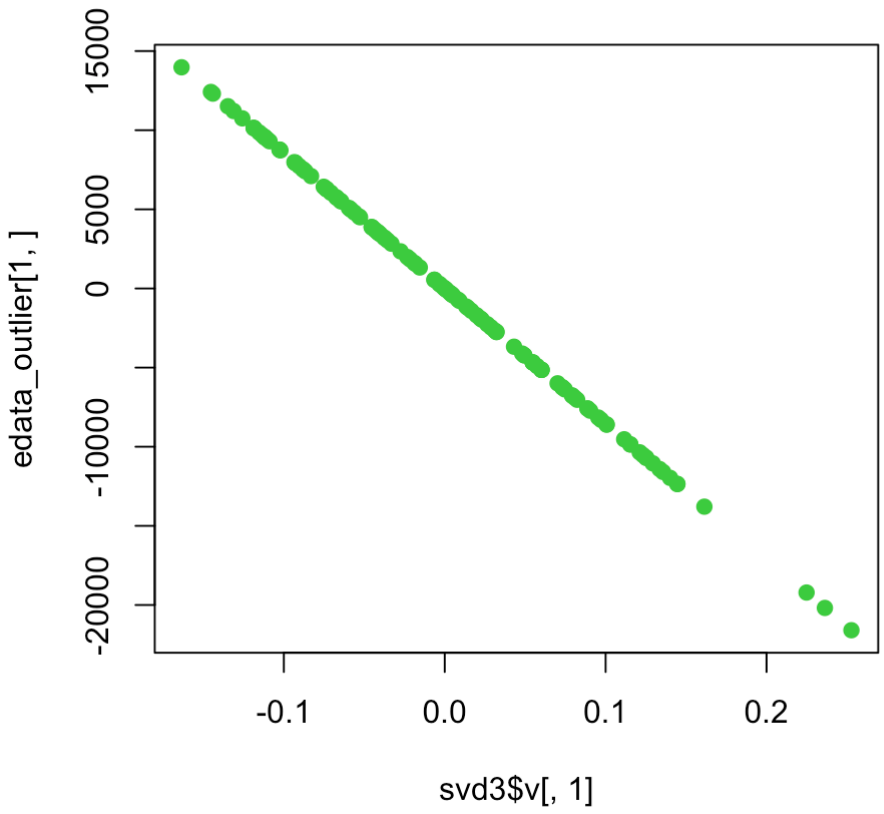
What happens if we introduce a single outlying gene

```
edata_outlier = edata_centered
edata_outlier[1,] = edata_centered[1,] * 10000
svd3 = svd(edata_outlier)
par(mfrow=c(1,2))
plot(svd1$v[,1],col=1,main="Without outlier")
plot(svd3$v[,1],col=2,main="With outlier")
```



It turns out the new top singular vector is perfectly correlated with the outlying gene

```
plot(svd3$v[,1],edata_outlier[1,],col=4)
```



Further resources

There are a large number of resources available about PCA and SVD but the lecture notes from Advanced Statistics for the Life Sciences (http://genomicsclass.github.io/book/) are the best set of lecture notes focused on genomics currently available.

Session information

Here is the session information

```
devtools::session_info()
```

```
##  setting  value
##  version  R version 3.2.1 (2015-06-18)
##  system   x86_64, darwin10.8.0
##  ui       RStudio (0.99.447)
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##
##  package           * version     date
##  acepack             1.3-3.3     2014-11-24
##  annotate            1.46.1      2015-07-11
##  AnnotationDbi     * 1.30.1      2015-04-26
##  assertthat          0.1         2013-12-06
##  BiasedUrn         * 1.06.1      2013-12-29
##  Biobase           * 2.28.0      2015-04-17
##  BiocGenerics      * 0.14.0      2015-04-17
##  BiocInstaller     * 1.18.4      2015-07-22
##  BiocParallel        1.2.20      2015-08-07
##  biomaRt             2.24.0      2015-04-17
##  Biostrings          2.36.3      2015-08-12
##  bitops              1.0-6       2013-08-17
##  bladderbatch      * 1.6.0       2015-08-26
##  broom             * 0.3.7       2015-05-06
##  caTools             1.17.1      2014-09-10
##  cluster             2.0.3       2015-07-21
##  colorspace          1.2-6       2015-03-11
##  corpcor             1.6.8       2015-07-08
##  curl                0.9.2       2015-08-08
##  DBI               * 0.3.1       2014-09-24
##  dendextend        * 1.1.0       2015-07-31
##  DESeq2            * 1.8.1       2015-05-02
##  devtools          * 1.8.0       2015-05-09
##  digest              0.6.8       2014-12-31
##  dplyr             * 0.4.3       2015-09-01
##  edge              * 2.1.0       2015-09-06
##  evaluate            0.7.2       2015-08-13
##  foreign             0.8-66      2015-08-19
##  formatR             1.2         2015-04-21
##  Formula           * 1.2-1       2015-04-07
##  futile.logger       1.4.1       2015-04-20
##  futile.options      1.0.0       2010-04-06
##  gdata               2.17.0      2015-07-04
##  genefilter        * 1.50.0      2015-04-17
##  geneLenDataBase   * 1.4.0       2015-09-06
##  geneplotter         1.46.0      2015-04-17
##  GenomeInfoDb      * 1.4.2       2015-08-15
##  GenomicAlignments   1.4.1       2015-04-24
##  GenomicFeatures     1.20.2      2015-08-14
##  GenomicRanges     * 1.20.5      2015-06-09
##  genstats          * 0.1.02      2015-09-05
##  ggplot2           * 1.0.1       2015-03-17
##  git2r               0.11.0      2015-08-12
##  GO.db               3.1.2       2015-09-06
```

```
##  goseq             * 1.20.0       2015-04-17
##  gplots            * 2.17.0       2015-05-02
##  gridExtra           2.0.0        2015-07-14
##  gtable              0.1.2        2012-12-05
##  gtools              3.5.0        2015-05-29
##  highr               0.5          2015-04-21
##  HistData          * 0.7-5        2014-04-26
##  Hmisc             * 3.16-0       2015-04-30
##  htmltools           0.2.6        2014-09-08
##  httr                1.0.0        2015-06-25
##  IRanges           * 2.2.7        2015-08-09
##  KernSmooth          2.23-15      2015-06-29
##  knitr             * 1.11         2015-08-14
##  lambda.r            1.1.7        2015-03-20
##  lattice           * 0.20-33      2015-07-14
##  latticeExtra        0.6-26       2013-08-15
##  lazyeval            0.1.10       2015-01-02
##  limma             * 3.24.15      2015-08-06
##  lme4                1.1-9        2015-08-20
##  locfit              1.5-9.1      2013-04-20
##  magrittr            1.5          2014-11-22
##  MASS              * 7.3-43       2015-07-16
##  Matrix            * 1.2-2        2015-07-08
##  MatrixEQTL        * 2.1.1        2015-02-03
##  memoise             0.2.1        2014-04-22
##  mgcv              * 1.8-7        2015-07-23
##  minqa               1.2.4        2014-10-09
##  mnormt              1.5-3        2015-05-25
##  munsell             0.4.2        2013-07-11
##  nlme              * 3.1-122      2015-08-19
##  nloptr              1.0.4        2014-08-04
##  nnet                7.3-10       2015-06-29
##  org.Hs.eg.db      * 3.1.2        2015-07-17
##  plyr                1.8.3        2015-06-12
##  preprocessCore    * 1.30.0       2015-04-17
##  proto               0.3-10       2012-12-22
##  psych               1.5.6        2015-07-08
##  qvalue            * 2.0.0        2015-04-17
##  R6                  2.1.1        2015-08-19
##  RColorBrewer        1.1-2        2014-12-07
##  Rcpp              * 0.12.0       2015-07-25
##  RcppArmadillo     * 0.5.400.2.0  2015-08-17
##  RCurl               1.95-4.7     2015-06-30
##  reshape2            1.4.1        2014-12-06
##  rmarkdown           0.7          2015-06-13
##  rpart               4.1-10       2015-06-29
##  Rsamtools           1.20.4       2015-06-01
##  RSkittleBrewer    * 1.1          2015-09-05
##  RSQLite           * 1.0.0        2014-10-25
##  rstudioapi          0.3.1        2015-04-07
##  rtracklayer         1.28.9       2015-08-19
##  rversions           1.0.2        2015-07-13
##  S4Vectors         * 0.6.5        2015-09-01
##  scales              0.3.0        2015-08-25
```

```
##  snm                 1.16.0      2015-04-17
##  snpStats          * 1.18.0      2015-04-17
##  stringi             0.5-5       2015-06-29
##  stringr             1.0.0       2015-04-30
##  survival          * 2.38-3      2015-07-02
##  sva               * 3.14.0      2015-04-17
##  tidyr               0.2.0       2014-12-05
##  UsingR            * 2.0-5       2015-08-06
##  whisker             0.3-2       2013-04-28
##  XML                 3.98-1.3    2015-06-30
##  xml2                0.1.2       2015-09-01
##  xtable              1.7-4       2014-09-12
##  XVector             0.8.0       2015-04-17
##  yaml                2.1.13      2014-06-12
##  zlibbioc            1.14.0      2015-04-17
##  source
##  CRAN (R 3.2.0)
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.2)
##  Github (jdstorey/edge@a1947b5)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
##  Bioconductor
```

```
##  Bioconductor
##  Bioconductor
##  local
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.2)
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.1)
##  Bioconductor
##  Github (alyssafrazee/RSkittleBrewer@0a96a20)
```

```
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.2)
##  Bioconductor
##  Bioconductor
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  Bioconductor
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  CRAN (R 3.2.1)
##  CRAN (R 3.2.2)
##  CRAN (R 3.2.0)
##  Bioconductor
##  CRAN (R 3.2.0)
##  Bioconductor
```

It is also useful to compile the time the document was processed. This document was processed on: 2015-09-06.