PM 579: Statistical Analysis of High-Dimensional Data | Homework 3

Saket Choudhary skchoudh@usc.edu

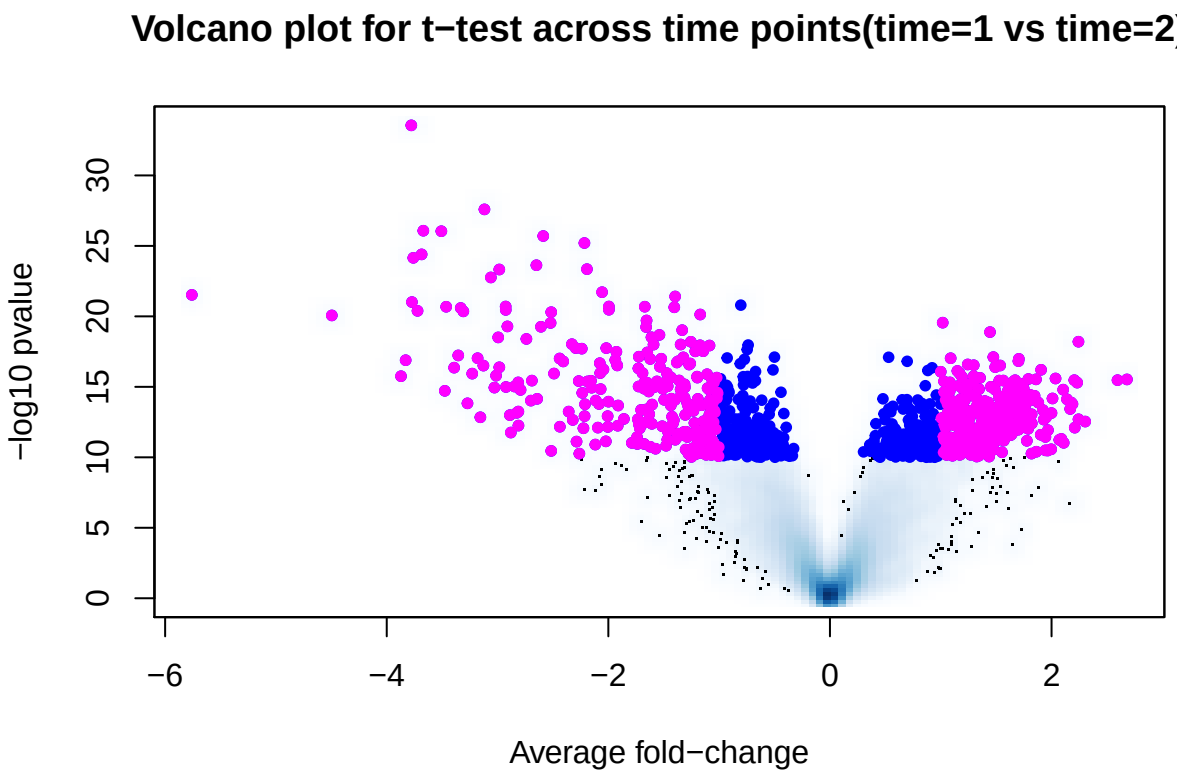
6/23/2016

Read Data

```
library(limma)
library(knitr)

load('stallcupdat.Rdata')
```

Volcano Plots

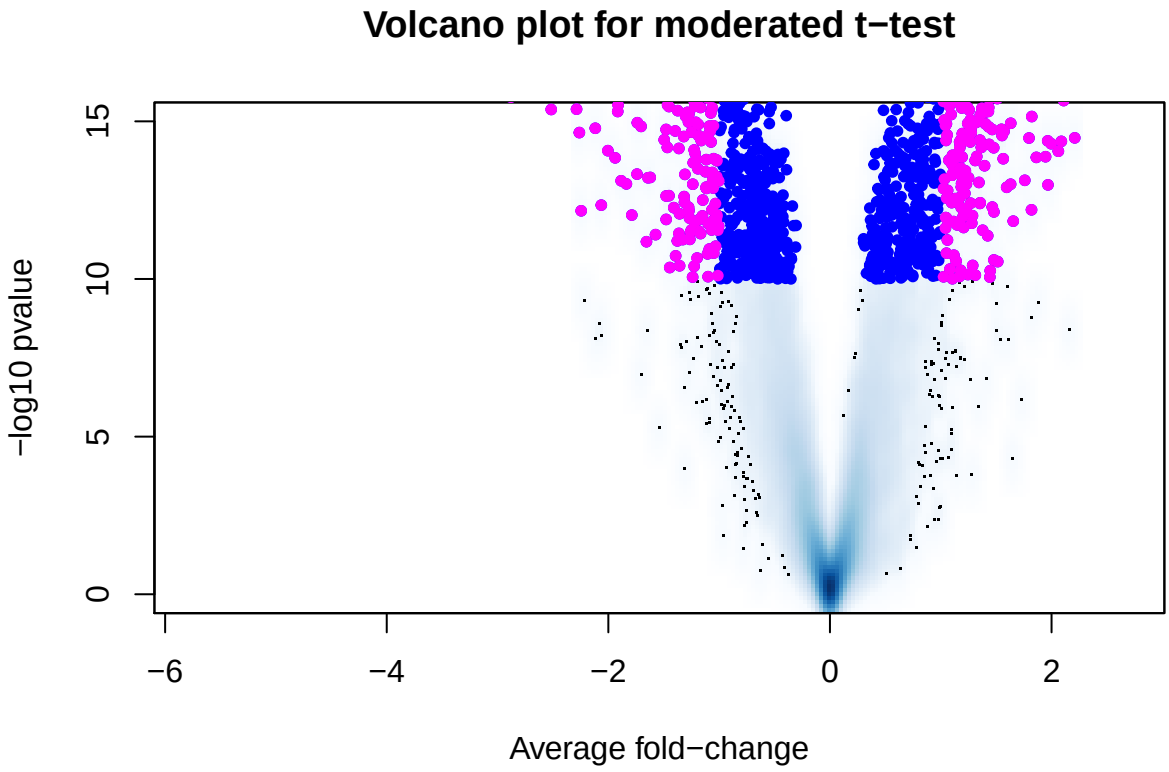
```
Index1 <- which(stallcupdat$target$time==1)
Index2 <- which(stallcupdat$target$time==2)
Index3 <- which(stallcupdat$target$time==3)

d <- rowMeans(stallcupdat$E[,Index2])-rowMeans(stallcupdat$E[,Index1])

tt <- rowttests(stallcupdat$E[, c(Index1,Index2)], factor(stallcupdat$targets$time[c(Index1,Index2)]))
lodt <- -log10(tt$p.value)
smoothScatter(d,
              lodt,
              nrpoints=500,
              xlab="Average fold-change",
              ylab="-log10 pvalue",
              main="Volcano plot for t-test across time points(time=1 vs time=2)")
points(d[lodt>10], lodt[lodt>10], pch=20, col=4)
points(d[lodt>10 & abs(d)>1], lodt[lodt>10 & abs(d)>1], pch=20, col=6)
```



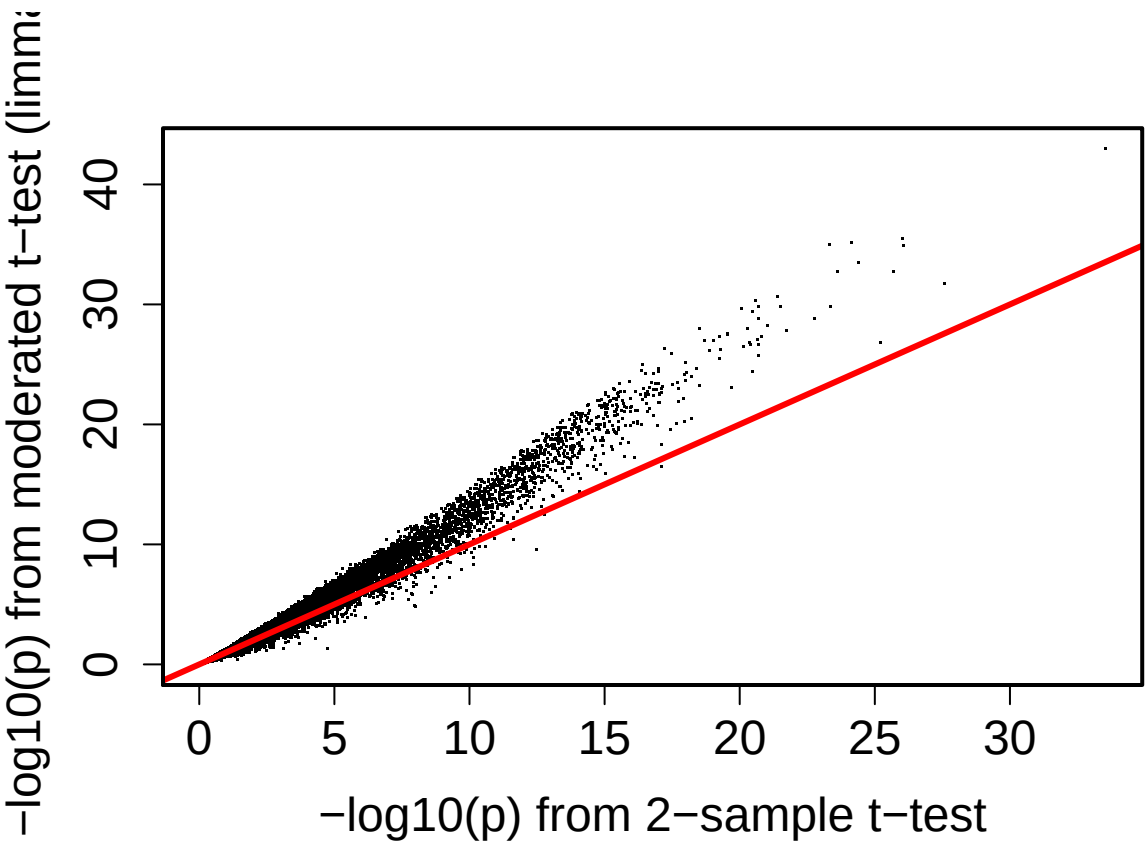
Based on t-tests, there seem to be enough differentially expressed genes between time points 1 and time points 2 (Across treatments+control)

Moderated t-tests

```
design <- model.matrix(~factor(stallcupdat$target$time))
fit <- lmFit(stallcupdat$E, design)
efit <- eBayes(fit)
lodmt <- -log10(efit$p.value[,2])

d <- rowMeans(stallcupdat$E[,Index2])-rowMeans(stallcupdat$E[,Index1])

smoothScatter(d,
              lodmt,
              nrpoints=500,
              main="Volcano plot for moderated t-test",xlab="Average fold-change",
              ylab="-log10 pvalue",ylim=c(0,15))
points(d[lodmt>10], lodmt[lodmt>10], pch=20, col=4)
points(d[lodmt>10 & abs(d)>1],lodmt[lodmt>10 & abs(d)>1], pch=20, col=6)
```




```
par(mar=c(5,5,3,2))
plot(lodt,lodmt,pch=".",cex.axis=1.5,cex.lab=1.5,
     xlab="-log10(p) from 2-sample t-test",
     ylab="-log10(p) from moderated t-test (limma)")
abline(0,1,col=2,lwd=3)
box(lwd=2)
```



As there are too few replicates(12 here) so the variance estimates can be stabilised using moderated t-tests. Clearly. in this case pooling variance information from similar genes helps as they p-value estimates improve.

Finding DE genes

```
time <- stallcupdat$targets$time
trts <- factor(stallcupdat$targets$treatment,
               levels=unique(stallcupdat$targets$treatment))
design <- model.matrix(~0+trts)
colnames(design) <- levels(trts)
kable(design)
```

Control2	X	Control1	Y
1	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	0	1	0

Control2	X	Control1	Y
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1
1	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1
1	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1
1	0	0	0
1	0	0	0
0	0	1	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1
1	0	0	0
1	0	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0

We define a more explicit design matrix without intercept.

Differential genes at time point 2

We define an contrast matrix which is more easy to interpret than using vectors

```
fit <- lmFit(stallcupdat$E[,c(time==2)],design[c(time==2),])

contr.matrix <- makeContrasts(XYsC1C2 = (X+Y)/2 - (Control1+Control2)/2,
```

```
                              XvsC1C2 = X-(Control1+Control2)/2,
                              YvsC1C2 = Y-(Control1+Control2)/2,
                              levels = design)
kable(contr.matrix )
```

	XYsC1C2	XvsC1C2	YvsC1C2
Control2	-0.5	-0.5	-0.5
X	0.5	1.0	0.0
Control1	-0.5	-0.5	-0.5
Y	0.5	0.0	1.0

```
fitgpd=contrasts.fit(fit,contr.matrix)
fitgpd=eBayes(fitgpd)
topTable(fitgpd,n=10)
```

```
##                 XYsC1C2    XvsC1C2    YvsC1C2     AveExpr        F
## ILMN_2392261  1.4636316  1.9561086  0.9711547  0.05572692 90.92310
## ILMN_1760103  1.5530846  2.0300292  1.0761400  0.18434020 79.23543
## ILMN_1692223 -1.9220230 -1.8575570 -1.9864891 -0.25493441 54.06754
## ILMN_1688184  0.7892799  0.7565609  0.8219990  0.21793060 53.20783
## ILMN_1721559  1.5286137  1.9785469  1.0786805  0.06988887 51.27430
## ILMN_1678143  1.1178679  1.4761343  0.7596015  0.23695775 50.49902
## ILMN_1767685  1.5999404  1.6852403  1.5146405  0.30349530 49.89704
## ILMN_1765668  1.1264200  1.3279808  0.9248592 -0.12808822 48.10125
## ILMN_1744765 -2.1002668 -2.1216260 -2.0789077  0.22195917 46.93273
## ILMN_1659610 -1.2895246 -1.4403356 -1.1387137 -0.17105952 45.92719
##                   P.Value    adj.P.Val
## ILMN_2392261 6.323691e-09 0.0001550948
## ILMN_1760103 1.583799e-08 0.0001942212
## ILMN_1692223 1.912231e-07 0.0011159689
## ILMN_1688184 2.118156e-07 0.0011159689
## ILMN_1721559 2.680453e-07 0.0011159689
## ILMN_1678143 2.952292e-07 0.0011159689
## ILMN_1767685 3.185103e-07 0.0011159689
## ILMN_1765668 4.013873e-07 0.0012305530
## ILMN_1744765 4.684749e-07 0.0012766462
## ILMN_1659610 5.365566e-07 0.0013159586
```

Here XvsC1C2 implies that differential expression was calculated by averaging over Control1 and Control2 values.

A total of 193 genes are upregulated and 158 downregulated when comparing average X+Y over average C1+C2 at time point 2 alone (i.e. not accounting for common genes with X vs C1C2 or YvsC1C2)

Also, around 735 genes are up and 582 genes are downregulated in the common region of XvsC1C2, XYvsC1C2, YvsC1C2 indicating that X,Y are similar expression wise.

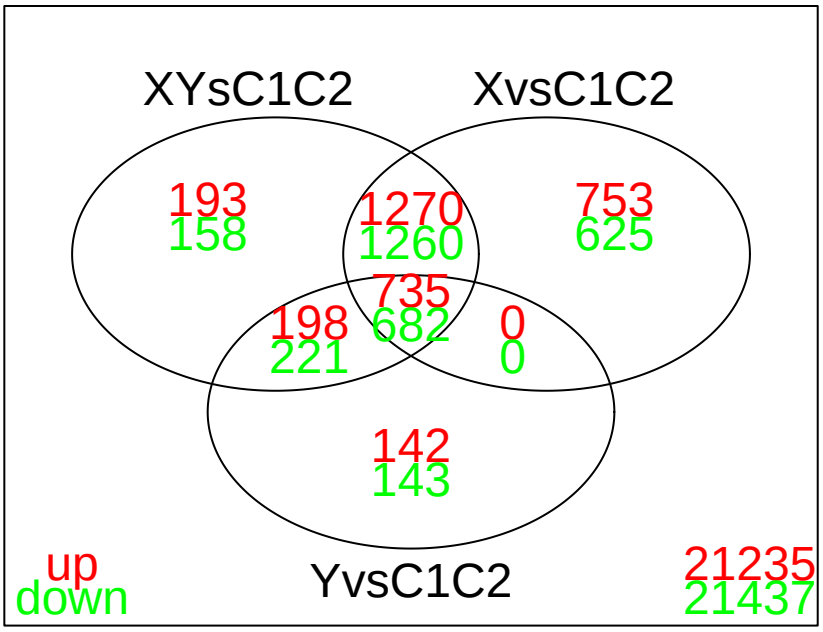
```
results <- decideTests(fitgpd,adjust.method="none")
a <- vennCounts(results)
print(a)
```

```
##   XYsC1C2 XvsC1C2 YvsC1C2 Counts
## 1       0       0       0  18148
## 2       0       0       1    283
## 3       0       1       0   1376
## 4       0       1       1      2
## 5       1       0       0    351
## 6       1       0       1    419
## 7       1       1       0   2530
## 8       1       1       1   1417
## attr(,"class")
## [1] "VennCounts"
```

```
head(results,n=10)
```

```
##               Contrasts
##                XYsC1C2 XvsC1C2 YvsC1C2
##   ILMN_1762337       0       0       0
##   ILMN_2055271       0       0       0
##   ILMN_2383229       0       0       0
##   ILMN_1806310       0       0       0
##   ILMN_1779670       0       0       0
##   ILMN_2321282       0       0       0
##   ILMN_1772582       0       0       0
##   ILMN_1717783       0       0       0
##   ILMN_1814316       0       0       0
##   ILMN_2359168       0       0       0
```

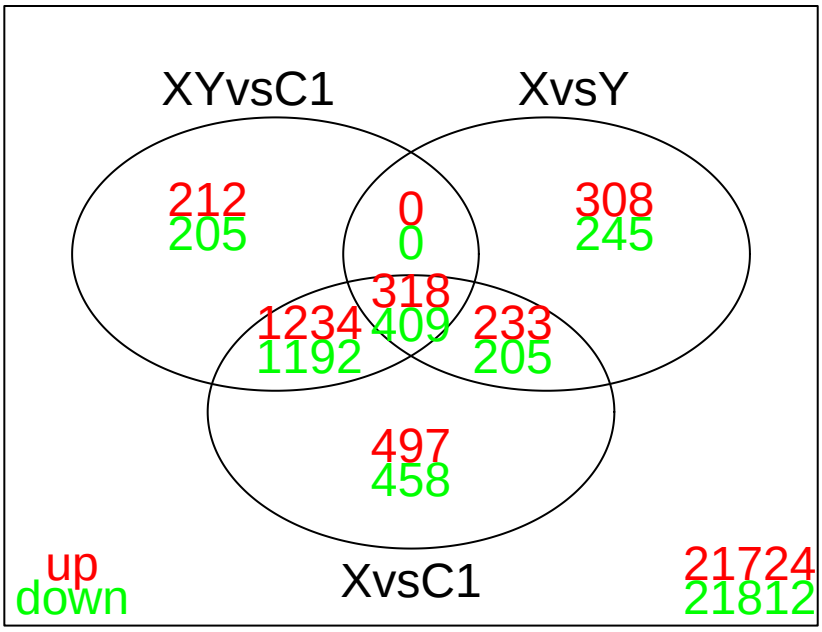
```
vennDiagram(results,include=c("up","down"),counts.col=c("red","green"), main="Time point = 2")
```



Let's compare XvsY and XvsC1 and YvsC1

```
contr.matrix <- makeContrasts(XYvsC1 = (X+Y)/2 - (Control1), XvsY = X-Y, XvsC1= X-Control1, levels = des
fitgpd <- contrasts.fit(fit,contr.matrix)
fitgpd <- eBayes(fitgpd)
results <- decideTests(fitgpd, adjust.method="none")
cs <- vennCounts(results)
vennDiagram(results, include=c("up","down"),counts.col=c("red","green"), main="Time point = 2")
```

Time point = 2




```
head(results,n=10)
```

```
##                Contrasts
##                 XYvsC1 XvsY XvsC1
##   ILMN_1762337       0    0     0
##   ILMN_2055271       0    0     0
##   ILMN_2383229       0    0     0
##   ILMN_1806310       0    0     0
##   ILMN_1779670       0    0     0
##   ILMN_2321282       0    0     0
##   ILMN_1772582       0    0     0
##   ILMN_1717783       0    0     0
##   ILMN_1814316       0    0     0
##   ILMN_2359168       0    0     0
```

```
print(vennCounts(results,include=c("up")))
```

```
##   XYvsC1 XvsY XvsC1 Counts
## 1      0    0     0  21724
## 2      0    0     1    497
## 3      0    1     0    308
## 4      0    1     1    233
## 5      1    0     0    212
## 6      1    0     1   1234
## 7      1    1     0      0
## 8      1    1     1    318
## attr(,"class")
## [1] "VennCounts"
```

```
print(vennCounts(results,include=c("down")))
```

```
##   XYvsC1 XvsY XvsC1 Counts
## 1      0    0     0  21812
## 2      0    0     1    458
## 3      0    1     0    245
## 4      0    1     1    205
## 5      1    0     0    205
## 6      1    0     1   1192
## 7      1    1     0      0
## 8      1    1     1    409
## attr(,"class")
## [1] "VennCounts"
```

A striking number that stands out is 0 up and downregulated genes between XYvsC1 and XvsY, thus indicating that there are no such genes which are diff expressed in (X+Y)vsC1C2 or XvsY or XvsC1 however there are genes which are in common up or down regulated in all these groups.

Differential genes at time point 3

```
fit <- lmFit(stallcupdat$E[,c(time==3)],design[c(time==3),])

contr.matrix <- makeContrasts(XYsC1C2 = (X+Y)/2 - (Control1+Control2)/2,
                              XvsC1C2 = X-(Control1+Control2)/2,
                              YvsC1C2 = Y-(Control1+Control2)/2,
                              levels = design)
kable(contr.matrix )
```

	XYsC1C2	XvsC1C2	YvsC1C2
Control2	-0.5	-0.5	-0.5
X	0.5	1.0	0.0
Control1	-0.5	-0.5	-0.5
Y	0.5	0.0	1.0

```
fitgpd=contrasts.fit(fit,contr.matrix)
fitgpd=eBayes(fitgpd)
kable(topTable(fitgpd,n=10))
```

	XYsC1C2	XvsC1C2	YvsC1C2	AveExpr	F	P.Value	adj.P.Val
ILMN_1692223	-1.8940198	-1.9449983	-1.8430414	-0.4530204	82.71110	0e+00	0.0001927
ILMN_1721559	1.4268679	1.6331010	1.2206347	0.1593326	71.53020	0e+00	0.0002601
ILMN_1701424	1.0294645	1.2389639	0.8199651	0.2988811	65.88971	0e+00	0.0003023
ILMN_2382290	0.8318677	0.8472919	0.8164435	0.2804587	52.94152	2e-07	0.0009729
ILMN_2096985	-0.8602917	-1.0186670	-0.7019164	-0.1337231	50.87320	2e-07	0.0010107
ILMN_1755811	-0.7159714	-0.8606184	-0.5713244	-0.0355728	48.77121	3e-07	0.0011091
ILMN_1792679	0.6616360	0.7118095	0.6114625	0.2222735	45.79061	4e-07	0.0014301
ILMN_2072622	0.7368801	1.0632469	0.4105133	0.3582748	43.31539	6e-07	0.0015061

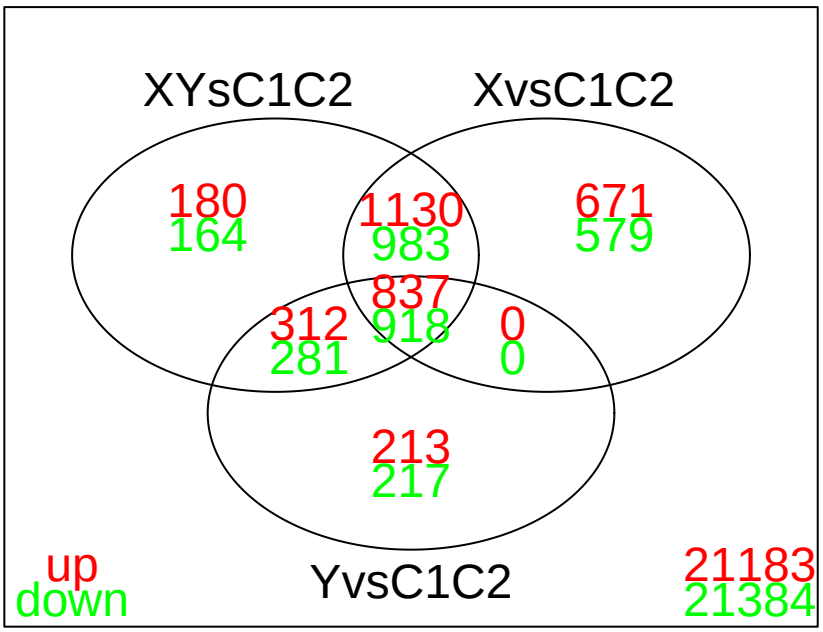
	XYsC1C2	XvsC1C2	YvsC1C2	AveExpr	F	P.Value	adj.P.Val
ILMN_1702301	0.4475621	0.4663214	0.4288029	-0.0008716	41.71393	7e-07	0.0015061
ILMN_1775829	-0.7633448	-0.9708421	-0.5558475	-0.0260578	41.60169	8e-07	0.0015061

Lots of common genes between XYvsC1C2 and XvsC1C2

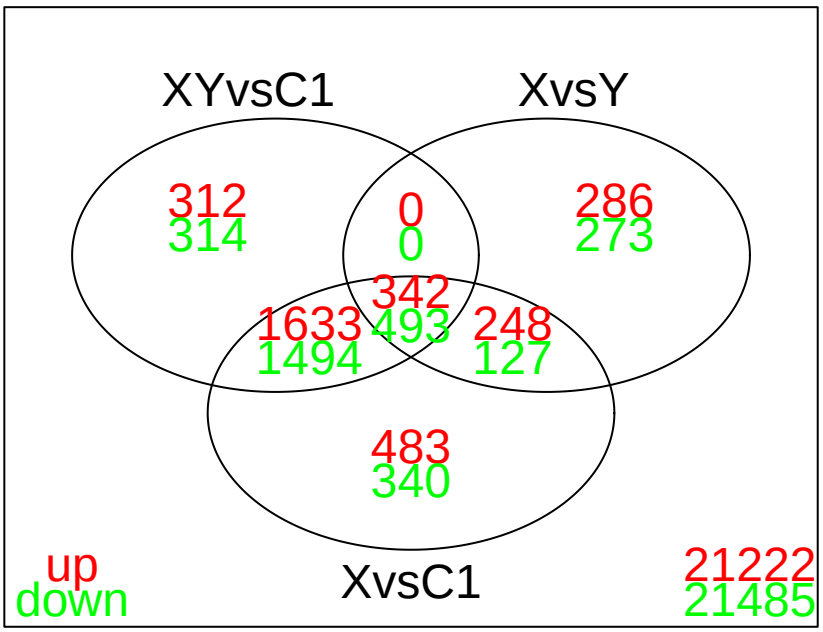
```
results <- decideTests(fitgpd,adjust.method="none")
a <- vennCounts(results)
head(results,n=10)
```

```
##                Contrasts
##                 XYsC1C2 XvsC1C2 YvsC1C2
##    ILMN_1762337       0       0       0
##    ILMN_2055271       0       0       0
##    ILMN_2383229       0       0       0
##    ILMN_1806310       0       0       0
##    ILMN_1779670       0       0       0
##    ILMN_2321282       0       0       0
##    ILMN_1772582      -1      -1       0
##    ILMN_1717783       0       1       0
##    ILMN_1814316       0       0       0
##    ILMN_2359168       0       0       0
```

```
vennDiagram(results,include=c("up","down"),counts.col=c("red","green"), main="Time point = 3")
```



```
contr.matrix <- makeContrasts(XYvsC1 = (X+Y)/2 - (Control1), XvsY = X-Y, XvsC1= X-Control1, levels = des
fitgpd <- contrasts.fit(fit,contr.matrix)
fitgpd <- eBayes(fitgpd)
results <- decideTests(fitgpd, adjust.method="none")
cs <- vennCounts(results)
vennDiagram(results, include=c("up","down"),counts.col=c("red","green"), main="Time point = 3")
```




```
head(results,n=10)
```

```
##                Contrasts
##                 XYvsC1 XvsY XvsC1
##   ILMN_1762337       0    0     0
##   ILMN_2055271       0    0     0
##   ILMN_2383229       0    0     0
##   ILMN_1806310       0    0     0
##   ILMN_1779670       0    0     0
##   ILMN_2321282       0    0     0
##   ILMN_1772582      -1    0    -1
##   ILMN_1717783       0    0     0
##   ILMN_1814316       0    0     0
##   ILMN_2359168       0    0     0
```

```
print(vennCounts(results,include=c("up")))
```

```
##   XYvsC1 XvsY XvsC1 Counts
## 1      0    0     0  21222
## 2      0    0     1    483
```

```
## 3      0    1     0    286
## 4      0    1     1    248
## 5      1    0     0    312
## 6      1    0     1   1633
## 7      1    1     0      0
## 8      1    1     1    342
## attr(,"class")
## [1] "VennCounts"
```

```
print(vennCounts(results,include=c("down")))
```

```
##   XYvsC1 XvsY XvsC1 Counts
## 1      0    0     0  21485
## 2      0    0     1    340
## 3      0    1     0    273
## 4      0    1     1    127
## 5      1    0     0    314
## 6      1    0     1   1494
## 7      1    1     0      0
## 8      1    1     1    493
## attr(,"class")
## [1] "VennCounts"
```

Again no common genes between XYvsC1 and XvsY which are not in XvsC1.

DE genes associated with time

Reference: https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf Section 9.6

```
time <- factor(stallcupdat$targets$time)
trts <- factor(stallcupdat$targets$treatment,
               levels=unique(stallcupdat$targets$treatment))

design <- model.matrix(~0+trts*time)
colnames(design) <- c('Control2', 'X', 'Control1', 'Y', 'time2', 'time3', 'X.time2', 'Control1.time2',
fit <- lmFit(stallcupdat$E,design)
contr.matrix <- makeContrasts(XYvsC1C2  = (X+Y)/2-(Control1+Control2)/2,
                              X3Y3vsX2Y2 = (X.time3+Y.time3)/2-(X.time2+Y.time2)/2,
                              X3vsX2 = X.time3-X.time2,
                              levels = design)
kable(contr.matrix)
```

	XYvsC1C2	X3Y3vsX2Y2	X3vsX2
Control2	-0.5	0.0	0
X	0.5	0.0	0
Control1	-0.5	0.0	0
Y	0.5	0.0	0
time2	0.0	0.0	0
time3	0.0	0.0	0
X.time2	0.0	-0.5	-1

	XYvsC1C2	X3Y3vsX2Y2	X3vsX2
Control1.time2	0.0	0.0	0
Y.time2	0.0	-0.5	0
X.time3	0.0	0.5	1
Control1.time3	0.0	0.0	0
Y.time3	0.0	0.5	0

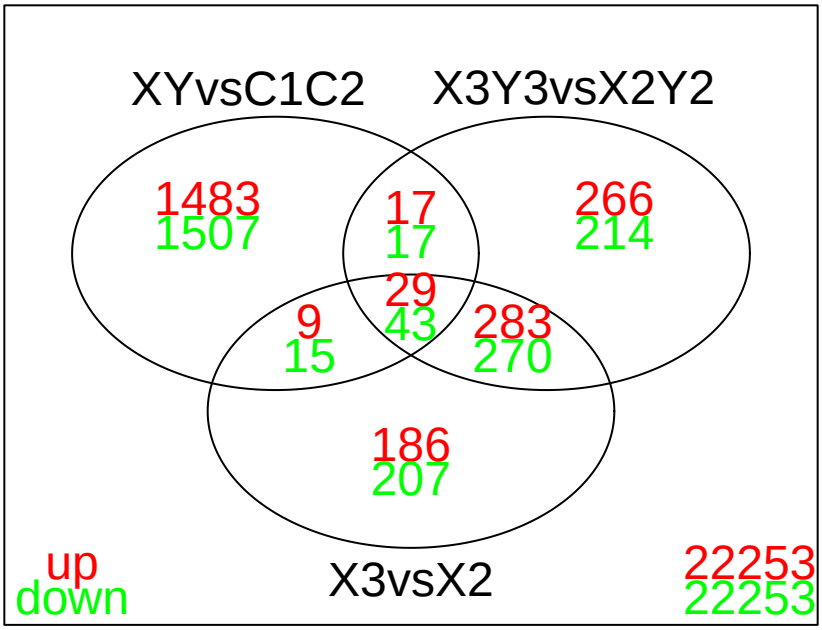
```
fitgpd=contrasts.fit(fit,contr.matrix)
fitgpd=eBayes(fitgpd)
kable(topTable(fitgpd,n=10))
```

	XYvsC1C2	X3Y3vsX2Y2	X3vsX2	AveExpr	F	P.Value	adj.P.Val
ILMN_2167758	-1.3896028	-0.0688103	-0.1023459	0	44.19823	0	0.00e+00
ILMN_1692223	-2.1329445	0.0127492	-0.1026954	0	41.95023	0	0.00e+00
ILMN_1655595	1.0482834	0.0033124	0.0069149	0	40.00381	0	0.00e+00
ILMN_1767685	1.7273583	-0.1139329	0.0082781	0	35.23824	0	2.00e-07
ILMN_1716658	-0.9511261	0.0106308	-0.0062736	0	34.50571	0	2.00e-07
ILMN_1755796	-1.3112695	-0.6494273	-0.5966956	0	32.94084	0	4.00e-07
ILMN_1765668	1.3044230	-0.2756104	-0.3990895	0	32.56274	0	4.00e-07
ILMN_1768577	-1.0893185	-0.2411923	-0.2485188	0	28.40039	0	2.10e-06
ILMN_1652407	0.6911198	-0.1529714	-0.1781824	0	25.46233	0	7.60e-06
ILMN_2388547	0.7136707	0.3246403	0.6653724	0	24.63418	0	1.04e-05

```
results <- decideTests(fitgpd,adjust.method="none") # doesn't subset of F<0.05'
a <- vennCounts(results)
print(a)
```

```
##   XYvsC1C2 X3Y3vsX2Y2 X3vsX2 Counts
## 1        0          0      0  20073
## 2        0          0      1    366
## 3        0          1      0    447
## 4        0          1      1    520
## 5        1          0      0   2897
## 6        1          0      1     51
## 7        1          1      0     67
## 8        1          1      1    105
## attr(,"class")
## [1] "VennCounts"
```

```
vennDiagram(results, include=c("up","down"),counts.col=c("red","green"))
```



The number of up or down regulated genes are more affected by the treatment type than by time as these numbers are less

when comparing based on time points.