```
Coefficients of linear discriminants:
              LD1     LD2     LD3     LD4     LD5     LD6
hdlngth  -0.15053  0.0583 -0.2557 -0.0124 -0.0819 -0.1871
skullw   -0.02653  0.0399 -0.2498  0.1245 -0.1365  0.1438
totlngth  0.10696  0.2792  0.3052 -0.1849 -0.1390 -0.0892
taill    -0.45063 -0.0896 -0.4458 -0.1730  0.3242  0.4951
footlgth  0.30190 -0.0360 -0.0391  0.0756  0.1191 -0.1261
earconch  0.58627 -0.0436 -0.0731 -0.0882 -0.0638  0.2802
eye      -0.05614  0.0892  0.7845  0.4644  0.2848  0.2972
chest     0.09062  0.1042  0.0498  0.1275  0.6475 -0.0787
belly     0.00997 -0.0517  0.0936  0.1672 -0.2903  0.1939

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6
0.8927 0.0557 0.0365 0.0082 0.0047 0.0022
```

The “proportion of trace” figures are the proportions of the between class variance that are explained by the successive linear combinations. For these data, the first linear discriminant does most of the discriminating. The variables that seem important are `hdlngth`, and `taill`, contrasted with `totlngth` and `footlgth`.

We invite the reader to repeat the analysis with the argument `CV=TRUE` in the call to `lda()`, in order to obtain a realistic predictive accuracy estimate.

## 9.3 *High-dimensional data — RNA-Seq gene expression

Subsection 2.8.3 demonstrated how, given numerous $p$-values that were generated in the same experiment or series of experiments, false discovery rate (FDR) estimates could be obtained. Here, we resume that discussion, using data from an experiment where plants were exposed to one of three treatments. Data are from Peter Crisp, obtained in the course of his PhD work in the ARC Centre of Excellence in Plant Energy Biology at Australian National University. The treatments were:

- a control;
- light stress, i.e., one hour of continuous exposure to light at ten times the level that the plants are normally grown under;
- drought stress,i.e., nine days without water, causing wilting of the leaves.

The interest is in how light stress and drought stress affect gene expression to produce proteins. Gene activity in the production of proteins was monitored by using RNA sequencing technology to determine, for each gene, the number of mRNA (messenger RNA) sequences, in a sample of plant tissue, that carry that gene’s information. (It is well to note that the mRNA counts measure only the activity of the cell machinery in production of protein. The relationship to protein production will be different for different genes, and affected also by other factors.)

After removing very sparse counts there were counts, for each of the three treatments, for 18,568 genes. On average, in the absence of any real effect, 5% or approximately 928 of the genes can be expected to show a difference at the 5% significance level. We can then note the number $m$ of differences observed at the 5% level, and take those with the smallest $m$-928 $p$-values as probably real. This is the basis for the Benjamini-Hochberg False Discrimination Rate (FDR) approach that was introduced in Subsection 2.8.3.

### *Brief note on mRNA technical issues

The role of the mRNA is to carry information encoded in the genes to the cell factories (ribosomes) that manufacture the proteins. Links on `https://ghr.nlm.nih.gov/handbook/hgp/genome`, or on another such site, can be consulted for more details on the biological mechanisms. The counts are, inevitably, sums over a very large number of plant cells.

### 9.3.1 Data and design matrix setup

The counts are in the matrix `DAAGbio::plantStressCounts`. The column names identify the samples:

```
counts <- DAAGbio::plantStressCounts
colSums(counts)
```

```
    CTL1     CTL2     CTL3   Light1   Light2   Light3 Drought1 Drought2 
  933573   943262   944871   946921   926570   931086   925995   915023 
Drought3 
  930588 
```

The column sums are in each case close to a million.

The function `cpm()` (counts per million reads) standardizes counts for each gene by dividing by entries in the `lib.size` column and multiplying by $10^6$. Here, the `lib.size` values are just the column totals. We retain the genes with a count per million of at least one in three of the 9 samples:

```
## Require at least 3 counts per million that are > 1
keep <- rowSums(counts)>=3
counts <- counts[keep,]
```

The `limma` package is set up to work with a design matrix that specifies the treatment structure. As there are just three treatments to compare, the design matrix can be set up to have one column for each treatment.

```
treatment <- factor(rep(c("CTL", "L", "D"), rep(3,3)))
design <- model.matrix(~0+treatment)
colnames(design) <- levels(treatment)
```

Note the use of L for Light stress and D for Drought stress.

Now apply the function `limma::voom()`, which transforms the counts to $\log(\text{count} + 1)$. At the same time, it estimates, as a smooth curve, the mean-variance relationship. Optionally, it will output a plot that shows the mean-variance relationship. Weights (= inverse of variance) for individual genes will be required for fitting the model.

```
library(limma)
v <- voom(counts, design)
```

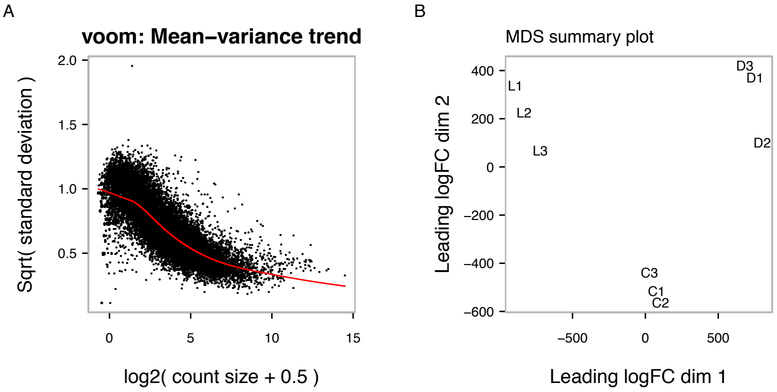

Figure 9.6: The left panel shows the mean-variance relationship. The right panel plots results from the use of multi-dimensional scaling to locate samples in two-dimensional space. By default, the 'top' 500 genes are used. As there are three groups, two dimensions suffice.

### *A two-dimensional representation*

Multi-dimensional scaling, using the function `plotMDS()`, can be used to give a broad overall comparison of the nine samples. Figure 9.6A shows this plot. Panel B shows a summary plot that was obtained using multi-dimensional scaling.

```
v <- voom(counts, design, fg="gray", plot=TRUE)
firstchar <- substring(colnames(counts),1,1)
plotMDS(counts, labels=paste0(firstchar, rep(1:3,3)), cex=0.8)
```

Plots such as are shown in Figure 9.6B can be important in drawing attention to samples where something is clearly wrong. There is an example in the limma user guide, where there appears to be a batch effect that is associated with sequencing type and date. Under Section 18.2 on "Differential Splicing after Pasilla Knockdown", see Subsection 18.2.8 on “Scale Normalization”. (The section and subsection numbers are for the 17 June 2014 version of the manual.)

### *Fitting the model*

We then proceed to fit the model:

```
fit <- lmFit(v, design)
```

We then specify the treatment contrasts in which we are interested, and extract information about these:

```
contrs <- c("D-CTL", "L-CTL", "L-D")
contr.matrix <- makeContrasts(contrasts=contrs,
                              levels=levels(treatment))
fit2 <- contrasts.fit(fit, contr.matrix)
efit2 <- eBayes(fit2)
```

Recall that `L` is light stress and `D` is drought stress. The function `eBayes()` uses an empirical Bayes method to shrink the probe-wise sample variances in towards a common value, with a consequent increase in the degrees of freedom for the individual variances. It is required here as a preliminary to examining output from efit2.

### 9.3.2 From p-values to false discovery rate (FDR)

The function topTable() sorts output according to whatever criterion is used to determine "top"; the default is to take first genes where *p*-values are smallest. Included in the output is a column adj.P.val. With the default argument adjust.method="BH", this returns the Benjamini-Hochberg false discovery rate (FDR) estimates that were discussed in Section 2.8.3. The FDR estimates can alternatively be obtained by direct use of the function call p.adjust(p=fit2$p.value[,1], method="BH").

```
## First contrast only; Drought-CTL
print(round(topTable(efit2, coef=1, number=4),15), digits=3)
```

```
          logFC AveExpr    t           P.Value         adj.P.Val    B
Gene24491  3.69     9.6 35.5 0.000000000000000 0.000000000000039 32.2
Gene13749  2.23    10.3 28.4 0.000000000000000 0.000000000001099 28.4
Gene10904  2.62    10.6 26.3 0.000000000000000 0.000000000002088 27.1
Gene13210  2.58    10.0 26.1 0.000000000000001 0.000000000002088 26.9
```

An alternative is to do an ANOVA-like overall check for differential expression, thus:

```
print(round(topTable(efit2, number=4), 16), digits=3)
```

```
          D.CTL  L.CTL   L.D AveExpr   F P.Value          adj.P.Val
Gene10714 -3.25  2.208  5.46    8.69 865       0 0.0000000000000063
Gene24491  3.69 -0.153 -3.85    9.60 858       0 0.0000000000000063
Gene11934 -1.03  3.386  4.41    8.19 802       0 0.0000000000000077
Gene13377 -2.30  3.273  5.57    8.58 764       0 0.0000000000000090
```

Notice that the *p*-values and "adjusted" *p*-values are now, in the cases shown, smaller. This is to be expected, for most but not all cases, because the $F$-statistic summarizes evidence for differential expression across all three contrasts.

The function decideTests() provides, by default with an adjusted *p*-value or FDR of 0.05, a matrix that scores each of the three contrasts as -1 (effect, in opposite direction to contrast), 0 (as judged by the chosen criterion, no effect), and 1 (effect, in same direction as the contrast). The first five rows of output are:

```
head(decideTests(fit2),5)
```

```
       Contrasts
        D-CTL L-CTL L-D
  Gene1     0     0   0
  Gene2     0     1   1
  Gene3     0     0   1
  Gene4     0     0   0
  Gene6     0     0   0
```

A summary is:

```
summary(decideTests(fit2))
```

```
   D-CTL L-CTL   L-D
-1  3215  2538  3162
0  13445 15134 13156
1   2291  1279  2633
```

```
## Try also
## summary(decideTests(fit2, p.value=0.001))
```

This very high level of differential expression is surely a comment on the plant's priorities.

## 9.4 High dimensional data from expression arrays

Each of the 7129 rows of the matrix hddplot::Golub holds "gene expression indices" (variables, or "features") that are measures of the biological activity of a gene. (The technology used is now to an extent superseded by RNA-based approaches, or by more direct measurement of the protein created.) The dataset hddplot::Golub is the result of some further pre-processing of data that is in the golubEsets package. See Golub et al. (1999) for details of the work that generated the golubEsets data.

Following a terminology that is common for such data, the variables will be called features. Each of the 72 observations (columns of Golub) is from a tissue sample from a cancer patient. The 72 observations are classified into one of the three cancer types: ALL B-type (coded allB), ALL T-type (coded allT) and AML (coded aml). ALL is Acute Lymphoblastic Leukemia (lymphoblastic = producing lymph tissue), while AML is Acute Myoblastic Leukemia (myoblastic = producing muscle tissue).

The data frame golubInfo has information on the tissue samples, which are the observations. As well as different sexes, there are two different body tissues (bone marrow and peripheral blood). There may also be variation because the tissues came from four different hospitals; this will not be pursued here. These differences within the cancer types are a complication for investigating differences between the cancers.

The presence of these other factors makes graphical exploration especially important, with the initial focus (and here, the only focus) on differences within cancer types. Finding suitable views of the data, inevitably low-dimensional, is a challenge. Views are required that may help reveal subgroups in the data, or points that may have been misclassified, or between group differences in the variance-covariance structure. Graphs should be revealing, without serious potential to mislead.

Note: One use for data of this general type might be the finding of a discrimination rule that, using a small subset of the features, will allow discrimination between the different cancer types. A diagnostic device (a "probe") could then be designed that, given a new sample, could determine the cancer type. Note however that any classification of cancers is likely to conceal large individual differences that, in many cancers, arise from random differences in the timing and outcome of trigger points in a cascade of genomic damage and disruption.

The papers Maindonald and Burden (2005), Ambroise and McLachlan (2002) and Zhu et al. (2006) are useful background reading for the discussion of this section.