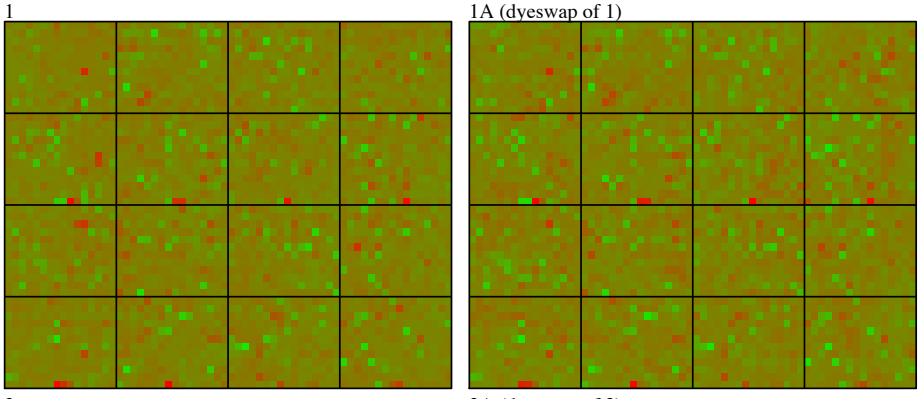

Figure 2.22: This false color image shows the intensity of the post signal (red), relative to the pre signal (green), for the first two of six half-slides ("panels") in a two channel microarray gene expression experiment. Use of one dye-swap pair per slide was designed to allow adjustment for any systematic red-green bias.

2.8.3 Severe multiplicity — the false discovery rate*

The dataset `DAAG::coralPval` that is the subject of the following discussion was generated using a microarray gene expression technology. Microarrays are now increasingly being replaced by the more direct measurements of gene activity in the cell that the RNA-Seq technology provides. In either case, a single experiment may yield information on thousands, or tens of thousands, of genes. The present data are from experimental work that was designed to compare gene expression, for the 3042 genes investigated, between two life-stages of coral — the pre-settlement free-swimming stage, and post-settlement. Each of the full complement of six panels (two only are shown in Figure 2.22) had 3072 spots; this included 30 blanks. Where there was an increase, the spot should be fairly consistently red, or reddish, over all six panels. Where there was a decrease, the spot should be fairly consistently green, or greenish. Results from the six sets of comparisons were used to generate 3042 p-values, one for each of 3042 sets of spots.

The methodology that will be described has wide application, to any form of comparison that generates large numbers of p-values — hundreds, or thousands, or more. The multiplicity of p-values allows inferences that an individual p-value does not provide. It allows the estimation of a false discovery rate (FDR).

**Microarrays and alternatives — technical note*

In the experimental procedure and subsequent processing that led to the plots shown in Figure 2.22, the slides are first printed with probes, with one probe per spot, each designed to check for the expression of one gene. The two samples carry labeling with separate fluorescent dyes so that when later a spot "lights up" under a scanner, the relative intensities of the two dye frequencies will provide a measure of differences in the signal intensity.

After labeling the separate samples, mixing them, and wiping the mixture over the slide or half-slide, and various laboratory processing steps, a scanner was used to determine, for

each spot, the intensities generated from the two samples. Various corrections are then necessary, leading finally to the calculation of logarithms of intensity ratios. Essentially, it is logarithms of intensity ratios that are shown in Figure 2.22.

For further information on the statistical analysis of microarray data, see Smyth (2004). With suitable pre-processing of the data, the methods carry over to the analysis of RNA-Seq data. See Law et al. (2014). For background on the coral data, see Grasso et al. (2008).

The false discovery rate (FDR)

The object `DAAG::coralPval` has 3072 p-values from the gene expression data represented in Figure 2.22.

The following calculates, for several different thresholds `pcrit` = p_{crit}, the total number of genes detected as differentially expressed with threshold as the threshold:

```
coralPval <- DAAG::coralPval
pcrit <- c(0.05, 0.02, 0.01, 0.001)
under <- sapply(pcrit, function(x)sum(coralPval<=x))
```

The numbers expected under the null hypothesis, in each case, are:

```
expected <- pcrit*length(coralPval)
```

These numbers can be conveniently set out in a table, allowing us to examine the implications of choosing one or other of these thresholds.

```
fdrtab <- data.frame(Threshold=pcrit, Expected=expected,
                     Discoveries=under, FDR=round(expected/under, 4))
print(xtable::xtable(fdrtab), include.rownames=FALSE, hline.after=FALSE)
```

Threshold	Expected	Discoveries	FDR
0.05	153.60	1310	0.12
0.02	61.44	1068	0.06
0.01	30.72	900	0.03
0.00	3.07	491	0.01

The column headed FDR is just the number of detections ("discoveries") expected under the null hypothesis, divided by the actual number detected. Although often described as an adjusted p-value, the result of the adjustment is not a p-value, but an estimate of the false discovery rate. For the false discovery rate to equal 0.05, the unadjusted p-value threshold should be set somewhere between 0.01 and 0.02.

The Benjamini-Hochberg method for adjusting p-values relies, in essence, on the argument just given. Rather than finding an unadjusted p-value threshold, it is however more straightforward to work directly with adjusted values, calculated as will now be described. After sorting the p-values from smallest to largest, the calculation is:

$$p_{adj[i]} = \frac{m}{i} p_i; \quad i = 1, 2, \ldots m$$

A further tweak is to set each $p_{adj[i]}$ to the smallest value, if any, that appears later in the sequence. This ensures that $p_{adj[i]}$ is a monotonic function of p_i. (Also, any value that is

greater than 1.0 is set to 1.) The function `p.adjust()` (`stats` package in base R), can be used (specify `method="BH"`) to do the adjustments, thus:

```
fdr <- p.adjust(coralPval, method="BH")
```

Here are numbers that fall under thresholds 0.05, 0.04, 0.02, and 0.01:

```
fdrcrit <- c(0.05, 0.04, 0.02, 0.01)
under <- sapply(fdrcrit, function(x)sum(coralPval<=x))
setNames(under, paste(fdrcrit))
```

```
0.05 0.04 0.02 0.01
1310 1234 1068  900
```

The FDR for a cutoff of 0.05 is a composite value, with some genes that fall under this threshold having a FDR much greater than 0.05, and many more having an FDR that is much less. The discussion that now follows shows how this composite FDR can be broken apart. Take p_{45} as the false discovery rate for genes in the range 0.04 < fdrcrit <= 0.05. Then the 1310 genes with fdrcrit <= 0.05 are comprised thus:

- 1310 - 1234 = 76 genes with an average FDR of p_{45}
- 1234 genes with an average FDR of 0.04

Then

$$p_{45} \times 76 + 0.04 \times 1234 = 0.05 \times 1310$$

Solving for p_{45} yields, rounded to two decimal places

$$p_{45} = 0.21$$

As with use of the p <= 0.05 criterion for a single p-value, it is tempting to place greater weight than is warranted on an FDR statistic that falls just under 0.05.

The average estimated false discovery rate for genes with 0.01 < fdrcrit <= 0.02, is:

$$\frac{0.02 \times 1068 - 0.01 \times 900}{1068 - 900} = 0.07$$

This line of argument can be combined with the assumption of a smooth change in the FDR to provide a local false discovery rate estimate. These bear the same relationship to the false discovery rate that a density, for the relevant distribution, bears to the corresponding p-value, i.e., to an area in the tail or tails of the distribution. The function `locfdr::locfdr()` is designed to provide, as well as estimates of local FDRs, an estimate of the proportion of p-values that correspond to cases where the null hypothesis is true. Estimates of the proportion of nulls may vary widely, depending on the method used.

As noted, the false discovery rate estimates are not p-values in the conventional sense. They give a frequency based probability that a detected difference is a real difference – information that p-values do not provide. Why insist on working with p-values when there is a better alternative? When m is large, the estimate provided has high statistical accuracy.

The `p.adjust()` FDR estimate remains valid in a wide range of contexts where p-values are positively correlated. Also available is `method="BY"`, designed for contexts where there may be quite general dependence structures. This gives a very conservative adjustment. Other available adjustment methods (see `?p.adjust`) are more in the style of p-values. .

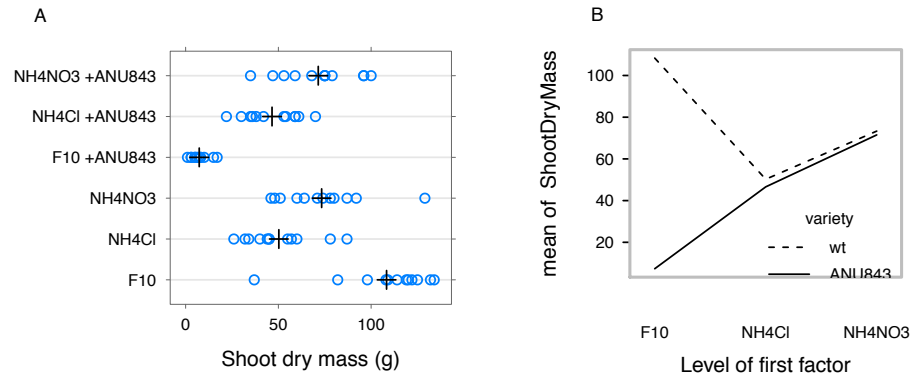

Figure 2.23: Both panels are for rice shoot dry mass data. Panel A shows a one-way strip plot, with different strips for different treatment regimes. Treatment means are shown with a large +. The interaction plot in Panel B shows how the effect of fertilizer (the first factor) changes with variety (the second factor). Data relate to Perrine et al. (2001)

As noted, 3072 p-values is small, by the standards of much other expression array data. The experimental data that will be considered in Section 9.3, from an experiment with RNA-Seq data, yielded 18658 p-values for each comparison of interest.

2.8.4 Data with a two-way structure, i.e., two factors

Consider now data from an experiment that compared wild type (`wt`) and genetically modified rice plants (`ANU843`), each with three different chemical treatments. A first factor relates to whether `F10` or `NH4Cl` or `NH4NO3` is applied. A second factor relates to whether the plant is wild type (`wt`) or `ANU843`.

There are 72 sets of results, i.e., two types (`variety`) × three chemical treatments (`fert`) × 6 replicates, with the experimental setup repeated across each of two blocks (`Block`). Figures 2.23A and 2.23B show alternative perspectives on these data.

Figure 2.23B shows a large difference between `ANU843` and wild type (`wt`) for the `F10` treatment. For the other treatments, there is no detectable difference. A two-way analysis will show a large interaction.[13]

Note, finally, that the treatments were arranged in two blocks. In general, this has implications for the analysis. This example will be discussed again in Chapter 4, where block effects will be taken into account.

2.8.5 Presentation issues

The discussion so far has treated all comparisons as of equal interest. Often they are not. There are several possibilities:

[13]
```
## Simplified version of code, Panel B only
with(rice, interaction.plot(fert, variety, ShootDryMass,
                            xlab="Level of first factor"))
```