RNA Workshop Monday Afternoon Practical

This series of exercises gives an opportunity to get hands-on experience identifying potential QC issues in differential expression projects. Most of these data sets are simulated and are simpler than real-world examples. The QC exploration will be conducted in R . I am not sure how many it will be possible to get through in this session and it is not required to complete all of them. Feel free to attempt as many or as few as you wish.

Software Required

Within R, the following packages need to be installed:

- ggplot2
- reshape2

If these have not already been installed, then as long as you have an internet connection they can be installed using the commands:

```
> install.packages("ggplot2")
```

```
> install.packages("reshape2")
```

I have also provided a script that creates pdf documents with a few simple QC plots. This will be the primary means of exploring these data sets. To load the script into your R session, use the command:

```
>
source( "http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Data_Analysis/Monday_pm/simple_RNA_QC.R" )
```

Tasks

For each example, create the simple QC plots and consider the following:

- What potential problems can you identify from the plots?
- What cold have caused those problems?
- What additional information regarding the project would you need to confirm those causes?
- What would you do to account for any QC issues when analysing the data?

Example 1

This example uses a cancer cell line data set that will be revisited in tomorrow's practical session. We'll go through this example in depth.

The data comes in two parts:

- The count table as generated by the pipeline used for this data.
- A table mapping the sample names to their experimental group.

In this set, there are three different cancer cell lines.

First, load the data into R:

```
> example1 <-
read.table( "http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Data_Analysis/Monday_pm/practical/Cancer_gene_expression_dataset.txt", header=T,
row.names=1, sep="\t" )
```

```
> example1.info <-
read.table( "http://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Data_Analysis/Monday_pm/practical/Cancer_sample.info.txt", header=T, sep="\t" )
```

You can inspect the data in R if you wish.

Now you can run the script that generates the plots:

```
> simple.RNA.QC( example1, example1.info, “Example1” )
```

This will create a pdf document called “Example1.pdf” you should be able to open and inspect. If you are working in R Studio and prefer to look at the plots from there, omit the last argument:

```
>  simple.RNA.QC( example1, example1.info )
```

In this example, the count table does not include any special categories for reads so there are no values for 'Unmapped' in the bar chart. There is also quite a lot of variability in the number of reads per sample, but this is not unexpected in cancer cell line data.

The heatmap shows that one sample, 'A7.A3J0', is more distant than the others in general, and might be considered an outlier, but given the inherent variability in cancer cell lines it might be a genuine difference rather than a technical failure.

The two PCA plots reiterate this pattern, with A7.A3J0 being isolated in the plot. The second plot also shows that there seem to be some clustering based on group, although it is difficult to pick out due to this potential outlier dominating the PCA plot.

Now, let's remove the outlier and recreate the plots:

```
> example1a <- example1[,-2]
```

```
> example1a.info <- example1.info[-2,]
```

```
> simple.RNA.QC( example1a, example1a.info, “Example1a” )
```

Revisiting the plots with the potential outlier excluded, there is a much more noticeable pattern in the PCA. Although some samples are more distant than others, they the different cell lines seem to form lines in the plot which is the sort of pattern expected from cancer cell line data.

Whether or not to include the possible outlier in any analysis is something of an open question. In non-cancer projects, it would be a clear candidate for exclusion based on how different it is, but here it is less certain. One possibility is to perform two analyses, one with the sample and one without, and see if there any significant differences in the gene lists produced.

Example 2

This dataset is for a simulated gene knockout experiment with 2 groups.

- Example2_dataset.txt
- Example2_sample_info.txt

Use Example 1 as the model to work from and check the QC plots for this data. The other examples will all use the same naming format as this one.

Example 3

Three groups, two with distinct treatments and one untreated.

Example 4

Two groups, depleted and undepleted.

Example 5

Two groups, normoxia and hypoxia.

Example 6

Two groups, wildtype and knockout.

Example 7

A two-factor model: low phosphate versus high phosphate, in untreated and treated samples, for four groups in total.

Example 8

A two-factor model: treated versus untreated, in wildtype and knockout samples, for four groups in total.

Example 9

Two types of treatment, with samples taken at 3 different time points, for 6 groups in total.

Example 10

Two groups, treated and untreated.

Example 11

Two groups, treated and untreated.

Example 12

Treated and untreated samples, across three time points, for 6 groups in total.

Example 13

Three different treatment groups.

Example 14

Treated and untreated samples, across three time points, for 6 groups in total.

Example 15

Two treatment groups.