Variant analysis in RNAseq showcase

Jan Oosting

2018 august 28

Introduction

In this practical a few kinds of analysis are presented that involve variant analysis. The practical starts as an unfinished RMarkdown document that is being completed. In RMarkdown you can create reports where the description of an analysis, the R code to perform the analysis, and the results are integrated.

RStudio contains help for editing RMArkdown under the Help menu with the Markdown quick reference and with a few cheatsheets. The template script Variant Analysis with RNA seq.rmd already contains some text and some chunks with R-code. The code in the chunks contains errors, and missing pieces. It is your task to fix the errors and amend the code. Initially all code chunks have an option eval=FALSE. This is to ensure that the code is not executed and to ensure that an initial report is being made when you press the Knit button. At the end of the practical the option eval should be set to TRUE and a report should be made.

Copy the file Variant Analysis with RNA seq.rmd to your working directory and press Knit. After a few moments a report should be shown with already a lot of text.

eQTL analysis

Chunk eqtl1

In chunk eqtl1 change the option eval=FALSE to eval=TRUE. When you now knit the report, an error will appear in .checkkeys. Apparently <select a gene> is not recognized. This value is assigned to the variable geneofinterest in the first line of the chunk.

Change <select a gene> to one of the gene names from the remark at the end of the line, and press Knit again.

If you did not make a spelling mistake, a plot will appear in the report, which displays the p-values of the correlation between genotypes of SNPs on chromosome 20 and the expression level of the chosen gene. The X-axis is the position of the SNP, and Y-axis the negative log of the p-value. In red above the plot the location of the selected gene is chosen.

You could do this for all 4 genes, but while exploring the data it more convenient to run the code just in RStudio. When working on a chunk you should first execute all previous chunks. Loading of necessary packages is usually done in one of the first chunk.

Put the cursor in the qetl1 chunk, and select Run all chunks above from the Run button. Now you can run single lines of code with <Ctrl-Enter>, or select code and run it with <Ctrl-Enter>. Alternatively you can run all lines of a chunk with <Ctrl-Shift-Enter>.

Check the plot for all genes in the first line, and select the most interesting. Describe your findings below the chunk.
Chunk eqtl2

Show the expression values of the gene for all genotypes of the best ranking SNP on chromosome 20
Table ts1 was produced, and shown in chunk eqtl1. It contains the names and p-values of the best ranking SNPs.

Set eval=TRUE in the chunk options. It seems an incorrect SNP name is taken from table ts1.
Correct the code so it uses the top-ranking snp name

Chunk eqtl3

In chunk 3 the associations between SNPs and genes is calculated for a small number of genes. The 6 best ranking genes are selected. The result is a vector of strings where the names of the probes for the genes are combined with the SNP using a dot.
The following code converts this vector into a table tab with a snp column, and a score column

```r
nms = strsplit(names(top6), "\.")
gn = sapply(nms,"[",1)
sn = sapply(nms,"[",2)
tab = data.frame(snp=sn,score=as.numeric(top6))
rownames(tab) = gn
```

Insert the code into the proper place in chunk eqtl3

Chunk eqtl4

We can corroborate our findings by comparing to results that were published previously. In this case we can match the SNPs we found to a list of published results that is made available in the GGTools package. The table you created in the previous chunk is used here.

How many lines does this table strMultPop have?

Write your conclusion below the chunk. Which SNPs were common? What does this say about the SNPs that were excluded?

Allele specific expression

Chunk ase1

Data for a few SNPs is loaded into the variable a.simple. Can you find out how many SNPs and samples are in the data set?. The ASEset class has a `dim()`, `ncol()`, and `nrow()` methods, which can help.

In markdown you can include R expressions by using single back-quotes around the R expression `r
<expression>`

Under the chunk fill in the correct R expressions to show the number of SNPs and samples in the dataset. Don’t forget to remove the double quotes around the place holders

Chunk ase2

In this chunk a lot of preparation is done, at the end some samples have to be excluded. Determine a threshold for the minimum number of reads, and whether all samples should have that many reads, or at least one.
You can use `apply(countsPerSample(a.simple), 1, <yourfilterfunction>)` to create a logical vector, and use that vector to subset the dataset `a.simple<-a.simple[, <selectionvector>]`
Create the filter, and subset the dataset

**Chunk ase3**

The allele specific expression can now be visualized. In the chunk a plot for only one SNP is shown

Make plots for all 3 SNPs. Describe your findings below the chunk.

**Chunk ase4**

In chunk 3 a chi-square test was performed for each sample to test the probability that both genotypes were expressed equally. It could be argued that the binomial test is better suited to test this. In this chunk the binomial test is applied to the expression data for each SNP

Make plots for all 3 SNPs. Describe your findings below the chunk.

**Fusion visualization**

The *chimeraviz* package contains a small number of example fusions.

**Chunk fusion1**

Defuse was used to find fusions in a sample. The *defuse* data is loaded and a circos plot is created to make a genome wide plot to show which parts of the genome are connected through the fusions.

Calculate the number of fusions below the chunk. It is the length of the fusions variable

**Chunk fusion2**

Defuse has determined a fusion between the RCC1 and HENMT1 gene. The plot shows a number of reads supporting this fusion. Can you calculate the number of spanning and split reads? Which reads are plotted?

Calculate the number of split and spanning reads, describe the plot.

**Chunk fusion3**

Describe the fusion, can you say anything about the function considering the functions of the gene?

**Bonus exercises**

**Improve layout**

R Markdown has lots of ways to improve the way the report is presented. In general this starts with deciding what is shown from all the code chunks with options `warning`, `echo`, `message`, `include` and `results`. Size of the images in chunks can be controlled with the `fig.width` and `fig.height` options.

The describing text can be formatted, see for options the menu item `help|Markdown Quick reference`
Data dependency in eQTL

The dataset for the eQTL contains trios of hapmap samples. It could be argued that the genetic dependency of the children on their parents create correlations that are hard to handle with the statistical procedures. Can you remove the children from the dataset, and redo the analysis? Is there a difference between the results including and excluding the children?