

Workshop Rmd example

HBC Training Team

2020-06-02

Project details

In this example report we are using a *toy dataset* to determine the impact of age, genotype and cell type on the average gene expression in mice. This toy study has 12 mice from **2 genotypes** (KO and Wt) and **2 cell types** (typeA and typeB).

Setup

Load Libraries

knitr and ggplot2 are the 2 libraries required to run this report.

```
> ### Bioconductor and CRAN libraries used
> library(knitr)
> library(ggplot2)
```

Load data

In addition to the normalized expression data, we need to make sure we have the appropriate metadata information loaded.

```
> data <- read.csv("data/counts.rpkm.csv", header = T, row.names = 1)
> meta <- read.csv("data/mouse_exp_design.csv", header = T, row.names = 1)
> kable(meta, format = "markdown")
```

	genotype	celltype	replicate
sample1	Wt	typeA	1
sample2	Wt	typeA	2
sample3	Wt	typeA	3
sample4	KO	typeA	1
sample5	KO	typeA	2
sample6	KO	typeA	3
sample7	Wt	typeB	1
sample8	Wt	typeB	2
sample9	Wt	typeB	3
sample10	KO	typeB	1
sample11	KO	typeB	2
sample12	KO	typeB	3

Modify the metadata data frame

The original metadata file did not have the average expression for each sample. So we will use the counts data to generate that information and add it to the metadata data frame along with the age of the mice (in days).

```
##          sample1 sample2 sample3 sample4 sample5
## ENSMUSG000000000001 19.7848000 19.265000 20.889500 24.076700 23.7222000
## ENSMUSG000000000003 0.0000000 0.000000 0.000000 0.000000 0.0000000
## ENSMUSG000000000028 0.9377920 1.032290 0.892183 0.827891 0.8269540
## ENSMUSG000000000031 0.0359631 0.000000 0.000000 0.000000 0.0000000
## ENSMUSG000000000037 0.1514170 0.056033 0.146196 0.180883 0.0473238
## ENSMUSG000000000049 0.2567330 0.258134 0.421286 2.191720 1.0730200
##          sample6 sample7 sample8 sample9 sample10 sample11
## ENSMUSG000000000001 20.8198000 2.611610 5.8495400 6.5126300 24.198100 24.0465000
## ENSMUSG000000000003 0.0000000 0.000000 0.0000000 0.0000000 0.000000 0.0000000
## ENSMUSG000000000028 1.1686300 1.134410 0.6987540 0.9251170 1.045920 0.9753270
## ENSMUSG000000000031 0.0511932 0.000000 0.0298449 0.0597726 0.000000 0.0000000
## ENSMUSG000000000037 0.1438840 0.000000 0.0685938 0.0494147 0.017004 0.0206405
## ENSMUSG000000000049 1.6853800 0.252342 0.2970320 0.2082800 0.369550 0.0634322
##          sample12
## ENSMUSG000000000001 26.9158000
## ENSMUSG000000000003 0.0000000
## ENSMUSG000000000028 0.6735630
## ENSMUSG000000000031 0.0204382
## ENSMUSG000000000037 0.0662324
## ENSMUSG000000000049 0.1161970
```

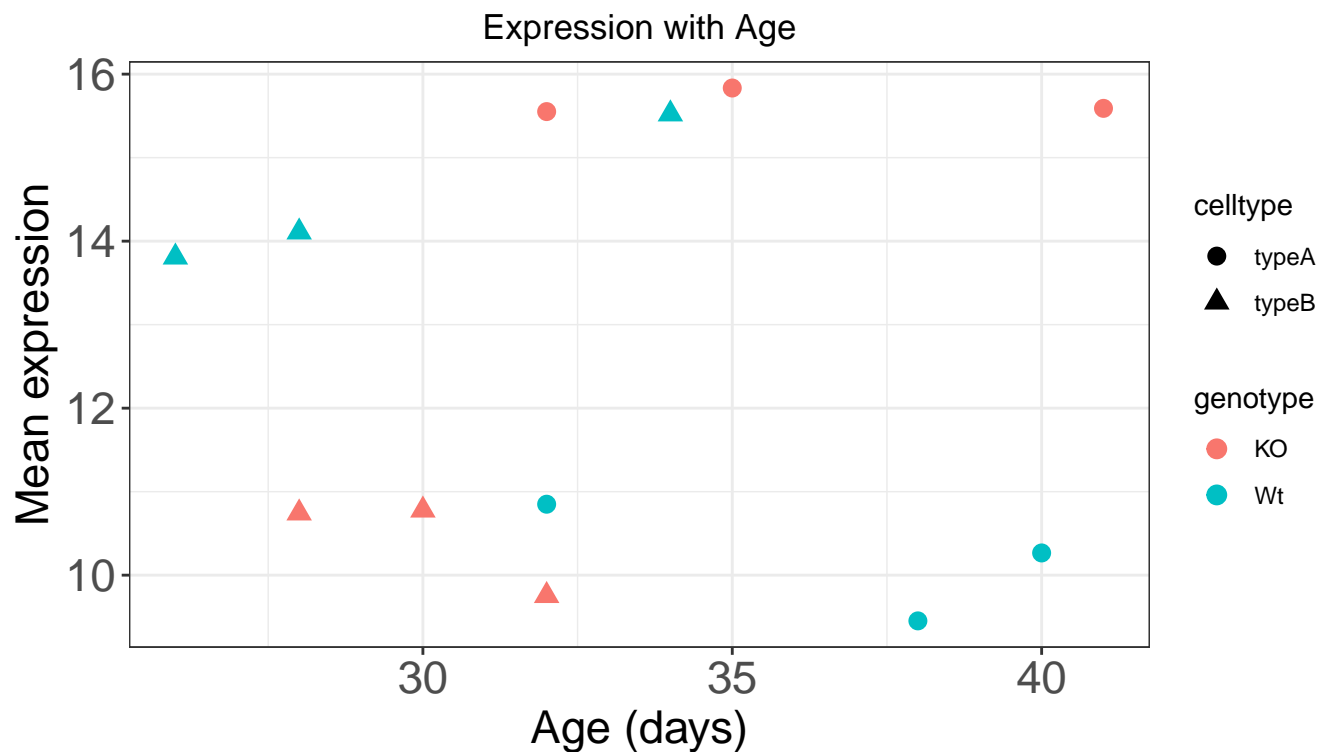
```
> # generate samplemeans for the data
> samplemeans <- apply(data_ordered, 2, mean)
>
> # Create a numeric vector with ages to add to meta (Note that there are 12
> # elements here)
> age_in_days <- c(40, 32, 38, 35, 41, 32, 34, 26, 28, 28, 30, 32)
>
> # Add samplemeans and age to the meta table
> new_meta <- cbind(meta, age_in_days, samplemeans)
>
> # print metadata dataframe
> kable(new_meta, format = "markdown")
```

	genotype	celltype	replicate	age_in_days	samplemeans
sample1	Wt	typeA	1	40	10.266102
sample2	Wt	typeA	2	32	10.849759
sample3	Wt	typeA	3	38	9.452517
sample4	KO	typeA	1	35	15.833872
sample5	KO	typeA	2	41	15.590184
sample6	KO	typeA	3	32	15.551529
sample7	Wt	typeB	1	34	15.522219
sample8	Wt	typeB	2	26	13.808281
sample9	Wt	typeB	3	28	14.108399
sample10	KO	typeB	1	28	10.743292

	genotype	celltype	replicate	age_in_days	samplemeans
sample11	KO	typeB	2	30	10.778318
sample12	KO	typeB	3	32	9.754733

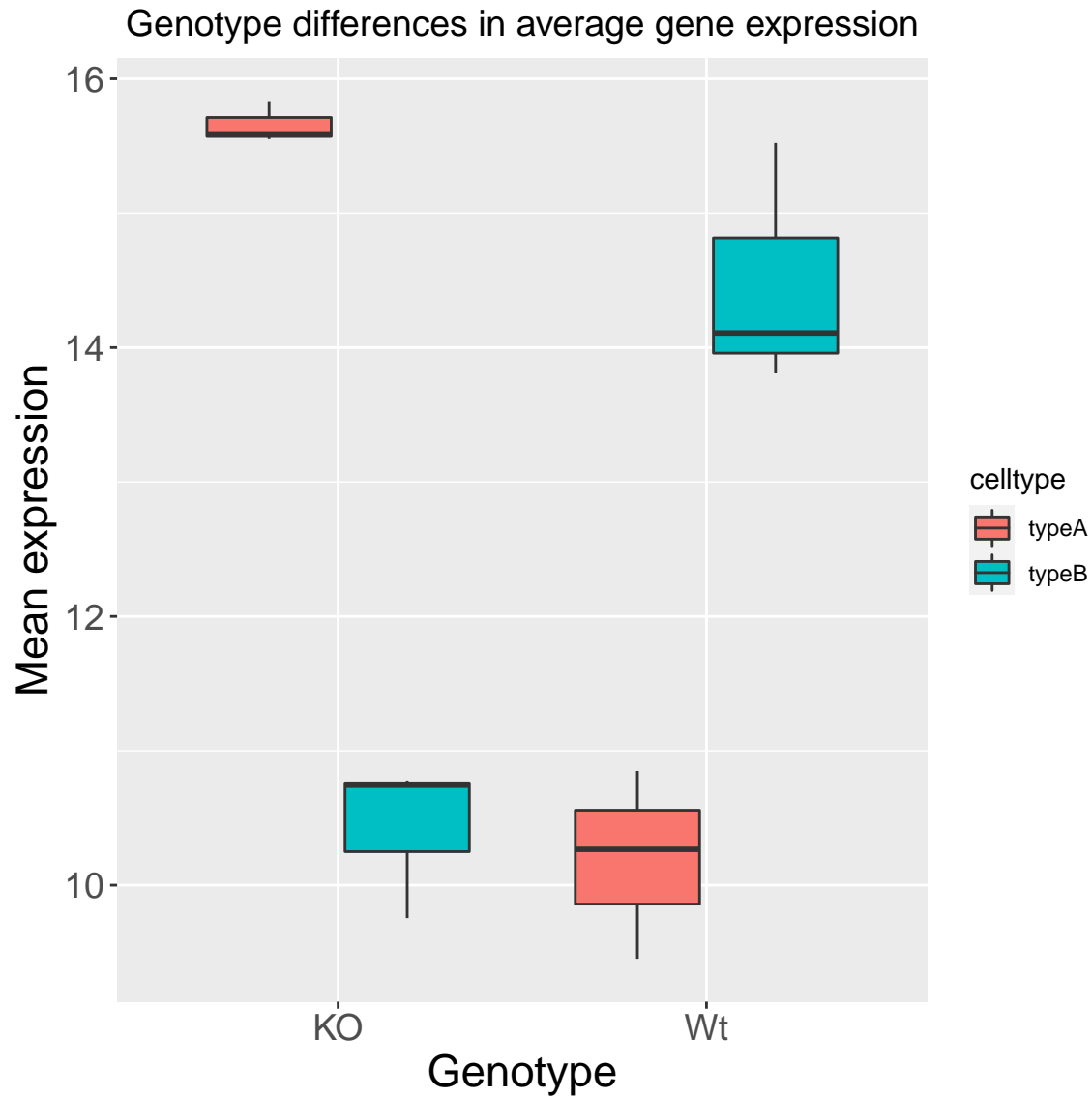
Does average expression change with age of mouse?

```
> ggplot(new_meta) + geom_point(aes(x = age_in_days, y = samplemeans, color = genotype,
+   shape = celltype), size = rel(3)) + theme_bw() + theme(axis.text = element_text(size = rel(1.5)),
+   axis.title = element_text(size = rel(1.5)), plot.title = element_text(hjust = 0.5)) +
+   xlab("Age (days)") + ylab("Mean expression") + ggtitle("Expression with Age")
```



Distribution of expression in the 2 genotypes

```
> ggplot(new_meta) + geom_boxplot(aes(x = genotype, y = samplemeans, fill = celltype)) +
+   ggtitle("Genotype differences in average gene expression") + theme(axis.text = element_text(size = rel(1.5)),
+   axis.title = element_text(size = rel(1.5)), plot.title = element_text(hjust = 0.5,
+   size = rel(1.25))) + xlab("Genotype") + ylab("Mean expression")
```



Tables

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1