R: fundamental skills for biologists

1-22 June 2022

This is an archive of questions asked via Slack during the workshop.

Question	Answer
Session 1: Spreadsheets, organising data and first steps with R	
Is it recommended that missing data is stated as NA as it is read as a non-value? What about using symbols instead, such as a dash that is used at times as well, would this cause issues?	Specifically in R - R knows how to handle NAs. Using a dash would cause a silent data mutation in R
In what cases is R better than python?	***Trainer A***: I'm a bit biased because I learnt R before I learnt a bit of python and I just find the R documentation a lot better overall for beginners than Python does. There's always this layer of assumed knowledge that I find frustrating whenever I'm trying to do something in Python. The plotting also I think is easier overall in R - python has 3 or more ecosystems for that iirc ***Trainer B***: Equally, the large R community means that there are a lot more pipelines available in R rather than python
What can I do to see the console?	You might have minimised your console window - there is a minimise/maximise option in the pane
Is there a reason you would write in the console vs the script? Or is it an individual preference?	***Trainer A:*** Use the console to try code out, script to record your code so you can re run it ***Trainer B:*** Generally, I write in the script because then you have a record of the analysis that you are doing. I write in the

	console when I have something that I want to try but don't necessarily think it'll be something that I want a record of ie, I'm quickly testing something. ***Trainer C:*** Write in the console if it's something you know is not relevant to your analysis - for example, calling help functions. Otherwise even when testing things I'd personally put them in the script (so if they do work, I can go back and reuse or edit those commands). And also when you are doing once only actions like installing packages
What does the bioc-intro.Rproj file do?	The project creates a directory that will act as a starting point for running .R scripts. The project structure also allows R to store information of each session that you run, such as the objects that you create and the history of commands that you have run
Is there a way to move these files into our own personal folders OR is there a way to save into these folders upfront from the outset?	You can export the data. Click the checkbox for the folder you want to keep, click the `More` button for the dropdown menu, then click `Export`
What makes data visualisation in R better than something like Prism? Example with your BarGraph Plot	***Trainer A:*** It is much much more customizable ***Trainer B***: I have not used Prism, but R is very flexible. Some packages in R can produce similar graphs in Prism, with but more complex graphs R can do a lot better ***Trainer C:*** Will be able to reproduce much more easily/send to someone else who will be able to generate the same plot if given the data :) ***Trainer D:*** huge variety and flexibility of plotting types. check out https://r-graph-gallery.com/ (includes example code for each) The R Graph Gallery The R Graph Gallery – Help and inspiration for R charts The R graph gallery displays hundreds of charts made with R, always providing the reproducible code. ***Trainer E:*** It's easier to wrangle data into the format you need to plot in R e.g. to plot different subsets of the data

	Trainer F: Also with the graphs generated in R you can then use the sister programs such as R shiny - where you can create interactive plots
I cannot see an output for any of these simple commands; eg 3 + 5 etc	You need to select 'Run' in the right-hand corner otherwise your code will not run into the console Or within the script, place your cursor at the line and then hit ctrl+enter / command + enter
Is there a point in writing "x=", given excluding it spits out the same result? e.g. round(digits=4, 3.1415926)	It's for making reading code easier but not strictly necessary. As long as you specify what the other arguments are, you won't have to specify 'x'
Can you use other script programs to write code e.g. VSCode then paste this into RStudio?	Yes, you can write your code in another program (or if it is already written in another format) and copy it over. Provided the language is R that you are copying over. Just note you would need to install the R packages you're using in both VStudio and RStudio if you wanted to run the code in both https://code.visualstudio.com/docs/languages/r
Can you add a numeric value to a character vector and visa versa?	You can but it will change the data type - this is called coercion
How can we omit a number?	`weight_g <- weight_g[-1]` will delete the first entry
What does double vector mean?	A 'double precision floating point number'. https://en.wikipedia.org/wiki/Double-precision_floating-point_format. Basically a computer term for a value with some number of decimal places
Rule of thumb: round brackets - indicate a function and will need to have arguments within the brackets. e.g., `round(5.56831, digits=2)`	

square brackets - are for sub-setting objects (I like to think of orientating within box or drawer); where you specify the location of the element you're searching for e.g., `molecules[2]`	
Session 2: Manipulating and analysing data with dplyr	
When I have a data spreadsheet with colour and italic or bold text, will it confuse R or will the computer just ignore these characteristics? Or in other words: Can I structure my table for optical reasons for me to see things more easily (not store data in colour or format) or does that cause problems in R?	In general, it's just ignored. If you save as a .csv, the colour information is stripped out. If it's saved as an .xslx file, the colour is there, but note that you need to use a special package to read these (e.g. readxl).
Tab complete is not prompting automatically.	That normally indicates that you are not in the right directory Check/set your working directory Might not be the issue but: you can only tab-autocomplete paths when they are within quote marks. e.g. r`na <- read.csv("course-data/<hit tab>` You can also try just typing `read_csv("")`. Then with your cursor in between the quotation marks, hit tab. It should create a drop down menu of what is available
Can you count unique genes by the column header name rather than column number?	Absolutely! `length(unique(rna[,c("gene")]))` This is called subsetting with column names
What is actually displayed if we only did `rna[1]` instead of `rna[,1]`?	`rna[1]` will show the column 1 in data.frame format whereas `rna[,1]` will show the contents of the column 1 as a vector
Is it always rows first and columns second in factors and the other way round in matrix?	In the `matrix()` function it doesn't matter if you give `ncol=` first or `nrow=` as you've named the arguments
Can you subset a list the same way you do a data frame?	First you have to specify the 'element' you want to subset. e.g. to get the 4th row of the

	cars element (which is the 4th in the list), you would use `l[[4]][4 , ]` or to get the first two elements, it's a more familiar subset command: `l[c(1,2)]`
Can we save lists as xlsx file?	***Trainer A***: Thought it is always possible to combine all elements of a list in a dataframe with the `bind_rows()` function and export that as a xlsx file I recommend it only for specific reasons though, because the structure of your initial list will completely change ***Trainer B***: I find it useful sometimes to save a list of data.frames into a multi-tabbed excel file, but you need to use a special package (maybe xlsx??) to do so. Usually .tab or .tsv is simpler
Is filtering case sensitive? male (didn't work) vs Male	It is case sensitive when it comes to the filtering, but when I compare `select(rna, starts_with("T"))` and `select(rna, starts_with("t"))`, I get the same output The select helpers are not case sensitive by default, you change that by adding `ignore.case=FALSE` https://tidyselect.r-lib.org/reference/starts_with.html
So mutate always adds the new column to the end of the data you start with? sounds dangerous. What if you (by mistake) run the same mutate command again? (error? extra columns? overwrite?)	***Trainer A***: At this point of the exercise we haven't change our initial object yet **Trainer B**: Mutate will change an existing column if you ask it to, but it only does it on the copy of the data you're working on. **Trainer C:** Also good with the 'pipes' you can see in one place all the changes you have made and you will not make a new object until you assign it again... **Follow up question:** So if I want to update the data object, I have to pipe it back to that object name?

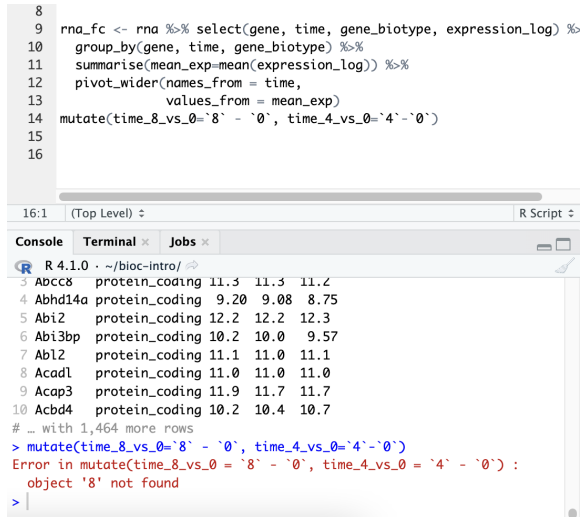
	Trainer D: Yes and it would not change your underlying raw data, so you can always start over ***Trainer E:*** If you run the same mutate command again (and you've saved the output as an object) it will replace the column as it has the same name ***Follow up question:*** so if I just leave a pipe hanging, that version of the data just disappears? it's not held as an object? ***Trainers B and C***: Yes ! Just displays to screen **Trainer D**: it is like running a command without assigning it to an object
Session 3: Data visualisation	

I got this error.
 	
`> ggplot(data=rna)`
`Error in .Call.graphics(C_palette2, .Call(C_palette2, NULL)) :`
`  invalid graphics state`
	Try restarting RStudio

Can you easily modify upper and lower boundaries (i.e. 0- 50 000 instead of 100 000 on x axis in this example)?	You can also use `scale_x_continuous(limits = c(0, 50000))`

What does this message mean?
 	
`> rna_fc <- rna %>% select(gene, time, gene_biotype, expression_log) %>%`
`+  group_by(gene, time, gene_biotype) %>%`
`+ summarize(mean_exp=mean(expression_log))`
`` `summarise()` has grouped output by 'gene', 'time'. You can override using the `.groups` argument. ``
	This is a standard warning from using `group_by` (it's about different ways of ungrouping data later)
	
	When we use group_by() , the output will retain the same grouping structure. This can be avoided if we use `%>% ungroup()` at the end, OR, use `group_by(gene, time) %>% summarize(mean_exp = mean(expression), .groups = 'drop')`

I have this error message coming up:	You're missing the pipe after pivot_wider

<table>
<tr>
<td> 
</td>
<td></td>
</tr>
<tr>
<td>Can we input hexadecimals for the colours?</td>
<td>Yes, see here http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/#hexadecimal-color-code-chart

There's also a table of standard R colours with names (like 'blue' : http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf</td>
</tr>
<tr>
<td>What is the alpha?</td>
<td>Transparency: 1 is opaque, 0 fully transparent

https://ggplot2.tidyverse.org/reference/aes_colour_fill_alpha.html#:~:text=Alpha%20refers%20to%20the%20opacity,corresponding%20to%20more%20transparent%20colors</td>
</tr>
<tr>
<td>How do you get the sample names to actually align with each boxplot or column for that matter?</td>
<td>Try adding:
<code>theme(axis.text.x = element_text(angle = 90, hjust = 0.5, vjust = 0.5))</code>

e.g. usually googling 'align axis labels ggplot' can find something (I always forget this)
https://stackoverflow.com/questions/37488075/align-axis-label-on-the-right-with-ggplot2</td>
</tr>
<tr>
<td>I got an error

<code>> ggplot(data=mean_exp_by_time,
+ mapping=aes(x=time, y=mean_exp, group=gene)) + geom_line()
Error in ggplot(data = mean_exp_by_time, mapping = aes(x = time, y = mean_exp, :
object 'mean_exp_by_time' not found</code></td>
<td>This is usually due to a mismatch between the object name in your code and what you have actually called the object. Check what you have called it in your environment in RStudio and change your code as needed.</td>
</tr>
</table>

Does it matter how you spell colour? color?	No, it doesn't matter. ggplot is US/UK English agnostic and you can even mix and match.
Why when I write `facet_wrap(~gene)`, it does not show any changes in the plot area?	You need to include the + symbol on the line before this code
Why this time around the tilde is after sex and previously it was before gene?	It's related to the way that the plot displays a variable. See also: https://benwhalley.github.io/just-enough-r/ggplot-details.html
I am too deep in the woods. Why do I keep getting this message " `geom_path: Each group consists of only one observation. Do you need to adjust the group aesthetic?`	You might be missing a group or colour call
Session 4: SummarizedExperiment and Bioconductor	

For SummarizedExperiment:
 	
I got this error
 	
`> se <- readRDS("course-data/data/GSE96870/GSE96870/se2.rds")`
`Error in gzfile(file, "rb") : cannot open the connection`
`In addition: Warning message:`
`In gzfile(file, "rb") :`
`  cannot open compressed file 'course-data/data/GSE96870/GSE96870/se2.rds', probable reason 'No such file or directory'`
	Make sure you are implementing your command from the right level of the directory. Check your directory using `getwd()`
	It might help running the full pathway (changing user to match your directory structure):
	`se <- readRDS("/home/user/Bioc_Intro/course-data/data/GSE96870/se2.rds")`

I'm getting this error:
 	
`se_created <- SummarizedExperiment(assays = SimpleList(counts=count_matrix)`
`,`
 	
`colData = sample_metadata,`
 	
`rowRanges = gene_metadata)`
	try rowData instead of rowRanges
	
	for the `SummarizedExperiment()` `rowRanges=` you need to supply a ranges (Granges/GenomicRanges) object. If you made one with `gene_metadata <- rowRanges(se)` and gave that to `rowRanges=` that would work and that's what I did in the material. But in the demo I accidentally did `gene_metadata <-`

`Error in validObject(.Object) : invalid class "RangedSummarizedExperiment" object: 1: invalid object for slot "rowRanges" in class "RangedSummarizedExperiment": got class "DFrame", should be or extend class "GenomicRanges_OR_GRangesList" invalid class "RangedSummarizedExperiment" object: 2: 'x@assays' is not parallel to 'x'`	`rowData(se)` which has no ranges 🤫 and it's a dataframe, and that you give to `SummarizedExperiment() rowData=`
Once tidySummarizedExperiment is loaded, is the dataset automatically viewed as a tibble?	Yes. You can change whether you want this functionality by changing options `("restore_SummarizedExperiment_show" = TRUE)`

I can't find .sample
 	
`> se %>% group_by(.samples) %>%`
`+ summarise(total_counts = sum(counts))`
`tidySummarizedExperiment says: A data frame is returned for independent data analysis.`
`Error in `dplyr::group_by()`:`
`! Must group by variables found in `.data`.`
`✖ Column `.samples` is not found.`
	You made a typo. Instead of `.samples` try `.sample`

What about proteomics annotations? I could not find any in bioconductor	What about this? https://www.bioconductor.org/packages/release/data/experiment/vignettes/RforProteomics/inst/doc/RforProteomics.html#7_Annotation I just googled proteomics annotation bioconductor 🙂 but looking at the url I think it's not a package it's experiment data so in that section of bioconductor, see here https://www.bioconductor.org/packages/release/data/experiment/