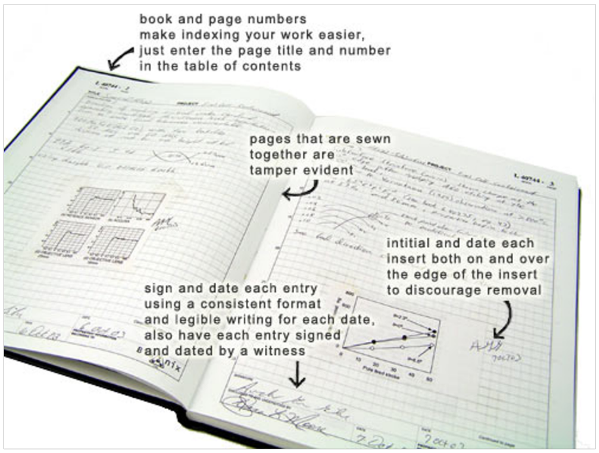
R Workshop Session 3 Exercise

Exercise 1:

I am sure you will agree that there is a big difference in how you record your steps in a lab notebook and the methods section of a manuscript. In your lab notebook you will documents the incremental progress, the many failures, new ideas in them and get them countersigned by your supervisor etc. However, in a manuscript, you will present a concise summary that can be followed and reproduced by the reader.


book and page numbers make indexing your work easier, just enter the page title and number in the table of contents
pages that are sewn together are tamper evident
intitial and date each insert both on and over the edge of the insert to discourage removal
sign and date each entry using a consistent format and legible writing for each date, also have each entry signed and dated by a witness

March 2018 | volume 15 | number 3
nature methods
www.nature.com/naturemethods
Techniques for life scientists and chemists
Direct RNA sequencing in nanopores
Highly multiplexed Raman imaging
Proteomic phosphohistidine probing
Capturing the RNA interactome
Perspective on STED nanoscopy

Scientific lab notebook (messy, ideas, incremental)

Publication (polished, concise, reproducible)

Coding is no different. We start with a messy incremental codes like we had in the Session 3 classroom. The codes get modified as we discover new things to fix in our exploration stage. We need to cleanup the codes to reduce mistakes amd to make them readable to others (especially yourself in the future). Good coding practice starts now!

I will show you how to do this for the expression data which goes from this:

```r
fn   <- "data_tcga/brca_RNA_Seq_v2_expression_median_sel.txt"
expr <- read_tsv(fn)
expr <- expr %>% column_to_rownames("Hugo_Symbol")
dim(expr)

expr[ , 1:3]

expr <- t(expr)

dim(expr)

head(expr, 3) %>% round(1)


expr <- log2( expr + 1 )
head(expr, 3) %>% round(1)
```

to this:

```
#################################
## Cleanup the expression data ##
#################################

fn <- "data_tcga/brca_RNA_Seq_v2_expression_median_sel.txt"

expr <- read_tsv(fn) %>%
  column_to_rownames("Hugo_Symbol") %>%
  t()

expr <- log2( expr + 1 )
```

Your task is to make your codes for cleaning the clinical data from TCGA Breast Cancer more concise and readable.

Exercise 2:

FASTQ files contain the base calls and the quality score (i.e. predicted error rate in base calling) in an encoded format[1]. For this exercise, it suffices to understand the following table[2].

Quality Score	Error Probability
Q40	0.0001
Q30	0.001
Q20	0.01

What is the probability of observing 3 or more incorrect bases in

1) a 150-base pair read length assuming all bases have Q30 score?
2) a 75-base pair read length assuming all bases have Q30 score?
3) a 75-base pair read length assuming all bases have Q40 score?

(Assume the errors in the bases are independent of each other).

[1] https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

[2] https://www.illumina.com/documents/products/technotes/technote_understanding_quality_scores.pdf