Integrative Analysis of Omics Data with Biological Knowledge in Translational Medicine

Ferran Briansó

Facultat de Biologia
Departament de Genètica, Microbiologia i Estadística

Universitat de Barcelona

A thesis submitted for the degree of

Doctor of Philosophy

XXXX XX 2023

For XXXXX XXXXXX

Acknowledgements

This is where you will normally thank your advisor, colleagues, family and friends, as well as funding and institutional support. In our case, we will give our praises to the people who developed the ideas and tools that allow us to push open science a little step forward by writing plain-text, transparent, and reproducible theses in R Markdown.

We must be grateful to John Gruber for inventing the original version of Markdown, to John MacFarlane for creating Pandoc (`http://pandoc.org`) which converts Markdown to a large number of output formats, and to Yihui Xie for creating `knitr` which introduced R Markdown as a way of embedding code in Markdown documents, and `bookdown` which added tools for technical and longer-form writing.

Special thanks to Chester Ismay, who created the `thesisdown` package that helped many a PhD student write their theses in R Markdown. And a very special thanks to John McManigle, whose adaption of Sam Evans' adaptation of Keith Gillow's original maths template for writing an Oxford University DPhil thesis in LaTeX provided the template that I in turn adapted for R Markdown.

Finally, profuse thanks to JJ Allaire, the founder and CEO of RStudio, and Hadley Wickham, the mastermind of the tidyverse without whom we'd all just given up and done data science in Python instead. Thanks for making data science easier, more accessible, and more fun for us all.

Ferran Brianso
Mataro, BCN
XX XXXXXX 2023

Abstract

The general concept of Data Integration can be defined as the combination of data residing in different sources in order to provide the users with a unified view of these data [1]. However, the practical meaning of the term Integration may vary from, for instance, the computational combination of data, to the combination of studies performed independently, the simultaneous analysis of multiple variables on multiple datasets, or any possible approach for homogeneously querying heterogeneous data sources. Therefore, in many cases, an integrative analysis may be preferable than a simple combination of data from distinct sources. Integrative analysis allows not only for the combination of heterogeneous data, but also for the combined use of these data in order to get the most relevant information and, what is better, to be able to extract some information that could not be unveiled by the separated analysis of each of the original data types.

Over the past decade, advancements in omics technologies have facilitated the high-throughput monitoring of molecular and organism processes. These techniques have been widely applied to identify biological agents and to characterize biochemical systems, often focusing on the discovery of therapeutic targets and biomarkers related with specific diseases [2,3,4]. While many single-omic approaches target comprehensive analysis of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) among other, there is still field to improve omics data analyses through integrative methods [5,6]. In this sense, the integrative point of view defined in the paragraph above, applied to multi-omics data, is a promising approach to achieve better biomarker development in biomedical research projects, and this is the core idea of this work.

As the field of omics has evolved from analyzing a unique type of data to multiple types, it has been natural to extend the previous use of multivariate techniques to this new situation. With this aim classical and new multivariate techniques have been applied to the analysis of multi-omics datasets. Many of these techniques are dimension reduction methods that aim at finding main sources of variability in the data while maximizing some information characteristic such as the variance of each dataset, the correlation between groups of variables or other. Examples of such techniques are well consolidated methods such as Principal

Component Analysis (PCA), Singular Value Decomposition (SVD), Correspondence Analysis (CA), and Partial Least Squares (PLS). Besides these more "novel" approaches have been used such as: Principal Components Regression, Coinertia and Multiple Coinertia Analysis, Generalized SVD, Sparse PLS, Multiple Factor Analysis (MFA), or combined versions of them [7,8,9]. Meng [10], Cavill [11], Wu [12], Subramanian [30], Krassowski [31], and Cantini [32], are good reviews of the state of the art of using multivariate and joint reduction methods for Integrative Multi-Omics Analysis.

Dimension reduction methods, especially those that are able to deal with situations that are typical from the omics context (with many more variables than samples, or possibly sparse matrices with many missing values), have been of great help in visualizing datasets or even for performing variable selection to find biomarkers for a given situation [12]. There is however one point where they underperform other approaches, that is, the difficulty in interpreting results from a biological point of view. This is relatively reasonable, because the most of these methods work by creating new variables that are some type of linear combination from the original ones. While this is useful, for example, for removing redundancy, this does not provide any clues on what these new dimensions may mean from a biological point of view.

This problem has been known since the beginning of using multivariate methods with omics data, but only a few approaches have been taken to deal with this. The first attempts to introduce biological information in the analyses consisted of using the most well-known database of biological functions, the Gene Ontology (GO) [13]. Fellenberg [14] introduces a way to integrate Gene Ontology information with Correspondence Analysis to facilitate the interpretation of microarray data. De Tayrac et al. [15] applies multiple factor analysis to the integrative analysis of microarray and DNA copy number data. They apply GO Terms on data visualizations by treating these terms as supplemental information. In recent years the representation of biological knowledge has shifted from Gene Ontology to using Gene Sets [16]. Meng and Culhane [10] have introduced the Integrative Clustering with Gene Set Analysis where gene set expression analysis is performed based on multiple omics data; and Tyekucheva et al. [17], go one step further and use the results of Gene Set Expression Analysis (GSEA) to integrate different omics data.

Altogether, the previous approaches show several things: Although the idea that integrating quantitative data with biological knowledge may increase interpretability, the number of successful attempts to do this is still small. In this thesis, the use of either classical GO Terms or more flexible annotations (Gene Sets or custom annotations), will be combined with different approaches, and combinations of them if needed, to guide integrative analysis and to improve its biological interpretability from the point of view of the biomedical researchers.

Contents

List of Figures

List of Tables

List of Abbreviations

1-D, 2-D . . . One- or two-dimensional, referring **in this thesis** to spatial dimensions in an image.

Otter . . . . . One of the finest of water mammals.

Hedgehog . . . Quite a nice prickly friend.

1 Introduction

Contents

1.1 Content of the introductory text (WIP)

The general concept of Data Integration can be defined as the combination of data residing in different sources in order to provide the users with a unified view of these data [1]. However, the practical meaning of the term Integration may vary from, for instance, the computational combination of data, to the combination of studies performed independently, the simultaneous analysis of multiple variables on multiple datasets, or any possible approach for homogeneously querying heterogeneous data sources. Therefore, in many cases, an integrative analysis may be preferable than a simple combination of data from distinct sources. Integrative

analysis allows not only for the combination of heterogeneous data, but also for the combined use of these data in order to get the most relevant information and, what is better, to be able to extract some information that could not be unveiled by the separated analysis of each of the original data types.

Over the past decade, advancements in omics technologies have facilitated the high-throughput monitoring of molecular and organism processes. These techniques have been widely applied to identify biological agents and to characterize biochemical systems, often focusing on the discovery of therapeutic targets and biomarkers related with specific diseases [2,3,4]. While many single-omic approaches target comprehensive analysis of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) among other, there is still field to improve omics data analyses through integrative methods [5,6]. In this sense, the integrative point of view defined in the paragraph above, applied to multi-omics data, is a promising approach to achieve better biomarker development in biomedical research projects, and this is the core idea of this work.

As the field of omics has evolved from analyzing a unique type of data to multiple types, it has been natural to extend the previous use of multivariate techniques to this new situation. With this aim classical and new multivariate techniques have been applied to the analysis of multi-omics datasets. Many of these techniques are dimension reduction methods that aim at finding main sources of variability in the data while maximizing some information characteristic such as the variance of each dataset, the correlation between groups of variables or other. Examples of such techniques are well consolidated methods such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Correspondence Analysis (CA), and Partial Least Squares (PLS). Besides these more "novel" approaches have been used such as: Principal Components Regression, Coinertia and Multiple Coinertia Analysis, Generalized SVD, Sparse PLS, Multiple Factor Analysis (MFA), or combined versions of them [7,8,9]. Meng [10], Cavill [11], Wu [12], Subramanian [30], Krassowski [31], and Cantini [32], are good reviews of the

state of the art of using multivariate and joint reduction methods for Integrative Multi-Omics Analysis.

Dimension reduction methods, especially those that are able to deal with situations that are typical from the omics context (with many more variables than samples, or possibly sparse matrices with many missing values), have been of great help in visualizing datasets or even for performing variable selection to find biomarkers for a given situation [12]. There is however one point where they underperform other approaches, that is, the difficulty in interpreting results from a biological point of view. This is relatively reasonable, because the most of these methods work by creating new variables that are some type of linear combination from the original ones. While this is useful, for example, for removing redundancy, this does not provide any clues on what these new dimensions may mean from a biological point of view.

This problem has been known since the beginning of using multivariate methods with omics data, but only a few approaches have been taken to deal with this. The first attempts to introduce biological information in the analyses consisted of using the most well-known database of biological functions, the Gene Ontology (GO) [13]. Fellenberg [14] introduces a way to integrate Gene Ontology information with Correspondence Analysis to facilitate the interpretation of microarray data. De Tayrac et al. [15] applies multiple factor analysis to the integrative analysis of microarray and DNA copy number data. They apply GO Terms on data visualizations by treating these terms as supplemental information. In recent years the representation of biological knowledge has shifted from Gene Ontology to using Gene Sets [16]. Meng and Culhane [10] have introduced the Integrative Clustering with Gene Set Analysis where gene set expression analysis is performed based on multiple omics data; and Tyekucheva et al. [17], go one step further and use the results of Gene Set Expression Analysis (GSEA) to integrate different omics data.

Altogether, the previous approaches show several things: Although the idea that integrating quantitative data with biological knowledge may increase interpretability, the number of successful attempts to do this is still small. In this thesis, the

use of either classical GO Terms or more flexible annotations (Gene Sets or custom annotations), will be combined with different approaches, and combinations of them if needed, to guide integrative analysis and to improve its biological interpretability from the point of view of the biomedical researchers.

1.2 Background/State of the Art

Falta desenvolupar punts

1.2.1 Omics data analyses

3 problemes esencials (veure projecte recerca Alex):

- Omics data may be partly incomplete, especially in multiomics studies, where not all types of data are usually available for all individuals.
- The results of these analyses are difficult to interpret. If we agree that the ultimate goal of many analyzes is a better understanding of the underlying biological processes, for example, in a disease study context, it should be possible to establish a clear relationship between the outcome of an analysis and what this means biologically. And this is not always so.
- These kind of data analytics are difficult to standardize, as it is not easy to make complex pipelines of multi-omics analyses, which integrate multiple processes with multiple sources, easy to reproduce or communicate.

Més el problema de la p»n (Dimensionallity Reduction Techniques; The p»n situation)

1.2.2 Integrative analyses

Allows the combination of distinct omics data.

The blind men and the elephant https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

Interpretability is a weak point of most multi omics approaches.

Methods focus much more on feature selection discovery and interaction highlighting measurement than on clinical or biological interpretability.

1.2.3 Review of existing approaches for multi-omics data integration

MCIA, RGCCA, MFA... Cavill, 2016; Culhane 2003...

1.3 Content from template

Welcome to `oxforddown` (Lyngs, 2019), a thesis template for R Markdown that I created when writing my own PhD thesis at the University of Oxford. This template allows you to write in R Markdown, while formatting the PDF output with the beautiful and time-tested OxThesis LaTeX template. The sample content is partly adapted from `thesisdown` .

Hopefully, writing your thesis in R Markdown will provide a nicer interface to the OxThesis template if you haven't used TeX or LaTeX before. More importantly, *R Markdown* allows you to embed chunks of code within your thesis and generate plots and tables directly from the underlying data, avoiding copy-paste steps. This gets you into the habit of doing reproducible research, which will benefit you long-term as a researcher, and also help anyone that is trying to reproduce or build upon your results down the road.

1.4 Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* allows you to read in your data, analyze it and to visualize it using **R**, **Python** or other languages, and provide documentation

and commentary on the results of your project.

Further, it allows for results of code output to be passed inline to the commentary of your results. You'll see more on this later, focusing on **R**. If you are more into **Python** or something else, you can still use *R Markdown* - see 'Other language engines' in Yihui Xie's *R Markdown: The Definitive Guide*.

Using LaTeX together with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may never have had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about reproducibility in research can benefit from using *R Markdown*. If you are working in 'softer' fields, the user-friendly nature of the *Markdown* syntax and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should still make it of great benefit to your thesis project.

2
Objectives

The main objectives of this work are the following:

1. To make an empirical comparison of some of the currently available dimension reduction techniques applied for the integration of omics data, focused on their ability to include biological annotations,

2. To develop methods and workflows able to apply these techniques, focusing on the matching of distinct omics datasets relying on biological knowledge,

3. To apply these methods to specific translational biomedical research cases, such as an integrative analysis of transcriptomics and proteomics data to study ischemic stroke, as well as to public datasets, which can be easily shared and are not as restricted by sample sizes as other projects.

4. To implement the knowledge acquired with this work into the appropriate bioinformatics tools, e.g. R packages or web-based tools, that will be used in future biomedical research projects for providing a better interpretation of this kind of studies.

All these objectives are in agreement with the tasks defined within a project partially supported by Grant MTM2015-64465-C2-1-R (MINECO/FEDER) from

the Ministerio de Economía y Competitividad (Spain), to which the PhD Thesis proposed here is related.

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit...

There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain...

— Cicero's *de Finibus Bonorum et Malorum.*

3 Methodology

Contents

3.1 Working phases

Working phases, with the corresponding steps, followed in order to achieve the above objectives:

1. Application of integrative multi-omics methods to (I) the analysis of specific data sets provided by research units from our former affiliation center, VHIR, and other research institutions that we collaborate with [34, 36, 37] and (II) to the integrative analysis of larger data sets from public data bases, such as Breast Cancer samples from the TCGA project [18, 19].

2. Development of methods, either in terms of new algorithms or in terms of combinative workflows, which will be able to improve, and facilitate, the analysis and biological interpretation of those data sets to be integrated.

3. Implementation of the methods developed for this study in the appropriate bioinformatics tools, such as an R package or a web-based application, to facilitate their use in the context of biomedical research projects.

Here follows a brief description of these main five activities, the methods in which they are initially based, the objectives that they are related to, and the corresponding results:

1. Application of some state-of-the-art methods for integrative multi-omics data analysis to the study of human brain tissue samples, collected by the Neurovascular Diseases Laboratory at Vall d'Hebron Research Institute. This part is already finished, and led to publications in 2018 and 2021 [37, 38]. Researchers obtained different omics data from necropsies, which had been processed to obtain mRNA, microRNA and protein expression values. Each dataset had been first analyzed independently using standard bioinformatics protocols [20]. These analyses allowed selecting subsets of relevant features, for each type of data, to be used in the integrative analysis. Among all available options, we decided to use two distinct and complementary approaches: (I) Multiple Co-inertia Analysis implemented in Bioconductor packages made4 [21] and mogsa [22], and (II) Regularized Canonical Correlation Analysis with Sparse Partial Least Squares regression (sPLS), provided by mixomics R package [23]. This work had been presented at some meetings [39, 40, 41, 43] and in an already published extended abstract's series book [35]. This step had been obviously useful for the achievement of the objective number 3 explained in the previous section, which aims on the study of the regulome's response to ischemic stroke, but also useful for detecting the advantages and drawbacks of the methods applied, thus setting the basis for the work regarding to objective number 2.

2. Reproduction of the same analyses steps performed in point 1) above with publicly available databases, such as distinct omics data from 150 samples from the TCGA-BRCA collection. This data set contains the expression or abundance of mRNA, miRNA and proteomics for 150 breast cancer samples previously prefiltered, as explained in Rohart et al. [29], and allows identifying a good multi-omics signature to discriminate between Basal, Her2 and Luminal A breast cancer subtypes. This work is already finished, and complies with objectives 3 and 2.

3. Use of all the data sets analyzed up to this point to make a comparison of results between the main implemented methods, and eventually some others, which is the aim of objective 1. This is based on quantitative and qualitative comparison and visualization methods, such as those explained by Thallinger [24] and Martin [25], going from simple Venn diagrams to more complex, network analysis, software such as some specific R packages [20] or Cytoscape [26]. The focus here is to use graphical visualization elements to compare the results of the analyses with and without the addition of biological information.

4. Development of new methods and/or workflows in order to improve and/or combine the benefits from the selected approaches, with focus in those allowing the addition of biological significance to the integration process. Here follows an overview of the methods developed to expand the original datasets (X, Y) with annotations (Ax, Ay) to obtain new blocks of data (Nx, Ny,and Nxy). And the workflow has been implemented adapting the integrative pipelines applied so far to the R targets package [33], a pipeline toolkit that improves reproducibility, skipping unnecessary steps already up to date and showing tangible evidence that the results match the underlying code and data. The development of this targets workflow is intended to comply with the objective number 2 of this working plan.

5. Implementation of the methods resulting from 4) as a new R package to be submitted to Bioconductor repository [27], and, finally, to complete objective 4 of this thesis plan, as a web application [28] to be used in further steps of the current biomedical research projects in which our collaborators are implied, as well as in future studies.

3.2 Explanation of the methods

The addition of biological annotations to the data sets being integrated, prior to the integrative analysis itself, can be useful to improve the integration/analysis outcomes as well as their biological interpretability.

Passos principals explicats aqui:

A. Pre process omics datasets in order to include biological information before the joint analysis –> Expanded datasets

B. Analysis of the expanded datasets by the use of contrasted joint Dimensionallity Reduction techniques

C. Process semi automation in ease to use tools

Start the process already having a couple [punt de millora: admetre 3 o + inputs] of data sets from distinct 'omics sources, mapped to gene ids (if GO annotation has to be performed), containing the results from a selection of differentially expressed genes or most relevant proteins analysis, or similar. [explicar aquí els requeriments de format dels data sets d'entrada!!]

For each input data set, if annotations are not already provided, two distinct basic annotation methods can be performed:

(i) a basic GO mapping, returning annotations to those GO entities for which we find more than a certain number of features (gene ids coming from our data set) annotated to them, [mostrar formula] [mostrar exemple]

(ii) a Gene Enrichment Analysis (based on Hypergeometric tests against all GO categories, with FDR correction[ref clusterProfiler]) is performed in order to

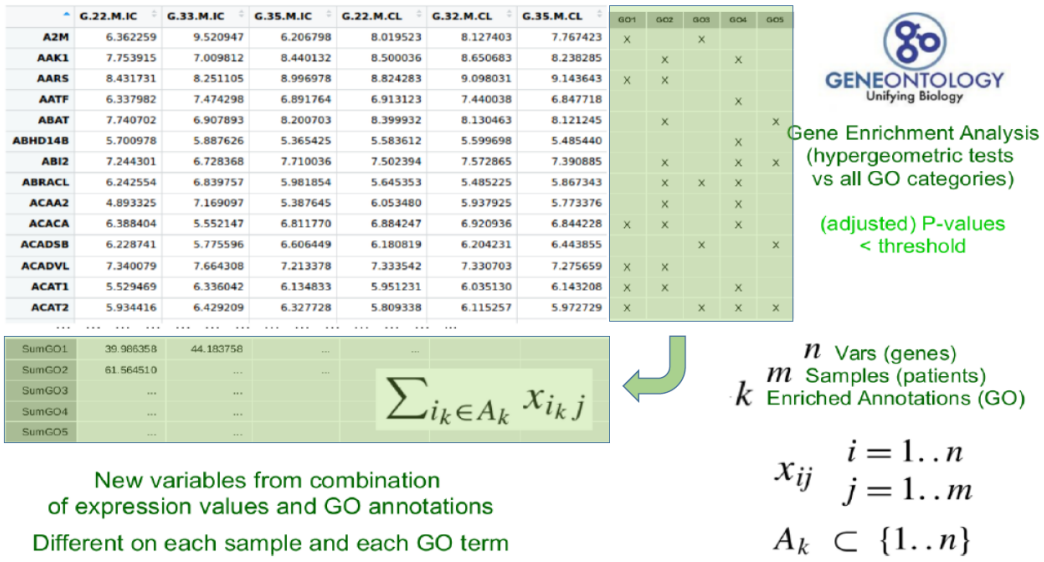

Figure 3.1: Addition of GO terms

retrieve the most relevant annotations to that set of genes/features. [mostrar exemple] [afegir aquí la opció d'afegir les anotacions com a individus suplementaris enlloc de variables]

Figure 3.1 is an example.

Alternatively, manual annotations can be provided (eg. GO terms, canonical pathways, or even annotation to custom entities) as an optional input file. [mostrar el format requerit].

Other annotation methods can be implemented, as functions to be used by the main pipeline, if more complex methods for biological information addition are required.

[Mostrar el format final de les anotacions, com a matrius dels data sets amb anotacions binàries 1/0 com a columnes extra]

Once the annotations are already computed, mapping each feature of the input data set to the corresponding biological entity, they can be used to generate new features (as new rows), computing the average value [punt de millora: £funció de ponderació?] of the expression/intensity values from all original features being mapped to the annotated biological entities.

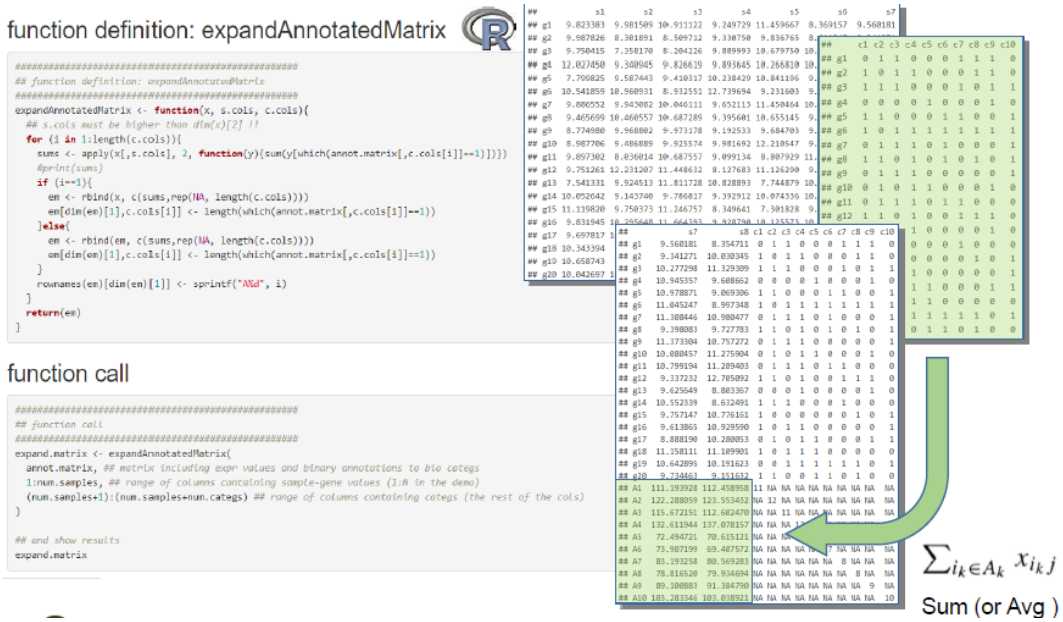

Figure 3.2: Addition of news feats

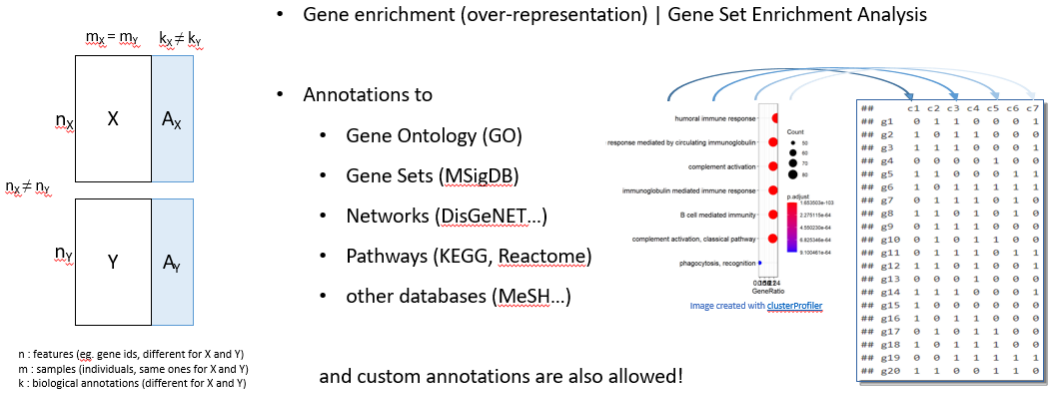

Figure 3.3: Gene enrichment diagram

(Traduir) Una vez tenemos las matrices anotadas (Figura 3.4, parte azul) pasamos a generar las matrices Expandidas (en verde) numerizando estas anotaciones, o sea, calculando la media de las expresiones numéricas de cada individuo para las variables anotadas a cada categoría. Esto se realiza con el producto matricial de los valores numéricos iniciales (expresion, proteinas...) con las matrices traspuestas de sus anotaciones, y luego con la matriz inversa de una matriz diagonal del conteo de cuantas anotaciones ha tenido cada categoría o entidad anotada.

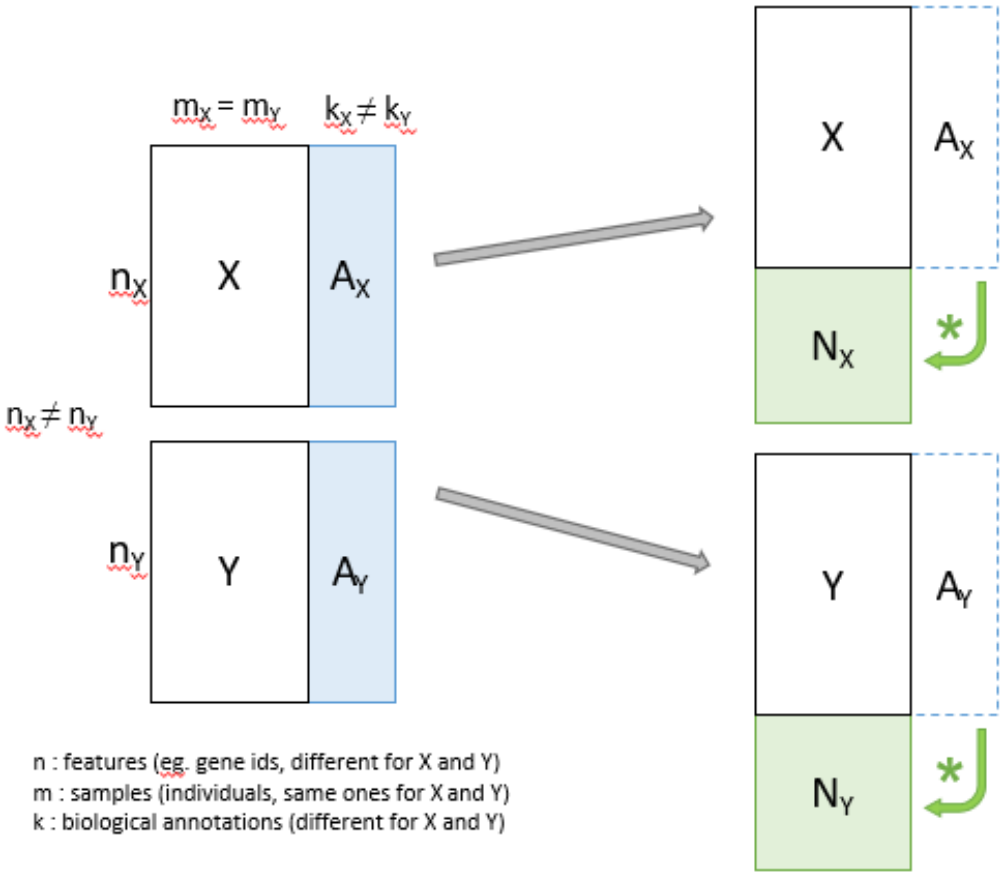

Figure 3.4: Matrix expansion diagram

3.2.1 Detail of the integrative data analyses applied...

Mètodes:

1- Significació biològica, com faig les anotacions 2- Expansió de les matrius (creació de noves vars a partir de les anotacions) 3- Anàlisi factorial en detall, + MCIA + RGCCA 4- TFM sobre workflows i automatització -> Paquet targets en general

3.2.2 Targets PIPELINE concept:

Targets workflow diagram (Figure 3.7) showing the steps corresponding with the complete process: The pipeline starts from (A) a couple of 'omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices). These are converted to R data frames with features in rows and samples in columns.

$$N_G = \phi(G, A_G), \quad G \in X, Y, \ldots,$$

$$N_X = \frac{1}{r_X}(X \times A'_X),$$

$$N_Y = \frac{1}{r_Y}(Y \times A'_Y),$$

Figure 3.5: Addition of new feats (2)

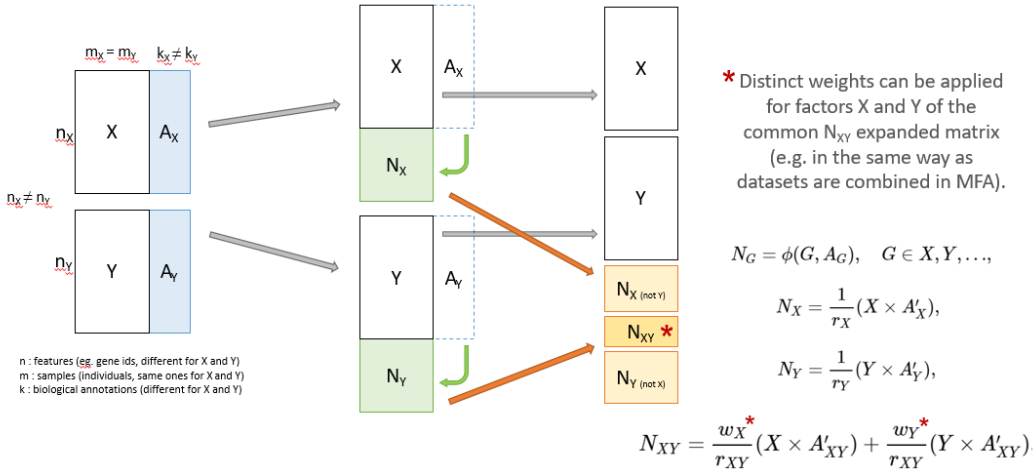

Figure 3.6: Matrix expansion diagram (2)

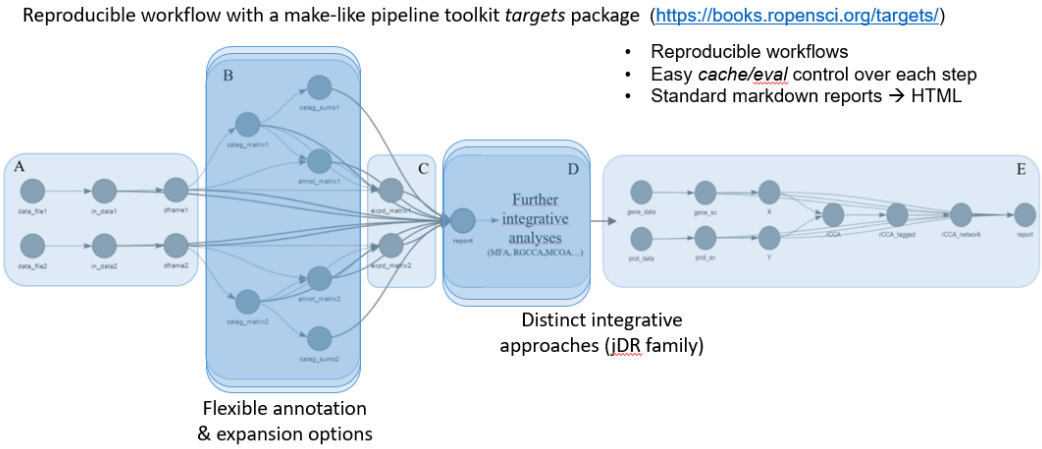

Figure 3.7: Workflow overview

Then, a data frame containing related annotations (B) is created, or loaded, for each given input matrix, and used to expand these original data, in order to end up with a pair of data frames (C) containing the original values plus the average expression/abundance values of the features related to each annotation as new features in additional rows. After that, distinct Dimension Reduction Methods are applied to perform the integrative analysis (D), and finally, an R markdown report (E) is rendered to show steps and main results of the full process.

3.2.3 Comparison of ODA results

3.2.4 Numeric measurement

% variabilitat explicat segons la estructura de la intersecció de les 2 taules

3.2.5 Biological interpretation

3.2.6 R package creation

4

Results

Contents

Text de presentacio dels resultats...

4.1 Results from the analysis of human brain tissue samples

4.2 Results from the expansion of omics data with biological annotations

Figure 4.1 is an snapshot (F) of one of the heat maps created to show the expanded matrices obtained in (Figures 3.4 i 3.5 prèvies, de Methods).

4.3 Results from the analysis of 150 TCGA-BRCA samples

Figure 4.2 contains some of the graphical results of the analysis of the 150 samples from TCGA-BRCA: Heat maps (A, C) and association networks (B, D) resulting from the integration by Regularized Canonical Correlations Analysis with mixomics R package. Performed with the original data sets (A, B) or using data expanded with biological annotations to Gene Ontology (C, D), so adding some GO terms to the features from each source, where the outputs contain higher level of information (higher density in both type of plots).

4.4 Results from the application of MFA on TCGA-BRCA data with, and without, expanded data

Figure 4.3 includes a Correlation Circle (left), with most relevant genes, proteins and added GO annotations. Distribution of samples (right) along the first two

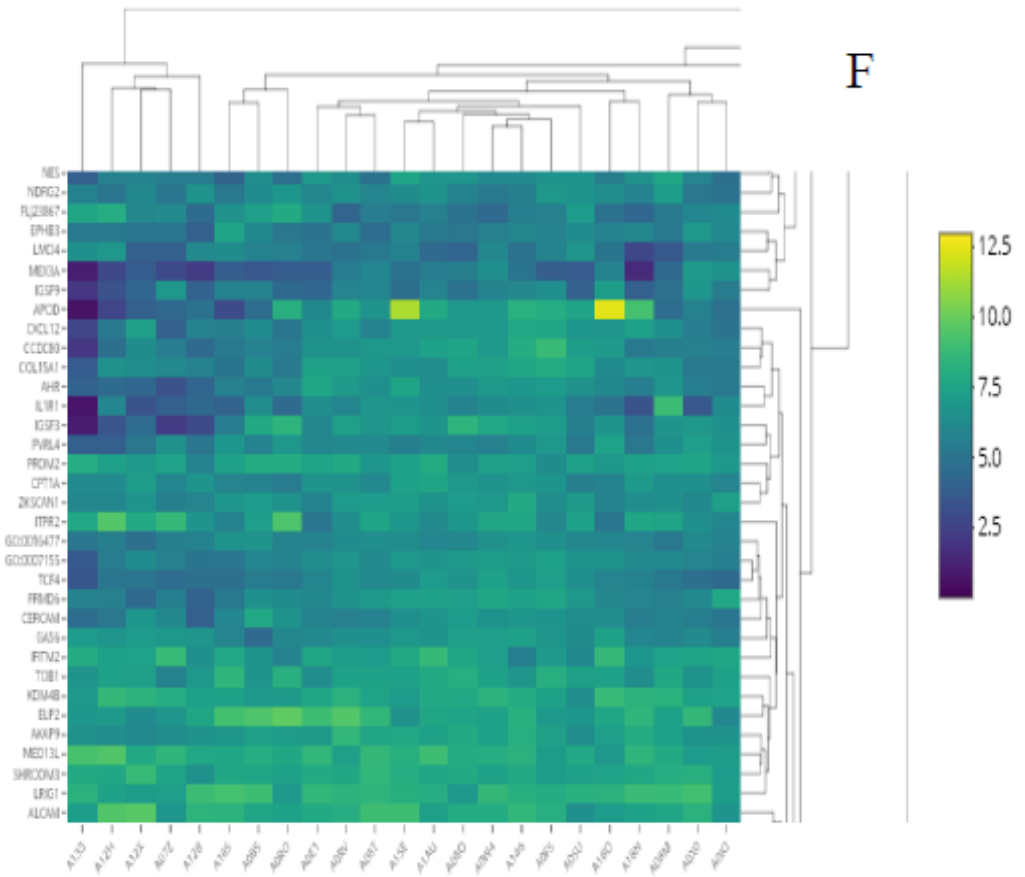

Figure 4.1: Heapmap of an expanded matrix

plotted dimensions. Both results coming from the application of Multiple Factor Analysis (FactoMineR and factoextra R packages) performed on the same 150 samples (Basal, Her2 and LuminalA conditions) from TCGA-BRCA.

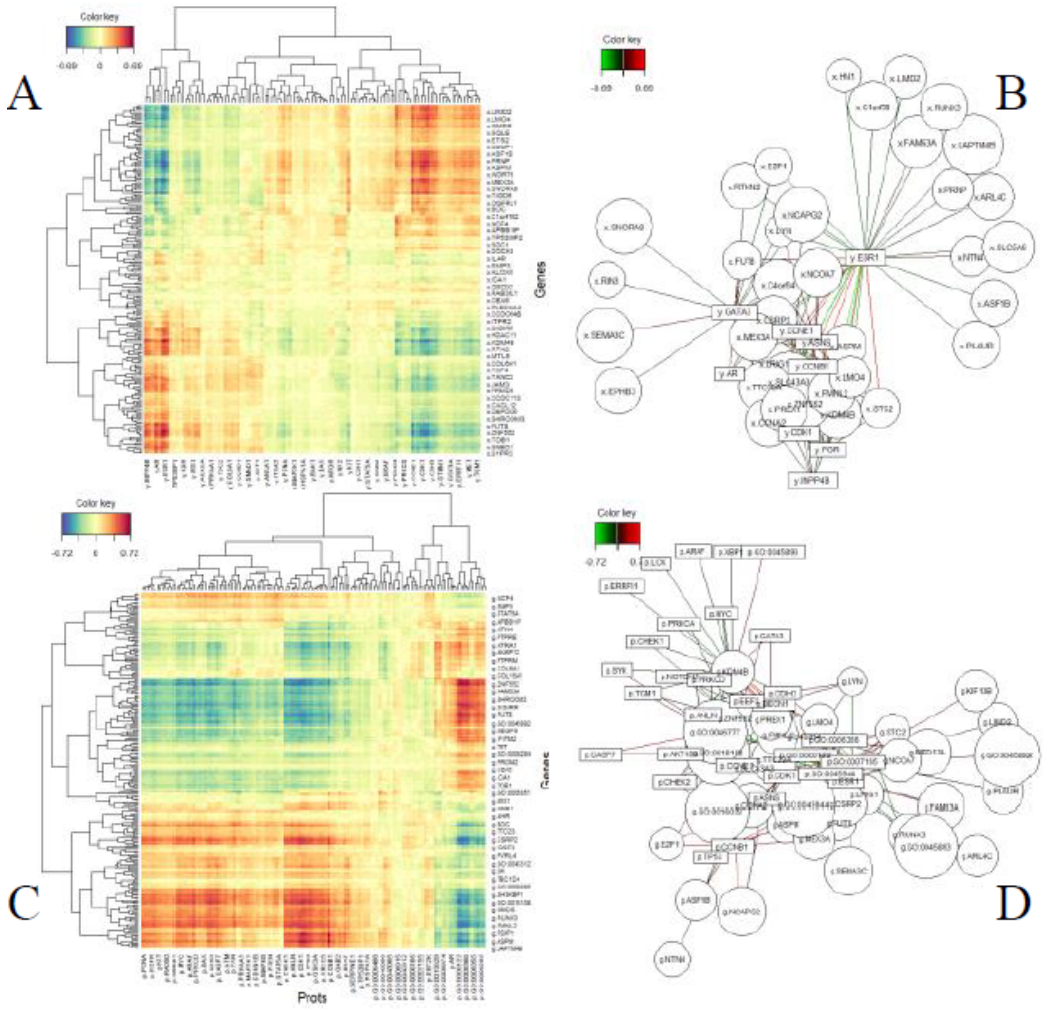

Figure 4.2: BRCA results overview

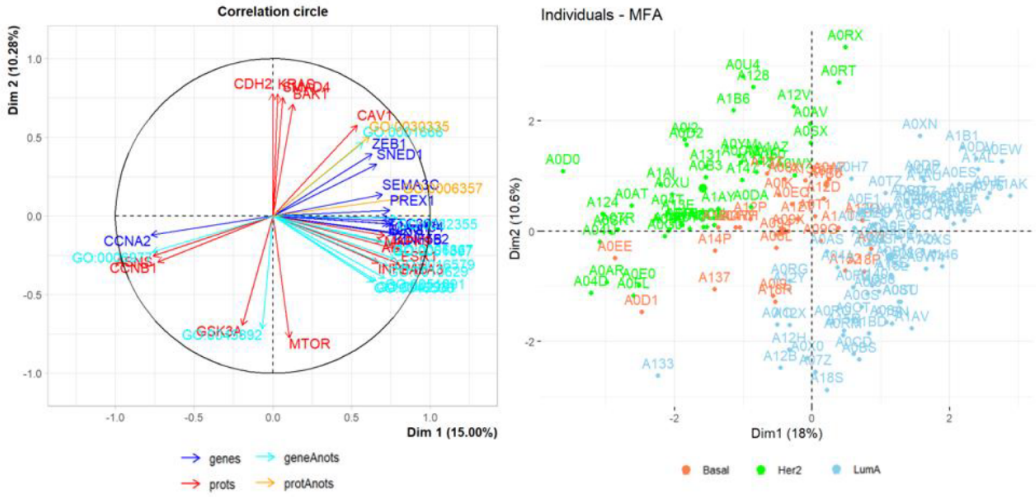

Figure 4.3: BRCA results with MFA

4.5 Citations

The usual way to include citations in an *R Markdown* document is to put references in a plain text file with the extension **.bib**, in **BibTex** format.[1] Then reference the path to this file in **index.Rmd**'s YAML header with `bibliography: example.bib`.

Most reference managers can create a .bib file with you references automatically. However, the **by far** best reference manager to use with *R Markdown* is Zotero with the Better BibTex plug-in, because the `citr` plugin for RStudio (see below) can read references directly from your Zotero library!

Here is an example of an entry in a **.bib** file:

```
@article{Shea2014,
  author =        {Shea, Nicholas and Boldt, Annika},
  journal =       {Trends in Cognitive Sciences},
  pages =         {186--193},
  title =         {{Supra-personal cognitive control}},
  volume =        {18},
  year =          {2014},
  doi =           {10.1016/j.tics.2014.01.006},
}
```

In this entry highlighed section, 'Shea2014' is the **citation identifier**. To default way to cite an entry in your text is with this syntax: `[@citation-identifier]`.

So I might cite some things (Lottridge et al., 2012; Mill, 1965 [1843]; Shea et al., 2014).

4.5.1 Appearance of citations and references section (pandoc)

By default, `oxforddown` lets Pandoc handle how citations are inserted in your text and the references section. You can change the appearance of citations and references by specifying a CSL (Citation Style Language) file in the `csl` metadata

[1] The bibliography can be in other formats as well, including EndNote (**.enl**) and RIS (**.ris**), see rmarkdown.rstudio.com/authoring_bibliographies_and_citations.

field of **index.Rmd**. By default, `oxforddown` by the Americal Psychological Association (7th Edition), which is an author-year format.

With this style, a number of variations on the citation syntax are useful to know:

- Put author names outside the parenthesis
  - This: `@Shea2014 says blah.`
  - Becomes: Shea et al. (2014) says blah.
- Include only the citation-year (in parenthesis)
  - This: `Shea et al. says blah [-@Shea2014]`
  - Becomes: Shea et al. says blah (2014)
- Add text and page or chapter references to the citation
  - This: `[see @Shea2014, pp. 33-35; also @Wu2016, ch. 1]`
  - Becomes: Blah blah (see Shea et al., 2014, pp. 33–35; also Wu, 2016, ch. 1).

If you want a numerical citation style instead, try `csl: bibliography/transactions-on-comput` or just have a browse through the Zotero Style Repository and look for one you like. For convenience, you can set the line spacing and the space between the bibliographic entries in the reference section directly from the YAML header in **index.Rmd**.

If you prefer to use `biblatex` or `natbib` to handle references, see this chapter.

4.5.2 Insert references easily with RStudio's Visual Editor

For an easy way to insert citations, use RStudio's Visual Editor. Make sure you have the latest version of RStudio – the visual editor was originally really buggy, especially in relation to references, but as per v2022.02.0, it's great!

4.6 Cross-referencing

We can make cross-references to **sections** within our document, as well as to **figures** (images and plots) and **tables**.

The general cross-referencing syntax is `\@ref(label)`

4.6.1 Section references

Headers are automatically assigned a reference label, which is the text in lower caps separated by dashes. For example, `# My header` is automatically given the label `my-header`. So `# My header` can be referenced with `\@ref(my-section)`

Remember what we wrote in section 4.5?

We can also use **hyperlink syntax** and add # before the label, though this is only guaranteed to work properly in HTML output:

- So if we write `Remember what we wrote up in [the previous section](#citations)?`
- It becomes Remember what we wrote up in the previous section?

Creating custom labels

It is a very good idea to create **custom labels** for our sections. This is because the automatically assigned labels will change when we change the titles of the sections - to avoid this, we can create the labels ourselves and leave them untouched if we change the section titles.

We create custom labels by adding `{#label}` after a header, e.g. `# My section {#my-label}`. See our chapter title for an example. That was section **??**.

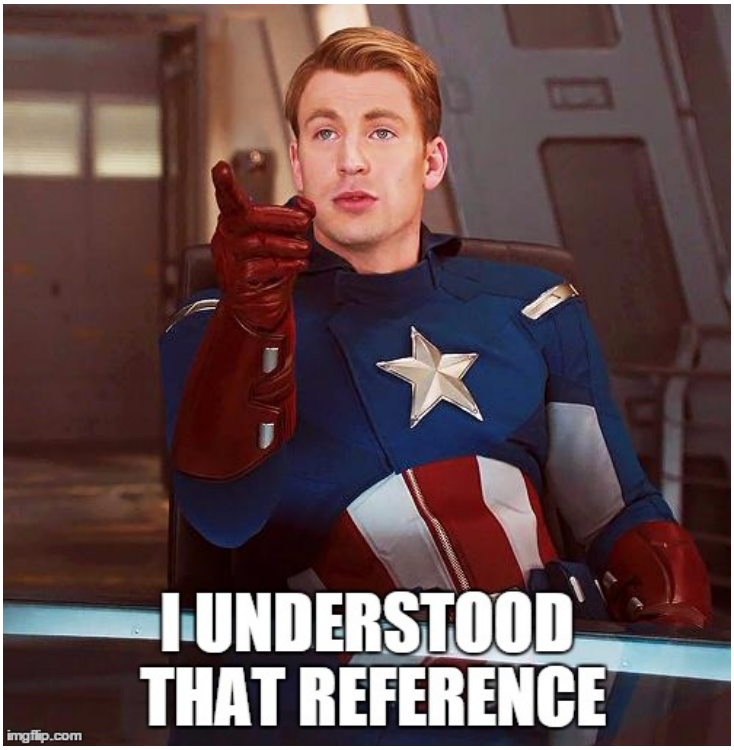

Figure 4.4: A marvel-lous meme

4.6.2 Figure (image and plot) references

- To refer to figures (i.e. images and plots) use the syntax `\@ref(fig:label)`
- **GOTCHA**: Figures and tables must have captions if you wish to cross-reference them.

Let's add an image:

```
knitr::include_graphics("figures/sample-content/captain.jpeg")
```

We refer to this image with `\@ref(fig:captain)`. So Figure 4.4 is this image. And in Figure 5.5 we saw a cars plot.

4.6.3 Table references

- To refer to tables use the syntax `\@ref(tab:label)`

Let's include a table:

Table 4.1: Stopping cars

speed	dist
4	2
4	10
7	4
7	22
8	16

```
knitr::kable(cars[1:5,],
             caption="Stopping cars")
```

We refer to this table with `\@ref(tab:cars-table2)`. So Table 4.1 is this table.

And in Table 5.6 we saw more or less the same cars table.

4.6.4 Including page numbers

Finally, in the PDF output we might also want to include the page number of a reference, so that it's easy to find in physical printed output. LaTeX has a command for this, which looks like this: `\pageref{fig/tab:label}` (note: curly braces, not parentheses)

When we output to PDF, we can use raw LaTeX directly in our .Rmd files. So if we wanted to include the page of the cars plot we could write:

- This: `Figure \@ref(fig:cars-plot) on page \pageref(fig:cars-plot)`
- Becomes: Figure 5.5 on page 54

Include page numbers only in PDF output

A problem here is that LaTeX commands don't display in HTML output, so in the gitbook output we'd see simply "Figure 5.5 on page".

One way to get around this is to use inline R code to insert the text, and use an `ifelse` statement to check the output format and then insert the appropriate text.

- So this: `` `r ifelse(knitr::is_latex_output(), "Figure \\@ref(fig:cars-plot) on page \\pageref{fig:cars-plot}", "")` ``

- Inserts this (check this on both PDF and gitbook): Figure 5.5 on page 54

Note that we need to escape the backslash with another backslash here to get the correct output.

4.7 Collaborative writing

Best practices for collaboration and change tracking when using R Markdown are still an open question. In the blog post **One year to dissertate** by Lucy D'Agostino, which I highly recommend, the author notes that she knits .Rmd files to a word document, then uses the `googledrive` R package to send this to Google Drive for comments / revisions from co-authors, then incorporates Google Drive suggestions *by hand* into the .Rmd source files. This is a bit clunky, and there are ongoing discussions among the *R Markdown* developers about what the best way is to handle collaborative writing (see issue #1463 on GitHub, where CriticMarkup is among the suggestions).

For now, this is an open question in the community of R Markdown users. I often knit to a format that can easily be imported to Google Docs for comments, then go over suggested revisions and manually incorporate them back in to the .Rmd source files. For articles, I sometimes upload a near-final draft to Overleaf, then collaboratively make final edits to the LaTeX file there. I suspect some great solution will be developed in the not-to-distant future, probably by the RStudio team.

4.8 Additional resources

- *R Markdown: The Definitive Guide* - `https://bookdown.org/yihui/rmarkdown/`
- *R for Data Science* - `https://r4ds.had.co.nz`

This chapter describes a number of additional tips and tricks as well as possible customizations to the `oxforddown` thesis.

4.9 Chunk caching and the __bookdown_files folder

If you set `cache=TRUE` in a code chunk, in order to cache its results if it's time-consuming to run see the R Markdown documentation, then the files for the caching are stored in the **__bookdown_files** folder.

If you don't use caching and you would like to just have the **__bookdown_files** folder deleted after the build process is complete, then set `allow_cache = FALSE` in **index.Rmd**'s call to `knit_thesis`.

That is, your YAML should then look like this:

```
knit: (function(input, ...) {
    thesis_formats <- "pdf";

    source("scripts_and_filters/knit-functions.R");
    knit_thesis(input, thesis_formats, allow_cache = FALSE, ...)
  })
```

4.10 Front matter

4.10.1 Shorten captions shown in the list of figures (PDF)

You might want your list of figures (which follows the table of contents) to have shorter (or just different) figure descriptions than the actual figure captions.

Do this using the chunk option `fig.scap` ('short caption'), for example `{r captain-image, fig.cap="A very long and descriptive (and potentially boring) caption that doesn't fit in the list of figures, but helps the reader understand what the figure communicates.", fig.scap="A concise description for the list of figures"`

4.10.2 Shorten captions shown in the list of tables (PDF)

You might want your list of tables (which follows the list of figures in your thesis front matter) to have shorter (or just different) table descriptions than the actual table captions.

If you are using `knitr::kable` to generate a table, you can do this with the argument `caption.short`, e.g.:

```
knitr::kable(mtcars,
             caption = "A very long and descriptive (and
             ↪  potentially
             boring) caption that doesn't fit in the list of
↪  figures,
             but helps the reader understand what the figure
             communicates.",
             caption.short = "A concise description for the list of
             ↪  tables")
```

4.11 Shorten running header (PDF)

You might want a chapter's running header (i.e. the header showing the title of the current chapter at the top of page) to be shorter (or just different) to the actual chapter title.

Do this by adding the latex command `\chaptermark{My shorter version}` after your chapter title.

For example, chapter **??**'s running header is simply 'Cites and cross-refs', because it begins like this:

```
# Citations, cross-references, and collaboration {#cites-and-refs}
\chaptermark{Cites and cross-refs}
```

4.12 Unnumbered chapters

To make chapters unnumbered (normally only relevant to the Introduction and/or the Conclusion), follow the chapter header with `{-}`, e.g. `# Introduction {-}`.

When you do this, you must also follow the heading with these two latex commands:

```
\adjustmtc
\markboth{The Name of Your Unnumbered Chapter}{}
```

Otherwise the chapter's mini table of contents and the running header will show the previous chapter.

4.13 Beginning chapters with quotes (PDF)

The OxThesis LaTeX template lets you inject some wittiness into your thesis by including a block of type `savequote` at the beginning of chapters. To do this, use the syntax ```` ```{block type='savequote'} ````.[2]

Add the reference for the quote with the chunk option `quote_author="my author name"`. You will also want to add the chunk option `include=knitr::is_latex_output()` so that quotes are only included in PDF output.

It's not possible to use markdown syntax inside chunk options, so if you want to e.g. italicise a book name in the reference use a 'text reference': Create a named piece of text with '(ref:label-name) My text', then point to this in the chunk option with `quote_author='(ref:label-name)'`.

4.14 Highlighting corrections (HTML & PDF)

For when it comes time to do corrections, you may want to highlight changes made when you submit a post-viva, corrected copy to your examiners so they can quickly verify you've completed the task. You can do so like this:

4.14.1 Short, inline corrections

Highlight **short, inline corrections** by doing `[like this]{.correction}` — the text between the square brackets will then be highlighted in blue in the output.

Note that pandoc might get confused by citations and cross-references inside inline corrections. In particular, it might get confused by `"[what @Shea2014 said]{.correction}"` which becomes what Shea et al. (2014) said In such cases,

[2] For more on custom block types, see the relevant section in *Authoring Books with R Markdown*.

you can use LaTeX syntax directly. The correction highlighting uses the soul package, so you can do like this:

- If using biblatex for references, use `"\hl{what \textcite{Shea2014} said}`
- If using natbib for references, use `"\hl{what \cite{Shea2014} said}`

Using raw LaTeX has the drawback of corrections then not showing up in HTML output at all, but you might only care about correction highlighting in the PDF for your examiners anyway!

4.14.2 Blocks of added or changed material

Highlight entire **blocks of added or changed material** by putting them in a block of type `correction`, using the syntax ```` ```{block type='correction'} ````.[3] Like so:

For larger chunks, like this paragraph or indeed entire figures, you can use the `correction` block type. This environment **highlights paragraph-sized and larger blocks** with the same blue colour.

Note that correction blocks cannot be included in word output.

4.14.3 Stopping corrections from being highlighted

To turn off correction highlighting, go to the YAML header of **index.Rmd**, then:

- PDF output: set `corrections: false`
- HTML output: remove or comment out `- templates/corrections.css`

[3] In the **.tex** file for PDF output, this will put the content between `\begin{correction}` and `\end{correction}`; in gitbook output it will be put between `<div class="correction">` and `</div>`.

4.15 Apply custom font color and highlighting to text (HTML & PDF)

The lua filter that adds the functionality to highlight corrections adds two more tricks: you can apply your own choice of colour to highlight text, or change the font color. The syntax is as follows:

> Here's `[some text in pink highlighting]{highlight="pink"}`
> Becomes: Here's some text in pink highlighting.
>
> `[Here's some text with blue font]{color="blue"}`
> Becomes: Here's some text with blue font
>
> Finally — never, ever actually do this – `[here's some text with black highlighting and yellow font]{highlight="black" color="yellow"}`
> Becomes: here's some text with black highlighting and yellow font

The file **scripts_and_filters/colour_and_highlight.lua** implements this, if you want to fiddle around with it. It works with both PDF and HTML output.

4.16 Adding a second abstract (PDF)

You may need two abstracts in your thesis, if you e.g. need both an abstract in English and some other language.

You can add a second abstract in **index.Rmd** like so:

```
abstract-second-heading: "Resumé"
abstract-second: "This is the second abstract, for example in
↪  beautiful French."
```

4.17 Including another paper in your thesis - embed a PDF document

You may want to embed existing PDF documents into the thesis, for example if your department allows a 'portfolio' style thesis and you need to include an existing typeset publication as a chapter.

In gitbook output, you can simply use `knitr::include_graphics` and it should include a scrollable (and downloadable) PDF. You will probably want to set the chunk options `out.width='100%'` and `out.height='1000px'`:

```
knitr::include_graphics(
  "figures/sample-content/pdf_embed_example/Lyngs2020_FB.pdf")
```

In LaTeX output, however, this approach can cause odd behaviour. Therefore, when you build your thesis to PDF, split the PDF into an alphanumerically sorted sequence of **single-page** PDF files (you can do this automatically with the package `pdftools`). You can then use the appropriate LaTeX command to insert them, as shown below (for brevity, in the `oxforddown` PDF sample content we're only including two pages). *Note that the chunk option* ***`results='asis'`*** *must be set.* You may also want to remove margins from the PDF files, which you can do with Adobe Acrobat (paid version) and likely other software.

```
# install.packages(pdftools)
# split PDF into pages stored in
  figures/sample-content/pdf_embed_example/split/
#
  pdftools::pdf_split("figures/sample-content/pdf_embed_example/Lyngs2020_FB.pd
#       output =
  "figures/sample-content/pdf_embed_example/split/")

# grab the pages
pages <-
  list.files("figures/sample-content/pdf_embed_example/split",
  full.names = TRUE)

# set how wide you want the inserted PDFs to be:
# 1.0 is 100 per cent of the oxforddown PDF page width;
# you may want to make it a bit bigger
pdf_width <- 1.2

# for each PDF page, insert it nicely and
# end with a page break
cat(stringr::str_c("\\newpage \\begin{center}
  \\makebox[\\linewidth][c]{\\includegraphics[width=", pdf_width,
  "\\linewidth]{", pages, "}} \\end{center}"))
```

‘I Just Want to Hack Myself to Not Get Distracted’: Evaluating Design Interventions for Self-Control on Facebook

Ulrik Lyngs[1], **Kai Lukoff**[2], **Petr Slovak**[3], **William Seymour**[1], **Helena Webb**[1], **Marina Jirotka**[1], **Jun Zhao**[1], **Max Van Kleek**[1], **Nigel Shadbolt**[1]
[1]Department of Computer Science, University of Oxford, UK, {first.last}@cs.ox.ac.uk
[2]Human Centered Design & Engineering, University of Washington, Seattle, US, kai1@uw.edu
[3]Department of Informatics, King’s College London, UK, petr.slovak@kcl.ac.uk

ABSTRACT
Beyond being the world’s largest social network, Facebook is for many also one of its greatest sources of digital distraction. For students, problematic use has been associated with negative effects on academic achievement and general wellbeing. To understand what strategies could help users regain control, we investigated how simple interventions to the Facebook UI affect behaviour and perceived control. We assigned 58 university students to one of three interventions: goal reminders, removed newsfeed, or white background (control). We logged use for 6 weeks, applied interventions in the middle weeks, and administered fortnightly surveys. Both goal reminders and removed newsfeed helped participants stay on task and avoid distraction. However, goal reminders were often annoying, and removing the newsfeed made some fear missing out on information. Our findings point to future interventions such as controls for adjusting types and amount of available information, and flexible blocking which matches individual definitions of ‘distraction’.

Author Keywords
Facebook; problematic use; self-control; distraction; ICT non-use; addiction; focus; interruptions

CCS Concepts
•**Human-centered computing** → **Empirical studies in HCI;**

INTRODUCTION

Research on ‘Problematic Facebook Use’ (PFU) has investigated correlations between Facebook use and negative effects on outcomes such as level of academic achievement [35] and subjective wellbeing [58, 57]. A cross-cutting finding is that negative outcomes are associated with difficulty at exerting self-control over use, as well as specific use patterns including viewing friends’ wide-audience broadcasts rather than receiving targeted communication from strong ties [13, 58].

CHI ’20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04.
http://dx.doi.org/10.1145/3313831.3376672

Much of this work has focused on self-control over Facebook use in student populations [2, 44, 46], with media multitasking research finding that students often give in to use which provides short-term ‘guilty pleasures’ over important, but aversive academic tasks [76, 88, 60]. In the present paper, we present a mixed-methods study exploring how two interventions to Facebook — goal reminders and removing the newsfeed — affect university students’ patterns of use and perceived control over Facebook use. To triangulate self-report with objective measurement, our study combined usage logging with fortnightly surveys and post-study interviews.

We found that both interventions helped participants stay on task and use Facebook more in line with their intentions. In terms of use patterns, goal reminders led to less scrolling, fewer and shorter visits, and less time on site, whereas removing the newsfeed led to less scrolling, shorter visits, and less content ’liked’. However, goal reminders were often experienced as annoying, and removing the newsfeed made some participants fear missing out on information. After the study, participants suggested a range of design solutions to mitigate self-control struggles on Facebook, including controls for filtering or removing the newsfeed, reminders of time spent and of use goals, and removing features that drive engagement. As an exploratory study, this work should be followed by confirmatory studies to assess whether our findings replicate, and how they may generalise beyond a student population.

RELATED WORK

Struggles with Facebook use

Whereas many uses of Facebook offer important benefits, such as social support, rapid spread of information, or facilitation of real-world interactions [78], a substantial amount of research has focused on negative aspects [58]. For example, studies have reported correlations between patterns of Facebook use and lower academic achievement [77, 86], low self-esteem, depression and anxiety [51], feelings of isolation and loneliness [2], and general psychological distress [15]. Such ‘Problematic Facebook Use’ (PFU) has been studied under various names (including ‘Facebook dependence’ [87] and ‘Facebook addiction’[5]), but a recent review summarised a common definition as ‘problematic behaviour characterised by addictive-like symptoms and/or self-regulation difficulties related to Facebook use leading to negative consequences in personal and social life’ [58].

REFERENCES

[1] Alexander T. Adams, Jean Costa, Malte F. Jung, and Tanzeem Choudhury. 2015. Mindless Computing: Designing Technologies to Subtly Influence Behavior. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 719–730. DOI: http://dx.doi.org/10.1145/2750858.2805843

[2] Sami Abdo Radman Al-Dubai, Kurubaran Ganasegeran, Mustafa Ahmed Mahdi Al-Shagga, Hematram Yadav, and John T. Arokiasamy. 2013. Adverse Health Effects and Unhealthy Behaviors among Medical Students Using Facebook. https://www.hindawi.com/journals/tswj/2013/465161/. (2013). DOI:http://dx.doi.org/10.1155/2013/465161

[3] All Party Parliamentary Group on Social Media and Young People's Mental Health and Wellbeing. 2019. *#NewFilters to Manage the Impact of Social Media on Young People's Mental Health and Wellbeing*. Technical Report. UK Parliament.

[4] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2019. *The Welfare Effects of Social Media*. Working Paper 25514. National Bureau of Economic Research. DOI: http://dx.doi.org/10.3386/w25514

[5] Cecilie Schou Andreassen, Torbjørn Torsheim, Geir Scott Brunborg, and Staale Pallesen. 2012. Development of a Facebook Addiction Scale. *Psychological Reports* 110, 2 (apr 2012), 501–517. DOI: http://dx.doi.org/10.2466/02.09.18.PR0.110.2.501-517

[6] Yummy Apps. 2019. Todobook. (May 2019).

[7] Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American Psychologist* 37, 2 (1982), 122–147. DOI: http://dx.doi.org/10.1037/0003-066x.37.2.122

[8] Fanni Bányai, Ágnes Zsila, Orsolya Király, Aniko Maraz, Zsuzsanna Elekes, Mark D. Griffiths, Cecilie Schou Andreassen, and Zsolt Demetrovics. 09-Jan-2017. Problematic Social Media Use: Results from a Large-Scale Nationally Representative Adolescent Sample. *PLOS ONE* 12, 1 (09-Jan-2017), e0169839. DOI: http://dx.doi.org/10.1371/journal.pone.0169839

[9] Elliot T Berkman, Cendri A Hutcherson, Jordan L Livingston, Lauren E Kahn, and Michael Inzlicht. 2017. Self-Control as Value-Based Choice. *Current Directions in Psychological Science* 26, 5 (2017), 422–428. DOI: http://dx.doi.org/10.1177/0963721417704394

[10] Walter R. Boot, Daniel J. Simons, Cary Stothart, and Cassie Stutts. 2013. The Pervasive Problem with Placebos in Psychology. *Perspectives on Psychological Science* 8, 4 (jul 2013), 445–454. DOI: http://dx.doi.org/10.1177/1745691613491271

[11] Amara Brook. 2011. Ecological Footprint Feedback: Motivating or Discouraging? *Social Influence* 6, 2 (April 2011), 113–128. DOI: http://dx.doi.org/10.1080/15534510.2011.566801

[12] Gharad Bryan, Dean Karlan, and Scott Nelson. 2010. Commitment Devices. *Annual Review of Economics* 2, 1 (Sept. 2010), 671–698. DOI:http://dx.doi.org/10.1146/annurev.economics.102308.124324

[13] Moira Burke and Robert E. Kraut. 2016. The Relationship Between Facebook Use and Well-Being Depends on Communication Type and Tie Strength. *Journal of Computer-Mediated Communication* 21, 4 (2016), 265–281. DOI: http://dx.doi.org/10.1111/jcc4.12162

[14] Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social Network Activity and Social Well-Being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1909–1912. DOI: http://dx.doi.org/10.1145/1753326.1753613

[15] Wenhong Chen and Kye-Hyoung Lee. 2013. Sharing, Liking, Commenting, and Distressed? The Pathway between Facebook Interaction and Psychological Distress. *Cyberpsychology, Behavior and Social Networking* 16, 10 (oct 2013), 728–734. DOI: http://dx.doi.org/10.1089/cyber.2012.0272

[16] Justin Cheng, Moira Burke, and Elena Goetz Davis. 2019. Understanding Perceptions of Problematic Facebook Use: When People Experience Negative Life Impact and a Lack of Control. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 199:1–199:13. DOI: http://dx.doi.org/10.1145/3290605.3300429

[17] Jacob Cohen. 1992. A Power Primer. *Psychological Bulletin* 112, 1 (1992), 155–159. DOI: http://dx.doi.org/10.1037/0033-2909.112.1.155

[18] Anna L Cox, Sandy J J Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1389–1397. DOI: http://dx.doi.org/10.1145/2851581.2892410

[19] Helen Creswick, Liz Dowthwaite, Ansgar Koene, Elvira Perez Vallejos, Virginia Portillo, Monica Cano, and Christopher Woodard. 2019. "... They don't really listen to people". *Journal of Information, Communication and Ethics in Society* 17, 2 (May 2019), 167–182. DOI: http://dx.doi.org/10.1108/jices-11-2018-0090

[20] Angela L. Duckworth, Katherine L. Milkman, and David Laibson. 2018. Beyond Willpower: Strategies for Reducing Failures of Self-Control. *Psychological Science in the Public Interest* 19, 3 (Dec. 2018), 102–129. DOI: http://dx.doi.org/10.1177/1529100618821893

5 Discussion

Contents

Potser no cal posar la TOC aqui£?

Resum de l'article. Apuntant a les conclusions. Comentant problemes i limitacions (emprar combinacions lineals de variables per crear-ne de noves).

Possibles extensions [punts de millora] Comentar i descriure cadascun d'ells:

- Poder fer servir 3 o més conjunts de dades
- Poder ponderar els pesos de les anotacions, segons tipus, data set d'origen, etc.

- Permetre treballar amb dades faltants o, fins i tot, blocs de dades faltants.
- Millorar les opcions del paquet: mètodes d'anotació bio, mètodes d'integració, tipus de gràfics resultants...

5.1 Making LaTeX tables play nice

Dealing with tables in LaTeX can be painful. This section explains the main tricks you need to make the pain go away.

(Note: if you are looking at the ebook version, you will not see much difference in this section, as it is only relevant for PDF output!)

5.1.1 Making your table pretty

When you use `kable` to create tables, you will almost certainly want to set the option `booktabs = TRUE`. This makes your table look a million times better:

```
library(knitr)
library(tidyverse)

head(mtcars) %>%
  kable(booktabs = TRUE)
```

	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

Compare this to the default style, which looks terrible:

```
head(mtcars) %>%
  kable()
```

	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

	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

5.1.2 If your table is too wide

You might find that your table expands into the margins of the page, like the tables above. Fix this with the `kable_styling` function from the `kableExtra` package:

```
library(kableExtra)

head(mtcars) %>%
  kable(booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down")
```

This scales down the table to fit the page width.

5.1.3 If your table is too long

If your table is too long to fit on a single page, set `longtable = TRUE` in the `kable` function to split the table across multiple pages.

```
a_long_table <- rbind(mtcars, mtcars)

a_long_table %>%
  select(1:8) %>%
  kable(booktabs = TRUE, longtable = TRUE)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs

Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1
Mazda RX41	21.0	6	160.0	110	3.90	2.620	16.46	0
Mazda RX4 Wag1	21.0	6	160.0	110	3.90	2.875	17.02	0
Datsun 7101	22.8	4	108.0	93	3.85	2.320	18.61	1
Hornet 4 Drive1	21.4	6	258.0	110	3.08	3.215	19.44	1
Hornet Sportabout1	18.7	8	360.0	175	3.15	3.440	17.02	0
Valiant1	18.1	6	225.0	105	2.76	3.460	20.22	1
Duster 3601	14.3	8	360.0	245	3.21	3.570	15.84	0
Merc 240D1	24.4	4	146.7	62	3.69	3.190	20.00	1
Merc 2301	22.8	4	140.8	95	3.92	3.150	22.90	1
Merc 2801	19.2	6	167.6	123	3.92	3.440	18.30	1

Merc 280C1	17.8	6	167.6	123	3.92	3.440	18.90	1
Merc 450SE1	16.4	8	275.8	180	3.07	4.070	17.40	0
Merc 450SL1	17.3	8	275.8	180	3.07	3.730	17.60	0
Merc 450SLC1	15.2	8	275.8	180	3.07	3.780	18.00	0
Cadillac Fleetwood1	10.4	8	472.0	205	2.93	5.250	17.98	0
Lincoln Continental1	10.4	8	460.0	215	3.00	5.424	17.82	0
Chrysler Imperial1	14.7	8	440.0	230	3.23	5.345	17.42	0
Fiat 1281	32.4	4	78.7	66	4.08	2.200	19.47	1
Honda Civic1	30.4	4	75.7	52	4.93	1.615	18.52	1
Toyota Corolla1	33.9	4	71.1	65	4.22	1.835	19.90	1
Toyota Corona1	21.5	4	120.1	97	3.70	2.465	20.01	1
Dodge Challenger1	15.5	8	318.0	150	2.76	3.520	16.87	0
AMC Javelin1	15.2	8	304.0	150	3.15	3.435	17.30	0
Camaro Z281	13.3	8	350.0	245	3.73	3.840	15.41	0
Pontiac Firebird1	19.2	8	400.0	175	3.08	3.845	17.05	0
Fiat X1-91	27.3	4	79.0	66	4.08	1.935	18.90	1
Porsche 914-21	26.0	4	120.3	91	4.43	2.140	16.70	0
Lotus Europa1	30.4	4	95.1	113	3.77	1.513	16.90	1
Ford Pantera L1	15.8	8	351.0	264	4.22	3.170	14.50	0
Ferrari Dino1	19.7	6	145.0	175	3.62	2.770	15.50	0
Maserati Bora1	15.0	8	301.0	335	3.54	3.570	14.60	0
Volvo 142E1	21.4	4	121.0	109	4.11	2.780	18.60	1

When you do this, you'll probably want to make the header repeat on new pages. Do this with the `kable_styling` function from `kableExtra`:

```
a_long_table %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = "repeat_header")
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

(continued)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2
Mazda RX41	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag1	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 7101	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive1	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout1	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant1	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 3601	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D1	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 2301	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 2801	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C1	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE1	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL1	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC1	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood1	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental1	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial1	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4

(continued)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Fiat 1281	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic1	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla1	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona1	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger1	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin1	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z281	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird1	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-91	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-21	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa1	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L1	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino1	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora1	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E1	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Unfortunately, we cannot use the `scale_down` option with a `longtable`. So if a `longtable` is too wide, you can either manually adjust the font size, or show the table in landscape layout. To adjust the font size, use kableExtra's `font_size` option:

```
a_long_table %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(font_size = 9, latex_options = "repeat_header")
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4

(continued)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2
Mazda RX41	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag1	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 7101	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive1	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout1	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant1	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 3601	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D1	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 2301	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 2801	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C1	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE1	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL1	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC1	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood1	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental1	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial1	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 1281	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic1	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla1	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona1	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger1	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin1	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z281	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird1	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-91	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-21	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa1	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L1	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino1	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora1	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E1	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

To put the table in landscape mode, use kableExtra's `landscape` function:

```
a_long_table %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = "repeat_header") %>%
  landscape()
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2

(continued)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2
Mazda RX41	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag1	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 7101	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive1	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout1	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant1	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 3601	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D1	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 2301	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 2801	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C1	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE1	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL1	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC1	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood1	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental1	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial1	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4

(continued)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Fiat 1281	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic1	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla1	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona1	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger1	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin1	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z281	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird1	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-91	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-21	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa1	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L1	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino1	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora1	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E1	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

5.1.4 Max power: manually adjust the raw LaTeX output

For total flexibility, you can adjust the raw LaTeX output from `kable`/`kableExtra` that generates the table. Let us consider how we would do this for the example of adjusting the font size if our table is too wide: Latex has a bunch of standard commands that set an approximate font size, as shown below in Figure 5.1.

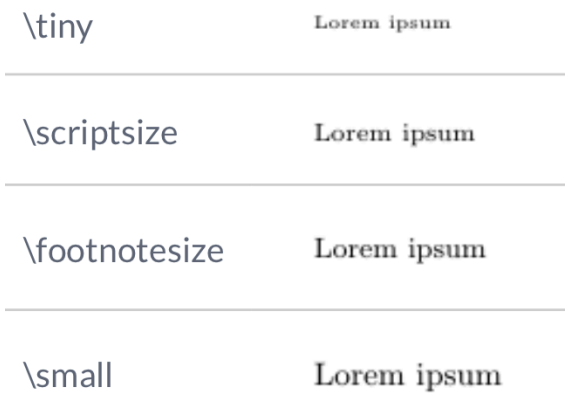

Figure 5.1: Font sizes in LaTeX

You could use these to manually adjust the font size in your longtable in two steps:

1. Wrap the longtable environment in, e.g., a `scriptsize` environment, by doing a string replacement in the output from `kable`/`kableExtra`
2. Add the attributes that make R Markdown understand that the table is a table (it seems R drops these when we do the string replacement)

```
our_adjusted_table <- a_long_table %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = "repeat_header") %>%
  # wrap the longtable in a tiny environment
  str_replace('\\\\begin\\{longtable\\}',
              '\\\\begin\\{scriptsize\\}\n\\\\begin\\{longtable\\}')
               ↪  %>%
  str_replace('\\\\end\\{longtable\\}',
              '\\\\end\\{longtable\\}\n\\\\end\\{scriptsize\\}')

#add attributes to make R Markdown treat this as a kable LaTeX
↪  table again
```

```
our_adjusted_table %>%
  structure(format = "latex", class = "knitr_kable")
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2
Mazda RX41	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag1	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 7101	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive1	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout1	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant1	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 3601	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D1	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 2301	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 2801	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C1	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE1	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL1	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC1	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood1	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental1	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial1	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 1281	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic1	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla1	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona1	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger1	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin1	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2

(continued)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Camaro Z281	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird1	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-91	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-21	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa1	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L1	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino1	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora1	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E1	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

5.2 Executable code chunks

The magic of R Markdown is that we can add executable code within our document to make it dynamic.

We do this either as *code chunks* (generally used for loading libraries and data, performing calculations, and adding images, plots, and tables), or *inline code* (generally used for dynamically reporting results within our text).

The syntax of a code chunk is shown in Figure 5.2.

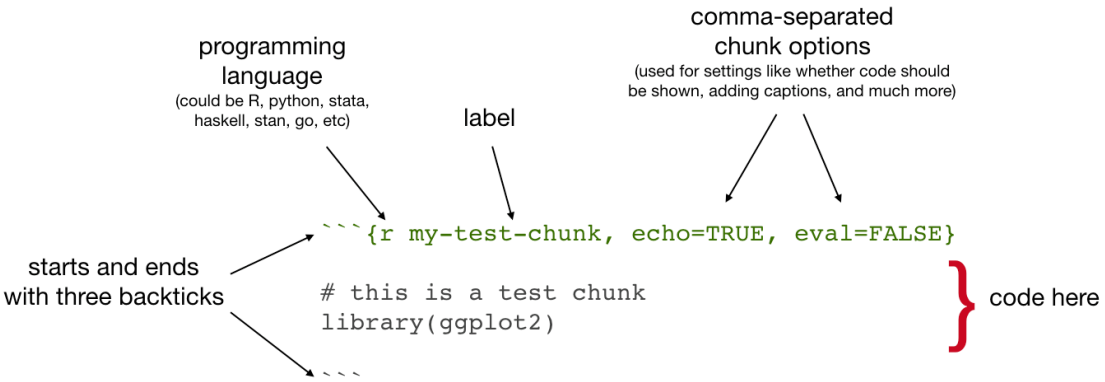

Figure 5.2: Code chunk syntax

Common chunk options include (see e.g. bookdown.org):

- `echo`: whether or not to display code in knitted output
- `eval`: whether or to to run the code in the chunk when knitting
- `include`: whether to include anything from the from a code chunk in the output document

- `fig.cap`: figure caption
- `fig.scap`: short figure caption, which will be used in the 'List of Figures' in the PDF front matter

IMPORTANT: Do *not* use underscoores in your chunk labels - if you do, you are likely to get an error in PDF output saying something like "! Package caption Error: \caption outside float".

5.2.1 Setup chunks - setup, images, plots

An R Markdown document usually begins with a chunk that is used to **load libraries**, and to **set default chunk options** with `knitr::opts_chunk$set`.

In your thesis, this will probably happen in **index.Rmd** and/or as opening chunks in each of your chapters.

5.2.2 Including images

Code chunks are also used for including images, with `include_graphics` from the `knitr` package, as in Figure 5.3

```
knitr::include_graphics("figures/sample-content/beltcrest.png")
```

Useful chunk options for figures include:

- `out.width` (use with a percentage) for setting the image size
- if you've got an image that gets waaay to big in your output, it will be constrained to the page width by setting `out.width = "100%"`

Figure rotation

You can use the chunk option `out.extra` to rotate images.

The syntax is different for LaTeX and HTML, so for ease we might start by assigning the right string to a variable that depends on the format you're outputting to:

Figure 5.3: Oxford logo

```
if (knitr::is_latex_output()){
  rotate180 <- "angle=180"
} else {
  rotate180 <- "style='transform:rotate(180deg);'"
}
```

Then you can reference that variable as the value of `out.extra` to rotate images, as in Figure 5.4.

5.2.3 Including plots

Similarly, code chunks are used for including dynamically generated plots. You use ordinary code in R or other languages - Figure 5.5 shows a plot of the `cars` dataset of stopping distances for cars at various speeds (this dataset is built in to **R**).

```
cars %>%
  ggplot() +
    aes(x = speed, y = dist) +
    geom_point()
```

Under the hood, plots are included in your document in the same way as images

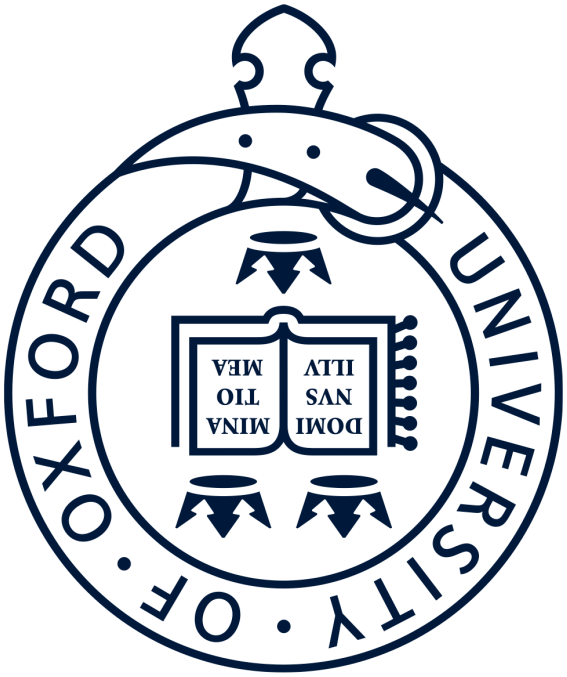

Figure 5.4: Oxford logo, rotated

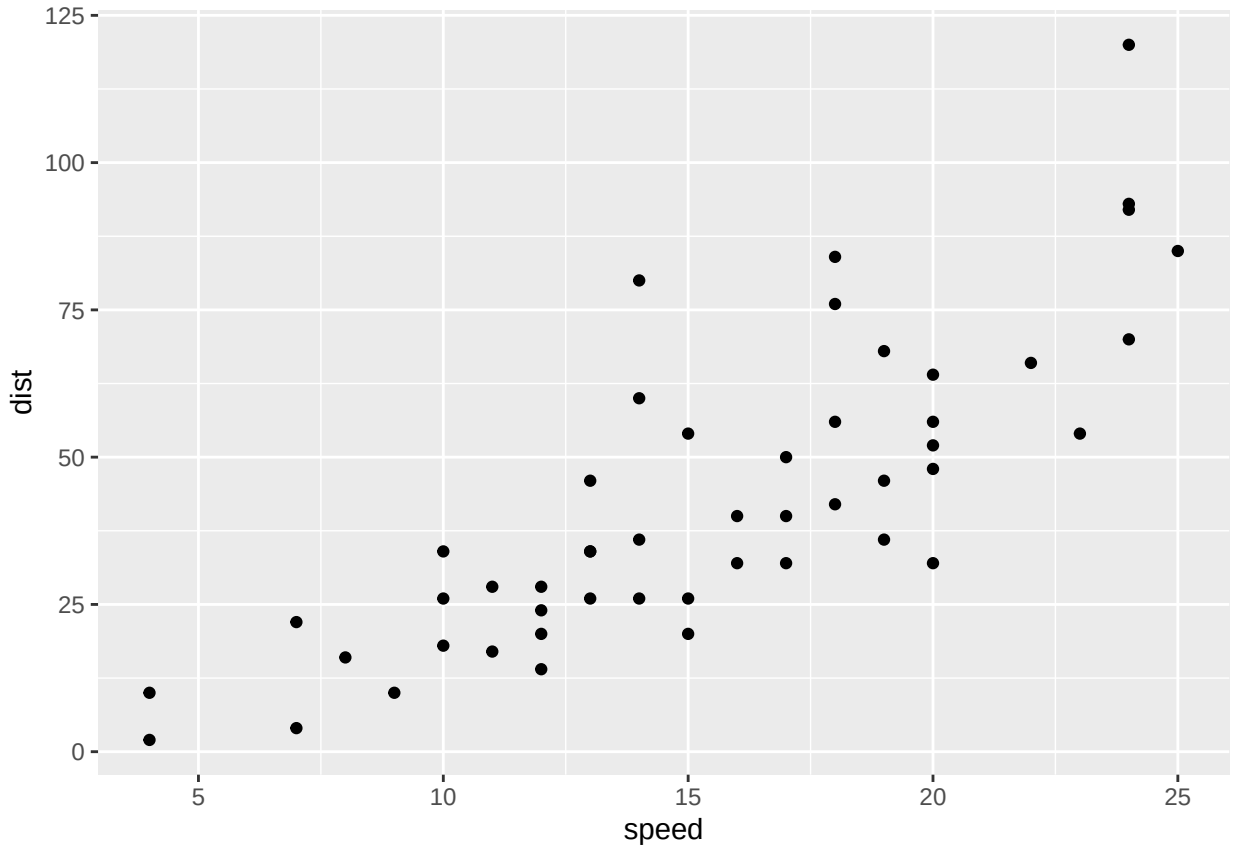

Figure 5.5: A ggplot of car stuff

Table 5.6: A knitr kable table

speed	dist
4	2
4	10
7	4
7	22
8	16
9	10

- when you build the book or knit a chapter, the plot is automatically generated from your code, saved as an image, then included into the output document.

5.2.4 Including tables

Tables are usually included with the `kable` function from the `knitr` package.

Table 5.6 shows the first rows of that cars data - read in your own data, then use this approach to automatically generate tables.

```
cars %>%
  head() %>%
  knitr::kable(caption = "A knitr kable table")
```

- Gotcha: when using `kable`, captions are set inside the `kable` function
- The `kable` package is often used with the `kableExtra` package

5.2.5 Control positioning

One thing that may be annoying is the way *R Markdown* handles "floats" like tables and figures. In your PDF output, LaTeX will try to find the best place to put your object based on the text around it and until you're really, truly done writing you should just leave it where it lies.

In general, you should allow LaTeX to do this, but if you really *really* need a figure to be positioned where you put in the document, then you can make LaTeX attempt to do this with the chunk option `fig.pos="H"`, as in Figure 5.6:

```
knitr::include_graphics("figures/sample-content/beltcrest.png")
```

Figure 5.6: An Oxford logo that LaTeX will try to place at this position in the text

As anyone who has tried to manually play around with the placement of figures in a Word document knows, this can have lots of side effects with extra spacing on other pages, etc. Therefore, it is not generally a good idea to do this - only do it when you really need to ensure that an image follows directly under text where you refer to it (in this document, I needed to do this for Figure 5.1 in section 5.1.4). For more details, read the relevant section of the R Markdown Cookbook.

5.3 Executable inline code

'Inline code' simply means inclusion of code inside text. The syntax for doing this is `` `r R_CODE` `` For example, `` `r 4 + 4` `` will output 8 in your text.

You will usually use this in parts of your thesis where you report results - read in data or results in a code chunk, store things you want to report in a variable, then insert the value of that variable in your text. For example, we might assign the number of rows in the `cars` dataset to a variable:

```
num_car_observations <- nrow(cars)
```

We might then write:

"In the `cars` dataset, we have `` `r num_car_observations` `` observations."

Which would output:

"In the `cars` dataset, we have 50 observations."

5.4 Executable code in other languages than R

If you want to use other languages than R, such as Python, Julia C++, or SQL, see the relevant section of the *R Markdown Cookbook*

There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

— Charles Darwin (Darwin, 1859)

6 Conclusions

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

Conclusion 1

The need for a better biological interpretation of multi-omics integrative methods let us to consider the inclusion of biological information during (not after) the analysis process

Conclusion 2

We propose a method focused on the expansion of the starting omics datasets, by adding new annotation-derived features to those matrices, before applying the integrative analysis

Conclusion 3

This approach allows the inclusion of relevant information from the main biological annotation tools, as well as any custom annotation, combined with the use our preferred Dimension Reduction techniques

Conclusion 4

We have implemented a pipeline for reproducible and easy-to-use execution, that facilitates the control of each step, the visualization of results and their reporting to PDF/HTML formats.

Text from Template: More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

This paragraph, by contrast, *will* be indented as it should because it is not the first one after the 'More info' heading. All hail LaTeX. (If you're reading the HTML version, you won't see any indentation - have a look at the PDF version to understand what in the earth this section is babbling on about).

Appendices

A

The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

```
library(tidyverse)
knitr::include_graphics("figures/sample-content/chunk-parts.png")
```

And here's another one from the same chapter, i.e. Chapter 5.2:

```
knitr::include_graphics("figures/sample-content/beltcrest.png")
```

B

The Second Appendix, for Fun

References

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. John Murray.

Lottridge, D., Marschner, E., Wang, E., Romanovsky, M., & Nass, C. (2012). Browser design impacts multitasking. *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting*. https://doi.org/10.1177/1071181312561289

Lyngs, U. (2019). Oxforddown: An oxford university thesis template for r markdown. In *GitHub repository*. https://github.com/ulyngs/oxforddown; GitHub. https://doi.org/10.5281/zenodo.3484682

Mill, J. S. (1965 [1843]). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Longmans.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Wu, T. (2016). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. Knopf Publishing Group.