BIO782P Statistics for Bioinformaticians Assignment 2021

### Assessment 1

# Introduction

You have been allocated two unique datasets to analyse. To find yours, look at the table at the end of this document. Having read the description of the way the data were collected, you need to import each dataset into R. Remember to check the datasets when you import them for things like proper allocation of each variable as a factor or a numeric variable etc, and to carry out proper exploratory data analysis. Once you're happy that you've got the data into a condition where it can be analysed, fit a model and use that to try to answer the questions given below. For each dataset, your writeup should consist of a properly annotated R script that will allow me to exactly replicate your analysis, and the results section of a paper describing your findings. The latter should be no more than 800 words for exercises 1 and 2 combined (**and can be a lot less...)** and should not contain any unnecessary figures or tables: I would recommend no more than two figures. If in doubt have a look at some papers in (for example) *Proceedings of the Royal Society B* or *Molecular Ecology* to see how it's done. Please don’t paste output from R directly into your report, by which I mean don’t just paste a summary table or an ANOVA table in, not that you shouldn’t use graphics from R. **You must also include a copy of your datasets inside your submission folder**.

# Probiotics and synbiotics

Complex communities of microbiota colonise the human gastrointestinal tract and play an important part in human health. The intestinal microbiota is essential for immune function, digestion, and metabolic processes. Examples of immune function by the gastrointestinal microbiota include the production of antimicrobial peptides and bacteriocins.

When antibiotics are administered to treat an infection, this often has unintended negative consequences for commensal or beneficial microorganisms in the gastrointestinal system (perturbation of the gastrointestinal microbiota is known as “dysbiosis”). The functions of the gastrointestinal microbiota are particularly affected in elderly patients. Some recent work, however, suggests that the administration of probiotics can help recover human gastrointestinal microbiota. A number of different bacterial strains have been investigated for inclusion in probiotics, and multispecies probiotics are being developed where bacterial strains have an additive or synergistic effect on each other. The effect of probiotics can be enhanced by adding non-digestible food material (such as high fibre) to the diet. A treatment consisting of a combination of probiotics and dietary fibre is known as a synbiotic.

In Investigation 1 (dataset 1), 80 hospitalised patients (age range 70-81 years) suffering from gastrointestinal infectionswere treated with antibiotics. Patients ranged in the severity of symptoms before treatment, which was recorded as the number of days with diarrhoea and vomiting pre-treatment. 40 patients were administered with a multispecies probiotic treatment (consisting of combined bacterial strains *Lactobacillus rhamnosus* and *Bifidobacterium lactis*). The other patients received a placebo. After five days of treatment, five intestinal mucus samples were taken per patient during an intestinal endoscopy to monitor recovery of the gastrointestinal microbiota. The amount of antimicrobial peptides within the mucus was measured using flow cytometry and an average value (mg/ml) was calculated per patient.

1. Explore how the severity of infection is related to the production of antimicrobial peptides by the human gastrointestinal microbiota - is this relationship dependent on the administration of probiotics and if so, how?

In Investigation 2 (dataset 2), a follow-up study was conducted to measure the effects of synbiotics on the diversity of gastrointestinal microbiota following treatment with antibiotics after gastrointestinal infection. A fully factorial design was implemented to test the effect of probiotics in combination with high dietary fibre using 120 patients. After ten days on the treatment, two stool samples were taken per patient with 6 replicates each. Deep amplicon sequencing of the V4 region of the 16S marker gene was performed using Illumina MiSeq 350bp technology. Shannon diversity was calculated from Amplicon Sequencing Variants and an average value was calculated per patient.

1. What are the effects of probiotics and diet on Shannon diversity of recovered gastrointestinal microbiota?

# **ADVICE FOR ASSESSMENT 1**

### Your report should consist of two things: firstly an annotated R script so that I can replicate your analysis, and secondly the results section of a paper describing your findings, with suitable figures and summaries of your analysis.

You will be graded on:

* A clear, functioning R script which is adequately annotated.
* Choice and implementation of analysis appropriate to the datasets.
* A complete and accurate description of the results.
* Selection and presentation of figures.
* Appropriate formatting and presentation of the report.

**You must also include a copy of your datasets inside your submission folder.**

# General feedback from previous years

This is a feedback document which was written for the class in a previous year. It might be helpful.

Overall, I was impressed with the standard of most of these reports. Most of you managed to get the analyses more or less right, and most of you produced good-quality scripts. None of them had too much annotation - remember that when you’re annotating a script you’re really writing a guide to yourself describing what it does. Think about the information that might be useful if you come back to the analysis a year or two in the future.

Regarding the “results section” of the reports, these were a lot more variable. Many of you put too much analysis into these, and in particular included material that would not normally go in a journal results section. Preliminary and exploratory analysis, diagnostic plots and the like are not generally included in journal results sections. Many of you had redundant graphs – you shouldn’t show the reader the same set of data twice. A common problem was too much focus on the statistics and not enough on what the statistical results mean in terms of the biology of the system, or in terms of effect sizes: don’t just tell me that there’s an effect, tell me how big the effect is, and put it in meaningful biological terms.

Figure captions are something that most of you need to work on. When you’re writing a figure caption, it’s a good idea to try to write it so that a casual reader who is skimming through the paper can look at the figure, have a look at the caption and have at least a rough idea about what the figure is showing. That doesn’t mean that each figure should have the whole methods section reproduced, but you can include a sentence or two that gives the casual reader a basic idea of what’s going on.

Other common problems included:

* Including multiple tests of the same thing
* Carrying out tests for normality on data prior to analysis
* Mixing p-values and AIC as criteria for model selection - the philosophy behind these is very different and you should use one or the other but not both.
* Giving significance levels for main effects when they are also included in higher-order interaction terms
* Including graphs showing no effect. There are circumstances when you might want to do this, if you’re reporting a negative result and you think it’s necessary to make a point about how little relationship there is, but in general we wouldn’t put this sort of thing in.
* Including code, function names etc. from R. Results sections from journals wouldn’t usually include this kind of material unless you’re describing an esoteric analysis that the readers will not be familiar with, in which case you might say “We fitted a generalised additive model to the data using the gam() function as implemented in the mcgv package (Wood 2014)” but usually you would just tell the reader the type of analysis used.
* Referring to “non-significant” results as “insignificant”. Don’t do this - have a think about why.
* No references! Not a single person put a reference in their results section. You don’t necessarily need references in a results section but you need them for any software you use, including R packages that aren’t part of the base R installation and for any methods which are not sufficiently common as to be considered standard. As an example, you wouldn’t use a reference for a standard linear model but you might for something like an MCMC GLMM.