Big Data Biology (BIO00047I)
Report Guide and Marking Scheme

Updated: 18/1/2022

Basic information

- **This assessment is due before:** Thursday 21nd April 2022 at 11am.
- **Word limit and document formatting:**
    - The report and the appendix R code should be submitted as **one pdf document**.
    - The maximum length is 1500 words.
    - **The 1500 word limit includes:** title, abstract, introduction, main text (results & discussion), conclusion.
    - **It does not include:** plot legends, the R code in the supplementary methods, the references.
- **Choose one data set to work on for your report.**
    - The data sets are described here. https://www-users.york.ac.uk/~dj757/BIO00047I/misc/BIO00047I-data-description-2021.pdf
- **Reports must contain some new analysis of the data.** Reports that merely replicate the commands from the workshops will be penalised. For example, Brassica analysis must use the full data set, not the sample data set provided in the workshops.

Guidance about data sets analysis

Below we provide some hints an ideas about how to make new and interesting analysis of each of the data sets. Remember – you only need to analyse one data set, and you do not need to follow all the idea listed here. Any one of them, carefully analysed could make an excellent report.

Brassica data

If you choose this data, be sure to analyse the full data set, not the sample data. The full data set is available here: http://www-users.york.ac.uk/~ah1309/BigData/data/OSR101_RPKM.txt

With this data, you should seek to answer some of these questions:

1. Are there any regions of the genome that are significantly associated with glucosinolate content of the seed? Where are they?
2. How many gene expression markers pass the multiple test correction threshold?
3. Can any candidate genes for this trait be identified?
4. Do they have a positive or negative correlation with glucosinolate content?
5. What do these correlations mean?

Yeast data

The yeast data is simpler to process (simpler R commands), but requires a little more independent thought. Merely replicating the analysis you performed in the workshop is not sufficient.
There are two type of data columns:

- Quantitative data (like gene expression level in RPKM).
- Categorical data (puts genes in to categories, like essential/non-essential)

Many different analyses of this data are possible, but we suggest that you analyse this data in one of two ways:

A. Choose a quantitative data column and explore it. Find out what other data columns it is correlated with, and if it is different in different categorical subsets of the data (like essential genes, or genes that are present in the nucleus, for example).

B. Choose a category of genes (like Golgi, Mitochondrion, Nuclear_dots, Nuclear_envelope, Nucleolus, Nucleus, Vacuole, essential, protein_coding) and describe how they differ.

Here are some examples. Addressing any **one** of these example is enough for your report. Other analysis are possible, including downloading other data from Pombase/Angeli, to include in your report.

1. The data column mRNA.stabilities describes how stable particular transcripts are. What other quantitative data is correlated with mRNA stabilities? Are particular categories of genes more or less stable? For example, are mitochondrial transcripts more/less stable that cytoplasmic transcripts? Why might this be? What does this tell you about the cell?
2. The quantitative data column conservation.phyloP describes how rapidly each gene has changed over evolutionary time. Is conservation correlated with gene expression? With gene knockout fitness? (solid.media.KO.fitness). Are particular categories of genes more or less conserved? Why might this be? What does this tell you about the cell?
3. As for #1 or #2, but applied to another quantitative variable.
4. The categorical data column Golgi describes which proteins are present in the nucleus. How many are there? Are they more highly expressed than proteins that are not found in the nucleus in the? Are they more highly conserved? Are there other quantities columns that differ? Do they have more introns (NumberIntrons). Why might this be? What does this tell you about the cell?
5. As for #4, but applied to another categorical value. For example, Nucleus, Mitochondrion, Nuclear_dots, Nuclear_envelop, essential, protein_coding. Why might this be? What does this tell you about the cell?

Fungal metagenomic data

Here are a list of tangents that you can take when performing the analysis on the fungal ecology data set. You won't have to do *all* of these (or any of these), but hopefully this will spark some ideas for personalising your final report.

In the workshops, we compared the distribution of phyla between the environmental samples. Here are some questions you could examine:

- How does the distribution of order/class/family/genus/species vary across samples?
- How does the distribution of phyla vary by latitude? Longitude?

For those of you that are interested in **evolution.**
In workshop 2, you learned how to export the DNA sequences as a fasta file, using 'write.fasta'. You can use a fasta file to draw a phylogenetic (evolutionary) tree using this website: https://www.ebi.ac.uk/Tools/msa/clustalo/

Try drawing evolutionary trees for each phyla—what does this tell you about the diversity of the OTUs that come from each phyla?

For those of you that are interested in **ecology.**
In workshop 4, we learned about the Simpson's Index of Diversity. The function 'diversity' can also calculate the 'Shannon index' (index="shannon"). What is the difference between the Simpson's index of diversity and the Shannon index? Is there a significant difference in the Shannon index of diversity across the different fields?

For those of you that are really interested in **ecology** or **mathematical biology.**
In the first workshop, we noticed that some samples had many more observed OTUs than others. If too few OTUs were sequenced, then we may not have identified all (or even most!) of the fungi in the sample. We can use a technique called rarefaction analysis to compare the species richness between samples where different numbers of OTUs were sampled: https://en.wikipedia.org/wiki/Rarefaction_(ecology). As a hint, this line of R code draws rarefaction curves:

```
rarecurve(t(hedgerow_counts_by_phylum))
```

The rarecurve function is part of the vegan package. To install this package, use this command

```
install.packages("vegan")
```

For those of you that are interested in **learning extra programming**
Here is a snippet of code that uses nested loops, and some functions that you may not have seen before (cor, heatmap, cutree, as.hclust).
Try to figure out (i) what does this code do? (you can 'google' the functions you don't know!) (ii) what biological questions is it trying to answer?

This code produces two vectors (clusts_by_OTU and clusts_by_sample) which could be used in further analysis to answer the following questions:

- Do environmental samples that have similar OTUs come from the same fields?
- Which species tend to be found together across the environmental samples? Do they come from the same phyla or different phyla?

```
a=apply(ITS_counts[,1:47], 1, function(i){apply(ITS_counts[,1:47], 1,
function(j){cor(i,j)})})
a_res=heatmap(a, keep.dendro=TRUE)

b=apply(ITS_counts[,1:47], 2, function(i){apply(ITS_counts[,1:47], 2,
function(j){cor(i,j)})})

b_res=heatmap(b, keep.dendro=TRUE)

clusts_by_OTU=cutree(as.hclust(a_res$Rowv), k=8)

heatmap(a, RowSideColors = rainbow(8)[clusts_by_OTU])

clusts_by_sample=cutree(as.hclust(b_res$Rowv), k=10)

heatmap(b, RowSideColors = rainbow(11)[clusts_by_sample])
```

Pseudosuchia macroevolutionary data

If you choose to analyse the Pseudosuchia data for your report you should be thinking about the following broad questions when you write your report:

- Has climate change driven speciation in Pseudosuchia over macroevolutionary time scales? If so, how?
- Why do we see this interaction between climate change and diversification? What are the mechanisms driving any patterns we see?
- Can we make generalisations about climate change and biodiversity change from these results? Why?
- What does it mean (if anything) for biodiversity experiencing climate change today?

Just following the Pseudosuchia workshops should give you some sensible results for initial interpretation, however, your results will be quite limited and you will be unable to make comparisons across taxa and inferences about ecology and climate change. Therefore, your report should contain some new analysis and there are a number of options you can take. Below is a list of some suggestions, you do not need to try them all and if you have other ideas you should feel free to explore those as well/instead. I recommend at least analysing the marine data for you to be able to fully address the broad questions listed above but you may also choose to explore some of the other suggestions:

- Analyse the marine data as well as the terrestrial. Are they different? If so, why?
- Explore what happens when you change the number of correlations carried out or if you change the starting seed. Does it affect the robustness of your results?
- Carry out correlations for extinction as well as speciation. Can we say anything about the drivers of extinction as well as the drivers of speciation? Are they different? Hint: in BAMMtools speciation is expressed as lambda and extinction is expressed as mu.

What if you carry out the correlations for the whole tree instead of partitioning by habitat? How does it affect your results? What does this tell you about the way in which we should approach macroevolutionary questions?

A little more ambitious for those who want to learn more R: using BAMMtools can you explore rate shifts on the phylogeny? When did these happen and what do they mean? Are they related to climate change or some other driver? Hint: You can look up the online documentation to find examples of what code to use to carry out additional macroevolutionary analyses (BAMM documentation)

This one is a bit more ambitious again but can you explore lineages through time using the ape package in R? If you extract the numbers of lineages through time you can correlate them against the speciation (and extinction) rate time series. This tells you whether the number of lineages drove speciation and/or extinction and can be an indication of whether clade competition played a role in shaping diversity. Tip: ape is a requirement of BAMMtools so you do not need to install any new packages for this option though you will want to look up the documentation (ape documentation).

Remember that these are just suggestions, you can choose to focus on whatever aspect you find interesting for your report. **It is not necessary to carry out all of these to get a good mark!**

Additional information

There is some useful background on pseudosuchian evolution and climate change in this paper (Mannion *et al*., 2015) that looked at similar questions using a different methodology whereas this paper (Davis *et al*., 2016) used the phylogenetic methods you will be using in this workshop to address a similar question for a different group of organisms. The data set description document and the workshops also contain further resources that you might find useful.

References

ape documentation: https://cran.r-project.org/web/packages/ape/ape.pdf

BAMM documentation: http://bamm-project.org/documentation.html

General guidance

Your report must have these sections:

1. Title
2. Abstract. A concise summary of the background and main results, at most150 words (this does count towards the 1500 words).
3. Main text. This section should start with a brief (one-paragraph) introduction, and then describe results and discussion.
   The results/discussion section should have titled sub-sections.
4. Conclusion.
5. References.
6. Supplementary methods. This section must contain all the R code used for the analysis. Code must be commented (this section does not count towards the 1500 words).

What the report should contain.
This section describes how we will grade your report, and what each section should contain. We show the marking scheme we will use to generate marks, so you can see exactly how they will be allocated.

Title
The title should describe your main question and/or your main result and what species you are using. We encourage interesting and enticing titles.

Abstract [maximum 150 words]
Your abstract should be concise summary of the background and main results. The best method to write an abstract it to include sentences that contain; the broad background to the topic, the narrow (specific) background to the topic, the main question/analysis that you are concerned with, what analysis you did, what result(s) you found, and finally what the biological implications of this analysis are.

Main text
This is the most important part of the report.

This section should start with a brief (one-paragraph) introduction, and then describe results and discussion.

The 'letter' style manuscripts differ from article style, in that the main text contains both results *and* discussion, blended together. To achieve this, each paragraph or section should contain:

- a question, query or hypothesis
- a test or analysis you performed
- the result (described verbally and/or in plots/tables)
- a brief discussion/comment (1-2 sentences) discussing the biological implications of the result, similar observations, mentioning any caveats, and/or new questions, etc.

e.g.

[hypothesis]* I hypothesised that essential genes would be more highly expressed than non-essential genes, because they are involved in central cellular processes for which abundant proteins would be required. [test] To test this hypothesis, I compared the protein expression levels of essential and non-essential genes using a Wilcoxon signed rank test. [result] This analysis showed that essential genes have slightly higher protein expression levels than non-essential genes ($P = 1 \times 10^{-3}$, Figure 1). [discussion/comment]* Similar results have been observed in other species [citations], showing that this is a general trend of molecular biology. It is possible that the majority of this result is due to ribosomal proteins. [next paragraph could examine this hypothesis].

* Don’t put these red markers in your own text.

The results/discussion section should have titled sub-sections. The titles of these sections should tell the reader what the section is about.

Conclusion
The conclusion should briefly reiterate the major findings of your study.

References
You should cite in your text, the original source of all data you use, and any other articles that are relevant to the topic and your enquiries.

Supplementary methods.
This section must contain all the R code used for the analysis. Code must be commented (this section does not count towards the 1500 words).

The marking scheme is on the next page.

Section*	Marking guide.	Marks	Total
Abstract*	Defines the broad and then the narrow background to the topic.	2	**5**
	Defines the question or problem that is addressed.	1	
	Describes the analysis and results.	1	
	Makes a conclusion, or summary.	1	
Introduction*	Is a clear summary of the background to the topic.	5	**10**
	Cites relevant published articles and reviews.	5	
Main text *	Hypothesis-testing evident with result paragraphs following the pattern of: Hypothesis (or question), test, result, conclusion.	10	**15**
	Writing style is clear, well-referenced, free of spelling/grammar errors and has subheadings.	5	
Plot(s)*	Plots are well-presented and illustrate data clearly.	5	**10**
	Figure legends meaningful & precise.	5	
Data analysis*	Statistical tests used appropriately (correct tests for the data).	5	**15**
	Conclusions/interpretations are well-supported by the data analysis.	10	
Conclusion*	Briefly reiterates the results.	1	**5**
	Mentions any caveats/limitations	2	
	Concludes something meaningful.	2	
Background reading**	Uses previous published results to back up, explain and/or contrast own analysis.	5	**5**
Supplementary methods**	Includes R commands used for project.	5	**10**
	R commands are clearly commented.	5	
All sections**	Rationale for the main enquiry/question is clear and interesting.	5	**25**
	Points for biological insight and/or technical skill.	20	

* Sections that are included in the 1500-word count. We do not count text in figure legends towards the 1500-word count. ** These are marks that we assign from various sections of the report, not explicit sections of the report. So they will contribute to the word count (in the Introduction, Abstract, Main text etc).