Thesis Amendments

Stuart Lee 29 October 2020

First, I would like to thank my examiners for their constructive feedback and insights. I have incorporated most of the comments into the thesis revisions. Where changes have not been made I have described my plans to incorporate them into future work. The point by point description of changes is below: the examiners' comments are in black and my response is in red.

Professor Levi Waldron

p. 1 Introduction: it's -> its

Amended.

p. 16 'Using plyranges and the the'

Amended.

p. 34: use of `dir` for a directory path string when `dir()` is a function returning contents of a directory path could be confusing. However, this chapter is already peer-reviewed and published so I would not change it.

Amended - since I can update on F1000 at some point.

p. 43-44: Why is the genome ('hg38') set using an endomorphic function, but style ('UCSC') set using a replacement function (`seqlevelsStyle`)?

No change was made. **plyranges** does not have any genome restyling functions, hence why the replacement function was used.

p. 77: Tensorflow 'non-liner embedding' p. 77-78" 'a lot of' -> many

Amended.

p. 80: 'its bounded above' ->it's

Amended.

p. 82: 'Like, when, using' -> Like when using "While our software, liminal is able " -> While our software, liminal, is able

Amended.

Section 5.3.1 'Finding Gestalt: focus and context' - this is probably a clever metaphor, but I didn't understand it.

This is a reference to Buja, Cook & Swayne (1996) paper mentioned in the paragraph above. I have amended section 5.3 to add what I mean by 'Find Gestalt' which refers to identifying patterns in visual forms.

p. 84: 'Whether, points are near, or far' -> remove both commas

Amended.

Figure 5.2 video https://player.vimeo.com/video/43963590 does not exist. The URL given in Table 5.1 is different and correct.

Amended.

p. 98: 'We have shown in the case studies, that' -> no comma

Amended.

Assistant Professor Stephanie C. Hicks

Addressing General Comments

On Scalability

There are two main reasons to not talk about performance benchmarking. First, the focus has been on the development of programming and user interfaces and performance benchmarking would distract from that. Second, my implementation of **plyranges** is deeply entwined with the underlying core Bioconductor infrastructure, especially the data structures and methods found in the **IRanges** and **GenomicRanges** packages. This would result in any benchmarking of **plyranges** turning into benchmarking of those packages which has been done elsewhere.

Regarding out of memory data structures like `DelayedArray`, it is possible to incorporate them as columns to a `GRanges` object and then the full **plyranges** API would be available. Currently, there appear to be no plans in Bioconductor for out of memory `Ranges`, and indeed there are times where it is more appropriate to switch to a rectangular data layout for computation, as is reflected in the workflow described in chapter 3 of the thesis. The development of a scalable out of memory interface for genomic data analysis will require the transformation between data measured along the genome (covered by the **plyranges** API) and the rectangular layouts that enable out of disk computing like `SummarizedExperiment`. This is an exciting research area that I would like to explore in the currently under development **plyexperiment**.

On alignment vs. pseudoalignment

I have amended the introduction to include a note about alignment. It is out of scope to go into the finer details of when to use alignment or pseudo-alignment methods to obtain counts for RNA-seq data. I have included results from both approaches to highlight that **plyranges** is agnostic to the method used and is flexible for many types of data.

Chapter 1 comments

As you introduced a 'biological data science workflow' in this chapter and did a great job of walking the read through the wrangling, combining and visualizing components of the workflow, I was hoping you would touch on tools for communication too. I understand this is outside the scope of the thesis, but I still think it would be nice to at least mention work others have done in this area.

This work focuses on exploratory tools, sometimes people confuse graphics used for exploratory work and graphics used for communication. It is not in the scope of the thesis to include communication graphics, so I have not made changes. The distinction is already provided in Figure 1.1.

Pg 1: I thought it was strange that the words 'figure' in e.g. 'figure 1.1' were not capitalized. Would recommend changing to 'Figure 1.1' throughout. Pg 1: Recommend capitalizing the word 'chapter' when referring to specific chapters e.g. 'chapter 2'.

Amended. I have made this consistent throughout the thesis now.

Pg 2: Figure 1.1 has been modified to include 'combine', but the caption does not reflect what is meant by that. Would recommend modifying caption to reflect importance of combining experimental assays together to gain biological insight.

Amended.

Pg 3: you introduce 'range-based genomics data', but it is not clear what is meant by that. More explanation would be helpful.

Amended. It now refers to data measured along the genome.

Pg 4: In Section 1.2, could you add a few more references on existing methods for integration of genomic data and then re-emphasize the point that however existing methods have not thought much about the interoperability between the tidyverse and Bioconductor approaches?

I have expanded this section to reflect these comments.

Chapter 2 comments

This chapter has already been published and the comments pertain to the use and learning of the **plyranges** package. They are best addressed through the **plyranges** documentation and vignette. Consequently, I have amended the thesis to include the package vignette as an appendix.

Chapter 3 comments

Almost all of Chapter 2 discusses the use of GRanges objects. As most of Chapter 3 discusses the use of the SummarizedExperiment object, could you add a motivating figure at the start explaining the relationship between a SummarizedExperiment and GRanges object (i.e. rowRanges)? I do not think this is obvious to a reader who may not very familiar with these objects. However, I think it is an important aspect to explain in this chapter.

I have amended this by adding a figure which explains what a `SummarizedExperiment` is and how it is related to a `GRanges`.

Pg 31: Figure 3.1 is fairly low resolution. Is it possible to make a higher resolution version?

Amended.

Pg 31: Figure 3.1 demonstrates how you use plyranges to integrate the results from DE and DA (i.e. finding overlaps). Would it also make sense to the integration step prior to modelling? I am thinking about this as a way to reduce the amount of tests performed for DE and DA.

No changes made. It would make sense but it is out of scope for this paper.

Pg 32: The fluentGenomics package needs a reference

Amended.

Pg 39: Could you add 1-2 sentences (and maybe a reference) about what is a MA plot and why it's useful for visualizing DE results?

Amended.

Pg 40: I think it would be useful if you could comment on the results() function a bit more to explain what is happening under the hood there. It's a pretty critical step for everything downstream after identifying the DE genes in preparation for the integration step. As it stands, I felt like it was kind of glossed over.

Amended.

Pg 45: More motivation could be provided on why you want to generate random samples of size equal to the number of DE-genes from the other_genes with your larger question in mind of 'how many DA peaks are near DE genes relative to'other" non-DE genes¿'. It is stated on pg 46"we minimize the variance on the enrichment statistics induced by the sampling process", but it might be good to explain this in greater detail at a higher level.

I have added a few more sentences to this section which describes the motivation for selecting sets of the same size between non-DE and DE genes, and why we want to do sub-sampling.

Chapter 4 comments

Much of this chapter depends on understanding what is meant by a coverage trace plot. You do have examples (e.g. Figures 4.4-4.7), but this is at the very end of the chapter. It would be good to provide a motivating figure be added at the start to explain what is meant by 'coverage' in this context?

Pg 63: Introduce what you mean by coverage traces and provide more motivation why you would want this over just numerical summaries of the data.

Amended. There is a now a motivating figure that explains how to interpret coverage scores.

Pg 66: You state 'resulting GRanges in memory can be compressed using run-length encoding for any categorical variable'. Could you give a numeric example of the amount of compression that can be obtained with RLEs? Also, maybe add a brief sentence explaining what are RLEs.

Amended.

Pg 68: Can you comment on why the GFF files from RefSeq were used?

I have amended this to comment that RefSeq is standard for zebrafish unlike mouse or human.

Pg 68: How did you handle reads that map over boundaries for introns and exons? or do you only keep the ones that are entirely inside an intron or exon?

I have amended this. Reads that are split are across exon / intron boundaries are counted towards both categories.

Pg 69: In the facets, you used a short descriptor, but could you write out what is each facet label is in the legend?

No changes made. The facet labels need to be short and the details are provided in the figure captions and main text.

Chapter 5 comments

One question I had as I was reading throughout this chapter, was what is your recommendation for users who want to calculate distances? can the liminal framework be used to better approximate distances instead of users who perform clustering directly on top of the tSNE or UMAP spaces? While is not ideal, and it is recommended clustering should performed in the PC space, it happens all the time. Another way of framing this question is can your framework be used to improve distance estimates in the tour with a NLDR method as you note it preserves global structure?

This is a really interesting question. I have amended the clustering case study to provide more details on performing cluster verification. The suggestions for using the tour to calibrate distances obtained from an NLDR method (or the other way around) is something that will be incorporated into future work.

Pg 77: 'non-liner' should be 'non-linear'

Amended.

Pg 78: 'produces' should be 'produce'

Amended.

Pg 78: Could you add a reference for PCA?

Amended.

Pg 80 (also in 81): The way 'tSNE' was written in 'this essentially turns t-SNE' looks very different than previously written stylistically. Could this be converted to the way you were originally referring to it?

Amended. I now use 't-SNE' throughout.

Pg 87: 'inn' should be 'in'

Amended.

Pg 88: Inside Figure 5.1A and B, could you have liminal label which plot is which i.e. t-SNE vs tour view. Also, could axis labels be added? Maybe also add this information to the Figure legend? – Pg 89 – same comment

When using the **liminal** package it is obvious that the embedding view is always on the left, and the tour view is on the right (it will be the view that is animated). I have annotated the static images of the interface that appear in the paper to make this clear to the reader which view is which. Regarding axis labels, they don't make sense for the tour since it's really a linear combination of multiple variables (there's an option to switch on a biplot view for the tour). It is also not necessary to include them for t-SNE because they can be misleading for the reasons mentioned in the chapter.

Pg 90: I believe there should be a period after 'tree structure data (figure 5.3)' instead of a comma

Amended.

Pg 94: Can you comment on the scRNA-seq example where you only pick first 5 PCs that explained only 20% of the variance. Or more generally, what is the trade-off on the liminal view by picking more or less PCs along with a NLDR method. Do you have recommendations for the user on how many PCs to pick for and optimal liminal view?

I have amended section 5.5.3 to provide more details about why the first five components were selected, and how to choose variables to tour.

Pg 94: Can you comment on the limits of the liminal view in terms of number of observations that one can visualize in a given tour? Have you thought about strategies for making this scale with more observations? – Ah, I see you touched on this at the end of the discussion of the capture. I think this is a really important point that would be greatly appreciated going forward. – And it was talked about in Chapter 6 in Future Work. Incorporating this more in the middle of the chapter with tables specifying what is or is not reasonable to plot in terms of the number of observations, would be helpful.

I have amended section 5.5.3 to discuss a strategy in the specific case of cluster analysis. I have also amended the discussion of the chapter with specific work that can mitigate the need to show all of the data. In the case of tasks performed for clustering it is not necessarily useful to show all points in the scatter plot when we are most interested in density and shapes. In future, I will incorporate appropriate sampling strategies or aggregations that could be done prior to producing a visualisation.

Chapter 6 comments

Pg 102: Instead of https://bioconductor.org/packages/release/bioc/html/plyranges.html, you can always use the short version https://bioconductor.org/packages/plyranges which points to the release version.

Amended.

References

Buja, A, D Cook & DF Swayne (1996). Interactive High-Dimensional Data Visualization. *J. Comput. Graph. Stat.* **5**(1), 78–99.