Exploring CD8+ T-cell Specificities using Single Cell Immune Profiling with an Outlook to Reproducible Bio Data Science

William Hagedorn-Rasmussen

Table of contents

1 Abstract

This is supposed to be an abstract.

2 Introduction

With all the advancements in recent years in regards to computational power, laboratory techniques and -equipment, the amount of produced data is ever increasing. The focus in this article will be on biological data, even though it applies to a range of fields. For a reference as to how fast the development is, the cost of sequencing the entire human genome decreased from $70M to $2000 in a matter of 15 years **Lunshof2022empty citation** (Lunshof, INSERT reference). As a more recent discovery, the field of Single Cell Transcriptomics is emerging with a high increase in number of associated publications (Angerer, INSERT reference).

When a bioinformatician works with data, it is important it is done in a reproducible fashion. Exactly as in the laboratory, it is paramount that other scientists is able to achieve the same results and to come to the same conclusions. That tends to be an issue, as the code used for working with data is not necessarily made available. If that's the case, how can you ever be sure, the no mistakes were made, or that any other problem occurred? Thereby, it is not only important to be working reproducibly, but also transparently.

A part of working with data is to find meaningful correlations, make new discoveries or support known hyptheseses. These findings should be presentable as to better share said findings. Especially when presenting to non-data stakeholders this becomes important, as they won't necessarily have the same foundation of knowledge as your peers.

We can further extent the issue of presenting to stakeholders. It is likely that they want to see certain parts of the data which wasn't included in the presentation. Then, if the data is made interactive, the stakeholders is enabled to find this themselves. Even without the need of writing a single line of code. In the end, this saves time and effort for both parties.

All of the above-mentioned issues I wish to address and take into account when working with this project. This is e.g. done by working transparently. Every pieces of code from the data preparation to the modelling is gathered in a package which is made available on GitHub. The functions written also include thorough documentations to make understanding the code easier. Through this, the data analyses are fully reproducible and can be reused to try and handle the increasing amount of data. Finally, the results of the modelling is made available through a Shiny App to make the data handling interactive.

- What is the situation now / why is this important?
  - Difficult to reproduce data analyses
  - Amount of available data grows
  - The presentation of data to non-data stakeholders
    - E.g. in industry
    - Make it interactive to reduce work load for bioinformatician and stakeholders
- What can this project do to help the issue?
- What kind of data am I working with?

  - Needed theory to understand
    * How are T-cells formed
    * What determines the TCR-sequence and what is the TCR
    * Alpha- beta sequence description
    * MHC, pMHC
    * Single Cell Immune Profiling

3 Methods

- Description of the structure of the package?
- Description of data set?
- How I cleaned
  - Describe dimensions of the data set(s)
- How I augmented
  - Describe dimensions of the data set(s)
- The integration into Shiny?

4 Results

- Mention pointers from method
  - What can the package offer
- Presentation of models
- How does the Shiny app work / what can it do

5 Discussion

6 Conclusion