Project specification **Ava Spataru**
3^{rd} Year Project u1606684
Prof. Sascha Ott Due Date: 12/10/2018

Analysis of Ovarian Cancer Single-cell RNA-seq Data

Main project idea

The main goal of this project is to analyse cell data from multiple patients suffering from ovarian cancer and classify cells as parts of different types, as well as understand how the identified types transform after treatment.

Background information

Every cell in the human body has multiple types of RNA. The RNA is a polymeric molecule (specific to each cell) that will differ based on several factors, including the function of the cell. A type of RNA (mRNA) moves from the nucleus of the cell to different parts of it or to other cells, to transmit information. Drop-seq is a strategy used to analyse these "messages" (mRNA transcripts) for each cell and identify similar transcripts as belonging to the same cell population.[1] This is based on the idea that similar transcripts indicate that the cells fulfill similar functions and therefore are similar.

After data sequencing, the result is a set of objects with multiple dimensions of information that is to be analysed. There are multiple methods that allow us to investigate highly dimensional data, such as scatterplots and heat maps, but the most important one for the scope of this project is t-SNE (t-Distributed Stochastic Neighbour Embedding). T-SNE is a fairly new approach to representing highly-dimensional data in a space with lower dimensions, that can be more easily interpreted [7]. The algorithm employs techniques of machine learning to adjust to each set of given data, while considering 2 parameters: perplexity and epsilon. Even tough the outcome of the algorithm is very useful for analysis, it can be erroneous. There is an online article [3] that explains how t-SNE plots might be misleading. For example, the perplexity parameter seems to be most reliable between 5 and 50, however a small perplexity leads to the domination of local variation and a perplexity larger than the number of points will not reach stability as often. There are a multitude of factors, such as cluster size, distance between clusters etc. with unclear behaviour, that are relevant to understand how t-SNE works.

Project motivation

There are multiple reasons why attempting to classify cells is important. Besides this being a next step towards understanding how organisms work, it could have significant impact on how treatment is conducted against cancer. If the classification is a success, on even a small set of data, the approach can then be generalised to different types of tissue.

One in five patients suffers through chemotherapy without any positive outcome, because sometimes the tumor is not affected in any way by the treatment. Currently, the doctors cannot determine whether the treatment will reduce the size of the tumor or have no significant impact. If cells could be clearly separated into types, then doctors would be able to associate similar tumors and decide if the most beneficial treatment is chemotherapy or surgery.

If cells could be clearly separated in types and subtypes, then that could be the beginning of designing medical treatment specific to each patient. If the doctors could see the types of cells and how each type responds to different medication, then they could start prescribing different medical treatments based on the combination of cell types presented.

Personal motivation

After a discussion with my supervisor, it was decided that the project is feasible, despite the lack of previous knowledge in the domain. I was really interested in learning more about the world of computational biology, which meant that this topic was a good fit. The domain is quite broad and requires a lot of prior understanding in order to conduct research in it, which is why the general area was decided on in the early stages of the project, but more specific details were discussed after reading more about the area and getting a grasp of what the research involves. This allowed me to focus on a more specific part of the research domain, that I considered more interesting for a 3rd year project.

Project goals

The primary aim of the project is to analyse the existing sets of data (from 2 patients) and successfully classify the cells as part of different types. The secondary aim of the project is to develop a deeper understanding of which genes modify during chemotherapy and how this affects the type classification.

Since the aim of the project is complex, it has been divided in sub-tasks/milestones that will help measure the success of this project. However, these could change as the project advances, to accommodate for a different, more specific approach.

The software packages that will be used to achieve these goals are: RStudio[4] and Seurat[5], which might be supplemented by others, such as Monocle[6].

A Main goals

The following table outlines some of the main goals of this project, as well as an initial approach and the number of hours assigned to achieving each goal. The approach and the estimated time might change as the project progresses.

A total number of 130 hours has been assigned to these tasks. Since the project itself was estimated to 300 hours, the rest of 170 hours will be allocated to reading, researching, report writing and improving the analysis. The estimations will probably change, because as the project progresses, different objectives may appear.

ID	Description	Approach	Hours
G1	Use t-SNE to make an initial separation of cells into types (on both data sets).	This can be done after some practice with applying t-SNE and installing all the tools needed.	20
G2	Determine what the parameters for t-SNE should be and why.	“Tweaking” parameters and observing the behaviour of the data sets on multiple runs with different values. When observing, the points made on the webpage “How to use t-SNE effectively” [3] should be considered.	5
G3	Compare similar behaviors on t-SNE between these two sets of data and other already existing findings.	Done by simple observation. The idea behind is that similar behaviors on the same algorithm might help identify common properties.	10

ID	Description	Approach	Hours
G4	Compare the groups identified in each set of data with the other ones, while emphasising on the most common genes in each.	Adjust certain parts of the data to evidentiate different genes. For example, observe how the data sets act if 100 more cells,which are mostly made up of gene A, are added.	30
G5	Identify which cells have changed from the initial sample and which cells have not. Determine if the remaining cells are unwanted or which unwanted cells did the treatment remove.	Done by observation and reasoning about the behaviour of certain genes.	5
G6	Identify genes that completely disappeared from the cell types or genes that transformed.	This step involves an in-depth analysis of each initial type and the expected behaviour of the composing genes after treatment.	5
G7	Roughly estimate the number of cells in each tissue that completely disappeared.	The aim of this stage is to estimate a "percentage of change" for separate parts of the data set and determine to what extent the treatment has equivalent effect upon different datasets.	5
G8	Try and create a mapping based on different assumptions of which gene changed to which gene. Repeat the assumptions until a stable result is obtained.	Create a set of assumptions, such as: If gene A in this cell changes to gene C, then gene B must have changed to gene D etc.	20
G9	Attempt to classify the cells by prioritising different genes.	Experiment with different ways of classification by accentuating certain properties. For example, classify based on the existence of a certain gene and the percentage of that gene within each cell.	30

B Stretch goals

The project is primarily a data analysis project, however some parts of it could be adjusted to allow for a larger amount of application development. For this reason, there is a possible stretch goal to develop a tool that would run the t-SNE multiple times in parallel, on the same set of data, while also changing parameters and flagging any noticeable behavior. The tool could also identify the run which presents the most meaningful visualisation of the data, based on the observations from the previously mentioned webpage: "How to use t-SNE effectively" [3].

The idea behind this tool is that it would reduce the number of human misinterpretations, caused by the misleading behaviour of t-SNE, while also allowing fast multiple runs and storing the information from previous runs. By expanding goal G3 (described above) there is a possibility that similar sets of data will act in similar ways on t-SNE, meaning that if the tool could also compare runs of t-SNE and identify similar ones, then it could potentially identify similar sets of data, thus indicating that the same treatment might be appropriate.

Another addition to the tool could be presenting multiple ways to analyse the same set of data. For example it could include, besides t-SNE, a number of different methods of high-dimensional data visualisation, such as the ones described in the paper "High-dimensional visualizations".[2]

This tool would provide storage of previous runs of t-SNE on the same data set, as well as give a high-level overview of the differences between runs and highlight unexpected behaviours. This will make it easier and faster for researchers to visualise and compare their data sets with previous ones, while logging a history of all previous runs.

However, the tool is only listed as a stretch goal because it has only been briefly mentioned in a supervision meeting and a more detailed discussion is needed to determine whether it would be achievable and meaningful as a part of a 3rd year project.

C Personal goals

In terms of personal development, this project should help me understand more about the world of computational biology and give me a good insight into what real-life research/data analysis projects are and how they are conducted.

Project challenges

One of the main challenges with this project is the lack of prior knowledge in the domain of computational biology, as well as little experience with data analysis. However, in the early stages of the project, the supervisor suggested a number of papers and online resources that proved very useful to filling in the knowledge gaps.

The classification of cells is a very difficult task, mostly because it involves analysing a large amount of complex data, based on various criteria. The size of the data set is not the only challenging part of the project, another difficulty being the limited understanding of what each bit in the transcript actually represents. The transcripts contain a large amount of information, that sometimes may present similarities for two cells, but carries very different meanings. There exists a level of understanding what information should be carried in each cell, but not enough to be able to quickly identify errors.

Multiple experiments have been conducted before in this area, attempting to classify cells of patients before and after chemotherapy. Unfortunately, the results are hard to interpret. In the beginning there were four types of cells identified and after the treatment, cells were grouped into five types. Out of these five groups, only one can be mapped to an initial group. This could be interpreted in multiple ways, for example: perhaps some of the groups of cells changed beyond recognition. Another possibility is that some groups of cells divided into smaller ones, but it is also equally possible that the initial groups of cells completely disappeared and were replaced by new ones.

Testing is another challenge with this project, because there is no straight-forward way of knowing if the identified types are actual correct types. Currently there is no knowledge of how many cell types should be identified, which means that there is no easy way to determine if two types are actually just one group or should be divided into smaller groups. Some cell data may differ for each person, meaning that there is a possibility the results of this project are not widely applicable.

The current sets of data have been maintained in a petri dish, which means that they may have developed differently than the ones inside a human body. For this reason, the actual analysis might turn out slightly different than the data of an actual patient. However, there exists a chance that a new set of data will be made available for analysis during the project, which will have been recently extracted and therefore more accurate.

Timeline and methodology

Figure 1 shows the initial plan for the project, consisting of three main phases, each of them divided in sub-tasks. The bulk of the work is assigned to Phase 3, because the most challenging parts of the project will be at the start, when the specifics are still to be decided upon and fully understood.

The methodology used is Agile, mostly due to the fact that the plan should be easily adapted to any unexpected findings or project scope changes.

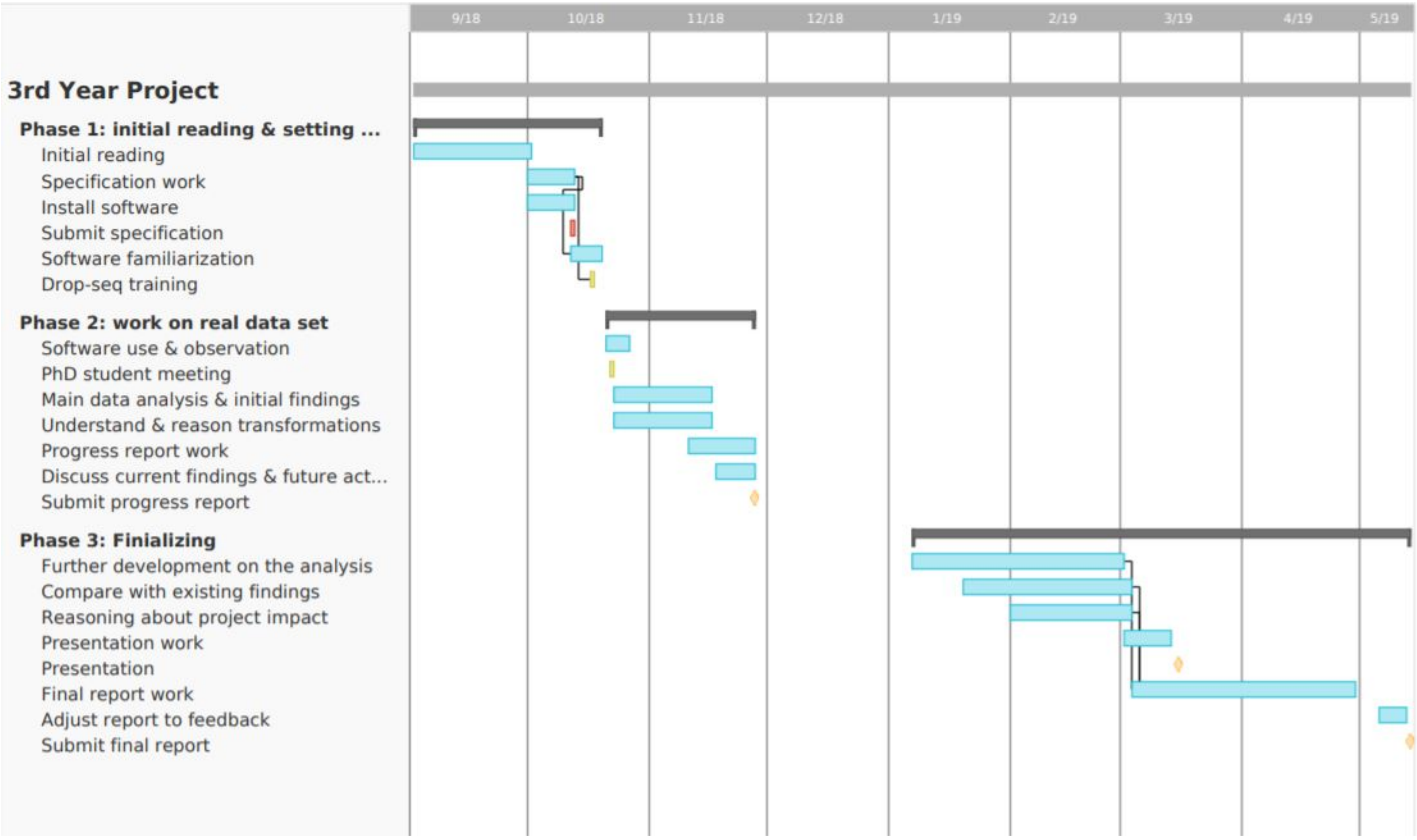

Figure 1: Project plan

Risk Assessment

The following table outlines the main risks of the project, along with a suggested solution for each.

ID	Description	Solution
R1	Software not functioning as expected. (Main software to be used: RStudio, Seurat)	Alternative software packages (Monocle, SC3, MAGIC) have been made available and could potentially be used throughout the project.
R2	Incomplete understanding of the project or problems using the software.	Maintaining a good relationship with the project supervisor, which will allow for me to ask for help and support with any such issues.
R3	Illness	The work on the project is aimed to finish before the deadline, such that the schedule can be easily deferred in case of unexpected events.

Ethical Considerations

The data to be analysed throughout the project is real data from patients at the Birmingham City Hospital, which involves some ethical considerations. However, the patients are entirely anonymised and have given their consent by signing an ethical approval, which is how they have volunteered to be part of the research.

Some of the analysed data will eventually be made public, when the results of the research are published, but the patients have all consented to this, meaning that there are no more further considerations for the scope of the project.

References

[1] *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets* [online]. HHS Public Access, 2015. [cited 11 October 2018]. Available from the World Wide Web: (https://www.ncbi.nlm.nih.gov/pubmed/26000488).

[2] Grinsteing G., Trutschl M., Cvek U. *High Dimensional Visualizations* [online]. Institute for Visualization and Perception Research University of Massachusetts Lowell and AnVil Informatics, Inc. [cited 11 October 2018]. Available from the World Wide Web: (https://pdfs.semanticscholar.org/43f7/66c06e2a7770d9f37dcd9cfff5bd5dcfc22f.pdf).

[3] Google Brain Team *How To Use t-SNE Effectively* [online]. Distill team, 2016 [cited 11 October 2018]. Available from the World Wide Web: (https://distill.pub/2016/misread-tsne/).

[4] RStudio. Available from the World Wide Web: (https://www.rstudio.com/products/rstudio/download/).

[5] Seurat. Available from the World Wide Web: (https://satijalab.org/seurat/).

[6] Monocle. Available from the World Wide Web: (http://cole-trapnell-lab.github.io/monocle-release/).

[7] *Visualizing Data Using t-SNE* [online]. Google Tech Talks, 2013. [cited 11 October 2018]. Available from the World Wide Web: (https://www.youtube.com/watch?v=RJVL80Gg3lA).