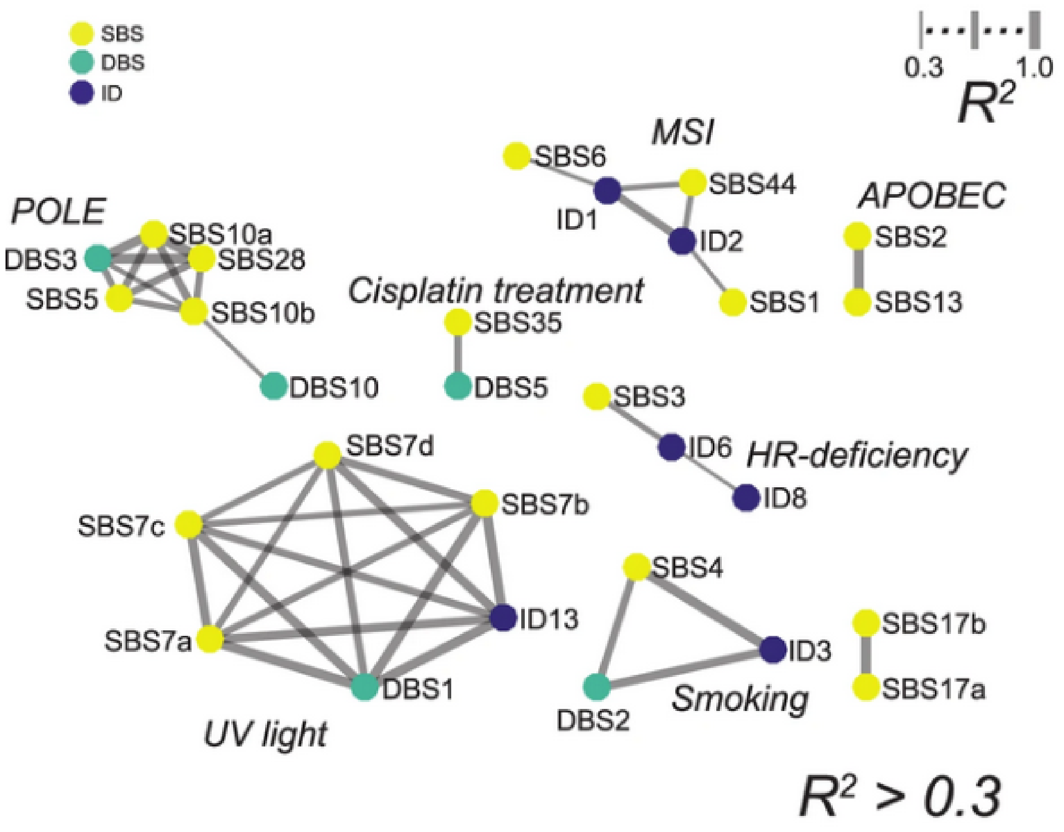
September t/m Januari 2024

Building mutational signatures in cancer using deep Bayesian neural nets

Plan of Approach



Student Jan Alfonso Busker j.a.busker@st.hanze.nl
Supervisor Gerton Lunter g.a.lunter@umcg.nl
Supervisor Hylke Donker h.c.donker@umcg.nl
Supervisor Michiel Noback m.a.noback@pl.hanze.nl

Introduction

Background

Cancer is predominantly considered a genomic disorder [1-3]. Cancer is distinguished by the unregulated proliferation of cells. In order to maintain this rate of proliferation, the cancer cells have developed many mechanisms, including the ability to evade apoptosis and the formation of vascular networks. The acquisition of most of these capabilities is primarily attributed to mutations in the DNA. The accumulation of these mutations has occurred as a result of exposure to endogenous and exogenous DNA-damaging mechanisms during the course of an individual's lifespan. It has been discovered that these deleterious processes result in the formation of mutations, which in turn leave a discernible mark on the genome [4]. The occurrence of "genomic scars" is attributed to the selective binding of certain substances to specific nucleotide sequences of the DNA. By examining the immediate environment of the mutations, one might deduce the manner in which these mutations were generated (aetiology). In addition you can also quantify how much the cancer cells (and its descendents) have been exposed to specific processes.
The trick is to count the number of mutations per context, and then to factorize this matrix using non-negative matrix factorisation. To date, the majority of researchers have focused their investigations on the dynamic nature of mutations, namely those involving the alteration of a single nucleotide within the surrounding environment [4][5].

Project Goals

The goal of this project is to refine: (1) the statistical model and (2) the current representation of the mutations.

1. **Refined model:** The current interpretation of mutational signatures is mostly based on post hoc analysis. The results of the different methods used to replicate the mutational signatures will be compared to determine the (meta) parameters.

2. **Refined context:** The current single base substitution signatures are based on one flanking nucleotide left and right of the substitution, which is not enough context to discriminate between different types of DNA damage. The goal is to increase the context size by adding at least one extra context letter and capture contextual imprints at periodicities corresponding to DNA loops. This can be achieved by increasing the space of the dimensions of the mutations by dividing mutations further according to DNA annotations.

Approach and Methods

To achieve the goals of refining the statistical model and the current representation of mutations in building mutational signatures in cancer using deep Bayesian neural nets, the following approach and methods will be used:

- Curate and quality control the samples and mutations: Data from ICGC and TCGA will be used, which is publicly available. Additional sources of data from HMF and PCAWG will be obtained to increase the dataset.

- Replicate the mutational signatures using non-negative matrix factorisation: The current single base substitution signatures based on one flanking nucleotide left and right of the substitution will be replicated. The signatures will be compared with signatures from 1-layer and 2-layer Bayesian neural networks.

- Determine priors for the expanded mutation representation: Based on existing COSMIC signatures, determined signatures of larger dimensions will be contrasted with the current 96 feature representation.

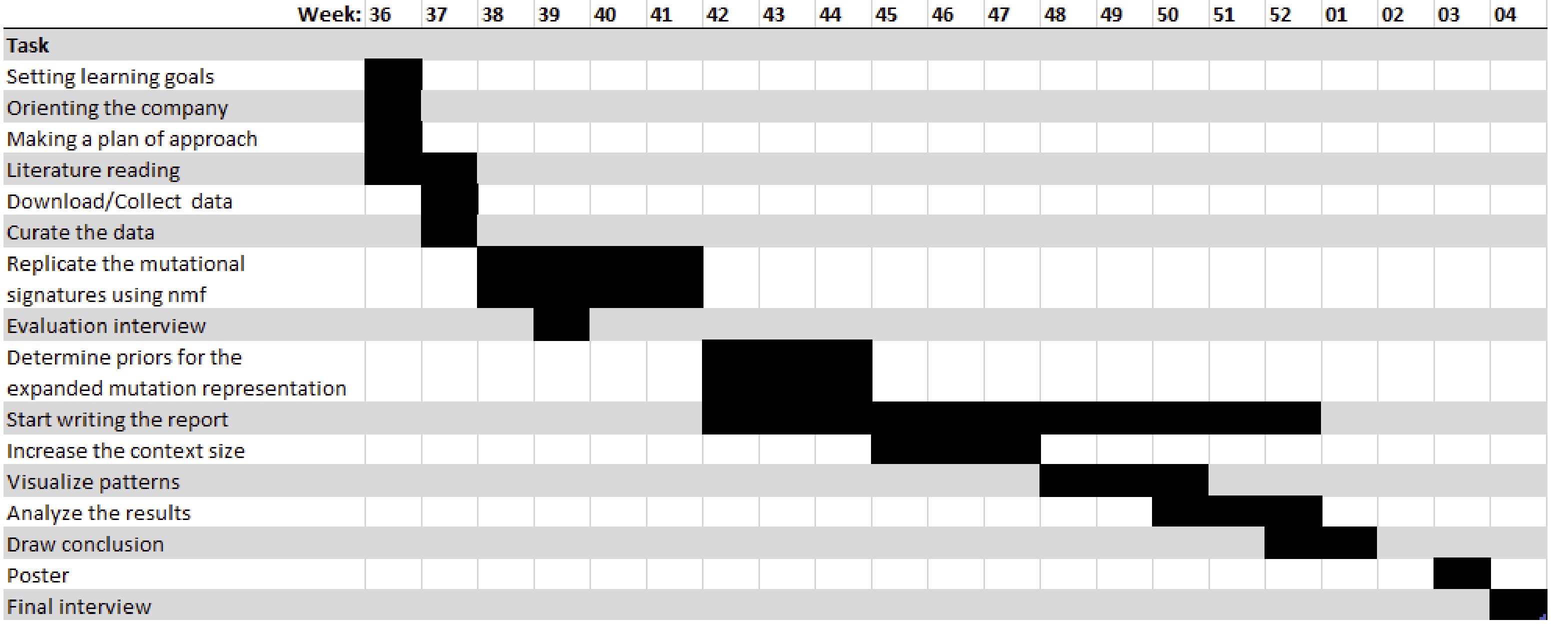
- Increase the context size: The context size will be increased by adding at least one extra context letter to discriminate between different types of DNA damage. The contextual imprints at periodicities corresponding to DNA loops will be captured by increasing the context size. This can be achieved by using latent dirichlet allocation (LDA) to divide mutations further according to DNA annotations.

- SigProfilerMatrixGenerator: SigProfilerMatrixGenerator will be used to visualize and explore patterns of small mutational events, including single base substitutions, doublet base substitutions, and small insertions and deletions. The generated matrices will be analyzed to identify patterns and signatures in the small mutational events, which can help refine the current representation of mutations.

- Analyze the results: The results of the different methods used to replicate the mutational signatures will be compared to determine which method is the most effective.

- Draw and communicate conclusions: Based on the results of the analysis, conclusions will be drawn about the effectiveness of the different methods used to replicate the mutational signatures and the potential for further research in this area. The results of the analysis will be communicated to the scientific community through publications and presentations.

Schedule



Week:	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	01	02	03	04
Task																					
Setting learning goals	■																				
Orienting the company	■																				
Making a plan of approach	■																				
Literature reading	■	■																			
Download/Collect data		■																			
Curate the data		■																			
Replicate the mutational signatures using nmf			■	■	■	■															
Evaluation interview				■																	
Determine priors for the expanded mutation representation							■	■	■												
Start writing the report							■	■	■	■	■	■	■	■	■	■	■				
Increase the context size										■	■	■									
Visualize patterns													■	■	■						
Analyze the results															■	■	■				
Draw conclusion																	■	■			
Poster																				■	
Final interview																					■

Organisation

This assignment is given by UMCG. The Department of Epidemiology is a major driving force in initiating and conducting life course research and is instrumental to the clinical research within the UMCG's main theme of Healthy Ageing. The research group 'Medical statistics and decision making' is part of the Department of Epidemiology. They focus on developing methods for statistical modeling in clinical and epidemiological studies and analyzing large cohort data. And develop decision analysis techniques to support benefit-risk assessments of medicines and medical decision-making.
Internal guidance (from the Hanze) is provided by Dr. M.A. Noback. And external guidance is provided by Prof. G.A. Lunter and Dr. Hylke C. Donker.

References

[1] Hanahan, Douglas, and Robert A. Weinberg. "The hallmarks of cancer." cell 100.1 (2000): 57-70.
[2] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144.5 (2011): 646-674.
[3] Hanahan, Douglas. "Hallmarks of cancer: new dimensions." Cancer discovery 12.1 (2022): 31-46.
[4] Alexandrov, Ludmil B., et al. "The repertoire of mutational signatures in human cancer." Nature 578.7793 (2020): 94-101.
[5] Degasperi, Andrea, et al. "Substitution mutational signatures in whole-genome–sequenced cancers in the UK population." Science 376.6591 (2022): abl9283.