Novel method to examine strong cancer driver mutations

Master Thesis

Sarah Elna Valentin | 201905732

Aarhus University | Bioinformatics Research Centre (BiRC)

Supervisor: Nicolai Juul Birkbak

Associate Professor

Department of Molecular Medicine

Aarhus University

Abstract

In this thesis a novel method was created to examine strong driver mutations in the context of the immune system hypothesis, which links cancer development to the immune system.

This was done by creating an expectation based on a linear regression on the mutational burden of patients combined with a gene weight. The expectation was found for a general dataset and 17 cancer-specific datasets. The expectation was tested with a χ^2-goodness-of-fit test. The alone-factor was also defined as the inverse of the mutational burden for a tumour sample and was used to filter potentially strong oncogenes from the results of the test.

The results of the χ^2-goodness-of-fit test found many significant genes. Of those, some were selected to be examined further based on how often they were found to be significant, how different they were from the expected, and them also having a high alone factor. These genes included TP53, KRAS, EGFR and PIK3CA, which were all known to be associated with cancer and identified as either tumour suppressor genes and/or oncogenes.

The genes found using this new method, combined with the alone-factor, had strong driver mutations. This supports that the method is able to identify strong driver mutations.

Acknowledgements

This project would not have been possible without many peoples' help, on a professional and personal level. I would like to thank Professor Nicolai Juul Birkbak for creating the project, and his continued guidance and advice. I would also like to thank Johanne Ahrenfeldt (postdoc) and Randi Istrup Juul (postdoc) for acting as co-advisors and sharing their expertise with me for this project.

I would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing.

All of the computing for this project was performed on the GenomeDK cluster. I would like to thank GenomeDK and Aarhus University for providing computational resources and support that contributed to the results of this project.

I am truly grateful for my family and friends who have supported me throughout my education. This especially goes for my husband, Jacob Løwe Valentin, for his daily encouragement and laughter. Finally, I would like to dedicate this thesis to 'bean', the little human growing inside me, who I am looking forward to meeting sometime close to the end of the year.

Contents