Identifying Copy Number Signatures by Latent Dirichlet Allocation

An Analysis Performed on Data from the Cancer Genome Altas

Janne Auning Engestoft

Supervisor: Nicolai Juul Birkbak

Co-supervisor: Mateo Sokač

Bioinformatics Research Center
Aarhus University
June 2022

Student number: 201708061

Table of contents

Abstract

Copy number variations (CNVs) play a key role in the instigation and progression of cancer and are associated with chromosomal instability (CIN), a state known to often confer poor patient prognosis. In fact, the continual gain and loss of chromosomes and chromosome segments, which is what characterises CIN, is thought to be the driving factor in several specific cancer types, such as high-grade serous ovarian carcinoma (HGSC).

Contrary to the mechanisms with which less complex genetic alterations are created, the underlying patterns of CNVs are not yet well characterised. This poses a challenge in the advancement of treatment in cancers with high rates of genetic aberrations, and thus the need for development of novel computational approaches is paramount.

This has led to the creation of this project, where the aim has been to identify the latent patterns of CNVs as signatures using latent Dirichlet allocation (LDA).

This resulted in the identification of 5 signatures, two of which were correlated with various CIN measures such as LST, HRD, and telomeric allelic imbalance. Simultaneously, three of the signatures were significantly associated with patient outcome across cancer types, where two of them were associated with survival, while a single one was associated with poor patient outcome.