A reference-free strategy for detecting circulating tumor DNA

Masters thesis
30 ECTS
Bioinformatics Research Center
Aarhus University

Anika Gottschalk and Carmen Oroperv

Page count: 89
Supplementary material: 12 pages
Handed in 15-06-2022

Abstract

Cancer is one of the most common causes of death, and if disease recurrence after initial treatment, such as tumor resection surgery, is detected fast, that often increases treatment success and overall survival of the patient. Efficient and precise strategies to detect relapse are therefore crucial. Cell-free DNA (cfDNA) is DNA fragments released into the blood during degradation of cells, and in cancer patients, a fraction of the cfDNA will be originating from the tumor: this part of cfDNA is called circulating tumor DNA (ctDNA). Several approaches have been proposed to use the fraction of ctDNA as a biomarker in cancer detection and treatment. Often, known somatic point mutations are used to detect ctDNA in blood samples. However, these point mutations become less and less reliable when the amount of input data is low and sequencing is also carried out with low coverage. The frequency of point mutations in these cases can become close to the error rate of the sequencing. It has therefore been proposed that clustering point mutations and structural variants could act as more reliable biomarkers to detect ctDNA in cancer treatment follow-up.

The goal of this thesis project was to develop a reference-free approach to identify tumor specific somatic variation from cfDNA samples. The approach was to identify k-mers that are unique to the given patients cancer genome, and then filter the cfDNA samples for these k-mers to detect ctDNA. To find the k-mers that are unique to the tumor, k-mers found in the germline were subtracted from k-mers found in the tumor samples before intersecting with the cfDNA samples to detect possible ctDNA. To improve detection, different restrictions and filtering approaches were applied to the data sets. The best combination of filters for the tumor k-mer set included a minimum counter of five on the tumor k-mers and quality filtering of the tumor reads before counting k-mers. To ensure that as much germline information is removed from the set of tumor k-mers, the germline k-mers of all patients were combined, and also merged with k-mers counted from the reference sequence and k-mers from a consensus sequence of the called germline variants and the reference genome, which was created to deal with k-mers that would appear at breakpoints between reads. The resulting numbers of unique tumor k-mers found in the cfDNA samples were used to calculate estimates of ctDNA fraction. A threshold of this fraction was defined to classify patients into relapsing and not relapsing patients, and define the time point when relapse is detected.

After finding the set of restrictions and filters that performed best on a training partition of the data (phase I patients), using this approach on a test split (phase II patients) could identify 11 of 28 relapsing patients. The data used in this project were generated by Claus Lindbjerg Andersens reseach group at Molekylær Medicinsk Afdeling at Aarhus University Hospital. In their project, they were able to identify 16 of 28 relapsing patients, though while producing a larger amount of false positives than the method developed in this project.

Further research and tests in clinical settings are needed before the method developed here could be applied in practice - if not instead of the follow-up based on imaging that is used now, then as a supplement that possibly could detect a relapse earlier.

Table of contents