

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 research 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

1	Introduction	1
1.1	Cell-free DNA and circulating tumor DNA	1
1.1.1	ctDNA and cancer treatment follow-up	2
1.1.2	Challenges in ctDNA detection	3
1.2	Mutations in cancer genomes	4
1.2.1	Single nucleotide variants	6
1.2.2	Structural variants	7
1.2.3	Foreign DNA in cancer genomes	8
1.3	Detecting somatic variation in cancer genomes and ctDNA	10
1.4	Current state of the art approaches and limit of detection (LOD) for detecting and quantifying ctDNA	12
1.5	Problem statement	16
2	Methods	18
2.1	Data	18
2.2	K-mers	19
2.3	Identifying unique tumor k-mers	20
2.4	Identifying unique tumor k-mers in the cfDNA	20
2.5	Estimating ctDNA levels in cfDNA	21
2.6	Development of the final workflow	22
2.7	Calling germline variants and creating consensus sequences	24
2.8	Visualization of the results	25
2.9	Analysing correlations and bias of the k-mer sets	25
2.9.1	Empirical analysis of the estimated ctDNA levels	25
2.9.2	Correlation between k-mer counts in the tumor and cfDNA	26
2.9.3	Correlation between cfDNA read count and estimated ctDNA levels	26
2.9.4	cfDNA k-mers not seen in germline or tumor k-mer sets	27
2.10	Mapping unique tumor k-mers to the human reference	27
2.11	Identifying k-mers originating from other organisms	27
2.12	Code availability and software	28
3	Results and discussion	30
3.1	Baseline workflow	30
3.1.1	Counting k-mers from germline, tumor and cfDNA reads	30
3.1.2	Detecting and counting unique k-mers in cfDNA samples	32
3.1.3	Estimating ctDNA levels based on the k-mer counts	33
3.2	Experiments on the germline	35
3.3	Experiments on the tumor k-mer sets	36
3.4	Experiments on the cfDNA k-mer sets	38
3.5	Changes in the unique tumor k-mer sets	39

3.6	Changes in the set of ctDNA k-mers	41
3.7	Analysing correlations and bias of the k-mer sets	42
3.7.1	Empirical analysis of the estimated ctDNA levels	42
3.7.2	Correlation between k-mer counts in the tumor and cfDNA	45
3.7.3	Correlation between cfDNA read count and estimated ctDNA levels	48
3.7.4	cfDNA k-mers not seen in germline or tumor k-mer sets	50
3.8	Final workflow	53
3.9	Mapping unique tumor k-mers to the human reference	60
3.10	K-mers originating from other organisms	62
3.11	Results of the phase II patients	67
3.12	Possible further improvements	74
4	Conclusion	76
	Bibliography	77
	Appendix A Appendix	90
A.1	Data	90
A.2	Germline reads	93
A.3	cfDNA reads	93
A.4	Empirical analysis of the estimated ctDNA levels	94
A.5	Correlation between cfDNA read count and estimated ctDNA levels	95
A.6	K-mers originating from other organisms	97
A.7	Phase II	98