Lengths of circulating DNA fragments as a promising predictor of cancer stage

Marek Štrba[1,2], **Jaroslav Budiš**[1,3,4], **Werner Krampl**[1,3,5], **Tomáš Sládeček**[1], **Ondrej Pös**[1,3,5], **Mária Lucká**[6,7], **Tomáš Szemes**[2,3,5]

[1]Geneton Ltd., Bratislava, Slovakia
[2]Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia
[3]Comenius University Science Park, Bratislava, Slovakia
[4]Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia
[5]Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava
[6]Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
[7]Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

Bioinformatics has become one of the key scientific disciplines during the 21st century. The aim of this work was to utilize state-of-the-art tools and help further the research of oncologic diseases while improving the quality of life of people by contributing to the research of liquid biopsy. To achieve this goal we analysed dataset of oncologic samples Graz with focus on DNA fragment length profiles in blood plasma. Our hypothesis was that, the oncologic samples will have shorter fragments due to the ctDNA being released into the bloodstream. The results from the Graz dataset confirmed our hypothesis. After confirming our hypothesis we tested different ML models and feature selection methods on the Graz dataset samples with the aim of creating the best possible predictor for differentiating oncologic samples from healthy samples. The result of our work was creation of a SVM predictor, which offered prediction with satisfactory accuracy.

Keywords: Liquid biopsy, Machine Learning, SVM, Fragment length profiles, cfDNA, ctDNA

Dĺžkové profily fragmentov cirkulujúcej DNA ako potenciálny prediktor štádia onkologických ochorení

V 21. storočí sa stala Bioinformatika jednou z popredných vedeckých disciplín. Cieľom tejto práce bolo využiť najmodernejšie dostupné nástroje a pomôcť svojim výsledkom k výskumu onkologických ochorení a prispieť tak k zlepšeniu kvality života ľudí tým, že prispejeme k výskumu problematiky tekutej biopsie. Za týmto cieľom sme postupne analyzovali dataset onkologických vzoriek Graz so zameraním sa na dĺžkové profily DNA fragmentov v krvnej plazme. Našou hypotézou bolo, že vzorky onkologických pacientov budú mať kvôli uvoľňovaniu ctDNA do krvi kratšiu dĺžku fragmentov. Výsledky analýzy Graz datasetu našu hypotézu potvrdili. Po potvrdení našej hypotézy sme otestovali rôzne ML modeli a rôzne „feature selection" metódy na vzorkách Graz datasetu za cieľom vytvorenia čo najlepšieho prediktora na rozlíšenie onkologických a zdravých vzoriek. Výsledkom bolo vytvorenia prediktora na báze SVM, ktorý nám ponúkol dostatočne uspokojivú presnosť predikcie.

Kľúčové slová: krvná biopsia, strojové učenie, SVM, dĺžkové profily fragmentov, cfDNA, ctDNA

Newslab, 2021; roč. 12 (2): 58 – 63

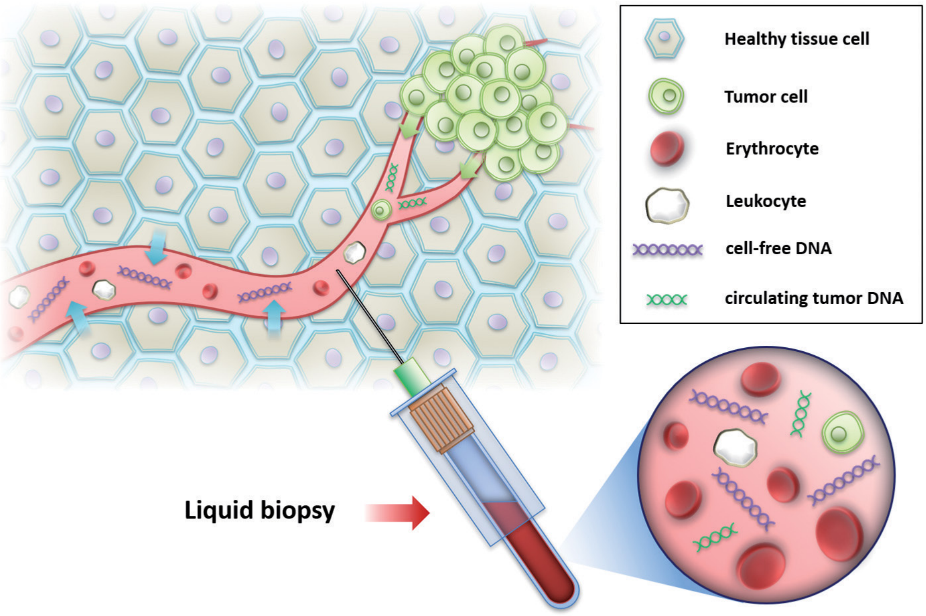
Introduction

Biopsy has been one of the leading diagnostics approaches to characterize oncology disease for many years. This well known procedure has saved many lives, however it has multiple crucial drawbacks including the high cost, discomfort it brings to the patient, difficulty of obtaining the samples in more complicated diagnosis and lastly it does not show the whole image of the tumour based on the tumour's heterogeneous nature[(1)]. The answer to these problems may in future be the liquid biopsy method. This method relies on the analysis of blood samples from patients, specifically the occurrence of circulating tumour DNA (ctDNA) in the sample[(2)] as can be seen in **Figure 1**. The goal of liquid biopsy is to monitor the development and progress of malignant disease and it's reaction to therapy by analyzing blood samples of the patient. Furthermore, liquid biopsy proved to be a useful tool in the process of targeted therapy - changing of medicine and therapy based on changes in blood test results[(1,2)].

Cell apoptosis and/or cell necrosis leads to release of short nucleic fragments(<166bp) into plasma[(3)]. These fragments are referred to as cfDNA - circulating cell-free

DNA. The tumour derived subset of cfDNA, caused by high turnover rate of tumour cells is called ctDNA - circulating tumour DNA[(3)]. Circulating Tumour Cells (CTC) are rare cells released into patient's bloodstream. These cells then extravasate to different organs, which in small fraction leads to

Figure 1. Liquid biopsy schema. Cell free DNA (cfDNA) is released into blood circulation by cells. The majority of these fragments is in length of 150-180bp and come from hematopoietic cells. However the DNA is also released by surrounding tissues potentially including tumour cells in the bodies of oncologic patients. These circulating tumour DNA fragments (ctDNA) are generally shorter than 145bp. The tumour releases not only the nucleic acids, but also the circulating tumour cells (CTC) which can also be used in the process of genetic analysis. The blood also contains leukocytes, which after the centrifugation of blood create so called "buffy coat", routinely used in the extraction of genomic DNA:



formation of metastasis[1]. Detection of these cells after the first cycles of therapy, combined with their short lifespan indicates a futility of the treatment[2].

There are several big challenges, that must be overcomed before liquid biopsy could be widely used in clinical practice.

Firstly, the concentration of ctDNAs in blood is very low, which leads to the need of extremely sensitive and specific analytic methods for the characterisation and even detection of these fragments. The concentration of ctDNA is even lower in samples of patients in early stages of cancer[2].

Secondly, even though the analysis of plasma samples, used in the ctDNA method is easier, the samples may be contaminated by cfDNA released during innate processes of cell life-cycle. The ctDNA method also severely lacks required standardisation of techniques.

Thirdly, the differences between healthy cfDNA and oncologic ctDNA are not significant which leads to problems in distinguishing them from each other.

Lastly, the liquid biopsy is not usable in testing of every cancer type or every patient, diagnosed with a given type of cancer. The main problem being the insufficient concentration of CTCs and ctDNA fragments into the bloodstream of the patient[4].

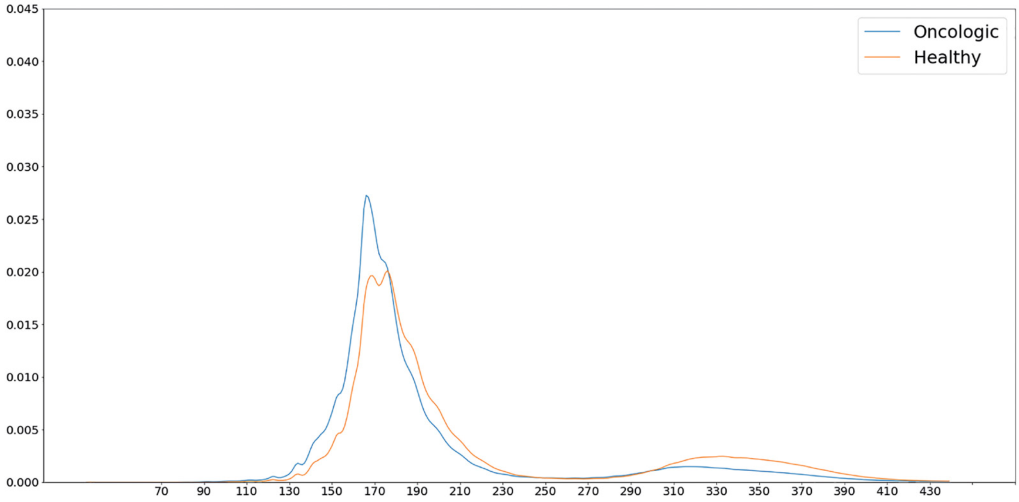
Multiple studies were conducted with the goal of confirming that there is a difference between the length of the ctDNA fragments and regular cfDNA fragments. Oncologic patients were proven to have increased quantity of cfDNA, which can be result of tumour shedding ctDNA fragments into the bloodstream[5].

The more valuable information is however, that the average size of ctDNAfragments is smaller than the size of a noncarcinome cfDNA[5,6]. Some specific examples include:

Human Melanoma (Healthy-165bp, Oncologic-145bp[5], Human Lung Cancer (Healthy-283.7±4.1 bp, Oncologic-277.0±4.7bp)[5], Animal models of GBM (Most common length of Healthy fragments - 167bp, most common length of Oncologic fragments - 134 and 144bp)[5]. It has also been proven, that the modal size of ctDNA fragments for most of the cancer types is bellow 167bp[6]. The typical size of cfDNA fragments is between 150 and 200bp[7] which provides only small intersection of sizes. The shorter length of fragments originating from tumours compared to healthy fragments has been also confirmed by a study of dried blood samples, in which it was determined that the modal length of ctDNA fragments was 150bp compared to 170bp of healthy cfDNA fragments[8] Furthermore a study from 2015, targeted at patients suffering from HCC observed a positive correlation between the proportion of DNA fragments less than 150bp and the tumour DNA fraction in plasma as well as negative correlation between the proportion of DNA fragments larger than 180bp and tumour DNA fraction.[9] The same study also determined based on the difference between cumulative frequencies between 8p deletions and 8q amplifications in plasma DNA of samples that since the difference attains maximum at 166bp the key difference between tumorous and nontumorous plasma DNA is the relavite abundance of DNA <166bp and 166bp[9].

It has been also proven, that not only the ctDNA fragments have shorter average length, but also, that the length of the fragment is clearly tied to the frequency of mutations on allele. In Melanoma Patients samples, the mutations occured in highest frequency on fragments of length between 110bp and 140bp[5]. Lung cancer samples also showed higher frequency of mutations in shorter fragments[5].

Figure 2. Here we can see the mean oncologic and healthy fragment length profile calculated from samples in our dataset. The biggest difference is in a dominant peak around length 165 bp in oncologic samples which is not present in the healthy samples.Furthermore the healthy samples show higher count of fragments of longer length. These observations support the hypothesis of oncologic samples having shorter fragment length profiles.



Fragmentation of mutated ctDNA is much more prevalent than the fragmentation of healthy cfDNA[6] which could explain the overal shorter length of the ctDNA fragments.

It has also been suggested, that the maximum enrichment of the ctDNA occurs in fragments of length between 90 and 150bp, with a secondary group in ranges of 250 to 320bp[6].

The contradicting evidence or facts, that could prove problematic are less numerous, yet they are present. It has been concluded that fraction of ctDNA in early stages of cancer is very low[14] which could lead to fragments of ctDNA that are longer than expected.

In the process of testing the lung cancer samples, it was discovered that, there is a considerable overlap in peak values of healthy and oncologic fragments[5].

The studies that were analysed for this study overwhelmingly support the use of fragment length of the DNA fragments in plasma as a partial or independent predictor for differentiating healthy samples from carcinoma samples. This conclusion was achieved based on several independent studies, which provided result determining clear difference of length of ctDNA and cfDNA fragments[5-9].

The goal of our study was to learn whether the fragment length profiles of samples could be used as a partial predictor for machine learning based liquid biopsy. Fragment length profiles represent the amount of DNA fragments of every given length in bp (base pairs) found in sequenced samples. We aimed to analyse differences in length profiles between samples and suggest whether these profiles could be used in the process of diagnosis and monitoring of oncologic diseases.

Material and Methods

Data acquisition

We downloaded the dataset of 116 oncologic samples from 44 patients (multiple samples were obtained from several patients as the disease progressed) and 22 control healthy samples[10] (EGA Study ID EGAS00001003530). All four grades of tumour severity[11] were present among the samples retrieved from patients.

Analysis

The first stage of data processing was carried out as previously described[12]. NextSeq-produced fastq files (two per sample) were directly mapped using the Bowtie 2[13] algorithm with --very-sensitive option to the human reference genome hg19 (GRCh37). Reads with mapping quality of 40 or higher were retained for further data processing. Length of a DNA fragment was determined as the difference of the leftmost and the rightmost mapped base of the corresponding read pair.

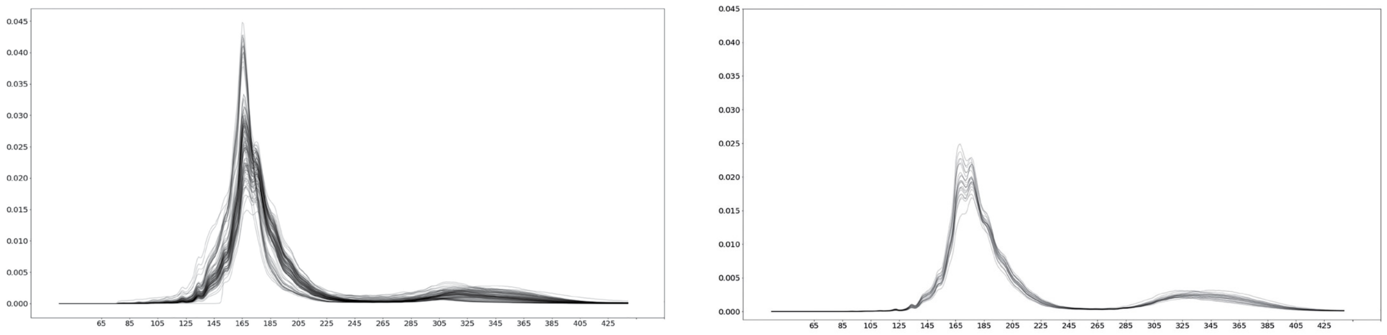
As the first step we decided to calculate weighted median and mean of fragment lengths of the samples. We used the functions provided by the Weighted stats library (https://pypi.org/project/weightedstats/).

As we can see from results, there is a clear difference between oncologic and control samples in their length. Medians of both metrics are significantly lower (U=214.15, $p=3.15 \times 10^{-10}$) for oncologic samples (162-193 bp, mean 174.28 bp) than healthy (177-192 bp, mean 183.18 bp), suggesting that the overall length of fragments in oncologic samples is shorter than the length in control samples. The only statistic contradicting this conclusion is the maximum weighted median of fragment length in oncological samples, which is higher than the healthy one.

However this phenomena can be easily attributed to the existence of an outlier in oncologic samples. The difference between healthy and oncologic samples can be clearly seen on mean fragment length profiles shown in **Figure 2**.

After the initial analysis of median and mean fragment lengths suggested a clear difference between oncologic and control samples we decided to normalize the count of fragments of given lengths in the samples (we divided the number of fragments of every given length by the sum of all fragments in the sample in effort to ease comparison between the samples). Following the normalization we plotted all on-

Figure 3. Plots (a) and (b) show all fragment length profiles of oncologic and healthy respectively. The darker shade of the line indicates that the given part of profile was shared by higher number of samples when compared to lighter lines. These plots, same as the **Figure 2** *show the dominant peak around 165bp in oncologic samples as well as the higher number of longer fragments in healthy samples.*



(a) **(b)**

Figure 4. Boxplots shown in this figure represent the declining trend of median fragment length with the rising severity of tumour (tumour grade). The higher values in G3 samples when compared to G2 and G4 can be attributed to the early stages of metastasis. The metastatic secondary tumours release longer fragments compared to the main tumour, which dilute the readings leading to a higher median fragment length. Other grades however all show signs that lower median fragment length could be linked to higher severity of the tumour.

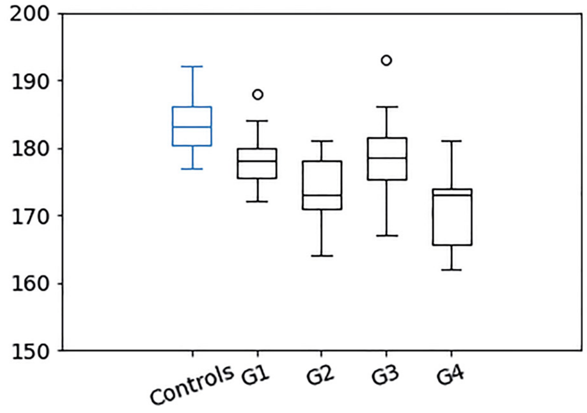
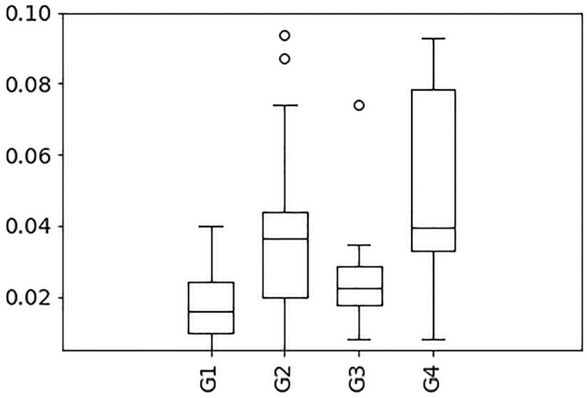


*Figure 5. In effort to reduce the undeniable visual difference between oncologic fragment length profiles and a mean healthy fragment length profile into just one number we calculated their euclidean distance. When grouped by the tumour severity we arrived to similar findings as with the median fragment length (***Figure 4***). The bigger the difference (higher number) the higher tumour severity. G3 proved to be an outlier similarly to the median fragment length case.*



cologic and all healthy fragment length profiles to allow for visual comparison.

The difference between Oncologic samples and Control samples is clear. Oncologic samples have very high and sharp peak in lengths between 155 bp and 165 bp (**Figure 2 and 3(a)**), which is completely absent in the control samples. On the other hand the control samples contain a higher number of fragments of length between 285 bp and 395 bp, which we can clearly see represented by the higher ridge on the right tail of the plot (**Figure 2 and 3(b)**).

We also compared the oncologic samples between each other in effort to find differences in their length profiles based on the grade of the tumour. We compared the median fragment lengths by the grade (**Figure 4**) and calculated euclidean distance of fragment length profiles from mean control healthy profiles (**Figure 5**).

We found that the G1 (the lowest tumour grade) samples were the most similar to control samples both in the median values and according to calculated euclidean distances. On the contrary the G4 (the highest tumour grade) samples proved to deviate from the control samples more significantly.

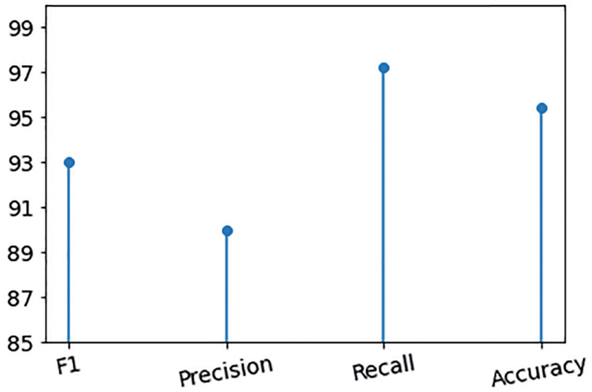
Machine learning

Based on the results from data analysis we decided to create a prototype of machine learning model which would differentiate the oncologic samples from healthy controls.

The crucial part of any machine learning is the feature selection. Based on the analysis results we were convinced that the calculated weighted median and weighted mean 35 had to be included as features. No matter how clear was the difference in weighted means and medians between oncologic samples and healthy controls these two features aren't enough to show the whole characteristic of a length profile. We tested several different approaches to select the additional features:

- **k-best** - using the sklearn's univariate feature selection (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html) which selects

Figure 6. The model achieved satisfactory results on the test datasets with scores of: F1 - 0.9301, Precision - 0.9 (ability to classify an oncologic sample as oncologic), Recall - 0.972 (ability to classify a healthy sample as healthy) and Accuracy - 0.9545. All scores were at least 0.9 which shows great promise in further usage of fragment length profiles in machine learning models used in oncology.



the best features based on univariate statistical tests we chose the 10 best features which had the best results in the given tests. 12 features total.

- **medians** - we divided the fragment length profile of a sample into ten bins by the
- fragment length. Each bin had the same range (first bin: 1bp-100bp, second bin: 101bp-200bp and so on…). After the division into bins we calculated the weighted median of every bin and added it as a feature. This method produced 10 additional features. 12 features total.
- **sums** - we divided the fragment length profile into bins same as when we calculated the medians, however in this method we summed the number of fragments
- in given bin instead of calculating the weighted median of the length in the bin. This method also provided 10 additional features with 12 total.
- **quantiles** - when using this method we calculated the quantiles of the fragment length profile with the step 0.1 (calculating the Q(0.1), Q(0.2)…Q(0.9)). This resulted in 9 additional features, 11 total.
- **peaks** - last method we decided to test was splitting the fragment length profiles into bins with unequal range based on the occurrence of peaks in the mean plotted profiles. The ranges we chose were:
  1. 1bp-159bp - the pre-peak area
  2. 160bp-174bp - main peak
  3. 175bp-189bp - secondary peak
  4. 190bp-229bp - downwards slope
  5. 230bp-279bp - trough
  6. 280bp-409bp - last low peak
  7. 410bp-1000bp - tail

  This method resulted in fewest additional features - 7, totaling only 9.

All additional features were calculated from the normalized values.

We tested machine learning models based on Decision Trees, Random Forests, SVMs and XGBoosting with multiple different hyperparameter settings. We compared the models in combination with all feature selection models using the GridSearch method based on 10 fold cross validation with the deciding metric being the F1 score because of the unbalanced nature of our dataset. The selection process was done using 85% of our data as the training sub-dataset.

The SVM model combined with the “medians” feature selection method provided the best result of F1 Score equal to 0,9325. We used this model to predict the remaining 15% of the dataset with results captured in **Figure 6**.

Discussion

With the results we acquired during the research it is safe to claim that fragment length profiles are a viable indicator of oncologic disease in the organism. However no matter how promising the results may seem it is crucial to test this conclusion on more data, ideally from multiple different diagnoses. Oncologic disease show great variance in their behaviour and structure which is the main reason behind the need for more testing.

The scope of this work did not allow for testing of other machine learning algorithms and neural networks or deep learning algorithms, which could potentially bring even better results. However even without testing these methods we were able to prepare a machine learning model with results satisfactory enough to be used in combination with other partial predictors in a larger meta-predictor. This was the main goal of this study, which means it is safe to say we were successful.

During our research we also learned that samples with different tumour grades deviate in their fragment length profiles. The level of deviation varied, however it certainly showed a great potential for further research. The important finding was, that the Euclidean distance from mean control profile grew with the severity of the disease. Based on our analysis the biggest challenge, when solving this differentiation, will most likely be the differentiation between grades G2 and G3 since the samples of these grades had often overlapping features.

What we deem as the biggest achievement of this study is that it showed a real possibility of liquid biopsy becoming a reality in the relatively near future. This would noticeably increase the quality of life for oncologic patients and also increase their chances of remission and curing.

Acknowledgment

This article was created with the support of the OP Integrated Infrastructure for the project: Long term strategic research and development focused on the occurrence of Lynch syndrome in the Slovak population and possibilities of prevention of tumors associated with this syndrome, ITMS: 313011V578, co-financed by the European Regional Development Fund.

Poďakovanie

Táto publikácia vznikla vďaka podpore v rámci Operačného programu Integrovaná infraštruktúra pre projekt: Dlhodobý strategický výskum a vývoj zameraný na výskyt Lynchovho syndrómu v populácii SR a možnosti prevencie nádorov spojených s týmto syndrómom , kód ITMS: 313011V578, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

REFERENCES

1. E. Crowley, F. Di Nicolantonio, F. Loupakis, andA. Bardelli, “Liquid biopsy: Monitoring cancer-genetics in the blood,”Nature reviews Clinical oncology, vol. 10, no. 8, p. 472, 2013.

2. K. Pantel and C. Alix-Panabi`eres, “Real-time liquid biopsy in cancer patients: Fact or fiction?” Cancer research, vol. 73, no. 21, pp. 6384–6388, 2013.

3. L. Calapre, L. Warburton, M. Millward, M. Ziman, and E. S. Gray, “Circulating tumour dna (ctdna) as a liquid biopsy for melanoma,” Cancer Letters, vol. 404, pp. 62–69, 2017.

4. I. S. Haque and O. Elemento, “Challenges in using ctdna to achieve early detection of cancer,” BioRxiv, p. 237 578, 2017.

5. H. R. Underhill, J. O. Kitzman, S. Hellwig, N. C. Welker, R. Daza, D. N. Baker, K. M. Gligorich, R. C. Rostomily, M. P. Bronner, and J. Shendure, “Fragment length of circulating tumor dna,” PLoS genetics, vol. 12, no. 7, e1006162, 2016.

6. F. Mouliere, D. Chandrananda, A. M. Piskorz, E. K. Moore, J. Morris, L. B. Ahlborn, R. Mair, T. Goranova, F. Marass, K. Heider, J. C. M.Wan, A. Supernat, I. Hudecova, I. Gounaris, S. Ros, M. Jimenez-Linan, J. Garcia-Corbacho, K. Patel, O. Østrup, S. Murphy, M. D. Eldridge, D. Gale, G. D. Stewart, J. Burge, W. N. Cooper, M. S. van der Heijden, C. E. Massie, C.Watts, P. Corrie, S. Pacey, K. M. Brindle, R. D. Baird, M. Mau-Sørensen, C. A. Parkinson, C. G. Smith, J. D. Brenton, and N. Rosenfeld, “Enhanced detection of circulating tumor dna by fragment size analysis,” Science Translational Medicine, vol. 10, no. 466, 2018, issn: 1946-6234. doi: 10.1126/scitranslmed.aat4921. eprint: https://stm.sciencemag.org/content/10/466/eaat4921.full.pdf. [Online]. Available: https://stm.sciencemag.org/content/10/466/eaat4921.

7. M. Fleischhacker and B. Schmidt, “Circulating nucleic acids (cnas) and cancer—a survey,” Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, vol. 1775, no. 1, pp. 181–232, 2007.

8. K. Heider, J. C.Wan, J. Hall, J. Belic, S. Boyle, I. Hudecova, D. Gale, W. N. Cooper, P. G. Corrie, J. D. Brenton, et al., “Detection of ctdna from dried blood spots after dna size selection,” Clinical Chemistry, vol. 66, no. 5, pp. 697–705, 2020.

9. P. Jiang, C.W. Chan, K. A. Chan, S. H. Cheng, J.Wong, V.W.-S.Wong, G. L.Wong, S. L. Chan, T. S. Mok, H. L. Chan, et al., “Lengthening and shortening of plasma dna in hepatocellular carcinoma patients,” Proceedings of the National Academy of Sciences, vol. 112, no. 11, E1317–E1325, 2015.

10. C. G. Smith, T. Moser, F. Mouliere, J. Field-Rayner, M. Eldridge, A. L. Riediger, D. Chandrananda, K. Heider, J. C. Wan, A. Y. Warren, et al., “Comprehensive characterization of cell-free tumor dna in plasma and urine of patients with renal tumors,” Genome medicine, vol. 12, no. 1, pp. 1–17, 2020.

11. N. C. I. USA, https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet, Last visited: 18. 10. 2020.

12. Minarik, G.; Repiska, G.; Hyblova, M.; Nagyova, E.; Soltys, K.; Budis, J.; Duris, F.; Sysak, R.; Gerykova Bujalkova, M.; Vlkova-Izrael, B.; et al. Utilization of Benchtop Next Generation Sequencing Platforms Ion Torrent PGM and MiSeq in Noninvasive Prenatal Testing for Chromosome 21 Trisomy and Testing of Impact of In Silico and Physical Size Selection on Its Analytical Performance. *PLoS ONE* 2015, *10*, e0144811. [Google Scholar] [CrossRef]

13. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med. 2010;2: 61ra91. pmid:21148127

14. C. -C. Huang, M. Du, and L. Wang, “Bioinformatics analysis for circulating cellfree dna in cancer,” Cancers, vol. 11, no. 6, p. 805, 2019.

Bc. Marek Štrba
Geneton s.r.o.
Ilkovičova 8, 841 04 Bratislava
email: marek.strbaa@gmail.com