CfDNAfragmentomics R and Python Package for Analysing cfDNA data

Yasamin Nouri Jelyani[*]

University of Toronto, Canada
yasamin.nourijelyani@mail.utoronto.ca

Abstract. cfDNA (cell-free DNA) fragments in liquid biopsies have a great potential to be used as a diagnostic tool for diseases like cancer. These cfDNA molecules have properties such as fragment length, methylation, and nucleosome occupancy that are intensely explored in research. This information can be used as biomarkers for detecting and subtyping cancer. ctDNA (circulating tumor DNA) are cfDNA fragments that circulate the blood with fragmentation features that are specific to blood samples coming from tumor effected patient. This paper presents a novel package, called cfDNAfragmentomics that utilizes fragmentation features of circulating cell-free DNA to detect the presence of circulating tumor DNA (ctDNA) and identify cancerous blood samples. This package includes a R Shiny app for an easy to use graphic user interface for users to input their data. CfDNAfragmentomics analyses the length distribution of cfDNA fragments, and utilizes nucleosome occupancy features which are known to be altered in cancer patients. Using naive statistical models, cfDNAfragmentomics identifies ctDNA and classifies cancerous versus non-cancerous blood samples. We evaluated the performance of cfDNAfragmentomics using real-world datasets from Snyder et. al. and Cristiano et. al. and demonstrated its sensitivity and specificity compared to existing methods. Furthermore, cfDNAfragmentomics provides a user-friendly interface for data input and result interpretation, making it accessible to researchers and clinicians with varying levels of expertise. We believe that cfDNAfragmentomics has the potential to advance the field of non-invasive cancer detection and personalized medicine, by providing a reliable and cost-effective tool for early diagnosis and monitoring of cancer patients.

Keywords: cfDNA, ctDNA, R package, Python, Liquid Biopsies

[*] Supervisor: Dr. Jared Simpson.
Contributions: Thank you to Dr. Jared Simpson and Jonathan Broadbent for significant contribution to the author's learning. Thank you to Jonathan Broadbent for the significant contribution to the development of the nucleosome-occupancy.py script.

1 Introduction

Cancer is a life threatening condition with an estimated burden of almost 10 million deaths in the year 2020 alone (21). Detecting the early signs of cancer and providing treatment for it before its evolution can have a significant impact on the survival and improvement of the condition of the patient, as well as helping reduce the rate of mortality. This is due to the fact that early detected cancer is often more curable and manageable, needing less life-threatening treatments such as chemotherapy, radiation therapy, and surgery (22). On the other hand, late stage cancer is more likely to evolve quickly, can spread more rapidly throughout the body due to metastasis, and is less likely to have a treatment that targets the cancerous cells directly (1). Late stage cancer can reduce the chances of survival of the patient and can have a significant financial impact on the healthcare system.

There is a large body of evidence supporting the necessity of early cancer detection to improve the negative effects of it on the patient. For instance, a paper published by Myers et.al. discussed that screening for breast cancer reduced the mortality of patients by approximately 20 percent (23). There is a growing body of research analysing less invasive screening strategies such as using blood samples in order to detect cancer early on in the tumor's life span (3). The use of liquid biopsies and the analysis of cfDNA (cell-freeDNA) in the blood using machine learning models have shown to have great potential for early cancer detection. These machine learning models can predict the possibility of the blood being originated from a cancer effected body versus a healthy blood sample. Examples of three such models include Griffin, DELFI, and ichorCNA which are explained in further detail under section 1.2, Current Technologies.

1.1 Motivation

This article is motivated by a proposal written by the author. This section refreshes some details of the introduction to the research geared towards the development of the cfDNAfragmentomics package.

The early detection and monitoring of cancer is crucial in the long-term survival of patients (1). Cancer subtyping is also an essential step in clinical oncology to identify the tissue of origin and develop a treatment regimen (2). Tissue biopsies are important for tumor subtyping and can guide treatment and prognosis strategies (3). However, there exist clinical impracticalities in using tissue biopsies, such as the complications caused for immunosuppressant patients (4). This makes tissue biopsies less common to be performed in the clinic and can result in a diagnosis to be based on the primary tumor (3). In metastatic cancer, basing the treatment on the primary tumor can be lethal as the tumor might evolve, and using previous biopsy analyses may lead to treatment resistance (3). There exist minimally invasive techniques such as liquid biopsies, which allow for the detection of cancer and are practical for use in the clinic (5). Liquid biopsies are samples from the blood, which contain small fragments of DNA called cfDNA (cell-free DNA) (5). These short DNA fragments are released into the blood due to cell death, are commonly 150-200 base pairs in length, and are packaged as nucleosomes (5). In healthy individuals, the cfDNA typically originates from hematopoietic cells because they are abundant, have high turnover, and have close access to the vasculature (5). In cancer patients, there exists a subset of cfDNA called ctDNA (circulating tumor DNA) (5). The ctDNA cell of origin affects their fragmentation pattern which can be useful in cancer subtyping (5). The phenotype of ctDNA is often very different from healthy cfDNA because of the epigenetic and genetic alterations in tumor cells and many chromosomal abnormalities (14).

cfDNA is typically originated from apoptosis of cells in the body that are ruptured to maintain the homeostasis for the number of cells that exist in a healthy individual(6, 15). In apoptosis, cfDNA is fragmented by intracellular endonucleases and is commonly short ($<$ 150 base pairs) in length (6, 15). Healthy tissues can also undergo damage which can lead to necrosis (unplanned cell death). In this case, the cell ruptures, and long DNA fragments exit the cell (4).

One of the hallmarks of cancer is the inhibition of apoptosis to induce tumorigenesis (7). When apoptosis is suppressed, there is an increase in metabolic stress due to ATP depletion that can increase necrosis (8). This may result in increased concentrations of long ($>$ 150 base pairs) cfDNA fragments in the plasma of cancer patients (15). Long fragments of cfDNA due to necrosis are potential biomarkers in cancer, and their high concentrations have been correlated with an increased chance of cancer recurrence (15). Due to the low concentration of cfDNA in the blood, especially in early-stage cancer, there is a need for very sensitive detection methods to be able to identify these molecules (5). There are several challenges to using cfDNA for cancer detection, including the low abundance of tumor-derived cfDNA in the bloodstream, the need for sensitive detection methods, and the potential for false positives and negatives.

Methods of detection for cfDNA include Polymerase Chain Reaction (PCR) assays and Next Generation Sequencing (NGS) (5). Although NGS has less sensitivity, it is preferred over PCR because it allows for Whole Genome Sequencing (WGS) and a broad assessment of the genome (5). Moreover, in cfDNA, Transcription Factor (TF) binding sites can be aggregated and used to find sequencing coverage of the cfDNA and identify the tissue of origin (3). Hence, NGS can be useful in the detection of cancer and subtyping. To sequence genes, technologies such as Illumina which uses bisulfite approaches for methylation analysis, and Oxford Nanopore Technologies (ONT) are used (9). ONT is used for shallow whole genome sequencing to find genetic differences among DNA molecules using features such as methylation and fragmentation (9). Hence, ONT can be used to identify cancer-associated fragmentation signatures. ONT is an ion membrane channel that receives blood samples, captures the DNA fragments, and allows for a single strand of DNA to enter the nanopore (10). The ionic current of each nucleotide is detected in the pore and is used for sequencing (10).

Another method for detecting ctDNA is using nucleosome occupancy profiles. Nucleosomes are the basic units of chromatin, which form the complex of DNA and chromosomes in eukaryotic cells. Each nucleosome consists of a segment of DNA wrapped around a core of histone proteins (25). The spacing and occupancy of nucleosomes along the DNA strand can have a significant impact on gene expression, transcription factor binding, and cellular function (25). Nucleosome occupancy refers to the frequency and distribution of nucleosomes along a particular region of the genome (3). High nucleosome occupancy corresponds to regions of DNA that are tightly packaged and inaccessible, called heterochromatin which are regions with low transcriptional activity. Conversely, low nucleosome occupancy often corresponds to regions that are more accessible and actively transcribed (25). Every tissue has specific patterns in which the cfDNA is packaged around the nucleosomes (3). Different nucleosome profiles exist because each tissue requires its DNA to interact with a different set of TFs (Transcription Factors) necessary for transcription (3). The regions for TF binding sites are not tightly packed with nucleosomes to allow for binding (3). At Euchromatin (open chromatin regions that contain genes for transcription and TF binding), DNA is exposed to endonuclease molecules and hence it is vulnerable to degradation (3). In these regions, nucleosome profile plots show loss of coverage, whereas these plots show high coverage at the Heterochromatin regions of DNA that are tightly bound to nucleosomes (3). Generating plots for cfDNA coverage can be used to detect TF binding sites. Since each cell which has a specific function, has different nucleosome packaging, the pattern of the cfDNA nucleosome coverage is indicative of the cell type and tissue of origin of the cfDNA molecule.

Katsman et.al. demonstrated that cfDNA fragments are primarily composed of mono-nucleosome fragments, although di-nucleosomes also exist in plasma samples (9). The positioning of the nucleosomes has shown to be cell-type specific, and the patterns are used to detect the cell of origin of the cfDNA (9). In fragmentation length analysis, ctDNA has been shown to have specific fragmentation features (9). These features include having a higher ratio of shorter mono-nucleosome cfDNA in cancer patients (100- 150 BP) than healthy mono-nucleosomes (100-220 base pairs) (9). Also, ctDNA contains a higher fraction of shorter di-nucleosomes (275-325 BP) than healthy di-nucleosomes (275-400 BP) (9). This is a biomarker in cancer that can be used to decipher cancerous versus healthy blood samples. Nucleosome occupancy plots can be generated to provide a pattern and detect differences between cancer and healthy blood samples(3).

Our R package, utilizes the fragmentation patterns and nucleosome occupancy of circulating cell-free DNA (cfDNA) to detect the presence of circulating tumor DNA (ctDNA) and identify cancerous blood samples. The package analyses the length distribution of cfDNA fragments and incorporates nucleosome occupancy analysis to identify characteristic changes in chromatin structure that are specific to cancer cells. The combination of nucleosome occupancy with fragmentation data analysis enhances the performance of our R package, as changes in nucleosome occupancy patterns and fragmentation have been shown to be associated with cancer development and progression. By leveraging the fragmentation patterns and nucleosome occupancy of cfDNA, cfDNAfragmentomics has the potential to advance the field of non-invasive cancer detection and personalized medicine, providing a reliable and cost-effective tool for early diagnosis and monitoring of cancer patients.

1.2 Current Technologies

Due to the great potential in tumor detection and subtyping using cfDNA, there are many technologies developed to use properties of cfDNA to perform analysis on cancer samples. Three tools, including DELFI (14), Griffin (3), and ichorCNA (used by Berman et. al., and developed by Adalsteinsson et. al) are explored below (24). These tools were used as inspiration for the cfDNAfragmentomics platform, to bring together different methods of analysing cfDNA from blood samples into one program that the user is able to input their bed files of their samples of interest, and receive statistics and graphs corresponding to their data. These output statistics can be used as a predictor for whether or not the liquid biopsy of interest is cancerous.

Methylation features and fragmentation patterns are both common cfDNA biomarkers in cancer detection (9). Analysis of fragmentation features alongside methylation allows for accurate detection of ctDNA and identification of the cell of origin (9). Technologies such as Griffin and DELFI use WGS to detect ctDNA and their tissues of origin (3, 14). Firstly, Griffin is a computational framework that uses nucleosome profiling to classify the tumor subtypes. (3). The tissue source of DNA is often detected from transcription regulation and nucleosome patterns because each tissue has a certain function, and based on that function, the accessibility of its DNA is modified for transcription (3). Griffin detects differences in chromatin accessibility and uses it to determine transcription regulation and the cell of origin of the DNA (3). Another technology in the fragmentation profile detection of cfDNA is called DELFI: a machine learning model that detects cancer and healthy cells with high sensitivity and specificity (sensitivity of 57-99, specificity of 98. Detected 91 of patients with cancer) (14). DELFI gives a score to the cfDNA to classify it as cancerous or healthy (14). For analysis with DELFI, blood is first collected from cancer patients and healthy individuals to train the machine learning model (14). Then, the cfDNA to be classified is given to the model to identify it as healthy or cancerous, and to detect the tissue of origin (14). The sensitivity of the model depends on the number of genomic and epigenomic alterations assessed (14). Hence, it is crucial to examine high numbers of abnormalities in the cfDNA when working with low coverage WGS (14, 16). Cristiano et. al. demonstrated that ctDNA fragment length is more variable compared to cfDNA because of the epigenomic, genomic, and chromosomal abnormalities in cancer (14). Also, cancer cells have altered nucleosome patterns and transcription start and end sites (14). These unique features in cancerous ctDNA are useful in DELFI to identify cancerous cells (14). ichorCNA was also used to perform cfDNA analysis. However, it uses statistical models to identify tumor fraction in patients. The three technologies are described in more detail below.

DELFI (DNA evaluation of fragments for early interception) The authors of the paper: Genome-wide cell-free DNA fragmentation in patients with cancer, developed a machine learning model to analyse the fragmentation pattern of cfDNA in cancer patients and used this method to identify tumor-specific fragmentation patterns. This approach allowed them to distinguish between tumor-derived and non-tumor-derived cfDNA, and to detect genomic alterations that are characteristic of cancer. The researchers analysed the cell-free DNA (cfDNA) from the blood of patients with different types of cancer, and compared it to the cfDNA of healthy individuals. The researchers used whole-genome sequencing to analyse the fragmentation pattern of cfDNA, and found that the cfDNA of cancer patients had an altered nucleosome pattern of white blood cells as compared to healthy individuals. Furthermore, the degree of fragmentation was found to be specific to the type of cancer, and was aberrant in cancer patients as compared to healthy cells, which suggests that the fragmentation pattern of cfDNA can be used as a biomarker for cancer diagnosis. DELFI showed to have high sensitivity as compared to other approaches in detecting cancer blood samples. Overall, the study provides important insights into the characteristics of cfDNA in cancer patients, and suggests that genome-wide cfDNA fragmentation analysis may have diagnostic and prognostic potential in cancer management. The findings of this study has implications for the development of non-invasive machine learning methods for cancer diagnosis and monitoring (14).

GRIFFIN The technology described in the paper by Doeably et.al. is a computational framework for analyzing nucleosome profiles of cell-free DNA (cfDNA) to classify cancer subtypes. The software uses machine learning algorithms to identify nucleosome fragmentation patterns associated with different cancer subtypes. The authors of the paper used a technique called ultra low-pass whole-genome sequencing to obtain nucleosome profiles from cfDNA samples collected from patients. They then used these profiles to train a machine learning model to classify the cancer samples into different subtypes. The Griffin framework has the potential to be used as a non-invasive diagnostic tool for cancer, because it can identify cancer subtypes from a less invasive blood test. This could be useful for patients who are unable to undergo invasive tumor biopsies, or for monitoring cancer progression and treatment response over time. Nucleosome profiling can be used to identify fragmentation patterns of different cancer subtypes because cancer cell have modifications in chromatin structure which can result in changes in nucleosome positioning and fragmentation patterns. The authors of the Griffin framework designed the algorithms to identify nucleosome fragmentation patterns associated with different cancer subtypes, and developed a model for subtyping cancer based on these patterns. The work presented in the Griffin paper is focused on developing a computational framework for subtyping cancer based on nucleosome profiling of cfDNA samples. This approach has the potential to improve cancer diagnosis and monitoring, particularly for patients who are unable to undergo invasive tissue biopsies (3).

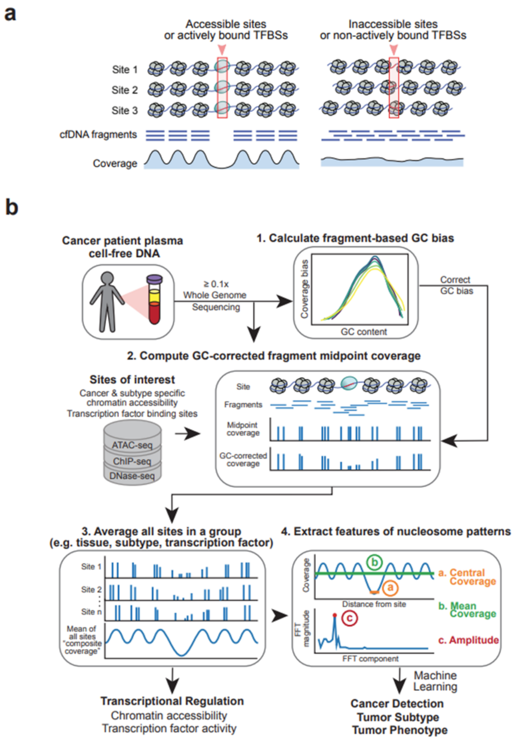


Fig. 1. We developed sequence coverage plots and TF binding sites taking inspiration from the pipeline described in Doebley et. al (Image from: Figure 1: Doebley et. al.)

ichorCNA and Berman Analysis ichorCNA is discussed by Adalstensson et. al. and is a software tool that identifies tumor cells from ultra Low Pass WGS of cfDNA data. ichorCNA uses a hidden Markov model (HMM) and a Bayesian statistical framework to predict Copy Number Alterations (CNAs) and to estimate tumor fraction of the cfDNA data. In the paper by Berman et. al., they create a a Fragment length density plot for the healthy and Lung cancer Adenocarcinoma (LUAD) cancerous cfDNA samples. The fraction of tumor cfDNA was then estimated using ichorCNA software (9, 24).

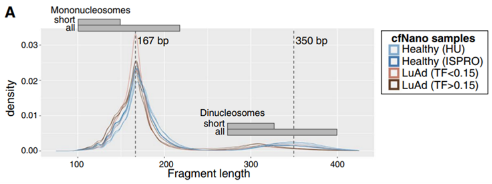


Fig. 2. We developed genome wide fragment size distribution for software validation of cancerous samples taking inspiration from ichorCNA and Berman et.al fragmentation analysis (Image from: Figure 4: Katsman et.al.)

2 Methods

Drawing influence from different tools for analyzing cancer cells using cfDNA such as DELFI, Griffin, and ichorCNA, we have created a comprehensive platform for cancer analysis that integrates some features of each tool. Our platform can be extended to provide an accurate and reliable prediction of cancer diagnosis and progression. We have incorporated novel features such as analysis of nucleosome occupancy and cfDNA fragmentation patterns to enhance the sensitivity and specificity of cancer detection, taking inspiration from different analysis methods, and clustering them to a user-friendly interface. The software to analyse tool was originally only developed in R. However, with larger datasets, python scripts were written to be run from the command line. We will discuss the development of our R package, python scripts, as well as the shiny app feature, the input of our tool, the output, and the results that are analysed and provided for the user.

In order to install the cfDNAfragmentomics R package, the following code must be run:

```
require("devtools")
devtools::install_github("Yasamin-Nourijelyani/CfDNAfragmentomics")
library("CfDNAfragmentomics")
```

To run the Shiny app:

```
runShinyCfDNAfragmentomics()
```

To get an overview of the available functions:

```
ls("package:CfDNAfragmentomics")
```

To view the documentation of each R function, run:

```
?<function_name>
# for example
?nucleosomeRatio
```

2.1 R Package Development

Our R package is designed with a focus on usability and functionality. One of the main features of the package is its well-documented functions, which provide clear and concise explanations of their usage, inputs, and examples. This makes it easy for users to quickly learn how to use the package and integrate it into their workflows. The package includes functions for analyzing bed files or tsv files, a common format for storing genomic data, allowing for analysis and manipulation of such files. Overall, the package is designed to make genomic analysis more accessible and user-friendly, helping researchers and clinicians to more effectively explore and understand their data.

The R package was developed using devtools, roxygen2, testthat, knitr, and shiny packages.

2.2 Shiny App

Our R Shiny app is designed to provide a user-friendly interface for comparing nucleosome length ratios between control and sample data. Note that nucleosome occupancy plots were not generated in the R shiny app and are created only using the python scripts. The significance of the difference in the fragmentation ratios will allow us to determine whether or not the sample of patient cfDNA are potentially cancerous. The app takes as input two tab-separated bed files containing data from the control and sample experiments. These files must include the fragment start and end positions as the second and third column of the tsv file which is standard practice in cfDNA data storage. The user can then specify the p-value that is deemed significant in their analysis, and the app will calculate the mono-nucleosome and di-nucleosome length ratios for both the control and sample data. The app presents the results in a plot, showing the distribution of nucleosome length ratios for both datasets (figure4), along with summary statistics of the Wilcoxon rank sum test and a boolean output answering the question: Is the sample provided considered to be cancerous? This app provides a simple and intuitive way to analyse nucleosome length ratios, and can be a useful tool for researchers studying epigenetic biomarkers in cancer.

2.3 Input Data

All data performed in this analysis are from the finaleDB website (20), and the test data for the shiny app come from the Alkdosi et. al. R package (18). The queries from finaleDB are lung cancer and healthy samples originating from the paper by Snyder et.al. and Cristiano et.al. These data are downloaded and used in the command line for analysis. The data downloaded from finaleDb is Fragment tsv files using hg38 latest assembly of the human genome (20).

2.4 Processing of Data

The data that is inputted to cfDNAfragmentomics is analysed for fragmentation features and nucleosome occupancy. These are potential cancer biomarkers and can be used to detect ctDNA in the blood.

Fragment Analysis In our R package, we utilized the Wilcoxon rank sum test to compare mononucleosome and dinucleosome fragment lengths in cancerous samples against those in healthy samples. The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is a non-parametric test used to determine significance of the difference in fragment size of the control and patient cfDNA lengths to determine if the patient data contains a cancer biomarker (19). The Wilcoxon statistical test is particularly useful for handling non-normally distributed data because it does not make any assumptions about the distribution of the data, making it ideal to analyse fragment lengths for cfDNA (19). Another advantage is that it does not require equal variances between the two populations, unlike some parametric tests such as the t-test (19). By comparing fragment lengths between cancerous and healthy samples, we can identify significant differences in fragmentation patterns that are indicative of cancerous liquid biopsies. This analysis is a crucial step in identifying potential cancer biomarkers that can be used to develop more accurate and effective cancer detection strategies. By utilizing the Wilcoxon rank sum test in our package, we are able to provide a powerful statistical tool for the detection and analysis of cfDNA fragmentation patterns. Here, we are comparing two independent samples, which are the population of healthy cfDNA data and population of patient cfDNA data, and we are making inference about the state of being cancer positive or negative for the population of cfDNA molecules in the patient. Both control and patient data inputted to the functions are assumed to be reads for the specific loci corresponding to the cancer type of interest. This analysis also assumes that the data is real human data and contains both mono-nucleosome and di-nucleosome length data. Similar to our analysis, in the study by Katsman et al., the Wilcoxon rank sum test was used to identify variability of methylation of cfDNA between cancer patients and healthy controls (9).

To perform fragmentation length analysis, the R shiny interface takes as input the sample and control data as a bed file or a tsv file (tab separated file). The sample data that is inputted contains potentially cancerous blood samples from the patient which we want to perform statistical analysis on, to determine whether or not the given sample is from a cancerous source. A control bed file is also passed into the function which is cfDNA from a healthy blood sample used to compare the mono-nucleosome and di-nucleosome fragment lengths with the patient sample. The Wilcoxon rank sum test is then performed to determine if the sample patient data has significantly shorter cfDNA mononucleosome and dinucleosome fragment lengths as compared to the control. If sample patient cfDNA has significantly shorter mononucleosome and dinucleosome fragment lengths, than the sample data output returns to the user that the sample likely comes from a cancerous patient. Threshold p-values are also passed into the function as optional parameters to output a decision variable output identifying whether or not the sample patient fragment lengths are significantly shorter in mononucleosome and dinucleosome ratios.

Due to the limitation of the size of the files that can be inputted into the R Shiny app, a command line script using python was developed for analysing large bed file fragment length ratios coming from the finaleDB database. The script is called nucratio.py, and is used to perform the same Wilcoxon rank sum test analysis as the

R shiny interface to compare the mononucleosome and dinucleosome fragmentation lengths. The analysis also determines whether or not the mono-nucleosome and di-nucleosome lengths are shorter in the patient as compared to the healthy data. If yes, then the output mentions that the user can check the p-value results. If significant, then the mono-nucleosome or di-nucleosome lengths are significantly shorter in patient as compared to the healthy cfDNA data, which indicates that the patient samples likely come from a cancer source.

The analysis can be performed using the command line code:

```
python <./location/nuc_ratio.py> <control_data.tsv> -s <sample_data.tsv>
```

The input parameters to the command are the healthy cfDNA samples, and the $-s$ parameter are the patient cfDNA fragmentation data.

Nucleosome Occupancy The analysis of the nucleosome occupancy is performed by the nucleosome-occupancy.py python script.

To run the script, put the following code is inputted to the command line:

```
python <./location/nucleosome_occupancy.py> -t <./location/TFBS_loc.tsv>
-w <Window size> -n <name> <./locationdata/gene.hg38.frag.bed>
```

for instance, in the directory

```
~/ynourijelyani/CfDNAfragmentomics/dev
```

We can run:

```
python ../python/nucleosome_occupancy.py -t TFBS_loc.tsv
-w 500 -n example data/EE87865.hg38.frag.tsv
```

The files that are used as the input parameter bed files from finaleDB come from the lung cancer or healthy samples from Snyder et.al. 2016 and Cristiano et.al. 2019 data files. These data are in .bgz format. To make the data usable by this function, it must be processed into bed files in the command line:

```
mv <filename.bgz> <filename.gz>
gunzip <filename.gz>
mv <filename.tsv> <filename.bed>
```

This way, bed files are generated from the finaleDB data, and can be used for the nucleosome coverage plotting analysis. This analysis takes as input the transcription factor binding site (TFBS) locations from the csv file retrieved from the paper by Fang et.al., additional file 4 (17). This file provides the midpoint of the TFBS and is used to aggregate coverage of the cfDNA input data. Note that the TFBS data for the specific cancer type being analysed should be used.

The script also takes as input a bed file containing genomic intervals, and uses this to calculate the coverage at every locus within the genome, given the midpoint of the active transcription factor binding site from the Fang et. al. TFBS locations and a specific window size (17). The window size is the number of base pairs that the coverage is aggregated to the left and right of the TFBS midpoint location. The coverage is stored in a dictionary where each chromosome is a key and the corresponding value is a numpy array representing the coverage at each position along the chromosome given the TFBS location. For every TFBS we aggregate the fragment coverage in a surrounding window. The size of the window is specified by the user as an argument (i.e. 1000bp) (16). The script uses the aggregated coverage vector to plot the coverage of the nucleosome occupancy in a line plot (16). The plot is smoothed using a Savitzky-Golay filter before being saved as a PNG file as seen in figure6 and figure7.

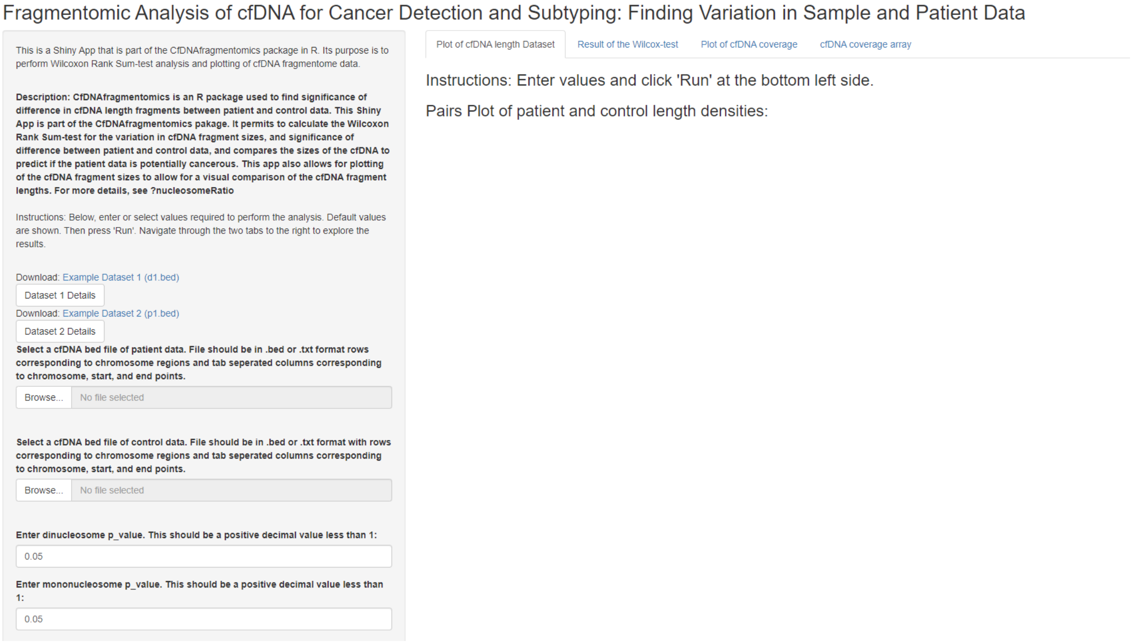
3 Results

Fragment Analysis The fragmentation analysis in our R shiny package was initially performed using small test data samples from Alkodsi et. al. package (18), located in the /inst/extdata folder to demonstrate the results from the cfDNAfragmentomics R shiny package (fig 3, 4, 5). For instance, figure 4 shows some plot generated from running p1.bed for patient data and d1.bed for control data in the shiny app. However, to validate the robustness and applicability of the package to larger datasets, we also used our method to analyse the nucleosome fragmentation patterns with the Snyder et.al. data from the Finale DB database. For this, we employed the Python nuc-ratio.py script to compare nucleosome length distributions in the healthy and cancer samples using a Wilcoxon rank sum test. The integration of these different methods and data sets in our analysis showcases the versatility and usefulness of our package in diverse genomics applications.

The python script takes as input the control and patient sample data as bed files, calculates the length of the fragments, separates the lengths to mono-nucleosome and di-nucleosome arrays, and compares the control fragment lengths with the patient cfDNA fragment lengths. The sample output includes a description of the length of the patient data and determines if mono-nucleosome or the di-nucleosome fragment lengths of the patient data are shorter than the control. The results also show the p-value and test statistics of the comparison to determine whether or not the difference is statistically significant (figure 6).

Nucleosome Occupancy The python script nucleosome-occupancy.py was used to develop the nucleosome coverage plots. This script uses the patient cfDNA bed file to obtain the coverage at each locus in the specified interval, and then smooths the coverage values using a Savitzky-Golay filter. The script then aggregates the coverage vectors across every window in the transcription factor binding sites (TFBS) file to provide an aggregated coverage vector. Finally, it plots the nucleosome occupancy line plot using the matplotlib library. By using this script, we were able to generate informative and visually appealing nucleosome occupancy plots for our analysis.

In figure 67, the unique fingerprint of the nucleosome occupancy for lung cancer from the Cristiano et.al. data is demonstrated (20). Also, figure 8 represents the same nucleosome occupancy plot using lung cancer data from Snyder et. al.(20). The dips in the plot demonstrate the lack of nucleosome coverage which causes the cfDNA fragments to be susceptible to nuclease enzymes which lead to low coverage of the data in those regions. The peaks represent high nucleosome coverage due to the tight nucleosome binding, resulting in euchromatin regions.The TFBS locations we used were



Fragmentomic Analysis of cfDNA for Cancer Detection and Subtyping: Finding Variation in Sample and Patient Data

This is a Shiny App that is part of the CfDNAfragmentomics package in R. Its purpose is to perform Wilcoxon Rank Sum-test analysis and plotting of cfDNA fragmentome data.

Description: CfDNAfragmentomics is an R package used to find significance of difference in cfDNA length fragments between patient and control data. This Shiny App is part of the CfDNAfragmentomics pakage. It permits to calculate the Wilcoxon Rank Sum-test for the variation in cfDNA fragment sizes, and significance of difference between patient and control data, and compares the sizes of the cfDNA to predict if the patient data is potentially cancerous. This app also allows for plotting of the cfDNA fragment sizes to allow for a visual comparison of the cfDNA fragment lengths. For more details, see ?nucleosomeRatio

Instructions: Below, enter or select values required to perform the analysis. Default values are shown. Then press 'Run'. Navigate through the two tabs to the right to explore the results.

Download: Example Dataset 1 (d1.bed)

Download: Example Dataset 2 (p1.bed)

Select a cfDNA bed file of patient data. File should be in .bed or .txt format rows corresponding to chromosome regions and tab seperated columns corresponding to chromosome, start, and end points.

Select a cfDNA bed file of control data. File should be in .bed or .txt format with rows corresponding to chromosome regions and tab seperated columns corresponding to chromosome, start, and end points.

Enter dinucleosome p_value. This should be a positive decimal value less than 1: 0.05

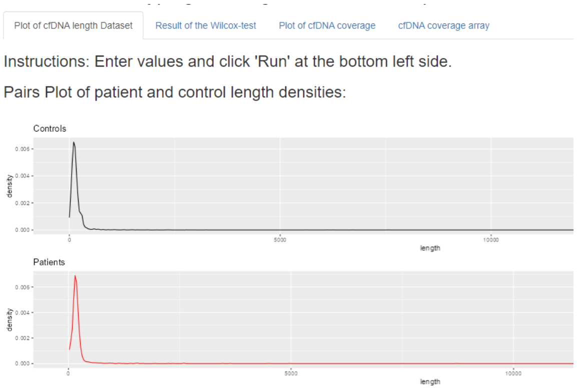
Enter mononucleosome p_value. This should be a positive decimal value less than 1: 0.05

Instructions: Enter values and click 'Run' at the bottom left side.

Pairs Plot of patient and control length densities:

Fig. 3. R shiny app interface for running the files

from the data in the additional file 4 of the Fang et.al. paper (17). The peaks and dips in the plot create a unique fingerprint of the tissue of origin for the cancer samples, which allows for subtyping of the samples and identifies the type of cancer corresponding to the coverage plots. In both figures 7 and 8, peaks of nucleosome occupied regions are seen at every 150 base pairs as expected.



Instructions: Enter values and click 'Run' at the bottom left side.

Pairs Plot of patient and control length densities:

Fig. 4. Visualization of fragment ratios, where the first peak represents mononucleosome fragment lengths and the second peak, if exists, is dinucleosome fragment lengths. Sample and control data plots using the data from the ctDNAtools package (18), used to visually compare the healthy and sample data. These plots were generated following the plots that were developed in the paper by Berman et.al. seen in figure2

4 Discussion

There are several positive aspects of our cancer analysis package, cfDNAfragmentomics. Firstly, the integration of different techniques and algorithms from existing tools such as DELFI, Griffin, and ichorCNA allows our package to benefit from the strengths of each approach. This results in a comprehensive and robust platform for cancer analysis. Additionally, our package includes innovative features such as the analysis of nucleosome occupancy and DNA fragmentation patterns, which can enhance the sensitivity and specificity of cancer detection. Another positive aspect of our package is its user-friendly interface developed using the R shiny app, which makes it accessible to researchers and clinicians with varying levels of expertise. By offering a powerful yet easy-to-use tool for cancer analysis, our package has the potential to accelerate progress in the field of cancer research and improve patient outcomes.

The code for this package uses clean architecture techniques to ensure that it is maintainable in the future. Clean architecture emphasizes the separation of concerns, with each layer of the system responsible for a specific aspect of the software (26). In our package, the core logic is separated from the user interface. The core logic layer contains the business logic and the algorithms used for cfDNA fragmentation analysis, while the user interface layer contains the code responsible for interacting with users in the R shiny app.

By separating these concerns, our code becomes modular, making it easier to test, maintain and extend. Additionally, clean architecture helps reduce dependencies between different components of the system, making it easier to make changes to the code without impacting other parts of the software. Our code also follows the Single Responsibility Principle (SRP), which is a key aspect of clean architecture (26). Each function in the code has only one responsibility, which helps to keep the code easy to maintain. The use of clean architecture allows our tool to be easily extended to analyse more cfDNA fragment features and epigenetic alterations of cfDNA such as methyla-

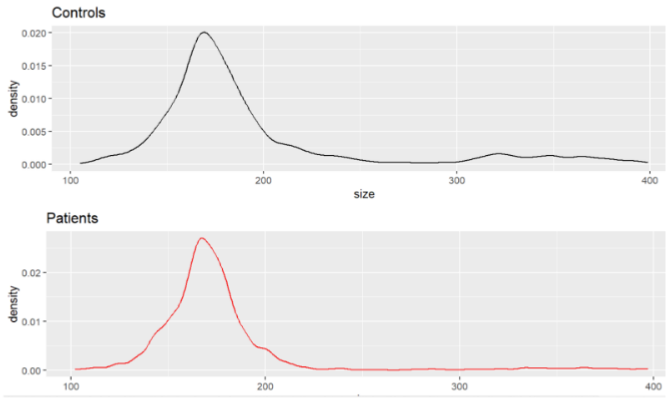
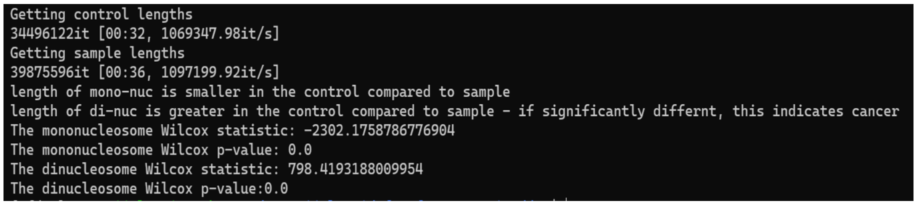


Fig. 5. Visualization of fragment ratios, where the first peak represents mononucleosome fragment lengths. Sample and control data plots using the data from the ctDNAtools package (18). Used to visually compare the healthy and sample data. These plots follow the plots that were developed in the paper by Berman et.al. seen in figure2 using different test data as compare to figure4

tion and fragment end features in the future developments of the package, to increase the sensitivity and reduce the likelihood of false positives in the data.

One potential weakness of our package is that it utilizes naive statistical models and does not incorporate machine learning algorithms. While statistical models are powerful tools for data analysis, they may not be as effective at identifying complex patterns in large cfDNA datasets. By not including machine learning algorithms in our package, we may be limiting its ability to accurately predict complex cancer epigenomic features and outcomes. Additionally, R is not powerful enough to analyse very large cfDNA datasets using its interface. Hence, the large data files cannot be processed by the R scripts. Python files were developed to ensure that the user can obtain output for larger datasets such as the data coming from the finaleDB database. However, our package is still useful for certain types of cancer analysis and may provide a valuable contribution to the field of cancer research. Overall, we are optimistic about the potential of our package to make a positive impact in analysing cfDNA data.



```
Getting control lengths
34496122it [00:32, 1069347.98it/s]
Getting sample lengths
39875596it [00:36, 1097199.92it/s]
length of mono-nuc is smaller in the control compared to sample
length of di-nuc is greater in the control compared to sample - if significantly differnt, this indicates cancer
The mononucleosome Wilcox statistic: -2302.1758786776904
The mononucleosome Wilcox p-value: 0.0
The dinucleosome Wilcox statistic: 798.4193188009954
The dinucleosome Wilcox p-value:0.0
```

Fig. 6. Output results of the nuc-ratio.py analysis using healthy and Lung cancer data from Snyder et. al. (20) The results show that the di-nucleosome ratios are significantly shorter in the patient cfDNA, which indicates that if the p-values are small, then the result is likely statistically significant. Since the p-values are very small in this analysis, it indicates that the patient cfDNA is likely from a cancer source.

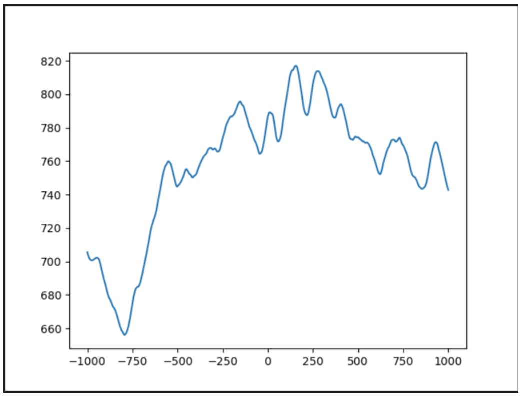


Fig. 7. Nucleosome Occupancy plots for lung cancer data from Christiano et.al finaleDB for sample EE88183.hg38.frag. Peaks are detected at every 150 base pairs as expected.

5 Conclusion

In conclusion, cfDNAfragmentomics package which uses fragmentation and nucleosome coverage analysis is a promising approach for cancer detection, and our R package provides a powerful tool for researchers and clinicians working in this area. By using innovative techniques such as nucleosome occupancy analysis and DNA fragmentation pattern analysis, our package provides a comprehensive platform for the detection and analysis of ctDNA and cancerous blood samples. Although our package may have limitations, such as its reliance on statistical models rather than machine learning algorithms, it has the potential to make a significant contribution to the field of cancer research. As cfDNA data slowly leads its way to clinical settings and starts being used by clinicians to analyse patient liquid biopsies for detecting potentially cancerous blood samples, we hope that this package inspires researchers to create user-friendly and easy to use packages for the analysis of cfDNA data. These packages should involve great documentation, an interactive app, as well as command line features to allow for easy use and analysis geared towards people who are not familiar with data manipulation. We hope to develop more powerful tools for analysing cfDNA fragments in the future

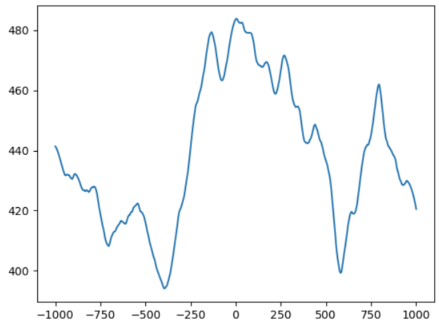


Fig. 8. Nucleosome Occupancy plots for lung cancer data from Snyder et.al finaleDB for sample EE86229.hg38.frag.bed. Peaks are detected at every 150 base pairs as expected.

such as algorithms which use better analysis techniques, include more cfDNA cancer hallmarks such as methylation patterns for more sensitive detection, and make the analysis more optimized in terms of speed. Overall, we are optimistic that this package will help accelerate progress in the accessibility and availability of cancer detection tools, ultimately improving outcomes for patients worldwide and easing the process of early cancer detection.

6 Data Availability

Data is accessed from of lung cancer and healthy from Snyder et.al 2016 paper available on finaleDB and from Christiano et.al. 2019 also available on finaleDB (20). Also, TFBS data is available from the paper by Fang et.al. (17) under additional files 4.

7 Code Availability

All code is available from https://github.com/Yasamin-Nourijelyani/CfDNAfragmentomics.git

8 References

1. Cree, I. A., Uttley, L., Buckley Woods, H., Kikuchi, H., Reiman, A., Harnan, S., Whiteman, B. L., Philips, S. T., Messenger, M., Cox, A., Teare, D., Sheils, O., Shaw, J., UK Early Cancer Detection Consortium (2017). The evidence base for circulating tumour DNA blood-based biomarkers for the early detection of cancer: a systematic mapping review. BMC cancer, 17(1), 697. https://doi.org/10.1186/s12885-017-3693-7
2. Lan Zhao, Victor H F Lee, Michael K Ng, Hong Yan, Maarten F Bijlsma, Molecular subtyping of cancer: current status and moving toward clinical applications, Briefings in Bioinformatics, Volume 20, Issue 2, March 2019, Pages 572–584, https://doi.org/10.1093/bib/bby026
3. Anna-Lisa Doebley, Minjeong Ko, Hanna Liao, A. Eden Cruikshank, Caroline Kikawa, Katheryn Santos, Joseph Hiatt, Robert D. Patton, Navonil De Sarkar, Anna C.H. Hoge, Katharine Chen, Zachary T. Weber, Mohamed Adil, Jonathan Reichel, Paz Polak, Viktor A. Adalsteinsson, Peter S. Nelson, Heather A. Parsons, Daniel G. Stover, David MacPherson, Gavin Ha.(2021) Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA, medRxiv 2021.08.31.21262867; doi: https://doi.org/10.1101/2021.08.31.21262867
4. Bouzidi, A., Labreche, K., Baron, M., Veyri, M., Denis, J. A., Touat, M., Sanson, M., Davi, F., Guillerm, E., Jouannet, S., Charlotte, F., Bielle, F., Choquet, S., Boëlle, P. Y., Cadranel, J., Leblond, V., Autran, B., Lacorte, J. M., Spano, J. P., Coulet, F., . . . IDEATION study group (2021). Low-Coverage Whole Genome Sequencing of Cell-Free DNA From Immunosuppressed Cancer Patients Enables Tumor Fraction Determination and Reveals Relevant Copy Number Alterations. Frontiers in cell and developmental biology, 9, 661272. https://doi.org/10.3389/fcell.2021.661272
5. Chin, R. I., Chen, K., Usmani, A., Chua, C., Harris, P. K., Binkley, M. S., Azad, T. D., Dudley, J. C., Chaudhuri, A. A. (2019). Detection of Solid Tumor Molecular Residual Disease (MRD) Using Circulating Tumor DNA (ctDNA). Molecular diagnosis therapy, 23(3), 311–331. https://doi.org/10.1007/s40291-019-00390-5
6. D'Arcy M. S. (2019). Cell death: a review of the major forms of apoptosis, necrosis and autophagy. Cell biology international, 43(6), 582–592. https://doi-org.myaccess.library.utoronto.ca/10.1002/cbin.11137
7. Liu, Z. G., Jiao, D. (2019). Necroptosis, tumor necrosis and tumorigenesis. Cell stress, 4(1), 1–8. https://doi-org.myaccess.library.utoronto.ca/10.15698/cst2020.01.208
8. Altman, B. J., Rathmell, J. C. (2012). Metabolic stress in autophagy and cell death pathways. Cold Spring Harbor perspectives in biology, 4(9), a008763. https://doi-org.myaccess.library.utoronto.ca/10.1101/cshperspect.a008763
9. Katsman, E., Orlanski, S., Martignano, F., Fox-Fisher, I., Shemer, R., Dor, Y., Zick, A., Eden, A., Petrini, I., Conticello, S. G., Berman, B. P. (2022). Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. Genome biology, 23(1), 158. https://doi-org.myaccess.library.utoronto.ca/10.1186/s13059-022-02710-1
10. Jain, M., Olsen, H. E., Paten, B., Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology, 17(1), 239. https://doi-org.myaccess.library.utoronto.ca/10.1186/s13059-016-1103-0
11. Modi, A., Vai, S., Caramelli, D., Lari, M. (2021). The Illumina Sequencing Protocol and the NovaSeq 6000 System. In: Mengoni, A., Bacci, G., Fondi, M. (eds) Bacterial

Pangenomics. Methods in Molecular Biology, vol 2242. Humana, New York, NY. https://doi-org.myaccess.library.utoronto.ca/10.1007/978-1-0716-1099-22

12. Sequencing short fragments with Nanopore Technology. Oxford Nanopore Technologies. (2022, August 9). Retrieved October 12, 2022, from https://nanoporetech.com/applications/short-fragment-mode
13. Delahaye, C., Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. PloS one, 16(10), e0257521. https://doi.org/10.1371/journal.pone.0257521
14. Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. ., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., Anagnostou, V., Forde, P., Naidoo, J., Marrone, K., Brahmer, J., Woodward, B. D., Husain, H., van Rooijen, K. L., … Velculescu, V. E. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 570(7761), 385–389. https://doi-org.myaccess.library.utoronto.ca/10.1038/s41586-019-1272-6
15. Ko, K., Kananazawa, Y., Yamada, T., Kakinuma, D., Matsuno, K., Ando, F., Kuriyama, S., Matsuda, A., Yoshida, H. (2021). Methylation status and long-fragment cell-free DNA are prognostic biomarkers for gastric cancer. Cancer medicine, 10(6), 2003–2012. https://doi-org.myaccess.library.utoronto.ca/10.1002/cam4.3755
16. Broadbent, J. Personal Communication, October, 2022-April 2023, University of Toronto.
17. Fang, C., Wang, Z., Han, C. et al. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. Genome Biol 21, 247 (2020). https://doi.org/10.1186/s13059-020-02152-7
18. Alkodsi A, Meriranta L, Pasanen A, Leppä S (2020). “ctDNAtools: An R package to work with sequencing data of circulating tumor DNA.” bioRxiv.
19. “Mann Whitney U Test in R Programming.” GeeksforGeeks, 28 Dec. 2021, https://www.geeksforgeeks.org/mann-whitney-u-test-in-r-programming/.
20. FinaleDB, http://finaledb.research.cchmc.org, Cristiano et.al. Snyder et.al
21. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660
22. American Cancer Society. Cancer Facts Figures 2021. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf. Accessed March 24, 2023.
23. Myers, E. R., Moorman, P., Gierisch, J. M., Havrilesky, L. J., Grimm, L. J., Ghate, S., Davidson, B., Mongtomery, R. C., Crowley, M. J., McCrory, D. C., Kendrick, A., Sanders, G. D. (2015). Benefits and Harms of Breast Cancer Screening: A Systematic Review. JAMA, 314(15), 1615–1634. https://doi.org/10.1001/jama.2015.13183
24. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, Gydush G, Reed SC, Rotem D, Rhoades J, Loginov D, Livitz D, Rosebrock D, Leshchiner I, Kim J, Stewart C, Rosenberg M, Francis JM, Zhang CZ, Cohen O, Oh C, Ding H, Polak P, Lloyd M, Mahmud S, Helvie K, Merrill MS, Santiago RA, O’Connor EP, Jeong SH, Leeson R, Barry RM, Kramkowski JF, Zhang Z, Polacek L, Lohr JG, Schleicher M, Lipscomb E, Saltzman A, Oliver NM, Marini L, Waks AG, Harshman LC, Tolaney SM, Van Allen EM, Winer EP, Lin NU, Nakabayashi M, Taplin ME, Johannessen CM, Garraway LA, Golub TR, Boehm JS, Wagle N, Getz G, Love JC, Meyerson M. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017 Nov 6;8(1):1324. doi: 10.1038/s41467-017-00965-y. PMID: 29109393; PMCID: PMC5673918.

25. Tamaru H. Confining euchromatin/heterochromatin territory: jumonji crosses the line. Genes Dev. 2010 Jul 15;24(14):1465-78. doi: 10.1101/gad.1941010. PMID: 20634313; PMCID: PMC2904936.
26. Martin, Robert C. Clean Architecture: a Craftsman's Guide to Software Structure and Design. London, England: Prentice Hall, 2018.