AMR FAROOQ

Special thanks to
Rami Naji, David Peta, and Franscizek Latala

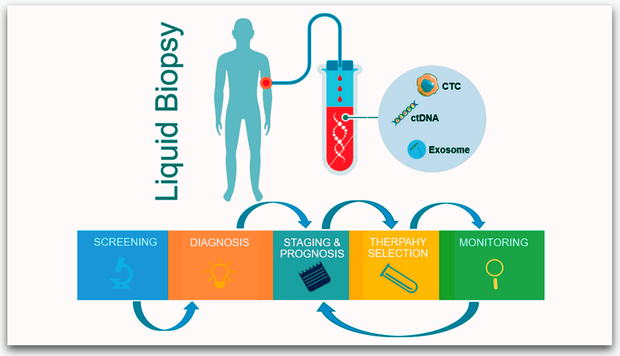
DETECTION OF CANCER USING BLOOD

Quantifying complementarity between different cfDNA features

TU Delft

SUPERVISORS

Marcel Jt T. Reinders, Stavros Makrodimitris, Bram Pronk, Daan Hazelaar



Kulasinghe, A., Wu, H., Punyadeera, C., & Warkiani, M. E. (2018). TClinical applications of liquid biopsy from blood circulating markers. The genomics and immunology information derived from liquid biopsy samples can be used for continuous monitoring, from early stage disease screening, assistance diagnosis, personalized therapy selection, to recurrence monitoring. CTC—circulating tumor cells; ctDNA—circulating tumor DNA.. https://doi.org/10.3390/mi9080397

01. BACKGROUND

Recent research has indicated attributes of cell-free DNA (cfDNA) called fragmentomics as a promising method for late stage cancer detection in a non-invasive manner.

The primary objective of this research is to uncover hidden patterns and interactions that could enhance the accuracy and sensitivity of blood-based cancer diagnostics (Liquid Biopsies).

This study explores he complementarity between three fragmentomics features; fragment length distribution, and nucleotide fragment end sequence diversity and nucleosome positioning for four different sample groups; breast cancer (BRCA) , colorectal cancer (CRC) , lung cancer (LUAD) and healthy controls.

Various machine learning techniques such as linear regression were employed to quantify any complementary relationships between the features

02. RESEARCH QUESTION

Explore the complementarity of various fragmentomics features

03. SETUP

The research is divided into three parts:

- **Identification**
  - Identifying which fragmentomics features to use
  - Extracting the features from the data set
- **Processing**
  - Find the most appropriate manner to combine feature values for all samples for the same dataset
- **Evaluation**
  - Selecting specific metrics from the processed data to assess the complementarity of the identified features.

04. IDENTIFICATION

LOG2(SHORT-LONG RATIO) OF FRAGMENT LENGTHS [2]

The fragment length ratio is calculated as:

$$ratio = \log_2\left(\frac{short\_count}{long\_count}\right)$$

$short\_count$ = number of short fragments (100-150 bp)
$long\_count$ = number of long fragments (150-220 bp)

All ratios were standardized using z-scores.

THE TRINUCLEOTIDE FRAGMENT END SEQUENCE DIVERSITY

- **Divide the genome into 5Mb bins (per chromosome) and count the trinucleotide frequency**
- **Trinucleotide:**
  - **A sequence of three consecutive nucleotides.**
  - **DNA is made up of four types of nucleotides: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G).**
  - **A trinucleotide could be any combination of these, like AAA, CCT, or GTC.**
- **Use the Gini index to get a value G which is a measure of statistical dispersion for every bin.**

The 5' trinucleotide fragment end sequence diversity is calculated for every input sample as the Gini index using the formula:

$$G = 1 - \sum_{i=1}^{64} P_i^2$$

where P_i is the frequency of a specific i trinucleotide endings [5].

Chromosome	Bin Start	Bin End	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC
chr1	0	5000000	251594	136372	198380	161128	244989	187726	62349	156462	239963	239703
chr1	5000000	10000000	414184	196886	280607	251944	305098	213175	57883	223899	321862	265495
chr1	10000000	15000000	388912	183213	271966	252019	265607	197661	47510	219402	311632	247405

TGA (...	TGC (...	TGG (...	TGT (...	TTA (...	TTC (...	TTG (...	TTT (...
TGA	TGC	TGG	TGT	TTA	TTC	TTG	TTT
225157	233607	321467	240797	127960	196058	184676	264128
281205	256475	345974	280456	199353	263939	246810	395759
289728	248980	346553	291860	209960	269310	276116	433589

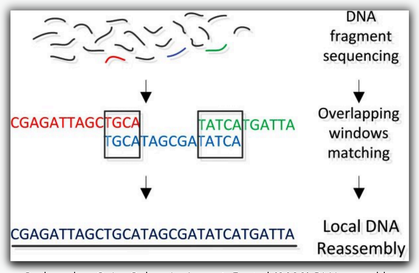
NUCLEOSOME POSITIONING PATTERNS

- **Observed using a numerical value known as The Windowed Protection Score (WPS).**
- **Defined as the difference between the number of DNA fragments completely spanning a 120 bp window centered at a given genomic coordinate and the number of fragments with an endpoint within that same window [3]**

```
FUNCTION calculate_wps(sample):
    INITIALIZE an empty list wps_data to store WPS results

    FOR each chromosome in sample:
        Determine the length of the chromosome

        FOR each 5Mb bin in the chromosome:
            INITIALIZE an empty list wps_values to store WPS for
               positions within the bin

            FOR each base pair position within the bin:
                INITIALIZE spanning_count and endpoint_count to 0

                Define a 120 bp sliding window centered at the current
                   position

                FOR each read within the 120 bp window:
                    if the read is properly mapped and is a proper pair:
                        Determine the fragment start and end positions

                        if the fragment completely spans the window:
                            Increment spanning_count

                        if the fragment has an endpoint within the window
                            :
                            Increment endpoint_count

                Calculate WPS for the current position as spanning_count
                   - endpoint_count
                Append the WPS value to wps_values

            Calculate the average WPS for the current bin from wps_values
            Store the chromosome, bin start, bin end, and average WPS in
               wps_data

    RETURN wps_data
```

05. PROCESSING

- *EXTRACT FEATURES FROM DATA:*
- *DATA WAS BINNED INTO NON OVERLAPPING WINDOWS OF 5-MEGABASE (MB)*
- *OVER 500 FEATURES OF A TYPE PER SAMPLE*
- *ONE FEATURE - PER SAMPLE PER BIN E.G.*
- *CHR1:0-5000000, ONE MEASUREMENT*
- *ALL FEATURE VALUES FOR A DATASET COMBINED INTO ONE FEATURE MATRIX: PER DATATSET PER FEATURE TYPE*



Pavlopoulos, G.A., Oulas, A., Iacucci, E. et al (2013) DNA assembly https://doi.org/10.1186/1756-0381-6-13

Sample Name	Chr1 0:5000000	Chr1 5000000:10000000	...	Chr22 ...
Sample 1	value 1	value 2	...	value n
Sample 2	value 3	value 4	...	value m
Sample 3	value 5	value 6	...	value o
...	...	...	...	...
Sample N	value x	value y	...	value z

06. EVALUATION

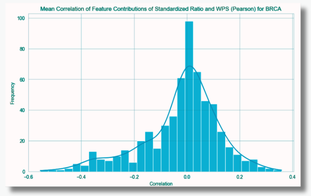
ALGORITHIMS

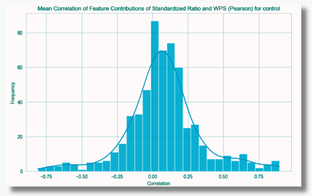
Linear Regression
Use Linear regression to confirm if its possible to predict one feature from another

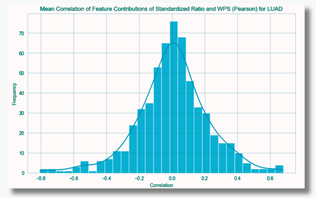
Multi-Omics Factor Analysis (MOFA+)
MOFA+ exploits the dependencies between the features to create a simplified representation of the larger dataset defined by multiple latent factors.
These factors capture the global sources of variability in the data [5].
Each factor has weights that highlight how important each feature is in determining the factor's value.
MOFA+ can use these factors to determine which features contribute to the same latent factor thus, indicating relationships like complementarity.

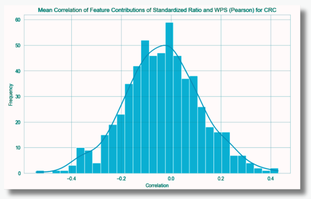
07. RESULTS & CONCLUSIONS

SHORT-LONG RATIO OF THE FRAGMENT LENGTHS AND WPS





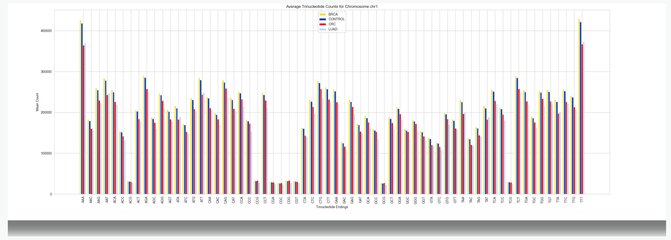


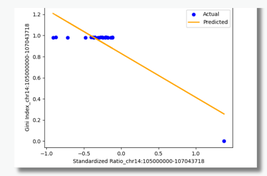


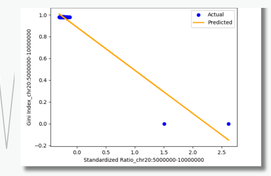
- Across all four groups - BRCA, Healthy controls, CRC and LUAD, our findings consistently indicated that the two feature types were largely independent from each other, suggesting that short-long ratios and the WPS do not significantly influence one another.
- Mmajority of the correlation values are centered around zero for all groups, suggesting a poor linear relationship between short-long ratios and the WPS.
- **<u>WE ASSERT THAT THESE TWO FEATURE TYPES EXHIBIT A HIGH DEGREE OF COMPLEMENTARITY, AS THEY PROVIDE UNIQUE AND NON-OVERLAPPING INFORMATION.</u>**

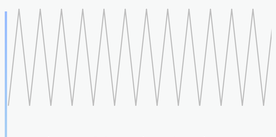
SHORT-LONG RATIO OF THE FRAGMENT LENGTHS AND 5' TRINUCLEOTIDE FRAGMENT END SEQUENCE DIVERSITY

On average counts per trinucleotide ending are close to identical per sample group for each chromosome with endings such as AAA, and TTT regularly having large counts and TCG and CGA consistently showing low counts.









For the BRCA and LUAD datasets we consistently saw low R-Squared scores throughout indicating that the two features are very independent from each other.

The healthy control and CRC datasets some chromosomes had large scores.
We observe a consistent negative correlation between the two feature types across both the CRC and healthy control datasets.

However, the actual values for the Gini index are all closely clustered together, It is challenging to draw conclusions why this feature is so uniform from a purely computer science prospective.

<u>WE THEREFORE CONCLUDE OUR INVESTIGATE AND ADJUDICATE THE GINI INDEX [6] DOES NOT CAPTURE UNIQUE FRAGMENTOMIC FEATURES, AND SHOULD NOT PURSED AS A BIOMARKER FOR THE DETECTION OF CANCER.</u>

08. REFERENCES

1. Kulasinghe A, Wu H, Punyadeera C, Warkiani ME. The Use of Microfluidic Technology for Cancer Applications and Liquid Biopsy.*Micromachines*.2018;9(8):397.https://doi.org/10.3390/mi9080397
2. Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, et al. Genome-wide cell-free dna fragmentation in patients with cancer. Nature, 570(7761):385–389, 2019
3. Matthew W Snyder, Martin Kircher, Andrew J Hill, Riza M Daza, and Jay Shendure. Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell, 164(1):57–68, 2016
4. Pavlopoulos, G.A., Oulas, A., Iacucci, E. et al. Unraveling genomic variation from next generation sequencing data. BioData Mining 6, 13 (2013). https://doi.org/10.1186/1756-0381-6-13
5. Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome biology, 21:1–17, 2020.
6. Norbert Moldovan, Ymke van der Pol, Tom van den Ende, Dries Boers, Sandra Verkuijlen, Aafke Creemers, Jip Ramaker, Trang Vu, Sanne Bootsma, Kristiaan J Lenos, et al. Multi-modal cell-free dna genomic and fragmentomic patterns enhance cancer survival and recurrence analysis. Cell Reports Medicine, 5(1), 2024