Enzymatic enrichment of unmethylated cell-free DNA improves the sensitivity of fragmentomic analysis for the diagnosis of patients with lung cancer

Valentina Miano[1]*, Paulina Siejka-Zielinska[1]*, Calum Mould[1], Seamus Grundy[2], Sean Knight[3], Robert K. Neely[1,4]
[1] Tagomics Ltd, Cambridge UK, [2] Northern Care Alliance NHS Foundation Trust, Manchester UK, [3] University of Manchester, Manchester UK, [4] University of Birmingham, Birmingham UK
* These authors contributed equally

Poster # 3209

Abstract

Global loss of genomic DNA methylation is associated with the progression of cancer. This genome-wide hypomethylation is thought to result from increased rates of genome replication in rapidly dividing cancer cells. We leveraged this aspect of cancer biology to address the challenge of early disease detection in cell-free DNA from a cohort of lung cancer patients. We combined Tagomics' Activace platform, which enriches unmethylated CpG sites of the genome for sequencing, with genome-wide, fragmentomic analysis on this hypomethylated DNA fraction. Using a single sample input (>5 ng of cell free DNA) Activace enables a multiomic (epigenomic and fragmentomic) read-out with only 70 M sequencing reads.

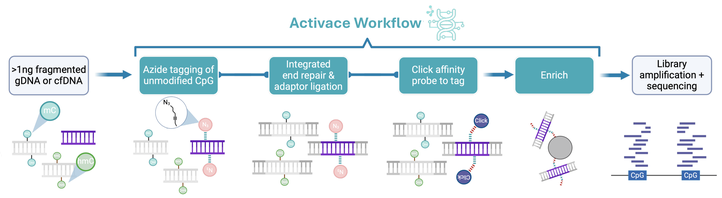
We tested the utility of this approach on a small cohort of 36 patients, who presented at the clinic with symptoms of lung cancer, had a blood sample taken and went on to receive a diagnosis using gold-standard testing (imaging, pathology). The cohort contained 14 cancer-free patients, 12 patients with an early-stage (Stage I & II) diagnosis of non-small cell lung cancer and 10 patients with a late-stage (Stage III & IV) diagnosis of non-small cell lung cancer. We sequenced the cfDNA obtained from the plasma of these patients; and enriched and sequenced the unmethylated DNA fraction from the same patients, using Tagomics' Activace platform. We compared the fragmentomics profiles of both the enriched, unmethylated DNA and the whole cfDNA and evaluated the impact of the enrichment on our ability to classify the cohort.

Using a leave-one-out approach, a logistic regression classifier, trained on the end motif profiles of the Activace dataset (enriched for unmethylated CpG sites), we correctly classified all but two of the early-stage lung cancer cases (91% sensitivity, 100% specificity). By contrast, using an analogous approach in the WGS dataset (whole cfDNA) eight lung cancer cases (seven early-stage and one late-stage) were misclassified (64% sensitivity, 100% specificity).
Multiomics tools, such as the Tagomics' Activace platform enable information-rich insight into disease biology and will likely outperform single analyte-based tests, in the future. The enrichment of DNA containing unmethylated CpG sites improves sensitivity of fragmentomic analysis, relative to analogous analysis of the whole genome, likely due to the enrichment of tumour-derived, hypomethylated DNA fragments in patients with cancer.

Tagomics' Activace platform

Activace enzymatic enrichment of the unmethylated DNA from low input cfDNA samples

Non-damaging process targeting unmethylated DNA



Multiomic read out: epigenomics and fragmentomics in one sample

Low sequencing depth per sample 70M

Does unmethylated DNA enrichment improve early-cancer detection when combined with a fragmentomic analysis?

Proof of Concept Cohort overview:

- 36 cfDNA samples from NCA cohort:
  - 14 non-cancer controls
  - 12 early-stage (Stage I & II) LUAD samples
  - 10 late-stage (Stage III & IV) LUAD samples

WGS
Whole Genome Sequencing

Activace
Unmethylated DNA enrichment and sequencing

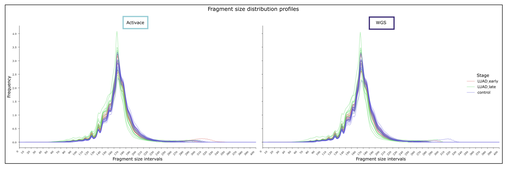
Define **WGS** and **Activace** fragmentomic profiles

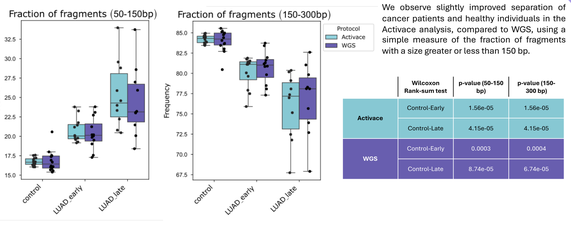
Identification of the best features for **correctly classifying the cohort**

Training **ML models** on fragmentomics' features

1. Fragmentomics features of unmethylated DNA

Activace recapitulates the fragment size profiles of WGS data with fewer reads

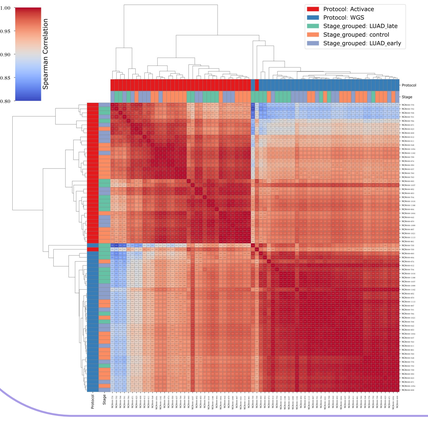




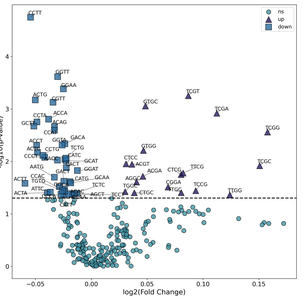
We observe slightly improved separation of cancer patients and healthy individuals in the Activace analysis, compared to WGS, using a simple measure of the fraction of fragments with a size greater or less than 150 bp.

	Wilcoxon Rank-sum test	p-value (50-150 bp)	p-value (150-300 bp)
Activace	Control-Early	1.56e-05	1.56e-05
	Control-Late	4.15e-05	4.15e-05
WGS	Control-Early	0.0003	0.0004
	Control-Late	8.74e-05	6.74e-05

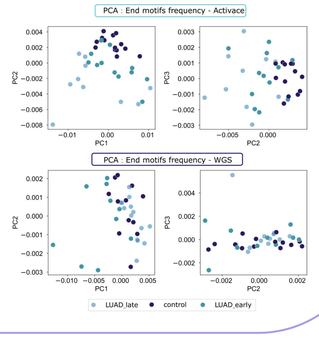
End-motif profiles of Activace and WGS are distinct and cluster separately



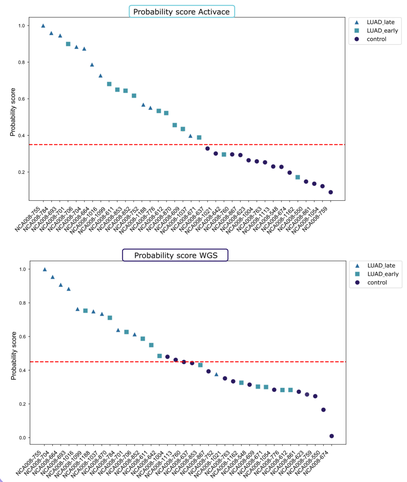
Differential end-motif enrichment between LUAD (all stages) and controls in Activace data



Improved separation between cancer and control samples from Activace data → unmethylome enrichment can help capture subtle cancer-specific changes in cfDNA end motif profiles



2. Unmethylated DNA enrichment improves cancer detection using end motif frequencies

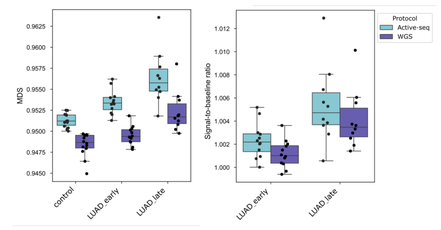


Probability threshold	Accuracy - Activace	Accuracy - WGS
0.3	0.89	0.69
0.35	0.94	0.69
0.4	0.89	0.72
0.45	0.86	0.75
0.5	0.83	0.78
0.55	0.75	0.78
0.6	0.72	0.72

- Leave-One-Out (LOO) cross-validation performed using a logistic regression model based on frequencies of all 256 4-mer end motifs
- **Significantly improved classification and accuracy in Activace dataset**
  - Probability thresholds corresponding to 100% specificity was selected for performance comparison
  - **Activace** (94% accuracy) - misclassified 2 early-stage LUAD cases
  - **WGS** (78% accuracy) misclassified 7 early-stage LUAD and 1 late-stage LUAD case

3. Unmethylated DNA end-motif features improving cancer prediction

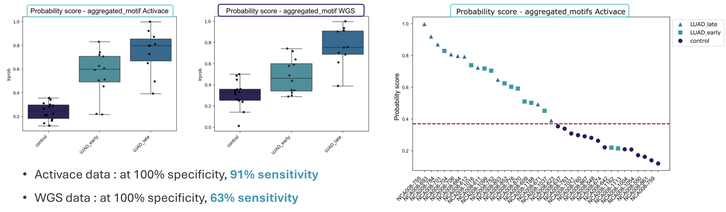
The **Motif Diversity Score** (MDS) is a measure of the diversity of end motifs in the cfDNA samples. This metric is typically increased in cancer samples using WGS data.



Comparison of motif diversity scores between LUAD and control samples and between WGS and Activace datasets:

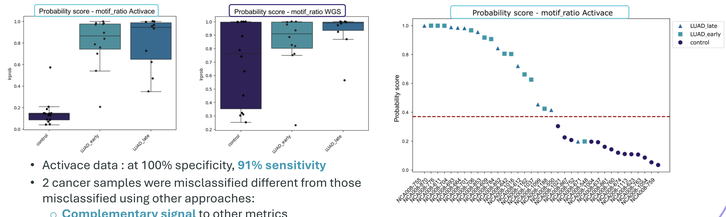
- We observed expected increase in MDS in cancer cfDNA, especially late stage LUAD, in both protocols (Activace P = 0.00014; WGS P = 4.15e-05)
- Only in Activace differences between controls and **early stage** LUAD were statistically significant (Activace P = 0.00014)
- Signal-to-baseline ratios were higher in Activace dataset compared to WGS (a higher ratio indicates a stronger deviation of the signal in cancer cfDNA compared to controls) which highlights improved detection of increased end motif variability at early stage LUAD.

Cumulative end motif frequency in promoter regions showed promising predictive potential using leave-one-out cross-validation with a logistic regression model only in Activace data, not in WGS, suggesting that unmethylome enrichment enhances the detection of potential cancer-associated changes in end motif profiles within these regions



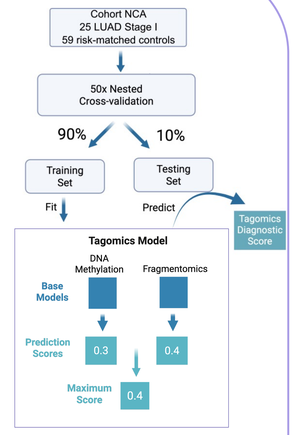
- Activace data : at 100% specificity, **91% sensitivity**
- WGS data : at 100% specificity, **63% sensitivity**

Per region cancer score is defined as the ratio of the frequency of cancer-associated motifs (top 5 motifs upregulated in LUAD and 3 motifs previously reported as upregulated in cancer) to healthy-associated motifs (top 5 motifs downregulated in LUAD and 3 motifs previously reported as downregulated in cancer) **in promoter regions**. We hypothesised that hypomethylated regions shed from tumour cells contain a higher proportion of fragments with end motifs upregulated in cancer. By calculating the ratio between putative cancer- and healthy-associated motifs, we improved the detection of cancer-specific signals in cfDNA.



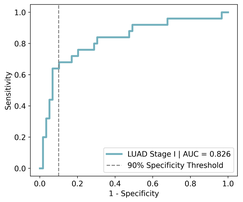
- Activace data : at 100% specificity, **91% sensitivity**
- 2 cancer samples were misclassified different from those misclassified using other approaches:
  - **Complementary signal** to other metrics
  - **Potential useful component for multiomic models**

4. Machine learning models to detect early-stage cancer



- Activace data from 25 patients with Stage I LUAD and 59 risk-matched controls (minimum 70 million reads)
- Machine learning models were designed to process DNA methylation and fragmentomics features. Models were fit to training splits and evaluated on testing splits.
- Models were evaluated by 50-fold nested cross-validation using a 90:10 partition of samples into training and testing, respectively.
- **Tagomics Diagnostic Score:** The final ML classifier was built by taking the maximum prediction score from an ensemble of base models.

- Pooled Receiver Operating Characteristic (ROC) curve showing performance of the best multi-omic model on early-stage LUAD compared to risk-matched controls. A single pooled ROC curve and area under the ROC (AUC) score was derived from the average testing set prediction score per sample across 50 cross-validation runs.



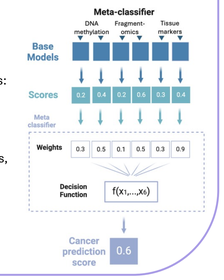
- The best **multiomic** model returned an AUC of 0.826 by taking the maximum prediction from **fragmentomics** and **DNA methylation** models. This model returned a 64% sensitivity at 90% specificity for the detection of stage I LUAD.

Tagomics Oncology Classifier

We are building a **Meta-classifier** based on 6 **Activace-derived features** (Base model probabilities) and including the following models:

- Logistic regression: Weighted sum
- Random forest: Ensemble of trees
- Naïve bayes: Joint distribution of features, assuming independence and gaussian
- Beta Naïve bayes: Joint distribution of features, assuming independence and beta

COMING SOON !



Conclusions

Disease-specific changes in **frequency of short and long cfDNA fragments** were detected in both Activace and WGS data, which confirms that our data processing pipeline is compatible with fragmentomics analysis and that an established fragmentomic biomarker is detected in the Activace dataset.

Significant differences in **end-motif profiles** between the Activace and WGS datasets indicate that the unmethylome enrichment enables access to a distinct subpopulation of cfDNA molecules (clear clustering by protocol).
Notably, end motif profiles derived from the Activace data **more effectively separated cancer from healthy** cfDNA samples, particularly **in early-stage** LUAD samples.

Using a logistic regression classifier trained on Activace-enriched end-motif, we correctly classified all but two of the early-stage lung cancer cases and, at 100% specificity, we **improved the accuracy to 94%**, compared to 78% for WGS data.

Features observed only in Activace data, such as disease-specific changes in **MDS** and the predictive value of end motifs and per-region-cancer-score in **promoter regions**, indicated that **unmethylome-enrichment enhanced** the detection of cancer-associated changes in end motif profiles and **outperformed** WGS data in predicting cancer-status.

Integration of **fragmentomics** and **epigenetics** modules into a **ML model improved the detcetion of cancer**, paving the way for the development of **the Tagomics Oncology Classifier for early-cancer detection**.