Machine Learning Model for Gastric Cancer Prediction using Cell-Free DNA

UCLA Dentistry

Trinity Chan[1], Kevin Trochez[1], Irene Choi[2], Neeti Swarup[2], David T.W. Wong[2]

[1]BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2]School of Dentistry, University of California, Los Angeles, Los Angeles, California, 90095, United States of America,

ABSTRACT

Gastric cancer (GC) is often diagnosed at late stages, leading to a poor prognosis. Cell-free DNA (cfDNA), which are small fragments of circulating DNA, in saliva contains rich, non-invasive biomarkers for the early detection of GC. In particular, we investigate the alterations of cfDNA fragmentation patterns caused by GC using BRcfDNA-Seq. Utilizing mapDamage to quantify nucleotide misincorporation patterns. Preliminary analysis suggests a significant difference between GC and non-GC patients. Using these non-mutational features, we employed a machine learning classifier to aid in the detection of GC. Class and models were optimized using SMOTE, ensemble methods, domain-specific engineering, and grid search with cross-validation using features with the most significant difference, including fragmentomics, chromosomal coverage, motif, and microbial abundance. These features form the basis for a classification model which ultimately achieved an AUC of **0.81** suggesting the model can accurately classify GC and non-GC patients.

BACKGROUND / MOTIVATION

DNA was extracted from ~500uL of saliva supernatant using Qiagen miRNA protocols. Libraries were built using Claret Biosciences single-stranded library kit that permits the incorporation of nicked, jagged, and double-stranded DNA. Saliva cfDNA was sequenced using NovaSeq X Plus for paired-end sequencing at 2 x 150 bp. Analysis of features was performed for both clinical cohorts.

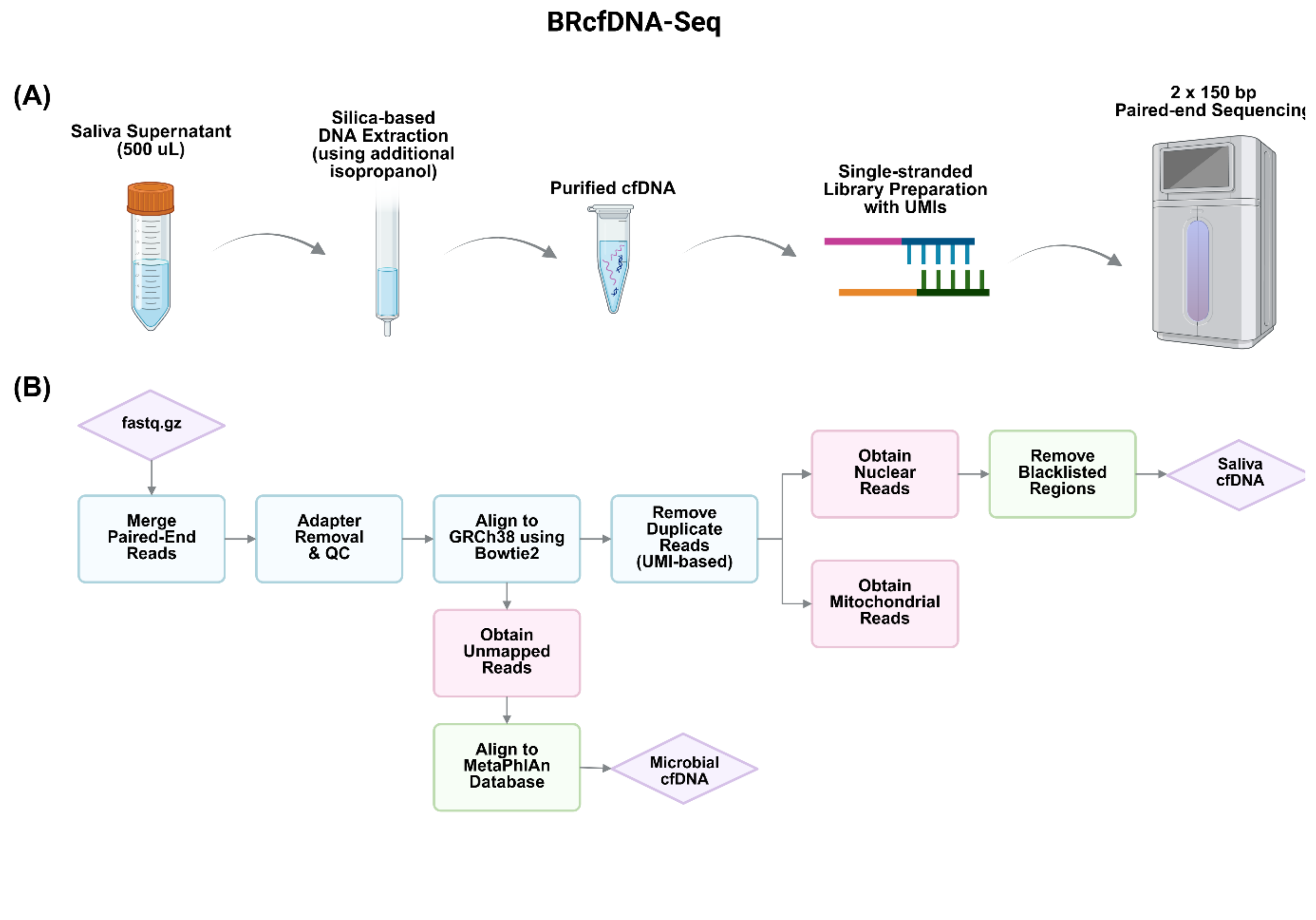
BRcfDNA-Seq
(A)
Saliva Supernatant (500 uL)
Silica-based DNA Extraction (using additional isopropanol)
Purified cfDNA
Single-stranded Library Preparation with UMIs
2 x 150 bp Paired-end Sequencing
(B)
fastq.gz
Merge Paired-End Reads
Adapter Removal & QC
Align to GRCh38 using Bowtie2
Remove Duplicate Reads (UMI-based)
Obtain Nuclear Reads
Remove Blacklisted Regions
Saliva cfDNA
Obtain Mitochondrial Reads
Obtain Unmapped Reads
Align to MetaPhlAn Database
Microbial cfDNA

__Figure 1. Broad Range Cell-Free DNA Sequencing Workflow.__ __(A)__ BRcfDNA-Seq combines extraction methods optimized for low-molecular weight DNA (Qiagen miRNA protocol) with a single-stranded library preparation (Claret BioSciences PicoPlus). __(B)__ Bioinformatic pre-processing pipeline designed for BRcfDNA-Seq, capturing saliva cfDNA captures saliva cfDNA, microbial cfDNA, and mitochondrial cfDNA.

APPROACH / METHODOLOGY

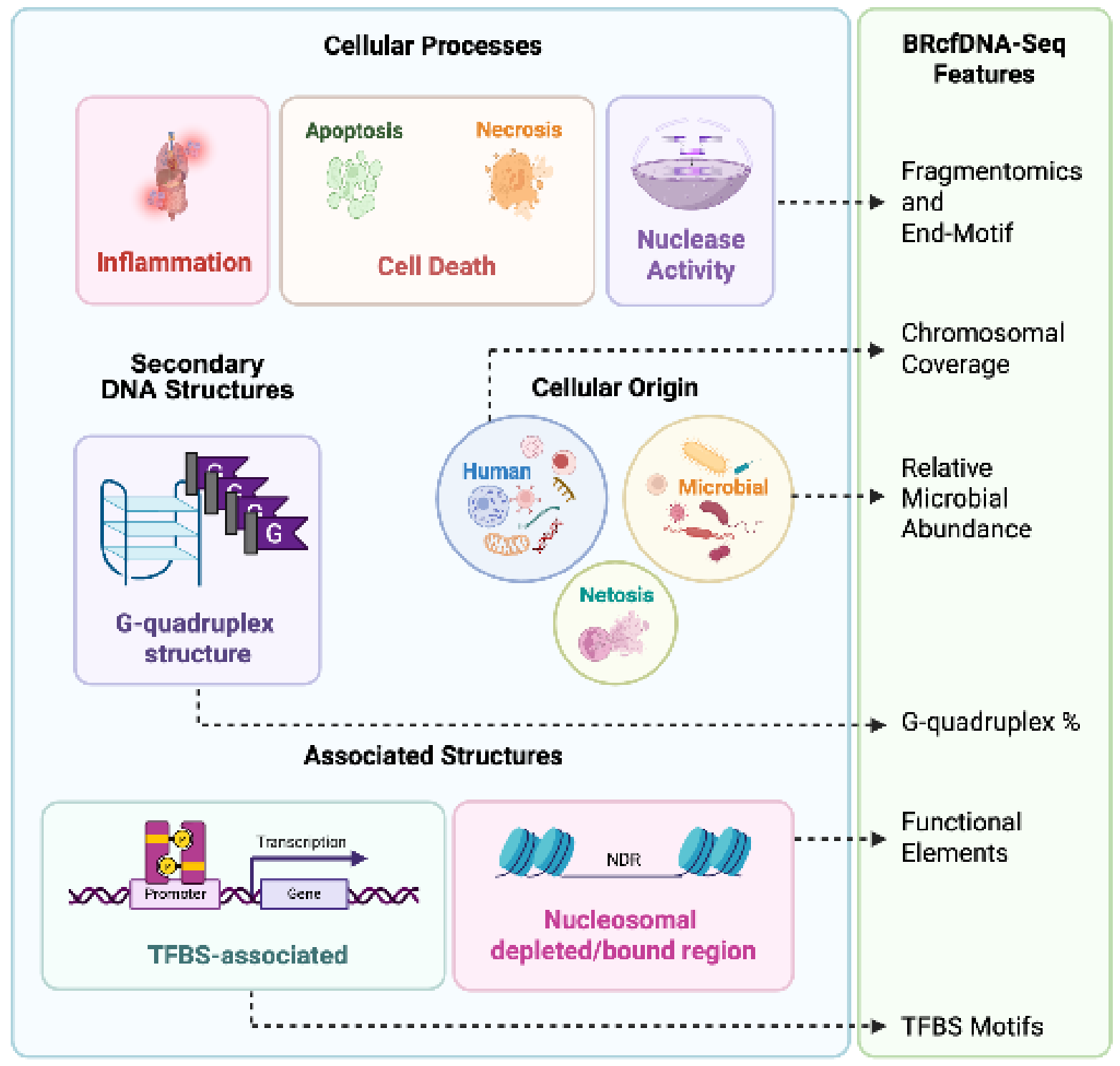

__Figure 2. Factors affecting cfDNA characteristics.__ Processes and structural features that lead to cfDNA formation; biological and cellular processes, genomic structures, nucleosomal associated structures or epigenomic modifications.

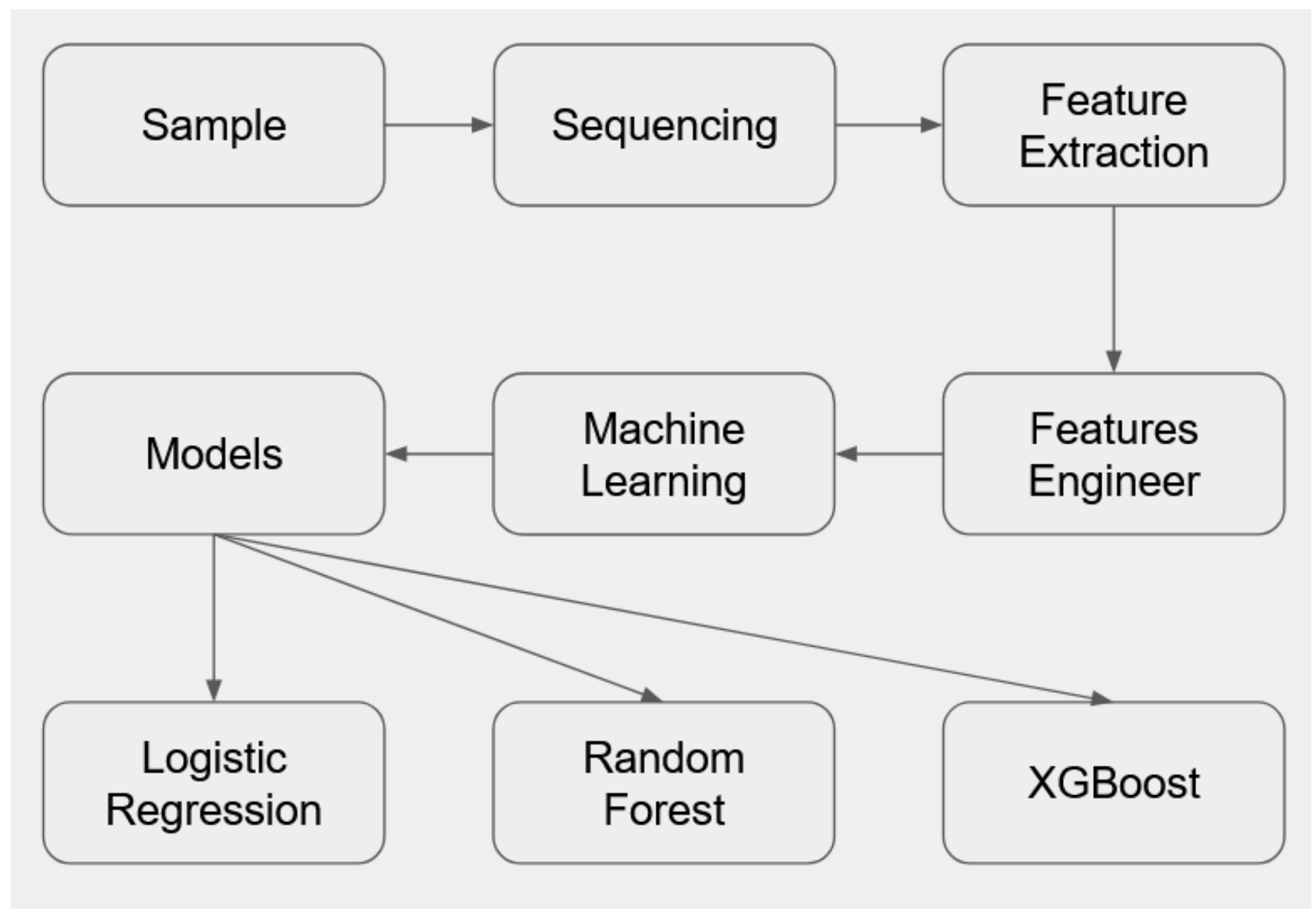

__Figure 3. Workflow Diagram.__ modifications Saliva cfDNA is sequenced using BRcfDNA-Seq and analyzed for fragmentomics, motif, chromosomal coverage, microbial features used for machine learning models.

RESULTS

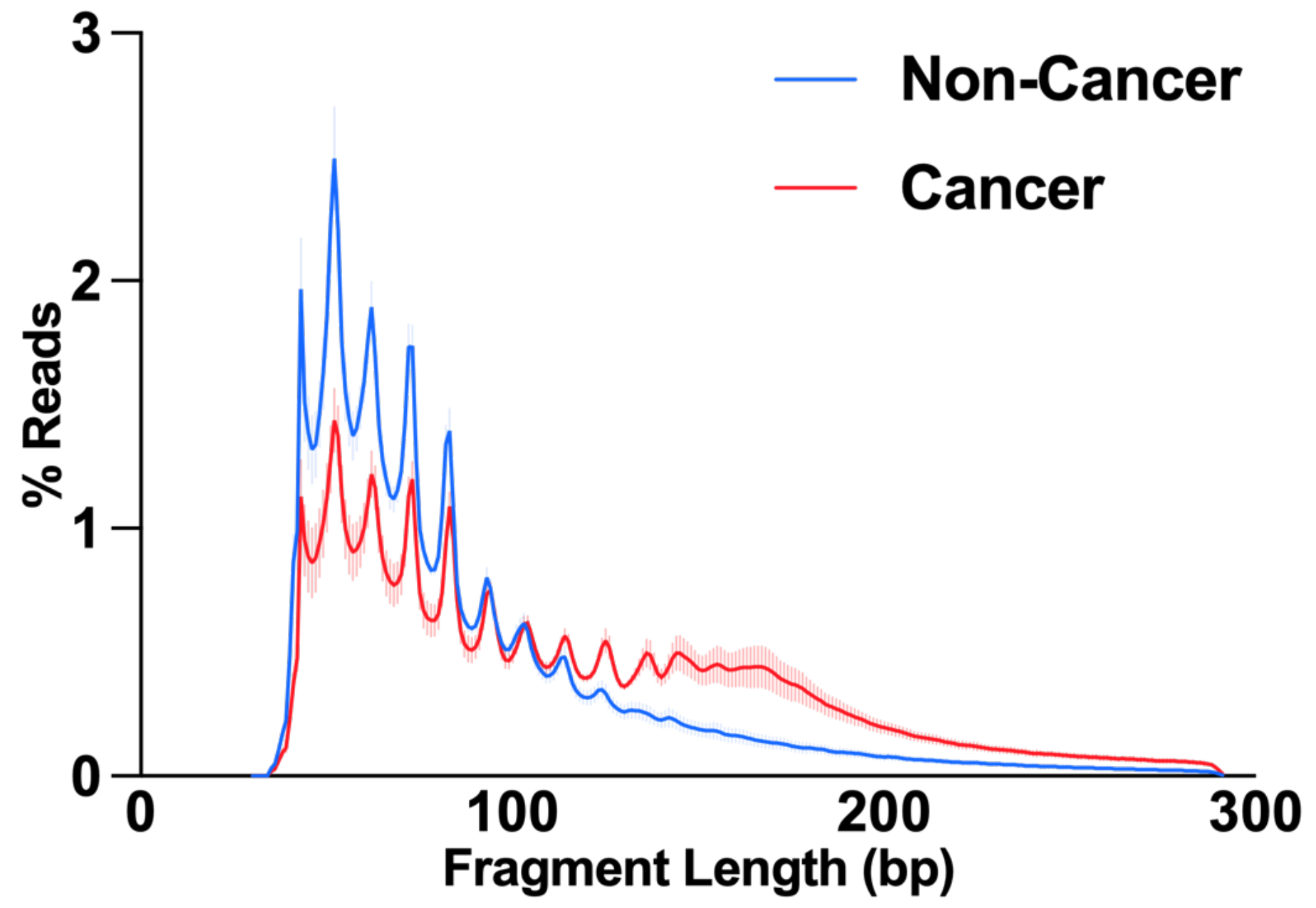

__Figure 4. BRcfDNA-seq reveals a unique jagged profile of ScfDNA.__ ScfDNA demonstrates equidistant (~10bps) multiple peaks, below 100 bps. Fragment profile is different between gastric cancer and non-gastric cancer patients.

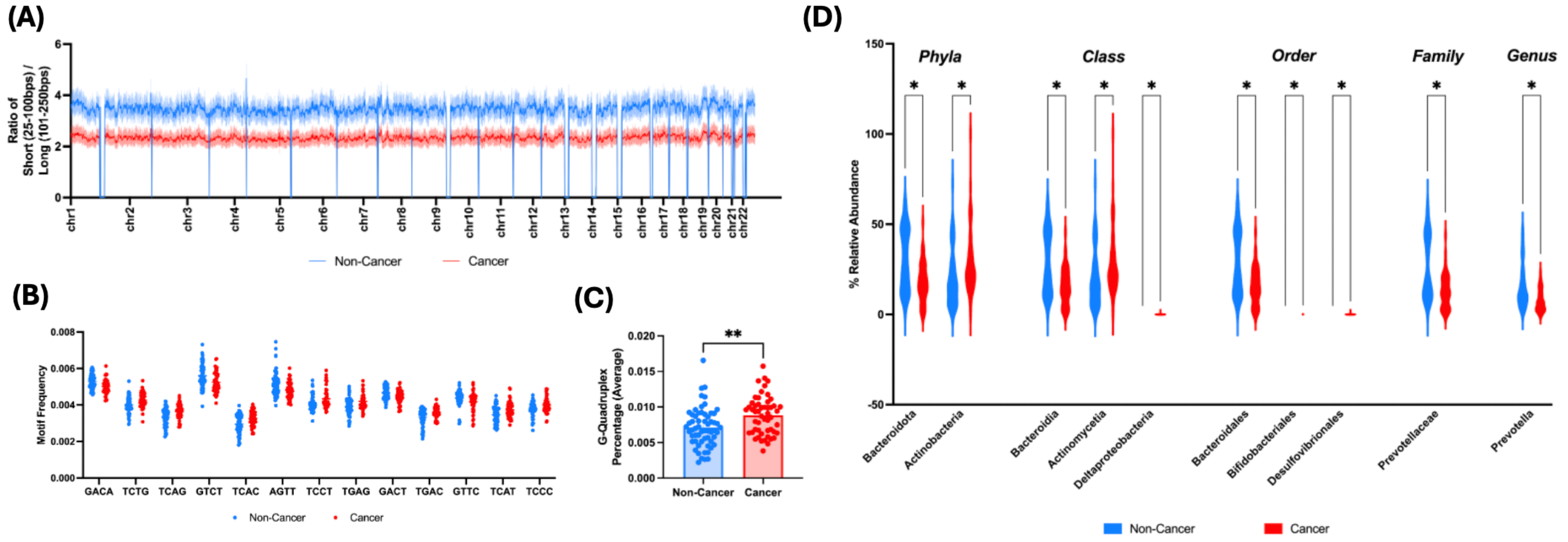

__Figure 6. Salivary cfDNA BRcfDNA-Seq features differentiate gastric cancer patients from non-cancer controls__. __(A)__ Non-cancer individuals exhibit a significantly higher fragmentomic ratio of short (25–100 bp) to long (101–250 bp) cfDNA fragments. __(B)__ Top discriminatory 4-mer end-motif frequencies show distinct patterns between gastric cancer and non-cancer samples. __(C)__ Gastric cancer patients display a significantly increased abundance of G-quadruplex structures. __(D)__ Relative abundances of salivary microbiome taxa at the phylum, class, order, family, and genus levels show significant differences between gastric cancer and non-cancer groups.

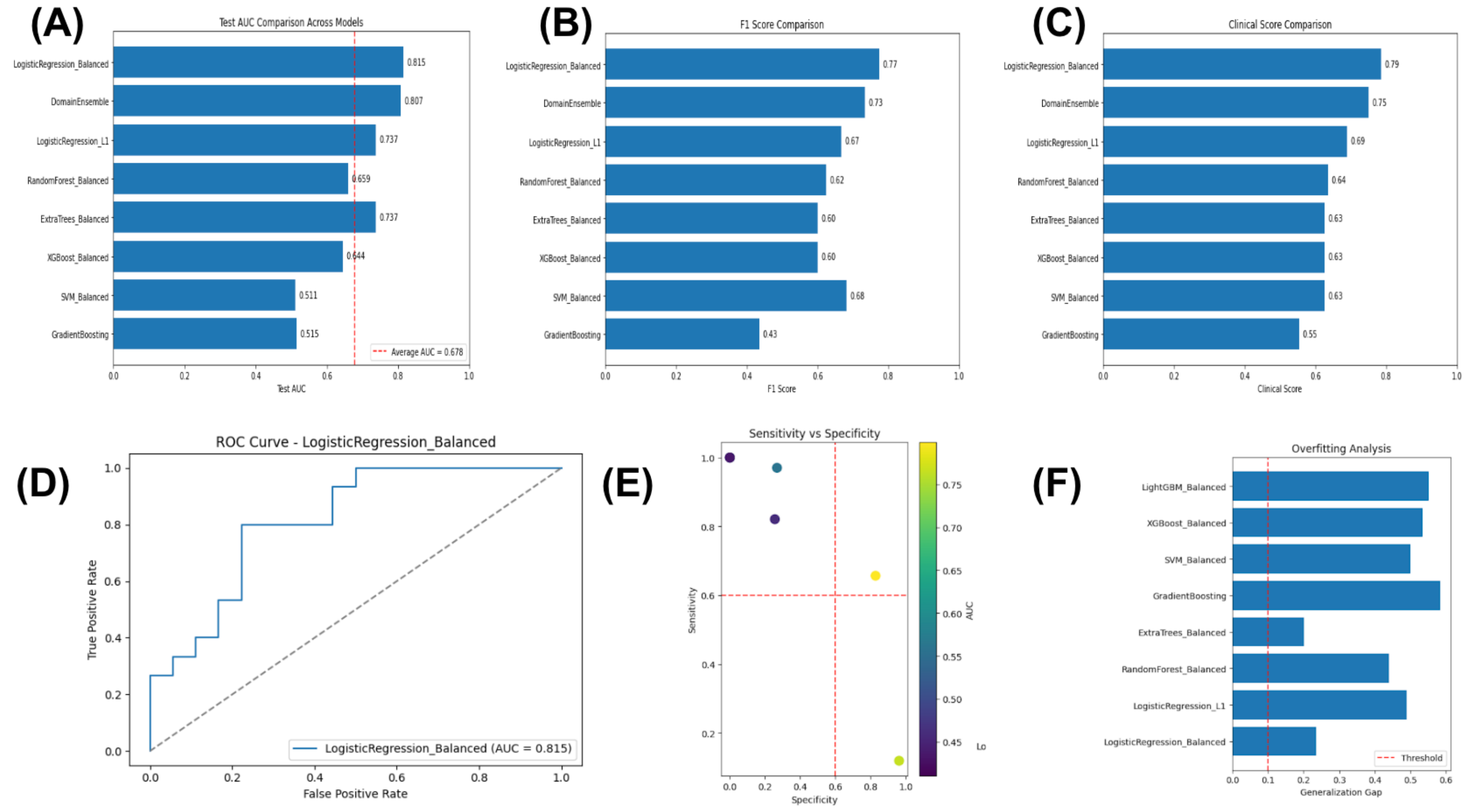

Figure 7. Performance comparison of multiple machine learning models for cancer classification. Bar plots compare model performance across three metrics: Test AUROC **(A)**, F1 Score **(B)**, and a custom Clinical Score **(C)**, which reflects clinical relevance or interpretability. The Logistic Regression **(D)** model achieved the highest performance in all three categories, with an AUROC of 0.815, F1 score of 0.77, and Clinical Score of 0.79. The Domain Ensemble model also performed competitively. Models such as Gradient Boosting and SVM performed poorly across all metrics

CONCLUSION

- We introduce a novel machine learning model for gastric cancer prediction using cfDNA data obtained through saliva.
- Pilot and training data suggests that ScfDNA differs in gastric cancer and non-cancer saliva, in terms of origin (human and microbial), fragmentomic profile, human genomic elements, and genes contributing to **ScfDNA** and 4-mer end motifs of **ScfDNA**.
- Features of plasma cfDNA are a result of variety of nuclease enzymes, similar to how the activity of nuclease enzymes in the oral cavity, salivary gland, and oral microbiota may be contributory to discriminatory non-somatic features of **ScfDNA**.
- Feature engineering was performed to normalize sequencing depth, reduce batch effects, and highlight biologically relevant variation. This included log transformations, z-score normalization, and domain-informed binning for chromosomal coverage and motif features
- Various classical machine learning models were tested with features derived from **BRcfDNA-Seq**.
- Logistic Regression Balanced model performed the best on the validation dataset giving an ROC of 0.81 indicating the model's performance to accurately predict gastric cancer

FUTURE DIRECTIONS

To build on our initial promising results, we could conduct more feature engineering or test more machine learning models to improve the validation accuracy. Adding more samples to both the training and validation dataset will increase its potential even further.

REFERENCES

- PMC4371901
- PMC6935139
- PMC6774252
- PMC2928508
- PMC4489295
- PMID: 32004449
- PMID: 40237938

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health (NIH) (R25 5R25DE030117-05). In addition, we acknowledge the donation made by Ronnie James Dio Stand Up and Shout Cancer Fund.