Predicting High Confidence ctDNA Somatic Variants with an Ensemble Machine Learning Model

R. Maruzani, L. Brierley, A. Jorgensen & A. Fowler
Department of Health Data Science, University of Liverpool

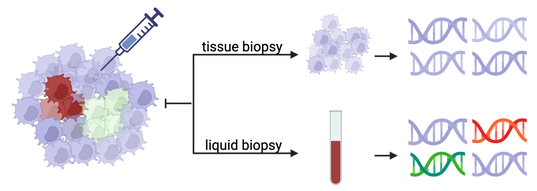
Medical Research Council

nonacus
ADVANCING NON-INVASIVE HEALTHCARE

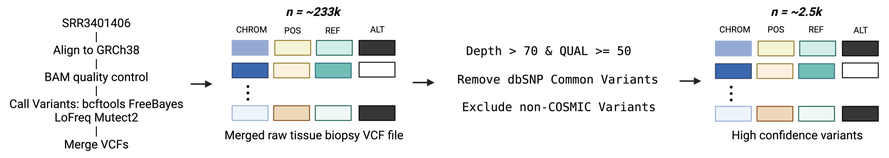
Introduction

Tumour variants can predict response to treatment and inform personalized treatment for cancer patients. Tissue biopsies provide the gold standard for the identification of tumour variants, however these are typically difficult to obtain and require an invasive procedure. Circulating tumour DNA (ctDNA) is a promising, minimally invasive cancer biomarker that can be used to inform treatment of cancer patients.
ctDNA is released from tumour cells into the bloodstream. ctDNA is easily accessible and contains tumour variants. Detecting real ctDNA variants with Next Generation Sequencing (NGS) technology can be a challenge due to the low abundance of ctDNA in the pool of cell free DNA in the bloodstream. Rule-based filtering strategies either remove a substantial number of true positive ctDNA variants along with false variant calls, or retains an implausibly large number of total variants.

Machine Learning (ML) enables identification of complex, non-linear patterns which may improve ability to distinguish between real low-frequency ctDNA variants, and false positive calls arising from sequencing errors. The aim of this study is to develop a machine learning model for detecting high-confidence ctDNA somatic variants in the absence of a matched tissue sample. We used a ctDNA dataset (SRP073475) sequenced with matched tissue samples (1) to train and test models. Samples in this dataset were whole exome sequenced at an average depth of 70x. Patients included in the study were diagnosed with either Stage III squamous cell carcinoma or Stage III lung adenocarcinoma

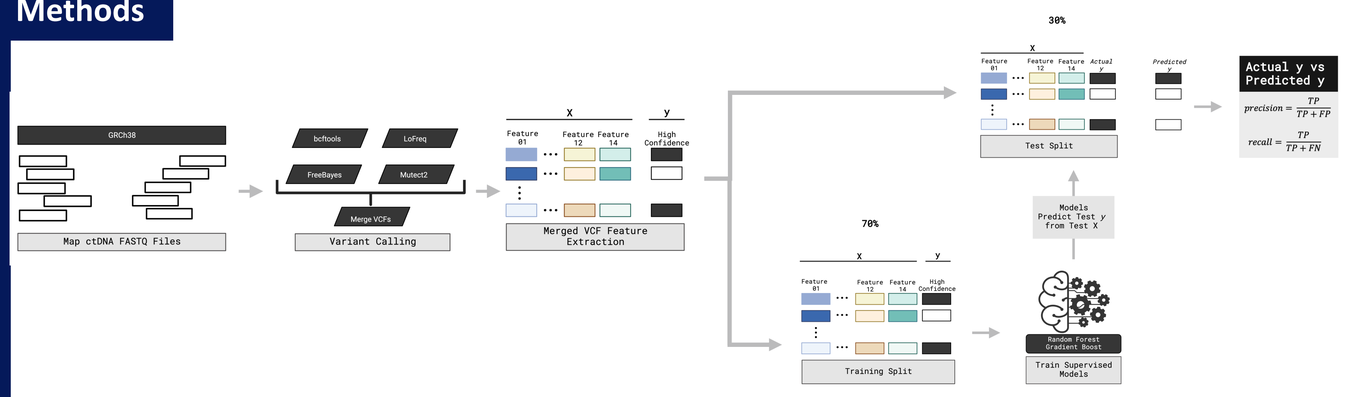


Advantages of Liquid Biopsy versus Tissue Biopsy for NGS analysis.



Workflow to identify high-confidence tumour variants. Variants called by 4 callers were filtered using stringent filters to obtain the high-confidence tumour variants. Variants in this set observed in the matched ctDNA sample were class 1, all other variants were class 0

Methods



Workflow for developing and evaluating ML models built to predict high-confidence ctDNA variants. ctDNA FASTQ files were mapped to a GRCh38 using bwa-mem at default parameters. Duplicate reads were marked with GATK MarkDuplicates and base quality scores were recalibrated with GATK ApplyBQSR. Variants were called with bcftools, LoFreq, FreeBayes, and Mutect2, before merging raw VCF files. For each variant, a set of 14 features were extracted and data was split into training and test sets. The training set was composed of 70% of the complete dataset. Training data was input for Random Forest and Gradient Boosting models (GBMs). Models were evaluated on test data, composed of 30% of the full dataset.

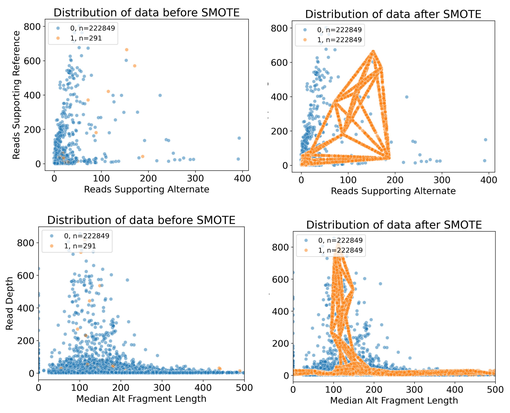
Features used in machine learning models

Features	Feature Description
dbSNP	Is variant described in dbSNP v155 database
COSMIC	Is variant described in the Catalogue of Somatic Mutations in Cancer v98 database
Strand Bias	Strand bias estimated using Fisher's exact test, Phred-scaled p-value
Median Fragment Length	Median fragment length of alternate reads subtracted from reference reads
Weighted Homopolymer Rate	The sum of squares of the homopolymer lengths, divided by the number of homopolymers in a 20 nucleotide region surrounding the variant. Homopolymers defined as 4-mers and above. Weighted Homopolymer Rate equation proposed by the Broad Institute (2)
GC Percentage	GC percentage in a 20 nucleotide region surrounding the variant
Reads Supporting Reference	Number of reads supporting the reference allele
Reads Supporting Alternate	Number of reads supporting the ALT allele
Mapping Quality	Median mapping quality of reads supporting ALT allele
Allele Frequency	Variant allele frequency calculated as reads supporting alternate/(reads supporting reference + reads supporting alternate
bcftools	Is variant detected by bcftools
FreeBayes	Is variant detected by FreeBayes
LoFreq	Is variant detected by LoFreq
Mutect2	Is variant Detected by Mutect2

Features and description used to predict high-confidence ctDNA variants

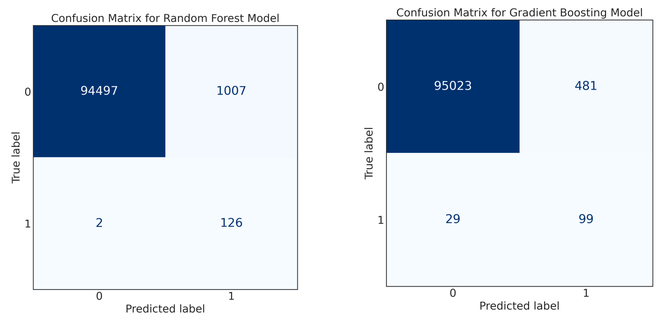
$$WHR = \frac{\sum_{i}^{n=1} l_i^2}{n}$$

Results



SMOTE was used to balance the training data. SMOTE balances classes by generating synthetic minority class data

Random Forest Model Outperformed Gradient Boosting Model in Detecting High-Confidence True Positive Variants



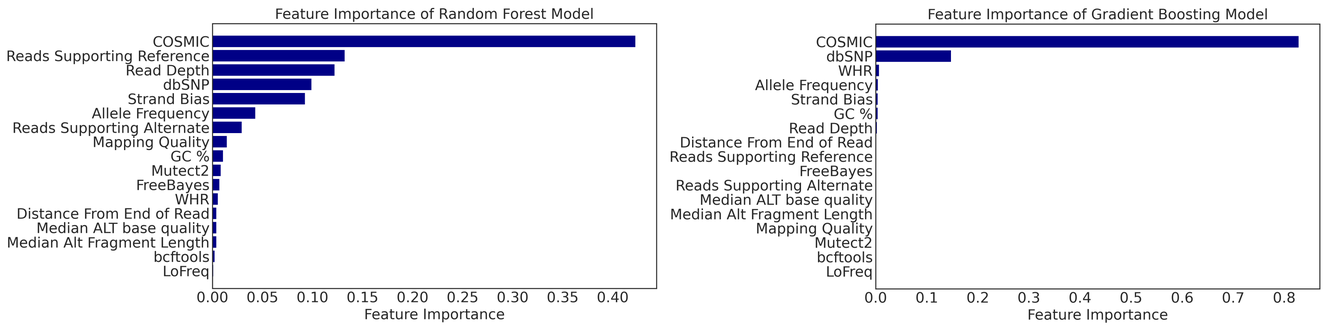
Confusion matrices of Random Forest and Gradient Boosting models on test data. Random Forest detected more true positive variants while the Gradient Boosting model predicted fewer false positive variants

Machine Learning Models Outperformed Soft and Hard Filtering in Predicting True Positive Variants

Method	Recall	Precision
Hard Filter	0.265	0.434
Soft Filter	0.507	0.228
Gradient Boosting	0.773	0.170
Random Forest	0.980	0.110

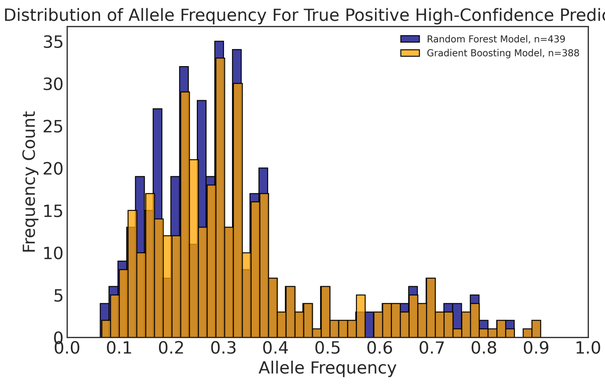
Comparison of Precision and Recall Between Random Forest, Gradient Boosting, Soft and Hard Filters. Hard filtering included removing variants with Phred Quality scores of less than 50, read depth of less than 20 and retaining only variants present in the COSMIC database. Soft filters lowered the Quality threshold to 40, and depth to 10.

COSMIC Membership was the Most Important Feature for Both Random Forest and Gradient Boosting Models



Feature Importance of Machine Learning Models: Random Forest feature importance is calculated as mean accumulation of the impurity decrease within each tree. Gradient Boosting feature importance is calculated as total reduction of the criteria brought by that feature.

Random Forest Model Predicted More Low Frequency Variants Than Gradient Boosting Model



Distribution of allele frequencies for true positive predictions in complete dataset. The Random Forest model predicted more high-confidence variants at lower frequencies compared to the Gradient Boosting model.

Discussion

We developed Machine Learning models to classify high-confidence ctDNA somatic variants in the absence of a matched tissue sample.

Circulating Tumour DNA fragments tend to be shorter than healthy cfDNA fragments. We therefore included median fragment length of reads supporting the alternate allele as a feature. Feature importance analysis however, showed median fragment length did not contribute significantly to Random Forest nor Gradient Boosting model performance. This may be due to the overlap in fragment lengths of healthy cfDNA and ctDNA fragments.

The variant presence in COSMIC feature scored highest in both Random Forest and Gradient Boosting models for feature importance. This is unsurprising as COSMIC is a manually curated database of somatic variants associated with cancer. A variant reported in COSMIC is therefore likely to be a high-confidence variant. The Random Forest model feature importance and allele frequency results suggest other features are important in predicting low frequency high-confidence variants.

Models were trained on single nucleotide variants only. While these variant types make up the majority of cancer associated mutations, small insertions and deletions can be linked to disease. Future iterations of models will include INDEL predictions.

Future Work

Cell free DNA end motif frequencies have recently been shown to differ in cancer patients compared to healthy subjects. The 4-mer CCCA in particular is less abundant in cancer patient cell free DNA (3). We will explore if this feature can contribute to the performance of our models.

We employed SMOTE to address the imbalance in classes in our data and GridSearchCV for parameter optimization. Future iterations of models will investigate the impact of other methods on performance.

The input data to the models presented here was from whole exome sequencing. Next, we will build models with data from targeted sequencing with matched normal samples. The targeted sequencing data will be ultrahigh depth and the matched normal will allow for more accurate removal of germline variants.

References
1) Dietz S., *et al.*, 2016 Low Input WholeExome Sequencing to Determine the Representation of the Tumor Exome in Circulating DNA of Non-Small Cell Lung Cancer Patients. PLoS ONE 11, e0161012. doi:10.1371/journal.pone.0161012
2) Seo, H., *et al.*, 2017. Evaluation of exome variants using the Ion Proton Platform to sequence error-prone regions. PLoS One 12, e0181304. https://doi.org/10.1371/journal.pone.0181304
3) Jiang, P., *et al.*, 2020. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discovery 10, 664–673. https://doi.org/10.1158/2159-8290.CD-19-0622