Testing the generalizability of cfDNA fragmentomic features across different studies for cancer early detection

FPN: 1681P

Yulong Xuan[1], Shu Su[2], Xiaojun Fan[3], Haimeng Tang[3], Hua Bao[3], Xin Lv[2], Wei Ren[2], Fangjun Chen[2], Xue Wu[3], Yang Shao[3], Tao Wang[1], Lifeng Wang[2]

[1] The Comprehensive Cancer Centre of Nanjing Drum Tower Hospital, Medical School of Nanjing University & Clinical Cancer Institute of Nanjing University, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School;

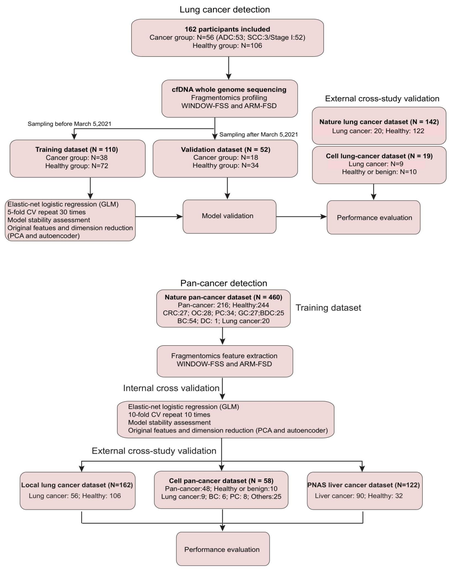
[2] Department of Thoracic and Cardiovascular Surgery, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School; [3] Geneseeq Research Institute, Nanjing Geneseeq Technology Inc., Nanjing 210044, Jiangsu, P. R. China

BACKGROUND

Lung cancer has the highest incidence and mortality worldwide. Early detection of lung cancer is critical for improving the prognosis of patients. Cell free DNA fragmentomics has shown potential in the detection of lung cancer. Recently, the most common approach involves the profiling of short (100-150bp) and long (151-220bp) fragment distributions to differentiate between healthy individuals and cancer patients. However, existing predictive models lack extensive cross-study validation, and the robustness and generalizability of the features and models should be tested and improved.

METHODS

The local lung cancer cohort consisted of 56 patients with lung cancer (93% stage I) and 106 healthy volunteers. The 162 subjects were divided into a training cohort (n=110) and a validation cohort (n=52). The samples were subjected to whole-genome sequencing. Two types of cfDNA fragment features were extracted: window-level fragment size summary (WINDOW-FSS), which summarizes the fragment sizes as short fragment (100-150bp) and total fragment (100-220bp) coverage at 5MB window level, and commonly used in previous studies, and arm-level fragment size distribution (ARM-FSD), which separately calculates the coverage of fragments with 5bp as a step at arm level while retaining the size distribution information. In addition, the derived PCA components and autoencoder deep features were also extracted and used to construct machine learning models for lung cancer prediction. The performance of the models was validated by the validation cohort and two independent external on-line cohorts (n=142 and 19 respectively). The value of the two features in pan cancer detection was assessed by an on-line pan-cancer cohort (n=460) as training and validated by another on-line pan-cancer cohort (n=58), an on-line liver cancer cohort (n=122) as well as the local lung cancer cohort (n=162).



RESULTS

Training and validation of ARM-FSD in lung cancer cohorts

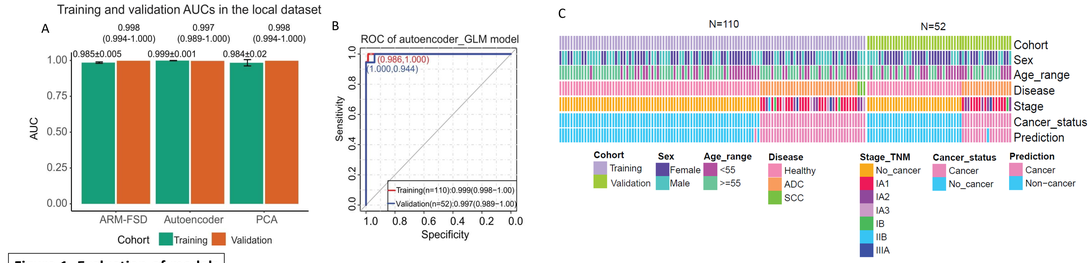


Figure 1. Evaluation of models in the detection of lung cancer. (A) AUCs of 5-fold cross repeating 30 times byvalidation the training cohort and validated by the validation (B) ROC curve of the training cohort based on average predicted probability of cancer by cross validation (red), and ROC curve of the validation cohort based on predicted probability of cancer (blue). (C) Prediction of lung cancer in patients in the training cohort and validation cohort by the autoencoder model.

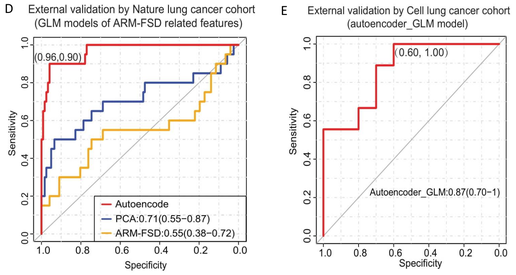
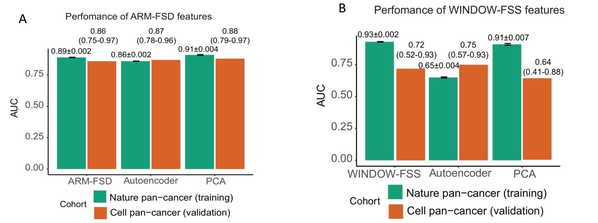


Figure 1 Cont'd. (D) External validation of the GLM models based on original ARM-FSD feature, autoencoder of ARM-FSD, or PCA of ARM-FSD by the Nature lung cancer cohort. (E) External validation of autoencoder of ARM-FSD model by Cell lung cancer cohort.

Performance of ARM-FSD compared to WINDOW-FSS



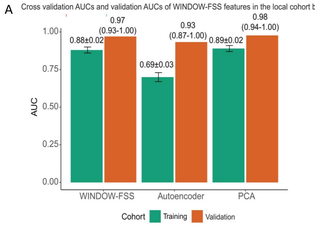
Validation AUCs of 9 non-linear models

Feature	GBM	RF	DL
WINDOW-FSS	0.44(0.23-0.65)	0.60(0.35-0.86)	0.60(0.44-0.76)
Autoencoder	0.48(0.29-0.68)	0.62(0.47-0.78)	0.68(0.45-0.90)
PCA	0.46(0.29-0.62)	0.68(0.49-0.87)	0.72(0.55-0.88)

Figure 3. Pan-cancer detection by ARM-FSD and WINDOW-FSS features. (A) AUCs of 10-fold cross validation repeating 10 times by the Nature pan-cancer training cohort (n=460) and externally validated by the Cell pan-cancer cohort (n=58).

Figure 3 Cont'd. (B) AUCs of 10-fold cross validation repeating 10 times by the Nature pan-cancer training cohort (n=460) and validated by the Cell pan-cancer external validation cohort (n=58). AUCs of 9 non-linear models (3 WINDOW-FSS features × 3 non-linear algorithms).

Training and validation of WINDOW-FSS in lung cancer cohorts



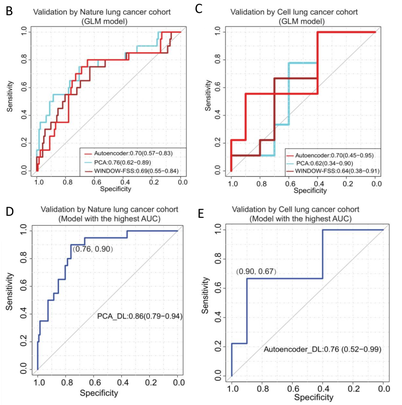
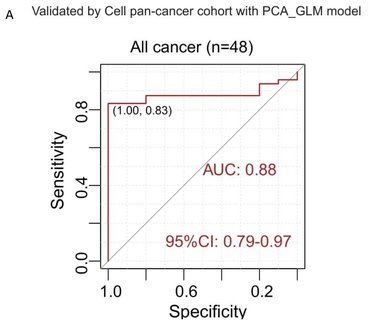


Figure 2. Evaluation of models incorporating WINDOW-FSS, autoencoder of WINDOW-FSS, or PCA of WINDOW-FSS in the detection of lung cancer. (A) AUCs of 5-fold cross validation repeating 30 times by the training cohort (n=110) and validated by the validation cohort (n=52) for the local lung cancer cohort. (B) External validation of the GLM models based on original WINDOW-FSS feature (WINDOW-FSS), autoencoder of WINDOW-FSS, or PCA of WINDOW-FSS by the Nature lung cancer cohort. (C) External validation of the GLM models based on original WINDOW-FSS feature, autoencoder of WINDOW-FSS, or PCA of WINDOW-FSS by the Cell lung cancer cohort.

Figure 2 Cont'd. (D) The model with the highest AUC validated by Nature lung cancer external cohort. PCA_DL means deep learning model with PCA of WINDOW-FSS. Numbers in parentheses indicate the specificity and sensitivity, respectively, at Youden index for each ROC. (E) The model with the highest AUC validated by Cell lung cancer external cohort. Autoencoder_DL means deep learning model with autoencoder of WINDOW-FSS.

Independent validation using an external pan cancer cohort



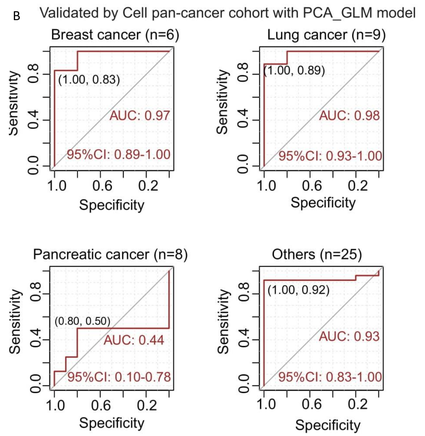


Figure 4. (A) Overall validation ROC curve derived from the model including PCA of ARM-FSD. Numbers in parentheses indicate the specificity and sensitivity, respectively, at Youden index for each ROC. (B) ROC curves stratified by cancer types derived from the ARM-FSD (PCA-transformed) model.

CONCLUSIONS

- A possible future direction could be to evaluate stability of other features, such as end motifs, in external cross-study validations.
- Our newly-developed ARM-FSD feature set is a robust and generalizable biomarker and has potential in the early detection of lung cancer and pan cancers.

CONTACT:

Dr. Lifeng Wang, Nanjing Drum Tower Hospital
Email: lifengwang@nju.edu.cn

Conflicts of interest:

The presenting author declares no conflict of interest.

Abbreviations: ARM-FSD: arm-level fragment size distribution; WINDOW-FSS: window-level fragment size summary; ACC: adenocarcinoma; SCC: squamous cell carcinoma; PCA: principal component analysis; CRC: colorectal cancer; OC: ovarian cancer; PC: pancreatic cancer; GC: gastric cancer; BDC: bile duct cancer; BC: breast cancer; DC: duodenum cancer; ROC: receiver operating characteristic; AUC: area under the ROC curve