



Construction of a Lung Cancer Diagnosis Model by Multi-omics Technology

MGI DNBSEQ sequencing platform enables the development of a lung cancer diagnosis multi-omics model

A research team mainly from the Peking University People's Hospital conducted a high-throughput sequencing research using the MGI DNBSEQ-G400 sequencer and explored liquid biopsy markers of lung cancer and benign lung nodule with multi-omics technology. The study revealed the application value of multi-omics liquid biopsy markers in the auxiliary diagnostic and prognostic evaluation of lung cancer.

The relevant results of this study were published in 2021 in the journal *Molecular Cancer*, under the title "Non-invasive lung cancer diagnosis and prognosis based on multi-analyte liquid biopsy"¹.

Recommended application: Tumor omics (lung cancer)

Recommended model: DNBSEQ-G400RS

- High quality of sequencing data

With its high accuracy, low repeat rate, and low index hopping rate, DNBSEQ sequencing technology provides accurate sequencing data for cancer research.

- Better diagnosis results from multi-omics model

Compared to models with a single parameter or without CEA, a diagnostic model based on both the genetics and epigenetics of circulating cell-free DNA (cfDNA) and the serum protein marker CEA has the best ability to classify malignant and benign cases.



Background

Lung cancer is a malignant tumor with the highest incidence and fatality rate in China². Risk factors include smoking, genetics, poor diet, air pollution and occupational or mixed occupational exposure to asbestos, metal, silicon dioxide, polycyclic aromatic hydrocarbon (PAH), and diesel exhaust. Early diagnosis³, early treatment⁴, and early monitoring⁵ are key to improving patient survival and reducing the economic burden to society and family. At present, low-dose computed tomography (CT) is the most widely used method for lung cancer scanning. However, it has some limitations. Not only does it incur radiation exposure, but it may also not allow clinically distinguishing between benign and malignant nodules leading to false positive results. In addition, the prognosis indices for lung cancer rely on the TNM staging of lung cancer in clinical practice, but patients at the same stage often have different prognoses, making it necessary to find additional evaluative parameters.

Molecular changes such as the tumor driver gene mutation status and expression features are closely related to the prognosis of lung cancer (LC). Recent evidence supports the prognostic value of epigenetic alternations, but these remain to be expounded. Therefore, new liquid biopsy markers suitable for the diagnostic and prognostic evaluation of LC have important clinical value and have become a research hotspot in recent years. Compared with traditional cancer diagnosis using tissue biopsy, liquid biopsy is more feasible, less invasive, and can deliver a more comprehensive characterization of tumor heterogeneity⁶ because circulating tumor DNA (ctDNA) is released into the bloodstream by all tumor sites.

Study description

ctDNA in the serum of cancer patients provides valuable information about cancer genomes, and shows great promise in non-invasive cancer detection^{7,8}. However, since ctDNA is diluted by a large amount of non-oncogenic cfDNA, its detection can be challenging, especially in the early stages of cancer where the tumor mass is small.

In this study, researchers developed a set of experiments and computing tools to measure the genetic and epigenetic signals of plasma cfDNA of patients with LC and patients with benign lung nodules (BLN) through high-throughput sequencing. They constructed a model based on sequencing data and the serum protein marker CEA, aiming to explore the potential application value of blood biomarkers in LC diagnosis and prognosis⁹.

Materials and Methods

Sample collection

In this study, DNA was extracted from preoperative blood samples, intraoperative LC tissues, and paracarcinoma normal tissue samples from 128 patients with LC and 94 patients with BLN. Concurrently, leukocyte genomic DNA was extracted to account for contamination by white blood cells (WBC gDNA).

Library preparation and sequencing

a. Design of probe panel

A whole-cancer panel containing 139 genes was designed based on the TCGA and COSMIC databases for subsequent targeted ultra-deep sequencing.

b. Targeted sequencing of WGS library

To reduce the impact of PCR or sequencing on the prepared DNA samples, an ultra-deep targeted sequencing library (NGS) was prepared with a paired-end unique molecular identifier (UMI) strategy.

The libraries were also prepared with this strategy for cfDNA, LC tissues, paracarcinoma normal tissues, and WBC gDNA. The targeted capture experiment was conducted for the WGS library preparation using IDT's xGen[®] Lockdown[®] reagent. Specific instructions are available in the relevant user manual. Paired-end sequencing of 100 basic groups (PE100) for the obtained library was completed on the DNBSEQ-G400 sequencer.

c. Targeted sequencing of WGBS library



Meanwhile, to detect the changes of LC-specific epigenetics, a whole-genome hydrosulphite sequencing library (WGBS) was prepared in a single-stranded DNA library preparation strategy for cfDNA, LC tissues, paracarcinoma normal tissues, and WBC gDNA. A targeted capture experiment was conducted for the WGBS library preparation using Roche's SeqCap Epi CpGiant probe. Specific instructions are available in the relevant user manual. Paired-end sequencing of 100 basic groups (PE100) for the obtained library was completed on the DNBSEQ-G400 sequencer.

Data analysis

Raw data from the DNBSEQ-G400 sequencer were processed as follows. After UMI cut and filtration of low-quality reads using FASTQ, the remaining reads were aligned to the human genome Hg19 to identify hotspot mutations.

Exon mutations of 139 cancer driver genes selected based on TCGA and COSMIC databases were analyzed, and differences in mutations of benign versus malignant lung nodules were determined according to mutation scores. In addition, WGBS targeted sequencing was performed to detect and identify the 315 LC-specific methylation positions (GpC) of LC tissues and paracarcinoma tissues. The subsequent analysis on the whole-genome methylation sequencing results covered 5.6 million GpC

positions, and the regional methylation ratio of each cfDNA was calculated to determine the effectiveness of methylation markers. Finally, a classification model was constructed via machine learning methods, where serum protein markers were combined with methylation and other aspects to screen the methylation positions and serum protein markers that could be used to diagnose benign and malignant lung nodules.

Sample collection	Library preparation and sequencing	Bioinformatics Analysis	Result analysis
Preoperative blood samples, intraoperative LC tissues, paracarcinoma normal tissue samples, and leukocyte genomic DNA from 128 patients with LC and 94 patients with BLN	<div>  Preparation of WGS and WGBS targeted sequencing libraries </div> <div>  DNBSEQ-G400 genetic sequencer </div>	FASTQ SOAPnuke-2.0.3 BWA-MEM BitMapperBS MethylDackel	Differences in mutations of benign and malignant tumors, LC-specific methylation positions, Effectiveness of methylation markers, Construction of a classification model via machine learning methods

Results

Detecting different mutation spectra of plasma cfDNA and WBC gDNA by targeted ultra-deep sequencing

The 128 patients with LC represented the natural distribution of tumor stages (with 66% in stage 0 or stage I). 94 patients with BLN were also included in the study (Fig. 1a). Ultra-deep targeted sequencing with the tumor driver gene panel (average de-duplication depth reaching 5000X) was performed on cfDNA and leukocyte genomic DNA from the same patient using paired-end molecular tags (UMI). The results showed that mutations detected in cfDNA and leukocytes were highly correlated (Fig. 1b), suggesting that these mutations were from leukocyte genomic DNA, and thus should be filtered in subsequent analyses. After filtration, a total of 153 mutations were detected in 67 (60.36%) samples of the 111 LC samples (Fig. 1c). A total of 28 mutations were detected in 23 (29.49%) samples of the 78 BLN samples (Fig. 1e). Mutations from cancer driver genes were also detected in the cfDNA of

patients with BLN, but the frequency of these mutations was relatively low and the mutation spectra were different from those of patients with LC (Fig. 1d).

Distinguishing between LC and BLN based on the classification model of somatic mutation

Prediction models were established based on detected mutations to distinguish between LC and BLN. The SUMAF model had an AUC of 0.67, a sensitivity of 55.9%, and a specificity of 76.9%. The weighted_SUMAF model had an AUC of 0.68, a sensitivity of 59.5%, and a specificity of 71.8% (Fig. 1f). These results indicated that classification models based only on mutation scores were limited in classifying LC and BLN plasma. Changes in the genomic sequence of cancer driver genes carried by BLN cfDNA might be more common than previously thought, limiting the effectiveness of mutation-based diagnostic analysis. Multi-analysis methods were likely to improve the detection of cancer signals.

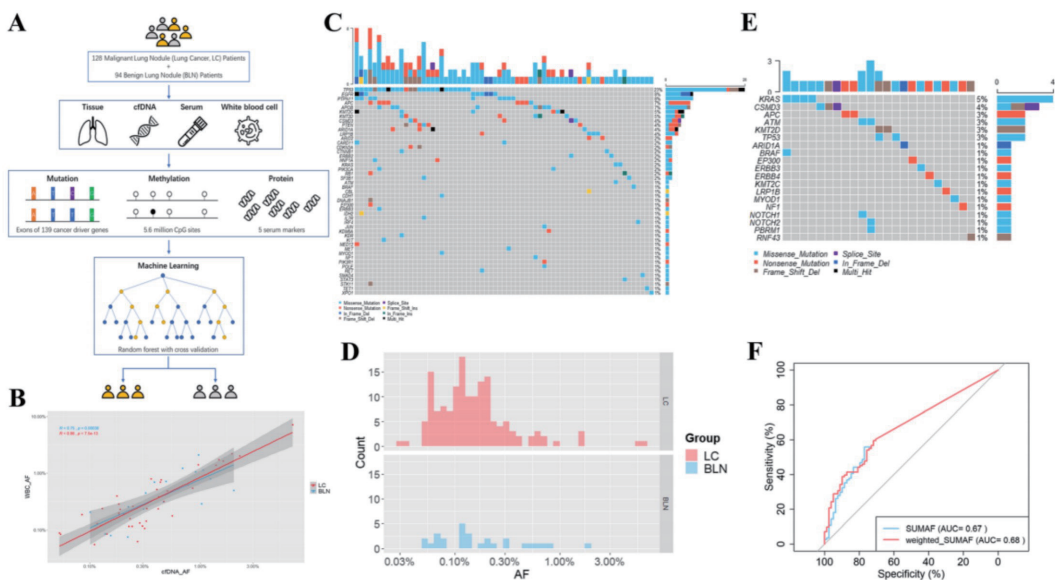


Fig.1. Research design and classification model for variations in plasma cfDNA, where shared variations are screening by matched WBC samples.

Classifying LC and BLN plasma based on cfDNA methylation data

To identify LC-specific epigenomics variation, the researchers performed whole-genome methylation sequencing on LC and paracarcinoma tissues (Fig. 2a) and identified a total of 315 LC-specific differentially methylated regions (DMRs; Fig. 2b). These included 293 highly methylated regions and 22 lowly methylated regions. GO (Gene Ontology) analysis suggested that the 293 highly methylated regions were mostly located in the transcription regulation areas (Fig. 2c). Subsequently, the researchers performed ultra-deep methylation targeted sequencing on the cfDNA from patients with LC and those with BLN, and screened methylation markers that could be used to distinguish between benign and malignant lung nodules via machine learning methods.

Identifying LC and BLN plasma by multi-omics analysis

Concurrently, the researchers also detected the presence of five protein markers in the plasma of patients, namely CEA, CYFRA21-1, NSE, CA19-9 and CA125. Only CEA levels were significantly higher in patients with LC than in patients with BLN ($p=0.04$, Student's t-test), with a staging AUC of 0.66 (Fig. 3 and 4). Thus CEA was included in the diagnosis model.

The results showed that in the same group of samples, the multi-omics prediction model based on wSUMAF mutation scores, regional methylation rates of 54 selected DMRs, and serum CEA levels could distinguish between patients with LC and those with BLN more effectively (AUC=0.78, Fig. 2d; AUC=0.74, Fig. 2e) than the model without CEA. The LC prognosis model based on cfDNA mutations and methylation markers was also more effective than the single-omics model (Fig. 2f, g; AUC=0.74, Fig. 2h).

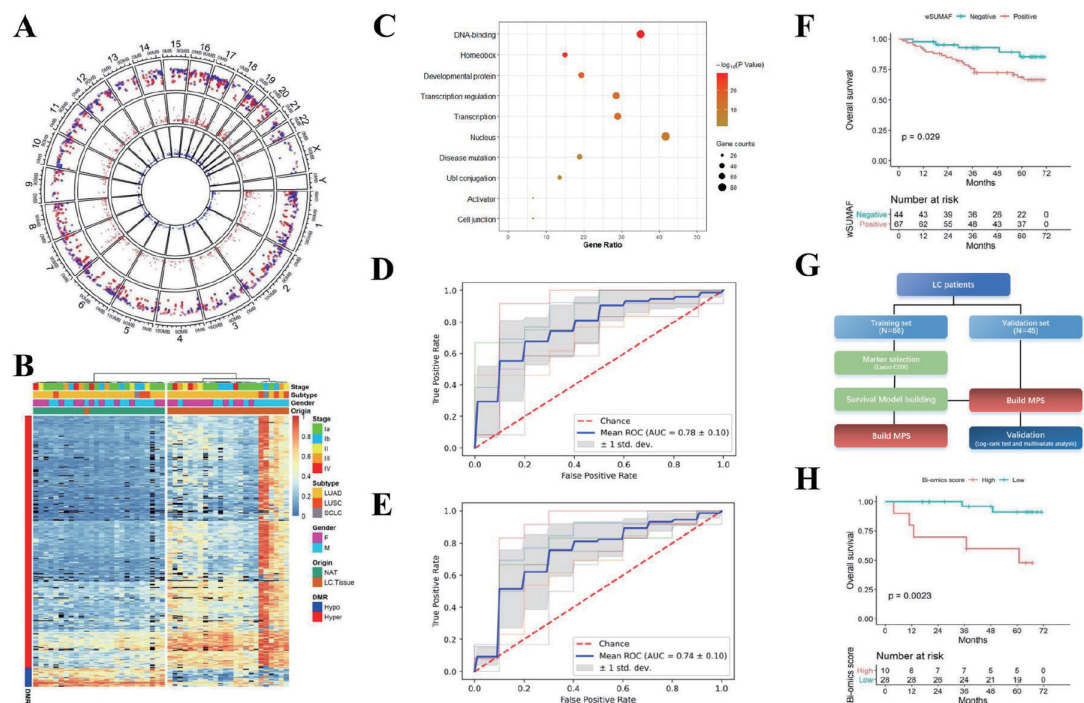


Fig. 2. DNA methylation and multi-omics analysis.

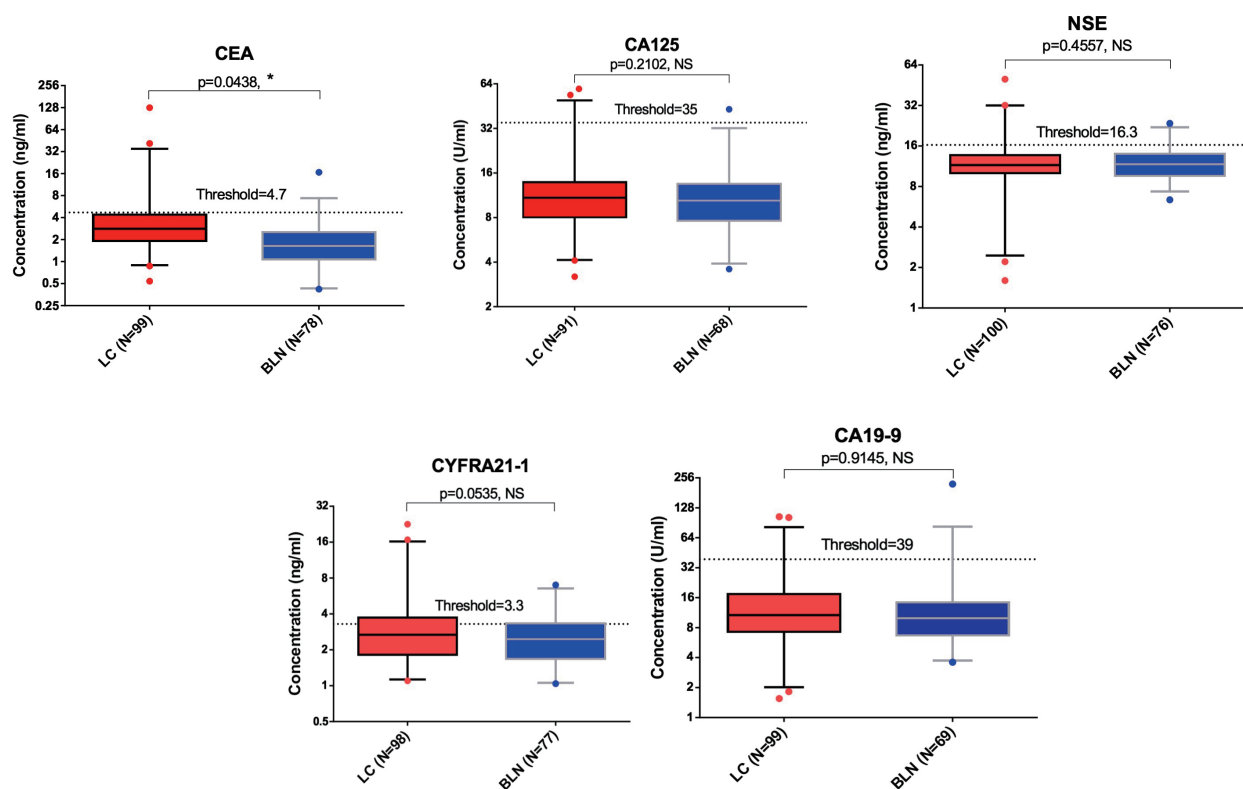


Fig. 3. Comparison of serum levels of five protein markers in patients with LC and BLN (t-test). The dashed line indicates the threshold for each marker commonly used in clinical practice.

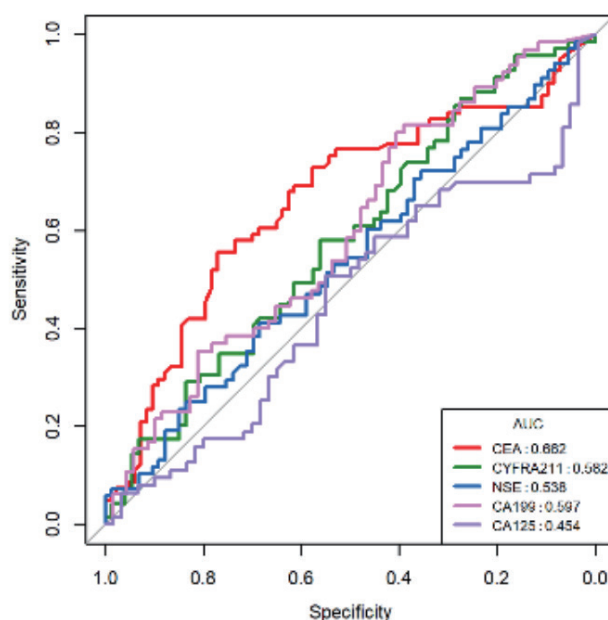


Fig. 4. Prediction model to distinguish between LC and BLN plasma cfDNA based on a single serum protein marker.

Summary

The joint model using genetic and epigenetic characteristics of cfDNA plus the serum protein marker CEA showed the best performance in classifying malignant and benign cases. In addition, the integrated model combining the cfDNA mutation status and the methylation-based prognostic marker might improve the prediction of survival for patients with LC.

These results were achieved using MGI's DNBSEQ-G400 sequencer. The researchers prepared the ultra-deep targeted sequencing library to detect changes in the genomic sequences of cfDNA and WBC gDNA. They also performed WGBS on LC tissues and paracarcinoma normal tissues to determine changes in LC-specific epigenetics. After amplification of the prepared library, PE100 sequencing was completed on the DNBSEQ-G400RS sequencer. After data analysis, the conclusion was drawn.



DNBSEQ-G400RS Genetic sequencer

References

1. Chen, K. *et al.* Non-invasive lung cancer diagnosis and prognosis based on multi-analyte liquid biopsy. *Mol Cancer* 20, 23, doi:10.1186/s12943-021-01323-9 (2021).
2. Zhang, S., Sun, K., Zheng, R., Zeng, H. & He, J. Cancer incidence and mortality in China, 2015. *Journal of the National Cancer Center* (2020).
3. Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C. & Boffetta, P. Risk factors for lung cancer worldwide. *European Respiratory Journal*, 889 (2016).
4. Nooreldeen, R. & Bach, H. Current and future development in lung cancer diagnosis. *International Journal of Molecular Sciences* 22, 8661 (2021).
5. Cipriano, L. E. *et al.* Lung cancer treatment costs, including patient responsibility, by disease stage and treatment modality, 1992 to 2003. *Value in Health* 14, 41-52 (2011).
6. Siena, S. *et al.* Dynamic Molecular Analysis and Clinical Correlates of Tumor Evolution Within a Phase 2 Trial of Panitumumab-Based Therapy in Metastatic Colorectal Cancer. *Annals of Oncology Official Journal of the European Society for Medical Oncology* (2017).
7. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359, eaar3247 (2018).
8. Lennon, A. M., Buchanan, A. H., Kinde, I., Warren, A. & Papadopoulos, N. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science* 369, eabb9601 (2020).
9. Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *American Association for the Advancement of Science* (2017).

Recommended Ordering Information

Category	Product	Cat. NO.
Instruments	DNBSEQ-G400RS Genetic Sequencer	900-000170-00
	MGISP-100RS Automated Sample Preparation System	900-000206-00
	MGISP-960RS Automated Sample Preparation System	900-000146-00
Software	MegaBOLT Bioinformatics analysis accelerator	900-000555-00
Library Prep	MGIEasy Whole Genome Bisulfite Sequencing Library Prep Kit (16 RXN)	1000005251
Sequencing Reagents	DNBSEQ-G400RS High-throughput Sequencing Set (FCL PE100)	1000016950

MGI Tech Co.,Ltd

Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA, 518083

+86-4000-688-114

en.mgi-tech.com

MGI-service@mgi-tech.com

The copyright of this brochure is solely owned by MGI Tech Co. Ltd.. The information included in this brochure or part of, including but not limited to interior design, cover design and icons, is strictly forbidden to be reproduced or transmitted in any form, by any means (e.g. electronic, photocopying, recording, translating or otherwise) without the prior written permission by MGI Tech Co., Ltd.. All the trademarks or icons in the brochure are the intellectual property of MGI Tech Co., Ltd. and their respective producers.

*1. For StandardMPS and CoolMPS: Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, Spain, UK, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland, Portugal, Austria and Romania. Unless otherwise informed, StandardMPS sequencing reagents, and sequencers for use with such reagents are not available in Hong Kong. No purchase orders for StandardMPS products will be accepted in the USA until after January 1, 2023.

2. For HotMPS sequencers: This sequencer is only available in selected countries, and its software has been specially configured to be used in conjunction with MGI's HotMPS sequencing reagents exclusively.

3. For HotMPS reagents: This sequencing reagent is only available in selected countries.