Machine learning of serum metabolic patterns encodes early-stage lung adenocarcinoma

Lin Huang
State Key Laboratory for Oncogenes and Related Genes, School of Biomedical Engineering, Shanghai Jiao Tong University

Kun Qian (k.qian@sjtu.edu.cn)
State Key Laboratory for Oncogenes and Related Genes, School of Biomedical Engineering, Shanghai Jiao Tong University

Method Article

Keywords: diagnostics, machine learning, cancer, metabolites, laser desorption/ionization mass spectrometry

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.pex-963/v1

Version of Record: A version of this preprint was published at Nature Communications on July 16th, 2020. See the published version at https://doi.org/10.1038/s41467-020-17347-6.

Abstract

Early cancer detection greatly increases the chances for successful treatment, but available diagnostics for some tumours, including lung adenocarcinoma (LA), are limited. An ideal early-stage diagnosis of LA for large-scale clinical use must address quick detection, low invasiveness, and high performance. Here, we conduct machine learning of serum metabolic patterns to detect early-stage LA. We extract direct metabolic patterns by the optimized ferric particle-assisted laser desorption/ionization mass spectrometry within 1 second using only 50 nL of serum. We define a metabolic range of 100-400 Da with 143 m/z features. We diagnose early-stage LA with sensitivity~70-90% and specificity~90-93% through the sparse regression machine learning of patterns. We identify a biomarker panel of seven metabolites and relevant pathways to distinguish early-stage LA from controls ($p < 0.05$). Our approach advances the design of metabolic analysis for early cancer detection and holds promise as an efficient test for low-cost rollout to clinics.

Introduction

Here, we optimize the laser desorption/ionization mass spectrometry (LDI MS) approach to analyse a large range of metabolites (including biologically relevant metabolites) as metabolic patterns from serum samples without pretreatment by improving the substrate used. Further encoded by machine learning algorithm, the serum metabolic patterns achieve high specificity and sensitivity diagnosis of early-stage LA and enable large-scale and low-cost rollout for use in clinics.

Reagents

Ferric chloride (purity > 97%), trisodium citrate (purity > 99%), ethylene glycol, sodium acetate (purity > 99%), tetraethyl orthosilicate (TEOS, purity > 96%), absolute ethanol (EtOH), trifluoroacetic acid (TFA, purity > 99%), and ammonium hydroxide (purity > 10%-35%) were purchased from Sinopharm Chemical Reagent Beijing Co. Ltd. (Beijing, China). Resorcinol (purity > 99%) was purchased from J&K China Chemical Ltd (Shanghai, China). Albumin from bovine serum (BSA, purity > 98%), α-cyano-4-hydroxycinnamic acid (CHCA, purity > 99%), Acetonitrile (ACN, purity > 99%), standards including cysteine (purity > 99%), uric acid (purity > 99%), D-glucose (purity > 99.5%), sucrose (purity > 99%), D-mannitol (purity > 99%), L-leucine (purity > 98%), L-cellobiose (purity > 99.5%), L-lysine (purity > 98%), valine (purity > 99%), DL-phenylalanine (purity > 99%), and arginine (purity > 99%) were purchased from Sigma, USA. Formaldehyde solution (CH_2O, purity > 36.0%) and standards including histamine (purity > 99%), uracil (purity > 99%), 3-hydroxypicolinic acid (purity > 99%), indoleacrylic acid (purity > 99%), and fatty acid (18:2) (FA) (purity > 99%) were purchased from Shanghai Aladdin Reagent Co. Ltd. (Shanghai, China). All aqueous solutions were prepared using deionized water (18.2 MΩ cm, Milli-Q, Millipore, GmbH).

Equipment

Oven

Centrifuge

Magnetic stirrer

5800 Proteomics Analyzer (Applied Biosystems, Framingham, MA, USA)

Procedure

Synthesis of substrate materials

1. Ferric chloride was first dissolved in ethylene glycol solution.

2. Trisodium citrate (weights from 0 to 0.8 g) was then added to tune the surface charge of the products.

3. Sodium acetate was added to the mixture and sonicated at room temperature for 30 min.

4. The reaction mixture was transferred to a Teflon-lined stainless-steel autoclave (capacity 50 mL) and held at 200°C for 10 h for the formation of ferric particles.

MS data acquisition

1. Ferric particles were dispersed in water as the matrix for LDI MS analysis at a concentration of 1 mg mL^{-1}.

2. 500 nL of matrix slurry was mixed with 50-500 nL of analyte solution on the plate and dried for LDI MS analysis.

3. Set the 5800 Proteomics Analyzer with a repetition rate of 200 Hz and an acceleration voltage of 20 kV. The delay time for this experiment was optimized to 250 ns.

4. Collected the raw MS data based on the experimental parameters above.

Preparation of clinical samples

1. All blood samples were drawn by venepuncture and clotted at room temperature within 40 minutes.

2. Serum samples were obtained by centrifuging at 5,100 xg and 4°C for 10 minutes.

3. After centrifugation, the precipitate was discarded and the supernatant serum was stored at -80°C immediately (within 15 minutes).

4. The elapsed time was within 1 hour between blood draw, centrifugation, and ultimate storage at -80°C.

Machine learning and computer-assisted diagnosis

1. Pre-processing of the raw mass spectra data, including baseline correction, peak detection, extraction, alignment, normalization, and standardization, was carried out by MATLAB (R2016a, The MathWorks, Natick, MA) prior to pattern recognition analysis.

2. The pre-processed MS data were considered as the inputs (serum metabolic patterns) to train and test the classifier. A 5-fold cross-validation approach was performed to estimate the performance of the classifier for both the inner-loop and outer cross-validation (20 rounds for each fold, thus 100 models for outer cross-validation in total).

3. An external double-bind test for differentiating early-stage LA from healthy controls were conducted based on the as-trained classifier. The disease labels of the double-bind test cohort were unknown and predicted by the classifier.

Troubleshooting

Time Taken

1. The time for the synthesis procedure is approximately to be 12 hours.

2. The preparation of matrix and analytes for LDI MS analysis takes an experienced researcher 1 minutes.

3. The collection procedure of MS data under laser irradiation takes 1 second.

4. The machine learning of MS data takes 30 seconds.

5. All in all the whole procedure for serum metabolic pattern based diagnosis can be comfortably completed in 3 minutes by an experienced researcher.

Anticipated Results

A computer-aided diagnosis based on serum metabolic patterns (obtained from ferric particle-assisted LDI MS) can be used to detect early-stage lung adenocarcinoma (LA).

References

Huang, L. et. al. Plasmonic silver nanoshells for drug and metabolite detection. Nat. Commun. **8**, 220 (2017)

Sun, X. et. al Metabolic fingerprinting on a plasmonic gold chip for mass spectrometry based in vitro diagnostics. ACS Central Sci. 4, 223-229 (2018)

Acknowledgements

We are grateful for the financial support from Project 81971771 and 81771983 by National Natural Science Foundation of China (NSFC), Project 2017YFE0124400 and 2017YFC0909000 by Ministry of Science and Technology of China, Innovation Group Project of Shanghai Municipal Health Comission (2019CXJQ03), and Project 16CR2011A by Clinical Research Plan of SHDC. This work was also sponsored by the Shanghai Rising-Star Program (19QA1404800) and Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.