Electronic Supplementary Material (ESI) for ChemComm.

Towards Rapid Prediction of Drug-resistant Cancer Cell Phenotypes: Single Cell Mass Spectrometry Combined with Machine Learning

Renmeng Liu, † Genwei Zhang, † and Zhibo Yang*

Electronic Supplementary Information (ESI)

Table of Content

Materials and Chemicals

Single-probe fabrication. Fused silica capillaries (O.D. = 103.0 μm, I.D. = 39.0 μm) (Polymicro Technologies, Phoenix, AZ). Dual bore quartz tubing, 1.120''×0.005"×12" (Friedrich & Dimmock Inc., Millville, NJ). UV epoxy (Prime Dental Inc., Chicago, IL). UV lamp (Foshan Liang Ya Dental Equipment, China). Laser-Based Micropipette Puller (P-2000, Sutter Instrument, Novato, CA). Conductive union (IDEX Health & Science LLC, Oak Harbor, WA).

Cell culture and sample preparation. Chronic myelogenous leukemia cell line (K-562, ATCC, Manassas, VA). RPMI 1640 culture medium (Gibco by Life Technologies, Long Island, NY). Fetal bovine serum (FBS) (Gibco by Life Technologies, Long Island, NY). Penicillin/streptomycin (Gibco by Life Technologies, Long Island, NY). Phosphate buffered saline (PBS) buffer (Gibco by Life Technologies, Long Island, NY). Fibronectin (Millipore Sigma, St. Louis, MO). Bovine serum albumin (Millipore Sigma, St. Louis, MO). Trypan blue solution 0.4% (Gibco by Life Technologies, Long Island, NY). Methanol (UHPLC-MS, Fluka Analytical, Mexico City, Mexico). Formic acid (0.1%) in water (LC-MS, Honeywell, Morris Plains, NJ). Chloroform (HPLC, Millipore Sigma, St. Louis, MO). Hemacytometer (Hausser Scientific, Horsham, PA). Cell culture flask (Cellstar, Greiner Bio-One North America Inc., Monroe, NC). Centrifuge tube (15 mL, Corning Co., Corning, NY). Thermanox coverslip (15 mm in diameter, Thermofisher Scientific Inc., Waltham, MA). 12-well plates (Cellstar, Greiner Bio-One North America Inc., Monroe, NC).

SCMS experiments. XYZ-manipulator (M-MT-XYZ, Newport Co., Irvine, CA). Syringe pump (Nexus 3000, Chemyx Inc., Stafford, TX). High resolution stereo microscope (Shenzhen D&F Co., China). Motorized XYZ-stage (MFA-CC, Newport Co., Irvine, CA) with Labview software package.[1] Thermo LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA). Acetonitrile (UHPLC-MS, Fluka Analytical, Mexico City, Mexico). Formic acid (AR, Avantor Performance Materials, LLC., Center Valley, PA).

Cell Culture and Sample Preparation

K-562 cells were cultured using RPMI 1640 medium (with 10% FBS and 100 U/mL penicillin/streptomycin antibiotic solution) in a humidified cell culture incubator (Heracell, Thermo Scientific Inc.) supplied with 5% CO_2 at 37 °C, and they were subcultured every 4 days according to the recommended protocols provided by ATCC. When K-562 cells were in logarithmic growth phase and reached optimal cell density, we performed sample preparation based on published protocols[2] with adaptions to suit for our single cell mass spectrometry (SCMS) analysis of both phenotypes (i.e., phenotype I and phenotype II, see main text). In particular, we placed Thermanox coverslips in the 12-well plate and coated them with 400 μL 40 μg/mL fibronectin (FN) solution (prepared in PBS) per well for 12 h at room temperature.

We subsequently removed FN solution and added 400 μL 0.1% BSA solution (prepared in PBS) per well to eliminate nonspecific bindings for 1 h. Meanwhile, we centrifuged K-562 cells from the cell culture flask and washed them twice using FBS-free RPMI 1640 medium. Washed cells were resuspended in FBS-free RPMI 1640 medium and were seeded (~1×10^5 cells per well) onto the FN-coated coverslips. Seeded cells were maintained in the incubator for 10 h to develop integrin-ECM interaction. We then aspirated and collected culture medium with suspended cells (phenotype II), and used FBS-free medium to wash the coverslips twice before immediately subjecting adherent cells on the coverslip (phenotype I) to the following SCMS analysis. Collected phenotypic II cells were transferred to clean coverslips and subsequently subjected to SCMS analysis. During one specific sample preparation procedure, we used trypan blue solution to stain cells on the coverslips and examined cell viability for both phenotypes. We found nearly all cells were alive prior to SCMS experiments.

SCMS Experiments

We used a microscale multifunctional device, the Single-probe, to interrogate individual cells and obtained metabolomic information *in situ*, in real time, and in ambient conditions. Detailed working mechanisms of the Single-probe were reported in our previous publications,[3] and only a brief description is provided here. Sampling solvent (acetonitrile with 0.1% formic acid) was introduced through the solvent-providing capillary (flow rate = 0.1 μL/min) and formed a liquid junction at the tip of dual-bore quartz tubing (tip size ~8 μm). When the tip was inserted into a specific cell, the sampling solvent mixed with cellular contents and subsequently extracted them from the cell. Such mixture was immediately drawn towards the tip of the nano-ESI emitter, where ionization occurred. We applied ionization voltage (+4.5 kV) to the sampling solvent through a conductive union to maintain a stable MS signal. Coupled to a mass spectrometer (LTQ Orbitrap XL), the Single-probe was used to acquire background signals, metabolomic profiles of phenotype I cells (n=100), and metabolomic profiles of phenotype II cells (n=108), respectively. The following SCMS experiments were performed with MS settings as: ionization voltage +4.5 kV, mass range 150–1200, mass resolution 60,000 at *m/z* (mass-to-charge ratio) 400, 1 microscan, 100 ms max injection time and automatic gain control (AGC) on.

MS data of background. Background signals originated from the sampling solvent, cell culture medium and substrates (FN-coated coverslips) may affect our downstream data analysis (i.e., inducing type I error), and therefore, should be excluded from our SCMS datasets. In particular, we placed a FN-coated coverslip with a droplet of FBS-free RPMI 1640 medium underneath the Single-probe, and we then placed the tip of the Single-probe onto the coverslip to detect any species under the SCMS experimental condition described above. The acquired MS signals were regarded as background.

SCMS data acquisition of phenotype I cells. Prepared coverslip with phenotype I cells were placed underneath the Single-probe. Through a high resolution stereo microscope (Figure 2A), we were able to locate individual cells (Figure 2B), and precisely insert the Single-probe tip into a target cell by manipulating the motorized XYZ-stage (increment step size = 0.1 μm). Meanwhile, we could observe the profile change of mass spectra from background to cellular contents.

SCMS data acquisition of phenotype II cells. Phenotype II cells suspended in culture medium were transferred to a clean coverslip placed underneath the Single-probe. As the culture medium partially evaporated, cells can be immobilized when they were still surrounded by a small amount of culture medium. We then followed similar measurement process as phenotype I cells to acquire metabolomic information of phenotype II cells.

SCMS Data Analysis

We performed a comprehensive data analysis of acquired raw SCMS data sets to obtain biological information of intracellular species while excluding interference from background and noise, as described in the following.

Data pre-treatment. Multiple steps are involved in SCMS data pre-treatment procedure. The pretreated datasets can then be used for the following statistical analysis and machine learning methods. First, $[PC(34:1) + Na]^+$ (*m/z* = 782.5676), a commonly detected cellular species,[3a, 4] was selected as the indicator of time spans when single cells were under SCMS measurements. Second, we exported MS peaks of each single cell from Xcalibur to Excel, followed by background extraction (i.e., elimination of MS signals detected from background such as the sampling solvent and RPMI culture medium) and noise reduction (i.e., elimination of MS peaks of low abundance, typically those with ion intensity < 10^3) using our in-home developed software. Third, we normalized the ion intensity of each cellular metabolite to the total ion intensity of all detected species in each single cell after background extraction and noise reduction. Fourth, we employed Geena 2 (http://bioinformatics.hsanmartino.it/geena2),[5] an online peak alignment tool, to align all single cell MS datasets. Last, we used MetaboAnalyst (http://www.metaboanalyst.ca),[6] an online metabolomics data analysis software, to select cellular metabolites in our datasets given different missing value thresholds (MVTs, see main text), followed by log-transformation of normalized ion intensity of each metabolite.

Statistical analyses. Multiple statistical analyses were performed to gain biological insights. Specifically, *t*-distributed stochastic neighbor embedding (*t*-SNE), provided as a built-in function (tsne()) in Statistics and Machine Learning Toolbox of MATLAB (MathWorks, 2017a), was used to visualize metabolomic profiles of single cells corresponding to both phenotypes. Using gscatter() function in MATLAB, the discrimination between two phenotypes after dimensionality

reduction can be achieved in 2D space (Fig. 3). The optimized parameters of *t*-SNE are: maximum iterations = 5000, perplexity = 30, number of PCA components = 50, verbose = 1, learn rate = 500, and exaggeration = 6. In addition, we performed two-sample *t*-test and principle component analysis (PCA) as two separate approaches to discover biomarkers characteristic to phenotypes I and II. In the first approach, we performed Student's *t*-test and Welch's *t*-test respectively for cellular metabolites with equal and unequal variance (tested through Levene's *F*-test), respectively, using an in-house developed software (source code available upon request). Metabolites with *t*-test *p*-values < 0.05 were marked as metabolic biomarkers (i.e., species with significant difference in abundance between two phenotypes). In the second approach, we performed PCA of pre-treated SCMS metabolomics dataset with 40% MVT, followed by selection of metabolites with high PC1 and PC2 loading scores as biomarkers.[7] All datasets were subjected to log-transformation and pareto scaling prior to PCA.

Machine Leaning (ML) and Model Evaluation

In this work, we implemented modern ML algorithms for the analysis of pre-treated SCMS datasets, and predicted drug-resistant phenotypes (i.e., with or without CAM-DR) of single cells in a rapid and reliable fashion.

ML model construction. Three ML methods, i.e., random forest (RF), penalized logistic regression (LR), and artificial neural network (ANN), were used and compared using in-house developed R scripts. The original data was randomly split with 80% single cells being selected as the training data and the rest 20% cells being selected as the testing data. Training data were used to train the model, while testing data were used to evaluate the accuracy of the model prediction. Since our datasets contain cells from two phenotypes, binary classification response was chosen for all three aforementioned approaches. The optimized parameters of RF (R package 'randomForest') were: mtry = 7, ntree = 500, and type = classification. Penalized LR was chosen in our study because full LR approach shown severe over-fitting issue when the number of variables increases. Elastic net LR (with $\alpha = 0.5$) was used due to its robustness and high performance on our dataset compared with lasso and ridge LR. ANN was performed using R package 'neuralnet'. One hidden layer and ten neurons were optimized to achieve the best prediction on the data, and the logistic function was chosen as the activation function to smooth the results of the cross product of the neurons and the weights. All source codes are available upon request.

ML model evaluation. The receiver operating characteristic (ROC) curve was constructed to illustrate the capability of the constructed binary classifier. Sensitivity was represented by the ratio of true positive to the sum of true positive and false negative, whereas specificity was represented by the ratio of true negative to the sum of true negative and false positive.[8] The area under the curve (AUC) for each model was calculated using the sum of the area underlying polygons.

In addition, k-fold (k = 5) cross-validation (CV) was performed to validate the robustness of each model and avoid overfitting. The final output was the averaged misclassification error of five independent validations.

Method Validation

To validate our method of combined SCMS experiment and ML models, we prepared another batch of K-562 cells and subjected them to SCMS measurements and ML data analysis as described in the main text (n = 15 for phenotype I cells, n = 16 for phenotype II cells). The acquired dataset with 40% MVT was subjected to ML model prediction.

Identification of Metabolic Biomarkers

To further identify the detected metabolic biomarkers of cells possessing CAM-DR, we performed LC-MS/MS analysis of cell lysate. According to our SCMS data, these metabolites exist in both phenotypes, whereas their relative abundances are different for each phenotype. Therefore, cells were cultured without being separated into two different phenotypes prior to the preparation of cell lysates. Cell lysate preparation was conducted by following a conventional protocols used for LC-MS metabolomics studies[9] with minor adaptions. Briefly, 5×10^5 K-562 cells were centrifuged and washed twice with PBS before being resuspended in 1 mL of extraction solvent (methanol:water:chloroform = 1:1:1, v/v/v). After vortexing on ice for 10 min, the mixture was centrifuged at 14000 rpm for 15 min. Extracted metabolites in methanol/water and chloroform were transferred to Eppendorf tubes separately followed by solvent evaporation in SpeedVac. Dried samples were reconstituted in 120 μL methanol followed by vortexing and centrifugation. Last, 100 μL of supernatant was transferred to LC-MS sample vials. The analytical column was a Waters Acquity UPLC HSS T3 (1.8 μm, 300 μm × 100 mm). The mobile phase was: A (water with 0.1% formic acid) and B (acetonitrile with 0.1% formic acid). The separative gradient was: 0 min 20% B, 3.5 min 35% B, 18 min 65% B, 21 min 99% B, 34 min 99% B with a constant flow rate of 12 μL/min. LC-MS/MS analysis was carried out using a Synapt G2-Si high resolution quadrupole time-of-flight mass spectrometer (Waters Corp., Milford, MA). The instrumental parameters were listed as follows: source voltage +3.2 kV, sampling cone 30, source offset 50, source temperature 80 °C, cone gas 50 L/h, mass range 150–1200, mass resolution 50000, scan time 1 s, collisional energy ramp 25–35 eV. Acquired LC-MS/MS spectra were searched through online metabolome databased such as METLIN (https://metlin.scripps.edu) and HMDB (http://www.hmdb.ca).[10] Metabolic biomarkers with matched accurate mass (within ±5 ppm) and MS/MS fragmentation patterns were identified as shown in Table S4.

Model Comparison

To compare the predictive accuracy between the ANN model constructed using SCMS dataset with 40% MVT (Model A) and the model based on metabolic biomarkers discovered through *t*-test (Model B) or loadings of PCA (Model C), we constructed both models (procedures are described in the "ML model construction" section), and used the same testing set to evaluate the model performance.

Model A *vs.* Model B. Model A contains 131 variables (metabolites), whereas Model B contains 70 variables (biomarkers discovered using two-sample *t*-test). Through 8 independent predictions, Model A generated 95.7 ± 2.6% predictive accuracy, whereas Model B generated 91.9% ± 4.4% predictive accuracy. By performing Welch's one-tail *t*-test, we conclude that Model A grants significantly higher predictive accuracy (p-value = 0.029, $\alpha = 0.05$).

Model A *vs.* Model C. Model A contains 131 variables (metabolites), whereas Model C contains 86 variables with high PC1 and PC2 loadings (biomarkers discovered using PCA loading plot, see Fig. S1). Through 8 independent predictions, Model A generated 95.7 ± 2.6% predictive accuracy, whereas Model C generated 92.6% ± 4.1% predictive accuracy. By performing Welch's one-tail *t*-test, we conclude that Model A grants significantly higher predictive accuracy (p-value = 0.046, $\alpha = 0.05$).

Supporting Tables

Table S1. RF model evaluation.

Missing value threshold (MVT)	Number of variables	Predictive accuracy	Computing time*	5-fold CV error rate
0%	7	77.1% ± 10.2%	7 s	0.206
20%	58	83.8% ± 7.0%	16 s	0.139
40%	131	91.9% ± 4.3%	30 s	0.106
50%	203	94.8% ± 4.2%	45 s	0.086
60%	358	92.9% ± 2.9%	80 s	0.043
70%	641	93.8% ± 2.7%	135 s	0.043
80%	1296	94.3% ± 2.7%	300 s	0.048
90%	3232	94.3% ± 2.7%	930 s	0.048

*Computation was performed using iMac18,1, Intel Core i5 processor with 2.3 GHz and 8 GB RAM.

Table S2. Penalized LR (elastic net) model evaluation.

Missing value threshold (MVT)	Number of variables	Predictive accuracy	Computing time*	5-fold CV error rate
0%	7	90.8% ± 2.7%	12 s	0.12
20%	58	88.1% ± 5.2%	68 s	0.158
40%	131	90.8% ± 3.2%	45 s	0.106
50%	203	91.5% ± 4.8%	32 s	0.067
60%	358	93.3% ± 2.6%	40 s	0.063
70%	641	92.2% ± 2.5%	60 s	0.057
80%	1296	94.7% ± 1.8%	120 s	0.048
90%	3232	94.7% ± 3.2%	330 s	0.043

*Computation was performed using iMac18,1, Intel Core i5 processor with 2.3 GHz and 8 GB RAM.

Table S3. ANN model evaluation.

Missing value threshold (MVT)	Number of variables	Predictive accuracy	Computing time*	5-fold CV error rate
0%	7	89.5% ± 3.2%	30 s	0.134
20%	58	92.9% ± 2.9%	4 s	0.115
40%	131	95.7% ± 2.6%	6 s	0.062
50%	203	93.4% ± 1.0%	9 s	0.082
60%	358	95.2% ± 2.4%	16 s	0.062
70%	641	94.8% ± 4.2%	40 s	0.071
80%	1296	96.2% ± 2.7%	145 s	0.067
90%	3232	93.8% ± 3.2%	924 s	0.074

*Computation was performed using iMac18,1, Intel Core i5 processor with 2.3 GHz and 8 GB RAM.

Table S4. Identified metabolic biomarkers using LC-MS/MS from cell lysate.

m/z	Name	Formula	Species
650.439	PC(25:0(CHO))	$C_{33}H_{64}NO_9P$	$[M + H]^+$
723.493	PA(36:2)	$C_{39}H_{73}O_8P$	$[M + Na]^+$
725.556	SM(d34:1)	$C_{39}H_{79}N_2O_6P$	$[M + Na]^+$
732.553	PE(35:1)	$C_{40}H_{78}NO_8P$	$[M + H]^+$
742.573	PE(O-35:0)	$C_{40}H_{82}NO_7P$	$[M + Na]^+$
744.590	PE(O-37:2)	$C_{42}H_{82}NO_7P$	$[M + H]^+$
746.569	PE(36:1)	$C_{41}H_{80}NO_8P$	$[M + H]^+$
754.535	PC(32:1)	$C_{40}H_{78}NO_8P$	$[M + Na]^+$
756.552	PC(32:0)	$C_{40}H_{80}NO_8P$	$[M + Na]^+$
758.569	PC(34:2)	$C_{42}H_{80}NO_8P$	$[M + H]^+$
760.585	PC(34:1)	$C_{42}H_{82}NO_8P$	$[M + H]^+$
766.535	PE(36:2)	$C_{41}H_{78}NO_8P$	$[M + Na]^+$
766.572	PE(P-37:1)	$C_{42}H_{82}NO_7P$	$[M + Na]^+$
768.588	PE(36:1)	$C_{41}H_{80}NO_8P$	$[M + Na]^+$
780.551	PC(34:2)	$C_{42}H_{80}NO_8P$	$[M + Na]^+$
782.567	PC(34:1)	$C_{42}H_{82}NO_8P$	$[M + Na]^+$
786.600	PC(36:2)	$C_{44}H_{84}NO_8P$	$[M + H]^+$
788.616	PC(36:1)	$C_{44}H_{86}NO_8P$	$[M + H]^+$
796.525	PE(37:2)	$C_{42}H_{80}NO_8P$	$[M + K]^+$
806.567	PC(36:3)	$C_{44}H_{82}NO_8P$	$[M + Na]^+$
808.582	PC(36:2)	$C_{44}H_{84}NO_8P$	$[M + Na]^+$
810.599	PC(36:1)	$C_{44}H_{86}NO_8P$	$[M + Na]^+$
814.632	PC(38:2)	$C_{46}H_{88}NO_8P$	$[M + H]^+$
822.540	PE(39:3)	$C_{44}H_{82}NO_8P$	$[M + K]^+$
824.556	PC(36:2)	$C_{44}H_{84}NO_8P$	$[M + K]^+$
826.572	PC(36:1)	$C_{44}H_{86}NO_8P$	$[M + K]^+$
836.615	PC(38:2)	$C_{46}H_{88}NO_8P$	$[M + Na]^+$
838.631	PC(38:1)	$C_{46}H_{90}NO_8P$	$[M + Na]^+$

Supporting Figures

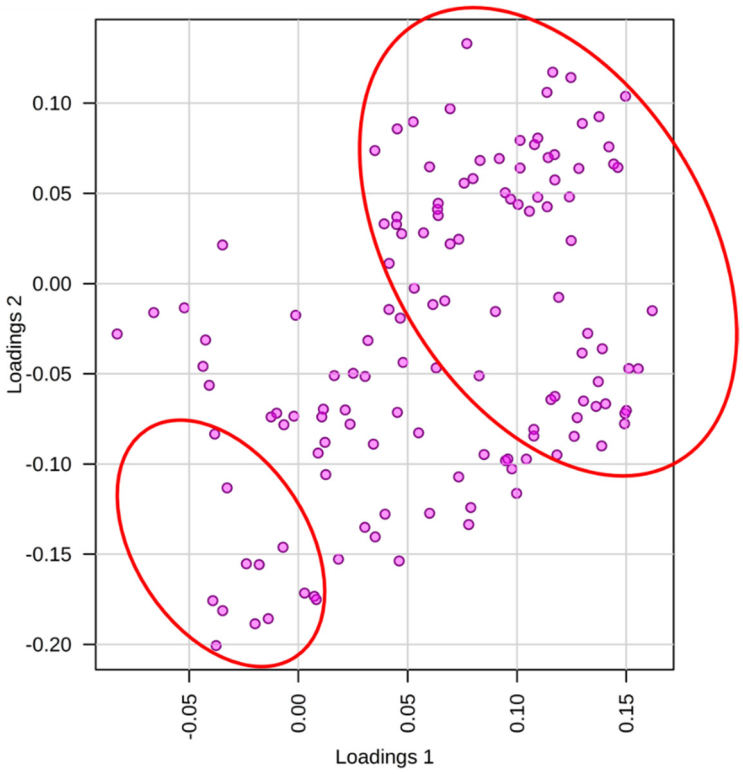

Fig. S1 Loading plot of PCA corresponding to SCMS metabolomics dataset with 40% MVT. A total number of 86 cellular metabolites (within red circles) with high PC1 and PC2 loading scores were selected as biomarkers.

References

1 I. Lanekoff, B. S. Heath, A. Liyu, M. Thomas, J. P. Carson and J. Laskin, *Anal. Chem.*, 2012, **84**, 8351-8356.
2 a) J. S. Damiano, A. E. Cress, L. A. Hazlehurst, A. A. Shtil and W. S. Dalton, *Blood*, 1999, **93**, 1658-1667; b) J. S. Damiano, L. A. Hazlehurst and W. S. Dalton, *Leukemia*, 2001, **15**, 1232.
3 a) N. Pan, W. Rao, N. R. Kothapalli, R. M. Liu, A. W. G. Burgett and Z. B. Yang, *Anal. Chem.,* 2014, **86**, 9376-9380; b) N. Pan, W. Rao, S. J. Standke and Z. B. Yang, *Anal. Chem.*, 2016, **88**, 6812-6819.
4 Y. Schober, S. Guenther, B. Spengler and A. Rompp, *Anal. Chem.*, 2012, **84**, 6293-6297.
5 P. Romano, A. Profumo, M. Rocco, R. Mangerini, F. Ferri and A. Facchiano, *Bmc Bioinformatics*, 2016, **17**, 61.
6 J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D. S. Wishart and J. Xia, *Nucleic Acids Res.*, 2018, **46**, W486-W494
7 P. Nemes, A. M. Knolhoff, S. S. Rubakhin and J. V. Sweedler, *Anal. Chem.*, 2011, **83**, 6810-6817.
8 J. G. Xia, D. I. Broadhurst, M. Wilson and D. S. Wishart, *Metabolomics*, 2013, **9**, 280-299.
9 X. Luo and L. Li, *Anal. Chem.*, 2017, **89**, 11664-11671.
10 D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach and A. Scalbert, *Nucleic Acids Res.*, 2018, **46**, D608-D617.