ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

АЛГОРИТМ ОБРАБОТКИ МАСС-СПЕКТРОМЕТРИЧЕСКИХ ДАННЫХ ДЛЯ ПОЛУЧЕНИЯ ДИАГНОСТИЧЕСКОЙ ПАНЕЛИ МОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ НА ПРИМЕРЕ ПОИСКА МАРКЕРОВ МЕТАСТАЗИРОВАНИЯ ПРИ РАКЕ МОЛОЧНОЙ ЖЕЛЕЗЫ

А.О. Токарева[1,2], В.В. Чаговец[3], А.С. Кононихин[1,3], Н.Л. Стародубцева[2,3], В.Е. Франкевич[3*], Е.Н. Николаев[1*]

[1]Сколковский институт науки и технологий, 121205, Москва,
Территория Инновационного Центра «Сколково», Большой бульвар д.30; стр.1. e-mail: e.nikolaev@skoltech.ru
[2]«Институт энергетических проблем химической физики имени В.Л. Тальрозе» Федерального исследовательского центра химической физики имени Н.Н. Семёнова Российской академии наук, 119334, Москва, Ленинский проспект 38 стр. 2
[3]Национальный медицинский исследовательский центр акушерства, гинекологии и перинатологии им. академика В.И. Кулакова, 117513, Москва, ул. Академика Опарина, 4, стр 2

Диагностика патологии по молекулярным маркерам является перспективным направлением клинической медицины, в котором масс-спектрометрия (МС) яслужит одним из методов, используемых для получения информации о молекулярных профилях. В контексте извлечения соединений, играющих ключевую роль для классификации патология/болезнь, важное значение имеет обработка полученных данных, часто включающих несколько сотен детектированных соединений. В данной работе предложен алгоритм обработки данных на примере данных МС, полученных при анализе опухолевой ткани и здоровой молочной железы, с целью выделения липидных маркеров метастазирования. В результате его реализации получен набор соединений, относящихся к классам липидов, связанных с процессами метастазирования и пролиферации, и позволяющих построить высокочувствительную диагностическую модель на основе логистической регрессии. Предложенный метод потенциально пригоден для обработки данных МС, полученных при анализе молекулярного профиля биоматериала другого базиса (метаболом, протеом, гликом).

Ключевые слова: масс-спектрометрия; обработка данных; биологические маркеры

DOI: 10.18097/BMCRM00156

ВВЕДЕНИЕ

Использование молекулярных маркеров является одним из способов диагностики заболевания. Масс-спектрометрия (МС) является чувствительным методом молекулярного анализа, который позволяет эффективно производить как целевой, так и нецелевой анализ образца, что делает его эффективным для поиска потенциальных маркеров заболевания в полученном молекулярном профиле образца. Использование предварительного хроматографического разделения увеличивает число соединений, регистрируемых масс-спектрометром в образце, и повышает точность идентификации соединений. Липиды – группа амфифильных соединений низкой молекулярной массы (меньше 2000 Да), участвующая в построении мембран клеток, передаче внутриклеточных и межклеточных сигналов и транспорте, запасании энергии и др. Уже показана взаимосвязь уровней отдельных липидов и классов липидов с наличием заболевания: ростом уровня фосфатидилхолинов при раке [1], падением уровня сфингомиелинов при болезни Альцгеймера [2]. Также в ряде работ получены панели липидов-маркеров для диагностики заболевания [3–5].

В данной работе предлагается последовательность действий для обработки данных, полученных с использованием хромато-масс-спектрометрического анализа для создания диагностической панели маркеров на примере диагностики метастазирования в регионарные лимфоузлы при раке молочной железы.

МАТЕРИАЛЫ И МЕТОДЫ

В работе использованы образцы ткани от 40 пациентов с раком молочной железы без регионарного метастазирования (средний возраст 56.3 года, у 20 пациентов стадия I, у 20 – стадия II. Рак у 20 пациентов относился по классификации TNM относился к группе T1N0M0, у 20 – к группе T2N0M0) и от 48 пациентов с раком молочной железы с регионарным метастазированием (средний возраст пациентов составил 56.7 лет, у 32 пациентов была II стадия рака, у 16 – III стадия рака. Рак у 10 пациенток по классификации TNM относился к группе T1N1M0, у 36 – к группе T2N1-3M0, у 2 – к группе T3N3M0): биопсийные материалы опухолевой ткани молочной железы и нормальной ткани молочной железы. Таким образом, пул образцов состоял из 4 клинических групп:

- группа 1 – нормальная ткань молочной железы от пациентов без метастазирования;
- группа 2 – ткань нормальной ткани молочной железы от пациентов с метастазированием;
- группа 3 – опухолевая ткань молочной железы от пациентов без метастазирования;
- группа 4 – опухолевая ткань молочной железы от пациентов с метастазированием.

Из образца тканей весом 40 г методом Фолча [6] выделяли липиды с последующим перерастворением в 200 мкл изопропанол/ацетонитрил 1/1. На основе 10 мкл. от каждого экстракта был создан образец контроля качества.

Полученные липидные экстракты (без образца контроля качества) были разбиты на 3 партии для анализа:

- партия 1: 16 образцов из группы 1, 14 образцов из группы 2, 16 образцов из группы 3, 14 образцов из группы 4;
- партия 2: 10 образцов из группы 1, 20 образцов из группы 2, 10 образцов из группы 3, 20 образцов из группы 4;
- партия 3: 14 образцов из группы 1, 14 образцов из группы 2, 14 образцов из группы 3, 14 образцов из группы 4.

Разделение липидных экстрактов осуществляли на хроматографе Dionex UltiMate 3000 («Thermo Scientific», Германия) с использованием обратно-фазовой колонки Zorbax C18 (длина 150 мм, внутренний диаметр 2.1 мм, размер частиц 5 мкм, «Agilent», США) и следующих элюентов в качестве подвижной фазы: элюент А - ацетонитрил/вода (60/40, о/о) с добавлением 0,1% муравьиной кислоты и 10 мМ формиата аммония; элюент В - ацетонитрил/изопропанол/вода, (90/8/2, о/о/о), с добавлением 0,1% муравьиной кислоты и 10 мМ формиата аммония. Скорость потока 35 мкл/мин, температура колонки 50°С. Доля градиента В изменялась по заданному алгоритму: 0-0.5 мин – 30% В, до 20-ой минуты росла до 99% и сохраняла значение до 30-ой минуты и за полминуты возвращалось к значению 30%. МС анализ производили с использованием прибора Maxis Impact («Bruker», Германия) со следующими настройками: диапазон 100-1800 *m/z*, с напряжением на капилляре 4.1 кВ в режиме положительных ионов, давлением распыляющего газа 0.7 бар, скорости потока осушающего газа 6 л/мин и температурой 200°С.

Выполнение тандемного МС анализа осуществляли с использованием зависимого сканирования, в котором после снятия спектра снимали спектры фрагментации при энергии столкновения в 35 эВ соединений, давших пять самых интенсивных пиков в спектре, с окном изоляции 5 Да и временем исключения 2 мин.

Анализ образцов контроля качества производился через каждые 10 исследуемых образцов.

Данные, полученные в ходе анализа в виде .d файлов, преобразовывали в формат MzXml посредством программного обеспечения msConvert (Proteowizard, 3.0.9987) и предобрабатывали с использованием алгоритма, предоставленного Koelmel [7], программного обеспечения MzMine [8]. Идентификацию липидов осуществляли при помощи программы Lipid Match [7]. Номенклатура ионов использована согласно Lipid Maps терминологии в сокращённой форме записи [9].

Обработку данных осуществляли посредством расчета для каждой партии: значения средней величины интенсивности пика каждого соединения

$$< I_{p,b} >= \frac{\sum_{i=1}^{N_b} I_{p,i,b}}{N_b} \quad (1)$$

и стандартного отклонения интенсивности пика каждого соединения

$$sd\left(I_{p,b}\right) = \sqrt{\frac{\sum_{i=1}^{N_b}\left(I_{p,i,b}-<I_{p,b}>\right)^2}{N_b-1}} \quad (2),$$

где N_b – число образцов в партии b, $I_{p,i,b}$ – интенсивность пика p в образце i партии b; для всего набора данных рассчитывали значения средней величины интенсивности пика каждого соединения

$$< I_p >= \frac{\sum_{i=1}^{N} I_{p,i}}{N} \quad (3)$$

и стандартного отклонения интенсивности пика каждого соединения

$$sd\left(I_p\right) = \sqrt{\frac{\sum_{i=1}^{N}\left(I_{p,i}-<I_p>\right)^2}{N-1}} \quad (4),$$

где N – общее число образцов $I_{p,i}$ – интенсивность пика *p* в образце *i*.На основе этих значений рассчитывали новые значения каждого пика в каждом образце по формуле

$$I^*_{p,s,b} = \frac{I_{p,s,b}-<I_{p,b}>}{sd(I_{p,b})} * sd\left(I_p\right)+< I_p > \quad (5),$$

т.е. осуществляли автомасштабирование полученных данных.

После выполнения нормировки набор данных был разбит на два: содержащий информацию об экстрактах здоровых тканей и содержащий информацию о липидных экстрактах опухолевых тканей. Для каждого набора данных было проделано вычисление значений проекций переменной с использованием метода ортогональных проекций на скрытые структуры [10,11]:

1. выполнено парето-масштабирование матрицы независимых переменных N*m **X**, где N – число образцов, m – число соединений

$$\mathbf{X}_i = \frac{\mathbf{X}_i-\langle\mathbf{X}_i\rangle}{\sqrt{sd(\mathbf{X}_i)}} \quad (6),$$

где $\mathbf{X}_i$– i-ый столбец матрицы **X**, $< \mathbf{X}_i >$– среднее значение i-того столбца и $sd(\mathbf{X}_i)$ – стандартное отклонение переменных в i-ом столбце;

2. выполнено парето-масштабирование столбца зависимых переменных y высотой N, где 0 обозначается состояние «отсутствие метастазирование», 1 – состояние «болезнь»

$$y = \frac{y-\langle y\rangle}{\sqrt{sd(y)}} \quad (7),$$

где <y> – среднее значение переменных отклика, sd(y) – стандартное отклонение переменных отклика;

3. рассчитаны веса для независимых переменных

$$w = \frac{\mathbf{X}^T y}{y^T y} \quad (8);$$

4. выполнена нормализация рассчитанного вектора

$$w = \frac{w}{\|w\|} \quad (9);$$

5. рассчитаны предсказательные счета

$$t = \mathbf{X}w \quad (10);$$

6. рассчитана предсказательная нагрузка независимых переменных

$$p^T = \frac{t^T\mathbf{X}}{t^T t} \quad (11);$$

7. вычислен вектор ортогональных нагрузок

$$w_o = p - w^T p w \quad (12);$$

8. выполнена нормализация рассчитанного вектора ортогональных нагрузок

$$w_o = \frac{w_o}{\|w_o\|} \quad (13);$$

9. рассчитаны ортогональные счета
$$t_o = \mathbf{X}w_o \text{ (14);}$$
10. рассчитана ортогональная нагрузка независимых переменных
$$p_o{}^T = \frac{t_o{}^T\mathbf{X}}{t_o{}^T t_o} \text{ (15);}$$
11. вычислены данные, не содержащие ортогональной составляющей
$$\mathbf{X}_p = \mathbf{X} - t_o{}^T p_o \text{ (16);}$$
12. вычислены предсказательные счета от независимых переменных, не содержащих ортогональной составляющей
$$t_p = \mathbf{X}_p w \text{ (17);}$$
13. вычислена предсказательная нагрузка от независимых переменных, не содержащих ортогональной составляющей
$$p_p{}^T = \frac{t_p{}^T\mathbf{X}_p}{t_p{}^T t_p} \text{ (18);}$$
14. нормирован вектор, содержащий предсказательную нагрузку от независимых переменных без ортогональной составляющей
$$p_p = \frac{p_p}{\|p_p\|} \text{ (19);}$$
15. рассчитан вектор, содержащий значения проекций переменной
$$VIP = p_p\sqrt{m} \text{ (20).}$$

Полученные значения проекции переменной (ПП) были использованы в качества критерия выбора соединений-потенциальных маркеров, где нижней границей значения проекции переменной для маркера принималась 1.

Далее произвольно была выбирана переменная из набора переменных, сформированного на основе значений ПП. После этого следует этап расчёта информационного критерия Акаике (ИКА):

1. выполняем поиск коэффициентов для логистической регрессии, где независимой переменной является выбранная выше переменная, переменной отклика – наличие или отсутствие метастазирование (1 или 0 соответственно) методом максимизации функции правдоподобия
$$l = \sum_{i=1}^{M} y_i(\beta^T x_i) - ln\left(1 + e^{\beta^T x_i}\right) \text{ (21),}$$
где y_i – переменная отклика, принимающая значения 0 или 1, x_i – объединённый вектор единицы и независимых переменных, β – объединённый вектор свободного члена и коэффициентов при переменных, M – число задействованных переменных;с использованием вычисленных коэффициентов вычисляем лог-функцию правдоподобия и ИКА
$$AIC = 2l - 2 * M \text{ (22) [12].}$$
2. повторяем п.п. 1 – 3 для всех m переменных.
3. выбираем переменную, для которой рассчитанное значение AIC будет максимальным (обозначим это значение как AIC');выполняем п.п. 1 – 5 для комбинации «выбранная ранее переменная + каждая из оставшихся переменных»;
4. сравниваем AIC со значением AIC';если AIC из п. 7 больше AIC', повторяем п.п. 1-7, имея в качестве постоянных переменных переменные, отобранные ранее и обозначив как AIC' значение из п. 7; Если AIC из п. 7 меньше AIC', п. 9; переменные, при которых было получено AIC' и рассчитанные для них коэффициенты используем дальше.

Далее осуществляли проверку на статистически значимое неравенство коэффициентов при переменных нулю с удалением переменных, не удовлетворявших этому условию:

1. для отобранных ранее переменных вычисляли их коэффициенты в логистической регрессией на основе максимизации функции правдоподобия;
2. подставив вычисленные значения β в матрицу
$$\mathbf{D} = \sum_{i=1}^{n} \frac{x_i x_i{}^T e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2} \text{ (23),}$$
вычисляли значения стандартной ошибки для коэффициента β^j
$$\mathrm{SE}\left(\beta^j\right) = \sqrt{\mathbf{D}_{jj}} \text{ (24),}$$
где j – порядковый номер коэффициента в векторе β;вычисляли вероятность отличия от нуля коэффициента β^j
$$p_j = \chi\left(\frac{\beta^j}{\mathrm{SE}(\beta^j)}\right)^2 \text{ (25),}$$
где $\chi(\theta)^2$– распределение квадрата независимой стандартной нормальной случайной величины θ;если $p_j > 0.05$, то переменная, соответствующая p_j и не являющаяся 1, исключается из задействованного набора переменных и действия 1- 4 повторяются.

Качество построенных моделей проверяли с использованием скользящего контроля по отдельным объектам, в котором модель тренировалась на N-1 объекте и тестировалась на оставшемся объекте N раз.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Изменение распределения координат образцов в первых трёх координатах главных компонент в результате нормализации представлены на рисунках 1 и 2. После нормализации относительное отклонение значения полного ионного тока для образцов контроля качества снизилось с 7% до 4%.

В случае опухолевой ткани молочной железы для 54 из 317 идентифицированных соединений значение ПП составило больше 1. Эти соединения преимущественно относятся к классам триацилглицеридов (31), фосфатидилхолинов (14), сфингомиелинов (6) и диацилглицеридов (3).

В случае нормальной ткани молочной железы для 60 из 317 идентифицированных соединений значение ПП больше оказалось больше 1. Данные соединения относятся к классам лизо- и фосфатидилхолинов (20), триацилглицеридов (18), диагцилглицеридов (12), сфингомиелинов (6), фосфатидилэтаноламинов (4).

После выбора переменных по значению ИКА и удалению переменных со статистически незначимо отличающимися от ноля коэффициентами для диагностики

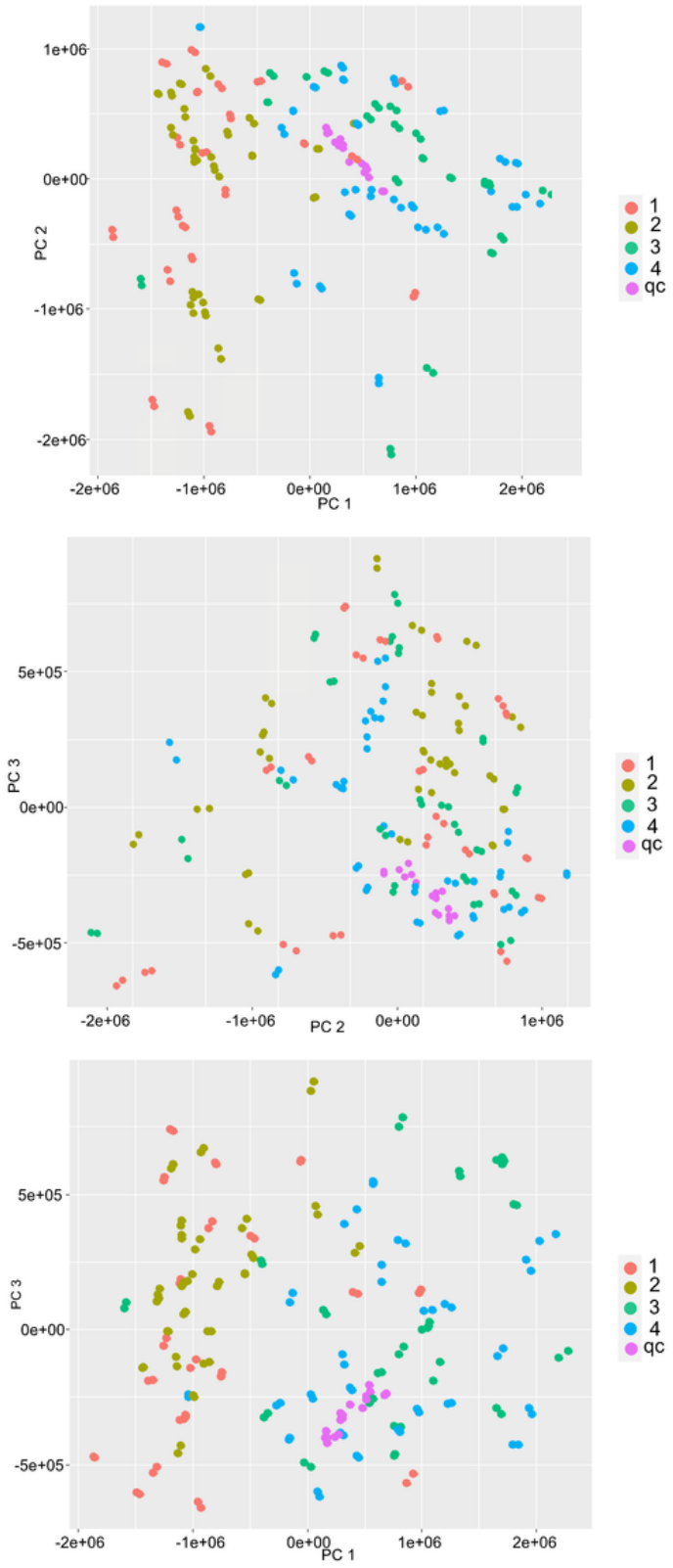

Рисунок 1. Распределение биологических образцов, не подвергавшиеся нормализации в пространстве трёх главных компонент. Группа 1 – нормальная ткань молочной железы от пациентов без метастазирования, группа 2 – ткань нормальной ткани молочной железыот пациентов с метастазированием, группа 3 – опухолевая ткань молочной железы от пациентов без метастазирования, группа 4 – опухолевая ткань молочной железы от пациентов с метастазированием, группа qc – группа образцов контроля качества.

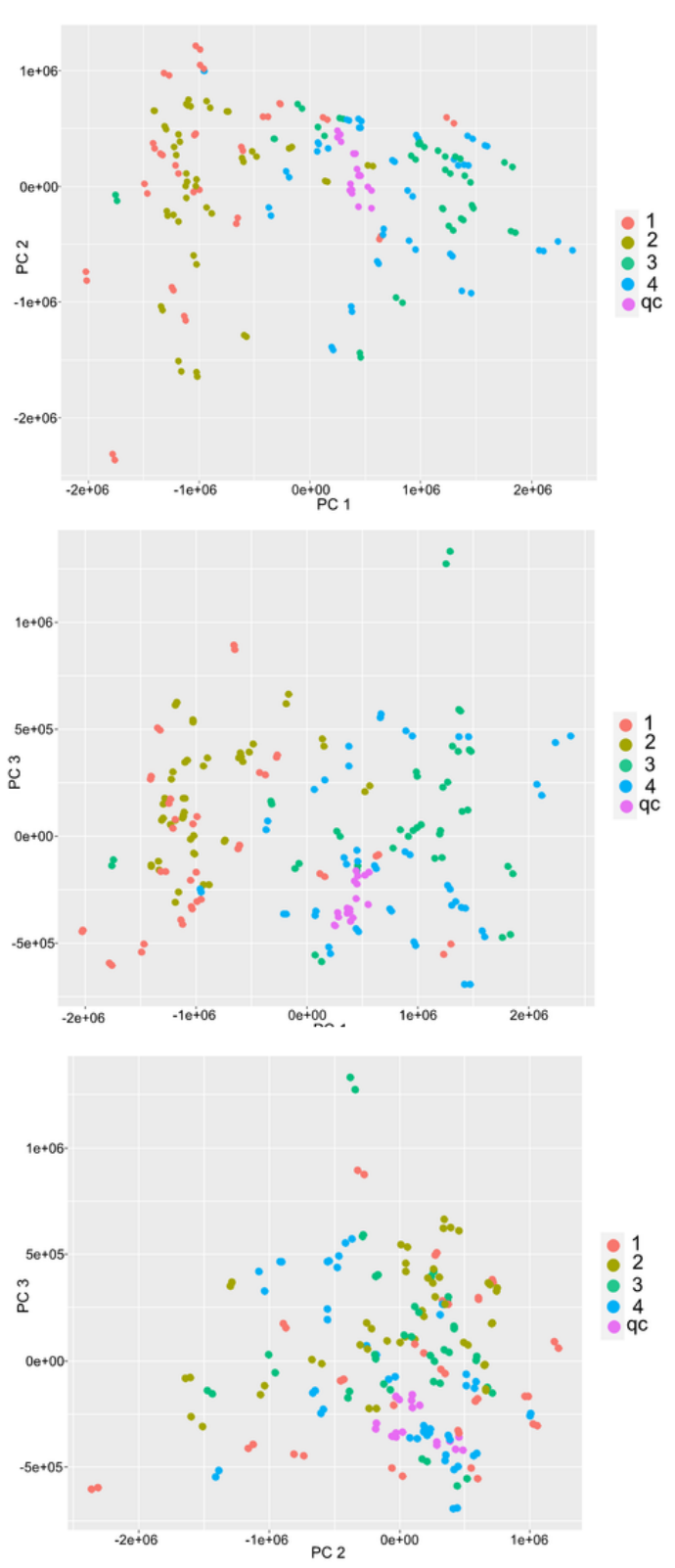

Рисунок 2. Распределение биологических образцов после автомасштабирования в пространстве трёх главных компонент. Группа 1 – нормальная ткань молочной железы от пациентов без метастазирования, группа 2 – ткань нормальной ткани молочной железы от пациентов с метастазированием, группа 3 – опухолевая ткань молочной железы от пациентов без метастазирования, группа 4 – опухолевая ткань молочной железы от пациентов с метастазированием, группа qc – группа образцов контроля качества.

Таблица 1. Коэффициенты при липидах-маркерах в опухолевых тканях β, их доверительный интервал (ДИ β), критерий Вальда и вероятность отличия коэффициента от нуля.

Липид	β	ДИ β	Критерий Вальда	*p*
Свободный член	0.3	-2.31 - 2.92	0.23	0.82
SM d18:2/16:0	-6.05E-05	-9.25E-05 - -3.73E-05	-4.38	<0.001
SM d18:1/16:0	9.87E-06	6.04E-06 - 1.53E-05	4.30	<0.001
SM d18:0/24:1	-5.82E-06	-1.21E-05 - -7.1E-07	-2.06	0.04
PC 16:0_18:2	-3.36E-06	-5.96E-06 - -1.30E-06	-2.90	0.004
TG 12:0_14:1_18:2	1.07E-05	2.01E-06 - 2.10E-05	2.24	0.02

Таблица 2. Коэффициенты при липидах-маркерах в здоровых тканях β, их доверительный интервал (ДИ β), критерий Вальда и вероятность отличия коэффициента от нуля.

Липид	β	ДИ β	Критерий Вальда	P
Свободный член	-2.46E+01	-4.53E01 - -1.13E01	-2.93	0.003
TG 10:0_18:1_18:3	3.11E-04	1.44E-04 - 5.70E-04	2.94	0.003
DG 18:0_18:1	-7.61E-05	-1.46E-04 - -2.79E-05	-2.59	0.01
SM d18:1/18:0	4.74E-05	6.05E-06 - 9.89E-05	2.07	0.04
LPC 16:0	-1.16E-04	-2.13E-04 - -5.37E-05	-2.93	0.003
TG 12:0_18:1_8:0	-1.28E-05	-2.44E-05 - -5.71E-06	-2.72	0.007
TG 10:0_18:2_18:2	6.93E-05	3.34E-05 - 1.27E-04	2.96	0.003
OxTG 18:1_18:2_18:3(OH)	-9.70E-05	-1.90E-04 - -3.77E-05	-2.51	0.01
PC P-16:0/20:4	9.30E-05	4.37E-05 - 1.73E-04	2.87	0.004
PC 12:0_14:1	-1.65E-05	-3.20E-05 - -5.74E-06	-2.52	0.01
DG 18:2_18:2	-1.67E-04	-3.38E-04 - -4.36E-05	-2.30	0.02

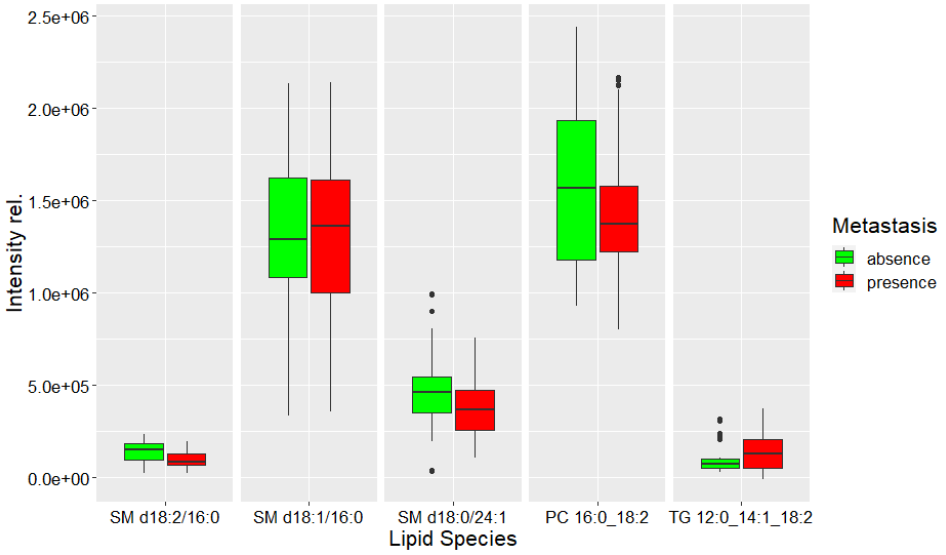

Рисунок 3. Диаграмма размахов относительных уровней липидов, являющихся маркерами регионарного метастазирования в опухолевых тканях при раке молочной железы. Зелёным обозначено отсутствие метастазирования, красным – наличие.

наличия регионарного метастазирования по опухолевой ткани получаем набор липидов {SM 18:2/16:0, SM 18:1/16:0, SM 18:0/24:1, PC 16:0_18:2, TG 12:0_14:1_18:2} (Таблица 1, рисунок 3). Тестирование модели дало значение площади под операционной кривой 0.81, чувствительность и специфичность 94% и 65% при пороге 0.39.

После выбора переменных по значению ИКА и удалению переменных со статистически незначимо отличающимися от ноля коэффициентами для диагностики наличия регионарного метастазирования по биопсии ткани молочной железы получаем набор липидов {TG 10:0_18:1_18:3, DG 18:0_18:1, SM d18:1/18:0, LPC 16:0, TG 12:0_18:1_8:0, TG 10:0_18:2_18:2, OxTG 18:1_18:2_18:3(OH), PC P-16:0/20:4, PC 12:0_14:1, DG 18:2_18:2} (табл. 2, рис. 4). Тестирование модели дало значение площади под операционной кривой 0.79, чувствительность и специфичность 88% и 58% при пороге 0.15.

Полученные модели характеризуются высокой чувствительностью. В модели, построенной для диагностики метастазирования по биопсии опухолевой

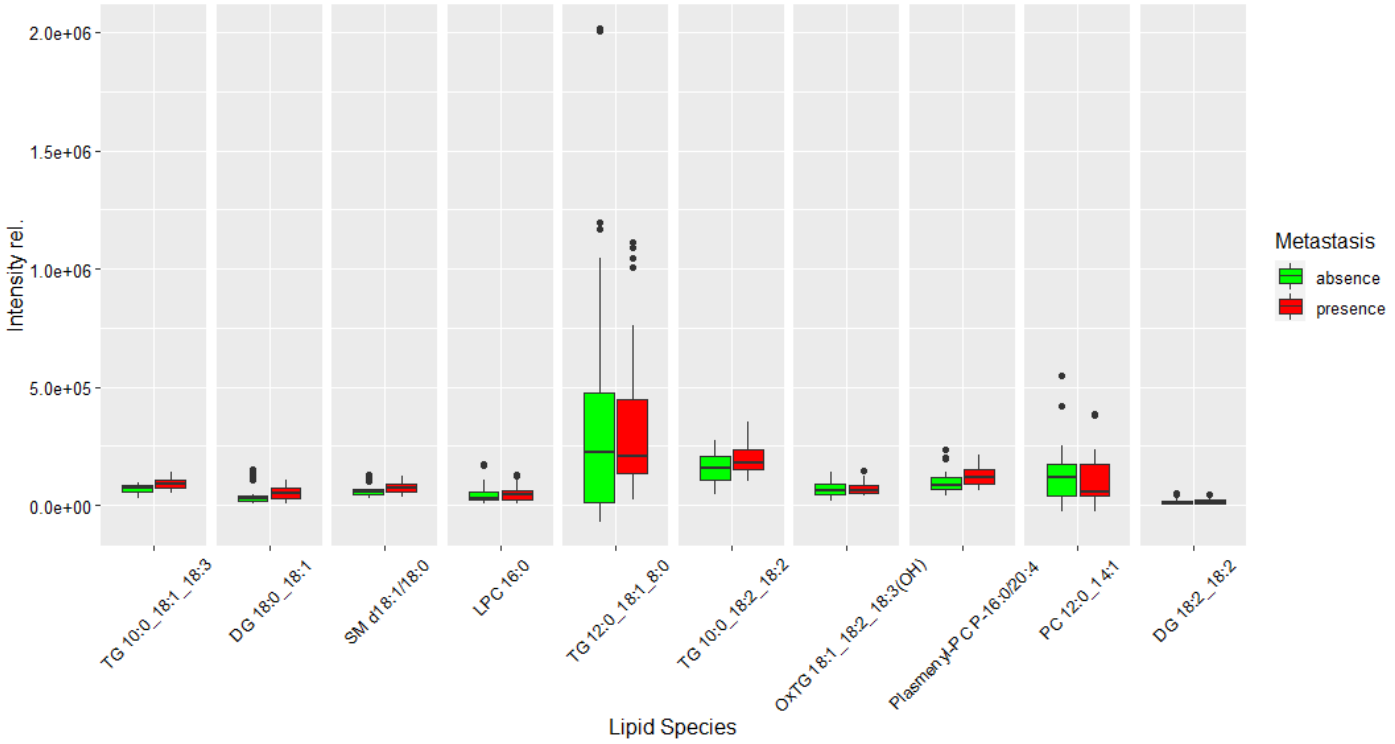

Рисунок 4. Диаграмма размахов относительных уровней липидов, являющихся маркерами регионарного метастазирования в тканях молочной железы при раке молочной железы. Зелёным обозначено отсутствие метастазирования, красным – наличие.

ткани, в качестве маркеров преобладают сфигномиелины (3 из 5), в модели, построенной для диагностики по биопсии ткани молочной железы, преобладают триглицериды (4 из 10) и фосфатидилхолины (3 из 10). Уже показано, что сфингомиелины являются маркерами метастатических процессов [13,14]. Триглицериды являются основным источником энергии в клетках [15]. Фосфатидилхолины связаны с пролиферативными процессами [16].

ЗАКЛЮЧЕНИЕ

Предложенный метод дал возможность построить диагностическую панель липидов для моделей высокой чувствительности. Наличие в моделях липидов, для чьих классов определена связь с метастатическими и опухолеобразующими процессами, говорит о релевантности предложенного метода обработки данных.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Все клинические исследования проводили в соответствии с принципами, изложенными в Хельсинкской декларации. Все пациенты прочитали и подписали информированное согласие, одобренное этическим комитетом Национального медицинского исследовательского центра акушерства, гинекологии и перинатологии имени ак. В.И. Кулакова (протокол №9 от 22.11.2018).

ФИНАНСИРОВАНИЕ

Работа выполнена при поддержке Мегагранта Министерства науки и высшего образования Российской Федерации (Соглашение со Сколковским институтом науки и технологий, № 075-10-2019-083 от 11 декабря 2019 г.)

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Podo, F., Canevari, S., Canese, R., Pisanu, M.E., Ricci, A., Iorio, E.* (2011) Tumour Phospholipid Metabolism. Exp. Oncol., **19**, 1–10.
2. *Wong, M.W., Braidy, N., Poljak, A., Pickford, R., Thambisetty, M., Sachdev, P.S.* (2017) Dysregulation of lipids in Alzheimer's disease and their role as potential biomarkers. Alzheimer's Dement., **13**(7), 810–27, DOI: 10.1016/j.jalz.2017.01.008
3. *Liu, X., Li, J., Zheng, P., Zhao, X., Zhou, C., Hu, C., Hou, X., Wang, H., Xie, P., Xu, G.* (2016) Plasma lipidomics reveals potential lipid markers of major depressive disorder. Anal. Bioanal. Chem., **408**(23), 6497–507, DOI: 10.1007/s00216-016-9768-5
4. *Anand, S., Barnes, J.M., Young, S.A., Garcia, D.M., Tolley, H.D., Kauwe, J.S.K., Graves, S.W.* (2017) Discovery and Confirmation of Diagnostic Serum Lipid Biomarkers for Alzheimer's Disease Using Direct Infusion Mass Spectrometry. J. Alzheimer's Dis., **59**(1), 277–90, DOI: 10.3233/JAD-170035
5. *Hogan, S.R., Phan, J.H., Alvarado-Velez, M., Wang, M.D., Bellamkonda, R. V., Fernández, F.M., Laplaca, M.C.* (2018) Discovery of Lipidome Alterations Following Traumatic Brain Injury via High-Resolution Metabolomics. J. Proteome Res., **17**(6), 2131–43, DOI: 10.1021/acs.jproteome.8b00068
6. Folch, J., Lees, M., Sloane Stanley, G.H. (1957) A simple method for the isolation and purification of total lipides from animal tissues. J. Biol. Chem., 226(1), 497–509.
7. *Koelmel, J.P., Kroeger, N.M., Ulmer, C.Z., Bowden, J.A., Patterson, R.E., Cochran, J.A., Beecher, C.W.W., Garrett, T.J., Yost, R.A.* (2017) LipidMatch: An automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. BMC Bioinformatics, **18**(1), 1–11, DOI: 10.1186/s12859-017-1744-3
8. *Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M.* (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics, **11**, 395. DOI: 10.1186/1471-2105-11-395
9. *Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Subramaniam, S.* (2007) LMSD: LIPID MAPS structure database. Nucleic Acids Res., **35**(SUPPL. 1), 527–32, DOI: 10.1093/nar/gkl838
10. *Wold, S., Sjöström, M., Eriksson, L.* (2001) PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**(2), 109–30, DOI: 10.1016/S0169-7439(01)00155-1

11. Galindo-Prieto, B., Eriksson, L., Trygg, J. (2015) Variable influence on projection (VIP) for OPLS models and its applicability in multivariate time series analysis. Chemom. Intell. Lab. Syst. 146, 297–304, DOI: 10.1016/j.chemolab.2015.05.001.
12. *Akaike, H.* (1998) Information Theory and an Extension of the Maximum Likelihood Principle. Sel. Pap. Hirotugu Akaike, 199–213, DOI: 10.1007/978-1-4612-1694-0_15
13. *Roy, J., Dibaeinia, P., Fan, T.M., Sinha, S., Das, A.* (2019) Global analysis of osteosarcoma lipidomes reveal altered lipid profiles in metastatic versus nonmetastatic cells. J. Lipid Res., **60**(2), 375–87, DOI: 10.1194/jlr.M088559
14. *Peng, W., Tan, S., Xu, Y., Wang, L., Qiu, D., Cheng, C., Lin, Y., Liu, C., Li, Z., Li, Y., Zhao, Y., Li, Q.* (2018) LC-MS/MS metabolome analysis detects the changes in the lipid metabolic profiles of dMMR and pMMR cells. Oncol. Rep., **40**(2), 1026–34, DOI: 10.3892/or.2018.6510
15. *Garrett, R.H., Grisham, C.* (2016) Biochemistry. 6th ed., CENGAGE Learning.
16. *Fagone, P., Jackowski, S.* (2013) Phosphatidylcholine and the CDP-choline cycle. Biochim. Biophys. Acta - Mol. Cell Biol. Lipids, **1831**(3), 523–32, DOI: 10.1016/j.bbalip.2012.09.009

Поступила: 02.09.2021
После доработки: 15.09.2021
Принята к публикации: 21.09.2021

PIPELINE OF MASS-SPECTROMETRY DATA PROCESSING FOR DIAGNOSTIC MOLECULAR MARKER PANEL OBTAINING USING THE EXAMPLE OF SEARCH MARKERS OF BREAST CANCER METASTASIS

A.O. Tokareva[1,2], V.V. Chagovets[3], A.S. Kononikhin[1,3], N.L. Starodubtseva[2,3], V.E. Frankevich[3], E.N. Nikolaev[1*]

[1]Skolkovo Institute of Science and Technology,
30 bld. 1 Bolshoy Boulevard,Moscow, 121205, Russia; *e-mail: e.nikolaev@skoltech.ru.
[2]V.L. Talrose Institute for Energy Problems of Chemical Physics, N.N. Semenov Federal Center of Chemical Physic, Russian Academy of Sciences, 38 bld. 2 Leninsky avenue, Moscow, 119334 Russia
[3]Academician V.I. Kulakov National Medical Research Center for Obstetrics, Gynecology and Perinatology, 4 bld. 2 Oparina str., Moscow, 117513 Russia

A pathology diagnostic using molecular marker is a perspective direction of clinical medicine. Mass-spectrometry (MS) is a one of methods, which are used for obtaining information about molecular profiles. Selection of species, essential for classification “case/control is an important task for data processing. Pipeline of data processing has been proposed using MS data, obtained during analysis of tumor breast tissue samples and health breast tissue samples, with the aim of metastasis marker selection. As a result, selection of lipid markers that belong to classes, related to metastasis and proliferation processes, makes it possible to create high sensitivity diagnostic model, based on logistic regression. The proposed method is applicable for data processing, obtained by MS analysis of other “omics”: metabolome, proteome, glycome.

Key words: mass-spectrometry; data processing; biological markers

FUNDING

This work was performed within the framework of the Megagrant of Ministry of Science and Higher Education of the Russian Federation (Agreement with Skolkovo Institute of Science and Technology, № 075-10-2019-083, December 11, 2019**)**

Received: 02.09.2021, revised: 15.09.2021, accepted: 21.09.2021