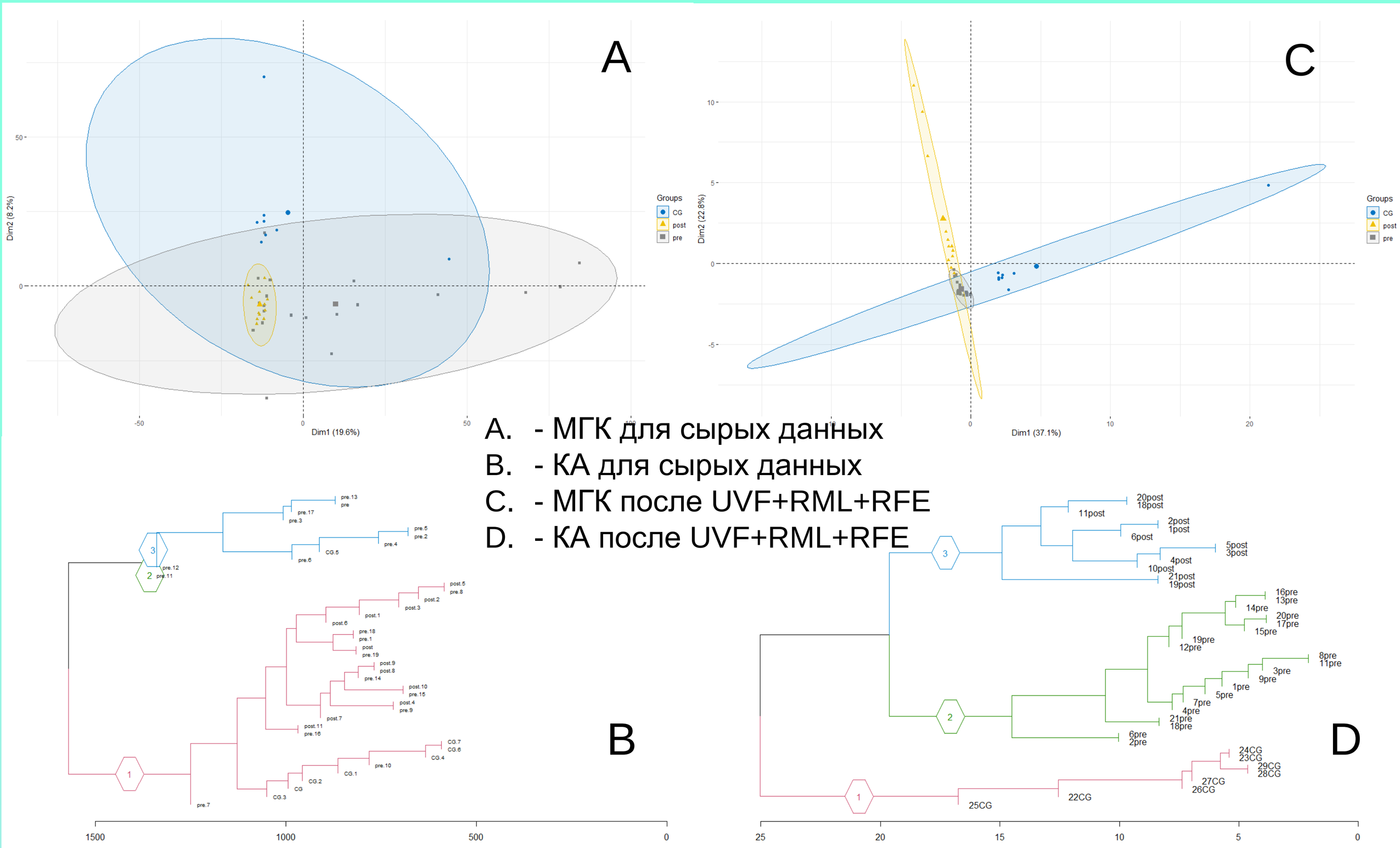
Хромато-масс-спектрометрическое ненаправленное профилирование и интеллектуальный анализ данных для выявления потенциальных биомеркеров

С1-47

Плющенко И.В., Родин И.А.

Московский Государственный Университет имени М.В. Ломоносова, Химический факультет

plyush1993@bk.ru



A. - МГК для сырых данных
B. - КА для сырых данных
C. - МГК после UVF+RML+RFE
D. - КА после UVF+RML+RFE

- Мультиядерные вычисления.

пакеты (“doParallel” , “parallel”). *статистические расчеты проводили на языке R.

- Загрузка таблицы пиков после интегрирования и выравнивания (iMet-Q).
- Дисперсионный анализ (UVF).

Последовательное применение: Shapiro-Wilk, Wilcox, Bartlett, Student тестов (Benjamini-Hochberg поправка на множественные сравнения).

пакет (“stats”)

- Повторяющиеся машинное обучение (RML)

4 модели (случайный лес, машина опорных векторов с радиальной ядерной функцией, k-ближайших соседей, проекция на скрытые структуры) с длиной тюнинга параметров 5 и троекратной 5 кросс-валидацией (повторение 10 раз).

пакет (“caret”).

- Рекурсивный отбор переменных (RFE)

Для моделей с точностью ≥ 80% генерировался набор уникальных переменных (начальное число признаков: 0.5*√общее число(ОЧ), √ОЧ, 2* √ОЧ, 3* √ОЧ, 5* √ОЧ) совместно со 100% частотой и средним рангом важности предикторов из 10 повторов каждой выбранной модели.

Отбор лучшего набора наивным байесовским классификатором (минимальное число предикторов при максимальной точности классификации).

пакеты (“caret”, “Deducer”)

- Обучение «без учителя»

(для кластерного анализа (КА): метрика = "canberra", объединение = "average") и метод главных компонент (МГК) (с масштабированием).

Для визуализации и проверки результатов.

пакеты (“stats”, “FactoMineR”, “dendextend”)

Оборудование	ЖХ	Полярность	Образцы	№ образцов	№ групп	Число переменных		
						Сырые данные	после UVF	после UVF+RML+RFE
LC-IT-TOF	ОФ С18	POS	моча	40	3	3441	1887	31