Санкт-Петербургский
Государственный
Политехнический
Университет

РАЗРАБОТКА МЕТОДА ПРЕДСКАЗАНИЯ ПРОЦЕНТНОГО СОДЕРЖАНИЯ ТИПОВ КЛЕТОК В ОБРАЗЦАХ КРОВИ ЧЕЛОВЕКА

Выполнила студентка гр. 63601-4: Иголкина А.А.
Руководитель: Самсонова М.Г.

30.05.2014

Обоснование для проведения работы

- Экспрессия гена: Ген → мРНК → Белок

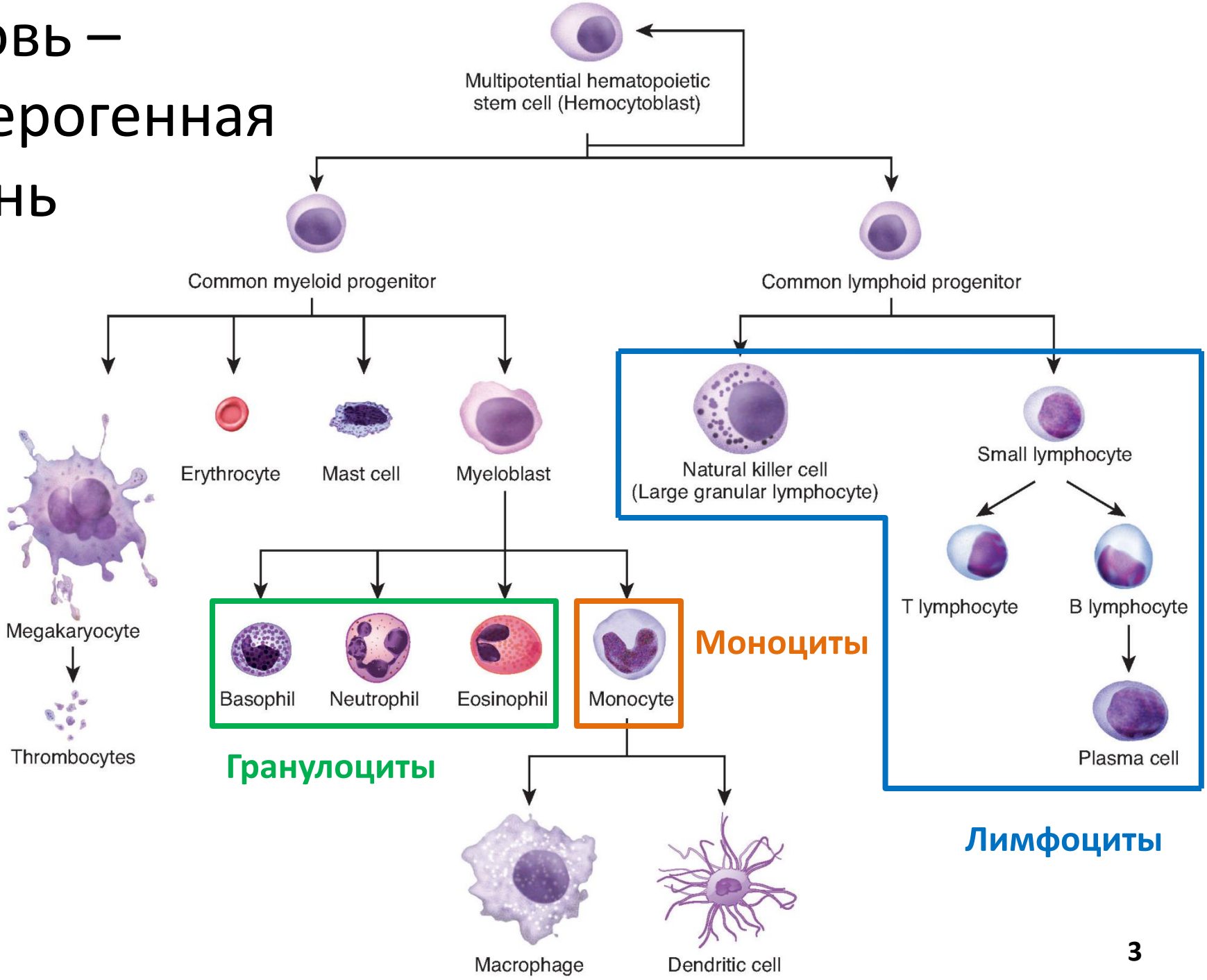
- eQTL (expression Quantitative Trait Loci) - генетические маркеры, влияющие на уровень экспрессии гена.

eQTL → ? Тип клетки →

- Как установить клеточную специфичность eQTL?
  Необходимо экспериментально определить:
  - Какие гены и как экспрессируются в клетках гетерогенной ткани
  - Процентное содержание отдельных типов клеток в гетерогенном образце.

Трудоемко
Дорогостояще
Иногда невозможно

Кровь – гетерогенная ткань



Исходные данные

Образцы крови
628 людей

Ген 1
Ген 2
Ген 3
….
…
Ген 31515

Уровни экспрессии генов

Процентные содержания типов клеток

Нейтрофилы
Базофилы
Эозинофилы
Моноциты
Лимфоциты

Данные получены в Медицинском Университете Гронингена
University Medical Centre Groningen (UMCG)

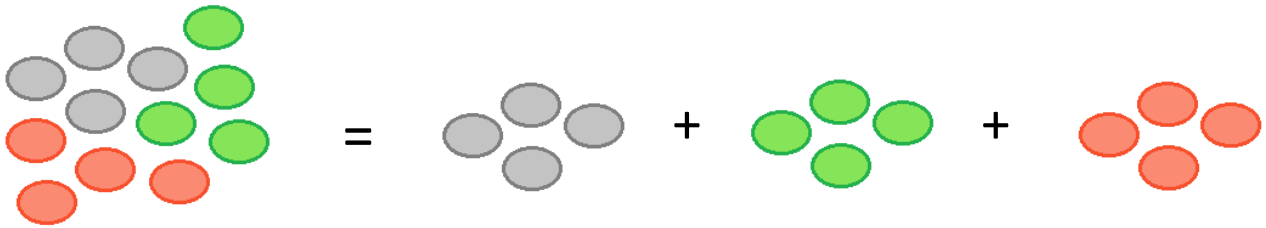
Цель

Предсказать процентное содержание отдельных типов клеток в образце гетерогенной ткани.

Задачи

- Построение модели, предсказывающей процентное содержание типов клеток в образце крови человека.
- Отбор групп генов для построения модели.

Стандартный метод деконволюции экспрессии генов



Метод требует дополнительных данных об экспрессии генов в отдельных типах клеток

$M_{i,j}$ - уровень экспрессии гена *j* в образце *i*.

$p_{i,k}$ - процентное содержание типов клеток *k* в образце *i*.

$S_{k,j}$ - уровень экспрессии гена *j* в типе клеток *k*.

$$M_{i,j} = \sum_{k=1}^{5} p_{i,k} \cdot S_{k,j}$$

$$p_{i,k} \geq 0$$

$$\sum_{k=1}^{5} p_{i,k} = 1$$

Нет надежных данных об экспрессии генов в отдельных типах клеток крови

Модель 1

- Предсказывает процентное содержание для каждого типа клеток в отдельности: фиксируем отдельный тип клеток.

- Идея: каждый ген аддитивно влияет на процентное содержание рассматриваемого типа клеток.

$M_{i,j}$ - уровень экспрессии гена *j* в образце *i*.

$p_{i,k}$ - процентное содержание типов клеток *k* в образце *i*.

$$p_{i,k} = \sum_{j=1}^{N_k} w_j \cdot M_{i,j}$$

- Метод оценки параметров регрессии - МНК (метод наименьших квадратов) с регуляризатором (гребневая регрессия):

$$\min \sum_{i=1}^{N} \left(p_{i,k} - \sum_{j=1}^{N_k} w_j \cdot M_{i,j} \right)^2 + \|w\|^2$$

Модель 2

- Предсказывает процентное содержание для каждого типа клеток в отдельности: фиксируем отдельный тип клеток.

$M_{i,j}$ - уровень экспрессии гена *j* в образце *i*.

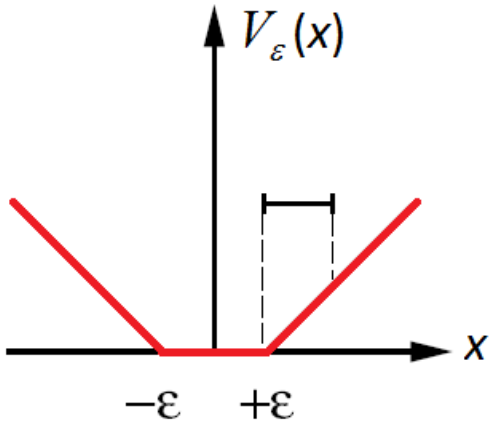
$$p_{i,k} = \sum_{j=1}^{N_k} w_j \cdot M_{i,j}$$

$p_{i,k}$ - процентное содержание типов клеток *k* в образце *i*.

- Оценка параметров – Метод машинного обучения: SVR (Support Vector Machine for Regression)

$$\min C \sum_{i=1}^{N} V_\varepsilon \left(p_{i,k} - \sum_{j=1}^{N_k} w_j \cdot M_{i,j} \right) + \frac{1}{2} \|w\|^2$$

$$w_j = \sum_{i=1}^{N} \alpha_i \cdot M_{i,j}, \alpha_i \in [0, C]$$



Модель 3

- Предсказывает процентное содержание всех типов клеток одновременно

- Определяет гипотетические уровни экспрессии генов в отдельных типах клеток.

$$M_{i,j} = \sum_{k=1}^{5} p_{i,k} \cdot S_{k,j}$$

Оценивается на тестовой выборке

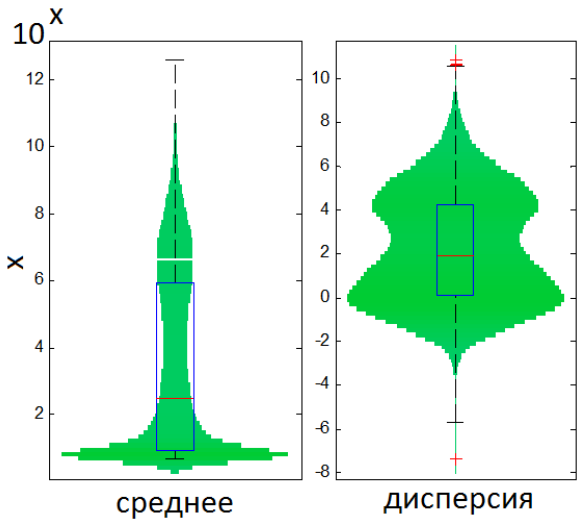
Оценивается на обучающей выборке

- Метод оценки параметров регрессии - МНК (метод наименьших квадратов) с регуляризатором (гребневая регрессия)

Процедуры отбора признаков (генов)

- Процедуры фильтрации: отсечение шума

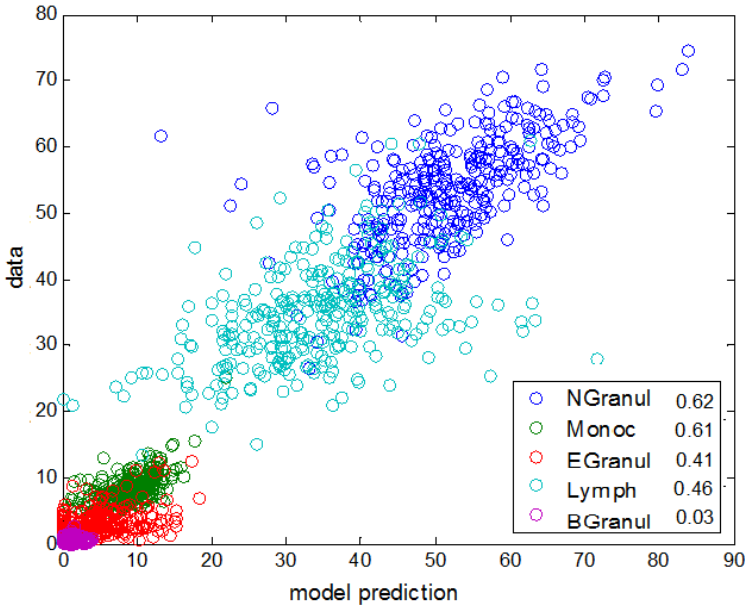
- Ранжирующие процедуры
  без использования модели:
  отбор наиболее информативных генов
  - Корреляция между уровнем экспрессии генов в образцах и процентным содержанием отдельных типов клеток
  - Метод главных компонент



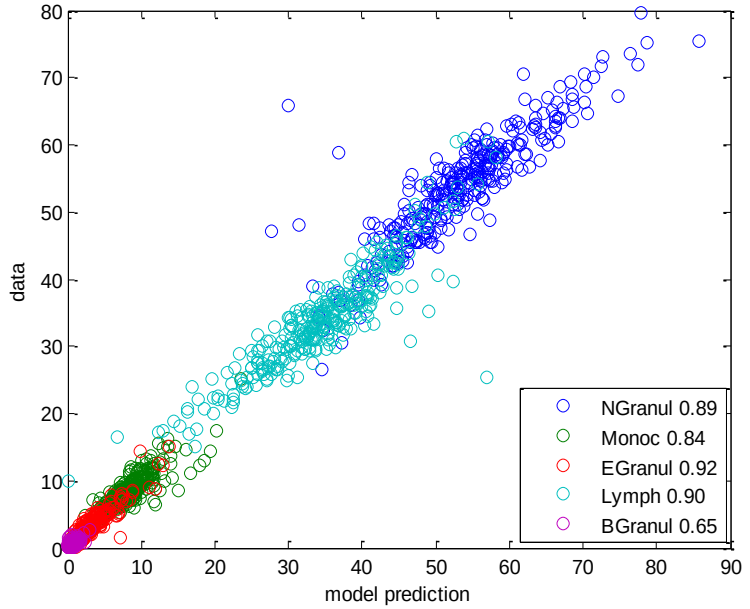
- Ранжирующие процедуры использующие модель: выбор наиболее предсказательных для модели признаков.
  - Лассо, эластичная сеть.
  - RFE (Recursive Feature Elimination)
  - Эвристический случайный выбор

Результаты

Модель 3: Гены из базы данных HaemAtlas



Модель 3: Эвр. Случайный выбор



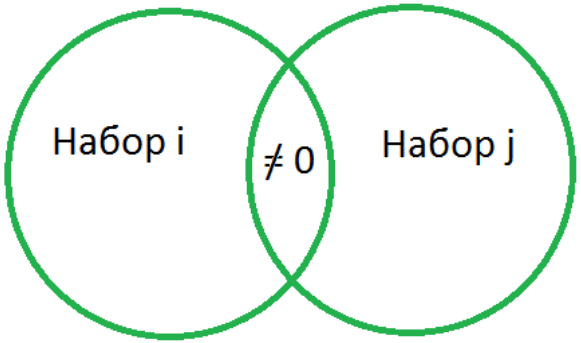
MSE – Mean Squared Error
CC – Коэффициент корреляции между предсказанием модели и данными

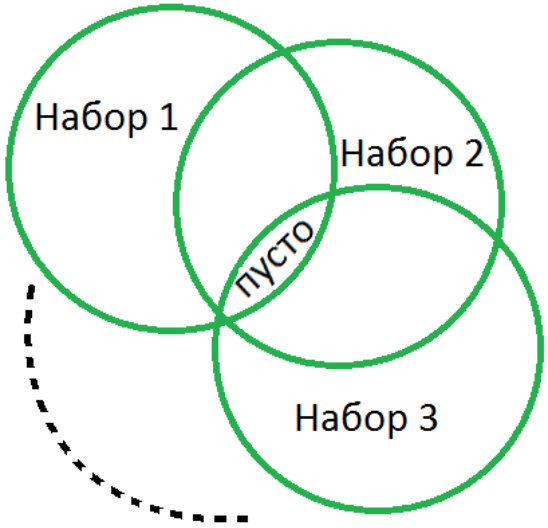
Предсказания для Нейтрофилов	MSE	CC
Модель 1: Лассо	54.89	0.71
Модель 1: Эластичная сеть	54.65	0.71
Модель 1: Лассо + RFE	31.45	0.76
Модель 2: Эвр. Случайный выбор	20.11	0.81
Модель 3: Эвр. Случайный выбор	18.85	0.89

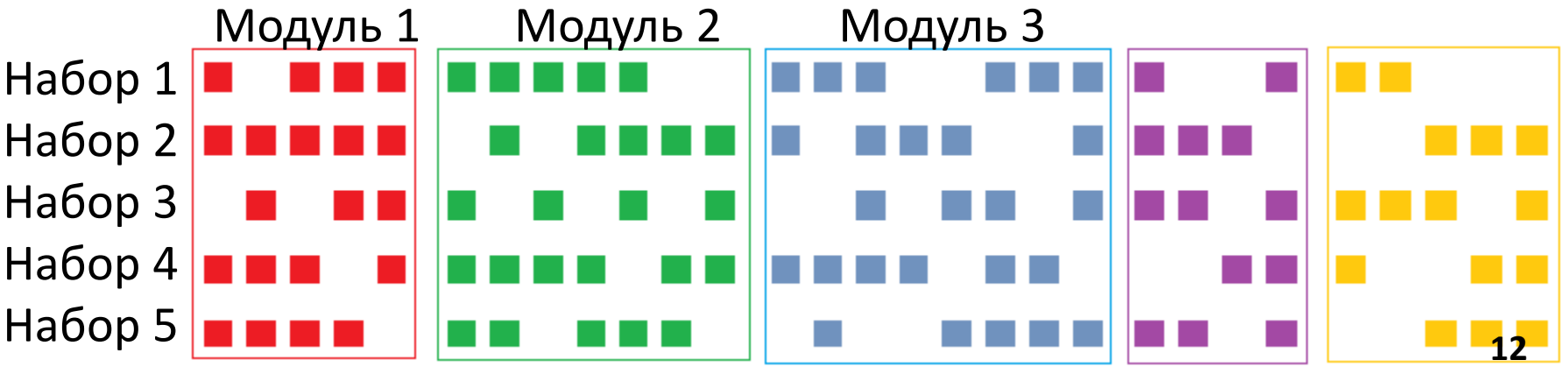
- Модель 2 (нелинейный случай) предсказывает лучше, чем Модель 1
- Модель 3 имеет наилучшую предсказательную способность
- Использование двух методов отбора лучше, чем одного.

Наборы генов, которые были отобраны для Модели 3

Мы получили 21 набор генов для Модели 3: CC=0.85-0.95







Заключение

- Мы разработали три модели для предсказания процентного содержания типов клеток в образцах крови человека. Все они показали высокую предсказательную способность.

- Наилучшую предсказательную способность показала Модель 3: CC = 0.85-0.95.

- Сформирован 21 набор генов для Модели 3. Каждая пара наборов имеет ненулевое пересечение. Пересечение всех наборов – пустое множество.

- Модель 3 может быть применена для предсказания процентных содержаний типов клеток в других гетерогенных тканях. Её предсказания будут использованы для атрибутирования eQTL типам клеток крови.

Публикации

• A.A.Igolkina, M.G.Samsonova, *Method to predict cell type percentage in human blood,* сборник докладов конференции BGRS\SB-2014.

• A.A.Igolkina, M.G.Samsonova, *Method to predict cell type percentage in human blood,* сборник докладов летней школы SBB-2014, работа заняла 2 место в конкурсе лучших устных докладов молодых ученых.

Благодарности

Самсонова М.Г.

Кадырова Н.О.

Березин С.

Козлов К.Н.

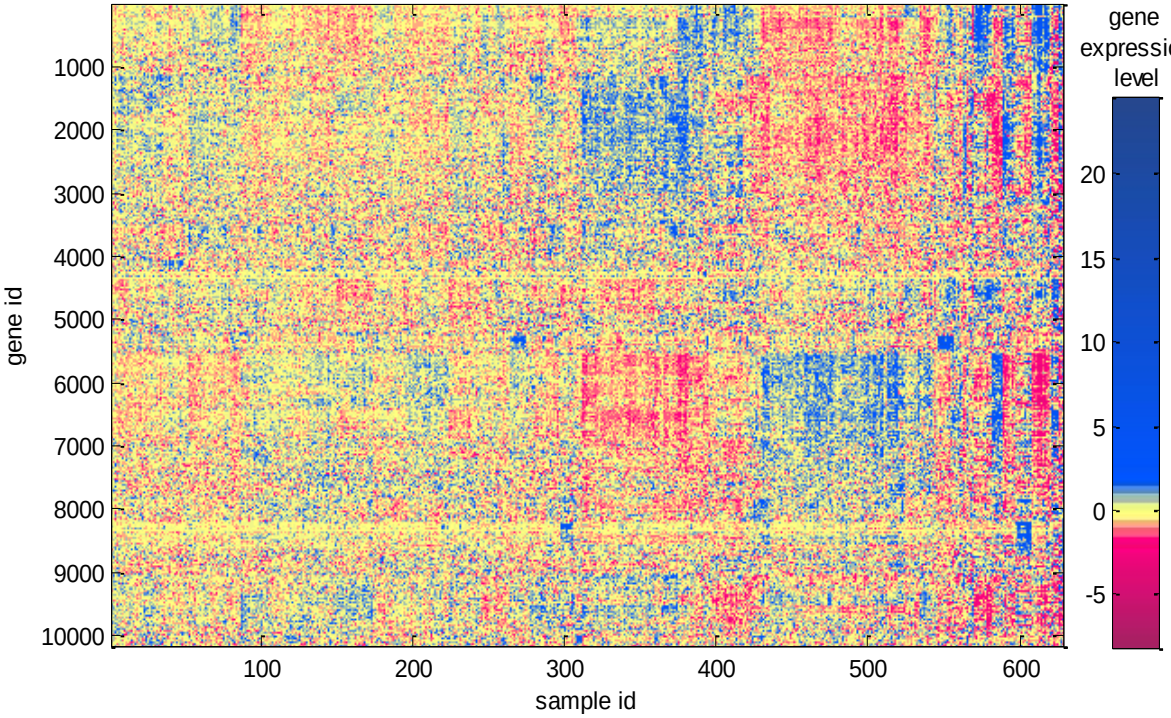
Писарев А.С.

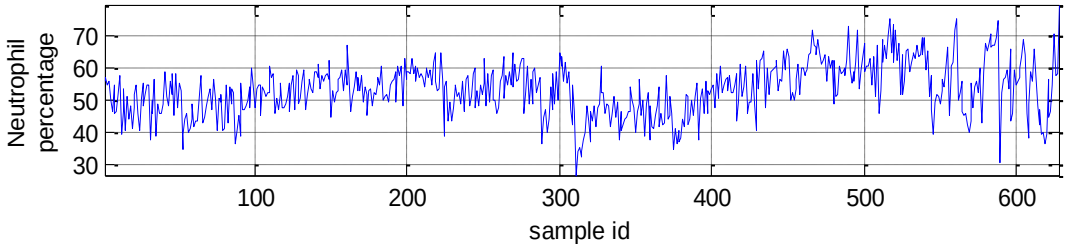
Lude Franke

Yang Li

Жернакова А.П.

Жернакова Д.





$$p_{i,k} = \sum_{j=1}^{N_k} w_j \cdot M_{i,j} = \langle w, M_i \rangle$$

$$w_j = \sum_{i=1}^{N} \alpha_i \cdot M_{i,j}, \alpha_i \in [0, C]$$

$$p_{i,k} = \sum_{i^*=1}^{N} \alpha_{i^*} \cdot \langle M_{i^*}, M_i \rangle$$

$$p_{i,k} = \sum_{i^*=1}^{N} \alpha_{i^*} \cdot k(M_{i^*}, M_i)$$