Выявление биомаркеров с гендерно-специфичной вариабельностью на основе анализа уровня метилирования ДНК

Выполнил: студент группы 381606-2
Солуянов Алексей Александрович
Научный руководитель: Заведующий кафедрой прикладной математики, д. ф.-м.н.
Иванченко Михаил Васильевич

Введение

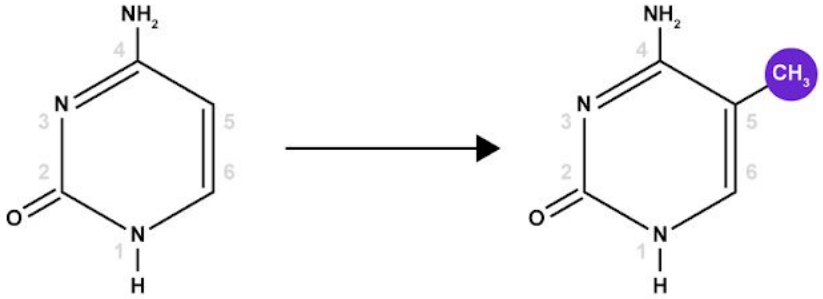
- Ускоренное старение организма.
- Развитие опасных заболеваний (гастроинтестинальный рак, рак легких, груди, карцинома и др.).
- Риск возникновения и развития многих заболеваний различен у мужчин и женщин, поэтому пол является важным аспектом любого исследования этиологии заболеваний.
- Различия в продолжительности жизни и ее качестве у мужчин и женщин.

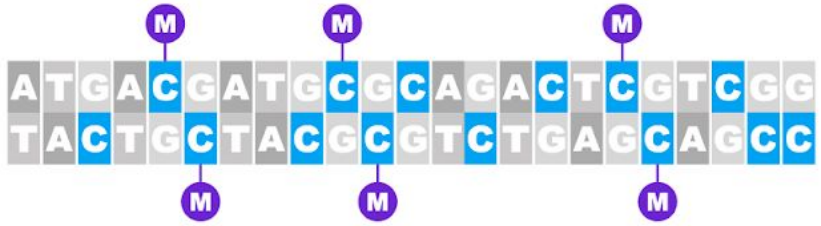
[1] Horvath, S. Genome Biol. 14, R115 (2013).

Метилирование ДНК

Модификация молекулы ДНК без изменения самой нуклеотидной последовательности ДНК.

Метилирование ДНК заключается в присоединении метильной группы (CH3) к цитозину в составе CpG сайта в позиции C5 цитозинового кольца.





CpG-сайт – область ДНК, где цитозин предшествует гуанину. Результат присоединения метильной группы к некоторому CpG-сайту является одной из основных метрик анализа метилирования ДНК.

https://en.wikipedia.org/wiki/CpG_site

Входные данные

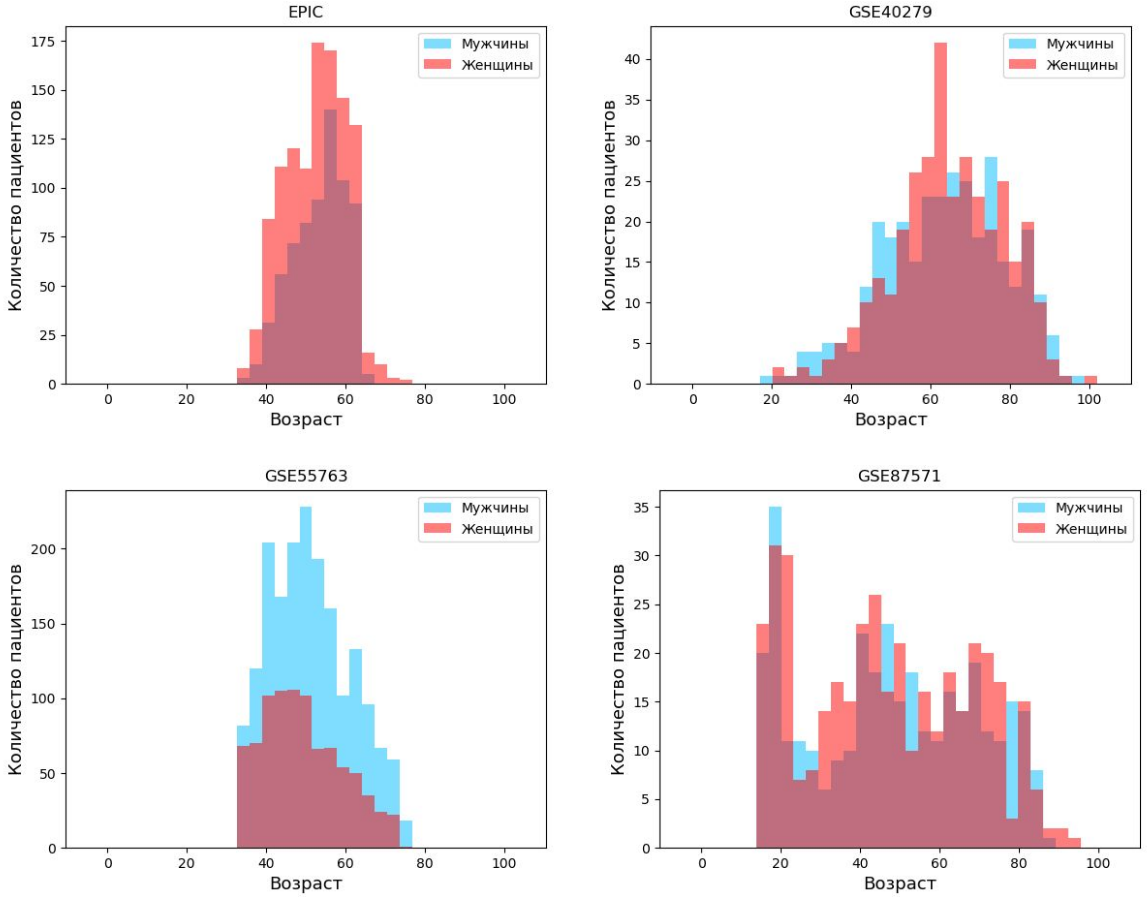
- Открытые базы данных, включающие наибольшее количество выборок: GSE87571[2], GSE40279[3] и GSE55763[4], EPIC[5], содержащие клиническую информацию о пациентах, информацию об уровне метилирования ДНК в более чем 400 000 CpG-сайтах.

CpG	1	2	…	n
1				
2				
…				
m				

Пусть у пациента №1 было проведено исследование в 1000 клетках. Метильная группа была присоединена к CpG №1 в 400 из них. Тогда уровень метилирования данного CpG-сайта у пациента №1 равен

$$\frac{400}{1000} = 0.4$$

- Таблица с основными характеристиками CpG-сайта (принадлежность к гену, координаты, уникальный идентификатор и др.)

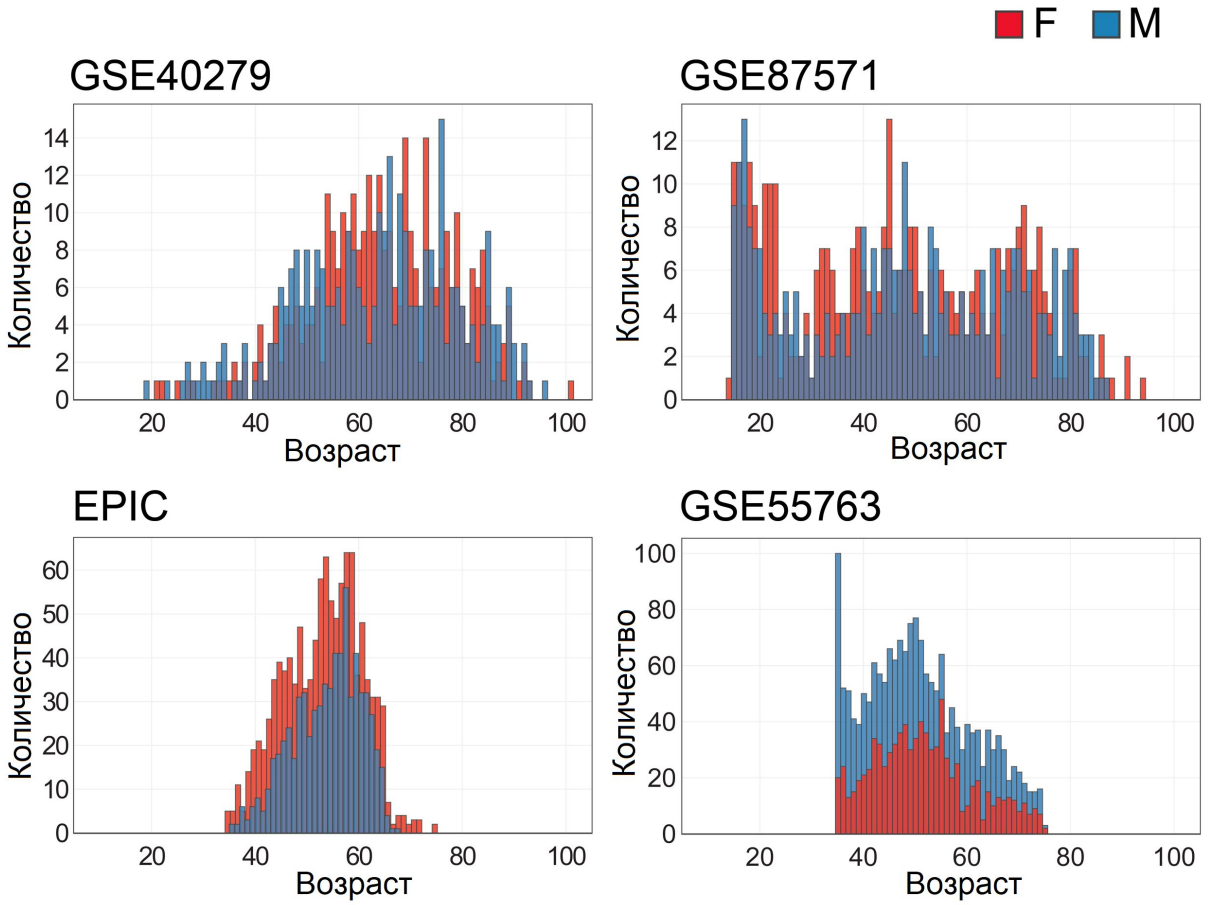


[2] A. Johansson et al, PloS One. 8 (2013) e67378
[3] G. Hannum et al, Mol. Cell. 49 (2013) 359–367.
[4] B. Lehne et al, Genome Biol. 16 (2015) 37.
[5] D. Palli et al, Tumori. 89 (2003) 586–593.

Метилирование ДНК

- Метилирование ДНК — добавление метильной группы (CH_3) к определенным участкам ДНК
- CpG-сайт – область ДНК, где гуанин (G) следует за цитозином (C)

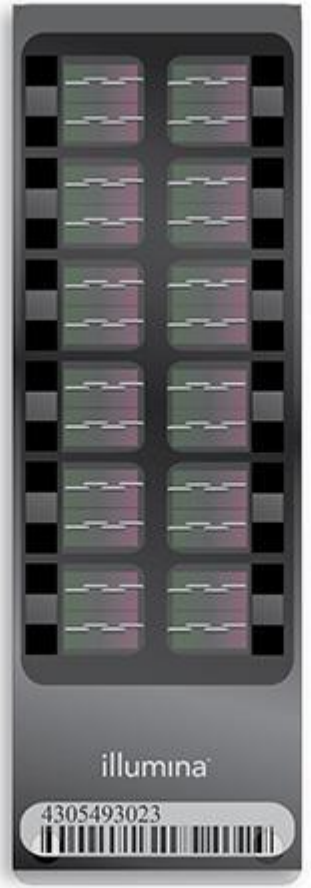
Пациент / CpG	1	2	…	m
1				
2				
…				
n				



База данных	Количество CpG-сайтов	Количество пациентов	Минимальный возраст	Максимальный возраст
GSE40279	473034	656	19	101
GSE87571	485512	729	14	94
EPIC	401506	1803	34	74
GSE55763	421906	2711	23	75

Illumina 450k

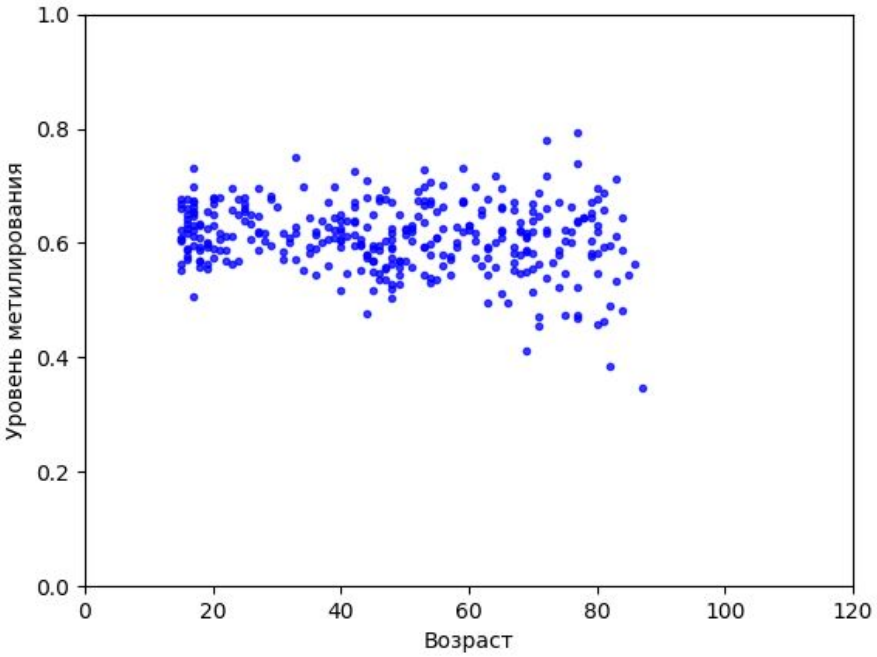
- В работе исследуются профили метилирования ДНК для клеток крови.
- Профили получены при помощи технологии Illumina Infinium HumanMethylation450 BeadChip (Illumina 450k).
- Illumina 450k измеряет уровень метилирования 485577 CpG проб.



Задача

- Разработать алгоритм анализа данных метилирования и определения биомаркеров с гендерно-специфичной вариабельностью.
- Разработать программный комплекс, реализующий данный алгоритм.

Вариабельность тесно связана с эпигенетическими мутациями, поэтому ее рост может сигнализировать о неестественных процессах в организме, которые могут быть вызваны развитием опасных заболеваний, таких как рак.



Линейная регрессия

- $$y = b_0 + b_1 x,$$

где b_i - параметры регрессии, x - независимая переменная (фактор модели), y – зависимая переменная.

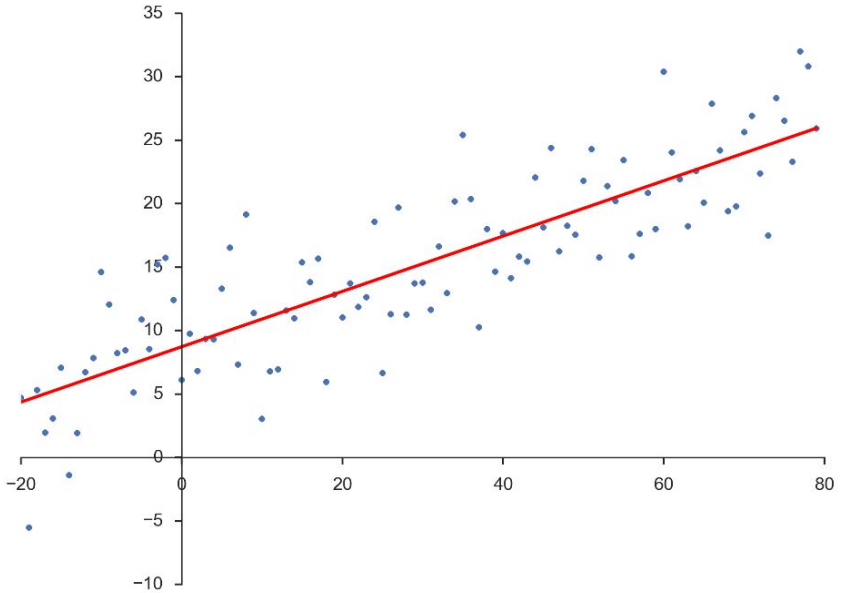
Задача - поиск прямой, являющейся лучшей аппроксимацией между переменными x и y.

Главная метрика метода – коэффициент детерминации:

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2},$$

где y_i – действительные значения y в каждом наблюдении, $\hat{y}_i$ – значения, предсказанные моделью, $\bar{y}$ – среднее по всем действительным значениям y_i.

Коэффициент детерминации показывает, насколько условная дисперсия модели отличается от дисперсии реальных значений y.

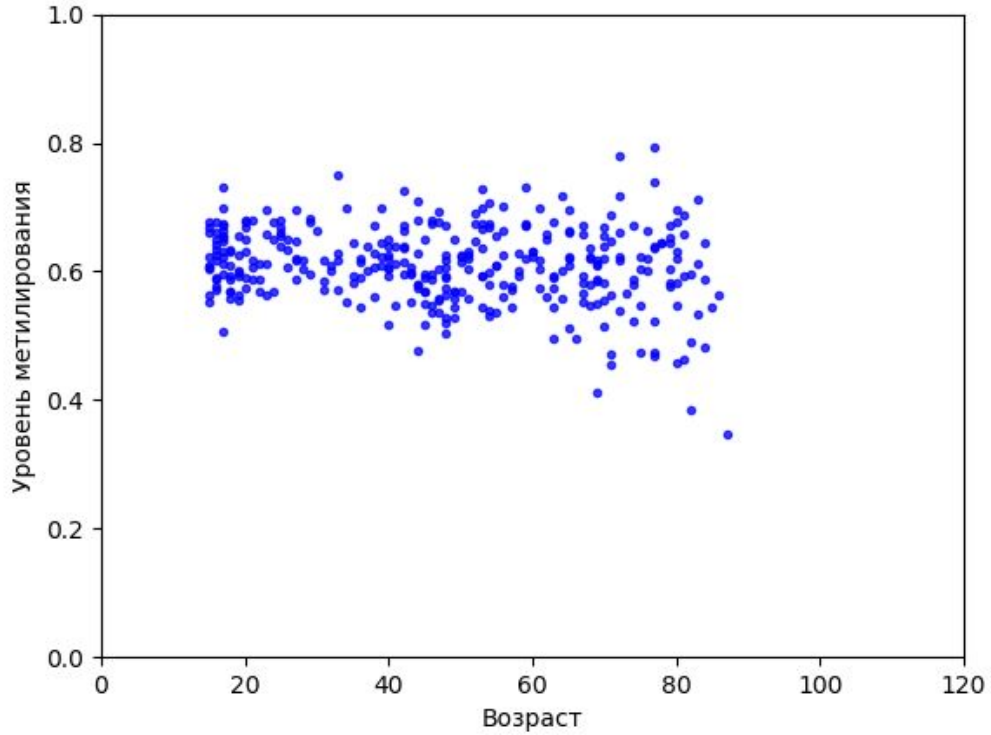


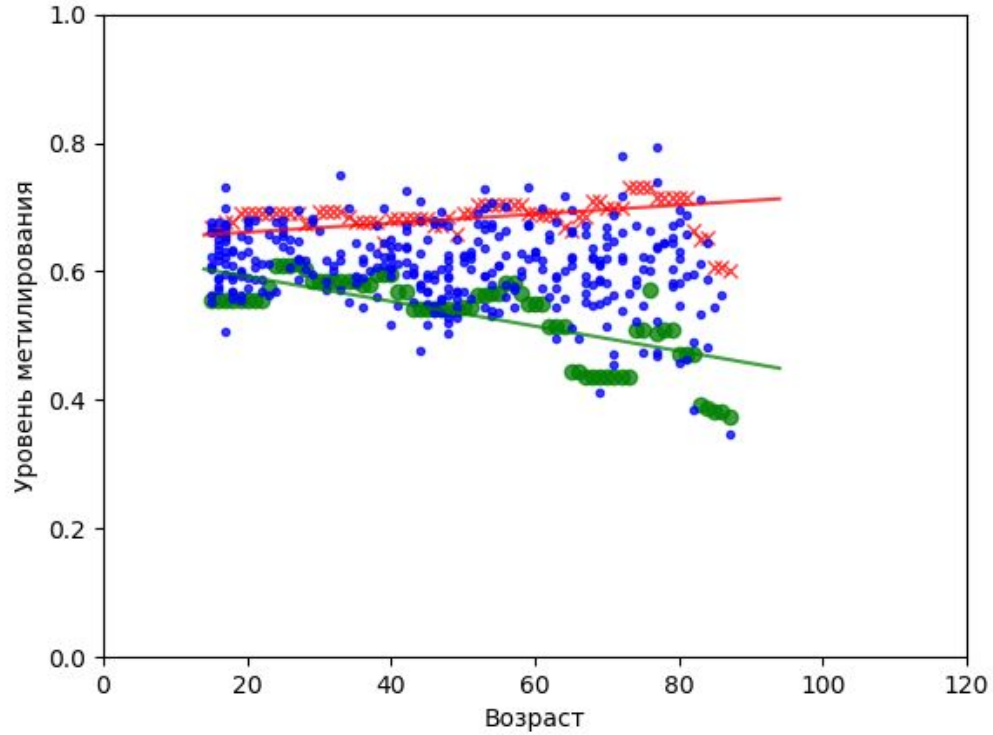
Применение линейной регрессии к данным метилирования

- $$\overline{CpG} = b_0 + b_1 \cdot \text{возраст}$$
$$\underline{CpG} = b_0 + b_1 \cdot \text{возраст, где}$$

$\overline{CpG}$ ($\underline{CpG}$) – значение 95(5)-процентного квантиля уровня метилирования данного CpG-сайта.

Главная метрика – коэффициент детерминации R^2.





Применение линейной регрессии к данным метилирования

Построение линейной регрессии проводится в четырех осях: lin-lin, lin-log, log-lin, log-log для определение максимального коэффициента детерминации.

$$R^{2\,M} = \min(R^{2\,b,M}_{\text{лучш}}, R^{2\,t,M}_{\text{лучш}})$$

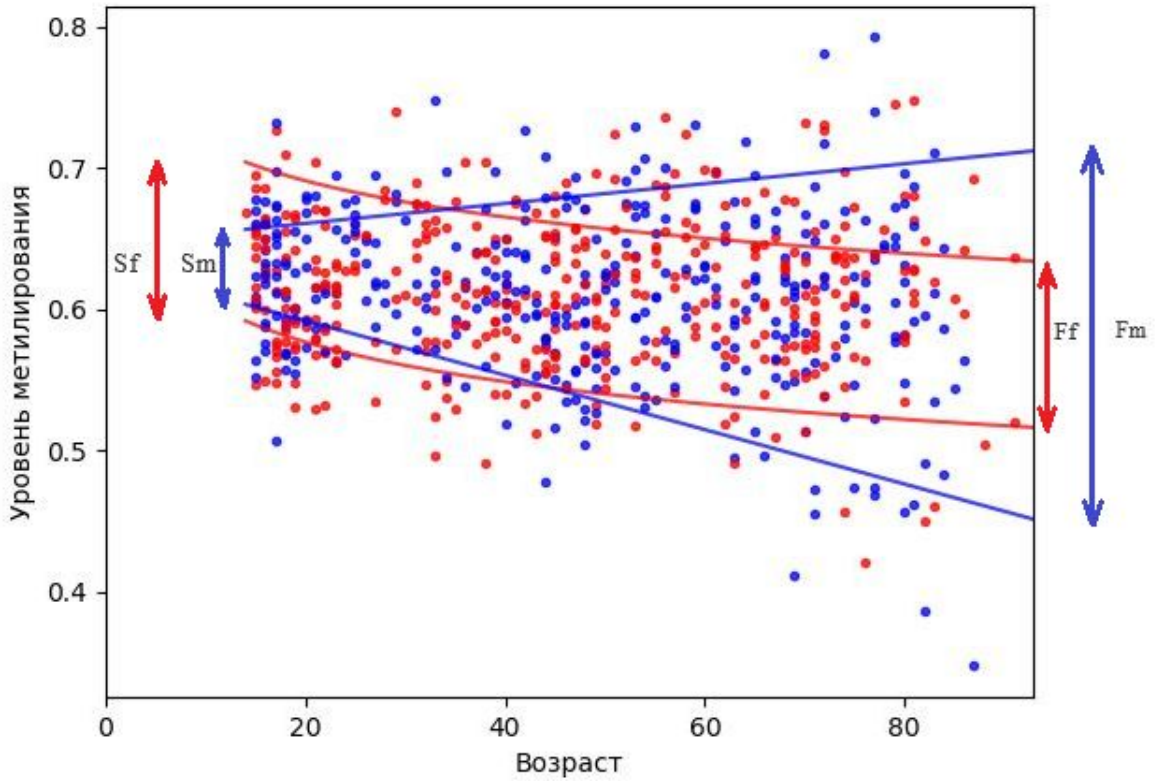
– минимальный коэффициент детерминации из максимальных по каждой из двух моделей для данных мужского пола конкретного CpG-сайта.

$$Z = \min(R^{2\,M}, R^{2\,F})$$

– метрика CpG-сайта на основе значений коэффициента детерминации построенных моделей для обоих полов.

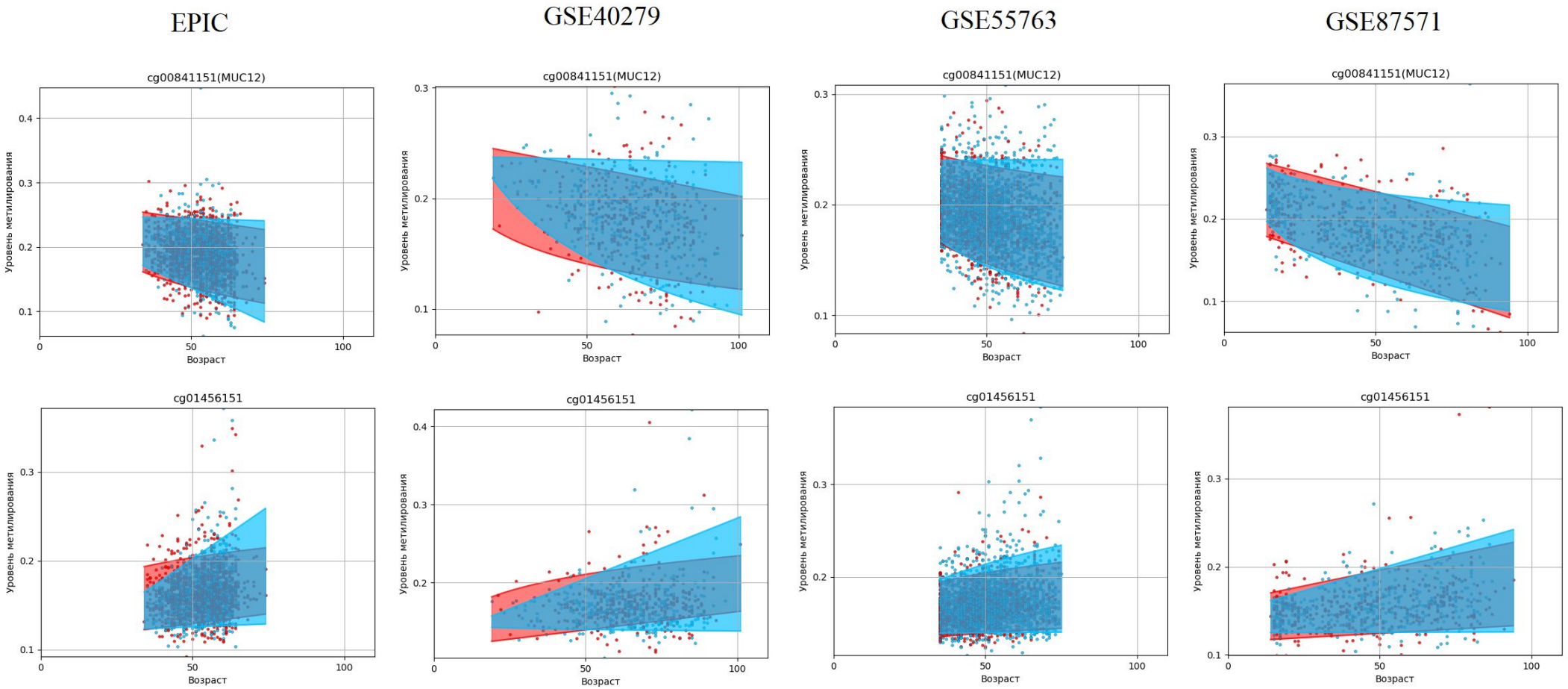
Анализ и сравнение областей, заключенных между полученными моделями для мужчин и женщин, позволит выявить возможное наличие гендерной специфичности в уровне метилирования отдельных CpG-сайтов.

Величины $I_m = \max\left(\frac{S_m}{F_m}, \frac{F_m}{S_m}\right)$ и $I_f = \max(\frac{S_f}{F_f}, \frac{F_f}{S_f})$ описывают величину расширения (сужения) областей данных для каждого пола.



Результаты

- В качестве лучших биомаркеров было взято пересечение наиболее ярких исходя из вышеописанного алгоритма для каждой из базы данных CpG-сайтов.
- При пересечении результатов были получены 45 CpG-сайта, демонстрирующие различную вариабельность для мужчин и женщин, 33 из которых имеют принадлежность к какому-либо гену.

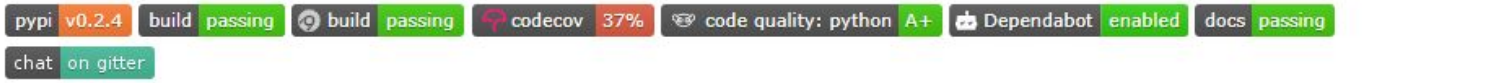


Функции генов

- Например, с кодирующим белок геном PRDM14 связаны такие заболевания как герминома головного мозга, пинеальной области, по некоторым данным которые чаще отмечается у пациентов мужского пола.
- Дисрегуляция гена IGFBP3 вовлечена в развитие рака молочной железы.

Программный комплекс

- Python 3.6/3.7
- Модульные тесты
- Использование оптимизированных структур данных
- Автоматическое тестирование Travis (Linux), Appveyor (Windows)
- Развернут на Python Package Index: `pip install pydnameth`
- Документация



pydnameth

DNA Methylation Analysis Package

Authors: Yusipov Igor, Kalyakulina Alena, Timakin Nikita, Soluyanov Alexey

This package provides some pipelines for analysis of methylation data. The main goal is to find correlations between methylation on the one hand, and age, sex, disease, and other characteristics of subjects on the other hand.

Examples of free-access methylation datasets:

- GSE40279
- GSE87571

Documentation

Available at https://pydnameth.readthedocs.io.

Features

- Defining best age-predictors CpGs for different subjects subsets.
- Defining best observable-specic CpGs (sex-specific, disease-specific) which are differently methylated with age.
- Building Epigenetic Clock.
- Plotting subjects distribution depending on the observable (sex, disease).
- Plotting methylation profiles for CpGs.

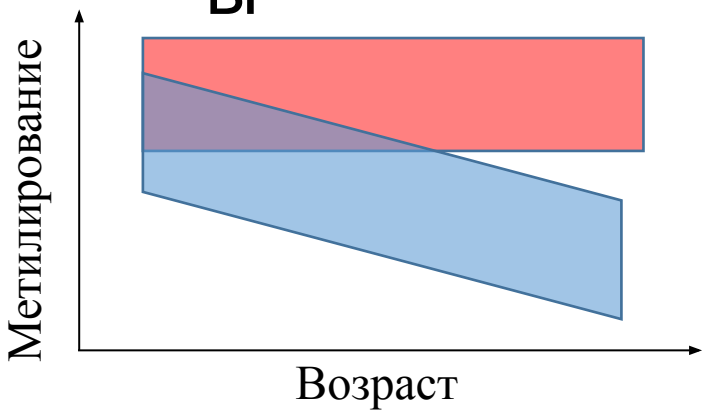
Copyrights

Free software: MIT license

Выводы

- Разработан алгоритм анализа данных уровня метилирования ДНК и отбора биомаркеров, имеющих гендерно-специфичную вариабельность.
- Создан программный комплекс, реализующий данный алгоритм, способный обрабатывать большой объем данных и визуализировать полученные результаты.

Тренды

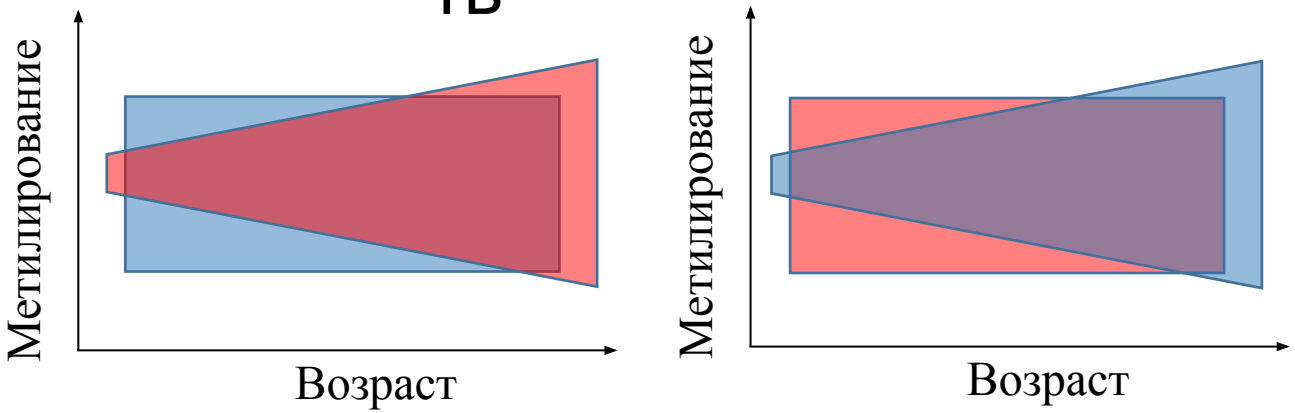


- Маркеры заболеваний, связанных с полом (маркеры мужского бесплодия, патологических процессов репродуктивных органов)

Singmann, Paula, et al. "Characterization of whole-genome autosomal differences of DNA methylation between men and women." *Epigenetics & chromatin* 8.1 (2015): 43.

- Рассматриваются различия между полами, без учёта зависимости от возраста

Вариабельность



- Маркеры заболеваний, связанных с накоплением эпигенетических мутаций (маркеры рака молочной железы, рака яичек)

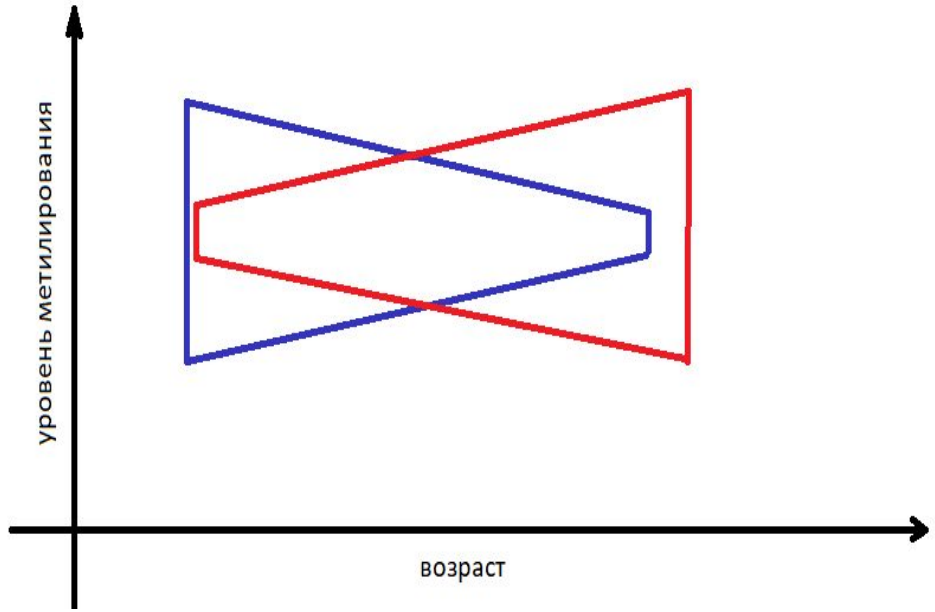
Slieker, Roderick C., et al. "Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms." *Genome biology* 17.1 (2016): 191.

- Рассматриваются возрастные тенденции вариабельности, без учёта пола

Спасибо за внимание!

Реализация

Сравнение только I_m и I_f не гарантирует полного результата, так как возможна ситуация, когда эти величины равны, но на графике картина выглядит следующим образом:



Подобная ситуация интересна для изучения, поэтому были введены дополнительные характеристики, позволяющие выявлять подобные картины:

$$S_I = \frac{S_f}{S_m}, \qquad F_I = \frac{F_m}{F_f}$$

Критерий Колмогорова-Смирнова

- Пусть X_n - выборка независимых одинаково распределённых случайных величин, $F_n(x)$ - эмпирическая функция распределения, $F(x)$ - предполагаемая функция распределения.
- Гипотезы: H_0- выборка взята из предполагаемого распределения, H_1 - выборка взята из другого распределения.
- α – уровень значимости критерия(вероятность принять H_1, когда верна H_0)
- Тогда $D_n = \sup |F_n(x) - F(x)|$ - статистика критерия.
- На основе D_n вычисляется p-значение (p-value).
- Если p-value > α, то H_0 не отвергается.