Аналитические методы поиска сайтов редактирования РНК в данных RNA-seq

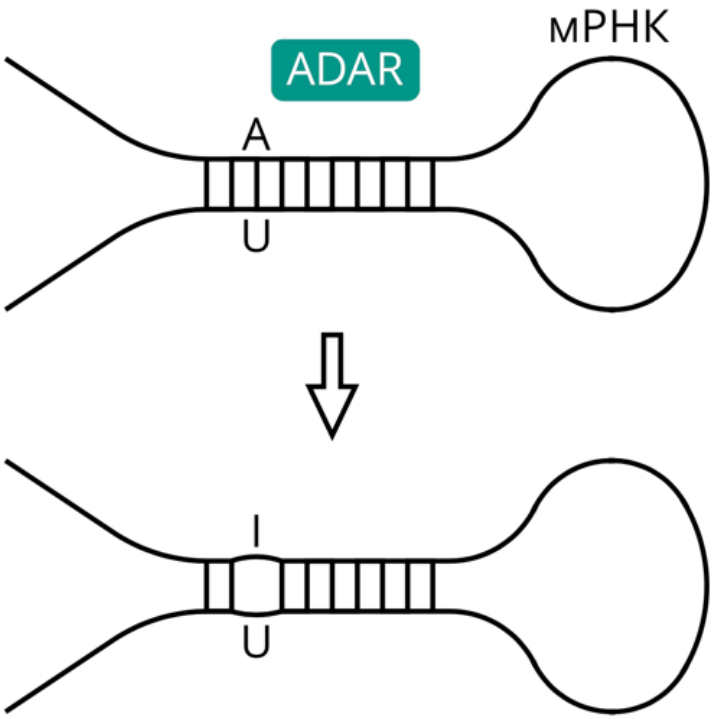
Щукина Ирина Алексеевна

научные руководители: А. Канапин, А. Самсонова

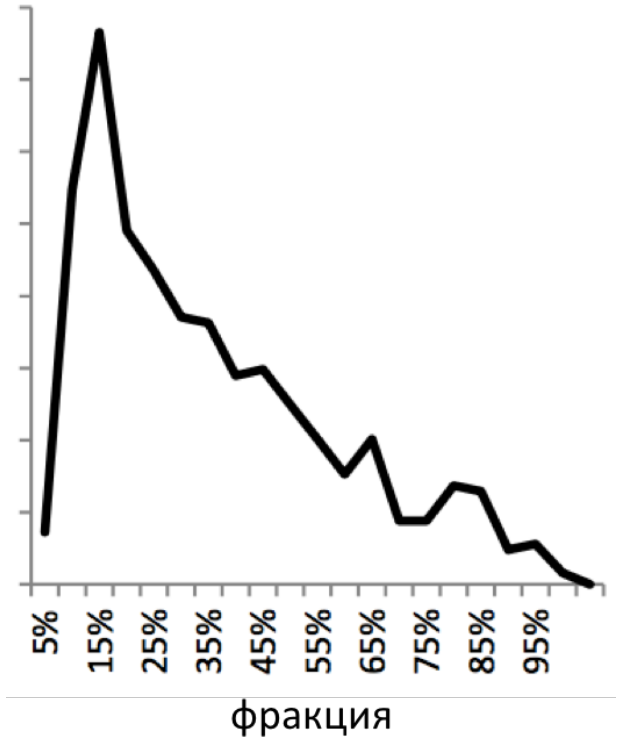
СПб АУ НОЦНТ РАН

14 июня 2017 г.

Редактирование РНК: A $\rightarrow$ I



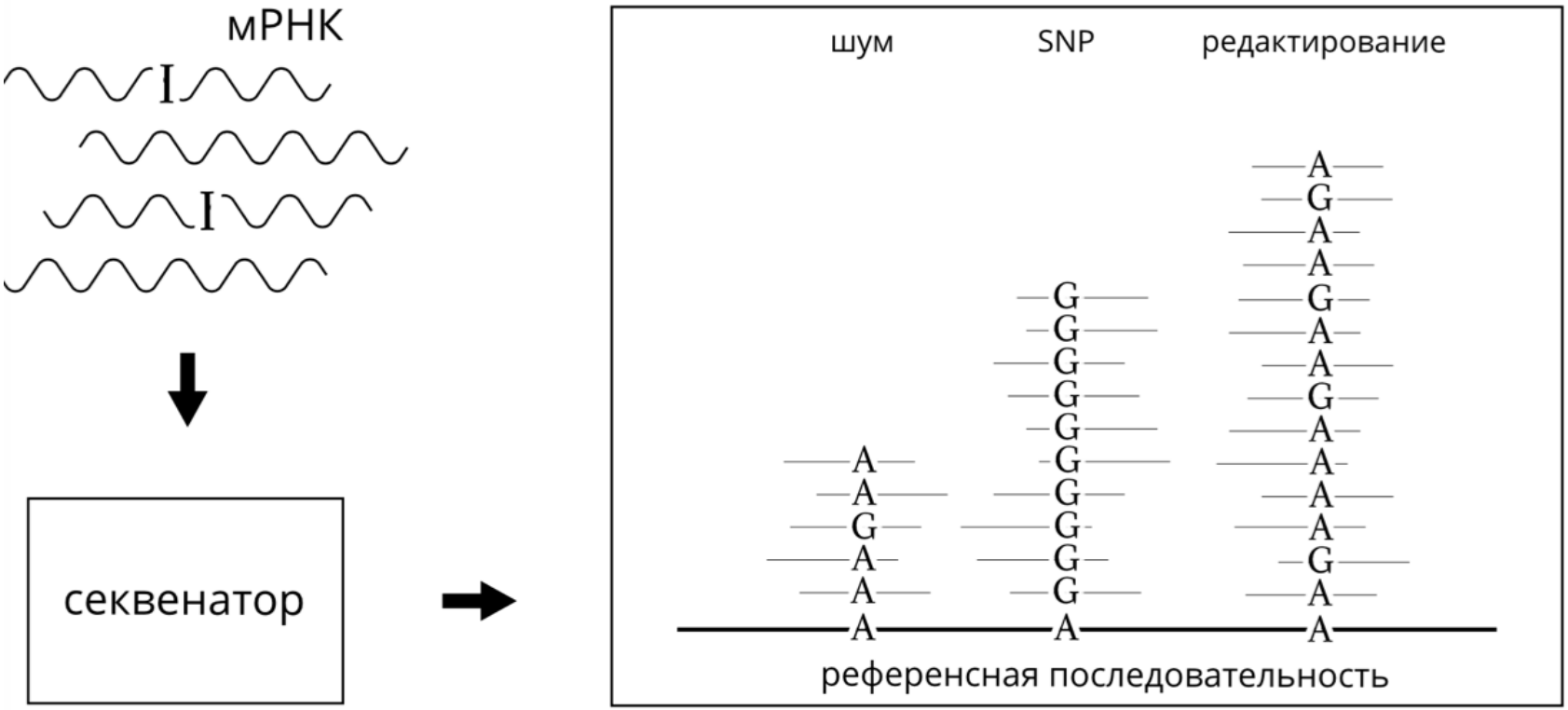
Клеточная машинерия и секвенатор интерпре-тируют инозин (I) как гуанин (G)



Ожидаемый вид распределения фракций отредактированных нуклеотидов

Редактирование в RNA-seq

AG = редактирование + технический шум + SNPs



Существующие подходы

- REDITools
- RED
- GIREMI
- Фильтрация по фракции/покрытию/качеству...
- Использование DNA-seq или нескольких реплик для отбора сайтов

REDITools denovo

Принцип работы:

1. Оценивает вероятность замены каждого типа по всему эксперименту

$$H_0 : p_{AG} = \frac{n_G}{n_G + n_A}$$
$$H_1 : p_{AG} < \frac{n_G}{n_G + n_A}$$

2. Точный тест Фишера для каждой позиции
3. Поправка на множественное тестирование

Цель и задачи проекта

Цель: разработать метод поиска сайтов редактирования РНК в данных RNA-seq

Задачи:

1. Изучить и модифицировать существующий алгоритм REDITools denovo
2. Применить байесовский подход к фильтрации технического шума
3. Разработать алгоритм для фильтрации SNP
4. Сравнить подходы с аналогами и произвести отбор

Модификация REDITools

- Знаем, что редактированию соответствует только замена типа AG. Остальные типы – шум, будем оценивать вероятности ошибки только по ним
- Дополнительный тест для фильтрации гомозиготных SNPs (оригинальный метод никак не затрагивает эту проблему).

Байесовский подход

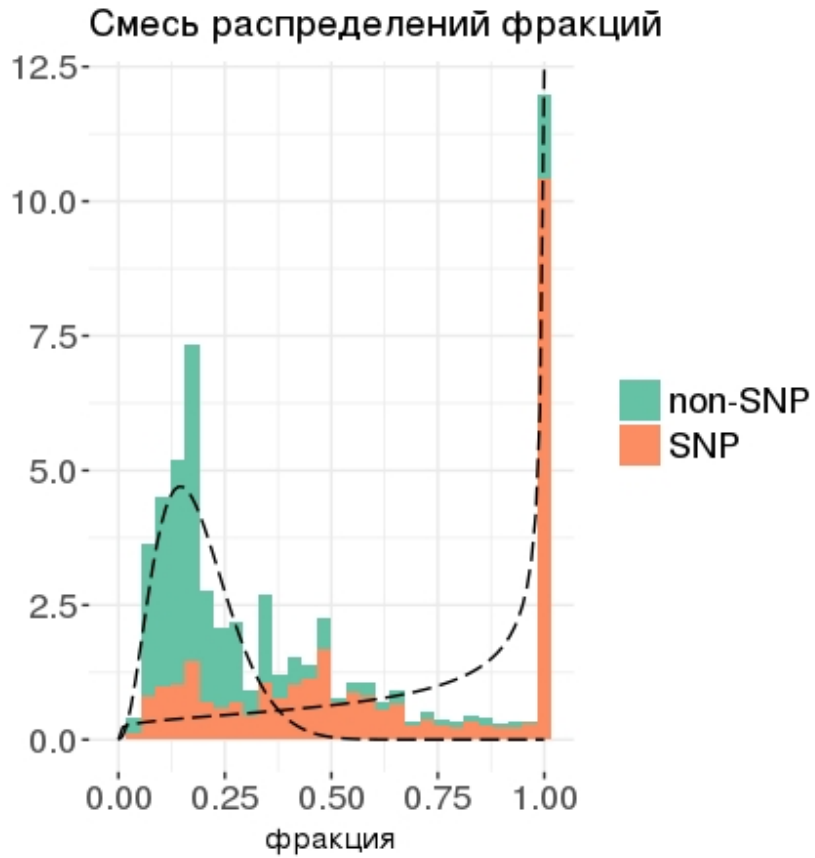
1. Оценим по шуму априорное распределение фракций
$$p_{AG} \sim \mathrm{B}(\alpha_0, \beta_0)$$
2. Для всех сайтов типа AG найдём апостериорное распределение
$$p_{AG} \,|\, n_A, n_G \sim \mathrm{B}(\alpha_0 + n_G,\ \beta_0 + n_A)$$
3. Оценим вероятность
$$\mathrm{P}[p_{AG} \,|\, n_A, n_G \leqslant \mathrm{thr}]$$
4. Установим FDR и отберём сайты в соответствии с ним

Данный метод позволяет избавиться от технического шума

Фильтрация SNP

Методы:

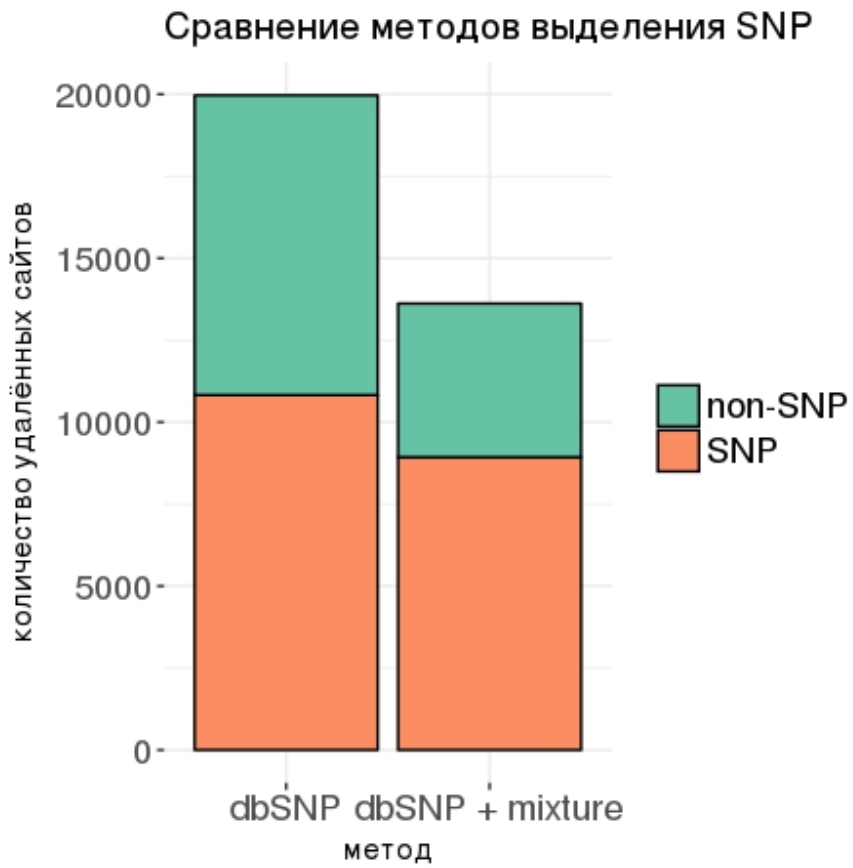
- Фильтрация против dbSNP — базы всех встречающихся SNP (избыточно)
- Смоделируем смесь распределений (EM алгоритм) и удалим позиции, из “SNP-компоненты”



Фильтрация SNP

Методы:

- Фильтрация против dbSNP — базы всех встречающихся SNP (избыточно)
- Смоделируем смесь распределений (EM алгоритм) и удалим позиции, из “SNP-компоненты”



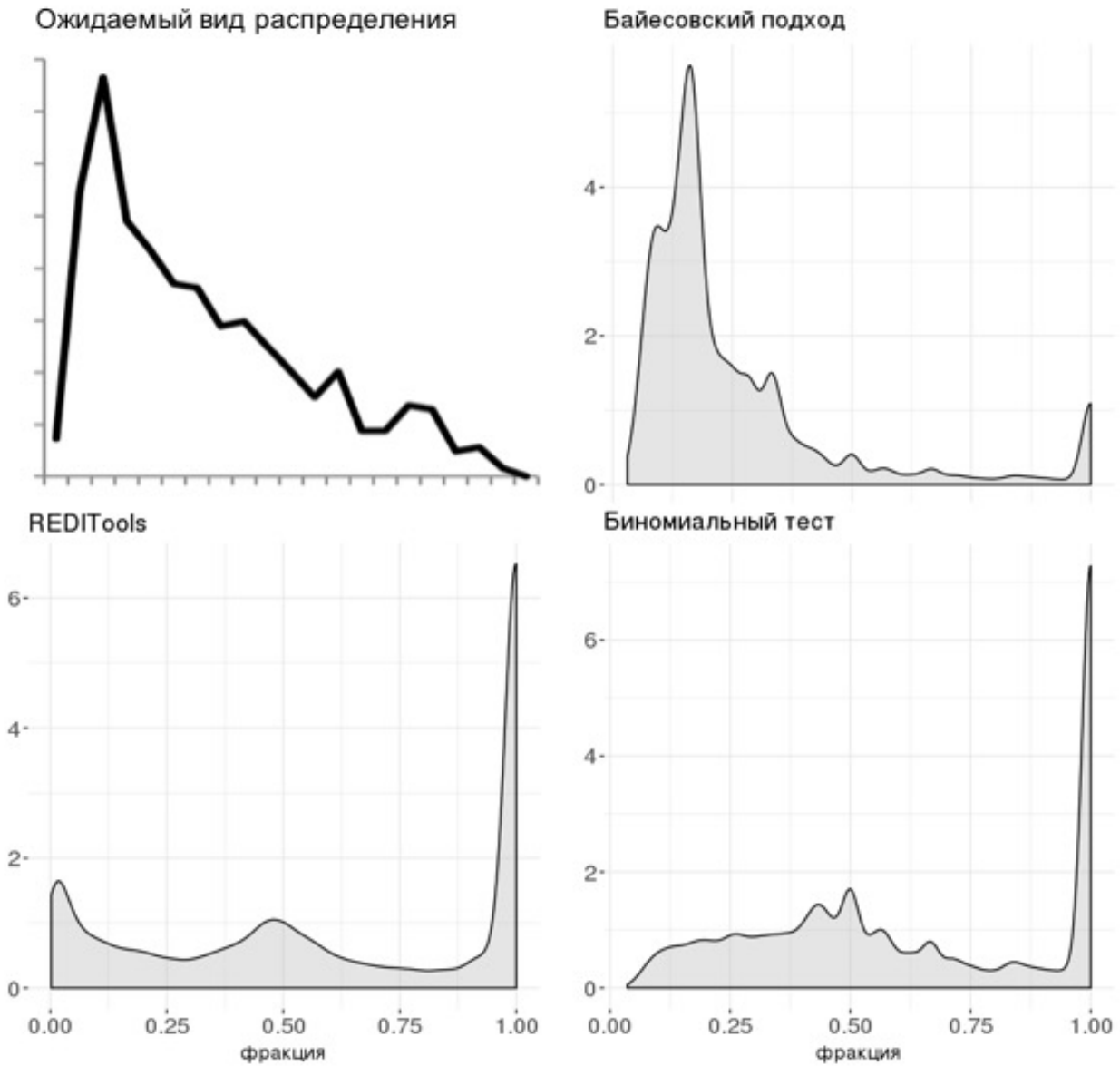
Сравнение методов

В роли baseline-метода использован биномиальный тест

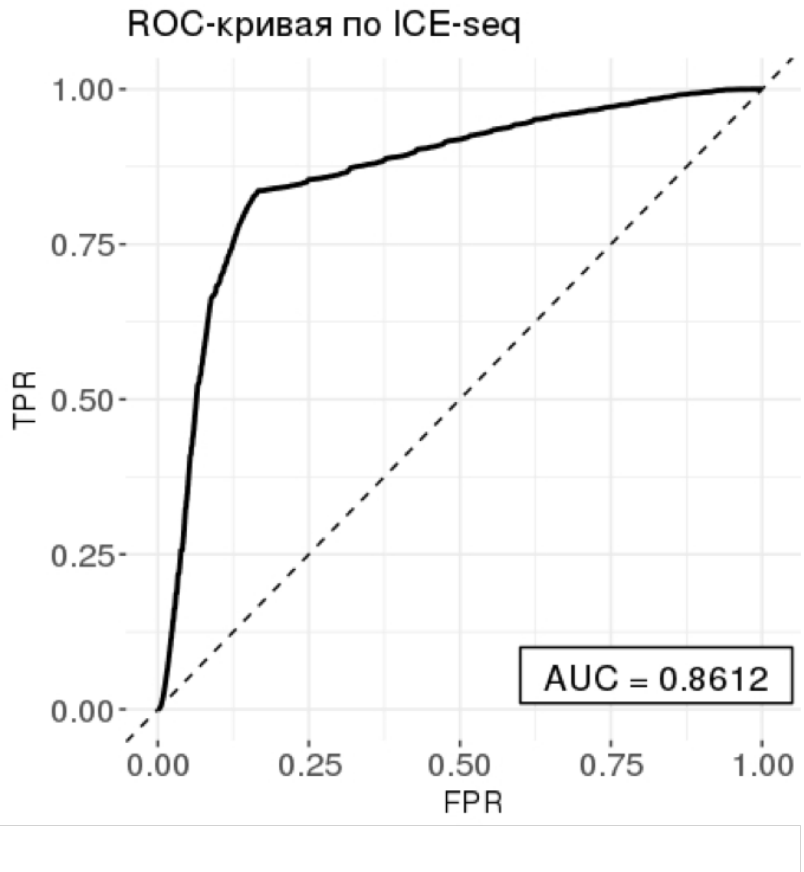
Данные:

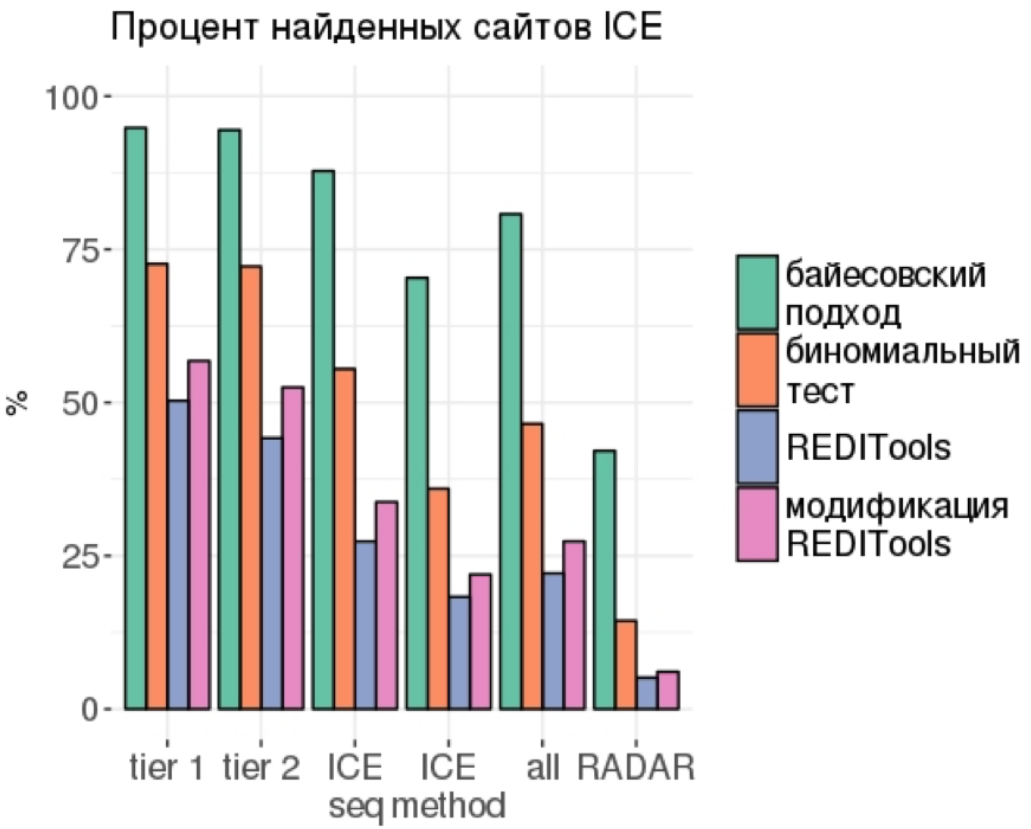
- GM12878 — достоверно известны SNP
- ICE-seq — экспериментальный протокол для поиска сайтов редактирования. Существует один датасет с набором (не полным) подтверждённых сайтов редактирования. Сайты разбиты на подгруппы в зависимости от степени валидированности.

Получаемые распределения фракций

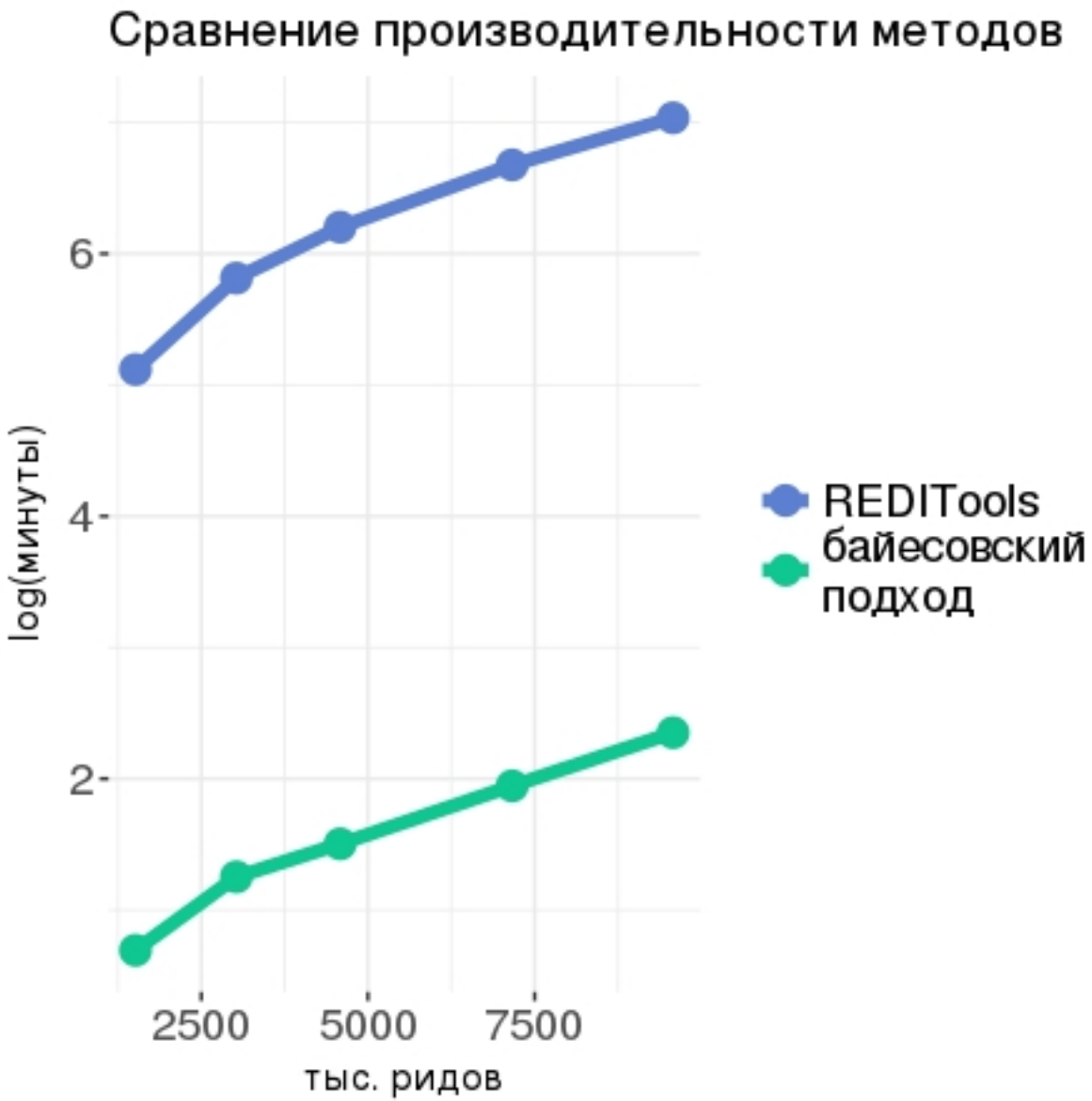


Сравнение методов





Сравнение методов



Дополнительные возможности

Помимо основного алгоритма реализованы:

- возможность фильтрации и определение стренда с помощью аннотации генома
- метрика для ранжирования сайтов на основе knn (значимое различие для TP сайтов ICE: $p < 2.2 \cdot 10^{-16}$)

Результаты

- модифицированы оценка вероятности и фильтрация SNP в существующем алгоритме REDITools
- применение байесовского подхода дало ощутимое улучшение результатов фильтрации технического шума
- разработан алгоритм для фильтрации SNP на основе базы dbSNP и разделения позиций по фракциям
- на основе отобранных методов реализован пайплайн для нахождения сайтов редактирования в данных RNA-seq

Сравнение методов