# Анализ ошибок амплификации

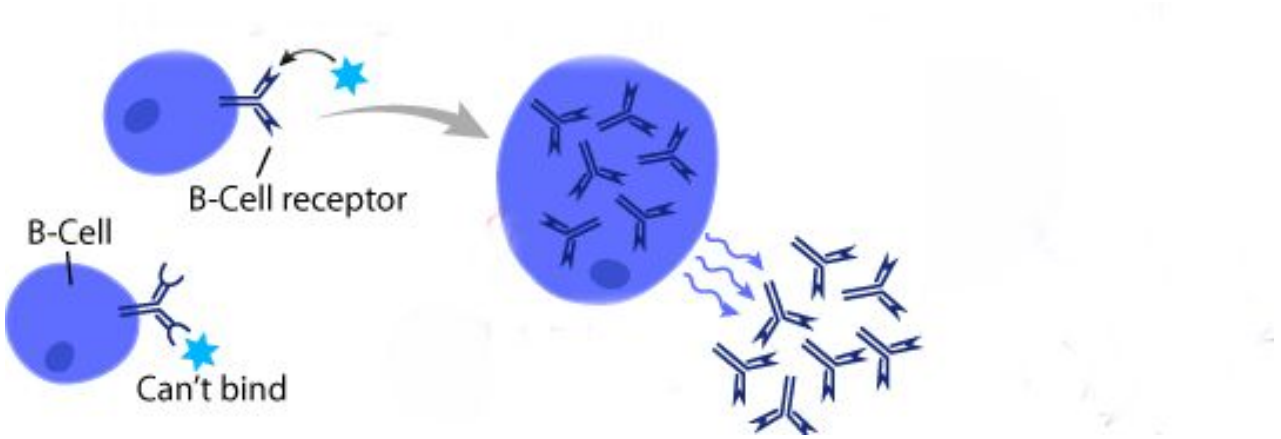
Денис Коноплев

Руководители:
Яна Сафонова
Александр Шлемов

Центр алгоритмической биотехнологии, СПбГУ

# Введение

- B-клетки - часть адаптивного иммунитета. Они вырабатывают белки, называемые антителами.
- Антитела связываются с вредными компонентами (антигенами) и обезвреживают их.
- Количество различных антигенов очень велико.
- Универсальность иммунного ответа обеспечивается формированием уникальных для B-клетки генов, производящих антитела.

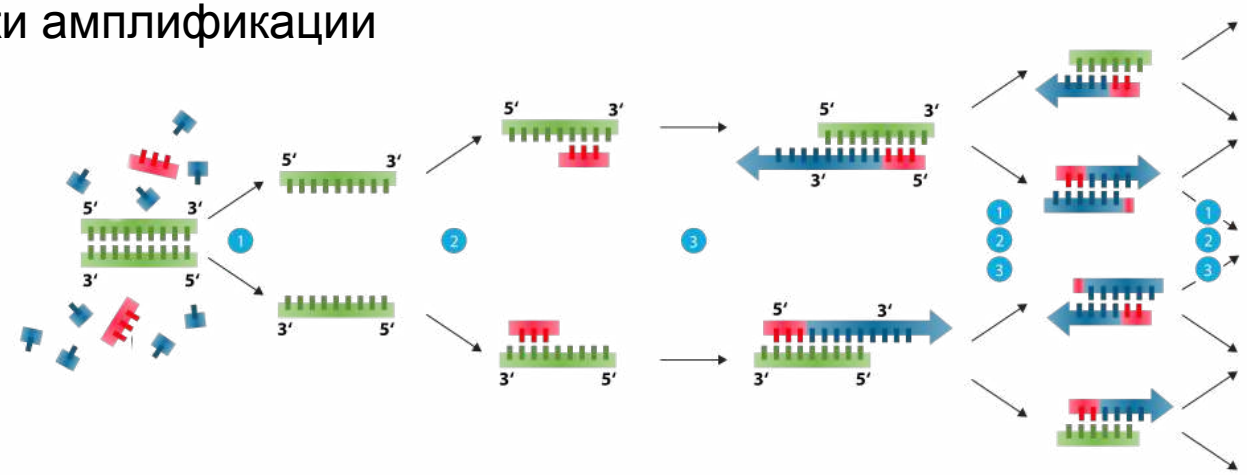


# Иммуносеквенирование

Иммуносеквенирование позволяет сканировать репертуар антител, что применяется при разработке лекарств и мониторинге лечения

Сложности, возникающие при анализе данных иммуносеквенирования

- Ошибки секвенирования
- Ошибки амплификации

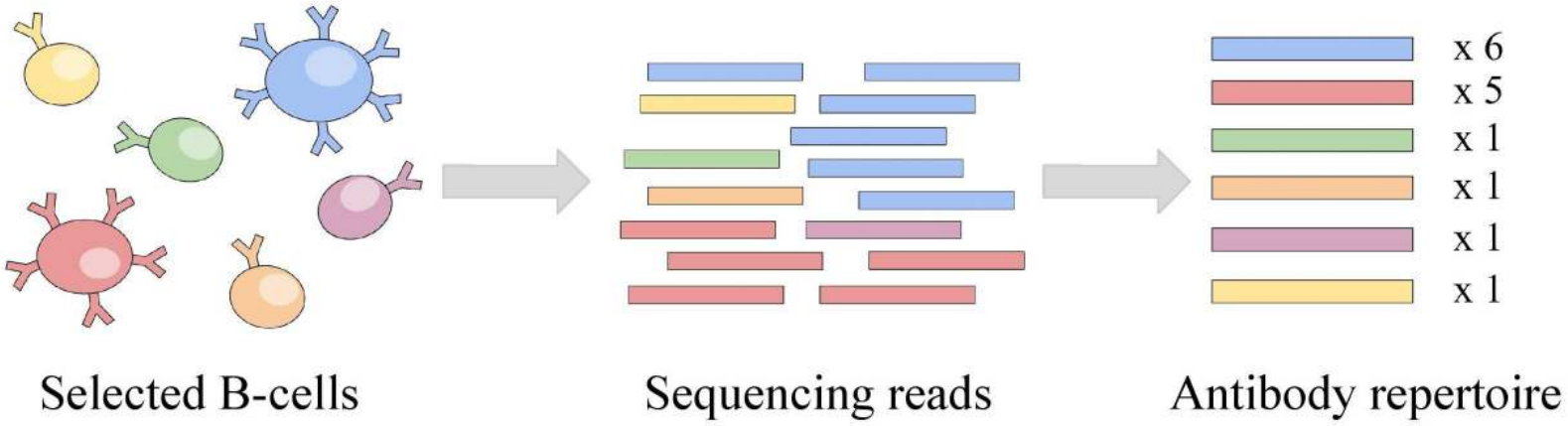



- Естественные мутации B-клеток порой очень похожи на ошибки # не ошибки!

# IgRepertoireConstructor

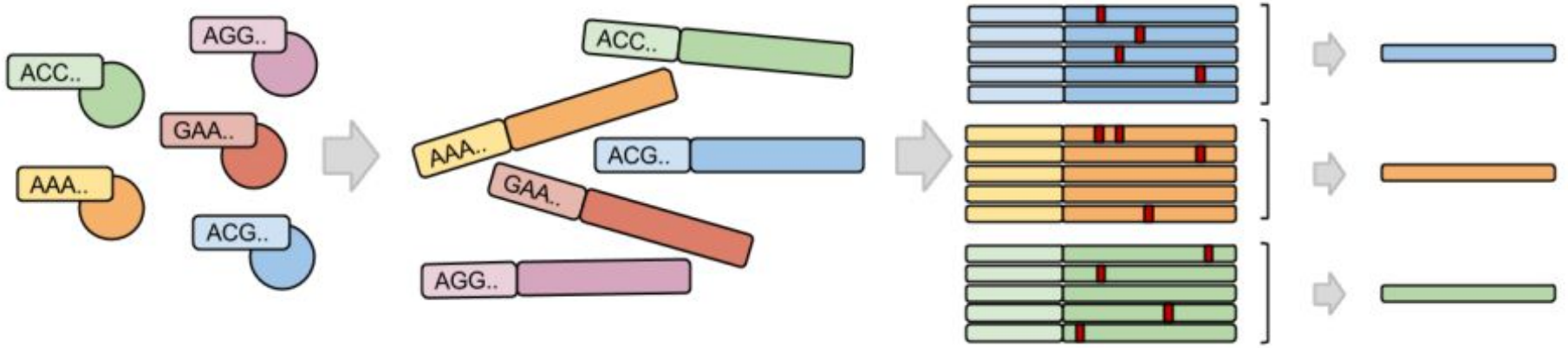
Инструмент, разработанный в лаборатории алгоритмической биотехнологии, позволяющий строить репертуар антител на основе данных иммуносеквенирования.

IgRepertoireConstructor исправляет ошибки секвенирования и амплификации и пытается восстановить истинные последовательности антител



# Анализ ошибок амплификации

- Амплификация вносит ошибки, которые могут содержаться во многих ридах.
- Чтобы исправлять их, нужно понять, на каком этапе произошла ошибка и какие замены происходят чаще.
- Предлагается работать с данными молекулярного баркодирования и на их основе предложить модель замен, производимых амплификацией.



# Задача

После выделения ридов, относящихся к одному антителу, нужно восстановить исходную последовательность.

Что делать когда риды разделяются на две равные части? Какой нуклеотид выбрать?

ACAGATCGAAACATACGTA

ACAGATCGAAACATACGTA

ACAGATCGATACATACGTA

ACAGATCGATACATACGTA

Текущая версия IgRepertoireConstructor не учитывает специфику ошибок амплификации при построении консенсуса.

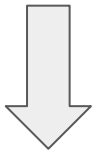
# Наивный метод

Нуклеотид в консенсусе определяется как наиболее частый нуклеотид на данной позиции.

ACAGATCGAAACATACGTA

ACAGATCGATACATACGTA

ACAGATCGATACATACGTA



ACAGATCGATACATACGTA

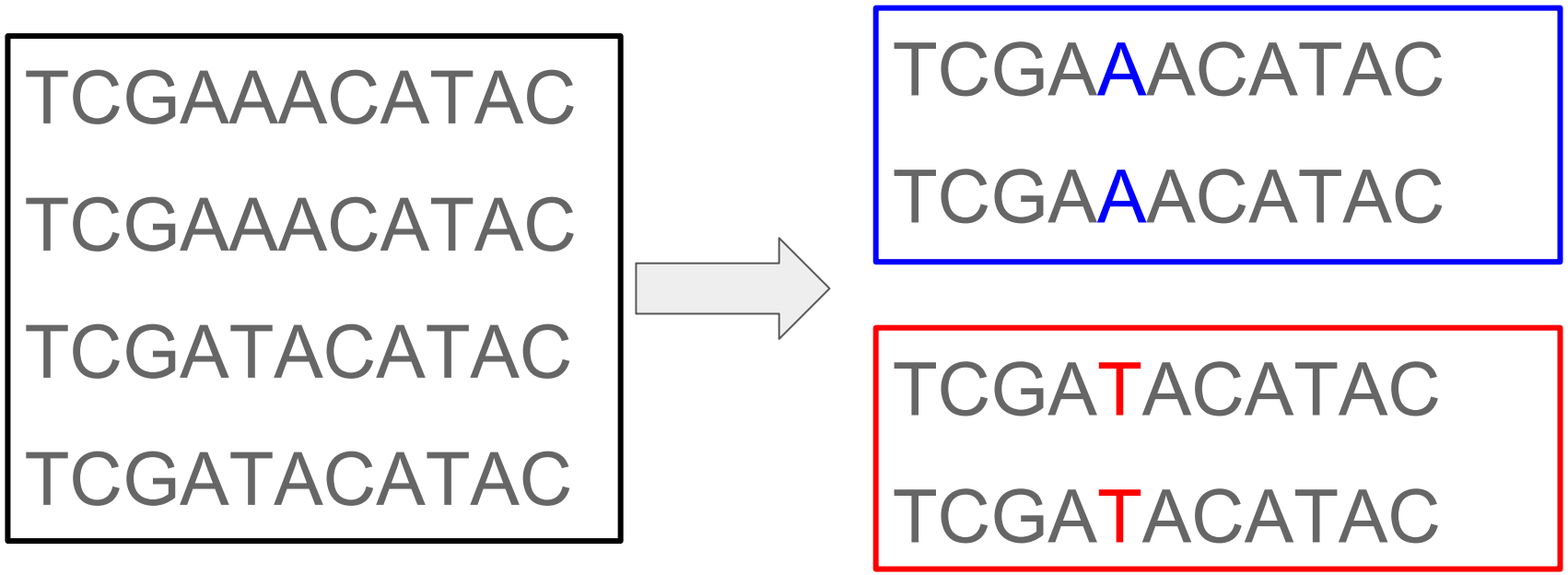
## Проблемы

- Модель не отражает реальность
- Непонятно, что делать с ситуацией когда риды разделяются пополам

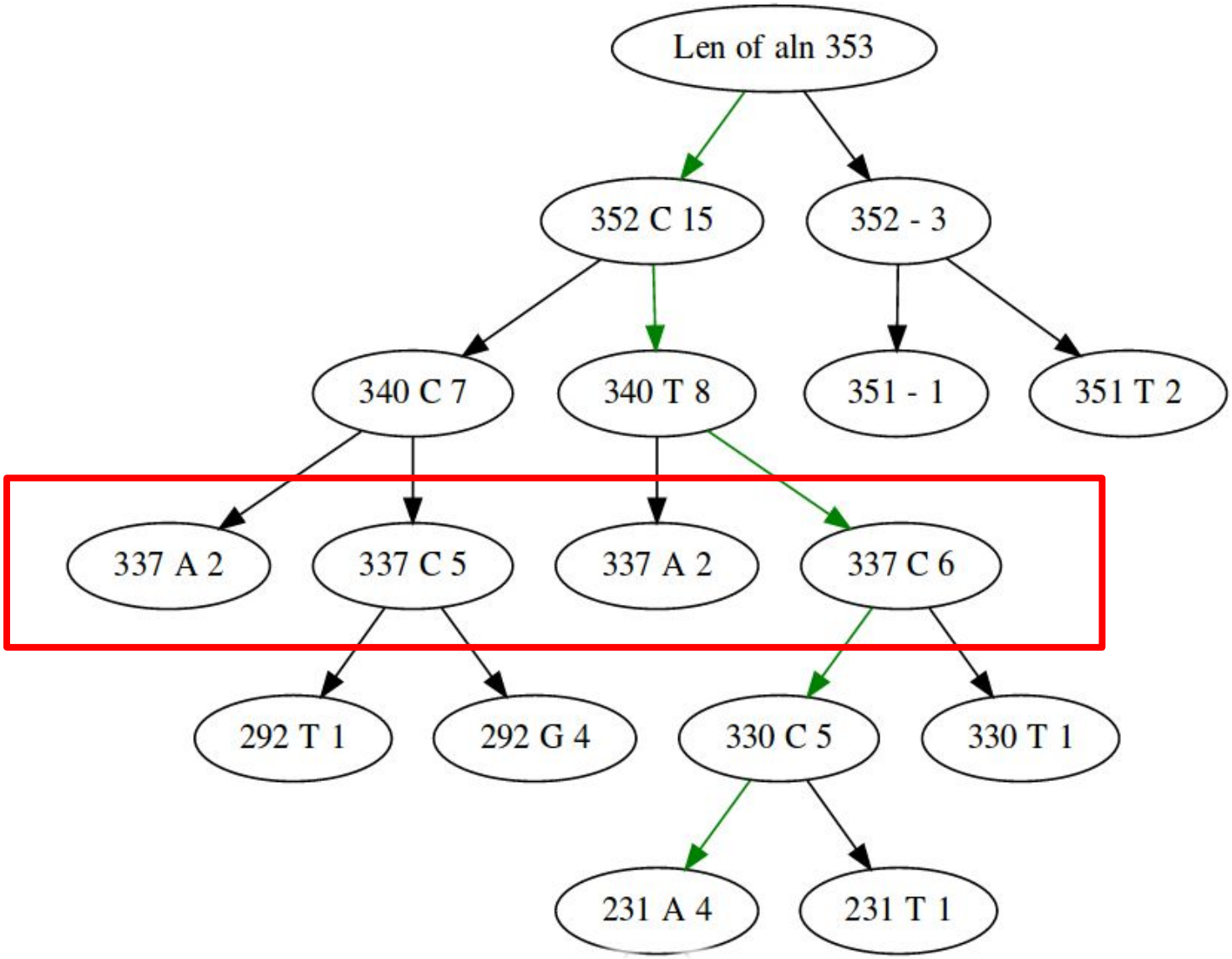
# Построение дерева мутаций

Идея – давайте рассматривать каждую ошибку амплификации, как ответвление от основной ветви. Данные будут разделяться на независимые части. К каждой из полученных частей применять тот же подход.

Предлагается посчитать количество различных мутаций, для того чтобы разрешать ситуации, когда риды делятся на примерно одинаковые части.

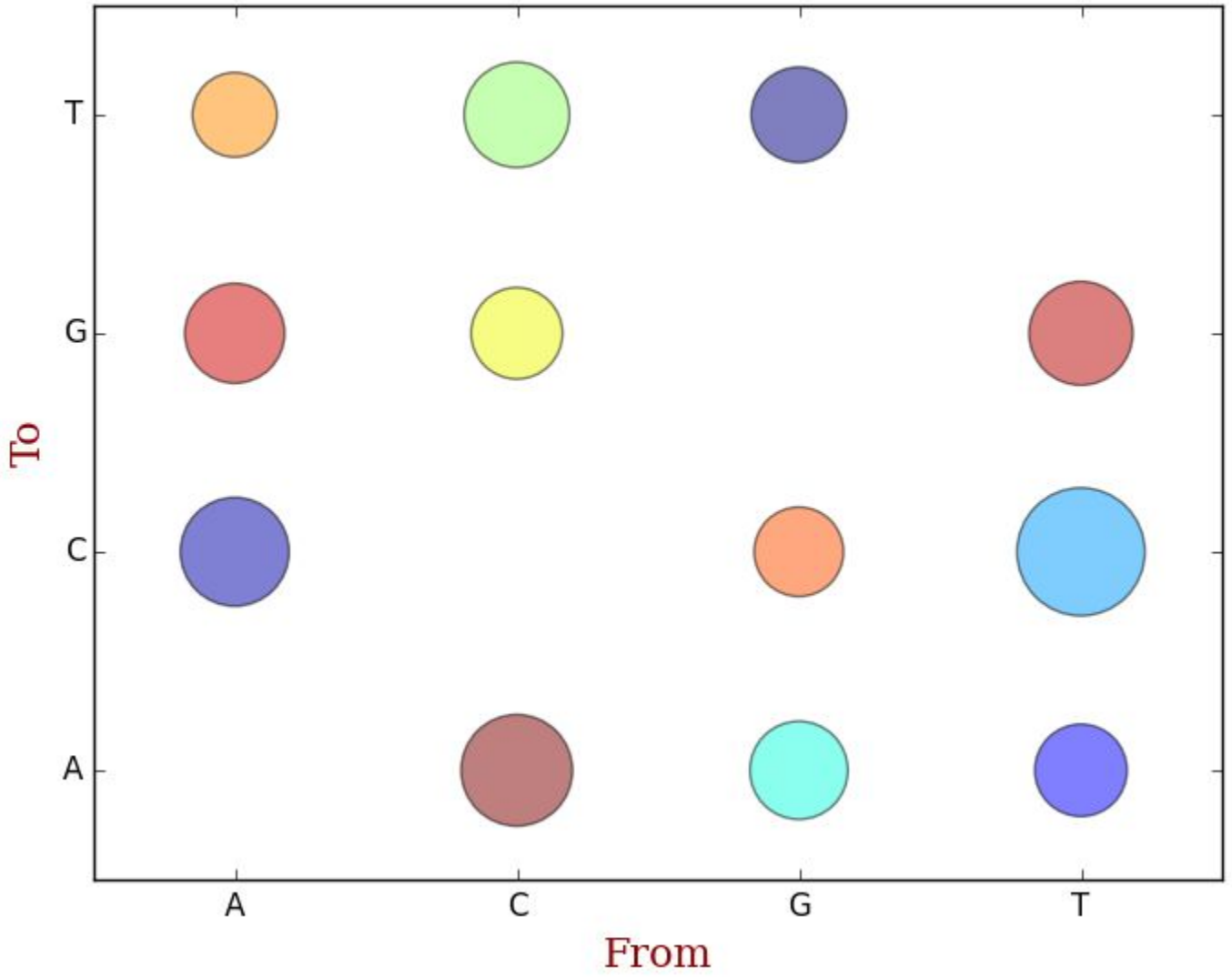


# Дерево мутаций

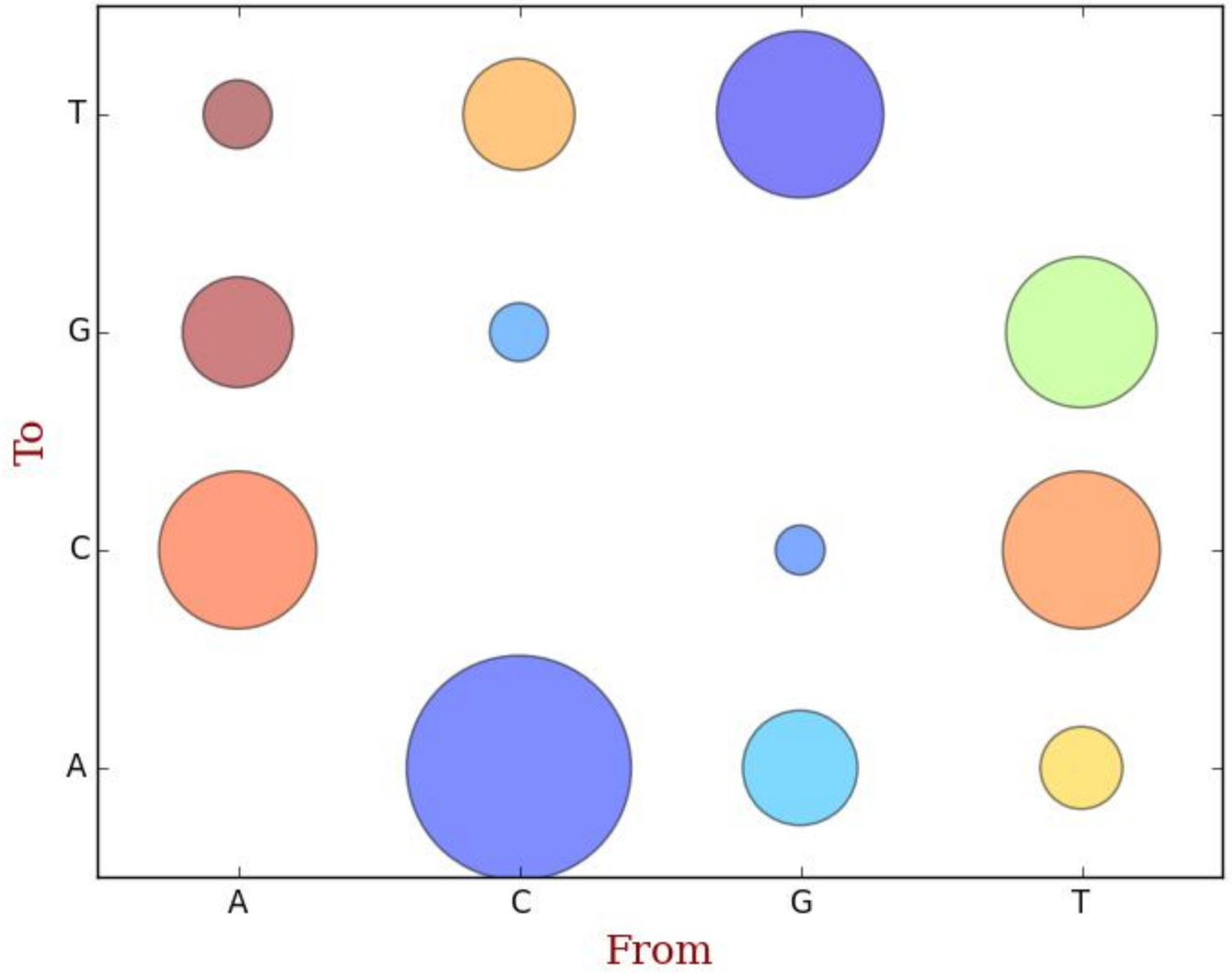


# Матрица замен для наивного подхода

Сложно выделить мутации, которые случаются чаще других.



# Матрица замен для дерева



# Верификация

В результате работы удалось выделить 5 из 6 пар, в которых можно принять решение при разделении ридов на равные части.

Чтобы проверить полученный результат, были взяты баркоды, относящиеся к контаминациям с известным референсом (РНК консервативных генов).
В 4 из 5 пар результат был подтвержден.

Пары, в которых один нуклеотид мутировал в другой значительно чаще

$$C \rightarrow A,\ G \rightarrow C,\ T \rightarrow C,\ G \rightarrow T$$

Пары, про которые нельзя сказать что-то определенное

$$G \sim A,\ A \sim T$$

# Результаты

- Реализованы два подхода построения консенсуса для амплифицированных данных
- Автоматизировано построение дерева мутаций
- Получены матрицы замен для двух подходов
- Верификация с помощью референса и биномиального теста подтвердила то, что в 4 из 6 замен являются значимыми

Спасибо за внимание