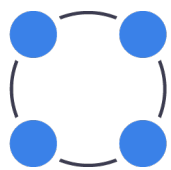
# Assembly of mammalian genomes using GemCode data

Кектеева Ангира

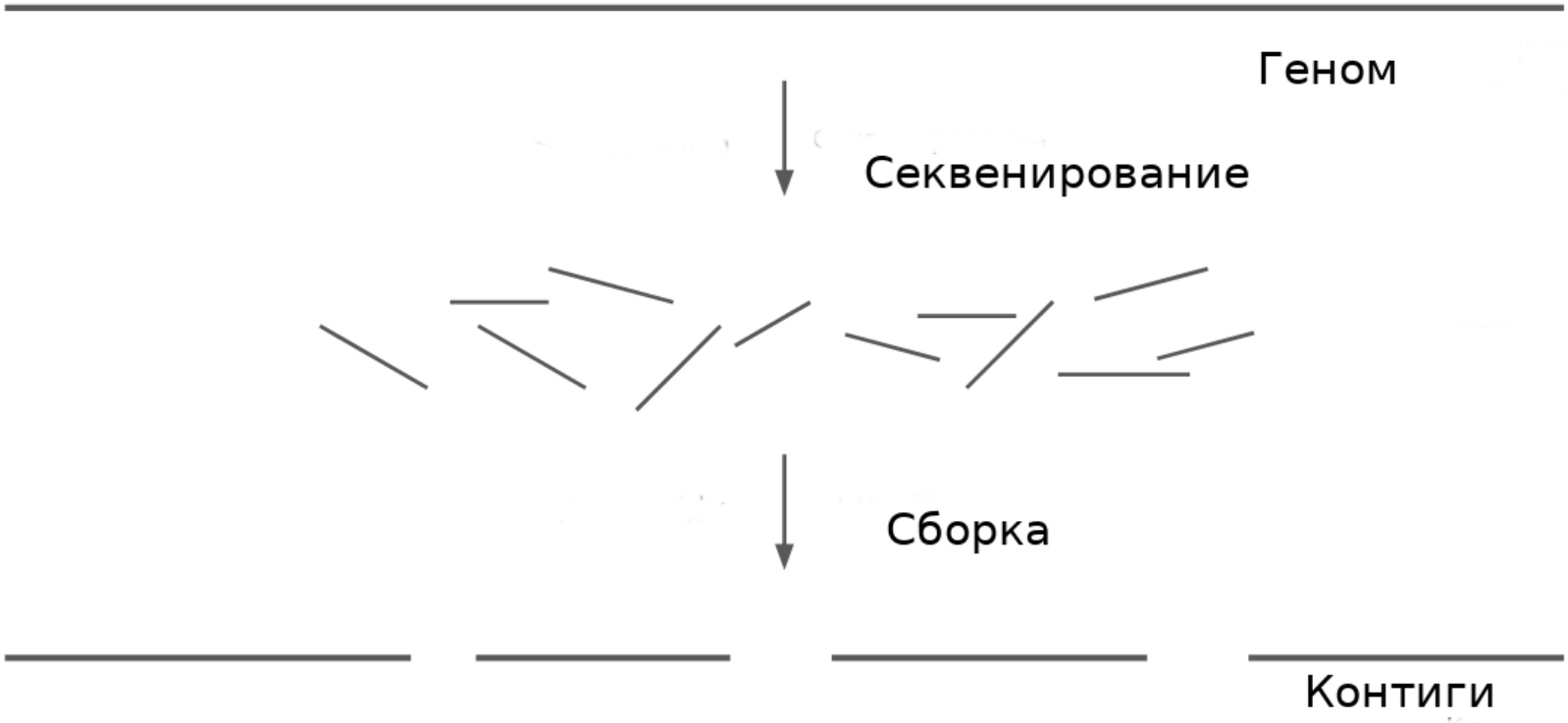
Руководители: Иван Толстоганов,
Антон Банкевич

16 декабря 2017

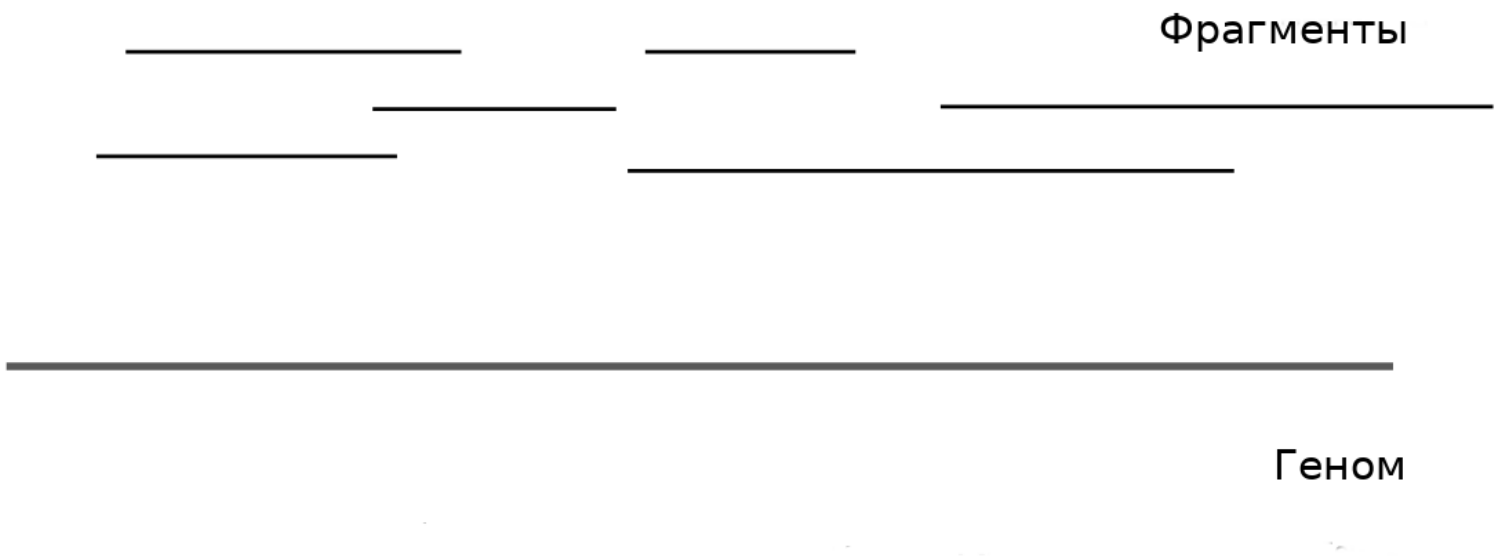


BIOINFORMATICS
INSTITUTE

# Сборка генома

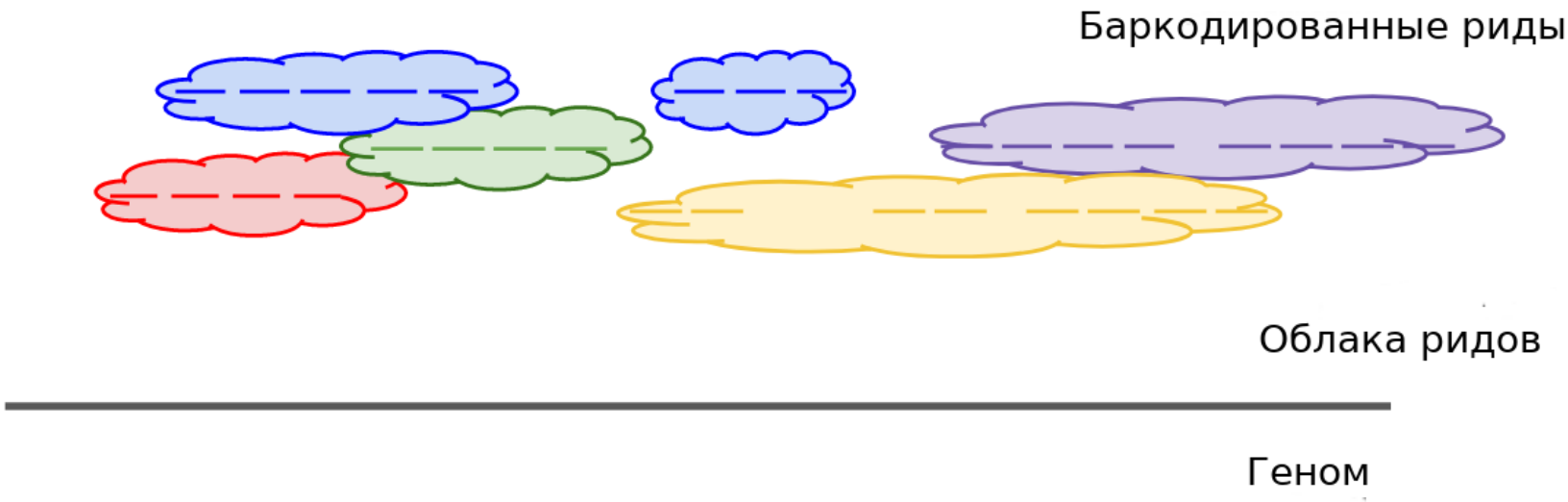


# Облака ридов




Последовательность ДНК "разрезана" на длинные фрагменты.

# Облака ридов




Фрагменты отмечены баркодами

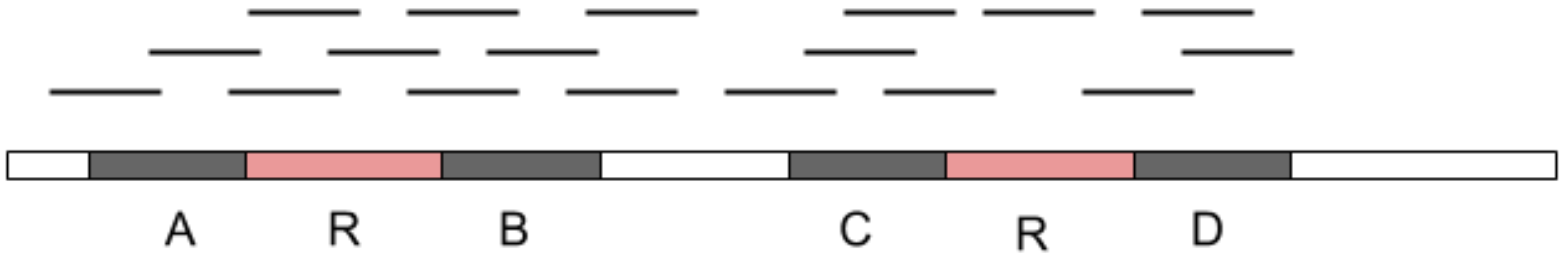
# Задача

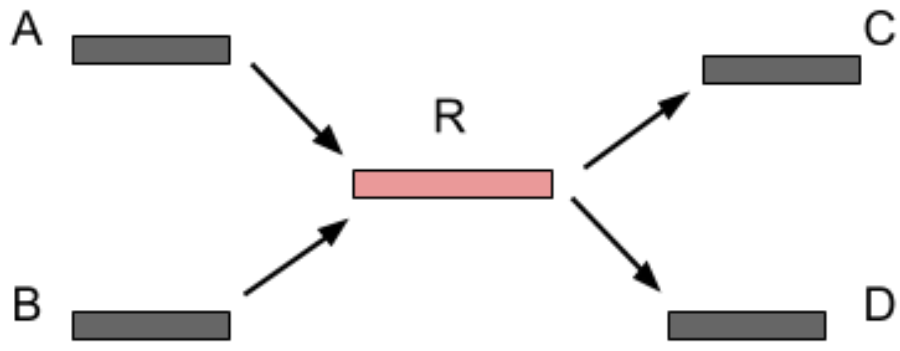
Основной задачей CloudSPAdes является сборка метагеномов, однако используемые в данном инструменте алгоритмы разрешения повторов в графе сборки могут быть применены и к сборке млекопитающих.

Задачи:

- Изучение существующих алгоритмов сборки с помощью баркодов
- Анализ недостатков применения текущей стратегии для больших геномов

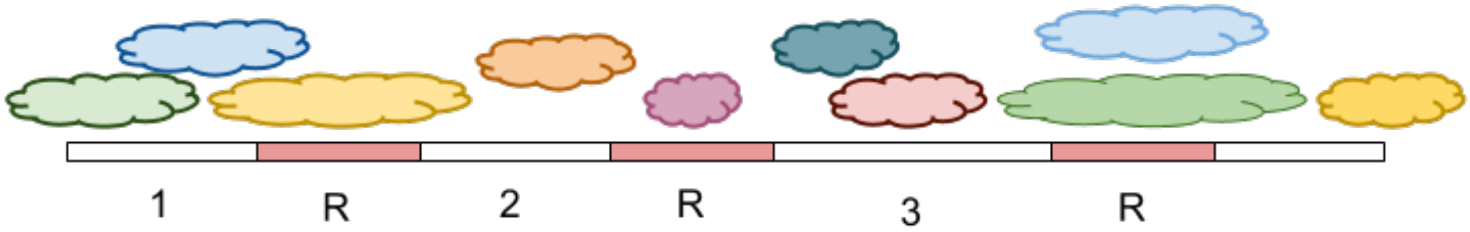
# Проблема повторов



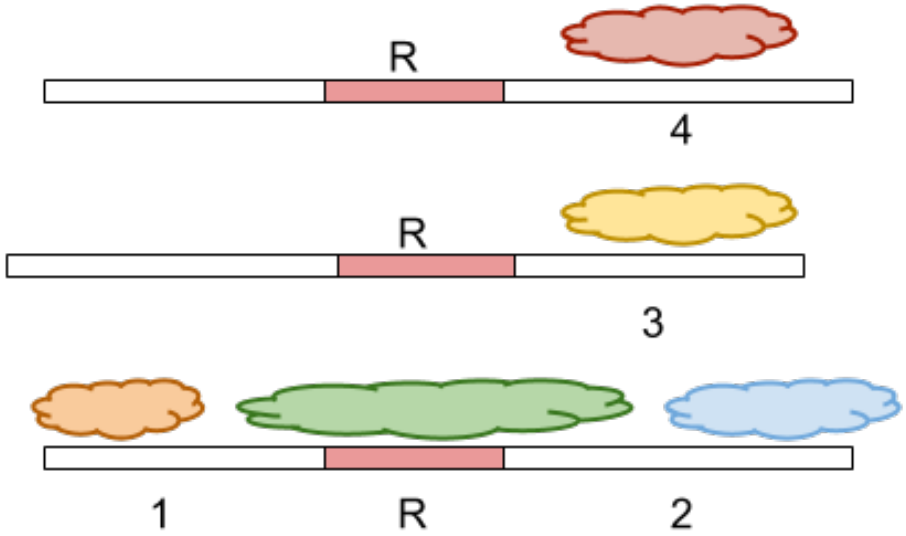



# Геном млекопитающих vs метагеном

Геном млекопитающих

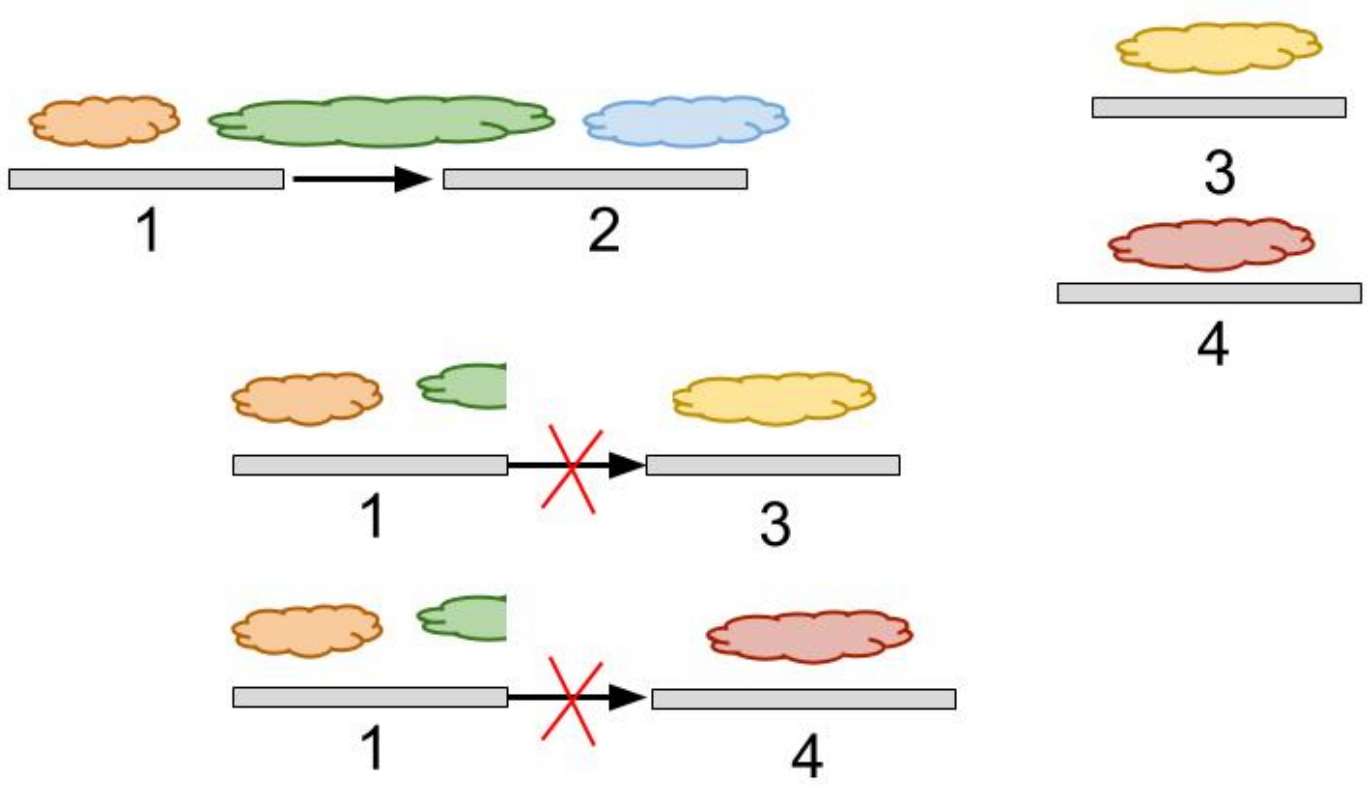



Метагеном



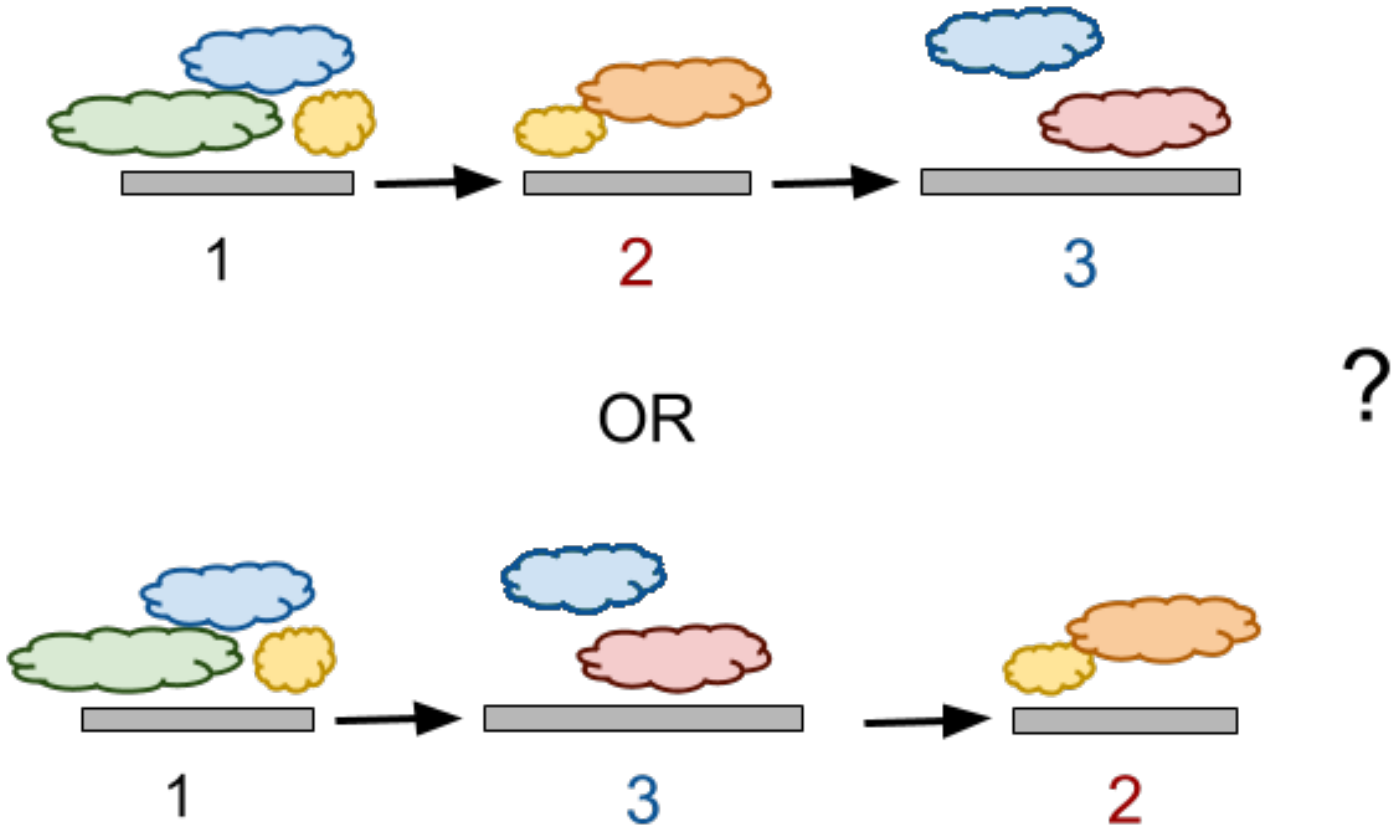
# Проблема упорядочивания фрагментов генома

В случае метагеномов сравнительно легко определить порядок следования фрагментов с помощью облаков.



# Проблема упорядочивания фрагментов генома

В случае геномов млекопитающих иначе: длинные ребра, связанные повторами в графе, находятся рядом в геноме.

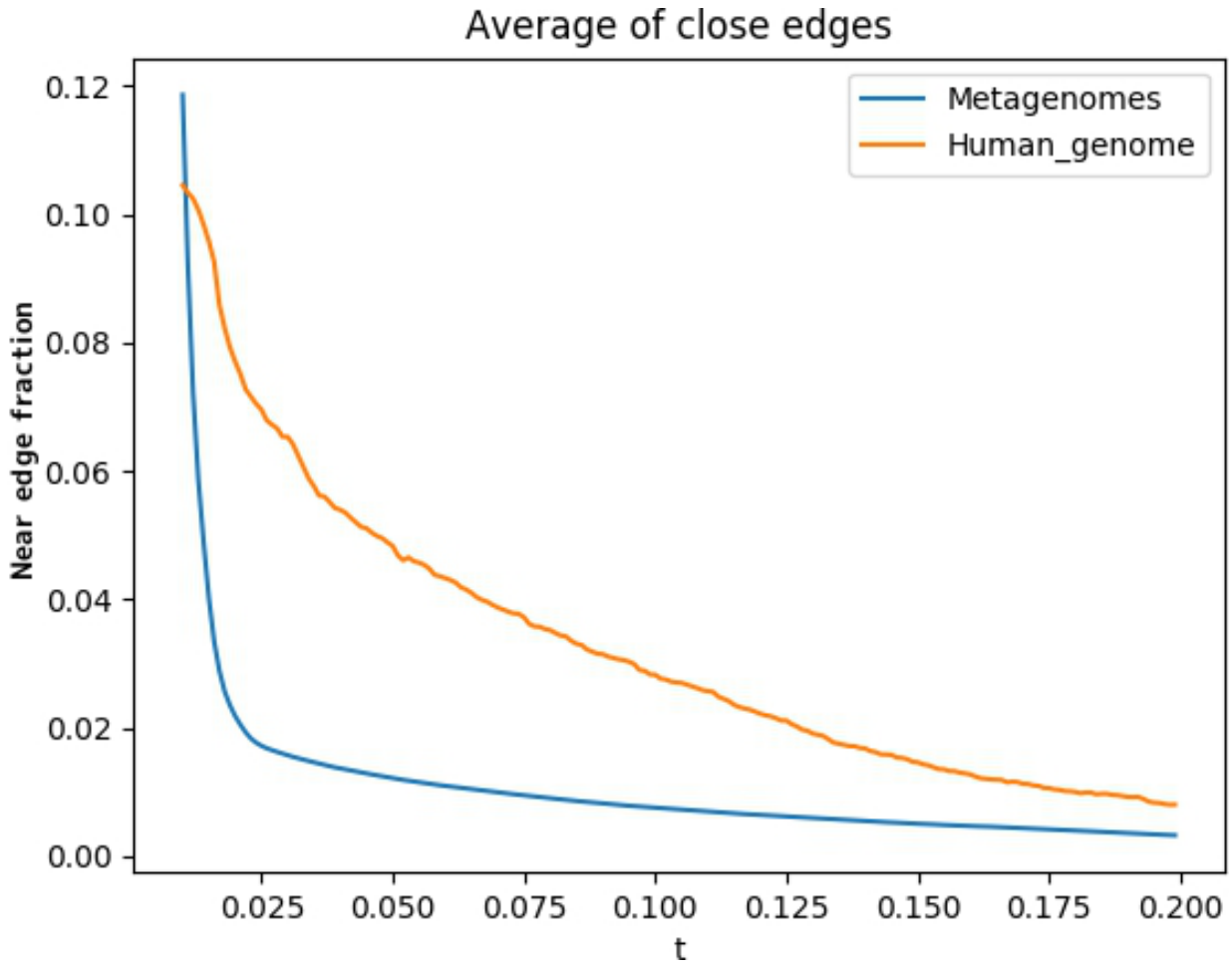


# Близкие рёбра

- $C(e_1, e_2)$ – мера схожести наборов баркодов на длинных рёбрах $e_1$, $e_2$
- $C(e_1, e_2) = \frac{|E_1 \cap E_2|}{\min(|E_1|,|E_2|)}$
- Для некоторого $t$ два ребра $e_1$, $e_2$ считаются близкими, если $C(e_1, e_2) > t$

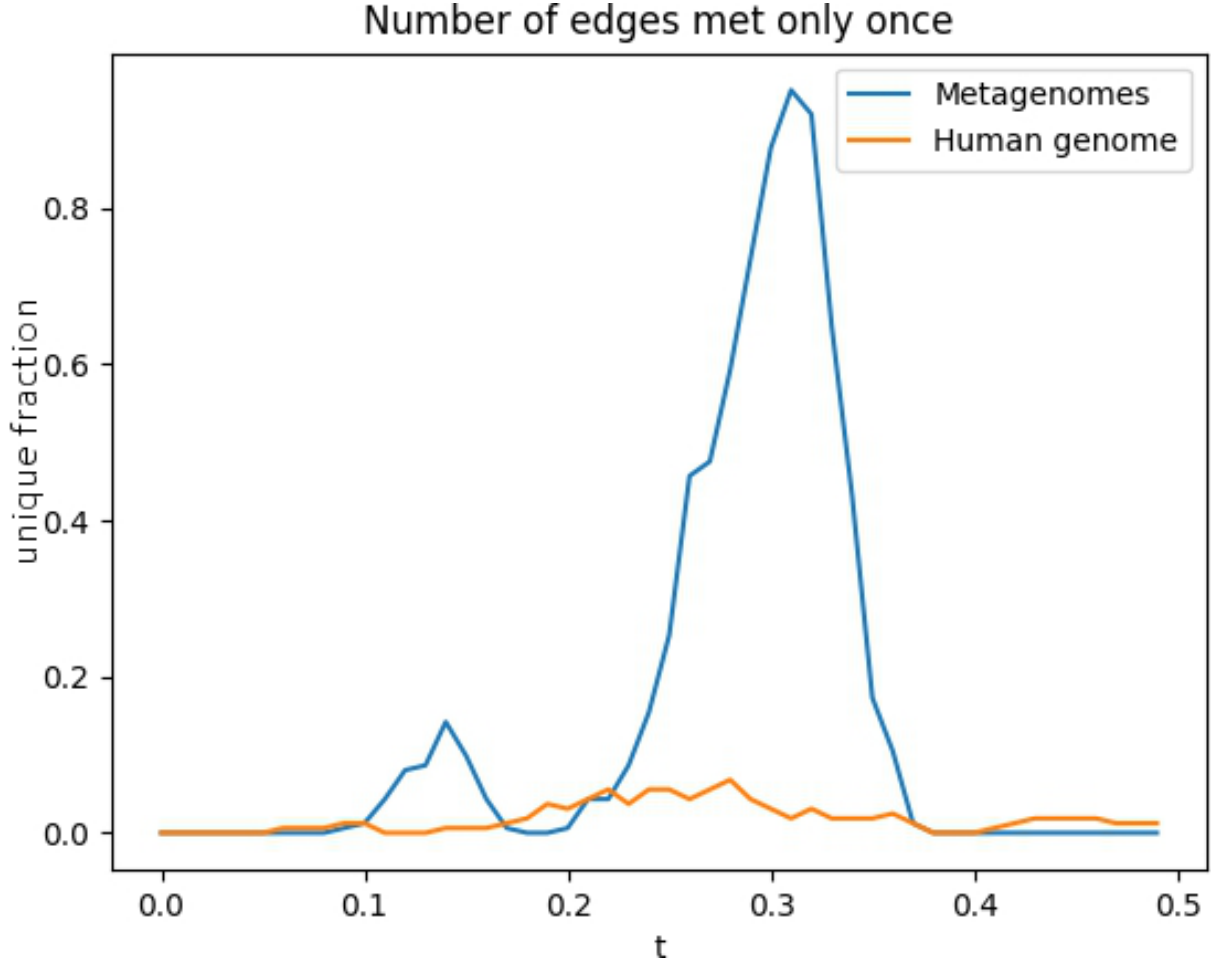
# Среднее количество близких ребер

- $Near(e, t)$ – доля рёбер, близких к $e$ для данного ребра $e$ и порога $t$
- Near edge fraction – среднее $Near(e, t)$ по всем рёбрам



# Количество ребер с однозначным продолжением

- Unique fraction – доля рёбер с единственным близким ребром для данного порога $t$



# Вывод

- Среднее количество близких ребер в человеческом геноме больше, в следствие чего требуется разработка дополнительных методов упорядочивания длинных рёбер в геномах млекопитающих.

# Результаты

- Были изучены существующие методы сборки метагеномов с помощью облаков ридов
- Исследованы недостатки стратегии разрешения повторов применительно к геномам млекопитающих

Спасибо за внимание!