Исследование хромосомных перестроек с помощью 3D структуры ДНК

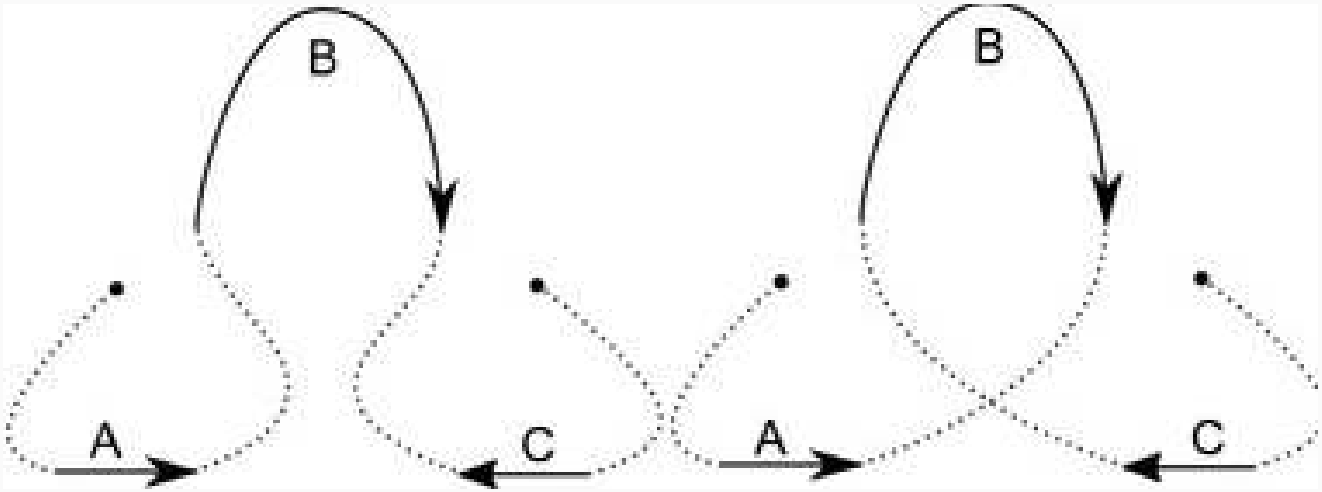
Елена Картышева, Дмитрий Орехов

Институт биоинформатики
Кураторы: Никита Алексеев, к. ф.-м. н., ИТМО; Алексей Сергушичев, к. т. н., ИТМО

Зачем?

- Перестройки могут быть причиной рака, поэтому умение находить перестройки генома пациента может помочь в диагностике рака на ранних стадиях.
- Зная последовательность перестроек, можем строить филогенетические деревья.

Double cut and join



Можем рассматривать хромосомные перестройки, как DCJ операции. Пример DCJ $\{A_h, B_t\}, \{B_h, C_h\} \longrightarrow \{A_h, B_h\}, \{B_t, C_h\}$ которая показывает инверсию блока B.

DCJ сценарий

Раскраска нужна, чтобы определить стоимость DCJ операции.

- DCJ между парами смежности одного цвета $\Longrightarrow cost = 0$
- DCJ между парами смежности разного цвета $\Longrightarrow cost = 1$

Цена последовательности DCJ операций, **DCJ сценарий**, это сумма цен всех этих операций. (MLPS) Minimum Local Parsimonious Scenario
INPUT: Набор пар смежности A и B с раскраской A.
OUTPUT: Самый короткий DCJ сценарий, превращающий A в B.
MEASURE: Вес DCJ сценария.

Что такое Hi-C?

Будем использовать Hi-C!

Как можно использовать Hi-C?

Идея

Чаще всего перестройки происходят между близкими в пространстве кусками хромосом $\Longrightarrow$ Можем использовать данные Hi-C как меру "похожести" для блоков смежности.

Problem

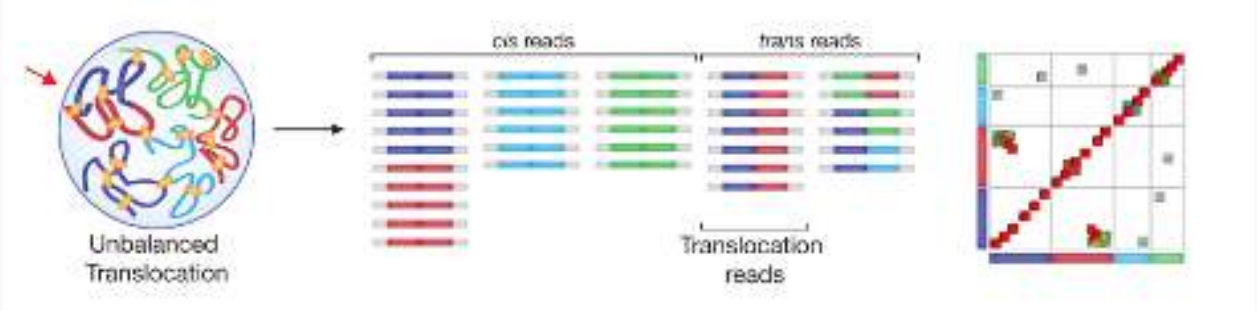
Не очень понятно, как именно связать значение Hi-C карты и расстояние в пространстве.

Solution

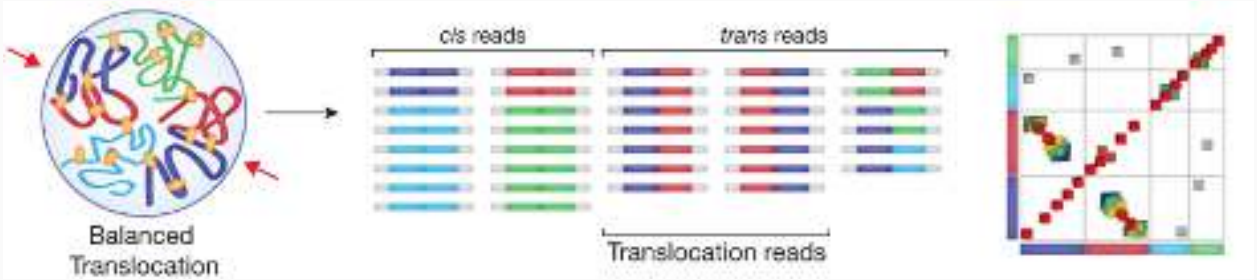
- Попробовать нормализовать Hi-C данные.
- Использовать уже исследованные зависимости между Hi-C и типами перестроек.

Basic patterns

Практически все значения на диагонали, нормальный геном



Маленький блок вне диагонали, несбалансированная перестройка.

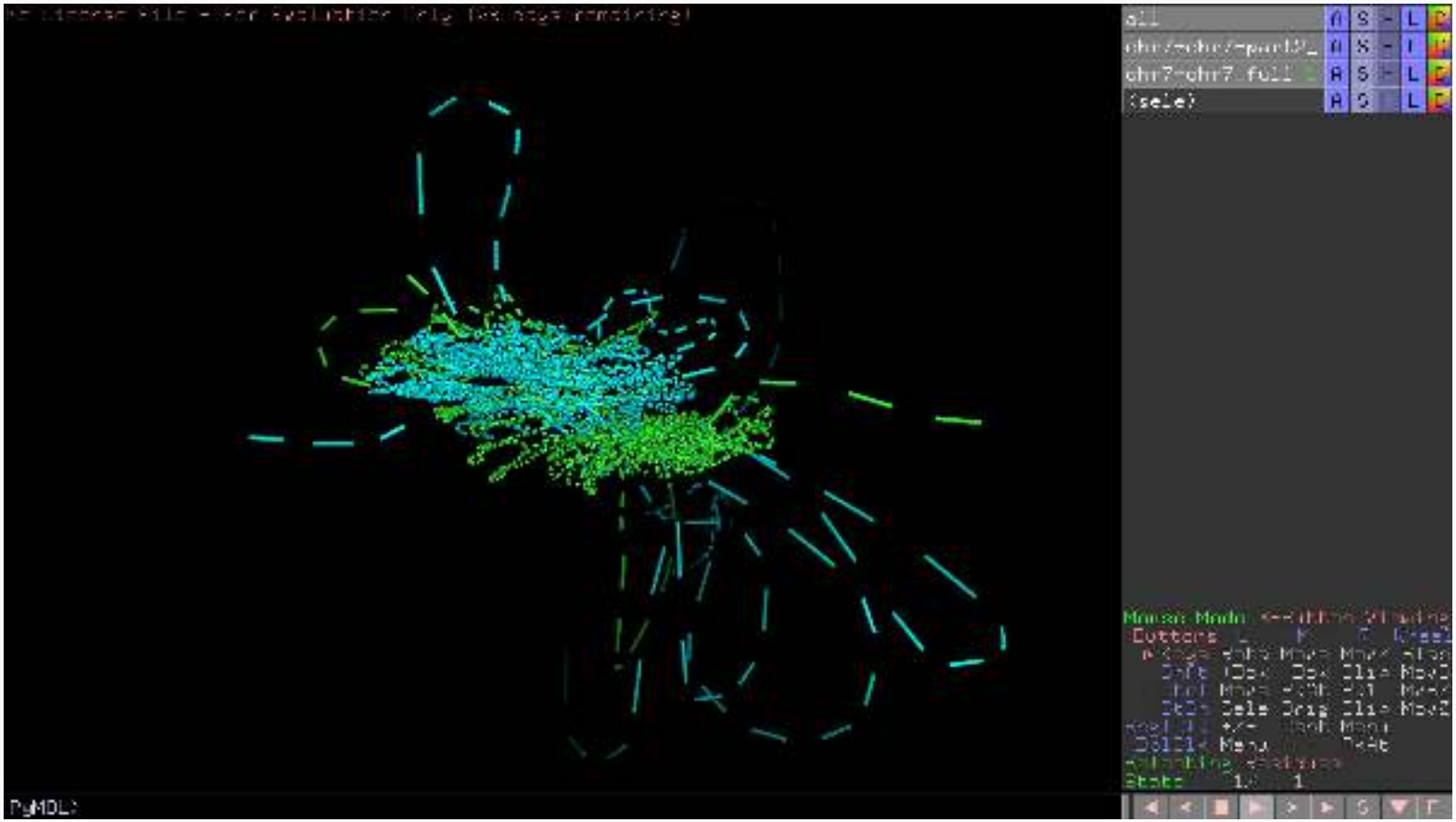


"Бабочка", сбалансированная перестройка.

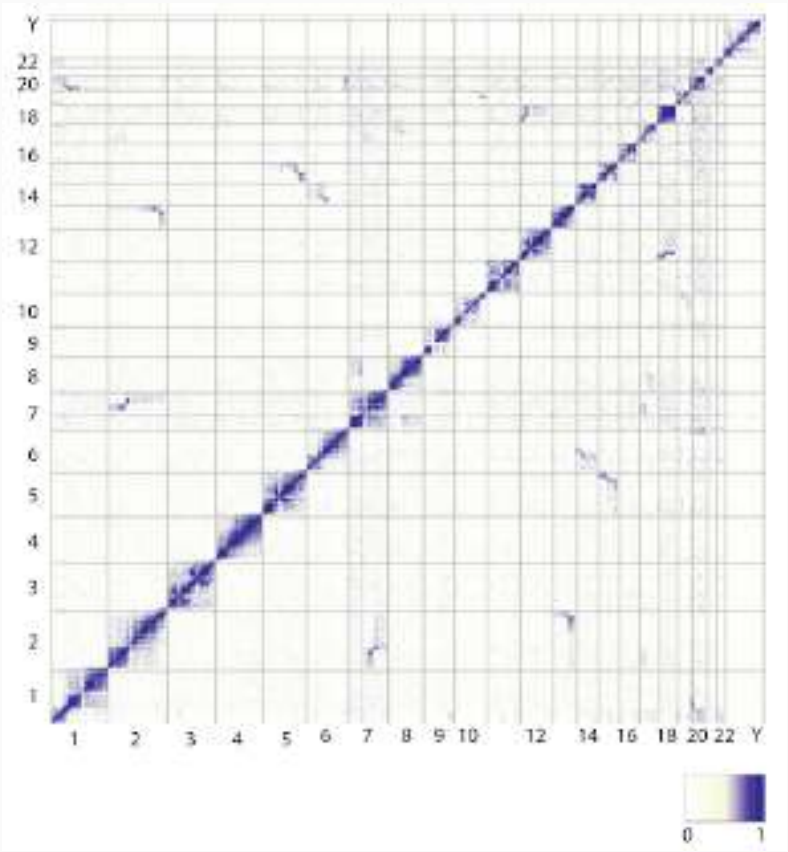

МНЕ КАЖЕТСЯ,
ИЛИ ЭТО ПЕРЕСТРОЙКА
risovach.ru

3d Max

По Hi-C можно строить 3D изображения хромосом.



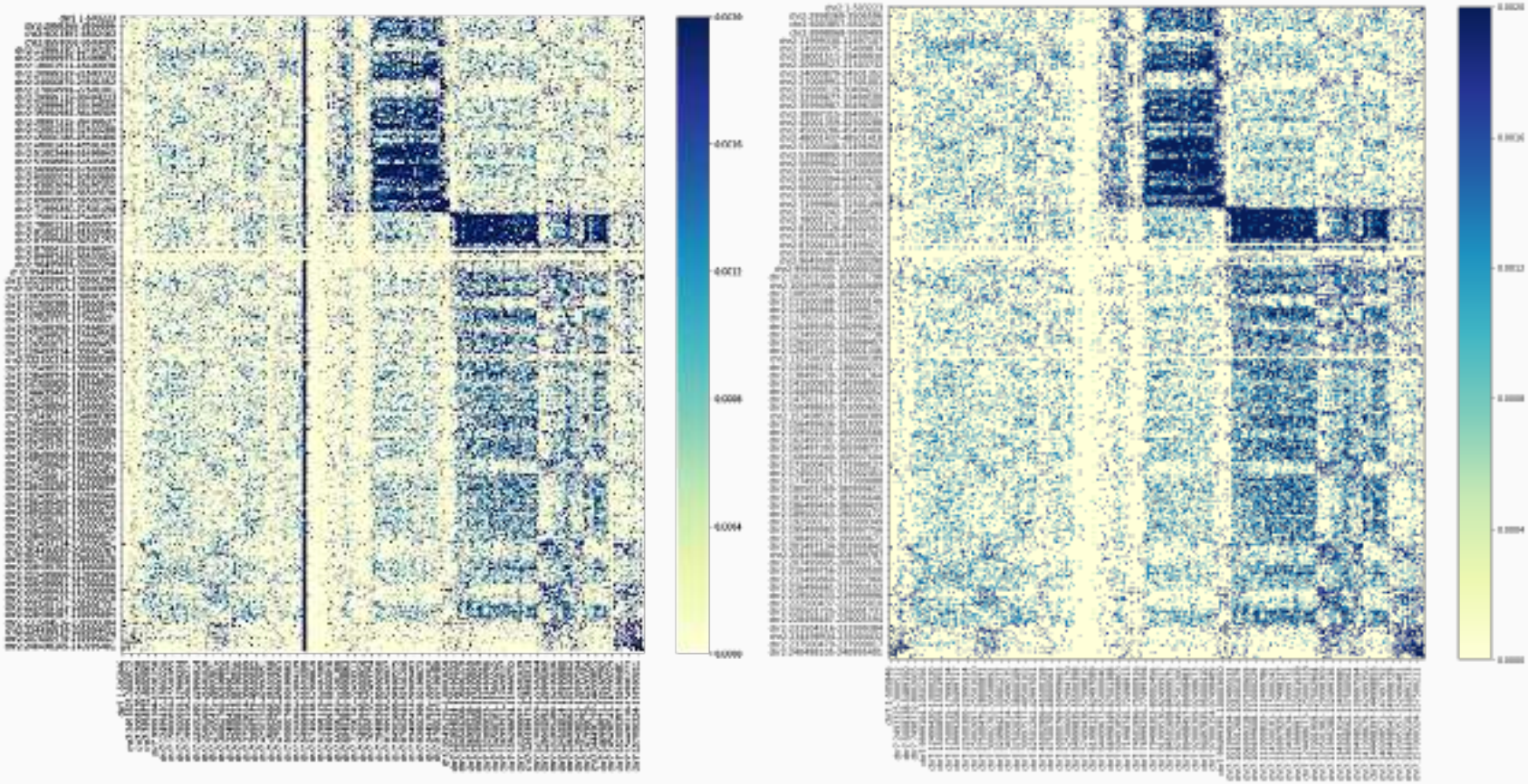
Данные



- Алгоритм тестировался на данных для глиобластомы человека.
- Вспользовались преобразованные Hi-C данные, linkage score plots, полученные Harewood et al [1].
- Производится подсчет интеракций между разбиениями генома на бины по 500kb, значение в каждом бине делится на число сайтов сцепления фермента HindIII.

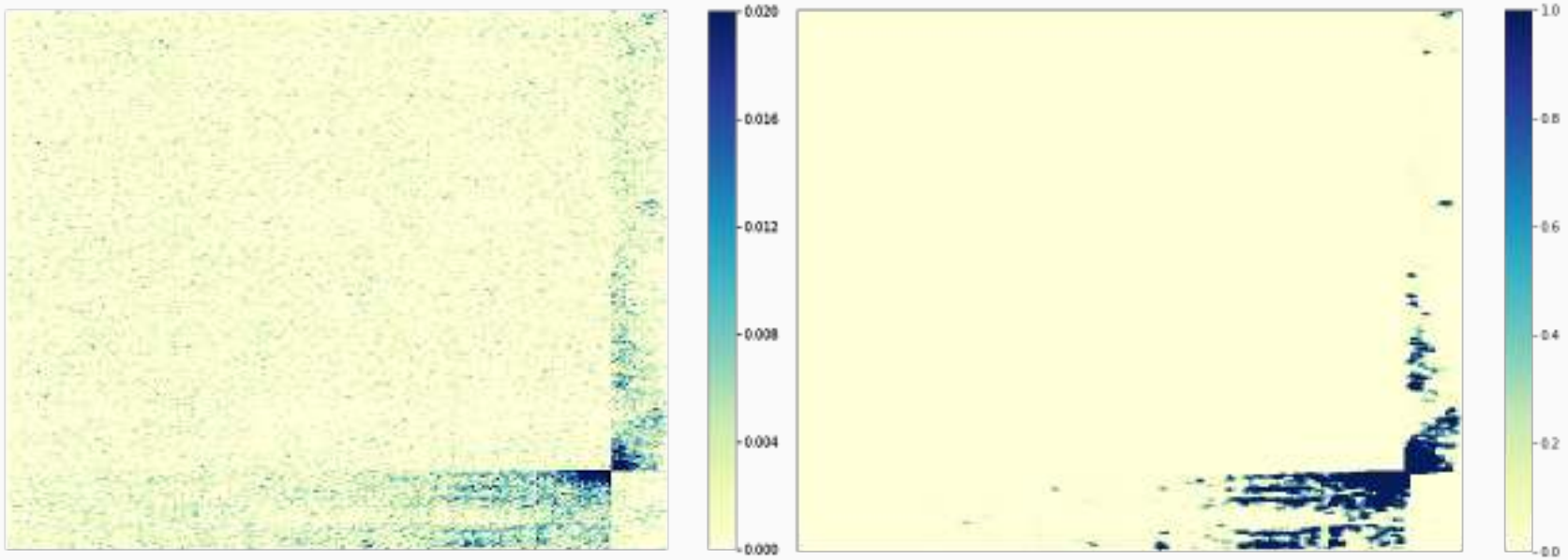
Поиск double minutes

- Простой подсчет суммы значений каждого из рядов/колонок.
- Определение сумм, значение которых больше среднего сумм по рядам/колонкам + трех стандартных отклонений, как аутлаеров.
- Заполнение данных рядов/колонок средним по таблице.



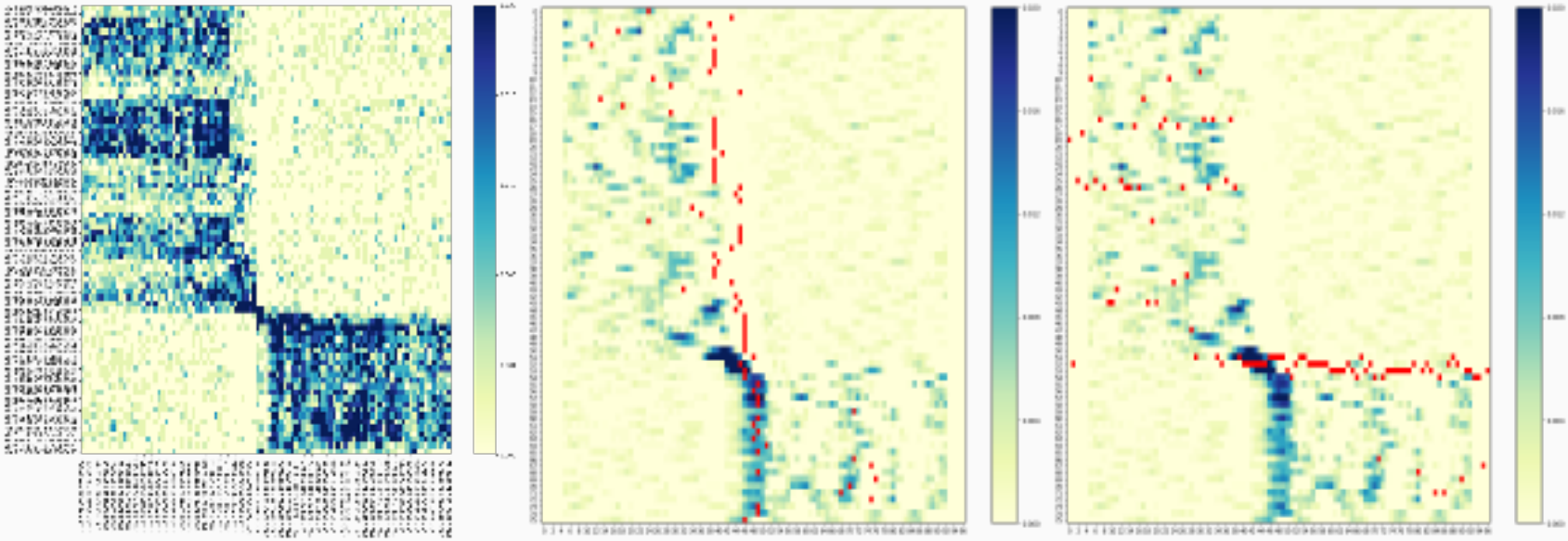
Фильтрация межхромосомных интеракций

- Для каждой пары хромосом получаем матрицу взаимодействий, берем log(x+1) от элементов матрицы.
- После этого используется свертку с окном 3x3 и единичными весами.
- Обучаем GMM с двумя компонентами и делаем hard clustering.
- На выходе бинарная матрицу взаимодействий для каждой пары хромосом.



Локализация перестроек

- Ищем островки в бинарной матрице, используя обход графа в глубину.
- Берем самый крупный островок для пары хромосом, получаем для него индексы границ и рассматриваем данный участок на оригинальной матрице.
- Считаем разницу между двумя скользящими средними, окна смежные, но не пересекаются.



Результаты

Результаты

- Изучили множество (прям реально кучу) подходов к изучению перестроек с помощью Hi-C.
- Получили список перестроек для GB176. Получилось найти все сбалансированные транслокации.
- Попробовали KMeans вместо GMM, фильтрует сильнее, но необходимо задать число кластеров.
- Попробовали свертку с большей длиной ребра окна (4x4, 5x5), свертку 3x3 с весами 0.5 для смежного соседа, 0.25 для соседа через одну клетку. Сильно смазывает, фильтрует не так хорошо, как окно 3x3.

Что дальше?

- Найти какой-то компьютерный подход к детекции и классификации перестроек.
- Научиться, соотнося Hi-C карту с референсом, предсказывать DCJ сценарий.
- Применить уже имеющееся решение для *Drosophila melanogaster* и *Drosophila yakuba* к нахождению перестроек между геномом здорового человека и больного раком.

Спасибо за внимание!

Вопросы?