Misassemblies detection without reference

Анна Лиознова

Руководитель:
Алексей Гуревич

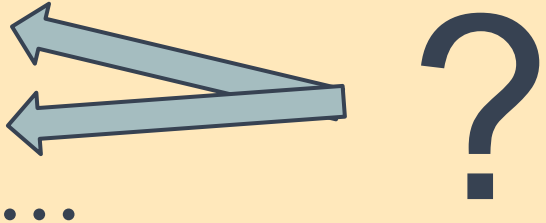
Сборка генома

Как оценить ее качество?

Как правильно?

…AACGTGACATTAGCCCTA…

…AACGTTAGCCGACATCTA…



?

Постановка задачи

Данные:

- сборка генома
- набор парных ридов

Задача:

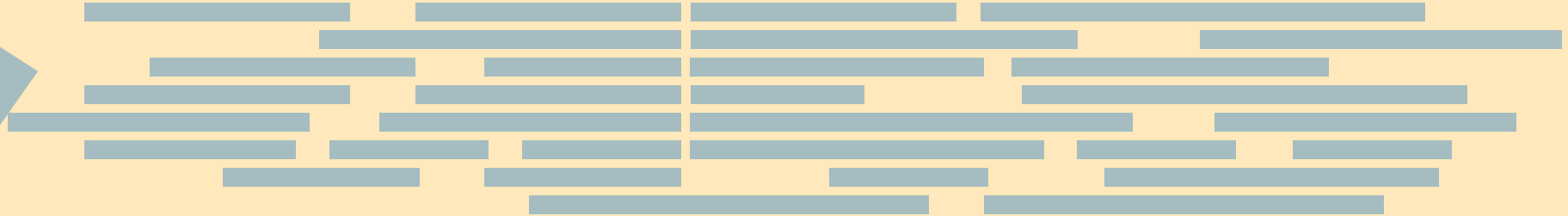
найти ошибки (misassemblies) в сборке

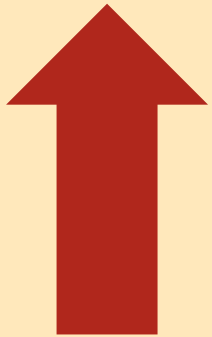
Идеи решения

#1 - начала и концы ридов

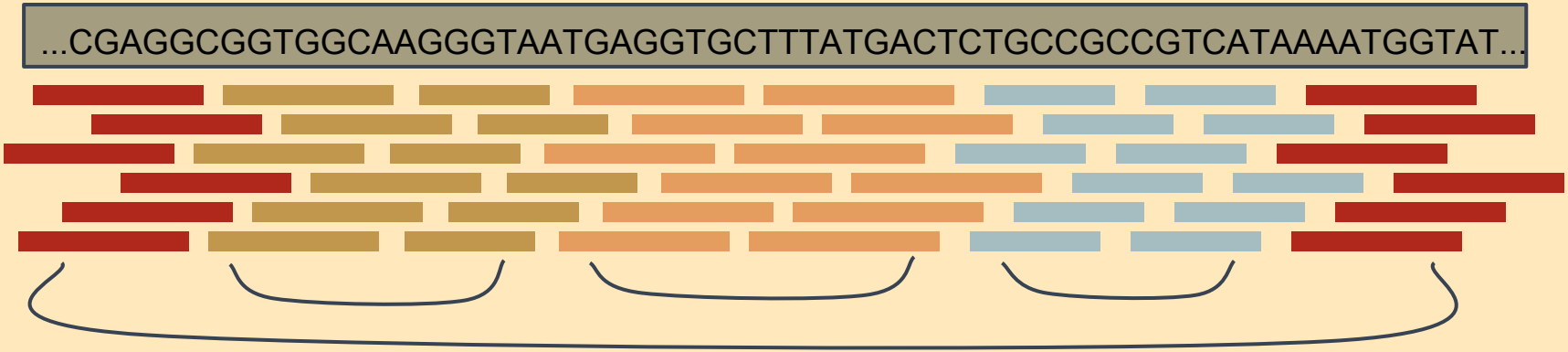
#1 - начала и концы ридов

...CGAGGCGGTGGCAAGGGTAATGAGGTGCTTTATGACTCTGCCGCCGTCATAAAATGGTAT...

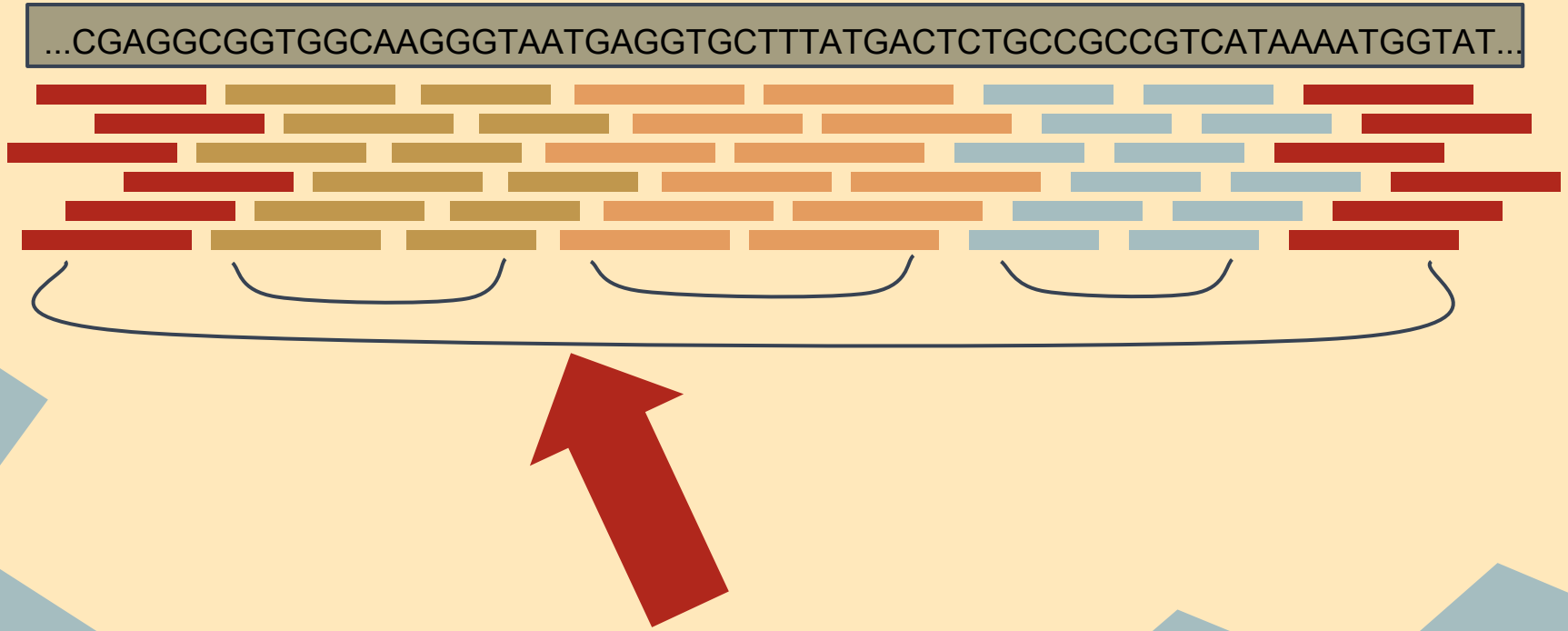




#2 - парные риды

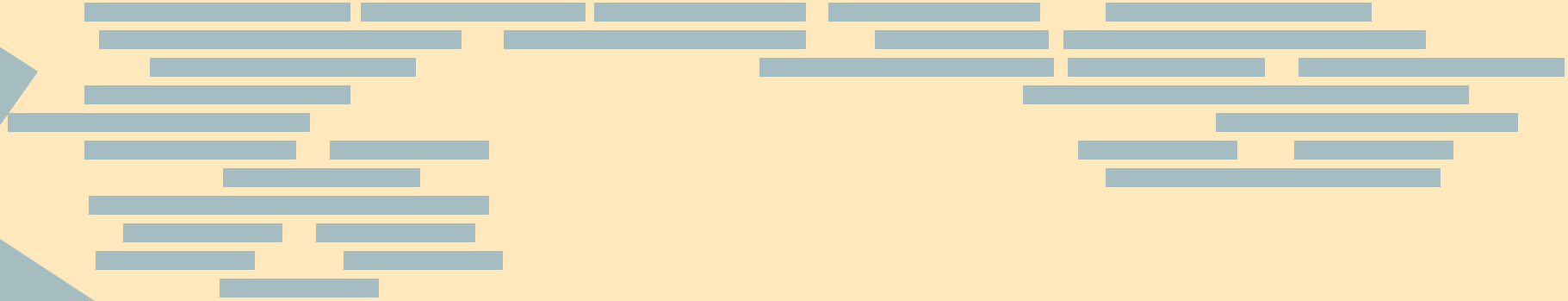


#2 - парные риды

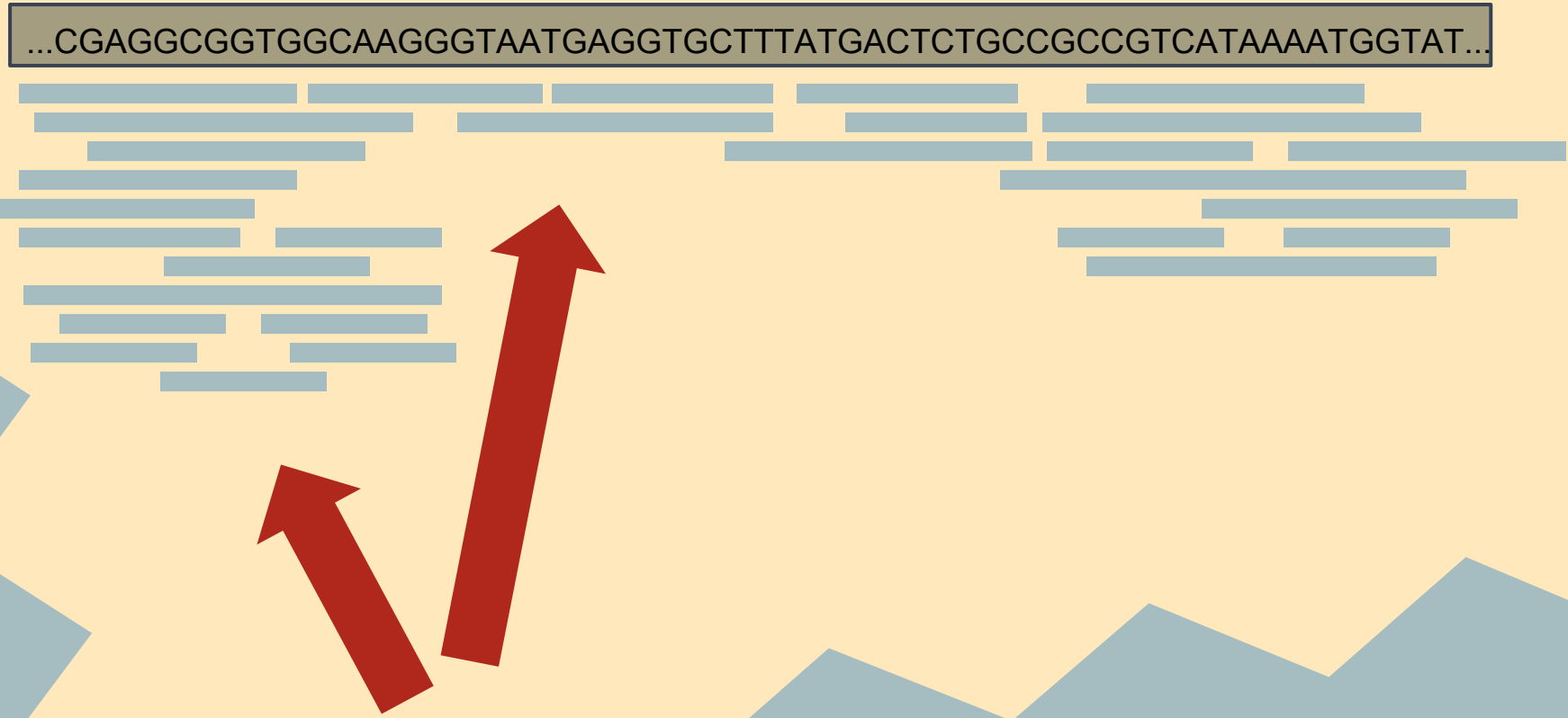


#3 - покрытие

...CGAGGCGGTGGCAAGGGTAATGAGGTGCTTTATGACTCTGCCGCCGTCATAAAATGGTAT...



#3 - покрытие


...CGAGGCGGTGGCAAGGGTAATGAGGTGCTTTATGACTCTGCCGCCGTCATAAAATGGTAT...

Разработанная утилита

- прикладывает риды к сборке
- находит подозрительные места в покрытии сборки ридами
- формирует отчет о возможных ошибках сборки

(утилита станет частью QUAST)

Что сделано

- Познакомилась с QUAST
- Написан скрипт на Python
- Выравнивание bwa, bowtie
- Тестирование на геноме *Enterobacteria phage lambda*
- Проверка на геноме *E. coli*
  - сборка EULER-SR
  - сборка SPAdes
  - референс

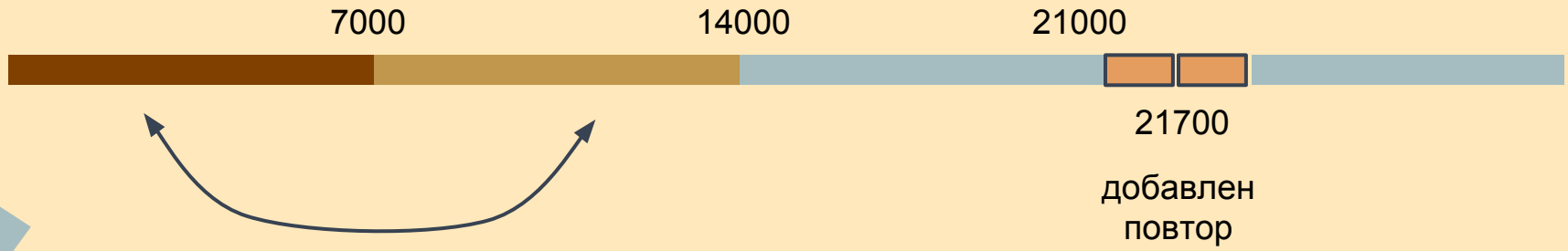
Тестовые данные

Организм: Enterobacteria phage lambda

Длина референса: 48,5 kbp

Число парных ридов: 10000

Тестовый запуск



Пример отчета

Positions where large number of reads ends and begins.

Position: 7001 number: 54

Position: 12814 number: 56

Positions: 14000-14001 number: 64

Пример отчета

Fragments with large TLEN.

Intervals: 1-200 :: 13755-13936 number: 35
Intervals: 6999-7164 :: 12814-12973 number: 20
Intervals: 6756-6952 :: 13999-14176 number: 27
Intervals: 12555-12755 :: 38078-38266 number: 22
Intervals: 36689-36877 :: 48976-49104 number: 15

Пример отчета

Fragments with low or high coverage. Average coverage is 42.3.

Start position: 5805 length: 61 average coverage: 62.3
Start position: 19087 length: 61 average coverage: 63.9
Start position: 20153 length: 73 average coverage: 61.0
Start position: 20573 length: 147 average coverage: 67.8
Start position: 20742 length: 93 average coverage: 63.5
Start position: 21593 length: 210 average coverage: 4.9
Start position: 28734 length: 71 average coverage: 63.2
Start position: 31475 length: 69 average coverage: 63.6
Start position: 46489 length: 63 average coverage: 63.6

Реальные данные

Организм: *E. coli*

Длина референса: 4.64 Mbp

Количество парных ридов: 28 428 648

Объем файла с выравниванием: 8,5 Гб

Время выравнивания: ~ 30 мин

Время работы утилиты: ~ 7 мин

Требуемая память: < 2 Гб

Результат (*E. coli*)

	reference	EULER	SPAdes
begin & end	1	50	12
TLEN	7	11	2
coverage	4	20	22
QUAST mis.	0 + 0 (local)	9 + 31 (local)	0 + 5 (local)

Дальшейшие шаги

- автоматизировать подбор констант
- сравнивать результаты по трем рассмотренным метрикам
- улучшить оценки средних значений параметров
- сравнить работу утилиты со сторонними программами
- опробовать утилиту на других данных


СПАСИБО ЗА ВНИМАНИЕ!
risovach.ru