Использование неравномерной глубины покрытия для разрешения повторов при сборке бактериальных геномов в одноклеточном секвенировании

Ксения Крашенинникова
научный руководитель Дмитрий Антипов

Санкт-Петербургский академический университет

krasheninnikova@gmail.com

3 июня 2013 г.

Определения

Геном - строка над четырехбуквенном алфавитом $\{A, C, G, T\}$.
Секвенирование генома - процесс чтения подпоследовательностей генома.
Рид - подстрока генома, полученная в результате секвенирования.

Данные секвенирования и сборка генома

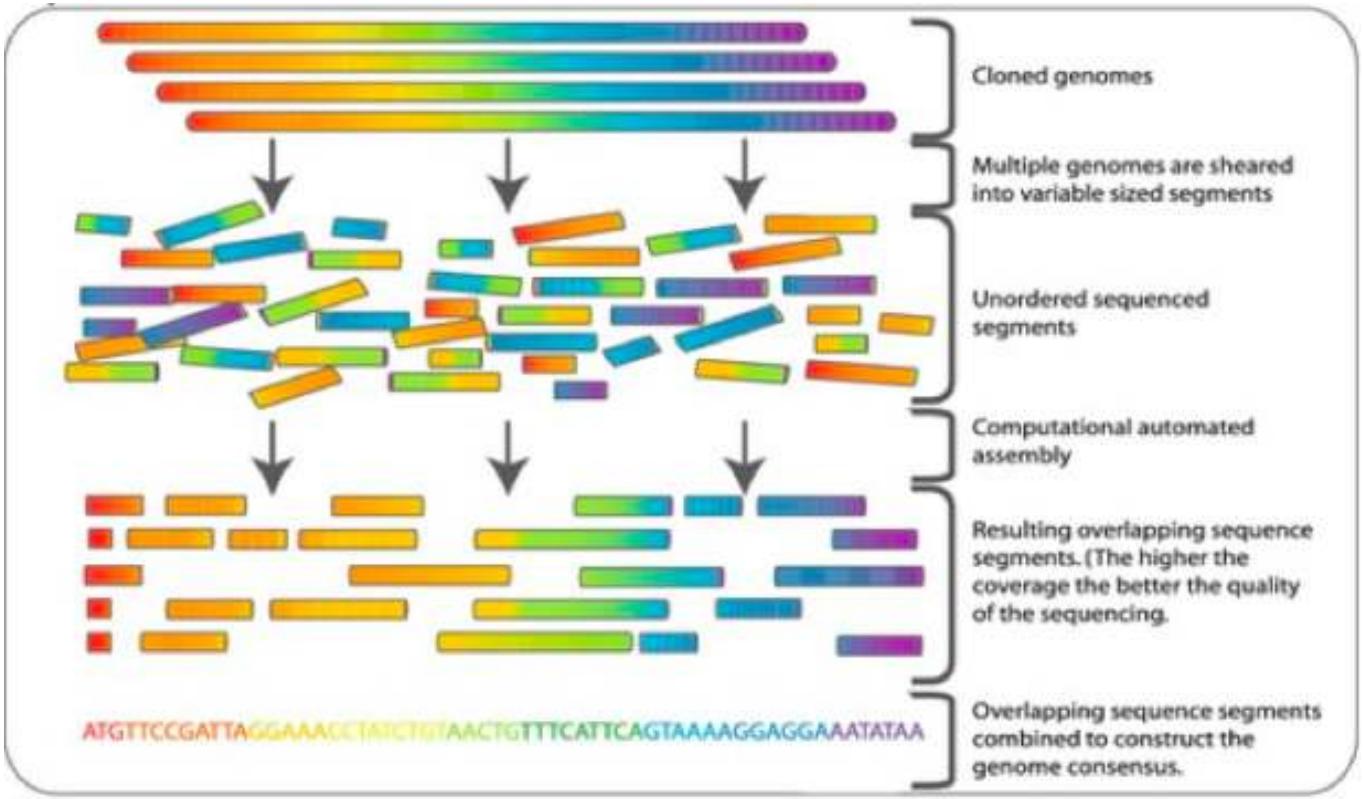


Рис. 1 : Задача сборки генома

Одноклеточное секвенирование. Артефакты MDA

Особое свойство данных MDA - неравномерное покрытие генома

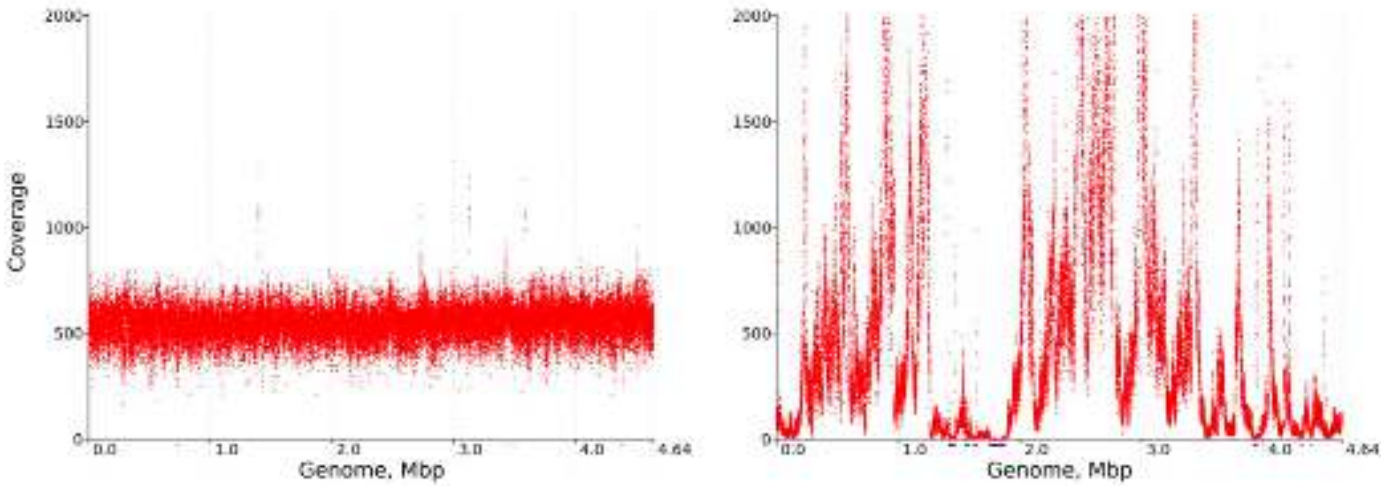
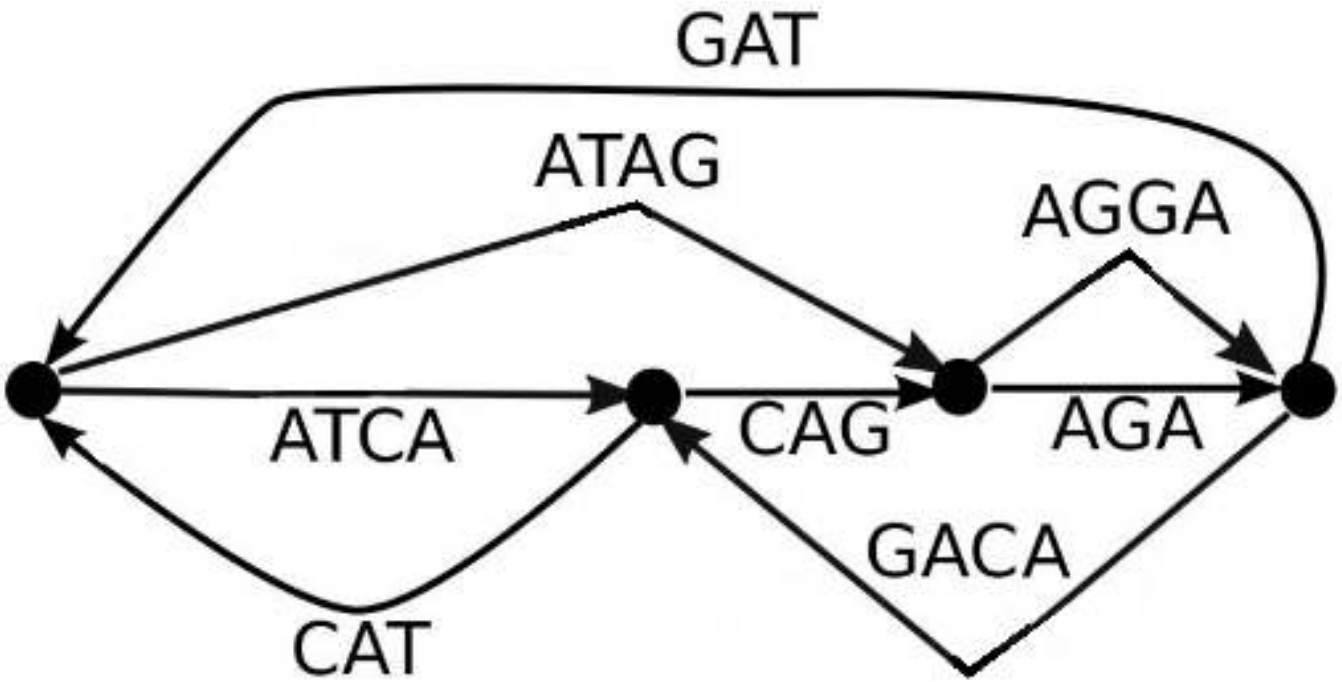


Рис. 2 : Сравнение данных традиционного (regular) и одноклеточного (single-cell) секвенирования для E. coli Illumina GA IIx парного секвенирования, длина ридов 100 bp, покрытие генома 600x

Сжатый граф де Брюйна



Этапы работы SPAdes

1. Коррекция ошибок
2. Построение сжатого графа
3. Упрощение сжатого графа
4. **Разрешение повторов на основе парной информации (distance estimation, repeat resolution with paired reads)**

Проблема разрешения повторов

Повтор - это ребро или группа ребер в сжатом графе де Брюйна, которая соответствует двум или более участкам генома.

Покрытие ребра вычисляется как средняя кратность k-меров, образующих последовательность ребра.

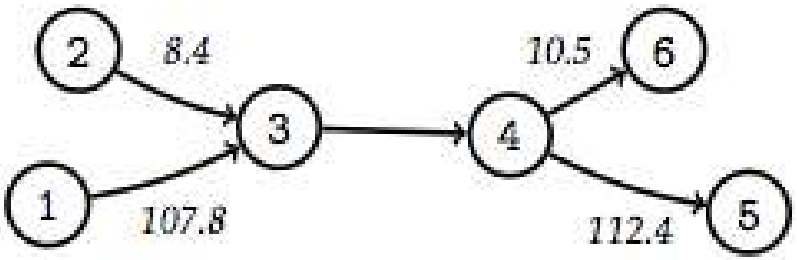


Рис. 3 : Простой повтор в сжатом графе де Брюйна
Ребра помечены соответствующими значениями покрытия

Подходы к разрешению повторов

Для разрешения геномных повторов используются следующие методы:

1. *Парные риды* - риды, расположенные в геноме на фиксированном расстоянии друг от друга. Прикладывая риды к ребрам сжатого графа де Брюйна, можно получить оценку на геномное расстояние между определенными парами ребер.
2. Сочетание длинных и коротких ридов.
3. **Неравномерное покрытие ридами генома при секвенировании**

Идея

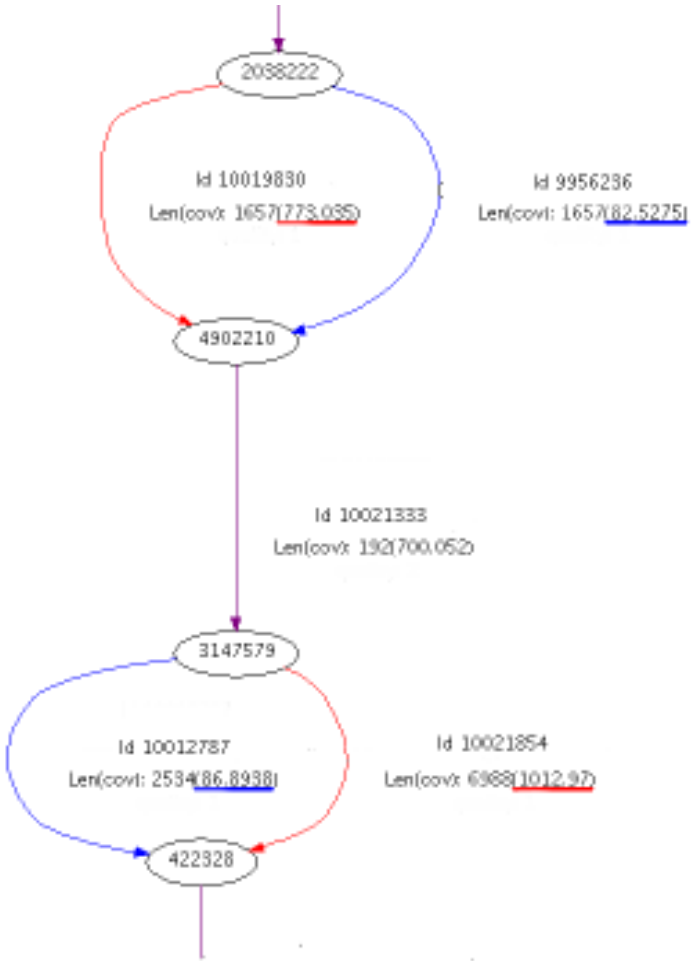


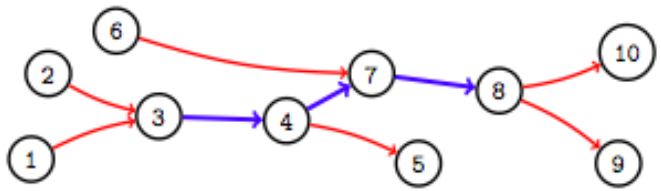
Рис. 4 : Фрагмент сжатого графа де Брюйна, который соответствует геномному повтору

Нахождение повторов в сжатом графе [1]

Разделим множество ребер в сжатом графе де Брюйна на два множества: *одиночные (уникальные)* ребра и группы ребер, образующих *повторные компоненты*.

Определение повторных компонент в сжатом графе с помощью информации о топологии:

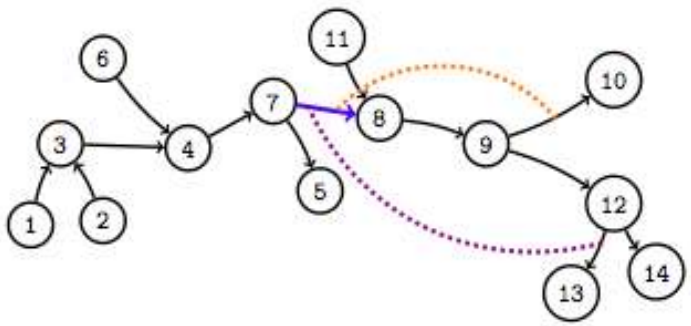
- Ребро $u = (u_1, u_2)$ назовем *уникальным*, если
  - $outdeg(u_1) > 1$ или $indeg(u_1) = 0$.
  - $indeg(u_2) > 1$ или $outdeg(u_2) = 0$.
- Если ребро не уникальное, то оно входит в повтор



Нахождение повторов в сжатом графе [2]

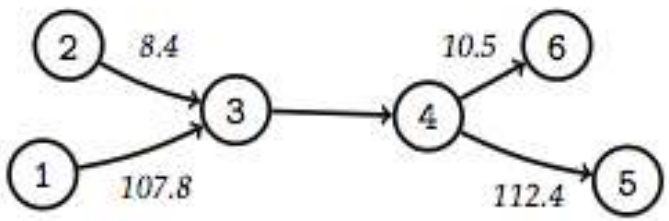
Определение повторных компонент в сжатом графе с помощью парной информации

- Рассмотрим две пары ридов (a_1, a_2) и (b_1, b_2) и ориентированные ребра u, v_1, и v_2 в сжатом графе де Брюйна.
- Пусть
  - риды a_1 и a_2 отображаются на ребра u и v_1 соответственно.
  - риды b_1 и b_2 отображаются на ребра u и v_2 соответственно.
- Ребро u в сжатом графе назвается *повторным*, если не существует достаточно короткого пути, который соединяет ребра v_1 и v_2.



Разрешение повторов

- Отсортировать оба множества ребер $\{in_1, ..., in_k\}$ и $\{out_1, ..., out_n\}$ по убыванию значений покрытия.
- Не рассматривать случаи, для которых $n \neq k$.
- Ввести отсечки, ограничивающие отношения:
  - $\forall i \in \{1, ..., n-1\} \; |1 - \frac{in_i}{in_{i+1}}| \gg 0$
  - $\forall i \in \{1, ..., n\} \; |1 - \frac{in_i}{out_i}| \to 0$



Разрешение тандемных повторов

Тандемный повтор - это паттерн в последовательности ДНК, состоящий из нескольких повторяющихся нуклеотидов, так что повторные участки расположены близко друг от друга.

Например, A-G-G-C-T-A-T-T-C-G-A-T-T-C-G-T-G-T

Тандемные повторы в сжатом графе де Брюйна можно разделить на простые и сложные.

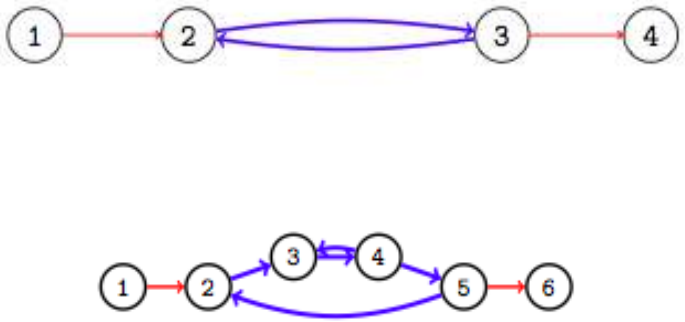


Рис. 5 : Простой и сложный тандемные повторы в сжатом графе де Брюйна

Этапы работы SPAdes

1. Коррекция ошибок
2. Построение сжатого графа
3. Упрощение сжатого графа (tip clipping, bulge removal, chimeric removal)
4. **Разрешение повторов на основе покрытия** (paired-read information)
5. Разрешение повторов на основе парной информации (distance estimation, repeat resolution with paired reads)

Результаты

Таблица 1 : Сравнение сборок датасета *E. coli single-cell*

Assembler	# contigs	NGA50 (bp)	Largest contig (bp)	Genome mapped (%)	MA	Complete genes
SPAdes (single reads)	357	53588	166064	94.191	0	3948
SPAdes + cov-based-rr (single reads)	336	62471	209317	94.276	0	3961
SPAdes +Path-Extend (paired reads)	246	97540	268493	94.859	2	4030
SPAdes + cov-based-rr + Path-Extend (paired reads)	240	108799	268493	94.914	2	4033

Спасибо!