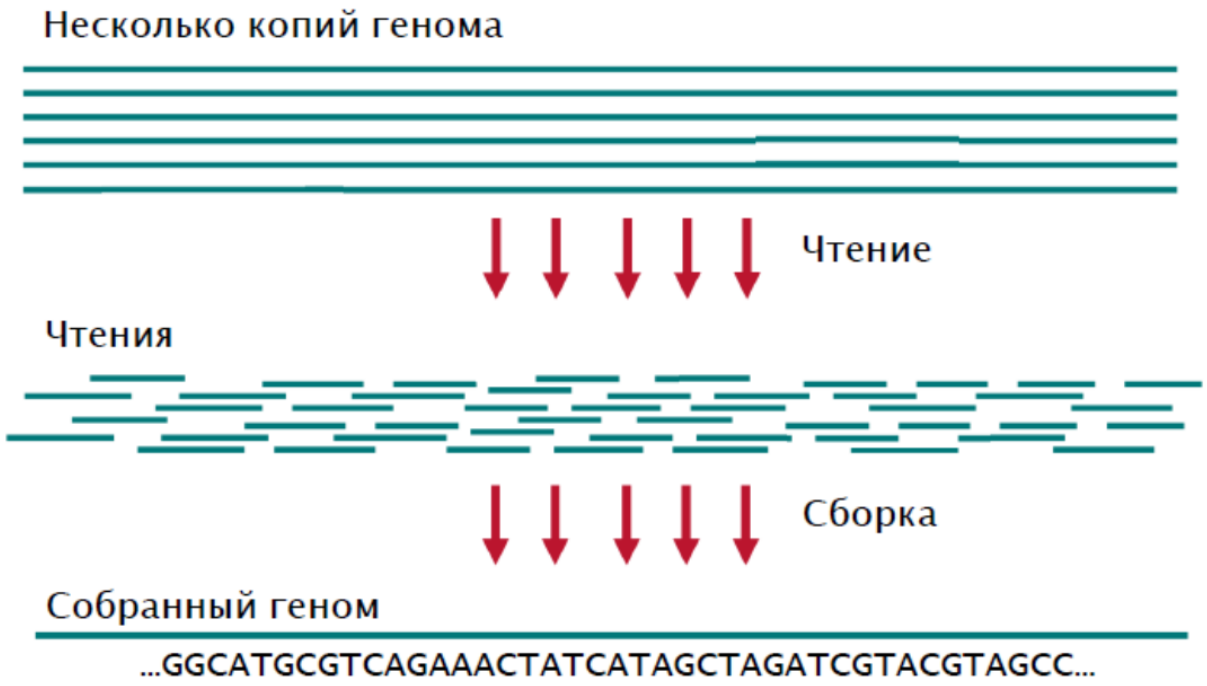
Разработка метрики оценивания сборки генома на основе принципа максимального правдоподобия.

Чтение и сборка генома в общем случае



Распространённые метрики сборки

- Длина кратчайшего контига/скэффолда
- Длина наибольшего контига/скэффолда
- Средняя длина контига/скэффолда
- N50/N90 — наибольшая длина контига такая, что в контигах не меньшей длины содержится 50/90% суммарной длины контигов
- NG50/NG90 — наибольшая длина контига такая, что в контигах не меньшей длины содержится 50/90% суммарной длины генома.

Проблемы

- Существующие на данный момент программы-сборщики работают на приближённых или эвристических алгоритмов

- Собирают не целую геномную последовательность, а последовательность контигов и скэффолдов.

Проблемы

- Наличие редких живых организмов, как следствие невозможность выделения нескольких копий их геномных последовательностей (single-cell)

- Распространенные метрики сборки генома почти не отражают качество сборки генома

Цель работы

- Попытаться вывести формулу на основе принципа максимального правдоподобия (далее ПМП), оценивающую качество сборки геномной последовательности

- Реализовать программу-оценщик, на основе этой формулы, принимающую на вход чтения и сборку генома.

Существующие метрики на основе ПМП

- CGAL
- de Novo
- ALE

CGAL

- $\boldsymbol{l(A,R) = log \prod_{i=1}^{N} p(r_i|A) \approx \sum_{i=1}^{N} log \sum_{j=1}^{M_i} p_F(l_{i,j}) p_S(s_{i,j}) p_E(r_i|a_{i,j}, e_{i,j})}$
  - R — множество чтений
  - A — сборка
  - M_i — количество возможных «соответствий» в сборке чтения r_i
  - $l_{i,j}, s_{i,j}, a_{i,j}$ и $e_{i,j}$ — соответственно длина чтения, позиция чтения, подпоследовательность сборки и ошибки для j-го «соответствия» i-го чтения

CGAL

- В данном методе предполагается, что чтения распределены равномерно по всей длине генома, что является существенным недостатком, поскольку в реальной ситуации чтения распределены чаще всего неравномерно.

de Novo

- A — событие, при котором сборка является изначальной геномной последовательностью
- R — событие, при котором исследуется определённый набор чтений
- Вероятность получить сборку по набору чтений:
  - $\boldsymbol{Pr}[\boldsymbol{A}|\boldsymbol{R}] = \frac{\boldsymbol{Pr}[\boldsymbol{R}|\boldsymbol{A}]\boldsymbol{Pr}[\boldsymbol{A}]}{\boldsymbol{Pr}[\boldsymbol{R}]}$
  - $Pr[A], Pr[R]$ — константы, требуется оценить $Pr[R|A]$.
  - $Pr[R|A] = \prod_{r \in R} Pr[r|A]$

de Novo. Оценка $Pr[r|A]$:

- **Динамическое программирование:**
  - Достаточно точное вычисление вероятностей ошибок для больших наборов чтений
  - Использует $O(n^2)$ памяти, где n — длина сборки
  - Сложный в реализации алгоритм

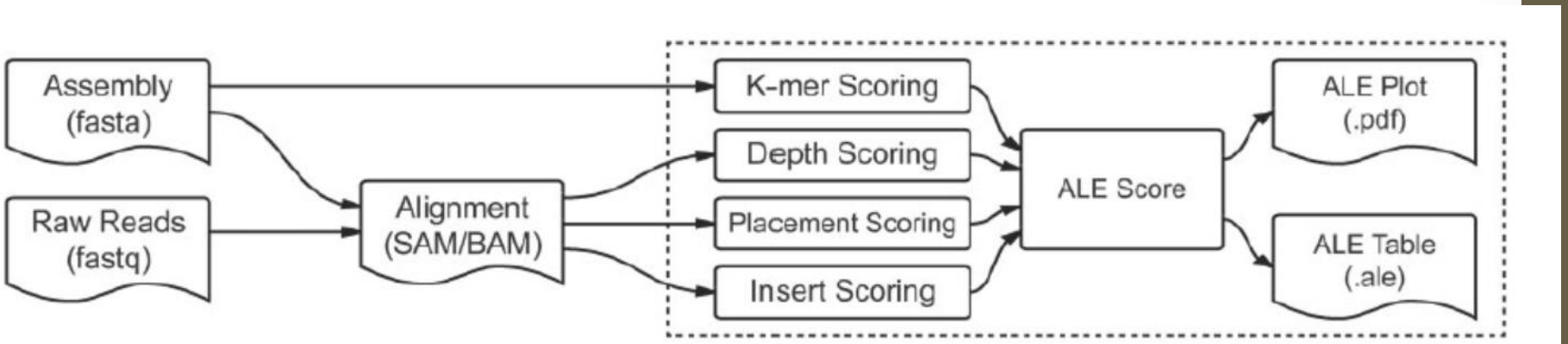
- **Метод «выстраивания чтений» (Aligner method):**
  - *«Выстраивает» чтения по всей длине сборки (SAM файл)*
  - *Считает множества соответствий для каждого чтения*
  - *На основе этих множеств считает требуемые вероятности.*
  - Эффективный по памяти
  - Простой в реализации

de Novo

- На данной нет ни одной рабочей реализации данного метода, в следствие чего невозможно проверить его работоспособность на практике

ALE

- Схема работы



ALE

- Схож по концепции с de Novo
- $Pr[R|A]$:
  - $\boldsymbol{Pr[R|A] = P_{placement}[R|A]\ P_{insert}[R|A]\ P_{depth}[R|A]}$
  - $P_{placement}[R|A]$ — оценивает, насколько хорошо чтения совпадают со сборкой
  - $P_{insert}[R|A]$ — оценивает, насколько хорошо априорными оценки расстояния между парными чтениями (insert length) совпадают с получившимися в результате сборки
  - $P_{depth}[R|A]$ — оценивает, насколько априорная глубина в каждой позиции совпадает с получившейся в результате сборки на основе GC-контента

ALE

- $P_{depth} = P_{depth}[R|A]$ оценивается на основе GC-контента. Такая оценка весьма специфична и не рассматривает общих случаев.

- В связи с этим в данной работе было предложено считать P_{depth} как разность распределений получившейся глубины (в результате сборки) и глубины из секвенатора.

ALE

- Разность априорного и апостериорного распределения глубин предлагается считать по следующей формуле:
- $\boldsymbol{P_{depth} = 1 - \mathrm{JSD}(S_{seq}||S_{real})}$
  - $\mathrm{JSD}(S_{seq}||S_{real}) = \frac{1}{2}D_{KL}(S_{seq}||M) + \frac{1}{2}D_{KL}(S_{real}||M)$, где $M = \frac{1}{2}(S_{seq} + S_{real})$
  - $D_{KL}(S||P) = \sum_i \log_2\left(\frac{S(i)}{P(i)}\right)S(i)$
  - $0 < \mathrm{JSD}(S_{seq}||S_{real}) < 1$, следовательно $P_{depth} \in [0; 1]$

Текущие результаты и перспективы

- Составлена формула на основе ALE
- Разрабатывается программа на R, рассчитывающая P_{depth}
- Планируется встроить эту программу в текущую программную реализацию ALE и сравнить результаты со старой версией ALE