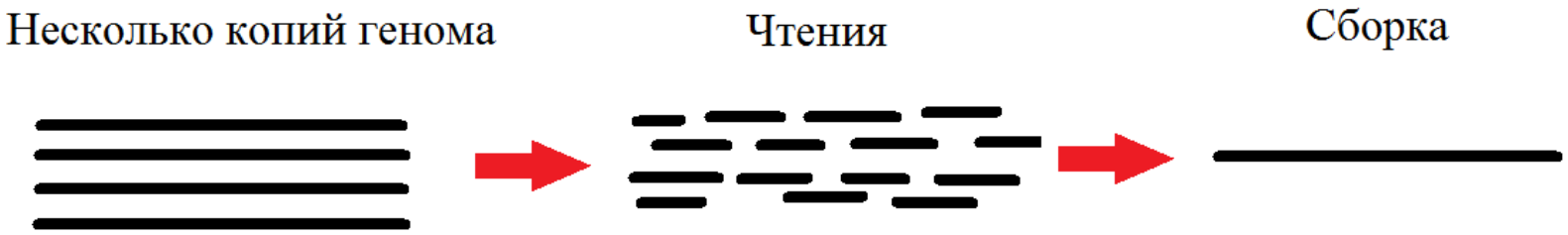
Разработка метода оценки качества сборки генома на основе принципа максимального правдоподобия

Муравьёв Сергей
Научный руководитель: к.т.н., доцент каф. КТ
Царев Ф. Н.

Университет ИТМО

19 июня 2014

Чтение и сборка генома в общем случае



Распространённые метрики сборки

- Длина кратчайшего/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.

Проблема

- Распространенные метрики оценки сборки генома почти не отражают качество сборки генома

Цель работы

- Научиться оценивать качество сборки генома

Задачи

- Разработать метод, оценивающий качество сборки геномной последовательности
- Реализовать программу-оценщик на основе данного метода, принимающую на вход чтения и сборку генома.

Log Likelihood

- $\boldsymbol{ln(P(A|R))}$
- $\boldsymbol{A}$ – событие, при котором сборка является «правильной» геномной последовательностью
- $\boldsymbol{R}$ – событие, при котором исследуется определённый набор чтений

Существующие методы, оценивающие log likelihood

- CGAL
- de Novo
- ALE

ALE

- $P(A|R) = \frac{P(R|A)P(A)}{P(R)}$
- $P(R|A)$:
  - $\boldsymbol{P(R|A) = P_{placement}(R|A)P_{insert}(R|A)P_{depth}(R|A)}$

    - $P_{placement}(R|A)$ – оценивает, насколько хорошо чтения совпадают со сборкой
    - $P_{insert}(R|A)$ – оценивает, насколько хорошо априорные оценки расстояния между парными чтениями (insert length) совпадают с получившимися в результате сборки
    - $P_{depth}(R|A)$ – оценивает, насколько априорная глубина покрытия в каждой позиции совпадает с получившейся в результате сборки

P_{depth}-score

- 100 множеств GC-контента: 0-1, 1-2,…99-100%.

- Глубина распределена по отрицательному биномиальному распределению $NB(r, p)$ для каждого множества.

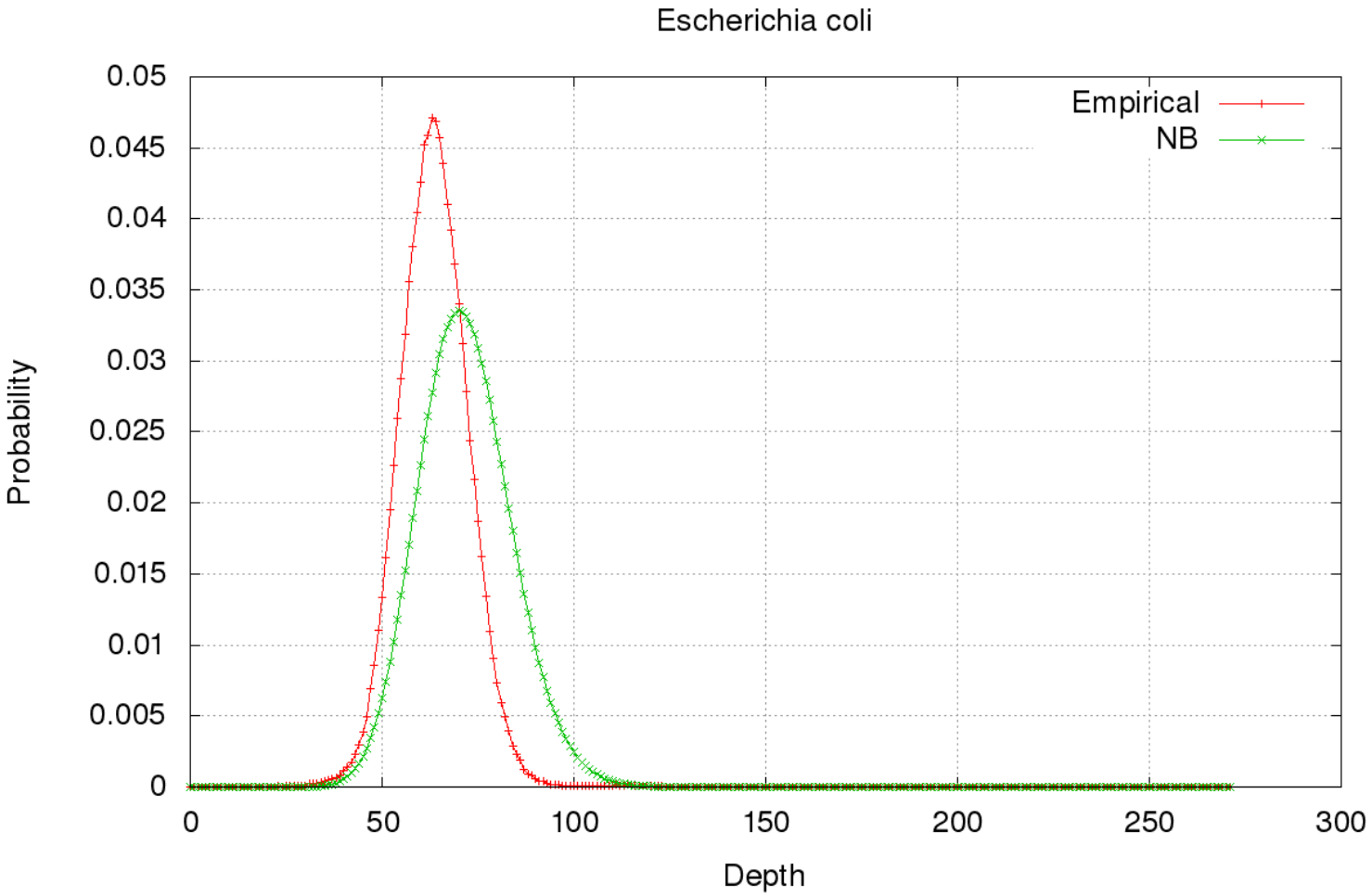
- $p = 0.5, r = \max(10,\ \mu_{depth(X_i)})$, где $\mu_{depth(X_i)}$ — среднее значение глубины для того множества GC-контента, куда попадает значение GC-контента X_i в позиции i

Проблемы ALE

- В ходе работы на данных из базы GAGE удалось установить, что $NB(r, p)$ не соответствует реальному распределению глубины покрытия

- Статистики, применявшиеся при сравнении:
  - χ^2-тест
  - Корреляция Пирсона и Спирмена
  - Приближение на основе принципа максимального правдоподобия

Эмпирическое распределение



Распределение глубин по NB(r,p) и эмпирически для GC-контента, которому соответствует наибольшее число позиций в сборке

Эмпирическое распределение

- Для определения вероятности глубин предлагается использовать эмпирическое распределение для каждого GC-контента

Эмпирическое распределение

- Аналогично ALE разобьём GC контент на 100 множеств: 0-1, 1-2,…99-100%
- Для каждого множества строится эмпирическое распределение глубин
- Чтобы реализовать обнаружение ошибок, позиции с нулевым покрытием не учитываются
- На основе составленных распределений строим таблицу $T[100][maxdepth]$

Тесты на синтетических данных

- Рассмотрена известная сборка *E. coli*

- В сборках искусственно создано несколько тысяч ошибок различных типов:
  - Замена нуклеотидов
  - Вставка/Удаление нуклеотидов
  - Разворот участка генома
  - Копия участка генома в другое место

Сравнительный анализ

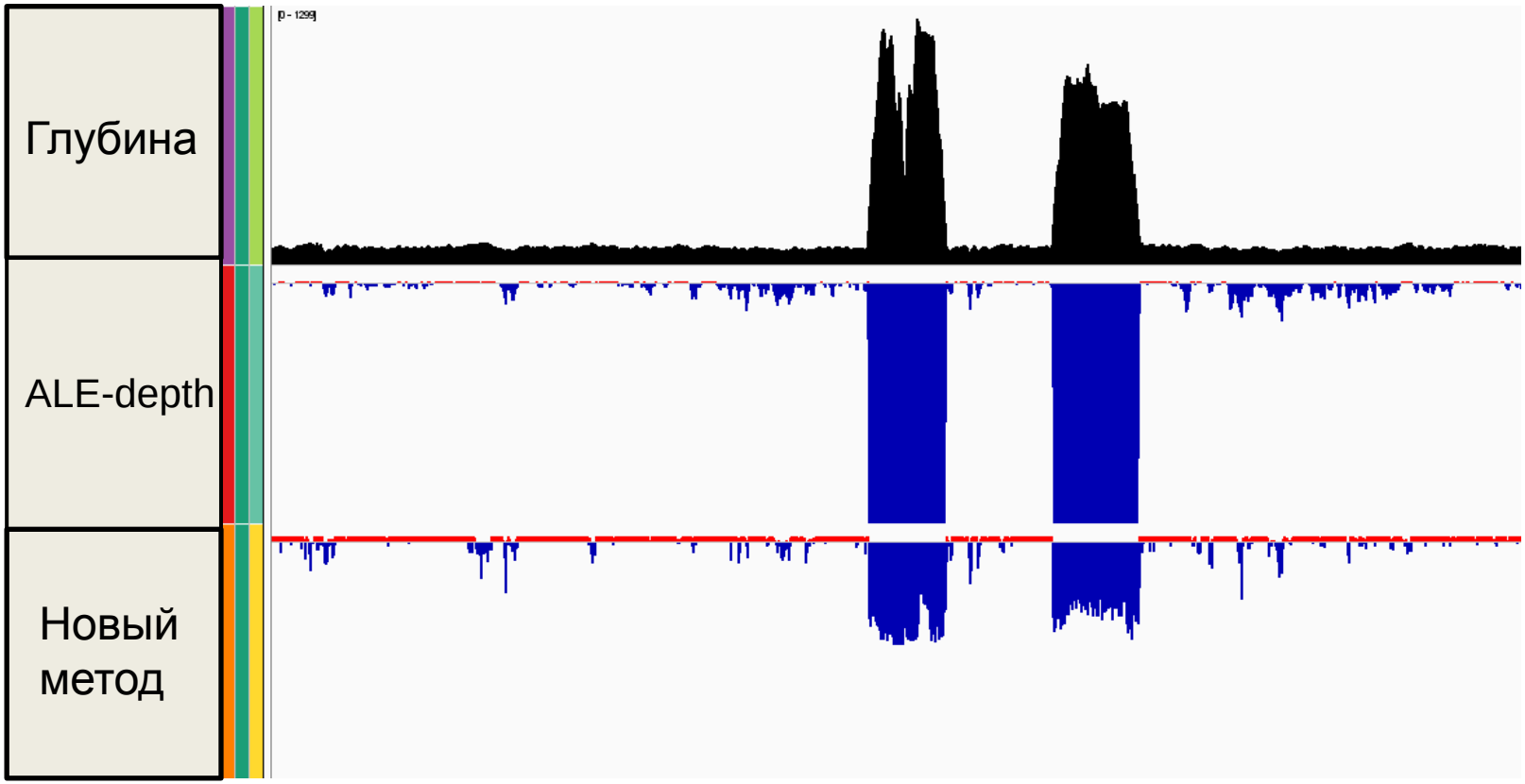
- Анализ сборки с ошибками проводился с помощью *Integrative Genome Viewer* — инструмента визуализации сборок

- Было выявлено, что предложенный метод:
  - Нашёл все внесённые ошибки
  - Обошёл «ложные срабатывания» ALE (возрастание глубины выше среднего на чистой сборке, обусловленное особенностями секвенирования)

Сравнительный анализ

Внесённые ошибки	Результат
10 ошибок замены по 1000 нуклеотидов	Найдены все позиции с ошибками
1000 ошибок вставки по 100 нуклеотидов	Найдены все позиции с ошибками
1000 ошибок удаления по 100 нуклеотидов	Определяется только левая граница ошибочного региона
1000 ошибок копирования участков генома длиной 100 нуклеотидов.	Определяется только правая граница ошибочного региона
1000 ошибок разворота участков генома длиной 100 нуклеотидов.	Найдены все позиции с ошибками

Ложные срабатывания



Сравнение работы предложенного метода и ALE при условии резкого возрастания глубины, обусловленного особенностями секвенирования, при помощи IGV.

Тесты на реальных данных

- Рассмотрены известные сборки организмов, а также сборки популярных сборщиков:
- *S. aureus*

Сборщик	**Достоверно известное количество ошибок замещения и вставки/удал. >5bp**	**ALE**		**Предложенный метод**	
		true neg.	false neg.	true neg.	false neg.
ABySS	19	19	3	19	0
Allpaths-LG	20	18	4	19	0
SGA	10	9	2	10	0
Velvet	42	40	3	41	1

Тесты на реальных данных

- *R. sphaeroides*

Сборщик	Достоверно известное количество ошибок замещения и вставки/удал. >5bp	ALE		Предложенный метод	
		true neg.	false neg.	true neg.	false neg.
ABySS	76	75	2	76	0
Allpaths-LG	49	45	5	48	0
SGA	12	12	1	12	0
Velvet	47	45	4	47	1

Тесты на реальных данных

- *H. sapies Chr. 14*

Сборщик	Достоверно известное количество ошибок замещения и вставки/удал. >5bp	ALE		Предложенный метод	
		true neg.	false neg.	true neg.	false neg.
ABySS	704	691	15	701	10
Allpaths-LG	2760	2584	352	2700	20
SGA	981	913	27	975	0
Velvet	4910	4567	560	4879	107

Результаты

- Разработан метод оценки на основе ALE
- Реализована программа на Python, рассчитывающая эмпирическое распределение глубин для каждой сборки и depth-score соответственно

Спасибо за внимание.

Вопросы?