Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Муравьёв Сергей Борисович

# Разработка метода оценки сборки генома на основе принципа максимального правдоподобия

Научный руководитель: доцент кафедры компьютерных технологий
Ф. Н. Царёв

Санкт-Петербург
2014

# Содержание

# Введение

Задача корректной оценки качества сборки геномной последовательности является важной частью биоинформатики. Задача состоит в том, чтобы по набору чтений и сборке цепи ДНК геномной последовательности определить качество сборки, не имея эталонного шаблона. Основными проблемами в этом процессе являются наличие большого числа ошибок в исходных данных, а также большой объем входных данных, исчисляющийся сотнями гигабайт.

Последние достижения в области секвенирования следующего поколения резко снизили стоимость секвенирования. С развитием сборщиков появилась возможность использовать большие объёмы секвенируемой информации. Благодаря технологии секвенирования методом дробовика стало появляться всё больше геномов различных организмов, от маленьких бактерий до млекопитающих. Несмотря на это, некоторые генетические последовательности получаются напрямую из материи, содержащей в себе несколько организмов, при помощи single-cell секвенирования и метагеномного секвенирования.

При сборке конкретного организма возникают ошибки, обусловленные короткой длиной чтений, шумах в информации, большим объёмом данных и особенностями сборщика. В случае с метагеномной сборкой возникают дополнительные трудности: глубина покрытия чтениями распределена неравномерано, неоднозначность в анализе повторяющихся областей в случае метагеномной сборки усугубляется. Было разработаено несколько инструментов для обнаружения ошибок при неметагеномной сборке. Однако такие средства используют в своём составе эталонную или близкую к эталонной сборку организма, геном которого хотят собрать. В случае отсутствия эталонной сборки подобные инструменты предоставляют косвенные оценки качества сборки генома, такие как N50, глубина покрытия, оценка расстояния между парными чтениями. Такие метрики предоставляют информацию о производительности сборщика, но не обеспечивают внутренних прямых измерений качества сборки.

В данной работе была разработана статистическая модель для оценки качества сборки генома в условиях отсутствия эталонной сборки. Данная модель представляет из себя формулу на основе формулы Байеса, которая вычисляет вероятность того, что сборка правильная, то есть соот-

ветствует эталону.

## Список литературы

1. Scala language. http://www.scala-lang.org/.