

## ОТЗЫВ

председателя диссертационного совета на диссертацию Михеенко Аллы Александровны на тему «Разработка вычислительных методов для анализа и визуализации эукариотических геномных сборок» на соискание ученой степени кандидата физико-математических наук по специальности 03.01.09 – Математическая биология, биоинформатика.

За прошедшие 20 лет было создано несколько десятков различных программ для сборки геномов и транскриптомов с нуля, что связано с появлением технологий секвенирования следующего поколения и, как следствие, снижением цены чтения. Постоянно обновляющиеся протоколы секвенирования и меняющиеся характеристики данных требуют поддержки существующего, а в некоторых случаях и разработки полностью нового программного обеспечения. Одна из самых больших проблем сборок геномов состоит в том, что прочитаны и собраны только эухроматические части генома, тогда как гетерохроматиновые (ГХ) области остаются несобранными. Прогресс в секвенировании не помогает решить задачу, проблема кроется в собственно составе ГХ и отсутствии методов сборки для районов с высоко повторяющимися последовательностями.

Основным компонентом ГХ являются тандемные повторы (ТП) (сателлитная ДНК). ТП представляют собой класс ДНК, который появляется у эукариот, но отсутствует у прокариот. Поля ТП состоят из многократно повторяющихся коротких последовательностей (мономеров). Содержание ТП в геномах высших эукариот может составлять десятки процентов. А.А.Михеенко оценила несобранную часть генома человека в 10%-15 % генома. Т.о. значительная часть генома нуждается во внимании автора диссертации.

В последние годы с открытием транскрипции ТП, отношение к этой части генома стало меняться, хотя долгое время ГХ относили к мусорной (junk) ДНК. Подавляющее большинство ТП до сих пор находится в неаннотированной, «темной» части генома. ТП по первичной последовательности различаются у разных видов вплоть до возможной видоспецифичности. Как правило, разные ТП расположены в центромерном (ЦЕН) и перичентромерном (периЦЕН) районах хромосом. ТП быстро эволюционируют, но сохраняют свои функции в кинетохоре. Несмотря на различия в последовательности ТП разных видов, ТП имеют общие черты – организацию в длинные гомогенные поля и длину мономера, часто соответствующую размеру нуклеосомной ДНК, склонность областей с ТП к формированию неканонических вторичных структур ДНК. Отдельные мономеры в составе поля ТП могут различаться по нуклеотидной последовательности заменами, делециями или инсерциями в нескольких нуклеотидов. Различные варианты повторов высшего порядка (HOR, high order repeat) характерны для отдельных хромосом и могут формировать длинные блоки в составе одной хромосомы. Экспрессия мажорного сателлита мыши (MaSat) оказалась необходима на двухклеточной стадии развития для того, чтобы произошла радикальная реорганизация генома. ТП – первые из последовательностей родительского генома, которые начинают экспрессироваться. Такие принципиальные открытия, касающиеся роли ТП, базируются на известной клонированной последовательности MaSat мыши. Для большинства других ТП невозможно определить их транскрипционный статус, так как до сих пор эти ТП не

описаны и не классифицированы. Недостаток информации о ТП затрудняет их исследования.

Диссертация А.А. Михеенко посвящена разработке новых методов оценки качества сборки геномов, программ для визуализации сборок и графов сборки, а также проблемам сборки ГХ районов с учетом особенностей строения ТП. Программы для сборки ГХ районов начали появляться, но их еще мало и трудно оценить качество их работы, т.е. качество конечной сборки. Ошибки в высшей степени вероятны в областях ТП. Программный пакет TandemTools, представленный в диссертации А.А. Михеенко, как раз и предназначен для анализа длинных тандемных повторов и обнаружения ошибок в сборках высокоповторяющихся областей. Результаты работы представлены в главе 2 (рис.2.9-2.11 и табл.8) и не вызывают сомнений в эффективности.

Качество проделанной работы отражено в трех главах с результатами, которые демонстрируют существенные преимущества разработанных методов. О важности и новизне диссертации можно судить по информации, представленной в введении и заключении. Приведены примеры использования созданных программ в реальных проектах. Результаты работы представлены в четырех серьезных статьях, опубликованных в журналах, индексируемых в базах данных Web of Science Core Collection и Scopus.

В порядке дискуссии хотелось бы задать автору несколько вопросов и высказать некоторые пожелания.

- Из текста диссертации складывается впечатление, что автор а) не делает различия между ЦЕН и периЦЕН ТП; б) за «нормальный» ТП принят альфа сателлит (а-сат) человека. Биологам-экспериментаторам хорошо известно, что даже а-сат существует как а-сатI (ЦЕН, содержит CENP-box) и а-сатII (периЦЕН, не содержит CENP-box) и они принципиально различны по, например, параметрам ДНК-белкового связывания. То же относится и к а-сат, как «образцу» ТП. Например, ЦЕН минорный сателлит (МиСат) мыши имеет мономер длиной 120 пн, а не ~170 нп как у а-сат. Возможно, для алгоритмов сборки эти различия не важны. Однако, они чрезвычайно важны для функционирования ЦЕН\периЦЕН и хочется надеяться, что будущие усилия будут направлены на создание программ, различающих разные ТП по формальным признакам.
- Отмечено, что программа StringDecomposer требует на вход мономеры. Для центромер человека мономеры, к счастью, известны, если использовать а-сат, как «образцовый» ТП. Означает ли это, что для других геномов, где неизвестен ЦЕН ТП (хомяки, свинья, коза, лошадь etc), пользователь, перед тем, как использовать программу StringDecomposer должен провести классификацию ТП в исходных ридх или контигах программой TRF (Benson, 1999) с набором определенных параметров, отсекающих мини- и микро-сателлиты
- В диссертации несобранную часть генома человека оценили в 10%-15 % генома. Обычно приводят гораздо более значительные цифры – около 80% генома составляют повторы разных типов. На чем основана эта конкретная цифра?
- На стр.41 сказано «минимайзеры скорее всего будут неравномерно распределены по сборке из-за локальных расширений идентичных тандемных повторов». Можно пояснить причины неравномерного распределения минимайзеров.

Текст диссертации поражает минимальным количеством опечаток.

Следует отметить, что возникшие вопросы ни в коей мере не умаляют высокого качества диссертации.

Диссертация Михеенко Аллы Александровны на тему «Разработка вычислительных методов для анализа и визуализации эукариотических геномных сборок» соответствует основным требованиям, установленным Приказом от 01.09.2016 № 6821/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Михеенко Алла Александровна заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 03.01.09. – Математическая биология, биоинформатика. Пункт 11 указанного Порядка диссертантом не нарушен.

Председатель диссертационного совета  
доктор биологических наук, профессор Кафедры цитологии и гистологии Биологического факультета СПбГУ



27 декабря 2020 года

Подгорная О.И.