ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 11 за 2017 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Романенков К.В.

Метод оценки качества сборки генома на основе частот k-меров

Рекомендуемая форма библиографической ссылки: Романенков К.В. Метод оценки качества сборки генома на основе частот k-меров // Препринты ИПМ им. М.В.Келдыша. 2017. № 11. 24 с. doi:10.20948/prepr-2017-11
URL: http://library.keldysh.ru/preprint.asp?id=2017-11

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Российской академии наук

К.В. Романенков

Метод оценки качества сборки генома на основе частот k-меров

Москва — 2017

Романенков К.В.

Метод оценки качества сборки генома на основе частот k-меров

Достаточно распространена ситуация, когда результаты применения геномных сборщиков или одного сборщика с разными параметрами существенно отличаются для одних и тех же входных данных, при этом в настоящее время не существует единой методики выбора наилучшей сборки. В данной работе предложен новый метод оценки качества геномной сборки организмов, для которых использование уже собранных геномов невозможно, с помощью анализа частот k-меров на основе программного средства Jellyfish. Предложенный метод устанавливает соответствие между набором коротких чтений, полученных в результате секвенирования, и собранным геномом, позволяя более точно оценивать результат геномной сборки. В результате проверки метода на различных сборках организма Encephalitozoon cuniculi fungus было установлено, что в большинстве случаев предложенная методика коррелирует с референс-зависимыми метриками и позволяет корректно определять лучшую сборку. При этом не была выявлена взаимосвязь между качеством сборки и стандартными метриками.

Ключевые слова: частоты k-меров, сравнение геномных сборок, оценка качества геномной сборки, Encephalitozoon cuniculi fungus

Kirill Vladimirovich Romanenkov

A new method of evaluating genome assemblies based on kmers frequencies

Running different genome assemblers or one genome assembler with different parameters on the same input data commonly leads to a great variety of results. However, there is no generally recognized method for choosing the best assembly. This article introduces a new reference-free method based on Jellyfish software for evaluating genome assembly by kmers frequencies analysis. The proposed method sets up a correspondence between short reads obtained from sequencer and assembled genome, which allows a more accurate genome assembly assessing. The method was validated on different assemblies of Encephalitozoon cuniculi fungus organism. It was found that in most cases it correlates with reference-dependent metrics and could correctly identify the best assembly. Furthermore, an interconnection between assembly quality and standard reference-free metrics was not observed.

Key words: kmers frequencies, genome assemblies evaluation, quality assessment, Jellyfish, Encephalitozoon cuniculi fungus

1. Введение

Современные высокотехнологичные автоматические методы расшифровки последовательностей ДНК (секвенирование) позволяют в течение относительно короткого времени (несколько дней) получить сотни миллиардов коротких последовательностей (длиной 100-500 символов) из четырех букв A, T, G, C, полученных прочтением фрагментов входного образца ДНК.

Восстановление последовательностей хромосом изучаемого организма по-прежнему остается весьма сложной задачей. Несмотря на успехи в расшифровке геномов человека, мыши и других организмов, большая часть видов еще даже не секвенирована. Ключевой этап сборки генома *de novo* (то есть сборки, при которой отсутствует ранее восстановленный геном особи того же вида, что и исследуемая) – это объединение прочитанных фрагментов генома («**ридов**») на основе их перекрывающихся участков. Другими словами, составление из перекрывающихся последовательностей, считанных из случайно выбранных мест генома, набора последовательностей большей длины («**контигов**»), которые соответствуют непрерывным фрагментам секвенируемой ДНК.

Для сборки генома используют специальные программы - сборщики генома, которые объединяют короткие фрагменты, полученные на этапе секвенирования. Большинство из них основано на концепции графа де Брюйна, которая заключается в следующем. Из коротких фрагментов, полученных на этапе секвенирования, формируются всевозможные подстроки длины k (k-меры). Таким образом, из строки длины l получается $l - \mathrm{k}+1$ k-меров. Затем сборщик строит граф де Брюйна, вершинами которого являются k-меры, а ориентированное ребро соединяет две вершины, если в соответствующих им k-мерах суффикс размера (k-1) первого из них совпадает с префиксом размера (k-1) второго, то есть между ними существует участок перекрытия размера (k-1). После этого выполняется упрощение этого графа, и все найденные в нем пути без разветвлений (контиги) попадают в ответ.

Обычно для работы геномных сборщиков необходимо задать несколько параметров. Достаточно распространена ситуация, когда результаты применения сборщиков или одного сборщика с разными параметрами существенно отличаются для одних и тех же входных данных. В настоящее время не существует единой методики выбора наилучшей сборки, одной из самых распространенных практик остается запуск сборщиков с различными параметрами, а затем выбор наилучшего варианта согласно различным метрикам. Тем не менее, эти метрики не учитывают степень близости полученного генома к набору коротких фрагментов и не позволяют получить полное представление о результатах сборки [1][24].

В данной работе предложен новый метод оценки качества геномной сборки при отсутствии «**референса**» (ранее восстановленный геном особи того же или родственного вида, к которому относится и исследуемый образец) с

помощью анализа частот k-меров, устанавливающий соответствие между набором коротких чтений, полученных в результате секвенирования, и собранным геномом, позволяя более точно оценивать результат геномной сборки.

2. Существующие методики оценивания качества геномной сборки

Результатом сборки генома *de novo* практически никогда не являются непосредственно последовательности хромосом, так как в ряде случаев не удается решить даже приближенную задачу восстановления последовательностей. На выходе сборщика обычно получается набор непрерывных, возможно перекрывающихся, фрагментов исследуемой ДНК - контигов. К настоящему времени разработано множество методик сравнения наборов контигов между собой. Их можно разделить на две основные категории: не требующие уже собранной каким-либо другим способом исследуемой последовательности («**референсная последовательность**») и требующие наличия референса.

2.1 Безреференсные методики

Довольно часто в задаче сборке генома *de novo* отсутствует референсная последовательность. Такая ситуация возникает, если секвенируется геном организма, который до этого не изучался, либо существует предположение, что существующая последовательность содержит серьезные ошибки. Ниже приведен список основных безреференсных методик.

Число контигов. Принято считать, что меньшее число контигов характеризует сборку в лучшую сторону. Однако это не всегда верно, например, в случае сравнения двух множеств контигов, при этом число последовательностей в первом множестве меньше, чем во втором, зато длина минимальной последовательности из второго множества превосходит длину максимальной последовательности из первого множества. В таком случае, второе множество контигов будет предпочтительнее.

Суммарная длина контигов. В идеале общее число символов, содержащееся в контигах, должно совпадать с числом символов референсной последовательности. На практике из-за перекрытия между контигами или непокрытия части генома ридами эти величины могут довольно значительно отличаться. Возможны ситуации, когда исследуемый геном включает в себя геном другого организма или просто загрязнен образцами чужой ДНК. Тогда сборщик восстанавливает части геномных последовательностей, относящихся к разным организмам, что, очевидно, ведет к увеличению суммарной длины контигов.

Nx, где $0 \leq x \leq 100$. Максимально возможная длина контига L, такая, что контиги с длиной $\geq L$ составляют по крайней мере $x\%$ от суммарной длины

контигов. Можно переписать эту метрику в формульном виде: пусть $C=\{c_1,\ldots,c_n\}$- упорядоченное множество контигов: $|c_i| \geq |c_{i+1}|$, $i=\overline{1,n-1}$. Тогда $Nx(C) = \max\limits_{1\leq k\leq n}(|c_k|)$:

$$\frac{\sum_{i=1}^{k}|c_i|}{\sum_{i=1}^{n}|c_i|} \times 100\% \geq x\%$$

Данная методика считается одной из основных при сравнении геномных сборок *de novo* и призвана отразить баланс между числом контигов, их средней и суммарной длиной. Стандартное значение x при этом равняется 50, то есть сборки сравниваются по длине контига, контиги длиннее которого составляют половину объема сборки. Большее значение, к примеру, N50 соответствует лучшей сборке и может считаться приближенным отражением фрагментированности набора контигов. Однако высокое значение N50 далеко не всегда говорит о качественной сборке: например, последовательно склеив все риды в одну большую строку, мы получим один контиг с максимально возможным значением N50 среди всевозможных сборок. Очевидно, что построенный таким образом контиг не имеет никакого отношения к восстановлению исходной строки.

Метрика ALE[2]. В отличие от вышеперечисленных метрик, данной метрике для оценки качества сборки помимо множества контигов требуется также набор входных ридов. ALE - это программа для качественной оценки сборки геномной последовательности на основе принципа максимального правдоподобия. В качестве ответа она выдает логарифм вероятности того, что сборка является верной при наличии заданного набора чтений. Для этого оценивается три фактора: насколько содержание чтений совпадает со сборкой, насколько априорные расстояния между парными чтениями совпадают с получившимися в результате сборки и насколько априорная глубина покрытия в каждой позиции совпадает с получившейся в результате сборки на основе GC-состава [2].

Данная метрика была выбрана в качестве эталонного программного средства, оценивающего качество сборки с помощью набора ридов, и в последующих разделах будет произведено сравнение предложенной методики и метрики ALE. Отметим, что в рамках дипломной работы Казакова «Разработка метода оценки качества сборки генома на основе принципа максимального правдоподобия», выполненной в НИУ ИТМО, было предложено улучшение этой методики, основанное на оценке глубины покрытия в каждой позиции с помощью эмпирического распределения, однако исходный код не был выложен в открытый доступ.

2.2 Методики, использующие референсный геном

Существование уже собранной последовательности секвенируемого организма позволяет гораздо точнее оценить корректность полученной сборки.

Кажущаяся на первый взгляд бессмысленной задача сборки уже восстановленной до этого последовательности имеет множество практических приложений. Геномы многих организмов, принадлежащих к одному виду, не являются точной копией друг друга, и именно участки отличия между ними представляют интерес для изучения. Когда говорят о расшифровке генома какого-то организма, имеется в виду сборка некоторого «абстрактного» генома, с которым впоследствии будут сравниваться другие собранные последовательности геномов этого же вида.

Анализ вариаций в последовательностях ДНК позволяет определять степень родства между организмами, помогает бороться с существующими генетическими заболеваниями или диагностировать ряд из них [3]-[5]. Процесс полной сборки референсной последовательности достаточно трудозатратен: обычно он включает в себя секвенирование с помощью различных технологий, ручной анализ полученных фрагментов, проведение дополнительных экспериментов по упорядочиванию набора контигов и формированию из них последовательностей хромосом. Полная сборка обычно не обходится без секвенирования по Сэнгеру, требующего значительных финансовых и временных ресурсов, поэтому геномов, для которых получена референсная последовательность, пока не так много. Тем не менее, в случае наличия такой последовательности нижеперечисленные метрики помогают оценить корректность полученной сборки.

Ошибки сборки. В этом и последующих пунктах будет использоваться термин **выравнивание** или **картирование** контигов или их частей на референсную последовательность, обозначающий размещение контигов над референсом так, чтобы друг под другом оказались сходные участки. Можно выделить три основных типа ошибок в контигах, подробнее см. [6]:

- *Relocation*. Ситуация, когда при картировании частей контига на референс промежуток между ними составляет более 1000 пар нуклеотидных оснований или они перекрываются на более чем 1000 пар нуклеотидных оснований.
- *Translocation.* Ситуация, когда части контига картируются на разные хромосомы референса.
- *Inversion.* Ситуация, когда части контига картируются на различные цепочки (стренды) ДНК.

NGx/NGAx, где 0 ≤ x ≤ 100. Значением NGx является максимально возможная длина контига L, такая, что контиги с длиной $\geq L$ составляют по крайней мере x% от длины референсного генома, а не суммарной длины контигов, как в случае Nx.

Метрика NGAx, которую еще называют **скорректированным N50**, является комбинацией NG50 и числа ошибок сборки и рассчитывается следующим образом. Прежде всего производится выравнивание контигов на референс. Если последовательность контига содержит ошибку сборки одного из вышеописанных типов, то в этой точке он разбивается на два отдельных блока,

для которых снова проводится процедура картирования на референс. Если какой-то из получившихся блоков невозможно выравнять на референс, то он исключается из рассмотрения. Для полученного таким образом набора контигов рассчитывается стандартная метрика NGx. Значение NGAx всегда меньше или равно значению NGx и, по сути, штрафует сборщик за склеивание участков, не следующих друг за другом в геноме. Более высокое значение NGAx обычно соответствует лучшей сборке.

Длиннейший непрерывный участок (ДНУ). Значение этой метрики - длина длиннейшего блока, образованного в результате процедуры разбиения контигов для вычисления NGAx. По сути, это скорректированная длина самого протяженного контига.

3. Предложенная методика оценки качества геномной сборки

3.1 Анализ числа уникальных k-меров

Практически любой геном содержит повторяющиеся участки, однако, начиная с определенного значения k, k-меры в некотором роде однозначно идентифицируют его; если посчитать числа встречаемости k-меров для достаточно большого k (ограниченного при этом сверху длиной рида), получится, что большая часть из них встречается в геноме в единственном экземпляре. Разумеется, истинность этого утверждения сильно зависит от структуры генома, но для геномов небольших и средних размеров с относительно малым числом повторов это в большинстве случаев так. Отметим, что *de novo* сборка больших геномов сложной структуры малоосмысленна и обычно подразумевает либо наличие референса, либо огромное число ридов, полученных с помощью различных технологий секвенирования. Покажем, что с увеличением значения k вероятность содержания случайного k-мера в геноме стремится к 0.

Утверждение 1. *Пусть дан случайный геном длины* G*, содержащий четыре вида нуклеотидов (символов). Вероятность встречи конкретного символа в геноме не зависит от позиции и составляет 0.25. Тогда вероятность встретить случайную строку длины* L *в последовательности генома хотя бы один раз составляет* $1 - \left(1 - \left(\frac{1}{4}\right)^L\right)^{G-L+1}$.

Доказательство. Приближенно оценим вероятность того, что строка не встретится в последовательности генома. Так как символы в строке и геноме случайны, вероятность совпадения строки длины L с некоторой подстрокой в геноме составляет $\left(\frac{1}{4}\right)^L$. Соответственно, вероятность несовпадения составляет $1 - \left(\frac{1}{4}\right)^L$. Предполагая, что позиции, в которых может быть найдена строка, не зависят друг от друга, получим вероятность того, что строка не содержится ни в

одной из возможных позиции генома: $\left(1-\left(\frac{1}{4}\right)^{L}\right)^{G-L+1}$. Из этого следует, что вероятность встретить строку в геноме хотя бы один раз составляет:

$$1-\left(1-\left(\frac{1}{4}\right)^{L}\right)^{G-L+1}. \quad (1)$$

Данное приближение является достаточно грубым и не учитывает структуру генома и строки. Тем не менее, этот способ вполне подходит для приближенной оценки такой длины подстроки, при которой вероятность её встречи в геноме достаточно мала.

Пусть $G = 10^9$, то есть порядок длины генома сравним с человеческим. Тогда по формуле (1) вероятность встретить случайную подстроку длины 14 хотя бы один раз составляет 0.975893. Для $k = 20$ эта же вероятность составляет 0.000909. Для геномов меньших размеров, например, бактерий или грибов, можно выбрать меньшее k, чтобы добиться схожей малой вероятности встречаемости строки.

К примеру, в геноме микроспоридии Encephalitozoon cuniculi fungus 2295551 из 2373670 21-меров встречаются в единственном экземпляре. Отметим, что этот геном в дальнейшем будет использоваться для проверки предложенной методики.

3.2 Подсчет числа уникальных k-меров

Задача подсчета встречаемости различных k-меров достаточно часто возникает в контексте сборки генома. Распределение частот используется для корректирования ридов разделением содержащихся в них k-меров на «доверенные» и «ошибочные» (то есть k-меров, последовательность которых содержит ошибку секвенирования) [19]. Также эта информация используется некоторыми сборщиками для определения того, является ли рассматриваемый участок повтором или нет. В ряде случаев анализ распределения частот k-меров позволяет находить ошибки сборки в уже сформированных контигах (проект AMOS [7]).

Разработано достаточное количество программных средств, предназначенных для подсчета числа k-меров в последовательностях: Jellyfish[8], BFCounter[9], DSK[10] и другие. Поскольку BFCounter не позволяет отслеживать k-меры, встречающиеся в единственном экземпляре, а DSK не смог корректно завершить свою работу на ряде входных данных, было принято решение использовать программу Jellyfish для подсчета числа встречаемости k-меров. Эта программа использует внутреннее сжатое представление k-меров, может работать в многопоточном режиме и требует относительно немного оперативной памяти.

Jellyfish принимает на вход последовательности ДНК, строит индекс для множества всех k-меров, содержащихся в них, и сохраняет его на диск. Функционал программы позволяет получать гистограмму встречаемости k-

меров, выводить k-меры заданной встречаемости, проверять, содержится ли заданный k-мер в множестве проиндексированных k-меров, и выполнять ряд других операций [8].

3.3 Предлагаемая методика

Основная идея предложенного метода оценки качества геномной сборки заключается в установлении соответствия между уникальными k-мерами в собранном геноме и k-мерами в ридах.

Сначала строится гистограмма встречаемости k-меров в ридах, полученных в результате секвенирования. Пример такой гистограммы для организма Encephalitozoon cuniculi fungus[17] при k = 21 изображен на рисунке 1а. По оси X отложены числа встречаемости k-меров в наборе чтений, а по оси Y отложено количество всевозможных k-меров с данной встречаемостью. Видно, что гистограмма имеет два пика: первый связан с ошибками чтения генома, а второй соответствует уникальным k-мерам в исходном геноме.

Большое количество ошибочных (ложных) k-меров в наборе чтений объясняется тем, что ошибка в прочтении даже одного символа (например, ошибка замены) приводит к потенциальному возникновению нового k-мера, до этого не присутствующего в этом множестве. Число встречаемости у второго пика (около 116) обусловлена покрытием при секвенировании генома: количеством прочтений каждого символа. На рисунке 1б изображена гистограмма встречаемости 21-меров собранного генома Encephalitozoon cuniculi fungus с помощью сборщика Velvet. Из-за того, что каждый участок генома прочитывается несколько раз, каждый уникальный k-мер из собранного генома встречается во множестве k-меров коротких чтений около 116 раз.

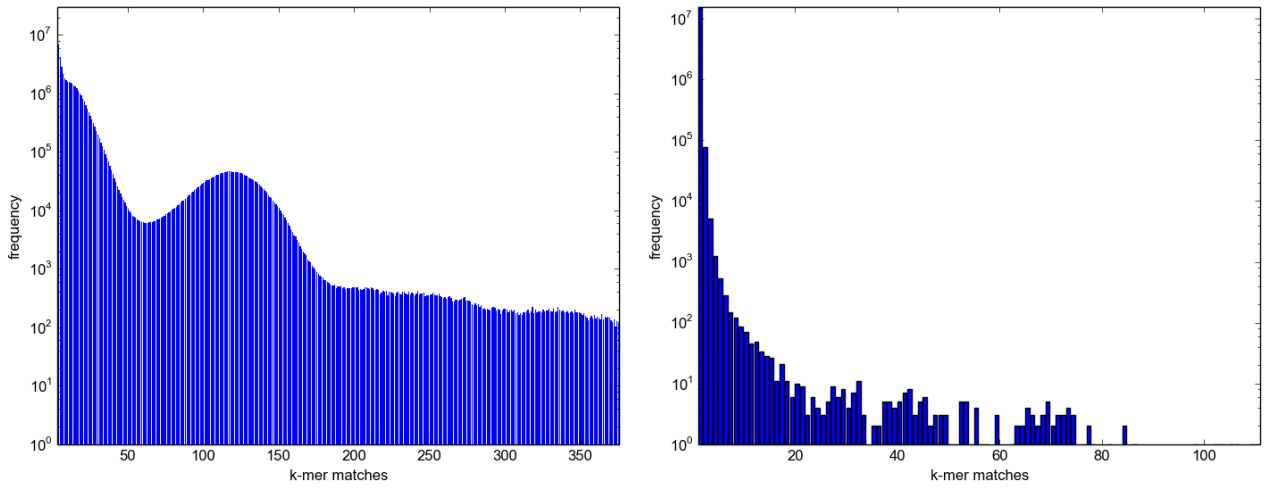

Рис. 1. Число встречаемости k-меров (k=21) в наборе чтений (а) и собранном геноме (б) Encephalitozoon cuniculi fungus. Логарифмическая шкала

Предлагается вычислять долю различных k-меров, взятых из некоторой окрестности второго пика на гистограмме встречаемости k-меров в чтениях, среди уникальных k-меров для собранного генома по следующей формуле:

$$Q = \frac{\sum_{i=1}^{|K_{\text{uniq\_reads}}|} \left[k^r(i) \in K_1^g\right]}{\left|K_{uniq\_reads}\right|}, \quad (2)$$

где $K_1^g = \{k^g : k^g \in K^g, abundance(k^g) = 1\}$, $K_i^r = \{k^r : k^r \in K^r, abundance(k^r) = i\}$, $K_{uniq\_reads} = \bigcup_{i\,=\,a_{p\_l}}^{a_{p\_r}} K_i^r$, $\left[a_{p\_l}; a_{p\_r}\right]$ - некоторая окрестность из второго пика на гистограмме встречаемости k-меров коротких чтений, $abundance(k)$ - число встречаемости k-мера k, K^r - множество всех k-меров коротких чтений, K^g - множество всех k-меров собранного генома.

Чем больше значение Q, тем лучше полученная сборка соотносится с результатами секвенирования. Таким образом, предложенная методика устанавливает соответствие между набором коротких чтений, полученных в результате секвенирования, и собранным геномом, позволяя более точно оценивать результат геномной сборки.

Сама процедура сравнения сборок согласно предложенной методике выглядит следующим образом:

1. Построение гистограммы встречаемости k-меров для ридов, полученных при секвенировании исследуемого генома. Так как молекула ДНК содержит две комплементарные цепочки, вводится понятие **канонического** k-мера: этот такой k-мер из пары k-мер и комплементарный k-мер, который меньше лексикографически. При подсчете чисел встречаемости k-меров сохраняются последовательности только канонических k-меров, при этом числа встречаемости для k-мера и его комплемента суммируются.
2. Выбор некоторой окрестности пика уникальных k-меров на гистограмме встречаемости k-меров в ридах.
3. Построение гистограммы встречаемости k-меров для каждой из полученных сборок.
4. Подсчет значения Q. Каждый k-мер из множества уникальных k-меров ридов проверяется на вхождение во множество уникальных k-меров собранного генома. Для этого подсчитываются числа встречаемости k-мера и его комплемента во множестве k-меров рассматриваемой сборки. Если сумма чисел их встречаемости не равна единице, то такой k-мер отбрасывается. Если же k-мер и его комплемент в сумме встретились ровно один раз, то считается, что данный k-мер принадлежит множеству уникальных k-меров собранного генома.
5. Выбор сборки с максимальным значением Q в качестве наилучшей.

Значение Q зависит от размера выбранной окрестности пика уникальных k-меров на гистограмме встречаемости для ридов: как показали эксперименты, оно уменьшается с увеличением размера окрестности. Это объясняется тем, что при расширении границ окрестности знаменатель дроби в формуле (2) увеличивается на число k-меров, попавших в расширенную окрестность, в то время как числитель увеличивается только на число добавленных k-меров, принадлежащих множеству уникальных k-меров собранного генома.

3.4 Выбор окрестности пика уникальных k-меров

Поскольку значение Q зависит от размера окрестности пика уникальных k-меров на гистограмме встречаемости для ридов, важно корректно выбрать ее границы. Если выбрать ее размер слишком маленьким, то значительная часть уникальных k-меров ридов не войдет в нее и значения Q будут достаточно близки к друг другу, что затруднит задачу их различения. Если же выбрать размер окрестности слишком большим, то помимо уникальных k-меров в нее войдут как ошибочные k-меры (число которых растет при расширении левой границы окрестности), так и k-меры, соответствующие повторяющимся участкам в исследуемом геноме (число которых растет при расширении правой границы окрестности). Исходя из этих соображений, выбор окрестности выглядит следующим образом:

1. На гистограмме встречаемости k-меров в ридах ищется второй пик, соответствующий уникальным k-мерам в исходном геноме, и локальный минимум между двумя пиками. Предполагается, что они существуют в гистограмме, то есть она имеет вид, приведенный на рисунке 1а. Такая структура гистограммы является достаточно типичной для секвенированных геномов небольших и средних размером и, таким образом, не сильно ограничивает область применимости предлагаемого метода. Пусть число встречаемости для пика уникальных k-меров равно p, а для локального минимума m.
2. Значение a_{p_r} устанавливается равным $2(p-10)-\alpha$: $\alpha \in [0;8], a_{p\_r} \vdots 10$. Выбор правой границы объясняется следующим образом: если уникальные k-меры из собранного генома встречаются во множестве k-меров ридов около p раз, значит k-меры, содержащиеся в собранном геноме два раза, будут встречаться во множестве k-меров ридов около $2p$ раз и так далее. То есть, правая граница выбирается таким образом, чтобы исключить из окрестности неуникальные k-меры собранного генома.
3. Для выбора левой границы приближенно оценивается размер исследуемого генома без учета повторов с помощью формулы, описанной в [11]:

$$G = \frac{S(L-k+1)}{LM}$$

где S - общий размер ридов, L - длина рида, k - размер k-мера, M - покрытие k-меров, которое определяется как число встречаемости для пика уникальных k-меров, т. е. $M = p$.

4. Выбирается сборка, общий размер контигов которой ближе всего к оцененному размеру генома, затем из сборки исключаются контиги короче 500 символов, а для оставшихся контигов подсчитывается число уникальных k-меров. Обозначим это число как K'^{g}_{1}, тогда левая граница вычисляется следующим образом:

$$a_{p\_l} = max \begin{cases} \max\limits_{x}: \left| \bigcup\limits_{i=x}^{a_{p\_r}} K_i^r \right| \geq 0.85 \left| K'^{g}_{1} \right| \\ m - \alpha: \alpha \in [1; 10], (m - \alpha) \vdots 10 \end{cases}.$$

То есть требуется, чтобы количество уникальных k-меров ридов составляло по крайней мере 85% от количества уникальных k-меров в контигах длиною ≥500 символов для сборки, размер которой ближе всего к размеру оцененного генома, и при этом левая граница не сильно отличалась от числа встречаемости локального минимума. Это ограничение необходимо, чтобы исключить из окрестности ошибочные k-меры.

Константы, используемые при выборе границ окрестности, получены эмпирически. Их выбор обусловлен необходимостью наличия достаточного количества уникальных k-меров в окрестности для различения сборок с одной стороны и необходимостью исключения из окрестности максимального количества ошибочных k-меров и k-меров, соответствующих повторяющимся участкам в исследуемом геноме, с другой.

4. Программная реализация

Для проверки метода оценки качества геномных сборок было разработано несколько программных модулей, условно разбитых на три группы. К первой группе относятся bash-скрипты, предназначенные для решения задачи сборки, а именно подготовки ридов для сборки (приведение к совместимому со сборщиками формату, очистка ридов от ошибок и шума) и запуска геномных сборщиков. Сборка генома — достаточно ресурсоемкая задача и требует большого количества как процессорного времени, так и оперативной памяти. Запуск всех сборщиков производился на суперкомпьютере «Ломоносов-1», при этом в ряде случаев в исходный код программ для сборки были внесены небольшие изменения для обеспечения совместимости с используемыми на кластере компиляторами.

Ко второй группе относятся вспомогательные программы, написанные на языке Python и предназначенные для анализа полученных сборок, а именно построения гистограммы встречаемости k-меров в ридах и контигах, выбора окрестности пика уникальных k-меров в ридах согласно разделу 3.4 и

сохранения проиндексированного множества k-меров на диск. Построение гистограммы и сохранение k-меров на диск проводилось средствами программы Jellyfish.

К третьей группе относится модуль, написанный на языке C++ и основанный на программном средстве Jellyfish. Данный модуль, используя внутреннее представление Jellyfish для k-меров, подсчитывает значение Q согласно пункту 4 из раздела 3.3 для каждой из рассматриваемых сборок.

Результатом объединения вышеперечисленных модулей стала система для запуска геномных сборщиков на вычислительном кластере и выбора оптимального параметра k, максимизирующего значение Q, схематично описанная в разделе 3.3

5. Проверка предложенной методики

Множество геномных сборщиков, основанных на концепции графа де Брюйна, достаточно велико [20]. Эти программные продукты разрабатываются как целыми отделами исследовательских центров, так и энтузиастами-одиночками, являются как свободно распространяемым ПО, так и коммерческими продуктами с закрытым исходным кодом, требуют для своей работы как маломощный персональный компьютер, так и высокопроизводительный вычислительный кластер. Отметим, что существуют сборщики, комбинирующие концепции графа де Брюйна и overlap-layout-consensus, например MaSuRCA[21] и ITMO Genome Assembler[22].

Отобранные в этой работе сборщики объединяют следующие критерии: они фигурируют в статьях, посвященных сравнительному анализу качества геномных сборок [12][23], используют концепцию графа де Брюйна, при этом не используя более одного значения k для сборки за раз (как это, например, делают сборщики A5, SPAdes и IDBA), имеют открытый исходный код, и им для работы достаточно одного набора ридов. Приведенный ниже список не претендует на полный охват всей области геномных сборщиков, тем не менее, он включает в себя программы, относительно успешно зарекомендовавшие себя в этом качестве. Всего рассматривается четыре сборщика: ABYSS[13], Ray[14], SOAPdenovo[15] и Velvet[16].

Для анализа корректности работы предложенной методики была проведена серия вычислительных экспериментов, в ходе которых производилась сборка генома микроспоридии Encephalitozoon cuniculi fungus [17]. Длина данного гаплоидного генома составляет около 2.5 млн нуклеотидов, он состоит из 11 хромосом, его N50 равен 218329. Парные риды длиной 100 символов, полученные в результате секвенирования генома данного организма по технологии Illumina HiSeq 2000 paired end, доступны на сайте Европейского биоинформатического института [18]; референс, состоящий из 11 последовательностей хромосом, опубликован на сайте NCBI [17].

Для сборщиков, использующих концепцию графа де Брюйна, важнейший параметр, позволяющий влиять на результат, - значение k, то есть длина тех кусков, на которые будут разбиты риды в процессе построения графа де Брюйна. Упрощенно говоря, повторы в геноме длиннее k могут усложнить структуру графа и затруднить извлечение из него контигов, таким образом, предпочтительнее выбирать k достаточно большим. С другой стороны, чем больше k, тем больше вероятность того, что последовательность k-мера будет содержать ошибку, таким образом, слишком большие значения k уменьшают долю корректных k-меров во входных данных. Еще один эффект, на который влияет выбор значения k, - связность получаемого графа де Брюйна. Если перекрытие между ридами составляет менее k символов, то соответствующие вершины в графе де Брюйна не будут соединены ребром и при восстановлении последовательности контига, включающего в себя эти вершины, отсутствующее ребро будет интерпретировано как нарушение покрытия. То есть, вместо одного контига сборщик получит две отдельные последовательности. Таким образом, выбор значения k - это компромисс между различными последствиями этого выбора.

Сборка генома микроспоридии производилась для пяти различных значений параметра k: 25, 33, 43, 51, 59. Выбор этих значений обусловлен следующими соображениями: при сборке генома *de novo* минимальное значение k обычно устанавливается равным 21 (например, в случае сборщика SPAdes), исходя из рассуждений, аналогичных утверждению 1. Верхняя граница параметра, как правило, не превосходит 60-70% от длины рида, в данном случае равной 100. Промежуточные значения призваны продемонстрировать необходимость выбора для каждого сборщика собственного значения k, обеспечивающего наилучшее качество сборки.

Так как для данного генома существует референс, то оценка качества работы предложенной методики выглядела следующим образом: каждый сборщик запускался с пятью различными значениями параметра k, для полученных сборок производился сбор статистик, описанных в разделе 2, с помощью программы Quast и подсчет значений Q, и пять сборок упорядочивались согласно подсчитанным значениям.

В качестве меры «правильности» полученных сборок была выбрана референс-зависимая метрика NGAx, позволяющая достаточно точно судить о качестве геномных сборок и описанная в пункте 2.2. Упорядоченный в соответствии с этой метрикой, набор сборок назовем **каноническим**. Упорядоченный в соответствии со значением Q, набор сборок сравнивался с каноническим набором. Ясно, что чем меньше различий существует между двумя наборами, тем лучше предложенная методика отражает качество геномной сборки, по крайней мере в терминах метрики NGAx.

Нижерасположенные разделы содержат результаты сравнения сборок организма Encephalitozoon cuniculi fungus различными сборщиками с различными значениями параметра k. Все метрики, за исключением Q, были

рассчитаны с помощью программы Quast. Программа Quast использует контиги длины ≥500 символов для сбора статистики, поэтому метрика Q также была вычислена на основе контигов длиннее 500 при k=21. Расчет значений Q проводился на машине с 32 Гб оперативной памяти с использованием всех 8 ядер процессора.

5.1 Анализ результатов работы сборщика Velvet

В таблице 1 показаны результаты сравнения сборок организма Encephalitozoon cuniculi fungus программой Velvet.

Как видно из таблицы, с точки зрения NGA50 лучшей является сборка при k=33, затем следуют k=43, 25, 51, 59. Это коррелирует со значениями Q, в отличие от метрики N50, значение которой у двух лучших сборок является самым низким. Несмотря на то, что количество и суммарная длина контигов при k=25 и 51 отличаются практически на порядок, это не оказало значительного влияния на предложенную методику, и значения Q в обоих случаях достаточно близки (как и соответствующие значения NGA5*0).*

Таблица 1

Результаты сравнения различных сборок Encephalitozoon cuniculi fungus программой Velvet

k	N50	Число контигов	NGA50	Длина контигов (в Mb)	ALE	Q
25	1175	15927	11316	17.7	-120*10^7	0.934086
33	928	14977	21452	14.5	-119*10^7	0.938702
43	937	8093	18836	8.3	N/A	0.937946
51	4091	2610	9237	3.7	N/A	0.92775
59	3339	1124	3189	2.4	-114*10^7	0.905697

На рисунке 2 приведены графики метрики NGAx для пяти сборок Encephalitozoon cuniculi fungus для различных значений параметра k. По оси абсцисс отложено x, а по оси ординат - соответствующее значение NGAx; цветные ломанные линии обозначают различные сборки. Более высокое значение NGAx отвечает более качественной сборке, то есть, согласно этой метрике, значения k, соответствующие различным сборкам в порядке убывания качества, расположены в следующем порядке: 33, 43, 25, 51, 59. Эта последовательность полностью согласуется с последовательностью, получаемой в результате анализа значений Q.

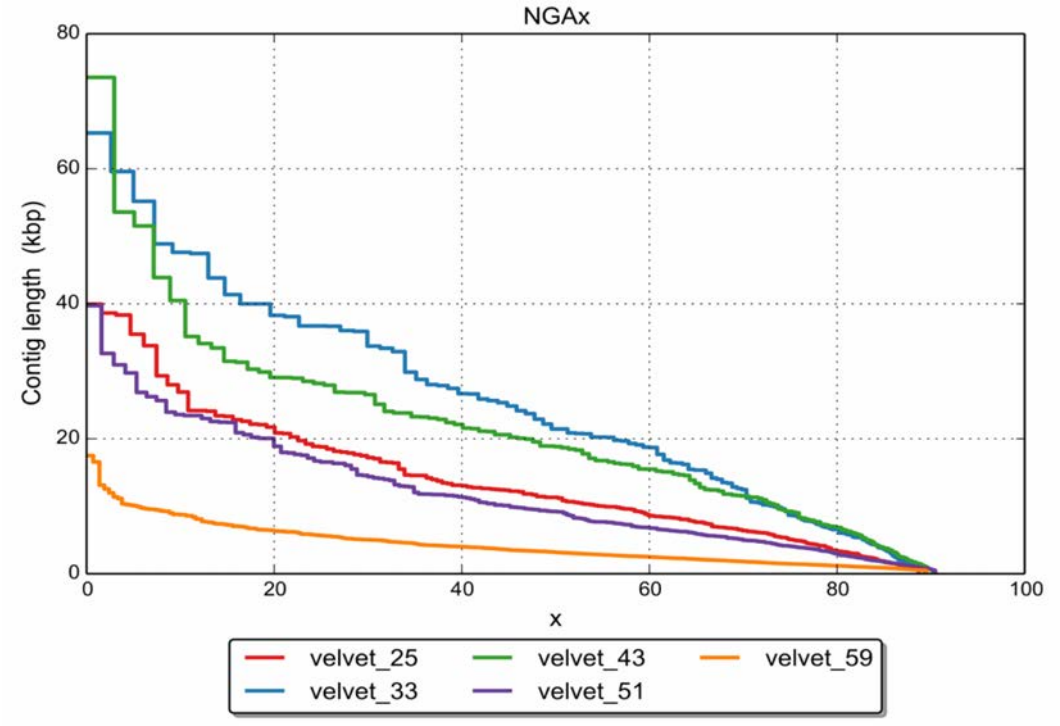

Рис. 2. График NGAx для результатов работы сборщика Velvet, запущенного с различными значениями параметра k

5.2 Анализ результатов работы сборщика Soap

В таблице 2 показаны результаты сравнения сборок организма Encephalitozoon cuniculi fungus программой SOAPdenovo.

Таблица 2

Результаты сравнения различных сборок Encephalitozoon cuniculi fungus программой SOAPdenovo

k	N50	Число контигов	NGA50	Длина контигов (в Mb)	ALE	Q
25	1101	16920	644	17.4	-121*10^7	0.639442
33	932	12249	2578	11.5	N/A	0.898704
43	932	3816	14839	4.7	N/A	0.936651
51	7480	1245	9248	2.8	-114*10^7	0.92729
59	3412	1086	3183	2.3	N/A	0.901179

Как видно из таблицы, с точки зрения NGA50 лучшей является сборка при k=43, затем следуют k=51, 59, 33, 25. Это коррелирует со значениями Q, в отличие от метрики N50, согласно которой k=51 или 59 предпочтительнее чем k=43. Значение Q при k=25 заметно выделяется из общего ряда. Это объясняется низким **покрытием** референсного генома контигами при k=25 (60% против ≈85.5% для остальных k). **Покрытие генома** определяется как

отношение количества нуклеотидов генома, на которые скартировался хотя бы один контиг из сборки, к общей длине генома. Данный факт свидетельствует о том, что предложенная методика соотносится с таким референс-зависимым показателем сборки, как покрытие.

Связь между покрытием генома и значением Q основывается на предположении о том, что риды равномерно покрывают геном и каждый нуклеотид генома содержится хотя бы в одном риде, то есть найдется содержащий его k-мер из множества k-меров ридов. Если же на какой-то символ генома не скартировался ни один контиг, значит этот символ не будет входить во множество уникальных k-меров собранного генома, что повлечет за собой уменьшение значения Q.

Также отметим, что отличия в числе и суммарной длине контигов практически на порядок для некоторых значений k не смещают значение Q. Это достигается за счет того, что при расчете Q вычисляется доля k-меров из некоторой окрестности пика уникальных k-меров чтений, присутствующих в уникальных k-мерах сборки, а не наоборот.

На рисунке 3 приведены графики метрики NGAx для пяти сборок Encephalitozoon cuniculi fungus для различных значений параметра k. Согласно этой метрике, значения k, соответствующие различным сборкам в порядке убывания качества, расположены в следующем порядке: 43, 51, 59, 33, 25. Эта последовательность полностью согласуется с последовательностью, получаемой в результате анализа значений Q.

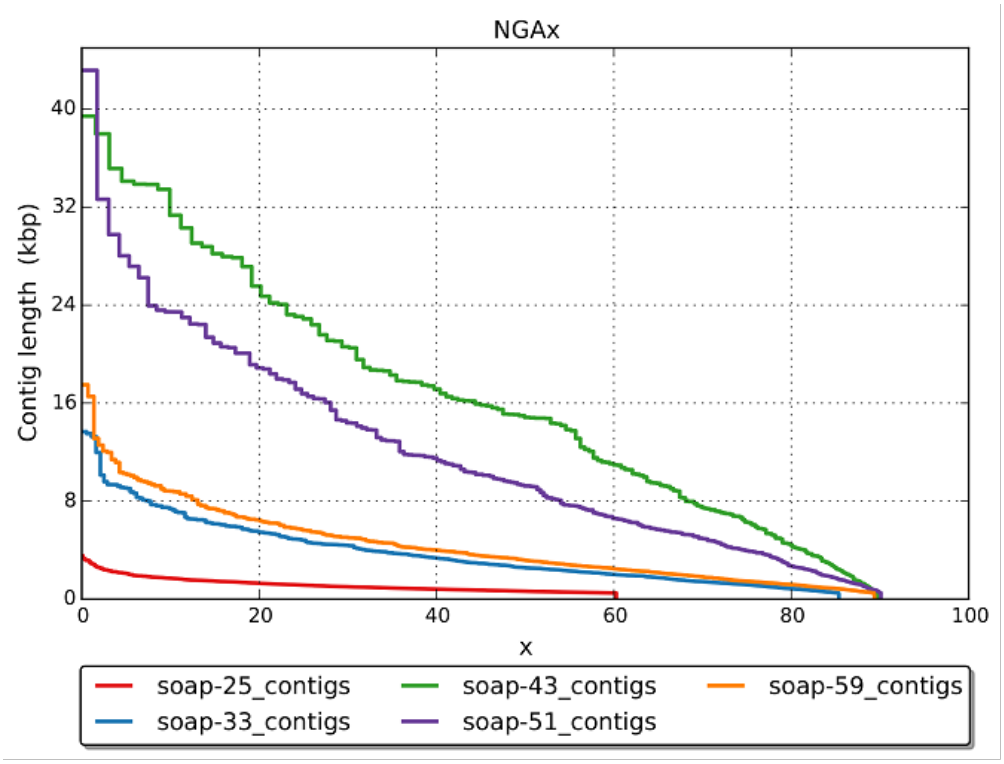

Рис. 3. График NGAx для результатов работы сборщика SOAPdenovo, запущенного с различными значениями параметра k

5.3 Анализ результатов работы сборщика Abyss

В таблице 3 показаны результаты сравнения сборок организма Encephalitozoon cuniculi fungus программой Abyss.

Таблица 3

Результаты сравнения различных сборок Encephalitozoon cuniculi fungus программой Abyss

k	N50	Число контигов	NGA50	Длина контигов (в Mb)	ALE	Q
25	1495	14981	28200	19.7	-120*10^7	0.932986
33	1110	12913	55107	14.3	N/A	0.940028
43	1082	4939	63324	5.8	N/A	0.940482
51	41991	1241	55270	3.1	N/A	0.934639
59	20966	488	20966	2.5	N/A	0.926928

На рисунке 4 приведены графики метрики NGAx для пяти сборок Encephalitozoon cuniculi fungus для различных значений параметра k, и в данном случае их достаточно сложно однозначно ранжировать согласно этой метрике. Вместе с тем на рисунке четко можно выделить две группы значений k, значительно отличающихся по качеству: k=25, 29 и k=43, 33, 51.

Несмотря на то, что значение NGA50 для k=51 выше, чем для k=33, NGAx для k=33 больше NGAx для k=51 для большего количества x и с большей разницей. Поэтому, согласно метрике NGAx, значения k, соответствующие различным сборкам в порядке убывания качества, расположены в следующем порядке: 43, 33, 51, 25, 59. Эта последовательность полностью согласуется с последовательностью, получаемой в результате анализа значений Q.

Отметим, что полученные значения Q не позволяют четко разделить сборки на две группы аналогично значениям NGAx, хотя сохраняют тот же относительный порядок. Проведенный анализ показал, что при изменении границ окрестности уникальных k-меров ридов, распределение значений Q приближается к распределению значений NGAx. Тем не менее, цель предложенной методики заключается именно в упорядочивании набора сборок, полученных в результате работы одного геномного сборщика, запущенного с различными параметрами k, а значения Q не обязаны находиться в линейной зависимости с NGAx.

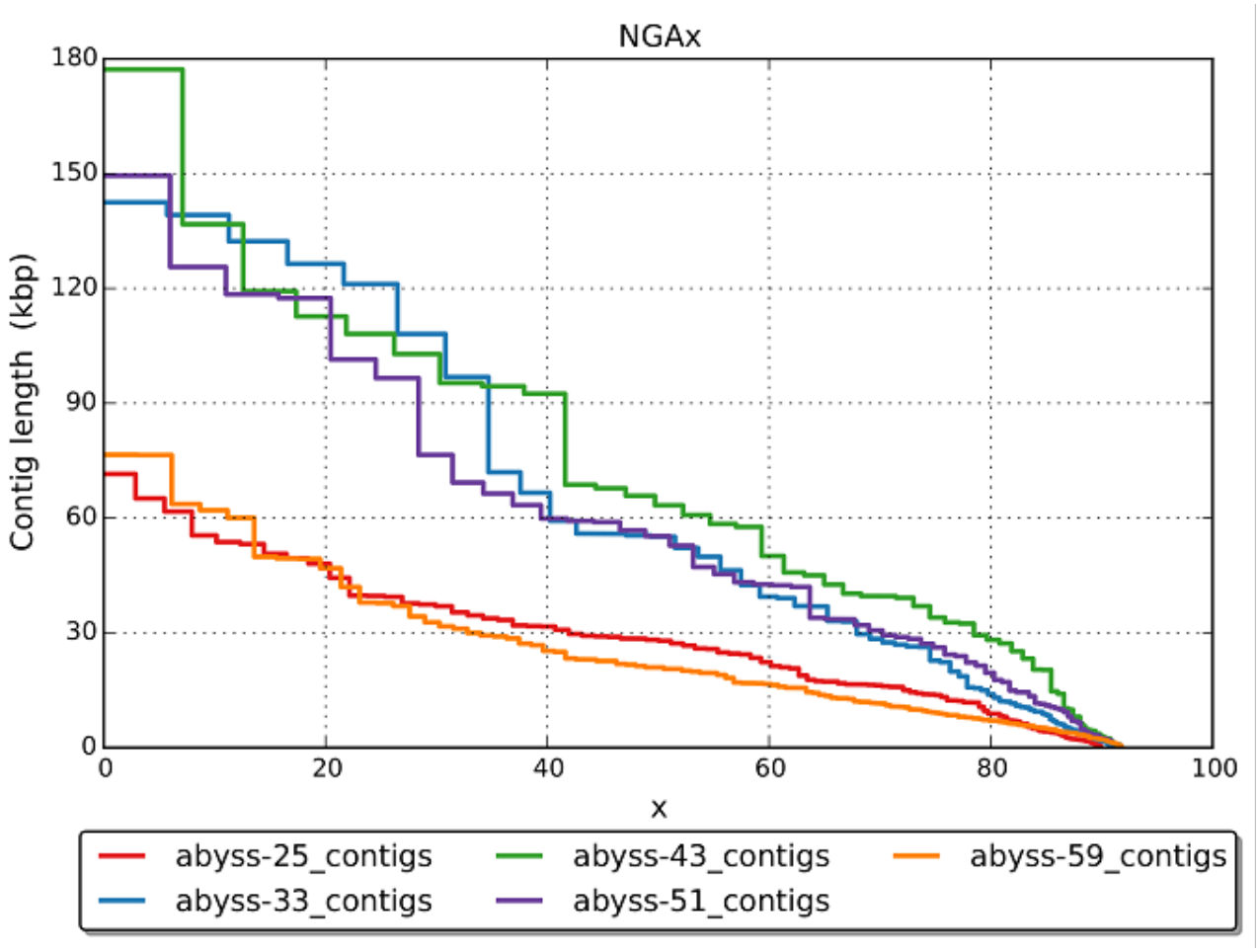

Рис. 4. График NGAx для результатов работы сборщика Abyss, запущенного с различными значениями параметра k

5.4 Анализ результатов работы сборщика Ray

В таблице 4 показаны результаты сравнения сборок организма Encephalitozoon cuniculi fungus программой Ray.

Таблица 4

Результаты сравнения различных сборок Encephalitozoon cuniculi fungus программой Ray

k	N50	Число контигов	NGA50	Длина контигов (в Mb)	ALE	Q
25	950	8820	78910	9.3	-116*10^7	0.895447
33	5780	3274	42981	4.6	-114*10^7	0.927833
43	14182	758	15087	2.7	N/A	0.92385
51	6909	339	4018	1.6	N/A	0.603678
59	2827	419	-	0.9	-115*10^7	0.364621

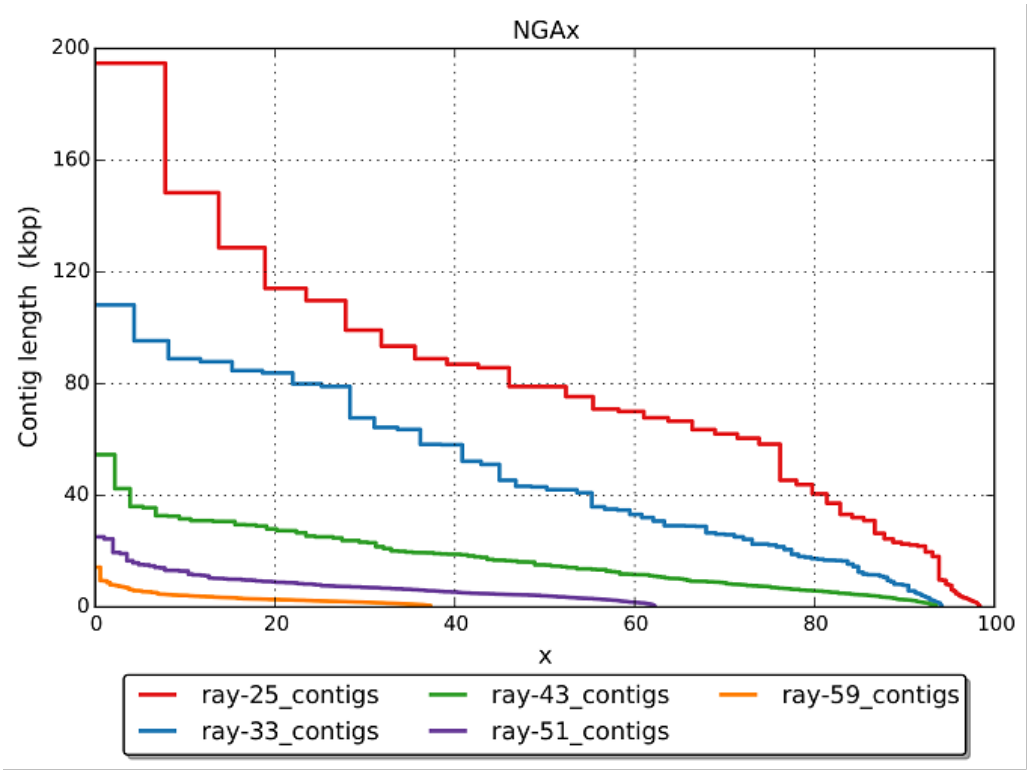

Рис. 5. График NGAx для результатов работы сборщика Ray, запущенного с различными значениями параметра k

На рисунке 5 приведены графики метрики NGAx для пяти сборок Encephalitozoon cuniculi fungus для различных значений параметра k. Согласно этой метрике, значения k, соответствующие различным сборкам в порядке убывания качества, расположены в следующем порядке: 25, 33, 43, 51, 59. Эта же последовательность, упорядоченная согласно метрике Q, выглядит так: 33, 43, 25, 51, 59. Ray - единственный сборщик, для которого эти последовательности не совпадают. Это объясняется особенностями сборки, построенной для k=25. Чтобы продемонстрировать их, введем вспомогательную величину $\tilde{Q}$, вычисление которой отличается от Q тем, что в числителе выражения находятся k-меры, которые встречаются в сборке хотя бы один раз.

$$\tilde{Q} = \frac{\sum_{i=1}^{|K_{\text{uniq\_reads}}|} [k^r(i) \in \tilde{K}_1^g]}{|K_{uniq\_reads}|},$$

где $\tilde{K}_1^g = \{k^g : k^g \in K^g, abundance(k^g) \geq 1\}$.

Значения Q и $\tilde{Q}$ в идеальных условиях не должны значительно отличаться, так как, согласно их определению, k-меры, проверяемые на вхождение во множество k-меров сборки, берутся из окрестности уникальных k-меров чтений и в большинстве своем должны встречаться в сборке в единственном экземпляре.

Очевидно, что Q всегда не превосходит $\tilde{Q}$, однако слишком большая разница их значений свидетельствует о потенциальных проблемах сборки. В случае Ray $\tilde{Q}$ при k=25 составляет 0.954664, то есть отличается от Q практически на 7%, в то время как в среднем разница между Q и $\tilde{Q}$ составляет 1-2%. Это означает, что достаточно большая доля уникальных k-меров чтений по каким-то причинам встречается в сборке более одного раза. Данный эффект может вызываться неверной оценкой кратности повторов в процессе работы сборщика, неудалением концевых перекрытий контигов, неудачным результатом применения упрощающих эвристик к графу де Брюйна и прочими факторами. Независимо от причины, качество такой сборки вызывает определенные вопросы.

Еще одной характеристикой сборки служит **показатель дублирования** (duplication ratio), который вычисляется как отношение количества выровненных символов контигов к количеству выровненных символов генома. В идеале, он должен быть равен 1, но если сборка содержит много контигов, покрывающих один и тот же участок референса, показатель ее дублирования может быть больше единицы. Для k=25 этот показатель составляет 1.063, то есть более 6% сборки составляют дубликаты, в то время как для остальных k его значение находится в интервале 1.007-1.025. К сожалению, метрика NGAx пропорциональна количеству повторяющихся фрагментов в сборке, так как при ее расчете контиги картируются на геном независимо, и их взаимное перекрытие никак не учитывается. Поэтому возможна ситуация, когда высокий показатель NGAx достигается не за счет хорошего качества сборки, а за счет наличия в ней большого числа повторов. Случай k=25 как раз попадает в эту категорию, поэтому он менее предпочтителен чем k=33 или 43, и значения Q вполне отражают эту ситуацию.

Значения $\tilde{Q}$ для k=25, 33, 43, 51, 59 равны 0.954664, 0.94906, 0.93624, 0.612756, 0.367249 соответственно. То есть последовательности k, упорядоченные согласно $\tilde{Q}$ и NGAx, совпадают. Однако в данном случае Q точнее отражает качество сравниваемых сборок, чем NGAx и $\tilde{Q}$.

6. Заключение

В работе предложена новая методика оценки качества геномной сборки при отсутствии референса с помощью анализа частот k-меров. Данная методика была проверена на геноме Encephalitozoon cuniculi fungus размера около 3*10^6 пар нуклеотидных оснований. Было установлено, что в большинстве случаев последовательность сборок, упорядоченная в порядке убывания значения NGAx, полностью согласуется с последовательностью сборок, упорядоченной в порядке убывания значения Q. Таким образом, показано, что предложенная методика коррелирует с референс-зависимыми метриками и позволяет корректно определять лучшую сборку. При этом не была выявлена взаимосвязь между качеством сборки и стандартными метриками.

Для проверки предложенной методики была разработана система для запуска геномных сборщиков на вычислительном кластере и выбора оптимального параметра k, максимизирующего значение Q

Также было проведено сравнение предложенной методики с существующим аналогом: метрикой ALE, предназначенной для оценивания качества сборки. Для сравнения ALE и предложенного метода для каждой из полученных сборок был посчитан ее логарифм вероятности с помощью программы ALE. Результаты этой оценки приведены в таблицах 1-4. К сожалению, на ряде данных ALE аварийно завершила свою работу, такие случаи обозначается в таблицах с помощью N/A.

Несмотря на то, что во многих случаях ALE не смогла выдать результат, полученные цифры свидетельствуют об отсутствии связи между логарифмом вероятности того, что сборка является верной, и NGAx. Похоже, что единственный показатель, с которым коррелирует оценка ALE - это суммарная длина контигов. Таким образом, для данного генома метрика ALE не отражает референс-зависимые метрики и не позволяет определять наиболее качественную сборку.

7. Библиографический список

1. Salzberg S.L., Phillippy A.M., Zimin A., Puiu D., Magoc T., et al GAGE: A critical evaluation of genome assemblies and assembly algorithms. // Genome Research. — 2012. — V. 22. — P. 557–567.
   URL: http://doi.org/10.1101/gr.131383.111
2. Clark S.C., Egan R., Frazier P.I., Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. // Bioinformatics. — 2013. — V. 29. — P. 435–443.
   URL: http://doi.org/10.1093/bioinformatics/bts723
3. Alföldi J., Lindblad-Toh K. Comparative genomics as a tool to understand evolution and disease. // Genome Research. — 2013. — V. 23, No. 7. — P. 1063–1068. URL: http://doi.org/10.1101/gr.157503.113
4. Баженова О., О'Брайен С. Применение биоинформатики в медицинских исследованиях // Здоровье – основа человеческого потенциала: проблемы и пути их решения. — 2014. — Т. 9, №1. — С. 102–104.
5. Gonzaga-Jauregui C., Lupski J.R., Gibbs R.A. Human Genome Sequencing in Health and Disease. // Annual review of medicine. — 2012. — V. 63. — P. 35-61. URL: http://doi.org/10.1146/annurev-med-051010-162644
6. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. // Bioinformatics. — 2013. — V. 29, No. 8. — P. 1072–1075. URL: http://dx.doi.org/10.1093/bioinformatics/btt086
7. Treangen T.J., Sommer D.D., Angly F.E., Koren S., Pop M. Next Generation Sequence Assembly with AMOS. // Current Protocols in Bioinformatics. — 2011. — V. 11, No. 11.8. — P. 1–18.
   URL: http://doi.org/10.1002/0471250953.bi1108s33

8. Marcais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. // Bioinformatics. — 2011. — V. 27, No. 6. — P. 764-770.
URL: http://doi.org/10.1093/bioinformatics/btr011
9. Melste P., Pritchard J.K. Efficient counting of *k*-mers in DNA sequences using a bloom filter. // BMC Bioinformatics. — 2011. — V. 12, No. 1. — P. 333-339. URL: http://doi.org/10.1186/1471-2105-12-333
10. Rizk G., Lavenier D., Chikhi R. DSK: k-mer counting with very low memory usage. // Bioinformatics. — 2013. — V. 29, No. 5. — P. 652–653.
URL: http://doi.org/10.1093/bioinformatics/btt020
11. Li X., Waterman M.S. Estimating the repeat structure and length of DNA sequences using L-tuples. // Genome Research. — 2003. — V. 13, No. 8. — P. 1916–1922. URL: http://doi.org/10.1101/gr.1251803
12. Koren S., Treangen T.J., Hill C.M., Pop M., Phillippy A.M. Automated ensemble assembly and validation of microbial genomes. // BMC Bioinformatics. — 2014. — V. 15, No. 5. — P. 126–134.
URL: http://dx.doi.org/10.1186/1471-2105-15-126
13. Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J., Birol I. ABySS: A parallel assembler for short read sequence data. // Genome Research. — 2009. — V. 19, No.6. – P. 1117–1123.
URL: http://doi.org/10.1101/gr.089532.108
14. Boisvert S., Laviolette F., Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. // Journal of Computational Biology. — 2010. — V. 17, No. 11. — P. 1519–1533.
URL: http://doi.org/10.1089/cmb.2009.0238
15. Luo R., Liu B., Xie Y., et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. — GigaScience. — 2012. — V. 1, No. 18. — P. 1–6. URL: http://doi.org/10.1186/2047-217X-1-18
16. Zerbino D.R., Birney E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. // Genome Research. — 2008. — V. 18, No. 5. — P. 821–829. URL: http://doi.org/10.1101/gr.074492.107
17. Encephalitozoon cuniculi GB-M1
URL: http://www.ncbi.nlm.nih.gov/genome/39?genome_assembly_id=22671 (accessed: 1.02.2017).
18. European Nucleotide Archive
URL: http://www.ebi.ac.uk/ena/data/view/SRR122309 (accessed 1.02.2017).
19. Александров А.В., Шалыто А.А. Метод исправления ошибок вставки и удаления в наборе чтений нуклеотидной последовательности. // Научно-технический вестник информационных технологий, механики и оптики. — 2016. — № 1. — С.108–114.
URL: http://doi.org/10.17586/2226-1494-2016-16-1-108-114
20. Miller J.R., Koren S., Sutton G. Assembly algorithms for next-generation sequencing data. // Genomics. — 2010. — V. 95, No. 6. — P. 315–327.

URL: http://doi.org/10.1016/j.ygeno.2010.03.001
21. Zimin A.V., Marcais G., Puiu D., Roberts M., Salzberg S.L., Yorke J.A. The MaSuRCA genome assembler. // Bioinformatics. — 2013. — V. 29, No. 21. —P. 2669–2677. URL: http://doi.org/10.1093/bioinformatics/btt476
22. Сергушичев А.А., Александров А.В., Казаков С.В., Царев Ф.Н., Шалыто А.А. Совместное применение графа де Брейна, графа перекрытий и микросборки для de novo сборки генома. // Известия Саратовского университета. Новая серия. Серия Математика. Механика. Информатика. — 2013. — Т. 13, вып. 2, ч. 2. — С. 51–57.
23. Phillippy A.M., Schatz, M.C., Pop M. (2008). Genome assembly forensics: finding the elusive mis-assembly. // Genome Biology. — V. 9, No. 3. — R55. URL: http://doi.org/10.1186/gb-2008-9-3-r55
24. Magoc T., Pabinger S., Canzar S., Liu X., Su Q., Puiu D., Tallon L.J., Salzberg S.L. GAGE-B: an evaluation of genome assemblers for bacterial organisms. // Bioinformatics. — 2013. — V. 29, No. 14. — P. 1718–1725. URL: http://doi.org/10.1093/bioinformatics/btt273

Оглавление