ISSN 0868–5886 *НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2022, том 32, № 1, с. 3–10*

СИСТЕМНЫЙ АНАЛИЗ ПРИБОРОВ И ИЗМЕРИТЕЛЬНЫХ МЕТОДИК

УДК 543.51+ 681.2–5

© А. Г. Бородинов, В. В. Манойлов, И. В. Заруцкий, А. И. Петров, В. Е. Курочкин, 2022

МЕТОДИКА ОЦЕНКИ КАЧЕСТВА ГЕНОМНОЙ СБОРКИ НА ОСНОВЕ АНАЛИЗА ЧАСТОТНОСТИ *K*-МЕРОВ В СЕКВЕНАТОРЕ ПАРАЛЛЕЛЬНОГО СЕКВЕНИРОВАНИЯ

В настоящее время в связи с развитием приборостроения для проведения генетического анализа существует острая необходимость в разработке методик оценки качества геномной сборки. Подсчет встречаемости различных *k*-меров часто возникает в задачах сборки генома. В данной работе на основе анализа различных программных средств выбраны программы, которые позволяют оценить качество геномной сборки. С помощью выбранных программ обработаны данные, полученные на отечественном секвенаторе параллельного секвенирования Нанофор СПС. На основе результатов обработки этих данных произведена оценка качества геномной сборки по методике анализа *k*-меров для прибора Нанофор СПС.

Кл. сл.: *k*-мер, NGS-методы, биоинформатика, сборка генома

ВВЕДЕНИЕ

K-мер — это просто последовательность из *k* символов в строке (или нуклеотидов в последовательности ДНК в задаче секвенирования). Разложение последовательности на ее *k*-меры позволяет анализировать этот набор фрагментов фиксированного размера, а не последовательность целиком, и это может быть более эффективным подходом. Простой пример: чтобы проверить, происходит ли последовательность S из организма A или из организма B, предполагая, что геномы A и B известны и достаточно разные, мы можем проверить, содержит ли S больше *k*-меров, присутствующих в A или в B.

Практически любой геном содержит повторяющиеся области, однако, начиная с определенного значения *k*, *k*-меры определенным образом однозначно идентифицируют его; если мы посчитаем количество появлений *k*-мер для достаточно большого *k* (ограниченного сверху длиной чтения), оказывается, что большинство из них находятся в геноме в единственном экземпляре. Например, если порядок длины генома сравним с человеческим, вероятность встретить случайную подстроку длины 14 хотя бы один раз составляет 0.975893 [1]. Для $k = 20$ эта же вероятность составляет 0.000909.

Подсчет встречаемости различных *k*-меров часто возникает в задачах сборки генома. Распределение частот встречаемости используется для процедуры корректирования ри́дов, что подразумевает разделение содержащихся *k*-меров на "доверенные" и "ошибочные" [1]. Подобная информация используется некоторыми программами сборки генома для определения того, является ли рассматриваемый участок повтором или нет.

В настоящее время в связи с развитием приборостроения для проведения генетического анализа существует острая необходимость в разработке методик оценки качества геномной сборки. Такие методики позволят оценить достоверность проведения генетического анализа в существующих и вновь разрабатываемых приборах. В данной работе на основе анализа различных программных средств выбраны программы, которые позволяют оценить качество геномной сборки в секвенаторах параллельного секвенирования. С помощью выбранных программ обработаны данные, полученные на отечественном секвенаторе параллельного секвенирования Нанофор СПС.

АНАЛИЗ ПРОГРАММНЫХ СРЕДСТВ ОЦЕНКИ КАЧЕСТВА СБОРКИ ГЕНОМА

Поскольку количество *k*-мер растет экспоненциально для значений *k*, подсчет *k*-мер для больших значений *k* является вычислительно сложной задачей. Хотя достаточно простые реализации работают для малых значений *k*, их необходимо адаптировать для приложений с высокой пропускной способностью или когда *k* велико. Для решения этой проблемы были разработаны различные инструменты:

• Jellyfish использует многопоточную хеш-таблицу без блокировок для подсчета *k*-мер и имеет реализации на Python, Ruby и Perl [2];

• KMC — это инструмент для подсчета *k*-мер, который использует многодисковую архитектуру для оптимизации скорости [3];

• Gerbil использует подход хеш-таблицы, но с дополнительной поддержкой ускорения графического процессора [4];

• K-mer Analysis Toolkit (KAT) использует модифицированную версию Jellyfish для анализа количества *k*-мер [5].

В качестве основного инструмента работы с *k*-мерами был выбран KAT (K-mer Analysis Tookit), представляющий эффективный набор средств для быстрого подсчета, сравнения и анализа спектров *k*-мер произвольной длины из данных генетических последовательностей.

Основным методом анализа при работе с *k*-мерами является проверка качества сборки генома путем сравнения характеристик *k*-меров совокупности анализируемых ри́дов с референтным образцом или с собранным геномом (при сборке *de novo*). Инструмент KAT **hist** — это графическое представление набора данных, показывающее, сколько коротких последовательностей фиксированной длины (*k*-мер) появляется определенное количество раз. Частота встречаемости нанесена на ось *x*, а число *k*-меров на оси *y*. Пример 31-mer spectrum of S.cerevisae S288C WGS приведен на рис. 1.

Инструмент KAT **comp** генерирует матрицу с *k*-мерным набором последовательностей частот *k*-меров на одной оси, а частотой встречаемости *k*-меров другого набора на другой оси. При сравнении набора ри́дов со сборкой KAT сначала вычисляет свойства и состав *k*-меров сборки. При представлении в виде стоковых гистограмм спектр *k*-меров для ри́дов разбивается по числу копий *k*-меров для сборки. Кроме того, KAT предоставляет инструмент **sect** для отслеживания покрытия *k*-мерами, исходя из рассчитанных спектров *k*-меров для совокупности ри́дов и референса. Это может помочь идентифицировать такие артефакты сборки, как события сворачивания и разворачивания, или обнаруживать повторяющиеся области в последовательности ДНК.

KAT также включает инструмент **hist** для вычисления спектра из одного набора последовательностей и инструмент **gcp** для анализа гуанин-цитозин содержания (GC-контента) в зависимости от частоты *k*-меров. Инструмент **filter** можно использовать для выделения последовательностей из полного набора в соответствии либо с покрытием *k*-мерами или GC-содержанием для заданного набора. Эти инструменты могут использоваться для различных задач, включая обнаружение и извлечение загрязняющих веществ (contaminant detection) как в необработанных ри́дах, так и в сборках (assemblies), анализ смещения по GC-составу и согласованность между парно-концевыми (paired end) ри́дами с чувствительностью по концентрациям примесей от 0.1 ppm.

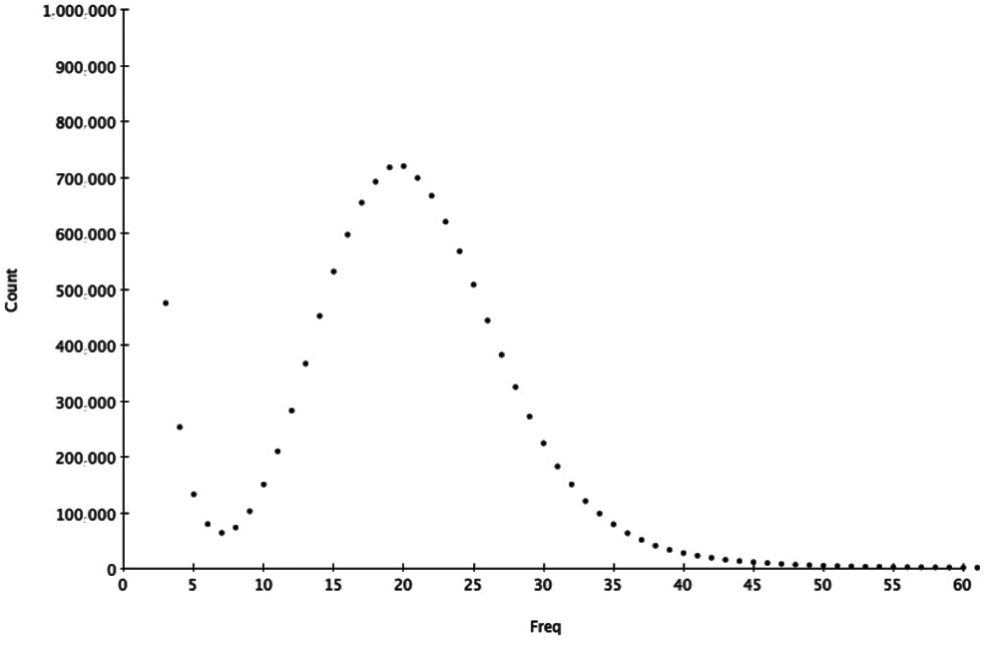

Рис. 1. Графическое представление набора данных KAT hist

KAT прост в использовании, обеспечивает высокую скорость анализа. Время получения результатов анализа составляет не более минуты.

МЕТОДИКИ РАБОТЫ С *K*-МЕРАМИ

В работе [1] предложен метод оценки качества геномной сборки, заключающийся в установлении соответствия между уникальными *k*-мерами в собранном геноме и *k*-мерами в ри́дах. Процедура выглядит следующим образом.

1. Построение гистограммы встречаемости *k*-меров для ри́дов.
2. Выбор окрестности пика уникальных *k*-меров на гистограмме встречаемости.
3. Построение гистограммы встречаемости *k*-меров для каждой сборки.
4. Расчет меры Q как доли различных *k*-меров, взятых из окрестности пика на гистограмме встречаемости *k*-меров в чтениях.
5. Выбор сборки с максимальным значением Q в качестве наилучшей.

В работе [6] предложен метод исправления ошибок, оптимизированный для работы с чтениями, содержащими как ошибки замены, так и ошибки вставки и удаления. Поскольку ошибки происходят с небольшой частотой, вероятность того, что один и тот же *k*-мер будет прочитан несколько раз с одинаковым набором ошибок, очень мала. Из этого вытекает, что те *k*-меры, которые встречаются в наборе чтений мало раз, являются ошибочными, остальные же являются реальными подстроками генома (рис. 2).

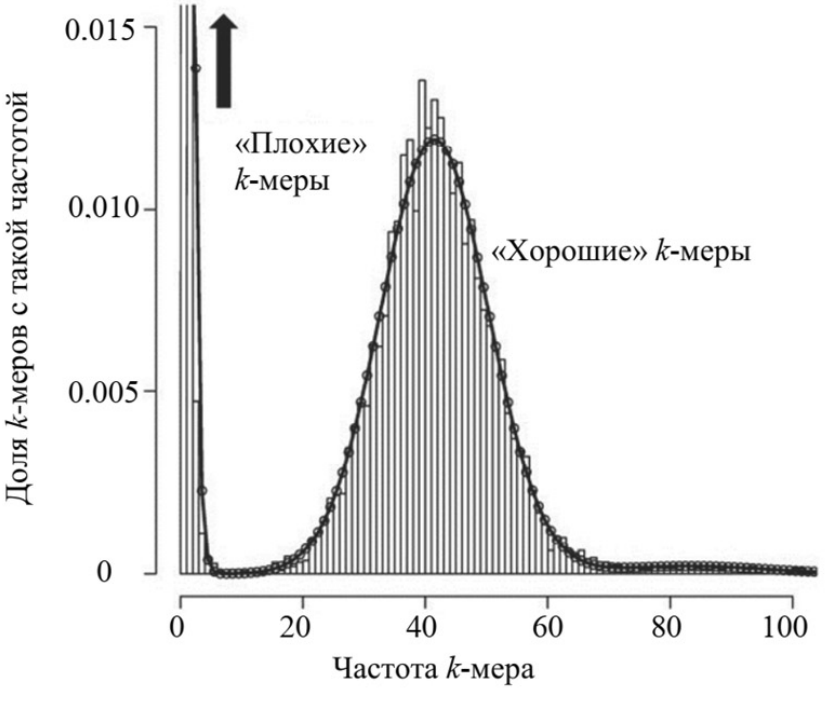

Рис. 2. Распределение частот *k*-меров в ри́дах [6]

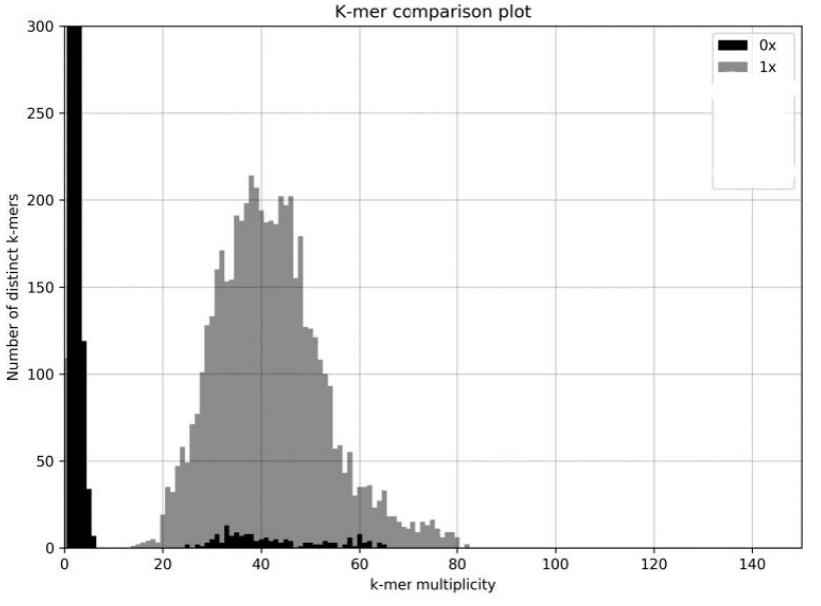

Рис. 3. Типичный *k*-mer comparison plot секвенирования Phix 174 на Illumina Miseq

ИСПОЛЬЗОВАНИЕ ПРОГРАММЫ KAT ДЛЯ ОБРАБОТКИ ДАННЫХ СЕКВЕНАТОРА НАНОФОР СПС

Для обработки данных секвенатора Нанофор СПС была использована опция программы KAT "K-mer comparison plot". По сути мы представляем, сколько элементов каждой частоты в спектре ри́дов оказались не включены в референтный геном (в нашем случае Phix174), включены один раз, включены дважды и т.д.

На рис. 3, 4 представлены k-mer comparison plot, полученные соответственно для приборов Illumina и Нанофор СПС. Показательно, что для сходных характеристик проточных ячеек запуск Нанофор СПС обеспечивает больший уровень покрытия ри́дами референсной последовательности (центр тяжести *k*-меров с уникальным покрытием).

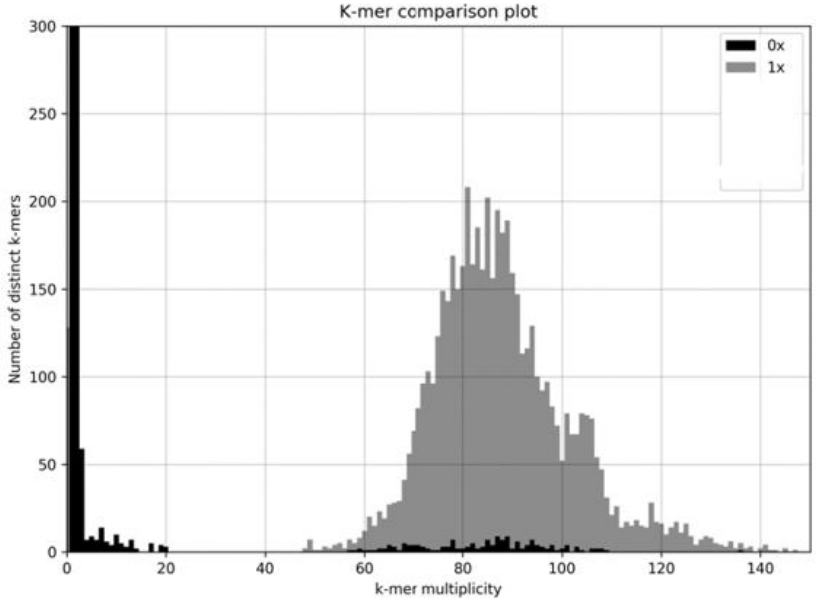

Рис. 4. Типичный *k*-mer comparison plot секвенирования Phix 174 на Нанофор СПС

ЗАКЛЮЧЕНИЕ

Проекты сборки генома обходятся дорого как по времени, так и по вложенным средствам. В этом случае выявление проблем с экспериментальными данными, обаруженных уже после сборки, может стать настоящей неудачей. С помощью K-mer Analysis Toolkit (KAT) исследователи могут получить доступ к качественным критериям и подтвердить свои результаты на более ранних этапах.

K-меры представляют собой небольшие фрагменты исходного генома с фиксированным числом оснований ДНК. Компьютер может эффективно работать с большим количеством *k*-меров, а затем идентифицировать связи между этими фрагментами, чтобы создать представление об исходном геноме. Основанные на *k*-мерах методы обычно используются для эффективного создания геномных сборок. KAT построен для изучения и сравнения наборов данных секвенирования с использованием основных свойств каждого отдельного *k*-мера, таких как частота встречаемости и нуклеотидный состав.

В первую очередь KAT может анализировать данные секвенирования для определения уровней случайных ошибок, систематических ошибок и контаминации. Информация, полученная в ходе этого анализа, может помочь исследователям решить, следует ли продолжать выполнение последующих задач, таких как сборка генома. Затем KAT может перепроверить проведенную сборку генома, определив полноту и точность сборки без каких-либо внешних справочных данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Романенков К.В.* Метод оценки качества сборки генома на основе частот *k*-меров, Препринт. ИПМ им. М.В. Келдыша, 2017.
2. *Marcais G., Kingsford C.* A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers // Bioinformatics. 2011. Vol. 27, is. 6. P. 764–770. DOI: 10.1093/bioinformatics/btr011
3. *Deorowicz S., Kokot M., Grabowski S., Debudaj-Grabysz A.* KMC 2: fast and resource-frugal *k*-mer counting // Bioinformatics. 2015. Vol. 31, is. 10. P. 1569–1576. DOI: 10.1093/bioinformatics/btv022
4. *Erbert M., Rechner S., Müller-Hannemann M.* Gerbil: a fast and memory-efficient *k*-mer counter with GPU-support // Algorithms for Molecular Biology. 2017. Vol. 12. Art. Num. 9. DOI: 10.1186/s13015-017-0097-9
5. *Mapleson D., Accinelli G.G., Kettleborough G.,Wright J., Clavijo B.J.* KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies // Bioinformatics. 2017. Vol. 33, is. 4. P. 574–576. DOI: 10.1093/bioinformatics/btw663
6. *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. Т. 16, № 1. С. 108–114. DOI: 10.17586/2226-1494-2016-16-1-108-114

Институт аналитического приборостроения РАН, Санкт-Петербург

Контакты: *Бородинов Андрей Геннадьевич*, borodinov@gmail.com

Материал поступил в редакцию 30.12.2021

METHODOLOGY FOR ASSESSING THE QUALITY OF GENOMIC ASSEMBLY BASED ON THE ANALYSIS OF THE FREQUENCY OF *K*-MERS IN A PARALLEL SEQUENCING SEQUENCER

A. G. Borodinov, V. V. Manoilov, I. V. Zarutskiy, A. I. Petrov, V. E. Kurochkin

Institute for Analytical Instrumentation of RAS, Saint-Petersburg, Russia

Counting the occurrence of different *k*-mers often causes problems of genome assembly. Analysis of the frequency distribution of *k*-mers makes it possible to find assembly errors in already formed contigs. Currently, in connection with the development of instrumentation for genetic analysis, there is an urgent need to develop methods for assessing the quality of genomic assembly. Such techniques will make it possible to assess the reliability of genetic analysis in existing and newly developed devices. In this work, based on the analysis of various software tools, programs were selected to assess the quality of genomic assembly in parallel sequencing sequencers. Using the selected programs, the data obtained on the domestic sequencer for parallel sequencing Nanofor SPS were processed. Based on the results of processing these data, the quality of the genomic assembly was assessed by the method of analysis of *k*-mers and recommendations were given for improving the hardware and software of the Nanofor SPS device.

Keywords: *k*-mers, NGS, bioinformatics, genome assembly

INTRODUCTION

A *k*-mer is simply a sequence of *k* symbols in a string (or nucleotides in a DNA sequence in the case of sequencing). The decomposition of a sequence into its *k*-mers allows one to analyze this set of fixed size fragments, rather than the whole sequence, and this may be a more efficient approach. A simple example: to check if the sequence S originates from organism A or from organism B, assuming that the genomes of A and B are known and quite different, we can check which *k*-mers contains S more of: those present in A or in B.

Almost any genome contains repeating regions, however, starting from a certain value of *k*, *k*-mers in a certain way uniquely identify it. If we count the number of occurrences of *k*-mers for a sufficiently large value of *k* (limited from above by the length of reads), it appears that most of them are in a single copy in the genome. For example, if the order of genome length is comparable with a human one, the probability of encountering a random substring of k =14 length at least once is 0.975893 [1]. For k = 20, the probability is 0.000909.

Counting the occurrence of different *k*-mers often arises in genome assembly tasks. The frequency distribution is used for the read correction procedure, which implies the seraration of the contained *k*-mers into "trusted" and "erroneous" ones [1]. This information is used by some genome assembly software programs to determine whether the region in question is a repeat or not.

Currently, due to the development of instrumentation for genetic analysis, there is an urgent need for the development of methods for assessing the quality of genomic assembly. Such techniques make it possible to assess the reliability of genetic analysis in existing and newly developed devices. In this work, based on the analysis of various software tools, programs were chosen that allow assessing the quality of genomic assembly in sequencers for parallel sequencing. Using the selected programs, the data obtained on the domestic sequencer Nanofor SPS [Нанофор СПС] for parallel sequencing were processed.

ANALYSIS OF SOFTWARE FOR ASSESSING THE QUALITY OF GENOME ASSEMBLY

Since the number of *k*-mers grows exponentially for values of *k*, calculating *k*-mers for large values of k is computationally challenging. While fairly simple applications work for small values of *k*, they need to be adapted when high throughput is needed or when *k* is large. Various tools have been developed to solve this problem:

- Jellyfish uses a multi-threaded, lock-free hash table for counting *k*-mers and has implementations in Python, Ruby, and Perl [2];
- KMC is a *k*-mer calculator that uses a multi-disk architecture to optimize speed [3];
- Gerbil uses a hash table approach, but with additional support for GPU acceleration [4];
- The K-mer Analysis Toolkit (KAT) uses a mod-

ified version of Jellyfish to analyze the number of *k*-mers [5].

As the main tool for working with *k*-mers, KAT (K-mer Analysis Tookit) was chosen, representing an effective set of tools for quickly calculating, comparing and analyzing the spectra of *k*-mers of arbitrary length from genetic sequence data.

The main analysis method when working with *k*-mers is to check the quality of genome assembly by comparing the characteristics of the *k*-mers of the set of analyzed reads with the reference sample or with the assembled genome (when assembling de novo). A KAT **hist** tool is a graphical representation of a dataset showing how many short, fixed-length sequences (*k*-mers) appear a specified number of times. The frequency of occurrence is plotted on the axis *x*, and the number of *k*-mers on the axis *y*. An example of 31-mer spectrum of S. cerevisae S288C WGS is given in Fig. 1.

Fig. 1. Graphical representation of the KAT hist dataset

A KAT **comp** generates a matrix with a *k*-mer set of frequency sequences of *k*-mers on one axis, and the frequency of occurrence of *k*-mers of another set on the other axis. When comparing a set of reads with an assembly, KAT first calculates the properties and composition of the *k*-mers of the assembly. When presented in the form of stock histograms, the spectrum of *k*-mers for reads is divided according to the number of copies of *k*-mers for assembly. In addition, KAT provides **sect** tool for tracking *k*-mer coverage based on calculated *k*-mer spectra for a set of reads and a reference. This can help identify assembly artifacts such as folding and unfolding, or detect repeating regions in a DNA sequence.

KAT also includes a **hist** tool for calculating a spectrum of a set of sequences and a **gcp** tool for analyzing guanine-cytosine content versus frequency of *k*-mers. A **filter** tool can be used to select sequences from the complete set according to either *k*-mer coverage or GC content for a given set. These tools can be used for a variety of tasks, including contaminant detection and extraction in both raw reads and assemblies, bias analysis over GC content, and consistency between paired end reads with sensitivity to impurity concentrations from 0.1 ppm. KAT is easy to handle, it provides high speed analysis. The time spent on obtaining the result of the analysis is no more than 1 min.

TECHNIQUES FOR WORKING WITH *K*-MERS

In [1], a method for assessing the quality of genomic assembly is proposed, which consists in establishing a correspondence between unique *k*-mers in the assembled genome and *k*-mers in reads. The procedure is as follows.

1. Construction of a histogram of the occurrence of *k*-mers for the reads.
2. Selection of the vicinity of the peak of unique *k*-mers on the histogram of occurrence.
3. Plotting a histogram of the occurrence of *k*-mers for each assembly.
4. Calculation of the measure Q as the fraction of different *k*-mers taken from the vicinity of the peak on the histogram of the occurrence of *k*-mers in reads.
5. Selection of the assembly with the maximum value of Q as the best.

In [6], an error correction method is proposed that is optimized for working with reads containing both substitution errors and insertion and deletion errors. Since errors occur with a small probability, the probability that the same *k*-mer will be read several times with the same set of errors is very small. It follows that those *k*-mers that occur a few times in the set of reads are erroneous, while the rest are real substrings of the genome (Fig. 2).

Fig. 2. Frequency distribution of *k*-mers in reads [6]

USING THE KAT SOFTWARE FOR SEQUENATOR NANOFOR SPS DATA PROCESSING

To process the data of the Nanofor SPS sequencer, the KAT program option **k-mer comparison plot** was used. In fact, we get a notion of how many elements of each frequency in the read spectrum were not included in the reference genom (in our case Phix174), included once, included twice, etc.

Figs. 3, 4 show the *k*-mer comparison plot results obtained with the Illumina and Nanofor SPS instruments, respectively. It is significant that Nanofor SPS provides a higher level of coverage of the reference sequence by reads (the centroid of *k*-mers with a unique coverage) in cases of similar characteristics of flow cells.

Fig. 3. Typical *k*-mer comparison plot results of Phix 174 sequencing using Illumina Miseq

Fig. 4. Typical *k*-mer comparison plot results of Phix 174 sequencing using Nanofor SPS

CONCLUSION

Genome assembly projects are costly in both time and investment. Identifying problems with experimental data discovered after assembly can be a real failure. With the K-mer Analysis Toolkit (KAT) researchers can access quality criteria and confirm the results in the earlier stages.

K-mers are small fragments of the original genome with a fixed number of DNA bases. A computer can efficiently work with a large number of *k*-mers and then identify the relations between these fragments to create an idea of the original genome. *K*-mer-based methods are commonly used to efficiently generate genomic assemblies. KAT is built to examine and compare sequencing datasets using the basic properties of each individual *k*-mer, such as frequency and nucleotide composition.

First of all, the KAT can analyze sequencing data to determine the levels of random errors, systematic errors and contamination. The information gained from this analysis can help researchers decide whether to continue with subsequent tasks, such as genome assembly. Then the KAT can re-check the performed assembly of the genome, determining the completeness and accuracy of the assembly without any external reference.

REFERENCES

1. Romanenkov K.V. [A new method of evaluating genome assemblies based on *k*-mers frequencies]. *Preprinty Instituta prikladnoi matematiki im. M.V. Keldysha RAN* [Preprints of the Keldysh Institute of Applied Mathematics], 2017, no. 11, 24 p. DOI: 10.20948/prepr-2017-11 (In Russ.).
2. Marcais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, 2011, vol. 27, is. 6, pp. 764–770. DOI: 10.1093/bioinformatics/btr011
3. Deorowicz S., Kokot M., Grabowski S., Debudaj-Grabysz A. KMC 2: fast and resource-frugal *k*-mer counting. *Bioinformatics*, 2015, vol. 31, is. 10, pp. 1569–1576. DOI: 10.1093/bioinformatics/btv022
4. Erbert M., Rechner S., Müller-Hannemann M. Gerbil: a fast and memory-efficient *k*-mer counter with GPU-support. *Algorithms for Molecular Biology*, 2017, vol. 12, art. num. 9. DOI: 10.1186/s13015-017-0097-9
5. Mapleson D., Accinelli G.G., Kettleborough G., Wright J., Clavijo B.J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, 2017, vol. 33, is. 4, pp. 574–576. DOI: 10.1093/bioinformatics/btw663
6. Alexandrov A.V., Shalyto A.A. [Error correction method for sequencing data with insertions and deletions]. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki* [Scientific and Technical Journal of Information Technologies, Mechanics and Optics], 2016, vol. 16, no. 1, pp. 108–114. DOI: 10.17586/2226-1494-2016-16-1-108-114 (In Russ.).

Contacts: *Borodinov Andrey Gennad'evich,*
borodinov@gmail.com

Article received by the editorial office on 30.12.2021