Аннотация работы Методы оценки геномных сборок и их реализация в пакете прикладных программ QUAST
Авторы: научные сотрудники лаборатории “Центр алгоритмической биотехнологии” Гуревич Алексей Александрович и Михеенко Алла Александровна

Для участия в конкурсе на соискание премии, присуждаемой Санкт-Петербургским государственным университетом за научные труды в 2018 году (в категории “за вклад в науку молодых исследователей”), выдвинуты следующие публикации:

Mikheenko A., Saveliev V., **Gurevich A**. MetaQUAST: evaluation of metagenome assemblies, *Bioinformatics* (2016) 32 (7): 1088-1090.
Mikheenko A., Valin G., Prjibelski A., Saveliev V., **Gurevich A**. Icarus: visualizer for de novo assembly evaluation, *Bioinformatics* (2016) 32 (21): 3321-3323.

Данный цикл работ посвящен решению задачи оценки качества геномных сборок. Геном содержит в себе наследственную информацию, определяющую развитие и жизнедеятельность живых организмов, поэтому знание последовательности генома является необходимым в широком спектре прикладных задач таких как генная инженерия, диагностика генетических заболеваний, производство новых лекарств и многих других. Несмотря на технологические прорывы последних лет, современные устройства все еще не позволяют считать полную последовательность генома за один проход даже для достаточно небольших организмов. Напротив, считываются небольшие фрагменты генома, которые затем вычислительно объединяются между собой в более длинные участки при помощи специальных программ, геномных сборщиков. Результатом работы таких программ является так называемая геномная сборка — черновой вариант последовательности генома организма.

Задача сборки генома из коротких фрагментов является вычислительно крайне сложной из-за большого объема обрабатываемых данных, ошибок в считанных фрагментах, их относительно короткой длины (обычно 100-1000 нуклеотидов — единичных блоков геномной последовательности) при большой длине полного генома (миллионы нуклеотидов для микроорганизмов и миллиарды нуклеотидов для животных и растений), а также высокой концентрации повторяющихся участков в геномах. Как следствие, точного решения данной задачи не существует и современные геномные сборщики используют различные эвристические алгоритмы, приводящие к расхождению в получаемых в итоге геномных сборках. На данный момент существует несколько десятков геномных сборщиков, среди которых нет ярко выраженного лидера, так как все данные программы показывают результаты разной степени достоверности в зависимости от класса исследуемых организмов и используемой технологии получения геномных фрагментов. Поскольку все методы последующего анализа сборок геномов чувствительны к качеству входных данных, крайне актуальной является задача оценки качества представленных геномных сборок и выявление наиболее достоверной из них в каждом конкретном наборе данных.

Представленные на конкурс работы являются расширением программы для оценки качества геномных сборок QUAST (Gurevich et al., Bioinformatics, 2013), которая была создана Гуревичем А.А. в 2012 году в качестве магистерской работы. Основное развитие проекта, которое обусловило его популярность и широкое международное признание, началось в 2015 году с приходом Гуревича А.А. и Михеенко А.А. в лабораторию “Центр алгоритмической биотехнологии” СПбГУ. В рамках представленного цикла работ были разработаны и реализованы расширения MetaQUAST (Mikheenko et al., Bioinformatics, 2016a) и Icarus (Mikheenko et al., Bioinformatics, 2016b), а также существенно улучшен базовый функционал QUAST и выпущена версия программы для работы с геномами большой длины, в том числе геномом человека, QUAST-LG (Mikheenko et al., Bioinformatics, 2018 -- не включена в заявку на конкурс, так как опубликована в 2018 году). Вся разработанная научно-техническая продукция находится в открытом доступе на платформах SourceForge и GitHub, а также доступна в виде веб-сервиса по адресу http://quast.sf.net/wi. Программы QUAST, MetaQUAST и Icarus зарегистрированы в

государственном реестре программ для ЭВМ (правообладатель: СПбГУ), для вышедшей в 2018 году QUAST-LG начата подготовка к процедуре регистрации.

Программа QUAST предназначена для оценки качества сборок бактериальных и небольших эукариотических геномов. В данном продукте реализованы как имевшиеся ранее теоретические наработки по данной задаче, так и новые подходы к оценке характеристик геномных последовательностей. В данной работе впервые введен показатель качества NGA50, который объединил в себе характеристики корректности и протяженности геномных сборок. Одной из важных отличительных особенностей QUAST является возможность работы как с уже хорошо изученными организмами (для которых известна референсная геномная последовательность), так и с теми, для которых геномная последовательность выявляется впервые.

Утилита MetaQUAST расширила возможности QUAST для его применения к изучению микробных сообществ, а не отдельных микроорганизмов. Данный режим работы особенно важен и востребован, так как подавляющее большинство бактерий (до 99%) не могут быть клонированы в лаборатории и основным, а порою и единственно возможным, методом их изучения является анализ образцов почвы, воды, органов животных и т.п., состоящих из совокупности большого числа микробов. В таких исследованиях идет обработка не отдельных геномов, а так называемого обобщенного “метагенома”. При этом существующие утилиты, предназначенные для оценки качества сборок отдельно взятых геномов, не могут быть применены для оценки и сравнения метагеномных сборок, так как они не учитывают всю специфику подобных данных. В MetaQUAST учтены и обрабатываются следующие особенности метагеномных наборов данных: (i) неизвестный видовой состав образца (соответствующие референсные последовательности идентифицируются вычислительными методами и загружаются из Интернета), (ii) огромное разнообразие представленных видов (предоставляются исчерпывающие отчеты отдельно для неограниченного количества геномов) и (iii) наличие близкородственных видов (обнаруживаются химерические соединения в сборках и помечаются как особый вид ошибок). Возможности MetaQUAST были продемонстрированы в статье Mikheenko et al., 2016a при сравнении результатов нескольких ведущих сборщиков на одном смоделированном и двух реальных наборах метагеномных данных.

Программа Icarus стала первым геномным браузером, предназначенным для оценки качества геномных сборок и визуализации ошибок сборки. Визуализация данных играет важную роль в анализе данных секвенирования генома. При этом технологии секвенирования и связанные с ними вычислительные методы развиваются намного быстрее, чем подходы к визуализации производимых ими данных. Для решения данной проблемы, мы разработали утилиту Icarus, которая позволяет в интерактивном режиме посмотреть на качество оцениваемых данных с различных точек зрения. Кроме того, Icarus может быть использован как в исследованиях с известным референсным геномом, так и при изучении новых видов организмов, что подробно продемонстрировано в работе Mikheenko et al., 2016b.

Annotation of work "Methods for genome assembly quality assessment and their implementation in the QUAST toolkit"
Authors: Alexey Alexandrovich Gurevich and Alla Alexandrovna Mikheenko, researchers, the laboratory “Center for Algorithmic Biotechnology”

To participate in the competition for the St. Petersburg State University award for scientific works in 2018 (in the category of “contribution to science made by young researchers”), the following publications are nominated:

Mikheenko A., Saveliev V., **Gurevich A**. MetaQUAST: evaluation of metagenome assemblies, *Bioinformatics* (2016) 32 (7): 1088-1090.
Mikheenko A., Valin G., Prjibelski A., Saveliev V., **Gurevich A**. Icarus: visualizer for de novo assembly evaluation, *Bioinformatics* (2016) 32 (21): 3321-3323.

This series of works is aimed to solve the genome assembly evaluation problem. The genome contains hereditary information that determines the development and vital functions of living organisms, therefore knowledge of the genome sequence is necessary for a wide range of applications such as genetic engineering, diagnosis of genetic diseases, drug development and many others. Despite recent technological breakthroughs, modern DNA sequencing technologies still cannot read the entire genome sequence of a chromosome, even for relatively small organisms. Instead, they generate large numbers of small genome fragments, which are further computationally combined into longer sequences using special software, *genome assemblers*. The output of such programs is called *genome assembly* and it represents a draft genome sequence of the organism.

The problem of assembling a genome from short fragments is computationally extremely difficult due to the large amount of processed data, errors in read fragments, their relatively short length (usually 100-1000 nucleotides — single blocks of the genomic sequence) comparing to a large length of the whole genome (millions of nucleotides for microorganisms and billions of nucleotides for animals and plants), as well as a high abundance of repetitive regions in the genomes. Therefore, there is no exact solution to this problem, and modern genome assemblers use different heuristic algorithms, which lead to discrepancies in the resulting genome assemblies. Dozen genome assemblers have been developed in recent years, but there is no clear winner among them since all programs show the results of varying quality depending on the class of studied organisms and the technology used to obtain genomic fragments. Since all methods of downstream analysis of genome assemblies are sensitive to the quality of input data, it is extremely important to assess the quality of genome assemblies and identify the most reliable one for each particular data set.

The works submitted for the competition represent extensions of QUAST -- quality assessment tool for genome assemblies (Gurevich et al., Bioinformatics, 2013) created by A. A. Gurevich in 2012 as his master's thesis. The main development of the project, which led to its popularity and wide international recognition, began in 2015 with A. Gurevich. and A. Mikheenko coming to the laboratory "Center for Algorithmic Biotechnology" of St. Petersburg State University. As a result of the presented list of works, extensions MetaQUAST (Mikheenko et al., Bioinformatics, 2016a) and Icarus (Mikheenko et al., Bioinformatics, 2016b) were developed and implemented, also the basic QUAST functionality was significantly improved and a version intended to work with large genomes, including the human genome, QUAST-LG was developed (Mikheenko et al., Bioinformatics, 2018 -- not included in the application for the competition, since it was published in 2018). All developed products are publicly available on the SourceForge and GitHub platforms and are also available as a web service at http://quast.sf.net/wi. QUAST, MetaQUAST, and Icarus are registered in the state register of computer programs (copyright holder: SPbU); for QUAST-LG, published in 2018, preparation for the registration procedure has begun.

QUAST is designed to evaluate bacterial and small eukaryotic genome assemblies. This tool implements both previously existing theoretical developments for this problem and new approaches to the assessment of genomic sequences. In this project, the quality metric NGA50 was first introduced, which combined the correctness and contiguity of genome assemblies. One

of the important QUAST features is the ability to work with already well-studied organisms (for which the reference genomic sequence is known), and with those for which the genomic sequence is detected for the first time.

MetaQUAST has extended QUAST to apply it to studies of microbial communities, rather than separate microorganisms. This mode is especially important and relevant, since the vast majority of bacteria (up to 99%) cannot be cloned in the laboratory and the main, and sometimes the only possible, method for studying them is to analyze samples of soil, water, animal organs, etc consisting of a large number of microbes. In such studies, not individual genomes are processed, but the so-called generalized “metagenome”. At the same time, existing utilities designed to assess the assembly quality of individual genomes cannot be used to perform evaluation and comparison between metagenomic assemblies, since these utilities do not take into account the specific features of such data. The following features of metagenomic data sets are taken into account and processed in MetaQUAST: (i) an unknown species composition of the sample (reference sequences are identified using computational methods and downloaded from the Internet), (ii) a huge variety of presented species (comprehensive reports are provided separately for an unlimited number of genomes), and (iii) the presence of closely related species (chimeric connections are found in assemblies and are marked as a special kind of errors). The capabilities of MetaQUAST were demonstrated in the paper Mikheenko et al., 2016a when comparing the results of several leading genome assemblers on one simulated and two real metagenomic data sets.

Icarus became the first genome browser designed to assess the quality of genomic assemblies and visualize assembly errors. Data visualization plays an important role in analyzing genome sequencing data. At the same time, sequencing technologies and related computational methods are developing much faster than approaches to visualizing the data produced by them. To address this problem, we developed the tool allowing to interactively assess the quality of the studied data from different points of view. In addition, Icarus can be used both in studies with a known reference genome, and in studies of new species, as demonstrated in details in Mikheenko et al., 2016b.