

# Мини-обзор генома *Burkholderia seminalis*

Artyom Tyukaev<sup>1</sup>

<sup>1</sup> Факультет биоинженерии и биоинформатики, МГУ им. М. В. Ломоносова

## АБСТРАКТ

В данном исследовании представлен обзор некоторых параметров генома *Burkholderia seminalis*, а именно нуклеотидный состав трех хромосом, анализ белок-кодирующих и РНК-кодирующих генов. Оценена вероятность случайного распределения генов между цепями ДНК. Был рассчитан GC-skew для каждой геномной молекулы. На основе результатов был выдвинут ряд предположений.

## ВВЕДЕНИЕ

Представители рода *Burkholderia* принадлежат к классу  $\beta$ -протеобактерий и имеют широкое распространение, повсеместно обитая в почве, воде и в ассоциации с растениями, грибами, животными и человеком. Некоторые виды *Burkholderia* являются патогенами растений во многих овощах и фруктах, в то время как другие зарегистрированы как условно-патогенные микроорганизмы человека и других животных<sup>[1]</sup>. Однако многие другие виды *Burkholderia* полезны для растений, подавляя болезни растений и стимулируя рост растений с помощью различных процессов, включая производство антибиотиков, секрецию аллелохимических веществ, индукцию устойчивости растений к патогенам, фиксацию азота или усиление поглощения питательных веществ растениями-хозяевами<sup>[2]</sup>.

*B. seminalis* TC3.4.2R3 имеет гены, кодирующие признаки, потенциально участвующие во взаимодействиях растений и микробов. А именно: 1) *B. seminalis* TC3.4.2R3 продуцирует ИУК и имеет гены ферментов биосинтеза ИУК. *Burkholderia* spp. 2) имеет ген ACC-дезаминазы, которая снижает уровень этилена, что индуцирует рост растения. 3) *B. seminalis* TC3.4.2R3 имеет полные консервативные кластеры для сидерофоров пиохелина и орнибактина, позволяющие эффективнее поглощать железо. 4) Геном также кодирует 21 предполагаемый TопВ-зависимый рецептор внешней мембраны, класс белков, которые функционируют при поглощении ресурсов из окружающей среды<sup>[3]</sup>.

## МАТЕРИАЛЫ И МЕТОДЫ

Данные по геному исследуемой бактерии были взяты с сайта Национального Центра Биотехнологической информации (NCBI). Для расчета нуклеотидного состава использовался язык программирования Python 3.9 и сервис Google colab. Для расчета параметра GC-skew использовался сервис Webskew. Чтобы рассчитать GC-skew для каждой из хромосом использовался скрипт seqret из пакета EMBOSS на сервере kodo ФББ МГУ, который разбивает файл генома на файлы с отдельными последовательностями. Для анализа данных использовались электронные таблицы Google Sheets.

## РЕЗУЛЬТАТЫ

### 1. Данные о геноме бактерии *Burkholderia seminalis* strain FL-5-4-10-S1-D7

Геном *Burkholderia seminalis* представлен тремя хромосомами. Информация о длине хромосом в п.н. бралась из assembly stats *Burkholderia seminalis* из базы данных NCBI (1). Как видно из данных, наименьшую длину имеет третья хромосома (1 093 546 п.н.). Общая длина генома бактерии составляет 7 648 944 п.н. Всего в геноме 6969 генов (включая белок-некодирующие последовательности)

Таблица 1. Данные о длине хромосом

| ДНК         | Длина (п.н.) |
|-------------|--------------|
| Хромосома 1 | 3 505 262    |
| Хромосома 2 | 3 050 136    |
| Хромосома 3 | 1 093 546    |

Для каждой из хромосом определен процентное содержание GC нуклеотидов (gc-перс), который рассчитывается по формуле (количество-GC-нуклеотидов) / (количество-ATCG-нуклеотидов) представленное в процентах. Расчет производился при помощи кода, написанного на Python (2).

Как видно по таблице, отличие в процентном содержании GC-нуклеотидов у второй хромосомы незначительно, и в среднем по всему геному gc-перс равен 67%.

Таблица 2. Процентное содержание GC-нуклеотидов

| ДНК             | Содержание GC-нуклеотидов (%) |
|-----------------|-------------------------------|
| Хромосома 1     | 67                            |
| Хромосома 2     | 68                            |
| Хромосома 3     | 67                            |
| Общее по геному | 67                            |

Высокое содержание GC-пар обеспечивает поддержание стабильности структуры ДНК при высоких температурах или других экстремальных для клетки условиях.

### 2. Гистограмма длин белков

Была составлена гистограмма длин белков *Burkholderia seminalis*. Судя по полученным данным, наиболее часто встречающаяся длина белкового продукта находится в диапазоне 100-400 а.к. Длину менее 60 а.к. имеет очень маленькое количество белков, и после значения длины,

равного 30, происходит резкий скачок. Можно выделить небольшой провал в диапазоне 120-210 с локальным минимумом в значении 180. Максимальная длина белкового продукта составляет 4503 а.к. Минимальная - 29 а.к. (4, лист Histogram)

### Гистограмма длин белков

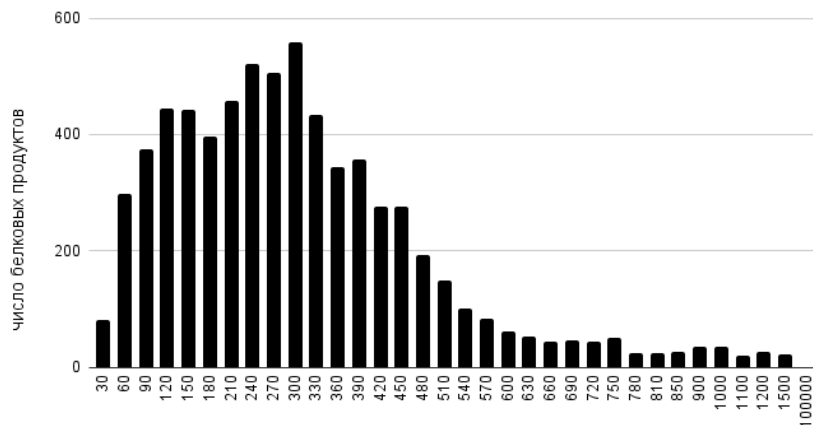


Fig. 1. Гистограмма для длин белковых продуктов кодирующих последовательностей.

### 3. Распределение генов белков между прямой и комплементарной цепями

Исходя из полученных данных (3, лист protein strain), в среднем по геному на прямой и комплементарной цепочке закодировано примерно равное количество белков.

Число генов белков на прямой и комплементарной цепочке

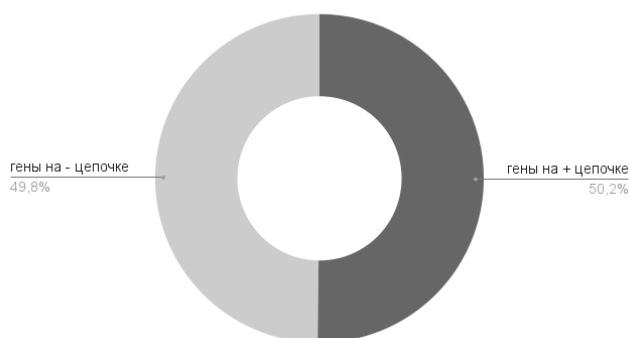


Fig. 2. Распределение белко-кодирующих генов между прямой и комплементарной цепочкой (+ and - strain).

### 4. Оценка случайности распределения генов по цепям ДНК

Для оценки случайности распределения генов на прямой и комплементарной цепи для каждой из хромосом мною был рассчитан статистический критерий Пирсона (хи-квадрат).

Используемый уровень значимости - 0.05. Поставим нулевую гипотезу о случайном распределении генов между цепями. Предельный уровень значимости при наличии одной степени свободы равен 3,84. Более подробно данные представлены в сопроводительных

материалах (3, лист хи-квадрат). Как мы видим, для третьей хромосомы гены распределены случайно (Таблица 7), так как значение критерия хи-квадрат меньше порогового. Для первой и второй хромосомы наблюдаются отклонения от случайного распределения (Таблица 5, Таблица 6).

Таблица 5. Расчет критерия Пирсона для Хромосомы 1

| Хромосома 1   | Количество генов (Ожидаемое) | Количество генов (Наблюдаемое) | Хи квадрат  |
|---------------|------------------------------|--------------------------------|-------------|
| Прямая цепь   | 3244                         | 3402                           | 15,39087546 |
| Обратная цепь | 3244                         | 3086                           |             |
|               |                              |                                |             |

Таблица 6. Расчет критерия Пирсона для Хромосомы 2

| Хромосома 2   | Количество генов (Ожидаемое) | Количество генов (Наблюдаемое) | Хи квадрат  |
|---------------|------------------------------|--------------------------------|-------------|
| Прямая цепь   | 2750                         | 2642                           | 8,482909091 |
| Обратная цепь | 2750                         | 2858                           |             |
|               |                              |                                |             |

Таблица 7. Расчет критерия Пирсона для Хромосомы 3

| Хромосома 3   | Количество генов (Ожидаемое) | Количество генов (Наблюдаемое) | Хи квадрат  |
|---------------|------------------------------|--------------------------------|-------------|
| Прямая цепь   | 975                          | 934                            | 3,448205128 |
| Обратная цепь | 975                          | 1016                           |             |
|               |                              |                                |             |

## 5. Типы белков протеома

1) Всего в геноме белок кодируют 6802 гена, из них 61 кодирует рибосомальные белки, то есть на их долю приходится 0,897% всех белок-кодирующих генов. Полный список белков с наименованиями содержится в сопроводительных материалах. (3, лист `ribosomal_proteins`)

2) Довольно значительная часть генов кодирует белки, функция которых не определена и существование которых под сомнением. Всего таких белков 730. В процентном соотношении о всех белок-кодирующих генов они составляют 10,7%. Полный список неопределенных белков находится в листе `hypotetical_protein`. Среди таких белков есть некоторые, которые имеют размер от 300 до 700 а.к., на основании чего можно сделать вывод о том, что значительная функциональная часть протеома исследуемой бактерии остается неисследованной. (3, лист `hypothetical_protein`)

3) Генов, кодирующих транспортные белки 804. В процентах они составляют 11,8% от всех белок-кодирующих последовательностей. (3, лист transport\_protein)

## 6. Статистические данные о генах РНК

В протеоме *Burkholderia seminalis* представлено 89 генов, кодирующих РНК, включая различные ее типы: rRNA, tRNA, ncRNA, tmRNA. От всех генов это составляет 1,28%. Это заметно меньше, чем количество белок-кодирующих генов. Соотношение РНК-кодирующих генов к белок-кодирующих равно 0,013, что соотносится примерно как 1 к 75. (3, лист RNA\_total)

Число генов рибосомальной РНК (rRNA) составляет 19, в процентах от всех РНК это 21,4%. Мною был обнаружен интересный факт особенности распределения генов РНК между тремя хромосомами исследуемой бактерии. Вторая и третья хромосомы несут по одной копии каждой rRNA, входящей в состав прокариотической рибосомы (5S RNA, 16S RNA, 23S RNA). В это же время первая хромосома несет по 3 копии каждой из вышеперечисленных RNA. Соотношение количества генов rRNA на первой хромосоме и ее длины не пропорционально соотносится с аналогичными параметрами для второй хромосомы, так как при относительно небольшой разнице в длине (500 тыс. нуклеотидов), первая хромосома несет в 3 раза больше копий генов rRNA. (3, лист rRNA)

Число генов транспортной РНК (tRNA) составляет 67. Это 75% от всех генов РНК в геноме. Причем 90% всех генов tRNA располагается на первой хромосоме (для сравнения третья хромосома кодирует только 3% всех tRNA). (3, лист tRNA)

Наибольшее количество генов среди РНК-кодирующих принадлежит генам tRNA, которые необходимы для синтеза белка. (3, лист RNA\_stat)

Таблица 3. Данные о РНК

| РНК  | Количество | Процент от всех РНК |
|------|------------|---------------------|
| рРНК | 19         | 21 %                |
| тРНК | 67         | 75 %                |

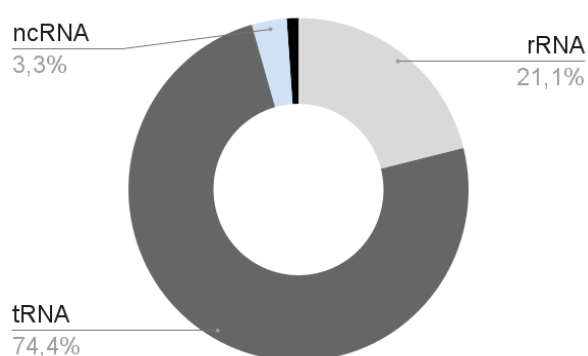


Fig. 3. Соотношение количества генов РНК в геноме.

## 7. Нуклеотидный состав геномной ДНК

С помощью программы на Python было посчитано количество нуклеотидов А, Т, С, G для каждой из хромосом и генома (2). Соотношение нуклеотидов А и Т, С и G согласуется с правилом Чаргаффа с погрешностью 0,32% для нуклеотидов G и C, что может быть связано с наличием модифицированных оснований в геноме.

## 8. Поиск участка начала (oriC) и терминации (ter) репликации

Был показан факт того, что с помощью вычисления соотношения между нуклеотидами G и C (так называемый GC-skew), можно определить место начала репликации (origin, далее - oriC) и ее конца (terminator, далее - ter). Место минимального значения GC-skew соответствует oriC, а максимального - ter. Действительно, ведь малая доля GC-пар в месте начала репликации значительно упрощает расплетение цепочки ДНК, что обеспечивает успешное формирование репликационных вилок. Высокий GC-состав в свою очередь будет способствовать терминации.

формула для вычисления GC-skew в определенном интервале:

$$\text{GC-skew} = (\#C - \#G) / (\#C + \#G)$$

где #C, #G - количество нуклеотидов C и G соответственно.

При расчете GC-skew на определенном участке интересующей нас последовательности получаются значения, которые затем взятием интеграла функции GC-skew образуют линию графика, минимум и максимум которого соответствуют расположению oriC и ter. Burkholderia seminalis имеет три хромосомы, поэтому поиск oriC и ter осуществлялся для каждой из них. Для расчёта GC-skew использовался сервис webskew (5).

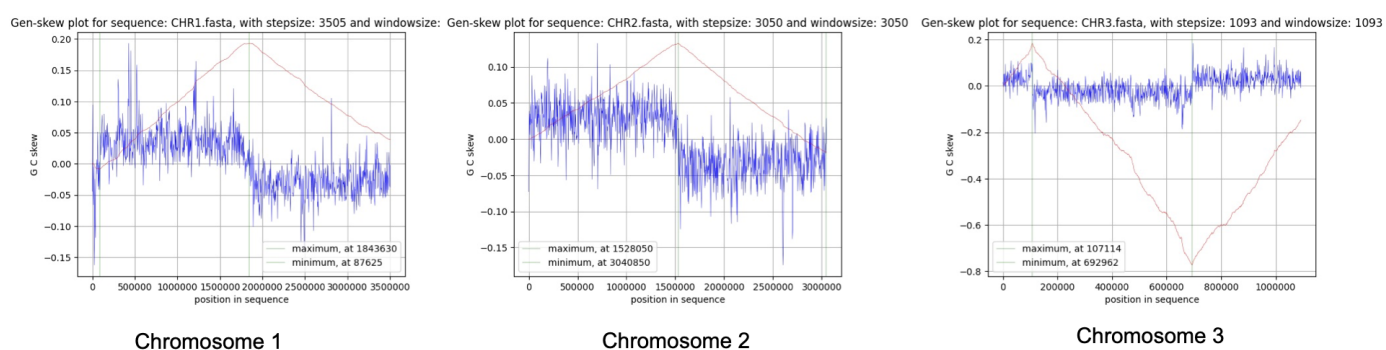


Fig. 4. график параметра GC-skew для хромосом Burkholderia seminalis.

Таким образом, предполагаемое расположение oriC и ter для каждой хромосомы представлены в таблице 4.

Таблица 4. oriC и ter хромосом

| Хромосома   | Положение oriC (в окрестности нуклеотида) | Положение ter (в окрестности нуклеотида) |
|-------------|-------------------------------------------|------------------------------------------|
| Хромосома 1 | 87 625                                    | 1 843 630                                |
| Хромосома 2 | 3 040 850                                 | 1 528 050                                |
| Хромосома 3 | 692 962                                   | 107 114                                  |

## СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

1. Данные NCBI по геному бактерии *Burkholderia seminalis*  
[https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/718/535/GCF\\_001718535.1\\_ASM171853v/](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/718/535/GCF_001718535.1_ASM171853v/)
2. Код на Python для вычисления количества нуклеотидов и содержания GC-нуклеотидов  
<https://colab.research.google.com/drive/15hG2LABKM7hfMtZeDeJtOf6G94YdwtTV#scrollTo=QSo3kYLgu3dQ>
3. Google таблица с исследованием  
<https://docs.google.com/spreadsheets/d/1Ph2VqYGbyz4LUOxfEA631vTBg1ZhkiRI60ioHqRtXfM/edit?usp=sharing>
4. Google таблица с гистограммой длин белковых продуктов  
[https://docs.google.com/spreadsheets/d/1ToCCiWqTKR4F6HjoyOCHkqQffNjUx\\_OFBItejuMKes/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1ToCCiWqTKR4F6HjoyOCHkqQffNjUx_OFBItejuMKes/edit?usp=sharing)
5. Сервис для расчета GC-skew  
<https://genskew.csb.univie.ac.at/webskew>

## СПИСОК ЛИТЕРАТУРЫ

- [1] Elshafie, H.S.; Camele, I. An overview of metabolic activity, beneficial and pathogenic aspects of *Burkholderia* spp. *Metabolites* **2021**, *11*, 321.
- [2] Hau-Hsuan Hwang, Pei-Ru Chien, Fan-Chen Huang, Shih-Hsun Hung, Chih-Horng Kuo, Wen-Ling Deng, En-Pei Isabel Chiang, Chieh-Chen Huang. A Plant Endophytic Bacterium, *Burkholderia seminalis* Strain 869T2, Promotes Plant Growth in Arabidopsis, Pak Choi, Chinese Amaranth, Lettuces, and Other Vegetables. *Microorganisms* **2021**, *9*(8), 1703.
- [3] Wellington L. Araújo, Allison L. Creason, Emy T. Mano, Aline A. Camargo-Neves, Sonia N. Minami, Jeff H. Chang and Joyce E. Loper. Genome Sequencing and Transposon Mutagenesis of *Burkholderia seminalis* TC3.4.2R3 Identify Genes Contributing to Suppression of Orchid Necrosis Caused by *B. gladioli*. *Molecular Plant-Microbe Interactions* **2016**, ISSN 0894-0282