

# АРТЕФАКТЫ СЕКВЕНИРОВАНИЯ В ДАННЫХ OXFORD NANOPORE TECHNOLOGY: ЛОЖНОПОЛОЖИТЕЛЬНОЕ ОБНАРУЖЕНИЕ NAEGLERIA FOWLERI И ЕГО БИОИНФОРМАТИЧЕСКАЯ ВЕРИФИКАЦИЯ

## OXFORD NANOPORE TECHNOLOGY DATA SEQUENCING ARTIFACTS: FALSE-POSITIVE DETECTION OF NAEGLERIA FOWLERI AND ITS BIOINFORMATICS VALIDATION

A. Zimina  
A. Grinev

**Summary.** In this work, a comprehensive analysis of sequencing artifacts generated using Oxford Nanopore Technology (ONT) is presented, exemplified by the erroneous detection of the free-living amoeba *Naegleria fowleri* in blood samples from patients with malaria (*Plasmodium falciparum*). In all four examined samples, sequences demonstrating nearly 100 % identity with the *N. fowleri* genome according to BLAST analysis were identified. However, subsequent multi-level verification, including phylogenetic analysis, open reading frame (ORF) prediction, and three-dimensional protein structure modeling using AlphaFold3, made it possible to establish the artifactual nature of these findings. The results obtained underscore the critical importance of comprehensive validation of ONT data in clinical studies, where false-positive results can lead to serious diagnostic errors. Particular attention is paid to the analysis of potential sources of artifacts, including systematic sequencing errors, characteristics of bioinformatic data processing, and issues with reference databases.

**Keywords:** Oxford Nanopore Technology, sequencing artifacts, *Naegleria fowleri*, bioinformatic analysis, false-positive results, primary amoebic meningoencephalitis, AlphaFold3, BLAST analysis.

**Зими́на Анастасия Алексе́евна**

Стажёр-исследователь,  
ФГАОУ ВО Первый Московский государственный  
медицинский университет им. И.М. Сеченова  
zimina\_a\_a@staff.sechenov.ru

**Гри́нев Алексе́андр Бори́сович**

Кандидат биологических наук,  
ФГАОУ ВО Первый Московский государственный  
медицинский университет им. И.М. Сеченова  
sagrinev@yandex.ru

**Аннотация.** В данной работе представлен комплексный анализ артефактов секвенирования, возникающих при использовании технологии Oxford Nanopore Technology (ONT), на примере ошибочного обнаружения свободноживущей амёбы *Naegleria fowleri* в образцах крови пациентов с малярией (*Plasmodium falciparum*). Во всех четырёх исследованных образцах были идентифицированы последовательности, демонстрирующие практически 100 % идентичность с геномом *N. fowleri* по данным BLAST-анализа. Однако последующая многоуровневая верификация, включающая филогенетический анализ, предсказание открытых рамок считывания (ORF) и моделирование трёхмерной структуры белков с использованием AlphaFold3, позволила установить артефактную природу этих находок. Полученные результаты подчёркивают критическую важность комплексной валидации данных ONT в клинических исследованиях, где ложноположительные результаты могут привести к серьёзным диагностическим ошибкам. Особое внимание уделено анализу возможных источников артефактов, включая систематические ошибки секвенирования, особенности биоинформатической обработки данных и проблемы референсных баз.

**Ключевые слова:** Oxford Nanopore Technology, артефакты секвенирования, *Naegleria fowleri*, биоинформатический анализ, ложноположительные результаты, первичный амёбный менингоэнцефалит, AlphaFold3, BLAST-анализ.

## Введение

Современные технологии секвенирования нового поколения, в частности Oxford Nanopore Technology (ONT), совершили революцию в геномных исследованиях благодаря своей способности генерировать сверхдлинные прочтения ДНК/РНК. Эта особенность делает ONT незаменимым инструментом для анализа структурных вариаций генома, сборки *de novo* и метагеномных исследований. Однако, несмотря на значительные преимущества, технология имеет существенное ограничение — относительно высокий

уровень ошибок по сравнению с платформами коротких чтений, такими как Illumina (средний Q-score около 15 против 30+ у Illumina).

В рамках нашего исследования, посвящённого изучению генетических особенностей *Plasmodium falciparum* у пациентов с малярией, мы столкнулись с неожиданным артефактом — последовательностями, идентифицируемыми как *Naegleria fowleri*. Этот свободноживущий протист известен как возбудитель первичного амёбного менингоэнцефалита (ПАМ) — редкого, но крайне тяжёлого заболевания центральной нервной системы. Биологи-

ческая вероятность обнаружения *N. fowleri* в образцах периферической крови пациентов с малярией крайне мала, так как этот патоген обычно локализуется в спинномозговой жидкости и тканях головного мозга, проникая в организм через обонятельный нерв при контакте с заражённой водой.

Данное наблюдение побудило нас провести всестороннее исследование природы этих последовательностей. Мы выдвинули гипотезу, что обнаружение *N. fowleri* является артефактом, связанным либо с особенностями технологии ONT, либо с ограничениями биоинформатических методов анализа. Для проверки этой гипотезы был применён комплексный подход, включающий филогенетический анализ, оценку функционального потенциала выявленных последовательностей и методы предсказания структуры белков.

### Материалы и методы

#### Характеристика биологического материала и секвенирование

Исследование проведено на образцах цельной крови, полученных от четырёх пациентов с подтверждённым диагнозом малярии, вызванной *Plasmodium falciparum*. Все пациенты были заражены в эндемичных регионах и не имели клинических признаков других инфекционных заболеваний. Выделение тотальной ДНК проводилось с использованием стандартных протоколов с последующей подготовкой библиотек для секвенирования на платформе MinION (Oxford Nanopore Technologies) в соответствии с рекомендациями производителя.

#### Биоинформатическая обработка данных

Первичная обработка сырых данных включала фильтрацию по качеству (Q-score > 7), обрезку адаптерных последовательностей и удаление низкокачественных прочтений с использованием программы Porechop. Для таксономической аннотации последовательности выравнивались с помощью BLASTn против базы данных NCBI nt (рис. 1). Для дополнительной верификации результатов был применён филогенетический анализ с построением дендрограмм на основе референсных последовательностей *N. fowleri*, поиск открытых рамок считывания (ORF) в выявленных последовательностях, предсказание структуры кодируемых белков с использованием AlphaFold3, визуализация и анализ предсказанных структур в программе PyMOL с оценкой параметров достоверности (pLDDT).

#### Сравнительный анализ с референсными данными

Для исключения возможности контаминации проведено детальное сравнение с геномом *N. fowleri* (изоляты

ATCC 30894 и ATCC 30863), клиническими изолятами *N. fowleri* из Таиланда [2], референсными последовательностями других видов *Naegleria*. Особое внимание уделено анализу консервативных участков, включая гены рибосомальных РНК и высококонсервативных белков цитоскелета.

### Результаты

#### Обнаружение последовательностей *N. fowleri*

Анализ данных ONT-секвенирования выявил присутствие последовательностей, соответствующих *N. fowleri*, во всех четырёх исследованных образцах. BLASTn-анализ показал 100 % идентичность с референсными последовательностями *N. fowleri* из базы данных NCBI. Однако несколько факторов вызвали сомнения в биологической достоверности этого результата:

1. Отсутствие клинических признаков ПАМ у пациентов.
2. Крайне низкая вероятность обнаружения *N. fowleri* в периферической крови.
3. Отсутствие эпидемиологических данных, указывающих на возможное сочетанное заражение сразу в четырёх разных регионах.

#### Филогенетический анализ

Филогенетическое дерево, построенное на основе последовательностей 12S рРНК и ITS1, показало, что выявленные последовательности формируют отдельную кладу, не соответствующую известным клиническим изолятам *N. fowleri* (рис. 2). Особенно показательным было отсутствие консервативных участков, характерных для патогенных штаммов этого вида.

#### Анализ белкового потенциала

ORF-анализ выявил несколько потенциальных белков, однако их функциональная аннотация показала отсутствие типичных для *N. fowleri* патоген-ассоциированных белков (например, Naegleriapore A), низкую консервативность предсказанных последовательностей и преобладание гипотетических белков с неизвестной функцией.

#### Моделирование структуры белков

Предсказание структуры с помощью AlphaFold3 и последующий анализ в PyMOL выявили низкие показатели достоверности (pLDDT < 50 для большинства доменов), отсутствие стабильных третичных структур и несоответствие предсказанных структур известным белкам *N. fowleri* (рис. 3).

Человек 1

| Description                                                                                                             | Scientific Name                   | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession                  |
|-------------------------------------------------------------------------------------------------------------------------|-----------------------------------|-----------|-------------|-------------|---------|------------|----------|----------------------------|
| <input checked="" type="checkbox"/> <a href="#">Naegleria fowleri Karachi_NF001 strain</a>                              | <a href="#">Naegleria fowleri</a> | 4614      | 9197        | 100%        | 0.0     | 99.45%     | 2541     | <a href="#">OD957322.1</a> |
| <input checked="" type="checkbox"/> <a href="#">Homo sapiens clone fosmid ABC8_4_1_000041143400_G2 genomic sequence</a> | <a href="#">Homo sapiens</a>      | 3940      | 8368        | 100%        | 0.0     | 89.20%     | 14353    | <a href="#">GQ222231.1</a> |
| <input checked="" type="checkbox"/> <a href="#">Homo sapiens clone BMMD-RP11_130B16_complete sequence</a>               | <a href="#">Homo sapiens</a>      | 2543      | 10360       | 100%        | 0.0     | 79.67%     | 188886   | <a href="#">KJ155474.1</a> |
| <input checked="" type="checkbox"/> <a href="#">Homo sapiens chromosome 22 clone c31p7 map 22q11_complete sequence</a>  | <a href="#">Homo sapiens</a>      | 2394      | 8821        | 100%        | 0.0     | 80.62%     | 28862    | <a href="#">AC012398.4</a> |
| <input checked="" type="checkbox"/> <a href="#">Pan troglodytes chromosome 22 clone CH251-171M16_complete sequence</a>  | <a href="#">Pan troglodytes</a>   | 2355      | 7913        | 100%        | 0.0     | 80.56%     | 153641   | <a href="#">AC275820.1</a> |

Человек 2

| Description                                                                         | Scientific Name                   | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession                  |
|-------------------------------------------------------------------------------------|-----------------------------------|-----------|-------------|-------------|---------|------------|----------|----------------------------|
| <a href="#">Naegleria fowleri Karachi_NF001 strain</a>                              | <a href="#">Naegleria fowleri</a> | 4614      | 9197        | 99%         | 0.0     | 99.45%     | 2541     | <a href="#">OD957322.1</a> |
| <a href="#">Homo sapiens clone fosmid ABC8_4_1_000041143400_G2 genomic sequence</a> | <a href="#">Homo sapiens</a>      | 3940      | 8368        | 99%         | 0.0     | 89.20%     | 14353    | <a href="#">GQ222231.1</a> |
| <a href="#">Homo sapiens clone BMMD-RP11_130B16_complete sequence</a>               | <a href="#">Homo sapiens</a>      | 2543      | 10360       | 99%         | 0.0     | 79.67%     | 188886   | <a href="#">KJ155474.1</a> |

Человек 3

| Description                                                                                | Scientific Name                   | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession                  |
|--------------------------------------------------------------------------------------------|-----------------------------------|-----------|-------------|-------------|---------|------------|----------|----------------------------|
| <input checked="" type="checkbox"/> <a href="#">Naegleria fowleri Karachi_NF001 strain</a> | <a href="#">Naegleria fowleri</a> | 3807      | 3807        | 100%        | 0.0     | 98.43%     | 5173     | <a href="#">OD956389.1</a> |
| <input checked="" type="checkbox"/> <a href="#">Naegleria fowleri Karachi_NF001 strain</a> | <a href="#">Naegleria fowleri</a> | 278       | 278         | 14%         | 5e-73   | 83.66%     | 332      | <a href="#">OD957990.1</a> |

Человек 4

Naegleria fowleri Karachi\_NF001 strain

Sequence ID: [OD956389.1](#) Length: 5173 Number of Matches: 1

Range 1: 2649 to 2903 [GenBank](#) [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#)

| Score         | Expect                                                        | Identities   | Gaps      | Strand     |
|---------------|---------------------------------------------------------------|--------------|-----------|------------|
| 261 bits(141) | 4e-69                                                         | 217/255(85%) | 0/255(0%) | Plus/Minus |
| Query 3       | CAAATAGATGCAATAAAAAATGATAAAGGGGATATCACCACAGATGCCACAGAAATTCAA  | 62           |           |            |
| Sbjct 2903    | CAAATAGACACAATAAAAAATGATAAAGGGGATATCACCACCGATCCCACAGAAATACAA  | 2844         |           |            |
| Query 63      | ACCATCATCAGAGATTATTACAAACAACCTCTATGCACATAAACTAGAAAACCTGGAAGAA | 122          |           |            |
| Sbjct 2843    | ACTACCATCAGAGAATACTACAAACACCTCTACGCAATAAACTAGAAAATCTAGAAGAA   | 2784         |           |            |
| Query 123     | ATGGATAAATTCCTGGACACTTGACCCCTCCCAAGCCTAAACCAGGAAGAAGTCAAAACC  | 182          |           |            |
| Sbjct 2783    | ATGGATAAATTCCTCGACACATACACCCTCCCAAGACTAAACCAGGAAGAAGTTGAATCT  | 2724         |           |            |
| Query 183     | CTGAATAGACGAATAATAAGGTCTGAAGTTGAGACAGCAATTAAGAGCCTACCACCCaaa  | 242          |           |            |
| Sbjct 2723    | CTGAATAGACCAATAACAGGCTCTGAAATTGTGGCAATAATCAATAGCTTACCAACCAAA  | 2664         |           |            |
| Query 243     | aaaaGCCCAGGTCCA                                               | 257          |           |            |
| Sbjct 2663    | AAGAGTCCAGGACCA                                               | 2649         |           |            |

Рис. 1. Результаты BLASTn-анализа последовательностей, идентифицированных как *N. fowleri*

| GOLD Organism ID          | Organism Name                                 | Phylum                 | NCBI Taxonomy ID     |
|---------------------------|-----------------------------------------------|------------------------|----------------------|
| <a href="#">Go0572210</a> | <a href="#">Naegleria fowleri NF_KFSHRC_1</a> | PROTISTS-HETEROLOBOSEA | <a href="#">5763</a> |
| <a href="#">Go0572156</a> | <a href="#">Naegleria fowleri ATCC 30894</a>  | PROTISTS-HETEROLOBOSEA | <a href="#">5763</a> |
| <a href="#">Go0384989</a> | <a href="#">Naegleria fowleri ATCC 30863</a>  | PROTISTS-HETEROLOBOSEA | <a href="#">5763</a> |
| <a href="#">Go0047971</a> | <a href="#">Naegleria fowleri</a>             | PROTISTS-HETEROLOBOSEA | <a href="#">5763</a> |

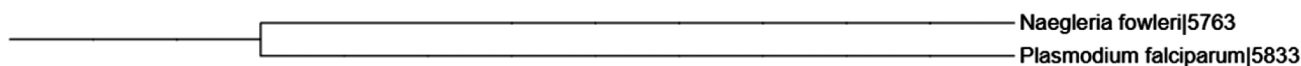


Рис. 2. Филогенетическое дерево на основе последовательностей 12S рРНК

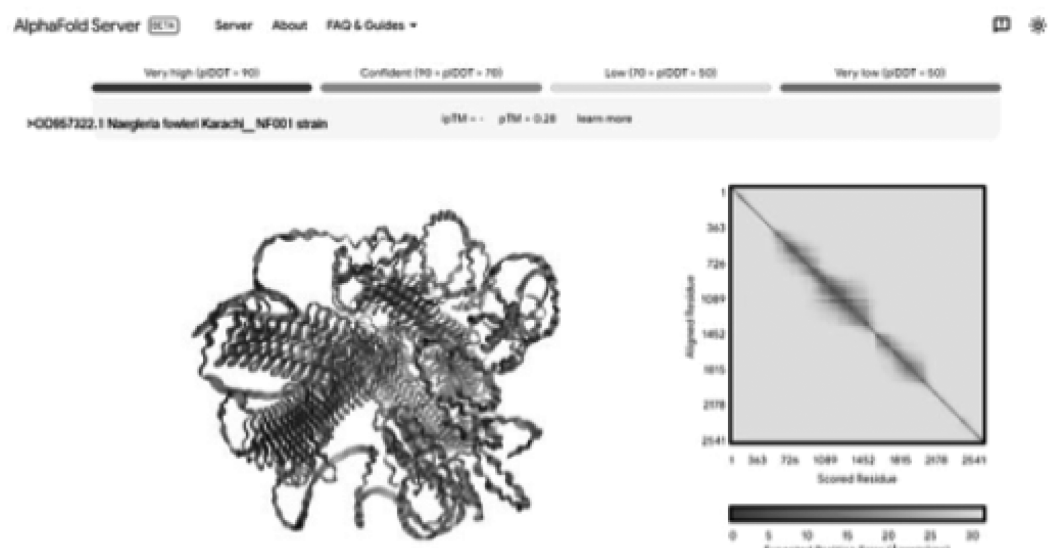


Рис. 3. Моделирование структуры белков с использованием AlphaFold3

## Обсуждение

Полученные результаты убедительно свидетельствуют, что обнаружение последовательностей *N. fowleri* в исследуемых образцах является артефактом, связанным с особенностями технологии ONT. Мы выделяем три основных возможных источника таких артефактов:

1. **Систематические ошибки ONT:** высокая частота инсерций/делеций, ошибки при фазировании гомополимерных участков, низкая точность определения нуклеотидной последовательности.
2. **Ограничения биоинформатических методов:** проблемы выравнивания коротких участков гомологии, ошибки аннотации в референсных базах, ограничения алгоритмов BLAST при работе с неполными и смешанными последовательностями.
3. **Биологические факторы:** наличие химерных последовательностей, горизонтальный перенос генов, наличие консервативных доменов у различных микроорганизмов.

Особое значение имеет сравнение с работой [1], где было показано, что даже высококачественные сборки генома *N. fowleri* требуют тщательной валидации. Наши результаты также согласуются с данными [2], подчёркивающими важность комплексного подхода при идентификации патогенных штаммов *N. fowleri*.

## Заключение

Проведённое исследование демонстрирует, что обнаружение последовательностей *N. fowleri* в образцах крови пациентов с малярией является артефактом, связанным с особенностями технологии ONT и обработки данных. Полученные результаты подчёркивают необходимость:

1. Разработки стандартизированных протоколов валидации данных ONT.
2. Использования дополнительных методов подтверждения наличия паразита у пациента (ПЦР, секвенирование на платформе Illumina).
3. Критической оценки клинического контекста при интерпретации результатов.
4. Совершенствования биоинформатических методов анализа длинных чтений.

Эти меры особенно важны в клинической диагностике, где ложноположительные результаты могут иметь серьёзные последствия для пациентов. Дальнейшие исследования должны быть направлены на разработку алгоритмов автоматического выявления и фильтрации подобных артефактов.

## ЛИТЕРАТУРА

1. Liechti N., Schürch N., Bruggmann R., Wittwer M. Nanopore sequencing improves the draft genome of the human pathogenic amoeba *Naegleria fowleri*. *Sci Rep.* 2019; 9:16040. DOI: 10.1038/s41598-019-52572-0.
2. Soontrapa P. et al. The First Molecular Genotyping of *Naegleria fowleri* Causing Primary Amebic Meningoencephalitis in Thailand. *Front Cell Infect Microbiol.* 2022; 12:931546. DOI: 10.3389/fcimb.2022.931546.
3. Харди А., Дюаркур С., Дефранс М. Фильтрация и анализ шаблонов модификации ДНК с использованием DNAModAnnot. *Методы Mol Biol.* 2023; 2624:87–114.
4. Kheirkhah O. et al. Population genetic structure of the fibrinogen-related protein 1 (FREP1) in Iranian isolates of *Anopheles stephensi*. *Gene.* 2025; 953:149436. DOI: 10.1016/j.gene.2025.149436.