МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра микробиологии и физиологии растений

СОЗДАНИЕ БАЗЫ ДАННЫХ ДЛЯ МУЛЬТИЛОКУСНОГО ТИПИРОВАНИЯ *LEGIONELLA PNEUMOPHILA*

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 3 курса 331 группы
направления 06.04.01 Биология
биологического факультета
Катышева Александра Дмитриевича

Научный руководитель
д-р биол. наук, доц. 26.11.25 Д.В. Уткин

Научный консультант
зав. отделом диагностики инфекционных болезней
ФКУН Российский противочумный
институт «Микроб»
Роспотребнадзора, канд. биол. наук С.А. Портенко

Заведующий кафедрой
д-р биол. наук, доц. 26.11.25. Д.В. Уткин

Саратов 2025

Введение

Актуальность темы. Все большее внимание в лабораторной диагностике и эпидемиологическом надзоре за инфекционными заболеваниями уделяется молекулярно-генетическим методам исследования. Методы, основанные на секвенировании нуклеиновых кислот, такие как мультилокусное сиквенс-типирование (MLST), анализ числа вариабельных тандемных повторов (MLVA, multiple locus variable-number tandem repeats analysis), определение однонуклеотидных полиморфизмов (SNP, single nucleotide polymorphism) и полногеномное секвенирование, играют особую роль в выявлении и описании генетической структуры инфекционных агентов, а также в мониторинге их генетической вариабельности и распространенности [1].

С появлением в начале 21 века платформ второго поколения для полногеномного секвенирования, а с 2010 года платформ третьего поколения, численность депонированных нуклеотидных последовательностей генов, локусов и полных геномов стремительно растет. Для бактерий *Salmonella* spp. число последовательностей полных геномов на сегодняшний день составляет более 530 тысяч, *E. coli* — более 260 тысяч, для *L. pneumophila* более 5 тысяч [2]. Для изучения свойств возбудителей на основе данных, полученных методом полногеномного секвенирования, разработаны стандартные молекулярно-генетические подходы, такие как: мультилокусное сиквенс-типирование, SNP-типирование.

В настоящее время эти подходы не могут быть использованы в полной мере, так как базы данных последовательностей, доступные ранее, сейчас заблокированы по ряду причин (ресурсы прекратили работу по экономическим или техническим причинам; пандемия COVID-19; санкции в связи с политической ситуацией в мире и т. д).

Цель и задачи исследования. Целью данной работы стала разработка и апробация схемы мультилокусного сиквенс-типирования штаммов *L. pneumophila* с использованием биоинформатических методов.

Для решения указанной цели были определены следующие задачи:

1. Анализ нуклеотидных последовательностей генов-мишеней, перспективных для мультилокусного сиквенс-типирования (MLST) штаммов *L. pneumophila*, представленных в базе данных NCBI GenBank, отбор и верификация генов-мишеней с помощью биоинформационных подходов.

2. Апробация разработанной схемы MLST-типирования на выборке, состоящей из геномов 24 штаммов *L. pneumophila* из Государственной коллекции патогенных бактерий ФКУН Российский противочумный институт «Микроб» Роспотребнадзора.

3. Разработка программы для загрузки полногеномных последовательностей, соответствующих заданным параметрам, из базы данных NCBI GenBank.

Материалы и методы исследования. Объектом исследования являлись 5657 нуклеотидных последовательности полных геномов изолятов *L. pneumophila* депонированных в международную базу данных NCBI GenBank, выделенных в разное время депонированные до 8 апреля 2024. Так же в работе были использованы 24 полногеномных последовательности штаммов из *L. pneumophila* из Государственной коллекции патогенных бактерий при ФКУН Российский противочумный институт «Микроб» Роспотребнадзора.

В работе были использованы молекулярно-генетические и биоинформатические методы исследования.

Молекулярно-генетические методы

Секвенирование ДНК производилось на базе лаборатории геномного и протеомного анализа, отдела микробиологии ФКУН Российский противочумный институт «Микроб» Роспотребнадзора на генетическом анализаторе: MGI (DNBSEQ-G50, Китай) с использованием наборов для пробоподготовки (MGIEasy FS DNA Library Prep Set) и секвенирования ДНК (DNBSEQ-G50 High-throughput Sequencing Set, PE150), а также на генетическом анализаторе Ion PGM (Ion Torrent). Среднее покрытие в зависимости от штамма варьировало от 50-кратного до 89-кратного. Последовательности были депонированы в базу нуклеотидных последовательностей NCBI GenBank.

Биоинформатические методы

С помощью программы fastp v. 0.24.0 [3] был проведен контроль качества полученных прочтений ДНК *L. pneumophila*, после чего проводили *de novo* сборку прочтений с помощью алгоритма unicycler v. 0.4.7.[4], а также с помощью программы Newbler GS Assembler v. 2.6.

Для проведения MLST-типирования в качестве референсной последовательности использовали полный геном штамма Philadelphia-1 с номером доступа в GenBank NC_002942, поиск гомологичных референсным генам последовательностей в выборке полных геномов был осуществлен с применением алгоритма неточного поиска BLAST+ [5] в операционной системе Linux (Ubuntu 20.04).

Загрузку и фильтрацию полных геномов по качеству прочтения производили с помощью авторского скрипта на языке программирования Python3.

Множественное выравнивание нуклеотидных последовательностей и последующий поиск единичных нуклеотидных полиморфизмов (SNP) были выполнены с использованием программы Snippy версии 4.6.0 [6]. Дальнейшую обработку и визуализацию полученных данных множественного выравнивания, анализ полиморфных позиций проводили с использованием пакета программ UGENE версии 45.0 [7]. Для визуализации данных множественного выравнивания с помощью программы IQ-TREE [8] были построены филогенетические деревья с применением алгоритма максимального правдоподобия. Итоговая визуализация и графическое представление полученного филогенетического дерева для публикации и анализа были выполнены с помощью веб-ресурса iTOL [9]

Структура и объем работы. Работа изложена на 61 странице, включает в себя введение, основной части, заключение, выводов, список использованных источников и приложение. Работа проиллюстрирована 4 таблицами и 13 рисунками. Список использованных источников включает 56 наименований.

Научная новизна

Впервые был разработан и реализован подход к мультилокусному

сиквенс-типированию *L. pneumophila* по 7 генам ассоциированным с вирулентностью с использованием мишеней, охватывающих всю область открытой рамки считывания (ORF).

Практическая значимость

Была разработана схема MLST-анализа *in silico* и база данных нуклеотидных последовательностей, содержащая аллельные варианты для каждого из генов-мишеней, что в свою очередь может быть реализовано в виде программного обеспечения, для определения сиквенс-типа по последовательности генома *L. neumophila*.

Разработана программа для загрузки полногеномных последовательностей из базы данных NCBI GenBank.

Основное содержание работы

В главе «Основная часть» представлен анализ литературных данных о метод мультилокусного сиквенс-типирования *in silico*, мультилокусное типирование *Legionella pneumophila, Escherichia coli, Salmonella spp*.

На первом этапе работы для апробации схемы MLST-типирования из базы нуклеотидных последовательностей GenBank были загружены все полные геномы штаммов *L. pneumophila,* депонированные до 8 апреля 2024 в количестве 5756 геномов. Загрузку последовательностей производили с помощью авторского скрипта на языке программирования Python3.

На рисунке 1 показано окно терминала в оперативной системе Ubuntu 20.04, с описанием работы программы. Для того, чтобы загрузить нужный нам таксон, в нашем случае это *L. pneumophila,* нужно через опцию --sp указать вид, а через опцию --outdir указать директорию для загрузки.

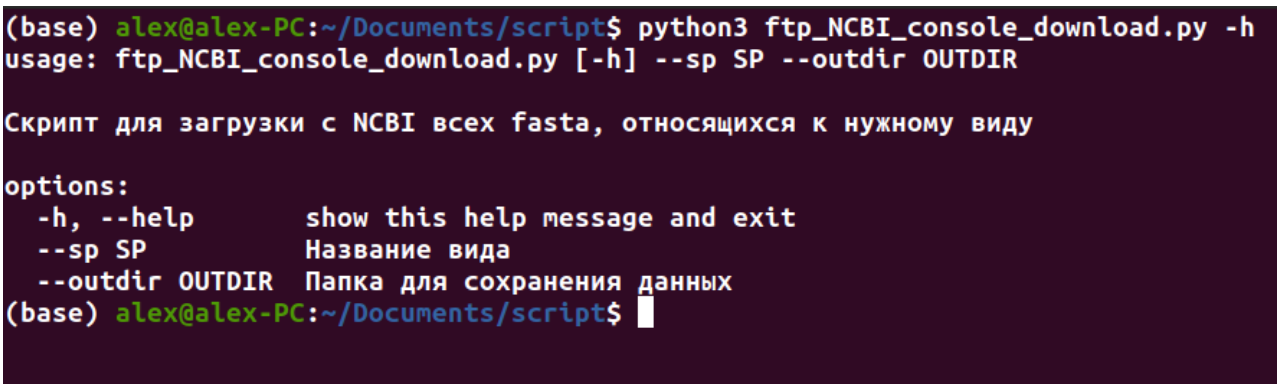

Рисунок 1 – Окно терминала в операционной системе Ubuntu 20.04, с выводом опций программы для загрузки данных в формате fasta из NCBI GenBank

По завершении работы программы в указанной папке будет содержаться 3 директории: fasta – будет содержать файлы формата .fasta, genbank – будет содержать файлы формата .gb. Формат файлов .gb предназначен для хранения аннотированных последовательностей ДНК/РНК и белков, разработанный и поддерживаемый Национальным центром биотехнологической информации США (NCBI). В директории номер 3 будут содержать архивы этих форматов, для экономии места. Файлы в директориях будут переименованы, их название будет содержать номер по базе GenBank и при наличии будут указаны: место выделение изолята, год выделения, авторское название штамма.

На втором этапе работы в качестве мишеней для схемы сиквенс-типирования геномов *L. pneumophila* было использовано 7 генов, ассоциированных с вирулентностью: *flaA*, *pilE*, *asd*, *mip*, *mompS*, *proA*, *neuA*

В процессе поиска и обработки данных по аллельным вариантам референсных генов для схемы сиквенс-типирования геномов *L. pneumophila* были получены следующие результаты: для гена *asd*, кодирующего аспартат ß–семиальдегид дегидрогеназу было найдено 5755 вариантов, среди них обнаружено 79 аллелей, для гена *flaA*, кодирующего бактериальный флагеллин – 5771 (81); *mip*, кодирующего потенциатор заражаемости макрофагов – 5755 (77); neuA, кодирующего цитидил-синтазу N-ацетилнейраминовой кислоты – 5183 (68), *pilE*, кодирующего пилин IV типа – 5684 (60), *proA*, кодирующего γ-глутамил-фосфатредуктазу – 5278 (89) и для гена *mompS*, кодирующего белок наружной мембраны

MOMP – 2804 (57).

Данные по количеству уникальных последовательностей приведены в Таблице 1.

Таблица 1 – Таблица общего количества найденных последовательностей

Ген	Общее число вариантов	Количество уникальных вариантов
flaA	5771	81
pilE	5684	60
asd	5755	79
mip	5755	77
mompS	2804	57
proA	5278	89

При анализе аллельного разнообразия, несмотря на значительные различия в количестве полиморфных сайтов, число уникальных аллелей для большинства генов оказалось статистически сопоставимым, что свидетельствует о неравномерном распределении мутаций по длине гена и наличии гипервариабельных участков. Ген *mompS* обладает максимальным числом полиморфных позиций, но формирует лишь 57 аллельных вариантов, тогда как ген *mip* с минимальным полиморфизмом образует 77 аллелей, что указывает на независимую мутацию разных генов. Также хотелось бы отметить, что в каждом из семи изученных генов вирулентности преобладали два основных аллельных варианта, которые вместе встречались более чем у 45% всех исследованных штаммов *L. pneumophila.*

Для визуализации степени эволюционного родства получившихся аллелей генов *neuA, flaA* было построено филогенетическое дерево с помощью метода UPGMA, на основе SNP и единичных INDEL. Филогенетические деревья показаны на рисунках 2 и 3.

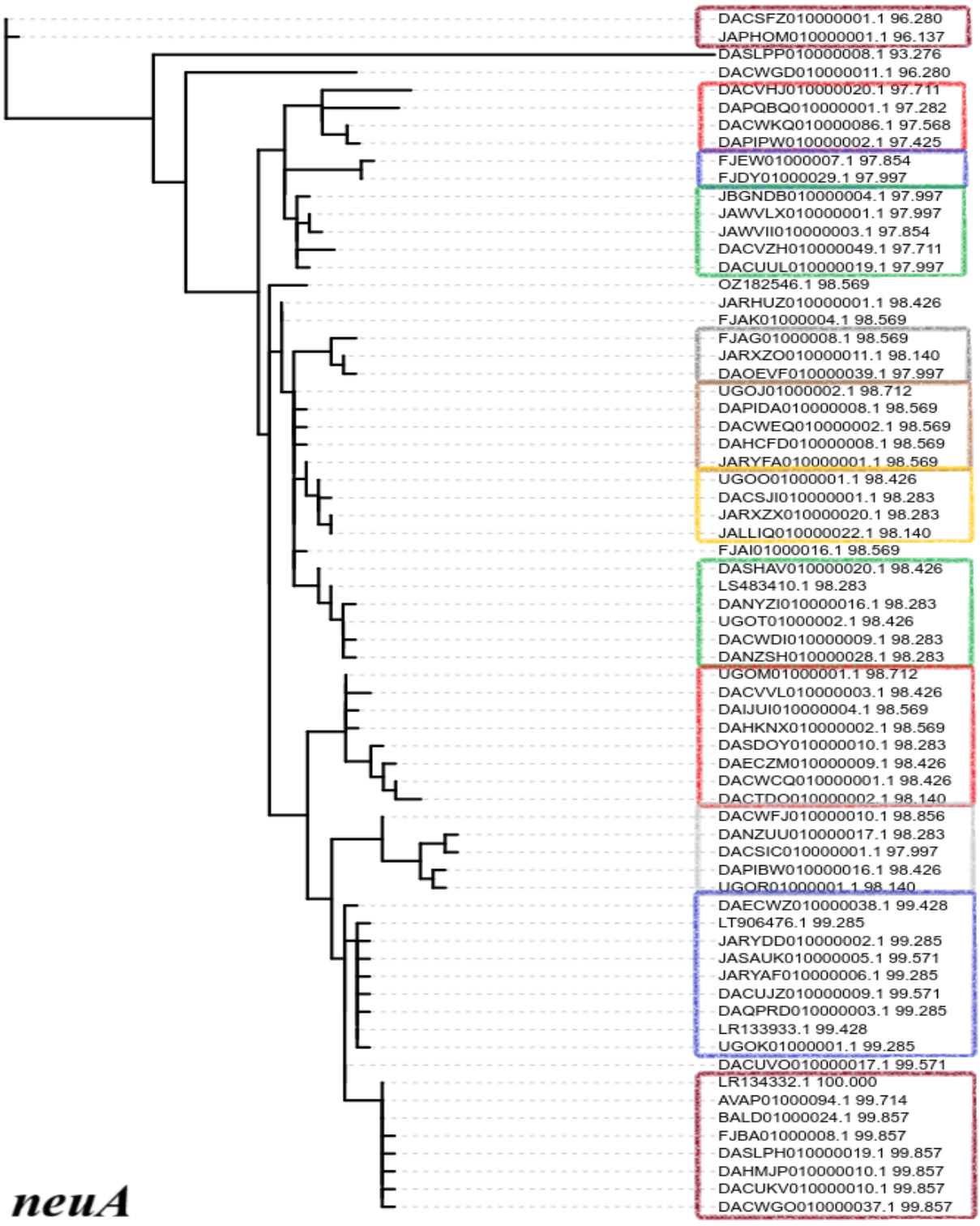

Рисунок 2 – Филогенетическое дерево вариантов гена *neuA*

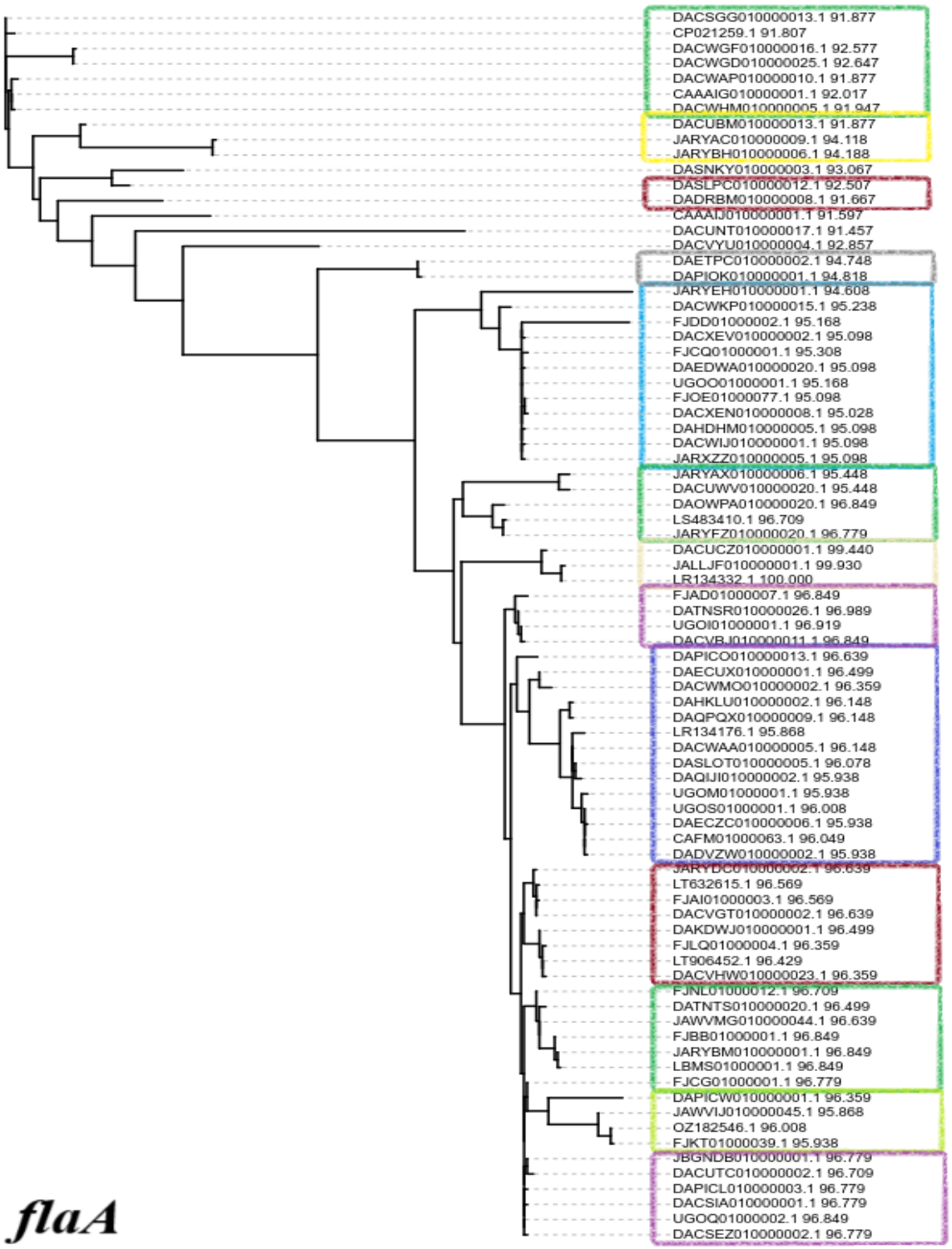

Рисунок 3 – Филогенетическое дерево вариантов гена *flaA*

Исходя из получившихся деревьев, мы можем наблюдать кластеризацию аллельных вариантов. На деревьях разноцветными рамками отмечены близкородственные варианты, исходя из этого можем сделать вывод об общем происхождении штаммов или общем географическом происхождении, а, также быть

дополнительной информацией для эпидемиологического прогнозирования, что может являться предметом изучения в дальнейшем.

На третьем этапе работы, для апробации схемы MLST-типирования были взяты в работу 24 полногеномные последовательности штаммов из Государственной коллекции патогенных бактерий при ФКУН Российский противочумный институт «Микроб» Роспотребнадзора, выделенные в городах Сочи и Астрахань. Нуклеотидные последовательности были получены на базе ФКУН Российский противочумный институт «Микроб» Роспотребнадзора в лаборатории геномного и протеомного анализа и депонированы в базу NCBI GenBank в период с 2022 по 2023 гг.

Для построения филогенетического дерева (рисунок 4) все семь выявленных аллелей группировали в один fasta-файл, тем самым оценивая сразу все отличия от референса для всех 7 мишеней. Также в выборку было дополнительно добавлены 9 штаммов со сходным профилем и указанием даты и места выделения.

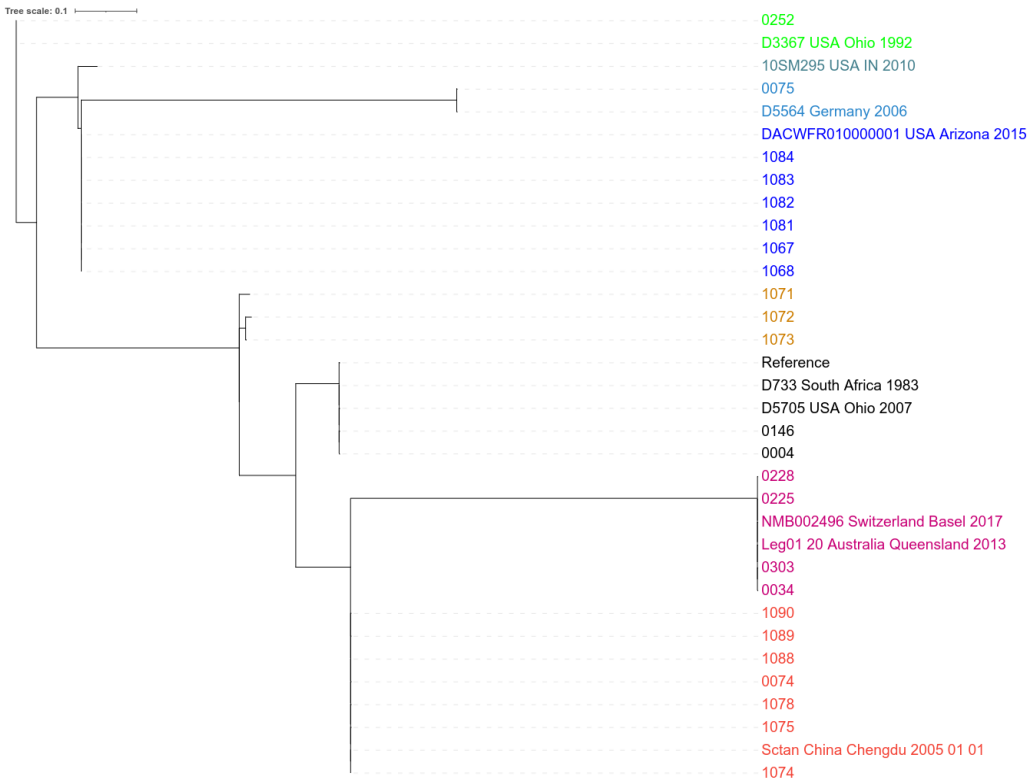

Рисунок 4 – Филогенетическое дерево по 7 генам-мишеням для штаммов, выделенных на территории Российской Федерации

Филогенетический анализ выявил два основных кластера штаммов *L. pneumophila*. Первый кластер (1067, 1068, 1081-1084) составляют штаммы серогруппы 1 с сохраненным геном *mompS*. Второй кластер объединяет штаммы серогрупп 1 и 2-14 с полной делецией гена *mompS* и разделяется на два субкластера: первый (1078, 1088, 1089, 74, 1074, 1075) с интактными генами *pilE* и *flaA* и второй (34, 225, 228, 303) с высокой вариабельностью этих. Отдельную филогенетическую ветвь формируют штаммы серогруппы 2-14 (75, 146, 4, 252) с уникальными генетическими профилями, характеризующиеся низкой гомологией по гену *neuA* и вариабельными показателями по другим генам. Выявленная кластеризация демонстрирует разделение штаммов по серогрупповой принадлежности и статусу гена *mompS*.

Полученные данные об аллельных вариантах каждого гена были сгруппированы в базу данных в таблице формата Excel, содержащую последовательность аллеля в нуклеотидах и нумерацию согласно списку выборки.

ВЫВОДЫ

1. Проанализированы нуклеотидные последовательности генов-мишеней: *flaA, pilE, asd, mip, mompS, proA, neuA*, перспективных для мультилокусного сиквенс-типирования штаммов *L. pneumophila*. На выборке из 5756 геномов, представленных в базе данных NCBI GenBank установлено: что, по результатам сравнения, исходя из количества полиморфных нуклеотидов наибольшей вариабельность характерна для генов *fla, mompS* и *proA*, средняя — для *asd, neuA, mip*, наименьшая — для *pilE. Б*ыло установлено наличие 81 аллеля гена *flaA*, 60 – *proA*, 79 – *asd*, 77 – *mip*, 57 – *mompS*, 89 – *proA*, 68 – *neuA*.

2. Гены-мишени для MLST-схемы типирования были апробированы на геномах 24 штаммов *L. pneumophila* из Государственной коллекции патогенных бактерий при ФКУН Российский противочумный институт «Микроб» Роспотребнадзора. Исследуемые генома на филогенетическом дереве образовали 4 кластера, характеризующиеся степенью гомологии генов-мишеней, а также отсутсвием или наличием гена *mompS*.

3. Разработана программа для загрузки полногеномных последовательностей из базы данных NCBI GenBank. В качестве входных данных программа принимает требуемый пользователем род и вид, в результате работы из базы NCBI GenBank будут загружены все имеющиеся в ней геномы, принадлежащие к данному виду в двух форматах: fasta (нуклеотидная последовательность) и gb (антторированная нуклеотидная последовательность, с координатами открытых рамок считывания и предсказанными белковыми продуктами).

Список использованных источников

1 Современные подходы к генотипированию возбудителей особо опасных инфекций / О.С. Бондарева [и др.] // Эпидемиология и инфекционные болезни. – 2014, № 1. – С. 34-44.

2 Legionella risk in evaporative cooling systems and underlying causes of associated breaches in health and safety compliance / B. Crook [*et al.*] // International journal of hygiene and environmental health. – 2020. – V. 224. – P. 113425.

3 fastp: an ultra-fast all-in-one FASTQ preprocessor / S. Chen [*et al.*] // Bioinformatics. – 2018. – V. 34, № 17. – P. 884–890.

4 Unicycler: resolving bacterial genome assemblies from short and long sequencing reads / R.R. Wick [*et al.*]// PLoS Computational Biology. – 2017. – V. 13, № 6. – P. 995-1002.

5 BLAST+: architecture and applications / Camacho C. [*et al.*] // BMC Bioinformatics. – 2009. – V. 10. – P. 421.

6 Snippy: Rapid haploid variant calling and core genome alignment [Электронный ресурс]. – URL: https://github.com/tseemann/snippy (дата обращения: 15.05.2024).

7 Unipro UGENE: a unified bioinformatics toolkit / K. Okonechnikov [*et al.*] // Bioinformatics. – 2012. – V. 28, № 8. – P. 1166–1167.

8 IQ-TREE 2: New models and efficient methods for phylogenetic in-fer-ence in the genomic era / B. Q. Minh [*et al.*] // Molecular Biology and Evolu-tion. – 2020. – V. 37, № 5. – P. 1530–1534.

9 Letunic, I. Interactive Tree Of Life (iTOL) v5: an online tool for phy-logenetic tree display and annotation / I. Letunic, P. Bork // Nucleic Acids Re-search. – 2021. – V. 49, № 1. – P. W293–W296.