ФЕДЕРАЛЬНОЕ КАЗЁННОЕ УЧРЕЖДЕНИЕ ЗДРАВООХРАНЕНИЯ

РОСТОВСКИЙ-НА-ДОНУ

ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ

НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ПРОТИВОЧУМНЫЙ ИНСТИТУТ

ФЕДЕРАЛЬНОЙ СЛУЖБЫ ПО НАДЗОРУ В СФЕРЕ ЗАЩИТЫ ПРАВ

ПОТРЕБИТЕЛЕЙ И БЛАГОПОЛУЧИЯ ЧЕЛОВЕКА

“УТВЕРЖДАЮ”

Директор ФКУЗ Ростовский-на-Дону

противочумный институт Роспотребнадзора,

к.м.н. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Титова С.В.

« 04 » декабря 2014 г.

Анализ результатов полногеномного секвенирования Vibrio cholerae и определения кратности вариабельных тандемных повторов (VNTR) с помощью программы SeqAnalyzer

Методические рекомендации

Ростов-на-Дону, 2014 г.

**Учреждения разработчики:**

ФКУЗ Ростовский-на-Дону противочумный институт Роспотребнадзора

ФБУН ЦНИИ Эпидемиологии Роспотребнадзора

**Авторы:**

Водопьянов А.С., Водопьянов С.О., Писанов Р.В., Олейников И.П., Кулешов К.В., Керманов А.В., Олейникова К.В.

**Рассмотрены** на заседании Ученого совета ФКУЗ Ростовский-на-Дону противочумный институт Роспотребнадзора протокол № 12 от 04.12.2014

**Уровень внедрения -** учрежденческий

**Область применения:** методические рекомендации предназначены для врачей, биологов, специалистов по молекулярной биологии, проводящих анализ данных полногеномного секвенирования штаммов холерного вибриона.

**Аннотация**

В методических рекомендациях предлагается компьютерная программа, позволяющая проводить анализ данных полногеномного секвенирования штаммов холерного вибриона. В основе метода лежит поиск заранее определенных референсных генов среди набора контигов, что позволяет определять вид, серогруппу, биовар анализируемого штамма, а также ряд мобильных генетических элементов и островков.

**Введение**

В настоящее время наблюдается бурный рост технологий высокопроизводительного секвенирования, что сопровождается заметным снижением стоимости определения первичной нуклеотидной структуры штаммов различных микроорганизмов. Однако при анализе данных возникают как чисто алгоритмические сложности, связанные с большим объемом данных, так и специфические сложности, связанные с характером конкретной биологической задачи. Ряд имеющихся программных продуктов и алгоритмов разработан под решение конкретных задач и не может быть применен в других условиях. В противовес этому, универсальные программные продукты для анализа данных полногеномного секвенирования, как правило, сложны в освоении и зачастую требуют большое количество дополнительной информации и (или) связи с зарубежными банками данных нуклеотидных последовательностей. Вместе с тем, последнее время все чаще возникает вопрос о возможном прекращении доступа к зарубежным банкам данным в связи с санкциями против Российской Федерации. По данным ряда СМИ, крупнейшие компании США, такие как Microsoft, Oracle, Symantec и Hewlett-Packard прекращают сотрудничество с банками и структурами, которые попали в санкционные списки США. Это обуславливает актуальность разработки отечественного программного обеспечения, независимого от зарубежных корпораций.

Немаловажным фактором является стоимость специализированного программного обеспечения, которое может многократно превышать стоимость оборудования. В ряде случаев, особенно в отношении государственных учреждений, это является критическим фактором, вынуждающим искать альтернативные методы анализа данных.

Для некоторых программных продуктов требуется сборка полного генома (de novo или с использованием референсной последовательности), что тоже представляет собой определенные сложности и затрудняет проведение анализа.

Однако учитывая высокую эпидемиологическую значимость холерного вибриона, скорость и оперативность проведения анализа являются довольно актуальными факторами.

Все вышеизложенное обуславливает необходимость разработки алгоритмов, позволяющего проводить быструю обработку данных полногеномного секвенирвования штаммов холерного вибриона.

Разработанная программа SeqAnalyzer позволяет проводить быстрый анализ результатов секвенирования Vibrio cholerae по заранее определенному набору генов. Важной особенностью является возможность определения кратности вариабельных тандемных повторов, что имеет существенное значение для эпидемиологического анализа.

Отличительными особенностями программы SeqAnalyzer являются:

1. Низкие требования к оборудованию – для работы требуется обычный «офисный» компьютер или ноутбук с 1 Гб оперативной памяти
2. Высокая скорость работы – на анализ данных одного генома (около 10 Мб) требуется 3-4 минуты
3. Простота управления – программа разработана по принципу «одной кнопки» - никаких дополнительных настроек или специальных знаний не требуется
4. Программа рассчитана на работу с набором контигов, что не требует проведения предварительной сборки генома
5. Простота оценки результатов – данные анализа данных полногеномного секвенирования представляют собой интуитивно понятный текстовый файл на русском языке.

**Алгоритм работы программы**

Поиск генов среди данных секвенирования проводится с использованием алгоритма локального выравнивая Смита-Ватермана с аффинными штрафами за делеции и вставки. При этом референс-последовательность каждого анализируемого гена ищется среди всех представленных контигов. Если один и тот же ген будет обнаружен программой в нескольких разных контигах, то в результирующий файл будет занесен лучший результат.

Для определения кратности вариабельных тандемных повторов используется алгоритм «виртуальной ПЦР» с праймерами, фланкирующими область тандемного повтора. При этом кратность (число повторов) определяется по формуле: *число повторов = (размер амплификата – размер фланкирующих участков) / кратность повтора*

**Установка и запуск программы**

SeqAnalyzer доступен для скачивания на сайте Ростовского противочумного института по адресу http://antiplague.ru/seqanalyzer в двух вариантах:

1. Полная версия (около 40 Мб) – предназначена для компьютеров с операционной системой Windows XP, 7, 8 без установленной виртуальной машины JAVA (JRE). В этом случае для запуска программы достаточно распаковать скачанный архив с программой и запустить файл SeqAnalyzer.bat
2. Только лишь исполняемый файл (около 1 Мб) – предназначен для компьютеров с установленной виртуальной машиной JAVA (JRE) на базе Windows, Linus, Mac OS X. При необходимости свежую версию JRE можно бесплатно скачать с официального сайта компании Oracle ([www.oracle.com](http://www.oracle.com)). При этом запускаемым файлом является SeqAnalyzer.jar

Для тестового запуска также доступен для скачивания Fasta-файл (301.fasta), содержащий данные полногеномного секвенирования токсигенного штамма Vibrio cholerae El Tor № 301 (выделенный в 2011 году в Таганрогском заливе)

**Входные данные**

В качестве исходных данных используется файл в fasta-формате, содержащий набор контигов (протяженных нуклеотидных последовательностей - результат сборки de novo коротких ридов, полученных в результате полногеномного секвенирования). Предпочтительное расширение файла «.fasta» или «.fa».

**Результат работы программы**

Результатом работы программы является текстовый файл, содержащий информацию о наличии в исследуемом наборе контигов анализируемых генов. В случае обнаружения гена для него указывается процент совпадения с референс-последовательностью, заложенной в программе (рисунок 1). Для большинства генов в качестве референсных последовательностей использованы гены штамма Vibrio cholerae El Tor №16961.

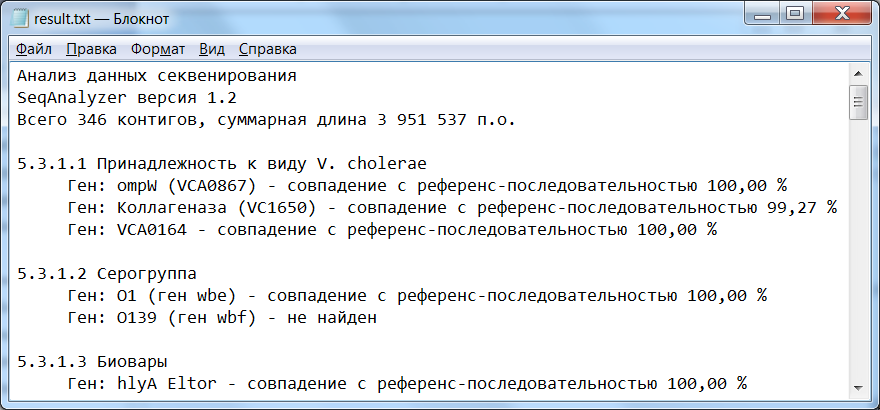


Рисунок 1 – Результат работы программы (фрагмент)

**Проведение анализа**

Внешний вид программы представлен на рисунке 2. Для выбора анализируемого файла необходимо нажать кнопку «Выбрать». При этом открывается стандартный диалог выбора файла, в котором необходимо выбрать файл, содержащий данные полногеномного секвенирования (контиги).

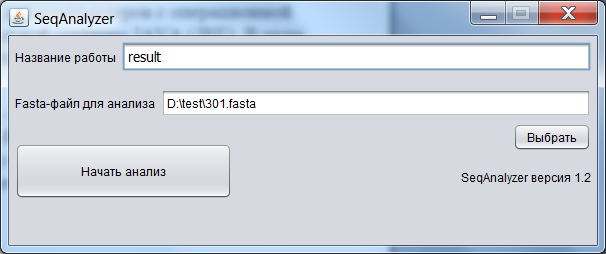


Рисунок 2 – Внешний вид программы

При необходимости можно задать название для проводимого анализа в поле «Название работы». В этом случае это же имя будет присвоено текстовому файлу с результатами (расположен в папке с программой).

Для проведения анализа необходимо нажать кнопку «Начать анализ». При этом в нижней части окна программы появляется индикатор готовности результатов. Ориентировочное время для анализа fasta-файла размером 10 Мб составляет 2-3 минуты. После окончания анализа автоматически открывается текстовый файл с результатами анализа (рисунок 1).

**Интерпретация результатов**

Результат работы программы представляет собой текстовый файл, в котором содержится информация о наличии набора анализируемых генов и процента сходства с референс-последовательностью.

Принадлежность к виду V. cholerae определяется по наличию генов ompW, гена коллангеназы и гена VCA0164. Наличие любого из этих генов подтверждает принадлежность к виду V. cholerae.

Серогруппа определяется по наличию генов wbe (V. cholerae О1) или wbf (V. cholerae О139).

Для определения биовара используется анализ генов гемолизина (hlyA), rtxC (присутствует только у V. cholerae El Tor) и cas3 (характерен для вибрионов классического биовара)

Немаловажным является определение различных мобильных генетических элементов. В программе предусмотрено выявление последовательностей профагов CTX и RSI, островков патогенности VPI-I и VPI-II, островков пандемичности VSP-I и VSP-II, кластеров генов RTX, mshA (манозо-чувтсвительный гемаглютинин), генов систем секреции третьего (T3SS) и шестого (T6SS) типа.

Для ряда штаммов холерного вибриона, выделяемых в последнее время, характерна множественная антибиотикоустойчивость. В связи с этим в программе SeqAnalyzer предусмотрено выявление генов устойчивости к триметоприму, сульфаметоксазолу, хлорамфениколу (левомицетину), стрептомицину, канамицину и к бета-лактамным антибиотикам.

В случае выявления в анализируемых контигах вариабельных тандемных повторов в программе предусмотрено автоматическое вычисление их кратности (числа повторов). Однако качество поиска тандемных повторов полностью зависит от качества и алгоритма сборки полногеномных последовательностей (контигов).

**Обратная связь**

Все пожелания, замечания, выявленные ошибки можно направлять по электронной почте ([alexvod@gmail.com](mailto:alexvod@gmail.com)), телефону (863) 240-22-66 или на почтовый адрес: 344002, г. Ростов-на-Дону, ул. Максима Горького, 117, Водопьянову Алексею Сергеевичу.

**Литература**

1. Водопьянов А.С. Вариабельные тандемные повторы, выявленные при компьютерном анализе генома Vibrio cholerae / А.С. Водопьянов, С.О. Водопьянов, М.Б. Мишанькин, И.Ю. Сучков, Б.Н. Мишанькин // Биотехнология. 2001. № 6. С. 85.
2. Осин А.В. Разработка алгоритма MLST-типирования пандемических и предпандемических штаммов Vibrio cholerae биовара эльтор / Осин А.В., Краснов Я.М., Гусева Н.П., Смирнова Н.И. // Проблемы особо опасных инфекций. 2011. № 1(107). С. 58-61.
3. Ceccarelli, D. Origin of Vibrio cholerae in Haiti / D. Ceccarelli, M. Spagnoletti, P. Cappuccinelli, V. Burrus, M.M. Colombo // Lancet Infect Dis. 2011 Apr;11(4):262. doi: 10.1016/S1473-3099(11)70078-0.
4. Safa A., Nair G.B., Alam M. Evolution of new variants of Vibrio cholerae O1. Trends Microbiol. 2010; 18:46–54.
5. Spagnoletti, M. Rapid detection by multiplex PCR of Genomic Islands, prophages and Integrative Conjugative Elements in V. cholerae 7th pandemic variants / M. Spagnoletti, D. Ceccarelli, M.M. Colombo // J Microbiol Methods. 2012 Jan;88(1):98-102. doi: 10.1016/j.mimet.2011.10.017. Epub 2011 Oct 28.
6. Teh CS. Genetic variation analysis of Vibrio cholerae using multilocus sequencing typing and multi-virulence locus sequencing typing // Teh CS, Chua KH, Thong KL // Infect Genet Evol. 2011 Jul;11(5):1121-8. doi: 10.1016/j.meegid.2011.04.005. Epub 2011 Apr 12.