«Сейфуллин оқулары-18(2): «XXI ғасыр ғылымы – трансформация дәуірі» Халықаралық ғылыми-практикалық конференция материалдары = Материалы международной научно-практической конференции «Сейфуллинские чтения – 18(2): «Наука XXI века - эпоха трансформации» - 2022.- Т.I, Ч.III. - С.95-96.

РАЗРАБОТКА ПАЙПЛАЙНА ДЛЯ ОБРАБОТКИ ДАННЫХ ГЕНОТИПИРОВАНИЯ БАКТЕРИАЛЬНЫХ ШТАММОВ

Шевцов В.А., докторант 2 курса
Казахский агротехнический университет им. С. Сейфуллина, г. Нур-Султан

Совмещение данных генотипирования микроорганизмов полученных в «догеномный» период с новыми геномными данными стало настоящей проблемой. Данная пропасть возникла в связи с различиями протоколов изучения ДНК микроорганизмов на протяжении развития технологий генотипирования. Генотипирование на основании фрагментации ДНК с разделением фрагментов в pulsed-field gel electrophoresis (PFGE), Random Amplified Polymorphic DNA (RAPD), Restriction Fragment Length Polymorphism (RFLP) сменились на более простые, точные и легко воспроизводимые методы как мультилокусный анализ вариабельных тандемных повторов (MLVA) и мультилокусное сиквенс типирование (MLST). В настоящее время существует переходный период, когда для полноценной эпидемиологии необходимо использовать данные, полученные по всему миру различными методами генотипирования, а также при ретроспективном анализе ранее полученных результатов генотипирования циркулирующих штаммов. В переходный период многие лаборатории вынуждены параллельно использовать полногеномное секвенирование, PFGE and MLVA для ряда патогенов, чтобы проводить полноценный эпидемиологический контроль. В связи с этим развитие методологии in silico генотипирования на полногеномных данных является актуальной задачей, особенно для патогенов, для которых MLVA и MLST рассматривался в качестве злотого стандарта генотипирования включая *Brucella spp*, *Bacillus anthracis*, *Yersenia pestis*, *Francisella tularensis* and *Neisseria meningitides*.[1]

В настоящее время полногеномное секвенирование входит в рутину большинства биологических исследований, позволяя получать большие массивы данных, которые необходимы обрабатывать, поэтому биоинформатическая обработка данных сейчас актуализируется и появляется необходимость совершенствования обработки данных процессов[2]. Кроме

кодирующих частей геномов бактерий, вирусов, растений, существует также множество некодирующих частей, включая так называемые сателлитные тандемные повторы, представляющие короткие последовательности повторяющихся фрагментов ДНК, которые нашли широкое применение в идентификации где до сих пор остается основным методом основанным на анализе коротких тандемных повторов. Так же данный метод нашел широкое приминение в молекулярно-генетическом анализе бактериальных штаммов патогенов, включая особо опасные инфекции такие как бруцеллез, сибирская язва, туляримия, чума, однако ранее данные повторы исследовались методом ПЦР амплификации и визуализации размеров бендов в электроарезе или капилярном электрафорезе, поэтому в геномную эру необходимо совмещать полученные ранее классическими методами данные с полученными данными с помощью метода NGS(New Generation Seqiencing) секвенирования, поэтому разработка программы, позволяющей определять размеры тандемных повторов является интересной и важной задачей для биологических наук.

В данной работе мы разработали скрипт, написанный на языке python3 позволяющий выявлять VNTR повторы на основании праймеров фланкирующих регион интереса использует риды. Преимущество нашего метода является возможность использования сырых данных, таким образом значительно снижается вероятность ошибок так как при использовании сборок, программы сборщики образуют конценсус в которой как правило образуются разрывы и последовательность с тандемным повтором может быть исключена из сборки или засчет нарушения размера k-мера или за счет гомоплазии данного региона. Кроме того данный скрипт позволяет определять количественное соотношение разных размеров повторов используемых в образце, что позволяет определять контаменированные образцы.

Скрипт был апробирован на 10 образцах полногеномных данных бруцелл, в результате метода с использованием 50 VNTR повторов найденых в геноме бруцелл все повторы были идентифицированы, а также в одном образце выявлена контаминация на основании соотношения разных длин 6 VNTR локусов, что свидетельствует о контаминации образца.

Список использованной литературы

1 Кузнецова, И. В. Молекулярно-Генетическая Диагностика Возбудителей Инфекционных Болезней.

2 Köser, C. U., Ellington, M. J., Cartwright, E. J., Gillespie, S. H., Brown, N. M., Farrington, M., ... & Peacock, S. J. (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology.