С.Сейфуллиннің 125 жылдығына арналған **«Сейфуллин оқулары – 15: Жастар, ғылым, технологиялар: жаңа идеялар мен перспективалар»** атты халықаралық ғылыми-теориялық конференциясының материалдары = Материалы Международной научно-теоретической конференции **«Сейфуллинские чтения – 15: Молодежь, наука, технологии – новые идеи и перспективы»**, приуроченной к 125 - летию С.Сейфуллина. - 2019. - Т.I, Ч.2 - С.148-151

РАЗРАБОТКА АЛГОРИТМОВ ДЛЯ ОБРАБОТКИ ПОЛНО ГЕНОМНЫХ ДАННЫХ БАКТЕРИЙ

Шевцов В.А.

В Казахстане идет бурное внедрение технологий полногеномного секвенирования бактериальных штаммов в клиническую эпидемиологию, фундаментальную и прикладную биотехнологию. В настоящее время полногеномные данные позволяют установить фактор патогенности или лекарственной устойчивости бактериальных патогенов и своевременно скорректировать терапию. Также в биотехнологии полногеномные данные позволяют отобрать высокоценные и перспективные штаммы. Тем не менее, биоинформатическая школа остается недостаточно развитой. В рамках данной работы будет оптимизирован алгоритм обработки полногеномных данных бактерий, который может на практике применятся в эпидемиологических и биотехнологических направлениях.

Определение нуклеотидной последовательности геномов в настоящее время является основной технологией в биологических исследованиях. Еще 20 лет назад изучение геномных данных казалась дорогой и трудно решаемой задачей именно в области определения нуклеотидной последовательности. Первичная стоимость генома человека оценивалась в 3 миллиарда долларов. Прогресс в развитии секвенаторов нового поколения (NGS) мгновенно изменили ситуацию, на данный момент геном человека за 1000 долларов США стал реальностью [1] .

Снижение стоимости «нить жизни» с аннотацией и предсказанием функции открытых рамок считывания. Существуют различные технологии в получении полногеномных данных, однако самые распространенные разработаны Illumina и LifeTechnology. Которые генерирует геномных исследований привело к бурному развитию данной области, и уже реализуются грандиозные проекты нацеленные на геномное секвенирование 5000 насекомых Arthropod Genomic Consortium, 2014, 5, десять тысяч геномов позвоночных Genome 10K Community of Scientists, 2009, миллионы микроорганизмов и др. Однако, генерация данных это только часть в представлении полногеномных данных, следующим не менее трудоемким шагом является сборка коротких последовательностей в цельную информативную «множество коротких фрагментов протяженностью 100-250 п.н. (Denovo Assemblingusing IlluminaReads, 2009) . Из-за коротких чтений возникают сложности в сборке генома связанные с тандемными повторами,

которые превышают длины чтения, палиндромами и др., Что не позволяет правильно сшить получаемые контиги и многие сборки геномов остаются, фрагментированы и с большим количеством ошибок в расположение фрагментов. Подтверждение этого описывается в работе Salzberg и Yorke где указывается на сотни неправильных соединений в геномах. Alkan с соавторами сообщили, что сборка «denovo» генома человека с использованием коротких фрагментов на 16% короче, чем геном, собранный с использованием гораздо более дорогих подходов. Сложность в обработки коротких фрагментов при сборке полных геномов спровоцировала разработку широкого спектра программ сборщиков. И все же, несмотря на количество доступных инструментов, нет универсальных алгоритмов обработки геномов и массивов данных получаемых с разных платформ.

Процесс и Данные Секвенирования. Данные после секвенирования представляют собой текст в котором присутствуют всего 4 символа: ATCG.

FASTQ: текстовый формат для хранения как последовательности ДНК, так и соответствующих показателей качества.

Платформы секвенирования. Возникающие в процессе секвенирования, могут привести к приблизительно 0,1–1% неправильным вызовам баз [1], которые известны как ошибки секвенирования. В таблице 1 показаны коэффициенты ошибок различных основных платформ NGS. Подготовка библиотеки также может привести к значительным ошибкам. Например, окисление гуанина является важным источником искусственных мутаций, потому что 8-oxoG, как правило, в паре с аденином вместо цитозина [2]. Долгое время инкубации, которые являются общими во многих выделения ДНК и гибридных протоколов захвата, может значительно увеличить число г!Цубституций. Недавно исследование показало, что процесс восстановления ДНК может устранить 77% и 82% G ! T и C ! Ошибками, соответственно [3]. Это исследование показывает, что повреждения ДНК могут вызвать большое количество ошибок. Помимо ошибок, вносимых в ходе подготовки и секвенирования проб, программное обеспечение и средства анализа также могут вносить ошибки. В частности, в референсном геноме можно вызвать ложноположительные варианты областей генома с гомологичными последовательностями и повторяющимися последовательностями.

Таблица 1 - Сравнение коэффициентов ошибок секвенирования

Платформа	Наиболее частые ошибки	Кoофицент ошибок	Кол-во данных/ запуск
Capillarysequencing	Одиночные нуклеотидные замены	10^{-1}	10 GB
IlluminaNextSeq	Одиночные нуклеотидные замены	10^{-3}	120 Gb
PacBio RS	CG удаления	10^{-2}	20 Gb
IonTorrent PGM		10^{-2}	200Gb

IlluminaMiSeq	Одиночные нуклеотидные замены	10^{-3}	15 Gb
IlluminaHiSeq	Одиночные нуклеотидные замены	10^{-3}	
454 GS Junior		10^{-2}	
Solid	A-T смещение	$2 * 10^{-2}$	

Бесклеточные фрагменты ДНК обычно короткие и имеют компактный пик около 167 bp [4]. Этот факт увеличивает вероятность того, что два разных исходных фрагментов безкоеточной ДНК имеют одинаковую последовательность и, следовательно, увеличивает сложность удаления этих дубликатов, поскольку алгоритмы дедупликации не смогут дифференцировать такие идентичные и дублированные чтения, вызванные усилением. Таким образом, обнаружение низкочастотных мутаций из шумных данных секвенирования кднк является сложной задачей. Обычные инструменты не могут хорошо справляться с задачами анализа цтДНК, поэтому необходимы более специализированные инструменты.

Сборка генома и аннотация генома - это области, где нет золотых стандартов. Проекты часто являются исследовательскими, и знание того, хороши ли результаты или плохи, часто трудно определить. Это особенно верно, если идет работа с организмами, которые лишь отдаленно напоминают уже секвенированные и опубликованные организмамы, в результате чего их мало с чем сравнивать.

С развитием новых технологий секвенирования качественная сборка генома становится более осуществимой задачей, чем когда-либо, и хорошо собранный и аннотированный геном будет ресурсом, который можно будет использовать в течение многих лет.

Список литературы

1. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA (2014) Accuracy of next generation sequencing platforms. Next Gener Seq Appl 1. https://doi.org/10.4172/jngsa.1000106

2. Arbeithuber B, Makova KD, Tiemann-Boege I (2016) Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. DNA Res 23 (6):547–559. https://doi.org/10.1093/dnares/dsw038

3. Lixin Chen PL (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355 (6326):752–756

4. Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, Gligorich KM, Rostomily RC, Bronner MP, Shendure J (2016) Fragment length of circulating tumor DNA. PLoS Genet 12(7):e1006162. https://doi.org/10.1371/journal.pgen.1006162