Отчет по работе на кластере НГУ

Тема работы.

Разработка программы для удаления последовательностей праймеров из прочтений массового параллельного секвенирования (NGS)

Состав коллектива (с указанием места учёбы/работы, учёных степеней и званий).

Кечин Андрей Андреевич, ассистент НГУ, м.н.с. Лаборатории фармакогеномики ИХБФМ СО РАН

Научное содержание работы:

Постановка задачи.

Разработка программы для удаления последовательностей праймеров из прочтений NGS

Современное состояние проблемы.

Завершающим этапом исследования генов *BRCA1/2* с помощью технологий NGS является обработка и анализ полученных после секвенирования прочтений. При этом могут быть использованы несколько разных подходов. Первым подходом является последовательный запуск всех программ через командную строку с выбором всех необходимых параметров их работы. Минус такого подхода – трудность его адаптации в других лабораториях. Вторым подходом является использование таких систем составления автоматических протоколов, как Galaxy (https://galaxyproject.org) и GeneXplain (http://genexplain.com), позволяющие с помощью компьютерной мыши легко составлять протоколы из тех программ, которые уже установлены на сервере. Однако при необходимости установки новых инструментов или модификации старых, возникает потребность в создании скриптов, интерфейсов и прочих настройках. Кроме того, подобные системы требуют постоянный доступ в интернет. Третьим подходом является использование готовых протоколов, в которых уже готовые программы объединены в единую оболочку с помощью BASH, Python, Perl и других скриптов. Такие протоколы легко могут быть модифицированы и не требуют доступа в интернет. В то же время, все параметры в них должны быть заранее подобраны и адаптированы к конкретной задаче на репрезентативной выборке результатов секвенирования, чтобы конечному пользователю не требовалось проводить подбор их значений параметров. На сегодняшний день доступны только коммерческие автоматические протоколы обработки данных. Поэтому актуальным является разработка свободно доступного протокола обработки NGS данных генов *BRCA1/2* и его проверка на представительной выборке образцов. Помимо объединения готовых программ в единый протокол, существует и необходимость разработки программы для удаления последовательностей праймеров из полученных прочтений. Это обусловлено тем, что созданные ранее программы адаптированы только для удаления последовательностей адаптеров, поскольку таргетное секвенирование,

основанное на амплификации (amplicon-based targeted NGS) получило широкое распространение только в последнее время.

Подробное описание работы, включая используемые алгоритмы.

Разработанную программу для удаления последовательностей праймеров из прочтений, названная cutPrimers (https://github.com/aakechin/cutPrimers) сравнивали с другими существующими программами (cutadapt, AlienTrimmer и BBDuk) по следующим параметрам:

1. процент прочтений с минимальной длиной не менее 80 нуклеотидов;
2. время обработки всех прочтений;
3. число ампликонов, покрытых не менее, чем 30 прочтениями (по каждой позиции);
4. медианное покрытие.

Сравнение проводили на трех независимых NGS-запусках и одном целом запуске (42 образца). В каждом наборе содержалось следующее число прочтений: 12757 (1 образец), 24376 (1 образец), 89919 (1 образец) and 16429892 (42 образца). После удаления последовательностей праймеров проводили их картирование с помощью программы BWA. Значения покрытия определяли программой samtools.

Полученные результаты.

Таблица 1. Сравнительный анализ четырех программ, используя четыре набора последовательностей из четырех независимых запусков NGS. Cutadapt была использована в двух режимах: с якорными символами (^ и $), которые заставляют cutadapt искать последовательности праймеров только на самых концах прочтений, и без них. BBDuk также запускалась в двух режимах: для поиска последовательностей праймеров с допуском только замен (hdist=3) и с допуском и замен, и инсерций/делеций (edist=2). Наилучшие варианты по каждому параметру сравнения выделены жирным шрифтом.

Число прочтений (число образцов)	**Программа**	**cutadapt**		**BBDuk**		**AlienTrimmer**	**cutPrimers**
	Режим	с якорными символами	без якорных символов	8 потоков, k=20, edist=2	8 потоков, k=20, hdist=3	парные прочтения	8 потоков, err=5
12757 (1)	**Время обработки, секунды**	32	113	9	56	3	**7**
24376 (1)		59	216	**9**	59	5	16
89919 (1)		211	824	**13**	**61**	18	70
16429892 (42)		40823	143653	**1109**	**3969**	3254	13739
12757	**Число оставшихся прочтений ≥90 н., % от всех прочтений**	8259 (64,7%)	3913 (30,7%)	9781 (76,7%)	9993 (78,3%)	4849 (38,0%)	**11182 (87,6%)**
24376 (1)		21854 (89,6%)	11897 (48,8%)	22733 (93,2%)	21876 (89,7%)	7829 (32,1%)	**22707 (93,2%)**
89919 (1)		64283 (71,5%)	31662 (35,2%)	67753 (75,3%)	65638 (73,0%)	27612 (30,7%)	**66885 (74,4%)**
16429892 (42)		14345458 (87,3%)	7800514 (47,5%)	15222691 (92,6%)	15217255 (92,5%)	2385222 (14,5%)	**15231108 (92,7%)**
12757 (1)	**Число покрытых ампликонов (>30 на позицию)**	105	65	111	137	6	**146**
24376 (1)		174	89	159	181	8	**183**
89919 (1)		181	116	185	188	21	**189**
16429892 (42)		181,0 (174-181)	118,5 (82-130)	188,0 (149-189)	183,0 (152-187)	41,5 (7-80)	**189,0 (182-189)**
12757 (1)	**Медианное покрытие**	37,0	9,0	35,0	60,0	0,0	**68,0**
24376 (1)		135,5	19,0	74,5	141,5	0,0	**145,5**
89919 (1)		431,0	97,0	247,0	456,5	3,0	**473,0**
16429892 (42)		1743,0	165	974,0	1387,0	3,0	**1885,5**

Программы cutadapt и AlienTrimmer оставляли значительно меньше прочтений по сравнению с cutPrimers и BBDuk. Основной проблемой AlienTrimmer и cutadapt (режим без якорных символов) было то, что они удаляли все последовательности, которые соответствуют какому-либо из 190 праймеров, которые вводит пользователь, в том числе находящиеся в середине прочтения. Это приводит к потере большого числа прочтений пересекающихся ампликонов и снижению числа правильно обработанных прочтений. В то же время, cutadapt с якорными символами находит последовательности праймеров, удаленных только на несколько нуклеотидов от конца, что может встречаться довольно часто: в случае, если длина ампликона меньше длины прочтения. В этом случае прибор NGS прочитывает последовательность адаптера и его символы рассматриваются программой cutadapt как ошибки (инсерции) при поиске последовательности праймера в прочтении.

В то же время, cutPrimers и BBDuk ищут последовательность праймера только в части прочтения и не удаляют последовательности, соответствующие последовательностям праймеров соседних ампликонов. Как следствие, такой подход дает больший выход правильно обрезанных прочтений. Особенно эта разница заметна при обработке прочтений целого запуска по числу ампликонов, покрытых минимум 30 прочтениями: для cutadapt не было ни одного образца, для которого были бы покрыты все 189 ампликонов (один ампликон не был покрыт совсем).

Несмотря на то, что BBDuk показал значительно более высокую скорость обработки, его использование имеет некоторые трудности. Во-первых, BBDuk имеет ограничение по числу допустимых ошибок (только 3 – для замен и только 2 – для замен + инсерций/делеций). Во-вторых, использование параметра «edist» приводит к вырезанию вместе с праймером лишней последовательности от 3’-конца, что изменяет значения покрытия крайней позиции в ампликоне. В-третьих, BBDuk использует для поиска последовательностей праймеров так называемые к-меры, что приводит к вырезанию лишней последовательности из прочтения.

Таким образом, cutPrimers представляет собой новую программу, позволяющую быстро и эффективно удалять последовательности праймеров из прочтений. Несмотря на то, что программа проигрывает некоторым из аналогичных программ по скорости работы, она превосходит их по точности идентификации и удаления последовательностей праймеров.

Иллюстрации, визуализация результатов. Нет

Эффект от использования кластера в достижении целей работы.

Использование кластера позволило ускорить весь процесс обработки благодаря более производительным процессорным ядрам по сравнению с нашим 8-ядерным ПК.

Перечень публикаций, содержащих результаты работы (если есть). Указать импакт-фактор журнала (Thomson Reuters, РИНЦ,…).

Kechin A., Boyarskikh U., Kel A., Filipenko M. cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing // J. Comput. Biol. – 2017. – V. 24. – №11. – C. 1138–1143.

Ваши впечатления от работы вычислительной системы и деятельности ИВЦ НГУ, а также Ваши предложения по их совершенствованию.

В процессе работы возникали некоторые трудности с запуском процессов, в том числе из-за отсутствия некоторых пакетов. Однако с помощью Владислава Калюжного удалось их обойти и прозвести необходимый анализ.

В качестве пожеланий хотелось бы, чтобы подробнее были описаны полезные команды для работы на кластере. Например, каким образом не отсоединяясь от сервера, проверить, сколько места из выделенной квоты осталось.