

Множественные дивергированные дисперсные повторы в геноме риса

В.М. Руденко*, Е.В. Коротков

ФИЦ «Фундаментальные основы биотехнологии» Российской академии наук, Москва, Россия

* v.m.rudenko@gmail.com

Цель: В настоящее время аннотация геномов предполагает не только определение местоположения генов, но также и различных видов повторов, включая дисперсные повторы (ДП). Выявление ДП важно по причине того, что большая часть из них является мобильными элементами, которые играют ключевую роль в пластичности генома, обеспечивая его наилучшую приспособленность к изменениям окружающей среды. Наибольшее число ДП обнаруживается в геномах растений. Цель работы состояла в выявлении и дальнейшем изучении сильнодивергированных дисперсных повторов в геноме *Oryza sativa* с помощью IP метода. Также необходимо было сравнить полученные результаты с известной аннотацией повторов риса, проведенной конвейером EDTA.

Материалы и методы: Все разработанные на сегодняшний день биоинформационные методы поиска ДП de novo имеют существенное ограничение. Они способны находить ДП с незначительной степенью дивергенции, определяемой числом замен на один нуклеотид x : $x \leq 1.0$. Разработанный нами IP метод позволяет выявлять сильнодивергированные ДП с $x \leq 1.5$, которые не могут быть обнаружены другими подходами. IP метод строит позиционно-весовые матрицы (ПВМ) семейств ДП, учитывающие частоты динуклеотидов в последовательностях повторов. ПВМ далее используются для поиска ДП методом динамического программирования.

В данной работе мы осуществляли поиск ДП в геноме *O. sativa* (сборка IRGSP-1.0). Геномная последовательность риса и аннотация, включающая местоположение генов были взяты с сайта: https://plants.ensembl.org/Oryza_sativa/Info/Index. Поскольку геном риса хорошо изучен, также имеются данные по наличию в нем различных видов повторов. Нами была рассмотрена для сравнения найденных IP методом ДП аннотация, полученная конвейером EDTA, который включает в себя 7 различных алгоритмов выявления повторов.

Результаты: Геном *O. sativa* содержит 12 хромосом общей длины около 370 Мб. Из-за большого размера генома для генерации ПВМ семейств ДП брались только 6 и 11 хромосомы. По этому обучающему множеству было построено 79 семейств повторов. Максимальная длина ДП ограничивалась 600 н.п. С помощью ПВМ семейств была просканирована геномная последовательность риса в прямом и обратном направлении. Для устранения избыточности поиск ДП производился с тем условием, чтобы повторы, расположенные на одной нити ДНК, не пересекались между собой более чем на 50 н.п. Для оценки ошибки предсказания проводился поиск ДП в случайных последовательностях. По каждому семейству ДП определялся уровень FDR (false discovery rate). На уровне FDR менее 5% в геноме риса было обнаружено 992,739 ДП, при этом число повторов на прямой и обратной нити ДНК примерно одинаково. Повторы покрывают чуть более 66% генома *O. sativa*. Основная часть ДП имеют длины в диапазонах 100–300 н.п., а также в районах 470 и 590 н.п. Исследование взаимного расположения экзонов и ДП показало, что преимущественно ДП располагаются в межгенных областях.

Конвейер EDTA смог обнаружить в геноме риса 386,672 ДП, которые покрывают 46.4% генома. Большинство ДП короткие 50-300 н.п. Мы также изучали пересечение ДП найденных методом IP и EDTA. В ряде случаев наблюдается значимая корреляция между построенными нами семействами и классами ДП, выявляемых EDTA.

Выводы: С помощью IP метода было обнаружено в 2,5 раза большее ДП в геноме *O. sativa*, чем было выявлено конвейером EDTA. Это сильнодивергированные ДП средней длины. В диапазоне коротких до 200 н.п. длин повторов метод IP использовать нецелесообразно. В этом случае при построении выравнивания по размытой профильной матрице с 16 строками невозможно достичь необходимого уровня статистической значимости. В то же время существующее ПО, что было показано на примере конвейера EDTA, ориентировано на определение достаточно коротких 50-250 н.п. консервативных повторов. В этой связи оптимальным решением для получения всеобъемлющей аннотации ДП в геномной последовательности будет использование популярных de novo подходов в совокупности с IP методом.

ДП покрывают значительную часть генома *O. sativa*. Нами дополнительно была проанализирована последовательность риса Индика R498. Сканирование этой сборки теми же ПВМ, которые мы построили ранее, продемонстрировало наличие более миллиона ДП, покрытие генома превышает 69%. Таким образом, множественное присутствие ДП является характерным для различных видов риса. Вероятная функция ДП в геноме состоит в формировании пространственной структуры молекулы ДНК.

Финансирование: Исследование поддержано грантом РНФ № 24-24-00031.