Новые технологии секвенирования – новый уровень исследования геномного разнообразия

Технологии секвенирования геномов третьего поколения позволяют исследовать последовательность ДНК с недоступной ранее точностью. Эти технологии использовали для секвенирования 64 гаплотипов от 32 индивидов из 25 популяций Африки, Европы, Северной и Южной Америки, Восточной и Южной Азии. В работе выявлено 107 590 структурных вариаций генома (инсерции, делеции, дупликации, инверсии), 68% которых не было описано ранее. В их числе авторы обнаружили 117 структурных вариаций, связанных с происхождением и имеющих разную частоту в популяциях; некоторые из них стали мишенями локальных адаптаций.

В журнале Science опубликовано самое точное на сегодняшний день исследование геномного разнообразия в человеческих популяциях, это работа большого международного коллектива при ведущей роли специалистов Медицинской школы Университета Мэриленд в США. Используя новейшие технологии секвенирования и картирования генома, ученые получили детальные карты 64 гаплотипов от 32 индивидов, относящихся к 25 популяциям Африки, Северной и Южной Америки (коренное население), Восточной Азии, Европы и Южной Азии. Это позволило описать огромное число структурных вариаций генома и мобильных элементов, ранее не найденных в эталонном геноме человека. При этом не использовался подход секвенирования «троек» (папа, мама, ребенок).

До сих пор секвенирование геномов было основано на прочтении и сборке коротких последовательностей (ридов). Это позволяло сравнивать геномы по однонуклеотидному разнообразию (SNP), но таким путем было трудно выявлять структурные вариации (SV) – инсерции, делеции, дупликации, инверсии, затрагивающие более крупные участки ДНК. Новые технологии секвенирования позволяют читать длинные риды и собирать геномы из них. В данной работе авторы использовали технологии одномолекулярного секвенирования в реальном времени (Pacific Biosciences), а также секвенирования одноцепочечной матрицы (Strand-seq). Этот метод был разработан в консорциуме Human Genome Structural Variation Consortium (HGSVC).

В результате генетикам удалось собрать 64 гаплотипа с исключительной точностью, эти гаплотипы они сравнивали с эталонным геномом человека и с данными проекта 1000 Genomes Project. Полученные результаты позволили описать весь спектр геномного разнообразия. В него вошли 107 590 структурных вариаций генома, из которых 68% не были описаны ранее, при секвенировании по технологии сборки коротких ридов; более 300 инверсий, 2,3 миллиона инделов и 15,8 миллионов SNV (single-nucleotide variant, замена одного нуклеотида). Они нашли также около 9 500 не описанных ранее мобильных элементов.

Структурные вариации неравномерно распределены по геному, они объединены в кластеры, и ученые обнаружили 278 «горячих точек» SV, где они затрагивают богатые генами участки ДНК. В кодирующей части генома нашли более 2 100 локусов количественных признаков (QTL), связанных со структурной изменчивостью, 42% этих локусов не было описано ранее. Эти локусы связаны с экспрессией 1 526 генов, многие из которых имеют отношение к различным заболеваниям. Таким образом, как пишут авторы, перекидывается мостик между молекулярной структурой и клинически значимыми фенотипами. Среди выявленных QTL 75% представляют собой инсерции, но есть и клинически значимые делеции, например делеция размером 1 069 kbp вблизи гена LIPI, который ассоциирован с сердечной недостаточностью.

Поскольку секвенированные геномы представляли 25 популяций с разных континентов, ученые провели поиск структурных вариаций, информативных для происхождения. Для этого они применили метод машинного обучения и скрытую марковскую модель. Анализ показал, что 87,2% изменчивости связано с происхождением. В геномах индивидов из Африки найдены наиболее длинные участки единого происхождения, в отличие от индивидов из Северной или Южной Америки, где преобладали участки смешанного происхождения, отражающие трансатлантические миграции в истории популяций.

В геномах разных популяций авторы провели поиск структурных вариаций, ставших мишенями для локальных адаптаций (популяционно-стратифицированные вариации). Их частота варьирует в различных популяциях, что отражает происхождение, прохождение через бутылочное горлышко и смешение популяций. Они нашли 117 таких популяционно-стратифицированных SV, различающихся по частоте в разных группах.

Один из примеров – инсерция размером 4,0 kbp в первом интроне гена толерантности к лактозе (LCT), которая содержит 11 сайтов связывания транскрипционных факторов. Эта инсерция имеет четкую связь с европейским происхождением. Вариации гена LCT – один из наиболее известных примеров адаптивной эволюции у европейцев. Другой пример – инсерция размером 2,8 kbp в районе гена CLEC16A, связанного с диабетом I типа, этот вариант с высокой частотой встречается в населении Америки, с максимальной — у жителей Перу, но довольно редок в остальных популяциях. Частота диабета I типа в Перу – одна из самых высоких.

Для сравнения, в проекте 1000 Genomes Project было найдено ограниченное количество наиболее частых структурных

вариаций генома — 69 тысяч против 107 136 в данной работе. Полученные результаты гораздо полнее отражают генетическое разнообразие популяций в разных регионах мира. Они позволят включить новые структурные вариации в полногеномные исследования ассоциаций (GWAS), что будет способствовать развитию персонализированной медицины.

текст Надежды Маркиной

Источник:

Peter Ebert et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation // Science 25 Feb 2021: eabf7117 DOI: 10.1126/science.abf7117

Статья в свободном доступе