УДК: 57.087

УВЕЛИЧЕНИЕ ЭФФЕКТИВНОСТИ ПОИСКА CNV ПО ДАННЫМ ЭКЗОМНОГО СЕКВЕНИРОВАНИЯ

В.Д. Гордеева[1,2], К.А. Бабалян[1,2], Р.И. Султанов[1,2], Г.П. Арапиди[1,2,3], Э.В. Генерозов[1,2], В.М. Говорун[1,2]

[1]ФНКЦ физико-химической медицины ФМБА России; [2]Московский физико-технический институт (Государственный университет); [3]Институт биоорганической химии им. М.М. Шемякина и Ю.А. Овчинникова РАН, Москва, Россия

NGS-технологии наряду с цитогенетическими и чиповыми методами способны идентифицировать участки ДНК, число копий которых варьируется в человеческой популяции (CNV) [1]. С помощью секвенирования полного экзома (WES) можно за сравнительно небольшие деньги охарактеризовать все кодирующие участки, нарушения в которых могут иметь серьезные последствия. Существующие алгоритмы поиска CNV по экзомным данным не способны предсказывать все вариации и их результаты достаточно плохо согласуются между собой, поскольку поиск усложняется из-за ряда факторов, таких как GC-состав, сложность участка ДНК, наличие повторов, дискретная структура экзома и т.д [2,3].

В рамках этой работы мы разрабатывали взвешенный подход к оценке CNV на основе подбора оптимального набора единичных алгоритмов (ансамбля) и грамотного выбора валидационных данных. Мы интегрировали данные 17 крупномасштабных исследований по поиску CNV [4] и определили наличие или отсутствие CNV более чем в 110 тысячах экзонов для образца "Геном в бутылке" (NA12878). Для построения модели были использованы данные полноэкзомного секвенирования 3 фазы проекта 1000 Геномов [5] и результаты 5 алгоритмов поиска CNV (EXCAVATOR2, ExomeDepth, XHMM, CONIFER, cn.MOPS). Каждый экзон был охарактеризован 203 факторами, среди которых были геномные характеристики локуса и нормализованное на разных масштабах покрытие.

По результатам этой работы мы предлагаем два альтернативных подхода построения оптимального ансамбля, один из которых идентифицирует больше достоверных вариаций (precision=49%, recall=18%, AUC=0.74), а другой дополняет результаты единичных алгоритмов (precision=59%, recall=2%, AUC=0.62).

Ключевые слова: алгоритм поиска вариации по числу копий (CNV), экзомное секвенирование, машинное обучение, ансамблевый подход

Литература

1. Lee, J. Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels / J. Lee, J. Jeon // Cytogenet Genome Res. – 2008. – Vol. 123, № 1–4. – P. 333–342.
2. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives / M. Zhao et al. // BMC Bioinformatics. – 2013. – Vol. 14 (Suppl 11). – S1.
3. Identification of copy number variants from exome sequence data / P. Samarakoon et al. // BMC Genomics. – 2014. – № 15. – P. 661.
4. The Database of Genomic Variants: a curated collection of structural variation in the human genome / J. MacDonald et al. // Nucleic Acids Res. – 2014. – № 42. – P. D986–92.
5. A global reference for human genetic variation / 1000 Genomes Project Consortium // Nature. – 2015. – № 526. – P. 68–74.