УДК 575.17

РАЗРАБОТКА МЕТОДА ДЛЯ ГЕНЕРАЦИИ GWAS ДАННЫХ ЧЕЛОВЕКА

Чангалиди А.И. (Университет ИТМО, Институт биоинформатики, Отдел геномной медицины ФГБНУ “НИИ АГиР им. Д.О. Отта”)
Научный руководитель – к.б.н., **Барбитов Ю.А.** (Отдел геномной медицины ФГБНУ “НИИ АГиР им. Д.О. Отта”)

Целью данного исследования является разработка метода для генерации данных полногеномного анализа ассоциаций (англ. genome-wide association studies, GWAS) с заданными параметрами (наследуемость, количество каузальных (причинных) вариантов, процент вариантов, располагающихся в каузальных генах), проверка производительности и корректности его работы.

Введение. В последние годы популярным стало такое исследование генома, как полногеномный поиск ассоциаций, в ходе которого исследователи идентифицируют вариации (в наиболее частом случае – однонуклеотидные полиморфизмы (англ. single nucleotide polymorphism, SNP)), которые ассоциированы с различными признаками (например, предрасположенность к заболеваниям, таким как рак, сахарный диабет, различным осложнения при беременности) [1]. Для разработки алгоритмов анализа нужны генотипические и фенотипические данные, на которых можно было бы тестировать разрабатываемые методы. Однако в данный момент лишь некоторые ресурсы предоставляют открытый доступ к таким данным. Существующие же алгоритмы генерации данных предоставляют очень малый выбор параметров симуляции, приводящий к генерации нереалистичных данных [2,3]. С целью увеличить доступность реалистичных данных с нужными характеристиками и был разработан предлагаемый алгоритм.

Основная часть. Алгоритм делится на 3 этапа: генерация данных генотипирования (вариации в геноме), генерация соответствующих фенотипических данных (признаков), а также проведение анализа ассоциаций между сгенерированными генотипами и фенотипами.

Первым шагом является генерация генотипических данных. Для того, чтобы результаты GWAS получались статистически достоверными, необходимо иметь выборку генотипов размером несколько тысяч человек. Однако, в открытом доступе имеется не так много публичных генотипических данных. Одним из таких является набор данных проекта “1000 Геномов” [4]. В нем около 500 образцов европейского происхождения. Чтобы получить больший объём выборки, предлагается с помощью HAPGEN v2.2.0 [5] осуществить симуляцию 5000 генотипов из имеющихся за счет случайных перестановок.

Далее необходимо провести симуляцию фенотипических данных. Для начала из полученных гаплотипов необходимо выбрать каузальные SNP из интересующих наборов генов, а также добавить случайные варианты в соответствии с требуемой долей каузальных вариантов из интересующего набора генов.

1. С помощью инструмента bedtools v2.30.0 [6] файлы с координатами SNP аннотируются: для каждого SNP получается информация о ближайшем гене, а так же его частота минорной аллели (англ. minor allele frequency, или MAF).
2. Далее происходит выборка SNP из полученного списка. Пусть в выбранном наборе генов есть n генов, и k SNP из этого набора должны быть значимы в генерируемых данных, всего же влияют на фенотип K SNP (включая k SNP из пути ($k \leq K$), остальные – случайные). Тогда алгоритм симуляции выглядит следующим образом:
    a. Случайно выбирается k генов из n генов, входящих в изучаемый набор;

b. Из таблицы, полученной на шаге (1), выбирается k SNP, соответствующих выбранным на шаге (2a) генам.

c. Добавляется $K - k$ случайных SNP (не принадлежащих n генам изучаемого набора)

d. Получается набор из K каузальных вариантов, k из которых соответствуют заданному набору генов

3. Используя PLINK v2.0 [7, 8] из bed-файла со всеми сгенерированными ранее генотипами извлекаются генотипы по отобранным SNP.

4. Используя этот файл, а также заданный коэффициент наследуемости, с помощью phenotypeSimulator (R-пакет, v0.3.4) [9] генерируются симулированные фенотипические данные.

Последним этапом является произведение анализа ассоциаций каждого варианта в сгенерированных генотипических данных с полученными значениями фенотипа.

Выводы. Для упрощения пользования алгоритмом, описанные ранее шаги были объединены в единый протокол, а также была произведена проверка корректности его работы и производительности.

Список использованных источников

[1] "Genome-Wide Association Studies". National Human Genome Research Institute. (2023) [Электронный ресурс]. Режим доступа: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet, свободный (15.02.2023).

[2] Li, C., & Li, M. (2008). GWAsimulator: a rapid whole-genome simulation program. (англ.) // Bioinformatics (Oxford, England), 24(1), 140–142. https://doi.org/10.1093/bioinformatics/btm549

[3] Fortune, M. D., & Wallace, C. (2019). simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics. (англ.) // Bioinformatics (Oxford, England), 35(11), 1901–1906. https://doi.org/10.1093/bioinformatics/bty898

[4] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. (англ.) // Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

[5] Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. (англ.) // Bioinformatics (Oxford, England), 27(16), 2304–2305. https://doi.org/10.1093/bioinformatics/btr341

[6] Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. (англ.) // Bioinformatics (Oxford, England), 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

[7] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. (англ.) // GigaScience, 4, 7. https://doi.org/10.1186/s13742-015-0047-8

[8] Authors : Purcell, S. M., Chang, C. C.. Package : PLINK 2.0. [Электронный ресурс]. Режим доступа: www.cog-genomics.org/plink/2.0/, свободный (27.02.2023).

[9] Meyer, H. V., & Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. (англ.) // Bioinformatics (Oxford, England), 34(17), 2951–2956. https://doi.org/10.1093/bioinformatics/bty197