

Оценка эффективности работы инструментов для функциональной аннотации результатов GWAS с использованием симулированных данных

Барбитов Ю.*, Чангалиди А., Глотов А.

ФГБНУ «НИИ акушерства, гинекологии и репродуктологии им. Д.О. Отта»,

Санкт-Петербург, Россия

* barbitoff@bk.ru

Ключевые слова: GWAS; сложные признаки; генетический вариант; симулированные данные

Мотивация и цель: Полногеномный анализ ассоциаций (genome-wide association studies, GWAS) является важнейшим подходом для поиска генетических основ сложных признаков человека и других организмов. За последние десятилетия были проведены сотни успешных GWAS-исследований; в то же время аннотация и интерпретация обнаруженных ассоциаций до сих пор остаются сложной задачей. Разработка эффективных методов для функциональной аннотации результатов GWAS затрудняется отсутствием готовых данных с известным набором причинных генетических вариантов и биологических процессов, вовлеченных в развитие изучаемого признака. Целью настоящей работы являлось создание метода для симуляции GWAS-данных с предварительно заданным набором причинных вариантов и/или процессов, а также применение этого инструмента для оценки точности работы инструментов для поиска обогащения наборов генов в результатах GWAS.

Методы и алгоритмы: В ходе работы был разработан новый инструмент, bioGWAS, позволяющий проводить симуляции реалистичных наборов GWAS-данных. bioGWAS был реализован в виде биоинформатического протокола на языке Snakemake, включающего в себя стадии симуляции генотипов (с использованием HAPGEN2 [1]), фенотипов (при помощи PhenotypeSimulator [2]) и результатов анализа ассоциаций между ними. Разработанный инструмент был применен для оценки эффективности работы часто используемых методов для поиска обогащенных наборов генов в результатах GWAS: MAGMA [3] и Pascal [4], а также разработанного ранее в нашей научной группе инструмента LSEA. Исходный код и текущая версия bioGWAS доступны по адресу <https://github.com/TohaRhymes/bioGWAS/> [5].

Результаты: Разработанный инструмент для симуляции GWAS-данных продемонстрировал свою высокую эффективность для генерации результатов GWAS с заданным набором причинных вариантов и наборов генов, в которых расположены данные варианты (показатель F1 при обнаружении заданных вариантов в качестве значимых ассоциаций выше 0.95). Использование данных, сгенерированных с помощью bioGWAS, для оценки эффективности идентификации заданных наборов генов инструментами MAGMA, Pascal и LSEA показал значительное влияние размера целевого набора генов на чувствительность обнаружения обогащения, а также позволил сделать вывод о значительном преимуществе метода LSEA над аналогами (чувствительность LSEA составила от

83 до 100 % для разных наборов генов, чувствительность других инструментов варьировала в диапазоне от 5 до 100 %).

Выводы: Разработанный в ходе работы инструмент bioGWAS расширяет арсенал подходов для создания и оценки эффективности методов аннотации результатов GWAS, а инструмент LSEA позволяет с высокой чувствительностью обнаруживать биологические процессы, вовлеченные в формирование сложных признаков.

Финансирование: Исследование поддержано Министерством науки и высшего образования Российской Федерации (проект «Многоцентровая научно-исследовательская коллекция биоресурсов «Репродуктивное здоровье человека»» (контракт № 075-15-2021-1058 от 28 сентября 2021 г.), а также стипендией по системной биологии, полученной Барбитовым Ю.А. в 2021–2023 гг.

Assessing the performance of tools for functional annotation of GWAS results using simulated data

Barbitoff Y.*, Changalidis A., Glotov A.

Research Institute of Obstetrics, Gynecology and Reproductology named after D.O. Ott, St. Petersburg, Russia

* barbitoff@bk.ru

Key words: GWAS; complex traits; genetic variant; simulated data

Motivation and Aim: Genome-wide association study (GWAS) is among the most important approaches for finding the genetic basis of complex traits in humans and other organisms. Hundreds of successful GWAS studies have been conducted over the past decades; however, annotation and interpretation of the detected associations is still a challenging task. The development of efficient methods for functional annotation of GWAS results is complicated by the lack of readily available datasets with a known set of causal genetic variants and biological processes involved in the development of the studied trait. The goal of this work was to create a method for simulating GWAS datasets with a predefined set of causal variants and/or processes, and to apply this tool to evaluate the accuracy of tools for finding gene set enrichment in GWAS results.

Methods and Algorithms: In this work, we developed a new tool, bioGWAS, to enable simulations of realistic GWAS datasets. bioGWAS was implemented as a bioinformatic pipeline using the Snakemake toolkit; it includes simulation of genotypes (using HAPGEN2 [1]), phenotypes (using PhenotypeSimulator [2]), and summary statistics of their association. The developed tool was applied to evaluate the performance of commonly used methods for finding enriched gene sets in GWAS results: MAGMA [3] and Pascal [4], as well as the LSEA tool developed earlier in our research group. The source code and the current version of bioGWAS are available at <https://github.com/TohaRhymes/bioGWAS/> [5].

Results: Our tool demonstrated its high efficiency for generating GWAS results with a given set of causal variants and gene sets in which these variants are located (F1 score for detecting predefined variants as significant associations is more than 0.95). We next used the data generated by bioGWAS to evaluate the efficiency of identification of given gene sets by MAGMA, Pascal, and LSEA tools. We found a significant effect of the size of the target gene set on the sensitivity of enrichment analysis; at the same time, results

showed that the LSEA method has a significant advantage over its competitors (LSEA sensitivity ranged from 83 to 100 % for different gene sets, while the sensitivity of other tools ranged from 5 to 100 %).

Conclusion: The bioGWAS tool expands the arsenal of approaches for development and evaluation of the performance of GWAS results annotation methods, while the LSEA tool allows for highly sensitive detection of biological processes involved in complex trait formation.

Funding: The study is supported by the Ministry of Science and Higher Education of the Russian Federation (project “Multicenter research bioresource collection” “Human Reproductive Health” contract No. 075-15-2021-1058 from September 28, 2021), as well as by Systems Biology Fellowship to Y.A.B. in 2021–2023.

Список литературы/References

1. Su Z., Marchini J., Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*. 2011;27(16):2304-2305. doi 10.1093/bioinformatics/btr341
2. Meyer H.V., Biner E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*. 2018;34(17):2951-2956. doi 10.1093/bioinformatics/bty197
3. de Leeuw C.A., Mooij J.M., Heskes T., Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol*. 2015;11(4):e1004219. doi 10.1371/journal.pcbi.1004219
4. Lamparter D. et al. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol*. 2016;12(1):e1004714. doi 10.1371/journal.pcbi.1004714
5. Changelidis A.I. et al. bioGWAS: A Simple and Flexible Tool for Simulating GWAS Datasets. *Biology (Basel)*. 2023;13(1):10. doi 10.3390/biology13010010