

Являются ли импутированные генотипы альтернативой данных экзомного секвенирования при анализе редких вариантов?

Зоркольева И.В.*, Аксенович Т.И., Цепилов Я.А.

Институт цитологии и генетики СО РАН, Новосибирск, Россия

* zor@bionet.nsc.ru

Ключевые слова: анализ ассоциаций на основе генов; биобанк Великобритании; бинарные признаки; количественные признаки; белок-кодирующие варианты; некодирующие варианты

Мотивация и цель: Полногеномные исследования ассоциаций (GWAS) – эффективный способ изучения ассоциаций генотип-фенотип. Этот подход позволил получить значительное количество надежных ассоциаций для многих сложных признаков и распространенных заболеваний человека. Но результаты GWAS объясняют лишь малую долю наследуемости. Для объяснения феномена «потерянной наследуемости» было предложено несколько подходов, в том числе анализ редких вариантов [1]. Такой анализ стал возможен с развитием технологий секвенирования экзонов. Однако широкое использование генотипов, полученных с помощью секвенирования, в настоящее время ограничено относительно небольшим количеством секвенированных образцов в биобанках и отсутствием свободного доступа к персональным генотипам. Напротив, результаты GWAS для тысяч признаков и заболеваний, полученные на сотнях тысяч образцов, сегодня находятся в свободном доступе (например, каталог GWAS) и могут быть использованы для поиска ассоциаций на основе генов (gene-based analysis), который является основным инструментом для анализа редких (частота минорного аллеля, $MAF < 0.01$) и ультраредких ($MAF < 0.001$) генетических вариантов [2, 3]. Целью данного исследования является проверка гипотезы о том, что импутированные генотипы могут являться дополнительным источником информации для анализа редких вариантов [4]. Для этого мы сравнили результаты анализа редких вариантов на основе генов с использованием импутированных и секвенированных генотипов, полученных на выборке одних и тех же участников проекта UK Biobank.

Методы и алгоритмы: Исследование проведено на выборке европейцев из проекта UK Biobank ($N = 188\,236$). Использованы секвенированные (11334327 аутосомных SNVs, $MAF < 0.01$) и импутированные генотипы (4258875 аутосомных SNPs, $INFO > 0.8$, missingness rate < 0.02 и $5 \times 10^{-5} < MAF < 0.01$).

Были проанализированы 20 бинарных (случай-контроль) и 27 количественных признаков. Одноточечный анализ ассоциаций для каждого признака проводили с помощью пакета fastGWA-GLMM, версия 1.94.0 beta [5].

Для анализа ассоциаций на основе генов использовали суммарные статистики (z-scores и размеры эффекта betas) и матрицы корреляций между генотипами всех вариантов в пределах гена. Анализ проводили для каждого из трех наборов вариантов: 1) несинонимичные (LoF + missense), 2) все кодирующие, 3) все

некодирующие варианты в гене. Перед анализом секвенированных данных генотипы ультраредких вариантов (число минорных аллелей, $MAC \leq 10$) были объединены в один вариант [6]. Использовались два регрессионных метода, основанных на суммарных статистиках: SKAT-O [7] и PCA [8], реализованных в пакете sumFREGAT R [3]. Результаты методов SKAT-O и PCA были объединены с помощью агрегированного омнибус-теста Коши, ACAT-O [9]. Анализ ограничивался белок-кодирующими генами с не менее чем двумя вариантами, имеющими суммарные статистики.

Результаты: Результаты анализа представлены в табл. 1.

Таблица 1. Сравнение числа сигналов ассоциации ($p\text{-value} < 2.5 \times 10^{-6}$) между импутированными (Имп) и секвенированными (Сек) генотипами для трех наборов вариантов в гене

Тип признака	Несинонимичные			Кодирующие			Некодирующие		
	Всего	Имп	Сек	Всего	Имп	Сек	Всего	Имп	Сек
Бинарный	25	2	24	20	3	18	18	14	4
Количественный	579	257	538	648	334	594	1610	1530	318

Количество сигналов ассоциации, выявленных для бинарных признаков, было меньше, чем для количественных, вследствие низкой распространенности бинарных признаков ($\min = 0.061$, $\text{mean} = 0.097$, $\max = 0.292$). Больше число ассоциаций было выявлено для несинонимичных и кодирующих вариантов при использовании секвенированных данных по сравнению с импутированными. Для количественных признаков секвенированные данные были примерно в два раза эффективнее импутированных данных по числу идентифицированных генов. Для некодирующих вариантов импутированные генотипы позволили выявить больше сигналов ассоциации, чем секвенированные данные. Для количественных признаков это соотношение составило 4.8. Некоторые сигналы ассоциации были обнаружены с помощью импутированных генотипов, но не обнаружены с помощью секвенированных.

Выводы: Мы показали, что импутированные генотипы могут быть использованы для анализа редких кодирующих вариантов, особенно когда секвенированные данные недоступны. Для некодирующих областей генов импутированные данные даже более информативны, чем данные, полученные с помощью секвенирования экзона. Мы считаем, что импутированные генотипы являются эффективным и при этом доступным источником данных для анализа редких вариантов.

Финансирование: Исследование поддержано грантом РФФИ № 23-25-00209.

Are imputed genotypes an alternative to exome sequencing data when analysing rare variants?

Zorkoltseva I.V.*, Axenovich T.I., Tsepilov Y.A.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* zor@bionet.nsc.ru

Key words: gene-based association analysis; UK biobank; binary traits; quantitative traits; protein-coding variants; non-protein-coding variants

Motivation and Aim: Genome-wide association studies (GWAS) are a powerful tool for studying genotype-phenotype associations. This methodology has generated a myriad of robust associations for a range of complex traits and common human diseases. However, GWAS results explain only a small fraction of heritability. Several approaches have been proposed to explain the phenomenon of missing heritability, including the analysis of rare variants [1]. Such analysis has become possible with the development of exome sequencing technologies. However, the widespread use of sequencing-derived genotypes is now limited by the relatively small number of sequenced samples in biobanks and the lack of free access to personal genotypes. On the contrary, GWAS results for thousands of traits and diseases from hundreds of thousands of samples are now freely available (e. g. GWAS catalogue) and can be used for gene-based association analysis, which is the main tool for analysing rare (minor allele frequency, $MAF < 0.01$) and ultra-rare ($MAF < 0.001$) genetic variants [2, 3]. The aim of this study is to test the hypothesis that imputed genotypes can be an additional source for rare variant analysis [4]. To do this, we compare the results of gene-based association analysis of rare variants using imputed and sequenced genotypes from the same participants in the UK Biobank project.

Methods and Algorithms: The study was conducted using individuals of white European ancestry from the UK Biobank with both exome sequenced (11334327 autosomal SNVs with $MAF < 0.01$) and imputed genotypes (4258875 autosomal SNPs with $INFO > 0.8$, missingness rate < 0.02 and $5 \times 10^{-5} < MAF < 0.01$) ($N = 188236$).

We analysed 20 binary (case-control) and 27 quantitative traits. A single variant association analysis for each trait was performed using the fastGWA-GLMM tool, version 1.94.0 beta [5].

For gene-based association analysis, summary statistics (z scores and effect sizes) and matrices of correlations between genotypes of all variants within a gene were used.

A gene-based association analysis was performed for each of three different variant annotations: LoF + missense, all coding and all non-coding variants in the gene. Before testing the sequenced data for each variant annotation, the genotypes of ultra-rare variants (minor allele count, $MAC \leq 10$) were collapsed into a single variant [6]. Two regression-based methods based on summary statistics were used: SKAT-O [7] and PCA [8], implemented in the sumFREGAT R package [3]. The results of SKAT-O and PCA were combined using the aggregated Cauchy omnibus test, ACAT-O [9]. Analysis was restricted to protein-coding genes with at least two variants that had the summary statistics.

The Bonferroni adjusted significance level for the total number of genes (20000) was defined as 2×10^{-6} .

Results: The results of the analyses are presented in Table 1.

Table 1. Comparison of the number of signals ($p\text{-value} < 2.5 \times 10^{-6}$) between imputed (imput) and sequenced (seq) genotypes for binary and quantitative traits. A comparison is shown for each annotation analysed

Trait	Lof+missence			Coding			Non-coding		
	All	Imput	Seq	All	Imput	Seq	All	Imput	Seq
Binary	25	2	24	20	3	18	18	14	4
Quantitave	579	257	538	648	334	594	1610	1530	318

The number of association signals observed for binary traits was lower than for quantitative traits due to the low prevalence of binary traits (min = 0.061, mean = 0.097,

max = 0.292). More associations were identified for non-synonymous and coding variants when using sequenced data compared to imputed data. For quantitative traits, sequenced data were about twice as efficient as imputed data in terms of gene number. For non-coding variants, the imputed genotypes detected more association signals than the sequenced data. For quantitative traits, this ratio was 4.8. There are some association signals that were detected with the imputed genotypes but not with the sequenced ones. **Conclusion:** We showed that the imputed genotypes can be useful for analysing coding variants, especially when sequenced data are not available. For non-coding regions of genes, imputed data are even more informative than data obtained by exome sequencing. We believe that imputed genotypes are a reasonable source for analysing rare variants. **Funding:** The study is supported by the Russian Science Foundation (RSF) grant No. 23-25-00209.

Список литературы/References

1. Wainschtein P., Jain D., Zheng Z. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet.* 2022;54(3):263-273. doi 10.1038/s41588-021-00997-7
2. 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
3. Svishcheva G.R., Belonogova N.M., Zorkoltseva I.V., Kirichenko A.V., Axenovich T.I. Gene-based association tests using GWAS summary statistics. *Bioinformatics.* 2019;35(19):3701-3708. doi 10.1093/bioinformatics/btz172
4. Wuttke M., Konig E., Katsara M.A., Kirsten H., Farahani S.K., Teumer A. et al. Imputation-powered whole-exome analysis identifies genes associated with kidney function and disease in the UK Biobank. *Nat Commun.* 2023;14(1):1287. doi 10.1038/s41467-023-36864-8
5. Jiang L., Zheng Z., Fang H., Yang J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet.* 2021;53(11):1616-1621. doi 10.1038/s41588-021-00954-4
6. Zhou W., Bi W., Zhao Z., Dey K.K., Jagadeesh K.A., Karczewski K.J. et.al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet.* 2022;54:1466-1469. doi 10.1038/s41588-022-01178-w
7. Li B., Leal S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008;83(3):311-321. doi 10.1016/j.ajhg.2008.06.024
8. Wang K., Abbott D. A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol.* 2008;32(2):108-118. doi 10.1002/gepi.20266
9. Liu Y., Chen S., Li Z., Morrison A.C., Boerwinkle E., Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet.* 2019;104(3):410-421. doi 10.1016/j.ajhg.2019.01.002