

Новая стратегия условного анализа генных ассоциаций

Свищёва Г.Р.*, Белоногова Н.М., Цепилов Я.А., Аксенович Т.И.

Институт цитологии и генетики СО РАН, Новосибирск, Россия

* gulsvi@bionet.nsc.ru

Ключевые слова: модель со случайными эффектами; анализ ассоциаций; условное распределение

Мотивация и цель: Анализ ассоциаций на основе генов (gene-based association analysis, GBAA) тестирует совместный эффект на признак всех однонуклеотидных полиморфизмов (SNP) внутри гена. Как правило, материалом для этого анализа являются z-статистики, полученные в результате полногеномного анализа ассоциаций (GWAS), и матрицы корреляций между генотипами SNP внутри генов. Как для GWAS, так и для GBAA обязательным элементом является условный анализ, позволяющий исключать влияние внегенных SNP, находящихся в неравновесии по сцеплению с SNP внутри гена. Одним из известных инструментов условного анализа является метод GCTA-COJO [1]. Этот метод корректирует z-статистики для каждого SNP внутри гена с учетом внегенных сигналов. Скорректированные z-статистики вместе с матрицами корреляций затем используются в GBAA. Простой и быстрой альтернативой COJO для GBAA является метод полигенного прунинга (polygenic pruning, PP) [2]. В его основе лежит исключение из рассмотрения внутригенных SNP, высоко коррелирующих с более значимо ассоциированными SNP вне гена. Оба этих метода изменяют число или значения z-статистик, участвующих в GBAA, не меняя матрицы генетических корреляций. Недостатком COJO являются вычислительные проблемы при обращении матрицы, а недостатком PP – потенциальное снижение мощности из-за удаления некоторых SNP. Здесь мы представляем новый метод условного анализа TauCOR, лишенный этих ограничений. Стратегия TauCOR состоит в том, что он не меняет значений z-статистик, а корректирует матрицы корреляций между генотипами SNP внутри гена. Такой подход позволяет избежать многих вычислительных проблем.

Методы и алгоритмы: Рассмотрим выборку генотипированных индивидов с измеренным фенотипом, для которой проведен GWAS. Для каждого гена мы определяем набор SNP внутри гена (g) и набор SNP из региона вокруг гена (r). Для этих наборов известны векторы z-статистик, z_r и z_g , и их корреляционные матрицы, U_g и U_r , соответственно. Кроме того, известна матрица корреляций между внутри- и внегенными SNP, U_{gr} . Задача состоит в том, чтобы скорректировать матрицу корреляций U_g , необходимую для GBAA, таким образом, чтобы учесть z_r , U_r и U_{gr} . В новом методе мы предполагаем, что эффекты SNP в регионе вокруг гена носят независимый случайный характер и распределены нормально с общим вкладом в изменчивость признака τ , а для эффектов SNP гена характер (случайный/фиксированный) на данном этапе не важен. Сначала мы анализируем SNP в регионе вокруг гена, чтобы оценить их вклад. Для этого мы строим распределение z_r с нулевым вектором матожиданий и корреляционной матрицей $U_r + \tau U_r U_r$. С помощью метода максимума правдоподобия, построенного для этого распределения, оцениваем τ . Затем анализируем ген интереса, рассмотрев для него

два потенциальных источника ассоциативного сигнала: SNP самого гена и SNP соседнего региона с уже известным вкладом τ . В этом случае нулевой гипотезой является предположение об отсутствии сигналов в гене, вызванных непосредственно SNP гена. В соответствии с нулевой гипотезой мы корректируем корреляционную матрицу U_g так, чтобы она отражала сигналы внегенных SNP, в результате чего получаем $U_g + \tau U_{gr} U_{gr}^T$. Далее выполняем GBAA с помощью любого из тестов, используя в качестве входных данных первичные результаты GWAS и скорректированные матрицы корреляций. Наиболее популярные GBAA-тесты представлены в [3].

Результаты: TauCOR был протестирован на наборе генов ‘золотого стандарта’, т. е. генов, эффект которых на признак однозначно установлен. Было отобрано 28 генов ‘золотого стандарта’ на 13 признаках. Мы использовали z-статистики из UK Biobank для этих признаков и сравнили новый метод с методами COJO и PP. Для каждого из 28 генов ‘золотого стандарта’ были сформированы локусы в пределах ± 1 Мб от границ гена. Всего в этих локусах содержалось 394 гена помимо генов из списка ‘золотого стандарта’. В противовес истинно-положительным генам ‘золотого стандарта’ в данном исследовании мы условно рассматриваем остальные гены в локусах как истинно-отрицательные для данных признаков. Число таких генов на локус варьировало от 1 до 59 со средним 14.6. GBAA проводили с помощью R-пакета sumFREGAT [4]. Для каждого гена SNP были отфильтрованы по частоте минорного аллеля $\geq 10^{-4}$, аннотированы с использованием VEP и организованы в три набора: некодирующие, кодирующие и несинонимические SNP. Для каждого набора SNP был выполнен тест ACAT-O, объединяющий результаты тестов SKATO и PCA. Всего, с учетом перекрытия анализируемых локусов, было протестировано 84 набора SNP для истинно-положительных генов и 1211 наборов для истинно-отрицательных генов. При проведении условного анализа для каждого гена рассматривали регион вокруг гена ± 5 Мб для COJO и PP и ± 1.5 Мб для TauCOR.

Прежде всего, мы провели GBAA и отобрали значимо ассоциированные наборы SNP для каждого гена, показавшие P-значения $\leq 2.5 \times 10^{-6}$. Среди 84 тестов, выполненных на наборах SNP для генов ‘золотого стандарта’, было 29 значимых, а среди 1211 тестов, выполненных на наборах SNP для генов из их окружения, было 97 значимых. Для этих значимых тестов проводили условный анализ. Фиксировали число тестов, прошедших условный анализ, для генов ‘золотого стандарта’, чтобы оценить чувствительность метода, и число тестов для генов из соседних регионов, чтобы оценить специфичность методов. Результаты приведены в табл. 1.

Таблица 1. Показатели эффективности различных методов

	GBAA	COJO	PP	TauCOR
Гены золотого стандарта	29	12	15	16
Соседние гены	97	17	22	5
Чувствительность		0.41	0.52	0.55
Специфичность		0.82	0.77	0.95

Как видно, число значимых тестов, полученных новым методом для генов ‘золотого стандарта’, лишь немного больше, чем у существующих методов. Это свидетельствует о его чувствительности, сравнимой с чувствительностью других

методов условного анализа. Однако специфичность нового метода оказалась значительно выше, чем у существующих. TauCOR показал ложно-положительный результат только в пяти случаях, тогда как другие методы показывали такой результат в четыре раза чаще.

Выводы: Стратегия TauCOR продемонстрировала сравнимую чувствительность условного анализа и значительно более высокую его специфичность по сравнению с COJO и PP. Метод может быть успешно использован для повышения точности анализа ассоциаций на геномном уровне.

Финансирование: Исследование выполнено при финансовой поддержке гранта Российского научного фонда (РНФ) № 23-25-00209.

A new strategy for conditional gene-based association analysis

Svishcheva G.R.*, Belonogova N.M., Tsepilov Y.A., Axenovich T.I.

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

* gulsvi@bionet.nsc.ru

Key words: random effects model; association analysis; conditional distribution

Motivation and Aim: Gene-based association analysis (GBAA) tests the joint effect of all single nucleotide polymorphisms (SNPs) within a gene on a trait. As a rule, the input data for this analysis are z-statistics obtained by genome-wide association analysis (GWAS) and correlation matrices between genotypes of SNPs within genes. For both GWAS and GBAA, conditional analysis is a mandatory step to exclude the influence of extra-gene SNPs that are in linkage disequilibrium with SNPs within a gene. One of the well-known conditional analysis tools is the GCTA-COJO method [1]. This method adjusts the z-statistics for each SNP within a gene to account for the extra-gene signals. These adjusted z-statistics together with correlation matrices are then used in GBAA. A simple and fast alternative to COJO for GBAA is the polygenic pruning (PP) method [2]. PP is based on the exclusion of intra-gene SNPs that are highly correlated with more significantly associated SNPs outside the gene. Both of these methods change the number or values of z-statistics involved in GBAA without changing the matrix of genetic correlations. The disadvantage of COJO is the computational problems in inverting the matrix, and the disadvantage of PP is the potential reduction in power due to the removal of some SNPs. Here we present a new conditional analysis method, TauCOR, devoid of these limitations. The strategy of TauCOR is that it does not change the values of the z-statistics, but adjusts the correlation matrices between genotypes of SNPs within a gene. This approach avoids many computational problems.

Methods and Algorithms: Consider a sample of genotyped individuals with a measured phenotype for which a GWAS was performed. For each gene, we define a set of SNPs within the gene (g) and a set of SNPs from the region around the gene (r). For these sets, we know the vectors of z-statistics, z_g and z_r , and their correlation matrices, U_g and U_r , respectively. Furthermore, the correlation matrix between intra- and extra-gene SNPs, U_{gr} , is known. The task is to adjust the correlation matrix, U_g , required for GBAA so as to account for z_r , U_r and U_{gr} .

In the new method, we assume that the effects of SNPs in the region around the gene are independently random and normally distributed with a common contribution to trait variability τ , and for the effects of SNPs in the gene, the type (random/fixed) is not

important at this stage. We first analyse the SNPs in the region around the gene to estimate their contribution. To do this, we construct a z_r distribution with zero mean vector and correlation matrix $U_r + \tau U_r U_r$. Using the maximum likelihood method constructed for this distribution, we estimate τ . We then analyse the gene of interest by considering two potential sources of association signal for it: SNPs of the gene itself and SNPs of a neighbouring region with an already known contribution τ . In this case, the null hypothesis is the assumption that there are no signals in the gene caused directly by intragenic SNPs. Under this null hypothesis, we adjust the correlation matrix U_g so that it reflects signals from extra-gene SNPs, resulting in $U_g + \tau U_{gr} U_{gr}^T$. We then perform GBAA with either test, using the primary GWAS results and the adjusted correlation matrices as input data. The most popular GBAA tests are presented in [3].

Results: TauCOR was tested on a set of ‘gold standard’ genes, i.e. genes whose influence on a particular trait has been unambiguously established. Twenty-eight ‘gold standard’ genes were selected for 13 traits. We used z-statistics from UK Biobank for these traits and compared the new method with COJO and PP methods.

For each of the 28 ‘gold standard’ genes, loci were generated within ± 1 Mb of the gene boundaries. These loci contained a total of 394 genes in addition to the ‘gold standard’ genes. As a contrast to the true-positive genes from ‘gold standard’, in this study we conditionally consider the remaining genes in the loci as true-negatives for these traits. The number of such genes per locus ranged from 1 to 59, with an average of 14.6.

GBAA was performed using the R package sumFREGAT [4]. For each gene, SNPs were filtered by minor allele frequency $\geq 10^{-4}$, annotated using VEP and divided into three sets: non-coding, coding and non-synonymous SNPs. For each set of SNPs, the ACAT-O test was performed combining the results of the SKATO and PCA tests. A total of 84 SNP sets for true-positive genes and 1211 SNP sets for true-negative genes were tested, taking into account the overlap of the loci analysed. In the conditional analysis for each gene, we considered a region around the gene in size of ± 5 Mb for COJO and PP and ± 1.5 Mb for TauCOR.

First of all, we performed GBAA and selected significantly associated SNP sets for each gene that showed P-values $\leq 2.5 \times 10^{-6}$. Among the 84 tests performed on SNP sets for ‘gold standard’ genes, there were 29 significant tests, and among the 1,211 tests performed on SNP sets for genes from their surroundings, there were 97 significant tests. Conditional analysis was performed for these significant tests. To assess the sensitivity of the methods, the number of tests that passed the conditional assay for ‘gold standard’ genes was counted, and to assess the specificity of the methods, the number of tests for genes from neighbouring regions was counted. The results are summarized in Table 1.

Table 1. Effectiveness indicators of various methods

	GBAA	COJO	PP	TauCOR
‘Gold standard’ genes	29	12	15	16
Neighbouring genes	97	17	22	5
Sensitivity		0.41	0.52	0.55
Specificity		0.82	0.77	0.95

As can be seen, the number of significant tests obtained by the new method for ‘gold standard’ genes is only slightly higher than that of the existing methods. This indicates its sensitivity comparable to that of other conditional analysis methods. However, the specificity of the new method was significantly higher than that of existing methods.

TauCOR showed false-positive results in only 5 cases, while other methods showed such results four times more often.

Conclusion: The TauCOR's strategy demonstrated comparable sensitivity and significantly higher specificity of conditional analysis in comparison to COJO and PP. The method can be successfully used to improve the accuracy of association gene-based analyses.

Funding: The study was financially supported by the Russian Science Foundation (RSF) grant No. 23-25-00209.

Список литературы/References

1. Yang J., Lee S.H., Goddard M.E., Visscher P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82
2. Belonogova N.M., Zorkoltseva I.V., Tsepilov Y.A., Axenovich T.I. Gene-based association analysis identifies 190 genes affecting neuroticism. *Sci Rep.* 2021;11(1):2484
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
4. Belonogova N.M., Svishcheva G.R., Kirichenko A.V., Zorkoltseva I.V., Tsepilov Y.A., Axenovich T.I. sumSTAAR: a flexible framework for gene-based association studies using GWAS summary statistics. *PLOS Comput Biol.* 2022;18(6):e1010172