ГЕНЕТИКА, 2024, том 60, № 7, с. 91–99

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ

УДК 575.1

РЕКОНСТРУКЦИЯ МАТРИЦЫ ГЕНОТИПИЧЕСКИХ КОРРЕЛЯЦИЙ МЕЖДУ ВАРИАНТАМИ ВНУТРИ ГЕНА ДЛЯ СОВМЕСТНОГО АНАЛИЗА ИМПУТИРОВАННЫХ И СЕКВЕНИРОВАННЫХ ДАННЫХ

© 2024 г. Г. Р. Свищёва[1, 2,*], А. В. Кириченко[1], Н. М. Белоногова[1], Е. Е. Елгаева[1, 3], Я. А. Цепилов[1], И. В. Зоркольцева[1], Т. И. Аксенович[1]

[1]*Федеральный исследовательский центр, Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, 630090 Россия*

[2]*Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, 119991 Россия*

[3]*Новосибирский государственный университет, Новосибирск, 630090 Россия*

**e-mail: gulsvi@mail.ru*

Поступила в редакцию 29.01.2024 г.
После доработки 27.02.2024 г.
Принята к публикации 25.03.2024 г.

При объединении импутированных и секвенированных данных в одном анализе ассоциаций на уровне генов возникает проблема реконструкции матриц генетических корреляций. Она связана с тем, что для гена известны корреляции между всеми импутированными генотипами вариантов и корреляции между всеми секвенированными, но неизвестны корреляции между генотипами вариантов, один из которых импутирован, а другой секвенирован. Для реконструкции этих корреляций мы предлагаем эффективный метод, основанный на максимизации детерминанта матрицы. Этот метод обладает рядом полезных свойств и имеет аналитическое решение для нашей задачи. Апробация предложенного метода была выполнена путем сравнения реконструированных и реальных корреляционных матриц, построенных на индивидуальных генотипах из Биобанка Великобритании. Сравнение результатов анализа ассоциаций на генном уровне, выполненного методами SKAT, BT и PCA на реконструированных и реальных матрицах с использованием смоделированных и вычисленных по реальным фенотипам суммарных статистик, показало высокое качество реконструкции и устойчивость метода к различным структурам гена.

Ключевые слова: импутированные и секвенированные генотипы, анализ ассоциаций на уровне генов, генетические варианты, суммарные статистики.

DOI: 10.31857/S0016675824070089 **EDN:** BHMPLU

Анализ ассоциаций на генном уровне является одним из самых эффективных статистических методов, используемых для идентификации генов, контролирующих различные признаки и болезни человека [1–4]. Для этого анализа используются два типа генотипических данных: импутированные и секвенированные. Первый тип получается путем генотипирования большого числа распространенных вариантов, расположенных равномерно по геному, и последующего восстановления (импутации) генотипов промежуточных вариантов [5]. Второй получается в результате полноэкзомного секвенирования [6]. Эти два типа данных частично перекрываются, однако по большей части они характеризуют разные варианты в гене. Секвенированные генотипы описывают преимущественно белок-кодирующие последовательности, а импутированные – интронные [7, 8]. Каждый из этих типов широко используется для анализа ассоциаций на уровне гена. Проблема возникает, когда мы хотим объединить оба типа генотипических данных в одном анализе. Известно, что в отсутствие доступа к индивидуальным генотипам и фенотипам для проведения анализа ассоциаций на генном уровне необходимы суммарные статистики – результаты одноточечного анализа ассоциаций (z-статистики и размеры эффектов) для всех вариантов, идентифицированных в гене, а также матрица корреляций между генотипами всех этих вариантов.

Для генетического варианта, генотипированного двумя способами, результат одноточечного анализа, а именно z-статистика, определяется как взвешенная сумма результатов, полученных для каждого генотипа. Однако восстановление корреляций между вариантами с импутированными и секвенированными генотипами представляет проблему.

Эта проблема, которая определяется как поиск значений, заполняющих недостающую информацию в матрице корреляций, давно изучается в линейной алгебре и в общем случае не имеет единственного допустимого решения. В этой связи были предложены численные методы решения, которые определяют диапазоны неизвестных элементов матрицы [9–11]. Однако эти методы были разработаны для матриц маленькой размерности (< 5). Для больших корреляционных матриц существующие численные решения требуют существенных вычислительных затрат [12, 13]. Учитывая, что для анализа ассоциаций необходимо реконструировать матрицы генотипических корреляций примерно для 20 000 генов, многие из которых содержат большое число генотипированных вариантов, численные методы не подходят, и нам необходимо найти легко реализуемое аналитическое решение.

Особенностью рассматриваемых нами матриц корреляций является то, что все их недостающие элементы можно поместить в один блок. Тогда саму восстанавливаемую матрицу можно представить в блочном виде 3 × 3, где неизвестный блок расположен вне главной диагонали, симметрично по обе стороны от нее. Это значительно упрощает задачу.

Цель данного исследования – поиск аналитического решения для восстановления матриц корреляций с учетом их специфической структуры и сравнение свойств реконструированной и реальной матриц с использованием данных из Биобанка Великобритании.

МЕТОД

Алгоритм реконструкции матрицы

Большинство методов реконструкции матрицы основаны на максимизации детерминанта матрицы [12–15]. Основным требованием для использования этого подхода является предположение о корректности реконструированной корреляционной матрицы (в частности, симметричность и положительная полуопределенность ее заданных диагональных подматриц [12]) и многомерной нормальности данных [14]. Метод реконструкции корреляционной матрицы, основанный на максимизации ее детерминанта, имеет ряд полезных теоретических свойств [14]:

существование и уникальность решения: существует ровно одна реконструированная корреляционная матрица, использующая принцип максимума детерминанта;

максимальная энтропия: максимум детерминанта ведет к максимизации энтропии;

максимальное правдоподобие: максимум детерминанта, по сути, это оценка максимального правдоподобия корреляционной матрицы для неизвестной базовой многомерной нормальной модели;

центр области допустимых решений: максимум детерминанта является центром области допустимых решений, ограниченной положительной полуопределенностью, и имеет прямое аналитическое решение для некоторых блочно-структурированных матриц.

На рис. 1 схематически изображены блоки частично заданной матрицы корреляций между тремя неперекрывающимися подмножествами генетических вариантов: M_1, M_2 и M_3 в рамках одного гена и соответствующие наборы z-статистик, полученные при объединении в один анализ импутированных данных (M_1 и M_2 подмножества вариантов) и экзомных данных (M_2 и M_3 подмножества вариантов).

Для удобства в соответствии с рис. 1 мы представили объединенную корреляционную матрицу U между тремя подмножествами генетических вариантов: M_1, M_2 и M_3, внутри гена в виде матричных блоков 3 × 3 с неизвестным блоком U_{13} (и сопряженным с ним блоком U_{31}, $U_{13} = U_{31}^T$):

$$U = \begin{array}{c} \begin{matrix} M_1 & M_2 & M_3 \end{matrix} \\ \begin{bmatrix} U_{11} & U_{12} & \boxed{U_{13} - ?} \\ U_{12}^T & U_{22} & U_{23} \\ U_{13}^T - ? & U_{23}^T & U_{33} \end{bmatrix} \end{array} \quad (1)$$

Тогда подматрица U_{13} может быть реконструирована через соседние с ней подматрицы, выделенные пунктирной линией в формуле (1):

$$U_{13} = U_{12} U_{22}^{-1} U_{23}. \quad (2)$$

Этот вывод, полученный в ряде работ [12, 14–16], следует из свойств детерминанта блочной матрицы вида $\begin{pmatrix} A & C \\ C^T & B \end{pmatrix}$, в частности:

a) $\det\begin{pmatrix} A & C \\ C^T & B \end{pmatrix} = \det(A) * \det\left(B - CA^{-1}C^T\right)$;

b) детерминант матрицы будет максимальным, если занулить подматрицу C, т. е. свести исходную матрицу к блочно-диагональному виду. Стоит отметить, что максимизация детерминанта матрицы, которая эквивалентна максимизации произведения собственных значений при сохранении фиксированного среднего собственных значений, равного единице, стремится сделать все собственные

значения равными, улучшая число обусловленности матрицы.

Легко показать, что в терминах принятых обозначений (рис. 1) максимальное значение детерминанта объединенной матрицы U выражается как

$$\det(U) = \det(U_{22})$$
$$\det\left(U_{11} - U_{12}U_{22}^{-1}U_{12}^{T}\right)\ \det\left(U_{33} - U_{23}^{T}U_{22}^{-1}U_{23}\right) = \quad (3)$$
$$= \det(U_{22})\ \det\left(U_{11}|U_{22}\right)\ \det\left(U_{33}|U_{22}\right)$$

или в терминах детерминантов исходных матриц корреляций как

$$\det(U) = \frac{\det\left(U_{imp}\right)\det\left(U_{seq}\right)}{\det\left(U_{shared}\right)},$$

где U_{imp}, как показано на рис. 1, представляет исходную корреляционную матрицу для вариантов гена с импутированными генотипами, U_{seq} – для вариантов гена с секвенированными генотипами и U_{shared} – для общих (перекрывающихся) вариантов гена.

Заметим, что решение (2) относится к аппроксимации, полученной путем удаления столбцов/строк с неизмеренными корреляционными коэффициентами:

$$U = UU^{-1}U \approx U_{*k}U_{kk}^{-1}U_{*k}^{T}. \quad (4)$$

Здесь U_{kk} заданная обратимая подматрица, где индекс k означает набор вариантов с генотипами, измеренными двумя способами: импутированным и секвенированным, а индекс * означает все варианты гена. Тогда в терминах наших обозначений (рис. 1) из формулы (4) получаем формулу $U_{13} = U_{12}U_{22}^{-1}U_{23}$, идентичную формуле (2).

Важно отметить, что решение (2) совпадает с условно независимым решением (см. формулу (3)), а значит реконструированные корреляции не могут превышать реальные значения [14]. Безусловно, значительное занижение реальных генотипических корреляций может привести к инфляции статистик при анализе ассоциаций на генном уровне. Однако поскольку секвенированные варианты являются редкими, их корреляции с распространенными импутированными вариантами будут близки к нулю [17], и можно ожидать, что статистическая инфляция будет минимальной.

Сравнение реальных и реконструированных матриц

Свойства предложенного метода, реконструирующего неизвестные коэффициенты корреляции между импутированными и секвенированными вариантами без использования информации о генотипах, были изучены путем сравнения восстановленных матриц с реальными матрицами, построенными по индивидуальным генотипам.

Оценка эффективности

Для оценки качества реконструкции корреляционных матриц был выбран показатель: оценка функции потерь, традиционно рассчитываемая как среднеквадратичная ошибка между фактическими значениями U_{13} и реконструируемыми $U_{13_{\text{rec}}}$:

$$L_1 = \text{mean}\left(\left(U_{13} \quad U_{13_{rec}}\right)^2\right).$$

Кроме того, поскольку наиболее популярным методом анализа ассоциаций на генном уровне является метод SKAT [18], который при использовании корреляционной матрицы ограничивается только собственными значениями ее спектрального разложения, в качестве дополнительных показателей эффективности мы использовали коэффициент регрессии RS_1 и коэффициент детерминации RS_2 регрессионного уравнения, сопоставляющего собственные значения реконструированной матрицы и реальной. В отличие от L_1 коэффициенты RS_1 и RS_2 рассчитываются на полной матрице U. Ожидается, что оба коэффициента будут близки к единице, а оценка функции потерь – к нулю.

Зависимость показателей эффективности была исследована относительно доли неизвестных коэффициентов среди всех (заданных и неизвестных) коэффициентов корреляции между импутированными и секвенированными вариантами. Эту долю можно представить геометрически как отношение площадей, s (рис. 1):

$$d = \frac{s\left(U_{13}\right)}{s\left(U_{22}\right) + s\left(U_{12}\right) + s\left(U_{23}\right) + s\left(U_{13}\right)}, d \in [0,1].$$

МАТЕРИАЛ

Генотипы и матрицы генотипических корреляций

Для построения объединенных матриц корреляций между вариантами внутри генов были использованы секвенированные и импутированные генотипы из Биобанка Великобритании (UK BioBank).

Оригинальные секвенированные генотипы (Data-Field 23156) были представлены в формате vcf в сборке GRCh38/hg38. Контроль качества этих данных осуществлялся с помощью пакета BCFtools версии 1.16.1 и по алгоритму, описанному на сайте https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=914. Секвенированные данные были преобразованы в формат pgen с помощью пакета plink версии 2.00a3.3LM.

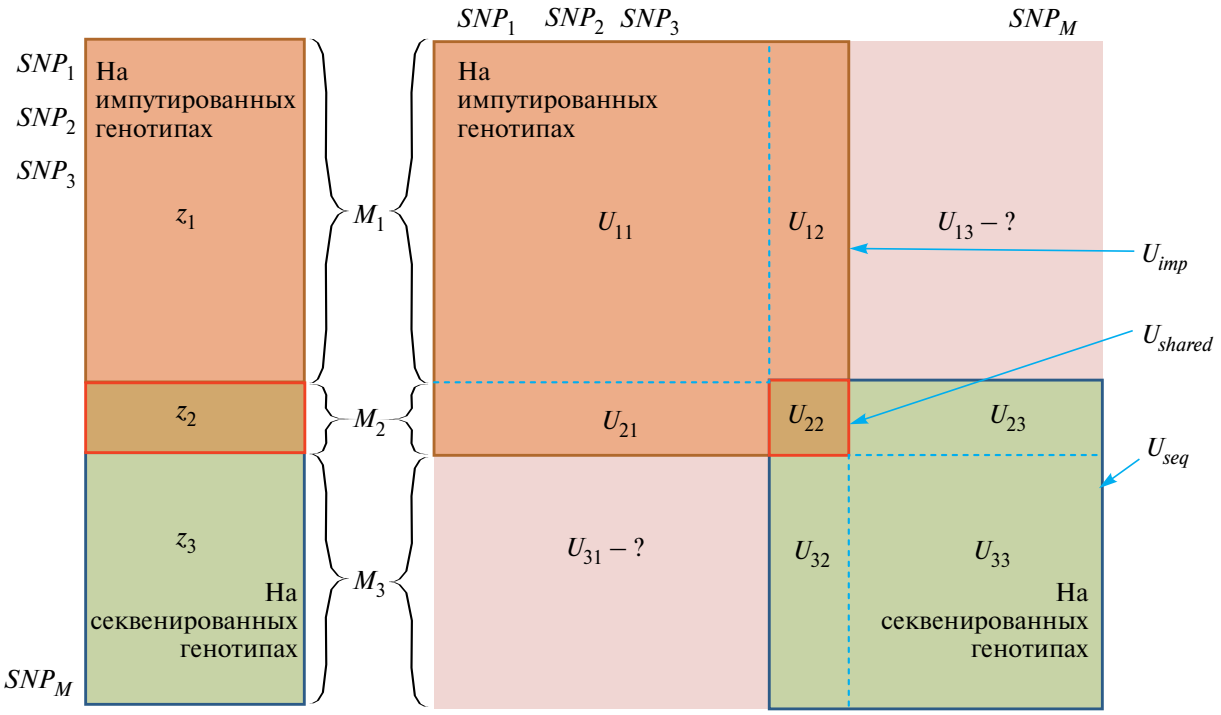

Рис. 1. Схематическое представление объединенной частично заданной матрицы корреляций между генетическими вариантами внутри гена. M_1 и M_3 обозначают неперекрывающиеся подмножества вариантов с импутированными и секвенированными генотипами соответственно, а M_2 – подмножество перекрывающихся вариантов. Суммарные статистики и матрицы генотипических корреляций для (M_1+M_2) вариантов посчитаны на импутированных данных, а для (M_2+M_3) вариантов – на секвенированных данных. Матрица корреляций между наборами M_1 и M_3 (блок U_{13}) неизвестна.

Оригинальные импутированные генотипы (Data-Field 22828), представленные в формате bgen в сборке GRCh37/hg19, были конвертированы в сборку GRCh38/hg38 с помощью пакета liftOver и в формат pgen с помощью пакета plink версии 2.00a. Дальнейшие манипуляции с данными проводились также с помощью пакета plink версии 2.00a.

Далее из обоих типов данных были выбраны генотипы только биаллельных маркеров для белых европейцев (Data-Field 21000), не связанных генетическим родством (Data-Field 22021). Из импутированных данных были исключены варианты, позиции которых совпадали с позициями секвенированных вариантов, чтобы отдать предпочтение секвенированным генотипам при объединении их с импутированными. Выборка была ограничена индивидами, у которых были оба типа генотипов, N ~ 200000 человек. Данные о генотипах обоих типов были объединены и сохранены в pgen-формате.

Для тестирования предложенного метода были выбраны гены на 21-й хромосоме. Для расчета корреляционных матриц была использована информация о позициях генов из данных Ensembl (https://www.ensembl.org/). Матрицы корреляций вычислялись с помощью пакета ldstore версии 2.0 (http://www.christianbenner.com/). Для каждого гена были вычислены три матрицы: по секвенированным, импутированным и объединенным генотипам. Для анализа отбирались гены с общим числом вариантов от 50 до 7000 в объединенной матрице и с обязательным наличием перекрывающихся вариантов. Кроме того, для каждого гена в корреляционную матрицу включали только те варианты, которые прошли установленные пороги по частоте минорного аллеля maf и показателю качества импутации info: для импутированных вариантов info > 0.8 и maf > 5×10^{-5}, а для секвенированных вариантов maf > 3.8×10^{-5}, что соответствовало количеству минорных аллелей, mac > 10. Всего в анализ были включены корреляционные матрицы для 124 генов. После фильтраций в объединенных матрицах корреляций число импутированных вариантов варьировалось от 2 до 1572, число секвенированных вариантов – от 6 до 352, а число перекрывающихся вариантов – от 1 до 96.

Фенотип и суммарные статистики

Для анализа реальных фенотипических данных были получены суммарные статистики для количественного признака, индекс массы тела (body mass index, BMI) (N = 187600). Чтобы привести распределение признака к нормальному, были

исключены из анализа 2247 человек со значениями признака, выходящими за пределы трех среднеквадратических отклонений от среднего значения, и было выполнено преобразование признака с помощью процедуры преобразования рангов (rank-transformation), после чего нормальность распределения признака была подтверждена критерием Колмогорова – Смирнова ($p = 1$).

Для получения суммарных статистик для BMI проводили одноточечный анализ ассоциаций для секвенированных и импутированных данных с помощью пакета fastGWA-GLMM, бета-версия 1.94.0 [19]. Анализ выполнялся с использованием опции – fastGWA-mlm и ограничивался редкими вариантами с частотой минорного аллеля (maf) в диапазоне от 5×10^{-5} до 0.01. Для исключения случайных эффектов, обусловленных родством между индивидами в выборке, была предварительно рассчитана матрица родства с использованием fastGWA-GLMM (опция – make-bK-sparse с параметрами по умолчанию) для всей выборки Биобанка Великобритании (N = 487000). В наш анализ в качестве ковариат мы включили пол, возраст и первые 10 главных генетических компонент, предоставленных Биобанком Великобритании.

Симуляционные суммарные статистики

Предложенный метод был также протестирован на симуляционных суммарных статистиках. Для каждого из выбранных генов 21-й хромосомы был смоделирован вектор z-статистик на реальной генотипической матрице корреляций U по определенному сценарию, учитывающему размер эффекта гена на признак, τ (1000 повторов для каждого сценария). Для симуляций суммарных статистик была рассмотрена модель, лежащая в основе метода SKAT, использующего суммарные статистики в качестве входных данных [20]

$$z \sim MN\left(0, U + \tau U^2\right).$$

Параметр τ прямо пропорционален наследуемости признака, обусловленной вариантами гена. Мы зафиксировали параметр τ как 0, 0.02, 0.05 и 0.1, где нулевой гипотезе, предполагавшей отсутствие ассоциаций между анализируемым геном и признаком, соответствует $\tau = 0$.

РЕЗУЛЬТАТЫ

Анализ показателей эффективности

В табл. 1 приведена статистика, описывающая структуру распределения выбранных показателей эффективности. Как видно, для большинства генов оценка функции потерь близка к нулю и в целом не превышает 0.11, а коэффициенты RS_1 и RS_2 близки к единице, что свидетельствует об очень хорошей сходимости собственных значений реконструированной корреляционной матрицы к собственным значениям реальной (рис. 2). При сравнении показателей эффективности между собой, как ожидалось, наблюдается сильная корреляция. Парные корреляции Пирсона составили: $cor(L_1, RS_1) = -0.752$, $cor(L_1, RS_2) = -0.816$ и $cor(RS_1, RS_2) = 0.739$.

Следует отметить, что мы не обнаружили явной зависимости показателей эффективности от доли неизвестных коэффициентов среди всех (заданных и неизвестных) коэффициентов корреляции между импутированными и секвенированными вариантами. У генов, для которых наблюдали “выбросы” (сильные отклонения от ожидаемых значений по всем трем показателям эффективности: L_1, RS_1 и RS_2), мы обнаружили сильное расхождение между секвенированными и импутированными генотипами для перекрывающихся вариантов, несмотря на фильтрацию импутированных генотипов по показателю info, что свидетельствует скорее о низком качестве импутации генотипов, чем о плохом качестве реконструкции.

Сравнение результатов анализа ассоциаций на генном уровне

После успешной апробации предложенного метода мы оценили качество реконструкции объединенной матрицы корреляций, выполнив анализ ассоциаций на генном уровне с использованием реконструированной и реальной матриц корреляций и сравнив результаты анализа – p-значения. Для анализа мы выбрали три популярных теста, SKAT (Sequence kernel association test) [18], BT (Burden test) и PCA (Principle component analysis) [21], использующих различные регрессионные модели и реализованных в R-пакете sumFREGAT [21, 22]. Все тесты выполнялись без взвешивания

Таблица 1. Показатели структуры распределения для оценок качества реконструкции корреляционных матриц

Статистика распределения	L_1	RS_1	RS_2
Минимум	5.00×10^{-8}	0.953	0.927
1-й квартиль	5.59×10^{-5}	0.999	1.000
Медиана	1.02×10^{-4}	1.000	1.000
Среднее	1.48×10^{-3}	0.999	0.998
3-й квартиль	2.32×10^{-4}	1.000	1.000
Максимум	1.06×10^{-1}	1.027	1.000

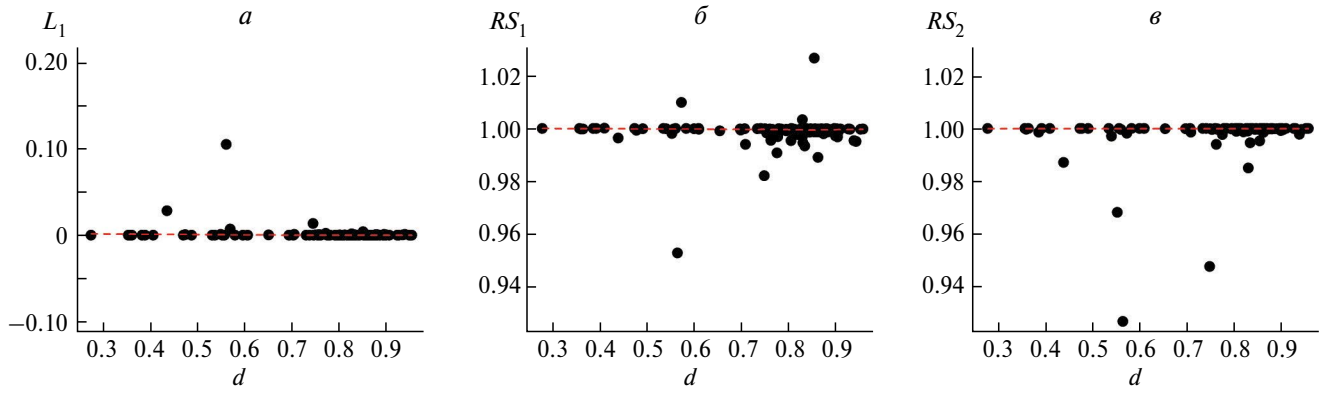

Рис. 2. Зависимости показателей эффективности реконструкции корреляционных матриц от доли неизмеренных коэффициентов среди всех (измеренных и неизмеренных) коэффициентов корреляции между импутированными и экзомными вариантами. По оси X откладывается доля неизмеренных коэффициентов среди всех коэффициентов корреляции между импутированными и экзомными вариантами, а по оси Y откладывается показатель эффективности.

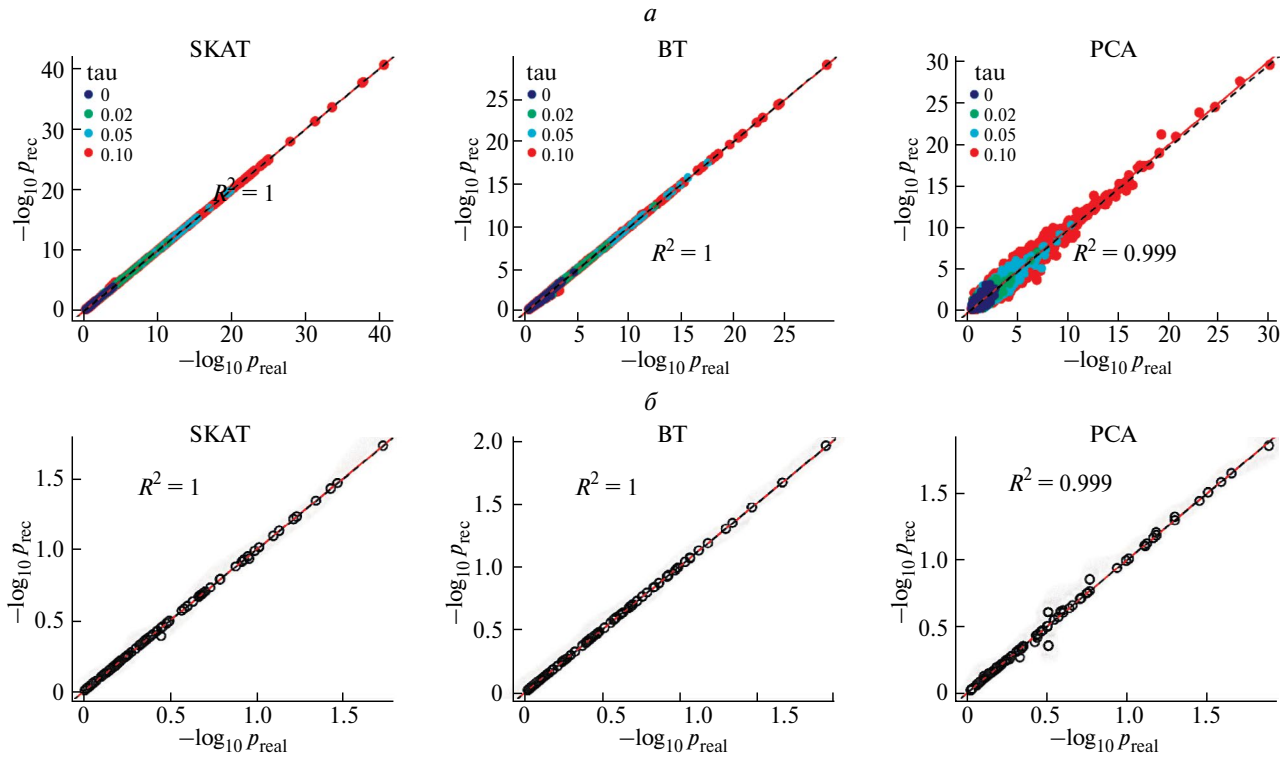

Рис. 3. Сопоставление результатов анализа ассоциаций на генном уровне, выполненного методами SKAT, BT и PCA на реконструированных и исходных матрицах корреляций: *а* – с использованием z-статистик, симулированных для различных сценариев относительно t; *b* – с использованием *z*-статистик, вычисленных на реальном признаке BMI.

суммарных статистик. Для PCA долю объясненной дисперсии брали 0.85.

По всем изученным генам 21-й хромосомы было получено полное соответствие p-значений с $R^2 = 1$ для методов SKAT и BT, использующих суммарные статистики как симулированные (рис. 3, а), так и вычисленные по реальному признаку BMI (рис. 3, б). Для метода PCA хорошее соответствие наблюдалось для всех генов на реальных данных ($R^2 = 0.999$) и для генов с маленькими *p*-значениями ($p < 1 \times 10^{-7}$) на симулированных данных. Для генов с большими *p*-значениями наблюдалось небольшое

рассеивание в виде облака без смещения регрессионной линии от ожидаемой ($R^2 = 0.999$).

ОБСУЖДЕНИЕ

Объединение импутированных и секвенированных данных для анализа ассоциаций на уровне генов может помочь обнаружить сигналы в генах, обусловленные совместным влиянием вариантов внутри генов. Если для генов исходные генотипы недоступны, корреляционные матрицы строятся на основе референсной выборки. В этом случае, как правило, корреляции между импутированными и секвенированными вариантами гена неизвестны. Для реконструкции этих корреляций мы предлагаем эффективный математический метод, основанный на максимизации детерминанта матрицы. Этот метод обладает рядом полезных свойств и имеет аналитическое решение для нашей задачи.

Ограничения и возможности предложенного метода были исследованы на реальных корреляционных матрицах, построенных на индивидуальных генотипах из Биобанка Великобритании. Очевидно, что одним из ограничений предложенного метода является невозможность его использования, если импутированные и секвенированные варианты гена не перекрываются. Кроме того, мы обнаружили, что еще одним ограничением метода служит сильное расхождение между импутированными и секвенированными генотипами перекрывающихся вариантов генов. Вероятно, такое расхождение обусловлено использованием в анализе редких вариантов с ошибочно импутированными генотипами. В дополнение следует отметить, что мы не выявили никаких ограничений, касающихся доли неизвестных корреляций в объединенной матрице.

Для апробации предложенного метода был проведен анализ ассоциаций на уровне генов, выполненный как на смоделированных суммарных статистиках, так и на суммарных статистиках, вычисленных на реальном признаке BMI из Биобанка Великобритании. Сравнение результатов анализа ассоциаций на генном уровне, выполненного популярными методами SKAT, BT и PCA на реконструированных и реальных матрицах, показало высокое качество реконструкции вне зависимости от доли перекрывающихся вариантов в гене и устойчивость метода к различным структурам генов.

Для перекрывающихся вариантов соответствие между корреляционными матрицами, построенными на импутированных данных, и корреляционными матрицами, построенными на секвенированных данных, может быть нарушено по ряду причин, например, использование референсных выборок разного размера или включение в анализ редких вариантов с генотипами низкого уровня импутации. Для решения этой проблемы можно применить процедуру регуляризации, которая для каждого гена “подгоняет” реконструированную корреляционную матрицу к соответствующим z-статистикам с помощью двух параметров регуляризации, настраиваемых персонально для импутированных и секвенированных данных. Регуляризация по типу Тихонова [23] использует спектральную фильтрацию собственных значений корреляционной матрицы. Подобная процедура нами уже была предложена в работе [21] при использовании референсных генотипических данных для анализа ассоциаций на уровне гена. Общая схема предлагаемой регуляризации для реконструированной матрицы корреляций, объединяющей импутированные и секвенированные генотипы, имеет вид:

$$E\left(zz^T\right) = \begin{bmatrix} \lambda_1 U_{11} & \sqrt{\lambda_1}\lambda_2 U_{12} & \sqrt{\lambda_1}\lambda_2 U_{13} \\ \sqrt{\lambda_2}\lambda_1 U_{12}^T & \lambda_2 U_{22} & \lambda_2 U_{23} \\ \sqrt{\lambda_2}\lambda_1 U_{13}^T & \lambda_2 U_{23}^T & \lambda_2 U_{33} \end{bmatrix} +$$

$$+ \begin{pmatrix} \left(1-\lambda_1\right) I_{M_1} & 0 & 0 \\ 0 & \left(1-\lambda_2\right) I_{M_2} & 0 \\ 0 & 0 & \left(1-\lambda_2\right) I_{M_3} \end{pmatrix},$$

здесь $U_{13} = U_{12}U_{22}^{-1}U_{23}\lambda_1$ и λ_2 – параметры регуляризации, вычисленные на импутированных и секвенированных данных соответственно, а I_{M_*} – единичные матрицы размерности M_*. Основываясь на предположении о многомерном нормальном распределении z-статистик, регуляризация корреляционных матриц обеспечит их устойчивость.

Таким образом, предложенный метод, который реконструирует неизвестные корреляции между импутированными и секвенированными генотипами вариантов в гене на основе корреляций между импутированными вариантами и корреляций между секвенированными вариантами без учета индивидуальных генотипов, может быть успешно применен для объединения разных типов генотипических данных с целью поиска сигналов, обусловленных совместным влиянием импутированных и секвенированных вариантов в генах.

Данное исследование проводилось с использованием ресурсов Биобанка Великобритании в рамках заявок № 18219 и № 59345.

Работа Г.Р. Свищёвой, И.В. Зоркольцевой, Н.М. Белоноговой и Е.Е. Елгаевой выполнена при финансовой поддержке гранта Российского научного фонда (РНФ) № 23-25-00209.

Настоящая статья не содержит каких-либо исследований с участием в качестве объектов людей.

Авторы заявляют, что у них нет конфликта интересов.

Вклад авторов. Г.Р. Свищёва: концептуализация, методология и апробация, исследование, визуализация, написание статьи; А.В. Кириченко: подготовка данных, исследование, написание статьи; Н.М. Белоногова: исследование, написание статьи; Е.Е. Елгаева: написание статьи; Я.А. Цепилов: концептуализация, написание статьи; И.В. Зоркольцева: подготовка данных, исследование, написание статьи; Т.И. Аксенович: концептуализация: научное редактирование и управление, написание статьи. Все соавторы обсуждали результаты и внесли вклад в подготовку окончательной версии рукописи.

СПИСОК ЛИТЕРАТУРЫ

1. *Eichler E.E., Flint J., Gibson G. at al.* Missing heritability and strategies for finding the underlying causes of complex disease // Nat. Rev. Genet. 2010. V. 11. № 6. P. 446–450. https://doi.org/10.1038/nrg2809
2. *Li B., Leal S.M.* Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data // The Am. J. Hum. Genet. 2008. V. 83. № 3. P. 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024
3. *Cirulli E.T.* The increasing importance of gene-based analyses // PloS Genetics. 2016. V. 12. № 4. https://doi.org/10.1371/journal.pgen.1005852
4. *Kang G., Jiang B., Cui Y.* Gene-based genomewide association analysis: A comparison study // Curr. Genomics. 2013. V. 14. № 4. P. 250–255. https://doi.org/10.2174/1389202913149990001
5. *Li Y., Willer C., Sanna S., Abecasis G.* Genotype imputation // Ann. Rev. Genomics and Hum. Genet. 2009. V. 10. P. 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242
6. *Uffelmann E., Huang Q.Q., Munung N.S. et al.* Genome-wide association studies // Nat. Rev. Methods Primers. 2021. V. 1. № 59. P. 1–21. https://doi.org/10.1038/s43586-021-00056-9
7. *Guo Y., Long J., He J. et al.* Exome sequencing generates high quality data in non-target regions // BMC Genomics. 2012. V. 13. № 1. P. 1–10. https://doi.org/10.1186/1471-2164-13-194
8. *Clark M.J., Chen R., Lam H.Y. et al.* Performance comparison of exome DNA sequencing technologies // Nat. Biotechnol. 2011. V. 29. № 10. P. 908–914. https://doi.org/10.1038/nbt.1975
9. *Stanley J.C., Wang M.D.* Restrictions on the possible values of r12, given r13 and r23 // Educational and Psychol. Measurement. 1969. V. 29. № 3. P. 579–581.
10. *Glass G.V., Collins J.R.* Geometric proof of the restriction on the possible values of rxy when rxz and ryz are fixed // Educational and Psychol. Measurement. 1970. V. 30. № 1. P. 37–39.
11. *Budden M., Hadavas P., Hoffman L., Pretz C.* Generating valid 4 × 4 correlation matrices // Applied Mathemat. E-Notes. 2007. V. 7. P. 53–59.
12. *Glunt W., Hayden T., Johnson C.R., Tarazaga P.* Positive definite completions and determinant maximization // Linear Algebra and its Applications. 1999. V. 288. P. 1–10. https://doi.org/10.1016/S0024-3795(98)10211-2
13. *Vandenberghe L., Boyd S., Wu S.-P.* Determinant maximization with linear matrix inequality constraints // SIAM J. Matrix Analysis and Applications. 1998. V. 19. № 2. P. 499–533. https://doi.org/10.1137/S0895479896303430
14. *Georgescu D.I., Higham N.J., Peters G.W.* Explicit solutions to correlation matrix completion problems, with an application to risk management and insurance // Royal Soc. Open Sci. 2018. V. 5. № 3. P. 172348.
15. *Grone R., Johnson C.R., Sá E.M., Wolkowicz H.* Positive definite completions of partial Hermitian matrices // Linear Algebra and its Applications. 1984. V. 58. P. 109–124.
16. *Popescu O., Rose C., Popescu D.C.* Maximizing the determinant for a special class of block-partitioned matrices // Mathem. Problems in Engineering. 2004. V. 2004. P. 49–61. https://doi.org/10.1155/S1024123X04307027
17. *Li B., Liu D.J., Leal S.M.* Identifying rare variants associated with complex traits via sequencing // Curr. Protocols in Hum. Genet. 2013. V. 78. № 1. P. 1–26. https://doi.org/10.1002/0471142905.hg0126s78
18. *Wu M.C., Lee S., Cai T. et al.* Rare-variant association testing for sequencing data with the sequence kernel association test // The Am. J. Hum. Genet. 2011. V. 89. № 1. P. 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029
19. *Jiang L., Zheng Z., Fang H., Yang J.* A generalized linear mixed model association tool for biobank-scale data // Nat. Genet. 2021. V. 53. № 11. P. 1616–1621. https://doi.org/10.1038/s41588-021-00954-4
20. *Svishcheva G.R.* A generalized model for combining dependent SNP-level summary statistics and its extensions to statistics of other levels // Scientific Reports. 2019. V. 9. № 1. P. 1–8. https://doi.org/10.1038/s41598-019-41827-5
21. *Svishcheva G.R., Belonogova N.M., Zorkoltseva I.V. et al.* Gene-based association tests using GWAS summary statistics // Bioinformatics. 2019. V. 35. № 19. P. 3701–3708. https://doi.org/10.1093/bioinformatics/btz172
22. *Belonogova N.M., Svishcheva G.R., Kirichenko A.V. et al.* sumSTAAR: A flexible framework for gene-based

association studies using GWAS summary statistics // PloS Comput. Biology. 2022. Т. 18. № 6. https://doi.org/10.1371/journal.pcbi.1010172

23. *Тихонов А.Н.* О решении некорректно поставленных задач и методе регуляризации // ДАН. 1963. Т. 151. № 3. С. 501–504.

Reconstruction Of a Matrix Of Genotypic Correlations Between Variants Within A Gene For Joint Analysis Of Imputed And Sequenced Data

G. R. Svishcheva[1, 2,*], A. V. Kirichenko[1], N. M. Belonogova[1], E. E. Elgaeva[1, 3], Ya. A. Tsepilov[1], I. V. Zorkoltseva[1], T. I. Axenovich[1]

[1] *Federal Research Centre, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia*
[2] *Vavilov Institute of General Genetics, Russian Academy of Sciences, 119991 Moscow, Russia*
[3] *Novosibirsk State University, 630090 Novosibirsk, Russia*
**e-mail: gulsvi@mail.ru*

When combining imputed and sequenced data in a single gene-based association analysis, the problem of reconstructing genetic correlation matrices arises. It is related to the fact that for a gene, we know the correlations between genotypes of all imputed variants and the correlations between genotypes of all sequenced variants, but we do not know the correlations between genotypes of variants, one of which is imputed and the other is sequenced. To recover these correlations, we propose an efficient method based on maximising the determinant of the matrix. This method has a number of useful properties and has an analytical solution for our task. Approbation of the proposed method was performed by comparing reconstructed and real correlation matrices constructed on individual genotypes from the UK biobank. Comparison of the results of gene-based association analysis performed by the SKAT, BT and PCA methods on reconstructed and real matrices, using modelled summary statistics and calculated summary statistics on real phenotypes, showed high quality of reconstruction and robustness of the method to different gene structures.

Keywords: imputed and sequenced genotypes, gene-based association analysis, genetic variants, summary statistics