Малов Сергей Васильевич,
Шевченко Андрей Константинович,
О'Брайен Стефан Джеймс

УДК 57.087.1, 519.24

ПОИСК ГЕНЕТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ*

Часть 2: Проблема интерпретации результатов множества тестов, компьютерные инструменты и приложения

Аннотация

В работе всесторонне изучается методология полногеномного поиска зависимости фенотипа с одним или несколькими генотипами. В данной части работы обсуждается проблема интерпретации результатов множества тестов, реализация статистических методов, описанных в первой части работы, на языке программирования **R** и особенности использования статистических методов в комплексных исследованиях заражаемости ВИЧ и развития СПИД. Также будут предложены некоторые соображения, касающиеся интерпретации результатов однотипных статистических тестов, полученных на базе независимых экспериментов.

Ключевые слова: полногеномный поиск закономерностей (ассоциаций), GWAS, проблема интерпретации результатов множества тестов, эпидемиология ВИЧ.

1. ВВЕДЕНИЕ

В первой части [2] работы рассматривались статистические тесты, пригодные для анализа генетических закономерностей (ассоциаций) по данным генетических исследований с клиническими данными различного типа. Общая проблема всех генетических исследований – интерпретация результатов множества статистических тестов. Отметим, что статистический эксперимент можно считать удачным, если удается найти зависимость исследуемых характеристик, то есть отвергнуть основную гипотезу о независимости. Еще в работе [43] отмечалось, что в среднем каждая двадцатая из верных гипотез ошибочно отвергается, если использовать стандартную 5 % верхнюю границу вероятности ошибки *I* рода при проведении множества тестов. В частности, наличие 12-ти статистически подтвержденных гипотез из 250 не дает никаких оснований считать, что хоть один из выводов верен. Существенная проблема, связанная с получением статистически значимых выводов, состоит в том, что результаты удачных экспериментов обычно замечают (публикуют), тогда как результаты неудачных экспериментов, в которых не удается статистически подтвердить то или иное предположение, остаются незамеченными. В связи с этим следует помнить, что

* Работа поддержана мегагрантом правительства Российской Федерации №~11.G34.31.0068.

доля статистически подтвержденных неверных выводов вполне может быть существенно больше ожидаемой величины. В задаче полногеномного анализа закономерностей данная проблема имеет особую актуальность, ибо число одновременно проводимых статистических тестов велико. Полногеномный поиск генетических закономерностей ассоциируется с поиском иголки в стоге сена. Для надежного выявления закономерности требуется тем больше статистических данных, чем больше генетических маркеров исследуется. Основные проблемы, связанные с интерпретацией результатов множества статистических тестов, и методы их решения описаны в разделе 2.

Ввиду огромного объема обрабатываемой информации современный поиск генетических закономерностей невозможен без использования компьютерных технологий. Поскольку задача непосредственно связана с проведением статистических исследований, выглядит естественным строить инструменты на базе статистических пакетов, однако специфика решаемых задач мотивирует создавать специализированные программы анализа генетических закономерностей. Среди прочих следует отметить PLINK software [37], PRESTO [8] и PERMORY [35]. Также отметим пакеты *GWAToolbox*, *SNPassoc*, *GenABEL* и *postgwas*, созданные на базе программного обеспечения **R** [38]. В данной работе мы обсудим возможность использования широкого набора инструментов языка программирования **R** для создания разнообразных инструментов поиска генетических закономерностей. Краткое описание **R** и его возможностей для задачи поиска генетических закономерностей дано в разделе 3.

Статистические методы широко используются для исследования данных эпидемиологических исследований, однако данные, получаемые в результате эпидемиологических исследований, далеко не всегда безоговорочно подходят для анализа известными статистическими методами. В разделе 4 будут рассмотрены методы анализа эпидемиологических данных по заражаемости ВИЧ и развитию СПИД, применимость различных статистических методов анализа данных типа времени жизни для анализа смешанных популяций, включающих серопревалентных индивидов (зараженных ВИЧ инфекцией к началу исследования) и сероконвертеров (заразившихся ВИЧ инфекцией в течение исследования).

Подходы к интерпретации результатов множества связанных и независимых тестов, полученных в независимых исследованиях, обсуждаются в разделе 5.

2. ПРОБЛЕМА ИНТЕРПРЕТАЦИИ РЕЗУЛЬТАТОВ МНОЖЕСТВА ТЕСТОВ

В последнее время статистическое обоснование выявленного феномена является необходимым условием каждого биомедицинского исследования. Обычно все ограничивается *P*-значением, вычисленным по статистике того или иного критерия. Результат считается статистически доказанным, если *P*-значение не превышает стандартного уровня значимости 0.05, тем самым вероятность ошибки I рода для данного утверждения не превышает 5 %. При этом следует принимать во внимание, что часть утверждений, полученных таким образом, могут будут ошибочными.

При проведении полногеномного тестирования на зависимость фенотипа от генотипов данная проблема обретает особую актуальность. Сигналом будем называть наличие зависимости фенотипа от выбранного генотипа. Проводится множество тестов на наличие сигнала для каждого из выбранных генотипов, вычисляются статистики критерия и *P*-значение. В современных полногеномных исследованиях число рассматриваемых генотипов достигает нескольких миллионов. Пусть $p_i, i = 1, ..., N$ – полученный набор *P*-значений, N – число генотипов. Известно, что в случае независимости фенотипа от i-го генотипа распределение p_i является равномерным на интервале [0, 1]. В случае независимых тестов и при отсутствии сигналов вероятность хотя бы одной ошибки I рода[1].

$$\mathrm{FWER} = \mathbb{P}\left(\cup_{i=1}^{N} \{p_i > x\}\right) = 1 - (1 - x)^N .$$

Таким образом, чтобы гарантировать отсутствие ошибок I рода в N тестах с вероятностью $(1-\alpha)$, потребуется установить порог для каждого теста

$$\alpha_{DS} = 1-(1-\alpha)^{1/N}.$$

Этот метод называется поправкой Данна–Шидака. В случае зависимых тестов можно использовать неравенство Буля

$$\mathbb{P}\left(\cup_{i=1}^{N}\{p_i > x\}\right) \le \sum_{i=1}^{N}\mathbb{P}(p_i > x) = n(1-x).$$

Чтобы гарантировать отсутствие ошибок I рода в N тестах с вероятностью $(1-\alpha)$, потребуется установить порог для каждого теста

$$\alpha_B = \alpha/N.$$

Полученный метод носит название поправки Бонферрони. Отметим, что

$$1-(\alpha+\alpha^2/2)/N \le (1-\alpha)^{1/N} \le 1-\alpha/N,$$

а следовательно, поправка Бонферрони практически совпадает с поправкой Данна–Шидака при небольших значениях α и больших N. Хольм [23] предложил усовершенствованную процедуру выбора, позволяющую контролировать FWER. Для этого все P-значения располагают в порядке возрастания $p_{(1)} \le ... \le p_{(N)}$. Зависимость фенотипа с соответствующим генотипом признается значимой, если $p_i < \alpha/(N-r_i+1)$, где r_i – ранг элемента p_i в вариационном ряду $p_{(1)}, ..., p_{(N)}$. Тем не менее, при большом количестве тестов только очень сильные связи (скажем, полный иммунитет к болезни при наличии определенного генотипа) могут быть выявлены по выборкам реального размера с использованием метода Хольма.

Какое-либо существенное улучшение данного метода возможно только в предположении сильной зависимости результатов тестов. В какой-то мере это относится к задаче генетических исследований, поскольку многие генотипы оказываются сцепленными. Наиболее эффективный метод, позволяющий учитывать зависимость результатов тестов, – критерии перестановок. Идея построения критерия перестановок состоит в том, что рассматриваются всевозможные соответствия (перестановки) имеющихся значений наблюдаемой переменной имеющимся ковариатам, в каждом случае вычисляется P-значение и строится вариационный ряд. Если P-значение, полученное при существующем соответствии наблюдаемой переменной, и ковариат, имеют ранг меньше $K*\alpha$, где K – число перестановок, то основная гипотеза независимости отвергается (см. [1]).

Отметим, что при сколько-нибудь большом наборе наблюдений перебрать все перестановки не представляется возможным, но описанный метод будет работать и в том случае, если сопоставлять генотипы фенотипам случайным образом. Применительно к генотипическим данным можно использовать следующий алгоритм:

1) выбрать достаточно большое значение R – размер рандомизованной выборки;

2) случайным образом сопоставить имеющиеся фенотипы имеющимся генотипам R раз и в каждом случае вычислить P-значения;

3) для i-го генотипа построить вариационный ряд полученных P-значений $p_{(i,1)} \le ... \le p_{(i,R)}$, найти $p_{i\alpha} = p_{(i,[R\alpha])}$;

4) построить вариационный ряд $p_{(1)} \le ... \le p_{(N)}$ истинных P-значений и признать значимыми отклонения от основной гипотезы для генотипов с индексами $j_1, ..., j_s$: $p_{j_1} \le ... \le p_{j_s} \le ... \le p_{j_{s+1}} \le p_j$ для любого $j \ne j_k$, $k = 1, ..., s+1$, и $p_{j_1} < p_{(1\alpha)}, ..., p_{j_s} < p_{(s\alpha)}$, $p_{j_{s+1}} \ge p_{(s+1\alpha)}$, где $p_{(1\alpha)} \le ... \le p_{(N\alpha)}$ – вариационный ряд, построенный на базе $p_{1\alpha}, ..., p_{N\alpha}$. Следует отметить, что данный метод слишком затратен с точки зрения вычислений, поэтому при изучении большого числа генотипов он неприменим.

---

[1] Family Wise Error Rate (*англ.*)

Пусть P_I – вероятность (неизвестная) хотя бы одной ошибки I рода (FWER). Попытки улучшить границы Бонферрони–Хольма за счет использования сцепленности генотипов предпринимались многими авторами, однако большинство методов имеет эвристический характер. Основная идея этих методов состоит в том, что тесты по сцепленным генотипам группируются в блоки. Отталкиваясь от формы Данна–Шидака представления P_I, можно поставить вопрос о том, сколько независимых блоков можно сопоставить исходному набору тестов

$$N_{\text{eff.}} = \log(1-P_I)/\log(1-\alpha).$$

Хорошая оценка для $N_{\text{eff.}}$ позволит построить эффективный критерий. Состоятельную оценку $N_{\text{eff.}}$ дает критерий рандомизации, но по описанным ранее причинам он неприменим для больших наборов наблюдений. В [9] предложена оценка

$$M_{\text{eff.}} = 1 + (N-1)(1-V_\lambda/N),$$

где $V_\lambda = (N-1)^{-1}\sum_{i=1}^{N}(1-\lambda_i)^2$, $\lambda_1, ..., \lambda_N$ – набор собственных чисел, построенных по матрице корреляций генотипов (неравновесий по сцеплению LDE[1]). Аналогичный подход использовался в [31] для бинарных (аллельных) генотипов. В [30] авторы предлагают использовать

$$M_{\text{eff.}} = 1 + N - N^{-1}\sum_{i=1}^{N}\lambda_i^2.$$

В [26] использовался иной подход к определению $M_{\text{eff.}}$:

$$M_{\text{eff.}} = \sum_{i=1}^{N} f(\lambda_i),$$

где $f(x) = \mathbb{1}_{\{x\geq 1\}} + (x-[x])$. В [21] используется

$$M_{\text{eff.}} = \min\{x : \sum_{i=0}^{x}\lambda_{(N-x)} \Big/ \sum_{i=0}^{x}\lambda_{(i)} > 0.995\},$$

где $\lambda_{(1)} \leq ... \leq \lambda_{(N)}$ – упорядоченный набор собственных чисел $\lambda_{(1)}, ..., \lambda_{(N)}$. Отметим, что все эти методы строятся на базе корреляций между генотипами, а не между статистиками критериев, а полученные значения $M_{\text{eff.}}$ носят чисто эвристический характер.

Если говорить о таблицах сопряженности 2×2, то статистика хи-квадрат критерия по сути представляет собой квадрат асимптотически нормальной статистики. Более того, совместное распределение этих статистик асимптотически нормально и при известных ковариационных характеристиках задача сводится к изучению максимума модуля компонент соответствующего многомерного нормального распределения. Задача о распределении максимума модуля компонент нормального распределения с произвольной матрицей ковариации на текущий момент не решена.

В [30] выдвигается гипотеза, что если вектор $\xi_1, ..., \xi_n$ имеет многомерное нормальное распределение $N(0,\Sigma)$, компоненты которого имеют стандартное нормальное распределение, то

$$\mathbb{P}(|\xi_1| < x \,|\, |\xi_2| < x, ..., |\xi_n| < x) \geq \mathbb{P}(|\xi_1| < x \,|\, |\xi_2| < x) \qquad (1)$$

при произвольном выборе матрицы ковариации Σ. Данная гипотеза выглядит вполне правдоподобно, но ее доказательство на текущий момент не построено. Более того, при определенных условиях данное утверждение по сути совпадает с недоказанным вариантом так называемой гауссовской корреляционной гипотезы. При выполнении (1) вычисления можно свести к двумерным распределениям. Пусть ξ_1, ξ_2 – случайный вектор, имеющий центрированное нормальное распределение, $D\xi_1 = D\xi_2 = 1$ и $\text{cov}(\xi_1, \xi_2) = \rho$. Тогда условное распределение ξ_1 при условии ξ_2 является нормальным с

[1] Linkage disequilibrium (*англ.*)

$$\mathbb{E}(\xi_1 \mid \xi_2) = \rho\xi_2 \text{ и } \mathbf{Var}(\xi_1 \mid \xi_2) = 1 - \rho^2.$$

Таким образом,

$$\mathbb{P}(|\xi_1| \le x \mid \xi_2) = \Phi\left(\frac{x - \rho\xi_2}{\sqrt{1-\rho^2}}\right) - \Phi\left(-\frac{x - \rho\xi_2}{\sqrt{1-\rho^2}}\right) = 1 - 2\Phi\left(\frac{\rho\xi_2 - x_2}{\sqrt{1-\rho^2}}\right),$$

где Φ – функция стандартного нормального распределения и

$$\mathbb{P}(|\xi_1| \le x \mid \xi_2 \le x) = \frac{1}{1-\alpha}\sqrt{\frac{2}{\pi}}\int_{-x}^{x}\Phi\left(-\frac{\rho\, t - x}{\sqrt{1-\rho^2}}\right)\exp(-t^2/2)dt\,.$$

С учетом данного соотношения и (1) в [30] была получена оценка для $N_{\text{eff}.}$

$$M_{\text{eff}.} = 1 + \sum_{j=2}^{N}\kappa_j\,,$$

где

$$\kappa_j = \frac{\log\left(1 - \frac{1}{1-\alpha}\sqrt{\frac{2}{\pi}}\int_{-x_\alpha}^{x_\alpha} e^{-x^2/2}\Phi\left(\frac{r_j x - x_\alpha}{\sqrt{1-r_j^2}}\right)dx\right)}{\log(1-\alpha)};$$

$r_j = \max_{1\le k\le j-1}|\rho_{kj}|$, $\rho_{kj} = \operatorname{cov}(\xi_k, \xi_j)$ – коэффициенты корреляции; $x_\alpha = \Phi^{-1}(1-\alpha)$. Для ускорения вычислений интеграла была применена формула $k_j \approx \sqrt{1 - r_j^{-1.31\log_{10}\alpha}}$.

Отметим также, что существует ряд методов непосредственного разбиения всего множества генотипов на блоки, носящих чисто эвристический характер.

Несмотря на все попытки оптимизировать метод идентификации сигналов, его практическая ценность остается достаточно низкой, и много значимых сигналов теряются. Обозначим N_{01} и N_{11} – числа корректно и ошибочно идентифицированных сигналов; $R = N_{01} - N_{11}$ – общее число идентифицированных сигналов. Отметим, что FWER = $\mathbb{P}(N_{01} > 0)$. Рассмотрим также характеристику FDR = $\mathbb{E}(N_{10}/R)$ – средний процент ошибочно идентифицированных сигналов при положительном числе идентифицированных сигналов.[1] В случае отсутствия идентифицированных сигналов FDR можно считать равным 0.[2] Идея использования данной характеристики была предложена в работе [4]. Пусть α – малое число; $p_{(1)}, \ldots, p_{(N)}$ – упорядоченные по возрастанию P-значения. Если выбрать все SNP, которым соответствуют значения $p_{(i)} \le \alpha\, i/N$, то в предположении независимости тестов значение FDR не будет превышать α. Выбор SNP, для которых $p_{(i)} \le \alpha\left(i\sum_{j=1}^{i}1/j\right)\Big/N$, позволяет контролировать FDR на уровне α в общем случае не обязательно независимых тестов. Отметим, что контроль FDR позволяет выбрать больше сигналов в случае их достаточно большого числа, но в случае малого числа сигналов такая постановка не дает больших преимуществ, граница для наименьшего P-значения по-прежнему остается в точности такой же, как и в случае контроля FWER.

В заключение данного раздела рассмотрим статистику Тьюки [43]

$$\text{HC}_\alpha = \sqrt{N}\,\frac{R_\alpha/N - \alpha}{\sqrt{\alpha(1-\alpha)}},$$

где R_α – число отвержений основной гипотезы на уровне α, применяемую в основном для

[1] False Discovery Rate (*англ.*).

[2] Корректнее использовать определения FDR, учитывающие случайное распределение числа идентифицированных сигналов. В этом случае возникает несколько близких по значению характеристик:
FDR = $\mathbb{E}(N_{10}/R \mid R > 0)\mathbb{P}(R > 0)$, pFDR = $\mathbb{E}(N_{10}/R \mid R > 0)$, mFDR = $\mathbb{E}(N_{10})/\mathbb{E}(R)$.

распознавания сигнала. Отметим, что при полном отсутствии сигналов и в случае независимости результатов тестов статистика HC_α асимптотически имеет стандартное нормальное распределение. В работе [11] изучается другая статистика $HC^* = \max_{0<\alpha<\alpha_0} HC_\alpha$, которая уже не зависит от α, но зависит от границы $\alpha_0 < 1$. Обобщение статистик HC и HC^* для зависимых тестов предлагается в работе [22]. Методы использования статистик такого типа для идентификации сигнала разработаны в работе [12].

3. СТАТИСТИЧЕСКИЙ ПАКЕТ R

R является свободной программной средой с открытым исходным кодом, ориентированной в первую очередь на обработку статистических данных [38]. **R** возник как свободный аналог одного из наиболее мощных статистических пакетов S-plus, однако на текущий момент возможности **R** существенно выше. В основе **R** лежит объектно-ориентированный язык программирования высокого уровня. Существует центральная система хранения CRAN (Comprehensive R Archive Network, http://cran.r-project.org). **R** состоит из стабильной базы и набора пакетов, ориентированных на решение конкретных задач. При инсталляции базовой части устанавливается только небольшой набор пакетов, остальные пакеты устанавливаются по требованию. При запуске **R** появляется консоль, команды можно вводить непосредственно в консоль или запускать их из текстового файла. С основами программирования в **R** можно познакомиться с использованием учебника «An introduction to R», включенного в базовую версию пакета в формате PDF, а также многочисленных пособий, часть из которых находится в открытом доступе. Здесь мы остановимся на описании методов обработки категориальных данных, данных типа времени жизни и коротких временных рядов, детально разбиравшихся в [2].

3.1. АНАЛИЗ КАТЕГОРИАЛЬНЫХ ДАННЫХ

Классические методы анализа таблиц сопряженности реализованы в базовой версии программы. Если все значения в таблице сопряженности достаточно велики, то можно использовать стандартный критерий хи-квадрат **`chisq.test()`**. Для таблиц сопряженности 2×2 с небольшими значениями в одной или нескольких ячейках реализован точный критерий Фишера **`fisher.test()`**. Входными данными для этих тестов могут быть либо два вектора наблюдаемых переменных, либо таблица сопряженности (матрица). При наличии малых значений в таблице сопряженности большего размера, чем 2×2, следует использовать опцию **`simulate.p.value=TRUE`**. По сути, используется критерий случайных перестановок на базе хи-квадрат. Параметр **`b`** позволяет выбирать размер генерируемой выборки. Отметим, что при использовании данной опции *P*-значение ограничено величиной $1/b$.

Альтернативно, обработку категориальных данных можно вести с использованием обобщенных линейных моделей. Функция **`glm()`** входит в базовый пакет. Предварительно потребуется задать формулу, скажем,

```
> reg<-formula(resp~gen)
```

где **`resp`** – наблюдаемая переменная, а **`gen`** – переменная типа «фактор», характеризующая генотип. Использование обобщенных линейных моделей позволяет включать в модель также факторы риска. В зависимости от типа наблюдаемой переменной, следует выбрать параметр **`family`**. Для бинарных откликов обычно выбирают **`family=binomial`**, в этом случае используется модель логистической регрессии. Также можно использовать пуассоновскую модель **`family=poisson`**, где в качестве наблюдаемой переменной будут значения из таблицы сопряженности. Выбор по умолчанию **`family=gaussian`** имеет смысл использовать в предположении, что наблюдаемая переменная имеет нормальное распределение.

3.2. АНАЛИЗ ДАННЫХ ТИПА ВРЕМЕНИ ЖИЗНИ

Методов обработки данных типа времени жизни в базовом пакете нет. Наиболее известные инструменты обработки данных типа времени жизни реализованы в пакете *survival*. Переменная типа времени жизни предварительно преобразуются в объект типа *Surv*. Команда **`Surv()`**, создающая объект класса *Surv*, имеет ряд аргументов: ***`time`***, ***`time2`***, ***`event`***, ***`type`*** и ***`origin`***. При наличии цензурированных справа данных для формирования данного объекта потребуется определить переменную ***`time`*** (время отказа или цензурирования) и индикатор ***`event:`*** ***`event`*`=1` в случае отказа и ***`event`*`=0` в случае цензурирования. Для интервально цензурированных данных требуется определить две временны́х переменных – левую и правую границы интервала. При выборе ***`type`*`=interval` переменная ***`event`*** помимо заявленных ранее значений может принимать значения ***`event`*`=2` (левое цензурирование) и ***`event`*`=3` (интервал). Значение ***`time2`*** используется только в случае ***`event`*`=3`. Альтернативно, если выбрать ***`type`*`=interval2`, то ***`time`*** и ***`time2`*** всегда воспринимаются как левая и правая границы интервала соответственно, значение ***`time2`*`=NA` интерпретируется как бесконечность (цензурирование справа). Для дальнейшего анализа генетических закономерностей составим формулу

```
> srv.f<-formula(srv~gen),
```

где **`srv`** – объект класса *Surv*, построенный по имеющимся данным типа времени жизни, **`gen`** – переменная, характеризующая генотип. Оценки Каплана–Мейера для наблюдений, группированных по генотипам, и много другой полезной информации дает **`survfit()`**. Функция **`survreg()`** позволяет получить результаты с использованием параметрических моделей. В частности, по умолчанию **`survreg(srv.f)`** использует семейство распределений Вейбулла. Чтобы поменять параметрическое семейство распределений Вейбулла на экспоненциальное, следует установить значение параметра **`dist="exponential"`**. Результат исполнения функции **`survreg.object`** может быть использован для получения оценок коэффициентов модели и их матрицы ковариаций, а также для дальнейшего проведения тестов проверки значимости влияния генотипа на фенотип. Модель Кокса реализована с помощью функции **`coxph()`**. Результат выполнения функции **`coxph.object`** содержит в себе информацию об оценках параметров модели и тестах проверки значимости влияния генотипа на фенотип.

Непараметрические методы работы с интервально-цензурированными данными реализованы в пакете ***`interval`***. Функция **`icfit()`** производит непараметрический анализ интервально-цензурированных данных. В качестве входных данных лучше использовать формулу, в левой части которой находится объект класса *Surv* с опцией **`type=interval2`**. Такой подход позволяет получать непараметрические оценки одновременно для нескольких групп, определяемых значением ковариаты. Альтернативно можно вводить правые и левые границы интервалов непосредственно. Результат выполнения функции **`icfit()`** (объект класса *icfit*) содержит необходимую информацию для построения непараметрических оценок с учетом неоднозначности. В частности, элемент **`$intmap`** объекта класса *icfit* содержит интервалы, в которых сконцентрированы вероятности, оценки которых содержатся в **`$pf`**. Для разделения на группы используется элемент **`$strata`**. Для получения графика функций отказа можно использовать непосредственно **`plot(icfit())`**. Выбор опции **`conf.int=TRUE`** позволяет получать доверительные интервалы с использованием бутстреп-метода, однако обычно это существенно повышает время выполнения команды.

Для проверки однородности групп с различными значениями ковариаты используется функция **`ictest()`** пакета *interval*. По сути, реализована идея, разработанная в [16], для пяти типов семипараметрических моделей **`scores=c("logrank1","logrank2", "wmw","normal","general")`**. Первые два значения параметра соответствуют моделям, рассмотренным в [40] и [18] соответственно, тогда как последние три – моделям из [16].

Детальное описание методов, использовавшихся при построении пакета ***interval*** можно найти в [17].

3.3. АНАЛИЗ КОРОТКИХ ВРЕМЕННЫХ РЯДОВ

Инструменты анализа с использованием обобщенных уравнений реализованы в пакетах *geepack* и *gee*. Предварительно требуется привести данные в формат **data.frame**, включающий **id** – номер индивида, **resp** – наблюдаемую переменную, **gen** – переменную типа «фактор», определяемую с использованием генотипа, **time** – время измерения или временную точку (соответствующую переменную назовем **d**) и определить модель

```
> gee.f<-formula(resp~gen*time).
```

С помощью функции **geeglm(gee.f,data=d,id=id,...)** пакета *geepack* проводится вычисление оценок с помощью обобщенных уравнений и их основных характеристик. При необходимости следует установить семейство распределений и функцию связи **family** (по умолчанию предполагается, что наблюдаемая переменная имеет нормальное распределение с тождественной функцией связи), а также корреляционную структуру **corstr**. Альтернативно можно использовать функцию **gee()** пакета *gee*.

Для анализа обобщенных смешанных моделей наиболее часто используется пакет *lme4*. Для анализа эффекта генотипа введем смешанные модели с простым эффектом индивида (с учетом эффекта генотипа и без него соответственно):

```
> gmm.f<-formula(resp~gen*time+(1|id));
> gmm.f0<-formula(resp~time+(1|id)).
```

Для проверки значимости эффекта генотипа запускается команда

```
> gs<-anova(lmer(gmm.f,d),lmer(gmm.f0,d))
```

и из объекта **gs** извлекается требуемое *P*-значение.

4. ПРИМЕНЕНИЕ В ИССЛЕДОВАНИИ ВИЧ ИНФЕКЦИИ И РАЗВИТИЯ СПИД

ВИЧ инфекция – это тяжелое хроническое заболевание, вызываемое вирусом иммунодефицита человека (ВИЧ), терминальной стадией которой является синдром приобретенного иммунодефицита (СПИД). Открытие ВИЧ и выдвижение предположения о его причастности к СПИД произошло в 1983 году одновременно в двух независимых лабораториях [3, 20]. Несмотря на серьезные усилия мирового сообщества, направленные на борьбу с эпидемией ВИЧ/СПИД, число ВИЧ-инфицированных в мире продолжает расти. В этой связи необходимо все более глубокое изучение как характеристик вируса, так и индивидуальных особенностей человека, влияющих на эффективность передачи, установления и развития ВИЧ инфекции. С момента инфицирования индивида его зараженные клетки становятся фабриками по производству новых вирусных частиц. В первые недели после появления в организме ВИЧ инфекции зараженные клетки могут производить несколько миллиардов вирусных частиц ежедневно, что приводит к увеличению концентрации циркулирующего в крови вируса, называемой вирусной нагрузкой (VL), до нескольких миллионов вирусных частиц в 1 мл крови (острая стадия). ВИЧ инфицирует Т-лимфоциты, несущие на своей поверхности рецепторы CD4 и CCR5, поскольку ВИЧ использует именно эти клеточные поверхностные белки как необходимые рецепторы для проникновения внутрь клетки. После острой стадии ВИЧ инфекции вирусная нагрузка снижается и стабилизируется у различных индивидов на различных уровнях (от сотен вирусных частиц в 1 мл крови до неопределимых количеств) на длительный период (хроническая стадия). У здорового человека, в 1 мкл крови содержится около 1200 CD4+ клеток. Однако на протяжении нескольких лет хронической стадии ВИЧ-инфекции у большинства пациентов, не проходящих лечение (ан-

тиретровирусную терапию – АРТ), концентрация CD4+ клеток постепенно убывает, падая ниже 200 кл./мкл, что является одним из критериев для констатации СПИД. CD4+ клетки необходимы для нормальной работы иммунной системы, и уменьшение их числа приводит к коллапсу иммунитета у пациентов, имеющих СПИД, делая их восприимчивыми к бактериальным, грибковым и вирусным оппортунистическим инфекциям, а также склонными к онкологическим и другим неинфекционным заболеваниям, которые у здорового индивида элиминируются иммунной системой. Эти болезни становятся фатальными для большинства не лечащихся пациентов со СПИД, что приводило к 95 % смертельных исходов до появления антиретровирусной терапии, впервые примененной в 1996.

Восприимчивость к инфекции, длительность периода до наступления СПИД у ВИЧ инфицированных, условия определения, последствия и осложнения СПИД и ответ (реакция) на применение АРТ различны у разных индивидов, что наводит на мысль о возможном наличии индивидуальных генетических детерминант в человеческом геноме, определяющих эти различия. Для поиска подобных детерминант, называемых ARG (AIDS Restriction Genes [32, 33], используются статистические подходы и методы полногеномного поиска генетических закономерностей на базе результатов когортных исследований.

Проведение комплексного исследования, направленного на изучение зависимости ВИЧ инфекции и динамики развития СПИД и наличия тех или иных генетических вариантов (SNP), требует много ресурсов и времени. В стандартной европейской популяции годовая заражаемость ВИЧ инфекцией обычно не превышает 1 %. По этой причине когорты для исследования заражаемости обычно формируются из представителей «групп риска», годовая заражаемость в которых может превышать 10 %. В ходе исследования, проходящего в течение определенного времени, рекрутированные участники регулярно проходят тестирование на наличие ВИЧ инфекции. С момента выявления у индивида ВИЧ инфекции можно изучать динамику развития СПИД. Как правило, часть индивидов оказываются инфицированными еще до начала исследования. Их называют серопревалентными (их число отражает уровень инфицированности в исследуемой популяции). Участников, инфицированных ВИЧ в ходе исследования, называют сероконвертерами. Остальные остаются серонегативными до конца исследования.

Важной составляющей при проведении статистических исследований является группировка индивидов по генотипу. Естественно классифицировать индивидов в три группы по генотипу: доминантная гомозигота, гетерозигота и рецессивная гомозигота. Для уточнения и более детальной интерпретации результатов имеет смысл провести дополнительные тесты с бинарной группировкой: доминантная гомозигота и гетерозигота против рецессивной гомозиготы, доминантная гомозигота против гетерозиготы и рецессивной гомозиготы, а также аллельный тест. В последнем случае каждому индивиду сопоставляются два одинаковых значения наблюдаемой переменной с соответствующими значениями аллелей, и по значениям аллелей производится группировка.

Основываясь только на результатах анализов в начальный момент времени (при формировании когорты), можно делать определенные выводы о SNP вариантах (аллельных вариантах или аллелях генов), способствующих или препятствующих заражению ВИЧ, путем проведения категориальных тестов. Считаем, что наблюдаемая переменная принимает значение 1 для всех серопревалентов и 0 – для остальных пациентов. Для проведения исследования заражаемости ВИЧ инфекцией можно использовать результаты последовательных анализов на ВИЧ, исключив при этом серопревалентных индивидов и индивидов, для которых имеется только один результат анализа в начальный момент времени.

Для изучения заражаемости ВИЧ инфекцией используются методы анализа данных типа времени жизни, где отказ – момент появления ВИЧ инфекции, отсчитываемый от момента проведения первого анализа, а цензурирование – время нахождения индивида под наблюдением. Для сероконвертеров наблюдается отказ, остальные индивиды считаются цензуриро-

ванными. Наиболее благоприятной, с точки зрения интерпретации результатов, является параметрическая экспоненциальная модель. Отметим, что начальный момент, начиная с которого может происходить инфицирование, не наблюдается, поэтому наблюдаемые моменты инфицирования, по сути, выбираются из усеченного на случайном уровне распределения времени отказа. Поскольку интенсивность отказа экспоненциального распределения постоянна, усеченное слева экспоненциальное распределение совпадает с исходным экспоненциальным распределением, что позволяет интерпретировать результаты в терминах распределения времени отказа. В случае если экспоненциальная модель плохо согласуется с экспериментальными данными, можно использовать распределение Вейбулла или семипараметрическую модель Кокса. При интерпретации результатов следует помнить, что полученные оценки будут относиться к распределению времени появления ВИЧ-инфекции случайного индивида данной когорты с момента формирования когорты. Это распределение может меняться, если поменять явные или неявные правила формирования когорты. Отметим, что для проверки значимости различий групп индивидов, классифицированных определенным образом по генотипам, полученные результаты могут быть с определенными оговорками интерпретированы и в терминах распределений времен отказов.

Объективно говоря, исходные данные лучше описывает модель данных типа времени жизни с интервальным цензурированием, поскольку время появления ВИЧ инфекции фиксируется по факту ее обнаружения в моменты обследований. Методы анализа интервально-цензурированных данных типа времени жизни хорошо работают в параметрическом случае, но для модели Кокса или непараметрических моделей их использование неэффективно. В случае применения модели с интервальным цензурированием лучше использовать параметрический подход. Непараметрические и семипараметрические методы лучше использовать в случае правого цензурирования. Для их использования точные времена инфицирования должны быть определены при подготовке данных. В качестве времени инфицирования обычно берут момент обнаружения заболевания или некоторую среднюю точку между моментом последнего обследования, на котором не было выявлено заболевания и моментом выявления заболевания. Последний способ выглядит более корректным, скажем, можно в качестве времени инфицирования использовать ожидаемое значение времени отказа при условии, что заболевание произошло в указанном интервале времени, в предположении параметрической модели, полученной по интервально-цензурированным данным. Проще, можно в качестве времени заболевания выбирать середину указанного интервала. Отметим, что в большинстве случаев такой подход не вполне обоснован, и полученные *P*-значения могут существенно зависеть от правила определения времени заболевания внутри соответствующего интервала и существенно отличаться от *P*-значений, полученных для интервально-цензурированных данных.

При исследовании динамики развития СПИД наибольший интерес представляет распределение времени с момента ВИЧ инфицирования до обнаружения СПИД, в первую очередь определяемое продолжительностью хронической стадии. Очевидно, что в этом случае применяются методы анализа данных типа времени жизни. Для анализа наиболее пригодны сероконвертеры, поскольку момент появления у них ВИЧ инфекции, с которого начинается развитие СПИД, известен (с точностью, определяемой временами между соседними тестированиями). Некоторую информацию о развитии СПИД можно получить и с использованием данных по серопревалентным индивидам. Использование серопревалентных индивидов наряду с сероконвертерами повышает мощность тестов, но интерпретировать результаты такого тестирования следует с осторожностью.

В рамках поставленной задачи рекомендуется провести исследование классическими методами категориального анализа с учетом серопревалентных и сероконвертировавших индивидов, а также методами анализа данных типа времени жизни только для сероконвертеров. В последнем случае удобно использовать модель Кокса, но использование категориальных методов анализа данных типа времени жизни предпочтительнее, поскольку влия-

ние неточности в измерении времени заболевания при выборе достаточно больших интервалов группировки становится меньше.

Наблюдения за вирусной нагрузкой и количеством CD4 клеток дают временные ряды (короткие временные ряды), для анализа которых обычно используют обобщенные линейные модели в предположении нормальности распределения наблюдаемых переменных, хорошо описывающие изменение изучаемых характеристик в хронической стадии. При изучении развития СПИД у серопревалентных индивидов имеет смысл использовать смешанную модель. Введение случайного фактора индивида позволяет учитывать, что течение процесса изучается не с самого момента инфицирования. Рассмотренные модели не подходят для описания острой стадии. В этом случае можно использовать смешанную модель нелинейной регрессии, где в качестве регрессии выглядит естественным выбор модели затухающих колебаний.

5. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ АНАЛИЗА

Результаты статистического анализа представляют собой результаты различных тестов на выявление зависимости отдельных генотипов с соответствующими фенотипами. Учитывая проблемы интерпретации результатов множества тестов, описанные в разделе 2, надежный выбор зависимых с фенотипом SNP возможен только в случае очень сильной связи (скажем, наличие гена CCR5 Δ 32 в гомозиготном варианте практически полностью исключает заражение ВИЧ инфекцией [10, 34]). Не столь сильные статистические зависимости в основном поглощаются случайными выбросами и их выявление не представляется возможным. Таким образом, основной целью полногеномного исследования закономерностей является скрининг, статистики критериев и P-значения осуществляют представление имеющихся статистических данных в компактной форме. Выбор SNP, которым соответствуют наименьшие P-значения, представляет интерес для дальнейшего анализа.

Успешность полногеномного анализа закономерностей во многом зависит от количества накопленной статистической информации. В связи с этим, особый интерес представляет обмен результатами статистических исследований между различными группами исследователей путем создания единой базы данных. Современные технологии Интернет такие возможности открывают. Для целей комплексного исследования создана программа визуального представления результатов множества тестов GWATCH [42], позволяющая объединить результаты множества статистических тестов на выявление зависимостей фенотипа с генотипами, полученные по результатам нескольких исследований, и выделить наиболее значимые закономерности и целые регионы малых P-значений, сопоставить результаты идентичных, уточняющих и дополняющих друг друга тестов для различных когорт индивидов.

Идеи комбинирования результатов нескольких независимых статистических тестов возникали еще в первой половине XX века в работах Фишера, Типпетта, Уоллиса, Е. Пирсона и ряда других авторов (см. [6]). Фишер [19] в качестве объединяющей статистики использовал половину суммы минус логарифмов P-значений, имеющую хи-квадрат распределение. Эта и некоторые другие объединяющие статистики изучались в работах [6, 27, 28]. Некоторые специальные случаи зависимых тестов рассматривались, в частности, в работах [7, 25]. Комбинирование результатов множества тестов и их совместная интерпретация называется мета-анализом. Имеется обширный круг работ по мета-анализу для различных статистических моделей (см. [41]). Рассмотрим некоторые эвристические аргументы, связанные с применением методов мета-анализа в задаче поиска генетических закономерностей.

Обнаружение малых P-значений для одних и тех же SNP как результатов двух или нескольких аналогичных тестов в независимых исследованиях может быть поводом для признания закономерности значимой или, как минимум, для дальнейшего ее изучения. Скажем, наличие двух P-значений соответствующих тестов, меньших, чем α , полученных по

результатам аналогичных статистических исследований на двух независимых когортах, фактически можно интерпретировать как значимую закономерность на уровне α^2. Метод использования r-й порядковой статистики, построенной по имеющимся P-значениям, для мета-анализа предлагается в работе [45]. Для интерпретации результатов анализа удобно использовать минус логарифм P-значения $s = -\log p$, большие значения этой характеристики соответствуют малым P-значениям. Имеет смысл рассмотреть суммарное значение и дисперсию характеристики s по всем результатам идентичных тестов, полученных из независимых экспериментов. Признанию закономерности значимой способствует большое суммарное значение при малой дисперсии s. Большая дисперсия s при большом суммарном значении, возможно, говорит о неоднородности влияния данного генотипа на фенотип в различных выборках, хотя при проведении множества тестов может быть обусловлена и случайными причинами. Отметим, что половина суммы значений s при справедливости основной гипотезы имеет распределение хи-квадрат, а число степеней свободы вдвое больше числа объединяемых результатов тестов. Альтернативно при интерпретации результатов идентичных тестов можно использовать непосредственно статистики критериев, а не P-значения. Следует предостеречь от искусственного разбиения исходной совокупности на группы, эффективность статистического анализа при этом не повысится.

Различные тесты для поиска генетических закономерностей по одним и тем же данным будем называть уточняющими. В пользу применения уточняющих тестов можно использовать следующие аргументы:

а) различную чувствительность уточняющих тестов к различным альтернативам;

б) бывает трудно обосновать соответствие условий выбранной модели реальным условиям, но при этом мощность статистических тестов в более жестких предположениях обычно выше.

Применение нескольких уточняющих тестов не является обязательным условием, однако их использование позволяет всесторонне взглянуть на выявленные закономерности. Сравнительный анализ результатов уточняющих тестов носит в большей степени визуальный характер и проводится для выбранных наиболее значимых хотя бы по одному из них закономерностей. В качестве примера приведем четыре уточняющих теста с различной группировкой вариантов генотипа, обсуждавшихся в разделе 4. Кодоминантная гипотеза однородности является наиболее жесткой, и наличие закономерностей в кодоминантном смысле влечет наличие различий в доминантном и рецессивном смыслах. Доминантный и рецессивный варианты тестов ориентированы на важность наличия соответствующей гомозиготы для изменения фенотипа, при этом проведение доминантного теста может решить проблему малых значений в таблице сопряженности с кодоминантной классификацией, так как число индивидов с рецессивной гомозиготой может быть достаточно малым. Аллельный тест носит особенный характер и характеризует закономерность в терминах наличия аллели у интересующего индивида. Другой пример – использование параметрических моделей, модели Кокса или непараметрической модели при анализе данных типа времени жизни. Наименьшие априорные предположения о распределении предполагает непараметрическая модель, что делает ее более универсальной по сравнению с остальными. С другой стороны, мощность непараметрических критериев обычно ниже, чем в случае модели Кокса или параметрической модели, что позволяет идентифицировать большее число потенциальных сигналов с использованием параметрических моделей или модели Кокса. Отметим, что согласованность результатов уточняющих тестов ожидаема и не может быть использована в контексте повышения значимости выявления закономерностей.

Наконец, при проведении комплексных исследований клинические данные могут содержать информацию о различных болезнях, связанных или не связанных между собой. Тесты, проводимые с использованием одних и тех же генотипических данных (для одной когорты пациентов), но для разного рода клинических данных будем называть дополняющими. Интерпретация результатов дополняющих экспериментов в наименьшей степени

поддается формализации. Тем не менее, идентификация сигналов на высоком уровне значимости двумя или несколькими дополняющими тестами одновременно дает почву для изучения схожести механизмов сопутствующих или препятствующих соответствующим заболеваниям или связи между этими заболеваниями.

Принимая во внимание потенциальную зависимость между результатами одного и того же теста для различных SNP, в первую очередь, обусловленную возможной сцепленностью близких SNP, помимо рассмотрения результатов отдельных тестов, интерес представляет изучение продолжительных сигналов – регионов с высокой концентрацией малых P-значений. Пусть $p_1,...,p_N$ – набор P-значений, $s_i = -\log p_i$, $i = 1, ..., n$. Для идентификации продолжительных сигналов можно использовать ядерное сглаживание,

$$d_i = \sum_j s_j K((i-j)/\sigma) \Big/ \sum_j K((i-j)/\sigma),$$

где K – ядро сглаживания, σ – параметр захвата. Выбор большого значения параметра σ обычно позволяет идентифицировать только продолжительные сигналы, фильтруя при этом отдельные экстремально-большие значения s_i, тогда как при малом значении σ приоритет отдается экстремально-большим значениям s_i в достаточно малом регионе. В качестве ядра сглаживания обычно выбирают четную функцию, финитную или с быстро убывающими хвостами. В частности, в случае выбора $K(u) = \mathbb{1}_{[-1,1]}(u)$, принимающей значение 1, если $u \in [-1, 1]$, и 0 – в противном случае, при $\sigma = r$ получаем усреднение $d_i = \sum_{j=i-r}^{i+r} s_j/(2r+1)$. Значения d_i в точках, недостаточно удаленных от концов интервала значений i, могут не вполне соответствовать поставленной задаче из-за отсутствия достаточного числа соседей, поэтому интерпретация больших значений d_i с близкими к концам рассматриваемого интервала значениями i, должна производиться с осторожностью.

Поскольку каждой SNP соответствует координата, то можно производить сглаживание с учетом расстояний между соседними SNP, а не с учетом их порядковых номеров. При этом

$$\widetilde{d}_i = \sum_j s_j K((c_i - c_j)/\sigma) \Big/ \sum_j K((c_i - c_j)/\sigma),$$

где c_i – координаты соответствующих SNP. Разумеется, параметр захвата должен соответствовать масштабу по координатной оси. Данный способ сглаживания выглядит более естественно, однако он имеет существенный недостаток, ибо при интерпретации полученного значения надо будет учитывать концентрацию SNP в окрестности выбранной координаты. В случае если концентрация SNP в выбранной окрестности мала, то роль отдельных SNP, входящих в данную окрестность, будет выше, чем для аналогичной окрестности с высокой концентрацией SNP.

Альтернативно для идентификации продолжительных сигналов выглядит естественным использование статистик типа HC.

Литература

1. *Леман Э.* Проверка статистических гипотез. 2-е изд. М.: Наука, 1979.

2. *Малов С.В., Шевченко А.К., О'Брайан С.Д.* Поиск генетических закономерностей. Часть 1. Статистические методы // Компьютерные инструменты в образовании, 2013. № 5. С. 17–32.

3. *Barre-Sinoussi F., Chermann J.C., Rey F., Nugeyre M.T., Chamaret S., Gruest J., Dauguet C., Axler-Blin C., Vezinet-Brun F., Rouzioux C., Rozenbaum W., Montagnier L.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS) // Science, 1983. Vol. 220, № 4599. P. 868–871.

4. *Benjamini, Y., Hochberg, Y.* Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the Royal Statistical Society: Series B, 1995. Vol. 57. P. 289–300.

5. *Benjamini Y., Yekutieli D.* The control of the false discovery rate in multiple testing under dependency // The Annals Statistics, 2001. Vol. 29. P. 1165–1188.

6. *Birnbaum A.* Combining Independent Tests of Significance // Journal of the American Statistical Association, 1954. Vol. 49, № 267. P. 559–574.

7. *Brown M.B.* A Method for Combining Non-Independent, One-Sided Tests of Significance // Biometrics, 1975. Vol. 31, № 4. P. 987–992.

8. *Browning B.L.* PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies // BMC Bioinformatics, 2008. Vol. 9. P. 309.

9. *Cheverud J.M.* A simple correction for multiple comparisons in interval mapping genome scans // Heredity, 2001. Vol. 87. P. 52–58.

10. *Dean M., Carrington M., Winkler C., Huttley G.A., Smith M.W., Allikmets R., Goedert J.J., Buchbinder S.P., Vittinghoff E., Gomperts E., Donfield S., Vlahov D., Kaslow R., Saah A., Rinaldo C., Detels R., O'Brien S.J.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. HGDS, MACS, MHCS, SF City Cohort, ALIVE Study // Science, 1996. Vol. 273. P. 1856–1862.

11. *Donoho D., Jin J.* Higher criticism for detecting sparse heterogeneous mixtures // The Annals of Statistics, 2004. Vol. 32, № 3. P. 962–994.

12. *Donoho D., Jin J.* Higher criticism thresholding: optimal feature selection when useful features are rare and weak // PNAS, 2008. Vol. 105, № 39. P. 14790–14795.

13. *Efron B.,Tibshirani R, Storey J.D., Tusher V.* Empirical Bayesan alysis of a microarray experiment // Journal of the American Statistical Association, 2001. Vol. 96. P. 1151–1160.

14. Farcomeni A. Some Results on the Control of the False Discovery Rate under Dependence // Scandinavian Journal of Statistics, 2007. Vol. 34, № 2. P. 275–297.

15. *Fahrmeir L., Kaufman H.* Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model // The Annals of Statistics, 1985. Vol. 13, № 1. P. 342–368.

16. *Fay M.P.* Rank invariant tests for interval censored data under the grouped continuous model // Biometrics, 1996. Vol. 52. P. 811–822.

17. *Fay M.P., Shaw P.A.* Exact and asymptotic weighted logrank tests for interval censored data: the interval R Package // Journal of Statistical Software, 2010. Vol. 36, № 2. P. 1–34.

18. *Finkelstein D.M.* A proportional hazards model for intervalcensored failure time data // Biometrics, 1986. Vol. 42. P. 845–854.

19. *Fisher R.A.* Statistical Methods for Research Workers. London: Oliver and Boyd, 11th ed., 1950. P. 99–101.

20. *Gallo R.C., Sarin P.S., Gelmann E.P., Robert-Guroff M., Richardson E., Kalyanaraman V.S., Mann D., Sidhu G.D., Stahl R.E., Zolla-Pazner S., Leibowitch J., Popovic M.* Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science, 1983. Vol. 220, № 4599. P. 865–867.

21. *Gao X., Starmer J., Martin E.R.* A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms // Genetic Epidemiology, 2008. Vol. 32. P. 361–369.

22. *Hall P., Jin J.* Innovated Higer Criticism for detecting sparse signals in correlated noise // The Annals of Statistics, 2010. Vol. 38, № 3. P. 1686–1732.

23. Holm S. A Simple Sequentially Rejective Multiple Test Procedure // Scandinavian Journal of Statistics, 1979. Vol. 6. P. 65–70.

24. *Ingster Y.I.* Some problems of hypothesis testing leading to infinitely divisible distribution // Mathematical Methods of Statistics, 1997. Vol. 6. P. 47–69.

25. *Kost J.T., McDermott M.P.* Combining dependent P-values // Statistics \& Probability Letters, 2002. Vol. 60. P. 183–190.

26. *Li J, Ji L.* Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix // Heredity, 2005. Vol. 95. P. 221–227.

27. *Littell R.C., Folks J.L.* Asymptotic Optimality of Fisher's Method of Combining Independent Tests // Journal of the American Statistical Association, 1971. Vol. 68, № 341. P. 802–806.

28. *Littell R.C., Folks J.L.* Asymptotic Optimality of Fisher's Method of Combining Independent Tests II. // Journal of the American Statistical Association, 1973. Vol. 68, № 341. P. 193–194.

29. *Malov S.V., O'Brien S.J.* On Survival Categorical Methods with Applications in Epidemiology and AIDS Research // In «Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control», Proceedings of the AMSA'13 conference (Novosibirsk, September 25–27, 2013). Новосибирск: НГТУ, 2013. P. 173–180.

30. *Moskvina V., Schmidt K.M.* On Multiple-Testing Correction in Genome-Wide Association Studies // Genetic Epidemiology, 2008. Vol. 32. P. 567–573.

31. *Nyholt D.R.* A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other // The American Journal of Human Genetics, 2004. Vol. 74. P. 765–769.

32. *O'Brien S.J., Hendrickson S.* Host Genomic Influences on HIV/AIDS // Genome Biology, 2013. Vol. 14:201.

33. *O'Brien S.J., Nelson G.W.* Human genes that limit AIDS // Nature Genetics, 2004. Vol. 36. P. 565–574.

34. *O'Brien S.J., Dean M.* In search of AIDS-resistance genes // Scientific American, 1997, Vol. 277. P. 44–51.

35. *Pahl R., Schafer H.* PERMORY: an LDexploiting permutation test algorithm for powerful genome-

wide association testing // Bioinformatics, 2010. Vol. 26. P. 2093–2100.

36. *Pollard K.S., van der Laan, M.J.* Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data // Journal of Statistical Planning and Inference, 2002. Vol. 125. P. 85–100.

37. *Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses // The American Journal of Human Genetics, 2007, Vol. 81. P. 559–575.

38. *R Core Team.* R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2011.

39. *Storey J.D.* A direct approach to false discovery rates // Journal of the Royal Statistical Society: Ser. B, 2002. Vol. 64. P. 479–498.

40. *Sun J.* A non-parametric test for interval-censored failure time data with applications to AIDS studies. // Statistics in Medicine, 1996. Vol. 15. P. 1387–1395.

41. *Sutton A. J., Higgins J.P.T.* Recent Developments in Meta-Analysis // Statistics in Medicine, 2008. Vol. 27. P. 625–650.

42. *Svitin A., Malov S.V., Cherkasov N., Dobrynin P., Guan Li, Geerts P., Troyer J., Hendrickson-Lambert S., Hutcheson-Dilks H., Oleksyk T.K., Donfield S., Gomperts E., Jabs D.A., Van Natta M., Harrigan R., Brumme Z.L., O'Brien S.J.* Gene Discovery and Data Sharing in Disease Association Analyses Across the Genome / to appear.

43. *Tukey J.W.* T13 N: The higher criticism. Course Notes, Statistics 411. Princeton Univ., 1976.

44. *Wang Z., Gardiner J.C., Ramamoorthi R.V.* Identifiability in interval censorship model // Statistics & Probability Letters, 1994. Vol. 21. P. 215–221.

45. *Wilkinson B.* A Statistical Consideration in Psychological Research // Psychological Bulletin, 1951. Vol. 48, 156–157.

46. *Yekutieli D., Benjamini Y.* Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference, 1999. Vol. 82. P. 171–196.

GENOME ASSOCIATIONS DISCOVERY. PART 2: MULTIPLE TESTING PROBLEM, COMPUTER TOOLS AND APPLICATIONS

Abstract

Methodology of genome association discovery is discussed comprehensively in this paper. The multiple testing problem, implementation of the statistical methods discussed in the first part of the paper and their applications in HIV-infection and AIDS progression studies are considered here. Several ideas on common interpretation results of independent or dependent similar tests will be given too.

Keywords: Whole genome association discovery, genome wide association study (GWAS), multiple testing problem, HIV/AIDS research.

Центр геномной биоинформатики им. Ф.Г. Добржанского:

Малов Сергей Васильевич, кандидат физико-математических наук, доцент, старший научный сотрудник лаборатории, malovs@sm14820.spb.edu,

Шевченко Андрей Константинович, лаборант-исследователь, andrey.k.shevchenko@gmail.com,

О'Брайен Стефан Джеймс (Stephen J. O'Brien), доктор философии в области биологии (PhD in biology), главный научный сотрудник, lgdchief@gmail.com.