Метод локализации сигнала с использованием скользящего среднего.

В результате выполнения процедуры множественного тестирования получены результаты N статистических тестов хи-квадрат – значения статистик $X_1^2, \ldots, X_N^2$. Известно, что при справедливости нулевой гипотезы соответствующяя статистика критерия имеет распределение $\chi^2_{(k-1)(d-1)}$, где k ($k = 2$ или $k = 3$) – число уровней генотипа, а d – число уровней фенотипа. Сделаем монотоное преобразование статистик $p_i = 1 - F(X_i)$, где F – функция распределения $\chi^2_{(k-1)(d-1)}$. Таким образом, результаты статистических тестов могут быть переписаны в виде $p_1, \ldots, p_N$. Метод Бонферрони идентификации сигнала заключается в выборе позиций (локусов), для которых $p_i \leqslant \alpha/N$.

Ослабим задачу. Предположим, что нам не нужно точно идентифицировать сигналы, а достаточно идентифицировать области, в которых присутствует хоть один сигнал (используем скользящие окна $1, \ldots, w$; $2, \ldots, w+1$; ... ;$N-w+1, \ldots, N$). В рамках каждого окна считается $T_k = -\sum_{i=k}^{k+w-1} \log p_i$. Известно, что распределение данной статистики в случае независимости результатов тестов $\Gamma(w, 1)$ – гамма распределение. Используя преобразование $p_i^* = 1 - G(T_i)$ получаем P-значения уже для скользящих средних (наивные). Практика показывает, что полученные P-значения обычно сильно занижены, поскольку результаты тестов часто зависимы. Требуется уточнить их с использованием методов Монте-Карло. Сделать это для всех значений невозможно, поскольку для этого требуется чрезмерно много вычислений. Для сокращения вычислений выбирают только окна с наименьшими наивными P-значениями (спокойно можно выбрать $p_i^* < 10^{-8}$, ибо меньше после уточнения они не станут). Считаем, что таблицу топов p_i^* мы умеем строить и на входе будет известно окно, в котором надо сделать точнение.

Критерий случайных перестановок. Пусть имеется набор значений неблюдаемого признака (фенотипа) - вектор Y и $(Z_1, \ldots, Z_w)$ - наборы генотипов в рамках выбранного окна. Если выбрать случайную перестановку индексов $\sigma = (\sigma_1, \ldots, \sigma_w)$, то величина $Y_\sigma = (Y_{\sigma_1}, \ldots, Y_{\sigma_n})$ будет независима с каждым Z_i. Для каждой случайной перестановки вычисляем статистику T_σ – получаем $T_{\sigma(1)}, \ldots, T_{\sigma(m)}$, где m – число перестановок. Добавляем в выборку истинное значение T и упорядочиваем по возрастанию. Если T – s-я порядковая статистика, то уточненное P-значение равно $p^*_{MC} = s/(m+1)$.

Следует отметить, что если реальное значение p^* мало, то для плучения корректного результата поребуется на порядок (или несколько порядков) больше перестановок, чем $1/p^*$. Адаптивный подход заключается в том, что m не фиксируется, а растет до тех пор пока s не достигнет некоторого напередред заданного значения (например, 10)

Алгоритм вычисления поправки P-value для "Density": категориальный тест χ^2.

Исходные данные:

1. Таблица наиболее значимых SNP (обязательные колонки: «snp.id»; «test.id», «lower», «upper», «naive.pv» или индекс) – S.
2. Генетические данные:

(а). Таблица генотипов (PLINK, файл ".bed") – G матрица $n \times N$;
(б). таблица "id"(PLINK, файл ".bim") – B вектор длины N;
(в). таблица фенотипов (PLINK, файл ".fam" или текстовый) – A вектор длины n.

Управляющие значения:
1) repl – наибольшее число репликаций;
2) k – натуральное число, для контроля остановки адаптивной поправки;
3) adp – TRUE/FALSE (adp можно сделать FALSE, если k-пустое значение и TRUE - иначе)

Цикл по строкам таблицы S (фиксируем номер строки $= i$):
{
Шаг 1: *(подготовка данных)*
1). Считываем lower[i]; upper[i] из таблицы S
2). Считываем test.id[i]
3). По 3-му полю test.id[i] выбираем форму группировки генотипа 'alt=(«cd»,«d»,«r»,«a»)'
4). Выбираем все столбцы G между 'lower[i]' и 'upper[i]' = матрица G'

Все готово для начала вычислений

Шаг 2: *(преобразование генотипа под тип теста)*:
Выбор:
alt=«cd»: $G^+ \equiv G'$

$$\text{alt=«r»: } G^+[i,j] = \begin{cases} 0, G'[i,j] = 0 | G'[i,j] = 1, \\ 1, G'[i,j] = 2, \\ 3, G'[i,j] = 3. \end{cases}$$

$$\text{alt=«d»: } G^+[i,j] = \begin{cases} 0, G'[i,j] = 0, \\ 1, G'[i,j] = 1 | G'[i,j] = 2, \\ 3, G'[i,j] = 3. \end{cases}$$

$$\text{alt=«a»: } G^+[i,j] = \begin{cases} 0, G'[i,j] = 0 | G'[i,j] = 1, \\ 1, G'[i,j] = 2, \\ 3, G'[i,j] = 3, \end{cases} ; \quad G^+[n+i,j] = \begin{cases} 0, G'[i,j] = 0, \\ 1, G'[i,j] = 1 | G'[i,j] = 2, \\ 3, G'[i,j] = 3 \end{cases} ;$$

$$A^+[i] = A[i]; \quad A^+[n+i] = A[i]; A = A^+.$$

Для дальнейших вычислений используем генотип G^+ и фенотип A.

Шаг 3: *(моделирование)*:
Тело цикла (подпрограмма вычисления среднего при каждой репликации)
Подпрограмма 1 (Аргументы=(G^+,A^σ)): Цикл *for (r in 1:w)*
{
а) выбираем $x = G^+[,r]$; $y = A^\sigma$
б) идентифицируем уровни x: $\{x_l\}$, и уровни y: $\{y_t\}$
в) вычисляем таблицу ν: $\nu[l,r] = \#\{x = x_l, y = y_r\}$

г) если обе d_1 и d_2 размерности ν больше 1, то вычисляем статистику χ^2 по формуле

$$X^2 = \sum_{l,r} \frac{(\nu[l,r] - \nu[l,\bullet]\nu[\bullet,r]/n)^2}{\nu[l,\bullet]\nu[\bullet,r]/n},$$

где $\nu[l,\bullet] = \sum_r \nu[l,r]$, $\nu[\bullet,r] = \sum_l \nu[l,r]$;
если $d1 = 1 | d2 = 1$ - вычисление невозможно, возврат.
д) Вычисляем P-value по формуле

$$P[l] = \gamma(s, 2X^2)/\Gamma(s)$$

где $s = (d_1 - 1)(d_2 - 1)$; $\Gamma(s) = \int_0^\infty u^{s-1}e^{-x}dx$ – гамма-функция; $\gamma(s,x)$ – верхняя неполная гамма функция

$$\gamma(s,x) = \int_x^\infty u^{s-1}e^{-x}dx.$$

}
2) вычисляем среднее значение $D = \sum_l(-\log P[l])/w^*$, w^* – число непустых значений.
3) выводим D.

Моделирование
$X = G^+$; $y = A$ (исходный усеченный набор генотипов и фенотип)
1. Запускаем подпрограмму 1 с аргументами X и y.
Результат: D_0
2. Выбор (m – число значений статистик, больших D_0)
adp=TRUE: $m = 0$; while (m<=k & m<=repl)
adp=FALSE: for (s in 0:repl)
{
а. $\sigma = (\sigma_1, \ldots, \sigma_n)$ «random permutation (1,. . .,n)»
б. $y = A[\sigma]$ (перестроеный фенотип)
в. Запускаем подпрограмму 1 с аргументами x и y.
Результат: D_s
г. Если подпрограмма 1 выдала результат и $D_s > D_0$ то $m = m + 1$
г.′ Если подпрограмма 1 выдала результат $s = s + 1$
}
3. Результат PVA=m/s (скорректированное P-value)
}