Актуальные задачи в полногеномных анализах ассоциаций

Александр Алексеевич Трушин, НГУ, Новосибирск

ОСНОВНЫЕ ПОЛОЖЕНИЯ

Single Nucleotide Polymorphism (SNP) - отличия последовательности ДНК размером в один нуклеотид между людьми или между парными хромосомами одного человека.

В силу диплоидности хромосомного набора каждый SNP представлен в любом человеке двумя аллелями из набора $\{A, C, T, G\}$:

$$\text{Генотип } X \in \{BB, Bb, bb\};\ B, b \in \{A, C, T, G\},\ B \neq b.$$

На выборке из n индивидов исследуется ассоциированность различных SNP с развитием признака человека (фенотипа).

АДДИТИВНАЯ SINGLE-SNP МОДЕЛЬ

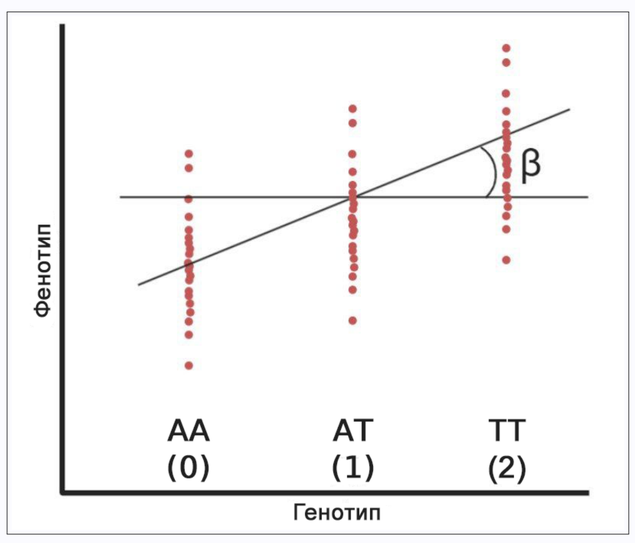
Пусть лишь один SNP оказывает влияние на фенотип.

Тогда можно поставить задачу поиска такого SNP.

Кодирование $\{BB, Bb, bb\} \to \{0, 1, 2\}$ позволяет применить модель линейной регрессии:

$$\vec{Y} = \mu\vec{1} + \beta\vec{X} + \vec{\varepsilon},$$

где $\vec{Y}$ - вектор фенотипов, $\vec{X}$ - вектор генотипов, β - линейный эффект.



Проверяется основная гипотеза $H_0 : \beta = 0$ против альтернативной $H_1 : \beta \neq 0$.

Статистика $t = \frac{\hat{\beta}}{s_{\hat{\beta}}} \left(\sim T_{n-2} \xrightarrow[n\to\infty]{\text{п.н.}} N_{0,1} \right)$, используемая для проверки H_0, определяет уровень значимости p как вероятность ошибки первого рода.

SNP с p-значением ниже порогового признаются значимыми.

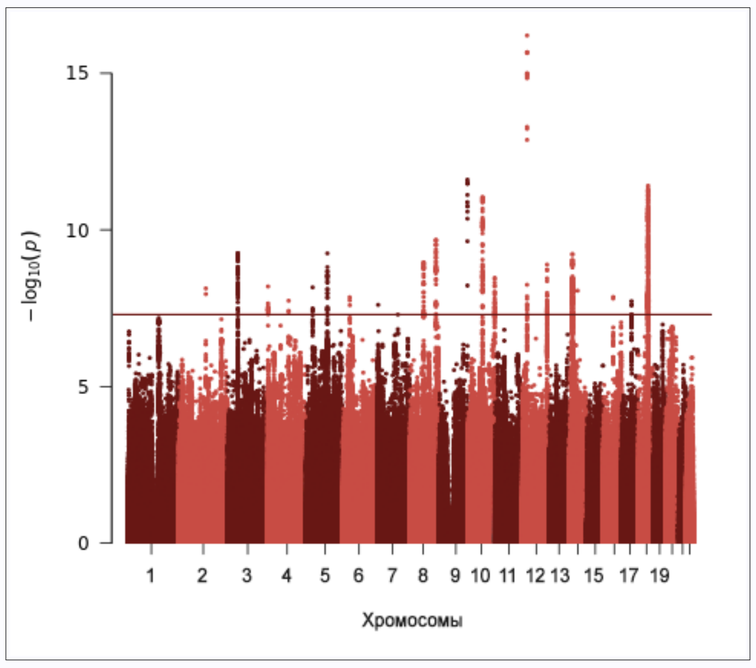


График Манхэттена полногеномного анализа ассоциаций хронической боли в спине
Каждый исследуемый SNP отмечен точкой на графике.
Горизонтальная линия отражает порог значимости для p-значения.

СУММАРНЫЕ СТАТИСТИКИ

С целью увеличения мощности исследования часто анализируются результаты многих исследований одного и того же признака.

Однако по причинам приватности индивидуальные данные не публикуются.

Вместо этого публикуются суммарные статистики - выжимка по каждому SNP [1]: геномные координаты, значение аллелей B и b и их частота в выборке, а также $\hat{\beta}$, $s_{\hat{\beta}}$, t и p-значение из Single-SNP анализа.

Методы полногеномных анализов ассоциаций, использующие ограниченные данные суммарных статистик вместо полных данных, позволяют значительно увеличить мощность исследования.

ПРОБЛЕМА НЕАДДИТИВНОСТИ ЭФФЕКТА

Кодирование $bb \mapsto 0,\ Bb \mapsto 1,\ BB \mapsto 2$ влечёт предположение о пропорциональности изменения эффекта количеству аллелей B.

Если есть основания полагать неаддитивность эффекта, можно использовать двухпараметрическую модель:

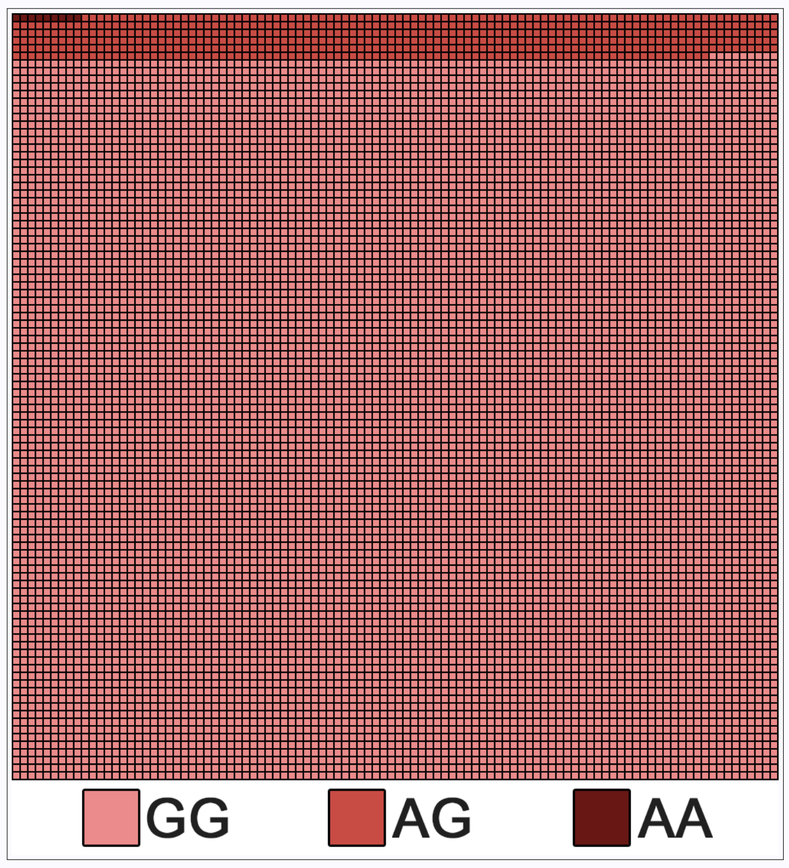
$$\vec{Y} = \mu\vec{1} + \beta_1\overrightarrow{\mathcal{I}(Bb)} + \beta_2\overrightarrow{\mathcal{I}(BB)} + \vec{\varepsilon}.$$

Гипотеза аддитивности $H_A : \beta_2 = 2\beta_1$, а также гипотезы других распространённых форм развития фенотипа (например, $H_D : \beta_1 = 0$ и $H_R : \beta_1 = \beta_2$) требуют дополнительных проверок [2].

ПРОБЛЕМА РЕДКИХ ГЕНОВ

Для каждого SNP номинально выделяют мажорный (B) и минорный (b) аллели - более и менее распространённые в выборке нуклеотиды соответственно.

В соответствие с частотой минорного аллеля (MAF) SNP делятся на обычные $(MAF \geq 3\%)$ и редкие $(MAF < 3\%)$. Анализ редких генов сложен из-за маленькой подвыборки индивидов с определённым генотипом.



Распределение генотипов rs1982243, в соответствие с законом Харди-Вайнберга.
В выборке размера 10^5 при частоте минорного аллеля 3% генотип AA
будет наблюдаться в среднем лишь у 9 индивидов.

ПРОБЛЕМА СОВОКУПНОГО АНАЛИЗА

В современных полногеномных исследованиях интерес представляют комплексно наследуемые признаки, на развитие которых влияет совокупность многих генов [3].

Таким образом, предположение о единственном SNP, оказывающим влияние на фенотип, становится недопустимым.

Логичный шаг — применение моделей многомерной линейной регрессии, однако на практике это затруднительно: исследуемых SNP на несколько порядков больше, чем индивидов в выборке [4].

БЛАГОДАРНОСТИ

Доклад основан на совместной работе с Е.И. Прокопенко и А.П. Ковалевским и подготовлен при помощи лаборатории рекомбинационного и сегрегационного анализа ИЦиГ СО РАН.

ЛИТЕРАТУРА

1. S. Burgess, A. Butterworth, S.G. Thompson, *Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data* , Genet Epidemiol., **37** (2013), 658-665.
2. Я.А. Цепилов, *Разработка и применение новых моделей в полногеномном анализе ассоциаций*, дис. канд. биол. наук, ФИИ ИЦиГ СО РАН (2016).
3. Y. Tsepilov et al., *Analysis of genetically independent phenotypes identifies shared genetic factors associated with chronic musculoskeletal pain conditions*, Commun Biol., **3** (2020), 329.
4. J. Yang et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits*, Nat Genet., **44** (2012), 369-375.

МАТЕМАТИЧЕСКИЙ ЦЕНТР В АКАДЕМГОРОДКЕ

ИНСТИТУТ МАТЕМАТИКИ им. С.Л. СОБОЛЕВА СО РАН 1957

ИЦГ

Novosibirsk State University *THE REAL SCIENCE