Теория селекции

Лекция 4. GWAS
Полногено́мный анализ ассоциа́ций

В.М.Ефимов, д.б.н.

Институт цитологии и генетики СО РАН

© ИЦиГ СО РАН

GWAS

Полногено́мный анализ ассоциа́ций (англ. GWAS, Genome-Wide Association Studies) — направление биологических (как правило, биомедицинских) исследований, связанных с исследованием ассоциаций между геномными вариантами и фенотипическими признаками. Часто под полногеномным анализом ассоциаций подразумевают только поиск связей между однонуклеотидными полиморфизмами и заболеваниями человека, однако термин употребим и к другим организмам.

(Википедия)

© ИЦиГ СО РАН

GWAS

Обычно сравнивают геномы группы больных людей с геномами контрольной группы, включающей в себя аналогичных по возрасту, полу и другим признакам здоровых людей. Материалом для исследования являются образцы ДНК каждого участника исследования. Если удается выявить варианты геномов (точнее, совокупность аллелей), которые значимо чаще встречаются у людей с данным заболеванием, то говорят, что такой вариант связан, или ассоциирован, с болезнью. В отличие от методов, которые проверяют один или несколько конкретных участков генома, полногеномный поиск ассоциаций использует полную последовательность ДНК. Следует отметить, что этот подход к исследованиям не выявляет мутации, ставшие причиной заболевания, а только более или менее значительную корреляцию с заболеванием или другим признаком

(Википедия)

© ИЦиГ СО РАН

GWAS

Вторая по важности область применения **GWAS** — фармакогенетика, то есть поиск аллелей, связанных с метаболизмом лекарственных препаратов и их побочными эффектами.

Первые успешные исследования полногеномных ассоциаций были проведены на больных макулодистрофией и опубликованы в 2005 году. Были обнаружены два аутосомных однонуклеотидных полиморфизма. К 2011 году были протестированы сотни тысяч людей, было проведено более 1200 исследований полногеномных ассоциаций для 200 заболеваний и фенотипических проявлений, в результате было найдено около 4000 однонуклеотидных полиморфизмов. Ряд исследований полногеномных ассоциаций был раскритикован за пренебрежение контролем качества. В целом методология до сих пор является предметом для споров.

(Википедия)

© ИЦиГ СО РАН

GWAS

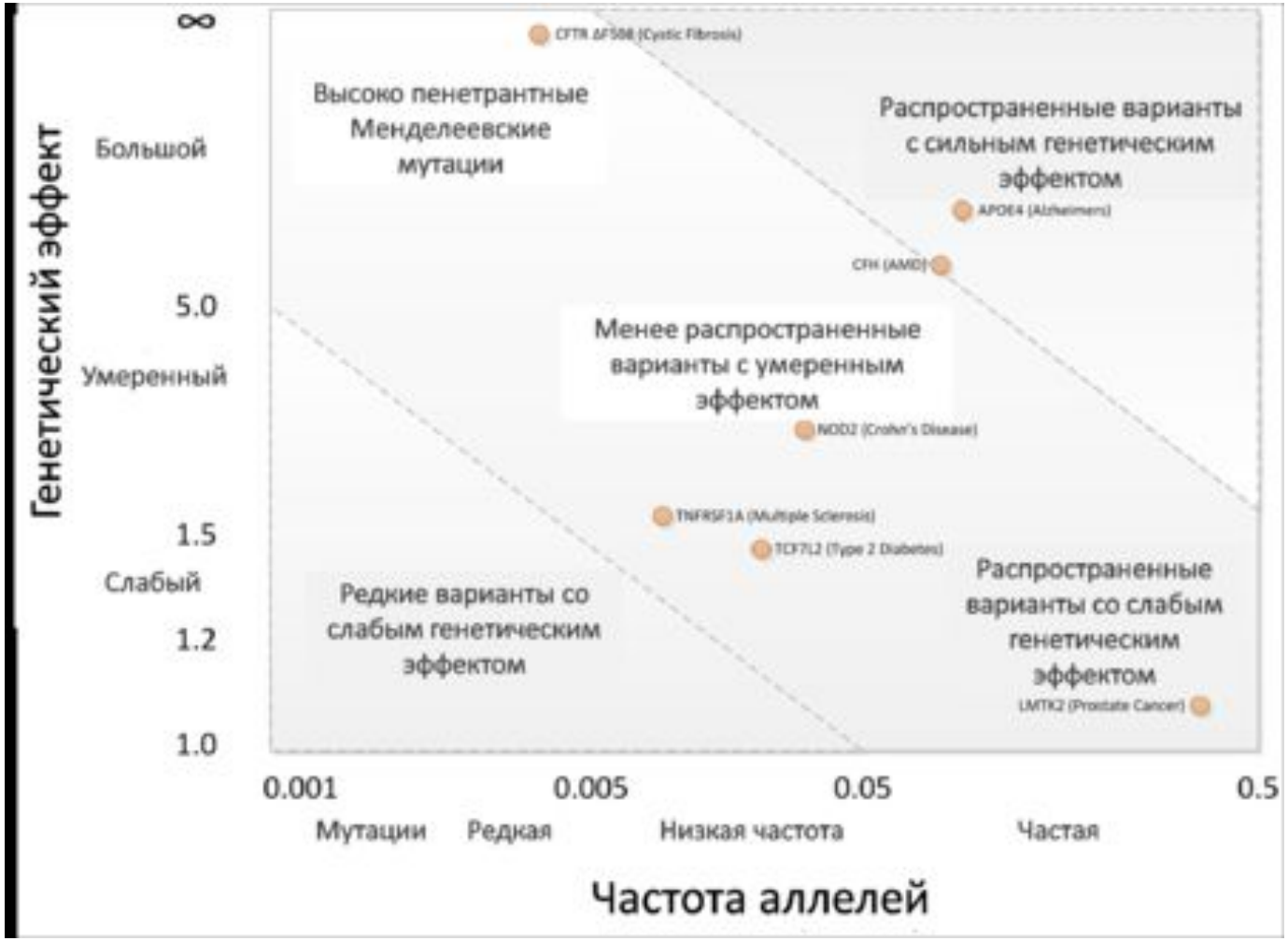
В основе поиска полногеномных ассоциаций как правило лежит сравнение геномов двух групп людей: носителей исследуемого фенотипа (заболевания) и контрольной группы. Для всех индивидуумов производится генотипирование для большинства известных однонуклеотидных полиморфизмов (SNP). Далее, для каждого SNP проверяется, насколько значимы различия в распределении частот аллелей между исследуемой и контрольной группой.

Альтернативой делению на две группы в полногеномных исследованиях является количественный анализ фенотипа, например, рост, концентрация биомаркера или экспрессия гена. Кроме того, могут быть использованы данные о пенетрантности исследуемых аллелей.

(Википедия)

© ИЦиГ СО РАН

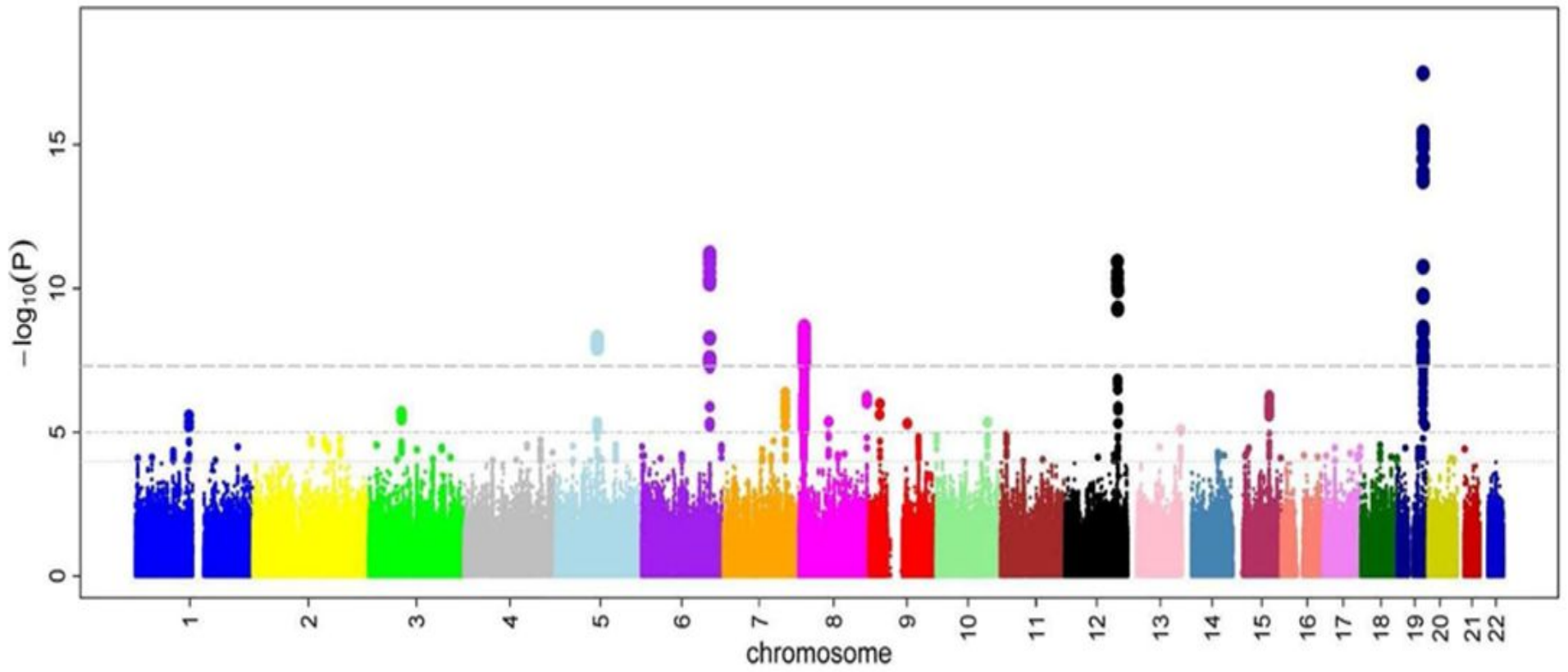
GWAS



(Википедия)

© ИЦиГ СО РАН

GWAS



Верхняя прямая − критерий Бонферрони, нижняя − FDR

© ИЦиГ СО РАН

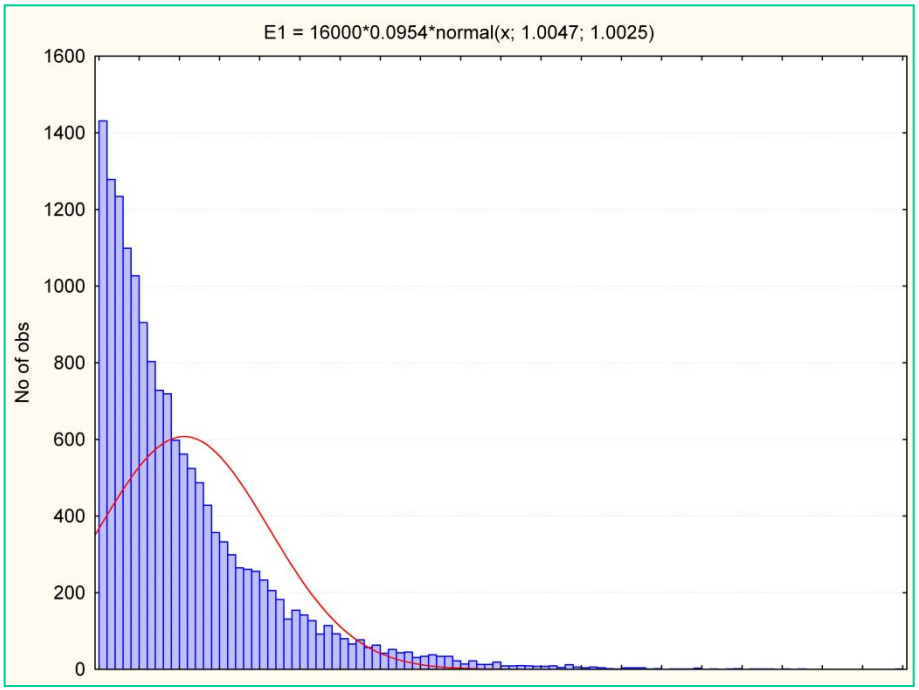
Если нам нужно сравнить средние двух заданных групп, обычно используется *t*-критерий Стьюдента или Уэлша. Если число групп больше двух, используется *F*-критерий Фишера (ANOVA).

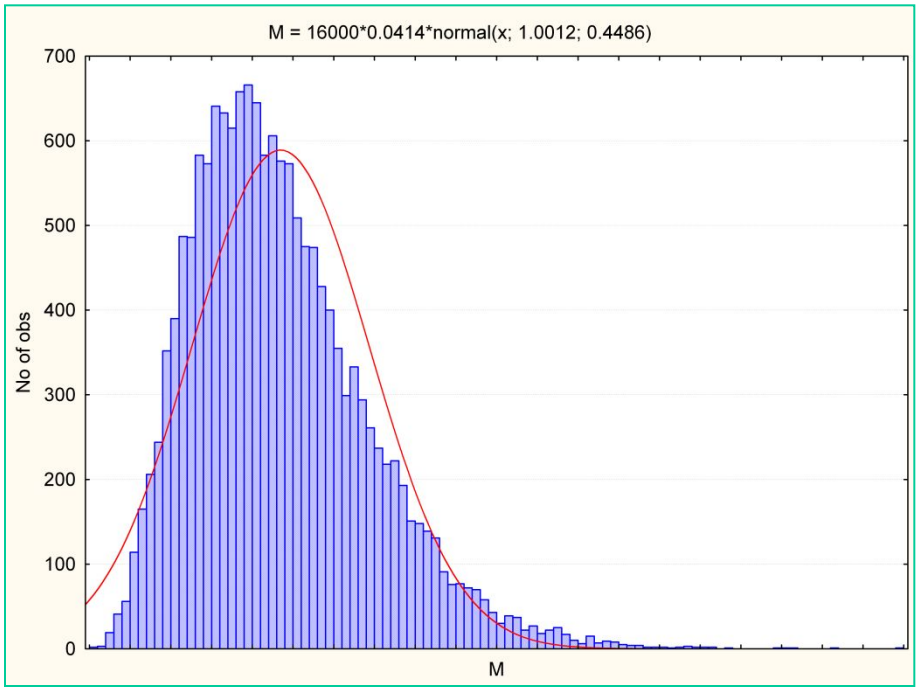
Распределение значений внутри групп предполагается нормальным. Строго говоря, только при справедливости этого предположения можно использовать *t*- или *F*-критерий.

© ИЦиГ СО РАН

На самом деле нормальность распределения требуется не для самих выборок, а только для их средних, а они всегда распределены приближенно нормально для любых распределений даже при небольших *N* в силу центральной предельной теоремы.

© ИЦиГ СО РАН





Слева: экспоненциальное распределение (λ=1). Справа: распределение выборочных средних из него объема n=10. Красная кривая: аппроксимация нормальным распределением.

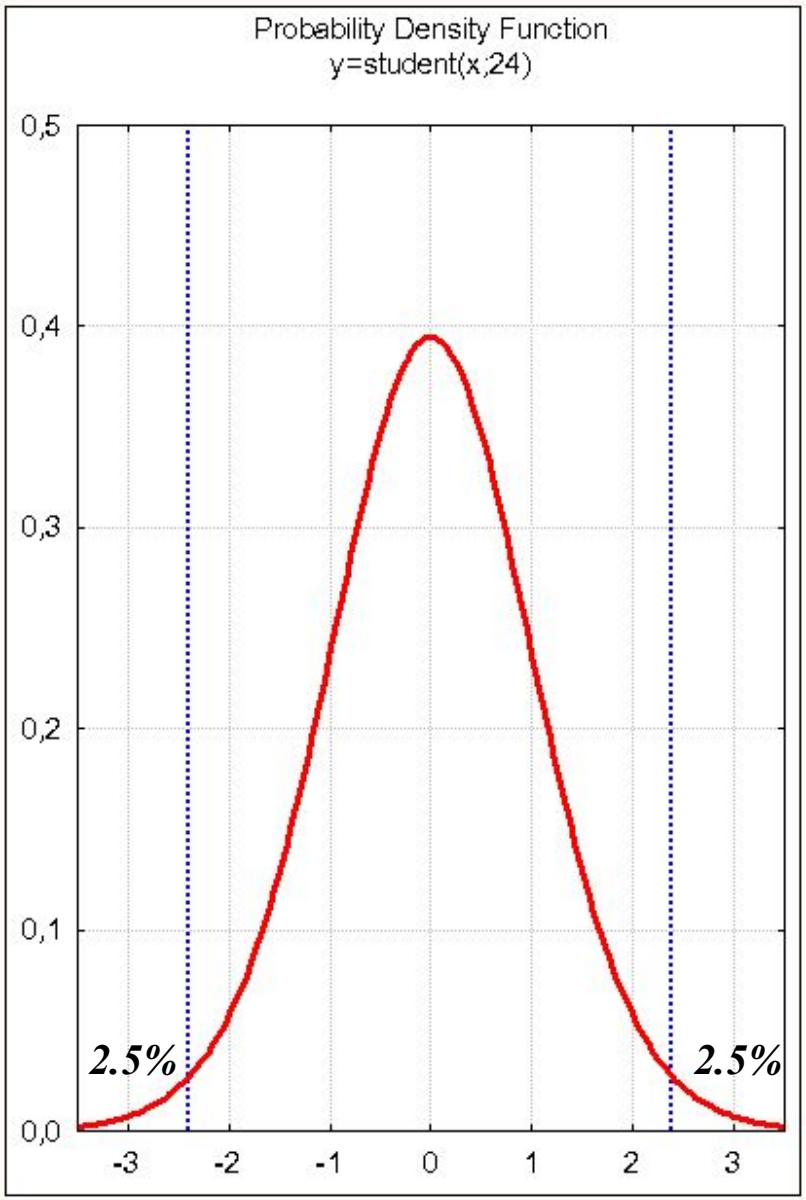
© ИЦиГ СО РАН

Нулевая гипотеза заключается в том, что различия между выборками являются **случайными** и все выборки на самом деле взяты из **одного и того же** генерального распределения.

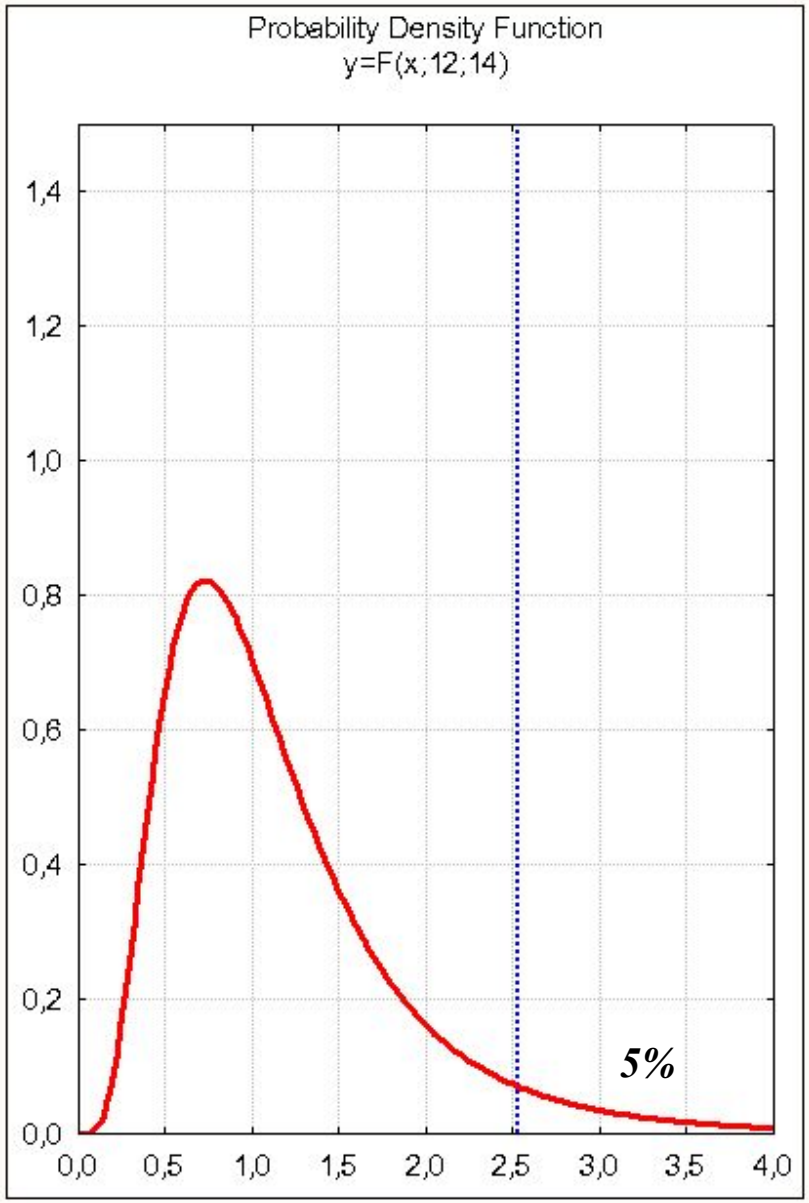
Если значения рассчитанного критерия превышают табличные значения для заданного уровня значимости α (**0.05; 0.01; 0.001**), то мы считаем, что нулевая гипотеза **не подтверждается**. Это ошибка **первого** рода.

© ИЦиГ СО РАН

Распределение t-критерия



Распределение F-критерия



© ИЦиТ СО РАН

Но это верно для одного эксперимента. Если эксперимент повторяется N раз, то вероятность, что хотя бы один выборочный критерий случайно превысит табличный уровень, равна $1-(1-\alpha)^N$. Если мы хотим гарантировать, что с вероятностью α ни один из N рассчитанных критериев не превысит табличный уровень, то мы должны выбрать табличный уровень, равный α/N. Это критерий Бонферрони.

© ИЦиГ СО РАН

Этот критерий слишком жесткий. Если, например, $\alpha = 0.05$ и $N = 10^6$, то $p = \alpha/N = 5 \cdot 10^{-8}$ и $lg(p) = -7.301$.

Через критерий Бонферрони проходит слишком мало полезной информации. Поэтому в 1995 году Бенджамини и Хохберг предложили критерий FDR (False Discovery Rate). Этот критерий сейчас, например, является основным в микрочиповых исследованиях.

© ИЦиГ СО РАН

FDR (false discovery rate)

Критерий FDR заключается в следущем. Произвольно выбираем уровень значимости p. Ожидаемое число случайных выборочных критериев, которые должны превысить соответствующее p табличное значение t_p, равно pN. Мы сравниваем это число с k -- реальным числом выборочных критериев, превысивших t_p. Если pN достаточно мало по сравнению с k, например, меньше 5%, то t_p в качестве порогового значения нас вполне устраивает.

© ИЦиГ СО РАН

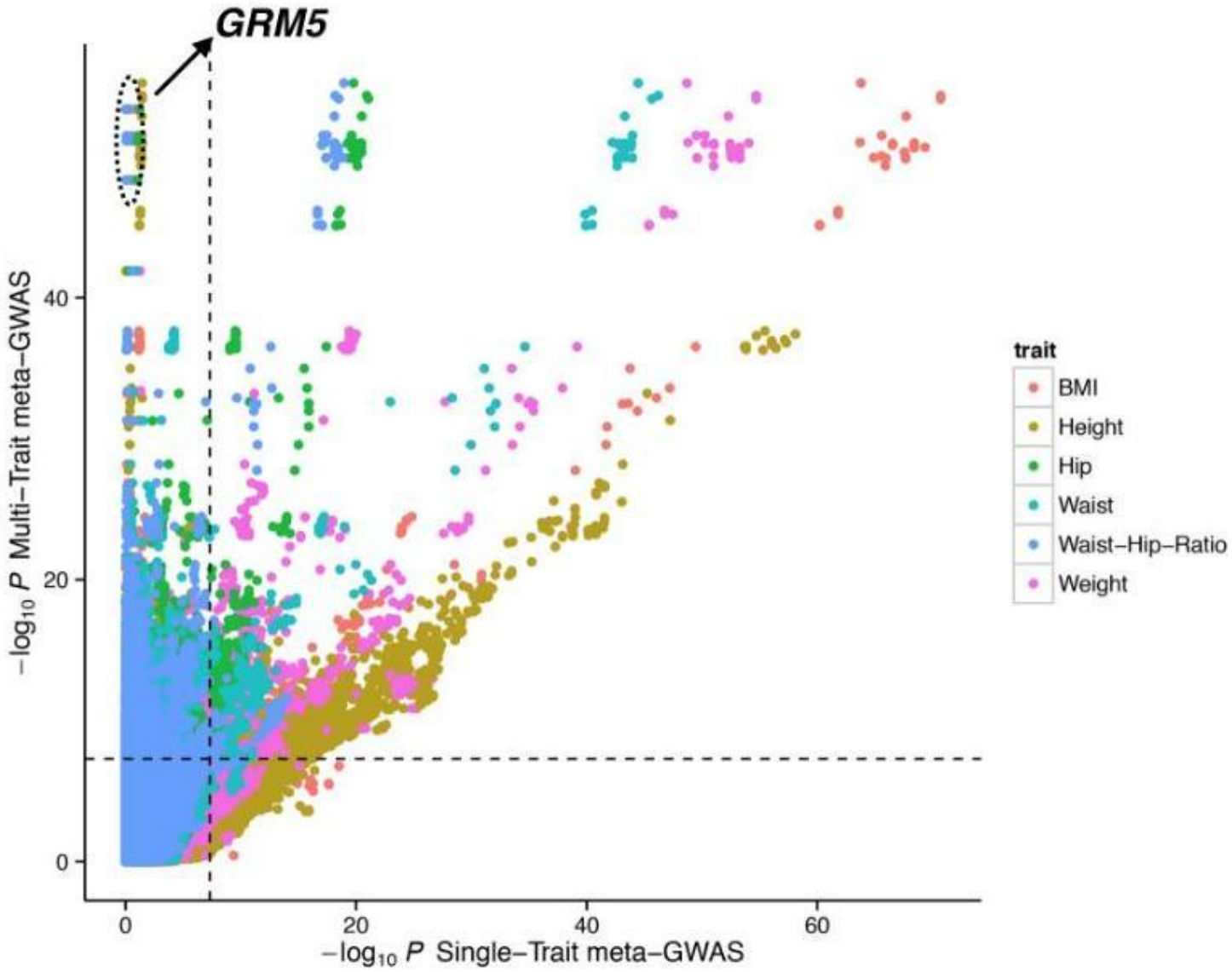
FDR (false discovery rate)

($N=10^6$)

t	p	lg(p)	pN	k
-13.358	2.41E-12	-11.62	2.41E-06	1
-12.488	1.36E-11	-10.87	1.36E-05	2
-15.238	1.55E-11	-10.81	1.55E-05	3
-14.264	2.31E-11	-10.64	2.31E-05	4
-12.227	6.73E-11	-10.17	6.73E-05	5
-11.223	7.15E-11	-10.15	7.15E-05	6
-11.558	7.53E-11	-10.12	7.53E-05	7
-11.525	9.69E-11	-10.01	9.69E-05	8
-12.962	1.07E-10	-9.97	1.07E-04	9
-10.704	1.30E-10	-9.88	1.30E-04	10

© ИЦиГ СО РАН

Multi-Trait meta-GWAS



© ИЦиГ СО РАН

Thank you for your attention!

感谢您的关注！

Спасибо за внимание!

© ИЦиГ СО РАН

GWAS

© ИЦиГ СО РАН

© ИЦиГ СО РАН