



# Теория селекции

## Лекция 4. GWAS

### Полногеномный анализ ассоциаций

*В.М.Ефимов, д.б.н.*

Институт цитологии и генетики СО РАН



Полногеномный анализ ассоциаций (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.



# 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 удастся 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



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**.

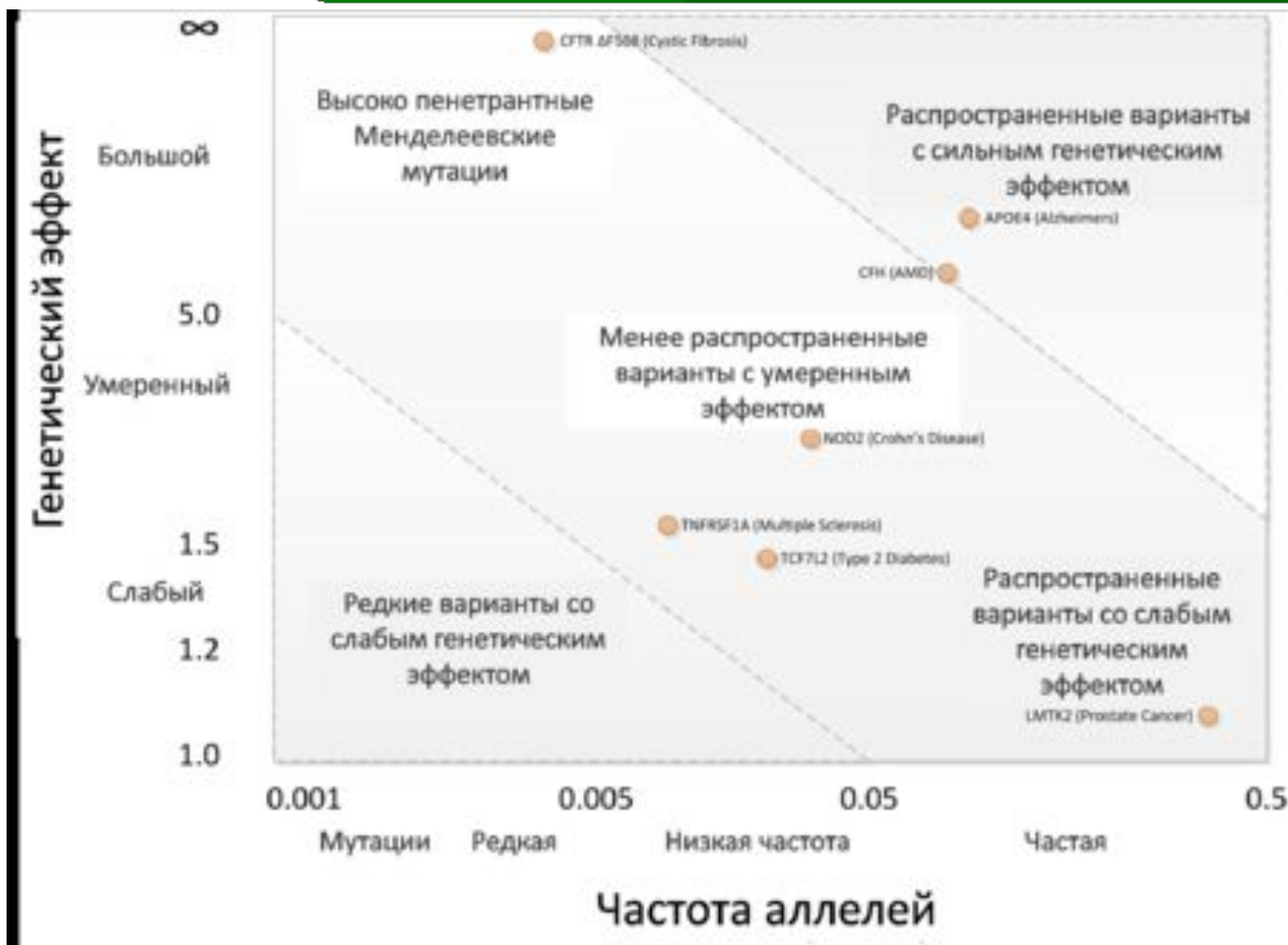


В основе поиска полногеномных ассоциаций как правило лежит сравнение геномов двух групп людей: носителей **исследуемого** фенотипа (заболевания) и **контрольной** группы. Для всех индивидуумов производится генотипирование для большинства известных однонуклеотидных полиморфизмов (SNP). Далее, для каждого 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.

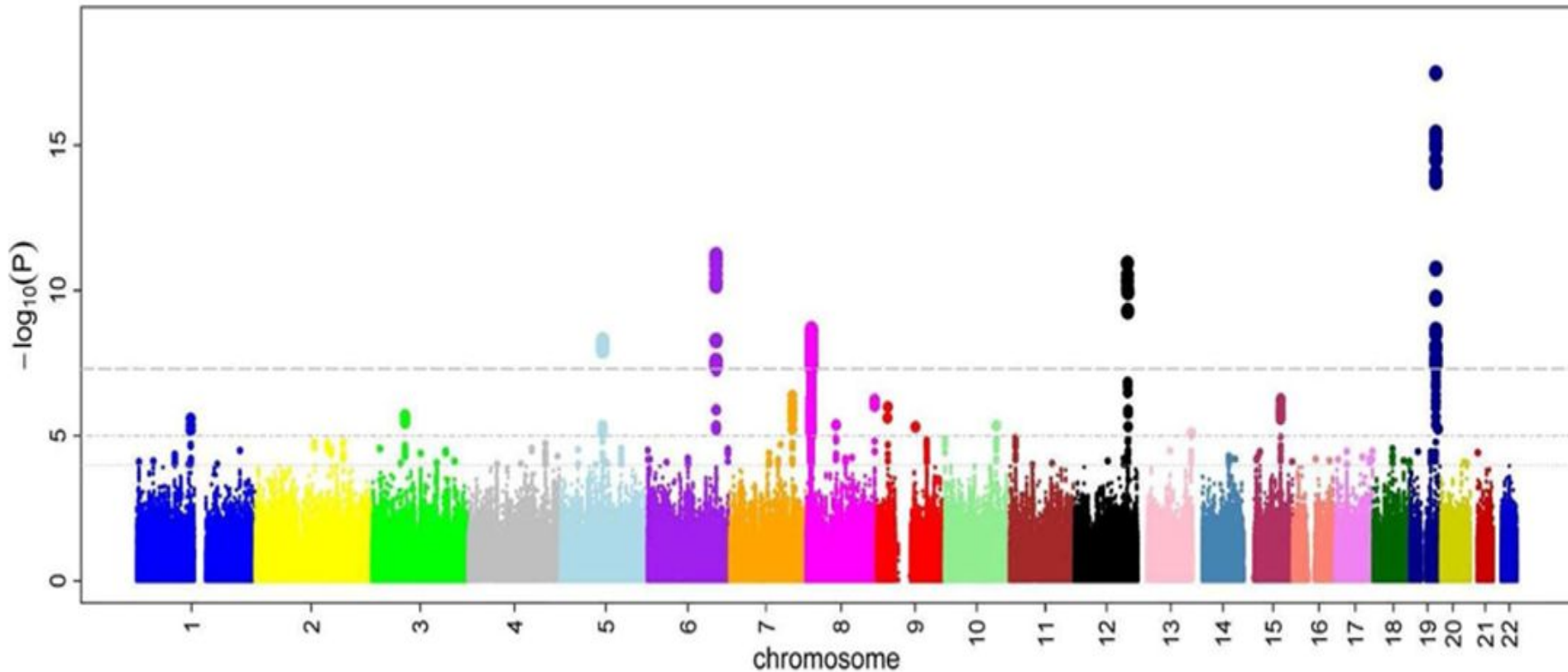


# GWAS





# GWAS



Верхняя прямая – критерий Бонферрони, нижняя – FDR



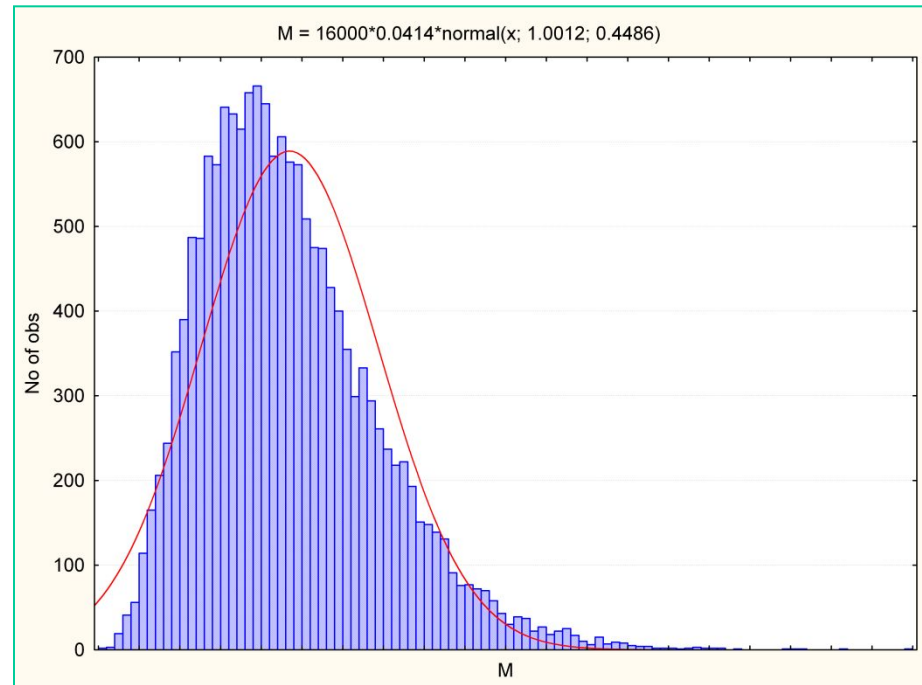
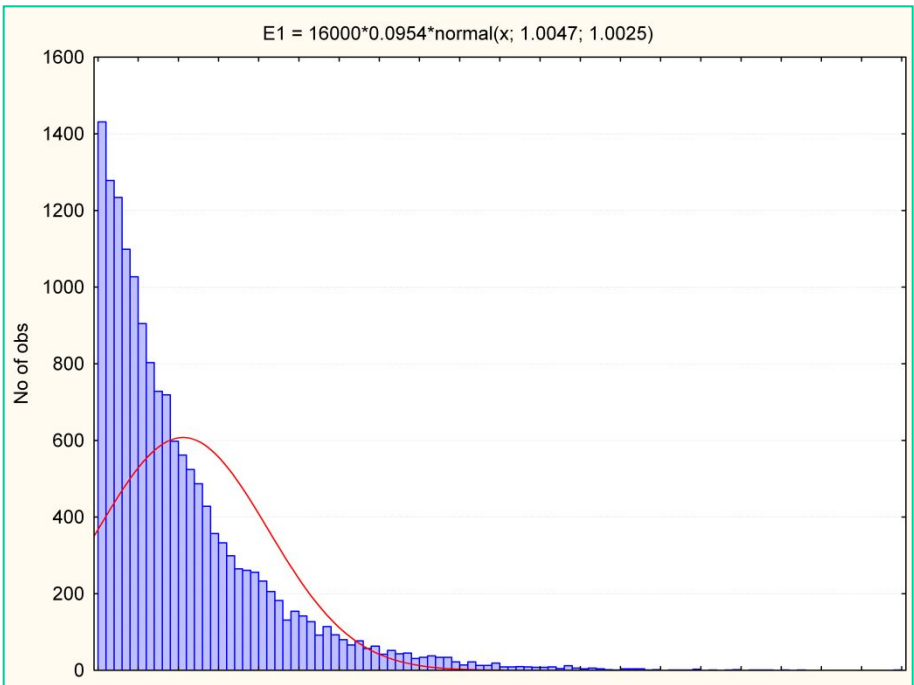
Если нам нужно сравнить **средние** двух заданных групп, обычно используется  $t$ -критерий Стьюдента или Уэлша. Если число групп больше двух, используется  $F$ -критерий Фишера (ANOVA).

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





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**.



Слева: **ЭКСПОНЕНЦИАЛЬНОЕ** распределение ( $\lambda=1$ ). Справа: распределение выборочных **средних** из него объема  $n=10$ . Красная кривая: аппроксимация **НОРМАЛЬНЫМ** распределением.

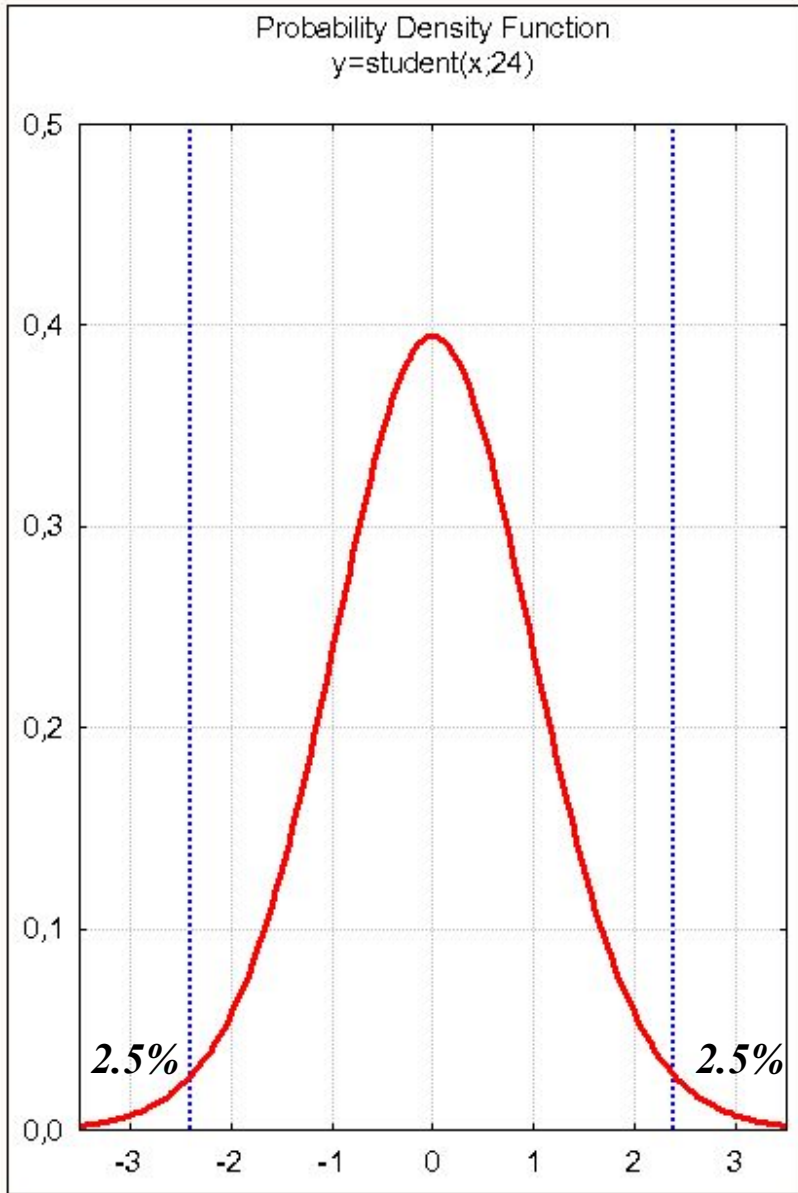


**Нулевая** гипотеза заключается в том, что различия между выборками являются **случайными** и все выборки на самом деле взяты из **одного и того же** генерального распределения.

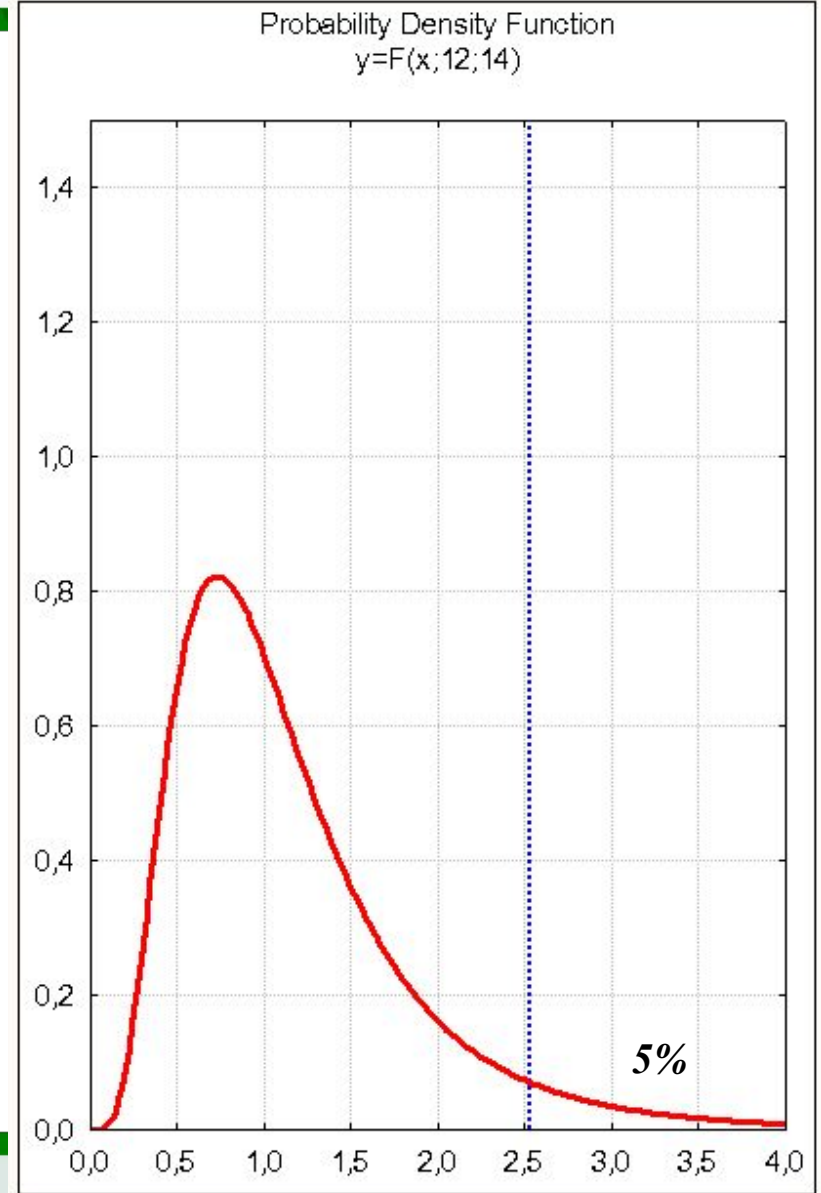
Если значения рассчитанного критерия превышают табличные значения для заданного уровня значимости  $\alpha$  (**0.05; 0.01; 0.001**), то мы считаем, что нулевая гипотеза **не подтверждается**. Это ошибка **первого** рода.



## Распределение t-критерия



## Распределение F-критерия





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  $\alpha$  ни один из  $N$  рассчитанных критериев не превысит табличный уровень, то мы должны выбрать табличный уровень, равный  $\alpha/N$ . Это критерий **Бонферрони**.



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.



## FDR (false discovery rate)



Критерий **FDR** заключается в следующем. 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.

# FDR (false discovery rate)

( $N=10^6$ )

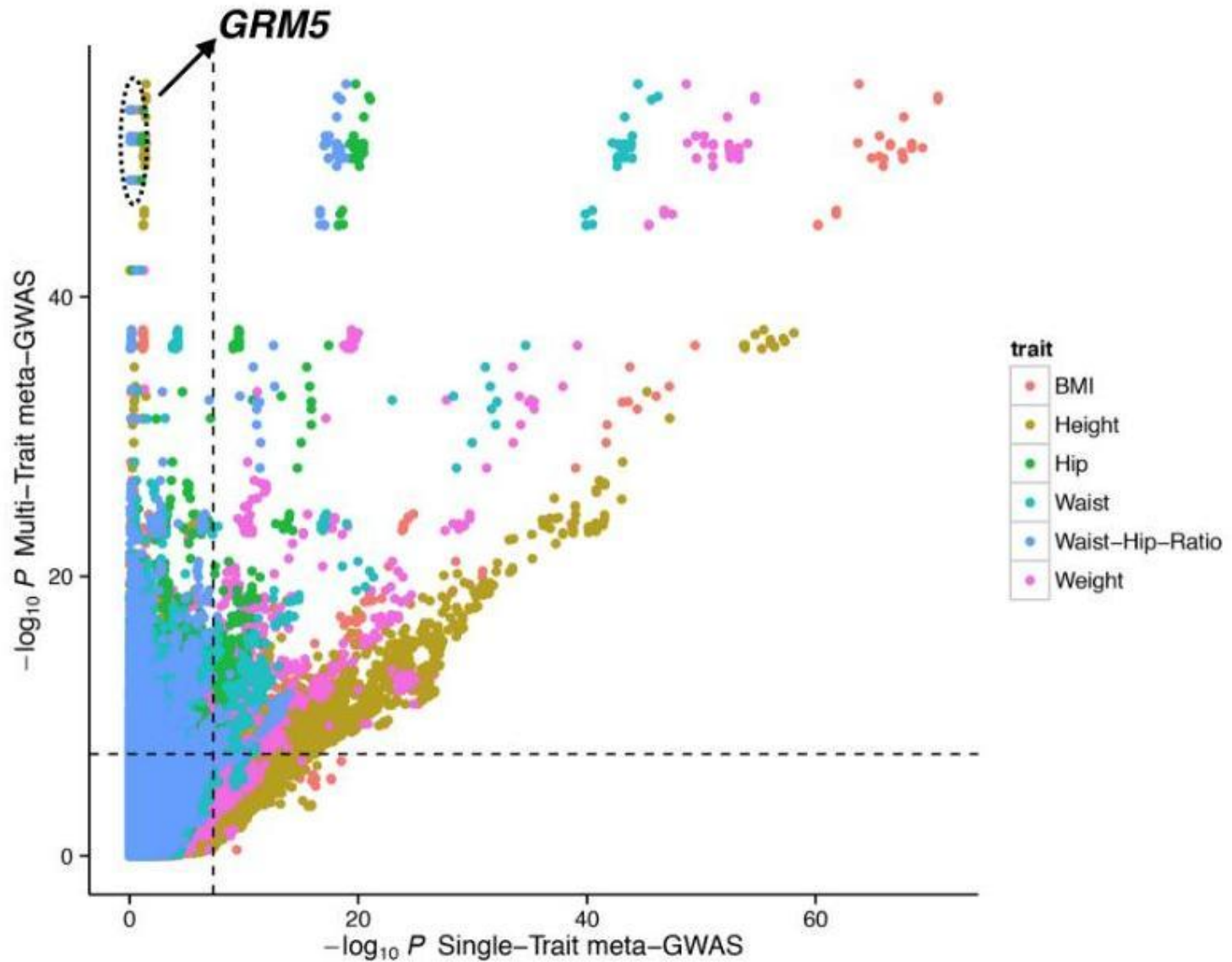


<b>t</b>	<b>p</b>	<b>lg(p)</b>	<b>pN</b>	<b>k</b>
<b>-13.358</b>	<b>2.41E-12</b>	<b>-11.62</b>	<b>2.41E-06</b>	<b>1</b>
<b>-12.488</b>	<b>1.36E-11</b>	<b>-10.87</b>	<b>1.36E-05</b>	<b>2</b>
<b>-15.238</b>	<b>1.55E-11</b>	<b>-10.81</b>	<b>1.55E-05</b>	<b>3</b>
<b>-14.264</b>	<b>2.31E-11</b>	<b>-10.64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





# Multi-Trait meta-GWAS





**Thank you for your attention!**

**感谢您的关注！**

**Спасибо за внимание!**





# GWAS



