

ОСНОВЫ БИОСТАТИСТИКИ

Александр Владимирович Рубанович

зав. лаб. экологической генетики ИОГен РАН

rubanovich@vigg.ru

тел. (499) 132-8958

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Выявление ассоциаций «генотип-фенотип»: минимальный набор действий

- Фенотип (на

Кроме этого в обоих случаях можно строить различные регрессионные модели: Зависимая переменная – признак (фенотип), независимыми переменные – генотипы. Например так: A/A - 0, A/T - 1, T/T - 2

н»)

а

- Фенотип - количественный признак (например: вес, содержание кальция, частота аббераций)



Вычисляем средние значения признака для разных генотипов; значимость по критерию Манна-Уитни

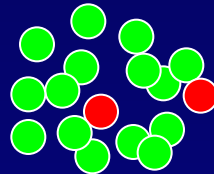
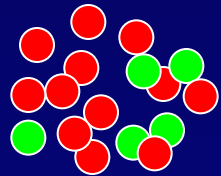
OR – количественная мера предрасположенности (**O**dd **R**atio)

OR – неперенный атрибут «**case-control association study**»
(выявление «генов предрасположенности» к заболеванию
путем сопоставлений частот генотипов у больных и здоровых)

OR показывает во сколько раз повышена вероятность
заболеть для носителя «плохого» генотипа

Группа больных

Контроль (здоровые)



● - генотип,
указывающий на
предрасположенность
к заболеванию

$P_{\text{больные}}$

\gg

$P_{\text{контроль}}$

$$OR = \frac{P_{\text{больные}} (1 - P_{\text{контроль}})}{P_{\text{контроль}} (1 - P_{\text{больные}})}$$

OR > 1 – генотип связан с болезнью

OR = 1 – нет связи между генотипом и болезнью

OR < 1 – протективный генотип

Soft для вычисления OR и проведения метаисследований

WinPepi Portal (2010) - computer programs for epidemiologists



Free!

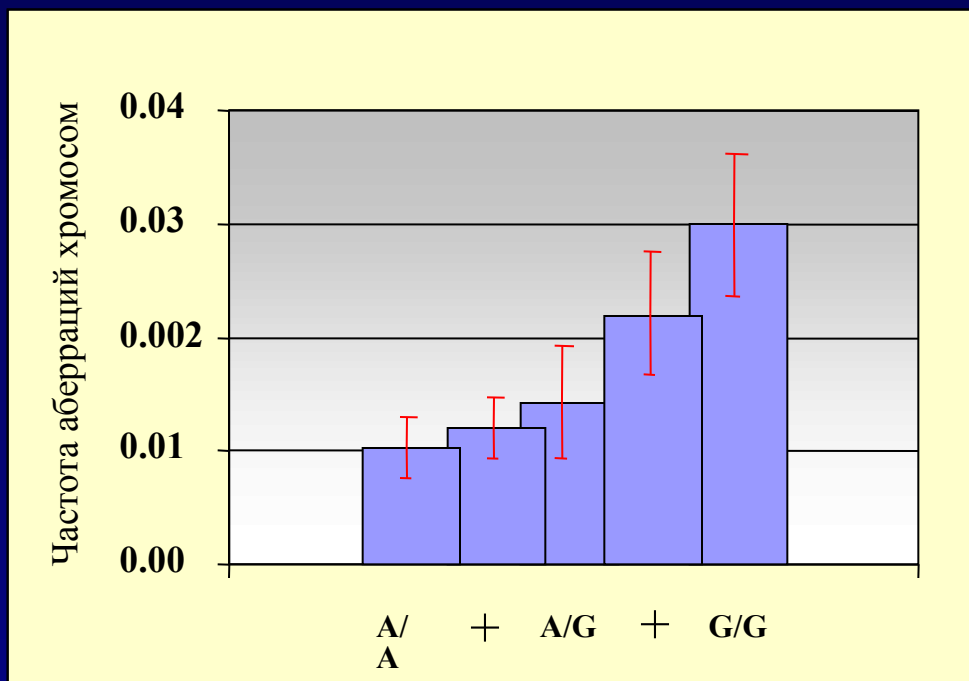


30 дней

Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений признака для носителей различных генотипов. Далее сравнение по непараметрическому тесту

Обычно стараются рассмотреть две группы



Гомозигота по ~~рецессивной~~ мажорному аллелю

дель

Доминантная форма

модель

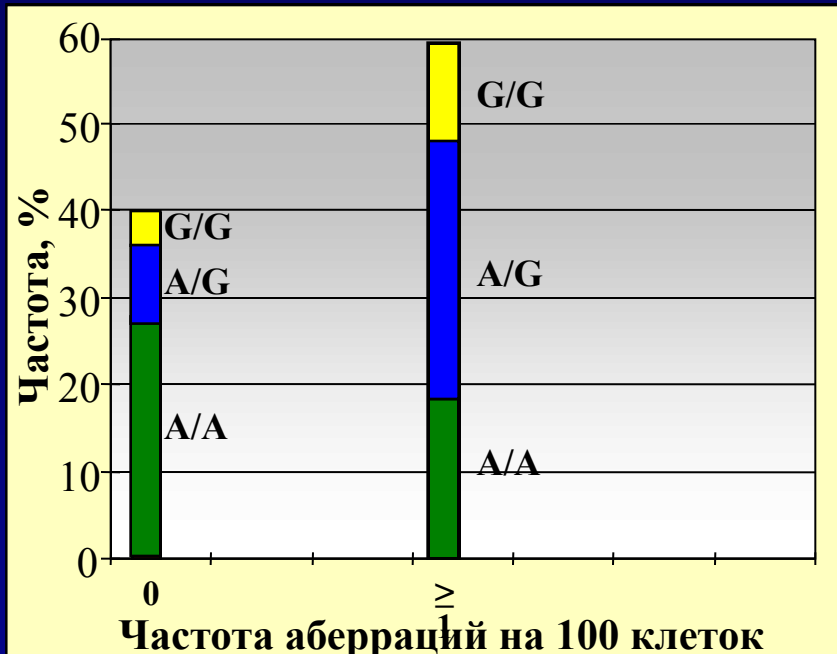
Гомозигота по минорному аллелю

Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений и дисперсий для каждого генотипа. Далее сравнение средних значений (не по Стрессу)

Далее вычисляется OR и значимость по точному критерию Фишера. В данном примере риск возникновения aberrаций у носителей минорного аллеля G равен $OR=2,1$ и $p=0,015$

- Сравнение частот генотипов для групп с низким (или высоким) значением признака



Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений признака для носителей различных генотипов. Далее сравнение по непараметрическому тесту (не по Стьюденту!)

- С зависимой переменной (признаком) (p) и независимыми переменными (генотипами) (x_i) . Например так: А/А - 0, А/Т - 1, Т/Т - 2

- Логистическая и пуассоновская регрессии

$$p = \frac{1}{1 + e^{a_1 x_1 + \dots + a_n x_n}}$$

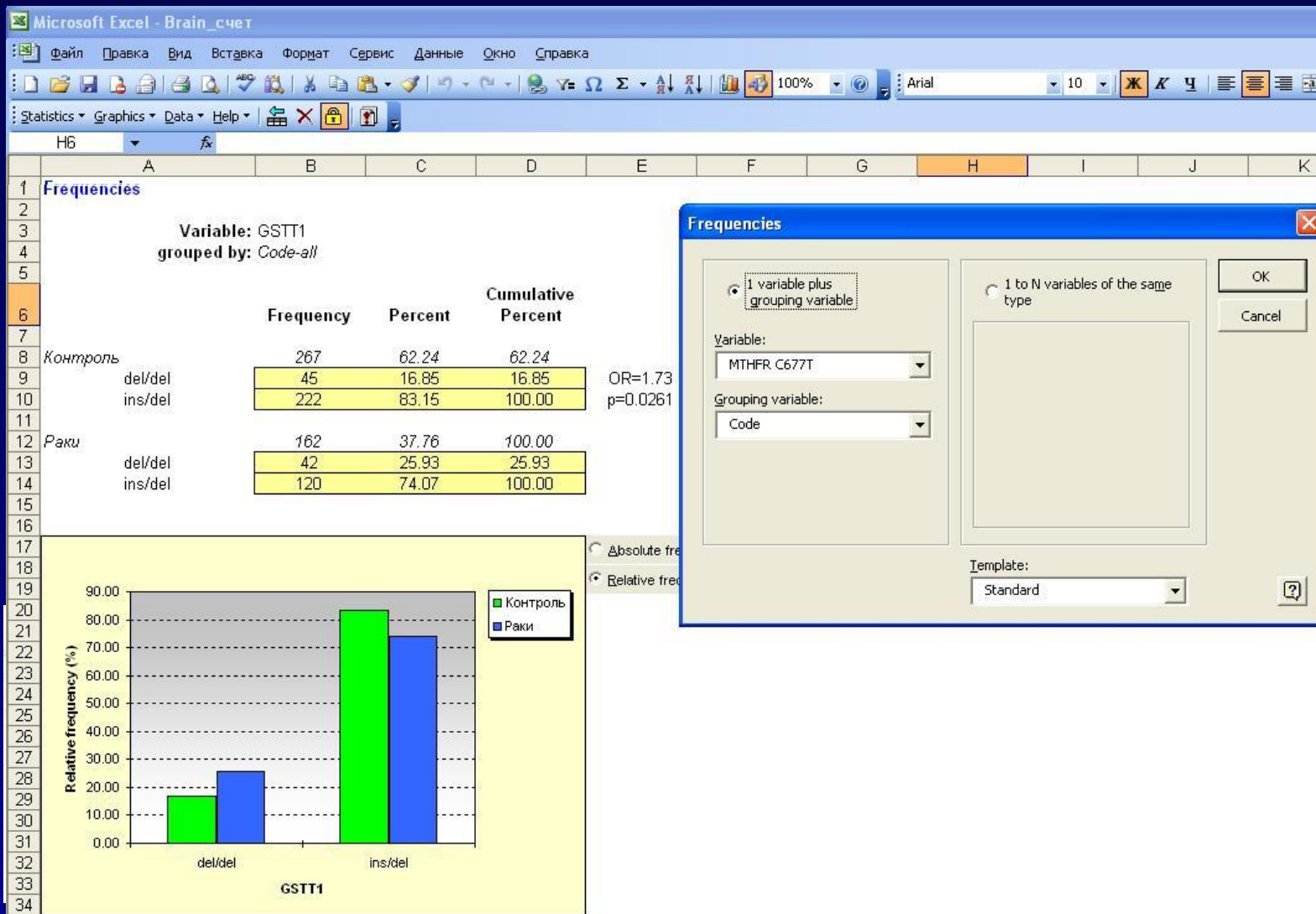
p – частота aberrаций
 x_i – генотип i -го локуса
 a_i – коэф. регрессии

Для логистической регрессии $a_i = \ln(\text{OR}_i)$

$$p = e^{a_0 + a_1 x_1 + \dots + a_n x_n}$$

Soft для работы с генотипами и гаплотипами

WinStat for Excel



Free!

Free!

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Чуть-чуть об ошибках статистических тестов

Нулевая
различия

Традиционно биолог ориентирован на контроль
ошибки I рода (через уровень значимости),
т.е. на гарантии отсутствия ложных открытий,

ии
ости

Ошибка I рода (α)

Вероятность отвергнуть правильную нулевую гипотезу =
Вероятность обнаружить различия там, где их нет = **Вероятность совершить фальшивое открытие**



Ошибка II рода (β)

Вероятность принять неправильную нулевую гипотезу =
Вероятность не обнаружить существующие различия =
Вероятность упустить открытие



Мощно
Вероятн
Вероятн

... и при этом мало заботится о возможности
упустить открытие (ошибка II рода)

От чего зависят ошибки статистических тестов?

❑ От размаха реально существующих отличий и разброса данных

❑ От объемов выборок

Ошибка I рода (вероятность фальшивого открытия)

слабо зависит от объемов выборок,
если они сравнимы по величине

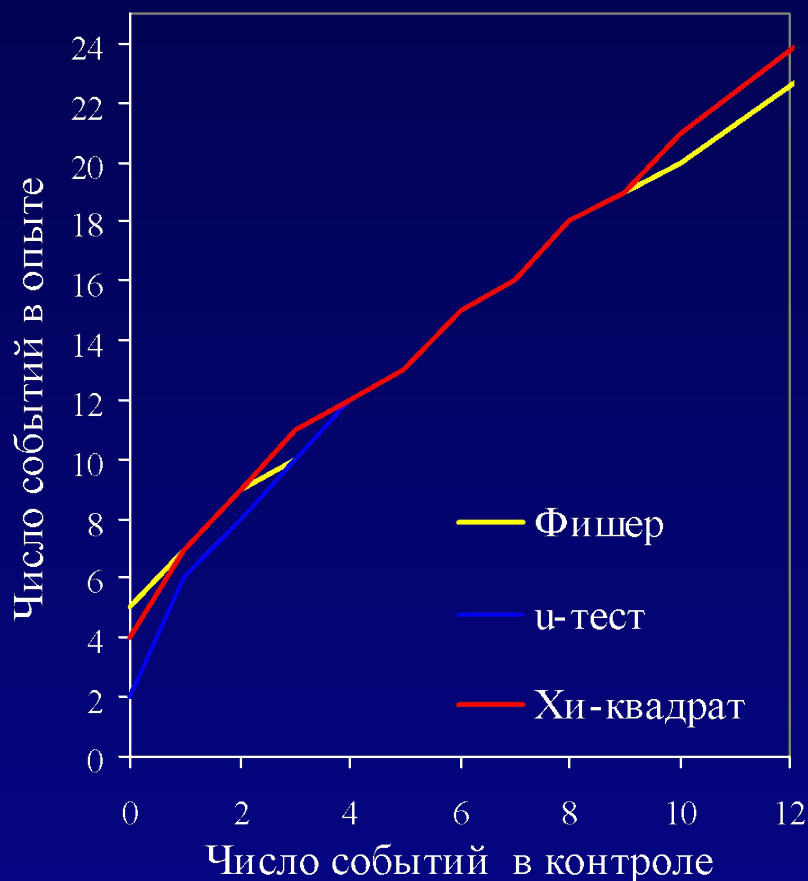
Крайний случай:

«критерий» св. Фомы Неверующего (0033)

Ошибка I рода = 0 \Leftrightarrow Ошибка II рода = 1

Сравнение частот при уровне значимости 0.05

Объемы выборок в опыте и контроле одинаковы



| Число событий в контроле | Минимальное число событий в опыте при значимом отличии от контроля | | |
|--------------------------|--|----------|-------|
| | Стьюдент | χ^2 | Фишер |
| 0 | 2 | 4 | 5 |
| 1 | 6 | 7 | 7 |
| 2 | 8 | 9 | 9 |
| 3 | 10 | 11 | 11 |
| 4 | 12 | 13 | 13 |
| 5 | 14 | 15 | 15 |
| 6 | 15 | 15 | 15 |
| 7 | 16 | 16 | 16 |
| 8 | 18 | 18 | 18 |
| 9 | 19 | 19 | 19 |
| 10 | 21 | 21 | 20 |
| 20 | 35 | 35 | 33 |
| 30 | 47 | 47 | 46 |

больше 5
независимо от объемов выборок
(100 или 1000)

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Проверка однородности материала и вычисление OR для нескольких выборок

□ Индекс рассеяния для биномиальных выборок

Можно ли объединить k независимых выборок и оценить частоту как

$$\bar{p} = \frac{\sum n_i}{\sum N_i}$$

| Объем выборки | Число мутаций | Частота |
|---------------|---------------|---------|
| N_1 | n_1 | p_1 |
| N_2 | n_2 | p_2 |
| | | |
| N_k | n_k | p_k |

Выборки можно объединять, если

$$\frac{\sum_i N_i (p_i - \bar{p})^2}{\bar{p}} < 2k$$

□ Mantel-Haenszel test



Mantel-Haenszel test

Comparison of two proportions or odds

[Back to "Comparison of..." menu](#)

Analyzes any simple 2×2 contingency table.

Check here for equivalence tests. Include missing data in analysis.

The group
For each

A:

B:

Stratified
strata hav

Comparison of two proportions or odds

[Back to "Comparison of..." menu](#)

Proportions (of "Yes"): A, 0.1000 B, 0.2200
If inverse sampling was used,

Exact tests:

Fisher's P:

One-tailed

Two-tailed

Double one-tail

Mid-P:

One-tailed

Two-tailed

Double one-tail

Overall's continuity

One-tailed

Two-tailed

Double one-tail

[New data](#)

Comparison of two proportions or odds

Stratum 2

[Back to "Comparison of..." menu](#)

Proportions (of "Yes"): A, 0.0674 B, 0.1429
If inverse sampling was used,
see results at end of output

Exact tests:

Fisher's P:

One-tailed:

Two-tailed:

Double one-tail

Mid-P:

One-tailed:

Two-tailed:

Double one-tail

Overall's continuity

One-tailed:

Two-tailed:

Double one-tail

[New data](#)

Comparison of two proportions or odds

Strata 1 to 2 combined

[Back to "Comparison of..." menu](#)

Значимость
гетерогенности
выборок

Вычисление OR для
совокупности выборок

Unadjusted odds ratio

Heterogeneity of odds ratios:

chi-sq (DF: 1) = 0.02 P = 0.888

Heterogeneity index (Higgins & Thompson's H):

H = 1.0

[A value above 1.5 suggests notable heterogeneity.]

Proportion of variation attributable to

heterogeneity (Higgins & Thompson's I-squared):

I-squared = 0.0%

Use scroll-bar or <PgDn> or <PgUp> to see other results.

[New data](#)

[Next stratum](#)

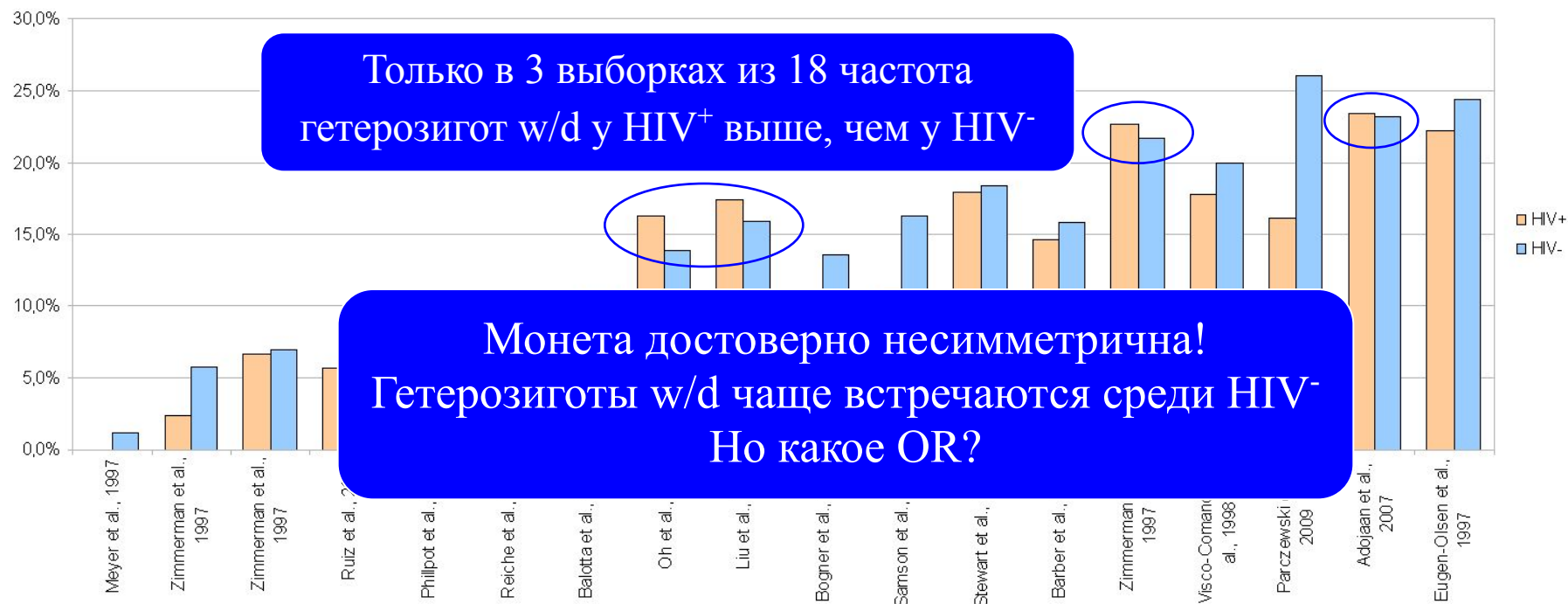
[All strata](#)

[Print](#)

Объединение выборок с неизвестными частотами

Если это принять за 4-ое превышение, то $p=0.015$

частота гетерозигот в выборках HIV+ и HIV-



Если ассоциации нет, то случаи «больше-меньше» должны появляться с вероятностью $\frac{1}{2}$

Вероятность выпадения 3 (и менее) орлов в 18 бросаниях монеты равна

$$p = C_{18}^3 \left(\frac{1}{2}\right)^{18} + C_{18}^2 \left(\frac{1}{2}\right)^{18} + C_{18}^1 \left(\frac{1}{2}\right)^{18} + C_{18}^0 \left(\frac{1}{2}\right)^{18} \approx 0.0038$$

| Мета-анализ | OR | RR = f_+ / f_- | $\Delta f = f_- - f_+$ |
|--|---|------------------|------------------------|
| Mantel-Haenszel оценка | 0.87 (1.15) | 0.887 | 0.016 |
| Unadjusted оценка (по всем данным) | 0.78 | 0.801 | 0.027 |
| 95%-довер. интервал | 0.77 - 0.97 | 0.81 - 0.98 | 0.007-0.023 |
| Значимость гетерогенности (p) | 0.131 | 0.236 | 0.451 |
| Число «null»-статей (OR=1) для ликвидации значимости | 7 | 2 | - |
| Значимость корреляции объемов выборок и эффектов (д.б. > 0.1) | 0.188 (Regression asymmetry test, Egger) 0.211 (Adjusted rank correlation, Begg&Mazumdar): | | |
| Итоговая значимость различий (Fisher's two-tailed) | 0.014 | | |

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Генерируем две одинаково распределенные выборки

Наблюдаем проявление фенотипов в популяции

Частоты минорных аллелей (в среднем 0.1)

| Ген | Больные | Здоровые | OR | p |
|-----|---------|----------|----|-----|
| 1 | 7 | 8 | | |
| 2 | 10 | 2 | | |
| 3 | 17 | 5 | | |
| 4 | 13 | 12 | | |
| 5 | 12 | 11 | | |
| 6 | 7 | 10 | | |
| 7 | 10 | 12 | | |
| 8 | 14 | 9 | | |
| 9 | 14 | 8 | | |
| 10 | 9 | 12 | | |
| 11 | 13 | 9 | | |
| 12 | 9 | 14 | | |
| 13 | 10 | 13 | | |
| 14 | 8 | 8 | | |
| 15 | 14 | 12 | | |
| 16 | 17 | 7 | | |
| 17 | 11 | 13 | | |
| 18 | 10 | 8 | | |
| 19 | 16 | 10 | | |
| 20 | 12 | 8 | | |

4

Должно быть OR=1

Значим

Сразу 3 локуса «ассоциированы» с заболеваемостью!

Как избежать фальшивых открытий?

❑ Правило Карло Бонферрони (1935):

При проведении m независимых статистических тестов значимы только те результаты, для которых

$$p < \frac{0.05}{m}$$

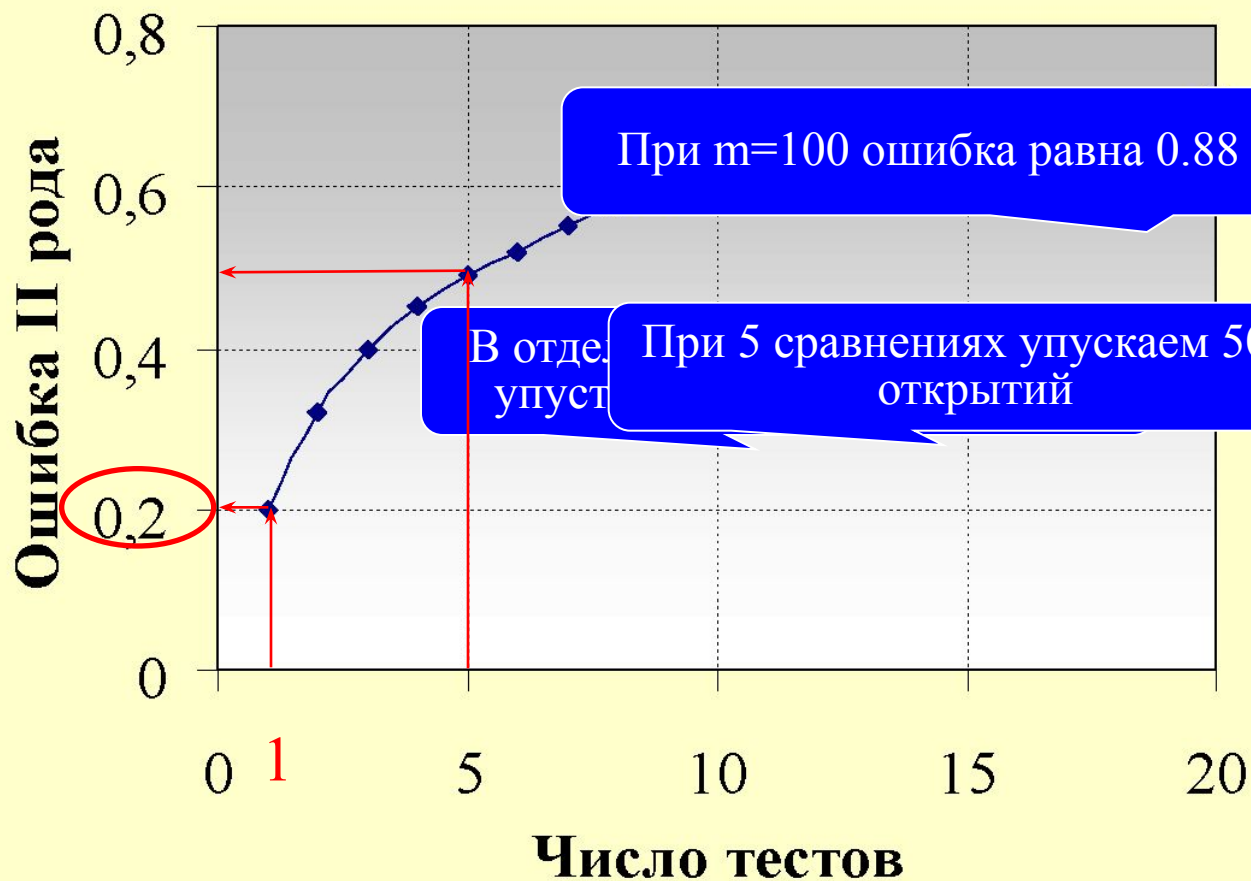
❑ **F**alse **D**iscovery **R**ate control: **FDR** - контроль

❑ **P**ermutation test

(компьютерная перестановка лэйблов «case-control»)

Зависимость ошибки II рода от числа тестов (SNP)

При 100 сравнениях ради того, чтобы гарантировать отсутствие хотя бы одного ложного результата, мы упускаем 88% открытий!



Новый принцип проверки статистических гипотез: FDR-контроль

False Discovery Rate control: Benjamini, Hochberg (1995)

Вероятность хотя бы одного
фальшивого открытия < Уровня значимости
Ошибка I рода < 0.05

Традиционный принцип
заменяется на



105 статей в



Средняя доля фальшивых открытий < Выбранный уровень

$$E\left(\frac{\text{Число неправильно отвергнутых нулевых гипотез}}{\text{Число отвергнутых нулевых гипотез}}\right) < 0.05$$

Пример: множественные сравнения по 10 тестам

| Тест | p_i | Корр Bonf | |
|------|--------|--------------|-------|
| 1 | 0,001 | 0,005 | 0,005 |
| 2 | 0,0055 | 0,005 | 0,010 |
| 3 | | | 0,015 |
| 4 | | | 0,020 |
| 5 | | | 0,025 |
| 6 | | | 0,030 |
| 7 | 0,3 | | |
| 8 | | | |
| 9 | | | |
| 10 | 0,8 | 0,005 | 0,050 |

Располагаем тесты в порядке увеличения p_i

Значимые различия после коррекции по FDR

В первой клетке —
во второй клетке —
втрое больше
и т.д.

Поправка Бонферрони
оставляет значимым лишь
первое сравнение

И это все!!!

Для 6-ого
ЭТОГО значения

НОСТЬ

Что делать, если FDR не помогает?

Permutation tests:

случайные перестановки пометок «case-control»
в компьютерных симуляциях по алгоритму:

- В исходной базе данных делаем случайную перестановку лейблов case-control

Точный тест Фишера – это тоже permutation test,
только реализованный аналитически (p
вычисляется
по формулам комбинаторной теории вероятностей)

- Вычисляем откорректированное p как

$$p' = \frac{\text{Число случаев } (p_{perm} \leq p)}{N}$$

Permutation test применительно к данным об ассоциации заболеваемости с 10 SNP

Переставляем отметки «case-control» 10000 раз. В результате получаем **коррекцию p**

| SNP | Частота минорного аллеля | | OR | p-value | |
|-----|--------------------------|---------------|-----|---------|-------|
| | Case (100) | Control (100) | | 0,0001 | 0,000 |
| 1 | 62 | 26 | 4,6 | 0,0001 | 0,000 |
| 2 | 50 | 30 | 3,7 | 0,009 | 0,010 |
| 3 | 40 | 30 | 2,8 | 0,011 | 0,007 |
| 4 | 30 | 30 | 1,0 | 0,023 | 0,025 |
| 5 | 20 | 30 | 0,9 | 0,071 | 0,109 |
| 6 | 10 | 30 | 0,3 | 0,096 | 0,098 |
| 7 | 10 | 20 | 0,5 | 0,103 | 0,058 |
| 8 | 10 | 10 | 0,2 | 0,120 | 0,067 |
| 9 | 10 | 5 | 0,4 | 0,571 | 0,476 |
| 10 | 10 | 0 | 0,0 | 0,911 | 1,000 |

Но так бывает не всегда

Значимо по Бонферрони

Совсем маленькая программка

```

simNum = 10000
surDif = Table[0, {Length[frCase], Length[frControl]}];
Do[1, RandomPermutation[2, volSample]];
tot = Join[ill, health];
ill = tot[[takeIll, volSample]];
health = tot[[takeH, volSample]];
genCase1 = gen[ill];
genControl1 = gen[health];
xiSq = genCase1[[genControl1]]^2 / (genCase1[[genControl1]] + N);
p1 = 1 - CDF[ChiSquareDistribution[1], xiSq];
surDif = surDif + {1, 1} * StepFunction[p1, {simNum}];
simp = surDif / simNum; N;
    
```

