

# ОСНОВЫ БИОСТАТИСТИКИ

## Александр Владимирович Рубанович

зав. лаб. экологической генетики ИОГен РАН

[rubanovich@vigg.ru](mailto: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

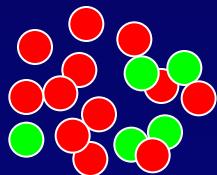
- Фенотип - количественный признак  
(например: вес, содержание кальция, частота aberrаций)
  - Кроме этого в обоих случаях можно строить различные регрессионные модели:  
Зависимая переменная – признак (фенотип), независимыми переменными – генотипы.  
Например так: A/A - 0, A/T - 1, T/T - 2
- Вычисляем средние значения признака для разных генотипов; значимость по критерию Манна-Уитни

# OR – количественная мера предрасположенности (Odd Ratio)

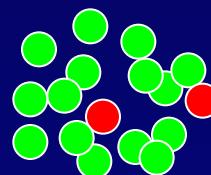
OR – непременный атрибут «case-control association study»  
(выявление «генов предрасположенности» к заболеванию  
путем сопоставлений частот генотипов у больных и здоровых)

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

Группа больных



Контроль (здоровые)



P<sub>больные</sub>

>>

P<sub>контроль</sub>



● - генотип,  
указывающий на  
предрасположенность  
к заболевания

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

OR>1 – генотип связан с болезнью

OR=1 – нет связи между генотипом и болезнью

OR<1 – протективный генотип

# Soft для вычисления OR и проведения математических исследований

WinPepi Portal (2010) - computer programs for epidemiologists

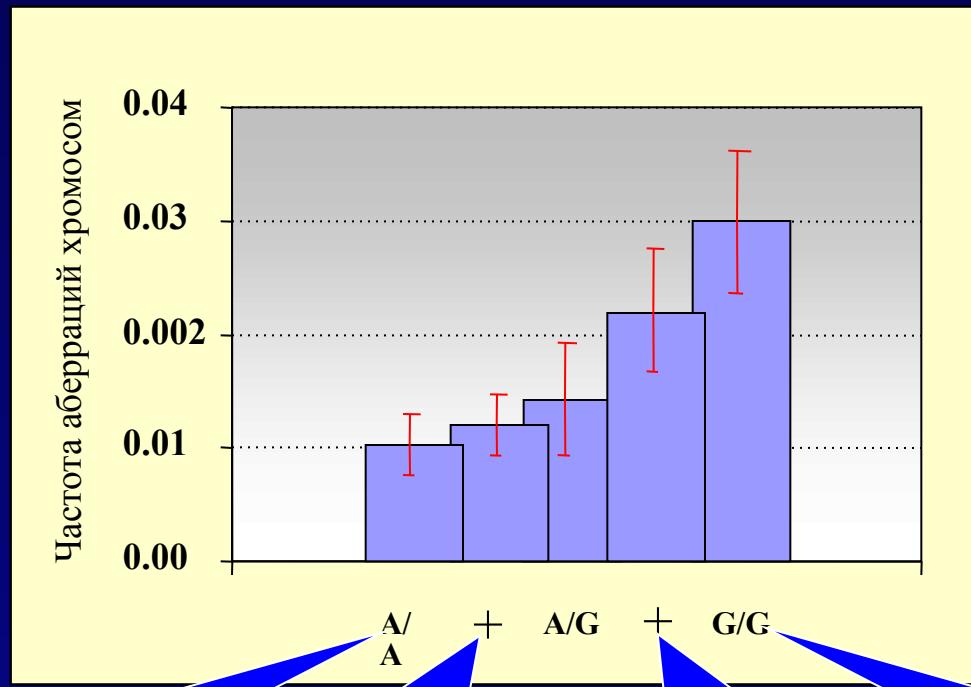


Free!

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  
мажорному аллелю

сивная  
дель

Доминальная  
модель

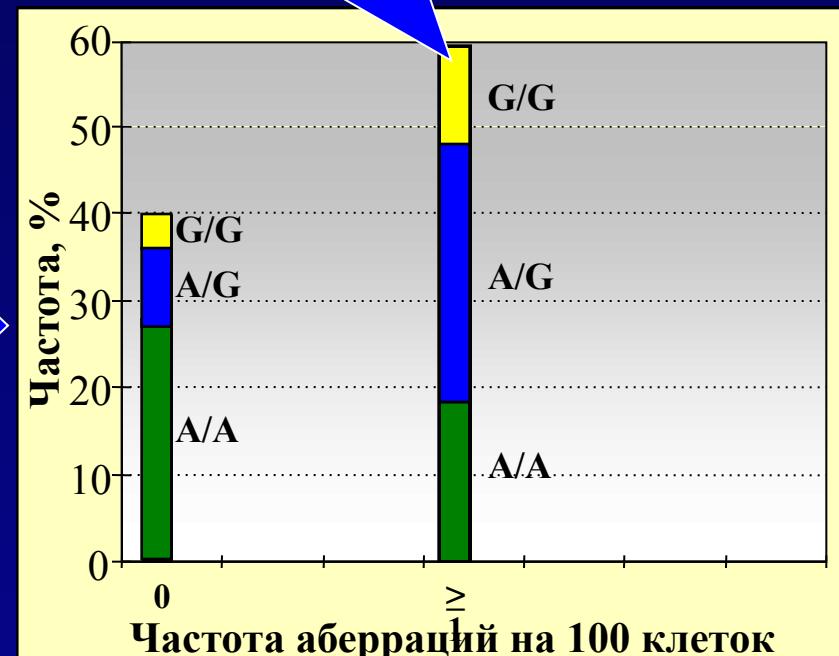
Гомозигота по  
минорному аллелю

# Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений генотипов.  
Далее сравниваются частоты генотипов (не по Стьюарту).

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

- Сравнение частот генотипов для 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!)
- Статистическая модель: зависимость количественного признака от генотипов. Зависимая переменная – признак ( $p$ ), независимые – генотипы ( $x_i$ ). Например так: A/A - 0, A/T - 1, T/T - 2
- Логистическая и пуассоновская регрессии

$$p = \frac{1}{1 + e^{a_1x_1 + \dots + a_nx_n}}$$

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

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

$$p = e^{a_0 + a_1x_1 + \dots + a_nx_n}$$

# Soft для работы с генотипами и гаплотипами

## □ WinStat for Excel

Microsoft Excel - Brain\_счет

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

Statistics Graphics Data Help

H6 f

1 Frequencies

2

3

4 Variable: GSTT1  
grouped by: Code-all

5

6

		Frequency	Percent	Cumulative Percent
Контроль	del/del	267	62.24	62.24
	ins/del	45	16.85	16.85
		222	83.15	100.00
Раки	del/del	162	37.76	100.00
	ins/del	42	25.93	25.93
		120	74.07	100.00

OR=1.73  
p=0.0261

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1010

1011

1012

1013

1014

1015

1016

1017

1018

# 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  
различие

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

## Ошибка I рода ( $\alpha$ )

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



## Ошибка II рода ( $\beta$ )

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



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

# 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

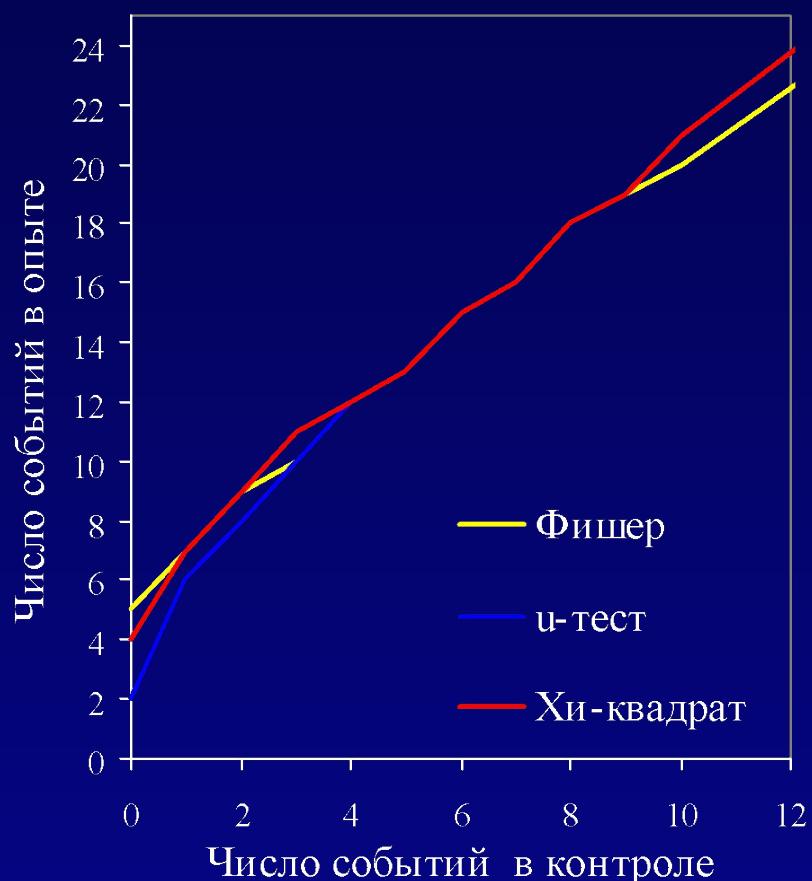
Крайний случай:

«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	$\chi^2$	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

Объем выборки	Число мутаций	Частота
$N_1$	$n_1$	$p_1$
$N_2$	$n_2$	$p_2$
....	....	....
$N_k$	$n_k$	$p_k$

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

Выборки можно объединять, если

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

## □ Mantel-Haenszel test



[Back to "Comparison of..." menu](#)Analyzes any simple  $2 \times 2$  contingency table.

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

## Mantel-Haenszel test

The group  
For each

A:

B:

Stratified  
strata have

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-tailed:  
 Mid-P:  
 One-tailed:  
 Two-tailed:  
 Double one-tailed:  
 Overall's continuity corrected:  
 One-tailed:  
 Two-tailed:  
 Double one-tailed

[New data](#)

## Comparison of two proportions or odds

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

Stratum 2  
Proportions (of "Yes"): A, 0.0674 B, 0.1429  
If inverse sampling was used,

Exact tests:  
 Fisher's P:  
 One-tailed:  
 Two-tailed:  
 Double one-tailed:  
 Mid-P:  
 One-tailed:  
 Two-tailed:  
 Double one-tailed:  
 Overall's continuity corrected:  
 One-tailed:  
 Two-tailed:  
 Double one-tailed

[New data](#)

## Comparison of two proportions or odds

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

Strata 1 to 2 combined  
Zначимость  
гетерогенности  
выборок

Вычисление OR для  
совокупности выборок

Unadjusted odds ratio = 0.727  
Heterogeneity of odds ratio = 0.888  
chi-sq (DF: 1) = 0.000 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 &lt;PgDn&gt; or &lt;PgUp&gt; to see other results.

# Объединение выборок с независимыми оценками

частота гетерозигот в выборках HIV+ и HIV-

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

Только в 3 выборках из 18 частота  
гетерозигот w/d у HIV<sup>+</sup> выше, чем у HIV<sup>-</sup>

Монета достоверно несимметрична!

Гетерозиготы w/d чаще встречаются среди HIV-  
Но какое OR?



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

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

Ma

Протективное действие гетерозиготы  
w/d CCR5 достоверно, но не велико: OR=1.15

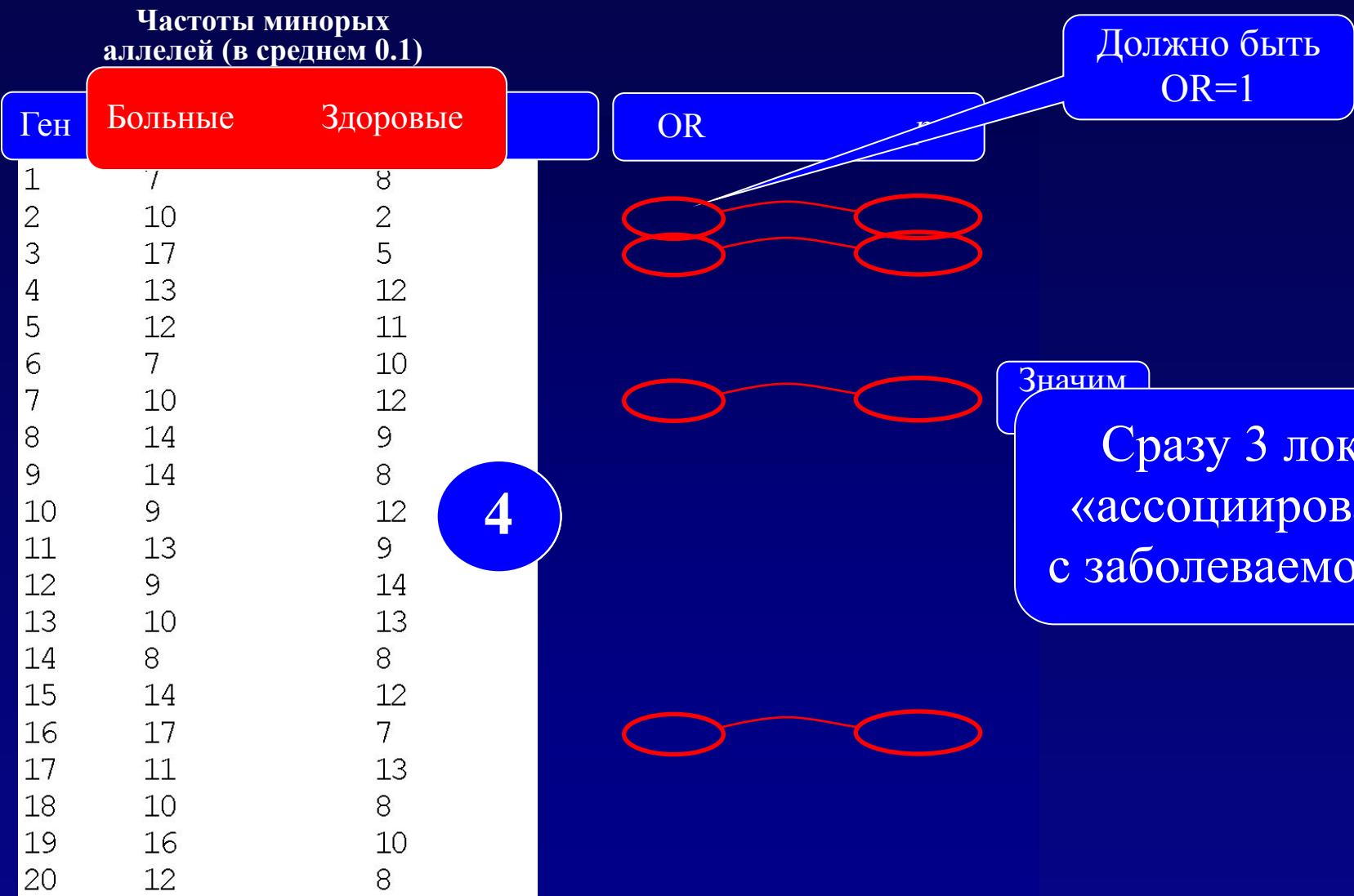
ты

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 однотипные выборки Наблюдаем спонтанное явление фальшивых ассоциаций (0.05%)



# Как избежать фальшивых открытий?

- Правило Карло Бонферрони (1935):

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

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

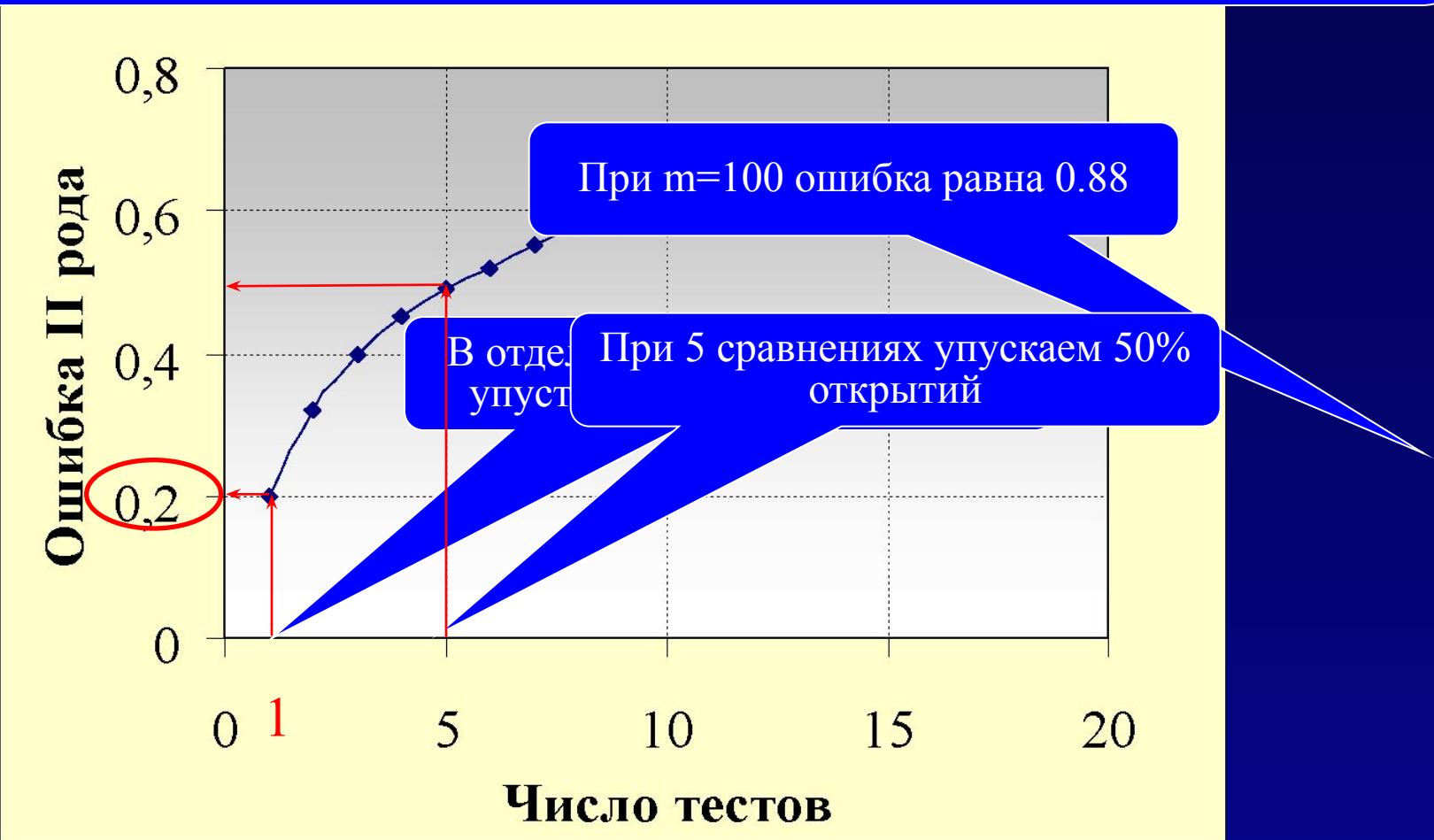
- False Discovery Rate control: FDR - контроль

- Permutation 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% открытых!

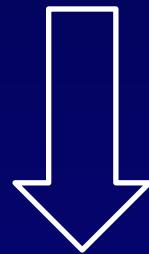


# Новый принцип проверки статистических гипотез: FDR-контроль

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. Вел.
1	0,001	0,005
2	0,0055	0,0055
3		0,015
4		0,015
5		0,015
6		0,030
7	0,3	
8		
9		
10	0,8	0,005

Располагаем тесты в порядке увеличения

Значимые различия после коррекции по FDR

В первой кл.

во второй кл.

втрое больше  
и т.д....

Коррекция Бонферрони оставляет значимым лишь первое сравнение

И это все!!!

Для 6-ого этого значения

значимость

# 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 (р  
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		
	Case (100)	Control (100)			Но так бывает не всегда
1	62	26	4,6	0,0001	0,000
2			2,7	0,009	0,010
3			2,8	0,011	0,007
			9	0,023	0,025
				0,071	0,109
				0,096	0,098
				0,103	0,058
				0,120	0,067
				0,571	0,476
				0,911	1,000

Совсем маленькая программка

```
simNm = 10000;
sumDif = Table[0, {Length[frCases]}];
Do[1] = RandomPermutation[200];
tot = Join[1, health];
ill = tot[[take[1, WolframSample]]];
health = tot[[take[1, WolframSample]]];
genCase1 = one.ill;
genControl1 = one.health;
chiSq = genCase1[[genControl1]] - genCase1;
genControl1 = N;
p1 = 1 - CDF[chiSquareDistribution[1, chiSq]];
sumDif = sumDif + StepFunction[p1, simNm];
simp = sumDif simNm/N;
```

