

Statistical Considerations for Design and Analysis of –Omics Studies

Aaron Isaacs, Ph.D.

"There are three kinds of lies: lies,
damned lies, and statistics."

- Samuel Clemens (a.k.a Mark Twain) who attributed it to Benjamin Disraeli

"Lies, Damned Lies, and Medical Science"

- Title of a 2010 article in The Atlantic magazine

Reproducibility crisis?

- Does it exist?
- It certainly does (or did) in genetics!

Statistics (A Brief Overview)

A word about R

- Free and open source
- Vast array of statistical tests available in the base package
- Huge number of specialized packages
- Scriptable – automate complicated pipelines

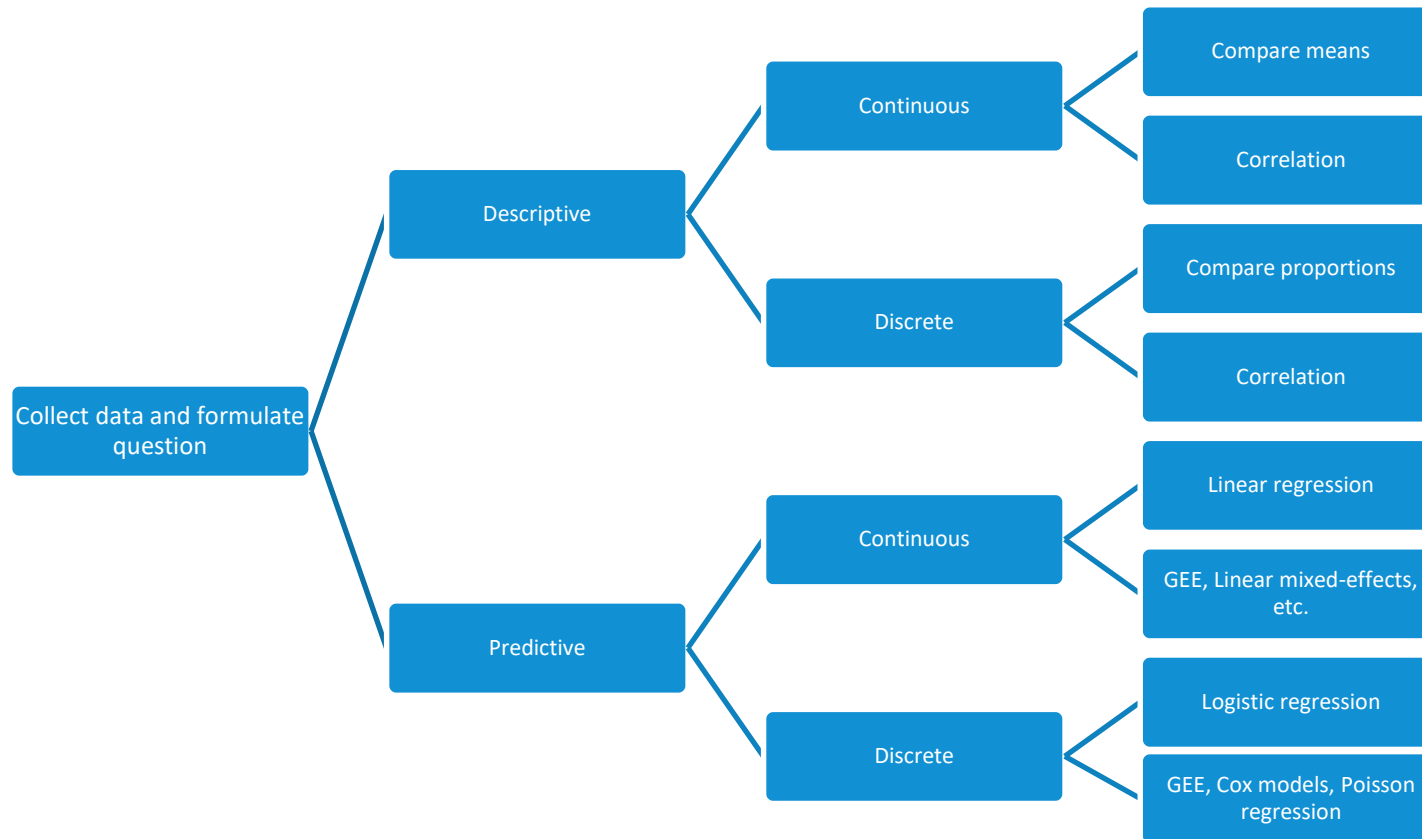
Why do we have to do statistics? (Sigh...)

- Formalizes a framework for the analysis and interpretation of scientific data
- Enhances the reproducibility of studies
- Establishes standards for reporting results
- (Helps) free researchers from personal biases

Classical or Bayesian?

- Classical (“frequentist”) analysis
 - Evaluate the likelihood that a finding in a study occurs by chance
- Bayesian analysis
 - Evaluate the likelihood that a particular hypothesis is correct given data collected in the study

Decisions, decisions, decisions...



ETC.!!!

Descriptive or predictive?

- Descriptive – describe the data collected
 - Measures of central tendency (mean, median), variance, correlation, distribution
- Predictive – make inferences based on the observations in the data
 - Regression (linear, logistic), survival (Cox)

What kind of numbers?

- Continuous – range of real numbers (such as cholesterol levels, QT interval on the ECG, age, etc.)
- Discrete
 - Dichotomous – two values (yes/no, high/low, diseased/healthy, etc.)
 - Categorical – multiple (unordered) values (USA/Netherlands/Germany, aspirin/vitamin K agonist/coumarin)
 - Ordinal – ordered (first/second/third, never/former/current smoker)

Categorize continuous traits?

- Clinical cut-off separating “diseased” and “healthy”
- Prior knowledge demonstrating that different strata face different risk (i.e. “high/medium/low”)
- Take “tails of the distribution” – may increase power

Parametric or non-parametric?

- Parametric – depends on validity of assumptions (such as normality, equal variances, etc.)
- Non-parametric – free from assumptions
- Parametric is typically more powerful if assumptions hold, whereas you can't go wrong with non-parametric

Is it normally distributed?

- A key assumption is often normality
- Normality can be somewhat assessed visually (the “bell” curve)
- Statistical test for normality (generally very sensitive in the case of large samples)
 - Shapiro-Wilk test
 - One sample Kolmogorov-Smirnov test

Comparing a group mean to a known value

- Parametric - one-sample t-test
- Non-parametric - one-sample Wilcoxon signed rank test
 - “According to the Centraal Bureau van Statistiek, middle-aged Dutch have an average HDL level of 53 mg/dL.”
 - “In our previous study of diabetes, the sample had a mean HDL level of 49 mg/dL.”

One sample t-test

```
R Console
> t.test(diabetes$hdl,mu=53)

One Sample t-test

data: diabetes$hdl
t = -2.9672, df = 401, p-value = 0.003185
alternative hypothesis: true mean is not equal to 53
95 percent confidence interval:
 48.75267 52.13788
sample estimates:
mean of x
 50.44527

> |
```

```
R Console
> t.test(diabetes$hdl,mu=49)

One Sample t-test

data: diabetes$hdl
t = 1.6786, df = 401, p-value = 0.094
alternative hypothesis: true mean is not equal to 49
95 percent confidence interval:
 48.75267 52.13788
sample estimates:
mean of x
 50.44527

> |
```

One-sample Wilcoxon signed rank test

```
Save workspace
R Console
> wilcox.test(diabetes$hdl,mu=53)

Wilcoxon signed rank test with continuity correction

data: diabetes$hdl
V = 28384, p-value = 8.281e-07
alternative hypothesis: true location is not equal to 53

> |
```

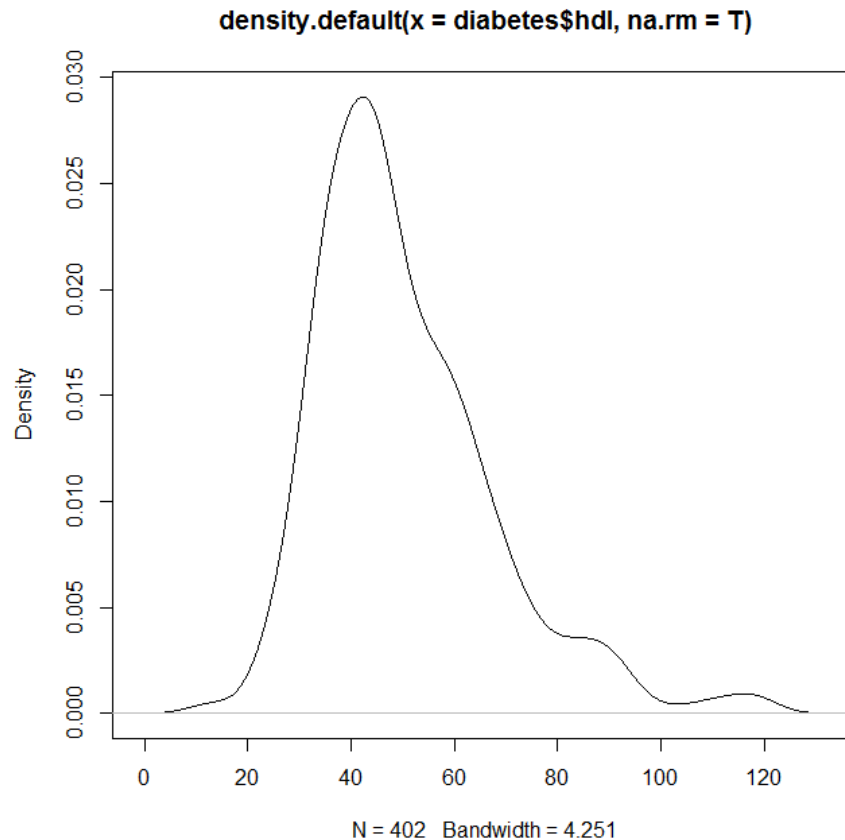
```
R Console
> wilcox.test(diabetes$hdl,mu=49)

Wilcoxon signed rank test with continuity correction

data: diabetes$hdl
V = 38282, p-value = 0.7171
alternative hypothesis: true location is not equal to 49

> |
```


Normality of HDL?



Shapiro-Wilk normality test

data: diabetes\$hdl

W = 0.92357, p-value = 1.918e-13

Comparing two group means

- Parametric – two sample t-test: to compare means between two groups
- Non-parametric – Mann-Whitney u-test
 - “Are total cholesterol levels higher in diabetic patients?”

Comparing two group means

```
R Console
> diabetes$t2d[diabetes$glyhb<=7]<-0
> diabetes$t2d[diabetes$glyhb>7]<-1
> t.test(diabetes$chol~diabetes$t2d)

Welch Two Sample t-test

data: diabetes$chol by diabetes$t2d
t = -3.2996, df = 70.828, p-value = 0.001518
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -40.451620 -9.976344
sample estimates:
mean in group 0 mean in group 1
      203.386      228.600

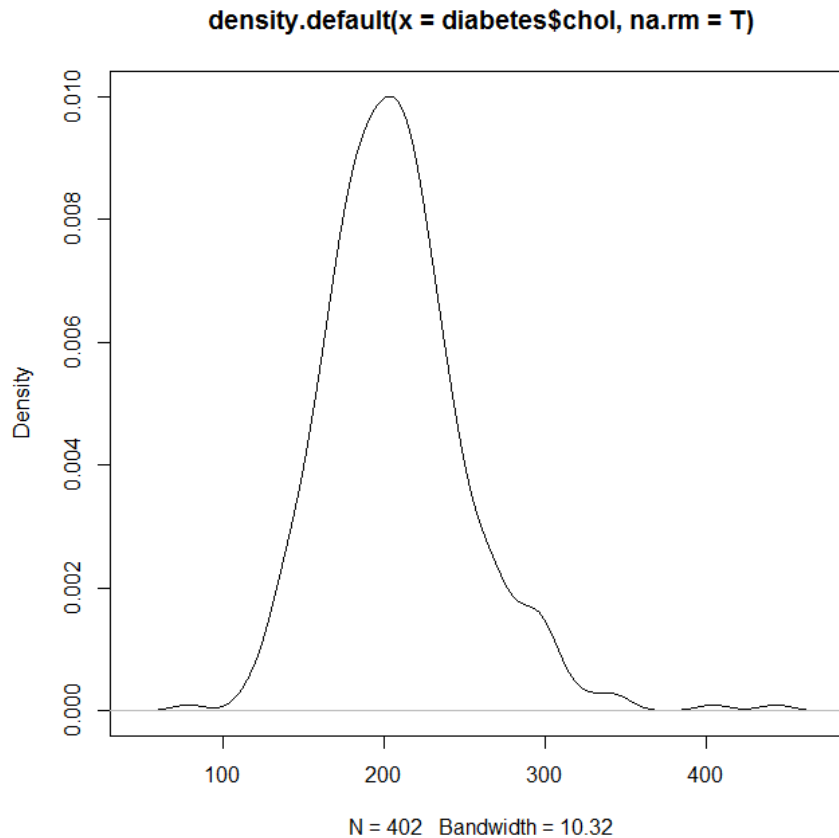
> wilcox.test(diabetes$chol~diabetes$t2d)

Wilcoxon rank sum test with continuity correction

data: diabetes$chol by diabetes$t2d
W = 6920.5, p-value = 0.0002314
alternative hypothesis: true location shift is not equal to 0

> |
```

Normality of cholesterol?



Shapiro-Wilk normality test

data: diabetes\$chol

W = 0.95939, p-value = 4.296e-09

Comparing paired means

- Measurements from the same individuals at two time points
- Parametric – paired sample t-test
- Non-parametric – Wilcoxon signed-rank test
 - “Did blood pressure values change between measurements?”

Look at normality first this time...

```
R Console

> shapiro.test(diabetes$bp.1s)

      Shapiro-Wilk normality test

data:  diabetes$bp.1s
W = 0.93766, p-value = 7.307e-12

> shapiro.test(diabetes$bp.1d)

      Shapiro-Wilk normality test

data:  diabetes$bp.1d
W = 0.99043, p-value = 0.01092

> shapiro.test(diabetes$bp.2s)

      Shapiro-Wilk normality test

data:  diabetes$bp.2s
W = 0.92742, p-value = 1.291e-06

> shapiro.test(diabetes$bp.2d)

      Shapiro-Wilk normality test

data:  diabetes$bp.2d
W = 0.98891, p-value = 0.324
|
```

Comparing paired means

```
R Console

> t.test(diabetes$bp.1s,diabetes$bp.2s,paired=T)

      Paired t-test

data:  diabetes$bp.1s and diabetes$bp.2s
t = 2.581, df = 140, p-value = 0.01088
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5559555 4.1958176
sample estimates:
mean of the differences
      2.375887

> wilcox.test(diabetes$bp.1s,diabetes$bp.2s,paired=T)

      Wilcoxon signed rank test with continuity correction

data:  diabetes$bp.1s and diabetes$bp.2s
V = 4998.5, p-value = 0.0003334
alternative hypothesis: true location shift is not equal to 0

> t.test(diabetes$bp.1d,diabetes$bp.2d,paired=T)

      Paired t-test

data:  diabetes$bp.1d and diabetes$bp.2d
t = 2.8855, df = 140, p-value = 0.004528
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6140184 3.2866908
sample estimates:
mean of the differences
      1.950355

> wilcox.test(diabetes$bp.1d,diabetes$bp.2d,paired=T)

      Wilcoxon signed rank test with continuity correction

data:  diabetes$bp.1d and diabetes$bp.2d
V = 4219.5, p-value = 0.01328
alternative hypothesis: true location shift is not equal to 0
```

Comparing > two group means

- Parametric – analysis of variance (ANOVA)
 - Not particularly well implemented in R
 - Doesn't identify *which* group is different; need pairwise tests for that (multiple comparisons)
- Non-parametric – Kruskal-Wallis test
 - Doesn't identify *which* group is different; need pairwise tests for that (multiple comparisons)

Comparing frequencies

- χ^2 test – 2 x 2 contingency table (note: this is identical to the test comparing two proportions)
- Fisher's exact test
 - Particularly useful for tables with small cell counts (typically meant as ≤ 5)
- 2 x k tables (and larger)

Comparing frequencies

```
R Console
> table(diabetes$t2d, diabetes$gender)

  male female
0  136    194
1   26     34
> chisq.test(table(diabetes$t2d, diabetes$gender))

Pearson's Chi-squared test with Yates' continuity correction

data: table(diabetes$t2d, diabetes$gender)
X-squared = 0.026997, df = 1, p-value = 0.8695
> fisher.test(table(diabetes$t2d, diabetes$gender))

Fisher's Exact Test for Count Data

data: table(diabetes$t2d, diabetes$gender)
p-value = 0.7772
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5078759 1.6704854
sample estimates:
odds ratio
 0.916939
> (136*34)/(194*26)
[1] 0.9167328
```

Confounding

- Confounding – “a variable that influences both the dependent variable and independent variable causing a spurious association”
- Genetics – population stratification: an allele frequency and disease frequency differ in two comingled populations

Population stratification

Population 1

	A+	A-	
D+	1200	300	1500
D-	2800	700	3500
	4000	1000	5000

$$f(A) = 0.8$$

$$f(D) = 0.3$$

$$N = 5000$$

$$OR = (1200 \cdot 700) / (300 \cdot 2800) = 1$$

$$\chi^2 = 0$$

$$P = 1$$

Population 2

	A+	A-	
D+	150	350	500
D-	1350	3150	4500
	1500	3500	5000

$$f(A) = 0.3$$

$$f(D) = 0.1$$

$$N = 5000$$

$$OR = (150 \cdot 3150) / (350 \cdot 1350) = 1$$

$$\chi^2 = 0$$

$$P = 1$$



	A+	A-	
D+	1200+150 1350	300 + 350 650	2000
D-	2800 + 1350 4150	700 + 3150 3850	8000
	5500	4500	10000

$$f(A) = 0.55$$

$$f(D) = 0.2$$

$$N = 10000$$

$$OR = (1350 \cdot 3850) / (650 \cdot 4150) = 1.93$$

$$\chi^2 = 157.828$$

$$P = 3.4 \times 10^{-36}$$

χ^2 test

	A+	A-	
D+	1350	650	2000
D-	4150	3850	8000
	5500	4500	10000

Expected:

$$A+/D+ = (2000/10000) * (5500/10000) * 10000 = 1100$$

$$A-/D+ = (2000/10000) * (4500/10000) * 10000 = 900$$

$$A+/D- = (8000/10000) * (5500/10000) * 10000 = 4400$$

$$A-/D- = (8000/10000) * (4500/10000) * 10000 = 3600$$

(Obs - Exp)²/Exp:

$$A+/D+ = (1350-1100)^2/1100 = 56.82$$

$$A-/D+ = (650-900)^2/900 = 69.44$$

$$A+/D- = (4150-4400)^2/4400 = 14.20$$

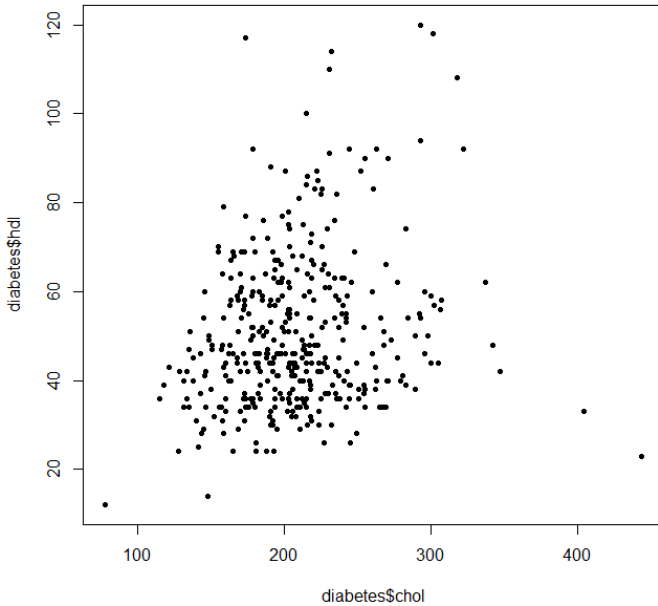
$$A-/D- = (3850-3600)^2/3600 = 17.36$$

$$\text{SUM} = 56.82 + 69.44 + 14.20 + 17.36 = 157.82$$

Correlation

- Parametric – Pearson's correlation
- Non-parametric – Spearman's correlation

Correlation



```
R Console Copy
> cor.test(diabetes$chol,diabetes$hdl)

Pearson's product-moment correlation

data: diabetes$chol and diabetes$hdl
t = 3.7983, df = 400, p-value = 0.0001683
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09042379 0.27929136
sample estimates:
      cor 
0.1865809

> cor.test(diabetes$chol,diabetes$hdl,method="s")

Spearman's rank correlation rho

data: diabetes$chol and diabetes$hdl
S = 9249200, p-value = 0.003401
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho 
0.1457572

Warning message:
In cor.test.default(diabetes$chol, diabetes$hdl, method = "s") :
  Cannot compute exact p-value with ties

> |
```

Linear regression

- Fits linear trends to individual variables and combines them
- $Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$

Linear regression

```
R Console Save workspace
> summary(lm(glyhb~age+gender+waist,data=diabetes))

Call:
lm(formula = glyhb ~ age + gender + waist, data = diabetes)

Residuals:
    Min       1Q   Median       3Q      Max
-3.3795 -1.1404 -0.4930  0.2881 10.3676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.043821   0.739574   1.411 0.158942
age           0.041907   0.006525   6.423 3.96e-10 ***
genderfemale -0.149831   0.215419  -0.696 0.487142
waist         0.070786   0.018588   3.808 0.000163 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.08 on 384 degrees of freedom
(15 observations deleted due to missingness)
Multiple R-squared:  0.1465,    Adjusted R-squared:  0.1398
F-statistic: 21.96 on 3 and 384 DF,  p-value: 3.791e-13

> |
```

Logistic regression

- Uses a link function to enable linear regression methods to be applied to dichotomous outcomes
- $\ln(P/(1-P)) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$
- Exponentiate both sides:
 - $P/(1-P) \sim \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)$
 - $P/(1-P) \sim \exp(\beta_0) * \exp(\beta_1 X_1) * \exp(\beta_2 X_2) * \dots * \exp(\beta_n X_n)$

Logistic regression

```
R Console
> summary(lm(t2d~age+gender+waist,data=diabetes))

Call:
lm(formula = t2d ~ age + gender + waist, data = diabetes)

Residuals:
    Min       1Q   Median       3Q      Max
-0.57617 -0.19328 -0.10169  0.00942  1.02597

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.558494   0.121004  -4.615 5.35e-06 ***
age           0.006041   0.001068   5.659 2.98e-08 ***
genderfemale -0.001372   0.035245  -0.039 0.968969
waist         0.011377   0.003041   3.741 0.000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3403 on 384 degrees of freedom
(15 observations deleted due to missingness)
Multiple R-squared:  0.1231,    Adjusted R-squared:  0.1162
F-statistic: 17.96 on 3 and 384 DF,  p-value: 6.264e-11

> |
```

Other extensions of regression

- Poisson regression – count data
- GLM – generalized linear models
- GEE – generalized estimating equations
- Linear mixed effects models – random effects
- Cox proportional hazards models – survival analysis

One-sided or two?

- One-sided P -values should only be used when any possible effect can only go in a single direction
- In practice, this is extremely rare!
- Most arguments for one-sided P -values are not particularly valid
- One example: weight at one year versus weight at birth!

Study Designs

Cohort study

- Random sampling of base population should be representative of that population (the validity of that assumption increases with sample size)
- Cross-sectional: assessment at one moment in time
- Prospective: follow participants over time
 - Incidence over time; “survival”

Cohort study properties

- Generalizable
- Incorporate observation over time
- Expensive to ascertain and follow-up
- Rare conditions or exposures difficult to study

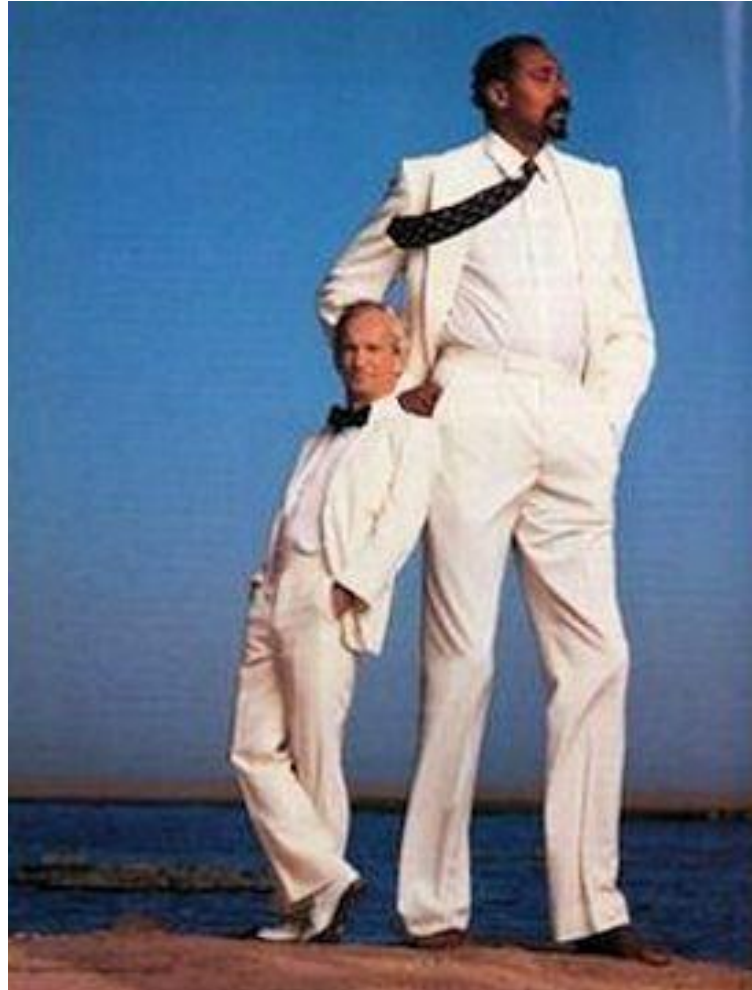
Case-control study

- Sample a case population and (appropriate) controls
- Cheaper than cohort
- Assess less frequent diseases/exposures
- May not be easily generalizable

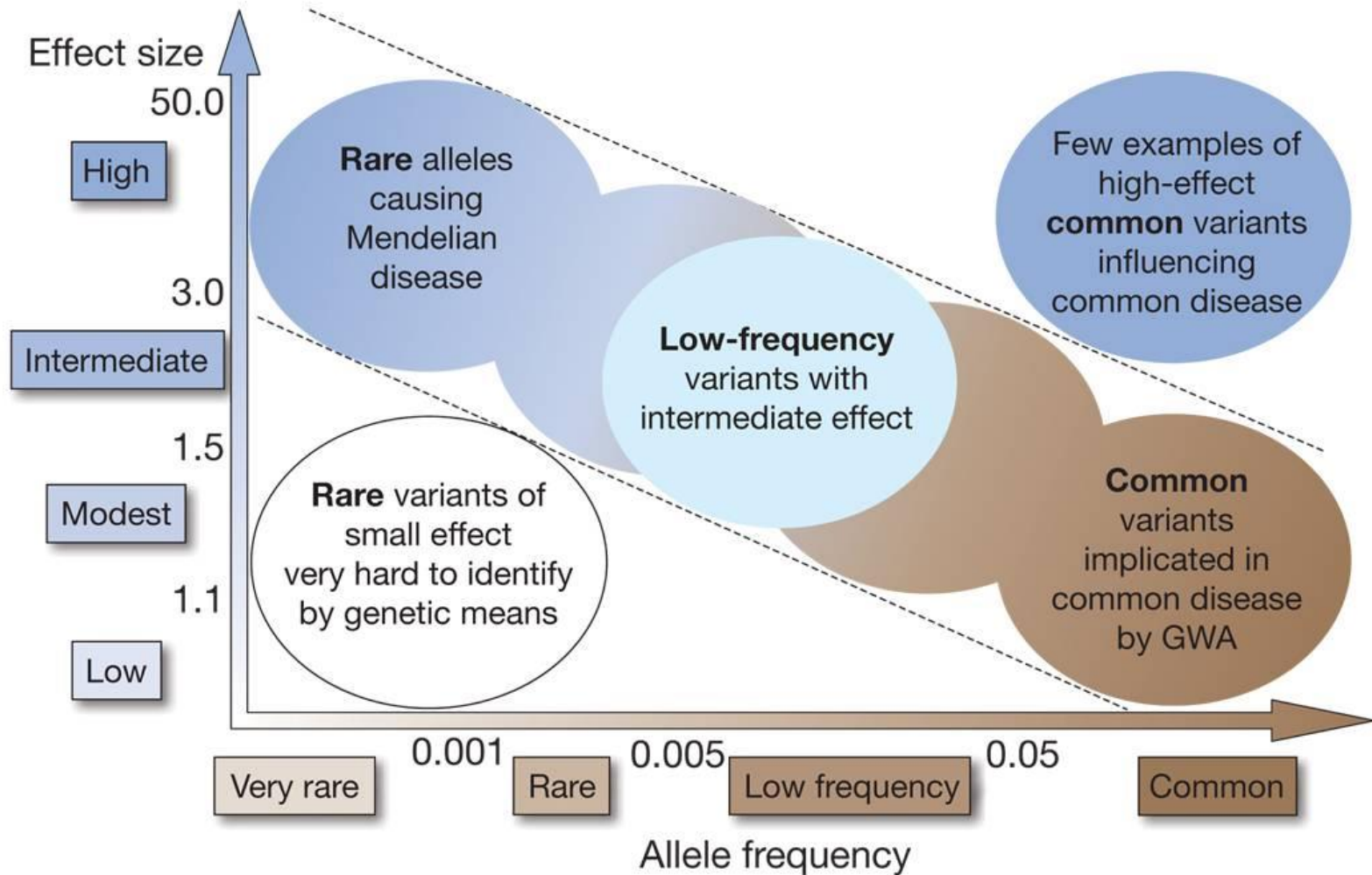
Others

- Case only designs
 - Cohort of cases
 - Case-control – “case cases” and “case controls”
- Case-cohort
- Sampling of extremes

Sampling extremes



Genetic studies – frequency vs. effect



Genetic studies

- Population-based
 - Embedded in other study designs
 - Standard analytical methods
- Family-based studies
 - Twin studies
 - Small pedigrees (such as trios & nuclear families)
 - Extended pedigrees
 - Require special analysis methods to account for relatedness

Some advantages of family studies

- Reduced genetic complexity
- Reduced environmental heterogeneity
- Enriched for rare alleles
- Enriched for rare phenotypes
- Robust to population stratification

Statistical Power and Power Calculation

Statistical errors

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

What is (statistical) power?

“In plain English, statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. If statistical power is high, the probability of making a Type II error, or concluding there is no effect when, in fact, there is one, goes down.”

Why is power important?

- Is it reasonable to proceed with a study?
 - Can differences be detected?
 - Is the necessary sample size achievable?
- Is it ethical to proceed with a study?
 - Will potentially deleterious sampling or intervention (or animal use) have the potential to yield useful results?

Power calculation

- Four parameters in typical power calculation:
 - Effect size
 - Sample size (n)
 - Significance level (α) [P(Type I error)]
 - Statistical power (β) [$1 - P(\text{Type II error})$]
- These parameters are related! If you know 3, you can calculate the fourth!

Power depends on...

- Since the significance level is (usually) pre-defined, that means power is crucially dependent on:
 - The effect size you hope to find and
 - The sample size (n)!

Power calculators

Note that different statistical tests require different power calculations!

- By hand
- R packages (base package, pwr, powerSurvEpi)
- Stand-alone software
- Online calculators

Online power calculators

- <http://powerandsamplesize.com/Calculators/>
- <http://www.sample-size.net/>
- A huge list: <http://statpages.info/>
- And many others...

R package pwr (a few examples)

- Student's t-test
 - > `pwr.t.test(n = , d = , sig.level = , power = , type = c("two.sample", "one.sample", "paired"))`
 - > `pwr.t2n.test(n1 = , n2 = , d = , sig.level = , power =)`
- ANOVA
 - > `pwr.anova.test(k = , n = , f = , sig.level = , power =)`
- Correlation
 - > `pwr.r.test(n = , r = , sig.level = , power =)`
- χ^2 test
 - > `pwr.chisq.test(w = , N = , df = , sig.level = , power =)`

Power calculation in genetics

- Allele frequencies can effectively reduce sample size, so power calculations need to account for this additional parameter
- Make sure to account for adjusted P -value threshold!

Genetic power calculators

- Sham and Purcell
 - Online tool
 - <http://zzz.bwh.harvard.edu/gpc/>
- Quanto
 - Stand-alone software
 - <http://biostats.usc.edu/Quanto.html>

Power calculation for RNAseq

- Power is dependent on frequency of transcription for a given transcript
- Sequencing depth is a related factor (very deep sequencing increases frequency of rare transcripts)
- RNAseq generates count data, which requires different distributional assumption (negative binomial or Poisson)

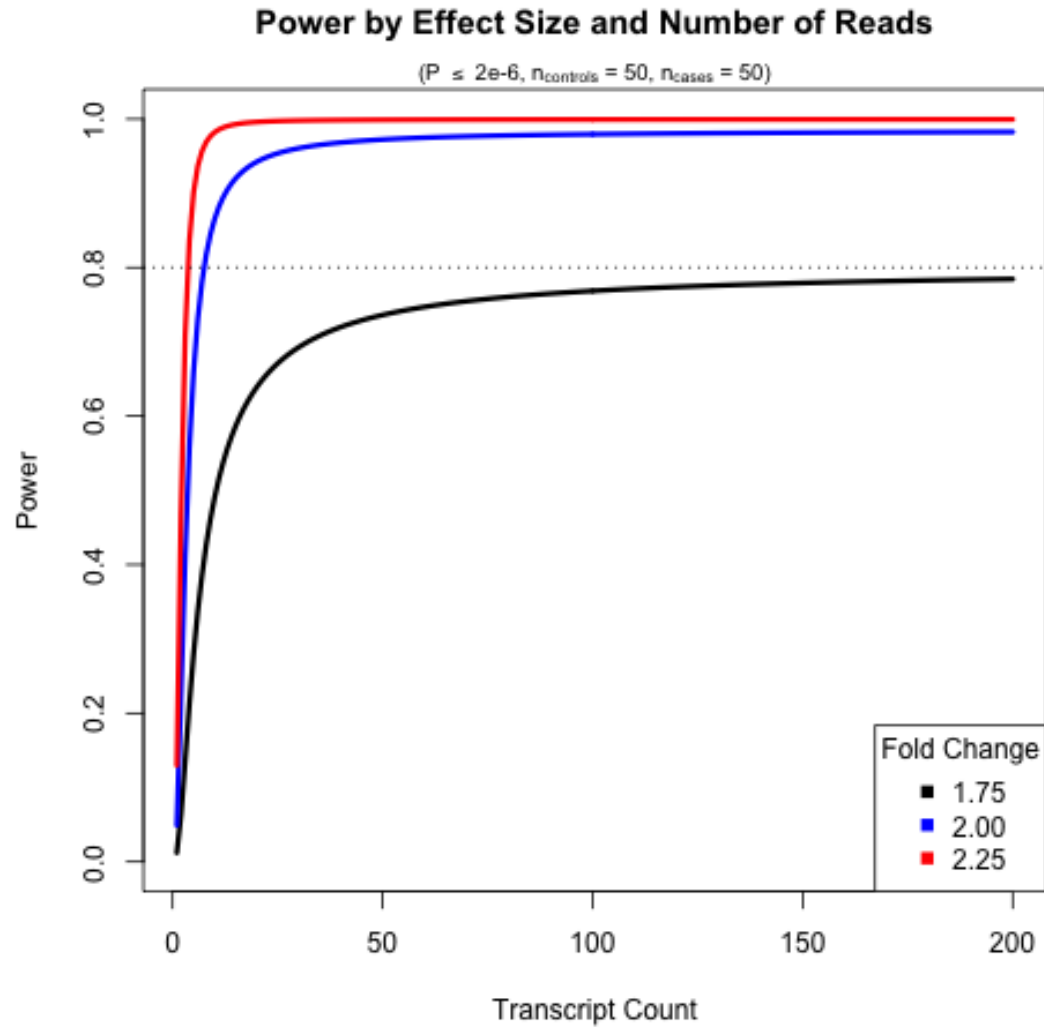
Power calculation for RNAseq

- Several available tools in R
- RNASeqPower is one option:
 - Uses negative binomial distribution appropriate for count data
 - Allows a range of transcript counts and effect sizes
 - Allows different sized case and control groups

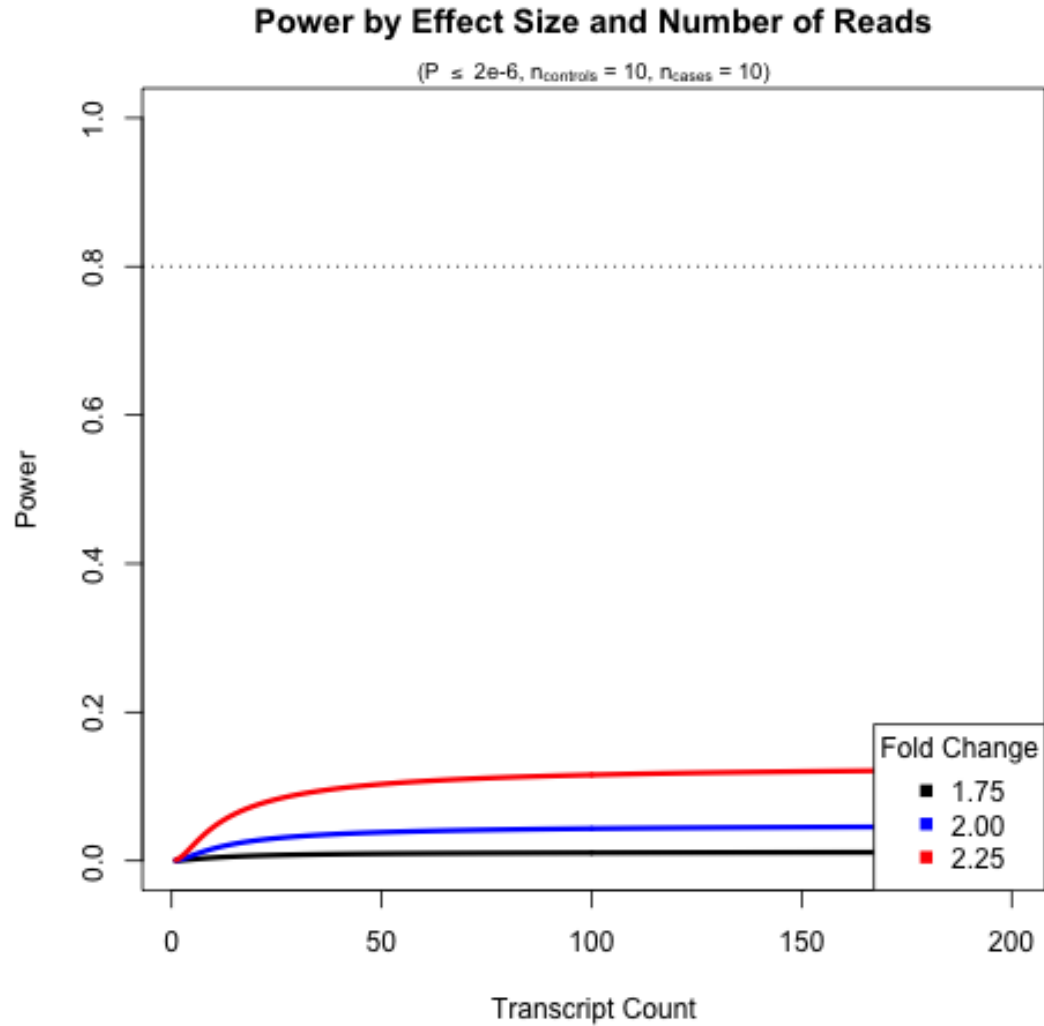
RNASeqPower

- > library(RNASeqPower)
- > a<-rnapower(depth=c(1:200),cv=0.5,effect=c(1.75,2,2.25),alpha=2e-6,n=50,n2=50)
- > png("power_rna_seq2019-07-03.png")
- > plot(a[,1],type="l",lwd=3,xlab="Transcript Count",ylab="Power",main="Power by Effect Size and Number of Reads",ylim=c(0,1))
- > points(a[,2],type="l",lwd=3,col="blue")
- > points(a[,3],type="l",lwd=3,col="red")
- > legend("bottomright",pch=15,col=c("black","blue","red"),legend=c("1.75","2.00","2.25"),title="Fold Change")
- > mtext(expression("(P " <= " 2e-6, n["controls"]*" = 50, n["cases"]*" = 50)"),cex=0.8)
- > abline(h=0.8,lty="dotted")
- > dev.off()

RNASeqPower



RNASeqPower

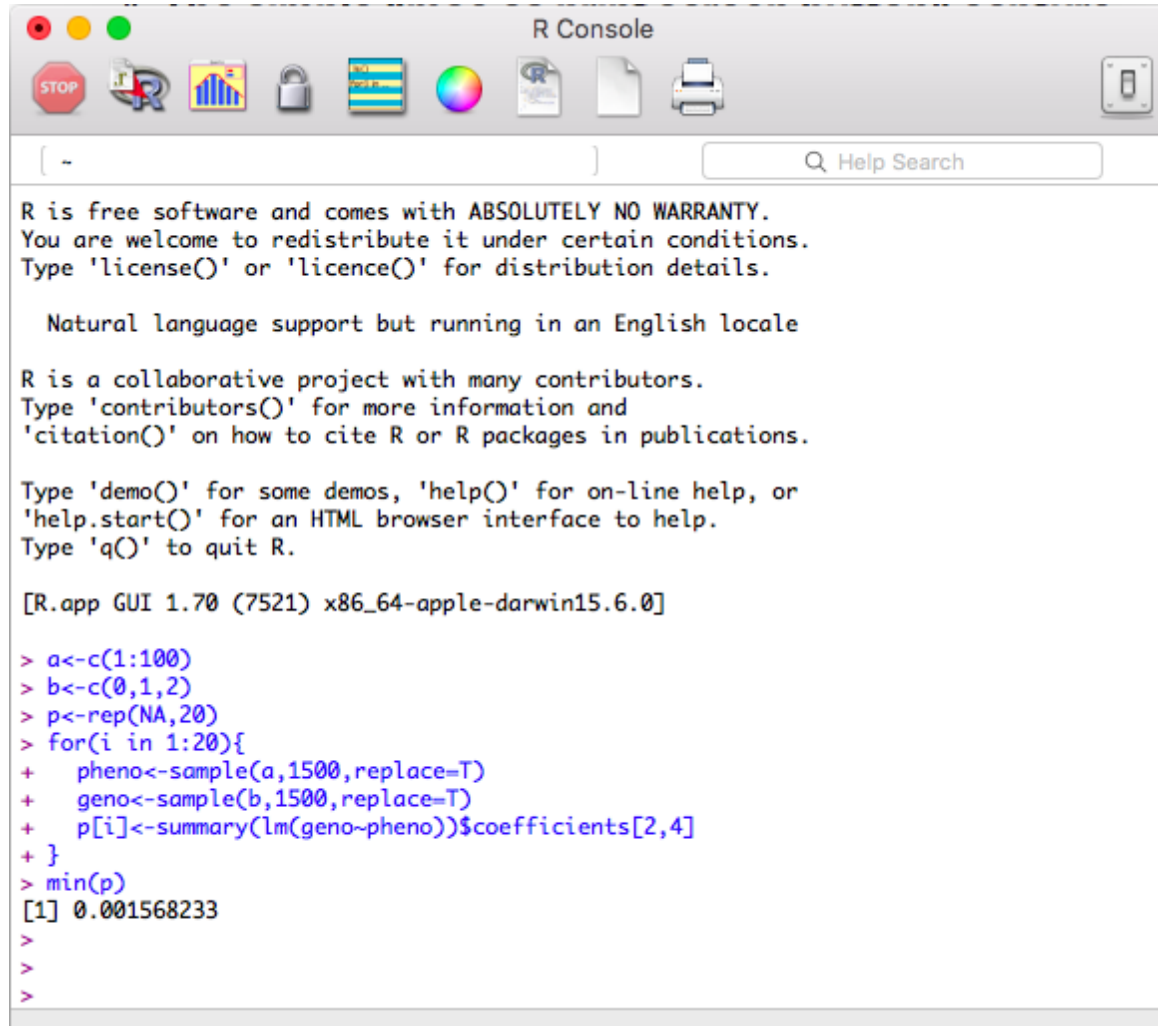


Multiple Comparisons and Corrections

The problem of multiple comparisons

- More tests = more Type I errors
- More Type I errors leads to:
 - Erroneous interpretation of results
 - Expenditure of money/time/effort chasing down false positives with follow-up experiments

Is "multiple testing" *really* a concern?



```
R Console

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

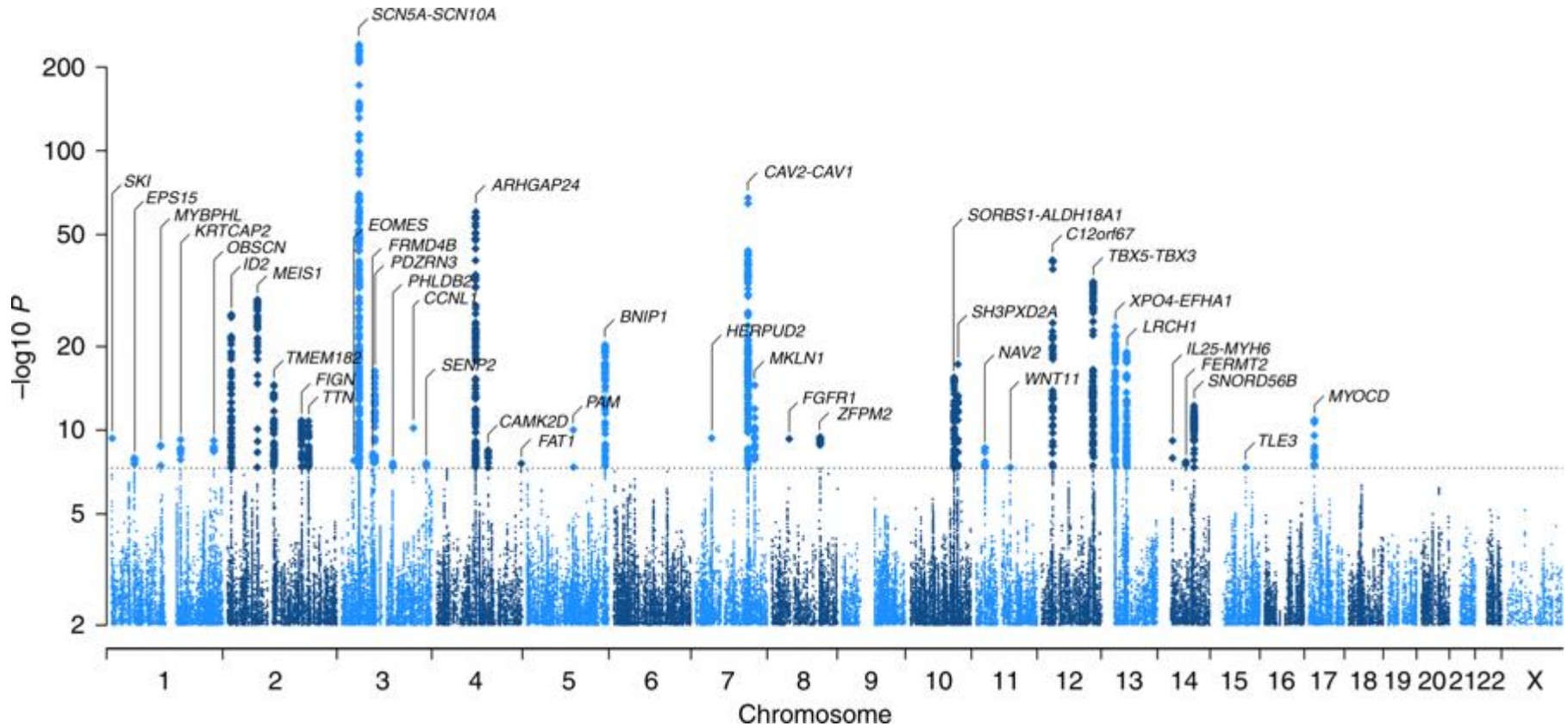
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.70 (7521) x86_64-apple-darwin15.6.0]

> a<-c(1:100)
> b<-c(0,1,2)
> p<-rep(NA,20)
> for(i in 1:20){
+   pheno<-sample(a,1500,replace=T)
+   geno<-sample(b,1500,replace=T)
+   p[i]<-summary(lm(geno~pheno))$coefficients[2,4]
+ }
> min(p)
[1] 0.001568233
>
>
>
```

Repeated 1000
iterations: $P \leq 0.05$
608 times and $P \leq 0.10$
871 times!

Multiple testing in a genetics study



A P -value of 0.05 corresponds to 1.3 on the y-axis!!!

Corrections: science or art?

- How strict should the corrections for multiple testing be?
 - For each predictor?
 - For each outcome?
 - For previous studies on the same topic?
 - For all studies in a cohort?
 - For all studies, ever, in the history of the world?
- Be “strict enough” and be able to motivate the choices made!

Adjustments for multiple comparisons

- Reduce false positive findings (perhaps at the expense of increasing false negatives)
- Enhance reproducibility of the results
- Increase believability and impact

Bonferroni correction

- Simple to implement ($P_{\text{adjusted}} = 0.05/n_{\text{tests}}$)
- Rigorous: one expected Type I error
- Crucially, assumes *independence* between tests (both predictors and outcomes)
- Increases Type II error rate (decreases power)

Benjamini and Hochberg (False Discovery Rate)

- Limits expected proportion of Type I errors
- Less stringent than Bonferroni
- Suitable for many tests
- Increased Type I error compared to Bonferroni

Matrix spectral decomposition

- Accounts for correlation between predictors
- Calculates an “effective” number of independent tests, which is then used in a Bonferroni correction
- Can also be applied to correlated outcomes
- Webtools and downloadable software (<https://sites.google.com/site/qutsgel/software>)

Permutation (the “gold standard”?)

- Perform analysis
- “Shuffle” outcomes or predictors and re-perform analysis – note lowest P -value
- Do this many times!
- Proportion of P -values less than original is permuted P

Permutation

- Carefully performed, it can preserve structure inherent to the data (linkage disequilibrium, relationship between outcome and covariates, etc.)
- Provides excellent estimate of how likely it is to observe a more extreme results *in the actual data*
- However: computationally intensive, particularly as the dataset grows

Other corrections

- Tukey's range test
- Benjamini and Yekutieli
- Holm
- Sudak
- Etc.

The bottom line?

- Think about the statistics BEFORE you start your study
- Calculate power
- Keep a stats book handy
- Seek out an expert if necessary!