Why is multiple testing a problem and what do I need to do about it?

Tiffany Timbers, Ph.D.

June 21, 2018
Omics @ SFU meeting

@TiffanyTimbers

Outline

- Review:
    1. hypothesis test
    2. test statistic
    3. p-value
- Multiple hypothesis test
- Multiple hypothesis testing problems
- Multiple hypothesis testing solutions

The hypothesis test paradigm in science

- **Scientific question:**

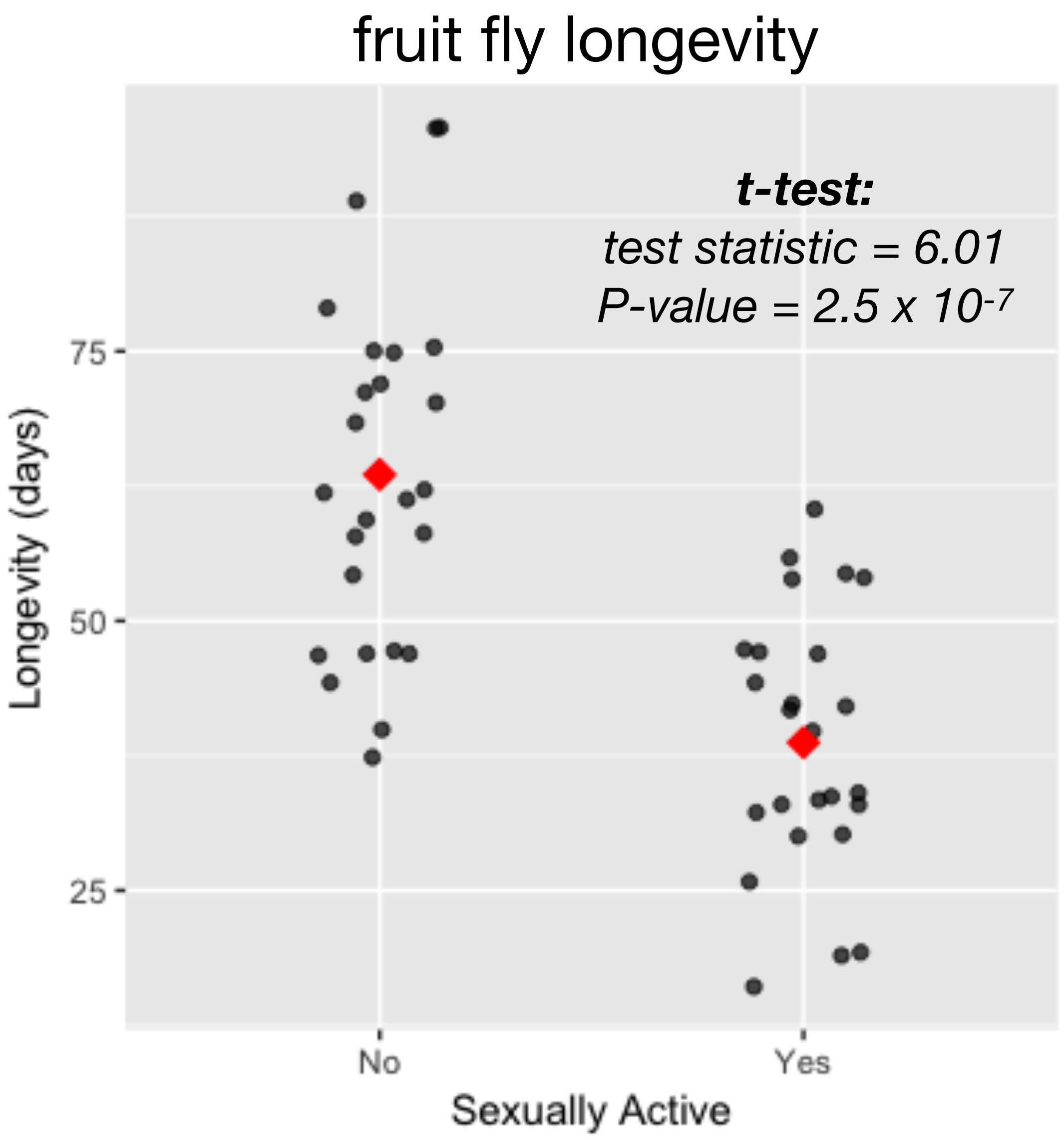
  on average, is longevity the same for sexually active and non-sexually active fruit flies?

- **Statistical Hypotheses**:

  H$_0$: $\mu_{No} = \mu_{Yes}$

  H$_A$: $\mu_{No} \neq \mu_{Yes}$

  where $\boldsymbol{\mu}$ *represents the population mean*



t-test:
test statistic = 6.01
P-value = 2.5 x 10^{-7}

Partridge & Farquhar, Nature, 1981

@TiffanyTimbers

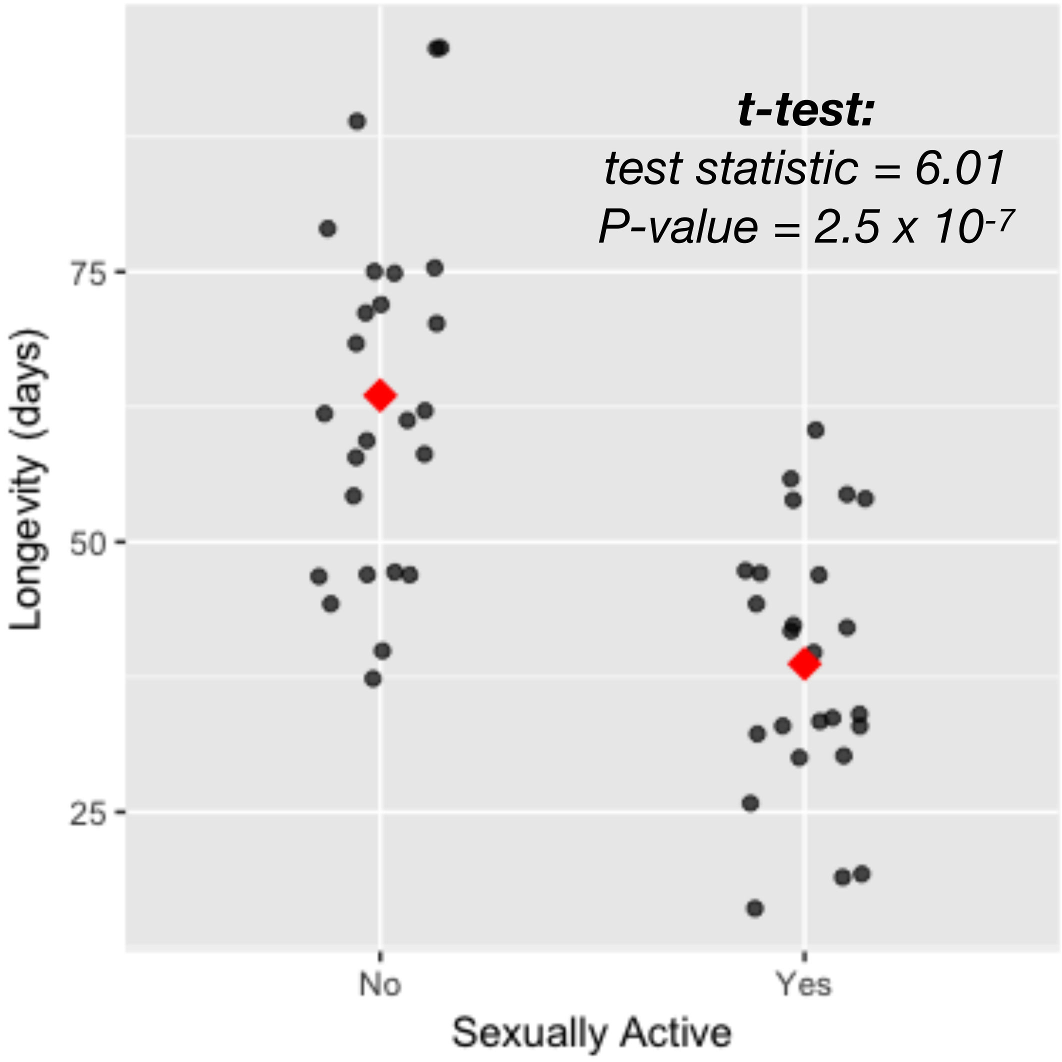
The hypothesis test paradigm in science

- **test statistic:**

  A test statistic measures the degree of agreement between the sample(s) of data and the null hypothesis.

- **P-value**:
  The probability of getting a test statistic at least as extreme as the one from your sample data, assuming the null hypothesis is true.

fruit fly longevity



t-test:
test statistic = 6.01
P-value = 2.5×10^{-7}

Partridge & Farquhar, Nature, 1981

The hypothesis test paradigm in science

Remembering:

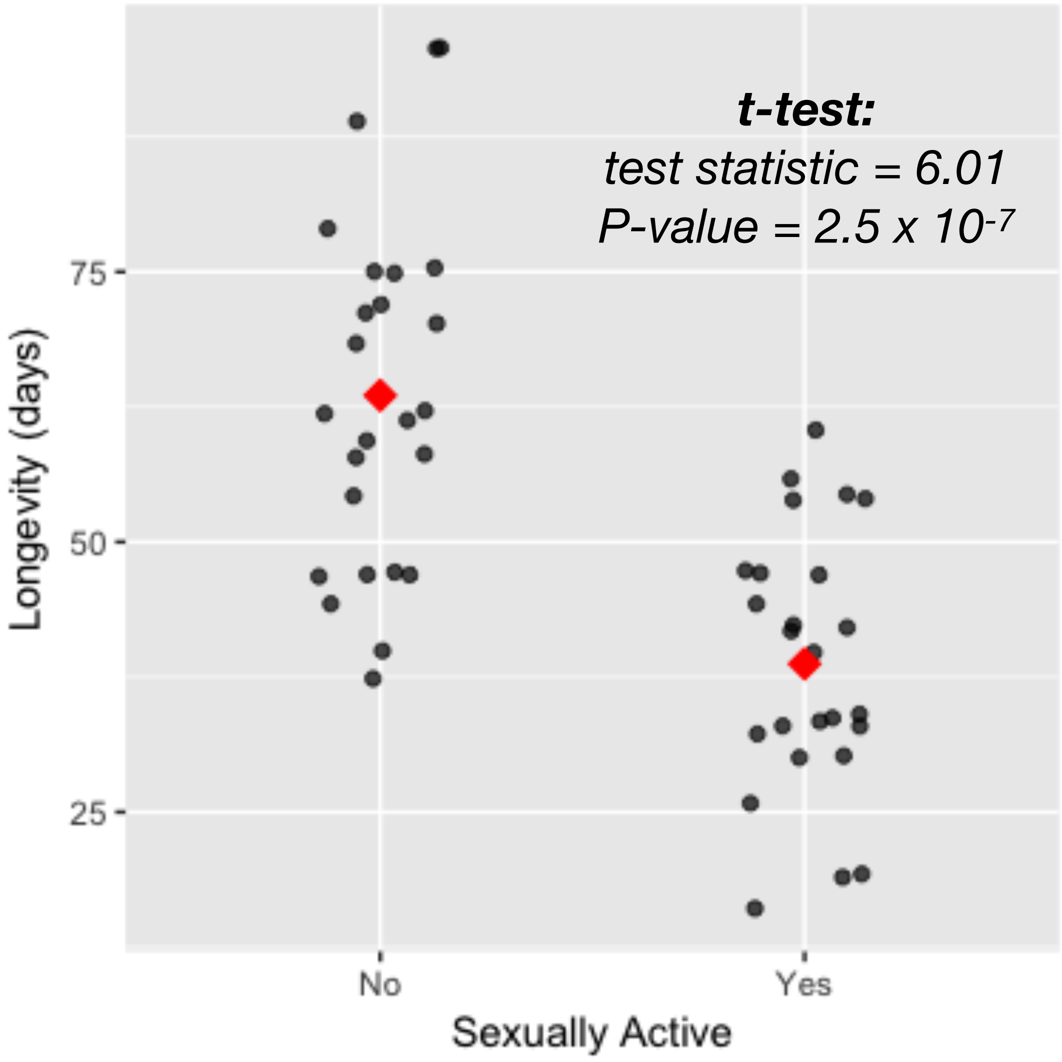
H_0: $\mu_{No} = \mu_{Yes}$

H_A: $\mu_{No} \neq \mu_{Yes}$

At a significance level $\alpha = 0.05$, testing procedure can be cast as:

- Reject the H_0 if P-value ≤ 0.05
- Don't reject the null otherwise.

fruit fly longevity



t-test:

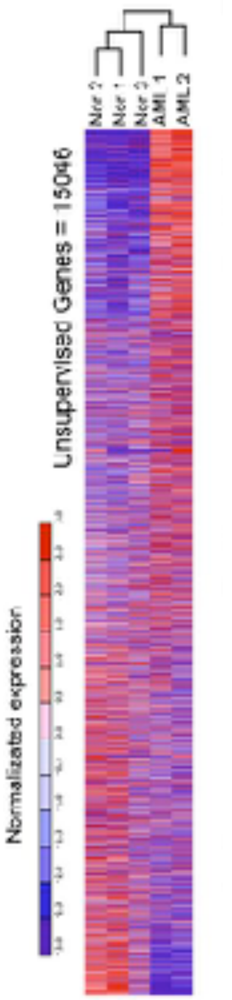
test statistic = 6.01

P-value = 2.5×10^{-7}

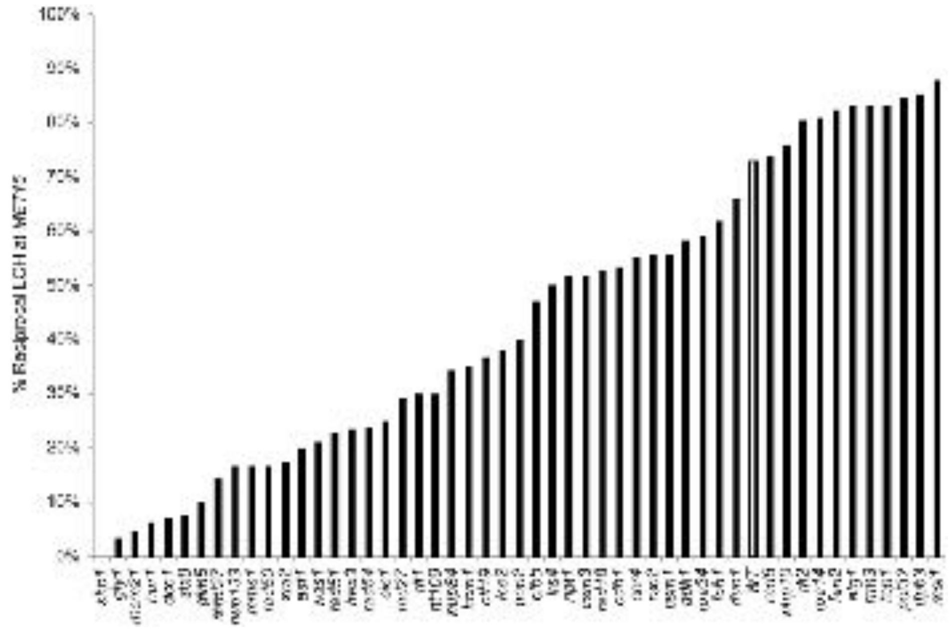
Partridge & Farquhar, Nature, 1981

@TiffanyTimbers

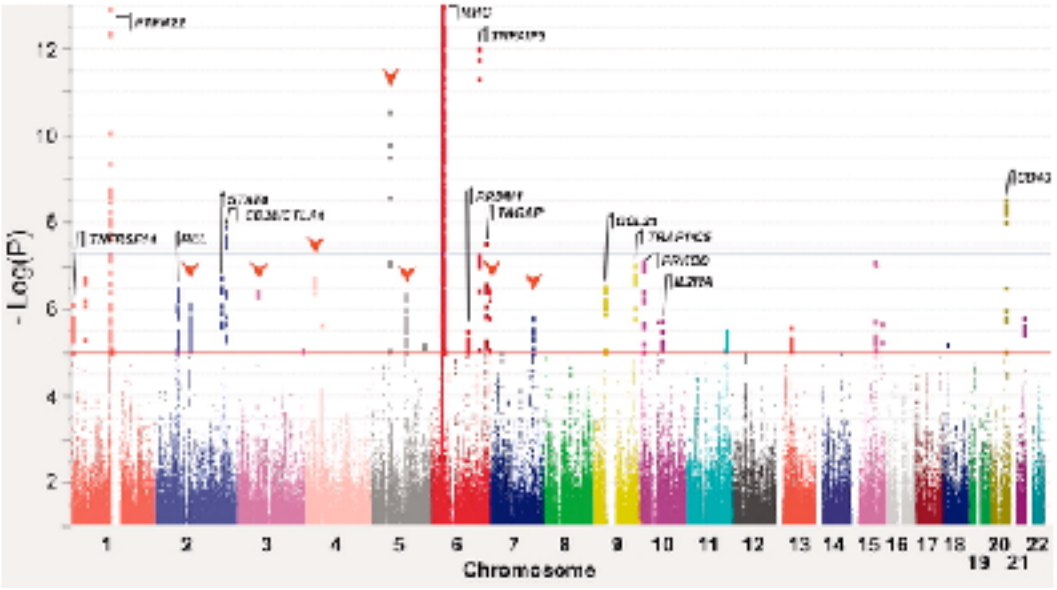
Multiple hypothesis tests in biology



https://doi.org/10.1182/blood-2013-03-489823



https://doi.org/10.1534/genetics.108.089250



https://doi.org/10.1534/genetics.110.120907

@TiffanyTimbers

The multiple hypothesis test problem

- When many hypotheses are tested simultaneously you increase the chance of false positives.

- ***Example:*** imagine you have a RNAseq experiment looking at the expression of **20,000 genes** and **not a single one is differentially expressed**. As usual, a significance level *α* = 0.05 is used.

    - By chance alone 20,000 x 0.05 = 1000 may have a P-value < 0.05

    - Thus here, individual P–values of 0.05 are no longer considered “significant” findings.

Need to adjust for multiple testing when assessing the statistical significance of findings!

Two multiple hypothesis test solutions

1. The Bonferroni correction
2. The False Discovery Rate

The Bonferroni correction

- If m hypothesis tests are to be done, then adjust significance threshold to be a/m
- From our example where we want to compare the differential expression of 20,000 genes where none are actually differentially expressed:
    - new significance threshold becomes: 0.05/20000 = 0.0000025
    - now, by chance alone there is only a ~ 5% chance that we will find a single significant finding:

$$
\begin{aligned}
Pr(\text{at least one significant result}) &= 1 - Pr(\text{no significant results}) \\
&= 1 - (1 - 0.0000025)^{20000} \\
&= 0.049
\end{aligned}
$$

The Bonferroni correction

- BUT there is no free lunch…

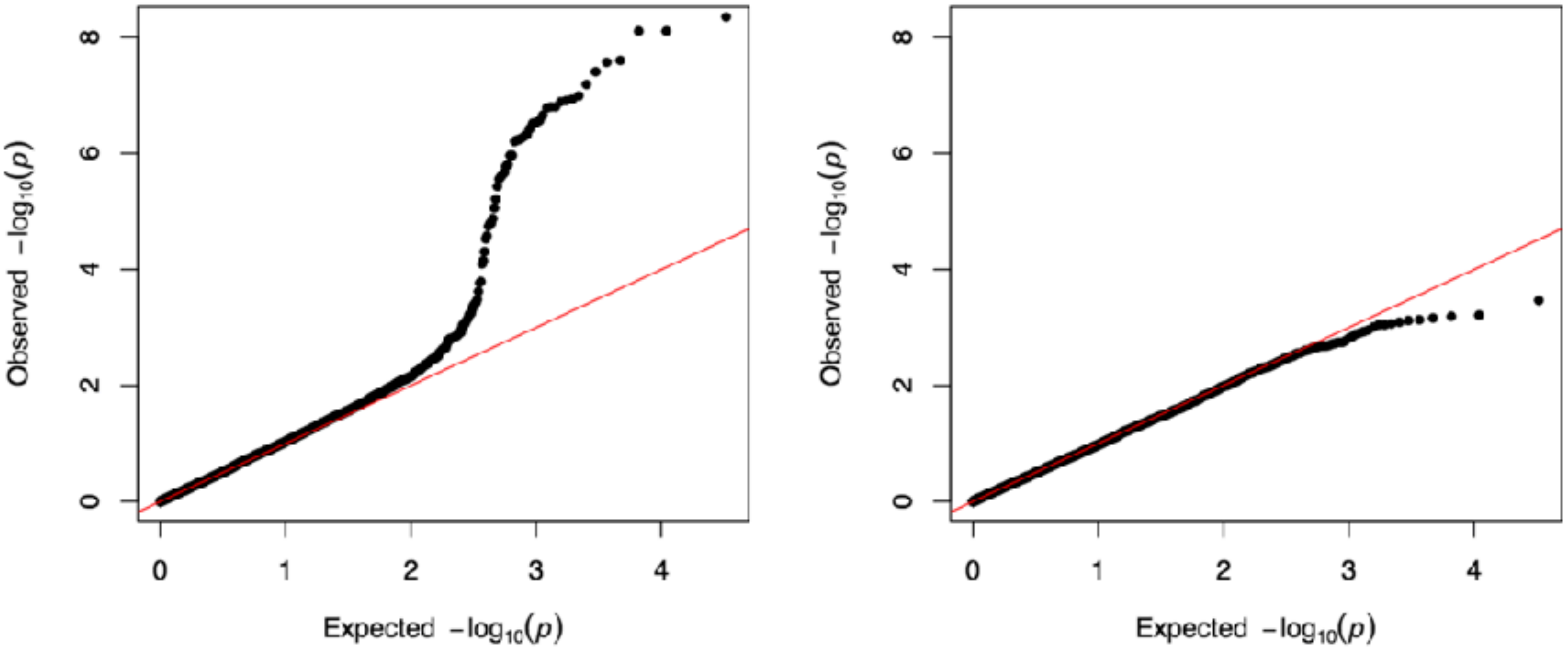
- In this example (with 20,000 hypothesis tests) to detect a "significant" difference I will need $P \leq 0.0000025$!!!

- This is possible if:

    - there is a large effect (e.g., large change in expression level)

    - you have a large sample size

The False Discovery Rate



Can we automate finding the “bend in the curve” and using this as the cutoff to label significant versus non-significant findings?

~~~
```{r}
library(qqman)
par(mfrow = c(1,2)
qq(gwasResults$P, ylim = c(0, 8.2))
qq(runif(16470), ylim = c(0, 8.2))
```
~~~

The False Discovery Rate

The Benjamini-Hochberg method:

- Choose a maximum tolerable false discovery rate, δ (e.g., 5%)
- Sort the P-values from the m hypothesis tests from smallest to largest:
    - $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \ldots \leq p_{(m)}$
- Let k^* be the biggest k for which $p_{(k)} < (\delta \,/\, m)\, k$
- Take $p_{(1)}, \ldots, p_{(k^*)}$ as the discoveries/significant findings

The False Discovery Rate

A simple example:

P-value	***k***	**(δ / *m*) *k***
0.0010	1	(0.05 / 10) x 1 = 0.005
0.0070	2	(0.05 / 10) x 2 = 0.010
0.0120	3	(0.05 / 10) x 3 = 0.015
0.0307	4	(0.05 / 10) x 4 = 0.020
0.1096	5	(0.05 / 10) x 5 = 0.025
0.2612	6	(0.05 / 10) x 6 = 0.030
0.4018	7	(0.05 / 10) x 7 = 0.035
0.5828	8	(0.05 / 10) x 8 = 0.040
0.7161	9	(0.05 / 10) x 9 = 0.045
0.9628	10	(0.05 / 10) x 10 = 0.050

- Let k^* be the biggest k for which $p_{(k)} < (\delta / m)\, k$
- Take $p_{(1)}$, … , $p_{(k^*)}$ as the discoveries/significant findings

@TiffanyTimbers

The False Discovery Rate

A simple example:

P-value	***k***	**(δ / *m*) *k***
0.0010	1	(0.05 / 10) x 1 = 0.005
0.0070	2	(0.05 / 10) x 2 = 0.010
0.0120	3	(0.05 / 10) x 3 = 0.015
0.0307	4	(0.05 / 10) x 4 = 0.020
0.1096	5	(0.05 / 10) x 5 = 0.025
0.2612	6	(0.05 / 10) x 6 = 0.030
0.4018	7	(0.05 / 10) x 7 = 0.035
0.5828	8	(0.05 / 10) x 8 = 0.040
0.7161	9	(0.05 / 10) x 9 = 0.045
0.9628	10	(0.05 / 10) x 10 = 0.050

k^* = 3

- Let k^* be the biggest k for which $p_{(k)} < (\delta / m)\, k$
- Take $p_{(1)}$, … , $p_{(k^*)}$ as the discoveries/significant findings

@TiffanyTimbers

The False Discovery Rate

A simple example:

P-value	***k***	**(δ / *m*) *k***
0.0010	1	(0.05 / 10) x 1 = 0.005
0.0070	2	(0.05 / 10) x 2 = 0.010
0.0120	3	(0.05 / 10) x 3 = 0.015
0.0307	4	(0.05 / 10) x 4 = 0.020
0.1096	5	(0.05 / 10) x 5 = 0.025
0.2612	6	(0.05 / 10) x 6 = 0.030
0.4018	7	(0.05 / 10) x 7 = 0.035
0.5828	8	(0.05 / 10) x 8 = 0.040
0.7161	9	(0.05 / 10) x 9 = 0.045
0.9628	10	(0.05 / 10) x 10 = 0.050

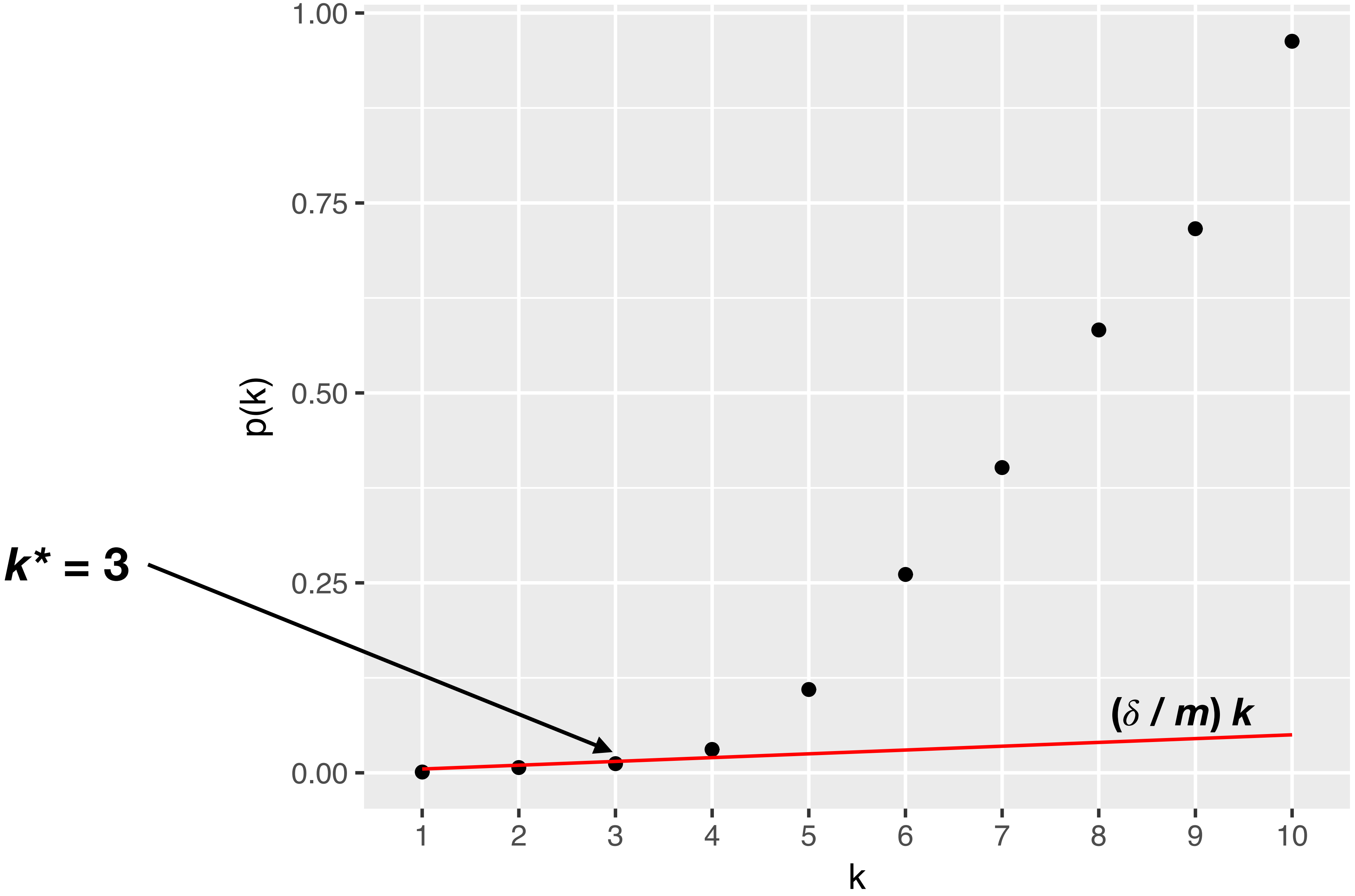
$k^* = 3$

- Let k^* be the biggest k for which $p_{(k)} < (\delta / m)\, k$
- Take $p_{(1)}, \dots, p_{(k^*)}$ as the discoveries/significant findings

@TiffanyTimbers

The False Discovery Rate

A simple example cont’d:



k* = 3

@TiffanyTimbers

Code for plot on previous slide

~~~
```{r}
library(tidyverse)

raw_pvalues = c(0.0010, 0.0070, 0.0120, 0.0307, 0.1096,
  0.2612, 0.4018, 0.5828, 0.7161, 0.9628)

pvalues <- data.frame(raw_pvalues, k = seq_along(raw_pvalues)) %>%
  mutate(bh_line = 0.05/nrow(.) * k)

ggplot(pvalues, aes(x = factor(k), y = raw_pvalues)) +
  geom_point() +
  geom_line(aes(x = k, y = bh_line), colour = "red") +
  xlab("k") +
  ylab("p(k)")
```
~~~

Two multiple hypothesis test solutions:

which to use?
and when?

Which to use? and when?

1. **The Bonferroni correction:** choose this if high confidence in all findings labelled as “significant” is needed (*i.e.,* if its better to be very conservative and have false negatives).

2. **The False Discovery Rate:** choose this if a certain proportion of false positives in all findings labelled as “significant” is tolerable (*i.e.,* if its better to be more liberal and have false positives).

@TiffanyTimbers

How to implement in R or Python?

R:

- p.adjust (base stats package)
- fdrtool package

Python:

- statsmodels package, specifically:
    - statsmodels.sandbox.stats.multicomp.multipletests

@TiffanyTimbers

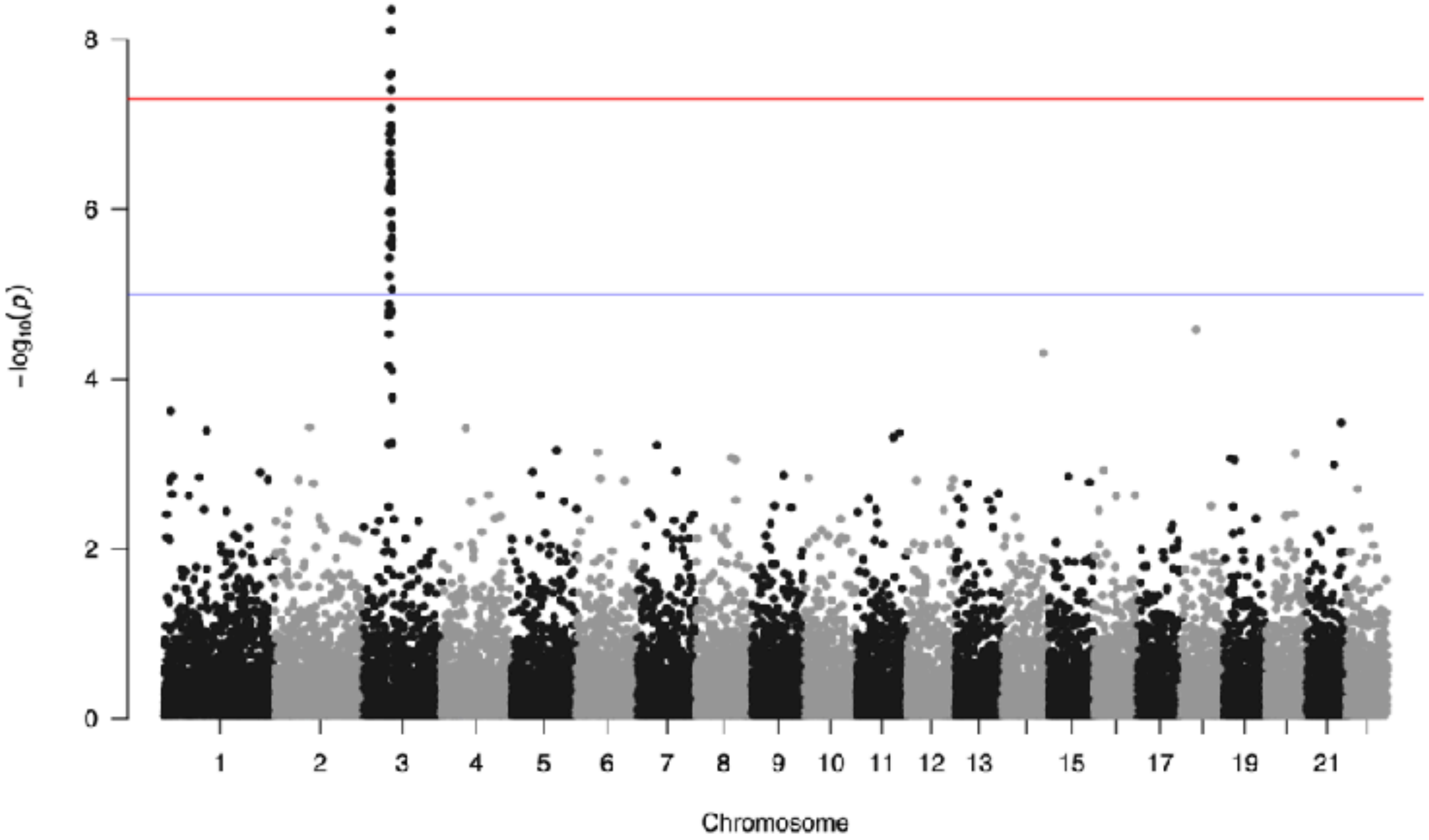
What we talked about

- Review:
  1. hypothesis test
  2. test statistic
  3. p-value
- Multiple hypothesis test
- Multiple hypothesis testing problems
- Multiple hypothesis testing solutions

Thanks!

@TiffanyTimbers

The False Discovery Rate



```{r}
library(qqman)
manhattan(gwasResults)
```