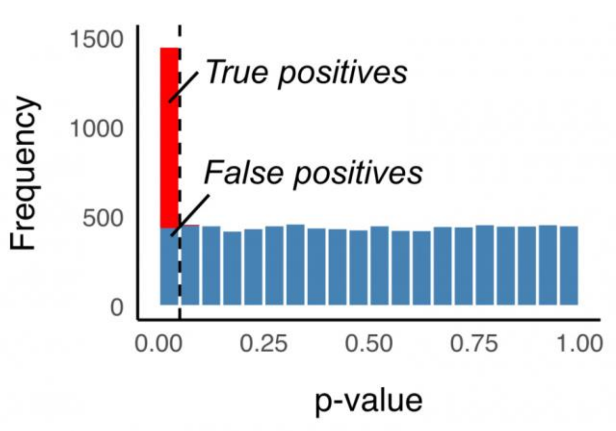
Statistical Analysis of Biological Data

- Multiple Testing -

Topic 12

Multiple Testing



Gregory R. Grant

Genetics Department

ggrant@pennmedicine.upenn.edu

Fall, 2023

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT YEARS AGO.

ITMAT Bioinformatics Laboratory

University of Pennsylvania

What is Multiple Testing?

- Multiple testing is not one thing or one problem.
- But in a nutshell, it arises whenever we try to draw conclusions that involve more than one test.
- Multiple testing causes the most difficulties when the balance between true positives and false positives gets out of whack, like it did in the AIDS testing example.
  - If the population has a lot of negatives and just a few positives, then a small percent of the false-positives can swamp out a large percent of the true-positives.

A matter of degree

How to handle multiple testing problems also depends on how many tests are being performed.

A typical paper based on PCR tests: Dozens, possibly a hundred tests, not thousands.

RNA-Seq DE analysis: Tens of thousands of tests (one for each gene).

We're going to look at both, but focus on the latter.

Review of AIDS Testing Example

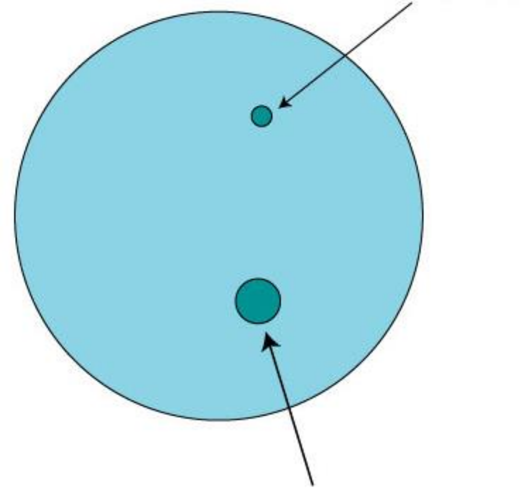
- A small percentage (FP=0.01) of a large population (those without AIDS) swamps out a large percentage (TP=.0.99) of a small population (those with AIDS).

- 1% of the healthy population falsely testing positive exceeds by fourfold the total number of true AIDS cases.
  - The first 0.01 “FP rate” is about a test on one individual.

    Prob(test positive | negative) = 0.01
  - The second 0.8 “FP rate” is about a population.

    Prob(negative | test positive) = 0.8

True Positives = 0.25%

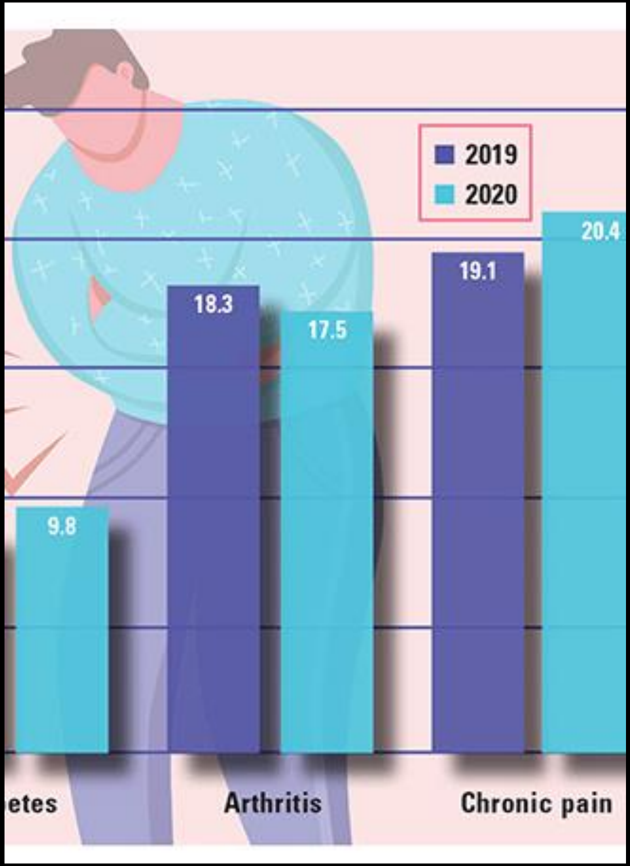


False Positives = 1%

- Therefore, the severity of the problem depends on something unknown: *the true population structure.*

Population Structure

- We assumed 1 in 400 going for testing are positive and concluded that 80% of those testing positive are actually false-positives.
- Suppose on the other hand that 40 of every 400 going for testing are positive.
- Now there'll be 40 true-positives and 0.01 x 360 = 3.6 false-positives.
- Now the probability of being negative if you tested positive has decreased from 80% to 10%.



Populations

- two extremes -

The nature of the multiple testing problem depends on the makeup of the population of things being tested.

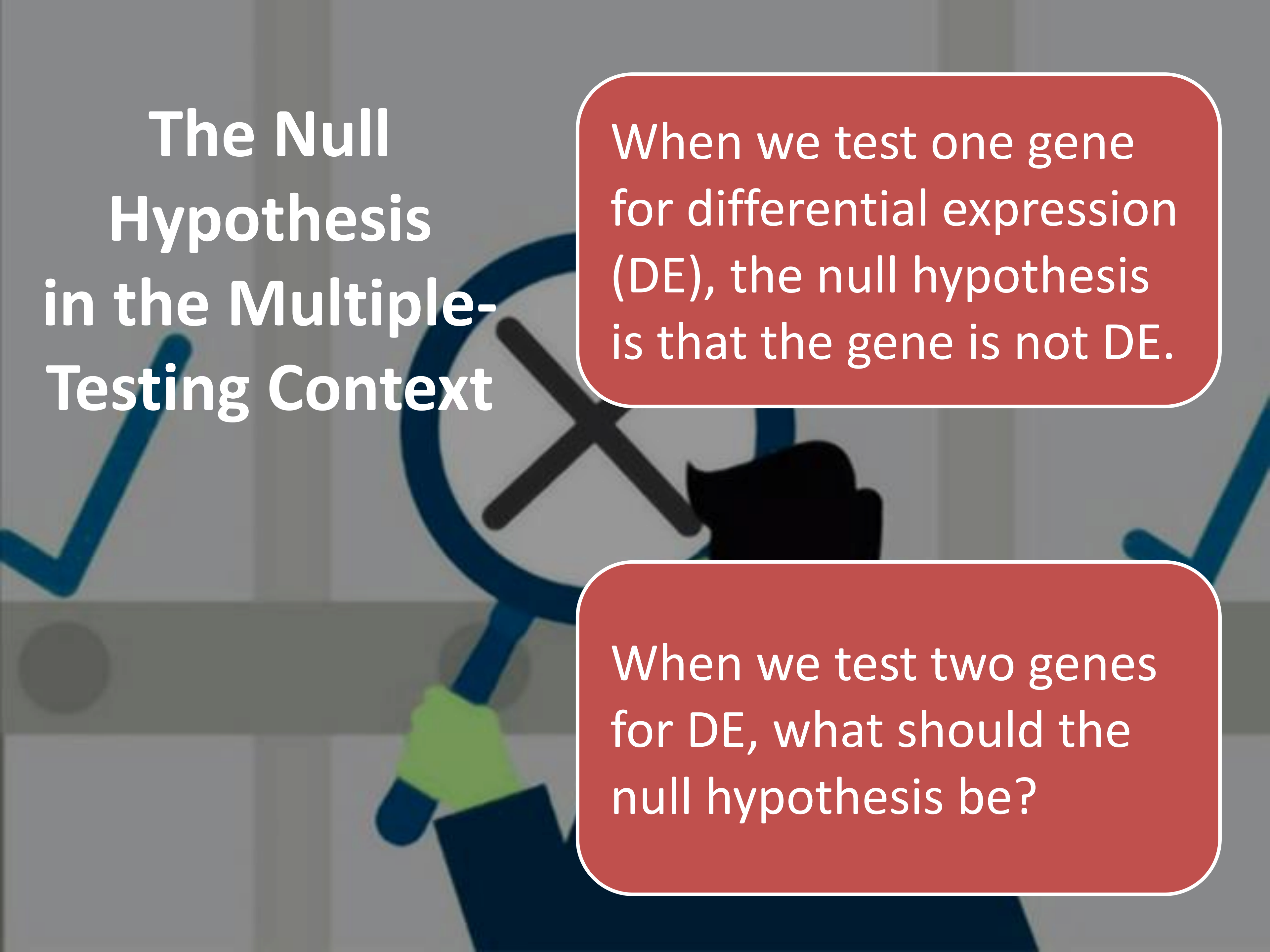
When doing PCR tests for DE for a study, it's likely that most of the null hypotheses are false.

- We tend to PCR test things we have a strong suspicion are DE.

When testing 30,000 genes for DE, it's likely that most of the null hypotheses are true.

- RNA-Seq is a fishing expedition, most genes are not DE.

The Null Hypothesis in the Multiple-Testing Context



When we test one gene for differential expression (DE), the null hypothesis is that the gene is not DE.

When we test two genes for DE, what should the null hypothesis be?

The Null Hypothesis for Two Genes

- **Option 1**
  - Make ***one*** null hypothesis:
  - H0: Neither gene is DE.
  - This is called “The Complete Null Hypothesis”.

- **Option 2**
  - Make ***two*** hypotheses:
    - $H0_1$: Gene one is not DE
    - $H0_2$: Gene two is not DE

The Complete Null

- Testing the complete null solves the multiple testing problem.

- But it doesn't tell us very much. If we reject H0, then we conclude that at least one is DE.

- This information is not very useful because we don't want just to know if some gene is DE, we want to know *which* genes are DE.

Multiple Hypotheses

- Two Tests -

- In this case we have two null hypotheses:

  $$H0_1 \text{ and } H0_2$$

- Suppose each test is run with a Type I error of 0.05.
- This leads to two *p*-values.
- Suppose we reject their respective null hypotheses if their *p*-value $p \leq 0.05$.
- How often will we reject *at least one* true null hypothesis?

Type I Error for Two Tests

- If both null hypotheses are true, then the probability of rejecting at least one of them is one minus the probability accepting both.

  (since the opposite of rejecting at least one, is rejecting none)

  - The probability of *not* rejecting one is 0.95, so the probability of not rejecting either is 0.95^2

- So, the probability of rejecting at least one true null hypothesis is

$$1 - 0.95^2 = 0.0975$$

- Our probability of making at least one mistake has almost doubled by doing two tests.
  - Which is not surprising.

Strong Control

- Suppose we are not willing to make *any* false rejections.
  - This is called “Strong Control” of the multiple testing problem.

- We could achieve this by lowering the per-test *p*-value cutoff to

$$\frac{0.05}{2} = 0.025$$

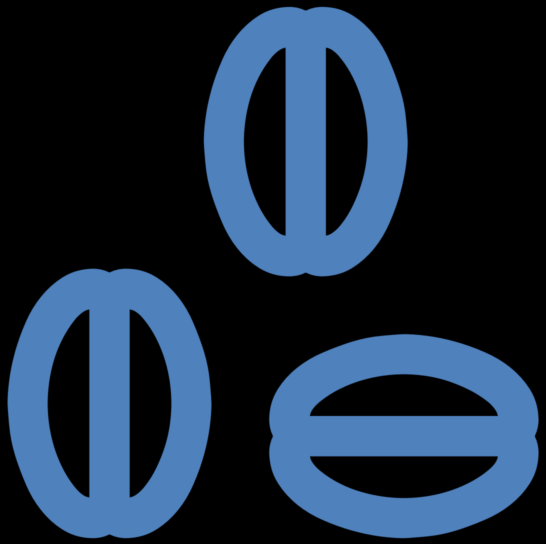
- Now the probability of making any false-positives at all is

$$1 - 0.975^2 = 0.049375$$

which just sneaks in under 0.05.

○ This is known as a “Bonferroni correction”.

➢ Let’s see what happens when there are three tests.



Strong Control with Three Genes

- If we had three genes, we'd have to make the cutoff even lower.

- We could achieve this by lowering the *p*-value cutoff to

$$\frac{0.05}{3} = 0.01\bar{6}$$

- Now the probability of making any false-positives at all is

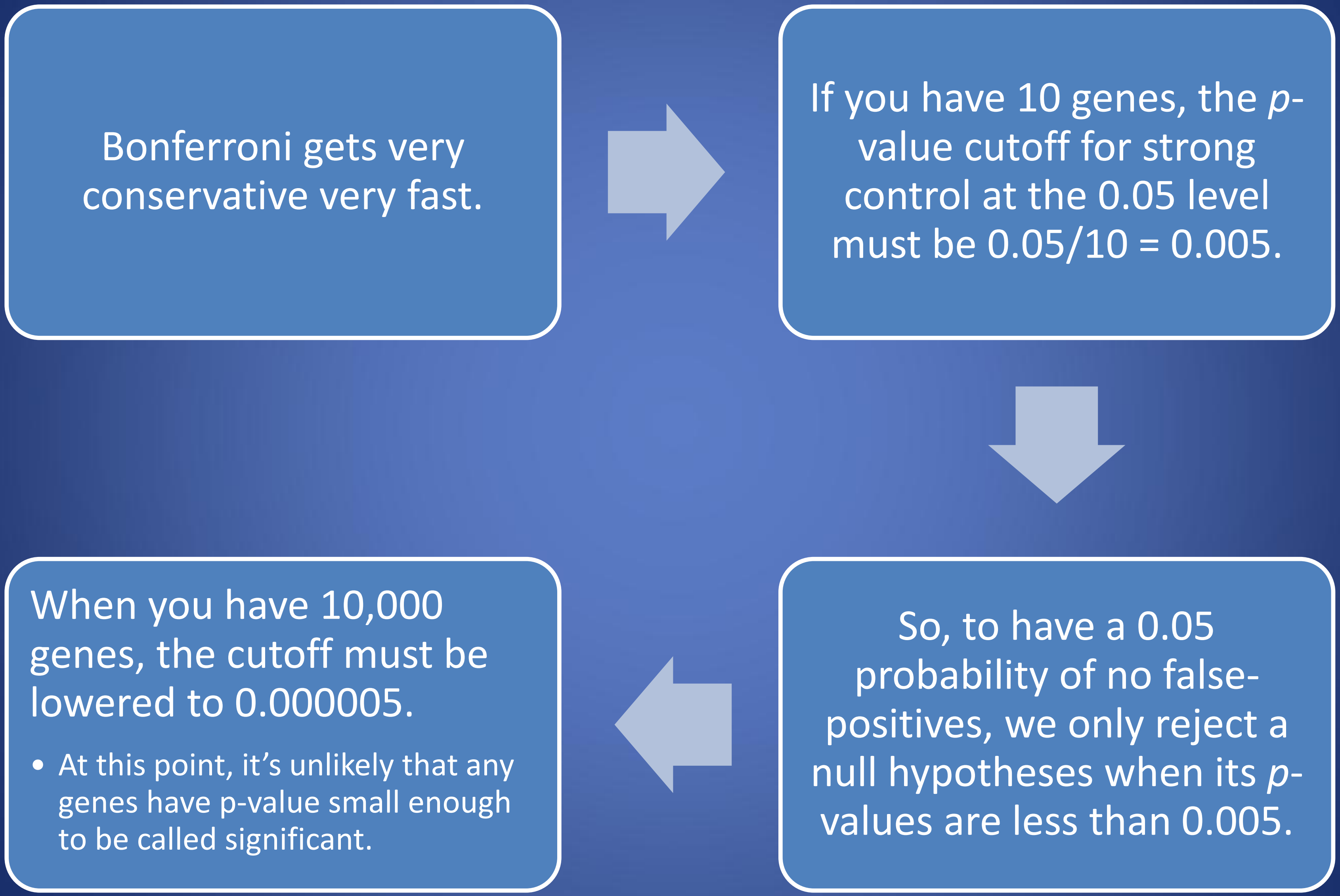
$$1 - 0.98\bar{3}^3 = 0.04917$$

which again sneaks in under 0.05.

The Bonferroni Correction

- This classical method says the following.
- Suppose we perform *N* tests (independent or not), and we want to control the *complete* null hypothesis at the level *C* (*e.g., C*=0.05).
  - Then, the probability that *any* ***true*** null hypotheses have *p*-value $\leq C/N$ is no more than *C*.
  - This means the probability of *any* false-positives is $\leq C$.

Bonferroni



Bonferroni gets very conservative very fast.

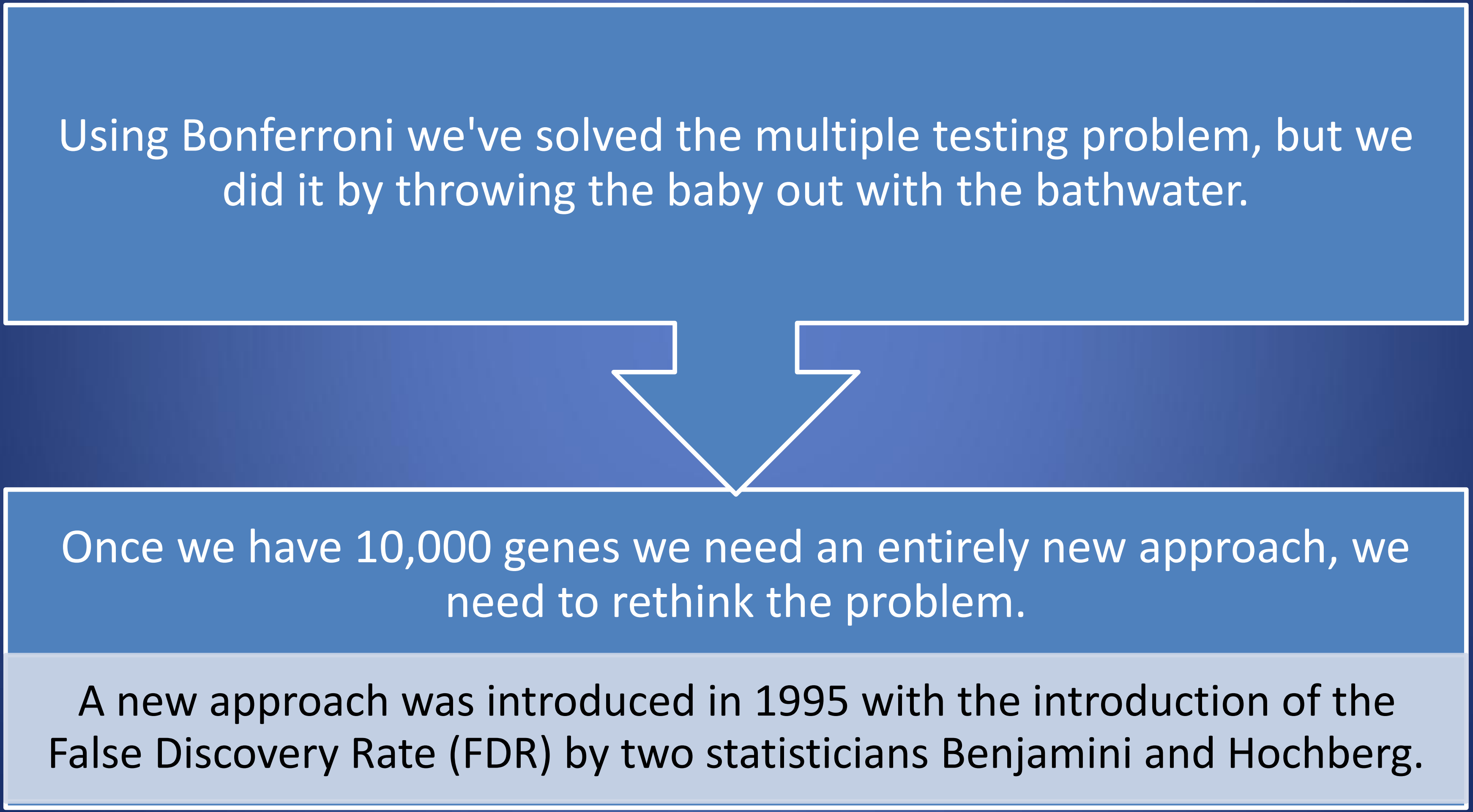
If you have 10 genes, the *p*-value cutoff for strong control at the 0.05 level must be 0.05/10 = 0.005.

So, to have a 0.05 probability of no false-positives, we only reject a null hypotheses when its *p*-values are less than 0.005.

When you have 10,000 genes, the cutoff must be lowered to 0.000005.

- At this point, it's unlikely that any genes have p-value small enough to be called significant.

Strong Control is too Strong



Using Bonferroni we've solved the multiple testing problem, but we did it by throwing the baby out with the bathwater.

Once we have 10,000 genes we need an entirely new approach, we need to rethink the problem.

A new approach was introduced in 1995 with the introduction of the False Discovery Rate (FDR) by two statisticians Benjamini and Hochberg.

FWER Recap

- Bonferroni controls the probability of making any errors at all.
- That is what's known as controlling “The Family-Wise Error Rate” (FWER).
- But the FWER approach is too strong for RNA-Seq DE analysis.
- We don't need to be 95% sure there are *no* false-positives on our list.
- Instead, we just need to keep *the percentage of false-positives* under control.

RNA-Seq

Suppose you are looking for the DE genes between two experimental conditions.

- Suppose there are 10 such genes.
- So, we're looking for 10 DE genes out of 30,000.
- Classic needle-in-haystack problem.

Suppose the bioinformatician can offer the biologist two options.

1. To provide a set of 2 genes that are 95% sure to all be truly DE (so we missed eight).
2. To provide a set of 20 genes that are expected to contain the 10 DE genes plus 10 non-DE genes.

The biologist would probably want both.

- The first is FWER controlled (at the 0.05 level).
- The second is FDR controlled (at the 0.5 level).

We've seen how to do FWER control (use Bonferroni).

- Next, we'll discuss the most popular way to do FDR control.

The False Discovery Proportion

When declaring 100 genes DE, we can tolerate a list with 90 true positives and 10 false positives.

90% true

What we don't want is a list with 90 false positives and only 10 true positives.

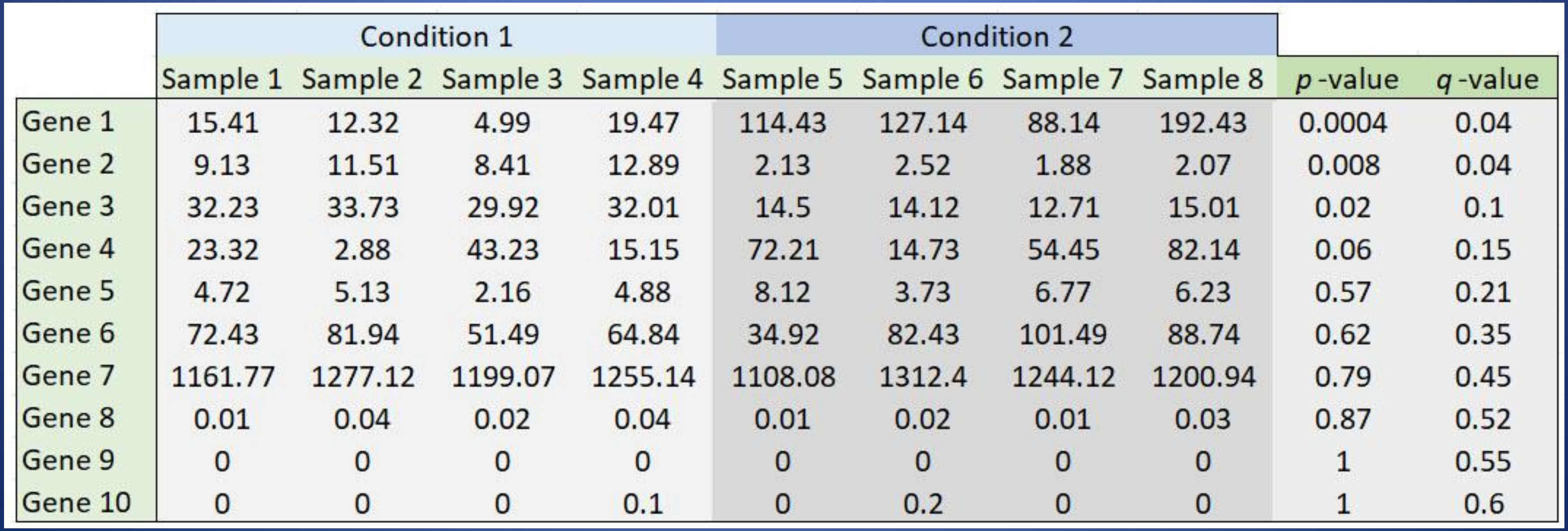
10% true

In the first case the False Discovery Proportion is 0.1, in the second case it’s 0.9.

Controlling this False Discovery Proportion is where *q*-values come in.

q-values

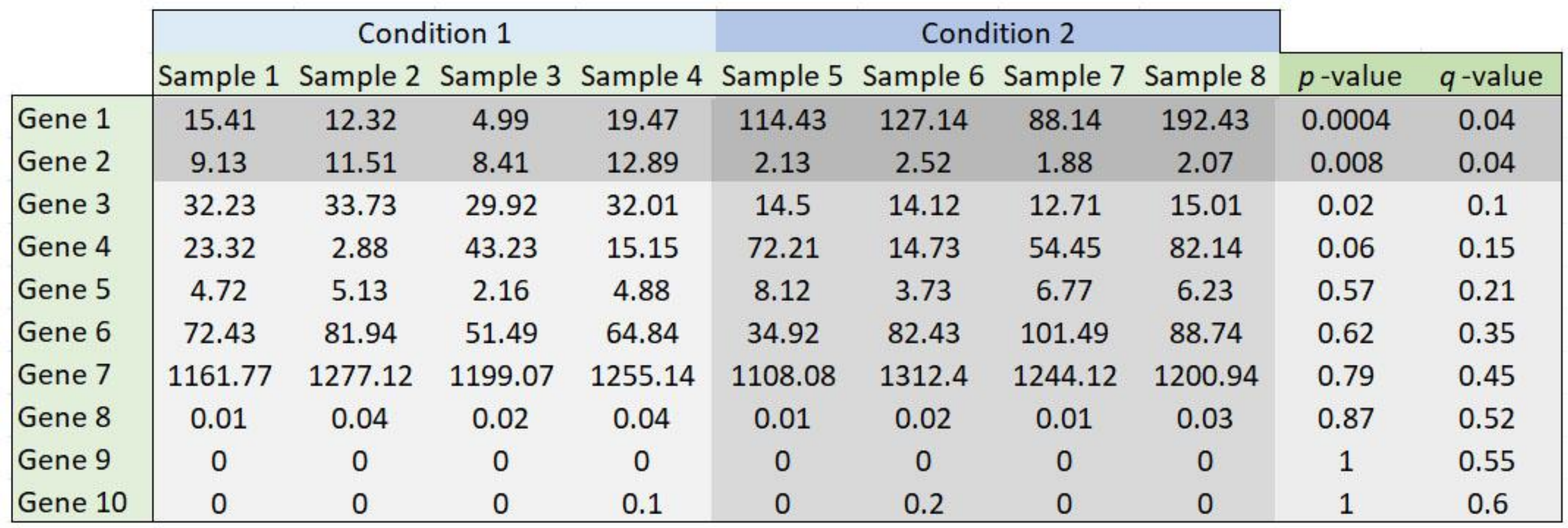
- *q*-values are between 0 and 1 just like *p*-values.
- And just as we get a *p*-value for each gene, there are methods that associate *q*-values to each gene.
- So after the statistical analysis, spreadhseets look like this:



	Condition 1				Condition 2					
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	*p* -value	*q* -value
Gene 1	15.41	12.32	4.99	19.47	114.43	127.14	88.14	192.43	0.0004	0.04
Gene 2	9.13	11.51	8.41	12.89	2.13	2.52	1.88	2.07	0.008	0.04
Gene 3	32.23	33.73	29.92	32.01	14.5	14.12	12.71	15.01	0.02	0.1
Gene 4	23.32	2.88	43.23	15.15	72.21	14.73	54.45	82.14	0.06	0.15
Gene 5	4.72	5.13	2.16	4.88	8.12	3.73	6.77	6.23	0.57	0.21
Gene 6	72.43	81.94	51.49	64.84	34.92	82.43	101.49	88.74	0.62	0.35
Gene 7	1161.77	1277.12	1199.07	1255.14	1108.08	1312.4	1244.12	1200.94	0.79	0.45
Gene 8	0.01	0.04	0.02	0.04	0.01	0.02	0.01	0.03	0.87	0.52
Gene 9	0	0	0	0	0	0	0	0	1	0.55
Gene 10	0	0	0	0.1	0	0.2	0	0	1	0.6

q-values in practice

- If you use a cutoff for the *q*-value of *C*, then the *expected* proportion of false-positives in the set of all significant genes is $\leq C$.



	Condition 1				Condition 2					
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	*p* -value	*q* -value
Gene 1	15.41	12.32	4.99	19.47	114.43	127.14	88.14	192.43	0.0004	0.04
Gene 2	9.13	11.51	8.41	12.89	2.13	2.52	1.88	2.07	0.008	0.04
Gene 3	32.23	33.73	29.92	32.01	14.5	14.12	12.71	15.01	0.02	0.1
Gene 4	23.32	2.88	43.23	15.15	72.21	14.73	54.45	82.14	0.06	0.15
Gene 5	4.72	5.13	2.16	4.88	8.12	3.73	6.77	6.23	0.57	0.21
Gene 6	72.43	81.94	51.49	64.84	34.92	82.43	101.49	88.74	0.62	0.35
Gene 7	1161.77	1277.12	1199.07	1255.14	1108.08	1312.4	1244.12	1200.94	0.79	0.45
Gene 8	0.01	0.04	0.02	0.04	0.01	0.02	0.01	0.03	0.87	0.52
Gene 9	0	0	0	0	0	0	0	0	1	0.55
Gene 10	0	0	0	0.1	0	0.2	0	0	1	0.6

- For example, for the set {Gene1, Gene2} we expect 2 x 0.04 = 0.08 false positives.
  - So considerably less than one. In other words, we expect both genes to be true positives.

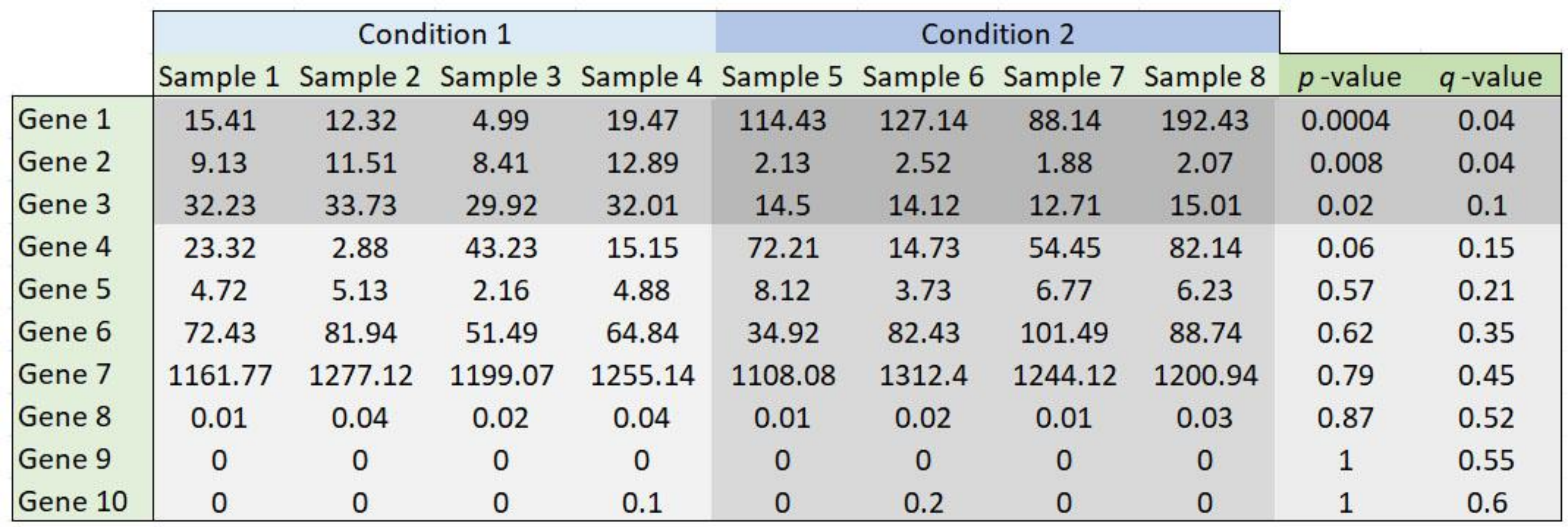
The False Discovery Rate

- In general, the False Discovery Rate (FDR) is the *expected* proportion of false-positives in the set of rejected hypotheses.

- We work with the FDR and not the False Discovery Proportion, because the latter is an unknown quantity.
  - We're just estimating that unknown quantity with the expected value which we can calcluate.

The FDR

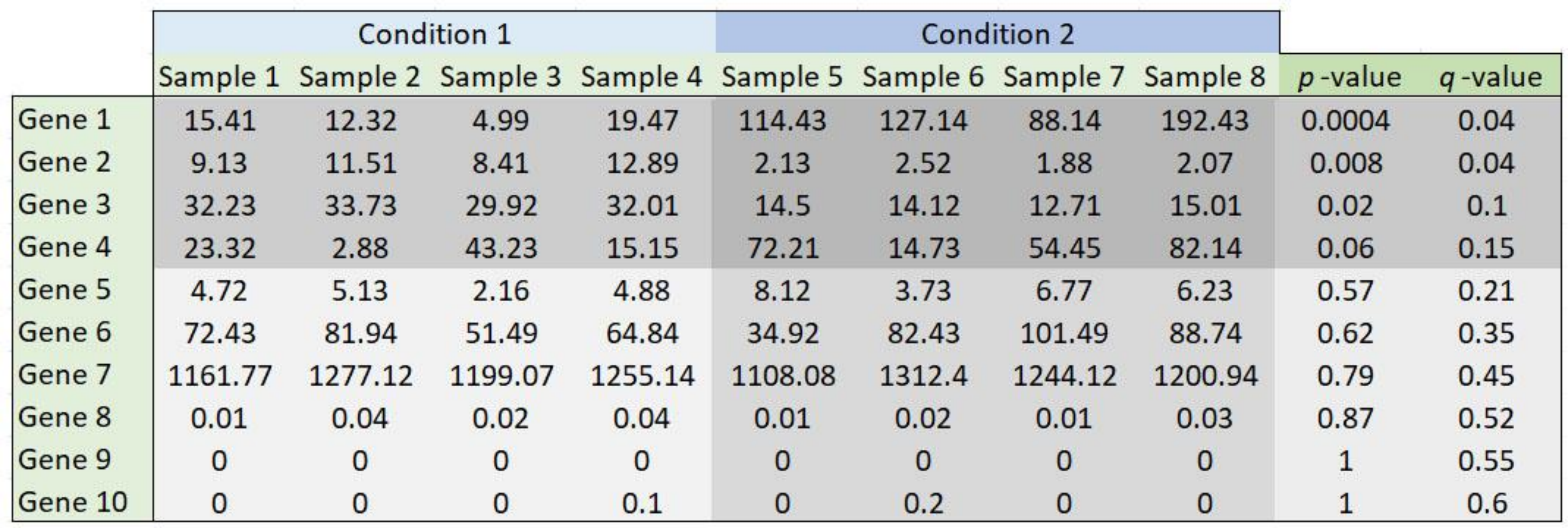
- The set {Gene1, Gene2, Gene3} is based on *q*-value cutoff of 0.1, so the set has an FDR of 0.1.
- Therefore, we expect 3 x 0.1 = 0.3 false positives in the set.
  - Still less than one.



	Condition 1				Condition 2					
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	*p* -value	*q* -value
Gene 1	15.41	12.32	4.99	19.47	114.43	127.14	88.14	192.43	0.0004	0.04
Gene 2	9.13	11.51	8.41	12.89	2.13	2.52	1.88	2.07	0.008	0.04
Gene 3	32.23	33.73	29.92	32.01	14.5	14.12	12.71	15.01	0.02	0.1
Gene 4	23.32	2.88	43.23	15.15	72.21	14.73	54.45	82.14	0.06	0.15
Gene 5	4.72	5.13	2.16	4.88	8.12	3.73	6.77	6.23	0.57	0.21
Gene 6	72.43	81.94	51.49	64.84	34.92	82.43	101.49	88.74	0.62	0.35
Gene 7	1161.77	1277.12	1199.07	1255.14	1108.08	1312.4	1244.12	1200.94	0.79	0.45
Gene 8	0.01	0.04	0.02	0.04	0.01	0.02	0.01	0.03	0.87	0.52
Gene 9	0	0	0	0	0	0	0	0	1	0.55
Gene 10	0	0	0	0.1	0	0.2	0	0	1	0.6

The FDR

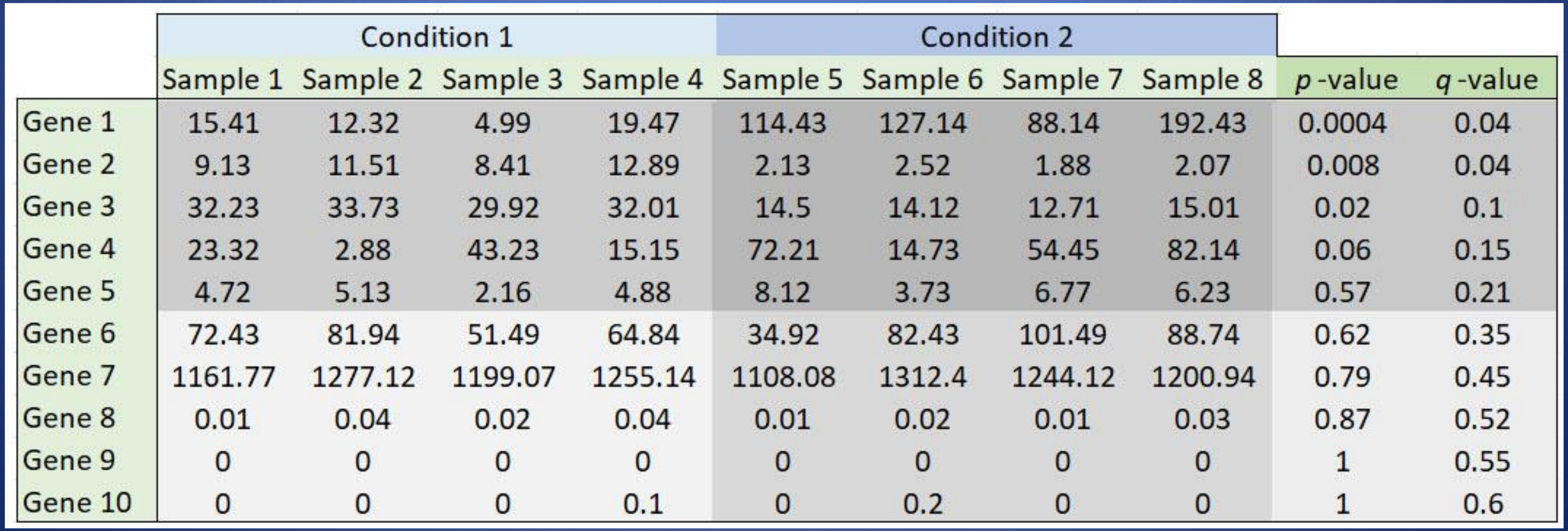
- The set {Gene1, Gene2, Gene3, Gene4} is based on *q*-value cutoff of 0.15, so the set has an FDR of 0.15.
- Therefore, we expect 4 x 0.15 = 0.6 false positives in the set.



	Condition 1				Condition 2					
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	*p* -value	*q* -value
Gene 1	15.41	12.32	4.99	19.47	114.43	127.14	88.14	192.43	0.0004	0.04
Gene 2	9.13	11.51	8.41	12.89	2.13	2.52	1.88	2.07	0.008	0.04
Gene 3	32.23	33.73	29.92	32.01	14.5	14.12	12.71	15.01	0.02	0.1
Gene 4	23.32	2.88	43.23	15.15	72.21	14.73	54.45	82.14	0.06	0.15
Gene 5	4.72	5.13	2.16	4.88	8.12	3.73	6.77	6.23	0.57	0.21
Gene 6	72.43	81.94	51.49	64.84	34.92	82.43	101.49	88.74	0.62	0.35
Gene 7	1161.77	1277.12	1199.07	1255.14	1108.08	1312.4	1244.12	1200.94	0.79	0.45
Gene 8	0.01	0.04	0.02	0.04	0.01	0.02	0.01	0.03	0.87	0.52
Gene 9	0	0	0	0	0	0	0	0	1	0.55
Gene 10	0	0	0	0.1	0	0.2	0	0	1	0.6

The FDR

- The set {Gene1, Gene2, Gene3, Gene4, Gene5} is based on *q*-value cutoff of 0.21, so the set has an FDR of 0.21.
- Therefore, we expect 5 x 0.21 = 1.05 false positives in the set.

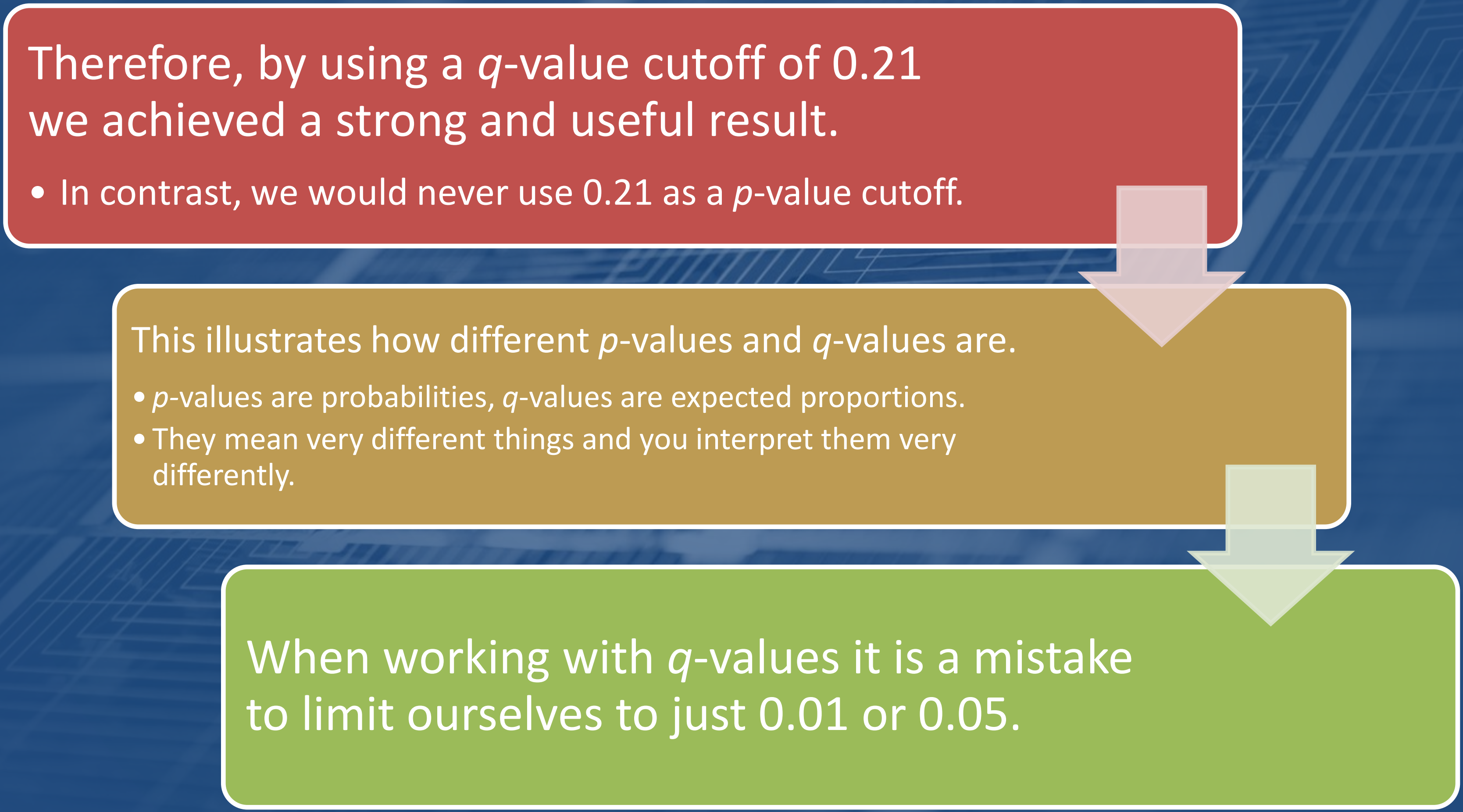


	Condition 1				Condition 2					
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	*p* -value	*q* -value
Gene 1	15.41	12.32	4.99	19.47	114.43	127.14	88.14	192.43	0.0004	0.04
Gene 2	9.13	11.51	8.41	12.89	2.13	2.52	1.88	2.07	0.008	0.04
Gene 3	32.23	33.73	29.92	32.01	14.5	14.12	12.71	15.01	0.02	0.1
Gene 4	23.32	2.88	43.23	15.15	72.21	14.73	54.45	82.14	0.06	0.15
Gene 5	4.72	5.13	2.16	4.88	8.12	3.73	6.77	6.23	0.57	0.21
Gene 6	72.43	81.94	51.49	64.84	34.92	82.43	101.49	88.74	0.62	0.35
Gene 7	1161.77	1277.12	1199.07	1255.14	1108.08	1312.4	1244.12	1200.94	0.79	0.45
Gene 8	0.01	0.04	0.02	0.04	0.01	0.02	0.01	0.03	0.87	0.52
Gene 9	0	0	0	0	0	0	0	0	1	0.55
Gene 10	0	0	0	0.1	0	0.2	0	0	1	0.6

q-values vs. *p*-values

- Suppose the biologists has 30,000 genes and wants to find which of them are DE.
- Suppose the underlying truth is that 4 of them are actually DE.
  - Initially, these four are like four needles in a haystack, lost among the 29,996 non-DE genes.
- Suppose the bioinformatician can deliver the biologist a set of 5 genes that contain the 4 DE genes.
- The biologist can then PCR validate these five to find out which are the 4 truly DE genes.

q-values vs. *p*-values



Therefore, by using a *q*-value cutoff of 0.21 we achieved a strong and useful result.

- In contrast, we would never use 0.21 as a *p*-value cutoff.

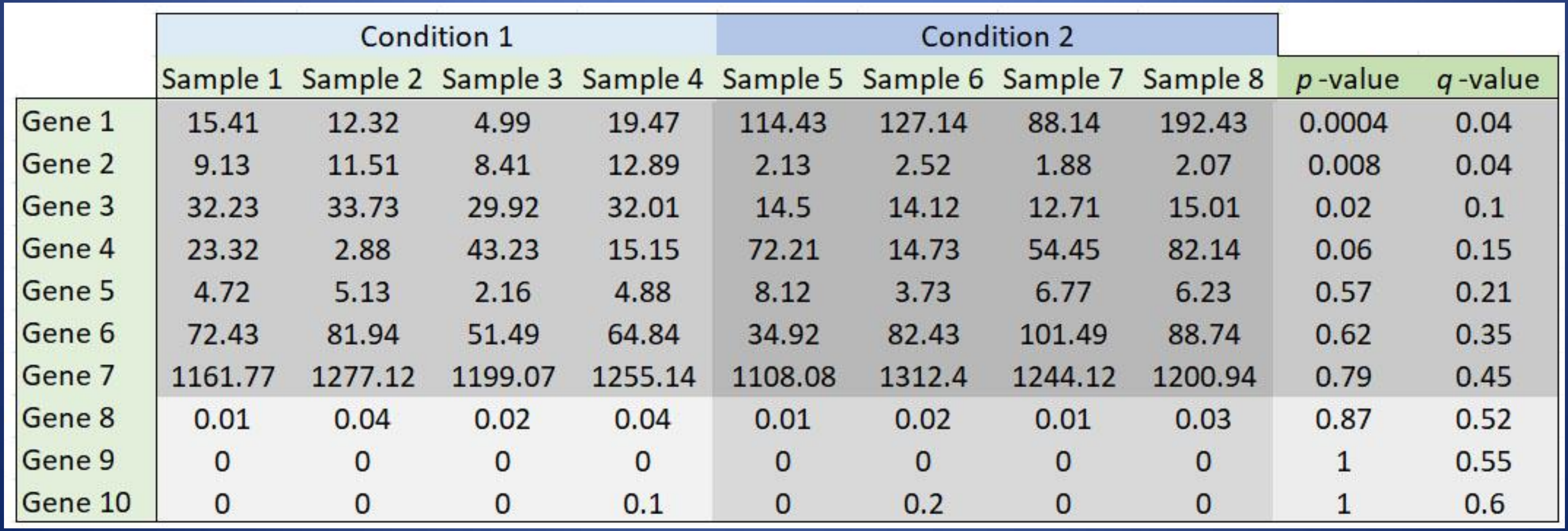
This illustrates how different *p*-values and *q*-values are.

- *p*-values are probabilities, *q*-values are expected proportions.
- They mean very different things and you interpret them very differently.

When working with *q*-values it is a mistake to limit ourselves to just 0.01 or 0.05.

The FDR

- The set {Gene1, Gene2, Gene3, Gene4, Gene5, Gene6, Gene7} is based on *q*-value cutoff of 0.45, so the set has an FDR of 0.45.
- Therefore, we expect 7 x 0.45 = 3.15 false positives in the set. So approximately 4 true positives and 3 false positives.



	Condition 1				Condition 2					
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	*p* -value	*q* -value
Gene 1	15.41	12.32	4.99	19.47	114.43	127.14	88.14	192.43	0.0004	0.04
Gene 2	9.13	11.51	8.41	12.89	2.13	2.52	1.88	2.07	0.008	0.04
Gene 3	32.23	33.73	29.92	32.01	14.5	14.12	12.71	15.01	0.02	0.1
Gene 4	23.32	2.88	43.23	15.15	72.21	14.73	54.45	82.14	0.06	0.15
Gene 5	4.72	5.13	2.16	4.88	8.12	3.73	6.77	6.23	0.57	0.21
Gene 6	72.43	81.94	51.49	64.84	34.92	82.43	101.49	88.74	0.62	0.35
Gene 7	1161.77	1277.12	1199.07	1255.14	1108.08	1312.4	1244.12	1200.94	0.79	0.45
Gene 8	0.01	0.04	0.02	0.04	0.01	0.02	0.01	0.03	0.87	0.52
Gene 9	0	0	0	0	0	0	0	0	1	0.55
Gene 10	0	0	0	0.1	0	0.2	0	0	1	0.6

In Practice

- We typically would not be using *q*-values to work with 10 genes like in the previous examples.
- *q*-values are used for RNA-Seq analyses with 30,000 genes.
- We use them also in many other data types and analyses.
  - Whenever there is a massive number of tests and a small proportion of the null hypotheses being null.

Allowing Mistakes - Validation -

- Another way to think of the FDR is as part of a validation approach; similar to the AIDS problem.

- Suppose for example that you are testing 30,000 genes for differential expression by RNA-Seq.

- *If* you can measure as many replicates as you want, then you can determine which are differentially expressed with high confidence.

- But in RNA-Seq nobody can afford to do “as many replicates as they want”.

- Suppose you cannot afford to do more than 3 replicates per group.

Allowing Mistakes The “validation” approach

- If we do a first pass with three replicates and find five genes, four of which are expected to be true positives (so FDR = 0.2).
  - Then we can perform more replicates by PCR *just of those five genes* to identify the false positive.
  - We’re assuming samples are cheap, PCR is cheap, but performing a full RNA-Seq assay of a sample is expensive.

- In the end the one false positive caused no harm.

- The trick is in finding the most powerful method to get from 30,000 down to a validation set with a small *proportion* of false positives.

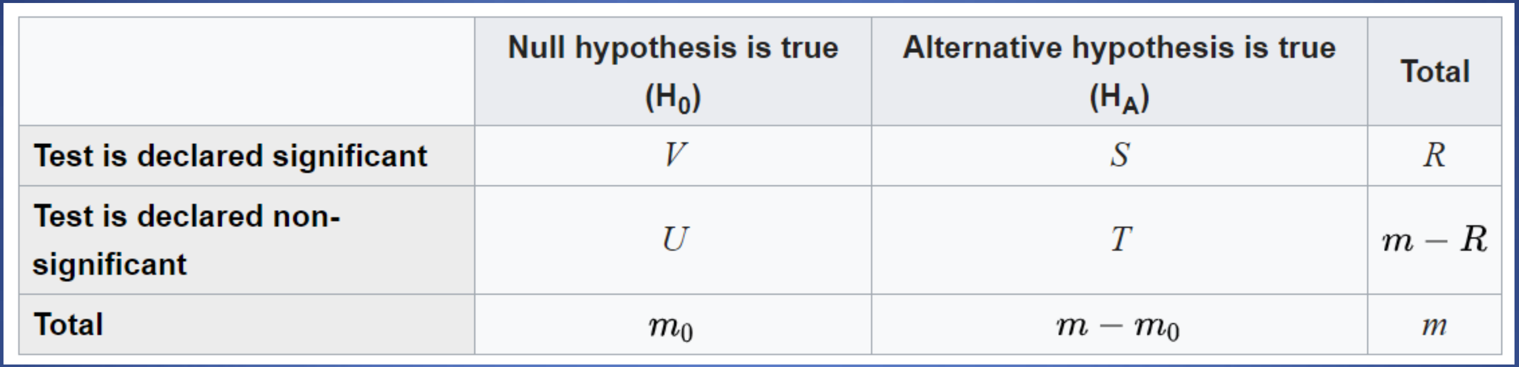
False Discovery Rate Subtlety

- Individual tests have false-***positive*** rates (*p*-values).
- But individual tests do not have false-***discovery*** rates.
  - A False Discovery Rates is about a *set* of tests, not a single test.
- We use *q*-values to determine sets of genes.
  - It's those sets to which we associate an FDR.

p-values are local
q-values are global

- Suppose we collect samples to perform an RNA-Seq Differential Expression (DE) analysis.
- Let *G* be a gene in the genome.
- Every gene will have a *p*-value.
- If the data change for other genes besides *G*, in theory the *p*-value for *G* will remain the same.
  - The T-Test and its *p*-value only depends on the data for *G*.
- In contrast, it could change *G*'s *q*-value.
  - That's because the *q*-value of any given gene depends globally on all other genes.
- Just like in the AIDS example. The confidence in a test for a single person depends on the population.

Mathematical Definition of the FDR



	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_A)	Total
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

FDR = $E[V/R \mid R>0] * \Pr(R>0)$

The FDR is associated to a *set* of hypotheses not a single hypothesis.

Algorithms and Assumptions

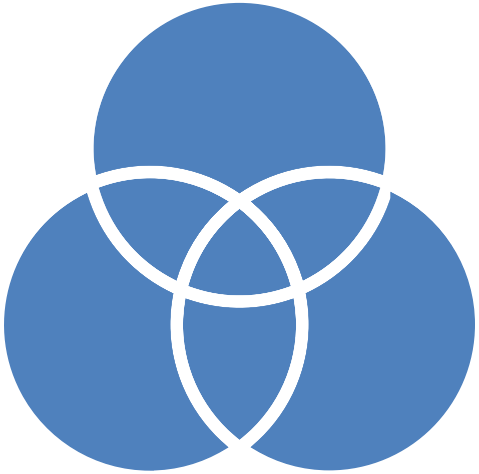
- Most of the algorithms for generating *q*-values start by generating the column of *p*-values. They then adjust (correct) those *p*-values to become *q*-values.

- The most popular is called Benjamini-Hochberg (B-H).
  – Be careful not to confuse “Benjamini-Hochberg” with “Bonferroni”.

Benjamini Hochberg

- This is the original method devised to control the FDR.
- It was first proposed in 1995.
- It works as follows:
- Suppose there are m (independent) null hypotheses.
- Suppose we want an FDR no greater than α.
  1. Compute p-values for each null hypothesis.
  2. Rank the p-values smallest to largest.
  3. Find the largest number k such that the k-th p-value on the ranked list is less than $k\alpha/m$.
  4. Reject the top k hypotheses on the ranked list.
- Then the FDR of these k hypothesis is at most α.
- *This is very far from obvious, and we won’t try to prove it.*

Limitations of Benjamini-Hochberg



- The BH method has its drawbacks.
- It assumes independent hypotheses.
  - Bonferroni, on the other hand, does not.
- It is only as good as the individual *p*-values.
- If the *p*-values are not reliable
  - For example, a *T*-test on data with only two replicates per group will have a lot of false negatives and BH will not fix that.

The BH method will inherit these issues and will also be underpowered.

q-values

- The *q*-value allows us to associate a level of confidence to individual hypotheses, as opposed to sets of hypotheses.

- The *q*-value for an observed value of the statistic *T* is:
  - $\min_{0<C<t}$ (Prob(H_0 True | $T > C$))
  - Don't worry about wrapping your head around this definition.
  - The point is, the set of hypotheses with *q*-value less than α has FDR $\leq$ α

Application of RNA-Seq

- The goal is to find the DE genes between two experimental conditions.
- We have seen how to get from raw data to a spreadsheet of expression values.
- There are several ways to normalize and generate *q*-values for each gene.
- This is a job for R and we will come back to it when we do R in a couple weeks.

Standard DE methods

- Most people use one of DESeq, edgeR or limma-Voom.
- These methods all produce *p*-values which are then transformed into *q*-values using Benjamini-Hochberg.
- Therefore, all methods are assuming independence between genes at some point or another.
  - This is done commonly, in spite of it being a strong assumption. People are always working on improvements.