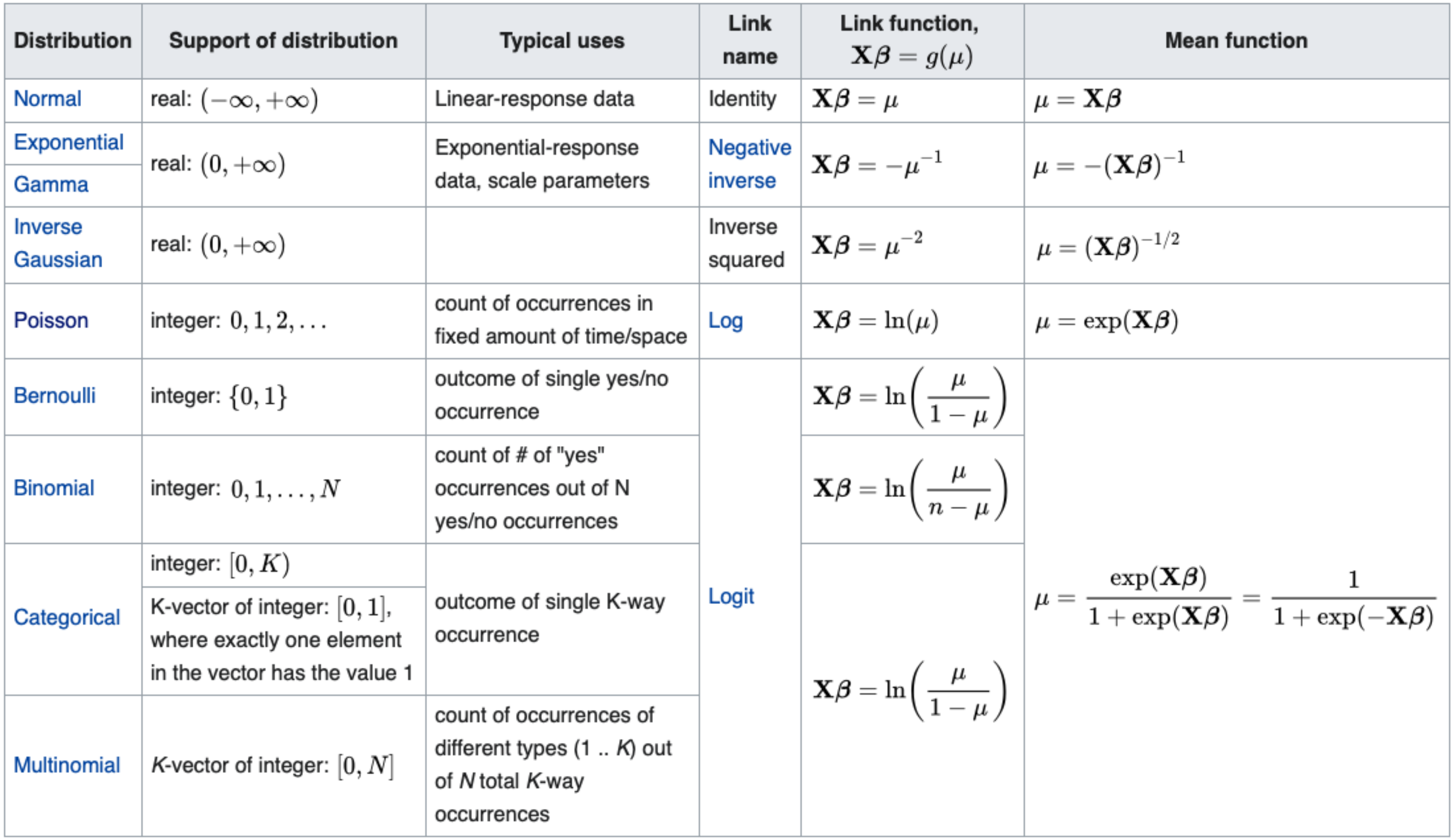
DGE (part2)

RNA-seq data analysis

Paulo Czarnewski

What is a GLM?

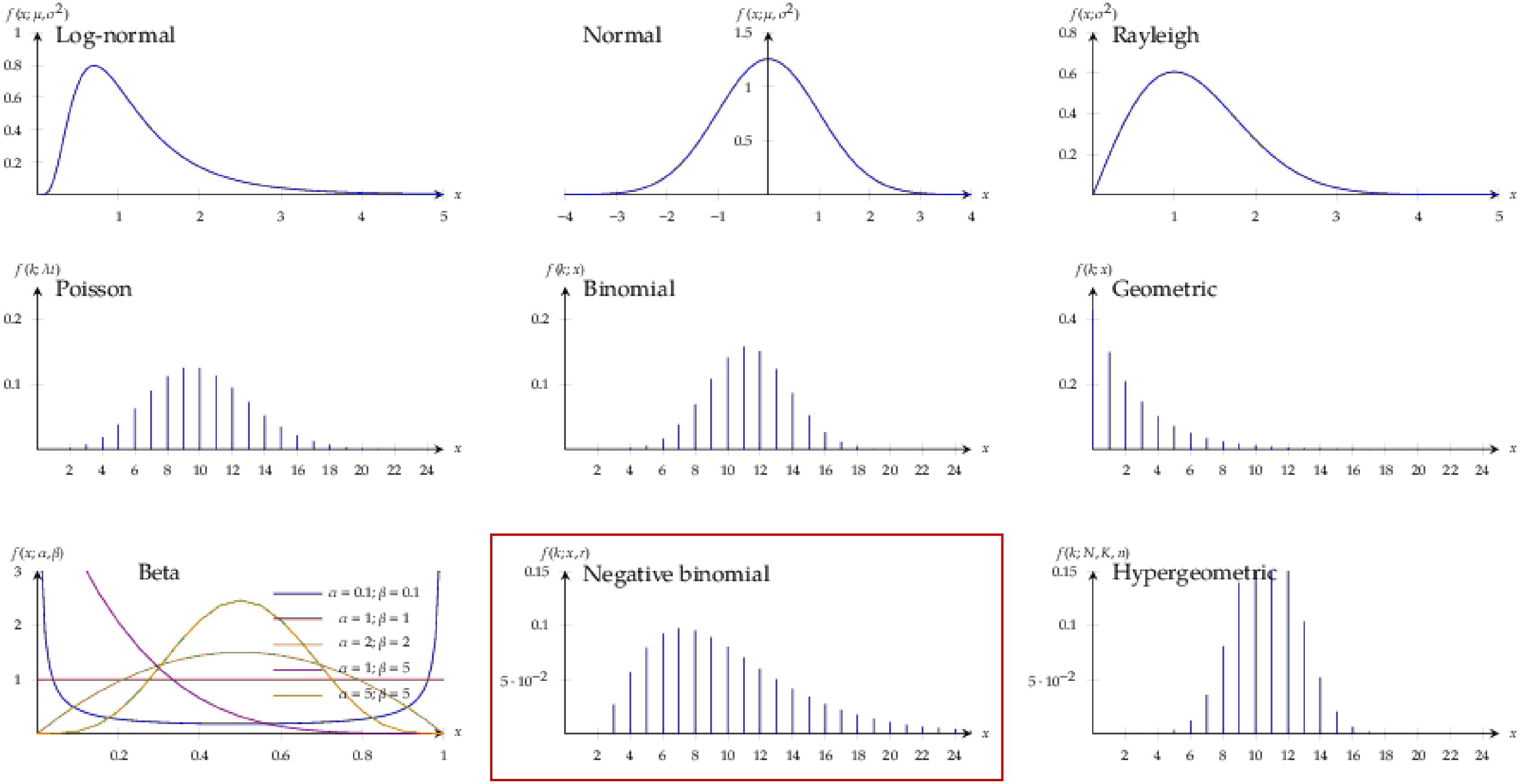
Generalized linear models (GLM) is a **flexible** generalization of ordinary *linear regression* that allows for response variables that have error distribution models other than a normal distribution.



Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model

GLM Distributions



DESeq2 and **EdgeR** are improved negative-binomial GLMs

https://en.wikipedia.org/wiki/Generalized_linear_model

Neg.Binomial vs Poisson Distributions

ψ is small
i.e. small sample size
i.e. low-count genes

ψ is large
i.e. large sample size
i.e. high-count genes

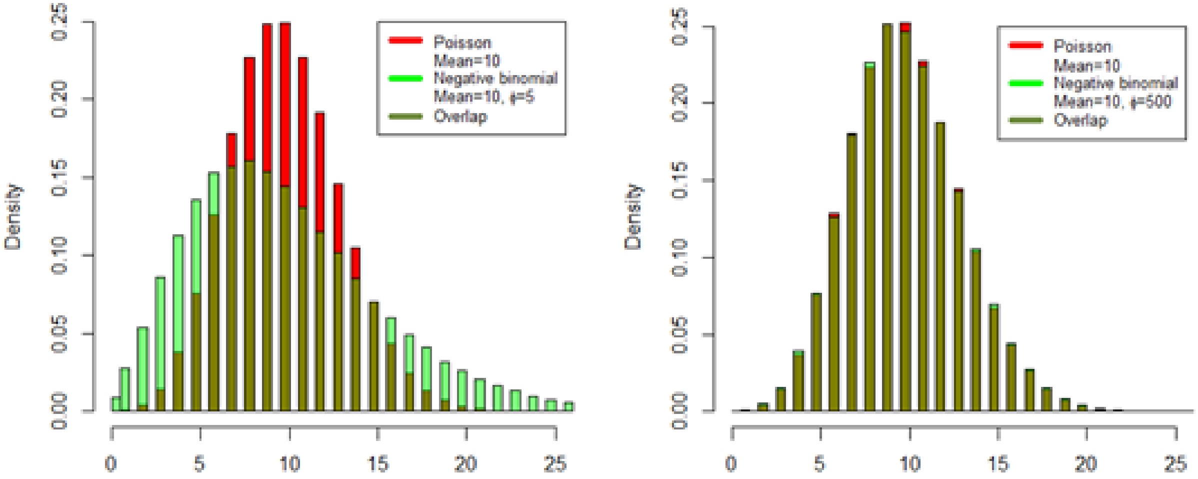


Figure shows that when ψ is small (e.g., ψ =5), a negative binomial distribution is more spread than a Poisson distribution with the same mean

"in case overdispersion exists, Poisson regression model might not be appropriate."

The negative binomial distribution will converge to a Poisson distribution for large ψ.

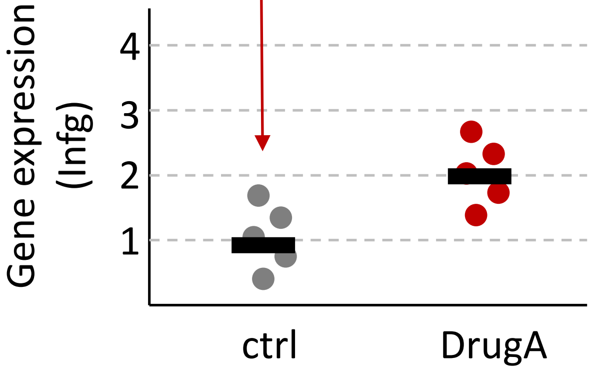
https://pulmonarychronicles.com/index.php/pulmonarychronicles/article/view/200/491

GLM intuition

What if I have 2 groups?

```
metadata$Drug <- factor( metadata$Drug ,
                 levels = c( "ctrl" , "DrugA" ) )
```

intercept group
1st level of factor



Comparison between groups

```
y ~ Drug
```

gene	DrugA
⋮	⋮
Infg	1
⋮	⋮

effect sizes / FC / logFC

```
y ~ 0 + Drug
```

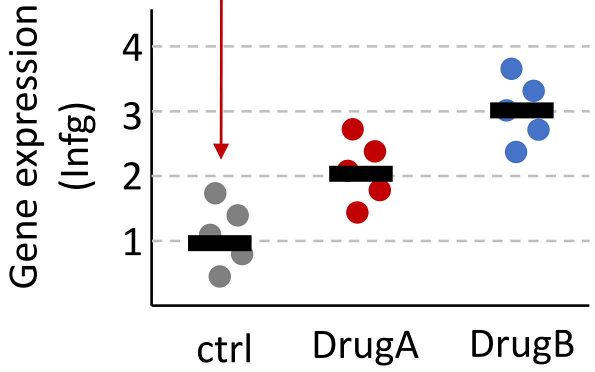
gene	ctrl	DrugA
⋮	⋮	⋮
Infg	1	2
⋮	⋮	⋮

Also testing if base expression is different than zero (not common)

What if I have 3 groups?

```
metadata$Drug <- factor( metadata$Drug ,
                         levels = c( “ctrl” , ”DrugA” , “DrugB” ) )
```

intercept group
1st level of factor



Comparison between groups

```
y ~ Drug
```

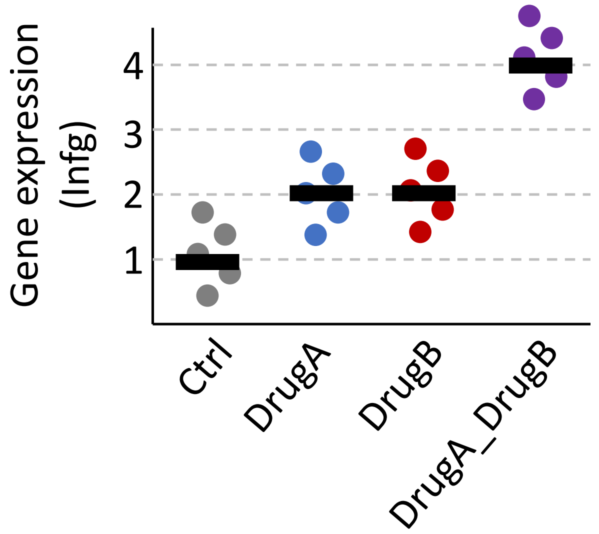
gene	DrugA	DrugB	pvalue
⋮	⋮	⋮	⋮
Infg	1	2	0.0001
⋮	⋮	⋮	⋮

Testing if the gene is significant
in any of the conditions listed.

What if I have 2 variable groups?

What you should **avoid** doing (whenever possible):

```
metadata$Drug <- factor( metadata$Drug ,
                         levels = c( “ctrl” , ”DrugA” , “DrugB” , “DrugA_DrugB”) )
```



```
y ~ Drug
```

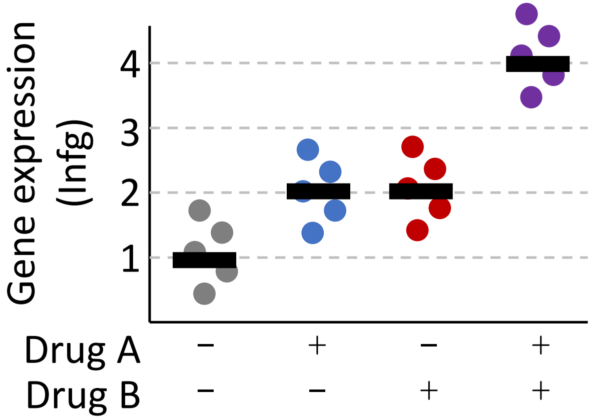
gene	DrugA	DrugB	DrugA_DrugB	pvalue
⋮	⋮	⋮	⋮	⋮
Infg	1	1	3	0.0001
⋮	⋮	⋮	⋮	⋮

What if I have 2 variable groups?

What you should do **instead** (whenever possible):

```
metadata$DrugA <- factor( metadata$DrugA ,
                          levels = c( "ctrl" , "DrugA" ) )

metadata$DrugB <- factor( metadata$DrugB ,
                          levels = c( "ctrl" , "DrugB" ) )
```



```
y ~ DrugA + DrugB
```

gene	DrugA	DrugB
⋮	⋮	⋮
Infg	1.5	1.5
⋮	⋮	⋮

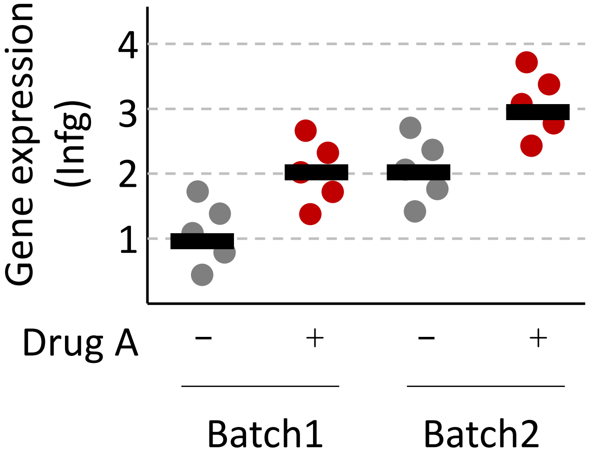
```
y ~ DrugA + DrugB + DrugA:DrugB
```

```
y ~ DrugA * DrugB
```

gene	DrugA	DrugB	DrugA:DrugB
⋮	⋮	⋮	⋮
Infg	1	1	1
⋮	⋮	⋮	⋮

Interaction effect

What if I have a batch effect?



```
y ~ Batch + DrugA
```

```
coef = 1:2
```

gene	Batch2	DrugA	pvalue
⋮	⋮	⋮	⋮
Infg	1	1	0.0001
⋮	⋮	⋮	⋮

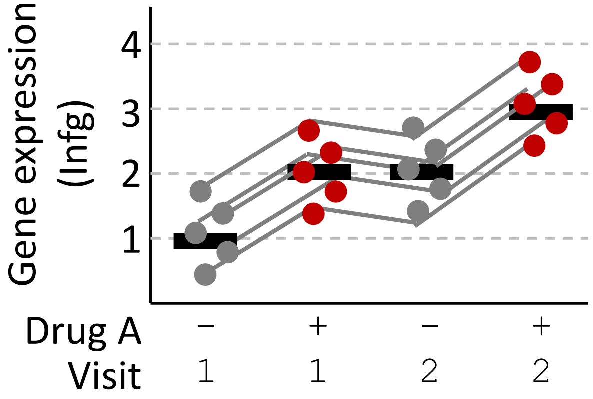
```
y ~ Batch + DrugA
```

```
coef = 2
```

gene	DrugA	pvalue
⋮	⋮	⋮
Infg	1	0.001
⋮	⋮	⋮

P-values will be different because you are testing different hypothesis!

What if I have a individual-matched samples, plus a Drug treatment in two clinical visits?



```
y ~ Patient + Visit + DrugA
```

gene	P2	P3	P4	P5	Visit2	DrugA	pvalue
⋮					⋮	⋮	⋮
Infg	.7	.5	.3	1	1	1	0.0001
⋮					⋮	⋮	⋮

```
y ~ Patient + Visit + DrugA
```

```
coef = 6
```

gene	DrugA	pvalue
⋮	⋮	⋮
Infg	1	0.001
⋮	⋮	⋮

What if I have time series (or other continuous)?

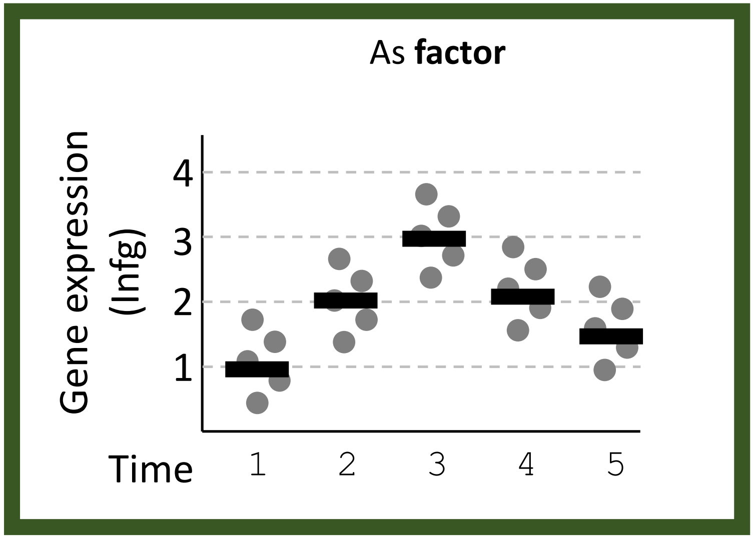
```
y ~ Time
```

IMPORTANT: set your **`Time`** variable as **factor** , so they are treated as categorical groups! e.g. by adding a string in the beginning

```
“day00” “day02” “day04” “day06” ...
```

Instead of :

```
0 2 4 6 ...
```



gene	day2	day3	day4	day5
⋮	⋮	⋮	⋮	⋮
Infg	1	2	1	.5
⋮	⋮	⋮	⋮	⋮

IMPORTANT: Other continuous covariates (such as **patient age** , **exposure time**, etc) should be used as **numeric** if they don’t represent grouping variables.

What if I have time series (or other continuous)?

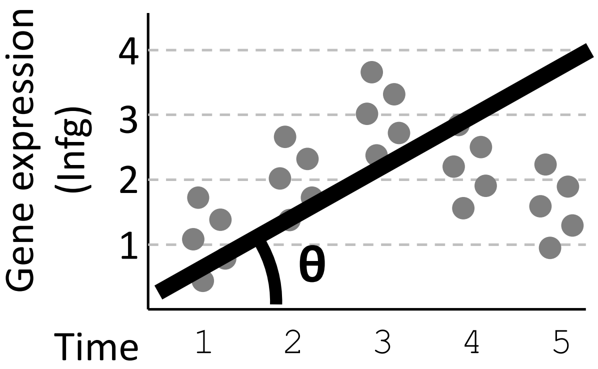
```
y ~ Time
```

IMPORTANT: set your **`Time`** variable as **factor** , so they are treated as categorical groups! e.g. by adding a string in the beginning

```
“day00” “day02” “day04” “day06” ...
```

Instead of :

```
0 2 4 6 ...
```



As **numeric**
= slope θ

gene	**day**
⋮	⋮
Infg	1
⋮	⋮

← Means:
“an increase of 1 per 1 day”

IMPORTANT: Other continuous covariates (such as **patient age** , **exposure time**, etc) should be used as **numeric** if they don’t represent grouping variables.

What if I have time series and a treatment?

```
y ~ Time * Treatment
```

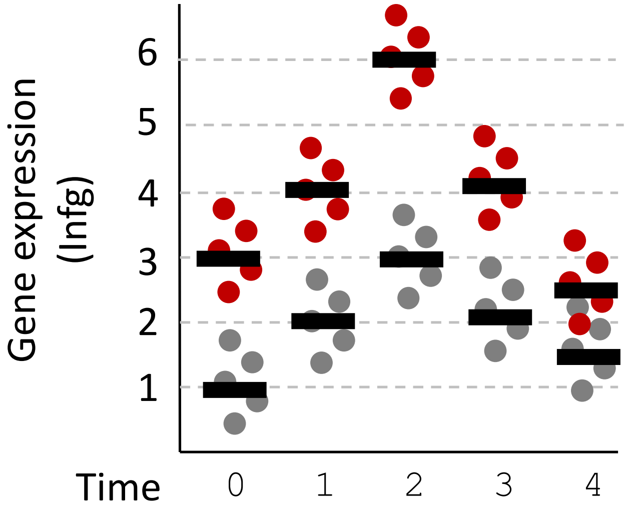
=

```
y ~ Time + Treatment + Time:Treatment
```

Overall DGE at any time point

Overall DGE between conditions

DGE between conditions specific to each time point



gene	1	2	3	4	A	1A	2A	3A	4A	pvalue
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Infg	1	2	1	.5	2	0	1	0	-1	0.0001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

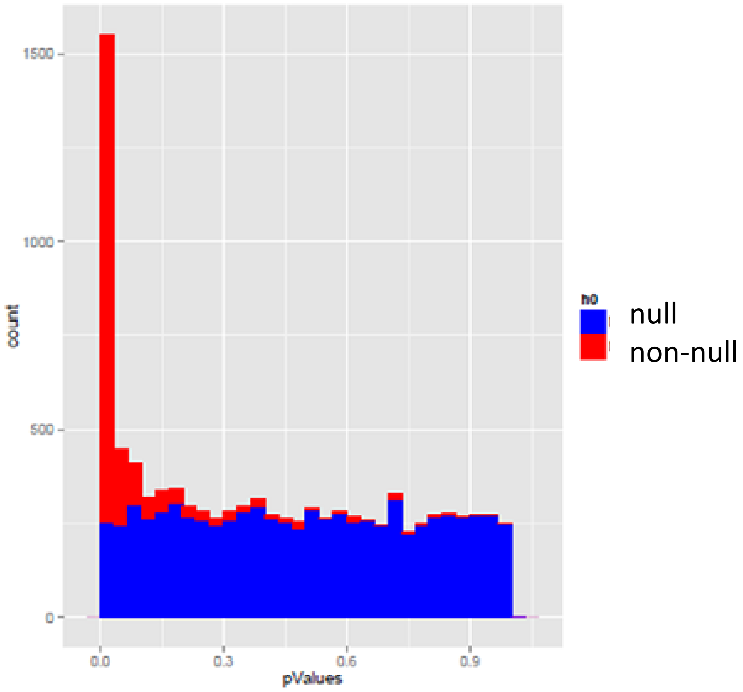
Test for specific contrasts

Testing if the gene is significant in any of the conditions listed.

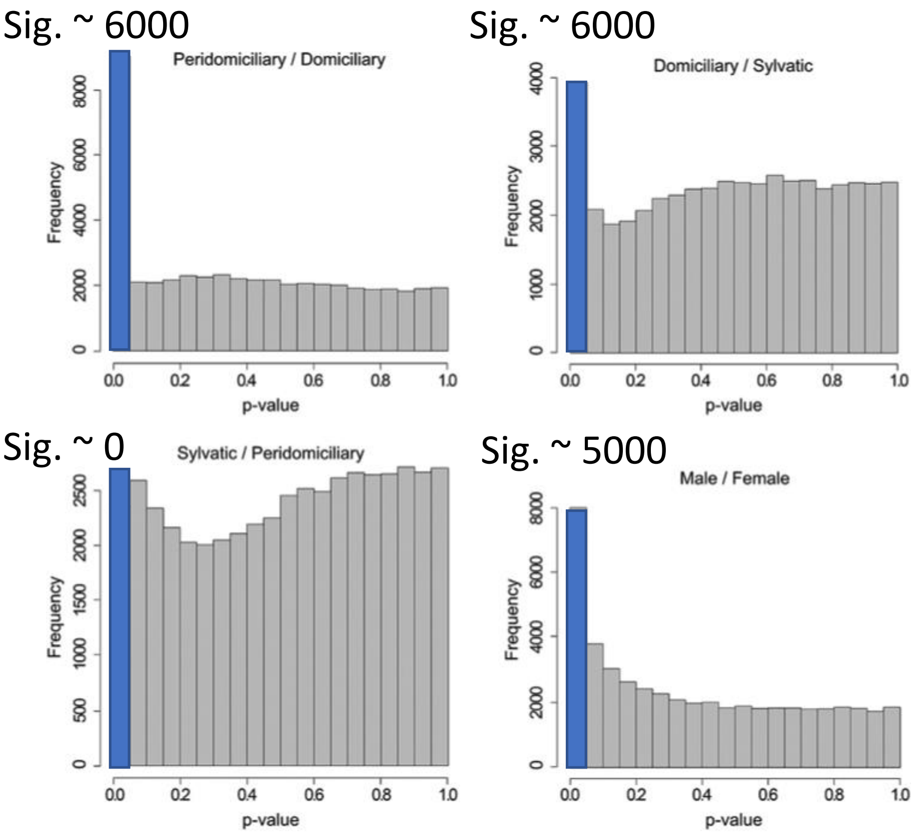
A reminder on the meaning of p-values

A reminder on the meaning of p-values

By chance, at least 5% of of the “significant” (p>0.05) are likely NOT significant (false positives)



https://online.stat.psu.edu/stat555/node/81/


Sig. ~ 6000

Sig. ~ 6000

Sig. ~ 0

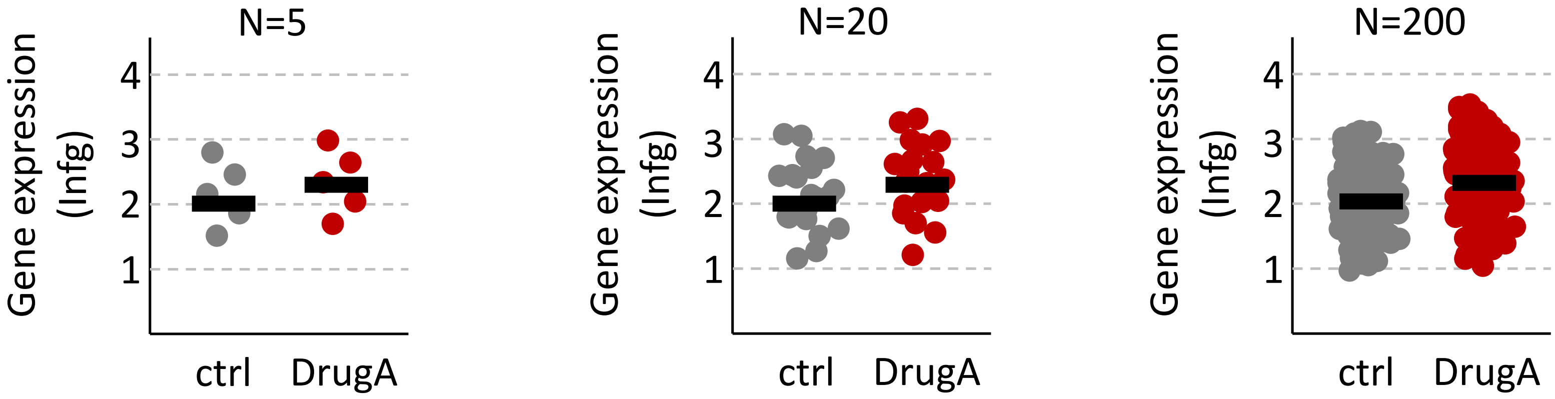
Sig. ~ 5000

That’s why we perform FDR correction on multiple testing, to adjust the p-values so that those 5% do not become significant at a **NULL** hypothesis.

A reminder on the meaning of p-values

p-values represent the confidence you have in your mean measurement, and **NOT** that the groups are different!

```
y ~ Drug
```



gene	DrugA	pvalue
⋮	⋮	⋮
Infg	.2	.1
⋮	⋮	⋮

gene	DrugA	pvalue
⋮	⋮	⋮
Infg	.2	.09
⋮	⋮	⋮

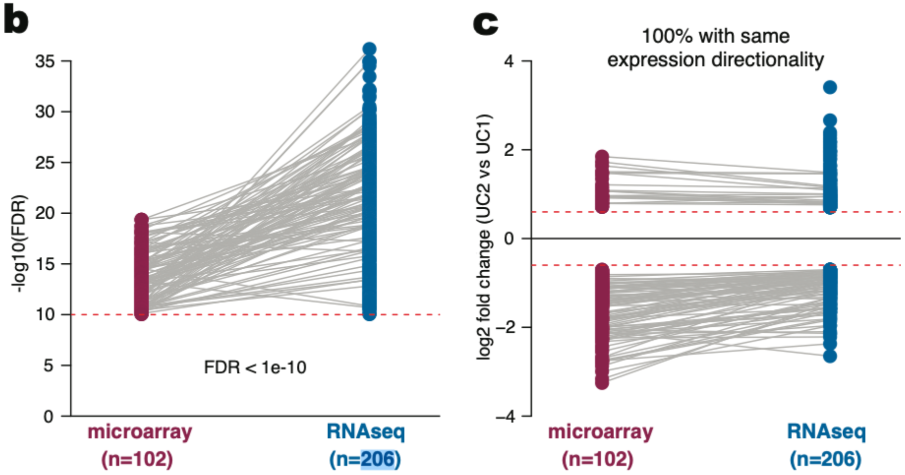
gene	DrugA	pvalue
⋮	⋮	⋮
Infg	.2	.001
⋮	⋮	⋮

That is why we **always** need to take the effect size (logFC) into consideration.
FDR does **NOT** correct for this!

A reminder on the meaning of p-values

p-values represent the confidence you have in your mean measurement, and **NOT** that the groups are different!

p-values become more “significant” as you increase the sample size, but fold changes remain constant



Czarnewski et al (2019) Nat Communications

Thank you. Questions?

Paulo Czarnewski