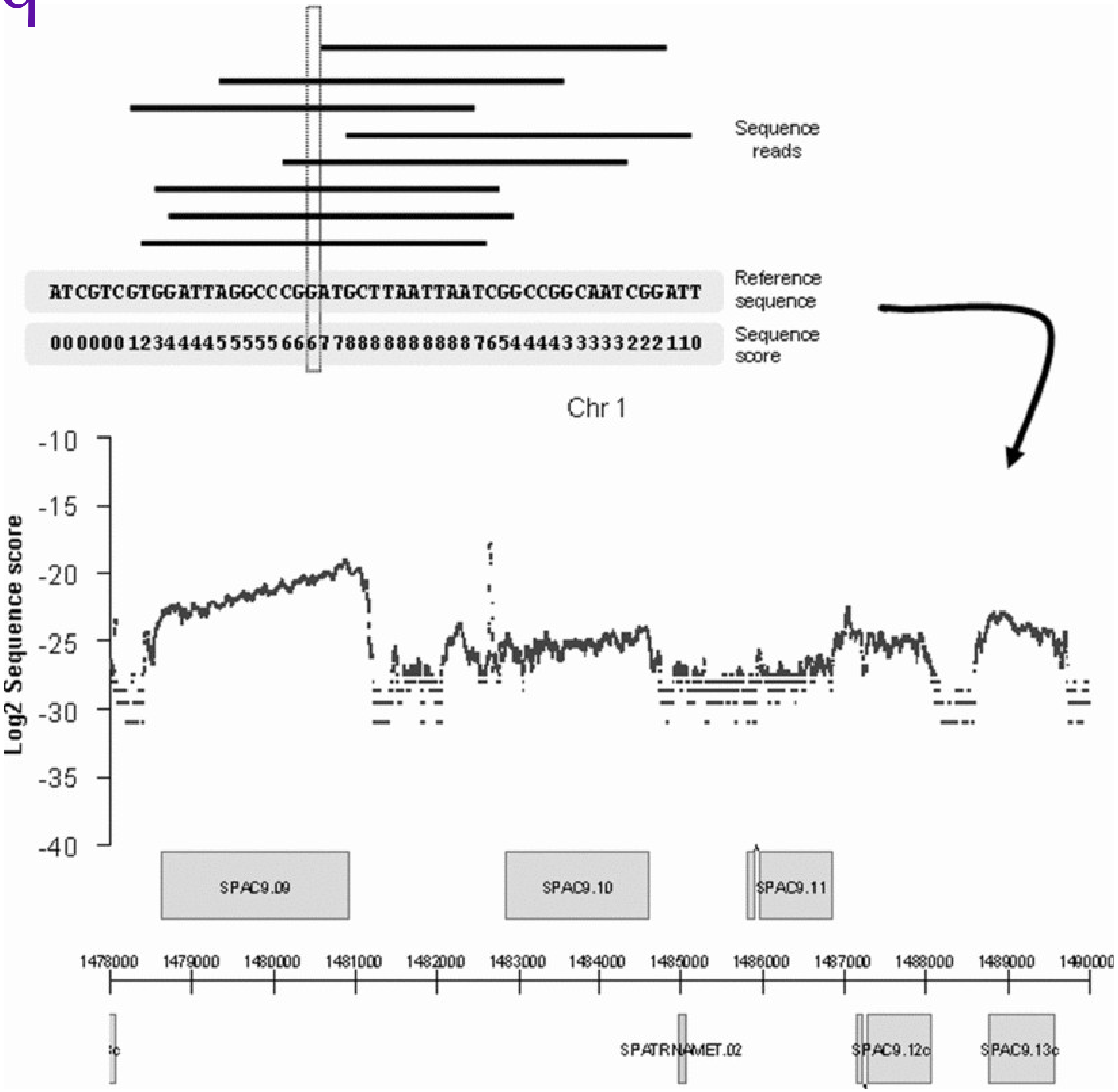
Differential expression analysis for sequencing count data

Simon Anders

EMBL

RNA-Seq


Sequence reads
ATCGTCGTGGATTAGGCCCGGATGCTTAATTAATCGGCCGGCAATCGGATT
Reference sequence
000000123444455555666778888888888765444433333222110
Sequence score
Chr 1
Log2 Sequence score
-10
-15
-20
-25
-30
-35
-40
SPAC9.09
SPAC9.10
SPAC9.11
1478000
1479000
1480000
1481000
1482000
1483000
1484000
1485000
1486000
1487000
1488000
1489000
1490000
SPATRNAMET.02
SPAC9.12c
SPAC9.13c

Count data in HTS

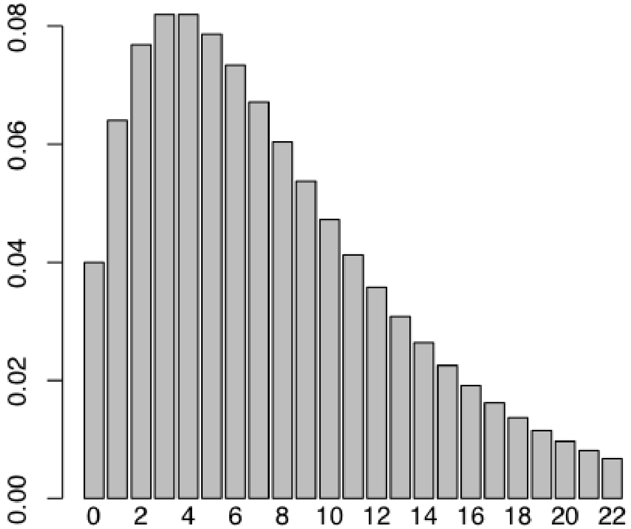
- RNA-Seq
- Tag-Seq

Gene	GliNS1	G144	G166	G179	CB541	CB660
13CDNA73	4	0	6	1	0	5
A2BP1	19	18	20	7	1	8
A2M	2724	2209	13	49	193	548
A4GALT	0	0	48	0	0	0
AAAS	57	29	224	49	202	92
AACS	1904	1294	5073	5365	3737	3511
AADACL1	3	13	239	683	158	40
[...]						

- ChIP-Seq
- Bar-Seq
- ...

Challenges with count data from HTS

discrete, positive, skewed
➡ no (log-)normal model



small numbers of replicates
➡ no rank based or permutation methods

sequencing depth (coverage) varies between samples
➡ "normalisation"

large dynamic range (0 ... 10^5) between genes
➡ heteroskedasticity matters

EMBL

Bioconductor packages

Bioconductor packages for testing for differential signal in sequencing count data:

- Based on negative-binomial distribution:
  - edgeR (Robinson, Mcarthy, Smyth)
  - DESeq (Anders, Huber)
  - BaySeq (Hardcastle, Kelly)

- Based on Poisson distribution:
  - DEGSeq (Wang et al.)

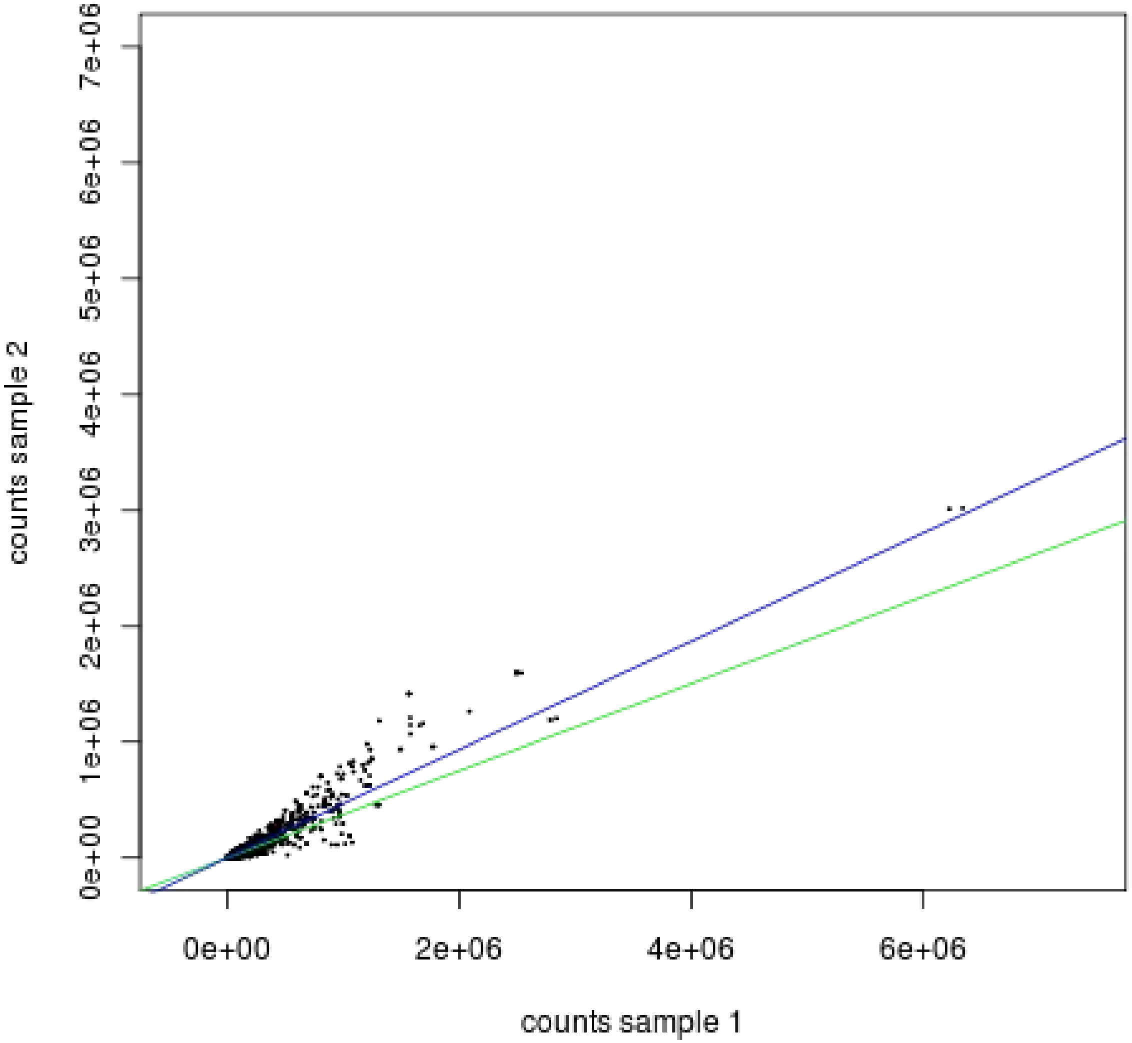
Normalisation for library size

- If sample A has been sampled deeper than sample B, we expect counts to be higher.

- Simply using the total number of reads per sample is not a good idea; genes that are strongly and differentially expressed may distort the ratio of total reads.

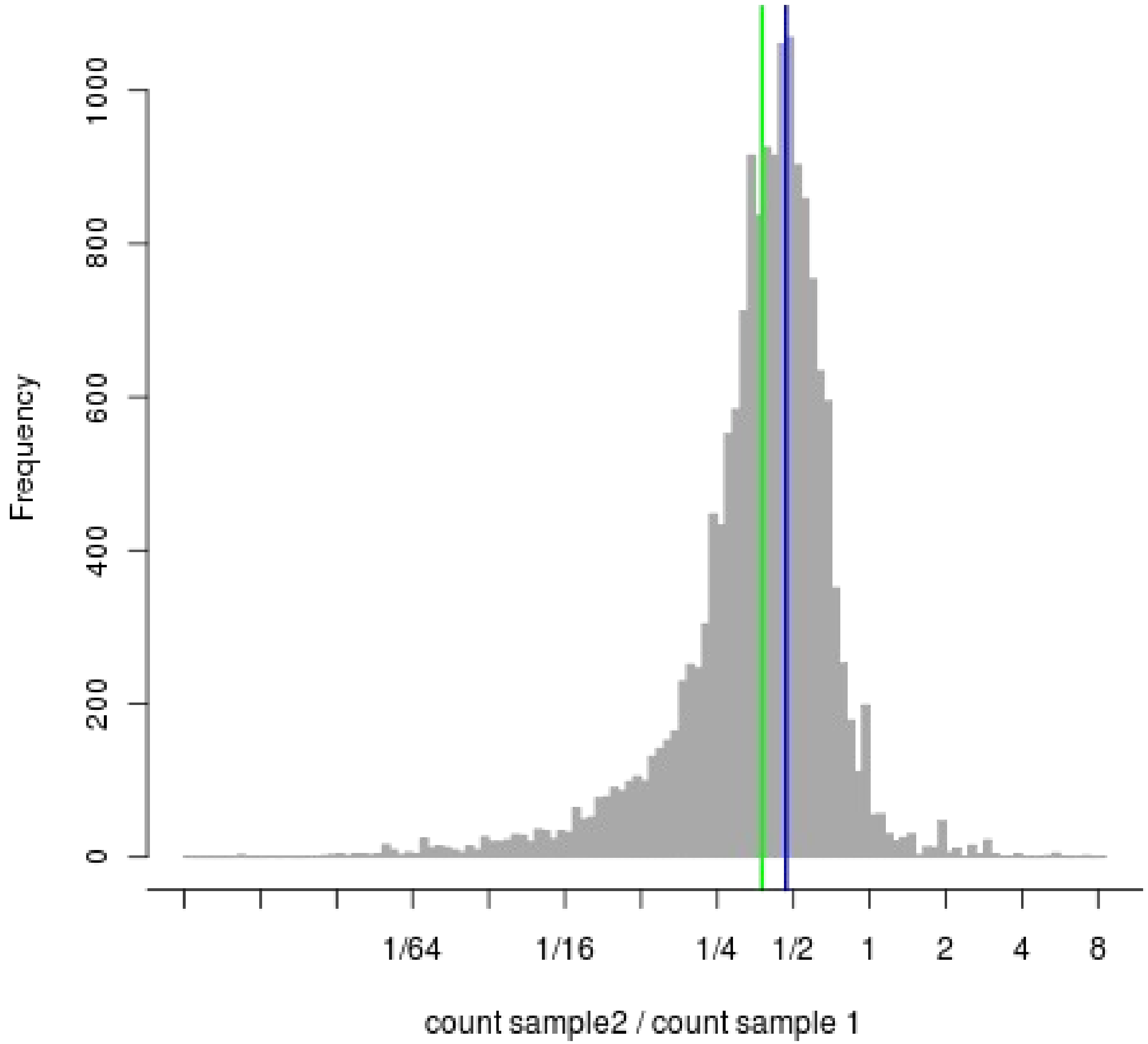
- By dividing, for each gene, the count from sample A by the count for sample B, we get one estimate per gene for the size ratio or sample A to sample B.
- We use the median of all these ratios.

Normalisation for library size



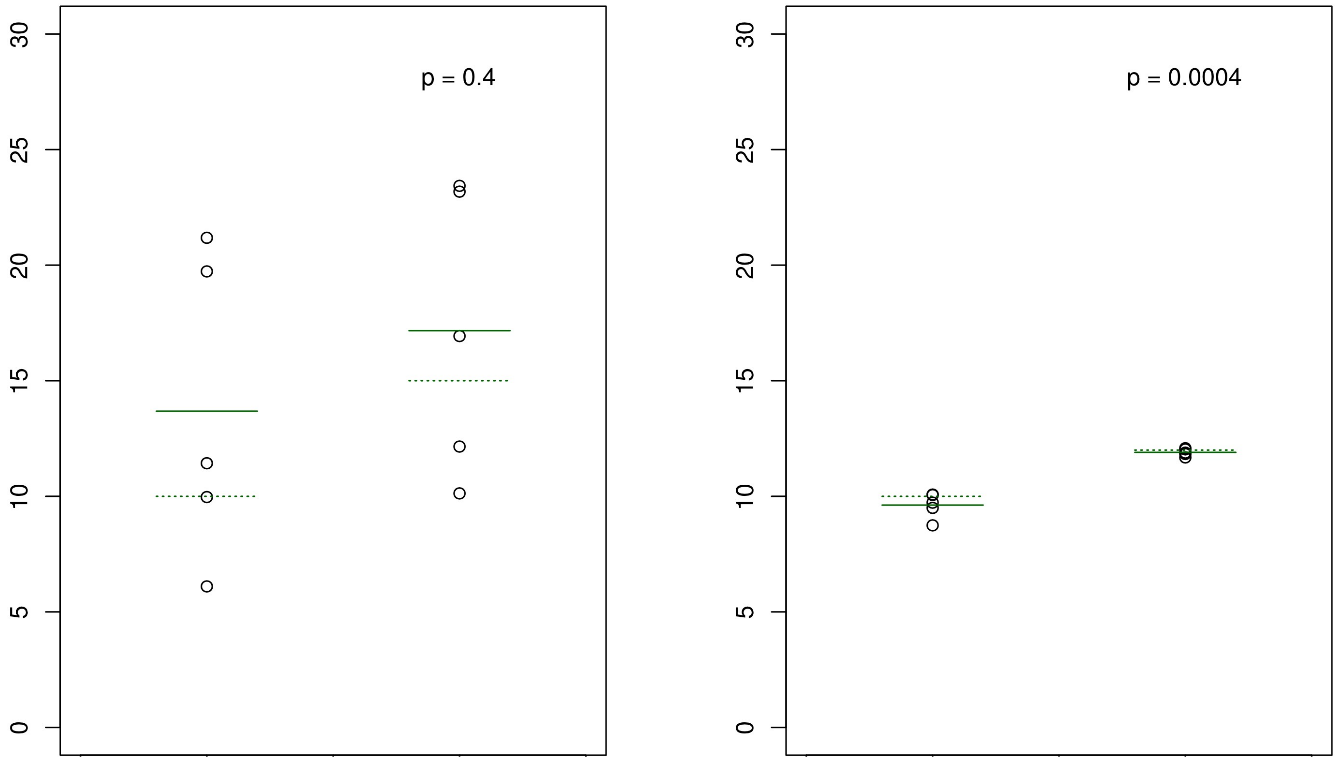
EMBL

Normalisation for library size

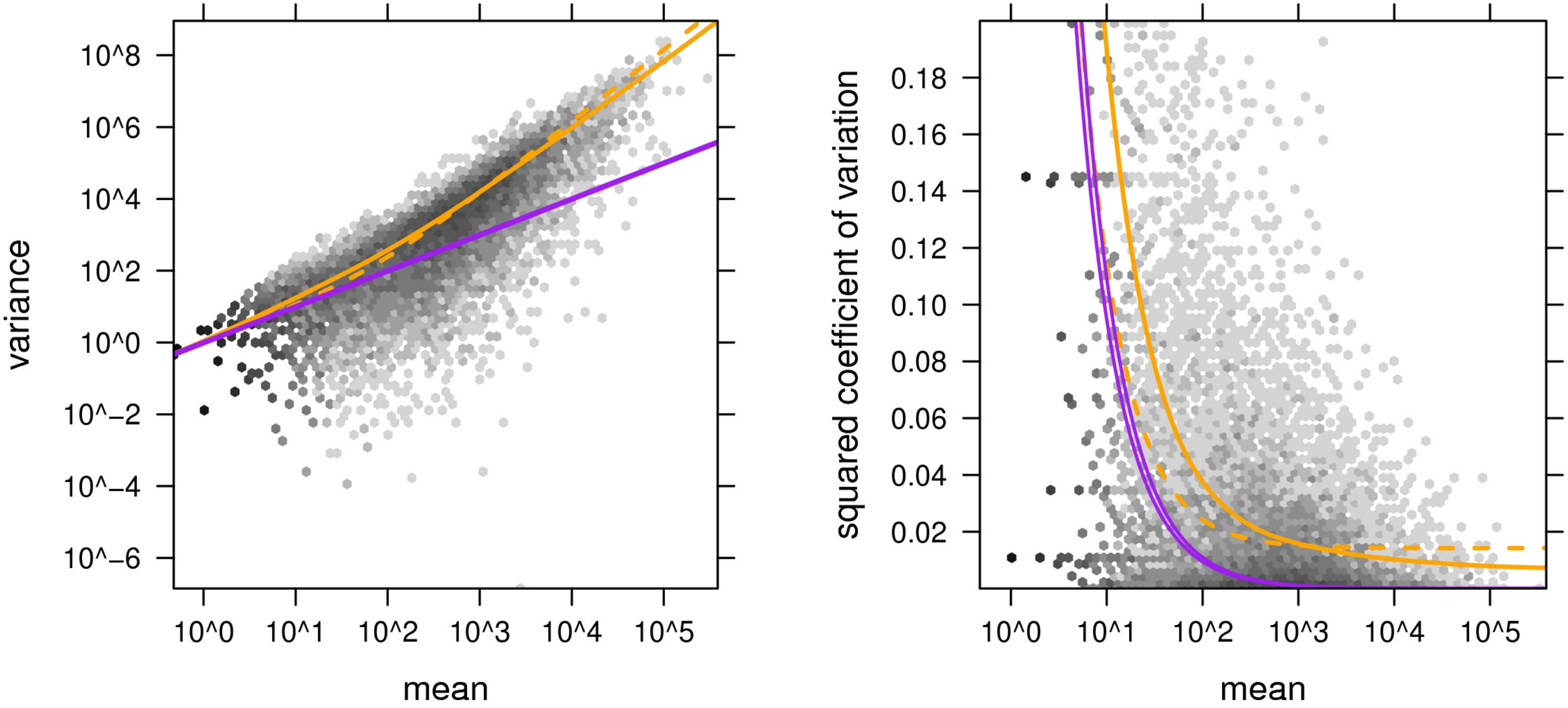


EMBL

Effect size and significance



Variance depends strongly on the mean

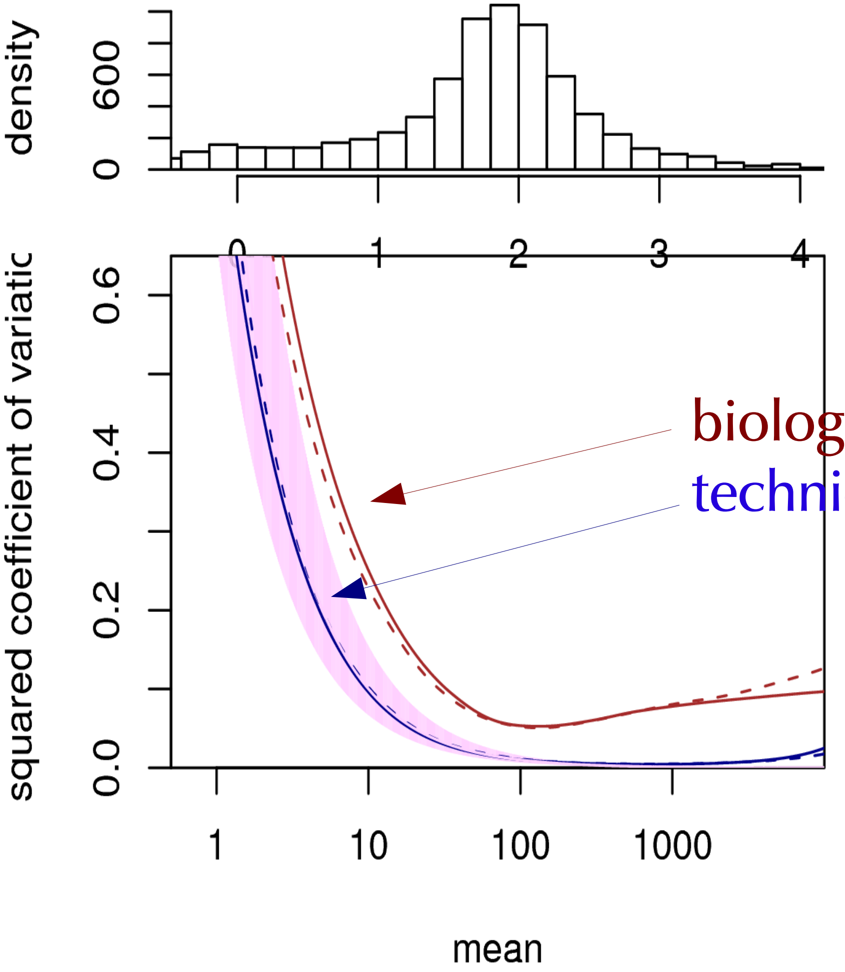


Variance calculated from comparing two replicates

Poisson	$v = \mu$	—
Poisson + constant CV	$v = \mu + \alpha \mu^2$	- - -
Poisson + local regression	$v = \mu + f(\mu^2)$	—

EMBL

Technical and biological replicates



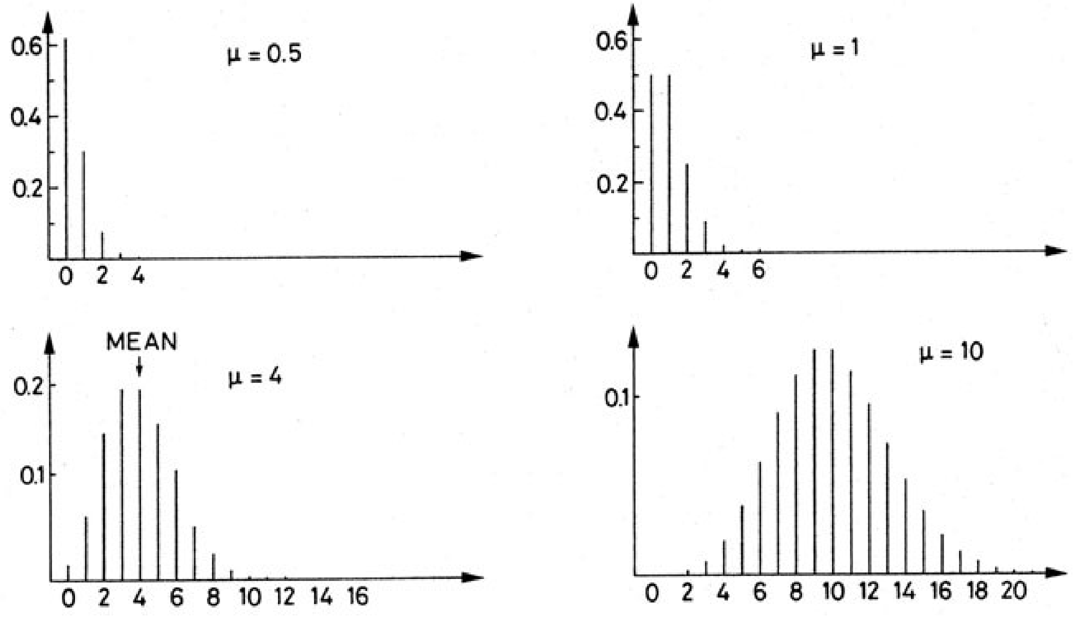
biological replicates

technical replicates

RNA-Seq of yeast [Nagalakshmi et al, 2008]

Poisson (I)

- The Poisson distribution turns up whenever things are counted
- Example: A short, light rain shower with *r* drops/m^2. What is the probability to find *k* drops on a paving stone of size 1 m^2?



Poisson (II)

For Poisson-distributed data, the variance is equal to the mean.

Hence, no need to estimate the variance
according to several authors: Marioni et al. (2008), Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010)

- Really?

Is HTS count data Poisson-distributed?

To sort this out, we have to distinguish *two* sources of noise.

Shot noise

- Consider this situation:
  - Several flow cell lanes are filled with aliquots of the *same* prepared library.
  - The concentration of a certain transcript species is *exactly* the same in each lane.
  - We get the same total number of reads from each lane.
- For each lane, count how often you see a read from the transcript. Will the count all be the same?

Shot noise

- Consider this situation:
  - Several flow cell lanes are filled with aliquots of the *same* prepared library.
  - The concentration of a certain transcript species is *exactly* the same in each lane.
  - We get the same total number of reads from each lane.
- For each lane, count how often you see a read from the transcript. Will the count all be the same?
- Of course not. Even for equal concentration, the counts will vary. This *theoretically unavoidable* noise is called *shot noise*.

Shot noise

- Shot noise: The variance in counts that persists even if everything is exactly equal. (Same as the evenly falling rain on the paving stones.)

- Stochastics tells us that shot noise follows a *Poisson distribution*.
- The standard deviation of shot noise can be *calculated*: it is equal to the square root of the average count.

Sample noise

Now consider

- Several lanes contain samples from biological replicates.
- The concentration of a given transcript varies around a mean value with a certain standard deviation.
- This standard deviation cannot be calculated, it has to be *estimated* from the data.

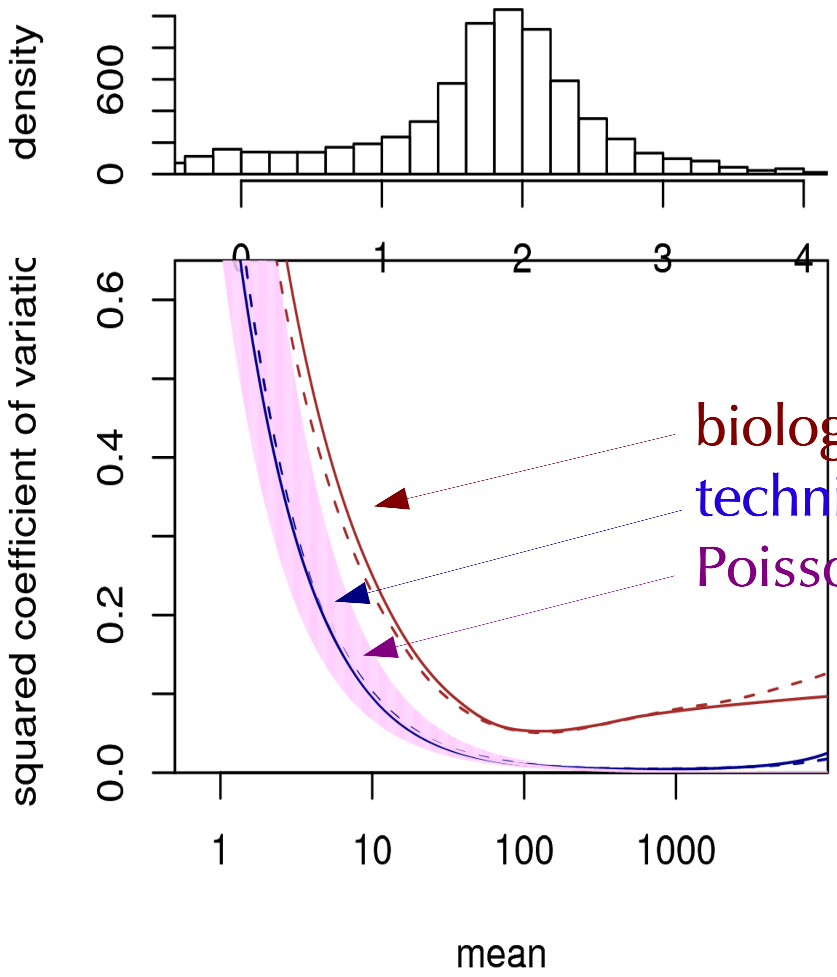
Technical and biological replicates

Nagalakshmi *et al.* (2008) have found that

- counts for the same gene from different *technical* replicates have a variance equal to the mean (Poisson).
- counts for the same gene from different *biological* replicates have a variance exceeding the mean (overdispersion).

Marioni *et al.* (2008) have looked confirmed the first fact (and confused everybody by ignoring the second fact).

Technical and biological replicates



biological replicates

technical replicates

Poisson noise

RNA-Seq of yeast [Nagalakshmi et al, 2008]

Summary: Noise

We distinguish:

- Shot noise
  - unavoidable, appears even with perfect replication
  - dominant noise for weakly expressed genes

can be computed

- Technical noise
  - from sample preparation and sequencing
  - negligible (if all goes well)
- Biological noise
  - unaccounted-for differenced between samples
  - Dominant noise for strongly expressed genes

needs to be estimated from the data

EMBL

Which null hypothesis?

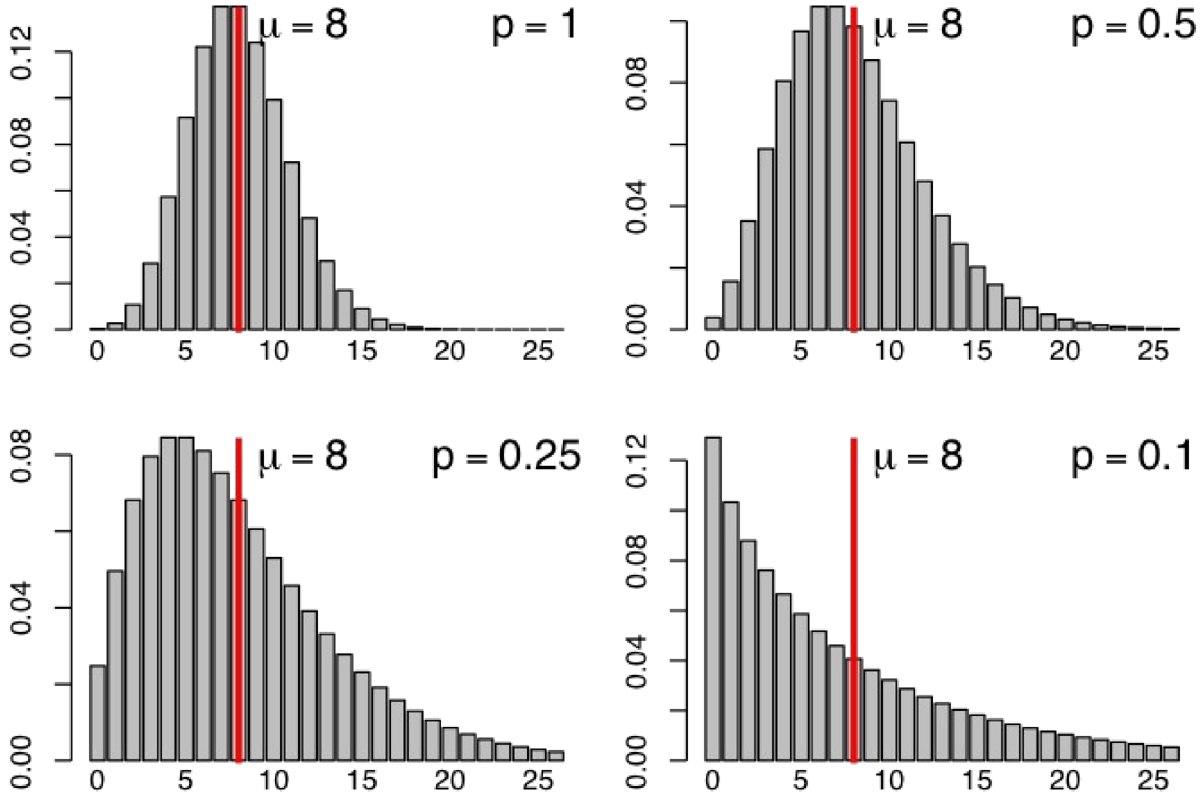
In a statistical test, we attempt to *reject* a null hypothesis.

Given two samples with different experimental conditions, which null hypothesis covers the question of biological interest?

- The concentration of transcripts from gene *i* is equal in the two samples.
  - → shot noise is all we need to know
- The difference of the concentrations is of a magnitude as is expected between replicate samples.
  - → Estimate of biological variability needed

The negative-binomial distribution

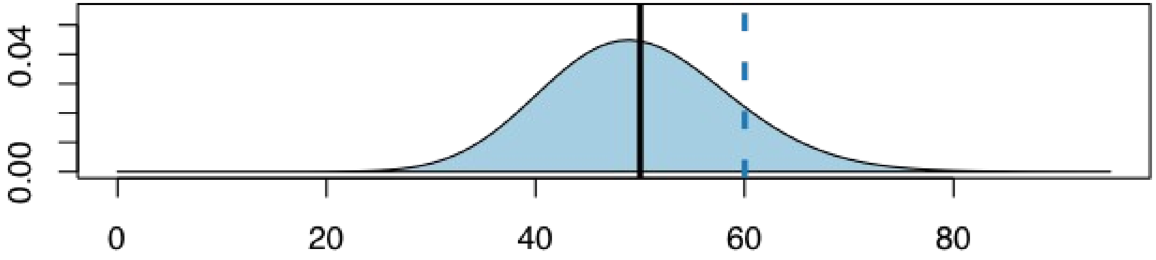
A commonly used generalization of the Poisson distribution with *two* parameters



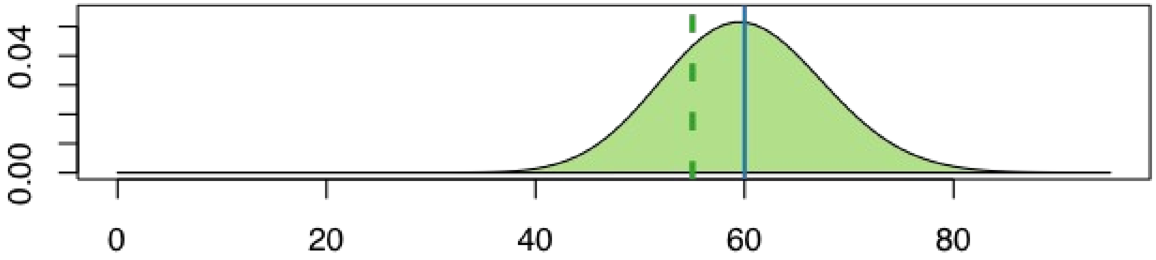
$$\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

EMBL

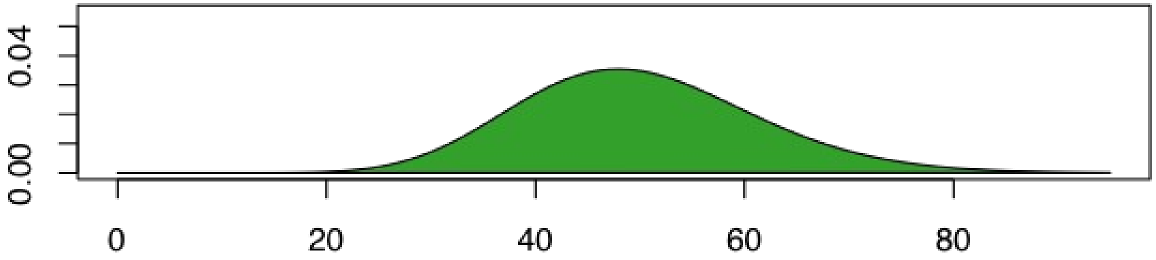
The NB distribution from a hierarchical model



Biological sample with mean μ and variance *v*



Poisson distribution with mean *q* and variance *q*.



Negative binomial with mean μ and variance $q+v$.

Testing: Null hypothesis

Model:

The count for a given gene in sample *j* come from negative binomial distributions with the mean $s_j \mu_\rho$ and variance $s_j \mu_\rho + s_j^2 v(\mu_\rho)$.

s_j	relative size of library *j*
μ_ρ	mean value for condition ρ
$v(\mu_\rho)$	fitted variance for mean μ_ρ

Null hypothesis:

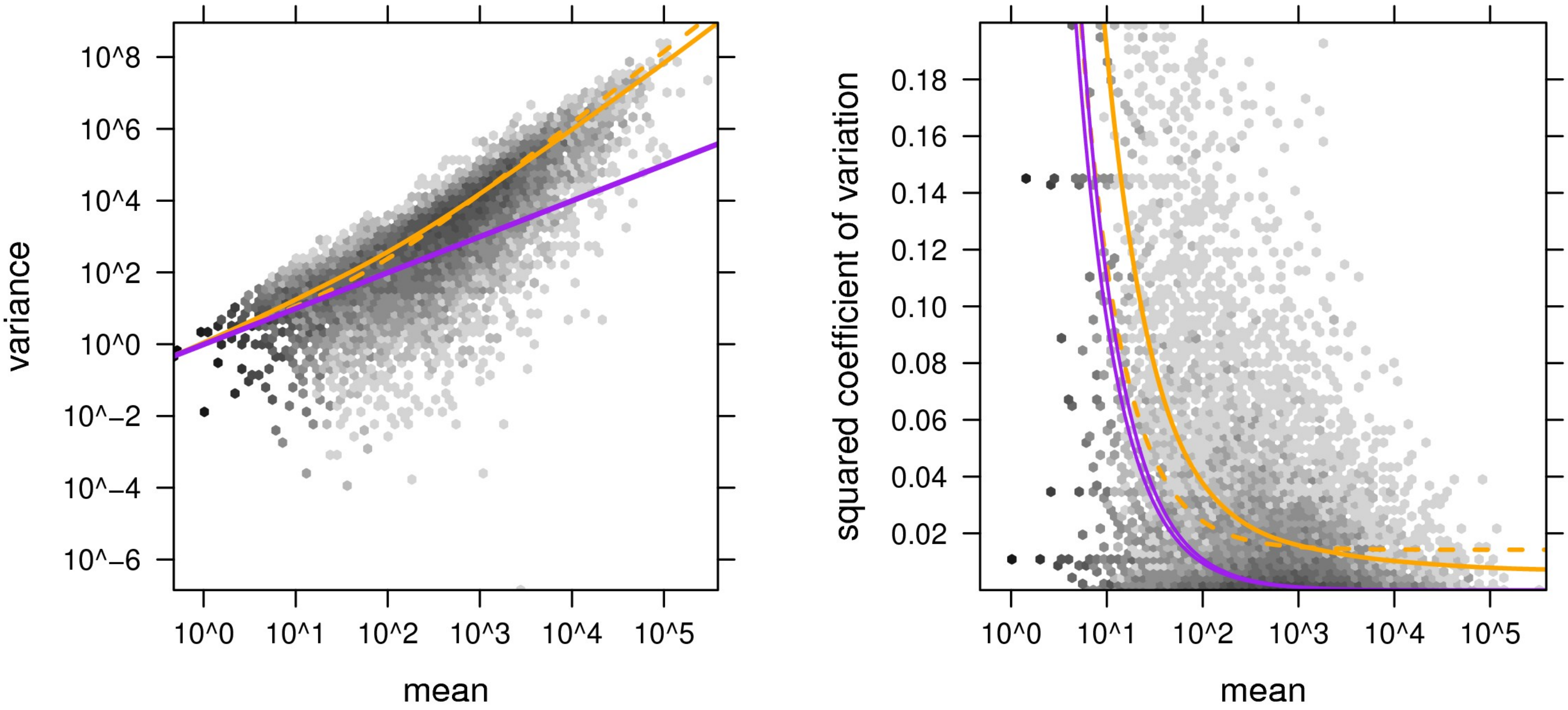
The experimental condition r has no influence on the expression of the gene under consideration:

$$\mu_{\rho_1} = \mu_{\rho_2}$$

EMBL

Model fitting

- Estimate the variance from replicates
- Fit a line to get the variance-mean dependence *v*(μ)
  (local regression for a gamma-family generalized linear model, extra math needed to handle differing library sizes)

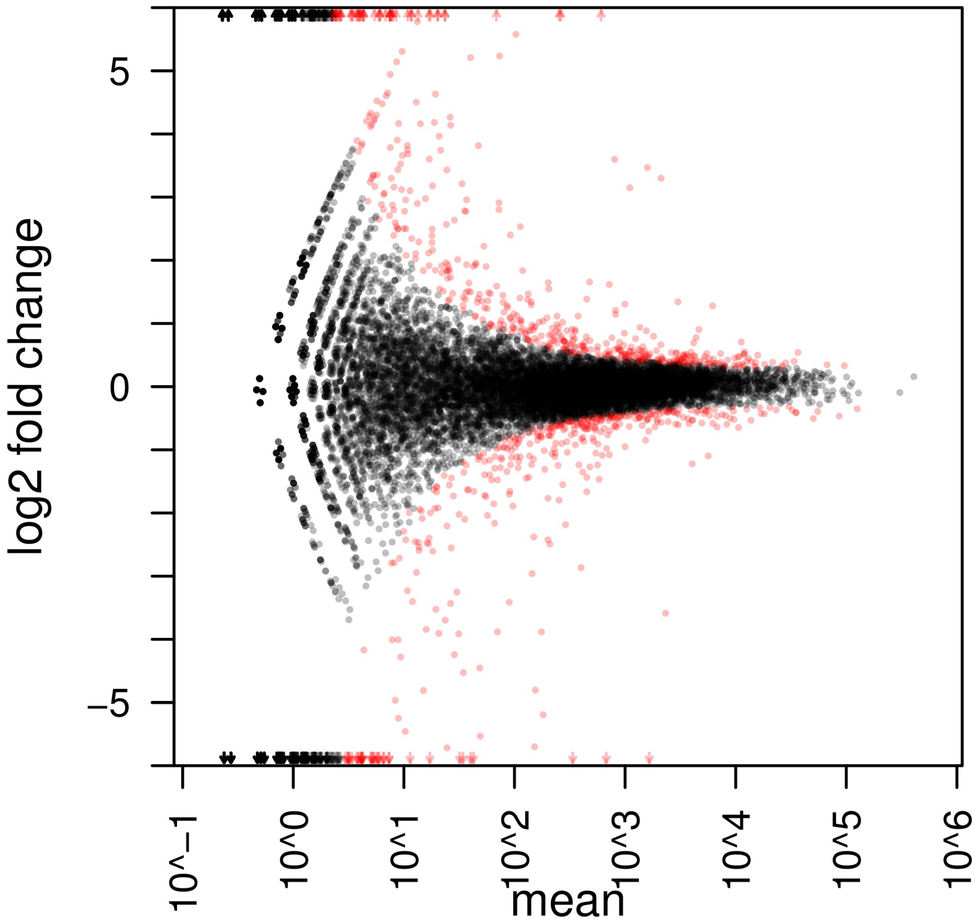


Testing for differential expression

- For each of two conditions, add the count from all replicates, and consider these sums $\mathrm{K}_{i\mathrm{A}}$ and $K_{i\mathrm{B}}$ as NB-distributed with moments as estimated and fitted.
- Then, we calculate the probability of observing the actual sums or more extreme ones, conditioned on the sum being $k_{i\mathrm{A}}+\mathrm{k}_{i\mathrm{A}}$, to get a *p* value.

(similar to the test used in Robinson and Smyth's *edgeR)*

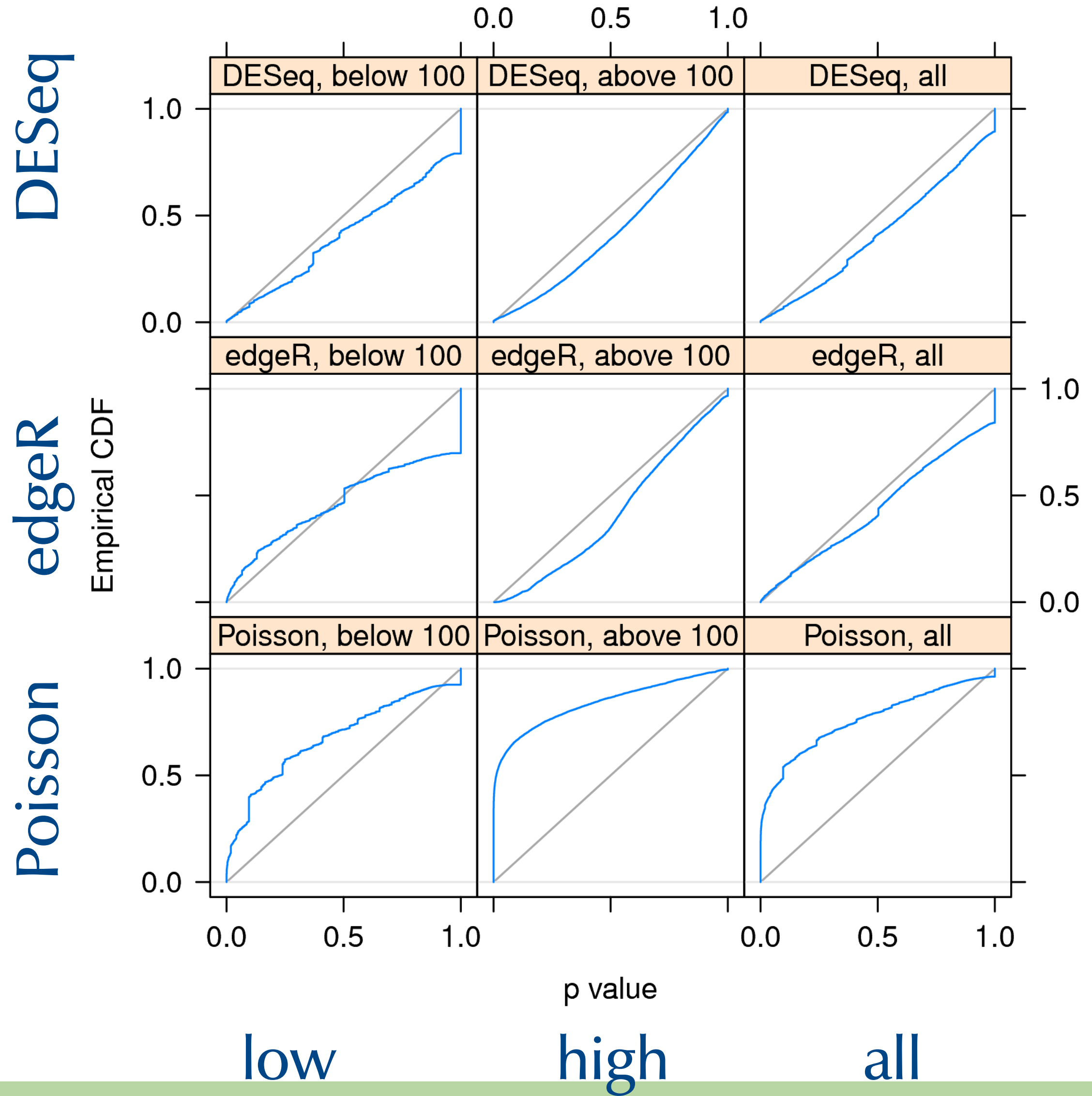
Differential expression



RNA-Seq data: overexpression of two different genes in flies [data: Furlong group]

EMBL

Type-I error control



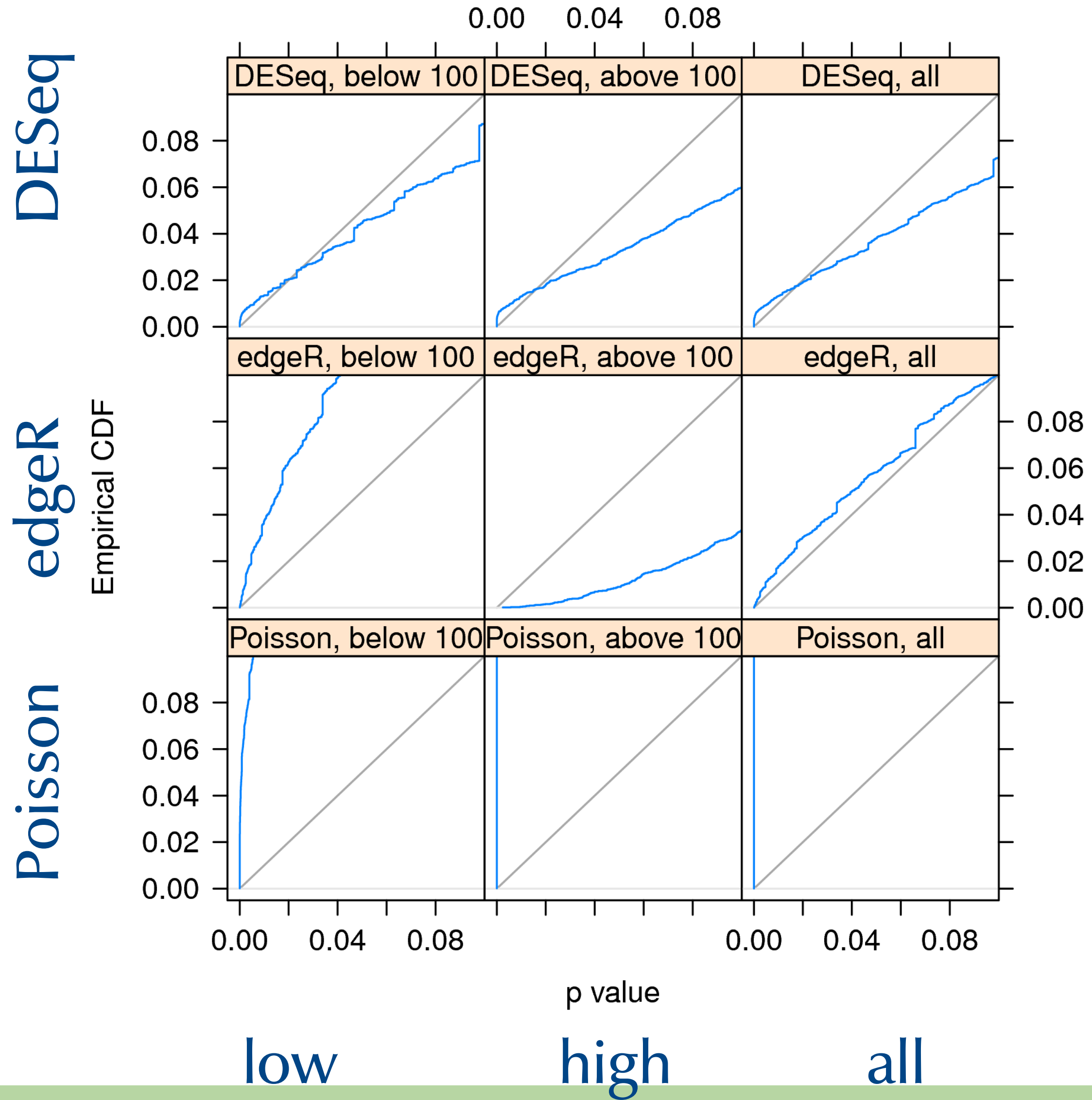
Comparison of replicates:

no differential expression,

expect uniform *p* values

EMBL

Type-I error control: Zoom



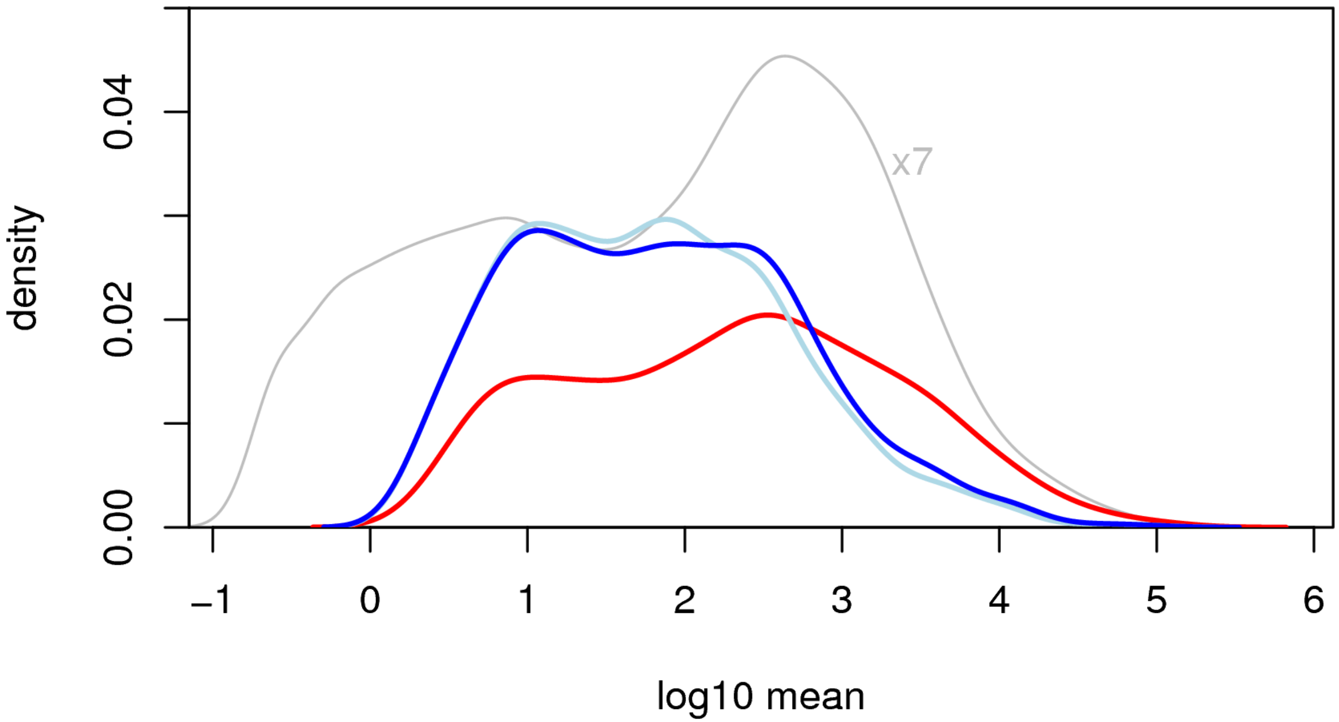
Comparison of replicates:

no differential expression,

expect uniform *p* values

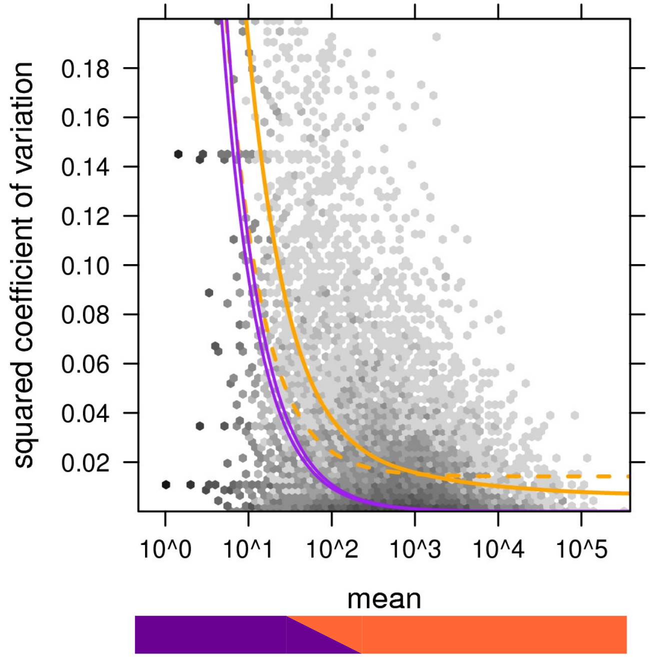
EMBL

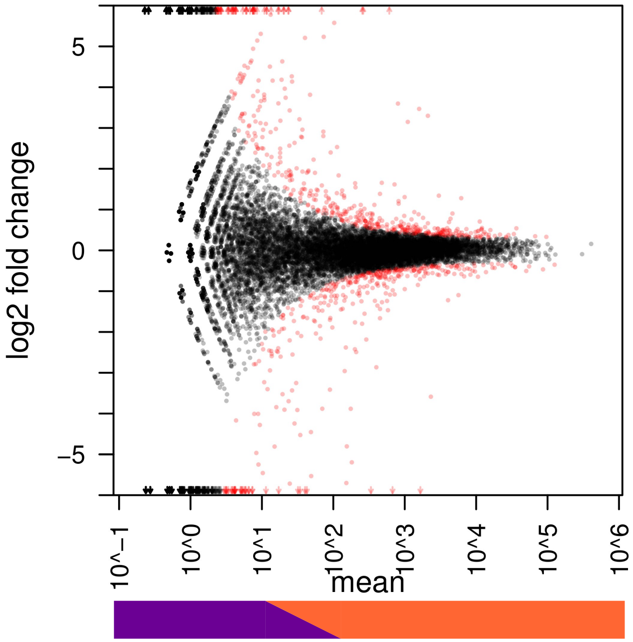
Distribution of hits along the dynamic range



▬ all genes
▬ differentially expressed according to DESeq
▬ differentially expressed according to edgeR

Two noise ranges





	dominating noise	*How to improve power?*
🟪	shot noise (Poisson)	deeper sampling
🟧	biological noise	more biological replicates

EMBL

Working without replicates

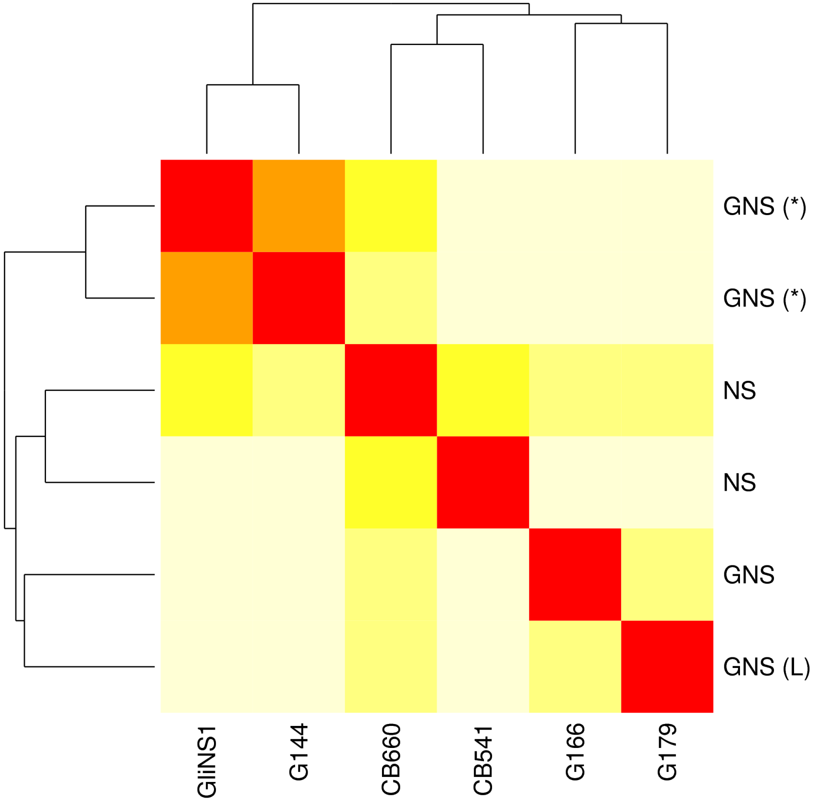
One can infer the variance from a comparison of different conditions.

- The variance will be overestimated, maybe drastically.
- The power is smaller, maybe much smaller.

Still, this is the best one can do without replicates.

Variance-stabilizing transformation

The estimated variance-mean dependence allows to derive a transformation that renders the count data approximately homoskledastic.



This is useful, e.g., as input for the `dist` function.

[Tag-Seq of neural stem cell tissue cultures, Bertone Group]

EMBL

Further use cases

Similar count data appears in

- comparative ChiP-Seq
- barcode sequencing
- …

and can be analysed with *DESeq* as well.

Alternative splicing

- So far, we counted reads in *genes*.
- To study alternative splicing, reads have to be assigned to *transcripts*.
- This introduces ambiguity, which adds uncertainty.
- Current tools (e.g., *cufflinks*) allow to quantify this uncertainty.
- However: To assess the significance of differences to isoform ratios between conditions, the assignment uncertainty has to be combined with the noise estimates.
- This is not yet possible with existing tools.

EMBL

Coming soon: GLMs for DESeq

- Until now, DESeq only supported simple comparisons.
- There are many use cases requiring more general models.
- Implementation of generalized linear models (GLMs) for DESeq now ready for use.

The model

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\text{with} \qquad \log \frac{\mu_{ij}}{s_j} = \eta_{ij} = \beta_{i0} + \sum_l \beta_{il} x_{jl}$$

K_{ij}	read count for gene i in sample j
s_j	size factor for sample j
α_i	dispersion for gene i
μ_{ij}	expectation for gene i in sample j
η_{ij}	linear predictor for gene i in sample j
x_{jl}	l-th for predictor sample j (indicator or quantitative)
β_{il}	l-th regression coefficient for gene i

The model

- This is a generalized linear model:
  - Link function: log link, with sample-dependent size factors
  - Family: negative binomial with known dispersion

- The dispersion is estimated from the fitted variance-mean relation, reading off at the sample average:

$$\tilde{\alpha}_i = v(\hat{q}_i)/\hat{q}_i^2 \qquad \text{with} \qquad \hat{q}_i = \sum_j \frac{k_{ij}}{s_j}$$

- The negative binomial is in the exponential family if the dispersion is given.

DESeq with GLMs: Applications

- Factorial designs with interactions
- Paired samples
- Regression of expression on genotype, etc.
- Alternative splicing
    - Count for each exon, then fit a model for each gene, and test the interaction between exon and experimental condition.
- Methylation (HELP)
- ChIP-Seq: ChIP vs input crossed with condition

EMBL

Conclusions

- Proper estimation of variance between *biological* replicates is vital. Using Poisson variance is incorrect.
- Estimating variance-mean dependence with local regression works well for this purpose.
- The negative-binomial model allows for a powerful test for differential expression

- Preprint on *Nature Preecedings*:
  "Differential expression analysis for sequence count data"
- Software (*DESeq*) available from Bioconductor and EMBL web site.

Google for DESeq

*

- Co-author: Wolfgang Huber

- Funding: European Union (Marie Curie Research and Training Network “Chromatin Plasticity”) and EMBL

Negative-binomial model (I)

- Suppose, we have *m* replicates of a given condition, and obtain counts for *n* genes.
- The concentration of gene *i* in replicate *j* is a random variable Q_{ij}, which is i.i.d. for $j=1,\dots,m$ with mean q_{i0} and variance σ_i^2.
- Let K_{ij} be the count value for gene *i* in replicate *j*. Its expectation value is $s_j\mu_i$ with size factor s_j.
- Given $Q_{ij}=q_{ij}$, the sequencing is a Poisson process and hence: $K_{ij} \sim \mathrm{Pois}(\, s_j q_{ij} \,)$.

Negative-binomial model (II)

- If Q_{ij} has mean μ_i and variance σ_i^2, what is the the marginal (“mixing”) distribution of $K_{ij} \sim \text{Pois}(\, s_j \text{q}_{ij}\,)$?
- If one assumes Q_{ij} to be gamma-distributed, the answer is:
- K_{ij} follows a negative binomial (NB) distribution with mean $s_j q_{i0}$ and variance $s_j q_{i0} + s_j \sigma_i^2$.

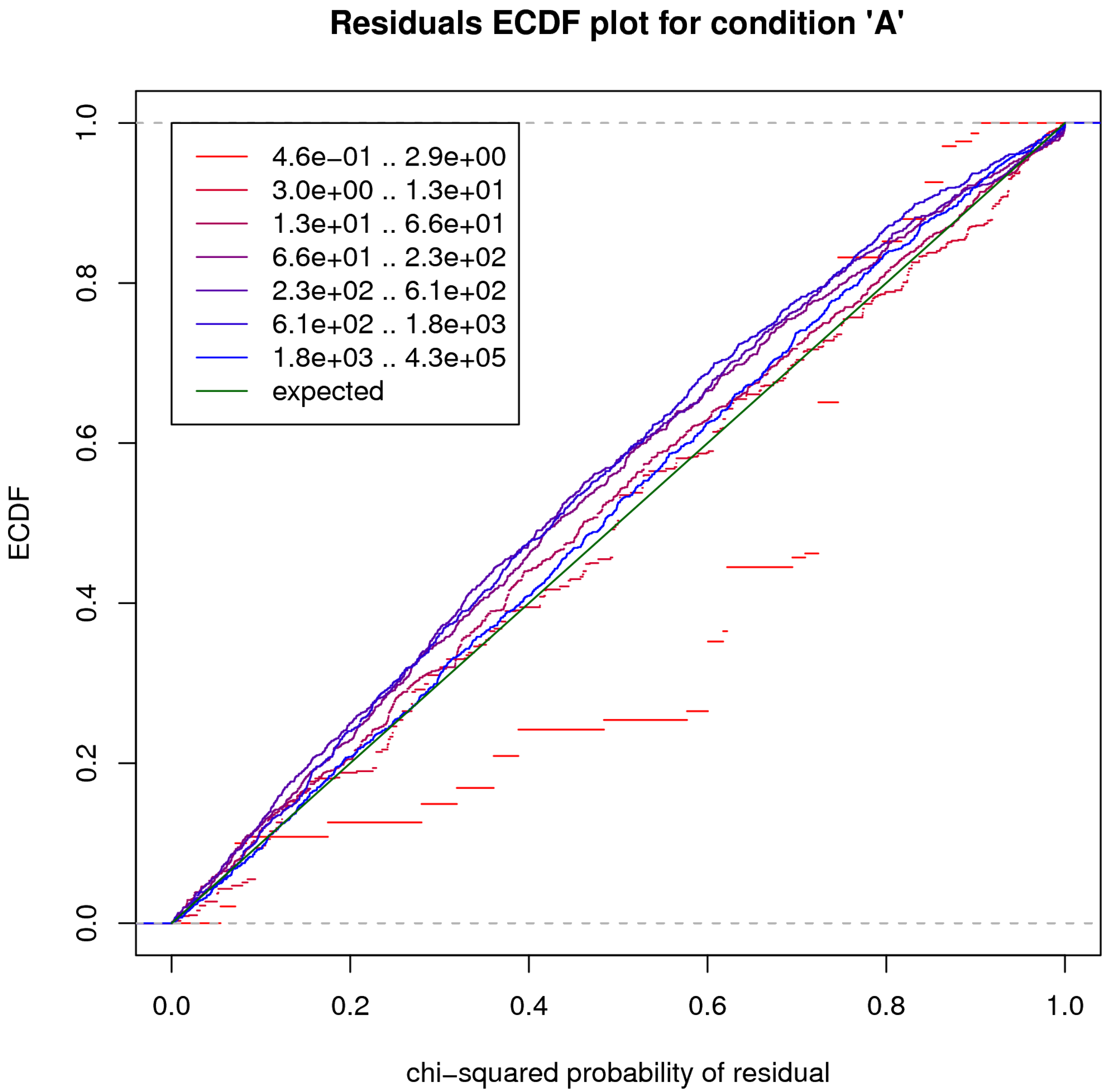
Model fitting

- Estimate relative library sizes s_j.
- Within a set of replicates, calculate for each gene sample mean and sample variance of k_{ij}/s_j.
- To get an unbiased estimate of σ_i^2, subtract an "average shot-noise" of $\frac{\hat{q}_i}{m} \sum \frac{1}{\hat{s}_j}$
- Fit a line through the graph of mean and variance estimates (with a gamma-family local regression).

Model:

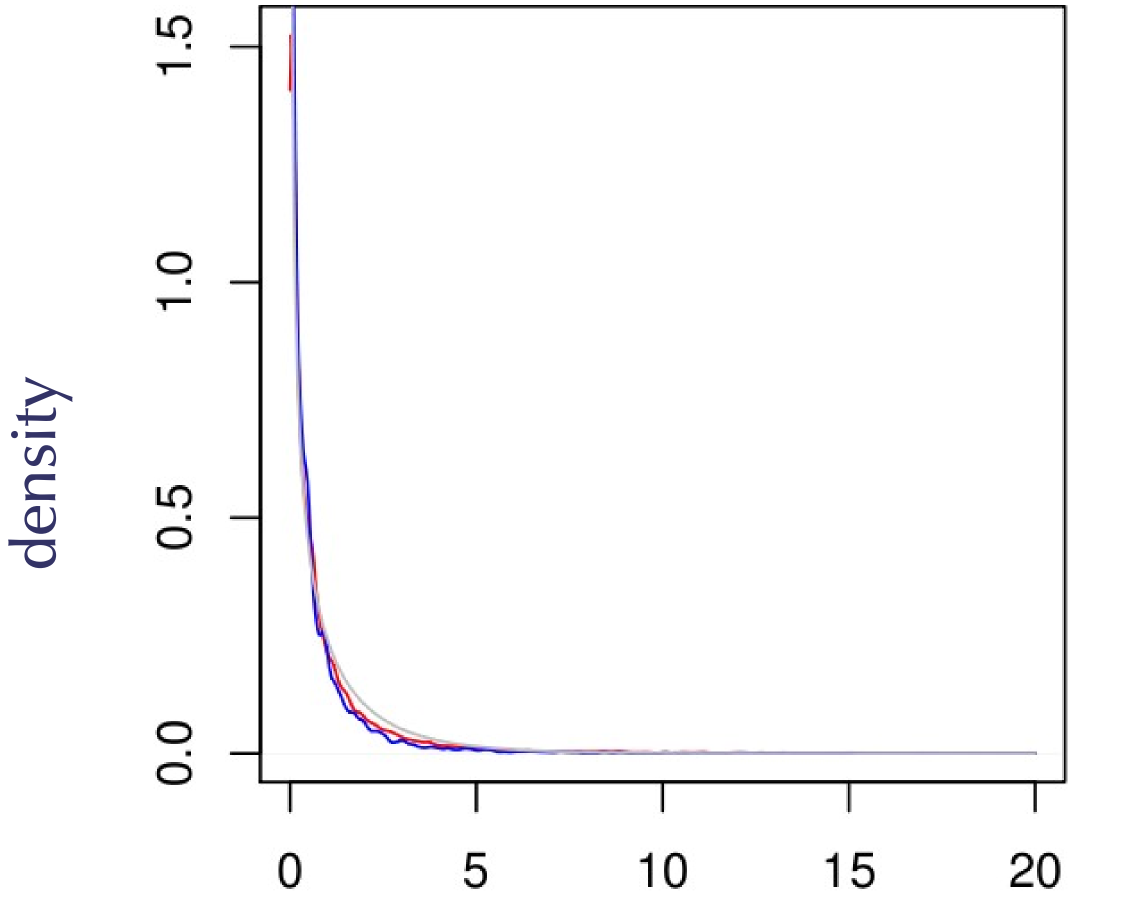
K_{ij} follows a negative binomial (NB) distribution with mean $s_j q_{i0}$ and variance $s_j q_{i0} + s_j \sigma_i^2$.

Diagnostic plot for variance fit



EMBL

Variance residuals distribution



per-gene sample variance / fitted variance

EMBL