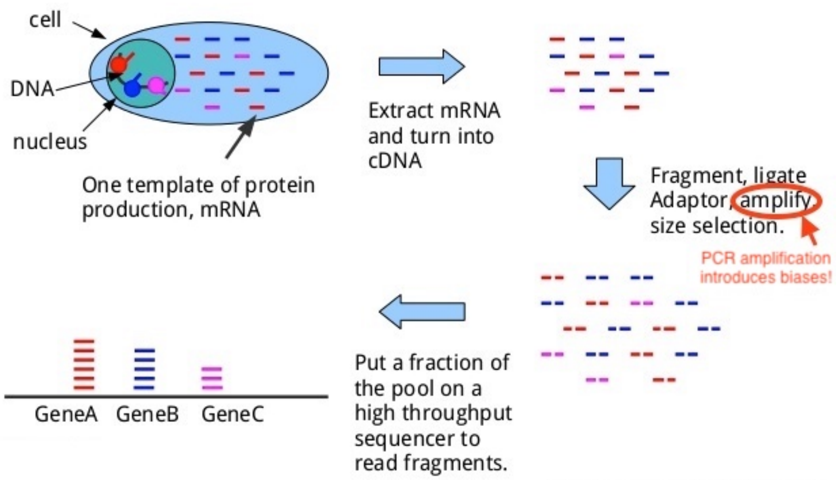
Correctly counting molecules with a little help from a well-known population model

Florian Pflug and Arndt von Haeseler

CIBIV
Center for Integrative Bioinformatics Vienna

Edinburgh, July 17th 2018

The Basic RNA-Seq Workflow



One template of protein production, mRNA

Extract mRNA and turn into cDNA

Fragment, ligate Adaptor, amplify, size selection.

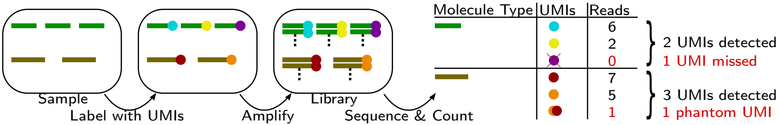
PCR amplification introduces biases!

Put a fraction of the pool on a high throughput sequencer to read fragments.

Nature Reviews Molecular Cell Biology 11, 467–478

RNA-Seq with Unique Molecular Identifiers (UMIs)

To measure absolute transcript counts, and avoid errors due to *PCR amplification bias*, mRNA transcripts are labelled with Unique Molecular Identifiers (UMIs; ●, ●, ●, ●, ●) *before amplification* ...



Molecule Type	UMIs	Reads
▬	●	6
	●	2
	●	0
▬	●	7
	●	5
	●●	1

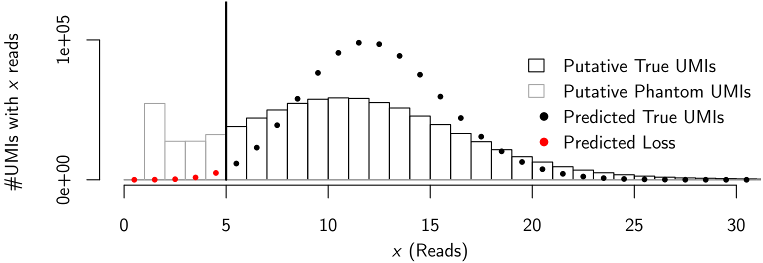
2 UMIs detected
1 UMI missed

3 UMIs detected
1 phantom UMI

... and after sequencing, not *Reads* but *Unique UMIs* are counted to measure transcript abundance

Towards a model to predict the reads/UMI distribution

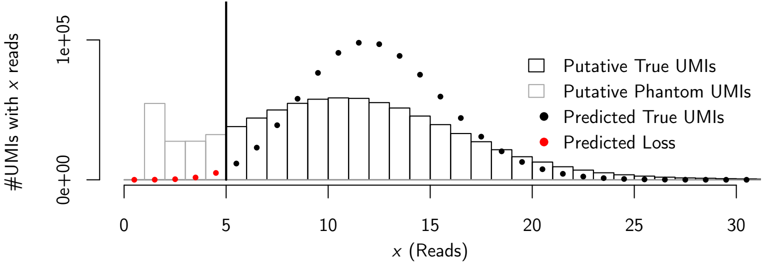
We sequence only a small percentage of all molecules...



.. but there's more dispersion than stochastic sampling can explain

Towards a model to predict the reads/UMI distribution

We sequence only a small percentage of all molecules...



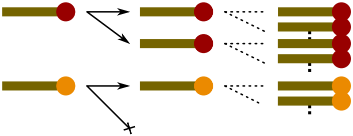
.. but there's more dispersion than stochastic sampling can explain

We must consider the stochasticity of the PCR

PCR as a supercritical Galton-Watson branching process

Of each UMI-labelled molecule there initially is a single copy.
During each cycle, each molecule is duplicated with probability E,

$$M_0 = 1, \qquad M_i = M_{i-1} + \mathrm{Binom}(M_{i-1}, E), \qquad \mathbb{E}M_i = (1+E)^i$$



after $1^{\mathbf{st}}$ cycle after n cycles

PCR as a supercritical Galton-Watson branching process

Of each UMI-labelled molecule there initially is a single copy. During each cycle, each molecule is duplicated with probability E,

$$M_0 = 1, \qquad M_i = M_{i-1} + \mathrm{Binom}(M_{i-1}, E), \qquad \mathbb{E}M_i = (1+E)^i$$

We normalize $M_0, M_1, \ldots$ to have expected value 1,

$$F_i = \frac{M_i}{(1+E)^i}$$

PCR as a supercritical Galton-Watson branching process

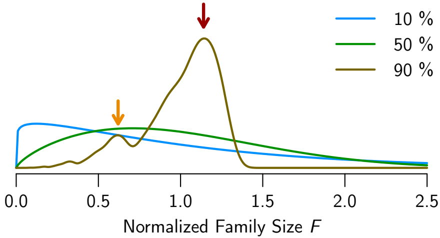
Of each UMI-labelled molecule there initially is a single copy. During each cycle, each molecule is duplicated with probability E,

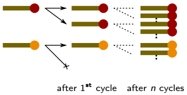
$$M_0 = 1, \qquad M_i = M_{i-1} + \mathrm{Binom}(M_{i-1}, E), \qquad \mathbb{E}M_i = (1+E)^i$$

We normalize $M_0, M_1, \ldots$ to have expected value 1,

$$F_i = \frac{M_i}{(1+E)^i}$$

And call the limit $F = \lim_{i\to\infty} F_i$ *(normalized) family size*.

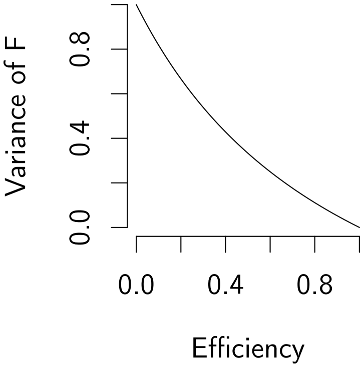




The family size distribution - Variance

While the density of the family size distribution doesn't seem to be analytically tractable, the variance has a simple analytic expression

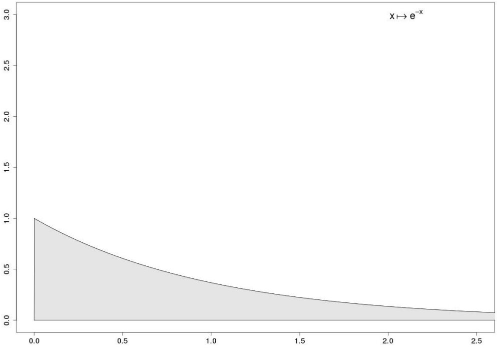
$$\mathbb{V}F = \frac{1 - E}{1 + E}$$



The family size distribution - Density

To compute the density, we must resort to numeric methods

We used simulations+KDE, but now a fast method developed by Straub and Neininger (Göthe-Universität Frankfurt) is available

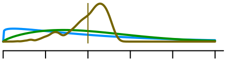


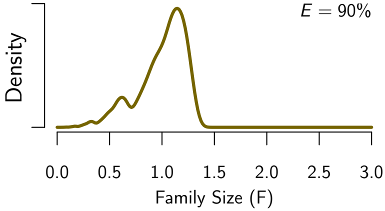
(Video due to Straub & Neininger)

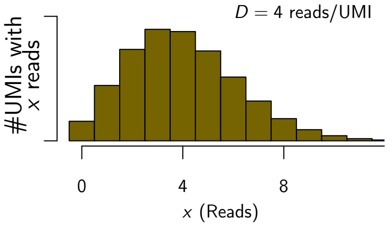
From the family size to the reads/UMI distribution

Sequencing is Poissonian sampling from families of unknown size

$$\mathbb{P}(k) = \int_0^\infty \underbrace{\mathbb{P}(k \mid \lambda = D \cdot x)}_{\text{Poisson}} \cdot \underbrace{\mathbb{P}(\text{Fam. Size} = c \mid E)} \, dx.$$







The complete model has two parameters, depth D and efficiency E.

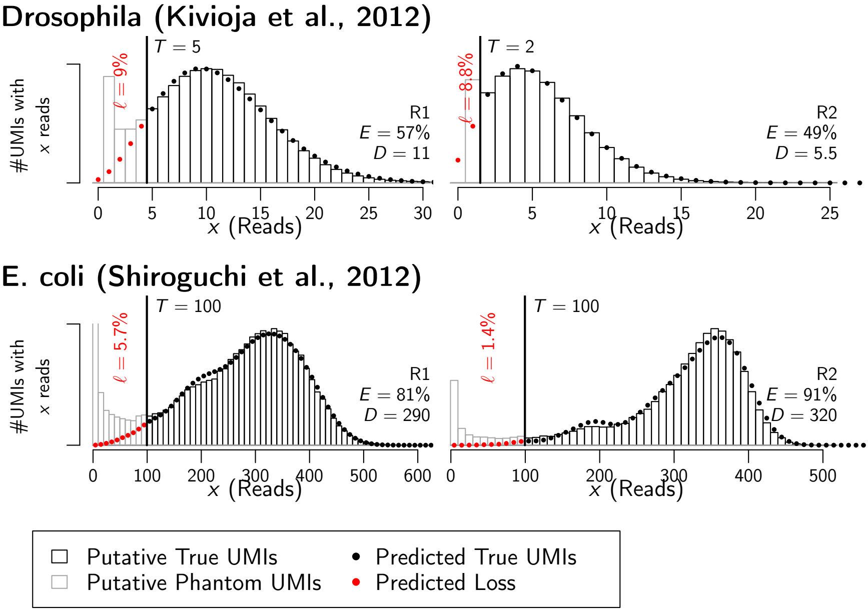
Parameter Estimation

For the reads count C per UMI, we can analytically find

$$\mathbb{E}(C) = D, \qquad \mathbb{V}(C) = D + D^2 \frac{1 - E}{1 + E}.$$

We can estimate D, E with the *method of moments*.

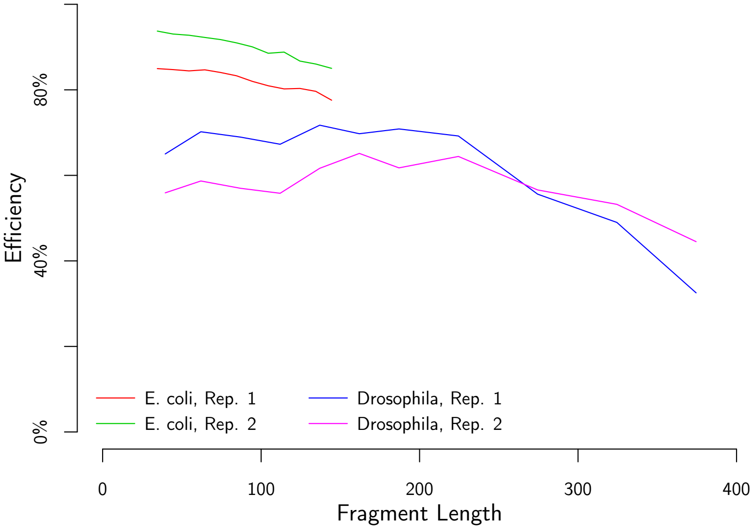
Observed & Expected Reads/UMI



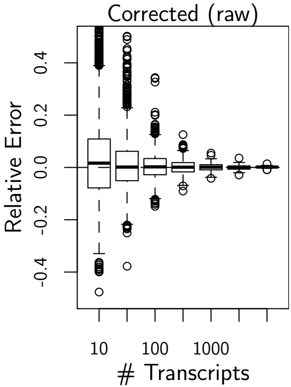
Drosophila (Kivioja et al., 2012)

E. coli (Shiroguchi et al., 2012)

PCR efficiency vs. length

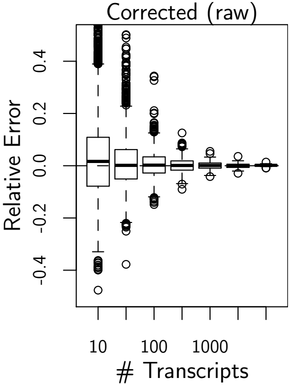


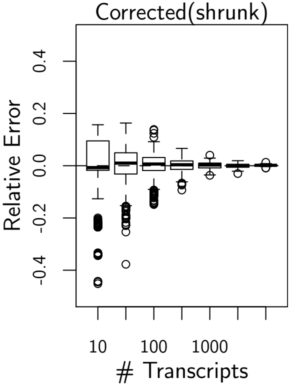
Correcting for gene-wise biases



For genes with few transcripts, we have little data to estimate D, E, and the correction hurts more than it helps...

Correcting for gene-wise biases





We *shrink* the gene-wise loss estimate $\hat{\ell}_g^{\text{raw}}$ towards global ones

$$\hat{\ell}_g^{\text{shrink}} = \lambda_g \cdot \hat{\ell}_g^{\text{raw}} + (1 - \lambda_g) \cdot \hat{\ell}_g^{\text{all}}$$

Summary

- The Galton-Watson branching process model captures the main stochastic properties of the PCR reaction
- while still allowing efficient parameter estimation
- and allows us to predict, detect & correct biases
- as well as studying of early-cycle PCR behaviour.

Most of this work was recently published in:

Florian G. Pflug and Arndt von Haeseler. TRUmiCount: Correctly counting absolute numbers of molecules using unique molecular identifiers. *Bioinformatics* (2018).

And we provide an R package `gwpcR` which implements the family size distribution, Poisson mixture, and parameter estimation

Acknowledgements

Every at the **CIBIV**, in particular:
Olga Chernomor
Celine Prakash
Luis Paulin-Paz

Goethe-Universität Frankfurt:
Jasmin Straub & Günther Neininger

VIENNA RNA BIOLOGY

mfpl MAX F. PERUTZ LABORATORIES

universität wien

MEDIZINISCHE UNIVERSITÄT WIEN

Thank You
for your Attention