






Correcting for unobserved Molecules in quantitative NGS Experiments

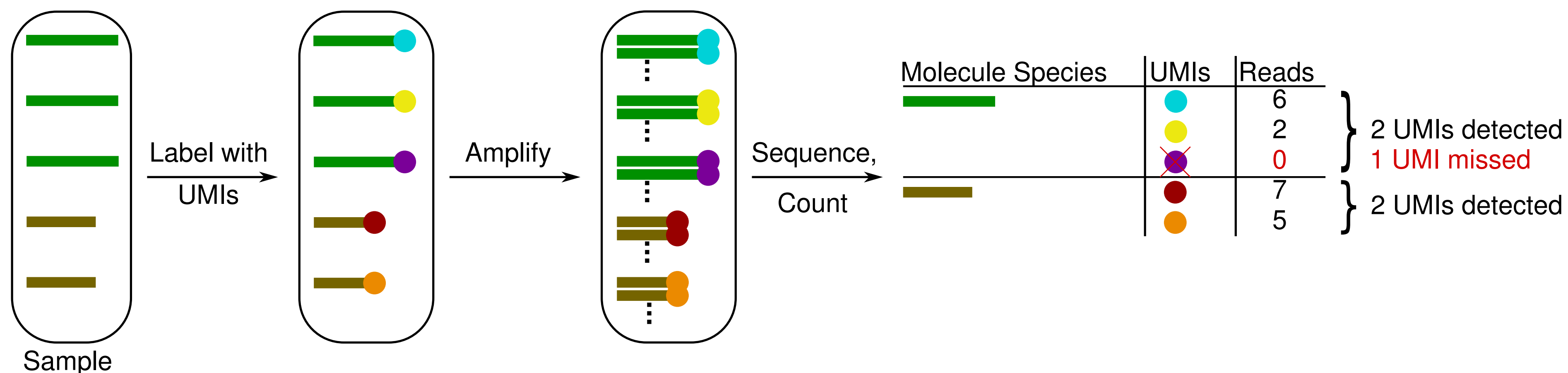
Florian Pflug and Arndt von Haeseler

1. INTRODUCTION – UNIQUE MOLECULAR IDENTIFIERS (UMIs)

RNA-Seq use Next Generation Sequencing (NGS) to measure the abundance of of each type of mRNA transcript present in a sample (3x  and 2x  in the figure).

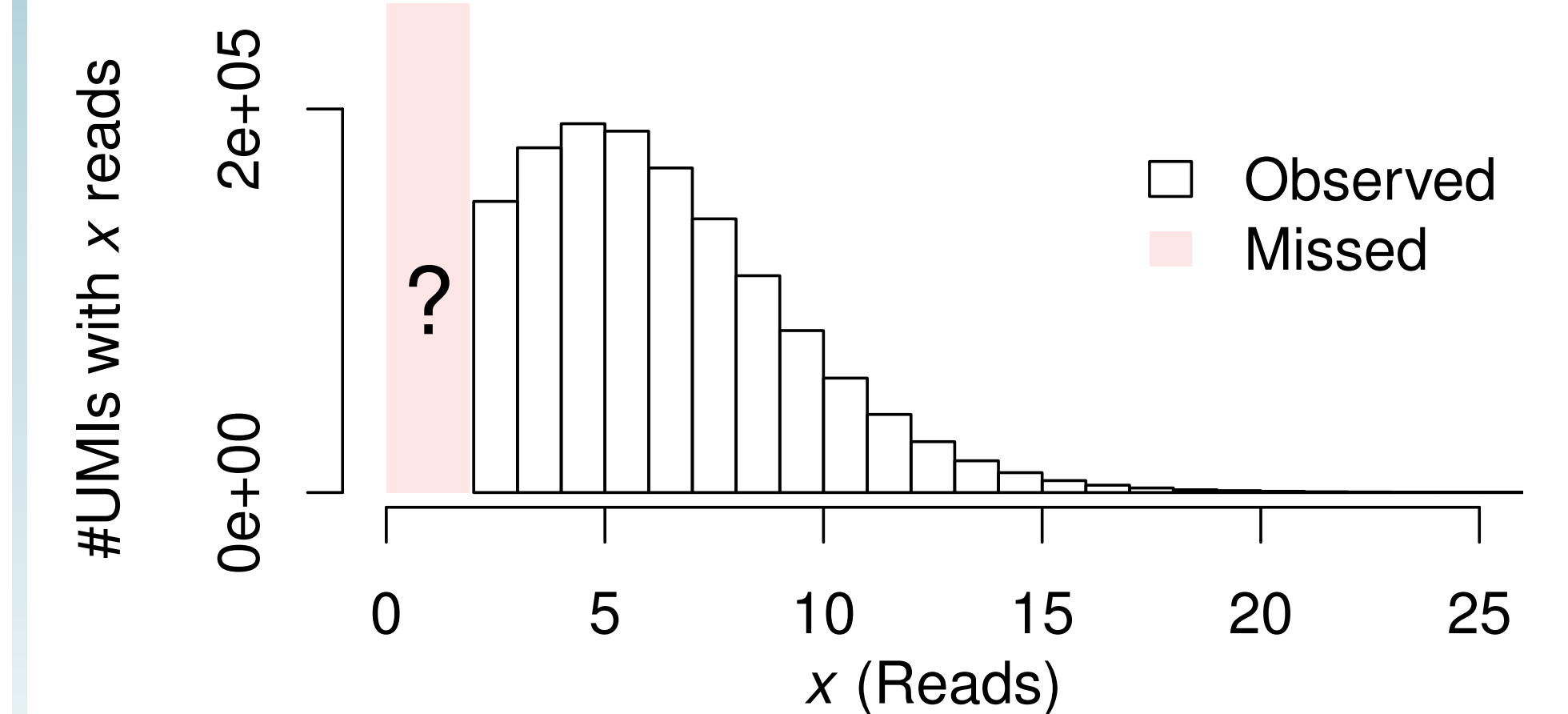
Workflow: The molecules in the sample are made distinguishable by adding *Unique Molecular Identifiers* (UMIs; , , , , ). The UMI-labelled molecules are then amplified and sequenced.

Counting distinct UMIs measures pre-amplification copy numbers, but only if no UMIs are missed.



2. MISSED UMIs

Reads/UMI for a *D. melanogaster* RNA-Seq experiment



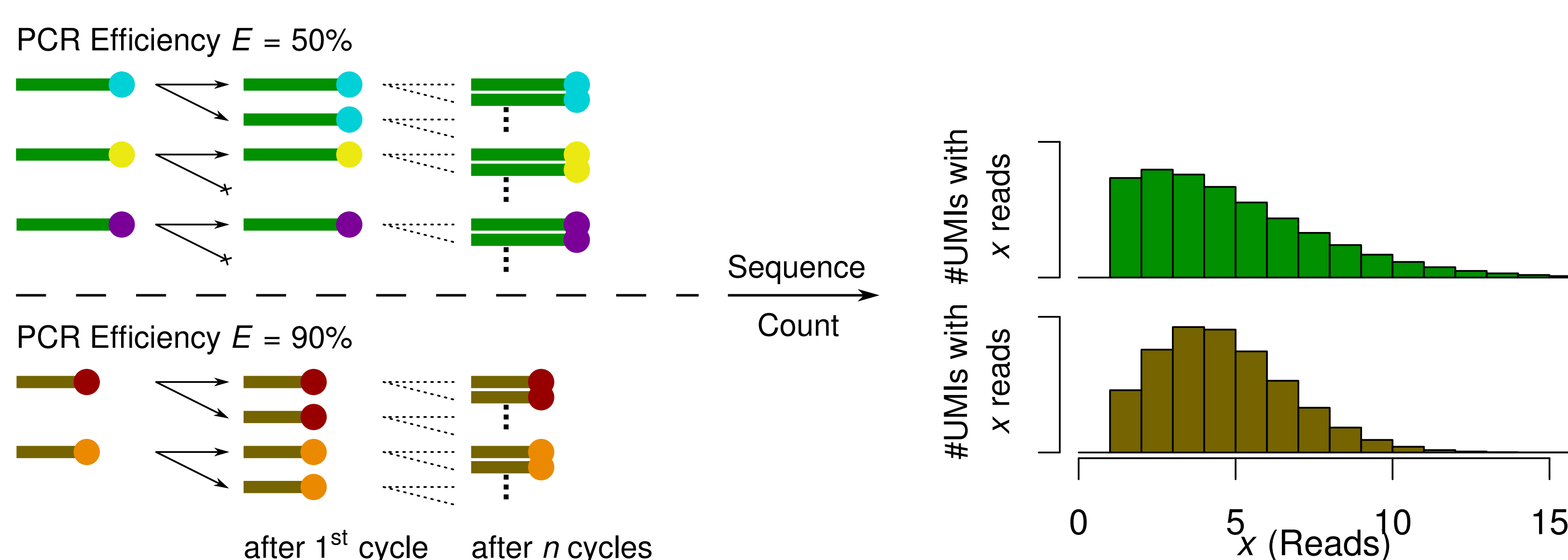
To estimate how many molecules were missed, we must extrapolate into the red area.

For this we must understand the distribution!

3. A STOCHASTIC PCR MODEL

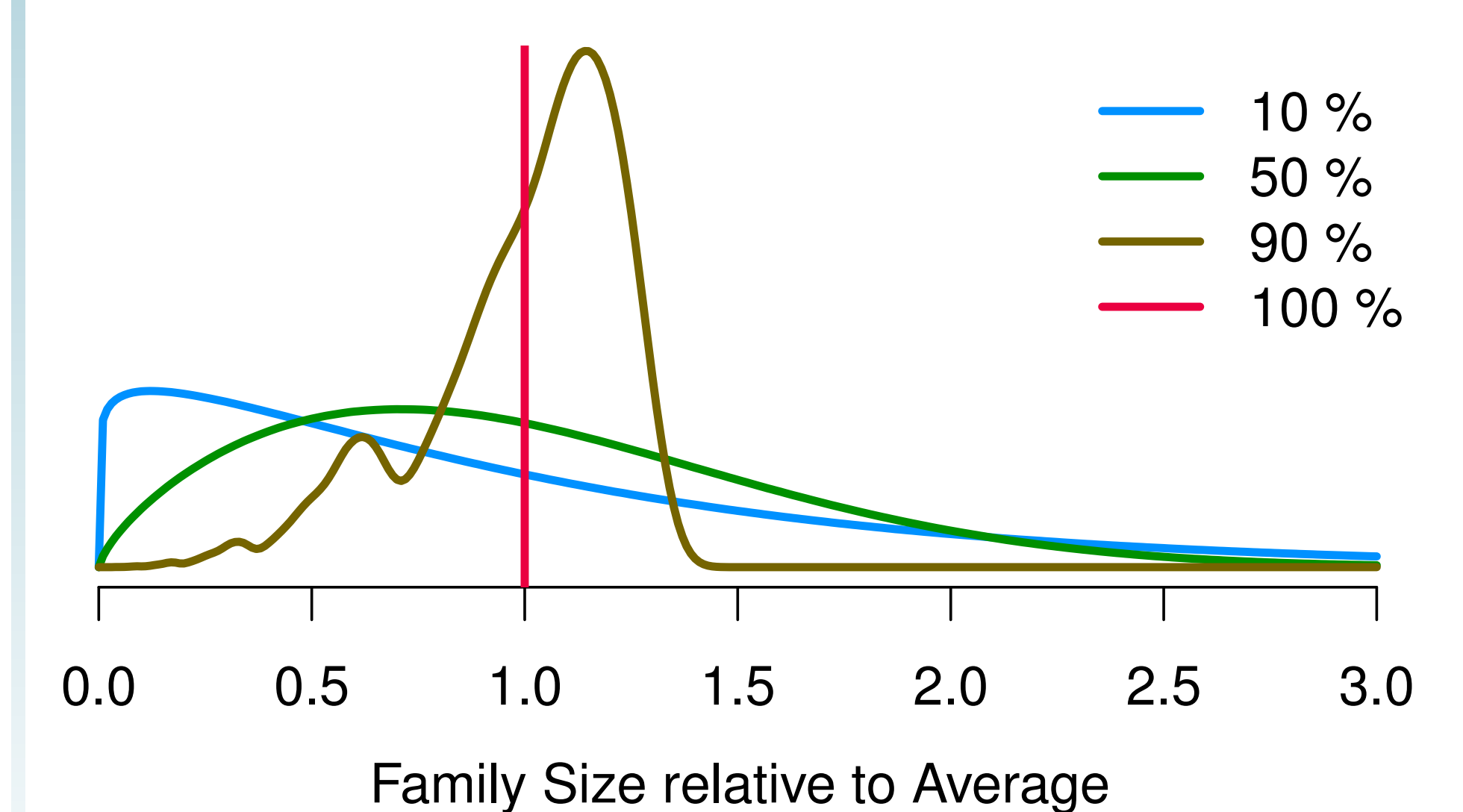
Starting from a single copy, the PCR duplicates each molecule during each cycle with efficiency E . This expands each original UMI-labelled molecule into a molecular family of identical copies.

If some molecules are not duplicated, later cycles have fewer templates to duplicate from, and the final family size is reduced. Simulated sequencing of these variably sized families produces Reads/UMI distributions similar to observed ones, with the shape depending on the PCR efficiency!



4. MOLECULAR FAMILY SIZES

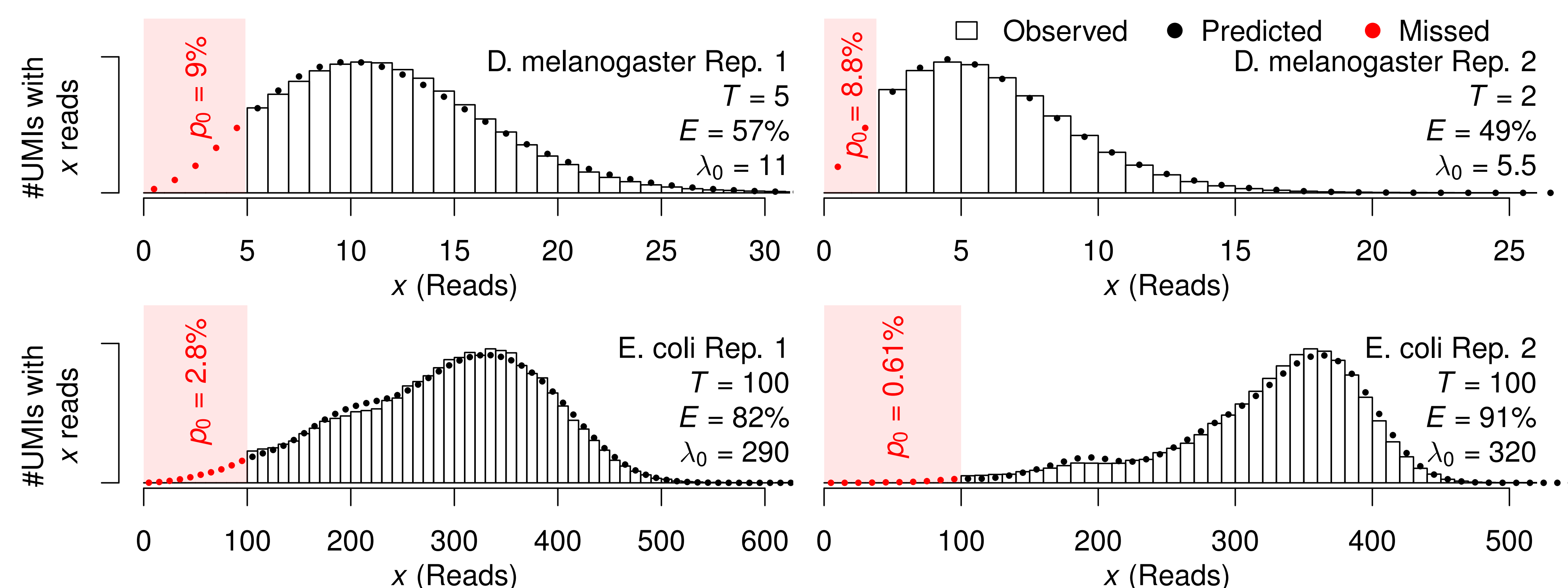
The distribution of molecular family sizes after amplification but before sequencing for different PCR efficiencies



The variability of family sizes grows as the efficiency drops!

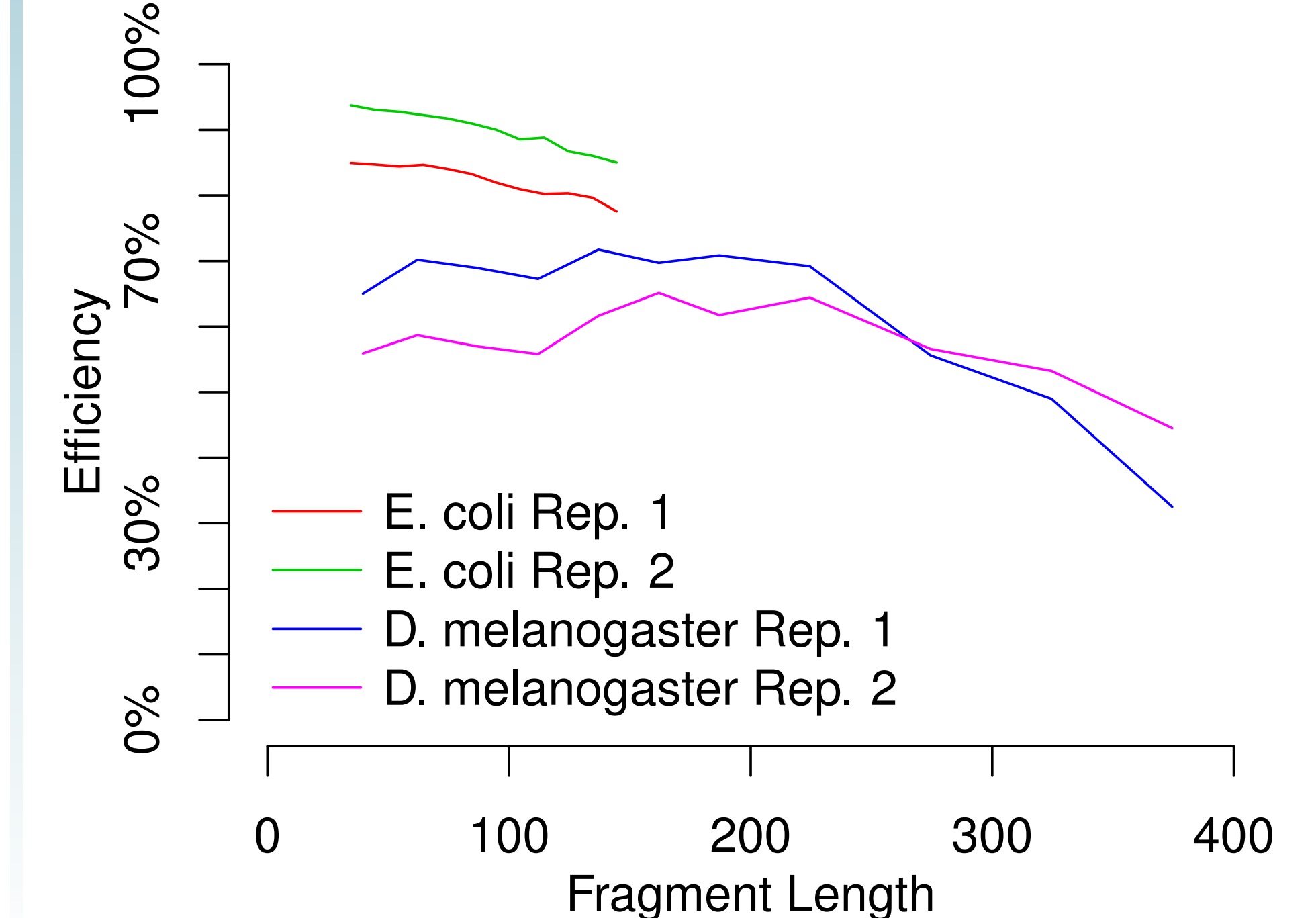
5. MODEL VS. EXPERIMENTAL DATA

The PCR efficiency is estimated from the observed Reads/UMI distribution of two RNA-Seq datasets (Kivioja *et al.* 2011 resp. Shiroguchi *et al.* 2012) by finding the efficiency for which the model fits best. The predicted distribution is then used to estimate the percentage of missed UMIs.



6. LENGTH DEPENDENCE

Fragments are binned by length and the efficiency is estimated separately for each bin by fitting the model.



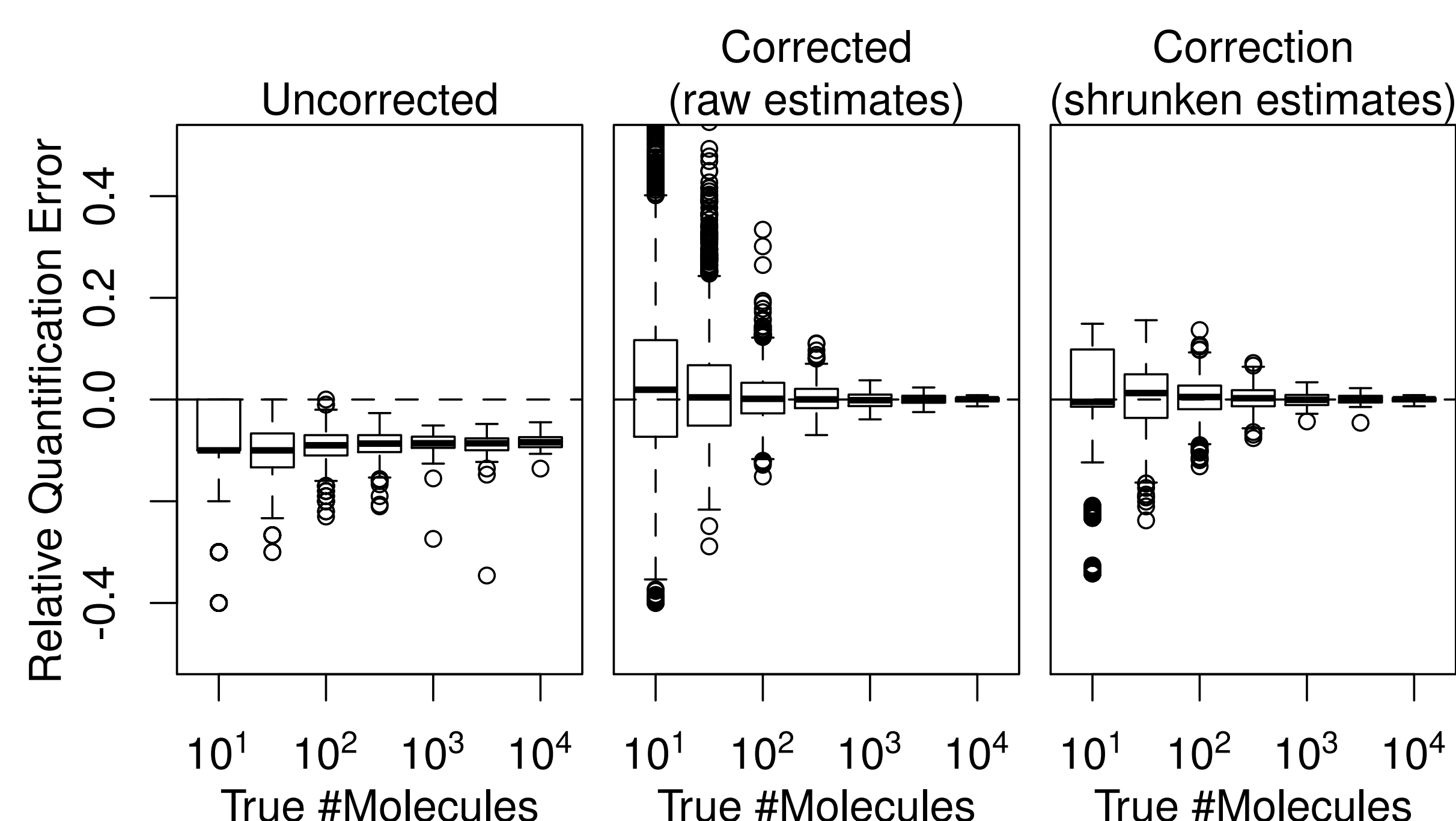
8. CORRECTING FOR GENE-SPECIFIC BIAS OF MISSED UMIs

Raw gene-specific estimates of ρ_0 suffer from a large estimation error if number of UMIs is small.

We *shrink* the raw gene-specific estimates towards the overall estimate

$$\hat{\rho}_0^{\text{shrink}} = \lambda \cdot \hat{\rho}_0^{\text{raw}} + (1 - \lambda) \cdot \hat{\rho}_0^{\text{all}}$$

In simulations $\hat{\rho}_0^{\text{shrink}}$ outperforms both $\hat{\rho}_0^{\text{raw}}$ and $\hat{\rho}_0^{\text{all}}$, having low variance and being asymptotically unbiased.



9. CONCLUSIONS & OUTLOOK

Stochastic PCR behaviour explains effects found in real-world RNA-Seq datasets, and a simple stochastic model offers accurate predictions.

We can correct for gene-specific biases in the number of missed UMIs, thus increasing the accuracy of expression comparisons *across genes*.