LEXOGEN
Enabling complete transcriptome sequencing

Estimating pre-PCR fragment numbers from post-PCR frequencies of unique molecular identifiers

Michael Moldaschl, Andreas Tuerk
LEXOGEN, Campus Vienna Biocenter 5, 1030 Vienna, Austria

Abstract

Polymerase chain reaction (PCR) introduces a sequence specific bias which can lead to inaccurate gene and isoform quantification in RNA-Seq. This effect can be mitigated by ligating random oligonucleotides, called Unique Molecular Identifiers (UMIs), to DNA fragments before PCR. UMIs can distinguish between pre-PCR fragments with identical sequence. Hence, counting for a pool of fragments with identical sequence the number of distinct UMIs after PCR can lead to less biased results than counting the fragments in the pool. This approach fails if UMIs are not evenly distributed or if their number is not sufficiently large. Here, we focus on computational methods which yield accurate estimates of pre-PCR fragment numbers for unevenly distributed or small sets of UMIs. We develop two types of methods. One depends on a reasonably accurate model for the PCR duplication process, the other is independent of such a model.

Introduction

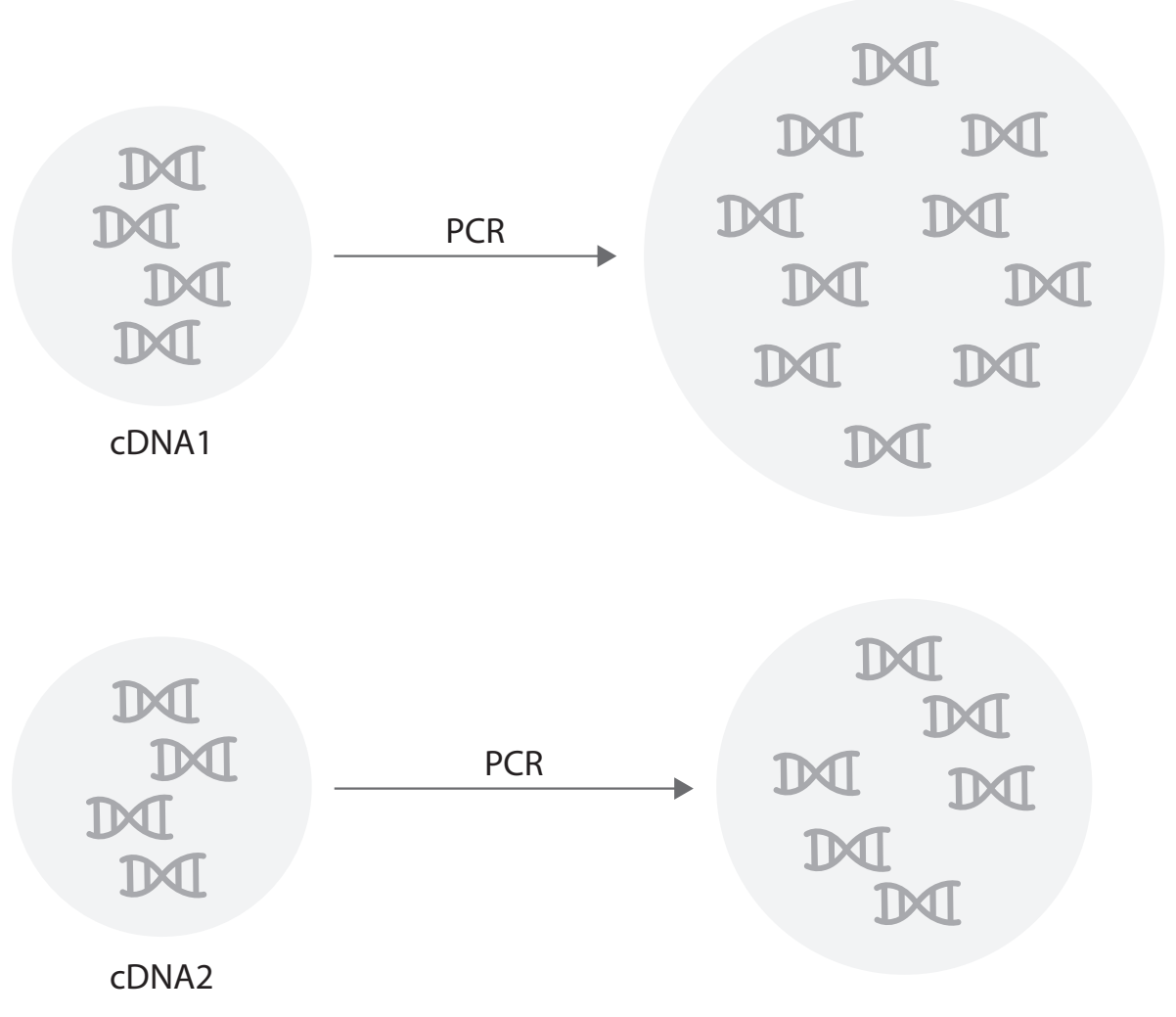

Figure 1 | PCR introduces a sequence specific bias. Two pools of identical cDNA fragments (labeled cDNA1 and cDNA2) with the same size are amplified by PCR. After PCR their sizes differ. This can lead to inaccurate gene and isoform quantification.

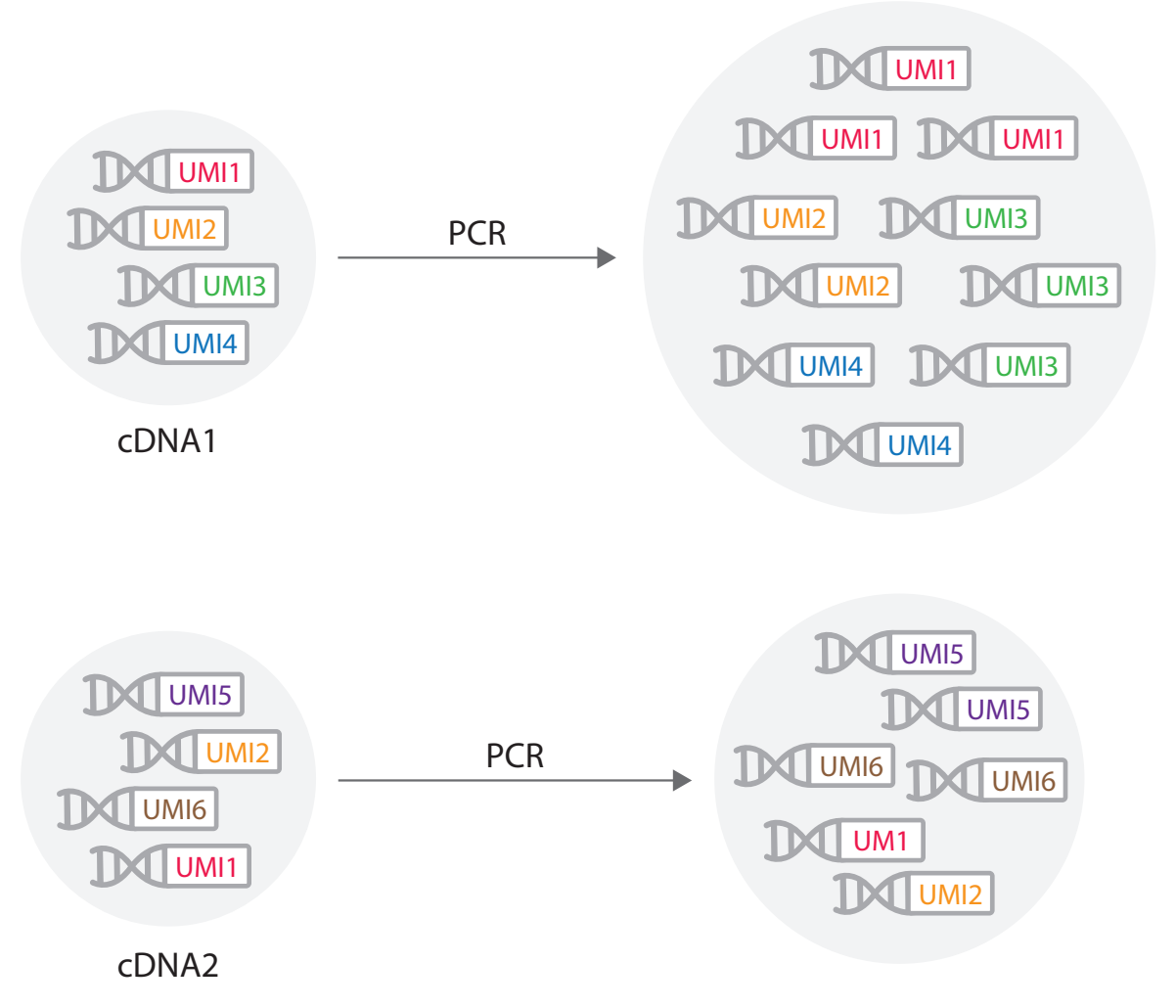

Figure 2 | UMIs distinguish between pre-PCR fragments with identical sequence. UMIs are ligated to cDNA fragments before PCR. If the set of UMIs is sufficiently large, the same UMI will not be ligated to two copies of a cDNA fragment. In this case, the number of distinct UMIs after PCR is the same as the number of fragment copies before PCR.

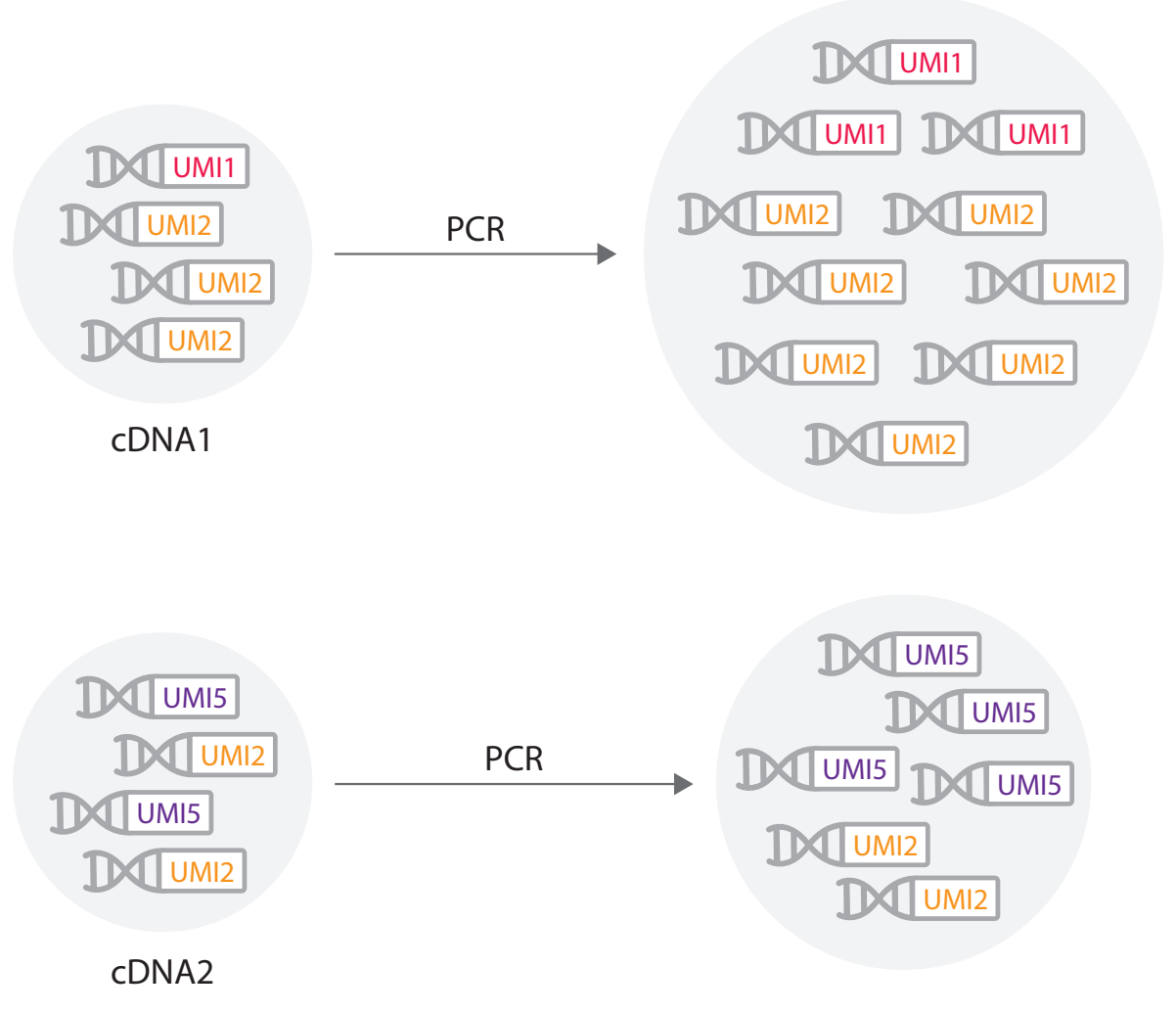

Figure 3 | Correct estimation of pre-PCR pool sizes requires large sets of evenly distributed UMIs. If either the UMI set is too small or the distribution not uniform then the same UMI can be ligated to two different fragment copies. Counting distinct UMIs after PCR will lead to an underestimate of the number of pre-PCR copies.

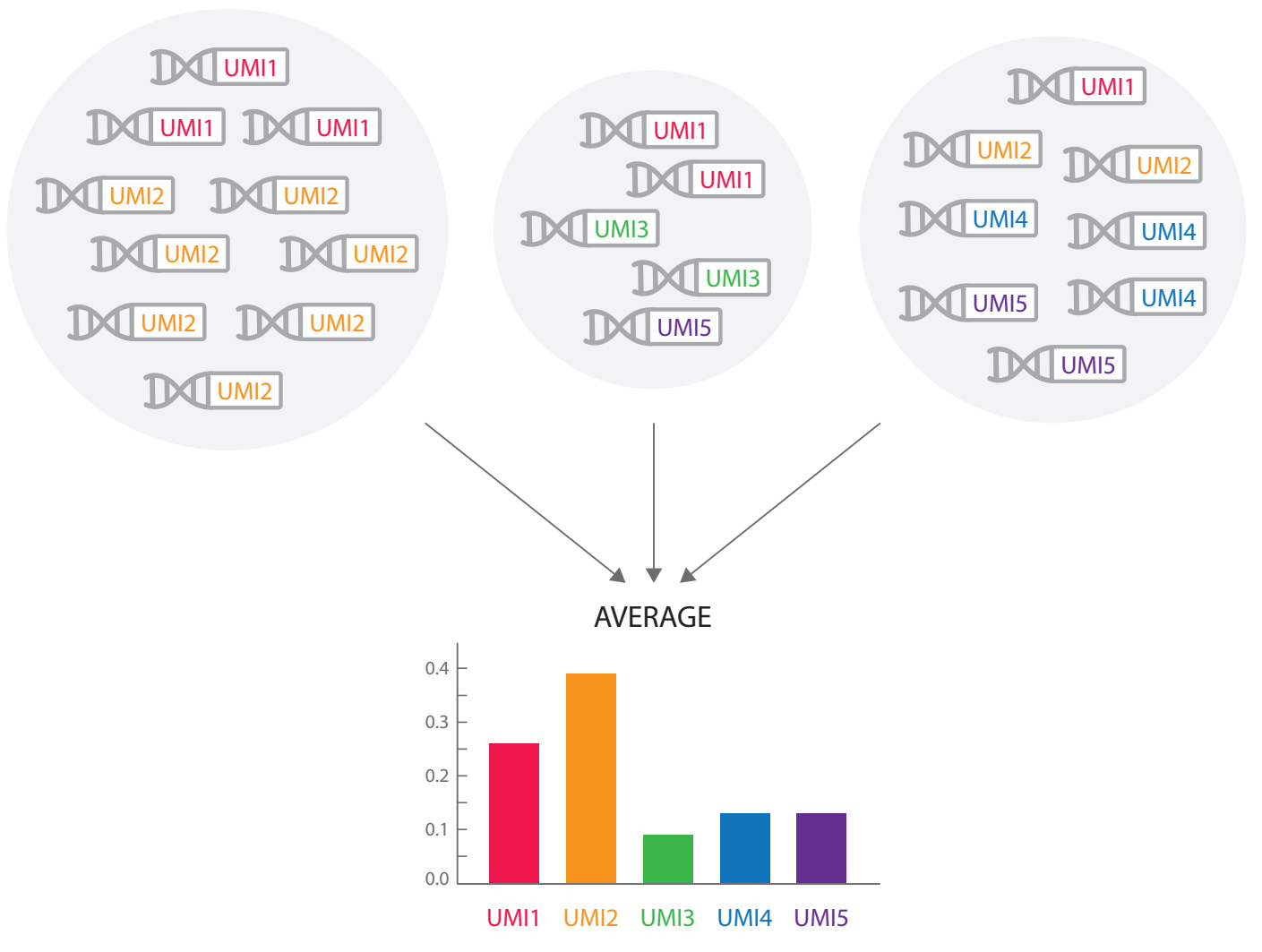

Figure 4 | Estimating the pre-PCR distribution of UMIs. For this purpose the post-PCR frequency of UMI counts is averaged over all fragment pools.

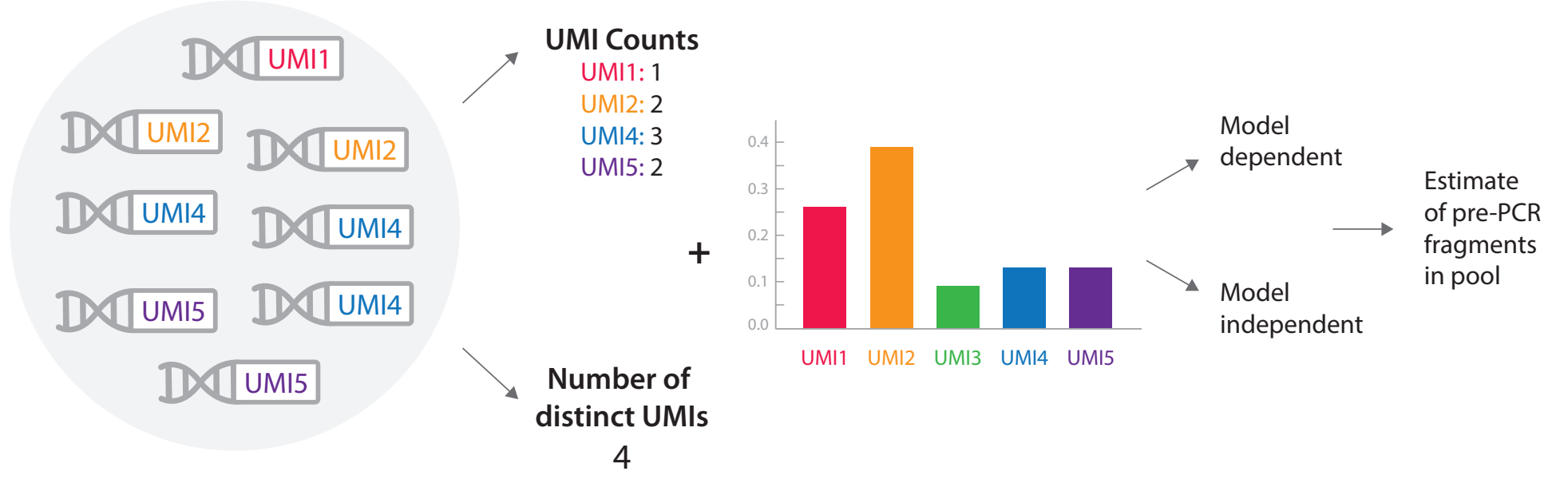

Figure 5 | Estimating the pre-PCR number of fragments in pool. Model dependent methods use UMI counts after PCR as input. Model independent methods use the number of distinct UMIs after PCR as input. Both types of methods use the estimated pre-PCR UMI distribution as input.

Model independent methods

UMI counting

UMI counting is the standard method for estimating K, the number of fragments before PCR in a pool of fragments with identical sequence [1,2]. If $\vec{n}=(n_1,\dots,n_B)$ are the counts in a pool for each UMI after PCR then UMI counting estimates K as the number of non-zero elements in $\vec{n}$, i.e.

$$K_{est} = DU(\vec{n}) = |\{i : n_i > 0\}|$$

UMI counting gives a good estimate only if the majority of fragments in a pool have UMIs ligated which are unique within the pool.

Frequency corrected UMI counting

Consider a multinomial distribution with B outcomes and outcome probability given by the pre-PCR UMI distribution $\vec{p}=(p_1,\dots,p_B)$. Frequency corrected UMI counting estimates K as the number of trials such that expected and observed number of $DU(\vec{n})$ coincide, i.e.

$$DU(\vec{n}) = E(\mathcal{DU}|K_{est},\vec{p})$$

with

$$E(\mathcal{DU}|K,\vec{p}) = \sum_{b=1}^{B}\left(1-(1-p_b)^K\right)$$

Frequency corrected UMI counting does not require unique UMIs for the majority of fragments within the pool.

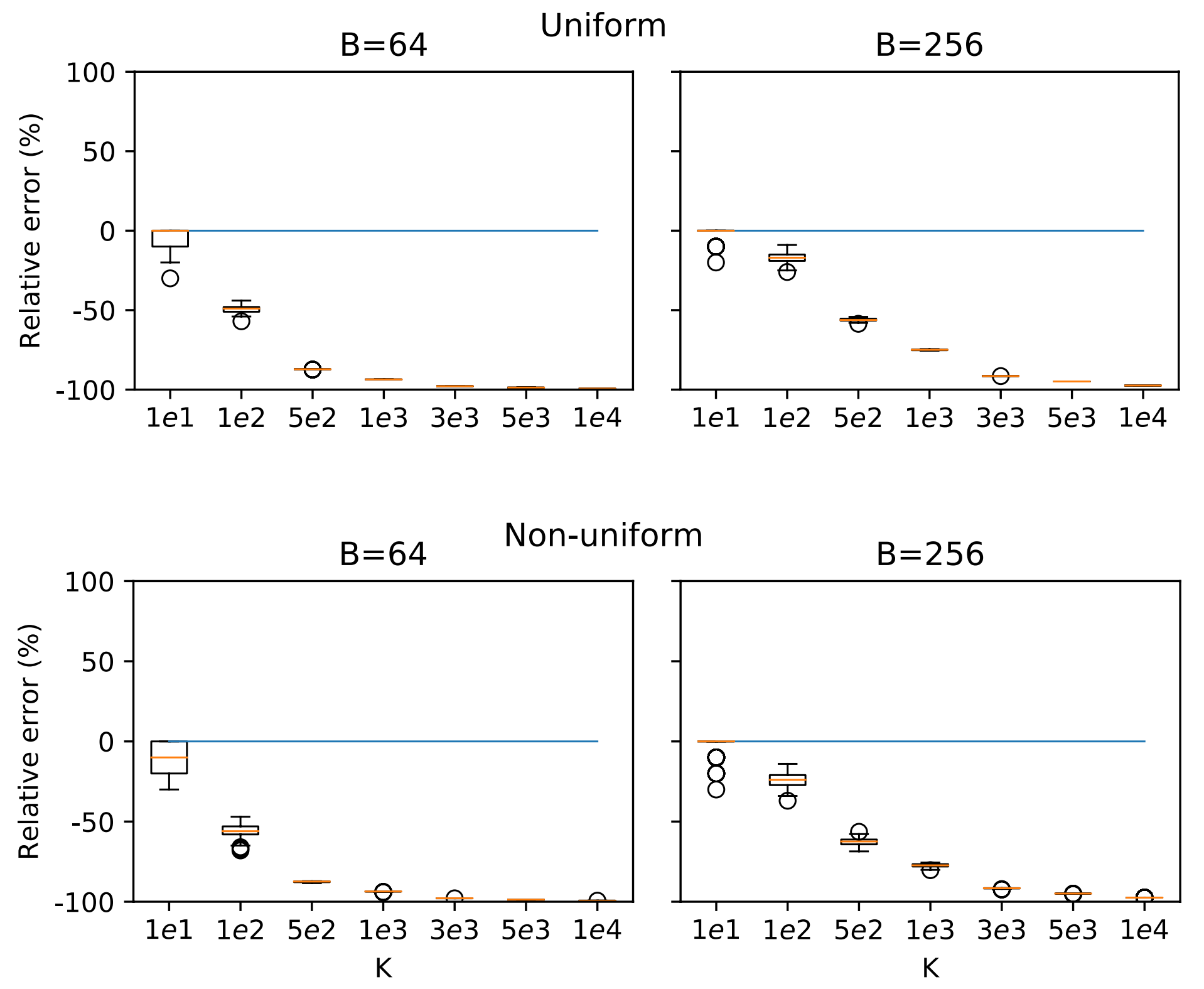

Figure 6 | Accuracy of UMI counting. Type of UMI distribution (uniform/non-uniform) and number of UMIs, B, are given in the title of the graphics. Relative error of estimated number of pre-duplication fragments, K, on y-axis, true number on x-axis.

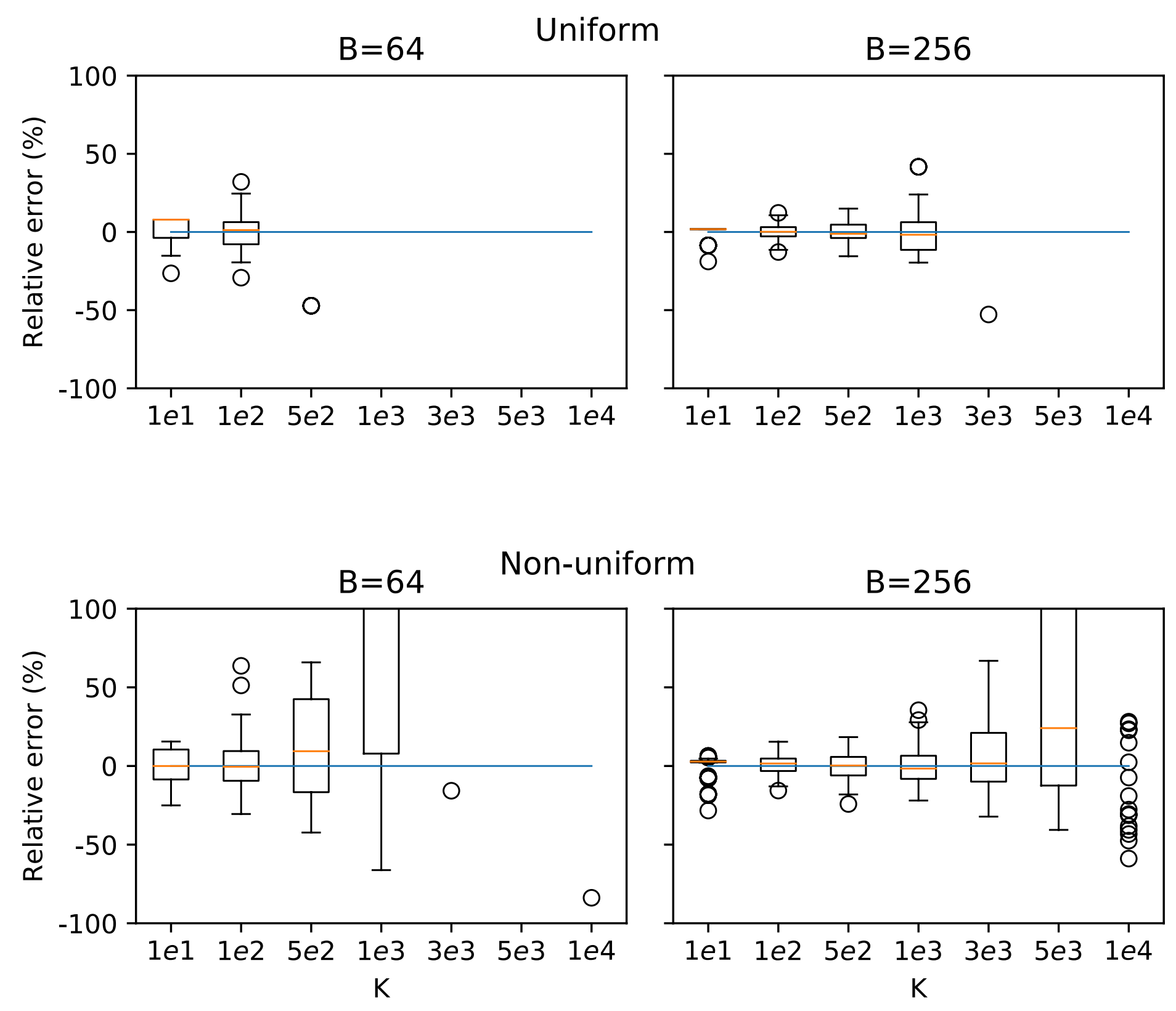

Figure 7 | Accuracy of frequency corrected UMI counting. Type of UMI distribution (uniform/non-uniform) and number of UMIs, B, are given in the title of the graphics. Relative error of estimated number of pre-duplication fragments, K, on y-axis, true number on x-axis.

Model dependent methods

- The distribution of post-PCR and pre-PCR UMI counts, $\vec{n}$ and $\vec{k}$, is the product of multinomial and PCR model with PCR efficiency p_d.
- Maximizing $p(\vec{n},|K,\vec{p},p_d)$ with respect to K and p_d is computationally not feasible. Instead assume p_d is a function of K and N and maximize.

$$\prod_{b=1,\dots,B} p(n_b|K,p_b,1-p_b,p_d)$$

PCR as a branching process

- In each cycle a fragment generates a new fragment with probability p_d.
- Model $p(n|K,p_b,p_d)$ with a multi-component mixture of the $p(n|k,p_b,p_d)$ for $k=1,\dots,K$. The $p(n|k,p_b,p_d)$ are either exact, normal or negative binomial.

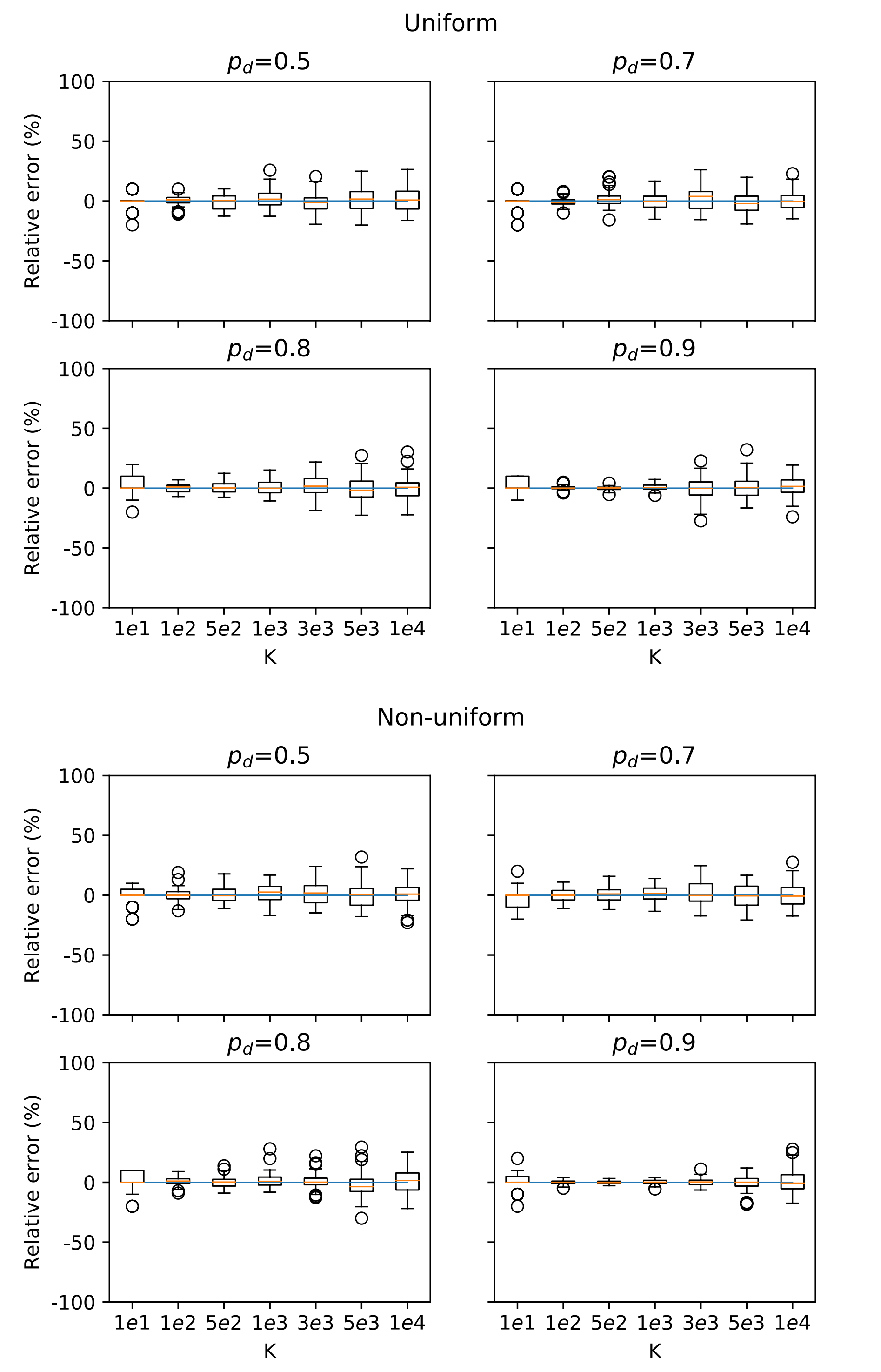

Figure 8 | Accuracy with multi-component mixture for $p(n|k,p_b,p_d)$. Data from PCR branching process with 256 UMIs and 15 cycles. Type of UMI distribution (uniform/non-uniform) and efficiency p_d are given in the title of the graphics. Relative error of estimated number of pre-duplication fragments, K, on y-axis, true number on x-axis.

Estimating pre-PCR UMI distributions

- If duplication follows a branching, Poisson or binomial process, p_b can be estimated by normalizing the post-PCR UMI counts n_b, because

$$E\left(\frac{n_b}{N}|N,K\right) = p_b$$

Experiments on RNA-Seq data

- Branching model approximates cumulative UMI count frequencies.
- Variance of counts for individual UMIs is higher than predicted by model
  » requires modification of branching model.

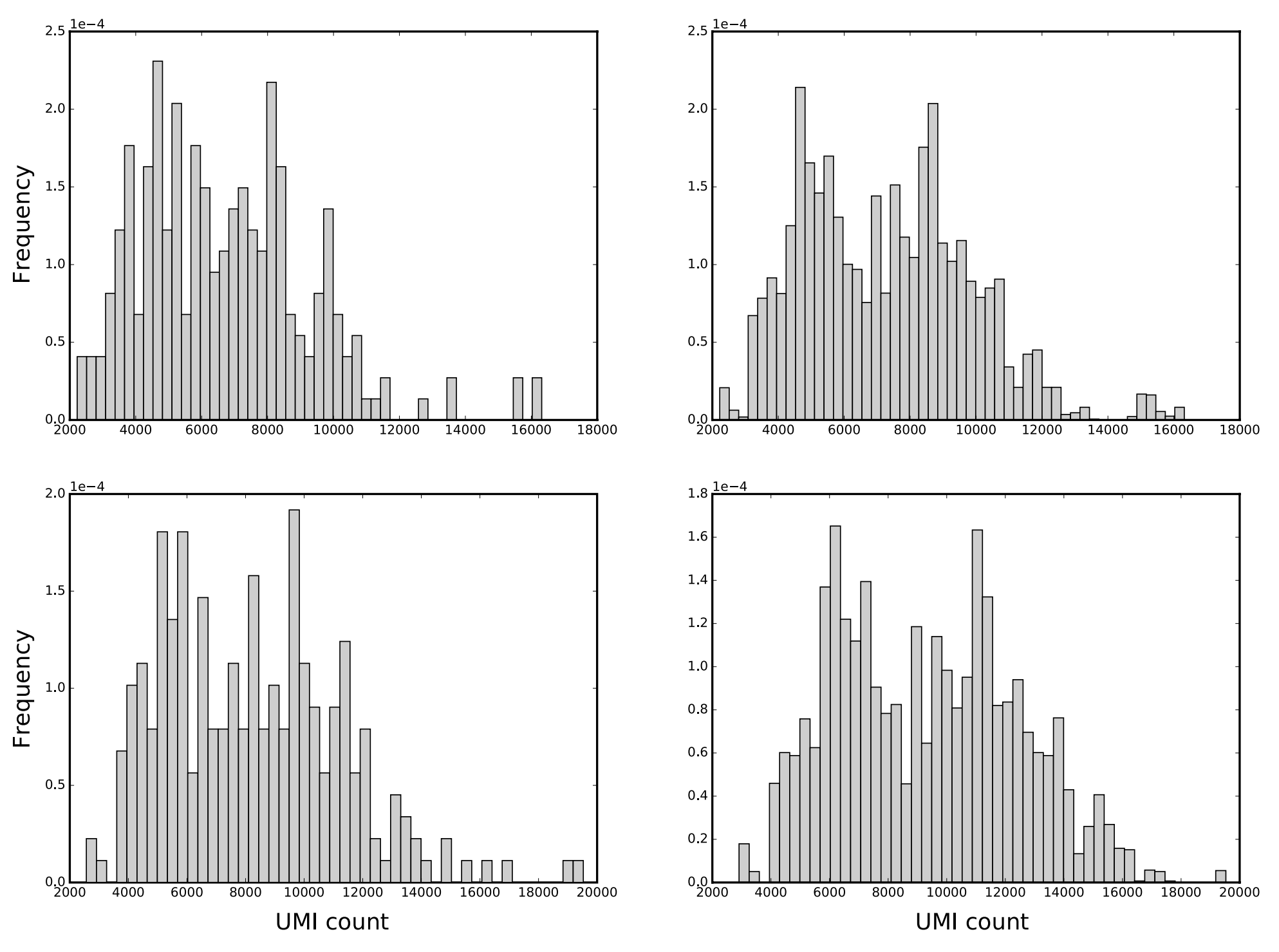

Figure 9 | Real and predicted UMI count frequencies on ERCC-00002 (top) and ERCC-00130 (bottom). Left panels show the histograms of UMI counts in RNA-Seq data. Right panels show the histograms of UMI counts generated by model for suitable parameters.

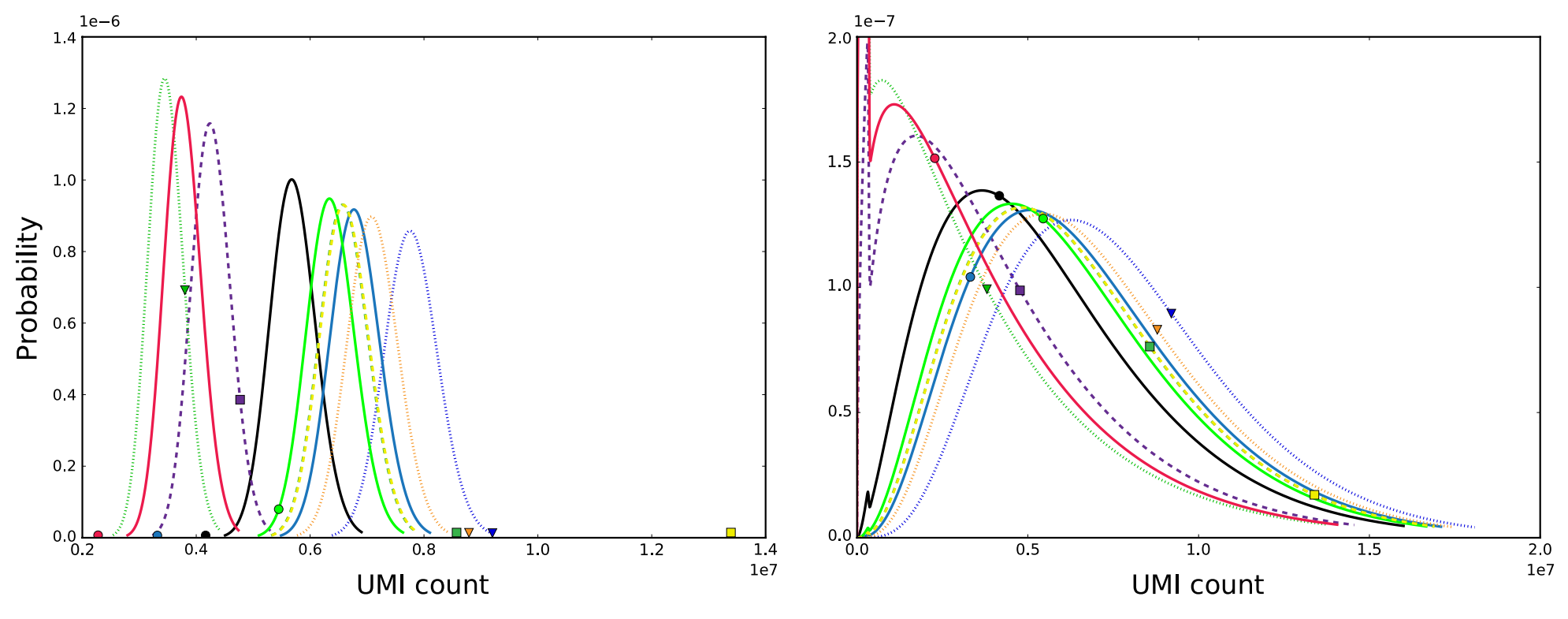

Figure 10 | Adjusting the branching model to account for high UMI count variance. Panels show a set of UMI counts (dots) and their likelihood under the original (left panel) and high variance branching model (right panel).

Conclusions

- UMI counting produces biased estimates for unevenly distributed or small numbers of UMIs.
- Frequency corrected UMI counting produces bias free estimates as long as the number of distinct observed UMIs, $DU(\vec{n})$, does not approach B.
- If $DU(\vec{n})$ approaches B:
  - UMI counting produces constant estimates independent of K.
  - Frequency corrected UMI counting fails to converge or strongly overestimates K.
- Model dependent methods:
  - Produce high accuracy estimates even if $DU(\vec{n})=B$.
  - Perform well for all efficiencies p_d.
  - Good approximation for cumulative UMI count frequencies on real RNA-Seq data.
  - High count variance of individual UMIs in real RNA-Seq data requires model modification.

References

[1] Glenn K. Fu, Jing Hu, Pei-Hua Wang, and Stephen P. A. Fodor. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proceedings of the National Academy of Sciences, 108(22):9026–9031, 2011.

[2] Glenn K. Fu, Weihong Xu, Julie Wilhelmy, Michael N. Mindrinos, Ronald W. Davis, Wenzhong Xiao, and Stephen P. A. Fodor. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proceedings of the National Academy of Sciences, 111(5):1891–1896, 2014.