

Gene expression

TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers

Florian G. Pflug^{1,*} and Arndt von Haeseler^{1,2}

¹Center for Integrative Bioinformatics Vienna (CIBIV), Joint Institute of the University of Vienna and Medical University of Vienna, Max F. Perutz Laboratories (MFPL), A-1030 Vienna, Austria and ²Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, A-1090 Vienna, Austria

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on February 2, 2018; revised on March 23, 2018; editorial decision on April 4, 2018; accepted on April 12, 2018

Abstract

Motivation: Counting molecules using *next-generation sequencing* (NGS) suffers from PCR amplification bias, which reduces the accuracy of many quantitative NGS-based experimental methods such as RNA-Seq. This is true even if molecules are made distinguishable using *unique molecular identifiers* (UMIs) before PCR amplification, and distinct UMIs are counted instead of reads: Molecules that are lost entirely during the sequencing process will still cause underestimation of the molecule count, and amplification artifacts like PCR chimeras create phantom UMIs and thus cause over-estimation.

Results: We introduce the TRUmiCount algorithm to correct for both types of errors. The TRUmiCount algorithm is based on a mechanistic model of PCR amplification and sequencing, whose two parameters have an immediate physical interpretation as PCR efficiency and sequencing depth and can be estimated from experimental data without requiring calibration experiments or spike-ins. We show that our model captures the main stochastic properties of amplification and sequencing, and that it allows us to filter out phantom UMIs and to estimate the number of molecules lost during the sequencing process. Finally, we demonstrate that the phantom-filtered and loss-corrected molecule counts computed by TRUmiCount measure the true number of molecules with considerably higher accuracy than the raw number of distinct UMIs, even if most UMIs are sequenced only once as is typical for single-cell RNA-Seq.

Availability and implementation: TRUmiCount is available at <http://www.cibiv.at/software/trumi-count> and through Bioconda (<http://bioconda.github.io>).

Contact: florian.pflug@univie.ac.at

Supplementary information: [Supplementary information](#) is available at *Bioinformatics* online.

1 Introduction

Experimental methods like RNA-Seq, ChIP-Seq and many others depend on *next-generation sequencing* (NGS) to measure the abundance of DNA or RNA molecules in a sample. The PCR amplification step necessary before sequencing often amplifies different molecules with different efficiencies, thereby biasing the measured abundances (Aird *et al.*, 2011). This problem can be alleviated by ensuring that all molecules are distinguishable before

amplification by some combination of factors comprising a *unique molecular identifier* (UMI) (Hug and Schuler, 2003; Kivioja *et al.*, 2012), which usually includes a distinct molecular barcode ligated to each molecule before amplification (Fig. 1A, colored dots; see Smith *et al.* (2017) for a more extensive history of the UMI method). After amplification and sequencing, instead of counting reads, reads are grouped by UMI, and each distinct UMI is taken to reflect a distinct molecule in the original sample (Fig. 1A). But while the number

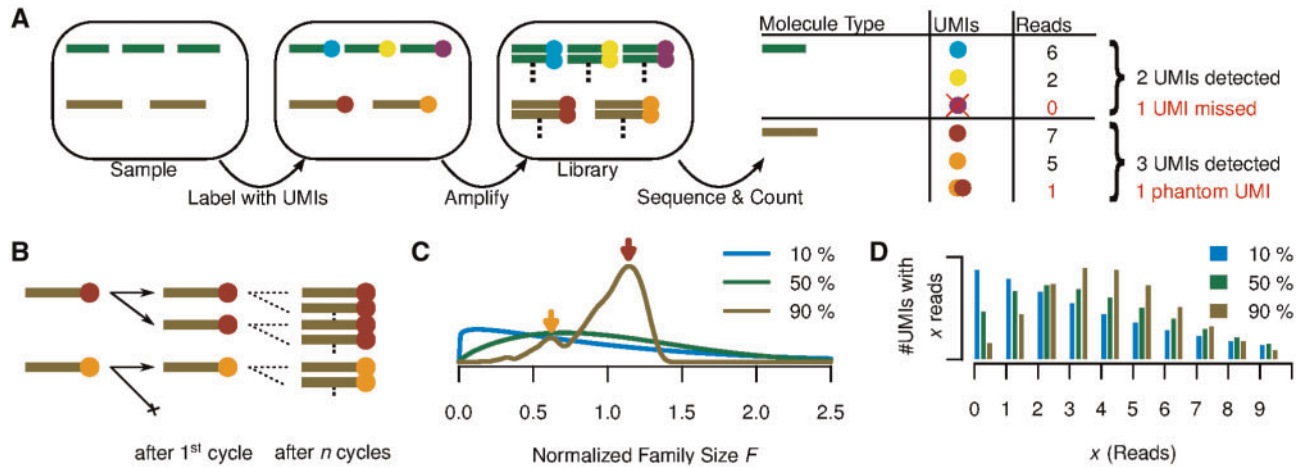


Fig. 1. (A) The relevant steps of library preparation when the UMI method is used. The sample initially contains three copies of molecule — and two copies of —, which are made unique by labelling with UMIs (●, ●, ●, ●, ●). Each of those molecules is expanded into a molecular family during amplification, and a random selection of molecules from those families is sequenced. Counting unique UMIs then counts unique molecules, unless UMIs have read-count zero (●) or phantom UMIs are produced (●). (B) PCR as a Galton-Watson branching process. Molecule — failed to be copied during the first PCR cycle and the final family size is thus reduced compared with —. (C) Normalized family size distribution for efficiency 10, 50 and 90%. The arrows mark the most likely normalized family sizes for the two molecules from (B), assuming a reaction efficiency of 90%, and taking their distinct fates during the first PCR cycle into account. (D) Distribution of reads per UMI for efficiency 10, 50 and 90% assuming $D = 4$ Reads per UMI on average

of distinct UMIs may be a better proxy for the molecule count, it is still biased, for two reasons:

- Molecules that are amplified with low efficiency will have fewer copies made, hence fewer reads per UMI, and thus a higher chance of being left entirely unsequenced (Fig. 1A, green transcript, violet UMI).
- Sequencing errors, PCR chimeras, and index miss-assignment (Sinha *et al.*, <http://www.biorxiv.org/content/early/2017/04/09/125724>) in multiplexed sequencing runs can produce *phantom UMIs* which do not correspond to any molecule in the original sample (Fig. 1A, orange/red phantom UMI).

Various methods have been proposed to counter-act these effects: Smith *et al.* (2017) proposed an algorithm for merging highly similar, erroneous versions of the same original UMI to correct for sequencing errors and single-nucleotide PCR amplification errors. To filter out more complex PCR artifacts, strand-specific UMI-labeling protocols were introduced (Shiroguchi *et al.*, 2012; Schmitt *et al.*, 2012) that allow filtering out artifacts based on whether UMIs for both strands of a template molecule were detected. A correction for molecules left entirely unsequenced is mentioned by Kivioja *et al.* (2012), but being based on the Poisson distribution, it severely under-estimates the amount of affected molecules; for their data by about an order of magnitude.

Instead of relying on sequence similarity or complicated strand-specific UMI-labeling protocols, we rely on the per-UMI read count to separate *true UMIs* (i.e. UMIs of actual molecules in the original sample) from *phantom UMIs*. Chimeric PCR products are typically produced during later reaction cycles, and can therefore be expected to have smaller copy numbers and hence a lower read-count than non-chimeric PCR products. Index miss-assignment and sequencing errors typically happen randomly, and are unlikely to produce a larger number of reads showing the same phantom UMI. For these reasons, phantom UMIs can be expected to have a markedly lower read count than most *true UMIs*, i.e. UMIs of actual molecules in the original sample.

Our bias-correction and phantom-removal algorithm TRUmiCount exploits this difference in expected read counts between phantoms and

true UMIs. It removes UMIs likely to be phantoms based on a read-count threshold, and then estimates and corrects the (gene-specific) loss, i.e. the fraction of molecules that were not sequenced or whose UMIs were mistaken for phantoms. For this correction TRUmiCount employs a model of PCR amplification that accounts for the stochasticity inherent to this amplification reaction.

2 Materials and methods

2.1 The TRUmiCount algorithm

The TRUmiCount algorithm consists of the following three steps:

1. We first filter out phantom UMIs by removing any UMI whose read count lies below a suitably chosen *error-correction threshold* (T).
2. We then estimate the loss (ℓ), i.e. the fraction of molecules that were not sequenced at all, or whose UMIs were removed by the error-correction threshold. This estimate is computed using a stochastic model of the amplification and sequencing process whose parameters are the PCR efficiency (E), and the sequencing depth (D), expressed as the average number of reads per UMI in the initial sample. From the observed distribution of reads per UMI, we estimate both (raw) gene-specific as well as library-wide values for these parameters, and compute corresponding estimates of the loss (see Section 2.2 for details).
3. Finally, we add the estimated number of lost UMIs back to the observed number of true UMIs (those UMIs with \geq threshold reads) to find the total number of molecules in the original sample. Since the loss can vary between genes, to yield unbiased counts, the correction must be based on gene-specific loss estimates. Due to the noise inherent to raw gene-specific estimates for genes with only few observed true UMIs, we employ a James-Stein-type (James and Stein, 1961) *shrinkage estimator*, adjusting the raw gene-specific parameter and loss estimates towards the library-wide ones (thus *shrinking* their difference). We choose the amount of shrinkage based on each estimate's precision, in such a way that the expected overall error is minimized (Carter and Rolph, 1974) (see Section 2.3).

2.2 Estimating the fraction of lost molecules

To estimate the loss, i.e. the fraction of molecules whose UMIs had a read count below the error-correction threshold, we model the distribution of per-UMI read counts by combining a stochastic model of PCR amplification with a model of NGS as random sampling.

2.2.1 A stochastic model of PCR amplification

To model PCR amplification, we use the *single-stranded* model of Krawczak *et al.* (1989), meaning we view PCR as a stochastic process that during each cycle duplicates each molecule independently with a particular probability E , called the reaction's *efficiency*. We further assume that a molecule is copied perfectly or not at all, i.e. that neither partial copies nor copies with a slightly different base-pair sequence are produced, that no molecules are destroyed or lost, and that the efficiency E stays constant throughout the reaction. Although this model has been extended by Weiss and von Haeseler (1997) to include the possibility of substitution errors during amplification, exhaustively modeling *all* possible sources of phantom UMIs seems futile. We therefore pursue a different approach, and model only the error-free case, trusting the error-correction threshold to remove phantoms. Over multiple cycles, each molecule is thus assumed to be expanded into a *molecular family* of identical copies. Since we use the single-stranded model, *molecule* for us always means a single-stranded piece of DNA, and we do not distinguish between a strand and its reverse complement. For our purposes, a piece of double-stranded DNA thus consists of two indistinguishable molecules.

Before amplification, we assume all molecules in the sample to be distinguishable by some UMI. During amplification, each of those molecules gives rise to a *molecular family* of (indistinguishable) copies. The initial size of such a family (i.e. the number of copies it is comprised of) is 1. During the first PCR cycle, the size increases to 2 if the single initial molecule is copied successfully, i.e. with probability E . Continuation of this process, always using all existing molecules as potential templates that are copied with probability E , produces a random sequence M_0, M_1, M_2, \dots of molecular family sizes after the 0th, 1st, 2nd, ... cycle. This sequence forms a Galton-Watson branching process (Weiss and von Haeseler, 1995), and follows the recursion

$$\begin{aligned} M_0 &= 1, & M_i &= M_{i-1} + \Delta_i \quad \text{where} \\ \Delta_i &\sim \text{Binom}(M_{i-1}, E). \end{aligned} \quad (1)$$

Although we are not aware of a way to obtain an explicit formula for the distribution of the family size M_i after i PCR cycles, the expected value and variance of M_i can be computed explicitly. According to Harris (1989, Ch. 1), Equation (5.3), $\mathbb{V}M_i = \frac{\sigma^2 m^i (m^i - 1)}{m^2 - m}$ where m and σ are the mean and SD of M_1 . In our case these are $m = 1 + E$ and $\sigma^2 = E \cdot (1 - E)$, thus we find

$$\mathbb{E}M_i = (1 + E)^i \quad (2)$$

$$\mathbb{V}M_i = \frac{1 - E}{1 + E} \cdot (1 + E)^i ((1 + E)^i - 1) \quad (3)$$

Equation 2 shows the well-known exponential growth of expected family sizes during PCR. But apart from recovering this well-known property of PCR, the Galton-Watson model also predicts the likelihood of *deviations* from this expectation due to random failures of copy operations, and by simulation allows us to find the actual distribution of M_i .

2.2.2 The normalized family size F

Due to the exponential growth of the expectation of M_i , the distribution of M_i depends heavily on the PCR cycle count i . That

dependency, however, affects mostly the *scale*, not the *shape* of the distribution of M_i . To see the effect on the shape more clearly, the effect on the scale is removed by replacing M_i with a re-scaled version which has an expected value of one,

$$\tilde{M}_i = \frac{M_i}{\mathbb{E}M_i} = \frac{M_i}{(1 + E)^i}. \quad (4)$$

These re-scaled family sizes can be sensibly compared across PCR cycles. We observe that with growing cycle counts, the additional stochasticity introduced by each additional cycle drops rapidly. The re-scaled family size after the first cycle varies by a factor of two depending on whether the (single) copy operation during the first cycle succeeds or fails. Later on there are more templates to copy from, and thus the success or failure to copy any particular molecule averages out, making the behavior of the reaction more deterministic. Finally, $\tilde{M}_i \approx \tilde{M}_{i+1}$, because the family size M_i increases during each cycle almost exactly by a factor of $1 + E$, which matches the decrease of the re-scaling factor in \tilde{M}_i . This informal argument can be turned into a formal proof (see Harris (1989), Ch. 1, Th. 8.1) of the convergence of the re-scaled family size as i tends towards ∞ , which allows us to remove the cycle count as a parameter entirely from what we call the *normalized family size*

$$F = \lim_{i \rightarrow \infty} \tilde{M}_i. \quad (5)$$

Although there is again no explicit formula known for the distribution of the normalized family size F , we find its variance from Equations (3–5) to be

$$\mathbb{V}F = \frac{1 - E}{1 + E}. \quad (6)$$

To quickly evaluate the density $f_F(x; E)$ of the distribution of F for a particular normalized family size x given reaction efficiency E , we interpolate using 2D polynomial interpolation (Akima, 1996) between pre-computed densities for different reaction efficiencies between 0 and 100% at different family sizes between 0 and 50 (see Supplementary Section S1.1 for details).

2.2.3 Modeling the sequencing process

The normalized family size distribution models the abundance of molecules with a particular UMI. To model the read count of a particular UMI after sequencing (i.e. the number of reads stemming from a particular pre-amplification molecule), we model NGS with a Poissonian sampling model (Marioni *et al.*, 2008). This amounts to assuming that (i) each individual copy has the same probability of being sequenced, (ii) this probability is small compared to the sequencing depth and (iii) there were many (distinguishable) original molecules. We further assume that a UMI is *on average* represented by D reads. Then the read count C of a UMI with known normalized molecular family size F is Poisson distributed,

$$\begin{aligned} C | F &\sim \text{Poisson}(F \cdot D), \\ \mathbb{P}(C = k | F) &= e^{-F \cdot D} \frac{(F \cdot D)^k}{k!}. \end{aligned} \quad (7)$$

In general, however, the exact family size F of any particular UMI is unknown—we only know the *distribution* of F . To compute the probability of a UMI having k reads, we average over all possible family sizes $x \in [0, \infty)$, weighting them with their respective density $f_F(x; E)$ in the distribution of the normalized family size F ,

$$\mathbb{P}(C = k) = \int_0^\infty \mathbb{P}(C = k | F = x) \cdot f_F(x; E) dx. \quad (8)$$

We note that while $\mathbb{P}(C = k)$ depends on D and E , we omit these dependencies for brevity of notation. To compute the probabilities $\mathbb{P}(C = k)$, we integrate numerically using the midpoint rule on the grid of family sizes x for which $f_F(x; E)$ was pre-computed. For the mean and variance of C we find the explicit expressions

$$\mathbb{E}(C) = D, \quad \mathbb{V}(C) = D + D^2 \frac{1 - E}{1 + E}. \quad (9)$$

Since we impose an error-correction threshold T and drop UMIs with fewer than T reads, the read-count distribution we actually observe is a *censored* version of C where the possible outcomes $C < T$ are removed. For the mean and variance of this censored distribution with threshold T we write

$$\mathbb{E}(C | C \geq T) = \frac{\sum_{k=T}^\infty k \cdot \mathbb{P}(C = k)}{\mathbb{P}(C \geq T)}, \quad (10)$$

$$\mathbb{V}(C | C \geq T) = \frac{\sum_{k=T}^\infty (k - \mathbb{E}(C | C \geq T))^2 \cdot \mathbb{P}(C = k)}{\mathbb{P}(C \geq T)}. \quad (11)$$

To compute $\mathbb{E}(C | C \geq T)$, we rewrite the infinite sum in Equation (10) to $\mathbb{E}(C) - \sum_{k < T} k \cdot \mathbb{P}(C = k)$, and similarly for $\mathbb{V}(C | C \geq T)$.

2.2.4 Computing the loss

The expected loss ℓ is the expected fraction of true UMIs that either remain completely unsequenced, or that are removed by the error-correction threshold. Since we treat each per-UMI read count, and hence each UMI's fate (to be filtered or not) as independent stochastic quantities, this expected fraction is simply the probability that a single UMI has a read-count below the threshold T , i.e.

$$\ell = \mathbb{P}(C < T). \quad (12)$$

2.3 Correcting for lost molecules

Given n^{obs} experimentally observed UMIs (after applying the error-correction threshold T to filter out phantoms) and their read count vector $\mathbf{c} = (c_1, \dots, c_{n^{\text{obs}}})$, we estimate the reaction efficiency E and the mean number of reads per UMI D . We use the *method of moments*, i.e. we find E and D such that the predicted mean equals the sample mean \hat{m} of \mathbf{c} , and the predicted variance its sample variance \hat{v} . Since we only take observed UMIs with at least T reads into account, we must compute the predictions using the censored distribution, i.e. find E , D such that $\hat{m} = \mathbb{E}(C | C \geq T)$ and $\hat{v} = \mathbb{V}(C | C \geq T)$

If $T = 0$, i.e. if \hat{m} and \hat{v} reflect the *uncensored* mean respectively variance, these equations can be solved explicitly by inverting Equation (9), which yields the method of moments estimates $\hat{D} = \hat{m}$ and $\hat{E} = \frac{1 - \hat{v}}{1 + \hat{v}}$, where $\hat{v}' = \frac{\hat{v} - \hat{m}}{\hat{m}}$ limited to the interval $[0, 1]$.

If $T > 0$, we solve the system of equations numerically to find E and D (see Supplementary Section S1.2). With these parameter estimates, we then compute an estimate $\hat{\ell}$ of the loss ℓ using Equation (12), and use it to correct for the expected number of lost molecules. Assuming that we observed n^{obs} UMIs and given $\hat{\ell}$, we estimate the total number of molecules in the original sample to have been

$$\hat{n}^{\text{tot}} = \frac{n^{\text{obs}}}{1 - \hat{\ell}}. \quad (13)$$

2.3.1 Gene-specific estimates and corrections

Since the reaction efficiency E and depth D , and hence also the loss, will usually vary between individual genes (or other genomic features of interest), to correct the observed number of transcripts of some gene $g \in 1, \dots, K$ for the loss, a *gene-specific* loss estimate $\hat{\ell}_g$ should be used. In principle, such estimates are found by applying the described estimation procedure to only the UMIs found for transcripts of gene g , i.e. by computing a gene-specific mean \hat{m}_g and variance \hat{v}_g of the number of reads per UMI, and solving for parameters E and D to find a gene-specific \hat{E}_g^{raw} and \hat{D}_g^{raw} , and computing $\hat{\ell}_g^{\text{raw}}$ using Equation (12). If the number n_g^{obs} of observed UMIs (i.e. transcripts) stemming from gene g is large, a correction based on $\hat{\ell}_g^{\text{raw}}$ yields an (approximately) unbiased and accurate estimate of the total number of transcripts of that gene. But if n_g^{obs} is small, the error of the estimator $\hat{\ell}_g^{\text{raw}}$ easily exceeds the variability of the true gene-specific value ℓ_g between genes. In such cases, correcting using the *library-wide* estimate $\hat{\ell}^{\text{all}}$ computed from *all* UMIs found in the library will yield a more accurate (although biased) estimate of the total number transcripts of gene g .

Interestingly, by combining these two flawed estimators of the true gene-specific loss ℓ_g , we obtain a *shrinkage estimator* $\hat{\ell}_g^{\text{shr}}$ that improves upon both in terms of *mean squared error* (MSE), see Carter and Rolph (1974) Equation (2.4),

$$\hat{\ell}_g^{\text{shr}} = \lambda_g \cdot \hat{\ell}_g^{\text{raw}} + (1 - \lambda_g) \cdot \hat{\ell}^{\text{all}}. \quad (14)$$

The gene-specific coefficient λ_g determines how much the raw gene-specific estimate is *shrunk* towards the global estimate, and its optimal choice (with respect to the MSE) depends on the variances the two constituent estimators. To determine the optimal λ_g we make the following assumptions about these estimators:

- i. the library-wide estimate $\hat{\ell}^{\text{all}}$ is a good proxy for the true *average* loss taken over all genes $1, \dots, K$. This seems reasonable given the size of a typical library, comprising millions of UMIs.
- ii. the estimator variance of the raw gene-specific estimator $\hat{\ell}_g^{\text{raw}}$ depends only on the number n_g^{obs} of observed UMIs for gene g , and does so in an inversely proportional manner. This is certainly true asymptotically for large numbers of observations, for small numbers Supplementary Figure S4 shows this approximation to be reasonable.

We write s for the variance of the true loss between genes (i.e. for the mean squared difference of ℓ_g and $\hat{\ell}^{\text{all}}$), and u for the proportionality constant between the estimator variance of $\hat{\ell}_g^{\text{raw}}$ and $1/n_g^{\text{obs}}$. According to Carter and Rolph (1974) Equation (2.4ff) the optimal choice for λ_g is then

$$\lambda_g = \frac{s}{s + u/n_g^{\text{obs}}} \quad (15)$$

To compute the gene-specific shrinkage estimators $\hat{\ell}_g^{\text{shr}}$, it remains to find constants u and s . Towards that end, we observe that the expected squared deviation of the raw gene-specific loss estimate $\hat{\ell}_g^{\text{raw}}$ from its average $\bar{\ell} = \frac{1}{n} \sum_{g=1}^K \hat{\ell}_g^{\text{raw}}$ is the total variance of $\hat{\ell}_g^{\text{raw}}$, which is comprised of the between-gene variance s and the estimator variance u/n_g^{obs} , or in other words $\mathbb{E}(\hat{\ell}_g^{\text{raw}} - \bar{\ell})^2 = s + u/n_g^{\text{obs}}$.

This allows us to estimate s and u using *least squares regression*, i.e. by minimizing

$$\sum_{g=1}^K ((\hat{\ell}_g^{\text{raw}} - \bar{\ell})^2 - s - u/n_g^{\text{obs}})^2 \cdot w(n_g^{\text{obs}}). \quad (16)$$

Without weighting (i.e. for $w(n) = 1$), the considerable drop in magnitude of $(\hat{\ell}_g^{\text{raw}} - \bar{\ell})^2$ as n_g^{obs} increases would allow genes with small

number of observations to yield an unduly large influence over the estimates. Since it is the genes with a low to moderate number of observations that benefit from shrinking, some modest bias of this sort is actually desired—but not as strong a bias as $w(n) = 1$ exhibits, and one not so purely focused on genes with very few observations. We therefore use the weights $w(n) = \frac{n}{1+n/100}$, which initially increase linearly with the number of observations, but eventually converge to 100 instead of increasing further. This has the desired effect of shifting the focus away from rarely observed genes, and concentrating it on genes with a moderate number of observations.

3 Results

3.1 PCR stochasticity versus efficiency

During PCR amplification, each uniquely labeled molecule is amplified into a molecular family of indistinguishable copies. Random successes or failures to copy molecules during early reaction cycles lead to a variation in the final family sizes (Fig. 1B), even between identical (expect for their molecular barcode) molecules. As the family size of each initial molecule grows, the proportion of successful copy operations approaches the efficiency E , therefore reducing the amount of noise added by each additional cycle. The total number of cycles thus has little influence on the final family size distribution, and is therefore not a parameter of our model. For the same reason, a plateau effect (i.e. diminishing reaction efficiency during later cycles) has little effect on the final family size distribution, and is thus not included in the model. The final distribution does, however, depend strongly on the reaction efficiency, with fluctuations in family size decreasing as the efficiency grows towards 100% (Fig. 1C).

For efficiencies close to 100%, most molecular families are thus of about average size, except for those ($\sim 100 - E$ percent) families for which the first copy failed. These are about half the average size, and form a distinct secondary peak in the family size distribution (Fig. 1C, brown curve). We emphasize that due to this, even at efficiencies close to 100%, the distribution still shows considerable dispersion, meaning that even at high efficiencies stochastic PCR effects are not negligible. At lower efficiencies, the family sizes vary even more wildly, as extreme family sizes (on both ends of the scale) become more likely (Fig. 1C, blue and green curves).

If we add sequencing to the picture, i.e. combine the stochastic PCR model outlined above with a model of sequencing as random Poissonian sampling (Marioni *et al.*, 2008), the variability of per-UMI read counts (Fig. 1D) then has two sources—the variability of molecular family sizes and the Poissonian sampling introduced by sequencing. Although the latter is reduced by increasing the sequencing depth, the former is independent of the sequencing depth but is reduced by increasing the reaction efficiency. For all reasonable error-correction thresholds T the predicted fraction of true UMIs filtered out by the error-correction step thus grows with diminishing efficiency E .

3.2 Model validation and phantom UMI removal

To validate our model of amplification and sequencing, we compared the predicted distribution of per-UMI read counts to the distribution observed in two published RNA-Seq datasets. Kivioja *et al.* (2012) labeled and sequenced transcripts in *Drosophila melanogaster* S2 cells using 10 bp random molecular barcodes from the 5' end. Shiroguchi *et al.* (2012) labeled and sequenced transcript fragments in *E.coli* cells on both ends, using (on each end) one of 145 molecular barcodes carefully selected to have large pairwise edit distances. The Y-shaped sequencing adapters used in the *E.coli*

experiment were designed such that each strand of a labeled double-stranded cDNA molecule produces a related but distinguishable molecular family.

To see whether our algorithm offers an advantage over existing UMI error-correction strategies, we pre-filtered the observed UMIs in each of the two replicates of these datasets using the following existing algorithms: We first merged UMIs likely to be erroneously sequenced versions of the same molecule, using the algorithm proposed by Smith *et al.* (2017). For the *E.coli* experiment we also removed UMIs for which the complementary UMI corresponding to the second strand of the same initial template molecule was not detected, as proposed by Shiroguchi *et al.* (2012). See Supplementary Section S1.4 for details on the analysis pipeline we used.

To this pre-filtered set of UMIs we then applied our algorithm. For each dataset, we manually chose an error-correction threshold by visually comparing read-count distribution and model prediction for different thresholds, and picking the lowest threshold that yielded a reasonably good fit. Above the error-correction threshold (Fig. 2a, black bars), the observed library-wide distribution of reads per UMI closely follows the model prediction, and the *E.coli* data even shows traces of the secondary peak that represents molecules not duplicated in the first reaction PCR cycle. Choosing a different threshold will change the number of UMIs surviving the error-correction filter, but has little influence on the estimated reaction efficiency and on the estimated total number of UMIs after loss correction (Supplementary Fig. S1). We thus conclude that our model captures the main stochastic behavior of the amplification and sequencing processes, and accurately models the read-count distribution of true UMIs.

The UMIs removed by our filter, i.e. those with fewer reads than the error-correction threshold demands, (Fig. 2A, gray bars) are over-abundant compared to our prediction. This over-abundance increases further as per-UMI read counts drop, indicating the existence of a group of UMIs with significantly reduced molecular family sizes. While we may expect some systematic variation of family sizes between true UMIs (on top of the stochastic variations that our PCR model predicts), we would expect these to be gradual and not form distinct groups. We conclude that the extra UMIs causing the observed over-abundance are indeed phantoms that are rightly removed by our algorithm. We note that none of these phantoms were removed by either the UMI merging algorithm of Smith *et al.* (2017), or (for the *E.coli* data) by filtering UMIs for which the complementary UMI (representing the second strand of the template molecule) was not detected.

For the *D.melanogaster* data, our loss estimates of 9% (R1) and 8.8% (R2) are about a magnitude higher than the 1% (R1) and 2% (R2) estimated using the (truncated) Poisson distribution suggested by Kivioja *et al.* (2012). Given that using a Poisson model amounts to assuming a 100% efficient duplication of molecules during each PCR cycle, this severe underestimation by the Poisson model shows that the inherent stochasticity of the PCR cannot be neglected.

3.3 Gene-specific quantification bias

The gene-specific (shrunk) estimates for amplification efficiency, average reads per UMI, and loss that our algorithm produces, vary between genes to different degrees (Fig. 2B). We observe the smallest amount of variation for the average number of reads per UMI (Fig. 2B, left)—the estimates of this parameter are virtually identical for a large majority of genes, and differs only for a few outliers.

The estimated amplification efficiencies on the other hand can vary substantially between genes (Fig. 2B, middle). For the two

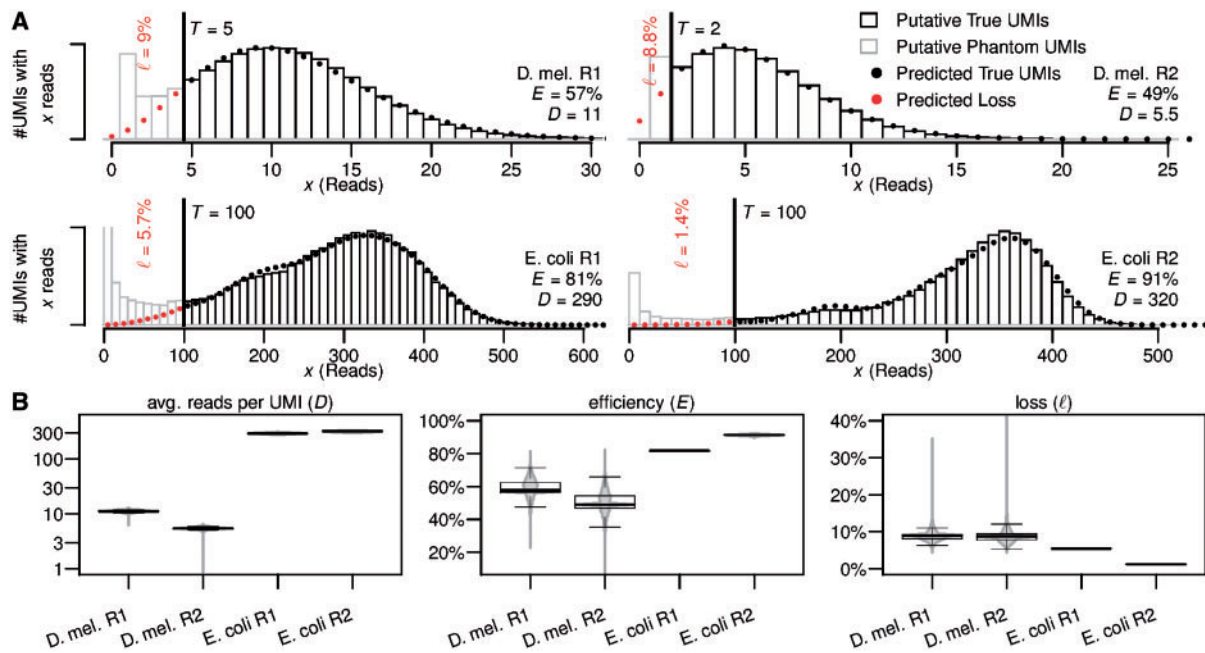


Fig. 2. (A) Observed and predicted library-wide distribution of reads per UMI and parameter and loss estimates. Filtered UMIs (grey bars, left of threshold T) are over-abundant and thus assumed to contain both phantom and true UMIs (red dots). UMIs surviving the filter (black bars) closely follow the predicted distribution (black dots) and are assumed to be true UMIs. (B) Variability of the (shrunk) model parameters and resulting loss between genes. Includes parameter for 7481 detected genes in *D.melanogaster* R1, 8001 genes in R2, 2380 genes in *E.coli* R1 and 2308 genes in R2. (Color version of this figure is available at *Bioinformatics* online.)

D.melanogaster replicates the range is 22–81% (R1) and 1–83% (R2). Considering that in this experiment only the 3' ends of transcripts were sequenced, and all fragments contributing to a gene hence share a similar sequence composition, this is not unexpected. These differences in efficiency cause the loss to vary heavily between genes as well (Fig. 2B, right), between 4 and 35% for R1 and between 4 and 89% for R2 (which has a much lower overall sequencing depth). Without gene-specific loss corrections, abundance comparisons between genes will thus suffer from systematic quantification bias against certain genes of up to $\approx 35-4\% = 31\%$ for R1 and up to $\approx 85\%$ for R2. The larger amount of systematic bias in *D.melanogaster* R2 is caused by the two-fold reduction of the number of reads per molecules in R2 compared with R1—due to the lower number of reads per molecule in R2, the same difference in amplification efficiencies between two genes translates into a larger difference of the number of lost molecules in R2 compared with R1.

In contrast, fragments from all parts of the transcript were sequenced in the *E.coli* experiments, and together with the high sequencing depth (≈ 300 reads per UMI), we now expect little variations of efficiency, and small and highly uniform losses across genes. Our efficiency and loss estimates reflects this (Fig. 2B, middle and right), and as the lack of outliers shows, they do so even for genes with only few UMIs. Yet for these genes, the raw (unshrunk) gene-specific estimates are noisy (Supplementary Fig. S3), proving that shrinking the raw estimates successfully reduces the noise to acceptable levels.

3.4 Bias-corrected transcript counts

To further verify the accuracy of the corrected transcript counts computed by our algorithm, we conducted a simulation study. We use the (loss-corrected) estimated total transcript abundances of *D.melanogaster* replicate 1, rounded to 10, 30, 100, 300, 1000, 3000 or 10 000 molecules as the true transcript abundances.

We then simulated amplification and sequencing of these transcripts, using for each gene the previously estimated gene-specific efficiency and average number of reads per UMI (Fig. 2B). To the resulting list of UMIs and their read-counts for each gene we applied our algorithm to recover the true transcript abundances (threshold $T = 5$ as before), and determined for each gene the relative error of the recovered abundances compared with the simulation input.

Figure 3 shows these relative errors (i) if no correction is done (ii) if the correction is based solely on the raw gene-specific loss estimates (i.e. no shrinkage) and (iii) for the full TRUmiCount algorithm (i.e. using shrunk loss estimates). The uncorrected counts systematically under-estimate the true transcript counts, in 50% of the cases by at least $\approx 10\%$, independent of the true number of transcripts per gene. And even at high transcript abundances, the relative error still varies *between* genes, biasing not only absolute transcript quantification, but also relative comparisons between different genes. The counts corrected using raw gene-specific estimates are unbiased and virtually error-free for strongly expressed genes, but exhibit a large amount of additional noise for weakly expressed genes. The full TRUmiCount algorithm successfully controls the amount of added noise, and shows no additional noise for weakly expressed genes, while still being unbiased and virtually error-free for more strongly expressed genes.

3.5 Performance for low sequencing depth

To assess the performance of the TRUmiCount algorithm at low sequencing depths such as are common for single-cell RNA-Seq experiments, we ran a second simulation with gene-specific depth parameters scaled such that the average across all genes was $D = 1$ read per molecule (Fig. 4). Under these conditions, the most likely outcome for a single molecule in the initial sample is to remain unsequenced (39% of molecules), and only 27% of molecules are found in more than one read. The library-wide efficiency estimate of 57%

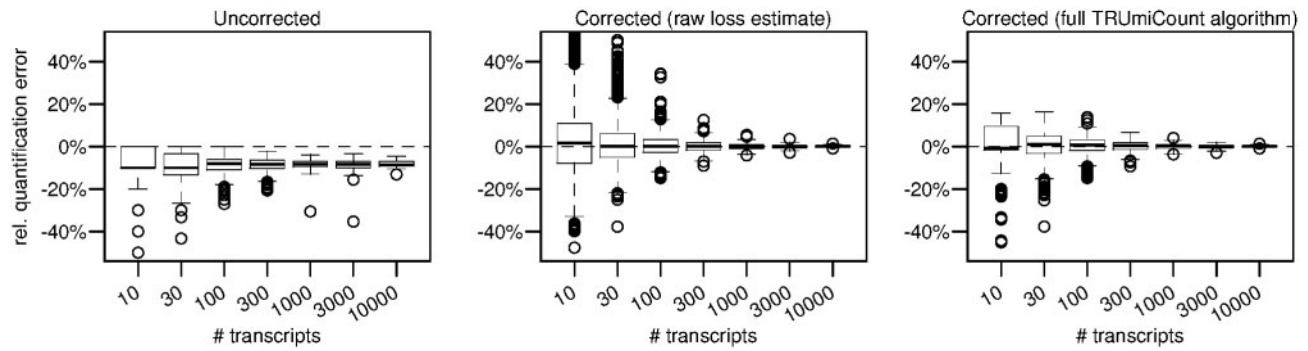


Fig. 3. Relative error of estimated total number of transcripts depending on the true number of transcripts. Left panel uses the observed number of UMIs without any correction. Middle panel uses the raw gene-specific loss estimates to correct for lost UMIs. Right panel uses the full TRUMiCount algorithm employing shrunken gene-specific loss estimates to correct for lost UMIs

(Fig. 4A) is nevertheless accurate, and identical to the one computed for the full dataset (*D.melanogaster* R1) that the simulation was based on (Fig. 2A).

For the relative error of the corrected transcript counts we observed a roughly 2-fold increase at low-sequencing depth (Fig. 4B) compared with the situation at original sequencing depth (Fig. 3, right), but still no systematic over- or under-estimation. We estimate that Poissonian sampling effects account for about a $\sqrt{1 - 0.09}/\sqrt{1 - 0.39} \approx 1.2$ -fold increase of the relative errors. The rest is probably due to the parameter estimation problem becoming harder at lower sequencing depths, particularly for weakly expressed genes. For more strongly expressed genes, the relative quantification error again drops towards zero, similar to the behavior at original sequencing depth.

4 Discussion

The TRUMiCount algorithm we presented successfully removes the biases inherent in raw UMI counts, and produces unbiased and low-noise measurements of transcript abundance, allowing for unbiased comparisons between different genes, exons and other genomic features. It does so even in the presence of various types of phantom UMIs and varying amplification efficiencies, both between samples and along the genome. Compared to other error-correction techniques, it is not restricted to particular types of phantom UMIs, or to a special Y-shaped design of the sequencing adapters.

Our model of the amplification and sequencing process is mechanistic, and its two parameters have an immediate physical interpretation. They can both be determined from the experimental data without the need for either guesses or separate calibration experiments. The TRUMiCount algorithm thus does not require any changes to library preparation over the basic UMI method. By inspecting the estimated parameters—in particular the amplification efficiency, the amplification reaction itself can be studied. For example, by estimating model parameters separately for sequenced fragments of different lengths, the drop of reaction efficiency with increasing fragment lengths can be quantified (Supplementary Fig. S2).

Although TRUMiCount requires that libraries are sequenced sufficiently deeply to detect at least some UMIs more than once, it can also deal with cases where a molecule is on average detected only by a single read, which is common e.g. for single-cell RNA-Seq. The performance of TRUMiCount is reduced a bit in such situations, but it still offers an improvement over uncorrected counts by removing

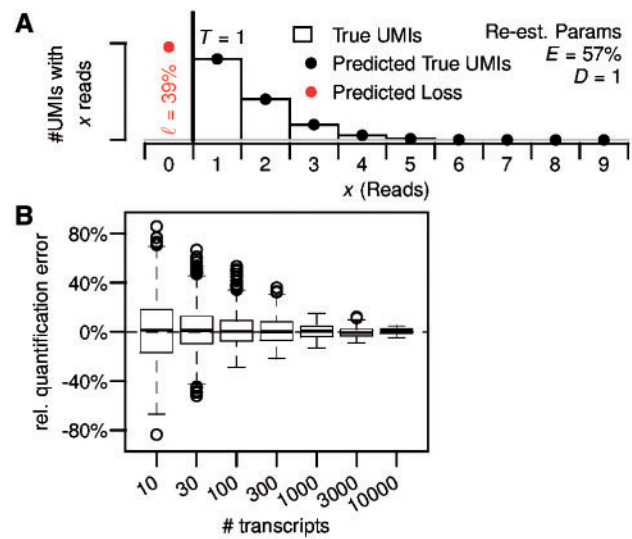


Fig. 4. TRUMiCount performance for low sequencing depth. (A) Overall distribution of observed and predicted reads per UMI for an average of $D = 1$ read per molecule. (B) Relative error of estimated total number of transcripts for different true numbers of transcripts and $D = 1$ read per molecules on average. (Color version of this figure is available at *Bioinformatics* online.)

systematic biases. For even lower read counts, where gene-specific bias correction becomes infeasible, we expect that TRUMiCount could still be used to correct for cell-specific (instead of gene-specific) biases, thus reducing the amount of technical noise when comparing absolute transcript counts of the same gene between individual cells.

The TRUMiCount algorithm can thus help to increase the accuracy of many quantitative applications of NGS, and by removing biases from comparisons between genes can aid in the quantitative unraveling of complex gene interaction networks. To make our method as easily accessible as possible to a wide range of researchers, we provide two readily usable implementations of our algorithm. Our R package *gupcR* enables a flexible integration into existing R-based data analysis workflows. In addition, we offer the command-line tool TRUMiCount which is designed to work in conjunction with the *UMI-Tools* of Smith *et al.* (2017). Together they provide a complete analysis pipeline which produces unbiased transcript counts from the raw reads produced by a UMI-based RNA-Seq experiment (<http://www.cibiv.at/software/trumicount>).

Acknowledgements

We would like to thank Armin Djamei and Simon Uhse for introducing us to the concept of UMIs, all members of the *Center for Integrative Bioinformatics Vienna* (CIBIV), in particular Luis Paulin-Paz, Celine Prakash and Philipp Rescheneder for their valuable feedback throughout all stages of this project, and Olga Chernomor for sharing her insights into shrinkage estimators. We also thank the “DoktoratsKolleg RNA Biology” for funding this work.

Funding

This work was supported by the Austrian Science Fund (FWF): W1207-B09.

Conflict of Interest: none declared.

References

- Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
- Akima,H. (1996) Algorithm 760; rectangular-grid-data surface fitting that has the accuracy of a bicubic polynomial. *ACM Trans. Math Softw.*, **22**, 357–361.
- Carter,G.M. and Rolph,J.E. (1974) Empirical bayes methods applied to estimating fire alarm probabilities. *J. Am. Stat. Assoc.*, **69**, 880.
- Harris,T.E. (1989). *The Theory of Branching Processes*. Dover, New York.
- Hug,H. and Schuler,R. (2003) Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J. Theor. Biol.*, **221**, 615–624.
- James,W. and Stein,C. (1961) Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379. University of California Press, Berkeley, CA.
- Kivioja,T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
- Krawczak,M. *et al.* (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucleic Acids Res.*, **17**, 2197–2201.
- Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
- Schmitt,M.W. *et al.* (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14508–14513.
- Shiroguchi,K. *et al.* (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 1347–1352.
- Smith,T. *et al.* (2017) UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
- Weiss,G. and von Haeseler,A. (1995) Modeling the polymerase chain reaction. *J. Comput. Biol.*, **2**, 49–61.
- Weiss,G. and von Haeseler,A. (1997) A coalescent approach to the polymerase chain reaction. *Nucleic Acids Res.*, **25**, 3082–3087.