Supplementary methods and results

Ligase reaction mechanism and kinetics

A series of elegant biochemical investigations in the mid 1970 to late 1970s(1–7), complemented by structural analyses in the first decade of this century(8–12), described the basic elements of the *intermolecular ligation reaction* between two (oligo/poly)nucleotide reactants. These investigations led to the proposal of a three step, ping-pong ordered kinetics(13) mechanism for this reaction(4, 6, 13, 14) . In the first step, RNA ligase binds *ATP* and transfers a nucleotidyl group from the trinucleotide cofactor to form a covalently adenylated enzyme (*E-AMP*) while releasing pyrophosphate (*PPi*). The second step involves the transfer of the AMP group from the enzyme to the 5′ phosphoryl of the donor (*D*) to form a pyrophosphate linked adenylate intermediate(1–4), *E·App-D*. The final step in the reaction is the formation of the phophodiester bond between the 3′ of the acceptor (*A*) and the 5′ of the donor sequence to form the product *AD* while *AMP* is released. In this reaction scheme, the presence of the acceptor sequence allows the final step of the reaction to proceed in the forward direction, because donor sequences that are too short to form circular products are not ligated unless an acceptor is present(7, 15). Furthermore, all three steps of the reaction are reversible(2, 16) supporting the following reaction scheme:

$$E + ATP \underset{k_{-1}}{\overset{k_1}{\leftrightharpoons}} E \cdot ATP \underset{k_{-1a}}{\overset{k_{1a}}{\leftrightharpoons}} E - AMP + PPi$$

$$E - AMP + D \underset{k_{-2}}{\overset{k_2}{\leftrightharpoons}} E \cdot App - D \underset{k_{-2a}}{\overset{k_{2a}}{\leftrightharpoons}} E + App - D$$

$$E \cdot App - D + A \underset{k_{-3}}{\overset{k_3}{\leftrightharpoons}} E \cdot App - D \cdot A \underset{k_{-3a}}{\overset{k_{3a}}{\leftrightharpoons}} AMP + E + AD$$

Even though this is a simplified schema e.g. there is evidence to suggest that the first step proceeds by two non-covalent enzyme, magnesium, ATP complexes(12), it nonetheless provides a minimal, accurate description of the stoichiometries involved and is compatible with existing data. When libraries are constructed for RNA-seq applications, the sequences of interest are ligated to the 3′ and 5′ either in a single step, or more commonly with the sequential addition of the 3′ and 5′ adapters. In either case, the RNA sequences in the biological complex play the role of both acceptor (3′ adapter ligation) and donor (5′ adapter ligation). It is not immediately clear how the reaction efficiency depends on the concentration of the RNA sequences to be ligated or even the composition of the sample. The latter may be an important factor since the second(first) step of the 3′(5′) adapter ligation is subject to multiple substrate inhibition, i.e. by all the RNA sequences in the sample of interest. We heuristically argue below that despite the highly non-linear nature of the ordinary differential equations describing the ligase reaction system, the efficiencies of the reaction itself may be treated as approximately constant for any given library preparation protocol.

When an excess of pre-adenylated adapters is used to drive 3′ ligation in the absence of *PPi*, the surplus of donors will establish a quasi-state with the non-covalent intermediate $E{\cdot}App\text{-}D$ and the adenylated enzyme *E-AMP* (note that due to the absence of *PPi*, the first step in the reaction will not be reversed). Simultaneously, the excess of donors will also drive the third step of the reaction forward. Hence after a brief transient period, the yield of the complex three-step reaction will be determined by the final step. The kinetic analysis of the last step is simplified by the large excess of adapters, which implies that their concentration is effectively constant over the course of the reaction, and the ping-pong mechanism of the reaction. These features allow one to use the quasi-steady state approximation(17, 18) to formulate a Michaellis-Menten type of equation for

the last step. Furthermore, the apparent reaction velocity and Michaellis-Menten constant depend on the kinetic constants (k_3, k_{-3}, k_4) and the nearly constant concentration of adenylated donor/enzyme complex. A similar argument may be used to simplify the analysis of the 5′ adapter ligation. In the latter case, the excess of ATP and acceptor (adapter) sequences will ensure that after a short transient the concentration of adenylated enzyme stays constant during the course of the reaction. This is because the free enzyme is consumed in the first step, is regenerated in the final step. At (quasi-)steady state, the output of the entire sequence of reactions will be determined by generation of the enzyme/adenylated donor(the original RNA sequence with the covalently attached 3′)/acceptor(5′ adapter) ternary complex. As in the case of 3′ ligation, the availability of excess *acceptor* sequences ensures that the reaction is driven forward.

Multiple substrate inhibition and ligation efficiency

When preparing libraries from a heterogeneous mixture of n RNAs, the ligation reaction of each of the distinct species is subject to competitive inhibition by all the other RNAs in the reaction mix. Substrate inhibition in both 5′ and 3′ ligation may be analyzed under the framework of multi-substrate competition for ordered ping-pong, bisubstrate reactions(19). Since one of the reactants (donor in 5′ ligation, acceptor in 3′ ligation) is in such large excess, its concentration may be absorbed into the maximum velocity and apparent Michaellis-Menten constant of the reaction. The expression for the reaction velocity is identical to the those obtained for the more familiar case of competitive inhibition in mono-substrate reactions. We will work out the expression for the efficiency of the 3′ adaptor ligation, noting that similar arguments may be used to derive an equivalent expression for the efficiency of the 5′ ligation. Under Michaelis-Menten kinetics, the reaction velocity for the i^{th} species (taken as a referent) may be written as:

$$V_i = \frac{V_i^{max} X_i}{X_i + K_M^i \left(1 + \sum_{j \neq i} \frac{X_j}{K_M^j}\right)} = \frac{V_i^{max} X_i}{K_M^i \left(1 + \sum_i \frac{X_j}{K_M^j}\right)} \quad \text{(Eq.1)}$$

The sum appearing in the denominator is over a large number of RNA-species and hence on accounts of the law of large numbers, would be expected to converge to the quantity $nE\left[\frac{X}{K_M}\right]$, where the expectation is taken over all RNA species present in the reaction sample. Using a Taylor series argument (see Appendix A), we can approximate this expectation with the ratio of the corresponding expectations:

$$V_i \approx \frac{V_i^{max} X_i}{K_M^i \left(1 + n \frac{E[X]}{E[K_M]}\right)} = \frac{V_i^{max} X_i}{K_M^i \left(1 + \frac{C_{Total}(t)}{E[K_M]}\right)} \quad \text{(Eq.2)}$$

where $C_{Total}(t)$ is the total RNA concentration in the reaction volume that has not been ligated up to time *t*. To characterize the order of the reaction, we need to consider the quantitative relation between the *initial concentration* of RNA against the expected Michaellis Menten constant. It is precisely this initial relation that determines the order of the reaction at all subsequent time points.

Small RNA input is less than 200 ng for most protocols in current use (e.g. the Illumina TrueSeq protocol suggests a minimum of 10-50 ng of purified small RNA), and can be as low as 1-10 ng for the Ion Total RNA-Seq Kit v2 and the NEXTflex Small RNA-Seq Kit v3. Such low input materials are typical of applications for biomarker discovery in clinical samples and extracellular fluid media. Ligation reaction volumes are between 10-20 microliters, yielding a final concentration of between 50.6-101.3 nM if 10 ng of small RNA, with an average length of 29b is diluted in 10-20 microliters. This should be considered a rather high estimate of the concentration, because of issues quantifying small amounts of nucleic acid in dilute samples(20),

and the unknown size distribution of small RNA. In our own experiments with synthetic mixes, the maximum amount of miRNA was 100 femtomole in 10 microliters, yielding an initial (total) concentration of 10 nM. This figures should be considered in light of the reported Michaelis – Menten constants for the enzymatic catalysis of the RNA ligation reaction which have been reported to be in the micromolar to millimolar range(2, 3, 13, 15, 21–23). As the concentration of RNA input is orders of magnitudes lower than the corresponding Michaelis-Menten constant, we can approximate the reaction velocity as a first order reaction:

$$V_i \approx \frac{V_i^{max} X_i}{K_M^i} \qquad \text{(Eq.3)}$$

As the concentration of reactants declines during the reaction's progression, the first order kinetic approximation becomes in fact more accurate. Hence, we can approximate the reaction velocity at all time points with the first order kinetic law. Integration of (Eq.3) up to a given reaction time yields T_R the following expression for the ligation reaction yield:

$$X_i \left\{ 1 - \exp\left(-\frac{V_i^{max}}{K_M^i} T_R \right) \right\} = X_i f_i^{5'} \qquad \text{(Eq.4)}$$

Examination of this equation reveals that the efficiency depends on the kinetic parameters of the ligation reaction (maximum velocity and Michaelis-Menten constant) which in turn depend on characteristics of the adapter and the sequence to be ligated, the reaction time the total RNA input of the protocol and concentration of adapter cofactors. To the extent that these are kept constant, we postulate that reaction efficiencies will be constant for a given RNA and independent of the initial abundance X_j or even the composition of the sample. A similar argument may be used to establish the near constancy of the 3′ ligation ($f_i^{3'}$). This allows us to

write the following, deterministic relationship between the input (X_i) and output (Λ_i) of the ligase reaction:

$$\Lambda_i = X_i f_i^{5'} f_i^{3'} = X_i f_i \qquad \text{(Eq.5)}$$

In the latter equation, the overall efficiency f_i is the product of the efficiencies of the two consecutive adapter ligations. Our hypothesis of constant ligation efficiencies, f_i, for any given library preparation protocol, leads to two testable predictions: the first concerns the form of a *linear-quadratic* (LQ) relation between the mean and the variance of sequence counts. We provide a mathematical proof of the LQ relation in a subsequent section. The second prediction specifies that sequencing data from an equimolar mix may be used to estimate relative efficiencies that are universally applicable to all datasets created with the same protocol. Consequently, these *bias correction factors* may be used to adjust the abundance estimates from other datasets in which these sequences were present in variable amounts.

Mean and Variance Relationships in Stochastic Branching Processes for PCR reactions

It is well recognized that the accumulation of the products of the PCR reaction may be stochastically modelled with a Galton-Watson branching process(24). We will consider the exponential phase of the PCR reaction, in which the reaction efficiency is constant. This assumption is likely to be verified in small RNA sequencing experiments, in which both the amount of starting material and the number of cycles are relatively small, e.g. less than 10 ng and fewer than 20 cycles respectively. It follows that the abundance of the i^{it} RNA species after the j+1 cycle, may be expressed by the Markovian relation(25–27):

$$\mathrm{L}_i^{j+1}|\mathrm{L}_i^j, q_i = \mathrm{L}_i^j + Binomial\left(\mathrm{L}_i^j, q_i\right) \qquad \begin{matrix} j = 0, N-1 \\ \mathrm{L}_i^0 = \Lambda_i \end{matrix} \tag{Eq.6}$$

The corresponding *mean* and *variance* of the counts at the N^{th} cycle, *conditional on the initial abundance* and *reaction efficiency* may be derived by standard branching theory results(24, 25) as:

$$\mu_i^N|\Lambda_i, q_i = \Lambda_i(1+q_i)^N \tag{Eq.7}$$

$$\sigma^{2}{}_i^N|\Lambda_i, q_i = \Lambda_i \frac{1-q_i}{1+q_i}(1+q_i)^N[(1+q_i)^N - 1] \cong \tag{Eq.8}$$

$$\Lambda_i \frac{1-q_i}{1+q_i}(1+q_i)^N(1+q_i)^N = \frac{1-q_i}{\Lambda_i(1+q_i)}{\mu_i^N}^2 \tag{Eq.9}$$

The error incurred in approximating (Eq.8) with (Eq.9) is very small, i.e. less than 0.8% for 12 or more PCR cycles and efficiency greater than 50%. Hence the variance of the distribution of the amplified PCR products is proportional to the square power of the mean at each cycle. Equivalently, the coefficient of variation is constant for a given efficiency of the reaction and an initial starting abundance. The values of the coefficient of variation for different combinations of starting materials and reaction efficiencies are shown in Supplementary Figure 1. The coefficient of variation is inversely related to the initial abundance and is substantially less than 5% for copy numbers that are likely to be present in ligation reaction volumes.

Numerical Approximation to the truncated Normal mixed Poisson distribution

Special function formulation: The truncated normal mixed Poisson distribution was investigated by several authors in the 1960s(28–30); interest in that distribution subsequently waned, since no

practical applications or “real data which follows this distribution”(30) were known at the time. We will follow the textbook notation (See page 398 in (31)), which was introduced in 1967(29) and define the probability mass function (p.m.f.) of this distribution as:

$$p(x|\mu,\phi) = exp\left(\frac{\mu^2\phi}{2} - \mu\right)\left(\mu\sqrt{\phi}\right)^x \frac{I_x\left(\mu\sqrt{\phi} - \frac{1}{\sqrt{\phi}}\right)}{I_0\left(-\frac{1}{\sqrt{\phi}}\right)} \quad \text{(Eq.10)}$$

where $I_0(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty exp\left(-\frac{t^2}{2}\right)dt$ and $I_r(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \frac{(t-x)^r}{r!} exp\left(-\frac{t^2}{2}\right)dt$.

(Eq.10) follows from the notation by Kemp and Kemp(29) by substituting the expression for the variance of the PCR reaction $\mu^2\phi$ for the variance of the Gaussian mixing distribution.

Using the definition of the cumulative density function of the standard normal distribution Φ(x), the relations between the *I* functions and the probability function $Hh_r(x)$: $Hh_r(x) = \sqrt{2\pi}I_r(x)$ (29) and between the probability function and the U parabolic cylinder function (12.7.7 (32, 33)) we can re-express the p.m.f. in (Eq.10) as

$$p(x|\mu,\phi) = exp\left(\frac{\mu^2\phi}{4} - \frac{\mu}{2} - \frac{1}{4\phi}\right)\frac{\left(\mu\sqrt{\phi}\right)^x}{\sqrt{2\pi}} U_{x+\frac{1}{2}}\left(\mu\sqrt{\phi} - \frac{1}{\sqrt{\phi}}\right)\frac{1}{1-\Phi\left(-\frac{1}{\sqrt{\phi}}\right)} \quad \text{(Eq.11)}$$

$$\approx exp\left(\frac{\mu^2\phi}{4} - \frac{\mu}{2} - \frac{1}{4\phi}\right)\frac{\left(\mu\sqrt{\phi}\right)^x}{\sqrt{2\pi}} U_{x+\frac{1}{2}}\left(\mu\sqrt{\phi} - \frac{1}{\sqrt{\phi}}\right)$$

We can considerably simplify (Eq.11), by taking advantage of the rapid convergence of the term involving the cumulative density function of the standard normal distribution to one. For just 3 copies and a PCR efficiency of >0.8 the last term in (Eq.11) differs from unity by less than 11/10000000 with the difference becoming smaller for higher abundances and efficiencies.

Therefore, this term may be ignored, reducing the problem of calculating the p.m.f. of the mixed Poisson distribution to that of computing the parabolic cylinder function U.

Approximation by the Negative Binomial I and the Linear Quadratic Normal Model: We undertook extensive numerical simulations to assess whether (Eq.11) can be numerically approximated by less complex functions. We assessed the degree to which the Negative Binomial I (NBI) distribution (as defined in the main text) and the Linear Quadratic Normal family (LQNO) defined as:

$$Normal\left(\mu, \sigma_{LQ}^2\right), \qquad \sigma_{LQ}^2 = \mu(1 + \phi\mu) \tag{Eq.12}$$

To carry out these evaluations, we simulated 10000 unique combinations of μ, ϕ uniformly in the range of $[1,10^6]$ x $[10^{-10},1]$ using a Sobol sequence for quasirandom numbers. Each combination of μ, ϕ defines a unique Truncated normal mixing distribution for a standard Poisson variate. Subsequently we simulated 50 values for the x in the interval $[\mu - 10\sigma_{LQ}, \mu + 10\sigma_{LQ}]$ for each such variate (a total of 500,000 evaluation points) and evaluated (Eq.11) using an arbitrary precision numerical library for the computation of the hyperbolic cylinder function(34). In case this approach failed to generate a numerical value, we numerically evaluated the integral defining the mixed distribution (Eq. 9 in the main text) by an arbitrary precision (20 decimal digits) double exponential integrator(35, 36). The value of the p.m.f. was then contrasted against the value of the probability density function of the NBI and the LQNO at the same values of μ, ϕ, x. The median (interquartile) absolute difference between the NBI and the truncated normal mixed Poisson distribution was 3.4×10^{-7} ($4\ 10^{-8}$ - $3.77\ 10^{-6}$), while that for the LQNO were 2.62×10^{-5} ($1.1\ 10^{-3}$-$4.1\ 10^{-7}$). In addition, we evaluated whether there are parameter combinations for which the LQNO distribution provides a better approximation to the mixed Poisson distribution

compared to the NBI. To do so, we built a logistic regression classifier to model the probability that the mixed Poisson p.m.f. was closer in absolute value to the LQNO than the NBI. This analysis (Supplementary Figure 4) demonstrates that depending on the combination of dispersion, mean value and signal generation probability, either one of these simpler distributions may provide a numerically superior approximation to the truncated normal mixed Poisson model. This analysis suggests that both the NBI and LQNO may be retained as alternatives to the more complex expression of (Eq.11) when modeling of RNA-seq experiments, since it will not always be evident the regime (upper yellow and lower red regions) in which an experiment is operating on.

Distributional regression models for the analysis of differential expression

The extension of the proposed framework to the problem of assessing differential expression changes between experimental conditions is straightforward. The two regression submodels in (Eq.12) **of the main text** (not to be confused with Eq.12 of the supplement), are augmented to account for differences in the abundances, i.e. the $\log X_i$ terms between two or more experimental conditions. This augmentation, takes the form of a simple linear model under the working assumption of $Q_{i,j} = Q_j$ and the reparameterization $Q_j = Q_j \times r_j$, when the relevant quantities are expressed in logarithmic scale. In this model, the mean parameter of the i^{th} sequence, from the j^{th} experiment in the k^{th} experimental condition, may be written as a function of the fold expression change of that sequence ($\Delta_{i,k}$) in that state relative to the (log-)expression against the referent state ($\log X_{i,0}$):

$$\begin{aligned} \log \mu_{i,j,k} &= \log X_{i,k} + \log Q_j + \log f_i \\ &= \log X_{i,0} + \log f_i + \log Q_j + \Delta_{\mathrm{i,k}} \\ &= \log \mu_{i,0} + \log Q_j + \Delta_{\mathrm{i,k}} \end{aligned} \qquad \text{(Eq.13)}$$

In moving from the second to the third equation, we adopt a viewpoint which considers the expression values in the referent condition as of a secondary interest relative to the differences in expression. This viewpoint allows us to absorb the ligase bias factors $\log f_i$ into the terms for the expression of the sequence in the referent state $\log X_{i,0}$ into composite terms $\log \mu_{i,0}$. Consequently, one may estimate log-fold changes even for sequences for which the bias correction factors are not available e.g. from calibration equimolar datasets.

Further structure may be imposed on (Eq.13) by adopting a *mixed effects* modeling perspective. This approach restricts the three terms to conform to a Normal distribution around their overall (grand) mean, while the latter ($\log \mu_{0,0}, \log Q_0, \Delta_{\mathrm{k}}$) are allowed to vary freely:

$$\begin{aligned} \log \mu_{i,0} &= \log \mu_{0,0} + m_{i,0}, \;\; m_{i,0} \sim Normal(0, \sigma^2_{\mu_0}) \\ \log Q_j &= \log Q_0 + w_j, \qquad w_j \sim Normal(0, \sigma^2_Q) \\ \Delta_{\mathrm{i,k}} &= \Delta_{\mathrm{k}} + \delta_{i,k}, k \neq 0, \;\; \delta_{i,k} \sim Normal(0, \sigma^2_k) \end{aligned}$$

This mixed effects model is a flexible approach which can readily accommodate global differential changes in expression level (Δ_{k}) that shift the expression level of every sequence by the same amount, while allowing sequence-specific variations ($\delta_{i,k}$) around this pattern. Furthermore, the model accommodates variation in the sequence counts of the referent state ($m_{i,0}$), relative to mean sequence count in that state ($\log \mu_{0,0}$), while also explicitly representing technical, library-specific variation in PCR efficiency and signal generation probabilities (w_j). In many applications, this technical variation, as quantified by the σ^2_Q term, will be of substantially smaller magnitude than the variation in differential expression (σ^2_k) and referent sequence count

($\sigma^2_{\mu_0}$). Hence, one may simplify this model by setting $\sigma^2_Q \to 0$, which implies that $w_j \to 0$. Then one may absorb the term $\log Q_0$ into the mean expression value of the referent group, $\log \mu_{0,0}$ which becomes the grand mean (intercept) term of the regression model: $\alpha = \log \mu_{0,0} + \log Q_0$. The reduced model assumes the much simpler form:

$$\begin{aligned} \log \mu_{i,j,k} &= \alpha + \Delta_{\mathrm{k}} + m_{i,0} + \delta_{i,k} \\ m_{i,0} &\sim Normal(0, \sigma^2_{\mu_0}) \\ \delta_{i,k} &\sim Normal(0, \sigma^2_k) \end{aligned}$$

A similar expression may be recovered, *mutatis mutandis* for the sub-model of the $\log \phi_{i,j,k}$ parameter. The GAMLSS model comprised of the two reduced sub-models on $\log \mu_{i,j,k}$ $\log \phi_{i,j,k}$ is the cornerstone of our approach to differential expression analysis.

Library Generation and Sequencing in the 4N random adaptor protocol

In this section, we describe the protocol for the generation of the libraries using the 4N randomized adaptor protocol

Input RNA

RNA inputs were pools of 962 (miRXplore; Miltenyi Biotec) or 286 synthetic RNA oligos (IDT). Each library was prepared from 0.1 femtomole to 100 femtomoles of RNA. Pools were either equimolar or ratiometric mixes as indicated. The composition of each pool can be found in Supplementary Table 1. Libraries and sequences were undertaken in three batches, with the identity of the samples (miRXplore v.s. 286 and ratiometric v.s. equimolar) randomly assigned to batches. Our randomization strategy, guards against the introduction of bias due to random variation or drift in laboratory practice or equipment performance.

Library preparation and sequencing

3' ligation:

3' ligation was carried out in thin-walled PCR tubes containing dried pellets of polyethelyne glycol (PEG) to give the indicated final PEG concentration (15%-25%) when resuspended in 10uL. For each library, the indicated amount of input RNA (0.1 femtomoles – 100 femtomoles) in 6uL of water was added to the PCR tube along with 1uL of 3' adapter such that the final amount of adapter in the ligation reaction was as indicated (10 picomoles – 50 picomoles) (3' adapter sequence: /5rApp/(N:25252525)(N)(N)(N)TGGAATTCTCGGGTGCCAAGG/3ddC/). Tubes containing RNA, adapter and PEG were denatured for 2 minutes at 70°C then chilled on ice. To each tube, 1uL of 10X T4 RNA ligase reaction buffer, 1uL of RNase inhibitor, and 1uL of T4 RNA ligase 2 truncated KQ were added and tubes were incubated for 2 hours at 25°C.

Adapter dimer removal:

Excess 3' adapter was removed enzymatically by addition of 1ug of E. coli single strand binding protein and incubation for 10 minutes at 25°C, followed by treatment with 1uL of 5' deadenylase for 1 hour at 30°C, then treatment with 1uL of RecJ for 1 hour at 37°C.

5' ligation:

In a separate tube, the indicated 5' adapter (5' adapter v1 sequence: rGrUrUrCrArGrArGrUrUrCrUrArCrArGrUrCrCrGrArCrGrArUrCr(N:25252525)r(N)r(N)r(N), 5' adapter v3 sequence: G*T*T*C*A*G*A*G*T*T*C*T*A*C*A*G*T*C*C*G*A*C*rGrArUrCr(N:25252525)r(N)r(N)r(N)) was denatured for 2 minutes at 70°C and chilled on ice. 1uL of 5' adapter to yield the

indicated final amount of 5' adapter (10 picomoles – 135 picomoles), 1uL of 10mM ATP and 1uL of T4 RNA ligase 1 was added to each tube and incubated at either 25°C or 37°C for 1-2 hours as indicated.

Reverse transcription:

6uL of ligated RNA and 1uL of reverse transcription primer (10uM; GCCTTGGCACCCGAGAATTCCA) were added to a new tube and denatured at 70°C for 2 minutes then chilled on ice. To each tube, 2uL of 5X first strand reaction buffer, 1uL of 100mM DTT, 0.5 uL of 12.5mM dNTP mix, 1uL of RNase inhibitor and 1uL of SuperScript III reverse transcriptase were added and the reaction was incubated at 55°C for 1 hour followed by 70°C for 15 minutes. To remove excess 5' adapter, 1uL of RiboShredder RNase blend was added and incubated at 37°C for 15 minutes.

PCR amplification #1:

25uL of 2X PCR master mix, 2uL of Illumina forward PCR primer (20uM; sequence: AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA) and 2uL of Illumina indexed reverse PCR primer (20uM; sequence: CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA, where XXXXXX indicates barcode), and 7.5uL of water was added to the cDNA. Libraries were amplified for 4 cycles. PCR reactions were cleaned using a PCR purification column and eluted in 20uL of elution buffer.

Size selection #1:

5uL of PippinHT loading buffer 30A was added to each purified PCR and libraries were loaded on a 3% agarose gel and size selected on a PippinHT automated electrophoresis instrument. The range of 127-156bp was collected. Eluted DNA was concentrated in a vacuum concentrator until the volume was less than 22.5uL.

PCR amplification #2:

Each library was transferred to a new PCR tube and the volume was adjusted to 22.5uL with water. 25uL of PCR master mix and 2.5uL of universal primer cocktail (20uM each primer; Forward primer sequence: AATGATACGGCGACCACCGAG, Reverse primer sequence: CAAGCAGAAGACGGCATACGA) was added to each tube. Libraries were amplified for a further 8-15 cycles. PCR reactions were cleaned using a PCR purification column and eluted in 20uL of elution buffer.

Size selection #2:

5uL of PippinHT loading buffer 30A was added to each tube and libraries were loaded on a 3% agarose gel and size selected on a PippinHT automated electrophoresis instrument. The range of 127-156bp was collected. Eluted DNA was concentrated in a vacuum concentrator until the volume was approximately 15uL. 1uL of each library was loaded onto a Bioanalyzer DNA 1000 chip to assess size and purity of the finished library.

Sequencing and RNA counts

Libraries were denatured, diluted and loaded onto a NextSeq 500 and sequenced according to manufacturer's instructions. Only those sequences that matched exactly those of the miRNAs in the miRXplore and 286 pools were counted and analyzed in this report. The following one-line bash script was used to generate RNA counts from a directory of fastq files:

LC_ALL=C grep -F -o -f /media/disk2/mirx/mirxplore.txt $FILE4 > pass/$PART4".extracted”

where mirxplore.txt is the list of full length sequences that are supposed to be in the pool.

Appendix

A. Expectation and variance of capture probabilities in RNA-seq experiments

In this section, we derive expressions for the expectation and the variance of ratios of random variables, such as the capture probabilities in (Eq. 3) of the **main text**. Previous treatments of this problem, consider the case of the ratio of two random variables (75–78). In our case, the denominator is the sum of the numerator and many other random variables, i.e. the abundances of all the other RNA species in the reaction. For clarity, we introduce the notation $L_{-i}^N = \sum_{\substack{j=1 \\ j\neq i}}^{n} L_j^N$ i.e. the sum of all products of the PCR reaction at the N^{th} cycle, except for the ith species.

Consider Taylor series expansions of the expected value of the ratios $s_i = \frac{L_i^N}{L_i^N + L_{-i}^N}$. Since the ratio will be small for almost all s_i, we have the result:

$$E[s_i] \approx E\left[\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]} + \frac{\partial}{\partial L_i^N}\left(\frac{L_i^N}{L_i^N + L_{-i}^N}\right)\Big|_{(E[L_i^N],E[L_{-i}^N])} (L_i^N - E[L_i^N]) + \frac{\partial}{\partial L_{-i}^N}\left(\frac{L_i^N}{L_i^N + L_{-i}^N}\right)\Big|_{(E[L_i^N],E[L_{-i}^N])} (L_{-i}^N - E[L_{-i}^N])\right] \Rightarrow$$

$$E[s_i] \approx \frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]} = \frac{E[L_i^N]}{E[L_i^N] + E[L_{-i}^N]} \qquad \text{(Eq.14)}$$

where the last relation follows because of the assumption of the statistical independence of PCR amplification, under the branching process model (44, 65, 79). By definition, the variance of the capture probability, s_i , is equal to $V[s_i] = E\left\{\left[s_i - E[s_i]\right]^2\right\}$. Using the expression for $E[s_i]$ and expanding s_i as a Taylor series to first order, one obtains

$$V[s_i] \approx E\left\{\left[\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]} + \frac{\partial}{\partial L_i^N}\left(\frac{L_i^N}{L_i^N + L_{-i}^N}\right)\Big|_{(E[L_i^N], E[L_{-i}^N])} (L_i^N - E[L_i^N])\right.\right.$$

$$\left.\left. + \frac{\partial}{\partial L_{-i}^N}\left(\frac{L_i^N}{L_i^N + L_{-i}^N}\right)\Big|_{(E[L_i^N], E[L_{-i}^N])} (L_{-i}^N - E[L_{-i}^N]) - \frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]}\right]^2\right\}$$

$$= E\left\{\left(\frac{E[L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2}\right)^2 (L_i^N - E[L_i^N])^2 + \left(-\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]^2}\right)^2 (L_{-i}^N - E[L_{-i}^N])^2\right.$$

$$\left. - 2\frac{E[L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2} \times \frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]^2} (L_i^N - E[L_i^N])(L_{-i}^N - E[L_{-i}^N])\right\}$$

$$= \left(\frac{E[L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2}\right)^2 V[L_i^N] + \left(\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]^2}\right)^2 V[L_{-i}^N]$$

$$- 2\frac{E[L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2} \times \frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]^2} Cov(L_i^N, L_{-i}^N)$$

$$= \left(\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]}\right)^2 \left\{\frac{V[L_i^N]}{E[L_i^N]^2} \times \left(\frac{E[L_{-i}^N]}{E[L_i^N + L_{-i}^N]}\right)^2 + \frac{V[L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2}\right\}$$

$$= \left(\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]}\right)^2 \left\{\frac{V[L_i^N]}{E[L_i^N]^2} \times \left(\frac{E[L_{-i}^N]}{E[L_i^N + L_{-i}^N]}\right)^2 - \frac{V[L_i^N]}{E[L_i^N + L_{-i}^N]^2} + \frac{V[L_i^N + L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2}\right\}$$

$$= \left(\frac{E[L_i^N]}{E[L_i^N + L_{-i}^N]}\right)^2 \left\{\frac{V[L_i^N]}{E[L_i^N]^2} \times \left(\frac{E[L_{-i}^N]^2 - E[L_i^N]^2}{E[L_i^N + L_{-i}^N]^2}\right) + \frac{V[L_i^N + L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2}\right\} \Rightarrow$$

$$V[s_i] \approx E[s_i]^2 \left\{\frac{V[L_i^N]}{E[L_i^N]^2} \times \frac{E[L_{-i}^N] - E[L_i^N]}{E[L_i^N] + E[L_{-i}^N]} + \frac{V[L_i^N + L_{-i}^N]}{E[L_i^N + L_{-i}^N]^2}\right\} \qquad \text{(Eq.15)}$$

In this derivation we used the statistical independence between the amplification reactions for the different RNA species to set the covariance term to zero. Furthermore, by defining$L^N = L_i^N + L_{-i}^N$, the total size of the PCR amplified library, we can further reduce the expression to:

$$V[s_i] \approx E[s_i]^2 \left\{ \frac{V[L_i^N]}{E[L_i^N]^2} \times \frac{E[L^N] - 2 \times E[L_i^N]}{E[L^N]} + \frac{V[L^N]}{E[L^N]^2} \right\} = E[s_i]^2 \phi_i \quad \text{(Eq.16)}$$

In all expressions, the notation $E[X]^2$ stands for the square of the expectation of X $E[X]$. The expectations and the variances appearing in the last two equations should be understood as the quantities characterizing the stochastic process of PCR amplification. For a given sample composition and efficiency of the PCR reaction, the term inside the brackets is a constant, *dispersion factor* (ϕ_i). Stated in other terms, the variance is proportional to the square of the mean and the proportionality constant is the dispersion factor.

Factoring common terms, and noting that the dispersion factors are in general functions of the expectation of the capture probabilities, allows us to derive two alternative, yet equivalent expressions for the variance as a function of the mean:

$$V[Y_i|K_k] = K_k E[s_i] + K_k^2 V[s_i] = E[Y_i|K_k](1 + \varphi_i E[Y_i|K_k]) \quad \text{(Eq.17)}$$

$$V[Y_i|K_k] = E[s_i]K_k + K_k^2 f(E[s_i]) \quad \text{(Eq.18)}$$

The functional forms of the mean/variance relations in the last two equations are the ones assumed in the edgeR (15, 18), and the deSEQx (14, 20) algorithms respectively.

B. Marginal distributions in chains of conditionally independent Binomial random variables

In this section, we derive a result that is used to determine the marginal distribution of the last node in a chain of conditionally independent Binomial random variables, i.e.:

$$\begin{gathered} B_0|M,p_0 \sim Binomial(M,p_0) \\ B_1|B_0,p_1 \sim Binomial(B_0,p_1) \\ \vdots \\ B_n|B_{n-1},p_n \sim Binomial(B_{n-1},p_n) \end{gathered} \tag{Eq.19}$$

In particular, we claim that:

$$B_n|M,p_0,\dots,p_n \sim Binomial(M,\prod_{i=0}^{n} p_i) \tag{Eq.20}$$

The statement holds true for a chain of just two variables B_0, B_1 e.g. see section on compound binomial distributions p374 in (49). An alternate proof of this assertion may be derived as follows: The sought after *marginal* distribution is obtained by marginalizing the joint distribution $P(B_0,B_1|M,p_0,p_1)$ over B_0 for all pairs of B_0, B_1 for which the joint distribution is positive, i.e. the set of values $0 \le B_1 \le B_0 \le M$. Hence:

$$P(B_1|M,p_0,p_1) = \sum_{B_0=B_1}^{B_0=M} P(B_0,B_1|M,p_0,p_1) = \sum_{B_0=B_1}^{B_0=M} P(B_1|B_0,p_1)P(B_0|M,p_0) \tag{Eq.21}$$

$$= \sum_{B_0=B_1}^{B_0=M} \binom{M}{B_0} p_0^{B_0}(1-p_0)^{M-B_0} \binom{B_0}{B_1} p_1^{B_1}(1-p_1)^{B_0-B_1}$$

$$= \frac{M!}{B_1!} \sum_{B_0=B_1}^{B_0=M} \frac{p_0^{B_0}(1-p_0)^{M-B_0} p_1^{B_1}(1-p_1)^{B_0-B_1}}{(M-B_0)!\,(B_0-B_1)!}$$

To proceed we make the substitution $B_0 = b + B_1$ and multiply both numerator and denominator by the factor $(M - B_1)!$ to obtain:

$$\frac{M!}{(M-B_1)!\,B_1!}\sum_{b=0}^{b=M-B_1}\frac{(M-B_1)!\;p_0^{b+B_1}(1-p_0)^{M-B_1-b}p_1^{B_1}(1-p_1)^b}{(M-B_1-b)!\,b!} \qquad \text{(Eq.22)}$$

$$=\frac{M!\,(p_0p_1)^{B_1}}{(M-B_1)!\,B_1!}\sum_{b=0}^{b=M-B_1}\frac{(M-B_1)!\;p_0^{b}(1-p_0)^{M-B_1-b}(1-p_1)^b}{(M-B_1-b)!\,b!}$$

$$=\frac{M!\,(p_0p_1)^{B_1}}{(M-B_1)!\,B_1!}\sum_{b=0}^{b=M-B_1}\binom{M-B_1}{b}(1-p_0)^{M-B_1-b}(1-p_1)^b\,p_0^b$$

$$=\frac{M!\,(p_0p_1)^{B_1}}{(M-B_1)!\,B_1!}\big((1-p_0)-(1-p_1)p_0\big)^{M-B_1}$$

$$=\frac{M!\,(p_0p_1)^{B_1}(1-p_0p_1)^{M-B_1}}{(M-B_1)!\,B_1!}=\binom{M}{B_1}(p_0p_1)^{B_1}(1-p_0p_1)^{M-B_1}$$

$$\Rightarrow B_1|M,p_0p_1\sim Binomial(M,p_0p_1)$$

For a chain of more than two binomials, the respective result follows by induction.

C. Marginal distribution of the hierarchical multinomial model

We derive the marginal distribution of the hierarchical multinomial model given in (Eq. 7) of the **main text** by marginalizing the probability $Y_j\big|B_j,r$ over the multinomial distribution of counts. In the second step, we marginalize the resulting distribution over the observed library depth from the two levels of the hierarchy to arrive at a model for $Y_1, Y_2, \dots, Y_n, Y_{n+1}$that conditions on the theoretical library depth K_0, the library depth variation probability, t, the capture probabilities $s_1, s_2, \dots, s_n$ and the signal generation probability r. This is an augmented probability space, in which the number of "missing" counts appears as another random variable (Y_{n+1}).

In the first step, marginalization of this multinomial is over all $B_1, B_2, \dots, B_n$ that satisfy the constraint $\sum B_i = K_k$ leads to the following expression after dropping the indexing variable in all product terms to reduce clutter:

$$\sum_{\substack{1 \le i \le n \\ \sum B_i = K_k}} (B_1, B_2, \dots, B_n | K_0, t, s_1, s_2, \dots, s_n) \prod (Y_i | B_i, r) =$$

$$\sum_{\substack{1 \le i \le n \\ \sum B_i = K_k}} \frac{K_0!}{\prod (Y_i!)\,(B_i - Y_i)!} \prod \left(r^{Y_i} s_i^{B_i} (1-r)^{B_i - Y_i} \right)$$

Since $B_i \geq Y_i$, we introduce the transformation $b_i = B_i - Y_i$ to re-write the sum as:

$$\frac{K_k!}{\prod (Y_i!)} \sum_{\substack{1 \le i \le n \\ b_i = 0}}^{\sum b_i = K_k - \sum Y_i} \prod \frac{\left(r^{Y_i} s_i^{b_i + Y_i} (1-r)^{b_i} \right)}{b_i!} = \frac{K_k!\,(r s_i)^{Y_i}}{\prod (Y_i!)} \sum_{\substack{1 \le i \le n \\ b_i = 0}}^{\sum b_i = K_k - \sum Y_i} \prod \frac{\left(s_i^{b_i} (1-r)^{b_i} \right)}{b_i!} =$$

$$\frac{K_k!\,(r s_i)^{Y_i}}{\prod (Y_i!)\,(K_k - \sum Y_i)!} \left\{ \sum_{\substack{1 \le i \le n \\ b_i = 0}}^{\sum b_i = K_k - \sum Y_i} \frac{(K_k - \sum Y_i)!}{b_i!} \prod s_i^{b_i} (1-r)^{b_i} \right\}$$

The expression inside the bracket is recognizable as the multinomial theorem expansion of the polynomial $(1 - \sum r s_i)^{K_k - \sum Y_i} = (1-r)^{K_k - \sum Y_i}$. Introducing this term into the last equation leads to the following expression:

$$Y_1, Y_2, \dots, Y_n, Y_{n+1} | K_k, r, s_1, s_2, \dots, s_n = \frac{K_k!\,(r s_i)^{Y_i} (1-r)^{K_k - \sum Y_i}}{\prod (Y_i!)\,(K_k - \sum Y_i)!}$$

This is a multinomial distribution over the set of *observed counts* augmented with the number of *missing counts* $Y_{n+1} = K_k - \sum Y_i$. Therefore:

$$Y_1, Y_2, \dots, Y_n, Y_{n+1} | K_k, r, s_1, s_2, \dots, s_n \sim Multinomial(K_k; r s_1, r s_2, \dots, r s_n, 1-r)$$

The marginal distribution of the latter expression over the binomial law describing the random variation in library depth is obtained as a single sum term, as there is only one value of K_k that is compatible with the multinomial model, i.e. $K_k = \sum_{i=1}^{i=n+1} Y_i$:

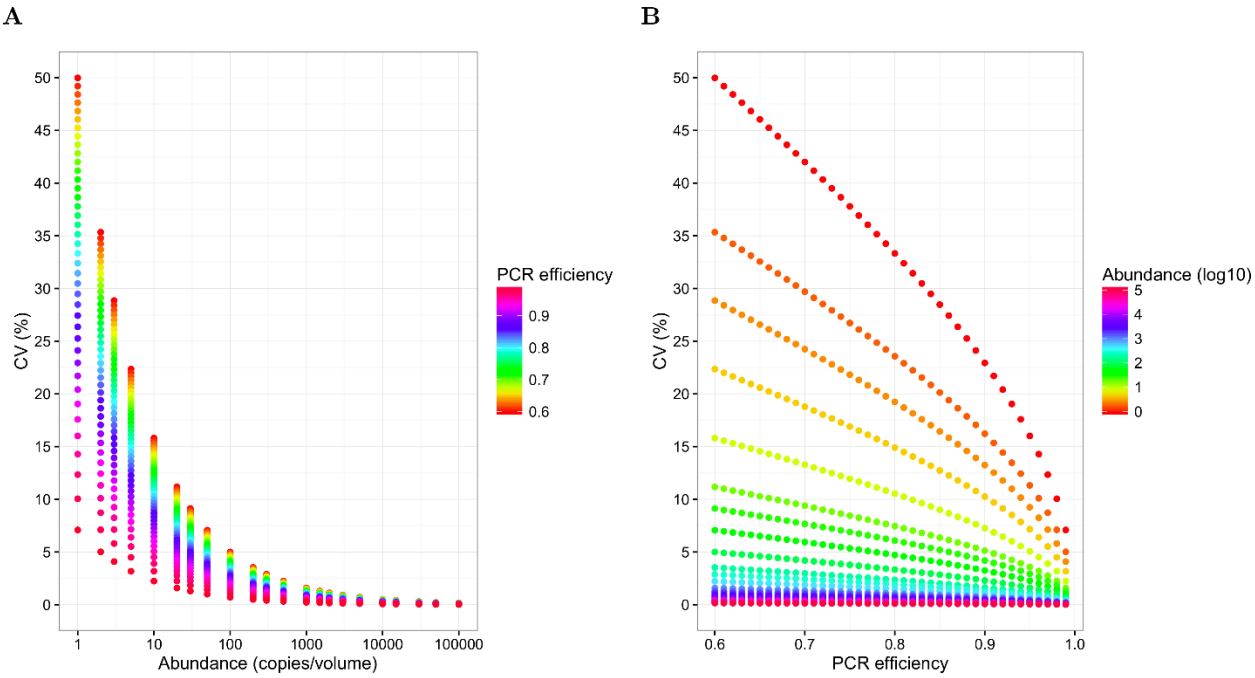
$$\frac{K_k!\,(rs_i)^{Y_i}(1-r)^{K_k-\sum Y_i}}{\prod(Y_i!)\left(K_k-\sum_{i=1}^{i=n}Y_i\right)!}t^{K_k}(1-t)^{K_0-K_k}\frac{K_0!}{K_k!\,(K_0-K_k)!}=$$

$$\frac{K_0!\,(rs_i)^{Y_i}(1-r)^{K_k-\sum_{i=1}^{i=n}Y_i}}{\prod(Y_i!)\left(K_k-\sum_{i=1}^{i=n}Y_i\right)!\,(K_0-K_k)!}(1-t)^{K_0-K_k}t^{K_k-\sum_{i=1}^{i=n}Y_i}\prod_{i=1}^{i=n}t^{Y_i}=$$

$$\frac{K_0!\,(trs_i)^{Y_i}(t(1-r))^{K_k-\sum_{i=1}^{i=n}Y_i}}{\prod(Y_i!)\left(K_k-\sum_{i=1}^{i=n}Y_i\right)!\,(K_0-K_k)!}(1-t)^{K_0-K_k}$$

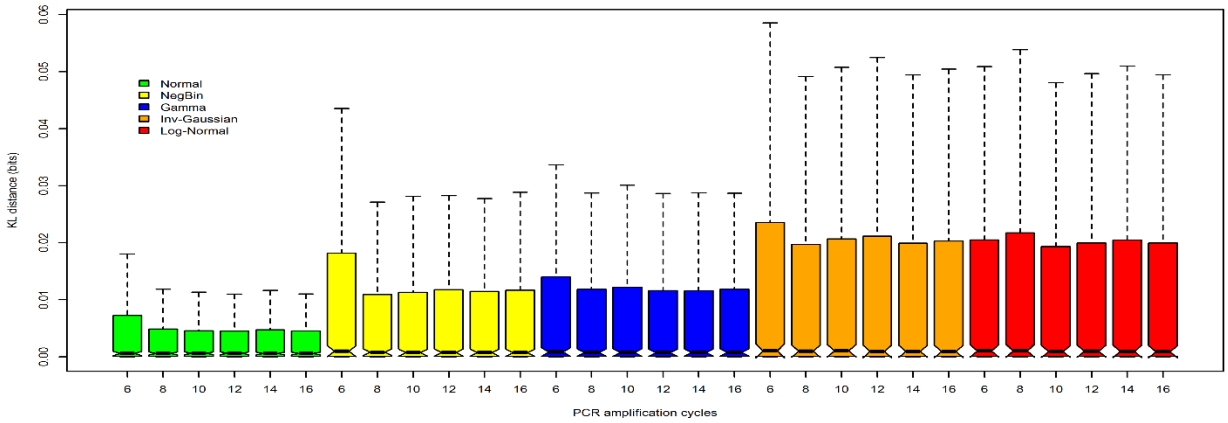
This is recognizable as the expression for a multinomial probability mass function of size K_0 with $n+2$ categories and corresponding probabilities: $trs_1, trs_2, \ldots, trs_n, t(1-r), 1-t$. The last two categories correspond to the "missing counts" from the two distinct processes of pre-analytical and post-analytical random library variation. By properties of the multinomial distribution, they can be combined into a single category with probability and counts equal to the sum of the respective quantities. By doing so, we prove the claim asserted in (Eq. 8) of the **main text**.

Supplementary Figures

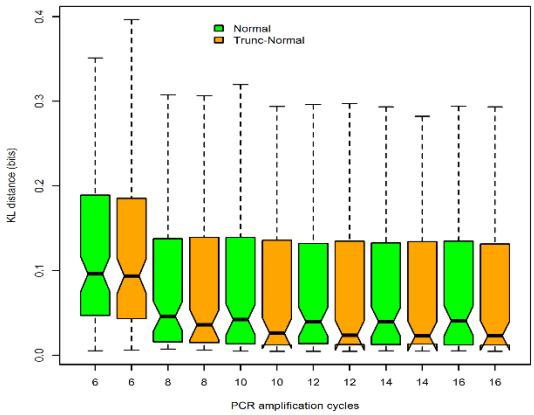


Supplementary Figure 1 Coefficient of variation (CV) of PCR counts for different amounts of starting material (A) and PCR efficiency (B). To generate these figures the PCR stochastic branching process model was simulated for 16 cycles of variable abundance and PCR efficiency. At each cycle 20,000 samples were drawn from the distribution defined by (Eq.6) and the mean and standard deviation were calculated for each combination of reaction efficiency and abundance. The figure summarizes the CV for the final cycle.

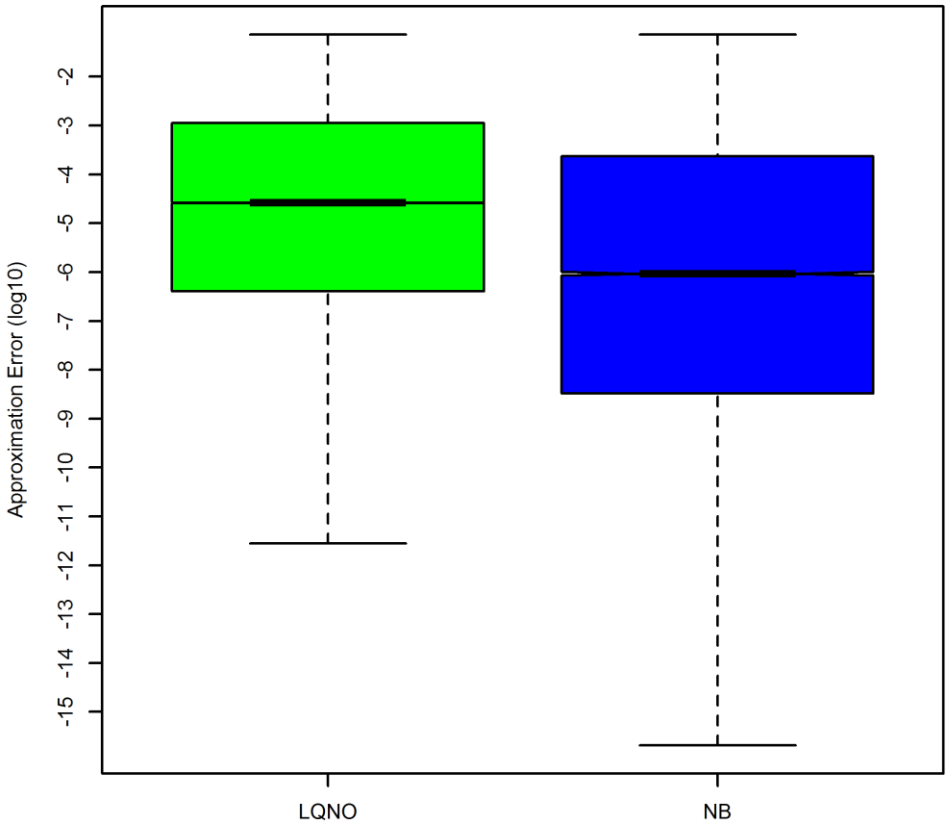
A



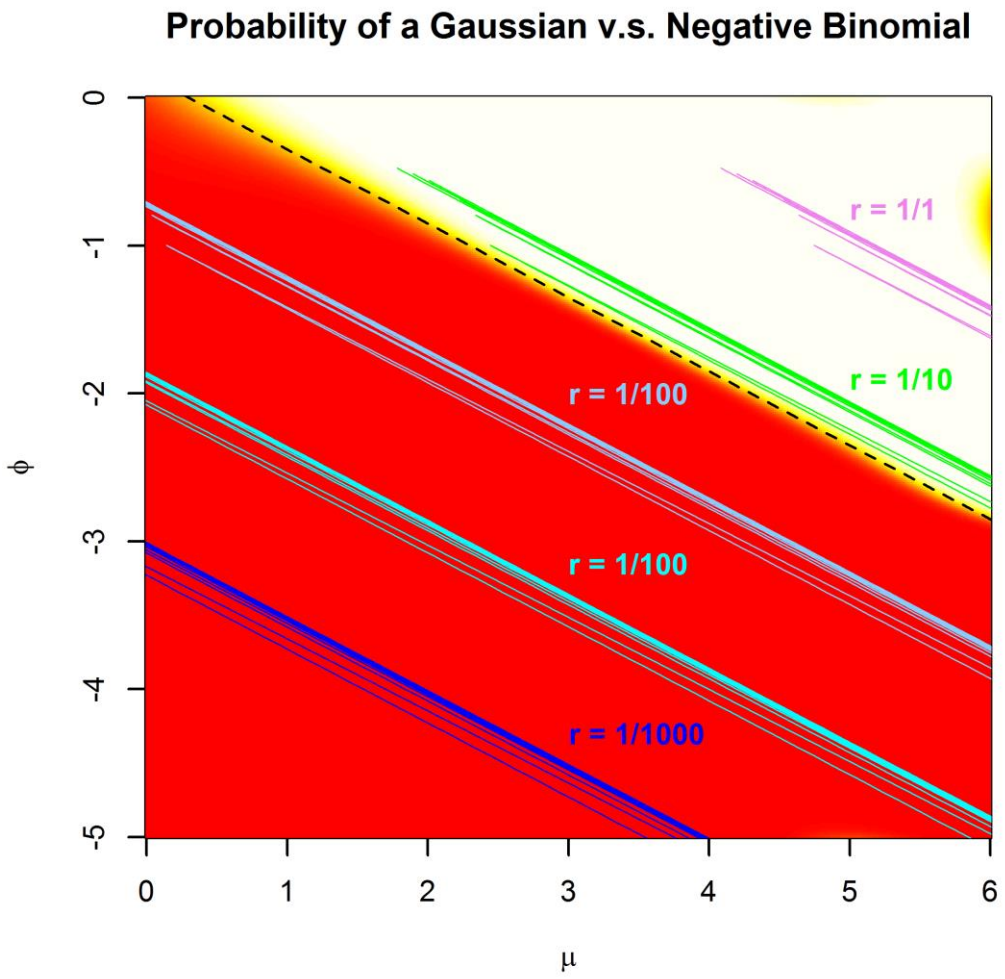
B



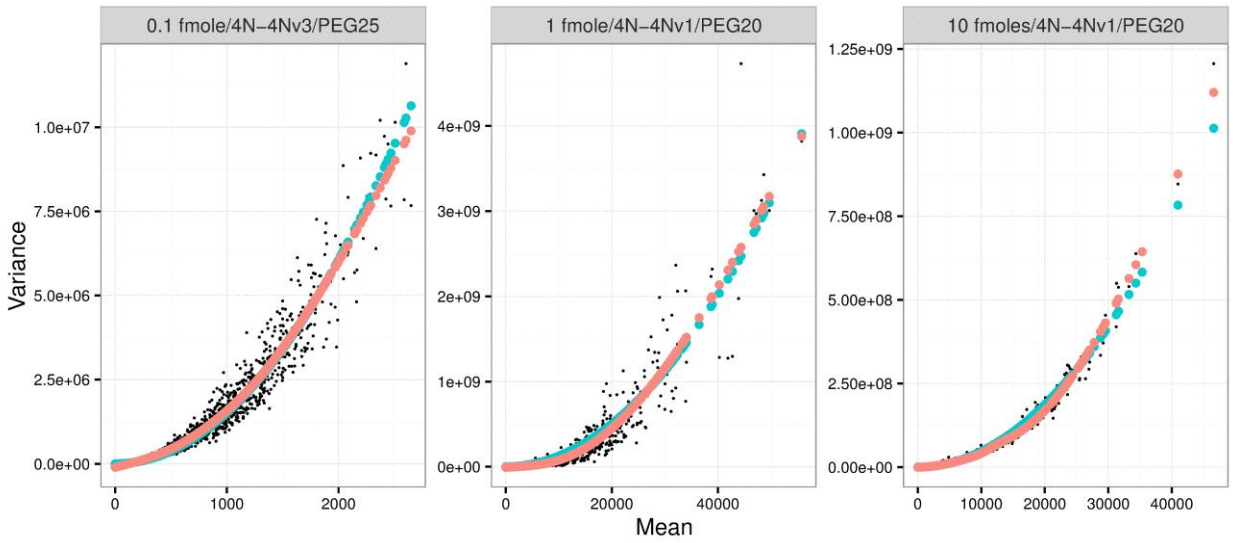
Supplementary Figure 2 KL distance is given in bits for different amplification cycles (ordinate in x axis). In each bar-graph the notch marks the median, and non-overlapping notches suggest that the difference of medians is statistically significant. Kullback Leibler (KL) distance between samples generated by the Galton-Watson PCR branching process and candidate distributions with the same mean and variance (in bits). Normal: Gaussian Distribution, NegBin: Negative Binomial I, Inv-Gaussian: Inverse Gaussian. See Table 1 in the main text for the definitions of these densities. Differences between the other distributions and the normal one were highly statistically significant. ($p<0.001$ mixed effects regression analysis) (A) KL distance between samples generated from the Galton Watson and either the Normal (Gaussian) distribution or its left-truncated version for less than 1000 copies of initial abundance (B)



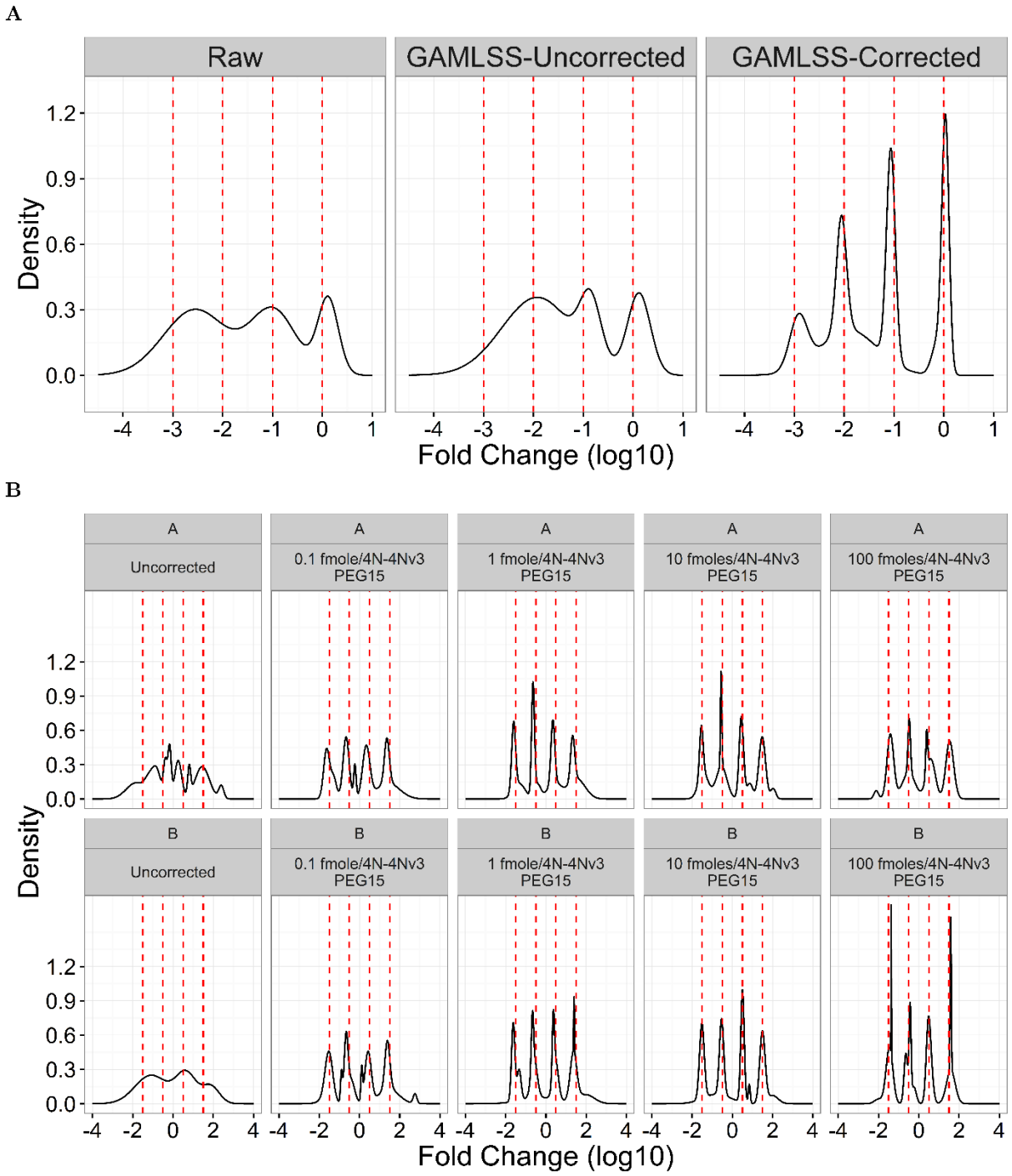
Supplementary Figure 3 Distribution of the absolute difference between the Negative Binomial (NB) or the Linear Quadratic Normal family (LQNO) and the truncated normal mixed Poisson distribution. To generate this figure, we simulated 10000 unique combinations of μ, ϕ uniformly in the range of $[1,10^6]$ x $[10^{-10},1]$. Subsequently we simulated 50 values for the x in the interval $[\mu - 10\sigma_{LQ}, \mu + 10\sigma_{LQ}]$ for each unique combination of μ, ϕ (a total of 500,000 evaluation points) and evaluated (Eq.11) using an arbitrary precision numerical library. The value of the p.m.f. was then contrasted against the value of the probability density function of the NBI and the LQNO at the same values of μ, ϕ, x.



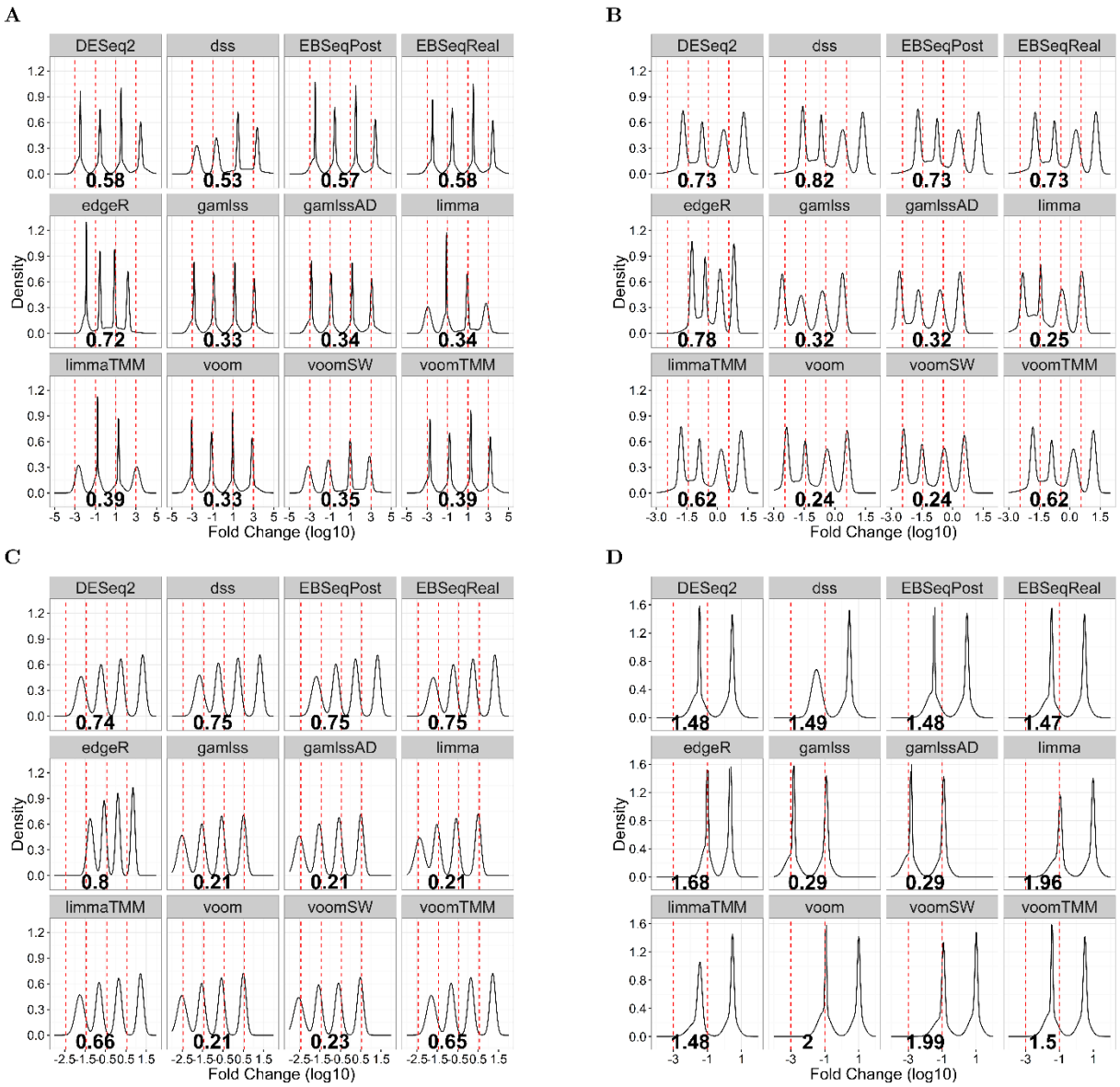
Supplementary Figure 4. Response surface yielding the probability that a mixed Poisson p.m.f is numerically closer to the Negative Binomial (red) v.s. the LQNO (yellow). There is a sharp demarcation of the regions in which the NBI performs better than the LQNO, given by the linear equation $\phi = 0.15 - 0.5\,\mu$, when both quantities are expressed in a logarithmic (base 10) scale. This is shown by the dashed line in the figure. The probability jumps from 1 to zero above this line, so that the LQNO provides a better approximation to the truncated normal mixed Poisson distribution for higher values of μ and smaller values of ϕ. We also plotted the range of values in the $\phi - \mu$ plane spanned by different values of the signal generation probabilities (r), PCR efficiencies (from 0.9 to 0.99) and abundances (from 1 to 10^{12}) for 16 amplification reaction cycles. Depending on the signal generation probability for a particular experiment, either distribution may yield a superior approximation to the truncated mixed Poisson distribution.



Supplementary Figure 5 Modeling of mean – variance relationship in 3 different RNA-seq experimental combinations (total of 32 RNA-seq datasets) involving different amounts of starting materials and 5p variations in the random 4N protocol. In each of these experiments, synthetic mixes of miRNAs (miRXplore) were used as input for library construction and sequencing. Within each experimental condition, we calculated the mean and the variance (over all the replicates) for each miRNA sequenced. Subsequently, we fit a linear – quadratic curve to these data-points (blue), estimating only the coefficient of the quadratic term from the data. These parametric curves are superimposed to flexible smoothing splines that were fit to the same data with smoothing regression models (red). There is remarkable agreement between the relations estimated by the flexible, data-driven curve and the LQ model. Data sources to derive this figure are detailed in Supplementary Table 1.

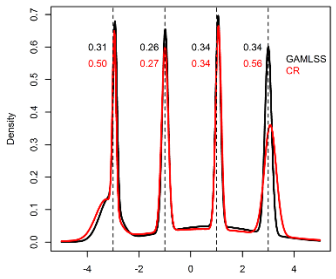


Supplementary Figure 6 Peak detection in expression profiles before and after bias correction Data of Haffner et all (A) and ratiometric series A and B in the validation 4N dataset (B). Relative to the uncorrected or raw data, application of bias correction results in expression profiles with peaks that coincide with the true expression patterns (dashed red lines). Sequence data (either the raw data or model estimates) were analyzed with finite Gaussian clustering methods to generate the profiles shown in the graph (black curves).

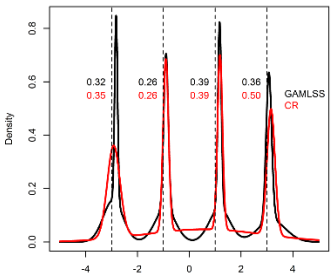


Supplementary Figure 7 Analysis of Differential Expression under scenarios of clustered symmetric DE without global changes in expression ratiometric A v.s. B (A), clustered asymmetric DE which shifts the global expression in one direction: equimolar v.s. ratiometric series A (B), equimolar v.s. ratiometric series B (C), ratiometric series A v.s. B in which the measurements of the overexpressed RNAs (subpools A and B) were omitted from the analyses (D). The dashed red lines are the true DE values, the numbers in bold the RMSE errors and the histograms are the model based clustering of the DE measures estimated by each method: *DESeq2*, *edgeR* (after Trimmed Mean Normalization, TMM), *EBSeq* (using either the Posterior Fold Changes, *EBSeqPost*, or the fold changes adjusted by normalization factors, *EBSeqReal*) , *dss*, limma (with or without TMM, limmaTMM), *voom* (unnormalized, or with sample weights, *limmaSW*, and after TMM, *limmaTMM*), *gamlss* (the reference implementation of the methods proposed in the text), *gamlssAD* (the fast implementation using Algorithmic Differentiation methods). This figure is based on the repeat sequencing of the validation datasets.

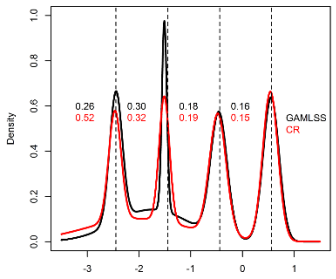
A


0.31
0.50
0.26
0.27
0.34
0.34
0.34
0.56
GAMLSS
CR
Density

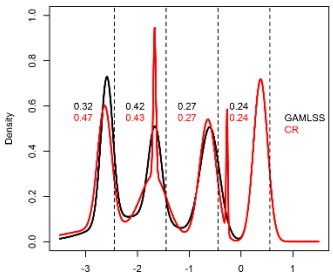
B


0.32
0.35
0.26
0.26
0.39
0.39
0.36
0.50
GAMLSS
CR
Density

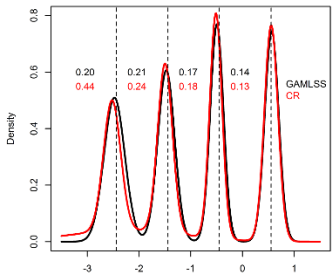
C


0.26
0.52
0.30
0.32
0.18
0.19
0.16
0.15
GAMLSS
CR
Density

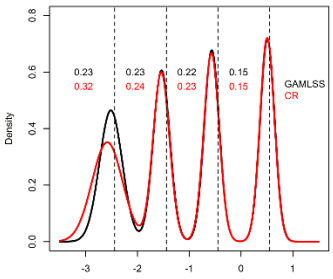
D


0.32
0.47
0.42
0.43
0.27
0.27
0.24
0.24
GAMLSS
CR
Density

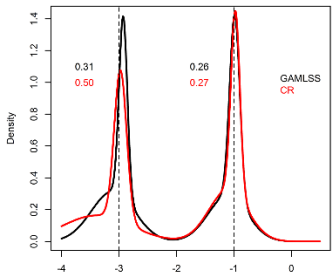
E


0.20
0.44
0.21
0.24
0.17
0.18
0.14
0.13
GAMLSS
CR
Density

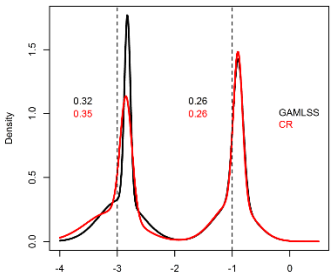
F


0.23
0.32
0.23
0.24
0.22
0.23
0.15
0.15
GAMLSS
CR
Density

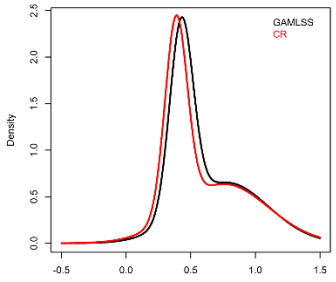
G


0.31
0.50
0.26
0.27
GAMLSS
CR
Density

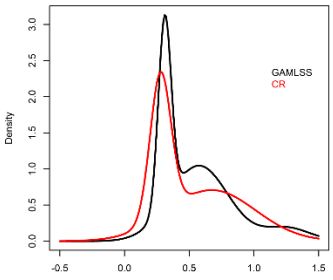
H


0.32
0.35
0.26
0.26
GAMLSS
CR
Density

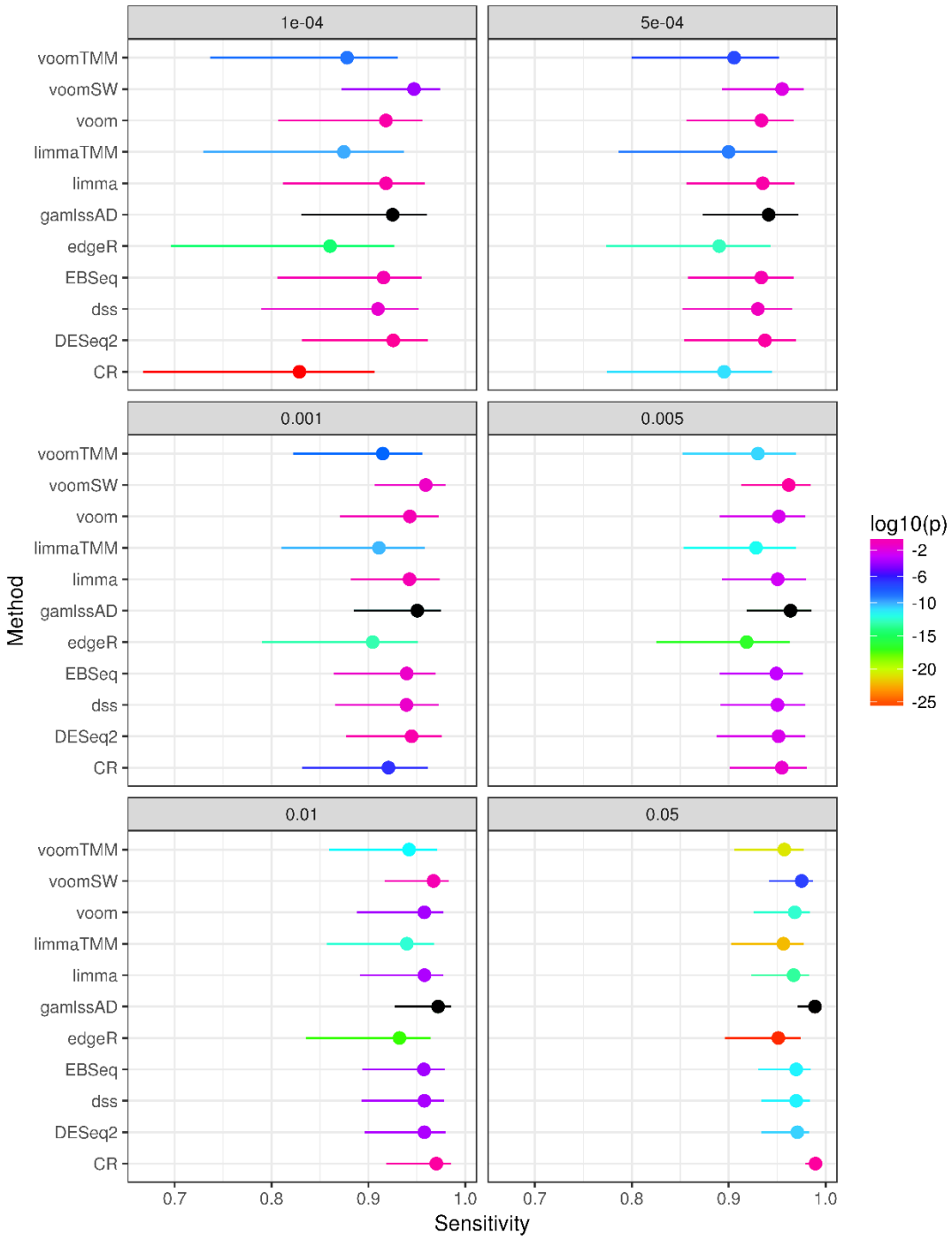
I


GAMLSS
CR
Density

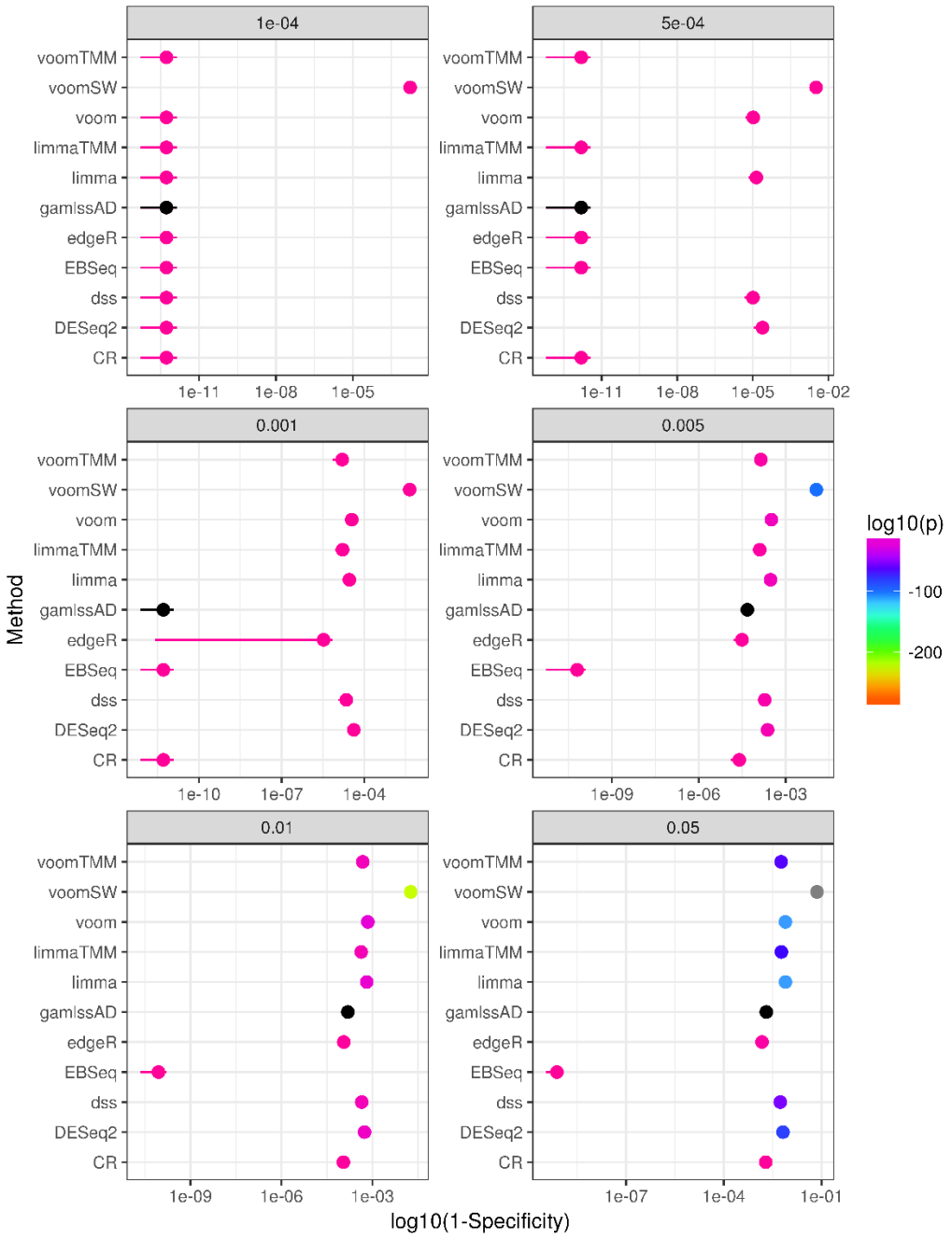
J


GAMLSS
CR
Density

Supplementary Figure 8 Analysis of Differential Expression with the gamlss and the Cubic Root (CR) transformation method under scenarios of clustered symmetric DE without global changes in expression ratiometric A v.s. B (A, B), clustered asymmetric DE which shifts the global expression in one direction: equimolar v.s. ratiometric series A (C,D), equimolar v.s. ratiometric series B (E,F), ratiometric series A v.s. B in which the measurements of the overexpressed RNAs (subpools A and B) were omitted from the analyses (G,H) and finally under a simulated global differential expression scenario (I,J). The images on the left column (A,C, E, G, I) were based on the initial sequencing of the validation libraries, while the ones on the right on the resequenced ones.



Supplementary Figure 9 Regression analysis of sensitivity as a function of the p-value cutoff from (0.05 to 0.0001). Confidence intervals and estimates were based on bootstrapping (200 samples) a mixed logistic regression analysis, with the proportion of significance tests as the dependent variable and method as a covariate. This analysis used the DE datasets considered in the text. The referent for this analysis is *gamlssAD* (shown in black). The color scale gives the p-value (in log10 scale) that each of the other methods had sensitivity different than *gamlssAD*.



Supplementary Figure 10 Regression analysis of specificity (in log10 scale) as a function of the p-value cutoff from (0.05 to 0.0001). Confidence intervals and estimates were based on bootstrapping (200 samples) a mixed logistic regression analysis, with the proportion of significance tests as the dependent variable and method as a covariate. This analysis used the datasets without any differential expression considered in the text. The referent for this analysis is *gamlssAD* (shown in black). The color scale gives the p-value (in log10 scale) that each of the other methods had specificity different than *gamlssAD*. For many of the p-value cutoffs, the confidence intervals fall below the spatial resolution of the figure and thus do not extent past the dot that gives the mixed model estimate

Supplementary Table 1 Characteristics of the miRexplore and 286 samples generated for the validation of the methodology proposed in the paper. The dilution column provides the dilution factor relative to a reference (input) concentration of 100 fmoles for each pool. The total input of the two ratiometric series was 100 fmoles.

Experimental Group	Dilution	N
miRExplore	1:10	10
286	1:1	8
	1:10	8
	1:100	8
	1:1000	8
Ratio Metric Series A (descending)	Mix of • 286 subpool A (1:1) • 286 subpool B (1:10) • 286 subpool C(1:100) • 286 subpool D (1:1000)	8
Ratio Metric Series B (ascending)	Mix of • 286 subpool A (1:1000) • 286 subpool B (1:100) • 286 subpool C(1:10) • 286 subpool D (1:1)	8
Total	7 groups	58 experiments

Supplementary Table 2 *Effects of bias correction in the equimolar 4N validation datasets*

Dataset	Correction Factor Dataset	RMSE	MAE	MAD	Prob(2Fold)	95% Range	99% Range
0.1 fmole	Uncorrected	1.100	0.940	1.338	0.154	2.492	3.031
0.1 fmole	1 fmole	0.113	0.091	0.110	0.990	0.430	0.581
0.1 fmole	10 fmoles	0.233	0.161	0.176	0.860	1.033	1.292
0.1 fmole	100 fmoles	0.367	0.260	0.263	0.717	1.366	1.694
1 fmole	Uncorrected	1.077	0.935	1.361	0.133	2.327	3.026
1 fmole	0.1 fmole	0.117	0.093	0.114	0.990	0.440	0.605
1 fmole	10 fmoles	0.110	0.081	0.095	0.972	0.464	0.603
1 fmole	100 fmoles	0.272	0.209	0.252	0.808	1.027	1.455
10 fmoles	Uncorrected	0.983	0.843	1.181	0.150	2.201	2.836
10 fmoles	0.1 fmole	0.222	0.150	0.144	0.888	0.888	1.237
10 fmoles	1 fmole	0.146	0.102	0.106	0.941	0.615	0.746
10 fmoles	100 fmoles	0.150	0.115	0.136	0.941	0.585	0.811
100 fmoles	Uncorrected	0.750	0.605	0.771	0.325	2.133	2.728
100 fmoles	0.1 fmole	0.346	0.257	0.292	0.710	1.392	1.702
100 fmoles	1 fmole	0.275	0.206	0.237	0.811	1.031	1.456
100 fmoles	10 fmoles	0.149	0.116	0.151	0.948	0.586	0.809

Each of the four equimolar series of varying RNA input (ranging from 0.1 to 100 fmoles) from the miRXplore pool in the column *Dataset* was prepared with the 4N protocol as detailed in the Supplementary Methods. Ligase bias metrics were calculated for each uncorrected dataset and for three corrected analyses. The latter used the empirical correction factors from the other three experiments. Bias reduction (assessed by any of the metrics) was highest when the dataset used for the calculation factors, differed up to an order of magnitude for the dataset that was corrected. In particular, RMSE was reduced from 77%-90% for these scenarios. Even when the RNA input in the correction dataset differed from the dataset to be corrected by 3 orders of magnitude, the % reduction in the RMSE was between 54% (correction of the 100 fmoles dataset by the 0.1 fmole) and 67% (correction of the 0.1 fmole dataset by the correction factors estimated from the100 fmoles one). P-values for the Flinger-Killeen, Ansari and Kolmogorov Smirnov tests for the comparison of variability reduction were all $<10^{-9}$. *RMSE*: Root Mean Square Error, *MAE*: Mean Absolute Error, *MAD*: Median Absolute Deviation, *Prob(2Fold)*: Probability of a short RNA to be expressed within two folds of the average (referent) value for its group, *95% and 99% Range:* Range of values for 95% and 99% of short RNAs assessed in each dataset. The last three metrics were calculated from the empirical Cumulative Density Function of GAMLSS estimates. For the calculation of RMSE, MAD, MAE see the text.

Supplementary Table 3 *Effects of bias correction in the case of empirical factors from samples of heterogeneous composition (equimolar series)*

Dataset	Correction Factor Dataset	miRNA subset	RMSE		MAE		MAD		Prob(2Fold)		95% Range		99% Range	
			Corr.	*Uncor.*	*Corr.*	*Uncor.*	*Corr.*	*Uncor.*	*Corr.*	*Uncor.*	*Corr.*	*Uncor.*	*Corr.*	*Uncor.*
miRXplore	0.1 fmole	Common	0.360	0.621	0.262	0.500	0.302	0.645	0.706	0.315	1.464	2.259	2.028	3.075
miRXplore	0.1 fmole	Unique	0.667	0.670	0.514	0.515	0.633	0.642	0.365	0.367	2.447	2.446	4.156	4.173
miRXplore	1 fmole	Common	0.331	0.621	0.237	0.500	0.253	0.645	0.736	0.315	1.347	2.259	1.962	3.075
miRXplore	1 fmole	Unique	0.666	0.670	0.514	0.515	0.634	0.642	0.365	0.367	2.446	2.446	4.154	4.173
miRXplore	10 fmoles	Common	0.319	0.621	0.234	0.500	0.255	0.645	0.741	0.315	1.192	2.259	1.662	3.075
miRXplore	10 fmoles	Unique	0.666	0.670	0.514	0.515	0.634	0.642	0.367	0.367	2.446	2.446	4.153	4.173
miRXplore	100 fmoles	Common	0.361	0.621	0.272	0.500	0.322	0.645	0.655	0.315	1.302	2.259	1.847	3.075
miRXplore	100 fmoles	Unique	0.666	0.670	0.514	0.515	0.634	0.642	0.367	0.367	2.446	2.446	4.156	4.173
0.1 fmole	miRXplore	Common	0.360	1.164	0.266	1.024	0.287	1.541	0.670	0.117	1.465	2.475	2.023	2.712
0.1 fmole	miRXplore	Unique	0.677	0.943	0.548	0.754	0.692	0.926	0.371	0.236	2.576	2.564	3.077	3.097
1 fmole	miRXplore	Common	0.326	1.137	0.244	1.011	0.276	1.519	0.701	0.096	1.345	2.320	1.940	2.697
1 fmole	miRXplore	Unique	0.626	0.931	0.502	0.766	0.623	0.981	0.371	0.213	2.312	2.317	2.985	3.002
10 fmoles	miRXplore	Common	0.312	1.046	0.242	0.919	0.305	1.283	0.731	0.117	1.161	2.197	1.685	2.643
10 fmoles	miRXplore	Unique	0.558	0.825	0.447	0.675	0.502	0.797	0.449	0.225	2.056	2.053	2.776	2.777
100 fmoles	miRXplore	Common	0.351	0.803	0.276	0.672	0.347	0.910	0.655	0.254	1.306	2.138	1.847	2.469
100 fmoles	miRXplore	Unique	0.513	0.618	0.404	0.457	0.494	0.491	0.449	0.483	1.982	1.983	2.591	2.596

Effects of bias correction when empirical factors are estimated from a sample with a different composition than the target one. The column "Corr." gives the metric for the corrected estimate for each series (column "Dataset") using the correction factor from the series listed under the column "Correction Factor Dataset". Column "Uncor." tabulates the uncorrected estimate for each dataset. The series miRXplore corresponds to the experiments with the 962 pool in the validation dataset. Series 0.1-100 fmoles are the series with the 286 from the validation 4N experiments with the stated RNA input. Bias metrics are calculated separately for the miRNAs that are shared between the target ("common") and the correction factor datasets, and those that only appear in the target dataset ("unique"). It is not possible to apply bias correction to the unique dataset. In the analysis of these equimolar samples, RMSE was reduced by 47.2% ± 12.9%, the MAE by 51.3% ± 13.5%, the MAD by 56.2% ± 13.3% for the miRNAs that were common between the target and

correction factor datasets. The percentage of miRNAs with expression level within two fold of the group mean increased from 23.0% ± 9.5% (uncorrected) to 69.9% ± 3.3%. Simultaneously, the 95% and 99% range were decreased by 36.6%± 8.8% and 33.8% ± 8.4% respectively. There was no change in the bias metrics for miRNAs which were not corrected. P-values for the Flinger-Killeen, Ansari and Kolmogorov Smirnov tests for the comparison of variability reduction were all <10-4 for the common subset. *RMSE*: Root Mean Square Error, *MAE*: Mean Absolute Error, *MAD*: Median Absolute Deviation, *Prob(2Fold)*: Probability of a short RNA to be expressed within two folds of the average (referent) value for its group, *95% and 99% Range*: Range of values for 95% and 99% of short RNAs assessed in each dataset. The last three metrics were calculated from the empirical Cumulative Density Function of GAMLSS estimates. For the calculation of RMSE, MAD, MAE see the text.

Supplementary Table 4 *Execution times required for the learning and application of bias correction factors*

implementation	Operating System	R version	Distribution	Processor	Dataset	Mean Learn	SD Learn	Mean Apply	SD Apply
Reference	Windows 10 x64 (build 9200)	3.3.1 (2016-06-21)	LQNO	i7-5960X	286	30.58	0.55	33.05	0.12
Reference	Windows 10 x64 (build 9200)	3.3.1 (2016-06-21)	LQNO	i7-5960X	miRXplore	99.87	0.57	86.31	0.42
Reference	Windows 10 x64 (build 9200)	3.3.2 (2016-10-31)	LQNO	T9300	miRXplore	119.62	60.67	138.99	46.46
Reference	Windows 7 x64 (build 7601)	3.3.2 (2016-10-31)	LQNO	i7-3770	286	35.43	0.26	45.08	0.21
Reference	Windows 7 x64 (build 7601)	3.3.2 (2016-10-31)	LQNO	i7-3770	miRXplore	72.92	0.49	87.58	1.59
TMB	Windows 10 x64 (build 9200)	3.3.1 (2016-06-21)	LQNO	i7-5960X	286	4.88	0.10	6.39	0.23
TMB	Windows 10 x64 (build 9200)	3.3.1 (2016-06-21)	LQNO	i7-5960X	miRXplore	22.48	0.49	21.21	0.46
TMB	Windows 10 x64 (build 9200)	3.3.1 (2016-06-21)	NBI	i7-5960X	286	15.76	0.56	18.57	0.45
TMB	Windows 10 x64 (build 9200)	3.3.1 (2016-06-21)	NBI	i7-5960X	miRXplore	62.77	2.32	60.46	1.62
TMB	Windows 10 x64 (build 9200)	3.3.2 (2016-10-31)	LQNO	T9300	286	10.88	0.51	14.39	0.40
TMB	Windows 10 x64 (build 9200)	3.3.2 (2016-10-31)	LQNO	T9300	miRXplore	47.95	0.61	44.01	0.67
TMB	Windows 7 x64 (build 7601)	3.3.2 (2016-10-31)	LQNO	i7-3770	286	5.93	0.38	7.03	0.07
TMB	Windows 7 x64 (build 7601)	3.3.2 (2016-10-31)	LQNO	i7-3770	miRXplore	27.05	0.28	25.52	0.28

Timings (means and standard deviations over 20 repeat runs, in seconds) for the learning and application of correction factors in two different datasets. All timings were obtained under Windows 64bit operating systems running multithreaded versions of R (Microsoft R Open). The Reference implementation can leverage multicore processors to speed up calculations, whereas the TMB implementation (gamlssAD) is inherently a single core program.

References

1. Kaufmann,G. and Littauer,U.Z. (1974) Covalent Joining of Phenylalanine Transfer Ribonucleic Acid Half-Molecules by T4 RNA Ligase. *Proc Natl Acad Sci U S A*, **71**, 3741–3745.

2. Cranston,J.W., Silber,R., Malathi,V.G. and Hurwitz,J. (1974) Studies on Ribonucleic Acid Ligase CHARACTERIZATION OF AN ADENOSINE TRIPHOSPHATE-INORGANIC PYROPHOSPHATE EXCHANGE REACTION AND DEMONSTRATION OF AN ENZYME-ADENYLATE COMPLEX WITH T4 BACTERIOPHAGE-INDUCED ENZYME. *J. Biol. Chem.*, **249**, 7447–7456.

3. Walker,G.C., Uhlenbeck,O.C., Bedows,E. and Gumport,R.I. (1975) T4-induced RNA ligase joins single-stranded oligoribonucleotides. *Proc Natl Acad Sci U S A*, **72**, 122–126.

4. Ohtsuka,E., Nishikawa,S., Sugiura,M. and Ikehara,M. (1976) Joining of ribooligonucleotides with T4 RNA ligase and identification of the oligonucleotide-adenylate intermediate. *Nucl. Acids Res.*, **3**, 1613–1624.

5. Last,J.A. and Anderson,W.F. (1976) Purification and properties of bacteriophage T4-induced RNA ligase. *Archives of Biochemistry and Biophysics*, **174**, 167–176.

6. Higgins,N.P., Geballe,A.P., Snopek,T.J., Sugino,A. and Cozzarelli,N.R. (1977) Bacteriophage T4 RNA ligase: preparation of a physically homogeneous, nuclease-free enzyme from hyperproducing infected cells. *Nucl. Acids Res.*, **4**, 3175–3186.

7. Sugino,A., Snoper,T.J. and Cozzarelli,N.R. (1977) Bacteriophage T4 RNA ligase. Reaction intermediates and interaction of substrates. *J. Biol. Chem.*, **252**, 1732–1738.

8. Nandakumar,J., Ho,C.K., Lima,C.D. and Shuman,S. (2004) RNA Substrate Specificity and Structure-guided Mutational Analysis of Bacteriophage T4 RNA Ligase 2. *J. Biol. Chem.*, **279**, 31337–31347.

9. Nandakumar,J., Shuman,S. and Lima,C.D. (2006) RNA Ligase Structures Reveal the Basis for RNA Specificity and Conformational Changes that Drive Ligation Forward. *Cell*, **127**, 71–84.

10. Yin,S., Ho,C.K. and Shuman,S. (2003) Structure-Function Analysis of T4 RNA Ligase 2. *J. Biol. Chem.*, **278**, 17601–17608.

11. Ho,C.K., Wang,L.K., Lima,C.D. and Shuman,S. (2004) Structure and Mechanism of RNA Ligase. *Structure*, **12**, 327–339.

12. Cherepanov,A.V. and de Vries,S. (2003) Kinetics and thermodynamics of nick sealing by T4 DNA ligase. *European Journal of Biochemistry*, **270**, 4315–4325.

13. Raae,A.J., Kleppe,R.K. and Kleppe,K. (1975) Kinetics and Effect of Salts and Polyamines on T4 Polynucleotide Ligase. *European Journal of Biochemistry*, **60**, 437–443.

14. Uhlenbeck,O.C. (1983) T4 RNA ligase. *Trends in Biochemical Sciences*, **8**, 94–96.

15. Snopek,T.J., Sugino,A., Agarwal,K.L. and Cozzarelli,N.R. (1976) Catalysis of DNA joining by bacteriophage T4 RNA ligase. *Biochem. Biophys. Res. Commun.*, **68**, 417–424.

16. Krug,M. and Uhlenbeck,O.C. (1982) Reversal of T4 RNA ligase. *Biochemistry*, **21**, 1858–1864.

17. Segel,L. and Slemrod,M. (1989) The Quasi-Steady-State Assumption: A Case Study in Perturbation. *SIAM Rev.*, **31**, 446–477.

18. Segel,L.A. (1988) On the validity of the steady state assumption of enzyme kinetics. *Bull. Math. Biol.*, **50**, 579–593.

19. Chou,T.C. and Talaly,P. (1977) A simple generalized equation for the analysis of multiple inhibitions of Michaelis-Menten kinetic systems. *J. Biol. Chem.*, **252**, 6438–6442.

20. Laurent,L.C., Abdel-Mageed,A.B., Adelson,P.D., Arango,J., Balaj,L., Breakefield,X., Carlson,E., Carter,B.S., Majem,B., Chen,C.C., *et al.* (2015) Meeting report: discussions and preliminary findings on extracellular RNA measurement methods from laboratories in the NIH Extracellular RNA Communication Consortium. *Journal of Extracellular Vesicles*, **4**.

21. Kaufmann,G., Klein,T. and Littauer,U.Z. (1974) T4 RNA ligase: Substrate chain length requirements. *FEBS Letters*, **46**, 271–275.

22. Hinton,D.M., Baez,J.A. and Gumport,R.I. (1978) T4 RNA Ligase joins 2'-deoxyribonucleoside 3',5'-bisphosphates to oligodeoxyribonucleotides. *Biochemistry*, **17**, 5091–5097.

23. McCoy,M.I. and Gumport,R.I. (1980) T4 ribonucleic acid ligase joins single-strand oligo(deoxyribonucleotides). *Biochemistry*, **19**, 635–642.

24. The Theory of Branching Processes | Theodore Edward Harris | Springer.

25. Stolovitzky,G. and Cecchi,G. (1996) Efficiency of DNA replication in the polymerase chain reaction. *PNAS*, **93**, 12947–12952.

26. Lalam,N. (2007) Statistical Inference for Quantitative Polymerase Chain Reaction Using a Hidden Markov Model: A Bayesian Approach. *Statistical Applications in Genetics & Molecular Biology*, **6**, 19–33.

27. Krawczak,M., Reiss,J., Schmidtke,J. and Rösler,U. (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucl. Acids Res.*, **17**, 2197–2201.

28. Kemp,A.W. and Kemp,C.D. (1966) An Alternative Derivation of the Hermite Distribution. *Biometrika*, **53**, 627–628.

29. Kemp,C.D. and Kemp,A.W. (1967) A Special Case of Fisher's 'Modified Poisson Series'. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **29**, 103–104.

30. Patil,G.P. (1964) On Certain Compound Poisson and Compound Binomial Distributions. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **26**, 293–294.

31. Johnson,N.L., Kemp,A.W. and Kotz,S. (2005) Univariate Discrete Distributions John Wiley & Sons.

32. NIST Digital Library of Mathematical Functions.

33. Olver,F.W.J., Lozier,D.W., Boisvert,R.F. and Clark,C.W. eds. (2010) NIST Handbook of Mathematical Functions Cambridge University Press, New York, NY.

34. Johansson,F. and others (2013) mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.18).

35. Takahasi,H. and Mori,M. (1974) Double Exponential Formulas for Numerical Integration. *Publications of the Research Institute for Mathematical Sciences*, **9**, 721–741.

36. Bailey,D.H., Jeyabalan,K. and Li,X.S. (2005) A comparison of three high-precision quadrature schemes. *Experiment. Math.*, **14**, 317–329.