KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-023-00821-w

TECHNICAL CONTRIBUTION

Two-Step Parameter Estimation for Read Feature Models

Florian Erhard[1]

Received: 31 March 2023 / Accepted: 28 November 2023

Abstract
Over the last two decades, the field of molecular biology has witnessed a revolution due to the development of next generation sequencing (NGS) technologies. NGS enables researchers to routinely generate huge amounts of data that can be used to pursue a large variety of questions in diverse biological systems. The development of these techniques has propelled the emergence of a sub-discipline within computational biology that is concerned with developing methods and statistical models to derive quantitative information from the complex and often indirect data that are generated by NGS. Often, NGS analysis results in particular patterns per biological entity that can be exploited to estimate quantitative parameters of biological interest. Here, I define read feature models (RFMs) as a general framework for such data. RFMs entail global, genome-wide parameters as well as parameters per biological entity, suggesting a two-step procedure for parameter estimation. I describe the analysis of metabolic RNA labeling data as an example of an RFM and analyze and discuss the merits and shortcomings of the two-step estimation.

1 Introduction

By the rapid and still ongoing development of next generation sequencing (NGS) technologies it is now possible to obtain the nucleotide sequences of currently billions of DNA molecules from a sequencing library in the matter of hours on a single machine [1, 2]. Here, I focus on second generation sequencing which generates reads with lengths ranging from 50 to 250 nucleotides, but the general concepts proposed here also apply to third generation sequencing which produces much longer reads but at currently lower throughput [3, 4].

While sequencing libraries consist of DNA, they can also be generated from RNA using reverse transcription [5]. Sequencing RNA provides numerous opportunities to study dynamic processes that occur in living cells. Arguably, the most prominent application is called RNA-seq. The sequences in an RNA-seq library correspond to RNA fragments that have been randomly sampled from all mRNAs extracted from a biological sample. After sequencing, reads are mapped to a reference sequence such as the genome, and the number of reads per gene is determined. Such read counts from an RNA-seq experiment thus approximate the individual expression levels of all genes.

The fundamental principle of RNA-seq is to obtain estimates of quantitative biological parameters based on counting specific sequences. It is, however, not the only example: NGS has been used to quantitatively measure binding of transcription factors to their target sites [6], initiation and elongation rates of RNA polymerases [7], rates of splicing [8] and RNA export from the nucleus [9], translation rates [10] and RNA decay [11], the thermodynamic ensemble of RNA structures [12] and interactions among RNAs and RNA binding proteins [13] among many other applications [14]. All these examples of quantitative NGS have in common that due to the biochemical steps performed in the wet-lab, information on particular parameters is introduced into the sequencing library. Which parameters can be measured by sequencing is virtually only limited by the creativity of the researcher [14]. The purpose of data analysis then is to extract this information from the sequencing reads by employing statistical models.

Here, we differentiate between two kinds of statistical models for quantitative NGS, namely read count models (RCMs) and what I here introduce as read feature models (RFMs). In this article, after defining the different scopes of RCMs and RFMs, I will formally introduce RFMs. Our recently developed GRAND-SLAM [15] method for the analysis of nucleotide conversion RNA-seq data fits into this

✉ Florian Erhard
Florian.Erhard@informatik.uni-regensburg.de

[1] Chair of Computational Immunology, University of Regensburg, 93053 Regensburg, Germany

Published online: 04 January 2024

statistical framework of RFMs, and I use this example to discuss advantages and potential shortcomings of a two-step approach for parameter estimation for RFMs.

2 Read Count Models

There is ample literature about RCMs [16–18]. RCMs are concerned with modeling the number of reads per biological entity using replicated biological samples. A frequently used model is the negative binomial distribution for which the mean and dispersion parameters can in principle be estimated independently for each gene from a large enough number of replicates. Importantly, however, only two or three replicates are common practice, which would result in highly variable dispersion estimates. For that reason, shrinkage estimators that share information across genes are widely used under the assumption that overdispersion is the same [19] or at least similar [20] for genes with similar expression level.

A simple application of RCMs is testing for differential gene expression in a pairwise comparison between two conditions using RNA-seq, e.g. treatment T vs control C. This can be accomplished by a hypothesis test asking whether $\mu_T = \mu_C$. Generalized linear models are a convenient framework for such tests and can also be used to analyze more complex experimental scenarios such as multiple conditions or multifactorial designs [20].

The data generation process for RNA-seq is highly complex and consists of biochemical reactions taking place during fragmentation of RNA, reverse transcription, adapter ligation, amplification using polymerase chain reaction and sequencing [5, 14]. The negative binomial distribution is not only an appealing model because it is able to handle the overdispersion that is observed for such data but can be seen to resemble these complex steps of data generation in a coarse-grained manner: If we assume the RNA level for a gene among replicated experiments to be gamma distributed, and consider the generation of reads for this gene to be a random sampling process from this RNA level (in competition with the total levels of all other genes), then a gamma-Poisson mixture distribution emerges for the read count, which is the negative binomial distribution. Of note, the overdispersion likely also includes technical variance due to library preparation in addition to biological variability.

RCMs can also be used for other applications than RNA-seq, e.g. to compare transcription factor occupancy on binding sites using ChIP-seq [21] or the strength of translation using Ribo-seq [22]. There are also scenarios where the assignment of reads to biological entities is not unique. For instance, genes of higher eukaryotes have different transcript isoforms, which often share large parts of their sequences. For reads corresponding to such sequences it is a priori not clear from which transcript isoform they originate. Isoform level quantification can be performed by treating the assignment of reads to isoforms as latent variable and using the EM algorithm or variational Bayes for inference [23, 24].

In summary, RCMs model read counts that belong to biological entities and are concerned with differences between biological conditions. However, many NGS applications generate patterns in the data that can be used to make more fine-grained inferences for each individual sample. This is where read feature models (RFM) come into play which are concerned with recognizing and exploiting these patterns.

3 Read Feature Models

NGS data derived from a single biological sample consists of short reads R. Each read $r \in R$ belongs to a biological entity, and we denote all reads belonging to the biological entity i as R_i. Usually, only specific features $s(r_j)$ of a read r_j are relevant and provide the sufficient statistics $D_i = \{s(r_j) | r_j \in R_i\}$ for parameter estimation for a biological entity i. An RFM consists of a parametric family $\mathcal{F}$ and a parameter vector $\theta = (\theta_G, \phi_1, \ldots, \phi_N)$ involving global parameters θ_G and parameters ϕ_i for the N individual biological entities. Each $d_j = s(r_j) \in D_i$ is modeled by a probability distribution from the parametric family $\mathcal{F}$ with parameters θ_G and ϕ_i, i.e. $d_j \sim \mathcal{F}(\theta_G, \phi_i)$. Thus, each read, or at least the features relevant for parameter estimation, emerge from a probability distribution that depends on a set of global parameters and a gene specific parameter but is independent of the specific parameters from other genes. Often, ϕ_i is one-dimensional and represents an activity or abundance of some sort for biological entity i, and is usually the biological parameter of interest. The global parameters θ_G by contrast often represent the stochastic behavior of the biochemical procedures that are used to generate the sequencing library. Thus, like RCMs, RFMs do not only try to fit observed data, but can be considered to model the actual data generation process in a coarse-grained manner.

There are, however, many fundamental differences between RCMS and RFMs. RCMs are used to compare quantities such as RNA levels (RNA-seq) or occupancies (ChIP-seq) across replicates and conditions. Thus, RCMs are concerned with the number of reads for a biological entity across replicated experiments. By contrast, the purpose of RFMs rather is to extract qualitative or quantitative information introduced into the sequencing library by the biochemical steps taken to generate the library. RFMs therefore focus on a single biological sample and model the features D_i of

all the reads mapped to a single biological entity i instead of their number. The function s might extract features such as the read length, its positioning within the entity or patterns of mismatched nucleotides.

An example of an RFM is implemented in our PRICE method [25]: PRICE aims to find stretches on RNAs called open reading frames (ORFs) that are translated by ribosomes into proteins based on data generated by a technique called Ribo-seq [10]. Due to the way RNA is prepared for sequencing, Ribo-seq reads corresponding to actively translating ribosomes have specific lengths and have a periodic pattern with regard to their positions along such stretches. PRICE learns the global parameters of an RFM using the ORFs of known proteins and can then be used to predict so far unknown translated ORFs. PRICE has been used by us and others to identify thousands of short ORFs in the human genome [26] and dozens to hundreds in clinically relevant viruses such as SARS-CoV-2 [27] and human cytomegalovirus [25, 28]. Moreover, PRICE enabled us to show that peptides originating from such short ORFs are presented via the major histocompatibility complex I (MHC-I) [29], defining a new class of antigens that might play a hitherto unknown role in the T cell mediated defense against infection and cancer [26].

A second use case for RFMs is PAR-CLIP, which is a quantitative NGS technique for the discovery of the binding sites on mRNAs of an important class of short regulatory RNAs called microRNAs [13]. A microRNA binding site can be as short as six consecutive nucleotides on an mRNA. PAR-CLIP generates clusters of reads at such binding sites with a specific pattern of start and end positions, and additionally induces specific mismatches close to the microRNA binding site. Our PARma method [30] utilizes an RFM to learn these patterns of positions and mismatches to precisely define the binding site within a cluster and by sequence complementary also the microRNA that binds there. An analysis of several data sets including several PAR-CLIP experiments revealed that microRNA binding to an mRNA generally is a context-dependent phenomenon adding an additional layer of complexity to the gene regulatory network [31].

These examples demonstrate that parameter estimation for RFMs can be done in a two-step process: First, the global parameters θ_G are estimated using the pooled data across many or all biological entities. θ_G is then considered constant for the estimation of gene-wise parameters in the second step. These examples also show that for building RFMs, a detailed understanding of the data generation process for a particular type of experiment is necessary.

4 An RFM for Nucleotide Conversion RNA-seq

Being able to quantify RNA that was synthesized during a defined period in addition to total RNA levels has many advantages over standard RNA-seq. For instance, this allows to estimate parameters describing the kinetics of gene expression (synthesis rates, degradation rates) [11, 32], and it enables to reveal short-term regulatory changes of gene expression in much greater detail than normal RNA-seq [33]. The most widely used methods for quantifying newly synthesized RNA are based on metabolic RNA labeling.

Metabolic RNA labeling utilizes nucleoside analogs such as 4-thiouridine (4sU) that are supplied to a cell culture for a defined period (e.g. 2h). Cells take up the 4sU and incorporate it into newly synthesized RNA instead of normal uridine (U). After e.g. 2h, RNA is extracted and treated with compounds that result in 4sU being sequenced as cytosine (C) [11]. The reads are then mapped to the genome sequence, where the U found on RNA corresponds to thymine (T). Thus, the incorporation of 4sU and its conversion in the RNA gives rise to a T-to-C mismatch in the mapped reads. Such T-to-C mismatches therefore provide evidence for the read originating from an RNA molecule that was transcribed during the last 2h.

The parameter of interest is the gene-wise new-to-total RNA ratio (NTR). The NTR is the starting point to derive other, biologically relevant parameters. For instance, there is a 1-to-1 correspondence between the NTR and the kinetic rate of RNA degradation [15]. Estimating the NTR is non-trivial for two reasons: First, library preparation and sequencing can also introduce mismatches, including T-to-C. Thus, a mismatch in a read can either be due to such an error, or because of the conversion of an incorporated 4sU. Second, and more importantly, only a small and typically unknown percentage of U are substituted by 4sU during transcription. Consequently, many reads that indeed originate from a newly synthesized RNA might not cover any site of 4sU incorporation by chance. We estimated that for published data [11], more than 75% of all reads originating from a newly synthesized RNA does not exhibit any T-to-C mismatch [34]. Thus, the fraction of reads having T-to-C mismatches among all reads belonging to a gene is a biased estimator of the NTR: Due to sequencing errors, it might overestimate the NTR, and due to reads not covering 4sU sites by chance, it might also underestimate the NTR. We previously proposed our GRAND-SLAM approach to estimate NTRs in an unbiased manner [15].

To define the model behind GRAND-SLAM in the framework of RFMs, we denote the probabilities of a T-to-C mismatch for reads originating from a newly synthesized RNA or pre-existing RNA molecule p_{new} and p_{old}, respectively. Thus, p_{old} corresponds to the probability of a sequencing error or any other base substitution that can happen during library preparation. By contrast, $p_{new} = p_{old} + p_{4sU}$, i.e. p_{new} includes the probability of errors and of the incorporation and conversion of a 4sU. Both p_{new} and p_{old} are global parameters and are the same for all genes. By contrast, the parameters $\nu_1, \ldots, \nu_N$ represent the gene specific NTRs for all genes.

The features extracted for a read are $s(r) = k_r$, with k_r being the number of T-to-C mismatches observed for read r. The parametric family of the RFM is a two-component binomial mixture model $BinomMix\left(p_{old}, p_{new}, \nu, n\right)$ with probability mass function

$$P\left(k_r; n, p_{old}, p_{new}, \nu\right) = (1-\nu) \cdot Binom\left(k_r; n, p_{old}\right) + \nu \cdot Binom\left(k_r; n, p_{new}\right)$$

$$Binom(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Thus, the global parameters of the RFM are $\theta_G = \left(p_{old}, p_{new}\right)$, the gene-wise parameters are the NTRs $\left(\nu_1, \ldots, \nu_N\right)$, and the sufficient statistic is $k_r \sim BinomMix\left(p_{old}, p_{new}, \nu_i, n_r\right)$ which is distributed according to the parametric family defining the RFM. Note that n_r here is the number of T covered by the read r in the genome, i.e. the maximal number of possible T-to-C mismatches, which can be considered a constant.

5 RFM Parameter Estimation Using a Two-Step Approach

Computing maximum likelihood estimators (MLE) or the Bayesian posterior distribution for the high dimensional parameter θ of an RFM is conceptually straightforward and could be done by numerical optimization to obtain the MLE or Markov chain monte carlo (MCMV) sampling for approximating the posterior. Of note, N can be quite large, making numerical optimization or MCMC computationally challenging. However, the special structure of RFMs suggests a two-step parameter estimation procedure that is computationally much more efficient: First, by pooling data from all biological entities, the global parameters θ_G are estimated and then considered constants. With that, the high-dimensional estimation problem decomposes into N independent low-dimensional problems.

For the nucleotide conversion RNA-seq RFM, p_{old} can be estimated from control samples that were not labeled with 4sU. Such control samples are usually included into experiments to test for 4sU induced effects on the biology of the cells. Since there is no 4sU, the mixture model reduces to a binomial distribution making estimation of p_{old} straightforward [15]. p_{new} can be estimated by introducing the nuisance parameter ν, which is the global NTR, i.e. the fraction of labeled RNA across all genes. This two-dimensional estimation problem can efficiently be solved using numerical optimization [15]. Once point estimates for the parameters p_{old} and p_{new} are available, they are treated as constant and only the gene specific NTR ν_i must be estimated for each gene. In GRAND-SLAM, the full posterior distribution of each ν_i is computed by numerical integration.

6 The Two-Step Approach Introduces Negligible Bias

Using point estimates for the global parameters θ_G and considering them as constants for the second step comes with the danger of introducing bias into the estimator of the gene specific parameters. For instance, for the GRAND-SLAM RFM, if p_{new} is overestimated, the ν_i are expected to be underestimated: Consider a gene with a true NTR of 1, i.e. all reads indeed originate from a labeled RNA molecule. The expected overall percentage of T-to-C mismatches for this gene therefore is equal to the true p_{new}. If the $\hat{p}_{new}$ is overestimated, i.e. $\hat{p}_{new} > p_{new}$, the required percentage of T-to-C mismatches to achieve $\nu_i = 1$ is $\hat{p}_{new}$, which is greater than p_{new}. Thus, overestimated p_{new} bias the ν_i towards 0. To investigate the magnitude of such bias empirically, I conducted simulation experiments.

Data were simulated from a *BinomMix* model with $N = 2.5 \cdot 10^7$, $p_{old} = 4 \cdot 10^{-4}$, $p_{new} = 0.02$ and $\nu = 0.15$, all reflecting realistic values for the total number of reads for a single sample, sequencing errors, 4sU incorporation and typical RNA turnover in mammalian cells for 1h of labeling [32]. For each read the number of T positions n was drawn from a distribution reflecting a read length of 100. To mimic the estimation of p_{old} by an additional, 4sU naïve sample, it was treated as a constant. The joint posterior distributions indeed show anticorrelation of p_{new} and ν (an example is shown in Fig. 1A), demonstrating that ν is biased towards 0 if p_{new} is overestimated. The 95% credible interval (CI) computed from the marginal posterior for the example in Fig. 1A was approximately [0.01996,0.02006], i.e. the relative uncertainty defined as the size of the 95% CI divided by the true value 0.02 was in the range of 0.5%.

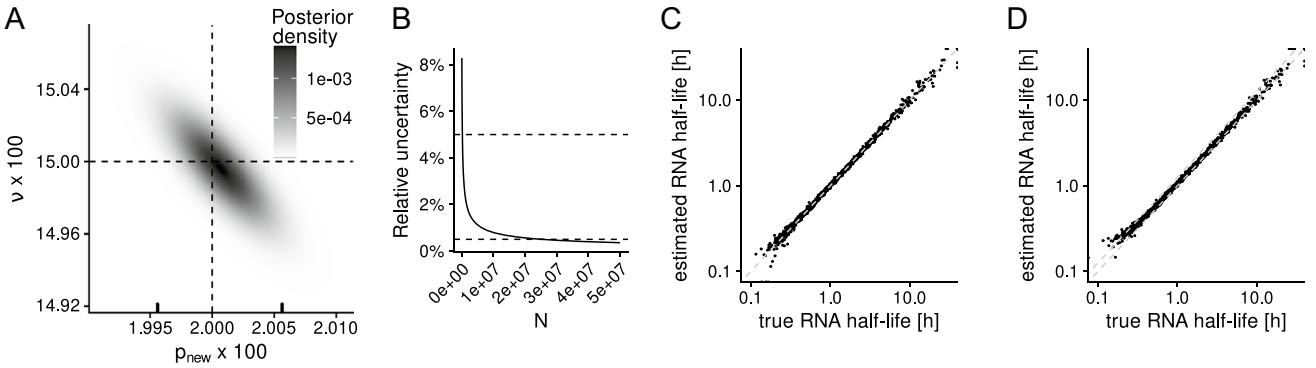

Fig. 1 **A** The joint posterior density distribution of data simulated with $p_{new} = 0.02$ and $\nu = 0.15$ is shown. The true values are marked by dashed lines, and the 95% credible interval (CI) of the marginal posterior for p_{new} is indicated at the bottom. **B** The relative uncertainty defined as the size of the 95% CI divided by the true value ($p_{new} = 0.02$) is shown for multiple simulations with total read counts N ranging from 300.000 to 50 mio reads. Relative uncertainty cutoffs of 0.5% and 5% are indicated. **C**–**D** Half-lives simulated for individual genes (n = 1000) are scattered against their estimated half-lives with a p_{new} that is overestimated by 0.5% (**C**) or 5% (**D**). The main diagonals (dashed line) representing no bias are indicated. For (**D**), a second dashed line above the main diagonal represents a $\log_2$ fold change of 0.1 between simulated and estimated half-lives

The accuracy of the point estimate for p_{new} mostly depends on the total number of reads N. Thus, additional experiments with N ranging from 300.000 to 50 mio reads were simulated, and the 95% CI of p_{new} and the relative uncertainty as defined above were computed (Fig. 1B). Relative uncertainties dropped steeply with increasing N and were below 1% with 6.3 mio reads. More reads improved the uncertainty only marginally. Thus, based on these empirical analyses, p_{new} is estimated with high relative accuracy for standard experimental setting with > 20 mio reads per sample.

To evaluate the effects of these uncertainties in the subsequent estimation of gene-wise RNA half-lives, which is a biologically relevant parameter and has a 1-to-1 correspondence to ν_i [15], the read simulator built into our grandR package [32] was used to generate data for individual genes with $p_{new} = 0.02$. Then, the ν_i were estimated based on an overestimated p_{new}. To reduce the effect of variance in the estimates of ν_i, 10.000 reads were simulated for each gene. With a relative overestimation of 0.5%, no bias in the RNA half-life estimates was discernable, i.e. the effects of an overestimated p_{new} was much smaller than the variance in the ν_i estimates even for genes with 10.000 reads (Fig. 1C). With a relative overestimation of 5%, however, especially short half-lives were clearly overestimated (Fig. 1D). The magnitude of the overestimation, however, was low, with most genes having a $\log_2$ fold change of estimated vs true RNA half-life below 0.1.

In summary, this simulation study indicates that inaccurate estimation of p_{new} in the first step has little to no effect on the estimates of ν_i in the second step for realistic data sets.

7 Discussion

There are many applications of NGS that result in specific patterns of sequencing reads mapped to the biological entities of interest. RFMs focus on modeling these patterns to extract biologically meaningful information from sequencing data. GRAND-SLAM is an RFM for nucleotide conversion RNA-seq to estimate the gene-wise NTR, which provides information about the dynamics of gene expression. When gene expression is at steady-state, the NTR can be transformed into the RNA half-life [15]. With few reads for a gene, the NTR cannot be estimated accurately, and if the NTR is close to 0 or 1, the transformation into the RNA half-life inflates even slight inaccuracies substantially [15, 35]. It is therefore important to quantify the uncertainty in these parameters, e.g. using Bayesian posteriors. Even when gene expression is not at steady-state, RNA half-lives can be estimated if gene expression from an additional prior timepoint is known [32], which introduces another source of uncertainty in the estimation.

The special structure of RFMs greatly facilitates the estimation of posteriors, since in the two-step approach, the NTR is estimated per gene by solving a univariate parameter estimation problem in the second step. This enables GRAND-SLAM to efficiently compute exact posteriors without MCMC sampling. Here, I investigated whether inaccurate estimation of global parameters introduce bias into the estimation of the gene-wise parameters in this two-step process. The empirical analyses presented here indicate that for realistic data sets inaccuracies of the global parameters only have negligible effects on the estimates of the

gene-wise parameters for nucleotide conversion sequencing RNA-seq.

In the definition of RFMs I explicitly made the assumption that each sequencing read is uniquely assigned to a single biological entity. A similar assumption has been made for the fundamental RCMs modeling RNA abundance [20]. There are scenarios, where this is not the case: Typically, short RNA-seq reads map to a single gene but occur in multiple isoforms of this gene in higher eukaryotes. Thus, for estimating RNA abundances (using RCMs) for all individual transcript isoforms, methods have been developed that treat the assignment of reads to isoforms as latent variable [23, 24]. The same approaches can also be implemented for RFMs, i.e. the latent variable can be integrated into the model, making the estimation slightly more complicated. Alternatively, the estimate of the latent variable could be used as a probabilistic but fixed assignment of reads to isoforms. Computing this probabilistic assignment as an additional prior step would make any RFM directly applicable to cases with non-unique reads. However, in contrast to integrating the latent variable into the RFM model, this procedure would not make full use of the patterns that are modeled by the RFM for the probabilistic read assignment. It is an interesting future direction to integrate latent variables into specific RFM models and evaluate whether treating the assignment as a separate first step provides sufficiently accurate results.

While the simulation approach proposed here demonstrates minimal bias introduced by the two-step approach for the GRAND-SLAM model, it is important to note that this methodology might not be as robust for other RFMs. The two-step approach proposed here should not be employed if global parameters estimated from it differ significantly to those obtained via a joint estimation. The susceptibility to this discrepancy is inherently associated with the specific RFM in use. Therefore, it is important to conduct a rigorous evaluation for specific RFMs to ascertain the reliability of the two-step approach. The simulation methodology presented here offers an empirical framework to probe such potential challenges.

Acknowledgements The author received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) by project grant ER 927/2-1 and in the framework of the Research Unit FOR5200 DEEP-DV (443644894) project ER 927/4-1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author declares that he has no conflicts of interest.

Code availability Code to reproduce the simulations and all figures is available on Zenodo (https://doi.org/10.5281/zenodo.10015199).

References

1. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46. https://doi.org/10.1038/nrg2626
2. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49
3. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends Genet 34:666–681. https://doi.org/10.1016/j.tig.2018.05.008
4. Wang Y, Zhao Y, Bollas A et al (2021) Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39:1348–1365. https://doi.org/10.1038/s41587-021-01108-x
5. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226
6. Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 13:840–852. https://doi.org/10.1038/nrg3306
7. Schwalb B, Michel M, Zacher B et al (2016) TT-seq maps the human transient transcriptome. Science 352:1225–1228. https://doi.org/10.1126/science.aad9841
8. Windhager L, Bonfert T, Burger K et al (2012) Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. Genome Res 22:2031–2042. https://doi.org/10.1101/gr.131847.111
9. Lefaudeux D, Sen S, Jiang K, Hoffmann A (2022) Kinetics of mRNA nuclear export regulate innate immune response gene expression. Nat Commun 13:7197. https://doi.org/10.1038/s41467-022-34635-5
10. Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet 15:205–213. https://doi.org/10.1038/nrg3645
11. Herzog VA, Reichholf B, Neumann T et al (2017) Thiol-linked alkylation of RNA to assess expression dynamics. Nat Methods 14:1198. https://doi.org/10.1038/nmeth.4435
12. Strobel EJ, Yu AM, Lucks JB (2018) High-throughput determination of RNA structures. Nat Rev Genet 19:615–634. https://doi.org/10.1038/s41576-018-0034-x
13. Hafner M, Landthaler M, Burger L et al (2010) Transcriptome-wide identification of RNA-binding protein and MicroRNA target sites by PAR-CLIP. Cell 141:129–141. https://doi.org/10.1016/j.cell.2010.03.009

14. Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. Nat Rev Genet 20:631–656. https://doi.org/10.1038/s41576-019-0150-2
15. Jürges C, Dölken L, Erhard F (2018) Dissecting newly transcribed and old RNA using GRAND-SLAM. Bioinformatics 34:i218–i226. https://doi.org/10.1093/bioinformatics/bty256
16. Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinform 14:91. https://doi.org/10.1186/1471-2105-14-91
17. Corchete LA, Rojas EA, Alonso-López D et al (2020) Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci Rep 10:19737. https://doi.org/10.1038/s41598-020-76881-x
18. Rapaport F, Khanin R, Liang Y et al (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14:3158. https://doi.org/10.1186/gb-2013-14-9-r95
19. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106. https://doi.org/10.1186/gb-2010-11-10-r106
20. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8
21. Eder T, Grebien F (2022) Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection. Genome Biol 23:119. https://doi.org/10.1186/s13059-022-02686-y
22. Zhong Y, Karaletsos T, Drewe P et al (2017) RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. Bioinformatics 33:139–141. https://doi.org/10.1093/bioinformatics/btw585
23. Glaus P, Honkela A, Rattray M (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 28:1721–1728. https://doi.org/10.1093/bioinformatics/bts260
24. Trapnell C, Hendrickson DG, Sauvageau M et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31:46–53. https://doi.org/10.1038/nbt.2450
25. Erhard F, Halenius A, Zimmermann C et al (2018) Improved Ribo-seq enables identification of cryptic translation events. Nat Methods 15:363–366. https://doi.org/10.1038/nmeth.4631
26. Ouspenskaia T, Law T, Clauser KR et al (2022) Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol 40:209–217. https://doi.org/10.1038/s41587-021-01021-3
27. Finkel Y, Mizrahi O, Nachshon A et al (2021) The coding capacity of SARS-CoV-2. Nature 589:125–130. https://doi.org/10.1038/s41586-020-2739-1
28. Stern-Ginossar N, Weisburd B, Michalski A et al (2012) Decoding human cytomegalovirus. Science 338:1088–1093. https://doi.org/10.1126/science.1227919
29. Erhard F, Dölken L, Schilling B, Schlosser A (2020) Identification of the cryptic HLA-I immunopeptidome. Cancer Immunol Res 8:1018–1026. https://doi.org/10.1158/2326-6066.CIR-19-0886
30. Erhard F, Dolken L, Jaskiewicz L, Zimmer R (2013) PARma: identification of microRNA target sites in AGO-PAR-CLIP data. Genome Biol 14:R79. https://doi.org/10.1186/gb-2013-14-7-r79
31. Erhard F, Haas J, Lieber D et al (2014) Widespread context dependency of microRNA-mediated regulation. Genome Res. https://doi.org/10.1101/gr.166702.113
32. Rummel T, Sakellaridi L, Erhard F (2023) grandR: a comprehensive package for nucleotide conversion RNA-seq data analysis. Nat Commun 14:3559. https://doi.org/10.1038/s41467-023-39163-4
33. Muhar M, Ebert A, Neumann T et al (2018) SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. Science. https://doi.org/10.1126/science.aao2793
34. Erhard F, Saliba A-E, Lusser A et al (2022) Time-resolved single-cell RNA-seq using metabolic RNA labelling. Nat Rev Methods Primers 2:1–18. https://doi.org/10.1038/s43586-022-00157-z
35. Uvarovskii A, Vries ISN, Dieterich C (2019) On the optimal design of metabolic RNA labeling experiments. PLoS Comput Biol 15:e1007252. https://doi.org/10.1371/journal.pcbi.1007252