BMC Bioinformatics

RESEARCH

Open Access

Internal and external normalization of nascent RNA sequencing run-on experiments

Zachary L. Maas[1,2] and Robin D. Dowell[1,2,3*]

*Correspondence:
robin.dowell@colorado.edu

[1] Department of Computer Science, University of Colorado, Boulder, USA
[2] BioFrontiers Institute, University of Colorado, Boulder, USA
[3] Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, USA

Abstract

In experiments with significant perturbations to transcription, nascent RNA sequencing protocols are dependent on external spike-ins for reliable normalization. Unlike in RNA-seq, these spike-ins are not standardized and, in many cases, depend on a run-on reaction that is assumed to have constant efficiency across samples. To assess the validity of this assumption, we analyze a large number of published nascent RNA spike-ins to quantify their variability across existing normalization methods. Furthermore, we develop a new biologically-informed Bayesian model to estimate the error in spike-in based normalization estimates, which we term Virtual Spike-In (VSI). We apply this method both to published external spike-ins as well as using reads at the 3′ end of long genes, building on prior work from Mahat (Mol Cell 62(1):63–78, 2016. https://doi.org/10.1016/j.molcel.2016.02.025) and Vihervaara (Nat Commun 8(1):255, 2017. https://doi.org/10.1038/s41467-017-00151-0). We find that spike-ins in existing nascent RNA experiments are typically under sequenced, with high variability between samples. Furthermore, we show that these high variability estimates can have significant downstream effects on analysis, complicating biological interpretations of results.

Keywords: Nascent RNA sequencing, Normalization, Bayesian

Introduction

Effective normalization is essential for rigorous analysis of high throughput sequencing data. In sequencing data, normalization identifies a set of features that are expected to be invariant between two data sets and leverages these to counteract the effects of systematic experimental bias and technical variation. Broadly, there are only two possibilities for the source of these invariant features: external spike-in controls or an internal invariant set [1, 2]. Whenever possible, external spike-in controls are preferred [3], as they control for more sources of variation by adding a presumably invariant set of data across samples. However, not all data sets contain external spike-ins and they cannot be added *post-facto*. Consequently, a variety of internal normalization methods have been developed [3, 4] which assume some internal feature of the data—typically a set of genes—is invariant between data sets. While most of these techniques were developed for microarrays or RNA-seq, they have been broadly applied to a variety of sequencing assays.

One set of protocols in particular—nascent RNA sequencing methods—are prone to large amounts of technical variation [5]. Nascent RNA sequencing protocols, such as global run-on sequencing (GRO-seq) [6], precision run-on sequencing (PRO-seq) [7] and their variations [8, 9], isolate small quantities of recently produced RNAs from actively engaged RNA polymerases [10]. Nascent RNA sequencing samples have a distinct profile relative to RNA-seq (Fig. 1A), resulting from the different phases of the RNA life cycle that they capture. RNA-seq samples from the pool of stable, messenger RNAs (mRNAs) which are predominantly spliced and polyadenylated. These RNAs originate from a relatively small fraction of the genome (exons and UTRs). In contrast, nascent RNA sequencing protocols capture RNA that is still actively engaged with RNA polymerases, meaning the RNAs are pre-splicing and need not be stable. As much of the genome is actively transcribed, nascent transcription protocols recover reads from much larger proportion of the genome (not only exons and introns, but also numerous intergenic regions). Consequently, if both assays are sequenced to the same depth, the equivalent nascent transcription data would have a lower per position depth.

External spike-ins in nascent RNA sequencing are also inherently different than in RNA-seq, leading to more uncertainty in the normalization process (Fig. 1A). The gold standard for spike-ins in RNA-seq is an External RNA Controls Consortium (ERCC)

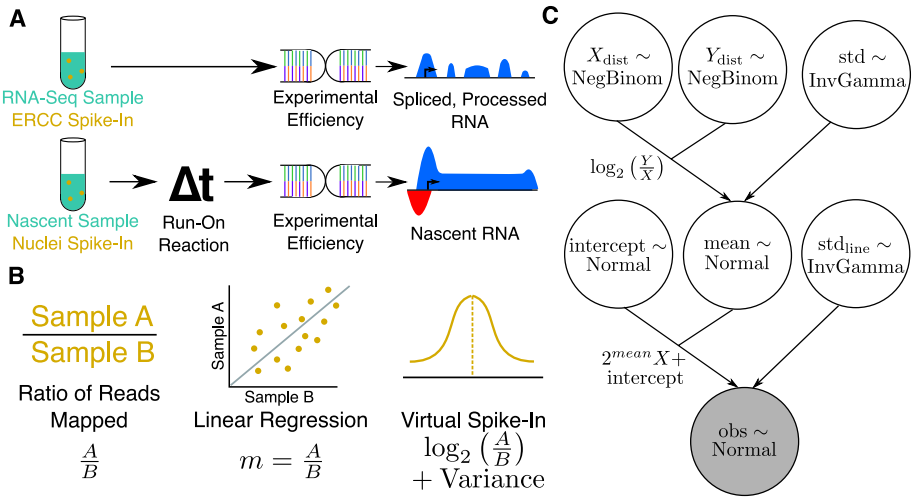

Fig. 1 A Bayesian model describing normalization data for nascent RNA sequencing data. **A** Schematic showing typical external control, handling, and resulting data profile differences between RNA-seq (top) and run-on nascent RNA sequencing assays (bottom). Note that run-on efficiency is assumed to be equivalent between spike-in nuclei and experimental nuclei. **B** Quantifying a normalization factor is accomplished either by a naive ratio of total reads approach (left), linear regression (middle), or by the Bayesian model proposed here (right). Linear regression (middle) is more resistant to noise and outliers, but does not provide a reliable way to measure the variance of the normalization estimate. The Bayesian model (right) converts the slope $m = \frac{A}{B}$ to log space, converting the multiplicative nature of the normalization factor to a linear one, for which normalization factors can be readily inferred as a normal distribution with variance. **C** A plate diagram showing the VSI model as implemented in pymc3. Briefly, we estimate our count distributions *X* and *Y* (top row) with a negative binomial. The ratio of two negative binomial distributions is approximately log-normal, so we derive a normal distribution called *mean* (middle) as the log of the ratio of *Y* and *X* with some variance (top right), estimated as an inverse gamma distributed random variable. With the estimation of the mean established, we then add additional parameters to describe the intercept, and variance of the actual line of best fit. This is done so that the parameter *mean* is estimating an error in log-transformed space, as discussed in Panel (**B**)

library, which uses a fixed amount of known RNAs which are added to the sample to quantify the variation introduced during sample handling, library preparation and sequencing. Crucially, this RNA spike-in library is introduced in known quantities prior to the experiment. Run-on centric nascent RNA protocols seek to identify the locations of actively engaged RNA polymerases by using marked nucleotides and a run-on reaction. Hence the ERCC spike-ins, by virtue of being mature RNAs, are incompatible with the run-on reaction. Instead, fixed amounts of nuclei from an external organism are typically added to the sample nuclei and then the run-on reaction is employed on the combination of cell types. Thus, the quantity of RNA from the external spike-in is determined by not only the efficiency of the protocol and sequencing, but also the efficiency of the run-on reaction. A necessary but potentially flawed assumption, then, is that all of the run-on reactions have the same efficiency, allowing the reads mapping to the spiked-in nuclei to be treated with the same reliability as an ERCC spike-ins. If an external spike-in is not used, many off-the-shelf RNA-seq tools are used directly for internal normalization [2, 11–13].

Critical to the effectiveness of internal or external normalization are the assumptions about what remains invariant. Notably, when run-on reactions are performed in the presence of a perturbation, nascent RNA sequencing contains a unique internal set of invariant data. RNA Polymerase II loads at the 5′ end of a gene and then proceeds through the gene with a relatively consistent processivity [10]. Thus, as first described by Mahat [14], at short time points after a perturbation, changes in transcription are not expected to have reached the 3′ end of long genes. Prior work on 3′ end normalization applied linear regression to the set of 3′ invariant ends and showed this approach was similar to other, presumably invariant, internal gene sets [14, 15]. However, they did not directly compare the approach to external spike-in controls or establish uncertainty bounds on their estimates.

In this work, we set out to compare run-on based 3′ normalization to external spike-ins. To this end, we developed a method for quantifying error in the estimation of spike-in normalization. Using this method we compare external spike-ins to internal invariant sets, focusing on the 3′ subset. We uncovered that most external spike-ins in nascent RNA assays are under-sequenced and potentially unreliable. Additionally, we find that when external spike-ins are of adequate depth and the assumptions of the 3′ normalization approach are met, the two methods show high correspondence.

Results

An algorithm to quantify error in spike-in normalization estimates

When normalizing between samples, there are different approaches to computing normalization factors from the invariant set, whether that set is an external spike-ins or internal [3] (Fig. 1B). The most naive of these is to take the simple ratio of reads mapping to the invariant set between two samples and use that as a normalization factor (Fig. 1B, left). However, this reduces the information contained within the set to a single summary value. The alternative approach is linear regression, where estimates of counts per invariant entity, typically genes, are used as data points for the fitting algorithm and the resulting slope is used as an the normalization factor [3] (Fig. 1B, middle). In this way, transcription levels across different orders of magnitude

can be leveraged to give a more accurate normalization factor. Thus, prior work in nascent transcription has often used naive linear regression to estimate normalization factors instead of a simple point estimate [14, 15]. However, to use linear regression, a sample's spike-ins must be of sufficient depth that a linear relationship exists in the count data. Additionally, naive linear regression does not provide error bounds.

To quantify the error inherent in estimating a normalization factor from data, we developed a hierarchical Bayesian version of the linear regression framework (Fig. 1B, right). Typically linear regression is formulated as:

$$y \sim mx + b$$

which describes the relationship between counts in two samples x and y in terms of two variables (m and b) the slope and intercept, respectively. In this framework, the slope (m) is interpreted as the best normalization factor between the two samples. In the naive context of normalizing to a spike-in (without considering the error of the estimate), this typically works well, as counts span multiple orders of magnitude and typically form a linear relationship between samples [3]. However, in standard linear regression only a single point estimate for the parameters is obtained.

To quantify the error in the estimated normalization factor, we extend the naive linear model above to incorporate an estimation of the error in log-space, backed by biologically informed count distributions. In the simplest terms, we generate a linear model whose mean is a normal distribution defined by the log-transformed ratio of our read counts [16, 17], plus an intercept term. By using the log-transformed ratio of read counts, we can assume the slope is normally distributed:

$$\mu_{\text{slope}} \sim \text{Normal}\left(\text{mean} = \log_2 \frac{Y}{X}, \sigma_{\text{mean}}\right)$$

Where μ_{slope} is the desired mean (normalization factor) and the resulting variance estimate σ_{mean} is then used as an estimate on the error of that normalization factor.

Our model is shown formally as a plate diagram in Fig. 1C. To fully specify the model, we assume the intercept follows a Normal distribution, Intercept $\sim$ Normal. The input data for this model is formally a counts matrix, M where $M_{i,j}$ represents the number of reads in sample i in region j. For all samples M_i, we select a single reference sample M_r to normalize against. We first model the count data over regions of interest as a Negative Binomial Distribution, as we expect the count distribution to be over-dispersed. This yields two variables—X and Y which describe to the count distribution of each sample input to the model. Priors for σ variables are selected to be uninformative using the conjugate InvGamma(1, 1) [18], while priors for X and Y are defined as $\mu_X = \text{mean}(M_i + 1)$ and $\mu_Y = \text{mean}(M_r + 1)$ to reflect the log-transformed ratio of Laplace smoothed count data.

We call our new method Virtual Spike-In (VSI) and leverage Markov Chain Monte Carlo (MCMC) methods to fit the underlying distributions. The input to the model is a set of data points between two samples, thus this model can also be applied to both external spike-ins and internal invariant sets of regions, such as the unperturbed 3′ end of long genes, or to any other set of invariant regions shared between two

samples that behave as count data. A technical discussion of implementation details for this model is available in the "Methods" section of this paper.

Confidence in normalization factor estimates depends on adequate spike-in depth

To assess the correctness of our VSI implementation and approach, we first compare the method to the standard linear regression approach. To this end, we processed samples of human cells with Drosophila spike-ins from a number of previously published studies employing nascent RNA sequencing data [19–32]. After filtering for samples with replicates and a nonzero number of reads mapping to the dm6 Drosophila genome, we were left with $n = 180$ samples (Additional file 1: Table S1, see Methods for complete details on data processing).

When running the VSI model on external spike-ins from published data [19–32], we find that it reliably recapitulates the results of naive linear regression (Fig. 2A), but now provides error bars on these estimates. In the regime of small normalization factors (values near zero), both linear regression and the VSI model perform essentially identically. Importantly, when the absolute value of linear regression estimates are large, the VSI approach tends to recover a comparatively lower normalization factor, likely a consequence of the model being more resistant to noise and extreme values than linear

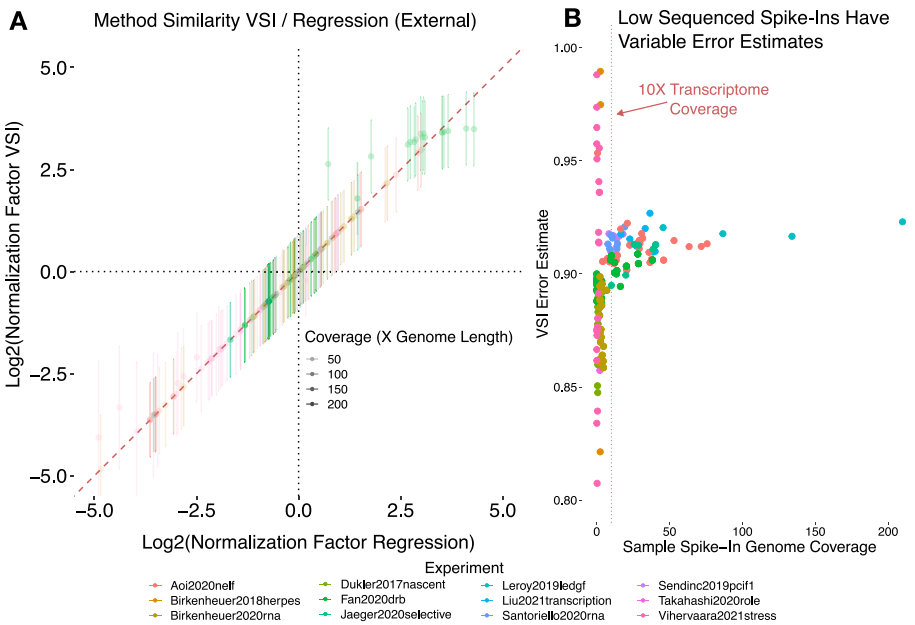

Fig. 2 Spike-ins have unusual behavior at the extremes. To assess where our model diverges in behavior from linear regression, we ran the VSI model on data from a number of published experiments [19–32]. Within each experiment, samples were grouped by condition and analyzed within those groups. All samples had Drosophila spike-ins, so annotated Drosophila genes were selected as the invariant set to count over. **A** Comparison of regression factors inferred by linear regression (x-axis) to those inferred by the Bayesian VSI model (y-axis). Estimates are shown along with an error bound of $\pm\sigma$. Notably, the regression estimate (x-axis) and VSI estimate (y-axis) deviate most dramatically when the absolute value of the normalization factor is large. **B** When we plot the depth of coverage of the spike-in (x-axis) against the VSI error estimate (y-axis) shows samples with less than 10× spike-in transcriptome coverage are less consistent than those above this threshold (dotted red line). Of note, error estimates range between 0.8 and 1.0, but when applied to the data they must be converted out of log2 space and multiplied by the normalization factor. Hence the impact of the error will scale with the normalization factor size. In a biological context, this is good—samples with large normalization factors have less confidence indicating poorer experimental efficiency and reproducibility

regression alone. However, large normalization factors suggest extreme differences in sample efficiencies which should call into question whether the data and spike-in are of sufficient depth and quality to be trusted. A detailed examination of the posterior distribution variance shows higher variability at low spike-in sequencing depth (Fig. 2B). The posterior variance (the variance of the estimated normalization factor after fitting the model) generally improves at depths greater than 10X the dm6 reference transcriptome, using a Drosophila transcriptome length of 30Mb [33]. Unfortunately, the majority of published samples are below this spike-in depth (Additional file 1: Fig. S1). This suggests that most published nascent RNA sequencing experiments using external spike-ins are under-sequenced, which may be a consequence of either an ineffective run-on reaction or a choice to prioritize sample read depth over spike-in read depth.

Evaluation of error in external and internal normalization

Normalization across invariant regions need not be limited to a spike-in, although an external spike-in is typically preferred. In theory, any set of invariant regions in a sequencing data set that follow a count distribution can be used to estimate a normalization factor between samples. This makes the Virtual Spike-In a versatile and widely useful model for quantifying normalization error across invariant regions.

As an example, our model can leverage reads at the 3′ end of long annotated genes, building on prior work [14, 15] (Fig. 3A). Nascent RNA assays survey engaged RNA polymerases genome-wide, which for any singular time point can be anywhere along the gene. However, in the presence of a perturbation, changes in transcription levels must originate at the 5′ end of genes, either by altering RNA polymerase II's loading and/or release from pausing. Once released, RNA polymerase II then proceeds through the gene at a relatively consistent rate [10, 15]. For example, in human cells RNA polymerase II has an elongation rate of roughly $2-3\frac{\text{kb}}{\text{min}}$ [34–38], although this rate can be highly variable. Therefore, at short time points, there is insufficient time to alter RNA polymerase II profiles at the 3′ ends of a long gene (see Fig. 3A).

Under this model, we note that RNA polymerase II profiles at genes past $\text{Length Threshold} = \text{Elongation Rate} \cdot \text{Time Point}$ should retain a consistent level of baseline transcription unperturbed by the experiment. Using this assumption, the invariant 3′ gene regions can be used for normalization between samples. Previous work [14, 15, 39], used a simple linear regression model to determine a normalization factor, defined by the slope of the best fit line, between the two samples using 3′ regions. However, these models did not establish uncertainty bounds on the accuracy of their normalization factors and did not compare their methodology to external biological spike-ins to quantify its effectiveness.

We leveraged the VSI approach to compare the 3′ normalization to external spike-in controls (Fig. 3B). For consistency of comparison between different experiments, and considering the typical timelines used, we selected a 180kb ($60\text{min} \cdot 3\frac{\text{kb}}{\text{min}}$) threshold for all samples when looking at the 3′ invariant region. We also exclude the last 500bps of the annotated gene from our normalization to reduce variance from the characteristic 3′ bump associated with termination in nascent RNA sequencing experiments. This results in 1198 3′ invariant regions used for normalization by the VSI model (roughly 10% of annotated RefSeq genes). Using this set, we found that

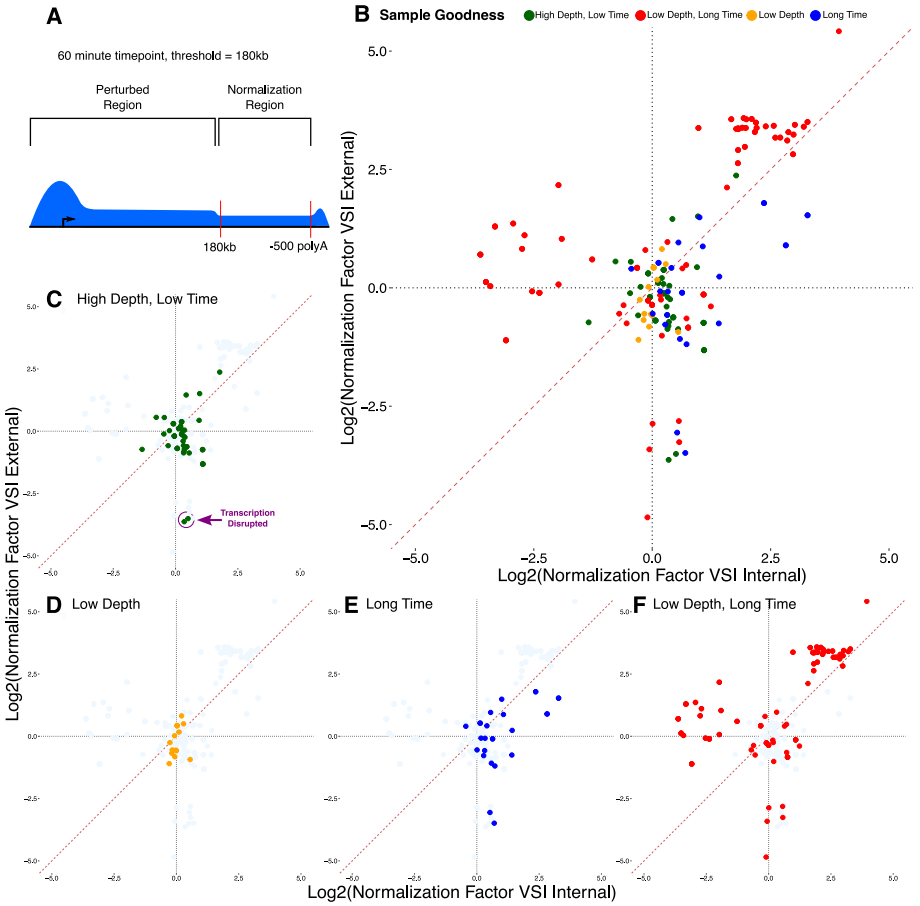

Fig. 3 3′ Normalization estimates depend on assumed polymerase elongation behavior and sequencing depth **A** A cartoon showing the characteristic shape of nascent RNA sequencing samples after a perturbation. RNA polymerase II loads at the 5 end of genes, thus after a perturbation alterations in transcription levels can only reach a distance that depends on the processivity of RNA polymerase II. In this work we assume 3kb/min and hence for a 60 min experiment the perturbation influences the first 180kb ($60\text{min} \cdot 3\frac{\text{kb}}{\text{min}}$). **B** We compared external spike-ins (y-axis) to 3 internal normalization across a large collection of previously published data. Samples are colored by whether they **C** meet both time point and depth assumptions (green), **D** have low sequencing depth (< 10X spike-in transcriptome) (orange), **E** have time points beyond the 3′ assumed 60 min (blue), or **F** meet neither assumption, being of both low spike-in depth and long time point (red). Notably, two samples in (circled in **C**) meet the coverage and time constraints of the 3′ normalization approach but involve depletion of NELF under heat shock conditions, which likely alters RNA polymerase elongation

the correspondence between the 3′ normalization approach and external spike-ins (Fig. 3B) showed extensive variation. In fact, the internal and external normalization factors were only rarely the same (diagonal line). Thus, we next sought to determine which factors influence the 3′ normalization method's fidelity.

We first consider time points below the 60 min threshold utilized. As the posterior estimate of the normalization factor varies dramatically below 10X spike-in coverage (Fig. 2B), we first consider only samples with stable estimates (spike-in coverage > 10X). For these samples, there is generally good concordance—small differences as most points are near the origin—between the 3′ normalization and external spike-in approach (Fig. 3C). Notably, two data sets show strikingly lower concordance between

the two methods. These two data sets were samples where NELF (negative elongation factor) was depleted and the cells were subjected to heat shock [19]. The lack of concordance between the methods suggests that the depletion of NELF may have had genome-wide effects on RNA polymerase, a condition that calls into question the invariant nature of any internal set.

At low external spike-in depth, inadequate spike-in data may exist for confidence in linear regression. Consistent with this notion, low depth spike-in samples have higher posterior estimate variance (Fig. 2B). However, despite this increased uncertainty, we found good concordance between the spike-in and the 3′ normalization estimates (Fig. 3D).

Importantly, the 3′ normalization approach inherently assumes that portions of genes are unreachable at the specified time point of the experiment. By using a uniform 60 min assumption, we could determine whether the concordance between the 3′ approach and external spike-ins breaks down at longer time points, when the assumed invariant regions can no longer be assured to be unchanged. As expected, when the internal set contains regions that could be varying between the samples (e.g. the time point is longer than the 60 min assumption), there was increasing discordance between the two normalization methods (Fig. 3D,E), particularly when long time points co-occurred with low coverage (Fig. 3F). Intriguingly, even in the data that fail to meet our assumptions (low depth + long time, Fig. 3F) we observe a small cluster of samples close to the origin of the plot. In these scenarios, we achieve concordance between internal and external spike-ins even when all assumptions are violated, as in these cases the perturbation happens to not strongly impact the long gene set used by the VSI normalization.

Collectively, these results suggest that the 3′ internal normalization approach gives results similar to the linear approximation of external spike-ins when the assumptions of the model are met. This is particularly true when the normalization factors are small (e.g. near the origin in Fig. 3B–F). When the assumptions of the VSI model are violated, either with long time points or disruptions that alter RNA polymerase itself, the two models strongly disagree.

To further characterize this pattern, we next turned our attention to the examination of a single high quality data set that contains multiple time points and roughly average spike-in sequencing depth (GSE96869) [23]. In this study, Dukler et al. treated K562 cells with the natural drug Celastrol, which activates mammalian heat shock response [23]. Cells were then assayed at several time points including 10 min, 20 min, 40 min, 60 min and 160 min. This PRO-seq data set has spike-in sample depth ranging from 0.7 to 1.1X Drosophila transcriptome coverage. Importantly, the cells undergo replicative arrest around the 40 min time point. As before, we employ a 180kb ($60\text{min} \cdot 3\frac{\text{kb}}{\text{min}}$) threshold for all samples when looking at the 3′ invariant region. For each sample, we compared normalization results using the 3′ internal normalization to external spike-ins, using both linear regression (VSI) and the ratio based point estimate.

We observe that the VSI model shows good concordance between internal (3′) and external spike-in estimates of the normalization factor, particularly at early time points (Fig. 4). After the onset of replicative arrest (t=40 min), the internal and external normalization factors begin to diverge, though only modestly in both the 40 min and one of the 60 min time point replicates. As expected, the largest deviations between the 3′

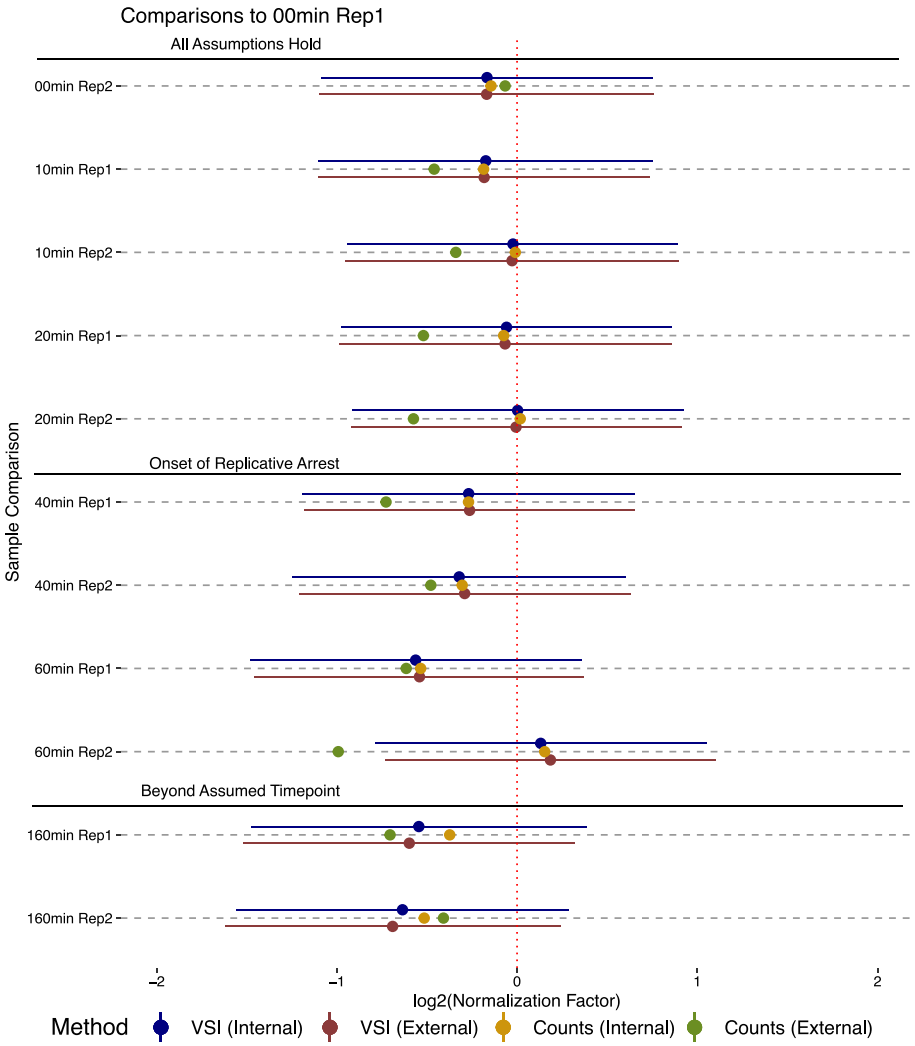

Fig. 4 Comparison of all normalization methods on good quality data. We compare normalization factors on a high quality data set [23] (GSE96869) computed by four distinct methods: VSI applied to the internal 3 invariant gene set (blue), VSI applied to an external Drosophila spike-in (red), the ratio approach applied to the 3 invariant gene set (yellow), and the ratio approach applied to the Drosophilia spike-in (green). Error bars are shown for the VSI estimates. The 3 invariant set uses a threshold of 180 kb (60 min), regardless of the data time point. For orientation, we note the normalization factor of zero (red dotted line), the onset of biological replication arrest and the assumed time point for the 3 invariant gene set

and external spike-in are observed at 160 min, when the time point is well beyond the 60 min assumed by the internal normalization. At all time points, the single point estimate of the external spike-in deviates substantially from both the linear model estimate of external spike-in and the 3′ approach, consistent with prior work on normalization approaches [3].

Downstream effects of normalization

Normalization factors are crucially important in downstream analyses of high throughput sequencing data. To that end, we next compared the results of differential expression analysis on the Dukler data set [23]. For differential expression analysis, we used DESeq2 [2], which uses an internal normalization approach. Specifically, DESeq2 calculates a

size factor as the median ratio of counts over every gene in the sample divided by the geometric mean of counts at that gene over all samples. The result is an effective method for normalization that implicitly assumes that most genes are unchanged across the comparison.

We sought to compare the default DESeq2 size factor approach to the 3′ internal normalization method. For this comparison, we performed differential expression analysis between the 0 min and 60 min time points (Fig. 5A, 40 min comparison shown in Additional file 1: Fig. S3). We observed that the posterior point estimate for the normalization factor recapitulate a strict subset of genes called as differentially expressed by the automatically estimated size factors (Fig. 5). In simpler terms, it appears that 3′ internal

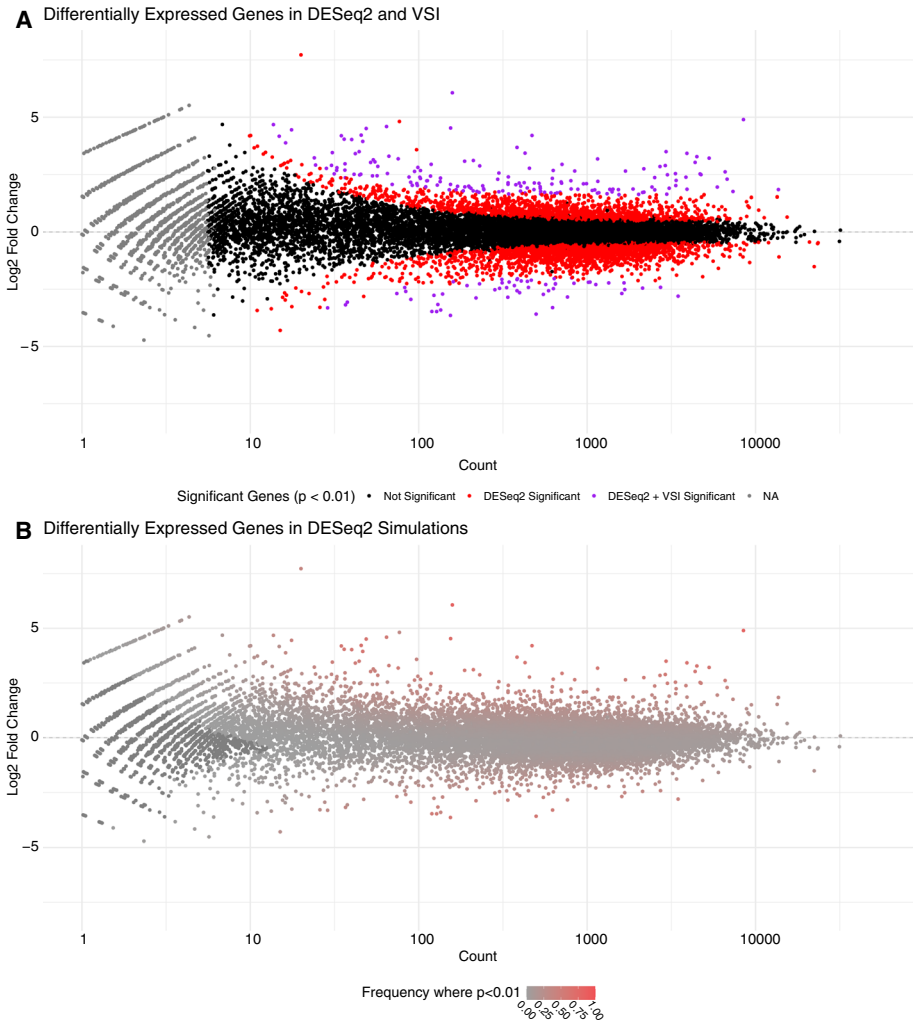

Fig. 5 Estimated Normalization Factors Provide Strict Cutoffs for DESeq2 **A** Differential expression analysis by DESeq2 (adj. p-val < 0.01) using size factors estimated from DESeq2 (red) and the VSI model (purple) on 3′ invariant regions. Note that DESeq2 calls normalization factors "size factors". The more conservative VSI identified set (purple) is a strict subset of the DESeq2 identified significant set. **B** Consistency of differential expression calls across a broad range of plausible normalization factors. Genes are colored based on the reproducibility of statistically significant differential expression (DESeq2, adj. *p* val < 0.01) across 1000 iterations where normalization factors were sampled from the posterior distribution estimated by VSI. Points that appear as significant most often are also those that are called as significant using both DESeq2 size factors and VSI 3 normalization (Panel **A**, purple)

estimated normalization factors are more conservative, effectively decreasing the set of genes called as significant. Arguably the VSI set is both more conservative and based on a biologically principled invariant set of data compared to the DESeq2 method.

In both cases, a single normalization term is calculated and presumed to be correct. Our earlier comparison to external spike-ins (Fig. 3) suggests two estimators may reach similar but not quite the same normalization factor. Therefore, we next sought to ascertain the extent to which minor, plausible fluctuations in the calculated normalization factor might influence differential expression analysis. To this end, we use a sampling approach. We ran 1000 simulations sampling normalization factors from the posterior distribution estimated by VSI for each of the 4 samples (10 min, 60 min; 2 replicates at each time point). We then ask how often a particular gene is called as significant across the samples. We observe that many of the genes called by DESeq2 as differentially expressed (red dots in Fig. 5A) have relatively low reproduciblity across the range of plausible normalization factors (Fig. 5B) and are therefore potential false positives. Notably, the genes with the highest reproducibility are those found by the VSI 3′ point estimate (purple dots in Fig. 5A correspond to red dots in Fig. 5B).

Discussion

We present Virtual Spike-In, a novel approach that uses a hierarchical Bayesian regression model to calculate normalization factors and quantify their uncertainty for nascent transcription datasets. We use this method to compare 3′ end normalization in run-on based nascent RNA sequencing experiments to external spike-in controls. We find that while the internal and external normalization rarely perfectly agree, the 3′ end normalization shows high concordance to external spike-in controls when assumptions of the method are met. Additionally, normalization is known to have strong effects on analysis results [3], and our work further supports this conclusion (Additional file 2).

While external spike-ins are typically assumed to be the gold standard for normalization of sequencing samples, we find that external spike-ins in published run-on based nascent RNA sequencing experiments are typically under sequenced. Importantly, external spike-ins in nascent RNA sequencing are not the same as those in RNA-seq. This makes the entire normalization process significantly more challenging. Using spike-in nuclei inherently assumes that for every sample, the efficiency of the run-on in the spike-in nuclei closely matches the efficiency of the run-on in the experimental nuclei. There is no reliable mechanism to determine if this assumption is correct. This problem is exacerbated by the relatively low read depth of most external spike-ins in nascent RNA assays. It is critically important that any normalization technique be based on adequate data, as even the best normalization model is limited by the available data.

The alternative to external spike-ins is to use an internal invariant set. Run-on based nascent transcription coupled to a perturbation has a unique invariant set in the 3′ ends. While 3′ end normalization is powerful, it has a number of important limitations compared to an external spike-in. First, the elongation rate of RNA polymerase II in the organism must also be known. At any given elongation rate and time point, a reasonable proportion of genes in the genome must be sufficiently long that invariant regions exist at the time point of interest. While this works well in the human genome, it is likely not the case for organisms with smaller genes and genomes. Even in the human genome,

when the normalization factor is estimated on later time points, it is based on increasingly smaller quantities of data, leading to less certainty. With that said, the use of a Bayesian model in this context does make the model robust to a small number of genes to be normalized against. Finally, the 3′ end approach cannot be used in the absence of a perturbation or if the perturbation could alter previously loaded RNA polymerase.

In addition to the assumptions made about the model, it is also important to consider the assumptions made about the selected 3′ regions if performing normalization internally. First and foremost, low expression is a persistent concern across all experiments and must be considered here. Undersequencing is, in general, a problem for normalization (of external spike-ins or of 3′ invariant regions) and downstream analysis. Consequently, if a sample is of low sequencing depth, either generally or particularly at the 3′ end of genes, we recommend it be excluded from further analysis for quality concerns. Likewise, our 3′ assessment depends on the accuracy of gene annotations and the presumption that long genes are not in some way atypical. Finally, the presence of intronic bidirectional signals (e.g. same strand overlapping transcription) could be problematic if the bidirectionals both reside within the invariant 3′ region and are themselves differentially transcribed. Despite these caveats, one benefit of 3′ end normalization is that it can be applied to many previously published run-on based nascent RNA sequencing data sets where an external spike-in is not present.

There are a number of nascent transcription assays that do not use a run-on step, and normalization for these assays present distinct challenges. Metabolic labeling approaches expose live cells to marked nucleotides over some time frame before the experiment [8, 40]. As such, both the profile and signal to noise characteristics of the data are influenced by the time and efficiency of the labeling process. In contrast, mammalian native elongating transcript sequencing (mNET-seq) [41] uses an antibody to pull down a component of the RNA polymerase II complex. As such, normalization of mNET-seq data is conceptually similar to ChIP-seq and should account for antibody efficiency. Further work is needed to characterize both internal and external normalization strategies for metabolic labeling and antibody oriented nascent transcription assays.

The Virtual Spike-In model is versatile. As the input to normalization is counts over a collection of regions, the VSI method can be applied to both internal invariant sets, such as the 3′ end normalization used here, and to external spike-in controls. Another notable advantage to the VSI technique is that it establishes error bounds on the calculated normalization factors, an important but often overlooked aspect of the data analysis. Effectively quantifying error in the point estimations of normalization factors is an important addition over the naive linear model. Quantification of error is essential to analyzing nascent RNA sequencing data rigorously. Ultimately, nascent RNA sequencing experiments appear to need a more reliable mechanism for external normalization, which is challenging given the limitations of the underlying protocols.

Methods

Our model is implemented in the Python programming language using the pymc3 MCMC library [42]. Inference is performed using an adaptive sampler, combining the No-U-Turn Sampler [43] (NUTS) for continuous variables with a Metropolis-Hastings Sampler [44, 45] for discrete variables, using 25,000 iterations after a burn-in period

of 2,500 samples. The number of iterations can be increased if a greater assurance of convergence is desired. A larger number of iterations are required for convergence of the discrete distribution due to the use of a Metropolis sampler instead of NUTS (Additional file 1: Fig. S5). Source code is available at https://github.com/Dowell-Lab/virtual_spike_in.

For both the human cell lines and *Drosophila* spike-in, reads were mapped to the hg38 and dm6 reference genomes respectively using the Nascent-Flow pipeline [46]. Counts were determined for all genes using featureCounts [47], considering only the maximally expressed isoform and counting reads per gene including exons and introns.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05607-3.

Additional file 1. This file contains supplementary material (additional description of the analysis pipelines) and supplemental figures.

Additional file 2. Contains a comma delimited table of the samples and papers that were used for analysis in this study, along with metadata on cell type and whether all analysis steps could be successfully completed on those samples.

Acknowledgements
Dr. Nina Ripin shared data that was used for initial development of the model. Dr. Dylan Taatjes proposed the name "Virtual Spike-In". Dr. Lynn Sanford suggested published datasets for model testing.

Author Contributions
ZLM developed the model, conducted the analysis and drafted the initial manuscript. RDD supervised the work. Both authors revised the manuscript.

Funding
This work was funded by the National Science Foundation under Grants ABI1759949 and HDR2022138 and the National Institutes of Health Grant GM125871 and AI156739.

Availability of data and materials
Data for this project was from previously published experiments, with additional metadata available in Additional file 1: Table S1. The datasets analysed in the current study are available in the GEO repository, with the accession numbers GSE144786, GSE143844, GSE106126, GSE130342, GSE96869, GSE141377, GSE139468, GSE117155, GSE150530, GSE104334, GSE128086, GSE122803, GSE121024, GSE127844, and GSE154746.

Declarations

Ethics approval and consent to participate
Not applicable to this study.

Consent for publication
Not applicable to this study.

Competing interests
Zachary Maas declares that he has no competing interests. Robin Dowell is a founder of Arpeggio Bioscience who uses nascent RNA sequencing protocols routinely.

Received: 21 March 2023 Accepted: 7 December 2023
Published online: 12 January 2024

References

1. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21(9):1543–51. https://doi.org/10.1101/gr.121095.111.
2. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
3. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. Mol Cell Biol. 2016;36(5):662–7. https://doi.org/10.1128/MCB.00970-14.
4. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2017;19(5):776–92. https://doi.org/10.1093/bib/bbx008.

5. Hunter S, Sigauke RF, Stanley JT, Allen MA, Dowell RD. Protocol variations in run-on transcription dataset preparation produce detectable signatures in sequencing libraries. BMC Genomics. 2022;23(1):187. https://doi.org/10.1186/s12864-022-08352-8.
6. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322(5909):1845–8. https://doi.org/10.1126/science.1162228.
7. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016;11(8):1455. https://doi.org/10.1038/nprot.2016.086.
8. Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P. TT-seq maps the human transient transcriptome. Science. 2016;352(6290):1225–8. https://doi.org/10.1126/science.aad9841.
9. Wissink EM, Vihervaara A, Tippens ND, Lis JT. Nascent RNA analyses: tracking transcription and its regulation. Nat Rev Genet. 2019;20(12):705–23. https://doi.org/10.1038/s41576-019-0159-6.
10. Cardiello JF, Sanchez GJ, Allen MA, Dowell RD. Lessons from eRNAs: understanding transcriptional regulation through the lens of nascent RNAs. Transcription. 2020;11(1):3–18. https://doi.org/10.1080/21541264.2019.1704128.
11. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):47. https://doi.org/10.1093/nar/gkv007.
12. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3. https://doi.org/10.1093/bioinformatics/bts034.
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27. https://doi.org/10.1093/biostatistics/kxj037.
14. Mahat DB, Salamanca HH, Duarte FM, Danko CG, Lis JT. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. Mol Cell. 2016;62(1):63–78. https://doi.org/10.1016/j.molcel.2016.02.025.
15. Vihervaara A, Mahat DB, Guertin MJ, Chu T, Danko CG, Lis JT, Sistonen L. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat Commun. 2017;8(1):255. https://doi.org/10.1038/s41467-017-00151-0.
16. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013;14(2):232–43. https://doi.org/10.1093/biostatistics/kxs033.
17. Choi Y, Coram M, Peng J, Tang H. A Poisson log-normal model for constructing gene covariation network using RNA-seq data. J Comput Biol. 2017;24(7):721–31. https://doi.org/10.1089/cmb.2017.0053.
18. Gelman A. Bayesian data analysis. 3rd ed. Boca Raton: Chapman & Hall/CRC Texts in Statistical Science. CRC Press; 2014.
19. Aoi Y, Smith ER, Shah AP, Rendleman EJ, Marshall SA, Woodfin AR, Chen FX, Shiekhattar R, Shilatifard A. NELF regulates a promoter-proximal step distinct from RNA Pol II pause-release. Mol Cell. 2020;78(2):261–2745. https://doi.org/10.1016/j.molcel.2020.02.014.
20. Barbieri E, Hill C, Quesnel-Vallières M, Zucco AJ, Barash Y, Gardini A. Rapid and scalable profiling of nascent RNA with fastGRO. Cell Rep. 2020;33(6): 108373. https://doi.org/10.1016/j.celrep.2020.108373.
21. Birkenheuer CH, Danko CG, Baines JD. Herpes simplex virus 1 dramatically alters loading and positioning of RNA polymerase II on host genes early in infection. J Virol. 2018;92(8):10–1128. https://doi.org/10.1128/JVI.02184-17.
22. Birkenheuer CH, Baines JD. RNA polymerase II promoter-proximal pausing and release to elongation are key steps regulating herpes simplex virus 1 transcription. J Virol. 2020;94(5):10–1128. https://doi.org/10.1128/JVI.02035-19.
23. Dukler N, Booth GT, Huang Y-F, Tippens N, Waters CT, Danko CG, Lis JT, Siepel A. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Res. 2017. https://doi.org/10.1101/gr.222935.117.
24. Fan Z, Devlin JR, Hogg SJ, Doyle MA, Harrison PF, Todorovski I, Cluse LA, Knight DA, Sandow JJ, Gregory G, Fox A, Beilharz TH, Kwiatkowski N, Scott NE, Vidakovic AT, Kelly GP, Svejstrup JQ, Geyer M, Gray NS, Vervoort SJ, Johnstone RW. CDK13 cooperates with CDK12 to control global RNA polymerase II processivity. Sci Adv. 2020;6(18):5041. https://doi.org/10.1126/sciadv.aaz5041.
25. Jaeger MG, Schwalb B, Mackowiak SD, Velychko T, Hanzl A, Imrichova H, Brand M, Agerer B, Chorn S, Nabet B, Ferguson FM, Müller AC, Bergthaler A, Gray NS, Bradner JE, Bock C, Hnisz D, Cramer P, Winter GE. Selective mediator dependence of cell-type-specifying transcription. Nat Genet. 2020;52(7):719–27. https://doi.org/10.1038/s41588-020-0635-0.
26. LeRoy G, Oksuz O, Descostes N, Aoi Y, Ganai RA, Kara HO, Yu J-R, Lee C-H, Stafford J, Shilatifard A, Reinberg D. LEDGF and HDGF2 relieve the nucleosome-induced barrier to transcription in differentiated cells. Sci Adv. 2019;5(10):3068. https://doi.org/10.1126/sciadv.aay3068.
27. Liu N, Xu S, Yao Q, Zhu Q, Kai Y, Hsu JY, Sakon P, Pinello L, Yuan G-C, Bauer DE, Orkin SH. Author Correction: Transcription factor competition at the γ-globin promoters controls hemoglobin switching. Nat Genet. 2021;53(4):586. https://doi.org/10.1038/s41588-021-00834-x.
28. Rao SSP, Huang S-C, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon K-R, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, Huang X, Shamim MS, Shin J, Turner D, Ye Z, Omer AD, Robinson JT, Schlick T, Bernstein BE, Casellas R, Lander ES, Aiden EL. Cohesin loss eliminates all loop domains. Cell. 2017;171(2):305–32024. https://doi.org/10.1016/j.cell.2017.09.026.
29. Santoriello C, Sporrij A, Yang S, Flynn RA, Henriques T, Dorjsuren B, Custo Greig E, McCall W, Stanhope ME, Fazio M, Superdock M, Lichtig A, Adatto I, Abraham BJ, Kalocsay M, Jurynec M, Zhou Y, Adelman K, Calo E, Zon LI. RNA helicase DDX21 mediates nucleotide stress responses in neural crest and melanoma cells. Nat Cell Biol. 2020;22(4):372–9. https://doi.org/10.1038/s41556-020-0493-0.
30. Sendinc E, Valle-Garcia D, Dhall A, Chen H, Henriques T, Navarrete-Perea J, Sheng W, Gygi SP, Adelman K, Shi Y. PCIF1 catalyzes m6Am mRNA methylation to regulate gene expression. Mol Cell. 2019;75(3):620–6309. https://doi.org/10.1016/j.molcel.2019.05.030.

31. Takahashi H, Ranjan A, Chen S, Suzuki H, Shibata M, Hirose T, Hirose H, Sasaki K, Abe R, Chen K, He Y, Zhang Y, Takigawa I, Tsukiyama T, Watanabe M, Fujii S, Iida M, Yamamoto J, Yamaguchi Y, Suzuki Y, Matsumoto M, Nakayama KI, Washburn MP, Saraf A, Florens L, Sato S, Tomomori-Sato C, Conaway RC, Conaway JW, Hatakeyama S. The role of mediator and little elongation complex in transcription termination. Nat Commun. 2020;11(1):1063. https://doi.org/10.1038/s41467-020-14849-1.
32. Vihervaara A, Mahat DB, Himanen SV, Blom MAH, Lis JT, Sistonen L. Stress-induced transcriptional memory accelerates promoter-proximal pause release and decelerates termination over mitotic divisions. Mol Cell. 2021;81(8):1715–17316. https://doi.org/10.1016/j.molcel.2021.03.007.
33. Daines B, Wang H, Wang L, Li Y, Han Y, Emmert D, Gelbart W, Wang X, Li W, Gibbs R, Chen R. The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. Genome Res. 2011;21(2):315–24. https://doi.org/10.1101/gr.107854.110.
34. Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife. 2014. https://doi.org/10.7554/eLife.02407.
35. Mimoso CA, Adelman K. U1 snRNP increases RNA Pol II elongation rate to enable synthesis of long genes. Mol Cell. 2023;83(8):1264–127910. https://doi.org/10.1016/j.molcel.2023.03.002.
36. Noe Gonzalez M, Blears D, Svejstrup JQ. Causes and consequences of RNA polymerase II stalling during transcript elongation. Nat Rev Mol Cell Biol. 2021;22(1):3–21. https://doi.org/10.1038/s41580-020-00308-8.
37. Fuchs G, Voichek Y, Benjamin S, Gilad S, Amit I, Oren M. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. 2014;15(5):69. https://doi.org/10.1186/gb-2014-15-5-r69.
38. Muniz L, Nicolas E, Trouche D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. EMBO J. 2021;40(15): 105740. https://doi.org/10.15252/embj.2020105740.
39. Fant CB, Levandowski CB, Gupta K, Maas ZL, Moir J, Rubin JD, Sawyer A, Esbin MN, Rimel JK, Luyties O, Marr MT, Berger I, Dowell RD, Taatjes DJ. TFIID enables RNA polymerase II promoter-proximal pausing. Mol Cell. 2020;78(4):785–7938. https://doi.org/10.1016/j.molcel.2020.03.008.
40. Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, Wlotzka W, von Haeseler A, Zuber J, Ameres SL. Thiol-linked alkylation of RNA to assess expression dynamics. Nat Methods. 2017;14(12):1198–204.
41. Nojima T, Gomes T, Carmo-Fonseca M, Proudfoot NJ. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. Nat Protoc. 2016;11(3):413–28.
42. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Comput Sci. 2016;2:55. https://doi.org/10.7717/peerj-cs.55.
43. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15(47):1593–623.
44. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21(6):1087–92. https://doi.org/10.1063/1.1699114.
45. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97–109. https://doi.org/10.2307/2334940.
46. Tripodi IJ, Gruca MA. Nascent-flow (2018). https://doi.org/10.17605/OSF.IO/NDHJ2
47. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.

Publisher's Note