Deconvolution of Nascent Sequencing Data Using Transcriptional Regulatory Elements

Zachary Maas

Department of Computer Science, BioFrontiers Institute, University of Colorado Boulder
596 UCB
Boulder, CO 80309, USA
E-mail: zachary.maas@colorado.edu

Rutendo Sigauke

BioFrontiers Institute, University of Colorado Boulder
596 UCB
Boulder, CO 80309, USA
E-mail: rutendo.sigauke@colorado.edu

Robin Dowell

Department of Molecular, Cellular, and Developmental Biology,
BioFrontiers Institute, Department of Computer Science, University of Colorado Boulder
596 UCB
Boulder, CO 80309, USA
E-mail: robin.dowell@colorado.edu

The problem of microdissection of heterogeneous tissue samples is of great interest for both fundamental biology and biomedical research. Until now, microdissection in the form of supervised deconvolution of mixed sequencing samples has been limited to assays measuring gene expression (RNA-seq) or chromatin accessibility (ATAC-seq). We present here the first attempt at solving the supervised deconvolution problem for run-on nascent sequencing data (GRO-seq and PRO-seq), a readout of active transcription. Then, we develop a novel filtering method suited to the mixed set of promoter and enhancer regions provided by nascent sequencing, and apply best-practice standards from the RNA-seq literature, using *in-silico* mixtures of cells. Using these methods, we find that enhancer RNAs are highly informative features for supervised deconvolution. In most cases, simple deconvolution methods perform better than more complex ones for solving the nascent deconvolution problem. Furthermore, undifferentiated cell types confound deconvolution of nascent sequencing data, likely as a consequence of transcriptional activity over the highly open chromatin regions of undifferentiated cell types. Our results suggest that while the problem of nascent deconvolution is generally tractable, stronger approaches integrating other sequencing protocols may be required to solve mixtures containing undifferentiated celltypes.

Keywords: Nascent Sequencing, Deconvolution

1. Introduction

One key problem of interest when studying transcription is the ability to capture the heterogeneity that exists in true biological samples.[1] Bulk sequencing samples from cells are an aggregate across a cellular population, and thus average out differences between individual cells to capture only an ensemble profile of a given sample. Notably in the case of samples taken from tissues composed of heterogeneous constituent cells, any celltype specific differences are not necessarily discernible in the heterogeneous mixture of expression data.

To some extent, this problem has been at least partially solved in the context of RNA-seq with the emergence of single cell RNA-seq protocols which allow for RNA content at the level of the individual cell to be measured.[2] However, the relatively high cost of sampling deeply limits the use of scRNA-seq in many contexts. Consequently, a great deal of work has been done to separate samples into constituent cell types *in silico*. This task is interchangeably referred to as deconvolution or microdissection. Deconvolution has been studied extensively in the context of both microarray data and in RNA-seq,[1,3–6] but has seen only limited application to other high throughput genomic data.

Nascent transcription protocols[7,8] are of particular interest for studies into transcriptional regulation.[9,10] Nascent sequencing protocols profile active RNA Polymerase II activity, which captures enhancer associated RNAs (eRNAs), short unstable transcripts that are often associated with transcription factor binding sites.[11] These eRNA transcript have proven to be highly informative markers of transcription factor activity.[9,10,12–16] Unfortunately RNA-seq, whether bulk or single cell, does not capture enhancer associated transcripts due to the fact they are unstable and not polyadenylated.[11] For this reason, the theoretical possibility of single cell measures of nascent transcription has tremendous potential for understanding regulation and transcription factor activity in key biological processes including development and disease progression.

Today, nascent sequencing protocols still operate only on the bulk level, largely because nascent protocols are relatively onerous, taking up to a week to process a set of samples.[7,8,17] Because nascent protocols capture RNA production, many of the signals arise from lowly abundant, highly unstable RNAs.[11] Furthermore, with current biochemical efficiencies, a single cell nascent sequencing protocol is likely infeasible, and thus deconvolution is needed to dissect nascent transcription profiles within tissues.

Nascent transcription data has relatively unique properties compared to RNA-seq. First, RNA-seq measures steady state mature, stable RNA levels which tend to be of relatively high abundance. In contrast, nascent sequencing protocols cover a much larger proportion of the genome ($\sim 40\%$ as opposed to $\sim 8\%$).[17] The consequence is that the average sequencing depth per transcript is typically lower in nascent data, in spite of often sequencing samples to a higher depth. Second, many transcripts measured in nascent protocols are unannotated, lowly transcribed, unstable eRNAs (Figure 1A).[11,17] In development, enhancer activities are the first changes detectable when a cell undergoes state change, suggesting their associated eRNAs have high potential as cell type markers.[18] Furthermore, enhancer associated RNAs tend to be more cell type specific than protein coding genes.[19] However, their low transcription levels lead to issues of reliable detection.[17] Thus methods developed for RNA-seq must be appropriately

adapted to use with nascent sequencing data.

Here, we use standardized methods for supervised deconvolution to nascent sequencing data, applying a newly developed filtering technique to solve problems presented by nascent data in the deconvolution context. We show that deconvolution of nascent sequencing data works reliably, albeit with different model performance than in RNA-seq. We find that eRNAs present an informative set of information for deconvolution that can be inferred without a reference annotation. Furthermore, we find that undifferentiated celltypes confound deconvolution of nascent sequencing data, likely because their transcriptional expression resembles that of an aggregate of different differentiated celltypes.

2. Results

The problem of supervised deconvolution with sequencing data is formulated as follows: *Given sequencing samples from homogenous cell types and a heterogenous sample made up of those cell types, can we estimate the mixing proportions of those constituent cell types?* The problem of supervised (or partial) deconvolution is typically formulated as a linear system (Equation 1).[5,20]

$$X = AS \tag{1}$$

Here, X is a single-row matrix with one column per region of interest (ROI) $(1 \times g)$, A is a single row matrix with one column per reference homogenous cell type $(1 \times s)$, and S is a matrix with one row per sample and one column per ROI $(s \times g)$. In most contexts, regions of interest (ROIs) correspond to annotated genes.

This is a overdetermined linear system, since the number of ROIs far exceeds the number of constituent cell types. Additionally, because these are biological values sampled from a noisy process, the key challenge is minimizing errors when solving the system. Most work in the literature has sought to solve the issues of this system in the context of RNA-seq or microarray[1,3–6,20–22] data, with limited applications of this approach to other kinds of sequencing data.

For RNA-seq, a large variety of tools and approaches have been developed,[1,3–5,5,6,20–22] which approach the problem using different models, constraints, and regularization approaches, as well as different ways to shrink the linear system. Many of these approaches claim to be the state-of-the-art, with most tools providing good performance. Consequently, we first examine the deconvolution problem on nascent sequencing using annotated genes and methods developed for RNA-seq.

2.1. *Deconvolution on annotated genes*

To evaluate existing deconvolution methods on nascent sequencing data, we first identified a number of high quality nascent sequencing data sets from a variety of cell types (see Table 2.1). Samples were processed using a standardized analysis pipeline[23] which includes quality control, mapping and bidirectional transcript identification. These bidirectional transcripts originate from both gene start sites and regulatory elements such as enhancers (Figure 1A). The non-gene associated bidirectionals are often referred to as enhancer associated RNAs, or eRNAs.

As a first test, we examined only annotated protein coding genes to mimic deconvolution analysis typically done in RNA-seq. Notably, nascent data differs from RNA-seq in that splicing information is not present in nascent sequencing experiments, as RNA is collected pre-splicing. Furthermore, consistent with standards in nascent transcription analysis,[17] we exclude the +300 initiation region of each gene when using featureCounts[24] to count reads (see Figure 1A), as this avoids the 5′ bidirectional peak.

To simulate a mixed sample, we generated 128 randomly mixed samples by subsampling reads from each reference sample. Samples used for all *in-silico* experiments in this paper were mixed proportionally from raw reads using samtools,[25] and are listed in Table 2.1. With these randomly mixed samples, we then performed supervised deconvolution using 4 different methods which are commonly discussed in the literature — Nonnegative-Least Squares Regression (NNLS), Ridge Regression, LASSO Regression, and ϵ-Support Vector Regression (SVR). For all methods tested, we apply a nonnegativity constraint (all mixing proportions must be at least zero) and a sum-to-one constraint (all mixing proportions must sum to one), as suggested in prior work.[1] These constraints serve to make results from various deconvolution procedures interpretable as mixing weights for the linear deconvolution system. Code and supplemental materials for this project are available at `https://github.com/Dowell-Lab/DeconvolutionNascent`. We find that these methods provide generally good accuracy on

Study	GEO Accession	SRR	Cell Type
Samples used in Figure 1–3			
Jiang 2018[26]	GSM3025555	SRR6789175	HCT116
Fei 2018[27]	GSM3100195	SRR7010982	HeLA
Andrysik 2017[28]	GSM2296635	SRR4090102	MCF7
Dukler 2017[29]	GSM2545324	SRR5364303	K562
Zhao 2016[30]	GSM2212033	SRR3713700	Kasumi-1
Danko 2018[31]	GSM3021718	SRR6780907	CD4+-T-cell
Chu 2018[32]	GSM3309955	SRR7616132	Jurkat-T-cell
Samples added for Figure 4			
Core 2014[33]	GSM1480326	SRR1552485	GM12878
Smith 2021[34]	GSM4214080	SRR10669536	ESC
Ikegami 2020[35]	GSM4207079	SRR10601203	BJ5ta

Table 2.1: Samples used in this study.

deconvolution on our 128 randomly generated mixtures, although it appears that regularized methods perform more poorly than naive NNLS (Figure 1B,C) in certain celltypes across these mixtures. In this context, it appears that regularization does not improve accuracy at the cost of significant computational slowdowns relative to NNLS. Given these promising initial results, we next sought to shift the focus away from annotated genes to the unannotated bidirectional transcripts present at both promoters and enhancers.

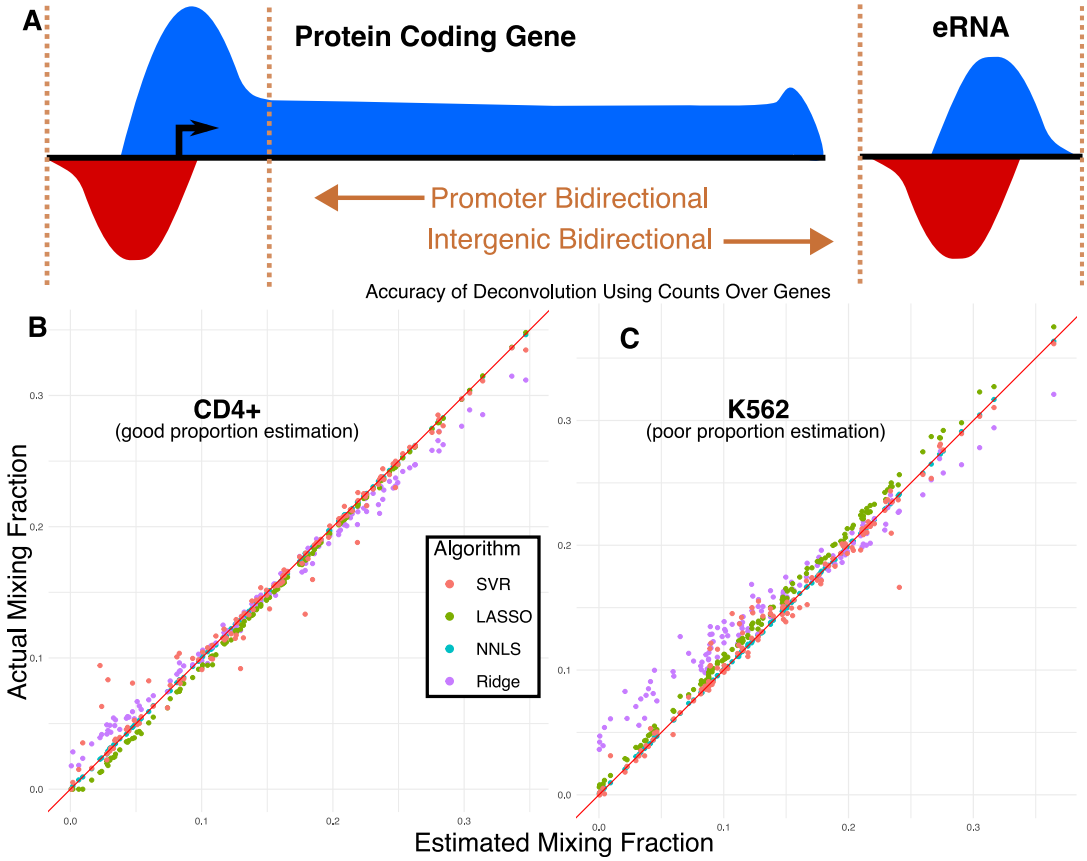


Figure 1. *A:* Nascent transcripts accumulate in a known bidirectional pattern around promoter sites as well as at enhancers.[7,36] These bidirectional regions are counted by convention around ±300bp from the site of RNA Polymerase initiation (roughly the center of the bidirectional).[9,10,36] For annotated genes, we exclude the initiation peak by counting +300 to the annotated transcription end site. *B:* Deconvolution was performed on random mixtures of cells from Table subsection 2.1. Some celltypes show highly accurate estimation of mixing proportion when doing deconvolution over all annotated genes, with most methods showing good linearity in their estimation. *C:* Other celltypes confound the regularized models used here, suggesting a systematic failure of regularization for proper estimation mixing proportion in this naive analysis. This failure appears to be more pronounced with L2 regularized methods and appears in all analyses conducted in this work, to some extent.

2.2. *Identifying bidirectionals as regions of interest*

In addition to transcription at annotated genes, nascent transcription data contains bidirectional transcription at both promoters and regulatory elements. While annotated genes are widely studied and the typical target for this class of deconvolution algorithms, the study of enhancer associated RNAs is important for understanding the regulatory landscape of the cell. Various methods exist to identify sites of bidirectional transcription[36–39] and to combine them across different samples.[10] As such, bidirectionals are an additional region of interest that we now consider in our deconvolution framework.

To this end, we use a combined set of 485,688 bidirectionals, identified by Tfit and dReg within the Nascent-flow framework, capturing both enhancer RNAs and promoter regions, for

all samples in Table 2.1.[36,38] Notably, this system is significantly larger than the set of protein coding genes (approximately 490,000 vs 20,000). In this work, we use the following terminology in reference to subsets of this system — Bidirectionals refers to any site of RNA polymerase II initiation and generally includes both promoters and enhancers; any bidirectionals whose 5′ end (+/-300bp annotated TSS) overlaps an annotated 5′ gene in the RefSeq hg38 annotation is called a promoter; all other bidirectionals are called enhancers. Given the large size of this system, we next turn our attention to filtering the set of bidirectionals, to shrink the size of the overdetermined system to make deconvolution more computationally feasible.

2.3. *Filtering methods are useful for shrinking the system*

In traditional deconvolution contexts like microarray and RNA-seq, patterns of differential expression are often leveraged to shrink the system. For example, CIBERSORT[40] uses an adaptive filtering method based on DESeq2 to find genes most indicative of specific celltypes. In the context of nascent sequencing data, however, tools like DEseq2 are problematic. The relatively low read coverage and cell type specificity of bidirectionals (e.g. inherent variability) leads DESeq2 to distrust these regions. To counter this, we developed a naive filtering scheme, selecting a fixed number of ROIs defined by the user for each homogenous reference sample where the reads for that sample were most different compared to all other samples. More formally, we define an algorithm for pruning the system of ROIs to a tractable level:

- Filter all ROIs to restrict them to regions where all celltypes have counts lower than the 99th percentile of reads in the sample. We do this to remove outliers whose extreme values could break the assumptions of a linear system.
- Generate transformed ratio T such that for each ROI (row), for each celltype (column), that entry is the log2 ratio of the count at that ROI over the maximum count for that ROI not in that celltype. This step generates a log2 transformed list of the ROIs that are the most specific to a single celltype.
- Order this list by the largest log2 ratio in any celltype in any ROI. Then, walk down this list keeping ROIs such that the number of ROIs for each celltype is approximately equal, up to some limit of elements. This generates a subset of the full system with the most celltype specific elements for each cell. The number of ROIs is approximate because the number of celltype specific elements varies per-celltype and can be exhausted at larger system sizes.

2.4. *Most linear methods perform with high accuracy on synthetic nascent data*

Given that bidirectional regions have distinct transcription characteristics compared to more robustly transcribed annotated genes, we first sought to assess deconvolution methods on the filtered bidirectional set. Using this set, we find that deconvolution achieves a high degree of accuracy (Figure 2A). Unexpectedly, we observe that across all sizes of system tested (including systems far in excess of the total number of genes in the human genome), non-negative least squares (NNLS) regression performs with the highest degree of accuracy. LASSO

(L1 regularized linear regression) has a close second in performance. This is likely because LASSO regularization will only drop out cell types that are unlikely to be present in the mixture. In contrast, Ridge Regression (L2 regularized linear regression) performs worse than all other tested methods for most system sizes. Similarly, ϵ-Support Vector Regression (ϵ-SVR) with L2 regularization also performs relatively poorly compared to NNLS, but relatively well compared to Ridge regression. Despite these differences in accuracy, all models perform reasonably well on our synthetic mixtures, achieving accuracy to within a few percent on randomized mixtures. This is notable because these deconvolution methods perform well both on systems much smaller and much larger than those typically used for deconvolution of RNA-seq data.

Interestingly, we find our subsetting method consistently selects a mixture of enhancers and promoters that does not significantly differ from the distribution expected by random chance (Figure 2B). Consequently, this procedure captures mostly eRNAs and not promoters, since the number of eRNAs far outnumbers the number of promoters. This suggests that certain enhancer-driven regulatory elements are highly informative in identifying celltype.

We next sought to determine which ROIs were most informative to the deconvolution problem. To answer this question, we utilized NNLS, the best performing method in our prior tests. Using NNLS, we compared the performance on bidirectionals (as in Figure 2A), to annotated genes (as in Figure 1B,C) and a combination of these features – selected using our region filtering approach (Figure 3). We find that these methods achieve high accuracy for both genes and bidirectionals across a number of system sizes, with somewhat reduced accuracy when combining these two sets of ROIs. This reduction in accuracy could be a result of colinearity in the combined set of ROIs, as some bidirectionals may be intronic and thus they are not a strictly non-overlapping set relative to annotated genes.

For the data tested and the size of system used, we found that certain methods in the literature were prohibitively slow for the large linear systems we tested. For example, a ν support vector regression (ν-SVR) approach as suggested by CIBERSORT[40] was too computationally expensive to test or benchmark reliably, taking more than 24 hours to do deconvolution on a single mixture of cells at large system sizes (approximately 100k ROIs or more). Due to these poor scaling characteristics, we instead chose to use an optimized implementation of the primal version of ϵ-SVR. This was chosen instead of a dual formulation to maintain computational tractability for the large number of samples relative to the number of features. In the context of nascent sequencing data, NNLS is likely the best model to use based on our benchmarking.

2.5. *Undifferentiated celltypes confound deconvolution of mixtures*

In the course of testing our model, we observed that certain celltypes strongly confounded all deconvolution models tested when using bidirectionals. To understand this puzzling behavior, we examined deconvolution in the presence and absence of these cell types. To do deconvolution of this system, we generated a titration curve, mixing celltypes from distinct separate mixing proportions into equivalent proportions for all celltypes.

We observed that both ESC cell lines and BJ5TA cell lines caused deconvolution to fail (Figure 4A,B). Specifically, inclusion of either cell line results in an overestimation of the mixing proportion for those cell types. We carefully examined these two cell lines to identify

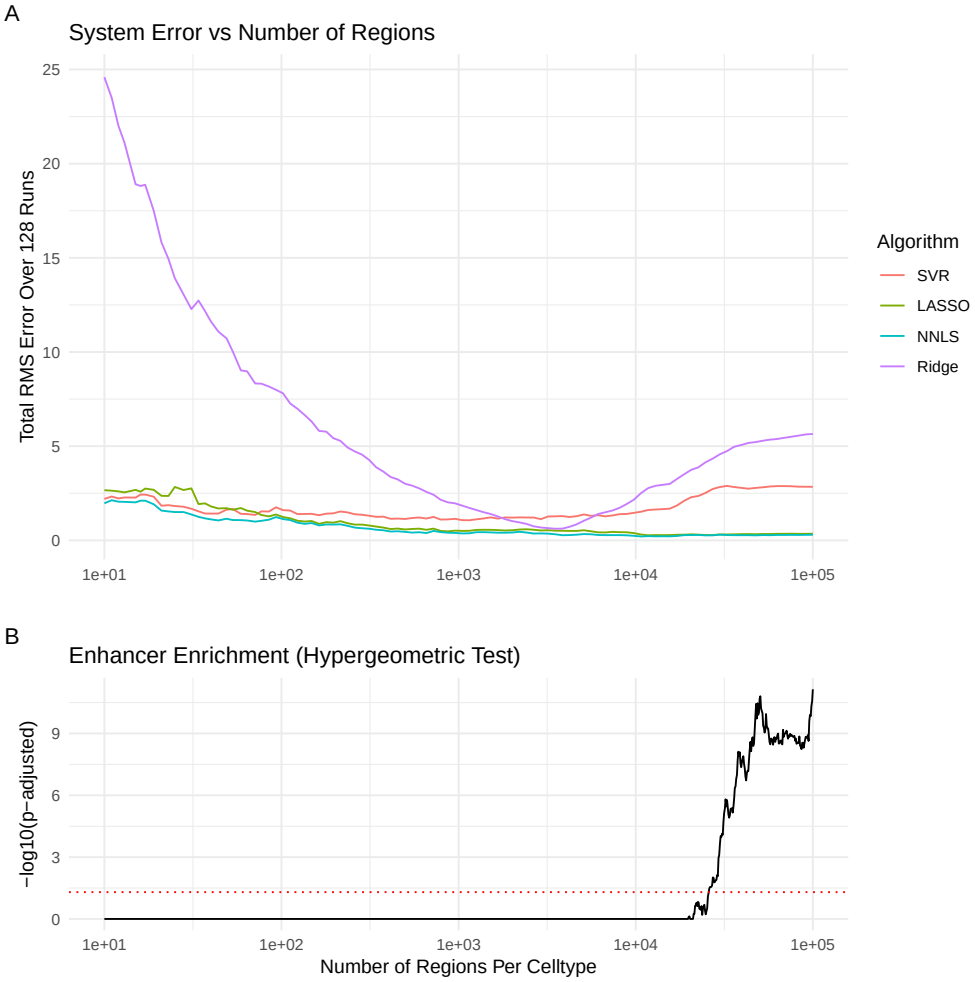


Figure 2. *A:* Models were tested using standard library implementations on a set of 128 randomly generated sets of mixing parameters. Each model was tested on 100 different subsets of ROIs selecting 10^n-many points for $n \in [1, 5]$ using linear spacing between subsequent n. Most models perform well in the intermediate region of 10^3–10^4 points selected per-sample, but diverge outside of that regime. For each set of ROIs selected, the same 128 randomly generated sets of mixing parameters were used as in Figure 2. We observe that for essentially all points, NNLS outperforms more complex models. *B:* To understand the selection process of our subsetting algorithm, we tested whether enhancers were selected from the full ROI set at a greater rate than would be expected by random. To do so, we performed a hypergeometric test with Bonferroni correction over all trials of our ROI subsets. We observe that for smaller system sizes the enhancer/promoter sampling ratio does not differ dramatically from that expected by random sampling. When the system size increases, enhancers become preferentially selected over promoters ($p < 0.05$), but this increase in the rate of enhancer selection does not correlate with the accuracy of any model.

distinguishing features relative to the other cell lines.

To determine whether the number of cell lines or cell line immortalization differences could be the source of the problem, we added lymphoblastoid cell lines immortalized (LCL) by EBV. Notably, LCLs do not confound the model and show excellent performance (Figure 4C). Both

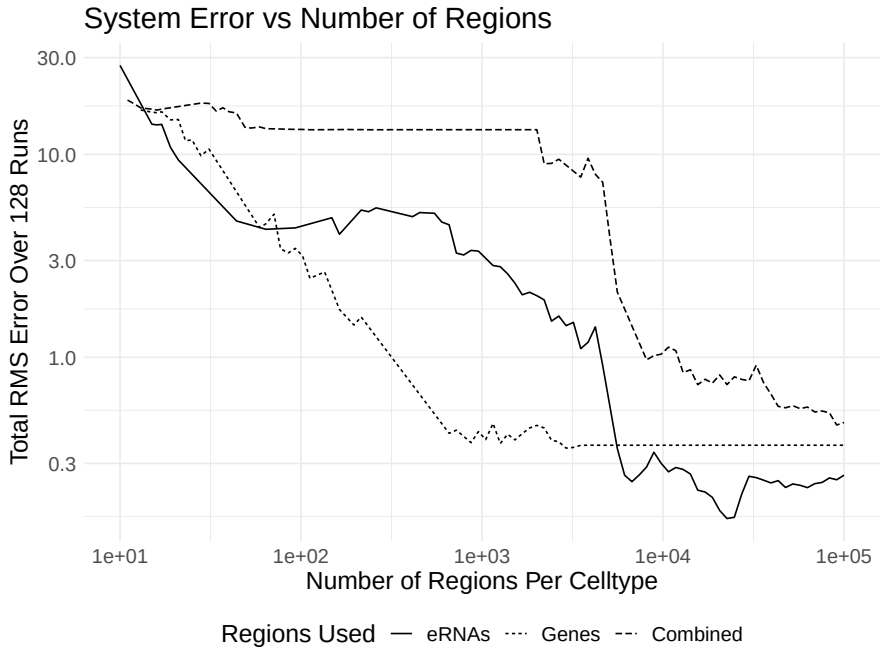
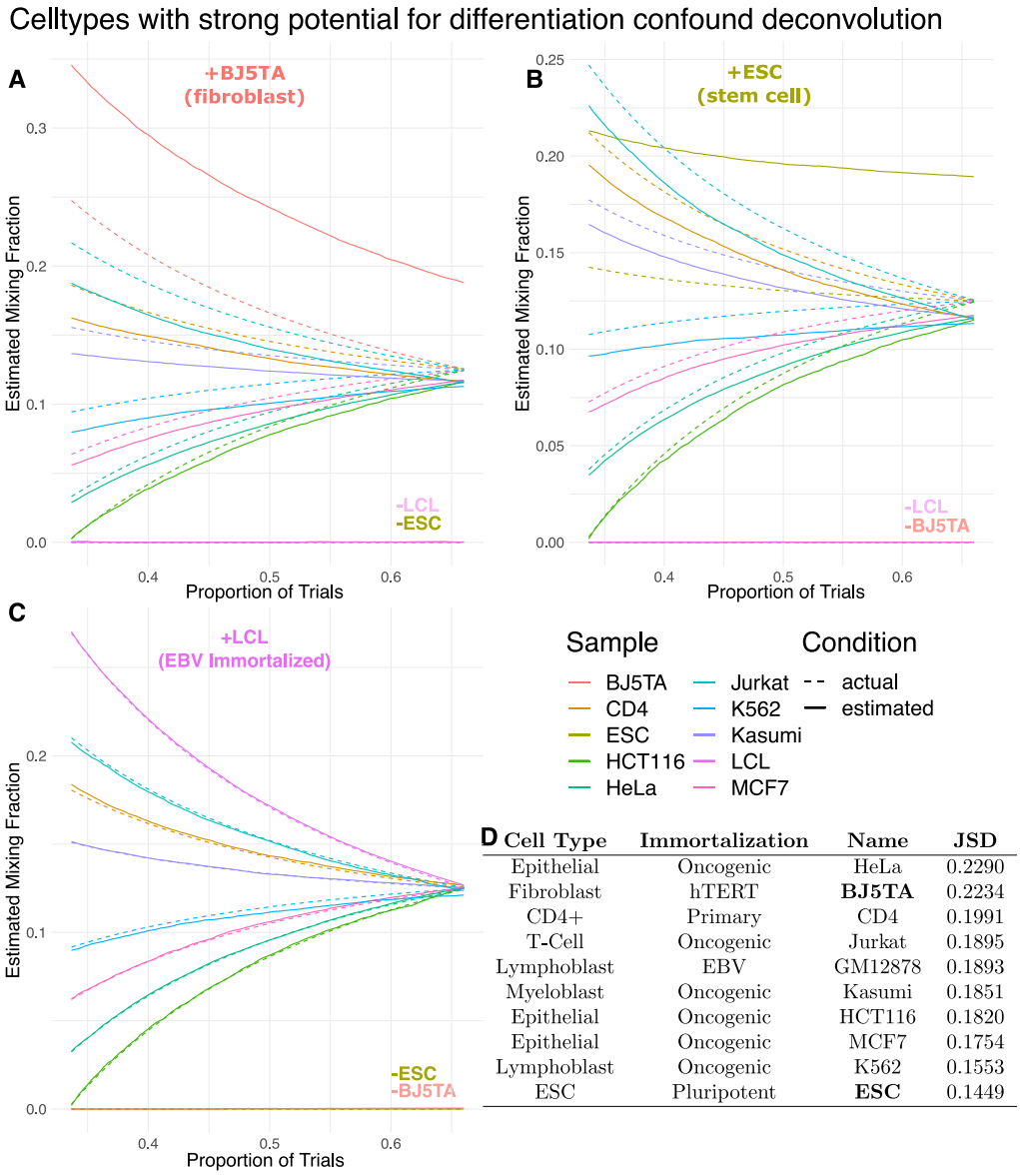


Figure 3. To compare the maximum theoretical accuracy of our system, we conducted the same analysis as in Figure 2 using either the region sets of bidirectionals, annotated genes, or a combination of the two, performing the same subsetting procedure as before. We observe that at smaller region sizes using genes alone provides a higher degree of accuracy than just bidirectionals, but that at larger sets of ROIs bidirectionals alone can achieve a higher absolute degree of accuracy. Somewhat unexpectedly, the combination of both sets of regions performs more poorly than each separate subset. Note that as system size increases, the accuracy of the set using annotated genes reaches a constant level purely because the total size of that system is exhausted by virtue of being an order of magnitude smaller than that of the bidirectionals or combined set.

ESC (embryonic stem cells) and BJ5TA (fibroblast derived) are non-terminally differentiated and non-oncogenic (Figure 4D). Furthermore, we see that even without regularization, NNLS successfully removes non-present celltypes (Figure 4A-C), meaning that undifferentiated cell-types will not be inferred in the mixing proportion if they are not present at all in the mixture. Furthermore, regularization techniques are not required to accomplish this removal of celltypes that are absent.

One alternative hypothesis to the source of this problem is that heterogeneity in the population of undifferentiated celltypes is the source. However, this would suggest that more heterogeneous cell populations should perform worse in deconvolution, as should cells from similar tissue types. Yet based on this data, this seems unlikely, given that both CD4+ and Jurkat cells, both peripheral blood mononuclear cells (PBMC) derived, are present in the mixture and are successfully estimated by our models. Since the addition of a lymphoblast cell line immortalized using EBV (GM12878) does not result in system failure in the same way that is observed with the non-differentiated cell-lines, we suspect that differentiation is the key issue here as opposed to heterogeneity. Our work suggests that undifferentiated or partially differentiated cell types pose a key challenge to the deconvolution of nascent sequencing data



D

Cell Type	Immortalization	Name	JSD
Epithelial	Oncogenic	HeLa	0.2290
Fibroblast	hTERT	**BJ5TA**	0.2234
CD4+	Primary	CD4	0.1991
T-Cell	Oncogenic	Jurkat	0.1895
Lymphoblast	EBV	GM12878	0.1893
Myeloblast	Oncogenic	Kasumi	0.1851
Epithelial	Oncogenic	HCT116	0.1820
Epithelial	Oncogenic	MCF7	0.1754
Lymphoblast	Oncogenic	K562	0.1553
ESC	Pluripotent	**ESC**	0.1449

Figure 4. To interrogate the effect of undifferentiated and partially differentiated celltypes on the performance of deconvolution, we performed a titration experiment, estimating mixing parameters for 100 different mixtures of celltypes as mixing proportions were taken from maximally separated to equivalent. For each trial n, the mixing proportions are equally spaced points in $[n\frac{1}{n_{\text{tot}}}, 1]$ that are then rescaled to sum to one. Each subset (A,B,C) was generated by holding out one celltype from the full mixture and renormalizing the adjusted mixing proportions to sum to one. *A,B:* Adding either BJ5TA or ESC cells into the mixture causes a higher-than-true proportion of those cells to be estimated. Neither of these cell lines are terminally differentiated. *C:* Addition of EBV immortalized LCL cells into the mixture does not result in failure of the deconvolution model, suggesting that the observed failures are not a function of how cells were immortalized. *D:* To understand if this failure could be attributed to celltype specificity, we calculated the mean Jensen-Shannon Divergence for each sample compared to all others. The pluripotent ESC cells show the lowest celltype specificity while the partially differentiated BJ5TA cells show the highest celltype specificity, with the exception of HeLa cells.

when using enhancers because their regulatory profile, particularly that of their enhancer regions, resemble an ensemble profile of multiple differentiated celltypes. In support of this, the problem does not seem to occur when using genes alone, suggesting that undifferentiated cells may lack the same level of specificity at bidirectionals as terminally differentiated cell types.

Our results suggest either very low or very high celltype specificity when looking at these samples' bidirectional ROIs (Figure 4D). When looking at the mean Jensen Shannon Divergence for each sample compared to all others, we observe that our undifferentiated cell lines are either the least specific (ESC) or the most specific (BJ5TA). Although HeLa cells show the highest degree of celltype specificity by this measure, HeLa cells are not representative of human cells, exhibiting notably different expression patterns[41] which would lead to a high degree of cell type specificity. Past work has shown that ESC cell lines have genome-wide transcriptional hyperactivity[42] that narrows as differentiation progresses. Additionally, work in hematopoetic cells has suggested that these undifferentiated cell lines are characterized by a high degree of fluidity in chromatin modification.[43] More work is required to definitively establish that differentiation is the source of the breakdown of deconvolution in this system, and will likely require significant work outside the scope of this preliminary study.

3. Conclusion

This work is the first to examine supervised deconvolution of heterogenous mixtures of nascent sequencing data. Deconvolution is an essential tool for the study of heterogenous samples, whether cell lines or tissues. While most work on deconvolution of heterogenous samples has moved on to focusing on single cell protocols, a single cell nascent sequencing protocol currently seems infeasible. Thus, nascent sequencing is limited to bulk experiments, which appear to be reliably separable by supervised deconvolution. We present here the use of nascent sequencing data as a testbed for this supervised deconvolution problem. We integrate best practices from the literature and develop new techniques to handle characteristics in nascent sequencing data where assumptions from the RNA-seq deconvolution literature do not hold.

To benchmark various deconvolution algorithms, we first developed a new algorithm to filter ROIs to only use regions with the most celltype specific expression. We find that this selection process does not preferentially select enhancer or promoter ROIs. That said, the number of enhancer associated bidirectionals far exceeds annotated genes, providing ample features from which to select regions of interest. Our proposed algorithm is simple, fast, and reliable, and establishes a strong first basis for the development of more specific ROI filtering tools for nascent deconvolution.

Using this algorithm, we compared standard methods used for solving the deconvolution problem. Specifically, we tested NNLS, Ridge, LASSO, as well as ϵ-SVR. We found that all methods reliably separate the nascent deconvolution system, with L2-regularized methods achieving comparatively poor performance to NNLS. Furthermore, we found that even a simple method like NNLS could reliably eliminate celltypes that were not present in the sample, suggesting regularization is not necessary for solving the deconvolution problem here. While we find that both annotated genes and bidirectionals can achieve high accuracy in supervised

deconvolution (with bidirectionals having an edge in absolute accuracy), it is worth emphasizing that bidirectionals are distinctly advantageous in that they are annotation-independent and discovered *de-novo* for each sample.

We show that the addition of undifferentiated samples to a nascent deconvolution system results in highly skewed mixing estimates, with undifferentiated celltypes predicted as far more likely than their actual frequency in the mixture. One possible reason for this is that undifferentated celltypes tend to show regulatory patterns akin to a combination of the regulatory patterns of each constituent celltype. It appears to be a necessary condition for some amount of the undifferentiated celltype to be present in the mixture in order for the system to fail.

One key issue in this work is the lack of availability of diverse high quality nascent sequencing data to perform simulations against. Although a large amount of nascent sequencing data is available and published, the number of cell types available is somewhat limited. Protocols aimed at extending run-on sequencing to a broader base of samples, such as ChRO-seq[44] show promise in alleviating this bottleneck. Importantly, many of the earliest nascent data sets lacked replicates – which excluded their usage here. Data quality and availability is often a limiting factor in computational studies, and this work is not an exception to that rule.

In this work, and generally for the supervised deconvolution problem, we assume that all cells in a sample are taken from an approximately homogeneous population. This is sometimes a reasonable assumption but is often not. One future frontier that could be highly beneficial to this project is the incorporation of single cell ATAC-seq (scATAC) as a secondary source of information to augment bulk nascent sequencing data. scATAC combines the chromatin accessibility readout provided by ATAC-seq (indicative of regions open to transcription) with the cell-specific information provided by modern single cell sequencing protocols. Tools are already well defined for clustering single cell sequencing data into constituent cell types, as individual cells can typically be separated using dimensionality reduction methods like PCA, tSNE, or UMAP.[45,46] Because transcription occurs in regions of open chromatin, which is what ATAC-seq measures, mixing fractions and celltype specific transcripts could be estimated more reliably using combined data from both protocols. Future work combining pairing single cell ATAC-seq data and nascent sequencing data could leverage techniques used by existing tools[21] to do deconvolution on a more granular level for individual samples, providing a strong complementary tool to the bulk deconvolution discussed here. While single cell approaches remain comparatively expensive, this combination would be a powerful tool for looking at transcriptional regulatory networks at the level of sub populations of samples.

Nascent sequencing is a powerful tool for the assessment of transcriptional regulatory networks, and when paired with deconvolution tools will also facilitate deeper understanding of those regulatory networks in heterogeneous cell populations. Leveraging a transcription oriented sequencing approach instead of an expression oriented (e.g. steady state) one provides myriad benefits — more thorough coverage of the genome, understanding of regulatory elements, and a deep view of underlying transcriptional dynamics — all of which can be integrated with different sequencing protocols to great effect. Supervised deconvolution represents an important preliminary foothold into this space, and this work shows that nascent sequencing data is well suited for that class of problems.

Bibliography

1. S. Mohammadi, N. Zuckerman, A. Goldsmith and A. Grama, A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues, *Proceedings of the IEEE* **105**, 340 (February 2017).
2. B. Hwang, J. H. Lee and D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, *Experimental & Molecular Medicine* **50**, 1 (August 2018).
3. S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis and A. J. Butte, Cell type–specific gene expression differences in complex tissues, *Nature Methods* **7**, 287 (April 2010).
4. Y. Zhong, Y.-W. Wan, K. Pang, L. M. Chow and Z. Liu, Digital sorting of complex tissues for cell type-specific gene expression profiles, *BMC Bioinformatics* **14**, p. 89 (March 2013).
5. F. Avila Cobos, J. Vandesompele, P. Mestdagh and K. De Preter, Computational deconvolution of transcriptomics data from mixed cell populations, *Bioinformatics* **34**, 1969 (June 2018).
6. A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan and H. F. Clark, Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus, *PLOS ONE* **4**, p. e6098 (July 2009).
7. L. J. Core, J. J. Waterfall and J. T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, *Science (New York, N.Y.)* **322**, 1845 (December 2008).
8. D. B. Mahat, H. Kwak, G. T. Booth, I. H. Jonkers, C. G. Danko, R. K. Patel, C. T. Waters, K. Munson, L. J. Core and J. T. Lis, Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq), *Nature Protocols* **11**, p. 1455 (August 2016).
9. J. G. Azofeifa, M. A. Allen, J. R. Hendrix, T. Read, J. D. Rubin and R. D. Dowell, Enhancer RNA profiling predicts transcription factor activity, *Genome Research* **28**, 334 (March 2018).
10. J. D. Rubin, J. T. Stanley, R. F. Sigauke, C. B. Levandowski, Z. L. Maas, J. Westfall, D. J. Taatjes and R. D. Dowell, Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment, *Communications Biology* **4**, 1 (June 2021).
11. J. F. Cardiello, G. J. Sanchez, M. A. Allen and R. D. Dowell, Lessons from eRNAs: Understanding transcriptional regulation through the lens of nascent RNAs, *Transcription* **11**, 3 (January 2020).
12. T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman and M. E. Greenberg, Widespread transcription at neuronal activity-regulated enhancers, *Nature* **465**, 182 (2010).
13. Z. Wang, T. Chu, L. A. Choate and C. G. Danko, Identification of regulatory elements from nascent transcription using dREG, *Genome Research* **29**, 293 (February 2019).
14. M. U. Kaikkonen, N. J. Spann, S. Heinz, C. E. Romanoski, K. A. Allison, J. D. Stender, H. B. Chun, D. F. Tough, R. K. Prinjha, C. Benner and C. K. Glass, Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription, *Molecular Cell* **51**, 310 (August 2013).
15. K. Kristjánsdóttir, A. Dziubek, H. M. Kang and H. Kwak, Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture, *Nature Communications* **11**, p. 5963 (November 2020).
16. S. Bae, K. Kim, K. Kang, H. Kim, M. Lee, B. Oh, K. Kaneko, S. Ma, J. H. Choi, H. Kwak, E. Y. Lee, S. H. Park and K.-H. Park-Min, Rankl-responsive epigenetic mechanism reprograms macrophages into bone-resorbing osteoclasts, *Cellular & Molecular Immunology* **20**, 94 (2023).
17. S. Hunter, R. F. Sigauke, J. T. Stanley, M. A. Allen and R. D. Dowell, Protocol variations in

run-on transcription dataset preparation produce detectable signatures in sequencing libraries, *BMC Genomics* **23**, p. 187 (March 2022).
18. F. Spitz and E. E. M. Furlong, Transcription factors: From enhancer binding to developmental control, *Nature Reviews Genetics* **13**, 613 (September 2012).
19. K. Lidschreiber, L. A. Jung, H. von der Emde, K. Dave, J. Taipale, P. Cramer and M. Lidschreiber, Transcriptionally active enhancers in human cancer cells, *Molecular Systems Biology* **17**, p. e9873 (January 2021).
20. T. Gong, N. Hartmann, I. S. Kohane, V. Brinkmann, F. Staedtler, M. Letzkus, S. Bongiovanni and J. D. Szustakowski, Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples, *PLOS ONE* **6**, p. e27156 (November 2011).
21. H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D. Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina, A. C. Adey, F. J. Steemers, J. Shendure and C. Trapnell, Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data, *Molecular Cell* **71**, 858 (September 2018).
22. D. D. Erdmann-Pham, J. Fischer, J. Hong and Y. S. Song, Likelihood-based deconvolution of bulk gene expression data using single-cell references, *Genome Research* **31**, 1794 (October 2021).
23. I. J. Tripodi and M. A. Gruca, Nascent-Flow (December 2018).
24. Y. Liao, G. K. Smyth and W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics (Oxford, England)* **30**, 923 (April 2014).
25. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin, The Sequence Alignment/Map format and SAMtools, *Bioinformatics* **25**, 2078 (August 2009).
26. W. Jiang, Z. Guo, N. Lages, W. J. Zheng, D. Feliers, F. Zhang and D. Wang, A Multi-Parameter Analysis of Cellular Coordination of Major Transcriptome Regulation Mechanisms, *Scientific Reports* **8**, p. 5742 (April 2018).
27. J. Fei, H. Ishii, M. A. Hoeksema, F. Meitinger, G. A. Kassavetis, C. K. Glass, B. Ren and J. T. Kadonaga, NDF, a nucleosome-destabilizing factor that facilitates transcription through nucleosomes, *Genes & Development* **32**, 682 (May 2018).
28. Z. Andrysik, M. D. Galbraith, A. L. Guarnieri, S. Zaccara, K. D. Sullivan, A. Pandey, M. MacBeth, A. Inga and J. M. Espinosa, Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity, *Genome Research* **27**, 1645 (October 2017).
29. N. Dukler, G. T. Booth, Y.-F. Huang, N. Tippens, C. T. Waters, C. G. Danko, J. T. Lis and A. Siepel, Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol, *Genome Research* **27** (October 2017).
30. Y. Zhao, Q. Liu, P. Acharya, K. Stengel, Q. Sheng, X. Zhou, H. Kwak, M. Fischer, J. Bradner, S. Strickland, S. Mohan, M. Savona, B. Venters, M.-M. Zhou, J. Lis and S. Hiebert, High-resolution mapping of RNA polymerases identifies mechanisms of sensitivity and resistance to BET inhibitors in t(8;21) AML, *Cell Reports* **16**, 2003 (August 2016).
31. C. G. Danko, L. A. Choate, B. A. Marks, E. J. Rice, Z. Wang, T. Chu, A. L. Martins, N. Dukler, S. A. Coonrod, E. D. Tait Wojno, J. T. Lis, W. L. Kraus and A. Siepel, Dynamic evolution of regulatory element ensembles in primate CD4+ T cells, *Nature Ecology & Evolution* **2**, 537 (2018).
32. T. Chu, E. J. Rice, G. T. Booth, H. H. Salamanca, Z. Wang, L. J. Core, S. L. Longo, R. J. Corona, L. S. Chin, J. T. Lis *et al.*, Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme, *Nature genetics* **50**, 1553 (2018).
33. L. J. Core, A. L. Martins, C. G. Danko, C. T. Waters, A. Siepel and J. T. Lis, Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and

enhancers, *Nature Genetics* **46**, 1311 (December 2014).
34. J. P. Smith, A. B. Dutta, K. M. Sathyan, M. J. Guertin and N. C. Sheffield, Quality control and processing of nascent RNA profiling data, *bioRxiv* **22**, p. 2020.02.27.956110 (February 2020).
35. K. Ikegami, S. Secchia, O. Almakki, J. D. Lieb and I. P. Moskowitz, Phosphorylated Lamin A/C in the Nuclear Interior Binds Active Enhancers Associated with Abnormal Transcription in Progeria, *Developmental Cell* **52**, 699 (March 2020).
36. J. G. Azofeifa and R. D. Dowell, A generative model for the behavior of RNA polymerase, *Bioinformatics* **33**, 227 (January 2017).
37. J. Azofeifa, M. A. Allen, M. Lladser and R. Dowell, FStitch: A Fast and Simple Algorithm for Detecting Nascent RNA TranscriptsBCB '14 (ACM, New York, NY, USA, Sept 2014).
38. C. G. Danko, S. L. Hyland, L. J. Core, A. L. Martins, C. T. Waters, H. W. Lee, V. G. Cheung, W. L. Kraus, J. T. Lis and A. Siepel, Identification of active transcriptional regulatory elements from GRO-seq data, *Nature Methods* **12**, 433 (May 2015).
39. Y. Zhao, N. Dukler, G. Barshad, S. Toneyan, C. G. Danko and A. Siepel, Deconvolution of expression for nascent RNA-sequencing data (DENR) highlights pre-RNA isoform diversity in human cells, *Bioinformatics* **37** (August 2021).
40. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn and A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles, *Nature Methods* **12**, 453 (May 2015).
41. J. J. M. Landry, P. T. Pyl, T. Rausch, T. Zichner, M. M. Tekkedil, A. M. Stütz, A. Jauch, R. S. Aiyar, G. Pau, N. Delhomme, J. Gagneur, J. O. Korbel, W. Huber and L. M. Steinmetz, The Genomic and Transcriptomic Landscape of a HeLa Cell Line, *G3: Genes—Genomes—Genetics* **3**, 1213 (March 2013).
42. S. Efroni, R. Duttagupta, J. Cheng, H. Dehghani, D. J. Hoeppner, C. Dash, D. P. Bazett-Jones, S. Le Grice, R. D. G. McKay, K. H. Buetow, T. R. Gingeras, T. Misteli and E. Meshorer, Global transcription in pluripotent embryonic stem cells, *Cell stem cell* **2**, 437 (May 2008).
43. Y. S. Chung, H. J. Kim, T.-M. Kim, S.-H. Hong, K.-R. Kwon, S. An, J.-H. Park, S. Lee and I.-H. Oh, Undifferentiated hematopoietic cells are characterized by a genome-wide undermethylation dip around the transcription start site and a hierarchical epigenetic plasticity, *Blood* **114**, 4968 (December 2009).
44. T. Chu, E. J. Rice, G. T. Booth, H. H. Salamanca, Z. Wang, L. J. Core, S. L. Longo, R. J. Corona, L. S. Chin, J. T. Lis, H. Kwak and C. G. Danko, Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme, *Nature Genetics* **50**, 1553 (2018).
45. L. van der Maaten and G. Hinton, Visualizing High-Dimensional Data Using t-SNE, *Journal of Machine Learning Research* **9**, 2579 (2008).
46. L. McInnes, J. Healy and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (September 2020).