Simplified motif background model provides significant speed-up for regulatory activity inference

Anonymous Authors[1]

Abstract

Regulatory microRNA activity can be inferred from mRNA data by evaluating the clustering of its target motif across expression-ranked mRNA sequences. Sequence-specific motif probabilities (SSPs) are required to distinguish functional motif enrichment from random motif occurrence driven by sequence length and nucleotide composition; however, current exact background Markov models scale poorly with k-mer motif size. Here, we introduce a binomial approximation parameterized by a single-site motif probability and sequence length for scalable SSP computation, which is at least 9-fold faster for k-mer screens while generating closely matching SSP values used for downstream Bayesian activity inference. The approximation is further extended to a 1st-order di-nucleotide background model, improving correction for compositionally biased sequences leading to better separation of AT- and GC-rich motif expectations for longer k-mers in AT-biased mRNA sequences. Applied to liver cancer samples, the binomial background approximation preserve downstream microRNA activity estimates, making it a better option for large-scale applications such as single-cell regulatory profiling.

1. Introduction

MicroRNAs (miRNAs) are short regulatory transcripts that fine-tune protein-coding mRNA abundance through target recognition and transcript degradation. They contribute to cell development, differentiation, integrity, and are frequently perturbed during cancer progression (Lee & Young, 2013; Slack & Chinnaiyan, 2019). During canonical targeting, miRNAs act through distinct $\sim$ 7 nucleotide (nt) long motifs located in the 3' untranslated regions (UTRs) of mRNAs (Figure 1a) (Van Roey & Davey, 2015; Gebert & MacRae, 2019). When a miRNA is active, transcripts containing its target motif tend to shift toward the lowly-expressed end of a ranked gene list. The motif-rank association enables unsupervised inference of miRNA activity from mRNA expression under conditions where the miRNA expression is otherwise unobserved (Van Dongen et al., 2008; Nielsen & Pedersen, 2021; Rasmussen et al., 2026).

This is particularly relevant for single-cell RNA-sequencing (scRNA-Seq), where miRNAs and their regulatory impact remain undetected outside specialized low-throughput protocols and primarily cell line settings (Isakova et al., 2021; Li et al., 2025). Methods such as Sylamer (Van Dongen et al., 2008) and miReact (Nielsen & Pedersen, 2021) infer activity by testing whether motif-containing genes significantly cluster along expression-ranked gene lists. Recently, bayesReact (Rasmussen et al., 2026) formulated a generative Bayesian hierarchical model of motif occurrence across ranked 3' UTR sequences. By modeling the motif distribution directly, bayesReact provides posterior uncertainty, data simulation, Bayes factor model comparison, and a statistical basis for extensions such as target efficiency and multi-modal regulatory modeling (Rasmussen et al., 2026).

Regulatory interpretation of motif distributions requires correcting for random occurrences, since short motifs appear frequently by chance, at rates varying with 3' UTR sequence lengths and nucleotide compositions. Thus, motif observations can reflect sequence context rather than functional targets. Activity inference, therefore, requires a background motif model to define null expectations conditional on each sequence context (Figure 1b,c). Motif distributions are then evaluated across sequence representations weighted by sequence-specific motif probabilities (SSPs), where a uniform distribution subsequently implies no activity (Rasmussen et al., 2026).

Markov models provide a principled solution to compute sequence-specific probabilities. Here, sequence generation is considered a finite discrete-time process, with state space defined by the motif of interest and transition probabilities given by nt frequencies or higher-order nucleotide contexts. Both Sylamer and miReact (through Regmex) use Markov corrections, with Regmex enabling large state spaces over

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

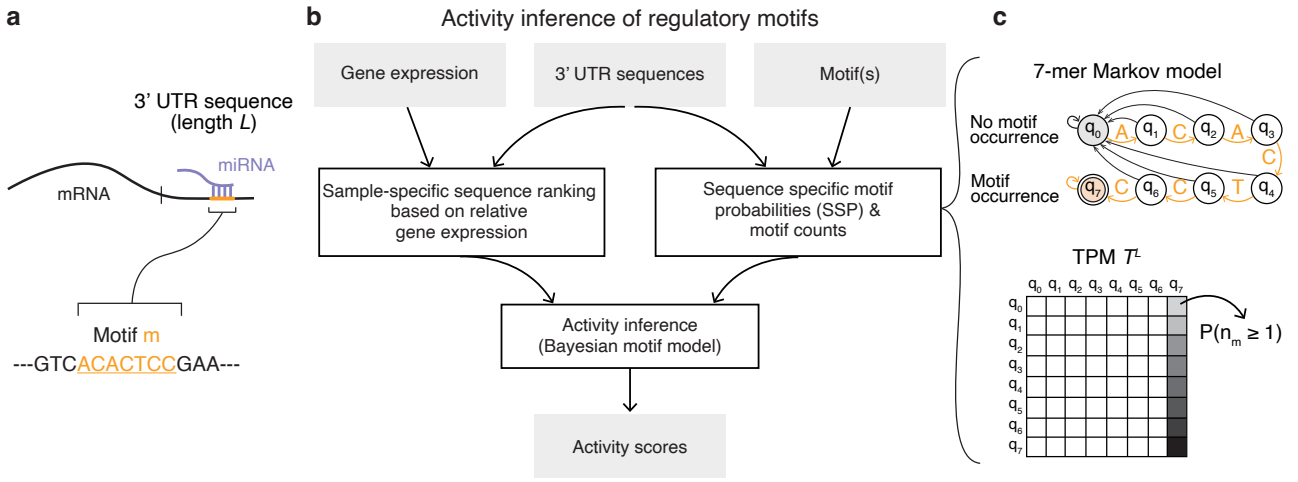

Figure 1. **Overview of motif model and regulatory activity inference.** (a) miRNA motif-binding model. UTR, untranslated region. (b) Steps for regulatory motif activity inference with input and output marked in grey. (c) Background motif model used to obtain sequence-specific motif probabilities. Depicted are the state space (top) and associated L-step transition probability (TPM) matrix (bottom). the L-step TPM is induced by a sequence with length L and motif m of length $k = 7$, and $P(n_m \geq 1)$ is the probability of observing m at least once in the sequence.

complex Regular expressions (REs) and embedded transition probability matrices (TPMs) to evaluate the SSP of any number of motif occurrences (Dongen & Enright, 2014; Nielsen et al., 2018). However, exact SSP evaluation becomes computationally expensive for large k-mer screens, longer motifs and sequences, due to large TPMs and high-dimensional matrix multiplication.

To address this issue, we implement a Binomial approximation of the Regmex background model, which is > 50-fold faster for 7-mer motifs and scales better with k-mer size, while producing closely matching SSP values. Furthermore, we show that extended di-nucleotide content improves the assessment of highly biased sequences, and better captures divergent SSP values between AT- and GC-rich motifs in 3' UTRs.

2. Methods

2.1. Data and processing

Human 3' UTR sequences were retrieved from the reference genome hg38 using BioMart and GENCODE v32 annotations (Yates et al., 2026; Frankish et al., 2021). We retained sequences with length between 20 and 10,000 nt and selected the longest 3' UTR isoform for each protein-coding gene. RNA sequences were represented on the cDNA alphabet $\Sigma = \{A, T, G, C\}$.

Paired mRNA and miRNA expression data were extracted from The Cancer Genome Atlas (TCGA), comprising bulk RNA-seq from 9,640 primary tumor samples across 32 cancer types (Weinstein et al., 2013). For each sample, mRNA counts were normalized by total read count, rescaled, and log2-transformed. For gene g in sample c, we computed a fold-change (FC) score

$$FC_{g,c} = x_{g,c} - \text{median}_{c'}(x_{g,c'}), \tag{1}$$

where $x_{g,c}$ denotes the normalized log-expression. Genes were ranked within each sample by decreasing $FC_{g,c}$, so 3' UTR sequences with lowest relative abundances are located at the end of the ranked list. We retained 18,559 3' UTR sequences with matched gene expression.

2.2. Background models for sequence-specific motif occurrence

To assess whether a motif is non-randomly distributed across a ranked sequence list, we need to correct for the confounding effects of sequence length and nucleotide composition through the sequence-specific probabilities of observing it at least once (Figure 1b). SSPs are obtained from a background Markov model over motif states, where sequence positions constitute discrete time steps and state transitions reflect partial motif matches. Motif occurrences are assumed to be non-overlapping, and the stochastic process always begins in an initial state q_0 (no motif observed), inducing an initial distribution $\pi = (1, 0, \ldots, 0)$. Regmex (Nielsen et al., 2018) represents a RE motif with a deterministic finite automaton (DFA) with state set Q, enabling evaluation of the probabilities of observing n_m motifs from 1 up to n_{obs} occurrences in a sequence by embedding the DFA n_{obs} times leading to a $Q \times \{1, \ldots, n_{obs}\}$ state space. The DFA can be interpreted as a Markov chain with transition probability matrix (TPM) T. However, for large embedded

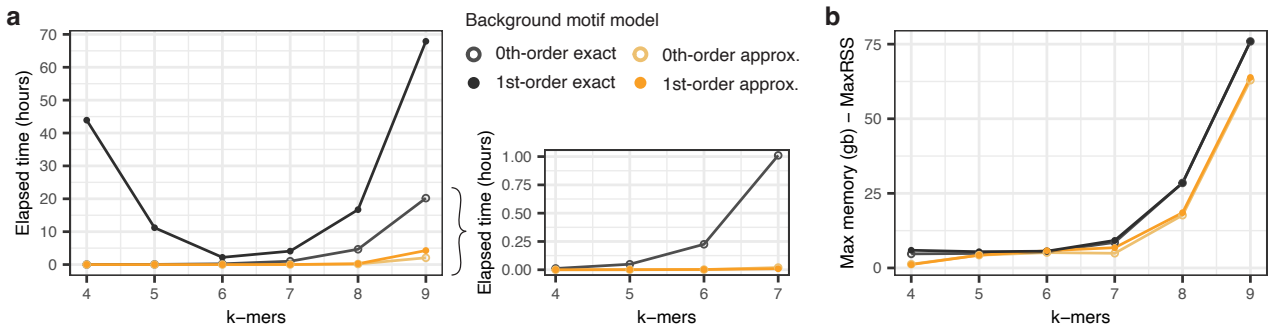

Figure 2. **Resource consumption of background motif models**. (a) Elapsed running times for different k-mer lengths and SSP computation; 0th- and 1st-order exact Markov models and corresponding binomial approximations. Full k-mer sets (4^k) are evaluated. (b) Maximum memory usage for the different model runs. MaxRSS, Maximum resident set size.

spaces the size of T explodes, and we consider instead the probability of observing a motif m at least once in sequence: $SSP = P(n_m \geq 1)$. We subsequently avoid extensive embedding by constructing a Markov chain that is absorbing in its final state q_K, where $K = k$ for any k-mer motif.

Let s be a sequence of length L over Σ with observed nucleotide frequencies $\{f_A, f_T, f_G, f_C\}$. Under a 0th-order Markov model, sequence positions are modeled as independent draws with $P(s_i = nt) = f_{nt}$. Thus, for a motif m occurring in s, the motif transition probabilities are given by:

$$T_{i,j} = \sum_{nt \in \Sigma} f_{nt} \cdot \mathbf{1}[\delta(q_i, nt) = q_j] \tag{2}$$

where $\delta(q_i, nt)$ is the deterministic transition function returning the next motif state after generating nucleotide nt from state q_i. The indicator expression equals one when a transition leads to q_j and zero otherwise. The SSP is then:

$$SSP = P(n_m \geq 1) = [T^L]_{q_0, q_K} \tag{3}$$

which corresponds to the probability of starting in q_0 (no motif observation) and transitioning to q_K (motif observation) after L steps (Figure 1c). In cases with long sequences and motifs, obtaining SSPs becomes computationally demanding and is especially problematic for large k-mer spaces, compared with evaluating a smaller set of user-defined biological REs. For scalable k-mer inference, we introduce a binomial approximation. The single-site motif probability under a 0th-order background model is

$$P(m) = \prod_{i=1}^{k} P(m_i) \tag{4}$$

Where $m = m_1 ... m_i ... m_k$ and $P(m_i) = f_{m_i}$. Under a binomial model, the number of possible motif start sites in s is $t = L - k + 1$. These are considered independent Bernoulli trials, and then $n_m \sim \text{Binom}(t, P(m))$:

$$\begin{aligned} P(n_m = 0) &= (1 - P(m))^t \\ SSP = P(n_m \geq 1) &= 1 - P(n_m = 0) \end{aligned} \tag{5}$$

To account for local nucleotide dependence in compositionally biased sequences, we also consider a 1st-order background model in which $P(m)$ is computed from conditional nucleotide probabilities:

$$\begin{aligned} P(m) &= P(m_1) \prod_{i=2}^{k} P(m_i | m_{i-1}) \\ &= f_{m_1} \prod_{i=2}^{k} \frac{f_{m_{i-1} m_i}}{f_{m_{i-1}}} \end{aligned} \tag{6}$$

Here, the conditional probability $P(m_i|m_{i-1}) = \frac{P(m_{i-1}, m_i)}{P(m_{i-1})}$, the marginal probability $P(m_{i-1}) = \sum_{nt \in \Sigma} P(m_{i-1}, nt)$, and the joint probability $P(m_{i-1}, m_i)$ is estimated from di-nucleotide frequencies. After defining $P(m)$, the binomial approximation is applied as in eq. 5.

The SSP is found for each sequence and motif pair, resulting in a **SSP** matrix with $S \times M$ entries, where S and M are the number of sequences and motifs, respectively. We use $S = 18,559$ and $M = 4^k$. Here, the binomial approximation trades exact Markov evaluation for $O(k)$ computation per sequence-motif pair, making it suitable for large screens where matrix exponentiation is the bottleneck.

2.3. Model comparisons

All background motif models were tested on complete k-mer sets with lengths 4–9, and run on the same high-performance computing (HPC) facility using 16 CPU cores.

We compared motif distributions under the different background models across FC-ranked 3' UTR sequences from

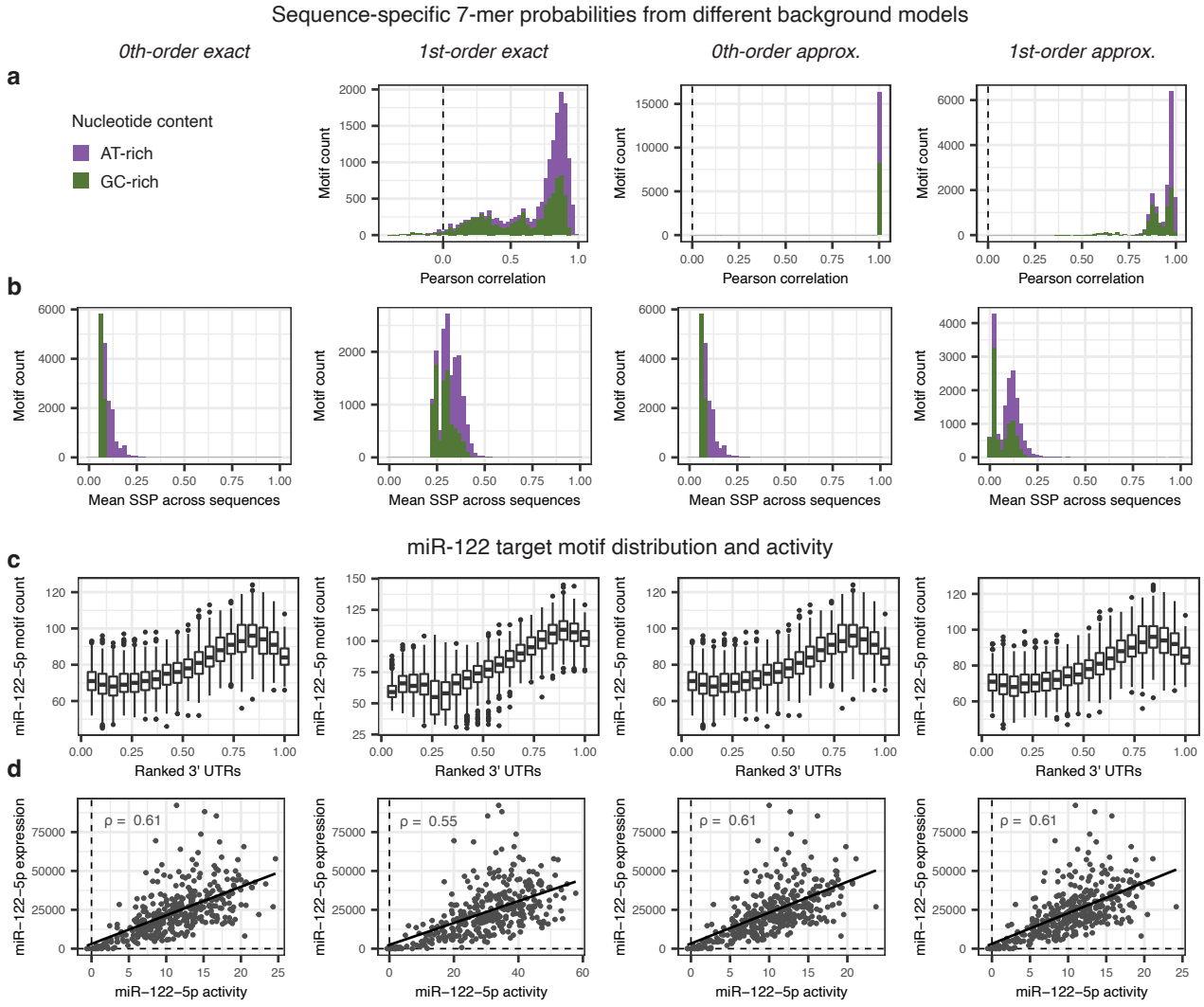

Figure 3. **7-mer motif probabilities and distribution**. (a) Pearson correlation between SSPs from 0th-order exact Markov model (default) against other background motif models (columns). Motifs are colored by nucleotide content, where AT-rich entails that A and T constitute the majority of the motif. (b) Distribution of mean sequence-specific probabilities (SSPs) for different background models. Motif mean is found across all 3' UTR sequences. (c) Liver-specific miR-122 target motif distribution across ranked 3' UTR sequences. The SSP-adjusted sequence interval $[0, 1]$ is divided into 20 bins, and the target motif occurrence is counted within each bin for each liver cancer sample (n = 367). (d) Inferred miR-122 activity against its observed normalized expression (transcripts per million, TPM). Linear regression line is shown, and Pearson ρ correlation is annotated.

TCGA primary tumor samples. Activity inference was then performed on the same ranked data using bayesReact with default settings (Rasmussen et al., 2026).

3. Results

3.1. Binomial approximation of motif background model

We first evaluated whether the exact 0th-order Markov background model could be replaced by a binomial approximation for exhaustive k-mer SSP computation. Across complete k-mer sets, the approximation reduced elapsed runtime by at least 9-fold relative to exact 0th-order evaluation, with increasing gains for larger k (Figure 2a). For the 7-mer set, which contains canonical miRNA target motifs, the full **SSP** matrix was computed in approximately one minute with a maximum memory use of 5 GB, making local execution feasible (Figure 2b). Meanwhile, the exact background models rely on Regmex Markov implementations. The 1st-order model constructs a full embedded TPM, whose state space grows with the observed number of motif occurrences, n_{obs}. This results in long run times for short k-mers as they can occur thousands of times in long 3' UTRs, producing large embedded state spaces (Figure 2a).

Under the 0th-order background model, sequence positions are independent conditional on sequence-specific nucleotide frequencies, and motif occurrences are evaluated under a non-overlapping counting convention. The resulting binomial SSP approximation thus closely matches the exact SSPs across motifs (Pearson correlation $\rho > 0.95$; Figure 3a, Extended Figure 4). Due to assumptions of non-overlapping motif counts, the close agreement is not surprising, as both models reduce motif occurrence to independent sequence-specific nucleotide draws parameterized by the same single-site motif probability $P(m)$. The 0th-order binomial model substantially reduces computational cost while preserving SSP values for downstream activity inference. This is evident when considering the liver-specific miR-122 (Coulouarn et al., 2009) where the motif distribution across SSP-adjusted and ranked 3' UTR representations remains nearly unchanged under the SSP approximation (Figure 3c). The resulting bayesReact activity estimates for miR-122 in liver cancer retained the same correlation with measured miR-122 expression for the exact and approximate 0th-order background models ($\rho = 0.61$; Figure 3d).

3.2. Accounting for extended sequence context

Human 3' UTRs are compositionally biased, with a mean AT content of $f_A + f_T = 56\%$. Under a 1st-order background model, motif probabilities are estimated from conditional nucleotide frequencies rather than marginal frequencies alone. Noteably, di-nucleotide frequencies can capture local sequence architecture not represented by nucleotide content alone, including homopolymeric AT tracts and other AT-rich elements that are common in 3' UTRs.

Consistently, GC-rich k-mers have lower mean SSPs in the 3' UTRs than AT-rich motifs, particularly under the 1st-order background models (Figure 3a,b). The 0th- and 1st-order models have the highest agreement for shorter k-mers, where fewer conditional transitions are accumulated. Disagreement increases for longer k-mers and compositionally biased motifs, especially GC-rich motifs (Extended Figure 4). This is expected because the 1st-order motif probability (see eq. 6) accumulates sequence-specific di-nucleotide effects multiplicatively across motif positions.

The 1st-order binomial approximation is comparable to the exact 1st-order model for shorter k-mers, but increasingly underestimates SSPs for longer k-mers (Extended Figure 5). This likely reflects the strong assumptions of the binomial model, in which motif opportunities are treated as independent Bernoulli trials with a single sequence-level motif probability, whereas the exact Markov model evaluates motif occurrence via the automaton state process. An approximation error can therefore accumulate with motif length and strong di-nucleotide dependence.

Bimodal SSP distributions are observed under the 1st-order background motif models, and are consistent with differing motif composition and entropy. Low-entropy AT-rich motifs have high expected occurrence probabilities in long AT-rich 3' UTRs, whereas higher-entropy or GC-rich motifs results in less frequent conditional transitions and therefore receive lower SSPs. For the C-rich miR-122 target motif, exact 1st-order background correction produce stronger clustering at the end of ranked 3' UTRs than the 0th-order correction, entailing increased clustering in lowly abundant sequences in the TCGA liver cancer samples (Figure 3c). Although this did not improve the correlation between inferred activity and measured miR-122 expression (Figure 3d), it increased activity magnitude, suggesting that higher-order background models can enhance motif activity signals by more accurately modeling sequence-specific null expectations.

4. Discussion

Here, we provide a binomial approximation to exact motif background models, useful for large-scale k-mer screens. The current approximation does not handle complex REs, where DFA construction in Regmex remains the primary solution and also enables overlapping motif occurrences if specified (Nielsen et al., 2018). However, since the binomial model is parameterized only by t and the single-site motif probability $P(m)$, it can potentially accommodate motif representations beyond exact k-mers used for miRNA target sites. Position weight matrix (PWM) motifs could be incorporated by defining motif observations through fixed PWM score thresholds and deriving corresponding SSPs under the approximate background model. This would enable activity inference for regulators with more degenerate binding preferences, such as RNA-binding proteins (Van Roey & Davey, 2015), while retaining the downstream Bayesian activity model (Rasmussen et al., 2026).

Minimizing computational resources is important for large-scale activity estimation, such as miRNA activity inference across millions of cells, and for enabling access to users without HPC capacity. For additional computational speed-up, dedicated k-mer counting could be considered as an alternative to the demanding exact RE matching performed under the current Regmex implementation. Maximum memory usage could also be further optimized by retaining only part of the $S \times M$ **SSP** matrix in memory, computing SSP values in chunks, which is possible since all sequence-motif pairs are assumed independent.

In conclusion, these results show that the binomial approximation of the background motif model is several-fold faster while preserving comparable SSP values. Since the SSPs provide a background correction for random motif occurrence used in downstream generative motif modeling and activity inference, the approximation enables regulatory activity inference in non-HPC settings and can provide a

starting point for other classes of motif inputs. Higher-order background models further show promise for correcting sequence bias in longer k-mers and in biological sequences with non-random local nucleotide structure.

Impact Statement

This work aims to improve the computational accessibility of sequence-specific motif background correction for regulatory activity inference. By reducing the resources required for large-scale motif probability evaluation, the proposed approximation may enable researchers to perform motif analyses on local machines and large single-cell datasets. The broader framework for activity inference is intended for regulatory motif screens and hypothesis generation.

References

Coulouarn, C., Factor, V. M., Andersen, J. B., Durkin, M. E., and Thorgeirsson, S. S. Loss of miR-122 expression in liver cancer correlates with suppression of the hepatic phenotype and gain of metastatic properties. *Oncogene*, 28(40):3526–3536, 2009.

Dongen, S. v. and Enright, A. J. Detecting microRNA signatures using gene expression analysis. In Kasabov, N. (ed.), *Springer Handbook of Bio-/Neuroinformatics*, pp. 129–150. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., et al. Gencode 2021. *Nucleic acids research*, 49(D1):D916–D923, 2021.

Gebert, L. F. and MacRae, I. J. Regulation of microRNA function in animals. *Nature reviews Molecular cell biology*, 20(1):21–37, 2019.

Isakova, A., Neff, N., and Quake, S. R. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proceedings of the National Academy of Sciences*, 118(51): e2113568118, 2021.

Lee, T. I. and Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013.

Li, J., Tian, J., and Cai, T. Integrated analysis of mirnas and mrnas in thousands of single cells. *Scientific Reports*, 15 (1):1636, 2025.

Nielsen, M. M. and Pedersen, J. S. miRNA activity inferred from single cell mRNA expression. *Scientific Reports*, 11 (1):9170, 2021.

Nielsen, M. M., Tataru, P., Madsen, T., Hobolth, A., and Pedersen, J. S. Regmex: a statistical tool for exploring motifs in ranked sequence lists from genomics experiments. *Algorithms for Molecular Biology*, 13(17):1–11, 2018.

Rasmussen, A. M., Bouchard-Côté, A., and Pedersen, J. S. bayesreact: expression-coupled regulatory motif analysis detects microrna activity across cancers, tissues, and at the single-cell level. *Nucleic Acids Research*, 54(4): gkag072, 2026.

Slack, F. J. and Chinnaiyan, A. M. The role of non-coding RNAs in oncology. *Cell*, 179(5):1033–1055, 2019.

Van Dongen, S., Abreu-Goodger, C., and Enright, A. J. Detecting microRNA binding and siRNA off-target effects from expression data. *Nature methods*, 5(12):1023–1025, 2008.

Van Roey, K. and Davey, N. E. Motif co-regulation and cooperativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation. *Cell Communication and Signaling*, 13(1):1–16, 2015.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Yates, A. D., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. Ensembl 2026. *Nucleic Acids Research*, 54(D1):D1053–D1060, 2026.

A. Extended Figures

A.1. Extended Figure 4

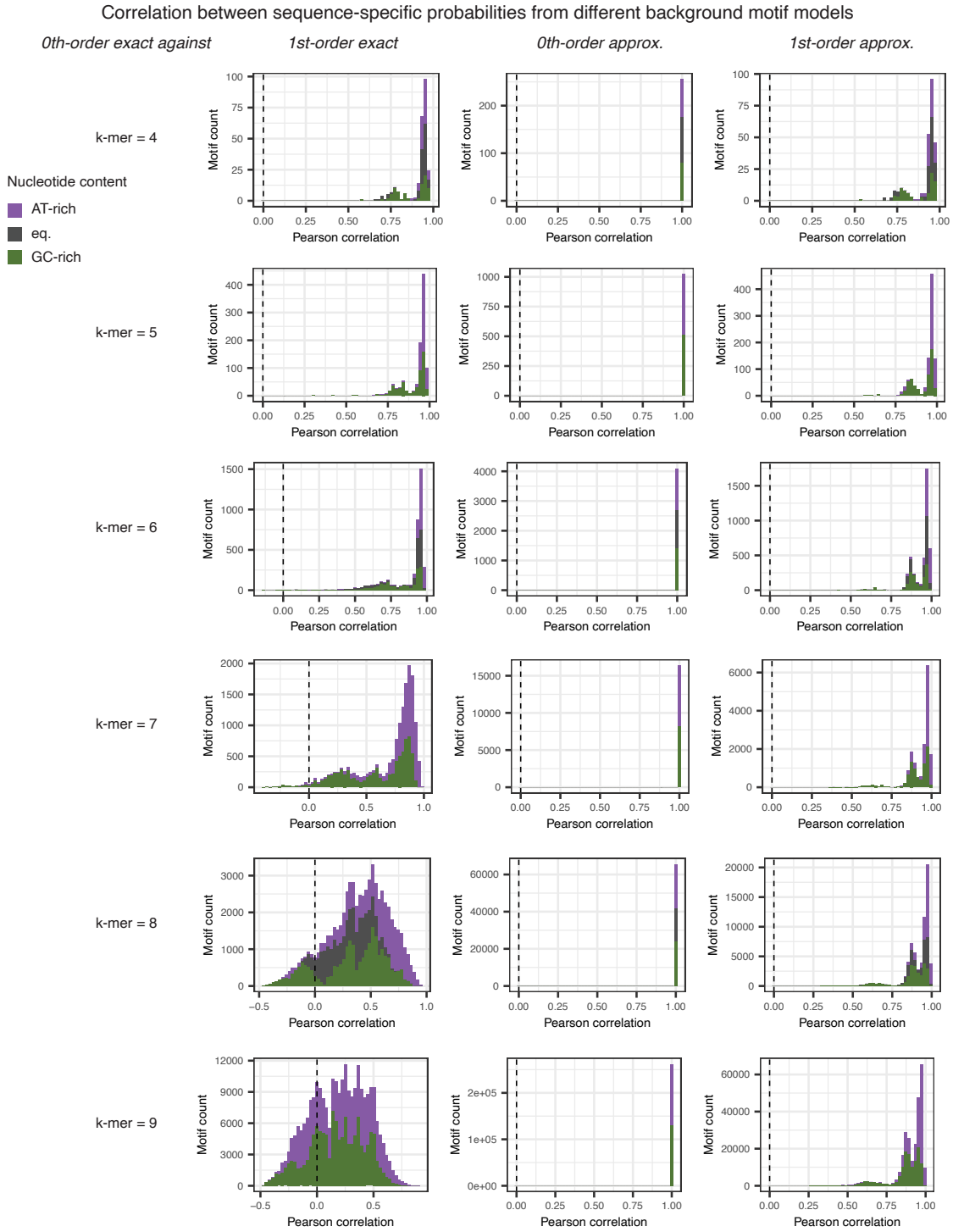

Figure 4. **Comparison of default 0th-order Markov model with extended and approximate background motif models.** Pearson correlation between 0th-order exact SSP for each motif in k-mer set (rows) against SSPs under other models (columns). Motifs are colored by nucleotide content, where AT-rich implies that A and T constitute the majority of the motif. eq., equal.

A.2. Extended Figure 5

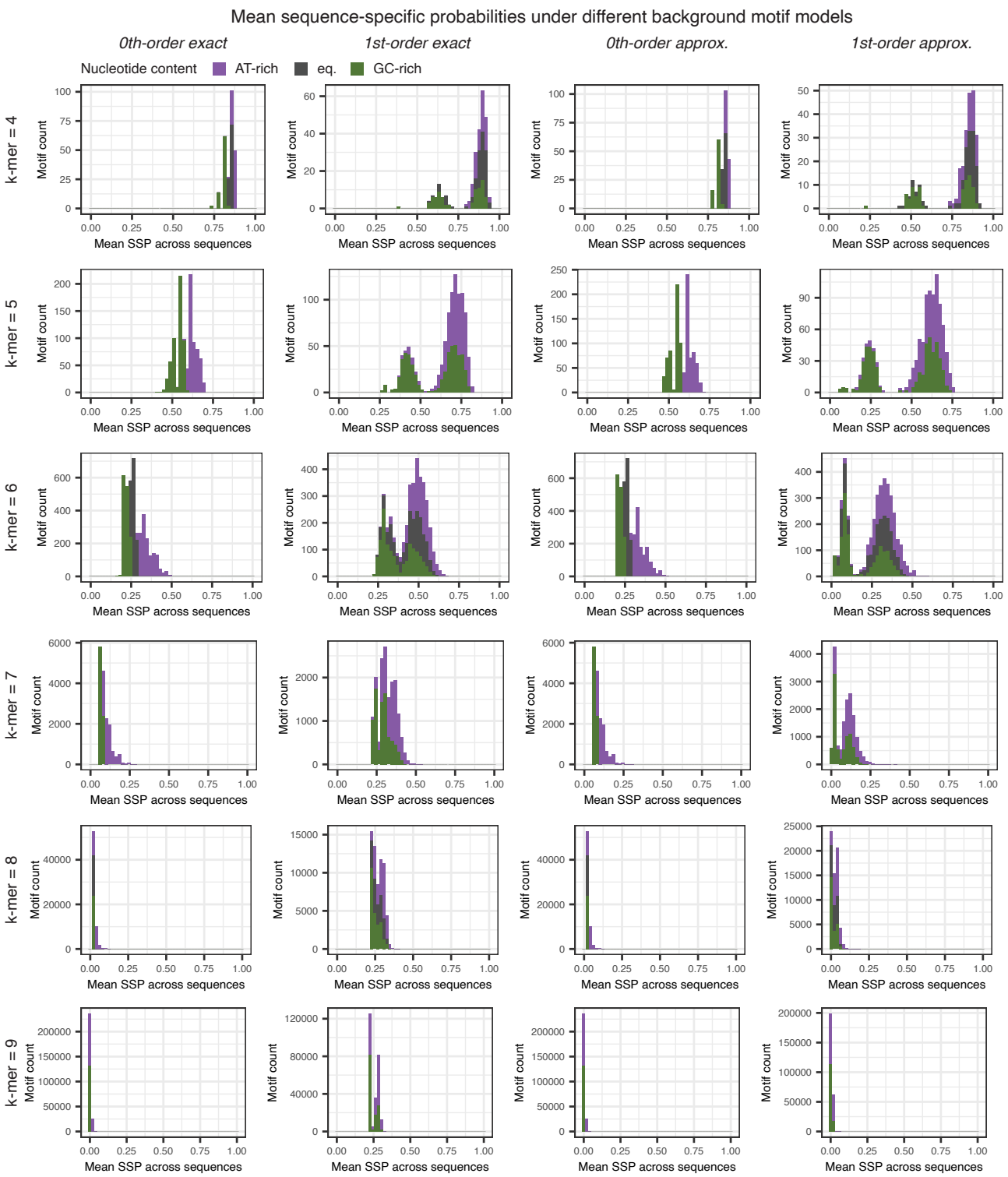

Figure 5. **Comparison of sequence-specific probabilities from different background motif models.** Mean sequence-specific probabilities (SSPs) for different background motif models. Motifs are colored by nucleotide majority.