frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fmolb.2022.1095193

OPEN ACCESS

EDITED BY
Guo-Wei Wei,
Michigan State University, United States

REVIEWED BY
Duan Chen,
University of North Carolina at Charlotte, United States
Yu Zhou,
Wuhan University, China
Jeffrey Wilusz,
Colorado State University, United States

*CORRESPONDENCE
Felipe-Andrés Piedra,
piedra@bcm.edu

SPECIALTY SECTION
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 10 November 2022
ACCEPTED 19 December 2022
PUBLISHED 09 January 2023

CITATION
Piedra F-A, Henke D, Rajan A, Muzny DM, Doddapaneni H, Menon VK, Hoffman KL, Ross MC, Javornik Cregeen SJ, Metcalf G, Gibbs RA, Petrosino JF, Avadhanula V and Piedra PA (2023), Modeling nonsegmented negative-strand RNA virus (NNSV) transcription with ejective polymerase collisions and biased diffusion. *Front. Mol. Biosci.* 9:1095193. doi: 10.3389/fmolb.2022.1095193

Modeling nonsegmented negative-strand RNA virus (NNSV) transcription with ejective polymerase collisions and biased diffusion

Felipe-Andrés Piedra[1]*, David Henke[1], Anubama Rajan[1], Donna M. Muzny[2], Harsha Doddapaneni[2], Vipin K. Menon[2], Kristi L. Hoffman[1], Matthew C. Ross[2], Sara J. Javornik Cregeen[1], Ginger Metcalf[2], Richard A. Gibbs[2], Joseph F. Petrosino[1], Vasanthi Avadhanula[1] and Pedro A. Piedra[1,3]

[1]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, Unites States, [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, Unites States, [3]Department of Pediatrics, Baylor College of Medicine, Houston, TX, Unites States

Infections by non-segmented negative-strand RNA viruses (NNSV) are widely thought to entail gradient gene expression from the well-established existence of a single promoter at the 3′ end of the viral genome and the assumption of constant transcriptional attenuation between genes. But multiple recent studies show viral mRNA levels in infections by respiratory syncytial virus (RSV), a major human pathogen and member of NNSV, that are inconsistent with a simple gradient. Here we integrate known and newly predicted phenomena into a biophysically reasonable model of NNSV transcription. Our model succeeds in capturing published observations of respiratory syncytial virus and vesicular stomatitis virus (VSV) mRNA levels. We therefore propose a novel understanding of NNSV transcription based on the possibility of ejective polymerase-polymerase collisions and, in the case of RSV, biased polymerase diffusion.

KEYWORDS

RNA viral genome, transcriptional regulation, biased diffusion, polymerase collisions, viral gene expression

Introduction

Viruses with nonsegmented negative-strand RNA genomes (NNSV) (all viruses of the order *Mononegavirales*) contain major pathogens such as Ebola, rabies, measles virus, respiratory syncytial virus (RSV), and vesicular stomatitis virus (VSV)—the latter is a highly studied bovine pathogen of the same family, *Rhabdoviridae*, as rabies virus.

The RNA genomes of NNSV are coated in nucleoprotein and support both whole genome replication and the transcription of subgenomic mRNAs by viral RNA-dependent RNA polymerases in the cytosol of infected cells. These genomes have a single promoter located at the 3′ end that is essential for both processes, presumably by facilitating the transient dissociation of terminal genomic RNA from nucleoprotein and the entry of viral polymerases, hitherto bound only to the nucleoprotein of the ribonucleoprotein (RNP) complex, into the RNA genome.

Every NNSV gene contains essential and highly conserved gene start (GS) and less highly conserved gene end (GE) signal sequences flanking the open reading frame (ORF).

Transcription is initiated at the GS signal which also serves as a capping signal on the 5' end of nascent mRNA (Barik, 1993; Liuzzi et al., 2005; Noton and Fearns, 2015). The polymerase then enters elongation mode until it reaches a GE signal, where it either continues translocating and transcribing (i.e., reads through) or it stops translocating and the mRNA is polyadenylated and released (i.e., terminates transcription) (Kuo et al., 1997; Noton and Fearns, 2015). In RSV, the two genes that are most 5' terminal have overlapping ORFs: the GE signal of matrix 2 (M2) occurs downstream of the GS signal of the last gene, the large polymerase (L) gene. Thus, for full-length L mRNA to be made, a polymerase must translocate 3' from the M2 GE signal (Fearns and Collins, 1999), suggesting that polymerases scan the RSV genome bidirectionally (i.e., diffuse) for a new GS signal after terminating transcription. Indeed, multiple studies suggest that scanning polymerase dynamics, or polymerase diffusion along the genome, may be a universal feature of NNSV transcription (Fearns and Collins, 1999; Kolakofsky et al., 2004; Barr et al., 2008; Noton and Fearns, 2015; Brauburger et al., 2016).

The still widely accepted textbook model of NNSV gene expression predicts a transcription gradient from 1) polymerase entry at the 3' end of the genome; 2) "obligatorily sequential" start-stop transcription in response to the conserved GS and GE signal sequences; and 3) transcriptional attenuation *via* an unknown mechanism between genes (Whelan et al., 2004; Noton and Fearns, 2015). However, multiple published studies show NNSV gene expression patterns—especially from RSV, which is one of its most highly studied members—that are either non-gradient, with one or more downstream genes appearing more highly expressed than upstream genes, or inconsistent with a simple gradient from a constant level of attenuation between genes (Krempl et al., 2002; Pagan et al., 2012; Aljabr et al., 2016; Levitz et al., 2017; Piedra et al., 2020a; Donovan-Banfield et al., 2022; Rajan et al., 2022). Regarding the latter, multiple studies show an abrupt and dramatic decrease in gene expression over the last two genes of the RSV genome (Krempl et al., 2002; Aljabr et al., 2016; Levitz et al., 2017; Donovan-Banfield et al., 2022; Rajan et al., 2022), the sole region of the genome containing overlapping ORFs—the textbook model of NNSV transcription offers no way of explaining this. In addition, the textbook model is devoid of potentially important biophysical phenomena: 1) polymerase (pol) diffusion along the viral genome; 2) potential interactions among pols (both diffusing and transcribing); and 3) stochastic transcription initiation and termination.

Here we implement a coarse-grained, mechanistic and stochastic computational model incorporating known and, ultimately, newly proposed features (ejective pol-pol collisions and 5' biased pol diffusion) of the underlying molecular biophysics to gain a deeper understanding of NNSV transcription and to capture, for the first

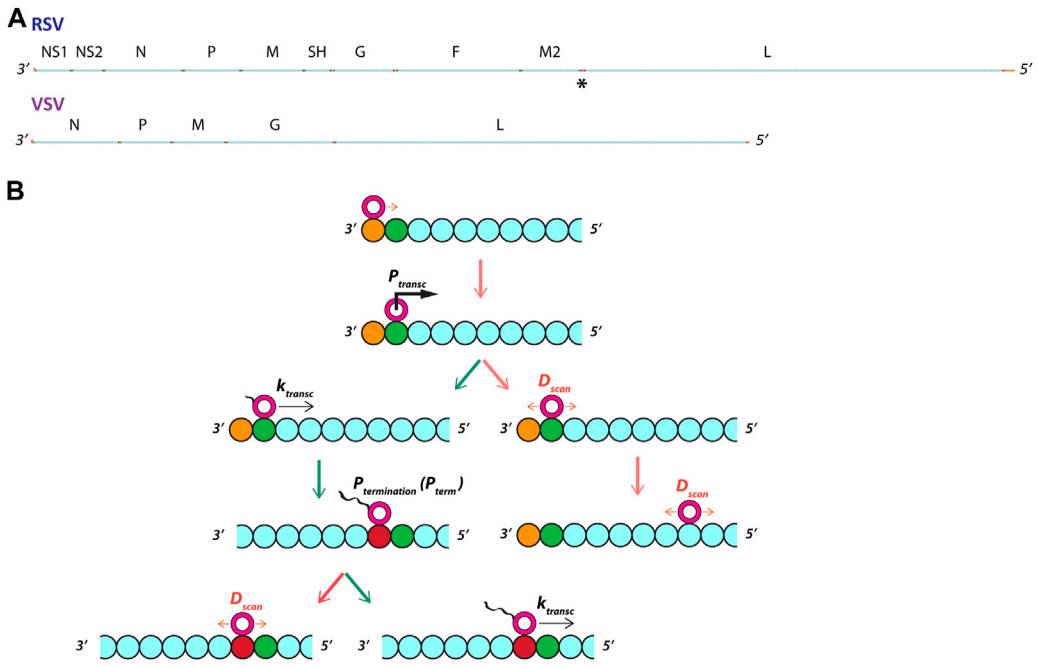

FIGURE 1
The model: linear respiratory syncytial virus (RSV) and vesicular stomatitis virus (VSV) genomes support the stochastic initiation and termination of transcription by a diffusing viral RNA-dependent RNA polymerase (pol). **(A)** The genetic structure of RSV and VSV genomes. The modeled RSV genome is 15,222 nt long and contains 10 ORFs with 8 gene junctions and a single short region (68 nt) of overlapping ORFs between genes M2 and L (see black asterisk). The modeled VSV genome is 11,152 nt long and contains 5 ORFs with 4 gene junctions. The genomes were divided into chunks approximating the size of a pol footprint (28 nts). Most of each genome is coding sequence (represented as cyan beads). **(B)** Essential model phenomena and parameters. A single RNA-dependent RNA polymerase (pol) starts an unbiased random walk at a rate D_{scan} (= 1 genomic chunk per event) at the most 3′ chunk (depicted as a burnt orange bead) of the modeled genome. Transcription initiation occurs with a probability P_{transc} when a pol diffuses onto a genomic chunk containing a gene start (GS) signal (depicted as a green bead). If transcription is not initiated, the unbiased random walk (i.e., diffusion) resumes. If transcription is initiated, the modeled pol state changes and the pol starts translocating 5′ down the genome at a rate k_{transc} (= x genomic chunks per event). Transcription termination occurs with a probability P_{term} when a transcribing pol translocates onto a genomic chunk containing a gene end (GE) signal (depicted as a red bead). If termination occurs, the pol state changes back to non-transcribing and resumes diffusion along the genome at a rate D_{scan}; if termination does not occur, the pol 'reads through' the GE signal and continues transcribing into the next ORF. (Cyan beads represent coding sequence).

TABLE 1 Gene start (GS) signal sequence-based constraints on RSV transcription initiation probabilities (P_{transc}) from Kuo et al. Most RSV/A genomes contain three different GS signal sequences. Kuo et al. performed minigenome studies to quantify the effects on gene expression of all single nt mutations within the GS signal (Kuo et al., 1997). The G gene GS signal contains a single mutation (relative to the most common GS signal sequence) at position 10 that reduced gene expression by ~35%. Kuo et al. reported that the L gene GS signal gave rise to a magnitude of gene expression equal to that of the most common GS signal. It is therefore reasonable to model RSV transcription with a single probability of transcription initiation at all GS signals except for G, where the probability should be multiplied by 0.65.

RSV gene			
NS1,NS2,N,P,M,SH,F,M2		G	L
GS signal sequence	CCCCGUUAU	CCCGUUUA**C**	CCC**U**GUUUU**A**
Effect on P_{transc}	--	0.65_X	1_X (no change)

time, experimentally observed non-gradient RSV and gradient VSV gene expression patterns.

Methods

The model

Computational models of RSV and VSV transcription were written in the Python programming language using the free and open-source Scientific Python Development Environment (Spyder version 3.3.2). The model code is freely available on GitHub: https://github.com/BCM-GCID/Publications/tree/main/Rethinking_NNSV_Gene_Expression.

In brief, the models simulate one or more viral RNA-dependent RNA polymerases (pols) entering a linear RSV or VSV genome at the 3' end and taking a random walk at a rate D_{scan} (units = "genomic chunks" per simulated event; D_{scan} = 1 throughout the results presented in this MS). A random walk is a simple model of diffusion where a simulated pol moves either one genomic chunk 5' or 3' along the genome. A parameter D_{bias} is used as a multiplicative factor (D_{scan}*D_{bias}) to 5' bias (or not) the random walk taken by modeled pols—i.e., $D_{bias} > 1$ biases pol movement 5'; $D_{bias} = 1$ results in an unbiased random walk. Each genome is divided into chunks of a size thought to reasonably approximate the footprint of a single RSV or VSV pol (28, 14, or 7 nt). Diffusing non-transcribing pols cannot "hop" over other pols and a single genomic chunk can only be occupied by a single pol at any one time.

Gene start (GS) and gene end (GE) signal sequences are modeled as separate genomic chunks positioned along the modeled genomes according to their known positions from sequencing data (Figure 1A). Transcription is initiated with a data-constrained probability (see Table 1) when a non-transcribing pol (pol_state = 0) diffuses onto a GS signal; termination of transcription or transcriptional readthrough occurs with a probability derived from published sequencing data when a transcribing pol (pol_state = 1) moving 5' at a rate k_{transc} (units = "genomic chunks" per simulated event) translocates onto a GE signal (Figure 1B). Initiations of transcription and transcriptional readthrough events are counted as gene expression events for the genes where they occur. For simulations incorporating multiple pols on a single genome, ejections of a non-

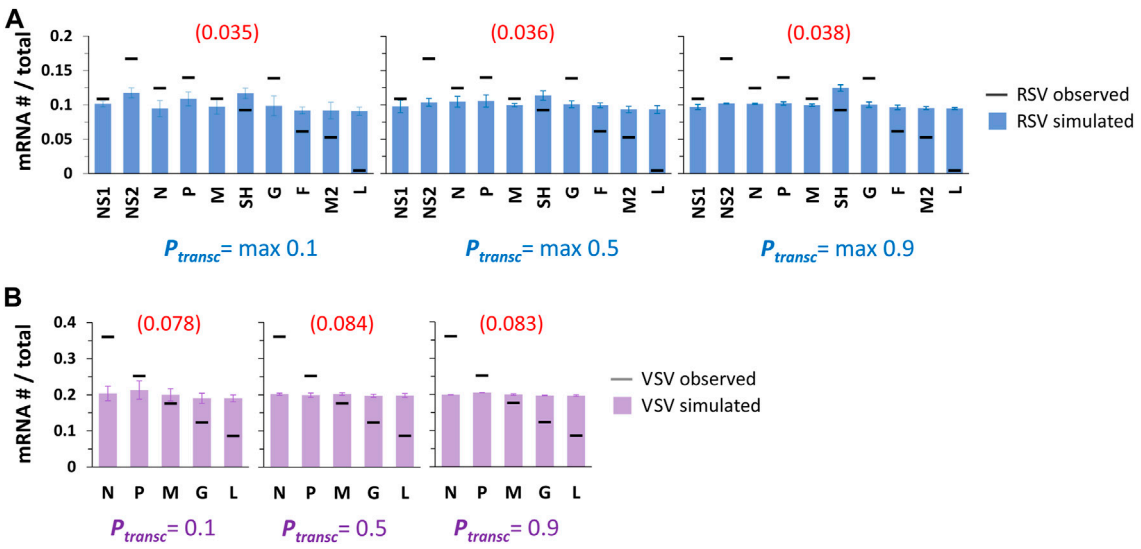

FIGURE 2
Single pol simulations produce flat patterns of gene expression across P_{transc} values tested. **(A)** Simulated RSV transcription. Histograms of mRNA # for each RSV gene divided by the total mRNA # show uniform gene expression across the 10 genes for all three sets of P_{transc} tested (max 0.1, max 0.5, and max 0.9). For each set of $P_{transc,}$ the max value equals the probability of transcription at every GS signal except for that of the G gene, which equals 0.65*max. Blue bars depict results from simulations; black horizontal bars depict average published experimentally observed values (Rajan et al., 2022). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above each histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations. **(B)** Simulated VSV transcription. Histograms of mRNA # for each VSV gene divided by the total mRNA # show uniform gene expression across the 5 genes for all three sets of P_{transc} tested (0.1, 0.5, and 0.9). For each set of $P_{transc,}$ the probability of transcription is the same at every GS signal. Lavender bars depict results from simulations; black horizontal bars depict average published experimentally observed values (Iverson and Rose, 1981). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above each histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations.

10.3389/fmolb.2022.1095193

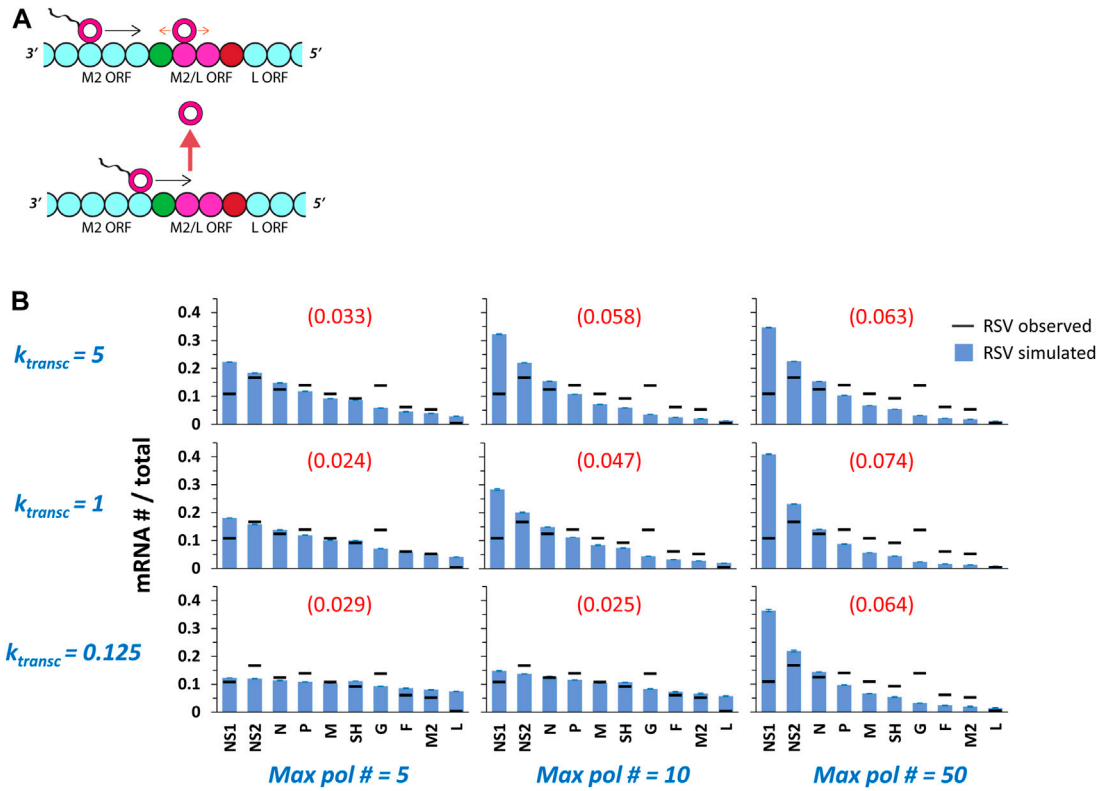

FIGURE 3
Multiple pols on a single genome undergoing ejective collisions between transcribing and non-transcribing pols produce gene expression gradients of increasing steepness with increasing 5′ translocation rate (k_{transc}) and increasing *maximum pol number (max pol #)*. **(A)** The M2/L overlap in ORFs. The final two genes of the RSV genome, M2 (which encodes both a transcription processivity factor and a regulatory factor that enhances replication) and L (which encodes the polymerase), share a 68 nt stretch (approximately two genomic chunks of 28 nts each—depicted as magenta beads) of ORF. This ORF overlap should be a hotspot for collisions between transcribing pols and non-transcribing pols diffusing in the neighborhood of the M2 GE signal (shown as red bead). The L gene GS signal is depicted as a green bead. **(B)** RSV gene expression patterns over a range of k_{transc} and *max pol #*. The parameter k_{transc} sets the rate at which transcribing pols move 5′ down the genome (units = genomic chunks per simulated event) and the parameter *max pol #* sets the maximum number of pols allowed on the genome at one time. Simulations of RSV transcription were performed at three different values of k_{transc} x three different values of *max pol #*. Histograms of mRNA # for each RSV gene divided by the total mRNA # depict results from the simulations (blue bars) and average published experimentally observed values (black horizontal bars) (Rajan et al., 2022). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above each histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations.

transcribing pol occur when a transcribing pol passes it. When a pol reaches the extreme 5' end of a modeled genome, it either diffuses 3' or dissociates from the genome.

The simulations occur one event at a time (i.e., time is modeled implicitly) whereby the positions and states (non-transcribing or transcribing) of the one or more modeled pols is stochastically updated according to the rules outlined above before proceeding to the next event. After simulating 10 s of thousands of events, each gene's mRNA level divided by the total mRNA level is outputted. These data are plotted to visualize a gene expression pattern.

Results and discussion

Determining the effects of stochastic transcription using a range of initiation probabilities

We took a heuristic approach to fitting actual observations of RSV and VSV gene expression and started by modeling a single pol taking an unbiased random walk down either genome and stochastically initiating and terminating transcription (Figures 1A,B).

In this simple case, the parameters to explore are probabilities of transcription initiation and termination. The termination probabilities can be derived directly from published sequencing data for RSV, as these are simply the complement of the published readthrough rates (Rajan et al., 2022). For VSV, we made use of estimates suggesting a very high probability of termination (0.99) for the GE signals modeled here (Barr et al., 1997). In contrast with termination probabilities, probabilities of transcription initiation are completely unknown. However, the three GS signals of the RSV genome modeled here have been tested in minigenomes for their relative strength of gene expression (Kuo et al., 1997). These relative strengths were used to constrain the ten transcription initiation probabilities of RSV (Table 1). The five GS signals of the VSV genome modeled here were all assumed to support an equal probability of transcription initiation.

Simulated patterns of RSV and VSV gene expression were essentially flat for all three sets of transcription probabilities (Figure 2). Standard deviations of individual mRNA levels were, as expected, highest for the lowest transcription probabilities tested

10.3389/fmolb.2022.1095193

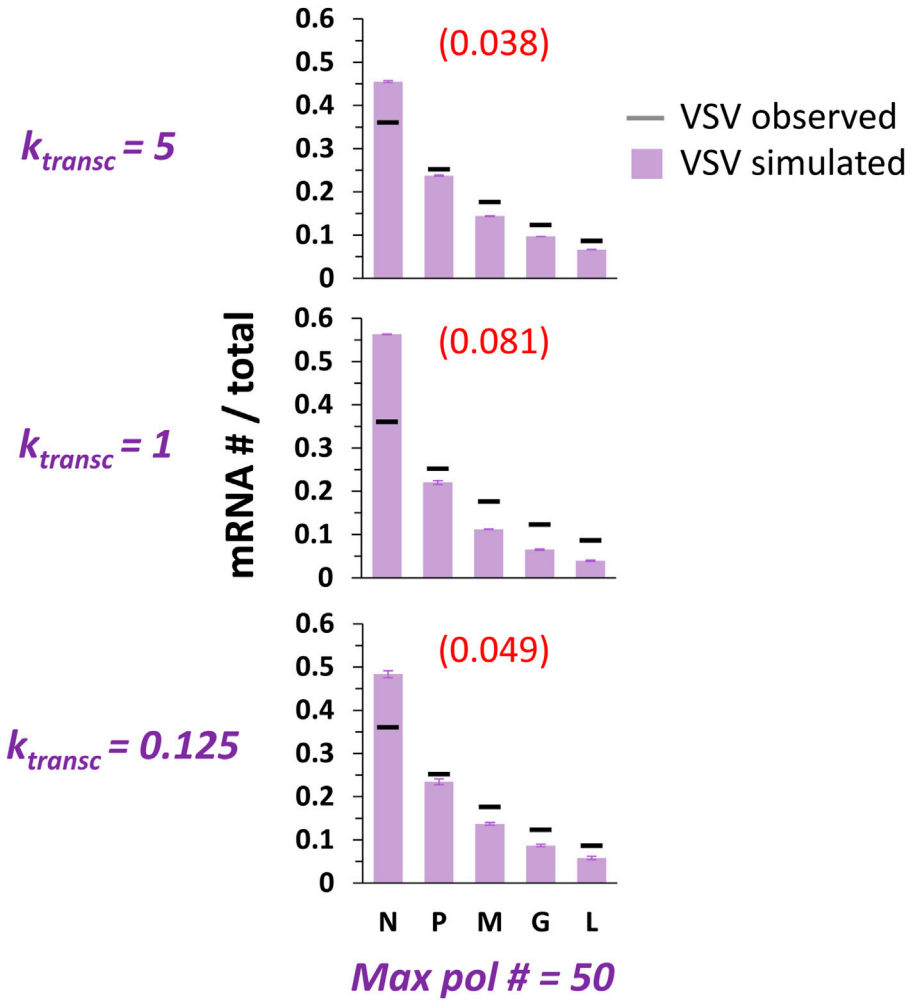

FIGURE 4
Simulations of at most 50 pols and collision-based pol ejections fit benchmark observations of VSV gene expression best at the highest k_{transc} tested. VSV gene expression patterns over a range of k_{transc} and a single *max pol #*. The parameter k_{transc} sets the rate at which transcribing pols move 5′ down the genome (units = genomic chunks per simulated event) and the parameter *max pol #* sets the maximum number of pols allowed on the genome at one time. Simulations of VSV transcription were performed at three different values of k_{transc}. Histograms of mRNA # for each VSV gene divided by the total mRNA # depict results from the simulations (lavender bars) and average published experimentally observed values (black horizontal bars) (Iverson and Rose, 1981). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above each histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations.

(Figure 2). In the case of RSV, a slight bump in gene expression occurs for the SH gene and becomes most visible at the highest transcription probabilities tested (Figure 2A). This is because of the lower rate of transcription initiation at the G gene GS signal (0.65x), which is directly downstream of the SH gene: the modeled pol occasionally fails to initiate transcription at the G gene before diffusing to the nearest GS signal, SH, where it is ~1.5x more likely to initiate transcription. We also calculated a root-mean-square deviation (RMSD) for each simulated gene expression pattern to quantify how well the model fit the observed *in vitro* gene expression patterns (Figures 2B,D).

Incorporating multiple polymerases into our model of NNSV transcription

Modeling a single pol diffusing along an RSV or VSV genome and stochastically starting and stopping transcription with the sequence-based probabilities used here cannot capture experimentally observed gene expression patterns. It is also well established that VSV virions contain 10 s of pols per genome (Thomas et al., 1985), making it very likely that both VSV replication and transcription involve multiple pols interacting with a single genome.

Thus, we decided to model multiple pols interacting with and transcribing single RSV and VSV genomes. This required conceiving of rules to govern interactions between the pols interacting with a single genome. We decided to implement one-by-one pol entry at the 3′ end of the genome, a variable maximum number of pols interacting with the genome at any one time, “soft” collisions between non-transcribing pols that prevent one pol from “hopping over” another, and hard collisions between 5′ translocating transcribing pols and diffusing non-transcribing pols resulting in the latter’s ejection from the genome.

The latter rule was partly inspired by observing that the steepest drop in RSV gene expression, a dramatic decrease reported by multiple

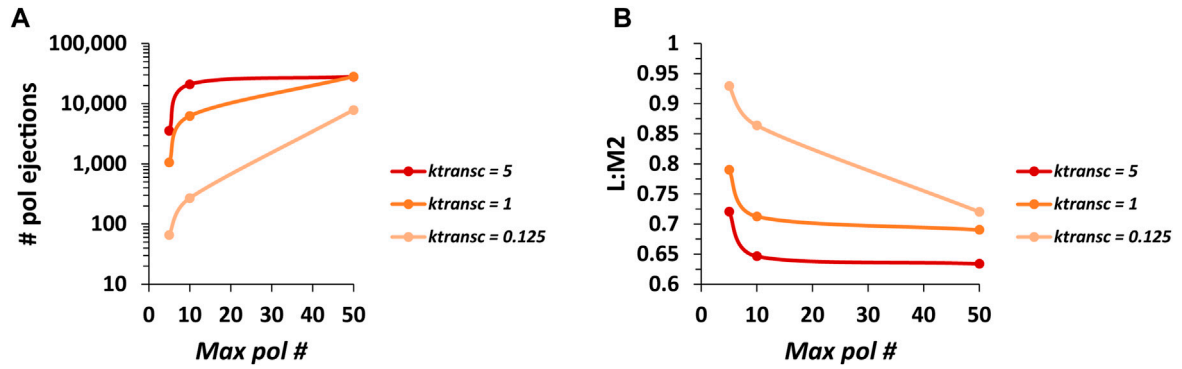

FIGURE 5
The number of pol ejections occurring over one run of the model and the ratio of RSV L mRNA to M2 mRNA levels (L:M2) produced are inversely related. **(A)** The number of pol ejections vs. *max pol #* for three different values of k_{transc}. Each data point is the average of three 100,000 event simulations. **(B)** The ratio of L mRNA to M2 mRNA levels (L:M2) vs. *max pol #* for three different values of k_{transc}. Each data point is the average of three 100,000 event simulations.

independent groups (Krempl et al., 2002; Aljabr et al., 2016; Levitz et al., 2017; Donovan-Banfield et al., 2022; Rajan et al., 2022), occurs over what should be a hot-spot for collisions between transcribing and non-transcribing pols: the overlap in the M2 and L gene ORFs (Figure 3A). We also took inspiration from work by Tang et al. (2014) reporting a very high affinity of VSV pols for the VSV ribonucleoprotein (RNP) complex and suggesting, through computational modeling, the importance of a class of ejective pol-pol collisions somewhat different from the class modeled here. Specifically, here we model two pol states—non-transcribing, which diffuse bidirectionally; and transcribing, which move only 5'—for pols that have gained access to the RNA genome through the 3' promoter; in contrast, Tang et al. modeled ejective collisions between pols that have accessed the RNA genome *via* the 3' promoter and pols "scanning" the VSV RNP complex *via* interactions between pol-bound P protein and N protein for the 3' promoter. We make no attempt to model the "scanning" pols that have yet to access the RNA genome of (Tang et al., 2014).

Because our model was modified to include multiple pols undergoing ejective collisions between transcribing and non-transcribing pols, it was necessary to explore another parameter, k_{transc}, setting the 5' translocation speed of a transcribing pol. We simulated RSV transcription under three different values each of k_{transc} and maximum pol number (Figure 3B), and VSV transcription under three different values of k_{transc} and a single maximum pol number (Figure 4). A single maximum pol number was used for VSV transcription because of published work suggesting approximately 50 VSV pols per VSV genome (Thomas et al., 1985); to our knowledge, this ratio is not known for RSV.

Simulated RSV gene expression patterns display a 3' to 5' gradient of increasing steepness with increasing maximum pol number and, for simulations with a maximum of 5 and 10 pols, with increasing k_{transc} (Figure 3B). The transcription gradient in our model is a consequence of a gradient in pol concentration emerging from ejective pol-pol collisions and obligatory pol reentry at the 3' end of the genome. In the case of simulations of at most 50 pols, the gene expression gradient is steepest at the middle value of k_{transc} because the higher value supports such a high frequency of ejective pol collisions that the actual number of modeled pols occupying a genome at steady-state tends to ~10, while the middle value leads to one of ~20 pols, which leads to a sharper pol concentration gradient along the genome and a steeper gene expression gradient. It is clear from both the calculated RMSD values and visually inspecting the fits that simulations incorporating a high maximum number of RSV pols per genome produce a gene expression pattern that is too steeply gradient; in contrast, simulations of at most 5 RSV pols per genome yield much better fits of the published data across the 20-fold range of k_{transc} values tested (Figure 3B).

We simulated VSV gene expression across the same 20-fold range of k_{transc} values and only one value of maximum pol number (Figure 4). At the highest value of k_{transc} tested, the model captures benchmark observations of VSV transcription fairly well. It is interesting that the middle value of k_{transc} results in the worst fit of the data; this results from the phenomenon described above for RSV transcription under the same maximum pol number: the highest value of k_{transc} tested leads to such a high frequency of pol collisions that the actual number of pols occupying the genome at steady-state is much lower than the maximum possible; because the lower value of k_{transc} leads to less frequent collisions and a concomitant increase in the number of pols occupying the genome, a steeper gene expression gradient results (Figure 4).

Further exploring the effects of collision-based pol ejections on RSV transcription

Thus, our simple model incorporating multiple pols undergoing random diffusion along the genome when not transcribing and ejective collisions when a transcribing and non-transcribing pol meet captures benchmark observations of VSV gene expression (Iverson and Rose, 1981) well while poorly fitting our published observations of RSV gene expression (Rajan et al., 2022). Furthermore, the model most poorly fits data coming from the last two genes of the RSV genome, where multiple groups report a dramatic decrease in gene expression. This is the sole region of the modeled genomes where two ORFs overlap; and this overlap helped inspire the addition of ejective pol collisions into our model. We decided to further investigate the effect of the modeled pol collisions on gene expression over the M2-L region of RSV by analyzing the relationships between 1) the number of pol ejections per run of our simulation and values of the maximum pol number and k_{transc}; and 2)

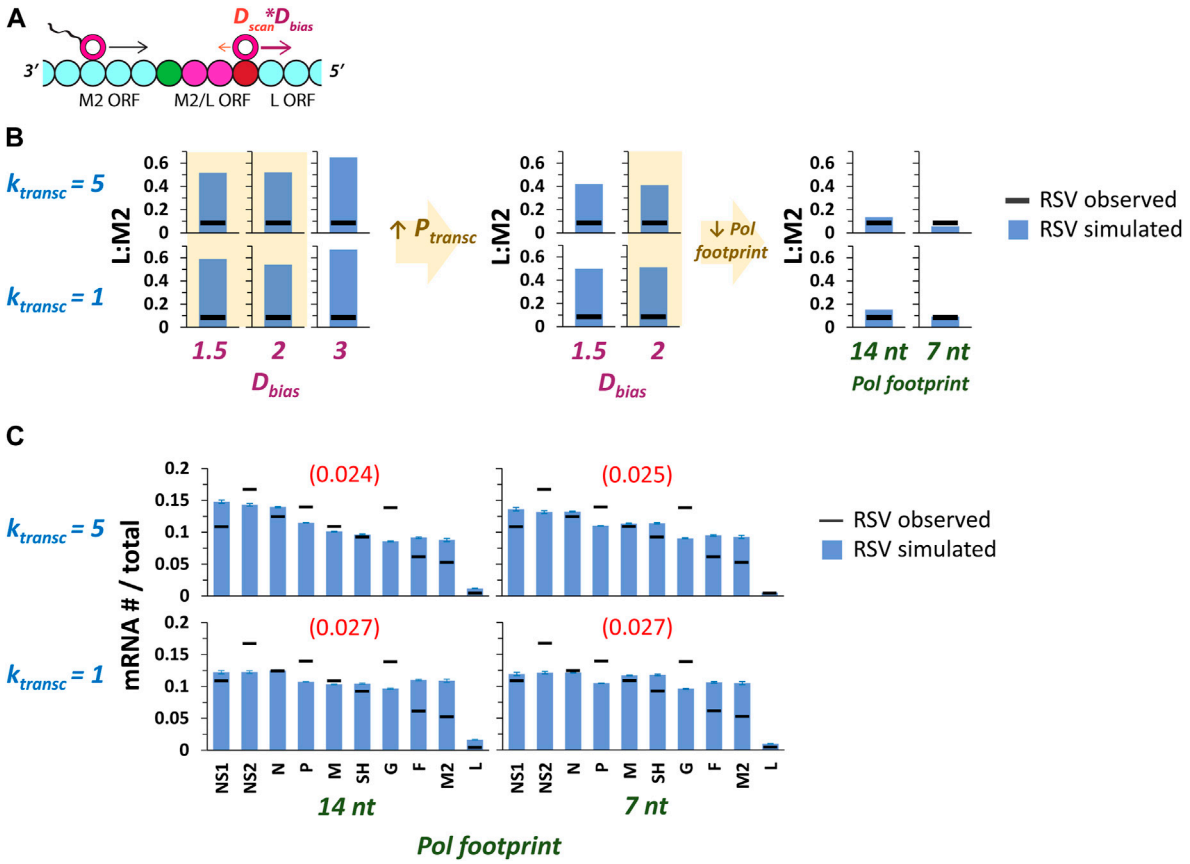

FIGURE 6
Including a 5′ pol diffusion bias (D_{bias}) can reproduce the experimentally observed drop in gene expression between RSV genes M2 and L. **(A)** D_{bias} is expected to have its largest effect on ratios of L mRNA to M2 mRNA levels (L:M2). M2 and L ORFs are depicted as cyan beads (bead size reflects a pol footprint size of 28 nt); the M2/L overlap is depicted as two magenta beads; the L GS signal is depicted as a green bead; the M2 GE signal is depicted as a red bead. **(B)** A scan of parameter values shows that moderate D_{bias} with reduced pol footprint size can reproduce the experimentally observed value of L:M2 (Rajan et al., 2022). Left panel: histograms show simulated (blue bar) and experimentally observed (horizontal black bar) L:M2 values for two different values of k_{transc} x three different values of D_{bias} (*max pol #* = 5). The two lower values of D_{bias} tested (highlighted in pale yellow) result in a slightly greater drop in L:M2 than D_{bias} = 3; these values were used in subsequent simulations. Middle panel: histograms show simulated (blue bar) and experimentally observed (horizontal black bar) L:M2 values for two different values of k_{transc} x two different values of D_{bias} under conditions of increased P_{transc} (= max of 0.9). As predicted, an increased P_{transc} resulted in a further decreased L:M2. Simulations with D_{bias} = 2 (results highlighted in pale yellow) were chosen for subsequent simulations. Right panel: histograms show simulated (blue bar) and experimentally observed (horizontal black bar) L:M2 values for two different values of k_{transc} x two different pol footprint sizes (14 and 7 nt) and D_{bias} = 2. A decreased pol footprint size increases the effective distance between the M2 GE and L GS signals, and results in simulated levels of L:M2 that closely match experimental observations. Each data point is the average of three 100,000 event simulations. **(C)** Global fits of the RSV gene expression data improve with the introduction of D_{bias} and reduced pol footprint size. Simulations of RSV transcription were performed at two different values of k_{transc} x two different values of *pol footprint* and D_{bias} = 2. Histograms of mRNA # for each RSV gene divided by the total mRNA # depict results from the simulations (blue bars) and average published experimentally observed values (black horizontal bars) (Rajan et al., 2022). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above each histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations.

ratios of L and M2 mRNA levels and values of the maximum pol number and k_{transc} (Figure 5).

The average number of pol ejections generally increases with increasing max pol # and increasing k_{transc} (Figure 5A). However, at the higher values of k_{transc} (1 and 5 genomic chunks per event), the average number of pol ejections starts to plateau beyond a max pol # of 10. This again shows that the steady-state number of pols bound to a genome in the model depends on k_{transc} and that this number is close to 10 at the highest value of k_{transc} tested (assuming a pol footprint of 28 nt). In addition, curves for the higher values of k_{transc} start to converge, suggesting that the model is reaching its maximum pol ejection frequency.

Consistent with the expected effect of the modeled pol collisions, ratios of L:M2 generally decrease with increasing max pol # and increasing k_{transc} (Figure 5B). However, L:M2 plateaus sharply for the higher values of k_{transc} tested beyond a max pol # of 10. In addition, curves for higher values of k_{transc} start to converge, suggesting that the model is reaching its minimum L:M2 which remains much higher than the average experimentally observed value of ~0.09 (Rajan et al., 2022). This suggested that the RSV model was missing something of fundamental importance.

Modifying the model to include 5′ biased diffusion of non-transcribing pols

We therefore decided to test the effects of including biased pol diffusion in our model, specifically a 5′ bias, which might help explain

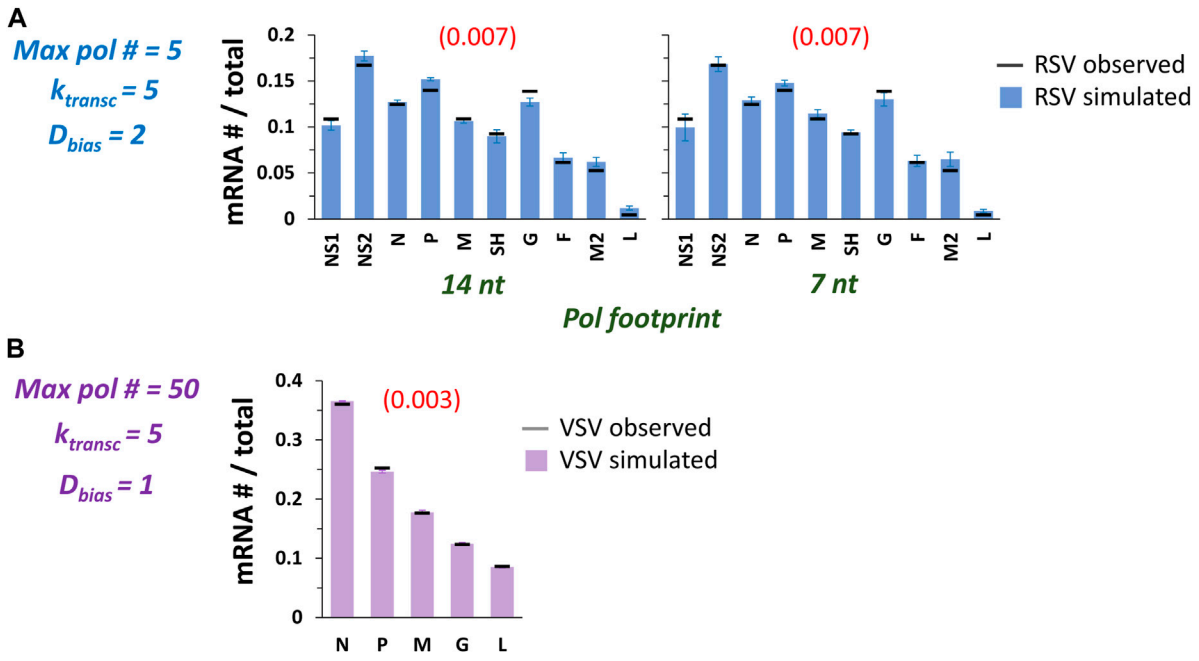

FIGURE 7
The model captures published observations of RSV and VSV transcription with adjustments to the underlying transcription probabilities (P_{transc}). **(A)** High quality fits of experimentally observed RSV gene expression patterns. P_{transc} were manually adjusted to achieve optimized fits at *max pol #* = 5, k_{transc} = 5, D_{bias} = 2, and *pol footprint* of 14 and 7 nt. Histograms of mRNA # for each RSV gene divided by the total mRNA # depict results from the simulations (blue bars) and average published experimentally observed values (black horizontal bars) (Rajan et al., 2022). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above each histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations. **(B)** A high quality fit of the benchmark experimentally observed VSV gene expression pattern. P_{transc} were manually adjusted to achieve an optimized fit at *max pol #* = 50, k_{transc} = 5, D_{bias} = 1 (i.e., NO 5′ bias), and *pol footprint* = 28 nt. The histogram of mRNA # for each VSV gene divided by the total mRNA # depicts results from the simulations (lavender bars) and average published experimentally observed values (black horizontal bars) (Iverson and Rose, 1981). Each data point is the average of three 100,000 event simulations; error bars show the standard deviation. The number in parentheses and red above the histogram is the root-mean-square deviation (RMSD) of the simulated gene expression pattern from the experimental observations.

TABLE 2 Major parameter values for high quality fits of different data sets by our model. Our model produces high quality fits of two different RSV data sets (Donovan-Banfield et al., 2022; Rajan et al., 2022) and benchmark observations of VSV gene expression (Iverson and Rose, 1981). Each list of transcription initiation probabilities (P_{transc}) contains the values used for every RSV or VSV GS signal following their 3′ to 5′ order along the genome.

Data set	Maxpol #	Pol footprint size(nt)	K_{transc}	D_{bias}	P_{transc}	RMSD
RSV piedra et al.	5	7	5	2	0.25,0.9,0.4,0.8,0.45,0.3,0.99,0.2,0.3,0.1	0.007
RSV piedra et al.	5	14	5	2	0.25,0.9,0.4,0.9,0.4,0.3,0.99,0.3,0.3,0.1	0.007
RSV Banfield et al.	15	7	5	2	0.9,0.4,0.5,0.8,0.9,0.4,0.5,0.7,0.9,0.9	0.013
VSV Iverson & Rose;	50	28	5	1	0.1,0.6,0.55,0.45,0.37	0.013

why gene expression is possible but falls off steeply when a pol must diffuse 3′ to reach the nearest GS signal after terminating transcription (as occurs in RSV from the M2-L overlap) (Figure 6A). The biased diffusion of proteins has been shown before (Ricchetti et al., 1988; Kwok et al., 2006; Powers et al., 2009), making this change to the model biophysically reasonable.

A 5′ pol diffusion bias was modeled by including a new parameter in the model, D_{Bias}, with a value used as a multiplicative factor for 5′ diffusion only (Figure 6A). Thus, a D_{Bias} value of 2 would result in a pol moving two steps (genomic chunks) with every 5′ movement while moving only one step (assuming D_{scan} = 1) with every 3′ movement; the probabilities of moving in either direction remain equal. This change could also be modeled by modifying the probabilities of 5′ vs. 3′ pol translocation and keeping each step size the same.

In order to test the effects of 5′ biased pol diffusion on gene expression in our model, we chose two of the parameter sets yielding fits with lower RMSDs from our first set of RSV transcription simulations involving multiple pols (Figure 3B), ran these with three different values of D_{Bias}, and looked for a drop in the predicted value of L:M2 mRNA levels (Figure 6B). The two lower values of D_{Bias} tested produced a greater drop in L:M2 than the highest value tested (Figure 6B). This is because under a maximum transcription initiation probability of 0.5, a high 5′ D_{Bias} leads to frequent “missing” of the M2 GS signal before transcription initiation

at the L GS signal. We therefore decided to increase the maximum transcription initiation probability to 0.9 and reran simulations at the lower values of 5' D_{Bias} tested. As expected, this resulted in a further drop in predicted L:M2 mRNA levels. However, simulated L:M2 levels remained much higher than our experimentally observed value (Figure 6B). Finally, we decided to decrease the pol footprint size by factors of 2 and 4, separately, knowing that this would increase the effective distance between the M2 GE signal and the L GS signal, and predicting a drop in L:M2 levels. The smallest pol footprint size tested, seven nts, is equal to the number of nucleotides bound by a single subunit of RSV nucleoprotein (N protein) and only three nts less than the size of the highly conserved RSV GS signal. Decreasing the pol footprint size yielded predicted L:M2 values that are very close to the experimentally observed value (Figure 6B); and global fits of the RSV gene expression data quantitatively improved for the higher value of k_{transc} tested and remained roughly the same for the lower value (Figure 6C).

Optimizing model fits

With the addition of D_{Bias} to our model of RSV transcription, it seemed that both RSV and VSV versions of the model were poised to capture experimentally observed patterns of gene expression. We therefore set about finding RSV and VSV transcription initiation probabilities that would produce optimal fits of the experimental data (Figure 7A, B). Using a set of transcription probabilities spanning a 10-fold range of values for a maximum pol number of 5, our RSV model yielded high quality fits of our experimental data (Figure 7A; Table 2). Our VSV model yielded a high quality fit of the experimental data with a set of transcription probabilities spanning a 6-fold range and a maximum pol number of 50 (Figure 7B; Table 2).

We also decided to fit the recently reported RSV long-read sequencing data of Donovan-Banfield et al., 2022. An increased *max pol #* and an approximately 2-fold range of P_{transc} were needed to capture their data (Table 2). These changes reflect the more gradient nature of the observed gene expression pattern, while our experimental observations showed much higher levels of G gene mRNA (Rajan et al., 2022).

A 5' diffusion bias was needed to capture both RSV data sets because of a common dramatic decrease in expression between genes M2 and L. In contrast, a 5' diffusion bias was not needed to capture the benchmark observations of VSV gene expression used here; however, including one has minimal effect on the model's output (data not shown). Thus, we simply cannot make a model-supported prediction about whether non-transcribing VSV pols diffuse with a 5' bias. Continuing with VSV, the high quality fit we report involves a 6-fold range of P_{transc}, but a quality fit can also be obtained with a 5-fold range of P_{transc} and less variation (= 0.1, 0.5, 0.5, 0.5, 0.5; RMSD = 0.009).

We believe the changes to transcription probabilities needed to produce high quality fits of the experimental data are reasonable. For instance, we have obtained preliminary data using RSV minigenomes encoding luciferase reporter genes showing that a single RSV GS signal sequence can support a 1.5-fold range of gene expression according to its alignment with bound nucleoprotein or *N-phase* (Piedra et al., 2020b). We do not know whether the reported N-phase-mediated changes to gene expression are exactly proportional to the changes in microscopic probabilities of transcription initiation modeled here because the former come from luciferase activity measurements and therefore reflect the addition of translation. Moreover, sequence changes outside of the highly conserved 10 nt stretch of the RSV GS signal can lead to gene expression changes (Kuo et al., 1997), and the VSV GS signal is less conserved than RSV's. However, we are not aware of minigenome studies exploring the effects of VSV GS signal sequence or N-phase on gene expression. Finally, it is also possible that the shape of the observed RSV and VSV gene expression patterns depends partly on differences in the underlying mRNA stabilities, which we make no attempt to model here; but we have shown previously that any such differences are unlikely to significantly affect experimentally observed RSV gene expression patterns (Piedra et al., 2020a). It is also worth mentioning that we make no attempt to model the potential effects of 1) variable nascent mRNA capping efficiency and 2) mRNA polyadenylation. Both could be modeled as a variable pause time at the start and end of transcription, respectively. However, we do not believe their inclusion would change the major results presented here.

Conclusion and limitations

Our model can capture observed RSV and VSV transcription patterns with biophysically reasonable parameters and parameter values. Our model makes the following major predictions in need of wet lab experimental testing: 1) ejective collisions occur between transcribing and non-transcribing NNSV pols; 2) non-transcribing RSV pols (and perhaps VSV pols) undergo 5' biased diffusion along the viral genome; and 3) an increase in the number of pols bound to and diffusing along an NNSV genome at any one time will lead to more frequent pol-pol collisions and a sharper transcription gradient. Sophisticated single molecule TIRF-based assays are needed to directly test predictions 1-2, while 3 can be tested using established minigenome or recombinant genome assays along with high throughput sequencing.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author.

Author contributions

FP wrote the MS and performed all work therein. Remaining authors reviewed the MS, offered critical commentary, played essential roles in obtaining the RSV sequencing data used, and are, along with FP, members of the same consortium: the Texas Medical Center Genomic Center for Infectious Diseases (NIH Grant # U19AI144297).

Acknowledgments

Thanks to PP and the Texas Medical Center Genomic Center for Infectious Diseases (TMC-GCID; NIH Grant # U19AI144297) for supporting this work. Thanks also to VM of Baylor College of Medicine's (BCM) Human Genome Sequencing Center (HGSC) and SJ of BCM's Alkek Center for Metagenomics and Microbiome

Research (CMMR) for uploading the manuscript's code to GitHub as one of multiple projects within the TMC-GCID.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aljabr, W., Touzelet, O., Pollakis, G., Wu, W., Munday, D. C., Hughes, M., et al. (2016). Investigating the influence of ribavirin on human respiratory syncytial virus RNA synthesis by using a high-resolution transcriptome sequencing approach. *J. Virol.* 90 (10), 4876–4888. doi:10.1128/JVI.02349-15

Barik, S. (1993). The structure of the 5' terminal cap of the respiratory syncytial virus mRNA. *J. Gen. Virol.* 74 , 485–490. doi:10.1099/0022-1317-74-3-485

Barr, J. N., Tang, X., Hinzman, E., Shen, R., and Wertz, G. W. (2008). The VSV polymerase can initiate at mRNA start sites located either up or downstream of a transcription termination signal but size of the intervening intergenic region affects efficiency of initiation. *Virology* 374 (2), 361–370. doi:10.1016/j.virol.2007.12.023

Barr, J. N., Whelan, S. P., and Wertz, G. W. (1997). cis-Acting signals involved in termination of vesicular stomatitis virus mRNA synthesis include the conserved AUAC and the U7 signal for polyadenylation. *J. Virol.* 71 (11), 8718–8725. doi:10.1128/JVI.71.11.8718-8725.1997

Brauburger, K., Boehmann, Y., Krahling, V., and Muhlberger, E. (2016). Transcriptional regulation in Ebola virus: Effects of gene border structure and regulatory elements on gene expression and polymerase scanning behavior. *J. Virol.* 90 (4), 1898–1909. doi:10.1128/JVI.02341-15

Donovan-Banfield, I., Milligan, R., Hall, S., Gao, T., Murphy, E., Li, J., et al. (2022). Direct RNA sequencing of respiratory syncytial virus infected human cells generates a detailed overview of RSV polycistronic mRNA and transcript abundance. *PLoS One* 17 (11), e0276697. doi:10.1371/journal.pone.0276697

Fearns, R., and Collins, P. L. (1999). Model for polymerase access to the overlapped L gene of respiratory syncytial virus. *J. Virol.* 73 (1), 388–397. doi:10.1128/JVI.73.1.388-397.1999

Iverson, L. E., and Rose, J. K. (1981). Localized attenuation and discontinuous synthesis during vesicular stomatitis virus transcription. *Cell* 23 (2), 477–484. doi:10.1016/0092-8674(81)90143-4

Kolakofsky, D., Le Mercier, P., Iseni, F., and Garcin, D. (2004). Viral DNA polymerase scanning and the gymnastics of Sendai virus RNA synthesis. *Virology* 318 (2), 463–473. doi:10.1016/j.virol.2003.10.031

Krempl, C., Murphy, B. R., and Collins, P. L. (2002). Recombinant respiratory syncytial virus with the G and F genes shifted to the promoter-proximal positions. *J. Virol.* 76 (23), 11931–11942. doi:10.1128/jvi.76.23.11931-11942.2002

Kuo, L., Fearns, R., and Collins, P. L. (1997). Analysis of the gene start and gene end signals of human respiratory syncytial virus: Quasi-templated initiation at position 1 of the encoded mRNA. *J. Virol.* 71 (7), 4944–4953. doi:10.1128/JVI.71.7.4944-4953.1997

Kwok, B. H., Kapitein, L. C., Kim, J. H., Peterman, E. J., Schmidt, C. F., and Kapoor, T. M. (2006). Allosteric inhibition of kinesin-5 modulates its processive directional motility. *Nat. Chem. Biol.* 2 (9), 480–485. doi:10.1038/nchembio812

Levitz, R., Gao, Y., Dozmorov, I., Song, R., Wakeland, E. K., and Kahn, J. S. (2017). Distinct patterns of innate immune activation by clinical isolates of respiratory syncytial virus. *PLoS One* 12 (9), e0184318. doi:10.1371/journal.pone.0184318

Liuzzi, M., Mason, S. W., Cartier, M., Lawetz, C., McCollum, R. S., Dansereau, N., et al. (2005). Inhibitors of respiratory syncytial virus replication target cotranscriptional mRNA guanylylation by viral RNA-dependent RNA polymerase. *J. Virol.* 79 (20), 13105–13115. doi:10.1128/JVI.79.20.13105-13115.2005

Noton, S. L., and Fearns, R. (2015). Initiation and regulation of paramyxovirus transcription and replication. *Virology* 479-480, 545–554. doi:10.1016/j.virol.2015.01.014

Pagan, I., Holmes, E. C., and Simon-Loriere, E. (2012). Level of gene expression is a major determinant of protein evolution in the viral order Mononegavirales. *J. Virol.* 86 (9), 5253–5263. doi:10.1128/JVI.06050-11

Piedra, F. A., Qiu, X., Teng, M. N., Avadhanula, V., Machado, A. A., Kim, D. K., et al. (2020). Genotype-dependent and non-gradient patterns of RSV gene expression. *Plos one*, 15, e0227558, doi:10.1371/journal.pone.0227558

Piedra, F. A., Qiu, X., Teng, M. N., Avadhanula, V., Machado, A. A., Kim, D. K., et al. (2020). Non-gradient and genotype-dependent patterns of RSV gene expression. *PLoS One* 15 (1), e0227558. doi:10.1371/journal.pone.0227558

Powers, A. F., Franck, A. D., Gestaut, D. R., Cooper, J., Gracyzk, B., Wei, R. R., et al. (2009). The Ndc80 kinetochore complex forms load-bearing attachments to dynamic microtubule tips via biased diffusion. *Cell* 136 (5), 865–875. doi:10.1016/j.cell.2008.12.045

Rajan, A., Piedra, F. A., Aideyan, L., McBride, T., Robertson, M., Johnson, H. L., et al. (2022). Multiple respiratory syncytial virus (RSV) strains infecting HEp-2 and A549 cells reveal cell line-dependent differences in resistance to RSV infection. *J. Virol.* 96 (7), e0190421. doi:10.1128/jvi.01904-21

Ricchetti, M., Metzger, W., and Heumann, H. (1988). One-dimensional diffusion of *Escherichia coli* DNA-dependent RNA polymerase: A mechanism to facilitate promoter location. *Proc. Natl. Acad. Sci. U. S. A.* 85 (13), 4610–4614. doi:10.1073/pnas.85.13.4610

Tang, X., Bendjennat, M., and Saffarian, S. (2014). Vesicular stomatitis virus polymerase's strong affinity to its template suggests exotic transcription models. *PLoS Comput. Biol.* 10 (12), e1004004. doi:10.1371/journal.pcbi.1004004

Thomas, D., Newcomb, W. W., Brown, J. C., Wall, J. S., Hainfeld, J. F., Trus, B. L., et al. (1985). Mass and molecular composition of vesicular stomatitis virus: A scanning transmission electron microscopy analysis. *J. Virol.* 54 (2), 598–607. doi:10.1128/JVI.54.2.598-607.1985

Whelan, S. P., Barr, J. N., and Wertz, G. W. (2004). Transcription and replication of nonsegmented negative-strand RNA viruses. *Curr. Top. Microbiol. Immunol.* 283, 61–119. doi:10.1007/978-3-662-06099-5_3