**S1 Text. First two moments of the nascent RNA distribution**

 In order to connect mechanisms of transcription initiation with nascent RNA distributions, in the results section we described a model of transcriptional dynamics with an arbitrarily complex initiation mechanism followed by an elongation process. We make the following assumptions in our model (Supplementary Figure 1B): after initiation, each RNA polymerase molecule elongates along the gene by hopping from one to the next base at a constant probability per unit time. RNAP molecules do not pause and nor do they collide with each other, while moving along the gene. We also take the size of the RNAP footprint to be one base, and we do not restrict the number of RNAPs at each base along the gene. Here we present details of the calculations leading up to analytic formulas for the first two moments of the nascent RNA distribution predicted by this model.

To compute the first two moments of the nascent RNA distribution for arbitrary promoter architecture, we consider a promoter that can exist in *N* possible states. The rate of going from the *s-th* to the *q-th* promoter state is *ks,q*, and the rate at which an RNAP molecule initiates transcription from the *s*-*th* promoter state is *ks,ini*. After initiation every RNAP molecule elongates by hopping from one base to the next at a constant rate *k* and the number of nascent RNAs at site *i* is *mi*. If the number of bases along the gene is *L,* thenthe number of nascent RNAs along the gene will be given by

 (3)

 Here *mi*’s are dependent random variables since at each base the number of nascent RNAs (*mi*) depends on the number of nascent RNAs present on the previous base (*mi-*1). The state of the gene is described by (*L* + 1) stochastic variables: the number of nascent RNAs (*m1,…,mL*) at each base along the gene, and the state of the promoter, *s*. Thus, the probability distribution function describing the gene is given by *P*(*s*, *m1,…,mL* ). The time evolution for this probability can be calculated using the chemical master equation approach:

 (4)

Similar approaches have been used previously in order to find the moments of mRNA and protein distributions[1–4]

. By defining the probability vector, we can rewrite equation (4) in compact matrix form as

  (5)

Here we introduce the following matrices:, (which captures the transition between different promoter states) whose elements are, if *q≠s* and ;  is a matrix that contains the rates of initiation from different promoter states. In the case of one-step initiation it is diagonal with the diagonal elements being the rates of initiation from different promoter states i.e.  . In the case of two-step initiation this matrix is off-diagonal owing to the fact that the promoter state changes after initiation. For instance if the promoter switches from the *s*-th to the *q*-th state after initiation, then the off-diagonal term that exists is, . , which captures the hopping process of RNAP molecules from one base to the next**,** is diagonal and the corresponding matrix elements are. In the steady state the left hand side of equation (5) is equal to zero.

It is to be noted that for if we consider the termination rate of an RNAP molecule from the last base to be different than the hopping rate *k*. If the rate of termination is *kterm*, then the last term of equation (5) will be different and the equation will look like the following,



For the sake of simplicity we don’t incorporate the termination rate for the subsequent analysis. However, even if we explicitly use the termination rate, the analysis will remain the same.

**Mean of the nascent RNA distribution**

From equation (3), the mean number  of nascent RNAs along the gene is given by

 (7)

To compute the mean of the steady-state probability distribution of nascent RNAs along the gene, we multiply both sides of equation (5) by *mi*, and sum over all values of *mi* from 0 to . We obtain two separate equations for  and , where *i* goes from *2* to *L.*

For  we find:

 (7)

Since none of the three matrices in equation (7) (, and ) are functions of *m1*, they can be taken out of the sums. Simplifying the above equation further we find:

 (8)

The equation above can be expressed in terms of the following partial moment vectors:

 (9)

These partial moment vectors are useful quantitates, as they are related to the moments of the probability distribution of *m1*. For example, the mean number of nascent RNAs at the first base is given by:

  (10)

We simplify equation (8) and express it in terms of the partial moment vectors, defined above. In order to do that we re-arrange the third, sixth and seventh term using the fact that *mi*’s are dummy variables. We also use the fact that the number of nascent RNA molecules at different bases can never fall below 0 (i.e., where *i* goes from 1 to *L*)$\vec{p}\_{(-1)}=0)$. Using equation (8) and equation (9) we obtain equations for the two partial moment vectors:

 (11) We can compute the mean of *m1*, by multiplying equation (11) with :

  (12)

Here the vector  contains the ordered list of rates of transcription initiation from each promoter state. To compute the mean of the nascent RNA number at the other bases we multiply equation (5) by *mi ,* where *i* goes from *2* to *L*.

 (13)

As before, we define the following partials moment vectors for convenience:

  (14)

Following the same procedure as *m1*,we can find the mean of *mi* from equation (13), in terms of the partial moments (equation (14)) of *mi* and *mi-1*:

 (15)

We compute the mean of *mi*, by multiplying equation (15) with . The transition matrix has the property that the sum of the elements of any one of its columns is always 0. Using this property, we find that, . Therefore, we find that the mean nascent RNA number at each base is the same as the mean of the previous one, i.e.,

 (16)

Using equation (14) and equation (15) we get:

 (17)

In other words in the steady state, the mean number of nascent RNAs at every base along the gene is the same. Hence the mean nascent RNA number along the gene is the sum of the mean of nascent RNA numbers at each base. Therefore from equation (3), equation (11) and equation (16), we find that the mean is proportional to the length of the gene:

 (18)

For instance, using this formula above one can find the mean number of nascent RNAs along a gene which initiates transcription in uncorrelated events at a constant rate *r*. For such a gene the mean number of nascent RNAs is given  This result can be understood from a simple consideration. Namely, *L/k* is the mean time that an RNAP spends on the gene before completing transcription. Since *r* is the rate at which RNAPs enter the elongation state, the mean number of RNAPs on the gene at any time is the product of the two.

**Variance of the nascent RNA distribution**

 In order to compute the variance, we need to find all the elements of the covariance matrix of the dependent random variables, (*m1,…mL*). Using equation (3), we find that the second moment is:



We expand it for convenience:

 (19) To find , we need to compute each term on the right hand side using equation (5). As before we define the following partial moments:

 (20)

Here *i* and *j* both range from 1 to *L.* We get a set of six equations for the partial moments defined above by multiplying equation (5) with *m12*, *m1m2* , *m1mi* (where *i* goes from 3 to *L*)*, mi2*

(where *i* goes from 2 to *L*), *mimi+1* (where *i* goes from 2 to *L-1*) and *mimj* (where *i* goes from *2* to *L-2* and *j* runs from *i+2* to *L)*

 (21)

 (22)

 (23)

 (24)

 (25)  (26)

From equation (21-26), we can compute the second order moments for the nascent RNA numbers at every base along the gene by multiplying the partial moment vectors with . In order to obtain the variance, we construct the covariance matrix, whose element in the *i*, *j* position is the covariance between the random variables characterizing the number of nascent RNAs at the  *i*-th and *j*-th bases along the gene, defined as

 (27)

Here we use the fact that . The dimension of the covariance matrix is  and since the number of bases *L* along a gene is typically of the order of few thousands, the number of second order moments to be evaluated to construct the covariance matrix is very high. However, it is readily evaluated with Mathematica using equation (21-26) and equation (27). Therefore the variance of the number of nascent RNA molecules along a gene is given by

 (28)

which is just the sum of all the elements of the covariance matrix.

**Supplementary Figures**



Figure 1: (A) **Two state promoter with traffic**. The promoter switches between two states: state *1*, from which transcription initiation occurs with a constant probability per unit time *k1,ini*, and state 2, from which transcription does not initiate. The promoter switches from state 1to state 2with probability per unit time *k12*, and from 2to 1with probability per unit time *k21*. After initiation each RNAP molecule hops from one base pair to the next along the gene at a rate *k* per unit time. Each RNAP molecule has a finite DNA footprint of *30 bp* and it can pause at any site with a rate *kP+* and come out of the pause with a rate *kP-*. An RNAP molecule cannot move forward if another one occupies the bases in front of it. The length of the gene is *L.* (B) **No-traffic model.** The elongation process is uniform with each RNAP molecule occupying one base pair. (C) Using the Gillespie algorithm [5,6], we simulate the traffic model where the promoter initiates at a constant rate, where we take four different values of initiation. Predictions of the no-traffic model for the mean and Fano factor agree well with the simulation results from the traffic model up to an initiation rate of 30 initiations/min. The elongation parameters used, are *kP-=* 4/sec, *kP+=*0.01/sec, *k=*80 bp/sec [7] as reported for ribosomal promoters in *E.Coli*. The length of the gene is *L*= 2000 bps and the footprint of one RNAP molecule is 30 bps.



Figure 2: **Mean number of nascent RNAs in yeast.** Here we plot the mean of the nascent RNA distribution for all twelve genes: MDN1, PRP8, RPB1, PUP1, PRE3, PRE7, KAP104, TAF5, TAF6, TAF12, RPB2, RPB3, reported by Gandhi et al. [8]. For six of these twelve genes, KAP104, TAF5, TAF6, TAF12, RPB2, RPB3, the mean increases linearly with the gene length, as shown in the main text in Fig. 2A. The other six genes do not follow the same trend line indicating that they have different initiation rates; PUP1, PRE3, PRE7, have similar initiation rates but different from the six genes analyzed in the main text.



Figure 3**: Fano factor for nascent RNA distributions in yeast.** Here we show 6 different constitutively expressed genes in yeast [8]: RPB1, MDN1, PUP1, PRE7, PRE3, PRP8 which have different initiation rates than the six genes shown in Figure 2, and another gene PDR5, known to be regulated [9]. The data for Fano factors measured in experiments are shown in comparison with the predictions for the one-step (red line) and two-step (area between red line and blue line) initiation models. The Fano factor for the PDR5 gene is greater than one, which is consistent with an ON-OFF model of transcription initiation. RPB1 has a Fano factor equal to 1, suggesting one-step initiation. All the other genes have Fano factors consistent with more than one rate-limiting step leading up to initiation. The blue line indicates the minimum possible Fano factor for the specific gene assuming two-step initiation and the measured average initiation rate, which we compute from the mean number of nascent RNAs. For the genes whose Fano factor is below this line (PUP1, PRE7, PRE3, PRP8) the data suggest that their transcription is initiated via three or more (similar in duration) steps.

**References**

1. Sanchez A, Choubey S, Kondev J (2013) Regulation of Noise in Gene Expression. Annu Rev Biophys 42: 469–491. doi:10.1146/annurev-biophys-083012-130401.

2. Sanchez A, Choubey S, Kondev J (2013) Stochastic models of transcription: From single molecules to single cells. Methods San Diego Calif 62: 13–25. doi:10.1016/j.ymeth.2013.03.026.

3. Sanchez A, Garcia HG, Jones D, Phillips R, Kondev J (2011) Effect of promoter architecture on the cell-to-cell variability in gene expression. PLoS Comput Biol 7: e1001100. doi:10.1371/journal.pcbi.1001100.

4. Kepler TB, Elston TC (2001) Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys J 81: 3116–3136.

5. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22: 403–434. doi:10.1016/0021-9991(76)90041-3.

6. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81: 2340–2361. doi:10.1021/j100540a008.

7. Klumpp S, Hwa T (2008) Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination. Proc Natl Acad Sci 105: 18159–18164. doi:10.1073/pnas.0806084105.

8. Gandhi SJ, Zenklusen D, Lionnet T, Singer RH (2011) Transcription of functionally related constitutive genes is not coordinated. Nat Struct Mol Biol 18: 27–34. doi:10.1038/nsmb.1934.

9. Zenklusen D, Larson DR, Singer RH (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. Nat Struct Mol Biol 15: 1263–1271. doi:10.1038/nsmb.1514.