nature structural & molecular biology

In vivo dynamics of RNA polymerase II transcription

Xavier Darzacq[1,2], Yaron Shav-Tal[1,3], Valeria de Turris[1], Yehuda Brody[3], Shailesh M Shenoy[1], Robert D Phair[4] & Robert H Singer[1]

We imaged transcription in living cells using a locus-specific reporter system, which allowed precise, single-cell kinetic measurements of promoter binding, initiation and elongation. Photobleaching of fluorescent RNA polymerase II revealed several kinetically distinct populations of the enzyme interacting with a specific gene. Photobleaching and photoactivation of fluorescent MS2 proteins used to label nascent messenger RNAs provided sensitive elongation measurements. A mechanistic kinetic model that fits our data was validated using specific inhibitors. Polymerases elongated at 4.3 kilobases min^{-1}, much faster than previously documented, and entered a paused state for unexpectedly long times. Transcription onset was inefficient, with only 1% of polymerase-gene interactions leading to completion of an mRNA. Our systems approach, quantifying both polymerase and mRNA kinetics on a defined DNA template *in vivo* with high temporal resolution, opens new avenues for studying regulation of transcriptional processes *in vivo*.

Transcription by RNA polymerase II (Pol II) is at the core of gene expression and hence is the basis of all cellular activities. Little information exists about the kinetics of this process in live cells[1], as understanding of gene expression regulation comes from studies using purified proteins. For instance, the subunits of the elongating Pol II are well known[2] and the crystal structure of this enzyme explains much of its behavior *in vitro*[3,4]. mRNA transcription can be deconstructed into a succession of steps: promoter assembly, clearance and escape[5], followed by elongation and termination. The process of transcriptional initiation involves several structural changes in the polymerase as the nascent transcript elongates[6]. Early in initiation, the polymerase can produce abortive transcripts[7,8]. These abortive cycles have been observed with a single prokaryote polymerase (RNAP) releasing several transcripts without escaping the promoter[9,10]. The elongation step can be regulated by pausing for various times, as demonstrated using prokaryotic polymerases *in vitro*[11,12]. For eukaryotic cells, attempts have been made to calculate the endogenous elongation speed using run-on assays[13], reverse-transcription (RT)-PCR[14] or fluorescence *in situ* hybridization (FISH)[15] on specific mRNAs, and these have yielded apparent elongation estimates ranging from 1.1 to 2.5 kilobases (kb) min^{-1}. To date, no assay has been developed to measure the various steps of Pol II transcription in a living cell. For instance, although abortive initiation is widely believed to occur, the dynamics of this event are unknown, including whether initiating polymerases are committed to entering processive elongation or whether they may dissociate from the DNA, and the probability of each event. Furthermore, no assay exists that can measure elongation speed on a chromatin template within a live cell. Accurate measurements of the kinetics of transcription are fundamental to the understanding of transcription assembly, transcriptional regulation and cross-talk with transcription-coupled processes.

Here we report accurate *in vivo* measurements of the mammalian Pol II engaged in each of the steps of active transcription. We previously developed a method for the *in vivo* labeling of mRNA transcripts containing a series of repeated stem-loops (from phage MS2), which are specifically bound by an MS2 coat protein fused to green fluorescent protein (GFP)[16]. The assay consists of a human cell line harboring a gene array into which these stem-loops have been integrated[17]. We have now used this system to follow the synthesis of RNA in real time. Our method allows direct measurement of Pol II initiation events as well as elongation in isolation from the other steps of transcription. By using a deterministic computational model constrained by extensive data sets and tested with transcription inhibitors, we were able to extract features of transcription heretofore unexplored and provide a guide for application of the method to other genes.

RESULTS

Kinetics of Pol II transcription

We used a cell line with a stable integration of approximately 200 repeats of a gene cassette at a single locus[17], each containing 256 upstream *lacO* repeats[18] and a minimal cytomegalovirus (CMV) promoter coupled to a tetracycline-operator cassette controlling a gene that encodes a functional mRNA with 24 MS2 repeats in its 3′ untranslated region[16,19] (**Fig. 1**). We could detect the locus using the lactose repressor fused to red (**Fig. 1b,e,h,k**) or cyan (**Fig. 1n,r,v**)

[1]Department of Anatomy and Structural Biology, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [2]Laboratoire de Génétique Moléculaire, Centre National de la Recherche Scientifique, UMR-8541, Ecole Normale Supérieure, 75005 Paris, France. [3]The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel. [4]Integrative Bioinformatics, Inc., Los Altos, California 94024, USA. Correspondence should be addressed to R.H.S. (rhsinger@aecom.yu.edu).

Received 18 April; accepted 28 June; published online 5 August 2007; doi:10.1038/nsmb1280

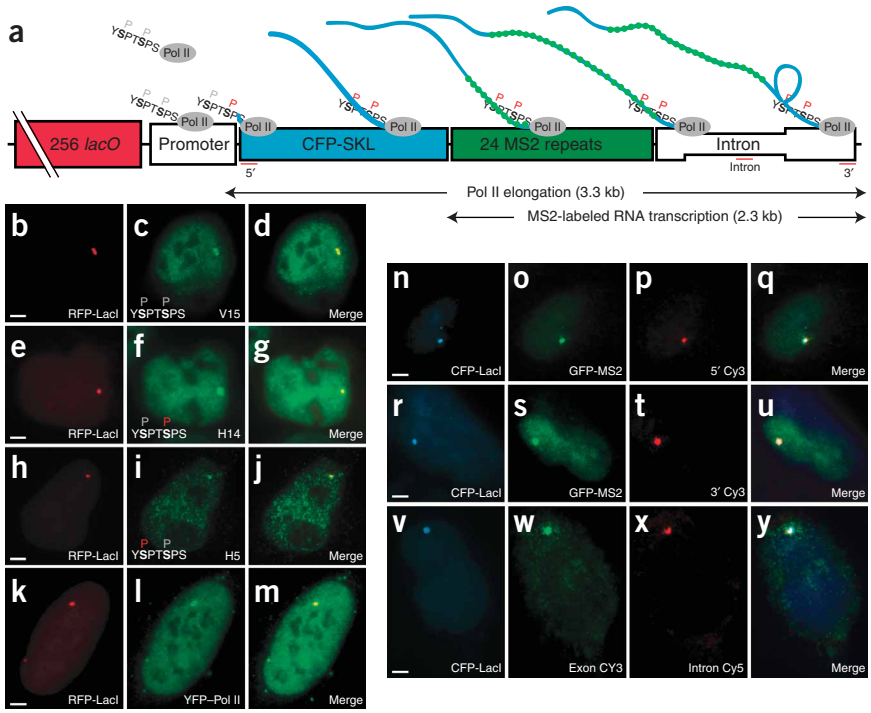

Figure 1 Detecting transcription *in vivo* using fluorescence microscopy. (**a**) Schematic of the gene cassette[17] stably integrated into chromosomes of human U2OS cells. P above protein sequence denotes Pol II phosphorylation state (red, phosphorylated). Reverse tet transactivator (rtTA) in the presence of doxycycline drives gene expression from a minimal CMV promoter[17]. Arrows indicate the 3.3-kb region transcribed by Pol II and the 2.3-kb region labeled by GFP-MS2 fusion proteins. Red lines indicate targets of FISH oligonucleotide probes. (**b**–**m**) Active transcription sites recruit Pol II. In **b**,**e**,**h**,**k**, RFP-LacI labels gene locus. Immunofluorescence (using indicated antibodies) reveals Pol II in three phosphorylation states: unphosphorylated (**c**), phosphorylated at Ser5 (**f**) and phosphorylated at Ser2 (**i**). **l** shows that the transcription site recruits YFP–Pol II (YFP-RPB1αAm^r). In **n**–**y**, nascent mRNAs were detected at active sites. In **n**,**r**,**v**, CFP-LacI labels gene locus. In **o**,**s**, mRNAs bound by GFP-MS2 were detected by FISH (probes at 5′ and 3′ ends are shown in **p**,**t**). FISH signals at exon (**w**) and intron regions (**x**) colocalize only at transcription site (see merge of each row, **q**,**u**,**y**). Scale bars, 5 μm.

fluorescent protein (RFP-LacI or CFP-LacI). Transcription was activated by the doxycycline-induced binding of a VP16 transactivation domain fused to a modified tetracycline repressor DNA-binding element. Upon transcriptional activation, we monitored the number of nascent precursor mRNAs (pre-mRNAs) at the locus using quantitative FISH[20] (**Fig. 1p**,**t**), the binding of MS2-GFP protein to the nascent mRNAs (**Fig. 1o**,**s**) or the recruitment of RNA polymerase to the site (**Fig. 1c**,**f**,**i**,**l**). The total number of mRNAs detected at the site ranged from 200 to 400, with an average of two polymerases per transcription unit. Probes directed to either exon (**Fig. 1w**) or intron sequences (**Fig. 1x**) demonstrated the presence and correct excision of the intron at the transcription site; exon probes detected the distribution of messenger ribonucleoprotein particles in the nucleoplasm[19] and cytoplasm[21], whereas intron probes detected only the transcription site[22,23].

Phosphorylation of the C-terminal domain (CTD) of the Pol II large subunit (RPB1) indicates its activity status[24]. Immunostaining with antibodies to the unphosphorylated CTD (**Fig. 1b–d**), CTD phosphorylated on Ser5 (**Fig. 1e–g**) or CTD phosphorylated on Ser2 (**Fig. 1h–j**) indicated that all polymerase activity states are present at the transcription site, suggesting that they participate in the three main processes of transcription: promoter binding, promoter clearance and elongation (**Fig. 1a**). A yellow fluorescent protein (YFP) fusion of the large subunit of Pol II was recruited and detected at the transcription site, allowing study of the dynamics of these three states (**Fig. 1k–m**).

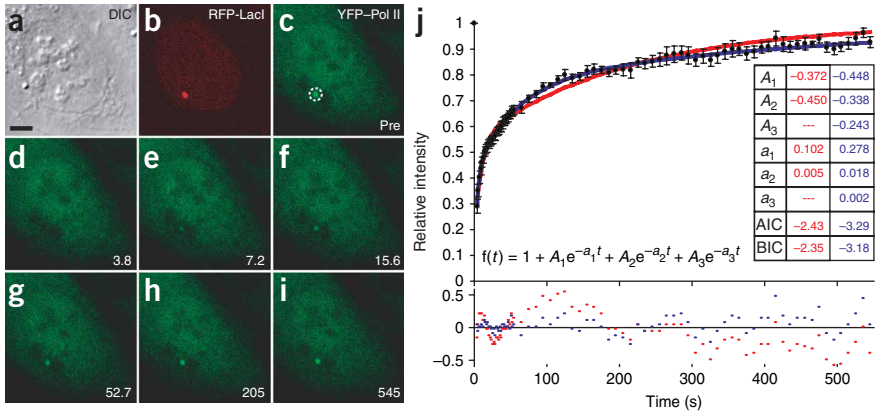

Figure 2 Quantifying Pol II transcription kinetics *in vivo*. Fluorescence recovery after photobleaching of the transcription site is shown in **a**–**i**. (**a**) Differential interference contrast images of live cells. (**b**) RFP-LacI labels gene locus. (**c**) Dashed circle indicates photobleached region. (**d**–**i**) Bleaching (**d**) and recovery (**e**–**i**) of YFP–Pol II[17] at active site, monitored for 545 s. Scale bar, 5 μm. (**j**) Pol II FRAP data (black; $n = 10$) fit to a sum of exponentials (see equation) to determine the minimal model complexity. This was done using generalized least-squares optimization as implemented in the SAAM II software package (http://depts.washington.edu/saam2/). Goodness of fit was evaluated by requiring that coefficients of variation on the parameter estimates were less than 30% and by checking for a random distribution of residuals around 0 (red and blue dots in lower chart represent residuals for two and three exponentials, respectively). By these criteria, a fit of the Pol II FRAP data requires three exponentials (blue), as residuals are not randomly distributed when fit to two exponentials (red). The Akaike information criterion (AIC)[59] and the Bayes-Schwarz information criterion (BIC)[60] for two- and three-exponential models are reported in the inset table. These standard quantitative measures of goodness-of-fit penalize additional model parameters. If the fit is sufficiently improved to justify the increased complexity of the model, then the AIC and BIC of the more complex model will be less than those of the simpler model. By this measure, three exponentials are superior to two in modeling our data. Error bars show s.e.m.

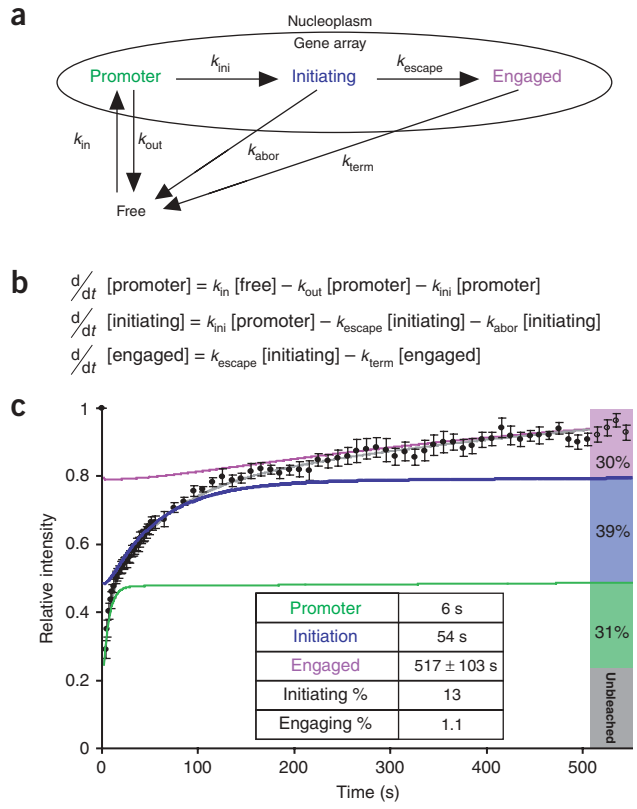

Figure 3 Polymerase II mechanistic kinetic model used to simulate the data. (**a**) Arrows labeled with rate constants represent transitions. (**b**) Differential equations simulating the mechanistic model in **a**, used to analyze the data in **c**. (**c**) Normalized fluorescence recovery of YFP–Pol II after photobleaching (black dots; data are the same as in **Fig. 2j**). The best-fit solution for the mathematical model (gray) characterizes three kinetically distinct states of Pol II (green, blue and purple, respectively) and predicts the steady-state fraction accumulating in each state (right bars). Inset table lists residence times for each state and probabilities for each step derived from the model in **a** and equations in **b**.

Real-time transcription was monitored via the dynamics of fluorescent fusion proteins[19,25]. By measuring several hundred engaged polymerases, we averaged out the stochastic 'noise' of individual gene expression. We selected a stable cell line expressing YFP fused to an α-amanitin–resistant RPB1 mutant (YFP-RPB1αAm^r). Under α-amanitin selection, endogenous RPB1 was degraded[26] and there was no detectable growth phenotype[27]. Doxycycline-induced cells expressing YFP-RPB1αAm^r showed an accumulation of the YFP–Pol II signal at the actively transcribing locus (**Fig. 1l**). Fluorescence recovery after photobleaching (FRAP) was measured by bleaching the YFP–Pol II signal enriched at transcription sites (**Fig. 2**) while the locus was continuously tracked using the RFP-LacI signal (**Fig. 2b**). A multiexponential fit of the recovery data revealed that a minimum of three kinetically distinct populations of polymerases were needed to describe the data (**Fig. 2j**). We postulated that the fraction of the recovery data showing the slowest kinetics represents post-initiation polymerases engaged with the gene and involved in elongation, the fastest fraction represents transient Pol II primary interactions at the promoter site, and the intermediate fraction represents initiation events. We then analyzed fluorescence recovery at the transcription site using mechanistic kinetic models[25] to quantitatively assess our hypothesis. We chose to apply a binding-dominated model, as we were able to verify that nucleoplasmic diffusion of YFP–Pol II did not influence our results (see below). The model (**Fig. 3a**) simulated the kinetics of Pol II assembly and elongation, and allowed the resolution of time constants for polymerases entering at a single site and the relative molecular flux for each step of the transcription process. A simple, linear sequential model could not fit the data, as this would mean that at steady state almost all the polymerases would accumulate at the slowest (rate-limiting) step, so three kinetically distinct populations would not be resolved.

Two types of models could fit the data: the first would simulate three independent populations of Pol II, each committed to performing only a specific step of transcription, whereas the second model would involve three interconnected and dependent populations. We considered the first solution biologically unrealistic, because then each Pol II subpopulation would have to be recruited independently for promoter binding, initiation and elongation. We therefore pursued the second model and formulated it as a system of ordinary differential equations (**Fig. 3b**). The model included the sequential steps of polymerase assembly, initiation and elongation as well as exit points for the polymerase from each step, and it therefore was not constrained to linearity (see above). These exit points represented abortive release from the pre-initiation or initiation complex[28] (**Fig. 3a**), as suggested by the observation that RNA polymerases have an intrinsic tendency to abort initiation after transcribing a short RNA molecule[5,29]. Parameter optimization constrained by the experimental data yielded a consistent solution (**Fig. 3c**, gray curve). In the solution, the mean residence times of the three kinetic fractions for promoter binding, initiation and elongation were 6, 54 and 517 s, respectively (**Fig. 3c**, green, blue and purple curves, respectively; **Table 1** summarizes all the measurements). These data indicate that, as in many nuclear events, the association of a molecule with its target is based on a series of transient interactions[30].

Pol II commitment to the gene is highly inefficient

The best-fit solution of this model predicts that only 13% of the polymerases interacting with the promoter are delivered to the initiation step and that only 8.6% of these engage in a longer-lasting process consistent with elongation (**Fig. 3c**). The net result of these sequential processes is that only 1 polymerase in 90 interaction events proceeds to elongation and produces an mRNA molecule. Despite the progressive order-of-magnitude increases in the residence time (6 s, 54 s and ~500 s (517 ± 103 s)) for their respective components, each of the polymerase populations represents about one-third of the steady-state polymerases (**Fig. 3c**, green, blue and purple bars). This equilibrium is the result of the balanced loss of polymerases during the steps of transcription.

In our analysis, we assumed that Pol II diffusion would not affect the measured recovery speed and therefore need not be included in the model; however, we tested this assumption using established methods[31,32]. First, we compared the dynamics of recovery of the gene array to the corresponding recovery of a nucleoplasmic region distant from the array, where it has been shown that the freely diffusing Pol II represents nearly three-fourths of the signal, whereas the engaged polymerases represent one-fourth[33]. This approach compared regions with different numbers of binding sites. The transcription site, containing 200 copies of the gene cassette, amplified the bound signal, whereas the nucleoplasmic region, expected to contain few active genes, reflected the recovery expected from predominately free rather than