Published online 24 December 2014

Nucleic Acids Research, 2015, Vol. 43, No. 2 ***699–707***
doi: 10.1093/nar/gku1338

Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing

Jeremy C. Davis-Turak[1], Karmel Allison[1,2], Maxim N. Shokhirev[1], Petr Ponomarenko[1], Lev S. Tsimring[1,3], Christopher K. Glass[1,2], Tracy L. Johnson[1,4,*] and Alexander Hoffmann[1,3,5,6,*]

[1]San Diego Center for Systems Biology (SDCSB), University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA, [2]Department of Cellular and Molecular Medicine, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA, [3]BioCircuits Institute, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA, [4]Department of Molecular, Cell, and Developmental Biology, University of California at Los Angeles, Los Angeles, CA 90095, USA, [5]Department of Microbiology, Immunology, and Molecular Genetics (MIMG), University of California at Los Angeles, Los Angeles, CA 90095, USA and [6]Institute for Quantitative and Computational Biosciences (QCB), University of California at Los Angeles, Los Angeles, CA 90095, USA

Received October 10, 2014; Revised December 07, 2014; Accepted December 10, 2014

ABSTRACT

When messenger RNA splicing occurs co-transcriptionally, the potential for kinetic control based on transcription dynamics is widely recognized. Indeed, perturbation studies have reported that when transcription kinetics are perturbed genetically or pharmacologically splice patterns may change. However, whether kinetic control is contributing to the control of splicing within the normal range of physiological conditions remains unknown. We examined if the kinetic determinants for co-transcriptional splicing (CTS) might be reflected in the structure and expression patterns of the genome and epigenome. To identify and then quantitatively relate multiple, simultaneous CTS determinants, we constructed a scalable mathematical model of the kinetic interplay of RNA synthesis and CTS and parameterized it with diverse next generation sequencing (NGS) data. We thus found a variety of CTS determinants encoded in vertebrate genomes and epigenomes, and that these combine variously for different groups of genes such as housekeeping versus regulated genes. Together, our findings indicate that the kinetic basis of splicing is functionally and physiologically relevant, and may meaningfully inform the analysis of genomic and epigenomic data to provide insights that are missed when relying on statistical approaches alone.

INTRODUCTION

Messenger RNA (mRNA) synthesis is a highly regulated process in which transcription factors and chromatin modifying factors coordinate with Pol II to produce a nascent strand of RNA. The nascent pre-mRNA is processed by 5′ capping, 3′ polyadenylation and pre-mRNA splicing—the removal of non-coding introns. Complete splicing is necessary for proper mRNA export, stability and protein function. RNA processing steps can in principle be initiated and completed during the transcription process, i.e. co-transcriptionally, but may also occur post-transcriptionally (1–9). It is now well established that a large fraction of splicing occurs co-transcriptionally in metazoan genomes (9,10).

Because of the constraints imposed on co-transcriptional splicing (CTS) by the parallel process of transcriptional elongation, an intron's fate may be dramatically affected by the elongation dynamics of Pol II. Indeed, a slower Pol II can result in increased use of a weak 5′ splice site in reporter gene constructs (11,12) or alternative exon skipping by allowing a negative factor to bind (13). Splice site choice can be altered in human cell lines by removing downstream pausing sites (14) or pharmacologically slowing down Pol II (15), and Pol II pausing may cause an increase in CTS (5,6,16). Also, spliceosome recruitment may be coordinated with transcription (17–21), for example via the carboxy-terminal domain (CTD) of Pol II (22), and Pol II mutants lacking the CTD produce splicing defects (23). Further, CTS may regulate chromatin modifications to reinforce transcription initiation (24) or may facilitate rapid gene induction during the inflammatory response (25,26).

*To whom correspondence should be addressed. Tel: +310-794-9925; Email: ahoffmann@ucla.edu
Correspondence may also be addressed to Tracy L. Johnson. Tel: +310-206-2416; Email: johnsont@ucla.edu

However, splicing may also occur post-transcriptionally (27–29), and studies of how splice patterns of specific genes are achieved have generally identified trans-acting splicing factors, which may function both co- and post-transcriptionally, as key determinants. Indeed, statistical machine learning approaches are the basis for computational tools that predict splice patterns based on splice factor expression and cognate binding site sequences (30).

Thus, whether the kinetic control implicit in co-transcriptional splicing is in general functionally important remains an open question. The answer determines (i) whether a kinetic modeling framework will be required to improve predictions of splice patterns. Indeed, no scalable mechanistic modeling framework amenable to genome-wide experimental testing has yet been established. It also determines (ii) whether the annotation of the genome itself may gain from applying kinetic considerations.

Here, we examined how the kinetics of transcription and splicing may affect the control of CTS within the normal ranges of metazoan physiologies. We constructed a scalable mathematical model of splicing coupled to transcriptional elongation in order to identify genetic and epigenetic features that regulate CTS. Using this model we developed methods to extract kinetic information from next generation sequencing (NGS) data (29,31), thus allowing us to parameterize the kinetic CTS model in a species-specific manner. We reasoned that while splicing of specific introns may be critically determined by *trans*-acting splicing factors, the common kinetic basis of CTS may be apparent when considering cohorts of introns.

Our analysis revealed a variety of gene features that contribute to CTS efficiency and that these are over-represented in specific cohorts of genes. By expanding the model we were able to simulate co-transcriptional outcomes of multi-intron genes genome-wide as a function of genomic and epigenomic determinants. Our results show that while genes may differ widely in their *cis*-determinants of CTS, the kinetic integration of transcription and splicing is an intrinsic feature of gene expression control, and that mechanistic mathematical models that account for these kinetic processes may form the basis for genome analysis tools that usefully complement statistical approaches.

MATERIALS AND METHODS

Computational modeling

We simulate the elongation of a single polymerase and the transformation of its associated transcript. The probability that an intron has spliced by the time that transcription has terminated is a function of the time it takes to cleave and polyadenylate the mRNA subsequent to the intron's synthesis, and the kinetic rate of splicing. Splicing can be modeled as a series of j sequential reactions (32). By assuming that the time of each reaction is an independent, exponentially distributed random variable with forward rate constant $k_s{*}j$, we can model the probability of splicing at time t as a gamma-distributed random variable with shape j and mean $1/k_s$. Thus, the probability P_i^t that an intron i has spliced by time t is the cumulative distribution:

$$P_i^t(j, k_s) = \sigma_i(t, j, k_s) = \frac{1}{\Gamma(j)} \int_0^{jk_s t} x^{j-1} e^{-x} dx$$

For a single-step reaction ($j = 1$), this simplifies to the exponential distribution:

$$\sigma_i(t, 1, k_s) = 1 - e^{-k_s t}$$

If we assume a constant elongation rate k_E, the total elongation time T_i downstream of intron I is proportional to the distance D_i from intron I to the poly(A) site:

$$\sigma_i(T_i, 1, k_s) = 1 - e^{-k_s T_i} = 1 - e^{-k_s D_i / k_E}$$

Splicing rate constants are reported in the manuscript as $k_E/(k_s{*}j)$.

Multi-intron model

Each potential transcript for a gene with N introns and $N + 1$ exons can be represented as a string $S = [S_1, S_2, \ldots, S_N]$, $S_i \in \{0, 1\}$, where $S_i = 1$ if intron i has been spliced out, and 0 if it is retained. Therefore the probability of each transcript S is:

$$P(S) = \prod_{i=1}^{N} [\sigma_i(T_i, j, k_s) * I(S_i = 1) + (1 - \sigma_i(T_i, j, k_s)) * I(S_i = 0)]$$

where $I(x)$ is the indicator function. To predict the abundance of each transcript at the end of CTS, we calculate $P(S)$ for all possible transcripts: co-transcriptional splicing efficiency was defined as the abundance of the transcript S whose introns have all been removed (all $S_i, = 1 \forall i \in \{1, 2, \ldots, N-1, N\}$). Therefore, CTS efficiency can be computed simply as:

$$\text{CTS efficiency} = \prod_{i=1}^{N} 1 - e^{-k_{s_i} D_i / k_E}$$

This is an O(n) operation and is therefore extremely fast, making this model scalable to genes of any complexity.

RESULTS

A model of co-transcriptional constitutive splicing

We first modeled co-transcriptional splicing of individual introns using a one- or two-step splicing model (cf. (32)). An intron's probability of being spliced co-transcriptionally σ is determined by its splicing rate constant and the duration of the transcriptional phase following the synthesis of the 3′ splice site but prior to mRNA polyadenylation (Supplementary Figure S1A). We next combined models of independent introns to generate a model of co-transcriptional constitutive splicing (CTCS). Our CTCS model enables simulations of multi-intron genes of any complexity and allows us to quantitatively assess the effects of genome structure and kinetic rates on genome-wide splicing outcomes (Figure 1A). Using parameters fit to RNA-seq data to simulate a test gene (see Supplemental Methods, and Supplementary

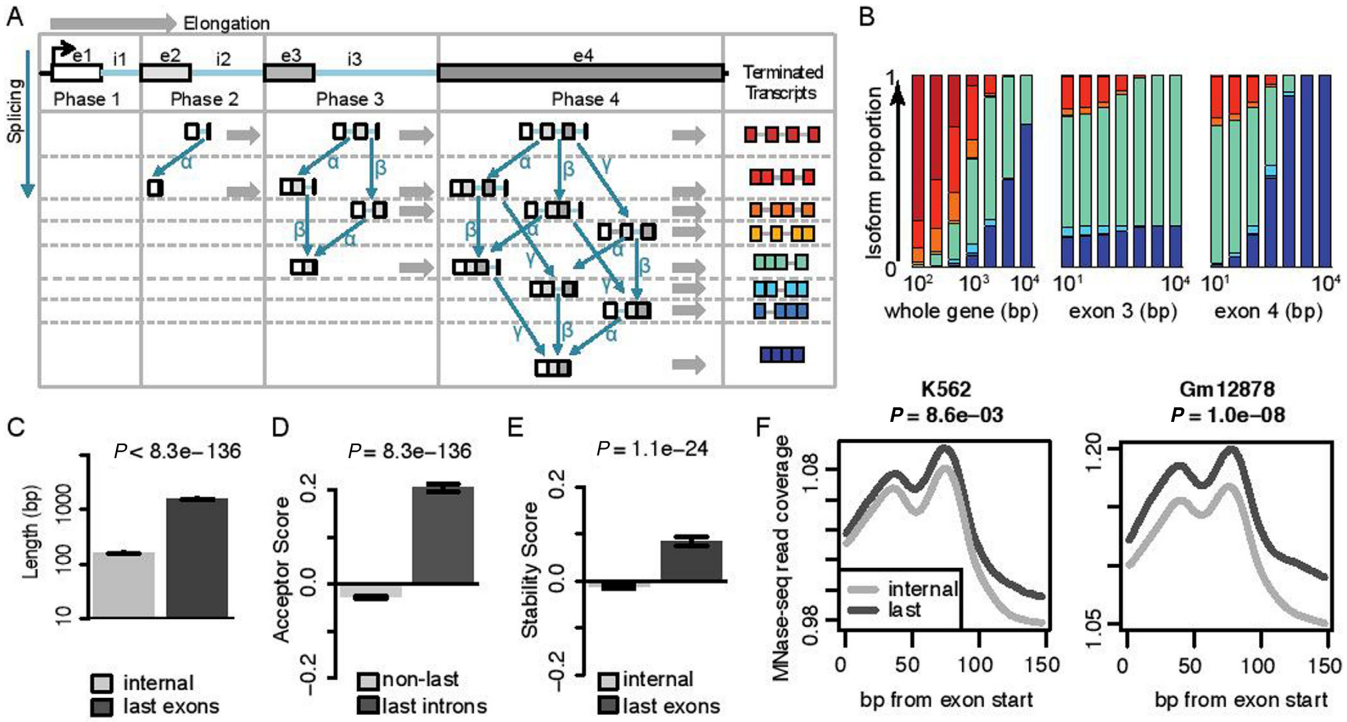

Figure 1. Model of co-transcriptional constitutive splicing (CTCS). (**A**) Model schematic showing all possible reactions and species for a 3-intron gene. The eight possible isoforms that can exist when transcription is complete are color-coded at right. (**B**) Model simulations of the 3-intron gene. Each column represents the distribution of the eight species after each simulation. *Left*: the lengths of all introns and exons were scaled up and down by a constant factor. *Middle*: length of exon 3 was varied; *Right:* length of exon 4 was varied. (**C**) Distribution of exon lengths among last and internal exons in the human genome. (All error bars indicate SEM; *P*-values are the results of *t*-tests). (**D**) Average splicing acceptor scores in last and non-last introns genome-wide. (**E**) Nucleosome stability scores in the first 147 bp of last versus internal exons. (**F**) Average genome-wide MNase-seq signal in K562 cells and GM12878 cells over internal exon starts versus last exon starts. *t*-Test was performed by averaging the signal across the 147 bp and comparing this average in internal versus last exons.

Figure S1B), the CTCS model recapitulated a central point of the kinetic theory of CTS control (6,7,9): namely, that long genes (Figure 1B, *left)*, and specifically genes with long last exons (Figure 1B, *right)* would favor CTS (because they provide more time for splicing), whereas the length of the penultimate exon (which has no influence on the splicing time of the last intron) would be less important (Figure 1B *middle*). These conclusions are robust to specific splicing parameter values (data not shown).

It has previously been suggested that long last exons may have evolved to optimize CTS (Figure 1C, Supplementary Figure S2A; (6)). We expected that if complex genomes evolved under pressure to maintain high CTS efficiency, other genomic signatures besides last exon length, that influence CTS, may be identifiable. Since our model predicted the excision of last introns to be the limiting step in determining the CTS efficiency, we compared acceptor splice site strengths across several genomes based on species-specific sequence motifs. Indeed, we found evidence for conservation of higher average acceptor scores in last introns compared to non-last introns in several vertebrate genomes (Figure 1D, Supplementary Figure S2B: $P = 8.3 \times 10^{-126}$). Since the presence of nucleosomes can inhibit transcription elongation (33) and thus provide more time for CTS, we next tested whether nucleosome stability was enriched at 3′ exons. We first evaluated nucleosome stability across several species using a simple algorithm based on biophysical considerations (34), and found that nucleosomes are indeed expected to be more stable at terminal exons than at internal exons (Figure 1E, Supplementary Figure S2C: $P = 1.1 \times 10^{-24}$). Furthermore, analysis of human cell MNase-seq data (ENCODE Project 35) revealed that nucleosomes are enriched in the proximity of last exons compared to internal exons (Figure 1F). Interestingly, it was previously observed that nucleosomes are present in higher abundance in exons flanking weaker splice sites, both in internal and last exons (36), reinforcing the hypothesis that nucleosome occupancy and splice sites may balance each other to control CTS.

Fitting the model to RNA-seq data

To parameterize our model, we took advantage of existing RNA-seq measurements of purified cellular compartments in K562 cells (31) to estimate the steady-state spliced fraction (Σ) for each intron (17 266 introns in 2768 genes; see 'Materials and Methods' section). For this analysis, we restricted our set of genes to those that use their most downstream annotated poly(A) site, as determined by RNA-seq of the cytoplasmic fraction (13 650 introns in 2136 genes). Examining non-poly-adenylated nuclear transcripts, median Σ strongly correlates with distance to the poly(A) site, in K562 cells, as reported previously (9), and in other cell types (Supplementary Fig-

ure S3). By fitting our model to the median Σ of introns binned according to distance from poly(A) site, we obtained a ratio of splicing rate to elongation speed (see Supplemental Methods). We also examined RNA-seq data from the chromatin fraction of mouse macrophages (29), but this dataset contains many polyadenylated mRNAs that are post-transcriptionally associated with chromatin (Supplementary Figure S4, (26,29)).

Fitting the CTCS model to the nuclear poly(A)-depleted K562 data resulted in a ratio of elongation rate to splicing rate of 615 bp/splicing event, equating to an intron half-life of 9 s if elongation is 3 kb/min (37). The fit was robust to the binning procedure used (Supplementary Figure S1B and C), and a two-step model did not improve the fit to the model (Supplementary Figure S1C). However, previous studies have derived estimates of co-transcriptional splicing rates in diverse organisms ranging from a 30 second half-life to a 5–10 min 'splicing completion time' (4,27,32,38).

A close inspection of the poly(A)-depleted nuclear RNA-seq data revealed that the CTCS model underestimated the steady-state spliced fraction of introns proximal to the poly(A) site (Figure 2A, Supplementary Figure S1B), similar to findings in yeast (6). This disconnect could be due in part to conditions that prolong association of nascent, un-polyadenylated RNA with the chromatin template beyond the time predicted by the poly(A) site (39), such as a transcriptional pause near 3′ ends (6), or transcriptional read-through past the poly(A) site. We therefore modified our CTCS model to include a post-poly(A) site time delay (model CTCS + T), and fit this model to the chromatin-associated RNA-seq data.

Remarkably, allowing for this additional time interval (T_{FIT}) in our model dramatically improved the fit to the data (Figure 2A, Supplementary Figure S3). The best fit was obtained when the median time delay was equivalent to elongating 4.7 kb past the poly(A) site (see Supplemental Methods), and with a new value of 3.1 kb/splicing event for the elongation to splicing ratio. Assuming an elongation rate of 3 kb/min, these values equate to a median 3′ delay of 94 s and a median intron half-life of 43 s. This second estimate of median splicing half-life is more consistent with, though still on the fast side of those previously reported (4,27,32,38). If elongation rates turned out to be slower, those half-life estimates would proportionally increase.

To investigate whether transcriptional read-through could account for the extra time observed in CTCS + T_{FIT}, we measured the extent of active transcription associated with each gene using a novel software tool (40) to analyze GRO-seq (41) data. We used our GRO-seq dataset (42) in mouse macrophages (Figure 2B), and an existing dataset of human MCF7 cells (43) to measure how far pol II activity extends. As cleavage and polyadenylation may occur prior to termination of pol II activity, these measurements put an upper limit on the pre-cleavage read-through distance.

Most genes showed pol II activity well beyond the annotated poly(A) site (Figure 2B) indicating median read-through distances in macrophages and MCF7 of 3.2 and 3.8 kb (equivalent to 68 and 76 s), respectively (Figure 2C). These data suggest that transcriptional read-through may contribute but does not fully account for the estimated delay in polyadenylation after traversing the poly(A) site.

To investigate the effect of the 3′ delay T on CTS, we calculated the CTS efficiency of all human and mouse genes using the CTCS and CTCS + T models and a splicing rate of 3.1 kb/splice (Figure 2D). CTS efficiency was defined as the fraction of transcripts in which all introns are removed prior to cleavage and polyadenylation (see Supplemental Methods), though some level of co-transcriptional splicing may be occurring even for transcripts that are scored as incompletely spliced. With no 3′ delay T, <50% of transcripts were predicted to be completely spliced upon polyadenylation. Genes with many introns, especially short genes, showed even lower CTS efficiency. With the 3′ delay equated to either the median fitted delay time (+T_{FIT}), or to the time equivalent of GRO-seq-measured read-through distances in individual genes (+T_{GRO}), resulted in an increase in CTS efficiency. However, CTS efficiency remained dependent on gene length and the number of introns, such that even these time delays are not sufficient to ensure that all introns are spliced in short genes, especially those with many introns.

Predicting selective Pol II pausing at 3′ ends

Our model revealed that some genes' structures predispose their transcripts for inefficient CTS. However, if efficient CTS were selected for during the evolution of complex genomes, we would expect to find compensatory signatures of other CTS determinants. Indeed, we found that nucleosome stability of genes is markedly higher in short genes than long genes in vertebrate genomes (Figure 3A, Supplementary Figure S5A). This trend could explain the finding that Pol II elongation rate is positively correlated with gene length (44). Furthermore, among short genes, those with high numbers of introns had very high average nucleosome stability scores. No similar compensatory signatures were observed for splice site scores (Supplementary Figure S5B), which correlate with intron length and are universally stronger in last introns (Supplementary Figures S2C and S5C). We examined nucleosome occupancy in K562 and GM12878 cells using the MNase-seq data. Within short genes, nucleosome density increased with increasing numbers of introns (Figure 3B).

We next tested whether we could find evidence of differential Pol II dynamics in long and short genes by examining Pol II CTD SerineS2 phosphorylation (PolS2) in K562 cells in the vicinity of poly(A) sites that were not within 1 kb of any other genes' starts or ends. In the 1 kb upstream of the poly(A) site, PolS2 read densities were higher for short genes than long genes (Figure 3C) (though PolS2 read densities also correlate with gene expression levels), and genes with more introns have disproportionately high PolS2 densities. Furthermore, short genes generally had prominent peaks of PolS2 signal after the poly(A) site (Figure 3D), whereas long genes had lower and broader peaks. These data indicate that differential regulation of Pol II elongation could be sufficient to confer high CTS efficiency to all genes, regardless of gene structure.

To test this hypothesis we simulated all human genes (Figure 3E) and mouse genes (Supplementary Figure S5E) with our CTCS + T_{FIT} model using variable elongation parameters. Using experimentally determined elongation rates of long genes in K562 cells (44), we tested that elongation rates

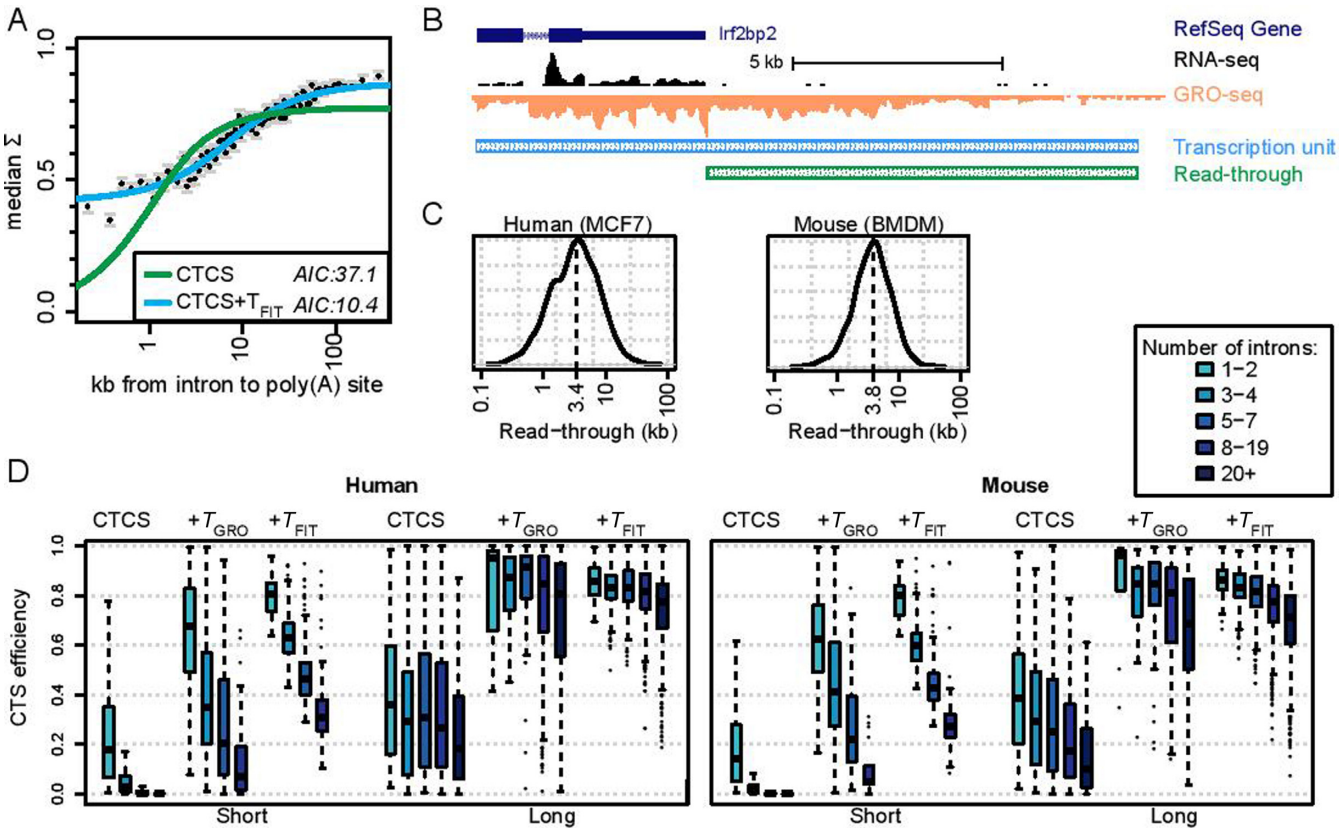

Figure 2. Pol II read-through may contribute to delay time following transcription of the poly(A) site. (**A**) Deriving kinetic parameters by fitting CTCS model to median spliced fraction Σ of 100 equally populated bins of introns. Green and blue lines are the fits to the CTCS or CTCS + T_{FIT} models, respectively. T_{FIT} refers to the average additional time after Pol II transcribes the poly(A) site, as determined by the model fit. (**B**) UCSC browser tracks showing GRO-seq traces in mouse macrophages for representative gene *irf2bp2*. Schematic boxes indicate the read-through lengths and transcription units determined computationally (40). (**C**) Distribution of transcriptional read-through as measured by GRO-seq in human MCF7 cells (top) and mouse macrophages (bottom). (**D**) Simulations of CTS in all human genes (left) and mouse genes (right). Genes were split into four evenly sized groups based on total gene length [short (<6444 bp), medium short (6444–20 252 bp), medium long (20 257–57 229 bp) and long (>57 229 bp)], and further subdivided by the number of introns. Boxplots of CTS efficiency for short and long groups are shown for simulations in three models: CTCS, CTCS + T_{FIT}, and CTCS + T_{GRO}. T_{GRO} refers to the additional time after Pol II transcribes the poly(A) site if transcription proceeds to the termination sites identified by GRO-seq, in individual genes for which GRO-seq measurements are available. All boxes show the extent of the 50% inter-quartile range and the notches estimate a 95% confidence interval for the median.

were negatively correlated with nucleosome stability (Supplementary Figure S5D), and we used this correlation to extrapolate elongation rates for each gene based on the nucleosome stability scores shown in Figure 3A (see Supplemental Methods). Allowing for a variable elongation parameter (+ Δelong) resulted in preferentially marked increases in CTS efficiency for short genes with many introns (Figure 3E). Next, we tested the effect of having stronger splice sites in last introns (+ Δsplice). This change increased CTS efficiency in most gene categories, but exacerbated the differences between genes with many or few introns. When we took into account both variable elongation and splicing rates (+ Δboth), there was an increase in CTS efficiency across all categories. These modeling results are consistent with a central role for elongation control in the regulation of CTS efficiency. Moreover, these data illustrate the power of modeling to elucidate the combined contributions of several factors, such as elongation, nucleosome stability, and splice site strength, for regulating co-transcriptional splicing.

Distinct genomic CTS-determinants of housekeeping genes

Since CTS efficiency depends on Pol II dynamics, we hypothesized that differentially regulated gene groups would show distinct signatures of CTS. We compared constitutively-expressed housekeeping (HK) genes with genes whose expression is more variable across cell types (non-HK; 45). Nascent RNA-seq revealed that HK genes have overall higher steady-state intron spliced fraction Σ than non-HK genes, especially for introns close to the poly(A) site (Figure 4A). The CTCS + T_{FIT} model, when fit to this data, predicts a significantly longer post-poly(A) site delay for HK genes (equivalent to 7.3 kb) compared to non-HK genes (3.8 kb). Interestingly, our mouse GRO-seq data, indicates that the average read-through is significantly longer in HK genes than non-HK genes (4.4 versus 3.6 kb, respectively: Figure 4B: $P = 1.6 \times 10^{-4}$).

Next, we analyzed how the combination of other CTS determinants could contribute to the higher CTS efficiency of HK genes. The model fit resulted in a higher elongation/splicing ratio in HK genes: 3.5 kb/splice

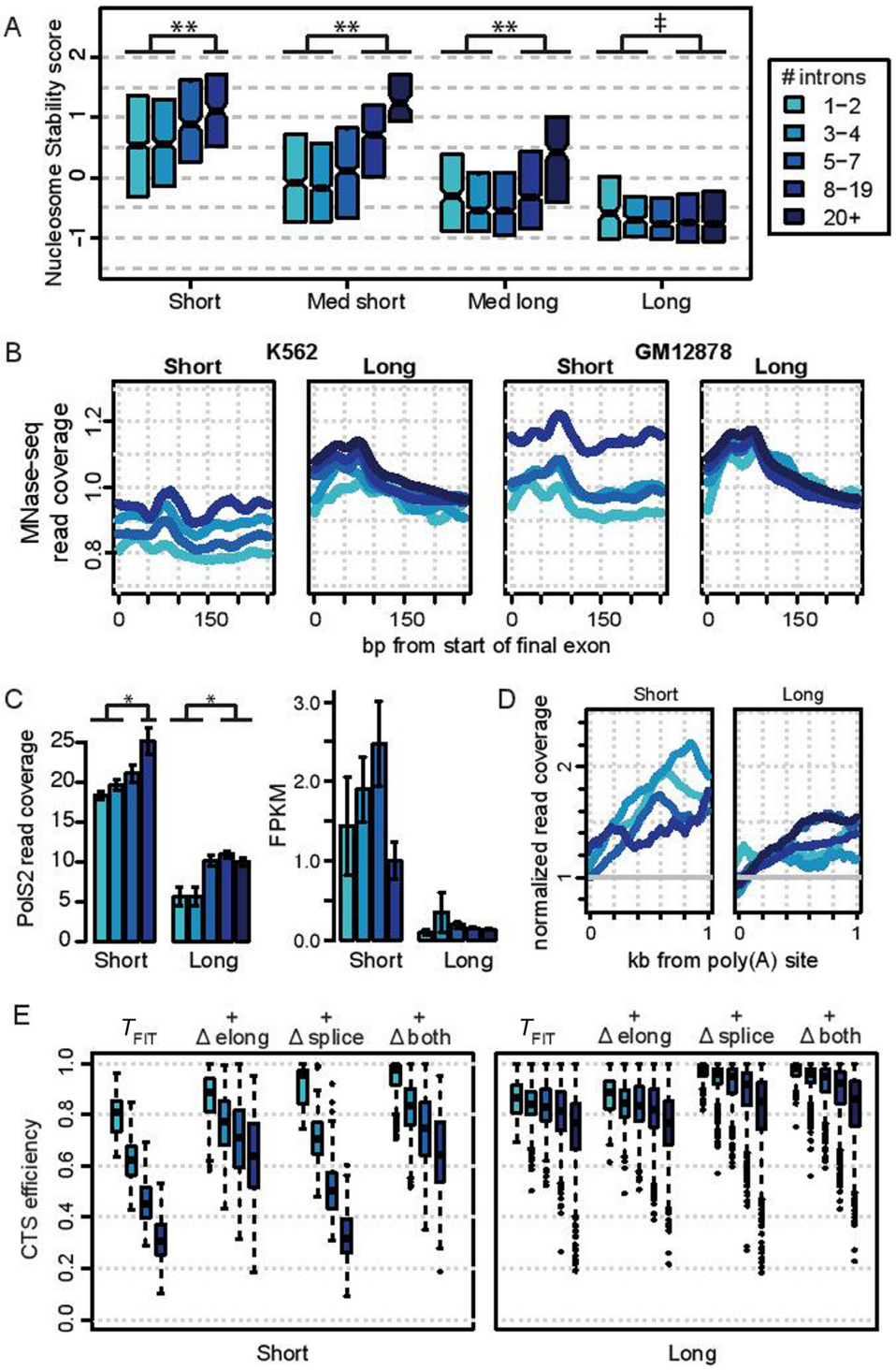

Figure 3. Variable Pol II elongation kinetics favor CTS. (**A**) Boxplots (50% inter-quartile range and 95% confidence interval of the median) of the nucleosome stability score of each gene in indicated gene categories based on gene length and intron numbers. Nucleosome scores were averaged over the region of the gene that encompassed the second through final exon. Stability scores of exons in genes with 1–4 introns are significantly lower ($P < 10e{-}10$, indicated by **), or higher ($P < 0.005$, indicated by ‡) than stability scores of exons in genes with 8+ introns (*t*-tests). (**B**) Average MNase-seq signal over all exon starts in K562 (*left*) and GM12878 cells (*right*) in short and long genes, split up by intron number. (**C**) Average PolS2 ChIP-seq signal in the 1 kb upstream of the poly(A) site (*left*), and gene expression (*right*) for short and long genes split up by number of introns, in K562 cells. PolS2 ChIP-seq signal in genes with 1–4 introns is significantly lower ($P < 0.005$ indicated by *) than PolS2 ChIP-seq signal in genes with 8+ introns (*t*-tests). FPKM stands for fragments per kb per million reads sequenced. (**D**) Average PolS2 signal downstream of the poly(A) site. Traces are normalized to the average of the 1 kb upstream of the poly(A) site for each category. (**E**) Simulations of CTS efficiency in short and long human genes using model CTCS + T_{FIT}. Boxplots of simulations in four separate modeling conditions (see Supplemental Methods) are shown: T_{FIT}: same as Figure 2D; + Δelong: elongation rate of each gene was modulated as an inverse function of nucleosome stability. + Δsplice: kinetic splicing rate was modulated so that last introns had a rate twice the speed of other introns. + Δboth: both elongation and splicing rates were modulated.

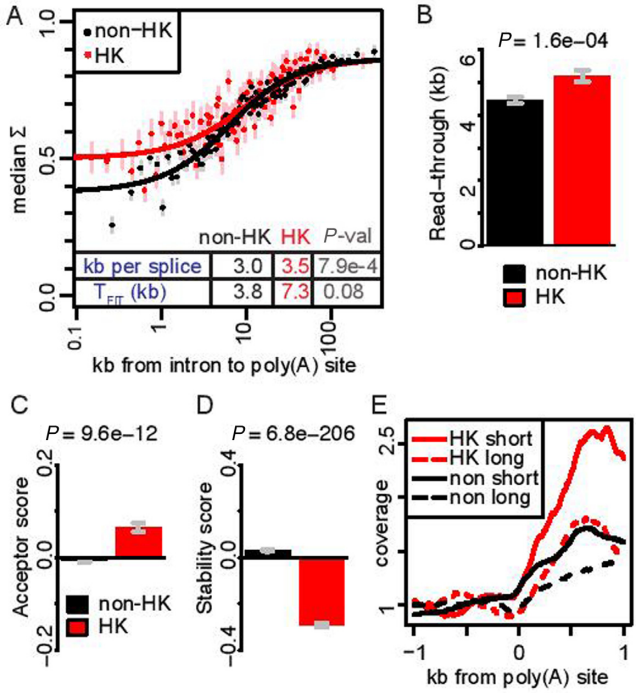

Figure 4. Housekeeping genes have distinct CTS determinants. (**A**) Splicing completion in nuclear poly(A)-depleted RNA-seq is higher in housekeeping genes (red) than other genes (black). Inset shows parameter fits to CTCS + T_{FIT}. The asterisks indicate statistical significance ($P < 0.001$). (**B**) Read-through in mouse genes as measured by GRO-seq. Error bars indicate SEM; *P*-values are the results of *t*-tests. Acceptor scores (**C**) and exonic nucleosome stability scores in the first 147 bp (**D**) in HK versus non-HK genes based on human genome sequence. (**E**) Normalized PolS2 ChIP-seq signal at the poly(A) site of HK and non-HK genes, for long and short genes.

compared 3.0 kb/splice in non-HK genes (Figure 4A). Interestingly, introns throughout HK genes in fact have stronger splice sites than non-HK gene introns (Figure 4C, Supplementary Figure S6A: $P = 9.6 \times 10^{-12}$): therefore the elongation/splicing ratio in HK genes is consistent with a faster elongation rate instead of a slower splicing rate. In support of this hypothesis, HK genes have on average lower nucleosome stability than non-HK genes (Figure 4D, Supplementary Figure S6B: $P = 6.8 \times 10^{-206}$). This faster elongation/splicing ratio implies that HK genes would have lower CTS efficiency were it not for a longer post-poly(A) site delay time. As transcriptional read-through measured by GRO-seq (4.4 kb) does not account for the expected delay (equivalent to 7.3 kb), we hypothesized that transcriptional pause sites may provide additional time. Strikingly, in support of this hypothesis, HK genes have much stronger PolS2 peaks downstream of their poly(A) sites in K562 cells than non-HK genes (Figure 4E).

DISCUSSION

In this study, we constructed a scalable computational model of CTS to identify and interpret the biological significance of genetic, sequence, and epigenetic features in vertebrate genomes. We found that distinct combinations of genetic and epigenetic determinants apply to different cohorts of genes; regulated genes appear to have evolved to rely more on nucleosomal control of pol II elongation to achieve high CTS efficiency than housekeeping genes, which contain more introns (46) but have higher splice site scores and exhibit longer Pol II read-through. As nucleosome density is a component of the regulated chromatin landscape, our observation suggests that splice patterns of non-HK genes are an integral part of their gene expression regulation.

Model simulations further suggested that many transcripts remain incompletely spliced when Pol II reaches the poly(A) site (Figure 2D), but that several mechanisms may contribute to ensuring completed splicing prior to transcript release. Indeed, our analysis estimates an average interval of ∼1.5 min between transcription of the poly(A) site and cleavage of the pre-mRNA (Figure 2A). One mechanism may involve transcriptional read-through by which Pol II terminates well past the poly(A) site (Figure 2C). A second may involve increased PolS2 occupancy indicative of pausing at the 3′ end of shorter genes. And third, several studies have shown that even cleaved and polyadenylated but incompletely spliced mRNAs are retained on the chromatin (26,29,47). The delay in transcript release could result from the complex requirements of termination (48), or perhaps reflects a checkpoint that prevents release of pre-mRNA transcripts (5). Even if the ultimate catalytic steps of splicing occur post-transcriptionally, the recruitment and assembly of splicing complexes likely occur co-transcriptionally (7,8,27,47), and are therefore subject to the kinetic considerations addressed here. Indeed, a detailed quantitative delineation of post-transcriptional events will require refinement of the current model so that the chemical reaction of splicing is delineated starting with the recruitment of splicing factors (cf. (49)). Similarly, the quantitative impact on CTS of other mechanisms such as splicing enhancers or suppressors and associated trans-acting factors (30,50), the chromatin-mediated recruitment of splicing factors, or alternative splicing may be studied by extending the current model formulation.

The present work suggests that mathematical modeling frameworks of splicing must include a kinetic component, and that kinetic considerations encoded in such models lead to new insights about the organization of the genome and epigenome, as well as the patterns of gene expression observed in RNAseq datasets. As such, the present work illustrates that simple kinetic considerations of gene expression processes can be brought to bear on the analysis of genome-wide datasets produced by Next Gen Sequencing approaches, revealing novel insights by mechanistically connecting measurements of gene structure, sequence elements, chromatin modification, and the abundance of mRNA isoforms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. Code is available at https://github.com/jdavisturak/CTCSmodel.

ACKNOWLEDGEMENTS

We would like to thank all participants of the ENCODE consortium for data made available to us, as well as Douglas Black (UCLA) for critical reading of the manuscript.

FUNDING

National Institutes of Health (NIH) grant funding, specifically the 'San Diego Center of Excellence for Systems Biology' [P50 GM085764 to A.H.]; University of California at Los Angeles to Center for 'Ribonomics of Gene Regulation' [U01 HG007912 to A.H.]; investigator grants [R01 GM085474 to T.L.J. and R01ES024995 to A.H.]; National Science Foundation Graduate Research Fellowship Program (to J.D.-T., M.N.S.). Funding for open access charge: NIH [P50 GM085764].

Conflict of interest statement. None declared.

REFERENCES

1. Shatkin,A. and Manley,J. (2000) The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.*, **7**, 1–5.
2. Neugebauer,K.M. (2002) On the importance of being co-transcriptional. *J. Cell Sci.*, **115**, 1–7.
3. Tennyson,C., Klamut,H. and Worton,R. (1995) The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat. Genet.*, **9**, 1–7.
4. Singh,J. and Padgett,R.A. (2009) Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.*, **16**, 1128–1133.
5. Alexander,R., Innocente,S., Barrass,J. and Beggs,J. (2010) Splicing-dependent RNA polymerase pausing in yeast. *Mol. Cell*, **40**, 582–593.
6. Carrillo,O., Preibisch,S. and Neugebauer,K. (2010) Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol. Cell*, **40**, 571–581.
7. Wetterberg,I., Bauren,G. and Wieslander,L. (1996) The intranuclear site of excision of each intron in Balbiani ring 3 pre-mRNA is influenced by the time remaining to transcription termination and different excision efficiencies for the various introns. *RNA*, **2**, 641–651.
8. Pandya-Jones,A. and Black,D.L. (2009) Co-transcriptional splicing of constitutive and alternative exons. *RNA*, **15**, 1896–1908.
9. Tilgner,H., Knowles,D.G., Johnson,R., Davis,C.A., Chakrabortty,S., Djebali,S., Curado,J., Snyder,M., Gingeras,T.R. and Guigo,R. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, **22**, 1616–1625.
10. Khodor,Y., Rodriguez,J., Abruzzi,K., Tang,C.-H., Marr,M.T. 2nd and Rosbash,M. (2011) Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev.*, **25**, 1–12.
11. de la Mata,M., Alonso,C.R., Kadener,S., Fededa,J.P., Blaustein,M.a., Pelisch,F., Cramer,P., Bentley,D. and Kornblihtt,A.R. (2003) A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell*, **12**, 525–532.
12. Howe,K.J., Kane,C.M. and Ares,M. Jr (2003) Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA*, **9**, 993–1006.
13. Dujardin,G., Lafaille,C., de la Mata,M., Marasco,L.E., Munoz,M.J., Le Jossic-Corcos,C., Corcos,L. and Kornblihtt,A.R. (2014) How slow RNA polymerase II elongation favors alternative exon skipping. *Mol. Cell*, **54**, 683–690.
14. Shukla,S., Kavak,E., Gregory,M., Imashimizu,M., Shutinoski,B., Kashlev,M., Oberdoerffer,P., Sandberg,R. and Oberdoerffer,S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74–79.
15. Ip,J.Y., Schmidt,D., Pan,Q., Ramani,A.K., Fraser,A.G., Odom,D.T. and Blencowe,B.J. (2011) Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res.*, **21**, 390–401.
16. Batsche,E., Yaniv,M. and Muchardt,C. (2006) The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat. Struct. Mol. Biol.*, **13**, 22–29.
17. Bentley,D. (2002) The mRNA assembly line: transcription and processing machines in the same factory. *Curr. Opin. Cell Biol.*, **14**, 336–342.
18. Close,P., East,P., Dirac-Svejstrup,A.B., Hartmann,H., Heron,M., Maslen,S., Chariot,A., Soding,J., Skehel,M. and Svejstrup,J.Q. (2012) DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. *Nature*, **484**, 386–389.
19. Gunderson,F.Q. and Johnson,T.L. (2009) Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. *PLoS Genet.*, **5**, e1000682.
20. Hirose,Y. and Manley,J. (2000) RNA polymerase II and the integration of nuclear events. *Genes Dev.*, **14**, 1415–1429.
21. Gornemann,J., Kotovic,K.M., Hujer,K. and Neugebauer,K.M. (2005) Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol. Cell*, **19**, 53–63.
22. de la Mata,M. and Kornblihtt,A.R. (2006) RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat. Struct. Mol. Biol.*, **13**, 973–980.
23. McCracken,S., Fong,N., Yankulov,K., Ballantyne,S., Pan,G., Greenblatt,J., Patterson,S., Wickens,M. and Bentley,D. (1997) The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, **385**, 357–361.
24. Bieberstein,N.I., Carrillo Oesterreich,F., Straube,K. and Neugebauer,K.M. (2012) First exon length controls active chromatin signatures and transcription. *Cell Rep.*, **2**, 62–68.
25. Hao,S. and Baltimore,D. (2013) RNA splicing regulates the temporal order of TNF-induced gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11934–11939.
26. Pandya-Jones,A., Bhatt,D.M., Lin,C.H., Tong,A.J., Smale,S.T. and Black,D.L. (2013) Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *RNA*, **19**, 811–827.
27. Tardiff,D., Lacadie,S. and Rosbash,M. (2006) A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Mol. Cell*, **24**, 1–22.
28. Vargas,D., Shah,K., Batish,M., Levandoski,M., Sinha,S., Marras,S., Schedl,P. and Tyagi,S. (2011) Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, **147**, 1054–1065.
29. Bhatt,D.M., Pandya-Jones,A., Tong,A.J., Barozzi,I., Lissner,M.M., Natoli,G., Black,D.L. and Smale,S.T. (2012) Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, **150**, 279–290.
30. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
31. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
32. Schmidt,U., Basyuk,E., Robert,M.-C., Yoshida,M., Villemin,J.-P., Auboeuf,D., Aitken,S. and Bertrand,E. (2011) Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. *J. Cell Biol.*, **193**, 819–829.
33. Subtil-Rodriguez,A. and Reyes,J.C. (2010) BRG1 helps RNA polymerase II to overcome a nucleosomal barrier during elongation, in vivo. *EMBO Rep.*, **11**, 751–757.
34. Vaillant,C., Audit,B. and Arneodo,A. (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.*, **99**.
35. Consortium,E.P. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
36. Tilgner,H., Nikolaou,C., Althammer,S., Sammeth,M., Beato,M., Valcarcel,J. and Guigo,R. (2009) Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, **16**, 996–1001.
37. Kwak,H. and Lis,J.T. (2013) Control of transcriptional elongation. *Annu. Rev. Genet.*, **47**, 483–508.
38. Aitken,S., Alexander,R. and Beggs,J. (2011) Modelling reveals kinetic advantages of co-transcriptional splicing. *PLoS Comp. Biol.*, **7**, e1002215.
39. Boireau,S., Maiuri,P., Basyuk,E., de la Mata,M., Knezevich,A., Pradet-Balade,B., Backer,V., Kornblihtt,A., Marcello,A. and Bertrand,E. (2007) The transcriptional cycle of HIV-1 in real-time and live cells. *J. Cell Biol.*, **179**, 291–304.
40. Allison,K.A., Kaikkonen,M.U., Gaasterland,T. and Glass,C.K. (2014) Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic Acids Res.*, **42**, 2433–2447.

41. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
42. Kaikkonen,M.U., Spann,N.J., Heinz,S., Romanoski,C.E., Allison,K.A., Stender,J.D., Chun,H.B., Tough,D.F., Prinjha,R.K., Benner,C. *et al.* (2013) Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol. Cell*, **51**, 310–325.
43. Li,W., Notani,D., Ma,Q., Tanasa,B., Nunez,E., Chen,A.Y., Merkurjev,D., Zhang,J., Ohgi,K., Song,X. *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, **498**, 516–520.
44. Veloso,A., Kirkconnell,K.S., Magnuson,B., Biewen,B., Paulsen,M.T., Wilson,T.E. and Ljungman,M. (2014) Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.*, **24**, 896–905.
45. Chang,C., Cheng,W., Chen,C., Shu,W., Tsai,M., Huang,C. and Hsu,I.C. (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One*, **6**, 1–10.
46. Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.
47. Brody,Y., Neufeld,N., Bieberstein,N., Causse,S.Z., Bohnlein,E.M., Neugebauer,K.M., Darzacq,X. and Shav-Tal,Y. (2011) The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol.*, **9**, e1000573.
48. Proudfoot,N.J. (2011) Ending the message: poly(A) signals then and now. *Genes Dev.*, **25**, 1770–1782.
49. Murugan,R. and Kreiman,G. (2012) Theory on the coupled stochastic dynamics of transcription and splice-site recognition. *PLoS Comput. Biol.*, **8**, e1002747.
50. Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.