Published in final edited form as:
Mol Cell. 2021 March 04; 81(5): 983–997.e7. doi:10.1016/j.molcel.2021.01.002.

Enhancers predominantly regulate gene expression during differentiation via transcription initiation

Martin S. C. Larke[1,2], **Ron Schwessinger**[1,2], **Takayuki Nojima**[3], **Jelena Telenius**[1], **Robert A. Beagrie**[4], **Damien J. Downes**[1], **A. Marieke Oudelaar**[1,2], **Julia Truch**[1], **Bryony Graham**, **M. A. Bender**[5], **Nicholas J. Proudfoot**[3], **Douglas R. Higgs**[4], **Jim R. Hughes**[1,2,*]

[1]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[2]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[3]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

[4]Laboratory of Gene Regulation, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

[5]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Summary

Gene transcription occurs via a cycle of linked events including initiation, promoter proximal pausing and elongation of RNA polymerase II (Pol II). A key question is how do transcriptional enhancers influence these events to control gene expression? Here we present an approach, which evaluates the level and change in promoter proximal transcription (initiation and pausing) in the context of differential gene expression, genome-wide. This combinatorial approach shows that in primary cells control of gene expression during differentiation is achieved predominantly via changes in transcription initiation rather than via release of Pol II pausing. Using genetically engineered mouse models, deleted for functionally validated enhancers of the α- and β-globin

Correspondence to: Douglas R. Higgs; Jim R. Hughes.

CORRESPONDING AUTHORS: doug.higgs@imm.ox.ac.uk, jim.hughes@imm.ox.ac.uk .
LEAD CONTACT
Jim R. Hughes jim.hughes@imm.ox.ac.uk

AUTHOR CONTRIBUTIONS
Funding; J.H, D.H and ML; Conceptualization, J.H, D.H and M.L.; Methodology, M.L and J.H; Validation, M.L; Formal analysis, R.S, J.Te and R.B; Resources; J.S, J.S.S and S.B; Investigation; M.L (T.N performed mNET-seq and DD performed NG-Capture-C); Software; J.Te, R.B, R.S, A.M.O; Writing –M.L, D.H and J.H, with N.P, M.A.B, R.B, A.M.O, D.D; Visualization; M.L, J.H and D.H; Data curation; M.L; Supervision; J.H and D.H; Project administration: M.L, J.H and D.H.

DECLARATION OF INTERESTS
Jim R. Hughes is a founder and director of Nucleome Therapeutics.

RESOURSE AVAILABILITY
Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jim R. Hughes (jim.hughes@imm.ox.ac.uk)

Materials availability
This study did not generate new unique reagents.

loci, we confirm that these elements regulate Pol II recruitment and/or initiation to modulate gene expression. Together, our data show that gene expression during differentiation is regulated predominantly at the level of initiation and that enhancers are key effectors of this process.

Introduction

Regulation of gene expression during lineage specification and differentiation occurs via the activation of promoters by enhancers both of which respond to transcriptional and epigenetic programs. There are many points in the transcription cycle at which enhancers could influence gene expression. Chromatin remodeling complexes generate accessible DNA at promoter regions, allowing transcription factor binding (Clapier et al., 2017; Venkatesh and Workman, 2015). Components of the general transcription machinery nucleate at promoters, forming pre-initiation complexes (PICs) and inducing conformational changes in the DNA helix to allow the binding of Pol II and concomitant synthesis of RNA from TSSs (Schilbach et al., 2017). A potential point of regulation lies +30-60 nucleotides (nts) downstream of the TSS where the elongation complex, including Pol II transiently pauses via the action of negative elongation factors (NELF and DSIF) (Wada et al., 1998; Yamaguchi et al., 1999). This paused complex is then converted into an elongation competent form via its phosphorylation by CDK9, the kinase component of the positive transcription elongation factor (P-TEFb). Transcriptional pausing is thereby released allowing productive transcriptional elongation (Adelman and Lis, 2012; Chen et al., 2018; Jonkers and Lis, 2015; Marshall and Price, 1995) before transcription is terminated (Czudnochowski et al., 2012; Kamieniarz-Gdula and Proudfoot, 2019; Laitem et al., 2015; Porrua and Libri, 2015). The key question addressed here is how do enhancers control the transcription cycle at promoters during differentiation?

ChIP-seq analysis has shown that Pol II accumulates across the promoter proximal region (~-100 to +300) including the TSSs and transcriptional pause site (TPS) of most coding genes (L. Core and Adelman 2019; Jonkers and Lis 2015; Kim et al. 2005; Muse et al. 2007; Rahl et al. 2010; Wissink et al. 2019; Zeitlinger et al. 2007). It has been proposed that this accumulation represents paused Pol II and its ratio to the Pol II density across the gene body (the pausing index) has been used as a measure of transcriptional pausing (Muse et al., 2007; Rahl et al., 2010; Zeitlinger et al., 2007). Inhibition of CDK9 causes this pausing index to increase, providing evidence for gene regulation via transcriptional pausing (Gressel et al., 2017; Jonkers et al., 2014; Rahl et al., 2010). Pausing index calculations have been applied to nascent RNA-based methods such as Start-seq (Henriques et al., 2013), Global Run On Sequencing (GRO-seq) (Core et al., 2008; Kwak et al., 2013) and its derivatives have led to the conclusion that pausing is the key regulatory step in the transcription cycle (Adelman and Lis, 2012; Chen et al., 2018; Jonkers and Lis, 2015) which may be influenced by enhancers.

Current interpretations of the transcription cycle depend on several caveats. 1) That CDK9 inhibition is specific and CDK9 solely regulates Pol II pause release *in vivo*. 2) That the Pol II pausing index is a good proxy for nascent RNA expression levels. 3) That the threshold at which genes are called paused is not arbitrary and represents a distinct functional state of Pol II activity. 4) That the methodology used to characterize the transcription cycle does not perturb cellular metabolism in such a way as to disrupt gene expression. These criteria are not always fully satisfied. For example, it has been shown that CDK9 plays a key role in regulation of transcription termination both at the 3' ends of genes and near the promoter (Laitem et al., 2015; Sansó et al., 2016), casting doubt on estimates of the genome-wide prevalence of Pol II pausing and the extent to which it regulates gene expression *in vivo*. The resolution provided by Pol II ChIP-seq is not best suited to resolve initiation and pausing which occur in a narrow window (within 60nt) (Core and Adelman, 2019; Wissink et al., 2019). In addition, ChIP-seq Pol II pausing index values do not correlate well with gene expression suggesting they are a poor predictor of gene expression state (Adelman and Lis, 2012; Rahl et al., 2010). Finally, due to the large number of factors and cofactors which are required in the process of transcription (Bernecky et al., 2016; Farnung et al., 2018; Nozawa et al., 2017; Sainsbury et al., 2015; Schilbach et al., 2017), there is concern that the normal transcription cycle may be perturbed if cells are subjected to extensive preparation and treatment, which is the case in many widely used run on methods to evaluate pausing (Chu et al., 2018; Core et al., 2008; Kwak et al., 2013; Tome et al., 2018).

To circumvent some of these issues, here we have estimated the extent of Pol II initiation and pausing and their contribution to gene expression in primary cells undergoing differentiation with minimal perturbation (referred to here as *in vivo*). We have examined initiation, pausing and nascent gene expression at thousands of genes, throughout mouse erythropoiesis using scaRNA-seq, a derivative of the Start-seq short capped RNA-seq assay (Henriques et al., 2013; Nechaev et al., 2010), in combination with measurements of nascent gene expression (mNET-seq and intronic RNA) (Boswell et al., 2017; la Manno et al., 2018; Nojima et al., 2015). While acknowledging the complexity and interdependence of the various stages of the transcription cycle, simultaneous analysis of initiation, pausing and elongation suggests that enhancer-driven transcription, during differentiation of primary cells is controlled predominantly via initiation rather than pause-release. To test this experimentally we investigated how deletion of the major enhancers of the well-characterized globin loci affects the transcription cycle in primary erythroid cells. Consistent with our genome-wide data, we show that in each case the enhancers primarily affect the stages of recruitment and/or initiation of transcription rather than pause-release or elongation.

Results

Analysis of transcription initiation and Pol II pausing *in vivo* using scaRNA-seq

To understand how initiation and promoter proximal pausing affect gene expression, it is necessary to measure these distinct stages of transcription quantitatively, at high resolution, and at single genes, *in vivo* (in primary cells during differentiation). The Start-seq assay has previously been used to detect and measure initiation and pausing by sequencing the ends of short capped RNA molecules (scaRNA-seq); importantly, this assay does not require extensive cell permeabilization or metabolic labelling (Henriques et al., 2013; Nechaev et al., 2010).

Total scaRNA molecule counts at a given promoter represent the sum of both nascent initiation and pausing. Positive changes in this count may imply more initiation, more pausing or a combination of both processes; negative changes represent the converse (Figure 1A). By generating a complementary measure of nascent gene expression such as intron coverage from RNA-seq (intronRNA) (Gaidatzis et al., 2015), changes in scaRNA level can be interpreted in a simple manner. For a given gene, concordant changes in scaRNA and intronRNA coverage imply regulation through initiation, and discordant changes imply regulation through pausing (Figure 1B).

To count the total number of scaRNA molecules at any given gene we modified the Start-seq assay (Figure S1A). Firstly, we adopted paired-end sequencing, allowing each scaRNA molecule to be "reconstructed" *in silico* from paired reads from either end of the sequenced molecule. Secondly, we isolated larger RNA transcripts than in the original assay (<=300nt compared to ~100nt), allowing us to examine scaRNA levels beyond the predicted range of Pol II pausing and the boundary of the +1 nucleosome (+ ~150nts relative to the TSS) which has been suggested to play a role in Pol II pausing (Jonkers and Lis, 2015; Mieczkowski et al., 2016; Voong et al., 2016; Yazdi et al., 2015). Finally, we removed the chromatin isolation step from the Start-seq protocol to reduce material losses enabling assay of smaller (primary) cell populations than previously analyzed (Core and Adelman, 2019; Henriques et al., 2013, 2018; Williams et al., 2015). We refer to this procedure as scaRNA-seq (Figure S1A).

Validating scaRNA-seq as a single molecule, genome-wide assay for initiation and pausing

We validated scaRNA-seq by analyzing K562 cells for which there are many published datasets documenting TSSs, initiation, and pausing (L. J. Core et al. 2014; Tome, Tippens, and Lis 2018). In addition, we wanted to ensure that scaRNA-seq would be applicable to a modest number of cells ($5x10^6$), facilitating our downstream analysis of primary mouse erythroid cells.

Firstly, we determined whether scaRNA-seq could be used to derive a count of the total number of promoter proximal transcripts (the sum of nascent initiation and pausing events) at single genes. We reconstructed RNA single molecules *in silico* using the mapped coordinates of sequenced paired reads located at their 5' and 3' ends. We then examined genes which have been previously characterized in K562 cells: in particular, we visualized the *HSPA1A* (Heat shock protein 70) and *MYC* genes (Figure 1C). We found that the scaRNA molecules align near to the annotated TSSs of these genes in a manner consistent with published expressed sequence tag (EST) and RNA-seq data sets. This established that scaRNA-seq can be used to count promoter proximal transcripts.

Using scaRNA-seq, we generated a pileup of the 5' or 3' ends of all single RNA molecules genome-wide. At each gene sampled we found a predominant, focused TSS (purple peaks) and the TPS (orange peaks) (Figure 1C). This confirmed that scaRNA-seq has the ability to resolve profiles of TSSs and TPSs in a similar manner to Start-seq, however this information is now obtained in one paired-end sequencing assay. To determine the accuracy and efficiency with which scaRNA-seq detects transcription initiation sites we used Homer (Heinz et al. 2010) to call TSSs and then selected those that lie within 500bp of previously annotated UCSC gene TSSs. We then identified each TSS by scaRNA-seq and the associated annotated TSS before searching for the presence of Initiator and TATA box motifs known to be enriched at promoters in eukaryotes (Vo Ngoc et al., 2017) and observed a strong (3-4 fold) enrichment at the TSSs called from scaRNA-seq data (Figure S1B).

Next, to assess the ability of scaRNA-seq to detect promoter proximal pausing genome-wide, we generated mNET-seq data, which identifies 3’ RNA ends associated with Pol II, in K562 cells (Nojima et al., 2015). We also analyzed published GRO-seq data, which maps nascent RNA 3’ ends using an *in vitro* run on assay (Core et al., 2008, 2014). The distribution of scaRNA 3’ ends corresponds to the profiles obtained using these previously validated approaches (Figure S1C) confirming that the 3’ ends of scaRNA represent Pol II pausing near the TSS genome-wide.

Finally, to validate that scaRNA-seq could detect transcription initiation and promoter proximal pausing we reanalyzed Pol II ChIP-exo data, which provides a high-resolution map of protein binding within chromatin (Figure S2A; Rhee and Pugh 2011). We plotted ChIP-exo data over two sources of TSS annotation, scaRNA derived (observed TSS) and UCSC (annotated TSS), displaying the data as a meta plot. An additional peak of Pol II occupancy at the TSS, indicative of Pol II loading is visible at the TSS when using observed TSSs (Figure S2B). This peak was previously undetectable using UCSC annotated TSSs (Figure S2B) and highlights the increased resolution provided by the use of scaRNA-seq derived TSSs as opposed to annotated TSSs. Together these findings show that scaRNA-seq accurately marks TSSs and TPSs genome-wide.

scaRNA-seq refines the annotation of promoter proximal transcription

Preliminary in-depth analysis of *HSPA1A* and *MYC* genes (Figure 1C) revealed an interesting phenomenon; at both genes, the predominant RNA 5’ end signal (observed using scaRNA-seq) is located downstream, of the annotated (UCSC) TSSs and is supported by mapped human ESTs and polyA+ RNA-seq (Dunham et al., 2012). The upstream UCSC-annotated *MYC* TSS appears to be a minor TSS in K562 cells confirmed by scaRNA, ESTs and polyA+ RNA-seq. For the *HSPA1A* gene the TSS observed with scaRNA-seq is poorly supported by any UCSC gene annotation. To investigate any systematic skew in annotated TSS positions versus scaRNA-observed TSSs, we plotted the relative positions of all TSSs observed in K562 cells using scaRNA-seq versus those annotated in UCSC. This showed that more than 97% of TSSs observed *in vivo* are located some distance away from the annotated (UCSC) TSSs. Less than 3% of the annotated (UCSC) TSSs are exactly coincident with an observed TSS (Figure 2A). Most observed TSSs lie downstream of the annotated TSS, indicating a systematic bias in the annotation of TSSs. Importantly, the TSSs called from scaRNA-seq data had a 3-4x higher enrichment for the TSS associated Initiator and TATA motifs than annotated UCSC TSSs (Figure S1B). This suggests that scaRNA-seq provides a more accurate representation of commonly used TSSs than the UCSC database gene annotations which are commonly used in the meta-analysis of genomics data.

To investigate whether there was a systematic skew in other commonly used TSS annotations we compared the position of scaRNA derived TSS to annotated TSSs from the Gencode (Frankish et al., 2019) and Refseq annotation projects (O’Leary et al., 2016)

(Figure 2A). This revealed annotated TSS positions are also found upstream of observed TSS positions in these commonly used gene annotations. The skew is greatest in the UCSC annotations and least in Refseq. What is the likely source of this skew? The pipeline for UCSC annotation, selects the "longest" and highest "quality" isoforms when assembling putative gene isoforms from raw RNA-seq data (Hsu et al., 2006). If an isoform of a given gene is expressed from a TSS in one tissue but in other tissues the isoform is expressed from a downstream TSS the less common, upstream coordinate is nevertheless selected as the primary TSS candidate. A further issue in the UCSC annotation pipeline may stem from the use of CAGE data in assembling putative Gene isoforms. CAGE is known to extend 5' cDNA beyond the position of the 5' cap, often by 10-100nts (Shiraki et al., 2003). Therefore, large-scale annotation projects which incorporated this data, such as UCSC (Hsu et al., 2006) may not represent the most commonly used TSSs. Our analysis suggests that different TSS annotations are skewed to different degrees and indicates that most TSSs are found within -50 to +300 relative to their annotated positions in K562 cells.

To investigate whether the use of different annotations would change the accepted view of promoter proximal transcription (Core et al., 2014) we plotted scaRNA molecules over observed, UCSC, Gencode and Refseq TSS annotations (Figure 2B). This revealed that different annotations appear to change the profile of transcription near the TSS with the primary differences being the degree to which transcription initiation appears focused rather than dispersed around the TSS and the point at which 3' RNA ends appear to accumulate (the site of Pol II pausing). Using observed scaRNA-seq or Refseq TSS revealed that RNA 3' ends consistently accumulate at +46nt downstream of the TSS (indicative of Pol II pausing) and then decay towards +150nt downstream of the TSS (Figure 2B).

To investigate the profile of initiation and pausing at single genes we plotted 5' and 3' ends of each scaRNA molecule as a heatmap, using the observed TSS derived from scaRNA-seq (Figure 2C). This confirms that most genes have focused initiation occurring in the few bps around the TSS and accumulations of 3' ends indicative of pausing occur at a sharply defined point (+46nt). This analysis suggests that transcription of capped RNA occurs in a focused manner with little evidence of dispersed transcription initiation. Relative to this initiation site, Pol II pauses in a range sharply demarked at +46nt and extending to ~150nt (coincident with the predicted position of the +1 nucleosome).

Gene expression is predominantly regulated via initiation rather than Pol II pausing.

It has been suggested that regulated Pol II pausing is a ubiquitous mechanism to control gene expression which relies on a *stably* paused Pol II complex which can be acted on to cause pause release (Adelman and Lis, 2012; Jonkers and Lis, 2015; Jonkers et al., 2014; Rahl et al., 2010). However, it has more recently been suggested that the high density of Pol II revealed by ChIP-seq reflects a dynamic process of premature termination (Erickson et al., 2018). This is supported by *in vivo* imaging studies showing that the majority of Pol II is associated with chromatin for seconds (Cisse et al., 2013; Li et al., 2019; Steurer et al., 2018) and modeling which revealed that the turnover of Pol II on housekeeping genes was comparable to highly "paused" genes such as HSPA1 (Krebs et al., 2017). The kinetics identified in these studies are not easily reconciled with a model of stably paused Pol II awaiting activation by cellular signals. Whilst it is clear from many studies that enriched levels of Pol II are often found near to TSSs, the question remains to what extent does this accumulation contribute to changes in gene expression?

To understand how levels of Pol II pausing relate to gene expression we used intronRNA as a proxy for nascent gene expression (Gaidatzis et al., 2015), a method that does not require extensive material preparation or metabolic labelling of cellular RNA using nucleotide analogues such as 4sU, which are cytotoxic (Burger et al., 2013). To further validate the use of intronRNA we compared its readout to other widely used nascent RNA methods GRO-seq, mNET-seq, polyA+, polyA- and total RNA (without polyA selection) in K562 cells (Figure S3A). This showed a high correlation between GRO-seq and mNET-seq (R=0.77). Total RNA correlates best with these established nascent RNA methods (R = 0.76 and 0.71) respectively. polyA+ also correlates highly (R= 0.62 and 0.65 respectively). polyA- also shows a good correlation (R= 0.65 and 0.54). To further examine whether introns derived from total RNA, polyA+ or polyA-fractions gave the best representation of nascent RNA we plotted intronRNA over all Refseq genes and reveal that polyA+ data provides the most uniform coverage of genes (Figure S3B). Together these data suggest that deriving intronRNA counts from polyA+ RNA represents a good measure of nascent expression when compared to GRO-seq and mNET-seq but does not require material preparation or labelling which may perturb transcriptional states.

To investigate how changes in the level of Pol II pausing contribute to changes in gene expression *in vivo* in mammalian cells, we used a mouse primary erythroid cell differentiation system (Hay et al., 2016), to generate datasets at 0 hour (0h) and 24 hour (24h) differentiation time points representing “early” and “late” erythropoiesis (Edling and Hallberg, 2007; Hay et al., 2016; McGrath et al., 2017). We generated polyA+ RNA-seq libraries at 0h and 24h time points and calculated intronRNA coverage of all genes. We excluded the first 300bp of each gene from analysis to avoid counting intronRNA which overlapped scaRNA at the promoter region, we then performed differential count analysis of the intronRNA counts identifying significant changes in gene expression genome-wide (Figure S4A and S4B) (Anders and Huber, 2010).

We generated corresponding scaRNA-seq libraries at 0h and 24h time points and called all TSSs to form a consensus observed TSS list. We paired each observed TSS to an annotated TSS (UCSC) if one fell within ±500bp and plotted the distance between observed and annotated TSSs (Figure S2C), revealing an enrichment of observed TSSs within ±100nt of the annotated TSS and suggesting in mouse fetal liver TSS annotation is skewed to a lesser extent than in K562. Based on this analysis, a window of ±500 bp was drawn around each annotated TSS to capture the observed TSS position and differential count analysis of scaRNA molecules between the 0h and 24h timepoints performed, revealing significantly differentially expressed TSSs genome-wide (Anders and Huber, 2010) (Figure S4C).

As outlined in Figure 1A and B by pairing changes in nascent initiation and pausing (measured by scaRNA-seq) and nascent gene expression (measured by intronRNA), the predominant mode of transcriptional regulation can be identified genome-wide. Concordant changes in these values indicate regulation through initiation whilst discordant imply regulation through pausing. We paired changes in scaRNA and intron RNA and plotted changes in these respective values to yield a scatter plot (Figure 3A). Importantly, we show that 97% (317/327) of genes which show statistically significant changes in gene expression (red dots) have concordant changes in scaRNA and intronRNA, indicating that regulation through initiation is the dominant regulatory step in achieving differential gene expression at these genes. For example, both *Npm1* and *Tfrc* have characteristic accumulations of scaRNA at their TSSs consistent with Pol II pausing. Nevertheless, the total level of scaRNA changes significantly and concordantly with intronic RNA level at these genes indicating regulation through initiation (Figure 1B, C) and (Figure 3B, C).

Only 3% (10/327) of significantly differently expressed genes have discordant changes in scaRNA and intronic RNA levels that would be consistent with regulation through changes in Pol II pausing. Comparing the magnitude of expression changes between genes regulated at initiation versus at pausing reveals the former class has a higher mean log fold change; 2.00 versus 1.30 and a larger range of expression change; -4.63 – 7.01 versus 1.03 – 1.84, indicating at the few genes where Pol II pausing does drive significant a change in gene expression change it does so in a more modest fashion than at genes where initiation is the primary regulatory stage.

Further information about the nature of gene regulation genome-wide can be gleaned from this analysis. For genes with statistically significant changes in expression (red dots) which have concordant scaRNA and intronRNA changes (regulated at initiation), there are marked differences in the spread of the data when comparing upregulated (scaRNA and intronRNA positive fold change) versus down regulated genes (scaRNA and intronRNA negative fold change). For upregulated genes there is good correlation between scaRNA and intronRNA level (R^2=0.49), whereas for genes which decrease in their expression there is a slight negative correlation between intronRNA and scaRNA (R^2=-0.24). This suggests that as genes are down regulated through initiation the relationship between scaRNA and intronRNA is less well correlated than when upregulated. However, despite these differences in the spread of data, as the relationship between intronRNA and scaRNA is concordant not discordant we can be confident that downregulation of expression is occurring through reduced initiation not increased pausing.

These data contrast with previous analyses which have suggested that ubiquitous stably paused Pol II acts as the gatekeeper of gene expression (Chen et al., 2018; Core and Adelman, 2019; Henriques et al., 2013; Jonkers et al., 2014; Shao and Zeitlinger, 2017). We show that that although Pol II pausing can be detected genome-wide it rarely significantly affects gene expression as it changes during differentiation *in vivo*. As 67% of all genes (and 97% of statistically significant changes in expression) have concordant scaRNA and intronRNA changes, the modulation of initiation alone can explain the majority of changes in gene expression *in vivo*. As these high-resolution data are generated using primary mouse erythroid tissues, without the use of chemical or genetic perturbation of the factors postulated to be involved in regulating pausing, we suggest that it represents a more physiological estimate of the role of Pol II pausing in modulating gene expression *in vivo*.

Globin enhancers modify transcription initiation to regulate gene expression

To examine this in more detail we investigated how the α- and β-globin genes are regulated by their well characterized enhancers (Bender et al., 2012; Gariglio et al., 1981; Lee et al., 2015; Sawado et al., 2003; Vernimmen, 2014). The mouse α-globin locus is regulated by five enhancers (R1-R4 and Rm) (Figure 4A). Two of the enhancers (R1 and R2) when deleted in combination (ΔR1R2) reduce nascent α-globin expression by ~95% (Hay et al., 2016). The mouse β-globin genes are regulated by six enhancers (HS1-HS6) (Figure 4A). Deletion of two of these (HS2 and HS3) in combination (ΔHS2HS3) results in a ~70% reduction in nascent β-globin expression (Bender et al., 2012). Therefore, R1 and R2, and HS2 and HS3 are major transcriptional enhancers of the α-globin and β-globin genes respectively. Previous work suggested that the human α-globin enhancers promote transcriptional initiation via recruitment of components of the pre-initiation complex and Pol II (Vernimmen et al., 2007). By contrast, the β-globin enhancers have been reported to facilitate release of paused polymerase (Bender et al., 2012; Gariglio et al., 1981; Hay et al., 2016; Sawado et al., 2003). Therefore, the α- and β-globin genes provide an ideal test bed to understand the mechanisms by which enhancers regulate the transcription cycles of their target genes.

To determine the effects of enhancer deletions on initiation and Pol II pausing, we performed scaRNA-seq in wildtype, ΔR1R2 homozygous mice (α-globin locus) and ΔHS2HS3 homozygous mice (β-globin locus) using primary erythroid cells from fetal liver. Deletion of enhancer elements at both loci results in a substantial decrease of scaRNA-seq transcripts at their TSSs, the total number of scaRNA molecules at the β globin genes reduced by 53% and at the α globin genes by 93% (Figure 4B, 4C and Figure S5) indicating that in the absence of their major enhancers, these genes either fail to recruit Pol II or initiate production of capped transcripts, or that these transcripts are more rapidly turned over via increased premature termination). These findings apply to both loci with a greater effect seen at the α globin locus. This is in line with previous work which estimated that ΔR1R2 has a greater effect on the total nascent RNA output of the α globin genes, than ΔHS2HS3 has on the β genes (95% and 70% reduction respectively) (Bender et al., 2012; Hay et al., 2016). In wildtype cells, the globin genes in both loci exhibit an accumulation of RNA 3' ends downstream of the promoters, indicative of Pol II pausing (Nechaev et al., 2010), yet increased levels of Pol II pausing do not occur in the enhancer knockouts, which would be represented as an increase in scaRNA transcripts (see Figure 1A and 1B).

To extend our previous observations that the ΔR1R2 deletions decrease the nascent output of the α-globin genes (Hay et al. 2016) we performed mNET-seq to measure RNA associated with Pol II through immunoprecipitation and sequencing (Nojima et al., 2015) to assess perturbation of the transcription cycle downstream of the promoter proximal region. These data corroborate our observations with scaRNA-seq. mNET-seq shows a uniform loss of both promoter proximal and gene body transcription. This shows that the enhancer deletions do not induce increased Pol II pausing at these loci, but rather lead to a total reduction in engaged Pol II (as measured by the RNA associated with it) at all points along the genes (Figure 4B and 4C).

The mechanism by which enhancers regulate the initiation of transcription

Given the loss of initiation we see in the ΔR1R2 and ΔHS2HS3 mouse models, we considered the potential mechanisms by which enhancers might affect the earliest stages of the transcription cycle. At many genes, close physical contact between the enhancer and promoter is associated with gene activation. Using Capture-C, we have previously shown that the remaining mouse α-globin enhancers maintain tissue-specific increased proximity to their cognate promoters when individual or pairs of enhancers are deleted from the cluster of enhancers (Hay et al. 2016; Figure 5A). Here we show, that this is also the case at the mouse β-globin cluster (Figure 5B). We also confirmed using ATAC-seq that the α- and β-globin enhancers are not required to maintain nucleosome free promoters in the ΔHS2HS3 (β-globin and ΔR1R2 (a-globin) models (Hay et al. 2016; Figure 5A, B and S6). Therefore, we conclude that the severe reduction in promoter proximal and full-length transcription seen in the absence of the globin enhancers is not simply due to a failure to form accessible chromatin or chromatin loops.

It seems more likely that the α- and β-globin enhancers act to recruit or stabilize Pol II at nucleosome free promoters. Previous studies at the human α-globin gene cluster showed a reduction of components of the PIC and Pol II in the absence of the enhancers (Vernimmen et al., 2007). To further test this hypothesis, we performed ChIP-seq of total Pol II in wildtype, ΔR1R2, and ΔHS2HS3 lines. This revealed that in the absence of the enhancers, at both the α- and β-globin loci, there is a reduction in the total level of Pol II both at the promoter and across the gene, but no evidence for a specific increased promoter proximal block to elongation (Figure 6A and Figure 6B). This suggests that in the absence of enhancers there is a failure to recruit or stabilize Pol II at the promoter leading to decreased initiation of transcription.

Discussion

It is generally accepted that genes are switched on during differentiation via interactions between tissue- and developmental stage-specific enhancers and their cognate promoters. By counting promoter proximal RNA transcripts and nascent RNA transcripts, the relative contributions of initiation and pausing become clear (Figure 1B) and we have shown that expression of RNA from most genes during differentiation is regulated via the recruitment and/or initiation of transcription rather than via Pol II pausing. We further support this model using detailed studies of mouse models of the globin genes with and without their major transcriptional enhancers. We show large uniform loses in nascent RNA (scaRNA-seq and mNET-seq) (Figure 4B and C) and Pol II (ChIP-seq) (Figure 6A and B) across both the promoter proximal regions and the associated gene bodies. This provides strong evidence that these enhancers act primarily at the stage of recruitment and/or initiation of transcription rather than through release of paused Pol II as previously postulated (Bender et al., 2012).

We hypothesize that the loss of transcription and Pol II occupancy seen in the globin enhancer deletion models (Figure 4B and C) may be due to a failure to efficiently form or stabilize the PIC and/or to recruit Pol II to these complexes. Deletion of the α- and β-globin enhancers did not abolish chromatin accessibility (Figure 5A and B), and we can therefore hypothesize that the protein complexes which enable recruitment of the

PIC may still bind and act on the promoter. This suggests that these enhancers may play a subsequent role in activating these factors, delivering them to open promoters or increasing the rate of Pol II recruitment. All three processes could result in higher rates of transcriptional initiation (Coulon et al., 2013; Li et al., 2016).The fact that the overall chromatin conformation of the α- and β-globin loci were largely unperturbed in deletion models of their primary transcriptional enhancers (Figure 5A and B) suggests that these elements either do not function or, perhaps more likely are redundant in the assembly of the chromatin compartments but act within the assembled compartment. Enhancers may provide a means to increase concentrations of general transcriptional machinery and Pol II in the proximity of a target gene, thus increasing the likelihood of initiation at that target gene (Coulon et al., 2013).

Others have previously demonstrated a potential regulatory role for Pol II pausing by inhibiting factors such as NELF and DSIF. However, it is not clear that inhibitors of these factors are specific. DRB, one widely used inhibitor used to measure the extent of Pol II pausing in previous work, is in fact an ATP analogue that may inhibit other energetic processes in the cell, making conclusions from these studies difficult to interpret (Bensaude 2011). A further assumption is that the function of the protein targeted by chemical inhibition (CDK9) solely regulates Pol II pausing *in vivo*, disregarding off target effects on other processes in the cell which have been shown to include direct regulation of transcription termination through XRN2 (Sansó et al., 2016). Therefore, experiments using inhibitors may have over-estimated the role of Pol II pausing *in vivo* compared to our data, which comes from the differentiation of primary cells and *in vivo* mouse models without using such inhibitors. Clearly, if CDK9 has roles beyond pausing, then inhibition of this protein may have led to overestimates of the prevalence and significance of Pol II pausing. Interestingly, modeling of ChIP-seq density (Ehrensberger et al., 2013) has also suggested that promoter proximal peaks of Pol II need not be explained by a stable regulated pause but other kinetic mechanisms such as the dynamic interactions between recruitment, initiation and pausing of Pol II. This model also suggests that initiation must be saturated before Pol II can become regulated at the pause release step, implying that regulation of gene expression via pausing would only occur at extremely highly expressed genes such as alpha and beta globin in erythroid tissues.

Recent work has suggested that an alternative mechanism of premature termination of Pol II near the promoter occurs at many genes where canonical pausing is predicted to be unstable (Elrod et al., 2019; Tatomer et al., 2019). This cycle of initiation and premature termination may explain why Pol II is often found to be enriched at promoters (by ChIP or run on sequencing) in the absence of productive elongation as an alternative to stably paused Pol II. However, it should be noted that Triptolide inhibition and the pausing index was used to infer changes in pausing (Elrod et al., 2019). Triptolide is known to stimulate rapid and substantial degradation of Pol II molecules through ubiquitination as well as inhibit Pol I transcription (Bensaude, 2011; Manzo et al., 2012; Steurer et al., 2018). In another study, a three-day RNAi knockdown of an Integrator subunit was used to investigate the role of this complex in premature termination (Tatomer et al., 2019). However, the integrator complex is known have additional roles in termination and 3’ end processing of snRNA and histone genes (Skaar et al., 2015; Yamamoto et al., 2014). Therefore, in both of these studies, cellular physiology and levels of general transcription machinery may have been broadly perturbed. The extent to which premature termination of transcription near promoters occurs as opposed to Pol II pausing remains to be determined. Live cell imaging studies have suggested that transcription factors and Pol II associate with chromatin for variable but often extremely short time periods, arguing against the formation of stable protein complexes such as paused Pol II on promoters and suggesting a stochastic process wherein increasing concentrations of factors results in increased rates of transcriptional activation (Cisse et al., 2013; Erickson et al., 2018; Li et al., 2019; Steurer et al., 2018).

Clearly, given the complexity of enhancer-driven gene regulation, enhancers may have multiple effects on the transcription cycle in which sequential steps appear to be interdependent. The effect of an enhancer may vary in in different cellular contexts (e.g. stress responses) and at different stages of commitment and lineage specification. However, here we show that during differentiation of a single lineage, enhancers appear to act predominantly via the recruitment and/or transcriptional initiation of Pol II. In summary, we suggest that during differentiation enhancers predominantly control gene expression via recruitment and/or initiation of transcription, rather than regulating promoter proximal pausing of Pol II.

Star Methods Text

Experimental Model and Subject Details

All primary mouse fetal liver tissues used in this study were generated by the WIMM transgenics facility. Cell lines used in this study (K562) were obtained from the MRC WIMM transgenics facility and grown in RPMI media (ThermoFisher Scientific) supplemented with 15% FBS (37°C, 5% CO_2). Cell density was maintained <=10^6 cells per mL. Cell density and viability were determined using the NucleoCounter NC-3000 platform (Chemometec).

Method Details

Data generation

<u>Mouse fetal liver culture and differentiation:</u> ΔR1R2 deletion lines were maintained as heterozygotes and crossed to yield homozygous ΔR1R2 and wildtype litter mates as described in (Hay et al., 2016). ΔHS2HS3 deletion lines carry a human YAC containing a rescue human beta globin allele and are maintained as homozygotes as described in (Bender et al., 2012). Mouse fetal livers were harvested at E12.5 and disaggregated in StemPro expansion media (ThermoFisher Scientific) supplemented with Eprex (Epo) @ 1U/mL (Janssen), SCF @ 50ng/mL (Peprotech), Dexamethasone @ 1μM (Hameln). Cells from individual fetal livers were grown for six days, maintaining a constant erythroblast cell count of <= 1 x10^6 cells per mL. Cells genotypes were confirmed during expansion. Cells were counted using the NucleoCounter before selection for CD117 positive cells using magnetic assisted cell separation LS columns and CD117 MicroBeads following the manufacturers protocol (Milltenyi). Cells were left to recover for 6 hours in expansion media at a density of 1x10^6 cells per mL. Aliquots of 5x10^6 cells were harvested as the ‘0h’ timepoint corresponding to early erythropoiesis whilst the remaining cells were re-suspended in differentiation media (which is expansion media with the following changes: increased levels of Eprex (5U/mL) and transferrin (0.5mg/mL), SCF and Dexamethasone removed). To induce erythroid differentiation cells were grown for 24 hours before harvesting for the ‘24h’ timepoint. At each time point, cell morphology was examined by preparing cytospins. Immunophenotypes were also determined by fluorescence activated cell sorting using the Attune NxT (ThermoFisher Scientific) using anti Ter119 (PE-Cy7 labelled) and anti CD117 (PE labelled) antibodies (Biolegend). 0h of erythropoiesis was defined as (CD117+, Ter119-) and 24h as (CD117-, Ter119+). Isolated cells were confirmed as having these appropriate erythroid stage markers (Edling and Hallberg, 2007; Hay et al., 2016; McGrath et al., 2017).

Short capped RNA sequencing (scaRNA-seq): 5x10^{6} cells were aliquoted per replicate, pelleted by centrifugation (500g, 5min, R.T) and washed once in 5mL 1xPBS before disaggregation and lysis in TRI Reagent (Merck). Samples were snap frozen in a dry ice ethanol bath and transferred to -80°C storage until use. RNA extraction was performed using phase lock gel tubes (VWR) following the manufacturers protocol resuspending in 13.5μL (DEPC treated) water. 0.5μL of RNA was diluted 1:10 in water for QC. RNA integrity number (RIN) was determined using the RNA ScreenTape System (Agilent) following the manufacturers protocol. RINs for all RNA samples used in this study were >= 9.6. RNA size selection was performed using a 10% TBE-Urea gel (Thermo Fisher Scientific) in combination with a low molecular weight ladder (NEB) to excise RNA <=300nt. The RNA was extracted from the gel using the crush and soak method (crushing of the gel fragment followed by overnight elution at 21°C with no shaking) and a final resuspension volume of 6μL. RNA 3’ adapter ligation was performed using the NEBNext Small RNA Library Prep Set for Illumina. A TRI Reagent extraction was performed, topping up the reaction to 200μL with water and using phase lock light tubes. The pellet was resuspended in 25.5μL. RNA Dephosphorylation was performed using Alkaline Phosphatase, Calf Intestinal (CIP, NEB) following the manufacturers protocol. A TRI Reagent extraction was performed as before, except the RNA pellet was resuspended in 26μL. RNA Decapping was performed in a reaction volume of 30μL for 30 mins @ 37°C using CAP-clip (Tebu bio). A TRI Reagent extraction performed as before except the pellet was resuspended in 17.5μL. 2μL T4 RNA ligase buffer (NEB), and 5μL PEG 8000 (NEB) were mixed with the resuspended RNA and used as an input for SR RT primer hybridization and 5’ SR adapter ligation using the NEBNext Small RNA Library Prep Set for Illumina. Reverse transcription of post adapter ligated RNA was performed using SuperScript III (ThermoFisher Scientific), in a final reaction volume of 50μL and incubating for 1 hour at 50°C. PCR of cDNA was performed using the NEBNext Small RNA Library Prep Set for Illumina in a final reaction volume of 120μL (60μL cDNA reaction, 60μL LongAmp Taq 2x, 5μL SR Primer, 5μL Index (X) Primer). PCR reaction cleanup was performed using 2:1 ratio of Ampure XP beads (Beckman) to PCR product and a final elution volume of 23μL 1 x T.E (Tris-HCl pH 8.0 (10mM), EDTA pH 8.0 (1mM)).

K562 scaRNA-seq libraries were produced without replication for meta-analysis. All mouse fetal liver scaRNA-seq libraries were produced in biological triplicate.

mNET-seq: mNET-seq was performed as in (Nojima et al., 2015) but with a staring input cell number of 5x10^6.

K562 mNET-seq libraries were produced without replication for meta-analysis. All mouse fetal liver mNET-seq libraries were produced in biological duplicate.

RNA-seq: 5x10^6 cells were aliquoted per replicate, pelleted by centrifugation (500g, 5min, R.T) and washed once in 5mL 1xPBS before disaggregation and lysis in TRI Reagent (Merck). Samples were snap frozen in a dry ice ethanol bath and transferred to -80°C storage until use. RNA extraction and DNase treatment were performed using the Direct-zol RNA Miniprep Kit (Zymo). The sample concentration was determined using the Qubit RNA BR Assay (ThermoFisher Scientific) and the RIN using the RNA ScreenTape System (Agilent). RINs for all RNA samples used in this study were >= 9.6. 2.0mg of Total RNA was depleted of ribosomal and globin RNA species with the Globin-Zero Gold rRNA Removal Kit (Illumina). polyA+ selection was performed using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB). RNA library preparation was performed using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB).

K562 scaRNA-seq libraries were produced without replication for meta-analysis. All Mouse fetal liver scaRNA-seq libraries were produced in biological triplicate.

ChIP-seq: 5x10^6 cells were aliquoted per replicate, pelleted by centrifugation (500g, 5min, R.T) and washed once in 5mL 1xPBS. Cells were resuspended in 9mL of RPMI media (+10% FBS). Cells fixation was performed by adding 1mL of fixation solution (50mM HEPES pH 8.0, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 100mM NaCl, 10% formaldehyde) and incubating on the roller at room temperature for 10 mins. Quenching was performed by adding of 1.25mL of fresh glycine solution (1M) and incubating on the roller at room temperature for 5 mins. Cells were centrifuged again and washed with ice cold 1xPBS and snap freezing in a dry ice ethanol bath. Samples were stored at -80°C until use. Cell pellets were lysed by resuspending in 200µL of cell lysis buffer (5mM PIPES pH 8.0, 85mM KCl, 0.5% IGEPAL-CA 630) and incubating on wet ice for 20 mins. The lysate was centrifuged (900g, 10 mins, 4°C), the supernatant discarded and the pellet resuspended in 100µL nuclear lysis buffer (50mM Tris-HCl pH 7.5, 10mM EDTA pH 8.0, 1% SDS) before incubating on ice for 5 mins. 30 µL of low salt RIPA buffer (140mM NaCl, 10mM Tris-HCl pH 7.5, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 1% Triton X 100, 0.1% Sodium deoxycholate, 0.1% SDS) was added to each sample. The samples were sonicated for 8 mins using a Covaris S220 sonicator (Duty cycle = 2%, Int = 3.0, Cycles/burst = 200, Power mode = frequency sweeping, Duration = 120s, Temp = 6°C). Insoluble material was pelleted by centrifugation (20,000g, 20 mins, 4°C) and the supernatant topped up to 1.1mL with low salt RIPA buffer without SDS. 100µL of the sample was frozen to use as an input control. 110µL of a 1:1 Protein A and Protein G Dynabead slurry (ThermoFisher Scientific) was prepared by combining equal volumes of the Protein A and G beads and washing twice in low salt RIPA buffer. The remaining sonicated sample was pre-cleared twice by incubation on a rotating platform (10RPM, 1 hour, 4°C) with 5µL of the A+G bead slurry. 10µL of Pol II antibody (N20 – sc899c, Santa Cruz) was conjugated to 100µL of the bead slurry on a rotating platform (10RPM, 1 hour, 4°C). The precleared sample and antibody conjugated beads were combined and incubated on a rotating platform (10RPM, 24 hours, 4°C). The beads were washed at 4°C on a magnetic Eppendorf rack; twice using low salt RIPA buffer (140mM NaCl, 10mM Tris-HCl pH 7.5, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 1% Triton X 100, 0.1% Sodium deoxycholate, 0.1% SDS), twice using high salt RIPA buffer (0.5M NaCl, 10mM Tris-HCl pH 7.5, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 1% Triton X 100, 0.1% Sodium deoxycholate, 0.1% SDS), once using LiCl RIPA buffer (250mM LiCl, 10mM Tris-HCl pH 7.5, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 1% Triton X 100, 0.1% Sodium deoxycholate, 0.1% SDS) and twice using 1xT.E (10mM Tris-HCl pH 8.0, 1mM EDTA pH 8.0). After the final wash the beads were resuspended in 150μL of elution buffer (20mM Tris-HCl pH 7.5, 5mM EDTA pH 8.0, 50mM NaCl, 1% SDS) and transferred to a fresh Eppendorf tube. The input samples were thawed and 200 μL of elution buffer added. RNAse A treatment was performed by adding 2μL or 1μL of RNAse A @ 10mg/mL (ThermoFisher Scientific) to the input or the beads respectively. Samples were incubated (1,000rpm, 30s on, 30s off (interval mix)) for 30 mins, at 37°C. 1μL or 2μL of Proteinase K (20mg/mL) was added to the beads or inputs respectively and samples were incubated (1,000rpm, 30s on, 30s off (interval mix)) at 65°C, overnight. Samples were centrifuged briefly to collect droplets. Samples containing beads were transferred to a magnetic rack and the supernatant was transferred to a new Eppendorf tube. A phenol chloroform extraction was performed using phase lock tubes followed by ethanol precipitation on samples and inputs. Samples and inputs were resuspended in 53μl 0.1 x T.E. Library preparation was performed using the NEB Ultra DNA library preparation kit for Illumina (NEB) and using 15 cycles of PCR followed by an Ampure XP bead (Beckman) clean-up.

All Mouse fetal liver ChIP-seq libraries were produced in biological triplicate.

ATAC-seq: ATAC-seq was performed as previously described (Buenrostro et al., 2013; Hay et al., 2016). Briefly, 100000 cells per biological replicate (x3) per sample were lysed in cold lysis buffer, nuclear pellets were obtained after 10 min centrifugation at 4°C at 500G and resuspended in 50ul of tagmentation mix (FC-121-1030, Illumina), then incubated for 30 min at 37°C. DNA was purified using the Qiagen MinElute columns (28004, Qiagen). Tagmented DNA was indexed with custom primers using NEB Next High-Fidelity 2x PCR Master Mix (M0541S, NEB). And purified with Qiagen PCR Cleanup Kit (28104, Qiagen). Samples were multiplexed sequenced on a next generation sequencing platform using the NextSeq® 500/550 High Output Kit v2 (75 cycles; FC-404-2005, Illumina) using paired-end reads.

NG-Capture-C: 5 x10^6 HS2HS3 cells were fixed using a final concentration of 2% formaldehyde in 1x PBS. 3C libraries were prepared using *DpnII,* as in (Davies et al., 2015), with the following modifications: no douncing was performed, all spins were performed at 300g, an additional centrifugation step (300g, 15 min, 4°C) was performed to pellet nuclei after ligation before resuspending in 300 μL 1x T.E (Sigma) for phenol chloroform extraction and each capture oligonucleotide was used at a working concentration of 2.9nM.

Digestion efficiency was determined by RT-qPCR with custom TaqMan probes. Ligation efficiency qualitatively determined by gel electrophoresis. Only 3C libraries with >70 % digestion efficiency were used. Target capture for specific genomic viewpoints was performed with 70mer biotinylated oligonucleotides designed using CapSequm (http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi) which targeted either promoter proximal or promoter distal *DpnII* fragments.

All NG Capture-C experiments were performed in biological triplicate.

Library QC and sequencing: For all libraries, the modal DNA library size and library concentrations were determined using the D1000 reagents on a Tapestation (Agilent) and the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) respectively. Libraries were quantified by qPCR using the NEBNext Library Quant Kit following the manufacturer's instructions. Library concentration by qPCR was used to calculate dilution required to produce 6nM libraries for sequencing on the NextSeq500/550 High Output v2 kit (75 cycles) (Illumina). Sequencing was performed in paired end mode using 40 cycles of sequencing per read and a 6-cycle index read.

Quantitation an Statistial Analyses

Deriving annotated genes from UCSC table browser—Lists of GenBank, Refseq and UCSC genes were download from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) in .bed format for mm9 or hg19 genome builds.

Generation of unique UCSC TSS lists—TSS of annotated transcripts were extracted and extended to 1 kb (+/- 500 bp). Chr M entries were discarded. Exact overlaps on the same strand were collapsed. Unique TSS were clustered within +-500 in a strand aware manner (bedtools cluster -s)(Quinlan and Hall, 2010) and the number of clustered TSS' was counted.

Generation of non-coding RNA list—Lists of Ensembl non coding gene coordinates (for mm9 and hg19 genes respectively) were downloaded from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables), using the following filtering terms; "miRNA", "tRNA", "snoRNA", "Mt_tRNA", "Mt_rRNA", "rRNA", "snoRNA", "snRNA". Lists of UCSC non-coding gene coordinates were also downloaded using the following filtering terms; "snRNA", "miRNA", "tRNA", "snoRNA", "Mt_tRNA", "Mt_rRNA", "rRNA", "snoRNA", "snRNA". All coordinates were combined into a consensus list for either the mm9 or hg19 genome and windowed by 100nt up or downstream using bedtools slop.

Generation of splice sites list—A list of UCSC genes were downloaded from the table browser (http://genome.ucsc.edu/cgi-bin/hgTables) in GTF format and splice sites extracted from the exon coordinates using AWK.

scaRNA-seq—The script scaRNAseq_pipe.pl was used to perform the following analysis procedures: alignment of the data to the relative reference genome using Bowtie/1.1.2 (Langmead et al., 2009). This involved three alignment stages (first pass alignment, second pass adapter trimming of reads which did not map in first pass and realignment, third pass FLASH-ing (Magoč and Salzberg, 2011) of reads which did not align second pass and realignment). All aligning reads were combined and remapped. Reads mapped in proper pairs were isolated from this file as a .bam file using samtools/0.1.19 view and flags (-bS -f 3)(Li et al., 2009). The bam file was converted into a bedpe file using bedtools/2.25.0 (Li et al., 2009; Quinlan and Hall, 2010) bamtobed. Chromosome, start, end and strand of each read was inferred from the. bedpe file allowing RNA molecules to be “reconstructed” an output in .bed6 file using AWK. Files were sorted (-k1,1 -k2,2n) and features overlapping noncoding RNAs and annotated exon 3’ splice sites were removed using bedtools/2.25.0 intersect. Total read counts at this stage were used to subsample down to the sample with the lowest read coverage before (Unix shuf). These files were converted into strand specific bigWigs of 5’ and 3’ most bases (four files in total) using bedtools genomecov and ucsctools bedGraphToBigWig (Kent et al., 2002). bigWigs for the same genotype, strand and end are merged bigWigMerge (Kent et al., 2002) before visualization on the UCSC genome browser.

For visualization of whole reconstructed RNA molecules, a region of interest was specified in. bed format and the script frags2bed.pl used to isolate RNA molecules overlapping this region entirely. RNA molecules were scored based on frequency to produce a color gradient before visualization on UCSC as .bed files. K562 scaRNA-seq data was not normalized prior to visualization.

scaRNA TSSs were called by creating a Homer Tag Directory (makeTagDirectory -format bed -sspe - keepAll -genome hg19/mm9) from a reconstructed RNA molecule file and calling peaks with (findPeaks - style tss) (Heinz et al., 2010). All peaks mapping to X, Y, M and random chromosomes were removed using grep -v.

Meta profiles of scaRNA-seq coverage were created by plotting the distribution of 5’ or 3’ ends of reads in 1bp bins relative to their respective TSS (+/- 500 bp) in a strand specific manner. Per TSS per position were restricted to a maximum of 3.

Observed TSS were defined as called scaRNA peaks which were located within +/- 500 bp from the 5’ end of an annotated unique UCSC gene using bedtools. To avoid any skews in the meta-profiles, observed TSS were filtered to retain only those with single a scaRNA peak within +/- 500 bp and only overlapped with a single unique annotated TSS. Annotated TSS were defined as the original 5’ end coordinate of the gene for which an observed TSS could be identified. observed TSS were filtered to retain only those with a single unique annotated TSS within +/- 500 bp and only overlapped with a single observed TSS. Distances between observed and annotated TSS were recorded and plotted.

Motif enrichment around observed or annotated TSSs was performed using a custom Perl script (SeqPile2.pl). The coordinates of the TSSs were used to retrieve sequence information for a user specified window around this position. The sequence was scanned in a user specified bin size (1bp) to identify specific user specified motifs (also using the reverse of the user specified motif to account for strand). The sum of the motif occurrence was plotted for each bin, using the 5’ end of the motif as the point plotted (i.e. a 2nt motif ends at the TSS are plotted at -2 relative to the TSS).

A consensus list of observed mouse fetal liver TSS coordinates was produced by calling TSS in each of the biological triplicates for the early (0h) and late (24h) erythroid stages (n=6 peaks calls) resulting in n= 65451 TSS. These TSSs were merged and exact overlaps collapsed. ChrM entries were discarded resulting in n = 51161 TSS. Observed and annotated TSS were defined using the clustering counts and intersects as described above.

Heatmaps of scaRNA-seq 5' and 3' ends were generated by producing bed files containing only the 5' or 3' most base of each reconstructed RNA molecule.

NG-Capture-C—Data were analyzed using scripts available at https://github.com/Hughes-Genome-Group/CCseqBasicF/releases and R was used to normalize data (Davies et al., 2015). Due to the inherent changes in linear proximity in the genome in the deletional models compared to wide-type, subtractive analysis was not performed.

ChIP-exo—ChIP-exo data (Pol II N20 antibody - sc899c) were downloaded from two separate studies; GSE108323 (Mchaourab et al., 2018) and SRA067908 (Pugh et al., 2013). Reads were aligned to hg19 using bowtie (Langmead et al., 2009) and PCR duplicates removed using samtools (Li et al., 2009). The resultant .bam file was converted into a "tag directory" using Homer makeTagDirectory.pl and the following flags (-format sam -sspe -keepAll -genome hg19). The distribution of 5' ends of reads, was plotted in 1bp bins around a given list of TSSs (+/-) 500bp) in a strand specific manner using Homer/4.8 annotatePeaks.pl and the following flags (-hist 1 -size 1000 -pc 3 -raw). Distribution of data on the reverse strand was multiplied by - 1 to flip the data on the X axis and allow easier visualization of regions occupied by Pol II.

RNA-seq—For published data, which was downloaded in fastq format, reads were trimmed to a maximum length of 50bp using AWK to remove lower quality bases from 3' ends of reads and improve alignment. New data generated in this study utilized 40bp paired end sequencing and so reads were not trimmed. For intronRNA extraction data were aligned to the relevant reference genome (mm9 or hg19) using bowtie (Langmead et al., 2009) and reads which mapped 100% to an intron and do not map to an annotated non-coding RNA were isolated using bedtools (Quinlan and Hall, 2010). The total number of reads in the resultant bam file was used to derive a reads per million (RPM) scaling factor which was applied when generating strand specific bedGraph coverage tracks with bedtools (Quinlan and Hall, 2010). Tracks from biological duplicates (K562 data) or triplicates (Mouse data) were merged prior to display using ucsctools (Kent et al., 2002).

For Total and polyA+ RNA-seq data were aligned to hg19 using STAR/2.4.2a (Dobin et al., 2013), only reads which did not map to annotated noncoding RNAs were retained. A coverage track without normalization was generated using bedtools (Quinlan and Hall, 2010). Meta profiles of intronic reads coverage over Refseq genes were generated from using ngsplots (Shen et al., 2014) and the following flags; -G hg19 -D Refseq -R genebody -RB 0.05 -MQ 0 -SE 0 -L 5000 -VLN 0 -MW 2.5.

Correlation analysis of RNA-seq techniques over annotated introns (obtained from the UCSC genome browser) was performed by counting reads mapping to introns, converting counts to RPKM values while omitting introns where zero coverage was observed in either data set under comparison. Values were plotted as a hexbin plot. Pearson correlations were calculated on the log10 transformed RPKM values.

Differential Expression Analysis—For differential expression analysis distances between observed and annotated TSS were recorded, showing an enrichment of TSSs fell within a window of -500 to +500 relative to the annotated TSS (Figure S2C). Therefore, a window of -500 to +500 was drawn relative to each UCSC gene (n = 55,397). TSS were clustered within +/- 500bp in a strand specific manner. Intron coordinates of all known isoforms were extracted from annotated introns (UCSC table browser). RNA seq coverage over introns was calculated with bedtools multicov in a strand specific manner. scaRNA fragment coverage over TSS was calculated using bedtools coverage in a strand specific manner. Within clustered annotated and observed TSS the transcript best supported by intron coverage was selected. Intron support was calculated as the average intronic read coverage per bp of combined intron length. For transcripts with equal intron support, the one with the highest scaRNA coverage support was selected, selecting one transcript randomly if those tied as well. This yielded a total of n=27,951 transcript TSS pairs. Differential expression analysis was performed with DEseq2 (Anders and Huber, 2010). TSS' with significantly differential counts (expression) were identified using a p value and false discovery threshold of <= 0.05.

ChIP-seq—ChIP-seq data were aligned to mm9 using bowtie (Langmead et al., 2009). Reads mapping in proper pairs were isolated, PCR duplicates removed using samtools (Li et al., 2009). Each read pair was converted into a reconstructed DNA fragment based their mapped coordinates. Peaks were called from sample ChIP-seq data and corresponding input controls using MACS2 (Quinlan and Hall, 2010). A consensus peak list was made from two sample sets (e.g. 3 x Wildtype versus 3 x R1R2) using bedtools, features were merged if <= 50bp apart on the same strand), and peaks which overlap input peaks were removed (Quinlan and Hall, 2010). Total millions of mapped reads under peaks was counted for each sample. Fragments were plotted as a coverage track which was normalized by fragments per million under Pol II peaks using bedtools (Kent et al., 2002). Tracks from biological triplicates were merged prior to display using ucsctools (Langmead et al., 2009).

ATAC-seq—ATAC-seq data were aligned to mm9 using bowtie (Quinlan and Hall, 2010). Reads mapping in proper pairs were isolated, PCR duplicates removed using samtools (Li et al., 2009). Each read pair was converted into a reconstructed DNA fragment based their mapped coordinates and also adjusted for Tn5 transposase cutting bias (forward strand fragments shift +4, reverse strand shift -5bp). Fragments were plotted as a coverage track which was normalized by fragments per million using bedtools (Kent et al., 2002) after the exclusion of mitochondrial reads (Figure S4). Tracks from biological triplicates were merged prior to display using ucsctools (Langmead et al., 2009).

mNET-seq—mNET-seq data were aligned to mm9 using bowtie (Quinlan and Hall, 2010). Reads mapping in proper pairs were isolated, PCR duplicates removed using samtools (Li et al., 2009). Reads mapping to annotated noncoding RNAs and splice sites were removed using bedtools/2.25.0 (Quinlan and Hall, 2010). Each read pair was converted into a reconstructed DNA fragment based their mapped coordinates using awk. The coordinate of the 3' end of each fragment (representing the active site of Pol II transcription) was extracted and windowed by +/-2nt before plotting as a normalized coverage track using bedtools (Kent et al., 2002). Tracks were normalized to the sample with the lowest total fragment count. Tracks from biological duplicates were merged prior to display using ucsctools (Langmead et al., 2009). Meta profiles of mNET-seq data over TSS were plotted over observed TSS as described for scaRNA-seq.

GRO-seq—Published GRO-seq data were downloaded and aligned to hg19 using bowtie (Quinlan and Hall, 2010). PCR duplicates were removed using samtools (Li et al., 2009) and reads mapping to non-coding RNAs removed using bedtools (Quinlan and Hall, 2010). Meta profiles of GRO-seq data over TSSs were plotted over observed TSS as described for scaRNA-seq.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank S. Butler, J. Sloane-Stanley and J. Sharpe of the WIMM Transgenics Facility, K. Clark, S-A. Clark and C. Waugh of the WIMM Flow Cytometry Facility and the MRC WIMM Centre for Computational Biology. This work was supported by the Medical Research Council grant (MC_UU_0009/4 and MR/T014067/1 to D.R.H. and MC_UU_00016/14 to J.R.H.). A Wellcome Trust Strategic Award (106130/Z/14/Z to D.R.H and J.R.H.). An ERC Advanced Grant [339270] and Wellcome Trust Investigator Award [107928|Z|15|Z] to T.N. and N.J.P. A Stevenson Junior Research Fellowship at University College, Oxford (to A.M.O.). A Sir Henry Wellcome Postdoctoral Fellowship (209181/Z/17/Z to R.A.B.). M.A.B was supported by NIH grant R37 DK44746 to M. Groudine.

Data and code availability

The datasets generated during this study are available at Gene Expression Omnibus under accession number: GSE138359. The following datasets were downloaded and reanalyzed as part of this study: K562: GRO-seq (GSE60456), ChIP-exo Pol II (GSE108323 and SRA067908), rRNA depleted polyA+ RNA-seq (GSE86660), rRNA depleted total RNA-seq (GSE90231), rRNA depleted poly- RNA-seq (GSE90231). GRO-cap (GSE60453), CAGE (ENCFF623BZZ). Mouse fetal liver: ATAC-seq (GSE78800) and Capture C (GSE78803). All novel data generated in this study are deposited in Gene Expression Omnibus under accession number: GSE138359.

All scripts used in this analysis are available at: https://github.com/martinlarke/Enhancers-predominantly-regulate-transcription-initiation-in-vivo

References

Adelman K, Lis JT. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nature Reviews Genetics. 2012; 13: 720–731.

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11 R106 [PubMed: 20979621]

Bender MA, Ragoczy T, Lee J, Byron R, Telling A, Dean A, Groudine M. The hypersensitive sites of the murine β-globin locus control region act independently to affect nuclear localization and transcriptional elongation. Blood. 2012; 119: 3820–3827. [PubMed: 22378846]

Bensaude O. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? Transcription. 2011; 2: 103–108. [PubMed: 21922053]

Bernecky C, Herzog F, Baumeister W, Plitzko JM, Cramer P. Structure of transcribing mammalian RNA polymerase II. Nature. 2016; 529: 551–554. [PubMed: 26789250]

Boswell SA, Snavely A, Landry HM, Churchman LS, Gray JM, Springer M. Total RNA-seq to identify pharmacological effects on specific stages of mRNA synthesis. Nature Chemical Biology. 2017; 13: 501–507. [PubMed: 28263964]

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013; 10: 1213–1218. [PubMed: 24097267]

Burger K, Mühl B, Kellner M, Rohrmoser M, Gruber-Eber A, Windhager L, Friedel CC, Dölken L, Eick D. 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. RNA Biology. 2013; 10: 1623–1630. [PubMed: 24025460]

Chen FX, Smith ER, Shilatifard A. Born to run: control of transcription elongation by RNA polymerase II. Nature Reviews. Molecular Cell Biology. 2018; 19: 464–478. [PubMed: 29740129]

Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, Longo SL, Corona RJ, Chin LS, Lis JT, et al. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. Nature Genetics. 2018; 50: 1553–1564. [PubMed: 30349114]

Cisse II, Izeddin I, Causse SZ, Boudarene L, Senecal A, Muresan L, Dugast-Darzacq C, Hajj B, Dahan M, Darzacq X. Real-time dynamics of RNA polymerase II clustering in live human cells. Science (New York, NY). 2013; 341: 664–667.

Clapier CR, Iwasa J, Cairns BR, Peterson CL. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nature Reviews Molecular Cell Biology. 2017; 18: 407–422. [PubMed: 28512350]

Core L, Adelman K. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes & Development. 2019; 33: 960–982. [PubMed: 31123063]

Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (New York, NY). 2008; 322: 1845–1848.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics. 2014; 46: 1311–1320. [PubMed: 25383968]

Coulon A, Chow CC, Singer RH, Larson DR. Eukaryotic transcriptional dynamics: From single molecules to cell populations. Nature Reviews Genetics. 2013; 14: 572–584.

Czudnochowski N, Bösken CA, Geyer M. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. Nature Communications. 2012; 3: 842.

Davies JOJ, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, Hughes JR. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nature Methods. 2015; 13: 74–80. [PubMed: 26595209]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. [PubMed: 23104886]

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489: 57–74. [PubMed: 22955616]

Edling CE, Hallberg B. c-Kit-A hematopoietic cell essential receptor tyrosine kinase. International Journal of Biochemistry and Cell Biology. 2007; 39: 1995–1998. [PubMed: 17350321]

Ehrensberger AH, Kelly GP, Svejstrup JQ. XMechanistic interpretation of promoter-proximal peaks and RNAPII density maps. Cell. 2013; 154: 713–715. [PubMed: 23953103]

Elrod ND, Henriques T, Huang K-L, Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. Molecular Cell. 2019; 76: 738–752. e7 [PubMed: 31809743]

Erickson B, Sheridan RM, Cortazar M, Bentley DL. Dynamic turnover of paused pol II complexes at human promoters. Genes and Development. 2018; 32: 1215–1225. [PubMed: 30150253]

Farnung L, Vos SM, Cramer P. Structure of transcribing RNA polymerase II-nucleosome complex. Nature Communications. 2018; 9 5432

Frankish A, Diekhans M, Ferreira A, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. 2019; 47: 766–773.

Fraser NW, Sehgal PB, Darnell JE. DRB-induced premature termination of late adenovirus transcription. Nature. 1978; 272: 590–593. [PubMed: 643052]

Gaidatzis D, Burger L, Florescu M, Stadler MB. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nature Biotechnology. 2015; 33: 722–729.

Gariglio P, Bellard M, Chambon P. Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. Nucleic Acids Research. 1981; 9: 2589–2598. [PubMed: 6269056]

Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, Cramer P. CDK9-dependent RNA polymerase II pausing controls transcription initiation. ELife. 2017; 6: 1–24.

Hay D, Hughes JR, Babbs C, Davies JOJJ, Graham BJ, Hanssen LLP, Kassouf MT, Oudelaar AM, Sharpe JA, Suciu MC, et al. Genetic dissection of the α-globin superen-hancer in vivo. Nature Genetics. 2016; 48: 895–903. [PubMed: 27376235]

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell. 2010; 38: 576–589. [PubMed: 20513432]

Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, Burkholder A, Fargo DC, Adelman K. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. Molecular Cell. 2013; 52: 517–528. [PubMed: 24184211]

Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. Widespread transcriptional pausing and elongation control at enhancers. Genes and Development. 2018; 32: 26–41. [PubMed: 29378787]

Hsu F, Kent JW, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics. 2006; 22: 1036–1046. [PubMed: 16500937]

Jonkers I, Lis JT. Getting up to speed with transcription elongation by RNA polymerase II. Nature Reviews Molecular Cell Biology. 2015; 16: 167–177. [PubMed: 25693130]

Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. ELife. 2014; 2014: 1–25.

Kamieniarz-Gdula K, Proudfoot NJ. Transcriptional Control by Premature Termination: A Forgotten Mechanism. Trends in Genetics. 2019; 35: 553–564. [PubMed: 31213387]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Research. 2002; 12: 996–1006. [PubMed: 12045153]

Kim TH, Barrera LO, Zheng M, Qu C, Singer Ma, Richmond Ta, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. Nature. 2005; 436: 876–880. [PubMed: 15988478]

Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Molecular Cell. 2017; 67: 411–422. e4 [PubMed: 28735898]

Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (New York, NY). 2013; 339: 950–953.

Laitem C, Zaborowska J, Isa NF, Kufs J, Dienstbier M, Murphy S. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. Nature Structural and Molecular Biology. 2015; 22: 396–403.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10 R25 [PubMed: 19261174]

Lee K, Hsiung CC, Huang P, Raj A, Blobel GA. Dynamic enhancer - Gene body contacts during transcription elongation. Genes & Development. 2015. 1992–1997. [PubMed: 26443845]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. [PubMed: 19505943]

Li J, Dong A, Saydaminova K, Chang H, Wang G, Ochiai H, Yamamoto T, Pertsinidis A. Single-Molecule Nanoscopy Elucidates RNA Polymerase II Transcription at Single Genes in Live Cells. Cell. 2019; 178: 491–506. e28 [PubMed: 31155237]

Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. Nature Reviews Genetics. 2016; 17: 207–223.

Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics (Oxford, England). 2011; 27: 2957–2963.

la Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. RNA velocity of single cells. Nature. 2018; 560: 494–498. [PubMed: 30089906]

Manzo SG, Zhou ZL, Wang YQ, Marinello J, He JX, Li YC, Ding J, Capranico G, Miao ZH. Natural product triptolide mediates cancer cell death by triggering CDK7-dependent degradation of RNA polymerase II. Cancer Research. 2012; 72: 5363–5373. [PubMed: 22926559]

Marshall NF, Price DH. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. The Journal of Biological Chemistry. 1995; 270: 12335–12338. [PubMed: 7759473]

McGrath KE, Catherman SC, Palis J. Delineating stages of erythropoiesis using imaging flow cytometry. Methods. 2017; 112: 68–74. [PubMed: 27582124]

Mchaourab ZF, Perreault AA, Venters BJ. ChIP-seq and ChIP-exo profiling of Pol II, H2A.Z, and H3K4me3 in human K562 cells. Scientific Data. 2018; 5 180030 [PubMed: 29509191]

Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S, Deaton AM, Urban JA, Larschan E, Park PJ, et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. Nature Communications. 2016; 7 11485

Muse GW, Gilchrist Da, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. RNA polymerase is poised for activation across the genome. Nature Genetics. 2007; 39: 1507–1511. [PubMed: 17994021]

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science (New York, NY). 2010; 327: 335–338.

Nilson KA, Guo J, Turek ME, Brogie JE, Delaney E, Luse DS, Price DH. THZ1 Reveals Roles for Cdk7 in Co-transcriptional Capping and Pausing. Molecular Cell. 2015; 59: 576–587. [PubMed: 26257281]

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. Cell. 2015; 161: 526–540. [PubMed: 25910207]

Nozawa K, Schneider TR, Cramer P. Core Mediator structure at 3.4 Å extends model of transcription initiation complex. Nature. 2017; 545: 248–251. [PubMed: 28467824]

O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016; 44: D733–45. [PubMed: 26553804]

Porrua O, Libri D. Transcription termination and the control of the transcriptome: why, where and how to stop. Nature Reviews Molecular Cell Biology. 2015; 16: 190–202. [PubMed: 25650800]

Pugh BF, Venters BJ, Pugh BF. Genomic Organization of Human Transcription Initiation Complexes. PloS One. 2013; 502 e0149339

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010; 26: 841–842.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. c-Myc regulates transcriptional pause release. Cell. 2010; 141: 432–445. [PubMed: 20434984]

Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011; 147: 1408–1419. [PubMed: 22153082]

Sainsbury S, Bernecky C, Cramer P. Structural basis of transcription initiation by RNA polymerase II. Nature Reviews Molecular Cell Biology. 2015; 16: 129–143. [PubMed: 25693126]

Sansó M, Levin RS, Lipp JJ, Wang VY-F, Greifenberg AK, Quezada EM, Ali A, Ghosh A, Larochelle S, Rana TM, et al. P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates. Genes & Development. 2016; 30: 117–131. [PubMed: 26728557]

Sawado T, Halow J, Bender MA, Groudine M. The β-globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. Genes and Development. 2003; 17: 1009–1018. [PubMed: 12672691]

Schilbach S, Hantsche M, Tegunov D, Dienemann C, Wigge C, Urlaub H, Cramer P. Structures of transcription pre-initiation complex with TFIIH and Mediator. Nature. 2017; 551: 204–209. [PubMed: 29088706]

Shao W, Zeitlinger J. Paused RNA polymerase II inhibits new transcriptional initiation. Nature Genetics. 2017; 49: 1045–1051. [PubMed: 28504701]

Shen L, Shao N, Liu X, Nestler E. Ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics. 2014; 15: 1–14. [PubMed: 24382143]

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences. 2003; 100: 15776–15781.

Skaar JR, Ferris AL, Wu X, Saraf A, Khanna KK, Florens L, Washburn MP, Hughes SH, Pagano M. The Integrator complex controls the termination of transcription at diverse classes of gene targets. Cell Research. 2015; 25: 288–305. [PubMed: 25675981]

Steurer B, Janssens RC, Geverts B, Geijer ME, Wienholz F, Theil AF, Chang J, Dealy S, Pothof J, van Cappellen WA, et al. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. Proceedings of the National Academy of Sciences. 2018; 115: 4368–4376.

Tatomer DC, Elrod ND, Liang D, Xiao MS, Jiang JZ, Jonathan M, Huang KL, Wagner EJ, Cherry S, Wilusz JE. The Integrator complex cleaves nascent mRNAs to attenuate transcription. Genes & Development. 2019; 33: 1525–1538. [PubMed: 31530651]

Tome JM, Tippens ND, Lis JT. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nature Genetics. 2018; 50: 1533–1541. [PubMed: 30349116]

Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nature Reviews Molecular Cell Biology. 2015; 16: 178–189. [PubMed: 25650798]

Vernimmen D. Uncovering Enhancer Functions Using the α-Globin Locus. PLoS Genetics. 2014; 10 e1004668 [PubMed: 25330308]

Vernimmen D, de Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. EMBO Journal. 2007; 26: 2041–2051.

Vo Ngoc L, Wang Y, Kassavetis GA, Kadonaga JT. The punctilious RNA polymerase II core promoter. Genes & Development. 2017; 31: 1289–1301. [PubMed: 28808065]

Voong LN, Xi L, Sebeson AC, Xiong B, Wang JP, Wang X. Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. Cell. 2016; 167: 1555–1570. e15 [PubMed: 27889238]

Wada T, Takagi T, Yamaguchi Y, Ferdous A, Imai T, Hirose S, Sugimoto S, Yano K, Hartzog GA, Winston F, et al. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes & Development. 1998; 12: 343–356. [PubMed: 9450929]

Williams LH, Fromm G, Gokey NG, Henriques T, Muse GW, Burkholder A, Fargo DC, Hu G, Adelman K. Pausing of RNA Polymerase II Regulates Mammalian Developmental Potential through Control of Signaling Networks. Molecular Cell. 2015; 58: 311–322. [PubMed: 25773599]

Wissink EM, Vihervaara A, Tippens ND, Lis JT. Nascent RNA analyses: tracking transcription and its regulation. Nature Reviews Genetics. 2019; 17: 19–21.

Yamaguchi Y, Takagi T, Wada T, Yano K, Furuya A, Sugimoto S, Hasegawa J, Handa H. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell. 1999; 97: 41–51. [PubMed: 10199401]

Yamamoto J, Hagiwara Y, Chiba K, Isobe T, Narita T, Handa H, Yamaguchi Y. DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes. Nature Communications. 2014; 5: 1–10.

Yazdi PG, Pedersen BA, Taylor JF, Khattab OS, Chen Y-H, Chen Y, Jacobsen SE, Wang PH. Nucleosome Organization in Human Embryonic Stem Cells. PloS One. 2015; 10 e0136314 [PubMed: 26305225]

Zeitlinger J, Stark A, Kellis M, Hong J-W, Nechaev S, Adelman K, Levine M, Young Ra. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nature Genetics. 2007; 39: 1512–15. [PubMed: 17994019]

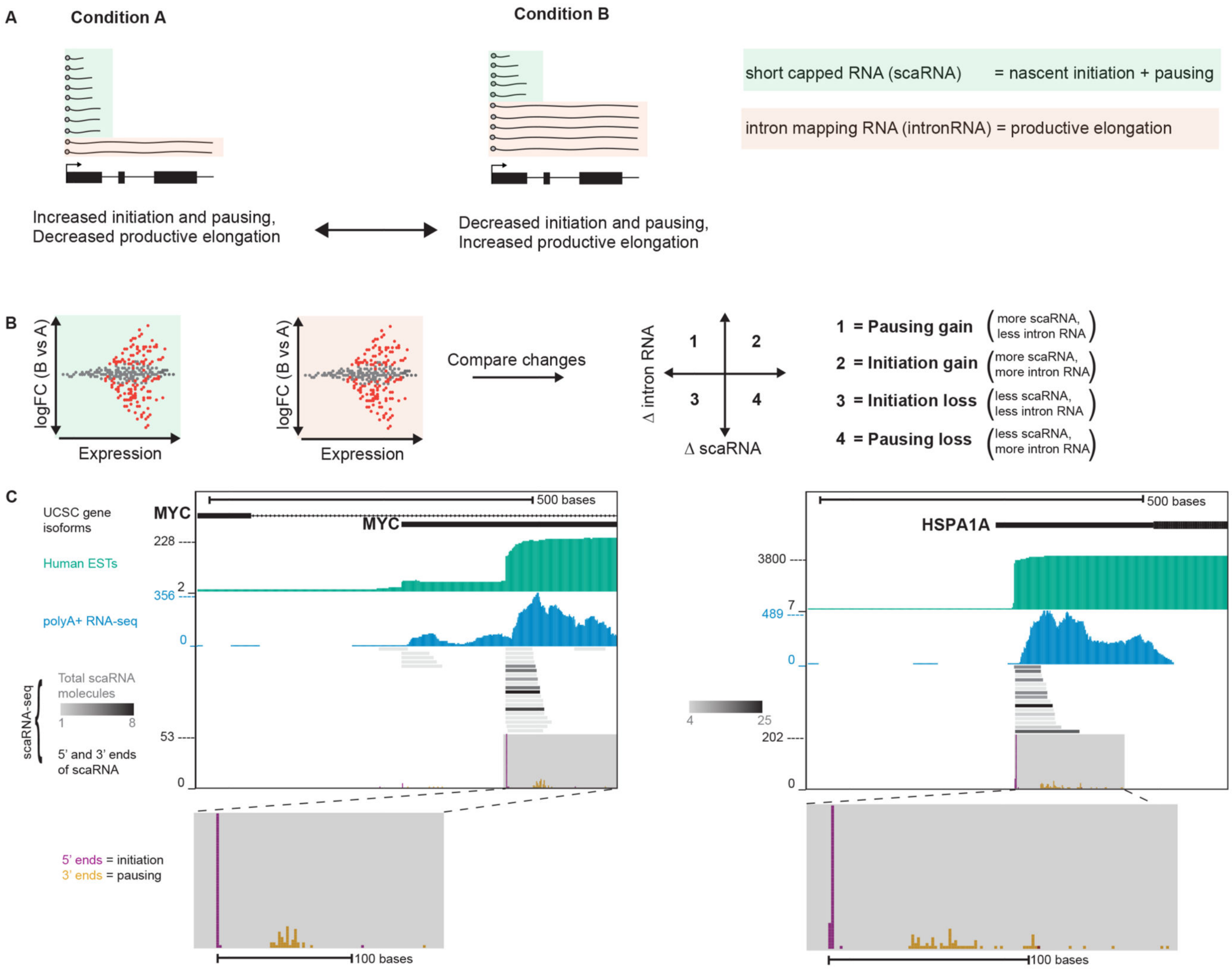
A
Condition A
Condition B
short capped RNA (scaRNA) = nascent initiation + pausing
intron mapping RNA (intronRNA) = productive elongation
Increased initiation and pausing,
Decreased productive elongation
Decreased initiation and pausing,
Increased productive elongation
B
logFC (B vs A)
Expression
logFC (B vs A)
Expression
Compare changes
Δ intron RNA
Δ scaRNA
1
2
3
4
1 = Pausing gain (more scaRNA, less intron RNA)
2 = Initiation gain (more scaRNA, more intron RNA)
3 = Initiation loss (less scaRNA, less intron RNA)
4 = Pausing loss (less scaRNA, more intron RNA)
C
UCSC gene isoforms
Human ESTs
polyA+ RNA-seq
scaRNA-seq
Total scaRNA molecules
1
8
5' and 3' ends of scaRNA
MYC
500 bases
MYC
228
2
356
0
53
0
4
25
500 bases
HSPA1A
3800
7
489
0
202
0
5' ends = initiation
3' ends = pausing
100 bases
100 bases

Figure 1. scaRNA-seq maps transcription initiation and pausing at a single molecule level *in vivo*.

(A) Total counts of scaRNA molecules in promoter proximal regions (0-300bp, relative to observed TSS) stand proxy for levels of initiation and pausing. Nascent gene expression was measured using reads mapping to introns from RNA-seq (excluding the first 300bp of a given gene) (la Manno et al., 2018).

(B) Overview of data analysis: scaRNA counts (scatter plot with green background) and intronRNA counts (scatter plot with orange background) are analyzed to find significant differential changes in expression (red dots) out of all genes (grey dots). Datasets are compared in situations when gene expression changes (0h and 24h of erythropoiesis). The change in scaRNA versus the change in intronRNA can be plotted as a scatter plot where each quadrant identifies a distinct class of regulation; 1 (pausing gain), 2 (initiation gain), 3 (initiation loss), 4 (pausing loss).

(C) scaRNA-seq in K562 cells to validate the assay with MYC (left) and HSPA1A (aka. Hsp70) (right). The figure shows (from top to bottom) annotated UCSC gene isoforms, human expressed sequence tags (ESTs), polyA+ RNA-seq in K562, the density of reconstructed RNA molecules and distributions of their RNA 5' and 3' ends. RNA

5’ and 3’ distributions indicate focused initiation of transcription and promoter proximal pausing downstream of the site of initiation (+20-60nt). All data are displayed as raw (unnormalized) counts. scaRNA-seq data were derived from $5x10^6$ K562 cells. polyA+ RNA-seq was downloaded from Encode and an isogenic duplicate merged prior to display. Data are displayed raw (without normalization). Human ESTs were obtained as a UCSC genome browser track.
See Figure S1 and S2

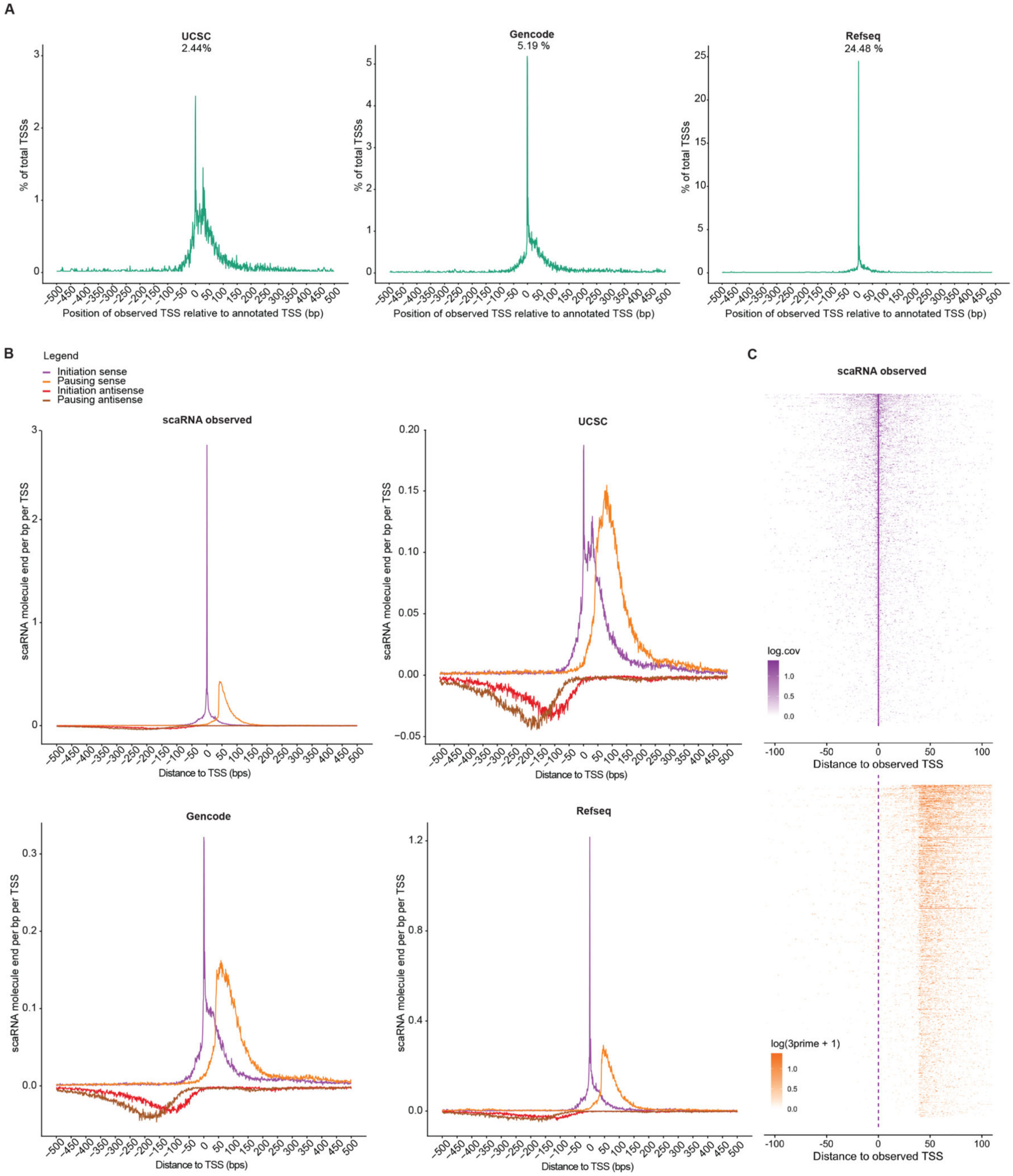

Figure 2. Annotated TSS positions are systematically skewed upstream of their *in vivo* locations.

(A) Plots of the distance between the position of observed TSSs (called from scaRNA-seq data) and annotated TSSs associated with UCSC, Gencode or Refseq gene annotations.

(B) A meta-analysis of scaRNA-seq data around observed TSS positions and associated annotated TSSs from three different annotations (Gencode, Refseq and UCSC). scaRNA observed and Refseq meta profiles show punctate initiation around the TSS, Gencode and UCSC show dispersed. 3’ RNA ends indicative of Pol II pausing peak between +50-100 bps and their distribution extends to around +150bps relative to the TSS. Antisense transcription and Pol II pausing are visible in all profiles, scaling on the Y axis gives the impression that antisense transcription occurs to a lower degree in Refseq and scaRNA observed than Gencode or UCSC. Data displayed as raw (unnormalized) coverage of RNA ends (reads per bp per TSS).

(C) Heatmaps of RNA 5’ ends (purple), RNA 3’ ends (orange). 0 bp position is highlighted with purple dotted line. Coverage is per bp from the observed TSS on the sense strand. Coverage is log (coverage +1) and capped at 55 reads per bp per TSS (log4) to aid in visualization of a wide dynamic range of coverage values as a heatmap. Distance to TSS is shown in bp.

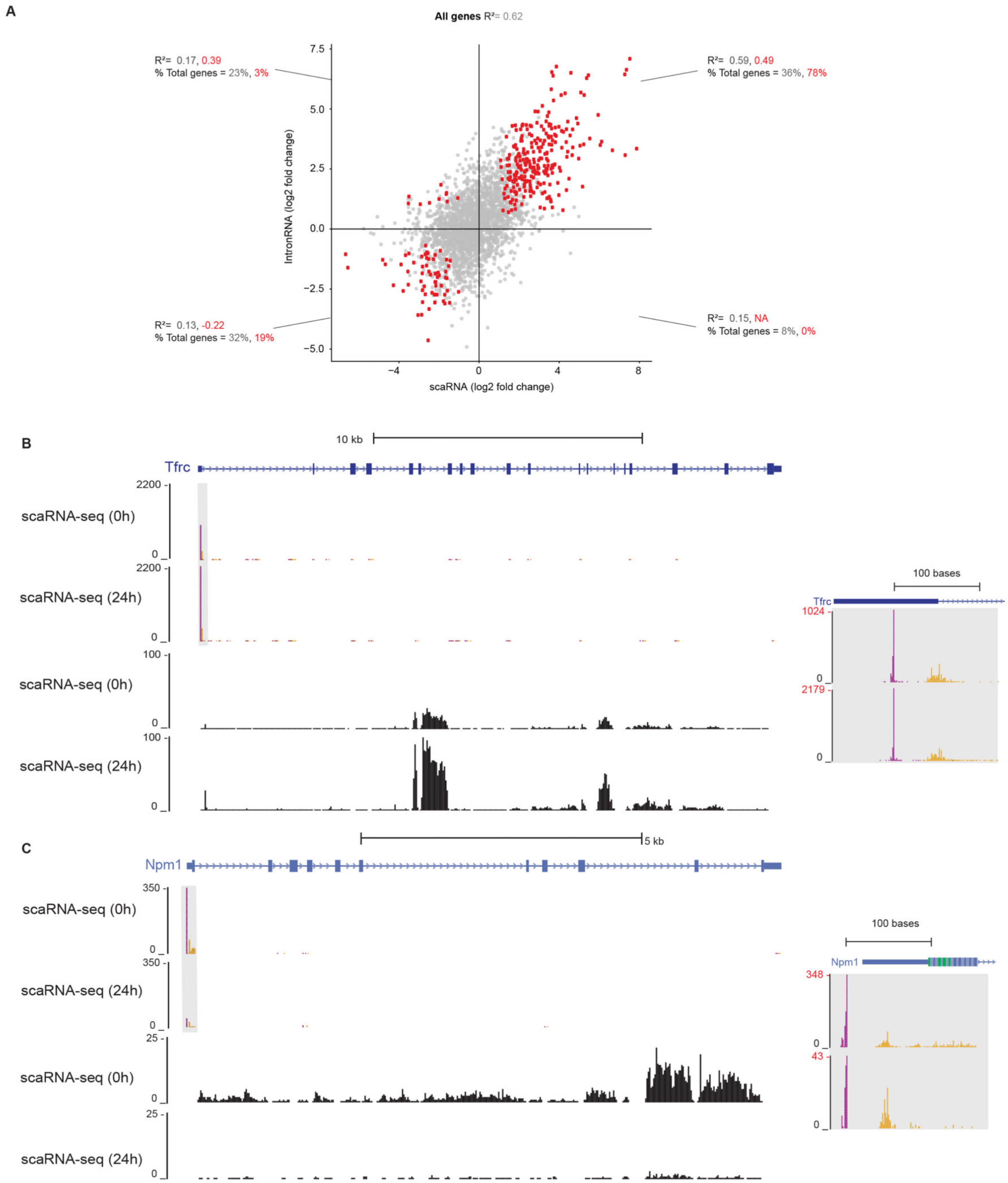

Figure 3. Transcription initiation is the predominant point of regulation in the transcription cycle.

(A) A scatter plot of log2 fold change in scaRNA molecules versus intronRNA for all detectably expressed genes (grey dots, n=2713). Genes which show significantly differential expression (red dots, n=327). The data show a positive correlation (R^2 of 0.62), with concordant changes in scaRNA and intronRNA indicating that regulation occurs predominantly via initiation. Each quadrant is marked with the correlation coefficient and percentages of genes (out of a total of 2713), all genes (grey) and significantly differentially expressed genes (red).

(B) *Tfrc* exemplifies increases in scaRNA across the region of promoter proximal transcription and intronRNA from the remainder of the gene, between 0h and 24h. This indicates an increase in initiation rather than a decrease in pausing (see Figure 1B). Set inset right for a zoomed view of scaRNA-seq data over promoter (grey box), with Y-axis scaled for visualization.

(C) *Npm1* exemplifies decreases in intronic RNA and scaRNA at 0h and 24h across the region of promoter proximal transcription. The level of scaRNA (5' and 3' end) transcripts decrease in parallel with intronic RNA at 24h indicating a decrease in initiation rather than increase in pausing which would be associated with an increase in scaRNA (see Figure 1B). Set inset right for a zoomed view of scaRNA-seq data over promoter (grey box), with Y-axis scaled for visualization.

All data were derived from biological triplicates at 0h and 24h of erythroid differentiation. scaRNA tracks were scaled to the sample with the lowest number of millions of mapped reads and a biological triplicate merged for visualization. IntronRNA tracks were normalized by reads per million and a biological triplicate merged for visualization. Differential count analyses were performed on raw unnormalized data using DEseq2 (Anders and Huber, 2010). See Figure S3 and 4.

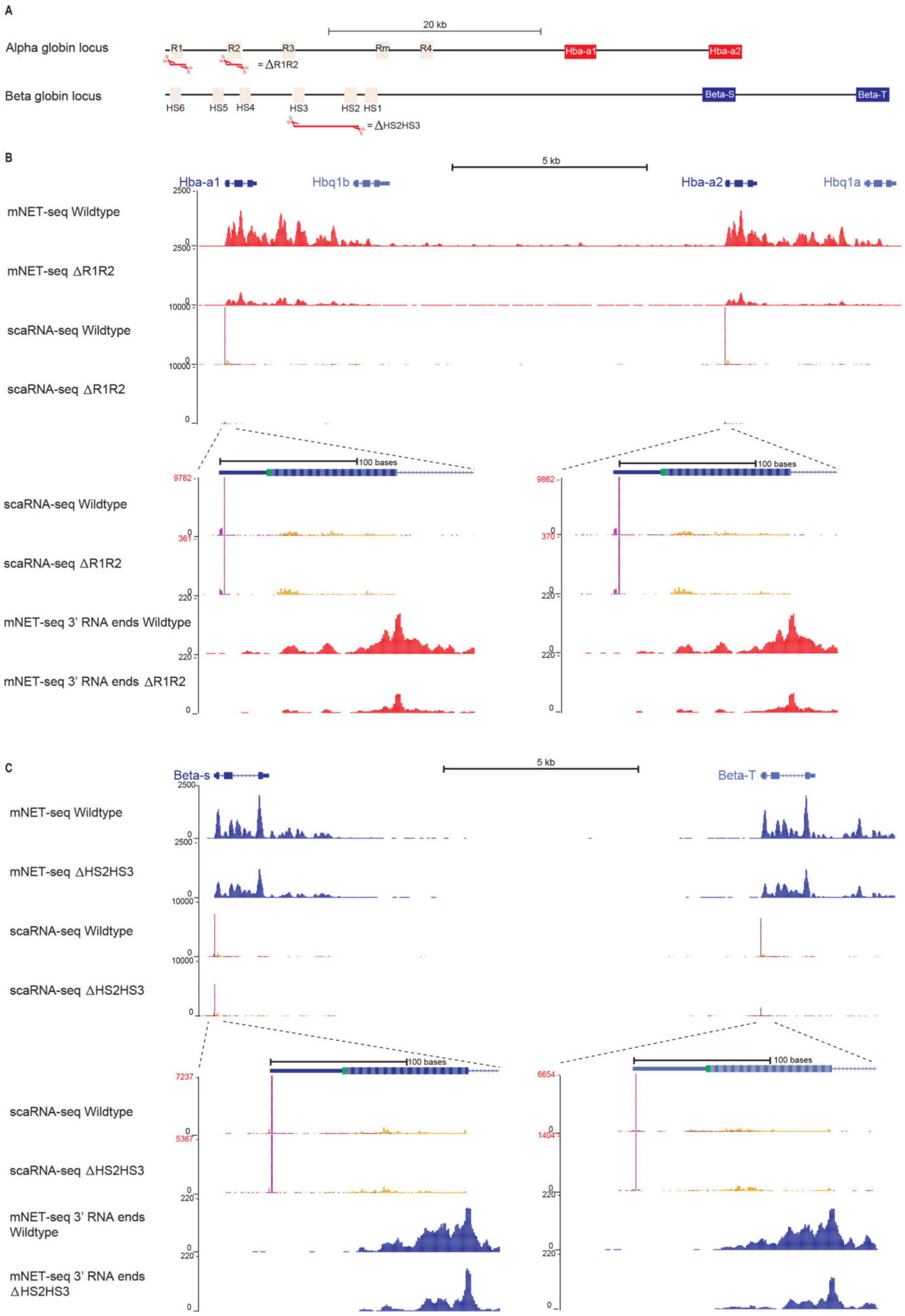

Figure 4. Enhancers regulate target gene expression via transcription initiation.

(A) The α- and β-globin loci represented to scale. Regulatory elements (enhancers) are indicated as orange boxes and the two copies of α-globin at the α-locus are highlighted in red. Two copies of β-globin at the β-globin locus are highlighted in blue. At the α-globin locus, deletion of the R1 and R2 enhancers (ΔR1R2) reduce nascent α-globin expression by 95% (Hay et al., 2016). At the β-globin locus, deletion of the HS2 and HS3 (ΔHS2HS3) reduce nascent β-globin expression by 70% (Bender et al., 2012).

(B) The α-globin locus in wildtype and ΔR1R2 cells, with the two α-globin genes (Hba-a1 and Hba-a2). scaRNA-seq and mNET-seq signals across the gene are both reduced showing that in ΔR1R2 there is a loss of transcription initiation. Two zoomed in views, highlight each gene’s promoter proximal region, with scaRNA-seq and mNET-seq. mNET-seq data represents the distribution of 3’ most base of each read (representing Pol II active site) with a 2nt window, to smooth data for visualization. scaRNA-seq data on zoomed view uses a scaled Y-axis (red values) to aid in visualization of the profile of Pol II pausing (3’ scaRNA ends).

(C) The β-globin locus in wildtype and ΔHS2HS3 cells, with the two β-globin genes (Beta-S and Beta-T). scaRNA-seq and mNET-seq signals across the gene are both reduced showing that in the ΔR1R2 model there is a loss of transcription initiation. Beta-T appears more affected than Beta-S. Two zoomed in views highlight each gene’s promoter proximal region. scaRNA-seq and mNET-seq. mNET-seq data represents the distribution 3’ most base of each read (representing Pol II active site) with a 2nt window, to smooth data for visualization. scaRNA-seq data on zoomed view uses a scaled Y-axis (red values) to aid in visualization of the profile of Pol II pausing (3’ scaRNA ends).

scaRNA-seq and mNET-seq tracks were scaled to the sample with the lowest number of millions of mapped reads and a biological triplicate merged for visualization. mNET-seq data were normalized by merging a replicate (n=2 for each genotype) and subsampling to the data set with the lowest total read count and visualization. See Figure S4 and 5.

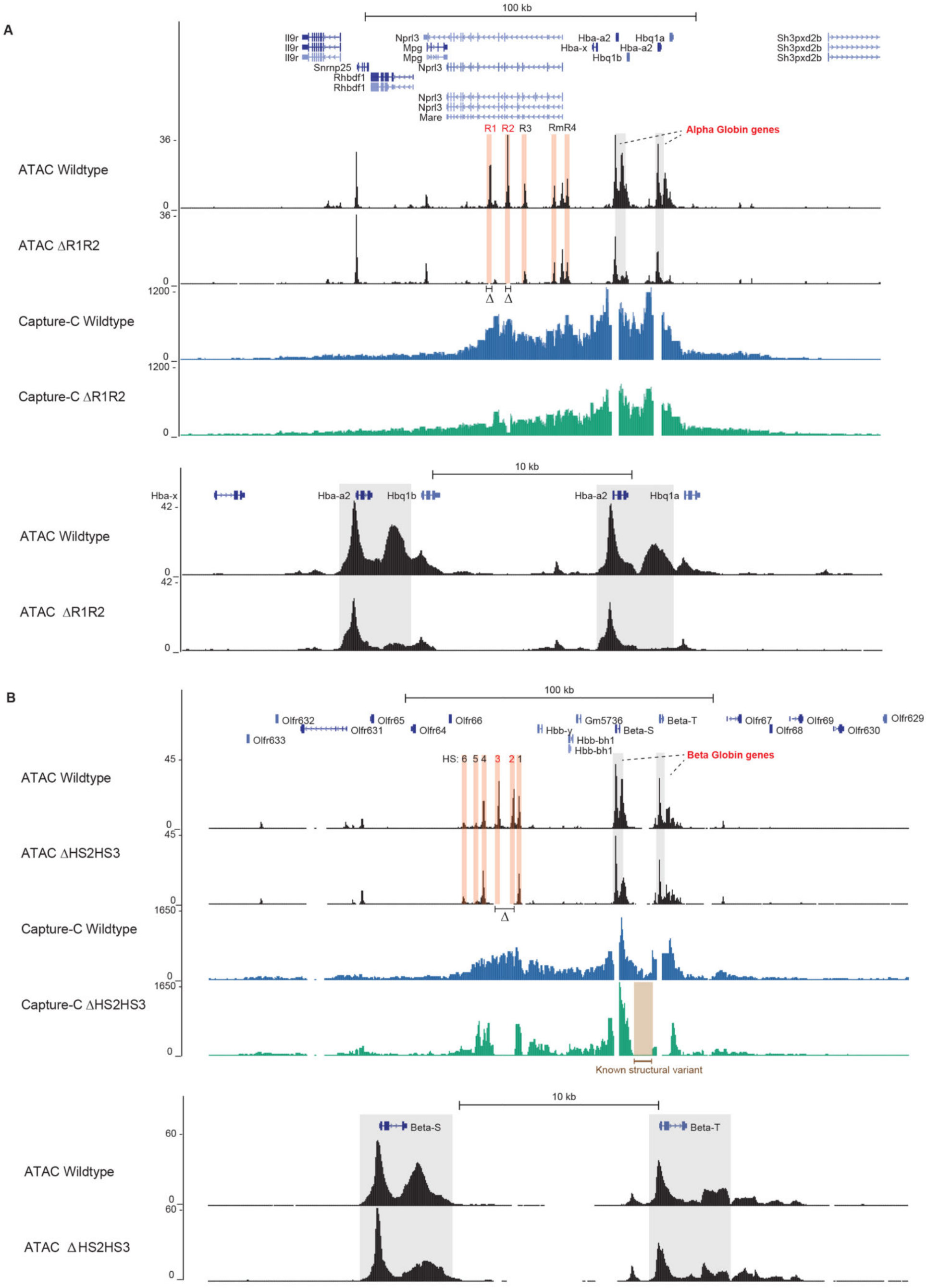

Figure 5. Promoter hypersensitivity and 3D chromatin structure are unperturbed by enhancer deletion.

(D) α-globin locus with two copies of α-globin (Hba-a1 and Hba-a2 – grey boxes) and regulatory elements (R1-R4 and Rm) highlighted as orange boxes. Deleted enhancers (R1 and R2) are shown in red text and deletions are visible in the ATAC-seq data as lack of coverage compared to wildtype. NG-Capture-C data show chromatin interactions from the viewpoint of the α-globin promoters (which appear as gaps in the Capture C track) interactions between the remaining enhancers and the genes persist in the ΔR1R2 deletion suggesting chromatin hub formation or maintenance is not affected by their deletion. Below, a zoomed in view of the α-genes shows that the promoters remain hypersensitive in the absence of the R1 and R2 enhancers. ATAC-seq signal in the 3’ UTR of the α-globin genes is also decreased in the ΔR1R2 mouse.

(E) β-globin locus with two copies of β-globin genes (Beta-S and Beta-T marked as grey boxes) and regulatory elements (R1-R6) highlighted as orange boxes. Deleted enhancers (HS2 and HS3) are shown in red text and deletions are visible in the ATAC-seq data as lack of coverage compared to wildtype. NG-Capture-C data show chromatin interactions from the viewpoint of the β-globin gene promoters (which appear as gaps in the Capture C track). Interactions between the enhancers and the genes persist in the ΔHS2HS3 deletion suggesting chromatin hub formation or maintenance is not affected by their deletion. A known structural polymorphism in the locus specific to this mouse strain is shown as a brown box and results in no NG-Capture-C coverage for the ΔHS2HS3 over this region but does not affect interpretation of the genes or enhancers. Zoomed view of the genes shown highlighting that promoters remain hypersensitive in the absence of R2 and R3 enhancers. All ATAC-seq data performed in triplicate were merged and normalized by reads per million. NG-Capture-C was also performed in triplicate and tracks show the mean interaction profile. NG-Capture C and ΔR1R2 ATAC-seq data for the α-globin locus are from published work (Hay et al., 2016). See Figure S5.

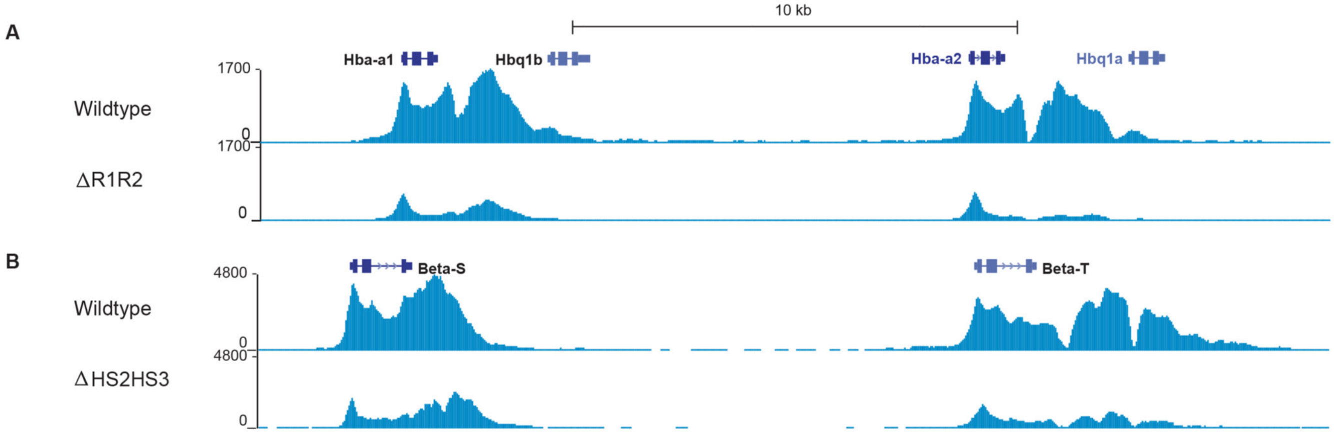

Figure 6. Enhancers stimulate Pol II recruitment to target genes

(F) Pol II ChIP-seq in wildtype and ΔR1R2 shows a total loss of Pol II across the α-globin genes (Hba-a1 and Hba-a2) consistent with defects in initiation not Pol II pause release.

(G) Pol II ChIP-seq in Wildtype and ΔHS2HS3 shows a total loss of Pol II across the β-globin genes (Beta-S and Beta-T) consistent with a defect in initiation not Pol II pause release

All ChIP-seq experiments were performed in biological triplicate and each replicate normalized by millions of mapped reads under Pol II peaks before being merged for visualization.