Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers

Jacob M. Tome, Nathaniel D. Tippens and John T. Lis

Eukaryotic RNA polymerase II (Pol II) has been found at both promoters and distal enhancers, suggesting additional functions beyond mRNA production. To understand this role, we sequenced nascent RNAs at single-molecule resolution to unravel the interplay between Pol II initiation, capping and pausing genome-wide. Our analyses identify two pause classes that are associated with different RNA capping profiles. More proximal pausing is associated with less complete capping, less elongation and a more enhancer-like complement of transcription factors than later pausing. Unexpectedly, transcription start sites (TSSs) are predominantly found in constellations composed of multiple divergent pairs. TSS clusters are intimately associated with precise arrays of nucleosomes and correspond with boundaries of transcription factor binding and chromatin modification at promoters and enhancers. TSS architecture is largely unchanged during the dramatic transcriptional changes induced by heat shock. Together, our results suggest that promoter- and enhancer-associated Pol II is a regulatory nexus for integrating information across TSS ensembles.
We have found that clusters of Pol II initiation are often more complex than divergent pairs. These clusters, which we call transcription initiation domains (TIDs), correspond to boundaries of evolutionary conservation, transcription factor binding and chromatin modification. Therefore, we propose that TIDs represent a prominent structural unit above divergent TSSs and facilitate a systems-level characterization of the immediate chromatin neighborhood around enhancers and promoters.

Results
Characterization of initiation, capping and pausing. We performed CoPRO with three separate RNA 5' state selection schemes to enrich for capped only, uncapped only, or capped and uncapped transcripts in biological duplicate in K562 cells. Additionally, CoPRO with selection for capped only was performed in MEFs before and after 60 min of heat shock at 42 °C. All figures show K562 data unless otherwise noted.

CoPRO provides a transcription profile for each nucleotide in the genome. Polymerases that initiated from the same transcription start nucleotide (TSN) can be seen at different positions along a vertical line in a heatmap of CoPRO for a single locus, while Pol II initiating from different TSNs, but pausing at the same nucleotide, creates a horizontal line (Fig. 1b). This characteristic pause and elongation pattern (a shared RNA 5' end with numerous 3' ends) provides a sensitive method to discriminate real initiation sites from background. As an example, the IPO7 promoter initiates at multiple positions within a ~50-nt TSS, with two preferred TSNs (see Fig. 1b and Supplementary Fig. 1 for additional examples). Notably, initiation
from nearby TSNs tends to pause at the same nucleotide (horizontal lines). A wider view (Fig. 1b, inset) shows elongation beyond the pause (within the insert size limit of Illumina sequencing\(^3\)), as well as antisense transcription. RefSeq annotations for start sites tend to be 5’ of the most used TSN detected by CoPRO. This is likely because annotations take the most 5’ site for which support was found to avoid mapping degradation products or are more representative of initiation sites used in a different cell type.

We developed a hierarchy to facilitate subsequent analyses (Fig. 1c). First, individual nucleotides where initiation occurs (TSNs) are identified by having multiple, distinct 3’ ends, among other criteria (Methods and Supplementary Fig. 2). We cluster nearby TSNs into TSSs, as they are probably driven by the same transcription factors and PIC. TSSs are further clustered into higher-order structures: divergent pairs are TSSs transcribing in opposite directions <300bp apart\(^9\),\(^{34}\) and TIDs are larger TSS clusters (<600-bp gaps) on either strand. We use available CAGE data to determine which TSNs and TSSs give rise to stable RNAs, such as mRNAs and long noncoding RNAs (lncRNAs)\(^1\). Most TSSs produce unstable RNAs of unknown function (for example, enhancer RNA, noncoding RNA, upstream antisense RNA). In total, CoPRO identified 458,549 TSNs, 78,467 TSSs, 18,411 divergent pairs and 27,705 TIDs in K562 cells.

As previously described\(^5\),\(^9\),\(^34\),\(^36\), most TSSs initiate at multiple TSNs and TSN number increases with TSS activity (Fig. 1d), although the number and distribution of TSNs varies widely (Supplementary Fig. 3). TSSs where initiation uses few TSNs, despite very high activity, tend to have exceptionally strong core promoter elements\(^35\), such as ACTB and HIST1H1 (Supplementary Fig. 1). While the basic unit of initiation is a divergent pair, each member of which has its own PIC\(^9\),\(^34\),\(^37\),\(^38\),\(^39\), most TSSs, especially at promoters, are found in larger clusters (Fig. 1c). Previous work has noted specific types of TSS clusters, such as divergent pairs of TSSs\(^3\),\(^4\), nearby alternative TSSs and head-to-head genes\(^34\), or convergent TSSs\(^5\). CoPRO, however, shows that large constellations are much more prominent when all unstable TSSs are considered: promoter TIDs, on average, contain 1.4 stable TSSs paired with 4.7 unstable TSSs.

With CoPRO, we are able to determine exactly how far a Pol II molecule has transcribed after initiation with every read pair and thus calculate the ‘distance transcribed’, defined as the exact
Pause site specification by sequence and local activity. A histogram of the most used pause position per TSN (maxPause) shows a bimodal distribution: one population from 20–32 nt (early) and another from 32–60 nt (late; Fig. 2a), although many TSNs have some pausing in both windows. After grouping TSNs by distance transcribed at maxPause, the sequence determinants of pausing become clear (Fig. 2b). Relative AT richness is seen at ~30 nt to the initiation site, particularly in the early-pausing TSNs (Fig. 2b), reflecting a stronger enrichment for the TATAWR motif (Supplementary Fig. 4). In *Drosophila*, Pol II that pauses proximal to initiation has been shown to remain in contact with the PIC at promoters with strong TATA boxes, both by perturbing this interaction with short DNA insertions between initiation and pause sites and by detecting co-occurrence of PIC components at the initiation site and Pol II at the pause site on the same DNA molecule. The association of early pausing with better TATAWR motifs and slightly higher TATA box–binding protein ChIP-seq signal (Supplementary Fig. 5) suggests that early-pausing Pol II may remain bound by the PIC in humans as well. Whereas early-pause TSNs are more AT-rich overall, late-pause TSNs are more GC-rich at +5 nt from the initiation site to 8 nt before the pause site. TSN initiation occurs preferentially on CA dinucleotides (the Inr motif) and pausing takes place during the incorporation of a cytosine residue (Fig. 2b, left). A capped CoPRO experiment in MEFs showed a remarkably similar pattern of GC content around initiation and pause sites, indicating that these mechanisms of pause site choice are conserved (Fig. 2b, right).

Pausing is thought to occur on recruitment of several pause factors such as negative elongation factor (NELF) and DSIF. Thus, we propose that this recruitment is most likely to occur during the slow step of cytosine incorporation, thus accentuating the tendency to observe Pol II there. These results refine previous kinetic models for pausing: the position of the pause is dictated by the energy landscape of PIC interactions, the transcriptional bubble and RNA–DNA hybrid, local availability of pause factors and nucleotide incorporation. A mixture of these features probably determines the position at which pause factors ultimately associate with Pol II. As predicted by this model, Pol II initiating at nearby TSNs tends to pause at the same position, and not simply a fixed distance after initiation (Fig. 2d). Individual examples, such as the exceptionally dispersed MAPK1 promoter (Fig. 2c), show that nearby TSNs share even minor pause sites.

To explore features of early and late pausing, we classified TSNs as predominantly early or late (first and last quartile by fraction of pausing in the early window; see Methods). Early-pause loci are enriched for enhancer- and lineage-specific transcription factors (GATA1, GATA2 and TAL1) and less active chromatin marks.
(Supplementary Fig. 5), are more likely to occur at TSSs that do not produce a stable RNA product, and are associated with lower overall expression and elongation (Supplementary Fig. 6). In contrast, late-pause loci are more enriched for activating transcription factors (ATF3 and ELF1), active chromatin marks (Supplementary Fig. 5), and high expression and pause release (Supplementary Fig. 6). Confirming the role of active features in the late pause class, we found a significant shift from late to early pausing on the widespread gene repression observed after heat shock in MEFs45 (Supplementary Fig. 6). Therefore, TSNs with late pausing are endowed with GC content, transcription factors and chromatin environments that facilitate productive elongation.

**Capping dynamics vary with pause location.** Early- and late-pause TSNs have different uncapped RNA distributions, indicating interplay between the processes of capping and pausing. TSNs of both classes have very short uncapped RNA (less than 22 nt), but late-pausing TSNs also have uncapped RNAs at the pause site (Fig. 3a, right) that are absent at early-pausing TSNs. In total, early-pause TSNs have much higher levels of uncapped RNA than late ones (Fig. 3b), but these uncapped transcripts are only observed at very short distances to the most active bin on the same strand. Convergent TSSs are the strongest antisense bin from 0 to –300 bp downstream of the maxTSS. The upstream divergent (sense) TSS is the strongest bin on the opposite strand from 0 to 1,000 bp downstream of the maxTSS.

**Fig. 4 | A global view of initiation shows rules for divergent pairing and widespread complex organization.** a, Model TID. Multiple TSSs by default create different relationships: divergent and convergent pairings are enumerated. A register is a series of uniformly phased nucleosomes downstream of a TSS, annotated here by the TSS that explains registration. b, CoPRO capped and MNase-seq in promoter TIDs, centered on the most active bin per TID (sense) and sorted by the distance to the most active bin on the opposite strand (antisense), N = 9,273 TIDs with >2 TSSs and >6 reads per strand (1,892 convergent maxTSSs, 5,749 divergent maxTSSs and 1,632 extended maxTSSs). c, Location of the strongest convergent TSS and its paired upstream divergent TSS (sense) relative to the maxTSS for all TIDs in b. Convergent TSSs are the strongest antisense bin from 0 to 700 bp downstream of the maxTSS. The upstream divergent (sense) TSS is the strongest bin on the same strand from 0 to –300 bp to the convergent TSS. d, Alternative registers shown in aggregate by re-sorting convergent TIDs by the location of different TSSs. MNase-seq metaplots are centered on the TSS used to assess its ability to explain nucleosome positioning. Top, sorted by convergent TSS (repeated from b). Middle, sorted by upstream divergent TSS (–300 to 0 bp, opposite strand). Bottom, sorted by strongest TSS from 100 to 1,000 bp downstream of the maxTSS (on the same strand). e, Epigenomics Roadmap histone modification data, sorted and oriented as in b.
Articles NATURE GENETICS

Heat shock specific

Both

Fig. 5 | Massive regulatory changes in heat shock occur by modulating the activity of pre-established TSSs. a, Fraction of TSSs identified by CoPRO capped before (control) and after 60 min of heat shock in MEFs. Despite significant changes in expression of 65% of genes, 96% of TSSs are observed under both conditions. b, Fraction of TIDs before and after heat shock. 98% of TIDs are detected under both conditions.

Initiation clusters are more complex than divergent pairs. Simple divergent pairs are not sufficient to characterize TID architecture: 71% of TSSs are within a TID containing >2 TSSs, and 80% of TIDs containing at least one stable TSS (promoter TIDs) have >2 TSSs. The largest TIDs are strong outliers, both by length and number of TSSs that they contain (Supplementary Fig. 7). TSS number increases with transcriptional activity (Fig. 1d) to some extent, indicating that more transcriptionally active loci use or create more TSSs.

While bidirectional pairing of mammalian TSSs has received considerable attention, larger TSS clusters, or TIDs, remain less well characterized (Fig. 4a). To illustrate, we identified the bin with the most CoPRO capped initiation signal on each strand of promoter TIDs and sorted them by distance between this pair of sites (Fig. 4b). The pattern is nearly identical in non-promoter TIDs (Supplementary Fig. 8a), indicating that this basic architecture is universal despite ‘non-promoters’ (for example, enhancers) having lower overall activity. Upstream antisense TSSs separated by >300 bp exhibit nucleosome occupancy and additional TSSs, with an additional sense peak flanking the antisense peak (Fig. 4b, extended maxTSSs), and are therefore best described as two divergent pairs, as shown previously with CAGE and Start-seq.

Arrangements of TSSs and nucleosomes are tightly coupled. Convergent TSSs are coordinately positioned between nucleosomes (Fig. 4b, antisense CoPRO capped between nucleosomes). The downstream antisense (convergent) bin with the greatest signal downstream of the most occupied TSS is preferentially located between nucleosomes phased to the strongest TSS (Fig. 4c, blue). This indicates that nucleosome phasing from the strongest TSS probably constrains the location of convergent TSSs. However, the downstream sense TSS pair with the convergent TSS (that is, TSS 4 if TSS 2 is the most active in Fig. 4a) appears less constrained (Fig. 4c, red versus blue), as its location is less peaked between nucleosomes phased to the strongest TSS. Interestingly, downstream sense TSSs lack antisense CoPRO signal until ~200 bp downstream of the maxTSS (Supplementary Fig. 8e). This indicates that the directionality of some minor TSSs could be enforced by the maxTSS + 1 nucleosome blocking antisense transcription. Within TIDs, we find increased conservation scores around most TSSs and less conservation of intervening sequences (Supplementary Fig. 8).

Nucleosome positioning follows two rules with respect to clusters of TSSs in TIDs: (1) nucleosomes are phased downstream of TSSs and (2) stronger TSSs dominate in phasing, thus constraining weaker TSSs. Therefore, the interplay between nucleosomes and TSSs forms intricate patterns within TIDs: registers of uniformly phased nucleosome are offset by the variable spacing between divergent pairs. Close examination of nucleosome positions shows alignment with the pause site rather than the initiation site (Supplementary Fig. 9), confirming an interplay between paused Pol II and the first nucleosome.

Sorting heatmaps of micrococcal nuclease sequencing (MNase-seq) relative to different TSSs shows their power in explaining registers (see Supplementary Figs. 8b–g and 10 for a summary). Two competing registers overlap in the space between convergent maxTSSs in TIDs where the most active antisense TSS is convergent to the most active sense TSS, providing an interesting subset for mechanistic understanding. Nucleosome occupancy between convergent maxTSSs is lower than that flanking divergent maxTSSs (Fig. 4b), confirming that nucleosome occupancy is reduced between convergently oriented active TSSs. Nucleosomes between convergent maxTSSs appear well-positioned to the sense maxTSS, but outer flanking nucleosomes do not, meaning that they are in registers not explained by these TSSs (Fig. 4d, top). Sorting these same data by the location of other TSSs shows registers in these regions that were randomly distributed in our first sort. Sorting on the upstream divergent TSS shows a register upstream (Fig. 4d, middle versus top). Likewise, the strongest downstream sense TSS is associated with a nucleosome register downstream (Fig. 4d, bottom), thus explaining more phasing in the downstream direction than the maxTSS alone. In summary, convergent maxTSSs have three prominent registers that are uncovered by aligning to the maxTSS, upstream divergent TSS and downstream sense TSS. Because most TIDs contain multiple divergent pairs, when all promoter TIDs are sorted similarly these patterns are observed genome-wide (Supplementary Fig. 8c). Thus, intricate registers of phased nucleosomes are common, but are hidden when they are randomly distributed throughout any single heatmap because we are only able to sort on one TSS pair at a time (Supplementary Fig. 10).

Chromatin environment and transcription are inexorably linked, and the comprehensive map of initiation afforded by CoPRO provides a framework to better understand their interplay. Histone modifications are distributed with distinct patterns relative to TSS pairs: H3K27ac is enriched after strong TSSs and between strong convergent pairs, H3K4me3 has a similar distribution with more spreading downstream of the maxTSS, and H3K4me1 is depleted near TSSs and enriched upstream (Fig. 4e).

TIDs are clusters of regulatory elements. Although dramatic changes in gene transcription occur after 1 h of heat shock, with 65% of all active genes showing significant up- or downregulation by PRO-seq, CoPRO in MEFs shows that this primarily occurs by adjusting activity at existing TSSs (Fig. 5a) and TIDs (Fig. 5b), and not by creation of new TSSs or TIDs. This supports our previous findings that pause release is targeted for regulation during heat shock.

Pol II is found at TID boundaries (oriented outward) and throughout their centers (Fig. 6, PRO-seq). In broad TIDs, the maxTSS tends to be centrally located (Supplementary Fig. 11). TIDs are DNase hypersensitive and conserved (PhastCons, Supplementary Fig. 12), suggesting functionality throughout these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSs and weaker occupancy inside these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSs and weaker occupancy inside these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSs and weaker occupancy inside these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSs and weaker occupancy inside these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSs and weaker occupancy inside these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSs and weaker occupancy inside these units.
nucleosomes, while the second nucleosome and beyond are less likely to be promoters (Fig. 6, top versus bottom), are lowly expressed (Fig. 1d) and span a single NFR. This strong exclusion from TIDs is more similar to the pattern of repressive marks like H3K9me3 and H3K27me3 (Supplementary Fig. 13) than to other active marks.

**TIDs demarcate chromatin domains at initiation sites.** Histone modifications and variants also show diverse patterning relative to TIDs. Repressive marks such as H3K9me3 are depleted within TIDs, and marks of elongation such as H3K36me3 are depleted within TIDs and enriched downstream of the maxTSS (Supplementary Fig. 13). Active marks are enriched within TIDs with variable spreading beyond their boundaries. The active histone variant H2A.Z (Fig. 6) shows enrichment at nucleosomes flanking narrow promoter TIDs, reflecting its enrichment flanking the maxTSS and its upstream divergent TSS (Fig. 4e). The active histone marks H3K27ac and H3K4me3 are enriched inside TIDs and just outside, with H3K4me3 showing greater enrichment downstream of the maxTSS. Nucleosome-resolution native MNase ChIP-seq strongly suggests that H3K27ac is primarily confined to nucleosomes immediately adjacent to TSSs (Supplementary Fig. 14). While nucleosome-resolution data are not available for H3K4me3 in K562 cells, two ChIP datasets show it extending past H3K27ac (Fig. 6 and Supplementary Fig. 14), and data in other cell types show that H3K4me3 peaks tend to be wider than H3K27ac peaks. Thus, it is tempting to speculate that both marks are usually deposited across TIDs and at the first nucleosome outside, while the second nucleosome and beyond are more likely to have H3K4me3 and not H3K27ac. Histone modifiers such as the H3K4me3 demethylase KDM1A (or LSD1) and the H3K9me3 demethylase PHF8 are positioned commensurate with their targeting and activity (Supplementary Data).

**Discussion**

We developed CoPRO to provide coordinated information about Pol II initiation and active site location, and cap status for single nascent transcript molecules genome-wide. We found evidence for two pause classes along a continuum of behaviors—early and late pausing. Generally, early pausing is associated with more enhancer-like loci, where entry into elongation and production of a stable RNA product may not be necessary for regulatory function. Importantly, we found that late-pausing sites switch to an early pause during the widespread repression of many genes that occurs during heat shock. Therefore, they probably represent different regulatory steps that Pol II must pass through to assemble a full elongation complex. Early pausing is linked more with initiation than with the transition to elongation, as it occurs more proximal to the initiation site and is less likely to lead to productive elongation. In contrast, late pausing may be more closely linked with elongation, as it occurs more distal to initiation and is associated with higher levels of elongation. In addition, these two pause classes have different 5′ capping profiles. Early pausing is associated with high levels of uncapped transcripts and a smooth transition of uncapped to capped transcripts (Fig. 3b). Late pausing is associated with lower levels of uncapped transcripts.
transcripts overall, but uncapped transcripts are found at the same locations as capped transcripts (Fig. 3b). This difference may be due to the fact that the capping enzyme is recruited by DSIF\(^\text{9,46}\), which itself is recruited to Pol II partly through interaction with the nascent RNA\(^\text{47}\). Because early-paused transcripts would only have 4–18 nt of RNA extending outside Pol II\(^\text{44}\), they may recruit DSIF and capping enzyme less efficiently than late-pause transcripts, which complete capping with a longer transcript. This shows that capping and pausing are tightly coupled, rather than simply sequential steps in transcription.

Mammalian genomes are extremely complex: the ENCODE Consortium identified almost 400,000 putative enhancers and over 70,000 promoters based on chromatin features across 147 cell types\(^\text{48}\). Thus, genome-wide functional data are critical for determining which of these elements are active in any one cell type and for characterizing responses to experimental perturbation. Analysis of such datasets is heavily dependent on the framework used: in the 3.2-billion-base-pair human genome, it is necessary to know where to look to interpret any new dataset. A priori knowledge of regulatory elements is often used as such a framework. This means relying on gene annotations for promoters and ChIP-seq peaks for chromatin modifications for enhancers. However, promoter annotations are often a poor match for precise initiation sites in an individual cell line\(^\text{48}\), and even the best ChIP-seq peak-calling algorithms identify binding sites with limited resolution\(^\text{49}\). To illustrate the impact of framework on interpretation, consider that Fig. 4c and the promoter portion of Fig. 6 for the same histone modifications show the same data at the same sites, but oriented and sorted differently to identify different aggregate patterns and thus different functional relationships. In this study, we characterized transcription initiation with minimal assumptions about the properties of underlying elements. When all TSSs are considered, regulatory elements are composed of clusters of TSSs that are much larger than previously appreciated. These clusters, or TIDs, are intimately coupled with important structures at many functional levels, including the patterns of transcription factor binding, precise arrays of nucleosomes and the extent of histone modifications. Thus, TIDs provide a systems-level perspective of transcribed regulatory domains and a novel framework for understanding genome regulatory architecture and function.

The ratio of H3K4me3 to H3K4me1 has been proposed to distinguish promoters from enhancers\(^\text{50}\). However, we have previously suggested that this ratio is correlated with the transcription level and so is not a basis for distinguishing a promoter from an enhancer. Here we find that patterns of chromatin modification are intimately coupled with TID architecture, in addition to transcription level. TSSs and nearby chromatin modifications are probably engaged in a feedback loop: transcription facilitates chromatin modification, which in turn enables subsequent rounds of transcription, as demonstrated previously for TFIIID recruitment and H3K4me3\(^\text{51,52}\). Thus, future studies are needed to examine the extent to which minor TSSs within TIDs shape chromatin or are opportunistic ‘passengers’ within a permissive environment driven by more active TSSs within the TID.

Nucleosomes at TSSs are highly dynamic\(^\text{53}\): thus, TIDs have the ostensibly paradoxical properties of low nucleosome occupancy and abundant active histone modifications. TID internal nucleosomes are seemingly too labile for detection by MNase-seq, yet are detectable by ChIP for their modifications\(^\text{54}\). Recently, it was shown that transcription increased the mobility of promoter and enhancer loci in nuclei\(^\text{55}\); similarly, Pol II–driven ‘molecular stirring’ within TIDs may help to account for their volatile chromatin state. TSS clusters delineate active compartments of enhancers and promoters across a wide spectrum of sizes, where abundant low-complexity protein domains\(^\text{56}\) (histone tails, TFIIID, Pol II CTD and cyclin T1\(^\text{57}\)) with a high potential for trans-activating interactions could establish phase-separated networks\(^\text{58}\). Although individual TSSs are driven by small units (150–300 bp) with limited information content\(^\text{59}\), TIDs provide a molecular framework for information integration within larger domains. Consistent with this idea, broad H3K4me3 domains tend to be located at the promoters of genes critical for cell identity and are associated with less cell-to-cell variability in mRNA levels\(^\text{60,61}\). This means that the clusters of initiation sites within TIDs may result in a greater level of regulatory control than could be encoded within smaller units such as individual promoters or divergent pairs.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0234-5.

Received: 13 April 2018; Accepted: 14 August 2018; Published online: 22 October 2018

**References**

1. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).

2. Seila, A. C. et al. Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).

3. Core, L. J. et al. Analysis of transcription start sites from nascent RNA supports a unified architecture of mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).

4. Lavender, C. A. et al. Downstream antisense transcription predicts genomic features that define the specific chromatin environment at mammalian promoters. *PloS Genet.* **12**, e1006224 (2016).

5. Henriques, T. et al. Widespread transcriptional pausing and elongation at enhancers. *Genes Dev.* **32**, 26–41 (2018).

6. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).

7. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).

8. Addman, K. & Liu, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).

9. Gilchrist, D. A. et al. Pausing of RNA polymerase II disrupts DNA-specific nucleosome organization to enable precise gene regulation. *Cell* **143**, 540–551 (2010).

10. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell* **53**, 819–830 (2014).

11. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).

12. Murakami, K. et al. Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* **342**, 1238724 (2013).

13. Sainsbury, S., Berneyck, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 129–143 (2015).

14. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).

15. Rasmussen, E. B. & Lis, J. T. In vivo transcriptional pausing and cap formation on three *Drosophila* heat shock genes. *Proc. Natl. Acad. Sci. USA* **90**, 7923–7927 (1993).

16. Mandle, S. S. et al. Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proc. Natl. Acad. Sci. USA* **101**, 7572–7577 (2004).

17. Leander, M., Leikis, K. & Cramer, P. Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. *Mol. Cell Biol.* **33**, 3805–3816 (2013).

18. Nilson, K. A. et al. THZ1 reveals roles for Cdk7 in co-transcriptional capping and pausing. *Mol. Cell* **59**, 576–587 (2015).

19. Mostek, S. & Price, D. Functional coupling of capping and transcription of mRNA. *Mol. Cell* **10**, 599–609 (2002).

20. Ramanathan, A., Robb, G. B. & Chan, S. mRNA capping: biological functions and applications. *Nucleic Acids Res.* **44**, 7511–7526 (2016).
Articles

Vihervaara, A. et al. Transcriptional response to stress is pre-wired by Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev.* **28**, 14–19 (2014).

Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat. Genet.* **49**, 1045–1051 (2017).

Krebs, A. R. et al. Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Mol. Cell.* **67**, 411–422 (2017).

Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3**, e02407 (2014).

Schwalb, B. et al. TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).

Proudfoot, N. J. Transcriptional termination in mammals: stopping the RNA polymerase II juggernaut. *Science* **352**, 715–718 (2016).

Nojima, T. et al. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).

Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541–554 (2015).

Nechaev, S. et al. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335–338 (2010).

Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

Fan, H. C., Blumenfeld, Y. J., Chikitaru, U., Hudgins, L. & Quake, S. R. Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clin. Chem.* **56**, 1279–1286 (2010).

Chen, Y. et al. Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.* **48**, 984–994 (2016).

Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).

Vo Ngoc, L., Wang, Y.-L., Kassavetis, G. A. & Kadonaga, J. T. The punctilious transcription-coupled to RNA processing. *Mol. Cell* **63**, 526–540 (2015).

van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).

Duttle, S. H. C. et al. Human promoters are intrinsically directional. *Mol. Cell* **57**, 674–684 (2015).

Gressel, S. et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eElife* **6**, 1–24 (2017).

Traut, T. W. Physiological concentrations of purines and pyrimidines. *Mol. Cell. Biol.* **140**, 1–22 (1994).

Ramani, V., Qiu, R. & Shendure, J. High-throughput determination of RNA structure by proximity ligation. *Nat. Biotechnol.* **33**, 980–984 (2015).

Song, Y., Liu, K. I. & Wang, T. H. Elimination of ligation dependent artifacts in 7T RNA ligase to achieve high efficiency and low bias microRNA capture. *PLoS ONE* **9**, e94619 (2014).

Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell. Biol.* **19**, 464–478 (2018).

Li, J. et al. Kinetic competition between elongation rate and binding of NELF controls promoter-proximal pausing. *Mol. Cell* **50**, 711–722 (2013).

Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. & Lis, J. T. Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. *Mol. Cell* **62**, 63–78 (2016).

Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).

Gilchrist, D. A. et al. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter proximal nucleosome assembly. *Genes Dev.* **22**, 1119–1133 (2008).

Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicke, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 2926–2931 (2010).

Vihervaara, A. et al. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nat. Commun.* **8**, 1–15 (2017).

de Dieuleveult, M. et al. Genome-wide nucleosome specificity and function of chromatin remodelers in ES cells. *Nature* **530**, 113–116 (2016).

51. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

52. Beagan, J. A. et al. YY1 and CTGF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* **27**, 1139–1152 (2017).

53. Weintraub, A. S. et al. YY1 is a structural regulator of enhancer-promoter loops. *Cell* **171**, 1573–1588 (2017).

54. Pradeepa, M. M. et al. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* **48**, 681–686 (2016).

55. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).

56. Engreitz, J. M. et al. Local regulation of gene expression by IncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).

57. Misra, A. & Gilmour, D. S. Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the *Drosophila* RNA polymerase II transcription elongation complex. *Proc. Natl. Acad. Sci. USA* **107**, 11301–11306 (2010).

58. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* **529**, 551–554 (2016).

59. Zhang, Y. et al. Model-based analysis of ChiP-seq (MACS). *Genome Biol.* **9**, R137 (2008).

60. Laubert, S. M. et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell* **152**, 1021–1036 (2013).

61. Vermeulen, M. et al. Selective anchoring of TFIIID to nucleosomes by 3′-triphosphorylation of histone H3 lysine 4. *Cell* **131**, 58–69 (2007).

62. Shivashbamy, S. et al. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).

63. Gu, B. et al. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**, 1050–1055 (2018).

64. Hughes, M. P. et al. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science* **359**, 698–701 (2018).

65. Lu, H. et al. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318–323 (2018).

66. Hinisz, D., Shrivivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).

67. Benyoun, B. A. et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673–688 (2014).

68. Chen, K. et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat. Genet.* **47**, 1149–1157 (2015).

Acknowledgements

We thank members of the Lis, Kwak and Danko laboratories for helpful discussions throughout the experimental design and analysis phases. J. Mahat assisted in designing CoPRO libraries for sequencing without PCR amplification and the heat shock protocol. The Cornell Sequencing core and especially P. Schweitzer were extremely helpful in designing libraries and eminently patient in accommodating our technical requests. We acknowledge funding from National Institutes of Health (NIH) grant nos. HG009393 and GM025232 (to J.T.L.). N.D.T. was supported by NIH training grant no. T32HD057854.

Author contributions

J.M.T. and N.D.T. conceived of CoPRO, carried out experiments and analyzed data. N.D.T. developed the computational framework for analysis. J.M.T., N.D.T. and J.T.L. interpreted results and prepared the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0234-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.T.L.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2018
Methods

Experimental design. CoPRO adapts PRO-cap\(^\text{15}\) for paired-end sequencing and includes a total of three different libraries that were enriched for capped nascent RNAs only, uncapped nascent RNAs only or both. With paired-end sequencing, each read tells us where an RNA polymerase molecule initiated and then where its active site is located. Comparison of the libraries for different capping states allowed us to identify the precise location of pausing and where nascent RNAs became capped across the tens of thousands of initiation sites that we detected. The paired nature of the data enabled identification of sites of transcription initiation with unprecedented precision: we were able to use the pattern of pausing and elongation as a sensitive way of calling real initiation sites and of filtering out termination by comparing capped and uncapped treatments. Because CoPRO maps nascent RNA, it is not affected by the post-transcriptional stability of the initiation and pausing events detected. Thus, it puts noncoding RNAs (such as enhancer RNAs and upstream divergent RNAs) on an equal footing with longer-lived RNAs (such as tRNAs and IncRNAs). Using our comprehensive maps of transcription initiation, we were able to compare the architecture of initiation sites with other features of the genome. For this, we chose human K562 cells for direct comparison with dozens of publicly available genome-wide datasets. See the Life Sciences Reporting Summary for a description of biological reagents used and replication.

Cell culture and CoPRO library preparation. K562 cells were obtained from ATCC and cultured antibiotic-free in accordance with their standards in DMEM, high glucose + HEPES (Thermo Fisher, 12430054) supplemented with 10\% FBS (Thermo Fisher, 10437028). Cultures were verified to be mycoplasma-free\(^\text{21}\) prior to library preparation and sequencing. Two biological replicates were cultured independently, generated by two passages, with library preparations done separately for technical and biological replicates. Cells were permeabilized and run-on reactions with all four biotin NTPs were carried out with 20 million cells per reaction, as described previously\(^\text{21}\). After isolating RNA from the run-on with TRIzol (Thermo Fisher, 10296028), three run-ons per biological replicate were pooled. Two adaptor ligations and reverse transcription were performed as described\(^\text{21}\), with custom adaptors detailed in Supplementary Table 1. Critically, no RNA fragmentation was done so that pairing of RNA 5′ ends remained biologically meaningful. The first ligation added a sample barcode to the library; TruSeq barcodes were chosen to minimize the predicted secondary structure of the adaptor\(^\text{21}\). Between adapter ligations and reverse transcription reactions were performed. The three cap state selections used a series of enzymatic treatments to reduce specific populations of RNAs to 5′ monophosphate, making them capable of ligation to an RNA adaptor by T4 RNA ligase (NEB, M0204S). All steps were performed using the manufacturer’s protocol, with phenol:chloroform extraction and ethanol precipitation between steps. We designed three separate 5′ state selections (Supplementary Fig. 15c).

1. Uncapped RNAs were selected by treating with calf intestinal alkaline phosphatase (CIP) (NEB, M0290S) and T4 polynucleotide kinase (PNK) 3′ phosphatase minus (NEB, M0236S), in order to reduce all uncapped RNAs to 5′ hydroxyl and then add 5′ monophosphate, but without removing the cap from capped RNA.

2. Capped RNAs were selected by treating with terminator 5′ phosphate–dependent exonuclease (Ep Berlin, TER21020) to degrade 5′ monophosphate RNA (from terminating polymerase) and CIP to reduce other uncapped RNAs to 5′ hydroxyl, making them incapable of ligation to 5′ adaptor. The 5′ cap was removed with RNA 5′ hydroxyl to monophosphate was assessed in two ways: first, using ligation of the 5′ hydroxyl ends for RNAs between 20 and 200 bp were gel purified. After 75 bp single-end sequencing, CoPRO showed pairing of 5′ and 3′ ends for RNAs between 20 and 68 bp in this dataset by only using reads where the whole insert was sequenced, so the adaptor was seen and trimmed from the end. This same method can be used to generate CoPRO-like information from a PRO-seq experiment.

In vitro tests of cap state selection. For tests of CoPRO treatments, radiolabeled RNA was made by incorporating 35S into the body of the RNA during transcription, using homemade T7 RNA polymerase and buffer (30 mM HEPES pH 7.8, 80 mM potassium glutamate, 15 mM MgAc, 0.25 mM EDTA, 5 mM DTT, 0.05% Tween–20, 2 mM spermidine, 2.5 mM ATP, GTP and UTP, 0.25 mM cold CTP), with YIPP (NEB, M2403S) and SUPERase-In (Thermo Fisher, AM2696) added as in the manufacturer’s protocol. Capped RNA was made by vaccinia capping (NEB, M2805S) but without subsequent uncapped clean-up; we estimate that capping was 80% complete (data not shown). Triphosphate requires no additional steps. The ability of different series of treatments to selectively reduce capped and uncapped RNA to monophosphate was assessed in two ways: first, using ligation of monophosphate to just as adaptor ligation requires a 5′ monophosphate in library preparation (Supplementary Fig. 15a) and, second, using ligation of the 5′ adaptor from a standard PRO-seq\(^\text{15}\) as the readout (Supplementary Fig. 15b).

Sequence alignment. Adaptor sequences were trimmed from paired-end reads with the cutadapt\(^\text{\textregistered}\) toolkit. Internal barcodes were used to demultiplex pooled libraries. Remaining sequence was trimmed with the bowtie2\(^\text{2}\) --very-sensitive option, which improved alignment of very short RNAs (<25 nt). We specified –X 1,000 (maximum insert size), –no-mixed (discard unpaired reads), –no-discard (no alignments >1 kb apart), and –unal (discard unaligned pairs). Alignments were performed against a pooled genome index that contained dmb and hg19 sequences. After alignment, the most 3′ nucleotide (corresponding to the nucleotide added during the run-on) was trimmed. For most analyses, we used reads shorter than 400 bp to avoid complications from co-transcriptional splicing. All scripts used for adaptor trimming and sequence alignment are provided (see Code availability).

Alignment statistics are summarized in Supplementary Table 2.

Read summarization and normalization. Alignments (for example, BAM files) were processed with a custom R script to summarize alignments sharing identical genomic coordinates (start and end) with an individual ‘count’ score. Reads >80 nt were weighted to correct for the known length bias of Illumina sequencers using previously reported weights\(^\text{3}\). For plots comparing absolute levels of capped and uncapped transcripts, such as Fig. 1f, the capped, uncapped, and capped and uncapped libraries were normalized such that capped + uncapped = capped and uncapped from 18 to 28 nt in length (at the set of maxTSNs used in Fig. 1f).

R scripts performing these summarization and normalization steps are available (see Code availability).

Defining TSNs, TSSs and TIDs. (See Supplementary Fig. 2 for a visual schematic.)

Start bases were identified as any 5′ end associated with at least three distinct pause positions (that is, 3′ ends within 55 nt), thus enriching for RNA polymerase pausing independently of total TSS activity and capping efficiency (Supplementary Fig. 2). The apparently distributed and ubiquitous nature of pausing ensures that even promoters where Pol II is frequently released to productive elongation are identified (Supplementary Fig. 2). We removed TSNs with less than 25% capped reads longer than 55 nt, as these were predominantly Pol II termination or RNA polymerase I and III transcription products. Sites with high levels of uncapped reads 300 bp upstream and downstream of a TSN were discarded as likely termination products (Supplementary Fig. 2). Chromosome 14 was excluded from all analyses.

Using these TSNs, TSSs were defined as TSN clusters on the same strand with no gaps larger than 60 nt. Similarly, TIDs were defined as TSS clusters on either strand with no gaps larger than 600 nt, beyond which enrichment for phased initiation is lost (Fig. 4c). These thresholds for grouping were chosen by looking at the separation of all possible pairings (Supplementary Fig. 2). Small changes to the threshold distances did not significantly affect the number of TSSs and TIDs called (data not shown).

TSSs were called as stable if they overlapped at least eight CAGE 5′ ends\(^\text{14}\). TSSs were classified as stable if at least eight CAGE 5′ ends overlapped them.

Some analyses were restricted to maxTSN or maxTSS to minimize effects from other nearby elements. We defined maxTSN as the TSN with the highest number of capped reads within a TSS. Similarly, maxTSS was defined as the TSS containing the highest number of total capped reads.

An R script and notebook to reproduce or refine these element definitions from our normalized K562 dataset is available (see Code availability).

Pause classification. To better understand pause behaviors, pause probability within the first 100 nt was computed for each maxTSN by dividing the TSN 3′ distribution vector (CoPRO capped read counts at each distance transcribed up to the first 100 nt of CoPRO total number of CoPRO read counts) by the overall probability of pausing between 20 and 200 nt and is the fourth quartile (highest probability of pausing between 33 and 60 nt). This classification is robust to experimental noise, as classes were consistent across our analyses.
experimental replicates and consistent with GRO-cap and PRO-seq measurements (Supplementary Fig. 17).

**Metaplots and heatmaps.** All metaplots in this work show a bootstrapped estimate of average signal from the sites summarized and the 87.5 and 12.5 percentiles. Briefly, this was done by taking 1,000 random samples of 10% of the data, averaging each subsample, and then calculating the median and confidence intervals from these 1,000-average profiles. Heatmaps in this work summarize sorted data into 200 lines by averaging every N/200 rows to produce a representative heatmap. This was initially developed for CoPRO data, as individual TSN pause profiles are sparse, and was subsequently used for all other data for which heatmaps are shown. The rationale for this is that the resolution of the files used for figures and the screen or printer used to display them would often only allocate a few hundred pixels to the heatmap. So it is best to intentionally bin data rather than allow binning to occur dependent on the mode of presentation. Where Roadmap Epigenomics datasets are referenced, we used the log fold change over input-normalized files provided by the NIH Roadmap Epigenomics Mapping Consortium. A list of ENCODE datasets used is provided in Supplementary Table 3.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** Useful scripts developed for this work are available on GitHub at https://github.com/ndt26/CoPRO.

**Data availability**

All sequencing data and processed bigwig and Rdata files have been deposited at GEO under accession GSE116472.

**References**

69. Young, L., Sung, J., Stacey, G. & Masters, J. R. Detection of *Mycoplasma* in cell cultures. *Nat. Protoc.* 5, 929–934 (2010).

70. Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* 11, 1455–1476 (2016).

71. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415 (2003).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- n/a
- Confirmed
- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

- bowtie2 2.2.6, samtools 0.1.19-96b5f2294a

Data analysis

- R version 3.3, gridExtra_2.3, doParallel_1.0.11, iterators_1.0.9, foreach_1.4.4, plyr_1.8.4, optparse_1.4.4, data.table_1.10.4-3, GenomicFile_1.10.3, rtracklayer_1.34.2, BiocParallel_1.8.2, GenomeInfoDb_1.10.3, SummarizedExperiment_1.4.0, Biobase_2.34.0, Rsamtools_1.26.2, Biostrings_2.42.1, XVector_0.14.1, viridis_0.5.1, viridisLite_0.3.0, latticeExtra_0.6-28, RColorBrewer_1.1-2, lattice_0.20-35, hexbin_1.27.2, bigWig_0.1-16, GenomicRanges_1.26.4, GenomeInfoDb_1.10.3, iRanges_2.8.2, S4Vectors_0.12.2, BiocGenerics_0.20.0, Python 2.7.6, numpy 1.8.2, matplotlib 1.3.1, pySam 0.9.1.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data will be made available on GEO. Useful code developed for this work will be available on GitHub. Useful intermediate files (including Rdata files with processed paired CoPRO data) will be made available upon request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences
☐ Behavioural & social sciences
☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size: N/A. This study utilizes a new functional genomic assay to characterize transcription initiation.

Data exclusions: No data were excluded.

Replication: CoPRO was done twice, with biological and technical replicates. Cells for replicates were separated by two passages, and library preparation was carried out during two separate weeks. After verifying that replicates were highly similar, they were pooled for subsequent analysis.

Randomization: N/A. The only difference between libraries is new enzymatic steps to examine 5' cap formation. Thus, samples had to remain identifiable throughout the protocol.

Blinding: N/A. This study focuses on development of a new functional genomic assay to examine transcription initiation and 5' cap formation in the well-characterized K562 cell line

Reporting for specific materials, systems and methods

Materials & experimental systems

Methods

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s): ATCC

Authentication: Cells were ordered fresh, morphology was consistent with K562

Mycoplasma contamination: Cells were negative for mycoplasma in a PCR test prior to the experiment. PRO-seq has detected mycoplasma in contaminated cells in the past: none was detected here.
Commonly misidentified lines (See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.