Published in final edited form as:

Nat Genet. 2018 November ; 50(11): 1533–1541. doi:10.1038/s41588-018-0234-5.

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

Jacob M. Tome[1,2,†], **Nathaniel D. Tippens**[1,3,4,†], and **John T. Lis**[5,6,*]

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA.

[2]Department of Genome Sciences, University of Washington, Seattle, WA, USA.

[3]Tri-Institutional Training Program in Computational Biology and Medicine, Cornell University, Ithaca, NY, USA.

[4]Department of Biological Statistics & Computational Biology, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA.

[5]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. jtl10@cornell.edu.

[6]Tri-Institutional Training Program in Computational Biology and Medicine, Cornell University, Ithaca, NY, USA. jtl10@cornell.edu.

Abstract

Eukaryotic RNA Polymerase 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 Pol2 initiation, capping, and pausing genome-wide. Our analyses reveal 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 (TSSes) 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 remarkably similar after the dramatic transcriptional changes caused by heat stress. Together, our results suggest that promoter- and enhancer-associated Pol2 is a regulatory nexus for integrating information across TSS ensembles.

*Correspondence to jtl10@cornell.edu.
†Contributed equally

Author Contributions

JMT and NDT conceived of CoPRO, carried out experiments, and analyzed data. NDT developed the computational framework for analysis. JMT, NDT, and JTL interpreted results and prepared the manuscript.

Competing Interests

The authors declare no competing interests.

Introduction

Mammalian gene promoters are often intimately coupled with upstream divergent transcription from a separate core promoter[1,2] that does not give rise to a stable RNA product. Similarly, many active enhancers are sites of divergent transcription with a basic architecture similar to gene promoters[3]. Transcribed element architecture can be more complex, including a TSS that is convergent to the main promoter, for example[4,5]. All of these transcripts are produced by the same enzyme: RNA Polymerase II (Pol2). Enhancer and upstream divergent transcription proceed through the same steps as gene promoter transcripts, including promoter proximal pausing and release[5], but lack many of the downstream features including absence of cleavage/polyadenylation sites and inclusion of +5' splice sites that are required to produce a stable transcript[3,5,6]. Sequence specific transcription factors (TFs) bind promoters and enhancers to regulate Pol2 transcription by acting at specific steps in the transcription cycle[7,8], and positioned nucleosomes flank these regions of transcription and factor binding[9–11]. From a mechanistic standpoint, the prevalence of transcription throughout the genome raises numerous questions about its role in gene regulation and chromatin architecture.

Transcription by Pol2 involves several steps[12], each of which is potentially rate limiting and regulated. Sequence-specific transcription factors bind DNA elements within accessible chromatin and recruit the general transcription machinery to promoters[7,8]. The general transcription factors assemble a pre-initiation complex (PIC) which in turn recruits Pol2 and facilitates initiation[13]. Next, Pol2 proceeds quickly to the site of the promoter proximal pause[8,14]: in some cases, proximally paused Pol2 may remain associated with the PIC[15]. 5' RNA capping takes place as Pol2 transitions from initiation to promoter proximal pausing[16], as the capping enzyme is recruited partly through association with the pausing and elongation factor DSIF[17,18], and is stimulated by Ser5 phosphorylation of the Pol2 CTD by TFIIH subunit Cdk7[19,20]. Capping involves addition of an inverted guanidine residue which is subsequently methylated[21]. Pol2 remains promoter proximally paused for varying amounts of time[22–25], providing an opportunity for regulatory input and assembly of a productive elongation complex. Pause release and entry into elongation is marked by P-TEFb mediated phosphorylation of Pol2 and pausing factors[26]. At the end of the transcription unit, cleavage and polyadenylation takes place, and Pol2 continues transcribing at a slower rate until termination is completed[27,28].

Methods that track the active site of Pol2[1,11,15,29–31] have identified promoter-proximal pausing as a key regulatory step for gene expression. Similarly, variants of these methods[3,9,32] map sites of transcription initiation with high sensitivity. Most studies have relied on aggregate molecular profiles from separate libraries mapping initiation and active sites to infer the distance that Pol2 travels before pausing. Here, we developed paired-end PRO-seq with selection for distinct 5' RNA capping states (Uncapped, Capped, or Capped and Uncapped), to provide coordinated information about 5' and 3' events during early Pol2 transcription genome-wide (Coordinated Precision Run-On and sequencing, or CoPRO, Figure 1a). With CoPRO, we examined promoter proximal pausing in both human K562 and mouse embryonic fibroblast (MEF) cells with single molecule resolution to understand determinants of pause location. By comparing separate libraries enzymatically selected for

either Uncapped or Capped 5’ ends, we map the process of RNA capping genome-wide in humans, revealing that capping dynamics vary depending on the distance that Pol2 transcribes before pausing. We leverage single-molecule profiles of pausing and elongation to identify transcription initiation sites with unprecedented specificity. We find that clusters of Pol2 initiation are often more complex than divergent pairs. These clusters, which we call Transcription Initiation Domains or TIDs, correspond with boundaries of evolutionary conservation, transcription factor binding, and chromatin modification. Therefore, we propose that TIDs represent a prominent structural unit above divergent TSSes, 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 minutes of heat shock at 42°C. All figures in this work use 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 Pol2 initiating from different TSNs but pausing at the same nucleotide create a horizontal line (Figure 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 (Figure 1b, Supplementary Figure 1 for additional examples). Notably, initiation from nearby TSNs tends to pause at the same nucleotide (horizontal lines). A wider view (Figure 1b, inset) reveals elongation beyond the pause (within the insert size limit of Illumina sequencing[33]), 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[9].

We developed a hierarchy to facilitate subsequent analyses (Figure 1c). First, individual nucleotides where initiation occurs (TSNs) are identified by having multiple distinct 3’ ends, among other criteria (Online Methods, Supplementary Figure 2). We cluster nearby TSNs into Transcription Start Sites (TSSes), as they are likely driven by the same transcription factors (TFs) and pre-initiation complex (PIC). TSSes are further clustered into higher-order structures: Divergent Pairs are TSSes transcribing in opposite directions <300 bp apart[3,9,34]; and Transcription Initiation Domains (TIDs) are larger TSS clusters (<600 bp gaps) on either strand. We use available CAGE data to determine which TSNs and TSSes give rise to stable RNAs such as mRNAs and lncRNAs[3]. Most TSSes produce unstable RNAs of unknown function (e.g., eRNA, ncRNA, upstream antisense RNA). In total, CoPRO

identified 458,549 TSNs, 78,467 TSSes, 18,411 divergent pairs, and 27,705 TIDs in K562 cells.

As previously described[35,36], most TSSes initiate at multiple TSNs, and TSN number increases with TSS activity (Figure 1d), though the number and distribution of TSNs varies widely (Supplementary Figure 3). TSSes where initiation uses few TSNs despite very high activity likely have exceptionally strong core promoter elements[35], such as *ACTB* and *HIST1H1*, (Supplementary Figure 1). While the basic unit of initiation is a divergent pair, each with its own PIC[3,9,34,37,38], most TSSes, especially at promoters, are found in larger clusters (Figure 1e). Previous work has noted specific types of TSS clusters, such as divergent pairs of TSSes[1,2], nearby alternative TSSes and head-to-head genes[34], or convergent TSSes[4], but CoPRO reveals that large constellations are much more prominent when all unstable TSSes are considered: promoter TIDs, on average, contain 1.4 stable TSSes paired with 4.7 unstable TSSes.

With CoPRO, we are able to determine exactly how far a Pol2 molecule transcribed after initiation with every read pair and thus calculate the ‘distance transcribed’, defined as the exact distance between the 5’ end (initiation site) and 3’ end (active site). The distance transcribed for the majority of Capped RNAs at TSNs are within the pause, reported to occur between 20–120 bp[3,15,39]. At single-molecule resolution, pausing is restricted to 20–60 nt (Figure 1f, blue) and often occurs at multiple non-contiguous positions (Supplementary Figure 3). Uncapped 5’ ends from TSNs are found associated with very short RNAs (Figure 1f, gold), confirming that capping begins as the 5’ triphosphate first emerges from polymerase[16,19].

Pause site specification by sequence and local activity

A histogram of the most used pause position per TSN (maxPause) reveals a bimodal distribution: one population from 20–32 nt (Early), and another from 32–60 nt (Late, Figure 2a), though many TSNs have some pausing in both windows. After grouping TSNs by distance transcribed at maxPause, the sequence determinants of pausing become clear (Figure 2b). Relative AT-richness is seen −30 nt to the initiation site, particularly in the Early pausing TSNs (Figure 2b), reflecting a stronger enrichment for the TATAWR motif (Supplementary Figure 4). In *Drosophila,* Pol2 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[15], and by detecting co-occurrence of PIC components at the initiation site and Pol2 at the pause site on the same DNA molecule[25]. Early pausing’s association with better TATAWR motifs and slightly higher TBP ChIP-seq signal (Supplementary Figure 5), suggest that early-pausing Pol2 may remain bound by the PIC in humans as well. Early pause TSNs are more AT-rich overall, whereas Late are more GC-rich +5 nt from the initiation site to 8 nt before the pause site. TSNs initiate preferentially on CA dinucleotides (the Inr motif[36]), and pause during the incorporation of a cytosine residue (Figure 2b, left). A capped CoPRO experiment in MEFs reveals a remarkably similar pattern of GC content around initiation and pause sites, indicating that these mechanisms of pause site choice are conserved (Figure 2b, right).

Paused Pol2 is most frequently found just *before* incorporating cytosine (Figure 2c). Cytosine is the least abundant nucleotide[40], and is therefore likely the slowest to be incorporated by Pol2. As a control, we examined the sequence around elongating Pol2 active site from CoPRO (reads with a distance transcribed >110 nt) or PRO-seq[3] (reads mapping outside of TIDs), and found a weaker preference for cytosine. This cannot be attributed to the biotin run-on in our protocol because a similar enrichment for pausing at cytosines has been observed with NET-seq[39], which does not include a run-on step. T4 RNA ligase has been found to have very little sequence bias[41,42], ruling out bias for C introduced during the rest of the library preparation. Pausing is thought to occur upon recruitment of several pause factors such as NELF and DSIF[43]. Thus, we propose this recruitment is most likely to occur during the slow step of cytosine incorporation, thus accentuating the tendency to observe Pol2 there. These results refine previous kinetic models for pausing: the position of the pause is dictated by the energy landscape of PIC interactions[15], the transcriptional bubble and RNA-DNA hybrid[31,39], local availability of pause factors[44], and finally nucleotide incorporation. A mixture of these features likely determines the position where pause factors ultimately associate with Pol2. As predicted by this model, Pol2 initiating at nearby TSNs tend to pause at the same position, and not simply a fixed distance after initiation (Figure 2d). Individual examples like the exceptionally dispersed *MAPK1* promoter (Figure 2e) 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 Online Methods). Early pause loci are enriched for enhancer- and lineage-specific TFs (GATA1, GATA2 and TAL1) and less active chromatin marks (Supplementary Figure 5), are more likely to occur at TSSes that do not produce a stable RNA product, and are associated with lower overall expression and elongation (Supplementary Figure 6). In contrast, Late pause loci are more enriched for activating transcription factors (ATF3 and ELF1), active chromatin marks (Supplementary Figure 5), and high expression and pause release (Supplementary Figure 6). Confirming the role of active features in the Late pause class, we find a significant shift from Late to Early pausing upon the widespread gene repression observed after heat shock in Mouse Embryonic Fibroblasts[45] (Supplementary Figure 6). Therefore, TSNs with Late pausing are endowed with GC content, TFs, and chromatin environments that facilitate productive elongation.

Capping dynamics vary with pause location

Early and Late pause TSNs have different uncapped RNA distributions, indicating an interplay between the processes of capping and pausing. TSNs of both classes have very short uncapped RNA (less than 22 nt), but Late also have uncapped RNAs at the pause site (Figure 3a, right) that are absent at Early pausing TSNs. In total, Early pause TSNs have much higher levels of uncapped RNA than Late (Figure 3b), but these uncapped transcripts are only observed at very short lengths. Capped and Uncapped CoPRO shows that the overall distribution of Pol2 at Early TSNs is continuous, transitioning from mostly uncapped to mostly capped RNAs (Figure 3b). Capped and uncapped distributions are best compared through joint probabilities, which show that capped RNAs in the Late pause mostly co-occur

with uncapped RNAs of the same length, while capped RNAs in the Early pause primarily co-occur with shorter uncapped RNAs (Figure 3c).

Initiation clusters are more complex than divergent pairs

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

While bidirectional pairing of mammalian TSSes has received considerable attention[9,34,38,46], larger TSS clusters, or TIDs, remain less well-characterized (Figure 4a). To illustrate, we identify the bin with the most CoPRO Capped initiation signal on each strand of promoter TIDs and sort them by distance between this pair of sites (Figure 4b). This pattern is nearly identical in non-promoter TIDs (Supplementary Figure 8a), indicating that this basic architecture is universal despite ‘non-promoters’ (e.g. enhancers) lower overall activity. Upstream antisense TSSes separated by >300 bp exhibit nucleosome occupancy and additional TSSes, with an additional sense peak flanking the antisense peak (Figure 4b, Extended maxTSSes), and are therefore best described as two divergent pairs, as shown previously with CAGE[34] and Start-seq[9].

Arrangements of TSSes and nucleosomes are tightly coupled

Convergent TSSes are coordinately positioned between nucleosomes[4] (Figure 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 (Figure 4c, blue). This indicates that nucleosome phasing from the strongest TSS constrains the location of convergent TSSes. However, the downstream sense TSS paired with the convergent TSS (i.e. TSS 4 if TSS 2 is the most active in Figure 4a) appears less constrained (Figure 4c, red vs blue), as their location is less peaked between nucleosomes phased to the strongest TSS. Interestingly, downstream sense TSSes lack antisense CoPRO signal until ~200 bp downstream of the maxTSS (Supplementary Figure 8e). This indicates that directionality of some minor TSSes could be enforced by the maxTSS +1 nucleosome blocking antisense transcription. Within TIDs, we find increased conservation scores around most TSSes and less conservation of intervening sequences (Supplementary Figure 8).

Nucleosome positioning follows two rules with respect to the clusters of TSSes in TIDs: i) nucleosomes are phased downstream of TSSes and ii) stronger TSSes dominate in phasing, thus constraining weaker TSSes. Therefore, the interplay between nucleosomes and TSSes 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 Figure 9), confirming an interplay between paused Pol2 and the first nucleosome[11,47]. Sorting heatmaps of MNase-seq relative to different TSSes reveals their power in explaining

registers (Supplementary Figure 8b-g, Supplementary Figure 10 for a summary). Two competing registers overlap in the space between Convergent maxTSSes 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 maxTSSes is lower than at those flanking Divergent maxTSSes (Figure 4b), confirming that nucleosome occupancy is reduced between convergently oriented active TSSes[4]. Nucleosomes between the convergent maxTSSes appear well-positioned to the sense maxTSS, but outer flanking nucleosomes do not, meaning that they are in registers not explained by these TSSes (Figure 4d, top). Sorting these same data by the location of other TSSes reveals registers in these regions that were randomly distributed in our first sort. Sorting on the upstream divergent TSS reveals a register upstream (Figure 4d, middle vs top). Likewise, the strongest downstream sense TSS is associated with a nucleosome register downstream (Figure 4d, bottom), thus explaining more phasing in the downstream direction than the maxTSS alone. In summary, convergent maxTSSes 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 Figure 8c). Thus, intricate registers of phased nucleosomes are common, but 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 Figure 10).

Chromatin environment and transcription are inexorably linked[48], and the comprehensive map of initiation afforded by CoPRO provides a framework with which to better understand their interplay. Histone modifications are distributed with distinct patterns relative to TSS pairs: H3K27ac is enriched after strong TSSes and between strong convergent pairs, H3K4me3 has a similar distribution with more spreading downstream of the maxTSS, and H3K4me1 is depleted near TSSes and enriched upstream (Figure 4e).

TIDs are clusters of regulatory elements

Though dramatic changes in gene transcription occur after one hour of heat shock, with 65% of all active genes showing significant up or downregulation by PRO-seq[45], CoPRO in MEFs shows that this primarily occurs by adjusting activity at existing TSSes (Figure 5a) and TIDs (Figure 5b), and not creation of new TSS or TIDs. This supports our previous findings that pause release is targeted for regulation during heat shock[45,49].

Pol2 is found at TID boundaries (oriented outward) and throughout their centers (Figure 6, PRO-seq). In broad TIDs the maxTSS tends to be centrally located (Supplementary Figure 11). TIDs are DNase hypersensitive and conserved (PhastCons, Supplementary Figure 12) suggesting functionality throughout these units. Nucleosome mapping by MNase-seq shows expected positioning beyond boundary TSSes, and weaker occupancy inside (Figure 6), similar to wide NFRs reported previously[50]. Many TFs bind TIDs with patterns consistent with their known functions (Supplementary Figure 12, Supplementary File for heatmaps of 171 ENCODE datasets[51]). Activating TFs such as MYC and ATF3 are enriched within promoter TIDs (Figure 6, Supplementary Figure 12). The enhancer and lineage-specific TFs GATA1, GATA2, and TAL1 are enriched at non-promoter TID boundaries (Figure 6 and

Supplementary File 1). Chromatin looping factors CTCF and Cohesin (RAD21) are enriched at boundaries, particularly upstream of small non-promoter TIDs, consistent with their roles as insulators and mediators of distal regulatory interactions. In contrast, a TF recently implicated in facilitating activating distal interactions, YY1[52,53], is enriched inside promoter TIDs.

TIDs demarcate chromatin domains at initiation sites

Histone modifications and variants also show diverse patterning relative to TIDs. Repressive marks like H3K9me3 are depleted within TIDs, and marks of elongation like H3K36me3 are depleted within TIDs and enriched downstream of the maxTSS (Supplementary Figure 13). Active marks are enriched within TIDs with variable spreading beyond their boundaries. The active histone variant H2A.Z (Figure 6) shows enrichment at nucleosomes flanking narrow promoter TIDs, reflecting its enrichment flanking the maxTSS and its upstream divergent TSS (Figure 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[54] strongly suggests that H3K27ac is primarily confined to nucleosomes immediately adjacent to TSSes (Supplementary Figure 14). While nucleosome-resolution data are not available for H3K4me3 in K562, two ChIP datasets show it extending past H3K27ac (Figure 6, Supplementary Figure 14), and data in other cell types show that H3K4me3 peaks tend to be wider than H3K27ac peaks[55]. 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 File 1).

Surprisingly, H3K4me1 is almost entirely excluded within TIDs (Figure 6), but rather flanks TIDs, with greatest enrichment surrounding narrow non-promoter TIDs. In general, narrow TIDs are less likely to be promoters (Figure 6, top vs bottom), are lowly expressed (Figure 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 Figure 13) than to other active marks.

Discussion

We developed CoPRO to provide coordinated information about Pol2 initiation and active site location, and cap status for single nascent transcript molecules genome-wide. We find 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 their regulatory function[56]. Importantly, we find that Late pausing sites switch to an Early pause during the widespread repression of many genes during heat shock. Thus, they likely represent different regulatory steps that Pol2 must pass through to assemble a full elongation complex. Early pausing is more linked 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.

Additionally, 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 (Figure 3b). Late pausing is associated with lower levels of uncapped transcripts overall, but their uncapped transcripts are found at the same locations as capped transcripts (Figure 3b). This difference may be due to the fact that capping enzyme is recruited by DSIF[17], which itself is recruited to Pol2 partly through an interaction with the nascent RNA[57]. Because Early paused transcripts would only have 4–18 nt of RNA extending outside of Pol2[58], they may recruit DSIF and capping enzyme less efficiently than Late pause sites, which complete capping with a longer transcript. This indicates that capping and pausing are tightly coupled, rather than simply acting as 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[51]. Thus, genome-wide functional data are critical for determining which of these elements are active in any one cell type, and characterizing responses to experimental perturbation. Analysis of such datasets is heavily dependent upon the framework utilized: in the 3.2 billion basepair human genome, one must know where to look to interpret any new dataset. *A priori* knowledge of regulatory elements is often used as such a framework. This usually 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[9], and even the best ChIP-seq peak calling algorithms identify binding sites with limited resolution[59]. To illustrate the impact of framework on interpretation, consider that Figure 4e and the promoter portion of Figure 6 for the same histone modifications show the same data at the same sites, but oriented and sorted differently to reveal different aggregate patterns and thus different functional relationships. In this study, we characterize transcription initiation with minimal assumptions about the properties of underlying elements. When all TSSes are considered, regulatory elements are composed of clusters of TSSes 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[55], however, we previously suggested this ratio is correlated with transcription level and not the basis for distinguishing a promoter from an enhancer[3]. Here, we find that patterns of chromatin modification are intimately coupled with TID architecture, in addition to transcription level. TSSes and nearby chromatin modifications are likely engaged in a feedback loop: transcription facilitates chromatin modification which in turn enables subsequent rounds of transcription, as demonstrated previously for TFIID recruitment and H3K4me3[60,61]. Thus, future studies are needed to examine the extent to which minor TSSes

within TIDs shape chromatin, or are opportunistic “passengers” within a permissive environment driven by more active TSSes within the TID.

Nucleosomes at TSSes are highly dynamic[62]; thus, TIDs have the ostensibly paradoxical properties of low nucleosome occupancy with abundant active histone modifications. TID internal nucleosomes are seemingly too labile for detection by MNase-seq, yet are detectible by ChIP for their modifications[50]. Recently, transcription was demonstrated to increase the mobility of promoter and enhancer loci in nuclei[63]: similarly, Pol2-driven ‘molecular stirring’ within TIDs may help 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[64] (histone tails, TFIID, Pol2 CTD, and Cyclin T1[65]) with a high potential for trans-activating interactions could establish phase-separated networks[66]. Although individual TSSes are driven by small units (150–300 bp) with limited information content[37], TIDs provide a molecular framework for information integration within larger domains. Consistent with this idea, broad H3K4me3 domains tend to be located at promoters of genes critical for cell identity and are associated with less cell-to-cell variability in mRNA levels[67,68]. 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.

Data Availability

All sequencing data, and processed bigwig and Rdata files are deposited under GEO accession GSE116472

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

Online Methods

Experimental Design

CoPRO adapts PRO-cap[15] for paired end sequencing, and includes a total of three different libraries that were enriched for either 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 allows us identify the precise location of pausing, and of where nascent RNAs become capped across the tens of thousands of initiation sites that we detect. The paired nature of the data enabled identification of sites of transcription initiation with unprecedented precision: we could use the pattern of pausing and elongation as a sensitive way of calling real initiation sites, and filter 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 non-coding transcription (such as eRNAs and upstream divergent RNAs) on an equal footing with longer lived RNAs like mRNA and lncRNAs. With our comprehensive maps of transcription initiation, we are able to compare the architecture of initiation sites with other features of the genome. For this purpose, we chose human K562 cells for direct comparison

with dozens of publically 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 (ThermoFisher cat. 12430054) supplemented with 10% FBS (ThermoFisher 10437028). Cultures were verified to be mycoplasma-free[69] prior to library preparation and sequencing. Two biological replicates were cultured independently, separated 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[70]. After isolating RNA from the run-on with Trizol (ThermoFisher, cat. 10296028), three run-ons per biological replicate were pooled. Two adapter ligations and reverse transcription were performed as described[70], with custom adapters detailed in Supplementary Table 1. Critically, no RNA fragmentation was done so that the pairing of RNA 5' and 3' ends remains biologically meaningful. The first ligation adds a sample barcode to the library; TruSeq barcodes were chosen to minimize predicted secondary structure of the adapter[71]. Between adapter ligations, cap state selection reactions were performed. The three cap state selections use a series of enzymatic treatments to reduce specific populations of RNAs to 5' monophosphate, making them capable of ligation to an RNA adapter by T4 RNA ligase (NEB, cat. M0204S). All steps were carried out following the manufacturer's protocol, with phenol:chloroform extraction and ethanol precipitation between steps. We designed three separate 5' state selections (Supplementary Figure 15c):

1. Uncapped RNAs were selected by treating with CIP, Calf Intestinal alkaline Phosphatase (NEB, cat. M0290S) and PNK, T4 Polynucleotide Kinase 3' phosphatase minus (NEB, cat. 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 (Epicentre, cat. TER51020) to degrade 5' monophosphate RNA (from terminating polymerase) and CIP to reduce other uncapped RNAs to 5' hydroxyl, making them incapable of ligating to 5' adapter. 5' cap was removed with RNA 5' pyrophosphohydrolase, RppH (NEB, cat. M0356S), using ThermoPol buffer (NEB, cat. B9004S). This treatment allows any transcript that has undergone the guanylation step of capping to be incorporated into this library. It is not sensitive to the subsequent methylation step of capping[21].
3. Capped + Uncapped RNAs were selected by treating with RppH, reducing both capped RNA and pre-capped RNA (5' triphosphate) to 5' monophosphate.

Each library for K562 CoPRO was sequenced on a NextSeq 500 as both cDNA (no PCR), and after PCR amplification with NEB Phusion (NEB, M0530) and PAGE purification as described[70]. See Supplementary Figure 16 for a schematic of library design.

MEFs were cultured in DMEM supplemented with 10% heat-inactivated FBS (v/v) and 1% penicillin/streptomycin (v/v) at 37°C with 5% CO_2 and 90% humidity. CoPRO was performed as described above using standard PRO-cap adaptor designs[70] instead of the PCR-free design. Libraries were PCR amplified for 15 cycles, and inserts from 10–200 bp were gel purified. After 75 bp single-end sequencing, CoPRO reveals pairing of 5' and 3' ends for RNAs between 20–68 bp in this dataset by only utilizing read where the whole insert was sequenced, so the adapter was seen and trimmed from the end. Note that 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 P^{32} into the body of the RNA during transcription, using home-made T7 RNA polymerase and buffer (30 mM HEPES pH 7.8, 80mM Potassium Glutamate, 15mM MgAc, 0.25 mM EDTA, 5 mM DTT, 0.05% Tween-20, 2 mM Spermidine, 2.5 mM ATP, GTP, and UTP, and 0.25 mM cold CTP), with YIPP (NEB cat. M2403S) and Superase-In (ThermoFisher cat. AM2696) added as per the manufacturer's protocol. Capped RNA was made by vaccinia capping (NEB cat. M2080S), 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 by using Terminator degradation as readout as it requires a 5' monophosphate for its exonuclease activity just as adapter ligation requires a 5' monophosphate in library preparation (Supplementary Figure 15a), and second by using ligation of the 5' adapter from a standard PRO-seq[70] as the readout (Supplementary Figure 15b).

Sequence alignment

Adaptor sequences were trimmed from paired-end reads with the 'cutadapt' toolkit. Internal barcodes were used to de-multiplex pooled libraries with a custom Python script. The remaining sequences were aligned with the bowtie2 --very-sensitive option, which improved alignment of very short RNAs (<25 nt). We specified –X 1000 (maximum insert size), --no-mixed (discard unpaired reads), --no-discordant (no alignments > 1 kbp apart), and --no-unal (discard unaligned pairs). Alignments were performed against a pooled genome index that contained dm6 and hg19 sequences. After alignment, the most 3' nucleotide (corresponding to the nucleotide added during the run-on) was trimmed. For most analyses, we only use reads shorter than 400 bp to avoid complications from co-transcriptional splicing. All scripts used for adaptor trimming and sequence alignment are provided (See Data Availability).

Alignment statistics are summarized in Supplementary Table 2.

Read summarization and normalization

Alignments (e.g. 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[33].

For plots comparing absolute levels of capped and uncapped transcripts such as Figure 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 Figure 1f).

R scripts performing these summarization and normalization steps are available (See Data Availability)

Defining transcription start nucleotides (TSNs), start sites (TSSs), and initiation domains (TIDs)

(See Supplementary Figure 2 for visual schematic)

Start bases were identified as any 5' end associated with at least 3 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 Figure 2). The apparently distributed and ubiquitous nature of pausing ensures that even promoters where Pol2 is frequenctly released to productive elongation are identified with this definition (Supplementary Figure 2). We removed TSNs with less than 25% capped reads longer than 55 nt, as these were predominantly Pol2 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 Figure 2). Chromosome M was excluded from all analyses.

Using these TSNs, TSSes 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 (Figure 4c). These thresholds for grouping were chosen by looking at the separation of all possible pairings (Supplementary Figure 2), and small changes to the threshold distances do not significantly affect the number of TSS and TIDs called (data not shown).

TSSes were called as stable if they overlapped at least eight CAGE 5' ends[46]. TSNs were classified as stable if at least eight CAGE 5' ends overlapped them.

Some analyses are 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 are 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 Data Availability).

Pause classification

To better understand pause behaviors, the pause probability within the first 100 nt was computed for each maxTSN by dividing the TSN's 3' distribution vector (CoPRO capped read counts at each distance transcribed up to 100 nt) by its total number of CoPRO capped reads up to 100 nt transcribed. Analysis of the cumulative distributions revealed that pausing

mostly occurs on 1–10 non-continuous major positions at mTSNs (Supplementary Figure 3). Further analysis of the max pause positions revealed a bimodal distribution (Figure 2a). To classify individual distributions as Early or Late pause, the probability of pausing anywhere between 20–32 nt was computed and divided into quartiles. Early is the first quartile (highest probability of pausing between 20–32 nt) and late is the fourth quartile (highest probability of pausing between 33–60 nt). This classification is robust to experimental noise, as classes were consistent across our experimental replicates and consistent with GRO-cap and PRO-seq measurements (Supplementary Figure 17).

Metaplots and heatmaps

All metaplots in this work show a bootstrapped estimate of average signal from the sites being summarized, along with 87.5 and 12.5 percentiles. Briefly, this is done by taking 1000 random samples of 10% of the data, averaging each subsample, and then calculating the median and confidence intervals from these 1000 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 TSNs' pause profiles are sparse, and was subsequently used for all other data for which heatmaps are shown. The rational here is that the resolution of the files used for figures, and the screen or printer used to display them often would only allocate at best a few hundred pixels to the heatmap, so it is best to intentionally bin data rather than allowing binning to occur depending on the mode of presentation.

Where Roadmap Epigenomics datasets are referenced, we are using the log fold change over input normalized files provided by the Consortium. A list of ENCODE datasets used is provided in Supplementary Table 3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank members of the Lis, Kwak, and Danko labs for helpful discussions throughout the experimental design and analysis phases of this work. Jay Mahat assisted in designing CoPRO libraries for sequencing without PCR amplification and the heat shock protocol. The Cornell Sequencing core and especially Peter Schweitzer were extremely helpful in designing libraries, and eminently patient in accommodating our technical requests. We acknowledge funding from NIH grants HG009393 and GM025332 to JTL. NDT was supported by National Institutes of Health (NIH) training grant T32HD057854.

References

1. Core LJ, Waterfall JJ & Lis JT Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–8 (2008). [PubMed: 19056941]
2. Seila AC et al. Divergent transcription from active promoters. Science (80-.). 322, 1849–1851 (2008).
3. Core LJ 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). [PubMed: 25383968]

4. Lavender CA et al. Downstream Antisense Transcription Predicts Genomic Features That Define the Specific Chromatin Environment at Mammalian Promoters. PLOS Genet. 12, e1006224 (2016). [PubMed: 27487356]
5. Henriques T et al. Widespread transcriptional pausing and elongation control at enhancers. 1–16 (2018). doi:10.1101/gad.309351.117.GENES
6. Almada AE, Wu X, Kriz AJ, Burge CB & Sharp PA Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature 499, 360–3 (2013). [PubMed: 23792564]
7. Vaquerizas JM, Kummerfeld SK, Teichmann SA & Luscombe NM A census of human transcription factors: Function, expression and evolution. Nat. Rev. Genet 10, 252–263 (2009). [PubMed: 19274049]
8. Adelman K & Lis JT Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet 13, 720–31 (2012). [PubMed: 22986266]
9. Scruggs BS et al. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol. Cell (2015). doi:10.1016/j.molcel.2015.04.006
10. Gilchrist D a et al. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell 143, 540–51 (2010). [PubMed: 21074046]
11. Weber CM, Ramachandran S & Henikoff S Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol. Cell 53, 819–30 (2014). [PubMed: 24606920]
12. Fuda NJ, Ardehali MB & Lis JT Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature 461, 186–92 (2009). [PubMed: 19741698]
13. Murakami K et al. Architecture of an RNA polymerase II transcription pre-initiation complex. Science (80-.). 342, (2013).
14. Sainsbury S, Bernecky C & Cramer P Structural basis of transcription initiation by RNA polymerase II. Nat. Rev. Mol. Cell Biol 16, 129–143 (2015). [PubMed: 25693126]
15. Kwak H, Fuda NJ, Core LJ & Lis JT Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339, 950–3 (2013). [PubMed: 23430654]
16. Rasmussen EB & Lis JT In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc. Natl. Acad. Sci. U. S. A 90, 7923–7 (1993). [PubMed: 8367444]
17. Mandal SS 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. U. S. A 101, 7572–7 (2004). [PubMed: 15136722]
18. Lidschreiber M, Leike 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, 3805–16 (2013). [PubMed: 23878398]
19. Nilson KA et al. THZ1 Reveals Roles for Cdk7 in Co-transcriptional Capping and Pausing. Mol. Cell 59, 576–587 (2015). [PubMed: 26257281]
20. Moteki S & Price D Functional coupling of capping and transcription of mRNA. Mol. Cell 10, 599–609 (2002). [PubMed: 12408827]
21. Ramanathan A, Robb GB, Chan S & Biolabs NE mRNA capping: biological functions and applications. 1–16 (2016). doi:10.1093/nar/gkw551
22. Henriques T et al. Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. Mol. Cell 52, 517–528 (2013). [PubMed: 24184211]
23. Buckley MS, Kwak H, Zipfel WR & Lis JT Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. Genes Dev. 28, 14–9 (2014). [PubMed: 24395245]
24. Shao W & Zeitlinger J Paused RNA polymerase II inhibits new transcriptional initiation. Nat. Publ. Gr (2017). doi:10.1038/ng.3867
25. Krebs AR et al. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Mol. Cell 1–12 (2017). doi:10.1016/j.molcel.2017.06.027
26. Jonkers I, Kwak H & Lis JT Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife 3, e02407 (2014). [PubMed: 24843027]
27. Schwalb B et al. TT-seq maps the human transient transcriptome. Science 352, 1225–8 (2016). [PubMed: 27257258]

28. Proudfoot NJ Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. Science (80-.). 352, 715–718 (2016).
29. Nojima T et al. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell 161, 526–540 (2015). [PubMed: 25910207]
30. Mayer A et al. Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. Cell 161, 541–554 (2015). [PubMed: 25910208]
31. Nechaev S et al. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science 327, 335–8 (2010). [PubMed: 20007866]
32. Andersson R et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]
33. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L & Quake SR Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. Clin. Chem 56, 1279–1286 (2010). [PubMed: 20558635]
34. Chen Y et al. Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. Nat. Genet 48, 984–994 (2016). [PubMed: 27455346]
35. Carninci P et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38, 626–635 (2006). [PubMed: 16645617]
36. Vo Ngoc L, Wang Y-L, Kassavetis GA & Kadonaga JT The punctilious RNA polymerase II core promoter. Genes Dev. 31, 1289–1301 (2017). [PubMed: 28808065]
37. Arensbergen J. Van et al. Genome-wide mapping of autonomous promoter activity in human cells. Nat. Publ. Gr 35, (2016).
38. Duttke SHC et al. Human Promoters Are Intrinsically Directional. Mol. Cell 57, 674–684 (2015). [PubMed: 25639469]
39. Gressel S et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. Elife 6, 1–24 (2017).
40. Traut TW Physiological concentrations of purines and pyrimidines i. 1–22 (1994).
41. Ramani V, Qiu R & Shendure J High-throughput determination of RNA structure by proximity ligation. Nat. Biotechnol 33, 980–984 (2015). [PubMed: 26237516]
42. Song Y, Liu KJ & Wang T-H Elimination of Ligation Dependent Artifacts in T4 RNA Ligase to Achieve High Efficiency and Low Bias MicroRNA Capture. PLoS One 9, e94619 (2014). [PubMed: 24722341]
43. Chen FX, Smith ER & Shilatifard A Born to run: control of transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol 1 (2018). doi:10.1038/s41580-018-0010-5
44. Li J et al. Kinetic Competition between Elongation Rate and Binding of NELF Controls Promoter-Proximal Pausing. Mol. Cell 50, 711–722 (2013). [PubMed: 23746353]
45. Mahat DB, Salamanca HH, Duarte FM, Danko CG & Lis JT Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. Mol. Cell 62, 63–78 (2016). [PubMed: 27052732]
46. Core LJ et al. Supplement: Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet 46, 1311–20 (2014). [PubMed: 25383968]
47. Gilchrist D a et al. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev. 22, 1921–33 (2008). [PubMed: 18628398]
48. Karlic R, Chung H-R, Lasserre J, Vlahovicek K & Vingron M Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci 107, 2926–2931 (2010). [PubMed: 20133639]
49. Vihervaara A et al. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat. Commun 8, 1–15 (2017). [PubMed: 28232747]
50. de Dieuleveult M et al. Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. Nature 530, 113–116 (2016). [PubMed: 26814966]
51. Consortium TEP An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]

52. Beagan JA et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. 27, 1139–1152 (2017). [PubMed: 28536180]
53. Weintraub AS et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell 1–16 (2018). doi:10.1016/j.cell.2017.11.008
54. Pradeepa MM et al. Histone H3 globular domain acetylation identifies a new class of enhancers. Nat. Genet 48, 681–686 (2016). [PubMed: 27089178]
55. Heintzman ND et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet 39, 311–318 (2007). [PubMed: 17277777]
56. Engreitz JM et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature (2016). doi:10.1038/nature20149
57. Missra A & Gilmour DS Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. Proc. Natl. Acad. Sci. U. S. A 107, 11301–6 (2010). [PubMed: 20534440]
58. Bernecky C, Herzog F, Baumeister W, Plitzko JM & Cramer P Structure of transcribing mammalian RNA polymerase II. Nature 1–14 (2016). doi:10.1038/nature16482
59. Zhang Y et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008). [PubMed: 18798982]
60. Lauberth SM et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 152, 1021–1036 (2013). [PubMed: 23452851]
61. Vermeulen M et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell 131, 58–69 (2007). [PubMed: 17884155]
62. Shivaswamy S et al. Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation. PLoS Biol. 6, e65 (2008). [PubMed: 18351804]
63. Gu B et al. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. Science 3136, eaao3136 (2018).
64. Hughes MP et al. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. Science 359, 698–701 (2018). [PubMed: 29439243]
65. Lu H et al. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. Nature 1 (2018). doi:10.1038/s41586-018-0174-3
66. Hnisz D, Shrinivas K, Young RA, Chakraborty AK & Sharp PA A Phase Separation Model for Transcriptional Control. Cell 169, 13–23 (2017). [PubMed: 28340338]
67. Benayoun BA et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell 158, 673–688 (2014). [PubMed: 25083876]
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). [PubMed: 26301496]
69. Young L, Sung J, Stacey G & Masters JR Detection of Mycoplasma in cell cultures. Nat. Protoc 5, 929–934 (2010). [PubMed: 20431538]
70. Mahat DB 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). [PubMed: 27442863]
71. Zuker M Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–15 (2003). [PubMed: 12824337]

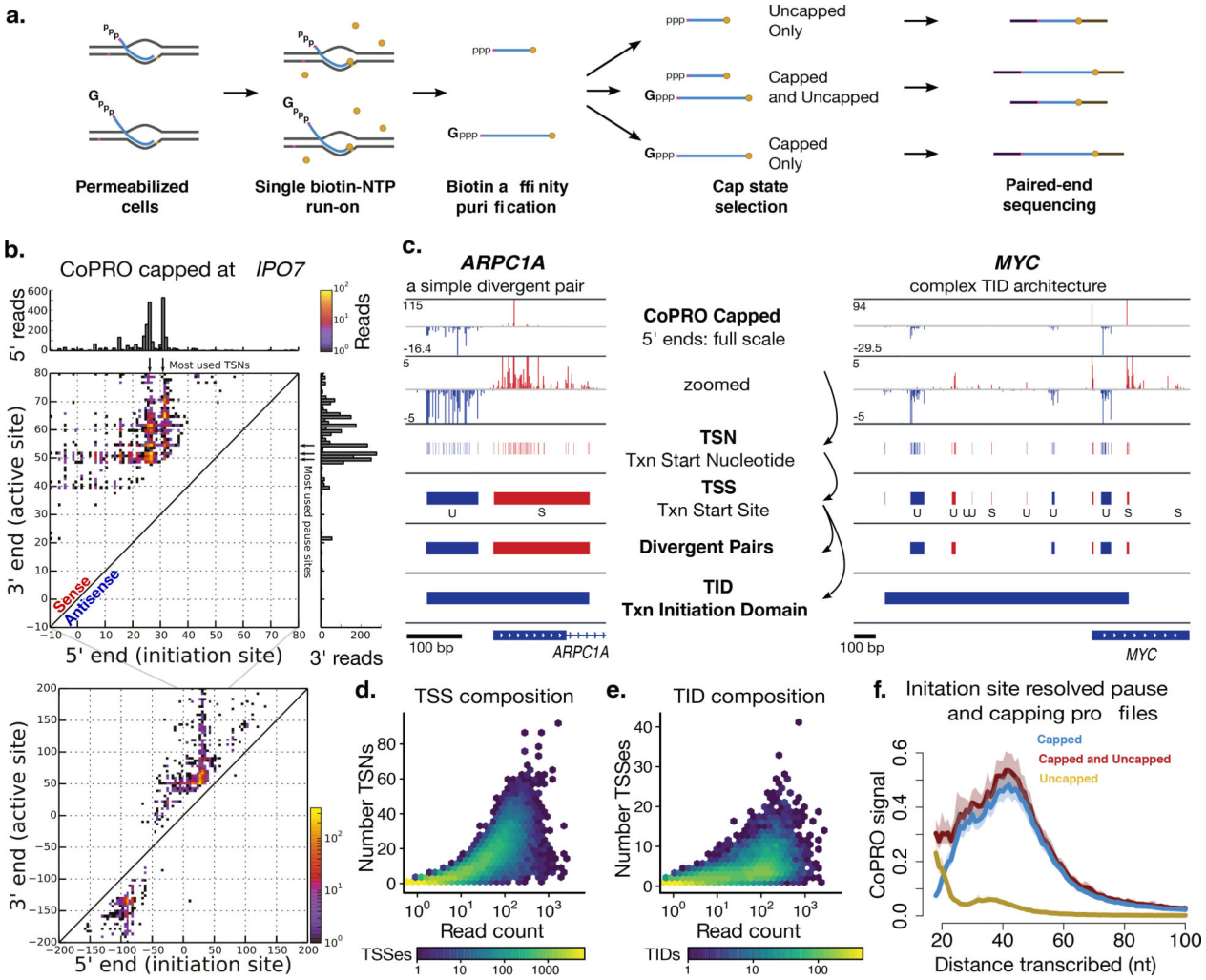

Figure 1: CoPRO simultaneously measures initiation and active site of Pol2 genome-wide.

(a) Overview of CoPRO. A single biotin NTP is incorporated by Pol2 by nuclear run-on, enabling biotin affinity purification of nascent RNA between steps in strand-specific RNA-seq library preparation. A series of enzymatic steps select for specific 5’ end states to make three different cap status-selected libraries (Online Methods). Paired-end sequencing maps initiation and active sites of individual Pol2 molecules. **(b)** CoPRO capped plot for the *IPO7* gene. 0 is the RefSeq TSS. Each bin represents a unique 5’ (initiation site, x axis) and 3’ (active site, y axis) pairing colored by number of reads mapped to that bin. Expanded view below. **(c)** Hierarchy of transcription initiation. CoPRO capped (top) is used to call TSNs, TSNs within 60 nt are grouped into TSSes, divergently oriented TSSes within 300 bp are grouped into TSS pairs, and TSSes within 600 bp are grouped into TIDs. **(d)** TSN number vs CoPRO capped expression at TSSes, N = 78,467. **(e)** TSS usage vs total CoPRO capped expression at TIDs, N = 27,705 **(f)** Pause profile of capped, uncapped, and capped and uncapped CoPRO at the TSN with highest CoPRO capped read count (maxTSN) in each TSS (only maxTSNs that are uniquely mappable at 18 nt are used) N = 30,240. Bootstrapped average signal, solid line; 12.5th and 87.5th percentiles, shaded (Online Methods). ‘Distance

transcribed' is the length of nascent RNA, and thus specifically tracks pausing from each initiation event.

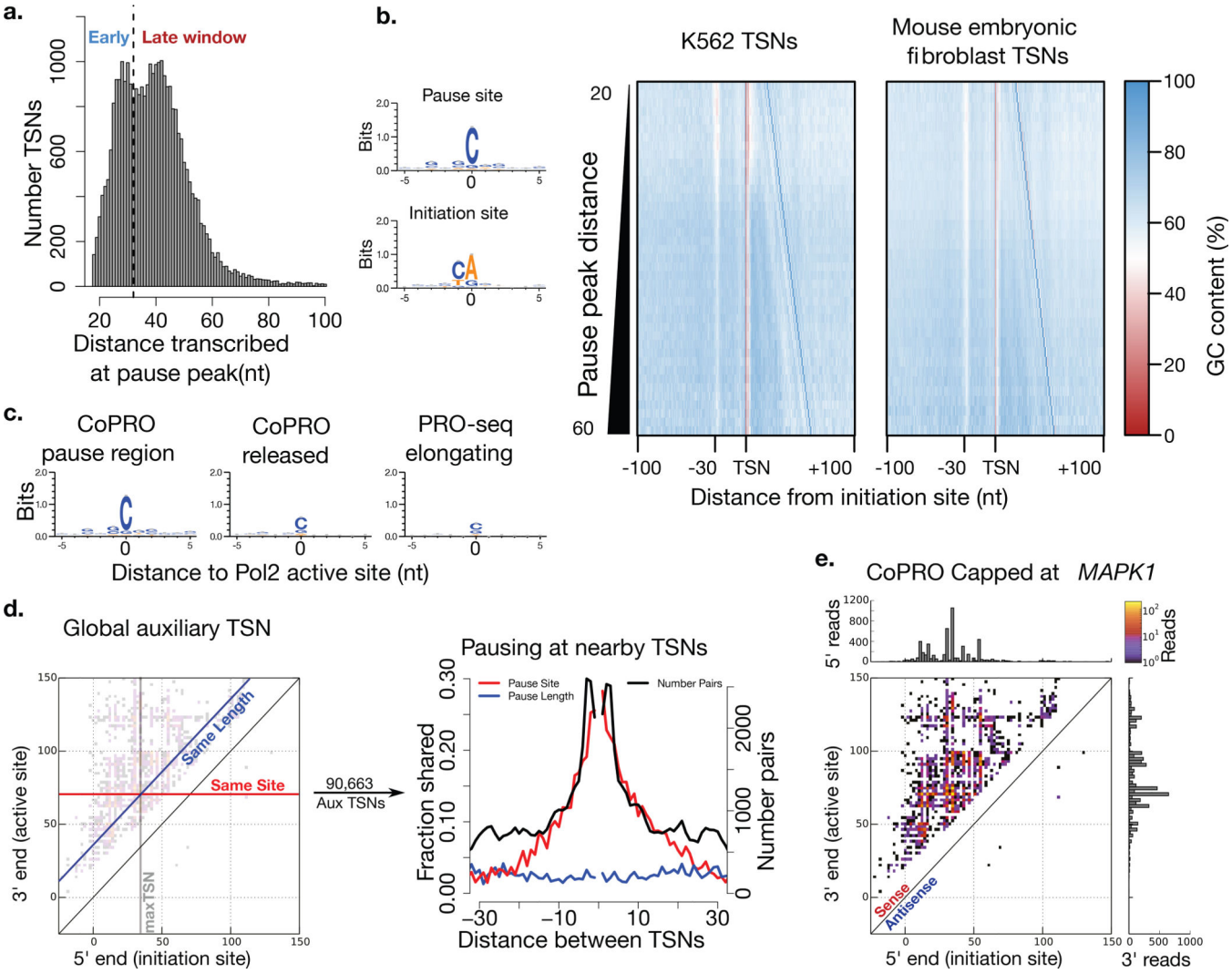

Figure 2: Sequence determinants of pause position choice.

(a) Histogram of most used pause position from maxTSNs, N = 30,240. **(b)** Average GC content around maxTSNs. Each row is the average of all maxTSNs sharing a maxPause distance from 20–60 nt. Sequence logo for initiation site and pause site in K562 at left. At right, sequence composition around mouse embryonic fibroblast TSNs from CoPRO capped data processed identically. N = 46,275. **(c)** Sequence logos for the top 3 most occupied 3' ends per maxTSN within the pause region (20–55 nt, CoPRO pause region), or after pause release (>110 nt, CoPRO released). 3' positions of elongating polymerase were found by taking PRO-seq reads from outside of TID boundaries (PRO-seq, right). Nucleotide 0 is the location of the active site of Pol2 (i.e. the nucleotide that was incorporated during the run-on). **(d)** Fraction of TSNs within the same TSS (auxiliary TSNs) that pause at the same nucleotide (red) and same length (blue) as their maxTSN, as a function of distance between the TSNs. Pause nucleotide is shared if the 3' nucleotide with the most CoPRO capped signal for the auxiliary TSN is the same as the maxTSN's (red). Pause length is shared when the distance transcribed at the 3' position with the most CoPRO capped signal is the same for the auxiliary and maxTSN (blue). Number of auxiliary TSNs observed at each distance is

shown in black. N = 90,663 auxiliary TSNs. **(e)** CoPRO capped RNA plot for the *MAPK1* gene.

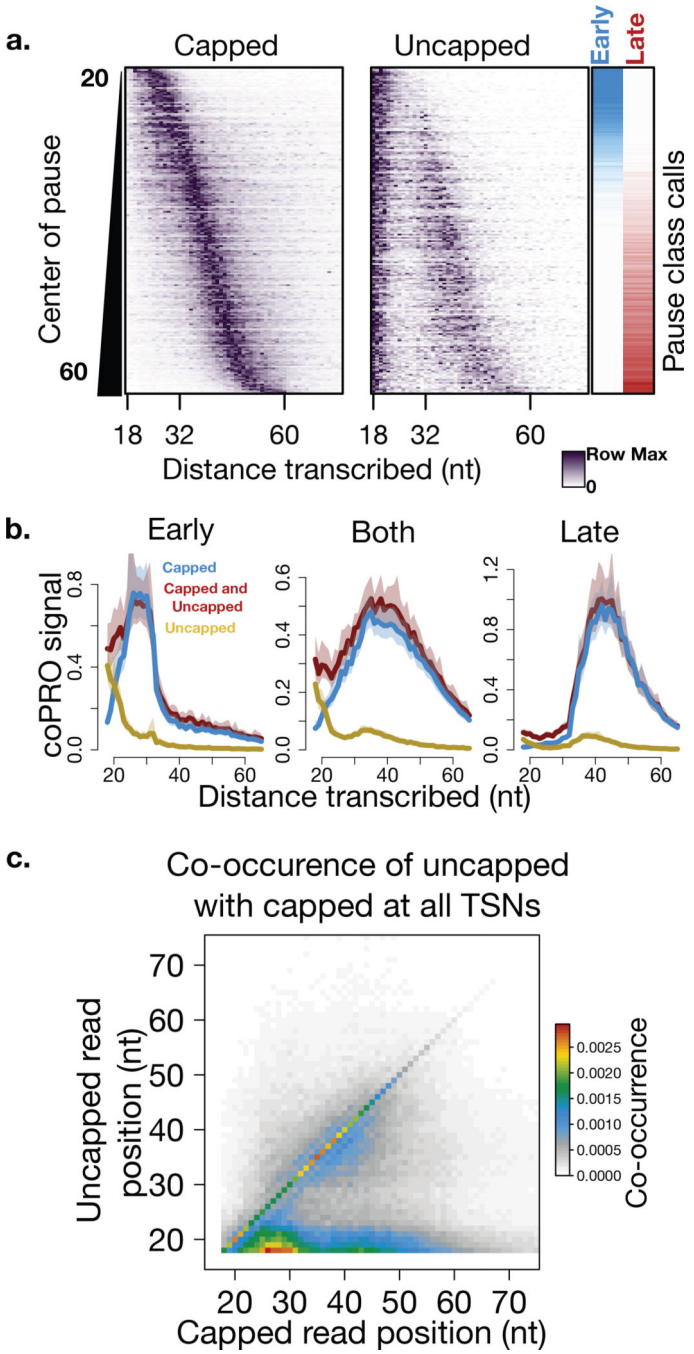

Figure 3: Late pausing TSNs are capped more efficiently, but later than early pause TSNs

(a) CoPRO capped and uncapped signal for maxTSNs sorted by the center of their pause, and row normalized so that maximum value of each row is 1. Early/Late called TSNs are indicated at the far right (Early/Late are top/bottom quartile by fraction of total CoPRO capped from 18–65 nt that falls within the Early window, 18–32 nt: see Online Methods). N = 30,240 **(b)** Metaplots of three cap-state-selected RNA libraries at maxTSNs mappable at 18 nt that were classified as either predominantly Early or Late pausing, or containing a mixture of pause classes (Both, middle 50 percent by fraction of pause in the Early window).

Same maxTSNs as **a**. Bootstrapped average, solid line, 12.5 and 87.5 percentiles, shaded. Early N = 7,320, Both N = 14,640, Late N = 7,320. **(c)** Joint probability distribution of uncapped CoPRO with capped CoPRO signal at the same TSN, from 18 to 100 nt. All maxTSNs mappable at 18 nt were used to generate a meta joint probability distribution (N = 30,240). A vector of the number of reads at each possible distance transcribed from 0 to 100 for each TSN is normalized to sum to 1 in both Capped and Uncapped CoPRO (representing probability) before calculating joint probability (dot product of each pair) and averaging. Thus, the entire distribution sums to 1.

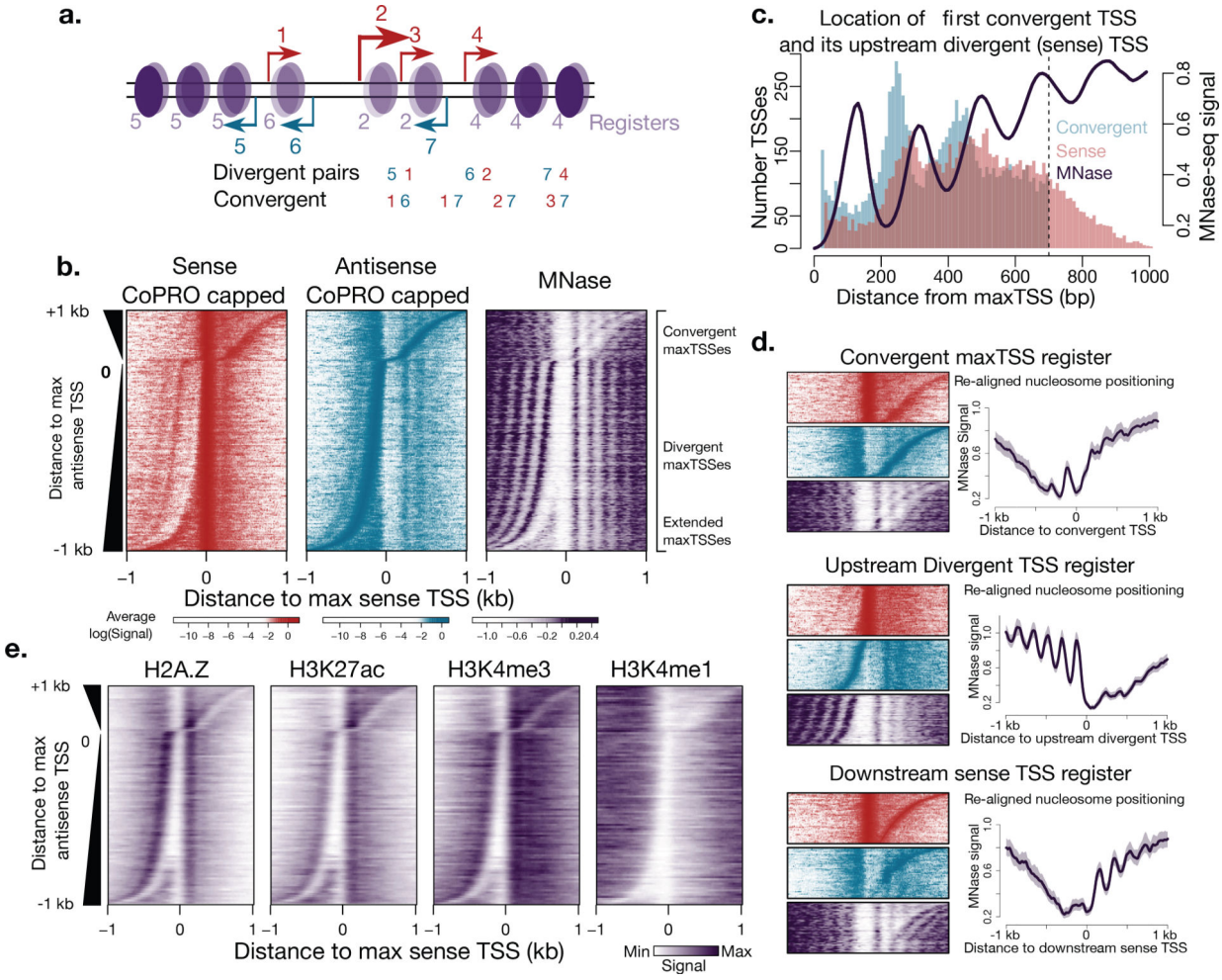

Figure 4: A global view of initiation shows rules for divergent pairing and reveals widespread complex organization.

(a) Model TID. Multiple TSSes 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 explaining 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), −/+1 kb. N = 9,273 TIDs with at >2 TSSes and >6 reads per strand (1,892 Convergent maxTSSes, 5,749 Divergent maxTSSes, and 1,632 Extended maxTSSes). **(c)** Location of the strongest convergent TSS and its paired upstream divergent TSS (‘Sense’) relative to maxTSS for all TIDs in **b**. Convergent TSSes are the strongest antisense bin from 0 to 700 bp downstream of maxTSS. The upstream divergent (‘Sense’) TSS is the strongest bin on the sense strand from 0 to −300 to the Convergent TSS. **(d)** Alternative registers revealed in aggregate by resorting convergent TIDs by the location of different TSSes. MNase-seq metaplots centered on the TSS used to sort assess its ability to explain nucleosome positioning. Top group, sorted by convergent TSS (repeated from b). Middle, sorted by the upstream divergent TSS (−300–0 bp, opposite strand). Bottom, sorted by strongest TSS from 100–1000 bp

downstream of the maxTSS (on the same strand). **(e)** Epigenomics Roadmap histone modification data, sorted and oriented as in **b**.

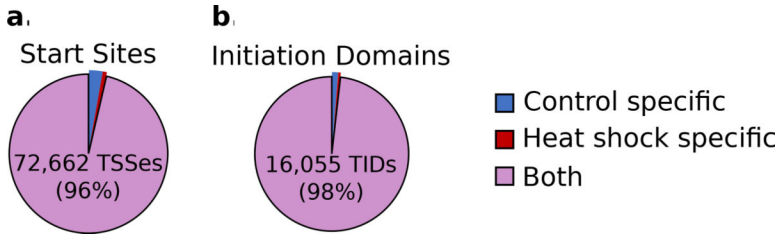

Figure 5: Massive regulatory changes in heat shock occur by modulating activity of pre-established TSSes

(a.) Fraction of TSSes identified by CoPRO capped before (control) and after 60 minutes of heat shock in MEFs. Despite significant changes in expression of 65% of genes[45], 96% of TSSes are observed under both conditions. **(b)** Fraction of TIDs before and after heat shock. 98% of TIDs are detected under both conditions.

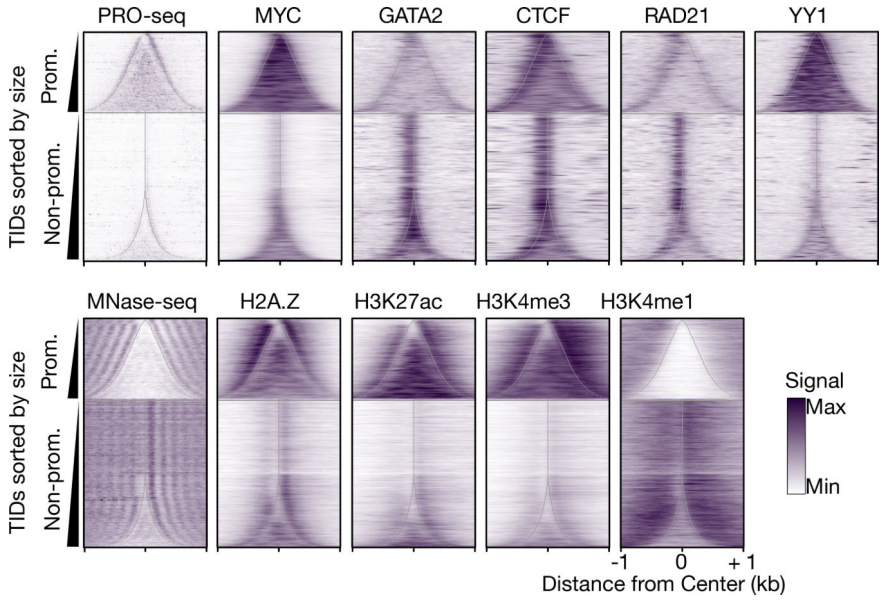

Figure 6: TID organization is linked to chromatin environment.

All heatmaps made as described in Supplementary Methods. Promoter TIDs (top) contain at least one TSS with enough CAGE signal to be called as stable, while Non-promoter TIDs contain no TSSes with significant evidence for accumulation of stable RNA products (i.e. putative enhancers). TIDs are sorted by width, centered, and oriented so the TSS with the highest CoPRO capped read count transcribes to the right. N = 27,705 TIDs. PRO-seq signal[3] is the sum of activity on both strands. Other than PRO-seq and histone modification ChIP-seq datasets (Roadmap Epigenomics), all others are from ENCODE. Histone modifications are on a linear color scale of normalized fold-change computed by Roadmap Epigenomics, all others are on a log color scale.