bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Genetic Dissection of the RNA Polymerase II Transcription Cycle

Shao-Pei Chou[1], Adriana K. Alexander[1,2], Edward J. Rice[1], Lauren A Choate[1], Paula E Cohen[2], and Charles G. Danko[1,2,*]

[1] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
[2] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

[*] **Address correspondence to:**
Charles G. Danko, Ph.D.
Baker Institute for Animal Health
Cornell University
Hungerford Hill Rd.
Ithaca, NY 14853
Phone: (607) 256-5620
E-mail: dankoc@gmail.com

Abstract

How DNA sequence affects the dynamics and position of RNA Polymerase II (Pol II) during transcription remains poorly understood. Here we used naturally occurring genetic variation in F1 hybrid mice to explore how DNA sequence differences affect the genome-wide distribution of Pol II. We measured the position and orientation of Pol II in eight organs collected from heterozygous F1 hybrid mice using ChRO-seq. Our data revealed a strong genetic basis for the precise coordinates of transcription initiation and promoter proximal pause, allowing us to redefine molecular models of core transcriptional processes. Our results implicate the strength of base pairing between A-T or G-C dinucleotides as key determinants to the position of Pol II initiation and pause. We report evidence that initiation site selection follows a stochastic process similar to brownian motion along the DNA template. We found widespread differences in the position of transcription termination, which impact the primary structure and stability of mature mRNA. Finally, we report evidence that allelic changes in transcription often affect mRNA and ncRNA expression across broad genomic domains. Collectively, we reveal how DNA sequences shape core transcriptional processes at single nucleotide resolution in mammals.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Introduction

Transcription by RNA polymerase II (Pol II) results in the synthesis of mRNAs encoding all protein-coding genes. Pol II transcription is a cyclic process that can be divided into stages representing transcription initiation, pause, elongation, and termination (Fuda et al., 2009; Jonkers and Lis, 2015). During the first stage of the transcription cycle, RNA polymerase is initiated on regions of accessible chromatin by the collective actions of transcription factors and co-factors that recruit the pre-initiation complex (PIC), melt DNA, and initiate Pol II (Grünberg et al., 2012; Haberle and Stark, 2018; Murakami et al., 2013; Tsai and Sigler, 2000). After initiation, Pol II pauses near the transcription start site of nearly all genes in metazoan genomes (Jonkers et al., 2014; Muse et al., 2007; Rougvie and Lis, 1988). Pol II is released from pause by transcription factors and a key protein kinase complex (P-TEFb), a tightly regulated step that controls the rates of mRNA production (Danko et al., 2013; Dig B. Mahat et al., 2016; Rahl et al., 2010; Zeitlinger et al., 2007). After pause release, Pol II elongates through gene bodies, which in some cases cover more than 1 MB of DNA in mammals (Carninci et al., 2005). Finally the co-transcriptionally processed pre-mRNA is cleaved from the elongating Pol II complex, allowing termination and recycling of Pol II (Cho et al., 1999; O'Sullivan et al., 2004; Rosonina et al., 2006). Intensive research efforts during the past thirty years have provided detailed knowledge of the proteins, RNAs, and other macromolecules that facilitate Pol II transcription (Gilchrist et al., 2010; Miller et al., 2001; Nechaev et al., 2010; Orphanides et al., 1998; Ranish et al., 1999).

We still have a relatively rudimentary understanding about how DNA sequence influences each step in the Pol II transcription cycle. Of all stages in Pol II transcription, the DNA sequence determinants of transcription initiation are perhaps the best characterized. DNA sequence motifs such as the TATA box, B recognition element (BRE), and the initiator motif are reported to bind proteins in the Pol II preinitiation complex (PIC) and thereby set the transcription start site (Carninci et al., 2006; Nilson et al., 2017; Smale and Baltimore, 1989). SNPs affecting these core transcription initiation motifs can affect the rates of mRNA production indicating the central importance of DNA sequence affecting transcription initiation (Kristjánsdóttir et al., 2020). In addition, DNA sequence motifs that correlate with Pol II pausing and termination have also been reported (Gressel et al., 2017; Schwalb et al., 2016; Tome et al., 2018). In all cases, these studies show that DNA sequence motifs involved in Pol II transcription are weak, degenerate, and spread across wide genomic regions. Therefore we are still a long way from developing predictive models that describe interactions between DNA sequence context and the steps and dynamics of Pol II transcription.

Here we used naturally occurring genetic variation between heterozygous F1 hybrid mice to understand how DNA sequence differences between alleles affect the Pol II transcription cycle. We generated an atlas of the position and orientation of RNA polymerase II in eight organs collected from three primary germ layers using ChRO-seq (Chu et al., 2018). Our detailed analysis reveals insight into how DNA sequence shapes each of the stages during the Pol II transcription cycle. For initiation, our results support a new model in which Pol II selects initiation sites through a process similar to brownian motion. We show that Pol II pauses on a C base following a G-rich stretch and is affected by short indels between the initiation and pause position, together explaining more than half of the observed changes in Pol II pause position. Finally, we reveal substantial allelic differences in the position of transcription termination, which in some cases

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

affect the primary structure and transcript stability of the mature mRNA. Collectively, we provide new insight into how DNA sequence shapes core transcriptional processes in mammals.

Results

Atlas of allele specific transcription in F1 hybrid murine organs

We obtained reciprocal F1 hybrids from two heterozygous mouse strains, C57BL/6 (B6) and Castaneus (CAST) (**Figure 1A**). Mice were harvested in the morning of postnatal day 22 to 25 from seven independent crosses (3x C57BL/6 x CAST and 4x CAST x C57BL/6; all males). We measured the position and orientation of RNA polymerase genome-wide in 8 organs using a ChRO-seq protocol designed to improve the accuracy of allelic mapping by extending the length of reads using strategies similar to length extension ChRO-seq (Chu et al., 2018) (**see Methods**). We obtained 376 million uniquely mapped ChRO-seq reads across all eight organs (21-86 million reads per organ; **Supplementary Table 1**) after sequencing, filtering, and mapping short reads to individual B6 and CAST genomes. Hierarchical clustering using Spearman's rank correlation of ChRO-seq reads in GENCODE annotated gene bodies (v.M25) grouped samples from the same organ. Additionally, organs with similarities in organ function clustered together, for instance: heart and skeletal muscle, and large intestine and stomach (**Figure 1B**).

We identified 1,374 genes and lincRNAs with strong evidence that transcription was significantly higher across the gene body on either the B6 or CAST allele in at least one organ, comprising about 8% of the 17,703 annotated genes. Visualizing the pattern of allelic transcription revealed that most allelic changes were specific to a single organ (**Figure 1C**). A minority of annotated genes showed evidence of genomic imprinting (n = 51), and these were generally imprinted in all organs analyzed here (**Figure 1C**). Organ-specific allelic biased transcription was not explained by either false-negatives in putatively unbiased organs (**Figure 1—figure supplement 1 A,B**) or by organ-specific differences in gene expression (**Figure 1—figure supplement 1C**). Notably, even organs with similar gene expression patterns (e.g., heart and skeletal muscle; large intestine and stomach) did not show much correlation in allelic differences between B6 and CAST. Taken together, these results show that most allelic differences in gene transcription were organ specific and showed heritability patterns that were consistent with a genetic cause.

Examination of genome browser tracks revealed that genes with allele specific transcription frequently clustered in genomic regions that had multiple separate transcripts with allelic differences. For instance, the imprinted domain associated with Angelman syndrome contained twenty ncRNAs and four genes that were transcribed more highly from the paternal allele and two genes transcribed from the maternal allele (**Figure 1D**). Multiple transcriptional changes occurring across a single locus in this example, and others like it that fit both imprinting and genetic inheritance patterns (**Figure 1—figure supplement 1D**), are consistent with differences affecting regulatory regions that control the activity of broad transcription domains. We asked whether genes located near one another shared allelic changes more frequently than expected if changes were caused by independent genetic or epigenetic differences. Indeed, we found that genes with allelic bias were significantly more likely to have allelic changes in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

expression of an adjacent transcript compared with genes which were not changed (brain: p-value = 5.35e-4, liver: p-value < 2.2e-16; Fisher's Exact Test).

To identify domains with evidence of allelic bias in an annotation-independent manner, we next analyzed ChRO-seq data from all eight organs using AlleleHMM (Chou and Danko, 2019). We identified 3,494 domains that showed consistent evidence of allelic imbalance (**Figure 1—figure supplement 1E, Supplementary Table 2**). The majority of these domains (n = 3,466) had consistent effects in each mouse strain (called strain-effect domains), the pattern expected if allelic imbalance was caused by DNA sequence differences between strains. Twenty-eight domains showed consistent evidence of genomic imprinting (imprinted domains). On average, both strain-effect and imprinted domains spanned broad genomic regions (~10-1,000 kb; **Fig. 1E**) that were frequently composed of two or more transcription units, including annotated genes and ncRNAs (either long intergenic ncRNAs and/ or enhancer-templated RNAs; **Fig. 1F**). Imprinted domains tended to be larger, have allelic differences detected across multiple tissues, and affected larger numbers of genes than strain effect domains (**Fig. 1E-F**). However, despite the overall trend toward larger imprinted domains, we nevertheless did identify many cases of larger domains containing allele-specific differences.

We conclude that allelic differences frequently alter regulatory processes that impact multiple transcription units across a locus. Potential mechanisms may include non-coding RNAs that act in a manner analogous to Xist in X-chromosome inactivation, differentially methylated regions involved in imprinting, or enhancers that regulate the expression of multiple transcripts simultaneously (Delaneau et al., 2019; Kumasaka et al., 2019; Rennie et al., 2018).

Widespread genetic changes in transcription initiation

Having validated our experimental dataset and examined the broad patterns of allelic differences across organs, we next set out to dissect the genetic basis for each stage of the Pol II life cycle. We first focused on defining allele specific patterns of transcription initiation. The 5' end of nascent RNA, denoted by the 5' end of paired-end ChRO-seq reads, marks the transcription start site (TSS) of that nascent RNA (Kwak et al., 2013; Tome et al., 2018). We identified TSSs in which at least 5 unique reads share the same 5' end inside of regions enriched for transcription initiation identified using dREG after merging all samples from the same tissue (Wang et al., 2018). Using a recently described hierarchical strategy (Tome et al., 2018), we grouped candidate TSSs into transcription start clusters (TSCs) and defined the maximal TSS as the position with the maximum 5' signal within each TSC (**Figure 2A; mid**). TSCs were grouped in turn into transcription start regions (TSRs), that are composed of multiple nearby TSCs based on the boundaries established by dREG (**Figure 2A; bottom**).

To verify that our pipeline identified TSSs accurately, even in the absence of enzymatic enrichment for capped RNAs, we examined whether candidate TSSs were enriched for the initiator DNA sequence element, a defining feature of Pol II initiation (Kaufmann and Smale, 1994; Smale and Baltimore, 1989). Most organs, especially brain and liver (which were the most deeply sequenced and are the focus of the analysis below unless otherwise specified), had high information content showing the initiator element at maxTSSs (**Figure 2—figure supplement 1A**). Moreover, the relationship between read depth and TSSs was similar to those recently reported in human cells (Tome et al., 2018) (**Figure 2—figure supplement 1B**). Finally, in K562

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

cells our pipeline identified TSSs that were supported by PRO-cap data better than existing human gene annotations (**Figure 2—figure supplement 1C**).

Allelic changes in TSSs were common, occuring in ~16-34% of TSCs tagged with SNPs (n = 1,109 - 5,793; binomial test FDR < 0.1; **Supplementary Table 3**). Changes in TSSs were highly enriched within strain effect domains identified by AlleleHMM (Chou and Danko, 2019) (odds ratios 3.5-5.4; p < 2.2e-16, Fisher's exact test). Therefore, many of these allelic changes in TSSs likely reflect allelic changes in the rates of transcription initiation on the gene secondary to allelic differences in transcription factor binding or other regulatory processes. These mechanisms have been explored extensively elsewhere (Battle et al., 2014; Chen et al., 2016; Lappalainen et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010). Throughout the remainder of this paper we focus on defining the effects of DNA sequence on core transcriptional processes.

We used genetic differences between alleles to define the relative strength of different initiator dinucleotides. The initiator motif is perhaps the best characterized feature of Pol II initiation and is most commonly characterized by a CA dinucleotide, but the sequence preferences of the initiator motif are weak and other dinucleotides are relatively common (Smale and Baltimore, 1989). We examined how changes between initiator dinucleotides affected the abundance of ChRO-seq reads. We identified the set of all max TSSs which had DNA sequence differences between B6 and CAST alleles. Our analysis revealed a hierarchy of initiator dinucleotides that impact initiation frequency with different magnitudes (**Figure 2B**). CA initiators which changed to TA had the lowest magnitude of shift in initiation frequency, whereas CA dinucleotides that change to CG have the largest magnitude. CA to TG changes were intermediate between CA to TA and CA to CG. The number of examples for CA to TG was much lower than other dinucleotide combinations because it required two DNA sequence changes. Therefore, we also examined TA to TG changes directly. This revealed that changes in initiator sequence from TA to TG were associated with a higher Pol II on the TA allele. Our results suggest a hierarchy of initiation frequency, in which Pol II prefers to initiate at a CA dinucleotide, followed by TA, TG, and finally CG.

Allelic changes in the shape of transcription initiation

We next asked how DNA sequence shapes which of the multiple independent initiator dinucleotides capture the majority of Pol II initiation events. We reasoned that cases in which Pol II initiation changed from one TSS to a nearby TSS would be highly informative about this initiation code. We therefore developed a statistical approach based on the Kolmogorov-Smirnov test to identify differences in the shape of the distribution of Pol II initiation (see Methods). Our approach identified 1,006 (brain) and 1,389 (liver) TSCs in which the shape of the 5' end of mapped reads within that TSC changed between alleles (FDR <= 0.1; Kolmogorov-Smirnov [KS] test) (see examples in **Fig. 3A-B**). We reasoned that changes in the shape of signal within TSCs may not always alter the total abundance of Pol II, as compensatory changes in the rates of Pol II initiation at nearby TSSs may compensate for one another. Consistent with this hypothesis, only ~10-20% of the TSCs identified using this approach were also found inside of strain effect domains, indicating that these changes in shape were fundamentally different from changes in initiation rates and reflected a different underlying biological process.

As TSCs typically span ~80 bp and are frequently comprised of multiple TSSs (Carninci et al., 2006; Tome et al., 2018), we divided changes in TSC shape into two classes: cases that

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

were driven predominantly by large changes in the abundance of Pol II at a single TSS position and cases in which multiple TSSs across the TSC contributed to changes in shape (see **Methods**). For example, the TSC giving rise to the transcription unit upstream and antisense to *Rps6kc1* had major differences in just one of the TSSs (**Figure 3A, arrow head**), whereas the promoter of *Zfp719* has multiple changes in TSSs within the same TSC (**Figure 3B, arrow heads**). In both examples, allelic changes result in differences in the position of the max TSS between alleles. Each of these two classes comprised approximately half of the changes in TSS shape in all organs (**Figure 3—figure supplement 1A**).

Next we examined the distribution of SNPs centered on allele specific max TSSs. To control for the ascertainment biases associated with having a tagged SNP in each allelic read, we compared the distribution of SNPs on allele specific max TSSs with a control set composed of TSSs that have no evidence of allele specificity. Single position driven allele-specific TSCs were associated with a strong, focal enrichment of SNPs around the max TSS (**Figure 3C** [liver] and **Figure 3—figure supplement 1B** [brain]). SNPs within 5 bp of the initiation site explain up to ~15-20% of single-base driven allele specific TSCs (**Figure 3E** and **Figure 3—figure supplement 1B**). This was predominantly explained by SNPs at the TSS itself, with an A highly enriched in the allele with highest max TSS usage at that position, consistent with the sequence preference of the initiator motif (**Figure 3D**). By contrast, the enrichment of C in the -1 position of the initiator motif was much weaker than the A at the 0 position. Although this result may partially be explained by a bias in which the C allele does not tag nascent RNA, it was consistent between organs (**Figure 3—figure supplement 1C**) and controlled based on the composition of the background set. Thus, our results suggest that the A in the initiation site may be the most important genetic determinant of transcription initiation, consistent with the Pol II initiation preference analysis conducted above (**Figure 2B**).

Multiple position driven allele specific TSCs had a weaker, broader enrichment of SNPs (**Figure 3E** [liver] and **Figure—figure supplement 1D** [brain], yellow shade indicates FDR <= 0.05; Fisher's exact test). Changes in TSC structure driven by multiple, separate TSSs, were enriched throughout the ~30 bp upstream, and to some extent downstream, potentially implicating changes in sequence specific transcription factors that influence the precise initiation site or changes that affect the binding of other components of the PIC (**Figure 3E** and **Figure 3—figure supplement 1D**).

Intriguingly, although the increased density of SNPs observed was largest near the Inr element, SNPs were enriched throughout the region occupied by the PIC in both single and multiple TSS driven allele specific TSCs (**Figure 3C,E**). Although the enrichment did not appear to be larger than surrounding DNA in the TATA box, the TATA box only occurs at ~10% of mammalian promoters (Carninci et al., 2006; Lenhard et al., 2012). Therefore, we conditioned on the presence of a clear TATA-like motif on at least one of the alleles and asked whether SNPs affecting the TATA box correlated with the magnitude of effect on initiation. We first used a TATA motif that had a general enrichment for AT content, consistent with the degenerate nature of the TATA box in mammals (see Methods). We found that SNPs which changed the TATA motif had a low, positive correlation with allele specificity, with a slightly stronger correlation in brain than in liver (Pearson's R = 0.09 [liver]; R = 0.18 [brain]). The positive correlation was marginally significant in the brain (p = 0.04), but not in the liver (p = 0.2). Similar results were also obtained with an additional TATA motif that was a stronger match to the classical TATA consensus

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

(TATAAA; p = 0.078 [liver]; p = 0.051 [brian]). These results are consistent with core promoter motifs playing an important role in the position and magnitude of transcription initiation, but they appear in our analysis to be weaker determinants of the precise initiation site than the initiator motif.

Next we examined other factors near the TSC, aside from DNA within the initiator motif or other known PIC components, that contributed to the position of TSSs. We noticed a higher frequency of A and T alleles downstream of the initiator motif on the allele with a higher max TSS (**Figure 3D**). We hypothesized that the lower free energy of base pairing in A and T alleles would make them easier to melt during initiation, and could therefore increase the frequency of TSS usage at these positions. Indeed, a more direct examination of AT content in 5 bp windows near the maxTSS identified a significantly higher AT content on the allele with the highest max TSS after masking DNA at positions -1 and 0 to avoid confounding effects of the initiator element (**Figure 3F** and **Figure 3—figure supplement 2E**). This enrichment of high AT content was consistent in both brain and liver tissue, but was unique to single TSS driven max TSSs (**Figure 3—figure supplement 1F-G**).

We conclude that multiple aspects of DNA sequence, including both sequence motif composition (especially the initiator element) and the energetics of DNA melting, influence TSS choice in mammalian cells.

Models of stochastic search during transcription initiation

Next we examined how SNPs that affect a particular TSS impact initiation within the rest of the TSC. In the prevailing model of transcription initiation in *S. cerevisiae*, after DNA is melted, Pol II scans by forward translocation until it identifies a position that is energetically favorable for transcription initiation (Braberg et al., 2013; Giardina and Lis, 1993; Kaplan et al., 2012; Kuehner and Brow, 2006; Qiu et al., 2020) (**Figure 4A, left**). In mammals, Pol II is not believed to scan, but rather each TSS is believed to be controlled by a separate PIC (Luse et al., 2020) (**Figure 4A, right**). We considered how mutations in a strong initiator dinucleotide (CA) would affect transcription initiation under each model. Under the yeast model, we expected CA mutations to shift initiation to the next valid initiator element downstream. Under the mammalian model, we expected each TSS to be independent and therefore a mutation in the TSS would have no effect on the pattern of nearby initiation sites.

We analyzed 277 and 372 TSSs in brain and liver, respectively, where the high allele contained a CA dinucleotide while the other allele did not. Candidate initiator motifs within 20 bp of the CA/ non-CA initiation site had slightly more initiation signal on the non-CA allele compared with the CA allele, consistent with the shooting gallery model but inconsistent with the prevailing model of independent TSSs expected in mammals (**Figure 4B; purple**). By contrast, TSSs where both alleles contained a CA dinucleotide (n = 8,147 [brain] and 10,113 [liver]) did not show this same effect (**Figure 4B; gray**). The difference was found for both adjacent CA dinucleotides and for weaker candidate (Py)(Pu) initiator elements, and was consistent across both single-base and multiple-base TSC configurations (**Figure 4—figure supplement 1**). These results appear more consistent with the yeast, rather than the current mammalian models, despite the fact that no scanning mechanism has been reported to date in mammals.

Unexpectedly, we also observed consistently higher initiation signals on the non-CA allele both upstream and downstream of the initiator dinucleotide (**Figure 4C**). The signal for an

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

increase on the non-CA allele stretched up to 20bp both upstream and downstream of the CA/ non-CA dinucleotide. For instance, a SNP in the initiator element nearly abolished the dominant max TSS of the protein coding gene *Smg9* in CAST (**Figure 4D, arrow**). Instead, initiation in CAST shifted to a new maxTSS located upstream (**Figure 4D, arrow head**), and also increased usage of several minor TSSs downstream (**Figure 4D**). This redistribution in both directions cannot be explained by the yeast uni-directional scanning model. Instead, we propose that after DNA melting and Pol II assembly on the template strand, Pol II rapidly and stochastically moves in both directions along the template strand scanning for an energetically favorable initiator dinucleotide in a process resembling brownian motion (**Figure 4E**).

We also considered an alternative interpretation, that DNA sequence preference in the initiator dinucleotide feeds back and affects the stability of the pre-initiation complex. Under this model, we expect that DNA sequence changes in a TSS would result in the redistribution of initiation signal to all of the potential TSSs within a TSR (assuming no change in the local concentration of components of Pol II, the pre-initiation complex, or other core transcription factors). To examine this model, we asked whether the redistribution in initiation signal observed near mutant TSSs was also found in all TSCs within the TSR. We found that the increased initiation signal was limited to within ~20bp of the CA/ non-CA initiation site (**Figure 4F**). In our view, this result supports the model proposed above, in which Pol II recruited by a single pre-initiation complex initiates at an energetically favorable initiator dinucleotide within a confined region of ~10-20 bp.

Correspondence and disconnect between allele specific TSS and pause position

We next examined allelic changes in the position of paused Pol II. To measure the position of the Pol II active site with single nucleotide precision, we prepared new ChRO-seq libraries in three organs (heart, skeletal muscle, and kidney from two female mice). New libraries were paired-end sequenced to identify the transcription start site and active site of the same molecule (Tome et al., 2018). New libraries clustered with those generated previously from the same organ (**Fig 5—figure supplement 1**). Using the same pipeline we developed for transcription initiation, we identified regions enriched for transcription initiation and pausing, and validated that candidate maxTSSs were enriched for the initiator motif (**Figure 5—figure supplement 2A**). Our analysis identified 2,260 dREG sites with candidate changes in the shape of paused Pol II, assessed using the position of the Pol II active site, defined as the 3' end of RNA insert (FDR <= 0.1; Kolmogorov-Smirnov [KS] test; see **Methods**).

Previous work has shown a tight correspondence between the site of transcription initiation and pausing, with pausing occurring predominantly in the window 20-60 bp downstream of the TSS (Tome et al., 2018). As expected, allelic changes in the position of paused Pol II were often coincident with changes in transcription initiation (**Figure 5A**), particularly when the changes were large. In addition to the main component of correlation between initiation and pausing, however, we also identified changes in both pause and initiation that were independent of the other step in the transcription cycle. In at least 111 cases (~31% of sites where paused Pol II changed shape, and the single RNA molecule was tagged by a SNP), the position of the pause was identical between alleles, but the position of the max TSS that initiated the paused Pol II changed by 1-32 bp (**Figure 5B, top**). Thus, in many cases where Pol II paused at the same position in both alleles, the RNA molecules were initiated at distinct TSSs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

More commonly (n = 269; ~52% of tagged RNA molecules with putative changes in pause shape), changes in Pol II pausing occurred between alleles despite being initiated from identical TSSs (**Figure 5B, bottom**). Although more frequent, changes in the position of paused Pol II that shared the same TSS were slightly smaller in magnitude, typically <10 bp. These cases in which the Pol II pause site changes without an accompanying change in the position of transcription initiation suggest the existence of a DNA sequence code that influences Pol II pausing.

DNA sequence determinants of promoter proximal pause position

To understand the genetic determinants of pausing, we analyzed cases in which the same TSS had different maximal pause sites in the CAST and B6 alleles. As tagged SNPs in the window between the initiation and pause site were relatively rare, we increased our statistical power by analyzing all three of the organs together after removing duplicate initiation sites when they overlapped. After filtering, we identified 269 candidate positions in which the same TSS gave rise to separate pause distributions on the B6 and CAST alleles. As a control, we used 1,396 TSS/ pause pairs that were tagged by SNPs but did not have allelic changes in the pause position.

We first examined how short insertions or deletions affect the position of paused Pol II. Paused Pol II is positioned in part through physical constraints with TFIID, a core component of the pre-initiation complex (Fant et al., 2020). As a result of such connections with the pre-initiation complex, short insertions or deletions affecting the distance between the TSS and the maximal pause site altered the frequency of pausing at a model gene (D. melanogaster *HSP70*; (Kwak et al., 2013)). In our dataset, changes in pausing were highly enriched for small insertions and deletions between the maxTSS and pause site (n = 56 (21%); expected = 22 (8%); p < 1e-5, Fisher's exact test). Changes in pause position were correlated with changes in the size of the insertion or deletion, such that the number of bases between the max TSS and the pause was typically less than 5 bp (**Figure 5C**). Although this result may be influenced by fragment length bias introduced during sequencing, it nevertheless provides additional support for a model in which paused Pol II is placed in part through physical constraints with the pre-initiation complex (Fant et al., 2020; Kwak et al., 2013).

Next we identified single nucleotide changes that affect the position of the pause site. Previous studies have defined a C nucleotide at the paused Pol II active site (Gressel et al., 2017; Tome et al., 2018), which we recovered by generating sequence logos of max pause positions in our three murine organs (**Figure 5D, bottom**). Additionally, we also observed a G in the +1 position immediately after the pause, and a G/A-rich stretch in the 10 bp upstream of the pause site that lies within the transcription bubble (**Figure 5D, bottom**). The 10 bp upstream of the pause position had a higher G content and a lower C content than the two surrounding windows (**Figure 5E**). Thus, our data show that Pol II pauses on the C position immediately after a G-rich stretch.

Allelic differences at the RNA polymerase active site (usually a C) had the strongest association with the pause, followed by SNPs at the -2 and -3 position relative to the pause (usually a G; **Figure 5D; Figure 5—figure supplement 2B**). We also noted enrichment of SNPs downstream of the pause, especially in the +1 position, although the number of SNPs supporting these positions were small. We also noted that multiple independent SNPs were frequently found in the same TSS/ pause pair (observed n = 10 (3.7%); expected = 1 (0.36%); p < 2e-5; Fisher's Exact Test; **Figure 5F**), suggesting that multiple changes in the weak DNA sequence motifs associated with pausing were more likely to affect the position of paused Pol II. Collectively, indels

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

and SNPs identified as enriched in the analysis above explained 49% of allele specific differences in the pause position (**Figure 5F**). Thus, the DNA sequence determinants of pause position are largely found either within the pause site, the transcription bubble of paused Pol II, or insertions or deletions between the pause and initiation site.

Pol II pause position is driven by the first energetically favorable pause site

We extended our analysis of allelic differences in pausing to determine how multiple candidate pause positions early in a transcription unit collectively influence the position of paused Pol II. As in the analysis above, we focused on the set of allelic differences in pause shape in which the two alleles had a distinct maximal pause position (n = 269). By definition, these transcription units had different maximal pause positions on the two alleles: on one allele the distance between the TSS and the pause position is shorter (which we call the 'short allele'), and on the other the distance between the TSS and the pause is longer ('long allele') (see cartoon in **Figure 5G, middle**). We found that the DNA sequence motif near the long pause position was similar on both alleles, recapitulating the C at the pause site and an enrichment of Gs in the transcription bubble (**Figure 5G, right**). By contrast, DNA sequence changes affecting pause position occurred at the short pause position on the long allele (**Figure 5G; bottom left**).

Our findings suggest a model in which single nucleotide changes that alter the free energy of the pause complex result in Pol II slipping to the next available position downstream. In favor of this model, when there was a SNP in the active site at the short pause, the max pause position moved downstream by <10bp, a relatively small amount compared to all changes in pause shape (**Figure 5H**). By contrast, indels between the initiation and short pause site tended to have a larger effect on the allelic difference between pause positions (**Figure 5I**). These observations support a model in which DNA sequence changes that disfavor pausing result in Pol II slipping downstream to the next pause site for which DNA sequence is energetically favorable.

Allelic changes in gene length caused by genetic differences in Pol II termination

Blocks with allelic bias identified by AlleleHMM, were frequently found near the 3' end of genes. We hypothesized that these blocks reflected allelic differences in the position of Pol II termination that altered the length of primary transcription units. For example, *Fam207a* had an excess of reads on the CAST allele without a new initiation site that could explain allelic differences (**Figure 6A**). We set out to identify protein-coding transcription units that have an allelic difference in the abundance of Pol II only at the 3' (and not the 5') end.

To approximate the boundaries of allelic differences in Pol II abundance at the 3' end of genes, henceforth called the allelic termination window, we used AlleleHMM blocks that begin inside of a transcription unit and end near or after the end of the same transcription unit. We identified 317-931 candidate allelic termination windows in each of the eight organs (total *n* = 3,450). Allelic termination windows varied in size between 1kb and 100kb, with a median size just under 10kb (**Figure 6B**; **Figure 6—figure supplement 1A**). Although the longest allelic termination windows likely reflect allelic changes in multiple genes, the median size is approximately consistent with the reported length of transcription past the polyadenylation cleavage site (Grosso et al., 2012). Several lines of evidence suggest that these allelic differences were enriched for bona-fide differences in the site of Pol II termination: First, the majority did not start near dREG sites (the start is marked by a triangle in **Figure 6A**), and second, the allele with

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

higher expression tended to have a similar ChRO-seq signal as the primary gene (**Figure 6—figure supplement 1B**).

Allelic differences in termination were consistent between different organs. For example, *Fam207a* had approximately the same allelic termination window in brain and liver (**Figure 6A**). To visualize the boundaries of allelic termination windows across larger numbers of transcripts, we used heat maps centered at the start of the allelic increase in expression (marked by a triangle) and sorted by the length of the allelic termination window (**Figure 6C**). Heatmaps showed a higher abundance of ChRO-seq reads in the allelic termination window on the allele with higher expression in the transcript body, as expected. Heat maps from brain and spleen using the same order as liver recovered similar patterns of allelic termination (**Figure 6C**). We conclude that allelic differences in termination were driven predominantly by DNA sequence.

Next we examined how allelic termination windows were associated with changes in nearby transcription. Allelic termination windows were more likely to occur in highly expressed transcripts, which may partially reflect increased statistical power for detecting changes supported by larger numbers of reads. Additionally, allelic termination windows were more likely to have allelic changes in the transcription level of an adjacent transcript compared with matched transcripts without an allelic termination window (liver: odds ratio = 1.65, p = 1.83e-15; brain: odds ratio = 1.23, p = 0.018; Fisher's exact test). We tested whether the association between allelic termination and adjacent transcript expression was also found when allelic termination and the adjacent transcript were encoded on opposite strands, hence avoiding the interpretation that AlleleHMM was more likely to detect allelic termination windows near an allelic difference with a large magnitude. The enrichments we observed were still significant in this more restrictive test (liver: odds ratio = 1.33, p = 1.32e-05; brain: odds ratio = 1.16 , p = 0.09; Fisher's exact test). Intriguingly, the expression of nearby transcription units was frequently highest on the allele that terminated early (liver: odds ratio = 4.97, p = 2.2e-16; brain: odds ratio = 9.67, p = 2.2e-16; Fisher's exact test). Taken together, these results suggest an association between the length of post-poly-A transcription and the expression of nearby transcripts.

Allelic changes in termination correlate with differences in mRNA stability

Allelic differences in termination may influence the primary structure (i.e., the RNA sequence) of the mature mRNA by affecting polyadenylation cleavage site or splice site use between the two alleles (Mittleman et al., 2021, 2020). To identify allelic differences in mRNA primary structure we sequenced poly-A enriched mRNA from two liver and brain samples. We identified multiple examples in which allelic termination differences showed clear evidence of differential use of candidate exons in the mRNA (**Figure 6—figure supplement 2A**), indicating underlying differences in mRNA primary structure. In some cases (e.g., *Sh3rf3*, **Figure 6—figure supplement 2A**), we identified unannotated 3' exons with differential inclusion in the mature mRNA between CAST and B6 alleles. A global analysis of the mRNA-seq data using AlleleHMM determined that 21-41% of transcription units with differences in allelic termination had clear evidence of changes in mRNA primary structure (**Figure 6D**), representing an enrichment of 2-4-fold relative to transcription units with no evidence of allelic termination ($p < 2.2e\text{-}16$; Fisher's exact test in both brain and liver). This indicates that changes in allelic termination often had a corresponding impact on the primary structure of the mature mRNA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Next, we examined whether changes in the mature mRNA primary structure affected the stability of the mature mRNA. We estimated mRNA stability in liver and brain using the ratio of signal in mRNA-seq and ChRO-seq data, which is correlated with the degradation rate of mature mRNA (Blumberg et al., 2021). We asked whether transcription units with differences in allelic termination which have differences in RNA primary structure have larger differences in mRNA stability between alleles compared with transcripts which do not have evidence of differences in RNA primary structure. We found that allelic changes in mRNA primary structure were more likely to have increased differences in mRNA stability between B6 and CAST alleles (Brain p = 4.28e-4, liver p = 2.69e-08; KS test; **Figure 6—figure supplement 2B-C**). Thus differences in allelic termination frequently alter the primary structure of the mature mRNA and in some cases impact mRNA stability.

Discussion

Despite much progress, we still have a relatively limited understanding of how DNA sequence influences the process of transcription initiation, pause, elongation, and termination by RNA Pol II. Our limited understanding reflects, in part, the fact that DNA sequence elements that influence transcription are highly degenerate and frequently spread across wide genomic regions, making them very difficult to identify and characterize. Here, we used naturally occurring genetic variation in F1 hybrids of two highly divergent mouse strains, CAST and B6. We generated and analyzed ChRO-seq data, which maps the location and orientation of RNA polymerase, in eight murine organs. This data provides a window into how short SNPs and Indels influence the position of Pol II that is robust to technical variation, providing new insights into Pol II dynamics.

We characterized the effects of DNA sequence variation on the position of the TSS during Pol II initiation. DNA sequence differences within ~5-10bp of the TSS were the most common determinant of the TSS. As expected, the initiator element plays a central role. The A base in the initiator motif, in particular, was the most important determinant of the TSS. Changes in the cytosine nucleotide at the -1 position had surprisingly little evidence for differences in our analysis, consistent with a hierarchy of transcription initiation that is more dependent on the A in a CA initiator dinucleotide. Unexpectedly, we found that AT nucleotides surrounding the initiation site were positively correlated with the use of that TSS between alleles. We speculate that AT content influences the TSS because AT-rich DNA requires lower free energy to melt (Breslauer et al., 1986). Thus, our study reveals both known and novel DNA sequence elements surrounding the TSS that influence which of the numerous potential TSSs within the TSC are selected for initiation.

Our results have led us to propose a new model of transcription initiation in which Pol II selects an initiation site through random motion on DNA resembling brownian motion. In the prevailing model of initiation in yeast, DNA is melted and Pol II scans by forward translocation until it identifies an energetically favorable TSS (Braberg et al., 2013; Kaplan et al., 2012; Qiu et al., 2020). Mammals are not believed to scan, but rather independent PICs are believed to support initiation from a narrow window (Luse et al., 2020). Although our study does support a role for DNA sequence motifs which control the location of the PIC in selecting the TSS, PIC motifs such as the TATA box were relatively less important for specifying the exact position of the TSS. Intriguingly, DNA sequence changes in the initiator element increased the use of nearby TSSs

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

both upstream and downstream in a window of ~20bp. We think the most straightforward interpretation of these results is that Pol II samples candidate initiation sites in both directions by a process resembling a one-dimensional brownian motion along the DNA template.

Another possible interpretation is that changes in the CA dinucleotide sequence alter DNA elements that support partially overlapping PICs (e.g., CA -> TA dinucleotide change makes the sequence slightly closer to a TATA box that could support initiation further downstream). Although we do not directly rule out this alternative interpretation, in our view it seems less likely because the effects we observe are constrained within a fairly narrow window, because the effects we observed are so consistent when conditioning on changes in the initiator DNA sequence motif, and because DNA sequence changes in the TATA box did not have as large of an impact on the initiation site. In either case, however, it is clear from our data that changes in the DNA sequence of initiator elements tend to increase the use of candidate initiators nearby.

By analyzing allelic changes in pause position from the same initiation site, we have learned much about how DNA sequence influences the precise coordinates of promoter proximal pause. Our results show that the pause position occurs at a C nucleotide downstream of a G-rich stretch, similar to motifs enriched at pause positions in prior studies (Gressel et al., 2017; Tome et al., 2018). As cytosine is the least abundant ribonucleotide (Traut, 1994), previous authors proposed it is the slowest to incorporate into nascent RNA, explaining its association with the Pol II pause (Tome et al., 2018). In addition, we also identified a G-rich stretch that coincides with the position of the transcription bubble, as well as a guanine nucleotide just downstream of the pause position. The enrichment of G nucleotides in the transcription bubble has a higher stability of RNA-DNA hybridization without using a cytosine nucleotide, and may serve to stabilize the RNA-DNA hybrid within the transcription bubble while Pol II remains paused.

We also noted widespread allelic differences in the site of transcription termination, which resulted in substantial differences in the length of primary transcription units between alleles. Allelic differences in transcription termination were largely similar between different organs, implicating DNA sequence differences as the major determinants of transcription termination. Although the majority (60-80%) of allelic differences in termination did not affect the primary structure (i.e., the set of all ribonucleotides in the mRNA), they often did have an impact. Moreover, when the primary structure of the mRNA was affected by allelic differences in Pol II termination, we found evidence that the transcript stability of the resulting mRNA was also affected.

Our study used an F1 hybrid cross between CAST and B6, which we contend is particularly well suited to problems in which the DNA sequence determinants are weak and spread across a large genomic region. Experimental batch variation is a major confounder in all genomic analyses. However, batch effects are shared between the B6 and CAST alleles, which are processed from the same nucleus and undergo exactly the same experimental processing and sequencing steps. Another advantage of using a F1 hybrid system is that by comparing the effect of SNPs or indels on otherwise very similar DNA sequences, we can reduce artifacts from larger changes in DNA sequence composition on the sequencing library itself. Intuitively, DNA sequence changes affecting the middle of an RNA insert, especially those which constitute only a single base difference between alleles, are unlikely to have a large impact on detection in ChRO-seq. As a result, we can be confident in differences observed in our analysis, even in cases where the differences between the CAST and B6 alleles were relatively small in magnitude.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

The use of an F1 hybrid system does, of course, have a number of limitations as well. Any DNA sequence variation between alleles that affects the probability of detection, including either DNA sequence variation or size bias in the RNA insert, could impact detection differently between alleles. One place where this might occur is ligation bias for SNPs on the 5' or 3' end of an RNA insert. However, we think the effect of such sequence bias is relatively small in our study. Results that might be affected, including the A base in the initiator element (at the 5' end of the RNA insert), and the C base in the Pol II active site (at the 3' end of the RNA insert), are supported by existing literature (Kaufmann and Smale, 1994; Kuehner and Brow, 2006; Smale and Baltimore, 1989; Tome et al., 2018). Major new results reported here generally reflect SNPs inside of the RNA insert, making differences in detection a less likely explanation.

In summary, our study dissects how DNA sequences impact steps during the Pol II transcription cycle. The dynamics of Pol II on each position of the genome can influence the rate of mRNA production and may impact organism phenotypes. Our work is a first step in understanding how DNA sequences influence stages in the Pol II transcriptional cycle and impact phenotypes in humans and other animals.

Methods

Experimental Methods:

Mouse experiments: The mice used in this study were reciprocal F1 hybrids of the strains C57BL/6J and CAST/EiJ. All mice were bred at Cornell University from founders acquired from the Jackson Laboratory. All mice were housed under strictly controlled conditions of temperature and light:day cycles, with food and water *ad libitum*. All mouse studies were conducted with prior approval by the Cornell Institutional Animal Care and Use Committee, under protocol 2004-0063.

Tissue collection: Mice were euthanized at 22 to 25 days of age by CO_2, followed by cervical dislocation. All mice were euthanized between 10 a.m. and 12 p.m., immediately after removal from their home cage. Whole brain, eye, liver, stomach, large intestine, heart, skeletal muscle, kidney, and spleen were rapidly dissected and snap frozen in dry ice.

*mRNA isolation/ RNA-seq library prep:*RNA was extracted from the brain and liver of a male and a female mouse (both 22d of age). Tissue samples were frozen using liquid nitrogen and pulverized using a mallet and a mortar. 100mg of each tissue was used for a TRIzol RNA extraction. Briefly, 1 mL of TRizol was added to each sample, chloroform was used for phase separation of the aqueous phase containing RNA, RNA was precipitated using isopropanol and washed with 75% ethanol. A total of 400 ng of RNA was input into the RNA-seq library prep. Poly-A containing mRNA was enriched for 2 rounds using the NEBNext Poly(A) mRNA Magnetic Isolation Module. Stranded mRNA-seq libraries were prepared by the Cornell TREx facility using the NEBNext Ultra II Directional RNA Library Prep Kit. Libraries were sequenced using an Illumina NextSeq500.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Chromatin isolation: Chromatin was isolated and ChRO-seq libraries were prepared following the methods introduced in our recent publication (Chu et al., 2018). Briefly, tissue was cryo-pulverized using a cell crusher (http://cellcrusher.com). Tissue fragments were resuspended in NUN buffer (0.3 M NaCl, 1 M Urea, 1% NP-40, 20 mM HEPES, pH 7.5, 7.5 mM MgCl2, 0.2 mM EDTA, 1 mM DTT, 20 units per ml SUPERase In Rnase Inhibitor(Life Technologies, AM2694), 1× Protease Inhibitor Cocktail (Roche, 11 873 580 001)).Samples were vortexed vigorously before the samples were centrifuged at 12,500 x g for 30 min at 4C. The NUN buffer was removed and the chromatin pellet washed with 1 mL 50 mM Tris-HCl, pH 7.5 supplemented with 40 units of RNase inhibitor. Samples were centrifuged at 10,000 x g for 5 minutes at 4C and the supernatant discarded. Chromatin pellets were resuspended in storage buffer (50 mM Tris-HCl, pH 8.0, 25% glycerol, 5 mM Mg(CH3COO)2, 0.1 mM EDTA, 5 mM DTT, 40 units per ml Rnase inhibitor) using a Bioruptor sonicator. The Bioruptor was used following instructions from the manufacturer, with the power set to high, a cycle time of 10 min (30s on and 30s off). The sonication was repeated up to three times if necessary to resuspend the chromatin pellet. Samples were stored at -80C.

ChRO-seq library preparation: ChRO-seq libraries were prepared following a recent protocol (Dig Bijay Mahat et al., 2016). We prepared some libraries to achieve single nucleotide resolution for the Pol II active site. In these cases, the chromatin pellet was incubated with 2x run-on buffer (10 mM Tris-HCl, pH 8.0, 5 mM MgCl2,1 mM DTT, 300 uM KCl, 20 uM Biotin-11-ATP (Perkin Elmer, NEL544001EA), 200 uM Biotin-11-CTP (Perkin Elmer, NEL542001EA), 20 uM Biotin-11-GTP (Perkin Elmer, NEL545001EA), 200 uM Biotin-11-UTP (Perkin Elmer, NEL543001EA)) for 5 minutes at 37 C. In some libraries we modified the run-on buffer to extend the length of reads for more accurate allelic mapping at the expense of single nucleotide resolution for the Pol II active site. In these cases, the run-on reaction was performed using a different ribonucleotide composition in the nuclear run-on buffer (10 mM Tris-HCl, pH 8.0, 5 mM MgCl2,1 mM DTT, 300 mM KCl, 200 μ M ATP (New England Biolabs (NEB), N0450S), 200 μ M UTP, 0.4 μ M CTP, 20 μ M Biotin-11-CTP (Perkin Elmer, NEL542001EA), 200 μ M GTP (NEB, N0450S)).The run-on reaction was stopped by adding Trizol LS (Life Technologies, 10296-010) and RNA was pelleted with the addition of GlycoBlue (Ambion, AM9515) to visualize the RNA. RNA pellet was resuspended in diethylpyrocarbonate (DEPC)-treated water. RNA was heat denatured at 65 C for 40 s to remove secondary structure. RNA was fragmented using base hydrolysis (0.2N NaOH on ice for 4 min). RNA was purified using streptavidin beads (NEB, S1421S) and removed from beads using Trizol (Life Technologies, 15596-026). We ligated a 3′ adapter ligation using T4 RNA Ligase 1 (NEB, M0204L). We performed a second bead binding followed by a 5′ decapping with RppH (NEB, M0356S). RNA was phosphorylated on the 5’ end using T4 polynucleotide kinase (NEB, M0201L) then ligated onto a 5’ adapter. A third bead binding was then performed. The RNA was then reverse transcribed using Superscript III Reverse Transcriptase (Life Technologies, 18080-044) and amplified using Q5 High-Fidelity DNA Polymerase (NEB, M0491L) to generate the ChRO-seq libraries. Libraries were sequenced using an Illumina HiSeq by Novogene. All adapter sequences and barcodes used for each sample are depicted in **Supplementary Table 4**.

Data analysis:

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Read mapping, transcription start site, and transcription unit discovery:

Processing and mapping ChRO-seq reads: Paired-end reads with single nucleotide precision were processed and aligned to the reference genome (mm10) with the proseq2.0 (https://github.com/Danko-Lab/proseq2.0). Libraries in which we tailored the run-on to extend the length of reads were pre-processed, demultiplexed, and aligned to the reference genome (mm10) with the proseqHT_multiple_adapters_sequencial.bsh. AlleleDB (Chen et al., 2016; Rozowsky et al., 2011) align the R1 reads to the individual B6 and Cast genomes. In brief, the adaptor sequences were trimmed with the cutadapt, then PCR duplicates were removed using unique molecular identifiers (UMIs) in the sequencing adapters with prinseq-lite.pl (Schmieder and Edwards, 2011). The processed reads were then aligned with BWA (mm10) in analyses not using individual genome sequences (Li and Durbin, 2009), or with bowtie (Langmead et al., 2009) as input for AlleleDB. When bowtie was used, we selected either the R1 or R2 files for alignment for analyses requiring either the 5' or 3' end of the RNA insert. All scripts for mapping can be found publicly at:
https://github.com/Danko-Lab/F1_8Organs/blob/main/00_F1_Tissues_proseq_pipeline.bash
https://github.com/Danko-Lab/utils/blob/master/proseq_HT/proseqHT_multiple_adapters_sequencial.bsh

Processing and mapping RNA-seq reads: We used STAR (Dobin et al., 2013) to align the RNA-seq reads. To avoid bias toward the B6 genome, we did not use any gene annotations for mapping, but used the list of splicing junctions generated by STAR. Mapping was performed in three stages: First, reads were first mapped without annotation and STAR generated a list of splicing junctions (sj1) from the data. Second, to identify potential allele specific splicing junctions, we performed allele specific mapping using STAR which takes as input a VCF file denoting SNPs differentiating Cast and B6, using the initial splice junction list (sj1). This personalized mapping was used to generate a more complete list of splice junctions (sj2). Third, we identified allele specific alignments by using the WASP option provided by STAR (van de Geijn et al., 2015). In this final mapping, we used the splice junction list (sj2) and a VCF file. This procedure generated a tagged SAM file (vW tag) providing the coordinates of allele specific alignments and their mapping position. Scripts can be found here: https://github.com/Danko-Lab/F1_8Organs/blob/main/termination/F1_RNAseq_forManuscript.sh

dREG: For each organ, we merged all reads from each replicate and cross to increase the power of dREG. BigWig files representing mapping coordinates to the mm10 reference genome were uploaded to the dREG web server at http://dreg.dnasequence.org (Wang et al., 2018). All of the output files were downloaded and used in subsequent data analysis. Scripts used to generate the BigWig files can be found at :
https://github.com/Danko-Lab/F1_8Organs/blob/main/F1_TSN_Generate_BigWig.sh

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Transcript unit prediction using tunits: We used the tunit software to predict the boundaries of transcription units *de novo* (Danko et al., 2018). We used the 5 state hidden Markov model (HMM), representing background, initiation, pause, body, and after polyadenylation cleavage site decay from tunits. To improve sensitivity for transcription unit discovery in each tissue, the input to tunits was the output of dREG and bigWig files that were merged across all replicates and crosses. Scripts can be found here: https://github.com/Danko-Lab/F1_8Organs/blob/main/Tunit_predict_manuscript.sh
https://github.com/Danko-Lab/F1_8Organs/blob/main/run.hmm.h5_F1bedgraph.R
https://github.com/Danko-Lab/F1_8Organs/blob/main/hmm.prototypes.R

Clustering: We used all transcripts that are 10,000 bp long from GENCODE vM25. Only reads mapped to the gene body (500bp downstream of the start of the annotation to the end of the annotation) were used. We filtered the transcripts and only kept those with at least 5 mapped reads in every sample. We export rpkm normalized expression estimates of each transcript. Morpheus was used to calculate and plot Spearman's rank correlation (https://software.broadinstitute.org/morpheus) with the following parameters: Metric = Spearman rank correlation, Linkage method = Average Linkage, distance.function.name= Spearman rank correlation. Scripts can be found here:
https://github.com/Danko-Lab/F1_8Organs/blob/main/getCounts_skipfirst500.R

Allele specific heatmaps in Fig. 1 C focused on all genes that are longer than 10,000 bp from GENCODE vM25 and were allele specific in at least 3 of the samples from each organ. Chromosomes M, X, and Y were excluded from the analysis. We computed the log-2 ratio of reads mapping uniquely to the B6 and CAST allele for each gene. Log-2 ratios were used as the input to Morpheus. Rows (representing genes) were ordered by 1 - Pearson's correlation across all samples. Columns were ordered manually. Organs used the same order as in the total gene expression clustering, above. Samples were ordered based on the direction of cross so that imprinted genes could easily be distinguished from strain-effect genes.

Testing positional correlation between transcripts: We asked whether adjacent transcription units shared the same allele specificity more frequently than chance. We identified transcription units that were allele specific based on a false discovery rate corrected binomial test cutoff less than 0.1. As a control set, we used all transcription units regardless of allelic bias. For each transcription unit in the allele specific and control set, we identified the number with a FDR corrected binomial test for allele specificity <0.1 (representing allele specific) or >0.9 (representing confident non-allele specific). The differences between groups were tested using a Fisher's exact test.

AlleleHMM: Maternal- and paternal- specific reads mapped using AlleleDB were used as input to AlleleHMM (Chou and Danko, 2019). We combined biological replicates from the same organ and cross, and used the allele-specific read counts as input to AlleleHMM. AlleleHMM blocks were compared with GENOCODE gene annotations to pick the free parameter, τ, which maximized sensitivity and specificity for computing entire gene annotations, as described (Chou

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

and Danko, 2019). Most organs used a τ of either 1E-5 (brain, liver, spleen, skeletal muscle) or 1E-4 (heart, large intestine, kidney, and stomach). As reported, the primary parameter that influenced τ was the library sequencing depth. AlleleHMM scripts can be found here:
https://github.com/Danko-Lab/AlleleHMM
https://github.com/Danko-Lab/F1_8Organs/blob/main/01_F1Ts_AlleleHMM.bsh

Discovering strain effect and imprinted domains: We used the following rules to merge nearby allele specific transcription events into starin effect or imprinted domains:

1. Identify candidate AlleleHMM blocks using pooled ChRO-seq reads from samples with the same organ and same cross direction.
2. Combine blocks above from the same organ (but different crosses). Combine p-values using Fisher's method for all biological replicates within the same tissue and cross direction. Keep blocks that are biased in the same direction with a Fisher's p-value <= 0.05.
3. Determine whether the blocks are under a strain effect (allelic biased to the same strain in reciprocal crosses) or parent-of-origin imprinted effect (allelic biased to the same parent in reciprocal crosses).
4. Merge overlapping strain effect blocks from different organs into strain effect domains; merge overlapping strain effects from the same imprinted blocks into imprinted domains.

Scripts implementing these rules can be found here: https://github.com/Danko-Lab/F1_8Organs/blob/main/Find_consistent_blocks_v3.bsh

After discovering blocks, we examined the number of gene annotations in each domain (Fig. 1E). We used GENCODE annotated genes (vM25). We kept all gene annotations and merged those which overlapped or bookended (directly adjacent to, as defined by bedTools) on the same strand so that they were counted once. All operations were performed using bedTools (Quinlan and Hall, 2010). Scripts can be found here:
https://github.com/Danko-Lab/F1_8Organs/blob/main/Find_consistent_blocks_v3.bsh
https://github.com/Danko-Lab/F1_8Organs/blob/main/Imprinted_figures.R

Determining the allelic bias state of annotated genes: We used GENCODE gene annotations representing protein-coding genes (vM20) in which the transcription start site overlapped a site identified using dREG (Wang et al., 2018). We used de novo annotations by the *tunits* package to identify unannotated transcription units, which do not overlap an annotated, active gene as a source of candidate transcribed non-coding RNAs. Transcription units from both sources were merged for downstream analysis. We determine if the gene/ncRNA are allelic biased by comparing mapped reads to the B6 and CAST genomes using a binomial test, retaining transcription units with a 10% false discovery rate (FDR). Allele specific transcription units were classified as being under a strain effect (allelic biased to the same strain in reciprocal crosses) or parent-of-origin imprinted effect (allelic biased to the same parent in reciprocal crosses). Scripts can be found: https://github.com/Danko-Lab/F1_8Organs/blob/main/Genetics_or_imprinting_v2.bsh

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Evaluate the contribution of false negatives to organ-specific allelic bias in organ-specific allelic biased domains (OSAB domain): In Supp Fig. 1A and B, we asked whether organs in which we did not identify allelic bias were false negatives. To do this we compared distributions of the transcription level in putatively unbiased organs. For each OSAB domain identified in at least one, but not in all organs, we examined the effect size of allelic bias in the organ with the highest expression that is putatively unbiased. We defined the effect size of allelic bias as the ratio between maternal and paternal reads in the candidate OSAB domain. If the allelic-biased organ was maternally biased, the effect size was calculated as maternal reads divided by paternal reads in the blocks, otherwise the effect size was calculated as paternal reads divided by maternal reads in the blocks. Scripts implementing this can be found here:
https://github.com/Danko-Lab/F1_8Organs/blob/main/AllelicBiase_expressionLevel.bsh
https://github.com/Danko-Lab/F1_8Organs/blob/main/getNonBiasedHighest_Biased_AllelicBiaseDistribution.R

Evaluate the contribution of expression to organ-specific allelic biased domains (OSAB domain): In Supp Fig. 1C, we asked whether OSAB domains were not actively transcribed in candidate unbiased organs. Using bedtools and in-house scripts, we calculated the rpkm (Reads per kilobase per million mapped reads) normalized transcription level of each strain effect block located within the OSAB domains in each organ. The full diploid genome was used for mapping. The non-allelic-biased organs with highest rpkm (nonBiasedH) were selected to compare with the rpkm of the allelic-biased organs in OSAB domains. Scripts implementing this can be found here:
https://github.com/Danko-Lab/F1_8Organs/blob/main/AllelicBiase_expressionLevel.bsh
https://github.com/Danko-Lab/F1_8Organs/blob/main/getNonBiasedHighest_Biased_TotalReadCountRatio.R

Analysis of allele-specific initiation:

Identification of candidate transcription initiation sites: We used 5 prime end of ChRO-seq reads (the R1 paired-end sequencing file, which represents the 5 prime end of the nascent RNA) to identify candidate initiation sites using methods adapted from (Tome et al., 2018). Briefly, candidate transcription start sites (TSS) from each read were merged into candidate transcription start clusters, in which the max TSS was identified. We identified candidate TSSs that fall within dREG sites and were supported by at least 5 separate reads. TSSs within 60bp of each other were merged into candidate TSCs. The TSS with the maximal read depth in each TSC was defined as the maxTSS for that TSC. We allow each TSC to have more than one maxTSSs if multiple TSSs share the same number of read counts in that TSC. To test whether the candidate maxTSSs represented bona-fide transcription start sites, we generated sequence logos centered on the maxTSS using the seqLogo R package (Bembom, 2019). We retained tissues in which the maxTSS contained a clearly defined initiator dinucleotide that reflects a

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

similar sequence composition as those previously reported (Tome et al., 2018). Additionally, we used an in-house R script to examine the relationship between TSS counts and Read counts of the TSC (Supp Fig2 B), and found a similar relationship to those reported (Tome et al., 2018).

Identify allele specific differences in TSCs abundance (ASTSC abundance): We used a binomial test to identify candidate allele specific transcription start clusters, with an expected allelic ratio of 0.5. We filtered candidate allele specific differences using a false discovery rate (FDR) corrected p-value of 0.1, corresponding to an expected 10% FDR.

Identify allele specific differences in TSC shape (ASTSC shape): We used a Kolmogorov-Smirnov (K-S) test to identify TSCs where the distribution of transcription initiation differed significantly between the B6 and CAST alleles (ASTSC shape). We used TSC sites with at least 5 mapped reads specific to the B6 genome and at least 5 mapped reads specific to CAST. Only autosomes were used. We corrected for multiple hypothesis testing using the false discovery rate (Storey and Tibshirani, 2003) and filtered ASTSC shapes using a 10% FDR. We further separated the ASTSC shape candidates into two groups: one driven by a single TSS (single TSS driven ASTSC shape), the other reflecting changes in more than one base in the TSC (multiple TSS driven ASTSC shape). To separate into two groups, we masked the TSS with the highest allelic difference (determined by read counts) within each TSC and performed a second K-S test. Multiple TSS driven ASTSC were defined as those which remained significantly different by K-S test after masking the position of highest allelic difference. Single TSS driven ASTSCs were defined as ASTSCs that were no longer significantly different by K-S test after masking the maximal position. In the second K-S test, we used the nominal p-value defined as the highest nominal p-value that achieved a 10% FDR during the first K-S test.

SNP analysis: We examined the distribution of single nucleotide polymorphisms (SNPs) near ASTSCs from each class. A major confounding factor in SNP distribution is the ascertainment bias of requiring at least one tagged SNP to define the allelic imbalance between the two alleles, resulting in an enrichment of SNPs within the read. To control for this bias, we compared the set of sites with a significant change in the TSC shape or abundance (FDR <= 0.1) with a background control set defined as candidate TSCs in which there was no evidence of change between alleles (FDR > 0.9) in all analyses. We display a bin size of 5 bp. To test for differences, we merged adjacent bins by using a bin size of 10bp to increase statistical power and tested for enrichment using Fisher's exact test, FDR cutoff = 0.05. (Fig. 3E,F). We also examined the difference in base composition between the allele with high and low initiation in each ASTSC shape difference centered on the position of the maxTSS in the allele with high initiation (in Fig. 3G). We determined the high/low allele based on the transcription level at maxTSS. If there are more than one TSSs with the max read counts, there will be more than one maxTSSs representing each TSC.

Comparison of SNPs in TATA motifs: We extracted the DNA sequence in the region between -35 to -20 bp upstream of each max TSN on both the B6 and CAST alleles. We used RTFBSDB (Wang et al., 2016) to compute the maximal score (defined using the log likelihood ratio of a motif match compared to a 1 bp Markov model as background) of a TATA motif (> 3) in this

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

region using two TATA motifs: M00216 (low information content) and M09433 (high information content). We compared the difference in the maximal score between the B6 and CAST alleles to the difference in max TSN usage between B6 and CAST alleles. Pearson's correlation between these two variables was tested using the cor.test function in R.

Comparison of AT content between alleles: As a proxy for melting temperature (in Fig. 3H), we examined the AT content in 5 bp windows around the maxTSS on alleles with high and low maxTSS usage. As in the SNP analysis (above), we compared the set of sites with a significant change in the TSC shape or abundance (FDR <= 0.1) with a background control set defined as candidate TSCs in which there was no evidence of change between alleles (FDR > 0.9). Computations were performed using R library TmCalculator (Li, 2019). We used Fisher's exact test to examine if there was an enrichment of AT(in the high allele) to GC (in the low allele) SNPs in each 5 bp bin, and adopted an FDR corrected p-value cutoff = 0.05. In all analyses, positions at -1 and 0 relative to the maxTSS were masked to avoid confounding effects of the initiator sequence motif on computed AT content.

Shooting gallery: In our analysis of the shooting glaray model, we focused on a subset of TSCs which do not appear to change expression globally, and which have a SNP in the initiator element. Toward this end, we identified TSCs which do not overlap AlleleHMM blocks. We set the allele with high and low expression based on allele specific reads in the maxTSS. Next, we divided data into a test and background control dataset in which the test set had a CA dinucleotide in the allele with high maxTSS use and any other combination except for CA on the other allele. The control set did not have a CA dinucleotide in the maxTSS initiator position. Next we computed the distance to the maxTSS and the allelic read count at other candidate initiator motifs (including CA, CG, TA, TG). In all analyses, we compared the set of maxTSSs with SNPs in the initiator position with the control set which did not have a SNP. Statistical tests used an unpaired Wilcoxon rank sum test. We corrected for multiple hypothesis testing using false discovery rate.

All scripts implementing analysis of allele-specific initiation can be found at: https://github.com/Danko-Lab/F1_8Organs/tree/main/initiation

Analysis of allele-specific pause:

Identification of allele specific differences in pause site shape: All pause analysis focused on ChRO-seq data in three organs (heart, skeletal muscle, and kidney) which used a single base run-on of all four biotin nucleotides. We first focused our analysis on dREG sites in each tissue to identify regions enriched for transcription start and pause sites. We retained dREG sites in which we identified at least 5 reads mapping from both B6 and Cast alleles. We performed a K-S test to identify all candidate dREG sites that contained a candidate difference in pause, filtering for a false discovery rate of 0.1 (n= 2784). To examine the relationship between initiation and pause, we identified the maxTSS and maxPause on the B6 and Cast allele separately using reads tagged with a SNP or indel. Since maxTSS and maxPause were defined

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

independently, the maxPause was not always correctly paired with the maxTSS (Tome et al., 2018). We therefore used 2260 dREG sites where allelic maxPause were 10 to 50 bp downstream of allelic maxTSS on both alleles. These analyses pertain to Fig. 5A and B.

Identify genetic determinants of pausing: To focus on the genetic determinants of pausing that were independent of initiation, we identified changes in which the same maxTSS had different allelic maximal pause sites between the Cast and B6 alleles as follows. We used a K-S test to identify maxTSSs with a difference in the maxPause site between alles, filtering for maxTSSs with a different maxPause between alleles and a 10% FDR in a K-S test (n = 269). In most analyses, we also draw a background set in which there was no evidence that sites sharing the same maxTSS had different maxPause sites between alleles, by identifying maxTSSs that have the same maxPause position and a K-S test FDR >0.9 (n = 1396). In all analyses, we also filtered for maxTSSs with at least 5 allelic reads and B6/CAST read ratio between 0.5 and 2.

Comparing GC content near the maxPause position: We compared the G, C, and GC content between alleles. We computed the G, C, and GC content as a function of position relative to maxPause. All of the G, C, and GC contents were combined across unique pause sites from all three organs for which we had single base resolution data (n = 3456). We compared three blocks: block 1 was 11 to 20 nt upstream of maxPause, block 2 was 1 to 10 nt upstream of maxPause, and block 3 was 1 to 10 nt downstream of maxPause.

All scripts implementing analysis of allele-specific pause can be found at: https://github.com/Danko-Lab/F1_8Organs/tree/main/pause

Analysis of allele-specific termination:

Definition of allelic differences in termination: We noticed frequent AlleleHMM blocks near the 3' end of annotated genes. We used the transcription units (tunits) predictions which overlapped annotated protein coding genes (vM25), as these generally retained the window between the polyadenylation cleavage site and the transcription termination site. We identified transcription units that have AlleleHMM blocks starting within the transcription unit and that end in the final 10% of the transcription unit or after the transcription unit. The overlapping region between the tunits and AlleleHMM blocks were called candidate allelic termination (AT) windows. To avoid obtaining candidate AT windows that reflected entire transcription units, we retain only AT windows whose length was less than or equal to 50% length of the host transcription unit.

RNA-seq analysis: Allele specific mapped RNA-seq reads using STAR (see above) were used as input to AlleleHMM to identify the region showing candidate allelic difference in mature mRNA. The transcription units that contain allelic termination windows, as defined above, were separated into two groups: One has an allelic difference in mature mRNA and the other does not. Those with an allelic difference in mature mRNA were defined as having the RNA-seq AlleleHMM blocks between 10Kb upstream of the AT windows to the end of the AT windows.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

RNA stability analysis: We asked whether there was an allelic difference in mRNA stability between transcription units in which the allelic differences in termination affects the mature mRNA and those in which it does not. The RNA stability was defined as in (Blumberg et al., 2021). The stability was defined as the ratio of RNA-seq read counts in exons to ChRO-seq read counts across the gene body. We used gene annotations from GENCODE (vM25). The RNA-seq reads were counted strand specifically using htseq-count. ChRO-seq reads were counted in a strand-specific fashion using in-house R scripts. After removing the genes with less than 10 B6-specific ChRO-seq and less than 10 CAST-specific ChRO-seq reads, the cumulative distribution functions were drawn. All differences were compared using a one-sided K-S test to compare differences in allelic RNA stability between groups.

All scripts implementing analysis of allele-specific termination can be found at: https://github.com/Danko-Lab/F1_8Organs/tree/main/termination

Acknowledgements

We thank Maria Garcia-Garcia, Abdullah Ozer, John Lis, Hojoong Kwak, Gilad Barshad, Alexandra Chivu, and all members of the Danko lab for valuable discussions and suggestions throughout the life of this project. We thank Peter Borst for help preparing and working with F1 hybrid mice and Jen Grenier and the Cornell TREx facility for preparing mRNA-seq libraries. We thank C. Kaplan (U. Pittsburgh) for rapid constructive comments based on our *bioRxiv* preprint. Work in this publication was supported by R01-HG010346 and R01-HG009309 (NHGRI) to CGD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health. Some of the figures in this manuscript were created using BioRender. All data are available at Gene Expression Omnibus under the accession number GSE174171.

Figure Captions:

Figure 1: Reciprocal hybrid cross to understand the Pol II transcription cycle.

(A) Cartoon illustrates the reciprocal F1 hybrids cross design between the strains C57BL/6J (B6) and CAST/EiJ (CAST). We have seven independent crosses (3x C57BL/6 x Cast and 4x Cast x C57BL/6).

(B) Spearman's rank correlation of ChRO-seq signals in gene bodies. The color on the top indicates the direction of crosses: Black is B6 x CAST, brown is CAST x B6. The cartoon on the right indicates the organ each sample was harvested from.

(C) Heatmap shows the allelic bias (log2[B6 / CAST]) in GENCODE annotated genes >10kb in size. Row order is determined by hierarchical clustering (using 1 - Pearson correlation); Column order was set manually. Several of the different gene cluster interpretations are shown by the writing on the right.

(D) The histogram shows the proportion of domains as a function of the domain length.

(E) The histogram shows the proportion of domains as a function of the number of gencode gene annotations in each domain.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

(F) The browsershot shows an example of ChRO-seq data that has an imprinted domain (top row). The second and third rows show the imprinted protein-coding genes and imprinted non-coding RNA (ncRNA) from all organs. (BN:brain, LV:liver, SK:skeletal muscle, GI: large intestine, HT: heart, KD: kidney, P: paternal, M: maternal). The yellow shade indicates the imprinted regions in the brain and liver.

Figure 1—figure supplement 1:

(A) The histogram shows the frequency of blocks within organ-specific allelic biased domains (OSAB domain) as a function of effect size. Red (Biased) is from the organ with OSAB domain. Blue (NonBiased) is from the organ with the highest expression that is putatively unbiased. If the allelic-biased organ was maternally biased, the effect size was calculated as maternal reads divided by paternal reads in the blocks, otherwise the effect size was calculated as paternal reads divided by maternal reads in the blocks.

(B) The histogram shows the frequency of blocks within the OSAB domain as a function of maternal reads ratio. Red (Biased) is from the organ with OSAB domain. Blue (NonBiased) is from the organ with the highest expression that is putatively unbiased.

(C) The histogram shows the frequency of blocks within OSAB domain as a function of the log2 ratio between the rpkm of the non-allelic-biased organs with highest rpkm (nonBiasedHighest) and the allelic-biased organs in OSAB domains.

(D) The browsershot shows an example of ChRO-seq data that has a strain effect domain (top row) (BN:brain, LV:liver, SK:skeletal muscle, GI: large intestine, HT: heart, KD: kidney, P: paternal, M: maternal). The second and third rows show alleleHMM blocks that are specific to liver within the domain. Tracks show ChRO-seq reads mapping to B6 and CAST genomes, SNPs, and a subset of RefSeq annotated genes.

(E) Cartoon depicts the methods used to identify allelic biased blocks, domains, and how they were classified as a strain-associated effect or imprinted.

Figure 2: Allelic changes in TSSs reveal a hierarchy of initiator dinucleotides.

(A) The ChRO-seq signals of the 5' end of the nascent RNA were used to define transcription start sites (TSSs), or the individual bases with evidence of transcription initiation. TSSs within 60bp were grouped into transcription start clusters (TSCs). The broad location of transcription start regions (TSRs), which comprise multiple TSCs, was determined using dREG.

(B) Violin plots show the ratio of allelic bias between the dinucleotides indicated on the X axis. Asterisk denote statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$) using a two-sided Wilcoxon rank sum test.

Figure 2—figure supplement 1:

(A) Sequence logos show the information content around the maxTSSs of each organ.

(B) Scatter plot shows the number of TSSs in a TSC as a function of the read counts in the TSC.

Figure 3: Allelic changes in the shape of transcription initiation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

(A) The browsershot shows an example of allelic differences in the shape of TSC that are predominantly explained by a single TSS position (arrowhead). ALL indicates signal from all reads, IDE indicates signal from reads that are not tagged with a SNP, B6 indicates signal from reads tagged with B6 SNP, CAST indicates signal from reads tagged with CAST SNP.

(B) The browsershot shows an example of allelic differences in TSC driven by multiple TSSs within the same TSC, arrowheads indicate several prominent positions with allelic differences in PolI abundance. ALL indicates signal from all reads, IDE indicates signal from reads that are not tagged with a SNP, B6 indicates signal from reads tagged with B6 SNP, CAST indicates signal from reads tagged with CAST SNP.

(C) Scatterplot shows the average SNP counts as a function of distance to the maxTSS at sites showing allelic differences in TSCs driven by a single TSS in the liver. Red denotes changes in TSC shape (Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSCs without evidence for differences in TSC shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05). Green and orange boxes denote the position of PIC binding motifs.

(D) The scatterplot shows the average difference in base composition between the allele with high and low TSS use around the maxTSS in single-base driven allele specific TSCs. The sequence logo on the bottom represents the high allele in single-base driven allele specific TSCs. The high/low allele were determined by the read depth at maxTSS.

(E) Scatterplot shows the average SNP counts as a function of distance to the maxTSS at sites showing allelic differences in TSC driven by multiple TSSs in the liver sample. Blue denotes changes in TSC shape classified as multiple TSS driven (Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSCs without evidence for differences in TSC shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)

(F) The scatter plot shows the difference of AT contents between the high and low alleles with the maxTSS and -1 base upstream maxTSS masked. Dots represent 5 bp non-overlapping windows. Red denotes single TSS driven allele specific TSCs; black denotes control TSCs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT(at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher's exact test, FDR <= 0.05).

Figure 3—figure supplement 1:

(A) Stacked bar chart shows the average proportion of allelic differences in TSCs driven by a single TSS (red). The specific proportion for each of the six organs are shown by dots. Six organs were selected that showed a strong signal of Inr motif at the maxTSSs (brain, heart, liver, large intestine, stomach, and spleen).

(B) Scatterplot shows the average SNP counts as a function of distance to the maxTSS at sites showing allelic differences in TSCs driven by a single TSS in the brain. Red denotes changes in TSC shape (Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSCs without evidence for differences in TSC shape (KS test; FDR > 0.90).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)

(C) The scatterplot shows the average difference in base composition between the allele with high and low TSS use around the maxTSS in single-base driven allele specific TSCs. The sequence logo on the bottom represents the high allele in single-base driven allele specific TSCs. The high/low allele were determined by the read depth at maxTSS. This figure denotes TSCs in the brain.

(D) Scatterplot shows the average SNP counts as a function of distance to the maxTSS at sites showing allelic differences in TSC driven by multiple TSSs in the brain sample. Blue denotes changes in TSC shape classified as multiple TSS driven (FDR corrected Kolmogorov-Smirnov (KS) test; FDR <= 0.10); black indicates TSCs without evidence for differences in TSC shape (KS test; FDR > 0.90). Dots represent non-overlapping 5 bp bins. Yellow shade indicates statistically significant differences (false discovery rate corrected Fisher's exact on 10 bp bin sizes, FDR <= 0.05)

(E) The scatter plot shows the difference of AT contents between the high and low alleles in the brain with the maxTSS and -1 base upstream maxTSS masked. Dots represent 5 bp non-overlapping windows. Red denotes singleTSS driven allele specific TSCs; black denotes control TSCs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT(at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher's exact test, FDR <= 0.05).

(F) The scatter plots show the difference of AT contents between the high and low alleles in liver with the maxTSS and -1 base upstream maxTSS masked. This plot shows the multiple TSS driven allele specific TSC in the liver samples. Dots represent 5 bp non-overlapping windows. Blue denotes multiple TSS driven allele specific TSCs; black denotes control TSCs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT(at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher's exact test, FDR <= 0.05).

(G) The scatter plots show the difference of AT contents between the high and low alleles with the maxTSS and -1 base upstream maxTSS masked. This plot shows the multiple TSS driven allele specific TSC in the brain samples. Dots represent 5 bp non-overlapping windows. Blue denotes multiple TSS driven allele specific TSCs; black denotes control TSCs with no evidence of allele specific changes. The yellow shade indicates a significant enrichment of AT(at high allele) to GC (at low allele) SNPs at each bin (size=5bp; Fisher's exact test, FDR <= 0.05).

Figure 4: Browian motion model of transcription start site selection.

(A) Cartoon shows our expectation of the effects of allelic DNA sequence variation on transcription start site selection based on the yeast "shooting gallery" model (left), or the mammalian model (right), in which Pol II initiates at potential TSSs (triangles) after DNA melting. We expected that mutations in a strong initiator dinucleotide (CA) on (for example) the CAST allele (bottom) would shift initiation to the initiator elements further downstream (yeast model), or would not affect the adjacent initiation sites (mammalian model). The size of the triangle indicates the strength of the initiator.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

(B) The violin plots show the distribution of ChRO-seq signal ratios on the candidate initiator motifs (including CA, CG, TA, TG) within 20bp of the maxTSSs that had a CA dinucleotide in the allele with high maxTSS (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray). Note that the central maxTSS was not included in the analysis. Wilcoxon rank sum test with continuity correction is p-value = 5.665e-10 for Brain and p-value < 2.2e-16 for liver.

(C) The box plots show the distribution of ChRO-seq signals ratios at TSSs with any YR dinucleotide (i.e., CA, CG, TA, TG) in both alleles as a function of the distance from the maxTSSs that had a CA dinucleotide in the allele with high maxTSS (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray). Yellow shade indicates Wilcoxon Rank Sum and Signed Rank Tests (SNP in Inr vs no SNP in Inr) with fdr <= 0.05. The TSCs were combined from the brain and liver samples.

(D) The browser shot shows an example of a maxTSS with increased initiation upstream and downstream of an allelic change in a CA dinucleotide. The arrow denotes a SNP at the maxTSS, in which B6 contains the high maxTSS with CA and CAST contains CG. The ChRO-seq signals at the alternative TSS with a CA dinucleotide (arrow head) upstream of the maxTSS were higher in the low allele (CAST in this case), resulting in a different maxTSS in CAST. Additional tracks show the B6 reference genome sequence, the position of all SNPs between B6 and CAST, and RefSeq gene annotations. All tracks line up with ChRO-seq data. ALL indicates signal from all reads, IDE indicates signal from reads that are not tagged with a SNP, B6 indicates signal from reads tagged with B6 SNP, CAST indicates signal from reads tagged with CAST SNP.

(E) Proposed model in which Pol II initiates from a PIC and selects an energetically favorable TSS by random movement along the DNA similar to brownian motion.

(F) Boxplots show the distribution of ChRO-seq signal ratios on the candidate initiator motifs (including CA, CG, TA, TG) in the same TSR in the allele with high maxTSS (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray). Note that the central maxTSS was not included in the analysis.

Figure 4—figure supplement 1:

Violin plots show the distribution of ChRO-seq signals ratios between high low alleles at the candidate initiator motifs (All Inr : CA, CG, TA, TG; OnlyWeak Inr : CG, TA, TG; and OnlyCA) that are within 20bp of the maxTSSs that had a CA dinucleotide in the allele with high maxTSS (SNP in Inr, purple) or had a CA dinucleotide in both alleles (No SNP in Inr, gray) in Brain(BN) or Liver(LV). Single: indicates Single TSSven allele-specific TSCs. Multiple: indicates Multiple TSS driven allele-specific TSCs. Yellow shade indicates Wilcoxon Rank Sum and Signed Rank Tests (SNP in Inr vs no SNP in Inr) with fdr <= 0.05.

Figure 5: Allele specific effects on the distribution of Pol II in the promoter proximal pause.

(A) Scatterplots show the relationship between distances of allelic maxPause and allelic maxTSS within dREG sites with allelic different pause (n = 2,260).

(B) Top histogram shows the number of sites as a function of the distance between allelic maxTSS in which the allelic maxPause was identical (n = 359). Bottom histogram shows

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

the number of sites as a function of the distance between allelic maxPause where the allelic maxTSS was identical (n = 823).

(C) Scatterplot shows the relationship between indel length and the allelic difference of the average pause position on the reference genome (mm10). The pause positions of CAST were first determined in the CAST genome and then liftovered to mm10. Only sites initiated from the maxTSS and with allelic difference in pause shape were shown (KS test, fdr <= 0.1; also requiring a distinct allelic maximal pause). Color indicates the organs from which the TSS-pause relationship was obtained.

(D) Top: scatterplot shows the average SNPs per base around the position of the Pol II in which the distance between the maxTSS and the max pause was lowest (short pause). Red represents sites with allelic difference in pause shape (Allelic pause difference, KS test, fdr <= 0.1 with distinct allelic maxPause, n = 269), Blue is the control group (No allelic pause difference, KS test, fdr > 0.9 and the allelic maxPause were identical, n = 1,396). Bottom: The sequence logo obtained from the maxPause position based on all reads (n = 3,456 max pause sites). Sites were combined from three organs, after removing pause sites that were identical between organs.

(E) Violin plots show the G content, GC content and C content as a function of position relative to maxPause defined using all reads (combined from three organs with duplicate pause sites removed, n = 3,456), block 1 was 11 to 20 nt upstream of maxPause, block 2 was 1 to 10 nt upstream of maxPause, and block 3 was 1 to 10 nt downstream of maxPause (* p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001). Block 2 had a higher G content and a lower C content than the two surrounding blocks.

(F) Pie charts show the proportion of different events around the maxPause of short alleles (short pause) with or without allelic pause differences. Sites were combined from three organs with duplicated pause sites removed.

(G) Sequence logos show the sequence content of short alleles and long alleles at 270 short pause sites and 278 long pause sites.

(H) The histograms show the fraction of pause sites as a function of distance between allelic maxPause, i.e. the distance between the short and long pause. The lines show the cumulative density function. Blue represents pause sites with allelic differences (n = 285); red is a subgroup of blue sites with a C to A/T/G SNP at the maxpause (n = 34). Two-sample Kolmogorov-Smirnov test p-value = 0.002694

(I) The histograms show the fraction of pause sites as a function of distance between allelic maxPause. The lines show the cumulative density function. Blue is pause sites with allelic differences (n = 285), green is a subgroup of blue sites that contain indels between initiation and long pause sites (n=60), Two-sample Kolmogorov-Smirnov test p-value = 0.02755.

Figure 5—figure supplement 1:
Spearman's rank correlation of the ChRO-seq data, including samples with a single base resolution for the Pol II active site. Samples with single nucleotide precision are shown on the top in red bold font and at the right with star. (BN:brain, GI: large intestine, ST: Stomach, LV:liver, KD: kidney, SP:spleen, HT: heart, SK:skeletal muscle, MB6: B6 x CAST, PB6: CAST x B6)

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Figure 5—figure supplement 2:

(A) Sequence logos show the information contents around the initiation site (see Methods).

(B) Scatter plots show the difference of nucleotide usage between short and long alleles as a function of distance to allelic maxPause.

Figure 6: Widespread allele specific differences in the Pol II termination site.

(A) The browser shot shows an example of allelic termination differences (yellow shade) in both brain and liver. Pol II terminates earlier on the B6 allele, resulting in a longer transcription unit on the CAST allele. The difference in allelic read abundance was identified by AlleleHMM. We defined the allelic termination difference (yellow shade) using the intersection between the transcription unit and AlleleHMM blocks. Tracks represent all ChRO-seq signal (top, marked ALL), reads mapping uniquely to the B6 or CAST allele (mid), the location of dREG, transcription units and AlleleHMM blocks (bottom). The position of the start of allelic termination is marked by an arrow. The allelic termination window is marked by yellow shading.

(B) The histogram shows the fraction of transcription units as a function of the length of allelic termination difference.

(C) Heatmaps show the raw read counts in transcription units (blue bar) with an allelic termination difference (yellow bar), centered at the beginning of allelic termination (solid triangle). The heatmap bin size is 500 bp, and 20kb is shown upstream and downstream. The rows were sorted by the length of allelic termination differences determined by ChROseq signals from Liver. The short and long alleles were determined based on analysis of the liver.

(D) Pie charts show the proportion of transcription units with allelic termination difference that also contains allelic difference in mature mRNA (orange).

Figure 6—figure supplement 1:

(A) Scatterplots show the relationship between the ChRO-seq signal in the transcription unit (defined using the tunits program; see Methods) and the region showing an allelic termination difference.

(B) Violin plots show the distribution of the length of allelic termination in eight organs.

Figure 6—figure supplement 2:

(A) The browser shot shows an example of allelic termination differences (yellow shade) in brain tissue. Pol II terminates earlier on the CAST allele, resulting in a longer transcription unit on the B6 allele. This example shows a punctate signal in the RNA-seq data that appears to be consistent with higher usage of an additional exon on the B6 allele. Tracks represent all ChRO-seq signal (top, marked ALL), reads mapping uniquely to the B6 or CAST allele (mid), transcription units, AlleleHMM blocks, all RNA-seq signal (marked ALL), and RNA-seq reads mapping uniquely to the B6 or CAST allele, and RefSeq gene annotations.

(B) Scatterplots represent the cumulative density function of allelic RNA stability differences in brain samples. Two-sample Kolmogorov-Smirnov tests, p-value = 4.3e-4.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

(C) The lines show the cumulative density function of the allelic RNA stability difference in the liver samples. Two-sample Kolmogorov-Smirnov tests, p-value = 2.69e-8.

References

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* **24**:14–24.

Bembom O. 2019. seqLogo: Sequence logos for DNA sequence alignments.

Blumberg A, Zhao Y, Huang Y-F, Dukler N, Rice EJ, Chivu AG, Krumholz K, Danko CG, Siepel A. 2021. Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *BMC Biol* **19**:30.

Braberg H, Jin H, Moehle EA, Chan YA, Wang S, Shales M, Benschop JJ, Morris JH, Qiu C, Hu F, Tang LK, Fraser JS, Holstege FCP, Hieter P, Guthrie C, Kaplan CD, Krogan NJ. 2013. From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell* **154**:775–788.

Breslauer KJ, Frank R, Blöcker H. 1986. Predicting DNA duplex stability from the base sequence. *Proceedings of the*.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SPT, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CAM, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y, FANTOM Consortium, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). 2005. The transcriptional landscape of the mammalian genome. *Science* **309**:1559–1563.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**:626–635.

Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. 2016. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* **7**:11101.

Cho H, Kim TK, Mancebo H, Lane WS, Flores O, Reinberg D. 1999. A protein phosphatase functions to recycle RNA polymerase II. *Genes Dev* **13**:1540–1552.

Chou S-P, Danko CG. 2019. AlleleHMM: a data-driven method to identify allele specific differences in distributed functional genomic marks. *Nucleic Acids Res*. doi:10.1093/nar/gkz176

Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, Longo SL, Corona RJ, Chin LS, Lis JT, Kwak H, Danko CG. 2018. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat Genet* **50**:1553–1564.

Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, Martins AL, Dukler N, Coonrod SA, Tait Wojno ED, Lis JT, Kraus WL, Siepel A. 2018. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nature Ecology & Evolution*. doi:10.1038/s41559-017-0447-5

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**:212–222.

Delaneau O, Zazhytska M, Borel C, Giannuzzi G, Rey G, Howald C, Kumar S, Ongen H, Popadin K, Marbach D, Ambrosini G, Bielser D, Hacker D, Romano L, Ribaux P, Wiederkehr M, Falconnet E, Bucher P, Bergmann S, Antonarakis SE, Reymond A, Dermitzakis ET. 2019. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**:eaat8266.

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

Fant CB, Levandowski CB, Gupta K, Maas ZL, Moir J, Rubin JD, Sawyer A, Esbin MN, Rimel JK, Luyties O, Marr MT, Berger I, Dowell RD, Taatjes DJ. 2020. TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. *Mol Cell* **78**:785–793.e8.

Fuda NJ, Ardehali MB, Lis JT. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**:186–192.

Giardina C, Lis JT. 1993. DNA melting on yeast RNA polymerase II promoters. *Science* **261**:759–762.

Gilchrist DA, Dos Santos G, Fargo DC, Xie B, Gao Y, Li L, Adelman K. 2010. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**:540–551.

Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, Cramer P. 2017. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* **6**. doi:10.7554/eLife.29736

Grosso AR, de Almeida SF, Braga J, Carmo-Fonseca M. 2012. Dynamic transitions in RNA polymerase II density profiles during transcription termination. *Genome Res* **22**:1447–1456.

Grünberg S, Warfield L, Hahn S. 2012. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat Struct Mol Biol* **19**:788–796.

Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

initiation. *Nat Rev Mol Cell Biol*. doi:10.1038/s41580-018-0028-8
Jonkers I, Kwak H, Lis JT. 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**:e02407.
Jonkers I, Lis JT. 2015. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**:167–177.
Kaplan CD, Jin H, Zhang IL, Belyanin A. 2012. Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo. *PLoS Genet* **8**:e1002627.
Kaufmann J, Smale ST. 1994. Direct recognition of initiator elements by a component of the transcription factor IID complex. *Genes Dev* **8**:821–829.
Kristjánsdóttir K, Dziubek A, Kang HM, Kwak H. 2020. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat Commun* **11**:5963.
Kuehner JN, Brow DA. 2006. Quantitative analysis of in vivo initiator selection by yeast RNA polymerase II supports a scanning model. *J Biol Chem* **281**:14119–14128.
Kumasaka N, Knights AJ, Gaffney DJ. 2019. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat Genet* **51**:128–137.
Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**:950–953.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**:R25.
Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506–511.
Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**:233–245.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760.
Li J. 2019. TmCalculator: Melting Temperature of Nucleic Acid Sequences.
Luse DS, Parida M, Spector BM, Nilson KA, Price DH. 2020. A unified view of the sequence and functional organization of the human RNA polymerase II promoter. *Nucleic Acids Res*. doi:10.1093/nar/gkaa531
Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. 2016. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* **11**:1455–1476.
Mahat DB, Salamanca HH, Duarte FM, Danko CG, Lis JT. 2016. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol Cell*. doi:10.1016/j.molcel.2016.02.025
Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF, Shilatifard A. 2001. COMPASS: A complex of proteins associated with a trithorax-related SET domain protein. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.231473398
Mittleman BE, Pott S, Warland S, Barr K, Cuevas C, Gilad Y. 2021. Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees. *Elife* **10**. doi:10.7554/eLife.62548
Mittleman BE, Pott S, Warland S, Zeng T, Mu Z, Kaur M, Gilad Y, Li Y. 2020. Alternative

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

polyadenylation mediates genetic regulation of gene expression. *Elife* **9**. doi:10.7554/eLife.57492

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**:773–777.

Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, Azubel M, Elmlund D, Levi-Kalisman Y, Liu X, Gibbons BJ, Levitt M, Kornberg RD. 2013. Architecture of an RNA Polymerase II Transcription Pre-Initiation Complex. *Science* **342**. doi:10.1126/science.1238724

Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. *Nat Genet* **39**:1507–1511.

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

Nilson KA, Lawson CK, Mullen NJ, Ball CB, Spector BM, Meier JL, Price DH. 2017. Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic Acids Res*. doi:10.1093/nar/gkx724

Orphanides G, LeRoy G, Chang CH, Luse DS, Reinberg D. 1998. FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell* **92**:105–116.

O'Sullivan JM, Tan-Wong SM, Morillon A, Lee B, Coles J, Mellor J, Proudfoot NJ. 2004. Gene loops juxtapose promoters and terminators in yeast. *Nat Genet* **36**:1014–1018.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**:768–772.

Qiu C, Jin H, Vvedenskaya I, Llenas JA, Zhao T, Malik I, Visbisky AM, Schwartz SL, Cui P, Čabart P, Han KH, Lai WKM, Metz RP, Johnson CD, Sze S-H, Pugh BF, Nickels BE, Kaplan CD. 2020. Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. *Genome Biol* **21**:132.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842.

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

Ranish JA, Yudkovsky N, Hahn S. 1999. Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. *Genes Dev* **13**:49–63.

Rennie S, Dalby M, van Duin L, Andersson R. 2018. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat Commun* **9**:487.

Rosonina E, Kaneko S, Manley JL. 2006. Terminating the transcript: breaking up is hard to do. *Genes Dev* **20**:1050–1056.

Rougvie AE, Lis JT. 1988. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54**:795–804.

Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M. 2011. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**:522.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**:863–864.

Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. *Science* **352**:1225–1228.

Smale ST, Baltimore D. 1989. The "initiator" as a transcription control element. *Cell* **57**:103–113.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad*

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Sci U S A **100**:9440–9445.

Tome JM, Tippens ND, Lis JT. 2018. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat Genet*. doi:10.1038/s41588-018-0234-5

Traut TW. 1994. Physiological concentrations of purines and pyrimidines. *Mol Cell Biochem* **140**:1–22.

Tsai FTF, Sigler PB. 2000. Structural basis of preinitiation complex assembly on human Pol II promoters. *EMBO J* **19**:25–36.

van de Geijn B, McVicker G, Gilad Y, Pritchard JK. 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**:1061–1063.

Wang Z, Chu T, Choate LA, Danko CG. 2018. Identification of regulatory elements from nascent transcription using dREG. *Genome Res* **29**:293–303.

Wang Z, Martins AL, Danko CG. 2016. RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics*. doi:10.1093/bioinformatics/btw338

Zeitlinger J, Stark A, Kellis M, Hong J-W, Nechaev S, Adelman K, Levine M, Young RA. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39**:1512–1516.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

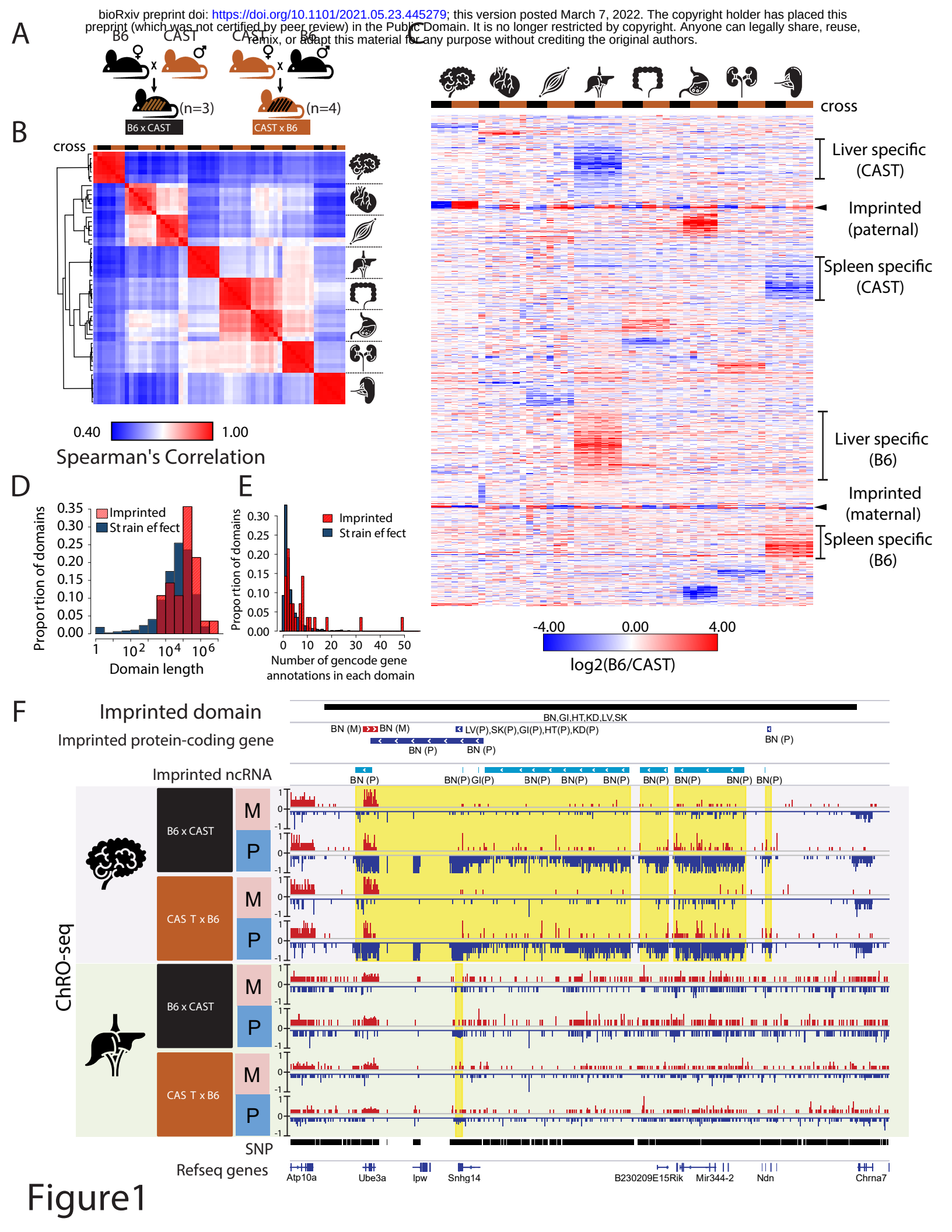
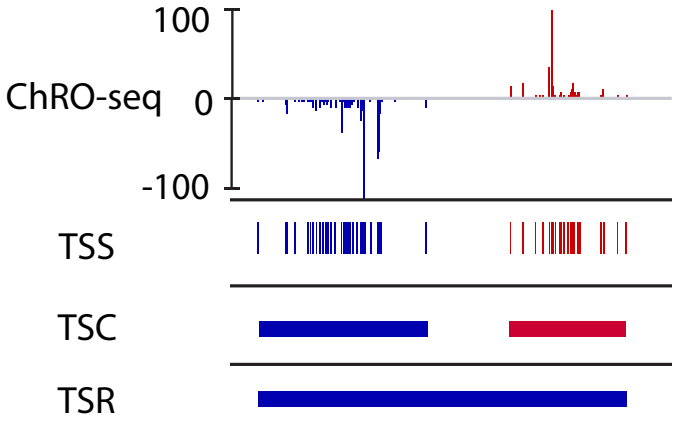


Figure1

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

A



B

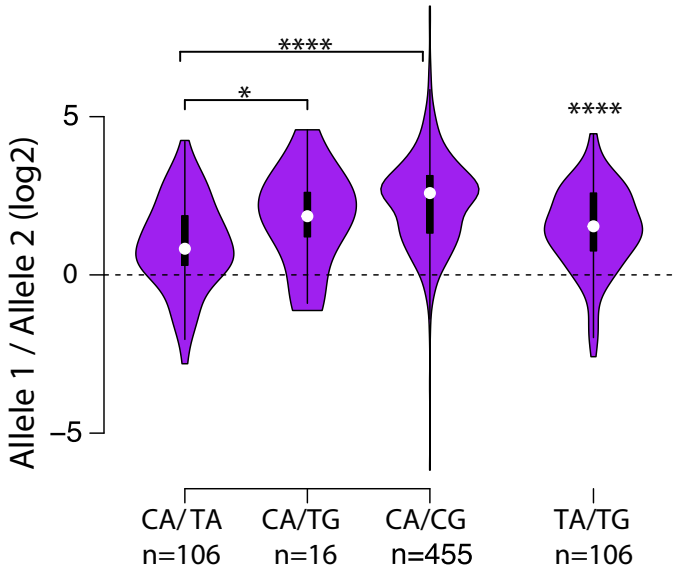


Figure2

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

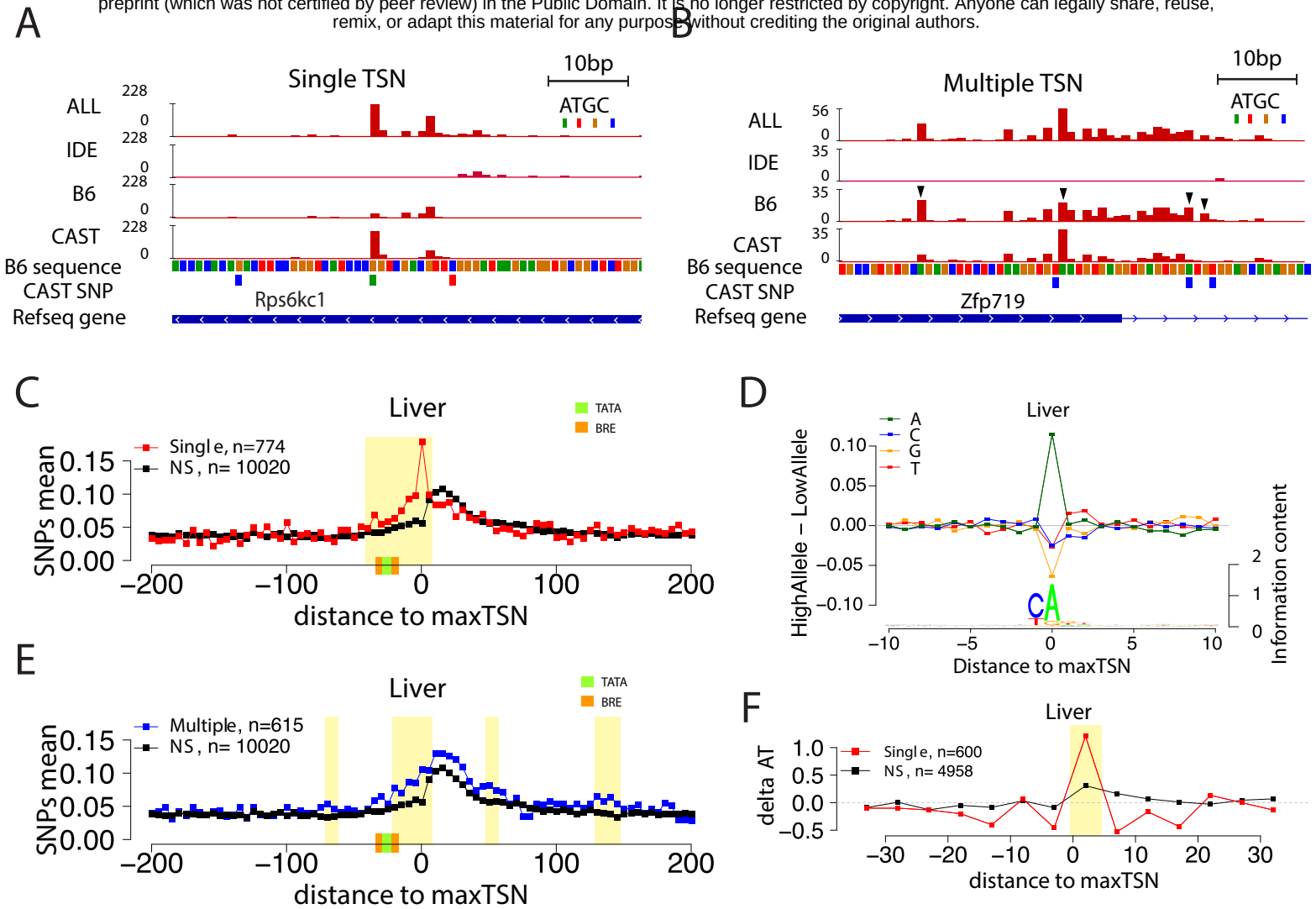


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

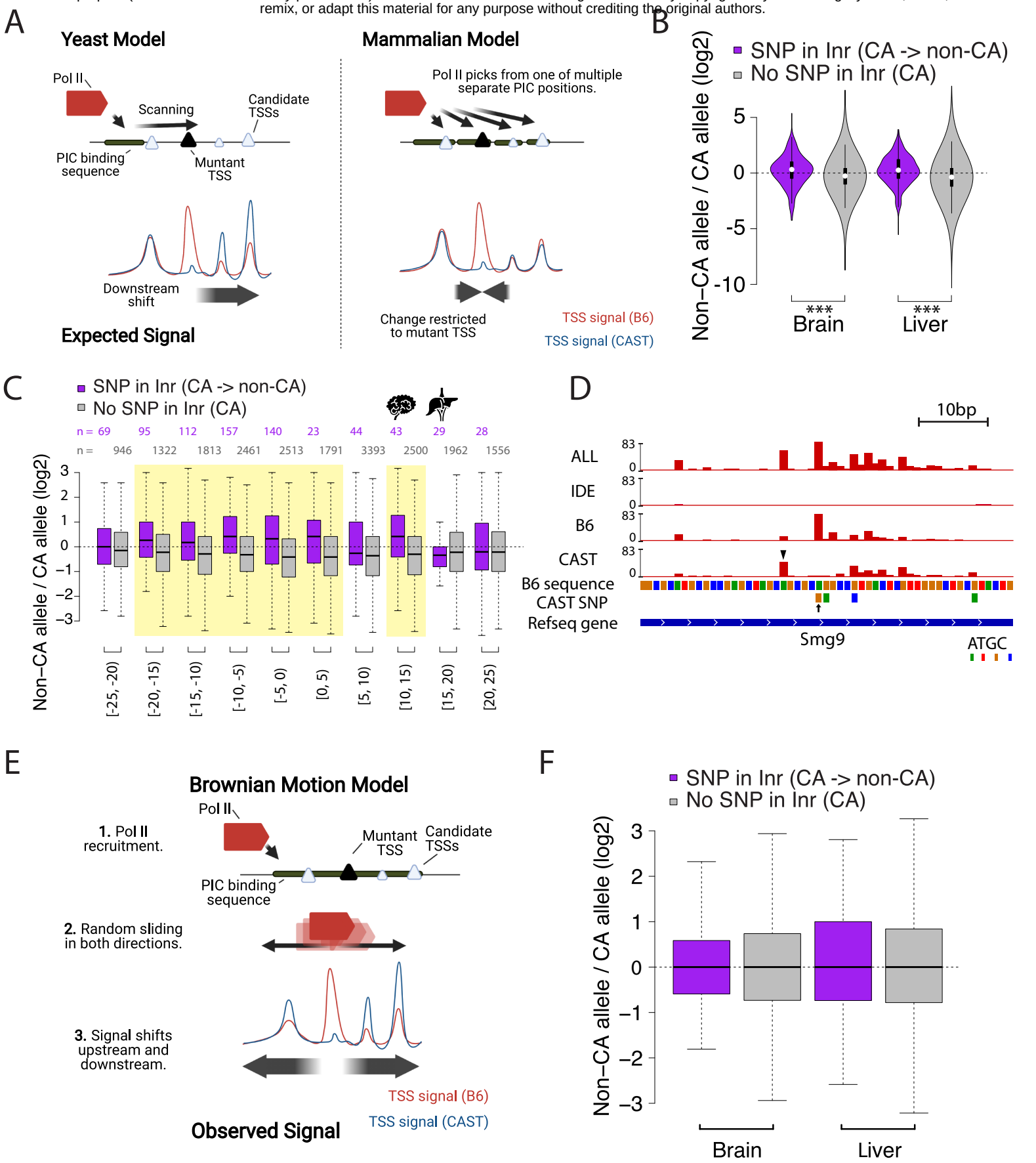


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

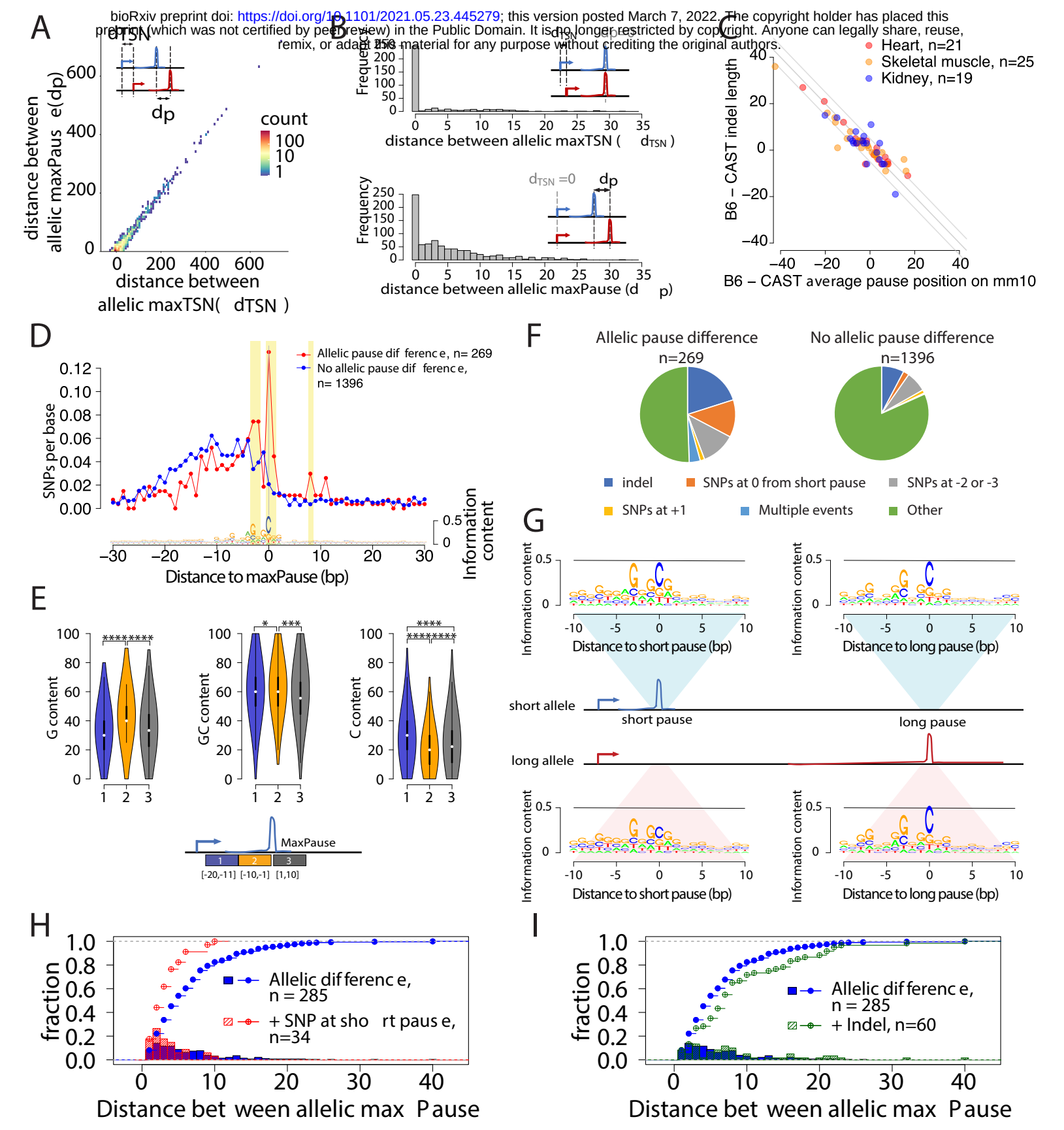
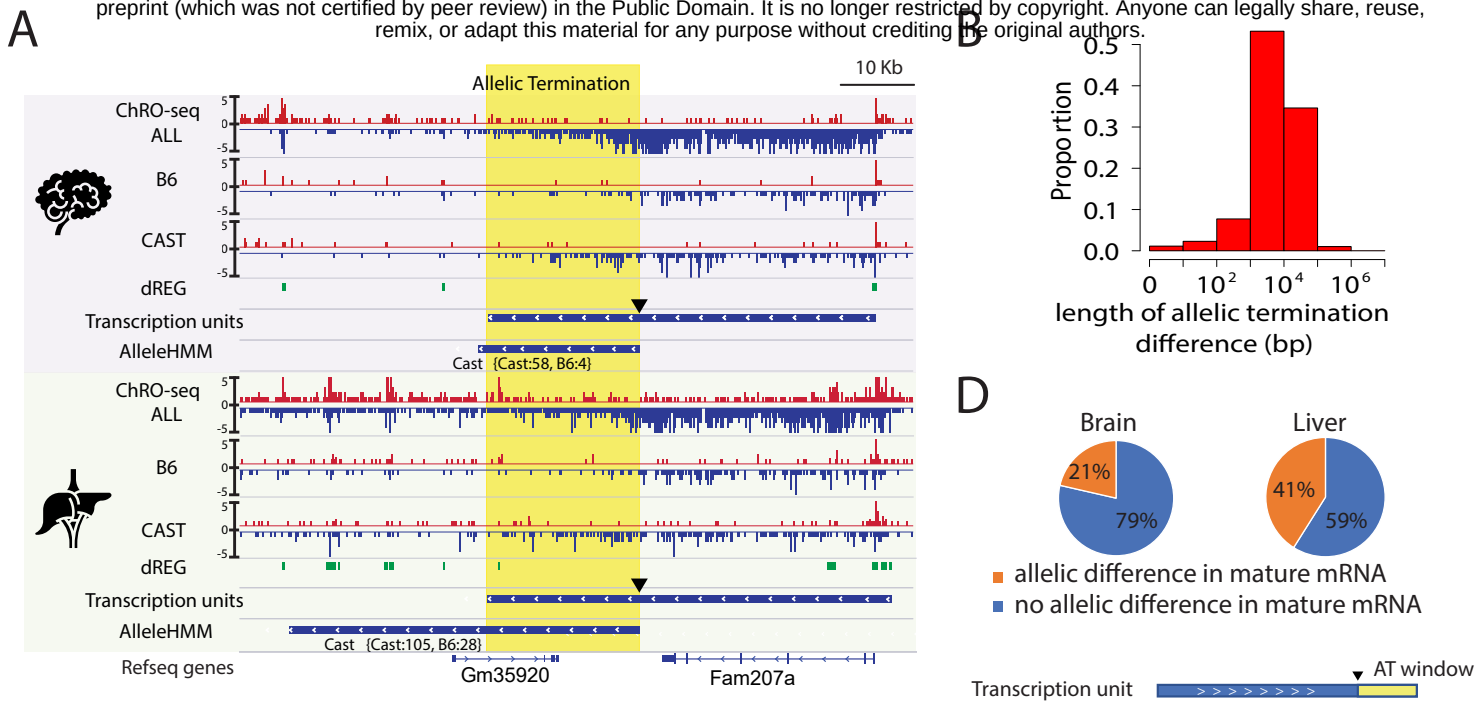


Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.23.445279; this version posted March 7, 2022. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.



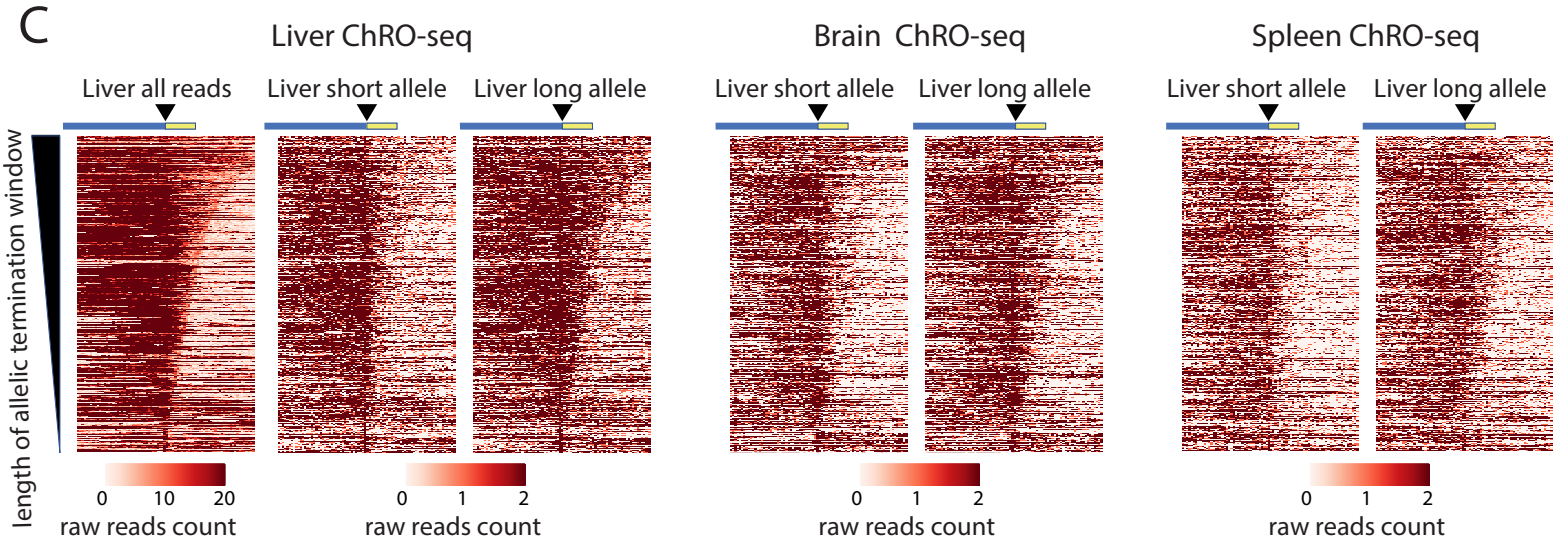


Figure 6