cis Elements that Mediate RNA Polymerase II Pausing Regulate Human Gene Expression

Jason A. Watts,[1] Joshua Burdick,[2] Jillian Daigneault,[3] Zhengwei Zhu,[3] Christopher Grunseich,[4] Alan Bruzel,[3] and Vivian G. Cheung[2,3,5,*]

Aberrant gene expression underlies many human diseases. RNA polymerase II (Pol II) pausing is a key regulatory step in transcription. Here, we mapped the locations of RNA Pol II in normal human cells and found that RNA Pol II pauses in a consistent manner across individuals and cell types. At more than 1,000 genes including *MYO1E* and *SESN2*, RNA Pol II pauses at precise nucleotide locations. Characterization of these sites shows that RNA Pol II pauses at GC-rich regions that are marked by a sequence motif. Sixty-five percent of the pause sites are cytosines. By differential allelic gene expression analysis, we showed in our samples and a population dataset from the Genotype-Tissue Expression (GTEx) consortium that genes with more paused polymerase have lower expression levels. Furthermore, mutagenesis of the pause sites led to a significant increase in promoter activities. Thus, our data uncover that RNA Pol II pauses precisely at sites with distinct sequence features that in turn regulate gene expression.

Introduction

RNA polymerase synthesizes RNA in a highly regulated manner to allow for a wide range of cellular functions.[1–6] One such regulation is the pausing of the RNA polymerase as it moves along the DNA. Dysregulation leads to too low or too high levels of gene expression that can have dire consequences. Human diseases from cancer to kidney disorders and neurodegeneration are known to result from defects in RNA processing.[7–9]

The discontinuous synthesis of RNA by RNA polymerase II was first noted in some selected genes but as advances allowed genome-wide studies, RNA Pol II pausing has been recognized as a general regulatory step. Paused RNA polymerase was found in the 5′ ends of β-globin,[10] *hsp70*,[11] and proto-oncogenes.[12,13] The resultant accumulation of RNA Pol II modulates gene expression at baseline and in response to stimuli. For example, in *Drosophila*, RNA Pol II pauses in the promoter region of *hsp70*, then upon heat induction, these paused polymerases are released and continue into RNA chain elongation.[11] As paused RNA polymerase was identified, studies began to determine how they are regulated. Early studies found two protein complexes, 5, 6-Dichloro-1-β-D-ribofuranosylbenzimidazole Sensitivity Inducing Factor (DSIF)[14,15] and Negative Elongation Factor (NELF)[16] that act *in trans* to retain the RNA Pol II in the promoter regions. In separate studies, a *cis*-element was identified at pause sites in *Drosophila*[17,18] and more recently in human cells.[19]

In parallel, our understanding of on the consequences of dysregulation of gene expression has grown. Those studies have led to therapeutics to restore aberrant gene expression, as exemplified by the antisense oligonucleotide Nusinersen for treatment of spinal muscular atrophy.[20] RNA-based therapeutics is a burgeoning modality. These drugs aim to change the expression levels of target genes. To achieve this goal optimally, it is critical to understand how human cells fine-tune gene expression. Polymerase pausing represents such a step. A few groups have already suggested manipulations of the DSIF subunit, Spt4, as a treatment for ALS and Huntington disease.[7,21] To advance these early proposals into the clinic, a deeper understanding of RNA polymerase pausing in normal human cells is needed. Studies of RNA Pol II pausing have largely relied on cancer and stem cells which are transcriptionally very active; thus, the pattern of pausing may be quite different than that in normal cells. Furthermore, for disease treatments, it is optimal to target RNA Pol II pausing at genes individually rather than manipulating the *trans*-acting protein complexes.

Recently, techniques such as GRO-seq,[22] PRO-seq,[23] NET-seq,[24,25] and Start-seq[26] have allowed the isolation of nascent RNA and the precise mapping of active RNA polymerases genome-wide. These advances greatly facilitate the quantitative assessment of RNA Pol II pausing. Here, we carried out PRO-seq in adult and neonatal skin as well as in kidney cells. We found that the RNA Pol II pauses in a highly regulated manner. In more than 1,000 sites, RNA Pol II paused at the same nucleotide positions among individuals and in different cell types. To identify the *cis*-elements that regulate RNA Pol II pausing, we found that the pause sites are found in regions with high GC content, a 9-mer sequence motif, and are predominantly cytosines. The genes with paused polymerases have lower gene expression levels. Perturbations of the cytosines in the 9-mer motif through natural sequence variants and site-directed mutagenesis show that RNA Pol II pausing decreases

[1]Department of Internal Medicine, Division of Nephrology, University of Michigan, Ann Arbor, MI, USA; [2]Howard Hughes Medical Institute, Chevy Chase, MD, USA; [3]Life Sciences Institute, University of Michigan, Ann Arbor, MI, USA; [4]National Institute of Neurological Disorders and Stroke, National Institute of Health, Bethesda, MD, USA; [5]Department of Pediatrics, Division of Neurology, University of Michigan, Ann Arbor, MI, USA
*Correspondence: vgcheung@med.umich.edu
https://doi.org/10.1016/j.ajhg.2019.08.003.

gene expression. These findings lay the foundation for specific targeting of RNA Pol II pausing to restore aberrant gene expression.

Material and Methods

Cell Culture

Skin fibroblasts from anonymized healthy adults collected as control subjects from an unrelated project[27] and foreskin fibroblasts from a healthy newborn (obtained from the University of Pennsylvania core, SBDRC, see Web Resources) were cultured in DMEM medium (Thermo-Fisher) with 10% fetal bovine serum at 37°C with 5% CO_2. Cells were passaged every 72 h using Trypsin-EDTA (0.05%). HK-2 cells (ATCC) were grown in keratinocyte serum-free media (GIBCO-BRL) 37°C with 5% CO_2, and media were changed 3 times per week. Adult fibroblast tissue samples were collected using a study approved by the NIH Combined Neuroscience Institutional Review Board, and informed written consent was obtained from all participants.

Precision Run-On Sequencing (PRO-Seq)

PRO-seq libraries were prepared as described previously.[23] Briefly, 5 × 10^6 nuclei were added to 2X Nuclear Run-On (NRO) reaction mixture (final concentrations: 10 mM Tris-HCl [pH 8.0], 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 0.03 mM each of biotin-11-A/C/G/UTP [Perkin-Elmer], 0.2 u/μL RNase inhibitor) and incubated for 3 min at 37°C. Nascent RNA was extracted by phenol (Trizol LS)/chloroform and then fragmented by base hydrolysis in 0.2 N NaOH on ice for 15 min. The reaction was neutralized by adding 0.7 × volume of 1 M Tris-HCl (pH 6.8). The fragmented nascent RNA was purified using 30 μL of Streptavidin M-280 magnetic beads (Invitrogen) and ligated with 3′ RNA adaptor (5′p-GAUCGUC GGACUGUAGAACUCUGAAC-/3InvdT/). Biotin-labeled products were recovered by streptavidin beads. For 5′ end repair, in PRO-seq the RNA products were successively treated with 5′ pyrophosphohydrolase (NEB) and polynucleotide kinase (NEB). 5′ repaired RNA was ligated to the 5′ RNA adaptor (5′-CCUUGGCACCCGAGAAUUCCA-3′). The products were further purified by the streptavidin beads. RNA was reverse transcribed using RT primer (5′-AATGATACGGCGACCACCGAGA TCTACACGTTCAGAGTTCTACAGTCCGA-3′). The product was PCR amplified, the resulting amplicons that are between 150 and 250 bp (insert > 70 bp) were purified using the BluePippin (Sage Science) agarose gel electrophoresis, and then sequenced on the HiSeq 2500 instrument (Illumina) to a depth of >150 million reads per sample (see Table S1).

Raw sequencing files were processed by trimming the adaptor sequences from the ends of reads using fastx_clipper from FASTX-Toolkit (Hannon Lab). Sequences with low-quality represented by a stretch of "#" in the quality score string in FASTQ file were removed. Reads that were >35 nt after trimming were included for downstream analysis. Reads were aligned to human reference genome (hg18) using GSNAP[28] (version 2013-10-28) with the following parameters: mismatches < [(read length +2)/ 12-2]; mapping score > 20; soft-clipping on (-trim-mismatch-score = −3). Bam files were generated and data normalized to reads per million mapped reads (RPM). Data were visualized using IGV.[29] For all the analyses, for each gene, we focused on the longest transcribed isoform.

Pausing Index

For each gene, we calculated the pausing index (PI) which is the ratio of normalized PRO-seq reads in a 1-kb window centered on the TSS to that in the rest of the gene with normalization for gene length.[22] We included the 14,503 genes that are more than 2 kb in length and more than 1 kb from another gene on the same strand. For downstream analyses, comparing gene expression to RNA Pol II pausing, we included genes with pausing indicies from 2 to 900.

Pause Sites

In the *in vitro* run-on portion of PRO-seq, RNA Pol II incorporates biotinylated bases and halts chain elongation; thus, the biotinylated bases are found at the 3′ end of each PRO-seq read. Accordingly, by determining the number of PRO-seq reads that ends at each nucleotide position, one obtains the number of RNA Pol II found at that position.[23] We determined the sites where most (top 20%) PRO-seq reads end and marked them as to where the RNA Pol II pauses. We then compared the five samples to identify where the RNA Pol II pauses at the same locations among them. Even though in PRO-seq, the RNA chain elongation is expected to halt upon addition of one biotinylated base, sometimes a few bases are added. To accommodate these extra bases, we allowed up to 3 bases among the samples in our comparisons. 1,367 RNA Pol II pause sites were identified with these criteria. The shared pause sites within 500 bp of an annotated transcription start site are considered as being in the promoter region.

To determine whether the overlap among individuals in these 1,367 pause sites were more frequent than would be expected by chance, we carried out a permutation test. We used the 6,372,215 sites with a PRO-seq read end in at least one of the 5 adult fibroblast samples. The read counts at each site were randomly shuffled within each of the 14,503 genes, keeping the same number of sites per gene as in the experimental data. The permutation was performed 10,000 times. After each iteration, we determined the number of pause sites in the randomized data using the criteria described above where a pause site has reads in the top 20% in all 5 adult fibroblast samples. This procedure gives 10,000 estimates of the number of pause sites that would be found under the null hypothesis that RNA Pol II pausing occurs randomly. The most number of pause sites found in the randomized data was 7 sites (in 2 of the 10,000 permutations), far less than the 1,367 that were found experimentally. Since none of the 10,000 permutations yielded the number of sites that we observed, we rejected the null hypothesis with a permuted p value < 0.0001.

NELF and DSIF Occupancy

Chromatin immunoprecipitation (ChIP) was performed as described previously.[30] Briefly, foreskin fibroblasts were crosslinked with 1% formaldehyde for 15 min. Cross-linking was stopped with 2.5 M glycine for 5 min. Nuclei were isolated by rotating crosslinked cells for 10 min at 4°C in 5 mL lysis buffer 1 (50 mM HEPES [pH 7.6], 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) followed by pelleting, and 10 min rotating in 5 mL lysis buffer 2 (200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 10 mM Tris [pH 8]). Nuclei were pelleted, then swelled in lysis buffer 3 (10 mM Tris [pH 8], 1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, 0.1% deoxycholic acid, 10% N-lauryl sarcosine) for 10 min, then sonicated on high setting (30 s on, 30 s off) for 5 min, 3 times, to shear chromatin to less than 500 bp with Bioruptor (Diagenode). After pelleting the

Table 1. Relative Contributions of the Sequence Feature at 1,367 Locations where RNA Pol II Pauses in Gene Promoters

Feature	Odds Ratio (95% CI)
9-mer Motif	2.02 (1.8, 2.27)
"Cytosine at pause site"	1.42 (1.18, 1.71)
"+1 Purine"	1.21 (1.01, 1.44)
GC skew	1.17 (1.08, 1.28)
GC content	1.14 (1.03, 1.25)

insoluble fraction, the supernatant was pre-cleared with Protein G agarose beads (Sigma) and anti-rabbit IgG (Sigma). 50 μg sheared chromatin was incubated in RIPA buffer (50 mM Tris [pH 8], 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS) with 5 μg rabbit IgG (Sigma), 5 μg NELFA (Santa Cruz) or 5 μg SPT5 (Santa Cruz) and recovered with Protein G agarose beads. Beads were washed twice with low-salt RIPA (150 mM NaCl) and twice in high-salt RIPA (300 mM NaCl), then eluted in 100 μL 1% SDS plus 100 mM sodium bicarbonate. After cross-link reversal, DNA was purified over QIAquick PCR Purification Kit (QIAGEN). Factor enrichment was verified with qPCR at the *HSP70* promoter (Forward-TCCAGTGAATCCCAGAAGACTC, Reverse-CCTGGGCTTTTATAAGTCGTCA) and gene body (Forward-GTTTGAGCACAAGAGGAAGGAG, Reverse-AGGAAATGCA AAGTCTTGAAGC). ChIP-seq libraries were prepared using the Ovation Ultralow Library system (NuGen). Libraries were sequenced on the HiSeq 2500 instrument (Illumina) and ~40 million 100-nt reads were generated per ChIP sample. Sequence pre-processing and alignment were performed as described for PRO-seq. NELFA or SPT5 enrichment compared to input DNA was determined using MACS under default settings.[31] MACS fold-enrichment > 5 with an FDR < 0.05 was considered as positive for factor occupancy.

Sequence Features

To identify a sequence motif, we used Weblogo[32] to analyze the 21 bases that flank each of the 1,367 pause sites. The resulting motif is the 9-mer shown in Figure 2A. Then to assess sequence features around the pause sites, we extracted reference genome sequence (hg18) corresponding to 500 bases upstream and 500 bases downstream of each of 1,367 pause sites. The motif scores were determined using the motifcounter package.[33] Then, GC content was determined as (G+C)/(A+T+C+G) in 100-bp sliding windows. GC skew was determined as (G−C)/(G+C) in 100-bp sliding windows. As background, we extracted sequences flanking 1,367 sites randomly selected from regions where RNA Pol II paused (average pausing index > 2 in fibroblast samples) but not at the same sites across the 5 individuals.

Linear Discriminant Analysis and Effect Size Determination

To determine whether the sequence features and *trans*-acting protein complexes allow us to classify sites as to where RNA Pol II pauses, we carried out conventional linear discriminant analysis with leave-one-out cross-validation (in the statistical package Minitab). The analyses were performed using data on NELFA and SPT5 abundance from ChIP assays, motif score, GC content, GC skew, as well as the presence or absence of cytosine at pause site and +1 purine on all 5 fibroblast samples. Then, to assess the effects of the sequence features on RNA Pol II pausing, we carried out stepwise binary logistic regression (in the statistical package Minitab). GC content, GC skew, and motif score were Z-transformed to allow for comparisons of their effect sizes. The results are reported as odds ratios in Table 1.

Gene Expression Analysis

RNA was isolated using RNeasy Mini-Kit (QIAGEN). Sequencing libraries were prepared from total RNA using TruSeq Stranded Total RNA Library Prep Kit (Illumina). Sequencing was performed on Illumina HiSeq 2500, and >135 million 100-nt reads were generated from each sample. Low-quality bases were trimmed from the 3′ end of reads and 3′ adaptor was trimmed using FASTQ/A Clipper with default settings (Hannon lab). Reads shorter than 35 bp were excluded from the analysis. Sequencing reads were aligned to human reference (hg18) using GSNAP (v.2013-10-28)[28] using the following parameters: mismatches $\leq$ [(read length+2)/12-2]; mapping score $\geq$ 20; soft-clipping on (-trim-mismatch-score = −3). Reads counts from each sample were normalized to the total number of mapped reads. Relative transcript abundance in fragments per kilobase mapped (FPKM) was determined using Cufflinks (v.2.2.1).[34] For analysis of allelic expression, we considered genes (N = 12) where C/T and C/C genotypes were represented at promoter pause sites. Expression of each gene was normalized across individuals by Z-score and then averaged by genotype.

Differential Allelic Pausing

To determine whether there are allelic differences in RNA Pol II pausing, we identified heterozygous sites from the DNA sequences of the five adult fibroblast samples. For each sample, sites with >10 read coverage and at least 25% of the reads showed an alternate base were considered as heterozygous. To avoid reference bias, we used the identified variants to construct an "alternate" genome where the reference alleles at the heterozygous sites were replaced with the alternate alleles. PRO-seq reads were then aligned to the reference and alternate genomes using GSNAP. We considered 134 sites in 83 genes where RNA Pol II pauses (sites in the top 50 percentile) in at least two samples. To assess for differential allelic pausing, the number of PRO-seq read-ends on the "C" versus "non-C" allele was determined and the group means were compared by t test. Results, including the ratio of read-ends on the "C" versus "non-C" allele, are reported in Figure 2H.

We queried the GTEx database for SNPs that overlap the RNA Pol II pause sites. We included SNPs with a minor allele frequency greater than 20%, those that coincide with where RNA Pol II pauses (sites in the top 50 percentile), and are enriched for SPT5 (per ChIP assays). This yielded seven SNPs: rs66966963, rs11547138, rs11248061, rs2303754, rs71476227, rs1049346, and rs35024348. Table 2 lists the genotypes, extent of differential allelic expression by tissue, and p values obtained from the GTEx portal v7 090617.

Luciferase Assay

The promoter region including the first exon for *MYO1E*, *BLCAP*, and *SESN2* were cloned into a Topo TA cloning vector (Invitrogen) following PCR amplification using primers: (*MYO1E*) 5′-GCTAG CTTGCTCACAATCCAGACGTAGG-3′, 5′-CTCGAGCACCCAAGC ACTCACAGGA-3′, (*BLCAP*) 5′-CTTTGAGCCACGAGAAGGTTTT-3′, 5′-CAGGAGTACTATGACCCACCTC-3′ and (*SESN2*) 5′-GCTAGCCT GTGTCTCGCATCTTTGGAG-3′, 5′-CTCGAGGCTTTGGTGCTGG

Table 2. Differential Allelic Expression at Heterozygous Pause Sites

SNP	Gene	Genotype	Tissue	DAE[a] (C versus Alternate Allele)	p Value
rs66966963	*ACTR3B*	C/T	prostate	−17%	4×10^{-5}
rs11547138	*AKIRIN1*	C/T	transformed fibroblasts	−12%	3×10^{-22}
rs11248061	*IDUA*	C/A	skin	−12%	8×10^{-10}
			stomach	−13%	6×10^{-12}
			nerve	−14%	4×10^{-19}
rs2303754	*POP4*	C/G	aorta	−4%	8×10^{-5}
			stomach	−4%	5×10^{-5}
			transformed fibroblasts	−4%	1×10^{-4}
			nerve	−4%	5×10^{-6}
			esophagus (muscularis)	−4%	9×10^{-7}
			breast	−6%	8×10^{-5}
			lung	−7%	1×10^{-6}
			coronary artery	−7%	3×10^{-5}
			cerebellum	−8%	3×10^{-5}
rs71476227	*ZDHHC21*	C/T	esophagus (mucosa)	−10%	2×10^{-9}
			esophagus (GEJ[b])	−13%	3×10^{-10}
			omentum	−14%	2×10^{-16}
			esophagus (muscularis)	−17%	5×10^{-24}
			spleen	−17%	1×10^{-7}
			adipose	−18%	1×10^{-33}
			whole blood	−19%	1×10^{-6}
			skin (leg)	−23%	7×10^{-51}
			skin (suprapubic)	−23%	4×10^{-39}

Data obtained from GTEx portal (see Web Resources). We identified seven SNPs that overlap the pause sites in this study. The five of them that showed significant allelic association with gene expression are shown in this table.
[a]Extent of differential allelic expression
[b]Gastresophageal junction

ACTCTTC-3′. Cloned promoters were verified by Sanger sequencing and then subcloned into pGL4.17 firefly luciferase plasmid (Promega). Point mutations were introduced by using Quickchange II site-directed mutagenesis kit (Agilent) and confirmed by Sanger sequencing. Primers for site-directed mutagenesis are listed in Table S2. 293T cells were cotransfected with 100 ng of pGL4-firefly luciferase and 50 ng pGL4.73-Renilla (Promega) using Lipofectamine 3000 (Invitrogen). Luciferase activity was determined 24 h post-transfection using the Dual Glo-Luciferase assay kit (Promega) and quantified on a Microplate Luminometer (Veritas). Differences in reporter activity were determined by t test.

Results

RNA Polymerase II Pauses at the Same Nucleotide Positions across Individuals and Cell Types

Using the precision nuclear run-on assay (PRO-seq), we determined the locations of transcriptionally engaged RNA Pol II at single-base resolution.[23,35] We carried out PRO-seq in skin fibroblasts from forearms of five adults and focused on genes that are at least 2 kb long and more than 1 kb from adjacent ones, to avoid transcription signals from neighboring genes. In PRO-seq, biotinylated ribonucleotides are used in the run-on assay, incorporation of these nucleotides inhibits the RNA polymerase from further chain elongation. Thus, the RNA Pol II is found at the 3′ end of each nascent transcript, and mapping of these transcripts provides the locations of the RNA Pol II. Accumulation of the polymerase in a region such as the promoter relative to the rest of the gene is used as an indicator of pausing.[22] Two examples of RNA Pol II pausing on *SESN2* and *SLC35D1* are illustrated in Figure 1A.

Visual inspection of the PRO-seq data showed that RNA Pol II pauses at very similar genic locations among individuals, and in some cases, at the same nucleotide positions (Figure 1B). To assess whether this is a general feature of RNA Pol II pausing, we asked whether this is seen more globally. For each individual and in each gene promoter

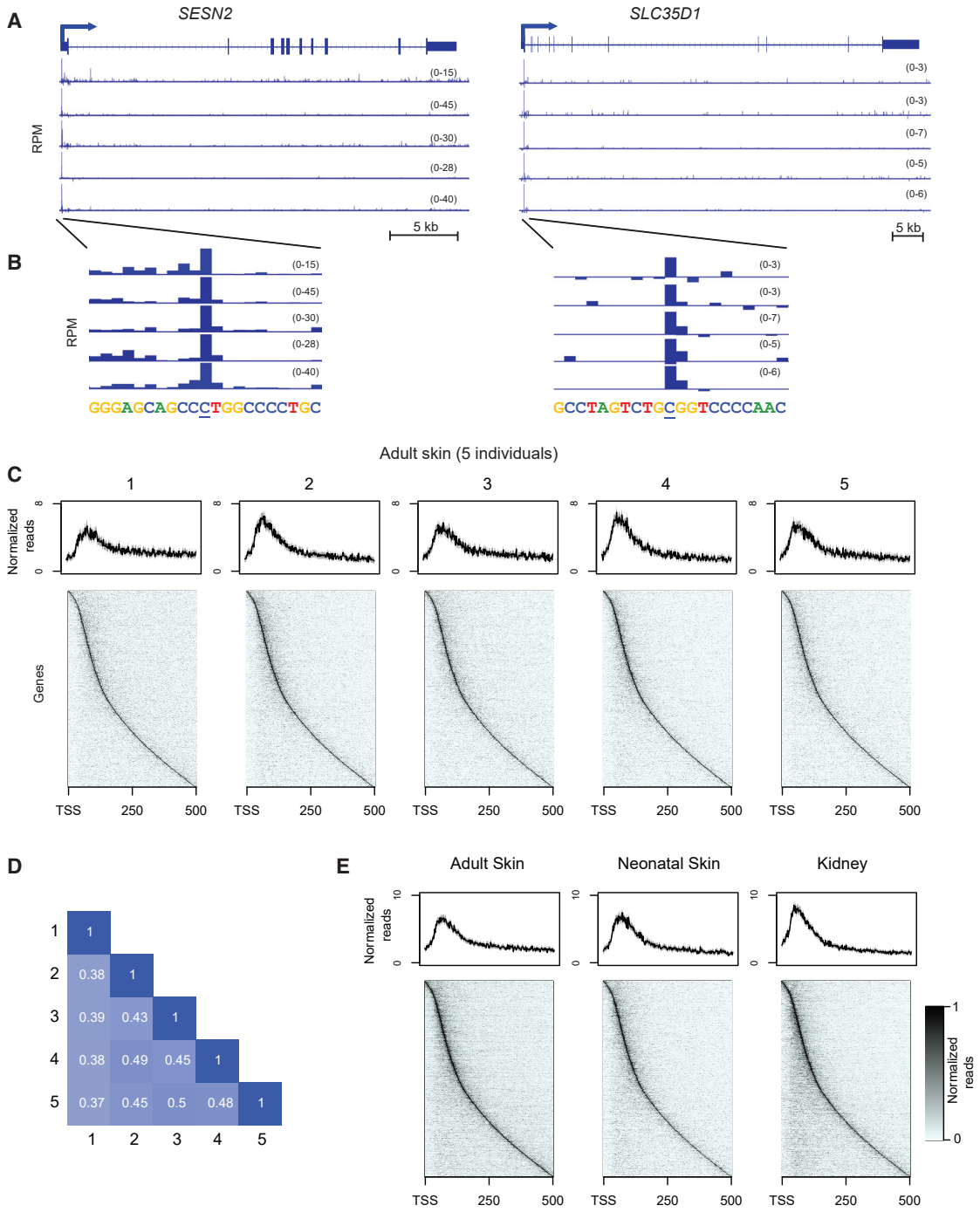

Figure 1. RNA Pol II Pauses at the Same Nucleotide Positions across Individuals and Cell Types
(A) Paused RNA Pol II at two genes in five individuals, highlighting similar RNA Pol II distribution. Scale bar 5 kb. y axis ranges are indicated in parentheses.
(B) 21-bp regions in *SESN2* and *SLC35D1* where polymerases are paused at the same base across individuals.
(C) Top: Profile of RNA Pol II in adult fibroblasts across five individuals. Bottom: Heatmap of the locations of RNA Pol II for 9,320 genes in adult fibroblasts; genes are plotted on the same rows for all individuals to allow direct comparisons.
(D) Pairwise correlation coefficients of the pause maxima between fibroblast samples ($p < 10^{-10}$; Spearman).
(E) Top: Profile of RNA Pol II in adult fibroblasts (average of the five individuals), neonatal fibroblasts, and kidney proximal tubular cells. Bottom: Heatmap of RNA Pol II positions for 7,760 genes; the genes are plotted on the same rows to allow direct comparisons.

region, the base positions with the highest number of paused RNA Pol II (or pause-maxima for short) were determined. We then compared the locations of the pause-maxima among the five individuals. Figure 1C shows across 9,320 gene promoters, RNA Pol II pauses at very similar locations among the five individuals; pairwise correlations (Figure 1D) are highly significant ($p < 10^{-10}$). We then extended the analysis to include skin samples from a

newborn. Figure 1E shows that the locations of paused RNA Pol II in newborn fibroblasts are also very similar to those in the adult samples ($p < 10^{-10}$). Next, we studied proximal tubule cells, HK-2, from the kidney. The locations of paused RNA Pol II for 7,760 genes that are expressed in the adult and newborn skin, as well as kidney cells, are plotted in Figure 1E which shows that across the different cell types, RNA Pol II pauses at very similar locations along the DNA. The correlations of the locations of the pause-maxima across cell types are highly significant ($p < 10^{-10}$). For 1,469 (19%) genes, the pause-maxima are within 3 nucleotides in the different cells.

We have so far focused on sites with the greatest number of paused polymerases. However, numbers that are at the far ends of distributions could result from various biases, so we carried out another analysis that examines not just one site but those in the top 20%. With this definition of paused RNA Pol II, we asked how often are the polymerases paused at the same locations across our five fibroblast samples. We found there were 1,367 sites where RNA Pol II paused at the same nucleotide locations in gene promoters across the five individuals (Table S3). We show these sites on the UCSC Genome Browser (see Web Resources). The precise pausing of RNA Pol II at these 1,367 sites did not occur by chance, as we performed a permutation test and found at most 7 sites (in two permutations) where RNA Pol II paused in the same locations among the five individuals in randomized data ($p < 0.0001$; see Material and Methods). Additionally, we looked at PRO-seq data from other groups and found that even though the studies were carried out in different labs on different cells, the RNA Polymerase paused at the same locations. Specifically, of the 1,367 sites, we found 1,086 sites are shared in common with PRO-seq data from Sistonen and colleagues,[36] while in 3 coPRO datasets from Lis and colleagues,[37] 599, 744, and 877 sites were shared. Collectively, these data show that during transcription, RNA Pol II pauses very precisely.

cis-Acting Elements that Characterize RNA Polymerase II Pause Sites

To determine the code that signals for the RNA polymerase to pause, we analyzed the 1,367 pause sites. The pausing complexes NELFA and SPT5 were found at 1,018 (74%) of these sites (Figures S1A and S1B), which is consistent with their role in mediating promoter-proximal pausing.[15,16,38,39] We then examined and found several sequence features. First, 906 (66%) of the pause sites are cytosines. Second, at 952 sites (69%), the next base to be added to the RNA chain after the pause is a purine; we will refer to these as "+1 purine." Third, there is a 9-mer sequence motif (Figures 2A and 2B; Table S2). Cramer and colleagues[19] used a different approach (NET-seq) to map promoter-proximal pauses and observed a similar motif (Figure S2). Fourth, in the 50-nucleotide regions around the pause sites, the GC-content is very high at 70% (Figure 2C). Fifth, there is GC skewing around the pause sites, indicating G-rich RNA (and G-rich non-template strand; Figure 2D).

Next, we asked whether these features allow us to identify pause sites. To assess this, we carried out a linear discriminant procedure using the 7 factors: the abundance of NELFA, SPT5, a cytosine at pause site, +1 purine, motif, GC content, and GC skew. When combined as a linear discriminant function, they correctly classified 72% of the pause sites. By cross-validation, we left out the site to be classified from the calculation of the discriminant function, and then assigned the site as paused or not based on the discriminant function of the remaining 1,366 sites. With this more stringent criterion, 71.8% of the promoter sites were still correctly classified.

Together these features allow identification of pause sites, but it is also important to know their relative contributions. However, it is very difficult to assess the relative effects with molecular approaches. Epidemiologic studies have identified risk factors and their effects on health conditions from heart disease to cancers.[40–42] Here, we took a similar approach and carried out regression analyses to assess the relative effects of the sequence features on pausing. We performed stepwise regression and found odds ratios range from 1.1 to 2.0 for the motif, "cytosine at pause site," "+1 purine," GC skew, and GC content (see Table 1).

Among these features, the one that is the most amenable for further investigation experimentally on a gene-by-gene basis is the cytosine at pause site. Additionally, this cytosine is conserved; as it was found to mediate pausing in *E. coli*.[43,44] To assess these cytosines, it would be best to focus on the cytosines while controlling other factors that may affect transcription. To accomplish this, we looked for pause sites where our subjects are heterozygous. To include more sites, we broadened the search to sites with paused polymerase in at least 2 (rather than 5) individuals, which yielded 134 heterozygous sites. We compared the number of paused RNA Pol II on the "C-allele" versus the other alleles. In all three comparisons, C versus A, C versus G, C versus T, there were more paused RNA Pol II on the C-alleles. An example is shown in Figure 2E, where at the pause site of *RABL2B*, two individuals are heterozygous for a C/T variant and RNA Pol II paused only on the C-allele. Similarly, RNA Pol II paused more often on the C-allele at the pause sites in *TTLL12* and *FRMD6* (Figures 2F and 2G). On average, RNA Pol II paused three times more often on the cytosine than the other alleles (Figure 2H). Thus, these sequence variants as experiments of nature show that the cytosine in the 9-mer motif is part of the *cis*-regulatory code that governs RNA Pol II pausing. They suggest that these pauses at specific genes can be averted if the cytosines are changed to another base.

Genes with More Paused RNA Polymerase II Have Lower Expression Levels

Next, we investigated the biological implications of RNA Pol II pausing by assessing its effect on gene expression. First, we obtained gene expression levels by sequencing

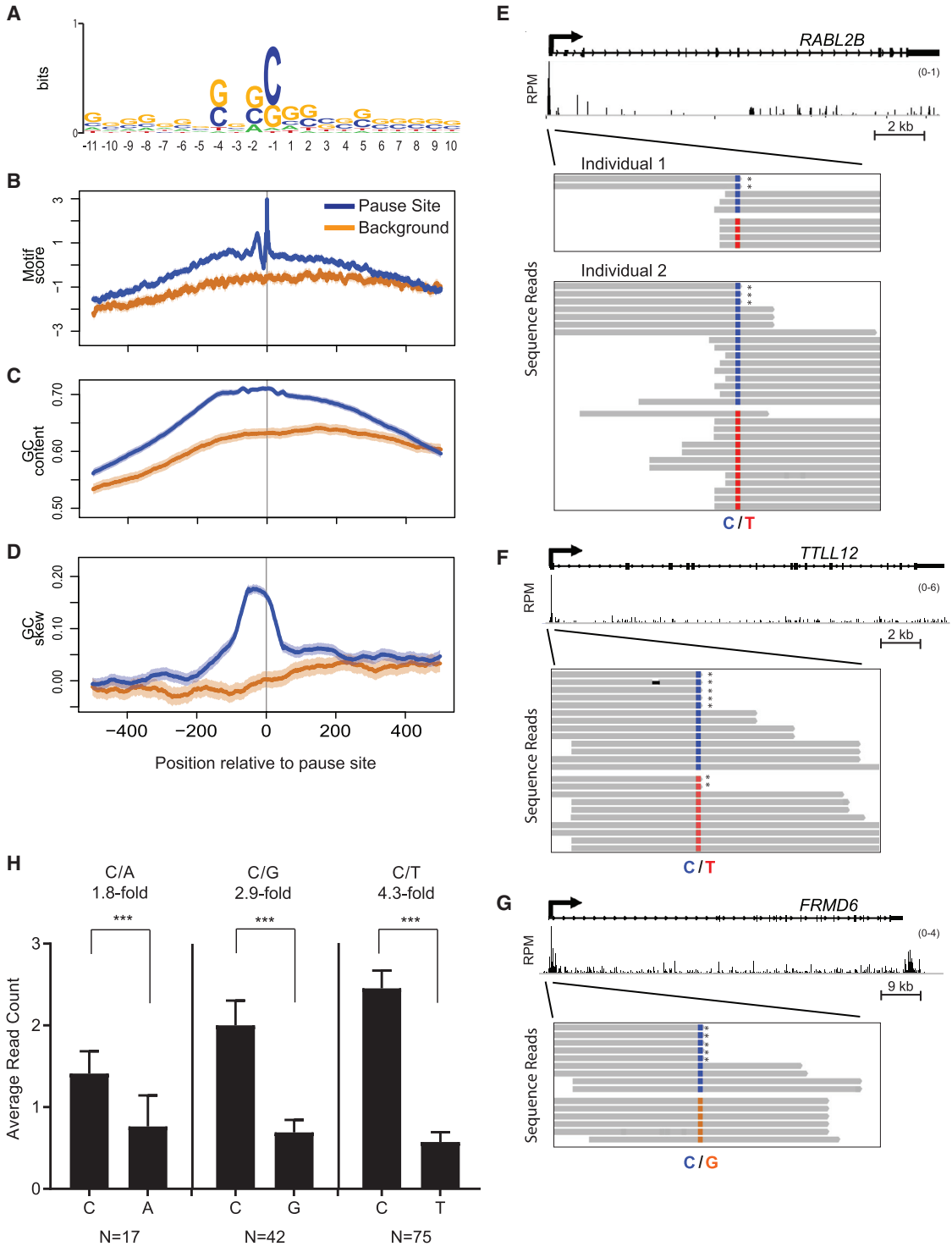

Figure 2. Sequence Features of RNA Pol II Pause Sites

(A) Sequence motif found at pause sites in gene promoters.

(B–D) Plots of motif score (B), GC content (C), GC skew (D) at sites with (blue) and without (orange) paused RNA Pol II.

(E) *RABL2B* gene model (top) and the average profile of RNA Pol II reads among five individuals (middle) are shown. Bottom: An example of differential pausing for two individuals heterozygous (C/T) at a pause site of *RABL2B*.

(F and G) Data presented as in (E), showing differential pausing for an individual at *TTLL12* (F) and *FRMD6* (G).

In (E)–(G), asterisk (*) marks truncated reads where the RNA Pol II have paused; C-allele in blue, T-allele in red. y axis ranges are indicated in parentheses.

(H) Average number of sequence reads that end at each allele of C/A, C/G, and C/T pairs (C versus A ***$p < 0.004$ after removing 1 outlier; C versus G ***$p < 0.0002$; C versus T ***$p < 10^{-12}$, t test). Error bars are the SEM. The extent of differential allelic pausing at cytosine compared to the alternate alleles are indicated.

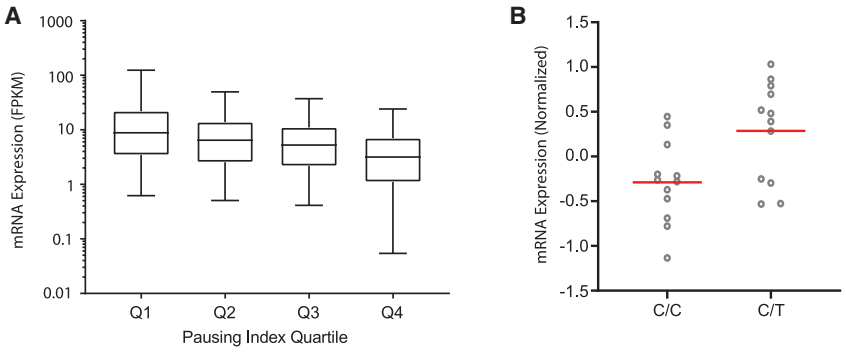

Figure 3. RNA Pol II Pausing Is Negatively Correlated with Gene Expression

(A) Genes with high pausing index have significantly lower expression levels ($p < 10^{-16}$, ANOVA, N = 5,620 genes). Boxes show 25th, 50th, 75th percentiles and whiskers are 5th and 95th percentiles. The ranges of pausing indices by quartiles: Q1 2.5–7.6, Q2 7.6–13, Q3 13–23, Q4 23–803.

(B) Normalized expression levels of genes (n = 12) with C/T variants at promoter pause sites, data for individuals with C/C or C/T genotype are plotted (red bar indicates the averages, $p < 0.006$; t test).

the mRNA from the same adult fibroblast samples that we had determined RNA Pol II pausing. Then to compare gene expression to the level of RNA Pol II pausing, we used the pausing index (PI), which is the abundance of RNA Pol II in the promoter relative to the entire gene.[22] We focused on 5,260 genes with paused RNA Pol II (PI > 2) in their promoters in all 5 individuals. We found that genes with more paused RNA Pol II have significantly ($p < 10^{-16}$) lower gene expression levels (Figure 3A). There is also a significant negative correlation ($R = -0.29$; $p < 10^{-83}$) between the extent of RNA polymerase II pausing and gene expression levels (Figure S3).

Next, we again leveraged sequence variants and assessed whether RNA polymerase pausing leads to lower gene expression. We examined the heterozygous pause sites to assess the effect of the paused polymerase on gene expression. Among our samples, we have the largest number of C/T heterozygotes, so we compared the expression of genes in individuals who are C/C homozygous to individuals who are heterozygous C/T at the promoter pause sites. The results show that gene expression levels are significantly lower ($p < 0.006$; t test) for individuals who have C/C genotypes compared to those with C/T genotypes (Figure 3B), thus confirming that the C-alleles with more paused polymerase are expressed at lower levels. To assess whether this relationship between pausing and gene expression can be generalized, we turned to the gene expression data collected by the Genotype-Tissue Expression Consortium, GTEx.[45] We searched the GTEx database for single nucleotide polymorphisms (SNPs) that overlap the pause sites identified in this study. We do not expect to find many SNPs that overlap our pause sites, but any of them allows us to ask whether findings from our samples can be generalized to a much larger dataset. We indeed found seven SNPs that overlap our pause sites. For five of the seven SNPs, the C-alleles are significantly associated with lower gene expression (see Table 2) across different tissues, beyond the skin and renal tubule cells that led us to the finding. For example, across nine tissue types, individuals with C/C genotypes in *ZDHHC21* have from 10% to 23% lower expression than individuals with T/T genotypes. Therefore, in a dataset with more individuals and cell types, C-alleles at sites where the polymerase is more likely to pause are associated with lower expression levels. Together, the results show that genes with more paused RNA Pol II have lower gene expression levels.

Mutagenesis of Regulators of RNA Polymerase II Pausing Changes Gene Expression

To validate experimentally the effect of RNA Pol II pausing on gene expression, we turned to luciferase reporter assays. We selected three genes, *MYO1E*, *BLCAP*, and *SESN2*, which have promoter-proximal pause sites. The promoters, including the first exons, were cloned into a luciferase reporter, and then by site-directed mutagenesis, at the pause sites, cytosines were converted to thymines. The single-base change from cytosine to thymine resulted in significantly higher promoter activity for *MYO1E* ($p < 0.05$), *BLCAP* ($p < 10^{-8}$), and *SESN2* ($p < 0.001$; Figures 4A–4C).

Next, we assessed other sequence features that were identified as *cis*-regulators. Work in bacteria has shown RNA polymerase pausing is affected not only by sequences at the pause site but also sequences upstream of the polymerase active site where the template and non-template DNA strands re-anneal.[43] Using the *SESN2* promoter, we changed the guanine at the "−11 position" to thymine and found that resulted in significantly ($p < 10^{-7}$; Table S5) higher promoter activity. Our regression analysis suggests sequence changes that deviate from the 9-mer motif would lead to less pausing and higher expression, whereas sequence changes which restore *cis* elements would lead to more pausing and lower expression. We mutated several of the sequences within the 9-mer motif in the *SESN2* promoter and all of them resulted in significantly higher promoter activity (Figure 4D). Additionally, changing the +1 thymine to guanine and thereby creating a +1 purine resulted in lower reporter activity ($p < 0.01$) as is consistent with the expected increase of RNA Pol II pausing (see Table S5). These results confirm that sequences in gene promoters that mediate RNA polymerase pausing regulate gene expression.

RNA Pol II Pausing Affects the Expression of Genes Including Those Mutated in Human Diseases

While our findings uncovered the contributions of sequences to RNA polymerase pausing, they also point to alteration of pausing as a potential treatment of genetic diseases that arise due to aberrant gene expression.

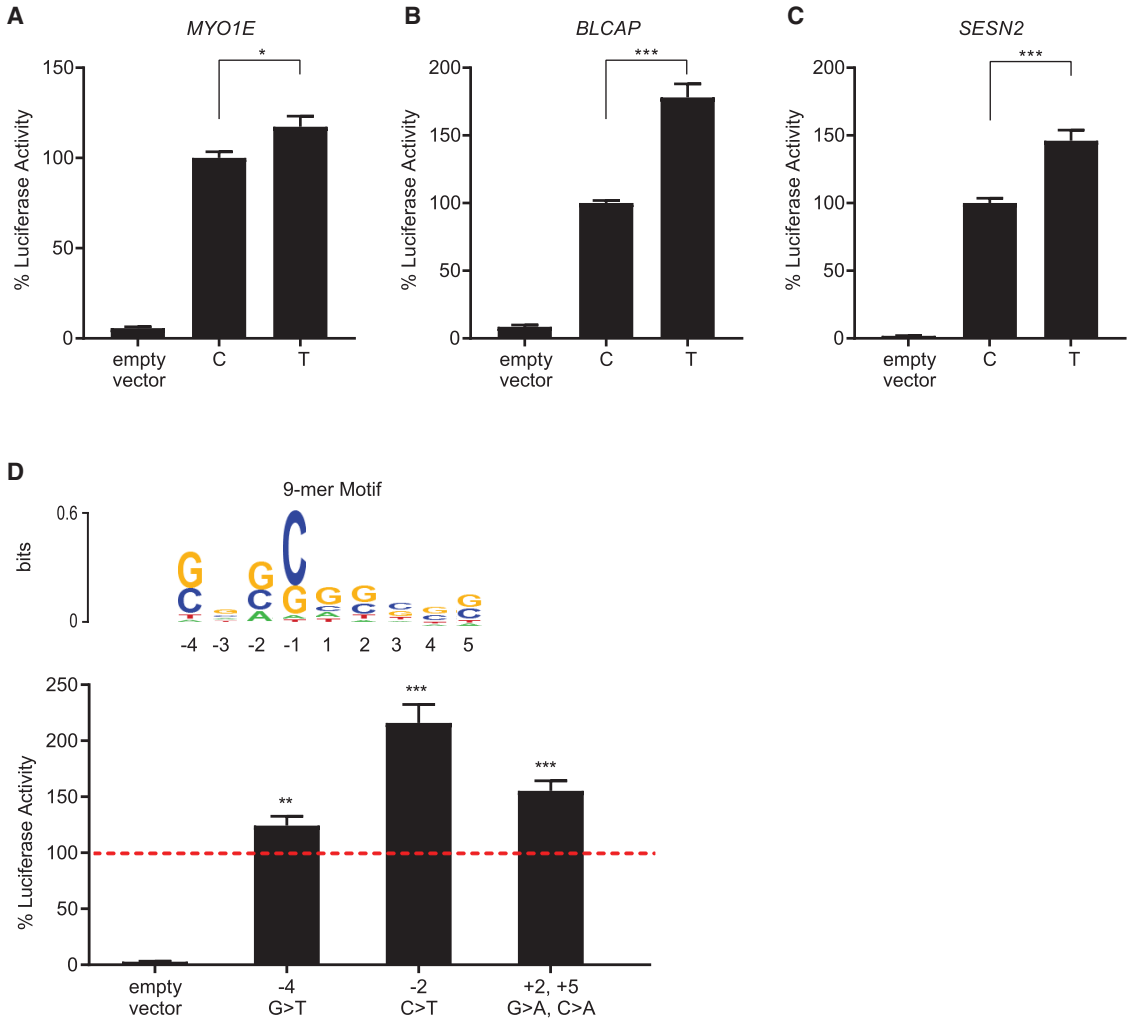

Figure 4. Mutagenesis of Sequences at or near RNA Pol II Pause Sites Changes Gene Expression
(A–C) Luciferase reporter activity of *MYO1E, BLCAP,* and *SESN2* promoters with cytosine at pause sites, compared to those where the pause sites were mutated to thymine (n = 15, *p < 0.05, ***p < 0.001; t test).
(D) Luciferase reporter activity of *SESN2* promoter with the indicated mutations (n = 9, −4T **p < 0.01; −2T ***p < 10^{-5}; +2A5A ***p < 10^{-4}; t test). Red line indicates luciferase activity of the wild-type sequence. The 9-mer motif is shown for reference.
Error bars are SEM.

Therapies have aimed at restoring gene expression since dysregulated gene expression is the mechanistic basis of many diseases. A recent success is Nusinersen that promotes the expression of *SMN2* in spinal muscular atrophy.[20]

The 1,367 pause sites that are characterized in this study are found in 1,141 genes. The Online Mendelian Inheritance in Man (OMIM) database shows that mutations in 347 of these genes (30%) are known to cause human diseases (see Table S6 or as an interactive table see Web Resources). Additionally, some of these mutations have already been shown to affect gene expression (see examples in Table 3). Restoring the expression levels of these genes through RNA polymerase pausing is a potential treatment option. The large number of these disease-causing mutant genes with paused RNA Pol II implies that approaches which target the regulatory sequences identified here could have broad applicability.

Discussion

In this study, we found that RNA Pol II pausing is highly regulated in normal human cells. Regardless of individuals or cell types, at more than 1,300 nucleotide locations of more than 1,000 genes, RNA polymerase pauses very precisely. These large number of sites allowed the identification of *cis* factors that regulate pausing of human RNA Pol II. Alteration of these regulators of RNA Pol II pausing affects expression levels of specific genes. These findings thus provide a basis for altering gene expression levels in mechanistic studies and for the development of expression-based therapeutics.

Aberrant gene expression leads to human diseases, including many single-gene disorders. To understand how gene expression is regulated, studies have yielded complex interactions between RNA polymerase, regulatory protein complexes, and underlying DNA sequences. While

Table 3. Diseases Characterized by Dysregulated Gene Expression

Disease	Gene	Reference
Multiple myeloma	*ELL2*	Li et al.[52]
Elliptocytosis	*EPB41L2*	Moriniere et al.[47]
Monocytopenia and mycobacterial infection syndrome (monoMAC)	*GATA2*	Johnson et al.[53]
Nephrotic syndrome	*KANK2*	Gee et al.[54]
Focal segmental nephrosclerosis	*MYO1E*	Mele et al.[48]
Diamond-Blackfan anemia	*RPL35A*	Noel[55]
Diamond-Blackfan anemia	*RPS19*	Gazda et al.[56]

these findings are elegant, they are also daunting since the regulation appears so complex that one wonders whether gene transcription can be targeted to restore aberrant gene expression as disease treatment. However, the development of Nusinersen for treatment of spinal muscular dystrophy by increasing the expression of SMN demonstrates that gene expression-based therapeutics is possible.[20] The urgent question is how to generalize the knowledge of transcription to develop treatments for other disorders.

The current approach often involves screening antisense oligonucleotides or small molecules for ones that alter the expression of specific genes. Thus, a study has to be designed for each disease. In addition, while these screening methods may yield ways to alter the expression of a gene, they do not provide any mechanistic information. In contrast, we identified DNA sequences that regulate polymerase pausing. These sequences can be targeted to alter the expression levels of more than 1,000 human genes where we find RNA Pol II pausing, including more than 300 genes that are known to be mutated in genetic diseases. Mutations that result in aberrant gene expression could be ameliorated by targeting RNA Pol II pausing as a means to restore gene expression levels, regardless of whether the causal mutation affects polymerase pausing. Gene therapy trials have shown that relatively modest changes in gene expression can be therapeutic, for example, an increase in the expression of Factor IX to about 10% of normal is sufficient in the treatment of hemophilia B.[46] There are other diseases where gene expression is the underlying cause and/or affects severity. For example, mutations in *EPB41L2* that alter splicing result in increased RNA turnover, lower gene expression, and hereditary elliptocytosis.[47] The severity of elliptocytosis is correlated to the expression of *EPB41L2*, so one can posit that a treatment can be developed that aims at deterring RNA Pol II from pausing to increase transcription and therefore gene expression.

Among the genes that we examined in our experimental validations are *MYO1E* and *SESN2*. By mutagenesis, we showed that changing the cytosines at the pause sites to thymine led to higher promoter activities for both genes. Loss-of-function mutations in *MYO1E* results in podocyte injury leading to nephrotic syndrome,[48] whereas overexpression of *MYO1E* can be protective against podocyte injury.[49] Our results here suggest abrogating the RNA Pol II pause would lead to upregulation of *MYO1E* expression, which could be protective against podocyte injury in nephrotic syndromes. Similarly, *SESN2* encodes an antioxidant enzyme that is protective against cellular stress in the liver[50] and is being considered as a target for the treatment of chronic liver disease.[51] This suggests that targeting RNA Pol II pausing may be used not only to correct pathogenic gene expression in Mendelian disorders, but it could also be used to change the expression of genes in chronic diseases.

The ability to alter the expression level of genes specifically is important not only in the therapeutic setting. For mechanistic studies, it is often necessary to manipulate the expression of a gene of interest. Methods such as overexpression and knockdown/knockout of genes often produce expression levels that are too high or too low. Targeting RNA polymerase pausing may allow experiments to be conducted with gene expression changes that are within a more physiologic range.

In conclusion, our study identifies the sequences that regulate RNA polymerase pausing on more than 1,000 human genes and show that these sequences can be altered to affect the expression level of specific genes. Thus, our finding provides a rationale to target RNA polymerase pausing in development of expression-based therapeutics for genetic disorders. Studies to elucidate how the sequence features promote the RNA Pol II to pause will expand our understanding of how nucleic acid sequence and most likely structure regulate transcription.

Accession Numbers

The deep sequencing data reported in this paper are deposited in dbGaP, accession number phs001322.12IV2, and sequence read archive, accession number PRJNA474118.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.08.003.

Acknowledgments

We thank John Lis and his lab for guidance on PRO-seq and insightful discussions. We thank David Levens for suggestions and thoughtful discussions. This work is supported by funds from the Howard Hughes Medical Institute (V.G.C.) and NIH T32 DK007378-38 (J.A.W.).

Received: April 22, 2019
Accepted: August 9, 2019
Published: September 5, 2019

Web Resources

dbGaP, https://www.ncbi.nlm.nih.gov/gap

GTEx, http://www.gtexportal.org/
Pause Sites, http://bit.ly/RNA_PolII_Pause_Sites
Penn Skin Biology and Diseases Resource-based Center, https://dermatology.upenn.edu/sbdrc
Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra
Table S6, http://bit.ly/Disease-Associated_Genes_with_RNA_PolII_Pause_Sites

References

1. Rougvie, A.E., and Lis, J.T. (1990). Postinitiation transcriptional control in Drosophila melanogaster. Mol. Cell. Biol. *10*, 6041–6045.
2. Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.
3. Marshall, N.F., and Price, D.H. (1992). Control of formation of two distinct classes of RNA polymerase II elongation complexes. Mol. Cell. Biol. *12*, 2078–2090.
4. Artsimovitch, I., and Landick, R. (2000). Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. Proc. Natl. Acad. Sci. USA *97*, 7090–7095.
5. Ring, B.Z., Yarnell, W.S., and Roberts, J.W. (1996). Function of E. coli RNA polymerase sigma factor sigma 70 in promoter-proximal pausing. Cell *86*, 485–493.
6. Krumm, A., Hickey, L.B., and Groudine, M. (1995). Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. Genes Dev. *9*, 559–572.
7. Kramer, N.J., Carlomagno, Y., Zhang, Y.J., Almeida, S., Cook, C.N., Gendron, T.F., Prudencio, M., Van Blitterswijk, M., Belzil, V., Couthouis, J., et al. (2016). Spt4 selectively regulates the expression of C9orf72 sense and antisense mutant transcripts. Science *353*, 708–712.
8. Lin, C., Smith, E.R., Takahashi, H., Lai, K.C., Martin-Brown, S., Florens, L., Washburn, M.P., Conaway, J.W., Conaway, R.C., and Shilatifard, A. (2010). AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. Mol. Cell *37*, 429–437.
9. Zhang, G., Liu, R., Zhong, Y., Plotnikov, A.N., Zhang, W., Zeng, L., Rusinova, E., Gerona-Nevarro, G., Moshkina, N., Joshua, J., et al. (2012). Down-regulation of NF-κB transcriptional activity in HIV-associated kidney disease by BRD4 inhibition. J. Biol. Chem. *287*, 28840–28851.
10. Gariglio, P., Bellard, M., and Chambon, P. (1981). Clustering of RNA polymerase B molecules in the 5′ moiety of the adult beta-globin gene of hen erythrocytes. Nucleic Acids Res. *9*, 2589–2598.
11. Rougvie, A.E., and Lis, J.T. (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.
12. Miller, H., Asselin, C., Dufort, D., Yang, J.Q., Gupta, K., Marcu, K.B., and Nepveu, A. (1989). A cis-acting element in the promoter region of the murine c-myc gene is necessary for transcriptional block. Mol. Cell. Biol. *9*, 5340–5349.
13. Bender, T.P., Thompson, C.B., and Kuehl, W.M. (1987). Differential expression of c-myb mRNA in murine B lymphomas by a block to transcription elongation. Science *237*, 1473–1476.
14. Fraser, N.W., Sehgal, P.B., and Darnell, J.E. (1978). DRB-induced premature termination of late adenovirus transcription. Nature *272*, 590–593.
15. Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F., et al. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev. *12*, 343–356.
16. Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell *97*, 41–51.
17. Lee, C., Li, X., Hechmer, A., Eisen, M., Biggin, M.D., Venters, B.J., Jiang, C., Li, J., Pugh, B.F., and Gilmour, D.S. (2008). NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. Mol. Cell. Biol. *28*, 3290–3300.
18. Lee, H., Kraus, K.W., Wolfner, M.F., and Lis, J.T. (1992). DNA sequence requirements for generating paused polymerase at the start of hsp70. Genes Dev. *6*, 284–295.
19. Gressel, S., Schwalb, B., Decker, T.M., Qin, W., Leonhardt, H., Eick, D., and Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. eLife *6*, 6.
20. Finkel, R.S., Chiriboga, C.A., Vajsar, J., Day, J.W., Montes, J., De Vivo, D.C., Yamashita, M., Rigo, F., Hung, G., Schneider, E., et al. (2016). Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study. Lancet *388*, 3017–3026.
21. Liu, C.R., Chang, C.R., Chern, Y., Wang, T.H., Hsieh, W.C., Shen, W.C., Chang, C.Y., Chu, I.C., Deng, N., Cohen, S.N., and Cheng, T.H. (2012). Spt4 is selectively required for transcription of extended trinucleotide repeats. Cell *148*, 690–701.
22. Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.
23. Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.
24. Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.
25. Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526–540.
26. Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., and Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev. *32*, 26–41.
27. Grunseich, C., Wang, I.X., Watts, J.A., Burdick, J.T., Guber, R.D., Zhu, Z., Bruzel, A., Lanman, T., Chen, K., Schindler, A.B., et al. (2018). Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. Mol. Cell *69*, 426–437.e7.
28. Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.
29. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.
30. Watts, J.A., Zhang, C., Klein-Szanto, A.J., Kormish, J.D., Fu, J., Zhang, M.Q., and Zaret, K.S. (2011). Study of FoxA pioneer

factor at silent genes reveals Rfx-repressed enhancer at Cdx2 and a potential indicator of esophageal adenocarcinoma development. PLoS Genet. *7*, e1002277.
31. Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat. Protoc. *7*, 1728–1740.
32. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.
33. Kopp, W., and Vingron, M. (2017). An improved compound Poisson model for the number of motif hits in DNA sequences. Bioinformatics *33*, 3929–3937.
34. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.
35. Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat. Protoc. *11*, 1455–1476.
36. Vihervaara, A., Mahat, D.B., Guertin, M.J., Chu, T., Danko, C.G., Lis, J.T., and Sistonen, L. (2017). Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat. Commun. *8*, 255.
37. Tome, J.M., Tippens, N.D., and Lis, J.T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nat. Genet. *50*, 1533–1541.
38. Yamaguchi, Y., Inukai, N., Narita, T., Wada, T., and Handa, H. (2002). Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA. Mol. Cell. Biol. *22*, 2918–2927.
39. Wu, C.H., Yamaguchi, Y., Benjamin, L.R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H., and Gilmour, D. (2003). NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. Genes Dev. *17*, 1402–1414.
40. Faggiano, F., Lemma, P., Costa, G., Gnavi, R., and Pagnanelli, F. (1995). Cancer mortality by educational level in Italy. Cancer Causes Control *6*, 311–320.
41. Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W.J., Burke, G., McBurnie, M.A.; and Cardiovascular Health Study Collaborative Research Group (2001). Frailty in older adults: evidence for a phenotype. J. Gerontol. A Biol. Sci. Med. Sci. *56*, M146–M156.
42. Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., and Kannel, W.B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation *97*, 1837–1847.
43. Imashimizu, M., Takahashi, H., Oshima, T., McIntosh, C., Bubunenko, M., Court, D.L., and Kashlev, M. (2015). Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. Genome Biol. *16*, 98.
44. Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., and Weissman, J.S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science *344*, 1042–1047.
45. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
46. Nathwani, A.C., Tuddenham, E.G., Rangarajan, S., Rosales, C., McIntosh, J., Linch, D.C., Chowdary, P., Riddell, A., Pie, A.J., Harrington, C., et al. (2011). Adenovirus-associated virus vector-mediated gene transfer in hemophilia B. N. Engl. J. Med. *365*, 2357–2365.
47. Morinière, M., Ribeiro, L., Dalla Venezia, N., Deguillien, M., Maillet, P., Cynober, T., Delhommeau, F., Almeida, H., Tamagnini, G., Delaunay, J., and Baklouti, F. (2000). Elliptocytosis in patients with C-terminal domain mutations of protein 4.1 correlates with encoded messenger RNA levels rather than with alterations in primary protein structure. Blood *95*, 1834–1841.
48. Mele, C., Iatropoulos, P., Donadelli, R., Calabria, A., Maranta, R., Cassis, P., Buelli, S., Tomasoni, S., Piras, R., Krendel, M., et al.; PodoNet Consortium (2011). MYO1E mutations and childhood familial focal segmental glomerulosclerosis. N. Engl. J. Med. *365*, 295–306.
49. Jin, X., Wang, W., Mao, J., Shen, H., Fu, H., Wang, X., Gu, W., Liu, A., Yu, H., Shu, Q., and Du, L. (2014). Overexpression of Myo1e in mouse podocytes enhances cellular endocytosis, migration, and adhesion. J. Cell. Biochem. *115*, 410–419.
50. Park, H.W., Park, H., Ro, S.H., Jang, I., Semple, I.A., Kim, D.N., Kim, M., Nam, M., Zhang, D., Yin, L., and Lee, J.H. (2014). Hepatoprotective role of Sestrin2 against chronic ER stress. Nat. Commun. *5*, 4233.
51. Kim, K.M., Yang, J.H., Shin, S.M., Cho, I.J., and Ki, S.H. (2015). Sestrin2: A Promising Therapeutic Target for Liver Diseases. Biol. Pharm. Bull. *38*, 966–970.
52. Li, N., Johnson, D.C., Weinhold, N., Kimber, S., Dobbins, S.E., Mitchell, J.S., Kinnersley, B., Sud, A., Law, P.J., Orlando, G., et al. (2017). Genetic Predisposition to Multiple Myeloma at 5q15 Is Mediated by an ELL2 Enhancer Polymorphism. Cell Rep. *20*, 2556–2564.
53. Johnson, K.D., Hsu, A.P., Ryu, M.J., Wang, J., Gao, X., Boyer, M.E., Liu, Y., Lee, Y., Calvo, K.R., Keles, S., et al. (2012). Cis-element mutated in GATA2-dependent immunodeficiency governs hematopoiesis and vascular integrity. J. Clin. Invest. *122*, 3692–3704.
54. Gee, H.Y., Zhang, F., Ashraf, S., Kohl, S., Sadowski, C.E., Vega-Warner, V., Zhou, W., Lovric, S., Fang, H., Nettleton, M., et al. (2015). KANK deficiency leads to podocyte dysfunction and nephrotic syndrome. J. Clin. Invest. *125*, 2375–2384.
55. Noel, C.B. (2019). Diamond-Blackfan anemia RPL35A: a case report. J. Med. Case Reports *13*, 185.
56. Gazda, H.T., Zhong, R., Long, L., Niewiadomska, E., Lipton, J.M., Ploszynska, A., Zaucha, J.M., Vlachos, A., Atsidaftos, E., Viskochil, D.H., et al. (2004). RNA and protein evidence for haplo-insufficiency in Diamond-Blackfan anaemia patients with RPS19 mutations. Br. J. Haematol. *127*, 105–113.