# Protocol Variations in Run-On Transcription Dataset Preparation Produce Detectable Signatures in Sequencing Libraries

**Samuel Hunter**
University of Colorado Boulder
**Rutendo Sigauke**
University of Colorado Boulder
**Jacob Stanley**
University of Colorado Boulder
**Mary Allen**
University of Colorado Boulder
**Robin Dowell** ( robin.dowell@colorado.edu )
University of Colorado Boulder

**Research Article**

**Keywords:** Run-on sequencing, PRO-seq, GRO-seq, library preparation

**Posted Date:** June 7th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-571377/v1



**Version of Record:** A version of this preprint was published at BMC Genomics on March 7th, 2022. See the published version at https://doi.org/10.1186/s12864-022-08352-8.

RESEARCH

# Protocol Variations in Run-On Transcription Dataset Preparation Produce Detectable Signatures in Sequencing Libraries

Samuel Hunter, Rutendo F. Sigauke, Jacob T. Stanley, Mary A. Allen and Robin D. Dowell*

*Correspondence:
robin.dowell@colorado.edu
Molecular, Cellular, and Developmental Biology, University of Colorado Boulder, 80301 Boulder, USA
Full list of author information is available at the end of the article

**Abstract**

**Background:** A variety of protocols exist for producing whole genome run-on transcription datasets. However, little is known about how differences between these protocols affect the signal within the resulting libraries.

**Results:** Using run-on transcription datasets generated from the same biological system, we show that a variety of GRO- and PRO-seq preparation methods leave identifiable signatures within each library. Specifically we show that the library preparation method results in differences in quality control metrics, as well as differences in the signal distribution at the 5′ end of transcribed regions. These shifts lead to disparities in eRNA identification. However, these differences do not impact analyses aimed at inferring the key regulators involved in changes to transcription.

**Conclusions:** Run-on sequencing protocol variations results in technical signatures that can be used to identify both the enrichment and library preparation method of a particular data set. These technical signatures are batch effects that limit detailed comparisons of pausing ratios and eRNAs identified across protocols. However, these batch effects have only limited impact on our ability to infer the regulators underlying observed transcriptional changes.

**Keywords:** Run-on sequencing; PRO-seq; GRO-seq; library preparation

## Background

The transcriptome dictates much of a cell's identity and behavior. As such, tracking how transcription patterns change in response to a biological perturbation is a popular approach to understanding molecular regulatory mechanisms. In particular, the population of newly transcribed RNAs provide a readout on the activity and regulation of cellular polymerases. Capturing and mapping these "nascent" transcripts provides a single base-pair resolution readout of the positions of RNA polymerases throughout the genome[1, 2, 3]. A portion of these nascent transcripts arise from enhancer regions and have been associated with transcription factor activity[4, 5, 6]. These enhancer RNAs (eRNAs) are unstable and thus not generally recovered by steady-state assays such as RNA-seq, which sample predominantly from pools of stable transcripts such as mRNAs[7].

To capture all RNAs arising from cellular RNA polymerases, several run-on transcription capture protocols, such as global run-on sequencing (GRO-seq) and precision run-on sequencing (PRO-seq), have been developed[1, 8, 3, 9, 10]. These protocols, collectively known as RO-seq, follow roughly a two step process: first, the

run-on RNA signal must be enriched above the background total RNA; second, the captured RNA must then be converted into a sequencing-ready cDNA library[1]. For the first step, run-on protocols share the same basic strategy, namely use an enrich-able nucleotide as a handle for distinguishing nascent RNA from previously produced RNA (Fig. 1A). Subsequently, sequencing adapters are added and the sample is reverse transcribed and amplified in preparation for sequencing. As these steps are somewhat modular, the process of enrichment is often interleaved with the various steps necessary for library preparation (Fig. 1B).

Similar to distinct RNA-seq library preparation methods, processing RNA through different RO-seq protocols is thought to leave technical artifacts within the library[11, 12, 13]; however, the extent to which these artifacts influence the resulting analysis has not been thoroughly explored. In this study, we sought to identify specific signatures and biases inherent to the protocol (enrichment strategy) and library preparation methods typically employed in RO-seq methods. For this comparison, we generated data from HCT116 cells treated for 1 hour with the p53 activator Nutlin-3a or a DMSO control, a well studied perturbation[4, 14]. Using these matched datasets, we find specific and reproducible biases in each respective dataset that influence both the quality metrics and 5′ distribution of reads. However, we find that these protocol and library specific effects do not strongly impact the inference of which transcription factor is driving the observed perturbation induced changes in transcription.

## Results

### Quality metrics are influenced by RO-seq transcription capture protocols

The ultimate goal of run-on protocols is to produce a dataset that accurately reflects the distribution of actively transcribing RNA polymerase [1, 15] genome wide. However, success in this endeavor depends greatly on the sequencing depth, library complexity, quality of enrichment, and transcription strength of the cell line [16]. To control for cell line differences, we generated run-on libraries from HCT116 cells using a previously employed perturbation strategy[4, 14]. Namely, we used global run-on (GRO) sequencing[1] with a Br-tagged UTP, and precision run-on (PRO) sequencing[2] with a Biotin to mark CTP [3] (Fig. 1A) as enrichment protocols. We then combined these enrichment protocols with one of four library preparation techniques: RNA adapter ligation (LIG)[1], Circularization (CIRC)[8], Random Priming (RPR)[17], or Template-Switching Reverse Transcription (TSRT)[9] (Fig. 1B) on HCT116 cells both with a 1 hr DMSO control and a 1 hr treatment with Nutlin-3a (see Materials and Methods). Samples were subsequently sequenced on an Illumina NextSeq 500 platform (RTA version: 2.4.11, Instrument ID: NB501447) using a single end strategy (50 or 75 bp lengths) to variable depths (summarized in Supplemental Table 1, see Materials and Methods).

The first noticeable difference between any two datasets (even with the same protocol/library preparation) is quality—both in terms of the depth of sequencing and complexity of the library. The depth of our samples range from 20 million reads to 170 million reads. We correct for the disparity in sequencing depth by combining the technical replicates of low-depth samples, and by subsampling deeply sequenced samples. As such, all subsequent comparisons were performed at equivalent depth (with a minimum of 75 million reads).

In contrast, library complexity cannot be corrected for computationally and ideally would be similar between library preparations before comparison. We use two metrics to assess complexity, the number of unique reads relative to the depth of the sample and the number of unique bases covered within the genome (Supplemental Table 1). While most of our libraries were comparably complex, we found that our libraries generated with a random-priming library kit were generally of lower complexity. A survey of previously published run-on datasets in the short read archive suggests that very few utilize random-priming based library preparation[18, 19]. Therefore, it is unclear whether the reduced complexity is a consequence of the library preparation method or a fault of our handling. However, public random primed datasets exhibited similar complexity metrics and read distributions to our dataset (Supplemental Fig. 1, Supplemental Table 1); therefore, we chose to include these libraries in our initial analyses to showcase their technical signatures and potential biases.

Notably, some library preparations result in clearly distinguishable sequence signatures within the acquired reads. In circularization (CIRC) libraries, regardless of the enrichment protocol, RNA is polyadenylated before reverse transcription, and the resulting cDNA is subsequently circularized via the enzyme circLigase[8]. As such, it is common to see many reads with long poly(A) tails before trimming (Fig. 2A). Additionally, the TSRT library preparation adds several C nucleotides to the end of each read[9]. Upon sequencing and adapter trimming, many read inserts showed an increased incidence of C nucleotides near the end of the read (Fig. 2A). In our samples, these sequence signals can effectively distinguish CIRC and TSRT libraries from the other library preparation methods. In contrast, LIG and RPR libraries show similar nucleotide composition across the reads. Likewise, GRO and PRO datasets constructed with matched library preparation methods are not distinguishable from sequence content signatures alone.

However, principal component analysis (PCA) of the read counts over genes tightly clusters based on library preparation and enrichment protocol, suggesting there are additional protocol-distinguishing features not evident in the average nucleotide composition of the dataset (Fig. 2B). Therefore, we next sought to identify whether enrichment quality metrics could be used to distinguish between the protocols. Quality control pipelines offer a way of quantifying steady-state RNA contamination by calculating the ratio of reads over exons and introns for each gene. While the specific value expected for this ratio depends on how reads are counted, a comparatively lower exon-intron ratio is indicative of less mRNA contamination[20]. But is this exon-intron ratio influenced by the choice of protocol? To answer this, we calculated log-normalized exon-intron ratios for every gene in each HCT116 control (DMSO) library. On average, PRO libraries showed a slightly lower amount of mRNA contamination across all genes relative to GRO libraries, consistent with the relative strength of the two enrichment strategies (Fig. 2C). Additionally, both CIRC and LIG libraries showed lower mRNA contamination relative to RPR libraries (Fig. 2D).

Sequence composition (Fig. 2A) can be utilized to identify CIRC and TSRT library preparation protocols with high confidence, while LIG and RPR libraries were more similar in sequence composition, albeit with some differences in complexity and quality metrics (Fig. 2C, Supplemental Table 1). However, the differences between

the enrichment protocols (GRO vs PRO) is less readily apparent from sequence composition or quality metrics alone (Fig. 2A,C, Supplemental Table 1). Yet, we wondered whether systematic signals exist within the data that could distinguish between the protocols. To this end, we applied a discrete wavelet transform (DWT) approach to the normalized coverage of each library (Fig. 2E). The DWT decomposes the signal in a region into low frequency signals (approximation coefficients) that capture consistent RNA polymerase signatures and high frequency signals (detail coefficients) that contain noise. The noise component captures both random noise and systematic noise. Because protocol specific signatures are a systematic source of noise, we reasoned that the high frequency signals may be able to distinguish between the protocols.

To test this hypothesis, we sought to evaluate the DWT on a set of genes where RNA polymerase signatures are the least influenced by library depth or complexity. Thus we identified a set of 210 highly transcribed genes that also had a low coefficient of variation across our datasets. Using the PyWavelets package in python, a symlet wavelet was scanned over the normalized coverage of each gene, effectively decomposing the signal into the two components (see Materials and Methods) (Fig. 2E)[21, 22]. Subsequently, we used principal component analysis (PCA) to cluster the detail coefficients. Overall, 87 genes (41.4%) separated the protocols (GRO vs PRO) directly on the first principle component whereas an additional 107 (50.95%) genes separated the protocols on a different plane within the PC1 and PC2 space (Fig. 2F, Supplemental Fig. 2, 3). These results suggested that the data sets contain a readily identifiable protocol signature. To confirm, we built a simple support vector machine classifier to determine whether the principle components of the wavelet analysis could be used to identify the protocol directly from the data (see Materials and Methods) (Supplemental Fig. 4). Using leave-one-out cross validation at the individual gene level, the classifier correctly identified the protocol >75% of the time (Fig. 2G). Furthermore, a simple majority rules voting scheme applied to the classifier results on all 210 genes correctly identified the protocol every time (100%), further confirming that each data set contains identifiable protocol specific signatures.

### Enrichment and Library Preparation Methods Significantly Shift 5′ Distribution

To better understand the protocol specific signatures within the data sets, we next examined annotated, protein-coding genes for systematic differences in their read distributions. At protein-coding genes, the behavior of RNA polymerase II is well characterized[23] which leads to repeatable patterns of read distribution throughout the gene (Fig 3A). Therefore, we sought to determine whether the protocol (GRO vs PRO) led to systematic differences in the detected 5′ initiation region or the elongation region. Counts across gene body regions suggested that elongation regions correlated well between protocol and library preparation differences (Supplemental Fig. 5, see also Materials and Methods); therefore, we subsequently focused our attention on the 5′ regions of genes.

To assess the differences in the 5′ distribution across protocols, we examined the read distribution of GRO and PRO libraries prepped from DMSO-treated HCT116 cells, with an otherwise similar library preparation protocol (LIG). Metagenes

revealed a shift in coverage near many transcription start sites (TSS) in PRO libraries that is not present in GRO libraries (Fig. 3B, Supplemental Fig. 6, 7). GRO and PRO libraries differ in the nucleotide analog used to enrich for nascent RNA; in PRO-seq, Biotin-NTP is added to the run-on mixture, which terminates transcription upon its incorporation into the nascent RNA. Conversely, in GRO-seq, BrUTP is added, which allows transcription to continue. Previous reports have suggested that changing the enrich-able nucleotide analog from Br-NTP to Biotin-NTP can result in gaps in coverage near TSSs[3]. We theorize that the shift in the 5′ region observed in PRO libraries therefore arises from early incorporation of Biotin-NTP near the TSS which leads to short, truncated reads that are not well mapped. As such, we reasoned that generating new libraries with a different ratio of Biotin-NTP/NTP in the initial run-on mixture would result in more reads captured around the 5′ end (Supplemental Fig 8). Metagenes indeed show a smaller shift with lowered Biotin-NTP concentration, although GRO-LIG libraries continued to show more signal in these regions than any PRO library.

We next reasoned that the observed differences in the detected 5′ read distribution at genes would commensurately affect the pausing index (PI), measured as the ratio of reads in the initiation region relative to the gene body[24]. We defined the initiation region as 50 bp upstream from the annotated TSS to 250 bp downstream of the TSS; gene body regions were defined as 251 bp downstream of the TSS to the annotated cleavage site. Using these regions, we calculated the PI for the longest isoform of each gene in both libraries. Consistent with our findings above, PI for individual genes were reasonably consistent across replicates (Supplemental Fig. 9, Spearman $R = 0.94$) but showed significant disparities between GRO and PRO libraries (Fig. 3C, $R = 0.59$, $p < 2.2e\text{-}16$). Spearman rank correlations for PI in both libraries were marginally higher ($R = 0.73$, $p < 2.2e\text{-}16$). While the PI is known to depend on the method is used to define the paused region[15], we found that the trends across protocols remained consistent even with different pause window calculations and counting software (Supplemental Fig. 10).

To ensure that our findings generalize to other data sets, we next examined publicly available datasets. While these data sets likely have larger batch effects arising from their preparation in distinct laboratories and cell types, we reasoned that the overall trend in 5′ end patterns should still be noticeable, albeit subject to more variance. GRO and PRO libraries obtained from other labs showed similar metagene distributions as observed in our data (Supplemental Fig. 11). In accordance with these findings, publicly available GRO and PRO libraries also showed similar PI distributions when compared with our libraries (Supplemental Fig. 11). We thus conclude that the underlying protocol-induced bias in 5′ ends remain detectable and consistent despite inherent batch effects.

Next, we evaluated the effects of library preparation on the 5′ end. To accomplish this, we constructed metagene summaries of our GRO-CIRC, GRO-LIG, and GRO-RPR libraries (Fig 3D). While CIRC and LIG libraries showed a similar distribution near the 5′ end, GRO-RPR libraries show a shift in coverage similar to PRO libraries. While it is unknown what leads to this shift, we theorize that random priming may have a length bias that is a contributing factor (i.e. the longer a RNA is the more likely a primer is to anneal to it).

Additionally, we found that the pause ratio is sensitive to which method is used to prepare the RNA. We compared pause index calculations for GRO-CIRC and GRO-LIG libraries. We found that, for each gene, pause indices tended to be larger for GRO-CIRC libraries compared to GRO-LIG libraries (Fig. 3E, $R = 0.57$, $p < 2.2e\text{-}16$). To assay whether this shift was systematic, we also computed the Spearman rank-correlation for these indices. Rank correlation between GRO-LIG and GRO-CIRC libraries was stronger than Pearson correlation; however, there were still many genes that showed disparate rankings across our datasets (Fig. 3E, $R = 0.77$, $p < 2.2e\text{-}16$).

### Changing library enrichment methods shifts intergenic read distributions and active enhancer detection

The bidirectional transcription typical of RNA polymerase initiation regions at the 5′ end of genes is also present at enhancers[25], albeit typically at much lower transcription levels. Therefore, we asked whether the patterns of enhancer transcription varied across protocols or library preparations. As a first pass inquiry that avoids reliance on enhancer annotations, we first compare the fraction of reads recovered from RefSeq annotated gene regions to reads recovered in intergenic regions for each data set. To ensure more statistical rigor, we included several publicly available datasets of different cell lines, along with six of libraries we generated from MCF10A cells prepped with PRO-TSRT (See Supplemental Table 1). When comparing GRO and PRO libraries (irrespective of cell type or library preparation method), we found that GRO libraries showed significantly more reads over gene regions compared to PRO libraries (Fig. 4A, $p < .01$, See Materials and Methods). Conversely, we found no significant differences when comparing library preparation methods (Fig. 4B).

The disparity in the gene-to-intergenic reads ratio in GRO and PRO libraries suggest their respective enrichment strategies may capture signal in unannotated regions at different rates. In particular, we were curious whether the capture of eRNAs would be affected by the choice of protocol. To investigate this possibility, we first examined annotated enhancers in the HCT116 cell line acquired from the FANTOM database (converted to hg38 coordinates using the online UCSC tool liftOver)[26]. The level of transcription between these enhancers was largely consistent between our datasets (Supplemental Fig 12). However, FANTOM annotated enhancers represent the comparatively stable enhancer transcripts arising from Cap Analysis Gene Expression (CAGE) data[27].

Therefore, we next sought to identify enhancers directly from the data using their characteristic bidirectional transcription signal[28]. Two algorithms have been developed to identify transcribed regulatory elements based on their bidirectional signal, dREG[29] and Tfit[30]. We employed both methods to annotate sites of bidirectional transcription in our GRO-CIRC, GRO-LIG, and PRO-LIG libraries. Strikingly, the identified regions varied substantially across protocol and library preparation for both algorithms (Supplemental Fig. 13). We hypothesized that these differences may be exaggerated by the sequencing depth, as eRNAs are lowly transcribed and therefore these regions are only consistently detectable at high sequencing depth. To this end, we combined replicates for PRO-LIG libraries to an effective depth of approximately 300 million reads, and replicates of GRO-CIRC

libraries to an effective depth of approximately 400 million reads. Transcribed regions identified in these combined libraries remained inconsistent; while many strong enhancers were called in both of these two deep data sets, other regions were exclusively found in only one (Fig. 4C).

This suggested the existence of transcribed regions whose signal is strongly dependent on the underlying experimental protocol. To confirm this possibility, we next sought to identify the set of transcribed regions with apparent differential transcription across protocols or library preparations. To compare enrichment protocols, we combined Tfit regions from PRO-LIG and GRO-LIG libraries (Fig 4D, Supplemental Fig. 14, see Materials and Methods), while library preparation methods were compared by combining Tfit regions from GRO-LIG and GRO-CIRC libraries (Fig 4E). In every case, regions were combined using *muMerge*[6] and differential read signal was assessed with DESeq1 analysis (Fig. 4F). We then constructed metagenes from set of regions with differential signal (Fig. 4G, H, Supplemental Fig. 15) and observed strong bidirectional signal in only one of the two datasets, while the other dataset showed signal only slightly above background. Manual inspection confirmed that these transcribed regions were only effectively captured by one library, even at high depths (Supplemental Fig. 16).

### Biological response to p53 activation is preserved across run-on transcription capture protocols

The protocol-specific nature of both pausing ratios and eRNA recovery led to concerns about whether the choice of experimental preparation influences commonly conducted downstream analyses, such as identifying which genes respond to a perturbation[4] and which transcription factors drive those changes[31, 32, 5, 6]. As such, we used the competitive MDM2 inhibitor Nutlin-3a, which has a known, specific, robust transcription response in human cells induced by the subsequent activation of the transcription factor p53[33, 4, 14].

First, we sought to determine the reproducibility of detecting differential gene transcription within our libraries. The precise identity of which genes respond to 1 hour of p53 activation is expected to vary across protocols and library preparations – as similar batch effects have been observed for RNA-seq libraries[34]. Thus, we focused specifically on whether the core p53 response program, i.e. the known targets of p53, was captured efficiently in each dataset. To this end we utilize the Gene Set Enrichment Analysis (GSEA) - Preranked[35, 36] tool on ranked, signed p-values obtained from DESeq2[37] (See Materials and Methods). Additionally, we expected that a substantial amount of variation between two libraries generated from different protocols would arise from the gene initiation region (Fig. 3). To confirm this, we subsequently examined two distinct methods of calculating differential gene transcription: the commonly used elongation-region-only approach and the full annotated gene region (Fig. 5A). Across all libraries and counting methods, the p53 pathway was the top hit in the GSEA-Preranked module (FDR q-val $< 0.001$, Fig. 5B, Supplemental Fig. 17), suggesting that each protocol, library preparation and counting method was capable of detecting the underlying biological perturbation in spite of technical signals introduced by protocol differences.

Next, we compared the correlation of the ranks of the genes in the Hallmark p53 pathway used by GSEA. We found that the majority of enriched genes were common

between each of the libraries (58.3% in GRO-LIG vs GRO-CIRC, 57.1% in GRO-LIG vs PRO-LIG) (Fig. 5B,C, Supplemental Fig. 18). However, there remained several genes that were only enriched in one of libraries. When only the elongation region was considered, the overlap improved (68.3% in GRO-LIG vs GRO-CIRC, 58.9% in GRO-LIG vs PRO-LIG), consistent with the 5′ initiation regions being the most variable portion of the gene between protocols. Perhaps unsurprisingly, excluding the 5′ initiation regions is the most common method of assessing differential transcription from run-on sequencing protocols[38, 39, 40, 41].

The second common use of run-on sequencing data is to infer which regulators are driving observed patterns of differential transcription[25, 29, 5]. Alterations in transcription factor activity can be detected by changes in the locations and levels of sites of bidirectional transcription[5, 6], the majority of which reside at enhancers[28]. Therefore we next sought to determine whether the alterations observed in eRNA detection (Fig. 4) impacted TF activity inference[6].

To this end, we used the Transcription Factor Enrichment Analysis (TFEA) tool to evaluate which transcription factor motifs are enriched at sites with altered transcription levels in response to Nutlin-3a[6]. In all cases, TFEA correctly identifies the p53 family (TP53, TP63, and TP73) as significantly upregulated, independent of the protocol and library prep used to generate the dataset (Fig. 5E,F, Supplemental Fig. 19). Upon closer inspection, 92.38% of p53-responsive enhancers responded similarly across protocols, but 7.62% of p53-responsive enhancers were unique to a particular protocol (Supplemental Fig. 16, 20).

## Discussion

We used multiple protocols and library preparations on HCT116 cells exposed to Nutlin-3a and determined that these experimental choices influences the signal of run-on sequencing libraries in systematic and predictable ways. The shape of the characteristic gene initiation peak is strongly influenced by the underlying protocol, while the signal at gene elongation regions remain largely consistent across protocols. Likewise, the recovery of many intergenic regions was protocol specific, even when at high depth. Despite these differences, the ability to detect p53 activation was unaffected by the choice of enrichment or library preparation protocol.

Promoter proximal pausing is a pervasive feature of RNA polymerase II activity[15]. Pausing is often quantified through calculations of the pausing index, the ratio of reads within the initiation region relative to the elongation region. While PI values are known to depend on the choices of windows used to define these regions[15], our work demonstrates that they also depend on the underlying protocol even when the details of the PI index calculation are held constant. Notably, genes sometimes appear to have an additional pause site downstream of the annotated TSS (Fig. 3E)[42]. However, we have found that these second pause sites are protocol dependent; as changes in the library preparation method shift or ablate the signal of this second peak. Collectively, our work indicates that read patterns at the 5′ end of genes are simply not comparable across protocols and library preparations. Additionally, our work indicates that larger conclusions about promoter proximal pausing must be reproduced with a distinct protocol to ensure they reflect the biology and not the technical patterns of a particular protocol.

Given the uniform activity of RNA polymerase II[43], the observed 5′ end protocol specific patterns we observed at genes should also impact enhancer associated transcripts. Indeed, we observe that some enhancers with relatively high read coverage in one library are not detectable using a different protocol. The most highly transcribed eRNAs (e.g. those annotated by FANTOM) are detected equally well by each protocol, but many eRNAs are lowly transcribed. We were surprised that increased depth did not resolve many of these protocol specific eRNAs.

This disparity in eRNA signal raises an intriguing question: which aspects of the protocols and resulting libraries contribute to the difference in eRNA capture rates? The slightly higher exon to intron ratio (Fig. 2D) of GRO-seq suggests this protocol contains a higher level of contaminating mRNA[44], consistent with Br-UTP antibody enrichment being a less efficient pull down method than Biotin-streptavidin enrichment. This bias may also explain why GRO-seq has a higher gene to intergenic ratio compared to PRO-seq (Fig. 4A). The use of Biotin halts polymerase elongation in PRO-seq, giving it a higher precision on RNA polymerase position[2]. However, this also results in short, unmappable fragments near the 5′ end of transcripts, which may limit the ability of PRO-seq to capture some eRNAs. Likewise, other factors probably contribute to the recovery of eRNAs[45], including sequence composition and biological variability.

Despite the observed protocol specific differences, our downstream analysis was consistent in detecting the underlying p53 perturbation. At genes, it is customary to exclude the initiation peak from differential gene transcription analysis[38, 39, 40, 41], and our work indicates this is a wise choice, as counting reads only over elongation regions gave more consistent results across the protocols. Yet even when using only elongation regions, the specific details of which genes respond varied between protocols, but the overall pattern of regulating the p53 pathway was consistent. Likewise, despite the differences in eRNA detection observed across protocols, p53 was readily detectable as the transcription factor activated by Nutlin-3a. Importantly, the observation that across protocols large scale conclusions are consistent but the specifics of individual genes (and eRNAs) are not fully consistent is not unique to run-on sequencing, as similar issues of reproducibility have been observed with RNA-seq protocols[46, 12].

## Conclusion

Protocol and platform differences have long been recognized as batch effect variables that introduce non-trivial experiment specific signals within high throughput sequencing data[47, 48]. Numerous efforts have focused on correcting batch effects, but it is always difficult to do so without some loss of biological signal[49, 50]. On the other hand, the distinct signals we detect raise an intriguing possibility that protocol and library preparation information can be inferred directly from the data itself. The noise component of the data can reliably differentiate between GRO- and PRO-seq datasets with remarkable accuracy, while sequence and quality signatures can often identify the library preparation methods used to prepare the dataset. Thus an automatic detection approach could be built to confirm or correct experimental information within the short read archive, at least for run-on assays[51]. Regardless, knowing the experimental details and managing associated batch effects is necessary when comparing in house data to previously published data sets.

## Materials and Methods

### Cell Culture Conditions

HCT116 and MCF10A cells were cultured in DMEM media supplemented with 10% FBS, 100 units/mL penicillin and 100 $\mu$g/mL streptomycin, at 37°C with 5% CO2. Cells were grown to a confluency of 60-70% in 15 cm culture dishes before passaging. Cells were passaged twice before harvesting, using PBS to wash and 0.05% w/v trypsin to detach the cells from the plate. Cells were aspirated and treated with media containing 10 $\mu$M Nutlin-3a (or DMSO) for 1 hour before harvest.

### Nuclei Isolation

Post-treatment, cells were placed on ice and washed three times with ice-cold PBS. Cells were incubated on ice in 10 mL ice-cold Lysis Buffer (10 mM Tris-HCl pH 7.5, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 2 U/mL SUPERase-IN, brought to volume with 0.1% DEPC DI-water, filtered before use) for 10 minutes. Cells were scraped and collected into 50 mL Falcon tubes, and centrifuged with a fixed-angle rotor at 1000 x g for 10 minutes at 4°C. Cells were resuspended with Lysis buffer with a wide-opening P1000 tip, and washed twice with 10 mL Lysis buffer (centrifuged at 1000 x g for 5 minutes at 4°C). After the second Lysis buffer wash, the samples were resuspended with 1 mL Freezing Buffer (50 mM Tris-HCl pH 8.3, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, brought to volume with 0.1% DEPC DI-water, filtered before use). Nuclei were centrifuged at 1000 x g for 5 minutes at 4$^{c}irc$C, and resuspended with 500 $\mu$L Freezing Buffer. Nuclei were then centrifuged for 2 minutes at 2000 x g, 4°C, and resuspended in 110 $\mu$L Freezing Buffer. 10 $\mu$L was retained for counting nuclei, while the remaining sample was snap-frozen in liquid nitrogen and stored at -80°C until use.

### GRO-seq and Library Preparation Methods

*Ligation (LIG)*

Run-on reactions were performed as in [1]. In brief, ice-cold isolated nuclei (100 $\mu$L) were added to 37°C 100 $\mu$L reaction buffer (Final Concentration: 5 mM Tris-Cl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 10 units of SUPERase In, 0.5% sarkosyl, 500 $\mu$M rATP, rGTP, and Br-UTP, 2 $\mu$M rCTP). The reaction was allowed to proceed for 5 min at 37°C, followed by the addition of 23 $\mu$L of 10X DNAseI buffer, and 10 $\mu$L RNase free DNase I (Promega). RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 $\mu$L GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 $\mu$L of DEPC-treated water. Libraries were prepared as in [1]. In brief, nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using Anti-BrdU beads and ligated with reverse 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/), and BrdU-labeled products were enriched by a second round of Anti-BrdU bead binding and extraction. For 5′ end repair, the RNA products were treated with tobacco acid pyrophosphatase (Epicenter) and T4 polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′ UGGAAUUCUCGGGUGCCAAGG) before being purified by a final round of Anti-BrdU bead binding

and extraction. RNA was reverse transcribed using 25 pmol of RP1 primer (5′AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) and size selected with 1X AMPure XP beads (Beckman) before being sequenced.

*Random Priming (RPR)*

Run-on reactions were performed as in [1]. In brief, ice-cold isolated nuclei (100 $\mu$L) were added to 37°C 100 $\mu$L reaction buffer (10mM Tris-Cl pH 8.0, 5 mM $MgCl_2$, 1 mM DTT, 300 mM KCl, 20 units of SUPERase In, 1% sarkosyl, 500 $\mu$M ATP, GTP, and Br-UTP, 2 $\mu$M CTP). The reaction was allowed to proceed for 5 min at 30°C, followed by the addition of 23 $\mu$L of 10X DNAseI buffer, and 10 $\mu$L RNase free DNase I (Promega). RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 $\mu$L GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 $\mu$L of DEPC-treated water. Libraries were prepared based on the NEBNext Ultra II Directional Library Preparation Kit. In brief, nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using Anti-BrdU beads (Santa Cruz Biotech, Santa Cruz, CA) 3 times. Samples were reverse-transcribed using random hexamers, and sequencing adapters added by PCR. The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) and size selected with 1X AMPure XP beads (Beckman) before being sequenced

## PRO-seq and Library Preparation Methods

*Ligation (LIG)*

Run-on reactions were adapted from [3]. In brief, ice-cold isolated nuclei (100 $\mu$L) were added to 37°C 100 $\mu$L reaction buffer (Final Concentration: 5 mM Tris-Cl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 10 units of SUPERase In, 0.5% sarkosyl, 125 $\mu$M rATP, 125 $\mu$M rGTP, 125 $\mu$M rUTP, 25 $\mu$M biotin-11-CTP (additionally, one library generated with 2.5 $\mu$M biotin-11-CTP, 25 $\mu$M rCTP). The reaction was allowed to proceed for 5 min at 37°C. RNA was extracted twice with Trizol, washed once with chloroform, and precipitated with 3 volumes of ice-cold ethanol and 1-2 $\mu$L GlycoBlue. The pellet was washed in 75% ethanol before resuspending in 20 $\mu$L of DEPC-treated water. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using streptavidin beads and ligated with reverse 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/), and biotin-labeled products were enriched by a second round of streptavidin bead binding and extraction. For 5′ end repair, the RNA products were treated with tobacco acid pyrophosphatase (Epicenter) and T4 polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′ UGGAAUUCUCGGGUGCCAAGG) before being purified by a final round of streptavidin bead binding and extraction. RNA was reverse transcribed using 25 pmol of RP1 primer (5′AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) and size selected with 1X AMPure XP beads (Beckman) before being sequenced.

*Template-Switch Reverse Transcription (TSRT)*

Template-Switch Reverse Transcription protocol (also known as uPRO), was adapted from [9]. Nuclei were incubated in the nuclear run-on reaction condition (5 mM Tris-HCl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 0.5% Sarkosyl, 0.4 units / l of SUPERase-In) along with biotin-NTPs and rNTPs (125 $\mu$M rATP, 125 $\mu$M rGTP, 25 $\mu$M rCTP, 125 $\mu$M rUTP, and 25 $\mu$M biotin-11-CTP) for 5 min at 37°C. Run-On RNA was extracted using TRIzol, and fragmented with 0.2 N NaOH for 10-12 min on ice. Fragmented RNA was neutralized with 1 M Tris-HCl pH 6.8, and buffer exchanged by passing through P-30 columns (Biorad). 3′ RNA adaptor (/5Phos/GAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/) is ligated at 5 $\mu$M concentration for 1 hour at room temperature using T4 RNA ligase (NEB), and nascent RNA was enriched twice with streptavidin beads. Extracted RNA was converted to cDNA using template switch reverse transcription with 1 $\mu$M RP1-short RT primer (5′ GTTCAGAGTTCTACAGTCCGA), 3.75 M RTP-Template Switch Oligo (5′ GCCTTGGCACCCGAGAATTCCArGrGrG), 1x Template Switch Enzyme and Buffer (NEB) at 42°C for 30 min. Resulting product was size selected with AMPure XP beads, and the cDNA was PCR amplified using primers compatible with Illumina Small RNA sequencing (TruSeq Small RNA primers RP1 and RPIn).

### Trimming, Mapping, Visualization, Quality Control

Resulting FASTQ files were trimmed and mapped to the GRCh38/hg38 reference genome and prepared for analysis and visualization through our in-house pipeline. In short, resulting FASTQ read files were first trimmed using bbduk (v38.05) to remove adapter sequences, as well as short or low quality reads. Reads were mapped with HISAT2 (v2.1.0), and resulting SAM files converted to BAM files using Samtools (v1.8). BedGraph files were generated using Bedtools (v2.25.0), and converted to TDF files for visualization using IGVtools (v2.3.75). Quality metrics were generated with FastQC (v0.11.8), Preseq (v2.0.3), RSeQC (v3.0.0), with figures generated through MultiQC (v1.6). For further version information and specific input information, see NextFlow pipeline found at https://github.com/Dowell-Lab/Nascent-Flow.git.

### Exon/Intron Ratio

RefSeq annotations were used to define exonic and intronic boundaries for each gene. The first exon of each gene was excluded (to avoid the initiation peak signal) in each calculation. Reads were counted using featureCounts from the R-Subread package (v1.6.0). Exonic and intronic reads were summed and normalized by RPKM, and a ratio for each gene is calculated. These ratios were log-normalized and the median ratio calculated for each set of libraries analyzed.

### Discrete Wavelet Transform

Samples with high coverage were used for this analysis. This included samples from the GRO-LIG, PRO-LIG, GRO-CIRC and PRO-TSRT libraries. The coverage over a gene transcript was normalized to 0-1 scale as show below:

$$c_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Where $x = (x_i, ..., x_n)$ represents read counts over a genomic location $n$, and $c_i$ is the normalized coverage per genomic location. A total of 210 genes with a CV less than 0.55 and average TPM greater than 150 were selected. Using the PyWavelet (version 1.0.3) API in python (version 3.6.3), the symlet 5 mother wavelet was scanned across the 210 genes, returning wavelet coefficients (approximation coefficient and detail coefficients) (Fig. 2E) [21, 22, 52]. After the first pass of wavelet transform, the detail coefficients were used as input for principal component analysis (PCA) using scikit-learn (version 0.20.2) [53]. So, for each gene and each sample, PC1 and PC2 values were returned. Genes were split into categories based on whether the protocols could be split on PC1 and PC2 or whether the gene could not separate the protocols in PC space. Plots were generated with matplotlib (version 3.3.4), ggplot2 (version 3.3.3) and cowplot (version 1.1.1) [54, 55, 56]. Code for the DWT analysis can be found on github (https://github.com/Dowell-Lab/Protocol-Comparisons).

### Support Vector Machine

Principal component analysis values (from PC1 and PC2) derived from the wavelet transform analysis pipeline were used as input to a support vector machine (SVM). In order to verify the performance of the classification, the leave-one-out cross validation (LOOCV) criteria was used (Supplemental Fig. 4). A linear kernel was chosen for the SVM using the e1071 (version 1.7-4) package in R (R version 3.6.0) [57, 58]. The folds for the LOOCV were created with the caret package (version 6.0-86) in R (version 3.6.0) and accuracy for each fold and gene was calculated [59]. A total of 18 folds were created, where each of the 18 samples was held out one at a time as the test sample in the SVM, while the remaining samples were used as a training set. This was done for all the 210 genes analysed and the evaluation determined the number of genes accurately predicting the protocol for each of the 18 samples. Plots were made using ggplot2 (version 3.3.3) and cowplot (version 1.1.1) [55, 56]. The jupyter notebook for the SVM LOOCV analysis can be found on github (https://github.com/Dowell-Lab/Protocol-Comparisons).

### Pause Index Calculations

Refseq annotations were used as the basis for pause index calculations. Counts were generated either from bedtools multicov (v2.28.0). The "paused" region was defined as -50 bp to 250 bp from the annotated TSS, and the elongation region was defined as 251 bp from the TSS to the annotated PolyA site. Reads from the same strand as the annotated gene were counted for the paused and elongation region, and calculated the index as follows:

$$\text{pausing index}(pi) = \frac{ReadCount(Pausing\ Region)/L1}{ReadCount\left(Gene\ Body\right)/L2}$$

Where L1 is the length of the pausing region (300 bp) and L2 is the length of the elongation region, measured from 251 bp past the TSS to the annotated cleavage site found in RefSeq. Only pause index values from a gene's longest isoform were considered. Genes shorter than 2000 bp were removed.

The above analysis was repeated on using featureCounts (v1.6.2) in the R-Subread package (v1.6.0), where the paused region was defined as -20 to +80 from the annotated TSS, and the elongation region as +81 from the TSS to -1000 from the annotated PolyA site. Genes shorter than 2000 bp were filtered out. These results are available in Supplemental Fig. 10

### Gene/Intergenic Reads Ratio Calculation

Genic and intergenic regions were determined by RefSeq (hg38, release number 109, downloaded August 14, 2019 from UCSC genome browser) annotation. Genic and intergenic read proportions were calculated by RSeQC (v3.0.0) read_distribution.py. Genic regions were defined as those overlapping a RefSeq annotation, including introns and untranslated regions. Intergenic regions were calculated as the remainder of reads not mapping to a gene region. The reads ratio of genic and intergenic regions can be found for each sample in Supplemental Table 1.

### Tfit

Tfit was used to identify regions of bidirectional transcription in each of our run-on sequencing libraries. Resultant BedGraph files from our samples were used as the input for the –bedgraph flag of the Tfit prelim module. The resultant preliminary region file was used as the –segment flag input for the Tfit model module, resulting in the final bidirectional calls used for analysis (see also https://github.com/Dowell-Lab/Tfit.git). Calls between replicates and treatments were combined using *muMerge*, generating a set of combined calls for each set of conditions (GRO-LIG, PRO-LIG, and GRO-CIRC). To compare library preparation methods, the above GRO-CIRC and GRO-LIG sets were combined together through bedtools merge (v2.28.0). Likewise, to compare enrichment methods, PRO-LIG and GRO-LIG sets were combined via bedtools merge (v2.28.0).

### dREG

We used dREG to identify regions of bidirectional transcription in each of our run-on sequencing libraries. Resultant BAM files from our samples were first converted to BigWig files compatible with dREG (see https://github.com/Danko-Lab/RunOnBamToBigWig.git). Using the online dREG portal at https://django.dreg.scigap.org, these files were used to generate dREG calls for bidirectional regions. Calls between replicates and treatments were combined using *muMerge*, generating a set of combined calls for each set of conditions (GRO-LIG, PRO-LIG, and GRO-CIRC). For comparative analyses between any of these sets, each set combined by *muMerge* was concatenated and used as the input for bedtools merge (v2.28.0), generating a consensus set of regions for those two sets.

### Differential Transcription Analysis

Differential transcription was performed using the DESeq2 (v1.26.0) R package (R version 3.6.3). DESeq2 no longer allows differential calls without replicates; thus, when comparing libraries where treatments and replicates were combined, the DESeq (v. 1.38.0) R package was used instead. Gene counts were generated using featureCounts (v1.6.2) from the R Subread package (v1.6.0), counting over the

entire gene body from RefSeq Annotations (release number 109, downloaded August 14, 2019 from UCSC genome browser). For featureCounts, BED6 region files were converted to SAF format with the following command: awk -F "\t" -v OFS="\t" 'print{$4, $1, $2, $3, $6}' region.bed > region.saf. Only the highest transcribed isoform of each gene was considered. Counts over Tfit, dREG, or FANTOM calls were generated with featureCounts.

## GSEA

DESeq2 gene results were ranked based on -log(P-value)/sign(Fold-Change). These ranked lists were used as the input for GSEA-preranked module (v4.1.0). The Hallmark v7.4 gene sets were used as the input database. Results were generated using 1000 permutations. Gene symbols were not collapsed.

## TFEA

Resulting Tfit bidirectional calls were used as the input for TFEA for each experiment (summarized in Supplemental Table 1). Calls were combined using *muMerge*. Transcription factor motifs were identified using FIMO (MEME Suite v5.1.1), using full human HOCOMOCO (version 11) motifs.

**Abbreviations**

RO-seq: Run-On sequencing. PRO-seq: Precision Run-On sequencing. GRO-seq: Global Run-On sequencing. CIRC: Circularization based library preparation. LIG: Ligation based library preparation. RPR: Random Priming based library preparation. TSRT: Template Switching Reverse Transcriptase based library preparation. DWT: Discrete Wavelet Transform. PCA: Principal Component Analysis. SVM: Support Vector Machine. LOOCV: Leave-One-Out Cross Validation. TSS: Transcription Start Site. eRNA: Enhancer RNA. GSEA: Gene Set Enrichment Analysis. TFEA: Transcription Factor Enrichment Analysis.

**Declarations**

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Dr. Dowell is founder of Arpeggio Biosciences, the other authors declare that they have no competing interests.

Acknowledgements

We thank artist David Deen for figure composition and refinement assistance. We also thank the BioFrontiers Institute Next-Gen Sequencing Core and the Biochemistry Shared Cell Culture Facility for their invaluable contributions to this study.

Author's contributions

This study was conceived by RDD, MAA and SH. Discrete wavelet transform analyses was conducted by JTS and RFS. GRO-seq libraries were generated by MAA. PRO-seq libraries were generated by SH and MAA. All other analyses and initial manuscript was written by SH. All authors reviewed and revised the manuscript.

Funding

This work was funded by a National Science Foundation (NSF) ABI grant number 1759949. We acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing high-performance computing resources (NIH 1S10OD012300) supported by BioFrontiers' IT staff.

Availability of data and materials

The datasets used in this study are summarized in Supplemental Table 1. Datesets generated for this study are available through the Sequence Read Archive, under the accession PRJNA722106.

**References**


1. Core, L., Lis, J.: Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science **319**, 1791 (2008)
2. Kwak, H., Fuda, N.J., Core, L.J., Lis, J.T.: Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science **339**(6122), 950–953 (2013). doi:10.1126/science.1229386. http://science.sciencemag.org/content/339/6122/950.full.pdf
3. 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., Lis, J.T.: Base-pair-resolution genome-wide mapping of active rna polymerases using precision nuclear run-on (pro-seq). Nature Protocols **11**(8), 1455–1476 (2016). doi:10.1038/nprot.2016.086
4. Allen, M.A., Mellert, H., Dengler, V., Andryzik, Z., Guarnieri, A., Freeman, J.A., Luo, X., Kraus, W.L., Dowell, R.D., Espinosa, J.M.: Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. eLife **3**, 02200 (2014). doi: 10.7554/eLife.02200
5. Azofeifa, J.G., Allen, M.A., Hendrix, J.R., Read, T., Rubin, J.D., Dowell, R.D.: Enhancer RNA profiling predicts transcription factor activity. Genome Research (2018). doi:10.1101/gr.225755.117
6. Rubin, J.D., Stanley, J.T., Sigauke, R.F., Levandowski, C.B., Maas, Z.L., Westfall, J., Taatjes, D.J., Dowell, R.D.: Transcription factor enrichment analysis (tfea): Quantifying the activity of hundreds of transcription factors from a single experiment. Nature Communications Biology (2021). doi:10.1038/s42003-021-02153-7
7. Rothschild, G., Basu, U.: Lingering questions about enhancer rna and enhancer transcription-coupled genomic instability. Trends in Genetics **33**(2), 143–154 (2017). doi:10.1016/j.tig.2016.12.002
8. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., Glass, C.K., Rosenfeld, M.G., Fu, X.-D.: Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature **474**(7351), 390–394 (2011)
9. Kim, S.S.-Y., Dziubek, A., Alisa Lee, S., Kwak, H.: Nascent rna sequencing of peripheral blood leukocytes reveal gene expression diversity. bioRxiv (2019). doi:10.1101/836841. https://www.biorxiv.org/content/early/2019/11/09/836841.full.pdf
10. Barbieri, E., Hill, C., Quesnel-Vallieres, M., Barash, Y., Gardini, A.: Rapid and scalable profiling of nascent rna with fastgro. bioRxiv (2020). doi:10.1101/2020.01.24.916015. https://www.biorxiv.org/content/early/2020/01/24/2020.01.24.916015.full.pdf
11. Shivram, H., Iyer, V.R.: Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies. RNA **24**(9), 1266–1274 (2018). doi:10.1261/rna.066217.118
12. Sarantopoulou, D., Tang, S.Y., Ricciotti, E., Lahens, N.F., Lekkas, D., Schug, J., Guo, X.S., Paschos, G.K., FitzGerald, G.A., Pack, A.I., Grant, G.R.: Comparative evaluation of rna-seq library preparation methods for strand-specificity and low input. Scientific Reports **9**(1), 13477 (2019). doi:10.1038/s41598-019-49889-1
13. Wang, L., Felts, S.J., Van Keulen, V.P., Pease, L.R., Zhang, Y.: Exploring the effect of library preparation on rna sequencing experiments. Genomics **111**(6), 1752–1759 (2019). doi:10.1016/j.ygeno.2018.11.030
14. Andrysik, Z., Galbraith, M.D., Guarnieri, A.L., Zaccara, S., Sullivan, K.D., Pandey, A., MacBeth, M., Inga, A., Espinosa, J.M.: Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. Genome research **27**(10), 1645–1657 (2017)
15. Adelman, K., Lis, J.T.: Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet **13**(10), 720–731 (2012)
16. Roberts, T.C., Hart, J.R., Kaikkonen, M.U., Weinberg, M.S., Vogt, P.K., Morris, K.V.: Quantification of nascent transcription by bromouridine immunocapture nuclear run-on rt-qpcr. Nature protocols **10**(8), 1198 (2015)
17. Orioli, A., Praz, V., Lhôte, P., Hernandez, N.: Human MAF1 targets and represses active RNA polymerase III genes by preventing recruitment rather than inducing long-term transcriptional arrest. Genome Research **26**(5), 624–635 (2016). doi:10.1101/gr.201400.115
18. Sasse, S.K., Gruca, M., Allen, M.A., Kadiyala, V., Song, T., Gally, F., Gupta, A., Pufall, M.A., Dowell, R.D., Gerber, A.N.: Nascent transcript analysis of glucocorticoid crosstalk with TNF defines primary and cooperative inflammatory repression. Genome Research (2019). doi:10.1101/gr.248187.119. http://genome.cshlp.org/content/early/2019/10/16/gr.248187.119.full.pdf+html
19. Bahat, A., Lahav, O., Plotnikov, A., Leshkowitz, D., Dikstein, R.: Targeting spt5-pol II by small-molecule inhibitors uncouples distinct activities and reveals additional regulatory roles. Molecular Cell **76**(4), 617–6314 (2019). doi:10.1016/j.molcel.2019.08.024
20. Smith, J.P., Dutta, A.B., Sathyan, K.M., Guertin, M.J., Sheffield, N.C.: Peppro: quality control and processing of nascent rna profiling data. Genome Biology **22**(1), 155 (2021)
21. Daubechies, I.: Ten Lectures on Wavelets. SIAM, ??? (1992)
22. Lee, G.R., Gommers, R., Waselewski, F., Wohlfahrt, K., O'Leary, A.: Pywavelets: A python package for wavelet analysis. Journal of Open Source Software **4**(36), 1237 (2019)
23. Jonkers, I., Kwak, H., Lis, J.T.: Genome-wide dynamics of pol ii elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife **3**, 02407 (2014). doi:10.7554/eLife.02407
24. Day, D.S., Zhang, B., Stevens, S.M., Ferrari, F., Larschan, E.N., Park, P.J., Pu, W.T.: Comprehensive analysis of promoter-proximal rna polymerase ii pausing across mammalian cell types. Genome Biology **17**(1), 120 (2016). doi:10.1186/s13059-016-0984-2
25. Kim, T.-k., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., Greenberg, M.E.: Widespread transcription at neuronal activity-regulated enhancers. Nature **465**(7295), 182–187 (2010)
26. Gao, T., Qian, J.: EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research **48**(D1), 58–64 (2019). doi:10.1093/nar/gkz980. https://academic.oup.com/nar/article-pdf/48/D1/D58/31697342/gkz980.pdf

27. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., Hayashizaki, Y.: Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences **100**(26), 15776–15781 (2003)
28. Cardiello, J.F., Sanchez, G.J., Allen, M.A., Dowell, R.D.: Lessons from eRNAs: understanding transcriptional regulation through the lens of nascent RNAs. Transcription **11**(1), 3–18 (2020)
29. Danko, C.G., Hyland, S.L., Core, L.J., Martins, A.L., Waters, C.T., Lee, H.W., Cheung, V.G., Kraus, W.L., Lis, J.T., Siepel, A.: Identification of active transcriptional regulatory elements from GRO-seq data. Nat Meth **12**(5), 433–438 (2015)
30. Azofeifa, J.G., Dowell, R.D.: A generative model for the behavior of RNA polymerase. Bioinformatics **33**(2), 227–234 (2016). doi:10.1093/bioinformatics/btw599. http://oup.prod.sis.lan/bioinformatics/article-pdf/33/2/227/25142928/btw599.pdf
31. Hah, N., Danko, C., Core, L., Waterfall, J.J., Siepel, A., Lis, J.T., Kraus, W.L.: A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell **145**(4), 622–634 (2011)
32. Hah, N., Murakami, S., Nagari, A., Danko, C.G., Kraus, W.L.: Enhancer transcripts mark active estrogen receptor binding sites. Genome Research **23**(8), 1210–1223 (2013)
33. Shen, H., Maki, C.G.: Pharmacologic activation of p53 by small-molecule mdm2 antagonists. Current pharmaceutical design **17**(6), 560–568 (2011). doi:10.2174/138161211795222603
34. Su, Z., Łabaj, P.P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., Wang, C., Schroth, G.P., Setterquist, R.A., Thompson, J.F., Jones, W.D., Xiao, W., Xu, W., Jensen, R.V., Kelly, R., Xu, J., Conesa, A., Furlanello, C., Gao, H., Hong, H., Jafari, N., Letovsky, S., Liao, Y., Lu, F., Oakeley, E.J., Peng, Z., Praul, C.A., Santoyo-Lopez, J., Scherer, A., Shi, T., Smyth, G.K., Staedtler, F., Sykacek, P., Tan, X.-X., Thompson, E.A., Vandesompele, J., Wang, M.D., Wang, J., Wolfinger, R.D., Zavadil, J., Auerbach, S.S., Bao, W., Binder, H., Blomquist, T., Brilliant, M.H., Bushel, P.R., Cai, W., Catalano, J.G., Chang, C.-W., Chen, T., Chen, G., Chen, R., Chierici, M., Chu, T.-M., Clevert, D.-A., Deng, Y., Derti, A., Devanarayan, V., Dong, Z., Dopazo, J., Du, T., Fang, H., Fang, Y., Fasold, M., Fernandez, A., Fischer, M., Furió-Tari, P., Fuscoe, J.C., Caimet, F., Gaj, S., Gandara, J., Gao, H., Ge, W., Gondo, Y., Gong, B., Gong, M., Gong, Z., Green, B., Guo, C., Guo, L., Guo, L.-W., Hadfield, J., Hellemans, J., Hochreiter, S., Jia, M., Jian, M., Johnson, C.D., Kay, S., Kleinjans, J., Lababidi, S., Levy, S., Li, Q.-Z., Li, L., Li, P., Li, Y., Li, H., Li, J., Li, S., Lin, S.M., López, F.J., Lu, X., Luo, H., Ma, X., Meehan, J., Megherbi, D.B., Mei, N., Mu, B., Ning, B., Pandey, A., Pérez-Florido, J., Perkins, R.G., Peters, R., Phan, J.H., Pirooznia, M., Qian, F., Qing, T., Rainbow, L., Rocca-Serra, P., Sambourg, L., Sansone, S.-A., Schwartz, S., Shah, R., Shen, J., Smith, T.M., Stegle, O., Stralis-Pavese, N., Stupka, E., Suzuki, Y., Szkotnicki, L.T., Tinning, M., Tu, B., van Delft, J., Vela-Boza, A., Venturini, E., Walker, S.J., Wan, L., Wang, W., Wang, J., Wang, J., Wieben, E.D., Willey, J.C., Wu, P.-Y., Xuan, J., Yang, Y., Ye, Z., Yin, Y., Yu, Y., Yuan, Y.-C., Zhang, J., Zhang, K.K., Zhang, W., Zhang, W., Zhang, Y., Zhao, C., Zheng, Y., Zhou, Y., Zumbo, P., Tong, W., Kreil, D.P., Mason, C.E., Shi, L., Consortium, S.E.Q.C..M.A.Q.C.-I.I.I.: A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nature Biotechnology **32**(9), 903–914 (2014). doi:10.1038/nbt.2957
35. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C.: Pgc-1$\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics **34**(3), 267–273 (2003). doi:10.1038/ng1180
36. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **102**(43), 15545–15550 (2005). doi:10.1073/pnas.0506580102
37. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology **15**(12), 550 (2014). doi:10.1186/s13059-014-0550-8
38. Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., Lis, J.T.: Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes & Development **25**(7), 742–754 (2011)
39. Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G., Lis, J.T.: Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. Molecular Cell **62**(1), 63–78 (2016). doi:10.1016/j.molcel.2016.02.025
40. Dukler, N., Booth, G.T., Huang, Y.-F., Tippens, N., Waters, C.T., Danko, C.G., Lis, J.T., Siepel, A.: Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Research **27**(11), 1816–1829 (2017). doi:10.1101/gr.222935.117
41. Booth, G.T., Parua, P.K., Sansó, M., Fisher, R.P., Lis, J.T.: Cdk9 regulates a promoter-proximal checkpoint to modulate rna polymerase ii elongation rate in fission yeast. Nature Communications **9**(1), 543 (2018). doi:10.1038/s41467-018-03006-4
42. Aoi, Y., Smith, E.R., Shah, A.P., Rendleman, E.J., Marshall, S.A., Woodfin, A.R., Chen, F.X., Shiekhattar, R., Shilatifard, A.: Nelf regulates a promoter-proximal step distinct from rna pol ii pause-release. Molecular Cell **78**(2), 261–2745 (2020). doi:10.1016/j.molcel.2020.02.014
43. Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., Sandelin, A.: Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun **5** (2014)
44. Wang, J., Zhao, Y., Zhou, X., Hiebert, S.W., Liu, Q., Shyr, Y.: Nascent rna sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics **19**(1), 633 (2018). doi:10.1186/s12864-018-5016-z
45. Wissink, E.M., Vihervaara, A., Tippens, N.D., Lis, J.T.: Nascent rna analyses: tracking transcription and its regulation. Nature Reviews Genetics **20**(12), 705–723 (2019). doi:10.1038/s41576-019-0159-6

46. Marioni, J., Mason, C., Mane, S., Stephens, M., Gilad, Y.: RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research **18**, 1509 (2008)
47. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics **11**(10), 733–739 (2010)
48. Goh, W.W.B., Wang, W., Wong, L.: Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology **35**(6), 498–507 (2017)
49. Somekh, J., Shen-Orr, S.S., Kohane, I.S.: Batch correction evaluation framework using a-priori gene-gene associations: applied to the gtex dataset. BMC Bioinformatics **20**(1), 268 (2019). doi:10.1186/s12859-019-2855-9
50. Zhang, Y., Parmigiani, G., Johnson, W.E.: ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics **2**(3) (2020). doi:10.1093/nargab/lqaa078. lqaa078. https://academic.oup.com/nargab/article-pdf/2/3/lqaa078/34054992/lqaa078.pdf
51. Sanitá Lima, M., Smith, D.R.: Don't just dump your data and run. EMBO reports **18**(12), 2087–2089 (2017). doi:10.15252/embr.201745118. https://www.embopress.org/doi/pdf/10.15252/embr.201745118
52. Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)
53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
54. Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007). doi:10.1109/MCSE.2007.55
55. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, (2016). https://ggplot2.tidyverse.org
56. Wilke, C.O.: cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2020). R package version 1.1.1. `https://CRAN.R-project.org/package=cowplot`
57. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. (2021). R package version 1.7-6. `https://CRAN.R-project.org/package=e1071`
58. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). R Foundation for Statistical Computing. https://www.R-project.org/
59. Kuhn, M.: Building predictive models in R using the caret package. Journal of Statistical Software, Articles **28**(5), 1–26 (2008). doi:10.18637/jss.v028.i05
60. Léveillé, N., Melo, C.A., Rooijers, K., Díaz-Lagares, A., Melo, S.A., Korkmaz, G., Lopes, R., Moqadam, F.A., Maia, A.R., Wijchers, P.J., Geeven, G., den Boer, M.L., Kalluri, R., de Laat, W., Esteller, M., Agami, R.: Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nature Communications **6**(1) (2015). doi:10.1038/ncomms7520
61. Niskanen, E.A., Malinen, M., Sutinen, P., Toropainen, S., Paakinaho, V., Vihervaara, A., Joutsen, J., Kaikkonen, M.U., Sistonen, L., Palvimo, J.J.: Global sumoylation on active chromatin is an acute heat stress response restricting transcription. Genome Biology **16**(1), 153 (2015). doi:10.1186/s13059-015-0717-y


**Figures**

**Figure 1 Summary of Run-On Sequencing (RO-seq) data sets.** (A) Summary diagram indicating enrichment steps for Global Run-On (GRO-seq, top) and Precision Run-On (PRO-seq, bottom) reactions. (B) Summary diagram for library preparation reactions. Blue bars: RNA; brown bars: cDNA; yellow/green bars: sequencing adapters. Library preparation enzymes are labeled and represented by blue shapes at each step.

**Figure 2 Quality Control metrics for varying library preparation and enrichment techniques.** (A) Nucleotide distribution of DMSO samples are plotted indicating the percent nucleotide representation (y-axis) versus the position within each read (x-axis). Library specific signatures are identifiable in CIRC and TSRT libraries (blue arrows). (B) Principal-Component Analysis of assorted library preparation and enrichment methods. Each library was prepped using HCT116 cells treated with either DMSO or Nutlin-3a for 1 hour. Log-normalized density plots of exon/intron ratios for each gene for each (C) enrichment method and (D) library preparation method. Mean indicated by vertical line for each respective distribution. (E) Schematic showing the wavelet transformation approach at the UBB locus. (F) Detail coefficients at UBB locus separates PRO and GRO libraries on PC1. (G) SVM classifier results for each tested library

**Figure 3 Analysis of gene transcription start sites among different protocols and library preparations.** (A) Genome viewer screenshot of 5′ end distribution among various library preparation and enrichment methods. (B) Metagenes constructed from GRO-seq (orange) and PRO-seq (blue) libraries (Ligation based library preparation, HCT116, DMSO 1hr). Genes shorter than 2000 bp were removed (n=15076). Vertical line indicates TSS annotated in RefSeq database. Distance from TSS is in bp, read depth was normalized by counts-per-million (CPM) (C) Pausing index calculations for top 500 most transcribed genes in GRO-seq and PRO-seq libraries, presented with Pearson (left) and Spearman (right) correlations (red line: y=x, black line: best fit). Pausing region is defined as -50 bp to 250 bp from annotated TSS (See Materials and Methods). (D) Metagenes constructed from Ligation (blue), Random-Primed (red), and Circularization-based (green) libraries (GRO-seq enrichment, HCT116, DMSO 1 hr). Genes shorter than 2000 bp were removed (n=15076). Vertical line indicates TSS annotated in RefSeq database. Distance from TSS is in bp, read depth was normalized by counts-per-million (CPM). (E) Pausing index calculations for Circularization and Ligation based libraries (GRO-seq, HCT116, DMSO 1 hr), graphed as in (C).

**Figure 4 Analysis of enhancer elements in multiple datasets.** (A,B) Number of reads counted over RefSeq annotated gene regions divided by the number of reads counted over intergenic (unannotated) regions, for each dataset analyzed. Datasets were first analyzed by enrichment method (GRO-seq (n=23) vs. PRO-seq (n=21), $p < .01$), then by library preparation method (LIG (n=17) vs CIRC (n=10) vs TSRT (n=10) vs RPR (n=7), $p > .05$). (C) Example section representing disparate representation of reads over an enhancer, even at high depths. (D, E) Scatterplots representing reads over Tfit (enhancer) calls (calls combined by MuMerge, counts normalized by TPM). (F) MA plot of calls found in (D). Red dots are significant ($p < .05$). (G, H) Metagenes of significant hits found in (F). Vertical line indicates the approximated center of the bidirectional transcripts as determined by Tfit. Distance from the center of the bidirectional is in bp, read depth was normalized by counts-per-million (CPM). (G): Calls that were differentially captured in GRO-LIG (n=224). (H): Calls that were differentially captured in PRO-LIG (n=480).

**Figure 5 TFEA and DESeq2 analyses of library preparation methods.** (A) Cartoon schematic demonstrating uncorrected (RefSeq Annotation) and 5′ corrected counting methods. (B) GSEA gene rank comparison of HALLMARK_P53 Gene set. Overlap is shown as genes that enrich in both datasets, genes that enrich in only one dataset, and genes that do not enrich in either dataset. (C) Scatterplot of comparative gene ranks for all p53 genes. Points in green indicate significant enrichment, as in (B). (D) Representation of nascent transcription data set. Bidirectional transcripts occur at active enhancer sites and gene start sites. Enhancer transcription co-occurs with upregulated gene transcription, indicating transcription factor activation. (E) TFEA results for GRO-LIG (Left) and GRO-CIRC (Right). p53 family (p53, p63, p73) highlighted by red dots

**Supplementary Information**
**Additional Files**
**Additional File 1**
**Supplemental Table 1 — Sample Information**
Sample information for all RO-seq libraries used in analyses. Information is as follows: cell type, treatment, time point, enrichment protocol, library preparation method, replicate number, depth, complexity metrics, and SRA identifiers

**Additional File 2**
**Supplemental Figure 1** Complexity curves and read distributions of public and in-house GRO-RPR datasets, indicating trends of lower quality for our libraries with this preparation.
**Supplemental Figure 2** Discrete wavelet transform PCA results for 210 highly transcribed genes, demonstrates 41.4% of genes separate on PC1.
**Supplemental Figure 3** DWT PCA Results of detail coefficients at UBB Locus. PCA results for UBB locus, as in Figure 2F, but results are colored by library preparation method. At this locus, the results cluster less distinctly by library preparation method, compared to the enrichment protocol.
**Supplemental Figure 4** Schematic for the Support Vector Machine (SVM) leave one out cross validation (LOOCV) analysis. Eighteen nascent RNA sequencing samples were used as input. Given a gene, each of the samples was selected as a test sample and the other samples as training set, the SVM classification was evaluated. Based on this criteria, a majority of the genes (>75%) accurately classified the protocol for the n=18 samples.
**Supplemental Figure 5** Rank correlation of elongation regions of genes in PRO-LIG and GRO-CIRC libraries. Initiation region was defined as in Fig. 3, (See also Materials and Methods). Genes smaller than 2000 bp were removed, and only the resulting top 500 transcribed genes (by TPM) were considered. Initiation region rank correlation (Top) is weaker than in the elongation region (Bottom), suggesting that most of the variability in these libraries lies near transcription start sites.
**Supplemental Figure 6** Read count heatmap of pause regions of genes in GRO-CIRC, GRO-LIG, GRO-RPR, and PRO-LIG libraries. (TSS +/- 500 bp, 10 bp per region). RefSeq hg38 gene annotations were used. Genes shorter than 2000 bp were not included. There is comparatively lower coverage near the TSS in many genes, representing the center of bidirectional transcription. This is especially prevalent in GRO-RPR and PRO-LIG libraries.
**Supplemental Figure 7** Read count heatmap of pause regions of genes in public MCF7 GRO-LIG and GRO-CIRC libraries (TSS +/- 500 bp, 10 bp per region) [14, 60]. RefSeq hg38 gene annotations were used. Genes shorter than 2000 bp were not included. These heatmaps reflect those generated from our own datasets, thus reinforcing that the patterns found in our datasets are not only a result of batch effects from our lab (See Supplemental Table 1).
**Supplemental Figure 8** Metagenes of PRO-LIG libraries with varying Biotin ratios. Libraries generated from HCT116 cell treated with DMSO, with PRO-LIG strategies. Libraries differed in the amount of available biotin added (25 $\mu$M vs 2.5 $\mu$M).
**Supplemental Figure 9** Pause Index Correlations of PRO_LIG Replicates.
**Supplemental Figure 10** Pause index (PI) and rank correlation of PI generated from GRO-CIRC and GRO-LIG libraries Pause indices generated with a different method than 3, using a different pause region definition (Pause region:TSS to +80, elongation region +81:TES-1000, genes shorter than 2000 bp were not included), and a different counting software (featureCounts). In spite of these changes, the relative distribution and correlation remains consistent with 3D, suggesting that these patterns are not merely a result of our software or PI region definitions.
**Supplemental Figure 11** (Top) Metagenes of public datasets[61, 40]. Libraries were generated from K562 Cells treated with DMSO and prepped with either PRO-LIG or GRO-CIRC methods. All 4 nucleotides added to the run-on reaction were biotin-NTPs. (Bottom) Public data[61, 40] were subjected to analysis as in Fig. 3C, left (see Supplemental Table 1). PI regions were defined as in Fig. 3. Notably, the rank correlation remains low (R=0.44) consistent with PI differences being driven by protocol.
**Supplemental Figure 12** Density plot of read counts (TPM) over HCT116 enhancers annotated in the FANTOM database. FANTOM annotations were generated from CAGE data, thus we reasoned that most FANTOM regions would overlap with relatively stable bidirectional transcription. As such, read counts over these regions is much more highly correlated between different protocols.
**Supplemental Figure 13** UpSet plot of Tfit and dREG calls among PRO-LIG, GRO-LIG, and GRO-CIRC libraries. Calls from replicates and treatments were combined using *muMerge* [6]. Much of the disparity in these overlaps can be attributed to Tfit or dREG failing to call bidirectional regions despite the presence of bidirectional transcription, as shown in 4C,D. However, there remains many enhancer regions not captured in one protocol at this depth.
**Supplemental Figure 14** Density Plot of Read Counts (TPM) over Tfit Calls between replicates Read counts for merged Tfit calls for PRO-LIG replicates. Counts are log(TPM) normalized to correct for depth. Very low read count calls (TPM $<$ .1) were excluded as likely false positives.
**Supplemental Figure 15** Metagene of enhancers differentially captured in either GRO-LIG or GRO-CIRC libraries Tfit calls across replicates and treatments were combined together using *muMerge* for both GRO-LIG and GRO-CIRC libraries. Metagenes of calls that were differentially captured (as determined by DESeq, see also 4, Materials and Methods) were generated for both GRO-CIRC (Top) and GRO-LIG (Bottom) Tfit calls. Reads counts were normalized by CPM.
**Supplemental Figure 16** Example enhancer region where libraries appear to disparately capture differential p53 enhancer activity. Darker colors represent transcription level in Nutlin-3a treated libraries, while lighter colors represent levels found in DMSO-treated libraries. Read counts are normalized by CPM.
**Supplemental Figure 17** Enrichment plot of GSEA results for GRO-LIG, PRO-LIG, and GRO-CIRC libraries Gene region definitions were adjusted as per 5A. In spite of library variations, the HALLMARK_P53_PATHWAY (red) is the strongest hit in all of our library comparisons.
**Supplemental Figure 18** Overlap of GSEA p53 genes in GRO-LIG and PRO-LIG libraries. Analysis was performed using counts over gene bodies (Left), and using a 5' correction (Right), as in 5A (see also Materials and Methods).

**Supplemental Figure 19** TFEA results for PRO-LIG libraries. Regions were combined by *muMerge*, as in Fig. 5. Red dots indicate transcription factors belonging to the p53 family (TP53, TP63, TP73).
**Supplemental Figure 20** Rank Differential of GRO-LIG and PRO-LIG enhancers. Ranks were determined within the TFEA through DESeq2. p53 enhancers which were more than 2 SD away from the mean were considered to be differentially captured in GRO-seq or PRO-seq.

# Figures

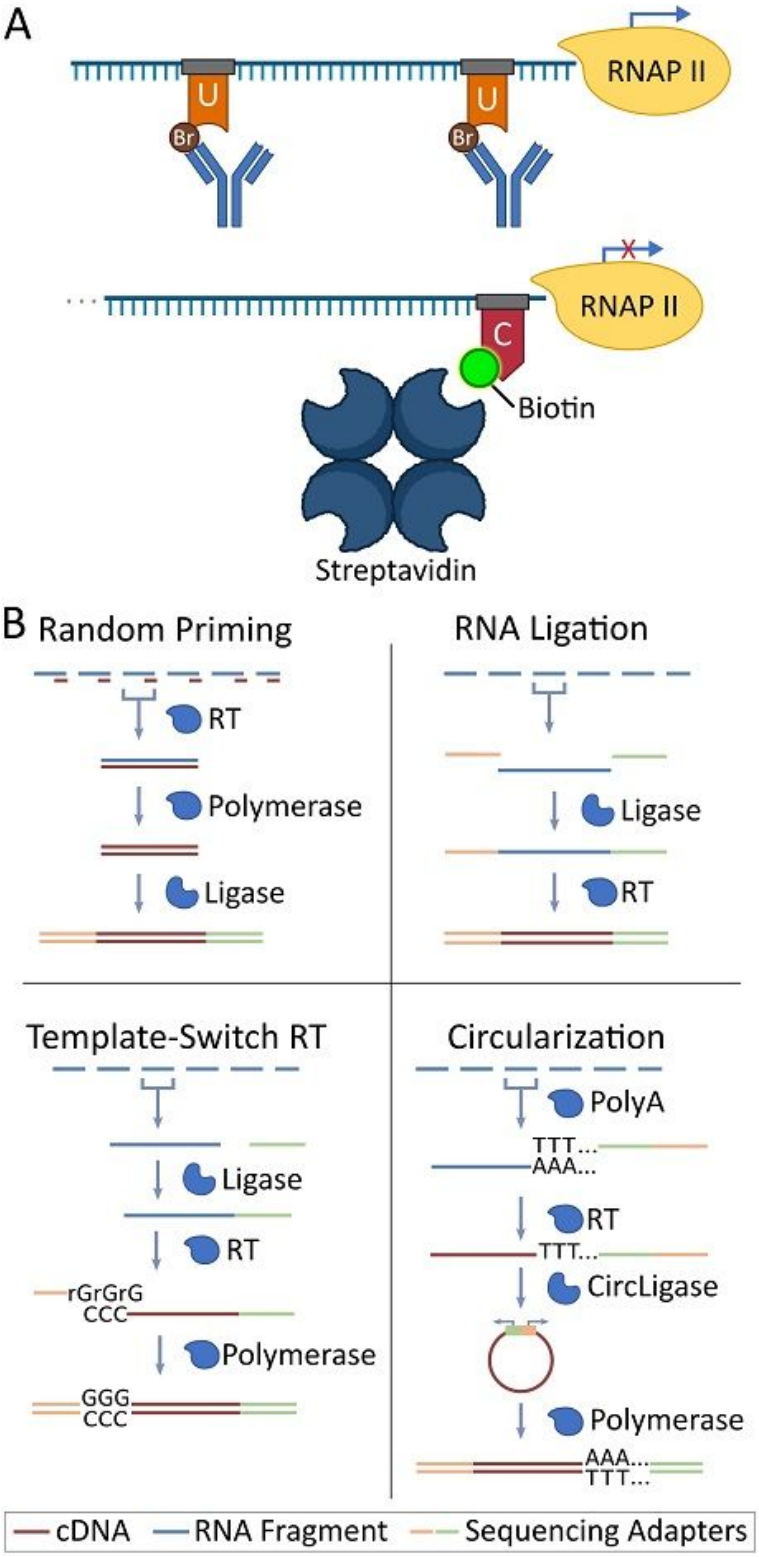
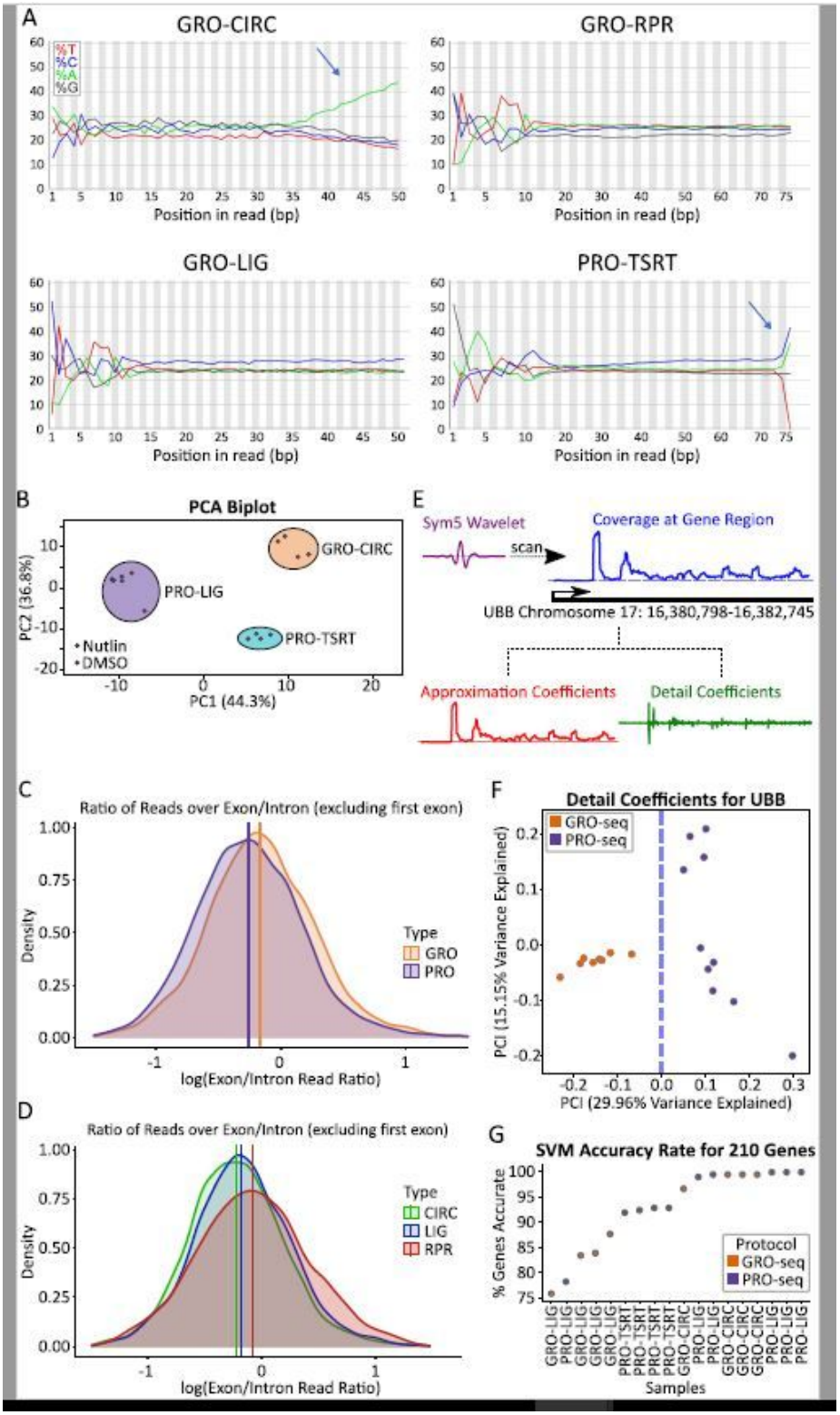



**Figure 1**

Summary of Run-On Sequencing (RO-seq) data sets. (A) Summary diagram indicating enrichment steps for Global Run-On (GRO-seq, top) and Precision Run-On (PRO-seq, bottom) reactions. (B) Summary diagram for library preparation reactions. Blue bars: RNA; brown bars: cDNA; yellow/green bars:

sequencing adapters. Library preparation enzymes are labeled and represented by blue shapes at each step.




## Figure 2

Quality Control metrics for varying library preparation and enrichment techniques. (A) Nucleotide distribution of DMSO samples are plotted indicating the percent nucleotide representation (y-axis) versus the position within each read (x-axis). Library specific signatures are identifiable in CIRC and TSRT

libraries (blue arrows). (B) Principal-Component Analysis of assorted library preparation and enrichment methods. Each library was prepped using HCT116 cells treated with either DMSO or Nutlin-3a for 1 hour. Log-normalized density plots of exon/intron ratios for each gene for each (C) enrichment method and (D) library preparation method. Mean indicated by vertical line for each respective distribution. (E) Schematic showing the wavelet transformation approach at the UBB locus. (F) Detail coefficients at UBB locus separates PRO and GRO libraries on PC1. (G) SVM classifier results for each tested library

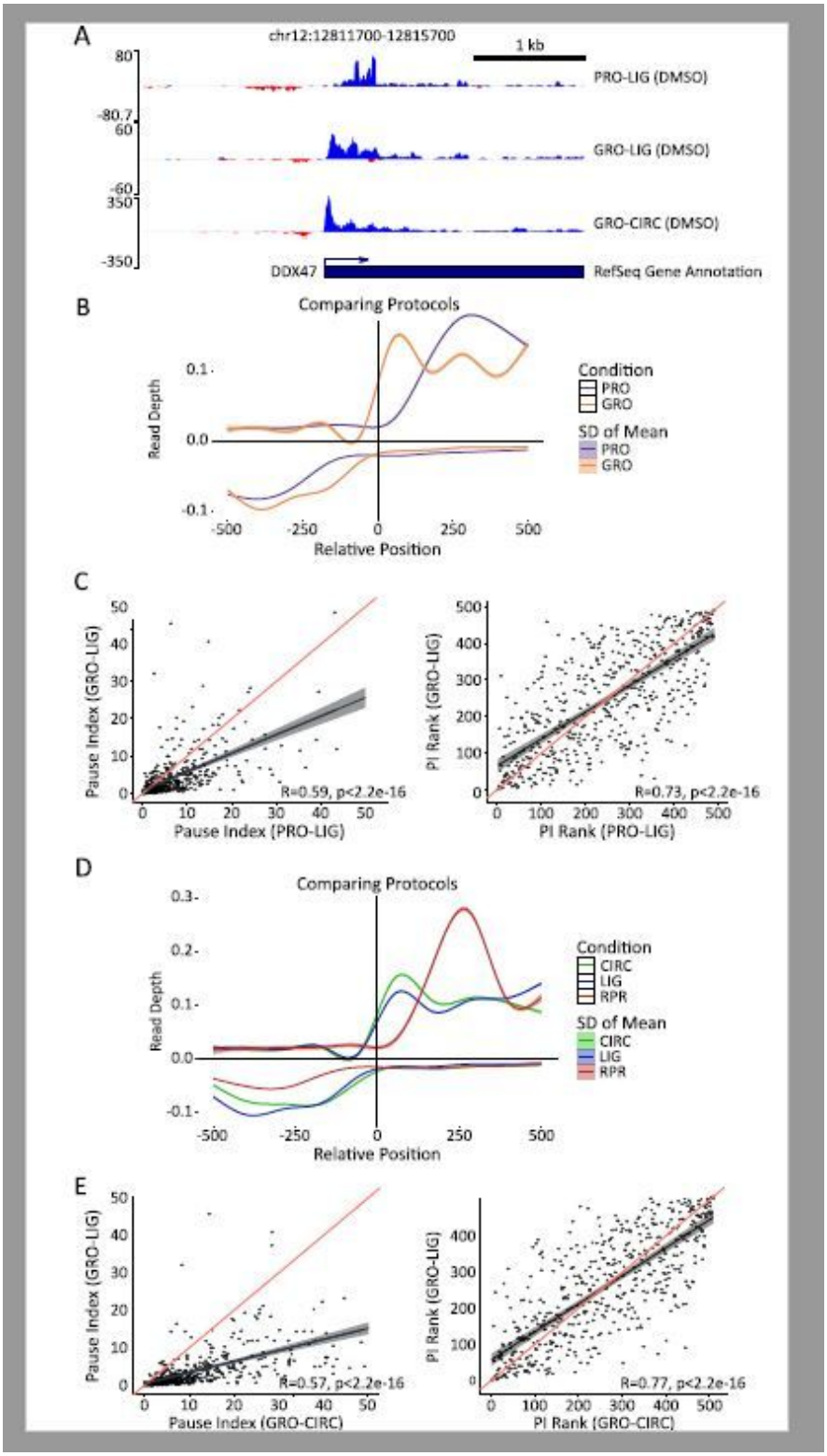
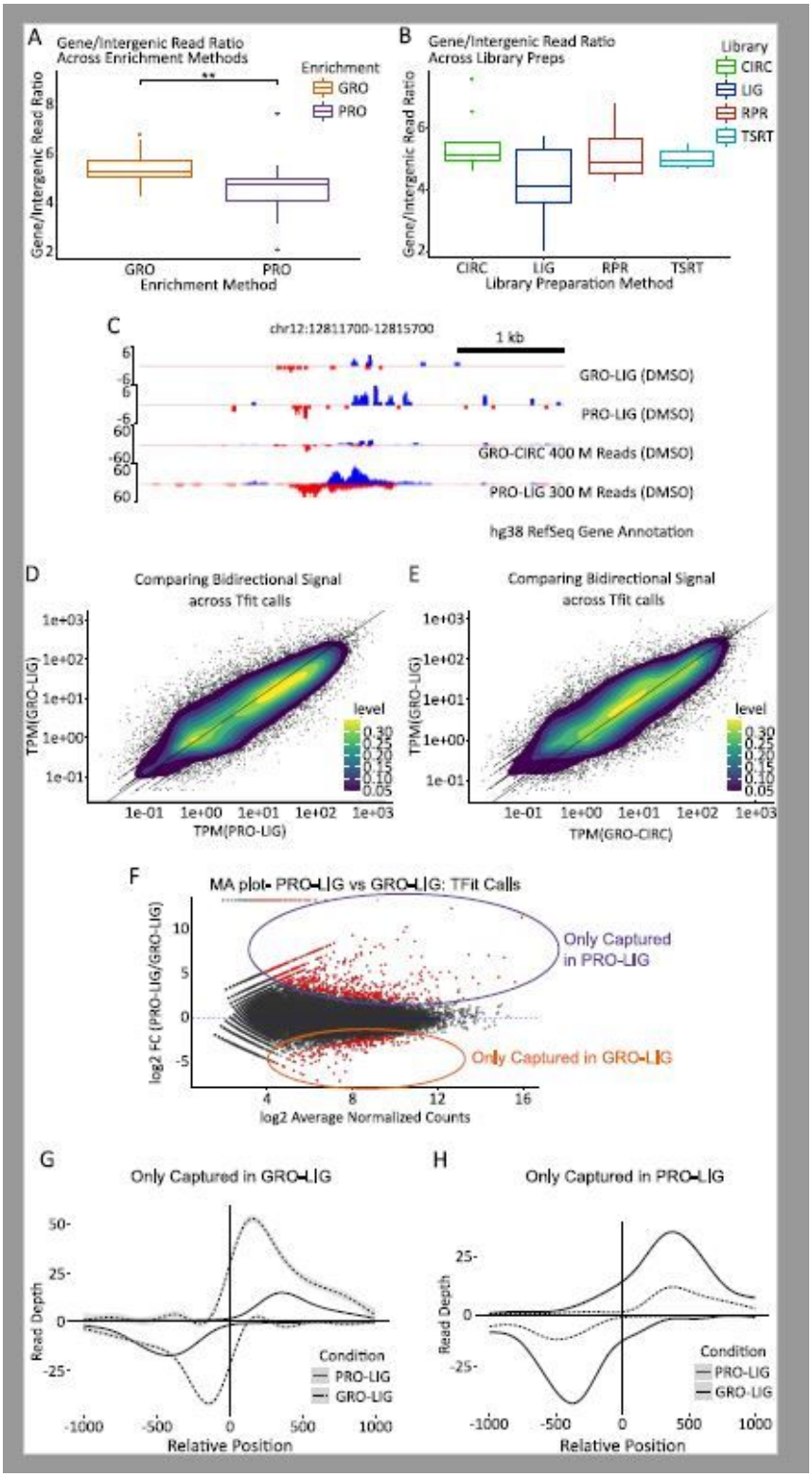



Figure 3

Analysis of gene transcription start sites among different protocols and library preparations. (A) Genome viewer screenshot of 50 end distribution among various library preparation and enrichment methods. (B) Metagenes constructed from GRO-seq (orange) and PRO-seq (blue) libraries (Ligation based library preparation, HCT116, DMSO 1hr). Genes shorter than 2000 bp were removed (n=15076). Vertical line indicates TSS annotated in RefSeq database. Distance from TSS is in bp, read depth was normalized by counts-per-million (CPM) (C) Pausing index calculations for top 500 most transcribed genes in GRO-seq and PRO-seq libraries, presented with Pearson (left) and Spearman (right) correlations (red line: y=x, black line: best fit). Pausing region is defined as -50 bp to 250 bp from annotated TSS (See Materials and Methods). (D) Metagenes constructed from Ligation (blue), Random-Primed (red), and Circularization-based (green) libraries (GRO-seq enrichment, HCT116, DMSO 1 hr). Genes shorter than 2000 bp were removed (n=15076). Vertical line indicates TSS annotated in RefSeq database. Distance from TSS is in bp, read depth was normalized by counts-per-million (CPM). (E) Pausing index calculations for Circularization and Ligation based libraries (GRO-seq, HCT116, DMSO 1 hr), graphed as in (C).




## Figure 4

Analysis of enhancer elements in multiple datasets. (A,B) Number of reads counted over RefSeq annotated gene regions divided by the number of reads counted over intergenic (unannotated) regions, for each dataset analyzed. Datasets were first analyzed by enrichment method (GRO-seq (n=23) vs. PRO-seq (n=21), $p < .01$), then by library preparation method (LIG (n=17) vs CIRC (n=10) vs TSRT (n=10) vs RPR (n=7), $p > .05$). (C) Example section representing disparate representation of reads over an enhancer,

even at high depths. (D, E) Scatterplots representing reads over Tfit (enhancer) calls (calls combined by MuMerge, counts normalized by TPM). (F) MA plot of calls found in (D). Red dots are significant ($p < .05$). (G, H) Metagenes of significant hits found in (F). Vertical line indicates the approximated center of the bidirectional transcripts as determined by Tfit. Distance from the center of the bidirectional is in bp, read depth was normalized by counts-per-million (CPM). (G): Calls that were differentially captured in GRO-LIG (n=224). (H): Calls that were differentially captured in PRO-LIG (n=480).

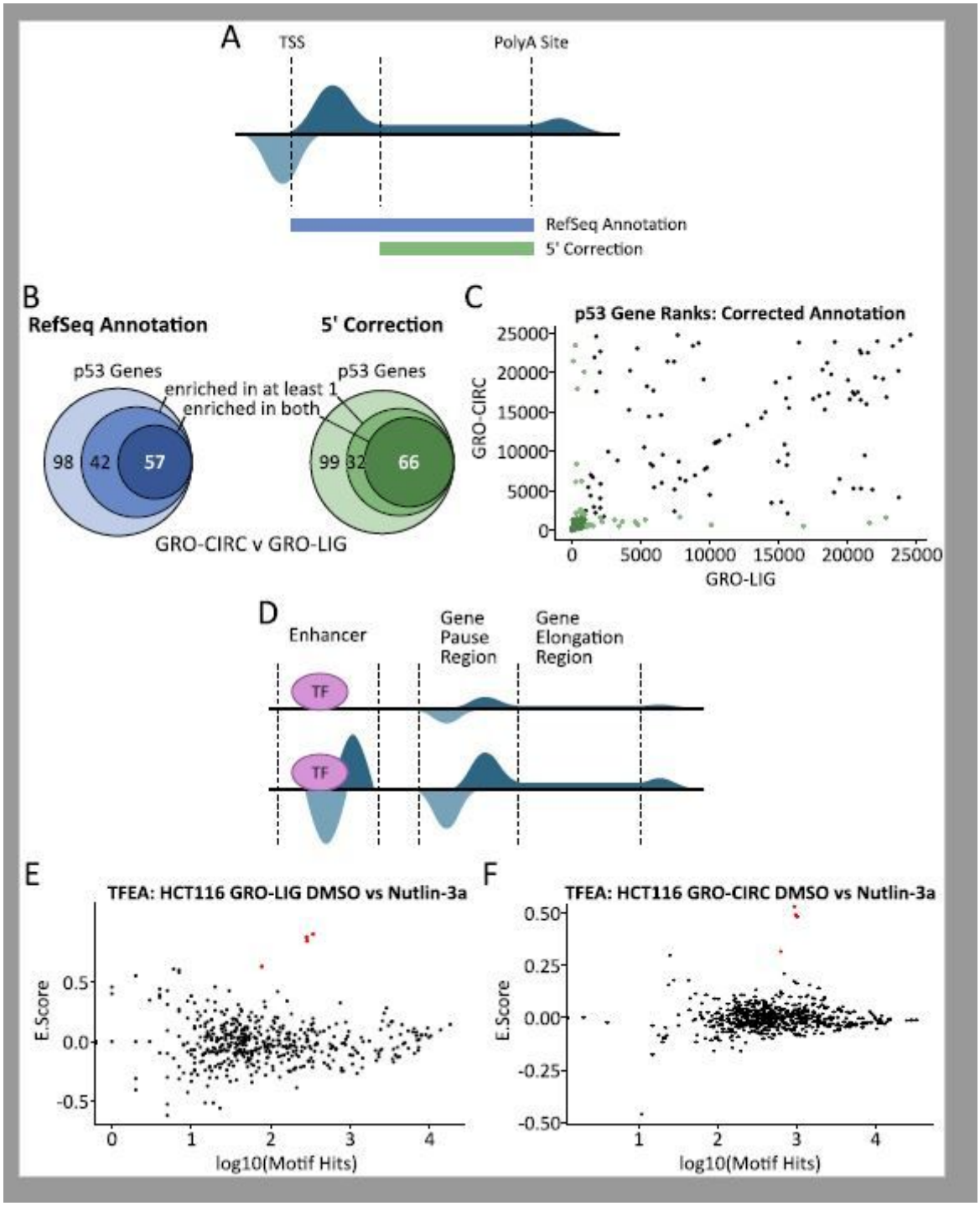



Figure 5

TFEA and DESeq2 analyses of library preparation methods. (A) Cartoon schematic demonstrating uncorrected (RefSeq Annotation) and 50 corrected counting methods. (B) GSEA gene rank comparison of HALLMARK_P53 Gene set. Overlap is shown as genes that enrich in both datasets, genes that enrich in only one dataset, and genes that do not enrich in either dataset. (C) Scatterplot of comparative gene ranks for all p53 genes. Points in green indicate significant enrichment, as in (B). (D) Representation of nascent transcription data set. Bidirectional transcripts occur at active enhancer sites and gene start sites. Enhancer transcription co-occurs with upregulated gene transcription, indicating transcription factor activation. (E) TFEA results for GRO-LIG (Left) and GRO-CIRC (Right). p53 family (p53, p63, p73) highlighted by red dots

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalfigures.pdf
- supplementaltable1.xlsx