Alternative Splicing Reporters Integrated Throughout the Genome Measure Splice Pattern Sensitivity to Local Chromatin Context

Matthew Scott Hirano

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Georg Seelig, Chair

Andre Berndt

Drew Sellers

Program Authorized to Offer Degree:

Bioengineering

University of Washington

Abstract

Alternative Splicing Reporters Integrated Throughout the Genome Measure Splice Pattern Sensitivity to Local Chromatin Context

Matthew Scott Hirano

Chair of the Supervisory Committee:

Georg Seelig

Electrical & Computer Engineering

Alternative splicing (AS) of transcripts into one or more isoforms is a major contributor to eukaryotic transcriptional diversity. Regulation of one splice isoform over another is important for downstream expression, and sequence and protein binding determinants of alternative splicing have been thoroughly studied. While gene position within the genome has been shown to be a powerful regulator of general transcription, the relationship between AS and epigenetic mechanisms like gene position or nuclear compartmentalization is still not well studied. To better understand position effects on AS, I apply a massively parallel reporter assay to integrate thousands of alternatively splicing reporters across the genome of K562 cells and measure alternative splicing activity by location. Five separate gene designs with different alternatively splicing exons were utilized to observe positional effects on single

exons with baseline inclusions ranging from high to low. By associating insertion locations with known DNA elements, epigenetic markers, and domain labels, the effects of general elements on a uniform gene's expression and splicing patterns within the same domain could be measured. Ultimately, measurements of integrated reporters across thousands of positions demonstrated little variance in alternative splicing and more minor variance in expression than previously observed with non-splicing reporters. Together, these results demonstrate the greater relative strength of exon, intron, and promoter sequence on transcriptional regulation. This thesis provides the groundwork for position effects on alternative splicing in the human genome, and highlights the need for such integrated studies with a wider variety of gene types.

Acknowledgements

Thank you to my advisor, Georg Seelig, for his mentorship over the course of my PhD, across several projects and through a global pandemic. He provided me with an opportunity to break into the field of synthetic biology and bioengineering in 2018, and without his support and guidance I could not have achieved the work that I have achieved since. I would also like to thank the members of my committee, Andre Berndt, Drew Sellers, and David Shechner.

I want to thank my parents and my sister for their unconditional support during this time as well, helping me with labor and love. Thank you to both my family and friends for bringing constant support and excitement to my life throughout these studies. Thank you too to my labmates, who all helped to make this time at UW not just educational and productive, but also enjoyable and thoroughly worthwhile experience.

Acknowledgement of Contributions

All of the projects in this thesis were performed with guidance from Georg Seelig, who has advised me on all aspects of research including design, execution, and communication. This thesis primarily demonstrates the results of my own work, with collaboration or advice from all members of the Seelig Lab. I want to acknowledge the contributions of **Nick Bogard** for initially proposing the self-reporting reporter assay concept, and **Johannes Linder** for his design of synthetic PolyA signal sequences in the self-reporting reporters as well as the development of APARENT2 and Borzoi predictive models used in both the self-reporting assay and the alternative splicing analysis, respectively. Thanks to **Gun Woo Byeon** for his assistance in applying Borzoi to my data. Additionally, much thanks to **Jason Hoffman** for his leadership in the SARS-CoV sensing project, which has been published in *The Science of the Total Environment*. Thank you to BPSD program coordinator Erin Kirschner and Bioengineering

program advisor for their support. Thank you to the UW Flow Cytometry Core Facility (DLMP) and their staff for assistance with cell sorting.

Table of Contents

List of Figures

List of Tables

Introduction

The majority of my PhD work has revolved around using massively parallel reporter assays (MPRAs) to directly study mechanisms of transcriptional regulation, as well as to improve upon existing technology for measuring transcription from the genome. MPRAs are particularly useful for studying complex, biological systems like gene transcription, where a large variety of regulatory sequences and gene arrangements can be tested in parallel. Thanks to the decreasing costs of high throughput sequencing and the increasing prevalence of machine learning models that can help to learn gene expression patterns from large amounts of data, MPRAs are becoming more useful than ever for better understanding transcriptional regulation within complex regulatory environments within the genome.

Alternative splicing (AS) is a major gene expression mechanism that contributes to eukaryotic transcriptional diversity and regulation. Studies have widely examined how *cis*-regulatory gene sequences and *trans*-regulatory elements such as protein factor expression directly drive alternative splicing patterns. While it is also known that the higher order, epigenetic structure of the chromatin landscape may interact with the mRNA splicing process, the significance of position-based chromatin structures and interactions in alternative splicing has not been thoroughly demonstrated.

While the majority of this thesis will discuss how we used MPRA approaches to specifically study alternative splicing regulation and genomic regulatory domains, I will also briefly cover experiences and challenges in applying large scale sequencing and quantification assays to disease-related applications. Due to overlap in sources, please note that the works cited section for chapters 1 and 2 are shared and presented at the end of chapter 2.

The topics that this thesis will cover are:

Chapter 1: A Massively Parallel Reporter Assay for Genomic Position Effects on Alternative Splicing

To better quantify the impact of a wide variety of genomic positions on local gene alternative splicing patterns, we present here the results of an MPRA that measures alternative splicing patterns in many novel genomic contexts. First, I describe an assay in which five synthetic reporters with varying predetermined alternative splicing activities are integrated across thousands of K562 chromatin contexts in order to observe position effects on expression. Measurements of splice isoform ratios from each individual reporter are cross-referenced with chromatin domains characterized by DNA-protein maps to infer domain effects on alternative splicing. The reporters ultimately demonstrate little predictable change in splicing patterns, and small yet consistent changes in transcriptional activity. Comparison of reporter features with endogenous gene expression and published reporters reaffirms that gene sequence is a stronger regulatory driver that may overwrite position effects on alternative splicing.

Chapter 2: A Self-Reporting Genomic Insert MPRA Integrates Positional Information into Transcripts with Truncated Polyadenylation Signal

In this chapter, I discuss the mechanism and validation of a novel MPRA Genomic Reporters Integrated in Parallel (GRiP), which aims to simplify workflows of genome-intragrated reporter library preparation and sequencing using a construct that self-reports its own integrated location in transcripts. Reporter gene cassettes are integrated into the genomes of K562 cells via the ubiquitous Non-Homologous End Joining (NHEJ) repair pathway, and include

a truncated Polyadenylation Signal (PAS) in the 3' untranslated region (UTR), causing transcription to terminate downstream of the reporter. Alignment of these downstream transcribed tags to the human genome demonstrated hundreds of reporter insertions across the genome, with individual locations verified by PCR and gDNA sequencing. I also demonstrated how using synthetic 3'UTRs designed by APARENT2, a deep residual network for engineering alternative polyadenylation, can also stall polyadenylation further to increase the length and mapping effectiveness of reporter-integrated genomic sequences. However, it is ultimately shown that truncated PAS elements also inhibit expression of the reporter gene. This chapter concludes with discussion of future possibilities for improving upon the MPRA design to circumnavigate these challenges in simultaneously measuring expression and integration from mRNA.

Chapter 3: Passively Sensing SARS-CoV-2 RNA in King County Public Transit Buses

I follow the theme of high throughput molecular assay design by describing a method for detecting SARS-CoV-2 RNA in public transit systems in order to approximately quantify the presence of disease-related particles across broadly defined areas. This work was published in the *Science of the Total Environment* journal in 2022 and co-first authored by Jason Hoffman and myself. Here, we placed sterile filters of various material types onto the air filters of King County Metro buses from August 2020 to March 2021 in order to catch viral particles as air is circulated. By collecting filters after the buses are in service for 2 weeks, we demonstrate that SARS-CoV-2 RNA particles may be accurately detected with phenol-chloroform isolation and qPCR, and quantification may be proportionally correlated to clinically reported cases in the

same county. This work demonstrates feasibility for detecting and tracking presence of viral particles in a sequence-specific fashion across public transit vehicles.

Massively Parallel Reporter Assay for Genomic Position Effects on Alternative Splicing

Summary

Alternative splicing of pre-mRNA into multiple isoforms is critical for transcript diversity in eukaryotes, but the wide variety of DNA, RNA, and protein interactions that regulate splicing makes it difficult to quantify each individual factor's contribution to isoform destiny. Splicing is most directly regulated *in cis* by the RNA transcript sequences that are necessary for assembly of the catalytic spliceosome, but importantly, interactions between splice factors and pre-mRNAs may also be regulated *in trans* in part by the three dimensional genome organization within the nucleus.

Here, we use a Massively Parallel Reporter Assay (MPRA) to study how alternative splicing patterns of a reporter gene vary in a context dependent manner by integrating cassettes randomly across the genome. By eliminating variance in gene promoter and sequence, the effects of surrounding sequences, DNA modifications, and chromatin organization on splice isoform preference and transcription can be better observed. Our results showed that while splice isoform ratios varied across different reporter cassette designs, the variance of splicing patterns and transcription was not greatly influenced by the genomic landscape around each reporter. Instead, gene promoter and sequence are shown to be most influential in determining alternative splicing and transcriptional activity. These results encourage further investigation of position effects on alternative splicing of introns with different splice strengths as well as genes with more chromatin state-sensitive promoters.

Introduction

Alternative splicing of mRNA allows one gene to produce multiple different transcripts, contributing greatly to eukaryotic gene diversity.[1–5] During pre-mRNA splicing, the 5' end of an intron termed the "splice donor" may ligate to more than one possible "splice acceptor" at the 3' end of the intron before the pre-mRNA length between those ends is spliced out and removed from the transcript. Exclusion of exons between the splice donor and acceptor potentially changes the coding sequence or reading frame of the final mRNA. In the most common splice pathway, the strength of each splice acceptor is directly dependent on how efficiently the spliceosome assembles on the intron, which is further dependent on a number of factors.[6]

Splicing is first initialized by recruitment of small nuclear ribonuclear proteins (snRNPs) to the pre-mRNA, which recognize the 5'-splice site donor AG dinucleotide, the 3'-splice site acceptor GU dinucleotide, the polypyrimidine tract (PPT) upstream of the splice acceptor, and the branchpoint sequence. The branchpoint sequence and location 40nt upstream of the 3'-splice site are generally conserved, but variance in these sequences also results in variable spliceosome efficiency or different mechanisms altogether, as in U2 or U12 spliceosomes, or ribosomal self-catalyzed splicing. Additional *cis*-regulatory sequences dubbed Exonic or Intronic Splice Enhancers (ESE, ISE) or Silencers (ESS, ISS) may also recruit SR proteins, splice factors, or inhibitory proteins to promote or silence assembly of the spliceosome[2,4,7,8].

Once assembled, the spliceosome complex catalyzes cleavage of the intron to be removed at its 5-splice site, then looping of the pre-mRNA into a lariat that brings the 5' end into close proximity with the 3'-splice site, and allows the GU guanine to covalently bond to the AG adenine, connecting the two exons that flank the spliced-out intron.[2,9–11] The importance of these various elements for efficient spliceosome assembly makes splice initiation heavily sequence-

dependent, so we next describe why we would expect for position effects to impact alternative splicing.

Splicing typically occurs co-transcriptionally, creating a complex environment where the pre-mRNA and spliceosome are interacting with the RNA polymerase and DNA. RNA Pol II itself may bind and recruit splice factors at its C-terminal end, or speed of elongation may in some cases limit the time in which the spliceosome can assemble around each splice junction of the emerging pre-mRNA. Transcription speed is in turn regulated by intron length, chromatin modifications, and nucleosome positioning, creating additional layers of potential regulation on top of the intrinsic intron sequences.[2,12]

Nucleosomes carrying histone modifications such as H3K4me1 are enriched around alternatively splicing sequences in endogenous genes. The contribution of these modifications may be explained by the "kinetic model" as contributing to alternative splicing by inhibiting or promoting polymerase speed, as well as recruiting splice factors. However, nucleosomes are also conserved outside of alternatively splicing genes. In the "chromatin-adaptor model", enrichment of certain chromatin signatures is associated with modified splice efficiency, theorized due to the recruitment of splice factors by chromatin modifications *in-trans*. It remains to be studied whether these sorts of modifications that are near but not directly in alternatively spliced genes may still interact and affect splicing patterns.[13,14]

While some epigenetic DNA modifications have been associated with increased or decreased splicing efficiency, it has been shown that the causal relationship between transcription and splicing may also work in reverse. Splicing of exons may activate weak promoters through a recruitment model in which splicing factors recruit transcription machinery, as demonstrated by insertion of an intron into a gene that results in increased upstream promoter

activity.[13,15] In rare instances, splicing-associated chromatin signatures have also been shown to appear after insertion of spliced introns, further obfuscating the role of splicing-related features on a case-by-case basis.[16]

Finally, there is a level of spatial and temporal coordination to these processes, as it has been shown that genes are compartmentalized within the nucleus to regions associated with varying levels of transcription or splicing. For example, dense regions of concentrated splice factors and transcription factors called nuclear speckles contain groups of genes that have higher splicing and transcriptional efficiency than genes outside of these speckles.[7,17–23]

Conversely, genes may be marginalized to the nuclear lamina in regions termed Laminar Associated Domains (LADs), where chromatin tends to be more compacted and protein factors may bind less efficiently. These domains have been strongly correlated with promoter repression or reorganization around promoters, but not thoroughly studied with splicing.[24–30] The influence that three-dimensional genome organization has on individual gene splicing and transcription raises the question of how alternative splicing behavior of a gene might change in a position-based manner, independent of the gene sequence and canonical splice junction motifs.

Here, we investigate how position effects may affect alternative splicing of a gene in thousands of different contexts using the massively parallel reporter assay Thousands of Reporters Integrated in Parallel (TRIP).[26,27] While this integrated reporter assay has previously been used to demonstrate domain-dependent promoter and gene transcriptional activity, the position effect on alternative splicing of reporter genes has yet to be shown. We used this established method to insert five types of alternatively splicing reporter cassettes randomly across the genome before measuring transcript expression and splice isoforms at each mapped location. Each three-exon reporter is designed to have an episomal baseline pattern of alternative

splicing, in which two possible isoforms may be expressed. Alternative splicing is then measured with the statistic "Percent Spliced In," or PSI, a ratio of the middle exon retention in final transcripts.

We observe that the rate of exon inclusion in a reporter gene's transcripts was not significantly influenced by position relative to defined topologically associated domains that have been previously shown to affect transcription and splicing (such as LADs and Speckle Associated Domains, or SPADs) nor chromatin markers associated with alternative splicing.

Understanding how position affects splicing could be potentially utilized in synthetic biology as transgene expression may vary depending on the surrounding context it lands into. Data from a reporter assay could also further refine predictive models that are otherwise trained on limited endogenous context-gene sequence combinations. In the future, better modeling of position-based interactions could possibly enable cell stage or tissue specific expression of synthetic or engineered genes that use alternative splicing to produce multiple transcripts.

Results

Integrations Map to Diverse Locations, Matching PiggyBac Behavior

Cassettes and Super PiggyBac Transposase vector plasmids (Systems Biosciences) were constructed and transfected into K562 cells as described in methods (**Figure 1.1A & B**). Pb transposase was expected to integrate ITR-flanked reporters into a wide variety of random positions besides the requirement of TTAA motifs[31,32]. Bowtie2 alignments of mapping libraries demonstrated consistent localization to endogenous TTAAs distributed equally across all chromosomes, as well as expected avoidance of centromeres as pb repressive regions (**Figure 1.1C**). Integrity of mapping was also compared to tendency to insert into LADS. Approximately

30% of mapped inserts aligned into LAD domains, matching previously published data in K562 cells[26]. Finally, barcoded cassettes in each replicate that independently mapped to the genome were compared to each other and found over 99.9% of aligned coordinates were in agreement within 50 bases. Cassettes that aligned to different coordinates in each replicate were discarded.

Transcript Quantification and Insert Mapping were Highly Correlated between Replicates

Expression of integrated reporters was measured through RNA-sequencing and identification of the exon in each transcript. The sum of transcripts sharing a unique combination of barcodes was used to calculate normalized expression and PSI, then paired with matching barcodes mapped uniquely to the genome. Barcodes with non-unique mapping data were thrown out due to uncertainty of position. Transcripts without paired mapping data were also not included in downstream calculations. Formulas for normalized expression calculation are provided in **Methods**. For barcodes that passed this filter in both replicates, associated transcript counts and gDNA normalization counts were highly correlated, demonstrating that expression was consistent across separate cultures and library preparations (**Figure 1.1B**).

Ultimately, we recorded 1,846 unique insertions across the genome with both mapping and normalized expression data, verified by two replicates (**Figure 1.1C**). An additional 4,030 cassettes had mapping and normalized expression data in one replicate, but did not appear in the other. Due to high correlation of all transcript, normalization , and mapping between replicates (**Figure 1.1B, 1.2A**), it is inferred that the majority of these additional cassette measurements are still accurate.

Notably, the individual transcripts and normalization counts per replicate had better correlation than the fully normalized expression values. While normalization is calculated to

account for differences due to uneven cell duplication in culture, uneven PCR cycling, or uneven numbers of sequencing reads, the high transcription correlation suggests that culturing and PCR factors were low-impact, and that both replicates were deeply sequenced for all cassettes.

Variation of Cassette Expression is Not Identical Across Reporter Types

Normalized expression is shown to generally correlate across replicates, demonstrating consistent transcriptional activity by position (**Figure 1.2A**). Because some cassettes consistently express more than others, we investigated whether the variation could be attributed to position effects and gene sequence. Accounting for all cassettes with and without replicates, the average normalized expression was equal for WDR35, ACADVL, and MYH11 cassettes, but consistently lower for COLQ and ASAH1 cassettes. To ensure that this transcriptional disparity was not due to uneven normalization counts, **Figure 1.2B** also demonstrates that the distribution of normalization counts per cassette is very similar across cassettes of all types. This suggests a sequence-dependent effect on transcription dictated entirely by the middle exon and flanking introns, despite the remainder of the gene sequence to be identical across all cassettes. It has yet to be shown whether PCR efficiency varies between cassettes due to this sequence difference. Potential RNA-binding protein motifs appear absent for COLQ in **Figure S1.3,** suggesting that the difference in expression may be instead splicing-related.

Transcription Variation is Affected by Repressive Topologically Associating Domains Less Than Previously Observed

To determine whether variation in individual cassette transcriptional activity is affected by position effects, we compared mean cassette expression for populations that align to known repressive domains like LADs to cassettes that aligned outside of these regions. It was previously

shown by Leemans, et al.[26] that integrated reporters in K562 cells have a lower mean expression when they align to regions defined as LADs by LaminB1 protein footprinting. Nucleolar Associated Domains (NADS) are also frequently associated with organizing repressive chromatin states[33], and so we identified cassettes with alignments that intersect both LADs and NADs with the expectation that they represented the most likely to be repressed according to their position (**Figure 1.2C**). We also considered SPADs as possible positional features that could be expected to positively influence general gene transcription in addition to efficient splicing[19,22]. Altogether, we observed that cassettes that align to either LADs alone or LADs+NADs do express less on average compared to those that do not. **Figure 1.2D** shows the population mean and median differences for cassettes within and outside of these regions for WDR35 and ASAH1. The median fold-changes were observed between 1.1 and 1.3 of the original normalized expression means before log transformation, with average means also higher in cassettes outside of LADs+NADs. This was compared to a greater median fold difference of 1.9 observed for hPGK1 promoter GFP reporters in K562 experiments by Leemans et al. This result shows that all of our five cassettes designs are less affected by laminar associated domains than previously published data for similar cassettes with the exception that those are unspliced reporters lacking introns.

Cassette expression was also checked against speckle-associated domains which are known to be regions where transcription factors are highly concentrated[19,22]. Because SPAD regions are nearly mutually exclusive from LADs, in order to separate their effects, only cassettes that do not align to LADs were checked for intersection with SPADs (**Figure 1.2D**). Cassette median expression appeared to be unaffected by these domains, and means were not statistically significantly different. Plots for all five reporter variants are shown in **Figure S1.2.**

Average PSI of cassettes was found to be highly stable across positions

It was found that the calculated PSI of all cassettes for each reporter construct was consistent between locations and replicates. Initially, we calculated the PSI as the ratio of transcripts with the variable exon spliced in over all transcripts with the same cassette barcode. A minimum threshold of transcripts is applied to remove outliers where the ratio of small integers may have high error. With a low minimum of 5 transcripts per cassette, comparing PSIs for cassettes across replicates shows high variance, but a high minimum of 50 transcripts per cassette demonstrates a much lower range and variance of cassettes. Overall, for all cassettes, the middle 50% of cassette PSIs was shown to be very stable around a single value for cassettes of each type (**Figure 1.3A**).

In WDR35 transcripts, the middle exon is notably spliced in 0% of the time exclusively. The remaining cassettes with a minimum of 50 transcripts demonstrated averaged values as follows: COLQ exons have an average PSI = 20.54%, ACADVL PSI = 34.31%, MYH11 PSI = 69.03%, and ASAH1 PSI = 81.45%. The rank of PSI from low to high matches the ranking measured from plasmid expression by Koplik, et al.[6] The precise average values between the two experiments did not match, though the standard errors of COLQ and ACADVL in the episomal experiment preclude this expectation (**Figure 1.1A**). While ASAH1 exons were observed to be spliced in 100% of the time with high confidence in the episomal experiment, this was not observed in our genomic experiment. While we posit that the majority of differences between genomic average PSI and episomal PSI is due to the genomic environment, as the intron and exon sequences are identical, though the episomal EF1α promoter has also been swapped out with an hPGK1 promoter in this experiment.

The range of PSIs for each reporter type is low, with interquartile ranges between 0% and 6.53% at most. This low range suggests that different genomic positions do not have a large average effect on PSI. Due to the transcript threshold, this value precludes cassettes with low transcripts. Some outlier ASAH1 and COLQ cassettes appear to have a consistent PSI value of 0, despite the high number of transcripts as well as consistency between replicates. These did not appear to insert into similar coordinates, but enrichment for domain-associated features is further investigated in **Figure 1.3C.**

H3K79me2, H3K27ac, and H3K9me3 Mean Signal Appears Weakly Enriched around Outlier PSI Inserts

Several splicing-associated chromatin signatures (SACS) have been identified in human cell lines to be associated with increased exon inclusion or exclusion in alternatively splicing genes. Here, we used DeepTools ComputeMatrix and PlotProfile functions to investigate whether some of these were enriched around cassettes that demonstrated above-average or below-average PSI values relative to other cassettes of the same type. The top 5th percentile and bottom 5th percentile were compared to cassettes within the interquartile range (between the first and third quartiles) against H3K9me3, H3K27me3, H3K4me1, H3K27ac, H3K36me3, and H3K79me2. H3K4me1 was previously shown to be associated with exon inclusion in endogenous K562 genes, and H3K79me2 was shown to be associated with exon exclusion, while the remainder are associated with both inclusion and exclusion of different genes[13]. Results for ASAH1 are shown as an example in **Figure 1.3C**, and the rest of cassettes are shown in **Figure S1.1.** We found that in all cassettes, H3K79me2 was slightly enriched around high PSI cassettes relative to all other cassettes, in contrast to the previous characterization with endogenous genes. H3K9me3 and H3K27ac were slightly depleted around both high and low PSI cassettes for

COLQ and ASAH1, but in ACADVL they were enriched upstream of high PSI cassettes for ACADVL only. In MYH11 cassettes, H3K27ac was also slightly enriched upstream of high PSI cassettes, and there was no differential presence of H3K9me3. While H3K4me1 is associated with exon inclusion in literature, it showed no significant relationship with high or low PSI cassettes in all designs.

Transcription Rate and Splicing Do Not Strongly Correlate in Reporters

After observing that there was no strong correlation between well-characterized topological domains and PSI or transcription, we checked whether splicing was affected by transcription in general, regardless of whether the expression rate was affected by position or only stochastic differences. Correlations of normalized expression and PSI in log scale are shown in **Figure 1.4A** which demonstrate trends with low Spearman rank correlation values between 0.09 and 0.27. These were observed as negative relationships for COLQ and ACADVL and positive for MYH11 and ASAH1, which by default have higher base exon inclusion. In order to directly observe any effects of transcriptional activity on splicing, we also checked for correlations between transcript count and PSI (**Figure 1.4B**). It was seen that PSI is more varied for cassettes with lower transcript counts as expected, and deviation decreases at higher transcript counts. Notably, ACADVL appeared to have a very small positive correlation between splicing and transcript counts at a very high mRNA count, but otherwise there was no statistically significant correlation between the two. Finally, to observe whether expression of one isoform over the other is significantly changing without calculating PSI which is subject to high variation at low transcript counts, we also checked the replicate consistency of transcription for each splice isoform separately (**Figure 1.4C**). These plots show us that for both transcripts spliced in and out, cassettes are generally consistent across cultured replicates at all levels of

transcription, with more inconsistency for low counts of transcripts with the exon spliced-in for COLQ, and low transcripts of either isoform for ASAH1. This observation reinforces the higher PSI variation in low normalized expression cassettes for COLQ and ASAH1 in **Figure 1.4A.**

Tet-Inducible Promoter Reporters Demonstrate No Change In Expression Within Laminar Associated Domains

Because variation in transcription or PSI for each cassette type was not as strongly influenced by topologically associated domains to the extent that has been previously demonstrated with similar integrated reporter assays, we performed a secondary experiment where the cassette promoter was replaced with a Tet-on TRE2GS promoter. We sought to investigate whether hPGK1, a promoter that endogenously appears in LADs and is less sensitive to LAD repression than other tissue-specific promoters[26,34], perhaps overrides position effects. In this experiment, a different promoter was selected with an inducible level of activity to observe whether cassettes would display stronger reactions to position effects.

The experiment was carried out in K562 cells in the same way as described in **Figure 1**, with the exception that one cultured set of cells received 200ng/mL doxycycline in culture after sorting, and the other was uninduced. Additionally, there was only one replicate for each experiment. Comparing normalized expression of cassettes that aligned within LADs to those that do not, we observed no change in median transcription between the two for both induced and uninduced cassettes, though the mean expression of uninduced cassettes outside of LADs was higher by 0.006 (a 1.2-fold change in mean) with a t-test p-value of 0.0123.

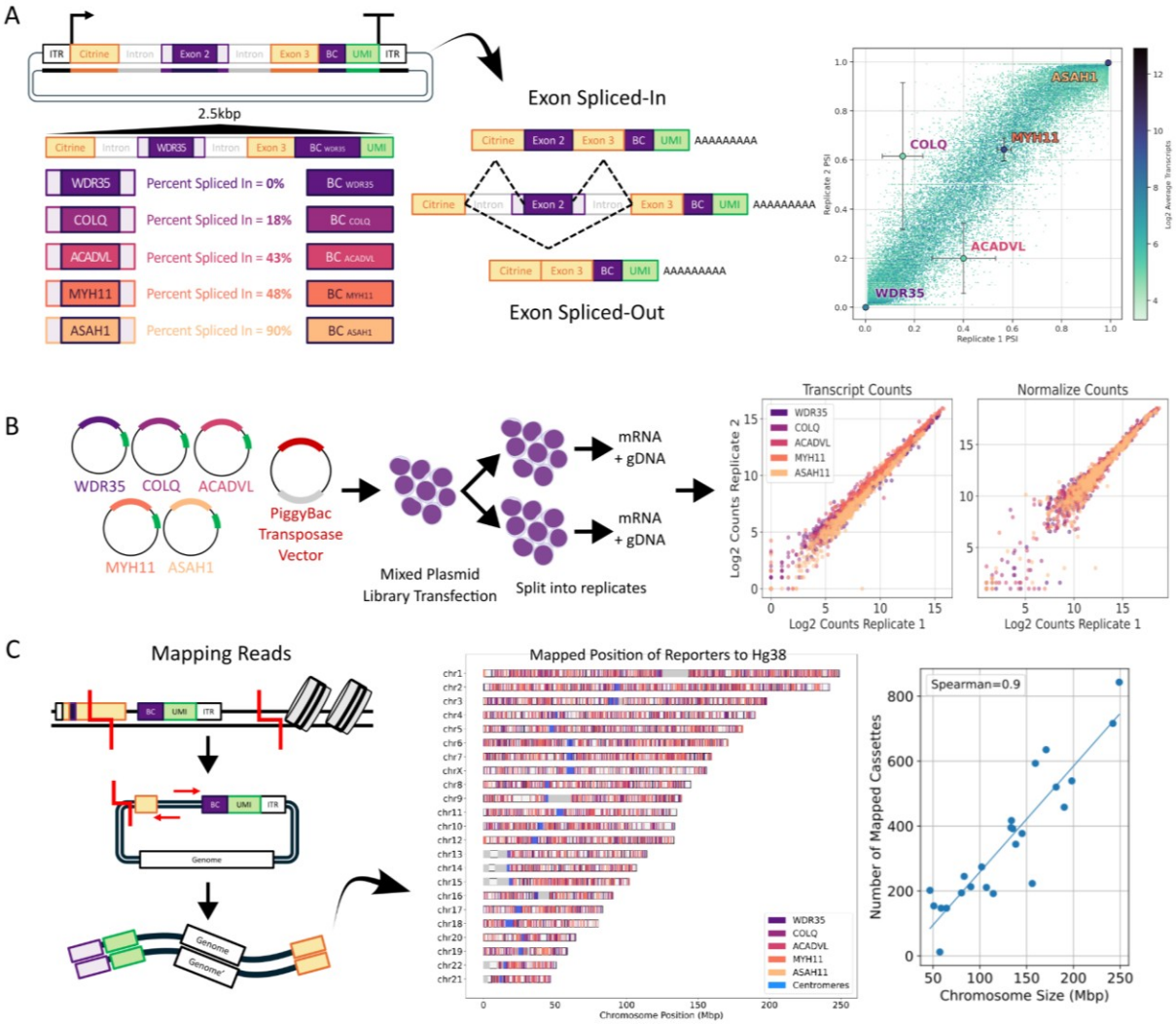


Figure 1.1. Overview of the reporter assay methodology

(A) Reporter cassettes carry a three-exon gene that expresses two possible splice isoforms, with the middle exon either included or excluded. The five reporter variants are distinguished by the middle exon and flanking partial intron sequences. Splice barcodes corresponding to the variable exon and a plasmid-specific UMI are located in the 3' UTR in order to identify variable exons that are spliced out in the final transcript. Reporter cassettes are flanked by ITRs recognized by Super PiggyBac transposase. Variant exons WDR35, COLQ, ACADVL, MYH11, and ASAH1 were chosen to represent a range of episomal PSI values from low to high. These episomal PSIs were previously measured in a similar reporter by Koplik et al[6]. **(B)** A mixed library of all five splice constructs and the Pb vector is transfected into K562 cells. Cassettes integrate into the host genome via transposition for 48 hours before sorting on mCherry fluorescence expressed by the Pb vector. Aliquots of 1,000 sorted cells are allowed to culture for one week then split into replicate pools. Replicates are cultured for another week before gDNA and mRNA are harvested from each replicate and expression libraries and normalization libraries are sequenced. Correlation of transcript and normalization counts from paired replicates are shown. **(C)** Part of the gDNA from (B) is used for a mapping library. Insert locations are determined by digesting

gDNA with DpnII which cuts upstream of the 3' UTR barcodes and randomly downstream in the genome, then ligating and amplifying these fragments with inverse-PCR, and finally sequencing the downstream genomic sequence. Karyotype plot shows locations of uniquely mapping sequences for each cassette type which also agree between replicates. The number of alignments per chromosome correlates with chromosome size.

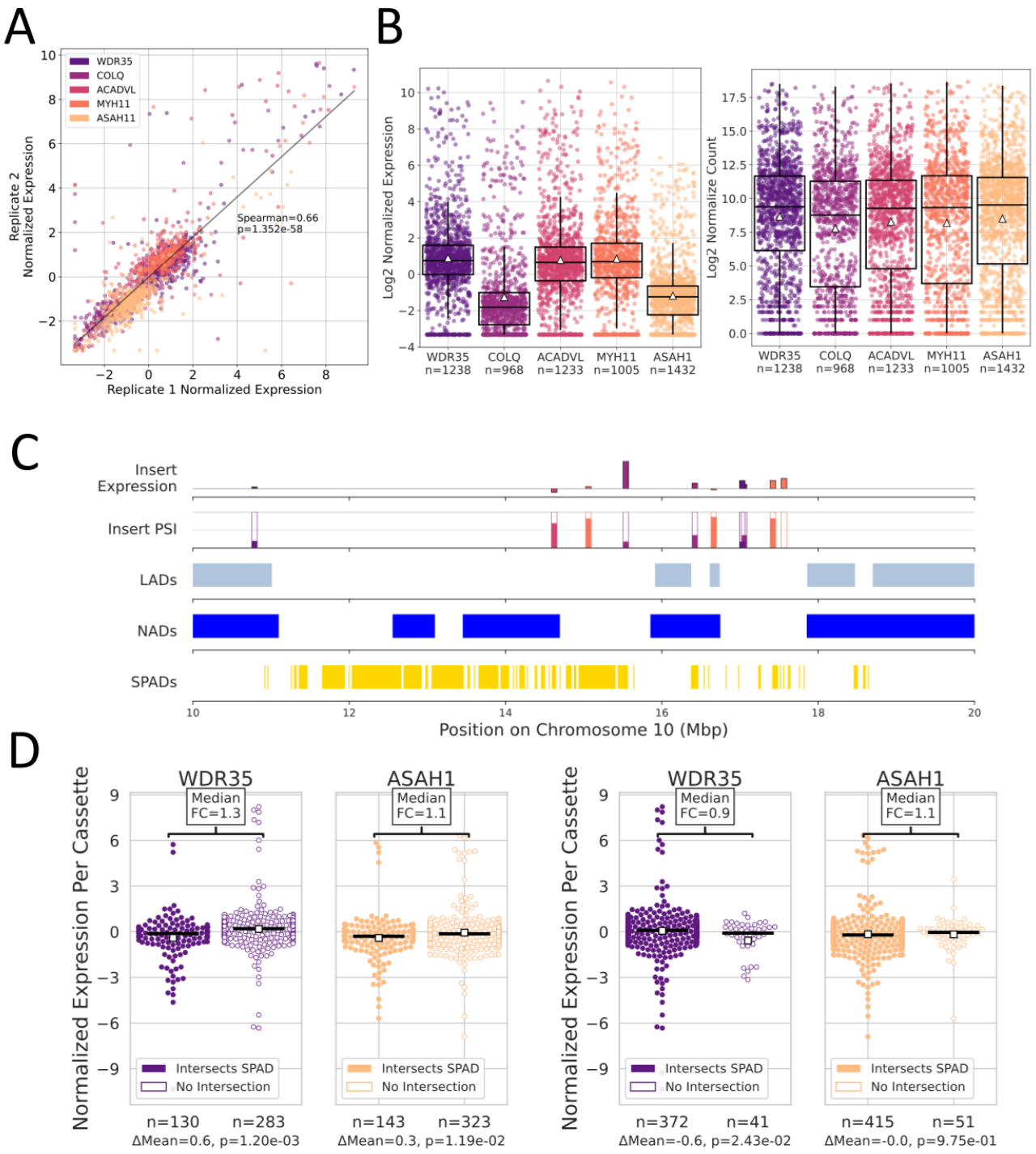


Figure 1.2. Transcription measurements by reporter type and location

(A) Correlation of normalized expression levels of each cassette present in both replicates. **(B)** Normalized expression for all cassettes is shown by cassette type, with averaged expression for cassettes that appeared in both replicates. Medians are the center bar of the box plot, and means are represented by a white triangle. COLQ and ASAH1 cassettes demonstrate lower expression on average. Normalization counts are plotted on the right to demonstrate similar counts of all insert types despite differing normalized expression means. **(C)** Uniquely mapped insert domains are defined here by overlap with discrete tracks obtained from ENCODE data sets (Table S1.3). LADs are defined by LaminB1 signal, NADs are defined by 4X-AP3 signal, and SPADs are defined by SON signal. Footprinting datasets are outlined in Table S3. **(D)** Left- Comparing the expression of cassettes that map to both LADs and NADs versus those that do not. Population medians are represented by the black bar, and means are represented by the white square. Significance of the mean difference is determined by independent t-Test. Cassettes express slightly more on average when outside of these repressive regions. Right- the same type of comparison is made for cassettes that intersect SPADs.

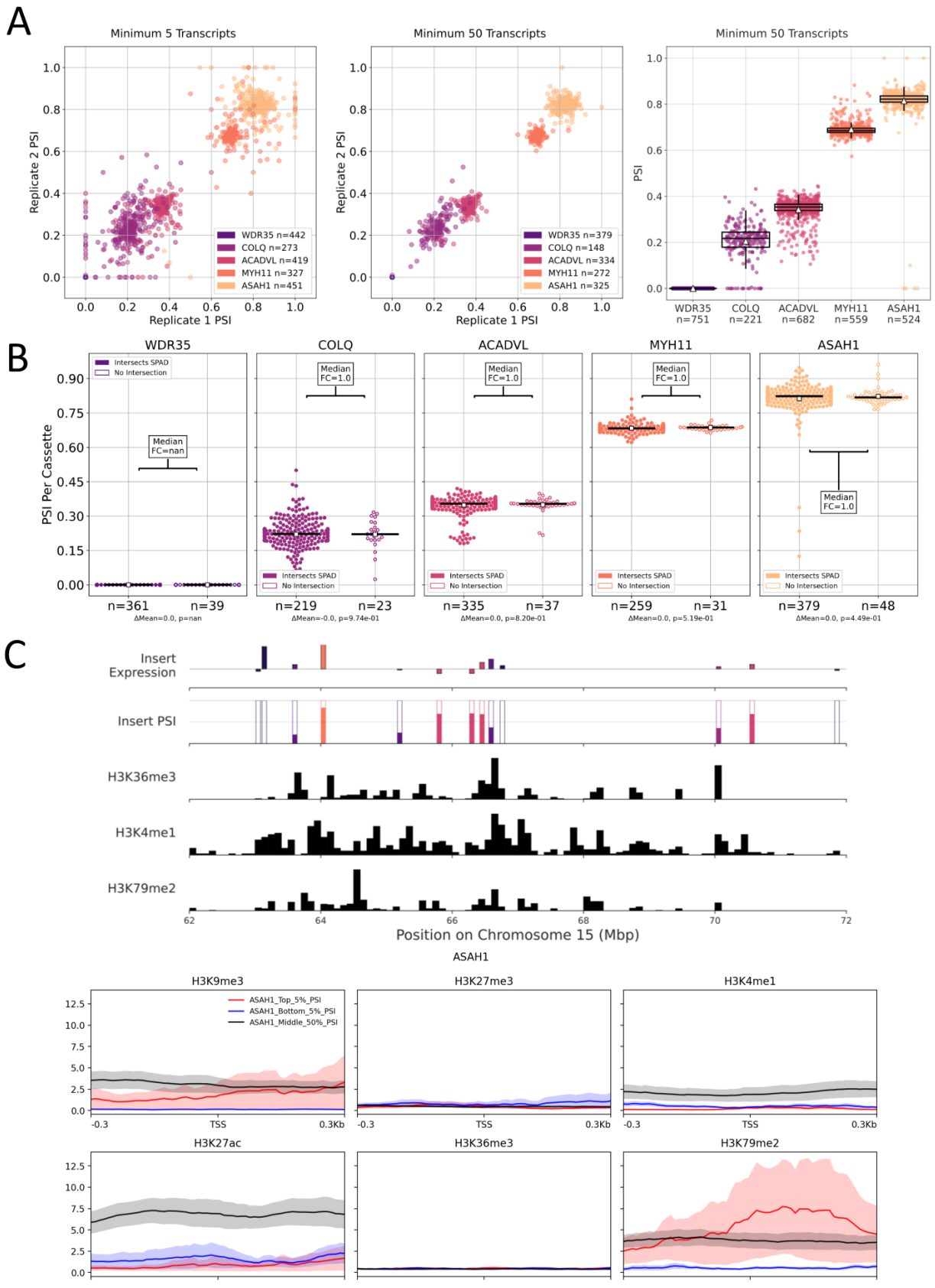

A
Minimum 5 Transcripts
Minimum 50 Transcripts
Minimum 50 Transcripts
Replicate 2 PSI
Replicate 1 PSI
PSI
WDR35 n=442
COLQ n=273
ACADVL n=419
MYH11 n=327
ASAH1 n=451
WDR35 n=379
COLQ n=148
ACADVL n=334
MYH11 n=272
ASAH1 n=325
WDR35 n=751
COLQ n=221
ACADVL n=682
MYH11 n=559
ASAH1 n=524
B
WDR35
COLQ
ACADVL
MYH11
ASAH1
PSI Per Cassette
Intersects SPAD
No Intersection
Median FC=nan
Median FC=1.0
n=361
n=39
n=219
n=23
n=335
n=37
n=259
n=31
n=379
n=48
ΔMean=0.0, p=nan
ΔMean=-0.0, p=9.74e-01
ΔMean=0.0, p=8.20e-01
ΔMean=0.0, p=5.19e-01
ΔMean=0.0, p=4.49e-01
C
Insert Expression
Insert PSI
H3K36me3
H3K4me1
H3K79me2
Position on Chromosome 15 (Mbp)
ASAH1
H3K9me3
H3K27me3
H3K4me1
H3K27ac
H3K36me3
H3K79me2
ASAH1_Top_5%_PSI
ASAH1_Bottom_5%_PSI
ASAH1_Middle_50%_PSI
TSS
-0.3
0.3Kb

Figure 1.3. PSI by reporter type and location
(A) Left- PSI of cassettes that map within each replicate are compared. Only cassettes with a minimum number of transcripts are used to calculate PSI. Middle- Higher thresholds result in less standard deviation. Right- Comparison of PSI for all cassettes by cassette type regardless of replicate status. Notably at a high threshold, some COLQ and ASAH1 cassettes report a consistent PSI of 0. **(B)** Cassettes in both replicates with an arbitrary minimum threshold of 20 transcripts are mapped against SPADs and PSIs for intersecting and non-intersecting cassettes are plotted. **(C)** Cassette insertions that land within regions with splicing-associated chromatin signature signals (Table S1.3) are checked for correlation between extreme splicing and SACS enrichment. Cassettes within the top 5th percentile, bottom 5th percentile, and middle 50th percentile (the second and third quartiles) are plotted for enrichment of six SACs. Lines with standard error ranges are plotted for distance relative to original insertion site (TSS), not including the ITR and promoter upstream of the TSS nor the length of the reporter gene downstream. Shown are the results of the ASAH1 traces. All other cassette plots are in Figure S1.1.

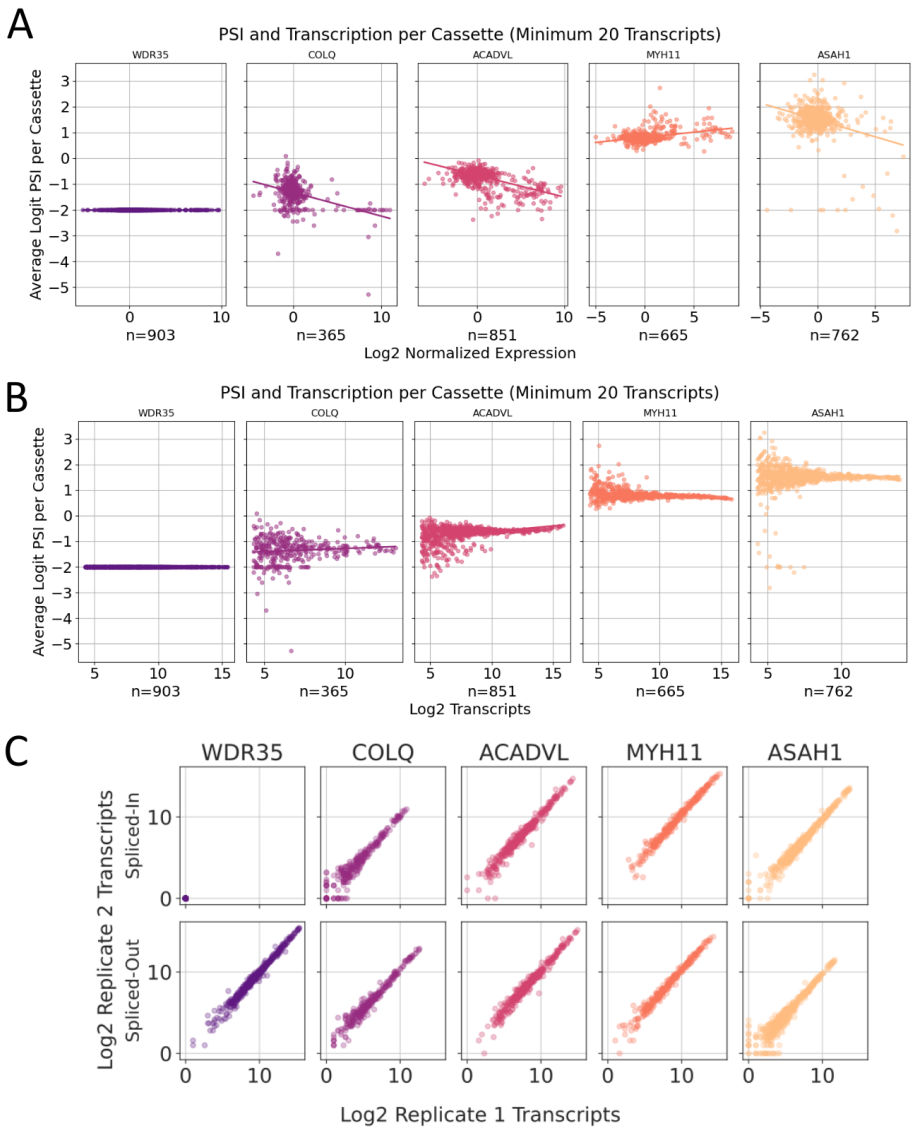


Figure 1.4. PSI does not strongly correlate with general transcription nor normalized expression

(A) Normalized expression is compared against logit PSI for each cassette with at least 20 transcripts. PSI is applied on the logit scale to better match the log scale on the x-axis. PSI becomes more distributed at higher normalized expression values. Line of best fit is plotted for each construct. Spearman correlations for COLQ, ACADVL, MYH11, and ASAH1 are -0.27, -0.23, 0.2, and -0.09, respectively. **(B)** Logit PSI is also correlated with direct log-scale transcript counts per cassette to avoid noise of normalization counts. In these plots, PSI is noisier at lower transcription counts, suggesting that the range of high expression cassettes from (A) is due to normalization data noise. **(C)** Correlations of transcripts with the variable exon spliced in and transcripts spliced out between each replicate for each construct are plotted. These show without directly calculating PSI that the number of any particular splice isoform is consistent for each insert.

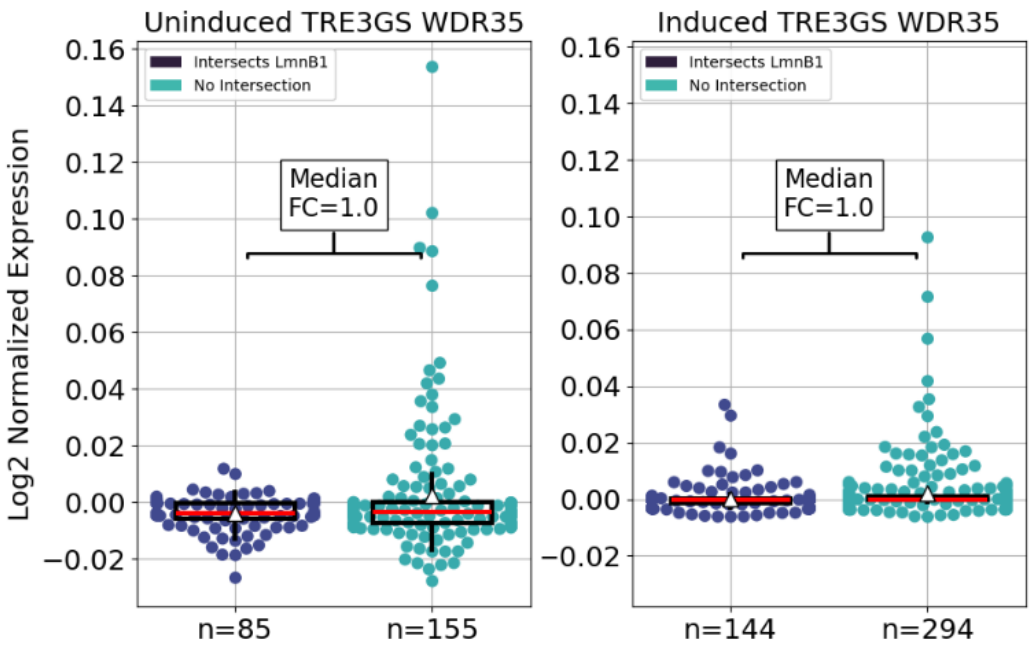


Figure 1.5. Dox-induced promoter reporters show no change in expression across LADs
Reporter cassettes with a TRE3GS doxycycline-inducible promoter were transfected and split into two sets of cells, one induced with doxycycline and one uninduced. In both experiments, the normalized expression median represented by the red bar in the boxplot did not change inside and outside of LAD domains. Means represented by the white triangle were different for the uninduced group by 0.006 with a t-test p-value of 0.0123.

Discussion

Here, we inserted reporter cassettes with previously measured ratios of two expressed splice isoforms across thousands of random positions in the human genome and observed small variances in transcription and PSI between reporters. While insertion into repressive regions like Lamina-Associated Domains was expected to have a significant effect on repressing transcription, there was only a minor difference in average expression between LAD and inter-LAD cassettes with the hPGK1 promoter. Additionally, splicing activity varied very little across all cassettes regardless of position.

While we expect gene sequence and promoter to be the strongest factors in determining transcription and splicing behavior of genes, there was reasonable expectation that transcription

should be sensitive to position due to previously demonstrated work that interactions with repressive topologically associated domains or enhancers should inhibit or enhance expression[26,34,35]. Because that was not seen here, we compare our reporter construct to existing methods and recognize that the splicing activity of the reporter in this work sets it apart from other integrated reporter experiments. It is possible that splicing activity may influence transcriptional activity of the gene due to recruitment of transcription factors by the spliceosome[2,15,21], so in order to test whether this occurs and obfuscates position effects on transcription, we will next aim to use an integrated reporter with no introns as a negative control.

The hPGK1 promoter, while demonstrated to show differential expression by position, may also be responsible for the low variation in transcription due to its status as an endogenous LAD promoter and highly active in all cell types (per Uniprot). While we also tested the TRE3GS promoter, both promoters appear to be less sensitive than others that are documented to be even more sensitive to position and heterochromatin. It has also recently been shown that promoter identity has specific effects on exon inclusion, regardless of expression level[36], so it would be prudent to test a variety of promoters to find one that is least likely to interfere with weaker position effects in order to improve this approach to studying how alternative splicing varies with position.

We selected our reporter variable exon and intron sequences to represent a range of baseline PSIs that may further be influenced by position effects. Besides such effects like inserting into regions near speckle domains and high concentrations of splice factors[18], histone modifications were also investigated as potential neighboring elements that could influence splicing of reporter cassettes. Several histone modifications are associated with alternative splicing and hypothesized to function by interacting with and recruiting protein factors involved

in transcription, splicing, and further genomic rearrangement[13,14]. Although these markers are primarily found within the actual exons and introns of the gene sequences that they are associated with, because they have been observed to be evolutionarily conserved outside of coding regions as well, we explored whether adjacency to our inserts could influence splicing, and it was found that H3K79me2, H3K27ac, and H3K9me3 markers were slightly correlated with exon inclusion or exclusion, depending on the exon in the reporter.

Beyond this correlation with a small number of outlier PSI inserts, the majority of reporters demonstrated very little deviation from a population PSI mean. We hypothesize that the intronic splice junctions that are included around each variable exon in the reporter constructs may be responsible for the consistent splice behavior, as the canonical splice junctions in each (**Figure S1.3**) may be unaffected by repressive domains, or capable of splicing efficiently even without being recruited to enhancing domains such as SPADs. In order to investigate why our alternative splicing reporter was rather insensitive to position effects, as well as to potentially develop a more sensitive reporter that is more suitable to testing these effects, we plan to test reporters with previously demonstrated weaker splice junctions such as CORO1B exon 5 from Bhat, et al.

Ultimately, this work demonstrated that we could successfully measure the alternative splicing activity of reporters integrated across the genome in high throughput, and that the genomic expression and splicing of these cassettes appears to be greatly dictated by the sequence of the introns and promoters more than position. In the future we hope to establish whether this paradigm applies to additional reporter designs, and to potentially develop a more position-sensitive reporter.

Methods

Plasmid Construction

Super PiggyBac Transposase Expression Vector plasmid (PB210PA-1) from System Biosciences was modified to add an mCherry reporter expressed from a CMV promoter for detection by flow cytometry.

Plasmids carrying five variants of a transposable, alternative splicing reporter cassette were constructed. The cassette contains a citrine reporter gene driven by the hPGK1 promoter and is flanked by PiggyBac ITRs. The citrine gene is split into two exons separated by the SMN2 intron 6, a variable sequence, and SMN2 intron 7. Each variable sequence consists of a single exon flanked by 20-76bp of its endogenous intronic sequences. The exons WDR35 exon11 (from RefSeq NM_001006657.2), COLQ exon 5 (from RefSeq NM_005677.4), ACADVL exon 3 (from RefSeq NM_000018.4), MYH11 exon 42 (from RefSeq NM_001040113.2) and ASAH1 exon 4 (from RefSeq NM_177924.5) were chosen based on their wide range of exon inclusion measurements from Koplik et al. (2025). The 3' UTR contains a 20nt splice barcode that uniquely corresponds to the variable middle exon, and a 10nt cassette barcode that is random for each plasmid. A NotI recognition sequence was inserted 79nt downstream of the reporter cassette's 3' ITR for digestion of cDNA transcribed from un-transposed cassettes. DpnII recognition sequences were carefully placed further downstream of the NotI site and upstream of the splice barcode within the second citrine exon. Importantly, no GACT recognition sequences are present between the barcodes at the 3' UTR and the downstream NotI site, so that DpnII digestion and circularization of the cassette would not disrupt the barcodes.

The Dox-Inducible construct includes a 2.3kb fragment between the ITR and the reporter gene that contains an rtTA gene expressed by an EF-1-α core promoter and a TRE3GS promoter for expressing the citrine reporter. The constitutively expressed construct instead includes a 548bp fragment between the 5' ITR and reporter gene that exclusively contains the hPGK1 promoter. This version also includes a NotI restriction motif downstream of the ITR at the 3' end of the reporter gene which cleaves untransposed cDNA during library preparation.

All plasmids were assembled with NEBuilder HiFi DNA Assembly mix and transformed into NEB 10-beta Electrocompetent E. coli with a BioRad Micropulser Electroporator. A subset of each culture was spread on an ampicillin plate overnight to estimate total barcode diversity between 500k and 20 million for each construct, as in Box 1 from Akhtar, et al. (2013). Plasmid libraries were purified using ZymoPURE II Plasmid Midiprep Kits to remove endotoxins and sequenced with Plasmidsaurus Whole Plasmid Sequencing service to ensure sequence fidelity.

Cell Culture

K562 cells were transfected according to the Thermofisher Neon Transfection protocol with 100µL tips. Briefly, one million cells were washed with Gibco PBS (without Ca^{2+} and Mg^{2+}) and passed through a 40µM cell strainer before re-suspending in 100µL Resuspension Buffer R and 9µg of DNA, which included 1.2µg each of five plasmid constructs in one mix, and 3µg of Super PiggyBac Transposase Expression Vector (System Biosciences) for a total 2:1 transposon:transposase ratio.

Transfected K562 cells were cultured in RPMI 1640 Medium (Thermo Scientific) supplemented with 10% FBS but no antibiotics in a 6-well plate. All cultures were maintained at 37 °C, 5% CO_2 in a humidified incubator. 48 hours after transfection, cells were sorted with a BD FACSymphony S6 and cells with intermediate mCherry signal corresponding to a medium

amount of transfected transposase were isolated. Aliquots of 10,000 sorted cells were then cultured separately in 6-well plates with the same RPMI 1640 Medium as before, replacing media every other day without discarding cells. 6 days after sorting, independent aliquots were split into two and separately cultured as replicates in 50mL flasks for another 12 days (18 total days since sorting, and 20 days since transfection).

Library Preparation

Cells were harvested using a Qiagen AllPrep kit to simultaneously extract gDNA and mRNA. Purified mRNA was reverse transcribed with SuperScript IV Reverse Transcriptase and a gene-specific primer that adds an 8nt UMI. Part of the gDNA was used to create the Mapping library by digesting with DpnII at 37°C overnight and ligated with NEB T4 DNA Ligase at 4°C overnight, cleaning the reaction with Zymo DNA Clean & Concentrator columns between steps. Another part of the gDNA was used to create the Normalization library.

Besides using column purification in place of phenol-chloroform extraction, the RT and PCR reactions followed the protocol described in steps 67-105 from Akhtar, et al. (2013). Notable differences are that we used a cassette-specific primer with a UMI for reverse transcription in place of an oligo(dT) primer. The polymerase was also replaced with SuperScriptIV and Phusion High-Fidelity DNA Polymerase.

Primers used for amplifying sequencing libraries are recorded in Table SX. The number of PCR cycles was determined by PCR gradient and observation of the expected bands on gels Figure SX.

Expression, Normalization, and Mapping Libraries were given different indexes and sequenced together using Novogene Pre-made Library Sequencing service.

Data Analysis

Because the TRIP data pipeline does not account for splice isoforms, an in-house pipeline was developed with similar alignment and clustering approaches. Mapped locations for each cassette barcode were filtered similarly, as each must be supported by at least 5 reads with identical alignments, and no more than 10% of reads may align to a secondary location. This ensures that mapping reads are Transcript reads that showed mismatches between the variable exon and its splice barcode were thrown out.

The full pipeline is visualized in Figure S1.3, and utilizes the following packages (Bowtie2, BedTools, FgBio, Cutadapt, Starcode). Following this filtering process, mapping locations were shown to be high-confidence because there was no disagreement in aligned locations for the same barcode present in two replicates.

FIMO was used for checking the presence of RBPs in each construct. PPT and branch point sequences were determined by similarity to human consensus sequence from Kadri, et al.[11]

Supplementary Data

Gene	RefSeq#	Exon	5' Intron Sequence	Exon	3' Intron Sequence
WDR35	NM_001006657.2	11	ATTGACGATTATTTGTATATAAGCCACTTTCATTAGGTTGAGTCATGTCTTTTTTG AGTTAACTGAATTTTTCTGAATTTCTCCTGTCATTTATATATTGCTTCAAAG	GAGGAGAACGAGATGGAGACATT TGGTGCAACG	GTAACAAATGAAATT AGCTC
COLQ	NM_005677.4	5	CAGCCAAGCAACTGAGCCACATGGCCATGCCCAGTCATCAGCTCTACTAAGGC TTCTGTTTCCAGAACATCTCTTCCAAAGCAGCCTCTCCTTACCTGTCTTCTCCAT CCATAG	GGGGAGCTTGGCCGACCAGGAA GGAAG	GTATGGTCTGCGCT GTTCTG
ACADVL	NM_000018.4	3	CGCGGCTCTCGCCTGTTCTCCCCTAGACACAGCGGAAGTCCCTTCCCTGAACT TGCTAACCGTCTCTTTTCCCAG	CTGGCTCTGGACAAGTCAGATTC CCACCCCTCTGACGCTCTGACCC GGAAAAAACCGGCCAAGGCG	GTAGGTAGCCCCGAG GCCAG
MYH11	NM_001040113.2	42	TTATTAAATCGAAGCAAGTCTTTACAAATAAGAATATACTTAACCGGGCTCCTTC TTGTTGCTGATGGTAACCTCCGGGGATTCTCTCTCTGTTTCAG	AGGGCCGCCCCCACAGGAAACTT CGCAGTGATGCACCAG	GTATCATGCCTTCAG CTTCG
ASAH1	NM_177924.5	4	AATCTTGACAGCATTTCATTATGTCTTTTAGTAAGCATTTTATTTATTTTTCAG	CTAAAGGTTATAGTGAATTCTCTG AAGAATATGATAAATACATTCGTG CCAAGTGGAAAAATTATGCAGGT GGTGGATGAAAAATTG	GTAAGAGATATTATT TCTCA

Table S1.1. Cassette variable sequences and sources

Name	Sequence
Citrine Exon 1	ATGGTGTCCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTCAGCGGCGAGGGCGA GGGCGATGCCACCTACGGCAAACTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCTTCGGCTACGGCCTGAT GTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGC AACTACAAGACCCGCGCCGAAGTGAAGTTCGAGGGCGACACCCTCGTGAACCGCATCGAGCTAAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCAC AAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAAGTGAACTTCAAGATCCGCCACAACATCGAG
SMN2 Intron 6	GTAAGTAATCACTCAGCATCTTTTCCTGACAATTTTTTTGTAGTTATGTGACTTTGTTTTGTAAATTTATAAAATACTACTTGCTTCTCTCTTTATATTACTAAAAAATAAAA ATAAAAAAATACAACTGTCTGAGGCTTAAATTACTCTCAACTTAATTTCTGATCATATTTTGTTGAATAAAATAAGTAAAATGTCTTGTGAAACAAAATGCTTTTTAACATC CATATAAAGCTATCTATATATAGCTATCTATATCTA
SMN2 Intron 7	AAAGTGAATCTTACTTTTGTAAAACTTTATGGTTTGTGGAAAACAAATGTTTTTGAACATTTAAAAAGTTCAGATGTTAGAAAGTTGAAAGGTTAATGTAAAACAATCAAT ATTAAAGAATTTTGATGCCAAAACTATTAGATAAAAGGTTAATCTACATCCCTACTAGAATTCTCATACTTAACTGGTTGGTTGTGTGGAAGAAACATACTTTCACAATAA AGAGCTTTAGGATATGATGCCATTTTATATCACTAGTAGGCAGACCAGCAGACTTTTTTTTATTGTGATATGGGATAACCTAGGCATACTGCACTGTACACTCTGACAT ATGAAGTGCTCTAGTCAAGTTTAACTGGTGTCCACAGAGGACATGGTTTAACTGGAATTCGTCAAGCCTCTGGTTCTAATTTCTCATTTGCAG
Citrine Exon 2	GACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCCGCCCTG AGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAG

Table S1.2. Cassette constant sequences

Data	Source	File Name
DHS DNase-seq	Stamatoyannopoulos, 2025	ENCFF988TCO.bed
LaminB1 DamID	Van Steensel, 2018	4DNFIV776O7C.bed
SON Ab2 TSA-seq	Belmont, 2019	4DNFIVZSO9RI.bw
4X-AP3 DamID	Van Steensel, 2020	4DNFIGM62KB2
TADs	Springer et al., 2026[37]	DataS6
H3K36me3 ChIP-seq	Stamatoyannopoulos, 2015	ENCFF215NKF.bigwig
H3K79me2 ChIP-seq	Bernstein, 2011	ENCFF334HSS.bigwig
H3K4me1 ChIP-seq	Farnham, 2011	ENCFF783QIW.bigwig
H3K27me3 ChIP-seq	Farnham, 2011	ENCFF470HOG.bigwig
H3K27ac ChIP-seq	Bernstein, 2011	ENCFF465GBD.bigwig
H3K9me3 ChIP-seq	Bernstein, 2011	ENCFF330EOT.bigwig

Table S1.3. Sources of DNA and chromosome feature data
All experiments cited are for K562 cultured cells and aligned to GRCh38 reference.

Primer Name	Use	Sequence
Inverse_fwd_outer	Nested gDNA PCR 1	GCATTGACAAGCACGCCTCACG
Inverse_rev_outer	Nested gDNA PCR 1	TTTATTAGGAAAGGACAGTGGGAGTGGC
Inverse_fwd_inner	Nested gDNA PCR 2 + PE1	CTTTCCCTACACGACGCTCTTCCGATCTCATGCGTCAATTTTACGCAGAC
Normalization_fwd	Normalization PCR + PE1	CCTACACGACGCTCTTCCGATCTCCGACCACTACCAGCAGAACACCCCCATC
PE2_UMI	Reverse transcription primer, gDNA PCR 2 reverse primer, Normalization reverse primer + PE2	TTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCAACAGATGGCTGGCAAC
Variable R1 PrimerN7	cDNA PCR 1 + PE1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNBNKBCAGAAGAACGGCATCAAAGTGAA
Variable R1 PrimerN10	cDNA PCR 1 + PE1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNBNKBCAGAAGAACGGCATCAAAGTGAA
Variable R1 PrimerN20	cDNA PCR 1 + PE1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNNNBNKBCAGAAGAACGGCATCAAAGTGAA

Table S1.4. Primers used for cDNA, mapping, and normalization library amplification.
NEBNext primers i500 and i700 series used for final adapter and paired end index additions onto PE1+PE2 sequences.

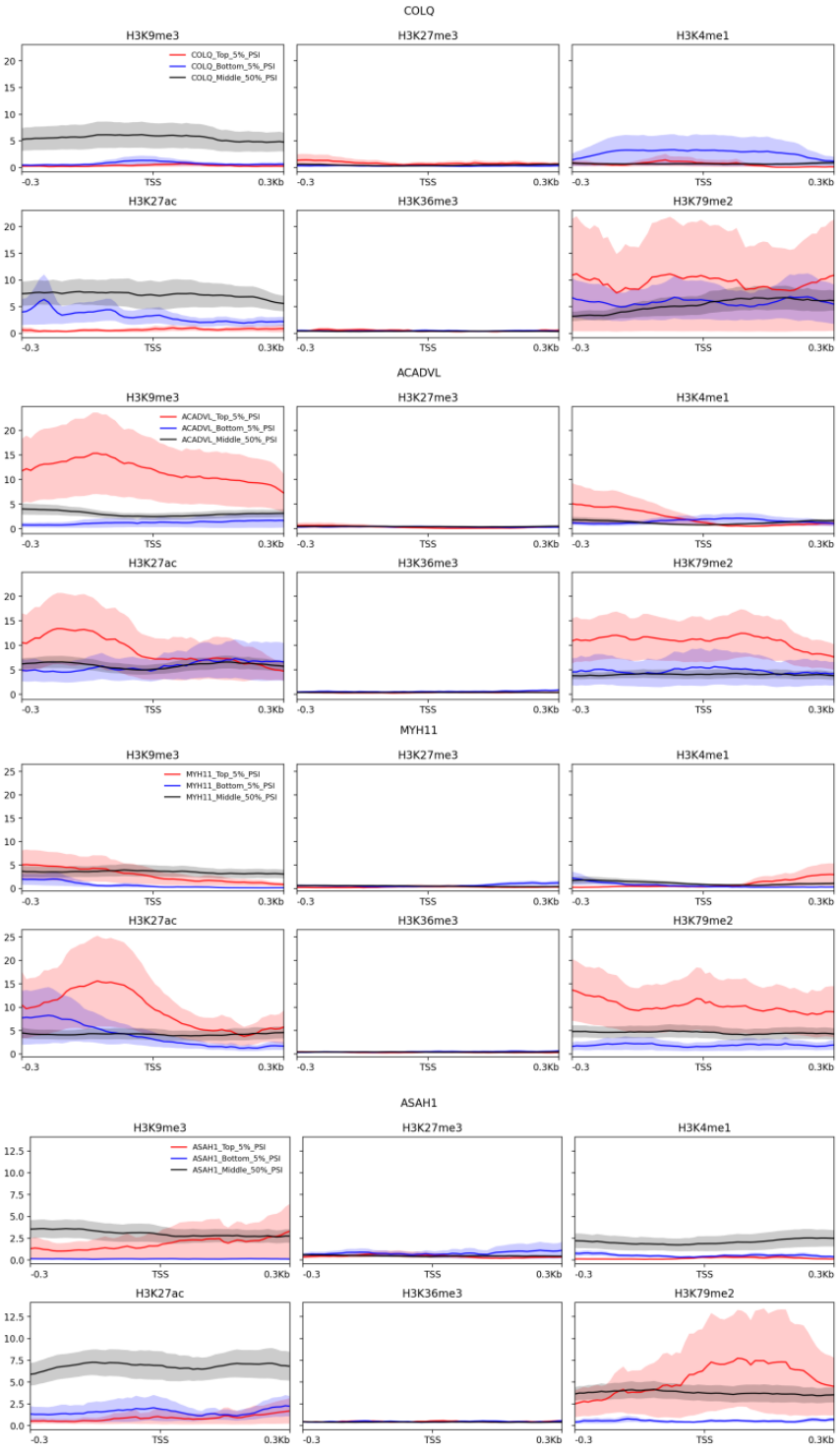


Figure S1.1. Deeptools ComputeMatrix profile plots for splicing-associated chromatin signatures

Enrichment of six SACS around top 5th percentile, bottom 5th percentile, and middle 50 percentile (second and third quartiles) for all cassette types are shown. TSS indicates place of insertion relative to mapped chromatin modification signal, not including ITR and promoter

upstream of actual transcription start site nor the length of the reporter gene downstream. Includes the ASAH1 plot from Figure 1.3.

Figure S1.2. Comparison of normalized expression for all cassette types against LAD, NAD, and SPAD domains.

Figure S1.3. Reporter spliced exon and introns are largely defined by canonical splice sequences

Variable sequences for each of the five cassette constructs are shown and putative canonical splice features are highlighted. Bolded sequences are exons surrounded by their endogenous introns. Canonical AG splice acceptor and GT splice donor dinucleotides are highlighted in yellow. Speculated polypyrimidine tracts adjacent to the splice acceptor are highlighted in blue. Speculated branch point sequences following the human conserved NNYTRAY pattern are highlighted in dark gray[11]. Additional RBP binding motifs recognized by FIMO are highlighted in light gray. WDR35's lack of a continuous PPT adjacent to the upstream intron's 3' end may be the reason for its consistent exclusion from nearly all transcripts.

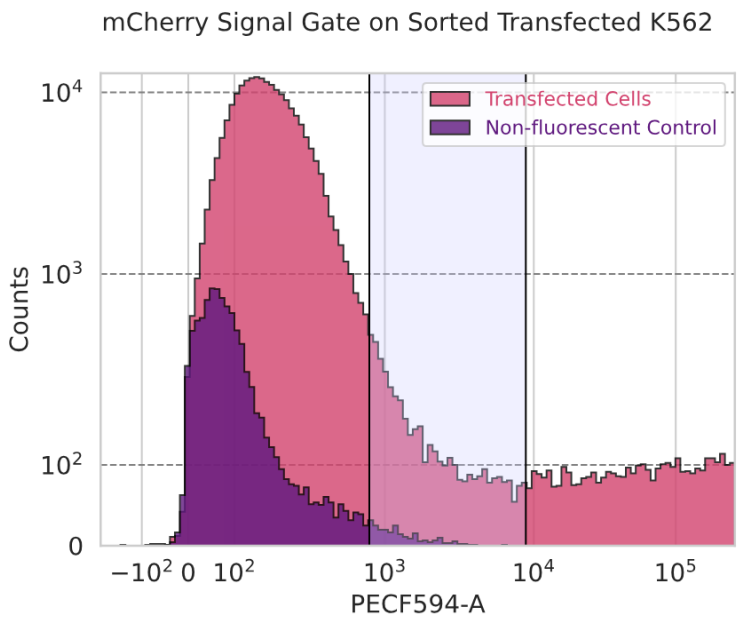


Figure S1.4. FACS gate selects cells with median piggyBac transfection
Cells were sorted with a BD FACSymphony S6 between 24-48h after transfection to create 1,000 cell aliquots with median mCherry signal shown by the gate, indicating successful transfection of pb transposase + mCherry vector. Fluorescence from citrine within splice cassette plasmids was checked but not used for sorting. Gate position was approximated based on fluorescence gating from Akhtar, et al.[38]

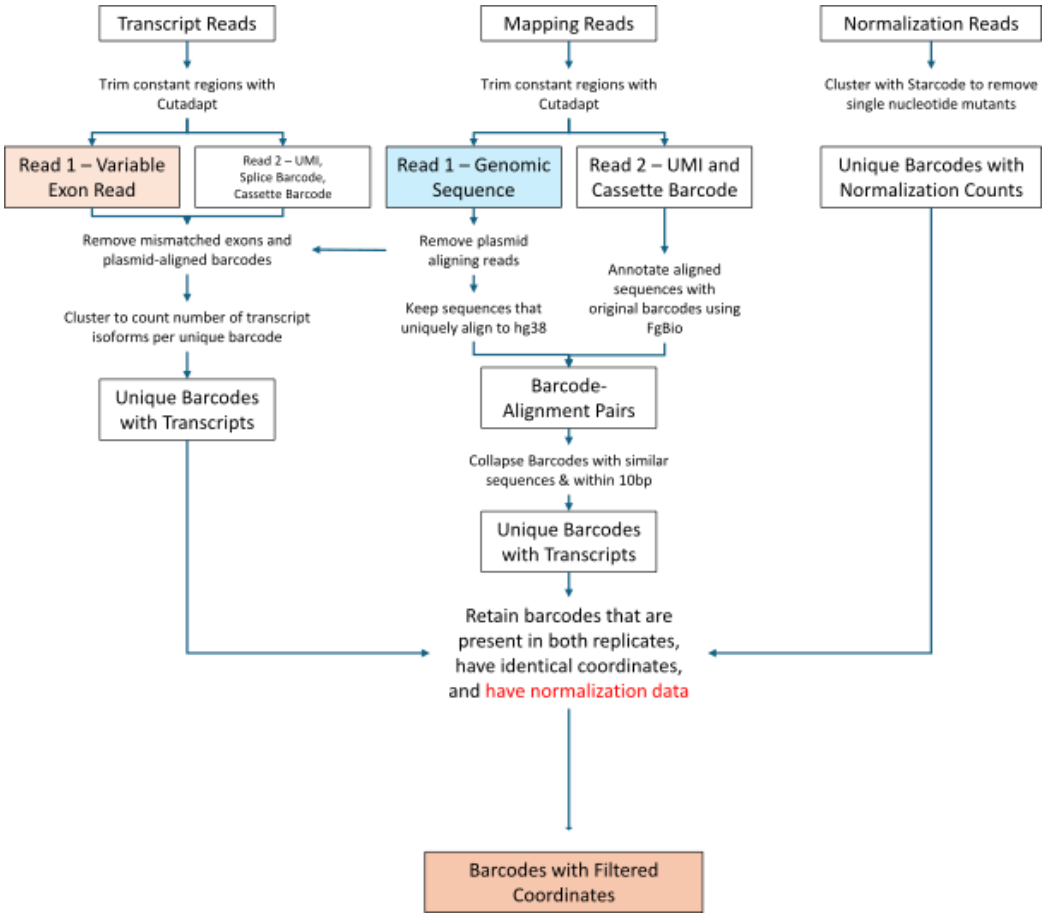


Figure S1.5. The custom TRIP-Splicing pipeline workflow mimics the TRIP pipeline
Each sequencing library for one sample is input as a separate FASTQ file and output into a final dataframe containing alignments, transcript counts, calculated PSI, normalization counts, and normalized expression values for each cassette barcode. Normalized expression for each set of barcodes from the same splicing reporter construct is calculated as $\log_2((1 + \text{transcripts}_{BC} / \Sigma \text{transcripts}_{\text{all PSI BC}})/(1 + \text{normalize}_{BC} / \Sigma \text{normalize}_{\text{all PSI BC}}))$. Cassettes with no normalization reads or no mapping reads are removed from any normalized expression calculations and analysis. Code is available upon request.

Self-Reporting Genomic Insert MPRA Integrates Positional Information into Transcripts with Truncated Polyadenylation Signal

Summary

Massively Parallel Reporter Assays that investigate transcriptional behavior of transgene reporters in the complex environment of the genome face a challenge in identifying where each reporter inserts within the genome after cell transfection. Whether the insertion is targeted (e.g. via CRISPR editing) or random (e.g. viral or transposon mediated), matching transcripts to locations is typically done by sequencing the genomic DNA, which requires additional library preparation and sequencing on top of the cDNA library. In order to simplify this process, we introduce a self-reporter assay that uses a reporter gene with a truncated polyadenylation signal which defers RNA transcription from termination at the cassette's 3' end and incorporates neighboring genomic sequences into the 3' end of each reporter transcript. This method eliminates the need for collection of genomic DNA for mapping information, reducing workflow steps and sequencing costs.

Introduction

There are hundreds of thousands of known enhancers and silencers in the human genome, which each may up or down-regulate nearby genes by recruiting transcription factors or repressors respectively to a promoter. These *cis*-regulatory elements (CREs) have been shown

via chromatin contact assays like Hi-C to bend in 3D space to make contact with promoters up to hundreds of thousands of bases away along the same chromosome[39–42].

However, most CREs have only been studied in their original contexts, where each gene only interacts with a limited number of enhancers or other elements. Given limited endogenous contexts of most unique genes, parallel reporter assays may be used to engineer a wide variety of gene combinations to test[34,35,43–47].

Some existing MPRAs aim to characterize specific CRE-promoter interactions by placing both elements into plasmids, where high-throughput construction of reporters is easily achieved with pooled DNA-synthesis and cloning. It has been demonstrated that individual promoter-CRE interactions correlate well between a plasmid environment and the genome. Still, studying interactions of promoters within the genome provides more direct characterization of both complex sequence elements and epigenetic elements.

In order to characterize the effect of various genomic domains on a particular gene, reporter cassettes may be inserted into the genome. In some cases, particular regions of interest may be targeted for insertion with DNA-editing methods such as CRISPR[48–53], but a wider variety of regions genome-wide may also be desirable to study, especially for the sake of training machine learning models to predict the expression of genes from novel positions. This approach demands a high throughput method for inserting reporters as well as retrieving expression and location information from each insert.

Here, we aimed to create a reporter cassette that could widely integrate across the genome in order to investigate the *cis*-regulatory impact of position effects on gene activity such as transcription, splice efficiency, and polyadenylation. This method, Genomic Reporters

Integrated in Parallel (GRIP), uses a reporter gene designed with a 3' UTR containing a cassette-specific barcode and ends with a truncated polyadenylation signal (PAS).

Transcription termination, 3' cleavage of the immature transcript, and polyadenylation are dictated by the 3' untranslated region (UTR) including a core AATAAA sequence surrounded by upstream and downstream sequences[54]. During transcription, these RNA elements recruit several protein factors that stimulate cleavage and polyadenylation. In this assay, it is demonstrated that a truncated PAS that lacks a controlled sequenced downstream of the core hexamer is still sufficient for 3' cleavage, so that cleavage and polyadenylation occur downstream of the cassette and the transcribed genomic sequence downstream of the integration point gets retained in transcripts. These transcripts may then be sequenced mapped against the genome to identify cassette insertion positions.

In contrast to existing integrated reporter assays such as TRIP[27], this approach theoretically does not require additional extraction and sequencing of genomic DNA to identify the locations of reporters integrated randomly via transposition. Targeted methods such as CRISPR editing or Prime editing are also used in genomic reporter assays, but the throughput is still limited by genomic sequencing or unique inserts for each location in order to trace transcripts to each known location. However, one downside of the GRIP approach is that integrations that do not express transcripts cannot be mapped without additional gDNA sequencing. This additional step reduces some of the workflow benefits of GRIP in cases where it is desirable to detect repressive domains or features.

Here we demonstrate how blunt-ended, double-stranded DNA cassettes may be integrated into the genome via the Non-Homologous End-Joining (NHEJ) pathway, a ubiquitous and highly active DNA-repair pathway in cells[55–57]. We also show how some integration may be

targeted by inducing double stranded breaks with CRISPR-cas9, as established by Geisinger, et al[52]. Finally, the design of truncated PAS sequences is shown to modulate the distance downstream that mRNAs are cleaved and polyadenylated at, thus including a genomic sequence adjacent to the point of insertion into transcripts. Using APARENT2, a residual neural network model that predicts 3'-cleavage and polyadenylation from sequence, we also demonstrate that synthetic PAS designs can be designed that further control the PolyA position downstream of the cassette sequence[45].

Ultimately, while this self-reporting method is able to produce transcripts that accurately self-report their position information in mRNA at verified locations within the genome, the efficiency of the reporter expression is currently insufficient for meaningful analysis of cis-regulatory elements. Delayed 3'cleavage inhibits transcription of the same reporter gene itself, which prevents characterizing activity of the gene promoter and local context. The GRIP method may be useful in the future if this inhibitory effect may be reduced or separated from activity of the actual gene of interest.

Results

Blunt-end Double Stranded DNA Cassettes Integrate into Genome

First, we investigated whether cassettes could insert into the genome via NHEJ, particularly at targeted sites where double stranded breaks are induced with CRISPR-cas9. Reporter cassettes were designed in a plasmid containing either a minP or hPGK sequence promoting expression of the hygromycin resistance gene followed by a full BGH polyadenylation sequence. The cassettes were produced by PCR using primers with 5' phosphorylation modification to allow blunt-end ligation into the genome. The reverse primer

also adds a 10nt random barcode into the 3' UTR 70 bases upstream. The final DNA cassette begins immediately at the 5' end of the promoter and ends at the core hexamer of the BGH sequence. Cassettes of each promoter type were transfected into K562 cells with a library of Px459 CRISPR-cas9 vectors targeting 4 random locations as well as the two safe-harbor sites AAVS1 and HPRT1, chosen for their demonstrated efficient CRISPR editing[58]. After insertion, reporters were allowed to express for 4 days before harvesting and sequencing mRNA. The "GRIP-tag" sequence between the AATAAA and the first instance of a PolyA tail (defined as ten consecutive adenines) was identified for each unique combination of barcodes and tag sequence (**Figure 2.1**).

Long GRIP-Tags Identify NHEJ Inserts At CRISPR Targets and Across Genome

The lengths of sequence tags used for mapping transcripts varied, including stochastic variation between cassettes that shared a barcode but yielded GRIP-tags of different lengths. For mapping, transcripts with a length of 10 or less were filtered out. The remainder were aligned against human genome reference GRCh38. We found that for each unique barcode, the majority of tag lengths were distributed between 10 and 20 bases, and that the proportion of tags that align uniquely increases with length (**Figure 2.2A**). This trend was consistent across different promoters. Of the six sites targeted by gRNAs for cleaving, the two safe harbor sites AAVS1 and HPRT1 as well as two intergenic locations on chromosome 1 demonstrated insertion and expression, identified by both alignment of GRIP-tags (**Figure 2.2B**) as well as PCR around the regions of interest (**Figure 2.3**). No tags aligned to the other two targeted regions on chromosome 6, which was verified with PCR and sanger sequencing to ensure that there were not simply inserts that did not express transcripts. Using Inference of CRISPR Editing (ICE), it was also seen that the efficiency of cleavage at all sites was low in no-cassette experiments[59]. In

an attempt to increase editing efficiency, AAVS1 was selected to have a paired gRNA to cut in two places 200 bases apart to encourage cassette insertion[52]. In this case, qPCR revealed no significant difference in insertion with one gRNA versus paired gRNAs (**Figure 2.3**).

Transcription of Integrated Reporters is Low

To count the number of transcripts that were captured from each individual integrated reporter, we took each unique cassette barcode that had a unique alignment and clustered them with all unique transcripts, including those which had a GRIP-tag that did not meet alignment requirements previously. This provided us with counts of barcodes with identical barcodes and similar genomic sequences, presumed to be transcripts that originated from the same insert. The distribution of numbers of transcripts per barcode is visualized for each promoter type in **Figure 2.2C**. Notably, inserts did not express very much on average, with both promoter types producing a median of 1 transcript per cassette. Because the transcript libraries were sequenced very deeply with a mean of 15,000 reads per unique transcript, and the cDNA libraries were only PCRed between 19 and 23 cycles, we interpreted the lack of transcripts as evidence that the integrated cassettes either had low transcriptional activity, or that their transcripts were very unstable.

APARENT2 Designed PAS Sequences Delay Polyadenylation and Decrease Transcription

It was demonstrated in **Figure 2.2A** that a truncated PAS is capable of producing transcripts carrying a GRIP-tag by delaying 3' cleavage and polyadenylation. Although the length is somewhat stochastic between 10bp and 20bp long and only one transcript per cassette must align uniquely to identify the position of each insert, there are still many cassettes that are

identified by a unique barcode that do not align to the genome. In order to improve the efficiency of GRIP-tag transcript reporting, we used APARENT2 to design ten synthetic PAS sequences that were predicted to have even more distal polyadenylation positions given the constraints of having any possible genomic sequence downstream of the core hexamer, and only the 70 bases upstream of the hexamer could be manipulated.

Cassettes with several of these designs were initially created and transfected into K562 cells as before, and integration into the AAVS1 locus was confirmed with PCR. However, these experiments failed to produce transcripts with which to measure the location of 3' polyadenylation. After this, we turned to testing the synthetic PAS constructs episomally rather than integrating them into the genome.

By constructing plasmids where the 3' UTR and truncated PAS designs were followed by a 124bp stretch of genomic sequence copied from four of the gRNA target sites, we were able to produce transcripts for each synthetic PAS design and measure the length of the GRIP-tags in four different contexts (**Figure 2.4**). The counts of transcripts for these were still low, indicating low transcriptional activity despite expression from plasmids. The GRIP-tag lengths were combined from the 10 synthetic designs to produce the histograms which visualize how the lengths do increase from the previously observed 10-20bp to distributions around 10-60bp long, and a small number of transcripts with even longer lengths. Notably, 60% of transcripts were mapped to the positive control consisting of a plasmid expressing with an untruncated BGH PAS, despite the control plasmid being transfected in equal proportion to every other individual plasmid. These transcripts also had a uniform length with no variable cleavage and polyadenylation. This high expression of the positive control supports the conclusion that truncating the PAS results in decreased expression.

Library Sequencing of Reporter Transcripts is Inefficient

One challenge observed in truncated-PAS reporters is the low amount of transcription per cassette. When sequencing the cDNA libraries, another challenge is presented in that a small number of reporters re-integrate into the CRISPR plasmids and express at a much higher level than the actually integrated reporters. Additionally, some cassettes appear to concatenate or insert into each other, as seen by GRIP-tags which align within the cassettes sequence itself. These reporters cannot be mapped to the genome and do not represent true genomic expression, but also express at a much higher rate (10,000 to 1,000,000 transcripts per integration compared to the median of 1 transcript for true integrations). This also occurred even when Hygromycin selection was removed from experiments to prevent selecting for cells with highly expressing events such as these. Finally, a small number of integrations that produced an aberrant number of transcripts had GRIP-tags that did not align to either the human genome nor the synthetic construct sequences. Together, these events made sequencing of the reporter assay severely inefficient and overly expensive.

Discussion

We set out to design and implement a novel reporter assay in which transcripts inserted into the genome with NHEJ may retain neighboring genomic sequences in their 3' ends, allowing mapping of the transcript location from the mRNA sequence alone. Double stranded DNA cassettes featuring 5' phosphorylated ends and a truncated PAS demonstrated that they were indeed capable of inserting in both a random and targeted manner as well as embedding genomic sequences for mapping into their 3' ends, though with varying efficiency. The truncated PAS induces delayed 3' cleavage and polyadenylation and enables longer GRIP-tags, but is also

responsible for greatly reducing the number of transcripts in our samples. Greater reporter mapping efficiency is thus counterproductive to the purpose of measuring the effects of genomic integration on reporter transcription, and this problem must be overcome to enable this approach for efficient simultaneous reporting of transcription and position.

Additionally, targeted NHEJ-mediated insertion with the use of CRISPR-cas9 plasmids was shown to be viable but not perfectly efficient, as a small library of six gRNA designs did not display consistent cutting and insertion into more than four locations. While the efficiency of DSBs may be improved independently, many other reporter integrations such as PiTCH utilize short homology arms on the cassette to increase targeted insertion efficiency, which may be scaled up with pooled oligo synthesis to produce cassettes with thousands of different homology arm designs. However, due to this method's approach of relying on the 10-20nt immediately downstream of the cassette's 3' end for mapping insertion location, then the homology arms could obfuscate the true position of cassettes if they insert into a location via the non-homologous pathway. These limitations introduce a challenge in increasing efficiency of targeted insertion with the GRIP method, though mostly random insertion is still a viable approach for studying genome-wide position effects.

The poor stability of mRNA with a truncated polyadenylation sequence that cleaves and polyadenylates pre-mRNA downstream of itself precludes it from being used for the actual reporter of transcription. For a potential future improvement on the GRIP assay, we propose a design pairing a location reporter with a separately expressed transcriptional reporter that are sufficiently insulated from each other on one large transposon (**Figure 2.5**). It has been demonstrated with the CallingCard mPRA[60] method that omitting a functional PAS and terminator from an integrated reporter gene entirely may result in transcription proceeding for

hundreds of bases until it reaches an endogenous or cryptic PAS, producing a long transcript with genomic sequence similar to GRIP-tags. A location reporter gene with this design could be activated from an inducible, powerful promoter to express transcripts for the sole purpose of transcribing mapping reporters separately from the regular transcriptional reporter. In the future, these experiments may provide the groundwork for reporter assays that are capable of simultaneously reporting transcription information and reporter integration information in cDNA alone.

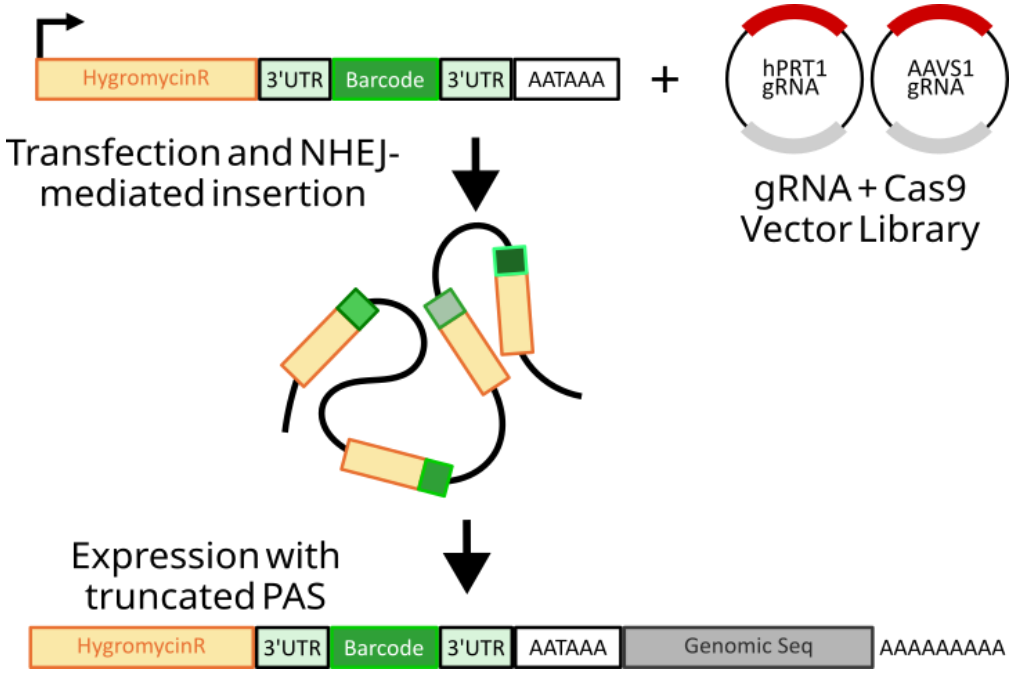


Figure 2.1. Overview of the self-reporting method

Double stranded DNA cassettes are transfected into cells together with CRISPR vector plasmids expressing designed gRNAs and cas9. Cassettes that insert into double stranded DNA breaks via NHEJ produce transcripts that include genomic sequence from the destination locus downstream of the PAS AATAAA hexamer. All transcripts sharing a barcode presumably transcribed from the same inserted cassette.

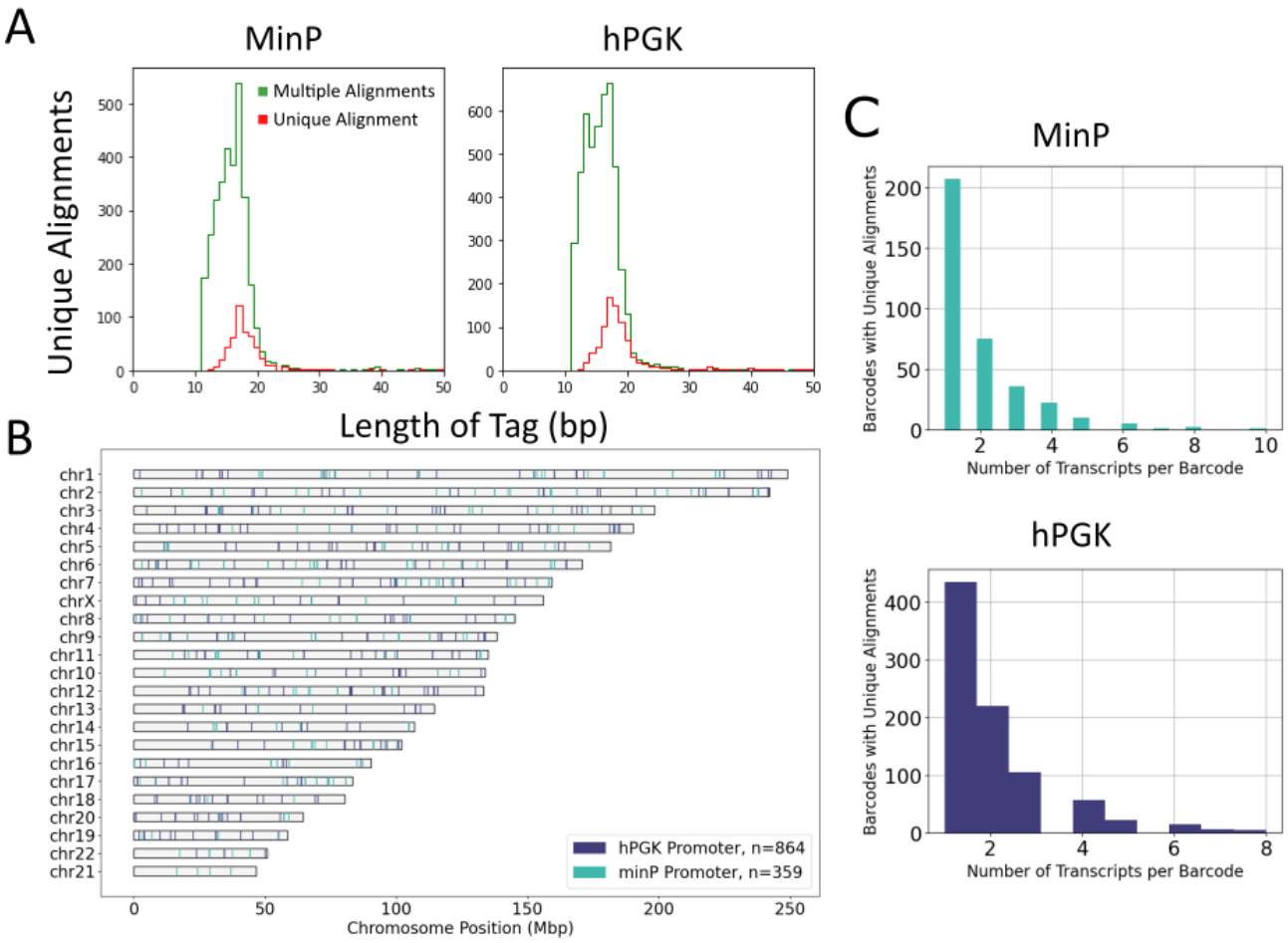


Figure 2.2. GRIP-tags are aligned to the genome and transcripts are grouped by aligned barcodes

(A) The sequence at the end of the BGH 3' UTR reporter transcripts located between the AATAAA hexamer and the polyA tail is aligned to the human genome with Bowtie2. Sequence tags may align uniquely or to multiple loci, and longer tags tend to align uniquely more frequently. After filtering out tags with a minimum length of 10, the majority of transcripts still do not carry a tag that uniquely aligns. **(B)** Alignments demonstrate cassette insertion into loci across the genome, in addition to those which were targeted by CRISPR. Of the regions targeted by gRNAs listed in Table S2.3, HMGA1 and SCIRT regions showed no insertions at the expected loci. hPGK1 and minP promoter cassettes were transfected in two separate experiments, and alignments overlapped here. **(C)** For each aligned barcode, the number of unique transcripts are counted to estimate relative expression per location. The median number of transcripts per insert for either promoter reporters = 1.

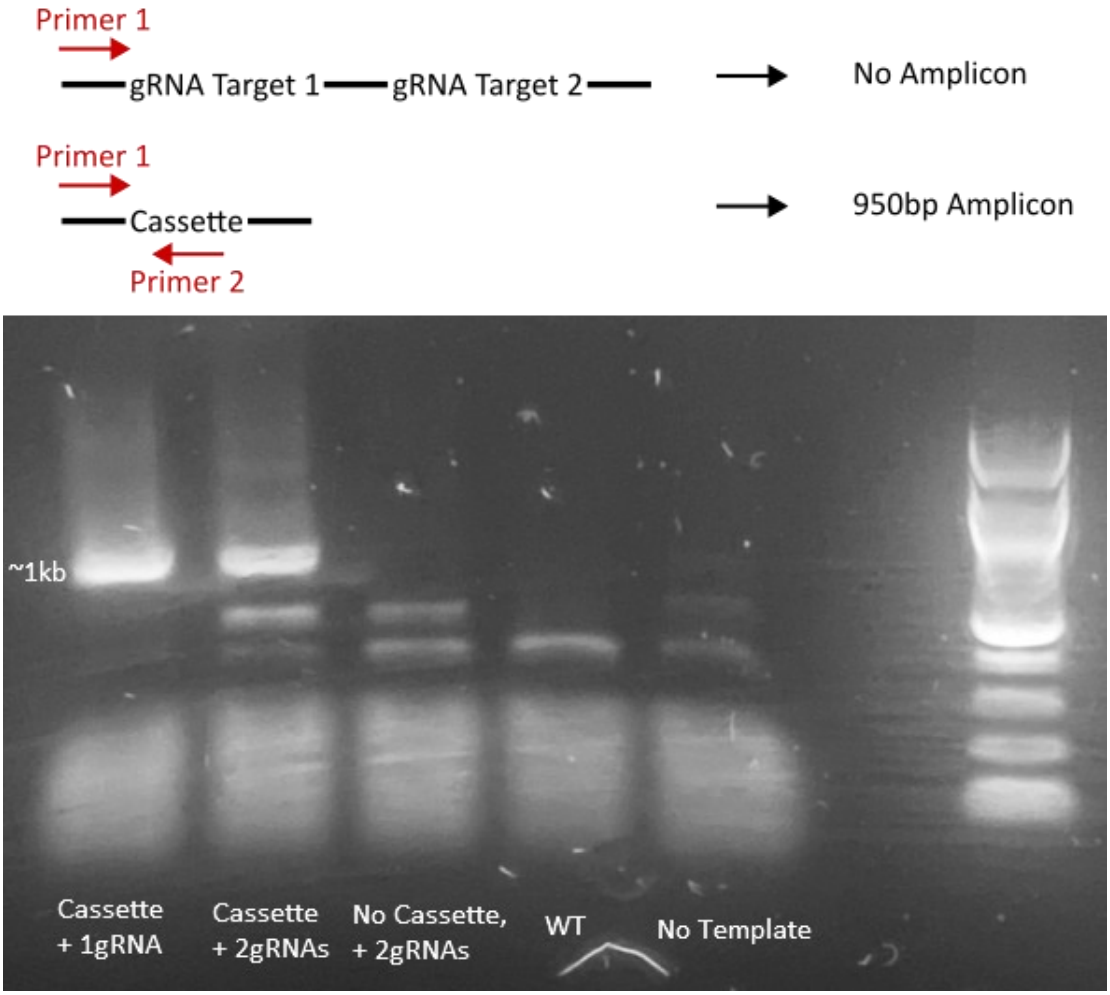


Figure 2.3 Insertion of cassettes into the AAVS1 locus was verified by PCR of gDNA
A primer located in the cassette five hundred bases from the end of the sequence and another locus-specific primer five hundred bases upstream amplify a 1kb amplicon when the cassette is successfully inserted in the gRNA target site 1. Insertion of the cassettes was successful with either a single gRNA (gRNA 1) or paired gRNAs (1 and 2) targeting the region. This process was done for HPRT1 and the remaining insertion sites were simply tested via PCR and sanger sequencing of the amplicons.

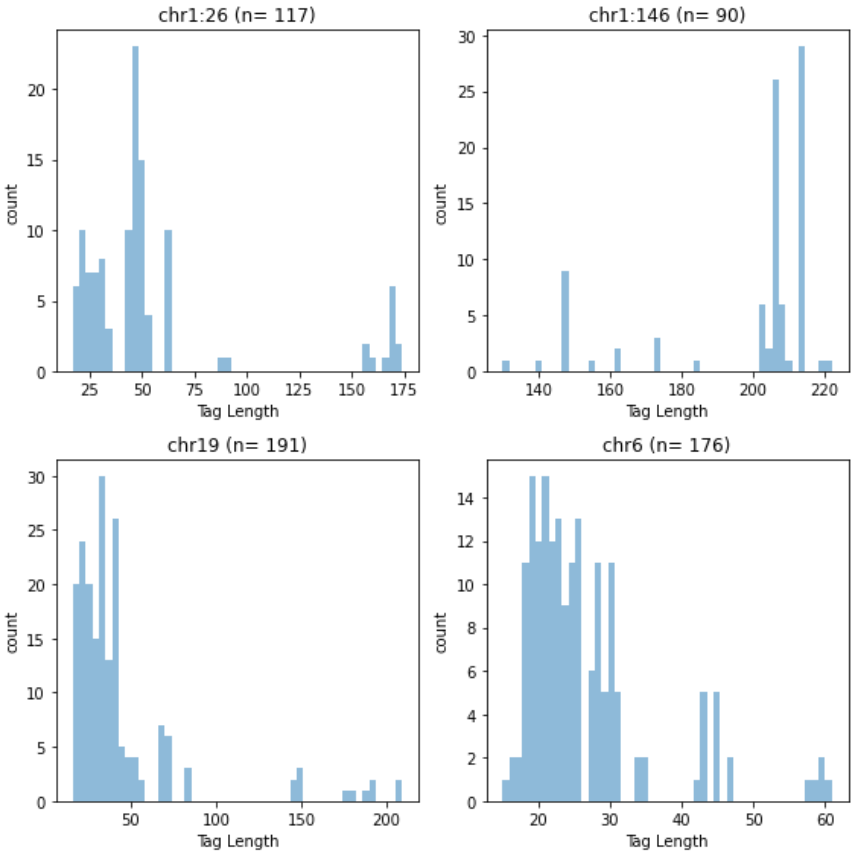


Figure 2.4. Lengths of sequences between the truncated PAS and the PolyA tail in transcripts expressed from plasmid Synthetic PAS constructs in K562 cells
Synthetic PAS sequences were tested by expressing the cassette in a plasmid with APARENT2 designed 3' UTRs followed by fifty bases of genomic sequence located immediately downstream. Shown are the sum of sequenced tags for 10 synthetic designs for each of four downstream designs, which were chosen to match previous genomic insertion sites. Lengths of the sequence following the PAS hexamer in the cassette sequence ranged from 20 to 220, demonstrating longer average tags than with a truncated BGH PAS.

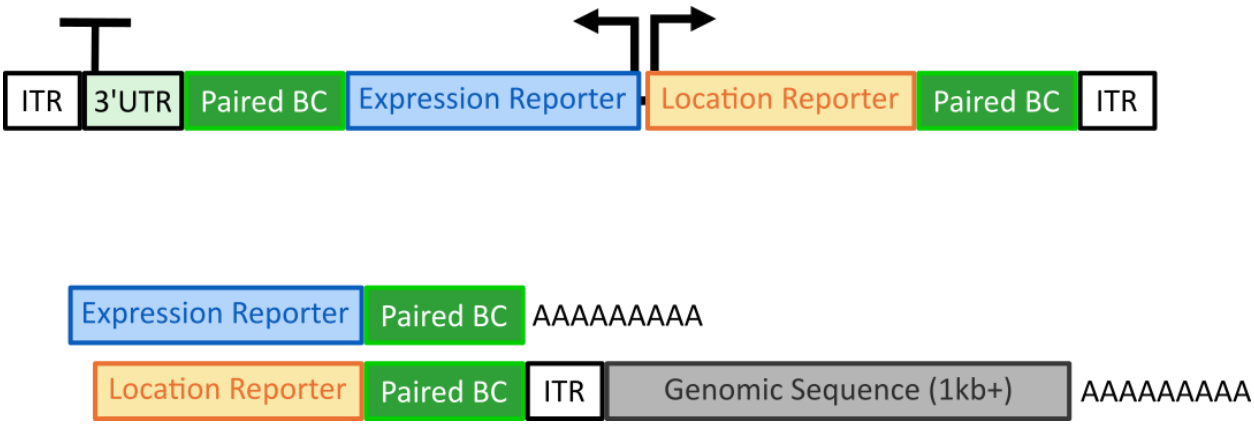


Figure 2.5. Proposed future design for self-reporting cassette.
The reporter lacking a terminator is separated from a reporter with full terminator, and both are flanked by ITRs for transposition. The location reporter gene transcribes beyond the ITR until it hits a cryptic or endogenous termination signal to produce self-mapping transcripts, while the expression reporter transcribes normally. Expression reporter locations may be determined by matching barcodes to the paired location reporter transcripts that uniquely align to the genome.

Methods

Cassette Construction

A plasmid with the CMV, MinP, or hPGK1 promoter with the citrine reporter gene was assembled with NEBuilder HiFi DNA Assembly mix and transformed into NEB 10-beta Electrocompetent E. coli with a BioRad Micropulser Electroporator. Cultures were spread on an ampicillin plate overnight and plasmids were purified using Qiagen Spin Miniprep Kits. IDT Ultramers were designed with phosphorylated 5' ends to prime and amplify cassettes that are capable of blunt-end ligation. The Ultramer for the 3' end of the reporter gene was designed to end in a PAS AATAAA hexamer, and included a random barcode consisting of five N bases located 70nt upstream of the 3' end. Following PCR amplification of the cassette plasmid template using the paired Ultramers, 1µL DpnI restriction enzyme was added to each 50µL of PCR reaction and incubated at 37°C for 30 minutes in order to digest methylated plasmid template. PCR reactions were then run on agarose gel electrophoresis, and cassettes of desired

length were cut out and purified from gel using Monarch DNA Gel Extraction Kit (NEB). Cassettes were further verified with sanger sequencing from GENEWIZ.

Synthetic PAS Plasmid Construction

Polyadenylation signal sequences were changed out for the complete BGH terminator in plasmids used as template for cassette amplification. In short, IDT ultramers containing each synthetic PAS design and 124 bases of the sequence from each chosen genomic region were designed and ordered. These ultramers were used as primers to amplify the entire plasmid along with plasmids to create a cassette fragment and two backbone fragments. All three fragments were re-assembled into one plasmid and verified using Plasmidsaurus plasmid sequencing. As a control, we also used a plasmid with the untruncated BGH PAS sequence.

sgRNA Design

AAVS1 and HPRT1 targeting guide RNA designs were generated from IDT. Other targeted locations were chosen at random and filtered with the same tool to select guides with low off-target scores. Oligos with guide sequences were cloned into Plasmid #62988 (Addgene, from Feng Zhang lab) that expresses chimeric gRNA plus a puromycin resistance marker and a human codon-optimized Cas9.

Cell Culture

HEK293 cells were transfected with Lipofectamine 3000 (Invitrogen) according to manufacturer protocol for 6-well plates, mixing 200ng double stranded cassette and 600ng gRNA in a 3:1 ratio of gRNA plasmid:cassette together with P3000 reagent in 200µL Opti-MEM (Thermo Scientific) per 300,000 cells. The ratio was taken from Geisinger, et. al (2016) while

maximizing the total amount of DNA that lipofectamine allows. HEK293 cells were reverse transfected by trypsinizing and washing with PBS before mixing suspended cells with the lipofectamine, P3000, and DNA mix and seeding into 2mL EMEM (VWR) without supplemental FBS nor antibiotics. All cultures were maintained at 37 °C, 5% CO_2 in a humidified incubator. In experiments where Hygromycin selection was used, Hygromycin B Gold (Fisher Scientific) was added to 400µg/mL 24 hours after transfection. Four days after transfection, cell fluorescence was measured. Cells were further cultured for RNAwas extracted from 2/3rds of cells using Monarch Total RNA Extraction and Isolation kit (NEB), and gDNA was extracted separately using Blood & Cell Culture DNA Mini Kit (Qiagen) from the remaining 1/3rd cells.

K562 cells were instead transfected according to the Thermofisher Neon Transfection protocol with 10µL tips. For each experiment, 200,000 cells were washed with Gibco PBS (without Ca^{2+} and Mg^{2+}) and passed through a 40µM cell strainer before re-suspending in 10µL Resuspension Buffer R and 150ng of purified cassette and equal ratios of cas9/gRNA plasmids mixed together to a total of 0.5µg. After transfection, cells were added to RPMI 1640 Medium (Thermo Scientific) with supplemental 10% FBS but without antibiotics. For experiments where Hygromycin selection was used, Hygromycin was added to 200µg/mL. Otherwise, culturing followed the same steps as with HEK293.

Verification of CRISPR Activity

In order to determine whether the cas9 plasmids were making double-stranded breaks at the target locations, qPCR on gDNA extracted from cells transfected with individual guide-cas9 plasmids alone was performed with primers designed to flank each targeted cut site and positioned approximately 250bp away from the guide target to produce a 500bp amplicon.

Amplicons were visualized on a 1% agarose gel and extracted with Monarch DNA Gel Extraction Kits (NEB) and sanger sequenced with GENEWIZ. Sanger trace files were then used with the ICE (Inference of CRISPR Edits) tool from Synthego to observe nucleotide mutations adjacent to the targeted cleavage site that is indicated of NHEJ-mediated repair.

GRIP-tag alignment

mRNA was harvested from cells with Monarch Magnetic mRNA Isolation Kit (NEB) and reverse transcribed with a reporter-specific primer that adds a UMI to each transcript. Libraries were read with paired-end Illumina sequencing and clustered to remove PCR duplicates with Starcode. All sequences upstream of the core hexamer and downstream of a polyA were trimmed using Cutadapt, and tags 10nt or longer were aligned to GRCh38 reference using Bowtie2 end-to-end alignment. Final insert locations were determined for each barcode when at least one of its transcripts aligns uniquely. Transcript libraries were sequenced with an Illumina High Output Nextseq Kit to an estimated average depth of 15,000 reads per transcript to ensure deep sequencing.

Works Cited

1. Ke, S. & Chasin, L. A. Context-dependent splicing regulation: Exon definition, co-occurring motif pairs and tissue specificity. *RNA Biology* **8**, 384–388 (2011).
2. Kornblihtt, A. R., De La Mata, M., Fededa, J. P., Muñoz, M. J. & Nogués, G. Multiple links between transcription and splicing. *RNA* **10**, 1489–1498 (2004).
3. Hertel, K. J. Combinatorial Control of Exon Recognition. *Journal of Biological Chemistry* **283**, 1211–1215 (2008).
4. Rogalska, M. E., Vivori, C. & Valcárcel, J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat Rev Genet* **24**, 251–269 (2023).
5. Liu, W. & Zhang, X. Single-cell alternative splicing analysis reveals dominance of single transcript variant. *Genomics* **112**, 2418–2425 (2020).
6. Koplik, S. E. *et al.* Massively parallel assay of human splice variants reveals cis-regulatory drivers of disease-associated and cell type-specific splicing regulation.
7. Pankivskyi, S. *et al.* U2AF2 controls alternative splicing in speckle-proximal regions in an RS domain-dependent manner. *Nucleic Acids Research* **54**, gkag143 (2026).
8. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* **163**, 698–711 (2015).
9. Murray, J. I., Voelker, R. B., Henscheid, K. L., Warf, M. B. & Berglund, J. A. Identification of motifs that function in the splicing of non-canonical introns. *Genome Biol* **9**, R97 (2008).
10. Moehle, E. A., Braberg, H., Krogan, N. J. & Guthrie, C. Adventures in time and space: Splicing efficiency and RNA polymerase II elongation rate. *RNA Biology* **11**, 313–319 (2014).
11. Kadri, N. K., Mapel, X. M. & Pausch, H. The intronic branch point sequence is under strong evolutionary constraint in the bovine and human genome. *Commun Biol* **4**, 1206 (2021).
12. Horn, T., Gosliga, A., Li, C., Enculescu, M. & Legewie, S. Position-dependent effects of RNA-binding proteins in the context of co-transcriptional splicing. *npj Syst Biol Appl* **9**, 1 (2023).
13. Agirre, E., Oldfield, A. J., Bellora, N., Segelle, A. & Luco, R. F. Splicing-associated chromatin signatures: a combinatorial and position-dependent role for histone marks in splicing definition. *Nat Commun* **12**, 682 (2021).
14. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in Alternative Pre-mRNA Splicing. *Cell* **144**, 16–26 (2011).
15. Uriostegui-Arcos, M., Mick, S. T., Shi, Z., Rahman, R. & Fiszbein, A. Splicing activates transcription from weak promoters upstream of alternative exons. *Nat Commun* **14**, 3435 (2023).
16. Kim, S., Kim, H., Fong, N., Erickson, B. & Bentley, D. L. Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13564–13569 (2011).
17. Barutcu, A. R. *et al.* Systematic mapping of nuclear domain-associated transcripts reveals speckles and lamina as hubs of functionally distinct retained introns. *Molecular Cell* **82**, 1035-1052.e9 (2022).

18. Bhat, P. *et al.* Genome organization around nuclear speckles drives mRNA splicing efficiency. *Nature* **629**, 1165–1173 (2024).
19. Chaturvedi, P. & Belmont, A. S. Nuclear speckle biology: At the cross-roads of discovery and functional analysis. *Current Opinion in Cell Biology* **91**, 102438 (2024).
20. Marie, P. *et al.* Gene-to-gene coordinated regulation of transcription and alternative splicing by 3D chromatin remodeling upon NF-κB activation. *Nucleic Acids Research* **52**, 1527–1543 (2024).
21. Wen, L. *et al.* RNA localization to nuclear speckles follows splicing logic. *Nucleic Acids Research* **54**, gkag174 (2026).
22. Zhang, L. *et al.* TSA-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes. *Genome Res.* **31**, 251–264 (2021).
23. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat Genet* **52**, 8–16 (2020).
24. Briand, N. & Collas, P. Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol* **21**, 85 (2020).
25. Lalanne, J.-B. *et al.* Multiplex profiling of developmental cis-regulatory elements with quantitative single-cell expression reporters. *Nat Methods* **21**, 983–993 (2024).
26. Leemans, C. *et al.* Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**, 852-864.e14 (2019).
27. Akhtar, W. *et al.* Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell* **154**, 914–927 (2013).
28. Alharbi, A. B., Schmitz, U., Bailey, C. G. & Rasko, J. E. J. CTCF as a regulator of alternative splicing: new tricks for an old player. *Nucleic Acids Research* **49**, 7825–7838 (2021).
29. Vester, K. *et al.* Recruitment of a splicing factor to the nuclear lamina for its inactivation. *Commun Biol* **5**, 736 (2022).
30. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**, eaaw1668 (2019).
31. Li, M. A. *et al.* The *piggyBac* Transposon Displays Local and Distant Reintegration Preferences and Can Cause Mutations at Noncanonical Integration Sites. *Molecular and Cellular Biology* **33**, 1317–1330 (2013).
32. Mossine, V. V., Waters, J. K., Hannink, M. & Mawhinney, T. P. piggyBac Transposon plus Insulators Overcome Epigenetic Silencing to Provide for Stable Signaling Pathway Reporter Cell Lines. *PLoS ONE* **8**, e85494 (2013).
33. Gupta, S. *et al.* The nucleolar granular component mediates genome-nucleolus interactions and establishes their repressive chromatin states. *Molecular Cell* **85**, 2165-2175.e6 (2025).
34. Hong, C. K. Y., Ramu, A., Zhao, S. & Cohen, B. A. Effect of genomic and cellular environments on gene expression noise. *Genome Biol* **25**, 137 (2024).
35. Hong, C. K. & Cohen, B. A. Genomic environments scale the activities of diverse core promoters.
36. Shanas, S., Schwartz, A. M., Nesher, N., Melnik, A. & Mikl, M. Promoter identity shapes splicing outcomes and fidelity. *Nat Commun* https://doi.org/10.1038/s41467-026-76370-1 (2026) doi:10.1038/s41467-026-76370-1.

37. Springer, S. M. *et al.* Identity rather than 3D position informs splicing of rare introns in the human genome. *iScience* **29**, 114532 (2026).
38. Akhtar, W. *et al.* Using TRIP for genome-wide position effect analysis in cultured cells. *Nat Protoc* **9**, 1255–1281 (2014).
39. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat Genet* **52**, 8–16 (2020).
40. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
41. Engel, K. L., Mackiewicz, M., Hardigan, A. A., Myers, R. M. & Savic, D. Decoding transcriptional enhancers: Evolving from annotation to functional interpretation. *Seminars in Cell & Developmental Biology* **57**, 40–50 (2016).
42. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* **22**, 108 (2021).
43. Akhtar, W. *et al.* Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell* **154**, 914–927 (2013).
44. Lalanne, J.-B. *et al.* Multiplex profiling of developmental cis-regulatory elements with quantitative single-cell expression reporters. *Nat Methods* **21**, 983–993 (2024).
45. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol* **23**, 232 (2022).
46. Jha, A., Gazzara, M. R. & Barash, Y. Integrative deep models for alternative splicing. *Bioinformatics* **33**, i274–i282 (2017).
47. Bulger, M. & Groudine, M. TRAPping enhancer function. *Nat Genet* **32**, 555–556 (2002).
48. Sakuma, T., Nakade, S., Sakane, Y., Suzuki, K.-I. T. & Yamamoto, T. MMEJ-assisted gene knock-in using TALENs and CRISPR-Cas9 with the PITCh systems. *Nat Protoc* **11**, 118–133 (2016).
49. Suzuki, K. *et al.* In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* **540**, 144–149 (2016).
50. Maresca, M., Lin, V. G., Guo, N. & Yang, Y. Obligate Ligation-Gated Recombination (ObLiGaRe): Custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome Res.* **23**, 539–546 (2013).
51. Kanca, O. *et al.* An efficient CRISPR-based strategy to insert small and large fragments of DNA using short homology arms. *eLife* **8**, e51539 (2019).
52. Geisinger, J. M., Turan, S., Hernandez, S., Spector, L. P. & Calos, M. P. *In vivo* blunt-end cloning through CRISPR/Cas9-facilitated non-homologous end-joining. *Nucleic Acids Res* **44**, e76–e76 (2016).
53. Nakade, S. *et al.* Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. *Nat Commun* **5**, 5560 (2014).
54. Millevoi, S. & Vagner, S. Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. *Nucleic Acids Research* **38**, 2757–2774 (2010).
55. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair* **7**, 1765–1771 (2008).
56. Sawatsubashi, S., Joko, Y., Fukumoto, S., Matsumoto, T. & Sugano, S. S. Development of versatile non-homologous end joining-based knock-in module for genome editing. *Sci Rep* **8**, 593 (2018).

57. van Overbeek, M. *et al.* DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell* **63**, 633–646 (2016).
58. Aznauryan, E. *et al.* Discovery and validation of novel human genomic safe harbor sites for gene and cell therapies. Preprint at https://doi.org/10.1101/2021.03.04.433856 (2021).
59. Hsiau, T. *et al.* Inference of CRISPR Edits from Sanger Trace Data.
60. Moudgil, A. *et al.* Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* **182**, 992-1008.e21 (2020).

Supplementary Data

3'UTR Name	3'UTR Sequence
apa18	AAAGGAATAGAACCAGAGCGTCAAGGAAGGAAAAAGCAAGACAGGACAAAGATCAGGAGGGGGGGGTTGTAATAAA
apa34	AAAGGAATAGGCCGGAAAGGGAGAAAGAAGAAGAAGGACCGACGTCAAAAGGCAACCGGAGGGGGGGGCTAATAAA
apa30	AAAGGAATAGGACAGGCAGATCAAGAAGAAGAAAGCAGAAAGAGAGGGGGGAGAGGTAGGTAGGGGTTGTAATAAA
apa91	AAAGGAATAGACGGACAGTCTCAAGAGAAAGAAAGGACCGGAAGGGAGGGAGGAAGGGCCCCGGGGGGGGAATAAA
apa04	AAAGGAATAGTCCGGAGAGCGAGAAAGAAGAAAAACCAAAGACGGGGGAAGAGAAGGAGGCCAGGGTGTTAATAAA
apa90	AAAGGAATAGGGTCCAGAGGCCGAAGAAGAAAAAGCGAAGAAGGGGGGGGTGGGCCCGGAGGGGGGGAGGAATAAA
apa11	AAAGGAATAGATTTTTAGCGAAGAAGGAGAAAAAAGAAGGACAAATAGGAAGGAAGGGGGGAGTGTGTATAATAAA
apa06	AAAGGAATAGGGTGGGCTAGAAGCAAGAAGAAGAAAGAAGACAGGGAAGGGGAAGGGAGGGTAGGGTTGTAATAAA
apa00	AAAGGAATAGGACAGGGACTCCAAGAGAAAAAAAAGACGAGAAGGAGAGGGGGGCCGGCGGGAGGGGAGGAATAAA
apa23	AAAGGAATAGACCGGTAGTATCGGACGAAGAAAAAGAAGGACGGCGGCGAGGGGAAGCCCCCCCCCTGTTAATAAA
BGH	CCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAA

Table S2.1. PolyA Signals tested for transcript termination distances

BGH is the 3' end of the whole BGH PAS used for the baseline reporter. In each synthetic PAS construct, the design from APARENT2 replaced this BGH sequence.

Sequence Origin	Sequence Downstream PAS
Aavs1_fwd (chr19)	CGGCGCTGCACCACGTGATGTCCTCTGAGCGGATCCTCCCCGTGTCTGGGTCCTCTCCGGGCATCTCTCCTCCCTCACCCAACCCCA TGCCGTCTTCACTCGCTGGGTTCCCTTTTCCTTCTC
chr1_146_fwd	GGTAGGTATTATTATCATAACCCCCTTGCATATGAAGAAGCCCAAATTTAGTAAATACTAAATAACTTTCCTTAATTTATACCCCCTGGAA AATGGGGGAAGAGCTTTGAATACCGTCAGTCA
chr1_26_fwd	CAAAGGCATCCTGGTCTAGCAGAAAAAGCCCAGGCTTTGGAGTAGCAGATGTTGTGTGGTAGCTCATGCCTGTAATCCCAGCACTCTG GGAGGCCGAGGTAGGTGGATCACCTGAGGTCAGGA
chr6_34_fwd	AGGTGGAAAAGACGAACGTTTCTTGGCCTCACTGGAAAGGGCCTCGGGCCTTCATTTAGTTCGTGGGGGGTTGGCTGAGGGGATTAA AGGGTCCCCAGCAGCAAGAGGTGGGGGGAGGCACCA
BGH	ATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGG GAAGACAATAGCAGGCATGCTGGGGACGACGGTGCTC

Table S2.2. Genomic sequences used for testing truncated PAS activity in plasmids

For each 3' UTR PAS sequence from Table S2.1., all 10 of these sequences were placed downstream in a sequence library for a total of 110 sequence combinations. Human sequences are taken from Human assembly GRCh38.

Target Name	Hg38 Coordinate	gRNA Strand	gRNA Sequence	PAM	Forward qPCR Primer	Reverse qPCR Primer
AAVS1a	Chr19:55,116,359	-	ACGTGGTGCAGCGCCGAGA	AGG	GGTGATGATGCAGGCCTACAAGAAG	GGTTCCGTCTTCCTCCACTCCCTC
AAVS1b	Chr19:55,116,359	+	GGGGCCACTAGGGACAGGAT	TGG	N/A	N/A
HPRT1	ChrX:134,473,411	+	ATTATGCTGAGGATTTGGAA	AGG	TACACGTGTGAACCAACCCG	GTAAGGCCCTCCTCTTTTATTT
HMGA1	Chr6:34,238,685	+	CTCTGATGTTCAGAAACAGG	TGG	N/A	N/A
Chr1 Intergenic	Chr1:146,936,193	-	TCATACTGTCCTCACGAGGT	AGG	N/A	N/A
Chr1 Intergenic 2	Chr1:26,126,791	+	GGGCAACTTGATATTACCAA	AGG	N/A	N/A
SCIRT	Chr6:43,947,697	+	TCCTAGATAGATGATCACAG	TGG	N/A	N/A

Table S2.3. CRISPR-cas9 gRNA designs and verification primers

Random Guide	Top Alignment	gRNA Strand	gRNA Sequence	PAM
OLS1	Chr10:44,446,619	+	ATCTTTGACAACAGCACAA	TGG
OLS2	Chr2:87,190,263	-	AAGAGCAACAAGGCAATC	AGG
OLS4	Chr14:53,940,061	+	CATGAAGAAAACAGTACAG	AGG
OLS8	Chr19:32,180,805	+	CACCCCGACCCCACCT	TGG
OLS10	Chr13:110,493,641	-	TCAACTAAATGAAAGATCA	CAG
OLS12	Chr15:62,864,980	+	ATATTCAACATTTTTAAAG	AAG

Table S2.4. Additional gRNA designs were used for experiments without sequencing

Passively sensing SARS-CoV-2 RNA in King County Public Transit Buses

Summary

Here, I will recount this work done in collaboration with Jason Hoffman and published in the *Science of the Total Environment* journal in 2022, in which SARS-CoV-2 RNA was collected from King County Metro bus air filters and quantified with qPCR in order to estimate viral presence by time and location in a low-cost approach that does not require specialized equipment. From August 2020 to March 2021, we placed non-invasive filters into the HVAC systems of public buses active in Seattle, Washington before collecting them every two weeks for phenol-based RNA extraction and qPCR quantification. It was found that samples taken from active buses showed significant viral amplification over controls that approximately matched county-wide counts of COVID-19 cases, indicating that this method is feasible for detection of viral particles on public transit. This method may be useful for future detection of specific COVID strains or other viral particles as well as extrapolating regional differences from each individual bus route. Due to the scope of the conclusions extending beyond my PhD work, that section from the published article has been excluded from this thesis.

Introduction

The global pandemic of COVID-19 has exceeded 50 million reported cases in the US and 270 million confirmed cases worldwide as of December 20th, 2021 (Centers for Disease Control and Prevention, 2020; World Health Organization and others, 2020). The virus causing COVID-

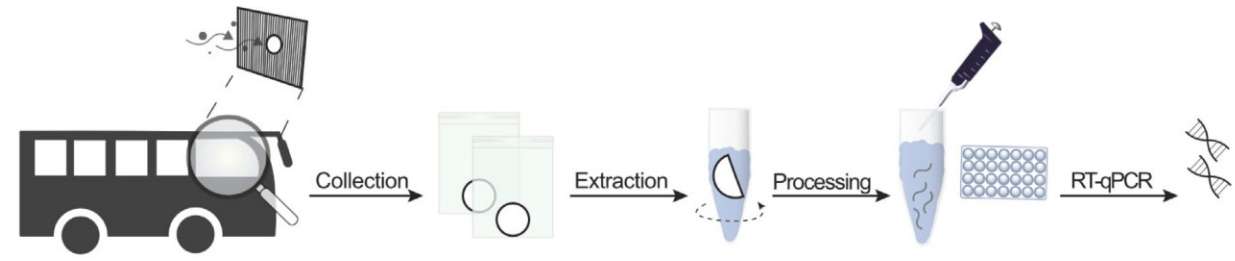
19, SARS-CoV-2, is primarily transmitted through airborne respiratory droplets via face-to-face contact (Buonanno et al., 2020; Zhang et al., 2020; Wiersinga et al., 2020) with asymptomatic or pre-symptomatic infected individuals. (Bai et al., 2020; Rothe et al., 2020; Yu et al., 2020; Hu et al., 2020a; Yu & Yang, 2020; Huff & Singh, 2020). Therefore, disease monitoring via viral presence testing is essential for managing potential outbreaks. Current disease monitoring is focused primarily on testing individual members of the population. However, frequent widespread testing across the entire population can be cost-prohibitive in many communities, even with pooled testing (Augenblick et al., 2020). While this resource intensive sampling strategy is useful for capturing the overall presence of a disease, alternative environmental sampling can serve as a warning sign of early-stage disease presence in a community prior to symptomatic patients testing positive (Daughton, 2020; World Health Organization, 2020).

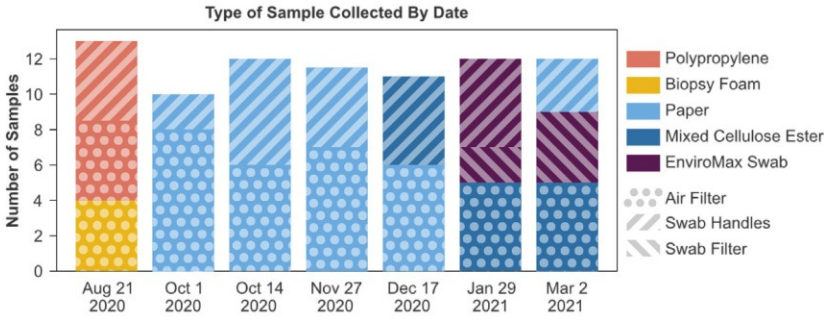
One example of passive viral sensing is testing for SARS-CoV-2 in community wastewater plants (Daughton, 2020; Wurtzer et al., 2020; Wu et al., 2020; Peccia et al., 2020; Medema et al., 2020). However, wastewater surveillance methods suffer from a fixed, coarse granularity since sampling happens far downstream from the individual source. Leveraging multiple environmental sampling techniques through additional infrastructural media, such as public transit, can make viral monitoring more robust. Wastewater sampling has been explored for commercial aircraft and cruise ships (Ahmed et al., 2020), but these approaches cannot be extended to public transport vehicles without wastewater management facilities. Public transit such as buses, light rails, and trains may be valuable targets for surveillance sampling, since they are linked to the population's geospatial mobility. The United Nations estimated >50% global population lives in urban centers (Neiderud, 2015), such as Seattle, where nearly 50% of urban commuters use public transit (S. D. of Transportation, 2019).

Viral particles expelled from the respiratory system of an infected individual can circulate through the air into Heating, Ventilation, and Air Conditioning (HVAC) systems, and have been detected in air filters in hospitals treating infected individuals (Kim et al., 2016; Ong et al., 2020; Guo et al., 2020; Chirico et al., 2020; Horve et al., 2021), suggesting a similar approach for public transit. The HVAC system in King County Metro buses, which involve an air intake, MERV-rated filter, and recirculation, pull air from the front or rear passenger-seating areas on buses and are designed to be running all-day long to provide fresh air for passengers. Therefore, they represent an opportunity for passive sensing if viral presence can be detected in the system. Prior work has examined risk of transmission for passengers on buses, trains, and airplanes at local, national, and international scale (Hu et al., 2020b; Zheng et al., 2020; Zhao et al., 2020; Browne et al., 2016; Hoehl et al., 2020; Shen et al., 2020; Luo et al., 2020; Bae et al., 2020; Yang et al., 2020; Kong et al., 2020; Kasper et al., n.d.; Harvey et al., n.d.; Silcott et al., 2020); however, prior work has not explored community monitoring on public transit. One potential reason for this is the cost and time associated with known sampling methods with adequate Limits of Detection (LOD) to sense the low number of copies of virus expected in filters without employing active systems of collection, such as environmental swabbing or vacuum-like Personal Environmental Monitor (PEM) equipment (Moreno et al., n.d.). Rapid and inexpensive RNA extraction methods have detected 10–20 copies/reaction, which may be above the viral copies recovered from passive HVAC systems in non-concentrated settings outside of hospitals (Panpradist et al., 2021a). Additionally, virus particles can remain viable for 7 days on porous surfaces, like air filters, and 3 days on non-porous surfaces, like metal hand-grips (Aboubakr et al., 2021; Matson et al., 2020). Therefore, air filters may accumulate and maintain virus over a longer time than swabbed surfaces, capturing data from more individuals with a single sample,

enabling pooled testing. Notably, viral presence does not imply infectivity, as a disinfected surface may still hold a dead virus with RNA that can be detected while the virus is no longer capable of infecting someone.

Here, we explore the feasibility of passive surveillance sampling in public buses by installing fabric sensors in vehicle air filtration systems. We demonstrate that sensitive methods of detection can be used to detect small virus copy numbers from samples collected from bus filters, using viral lysis, RNA extraction, and RNA detection via reverse transcription quantitative polymerase chain reaction (RT-qPCR) in a combination not proven in prior literature. In contrast to prior work studying building air filters (Horve et al., 2021; Rosario et al., 2018), we show that a novel combination of methods enables passive, scalable sampling while maintaining the potential for finer-grained community spread monitoring in localized areas via known bus routes. A requirement for scalability of this method is leveraging the already-operating HVAC systems in the bus for sensing while maintaining high analytical sensitivity and specificity for low concentration environments. Thus, we evaluated our in-house method in samples collected from actively circulating buses from August 2020 to March 2021 to demonstrate the detection of SARS-CoV-2 RNA in real-world environments, and we present herein an analysis of how this method may be related to citywide cases for future disease monitoring use cases.





Filter Type \ Bus Location	All Air Filter	Front Air Filter	Rear Air Filter	Swab - Rear Filter	Swab - Bus Handholds	Overall (% pos, CT mean +/- std)
MCE (Mixed Cellulose Ester)	18.2% (2/11) 37.3 +/- 0.04	N/A	18.2% (2/11) 37.3 +/- 0.04	N/A	0% (0/5) N/A	12/5% (2/16) 37.3 +/- .04
Paper	7.4% (2/27) 38.1 +/- 0.1	12.5% (2/16) 38.1 +/- 0.1	0% (0/11) N/A	N/A	62.5% (10/16) 35.0 +/- 1.8	30.0% (12/43) 35.5 +/- 2.0
EnviroMax Swab	N/A	N/A	N/A	33.3% (2/6) 36.3 +/- 1.1	20.0% (1/5) 32.4 +/- N/A	27.2% (3/11) 35.0 +/- 2.4
Polypropelene cloth	25.0% (1/4) 39.7 +/- N/A	50.0% (1/2) 39.7 +/- N/A	0% (0/2) N/A	N/A	40.0% (2/5) 38.6 +/- 1.1	33.3% (3/9) 39.0 +/- 1.0
Biopsy Foam	0% (0/3) N/A	0% (0/2) N/A	0% (0/1) N/A	N/A	N/A	0% (0/3) N/A
Overall (% pos, CT mean +/- std)	11.1% (5/45) 38.1 +/- 1.0	15.0% (3/20) 38.6 +/- 0.9	8.0% (2/25) 37.3 +/- 0.04	33.3% (2/6) 36.3 +/- 1.1	41.9% (13/31) 35.4 +/- 2.2	24.4% (20/82) 36.1 +/- 2.2

Figure 3.1. PCR assay methodology

Detection of the samples collected from the metro bus using our in-house extraction protocol. (**A**) Workflow for passive sensing SARS-CoV-2 RNA including sample collection, sample transfer from papers or swabs, RNA extraction, and RT-qPCR for detection. (**B**) Sampling occurred via two methods in different areas of the bus. We collected supplementary pre-filters after more than 7 days of being installed inside the HVAC systems of actively-used metro buses (blue). We also swabbed commonly-touched surfaces on the bus (red). (**C**) Sample types and collection methods used during the course of the study. (**D**) Positivity rate and average CT value breakdown by collection material and location. Sampling from both air filters as well as surfaces returned traces of SARS-CoV-2. Swabs from bus handholds made up the majority of SARS-CoV-2 detections with 42% positivity rate (13/31), while materials placed in air filters had the lowest positivity rate at 11% (5/45). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Results

Results on environmental samples from buses

Out of 82 samples (164 total with replicates) tested, 24% (20/82, 95% CI: 16–35) of samples tested positive for SARS-CoV-2 RNA (Fig. 1D), indicating viral presence but not necessarily infectivity. Samples collected by the swabbing method showed the highest positivity rate at 42% (13/31, 95% CI: 25–61) while filters had the lowest positivity rate at 11% (5/45, 95% CI: 4–24). 1 replicate amplified in all positive bus samples (Table A.3). This indicates that the viral particles may not be distributed evenly across the sampling material or that the sample methods may be sensing viral presence from different signal sources (i.e. riders who breathe may not touch the railing). Most positive samples had SARS-CoV-2 RNA near or below the LOD of the RT-qPCR assay (Fig. A.4) and thus were confirmed correct product sizes by fragment analysis (Fig. A.12).

Method comparison

We compared our TRIzol-based RNA extraction method with a more commonly used column-based RNA extraction from Qiagen. After dividing five filters in half and processing in parallel, we found a bus positivity rate of 60% (3/5) and 80% (4/5) (Table A.4) in TRIzol-based and column-based methods, respectively. Interestingly, our TRIzol-based method did not yield positive results from any samples collected by EnviroMax swabs, which were positive with the column-based method. We observed black particle residues in samples using EnviroMax swabs (Fig. A.13), which were filtered out by the column-based extraction. These residues ended up in the RT-qPCR reactions when EnviroMax samples were extracted by the TRIzol-based method,

which may have inhibited the RT-qPCR reaction. On the other hand, the column-based method displayed 0% positivity rate on all air filter samples, while the Trizol-based method detected 40% (2/5) positivity on the same air filters. We hypothesize that the debris broken from filters could interfere with the binding of SARS-CoV-2 RNA to the silica membrane in the columns or the SARS-CoV-2 RNA might remain trapped in the columns. These filters are made from mixed cellulose ester which have different surface properties and porous structures from those of polyurethane foam structures (EnviroMax swabs) which reportedly had a high release efficiency even with minimal agitation (Panpradist et al., 2014).

All air filters were installed on buses for more than 7 days, and thus can represent pooled samples of all riders for the prior 7 days (Fig. 2). One exception is that, for one sampling date on October 14, 2021, filters were installed and collected in one day. In one-day testing, 0 filters returned positive, indicating that one day may not be enough filter exposure time to build up a detectable viral load. However, a relatively high rate (60%) of swab samples returned positive, which may be attributable to lack of surface decontamination mid-day. Metro cleans buses nightly, and the morning sampling period for all 2-week samples occurred the morning following the decontamination, in between which no riders would have ridden the bus.

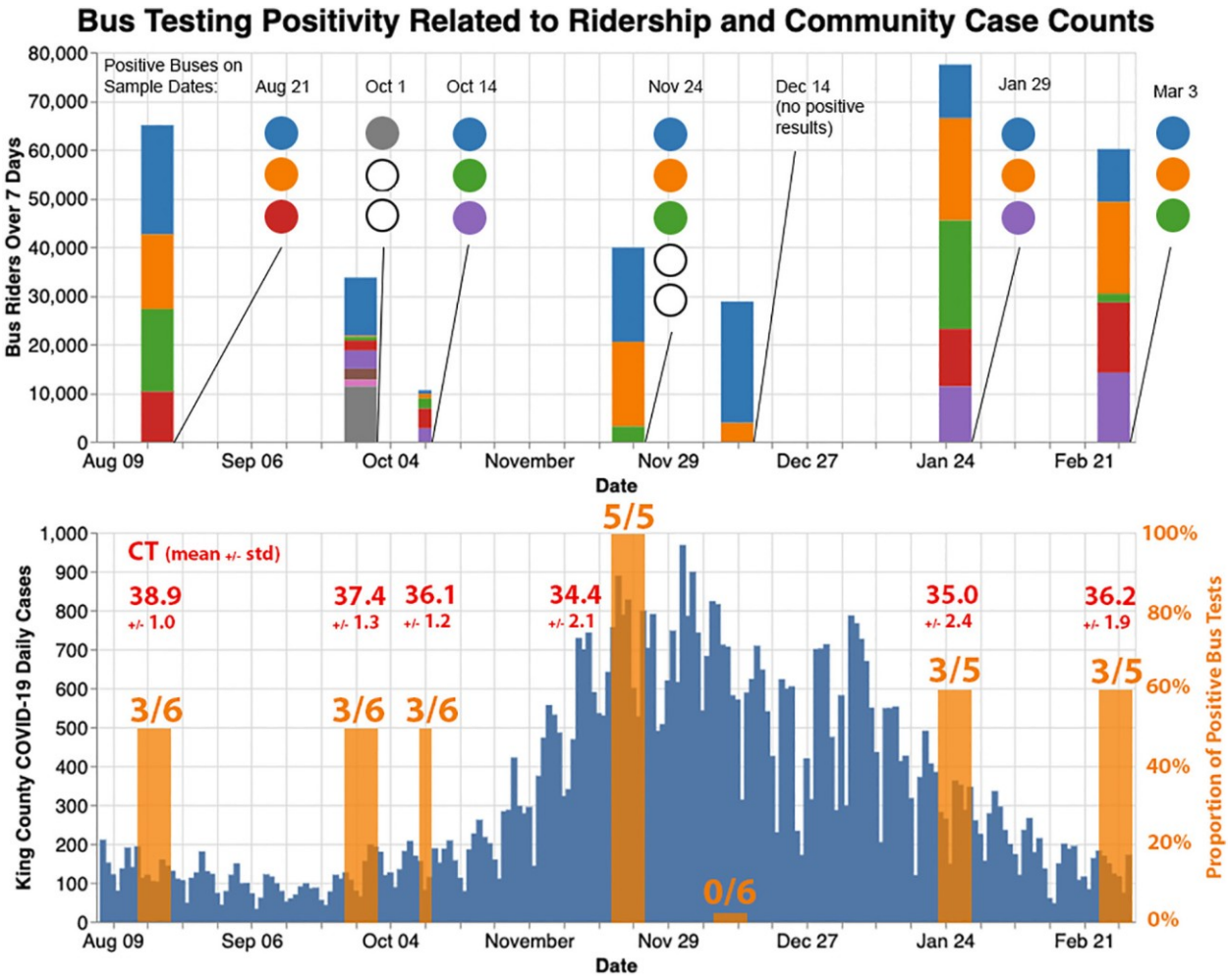


Figure 3.2. Quantification of viral RNA approximately correlated with regional reported cases

Top chart shows total riders per bus per 7-day period of filter installation before sampling. Each color denotes a unique bus that week. The color of associated circles denotes a positive result from that bus. An empty circle denotes a positive sample for a bus with 0 riders during that week. Bottom chart shows new cases of SARS-CoV-2 in King County (blue) superimposed with the proportion of buses sampled that week returning positive results (orange). CT values for positive results are listed by date (red), and showed a −0.687 Pearson's correlation when compared to King County individual testing positivity (Fig. A.7). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Results compared to population testing

Fig. 2 shows the bus testing results juxtaposed with the SARS-CoV-2 case counts and bus ridership counts for the 7 sampling periods. While the sample size is too low to demonstrate

positive correlation, we do see that a higher proportion of buses sampled return positive results when SARS-CoV-2 cases in King County were high, with the exception of December 14th, which showed no positive results. This was likely due to the fact that only 2/6 buses sampled had any ridership during that week, reducing exposure and decreasing likelihood of detection. October 1st and November 24th also returned positive results for buses which did not leave the station (zero riders), denoted by the empty circles in Fig. 2. This finding indicates that some results in this study may be a signal of continued viral RNA presence after more than 7 days or infected maintenance workers who entered a bus during the sample period.

Metro's ridership data was compared to bus positivity results to determine whether ridership was correlated with the positivity rate for buses, and a small Pearson's Correlation of 0.255 was observed. King County population testing data was also compared to average CT value of positive results, and a Pearson's Correlation of −0.687 was observed (Fig. A.7). The negative trend indicates that the strength of the detected signal on buses may increase as more people test positive in the city. The correlation observed may be reduced due to the small sample size and infection control measures taken by King County Metro. This included required rider masking and nightly surface disinfection (see Methods) throughout the entire sampling period, likely resulting in an increased number of false negative results due to a decreased number of viral particles escaping the mask of infected riders and reduced viral presence on hand rails. Ventilation and virus collection on filters may have been affected by driver and rider behavior, as well, as it is the driver's choice whether to activate the ventilation system (though most drivers do) and passengers may open the windows if they choose, which would modulate ventilation and likelihood of viral capture on filters. Additionally, the population of metro riders

from which the testing was sampled is not fully representative of the overall population from which individual testing was performed.

Control validation

In control experiments, extracting spike control from filters yielded 6/9 (66.7%) replicates and 7/9 (77.8%) replicates of spike controls directly in solution amplifying and displaying positive detection using our method, while only 1/18 (5.6%) of the replicates of negative controls returned positive with a high CT value of 37.3 (Fig. A.5). Based on a PCR standard curve from the same experiment (Fig. A.6), the extraction efficiency is approximately 33% (average of 134 out of 400 copies) for filter extraction controls and 115% (average of 459 out of 400 copies, which is within the error range for PCR extraction) for direct in solution extraction controls (Table A.2). This suggests that about one-third of viral material may be lost during the filter extraction step, but not much is lost during RNA extraction. This may have resulted in some false negatives in bus testing but demonstrates that positive detection results were likely a result of viral material collected from bus sampling.

Discussion

Here we show that passive infrastructure-mediated sensing of viral presence may be feasible. By leveraging a passive viral sensing method such as ours in parallel with other environmental and individual testing methods, epidemiologists could inexpensively monitor a community to identify locales of transmission and estimate case numbers within local regions. Our method may be more valuable when cases within the general community are low, leveraging the fabric sensors to passively pool respiratory droplet samples from riders temporally over the filter installation period and spatially over bus routes. This monitoring would be enabled by

sufficient resourcing to enable daily sampling and testing of a subset of bus routes to ensure adequate coverage, which was not possible with the small study team in this proof-of-concept study.

This study is limited by its sample size ($n = 39$ total buses). Sample filters were placed and recovered manually by the research team and metro collaborators, which could be scaled by larger research teams (Table A.5). False negatives in air filters may have been caused by the limited coverage of the sample media over the air vent and air currents diverting around the sample filter media, resulting in a lower positivity rate than swab samples. In addition, mask mandates were in effect for riders during the sample period, likely reducing the number of viral particles expelled into the air by breathing of infected riders landing on the filters. Finally, Metro's nightly cleaning process is designed to reduce the overall viral load in the bus, even if it did not completely remove the viral presence signal. Considering these effects, the small viral loads (Fig. A.4) of some samples are unsurprising, but may fall below the typical LOD for many commercial PCR kits, including our chosen Taqpath 1-step kit (C. for Disease Control, 2020). Alternative sensitive assays that amplify multiple regions in SARS-CoV-2 may be useful to detect these samples with low concentrations (Kline et al., 2021; U. Food, 2021). We also note that our method does not necessarily identify the risk to bus riders, but rather the presence of inactive SARS-CoV-2 RNA. Viral viability tests and more frequent sampling are needed to understand the risk to riders.

Future research into scalable, sensitive viral detection for environmental samples would enhance this approach. Studies evaluating filter placement, size, and material, incorporating further control experiments in simulated environments, could further validate the sensitivity of the method and optimal materials for sampling (Buonanno et al., 2020; Holmgren et al., 2010).

City-wide deployments enabled by scalable detection methods, such as rapid diagnostic lateral flow detection for on-site detection enabled by miniaturized amplification devices (Panpradist et al., 2021a; Panpradist et al., 2021b) or sequencers (Cardozo et al., 2021), could gather more data, enabling network analysis techniques to study probability of SARS-CoV-2 transmission on a neighborhood level. This method could be adapted and deployed to provide an early signal of community outbreak of SARS-CoV-2, or other viruses transmitted by respiratory droplets, when case counts are low in the population. This could provide a relatively inexpensive early warning system and ongoing monitoring insight into the local routes of viral transmission for current and future respiratory pathogens.

Methods

Sample collection from public buses

Between August 2020 and March 2021, environmental samples were collected from 15 actively deployed buses in the Seattle King County Metro fleet (Fig. 1A). Bus selection was narrowed down to the main bus depot that serviced the Downtown Seattle area, which has the highest ridership. Individual buses were selected to be sampled via a convenience sampling approach based on which buses could be made available at the depot on a regular basis between 7:00–9:00 AM for sample retrieval.

Detection of SARS-CoV-2 RNA in filters

Sample extraction for testing was performed within the same day of the sample collection from metro buses. Detection of SARS-CoV-2 RNA consisted of the following steps: viral

extraction and lysis, RNA isolation via phenol-chloroform isoamyl extraction, and RNA detection via RT-qPCR (Fig. 1A).

RNA extraction and isolation

Filters collected from buses were cut into 2-cm^2 pieces. Two pieces, considered sample replicates, were placed into microcentrifuge tubes containing 200 µL lysis buffer (50 mM EDTA pH 8.0, 250 mM Tris-HCl pH 8.0, 50 mM NaCl, 1% (*w*/*v*) SDS) (Miura et al., 2011). The tubes were placed on a foam tube rack attached to a vortexer and agitated for 15 min, at high speed, at room temperature. After vortexing, 600 µL TRIzol was added to each tube, pipette-mixed 10 times, and then the resultant 800 µL solution was transferred into a new tube. The solutions were incubated at room temperature for 5 min to allow complete dissociation of viral particles into the upper media and inactivation of any potentially remaining active virus in the solution. The RNA was then isolated from protein and DNA following the standard TRIzol phase separation procedure (Rio et al., 2010). Precipitation of RNA was carried out by adding 1 mL 200-proof ethanol and 1 µL RNA-grade glycogen (R0551, ThermoFisher) to each tube, followed by 1-min vortexing. Each tube was then incubated overnight at −20 ° C. Following overnight precipitation, the supernatant was discarded and residual ethanol was allowed to evaporate. The RNA pellet was washed following the standard TRIzol RNA wash procedure and subsequently re-suspended in 8 µL nuclease-free water.

Detection of SARS-CoV-2 RNA using RT-qPCR

Each 8 TRIzol isolation product was assayed with TaqPath 1-step RT-qPCR (A15299, ThermoFisher Scientific) in 20 µL reactions. We used probes from the CDC SARS-CoV-2 qPCR probe assay targeting two regions in the N gene, designated N1 and N2 (10,006,713, Integrated DNA Technologies), one for each sample replicate. To avoid cross-contamination, the reactions were loaded into non-adjacent wells in a 96-well plate on ice at a separate bench from where the RNA isolation step was performed. Wells also were covered with Parafilm between loading samples. RT-qPCR was carried out on a Quantstudio 3 (ThermoFisher) using the CDC-recommended protocol (C. for Disease Control, 2020).

Positivity determination

Positive results were determined by amplification before a specified PCR cycle threshold (CT). Raw RT-qPCR amplification data were loaded via a custom Python script (Supplementary Information) that determined CT based on a common threshold across all samples (50,000 RFU) which delineated between amplified samples and non-amplified samples in control experiments. For environmental samples from buses, a positive result for a bus was declared if one replicate from one of the sample methods from that bus had a CT <40. Positive and negative controls experiments were conducted by dripping AccuPlex enveloped RNA reference material (0505–0126, Seracare, Milford, MA) onto unused filter, and the full extraction and detection method was performed to confirm that positive results were a result of viral presence in the air filter.

Works Cited

1. Aboubakr H.A., Sharafeldin T.A., Goyal S.M. Stability of sars-cov-2 and other coronaviruses in the environment and on common touch surfaces and the influence of climatic conditions: a review. Transbound. Emerg. Dis. 2021;68(2):296–312. doi: 10.1111/tbed.13707.
2. Ahmed W., Bertsch P.M., Angel N., Bibby K., Bivins A., Dierens L., Edson J., Ehret J., Gyawali P., Hamilton K.A., Hosegood I., Hugenholtz P., Jiang G., Kitajima M., Sichani H.T., Shi J., Shimko K.M., Simpson S.L., Smith W.J.M., Symonds E.M., Thomas K.V., Verhagen R., Zaugg J., Mueller J.F. Detection of SARS-CoV-2 RNA in commercial passenger aircraft and cruise ship wastewater: a surveillance tool for assessing the presence of COVID-19 infected travellers. 2020;27 doi: 10.1093/jtm/taaa116.
3. Augenblick N., Kolstad J.T., Obermeyer Z., Wang A. National Bureau of Economic Research; 2020. Group Testing in a Pandemic: The Role of Frequent Testing, Correlated Risk, and Machine Learning, Technical Report.
4. Bae S.H., Shin H., Koo H.-Y., Lee S.W., Yang J.M., Yon D.K. Early release - asymptomatic transmission of SARS-CoV-2 on evacuation flight. Emerg. Infect. Dis. 2020;26(11) doi: 10.3201/eid2611.203353. n/a.
5. Bai Y., Yao L., Wei T., Tian F., Jin D.-Y., Chen L., Wang M. Presumed asymptomatic carrier transmission of COVID-19. JAMA. 2020;323:1406–1407. doi: 10.1001/jama.2020.2565.
6. Bartlett M.S., Vermund S.H., Jacobs R., Durant P.J., Shaw M.M., Smith J.W., Tang X., Lu J.-J., Li B., Jin S., et al. Detection of pneumocystis carinii dna in air samples: likely environmental risk to susceptible persons. J. Clin. Microbiol. 1997;35:2511–2513. doi: 10.1128/jcm.35.10.2511-2513.1997.
7. Browne A., Ahmad S.S.-O., Beck C.R., Nguyen-Van-Tam J.S. The roles of transportation and transportation hubs in the propagation of influenza and coronaviruses: a systematic review. J.Travel Med. 2016;23 doi: 10.1093/jtm/tav002.
8. Buonanno G., Morawska L., Stabile L. medRxiv; 2020. Quantitative Assessment of the Risk of Airborne Transmission of SARS-CoV-2 Infection: Prospective and Retrospective Applications. 2020.06.01.20118984.
9. Cardozo N., Zhang K., Doroschak K., Nguyen A., Siddiqui Z., Strauss K., Ceze L., Nivala J. Multiplexed Direct Detection of Barcoded Protein Reporters on a Nanopore Array. Nat. Biotechnol. 2021:1–5. doi: 10.1038/s41587-021-01002-6.
10. Centers for Disease Control and Prevention Cdc covid data tracker. 2020. https://covid.cdc.gov/covid-data-tracker/
11. Chirico F., Sacco A., Bragazzi N.L., Magnavita N. Can air-conditioning systems contribute to the spread of SARS/MERS/COVID-19 infection? Insights from a rapid review of the literature. Int. J. Environ. Res. Public Health. 2020;17 doi: 10.3390/ijerph17176052.
12. C. for Disease Control, Prevention Cdc 2019-novel coronavirus (2019-ncov) real-time rt-pcr diagnostic panel. 2020. https://www.fda.gov/media/134922/download
13. U. Food D. Administration Emergency use authorization (eua) of the amazon multi-target sars-cov-2 real-time rt-pcr test. 2021. https://www.fda.gov/media/151456/download Accessed: 2021–10-25.

14. Daughton C.G. Wastewater surveillance for population-wide Covid-19: the present and future. Sci. Total Environ. 2020;736 doi: 10.1016/j.scitotenv.2020.139631.
15. Guo Z.-D., Wang Z.-Y., Zhang S.-F., Li X., Li L., Li C., Cui Y., Fu R.-B., Dong Y.-Z., Chi X.-Y., Zhang M.-Y., Liu K., Cao C., Liu B., Zhang K., Gao Y.-W., Lu B., Chen W. Aerosol and surface distribution of severe acute respiratory syndrome coronavirus 2 in hospital wards, Wuhan, China, 2020. Emerg. Infect. Dis. 2020;26 doi: 10.3201/eid2607.200885.
16. Harvey A.P., Fuhrmeister E.R., Cantrell M.E., Pitol A.K., Swarthout J.M., Powers J.E., Nadimpalli M.L., Julian T.R., Pickering A.J. Longitudinal monitoring of SARS-CoV-2 RNA on high-touch surfaces in a community setting 8 168–175. https://pubs.acs.org/doi/10.1021/acs.estlett.0c00875 URL.
17. Hoehl S., Karaca O., Kohmer N., Westhaus S., Graf J., Goetsch U., Ciesek S. Assessment of SARS-CoV-2 transmission on an international flight and among a tourist group. JAMA Netw. Open. 2020;3 doi: 10.1001/jamanetworkopen.2020.18044.
18. Holmgren H., Ljungström E., Almstrand A.-C., Bake B., Olin A.-C. Size distribution of exhaled particles in the range from 0.01 to 2.0 m. J. Aerosol Sci. 2010;41:439–446.
19. Horve P.F., Dietz L., Fretz M., Constant D.A., Wilkes A., Townes J.M., Martindale R.G., Messer W.B., Wymelenberg K.V.D. 6th ed. Vol. 31. 2021. Identification of SARS-CoV-2 RNA in Healthcare Heating, Ventilation, and Air Conditioning Units; pp. 1826–1832. (Indoor Air). 2020.06.26.20141085.
20. Hu M., Lin H., Wang J., Xu C., Tatem A.J., Meng B., Zhang X., Liu Y., Wang P., Wu G., Xie H., Lai S. The risk of COVID-19 transmission in train passengers: an epidemiological and modelling study. Clin. Infect. Dis. 2020 doi: 10.1093/cid/ciaa1057.
21. Hu Z., Song C., Xu C., Jin G., Chen Y., Xu X., Ma H., Chen W., Lin Y., Zheng Y., Wang J., Hu Z., Yi Y., Shen H. Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. Sci. China Life Sci. 2020;63:706–711. doi: 10.1007/s11427-020-1661-4.
22. Huff H.V., Singh A. Asymptomatic transmission during the COVID-19 pandemic and implications for public health strategies. Clin. Infect.Dis. 2020 doi: 10.1093/cid/ciaa654.
23. Junter G.-A., Lebrun L. Cellulose-based virus-retentive filters: a review. Rev. Environ. Sci. Biotechnol. 2017;16:455–489. doi: 10.1007/s11157-017-9434-1.
24. Kasper M.R., Geibe J.R., Sears C.L., Riegodedios A.J., Luse T., Von Thun A.M., McGinnis M.B., Olson N., Houskamp D., Fenequito R., Burgess T.H., Armstrong A.W., DeLong G., Hawkins R.J., Gillingham B.L. An outbreak of Covid-19 on an aircraft carrier. http://www.nejm.org/doi/10.1056/NEJMoa2019375
25. Kim S.-H., Chang S.Y., Sung M., Park J.H., Bin Kim H., Lee H., Choi J.-P., Choi W.S., Min J.-Y. Extensive viable Middle East respiratory syndrome (MERS) coronavirus contamination in air and surrounding environment in MERS isolation wards. clinical infectious diseases: an official publication of the infectious diseases society ofAmerica. 2016;63:363–369. doi: 10.1093/cid/ciw239.
26. Kline E.C., Panpradist N., Hull I.T., Wang Q., Oreskovic A.K., Han P.D., Starita L.M., Lutz B.R. medRxiv; 2021. Multiplex Target-redundant rt-lamp for Robust Detection of sars-cov-2 Using Fluorescent Universal Displacement Probes.
27. Kong D., Wang Y., Lu L., Wu H., Ye C., Wagner A.L., Yang J., Zheng Y., Gong X., Zhu Y., Jin B., Xiao W., Mao S., Jiang C., Lin S., Han R., Yu X., Cui P., Fang Q., Lu Y., Pan

H. Clusters of 2019 coronavirus disease (COVID-19) cases in Chinese tour groups. Transboundary and Emerging Diseases. 2020 doi: 10.1111/tbed.13729. n/a.
28. Luo L., Liu D., Liao X., Wu X., Jing Q., Zheng J., Liu F., Yang S., Bi H., Li Z., Liu J., Song W., Zhu W., Wang Z., Zhang X., Huang Q., Chen P., Liu H., Cheng X., Cai M., Yang P., Yang X., Han Z., Tang J., Ma Y., Mao C. Contact settings and risk for transmission in 3410 close contacts of patients with COVID-19 in Guangzhou, China. Ann. Intern. Med. 2020 doi: 10.7326/M20-2671.
29. Matson M.J., Yinda C.K., Seifert S.N., Bushmaker T., Fischer R.J., van Doremalen N., Lloyd-Smith J.O., Munster V.J. Effect of environmental conditions on sars-cov-2 stability in human nasal mucus and sputum. Emerg. Infect. Dis. 2020;26:2276. doi: 10.3201/eid2609.202267.
30. Medema G., Heijnen L., Elsinga G., Italiaander R., Brouwer A. Presence of sars-coronavirus-2 rna in sewage and correlation with reported covid-19 prevalence in the early stage of the epidemic in the Netherlands. Environ. Sci.Technol. Lett. 2020;7:511–516. doi: 10.1021/acs.estlett.0c00357.
31. Miura T., Masago Y., Sano D., Omura T. Development of an effective method for recovery of viral genomic RNA from environmental silty sediments for quantitative molecular detection. Appl. Environ. Microbiol. 2011;77:3975–3981. doi: 10.1128/AEM.02692-10.
32. T. Moreno, R. M. Pintó, A. Bosch, N. Moreno, A. Alastuey, M. C. Minguillón, E. Anfruns-Estrada, S. Guix, C. Fuentes, G. Buonanno, L. Stabile, L. Morawska, X. Querol, Tracing surface and airborne SARS-CoV-2 RNA inside public buses and subway trains. 147, 106326. URL: https://www.sciencedirect.com/science/article/pii/S0160412020322819 10.1016/j.envint.2020.106326. doi: 10.1016/j.envint.2020.106326.
33. Neiderud C.-J. How urbanization affects the epidemiology of emerging infectious diseases. Infect. Ecol. Epidemiol. 2015;5 doi: 10.3402/iee.v5.27060.
34. Ong S.W.X., Tan Y.K., Chia P.Y., Lee T.H., Ng O.T., Wong M.S.Y., Marimuthu K. Air, surface environmental, and personal protective equipment contamination by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from a symptomatic patient. JAMA. 2020 doi: 10.1001/jama.2020.3227.
35. Panpradist N., Toley B.J., Zhang X., Byrnes S., Buser J.R., Englund J.A., Lutz B.R. Swab sample transfer for point-of-care diagnostics: characterization of swab types and manual agitation methods. PloS one. 2014;9 doi: 10.1371/journal.pone.0105786.
36. Panpradist N., Wang Q., Ruth P.S., Kotnik J.H., Oreskovic A.K., Miller A., Stewart S.W., Vrana J., Han P.D., Beck I.A., et al. Simpler and faster covid-19 testing: strategies to streamline sars-cov-2 molecular assays. EBioMedicine. 2021;64 doi: 10.1016/j.ebiom.2021.103236.
37. Panpradist N., Kline E., Atkinson R.G., Roller M., Wang Q., Hull I.T., Kotnik J.H., Oreskovic A., Bennett C., Leon D., et al. medRxiv; 2021. Harmony covid-19: A Ready-to-use Kit, Low-cost Detector, and Smartphone App for Point-of-care sars-cov-2 rna Detection.
38. Peccia J., Zulli A., Brackney D.E., Grubaugh N.D., Kaplan E.H., Casanovas-Massana A., Ko A.I., Malik A.A., Wang D., Wang M., Weinberger D.M., Omer S.B. medRxiv; 2020. SARS-CoV-2 RNA Concentrations in Primary Municipal Sewage Sludge as a Leading Indicator of COVID-19 Outbreak Dynamics. 2020.05.19.20105999.

39. Rio D.C., Ares M., Hannon G.J., Nilsen T.W. Purification of rna using trizol (tri reagent) Cold Spring Harb. Protoc. 2010;2010 doi: 10.1101/pdb.prot5439. pdb–prot5439.
40. Rosario K., Fierer N., Miller S., Luongo J., Breitbart M. Diversity of DNA and RNA viruses in indoor air as assessed via metagenomic sequencing. Environ. Sci. Technol. 2018;52:1014–1027. doi: 10.1021/acs.est.7b04203.
41. Rothe C., Schunk M., Sothmann P., Bretzel G., Froeschl G., Wallrauch C., Zimmer T., Thiel V., Janke C., Guggemos W., Seilmaier M., Drosten C., Vollmar P., Zwirglmaier K., Zange S., Wölfel R., Hoelscher M. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N. Engl. J. Med. 2020;382:970–971. doi: 10.1056/NEJMc2001468.
42. S. D. of Transportation 2019 Seattle center city commute mode split survey. 2019. https://www.commuteseattle.com/wp-content/uploads/2020/09/2019-Mode-Split-Final-Report.pdf URL: accessed: 2021–12-20.
43. Shen Y., Li C., Dong H., Wang Z., Martinez L., Sun Z., Handel A., Chen Z., Chen E., Ebell M.H., Wang F., Yi B., Wang H., Wang X., Wang A., Chen B., Qi Y., Liang L., Li Y., Ling F., Chen J., Xu G. Community outbreak investigation of SARS-CoV-2 transmission among bus riders in eastern China. JAMA Intern. Med. 2020 doi: 10.1001/jamainternmed.2020.5225.
44. Silcott D., Kinahan S., Santarpia J., Silcott B., Silcott R., Silcott P., Silcott B., Distelhorst S., Herrera V., Rivera D., et al. NEBRASKA UNIV AT OMAHA; OMAHA: 2020. TRANSCOM/AMC Commercial Aircraft Cabin Aerosol Dispersion Tests, Technical Report.
45. Wiersinga W.J., Rhodes A., Cheng A.C., Peacock S.J., Prescott H.C. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. JAMA. 2020 doi: 10.1001/jama.2020.12839.
46. World Health Organization . 2020. Status of Environmental Surveillance for SARS-CoV-2 Virus.
47. World Health Organization and others Coronavirus disease (covid-2019) situation reports. 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/ URL: accessed: 2021–12-20.
48. Wu F., Xiao A., Zhang J., Gu X., Lee W.L., Kauffman K., Hanage W., Matus M., Ghaeli N., Endo N., Duvallet C., Moniz K., Erickson T., Chai P., Thompson J., Alm E. medRxiv; 2020. SARS-CoV-2 Titers in Wastewater Are Higher Than Expected From Clinically Confirmed Cases. 2020.04.05.20051540.Wurtzer S., Marechal V., Mouchel J.-M., Maday Y., Teyssou R., Richard E., Almayrac J.L., Moulin L. medRxiv; 2020. Evaluation of Lockdown Impact on SARS-CoV-2 Dynamics Through Viral Genome Quantification in Paris Wastewaters. 2020.04.12.20062679.Yang N., Shen Y., Shi C., Ma A.H.Y., Zhang X., Jian X., Wang L., Shi J., Wu C., Li G., Fu Y., Wang K., Lu M., Qian G. In-flight transmission cluster of COVID-19: a retrospective case series. Infect. Dis. 2020:1–11. doi: 10.1080/23744235.2020.1800814. Yu P., Zhu J., Zhang Z., Han Y. A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period. J. Infect. Dis. 2020;221:1757–1761. doi: 10.1093/infdis/jiaa077.
49. Yu X., Yang R. COVID-19 transmission through asymptomatic carriers is a challenge to containment. Influenza Other Respir. Viruses. 2020;14:474–475. doi: 10.1111/irv.12743.

50. Zhang R., Li Y., Zhang A.L., Wang Y., Molina M.J. Identifying airborne transmission as the dominant route for the spread of COVID-19. Proc. National Academy of Sciences of the United States of America. 2020;117:14857–14863. doi: 10.1073/pnas.2009637117.
51. Zhao S., Zhuang Z., Ran J., Lin J., Yang G., Yang L., He D. The association between domestic train transportation and novel coronavirus (2019-nCoV) outbreak in China from 2019 to 2020: a data-driven correlational report. Travel Med. Infect. Dis. 2020;33 doi: 10.1016/j.tmaid.2020.101568.
52. Zheng R., Xu Y., Wang W., Ning G., Bi Y. Spatial transmission of COVID-19 via public and private transportation in China. Travel Med. Infect. Dis. 2020;34 doi: 10.1016/j.tmaid.2020.101626.

Supplementary Data

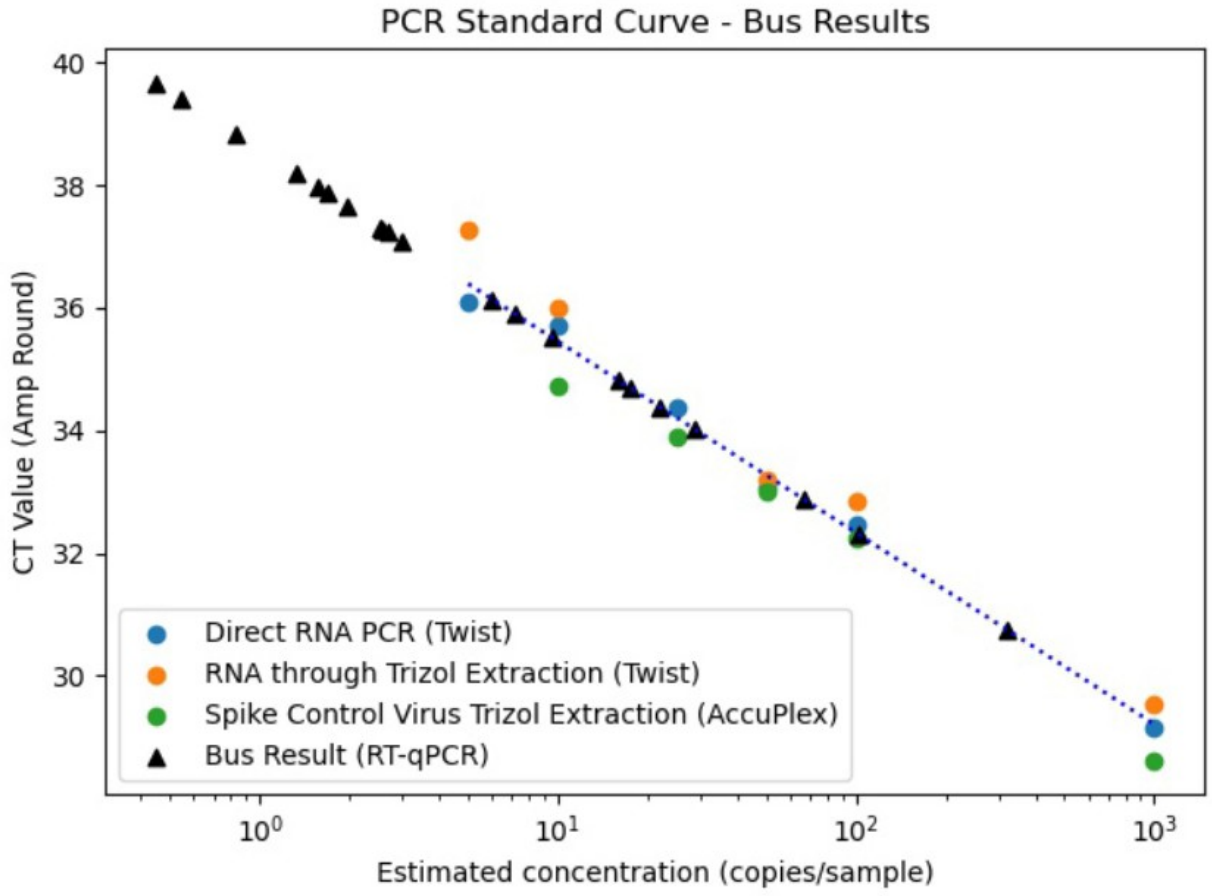


Figure S3.1. Estimated RNA quantification from buses is benchmarked by qPCR tests
For positive RNA controls from Twist and Accuplex, the calculated input RNA is compared to the amount of amplification on the Y-axis. Adding known amounts of RNA to filters and extracting them to test capture efficiency results in some RNA loss indicated by higher CT values. Results from bus samples are reverse calculated to estimate initial particle counts.

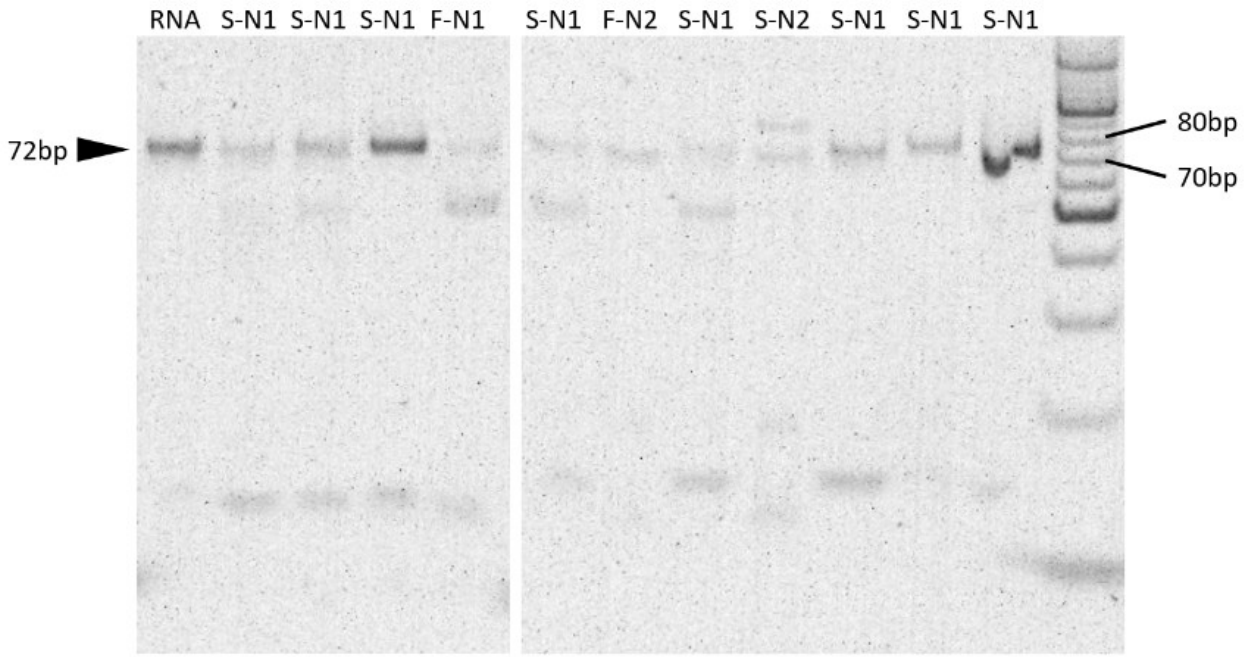


Figure S3.2. Verification of CoV-19 amplicon by acrylamide gel
Reference CoV-19 non-replicable RNA (Twist Bioscience) amplified to expected size at 72bp. Remaining samples from swabs (S) or filters (F) demonstrated amplicons are the expected size.