The Transcriptome Architecture of Polyomaviruses

Jason Nomburg[1,2,3], Wei Zou[4], Thomas C. Frost[1,3], Chandreyee Datta[5,6,7,8], Shobha Vasudevan[5,6,7,8], Gabriel J. Starrett[9], Michael J. Imperiale[4,10], Matthew Meyerson[1,2,11*], James A. DeCaprio[1,3,12*]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston MA

[2]Broad Institute of MIT and Harvard, Cambridge, MA

[3]Harvard Program in Virology, Harvard University Graduate School of Arts and Sciences, Boston, MA

[4]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan

[5]Massachusetts General Hospital Cancer Center, Harvard Medical School, 185 Cambridge St, CPZN4202, Boston, MA

[6]Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA

[7]Center for Regenerative Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA

[8]Harvard Stem Cell Institute, Harvard University, Cambridge, MA

[9]Laboratory of Cellular Oncology, CCR, NCI, NIH, Bethesda, MD, USA

[10]Rogel Cancer Center, Ann Arbor, MI

[11]Department of Genetics, Harvard Medical School, Boston, MA

[12]Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

*Correspondence to:

James A. DeCaprio

james_decaprio@dfci.harvard.edu

Matthew Meyerson

matthew_meyerson@dfci.harvard.edu

Abstract

Polyomaviruses (PyV) are ubiquitous pathogens that can cause devastating human diseases. Due to the small size of their genomes, PyV utilize complex patterns of RNA splicing to maximize their coding capacity. Despite the importance of PyV to human disease, their transcriptome architecture is poorly characterized. Here, we compare short- and long-read RNA sequencing data from eight human and non-human PyV. We provide a detailed transcriptome atlas for BK polyomavirus (BKPyV), an important human pathogen, and the prototype PyV, simian virus 40 (SV40). We identify pervasive wraparound transcription in PyV, wherein transcription runs through the polyA site and circles the genome multiple times. Comparative analyses identify novel, conserved transcripts that increase PyV coding capacity. One of these conserved transcripts encodes superT, a T antigen containing two RB-binding LxCxE motifs. We find that superT-encoding transcripts are abundant in PyV-associated human cancers. Together, we show that comparative transcriptomic approaches can greatly expand known transcript and coding capacity in one of the simplest and most well-studied viral families.

Introduction

Polyomaviruses (PyV) are ubiquitous pathogens that can cause devastating human diseases (Jiang et al., 2009a) including polyomavirus-associated nephropathy (PVAN), hemorrhagic cystitis, and bladder cancer associated with BKPyV (Starrett et al., 2021), progressive multifocal leukoencephalopathy caused by JCPyV, Merkel cell carcinoma caused by Merkel cell polyomavirus (MCPyV), and dermatosis caused by human polyomavirus 7 (HPyV7) (Jiang et al., 2009a; Nguyen et al., 2017). PyV have circular double-stranded DNA genomes and express viral genes with distinct “early” and “late” kinetics. Early and late transcripts are driven by a bi-directional central promoter, and each terminate at their own polyA signal sequence located between the early and late regions. The PyV early region encodes tumor or T antigens that promote cell cycle progression and facilitate replication of the viral genome by host DNA polymerase. The PyV late region originating from the common PyV promoter element on the opposite genome strand encodes the structural proteins required for the generation of progeny virions.

PyV transcripts undergo complex splicing to increase their coding capacity in the face of their small ~5kb genomes. In addition to the major large T (LT) and small T (ST) antigens, additional T antigen splice forms have been identified including transcripts that generate truncated versions of LT (17kT, 57kT, truncT, and T’ in SV40, MCPyV, BKPyV, and JCPyV respectively), a “superT” antigen that contains a duplicated LxCxE RB-binding motif in SV40, middle T (MT) in murine PyV (MPyV), and ALTO in MCPyV (Abend et al., 2009; Carter et al., 2013; Freund et al., 1992; Kress et al., 1979; Shuda et al., 2008; Smith et al., 1979; Trowbridge and Frisque, 1995; Zerrahn et al., 1993). Although the diversity of late transcripts has been explored in SV40 (Good et al., 1988), late transcript diversity in other PyV, including the major human pathogens, is poorly characterized. To address this lack of knowledge of PyV transcription and to discover unannotated biologically relevant PyV-encoded protein products, we used long- and short-read RNA sequencing technologies to characterize the transcriptomes of eight human and non-human PyVs.

Results

RNA sequencing expands PyV transcript diversity.

To expand known PyV transcript diversity, we conducted a series of viral infections followed by total or polyA short-read Illumina RNA sequencing (short-RNAseq (total) and short-RNAseq (polyA) respectively) (**Figure 1A**). We integrated this newly generated data with publicly available data from infected cell culture, human skin, and other settings (**Table 1**). Viruses studied include SV40, BKPyV Dunlop variant and Dik (wild type, or archetype), JCPyV, MPyV, MCPyV, HPyV7, and bark scorpion polyomavirus 1 (BSPyV1).

For SV40 and the BKPyV Dunlop variant, which replicate robustly in cell culture, we complemented short-read sequencing with Nanopore direct RNA sequencing (dRNAseq) and PacBio Single-Molecule Real-Time sequencing (SMRTseq) (**Figure 1A**), two long-read sequencing approaches for polyA RNA with distinct library preparations and sequencing strategies. Resultant RNAseq reads from long- and short-read sequencing strategies were mapped against the viral reference genome and grouped into transcript classes based on the presence of shared introns as detailed in the Methods (**Figure 1B**). For SV40, viral transcripts represented 11.6% and 8.8% of transcripts in dRNAseq and SMRTseq, respectively. For BKPyV Dunlop, viral transcripts represented 28.6% and 27.8% of transcripts in dRNAseq and SMRTseq, respectively. The total number of viral reads is detailed in **Figure S1A**. Transcripts within the same class contain the same introns but may have distinct transcript start sites (TSSs) and transcript end sites (TESs). For most transcriptomes, the majority of transcripts are members of the first few transcript classes (**Figure S1B**). To filter out erroneous splice sites, we required that all introns present in a dRNAseq or SMRTseq read must also be supported by at least 5 splice junction-spanning reads in short-RNAseq (total) data. Detailed information on this transcript class strategy is present in the Methods.

Comparison of read coverage from short-RNAseq (total), dRNAseq, and SMRTseq revealed that dRNAseq and SMRTseq were relatively consistent with read coverage, generally reflected expected patterns of exon usage (**Figure 1C**). In contrast, the read coverage of short-RNAseq (total) was less representative of expected viral exon usage and may reflect noise due to the amplification of smaller RNA fragments (**Figure 1C**).

For SV40 and BKPyV Dunlop, a transcript class (consisting of transcripts with shared introns) was considered a bona-fide viral transcript if it was at least 0.1% of late or early transcripts in dRNAseq or SMRTseq data as described in the Methods. For SV40, which has detailed splice annotations (Good et al., 1988), we found that dRNAseq and SMRTseq data are largely consistent with existing annotations. However, we identified five previously unannotated SV40 transcripts that were supported by both long-read sequencing approaches, plus one additional previously unannotated SV40 transcript class supported by SMRTseq and short-RNAseq (total) (**Figure 1D, Figure 2).**

In contrast to SV40 and despite its clinical importance, BKPyV transcripts have been poorly characterized. We identified a total of 23 transcripts, 21 of which are supported by both dRNAseq and SMRTseq data and only six of which were previously identified (Abend et al., 2009; Seif et al., 1979) (**Figure 1E, Figure 3**). While novel BKPyV late transcripts are often analogous to the characterized wraparound and non-wraparound transcripts previously identified in SV40, several additional and unexpected BKPyV early transcripts were identified. For example, an atypically early TSS revealed a splice donor that was used to generate transcript E3 (**Figure S4**). Early transcripts including E6, E9, and E11 are conserved across numerous PyV and lead to formation of novel ORFs - these are described in detail below.

We generated a comprehensive atlas of SV40 and BKPyV transcripts in **Figures S3 - S8**. Watch plots display the structure of each identified transcript, and read pileups show all transcripts identified in each transcript class. The relative abundance of each transcript as well as exact splice coordinates and abundance information for each identified transcript is provided in **Supplementary Tables 1 and 2**. Transcripts can also be explored using an interactive Google Colab notebook (https://colab.research.google.com/github/jnoms/SV40_transcriptome/blob/main/bin/colab/PyV_exploratory.ipynb). A comprehensive analysis of all splice sites detected in short-read short-RNAseq (total) and short-RNAseq (polyA) in eight PyV studied is presented in **Figure S9**.

To address the possibility that distinct transcript isoforms could be preferentially translated, we performed polysome profiling of SV40-infected cells coupled with dRNAseq of whole-cell and polysome-associated polyadenylated RNAs (**Figure 1F**). The ribosome occupancy, determined as the ratio between a transcript's normalized polysome abundance and its normalized whole-cell abundance, has a mean of slightly above 1 for host transcripts (**Figure S2D**). We found 11.2% of reads in the whole-cell fraction and 18.7% in the polysome fraction were viral, consistent with active translation of viral transcripts. For late transcripts, the relative abundance in the whole-cell fraction was tightly coupled to polysome relative abundance (**Figure 1G**), indicating limited preferential translation of late transcripts. In contrast to late transcripts, we found that the LT:ST ratio was 1.3:1 in the polysome fraction compared to a 3.4:1 ratio of LT:ST transcripts in the whole-cell fraction, indicating preferential translation of ST during infection.

Wraparound transcription is conserved across diverse PyV.

Long-read sequencing revealed the existence of many late transcripts that contain multiple copies of a duplicated leader exon. Leader-leader splicing is due to "wraparound transcription" of PyV transcripts that failed to terminate at the late polyadenylation signal and continue to circle the genome repeatedly. PyV wrapround transcription has been described previously although the structure and diversity of these RNA species is unknown (Adami et al., 1989; Garren et al., 2015; Luo and Carmichael, 1991; Reddy et al., 1978). We investigated these transcripts in dRNAseq data from SV40 and BKPyV. To supplement these data, we also performed dRNAseq on MPyV-infected cells. Wraparound transcription, defined by the presence of repetitive copies of a shared leader sequence, was found in long-read sequencing for all three PyVs (**Figure 4A, B, C**: note the presence of the leader-leader or repeated exon near the “11 o’clock” position in watch plots). In addition to this leader sequence repetition, there are diverse forms of wraparound transcripts that contain various combinations of subsequent introns and encode for distinct viral proteins (**Figure 2, 3**). While only 3.6% of SV40 transcripts originate from wraparound transcription, BKPyV and MPyV have markedly higher rates at 25% and 41% respectively (**Figure 4D**).

Next, we inferred the presence of wraparound transcription in diverse PyV by identifying short-RNAseq (total) reads that span the leader-leader junction (**Figure 4E**). Despite the limited length of these short reads, leader-leader junctions can be accurately identified within a single read through analysis of junction sites (**Figure 4F**). We found evidence of wraparound transcription in all eight PyV investigated here. This includes HPyV7 RNAseq from infected human skin and RNAseq data from a scorpion containing the highly divergent Bark scorpion polyomavirus 1 (BSPyV1), indicating that wraparound transcription occurs *in vivo* and is widely conserved across PyV.

Pervasive premature polyadenylation of early transcripts in SV40, BKPyV, and MPyV.

We found that many early transcripts in SV40 and BKPyV underwent alternative polyadenylation (APA) earlier than the canonical polyA site as indicated by premature transcript end positions near 3 o'clock in the watch plots (**Figure S10A, B**). Early transcript APA had been previously identified in MPyV, where there is a canonical polyA signal sequence (AATAAA) within the LT ORF (Kamen et al., 1980b; Norbury and Fried, 1987). Indeed, dRNAseq identified APA of early transcripts in MPyV-infected cells (**Figure S10C, D**). In contrast to MPyV, APA in SV40 and BKPyV may be driven by alternative polyA signal sequences to the 5' of the APA site (ATTAAA in SV40, AAGAAA or TATAAA in BKPyV). Assessment of the cumulative incidence of early transcript termination shows abrupt increases in transcript termination ~1500nt upstream of the canonical polyA site in all three viruses (**Figure S10D**). This APA appears to be similarly abundant in LT and ST transcripts. We found that transcripts with APA still contain a full polyA tail that, while shorter than the polyA tails of transcripts that use the canonical polyA site, still tend to be longer than the polyA tails of host transcripts (**Figure S10E, S10F, S10G, S2C**). The polyA tail length of a spike-in control RNA with a known 30-adenine polyA tail was correctly estimated by dRNAseq (**Figure S2C**). We find that transcripts containing APA can associate with polysomes (**Figure S10H, I**), indicating that these transcripts are translated.

Comparative analysis of short-RNAseq (total) data reveals conserved, unannotated splice-forms that may generate variant ORFs.

Next, we conducted a comparative analysis of PyV transcription from short-RNAseq (total) data (**Figure S9**), with the hypothesis that data from diverse PyV could reveal unannotated splice forms. This analysis led to the discovery of several unannotated but conserved splicing events that have the potential to expand the coding capacity of PyV (**Figure 5**).

We found that PyVs including HPyV7, MPyV, BKPyV Dunlop, and MCPyV express a transcript utilizing the LT first exon donor but an acceptor within the ST ORF leading to the generation of the ST2 ORF (**Figure 5A**). This splice occurs in-frame in HPyV7 and BKPyV resulting in an internal deletion within ST, while in MPyV and MCPyV this splice lands out of frame and results in the addition of novel C-terminal amino acids. The ST2 splice is highly abundant in HPyV7 representing over 20% of spliced early transcripts from HPyV7-infected human skin. ST2-encoding transcripts were detected in BKPyV dRNAseq and SMRTseq data (transcript E6).

MPyV encodes MT in addition to the LT and ST antigens common with other PyV. MPyV MT is generated from a splicing event that connects the ST ORF with an ORF in the alternative frame of the LT second exon. To our surprise, we found that BKPyV expresses low levels of a similar transcript containing a splice that connects the ST ORF with an MT-like ORF likewise in an alternative frame of the LT second exon (**Figure 5B**). This MT transcript was also detected in BKPyV dRNAseq and SMRTseq data (transcript E9).

JCPyV encodes two VP1 variants, VP1Xs, that consist of the N-terminal region of VP1 with novel C-termini that make up as much as 30% of late spliced transcripts in JCPyV (**Figure 5C**) and have been recently identified and validated by an independent group (Saribas et al., 2018). We found that VP1X-encoding transcripts were also produced by MCPyV, SV40, BKPyV, and MPyV, albeit at a lower abundance than in JCPyV. Except for one JCPyV VP1X-encoding splice, these transcripts were generated from splicing of wraparound transcripts that run through the late polyA signal sequence.

SuperT, a T antigen containing two RB-binding motifs, is present in multiple PyV and in PyV-associated human cancers.

Studies in SV40-transformed cells previously identified a superT antigen with higher molecular weight than LT, containing a duplicated region with two copies of the LxCxE RB-binding motif (Eul and Patzel, 2013). We found that a superT-specific splice was present in SV40, BKPyV (Dik and Dunlop variants), JCPyV, and MCPyV during viral infection (**Figure 5D**). The superT-specific splice originates from a splice donor canonically associated with a conserved truncated LT antigen (17kT in SV40, truncT in BKPyV, 57kT in MCPyV, and T’ in JCPyV), but uses the LT second exon acceptor available due to wraparound transcription. We find evidence of superT in the dRNAseq and SMRTseq data for SV40 and BKPyV Dunlop infections (transcripts E4 and E11 respectively). Western blot with an antibody reactive to LT in BKPyV Dik-infected cells revealed a band with slightly higher molecular weight than LT that is consistent with superT (**Figure 5F**). BKPyV Dik mutant M1, designed to remove ST by replacing the LT intron with an intron from the plasmid pCI (**Figure 5E**), also generated a superT band of expected size. BKPyV Dik mutant M2 was generated by removing the LT intron and adding the pCI intron just 5' of the LT first exon. Should the truncT donor be used to generate superT in this mutant, the only available acceptor is before the LT 1st exon, which would result in the formation of an aberrantly larger superT due to the inclusion of a second copy of the LT first exon (**Figure 5E**). short-RNAseq (polyA) analysis of cells infected with BKPyV Dik WT, M1, or M2 show junctions consistent with this model (**Figure S11**), and western blot revealed that the superT band in M2 is shifted to a higher molecular weight (**Figure 5F**). Together, these data indicate that superT is generated by BKPyV Dik during viral infection.

SuperT was initially identified as an unexplained higher-molecular weight T antigen present in many SV40-transformed cell lines (Kress et al., 1979; Smith et al., 1979). While superT can be generated during viral infection because of wraparound transcription, in SV40-transformed cells it would be possible to yield pre-mRNAs that can be spliced to form superT should the virus be integrated in tandem copies (**Figure 6A**). Indeed, we previously observed that MCPyV integration events in Merkel cell carcinoma (MCC) often lead to partial duplications of the viral genome and result in the tandem insertion of multiple copies of viral early genes (Starrett et al., 2020). Furthermore, the duplicated region in superT includes the RB-binding LxCxE motif, raising the possibility that superT can function as a potent oncogene. We therefore asked if there is evidence of superT in PyV-associated human cancers.

To address this question, we first analyzed short-RNAseq (total) data from five BKPyV-associated bladder cancers (Starrett et al., 2021). To our surprise, we found that short-RNAseq (total) data from two replicates of one BKPyV-associated bladder cancer contained a higher abundance of superT-specific splice than even the LT- or ST-specific splices, suggesting that a large fraction of "LT" in this tumor is superT (**Figure 6B**). We next analyzed short-RNAseq (polyA) data from a series of 30 MCPyV-positive MCCs and found evidence of superT in six cases (**Figure 6B**). Notably, the total number of viral reads in some MCPyV-positive but superT-null tumors was very low, leaving open the possibility that sequencing depth was insufficient to identify the superT splice in additional tumors. Using PCR and sanger sequencing, we confirmed the presence of the superT splice in MCC tumor J45_440 (**Figure S12A**).

We hypothesized that superT may be generated by cis-splicing due to concatemeric integration of multiple copies of the etiologic PyV in these tumors (**Figure 6A**). To address this hypothesis, we investigated three MCCs (J45_440, J17_296, J11_285) for which we possess short-read whole genome sequencing data. From J11_285, we were able to assemble the entire integration site, showing that MCPyV is integrated in a manner that could allow cis-splicing to generate superT (**Figure S12B**). For J45_440, we assembled a single viral block integrated in chromosome 7 (**Figure S12C**). We found that 1) there are likely 2 copies of the viral genome, and 2) the 5' viral integration site appears to fall on chromosome 7 "after" the 3' viral integration site, observations consistent with the existence of two copies of the viral genome in tandem separated by a small segment of host DNA at this integration site. For J17_296, from the assembly, we could infer three distinct segments of viral DNA with integration sites closely spaced within chromosome 2 (**Figure S12D**), indicating a complex integration pattern. The longer block contains two copies of the early region and can likely support superT generation through cis-splicing. The LT ORF of MCPyV is often truncated by premature stop codons or deletions in MCC. We found that a stop codon in J17_296 likely prevents expression of superT, but no stop codons occur before the superT splice in J45_440 or J11_285 (**Figure S12E**). Together, these data indicate that viral integration sites often could support cis-splicing to generate superT.

Two recent studies have found evidence of circular RNAs (circRNAs) that may be generated by MCPyV in MCC and may support the translation of ALTO (Abere et al., 2020; Yang et al., 2021). Of note, the major circRNA splice is equivalent to our proposed MCPyV superT splice - a short-RNAseq read spanning the proposed circRNA junction cannot be differentiated from a read spanning the superT junction. However, the MCC RNAseq samples in which we found superT are short-RNAseq (polyA), which should select against the potential circRNA due to its lack of a polyA tail. Furthermore, we detect the superT splice in short-RNAseq (polyA) of SV40 and BKPyV Dunlop infections in cell culture (**Figure S9**), although at around ~2/3 of its relative abundance in short-RNAseq (total). Finally, we identify full-length superT transcripts in dRNAseq data, which is highly unlikely to sequence circRNA since it is not polyadenylated. This leaves open the possibility that some superT-like splice in short-RNAseq (total) from viral infection originates from circRNA but suggests that most are from linear transcripts that contain a polyA tail.

Discussion

Here, we show that leveraging multiple long- and short-read RNA sequencing approaches across 8 polyomaviruses has allowed us to greatly expand known transcript diversity of this viral family. Short read RNAseq has limited capacity to characterize transcriptome diversity because only a small fraction of reads span splice junctions, and these junctions often cannot be phased with other junctions or to the transcript start and end sites. Integrating long-read sequencing has allowed sequencing of entire transcripts, including phasing of splice sites and transcript start and end positions. Recent studies have leveraged long read sequencing to shed light on exceptional complexity in the transcriptomes of diverse RNA and DNA viruses (Balázs et al., 2017; Depledge et al., 2019; Garalde et al., 2018; Keller et al., 2018; Kim et al., 2020; Nomburg et al., 2020; Price et al., 2020). We have expanded these studies to show that a comparative approach within a viral family can identify conserved transcripts that extend viral coding capacity.

Historically, studies of PyV transcripts were limited by the sensitivity and resolution of northern blots, or by the read length of short read sequencing. Despite these limitations, studies in the 70's and 80's were able to cumulatively characterize several SV40 late transcripts, including one containing leader-leader splicing (Ghosh et al., 1978; Good et al., 1988; Reddy et al., 1978). In contrast to SV40, the architecture of BKPyV late transcripts is poorly characterized - prior to this work, the two major classes of late transcripts (“16S” and “19S”, reflecting transcript size based on gradient sedimentation properties) were the primary late transcript classifications (Seif et al., 1979). Only recently did a study provide some evidence for leader-leader splicing in BKPyV (Zou et al., 2020). While the late transcripts of most PyV are thought to encode the canonical late viral proteins, a recent study in JCPyV identified two splice events that lead to the generation of novel proteins containing the N-terminal region of VP1 - one of which was validated through western blot (Saribas et al., 2018). We found that these transcripts (deemed “VP1X”) are highly expressed in JCPyV but are also expressed at lower level in BKPyV, MPyV, MCPyV, and SV40.

Leader-leader splicing is known to be highly prevalent in MPyV, where as many as 12 leader exons have been observed on a single RNA (Kamen et al., 1980a; Legon et al., 1979; Treisman, 1980) - in our data, we have identified over 15 leader exons in a single transcript. Furthermore, leader-leader splicing is required for stable accumulation of MPyV late transcripts, dependent on length but not nucleotide composition of the leader (Adami et al., 1989). Despite these observations, the exact structure, diversity, and conservation of wraparound transcripts was not understood. Here, we found that leader-leader splicing and wraparound transcription occurs in all PyV studied, including in the divergent Bark scorpion polyomavirus 1, and found that the prevalence of leader-leader splicing varies significantly between PyV. It is possible that this variation reflects differences in the strength of the late polyA signals of these PyV. We found a large diversity of wraparound transcripts containing variable numbers of the leader sequence and diverse patterns of subsequent exon usage.

While late and early transcripts are thought to primarily end at the canonical late or early polyadenylation sites, studies previously observed APA of early transcripts in MPyV (Kamen et al., 1980b; Norbury and Fried, 1987). Here, we likewise identify pervasive APA of early SV40 and BKPyV Dunlop transcripts and find that SV40 early transcripts with APA can associate with polysomes and are likely translated. In addition, polysome profiling revealed that SV40 transcripts are higher abundance in polysome-associated RNAs than in whole-cell RNA populations, indicating preferential translation of SV40 transcripts. The relative abundance of individual late viral transcripts in the polysome closely reflected their whole-cell abundance - conversely, ST transcripts were preferentially translated compared to LT transcripts. The mechanism driving this difference needs further study, as these transcripts differ only by a minor difference in splice donor usage.

In addition to the major early transcripts encoding LT and ST, other early transcripts have been identified in some PyV. MPyV encodes MT, generated by a splice connecting the ST ORF and an ORF overprinted with LT second exon. MT is a primary oncogene in MPyV and was thought to be largely restricted to rodent PyVs (Gottlieb and Villarreal, 2001). We found that BKPyV generates a MT-like ORF through splicing connecting ST and an ORF similarly overprinted with the LT second exon, showing that non-rodent PyVs may be capable of expressing MT-like ORFs. In addition, MPyV also encodes a tinyT antigen consisting largely of the LT first exon, resulting from a splice connecting the LT first exon donor and MT acceptor (Riley et al., 1997). We identified a novel T antigen, ST2, that is generated from a splice from the LT first exon donor to a splice acceptor within the ST ORF. This transcript is highly expressed in HPyV7 and present at lower levels in BKPyV, MPyV, and MCPyV. Many PyV encode a truncated variant of LT - this includes SV40 17kT, BKPyV truncT, MCPyV 57kT and JCPyV T' proteins (Abend et al., 2009; Shuda et al., 2008; Trowbridge and Frisque, 1995; Zerrahn et al., 1993). These transcripts contain a canonical LT splice and a subsequent splice that removes a large portion of the LT ORF.

We found that the same secondary splice sites responsible for truncated LT variants can be used to generate superT. superT was initially observed in many SV40-transformed cell lines (Kress et al., 1979; Smith et al., 1979) - in a similar manner, we find that concatemeric integration of BKPyV and MCPyV in human cancers can facilitate the generation of superT. We also find that superT is generated in lytic infections of SV40, BKPyV, MCPyV and JCPyV. Eul and colleagues have published several studies proposing that SV40 superT can be generated by trans-splicing between two separate pre-mRNAs in the context of artificial expression constructs encoding the SV40 early region (20, 31, 32). However, we find that in MCC tumors that generate superT and for which we can assemble the viral integration site, the viral genome is likely integrated in tandem in a way that could facilitate the cis-splicing of pre-mRNA that spans multiple genome copies. Thus, while we cannot rule out trans-splicing from these data, we believe cis-splicing is more likely. Future studies are necessary to understand the biology of superT including its oncogenic potential and ability to bind multiple RB molecules. Finally, efforts should be taken to understand if superT is expressed by PyV and contributes to disease in other contexts, such as by BKPyV in PVAN or JCPyV in PML.

We show that complex, uncharacterized splicing events are used by PyV to expand their protein coding capacity. Future work is necessary to understand the biological function of these transcripts and proteins. It is possible that unannotated splicing we identify here could be differentially abundant in other biological contexts, so it will be important to investigate PyV splicing in other infection contexts and human diseases. Future transcriptome analyses that integrate long and short reads from multiple viruses may have utility to expand characterized transcript and coding capacity in other viral families.

Conclusions

We provide a comprehensive transcriptome atlas for the prototype PyV SV40, as well as the critically important human pathogen BKPyV. Comparative analyses of PyV transcriptomes reveals conserved splice events that may expand PyV coding capacity. We find that superT, a transcript generated by SV40, BKPyV, JCPyV, and MCPyV that encodes a T antigen containing two RB-binding LxCxE domains, is present in several PyV-associated human cancers. Together, these data expand our understanding of PyV transcriptomes and uncover unannotated PyV-encoded proteins of potential relevance to human disease.

Materials and Methods

Data and code availability.

All code used in this project can be found at the zenodo and github links below. The zenodo repository also contains all processed data necessary to reproduce all analyses and figures. The main processing steps used to process RNAseq data are present as nextflow pipelines which call modular bash and python scripts.

Zenodo: https://doi.org/10.5281/zenodo.5593468

Github: https://github.com/jnoms/SV40_transcriptome

Furthermore, a series of interactive Google Colab notebooks can download all processed data from Zenodo and completely reproduce all analyses and non-schematic primary figures. The colab documents are stored on github at https://github.com/jnoms/SV40_transcriptome/tree/main/bin/colab. Direct links to the Google Colab documents are as follows:

Figure 1:
https://colab.research.google.com/github/jnoms/SV40_transcriptome/blob/main/bin/colab/Figure1.ipynb

Figure 4:

https://colab.research.google.com/github/jnoms/SV40_transcriptome/blob/main/bin/colab/Figure4.ipynb

Figure 6:

https://colab.research.google.com/github/jnoms/SV40_transcriptome/blob/main/bin/colab/Figure6.ipynb

A Google Colab notebook is available for interactive investigation of all SV40 and BKPyV viral transcript classes, and does not require computational skills to use: https://colab.research.google.com/github/jnoms/SV40_transcriptome/blob/main/bin/colab/PyV_exploratory.ipynb

All raw RNA sequencing data are available at the NCBI sequence read archive at accession XXXXXX.

Datasets

Information on all samples and viruses (excluding tumors) can be found in **Table 1**.

Table 1

Virus	Sequencing Type	Origin (Accession)	MOI / Timepoint	Host

SV40	dRNAseq (two replicates)	Generated here	MOI 1 / 48hpi	*C. Sabaeus*
SV40	SMRTseq	Generated here	MOI 1 / 48hpi	*C. Sabaeus*
SV40 (polysome input/whole-cell)	dRNAseq	Generated here	MOI 1 / 44hpi	*C. Sabaeus*
SV40 (polysome)	dRNAseq	Generated here	MOI 1 / 44hpi	*C. Sabaeus*
SV40	Short-RNAseq (total)	Generated here	MOI 1 / 48hpi	*C. Sabaeus*
SV40	short-RNAseq (polyA)	Generated here	MOI 1 / 48hpi	*C. Sabaeus*
BKPyV (Dunlop)	dRNAseq	Generated here	MOI 0.5 / 3dpi	Human
BKPyV (Dunlop)	SMRTseq	Generated here	MOI 0.5 / 3dpi	Human
BKPyV (Dunlop)	Short-RNAseq (total)	Generated here	MOI 0.5 / 3dpi	Human
BKPyV (Dunlop)	short-RNAseq (polyA)	Generated here	MOI 0.5 / 3dpi	Human
BKPyV (Dik) WT	Short-RNAseq (total)	Generated here	MOI 1 / 5dpi	Human
BKPyV (Dik) WT	Short-RNAseq (polyA)	Generated here	MOI 1 / 5dpi	Human
BKPyV (Dik) M1	Short-RNAseq (polyA)	Generated here	MOI 1 / 5dpi	Human

BKPyV (Dik) M2	Short-RNAseq (polyA)	Generated here	MOI 1 / 5dpi	Human
MPyV	dRNAseq	Generated here	Unknown / 28hpi	Mouse
MPyV	Short-RNAseq (total)	Garren et al. (Garren et al., 2015) (SRR2043214)	MOI 50 / 36hpi	Mouse
JCPyV	Short-RNAseq (total)	Assetta et al. (Assetta et al., 2016) (SRR9967610)	Unknown / 9dpi	Human
MCPyV (Synthetic genome)	short-RNAseq (polyA)	Theiss et al. (Theiss et al., 2015) (EBI: ERS760222)	200ng viral DNA / Unknown	Human
HPyV7	Short-RNAseq (total)	Rosenstein et al. (Rosenstein et al., 2021) (SRR11488976, SRR11488977)	From infected human skin	Human
BSPyV1	Short-RNAseq (total)	Identified by Schmidlin et al. (Schmidlin et al., 2021) (SRR5958578)	From whole scorpion	*C. sculpturatus*

Tumor samples

The BKPyV-associated bladder cancer is sample TBC03 that has been described (Starrett et al., 2021). This sample is stranded, short-RNAseq (total).

Merkel cell carcinoma samples: Sections of tissue were isolated from patient-derived tumor biopsies and suspended in RNAlater (Thermo Fischer) until further processing.

RNA and DNA was extracted from each section via the AllPrep DNA/RNA kit (Qiagen). Isolated RNA and DNA were each sequenced (PE150) on the NovaSeq 6000 platform (Illumina) for a depth of 50 M reads or 60x genomic coverage per sample, respectively (Novogene). RNAseq data are unstranded, short-RNAseq (polyA).

SV40 infection and RNA extraction

BSC40 cells (ATCC CRL-2761) were seeded on 150mm dishes at 5.37*10^6 cells per plate - about 70% confluence. After waiting 4 hours for the cells to adhere, cells were infected with SV40 at MOI 1 as previously described (Tremblay et al., 2001) with slight modification. In brief, maintenance media was removed, and each 150mm dish was inoculated with 6mL of virus stock diluted in DMEM + 2% FBS. Infection was allowed to proceed at 37°C, 5% CO2 for one hour, with the plates rocked every 15 minutes to ensure adequate coverage of the solution over the cell monolayer. At the end of this period, DMEM + 2% FBS was added to a final volume of 25mL per 150mm dish. Each dish was then incubated at 37°C, 5% CO2 for 48 hours. RNA was extracted using the QIAGEN RNeasy Mini Plus Kit (QIAGEN 74134). This total RNA was then subjected to Nanopore direct RNA sequencing and Illumina total- and polyA-RNA sequencing as described below.

BKPyV infection and RNA extraction

Archetype and rearranged BKPyV (Dik and Dunlop, respectively) were purified and titrated as described (Jiang et al., 2009b). RPTE-hTERT cells (Zhao and Imperiale,

2019) were plated in 6-well plate and prechilled for 15 min at 4°C and infected with Dik or Dunlop at a MOI of 1 and 0.5 fluorescence-forming unit (FFU)/cell, respectively. The cells were incubated at 4°C for 1 h with gentle shaking every 15 min. The virus was removed and fresh REGM medium was added to the cells. Dik and Dunlop infected cells were collected at 120 hpi and 96 hpi, respectively. Total RNA was extracted using the Direct-zol RNA MiniPrep kit (ZYMO Research, USA). This total RNA was then subjected to Nanopore direct RNA sequencing and Illumina total- and polyA-RNA sequencing as described below.

MPyV infection and RNA extraction

C57 mouse embryo fibroblasts (ATCC SCRC-1008) were plated on a 150mm dish at 40% confluence. After several hours of growth, the typical DMEM + 10% FBS media was replaced with serum free DMEM. The next day, the crude viral stock was thawed at 37°C, incubated at 45°C for 20 minutes to facilitate the final liberation of virus into the supernatant, and cell debris removed from the viral stock with centrifugation. The prepared virus stock was then diluted 1:10 with an absorption buffer consisting of HBSS with 10mM HEPES, 1% FBS, at pH 5.6. Media was removed from the target cells, and 6mL of diluted virus in absorption buffer was added. Infection was allowed to proceed at 37°C, 5% CO2 for one hour, with the plates rocked every 15 minutes to ensure adequate coverage of the solution over the cell monolayer. At the end of this period, the absorption buffer was removed and DMEM + 2% FBS was added to a final volume of 25mL per 150mm dish. Cells were inoculated for 28 hours at 37°C, 5% CO2, after which RNA was extracted using TRIzol (ThermoFisher 15596026) according to the manufacturer's instructions. This total RNA was then subjected to Nanopore direct RNA sequencing as described below.

The virus stock used here was kindly provided by the lab of Robert Garcea. This virus stock (viral strain NG59RA) was a crude supernatant from MPyV-infected cells originally generated by the lab of Thomas Benjamin on 02/08/2011 and was of unknown titer. This stock was subjected to a total of three freeze-thaw cycles before use.

SV40 polysome profiling

BSC40 cells were plated on 4 150mm dishes at 60% confluence. After waiting 4 hours for the cells to adhere, cells were infected with SV40 at MOI 1 as reported above. At 44 hours post infection cell culture media was replaced with media containing 100ug/mL cycloheximide and incubated for 5 minutes. Plates were placed on ice, media discarded, and cells were scraped into PBS containing 100ug/mL cycloheximide. Cells were spun down, the PBS discarded, and cells were lysed in a lysis buffer containing 10mM Tris (pH 8), 100mM KCl, 10mM MgCl2, 2mM DTT, 1% Triton X100, 100ug/mL cycloheximide, and 1unit/uL SUPERase RNase inhibitor (Thermo AM2694). Lysates were incubated on ice for 20 minutes with intermittent tapping, and then spun at 10,000g for 10 minutes at 4°C. The supernatant was loaded onto a 10-55% sucrose gradient followed by ultracentrifugation (Beckman Coulter Optima XPN-100 ultracentrifuge) at 32,500 × rpm at 4 °C for 80 minutes in the SW41 rotor. Gradients were prepared with a gradient mixer and pump. Samples were separated by density gradient fractionation system (Biocomp Piston gradient fractionator IP). RNA was extracted from reserved input ("whole-cell") lysate, as well as the polysome fraction using TRIzol. Equal volumes of each fraction containing heavy polysomes (>2) was pooled prior to extraction (Lee et al., 2020).

Western blotting

Infection of wildtype Dik and two Dik mutants in RPTE-hTERT cells was performed as mentioned above. Protein samples were harvested in E1A buffer with protease and phosphatase inhibitors, electrophoresed, transferred, and probed with large tumor antigen antibody (pAb416) as previously described (Zhao and Imperiale, 2019).

RNA sequencing

The concentration of total RNA was determined using the Qubit Fluorometer with the Qubit RNA HS Assay Kit (ThermoFisher Q32852). RNA quality was then assessed on an Agilent Bioanalyzer and the RNA 6000 Pico Kit (Agilent 5067-1513). PolyA RNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB E7490S) with an input of 5ug of total RNA - for SV40 and BKPyV Dunlop, up to 8 total reactions were used to yield sufficient polyA RNA (500ng) for subsequent protocols. In the case of MPyV, due to limited amounts of total RNA, three reactions were used to yield roughly 100ng of polyA RNA. PolyA RNA concentration was determined again using the Qubit RNA HS Assay Kit (ThermoFisher Q32852). PolyA RNA was then concentrated to 9uL using a centriVap.

500ng of polyA RNA (or, in the case of MPyV, 100ng) in 9uL was then processed using the Nanopore Direct RNA sequencing kit (SQK-RNA002). Resultant libraries were sequenced for up to 24 hours on a MinION using an R9.4.1 flow cell.

In the case of polysome profiling: extracted RNA from the input and polysomes were separately subjected to 5 reactions each of the NEBNext Poly(A) mRNA Magnetic Isolation Module using 5ug RNA input per reaction. All resultant polyA RNA was then processed using the Nanopore Direct RNA sequencing kit (SQK-RNA002). Resultant libraries were sequenced for up to 24 hours on a MinION-Mk1C using an R9.4.1 flow cell.

Illumina total RNA sequencing and polyA RNA sequencing of SV40-, BKPyV Dunlop-, and BKPyV Dik-infected cells was conducted by Novogene Corporation Inc. The QC for the RNA samples was performed using Qubit and Bioanalyzer instruments. Libraries were then prepared using NEBNext Ultra II with RiboZero Plus kit (for short-RNAseq (total)) and NEBNext Ultra II with PolyA Selection kit (for short-RNAseq (polyA)). Both library approaches are strand-specific. Library quality and concentration was assessed with Labchip and qPCR. Libraries were sequenced on NovaSeq6000 using PE150 sequencing.

PacBio SMRT sequencing of SV40 and BKPyV Dunlop was conducted by the Georgia Genomics and Bioinformatics Core. Each sample was subjected to IsoSeq library preparation and sequenced on an individual 8M SMRT cell for 26 hours on a Sequel-II machine.

Initial Sequence Processing

Raw Nanopore dRNAseq reads from standard SV40, BKPyV Dunlop, and MPyV infections were basecalled with Guppy version 4.2.2 with the following command:

```
guppy_basecaller -i fast5 -s basecalled --flowcell FLO-MIN106 --kit SQK-RNA002 -r --trim_strategy rna --reverse_sequence true --u_substitution true --cpu_threads_per_caller 10
```

Raw Nanopore dRNAseq reads from polysome profiling of SV40 transcripts were basecalled on a MinION-Mk1C using MinKNOW version 21.02.2.

PacBio SMRTseq subreads were processed using ccs (version 6.0.0). Full-length, nonchimeric reads were then generated using the lima (version 2.0.0) and Isoseq3 (version 3.4.0) packages provided by PacBio.

Stranded Illumina short-RNAseq (total) and short-RNAseq (polyA) reads were processed in the following way: Files containing read 1 (R1) and read 2 (R2) were trimmed and adapters removed using Trim Galore! (Krueger, 2016). Next, reads in R1 files were reverse complemented to orient the reads correctly relative to the transcript of origin, and all read headers in the R1 and R2 files were labeled with "_1" or "_2" respectively. The R1 and R2 files were then concatenated. This Illumina processing pipeline is available in process_illumina.nf.

The MCC tumor RNAseq assessed in this manuscript were short-RNAseq (polyA) that were NOT stranded. This means that the strand of origin of each read is unknown. To address this uncertainty, the complement AND reverse complement of both R1 and R2 were concatenated into the final FASTQ file. As described below in the section "Processing of short-read short-RNAseq (total) and short-RNAseq (polyA) span files", future processing kept the most-likely alignment strand for each read.

Sequence Alignment and Processing

Most long-read sequencing data and Illumina sequencing data were aligned to the appropriate viral genome using Minimap2 (Li, 2018). The exceptions are the short-RNAseq (total) JCPyV data from Assetta et al. (Assetta et al., 2016) and the HPyV7 data from Rosenstein et al. (Rosenstein et al., 2021) - these samples contained sequencing reads of 101bp or shorter and were instead mapped with STAR (Dobin et al., 2013). All non-primary alignments were discarded. Sequence alignments in BAM format were then converted to BED using bedtools (Quinlan and Hall, 2010). Here, bedtools considers any Minimap2- or STAR-called intron ("N" cigar flag) as an intron to split alignment segments. Parameters for alignment and bed conversion can be found in minimap2.sh and star.sh.

To capture transcripts that originate from a pre-mRNA that circled the viral genome more than once, and therefore contain repetitive sequences, all alignments were conducted against concatenated copies of the viral genome. In the case of short-read short-RNAseq (total) and short-RNAseq (polyA), the reference consisted of two concatenated copies of the viral genome. For long-read dRNAseq and SMRTseq, the reference consisted of twenty concatenated copies of the viral genome.

Because the references consisted of multiple copies of the same viral genome, mapped reads were assigned to a random copy of the genome. Therefore, all reads in resultant BED files were “slid” such that they started in the first genome copy of the reference using bed_slide_wraparound_reads.py.

All reference genomes can be found in resources/ref directory of the associated github repository. All references used contain the PyV late region at the start/5’ end of the reference on the “+” or sense strand, with the early region on the antisense or “-“ strand. The concatenated references are based on the following reference genomes collected from NCBI, with any modifications listed:

- SV40: NC_001669.1. The first 100 nucleotides were moved to the end of the sequence.
- BKPyV: KP412983.1
- JCPyV: NC_001699.1

- MPyV: NC_001515.2. The sequence was reverse-complemented to orient the late region towards the start of the reference.
- MCPyV: NC_010277.2
- HPyV7: NC_014407.1
- BSPyV1: LN846618.1

Next, a span file was generated from each slid BED file using bed_to_span.py. This script splits each read into "spans", where each span is an exon or an intron with all positions relative to the viral genome. The introns are defined by the Minimap2- or STAR-called introns ("N" cigar flag) as mentioned above. All regions between the start and end of the reads that are not introns were called as distinct exons. Transcripts were clustered into transcript classes based on introns as discussed below. A "tidy" output span file was then generated that contains the name, strand, and transcript class of a given read, with separate lines for the start and end of each span (e.g., exon or intron) within the sequencing read.

Alignment of repetitive regions

Reads that originate from a transcript that circles the genome more than once can be detected because there is one or more repetitive regions within the read. Alignment against multi-copy reference genomes (20 copies in the case of dRNAseq and SMRTseq) as described above sufficiently captured most of these transcripts, with some exceptions. First, BKPyV SMRTseq data had a poor alignment rate of the leader exon in late WA transcripts - this means that WA transcripts are underrepresented in the BKPyV SMRTseq data. Second, alignment of superT and superT* transcripts from SMRTseq and dRNAseq data was generally poor, with the repetitive region often failing to map via Minimap2. Potential superT and superT* reads in dRNAseq and SMRTseq data were identified through assessment of BAM files following mapping. Early reads that contain a CIGAR flag showing an insertion of 100 bases or more were flagged, and up to 50 of these transcripts were manually investigated through online BLASTN (Johnson et al., 2008) against the viral reference genome. Reads supporting superT in SV40 dRNAseq data, superT* in SV40 SMRTseq data, and superT in BKPyV SMRTseq data were initially missing from Minimap2 alignments but were identified via this approach. One transcript of each type was then repaired upon data import to R such that these transcripts are represented in downstream visualizations - these actions are clearly marked in UTILS_import_data.R. Thus, superT and superT* in SMRTseq and dRNAseq data are underrepresented in abundance plots (**Figure S3D, S4D**) and read pileups (**Figure S6, S8**) compared to their actual abundance in the cell due to these alignment challenges.

Generation of transcript classes

Transcript classes were generated during processing of BED files using bed_to_span.py. Each transcript class consists of sequencing reads that contain the same combination of introns. The transcript class number is based on the abundance of transcripts within a transcript class - e.g., transcript class 1 contains more transcripts than transcript class 2, and so on. Transcript class generation is similar for both long-

and short-read sequencing data, although short reads usually (but not always) tend to contain a maximum of one intron. Notably, transcript class assignment is independent of the transcript start and end positions, meaning that there can be heterogeneity of transcript start and end positions within a transcript class. For all SMRTseq and dRNAseq data, for a transcript class to be generated all introns contained within the transcript class were required to be supported by at least 5 junction-spanning reads within a short-RNAseq (total) dataset. For SV40 and BKPyV Dunlop SMRTseq and dRNAseq data, the short-RNAseq (total) data was generated from RNA from the same extraction. SV40 dRNAseq replicate 2 was corrected with the short-RNAseq (total) data from the first SV40 replicate. For the MPyV dRNAseq data, short-RNAseq (total) data from Garren et al. was used. If a transcript contained an intron that was not supported by at least 5 junction-spanning reads in the Illumina dataset, it was discarded. We opted to use this filtering strategy rather than implementing long-read correction because correction algorithms were unable to cope with wraparound transcripts.

There were limited circumstances where dRNAseq or SMRTseq transcript classes were removed manually during processing - this occurred to four transcript classes that made it through filtering. In these circumstances, alignments were deemed to be artifactual due to Minimap2 alignment errors. These instances are clearly programmatically marked in UTILS_import_data.R with specific rationale for each action.

Splice coordinate system

All splice or intron positions marked in any figure or table of this manuscript are **0-indexed positions of the intron**. To convert these coordinates to the 1-indexed/absolute position of the intron on the viral genome, add 1 to the intron start position. For example, for the intron 276-1600, viral genome nucleotide # 277 is the first nucleotide within the intron, and viral genome nucleotide # 1600 is the last nucleotide within the intron.

Processing of short-RNAseq (total) and short-RNAseq (polyA) span files

The majority of the short-read RNAseq data investigated here used a strand-specific sequencing strategy (except for the MCC tumor RNAseq). With this strategy, the strand of origin for the transcript yielding each read is known, and a read can be correctly assigned to the sense ("+" / late) or antisense ("-" / early) strand. However, a fraction of transcripts can be inaccurately stranded due to artifacts during library preparation. When there were many more late reads than early reads in a short-read dataset, a prohibitive fraction of "early" reads would be reads from late transcripts that were incorrectly stranded due to this artifact. To address this issue, short reads that aligned to the + strand were required to either start or end within the late region (defined as the first ½ of the genome), and short-reads that aligned to the - strand were required to either start or end within the early region (defined as the second ½ of the genome).

Transcript identification

For **Figure 1** and all supplementary figures, a SV40 or BKPyV transcript was identified and assigned a transcript ID if it was at least 0.1% of early or late strands in dRNAseq or SMRTseq data with one exception - SV40 transcript L8 had been previously identified and was kept despite being at only 0.06% abundance. Existing transcript names, where available, were taken from relevant studies (Abend et al., 2009; Good et al., 1988; Seif et al., 1979; Zerrahn et al., 1993). This assignment occurred from the span files, meaning that all sequencing reads in question were previously required to contain introns that were supported by at least 5 short-RNAseq (total) junction-spanning reads. For SV40, for which there was two dRNAseq replicates, identification of a sequencing read at 0.1% or greater in just one replicate was sufficient.

Transcript IDs (e.g., E1, E2, E3…, L1, L2, L3,…) consist of the kinetic class (E: Early, or L: Late) of the identified transcript followed by an integer value in ascending order of abundance. This abundance value was calculated by ordering the transcripts in order of the maximum observed relative abundance in dRNAseq or SMRTseq data.

Of note, the relative abundance of transcripts between dRNAseq and SMRTseq data is skewed by distinct read-length biases between the two approaches. The dRNAseq approach has a 3' bias and a bias towards shorter transcripts, while SMRTseq library preparation resulted in preferential sequencing of transcripts closer to ~2500bp in length. Resultant differences in the length of aligned reads can be seen in **Figure S1C**. The TSS distribution of SV40 late transcripts varies between transcript classes, while the late TSS distribution tends to be similar across transcript classes in BKPyV (**Figure S2A**).

Calculation of sequencing coverage

To determine the sequencing coverage for each sample (as in **Figure 1C**), BAM files from alignment were "slid" such that all transcripts must start in the first genome copy of the reference using bam_slide_wraparound_reads.py, in a similar manner as the beds were slid as described above. Forward and reverse strand reads were split, and the depth was calculated using the command `samtools depth -aa -d0` separately for forward and reverse reads. These processing steps are present in bam_coverage.nf. During plotting, the coverage for each strand was normalized to the maximum coverage at any position (e.g., the maximum coverage of the late and early strands was set to 1).

Watch plots

Each panel of a watch plot represents information for a single transcript class. The center "arms" of these plots are histograms detailing the distribution of start (blue) and end (red) positions for the transcripts within the transcript class. These histograms are normalized to the highest abundance position. The outer ring of each watch plot shows the viral ORF map. Each inner grey ring indicates the number of genomes spanned - all transcripts are displayed moving outwards from the center. Red segments indicate the **exons** of each transcript class. The first exon starts on the most-inner grey ring at the most common transcript start site for the transcript class, and the last exon ends on the most-outer grey ring at the most common transcript end site for the transcript class. The 3' end of the transcript is indicated by the red arrow at the end of the last exon. Thus, the transcripts spiral outwards from the center in the direction of the red arrow. **Figure 4** contains a schematic key describing watch plots.

Read pileup plots

Each square/rectangular panel of a read pileup plot shows the reads present in a single transcript class. The arrows at the top of each panel indicate the viral ORF map, with dashed lines indicating the end of each genome copy. Next, the lines indicate histograms of the transcript start (blue) and end (red) sites for the transcripts within the transcript class. Below the x-axis, each row indicates a single sequencing read. The spans in red indicate the exons inferred from a sequencing read, while the spans in pink indicate the introns/splice junctions. Sometimes the distribution of transcript end positions for a transcript class can be obscured by the thickness of the transcript lines - the histograms should always be consulted to assess abundance.

For SV40 dRNAseq watch and pileups: There were two SV40 dRNAseq replicates. Watch plots and read pileups are based on replicate 1, although missing transcripts that were identified in replicate 2 but not 1 were also plotted.

Short-read intron plots (**Figure S9, S11**)

In these plots, lines indicate specific introns. The upper and lower horizontal arrows indicate the viral ORF map - often, these ORF maps will indicate two concatenated viral reference genomes. The circles above or below each ORF map indicate the percentage of early or late introns that fall at each genome position. Early introns and percentages are colored red, while late introns and percentages are colored blue.

polyA tail length

polyA tail length was determined from dRNAseq data using the `polya` command of Nanopolish (Loman et al., 2015). To determine the polyA distribution of host transcripts, sequencing reads were aligned to the human GRCh38 (for BKPyV samples), *C. Sabaeus* (for SV40 samples), or mouse (for MPyV) cDNA transcriptomes downloaded from ensembl. Only reads with a Nanopolish QC tag of “PASS” were considered for downstream polyA tail length analyses.

The dRNAseq library preparation included the addition of the “RNA Control Standard” (RCS), which is a synthetic RNA based on yeast ENO2 containing a 30-adenine polyA tail. dRNAseq samples were mapped against ENO2 to assess the polyA tail length distribution of this control.

The cumulative incidence of transcript termination (**Figure S10D**) was calculated by determining, for each early read, how far the read's transcript end site is from the canonical polyA site position for each virus.

Polysome profiling analysis

To determine the ribosome occupancy of host genes, dRNAseq reads were aligned to the *C. Sabaeus* cDNA transcriptome downloaded from ensembl. The number of reads mapped to each transcript was extracted with `samtools idxstats`. Transcripts were filtered to include only those with at least 10 reads in both polysome and input fractions. The normalized abundance of each transcript in each fraction was defined as (# of mapped reads)/(total number of virus and host mapped reads). Ribosome occupancy of each transcript was determined as (normalized abundance in polysome)/(normalized abundance in whole-cell), where a value of >1 indicates preferential translation.

Ribosome occupancy of individual viral transcripts could not be calculated because of increased rates of transcript truncation in the polysome fraction compared to the whole-cell fraction. This was indicated by a nearly doubled proportion of unspliced reads with premature 5' ends in the polysome fraction compared to the whole-cell fraction, and likely indicates transcript degradation during sucrose centrifugation or fraction collection. Because viral transcripts are mostly identical and vary largely at a 5' splice site, elevated transcript truncation decreased the observed abundance of individual viral transcripts in the polysome fraction and make ribosome occupancy calculations for individual viral transcripts unreliable.

MCC440 superT PCR and sanger sequencing

Anchored poly-dT primers (Life Technologies) were used for specific reverse-transcription of full-length mRNA into cDNA. Primers were designed to uniquely amplify the super-LT junction through exploitation of repetitive sequences. Primer sequences were as follows (5’ -> 3’); Forward: CTGGACTGGGAGTCTGAAGC, Reverse: ACCCCTCCTCCATTCTCAAGA. Q5 polymerase (NEB) with standard reaction conditions was used for amplification.

Generation of integrated PyV structures and viral variant calling

Tumor WGS was aligned against a fusion reference genome containing hg38 and Merkel cell polyomavirus (NC_010277) using bowtie2 with default parameters. Integrated virus assembly graphs and annotations were generated using Oncovirus tools (https://github.com/gstarrett/oncovirus_tools). Assembly graphs were then manually interpreted to create linear integration structures for PyV-associated MCC.

Point mutations were called in the PyV genomes using lofreq with default parameters (https://csb5.github.io/lofreq/) (PMID: 23066108). Lofreq output was functionally annotated with SnpEff (http://pcingola.github.io/SnpEff/) (PMID: 22728672) using the relevant GenBank gene annotations for the above genomes. Variants were plotted out in R with the ggplot2 package.

List of abbreviations

APA - Alternative polyadenylation

BKPyV - BK Polyomavirus

BSPyV1 - Bark scorpion polyomavirus 1

dRNAseq - Nanopore direct RNA sequencing

HPyV7 - Human polyomavirus 7

JCPyV - JC Polyomavirus

LT - Large T antigen

MCC - Merkel cell carcinoma

MCPyV - Merkel cell polyomavirus

MPyV - Murine polyomavirus

MT - Middle T antigen

ORF - Open reading frame

SMRTseq - PacBio SMRT sequencing

ST - Small T antigen

SV40 - Simian virus 40

TES - Transcript end site

TSS - Transcript start site

PVAN - Polyomavirus-associated nephropathy

PyV - Polyomavirus

Declarations

Competing interests

M.M. receives research support from Bayer, Janssen, Ono; consults for Bayer, Interline, Isabl; and receives patent royalties from Labcorp and Bayer. J.A.D. has received research support from Rain Therapeutics, Inc. and is a consultant for Rain Therapeutics, Inc. and Takeda, Inc.

Funding

This work was supported in part by the US Public Health Service grants R35CA232128 and P01CA203655 and by the Bridge Project, a partnership between the Koch Institute for Integrative Cancer Research at MIT and the Dana-Farber/Harvard Cancer Center to

J.A.D.; a Cancer Grand Challenges OPTIMISTICC team award (C10674/A27140) and by National Cancer Institute grant R35CA197568 to M.M.; R01 AI060584 and R21 AI147155 to M.J.I; CD & SV were supported by R35GM134944; GJS is supported by the NIH Intramural Research Program.

Authors' Contributions

Conceptualization - J.N., M.M., J.A.D.

Data curation - J.N.

Formal analysis - J.N.

Funding acquisition - M.M, J.A.D., M.J.I., S.V., G.J.S.

Investigation - J.N., W.Z., T.C.F., C.D.

Methodology - J.N., M.M., J.A.D.

Project administration - J.N.

Resources - J.N., M.M., J.A.D., T.C.F., W.Z., M.J.I, S.V., C.D.

Software - J.N.

Supervision - M.M., J.A.D.

Validation - J.N.

Visualization - J.N.

Writing, Original draft - J.N., M.M., J.A.D.

Writing, review & editing - J.N., W.Z., T.C.F., C.D., S.V., G.J.S., M.J.I., M.M., J.A.D.

Acknowledgements

We thank Katie Vicari (www.KatieRisVicari.com) for her work illustrating Figure 1A, 2, 3, and S12B. Figure 1F was made in BioRender. We thank Mary O'Reilly with the Broad Research Communication Lab for helpful discussions on figure design. We thank Robert Garcea and Kimberly Erickson for kindly providing stocks of MPyV. Portions of this research were conducted using the O2 High Performance Computing Cluster, supported by the Research Computing Group at Harvard Medical School. Portions of this work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Figure legends

Figure 1 - RNA sequencing expands known SV40 and BKPyV transcript diversity.

A. Overview of experimental procedures. Cells were infected with a polyomavirus, and RNAs extracted. RNA was sequenced using long-read (Nanopore dRNAseq and PacBio SMRTseq) and short-read (Illumina short-RNAseq (total) and short-RNAseq (polyA)). Transcripts were analyzed, and the impact of observed splice events on viral open reading frames was assessed.

B. Mechanism of transcript clustering in this study. Transcripts were aligned to the viral genome and grouped into transcript classes based on the presence of shared introns. Thus, within a transcript class there may be variation in the exact transcript start and end positions. This clustering strategy was used for both long- and short-RNAseq data.

C. Viral RNA sequence coverage for SV40 and BKPyV as determined from dRNAseq, SMRTseq, and short-RNAseq (total) data. The Y axis indicates the scaled coverage, with X axis indicating the position on the viral genome. Coverage for late transcripts (mapping to the + strand) is above the x axis, while coverage for early transcripts (mapping to the - strand) is below the x axis. Coverage is scaled separately for each strand such that the maximum observed coverage for each strand is 1. Arrows at the top of the plot indicate the positions of viral genes.

D-E. UpSet plot indicating the overlap between existing transcript annotations, dRNAseq data, and SMRTseq data for SV40 (**D**) and BKPyV Dunlop (**E**). Bars indicating overlap with existing transcript annotations are black, while those indicating no overlap with existing annotations are blue. These blue bars indicate the number of novel, unannotated transcripts identified.

F. Overview of polysome profiling of SV40-infected cells. BSC40 cells were infected with SV40. Cells were lysed, and a portion of the lysate was subjected to dRNAseq (representative of the RNA content of the whole cell). The remaining lysates was centrifuged through a sucrose gradient, after which fractions containing RNA associated with two or more ribosomes were pooled and subjected to dRNAseq.

G. Relative abundance of SV40 early and late transcripts in the whole-cell and polysome fractions of SV40-infected cells. Y-axis indicates the percentage of early or late transcripts and is log scale. X axis indicates each transcript, with black dots indicating each transcript's whole-cell relative abundance and red dots indicating each transcript's polysome relative abundance.

Figure 2 - Annotated and novel SV40 transcripts.

A. Transcripts are shown relative to the viral genome. Each line is a viral transcript, with red lines indicating exons and dashed blue lines indicating introns. Spokes indicate the positions of common splice donors and splice acceptors. Transcripts that were annotated prior to this study are on a yellow background, and novel transcripts are on a while background. Wraparound transcription that results in multiple copies of a region is annotated with double lines, and the number of copies is indicated in parentheses. The line labeled "pA" indicates the approximate position of the polyA signal sequence.

Figure 3 - Annotated and novel BKPyV transcripts.

A. Transcripts are shown relative to the viral genome. Each line is a viral transcript, with red lines indicating exons and dashed blue lines indicating introns. Spokes indicate the positions of common splice donors and splice acceptors. Transcripts that were annotated prior to this study are on a yellow background, and novel transcripts are on a while background. Wraparound transcription that results in multiple copies of a region is annotated with double lines, and the number of copies is indicated in parentheses. The line labeled “pA” indicates the approximate position of the polyA signal sequence.

Figure 4 - Pervasive wraparound transcription across PyV

A-C. Watch plots indicating the top 4 highest abundance late wraparound transcript classes in dRNAseq data from SV40 (A), BKPyV Dunlop (B), and MPyV (C). The outer ring of each watch plot indicates the position of the viral ORFs. The inner arms are histograms detailing the distribution of transcript starts (in blue) and ends (in red) for transcripts within each transcript class. The red segments indicate exons. Transcripts start in the innermost ring - a second or third ring indicates that the pre-mRNA that generated the transcript must have circled the viral genome multiple times. The 3’ end of the transcript and the direction in which these plots are oriented is indicated by the red arrow at the end of the last exon segment. The red exon segments start at the most common transcript start site within the transcript class, and end at the most common transcript end site within the class. The watch plot key shows an example of the path of the pre-mRNA for SV40 transcript class L6_I.

D. Bar plots indicating the percentage of late transcripts that span a given number of genome lengths in SV40, BKPyV Dunlop, and MPyV dRNAseq data.

E. The leader-leader junction, that connects the pre-mRNA from one genome to the subsequent wraparound, was identified in Illumina short-RNAseq (total) data. The intron in question is plotted as a black line in this plot, with the x axis indicating the genomic position of the intron. The top late wraparound transcript for each virus was plotted. The gene map indicates the approximate gene position and is accurate for SV40 - the exact position of the viral genes varies between viruses. Percentages indicate the percentage of late junction-spanning transcripts that support the plotted wraparound leader-leader junction.

F. Schematic illustrating how leader-leader wraparound transcription can be detected from short read short-RNAseq (total). Leader-leader splicing can be seen as a repetitive exon in watch plots from long-read RNAseq data. Ultimately, there was an original processed mRNA in the cell that contained two tandem leader sequences. When this transcript of origin is sequenced via short read sequencing, reads will be generated across its length. A minority of these reads will span the leader-leader junction, and mapping against the viral reference genome can be used to uncover leader-leader splicing.

Figure 5 - Detection of novel, conserved splicing events that expand PyV coding capacity.

A-D. Schematics illustrating identified ORFs. Each row is a reading frame (except for ST and the LT 1st exon, which are in the same frame), and unannotated amino acids are represented by grey boxes. The measured intron is indicated by the red arrow. Colored ORFs are annotated, while grey ORFs are unannotated. Percentages on the right side of the figure are the percentage of spliced viral transcripts on the same strand as determined from short-read short-RNAseq (total) data. Numbers after each virus name indicate the transcript class within each short-RNAseq (total) dataset. The measured intron is indicated by the red arrow.

A) ST2: This ORF is generating from a splicing event that uses the LT first exon donor and an acceptor within the ST ORF. In HPyV7 and BKPyV Dunlop, the splice lands in frame and results in an internal deletion within ST. In MPyV and MCPyV the splice lands out of frame, resulting in an ORF that contains the N-terminal region of ST and novel amino acids at the C terminus.

B) MT: MPyV encodes a MT following splicing connecting the end of the ST ORF with an ORF in an alternate frame of the LT second exon. In BKPyV, a similar splice occurs connecting ST with an MT-like ORF in an alternative frame of the LT second exon.

C) VP1X: JCPyV encodes two VP1X ORFs generated by splicing within VP1 and landing in an alternative frame of VP1, or earlier in the late region due to wraparound transcription. While predominant in JCPyV, VP1X is likewise present in many other PyV.

D) superT: The superT-specific splice utilizes the splice donor canonically associated with truncated T antigens such as 17kT in SV40 and truncT in BKPyV. Due to wraparound transcription, a LT second exon acceptor is available to the 3' of this donor and acts as the acceptor. For the superT ORF to form, an initial LT splice is required. Ultimately, superT contains a duplication in part of the LT second exon that includes the RB-binding LxCxE motif.

E. Schematics detailing BKPyV Dik isolates used for querying the existence of superT. BKPyV WT is wild type virus. M1 contains a LT intron that has been replaced with an intron from the plasmid pCI. Both WT and M1 are expected to generate LT and superT of expected sizes. M2 has a completely removed LT intron, and the pCI intron is located directly 5' of the LT ORF. M2 is expected to encode LT of expected size, but a larger superT variant due to incorporation of a second copy of the LT first exon.

F. Western blot of cells infected with BKPyV Dik WT, M1, or M2 and probed with an antibody reactive against LT. The lower molecular weight band is LT, and the higher molecular weight bands are consistent with superT.

Figure 6 - Detection superT-encoding transcripts in PyV-associated cancers

A. Schematic detailing the generation of superT during lytic infection as compared to from integrated virus in cancer. During viral infection, the RNA polymerase can circle the viral genome multiple times, resulting in a pre-mRNA that can be spliced to generate superT. In the case of host integration, a polyomavirus can be integrated in tandem copies such that a pre-mRNA is generated with more than one copy of the viral early region. This pre-mRNA can be similarly spliced to generate a superT transcript.

B. Heatmap indicating the abundance of the superT, ST, and LT introns from RNAseq data from two replicates of a BKPyV-positive bladder cancer and six MCPyV-associated MCCs. Percentages indicate the percentage of spliced early viral reads for each sample. The splice measured in each row is indicated by the red arrow in the schematics on the right side of the figure.

Figure S1 - Sequencing statistics

A. The number of reads for all datasets studied here. For long-read dRNAseq and SMRTseq, this number includes spliced and unspliced reads. Because short reads are only useful for transcript characterization when they span a splice junction, the counts for short-reads represent the number of splice-junction-spanning reads.

B. The cumulative percentage of transcripts in each number of transcript classes, by strand. The X-axis indicated the total number of transcript classes. The Y axis indicates the cumulative percentage of transcripts within those transcript classes. These plots indicate that most transcripts in most samples are contained within the first few transcript classes.

C-E. The alignment length distribution of early, late, spliced, and unspliced transcripts for dRNAseq and SMRTseq data from SV40 (**C**), BKPyV Dunlop (**D**), and MPyV (**E**). The X axis indicates the length of a read's alignment, while the Y axis indicates the density/percentage of transcripts with a given alignment length. This plot shows that dRNAseq and SMRTseq data sample from RNA populations of different length.

Figure S2 - Transcript start sites and polyA tail lengths.

A, B. The distribution of transcript start sites for late (**A**) and early (**B**) transcripts for SV40 (left column), BKPyV Dunlop (middle column), and MPyV (right column). The arrows indicate the viral ORF positions.

C. The distribution of polyA tail lengths for the 30-adenine ENO2 control (black), host (red), and viral (yellow) transcripts for SV40, BKPyV Dunlop, and MPyV. The X axis indicates the length of the polyA tail, while the Y axis indicates the density/percentage of transcripts with each length.

D. Ribosome occupancy of host transcripts in SV40-infected cells. Each grey dot is a host transcript. The red, blue, and black dots are specifically noted host transcripts. Ribosome occupancy is on the Y axis, while the X axis does not hold value. Lines on the violin plot indicate 1st, 2nd, and 3rd quartiles.

Figure S3 - SV40 transcriptome atlas, watch plots

A-C. Watch plots indicating all identified transcripts in SV40. (**A**) and (**B**) show transcripts that were identified in both dRNAseq and SMRTseq data, while (**C**) shows transcripts identified in SMRTseq only.

E. Barplots that show the abundance of each transcript type in the dRNAseq and SMRTseq data. Here, there are two dRNAseq bars (one per replicate). The Y axis indicates the percentage of transcripts of the same strand. As discussed in the methods, alignment of superT and superT* was challenging, so the actual abundance of these transcripts is higher than reported here.

Figure S4 - BKPyV Dunlop transcriptome atlas, watch plots

A-C. Watch plots indicating all identified transcripts in BKPyV Dunlop. (**A**) and (**B**) show transcripts that were identified in both dRNAseq and SMRTseq data, while (**C**) shows transcripts identified in dRNAseq only.

F. Barplots that show the abundance of each transcript type in the dRNAseq and SMRTseq data. The Y axis indicates the percentage of transcripts of the same strand. As discussed in the methods, alignment of superT and superT* was challenging, so the actual abundance of these transcripts is higher than reported here.

Figure S5 - SV40 transcriptome atlas, late transcript read pileups

A, B. Read pileups showing the late transcripts identified in SV40 dRNAseq (A) and SMRTseq (B). The arrows at the top of the plot indicate the viral ORF positions. Below the X axis, each row is an individual transcript, with exons indicated in red and splice junctions/introns indicated in pink. Above the X axis are histograms indicating the transcript start (blue) and transcript end (red) sites. (U: unspliced).

Figure S6 - SV40 transcriptome atlas, early transcript read pileups

A, B. Read pileups showing the early transcripts identified in SV40 dRNAseq (A) and SMRTseq (B). The arrows at the top of the plot indicate the viral ORF positions. Below the X axis, each row is an individual transcript, with exons indicated in red and splice junctions/introns indicated in pink. Above the X axis are histograms indicating the transcript start (blue) and transcript end (red) sites. (U: unspliced).

Figure S7 - BKPyV Dunlop transcriptome atlas, late transcript read pileups

A, B. Read pileups showing the late transcripts identified in BKPyV Dunlop dRNAseq (A) and SMRTseq (B). The arrows at the top of the plot indicate the viral ORF positions. Below the X axis, each row is an individual transcript, with exons indicated in red and splice junctions/introns indicated in pink. Above the X axis are histograms indicating the transcript start (blue) and transcript end (red) sites. (U: unspliced).

Figure S8 - BKPyV Dunlop transcriptome atlas, early transcript read pileups

A, B. Read pileups showing the early transcripts identified in BKPyV Dunlop dRNAseq (A) and SMRTseq (B). The arrows at the top of the plot indicate the viral ORF positions. Below the X axis, each row is an individual transcript, with exons indicated in red and splice junctions/introns indicated in pink. Above the X axis are histograms indicating the transcript start (blue) and transcript end (red) sites. (U: unspliced).

Figure S9 - Intron plots for all datasets studied

A. Intron plots generated from short-read RNAseq. The arrows at the top and bottom of each panel indicate the position of viral ORFs. The lines indicate specific introns identified in the RNAseq data, with the 5' end on the top and the 3' end on the bottom. The blue color indicates late transcripts, with red indicating early transcripts. The size of the circles above and below the viral ORF maps indicate the percentage of junction-spanning reads with a 5' end (on top) or 3' end (on bottom) at that position. Junctions are plotted if they are at least 1% of early or late transcripts, except for the SV40 pA superT junction (transcript class 3) which is just below threshold but is of interest.

B. Another representation of intron plots for each virus. The top arrows indicate the position of viral ORFs. The X axis indicates the genomic position for each splice. The Y axis indicates a single transcript class, with that class' intron plotted as a line. The percentage of early or late transcripts is indicated with the numeric percentage. Junctions are plotted if they are at least 1% of early or late transcripts, except for the SV40 pA superT junction (transcript class 3) which is just below threshold but is of interest.

Figure S10 - Alternative polyadenylation of early transcripts in SV40, BKPyV, and MPyV.

A-C. Watch plots indicating the LT and ST transcripts for SV40 (**A**), BKPyV Dunlop (**B**), and MPyV (**C**). The focus of these plots is the distribution of transcript end positions, which are the inner red arms. The region of APA of highlighted in blue, with the canonical transcript end sites highlighted in red.

D. A cumulative incidence plot of transcript termination in SV40 (blue), BKPyV Dunlop (red), and MPyV (green). The X axis indicates the distance to the canonical polyA site, while the Y axis indicates the percentage of transcripts that have terminated by that position.

E-G. Density plots showing the distribution of polyA tail lengths for LT and ST transcripts that end at the canonical site (solid) or undergo APA (dashed) for SV40 (**E**), BKPyV Dunlop (**F**), and MPyV (**G**). The x axis indicates the length of the polyA tail, while the Y axis indicates the density/proportion of transcripts with the given length.

Figure S11 - short-RNAseq (polyA) analysis of BKPyV Dik WT, M1, and M2

A-C. Intron plots generated from short-read (polyA) RNAseq of cells infected with BKPyV WT, or the M1 or M2 mutants. The arrows at the top and bottom of each panel indicate the position of viral ORFs relative to the standard BKPyV genome - note that the genomes of mutants M1 and M2 are altered as indicated in Figure 5E. The lines indicate specific introns identified in the RNAseq data, with the 5' end on the top and the 3' end on the bottom. The size of the circles above and below the viral ORF maps indicate the percentage of junction-spanning reads with a 5' end (on top) or 3' end (on bottom) at that position. Only early junctions that are at least 1% of early early transcripts are plotted. The superT junction is colored in gold. (**A**) Intron plot for BKPyV Dik WT. (**B**) Intron plot for BKPyV Dik M1. (**C**) Intron plot for BKPyV Dik M2.

Figure S12 - superT in MCPyV-associated MCC

A. Sanger sequencing of an RT-PCR product from MCC J45_440, showing the superT-specific junction.
B. A schematic detailing the MCC 285 MCPyV integration site, showing how it is possible that superT is generated via cis-splicing.
C. The assembled viral block in MCC tumor J45_440. This integration site is based on de-novo assembly using short whole genome sequencing reads. Despite only assembling one viral block, we found that 1) there are likely 2 copies of the viral genome, and 2) the 5’ viral integration site appears to fall on chromosome 7 “after” the 3’ viral integration site, observations consistent with the existence of two copies of the viral genome in tandem separated by a small segment of host DNA at this integration site.
D. The assembled viral blocks in MCC tumor J17_296. The longest block contains two copies of the early region.
E. Lollipop plots showing identified SNPs in the MCPyV genomes of J45_440, J17_296, and J11_285. The gene-map below the figure indicates the position of viral ORFs. Each lollipop is colored according to the nucleotide substitution identified.

Abend, J.R., Joseph, A.E., Das, D., Campbell-Cecen, D.B., and Imperiale, M.J. (2009). A truncated T antigen expressed from an alternatively spliced BK virus early mRNA. The Journal of general virology *90*, 1238.

Abere, B., Zhou, H., Li, J., Cao, S., Toptan, T., Grundhoff, A., Fischer, N., Moore, P.S., and Chang, Y. (2020). Merkel Cell Polyomavirus Encodes Circular RNAs (circRNAs) Enabling a Dynamic circRNA/microRNA/mRNA Regulatory Network. Mbio *11*, e03059-03020.

Adami, G., Marlor, C., Barrett, N., and Carmichael, G.G. (1989). Leader-to-leader splicing is required for efficient production and accumulation of polyomavirus late mRNAs. Journal of virology *63*, 85-93.

Assetta, B., De Cecco, M., O'Hara, B., and Atwood, W.J. (2016). JC polyomavirus infection of primary human renal epithelial cells is controlled by a type I IFN-induced response. MBio *7*, e00903-00916.

Balázs, Z., Tombácz, D., Szűcs, A., Snyder, M., and Boldogkői, Z. (2017). Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. Scientific data *4*, 1-7.

Carter, J.J., Daugherty, M.D., Qi, X., Bheda-Malge, A., Wipf, G.C., Robinson, K., Roman, A., Malik, H.S., and Galloway, D.A. (2013). Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. Proceedings of the National Academy of Sciences *110*, 12744-12749.

Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., Mohr, I., and Wilson, A.C. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nature communications *10*, 1-13.

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

Eul, J., and Patzel, V. (2013). Homologous SV40 RNA trans-splicing: a new mechanism for diversification of viral sequences and phenotypes. RNA biology *10*, 1689-1699.

Freund, R., Sotnikov, A., Bronson, R.T., and Benjamin, T.L. (1992). Polyoma virus middle T is essential for virus replication and persistence as well as for tumor induction in mice. Virology *191*, 716-723.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., and Warland, A. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nature methods *15*, 201-206.

Garren, S.B., Kondaveeti, Y., Duff, M.O., and Carmichael, G.G. (2015). Global analysis of mouse polyomavirus infection reveals dynamic regulation of viral and host gene expression and promiscuous viral RNA editing. PLoS pathogens *11*, e1005166.

Ghosh, P., Reddy, V., Swinscoe, J., Lebowitz, P., and Weissman, S. (1978). Heterogeneity and 5′-terminal structures of the late RNAs of simian virus 40. Journal of molecular biology *126*, 813-846.

Good, P.J., Welch, R.C., Ryu, W.-S., and Mertz, J.E. (1988). The late spliced 19S and 16S RNAs of simian virus 40 can be synthesized from a common pool of transcripts. Journal of virology *62*, 563-571.

Gottlieb, K.A., and Villarreal, L.P. (2001). Natural biology of polyomavirus middle T antigen. Microbiology and molecular biology reviews *65*, 288-318.

Jiang, M., Abend, J.R., Johnson, S.F., and Imperiale, M.J. (2009a). The role of polyomaviruses in human disease. Virology *384*, 266-273.

Jiang, M., Abend, J.R., Tsai, B., and Imperiale, M.J. (2009b). Early events during BK virus entry and disassembly. Journal of virology *83*, 1350-1358.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic acids research *36*, W5-W9.

Kamen, R., Favaloro, J., and Parker, J. (1980a). Topography of the three late mRNA's of polyoma virus which encode the virion proteins. Journal of virology *33*, 637-651.

Kamen, R., Favaloro, J., Parker, J., Treisman, R., Lania, L., Fried, M., and Mellor, A. (1980b). Comparison of polyoma virus transcription in productively infected mouse cells and transformed rodent cell lines. Paper presented at: Cold Spring Harbor symposia on quantitative biology (Cold Spring Harbor Laboratory Press).

Keller, M.W., Rambo-Martin, B.L., Wilson, M.M., Ridenour, C.A., Shepard, S.S., Stark, T.J., Neuhaus, E.B., Dugan, V.G., Wentworth, D.E., and Barnes, J.R. (2018). Direct RNA sequencing of the coding complete influenza A virus genome. Scientific reports *8*, 1-8.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., and Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. Cell *181*, 914-921. e910.

Kress, M., May, E., Cassingena, R., and May, P. (1979). Simian virus 40-transformed cells express new species of proteins precipitable by anti-simian virus 40 tumor serum. Journal of virology *31*, 472-483.

Krueger, F. (2016). Babraham Bioinformatics-Trim Galore.

Lee, S., Micalizzi, D., Truesdell, S.S., Bukhari, S.I., Boukhali, M., Lombardi-Story, J., Kato, Y., Choo, M.-K., Dey-Guha, I., and Ji, F. (2020). A post-transcriptional program of chemoresistance by AU-rich elements and TTP in quiescent leukemic cells. Genome biology *21*, 1-23.

Legon, S., Flavell, A.J., Cowie, A., and Kamen, R. (1979). Amplification in the leader sequence of late polyoma virus mRNAs. Cell *16*, 373-388.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100.

Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature methods *12*, 733-735.

Luo, Y., and Carmichael, G.G. (1991). Splice site skipping in polyomavirus late pre-mRNA processing. Journal of virology *65*, 6637-6644.

Nguyen, K.D., Lee, E.E., Yue, Y., Stork, J., Pock, L., North, J.P., Vandergriff, T., Cockerell, C., Hosler, G.A., and Pastrana, D.V. (2017). Human polyomavirus 6 and 7 are associated with pruritic and dyskeratotic dermatoses. Journal of the American Academy of Dermatology *76*, 932-940. e933.

Nomburg, J., Meyerson, M., and DeCaprio, J.A. (2020). Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. Genome medicine *12*, 1-14.

Norbury, C.J., and Fried, M. (1987). Polyomavirus early region alternative poly (A) site: 3'-end heterogeneity and altered splicing pattern. Journal of virology *61*, 3754-3758.

Price, A.M., Hayer, K.E., McIntyre, A.B., Gokhale, N.S., Abebe, J.S., Della Fera, A.N., Mason, C.E., Horner, S.M., Wilson, A.C., and Depledge, D.P. (2020). Direct RNA sequencing reveals m 6 A modifications on adenovirus RNA are necessary for efficient splicing. Nature communications *11*, 1-17.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Reddy, V.B., Ghosh, P.K., Lebowitz, P., and Sherman M, W. (1978). Gaps and duplicated sequences in the leaders of SV40 16S RNA. Nucleic acids research *5*, 4195-4214.

Riley, M.I., Yoo, W., Mda, N.Y., and Folk, W.R. (1997). Tiny T antigen: an autonomous polyomavirus T antigen amino-terminal domain. Journal of virology *71*, 6068-6074.

Rosenstein, R.K., Pastrana, D.V., Starrett, G.J., Sapio, M.R., Hill, N.T., Jo, J.-H., Lee, C.-C.R., Iadarola, M.J., Buck, C.B., and Kong, H.H. (2021). Host-Pathogen Interactions in Human Polyomavirus 7–Associated Pruritic Skin Eruption. The Journal of investigative dermatology *141*, 1344-1348. e1348.

Saribas, A.S., DeVoto, J., Golla, A., Wollebo, H.S., White, M.K., and Safak, M. (2018). Discovery and characterization of novel trans-spliced products of human polyoma JC virus late transcripts from PML patients. Journal of cellular physiology *233*, 4137-4155.

Schmidlin, K., Kraberger, S., Cook, C., DeNardo, D.F., Fontenele, R.S., Van Doorslaer, K., Martin, D.P., Buck, C.B., and Varsani, A. (2021). A novel lineage of polyomaviruses identified in bark scorpions. bioRxiv.

Seif, I., Khoury, G., and Dhar, R. (1979). The genome of human papovavirus BKV. Cell *18*, 963-977.

Shuda, M., Feng, H., Kwun, H.J., Rosen, S.T., Gjoerup, O., Moore, P.S., and Chang, Y. (2008). T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. Proceedings of the National Academy of Sciences *105*, 16272-16277.

Smith, A.E., Smith, R., and Paucha, E. (1979). Characterization of different tumor antigens present in cells transformed by simian virus 40. Cell *18*, 335-346.

Starrett, G.J., Thakuria, M., Chen, T., Marcelus, C., Cheng, J., Nomburg, J., Thorner, A.R., Slevin, M.K., Powers, W., and Burns, R.T. (2020). Clinical and molecular characterization of virus-positive and virus-negative Merkel cell carcinoma. Genome medicine *12*, 1-22.

Starrett, G.J., Yu, K., Golubeva, Y., Lenz, P., Piaskowski, M.L., Peterson, D., Dean, M., Israni, A., Hernandez, B.Y., Tucker, T.C.*, et al.* (2021). Common Mechanisms of Virus-Mediated Oncogenesis in Bladder Cancers Arising In Solid Organ Transplant Recipients. medRxiv, 2021.2011.2011.21266080.

Theiss, J.M., Günther, T., Alawi, M., Neumann, F., Tessmer, U., Fischer, N., and Grundhoff, A. (2015). A comprehensive analysis of replicating Merkel cell polyomavirus genomes delineates the viral transcription program and suggests a role for mcv-miR-M1 in episomal persistence. PLoS pathogens *11*, e1004974.

Treisman, R. (1980). Characterisation of polyoma late mRNA leader sequences by molecular cloning and DNA sequence analysis. Nucleic acids research *8*, 4867-4888.

Tremblay, J.D., Sachsenmeier, K.F., and Pipas, J.M. (2001). Propagation of wild-type and mutant SV40. In SV40 Protocols (Springer), pp. 1-7.

Trowbridge, P.W., and Frisque, R.J. (1995). Identification of three new JC virus proteins generated by alternative splicing of the early viral mRNA. Journal of neurovirology *1*, 195-206.

Yang, R., Lee, E.E., Kim, J., Choi, J.H., Kolitz, E., Chen, Y., Crewe, C., Salisbury, N.J., Scherer, P.E., and Cockerell, C. (2021). Characterization of ALTO-encoding circular RNAs expressed by Merkel cell polyomavirus and trichodysplasia spinulosa polyomavirus. PLoS pathogens *17*, e1009582.

Zerrahn, J., Knippschild, U., Winkler, T., and Deppert, W. (1993). Independent expression of the transforming amino-terminal domain of SV40 large I antigen from an alternatively spliced third SV40 early mRNA. The EMBO journal *12*, 4739-4746.

Zhao, L., and Imperiale, M.J. (2019). Establishing Renal Proximal Tubule Epithelial-Derived Cell Lines Expressing Human Telomerase Reverse Transcriptase for Studying BK Polyomavirus. Microbiology resource announcements *8*, e01129-01119.

Zou, W., Vue, G.S., Assetta, B., Manza, H., Atwood, W.J., and Imperiale, M.J. (2020). Control of archetype BK polyomavirus microRNA expression. Journal of Virology *95*, e01589-01520.

Figure 1

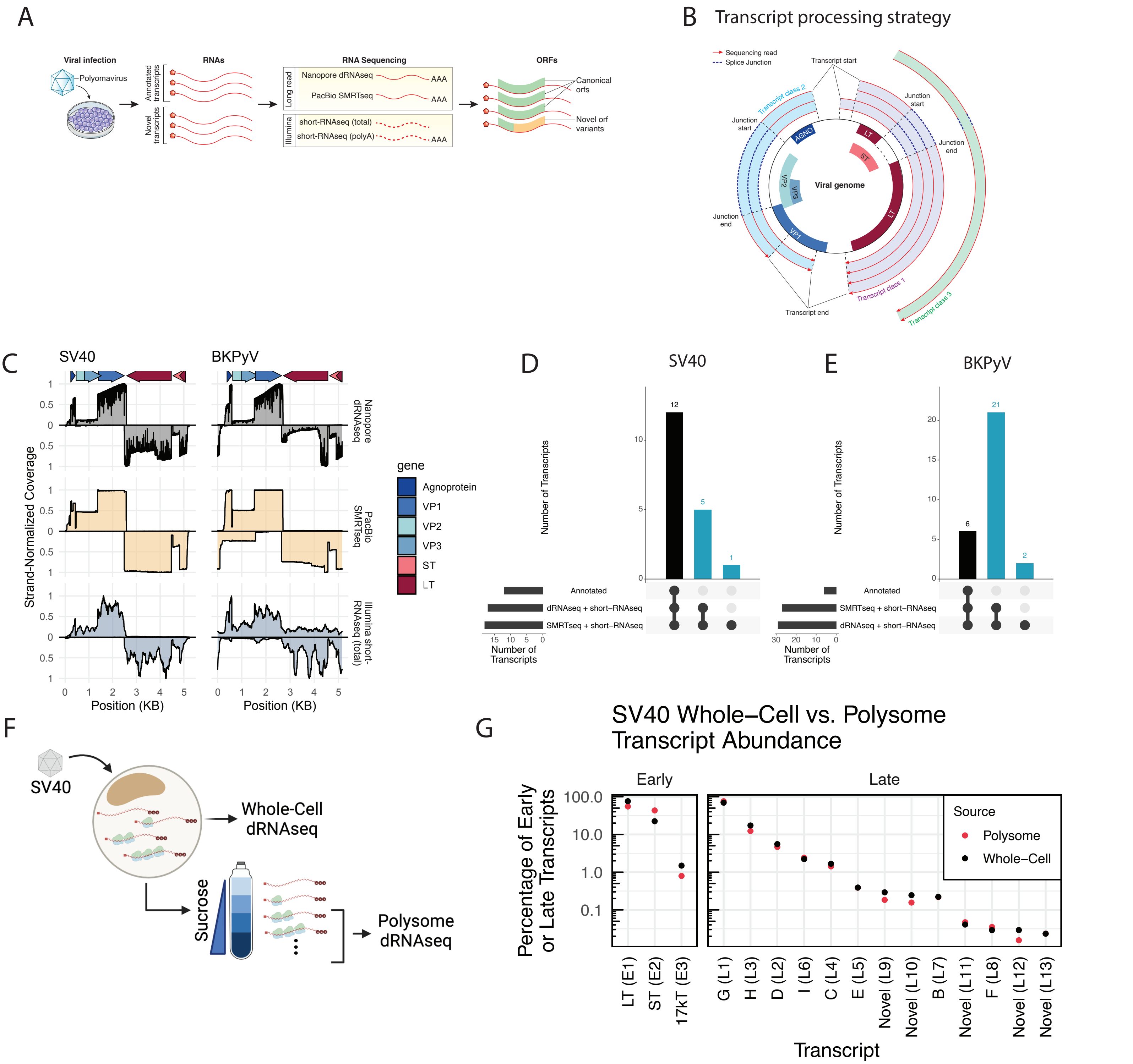


Figure 1 - RNA sequencing expands known SV40 and BKPyV transcript diversity.

A. Overview of experimental procedures. Cells were infected with a polyomavirus, and RNAs extracted. RNA was sequenced using long-read (Nanopore dRNAseq and PacBio SMRTseq) and short-read (Illumina short-RNAseq (total) and short-RNAseq (polyA)). Transcripts were analyzed, and the impact of observed splice events on viral open reading frames was assessed.

B. Mechanism of transcript clustering in this study. Transcripts were aligned to the viral genome and grouped into transcript classes based on the presence of shared introns. Thus, within a transcript class there may be variation in the exact transcript start and end positions. This clustering strategy was used for both long- and short-RNAseq data.

C. Viral RNA sequence coverage for SV40 and BKPyV as determined from dRNAseq, SMRTseq, and short-RNAseq (total) data. The Y axis indicates the scaled coverage, with X axis indicating the position on the viral genome. Coverage for late transcripts (mapping to the + strand) is above the x axis, while coverage for early transcripts (mapping to the - strand) is below the x axis. Coverage is scaled separately for each strand such that the maximum observed coverage for each strand is 1. Arrows at the top of the plot indicate the positions of viral genes.

D-E. UpSet plot indicating the overlap between existing transcript annotations, dRNAseq data, and SMRTseq data for SV40 (D) and BKPyV Dunlop (E). Bars indicating overlap with existing transcript annotations are black, while those indicating no overlap with existing annotations are blue. These blue bars indicate the number of novel, unannotated transcripts identified.

F. Overview of polysome profiling of SV40-infected cells. BSC40 cells were infected with SV40. Cells were lysed, and a portion of the lysate was subjected to dRNAseq (representative of the RNA content of the whole cell). The remaining lysates was centrifuged through a sucrose gradient, after which fractions containing RNA associated with two or more ribosomes were pooled and subjected to dRNAseq.

G. Relative abundance of SV40 early and late transcripts in the whole-cell and polysome fractions of SV40-infected cells. Y-axis indicates the percentage of early or late transcripts and is log scale. X axis indicates each transcript, with black dots indicating each transcript's whole-cell relative abundance and red dots indicating each transcript's polysome relative abundance.

Figure 2

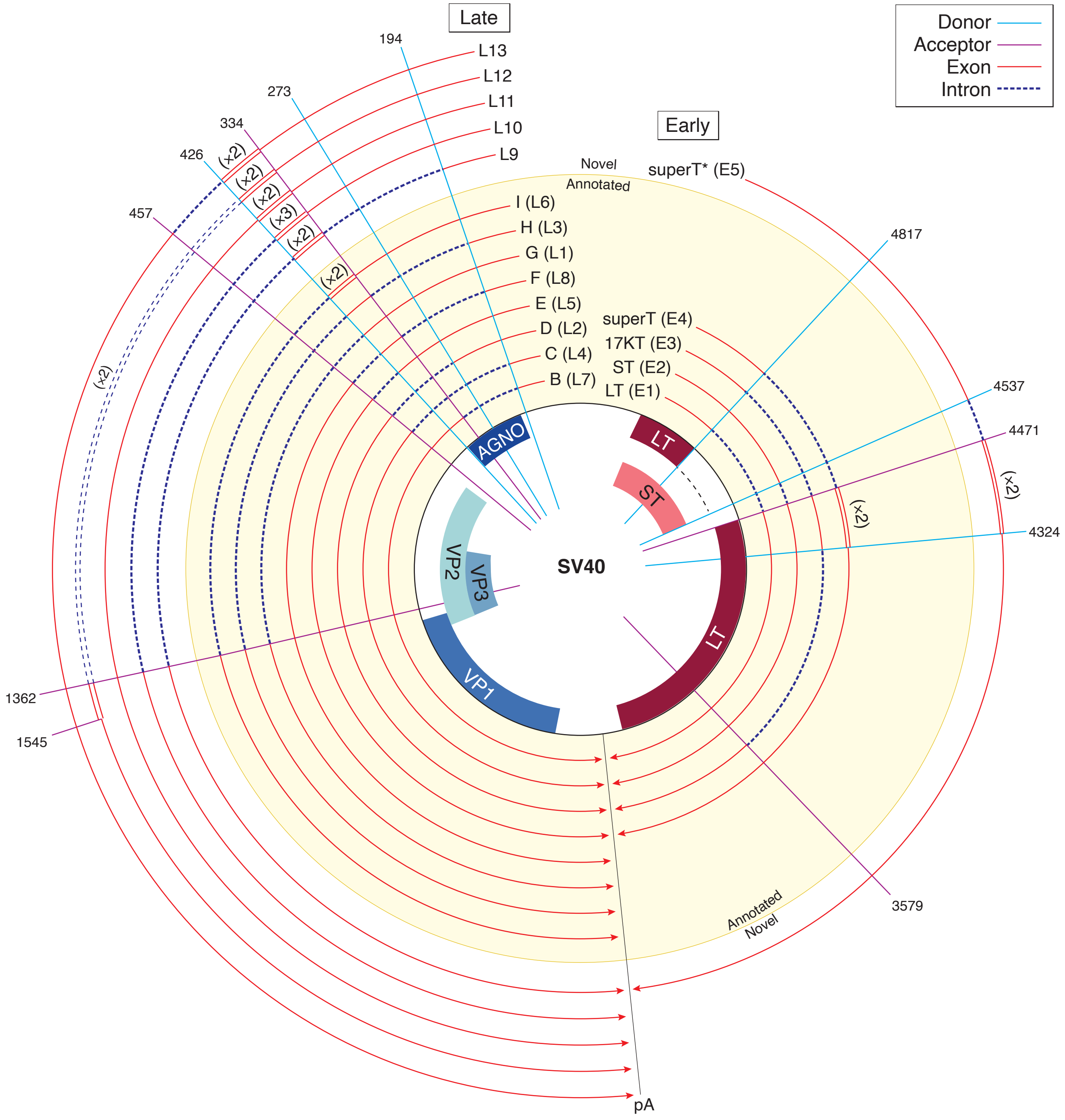


Figure 2 - Annotated and novel SV40 transcripts.

A. Transcripts are shown relative to the viral genome. Each line is a viral transcript, with red lines indicating exons and dashed blue lines indicating introns. Spokes indicate the positions of common splice donors and splice acceptors. Transcripts that were annotated prior to this study are on a yellow background, and novel transcripts are on a while background. Wraparound transcription that results in multiple copies of a region is annotated with double lines, and the number of copies is indicated in parentheses. The line labeled "pA" indicates the approximate position of the polyA signal sequence.

Figure 3

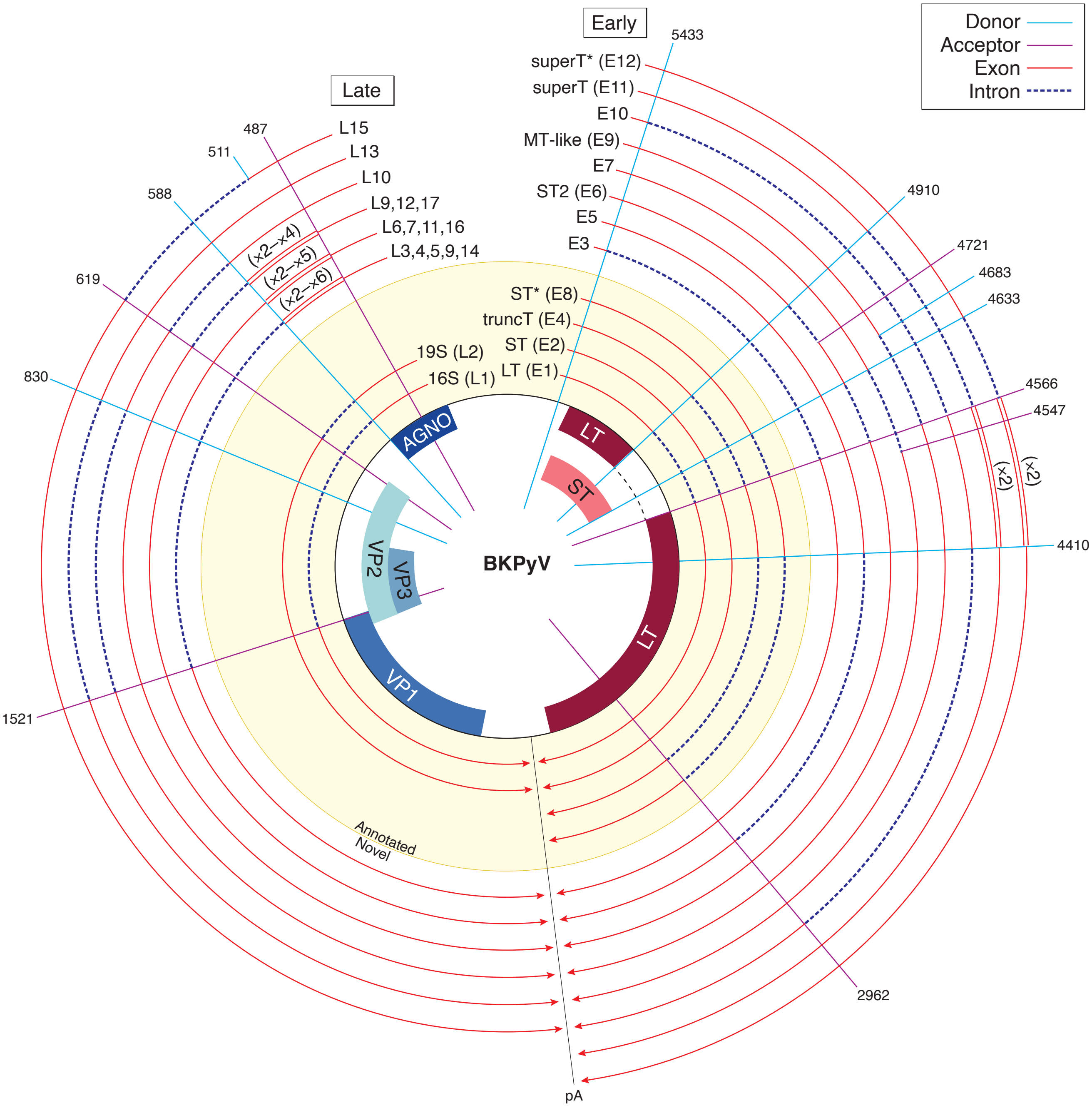


Figure 3 - Annotated and novel BKPyV transcripts.

A. Transcripts are shown relative to the viral genome. Each line is a viral transcript, with red lines indicating exons and dashed blue lines indicating introns. Spokes indicate the positions of common splice donors and splice acceptors. Transcripts that were annotated prior to this study are on a yellow background, and novel transcripts are on a while background. Wraparound transcription that results in multiple copies of a region is annotated with double lines, and the number of copies is indicated in parentheses. The line labeled "pA" indicates the approximate position of the polyA signal sequence.

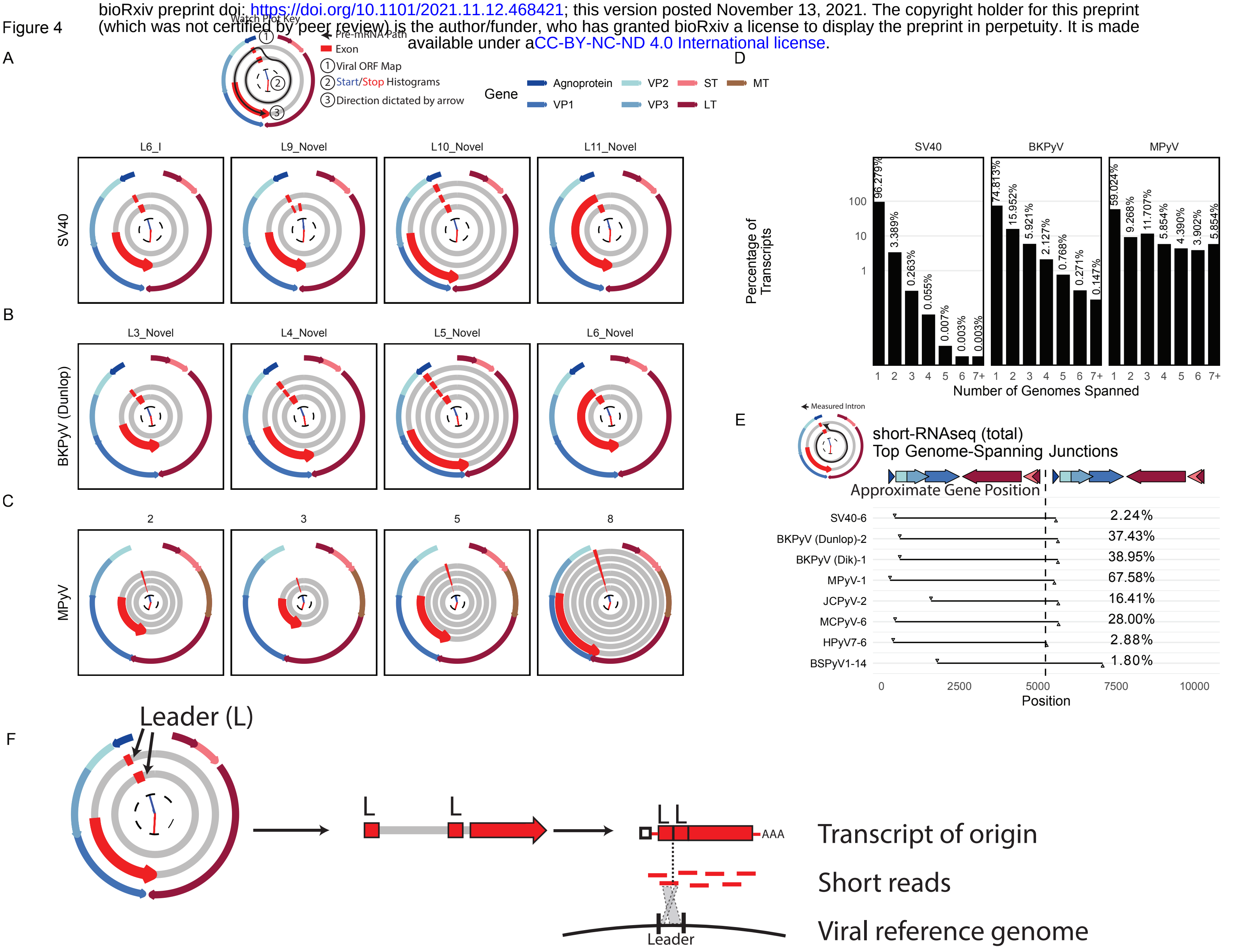


Figure 4 - Pervasive wraparound transcription across PyV

A-C. Watch plots indicating the top 4 highest abundance late wraparound transcript classes in dRNAseq data from SV40 (A), BKPyV Dunlop (B), and MPyV (C). The outer ring of each watch plot indicates the position of the viral ORFs. The inner arms are histograms detailing the distribution of transcript starts (in blue) and ends (in red) for transcripts within each transcript class. The red segments indicate exons. Transcripts start in the innermost ring - a second or third ring indicates that the pre-mRNA that generated the transcript must have circled the viral genome multiple times. The 3′ end of the transcript and the direction in which these plots are oriented is indicated by the red arrow at the end of the last exon segment. The red exon segments start at the most common transcript start site within the transcript class, and end at the most common transcript end site within the class. The watch plot key shows an example of the path of the pre-mRNA for SV40 transcript class L6_I.

D. Bar plots indicating the percentage of late transcripts that span a given number of genome lengths in SV40, BKPyV Dunlop, and MPyV dRNAseq data.

E. The leader-leader junction, that connects the pre-mRNA from one genome to the subsequent wraparound, was identified in Illumina short-RNAseq (total) data. The intron in question is plotted as a black line in this plot, with the x axis indicating the genomic position of the intron. The top late wraparound transcript for each virus was plotted. The gene map indicates the approximate gene position and is accurate for SV40 - the exact position of the viral genes varies between viruses. Percentages indicate the percentage of late junction-spanning transcripts that support the plotted wraparound leader-leader junction.

F. Schematic illustrating how leader-leader wraparound transcription can be detected from short read short-RNAseq (total). Leader-leader splicing can be seen as a repetitive exon in watch plots from long-read RNAseq data. Ultimately, there was an original processed mRNA in the cell that contained two tandem leader sequences. When this transcript of origin is sequenced via short read sequencing, reads will be generated across its length. A minority of these reads will span the leader-leader junction, and mapping against the viral reference genome can be used to uncover leader-leader splicing.

Figure 5

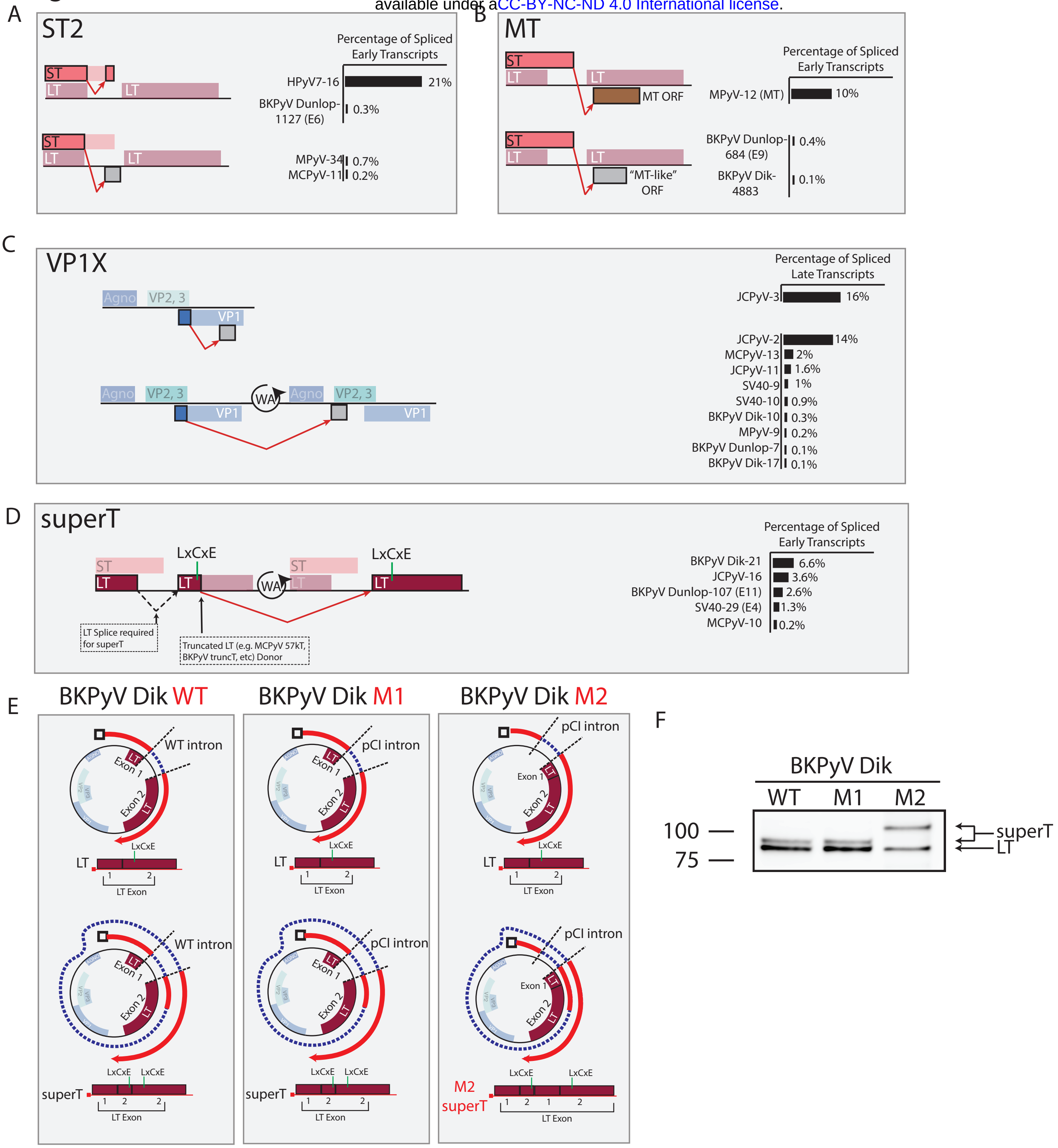


Figure 5 - Detection of novel, conserved splicing events that expand PyV coding capacity.

A-D. Schematics illustrating identified ORFs. Each row is a reading frame (except for ST and the LT 1st exon, which are in the same frame), and unannotated amino acids are represented by grey boxes. The measured intron is indicated by the red arrow. Colored ORFs are annotated, while grey ORFs are unannotated. Percentages on the right side of the figure are the percentage of spliced viral transcripts on the same strand as determined from short-read short-RNAseq (total) data. Numbers after each virus name indicate the transcript class within each short-RNAseq (total) dataset. The measured intron is indicated by the red arrow.

A) ST2: This ORF is generating from a splicing event that uses the LT first exon donor and an acceptor within the ST ORF. In HPyV7 and BKPyV Dunlop, the splice lands in frame and results in an internal deletion within ST. In MPyV and MCPyV the splice lands out of frame, resulting in an ORF that contains the N-terminal region of ST and novel amino acids at the C terminus.

B) MT: MPyV encodes a MT following splicing connecting the end of the ST ORF with an ORF in an alternate frame of the LT second exon. In BKPyV, a similar splice occurs connecting ST with an MT-like ORF in an alternative frame of the LT second exon.

C) VP1X: JCPyV encodes two VP1X ORFs generated by splicing within VP1 and landing in an alternative frame of VP1, or earlier in the late region due to wraparound transcription. While predominant in JCPyV, VP1X is likewise present in many other PyV.

D) superT: The superT-specific splice utilizes the splice donor canonically associated with truncated T antigens such as 17kT in SV40 and truncT in BKPyV. Due to wraparound transcription, a LT second exon acceptor is available to the 3′ of this donor and acts as the acceptor. For the superT ORF to form, an initial LT splice is required. Ultimately, superT contains a duplication in part of the LT second exon that includes the RB-binding LxCxE motif.

E. Schematics detailing BKPyV Dik isolates used for querying the existence of superT. BKPyV WT is wild type virus. M1 contains a LT intron that has been replaced with an intron from the plasmid pCI Both WT and M1 are expected to generate LT and superT of expected sizes. M2 has a completely removed LT intron, and the pCI intron is located directly 5′ of the LT ORF. M2 is expected to encode LT of expected size, but a larger superT variant due to incorporation of a second copy of the LT first exon.

F. Western blot of cells infected with BKPyV Dik WT, M1, or M2 and probed with an antibody reactive against LT. The lower molecular weight band is LT, and the higher molecular weight bands are consistent with superT.

Figure 6

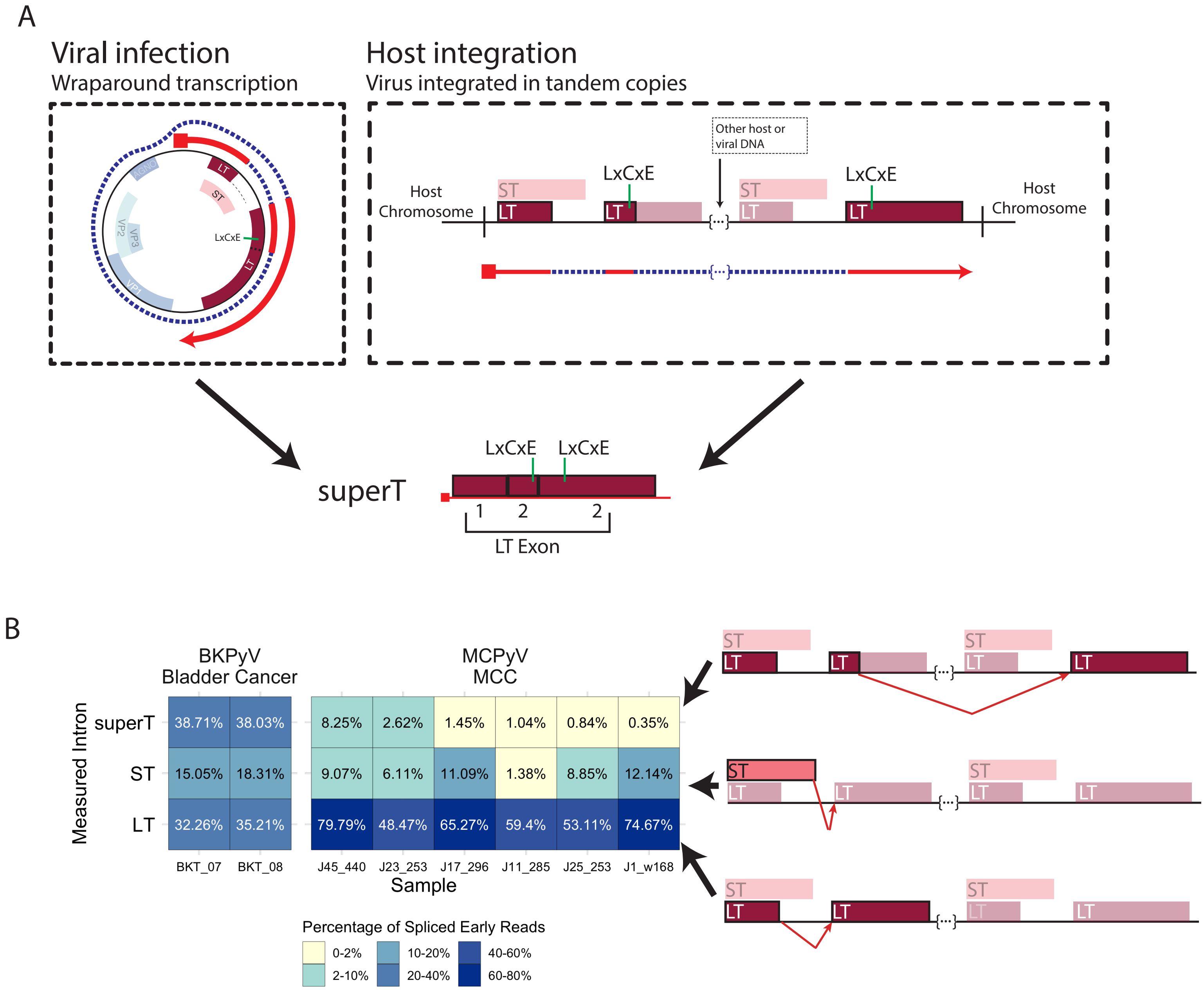


Figure 6 - Detection superT-encoding transcripts in PyV-associated cancers

A. Schematic detailing the generation of superT during lytic infection as compared to from integrated virus in cancer. During viral infection, the RNA polymerase can circle the viral genome multiple times, resulting in a pre-mRNA that can be spliced to generate superT. In the case of host integration, a polyomavirus can be integrated in tandem copies such that a pre-mRNA is generated with more than one copy of the viral early region. This pre-mRNA can be similarly spliced to generate a superT transcript.
B. Heatmap indicating the abundance of the superT, ST, and LT introns from RNAseq data from two replicates of a BKPyV-positive bladder cancer and six MCPyV-associated MCCs. Percentages indicate the percentage of spliced early viral reads for each sample. The splice measured in each row is indicated by the red arrow in the schematics on the right side of the figure.

Fig S1

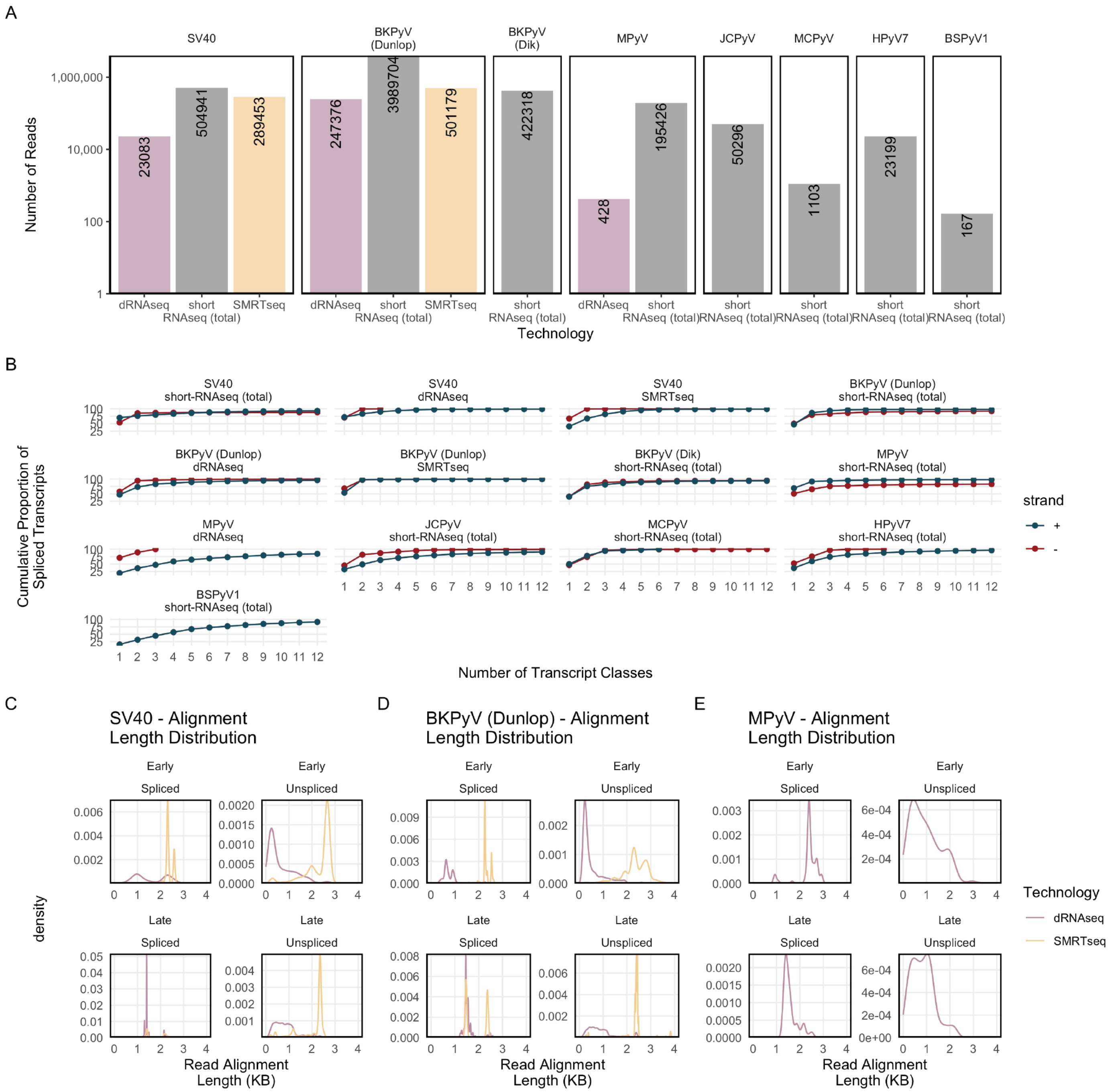
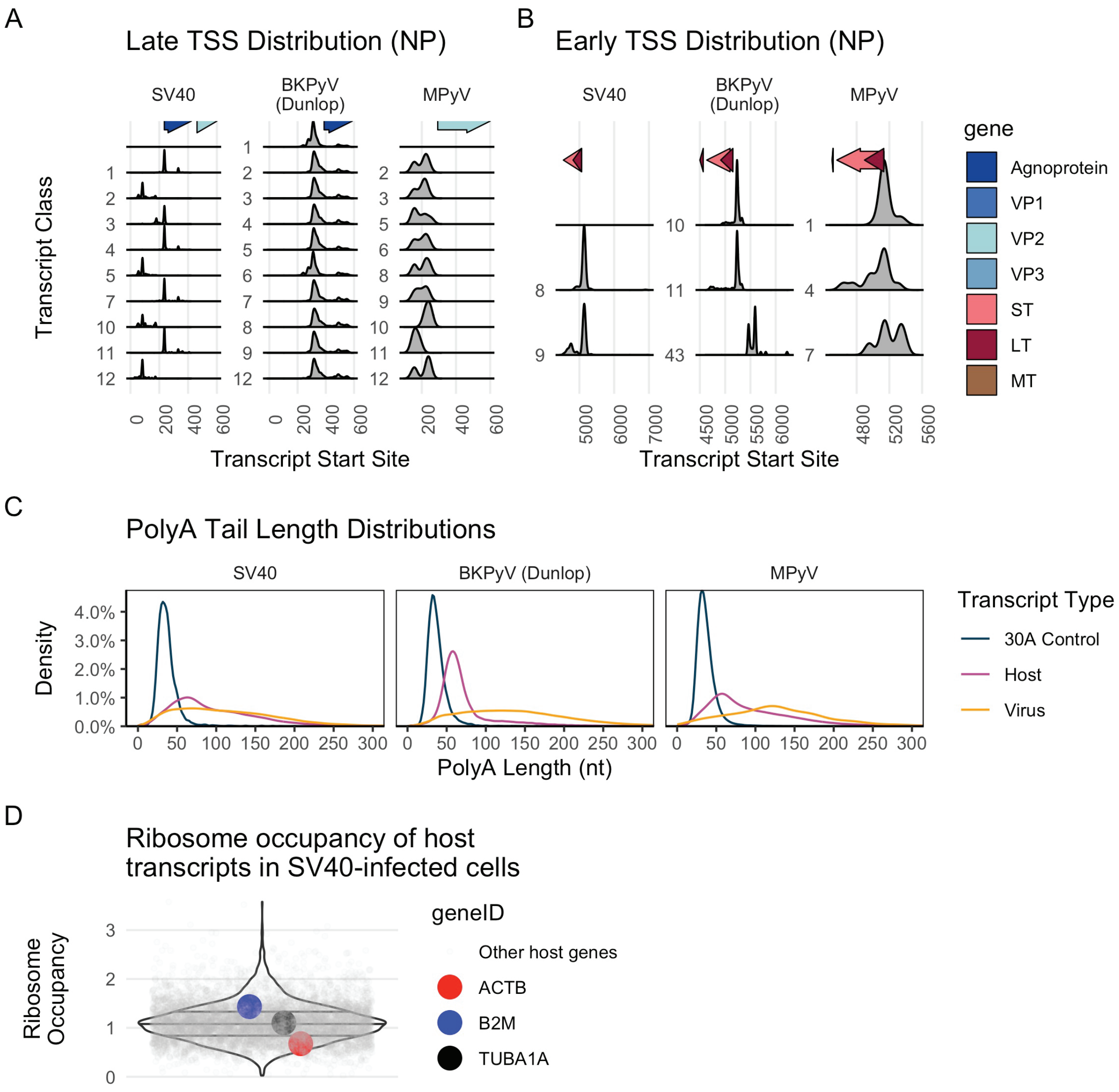
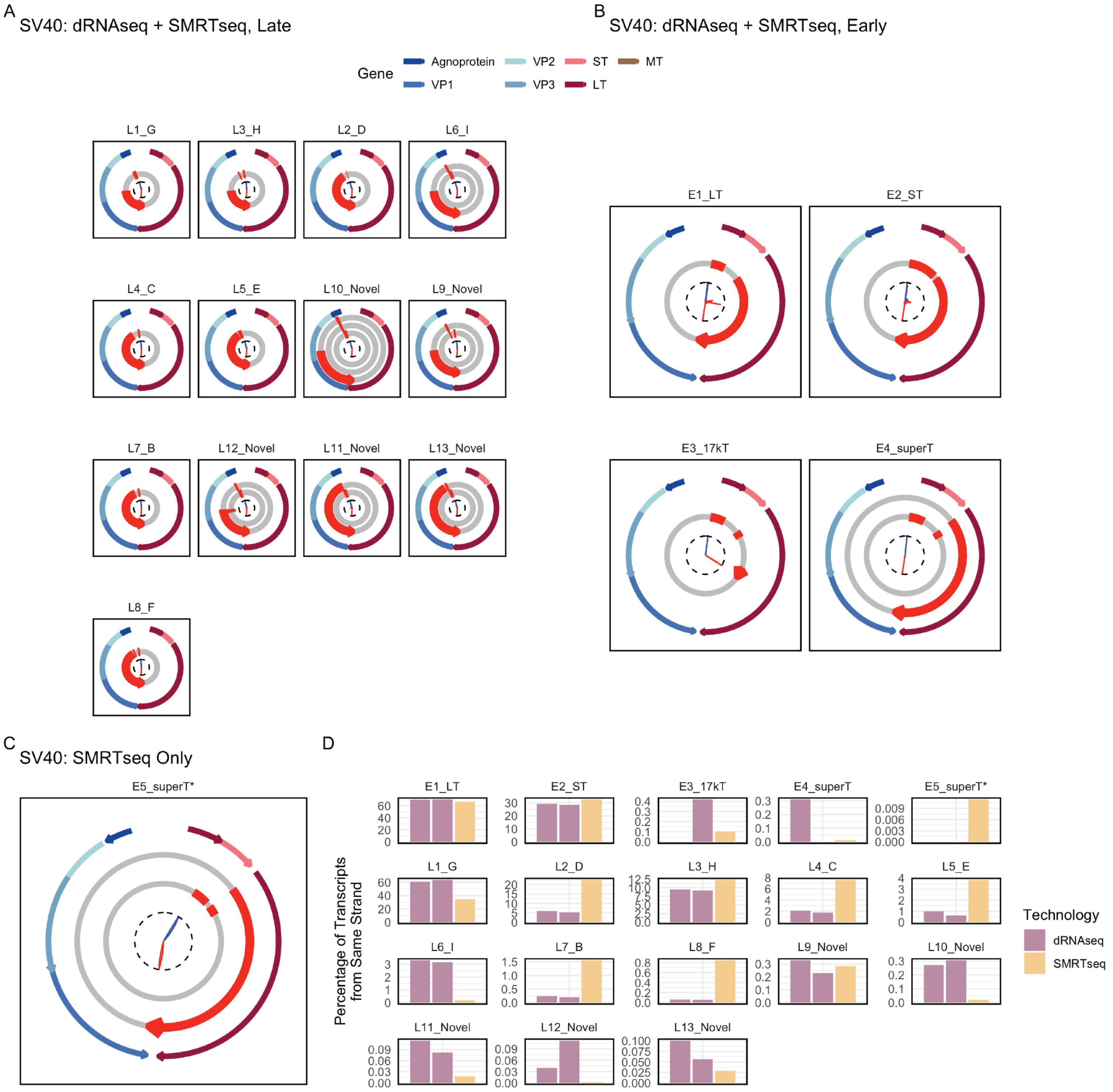

A
SV40
BKPyV (Dunlop)
BKPyV (Dik)
MPyV
JCPyV
MCPyV
HPyV7
BSPyV1
Number of Reads
1,000,000
10,000
100
1
23083
504941
289453
247376
3989704
501179
422318
428
195426
50296
1103
23199
167
dRNAseq
short RNAseq (total)
SMRTseq
Technology
B
SV40 short-RNAseq (total)
SV40 dRNAseq
SV40 SMRTseq
BKPyV (Dunlop) short-RNAseq (total)
BKPyV (Dunlop) dRNAseq
BKPyV (Dunlop) SMRTseq
BKPyV (Dik) short-RNAseq (total)
MPyV short-RNAseq (total)
MPyV dRNAseq
JCPyV short-RNAseq (total)
MCPyV short-RNAseq (total)
HPyV7 short-RNAseq (total)
BSPyV1 short-RNAseq (total)
Cumulative Proportion of Spliced Transcripts
100
75
50
25
1 2 3 4 5 6 7 8 9 10 11 12
Number of Transcript Classes
strand
+
-
C
SV40 - Alignment Length Distribution
D
BKPyV (Dunlop) - Alignment Length Distribution
E
MPyV - Alignment Length Distribution
Early Spliced
Early Unspliced
Late Spliced
Late Unspliced
density
0 1 2 3 4
Read Alignment Length (KB)
Technology
dRNAseq
SMRTseq

Figure S2


A
Late TSS Distribution (NP)
SV40
BKPyV (Dunlop)
MPyV
Transcript Class
1
2
3
4
5
7
10
11
12
1
2
3
4
5
6
7
8
9
12
2
3
5
6
8
9
10
11
12
0
200
400
600
Transcript Start Site
B
Early TSS Distribution (NP)
SV40
BKPyV (Dunlop)
MPyV
8
9
10
11
43
1
4
7
5000
6000
7000
4500
5000
5500
6000
4800
5200
5600
Transcript Start Site
gene
Agnoprotein
VP1
VP2
VP3
ST
LT
MT
C
PolyA Tail Length Distributions
SV40
BKPyV (Dunlop)
MPyV
Density
4.0%
3.0%
2.0%
1.0%
0.0%
0
50
100
150
200
250
300
PolyA Length (nt)
Transcript Type
30A Control
Host
Virus
D
Ribosome occupancy of host transcripts in SV40-infected cells
Ribosome Occupancy
3
2
1
0
geneID
Other host genes
ACTB
B2M
TUBA1A

FigS3: SV40 Transcriptome



FigS4: BKPyV (Dunlop) Transcriptome

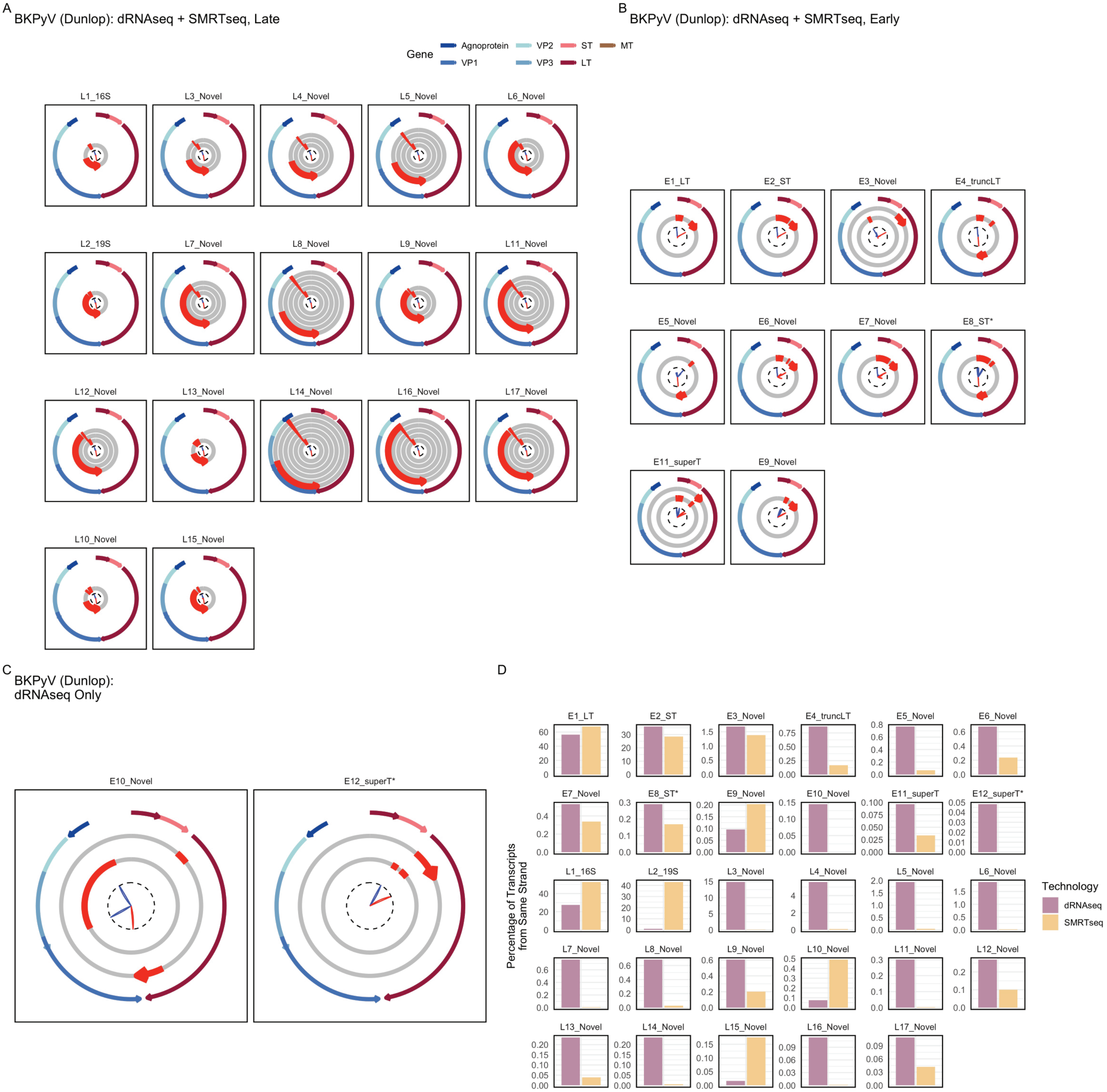


Fig S5: SV40 Late Transcript Pileups

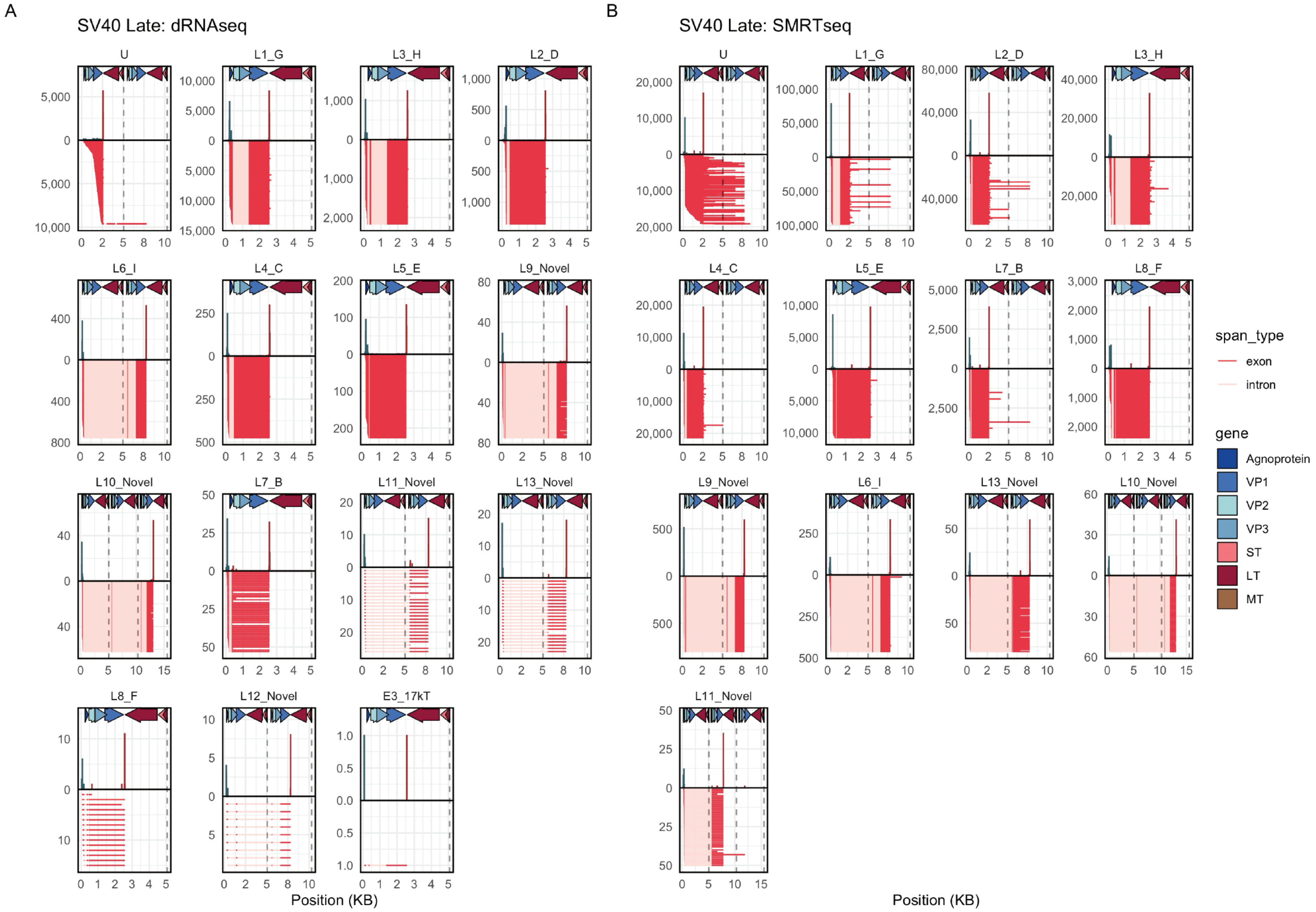


Fig S6: SV40 Early Transcript Pileups

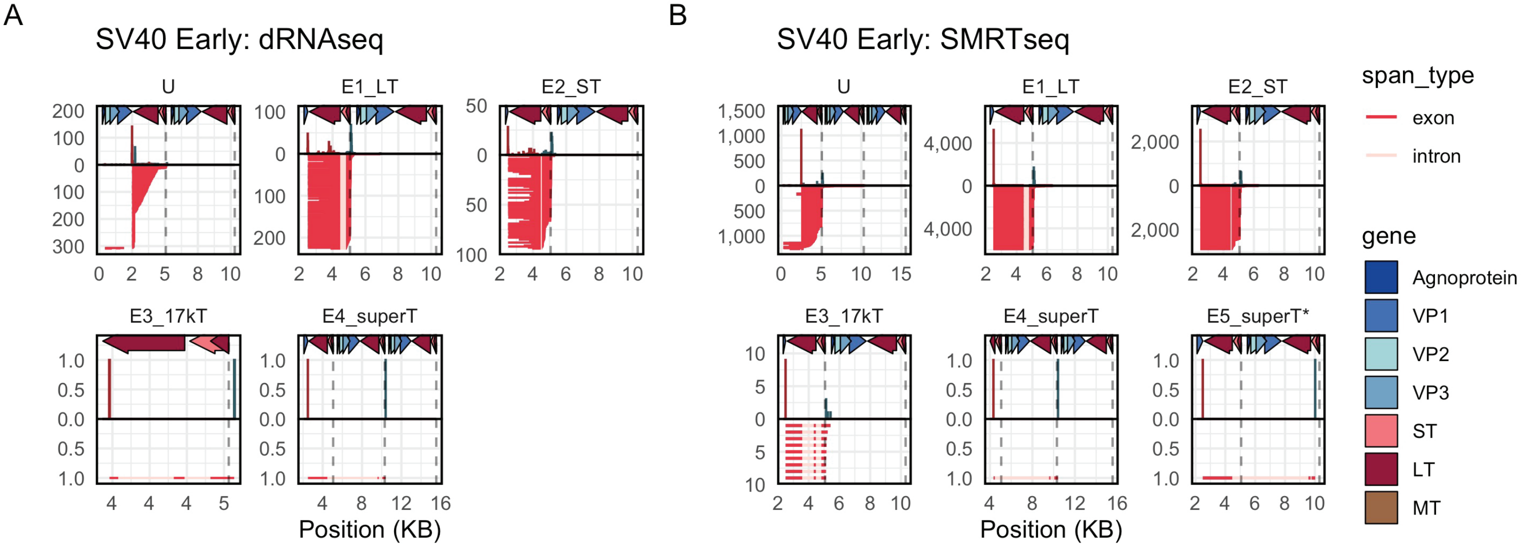


Fig S7: BKPyV (Dunlop) Late Transcript Pileups

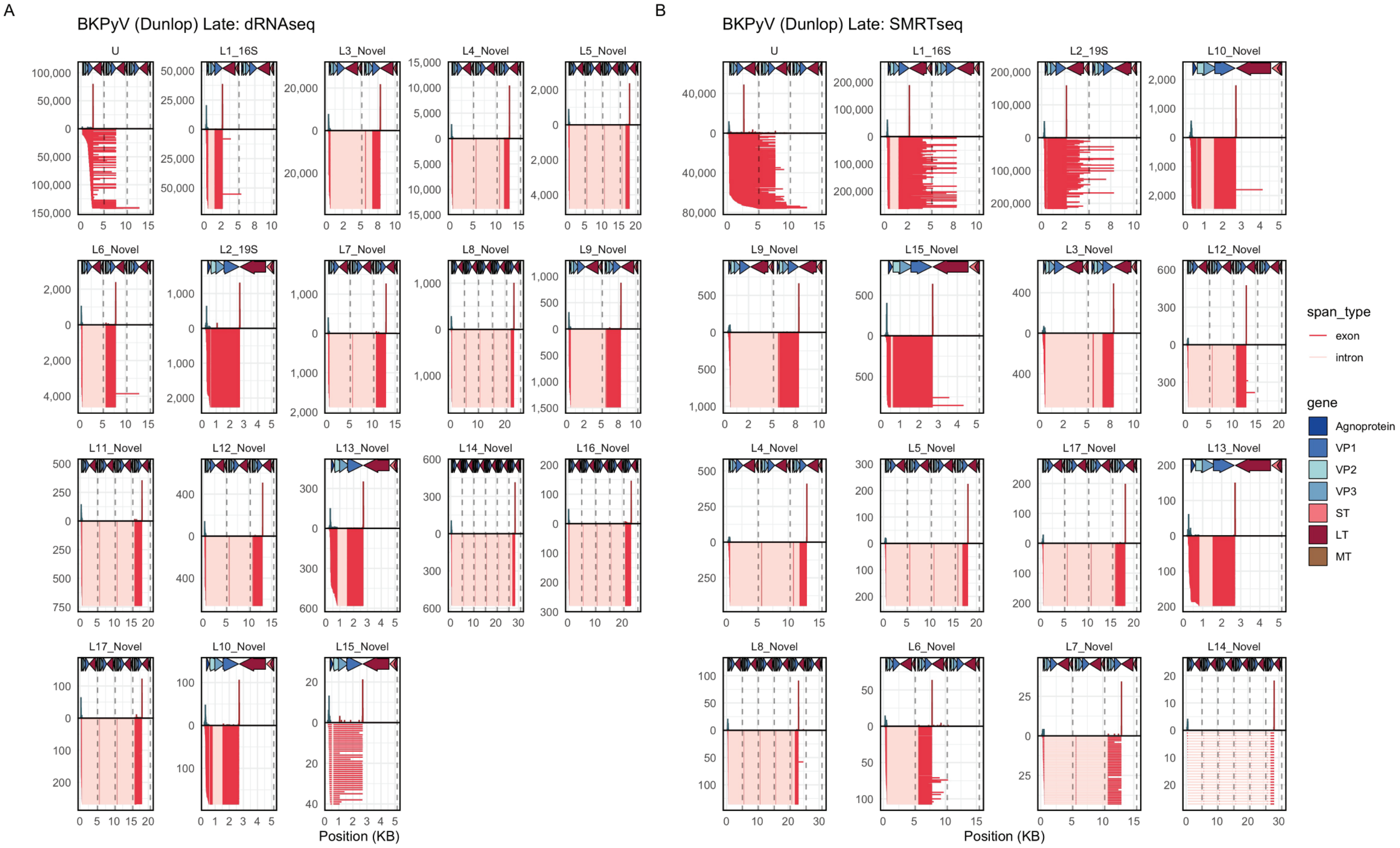


Fig S8: BKPyV (Dunlop) Early Transcript Pileups

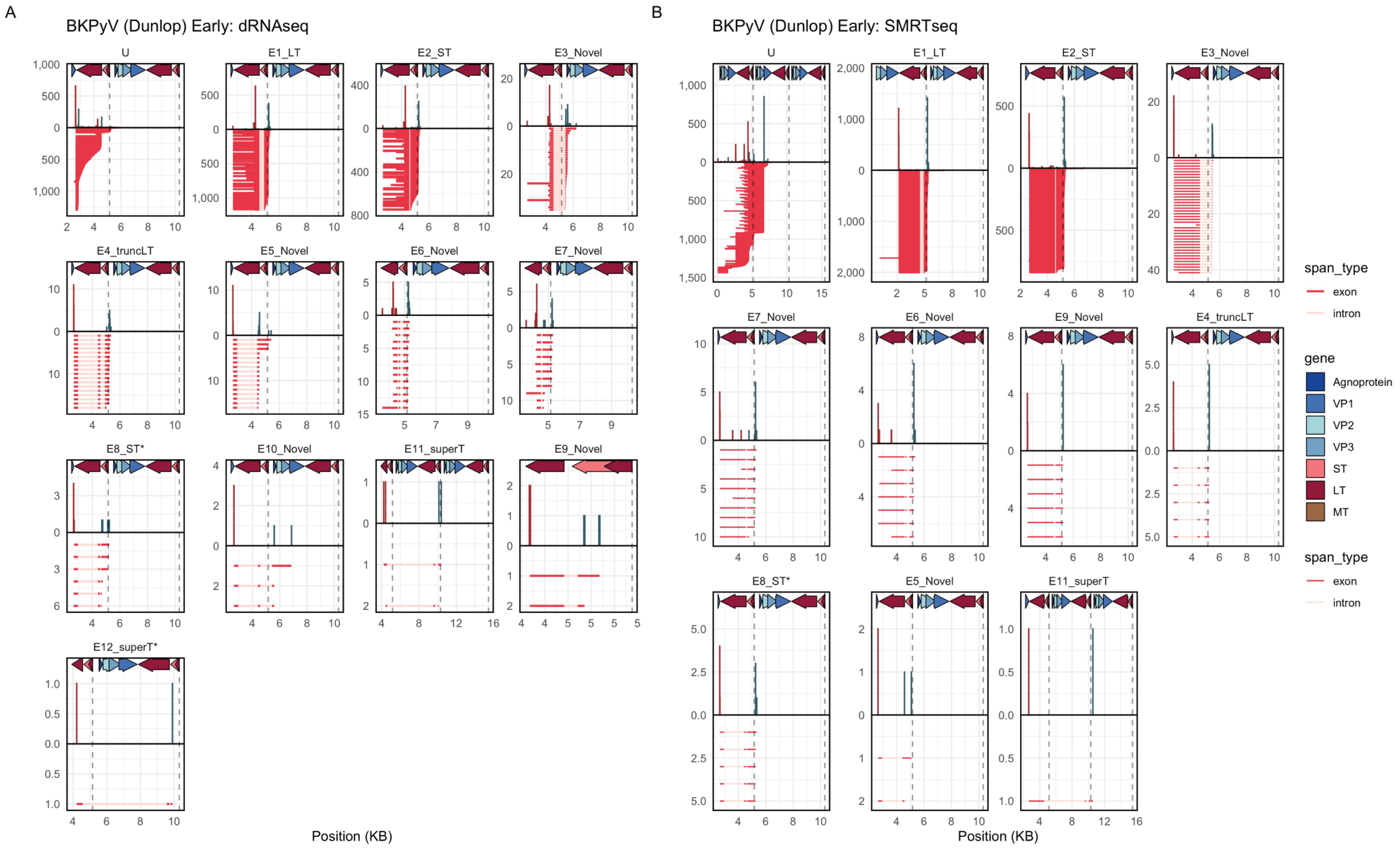


Fig S9

Illumina Introns >1% of Junction-spanning reads (by strand)

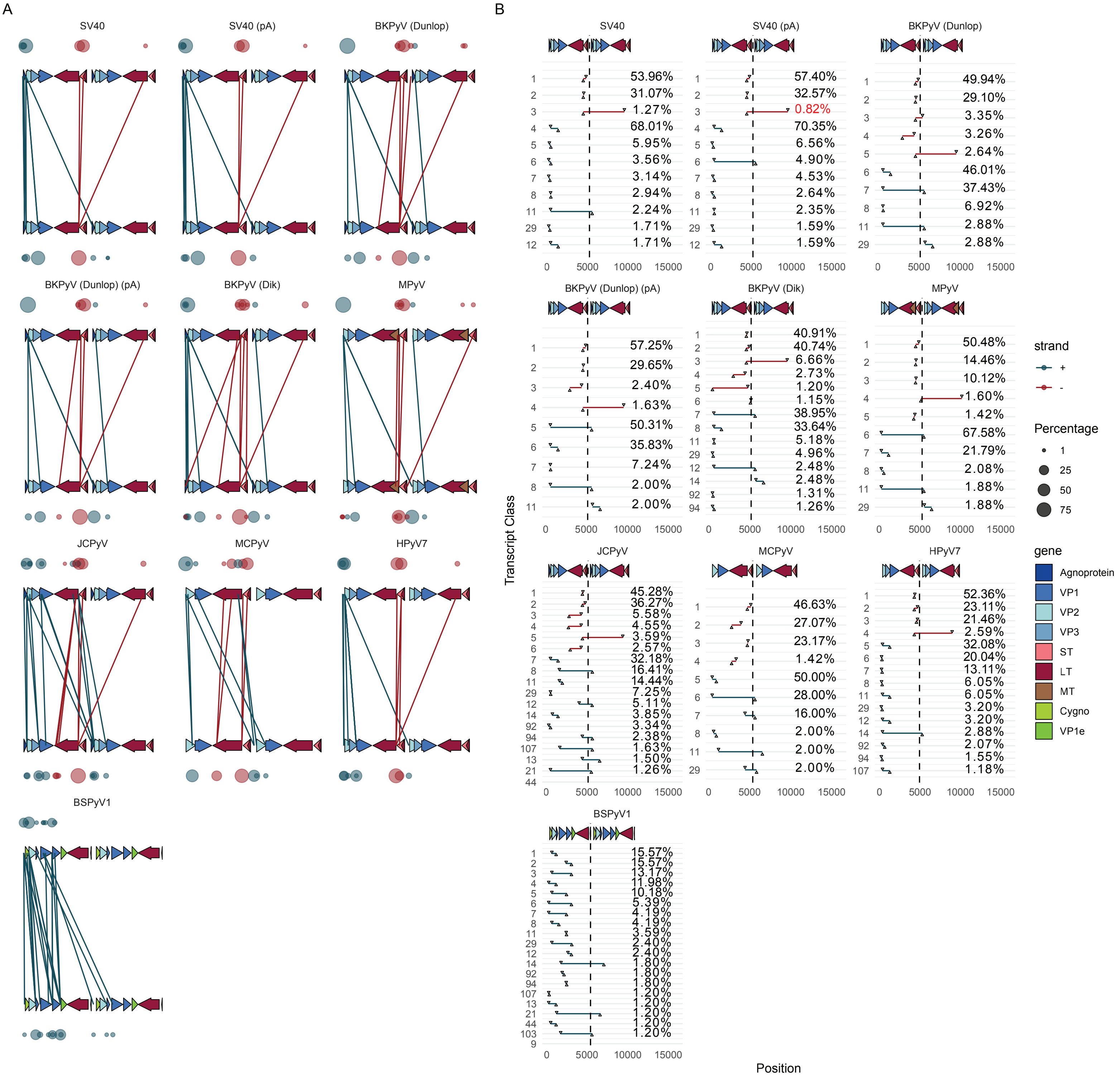


Figure S10

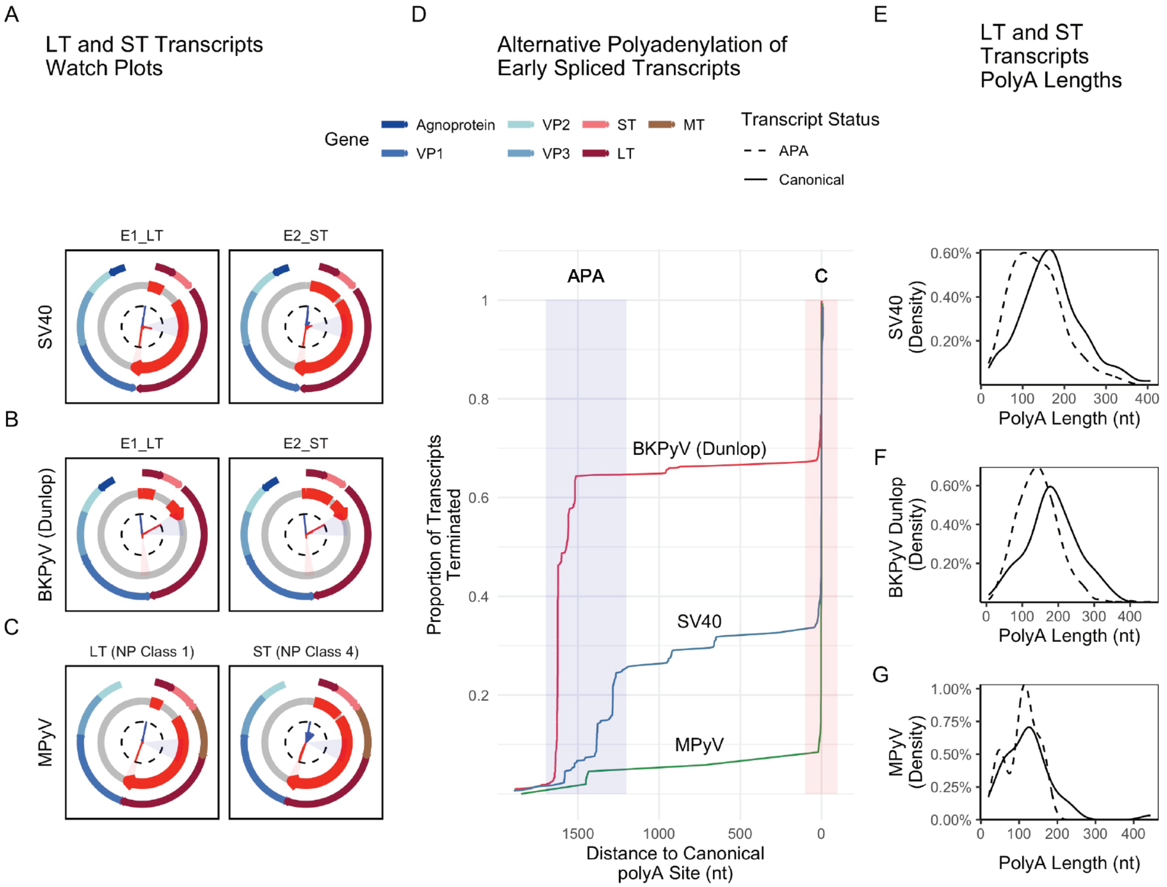

A
LT and ST Transcripts
Watch Plots
D
Alternative Polyadenylation of
Early Spliced Transcripts
E
LT and ST
Transcripts
PolyA Lengths
Gene
Agnoprotein
VP1
VP2
VP3
ST
LT
MT
Transcript Status
APA
Canonical
E1_LT
E2_ST
SV40
B
E1_LT
E2_ST
BKPyV (Dunlop)
C
LT (NP Class 1)
ST (NP Class 4)
MPyV
APA
C
1
0.8
0.6
0.4
0.2
Proportion of Transcripts
Terminated
BKPyV (Dunlop)
SV40
MPyV
1500
1000
500
0
Distance to Canonical
polyA Site (nt)
SV40
(Density)
0.60%
0.40%
0.20%
0
100
200
300
400
PolyA Length (nt)
F
BKPyV Dunlop
(Density)
0.60%
0.40%
0.20%
0
100
200
300
400
PolyA Length (nt)
G
MPyV
(Density)
1.00%
0.75%
0.50%
0.25%
0.00%
0
100
200
300
400
PolyA Length (nt)

Fig S11

BKPyV Dik short-RNAseq (polyA)

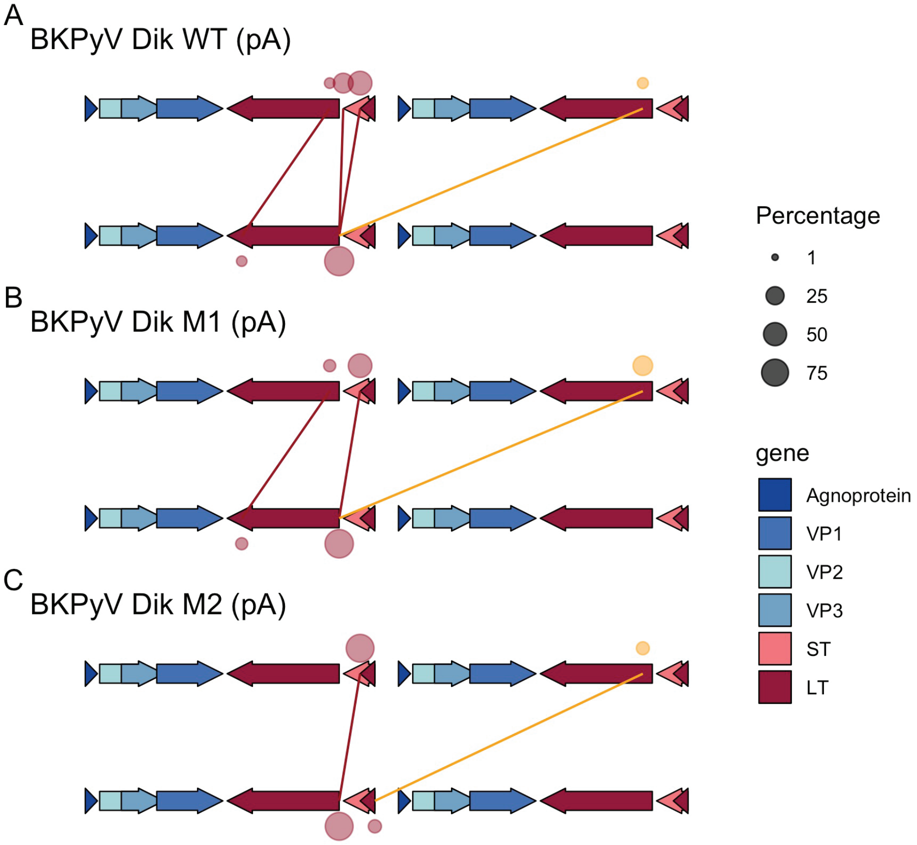
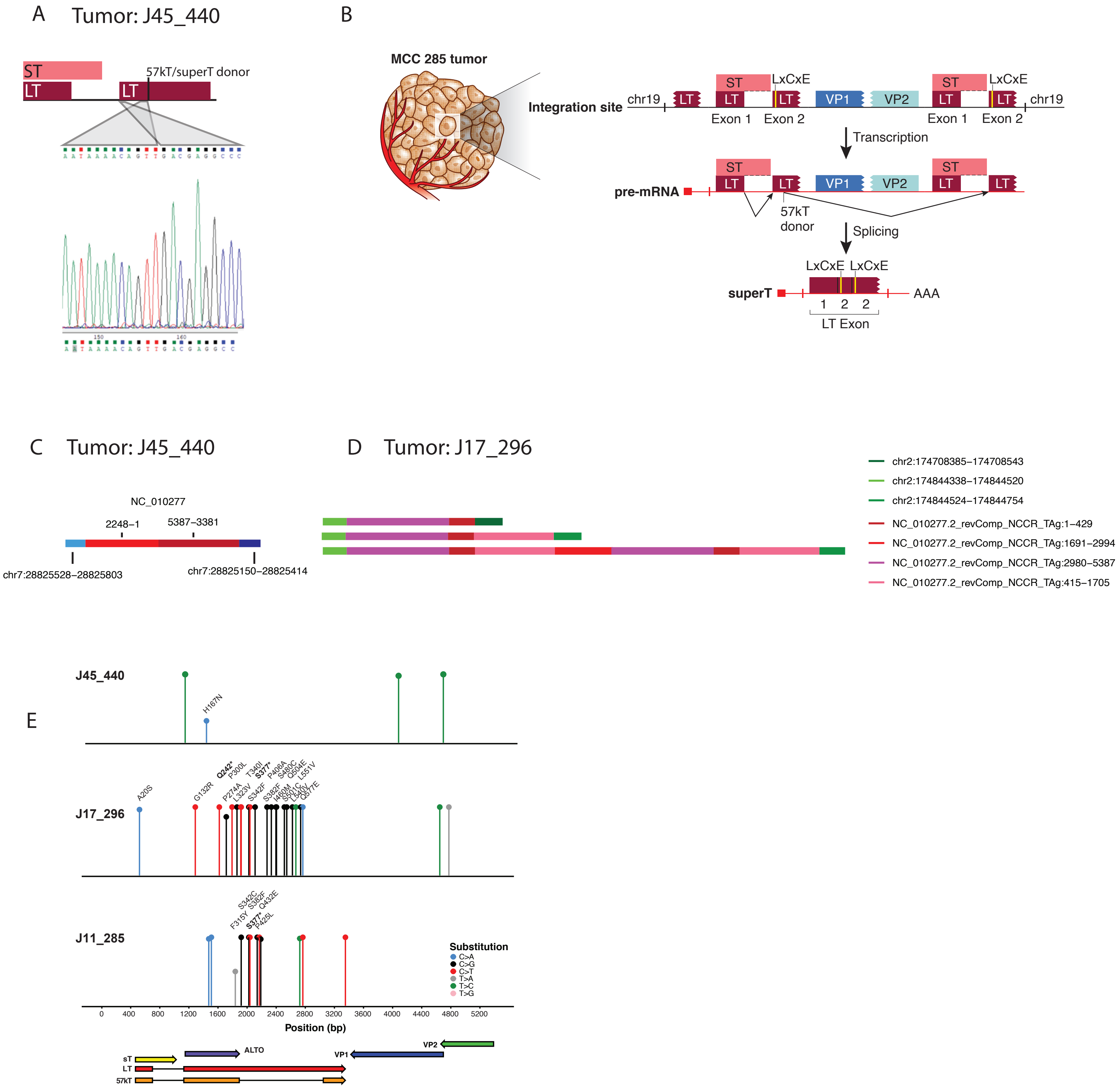


Fig S12


A
Tumor: J45_440
ST
LT
LT
57kT/superT donor
A A T A A A A C A G T T G A C G A G G C C C
150
160
A A T A A A A C A G T T G A C G A G G C C C
B
MCC 285 tumor
Integration site
chr19
LT
ST
LT
LxCxE
LT
VP1
VP2
ST
LT
LxCxE
LT
chr19
Exon 1
Exon 2
Exon 1
Exon 2
Transcription
pre-mRNA
ST
LT
LT
VP1
VP2
ST
LT
LT
57kT
donor
Splicing
LxCxE
LxCxE
superT
1
2
2
LT Exon
AAA
C
Tumor: J45_440
NC_010277
2248–1
5387–3381
chr7:28825528–28825803
chr7:28825150–28825414
D
Tumor: J17_296
chr2:174708385–174708543
chr2:174844338–174844520
chr2:174844524–174844754
NC_010277.2_revComp_NCCR_TAg:1–429
NC_010277.2_revComp_NCCR_TAg:1691–2994
NC_010277.2_revComp_NCCR_TAg:2980–5387
NC_010277.2_revComp_NCCR_TAg:415–1705
E
J45_440
H167N
J17_296
A20S
G132R
Q242*
P300L
P274A
L323V
T340I
S377*
S342F
P406A
S480C
Q504E
L551V
S382F
I460M
S501C
L540V
Q577E
J11_285
S342C
S382F
Q432E
F315Y
S377*
P425L
Substitution
C>A
C>G
C>T
T>A
T>C
T>G
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
5200
Position (bp)
sT
LT
57kT
ALTO
VP1
VP2