Title: Nanopore Guided Annotation of Transcriptome Architectures

Short Title: Nanopore-based Transcriptome Annotation

Authors: Jonathan S. Abebe[1], Yasmine Alwie[2#], Erik Fuhrmann[2#], Jonas Leins[2], Julia Mai[2,3], Ruth Verstraten[2,4], Sabrina Schreiner[2,3,5], Angus C. Wilson[1], Daniel P. Depledge[1,2,4,5*]

[1]Department of Microbiology, New York University School of Medicine, New York, NY, USA

[2]Institute of Virology, Hannover Medical School, Hannover, Germany

[3]Institute of Virology, University Medical Center, Albert-Ludwigs-University Freiburg, Freiburg, Germany

[4]German Center for Infection Research (DZIF), partner site Hannover-Braunschweig, Hannover, Germany

[5]Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany

[#]These authors contributed equally

*Corresponding author: email: depledge.daniel@mh-hannover.de

ORCIDs

JSA - 0009-0000-1268-0427

YA - 0000-0002-7001-5641

EF – 0009-0001-8512-9369

JL –

JM – 0000-0001-9501-5691

RV - 0009-0008-6046-6335

SS - 0000-0002-5744-7159

ACW - 0000-0002-5016-4164

DPD - 0000-0002-4292-0599

Keywords: Nanopore, Direct RNA Sequencing, Transcriptome, Annotation, Adenovirus, Herpesvirus, Coronavirus, HAdV-F41

ABSTRACT

High-resolution annotations of transcriptomes from all domains of life are essential for many sequencing-based RNA analyses, including Nanopore direct RNA sequencing (DRS), which would otherwise be hindered by misalignments and other analysis artefacts. DRS allows the capture and full-length sequencing of native RNAs, without recoding or amplification bias, and resulting data may be interrogated to define the identity and location of chemically modified ribonucleotides, as well as the length of poly(A) tails on individual RNA molecules. Existing software solutions for generating high-resolution transcriptome annotations are poorly suited to small gene dense organisms such as viruses due to the challenge of identifying distinct transcript isoforms where alternative splicing and overlapping RNAs are prevalent. To resolve this, we identified key characteristics of DRS datasets and developed a novel approach to transcriptome. We demonstrate, using a combination of synthetic and original datasets, that our novel approach yields a high level of precision and recall when reconstructing both gene sparse and gene dense transcriptomes from DRS datasets. We further apply this approach to generate a new high resolution transcriptome annotation of the neglected pathogen human adenovirus type F 41 for which we identify 77 distinct transcripts encoding at least 23 different proteins.

INTRODUCTION

The transcriptome architecture of a given organism denotes the full catalog of RNAs arising from the combined action of transcription and post-transcriptional processing. Of these, many RNAs are transcribed only in specific temporal or tissue contexts or in response to intrinsic or extrinsic stresses. The content and complexity of transcriptome architectures varies dramatically between different organisms and can be broadly classified as gene sparse or gene dense depending on the proportion of the genome that encodes transcripts. In contrast to the large gene sparse genomes of most eukaryotes and archaea, the genomes of viruses are generally small and gene dense (1). This poses a significant challenge to studies of gene regulation, transcription, and translation, particularly when using short-read sequencing approaches as these cannot adequately resolve alternative splicing and overlapping RNAs (2).

Long-read RNA sequencing enables the sequencing of full-length RNAs in the form of both native and recoded (cDNA) RNA using platforms developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) (3). These methodologies have significantly enhanced our ability to annotate transcriptomes of all sizes and complexities by (i) resolving simple and complex repeat regions, (ii) providing linkage between splice sites in studies

of alternative splicing, and (iii) enabling the discovery of new transcript isoforms. The specific attraction of nanopore direct RNA sequencing (DRS) (4), is the power to interrogate RNA biology at the level of individual molecules. In theory, each sequence read derived by DRS represents a single native RNA and thus contains all the information needed to identify (i) the corresponding genomic sequence from which it was transcribed, (ii) all modified ribonucleotides within the RNA molecule, and (iii) the length of the poly(A) tail (if present). This information can in turn guide predictions of secondary structure, stability, and ultimately, function. Our ability to perform such comprehensive analyses is steadily increasing with the development of computational approaches to extract such data (5–10). However, to successfully interrogate RNAs at the level of individual molecules, it is crucial that sequence reads can be unambiguously assigned to the correct transcript isoform – a process that requires a high-resolution annotation of the underlying transcriptome architecture. This has been demonstrated in a number of recent studies, all of which required the generation of high-resolution transcriptome annotations to facilitate the desired analysis (11–16). While many of these high-resolution annotations were obtained by laborious manual processing, this is neither a practical nor sustainable methodology. Several computational approaches capable of providing high-resolution transcriptome annotations and quantifications have recently been developed and have proven extremely powerful in the context of studying the gene sparse transcriptomes of higher eukaryotes (17). Examples include Stringtie2 (18), Bambu (19), and Isoquant (20). However, as these approaches appear designed with higher eukaryotic transcriptomes in mind, their utility in decoding the gene dense transcriptomes of viruses remains poor. This remains a significant issue for many viral pathogens including the adenovirus strain F serotype 41 (HAdV-F41) which is the primary cause of adeno-associated acute gastroenteritis of infants (21, 22) and more recently has been associated with adeno-associated virus (AAV)-driven cases of acute liver failure (23). HAdV-F41 differs from other human adenoviruses in terms of tropism and a detailed examination of its transcriptome and protein coding potential is urgently needed to provide further insight into its molecular behavior and pathogenicity.

To resolve this, we have developed a new computational approach entitled Nanopore Guided Annotation of Transcriptome Architectures (NAGATA) and showcase its ability to generate high-resolution transcriptome annotations from DRS datasets. Using both synthetic and real nanopore datasets, we demonstrate that NAGATA significantly outperforms other annotation tools in accurately reconstructing the transcriptomes of selected DNA and RNA viruses. We further present a new high-resolution transcriptome annotation for the neglected human pathogen, HAdV-F41.

Materials & methods

Publicly available datasets used in this study

Raw fast5 datasets for HAdV-C5 (PRJEB35667), VZV (PRJEB38829), and hCoV-OC43 (PRJEB42052) were downloaded from the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra)

Reference genomes and source annotations

The human genome assembly (GRCh38.p14) and GTF annotation files were obtained from Ensembl (https://www.ensembl.org/index.html). All viral reference genomes were downloaded from Genbank (https://www.ncbi.nlm.nih.gov/genbank/). The following accession numbers were used: HAdV-C5 (AC_000008.1), hAdV-F41 (ON561778.1), VZV strain Dumas (NC_001348.1), and hCoV-OC43 (NC_006213.1). The corresponding GFF3 annotation for HAdV-F41 was downloaded from the same source. GFF3 annotations for HAdV-C5 , SVV, and hCoV-OC43 were obtained from repositories associated with the recent reannotation efforts (14, 24, 25).

Generation of in silico datasets

Strand separated GFF3 files were converted, via genePred files, to BED12 files using UCSC tools (26). BEDtools v2.27.1 (27) getfasta [*-s -split*] was used to generate multi-FASTA files containing 200 copies of each transcript.

Generation of hAdV41 nanopore direct RNA sequencing datasets

A549 cells (ATCC, No. CCL-185) and HEK293 (ECACC European Collection of Authenticated Cell Cultures; Sigma-Aldrich, No. 85120602-1VL) were grown in Dulbecco's modified Eagle's medium supplemented with 5% fetal calf serum, 100 U of penicillin, 100 µg of streptomycin per mL in a 5% CO2 atmosphere at 37 °C. These cell lines are frequently tested for mycoplasma contamination. HAdV-F41 (wild-type Tak strain; (28)) was propagated and titrated in HEK293 cells by quantitative immunofluorescence staining of the hexon protein (8C4, Santa Cruz Biotechnology) at 48 hpi. as previously published (29). For HAdV-F41 infection, A549 cells were infected at an MOI of 50 in non-supplemented DMEM. After incubation for 1 h at 37°C, infection

was stopped by replacing virus containing medium with fresh DMEM medium with supplements. At defined times post infections, the supernatant was removed and cells lysed in 8 ml Trizol per 10 cm dish. After equilibration, 0.2 vol of chloroform was added followed by vigorous vortexing, a three-minute incubation at RT, and centrifugation at 12,000g for 15 mins at 4°C. The aqueous phase was collected and precipitated using 0.5 vol isopropanol and 1 ul of Glycoblue (Invitrogen) for 10 mins at RT prior to pelleting by centrifugation at 12,000g for 15 mins at 4°C. Pellets were washed in 75% ethanol and centrifuged at 12,000g for 5 mins at 4°C. The supernatant was removed and the pellet air-dried for 5 min before resuspending in RNAse-free water and incubated at 55°C for 10 min before quantification with a Qubit hsRNA kit (Invitrogen). Poly(A) selection was performed using Dynabeads (Invitrogen) with 133 µl beads added to 25 µg of total RNA. Nanopore DRS libraries were prepared according to the Deeplexicon multiplexing protocol (30) and sequenced for 24 hours on an R 9.4.1 flowcell using a MinION Mk.1b .

Basecalling and poly(A) tail estimation

For all DRS datasets, high-accuracy basecalling was performed with Guppy v6.5.7 [*-c rna_r9.4.1_70bps_hac.cfg -r --calib_detect --trim_strategy rna --reverse_sequence true*] and poly(A) tail analyses generated with nanopolish v0.14.

Alignment and downstream processing of in silico and DRS datasets

For HAdV-C5 , VZV, and HG38, reads in fastq files were aligned against the respective reference genome using minimap2 (31) [*-ax splice -k14 -uf --secondary=no*] and parsed to generate sorted BAM files using SAMtools v1.15 (32) in which only primary alignments were retained [*samtools view -F 2308*]. For hCoV-OC43, alternative minimap2 parameters were specified [*-ax splice -k 8 -w 3 -g 30000 -G 30000 -C0 -uf –no-end-flt –splice-flank=no*] to account for discontinuous transcription (33).

Transcriptome reconstruction parameters

To reconstruct the transcriptomes presented in this study, NAGATA was run with the following (default) parameters [*-s 5 -c 100 -t 50 -cg 50 -tg 30 -iso 15 -m 8 -a 1 -b 1*], except for several instances in which –c and –t parameters were reduced. For Stringtie2 v2.1.3 (18), the following flags were used [*--viral -L*]. For Isoquant v3.1.1 (20) [*--data_type nanopore* --

model_construction_strategy default_ont --splice_correction_strategy default_ont, --fl_data --matching_strategy loose --report_novel_unspliced]. For Bambu v3.3.5 (19), we followed the *de novo* transcript discovery approach described in their manual and included a flag for single exon discovery [*bambu(reads = "in.bam", annotations = NULL, genome = "ref.fasta", NDR = 1, quant = FALSE, opt.discovery = list(min.txScore.singleExon = 0))*]. For all four tools, the same sorted BAM files were used as input and resulting GTF files (Stringtie2, Isoquant, Bambu) were converted into bed files using UCSCutils (26) gtfToGenePred and genePredtoBed.

Overlap analyses

Transcript annotations produced by each tool were converted from GFF3/GTF to BED12 using gtfToGenePred and genePredToBed from the UCSCutils (26) package and compared against existing annotations in BED12 format using the custom python script post_intersect_processing_v4.1.py. This produces three BED12 outputs: annotation overlaps, tool-specific annotations, and annotation only. Each of these contains transcripts assigned to these three categories. To compare outputs from multiple tools (e.g. annotation overlaps from NAGATA, Bambu, Isoquant, and Stringtie2), we used the custom python script multiple-overlap.v1.py. Both custom scripts are available from https://github.com/DepledgeLab/NAGATA. F1 scores (2*(p*r)/(p+r)) were calculated using (*p*) precision (true positives (TP) + false positives (FP) + false negatives (FN)) and (*r*) recall (TP/(TP+FN) values.

Generation of R plots & R packages used

All plotting was performed using Rstudio (https://posit.co/download/rstudio-desktop/) with R v4.1.1 and the following packages: data.table (https://r-datatable.com), Gviz (34), GenomicFeatures (35), ggplot2 (36), UpSetR (37), dplyr (https://dplyr.tidyverse.org/), tidyr (https://tidyr.tidyverse.org/), and patchwork (https://github.com/thomasp85/patchwork).

Data availability

NAGATA is written in Python 3 and is available in the https://github.com/DepledgeLab/NAGATA repository, along with test datasets and accessory scripts. The deeplexicon multiplexed raw Fast5 dataset generated for hAdV41 is available from the ENA/SRA under the accession number PRJEB72818.

RESULTS

Characteristics of nanopore DRS genome alignments inform transcript boundaries

The aim of this study was to implement a new algorithm for generating high-resolution transcriptome annotations from DRS datasets using read alignments against a genome of interest. As standard DRS proceeds in a 3' -> 5' direction, first through the adapter, then the poly(A) tail, and finally the body of the RNA itself, all reads are expected to contain the poly(A) tail and the 3' end of the RNA. Processing of the raw nanopore signals allows segmentation of these three units (4). This, in theory, allows precise plotting of the cleavage and polyadenylation sites (CPAS). However, an analysis of multiple extant nanopore DRS datasets (14, 38, 39) using nanopolish (5) demonstrates that poly(A) tails can only be reliably detected in ~58 – 83% of the reads (Table S1). We theorized that reads for which a poly(A) tail could not be identified by nanopolish would likely be over- or under-trimmed and that this would impact on accurate mapping of CPAS. For the 5' end of RNAs, it has been observed that nanopore DRS will sequence to within a few nucleotides of the 5' end (14, 40). Given the continuous turnover of poly(A) RNA in the cell, combined with in vivo/vitro strand breakage and signal processing errors (40), only a proportion of sequenced RNAs are expected to be full length and thus would share near-identical 5' alignment ends that can be interpreted as transcription start sites (TSS). For non-full length RNAs that originate from multi-exon splicing this can create alignment artefacts where 5' ends cannot be extended across splice junctions. This in turn leads to extensive 5' soft clipping of the alignment and the clustering of many 5' alignment ends at the same location, thus giving rise to artifact TSS.

To examine this more closely, we used existing datasets from adenovirus type 5 (HAdV-C5) infected A549 cells and Varicella Zoster Virus (VZV) infected ARPE-19 cells for which high-resolution transcript annotations exist and for which the TSS and CPAS have been confirmed by orthologous methodologies (14, 24). We segregated reads according to the presence or absence of a detectable poly(A) tail and whether resulting alignments showed 5' soft clipping > 3 nt, and subsequently determined the closest annotated TSS and CPAS for each read. Soft clipping denotes portions of a read that cannot be aligned to the target, either due to sequence mismatch or, in the case of splice junctions, the inability to locate the 5' junction site. For both HAdV-C5 and VZV we observed that reads with 5' soft clipping > 3 nt could be associated with artefact TSS and produced high levels of noise in regions proximal to previously confirmed TSS (Fig. 1A-B, Fig. S1A-B). Similarly, alignments using reads without detectable poly(A) tails resulted in larger

numbers of 3' alignments that were > 50 nt from defined CPAS (Fig. 1C, Fig. S1C). Note that the CPAS used for HAdV-C5 and VZV were previously defined using Illumina RNA-Seq datasets (14, 24) in conjunction with ContextMap2 (41). TSS used for HAdV-C5 were previously defined by nanopore DRS while those used for VZV were defined by CAGE-Seq(14, 24) The latter is considered the most accurate method as nanopore DRS can only sequence to within 5-10 nt of the 5' cap (14, 15), hence why the 'distance to nearest TSS' shows a greater offset in the VZV data (Fig. S1A-B). Thus, defining TSS and CPAS using DRS alignments requires careful filtering of reads without measurable poly(A) tails and alignments showing 5' soft clipping, a procedure that is not currently utilized by existing transcriptome annotation softwares.

Nanopore Guided Annotation of Transcriptome Architectures (NAGATA)

Utilizing the characteristics described above, the NAGATA algorithm is designed to convert DRS alignments against a genome into corresponding transcriptome annotations. As input, it accepts a sorted BAM file containing genome-level primary alignments and a poly(A) output file from the nanopolish package (5). NAGATA subsequently functions through three distinct stages (i) prefiltering, (ii) TSS/CPAS definition, and (iii) isoform deconvolution and filtering (Fig. 2). The pre-filtering step masks alignments for which (i) a poly(A) tail could not be detected by nanopolish, and/or (ii) soft clipping above a specified threshold (default = 3) is observed at the 5' end of the alignment. Using the top strand of the recently reannotated adenovirus type 5 (HAdV-C5) transcriptome as an example (12), we demonstrate how raw DRS alignments lead to multiple artefact TSS and CPAS that are otherwise eliminated when applying our pre-filtering strategy (Fig. 2A-B). For the TSS/CPAS definition step (Fig. 2C), NAGATA first defines transcriptional units (TUs) by grouping alignments with identical 3' ends and determining the number of alignments in each group. Alignments generated from reads without detectable poly(A) tails are masked at this stage while reads with 5' soft clipping are retained. 3' end positions with an abundance count above a user-defined threshold (default = 50) are used as anchors and all alignments with 3' ends within a defined distance (default +/- 25 nt) of an anchor are added to the group. If two or more anchors are present in this range, the defined 3' end defaults to the anchor with the largest number of initial alignments (Fig. 3). Once 3' end grouping is complete, each anchor is defined as a CPAS and the 3' end of individual alignments in each group are corrected to match that of the anchor (Fig. 3). This process is subsequently repeated to define TSS by grouping the 5' ends of all alignments. 5' alignments within a defined distance (default +/- 12 nt) of an anchor are added to the same group and the 5' end of individual alignments corrected to match that of the anchor (Fig.

3). Here, alignments generated from reads with 5' soft clipping are masked while reads without detectable poly(A) tails are retained. The final step deconvolutes the isoforms present in each TU (Fig. 2D), a function that is performed by first segregating alignments by the number of exons present and subsequently by comparing the genomic position of the exons. Here a distance of up to 50 nt between exon start and end positions is allowed between alignments, again with a correction step based on the position with the most alignments. Finally, for each resulting isoform, we apply two filters to decide on the validity of the isoform. The first filters on the total number of supporting alignments (raw count) while the second calculates a TSS/CPAS ratio (number of supporting alignments / total alignments associated with the same TSS/CPAS). The latter specifically functions to identify and remove low abundance isoforms (by default < 1% frequency).

Benchmarking NAGATA using synthetic datasets

Transcriptomes vary in size and complexity between organisms but most analytical softwares appear designed and optimized for a specific organism e.g. *H. sapiens*, a process that may lead to suboptimal performances when applied to different transcriptome architectures. The aim of NAGATA was to implement an approach that is agnostic in regard to the underlying transcriptome architecture. To test this, we generated an *in silico* dataset comprising 200 copies of all RNAs encoded on chromosome 1 of the gene sparse human genome and aligned these back against the genome using Minimap2 (31). We calculated precision and recall (F1) scores for NAGATA and three popular annotation tools; Stringtie2 (18), Isoquant (20) and Bambu (19) and observed all four produced similar results (Fig. 4A). Surprisingly, no tool achieved a perfect score indicating that even with an idealized dataset, the process of alignment alone introduces artefacts and error into the final results. We next applied the same strategy to two DNA viruses (Adenovirus Type 5 and Varicella Zoster Virus (VZV)) with distinct gene dense transcriptome architectures (14, 24). HAdV-C5 transcriptomes consist of relatively few TUs and large numbers of alternatively spliced RNAs (Fig. S2), whereas VZV transcriptomes consist of large numbers of TUs, each predominantly comprised of multiple single-exon transcripts with unique TSS but shared 3' co-terminal ends (Fig. S3). NAGATA produced an F1 score of 0.99 for HAdV-C5 and was able to reconstruct 88/89 transcript isoforms with no false positives (Fig. 4B, Fig. S2), Both Stringtie2 (77/89 true positives, 13 false positives, F1 = 0.83), and Isoquant (71/89 true positives, 19 false positives, F1 = 0.79) were able to identify the relative positions of all canonical TSS and CPAS, apart from Stringtie2 failing to correctly identify transcripts of the E3 region (Fig. S2). Instead, the predicted transcripts have similar structures but were co-ordinate shifted relative to the canonical

transcripts. Bambu performed the least well (31/89 true positives, 0 false positives, F1 = 0.34). While Bambu and Isoquant were both run using parameters allowing for the detection of mono-exonic transcripts (e.g. pIX, E3.12k and E4orf1), neither tool was able to identify any. For VZV, NAGATA produced an F1 score of 0.97 with 135/137 transcript isoforms correctly identified and three false positives (Fig. 4C, Fig. S3). Stringtie2 (55/137 true positives, 3 false positives, F1 = 0.40), Isoquant (6/137, 0 false positives, F1 = 0.04), and Bambu (23/137, 25 false positives, F1 = 0.15) all performed poorly. Together, these results indicate that the underlying architecture of the selected gene dense viral transcriptomes can be resolved by NAGATA but not by other existing annotation tools.

Benchmarking NAGATA using real nanopore datasets

To verify that the results shown above were not biased by using synthetic (idealized) datasets, we next examined NAGATA's performance using real DRS datasets. We first downloaded a subset of the DRS data used to generate the most recent annotation of HAdV-C5 (12). These datasets derived from A549 cells infected with HAdV-C5 for either 12 or 24 hours and were analyzed individually (12h, 24h) and in combination (12h-24h). Using the 12h-24h combined dataset, NAGATA identified 144 transcripts, 71 of which were present in the existing annotation (Fig. 5A). Of the 73 novel transcripts identified by NAGATA, the majority could be classified as either incompletely spliced pre-mRNAs or alternatively spliced isoforms of known transcripts. Taking the E1 region as an example, NAGATA identified 11/13 annotated transcripts and two 'novel' transcripts that we classified as unspliced polyadenylated pre-mRNAs of comparatively low abundance (Fig. 5A&B). In total 13/73 transcripts were recorded as incompletely spliced pre-mRNAs (marked with asterisks in Fig 5A, the majority located in the L1 region). A further 16/73 novel transcripts matched existing transcript structured but additionally contained the 'i-leader' exon that is occasionally incorporated into the Major Late Promoter (MLP) tripartite leader (42). The remaining 44 newly identified transcripts were alternatively spliced isoforms of previously annotated transcripts. Notably, all were of relatively low abundance within a given transcription unit (Fig 5B). A further 16 transcripts in the current annotation were not detected. Visual inspection of the raw read data confirmed them to be either absent in that specific dataset or supported by only 1-2 reads and thus below the detection threshold of NAGATA. To examine the value of including multiple timepoints when running NAGATA, we compared the results obtained from the individual 12h (n = 75 transcripts) and 24h (n = 125 transcripts) datasets with the 12h-24h dataset (n = 144 transcripts) (Fig 5C). Unsurprisingly, merging of the datasets increased the number of

transcripts reported. To measure the impact of overall sequencing depth on NAGATA results, we randomly subsampled the HAdV-C5-12h-24h dataset to four different read depths and applied NAGATA using the default parameters. Intriguingly, the number of transcripts identified showed only a small increase between 100k to 250k viral reads, suggesting that subsampling approaches may be useful for identifying when sequencing depth has reached a saturation point for transcript detection (Fig 5D). Finally, we again compared NAGATA's performance to that of Stringtie2 and Isoquant and observed a large reduction in the numbers of transcripts identified that overlapped with the existing annotation and, in the case of Stringtie2, a number of novel transcripts that did not overlap with the novel transcripts identified by NAGATA (Fig 5E, Fig S4) and were not supported by the underlying read alignments.

For the second test, we downloaded and analyzed an VZV DRS dataset that was derived from ARPE-19 cells infected at low MOI with wild-type VZV strain EMC-1 for 96 hours (13). We aligned this dataset against the VZV strain Dumas genome and processed the output with NAGATA, comparing the results against the existing VZV transcriptome annotation (13) (Fig 6A). Following visual inspection of the prefiltered datasets (i.e., after removal of reads without well-defined poly(A) tails and removal of alignments with 5' soft-clipping values > 3), we reduced the thresholds for defining putative TSS (-t flag) and CPAS (-c flag) which increased the number of putative TSS peaks from 67 to 89 and CPAS peaks from 44 to 54 (Fig 6B, Fig. S5). This approach increased the total number of transcripts reported by NAGATA from 129 using default settings to 147 using optimized settings (Fig 6C). A corresponding increase in the number of transcripts overlapping with existing annotations (from 58 to 76) was also observed (Fig 6C). We examined the read alignments underlying the new transcripts (n = 71, Fig 6A) and confirmed 24 of these utilized 22 distinct TSS that were not previously reported while just two utilized CPAS that had not previously been described. The remaining new transcripts could be classified as alternatively spliced or single exon transcripts that utilized different combinations of existing TSS and CPAS. Visual inspection of the read data confirmed the newly identified TSS to be robust and further confirmed an absence of sufficient read data at previously reported TSS that were not identified by NAGATA. A total of 56 transcripts from the original annotation were not identified here. Notably, many of these were located in regions of low read coverage and thus were either not supported by enough individual reads or were absent entirely in the downloaded dataset. Of transcripts encoding known protein-coding ORFs, only four were not detected by NAGATA in this data (pORF28, pORF38, pORF55 and pORF56 (Fig 6A). Across all reported transcripts, the abundance value of newly identified transcripts (median read count = 90) was lower than for previously annotated transcripts (median read count = 300) (Fig 6D). Further analyses with

Stringtie2 and Isoquant again resulted in a small number of overlapping transcripts being identified and a large number of erroneous transcripts that were not consistent with the underlying read alignments (Fig 6E, Fig. S6).

Application of NAGATA to a cytoplasmic RNA virus

To expand beyond nuclear-replicating DNA viruses, we also examined the ability of NAGATA to reconstruct the transcriptome of the cytoplasmic betacoronavirus hCoV-OC43. Coronaviruses are members of the Nidovirales order which replicate through transcription of negative-sense RNA intermediates that serve as templates for positive-sense genomic RNA (gRNA) and sub-genomic RNAs (sgRNAs). sgRNAs are generated through a process termed discontinuous transcription that combines a leader sequence in the 5' UTR with varying regions from the 3' end of the genome (43). From a computational perspective, alignments of sgRNAs against a genome appear similar to those generated from spliced RNAs although care must be taken to ensure that alignment and downstream processing software accurately record the junctions between the leader sequence and body. The prior annotation of hCoV-OC43 identified nine sgRNAs in addition to the primary gRNA (44). Using a publicly available DRS dataset that was generated from hCoV-OC43 infected MRC-5 cells (25), we observed that NAGATA was able to reconstruct all reported sgRNAs in addition to two previously unannotated sgRNAs (Fig. 7A). Of these, both of which were low abundance (Fig. 7B), the first contained a 3' junction between those of sgRNAs encoding the M and N proteins while the second used the same 3' junction as the sgRNA encoding N protein but contained additional sequence in the 5' leader. While Stringtie2 was also able to reconstruct all sgRNAs (and the gRNA), it also reported a larger number of artefact transcripts that did not coincide with sgRNA junctions (Fig 7C). By contrast, Isoquant was only able to reconstruct three sgRNAs and produced over 20 novel transcripts that were not supported by the underlying data (Fig 7C).

Defining the transcriptome of human adenovirus F serotype 41

The linear dsDNA genome of HAdV-F41 has a length of 34188 bp and prior studies have indicated it shares a similar overall transcriptome architecture to other human adenoviruses (45), although many proteins are poorly conserved and differ in length (46–50). Despite increasing interest in this neglected human pathogen, the existing reference genome annotation (ON561778.1) contains just 33 reported coding sequences (CDS). To address this, we infected

A549 cells with HAdV-F41 at a multiplicity of infection (MOI) of 50 and collected total RNA at 12, 24, and 48 hours post infection. We isolated the poly(A) fractions and prepared multiplexed nanopore DRS libraries using the deeplexicon protocol (30) and sequenced these for 24 hours on a nanopore MinION. Following basecalling and demultiplexing, the datasets were aligned against the HAdV-F41 reference genome and processed using NAGATA. We observed ten-fold fewer read alignments against the reverse strand compared to the forward strand (Fig. 8B) and thus analyzed each strand individually with different values for -t and -c. In total, NAGATA reported 11 transcription units comprising a total of 77 transcripts, 70 on the forward strand and 7 on the reverse strand (Fig. 8A). For the forward strand, transcripts representing all major transcription units (E1A-B, L1-L5, E3) were identified and the only annotated CDS that could not be assigned to NAGATA-derived transcripts were E3-14.5K and E3-14.7K. We assigned CDS sequences for E1A-S and E1A-9s homologs that were not reported in the original annotation and also identified a putative N-terminal truncated Fiber^{long} isoform (Fig. 8A). While the overall architecture of the HAdV-F41 transcriptome mirrors that of the HAdV-C5 transcriptome in terms of transcription units, there are some notable differences. Specifically, the L2, L4, and L5 regions all possess dual CPAS. For L2, the upstream CPAS is strong, as evidenced by the higher abundance of transcripts terminating at this position compared to the downstream CPAS (Fig. 8C). For L4 the situation is reversed with the upstream CPAS being a weak terminator (Fig. 8C). By contrast, both L5 CPAS show similar strengths. In contrast to HAdV-C5, the L5 region of HAdV-F41 encodes two distinct Fiber isoforms (Fiber^{short} & Fiber^{long}). For the reverse strand, we identified two discrete transcripts encoding IVa2 that differed only in their TSS position (5377 vs 5413). We further identified four alternatively spliced transcripts encoding DBP, each with a unique TSS, while a fifth single-exon transcript with a TSS in the 3' exon of DBP, putatively encoding an N' terminal truncated DBP, was also identified (Fig. 8A). NAGATA was unable to identify transcripts encoding AdPol, TP, or any of the E4 region proteins. Close examination of the raw read data indicated low level transcription of these regions but at insufficient levels for NAGATA to decode. Taken together, NAGATA has significantly increased the resolution of the hAdV41 transcriptome and provides a foundation for future studies.

DISCUSSION

Decoding transcriptome architectures using nanopore DRS is essential for accurately interrogating RNA biology at single molecule resolution. Multiple approaches have been developed for this purpose and softwares such as Bambu (19), Isoquant (20), and StringTie2 (18)

are highly effective in reconstructing transcriptome annotation for gene sparse higher eukaryotic transcriptomes (20). However, as shown by our analyses here, these appear generally unsuited to reconstructing transcriptomes of gene dense genomes (e.g. viral genomes). This manifests as a failure to correctly separate transcript isoforms that share significant overlap and we interpret this limitation as being due to the presence of many overlapping RNAs with distinct TSS and co-terminal 3' ends, a feature of many gene dense genomes.

The NAGATA method was informed by specific characteristics of nanopore DRS datasets and resulting genome-level alignments. It provides an alternative approach to transcriptome annotation by grouping similar structural elements together, alignment-by-alignment. Alignments with similar TSS and CPAS values are used to define the initial set of transcriptional units with the most abundant TSS, CPAS positions used as 'anchors to collapse and correct all relevant alignments. The collapsing and correction steps increases sensitivity without biasing the identification of TSS and CPAS, as evidenced by comparisons to CAGE-seq and ContextMap2 results from prior studies (41, 51, 52). Similarly, isoform level deconvolution takes place by grouping alignments in a TU by the similarity of the positions and sizes of the exons present. By identifying and retaining only reads from full-length RNAs, between 40-60% of reads in most datasets are removed prior to analysis. This naturally limits the utility of NAGATA in settings where read depth is low (e.g., the E4 region of hAdV41, Fig. 8) unless integrated with approaches such as Nanopore ReCappable Sequencing (53), that preferentially capture full-length RNAs.

A limitation of all annotation tools, including NAGATA, is the underlying assumption that the transcriptome architecture remains consistent across a genome. While this holds true in many cases (e.g., *H. sapiens*, adenoviruses, coronaviruses), there are notable exceptions. The transcriptome architecture of VZV is dominated by single-exon transcripts with co-terminal 3' ends. However, there are also several small regions encoding multitudes of alternatively spliced multi-exon transcripts (Fig. 6). Accurately reconstructing 'transcriptional islands' such as this may require different parameters to the rest of the genome.

The development of NAGATA enabled us to generate a substantially improved annotation of the HAdV-F41. Here, the existing annotation comprised a handful of predicted CDS with no information on TSS or CPAS. Using NAGATA, we identified 77 transcript isoforms from 11 TUs and were able to assign new or known CDS to all of these. We further observed the presence of CPAS redundancies for the L2, L4, and L5 TUs (Fig. 7) that is not seen in the related HAdV-C5. The functional relevance of these is not known and thus bears further investigation.

In summary, NAGATA offers a novel and flexible approach to generating high-resolution transcriptome annotations from nanopore DRS datasets that can be applied against both gene sparse and gene dense organisms. Given increasing global efforts to sequence large number of genomes from all domains of life and the need to supplement these with accurate transcriptome maps, we offer NAGATA as a new approach to achieve this objective.

Funding

SS was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the framework of the Research Unit FOR5200 DEEP-DV (443644894) project 08. ACW is supported by grants from the National Institute of Allergy and Infectious Disease R01-AI170583 and R01AI176335. DPD is supported by a German Centre for Infection Research (DZIF) Associate Professorship and the NIAID grants R01-AI170583 and R01-AI152543. DPD and SS also receives funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2155 - project number 390874280.

References

1. Koonin,E.V. (2009) Evolution of genome architecture. *Int J Biochem Cell Biol*, **41**, 298–306.

2. Depledge,D.P., Mohr,I. and Wilson,A.C. (2019) Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes. *J Virol*, **93**, e01342-18.

3. Weirather,J.L., de Cesare,M., Wang,Y., Piazza,P., Sebastiano,V., Wang,X.-J., Buck,D. and Au,K.F. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res*, **6**, 100.

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

5. 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.

6. Krause,M., Niazi,A.M., Labun,K., Torres Cleuren,Y.N., Müller,F.S. and Valen,E. (2019) tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA*, **25**, 1229–1241.

7. Abebe,J.S., Price,A.M., Hayer,K.E., Mohr,I., Weitzman,M.D., Wilson,A.C. and Depledge,D.P. (2022) DRUMMER-Rapid detection of RNA modifications through comparative nanopore sequencing. *Bioinformatics*, 10.1093/bioinformatics/btac274.

8. Begik,O., Lucas,M.C., Pryszcz,L.P., Ramirez,J.M., Medina,R., Milenkovic,I., Cruciani,S., Liu,H., Vieira,H.G.S., Sas-Chen,A., *et al.* (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol*, 10.1038/s41587-021-00915-6.

9. Nguyen,T.A., Heng,J.W.J., Kaewsapsak,P., Kok,E.P.L., Stanojević,D., Liu,H., Cardilla,A., Praditya,A., Yi,Z., Lin,M., *et al.* (2022) Direct identification of A-to-I editing sites with nanopore native RNA sequencing. *Nat Methods*, **19**, 833–844.

10. Hendra,C., Pratanwanich,P.N., Wan,Y.K., Goh,W.S.S., Thiery,A. and Göke,J. (2022) Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat Methods*, **19**, 1590–1598.

11. Donovan-Banfield,I., Turnell,A.S., Hiscox,J.A., Leppard,K.N. and Matthews,D.A. (2020) Deep splicing plasticity of the human adenovirus type 5 transcriptome drives virus evolution. *Commun Biol*, **3**, 124.

12. Price,A.M., Steinbock,R.T., Lauman,R., Charman,M., Hayer,K.E., Kumar,N., Halko,E., Lum,K.K., Wei,M., Wilson,A.C., *et al.* (2022) Novel viral splicing events and open reading frames revealed by long-read direct RNA sequencing of adenovirus transcripts. *PLOS Pathogens*, **18**, e1010797.

13. Braspenning,S.E., Verjans,G.M.G.M., Mehraban,T., Messaoudi,I., Depledge,D.P. and Ouwendijk,W.J.D. (2021) The architecture of the simian varicella virus transcriptome. *PLoS Pathog*, **17**, e1010084.

14. Braspenning,S.E., Sadaoka,T., Breuer,J., Verjans,G.M.G.M., Ouwendijk,W.J.D. and Depledge,D.P. (2020) Decoding the Architecture of the Varicella-Zoster Virus Transcriptome. *mBio*, **11**, e01568-20.

15. 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. *Nat Commun*, **10**, 754.

16. Whisnant,A.W., Jürges,C.S., Hennig,T., Wyler,E., Prusty,B., Rutkowski,A.J., L'hernault,A., Djakovic,L., Göbel,M., Döring,K., *et al.* (2020) Integrative functional genomics decodes herpes simplex virus 1. *Nat Commun*, **11**, 2038.

17. Dong,X., Du,M.R.M., Gouil,Q., Tian,L., Jabbari,J.S., Bowden,R., Baldoni,P.L., Chen,Y., Smyth,G.K., Amarasinghe,S.L., *et al.* (2023) Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nat Methods*, 10.1038/s41592-023-02026-3.

18. Kovaka,S., Zimin,A.V., Pertea,G.M., Razaghi,R., Salzberg,S.L. and Pertea,M. (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, **20**, 278.

19. Chen,Y., Sim,A., Wan,Y.K., Yeo,K., Lee,J.J.X., Ling,M.H., Love,M.I. and Göke,J. (2023) Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods*, **20**, 1187–1195.

20. Prjibelski,A.D., Mikheenko,A., Joglekar,A., Smetanin,A., Jarroux,J., Lapidus,A.L. and Tilgner,H.U. (2023) Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol*, 10.1038/s41587-022-01565-y.

21. Uhnoo,I., Wadell,G., Svensson,L. and Johansson,M.E. (1984) Importance of enteric adenoviruses 40 and 41 in acute gastroenteritis in infants and young children. *J Clin Microbiol*, **20**, 365–372.

22. Mautner,V., Steinthorsdottir,V. and Bailey,A. (1995) Enteric adenoviruses. *Curr Top Microbiol Immunol*, **199 (Pt 3)**, 229–282.

23. Morfopoulou,S., Buddle,S., Torres Montaguth,O.E., Atkinson,L., Guerra-Assunção,J.A., Moradi Marjaneh,M., Zennezini Chiozzi,R., Storey,N., Campos,L., Hutchinson,J.C., *et al.* (2023) Genomic investigations of unexplained acute hepatitis in children. *Nature*, **617**, 564–573.

24. Price,A.M., Steinbock,R.T., Lauman,R., Charman,M., Hayer,K.E., Kumar,N., Halko,E., Lum,K.K., Wei,M., Wilson,A.C., *et al.* (2022) Novel viral splicing events and open reading frames revealed by long-read direct RNA sequencing of adenovirus transcripts. *PLoS Pathog*, **18**, e1010797.

25. Burgess,H.M., Depledge,D.P., Thompson,L., Srinivas,K.P., Grande,R.C., Vink,E.I., Abebe,J.S., Blackaby,W.P., Hendrick,A., Albertella,M.R., *et al.* (2021) Targeting the m6A RNA modification pathway blocks SARS-CoV-2 and HCoV-OC43 replication. *Genes Dev*, **35**, 1005–1019.

26. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler, and D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

27. Quinlan,A.R. (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*, **47**, 11.12.1-34.

28. Leung,T.K.H. and Brown,M. (2011) Block in entry of enteric adenovirus type 41 in HEK293 cells. *Virus Res*, **156**, 54–63.

29. Kindsmüller,K., Groitl,P., Härtl,B., Blanchette,P., Hauber,J. and Dobner,T. (2007) Intranuclear targeting and nuclear export of the adenovirus E1B-55K protein are regulated by SUMO1 conjugation. *Proc Natl Acad Sci U S A*, **104**, 6684–6689.

30. Smith,M.A., Ersavas,T., Ferguson,J.M., Liu,H., Lucas,M.C., Begik,O., Bojarski,L., Barton,K. and Novoa,E.M. (2020) Molecular barcoding of native RNAs using nanopore sequencing and deep learning. *Genome Res.*, 10.1101/gr.260836.120.

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

32. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

33. 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.e10.

34. Hahne,F. and Ivanek,R. (2016) Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol Biol*, **1418**, 335–351.

35. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol*, **9**, e1003118.

36. Wickham,H. (2016) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag New York.

37. Conway,J.R., Lex,A. and Gehlenborg,N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, **33**, 2938–2940.

38. Price,A.M., Hayer,K.E., McIntyre,A.B.R., Gokhale,N.S., Abebe,J.S., Della Fera,A.N., Mason,C.E., Horner,S.M., Wilson,A.C., Depledge,D.P., *et al.* (2020) Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. *Nat Commun*, **11**, 6016.

39. Burgess,H.M., Grande,R., Riccio,S., Dinesh,I., Winkler,G.S., Depledge,D.P. and Mohr,I. (2023) CCR4-NOT differentially controls host versus virus poly(a)-tail length and regulates HCMV infection. *EMBO Rep*, **24**, e56327.

40. Workman,R.E., Tang,A.D., Tang,P.S., Jain,M., Tyson,J.R., Razaghi,R., Zuzarte,P.C., Gilpatrick,T., Payne,A., Quick,J., *et al.* (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods*, **16**, 1297–1305.

41. Bonfert,T., Kirner,E., Csaba,G., Zimmer,R. and Friedel,C.C. (2015) ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics*, **16**, 122.

42. Falvey,E. and Ziff,E. (1983) Sequence arrangement and protein coding capacity of the adenovirus type 2 'i' leader. *J Virol*, **45**, 185–191.

43. Malone,B., Urakova,N., Snijder,E.J. and Campbell,E.A. (2022) Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat Rev Mol Cell Biol*, **23**, 21–39.

44. St-Jean,J.R., Jacomy,H., Desforges,M., Vabret,A., Freymuth,F. and Talbot,P.J. (2004) Human respiratory coronavirus OC43: genetic stability and neuroinvasion. *J Virol*, **78**, 8824–8834.

45. Yeh,H.Y., Pieniazek,N., Pieniazek,D. and Luftig,R.B. (1996) Genetic organization, size, and complete sequence of early region 3 genes of human adenovirus type 41. *J Virol*, **70**, 2658–2663.

46. Allard,A. and Wadell,G. (1988) Physical organization of the enteric adenovirus type 41 early region 1A. *Virology*, **164**, 220–229.

47. Allard,A. and Wadell,G. (1992) The E1B transcription map of the enteric adenovirus type 41. *Virology*, **188**, 319–330.

48. van Loon,A.E., Gilardi,P., Perricaudet,M., Rozijn,T.H. and Sussenbach,J.S. (1987) Transcriptional activation by the E1A regions of adenovirus types 40 and 41. *Virology*, **160**, 305–307.

49. van Loon,A.E., Ligtenberg,M., Reemst,A.M., Sussenbach,J.S. and Rozijn,T.H. (1987) Structure and organization of the left-terminal DNA regions of fastidious adenovirus types 40 and 41. *Gene*, **58**, 109–126.

50. Ishino,M., Ohashi,Y., Emoto,T., Sawada,Y. and Fujinaga,K. (1988) Characterization of adenovirus type 40 E1 region. *Virology*, **165**, 95–102.

51. Kawaji,H., Lizio,M., Itoh,M., Kanamori-Katayama,M., Kaiho,A., Nishiyori-Sueki,H., Shin,J.W., Kojima-Ishiyama,M., Kawano,M., Murata,M., *et al.* (2014) Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res*, **24**, 708–717.

52. Carninci,P., Kvam,C., Kitamura,A., Ohsumi,T., Okazaki,Y., Itoh,M., Kamiya,M., Shibata,K., Sasaki,N., Izawa,M., *et al.* (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327–336.

53. Ugolini,C., Mulroney,L., Leger,A., Castelli,M., Criscuolo,E., Williamson,M.K., Davidson,A.D., Almuqrin,A., Giambruno,R., Jain,M., *et al.* (2022) Nanopore ReCappable sequencing maps SARS-CoV-2 5′ capping sites and provides new insights into the structure of sgRNAs. *Nucleic Acids Res*, **50**, 3475–3489.

Figures & Figure Legends

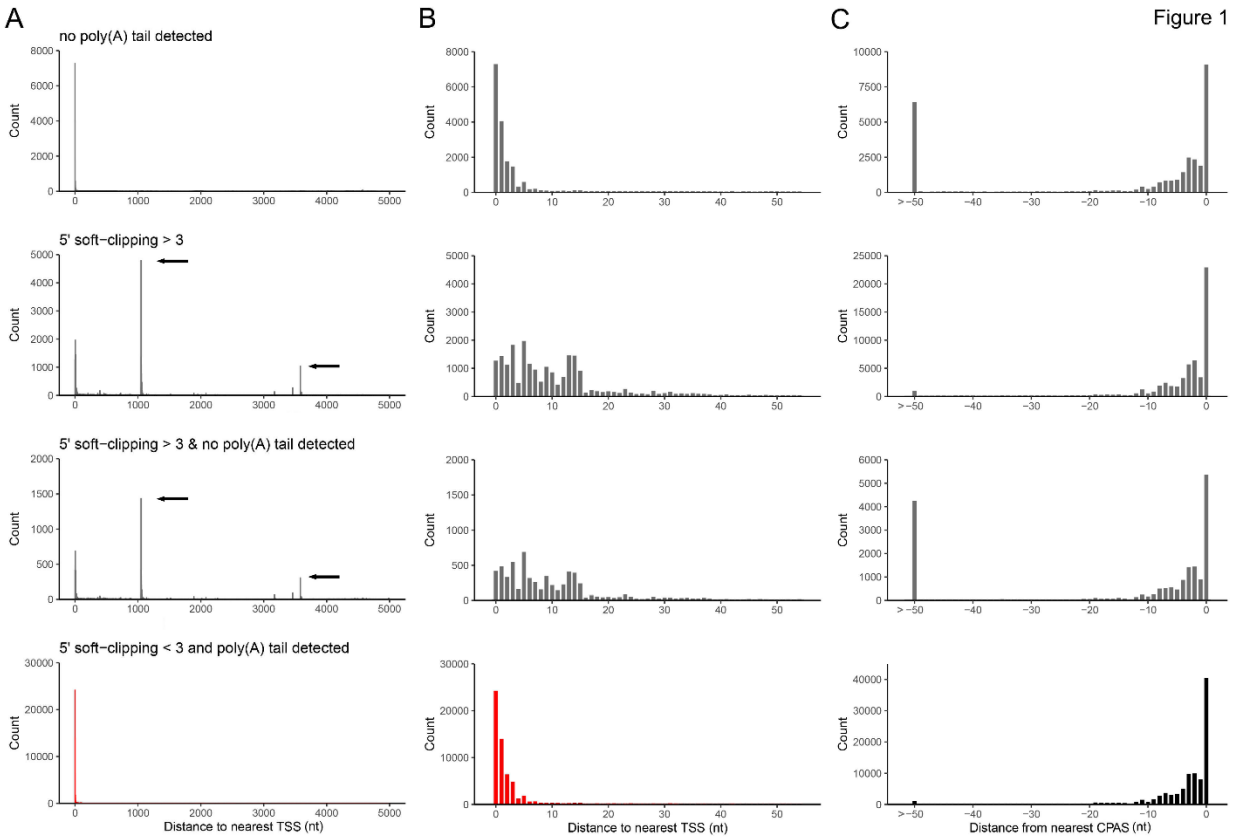

Figure 1: Characteristics of nanopore DRS alignments. Alignments of adenovirus type 5 DRS reads were segregated according to the presence/absence of detectable poly(A) tails and the presence of soft-clipping values > 3 at the 5' end. **(A-C)** The genomic location and read count of **(A-B)** 5' alignment ends relative to previously defined transcription start sites (TSS) or **(C)** 3' alignment ends relative to previously defined cleavage and polyadenylation sites (CPAS) were determined for each of four conditions (no poly(A) tail detected, 5' soft-clipping > 3, no poly(A) tail & 5' soft-clipping > 3, and poly(A) tail & 5' soft-clipping ≤ 3) across windows of **(A)** 5000 nt and **(B-C)** 50 nt. Black arrows indicate the location of artifact TSS derived from misalignment across splice junctions.

Figure 2

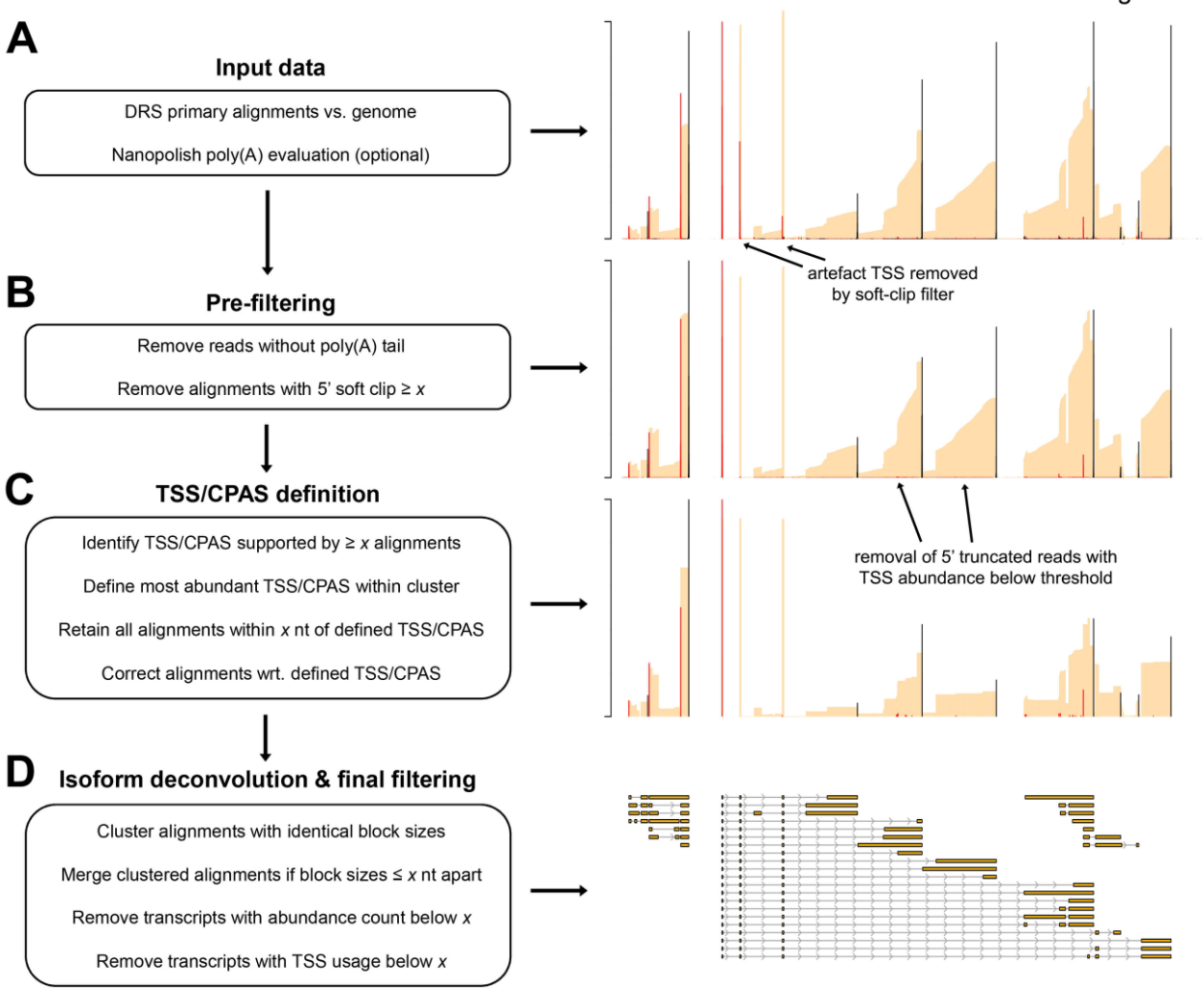

Figure 2: Overview of the NAGATA methodology. (A) DRS genome alignments (beige) that are filtered to retain only primary mappings and (optionally) nanopolish poly(A) output files are used as input for NAGATA. Putative Transcription Start Sites (TSS, red) and Cleavage and Polyadenylation Sites (CPAS, black) are defined (TSS/CPAS definition) by counting the number of alignments with identical 5' (TSS) or 3' (CPAS) ends. **(B)** Pre-filtering removes alignments with 5' soft-clipping values greater than a specified value and optionally removes read alignments for which poly(A) tails are not detected by nanopolish. **(C)** TSS and CPAS are defined (TSS/CPAS definition) by counting the number of alignments with identical 5' (TSS) or 3' (CPAS) ends and considering only those exceeding a specified count as valid. For TSS/CPAS that pass this threshold, all neighboring TSS/CPAS within a defined distance are retained and their co-ordinates adjusted to the dominant TSS/CPAS position. At this stage, transcription units (TU) are defined and all alignments sharing the same CPAS are considered part of the same TU (i.e. transcripts

with differing TSS but the same CPAS are considered part of the same TU). **(D)** For each resulting TU, transcript isoform deconvolution and final filtering is performed by first collapsing alignments if they share the same blockSize and blockStarts distribution and only those exceeding a specified count are considered valid. Alignments with similar blockSize/blockStart values (typically within 1-3 nt) are merged prior to filtering based on abundance counts. Finally, NAGATA applies a filter to remove transcripts with a TSS usage below a defined fraction.

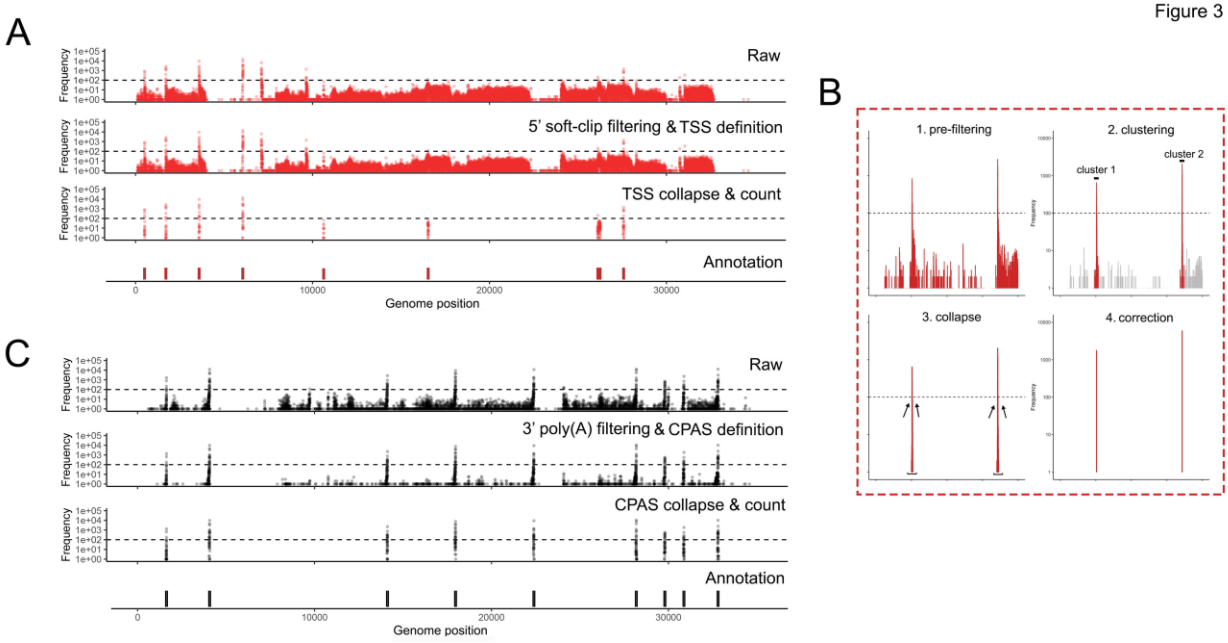

Figure 3: Prefiltering at 5' and 3' ends define robust TSS and CPAS. NAGATA masks alignments with 5' soft-clipping values above a user-defined value (default = 3) and, where nanopolish poly(A) output files are available, also removes alignment derived from reads without detectable poly(A) tails. **(A)** Visualization of TSS positions contained in the raw alignment (top), prefiltered alignments i.e., post 5' soft-clipping and 3' poly(A) tail filtering (second row), post-collapse of neighboring TSS (third row), and the existing TSS/CPAS annotation position (bottom row). **(B)** A specific close-up showing the pre-filtering, clustering, collapse, and correction steps. **(C)** Same as (A) but for CPAS.

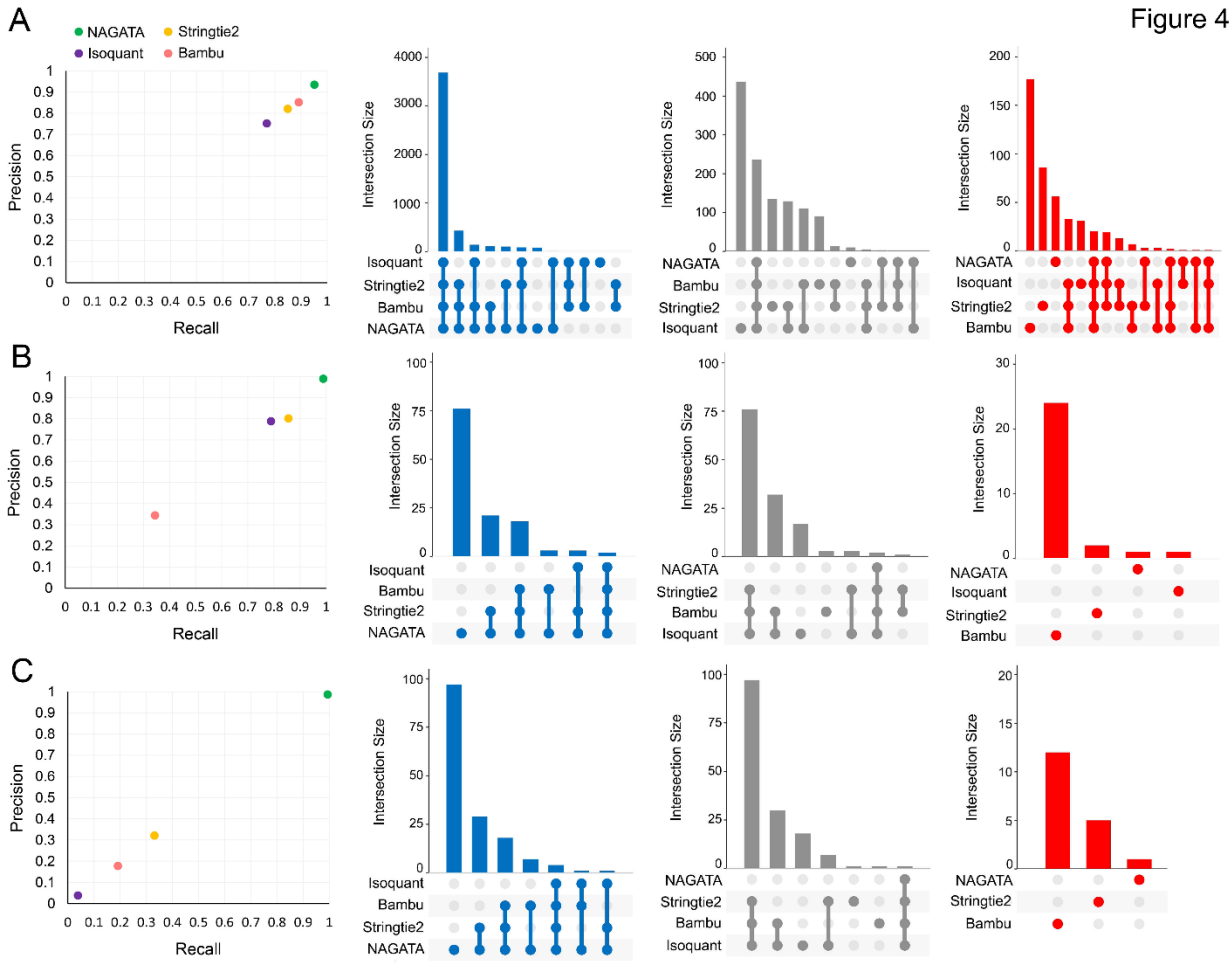

Figure 4: Benchmarking NAGATA using synthetic datasets. Scatter plots comparing precision and recall values for NAGATA and three additional softwares (StringTie2 (18), Isoquant (20), and Bambu (19)) capable of *de novo* transcriptome annotation are supported by UpSet plots denoting the breakdown of true-positives (blue), false-negatives (grey), and false-positives (red) are shown for **(A)** *H. sapiens* HG38 assembly chromosome 1, **(B)** Adenovirus type 5 (HAdV-C5), and **(C)** varicella-zoster virus (VZV) datasets.

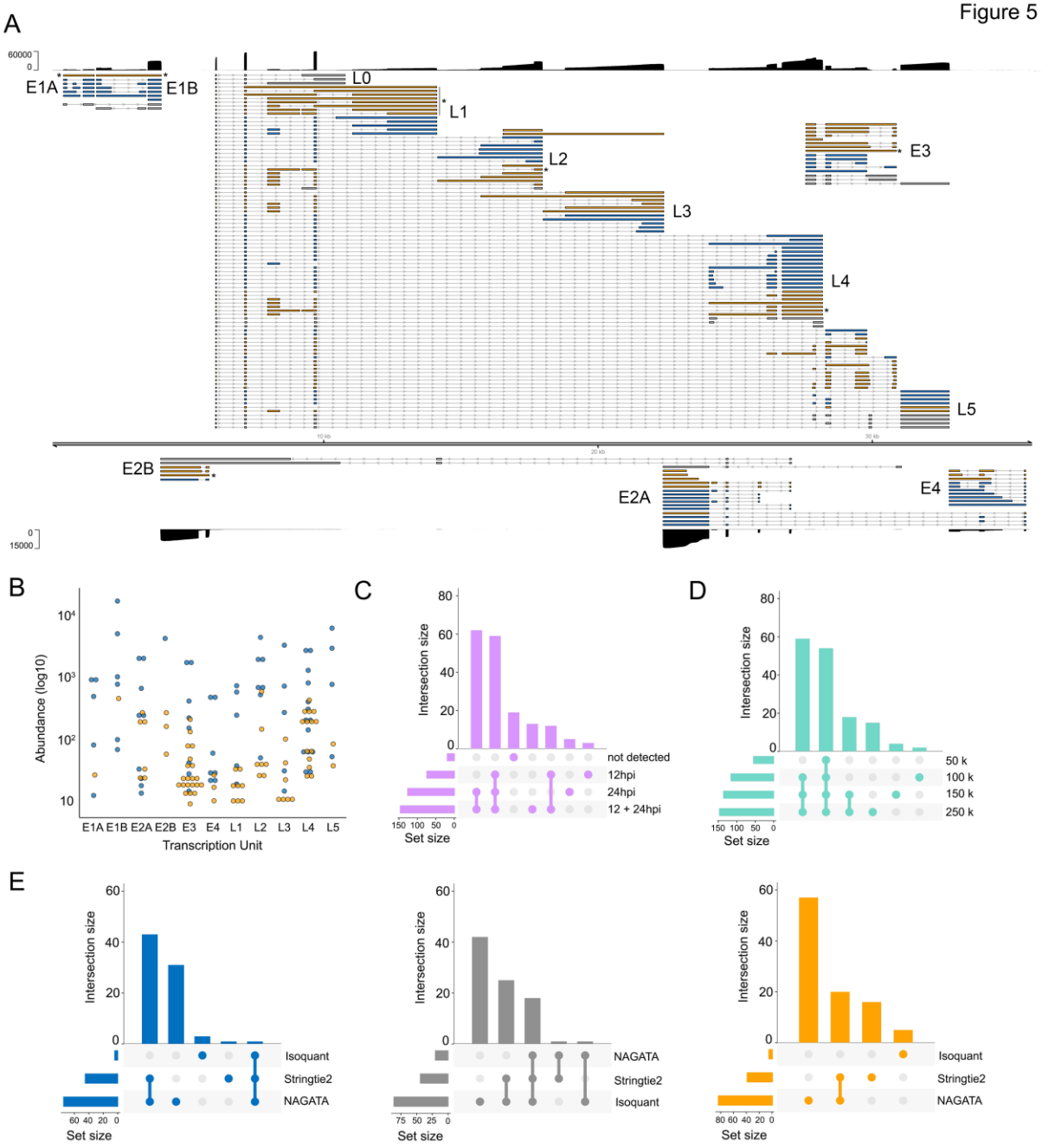

Figure 5: Reconstruction of the adenovirus type 5 transcriptome. (A) Schematic depicting the HAdV-C5 transcriptome as constructed by NAGATA using merged DRS datasets representing 12 and 24 hours post infection (hpi) of A549 cells. Read coverage is shown for each strand (black) with the y-axis denoting read depth. Major transcription units (e.g. E1A) are indicated in black text while individual transcripts are coloured according to classification (orange = not present in existing annotation, blue = present in existing annotation, grey = reported in existing annotation but not detected by NAGATA in this dataset). Wide and thin boxes indicate

canonical CDS domains and UTRs, respectively. Black asterisks denote transcripts putatively classified as unspliced pre-mRNAs. **(B)** For each detected transcript in each transcription unit, a raw abundance count was generated using NAGATA and colour-coded according to transcript classification. **(C)** Upset plot denoting the number of transcripts reported by NAGATA in the individual 12 hpi and 24 hpi datasets and merged 12 + 24 hpi dataset. **(D)** Upset plot denoting the number of transcripts reported by NAGATA in the merged 12 + 24 hpi dataset after random subsampling of reads. **(E)** Upset plots denoting the number of transcripts reported by NAGATA, StringTie2 (18), and Isoquant (20), segregated according to (blue) overlaps with existing annotation, (grey) not detected, and (orange) not present in original annotation.

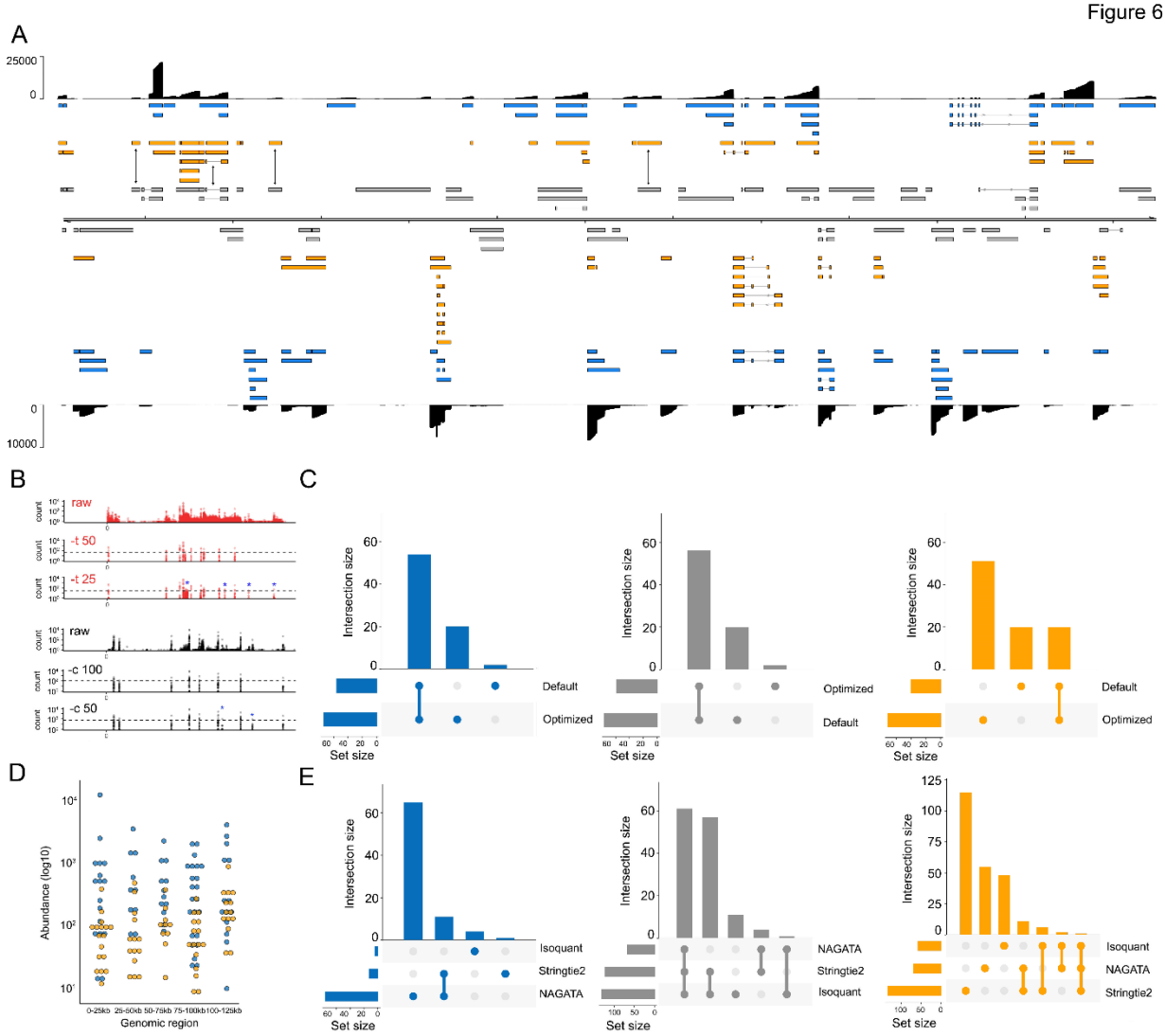

Figure 6: Reconstructing the varicella zoster virus transcriptome. (A) Schematic depicting the VZV strain Dumas transcriptome as constructed by NAGATA using a single previously published DRS dataset (14). Read coverage is shown for each strand (black) with the y-axis denoting read depth. Transcripts are coloured according to classification (orange = not recorded in existing annotation, blue = recorded in existing annotation, grey = reported in existing annotation but not detected by NAGATA in this dataset). Wide and thin boxes indicate canonical CDS domains and UTRs, respectively. **(B)** Close-up of the forward strand TSS (red) and CPAS (black) pileups across first 25kb of the VZV genome. Tracks shown include the raw (prefiltered) data and the effect of filtering using different values for TSS (–t) and –c. **(C)** Upset plot denoting the number of transcripts reported by NAGATA using the default and VZV-optimized configurations. **(D)** For each detected transcript in a given transcription unit, a raw abundance

count was generated using NAGATA and colour-coded according to transcript classification. For simplicity, the dot plot is divided into 25 kb windows. **(E)** Upset plots denoting the number of transcripts reported by NAGATA, StringTie2 (18), and Isoquant (20), segregated according to (blue) overlaps with existing annotation, (grey) not detected, and (red) not present in original annotation.

Figure 7

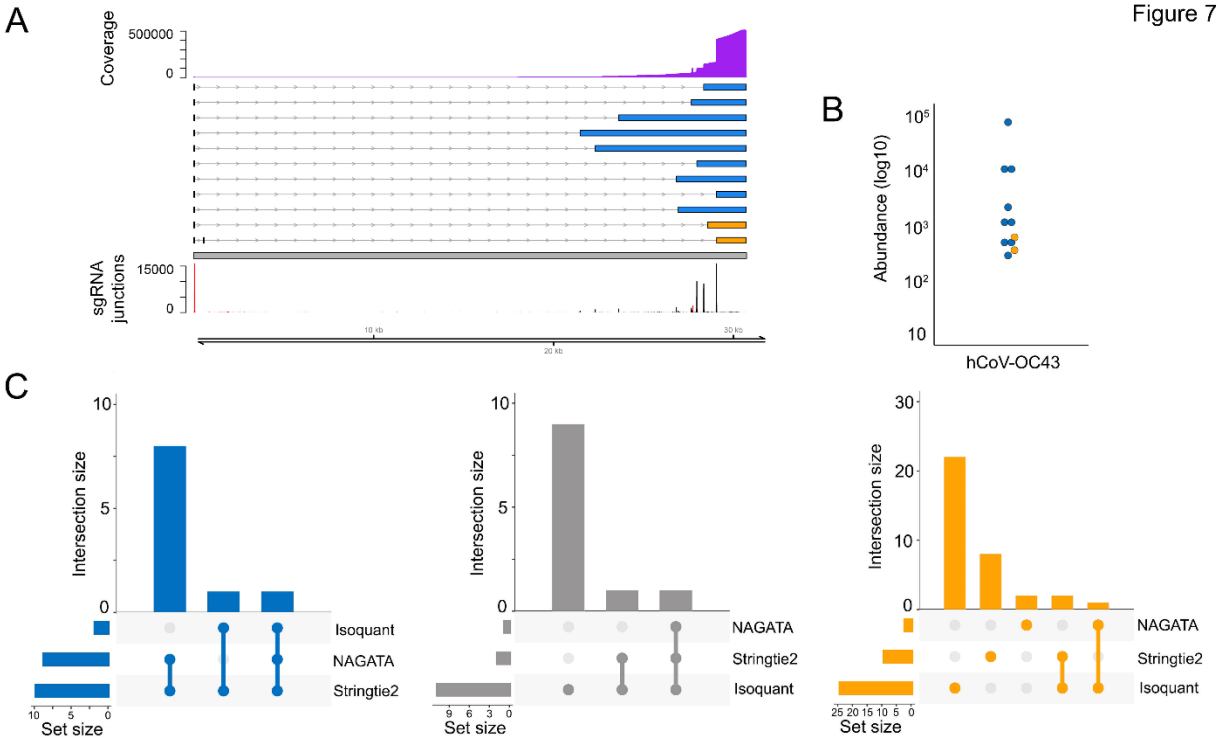

Figure 7: Reconstructing the hCoV-OC43 transcriptome. (A) Schematic depicting the hCoV-OC43 transcriptome as constructed by NAGATA using a single previously published DRS dataset (25). Read coverage is shown (purple) with the y-axis denoting read depth while the locations and abundances of sgRNA 5' (red) and 3' (black) junctions are also shown. Transcripts are coloured according to classification (orange = not recorded in existing annotation, blue = recorded in existing annotation, grey = reported in existing annotation but not detected by NAGATA in this dataset). Wide and thin boxes indicate canonical CDS domains and UTRs, respectively. **(B)** For each transcript, a raw abundance count was generated using NAGATA and colour-coded according to transcript classification. **(C)** Upset plots denoting the number of transcripts reported by NAGATA, Stringtie2, and Isoquant, segregated according to (blue) overlaps with existing annotation, (grey) not detected, and (red) not present in original annotation.

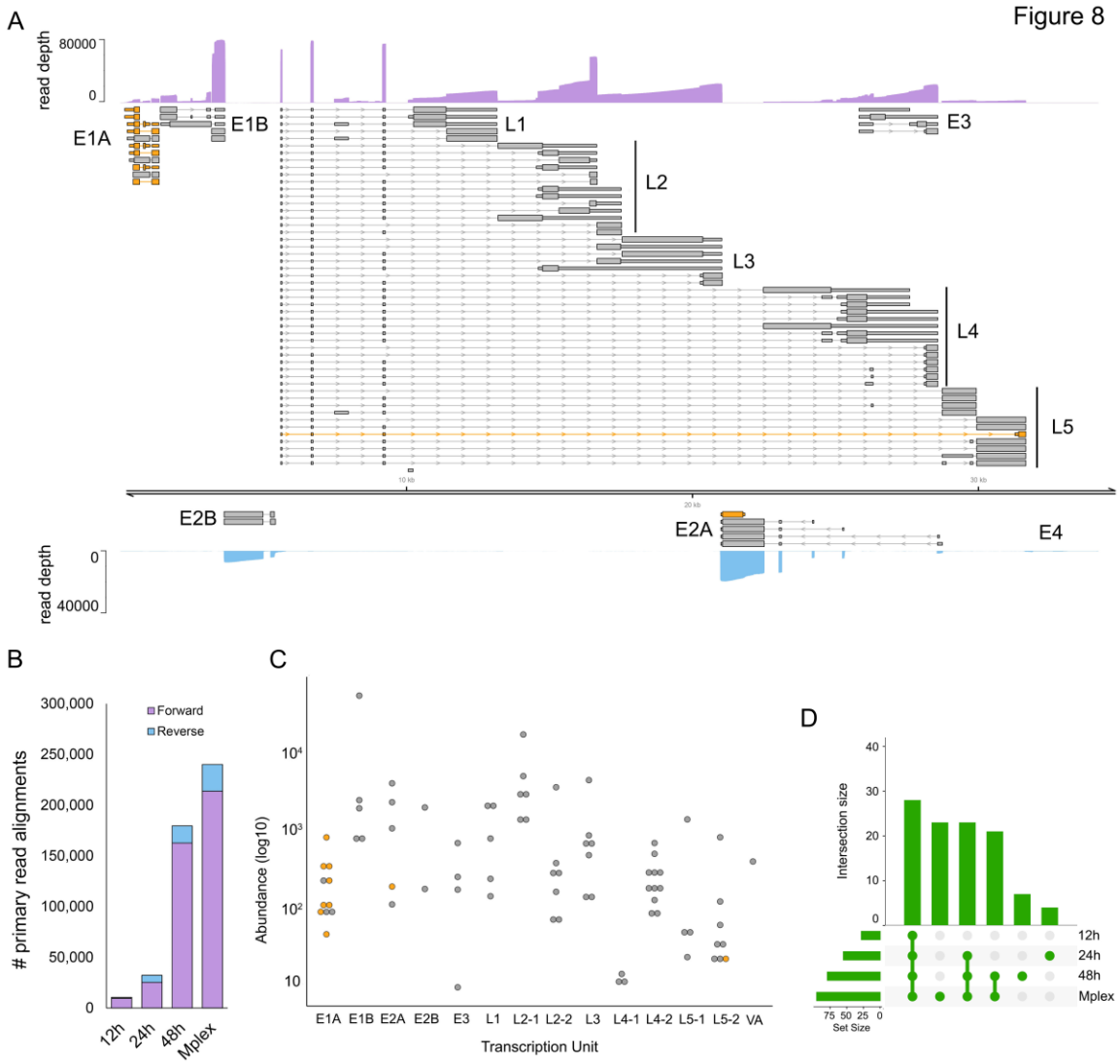

Figure 8: Annotation of the hAdV41 transcriptome. (A) The reannotated hAdV41 transcriptome encodes 77 transcripts, of which 67 encode 23 canonical ORFs with their UTRs defined (grey). Wide and thin boxes indicate canonical CDS domains and UTRs, respectively. A further 9 transcripts (pink), putatively encoded novel or *N*' terminal truncated protein isoforms. Nanopore DRS coverage plots (purple) are shown for the combined (Mplex) dataset (12 + 24 + 48 hpi) in a strand-specific manner. Y-axis values indicate the read depth. **(B)** The number of hAdV41 read alignments recorded against the forward and reverse strands, separated by dataset, show a strong bias toward transcription from the forward strand. **(C)** For each detected transcript in each transcription unit, a raw abundance count was generated using NAGATA and colour-coded according to transcript classification. **(D)** Upset plots denoting the number of transcripts reported by NAGATA in each of the individual, as well as the merged, datasets.

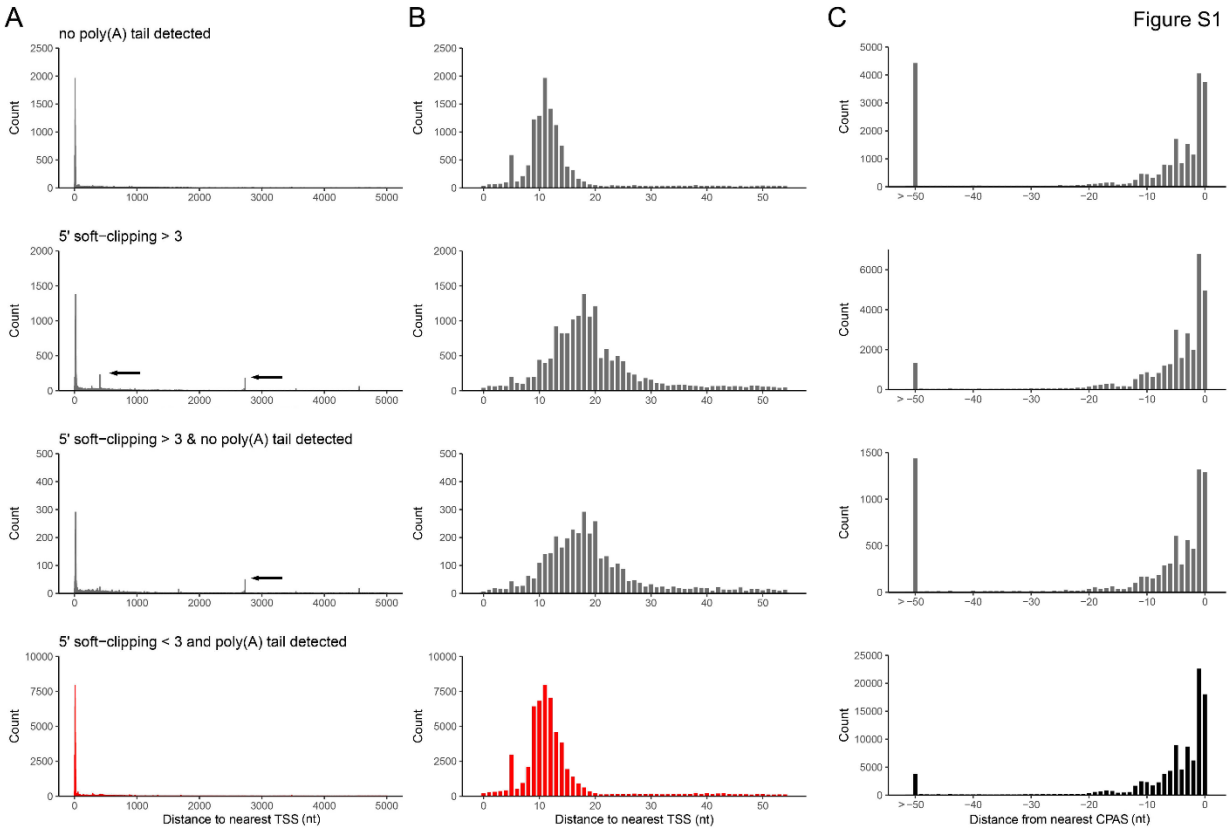

Figure S1: Further characteristics of nanopore DRS alignments. Alignments of varicella zoster virus DRS reads were segregated according to the presence/absence of detectable poly(A) tails and the presence of soft-clipping values > 3 at the 5' end. **(A-C)** The genomic location and read count of **(A-B)** 5' alignment ends relative to previously defined transcription start sites (TSS) or **(C)** 3' alignment ends relative to previously defined cleavage and polyadenylation sites (CPAS) were determined for each of four conditions (no poly(A) tail detected, 5' soft-clipping > 3, no poly(A) tail & 5' soft-clipping > 3, and poly(A) tail & 5' soft-clipping ≤ 3) across windows of **(A)** 5000 nt and **(B-C)** 50 nt. Black arrows indicate the location of artifact TSS derived from misalignment across splice junctions.

Figure S2

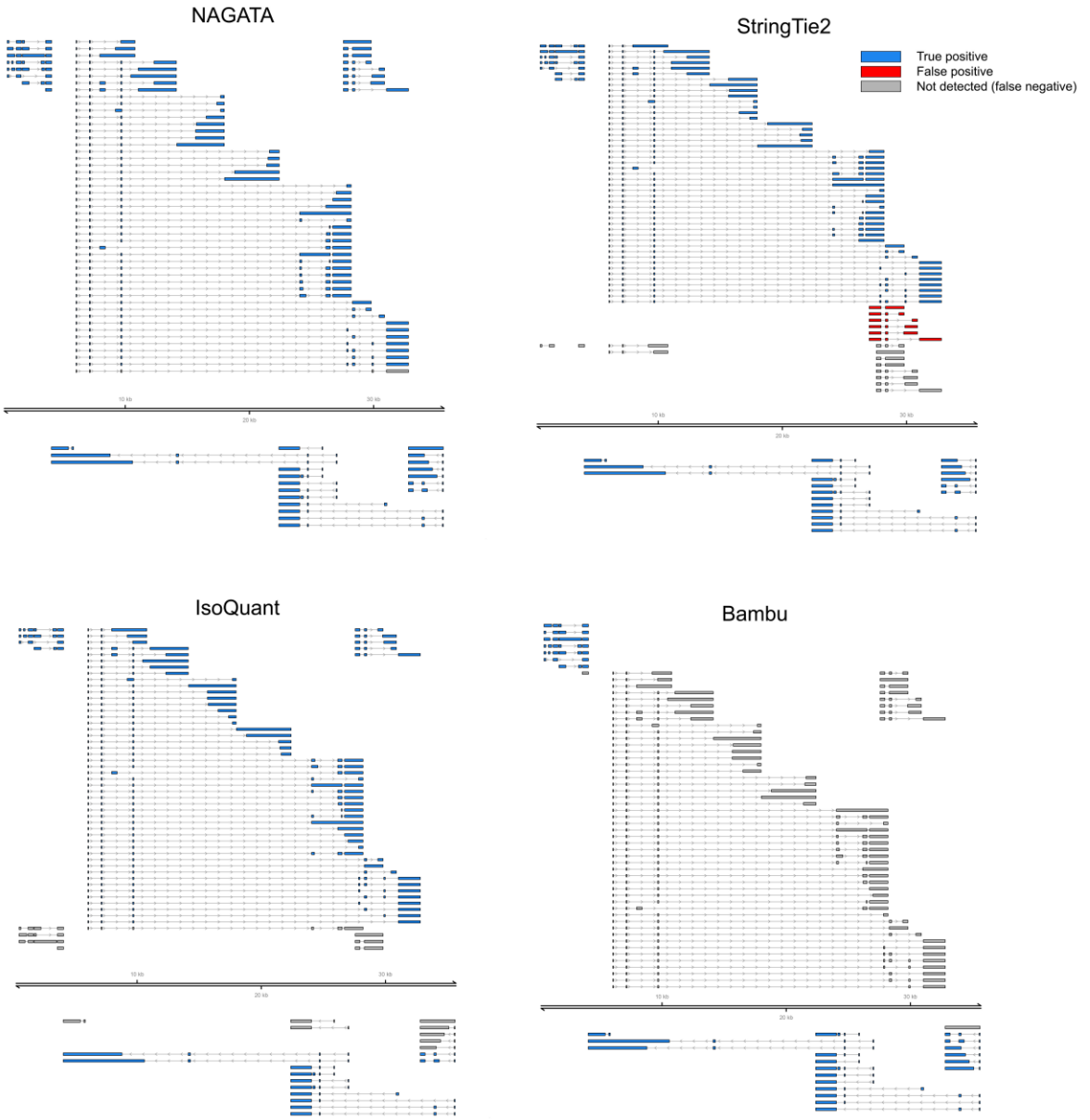

Figure S2: Adenovirus type 5 transcriptome reconstructions from synthetic datasets. The reconstruction of the HAdV-C5 transcriptome from synthetic datasets was performed using NAGATA, StringTie2 (18), Isoquant (20), and Bambu (19). Transcript isoforms are coloured according to status: true positive (blue), false positive (red), and false negative (grey).

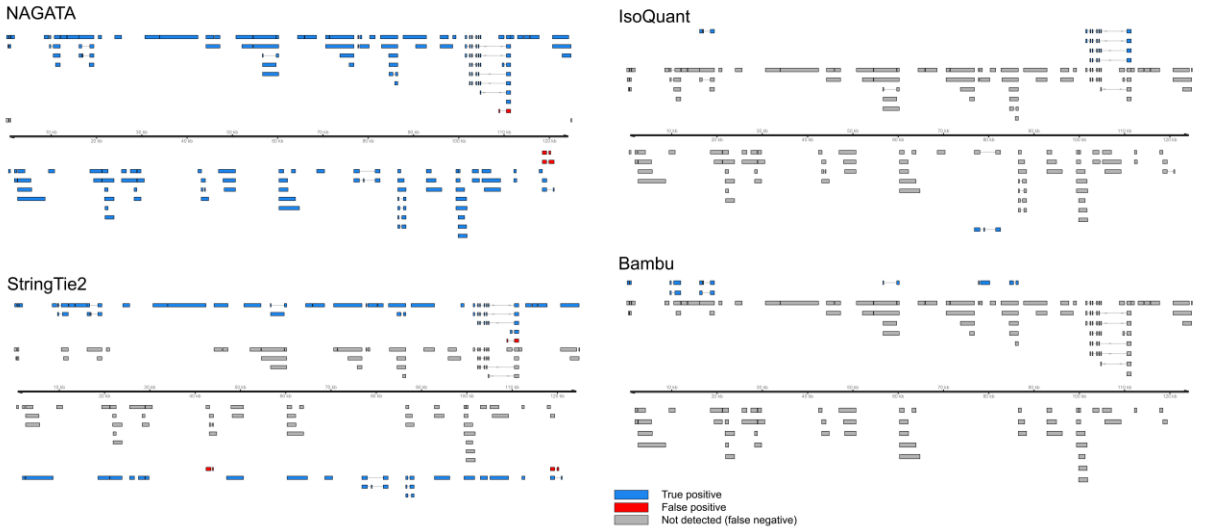

Figure S3: Varicella Zoster Virus transcriptome reconstructions from a synthetic dataset. The reconstruction of the VZV transcriptome from synthetic datasets was performed using StringTie2 (18), Isoquant (20), and Bambu (19). Note that Bambu (19) produced no true or false positives and thus is excluded. Transcript isoforms are coloured according to status: true positive (blue), false positive (red), and false negative (grey).

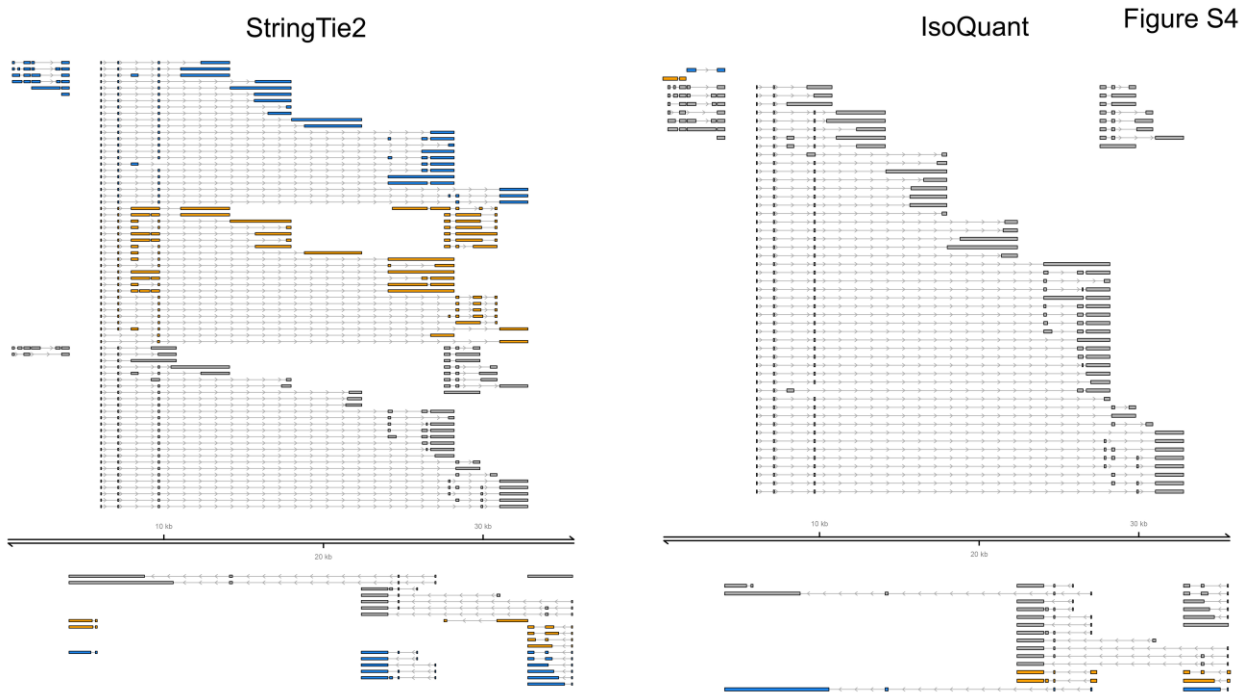

Figure S4: Adenovirus type 5 transcriptome reconstructions using a real DRS dataset. The reconstruction of the HAdV-C5 transcriptome was performed using NAGATA, StringTie2 (18), and Isoquant (20). Transcript isoforms are coloured according to status: overlap with existing annotation (blue), transcript not previously reported (red), and annotated transcript not detected (grey).

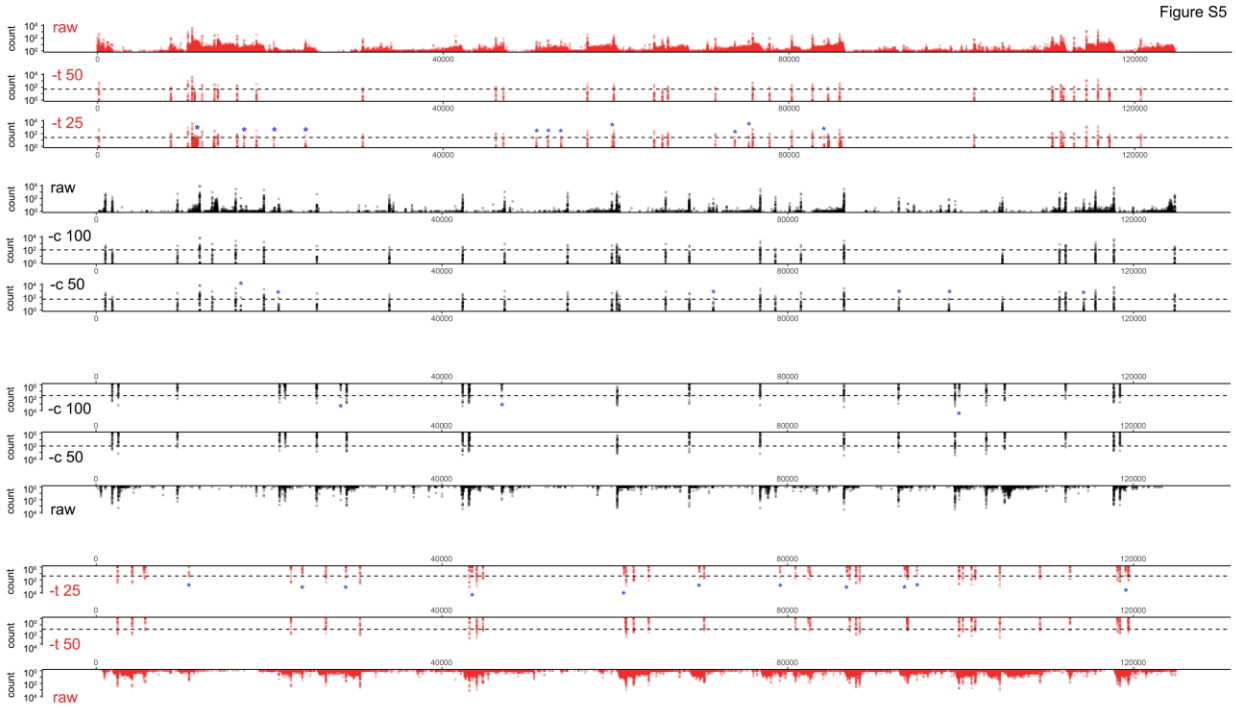

Figure S5: Analysis of TSS/CPAS identification using default and VZV-optimized NAGATA settings. TSS (red) and CPAS (black) pileups across both strands of the VZV genome are shown. Tracks include the raw (prefiltered) data and the effect of filtering using different values for –t and –c. TSS/CPAS that are identified in the optimized (-t 25, -c 50) run only are highlighted with a blue asterisk.

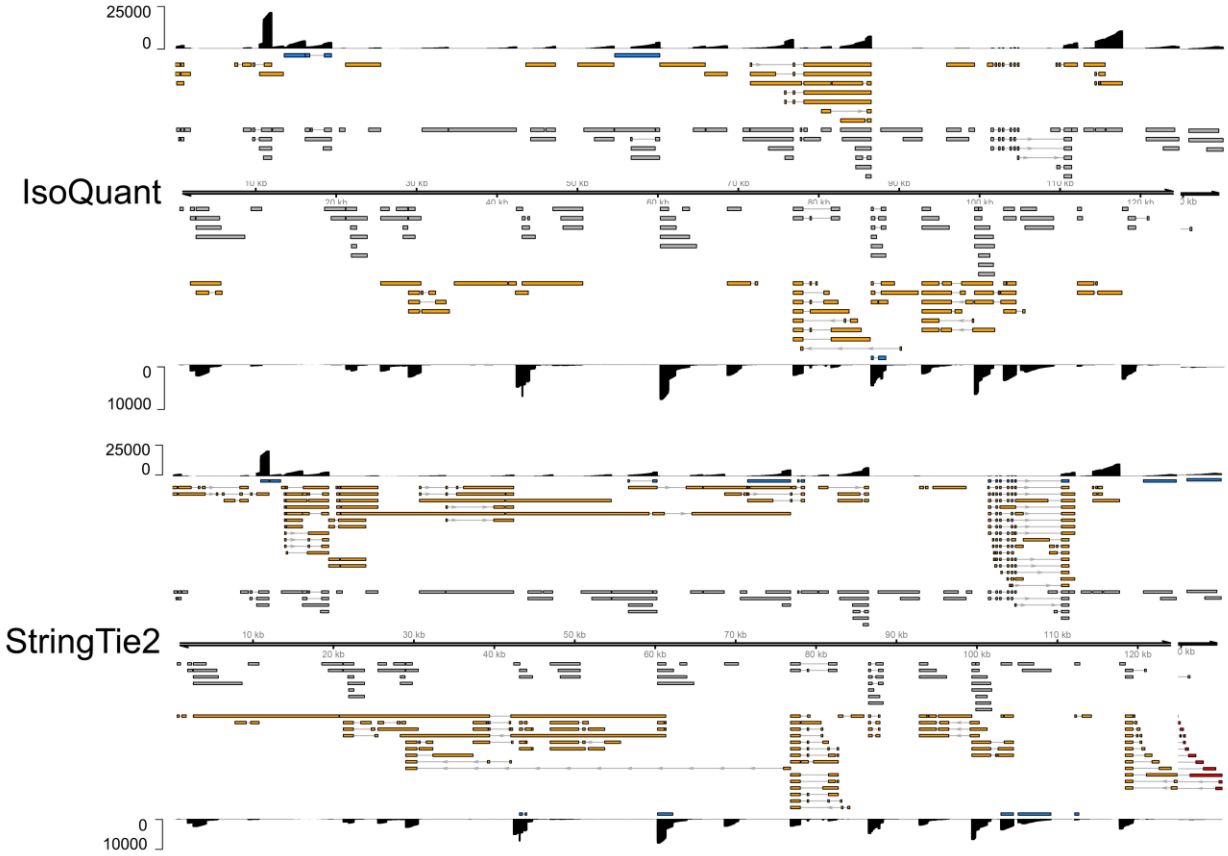

Figure S6: Varicella Zoster Virus transcriptome reconstructions using a real DRS dataset. The reconstruction of the VZV transcriptome was performed using NAGATA (Fig. 6), StringTie2 (18), and Isoquant (20). Transcript isoforms are coloured according to status: overlap with existing annotation (blue), transcript not previously reported (orange), and annotated transcript not detected (grey).