Title page

NanoTrans: an integrated computational framework for comprehensive transcriptome analysis with Nanopore direct RNA sequencing

Ludong Yang[1,^], Xinxin Zhang[1,^], Fan Wang[1,2,^], Li Zhang[1,*], Jing Li[1,*], Jia-Xing Yue[1,*]

[1]State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, China.

[2]Department of Medical Oncology, The Affiliated Huai'an No.1 People's Hospital of Nanjing Medical University, Huai'an, China.

^: contribute equally.
*: corresponding authors.

* E-mail: zhangli@sysucc.org.cn (LZ); lijing3@sysucc.org.cn (JL); yuejiaxing@gmail.com (JXY)

E-mail and ORCID for all authors:
Ludong Yang > ld_yang@outlook.com, 0009-0007-6684-5689
Xinxin Zhang > zhangxx@sysucc.org.cn, 0000-0002-2986-1002
Fan Wang > wangf357@mail2.sysu.edu.cn, 0000-0001-7420-2025
Li Zhang > zhangli@sysucc.org.cn,
Jing Li > lijing3@sysucc.org.cn, 0000-0001-5697-9527
Jia-Xing Yue > yuejiaxing@gmail.com, 0000-0002-2122-9221

Abstract

Nanopore direct RNA sequencing (DRS) provides the direct access to native RNA strands with full-length information, shedding light on rich qualitative and quantitative properties of gene expression profiles. Here with NanoTrans, we present an integrated computational framework that comprehensively covers all major DRS-based application scopes, including isoform clustering and quantification, poly(A) tail length estimation, RNA modification profiling, and fusion gene detection. In addition to its merit in providing such a streamlined one-stop solution, NanoTrans also shines in its workflow-orientated modular design, batch processing capability, all-in-one tabular and graphic report output, as well as automatic installation and configuration supports. Finally, by applying NanoTrans to real DRS datasets of yeast, *Arabidopsis*, as well as human embryonic kidney and cancer cell lines, we further demonstrated its utility, effectiveness, and efficacy across a wide range of DRS-based application settings.

Introduction

The recent years have witnessed a rapid advancement of long-read sequencing technologies such as PacBio and Oxford Nanopore in terms of both read length and read accuracy. Combined with their single-molecular nature, this opens up exciting opportunities for many applications in both basic and applied biomedical research (Logsdon et al., 2020). Long-read-based RNA sequencing, best represented by Oxford Nanopore's direct RNA sequencing (DRS), has also become increasingly popular, offering rich biological information while avoiding reverse transcription and amplification biases (Garalde et al., 2018). Accordingly, a number of dedicated bioinformatic tools have recently been developed for analyzing Nanopore DRS data with emphasis on specific applications, such as Mandalorion (Byrne et al., 2017) and Flair (Tang et al., 2020) for full-length isoform clustering, NanoCount (Gleeson et al., 2022) for expression quantification, nanopolish (Loman et al., 2015) and tailfindr (Krause et al., 2019) for poly(A) tail length estimation, EpiNano (Liu et al., 2019) and Xpore (Pratanwanich et al., 2021) for RNA modification, LongGF (Liu et al., 2020) and JAFFAL (Davidson et al., 2022) for gene fusion detection. While these tools greatly facilitated Nanopore DRS data analysis in their corresponding application scope, a unified framework that combines the power of these different tools are still badly needed. The Nextflow-based pipeline MasterOfPores is among the first to address this challenge (Cozzuto et al., 2020), but it lacks the support for important analysis such as differential gene/isoform expression detection, alternative splicing event evaluation, and fusion gene identification. More recent, a Snakemake-based pipeline named FASTdRNA (Chen et al., 2023) was developed with additional support for alternative splicing event evaluation, but still lacking important functionalities such as differential gene/isoform expression detection and fusion gene identification. Therefore, while these existing bioinformatic pipelines facilitated the wider adoption of Nanopore DRS, noticeable functional and practical limitations still exist in their current implementations.

In this study, we developed NanoTrans, an alternative solution for one-stop Nanopore DRS data analysis that cover diverse application scopes: expression quantification, isoform characterization, poly(A) tail length profiling, RNA modification evaluation, and fusion gene detection. As demonstrated with various real world testing examples and a head-to-head

benchmarking comparison with MasterOfPores and FASTdRNA, NanoTrans shines in its rich functionality, robust stability, and computational optimization. Therefore, NanoTrans represents a step up in addressing the practical needs of the field, which helps to make the Nanopore DRS technology accessible to a broader research community for taking advantage of its full potentials.

Results

The framework design of NanoTrans

NanoTrans is a Linux-based computational framework for automated high-throughput Nanopore DRS data analysis. NanoTrans is self-contained, with all dependencies automatically installed and configured via a pre-shipped installer script. The design of NanoTrans is workflow-orientated, with a series of task-specific modules numbered according to their processing order (Figure 1 and Figure S1). Briefly, NanoTrans first performs Nanopore reads basecalling and reference genome preprocessing with its two starting modules numbered with “00”. The basecalled fastq reads are subsequently mapped to the preprocessed genome in a splicing-aware manner (module “01”), after which isoform clustering and quantification are further performed accordingly (module “02”). Based on the clustered and quantified isoforms, NanoTrans can perform different application-specific analysis such as isoform expression and splicing comparison (module “03”), isoform RNA modification identification (module “04”), isoform poly(A) tail length profiling (module “05”). In addition, reference-based gene fusion detection (module “06”) can be applied as well. A user-defined master sample table is used for specifying sample list, experimental design, and reads locations, based on which automatic batch processing and between group comparison are natively supported. Finally, an all-in-one final report containing all major tabular and graphic outputs will be generated for easy result exploration (module “07”).

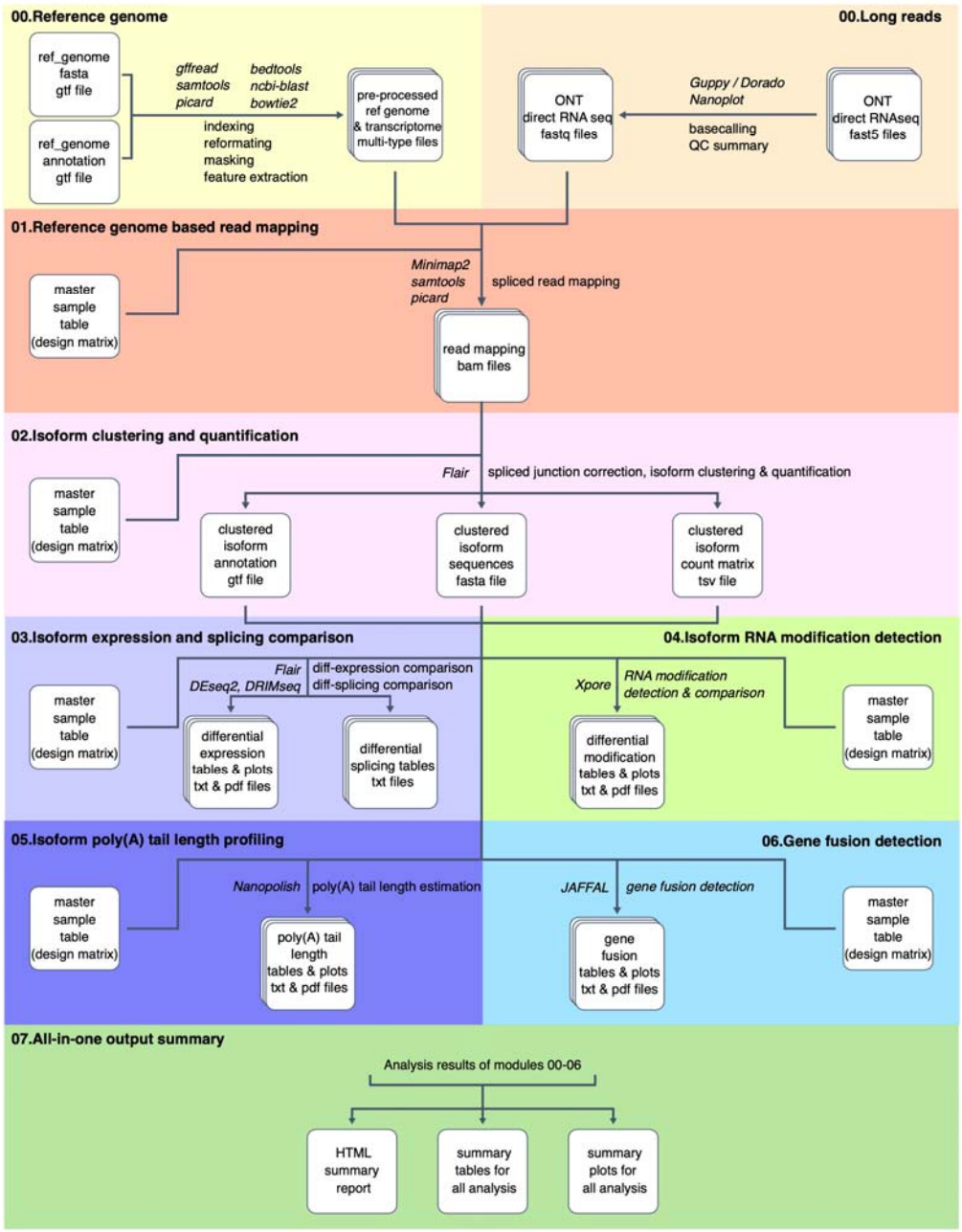


Figure 1. The framework design and application demonstration of NanoTrans. The overview of the NanoTrans framework. The names of third-party tools employed in each step are denoted in italics.

Application demonstration summary of NanoTrans

To demonstrate the application of NanoTrans in real case scenarios, we first applied it to four public Nanopore DRS datasets (Table S1) from the budding yeast *Saccharomyces cerevisiae* (Dataset 1) (Tudek et al., 2021), the mustard plant *Arabidopsis thaliana* (Dataset 2) (Parker et

al., 2020), the human embryonic kidney cell line (HEK293) (Dataset 3) (Chen et al., 2021), and the human lung adenocarcinoma (A549) and leukemia (K562) cancer cell lines (Dataset 4) (Gewartowska et al., 2021), respectively. These datasets were selected to demonstrate the performance of different functional modules of NanoTrans across different application settings (Table S2). With the yeast dataset (Dataset 1), we quantified the expression level and poly(A) tail length of each isoform and replicated the observation from the original study showing a negative correlation between RNA abundance and poly(A) tail length. With the *Arabidopsis* dataset (Dataset 2), we examined differential expression, differential splicing, and RNA modification. With the human embryonic kidney cell line dataset (Dataset 3), we verified the finding of the original study that TENT5A is responsible for the elongation of mRNA poly(A) tails. Finally, with the human cancer cell lines (Dataset 4), we identified a list of gene fusion alterations for the A549 and K562 cell lines respectively, highlighting the frequent chromosomal rearrangements in cancer genomes. Next, we present the performance of NanoTrans using the datasets above in details.

Application demonstration of NanoTrans in *Saccharomyces cerevisiae*

Applying NanoTrans into the yeast dataset (Dataset 1) (Tudek et al., 2021), we quantified the expression level and poly(A) tail length of each isoform. We tried to test whether the outputs by NanoTrans could replicate the observation from the original study showing a negative correlation between RNA abundance and poly(A) tail length.

To do this, the *S. cerevisiae* reference genome (R64-1-1) and its associated feature annotation were retrieved from Ensembl (Table S1) and preprocessed with the 00.Reference_Genome module of NanoTrans for preparing various input files needed for downstream analysis. Then, the raw fast5 reads for the *S. cerevisiae* dataset were retrieved (Table S1) (Tudek et al., 2021) and provided to the 00.Long_Reads module of NanoTrans for basecalling and quality assessment. NanoTrans directly generates summarized tables and plots of this assessment, including read length, read quality, number of reads, read length N50 and so on (Table S3, Figure S2). Afterwards, the basecalled reads were mapped to the preprocessed reference genome via the 01.Reference_Genome_based_Read_Mapping module of NanoTrans. Given

that this sequencing run contains admixed reads derived from both budding yeast and mouse samples, a stringent mapping quality filter of 20 (default: 0) was used at this step to filter out those non-yeast reads. Based on the read-mapping bam file, transcriptome clustering and quantification was further performed with the 02.Isoform_Clustering_and_Quantification module of NanoTrans. In the end, the clustered and quantified *S. cerevisiae* transcriptome isoform set was processed by the 05.Isoform_PolyA_Tail_Length_Profiling module for per-isoform poly(A) length estimation. By combining the quantification and poly(A) tail length profiling results, we examined their relationship. As reported by the original study that generated this Nanopore DRS dataset (Table S2) (Tudek et al., 2021), our analysis with NanoTrans recapitulated the negative relationship between the gene expression level and poly(A) tail length of each isoform (Figure 2A).

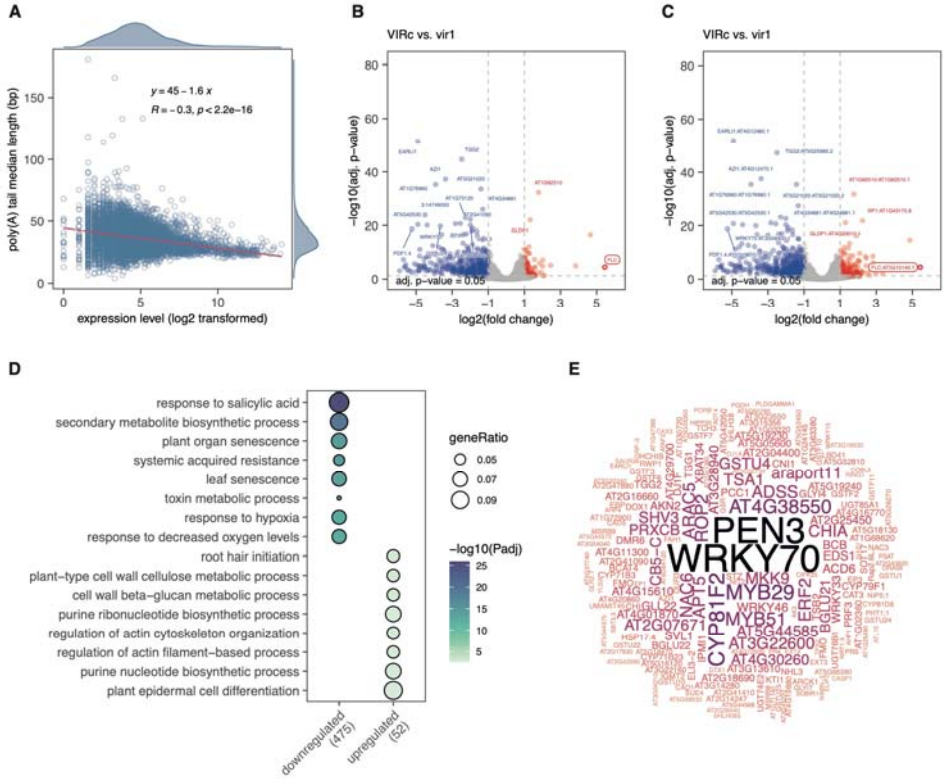


Figure 2. Applications of NanoTrans for isoform quantification and isoform expression comparison. (A) The negative correlation between isoform expression levels and poly(A) tail lengths identified in Dataset 1; the distribution of each isoform's expression level (log2 transformed) is plotted along the x-axis while the distribution of each isoform's median poly(A) length is plotted along the y-axis. (B-C) The differentially expressed genes (B) and isoforms (C) identified based on the comparison between the VIRc and *vir1* groups. False discovery rate (FDR) cutoff: 0.05. Fold change (log2 transformed) cutoffs: -1 and 1. (D)

Biological processes enriched by differentially expressed genes (VIRc vs. *vir1*, adjusted p value < 0.05). (E) The size of words showed the frequencies of DEGs in the enriched GO terms; for example, *WRKY70* has the highest frequency among the enriched terms.

Application demonstration of NanoTrans in *Arabidopsis thaliana*

To further demonstrating the performance of NanoTrans for analyzing differential expression, differential splicing, and RNA modification, we applied the corresponding modules in an *Arabidopsis thaliana* dataset (Dataset 2) (Parker et al., 2020). This dataset includes nanopore DRS reads of *VIRILIZER* loss of function mutant (*vir1*) and its full-function control (VIR complemented; VIRc), in which *VIRILIZER* is a conserved m^6A writer complex component. Both the mutant and control lines had four biological replicates.

We first retrieved all the raw fast5 reads and used 00.Long_Reads module of NanoTrans for basecalling and quality assessment (Table S4). The basecalled reads were mapped to the preprocessed reference genome (TAIR10) via the 01.Reference_Genome_based_Read_Mapping module of NanoTrans with default settings. Based on the read-mapping bam file, transcriptome clustering and quantification was further performed with the 02.Isoform_Clustering_and_Quantification module of NanoTrans by leveraging all eight samples.

Next we used the 03.Isoform_Expression_and_Splicing_Comparison module of NanoTrans to perform differential expression and splicing analysis between the two comparison groups: *vir1* and VIRc. Since both comparison groups have ≥ 2 replicates, both gene-level and isoform-level differential expression can be assessed, and the batch effect correction is automatically taken care of behind scenes. The differentially expressed genes and isoforms are reported in tab-separated tables (sorted by p-value) and further visualized with volcano plots (Figure 2B-C). In such volcano plots, differentially expressed genes are labeled with "gene_name" only (Figure 2B), while differentially expressed isoforms are labeled with "gene_name:isoform_id" tags (Figure 2C). NanoTrans recaptured the differential gene expression signal on the *FLC* (*AT5G10140*) gene reported by the original study (Table S2), highlighting the important impact of the m^6A machinery on flowering time regulation in

Arabidopsis. More interestingly, we made a previously unreported finding that the *WRKY70* gene is among the strongly differentially expressed genes with the lowest p-values, suggesting its change in expression heavily linked to the interruption of the m^6A machinery. Gene ontology (GO) enrichment of the differentially expressed genes further highlighted *WRKY70*'s prevalence across significantly enriched biological processes (Figure 2D-E). This finding is consistent with a recent report that m^6A regulates the normal development in *Arabidopsis* by destabilizing the senescence-related transcripts, including *WRKY70* (Sheikh et al., 2024).

Regarding the differential splicing analysis, four types of alternative splicing events are examined with NanoTrans: intron retention (ir), alternative 3' splicing (alt3), alternative 5' splicing (alt5), and cassette exons (es). A summary table with differentially spliced isoforms (sorted by p-values) is reported. The detailed isoform splicing forms and usage across all samples can be further visualized using the NanoTrans.02.Plot_Isoform_Usage.sh script in the 02.Isoform_Clustering_and_Quantification module of NanoTrans. In Figure 3, we showed an example of such visualization, in which a novel isoform and two alternative splicing events were identified for the *Arabidopsis POM1* (*AT1G05850*) gene (Figure 3A). Samples from the two compared groups show considerable differences in their isoform usages, with the *vir1* group showing uniformly less usage for this novel transcript (Figure 3B). Furthermore, the complexity and alternative splicing of transcripts were also achievable through Module 02 of NanoTrans. For example, NanoTrans accurately identified the mutually exclusive exons between exon 2 and exon 3 in the *FLM* (*AT1G77080*) gene (Figure 3C). Secondly, it can precisely reveal highly complex transcript, such as AT1G48090.6 (Figure 3D). These two instances were also highlighted in the original study that produced this DRS dataset (Parker et al., 2020).

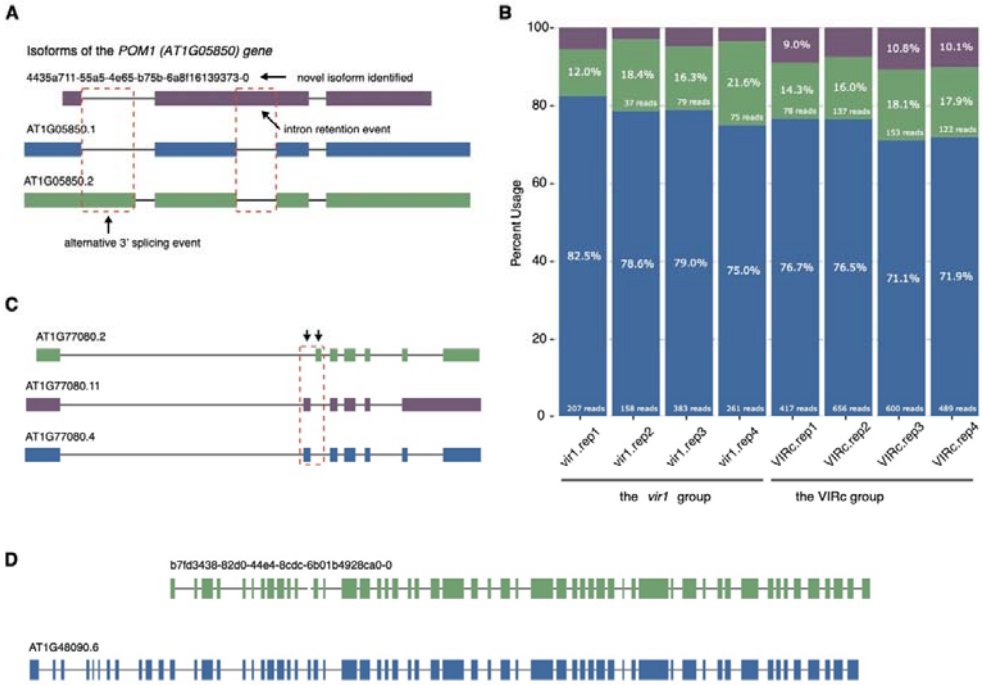


Figure 3. Examples of differentially spliced genes identified based on the comparison between the VIRc and *vir1* groups. (A) A novel transcript that is previously unannotated is identified for the *POM1* (*AT1G05850*) gene. The two identified intron retention events of this novel transcript are further highlighted based on its comparison to the two already annotated transcripts. (B) The different isoform usages of the three transcripts identified in panel A across all samples from the *vir1* and VIRc groups. (C) The mutually exclusive exons (exon 2 and exon 3, indicated by black arrows) in *FLM* (*AT1G77080*) gene is highlighted. (D) A complex transcript composed of a multitude of exons, AT1G48090.6, is identified.

We also used the 04.Isoform_RNA_Modification_Identification module of NanoTrans to evaluate transcript-specific kmers (5-mers) with differential RNA modification rates between the two defined comparison groups: *vir1* and VIRc (Figure 4A-B). Furthermore, the characteristics of differentially modified kmers can be explored and visualized using the script NanoTrans.04.Plot_characteristics_kmers.sh within the analysis module 04 of the NanoTrans framework. Specifically, the differentially modified kmers exhibited characteristics comparable to the original study, such as predominant distribution in 3' untranslated regions (UTRs), a pronounced peak downstream of stop codons, and substantial overlaps between the identified kmers and polyadenylation sites (PASs) (Figure 4C-G). Taken together, these observations recapitulated the characteristics of differentially modified kmers identified in

original study of this dataset (Parker et al., 2020).

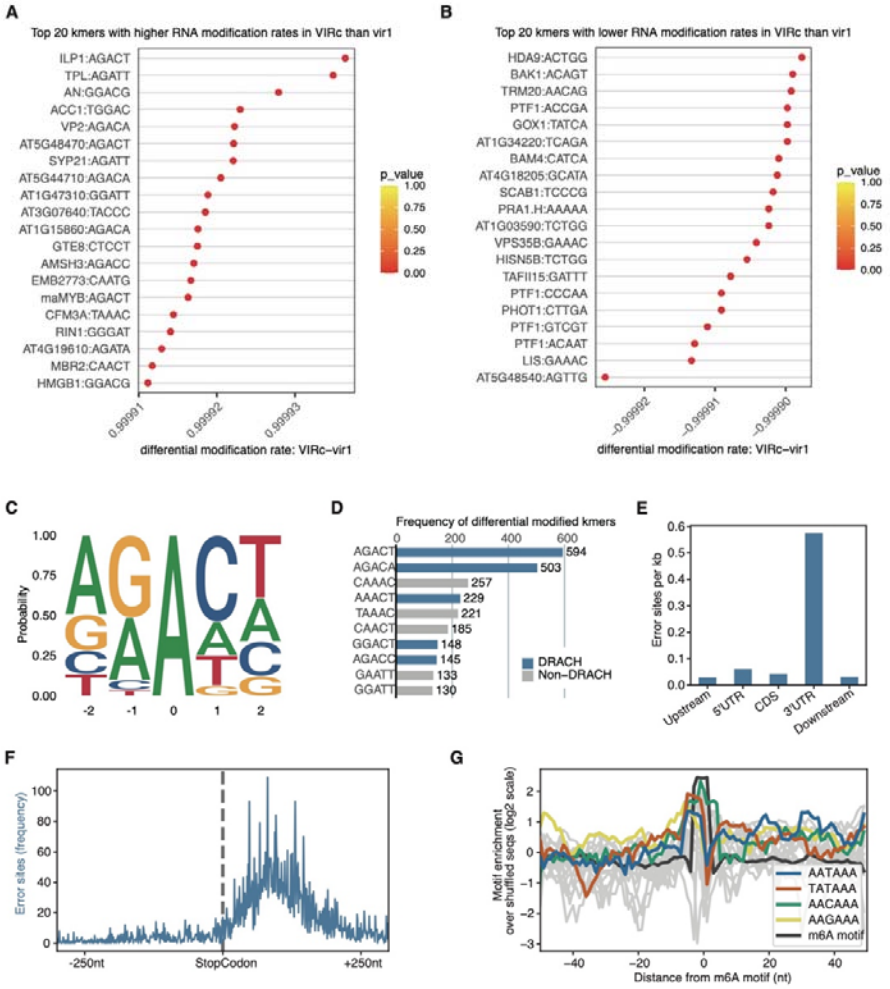


Figure 4. Characteristics of differentially modified kmers between the VIRc and *vir1* groups. The top 20 kmer-gene combinations with higher (A) or lower (B) RNA modification rates in the VIRc group than the *vir1* group. (C) The sequence logo showed the preference of filtered differentially modified kmers (DMKs, differential modification rates > 0.5, p-value < 0.05, and only kmers with center A were retained). (D) The top 10 DMKs with the highest frequency. (E) Distribution of the DMKs on distinct genomic features. (F) Frequency of DMKs around stop codons. (G) Overlaps between RNA modifications and poly(A) signal sites. NanoTrans directly outputs this kind of plot for each analyzed sample.

Application demonstration of NanoTrans in human embryonic kidney cell line

Dataset 3 (BioProject accession: PRJEB39819; Reads accession: ERR5534379-ERR5534382) contains reads derived from two comparison groups (TENT5A^{WT} vs. TENT5A^{MUT}) of the cell

samples with the human embryonic kidney cell line HEK293 background (Gewartowska et al., 2021). The samples from the TENT5A^{MUT} group express a mutant version of the TENT5A protein with mis-sense mutations in its putative catalytic residues (D139N; D141N). There are two replicates in each comparison group. We named the two samples from the TENT5A^{WT} group as TENT5A.WT.rep1 and TENT5A.WT.rep2, while naming the two samples from the TENT5A^{MUT} group as TENT5A.MUT.rep1 and TENT5A.MUT.rep2. The decompressed raw reads were provided to the 00.Long_Reads module of NanoTrans for basecalling and quality assessment. The read length and quality summary of the basecalling results are shown in Table S5. The human reference genome (GRCh38) and its associated feature annotation were retrieved from Ensembl (release 107) and preprocessed with the 00.Reference_Genome module of NanoTrans. The basecalled reads were mapped to the preprocessed reference genome via the 01.Reference_Genome_based_Read_Mapping module of NanoTrans with default settings. Based on the read-mapping bam file, transcriptome clustering and quantification was further performed with the 02.Isoform_Clustering_and_Quantification module of NanoTrans by leveraging all four samples. The clustered and quantified human transcriptome isoform set was processed by the 05.Isoform_PolyA_Tail_Length_Profiling module of NanoTrans for isoform poly(A) tail length estimation.

First, we compared the poly(A) tail lengths of all isoforms between the two comparison groups: TENT5A^{WT} and TENT5A^{MUT}. As Figure 5A and 5B show, the isoforms from the TENT5A^{WT} group showed slightly longer poly(A) tails in comparison to those from the TENT5A^{MUT} group across the whole genome (mean difference = 6.512; p-value $=1.923\times10^{-82}$). In addition to such genome-wide comparison, we also generated a selected gene list for genes that are associated with extracellular matrix constituents (GO: 0031012) by using the AmiGO2 web utility (http://amigo.geneontology.org). By further limiting our isoforms of interests to these genes, we saw much more pronounced differences in terms of poly(A) tail length between the two groups, with the genes from the TENT5A^{WT} groups showing longer poly(A) tails in general (mean difference = 11.705; p-value $=1.646\times10^{-7}$; Figure 5C-D). Such observation is consistent with the original study (Table S2) (Gewartowska et al., 2021).

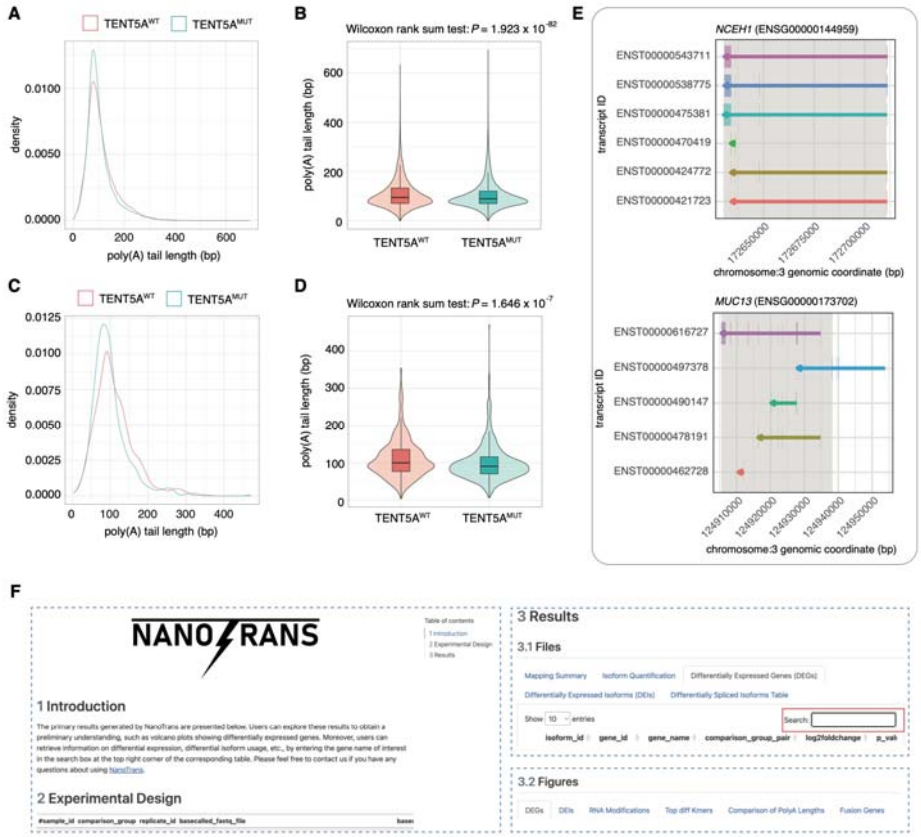


Figure 5. Applications of NanoTrans for poly(A) tail length estimation and gene fusion detection and its all-in-one summary report feature. (A) The density plot of genome-wide per-isoform poly(A) tail length distributions for the TENT5A^{WT} and TENT5A^{MUT} groups. (B) The violin plot of genome-wide per-isoform poly(A) tail length comparison between the TENT5A^{WT} and TENT5A^{MUT} groups. (C) The density plot of the per-isoform poly(A) tail length distribution of genes associated with extracellular matrix constituents (GO: 0031012) for the TENT5A^{WT} and TENT5A^{MUT} groups. (D) The violin plot of the per-isoform poly(A) tail length comparison of genes associated with extracellular matrix constituents (GO: 0031012) between the TENT5A^{WT} and TENT5A^{MUT} groups. (E) The *NCEH1*:*MUC13* gene fusion identified in the A549 human lung cancer cell line. The genome arrangements of all transcripts derived from the left donor (*NCEH1*) and right donor (*MUC13*) are presented with their corresponding transcription directions indicated by arrows. The genomic regions that constitute the fused transcript are highlighted in light grey. (F) An all-in-one HTML report that summarizing the results. The report contains three sections, in which the Results section gathered the primary files and figures generated by NanoTrans.

Application demonstration of NanoTrans in human cancer cell lines

Dataset 4 (BioProject accession: PRJEB44348; Reads accession: ERR9286473 and ERR9286521) contains reads derived from two human cancer cell lines: the lung

adenocarcinoma cell line A549 and the leukemia cell line K562 (Chen et al., 2021). We named these two samples as SGNex.A549.rep1 and SGNex.K562.rep1 respectively. The decompressed raw reads were provided to the 00.Long_Reads module of NanoTrans for basecalling and quality assessment. The read length and quality summary of the basecalling results are shown in Table S6. The human reference genome (GRCh38) and its associated feature annotation were retrieved from Ensembl (release 107) and preprocessed with the 00.Reference_Genome module of NanoTrans. We used the 06.Fusion_Gene_Detection module of NanoTrans to detect fusion genes based on the pre-processed reference genome and transcriptome by NanoTrans. For each sample, a summary table (sorted by the confidence level) was generated to report potential fusion genes and the associated transcript-based evidence.

Based on this dataset, we identified *NCEH1*:*MUC13* as a high-confidence fusion event for the A549 cell line while finding *BAG6:SLC44A4* and *GSE1:ATXN1-AS1* as high-confidence fusion event for the K562 cell line. These findings echoed well with the CCLE fusion gene call set (Ghandi et al., 2019)(Table S2). For each identified potential gene fusion event, NanoTrans also plots the genome arrangement of all transcripts associated with the left and right donor partners and highlights the corresponding genomic regions corresponding to the fused transcript. As an example, the *NCEH1*:*MUC13* gene fusion is graphically depicted, with the donor regions that make up the fused transcript highlighted in light grey (Figure 5E). This fusion should be triggered by an intrachromosomal rearrangement on chromosome 3.

An all-in-one HTML report for result summary

Finally, the Module 07 provides an HTML report rendered by Quarto (https://quarto.org/), integrating essential files and figures to offer users a convenient exploration of the primary results. The report consists of three main sections: Introduction, Experimental Design, and Results (Figure 5F). Within the Results section, files and figures from each module are organized across different page tabs, providing a streamlined presentation. Moreover, users can easily locate a specific gene by entering its names into the search box at the top-right corner of the corresponding table.

Benchmarking comparison between NanoTrans and similar existing tools

There are at least two bioinformatics tools, MasterOfPores (Cozzuto et al., 2020) and FASTdRNA (Chen et al., 2023), have been developed for Nanopore DRS data analysis. Hence, we compared these tools in terms of their functionality, ease of use, and computational efficiency.

Leveraged by functionality, NanoTrans, MasterOfPores, and FASTdRNA all covered the following functions in their latest implementations: 1) preprocessing (basecalling, quality control, reads mapping, reads counting); 2) RNA modification detection; 3) poly(A) tail length estimation (Table 1). In addition, NanoTrans and FASTdRNA have dedicate modules for alternative splicing detection in comparison to MasterOfPores. Most noticeably, only NanoTrans provides supports for differential gene/isoform expression detection and fusion gene identification analysis, which are important applications of DRS data (Table 1). In addition, while all three tools provide human-readable tabular results, NanoTrans stresses more on intuitive results summarization and visualization, which can greatly assist data interpretation towards better biological insights. Regarding ease of use, in addition to Linux command lines, MasterOfPores and FASTdRNA further require users' familiarity with workflow system tools such as Nextflow and Snakemake. Additionally, the setup of MasterOfPores requires the pre-installation of containers such as Docker or Apptainer (formerly Singularity). These prerequisites raised extra challenges for end users. In comparison, NanoTrans only relies on Linux command lines, with a design philosophy to minimize the requirement of users' bioinformatic knowledge when executing it. To gauge the computational efficiencies of NanoTrans, MasterOfPores, and FASTdRNA, we conducted a benchmarking comparison based on the same testing dataset (Dataset 5, reads quality shown in Table S7) for their commonly supported analysis, which include Nanopore reads preprocessing, RNA modification detection, and poly(A) tail length estimation. In comparison to NanoTrans, we found MasterOfPores took 18 additional hours to run through all analysis and consumed twice of much of disk space (Table 1). As for FASTdRNA, maybe because of its ongoing version updates, we cannot successfully run through its full workflow. With minor

bug fixing, we were able to execute its preprocessing module (dRNAmain) and found both NanoTrans and MasterOfPores significantly outperform FASTdRNA in terms of computational efficiency (Table 1).

Table 1. Comparison of NanoTrans, MasterOfPores, and FASTdRNA.

	Functionality					
	NanoTrans		MasterOfPores		FASTdRNA	
	Modules	**Output Reports**	**Modules**	**Output Reports**	**Modules**	**Output Reports**
Preprocessing (basecalling, reads mapping, reads counting)	00, 01, 02	Tables & Graphs	mop_preprocess	Tables & Graphs	dRNAmain	Tables & Graphs
Detecting & comparing differential gene expression	03	Tables & Graphs	/	/	/	/
Detecting & comparing alternative splicing	02	Tables & Graphs	/	/	dRNAas	Tables
Detecting & comparing RNA modifications	04	Tables & Graphs	mop_mod mop_consensus	Tables & Graphs	dRNAmodif	/
Detecting & comparing polyA tail lengths	05	Tables & Graphs	mop_tail	Tables & Graphs	dRNAtail	/
Detecting fusion genes	06	Tables & Graphs	/	/	/	/
Results summary	07	HTML	/	/	/	/
	Performance					
	NanoTrans		**MasterOfPores**		**FASTdRNA**	
	Elapsed Time*	**Disk Usage**	**Elapsed Time***	**Disk Usage**	**Elapsed Time***	**Disk Usage**
Preprocessing	1d 18h 30m 47s	**61.1G**	**1d 9h 4m 57s**	123G	2d 14h 17m 1s	140G
RNA modification detection	**18h 50m 28s**	**30G**	1d 2h 32m 16s	51G	/	/
Poly(A) tail length estimation	**11h 10m 9s**	**768M**	1d 6h 55m 9s	5.5G	/	/
Total	**3d 0h 31m 24s**	**91.9G**	3d 18h 32m 22s	179.5G	/	/

Footnote
/: indicates that the module was unavailable or that the step failed to run.
*: indicates that 12 threads were used throughout the benchmarking analysis.
bold: indicates that the corresponding tool performs better than the other two.

Discussion

In this work, we established a framework named NanoTrans to facilitate easy and standardized Nanopore direct-RNA-sequencing (DRS) data analysis for not only experienced bioinformaticians but wet-lab experimentalists with beginners' training on bioinformatics as well. NanoTrans alleviates the time and energy-consuming challenges of tools selection, parameter optimizing, workflow scripting, results visualization for DRS data analysis, offering a unified and smooth experience that enables the end users to better focus on biological insights. A detailed step-by-step manual with real world testing examples is further

provided to guide the users through the workflow even when they have limited bioinformatics knowledge. Under the hood, NanoTrans comprises seven modules that comprehensively cover different aspects of DRS data analysis, ranging from the pre-processing of raw data, such as basecalling and mapping, to dedicated downstream analysis, such as identifying RNA modifications and gene fusion events. In most cases, executing NanoTrans modules only requires executing a single command to run. Finally, NanoTrans provides an all-in-one HTML report that allows users to explore results conveniently and comprehensively.

In this study, we demonstrated the robustness and versatility of NanoTrans by applying it to four diverse DRS datasets generated based on yeast, *Arabidopsis*, and human cell lines. Important conclusions of original studies that generated these DRS datasets were validated during our re-analysis with NanoTrans (Table S2). In addition, novel biological insights such as the importance of *WRKY70* gene expression change in response to the m^6A machinery disruption is further revealed via NanoTrans during this process, which highlights the value of NanoTrans in facilitating biology-driven Nanopore DRS data analysis. Additionally, NanoTrans is designed and implemented to be as much user-friendly as possible, requiring minimal prior knowledge on Linux command lines from the end users. Technical complications of DRS data processing are automatically handled behind the scenes. Admittedly, Nanopore direct RNA sequencing and the best practice of its data analysis are both advancing very rapidly. In future, we will continue to maintain and update NanoTrans to keep it synchronized with the development of the field. The current implementation of the NanoTrans framework makes it very straightforward to incorporate newly developed tools when needed. We expected NanoTrans to grow with the field, continuously enabling researchers to make insightful biological discoveries via Nanopore DRS technology.

Materials and methods

Software prerequisites

NanoTrans is designed for a desktop or computing server running an x86-64-bit Linux operating system, preferably also with the multithreaded processor support. The installation process requires a stable internet connection and the following Linux software compilation prerequisites.

- bash (https://www.gnu.org/software/bash/)
- bzip2 and libbz2-dev (http://www.bzip.org/)
- gcc and g++ (https://gcc.gnu.org/)
- git (https://git-scm.com/)
- GNU make (https://www.gnu.org/software/make/)
- gzip (https://www.gnu.org/software/gzip/)
- libopenssl-devel (https://github.com/openssl/openssl)
- libcurl-devel (https://curl.se/libcurl/)
- java runtime environment (JRE) v1.8.0 (https://www.java.com)
- perl v5.12 or newer (https://www.perl.org/)
- tar (https://www.gnu.org/software/tar/)
- unzip (http://infozip.sourceforge.net/UnZip.html)
- wget (https://www.gnu.org/software/wget/)
- zlib and zlib-devel (https://zlib.net/)
- xz and xz-devel (https://tukaani.org/xz/)
- perl-devel (https://pkgs.org/download/perl-devel)
- R 3.6 or newer (https://www.r-project.org/)

Software installation and configuration

A bash script is pre-shipped with NanoTrans to perform automatic installation and configuration for NanoTrans' dependencies. A complete list of these dependencies is provided in Table S8.

Software implementation

NanoTrans is implemented using Perl, Python, R, and Bash scripts. The analysis codes are first encapsulated into functions using Perl, Python or R scripts. The Bash scripts are then used to package these functions along with their required parameters into individual modules. Each module can therefore be executed through one line bash command.

Hardware requirements

NanoTrans can be deployed on any Linux-based x86-64-bit computing platforms, including

cloud service such as Amazon AWS. The minimum hardware requirements for installing NanoTrans are listed below: 2 CPUs, 8 GB RAM, 20 GB of disk space. A minimum of four times the size of hard disk storage in relation to the size of raw fast5 read files is strongly recommended.

Nanopore DRS datasets reanalyzed in this study

Dataset 1

Dataset 1 (BioProject accession: PRJEB39819; Reads accession: ERR4977431) contains admixed reads derived from both the budding yeast and mouse transcriptomes (Tudek et al., 2021). All the reads were used for basecalling and reference-based mapping and only those reads that can be mapped to the budding yeast *Saccharomyces* cerevisiae reference genome (R64-1-1) were used for the downstream analysis. The sequenced yeast RNA was derived from the *S. cerevisiae* strain BY4741 growing on YPDA media under 30 °C and represented a single biological replicate, therefore we named this sample as BY4741.WT.30C.rep1. The sequencing device, sequencing flowcell version, and the sequencing kit used for this dataset are Oxford Nanopore MinION, FLO-MIN106, and SQK-RNA002 respectively.

Dataset 2

Dataset 2 (BioProject accession: PRJEB32782; Reads accession: ERR3764352- ERR3764359) contains reads derived from two comparison groups (vir-1 vs. VIR-complemented) of Arabidopsis thaliana samples with four replicates in each group (Parker et al., 2020). The samples from the vir-1 group are mutant defective in the function of the *VIRILIZER* gene, which is an important component of the *Arabidopsis* RNA m^6A writer machinery. We named these samples as vir1.rep1, vir1.rep2, vir1.rep3, and vir1.rep4 respectively. The samples from the VIR-complemented group all contain a complemented functional version of the *VIRILIZER* gene in combination with a fused Green Fluorescent Proteins (GFP) tag. We named these samples as VIRc.rep1, VIRc.rep2, VIRc.rep3, and VIRc.rep4 respectively. The sequencing device, sequencing flowcell version, and the sequencing kit used for this dataset are Oxford Nanopore MinION, FLO-MIN106, and SQK-RNA001 respectively.

Dataset 3

Dataset 3 (BioProject accession: PRJEB39819; Reads accession: ERR5534379-ERR5534382) contains reads derived from two comparison groups (TENT5A^{WT} vs. TENT5A^{MUT}) of the cell samples with the human embryonic kidney cell line HEK293 background (Gewartowska et al., 2021). The samples from the TENT5A^{MUT} group express a mutant version of the TENT5A protein with mis-sense mutations in its putative catalytic residues (D139N; D141N). There are two replicates in each comparison group. We named the two samples from the TENT5A^{WT} group as TENT5A.WT.rep1 and TENT5A.WT.rep2, while naming the two samples from the TENT5A^{MUT} group as TENT5A.MUT.rep1 and TENT5A.MUT.rep2. The sequencing device, sequencing flowcell version, and the sequencing kit used for this dataset are Oxford Nanopore MinION, FLO-MIN106, and SQK-RNA002 respectively.

Dataset 4

Dataset 4 (BioProject accession: PRJEB44348; Reads accession: ERR9286473 and ERR9286521) contains reads derived from two human cancer cell lines: the lung adenocarcinoma cell line A549 and the leukemia cell line K562 (Chen et al., 2021). We named these two samples as SGNex.A549.rep1 and SGNex.K562.rep1 respectively. The sequencing device, sequencing flowcell version, and the sequencing kit used for this dataset are Oxford Nanopore GridION, FLO-MIN106, and SQK-RNA001 respectively.

Dataset 5

Dataset 5 (BioProject accession: PRJCA017047; Reads accession: CRR786043, CRR786046, CRR786061, CRR786077) contains reads derived from two comparison groups (*WRKY7* mutant vs. wildtype) of the common wheat *Triticum aestivum* with two replicates in each group (Chen et al., 2023). The sequencing device, sequencing flowcell version, and the sequencing kit used for this dataset are Oxford Nanopore MinION, FLO-MIN106, and SQK-RNA002 respectively.

Reference genome and annotation of each dataset

By design, NanoTrans supports the reference genome and annotation file released by Ensembl

(https://www.ensembl.org) or its sister sites (e.g., Ensembl Fungi, Ensembl Plants, Ensembl Protists, and Ensembl Metazoa). Therefore, for the analysis of each dataset, we retrieved the corresponding reference genome (in FASTA format) and annotation file (in GTF format) from Ensembl (https://www.ensembl.org). Their downloading Uniform Resource Locators (URLs) are listed in Table S1.

Benchmarking comparison among NanoTrans, MasterOfPores, and FASTdRNA

Dataset 5 was used to compare the computational performance of NanoTrans (https://github.com/yjx1217/NanoTrans, with the committed number of b3112cc), MasterOfPores (https://github.com/biocorecrg/MOP2, with the committed number of b3d5cfb), and FASTdRNA (https://github.com/Tomcxf/FASTdRNA, with the committed number of 86685fb). For each tool, modules corresponding to the commonly supported functions including Nanopore reads preprocessing (basecalling, reads mapping, and reads counting), RNA modification detection, and poly(A) tail length estimation were executed. This benchmarking test was conducted on a Linux-based server with Intel(R) Xeon(R) Gold 6248R CPU (3.00GHz), 256 GB RAM, and the CentOS operating system (release 7.9.2009). A maximum of 12 threads were used throughout the benchmarking analysis.

Code and data availability

The NanoTrans software is freely distributed under MIT license at GitHub (https://github.com/yjx1217/NanoTrans).

CRediT authorship contribution statement

Ludong Yang: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - Original Draft, Writing - Review & Editing. **Xinxin Zhang**: Data curation, Formal analysis, Investigation, Validation, Writing - Review & Editing. **Fan Wang**: Data curation, Formal analysis, Investigation, Validation, Writing - Review & Editing. **Li Zhang**: Resources, Writing - Review & Editing, Funding acquisition. **Jing Li**: Data curation, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Funding acquisition. **Jia-Xing Yue**: Conceptualization, Supervision, Data curation,

Formal analysis, Investigation, Methodology, Software, Validation, Writing - Original Draft, Writing - Review & Editing, Funding acquisition.

Conflict of interest

The authors declare no competing interests.

Acknowledgements

We thank Dr. Song Gao (Sun Yat-sen University Cancer Center) for inspiring discussion. We thank Dr. Long Wang (Nanjing University) for raw fast5 reads downloading. We also thank the facility support from the Single-Molecule Sequencing Platform at Sun Yat-sen University Cancer Center. This work is supported by National Natural Science Foundation of China (32070592 to JXY, 32000395 to JL, 82272789 to LZ), Guangdong Basic and Applied Basic Research Foundation (2022A1515010717 and 2019A1515110762 to JXY and 2022A1515011873 to JL), Guangdong Pearl River Talents Program (2019QN01Y183 to JXY, 2021QN02Y168 to JL), Guangzhou Municipal Science and Technology Bureau (202102020938 to JL), and Young Talents Program of Sun Yat-sen University Cancer Center (YTP-SYSUCC-0042 to JXY and YTP-SYSUCC-0040 to JL).

References

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M., Vollmers, C., 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun 8, 16027. https://doi.org/10.1038/ncomms16027

Chen, X., Liu, Y., Lv, K., Wang, M., Liu, X., Li, B., 2023. FASTdRNA: a workflow for the analysis of ONT direct RNA sequencing. Bioinformatics Advances 3, vbad099. https://doi.org/10.1093/bioadv/vbad099

Chen, Y., Davidson, N.M., Wan, Y.K., Patel, H., Yao, F., Low, H.M., Hendra, C., Watten, L., Sim, A., Sawyer, C., Iakovleva, V., Lee, P.L., Xin, L., Ng, H.E.V., Loo, J.M., Ong, X., Ng, H.Q.A., Wang, J., Koh, W.Q.C., Poon, S.Y.P., Stanojevic, D., Tran, H.-D., Lim, K.H.E., Toh, S.Y., Ewels, P.A., Ng, H.-H., Iyer, N.G., Thiery, A., Chng, W.J., Chen, L., DasGupta, R., Sikic, M., Chan, Y.-S., Tan, B.O.P., Wan, Y., Tam, W.L., Yu, Q., Khor, C.C., Wüstefeld, T., Pratanwanich, P.N., Love, M.I., Goh, W.S.S., Ng, S.B., Oshlack, A., Göke, J., Consortium, S.-Ne., 2021. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. https://doi.org/10.1101/2021.04.21.440736

Cozzuto, L., Liu, H., Pryszcz, L.P., Pulido, T.H., Delgado-Tejedor, A., Ponomarenko, J., Novoa, E.M., 2020. MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets. Front Genet 11, 211. https://doi.org/10.3389/fgene.2020.00211

Davidson, N.M., Chen, Y., Sadras, T., Ryland, G.L., Blombery, P., Ekert, P.G., Göke, J., Oshlack, A., 2022. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. Genome Biology 23, 10. https://doi.org/10.1186/s13059-021-02588-5

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E.J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A.J., Turner, D.J., 2018. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 15, 201–206. https://doi.org/10.1038/nmeth.4577

Gewartowska, O., Aranaz-Novaliches, G., Krawczyk, P.S., Mroczek, S., Kusio-Kobiałka, M., Tarkowski, B., Spoutil, F., Benada, O., Kofroňová, O., Szwedziak, P., Cysewski, D., Gruchota, J., Szpila, M., Chlebowski, A., Sedlacek, R., Prochazka, J., Dziembowski, A., 2021. Cytoplasmic polyadenylation by TENT5A is required for proper bone formation. Cell Reports 35. https://doi.org/10.1016/j.celrep.2021.109015

Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., Hu, K., Andreev-Drakhlin, A.Y., Kim, J., Hess, J.M., Haas, B.J., Aguet, F., Weir, B.A., Rothberg, M.V., Paolella, B.R., Lawrence, M.S., Akbani, R., Lu, Y., Tiv, H.L., Gokhale, P.C., de Weck, A., Mansour, A.A., Oh, C., Shih, J., Hadi, K., Rosen, Y., Bistline, J., Venkatesan, K., Reddy, A., Sonkin, D., Liu, M., Lehar, J., Korn, J.M., Porter, D.A., Jones, M.D., Golji, J., Caponigro, G., Taylor, J.E., Dunning, C.M., Creech, A.L., Warren, A.C., McFarland, J.M., Zamanighomi, M., Kauffmann, A., Stransky, N., Imielinski, M., Maruvka, Y.E., Cherniack, A.D., Tsherniak, A., Vazquez, F., Jaffe, J.D., Lane, A.A., Weinstock, D.M.,

Johannessen, C.M., Morrissey, M.P., Stegmeier, F., Schlegel, R., Hahn, W.C., Getz, G., Mills, G.B., Boehm, J.S., Golub, T.R., Garraway, L.A., Sellers, W.R., 2019. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 569, 503–508. https://doi.org/10.1038/s41586-019-1186-3

Gleeson, J., Leger, A., Prawer, Y.D.J., Lane, T.A., Harrison, P.J., Haerty, W., Clark, M.B., 2022. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Research 50, e19. https://doi.org/10.1093/nar/gkab1129

Krause, M., Niazi, A.M., Labun, K., Cleuren, Y.N.T., Müller, F.S., Valen, E., 2019. tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. RNA 25, 1229–1241. https://doi.org/10.1261/rna.071332.119

Liu, H., Begik, O., Lucas, M.C., Ramirez, J.M., Mason, C.E., Wiener, D., Schwartz, S., Mattick, J.S., Smith, M.A., Novoa, E.M., 2019. Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun 10, 4079. https://doi.org/10.1038/s41467-019-11713-9

Liu, Q., Hu, Y., Stucky, A., Fang, L., Zhong, J.F., Wang, K., 2020. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. BMC Genomics 21, 793. https://doi.org/10.1186/s12864-020-07207-4

Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nature Reviews Genetics 1–18. https://doi.org/10.1038/s41576-020-0236-x

Loman, N.J., Quick, J., Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12, 733–735. https://doi.org/10.1038/nmeth.3444

Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D., Hall, A.J., Barton, G.J., Simpson, G.G., 2020. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife 9, e49658. https://doi.org/10.7554/eLife.49658

Pratanwanich, P.N., Yao, F., Chen, Y., Koh, C.W.Q., Wan, Y.K., Hendra, C., Poon, P., Goh, Y.T., Yap, P.M.L., Chooi, J.Y., Chng, W.J., Ng, S.B., Thiery, A., Goh, W.S.S., Göke, J., 2021. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol 39, 1394–1402. https://doi.org/10.1038/s41587-021-00949-w

Sheikh, A.H., Tabassum, N., Rawat, A., Almeida Trapp, M., Nawaz, K., Hirt, H., 2024. m6A RNA methylation counteracts dark-induced leaf senescence in Arabidopsis. Plant Physiology 194, 2663–2678. https://doi.org/10.1093/plphys/kiad660

Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., Brooks, A.N., 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun 11, 1438. https://doi.org/10.1038/s41467-020-15171-6

Tudek, A., Krawczyk, P.S., Mroczek, S., Tomecki, R., Turtola, M., Matylla-Kulińska, K., Jensen, T.H., Dziembowski, A., 2021. Global view on the metabolism of RNA poly(A) tails in yeast Saccharomyces cerevisiae. Nat Commun 12, 4951. https://doi.org/10.1038/s41467-021-25251-w

Tables

Table 1. Comparison of NanoTrans, MasterOfPores, and FASTdRNA.

Figure legends

Figure 1. The framework design and application demonstration of NanoTrans. The overview of the NanoTrans framework. The names of third-party tools employed in each step are denoted in italics.

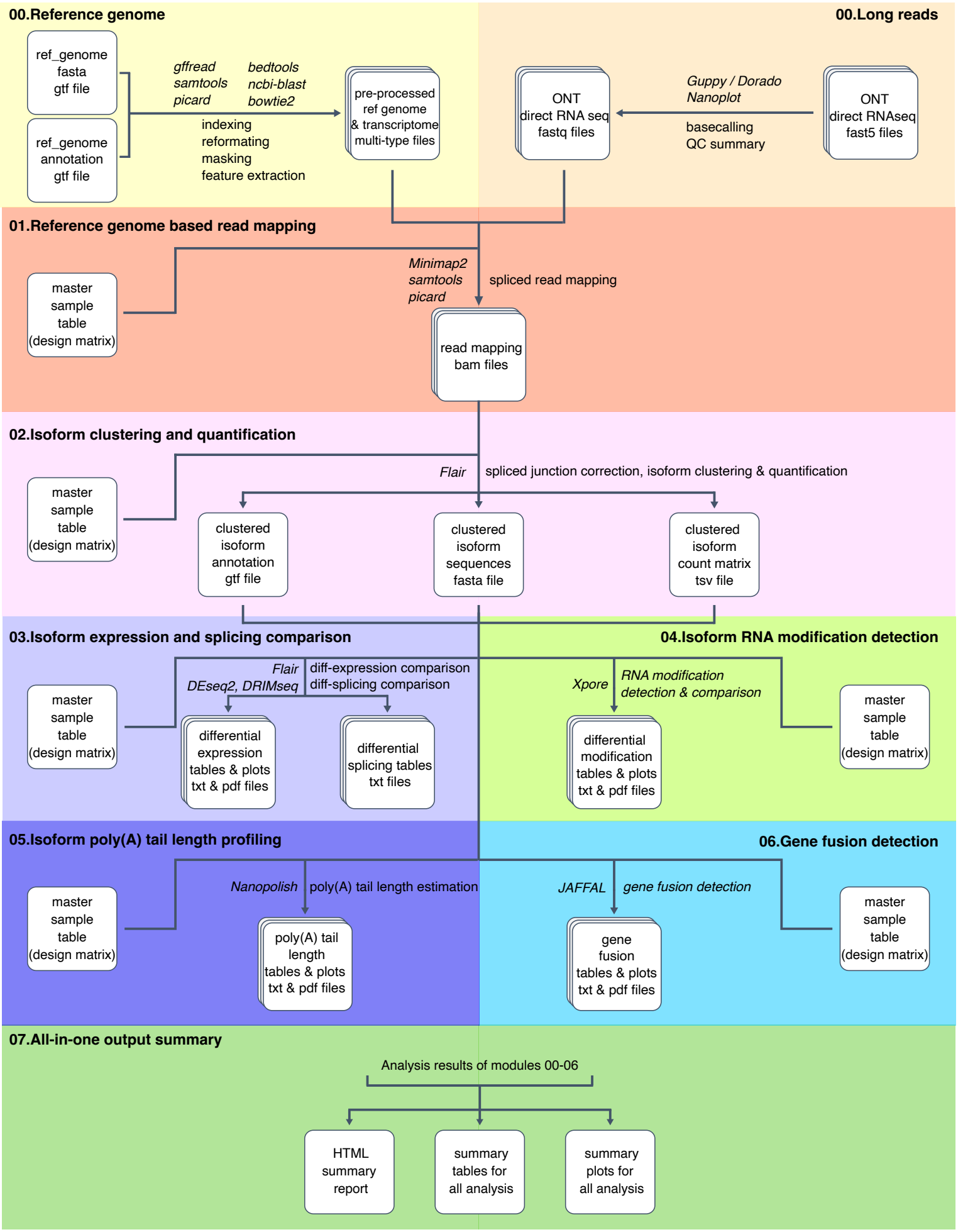
Figure 2. **Applications of NanoTrans for isoform quantification and isoform expression comparison.** (A) The negative correlation between isoform expression levels and poly(A) tail lengths identified in Dataset 1; the distribution of each isoform's expression level (log2 transformed) is plotted along the x-axis while the distribution of each isoform's median poly(A) length is plotted along the y-axis. (B-C) The differentially expressed genes (B) and isoforms (C) identified based on the comparison between the VIRc and *vir1* groups. False discovery rate (FDR) cutoff: 0.05. Fold change (log2 transformed) cutoffs: -1 and 1. (D) Biological processes enriched by differentially expressed genes (VIRc vs. *vir1*, adjusted p value < 0.05). (E) The size of words showed the frequencies of DEGs in the enriched GO terms; for example, *WRKY70* has the highest frequency among the enriched terms.

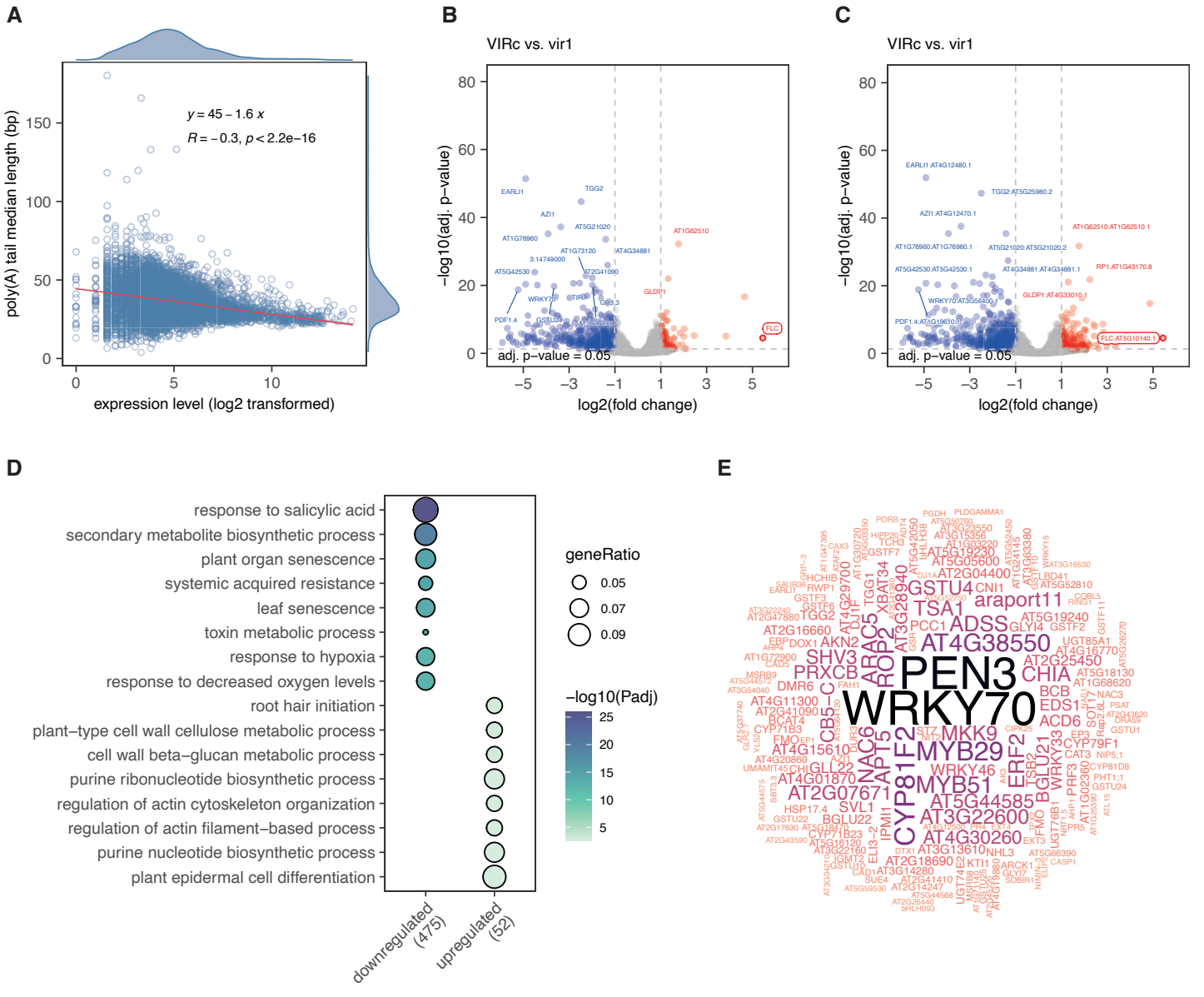
Figure 3. Examples of differentially spliced genes identified based on the comparison between the VIRc and *vir1* groups. (A) A novel transcript that is previously unannotated is identified for the *POM1* (*AT1G05850*) gene. The two identified intron retention events of this novel transcript are further highlighted based on its comparison to the two already annotated transcripts. (B) The different isoform usages of the three transcripts identified in panel A across all samples from the *vir1* and VIRc groups. (C) The mutually exclusive exons (exon 2 and exon 3, indicated by black arrows) in *FLM* (*AT1G77080*) gene is highlighted. (D) A complex transcript composed of a multitude of exons, AT1G48090.6, is identified.

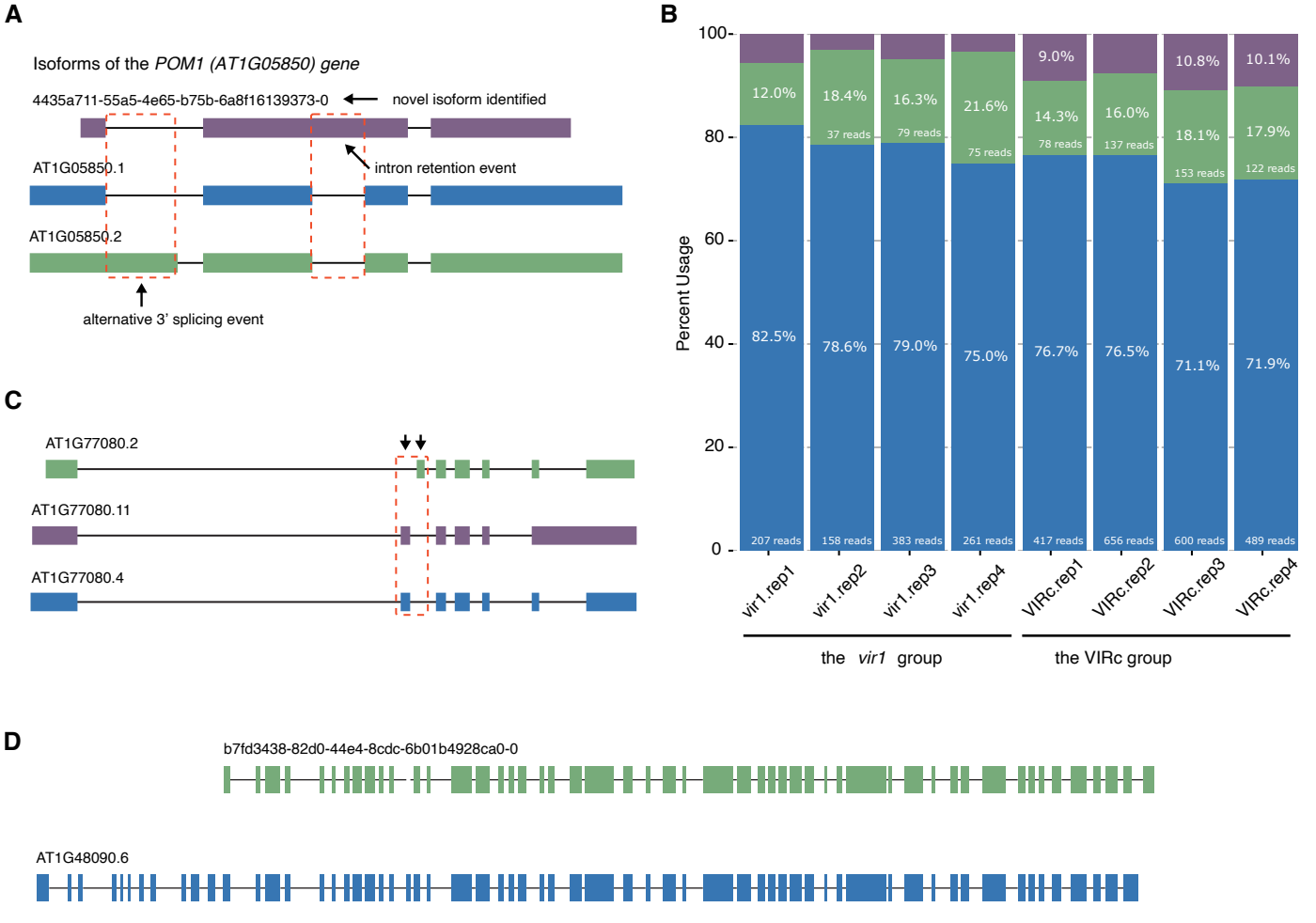
Figure 4. Characteristics of differentially modified kmers between the VIRc and *vir1* groups. The top 20 kmer-gene combinations with higher (A) or lower (B) RNA modification rates in the VIRc group than the *vir1* group. (C) The sequence logo showed the preference of filtered differentially modified kmers (DMKs, differential modification rates > 0.5, p-value < 0.05, and only kmers with center A were retained). (D) The top 10 DMKs with the highest frequency. (E) Distribution of the DMKs on distinct genomic features. (F) Frequency of DMKs around stop codons. (G) Overlaps between RNA modifications and poly(A) signal sites. NanoTrans directly outputs this kind of plot for each analyzed sample.

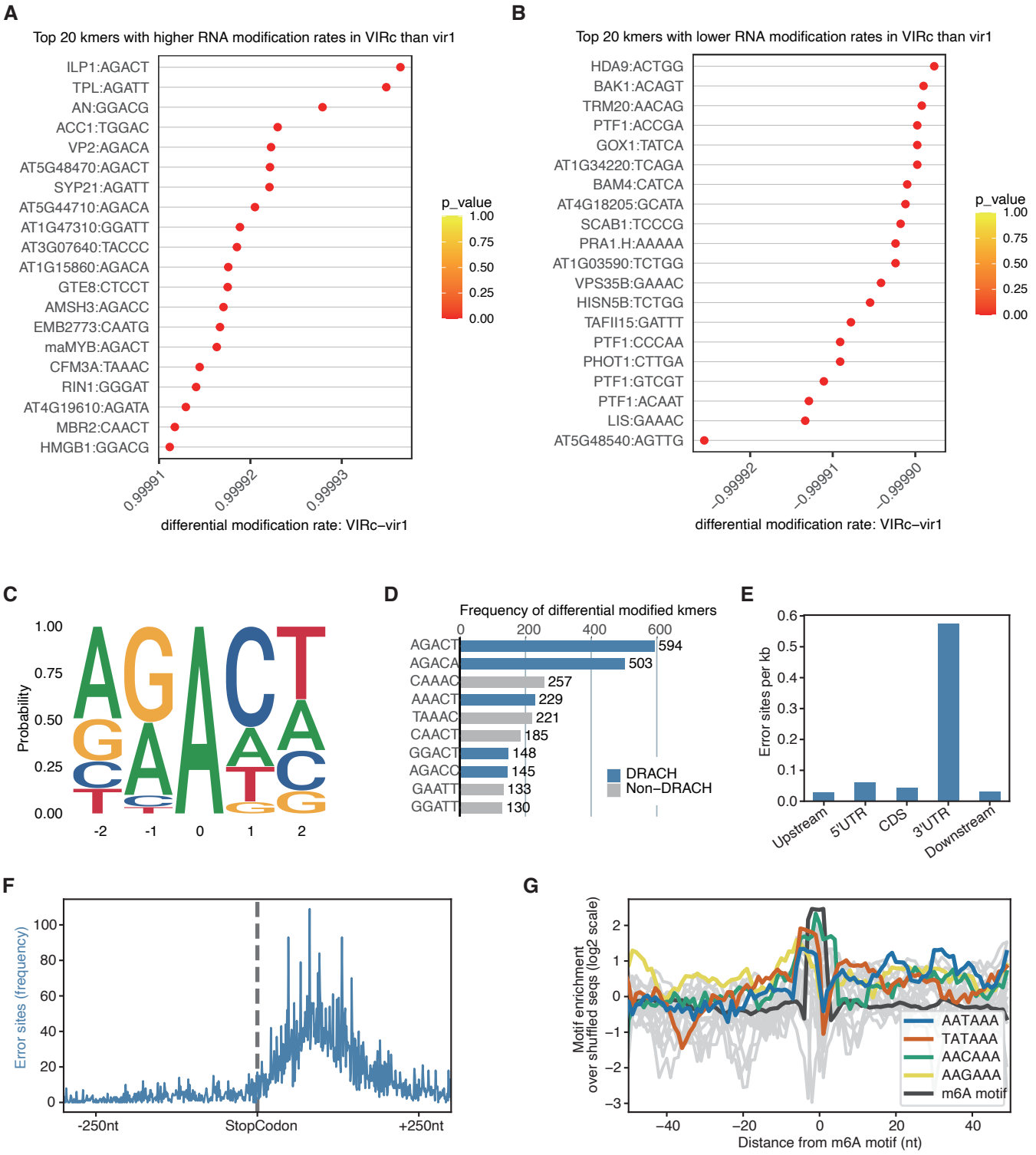
Figure 5. Applications of NanoTrans for poly(A) tail length estimation and gene fusion detection and its all-in-one summary report feature. (A) The density plot of genome-wide per-isoform poly(A) tail length distributions for the TENT5A^{WT} and TENT5A^{MUT} groups. (B)

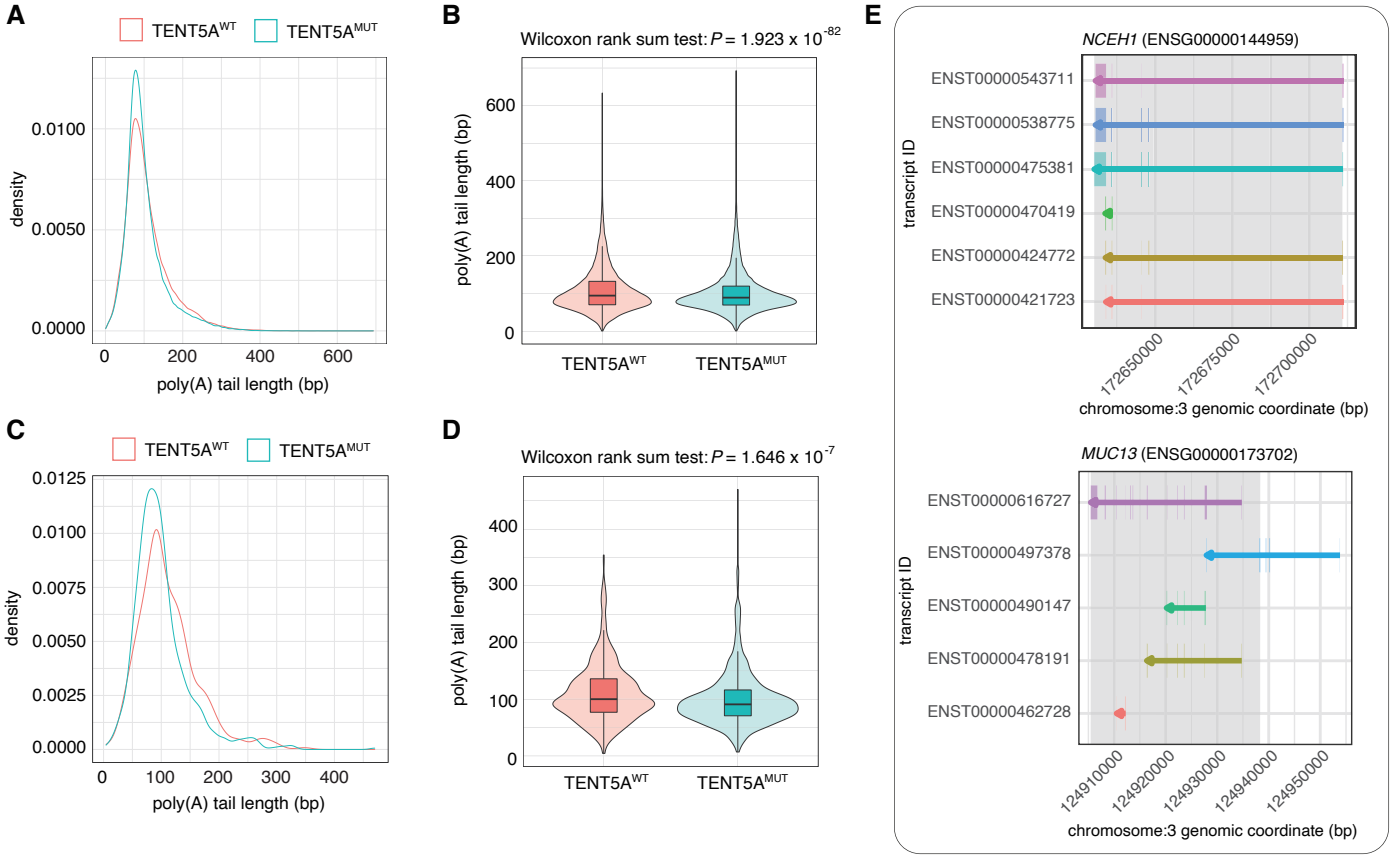
The violin plot of genome-wide per-isoform poly(A) tail length comparison between the TENT5A^{WT} and TENT5A^{MUT} groups. (C) The density plot of the per-isoform poly(A) tail length distribution of genes associated with extracellular matrix constituents (GO: 0031012) for the TENT5A^{WT} and TENT5A^{MUT} groups. (D) The violin plot of the per-isoform poly(A) tail length comparison of genes associated with extracellular matrix constituents (GO: 0031012) between the TENT5A^{WT} and TENT5A^{MUT} groups. (E) The *NCEH1*:*MUC13* gene fusion identified in the A549 human lung cancer cell line. The genome arrangements of all transcripts derived from the left donor (*NCEH1*) and right donor (*MUC13*) are presented with their corresponding transcription directions indicated by arrows. The genomic regions that constitute the fused transcript are highlighted in light grey. (F) An all-in-one HTML report that summarizing the results. The report contains three sections, in which the Results section gathered the primary files and figures generated by NanoTrans.


00.Reference genome
ref_genome fasta gtf file
ref_genome annotation gtf file
gffread
samtools
picard
bedtools
ncbi-blast
bowtie2
indexing
reformating
masking
feature extraction
pre-processed ref genome & transcriptome multi-type files
00.Long reads
ONT direct RNA seq fastq files
Guppy / Dorado
Nanoplot
basecalling
QC summary
ONT direct RNAseq fast5 files
01.Reference genome based read mapping
master sample table (design matrix)
Minimap2
samtools
picard
spliced read mapping
read mapping bam files
02.Isoform clustering and quantification
master sample table (design matrix)
Flair
spliced junction correction, isoform clustering & quantification
clustered isoform annotation gtf file
clustered isoform sequences fasta file
clustered isoform count matrix tsv file
03.Isoform expression and splicing comparison
04.Isoform RNA modification detection
master sample table (design matrix)
Flair
DEseq2, DRIMseq
diff-expression comparison
diff-splicing comparison
differential expression tables & plots txt & pdf files
differential splicing tables txt files
Xpore
RNA modification detection & comparison
differential modification tables & plots txt & pdf files
master sample table (design matrix)
05.Isoform poly(A) tail length profiling
06.Gene fusion detection
master sample table (design matrix)
Nanopolish
poly(A) tail length estimation
poly(A) tail length tables & plots txt & pdf files
JAFFAL
gene fusion detection
gene fusion tables & plots txt & pdf files
master sample table (design matrix)
07.All-in-one output summary
Analysis results of modules 00-06
HTML summary report
summary tables for all analysis
summary plots for all analysis

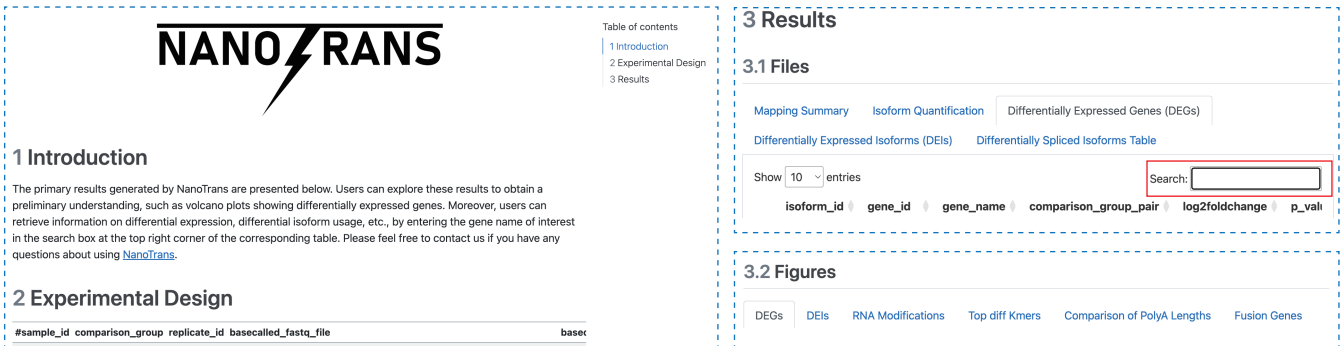

A
y = 45 − 1.6 x
R = −0.3, p < 2.2e−16
poly(A) tail median length (bp)
expression level (log2 transformed)
B
VIRc vs. vir1
−log10(adj. p−value)
log2(fold change)
adj. p−value = 0.05
C
VIRc vs. vir1
−log10(adj. p−value)
log2(fold change)
adj. p−value = 0.05
D
response to salicylic acid
secondary metabolite biosynthetic process
plant organ senescence
systemic acquired resistance
leaf senescence
toxin metabolic process
response to hypoxia
response to decreased oxygen levels
root hair initiation
plant−type cell wall cellulose metabolic process
cell wall beta−glucan metabolic process
purine ribonucleotide biosynthetic process
regulation of actin cytoskeleton organization
regulation of actin filament−based process
purine nucleotide biosynthetic process
plant epidermal cell differentiation
downregulated (475)
upregulated (52)
geneRatio
0.05
0.07
0.09
−log10(Padj)
25
20
15
10
5
E
PEN3
WRKY70
AT4G38550
MYB29
MYB51
CYP81F2
AT3G22600
AT4G30260
AT5G44585
AT2G07671
araport11
GSTU4
TSA1
ADSS
CHIA
EDS1
ACD6
ERF2
BGLU21
WRKY33
MKK9
APT5
NAC6
GLL22
AT4G01870
AT4G15610
AT2G16660
AT2G25450
AT5G18130
AT5G19240
AT5G05600
AT2G04400
AT5G19230
AT3G23550
AT3G15356
AT1G03220
AT4G11300
AT2G41090
BCAT4
CYP71B3
FMO
AT2G20860
UMAMIT45
CHI
SVL1
HSP17.4
GSTU22
BGLU22
CYP71B23
AT5G16120
AT3G22160
AT3G13610
NHL3
AT2G18690
KTI1
AT3G14280
AT2G41410
AT2G14247
AT5G44568
AT2G05440
bHLH093
AT5G66390
CASP1
ARCK1
GLYI7
SOBIR1
AT4G19960
AT5G59530
SUE4
UGT74E2
UGT76B1
EXT3
PR5
ATL15
AT1G02930
AT1G02920
GSTU24
PHT1;1
CYP81D8
NIP5;1
CAT3
CYP79F1
EP3
GSTU1
ORA59
PSAT
SOT1
NAC3
UGT85A1
AT4G16770
AT5G26270
AT5G52810
COBL5
RINO1
GSTF11
GSTF2
GLYI4
BCB
AT1G68620
DMR6
PRXCB
SHV3
AKN2
DOX1
TGG2
AT2G47880
AT2G29700
XBAT34
AT3G28940
ROP2
ARAC5
TGG1
PCC1
AT5G42050
AT5G52450
AT5G63380
AT5G24145
AT5G63390
AT5G16530
AT4G24780
AT5G52310
AT5G52300
AT5G52290
AT5G52270
AT5G52250
AT5G52240
AT5G52230
AT5G52220
AT5G52210
AT5G52200
AT5G52190
AT5G52180
AT5G52170
AT5G52160
AT5G52150
AT5G52140
AT5G52130
AT5G52120
AT5G52110
AT5G52100
AT5G52090
AT5G52080
AT5G52070
AT5G52060
AT5G52050
AT5G52040
AT5G52030
AT5G52020
AT5G52010
AT5G52000
AT5G51990
AT5G51980
AT5G51970
AT5G51960
AT5G51950
AT5G51940
AT5G51930
AT5G51920
AT5G51910
AT5G51900
AT5G51890
AT5G51880
AT5G51870
AT5G51860


A
Isoforms of the POM1 (AT1G05850) gene
4435a711-55a5-4e65-b75b-6a8f16139373-0
novel isoform identified
intron retention event
AT1G05850.1
AT1G05850.2
alternative 3' splicing event
B
Percent Usage
100
80
60
40
20
0
12.0%
82.5%
207 reads
18.4%
37 reads
78.6%
158 reads
16.3%
79 reads
79.0%
383 reads
21.6%
75 reads
75.0%
261 reads
9.0%
14.3%
78 reads
76.7%
417 reads
16.0%
137 reads
76.5%
656 reads
10.8%
18.1%
153 reads
71.1%
600 reads
10.1%
17.9%
122 reads
71.9%
489 reads
vir1.rep1
vir1.rep2
vir1.rep3
vir1.rep4
VIRc.rep1
VIRc.rep2
VIRc.rep3
VIRc.rep4
the vir1 group
the VIRc group
C
AT1G77080.2
AT1G77080.11
AT1G77080.4
D
b7fd3438-82d0-44e4-8cdc-6b01b4928ca0-0
AT1G48090.6


A
Top 20 kmers with higher RNA modification rates in VIRc than vir1
ILP1:AGACT
TPL:AGATT
AN:GGACG
ACC1:TGGAC
VP2:AGACA
AT5G48470:AGACT
SYP21:AGATT
AT5G44710:AGACA
AT1G47310:GGATT
AT3G07640:TACCC
AT1G15860:AGACA
GTE8:CTCCT
AMSH3:AGACC
EMB2773:CAATG
maMYB:AGACT
CFM3A:TAAAC
RIN1:GGGAT
AT4G19610:AGATA
MBR2:CAACT
HMGB1:GGACG
0.99991
0.99992
0.99993
differential modification rate: VIRc−vir1
p_value
1.00
0.75
0.50
0.25
0.00
B
Top 20 kmers with lower RNA modification rates in VIRc than vir1
HDA9:ACTGG
BAK1:ACAGT
TRM20:AACAG
PTF1:ACCGA
GOX1:TATCA
AT1G34220:TCAGA
BAM4:CATCA
AT4G18205:GCATA
SCAB1:TCCCG
PRA1.H:AAAAA
AT1G03590:TCTGG
VPS35B:GAAAC
HISN5B:TCTGG
TAFII15:GATTT
PTF1:CCCAA
PHOT1:CTTGA
PTF1:GTCGT
PTF1:ACAAT
LIS:GAAAC
AT5G48540:AGTTG
−0.99992
−0.99991
−0.99990
differential modification rate: VIRc−vir1
p_value
1.00
0.75
0.50
0.25
0.00
C
Probability
1.00
0.75
0.50
0.25
0.00
-2
-1
0
1
2
D
Frequency of differential modified kmers
0
200
400
600
AGACT 594
AGACA 503
CAAAC 257
AAACT 229
TAAAC 221
CAACT 185
GGACT 148
AGACC 145
GAATT 133
GGATT 130
DRACH
Non−DRACH
E
Error sites per kb
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Upstream
5'UTR
CDS
3'UTR
Downstream
F
Error sites (frequency)
100
80
60
40
20
0
-250nt
StopCodon
+250nt
G
Motif enrichment over shuffled seqs (log2 scale)
2
1
0
−1
−2
−3
−40
−20
0
20
40
Distance from m6A motif (nt)
AATAAA
TATAAA
AACAAA
AAGAAA
m6A motif


A
TENT5A^WT
TENT5A^MUT
density
0.0100
0.0050
0.0000
0
200
400
600
poly(A) tail length (bp)
B
Wilcoxon rank sum test: P = 1.923 x 10^-82
poly(A) tail length (bp)
600
400
200
0
TENT5A^WT
TENT5A^MUT
C
TENT5A^WT
TENT5A^MUT
density
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
100
200
300
400
poly(A) tail length (bp)
D
Wilcoxon rank sum test: P = 1.646 x 10^-7
poly(A) tail length (bp)
400
300
200
100
0
TENT5A^WT
TENT5A^MUT
E
NCEH1 (ENSG00000144959)
transcript ID
ENST00000543711
ENST00000538775
ENST00000475381
ENST00000470419
ENST00000424772
ENST00000421723
172650000
172675000
172700000
chromosome:3 genomic coordinate (bp)
MUC13 (ENSG00000173702)
transcript ID
ENST00000616727
ENST00000497378
ENST00000490147
ENST00000478191
ENST00000462728
124910000
124920000
124930000
124940000
124950000
chromosome:3 genomic coordinate (bp)

F


NANO RANS
Table of contents
1 Introduction
2 Experimental Design
3 Results
1 Introduction
The primary results generated by NanoTrans are presented below. Users can explore these results to obtain a preliminary understanding, such as volcano plots showing differentially expressed genes. Moreover, users can retrieve information on differential expression, differential isoform usage, etc., by entering the gene name of interest in the search box at the top right corner of the corresponding table. Please feel free to contact us if you have any questions about using NanoTrans.
2 Experimental Design
#sample_id comparison_group replicate_id basecalled_fastq_file
3 Results
3.1 Files
Mapping Summary
Isoform Quantification
Differentially Expressed Genes (DEGs)
Differentially Expressed Isoforms (DEIs)
Differentially Spliced Isoforms Table
Show 10 entries
Search:
isoform_id gene_id gene_name comparison_group_pair log2foldchange
3.2 Figures
DEGs
DEIs
RNA Modifications
Top diff Kmers
Comparison of PolyA Lengths
Fusion Genes