Title page

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

Ludong Yang[1,^], Xinxin Zhang[1,^], Fan Wang[1,^], 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.

^: 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, and all-in-one tabular and graphic report output, as well as automatic installation and configuration support. 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. Here we developed NanoTrans, an easy-to-use one-stop solution for general users to perform comprehensive Nanopore DRS data analysis that covers diverse application scopes: expression quantification, isoform characterization, poly(A) tail length profiling, RNA modification evaluation, and fusion gene detection. The advent of NanoTrans addresses the practical needs of the field and 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 analyses 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").

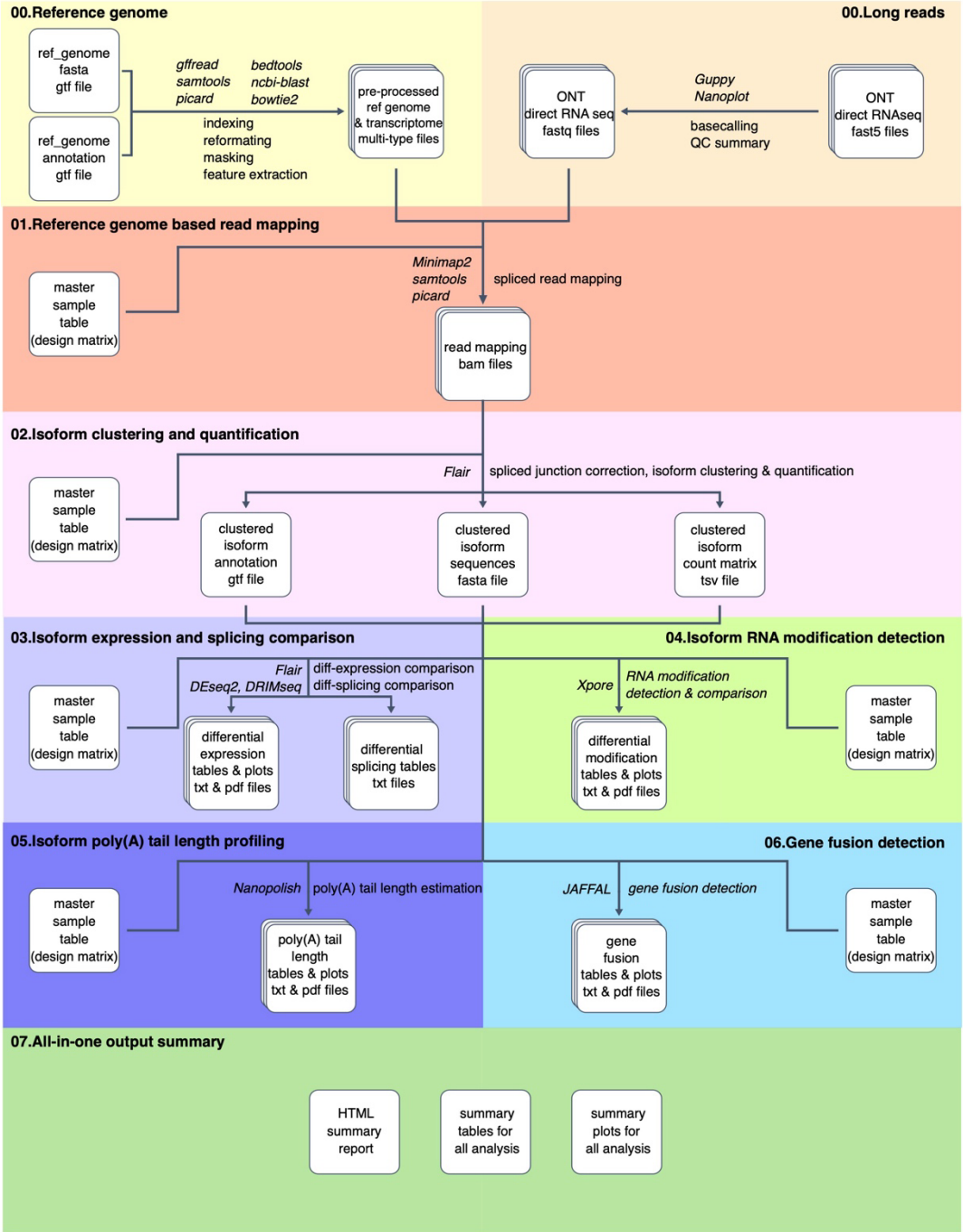


Figure 1. The framework design and application demonstration of NanoTrans. The overview of the NanoTrans framework. The names of third-party tools employed in each step are denoted in italics.

Application demonstration summary of NanoTrans

To demonstrate the application of NanoTrans in real case scenarios, we applied it to four public Nanopore DRS datasets (Table 1 and 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. 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.

Table 1. Public datasets used for application demonstration.

Organisms	**Raw data Bioproject ID**	**Analyzed samples**	**Demonstrated applications**
Yeast	PRJEB39819	BY4741.WT.30C.rep1	Isoform clustering and quantification, Isoform poly(A) length profiling
Arabidopsis	PRJEB32782	vir1.rep1, vir1.rep2, vir1.rep3, vir1.rep4, VIRc.rep1, VIRc.rep2, VIRc.rep3, VIRc.rep4	Isoform clustering and quantification, Isoform expression and splicing comparison, Isoform RNA modification detection

Human (embryonic cell lines)	PRJEB39819	HEK293.TENT5A.WT.rep1, HEK293.TENT5A.WT.rep1, HEK293.TENT5A.MUT.rep1, HEK293.TENT5A.MUT.rep2	Isoform clustering and quantification, Isoform poly(A) length profiling
Human (cancer cell lines)	PRJEB44348	SGNex.A549.rep1, SGNex.K562.rep1	Isoform clustering and quantification, Gene fusion detection

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, first, the *S. cerevisiae* reference genome (R64-1-1) and its associated feature annotation were retrieved from Ensembl (Table S2) 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 (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 2).

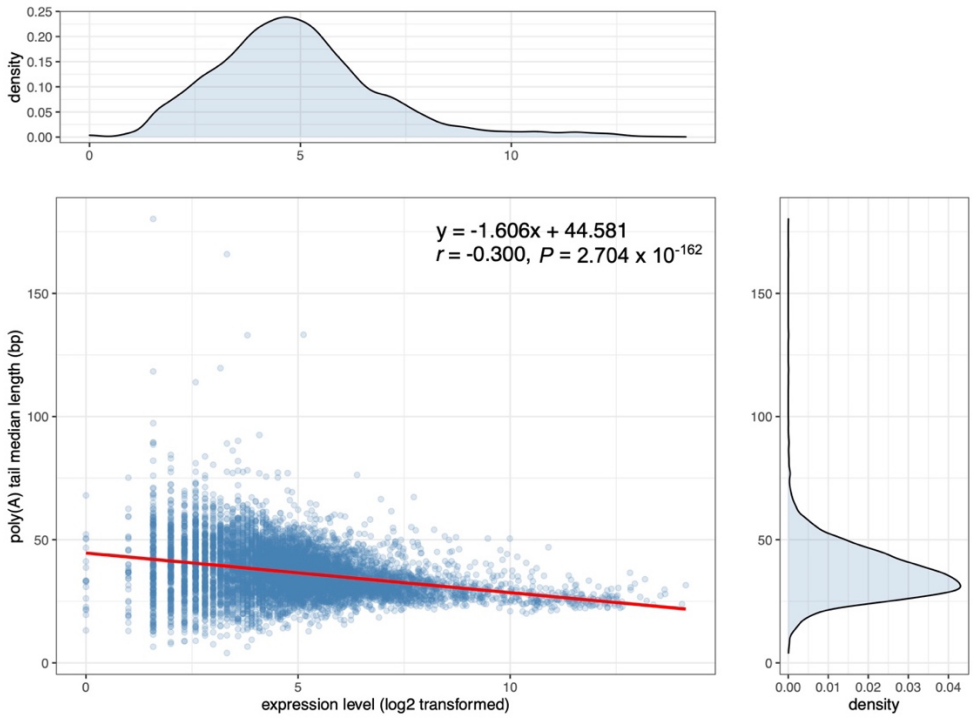


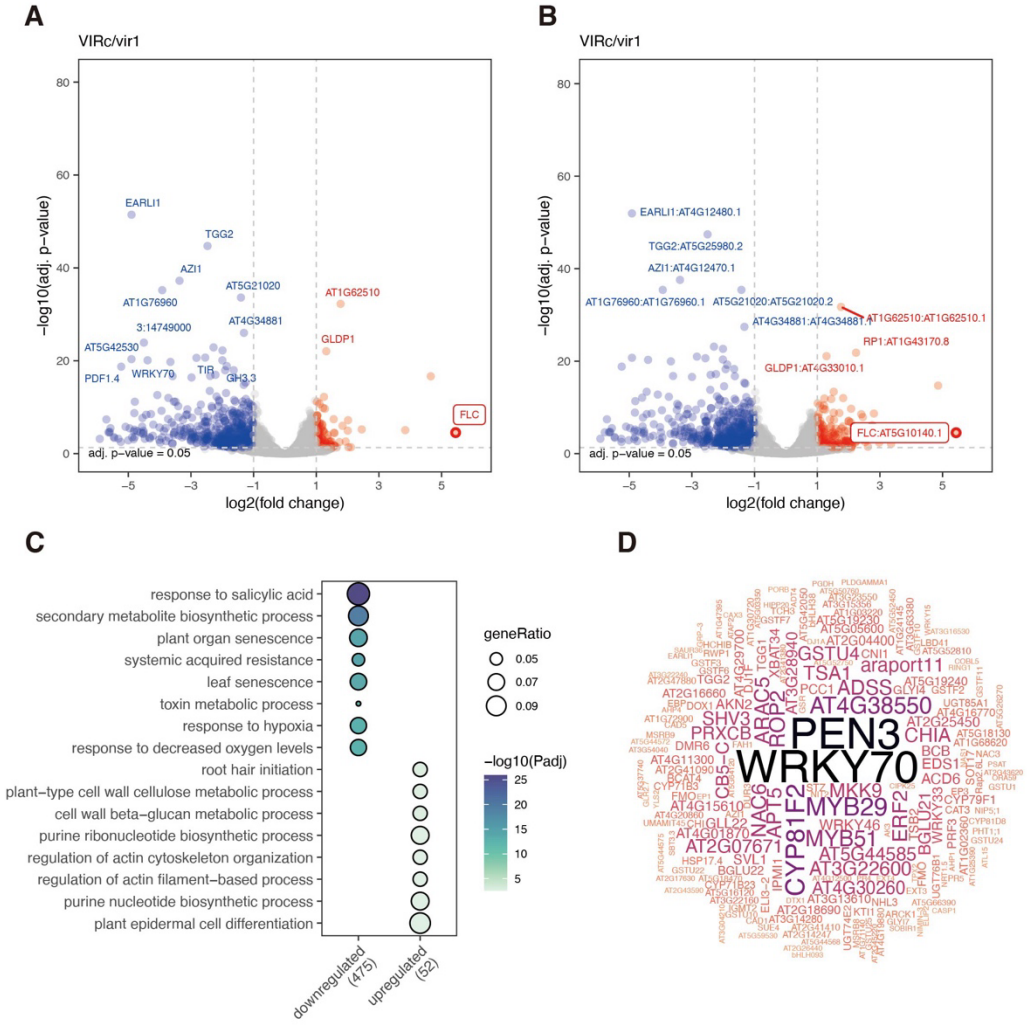
Figure 2. The negative correlation between isoform expression levels and poly(A) tail lengths identified in Dataset 1. The distribution of each isoform's expression level (log2 transformed) is plotted along the x-axis while the distribution of each isoform's median poly(A) length is plotted along the y-axis.

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. The differentially expressed genes and isoforms are reported in tab-separated tables (sorted by p-value) and further visualized with volcano plots (Figure 3A-B). In such volcano plots, differentially expressed genes are labeled with "gene_name" only (Figure 3A), while differentially expressed isoforms are labeled with "gene_name:isoform_id" tags (Figure 3B). We found the WRKY70 gene is among those strongly differentially expressed genes with the lowest p-values, suggesting its change in expression heavily linked to the interruption of the m6A machinery. Gene ontology (GO) enrichment of the differentially expressed genes further highlighted WRKY70's prevalence across significantly enriched biological processes (Figure 3C-D). Also, NanoTrans recaptured the differential gene expression signal on the *FLC* (*AT5G10140*) gene reported by the original study, highlighting the important impact of the m6A machinery on flowering time regulation in *Arabidopsis*. Taken together, these results demonstrated that NanoTrans can be very help in characterizing the expression and splicing changes at both the gene and isoform levels with its 03.Isoform_Expression_and_Splicing_Comparison module.


A
VIRc/vir1
−log10(adj. p−value)
log2(fold change)
EARLI1
TGG2
AZI1
AT5G21020
AT1G76960
AT4G34881
3:14749000
AT5G42530
PDF1.4
WRKY70
TIR
GH3.3
AT1G62510
GLDP1
FLC
adj. p−value = 0.05
B
VIRc/vir1
EARLI1:AT4G12480.1
TGG2:AT5G25980.2
AZI1:AT4G12470.1
AT1G76960:AT1G76960.1
AT5G21020:AT5G21020.2
AT4G34881:AT4G34881.4
AT1G62510:AT1G62510.1
RP1:AT1G43170.8
GLDP1:AT4G33010.1
FLC:AT5G10140.1
adj. p−value = 0.05
C
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
D
PEN3
WRKY70
CYP81F2
MYB29
MYB51
AT4G38550
GSTU4
araport11
TSA1
ADSS
CHIA
PRXCB
SHV3
AT2G07671
AT5G44585
AT3G22600
AT4G30260
AT4G01870
AT4G15610
MKK9
EDS1
ACD6
BCB
AT2G25450
AT2G16660
AT5G19240
AT4G11300
AT2G41090
SCA14
AT3G13610
AT2G18690
AT3G14280
AT2G31424
AT5G05600
AT2G04400
AT4G18230
AT3G28940
AT4G29700
AT1F
TGG1
ARAC5
ROP2
ARAC6
AP15
NAC6
CB5
AKN2
DOX1
PCC1
GLYI4
GSTF2
UGT85A1
AT4G16770
AT5G18130
AT1G68620
NAC3
PSAT
CYP79F1
CAT3
NIP5;1
GSTU1
GSTU24
CYP81D8
PHT1;1
ATL6
PRF3
WRKY33
BGL21
ERF2
EFR
WRKY46
GLL22
CHI
SVL1
BGLU22
HSP17.4
IPMI1
ELI3−2
CYP71B23
AT3G22160
AT5G16120
AT2G43590
AT2G17630
AT4G34410
AT2G31410
AT5G59530
AT2G25440
AT4G34568
AT2G26440
AT5G44572
DMR6
MSRB9
AT5G44640
AT2G22900
LAD5
AT1G72900
AT2G37880
TGG2
AT2G16660
AT5G19240
AT4G16770
AT2G25450
AT5G18130
AT1G68620
AT4G11300
AT2G41090
SCA14
AT4G15610
AT4G20860
UMAMIT45
CHI
GLL22
AT4G01870
AT2G07671
HSP17.4
SVL1
BGLU22
CYP71B23
AT2G43590
AT2G17630
AT3G22160
AT5G16120

events were identified for the *Arabidopsis POM1* (*AT1G05850*) gene (Figure 4). 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. 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 4C). Secondly, it can precisely reveal highly complex transcript, such as AT1G48090.6 (Figure 4D). These two instances were also highlighted in the original study that produced this DRS dataset (Parker et al., 2020).

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 5A-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 previous reports, such as predominant distribution in 3' untranslated regions (UTRs) and a pronounced peak downstream of stop codons (Figure 5C-F). Additionally, our analysis revealed substantial overlap between the identified kmers and polyadenylation sites (PASs) (Figure 5G). Taken together, these observations recapitulated the characteristics of differentially modified kmers identified in original study of this dataset (Parker et al., 2020).

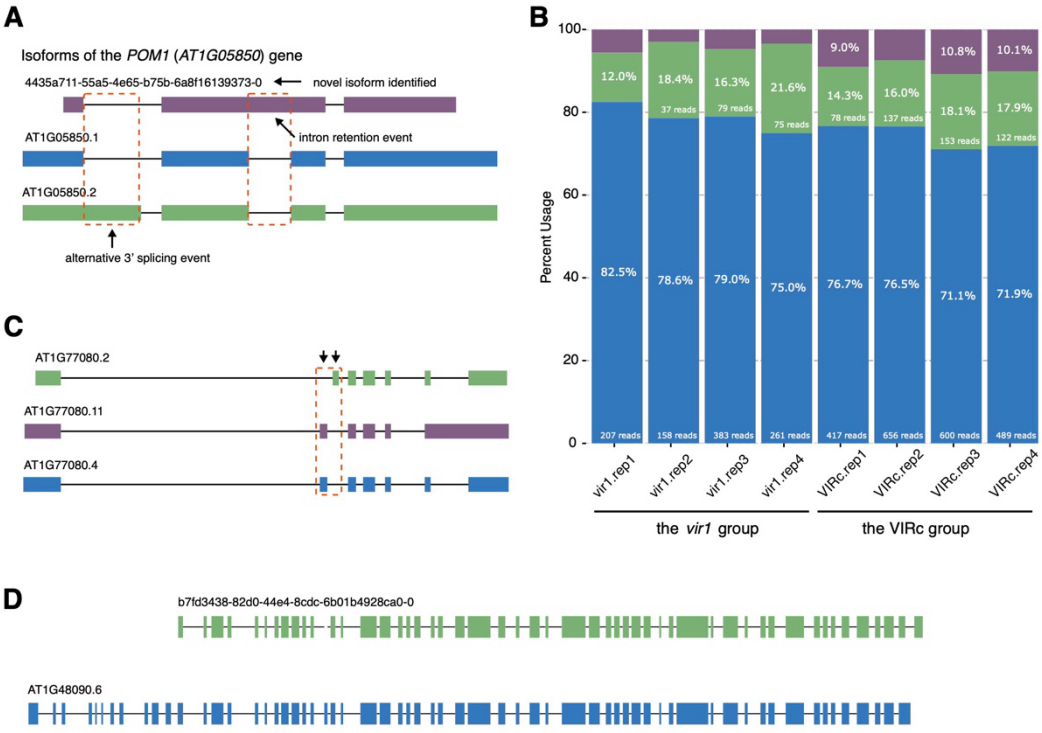


Figure 4. Examples of differentially spliced genes identified based on the comparison between the VIRc and *vir1* groups. (A) A novel transcript that is previously unannotated is identified for the *POM1* (*AT1G05850*) gene. The two identified intron retention events of this novel transcript are further 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.

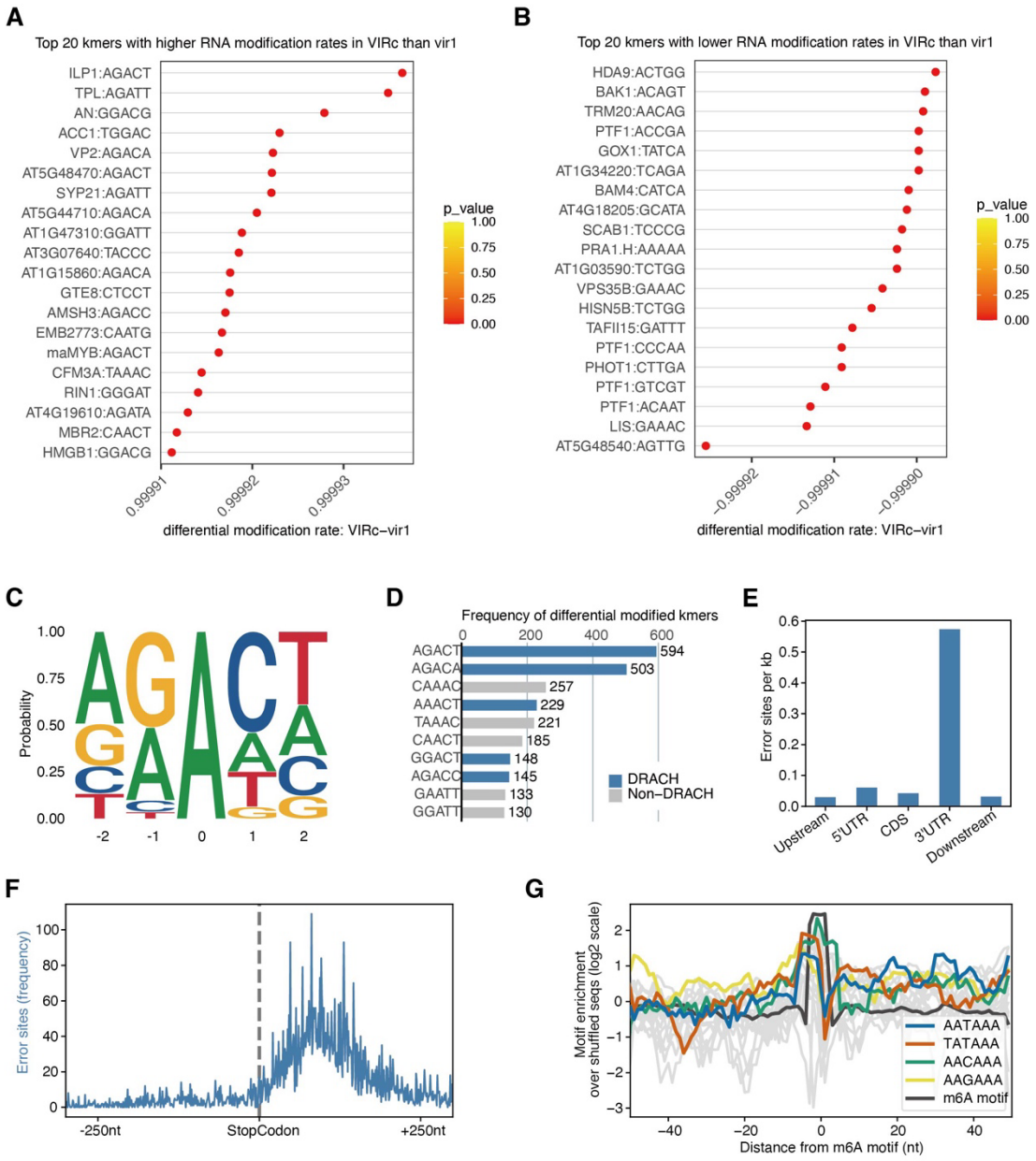


Figure 5. Characteristics of differentially modified kmers between the VIRc and vir1 groups. The top 20 kmer-gene combinations with higher (A) or lower (B) RNA modification rates in the VIRc group than the *vir1* group. (C) The sequence logo showed the preference of filtered differentially modified kmers (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 6A and 6B 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 =1.923×10^{-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 =1.646×10^{-7}; Figure 6C-6D). Such observation is consistent with the original study (Gewartowska et al., 2021).

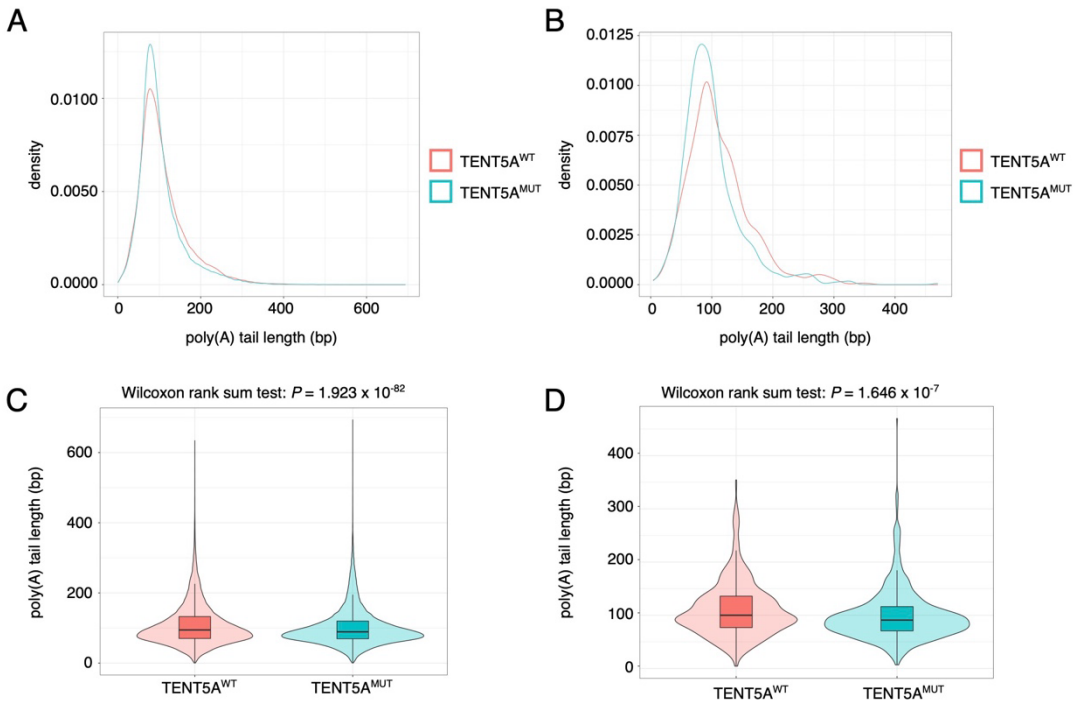


Figure 6. Poly(A) tail length differences between the TENT5A^{WT} and TENT5A^{MUT} groups. (A) The density plot of genome-wide per-isoform poly(A) tail length distributions for the TENT5A^{WT} and TENT5A^{MUT} groups. (B) The violin plot of genome-wide per-isoform poly(A) tail length comparison between 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.

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 retrieve 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. All these high-confidence fusion events were further validated in the CCLE fusion gene call sets (Ghandi et al., 2019). 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 depicted in Figure 7, with the donor regions that make up the fused transcript highlighted in light grey (Figure 7). This fusion should be triggered by an intrachromosomal rearrangement on chromosome 3.

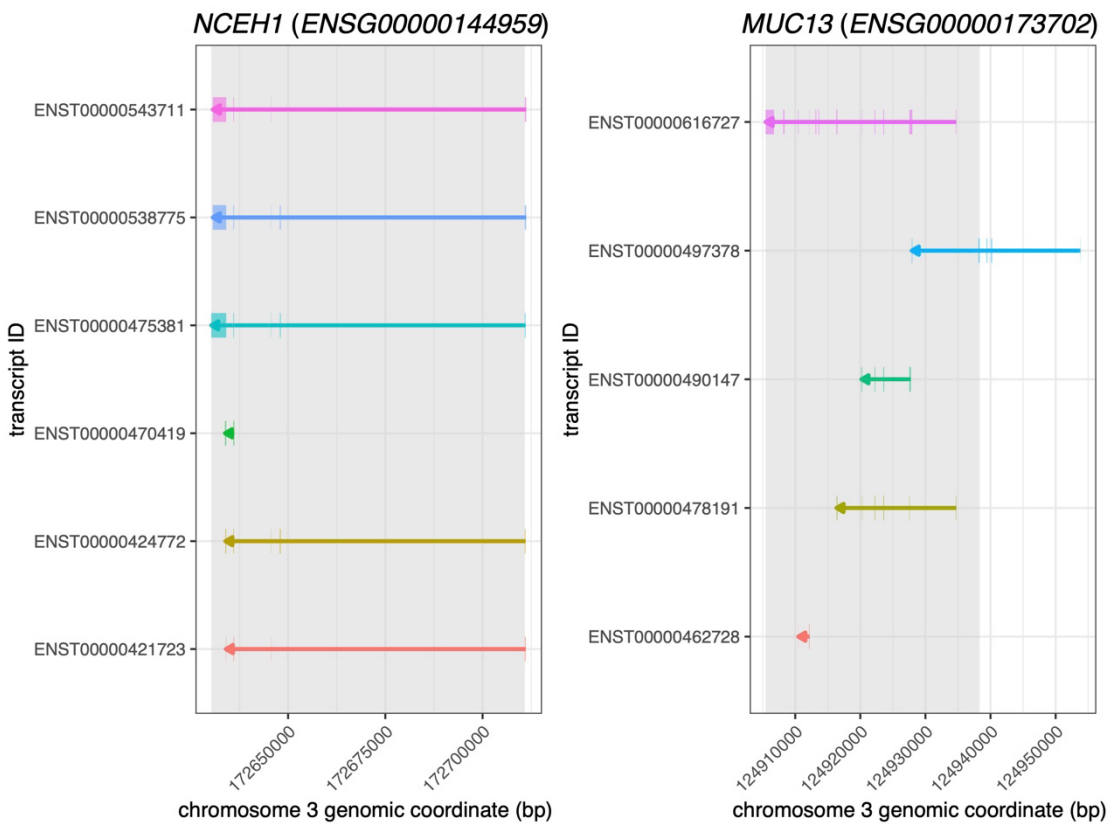


Figure 7. The *NCEH1*:*MUC13* gene fusion identified in the A549 human lung cancer cell line. The genome arrangements of all transcripts derived from the left donor (*NCEH1*) and right donor (*MUC13*) are presented with their corresponding transcription directions indicated by arrows. The genomic regions that constitute the fused transcript are highlighted in light grey.

An all-in-one HTML report for result summary

Finally, the Module 07 provides an HTML report rendered by Quarto, 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 8). 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.

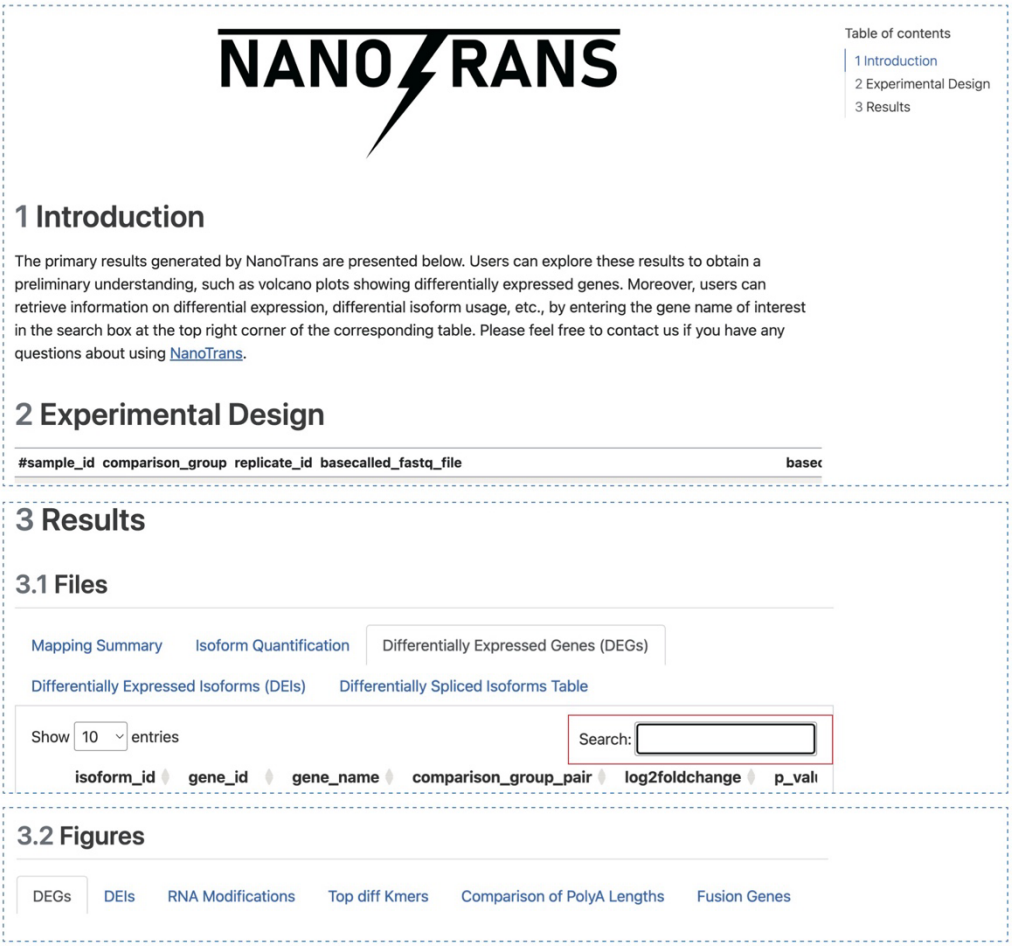


Figure 8. An HTML report summarizing results. The report contains three sections, in which the Results section gathered the primary files and figures generated by NanoTrans.

Discussion

We established a framework called NanoTrans to alleviate the time and energy-consuming challenges associated with installing and deploying tools for DRS data analysis. 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 analyses, such as identifying RNA modifications and gene fusion events. Moreover, NanoTrans provides an HTML report that allows users to explore results conveniently. 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 further validated during our re-analysis with NanoTrans. Additionally, NanoTrans is designed to be user-friendly, requiring minimal additional skills apart from basic Shell knowledge. In summary, NanoTrans provides a highly streamlined framework that automatically handles the technical complications of DRS data processing behind the scenes and allow the front users to focus on extracting the maximum of biological significant findings revealed by 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 S7.

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.

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.

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 S3.

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 - 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. **Jing Li**: Data curation, Formal analysis, Investigation, Resources, Writing - Review & Editing. **Jia-Xing Yue**: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - Review & Editing.

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), Natural Science Foundation of Guangdong Province (2022A1515010717 to JXY and 2022A1515011873 to JL), Guangdong Basic and Applied Basic Research Foundation (2019A1515110762 to JXY), 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).

Tables

Table 1. Public datasets used for application demonstration.

Figure legends

Figure 1. The framework design and application demonstration of NanoTrans. (A) The overview of the NanoTrans framework. The names of third-party tools employed in each step are denoted in italics. (B) Public datasets used for application demonstration.

Figure 2. 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.

Figure 3. The differentially expressed genes and isoforms identified based on the comparison between the VIRc and vir1 groups. (A) Differentially expressed genes. (B) Differential expressed isoforms. False discovery rate (FDR) cutoff: 0.05. Fold change (log2 transformed) cutoffs: -1 and 1. (C) Biological processes enriched by differentially expressed genes (VIRc vs. *vir1*, adjusted p value < 0.05). (D) 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 4. 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 5. 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 6. Poly(A) tail length differences between the TENT5A^{WT} and TENT5A^{MUT} groups.

(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.

Figure 7. 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.

Figure 8. An HTML report summarizing results. The report contains three sections, in which the Results section gathered the primary files and figures generated by NanoTrans.

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, 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

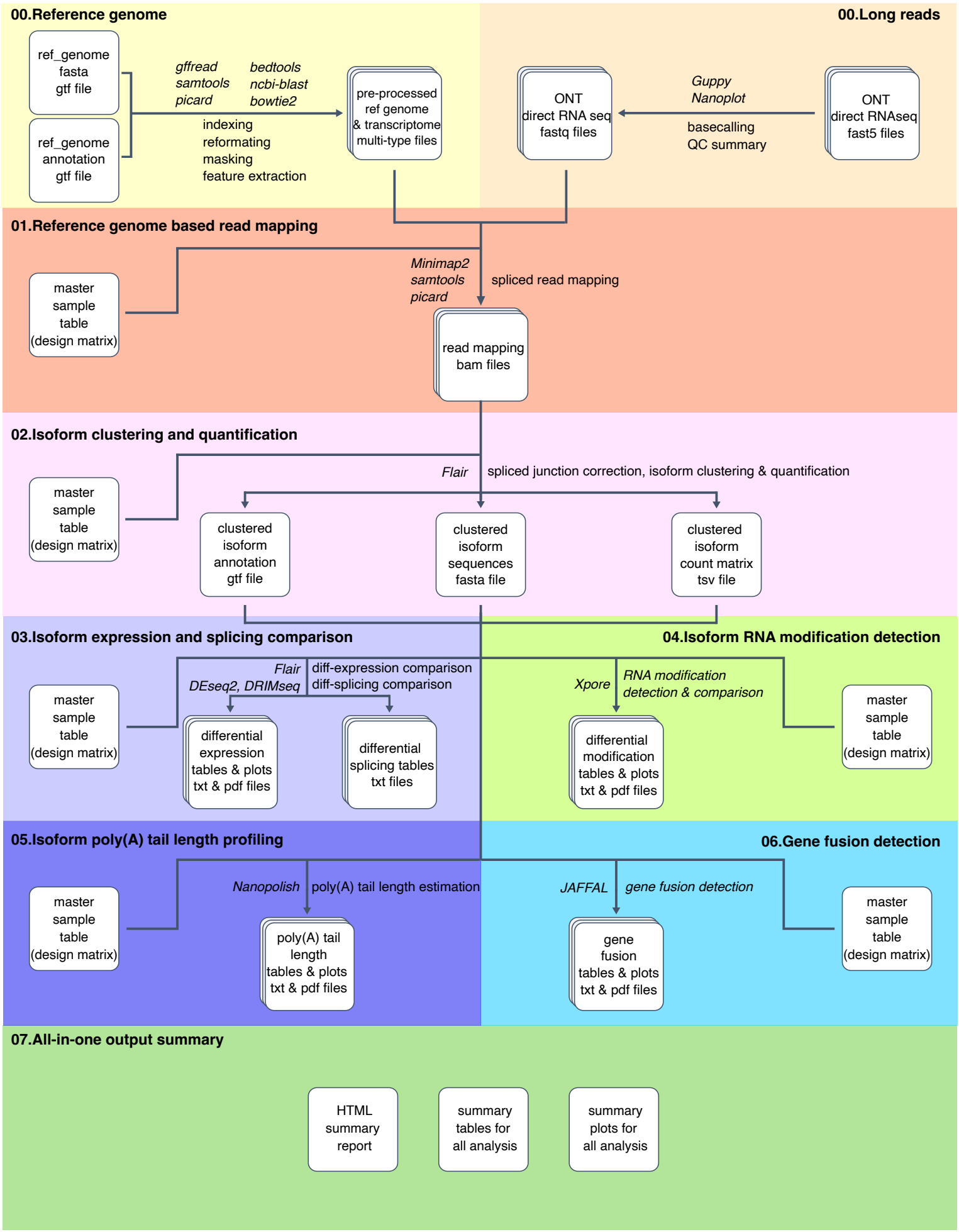
Loman, N.J., Quick, J., Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12, 733–735. https://doi.org/10.1038/nmeth.3444

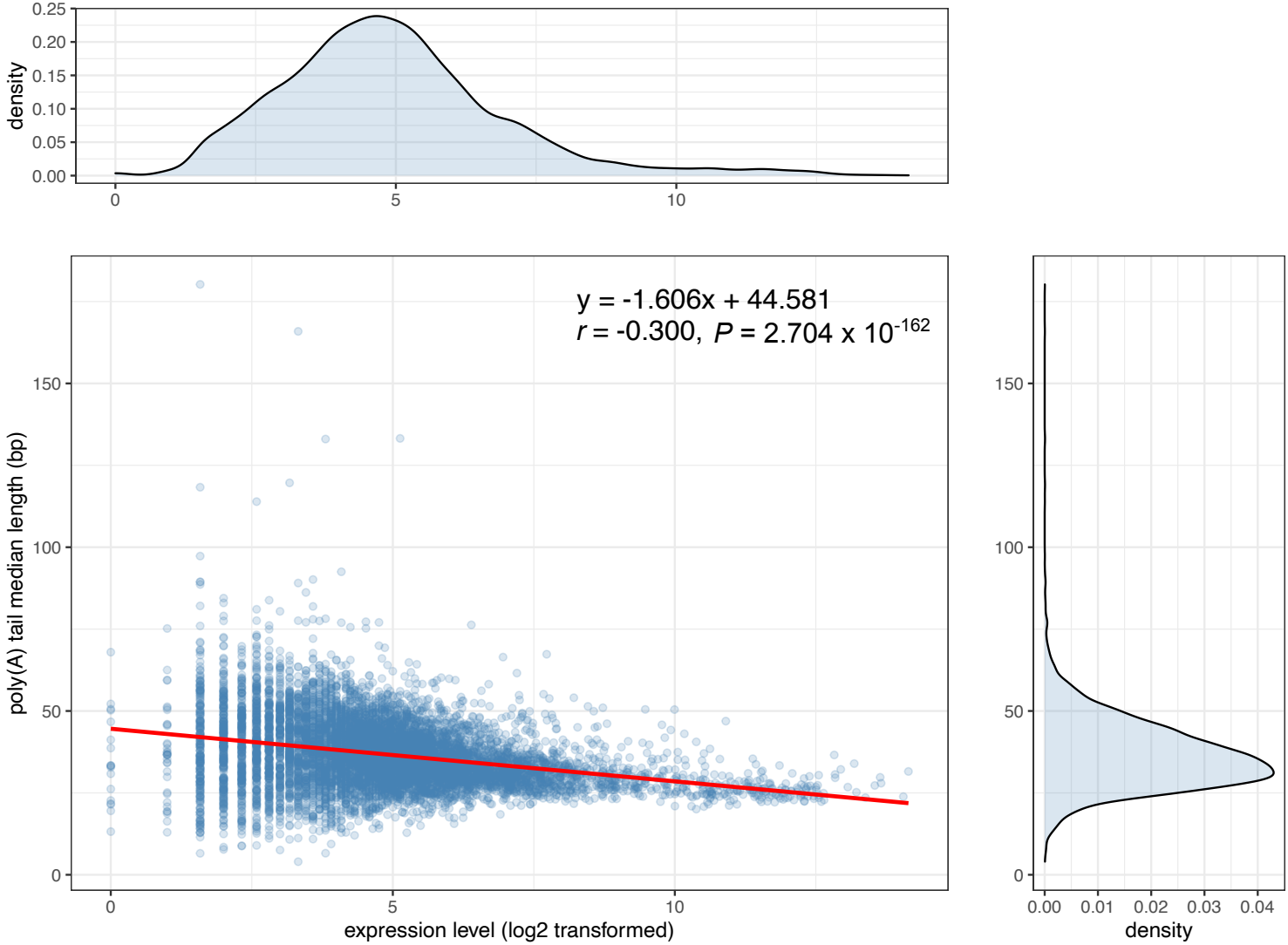
Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D., Hall, A.J., Barton, G.J., Simpson, G.G., 2020. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife 9, e49658. https://doi.org/10.7554/eLife.49658

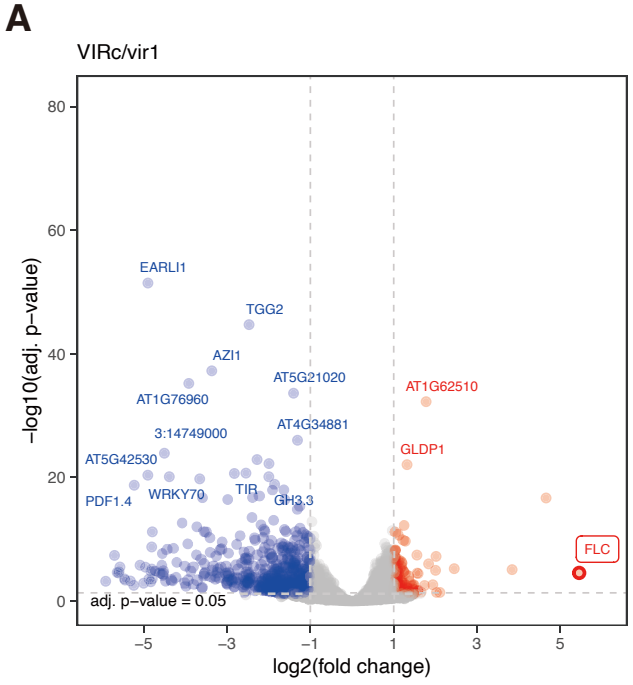
Pratanwanich, P.N., Yao, F., Chen, Y., Koh, C.W.Q., Wan, Y.K., Hendra, C., Poon, P., Goh, Y.T., Yap, P.M.L., Chooi, J.Y., Chng, W.J., Ng, S.B., Thiery, A., Goh, W.S.S., Göke, J., 2021. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol 39,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

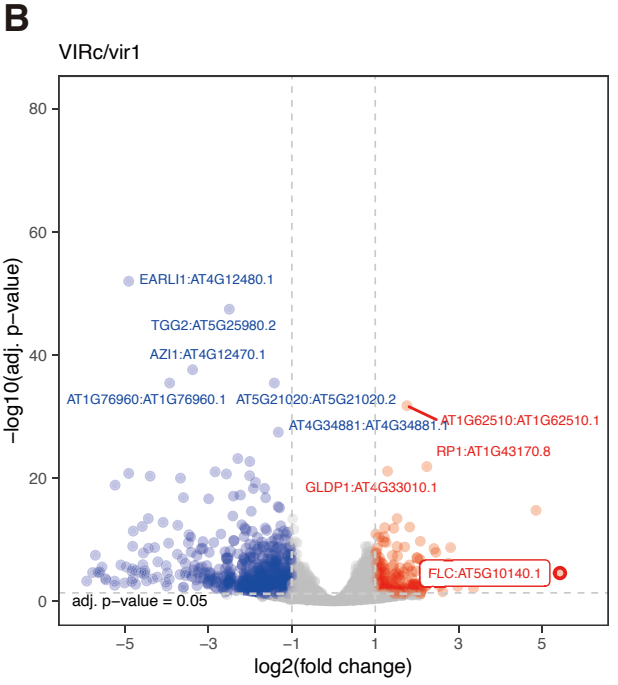
Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., Brooks, A.N., 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun 11, 1438. https://doi.org/10.1038/s41467-020-15171-6

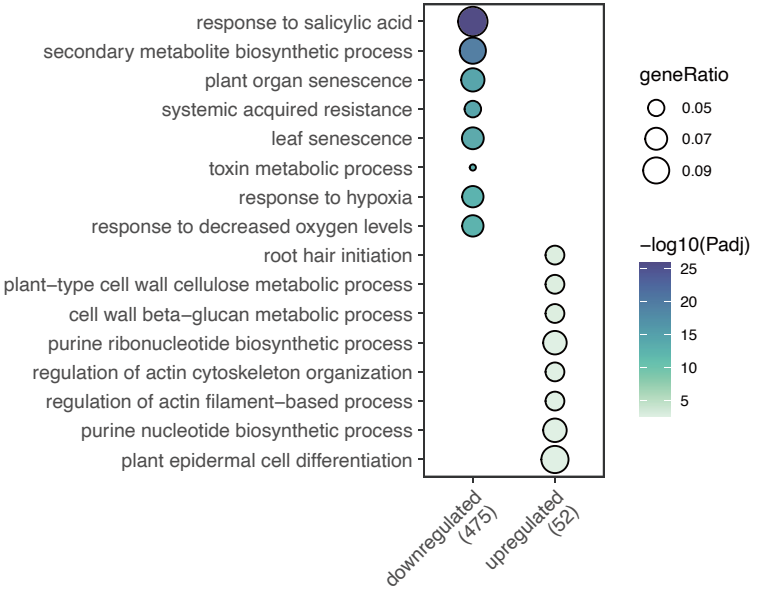
Tudek, A., Krawczyk, P.S., Mroczek, S., Tomecki, R., Turtola, M., Matylla-Kulińska, K., Jensen, T.H., Dziembowski, A., 2021. Global view on the metabolism of RNA poly(A) tails in yeast Saccharomyces cerevisiae. Nat Commun 12, 4951. https://doi.org/10.1038/s41467-021-25251-w

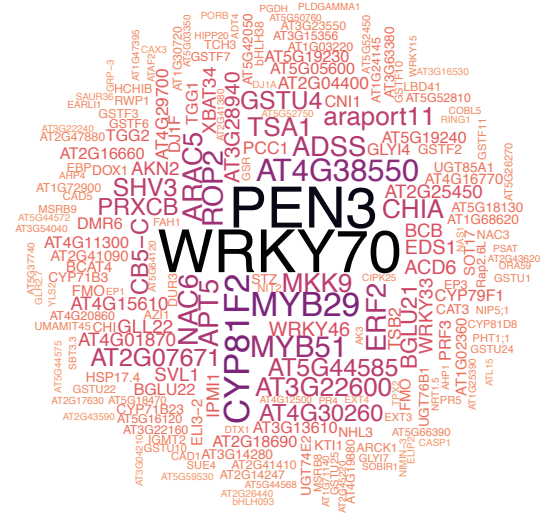

00.Reference genome
ref_genome fasta gtf file
ref_genome annotation gtf file
gffread
samtools
picard
bedtools
ncbi-blast
bowtie2
indexing
reformating
masking
feature extraction
pre-processed ref genome & transcriptome multi-type files
00.Long reads
ONT direct RNA seq fastq files
Guppy
Nanoplot
basecalling
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
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
04.Isoform RNA modification detection
Xpore
RNA modification detection & comparison
differential modification tables & plots txt & pdf files
master sample table (design matrix)
05.Isoform poly(A) tail length profiling
master sample table (design matrix)
Nanopolish
poly(A) tail length estimation
poly(A) tail length tables & plots txt & pdf files
06.Gene fusion detection
JAFFAL
gene fusion detection
gene fusion tables & plots txt & pdf files
master sample table (design matrix)
07.All-in-one output summary
HTML summary report
summary tables for all analysis
summary plots for all analysis

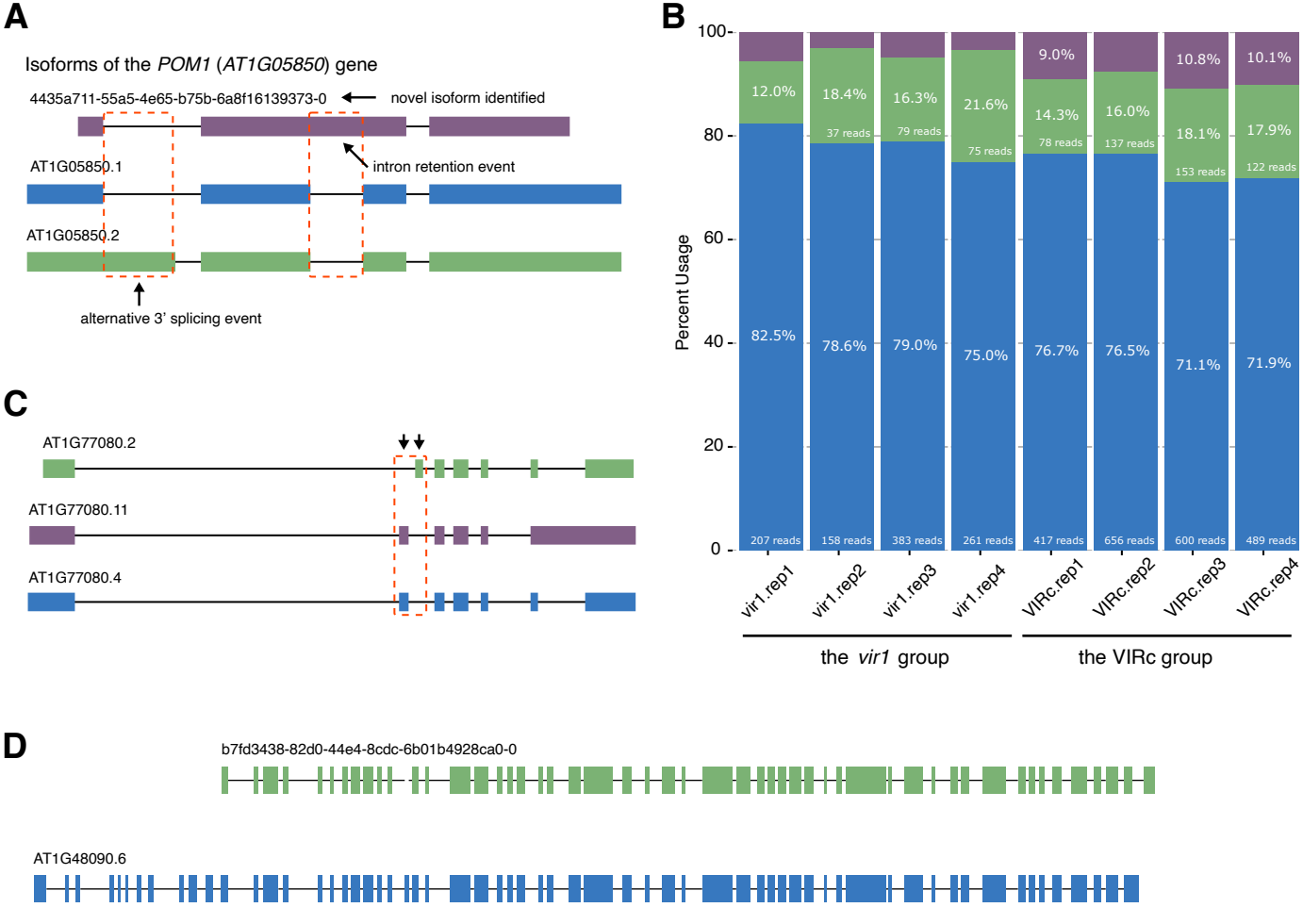

density
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
y = -1.606x + 44.581
r = -0.300, P = 2.704 x 10^-162
poly(A) tail median length (bp)
150
100
50
0
expression level (log2 transformed)
density
0.00
0.01
0.02
0.03
0.04

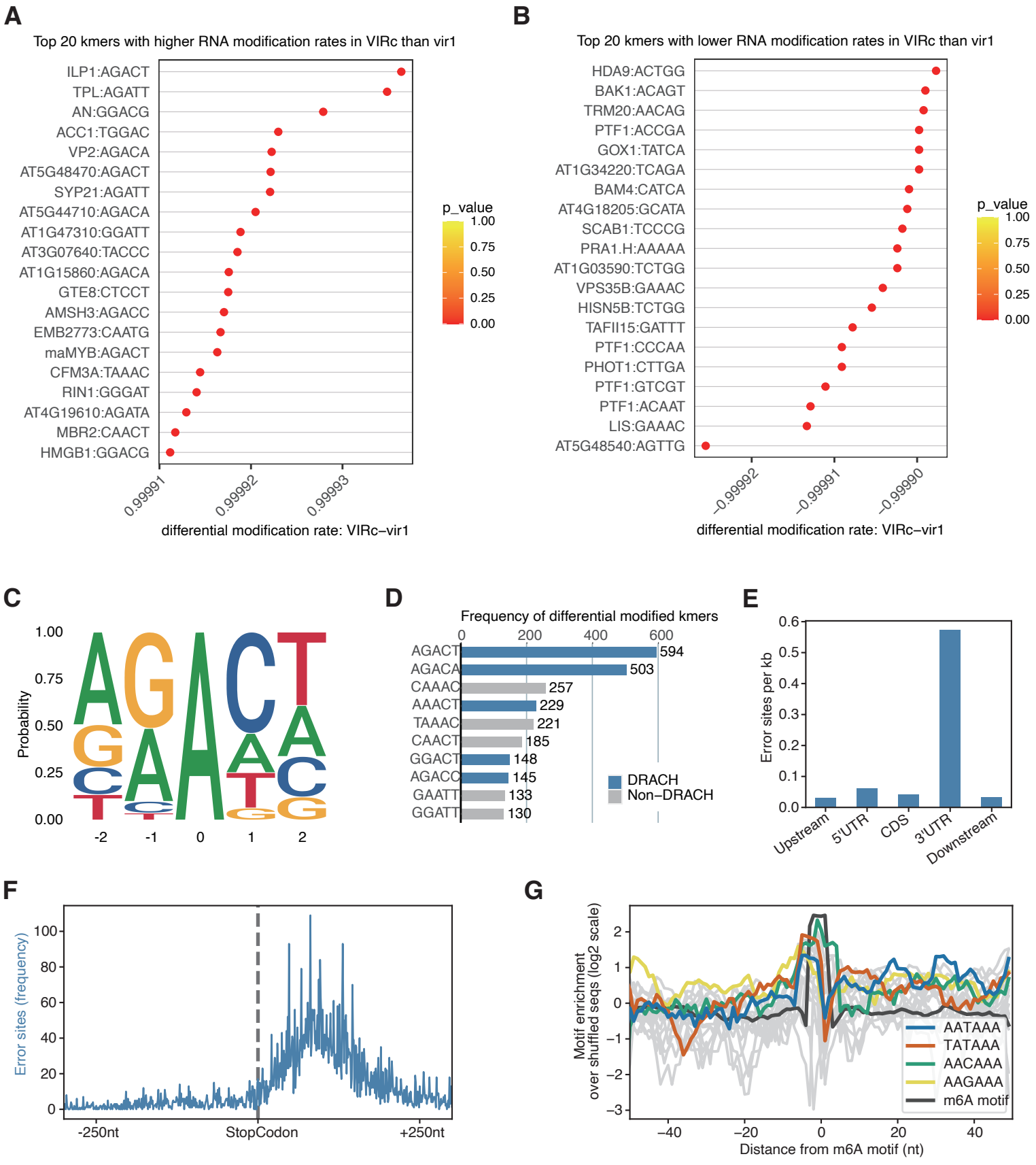

A
VIRc/vir1
−log10(adj. p−value)
log2(fold change)
adj. p−value = 0.05
EARLI1
TGG2
AZI1
AT5G21020
AT1G76960
AT4G34881
3:14749000
AT5G42530
PDF1.4
WRKY70
TIR
GH3.3
AT1G62510
GLDP1
FLC

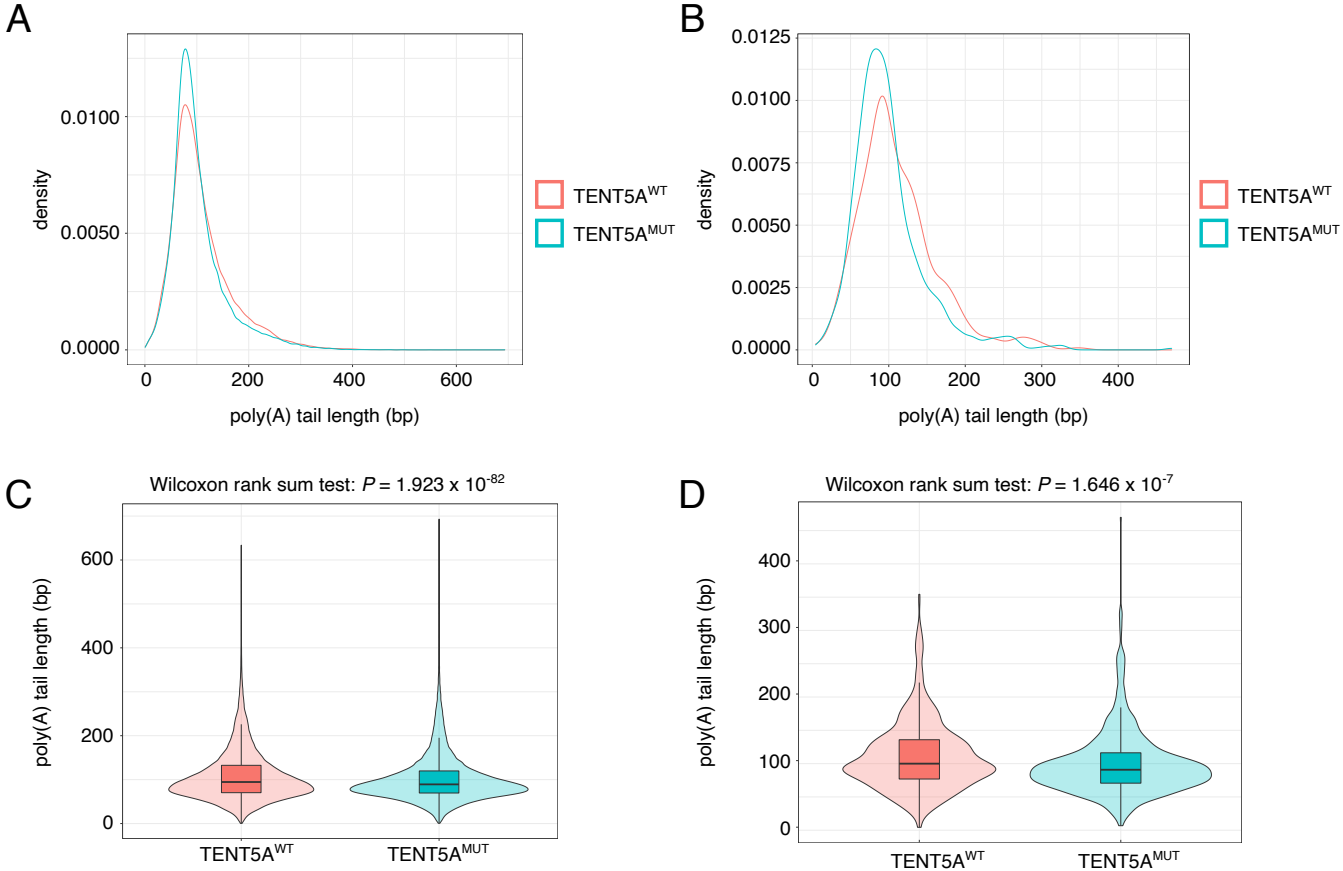

B
VIRc/vir1
−log10(adj. p−value)
log2(fold change)
adj. p−value = 0.05
EARLI1:AT4G12480.1
TGG2:AT5G25980.2
AZI1:AT4G12470.1
AT1G76960:AT1G76960.1
AT5G21020:AT5G21020.2
AT4G34881:AT4G34881.1
AT1G62510:AT1G62510.1
RP1:AT1G43170.8
GLDP1:AT4G33010.1
FLC:AT5G10140.1

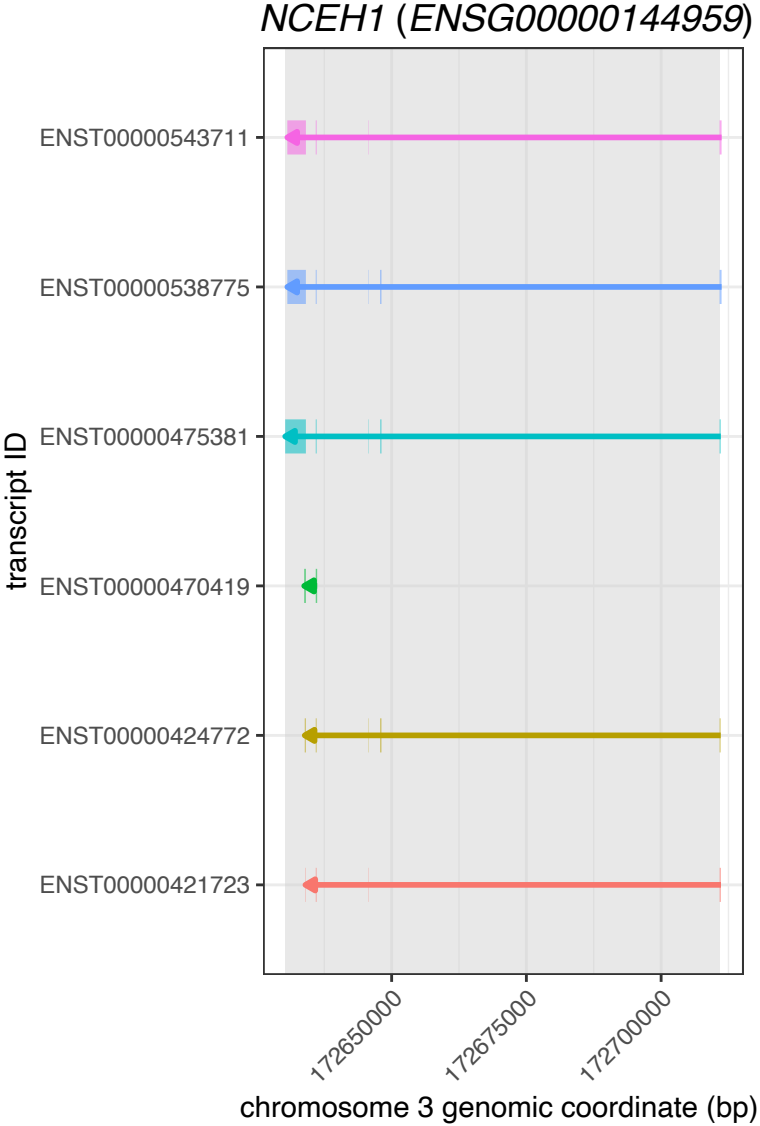

C
response to salicylic acid
secondary metabolite biosynthetic process
plant organ senescence
systemic acquired resistance
leaf senescence
toxin metabolic process
response to hypoxia
response to decreased oxygen levels
root hair initiation
plant−type cell wall cellulose metabolic process
cell wall beta−glucan 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

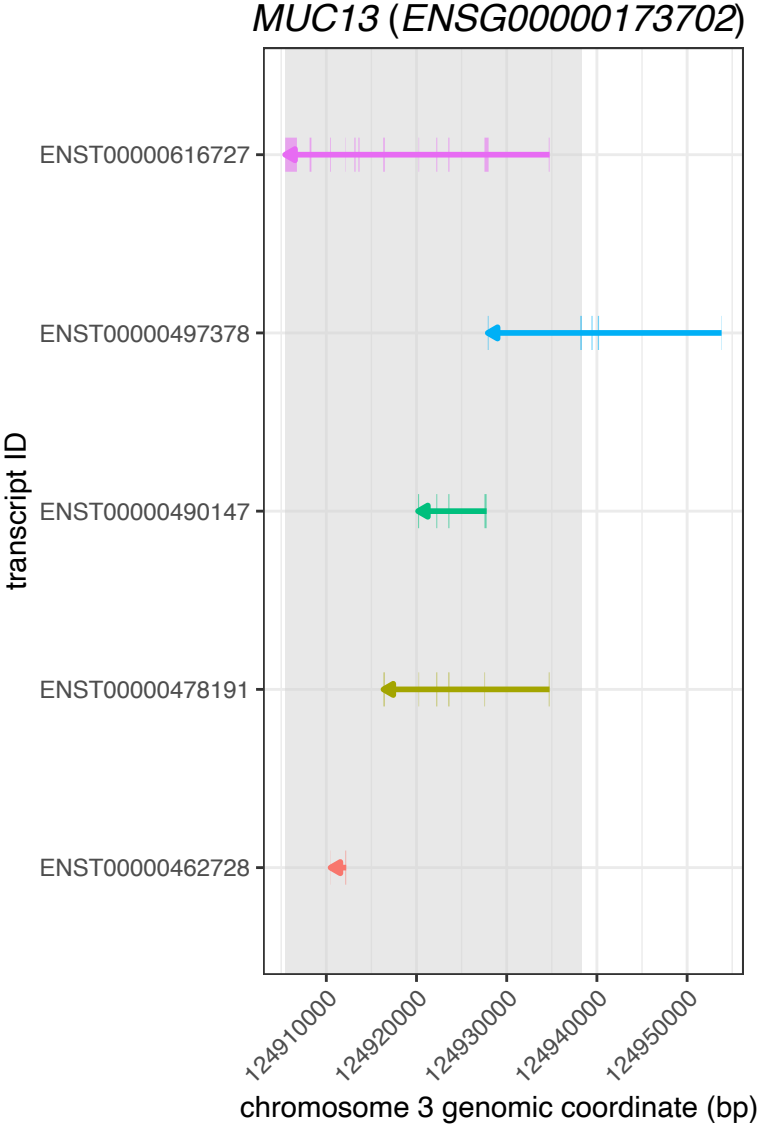

D
WRKY70
PEN3
AT4G38550
araport11
GSTU4
TSA1
ADSS
CHIA
SHV3
PRXCB
MYB29
MYB51
AT5G44585
AT3G22600
AT4G30260
CYP81F2
ERF2
BGLU21
WRKY33
WRKY46
AT2G07671
AT4G01870
AT4G15610
AT2G25450
AT2G04400
AT5G05600
AT5G19230
AT1G03220
AT3G23550
AT3G15356
AT3G28940
AT2G29700
AT2G16660
AT2G41090
AT4G11300
AT2G18690
AT3G13610
AT3G14280
AT2G14247
AT2G41410
AT5G44568
AT2G26440
AT5G59530
AT5G66390
AT5G52810
AT5G18130
AT1G68620
AT4G16770
AT5G19240
AT5G42050
AT5G42450
AT5G42460
AT5G42470
AT5G42480
AT5G42490
AT5G42500
AT5G42510
AT5G42520
AT5G42530
AT5G42540
AT5G42550
AT5G42560
AT5G42570
AT5G42580
AT5G42590
AT5G42600
AT5G42610
AT5G42620
AT5G42630
AT5G42640
AT5G42650
AT5G42660
AT5G42670
AT5G42680
AT5G42690
AT5G42700
AT5G42710
AT5G42720
AT5G42730
AT5G42740
AT5G42750
AT5G42760
AT5G42770
AT5G42780
AT5G42790
AT5G42800
AT5G42810
AT5G42820
AT5G42830
AT5G42840


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
density
0.0100
0.0050
0.0000
0
200
400
600
poly(A) tail length (bp)
TENT5A^WT
TENT5A^MUT
B
density
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
100
200
300
400
poly(A) tail length (bp)
TENT5A^WT
TENT5A^MUT
C
Wilcoxon rank sum test: P = 1.923 x 10^-82
poly(A) tail length (bp)
600
400
200
0
TENT5A^WT
TENT5A^MUT
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


NCEH1 (ENSG00000144959)
ENST00000543711
ENST00000538775
ENST00000475381
ENST00000470419
ENST00000424772
ENST00000421723
transcript ID
172650000
172675000
172700000
chromosome 3 genomic coordinate (bp)


MUC13 (ENSG00000173702)
ENST00000616727
ENST00000497378
ENST00000490147
ENST00000478191
ENST00000462728
transcript ID
124910000
124920000
124930000
124940000
124950000
chromosome 3 genomic coordinate (bp)

NANOTRANS

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	basec

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	p_valu

3.2 Figures

DEGs | DEIs | RNA Modifications | Top diff Kmers | Comparison of PolyA Lengths | Fusion Genes