BMC Bioinformatics

SOFTWARE

Open Access

CRESCENT, a comprehensive RNA-Seq expression, splicing, and coding/non-coding element network tool

Gilles Sireta[1], Gwendal Cueff[2], Vincent Darbot[3], Marie Lefebvre[4], Simon Amiard[1], Aline V. Probst[1] and Christophe Tatout[1*]

*Correspondence:
Christophe Tatout
christophe.tatout@uca.fr
[1]Université Clermont Auvergne, CNRS, INSERM, iGReD Clermont-Ferrand, Clermont-Ferrand, France
[2]UMR 1019, INRAE, Unité de Nutrition Humaine, Université Clermont Auvergne, 63000 Clermont-Ferrand, France
[3]Aviwell SAS, TWB Campus, 31077 Toulouse, France
[4]Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, 63000 Clermont-Ferrand, France

Abstract

Background Traditional short-read RNA-Seq analysis pipelines predominantly focus on protein-coding genes, often overlooking other genomic sequences such as transposable elements (TEs) and non-coding RNA dynamics and do not usually investigate splicing events or transcript usage. To fully capture the complexity of the transcriptome, and in particular transcriptomic regulation, it is crucial to adopt a comprehensive approach that integrates these diverse aspects, providing a more complete and nuanced understanding of expression dynamics in the studied organism.

Results To address these limitations, we present CRESCENT (Comprehensive RNA-seq Expression, Splicing, and Coding/non-coding Element Network Tool), a Snakemake workflow capable of performing a fully automated and comprehensive RNA-Seq analysis. CRESCENT integrates multiple tools at each step of the workflow and enables analysis of differential expression, differential alternative splicing, differential transcript usage, and gene ontology-based functional enrichment for all three. The workflow takes advantage of multiple Snakemake wrappers to minimize required installations for the user, integrating the latest versions of popular bioinformatic tools. It can be run for a complete analysis or for only a specific part in accordance with the configuration file provided by the user. The CRESCENT workflow was validated, demonstrating the pipeline's reliability, as differentially expressed protein-coding genes, TEs and differential alternative splicing events were consistent with previously published datasets. Finally, benchmarking CRESCENT performance indicated that it can be run on a personal computer or a remote server, including a high-performance computing cluster, allowing a user to process small single-end sequencing on species possessing a small genome like *Arabidopsis thaliana* to very large paired-end sequencing on polyploid species like wheat.

Conclusion and availability CRESCENT is a scalable solution for comprehensive transcriptomic profiling. It is freely available at https://github.com/gilless429/crescent.

Keywords RNA-Seq workflow, Snakemake, Alternative splicing, Protein coding genes, Transposable elements, Non-coding RNA

Background

Reproducibility is a well-known challenge in science [1] including bioinformatics, particularly for multi-step workflows [2], which are difficult to standardize due to the large range of available tools and the numerous parameters that must be configured at each step. Workflow Management Systems (WMSs) present a solution to the reproducibility challenge by wrapping complicated tasks and dependencies into modular, reusable pipelines [2]. Galaxy, Snakemake and Nextflow are among the most widely adopted WMSs in bioinformatics. Galaxy [3] adopts a graphical user interface (GUI) facilitating non-expert use and and facilitates the transfer of workflows from one user to another via web links. Snakemake [4] is based on a Python-based syntax, making workflows accessible and easy to adapt for bioinformaticians, further facilitated by the use of interoperable configuration file formats such as YAML. Nextflow [5] relies on containerization (Docker, Singularity) and environment management (Conda). It prioritizes scalability and portability while supporting reproducibility via community standards like nf-core [6]. Each system addresses reproducibility from a distinct angle: Galaxy by user accessibility through a simple GUI and automatic history recording, Nextflow via strong standards and portability particularly for scaling to larger computational environments, while Snakemake is more flexible and based on a Python-like syntax familiar to bioinformaticians.

Snakemake, like many WMSs supports and promotes reproducibility by following the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for scientific data management [7]. Its workflows are easily shared on git platforms, whether standards like github and gitlab, or dedicated platforms like WorkflowHub (https://workflowhub.eu/). In recent years, reproducibility was improved further by the addition of Snakemake wrappers [8] to simplify the integration and handling of compatibility for the bioinformatics tools used in workflows, reducing end user environment management requirements. By using wrappers (https://Snakemake-wrappers.readthedocs.io/en/stable/wrappers.html), workflows become more maintainable, reproducible, and easily shareable, while benefiting from community-contributed and tested implementations of common bioinformatics tools.

RNA-Seq analysis has become a standard method to study the transcriptome at the whole-genome scale. Several workflows have been published for RNA-seq analysis, including Galaxy workflows published through the popular Galaxy training network [17], several Snakemake-based workflows such as RASflow [9], ARMOR [10], VIPER [11], hppRNA [12] or transXpress [13], and Nextflow-based workflows like nf-core/rnaseq [6], nf-rnaSeqCount [14] or RNAflow [15] (**Supplementary Material 1** Table 1).

While these workflows demonstrate the versatility of WMSs in creating efficient RNA-seq analysis pipelines for gene expression analysis, they do not systematically explore key features such as Transposable Elements (TEs), non-coding RNAs (ncRNAs), nor provide Differential Alternative Splicing (DAS) and Differential Transcript Usage (DTU) tools and results. TEs, which may constitute a large fraction of eukaryotic genomes, not only drive genome evolution but also regulate gene expression and interplay with ncRNAs [16]. ncRNAs, varying in size and function (e.g., siRNA, piRNA, lncRNA), act as key regulators of transcriptional and post-transcriptional processes [17] while DAS and DTU are main contributors of transcriptome complexity and proteome diversity [18]. Comprehensive RNA-Seq analyses that incorporate these layers will contribute to deciphering regulatory mechanisms missed by conventional gene-based approaches.

Table 1 Comparative table for tools included in CRESCENT

Workflow Step	crescent Tools	Alternative tools	Wrapper Status	Rationale for Tool Choice
Sequencing QC	FastQC (2010)	RNA-SeQC (2011) RSeQC (2012) Trimmomatic (2014) FASTQE (2021)	Wrapper	Very popular and widely used for fast check of read quality Quick overview to detect any problem on reads HTML reporting
Read trimming	fastp (2018)	Trimmomatic (2014)	Wrapper	Ultra-fast (2–5 times faster than cutadapt) Complete bundle: quality controls, adapter trimming, quality filtering, per-read quality pruning
	cutadapt (2011)	AdapterRemoval (2016)		Compatible with multiple format (fastq, fasta, native or gunzipped) Features: adapter detection, adapter trimming
Genome-based mapping	HISAT2 (2015)	topHat2 (2012),	Wrapper	Graph FM index for alignment Fastest amongst aligners Optimized for single and paired-end reads
	STAR (2013)	Bowtie2 (2012)		Higher sensitivity Efficient for large size genomes Easy integration in workflows Faster than topHat2 or botwie2 Optimized for spliced reads
Transcript quantification	Salmon (2016)	eXpress (2013) Kallisto (2015)	Wrapper	Offers bias correction and selective alignment for higher quantification accuracy
BAM/SAM handling	SAMtools (2009) Sambamba (2015)	Picard tool (2011)	Wrapper	A standard library for manipulating sequence files, highly efficient and well-maintained Speed thanks to multithreading Coverage analysis Filtering capabilities
Read counting	featureCounts (2014)	Rcount (2015)	Wrapper	5–20× faster than htseq-count Uses less memory Proposes more options such as handling multimapping/overlapping
	htseq-count (2011)			Discard multi-mapping and overlapping reads Simple usage: straightforward reads summarisation
Differential Expression Analysis (DEA)	DESeq2 (2009) EdgeR (2013)	Cuffdiff 2 (2013) dearseq (2020)	no	Most used DEA tools Simple script Proposed as R packages
GO enrichment	clusterProfiler (2012)	g:Profiler2 (2024)	no	Flexible GO and KEGG for functional enrichment R scripts Large collection of graphical outputs Custom gene background
Differential Alternative Splicing (DAS)	rMATS turbo v4.3.0 (2020)	MAJIQ (2016) SUPPA2 (2017)	no	Standard method for DAS A new Nature Methods publication released in 2024
Differential Transcript Usage (DTU)	RATs (2019)	Cufflinks (2012) DRIM-Seq (2017)	no	Detection of DTU with transcript-level quantification Works with Kallisto Requires wasabi when using Salmon

Table 1 (continued)

Workflow Step	crescent Tools	Alternative tools	Wrapper Status	Rationale for Tool Choice
Reporting	MultiQC (2016)	fastp (2018)	Wrapper	Open-source tool Aggregating output logs from > 150 bioinformatics tools Single interactive HTML report with data export Best solution to integrate multiple sources

To address these limitations, we selected Snakemake for its Python-based syntax and its user-friendly debugging capabilities, which are generally simpler than those of Nextflow. We designed a new RNA-Seq workflow that enables exploration of genomic features including genes, TEs and ncRNA for differential expression analysis (DEA), as well as proposing investigation of RNA splicing, and gene ontology (GO) enrichment results for both expression and splicing results, providing a holistic approach to transcriptomics research. We have called this workflow CRESCENT for Comprehensive RNA-Seq Expression, Splicing, and Coding/non-coding Element Network Tool available at https://github.com/gilless429/crescent.

Implementation

CRESCENT is a Snakemake RNA-seq analysis workflow that offers flexibility in reference genome, transcriptome, and annotation usage including "gold-standard" tools in NGS bioinformatics, well known by the community, to facilitate collective contributions and further evolution of the pipeline (Table 1). CRESCENT also incorporates Snakemake wrappers [8] when their implementation matches the pipeline's needs. Tools installed via conda outside of the wrappers use the latest versions except for Snakemake (=8.20.3), Snakemake-minimal (=8.20.3) and Snakemake-wrapper-utils (=0.6.2) for which specification of versions was needed to ensure the full functionality of CRESCENT, especially as new versions of Snakemake become available and risk altering the workflow.

The workflow begins with quality control and trimming of sequencing reads, followed by read quantification or mapping to the reference, then feature assignment through read counting. For read counting, while htseq-count and featureCounts are made available, the latter is preferred for TE quantification due to its faster speed and flexible handling of multi-mapping and overlapping reads, including options to discard or fractionally assign multi-mapped reads. For ncRNA isoform ambiguity, Salmon uses probabilistic methods for read assignment, while CRESCENT supports strand-specific libraries, paired-end sequencing, and longer reads to better distinguish similar transcripts, ncRNA or repeated sequences during the mapping process. The workflow is designed for a comprehensive transcriptomic analysis including Differential Expression Analysis (DEA) as well as using Differentially Alternative Splicing (DAS) and Differential Transcript Usage (DTU) pipelines to study differential splicing between the compared groups of samples. DAS is performed using rMATS (replicate Multivariate Analysis of Transcript Splicing) [19], recently updated in its rMATs turbo v4.3.0 version running faster and implemented with new features [20] outperforming others among 21 different tools in a recent benchmarking assay [21]. DTU analysis is performed with the R package RATs (Relative Abundance of Transcripts) [22], which uses transcript quantification to identify genes with differentially abundant isoforms and define the proportion

of each isoform in respect to the total gene expression. RATs was chosen for its high sensitivity in detecting significant differential splicing, resulting in the largest number of DTU in benchmarking [21]. Note that RATs requires transcript-level quantification, for which we chose to use salmon [23] in spite of RATs natively handling quantification results from Kallisto [24], not salmon. Conversion between Kallisto and salmon's output formats has to be handled via the wasabi R package first. Salmon was chosen over Kallisto in spite of RATs expecting Kallisto output due to its more advanced features such as selective alignment (which improves quantification accuracy) and bias correction (GC bias for example), along with a wider adoption by the bioinformatic community at present. DESeq2 and edgeR were selected for DEA as robust, very popular and complementary statistical methods although dearseq, a more recent tool outperforms DESeq2 and EdgeR in recent benchmarking assay [25], could be a relevant alternative. Finally, Multiqc analysis is used to generate reports summarizing the results of sequencing quality control, trimming, mapping, transcript quantification and read counting. CRESCENT is also capable of using mapped bam files as input, though this only allows DEA and DAS analysis. Functional enrichment analysis of genes found by DEA, DAS and DTU is then performed via GO using ClusterProfiler package and its enrichGO() function, applying multiple testing corrections through the Benjamini-Hochberg (BH) false discovery rate (FDR). Finally, a report is generated at the end of the analysis as a comprehensive MultiQC summary of the bioinformatic steps taken (Fig. 1 and Table 1). CRESCENT is built for modular execution, allowing each analysis branch (DEA, DTU, DAS) to run independently or in combination, and analyses can be paused at intermediary steps. CRESCENT takes advantage of a configuration file containing many parameters, each accompanied by a detailed description of the required input. To ensure reproducibility, the complete CRESCENT configurations for the three datasets used in this study are available in **Supplementary Material 1–3**.

For more advanced users, extra parameters can be entered for each tool that possesses additional options to the ones used. The user can enter these parameters as an additional variable, written exactly as if they were given in the standard command line, and they will be applied. As fully automated workflows can introduce errors, CRESCENT is designed to identify issues, often specifying the problematic parameter if an error arises and listing the valid input options to allow corrections. Through Snakemake's input–output dependency handling, it resumes from the point of failure without restarting entirely, ensuring essential human quality control in bioinformatics analyses. Finally, to ensure that all necessary input data are properly located and to facilitate the systematic generation and storage of analysis outputs, the analysis is run using a specific folder structure, which needs to be set up at the beginning of the analysis (**Supplementary Material 5** Fig. 1).

Results

Differential expression analysis

To explore the functionalities of CRESCENT, we took advantage of a published RNA-Seq analysis investigating the consequences of a 24 h heat stress at 37 °C applied to *Arabidopsis thaliana* seedlings [26]. Sequencing data (poly-A RNA-Seq library, 150 bp paired-end, not stranded— **Supplementary Material 2**) were obtained from the Gene Expression Omnibus (GEO) database (project reference: GSE132415) in the form of

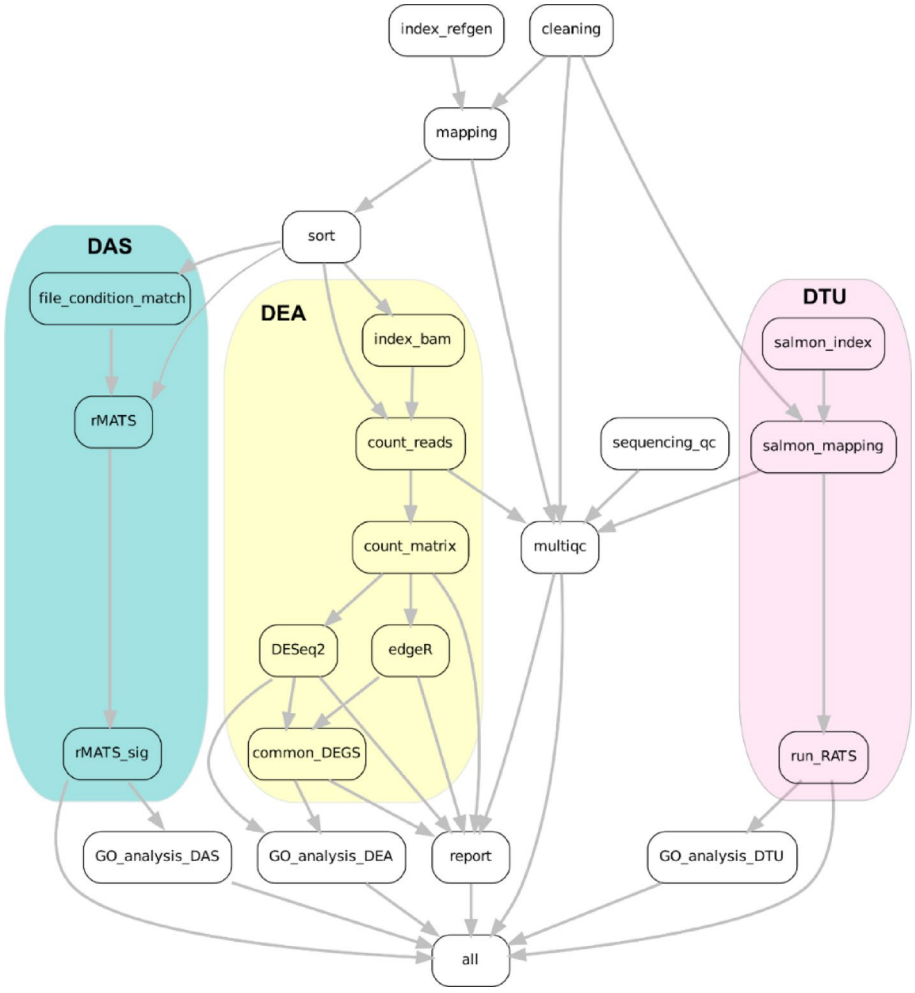

Fig. 1 Directed acyclic graph of the CRESCENT RNA-Seq workflow. CRESCENT begins with raw read quality control (sequencing_qc) and trimming (cleaning). Reference genome/transcriptome indices are generated if required (index_refgen, salmon_index), then reads are quantified using Salmon (salmon_mapping) for DTU (Differential Transcript Usage), or aligned to the reference genome (mapping) for DAS (Differential Alternative Splicing). DEA can use either salmon or genome alignment results. Aligned BAM files are processed through sorting (sort) and indexing (index_bam). For DEA, reads are assigned to features (count_reads), then a matrix unifying read counts is produced (count_matrix) so DEA proper can be carried out using DESeq2 and edgeR, producing each a list of DEFs, a normalized expression matrix, and a PCA plot. A common list to both analyses is produced (common_DEGs). DAS events are detected by rMATS turbo in a two-step process: first all detected DAS events are produced (rMATS) and then filtered using user-defined statistical significance thresholds (rMATS_sig). RATs and provides tables of the genes and transcripts involved in DTU. Functional enrichment is performed via Gene Ontology (GO_analysis) for DAS/DEA/DTU, providing plots for each type of ontology (Biological Process, Molecular Function, Cellular Compartment). QC summaries are generated (multiqc) at several steps and merged into a single report, ensuring reproducibility and transparency throughout the workflow

raw fastq.gz sequencing read files. Sequencing QC and trimming were performed via fastqc and cutadapt (removing all reads shorter than 70 bp as in [26]) followed by mapping using hisat2 (TopHat2 was used in [26]) and read counting with featureCounts. This process can be executed on fastq or fastq.gz files, but the reference genome, transcriptome, and annotation need to be uncompressed before the mapping step. The DEA was then performed between heat treated and control samples with DESeq2 (as in [26]), with a threshold of adjusted pValue (pADJ) < 0.05 and a threshold of $|\log2FC| > 1$ ($FC > 2$). 4621 Differentially Expressed Genes (DEGs), 2534 up and 2087 down, were identified. There was significant overlap with the original analysis [26] (4255/4621

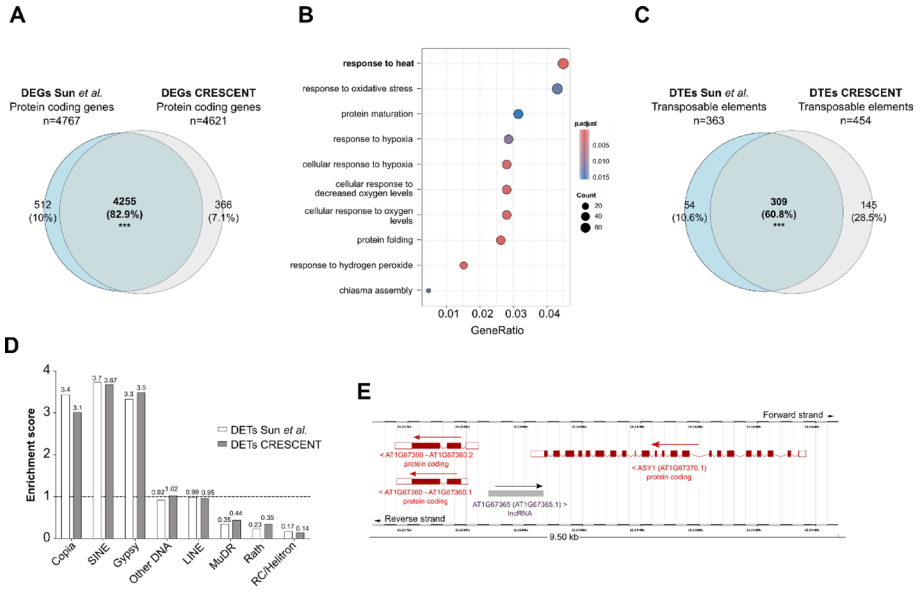

Fig. 2 Analysis of differential expressed genes and TEs. DEA of the RNA-Seq dataset GSE132415 from *Arabidopsis thaliana* comparing mock and heat shock treated seedlings [11]. It includes 4 paired-end unstranded sequencing samples, 2 of them from heat stress conditions (37 °C/22 °C 12 h alternation for 3 days; GSM3863058 and GSM3863059) and 2 samples at normal growth conditions (22 °C; GSM3863056 and GSM3863057). **A** Significant overlap between the Differentially Expressed Genes (DEGs) obtained with CRESCENT and found by Sun et al. 2020 using DESeq2 with log2FC ≥|1| and adjusted pValue ≤ 0.05 thresholds. **B** GO analysis from CRESCENT (clusterProfiler) highlighting the enrichment in the biological process "response to heat" within the upregulated DEGs. **C** Significant overlap of the Differentially expressed Transposable Elements (DTEs) between CRESCENT and [11]. **D** Distribution of TEs expressed as the percentage of TEs in Arabidopsis genomic sequences (white) and within DEGs defined in the CRESCENT output. Statistical significance in **A** and **B** was assessed by Fischer tests (overlap p-value ≤ 0.01: ***) with N at 21,112 for DEGs and 1982 for DTEs. Here, N is defined as the total number of genes or TEs with ≥ 10 reads in at least one of the samples. **E** Modified genome browser representation obtained from EnsemblPlant showing *AT1G67365-EL0117* (TAIR10-PLncDB_ID respectively) lncRNA located upstream and in antisense orientation of the *ASY1* locus. *AT1G67365* and *ASY1* are up-regulated during long heat-stress in the RNA-Seq dataset GSE132415

Table 2 Performance assay. CRESCENT performance assessed with four RNA-Seq datasets from three distinct species and two different computing environments with 4 threads & 8 GB RAM or 16 threads & 32 GB RAM. To simplify the processing of PRJEB59062, the analysis was restricted to one Fusarium strain (MDC_Fg1) at two timepoints (48- and 96-h post-infection), keeping only one technical replicate for each biological replicate. A total of eight paired-end sequencing samples were thus used: four for 48-h (ERR10886903, ERR10886906, ERR10886909, ERR10886912) and four for 96-h (ERR10886927, ERR10886930, ERR10886933, ERR10886936). Processing time was assessed for DEA, DTU and DAS analysis but does not include GO analysis as it is strongly dependent on the dataset's underlying biological reality (longer if number of DEGs is greater)

Accession	Species	Reference genome (GB)	Fastq.gz (GB)	4 threads/8 GB RAM	16 threads/32 GB RAM
Test dataset	*Arabidopsis thaliana*	0.116	0.193	0h8m37s	0h06m09
GSE132415	*Arabidopsis thaliana*	0.116	12.3	3h31m47s	0h39m53s
GSE18508	*Drosophila melanogaster*	0.140	4.7	1h18m36s	0h16m41s
PRJEB59062	*Triticum aestivum*	17	64.1	CAN'T	2h14m2s

DEGs, overlap pValue ≤ 0.01) (Fig. 2A and **Supplementary Material 1** Table 2), the observed differences most likely reflecting the impact of distinct mapping tools between the two workflows. GO analysis performed for up-regulated genes illustrates the successful identification of response to heat as one of the biological processes significantly

affected during long heat stress (Fig. 2B). 475 Differentially Expressed Transposable Elements (DTEs), mostly up-regulated (454 of 475), were identified with a high overlap with the initial study (309/454 DTEs, overlap pValue ≤ 0.01) (Fig. 2C**)** and with very similar enrichment scores for TE families (Fig. 2D and **Supplementary Material 1 Table 3**). The Ty1/Copia-type Long Terminal Repeat (LTR) retrotransposon family *ATCOPIA78*, also known as *ONSEN* was identified among the upregulated TEs (**Supplementary Material 1 Table 3**). This TE family is well-established to be derepressed during long heat stress [27, 28]. Finally, 116 differentially expressed ncRNA were identified allowing for investigation of the expression of microRNA (miR) genes, whose functions are often well conserved across species. Among the miR genes identified (**Supplementary Material 1 Table 4**), miR156c, miR159b, miR164b**,** and miR166b were previously described as key regulators of the heat stress response [29]. Among the 67 'other-rna' (TAIR10 annotation) transcripts affected by long heat stress (**Supplementary Material 1 Table 4**), 25 were annotated as lncRNA in the PLncDB database [30], with 10 of them previously reported as responsive to heat stress. Notably, of the heat-responsive lncRNAs, *AT1G67365* was significantly up-regulated during long heat stress and is located downstream of *ASYNAPTIC 1* (*ASY1*, *AT1G67370*), which was also up-regulated under this same stress condition (Fig. 2E). *ASY1* is a well-characterized meiotic gene essential for homologous chromosome pairing, synapsis, recombination, and crossover interference leading to meiotic defects under heat stress condition [31]. The co-regulation of *AT1G67365* and *ASY1* raises the hypothesis that this lncRNA may influence the heat stress response of *ASY1*. This observation exemplifies how CRESCENT analysis facilitates the identification of candidate regulatory relationships to guide subsequent functional studies. From this first DEA analysis, significant overlap was observed with previously published results in good agreement with the known characteristics of long heat stress response in *Arabidopsis thaliana*, validating the automated analysis performed by CRESCENT.

DAS and DTU analysis

To explore the relevance and validity of the workflow for DAS and DTU analysis, we used the RNA-Seq analysis performed by [32] studying the *Drosophila* homolog of the mammalian splicing regulators Nova-1 and Nova-2, which is also a popular reference-based RNA-Seq dataset of the Galaxy training network [17]. In this dataset (Poly-A RNA-Seq library, 50 bp single-read, not stranded— **Supplementary Material 3**), a Drosophila strain in which the *Pasilla* (*PS*) gene is downregulated by RNA interference (RNAi) was compared with the wildtype strain. Sequencing data were obtained from the NCBI GEO archive (project reference: GSE18508). CRESCENT detected 223 DTUs and 197 unique DAS (FDR ≤ 0.05) (**Supplementary Material 1 - Table 5 and 6**). DAS analysis revealed a large proportion of alternative splicing events to be skipped exons (DAS_SE: 46.1%) as previously reported for Drosophila [33] (Fig. 3A). 16.5% of the genes with alternative splicing events also display different transcript usage (47 genes, overlap pValue ≤ 0.01) (Fig. 3B). Overlap was observed between our DAS results and genes with differential splicing identified in the original study (104 genes, overlap pValue ≤ 0.01) **(**Fig. 3B**)**. More specifically, when considering skipped exon events (referred to as "cassette exon" in [32]) genes like *bmm* (FBgn0036449), *cg* (FBgn0000289), *osa* (FBgn0261885), *msn* (FBgn0010909) and *SesB* (FBgn0003360), discussed in [32], were included among the

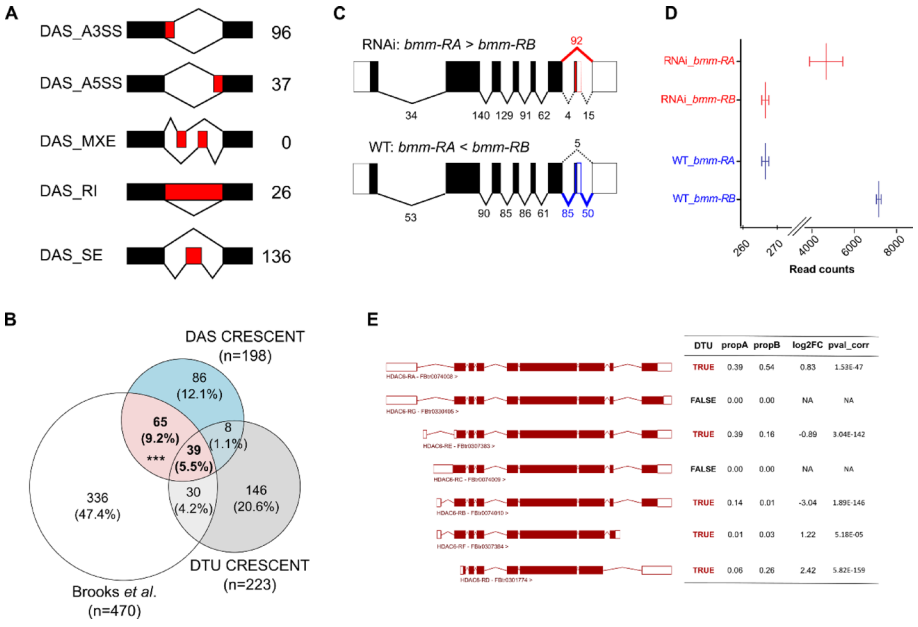

Fig. 3 Analysis of differential alternative splicing events and isoform abundance. DAS and DTU analysis for the GSE18508 dataset containing paired-end sequencing samples including 4 wild-type samples (GSM461176, GSM461177, GSM461178, GSM461182) and 3 *pasilla* RNAi samples (GSM461179, GSM461180, GSM461181). **A** Results of alternative splicing events analysis with constitutive exon(s) (black) and alternatively spliced exon(s) (red) corresponding to the five major types of alternative splicing patterns: alternative 5′ splice site (A5SS), alternative 3′ splice site (A3SS), mutually exclusive exons (MXE), retained intron (RI) and skipped exons (SE). Number of events detected by CRESCENT (FDR ≤ 0.05) is indicated on the right. **B** Significant overlap between DAS and DTU analysis performed by CRESCENT (FDR ≤ 0.05) and the DAS analysis performed by Brooks et al. 2011 (corrected p-value ≤ 0.05). The overlap between Brooks et al. 2011 and DAS CRESCENT (in pink) was statistically significant (overlap p-value ≤ 0.01) when assessed by Fischer test where N = 10,216 is the total number of expressed genes for which ≥ 10 reads were assigned in at least one sample. **C** Alternative splicing of the *bmm* locus from DAS analysis performed by CRESCENT. Schematic representation of the two *bmm* isoforms with *bmm-RA* (7 exons) and *bmm-RB* (8 exons) on the left with read count indicated for each splicing event. **D** Isoform abundance from the DTU analysis as read counts for RNAi (red) and WT (blue). **E** DTU analysis of the 7 alternative transcripts at the *Histone Deacetylase 6* locus identified as a DTU-only locus detected by CRESCENT. 5 out of 7 *HDAC6* transcripts are significantly differentially expressed (up or down as indicated by Log2FC) between RNAi-pa (propA) and WT (probB) genotypes

30 genes common to our analysis and theirs (**Supplementary Material 1 - Table 7**). Looking closely at *bmm*, we also confirmed as in [32] that the abundance of the two isoforms called *bmm-RA* (8 exons) and *bmm-RB* (7 exons), with and without the potentially skipped exon respectively, differed between *ps* RNAi and wild type strains. The *bmm-RA* isoform was more abundant in *ps* RNAi than in wild type in our results, just as in [32] (Fig. 3C and 3D). As DTU analysis was not included in [32], we used a recent study applying DTU detection in Drosophila, leading to the detection of 662 DTU events with a significance threshold of pVal < 0.05 [34]. We selected the *HISTONE DEACETYLASE 6* (*HDAC6*) gene, which is common to this study and the 146 DTU events detected by the CRESCENT analysis and illustrate in Fig. 3E a representative DTU output from CRESCENT. *HDAC6* expresses seven alternative transcripts, five of which exhibit significant DTU in the Brooks et al. dataset. Thus, CRESCENT 's DAS and DTU analyses effectively identified significant alternative splicing and transcript abundance changes in *Drosophila*, validated by overlap with previous findings and detailed isoform-level analysis, such as the differential inclusion of cassette exons in the *bmm* gene or differential transcript usage at the *HDAC6* gene.

Performance

To test the performance of CRESCENT, we used two distinct configurations in term of CPU threads and RAM and four datasets representing organisms with small (0.1–0.2 Gb) and large (16 Gb) genome sizes to illustrate the potential of CRESCENT. We used *Arabidopsis* (0.18 Gb) datasets, one small test dataset available on GitHub and the dataset analyzed in Fig. 2 [26], the *Drosophila* (0.13 Gb) dataset used in Fig. 3 [32] and a dataset from *Triticum aestivum* (wheat), a species with a large genome size (~ 16 Gb). For wheat, RNA-Seq data from the PathoV_NewMyco dataset (Poly-A RNA-Seq library, 150 bp paired-end, not stranded—**Supplementary Material 4**) [35] was obtained from the European Nucleotide Archive (ENA) database (project reference: PRJEB59062). The execution time ranged from 6 min to 2 h depending on the dataset and hardware (Table 2).

Conclusion

CRESCENT is a Snakemake workflow for transcriptome analysis. Like other Snakemake workflows such as ARMOR [10], hppRNA [12], RASflow [9], transXPress [13] or VIPER [11], CRESCENT performs classical DEA while offering flexibility in reference genome/transcriptome usage and employing multiple tools per analysis step, but distinguishes itself by extending to DAS and DTU analyses. The substantial overlap with the results of previous studies in differential expression analysis results with regards to protein coding genes and TEs as well as in alternative splicing analysis results observed during this analysis strongly corroborates the validity and efficacy of the automated CRESCENT analysis pipeline.

Availability and requirements

Project name: CRESCENT.

Project home page: https://github.com/gilless429/crescent

Operating system(s): Linux-based.

Programming language: Snakemake / Python.

Other requirements: Conda.

License: GPLv3.

Any restrictions to use by non-academics: none.

Abbreviations

CRESCENT	Comprehensive RNA-Seq Expression, Splicing, and Coding/non-coding Element Network Tool
DAS	Differential Alternative Splicing
DEA	Differential Expression Analysis
DEG	Differentially Expressed Gene
DTE	Differentially Expressed Transposable Element
DTU	Differential Transcript Usage
FDR	False Discovery Rate
GO	Gene Ontology
HPC	High-Performance Computing
lncRNA	Long non-coding RNA
ncRNA	Non-coding RNA
pADJ	Adjusted pValue
snoRNA	Small nucleolar RNA
RATs	RNA-seq Analysis of Transcript Splicing
rMATS	Replicate Multivariate Analysis of Transcript Splicing
TE	Transposable Element

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06368-5.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements
We thank Yoan Renaud from the iGReD Bioinformatic platform for technical advice, Ludovic Bonhomme from INRAe, GDEC for selecting the wheat dataset, Nadia Goue from the Auvergne Bioinformatic (AuBi) platform facility and the Mesocentre Clermont-Auvergne for providing computing and storage resources facilities.

Author contributions
AVP and CT supervised the study. GS designed the CRESCENT workflow and performed the bio-informatics analysis with the help of GC and VD. SA carried out user validation of the workflow. CT designed the graphical representation of the data, and CT and GS were the major contributors to manuscript writing. All authors proofread and approved the final manuscript.

Funding
This work was supported by Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Université Clermont Auvergne (UCA). The authors acknowledge the support received from the Agence Nationale de la Recherche of the French government through the program "Investissements d'Avenir" (16IDEX0001 CAP 2025 CIR 1, International Research Center on Sustainable AgroEcosystems) and network support from the GDR EPIPLANT.

Data availability
Code and documentation are available at https://github.com/gilless429/crescent. Sequencing data were obtained from the Gene Expression Omnibus (GEO) database (project reference: GSE132415 and GSE18508) and from the European Nucleotide Archive (ENA) database (PRJEB59062A). Additional information on processed data is provided in Supplemental files

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 5 June 2025 / Accepted: 2 January 2026
Published online: 21 January 2026

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–4. https://doi.org/10.1038/533452a.
2. Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, et al. Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. Future Gener Comput Syst. 2017;75:284–98. https://doi.org/10.1016/j.future.2017.01.012.
3. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451–5. https://doi.org/10.1101/gr.4086505.
4. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinform. 2012;28:2520–2. https://doi.org/10.1093/bioinformatics/bts480.
5. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9. https://doi.org/10.1038/nbt.3820.
6. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38:276–8. https://doi.org/10.1038/s41587-020-0439-x.
7. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
8. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021. https://doi.org/10.12688/f1000research.29032.2.
9. Zhang X, Jonassen I. RASflow: an RNA-Seq analysis workflow with Snakemake. BMC Bioinformatics. 2020;21:110. https://doi.org/10.1186/s12859-020-3433-x.
10. Orjuela S, Huang R, Hembach KM, Robinson MD, Soneson C. ARMOR: an automated reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data. G3 Genes Genomes Genet. 2019;9:2089–96. https://doi.org/10.1534/g3.119.400185.

11. Cornwell M, Vangala M, Taing L, Herbert Z, Köster J, Li B, et al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinform. 2018;19:1–14. https://doi.org/10.1186/s12859-018-2139-9.
12. Wang D. hppRNA—a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples | Briefings in Bioinformatics | Oxford Academic. Brief Bioinform. 2018;19:622–6.
13. Fallon TR, Čalounová T, Mokrejš M, Weng J-K, Pluskal T. transXpress: a Snakemake pipeline for streamlined de novo transcriptome assembly and annotation. BMC Bioinform. 2023;24:1–11. https://doi.org/10.1186/s12859-023-05254-8.
14. Mpangase PT, Frost J, Tikly M, Ramsay M, Hazelhurst S. Nf-rnaSeqCount: a Nextflow pipeline for obtaining raw read counts from RNA-seq data. S Afr Comput J. 2021;33:830. https://doi.org/10.18489/sacj.v33i2.830.
15. Lataretu M, Hölzer M. RNAflow: an effective and simple RNA-Seq differential gene expression pipeline using Nextflow. Genes (Basel). 2020;11:1487. https://doi.org/10.3390/genes11121487.
16. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86. https://doi.org/10.1038/nrg.2016.139.
17. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023;24:430–47. https://doi.org/10.1038/s41580-022-00566-8.
18. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18:437–51. https://doi.org/10.1038/nrm.2017.27.
19. Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111:E5593-5601. https://doi.org/10.1073/pnas.1419161111.
20. Wang Y, Xie Z, Kutschera E, Adams JI, Kadash-Edmondson KE, Xing Y. rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. Nat Protoc. 2024;19:1083–104. https://doi.org/10.1038/s41596-023-00944-2.
21. Jiang M, Zhang S, Yin H, Zhuo Z, Meng G. A comprehensive benchmarking of differential splicing tools for RNA-seq analysis at the event level. Brief Bioinform. 2023;24:bbad121. https://doi.org/10.1093/bib/bbad121.
22. Froussios K, Mourão K, Simpson G, Barton G, Schurch N. Relative abundance of transcripts (RATs): identifying differential isoform abundance from RNA-seq. F1000Res. 2019;8:213. https://doi.org/10.12688/f1000research.17916.1.
23. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9. https://doi.org/10.1038/nmeth.4197.
24. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7. https://doi.org/10.1038/nbt.3519.
25. Laval Q, Richert L, Hejblum B. CO7.2 - a robust pipeline for RNA-seq data and a benchmark of differential expression analysis methods. J Epidemiol Popul Health. 2025;73:203009. https://doi.org/10.1016/j.jeph.2025.203009.
26. Sun L, Jing Y, Liu X, Li Q, Xue Z, Cheng Z, et al. Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. Nat Commun. 2020;11:1886. https://doi.org/10.1038/s41467-020-15809-5.
27. Pecinka A, Dinh HQ, Baubec T, Rosa M, Lettner N, Scheid OM. Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. Plant Cell. 2010;22:3118–29. https://doi.org/10.1105/tpc.110.078493.
28. Pietzenuk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, et al. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. Genome Biol. 2016. https://doi.org/10.1186/s13059-016-1072-3.
29. Li S, Liu J, Liu Z, Li X, Wu F, He Y. HEAT-INDUCED TAS1 TARGET1 Mediates Thermotolerance via HEAT STRESS TRANSCRIPTION FACTOR A1a–Directed Pathways in Arabidopsis[C][W]. Plant Cell. 2014;26:1764–80. https://doi.org/10.1105/tpc.114.124883.
30. Jin J, Lu P, Xu Y, Li Z, Yu S, Liu J, et al. PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. Nucleic Acids Res. 2021;49:D1489–95. https://doi.org/10.1093/nar/gkaa910.
31. Huang J, Wang H, Wang Y, Copenhaver GP. Comparative transcriptomic analysis of thermally stressed *Arabidopsis thaliana* meiotic recombination mutants. BMC Genomics. 2021;22:1–13. https://doi.org/10.1186/s12864-021-07497-2.
32. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, et al. Conservation of an RNA regulatory map between *Drosophila* and mammals. Genome Res. 2011;21:193–202. https://doi.org/10.1101/gr.108662.110.
33. Gibilisco L, Zhou Q, Mahajan S, Bachtrog D. Alternative splicing within and between *Drosophila* species, sexes, tissues, and developmental stages. PLoS Genet. 2016;12:e1006464. https://doi.org/10.1371/journal.pgen.1006464.
34. Cmero M, Davidson NM, Oshlack A. Fast and accurate differential transcript usage by testing equivalence class counts. F1000Res. 2019. https://doi.org/10.12688/f1000research.18276.1.
35. Rocher F, Dou S, Philippe G, Martin M-L, Label P, Langin T, et al. Integrative systems biology of wheat susceptibility to *Fusarium graminearum* uncovers a conserved gene regulatory network and identifies master regulators targeted by fungal core effectors. BMC Biol. 2024;22:53. https://doi.org/10.1186/s12915-024-01852-x.

Publisher's Note