A comprehensive pipeline for genome annotation across species: a case study on *Schmidtea mediterranea*

Anastasiia Zaremba, Małgorzata Marszałek-Zeńczak, Annasha Dutta, Anna Samelak-Czajka, Paulina Jackowiak#

Laboratory of Single Cell Analyses; Institute of Bioorganic Chemistry Polish Academy of Sciences; Zygmunta Noskowskiego str. 12/14, 61-704 Poznań; Poland

azaremba@ibch.poznan.pl, marszalekgosia@gmail.com, adutta@ibch.poznan.pl, asczajka@ibch.poznan.pl,

#Corresponding author: paulinaj@ibch.poznan.pl

Abstract

Background

Despite advancements in genome annotation tools, challenges persist for non-classical model organisms with limited genomic resources, such as *Schmidtea mediterranea*.

Findings

To address these challenges, we developed a flexible and scalable genome annotation pipeline that integrates short-read (Illumina) and long-read (PacBio) sequencing technologies. The pipeline combines reference-based and *de novo* assembly methods, effectively handling genomic variability and alternative splicing events. Functional annotation is conducted to filter out low-

confidence transcripts and ensure biological relevance. Applying this pipeline to the asexual strain of *S. mediterranea* revealed thousands of previously undescribed putative genes and transcripts, and improved the existing gene models, highlighting its utility in annotating complex, underexplored genomes.

Conclusions

The modularity and comprehensiveness of our pipeline ensure its adaptability for genome annotation across diverse species, making it a valuable tool for annotating genomes of non-model organisms and supporting broader genomic research. The source code and implementation details are available at https://github.com/Norreanea/SmedAnno.

Keywords

Genome annotation; Transcriptome assembly; *Schmidtea mediterranea*; Non-model organisms; Functional annotation

Background

Understanding the genetic makeup of organisms is crucial for unraveling the complexities of their biology, making genome annotation a foundational aspect of molecular and genomic research. By precisely identifying coding regions, non-coding elements, and regulatory sequences, genome annotation transforms raw genomic data into a practical blueprint that opens up possibilities for studying living organisms at the molecular level.

Existing genome annotation methods can be broadly categorized into *ab initio* [1], homology-based [2], evidence-based (reference-based) [3], and hybrid approaches [4–6], each leveraging different strategies and tools to predict and validate genomic features. Each category has its unique advantages and challenges. *Ab initio* methods, which predict genes solely from intrinsic sequence signals without external evidence, are capable of identifying novel genes; however, they tend to produce a higher rate of false positives and inaccurate gene models—such as overpredicted exons, missed genes, and incorrect exon–intron boundaries—especially when applied to complex or poorly characterized genomes. Homology-based methods use evolutionary conservation to achieve high accuracy for well-conserved genes but often miss novel or rapidly evolving genes that lack close homologs. Evidence-based approaches utilize transcriptomic or proteomic data to support gene models, thereby enhancing annotation quality for expressed genes, yet they are inherently limited to regions where experimental data exist and are as robust as the underlying evidence. Hybrid approaches combine intrinsic signals with external evidence in an effort to balance sensitivity and specificity.

To streamline the annotation process, several genome annotation platforms and pipelines have been developed. For instance, the RefSeq eukaryotic annotation pipeline at NCBI [7] employs highly curated alignment-based evidence and gene prediction models, while another recently reported pipeline [8] aims to integrate evidence from multiple data sources for eukaryotic genome annotation. Web-based platforms such as Galaxy Genome Annotation (GGA) [9] provide an accessible interface for constructing annotation workflows, and tools like Prokka [10] offer rapid prokaryotic annotation that automates the integration of gene prediction and functional annotation, generating standardized output formats compatible with GenBank. Additionally, PAGIT (Post

Assembly Genome Improvement Toolkit) [11], the community-driven Nextflow-based approach nf-core GenomeAnnotator [12], and other pipelines [13–15] expand the available toolkit.

Despite these advancements, applying genome annotation methods to non-model or non-classical model organisms presents significant challenges. Due to limited reference data, tools must rely more heavily on ab initio methods, which—as noted above— are prone to errors, particularly for complex genomes, and often lack sufficient training data. Repetitive sequences and the presence of multiple isoforms complicate the annotation process. Genetic diversity and evolutionary distance from well-studied species can impede the identification of orthologous genes and conserved pathways, while limited experimental data often constrain the ability to validate or refine gene models. Such constraints are further compounded by annotation artifacts—including fragmented and chimeric transcripts—that can obscure true biological functions and introduce errors into downstream analyses.

These challenges are particularly relevant to research on the sexual (S2F2 clonal line) and asexual (CIW4) strains of *Schmidtea mediterranea*, a highly regenerative flatworm model organism [16,17]. Currently, reference genome sequences and comprehensive annotations are available primarily for the sexual strain, including contig-level genome assemblies such as SmedGD_c1.3 [18,19], dd_Smes_g4 [20,21], chromosome-scale genome assemblies schMedS2 [22] and recent schMedS3 [23], as well as corresponding annotations like SmedGD_c1.3, high-confidence SMESG, schMedS1, schMedS2, schMedS3_1 and schMedS3_2. Chromosome-scale assemblies of the sexual strain indicate that sexual and asexual *Schmidtea mediterranea* share high genomic similarity—they are both diploid with four pairs of chromosomes. Although the asexual strain possesses a chromosomal translocation between chromosomes 1 and 3, these reference sequences and annotations offer a strong starting point for transcriptome analysis. Nonetheless,

given potential strain-specific transcripts and structural variations, further refinement of the asexual strain's transcript set is essential [24]. Additionally, the asexual strain exhibits a simplified reproductive strategy [25–28] and it has been widely used in numerous studies, with extensive data deposited in public repositories. As of January 2, 2025, there are 122 DNA and 1,694 RNA datasets on NCBI [29] for this strain, the majority of which are short-read data. In comparison, for the sexual strain, there are only 18 DNA and 87 RNA datasets. The predominance of RNA-seq data highlights the extensive transcriptomic resources available for the asexual strain, which offers a strong foundation for gene expression and functional studies and underscores the potential for leveraging RNA sequencing to enhance genome annotation efforts. However, limited detailed reference resources for asexual strain remain a significant challenge. Currently, the scientific community must rely mostly on reference transcriptomic data—such as dd_Smed_v6.pcf.contigs.fasta [21] (a transcriptome assembly generated from RNA-Seq data) and a contig-based genome assembly and annotation SmedAsxl_genome_v1.1 [19]. Beyond that, existing transcriptome assemblies are fragmented and incomplete, lacking the high-resolution insights offered by newer sequencing technologies. To bridge this gap, our study aims to implement a sophisticated bioinformatics pipeline to improve annotated genes and transcripts in asexual *Schmidtea mediterranea* using the existing chromosome-scale genome of the sexual strain (schMedS2) as a reference and in-house generated comprehensive collection of transcriptomic data for asexual worms. Our dataset includes both short-read and long-read data, providing a significant advantage in resolving complex isoforms. Addressing these challenges requires the integration of various sequencing technologies and advanced bioinformatics tools. Short-read sequencing platforms, such as Illumina, offer high accuracy and depth of coverage, a low error rate, which is essential for identifying single-nucleotide polymorphisms, small indels, and importantly, splice

sites—ensuring accurate distinction of exon-intron boundaries [30]. Conversely, long-read technologies like PacBio provide valuable information on structural variants and complex transcript isoforms, enhancing the resolution of genome assemblies [31,32]. Combining these technologies benefits from their complementary strengths, enabling a more comprehensive and accurate genome annotation [4].

Here we present a universal, comprehensive, and adaptable pipeline for genome annotation to process and interpret the vast amounts of sequencing data generated, that integrates both short- and long-read sequencing data, employs state-of-the-art bioinformatics tools for transcriptome assembly and genome characterization, and incorporates robust methods for functional and homology annotation. By doing so, we aim to advance genome annotation efforts within *Schmidtea mediterranea* and to provide a scalable framework applicable to other non-model organisms, fostering broader genomic research and discovery.

Methods

Planarian culturing and sample collection

Asexual *Schmidtea mediterranea* were cultured as described in [33]. Animals were collected, immediately flash-frozen in liquid nitrogen to preserve RNA integrity, and stored at −80 °C until RNA isolation.

Dataset details

Full dataset-specific details including experimental treatments, RNA extraction, quality control, are provided in Table S1 and in the associated GEO submissions.

Validation with simulated data

A synthetic genome was created using a custom Python script, which constructed a single chromosome spanning 2 mega base pairs (bp). This genome included 1,000 genes randomly distributed along the chromosome, each comprising between 2 and 20 exons with lengths ranging from 100 to 1,000 nucleotides (nt) and introns of 100 to 1,000 nt. Paired-end short-read RNA-Seq data were simulated using ART Illumina tool, generating 100 bp reads at 10× coverage across six samples. Additionally, long-read sequencing data were produced using PBSIM tool with 150 bp reads at 10× coverage for one sample. To evaluate the pipeline's ability to handle incomplete annotations, a random subset of genes was removed from the annotation file using the custom bash script, resulting in a GTF file with missing gene information.

Reference annotation polishing

The reference genome SchMedS2 [21] was utilized for this study. High-confidence annotations for this genome were obtained from two sources: SmedSxl smed_chr_ref_v1 UCLA (SIMRBase (https://simrbase.stowers.org)) and schMedS2_PlanMine_high-confidence_liftover (PlanMine [21]). Both annotations are compatible with the genome sequence; however, due to minor differences between them, the genes were first unified, and the annotation was carefully reviewed and corrected. Through detailed analysis and comparison, 27,676 SMEST transcripts and 21,332 SMESG genes were identified as the most confident and consistent across both sources. Next, 169 transcripts with identical exon structures using the AGAT tool [34] were removed to eliminate redundancy and enhance annotation quality.

Assembly evaluation and benchmarking

The transcriptome assembly was evaluated using the SIRV-set 4 RNA spike-in controls to assess its accuracy, with specific StringTie2 parameters (-c, the minimum reads per bp coverage to consider for multi-exon transcript, and -f, the minimum isoform fraction) being tested. Completeness was assessed with BUSCO [35] (v5.7.1) with a Metazoa benchmark of 954 conserved genes, while assembly quality metrics were obtained using rnaQUAST [36] (v2.3.0) and Transrate [37] (v1.0.3). Benchmarking procedures were conducted by comparing the performance of the pipeline against existing annotations, with metrics such as the number of correctly identified transcripts, assembly scores, and completeness indicators being analyzed.

ncRNA identification

All ncRNA sequences were downloaded from RNAcentral [38] and supplemented with 457 tRNA sequences from an external publication. The sequences were linearized and deduplicated using seqkit rmdup. The initial dataset contained 27,390,440 sequences, which was reduced to 27,317,511 after deduplication. A BLASTn search was conducted to identify homologous sequences within the *Schmidtea mediterranea* nuclear and mitochondrial genomes. The following parameters were used: -query: 27,317,511 ncRNA sequences; -db: *Schmidtea mediterranea* genome (nuclear + mtDNA), -evalue: 1e-20, -word_size: 15, -qcov_hsp_perc: 90, -dust: no. The output was then processed to extract the original sequence length of hits for further filtering. A custom Python script was used to classify ncRNA types based on query ID patterns. Hits were retained if alignment coverage met predefined thresholds (95% for tRNA, 90% for miRNA, 70% for lncRNA, 90% for rRNA, 80% for other ncRNA). Then a custom R script was used for converting to gtf format, refining and merging overlapping regions to ensure only non-redundant ncRNA annotations were retained.

Analyses

Selection and validation of transcriptome assembly approach

One of the key steps in genome annotation is the assembly of the transcriptome. In our efforts to enhance genome characterization for the asexual strain of *Schmidtea mediterranea*, we employed StringTie2 [4], a widely recognized tool for transcriptome assembly. StringTie2 is versatile, supporting both short-read and long-read sequencing datasets, which makes it well-suited for comprehensive transcriptome analysis. However, StringTie2 has multiple versions, each offering distinct features and performance enhancements. Therefore, before integrating StringTie2 into our pipeline, we thoroughly optimized its parameters to ensure the highest accuracy and reliability in transcript predictions. To this end, we employed two complementary strategies.

In the first strategy, we generated an *in-silico* dataset to simulate genomic sequences with corresponding full and incomplete annotations and short-read RNA-Seq data (see Methods section). This synthetic data allowed us to create scenarios with known simple transcript structures, enabling a comprehensive evaluation of StringTie2 performance. Reads were aligned to the *in-silico* reference genome using STAR. We then compared two commonly available versions of StringTie2, 2.1.1 and 2.2.1, which allowed us to assess version-specific performance improvements. Our results indicated that StringTie2 version 2.1.1 provides a higher identification rate of short-read data in reference-based mode than its predecessor (Figure 1A). Although the StringTie2 documentation does not explicitly state that version 2.1.1 is more sensitive for short-read data, the combination of release note details and our experimental observation suggests that less stringent filtering of version 2.1.1 results in a higher count of assembled isoforms. Meanwhile,

version 2.2.1 has been optimized for hybrid-read assembly (with the --mix option) and incorporates more conservative measures to reduce false positives—which can lower the total transcript count when only short-read data are used. This trade-off between sensitivity and precision is common in transcript assemblers. When maximizing the number of detected isoforms from short reads is a priority, 2.1.1 may be advantageous, whereas 2.2.1's improvements (including its ability to process mixed-read data) make it preferable for datasets that include long reads. It should be noted that these version choices are based on their prevalence at the time of our study, and although our benchmarking indicates reliable performance, further testing in diverse datasets may be beneficial.

In the second strategy, we utilized Spike-In RNA Variants (SIRVs), a commercially available collection of RNA spike-in controls, that we added to the RNA isolated from planarians (see Table S1) prior to the preparation of sequencing libraries. These SIRVs include three distinct modules — ERCC Mix 1 (a set of 92 synthetic RNA molecules ranging between 250 and 2,000 nt in length to resemble natural eukaryotic transcripts), Isoform Mix E0 (referred here as multi-exon SIRVs), and long SIRVs (referred here as mono-exon SIRVs) — each designed to mimic specific aspects of transcriptome complexity, including transcript abundance, isoform diversity, and transcript length. RNA-Seq reads were aligned to the reference SIRVs sequences using STAR for short reads and Minimap2 for long reads. We then employed StringTie2 using the "insufficient annotation" as the reference provided by Lexogen (SIRV_I), where some SIRVs that are actually present in the mixes are not annotated, and compared the prediction to the correct annotation of all SIRVs (SIRV_C). Such a comparative analysis demonstrated how the pipeline could substantially improve genome annotations through a proper differentiation between known and novel transcript isoforms. Then, focusing on the mixture of synthetic RNA isoforms Iso Mix E0 category, we systematically assessed the performances of seven assembly approaches for transcript variant

detection, implementing different combinations of read types—short and mixed—and assembly modes—reference-based and *de novo*. These approaches included short reads in reference-based mode (SR_RB), short reads in *de novo* mode (SR_DN), mixed reads in reference-based mode (MR_RB), mixed reads in *de novo* mode (MR_DN), merged reads in *de novo* mode (MeR_DN), merged reads in reference-based mode (MeR_RB), and merged reads in both modes (MeR_RB_DN). For each approach, StringTie2 performance (Table S2) was quantified in terms of accuracy metrics, reflecting the ability to correctly identify transcript isoforms (Figure 1B). SR_DN approach achieved the lowest accuracy, ranging from 35.1% to 45.1% with its maximum at -c 1.5 and -f 0.02. Surprisingly, generally, the highest accuracy rate of 81.1% was obtained when predictions from both short and mixed reads in both modes were merged (MeR_RB_DN). This outcome underscores the synergistic benefits of integrating multiple read types and assembly strategies.

Consequently, our pipeline employs a dual-version strategy—by default, StringTie2 version 2.1.1 is used for assembling short-read data and version 2.2.1 for mixed-read data, ensuring optimal performance across a range of sequencing data types. Users should note that while these versions provided robust performance in our tests, alternative versions may be considered depending on specific experimental contexts and can be adjusted via command-line. We set -c 1.5 and -f 0.02 based on pilot experiments with SIRVs, which balanced sensitivity and precision in isoform detection (see Table S2). After completing individual assemblies, we incorporated the MeR_RB_DN approach to merge reference-based and de novo transcripts, resulting in a comprehensive and unified transcriptome annotation that captures a diverse array of transcript isoforms.

Identification of artifactual transcripts

In our tests with *in-silico* data, we identified several types of artifactual transcripts, including those with unusually long exons exceeding 100,000 nt, reversed duplicates (transcripts overlapping the same genomic regions as the reference ones but on the opposite strand), as well as fragmented and chimeric RNA. Fragmented transcripts are individual RNAs that have been split into multiple predicted gene models due to incomplete transcript reconstruction or assembly errors. Chimeric transcripts, on the other hand, erroneously combine sequences from multiple distinct biological genes. In eukaryotic genomes, exons generally range in size from approximately 50 to 5,000 nt, with the majority falling between 100 and 1,000 nt [39]. Instances of exons exceeding 10,000 nt are exceptionally rare and are usually associated with specific organismal features, such as immune-related genes or tandem duplications. Even in these rare cases, such lengths remain an order of magnitude below 100,000 nt [40]. While highly effective in many contexts, StringTie2 may incorrectly model transcripts, especially in regions with overlapping genes or incomplete stop codons. Such inaccuracies can extend exon boundaries far beyond biologically realistic lengths [41].

To address the identified challenges and eliminate artifacts, we added a custom function to the pipeline that ensured that transcripts with exons exceeding 10,000 nt and/or transcripts that surpassed 100,000 nt of the genomic region were removed, thus refining the annotation and preventing the propagation of erroneous data. Additionally, we included a function to identify and filter out reversed duplicates without inadvertently removing legitimate transcripts. We also implemented criteria to flag potential fragmented and chimeric RNA based on exon length, genomic proximity, and functional annotations.

In our pipeline, indicators for fragmentation include multiple predicted genes in close genomic proximity (within 1,000 nt) on the same strand and adjacent or overlapping genomic coordinates. The exon length and genomic proximity thresholds are based on general eukaryotic genome characteristics [40]. These parameters should be adjusted as needed based on the specific organism or dataset characteristics. Genes and transcripts are flagged as chimeric if they fulfill at least one of the following criteria: genes have multiple, unrelated functional annotations; display structural inconsistencies such as exons that are widely spaced or non-contiguous; span multiple distinct genomic regions. As part of our automated pipeline, we only added information indicating whether a gene or transcript is possibly fragmented or chimeric. The decision to remove corresponding genes or transcripts should be made by the user based on specialized evaluation. Incorporating functional annotations into the pipeline as a part of manual curation can significantly improve the accuracy of detecting artifacts.

Pipeline overview and workflow

Having resolved the challenges outlined above, we developed a versatile genome annotation pipeline and demonstrated its performance by applying it to the asexual strain of *Schmidtea mediterranea*. This pipeline integrates multiple sequencing technologies and bioinformatics tools to achieve comprehensive and accurate gene and transcript annotations across different organisms. The pipeline is divided into two main parts, each structured into distinct sequential steps.

Part I: Pre-processing and assembly

The first part of the pipeline is fully automated and encompasses pre-processing, alignment, assembly, and functional annotation. The input data for the pipeline can be either short-

read or mixed, the latter including both short and long reads. While the pipeline can process a single sample when only short reads are available, incorporating multiple samples is recommended to achieve more reliable and comprehensive results. Mixed-read datasets require at least one sample with short reads and one with long reads. Additionally, the automated pipeline is optimized for paired-end reads in short-read mode. Mapping and transcript assembly are performed independently for each sample. The pipeline begins by pre-processing the RNA-Seq data, focusing in particular on the removal of reads corresponding to ribosomal RNA (rRNA), based on their alignment to user-provided rRNA reference sequences (Figure 2). This step was found to be very important in the case of organisms with unusually high A/T content, such as *Schmidtea mediterranea*. Library preparation methods typically involve rRNA depletion or polyA enrichment to facilitate the study of protein-coding gene expression. However, commercially available rRNA depletion kits are mostly optimized for model species, and polyA enrichment is not effective in organisms with a high A/T content. Consequently, reads corresponding to rRNA can be abundant in sequencing data for non-model organisms. In scenarios where rRNA reference sequences are unavailable (not provided by the user), the pipeline bypasses the rRNA removal step, continuing with the whole RNA-Seq dataset. The pipeline then aligns the processed RNA-Seq reads to the reference genome using STAR for short reads and Minimap2 for long reads. The transcript assembly phase utilizes StringTie2: by default, version 2.1.1 is applied to short-read data, while version 2.2.1 is used for mixed-read data (which integrates both short and long reads). Users can specify alternative versions if desired. After assembly, the pipeline merges transcripts across all samples using StringTie2 to create unified sets of transcripts. This merging process is performed separately for reference-based (if reference annotation was provided) and *de novo* assemblies before combining them into the final complete annotation. This dual approach by performing both

reference-based and *de novo* annotation, provides comprehensive assembly, thus capturing a wide range of transcript isoforms. Subsequently, transcripts with excessively long exons or that cover unusually large genomic regions are filtered out. Next, we utilized gffcompare to compare the predicted annotations against the reference annotations, assessing the completeness and accuracy of the assembled transcripts. This comparison generates a report that includes data summaries and accuracy estimations. Additionally, it enriches the predicted annotations with detailed information about their overlaps with known reference transcripts and genes. However, the generated annotations may contain issues that can compromise the quality and usability of the annotation files. To address these issues, we employed AGAT (Another GTF/GFF Tool), which systematically identifies and corrects common syntax errors, ensuring that essential attributes such as gene_id and transcript_id are present for each feature, preventing incomplete annotations, correcting improper use of delimiters in attribute fields (such as replacing spaces with a semicolon), detecting and removing duplicate entries, and ensuring correct key-value pair syntax. This correction phase is crucial for maintaining high-quality genome annotation. The obtained set of transcripts is then subjected to functional annotation in a process that involves the sequential application of several tools, and which includes homology analysis and the identification of protein domains. To this end, first, the ORFs are predicted using TransDecoder. Following that, BLASTp and BLASTx are used for homology search through the SwissProt database. Finally, protein domains are identified using hmmscan through the PFAM database. This multi-faceted approach filters out low-confidence annotations, retaining only the transcripts supported by multiple lines of evidence. We utilized the SwissProt and PFAM databases as a part of the automated pipeline specifically for their low memory requirements. However, we recommend more sophisticated resources, for example eggNOG, NCBI, UniRef, KEGG, InterPro, RefSeq, COG, Reactome, and

BioCyc for conducting more comprehensive and in-depth homology analyses, as these databases provide broader coverage of protein functions, evolutionary relationships, and diverse protein characteristics. Next, the functional annotations are integrated into the GTF file. This integration applies specific thresholds: for PFAM predictions, a bit score greater than 20, a conditional E-value below 1e-5, and coverage exceeding 50%; for SwissProt predictions, a BLAST e-value threshold of 1e-5 and an identity threshold of 25%. These criteria ensure that low-confidence functional annotations are discarded. To this end, the pipeline adds new columns to the GTF annotation file, including has_ORF (TRUE/FALSE), best_SwissProt_blastp_hit, best_SwissProt_blastx_hit, and best_Pfam_hit domain information. Subsequently, users have the option to remove redundant transcripts and eliminate those lacking functional information, ensuring that each retained transcript is supported by at least one entry. The next step is the identification of overlapping genes and transcripts, which detects genes that share genomic regions and checks if each transcript in a gene has at least one overlapping exonic region with any other transcript in the same gene. Lastly, the pipeline identifies the aforementioned reversed duplicates and addresses fragmented and chimeric genes both with and without reference annotations. All these steps (Figure 2) are implemented in the SmedAnno tool, forming Part I of the entire pipeline. The source code can be accessed on GitHub (https://github.com/Norreanea/SmedAnno).

Part II: Manual curation and refinement

The second part of the pipeline involves user-driven inspection and refinement of the annotated data. Despite the implementation of several automated steps aimed at optimizing transcript assembly and obtaining biologically relevant transcripts that encode functional protein domains, resolving overlapping or fragmented transcripts remains a significant challenge that requires specialized approaches. While the primary focus is on transcript assembly, gene features

are subsequently derived from the assembled transcripts. Fragmented and chimeric transcripts, which primarily affect RNA-level annotations, are addressed by prioritizing the longest isoform to define each gene, thereby reducing redundancy and improving annotation accuracy. To accurately determine whether genes should be merged after being marked as potentially fragmented and/or overlapped, promoter and terminator sequences should be analyzed using Promoter 2.0 and ARNold tools (Figure 3). This analysis verifies the coherence of regulatory regions, ensuring that merged gene models maintain consistent promoter and terminator elements. These tools are accessible via their respective web pages and are not included in the automated pipeline. Users must extract sequences between inspected genes to check for the presence of promoters and terminators. After potential merges, manual adjustments are advised to finalize gene models, ensuring accurate gene length prediction and overall annotation reliability.

While genome annotation typically focuses on protein-coding genes, our pipeline also includes a step for annotating non-coding RNAs (ncRNAs), which have significant functional relevance. This step is optional, as it requires downloading the extensive (>20 GB) RNAcentral database. The ncRNA sequences are first linearized and deduplicated using seqkit rmdup to remove redundant entries. Homologous regions are then identified in the reference genome using BLASTn. After the identification step, ncRNA types are classified based on the sequence identifier. The matching results are then converted to GTF format. A multi-step filtering and refinement process is used to improve the accuracy of the annotations. This process addresses several common issues, including multiple alignments at identical genomic positions, inconsistent naming conventions, and clustering.

The final step integrates all gathered information into a unified genome annotation file, ready for downstream applications, including both coding and non-coding elements with

comprehensive functional insights. One additional step as part of the genome annotation extension involves incorporating Gene Ontology (GO) annotations based on the functional annotations obtained in Part I. However, this requires downloading the large goa_uniprot_all.gaf file (>200 GB after gunzip). While we provide the necessary code for this integration, it is not included as part of the automated Part I pipeline.

The bioinformatics tools utilized in each pipeline step, along with their respective versions are summarized in Table 1.

Table 1. Bioinformatics tools and pipeline steps

Step	**Tool**	**Version**
Step 1: Remove rRNA	STAR Minimap2 samtools	2.7.9a
Step 2: Align reads to the reference genome	STAR Minimap2	2.7.9a
Step 3: Assemble genes & transcripts	StringTie2	2.1.1 (short reads) 2.2.1 (mixed reads)
Step 4: Merge reference-based assemblies	StringTie2	2.2.1
Step 5: Merge *de novo* assemblies & create pre-final annotation	StringTie2	2.2.1
Step 6: Filter out transcripts artifacts	bash	
Step 7: Compare isoforms	gffcompare	0.12.6
Step 8: Correct & enhance annotation	AGAT	1.4.1
Step 9: Determine functional attributes	TransDecoder blastp blastx hmmscan	5.7.1 2.15.0+ 2.15.0+ 3.3.2
Step 10: Integrate functional attributes	R	4.4.1

Step 11: Identify overlapping genes & transcripts	R	4.4.1
Step 12: Identify reversed duplicates	R	4.4.1
Step 13: Identify fragmented & chimeric genes	R	4.4.1
Step 14: Analyse promoter & terminator sequences	Promoter 2.0 ARNold	No official version number No official version number
Step 15: Annotate non-coding RNA (ncRNA)	blastn seqkit	2.15.1+ 2.8.2
Step 16: Integrate collected information into a unified genome annotation file	R	4.4.1
Step 17: Incorporate gene ontology (GO)	R	4.4.1
Step 18: Perform final review and quality assurance	R	4.4.1

Identifying fragmented and chimeric artifacts

In the context of genomic annotation, artifacts refer to erroneous gene or transcript models that arise from limitations in the annotation process rather than representing true biological entities. Specifically, we refer here to fragmented artifacts as those that occur when a single gene is incorrectly split into multiple, separate gene models or transcripts. This fragmentation can result from incomplete transcript assembly or sequencing errors, leading to discontinuous or truncated representations of what is biologically a single gene. On the other hand, chimerism typically arises from misassembly during transcriptome reconstruction, especially in regions with high gene density or overlapping gene structures.

Our analysis employs flags such as: *is_potentially_fragmented_de_novo*, *is_potentially_chimeric_due_to_structure_de_novo*, and *is_potentially_chimeric_due_to_function_de_novo* to identify these artifacts based on functional

and structural criteria. Below, we discuss three illustrative examples (depicted in Figure 4) that highlight the complexities and implications of these annotation challenges. SMESGs here are reference genes and MSTRGs are putative genes.

Fragmented gene prediction in proximal gene regions

In one genomic region, three genes—SMESG000004200.1, SMESG000004201.1, and MSTRG.967—are located in close proximity on chromosome 1 (smed_chr1) (Figure 4A). Specifically, gene SMESG000004200.1 spans position 14,973,263 to 15,003,452 on the positive strand, while SMESG000004201.1 extends from 15,003,453 to 15,011,292 on the same strand. StringTie2, predicted an additional transcript, MSTRG.967, which encompasses the regions of both genes (14,973,256 to 15,011,292). Both SMESG000004200.1 and SMESG000004201.1 were independently annotated as "Epidermal growth factor receptors." However, the prediction by StringTie2 that merges these two genes into a single transcript is more biologically plausible. This is because the close genomic proximity and shared functional annotations suggest a possible gene fusion or a complex gene structure that is better represented by a single, contiguous transcript. Flagging these genes as potentially fragmented (is_potentially_fragmented_de_novo) highlights the necessity to consider merged annotations in densely packed genomic regions to avoid misinterpretation of gene boundaries.

Chimeric transcript prediction due to excessive intron length

The gene SMESG000065897.1 initially comprised a single transcript, SMEST065897001.1, spanning positions 325,241,809 to 325,242,781 on chromosome 1 (Figure 4B). However, StringTie2 predicted an additional transcript, MSTRG.16892.1, which extends dramatically from 325,241,809 to 325,347,749, introducing an intron exceeding 100,000 nt. This

unusually large intron length is atypical and suggests a misassembly, classifying MSTRG.16892.1 as potentially chimeric (is_potentially_chimeric_structure_de_novo).

The presence of such a large intron likely indicates that MSTRG.16892.1 erroneously combines sequences from multiple genomic loci or fails to accurately represent the true transcript structure of SMESG000065897.1. Consequently, the chimeric transcript should be disregarded, and the gene localization should remain based solely on the original transcript SMEST065897001.1.

Functionally chimeric gene due to divergent transcript annotations

The gene SMESG000077853.1 exhibits a complex transcriptional landscape with multiple transcripts (Figure 4C). While most transcripts, such as MSTRG.31576.1, MSTRG.31576.5, MSTRG.31576.6, and all reference SMEST transcripts are annotated as "ankyrin repeat proteins," one transcript, MSTRG.31576.11, is annotated as "Squamous cell carcinoma antigen recognized by T-cells 3 (SART-3)". This divergence in functional annotation across transcripts results in the gene being flagged as chimeric (is_potentially_chimeric_function_de_novo). These inconsistent functional annotations suggest that MSTRG.31576.11 may have been incorrectly assembled, possibly incorporating sequences from a different gene or representing a mispredicted splice variant. Given that "SART-3" functions are unrelated to "ankyrin repeat proteins," it is evident that MSTRG.31576.11 does not belong to SMESG000077853.1. Consequently, MSTRG.31576.11 was filtered out.

Detecting missing genes through functional annotation and homology assignments

Functional annotations and homology assignments were then carried out to explain the biological function of the assembled transcripts and identify orthologous genes by using

established databases and computational tools. This approach played a key role in detecting missing genes thus highlighting potential gaps and guiding further refinement of the transcriptome assembly.

To incorporate functional data into our newly developed *Schmidtea mediterranea* genome annotation (schMed_Pn), we utilized the following databases: Pfam, SwissProt, and non-redundant (nr) protein sequences from NCBI. Our predicted annotation comprises 25,265 genes, of which 4,011 are novel and 21,254 are identical to the reference genes. Each gene possesses at least one function-related entry, either through a database hit or by being classified as having an open reading frame (ORF). For 18,552 genes, we successfully identified Gene Ontology (GO) terms using homology-based strategies, which included the pfam2go package, UniProt GOA conversion, and a custom R script to retrieve GO annotations based on NCBI IDs. Notably, 2,427 of these GO-annotated genes are putative. To evaluate the similarity of GO terms between novel and reference genes, Venn diagrams were constructed, highlighting shared and unique functional categories across the gene sets. We assigned a total of 23,909 GO terms to genes within our predicted annotation. Specifically, 18,586 GO terms were assigned to SMESG (reference) genes, and 5,323 GO terms were assigned to MSTRG (putative) genes. Of these, 5,180 GO terms were shared between both gene categories, while 143 GO terms were uniquely assigned to MSTRG genes (Figure 5A). To visualize the functional distribution of the gene sets, we generated word clouds representing the most frequent unique GO terms for each group: MSTRG (Figure 5D) and SMESG (Figure 5B), as well as those that were common for two groups (Figure 5C).

Insights into planarian ncRNA diversity

Beyond refining the annotations of protein-coding genes, we also performed a preliminary assessment of non-coding RNAs (ncRNAs) in *Schmidtea mediterranea*. Although not an

exhaustive characterization, our analysis reveals a diverse repertoire of regulatory RNAs, highlighting significant differences between publicly available annotations and newly generated data. Publicly available datasets from RNAcentral assigned a total of 154 rRNAs, 578 miRNAs (including 269 precursors), 61 tRNAs, 2 snoRNAs, 12 snRNAs, and 2 SRP RNA hits to *Schmidtea mediterranea*. After deduplication, these numbers were refined to 129 rRNA, 571 miRNAs (with 156 precursors), 60 tRNAs, 2 snoRNAs, 12 snRNAs, and 2 SRP RNA hits. Additionally, according to recent literature [42], *Schmidtea* genome contains 457 tRNA hits, which we reduced to 434 after deduplication. In contrast, our study, utilizing our annotation pipeline, identified 166 rRNA hits across 7,335 genomic positions, 182 miRNA hits across 1,390 genomic positions, 487 tRNA hits distributed over 4,408 genomic positions, 2 snoRNA hits across 69 genomic positions, 18 snRNA hits across 407 genomic positions, 8 miscellaneous RNAs spread across 5,585 genomic positions, RNase P ncRNA components across 679 genomic positions, 28 lncRNA hits across 326 genomic positions and 2 sRNA hits across 617 genomic positions. The discrepancy between published records assigned to *Schmidtea mediterranea* and our annotated RNAs may be due to the fact that some sequences listed in RNAcentral may be partial or might simply be artifacts, inaccurate submissions, or too divergent to map unambiguously. Thus, not mapping to a reference genome does not necessarily imply these entries are universally spurious, but it does highlight the ongoing need for careful re-annotation and experimental validation in underexplored organisms like planarians.

In addition, we uncovered 40 copies of signal recognition particle RNAs (SRP_RNAs) situated on chromosome 3 within an interval flanked by the genes SMESG000020110.1 (BLAST-predicted as a lysine-specific demethylase) and MSTRG.33721 (BLAST-predicted as containing a thrombospondin type 3 repeat). Intriguingly, the two protein-coding genes are oriented in the

5′→3′ direction, whereas the SRP_RNAs are oriented 3′→5′. While this arrangement might suggest a localized regulatory cluster, SRP RNAs known to be evolutionarily conserved and often found in multiple genomic locations [43]. Their presence in compact regions does not necessarily imply a functional relationship with adjacent genes. Without further evidence, it is difficult to draw functional conclusions from their location alone. Experiments measuring SRP_RNA expression levels, interactions with neighboring genes, or chromatin accessibility in these regions would be needed to confirm whether these SRP_RNAs play a role in local regulation or are simply part of a repeated structural element.

Moreover, we identified 10 copies of TERC_RNA located on smed_chr4, which align with the *Mus musculus* (house mouse) partial telomerase RNA component (TERC). The presence of TERC_RNA homologous sequences suggests that *Schmidtea mediterranea* may have a functional TERC. However, *TERT* gene (SMESG000061912.1) is located on smed_chr3, indicating that *TERT* and *TERC* are on separate chromosomes. Yet, further analysis is needed to confirm if telomerase enzyme complex (TERT + TERC) in planarians interact in a functional context.

Benchmarking and genome annotation performance

To ensure the reliability of schMed_Pn genome annotation, we performed thorough performance benchmarking against existing genome annotations of *Schmidtea mediterranea*.

To assess the completeness and quality of the existing and generated in this study genome annotations, we analyzed transcript length distribution. It revealed that all annotations predominantly consist of shorter transcripts, as reflected by a sharp decline in cumulative transcript counts with increasing transcript length (Figure 6A). Specifically, dd_Smed_v6 showed the steepest decline in cumulative transcript counts with increasing length, suggesting a higher

fragmentation rate; however, caution is advised in interpreting these distributions, as there is no universally accepted “standard” shape for planarian transcript length. In contrast, schMed_Pn demonstrated a more gradual decline, which indicates a broader distribution of transcript lengths. Assemblies such as schMedS1, schMedS2, schMedS3_1, and schMedS3_2 exhibited moderate slopes with slight differences, indicating varying transcript length distributions among them. The SmedAsxl genome_v1.1 displayed a narrower range of longer transcripts. In contrast, SmedGD_c1 and SMESG.1, distributions similar to schMed_Pn, demonstrated relatively smooth distributions, signifying balanced transcript coverage with a significant proportion of longer transcripts.

The completeness of the transcriptome assembly was evaluated using BUSCO, which assesses the presence of conserved orthologs, by comparing a transcriptome of interest to those within the Metazoa lineage (Figure 6B). dd_Smed_v6 exhibited the highest proportion of complete (single-copy) orthologs, indicating a robust assembly, but also showed moderate levels of duplicated and missing orthologs, suggesting some redundancy and gaps. schMed_Pn maintained moderate proportions of both complete single-copy and complete duplicated orthologs while demonstrating low levels of fragmented orthologs, with missing orthologs comparable to dd_Smed_v6. Assemblies such as schMedS1, schMedS2, schMedS3_1, and schMedS3_2 generally displayed lower proportions of complete orthologs and higher duplication rates, particularly in schMedS3_1 and schMedS3_2, indicating possible redundancy. Furthermore, SMESG.1 achieved high complete orthologs and the lowest fragmentation rates, whereas SmedAsxl genome_v1.1 had fewer complete orthologs and higher proportions of fragmented and missing orthologs, highlighting areas that may require further refinement. From a BUSCO standpoint, SMESG.1 and dd_Smed_v6 both perform very well (albeit with different trade-offs in duplication and missing orthologs). schMed_Pn also ranks among the stronger annotations,

demonstrating moderate duplication and missing ortholog levels while retaining high completeness.

In addition, we employed rnaQUAST and Transrate to generate detailed assembly quality metrics (Table 2). Notably, schMed_Pn exhibited a high number of transcripts over 1,000 nt and maintained a competitive N50 value (the length of the shortest contig or scaffold covering 50% of the total assembly), reflecting effective transcript coverage and assembly quality.

Table 2. Comparative metrics for selected transcriptome assemblies of *Schmidtea mediterranea*

Metric	**dd_Smed_v6**	**schMed_Pn**	**schMedS1**	**schMedS2**	**schMedS3_1**	**schMedS3_2**	**SmedAsxl_v1.1**	**SmedGD_c1.3**	**SMESG.1**
No. of Transcripts	41,745	41,396	22,348	27,79	41,358	41,036	22,855	29,85	31,102
Min transcript length (nt)	201	200	267	267	145	131	75	87	267
Max transcript length (nt)	19,524	48,466	86,373	1,022,601	76,733	94,201	43,226	12,11	48,121
Total Bases (M)	53.23	97.09	40.12	57.93	72.60	71.81	25.55	23.15	65.55

Mean transcript length (nt)	1,275	2,345	1,795	2,084	1,755	1,75	1,114	770	2,108
No. of transcripts >1k	18,487	31,199	16,082	20,727	29,402	28,932	8,61	7,218	23,648
No. of transcripts >10k	94	702	67	293	116	119	47	2	346
No. of transcripts with ORF	20,461	32,798	18,694	23,854	31,503	31,147	12,937	17,567	26,665
N50 (nt)	1,931	3,070	2,159	2,535	2,133	2,141	1,653	1,011	2,570
GC (%)	0.34	0.33	0.33	0.33	0.33	0.33	0.34	0.36	0.33

Performance benchmarking enabled the identification of improvements and areas needing enhancement while also positioning schMed_Pn within the context of current resources. These analyses collectively demonstrate that our approach offers a balanced and complete annotation, effectively capturing a wide range of transcript lengths and maintaining high completeness.

Discussion

Numerous pipelines and tools have been developed to tackle genome annotation across diverse organisms. For well-studied species—such as *Drosophila melanogaster* or *Arabidopsis*

thaliana—community-maintained platforms (FlyBase [44] and TAIR [45,46] respectively) provide robust resources that streamline annotation. In prokaryotes, automated tools like Prokka [10] efficiently handle smaller, less complex genomes, while eukaryotic model organisms often benefit from RefSeq [7] or large consortia efforts, which incorporate extensive experimental validation. However, non-model eukaryotes, including many invertebrates and plants, typically face an incomplete reference, novel repetitive elements, and limited training data for *ab initio* predictions. Pipelines such as Maker [5,47], Braker [48,49], GGA [9], nf-core GenomeAnnotator [12], and others can automate workflow portions—especially when reference data or orthologous gene sets are available. However, in certain species with high genomic complexity or less functional annotation data, these tools may still underperform or require extensive manual curation.

In this study, we developed and validated a comprehensive genome annotation pipeline designed to address the challenges of working with non-classical model organisms, while generating high-quality annotations applicable to broader genomic research. Our pipeline uniquely combines: (i) the integration of multiple data types, (ii) automated artifact detection, (iii) dual approach assembly —reference-based plus *de novo*, and (iv) broad functional integration.

Our analysis identified 4,011 novel genes and multiple previously uncharacterized isoforms and ncRNAs in the asexual strain, enriching the genomic landscape of *Schmidtea mediterranea*. Notably, our GO term analysis of putative MSTRG genes revealed enrichment in diverse biological functions. Among these, “chondroitin 4-sulfotransferase activity” and “diterpenoid biosynthetic process” suggest that these novel genes might contribute to unique regulatory pathways or stress responses in the asexual strain. Although many of these assignments are based on low-frequency observations, the presence of functions related to RNA binding (box

C/D and H/ACA snoRNP complex binding) further underscores the potential regulatory complexity.

While our transcriptomic data offers strain-specific insights, relying on the genome of the sexual strain as the reference introduces challenges. Differences between strains, such as chromosomal translocation and other genetic variations, may compromise the accuracy of gene prediction and transcript assembly, leading to potential inconsistencies in annotation quality. Variations in gene expression, alternative splicing events, and structural genomic differences necessitate meticulous bioinformatics strategies to ensure reliable annotations.

A deeper understanding of the genomic arrangement highlights further complexities that demand careful interpretation. Genes located in close proximity can reflect various biological scenarios. One possibility is the presence of functionally independent genes that happen to be closely spaced, as observed in other genomes where gene density is high [50,51]. For example, two planarian genes—SMESG000043996.1 and SMESG000043998.1—encode kynureninase and synaptotagmin XVII, respectively, are unrelated but positioned close to each other, supporting their independence while reflecting high gene density. Alternatively, such arrangements may arise from evolutionary events such as gene duplication followed by functional divergence, where both copies remain in close physical proximity. Another scenario includes operon-like gene structures commonly seen in prokaryotes but also identified in some eukaryotic contexts [52]. These structures allow the co-regulation of adjacent genes through shared promoters or other regulatory elements. Bidirectional promoters, enabling two genes to be transcribed in opposite directions from the same regulatory region, represent another mechanism driving close gene arrangements [53].

Non-contiguous transcripts, where exons are unusually spaced or appear fragmented, often arise from structural variations, alternative splicing, or sequencing artifacts. Biologically, large introns or dispersed exonic regions can be linked to long-range regulatory elements or structural genome features such as transposon insertions. Moreover, such patterns may indicate complex gene models, as seen in genes encoding for proteins with repetitive domains [54,55] or in cases where alternative promoters and polyadenylation sites lead to diverse transcript forms [56,57].

Consequently, accurately capturing these intricate gene architectures often requires more comprehensive strategies than those employed in simpler genomes, underscoring the need for user scrutiny and flexibility in annotation pipelines. The primary advantage of our pipeline lies in its capacity to address these complexities. By combining reference-based and *de novo* methods and incorporating both coding and non-coding RNA annotations it addresses the challenges posed by limited reference genomes. However, the pipeline's reliance on user-defined thresholds and manual curation introduces subjectivity, potentially impacting reproducibility.

The existing, planarian-specific repositories—such as PlanMine [21], PLANAtools [58], WomBase [59], PlanoSphere [60], PlanNET [61], and various planaria single-cell atlas datasets [62–65]—provide essential resources for gene model comparisons, single-cell data, and functional information in *S. mediterranea*. Despite their value, differences in naming conventions, annotation updates, and coverage across sexual and asexual strains can lead to inconsistencies. These discrepancies highlight the importance of robust, well-documented pipelines that can reconcile or supplement these external databases, thus ensuring more accurate gene models.

Additionally, while our automated steps effectively identify fragmented and chimeric transcripts, resolving these artifacts requires specialized expertise and is not fully automated. Another limitation concerns functional annotations, which can lead to false positives in identifying

chimeric genes. For instance, the same protein from different species or proteins representing different subunits of a complex may be annotated as distinct entries in functional databases. This can lead to genes being incorrectly flagged as chimeric if the functional entries do not exactly match. As a result, refined algorithms that account for near-identical annotations and minor terminological differences are vital.

Future enhancements could include the integration of machine learning algorithms to automate artifact resolution and improve annotation consistency. Expanding the pipeline to support multi-omics data integration, such as proteomics and epigenomics, would further enhance its utility. Community-driven efforts to standardize annotation protocols and databases would ensure broader applicability and reproducibility. Finally, experimental validation of novel genes and isoforms identified through this pipeline will be critical to confirm their biological relevance.

Compared to pipelines that excel in either prokaryotic or model vertebrate annotation – where reference data is abundant and well-curated–, our approach fills the gap for non-classical model systems with more limited resources. In doing so, it advances genome annotation efforts for *Schmidtea mediterranea* and provides a robust framework for addressing similar challenges in other species, fostering broader genomic discoveries.

Availability of Source Code and Requirements

Project name: SmedAnno

Project homepage: https://github.com/Norreanea/SmedAnno

Operating system: Linux

Programming languages: shell, R, Python, awk

License: MIT License

Data Availability

The *in-silico* datasets supporting the results and all other code used in this article are available in the GitHub repository archived in Zenodo https://doi.org/10.5281/zenodo.15130394. RNA-seq data can be accessed from the following GEO repositories: GSE293999 and GSE293997 from NCBI BioProject PRJNA1243063; and GSE294569 from NCBI BioProject PRJNA1246370. The final genome annotation file is available in GSE294569.

List of Abbreviations

BP: base pairs; NT: nucleotide; RNA: ribonucleic acid; rRNA: ribosomal RNA; ncRNA: non-coding RNA; SIRV: Spike-In RNA Variants; ORF: open reading frame; GTF: Gene Transfer Format; BLAST: Basic Local Alignment Search Tool; STAR: Spliced Transcripts Alignment to a Reference; BUSCO: Benchmarking Universal Single-Copy Orthologs; GO: Gene Ontology; cDNA: complementary DNA; DNA: deoxyribonucleic acid; RIN: RNA Integrity Number; GGA: Galaxy Genome Annotation.

Competing Interests

Nothing to declare

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used the free version of OpenAI-ChatGPT c2022 (v.GPT-o3-mini, February 2025) in order to improve manuscript readability and language. After using this tool, the authors thoroughly reviewed and modified the manuscript and took full responsibility for the content of the publication. The literature survey was done only by the authors, with no input from the AI tools.

Funding

This work was financed by the National Science Centre, Poland, through a grant no. 2019/35/B/NZ2/02658 to PJ.

Authors' Contributions

AZ: Conceptualization, Methodology, Software, Visualization, Writing – Original Draft, Writing – Review & Editing; MMZ: Conceptualization, Methodology, Writing – Review & Editing;

AD: Investigation, Writing – Review & Editing; ASCz: Investigation, Writing – Review & Editing; PJ: Conceptualization, Supervision, Project administration, Funding Acquisition, Writing – Review & Editing.

References

1. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006; doi: 10.1093/nar/gkl200.

2. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.. BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009; doi: 10.1186/1471-2105-10-421.

3. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; doi: 10.1093/bioinformatics/bts635.

4. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol*. Public Library of Science; 2022; doi: 10.1371/journal.pcbi.1009730.

5. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al.. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 2008; doi: 10.1101/gr.6743907.

6. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al.. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 2008; doi: 10.1186/gb-2008-9-1-r7.

7. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al.. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. Oxford University Press; 2016; doi: 10.1093/nar/gkv1189.

8. Kurylo C, Guyomar C, Foissac S, Djebali S. TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data. *NAR Genom Bioinform*. Oxford University Press; 2023; doi: 10.1093/nargab/lqad089.

9. Hiltemann S, Rasche H, Gladman S, Hotz HR, Larivière D, Blankenberg D, et al.. Galaxy Training: A powerful framework for teaching! *PLoS Comput Biol*. Public Library of Science; 2023; doi: 10.1371/journal.pcbi.1010752.

10. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. Oxford University Press; 2014; doi: 10.1093/bioinformatics/btu153.

11. Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc*. 2012; doi: 10.1038/nprot.2012.068.

12. 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; doi: 10.1038/s41587-020-0439-x.

13. Wu S, Yuan S, Su Z. User-friendly genome assembly and gene annotation pipelines for vertebrates. bioRxiv. 2024; doi: 10.1101/2024.05.21.595213

14. Maia GA, Filho VB, Kawagoe EK, Teixeira Soratto TA, Moreira RS, Grisard EC, et al.. AnnotaPipeline: An integrated tool to annotate eukaryotic proteins using multi-omics data. *Front Genet*. 2022; doi: 10.3389/fgene.2022.1020100.

15. Humann JL, Lee T, Ficklin S, Main D. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. *Gene Prediction.* 2019; doi: 10.1007/978-1-4939-9173-0_3

16. Sánchez Alvarado A. The freshwater planarian Schmidtea mediterranea: Embryogenesis, stem cells and regeneration. *Curr Opin Genet Dev*. Elsevier Ltd; 2003; doi: 10.1016/S0959-437X(03)00082-0

17. Baguñà J, Carranza S, Pala M, Ribera C, Giribet G, Arnedo MA, et al.. From morphology and karyology to molecules. New methods for taxonomical identification of asexual populations of freshwater planarians. A tribute to Professor Mario Benazzi. *Italian Journal of Zoology*. Societa Paleontologica Italiana; 1999; doi: 10.1080/11250009909356258.

18. Robb SMC, Ross E, Alvarado AS. SmedGD: The Schmidtea mediterranea genome database. *Nucleic Acids Res*. 2008; doi: 10.1093/nar/gkm684.

19. Robb SMC, Gotting K, Ross E, Sánchez Alvarado A. SmedGD 2.0: The Schmidtea mediterranea genome database. *Genesis*. John Wiley and Sons Inc.; 2015; doi: 10.1002/dvg.22872.

20. Grohme MA, Schloissnig S, Rozanski A, Pippel M, Young GR, Winkler S, et al.. The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. *Nature*. Nature Publishing Group; 2018; doi: 10.1038/nature25473.

21. Rozanski A, Moon HK, Brandl H, Martín-Durán JM, Grohme MA, Hüttner K, et al.. PlanMine 3.0 - improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic Acids Res*. Oxford University Press; 2019; doi: 10.1093/nar/gky1070.

22. Guo L, Bloom JS, Dols-Serrate D, Boocock J, Ben-David E, Schubert OT, et al.. Island-specific evolution of a sex-primed autosome in a sexual planarian. *Nature*. Nature Research; 2022; doi: 10.1038/s41586-022-04757-3.

23. Ivanković M, Brand JN, Pandolfini L, Brown T, Pippel M, Rozanski A, et al.. A comparative analysis of planarian genomes reveals regulatory conservation in the face of rapid structural divergence. *Nat Commun*. 2024; doi: 10.1038/s41467-024-52380-9.

24. Rabiasz A, Ziętkiewicz E. Schmidtea mediterranea as a Model Organism to Study the Molecular Background of Human Motile Ciliopathies. Int J Mol Sci. Multidisciplinary Digital Publishing Institute (MDPI); *Int J Mol Sci* . 2023; doi: 10.3390/ijms24054472

25. Quinodoz S, Thomas MA, Dunkel J, Schötz EM. The More the Merrier?: Entropy and Statistics of Asexual Reproduction in Freshwater Planarians. *J Stat Phys*. Springer New York LLC; 2011; doi: 10.1007/s10955-011-0157-3.

26. Reddien PW. The Cellular and Molecular Basis for Planarian Regeneration. *Cell*. Cell Press; 2018; doi: 10.1016/j.cell.2018.09.021

27. Reddien PW, Sánchez Alvarado A. Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol*. 2004; doi: 10.1146/annurev.cellbio.20.010403.095114

28. Benham-Pyle BW, Mann FG, Brewster CE, Dewars ER, Vuu DM, Nowotarski SH, et al.. Planarians employ diverse and dynamic stem cell microenvironments to support whole-body regeneration. bioRxiv. 2023; doi: 10.1101/2022.03.20.485025

29. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al.. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. Oxford University Press; 2022; doi: 10.1093/nar/gkab1112.

30. Palmieri N, Schlötterer C. Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS One*. 2009; doi: 10.1371/journal.pone.0006323.

31. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018; doi: 10.1038/s41576-018-0003-4.

32. Chen Z, Ain NU, Zhao Q, Zhang X. From tradition to innovation: conventional and deep learning frameworks in genome annotation. *Brief Bioinformatics*. Oxford University Press; 2024; doi: 10.1093/bib/bbae138

33. Newmark PA, Sánchez Alvarado A. Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. *Dev Biol*. Academic Press Inc.; 2000; doi: 10.1006/dbio.2000.9645.

34. Dainat J, Hereñú D, Dr. Murray KD, Davis E, Ugrin I, Crouch K, et al.. NBISweden/AGAT: AGAT-v1.4.1 (v1.4.1). Zenodo;

35. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. Oxford University Press; 2015; doi: 10.1093/bioinformatics/btv351.

36. Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*. Oxford University Press; 2016; doi: 10.1093/bioinformatics/btw218.

37. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Res*. Cold Spring Harbor Laboratory Press; 2016; doi: 10.1101/gr.196469.115.

38. Sweeney BA, Petrov AI, Ribas CE, Finn RD, Bateman A, Szymanski M, et al.. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res*. 2021; doi: 10.1093/nar/gkaa921.

39. Francis WR, Wörheide G. Similar ratios of introns to intergenic sequence across animal genomes. *Genome Biol Evol*. Oxford University Press; 2017; doi: 10.1093/gbe/evx103.

40. Gudlaugsdottir S, Boswell DR, Wood GR, Ma J. Exon size distribution and the origin of introns. *Genetica*. 2007; doi: 10.1007/s10709-007-9139-4.

41. Bhasi A, Philip P, Manikandan V, Senapathy P. ExDom: An integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res*. 2009; doi: 10.1093/nar/gkn746.

42. Lakshmanan V, Sujith TN, Bansal D, Shivaprasad P V., Palakodeti D, Krishna S. Comprehensive annotation and characterization of planarian tRNA and tRNA-derived fragments (tRFs). *RNA*. 2021; doi: 10.1261/rna.077701.120.

43. Larsen N, Zwieb C. SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res*. 1991; doi: 10.1093/nar/19.2.209.

44. Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, et al.. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics*. 2024; doi: 10.1093/genetics/iyad211.

45. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al.. The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *genesis*. 2015; doi: 10.1002/dvg.22877.

46. Cheng C, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*. 2017; doi: 10.1111/tpj.13415.

47. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 2011; doi: 10.1186/1471-2105-12-491.

48. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. *Methods Mol Biol* . 2019; doi: 10.1007/978-1-4939-9173-0_5

49. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*. 2021; doi: 10.1093/nargab/lqaa108.

50. Arnone JT, Robbins-Pianka A, Arace JR, Kass-Gergi S, McAlear MA. The adjacent positioning of co-regulated gene pairs is widely conserved across eukaryotes. *BMC Genomics*. 2012; doi: 10.1186/1471-2164-13-546.

51. Gherman A, Wang R, Avramopoulos D. Orientation, distance, regulation and function of neighbouring genes. *Hum Genomics*. 2009; doi: 10.1186/1479-7364-3-2-143.

52. Kominek J, Doering DT, Opulente DA, Shen XX, Zhou X, DeVirgilio J, et al.. Eukaryotic Acquisition of a Bacterial Operon. *Cell*. Cell Press; 2019; doi: 10.1016/j.cell.2019.01.034.

53. Ahmad SS, Samia NSN, Khan AS, Turjya RR, Khan MAAK. Bidirectional promoters: an enigmatic genome architecture and their roles in cancers. *Mol Biol Rep*. Springer Science and Business Media B.V.; 2021; doi: 10.1007/s11033-021-06612-6.

54. Hoock TC, Peters LL, Lux SE. Isoforms of Ankyrin-3 That Lack the NH2-terminal Repeats Associate with Mouse Macrophage Lysosomes. *J Cell Biol*. 1997; doi: 10.1083/jcb.136.5.1059.

55. Yamankurt G, Wu HC, McCarthy M, Cunha SR. Exon Organization and Novel Alternative Splicing of Ank3 in Mouse Heart. *PLoS One*. 2015; doi: 10.1371/journal.pone.0128177.

56. Aísa-Marín I, García-Arroyo R, Mirra S, Marfany G. The Alter Retina: Alternative Splicing of Retinal Genes in Health and Disease. *Int J Mol Sci*. 2021; doi: 10.3390/ijms22041855.

57. Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC, et al.. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res*. 2011; doi: 10.1101/gr.120535.111.

58. Hoffman M, Wurtzel O. PLANAtools—An interactive gene expression repository for the planarian Schmidtea mediterranea. *Front Cell Dev Biol*. 2023; doi: 10.3389/fcell.2023.1149537.

59. Sternberg PW, Van Auken K, Wang Q, Wright A, Yook K, Zarowiecki M, et al.. WormBase 2024: status and transitioning to Alliance infrastructure. *Genetics*. 2024; doi: 10.1093/genetics/iyae050.

60. Nowotarski SH, Davies EL, Robb SMC, Ross EJ, Matentzoglu N, Doddihal V, et al.. Planarian Anatomy Ontology: a resource to connect data within and across experimental platforms. *Development*. 2021; doi: 10.1242/dev.196097.

61. Castillo-Lara S, Abril JF. PlanNET: homology-based predicted interactome for multiple planarian transcriptomes. *Bioinformatics*. 2018; doi: 10.1093/bioinformatics/btx738.

62. Cui G, Dong K, Zhou J-Y, Li S, Wu Y, Han Q, et al.. Spatiotemporal transcriptomic atlas reveals the dynamic characteristics and key regulators of planarian regeneration. *Nat Commun*. 2023; doi: 10.1038/s41467-023-39016-0.

63. King HO, Owusu-Boaitey KE, Fincher CT, Reddien PW. A transcription factor atlas of stem cell fate in planarians. *Cell Rep*. 2024; doi: 10.1016/j.celrep.2024.113843.

64. Fincher CT, Wurtzel O, de Hoog T, Kravarik KM, Reddien PW. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science (1979)*. 2018; doi: 10.1126/science.aaq1736.

65. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, et al.. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science (1979)*. 2018; doi: 10.1126/science.aaq1723.

Figure Legends

Figure 1. **Identification rates and accuracy metrics for *in-silico* and SIRVs datasets.**

(A) - Transcript identification rates using *in-silico* data. Circles depict the four approaches tested for short read (SR) data, using *de novo* (DN) or reference-based (RB) assembly and StringTie2 version 2.1.1 (211) or version 2.2.1 (221). (B) - Accuracy metrics of transcript isoform identification using SIRVs on SR, mixed reads (MR) and merged reads (MeR) approaches. Whiskers on the bar plots represent the variability in accuracy across five different StringTie2 parameter configurations tested.

Figure 2. **Workflow diagram for RNA-Seq pre-processing, transcript assembly, and functional annotation.**

Figure 3. **User-driven inspection and refinement.**

Figure 4. **Annotation artifacts due to errors.**

(A) - Potentially fragmented gene scenario with SMESG000004200.1, SMESG000004201.1, and MSTRG.967. (B) - The chimeric transcript prediction in SMESG000065897.1 due to an excessively large intron in MSTRG.16892.1. (C) - The functionally chimeric gene SMESG000077853.1 with divergent transcript annotations.

Figure 5. **Venn diagram and a word cloud of GO terms.**

(A) - Venn diagram of GO term overlaps between MSTRG (novel genes) and SMESG (reference genes). (B) - Word cloud of unique GO terms assigned to SMESG (reference) genes. (C) - Word cloud of common GO terms assigned to both SMESG and MSTRG genes. (D) - Word cloud of unique GO terms assigned to MSTRG (novel) genes.

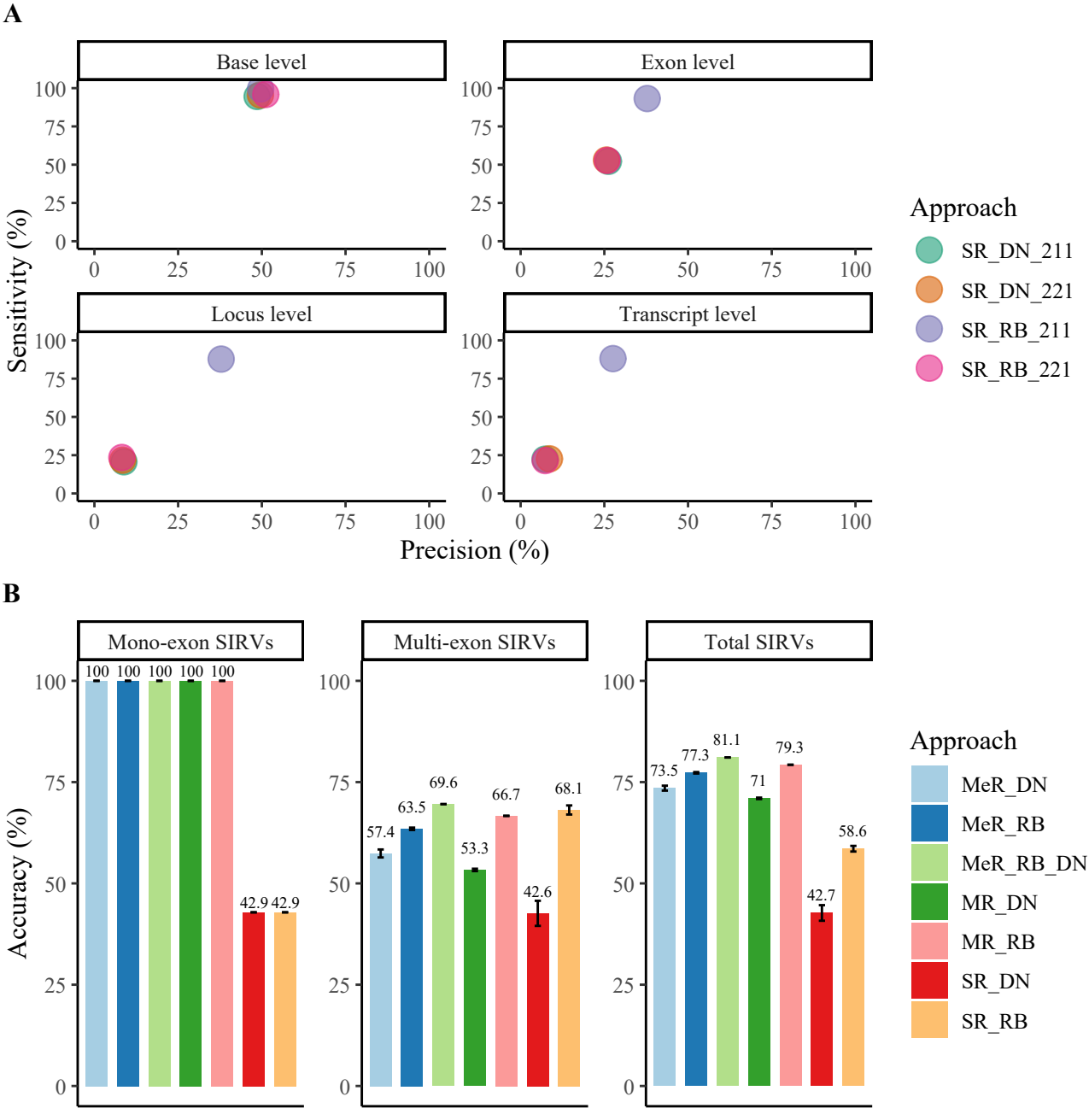
Figure 6. **Performance metrics and annotation features of transcriptome assemblies.**

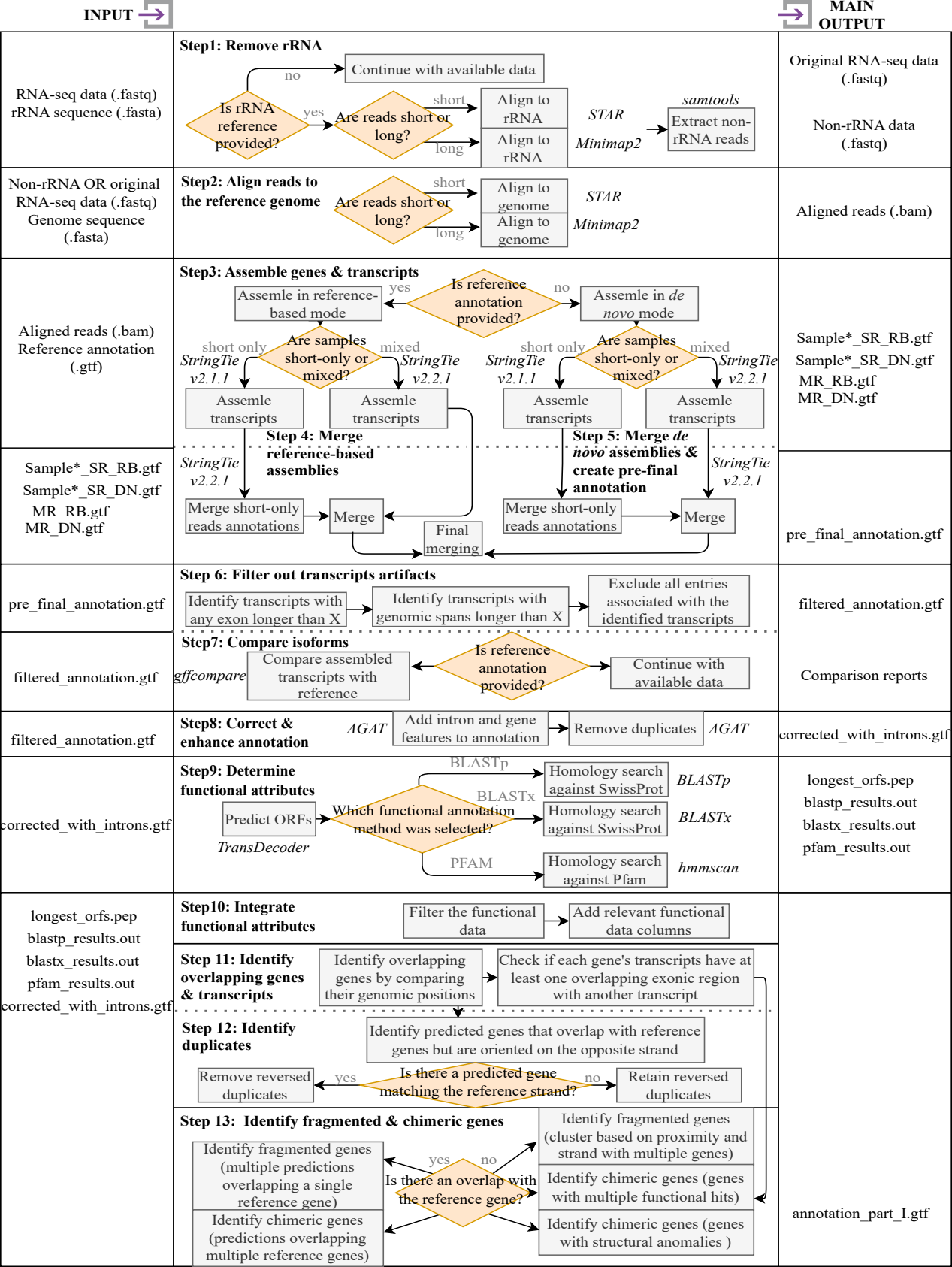
(A) - Cumulative distribution of transcript lengths in selected annotations. The y-axis (logarithmic scale) represents the cumulative count of transcripts/isoforms, indicating how many transcripts exceed each length threshold on the x-axis (linear scale). (B) - BUSCO evaluation of unigene assembly completeness.

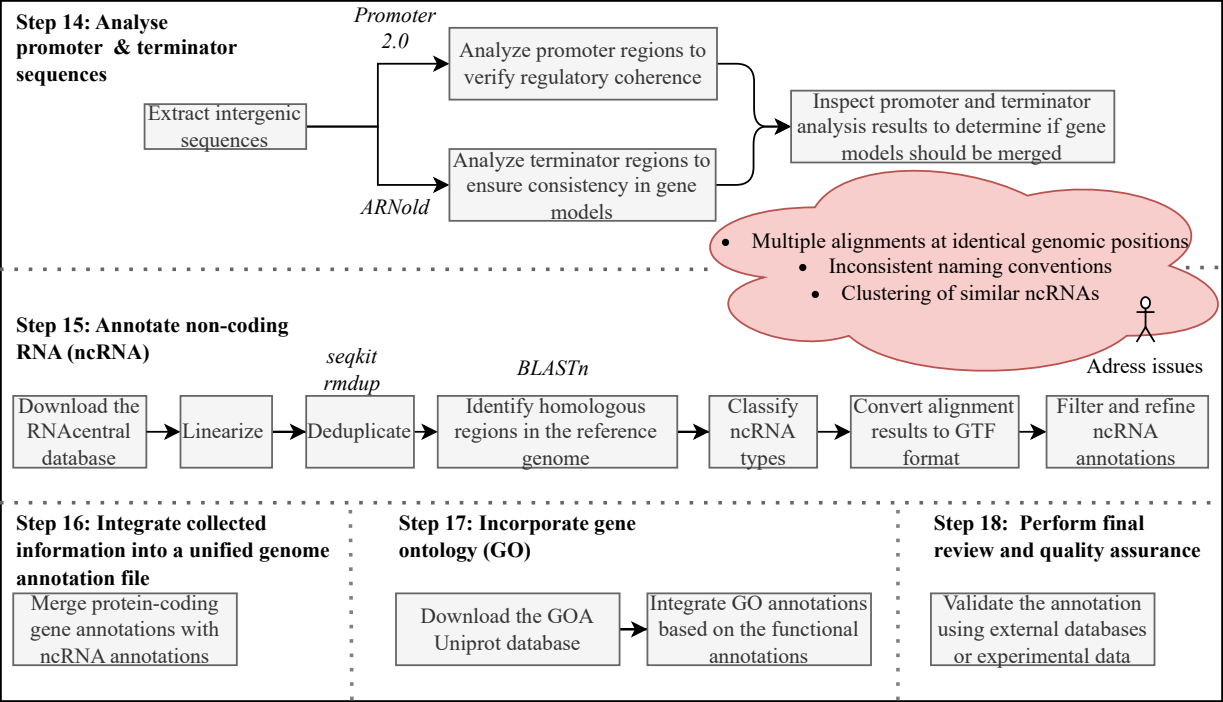
Additional Files

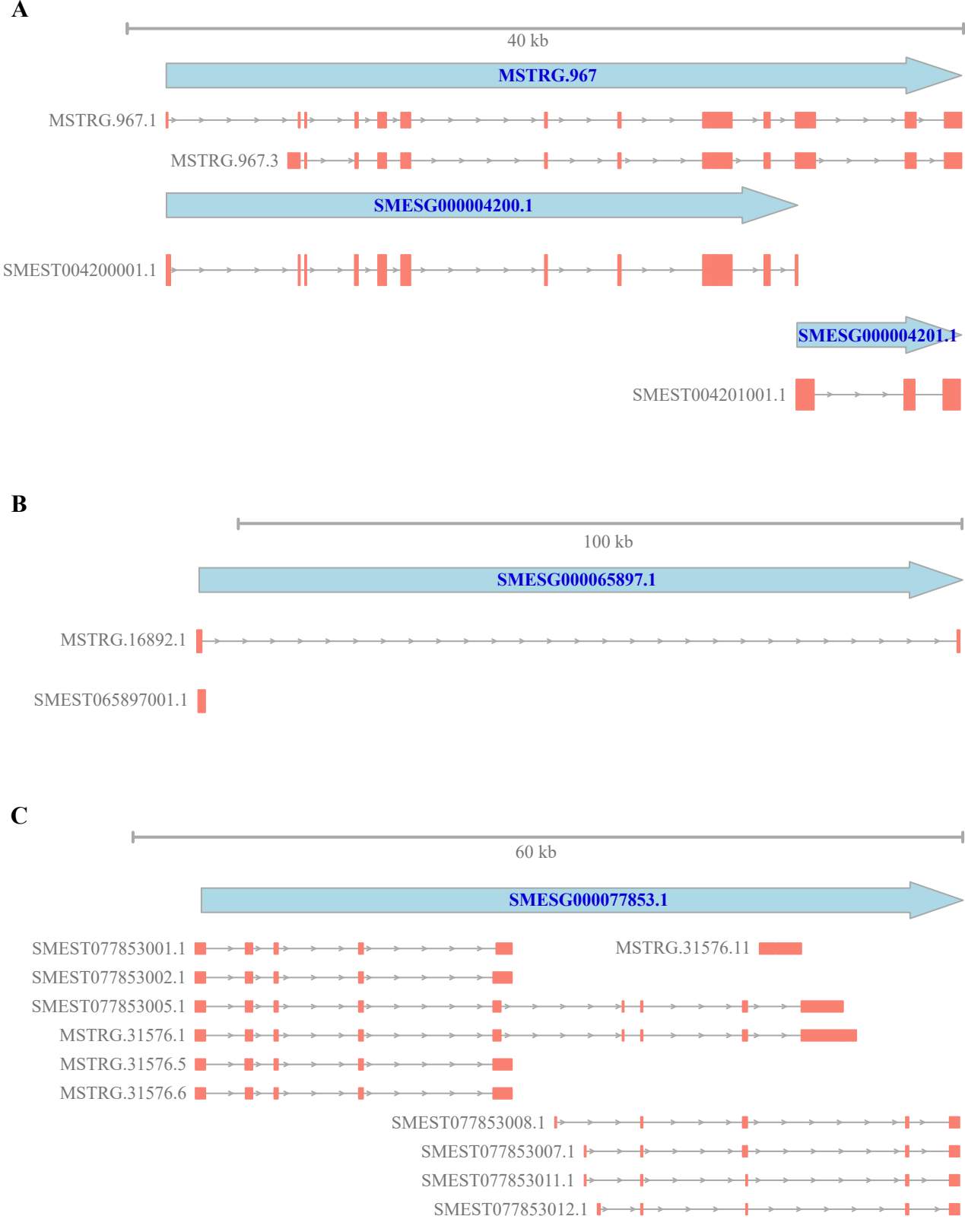
Supplementary Table S1. Sequencing samples summary.

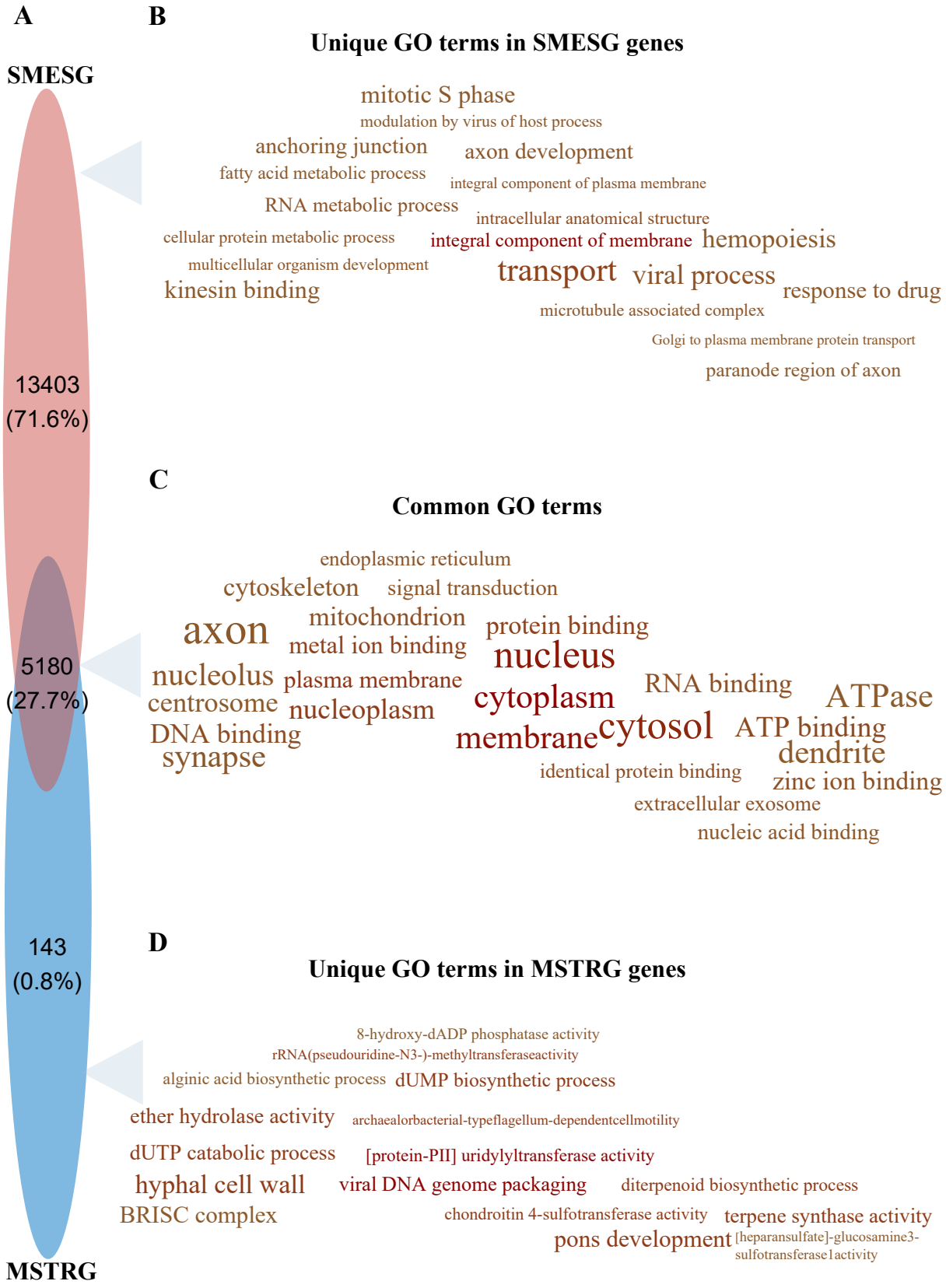
Supplementary Table S2. StringTie2 parameters tested on SIRVs.


A
Base level
Exon level
Locus level
Transcript level
Sensitivity (%)
Precision (%)
0
25
50
75
100
Approach
SR_DN_211
SR_DN_221
SR_RB_211
SR_RB_221
B
Mono-exon SIRVs
Multi-exon SIRVs
Total SIRVs
Accuracy (%)
100 100 100 100 100 42.9 42.9
57.4 63.5 69.6 53.3 66.7 42.6 68.1
73.5 77.3 81.1 71 79.3 42.7 58.6
Approach
MeR_DN
MeR_RB
MeR_RB_DN
MR_DN
MR_RB
SR_DN
SR_RB

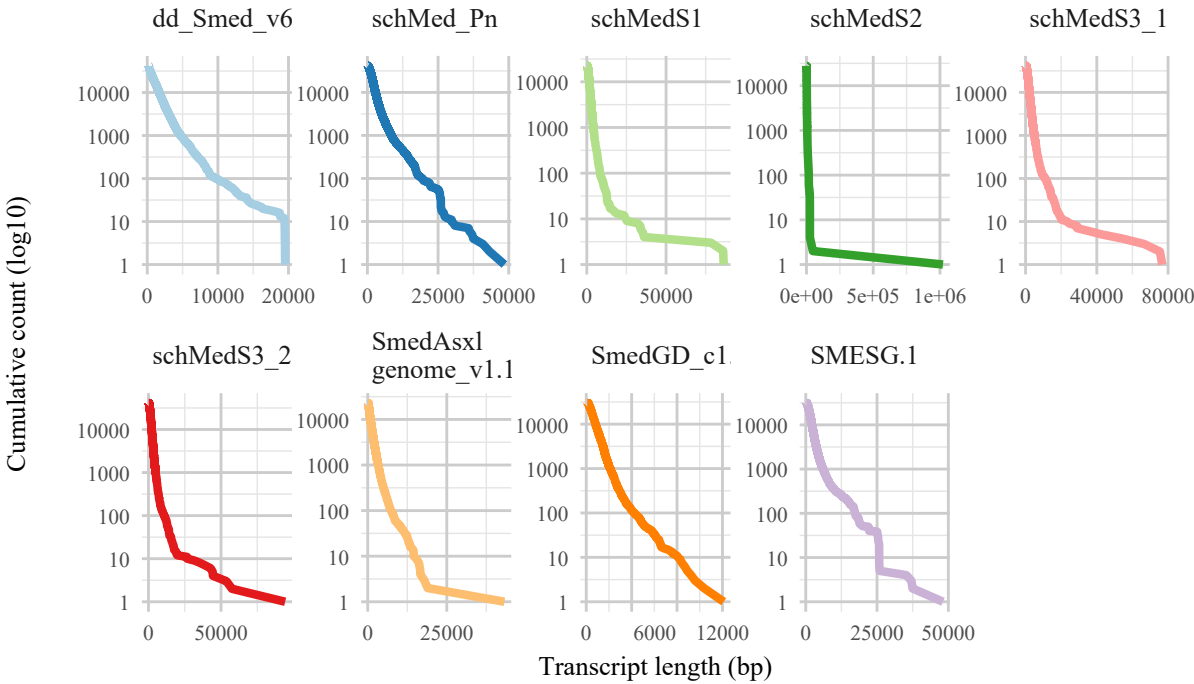

INPUT
MAIN OUTPUT
RNA-seq data (.fastq)
rRNA sequence (.fasta)
Step1: Remove rRNA
Is rRNA reference provided?
no
Continue with available data
yes
Are reads short or long?
short
long
Align to rRNA
Align to rRNA
STAR
Minimap2
samtools
Extract non-rRNA reads
Original RNA-seq data (.fastq)
Non-rRNA data (.fastq)
Non-rRNA OR original RNA-seq data (.fastq)
Genome sequence (.fasta)
Step2: Align reads to the reference genome
Are reads short or long?
short
long
Align to genome
Align to genome
STAR
Minimap2
Aligned reads (.bam)
Aligned reads (.bam)
Reference annotation (.gtf)
Step3: Assemble genes & transcripts
Is reference annotation provided?
yes
Assemle in reference-based mode
no
Assemle in de novo mode
Are samples short-only or mixed?
short only
mixed
StringTie v2.1.1
StringTie v2.2.1
Assemle transcripts
Assemle transcripts
Assemle transcripts
Assemle transcripts
Sample*_SR_RB.gtf
Sample*_SR_DN.gtf
MR_RB.gtf
MR_DN.gtf
Step 4: Merge reference-based assemblies
Step 5: Merge de novo assemblies & create pre-final annotation
Merge short-only reads annotations
Merge
Merge short-only reads annotations
Merge
Final merging
Sample*_SR_RB.gtf
Sample*_SR_DN.gtf
MR_RB.gtf
MR_DN.gtf
pre_final_annotation.gtf
pre_final_annotation.gtf
Step 6: Filter out transcripts artifacts
Identify transcripts with any exon longer than X
Identify transcripts with genomic spans longer than X
Exclude all entries associated with the identified transcripts
filtered_annotation.gtf
filtered_annotation.gtf
Step7: Compare isoforms
gffcompare
Compare assembled transcripts with reference
Is reference annotation provided?
Continue with available data
Comparison reports
filtered_annotation.gtf
Step8: Correct & enhance annotation
AGAT
Add intron and gene features to annotation
Remove duplicates
AGAT
corrected_with_introns.gtf
corrected_with_introns.gtf
Step9: Determine functional attributes
Predict ORFs
TransDecoder
Which functional annotation method was selected?
BLASTp
BLASTx
PFAM
Homology search against SwissProt
Homology search against SwissProt
Homology search against Pfam
BLASTp
BLASTx
hmmscan
longest_orfs.pep
blastp_results.out
blastx_results.out
pfam_results.out
longest_orfs.pep
blastp_results.out
blastx_results.out
pfam_results.out
corrected_with_introns.gtf
Step10: Integrate functional attributes
Filter the functional data
Add relevant functional data columns
Step 11: Identify overlapping genes & transcripts
Identify overlapping genes by comparing their genomic positions
Check if each gene's transcripts have at least one overlapping exonic region with another transcript
Step 12: Identify duplicates
Identify predicted genes that overlap with reference genes but are oriented on the opposite strand
Is there a predicted gene matching the reference strand?
yes
no
Remove reversed duplicates
Retain reversed duplicates
Step 13: Identify fragmented & chimeric genes
Is there an overlap with the reference gene?
yes
no
Identify fragmented genes (multiple predictions overlapping a single reference gene)
Identify chimeric genes (predictions overlapping multiple reference genes)
Identify fragmented genes (cluster based on proximity and strand with multiple genes)
Identify chimeric genes (genes with multiple functional hits)
Identify chimeric genes (genes with structural anomalies)
annotation_part_I.gtf


Step 14: Analyse promoter & terminator sequences
Promoter 2.0
Extract intergenic sequences
Analyze promoter regions to verify regulatory coherence
Analyze terminator regions to ensure consistency in gene models
ARNold
Inspect promoter and terminator analysis results to determine if gene models should be merged
Multiple alignments at identical genomic positions
Inconsistent naming conventions
Clustering of similar ncRNAs
Adress issues
Step 15: Annotate non-coding RNA (ncRNA)
seqkit rmdup
BLASTn
Download the RNAcentral database
Linearize
Deduplicate
Identify homologous regions in the reference genome
Classify ncRNA types
Convert alignment results to GTF format
Filter and refine ncRNA annotations
Step 16: Integrate collected information into a unified genome annotation file
Merge protein-coding gene annotations with ncRNA annotations
Step 17: Incorporate gene ontology (GO)
Download the GOA Uniprot database
Integrate GO annotations based on the functional annotations
Step 18: Perform final review and quality assurance
Validate the annotation using external databases or experimental data

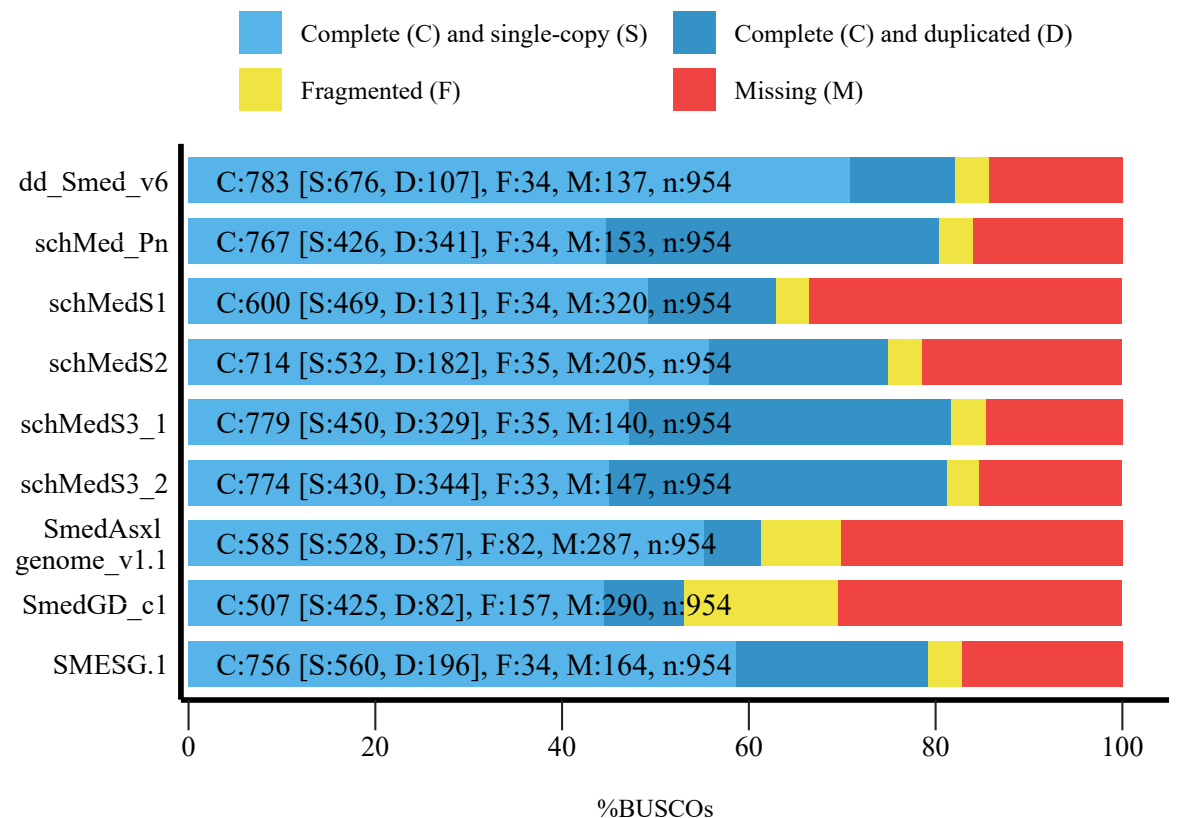

A
40 kb
MSTRG.967
MSTRG.967.1
MSTRG.967.3
SMESG000004200.1
SMEST004200001.1
SMESG000004201.1
SMEST004201001.1
B
100 kb
SMESG000065897.1
MSTRG.16892.1
SMEST065897001.1
C
60 kb
SMESG000077853.1
SMEST077853001.1
SMEST077853002.1
SMEST077853005.1
MSTRG.31576.1
MSTRG.31576.5
MSTRG.31576.6
MSTRG.31576.11
SMEST077853008.1
SMEST077853007.1
SMEST077853011.1
SMEST077853012.1


A
SMESG
13403
(71.6%)
5180
(27.7%)
143
(0.8%)
MSTRG
B
Unique GO terms in SMESG genes
mitotic S phase
modulation by virus of host process
anchoring junction
axon development
fatty acid metabolic process
integral component of plasma membrane
RNA metabolic process
intracellular anatomical structure
cellular protein metabolic process
integral component of membrane
hemopoiesis
multicellular organism development
transport
viral process
kinesin binding
response to drug
microtubule associated complex
Golgi to plasma membrane protein transport
paranode region of axon
C
Common GO terms
endoplasmic reticulum
cytoskeleton
signal transduction
axon
mitochondrion
protein binding
metal ion binding
nucleus
nucleolus
plasma membrane
RNA binding
ATPase
centrosome
nucleoplasm
cytoplasm
DNA binding
membrane
cytosol
ATP binding
synapse
dendrite
identical protein binding
zinc ion binding
extracellular exosome
nucleic acid binding
D
Unique GO terms in MSTRG genes
8-hydroxy-dADP phosphatase activity
rRNA(pseudouridine-N3-)-methyltransferaseactivity
alginic acid biosynthetic process
dUMP biosynthetic process
ether hydrolase activity
archaealorbacterial-typeflagellum-dependentcellmotility
dUTP catabolic process
[protein-PII] uridylyltransferase activity
hyphal cell wall
viral DNA genome packaging
diterpenoid biosynthetic process
BRISC complex
chondroitin 4-sulfotransferase activity
terpene synthase activity
pons development
[heparansulfate]-glucosamine3-sulfotransferase1activity

A


dd_Smed_v6
schMed_Pn
schMedS1
schMedS2
schMedS3_1
schMedS3_2
SmedAsxl genome_v1.1
SmedGD_c1
SMESG.1
Cumulative count (log10)
Transcript length (bp)

B


Complete (C) and single-copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)
dd_Smed_v6
C:783 [S:676, D:107], F:34, M:137, n:954
schMed_Pn
C:767 [S:426, D:341], F:34, M:153, n:954
schMedS1
C:600 [S:469, D:131], F:34, M:320, n:954
schMedS2
C:714 [S:532, D:182], F:35, M:205, n:954
schMedS3_1
C:779 [S:450, D:329], F:35, M:140, n:954
schMedS3_2
C:774 [S:430, D:344], F:33, M:147, n:954
SmedAsxl genome_v1.1
C:585 [S:528, D:57], F:82, M:287, n:954
SmedGD_c1
C:507 [S:425, D:82], F:157, M:290, n:954
SMESG.1
C:756 [S:560, D:196], F:34, M:164, n:954
%BUSCOs