Facilitating genome annotation using ANNEXA and long-read RNA sequencing

Nicolaï Hoffmann[1], Aurore Besson[1], Edouard Cadieu[2], Matthias Lorthiois[2], Victor Le Bars[2], Armel Houel[1], Christophe Hitte[1], Catherine André[1], Benoit Hédan[1], and Thomas Derrien[1]

https://doi.org/10.1101/2025.04.16.648718

Abstract

With the advent of complete genome assemblies, genome annotation has become essential for the functional interpretation of genomic data. Long-read RNA sequencing (LR-RNAseq) technologies have significantly improved transcriptome annotation by enabling full-length transcript reconstruction for both coding and non-coding RNAs. However, challenges such as transcript fragmentation and incomplete isoform representation persist, highlighting the need for robust quality control (QC) strategies. This study presents ANNEXA, a pipeline designed to enhance genome annotation using LR-RNAseq data while also providing QC for reconstructed genes and transcripts. ANNEXA integrates two transcriptome reconstruction tools, StringTie2 and Bambu, applying stringent filtering criteria to improve annotation accuracy. It also incorporates deep learning models to evaluate transcription start sites (TSSs) and employs the tool FEELnc for the systematic annotation of long non-coding RNAs (lncRNAs). Additionally, the pipeline offers intuitive visualisations for comparative analyses of coding and non-coding repertoires. Benchmarking against multiple reference annotations revealed distinct patterns of sensitivity and precision for both known and novel genes and transcripts and mRNAs and lncRNAs. To demonstrate its utility, ANNEXA was applied in a comparative oncology study involving LR-RNAseq of two human and eight canine cancer cell lines. The pipeline successfully identified novel genes and transcripts across species, expanding the catalog of protein-coding and lncRNA annotations in both species. Implemented in Nextflow for scalability and reproducibility, ANNEXA is available as an open-source tool: https://github.com/IGDRion/ANNEXA.

Keywords: Genome Annotation, long-read sequencing, lncRNAs, comparative transcriptomics

[1]Canine Genetics Team, CNRS, Univ Rennes – UMR6290, IGDR (Institut de Génétique et Développement de Rennes), 35043 Rennes, France, [2]IGDRion platform, CNRS, Univ Rennes – UMR6290, IGDR (Institut de Génétique et Développement de Rennes), 35043 Rennes, France

Correspondence
nicolai.hoffmann@univ-rennes.fr, thomas.derrien@univ-rennes.fr

Introduction

With the increasing availability of entire genome assemblies, i.e. telomere-to-telomere (T2T) (Nurk et al., 2022), one challenge in genome research is to move from improving genome completeness to refining genome annotation. High-quality genome sequences now enable a more precise characterisation of genes and transcripts, especially in repetitive regions of the genomes, making transcriptome-based annotation a critical step in the functional interpretability of the genomes (Nurk et al., 2022). To this end, RNA sequencing (RNA-seq) plays a central role in this process by providing direct transcript-level evidence, essential for defining gene structures, alternative splicing events, and non-coding RNA repertoires.

While short-read RNAseq (SR-RNASeq) has shown some limitations to reconstruct full-length transcripts (Steijger et al., 2013), long-read RNA sequencing (LR-RNAseq), provided by platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), has significantly advanced transcriptome annotation by providing reads that span repeats and also by allowing direct connectivity between distant exons of the same isoform (Conesa et al., 2016). However, despite these advantages, LR-RNAseq-based transcriptome reconstruction still remains prone to artifacts such as truncated transcript models and partial isoform representation (Sessegolo et al., 2019). To ensure accurate genome annotation, robust quality control (QC) and filtering strategies are needed to evaluate and refine transcriptome reconstructions.

Several tools have been developed to assemble long reads from LR-RNAseq data. A recent benchmark study from the LRGASP consortium has compared fourteen transcriptome reconstruction and quantification tools (Pardo-Palacios et al., 2024b) and showed that choosing the best program depends on the biological context of the study and the completeness of the reference annotation. Among the benchmarked tools, Bambu (Chen et al., 2023) and StringTie (Kovaka et al., 2019) consistently demonstrated strong performance across multiple metrics. However, one limitation of this benchmark is that it did not explicitly assess tool performance with respect to RNA biotypes, particularly in distinguishing between protein-coding transcripts (mRNAs) and long non-coding RNAs (lncRNAs). Given the complexity and heterogeneity of transcriptomes, this distinction could be important since different RNA biotypes may vary in expression levels, structural features, and evolutionary conservation, all of which can impact reconstruction accuracy. This consideration is especially important in light of the substantial expansion of the human lncRNA catalogue in recent Gencode releases (Kaur et al., 2024), underscoring the need for tools that can accurately reconstruct and annotate both known and novel transcripts across diverse RNA classes.

A number of pipelines dedicated to long-read RNA-seq analysis and annotation extension integrate some of the tools benchmarked in LRGASP but present certain limitations. For example, nf-core nanoseq (Chen et al., 2025) supports transcriptome reconstruction using Bambu or StringTie but it is restricted to Oxford Nanopore data and does not annotate long non-coding RNAs. TAGADA (Kurylo et al., 2023) and annotate_my_genomes (Farkas et al., 2022) offer lncRNA detection but rely exclusively on StringTie for transcript assembly. To overcome these limitations, we developed ANNEXA, a Nextflow-based pipeline specifically designed to extend reference annotations from long-read RNA-seq data while also assessing the quality of newly identified transcript models. ANNEXA integrates two transcriptome reconstruction tools, Bambu (Chen et al., 2023) and StringTie (Kovaka et al., 2019), and provides users with stringent filters to improve annotation accuracy. It incorporates a deep learning strategy to potentially remove truncated

transcript models by evaluating transcription start sites (TSSs) of all novel transcripts. ANNEXA is also designed to systematically annotate and evaluate the annotation of lncRNAs by incorporating the FEELnc program (Wucher et al., 2017). It provides intuitive visual representations of the annotation, facilitating comparative analysis of coding and non-coding repertoires. To illustrate the usability of ANNEXA in a comparative oncology project, we sequenced two human and eight canine cancer cell lines from mucosal melanomas, histiocytic sarcomas and osteosarcoma using ONT direct cDNA sequencing, and identified novel human and canine genes/transcripts, some of which being conserved in the two species.

By implementing a structured framework for quality control and annotation, ANNEXA enhances the reliability of long-read transcriptomics, ensuring more comprehensive and biologically meaningful extended genome annotations as a prerequisite for further functional studies.

Methods

Overview of ANNEXA

ANNEXA is a pipeline that extends user-provided reference annotations with novel genes and transcript isoforms from long-read sequencing data. It only uses three parameter files: a reference genome, a reference annotation and a text file listing the paths to BAM file(s). (**Fig. 1.A**). Unlike technology-specific pipelines such as nf-core/isoseq (Guizard et al., 2023), which is tailored for PacBio data, ANNEXA is compatible with both Nanopore and PacBio RNA sequencing technologies. The pipeline is organised into four main modules, described below:

(1) Transcriptome reconstruction and quantification
(2) Coding potential evaluation and transcript classification
(3) Transcript filtering and full-length assessment
(4) Quality Control

Transcriptome reconstruction and quantification

As recently demonstrated by the LRGASP consortium, the choice of the bioinformatic tool to model known and novel genes/transcripts depends on the biological context, and more particularly, on the expected levels of annotation precision (controlling the number of novel false-positive transcripts) and sensitivity/recall (controlling the number of novel false-negative transcripts). Ideally, users should be able to select the appropriate tool based on the expected number of novel transcripts and the quality or completeness of the reference genome and annotation. To support this, ANNEXA enables the reconstruction and quantification of both known and novel transcripts using two distinct tools: Bambu (Chen et al., 2023) and StringTie2 (Kovaka et al., 2019). When multiple input files are provided, ANNEXA pools together all samples to produce a single extended annotation. With StringTie, a first step is performed where each sample is processed separately, producing a GTF file specific to that sample. In a second step, all GTF files are merged together with the reference annotation using the `-merge` option from StringTie, producing a unified, non-redundant transcript set. With Bambu, it takes all samples together to process reads, and uses a machine learning algorithm to define a TPS (Transcript Probability Score) corresponding to the probability that a transcript is valid in at least one sample.
In ANNEXA, several options are available for Bambu, including the ability to adjust the Novel Discovery Rate (NDR) threshold, a key metric derived from the TPS that balances sensitivity and precision (Chen et al., 2023) across multiple samples. For instance, users can enhance sensitivity by

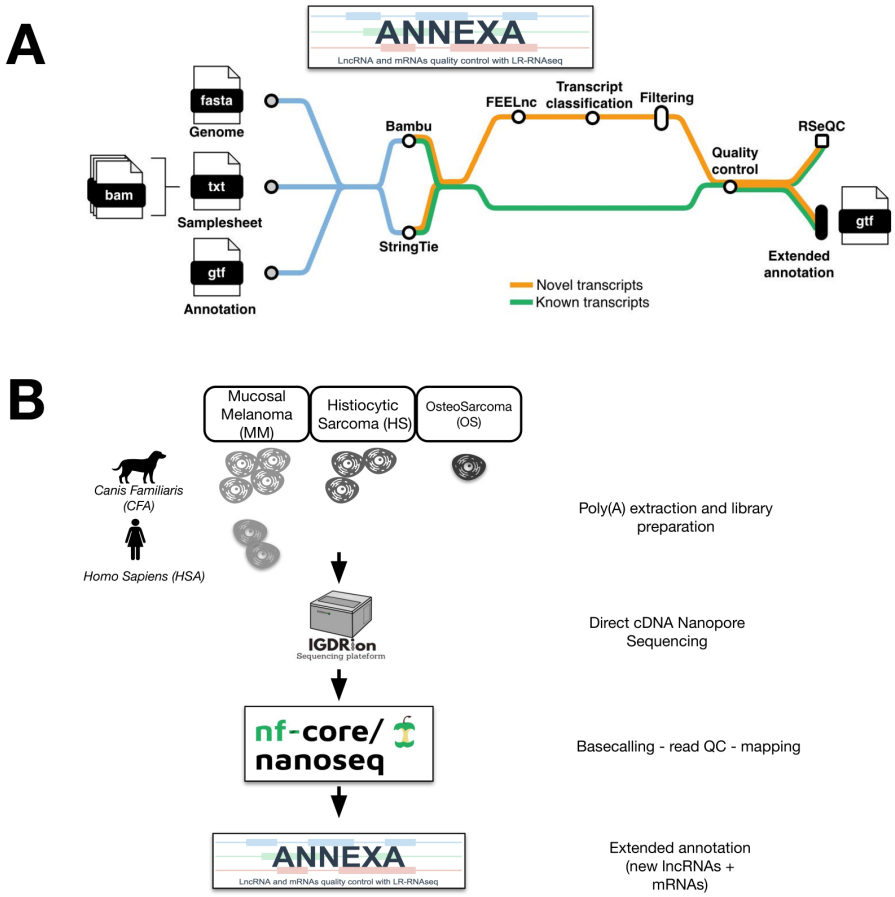


Figure 1 – **A**- Metromap of ANNEXA. **B**- Experimental design with species-specific input cancer cell lines and experimental and computational analyses.

increasing the default recommended NDR threshold with the `--bambu_threshold` option, and include single-exon transcripts with the `--bambu_singleexon` option. Alternatively, users can choose to use the NDR threshold recommended by Bambu by enabling the `--bambu_rec_ndr` option, which replicates Bambu's default behaviour.

For StringTie2, ANNEXA uses the long reads assembly mode (`-L`) for the reconstruction step on each individual sample and the `-merge` option to produce a unified transcript set.

Gene and transcript quantification is performed internally by both tools. Bambu directly produces two raw count table for gene and transcripts, while for Stringtie the normalised counts are converted to raw counts using the `extractGeneExpression` function from the IsoformSwitchAnalyzeR program (Vitting-Seerup and Sandelin, 2019) and all the expression tables are made available by default in the final output of ANNEXA.

At the end of this module, ANNEXA integrates all genes and transcripts from the reference annotation along with the novel transcript models from StringTie2 or Bambu into an unfiltered extended annotation.

Coding potential and transcript classification

Among the newly assembled transcripts, it is essential to annotate different RNA categories, particularly distinguishing protein-coding transcripts (mRNAs) from long non-coding RNAs (lncRNAs). To achieve this, ANNEXA integrates the FEELnc program (Wucher et al., 2017) to predict the coding potential of all novel transcripts from novel genes (yellow line, **Fig. 1.A**). To accelerate

the FEELnc process, we previously demonstrated that training it on a subset of known mRNAs and lncRNAs from the reference annotation preserves high predictive performance (Wucher et al., 2017), thereby allowing ANNEXA to use FEELnc with a, by default, random subset of 3,000 reference lncRNAs and mRNAs. However, in the instance where the user's reference annotation lacks lncRNAs or contains less than 3,000 transcripts, we have implemented the `--feelnc_mrna` and `--feelnc_lncrna` options which allow them to define a number of transcripts to use.
To assess the potential impact of genomic alterations (mutations) on newly identified protein-coding transcripts, ANNEXA employs TransDecoder (https://github.com/TransDecoder/TransDecoder) using the `TransDecoder.LongOrfs` module to extract longest ORFs and `TransDecoder.Predict` with the `-single_best_only` option to predict coding regions, ensuring that only the single best ORF per transcript is retained while reporting the corresponding CDS (Coding Determining Sequence) information in the final GTF file. FEELnc is applied to all novel transcripts, both from novel and known genes, while TransDecoder is applied to all novel transcripts classified as protein-coding by FEELnc. Additionally, to classify novel transcripts with respect to the input reference annotation, ANNEXA uses GffCompare (Pertea and Pertea, 2020), incorporating this information under the `class_code` attribute in the extended GTF. This enables users to efficiently extract specific transcript classes, such as `class k` corresponding to alternative isoforms extending known genes at the 5' or 3' ends or `class x`, corresponding to exonic antisense transcripts (often classified as antisense lncRNA biotype).

Transcript filtering and full-length assessment

While long-read RNA sequencing (LR-RNAseq) has significantly improved transcriptome reconstruction by capturing full-length isoforms, it still exhibits biases that lead to fragmented transcript models, particularly at the 5' end *i.e.* Transcription Start Sites (TSSs) (Tardaguila et al., 2018).
To refine the extended annotation file and exclude novel incomplete transcripts, ANNEXA employs two complementary filtering strategies: (i) the Novel Discovery Rate (NDR) cutoff, derived from the Bambu tool, which evaluates the confidence of novel transcript predictions and (ii) the TransforKmer cutoff which assesses the likelihood of novel transcript transcription start sites (TSSs) using a species-specific deep learning model. The TransforKmer model is pre-trained with DNABERT (Ji et al., 2021), a transformer-based architecture designed for genomic sequence analysis, and further fine-tuned on a classification task using labeled TSSs from the reference annotations (Karollus et al., 2024). By applying this cutoff, ANNEXA filters out novel transcripts with TSSs that are unlikely to be biologically valid, based on the model's predictions (**Supp. Fig. 1**).

TransforKmers performance was evaluated using annotated transcript TSSs as positive examples and two types of negative controls: shuffled sequences preserving nucleotide composition and fully random sequences. Model outputs show clear separation between true TSSs and negative controls (**Supp. Fig. 2.A**). In addition, TFK classification accuracy was stratified by transcript biotype and transcript support level (TSL) in human, revealing a correlation between TFK accuracy and well-supported transcripts (**Supp. Fig. 2.B**).
To adjust the stringency of the filtering process, users can choose between two modes: (i) the union of the two filters (filtering `operation = union`), which retains transcripts passing at least

one filter, or (ii) the intersection (filtering `operation = intersection`), which retains only transcripts passing both filters (**Supp. Fig. 3**). At the end of this module, ANNEXA outputs an additional more stringent filtered annotation. Because ANNEXA provides both a full and filtered annotations as output, researchers can flexibly tailor their analysis, prioritising either sensitivity or precision, based on their specific research goals and biological questions.

Quality Control

The two resulting annotation files (*full* and *filtered*) may contain thousands of novel isoforms, including both mRNAs and lncRNAs, which require thorough inspection for final quality control (QC). Inspired by the SQANTI tool (Pardo-Palacios et al., 2024a), ANNEXA computes multiple features by comparing known (*i.e.*, matching the reference annotation) and novel (*i.e.*, reconstructed by Bambu or StringTie) genes, transcripts, and exons, collectively referred to as annotated elements (AE). The QC metrics can be broadly categorised into two groups:

- Structural metrics of AEs: including the number of known and novel AEs, AE length distribution, proportion of single- versus multi-isoform genes (at the gene level), and proportion of single- versus multi-exonic transcripts (at the transcript level).
- Quantification-related metrics: such as the distribution of gene read counts across input samples or the breadth of expression of the AEs.

Additionally, an optional QC feature assesses gene body coverage by sample reads using the RSeQC pipeline (Wang et al., 2012), allowing the identification of potential 5' or 3' read-coverage biases across known and novel genes.
All QC indicators are generated as CSV files, which serve as input for ANNEXA to compile and visualise the data in a comprehensive final QC report. This report contains information at three AE levels (gene, transcript and exon) and summarises key metrics, such as the number of novel elements, their length distribution, and expression counts as illustrated in **Fig. 2**.

ANNEXA implementation

ANNEXA is implemented in Nextflow (Di Tommaso et al., 2017) and integrates scripts written in multiple languages, including R, Python and Bash. The Nextflow framework is designed to facilitate the development of reproducible, scalable, and portable analysis workflows. ANNEXA can use both Docker and Singularity containers and has been successfully executed on standalone computers and high-performance computing clusters using SGE or SLURM. The project is open-source, with its code accessible on GitHub : https://github.com/IGDRion/ANNEXA/.

Long-Read Sequencing protocol of cancer cell lines

Experimental methods. Ten LR-RNAseq experiments were conducted (**Fig. 1.B**) with eight canine cancer cell lines from three canine cancer types: Mucosal Melanoma (MM, n=4), Histiocytic Sarcoma (HS, n=3) and Osteosarcoma (OS, n=1) and two human cancer cell lines also originating from MM patients. For all samples, RNA was extracted from ∼15 million cells using the NucleoSpin RNA kit (Macherey-Nagel). Library preparation was performed according to manufacturer's protocol (Oxford Nanopore Technologies, ONT) with the direct cDNA Sequencing Kit (SQK-DCS109). Sequencing was done using MinION Flow Cells (FLO-MIN106D) with GridION device from the IGDRION platform (https://igdr.univ-rennes.fr/igdrion).

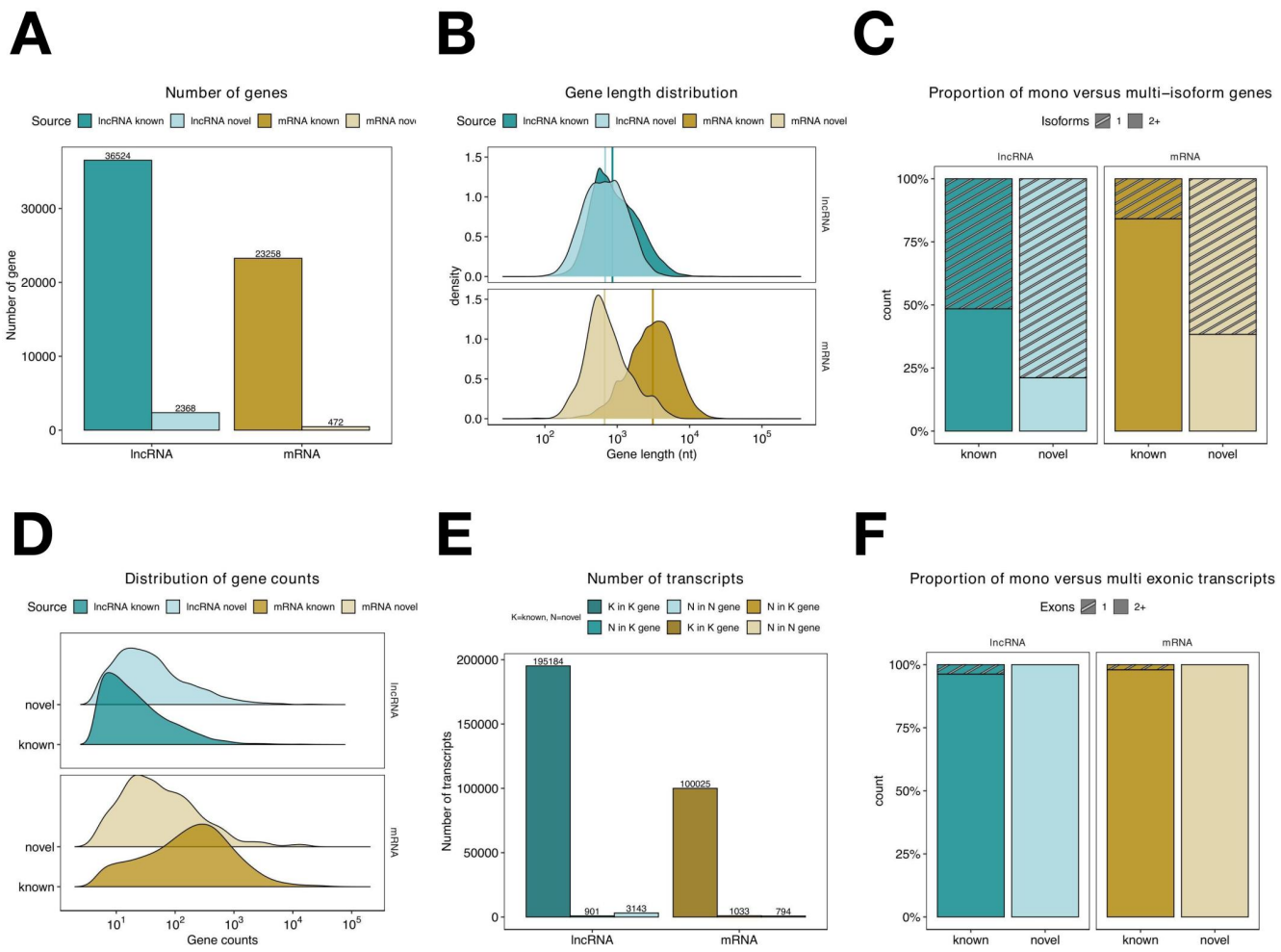


Figure 2 – Example of an automatic QC report from ANNEXA. This report was generated from the analysis of two human mucosal melanoma samples using Bambu with the Gencode annotation (the full PDF can be found in the GitHub repository of ANNEXA). **A**- Number of known and novel genes. **B**- Gene length distribution. **C**- Proportion of gene mono- versus multi-isoform(s). **D**- Distribution of gene read counts. **E**- Number of transcripts in known (K) and novel (N) genes. **F**- Proportion of transcript mono/single- versus multi-exonic.

Computational methods. Basecalling of `fast5` files was done with guppy (version 6.0.0) and the nf-core/nanoseq pipeline (version 3.1.0) (Chen et al., 2025) from the nf-core community (Ewels et al., 2020) was used to do all primary bioinformatic analyses. Briefly, this included the quality control (QC) of the reads with the nanoplot (version 1.41.6) (De Coster and Rademakers, 2023) and multiqc (version 1.9) (Ewels et al., 2016) programs and the mapping of the `fastq` files onto human and canine genomes using minimap2 software (version 2.15-r905) (Li, 2018). For the ten samples, we considered all reads having a Qscore >5 (**Table 1**). Intersection between all genomic features (i.e. TSS and CAGE data) was done using bedtools (version 2.27.1).

Benchmark analyses

Reference annotations and genomes. In the LRGASP study, only one reference annotation (Gencode) was used to benchmark tools in human LR-RNASeq samples. In our study, we benchmarked ANNEXA using several human and dog annotations in order to test the influence of the completeness of the input reference annotation in two species with high and medium quality

Sample	Cellosaurus ID	Total Reads	Median Length	Primary Mapped	% Primary Mapped
CFA-MM1	CVCL_0D14	4,002,891	806	3,408,273	85.15
CFA-MM2	CVCL_IZ11	8,843,341	769	6,253,908	70.72
CFA-MM3	Dog-OralMel-18395	7,108,351	943	5,901,302	83.02
CFA-MM4	Dog-OralMel-18333	6,877,092	896	5,696,283	82.83
CFA-HS1	Dog-HS-5472-lung	11,177,946	826	9,157,969	81.93
CFA-HS2	Dog-HS-5472-spleen	4,055,709	878	3,355,138	82.73
CFA-HS3	Dog-HS-7051	3,908,392	793	3,040,788	77.80
CFA-OS1	Dog-OSTEO-10188	6,816,094	1,050	5,707,582	83.74
HSA-MM1	CVCL_1282	5,457,058	886	4,865,343	89.16
HSA-MM2	CVCL_6797	4,734,796	852	4,025,056	85.01

Table 1 – Number of reads sequenced and aligned on canFam4 (CFA) and GRCh38 (HSA).

reference annotations. We compiled reference genomes and annotations downloaded from Ensembl (Dyer et al., 2025), Refseq (O'Leary et al., 2016), Gencode (Mudge et al., 2025) or original papers (Varabyou et al., 2023; Wang et al., 2021) (**Table 2**).

Species	Assembly	Reference Annotation (version)	Genes	Protein coding	lncRNA	Transcripts	Protein coding	lncRNA	Link
Dog	canFam4	Ensembl (113)	30,653	20,533	6,446	56,403	44,438	8,291	https://ftp.ensembl.org/pub/release-113/gtf/canis_lupus_familiarisgsd/
Dog	canFam4	UU (University Uppsala)	29,535	20,072	9,177	158,561	143,888	13,476	http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=canFam4
Dog	canFam4	Refseq	43,427	21,241	13,404	103,619	68,562	25,558	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/011/100/685/
Human	GRCh38	Gencode (47)	86,402	23,258	36,524	412,034	100,025	195,184	https://www.gencodegenes.org/human/release_47.html
Human	GRCh38	CHESS (3.0)	63,755	22,083	18,598	168,451	105,328	36,354	https://github.com/chess-genome/chess

Table 2 – Species, assembly, reference annotation with gene and transcript counts and annotation links.

The Gencode v47 annotation leverages the Capture Long-Read Sequencing (CLS) pipeline, combining Nanopore and PacBio sequencing to identify over 132,000 novel human lncRNAs, resulting in higher transcript counts compared to other annotations (Kaur et al., 2024). The CHESS annotation, by contrast, builds on Human RefSeq and incorporates RNA-seq data from the GTEx project, offering broad tissue representation but without RNA enrichment followed by long-read sequencing. For the canine genome, the annotation provided by the University of Uppsala (UU) integrates Nanopore and PacBio data from 40 canine tissues, providing a comprehensive long-read-based resource (Wang et al., 2021). Unlike UU, Ensembl and RefSeq annotations lack a standardised LR-RNASeq transcriptome dataset across canine assemblies, leading to variability in gene and transcript counts, as recently reviewed (Hoffmann et al., 2026a).

Transcriptome reconstruction benchmarking. ANNEXA was run using Bambu and StringTie on the eight canine samples (bam files) with the canFam4 genome assembly and the three reference annotations (Ensembl, RefSeq and UU), and on the two human samples with the GRCh38 assembly and CHESS or Gencode annotations. For the Bambu runs, single-exon transcripts were discarded because their inclusion interfered with the calculation of the Novelty Detection Rate (NDR) recommended by the tool, leading to unstable filtering behavior. In contrast, StringTie retains single-exon transcripts by default when they meet the minimum coverage threshold. The full and filtered extended annotations produced after the runs were filtered to only keep genes and transcripts expressed (raw count >0) in at least one sample and were compared to the initial reference annotations using the GffCompare tool version v0.12.6 (Pertea and Pertea, 2020) which provides sensitivity and precision metrics at the exon, transcript and locus/gene levels.

True Positives (TP), False Positives (FP) and False Negatives (FN) were defined as followed:

- TP: Reconstructed features (exons, transcripts, or loci) that match corresponding features in the reference annotation.
- FP: Reconstructed features that are absent from the reference annotation. These may represent novel transcripts or potential artifacts.
- FN: Features present in the reference annotation but missing in the extended annotations, indicating transcripts that were not recovered.

From these classifications, we thus computed Precision as TP /(TP + FP) and Recall or Sensitivity as TP/(TP+FN).

Orthology analysis. Canine and human novel genes and transcripts were mapped on target genomes (GRCh38 and CanFam4, respectively) using the liftoff program (Shumate and Salzberg, 2021) with default parameters. Then, query elements were classified into three classes with respect to the target reference annotation (Gencode and Ensembl, respectively): unmapped (query gene not mapped to target genome), mapped unknownGenes (query gene mapping to intergenic regions) and mapped knownGenes (query gene mapping by 50% coverage and 50% sequence identity to known genes from target annotation *e.g.* Gencode for query dog gene and Ensembl for human query gene).

Results

We produced LR-RNASeq data from both human (n=2) and canine (n=8) cancer cell lines (**Fig. 1.B**) and applied ANNEXA using different genome assemblies and reference annotations for each species in order to study the robustness of the pipeline in its ability to extend and to quality control these annotations.

Effect of annotation sources for the identification of known and novel genes/transcripts

Analysis of known and novel genomic features (genes and transcripts) reconstructed by Bambu and StringTie (with their default parameters) revealed striking differences with respect to the tool used and the completeness of the input reference annotation (**Fig. 3**).

Known genes and transcripts. For the canine genome (canFam4), Bambu reconstructed more known elements than StringTie, consistently across all three reference annotations tested (Ensembl, Refseq, University of Uppsala (UU)), with this difference being the most noticeable for the UU annotation (21,416 genes and 106,378 transcripts reconstructed with Bambu, 15,163 and 59,422 with StringTie) (**Fig. 3.A**). For the human genome (GRCh38), Bambu also identified more known elements than StringTie across the two annotations (CHESS and Gencode) with this difference being most pronounced for Gencode (27,341 genes and 87,249 transcripts reconstructed with Bambu compared to 18638 and 50,255 respectively with StringTie) (**Fig. 3.B**). Compared to dogs, the relatively low percentage of features reconstructed from a reference annotation in human (for example, only 21% of Gencode transcripts are reconstructed by Bambu) could be explained by the fact that human reference annotations are more exhaustive than in dogs and also because our experimental design contained more LR-RNAseq samples in dogs (n=8) versus human (n=2).

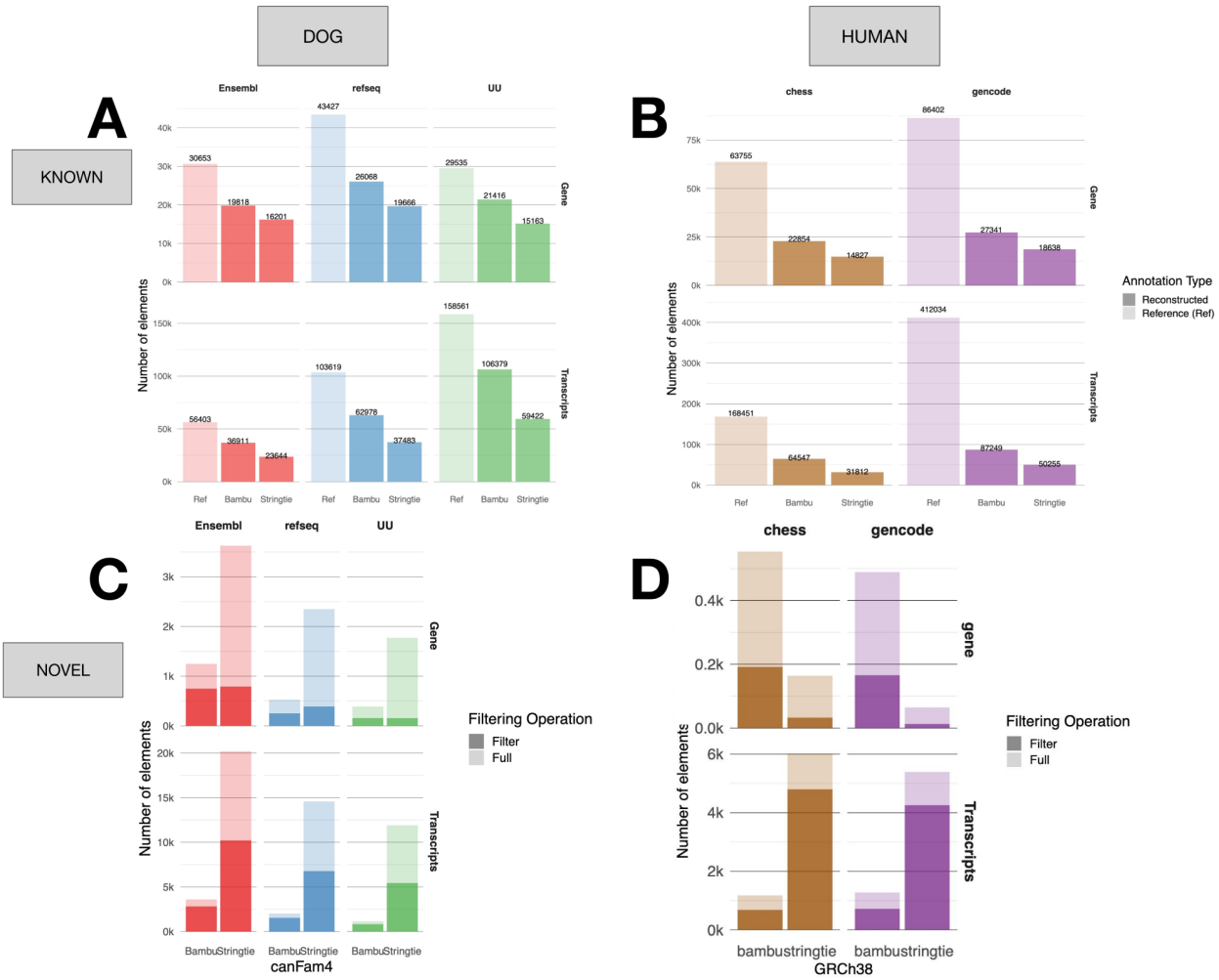


Figure 3 – Benchmark analysis between Bambu and StringTie for dog (left panels) and human (right panels) for known (top panel) and novel (bottom panel) genes (upper part of each panel) and transcripts (bottom part). For known elements (top), the number of genes/transcripts from the reference annotation (transparent bars) is separated from reconstructed genes/transcripts by Bambu and StringTie (solid bars). **A**- Number of known canine genes/transcripts retrieved from three reference annotations (Ensembl - red, refseq - blue and UnivUppsala - green). **B**- Number of known human genes/transcripts retrieved from two reference annotations (Chess - brown and Gencode - purple). **C**- Number of novel dog genes/transcripts. Full set of elements (light bar) are separated from ANNEXA's filtered (intersection) set (dark bar). **D**- Number of novel human genes/transcripts detected on canonical chromosomes (chr1-22, X, Y and M). Full set of elements (light bar) are separated from ANNEXA's filtered (intersection) set (dark bar).

Novel genes and transcripts. For the canine genome (canFam4), StringTie consistently identified more novel elements than Bambu, independently of the three reference annotations tested (Ensembl, RefSeq and UU). However, we observed that this difference is most pronounced using the Ensembl annotation, where Bambu detected 3,134 novel genes and 3,600 novel transcripts compared to the StringTie set of genes and transcripts (12,671 and 20,184, respectively). Using the UU canine annotation yielded the lowest number of novel genes and transcripts identified by both StringTie (7,791 and 11,900) and Bambu (1,070 and 1,150), likely due to the greater initial isoform completeness of the UU reference annotation compared to Ensembl and RefSeq, thereby reducing the number of newly detected transcripts (**Fig. 3.C**). In the human genome

(GRCh38), using the Gencode annotation as reference produced less novel elements than with CHESS with both Bambu and StringTie, as expected given the higher number of genes and transcripts catalogued in the latest version of Gencode (**Fig. 3.D**). Interestingly, we observe a difference in the novel gene reconstruction pattern between canine and human annotations, where StringTie reconstructs more genes in the three canine annotations and Bambu more in the two human annotations, despite StringTie consistently reconstructing more novel transcripts across all five annotations. This suggests that in a well annotated genome such as GRCh38, Bambu can still identify novel genes while StringTie primarily reconstructs novel isoforms of already known genes.

Regarding ANNEXA filtering operations, which evaluated the full-lengthiness of reconstructed genomic elements (see **Methods**), we observed that it dramatically reduced the number of novel elements identified by both tools, but the decrease is more pronounced for StringTie. On average in dogs, 50% of Bambu and 16% of StringTie novel genes are being conserved after ANNEXA's check for TSS validity, while in Humans 65% of Bambu genes and 46% of StringTie genes are conserved in CHESS, and 35% and 25% with Gencode. The proportion of filtered StringTie transcripts is more pronounced with Ensembl (dog) and CHESS (human) annotations, suggesting that these new transcripts may often be fragmented. Together, these findings highlight that reference annotation selection significantly impacts novel element discovery, with StringTie demonstrating particular sensitivity to this choice, especially in less well-annotated genomes such as canFam4. In addition, this shows that Bambu consistently demonstrated stronger performance in terms of precision and recall, particularly after applying ANNEXA's filtering steps for gene and transcript full-lengthiness (**Supp. Fig. 4**)

Effect of species-specific reference transcriptome for the identification of known mRNAs and lncRNAs

Compared to protein-coding genes (mRNAs), long non-coding RNAs (lncRNAs) are considered particularly challenging to annotate due to their low expression levels and tissue specificity (Derrien et al., 2012; Kainth et al., 2023). As in our previous analyses, we evaluated the ability of Bambu and StringTie to reconstruct known genes and transcripts starting from long-read RNA sequencing (LR-RNASeq) data, this time distinguishing lncRNAs from mRNAs (**Fig. 4**).

In dogs (canFam4, Ensembl annotation), we observed that Bambu reconstructed a greater number of mRNAs and lncRNAs from the reference annotation than StringTie, with 13,239 vs. 9,930 genes for mRNAs and 3,314 vs. 2,525 for lncRNAs (**Fig. 4.A**). At the transcript level, we also noticed that the difference between Bambu and StringTie in term of reconstructed transcripts is more pronounced for mRNAs than for lncRNAs biotypes. A similar trend was observed in humans (GRCh38, Gencode annotation), where Bambu identified more genes and transcripts from the latest Gencode annotation than StringTie (**Fig. 4.B**). However, the proportion of reconstructed known genes was lower in humans than in dogs, again likely due to the smaller number of human LR-RNASeq data available as input.

Using ANNEXA's transcript classification module, we then categorised both known (`class_code =`) and novel (all other `class_code`) mRNAs and lncRNAs annotated by the two tools from the eight canine and two human LR-RNASeq samples. Interestingly, this analysis revealed distinct patterns in the classification of novel canine lncRNAs: Bambu predominantly assigned them to `class_code u` (intergenic lncRNAs or lincRNAs, 14%) and `class_code x` (antisense lncRNAs,

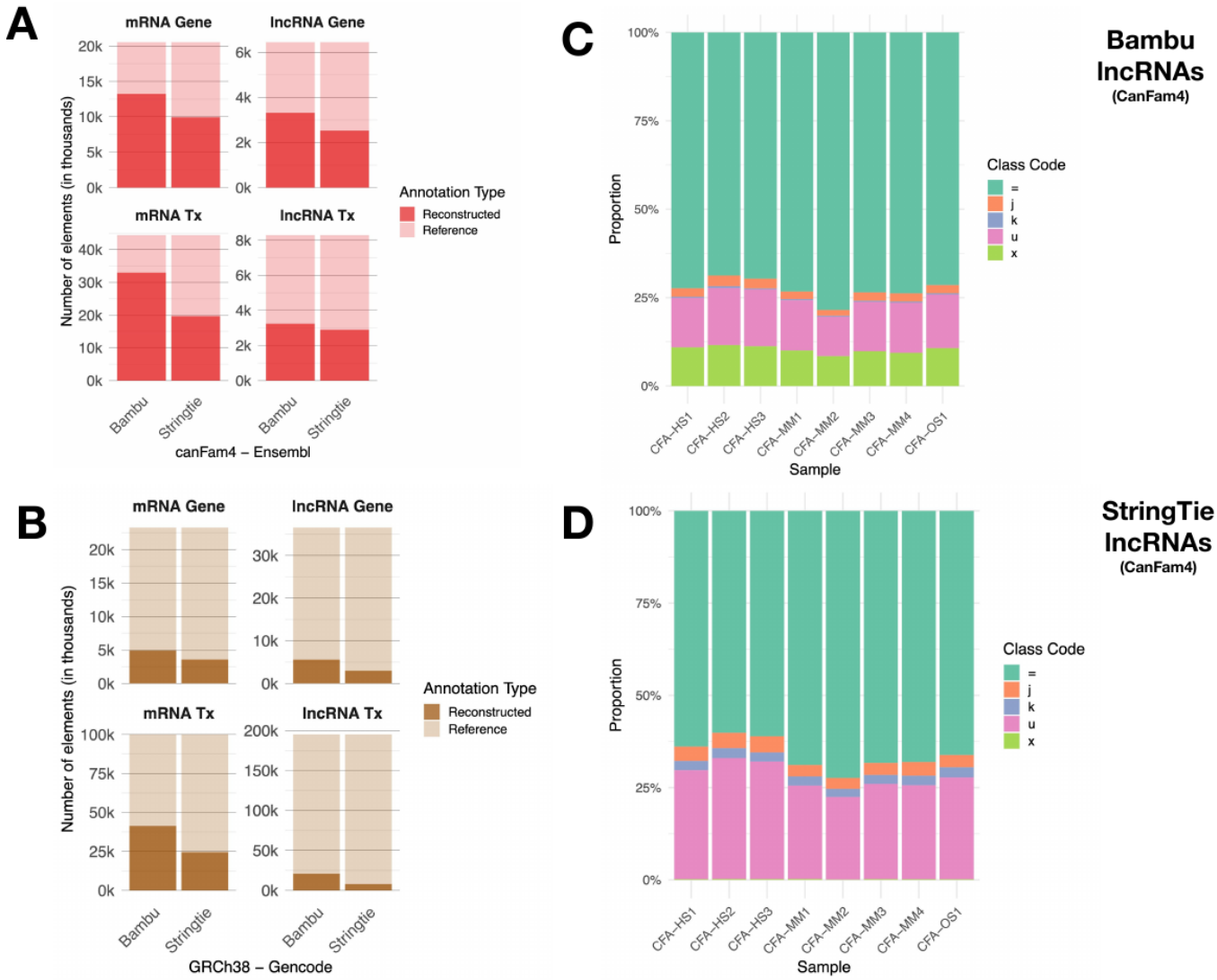


Figure 4 – Annotation of lncRNAs/mRNAs by Bambu and StringTie. **A**- Number of known canine genes (top) and transcripts (bottom) for mRNA (left) and lncRNAs (right) annotated by Ensembl. **B**- Number of known human genes (top) and transcripts (bottom) for mRNA (left) and lncRNAs (right) annotated by Gencode. **C**- Class code distribution of canine lncRNAs annotated by Bambu. **D**- Class code distribution of canine lncRNAs annotated by StringTie. For panel C and D, class codes = (green) : transcript isoforms matching the reference annotation, j (orange): novel spliced isoforms, k (purple): extension of reference gene, u (pink): intergenic transcripts and x (light green): antisense transcripts.

10%), whereas StringTie primarily annotated novel lncRNAs as lincRNA (27%) and extension of known lncRNAs (2.4%), with very few classified as antisense (n=4) in dogs (**Fig. 4.C** and **Fig. 4.D**). This discrepancy is further illustrated in **Supp. Fig. 5**, where novel antisense lncRNAs are consistently detected by Bambu in both species but are largely absent in StringTie's canine annotations, highlighting a tool-specific difference in antisense transcript reconstruction.

Biological application of ANNEXA

Over the past decade, canine models have gained recognition as valuable spontaneous and immunocompetent systems for studying human cancers, particularly histiocytic sarcomas (HS) (Hédan et al., 2020) and mucosal melanoma (MM) (Prouteau and André, 2019). For MM, although

rare in humans, it represents the most prevalent oral malignancy in dogs and exhibits notable clinical, biological, and genetic parallels with its human counterpart (Prouteau et al., 2022). Leveraging ANNEXA's ability to balance precision and recall, we applied it to both canine and human data. In dogs, using a relaxed Bambu NDR threshold (NDR = 1) allowed to identify 9,612 novel genes (8,713 lncRNAs and 899 mRNAs) across the eight long-read RNA-Seq datasets (see **Methods**). Applying ANNEXA's TSS validity filter significantly reduced the number of novel genes to 749 (595 lncRNAs and 154 mRNAs), indicating a likely high rate of false positives in the unfiltered set which could correspond to incomplete transcript models. Notably, the proportion of gene TSSs validated by orthogonal datasets, such as CAGE (Cap Analysis of Gene Expression) from the DogA consortium (Hörtenhuber et al., 2024), was significantly higher in the filtered set (52%, with 311 novel lncRNAs and 75 mRNAs validated) compared to the full dataset (9.3%, with 903 novel lncRNAs and 186 mRNAs validated). We also sought to assess whether novel genes in both humans and dogs could be conserved through evolution as a proxy for functional evidence (Ulitsky, 2016). We thus mapped the 9,614 novel canine genes (Ensembl - canFam4) on the human genome assembly used for the extended human annotation from Gencode and found 3,709 (38.6%) that could be mapped to GRCh38. Among these, 3,263 (88%) correspond to human genes from the Gencode reference annotation and thus represent novel orthologous relationships between novel dog and known human genes. Of these mapped genes, 2,901 (88.9%) were classified as lncRNAs, of which 931 matched with human lncRNA genes, while 362 (11.1%) were classified as protein coding and 240 of those were matching with human protein coding genes. Notably, we identified five novel canine lncRNA genes that also map to novel human genes (**Table 3**, **Supp. Fig. 6**), with two of them being classified as protein coding in humans. Although these genes were not supported by CAGE data, they exhibited relatively high expression levels in both species, with mean read counts of 90.3 in dog and 66.1 in human MM samples. These conserved and expressed genes across species provide novel candidates for future functional validation and underscore the value of cross-species annotation to uncover novel, potentially functional elements.

Dog (canFam4)								Human (GRCh38)							
Chrom	Start	End	Strand	Gene ID	Gene Biotype	Expression	FEELnc	Chrom	Start	End	Strand	Gene ID	Gene Biotype	Expression	FEELnc
5	32478938	32479297	-	BambuGene4459	lncRNA	48.25	0.086	10	80246557	80246921	+	BambuGene3464	lncRNA	23	0.13
13	38119835	38120449	+	BambuGene10705	lncRNA	7.625	0.05	KI270816.1 (8)	293424	294817	-	BambuGene7604	protein_coding	230.5	0.965
17	42888308	42888800	+	BambuGene12928	lncRNA	383.25	0.01	10	118932092	118950427	+	BambuGene3519	lncRNA	50.5	0.2492
33	24998438	25016350	+	BambuGene20393	lncRNA	9.5	0.036	10	127524536	127536860	-	BambuGene3611	lncRNA	16	0.284
X	92327013	92329767	+	BambuGene22331	lncRNA	3	0.0445	MU273394.1 (X)	67950	88065	+	BambuGene6983	lncRNA	39	0.206
X	92327013	92329767	+	BambuGene22331	lncRNA	3	0.0445	MU273394.1 (X)	69249	84497	-	BambuGene6987	protein_coding	10.5	0.932

Table 3 – Genomic coordinates, expression levels (raw counts) and FEELnc coding potential score of five novel orthologous genes identified in canine and human LR-RNASeq samples.

Discussion

In this work, we introduce ANNEXA, an all-in-one tool that not only performs transcriptome reconstruction but also simultaneously ensures quality control of extended annotations. Moreover, ANNEXA uniquely enables the comparative analysis of both novel mRNAs and lncRNAs derived from long-read RNA-seq data, thereby enhancing their functional profiling and refining our understanding of their potential biological relevance. The comparison of two main transcriptome discovery tools and quantification methods reveals distinct trends depending on the reference annotation used. Our results show that Bambu consistently reconstructed more known genes and transcripts than StringTie, regardless of the reference annotation or the species (dog

and human) analysed in this study. It also confirmed that the number of newly detected genomic elements is strongly influenced by the structure and completeness of the reference annotations.

The discrepancies between Bambu and StringTie could be attributed to their distinct methodological approaches. StringTie assembles transcripts by building splice graphs, without reference-based training, allowing it to retain a broader range of novel transcripts and include those who might deviate from known annotations. By contrast, Bambu uses a reference-guided machine learning algorithm to build a transcriptome, which prioritises precision by filtering out novel transcripts that are too dissimilar from the reference. This can explain why Bambu predicts fewer, higher-confidence transcripts, often limited to a single transcript per locus, while StringTie tends to predict multiple transcripts per locus, as observed in the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) consortium's study. It demonstrated that tools guided by a reference transcriptome, such as Bambu, perform optimally in well-annotated genomes and showed that Bambu predicted the fewest transcripts in de novo contexts, while StringTie predicted two or more transcripts for nearly half of the loci (Pardo-Palacios et al., 2024b).

The impact of reference annotation choice is also evident in the analysis of our human and canine long-read RNASeq data. As expected, using the latest Gencode release as reference produced fewer novel elements due to the higher number of genes and transcripts (86,402 and 412,034, respectively) already catalogued in its latest version. This underscores the importance of a comprehensive curated reference annotation to limit the detection of potentially artifact-driven elements.

One challenge in benchmarking transcriptome reconstruction tools lies in the classification of predicted transcripts. Our evaluation focused on the recovery of features present in reference annotations, and following GffCompare's documentation (Pertea and Pertea, 2020), we defined false negatives (FNs) as reference features absent from the extended annotations, and false positives (FPs) as novel features not present in the reference. In practice, some FPs may correspond to genuine, previously unannotated transcripts, while some FNs may represent transcripts not expressed in the analysed samples. Moreover, the exclusive use of reference annotations prevents the identification of true negatives, and thus precludes the computation of specificity. These limitations, which also apply for our TSS annotation with TransforKmers, highlight that performance estimates derived from reference-based benchmarking should be interpreted with caution, as they reflect both tool behaviour and the non-comprehensive nature of current annotations.

Using ANNEXA's quality control module for known and novel long non-coding RNAs, we also observed this consistent trend where Bambu reconstructed a greater number of non-coding genes and transcripts than StringTie, both in dogs (canFam4, Ensembl) and humans (GRCh38, Gencode). However, the proportion of reconstructed known genes was lower in humans than in dogs for both tools, likely due to the more limited availability of human long-read RNA sequencing data in our study design, combined with the higher completeness of human reference annotations (Kaur et al., 2024).

Additionally, the classification of novel lncRNAs revealed distinct tool-specific patterns between Bambu and StringTie. Bambu primarily categorised novel lncRNAs as intergenic (lincRNAs) or antisense, reflecting its ability to detect antisense transcripts, even in contexts where reference annotations are incomplete, such as in canine samples. In contrast, StringTie assigned a higher proportion of novel lncRNAs to lincRNAs and novel isoforms of known lncRNAs, with

only a few classified as antisense. This discrepancy is particularly pronounced in dog samples, where StringTie's reference-guided assembly may limit its ability to reconstruct antisense transcripts. Conversely, Bambu's approach appears less dependent on reference strand information, enabling consistent detection of antisense lncRNAs across species. These differences in classification outputs underscore the impact of transcriptome reconstruction tools and suggest that integrating multiple approaches may provide a more comprehensive representation of lncRNA diversity. Consequently, the choice of reconstruction tool should be guided by specific research objectives. For example, researchers focusing on natural antisense transcripts (NATs) in non-model organisms with incomplete reference annotation may benefit more from Bambu's capabilities, whereas those aiming to extend known lncRNA annotations might prefer StringTie.

Our analyses further emphasises the need to carefully evaluate novel annotations generated by transcriptome discovery tools. The increase in detected genes and transcripts should be considered within the context of annotation robustness and biological validation. The LRGASP Consortium recommends incorporating additional orthogonal data and replicate samples when aiming to detect rare and novel transcripts. In our comparative oncology project, we used CAGE data from both dog and human samples to validate the full-length nature of novel transcripts. We found that only less than 10% of unfiltered novel transcripts (both coding and noncoding) were validated by at least one CAGE signal. Several factors could explain this relatively low proportion of TSS validation, including the fact that the CAGE sample conditions did not match those of our LR-RNASeq cancer cell lines and/or the earlier version of the Nanopore kit for library preparation (DCS109) could have led to truncated reads and, consequently, incomplete transcripts (Sessegolo et al., 2019). However, the use of ANNEXA's module for TSS validation increased the proportion of validated genes to over 50%, highlighting the importance of combining complementary experimental and computational approaches to refine annotations.

In summary, this study highlights ANNEXA's modularity in adjusting precision and sensitivity for accurate, context-specific coding and non-coding annotation. This flexibility is essential for subsequent experimental validation and for advancing our understanding of the role of the non-coding genome in biological processes.

Availability and implementation

ANNEXA is written in Nextflow DSL2 (Di Tommaso et al., 2017). ANNEXA can run each process in conda environments as well as Docker or Apptainer containers, ensuring reproducibility and ease of use on different machines. The pipeline is available at `https://github.com/IGDRion/ANNEXA`. Input files and code to reproduce figures of this paper are also available here : `https://github.com/IGDRion/ANNEXA/tree/main/Paper_Figures`. The extended human and canine annotations from the results are available on Zenodo (DOI: 10.5281/zenodo.18195454).

Acknowledgements

The authors thank the GenOuest bioinformatics core facility (`https://www.genouest.org`) for providing them with the necessary storage and CPU/GPU resources to perform the analysis, the IGDRIon facility for long-read sequencing, members of the BIS group (`https://igdr.univ-rennes.fr/biologie-silico-0`) for useful discussions. Preprint version xxx[change to the correct number] of this article has been peer-reviewed and recommended by Peer Community

In XYZ[change to the name of the PCI] (https://doi.org/10.24072/pci.xxx[replace by the doi of the recommendation]; (**fake**)[replace by the citation of the recommendation]).

Fundings

This study was supported by IGDR, University of Rennes, Région Bretagne with financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, and funds from Inserm (C20013NS), Cancéropole Grand Ouest (CGO), Ligue Regionale contre le cancer du Grand Ouest and MSDAvenir foundation.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. TD is recommender of a PCI.

Data, script, code, and supplementary information availability

All fastq files of the 10 LR-RNASeq samples are available through the ENA database under study ID: PRJEB88593. Script and codes are available online (https://doi.org/10.5281/zenodo.15209930; Hoffmann et al., 2026b). Supplementary information is available online (https://www.biorxiv.org/content/10.1101/2025.04.16.648718v1; Hoffmann et al., 2025).

References

Chen Y, Davidson NM, Wan YK, Yao F, Su Y, Gamaarachchi H, Sim A, Patel H, Low HM, Hendra C, Wratten L, Hakkaart C, Sawyer C, Iakovleva V, Lee PL, Xin L, Ng HEV, Loo JM, Ong X, Ng HQA, et al. (2025). A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. *Nature Methods* **22**, 801–812. https://doi.org/10.1038/s41592-025-02623-4. URL: https://www.nature.com/articles/s41592-025-02623-4.

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J (2023). Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nature Methods* **20**. Number: 8, 1187–1195. https://doi.org/10.1038/s41592-023-01908-w. URL: https://www.nature.com/articles/s41592-023-01908-w.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13. https://doi.org/10.1186/s13059-016-0881-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/.

De Coster W, Rademakers R (2023). NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311. https://doi.org/10.1093/bioinformatics/btad311. URL: https://doi.org/10.1093/bioinformatics/btad311.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22**. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press, 1775–1789. https://doi.org/10.1101/gr.132159.111. URL: http://genome.cshlp.org/content/22/9/1775.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316–319. https://doi.org/10.1038/nbt.3820. URL: https://www.nature.com/articles/nbt.3820.

Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, Becker A, Bennett R, Beracochea M, Berry A, Bhai J, Bhurji SK, Boddu S, Branco Lins PR, Brooks L, Ramaraju SB, Campbell LI, Martinez MC, Charkhchi M, Cortes LA, et al. (2025). Ensembl 2025. *Nucleic Acids Research* **53**, D948–D957. https://doi.org/10.1093/nar/gkae1071. URL: https://doi.org/10.1093/nar/gkae1071.

Ewels P, Magnusson M, Lundin S, Käller M (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354. URL: https://doi.org/10.1093/bioinformatics/btw354.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**, 276–278. https://doi.org/10.1038/s41587-020-0439-x. URL: https://www.nature.com/articles/s41587-020-0439-x.

Farkas C, Recabal A, Mella A, Candia-Herrera D, Olivero MG, Haigh JJ, Tarifeño-Saldivia E, Caprile T (2022). annotate_my_genomes: an easy-to-use pipeline to improve genome annotation and uncover neglected genes by hybrid RNA sequencing. *GigaScience* **11**, giac099. https://doi.org/10.1093/gigascience/giac099. URL: https://doi.org/10.1093/gigascience/giac099.

Guizard S, Miedzinska K, Smith J, Smith J, Kuo RI, Davey M, Archibald A, Watson M (2023). nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. *Bioinformatics* **39**, btad150. https://doi.org/10.1093/bioinformatics/btad150. URL: https://doi.org/10.1093/bioinformatics/btad150.

Hédan B, Rault M, Abadie J, Ulvé R, Botherel N, Devauchelle P, Copie-Bergman C, Cadieu E, Parrens M, Alten J, Zalcman EL, Cario G, Damaj G, Mokhtari K, Le Loarer F, Coulomb-Lhermine A, Derrien T, Hitte C, Bachelot L, Breen M, et al. (2020). PTPN11 mutations in canine and human disseminated histiocytic sarcoma. *International Journal of Cancer* **147**, 1657–1665. https://doi.org/10.1002/ijc.32991. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.32991.

Hoffmann N, Besson A, Cadieu E, Lorthiois M, Bars VL, Houel A, Hitte C, André C, Hédan B, Derrien T (2025). Facilitating genome annotation using ANNEXA and long-read RNA sequencing. Pages: 2025.04.16.648718 Section: New Results. https://doi.org/10.1101/2025.04.16.648718. URL: https://www.biorxiv.org/content/10.1101/2025.04.16.648718v1.

Hoffmann N, Besson A, Hitte C, Plassais J, Derrien T (2026a). Long Noncoding RNAs in Dogs: Comparative Insights Across Species and Breeds. *Annual Review of Animal Biosciences* **14**, 85–106. https://doi.org/10.1146/annurev-animal-030424-073201. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-animal-030424-073201.

Hoffmann N, Derrien T, vlebars, Aurore-B, Lorthiois M (2026b). IGDRion/ANNEXA: v3.3.0. https://doi.org/10.5281/zenodo.18922629. URL: https://zenodo.org/records/18922629.

Hörtenhuber M, Hytönen MK, Mukarram AK, Arumilli M, Araujo CL, Quintero I, Syrjä P, Airas N, Kaukonen M, Kyöstilä K, Niskanen J, Jokinen TS, Mottaghitalab F, Takan I, Salokorpi N, Raman A, Stevens I, Iivanainen A, Yoshihara M, Gusev O, et al. (2024). The DoGA consortium

expression atlas of promoters and genes in 100 canine tissues. *Nature Communications* **15**, 9082. https://doi.org/10.1038/s41467-024-52798-1. URL: https://www.nature.com/articles/s41467-024-52798-1.

Ji Y, Zhou Z, Liu H, Davuluri RV (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120. https://doi.org/10.1093/bioinformatics/btab083. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11025658/.

Kainth AS, Haddad GA, Hall JM, Ruthenburg AJ (2023). Merging short and stranded long reads improves transcript assembly. *PLOS Computational Biology* **19**, e1011576. https://doi.org/10.1371/journal.pcbi.1011576. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011576.

Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J (2024). Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biology* **25**. Number: 1, 1–21. https://doi.org/10.1186/s13059-024-03221-x. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03221-x.

Kaur G, Perteghella T, Carbonell-Sala S, Gonzalez-Martinez J, Hunt T, Mądry T, Jungreis I, Arnan C, Lagarde J, Borsari B, Sisu C, Jiang Y, Bennett R, Berry A, Cerdán-Vélez D, Cochran K, Vara C, Davidson C, Donaldson S, Dursun C, et al. (2024). GENCODE: massively expanding the lncRNA catalog through capture long-read RNA sequencing. Pages: 2024.10.29.620654 Section: New Results. https://doi.org/10.1101/2024.10.29.620654. URL: https://www.biorxiv.org/content/10.1101/2024.10.29.620654v1.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**. Number: 1, 1–13. https://doi.org/10.1186/s13059-019-1910-1. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1.

Kurylo C, Guyomar C, Foissac S, Djebali S (2023). TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data. *NAR Genomics and Bioinformatics* **5**, lqad089. https://doi.org/10.1093/nargab/lqad089. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10578202/.

Li H (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191. URL: https://doi.org/10.1093/bioinformatics/bty191.

Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, Loveland JE, Arnan C, Barnes I, Bennett R, Berry A, Bignell A, Cerdán-Vélez D, Cochran K, Cortés LT, Davidson C, Donaldson S, Dursun C, Fatima R, Hardy M, et al. (2025). GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research* **53**, D966–D975. https://doi.org/10.1093/nar/gkae1078.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, et al. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)* **376**, 44–53. https://doi.org/10.1126/science.abj6987.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V,

Choi J, Cox E, Ermolaeva O, Farrell CM, et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745. https://doi.org/10.1093/nar/gkv1189. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702849/.

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, Tseng E, Conesa A (2024a). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nature Methods* **21**, 793–797. https://doi.org/10.1038/s41592-024-02229-2. URL: https://www.nature.com/articles/s41592-024-02229-2.

Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G, Behera AK, Gonzalez Martinez JM, Hunt T, Lagarde J, Liang CE, Li H, Meade MJ, Moraga Amador DA, Prjibelski AD, Birol I, et al. (2024b). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nature Methods* **21**, 1349–1363. https://doi.org/10.1038/s41592-024-02298-3. URL: https://www.nature.com/articles/s41592-024-02298-3.

Pertea G, Pertea M (2020). GFF Utilities: GffRead and GffCompare. https://doi.org/10.12688/f1000research.23297.1. URL: https://f1000research.com/articles/9-304.

Prouteau A, André C (2019). Canine Melanomas as Models for Human Melanomas: Clinical, Histological, and Genetic Comparison. *Genes* **10**. Number: 7, 501. https://doi.org/10.3390/genes10070501. URL: https://www.mdpi.com/2073-4425/10/7/501.

Prouteau A, Mottier S, Primot A, Cadieu E, Bachelot L, Botherel N, Cabillic F, Houel A, Cornevin L, Kergal C, Corre S, Abadie J, Hitte C, Gilot D, Lindblad-Toh K, André C, Derrien T, Hedan B (2022). Canine Oral Melanoma Genomic and Transcriptomic Study Defines Two Molecular Subgroups with Different Therapeutical Targets. *Cancers* **14**. Number: 2, 276. https://doi.org/10.3390/cancers14020276. URL: https://www.mdpi.com/2072-6694/14/2/276.

Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, Lacroix V, Aury JM (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific Reports* **9**, 14908. https://doi.org/10.1038/s41598-019-51470-9. URL: https://www.nature.com/articles/s41598-019-51470-9.

Shumate A, Salzberg SL (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016. URL: https://doi.org/10.1093/bioinformatics/btaa1016.

Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, Harrow J, Bertone P (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**, 1177–1184. https://doi.org/10.1038/nmeth.2714. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3851240/.

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, Edelmann M, Ezkurdia I, Vazquez J, Tress M, Mortazavi A, Martens L, Rodriguez-Navarro S, Moreno-Manzano V, Conesa A (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research* **28**, 396–411. https://doi.org/10.1101/gr.222976.117.

Ulitsky I (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature Reviews Genetics* **17**, 601–614. https://doi.org/10.1038/nrg.2016.85. URL: https://www.nature.com/articles/nrg.2016.85.

Varabyou A, Sommer MJ, Erdogdu B, Shinder I, Minkin I, Chao KH, Park S, Heinz J, Pockrandt C, Shumate A, Rincon N, Puiu D, Steinegger M, Salzberg SL, Pertea M (2023). CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *Genome Biology* **24**, 249. https://doi.org/10.1186/s13059-023-03088-4. URL: https://doi.org/10.1186/s13059-023-03088-4.

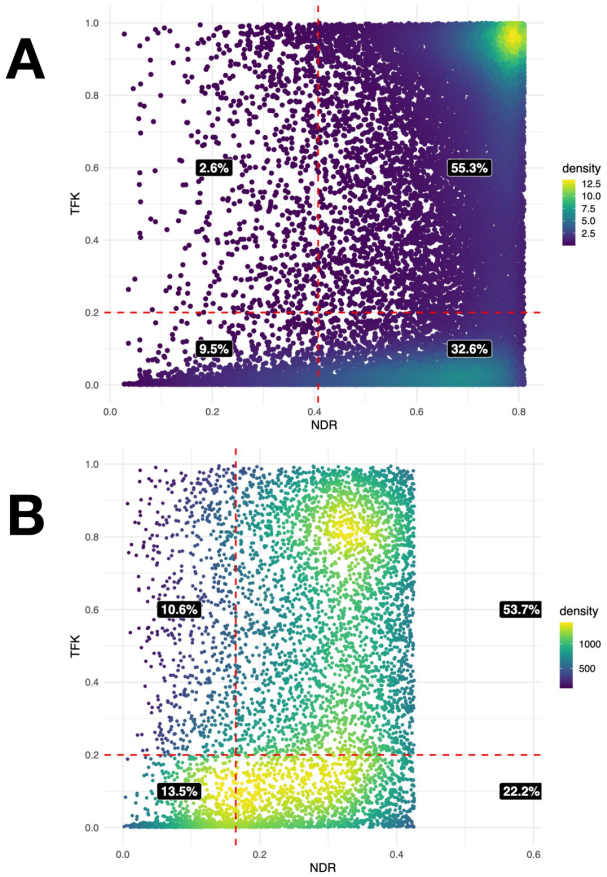
Vitting-Seerup K, Sandelin A (2019). IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics* **35**, 4469–4471. https://doi.org/10.1093/bioinformatics/btz247. URL: https://doi.org/10.1093/bioinformatics/btz247.

Wang C, Wallerman O, Arendt ML, Sundström E, Karlsson Å, Nordin J, Mäkeläinen S, Pielberg GR, Hanson J, Ohlsson Å, Saellström S, Rönnberg H, Ljungvall I, Häggström J, Bergström TF, Hedhammar Å, Meadows JRS, Lindblad-Toh K (2021). A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Communications Biology* **4**, 1–11. https://doi.org/10.1038/s42003-021-01698-x. URL: https://www.nature.com/articles/s42003-021-01698-x.

Wang L, Wang S, Li W (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185. https://doi.org/10.1093/bioinformatics/bts356. URL: https://doi.org/10.1093/bioinformatics/bts356.

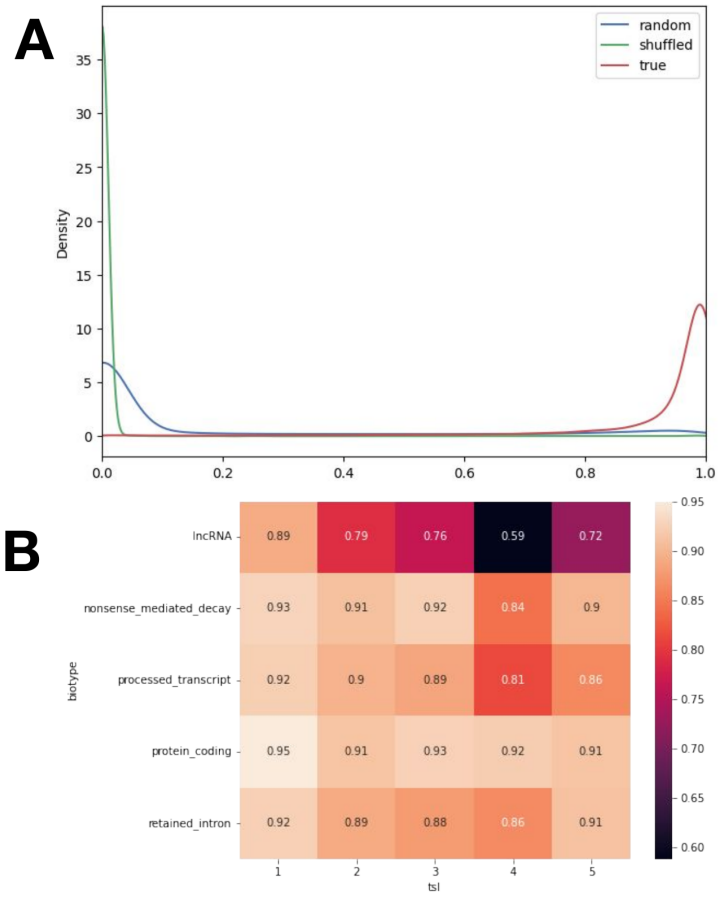
Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H, Cirera S, Fredholm M, Botherel N, Leegwater PA, Le Béguec C, Fieten H, Johnson J, Alföldi J, André C, Lindblad-Toh K, et al. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research* **45**, e57. https://doi.org/10.1093/nar/gkw1306. URL: https://doi.org/10.1093/nar/gkw1306.

Appendix A. Supplementary Figure 1



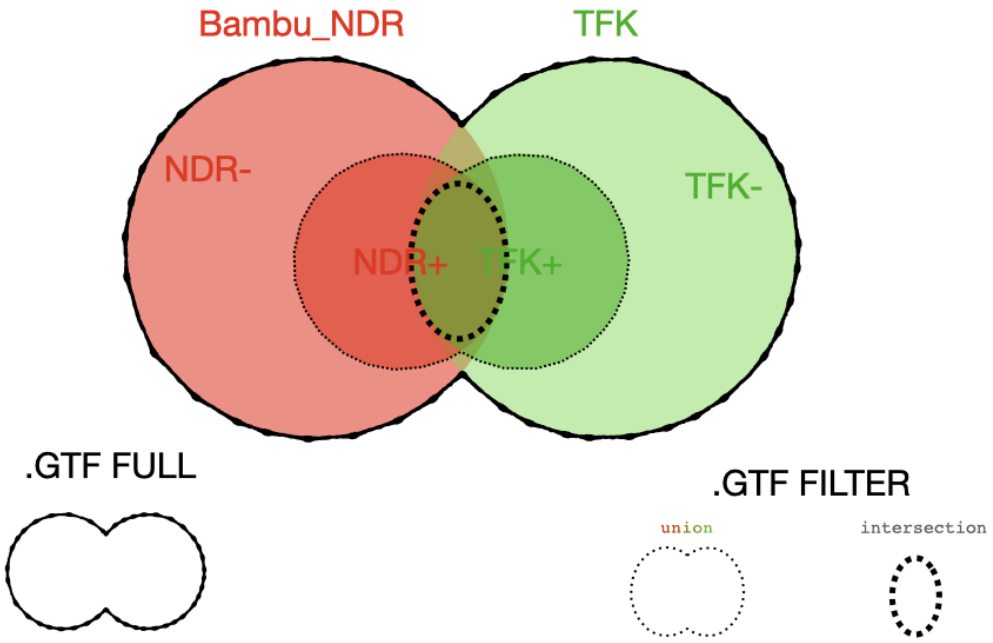
Supplementary Figures 1 – Distribution of novel transcripts identified by Bambu according to their TransforKmer score (TFK) and Novel Discovery Rate score (NDR) in canine samples with the Ensembl annotation (**A**) and human samples with the Gencode annotation (**B**). The red dotted lines represent the cutoffs used by ANNEXA to filter novel transcripts.

Appendix B. Supplementary Figure 2



Supplementary Figures 2 – **(A)** Density plot of TransforKmers (TFK) prediction probabilities stratified by TSS sequence origins: random sequences (blue), shuffled (green) and true/reference TSSs (red). **(B)** Heatmap of TFK classification accuracy in human relative to transcript biotypes and transcript support level (tsl).

Appendix C. Supplementary Figure 3

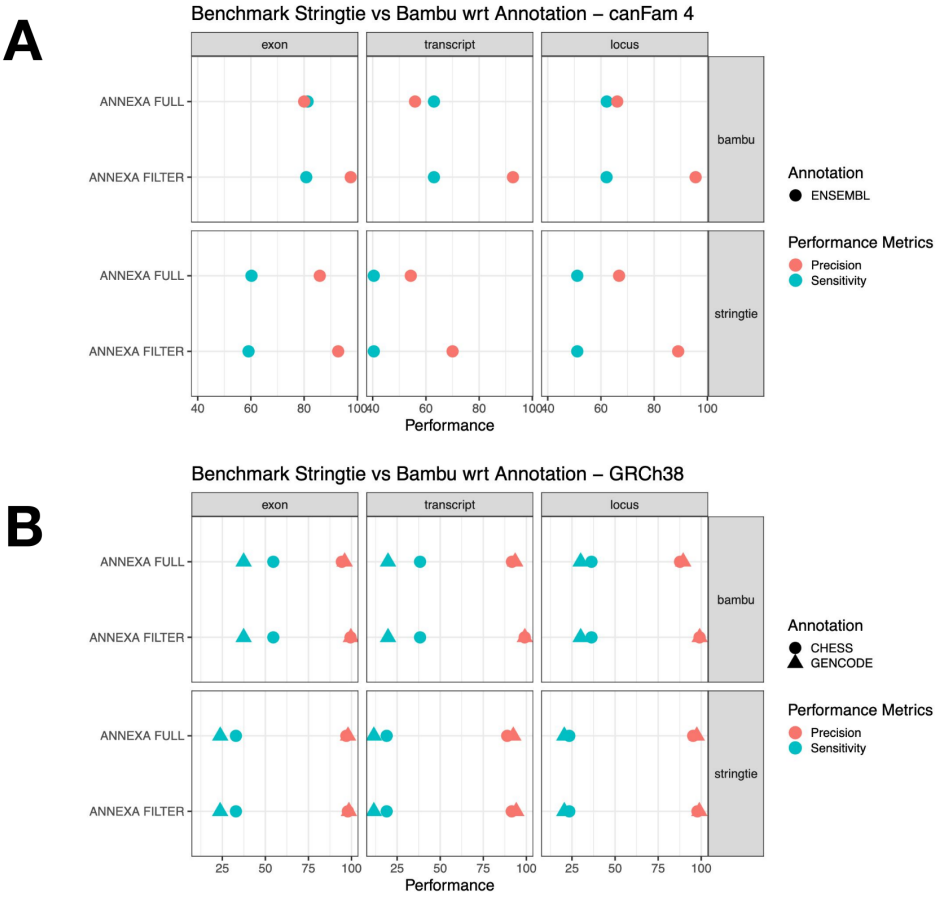



```
--filter              [boolean] Perform or not the filtering step. [default: true]

--operation           [string]  Operation to retained novel transcripts. 'union' retain tx validated by either
bambu or transforkmers, 'intersection' retain tx validated by both. (accepted: union, intersection) [default:
intersection]
```

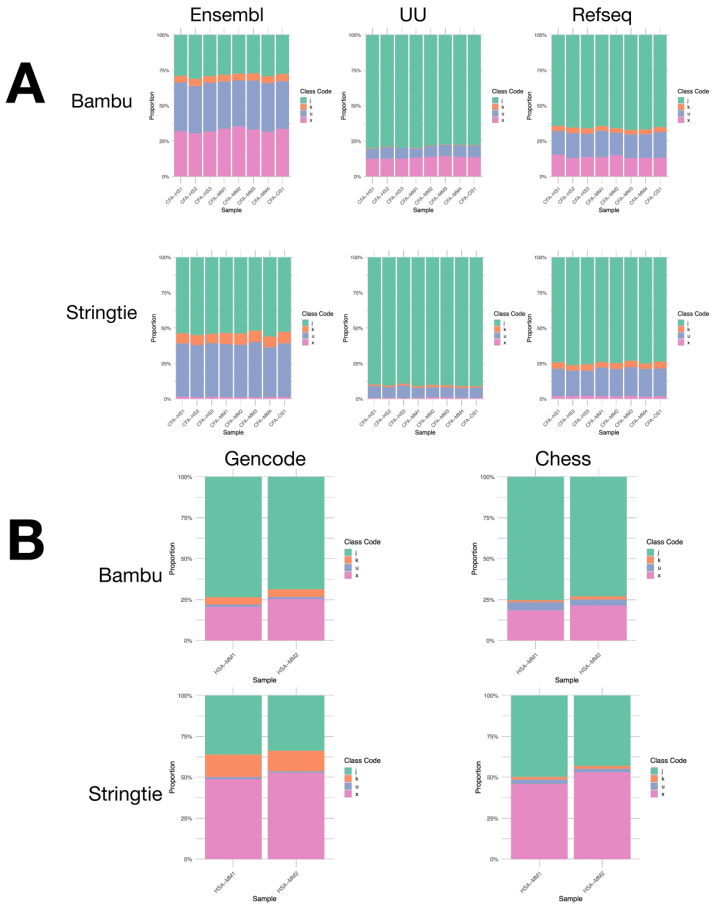
Supplementary Figures 3 – Filtering operations in ANNEXA. Novel transcripts can be filtered based on their NDR cutoff (in red, exclusive to Bambu) and/or TransforKmers (TFK) cutoff (in green, available with both Bambu and StringTie).

Appendix D. Supplementary Figure 4



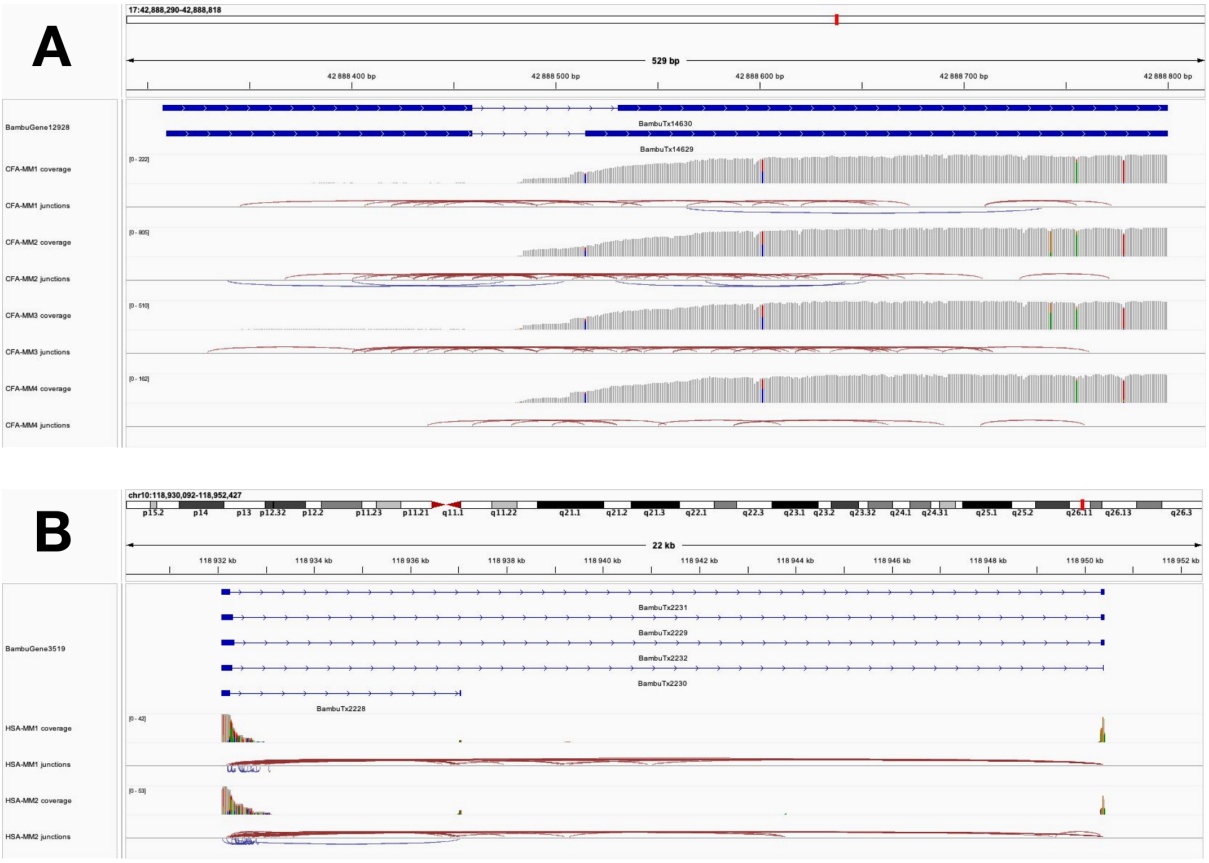
Supplementary Figures 4 – Comparison of precision (red) and sensitivity (green) metrics when using ANNEXA with Bambu or Stringtie on the canine (A) or human (B) LR-RNAseq data.

Appendix E. Supplementary Figure 5



Supplementary Figures 5 – Classification of novel transcripts in dogs (**A**) and humans (**B**) with j = novel alternatively spliced transcript (green), k= extension of known transcript (orange), u = intergenic transcript (blue) and x = antisense transcript (pink).

Appendix F. Supplementary Figure 6



Supplementary Figures 6 – IGV screenshot of a novel orthologous gene both annotated in dog (A) and human (B). Long reads aligned from four mucosal melanoma (MM) samples in dogs (canFam4) and two MM samples in human (GRCh38) are represent below annotations.