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], Christophe Hitte[1], and Thomas Derrien[1]

https://doi.org/10.5802/fake.doi

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 an updated version of 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 visualizations 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 plateform, 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), 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 characterization 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 noncoding 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 transcript fragmentation and incomplete isoform representation (Sessegolo et al., 2019). To ensure accurate genome annotation, robust quality control (QC) 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 LR-GASP 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 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.

To address these limitations, we present ANNEXA, a novel pipeline designed to extend reference annotation based on LR-RNASeq data and assess the quality of novel model transcripts. 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 also incorporates a deep learning strategy to potentially remove incomplete 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.

Methods

Overview 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 mapping files (**Fig.1.A**). ANNEXA, which can support both Nanopore and PacBio RNA sequencing data, contains four main modules described below:

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

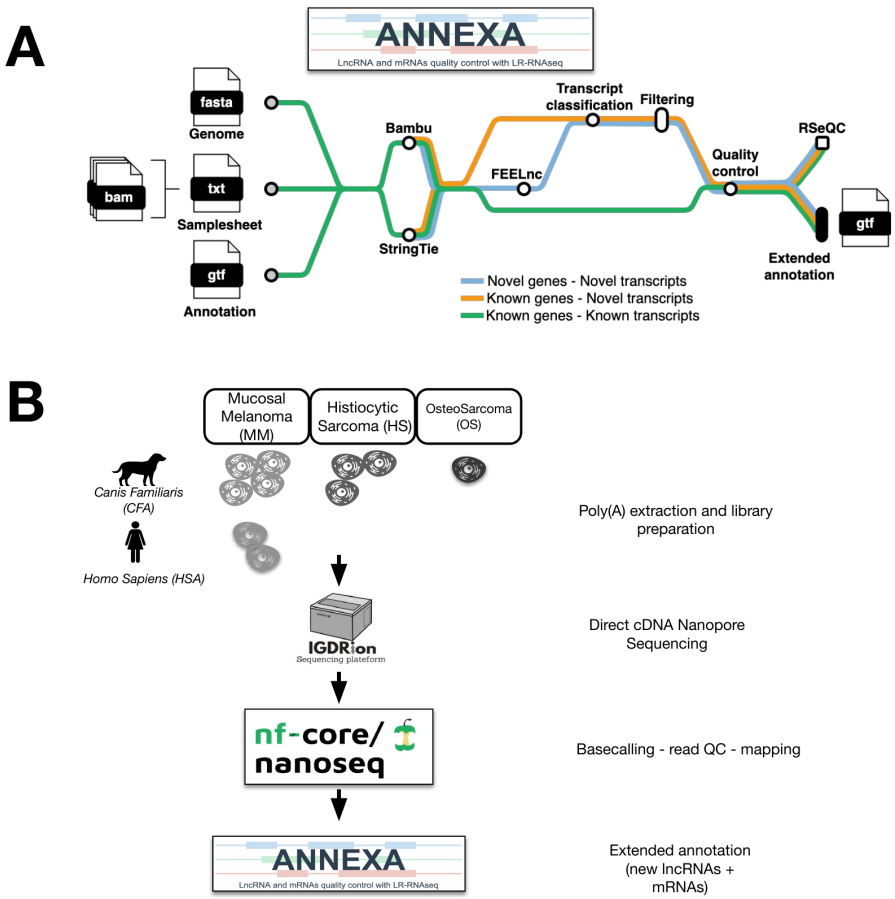


Figure 1 – **A**- Metromap of ANNEXA. **B**- Experimental design with species-specific input cancer cell lines and experimental and computational analyses.

Transcriptome reconstruction

As recently demonstrated by the LR-GASP 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). In ANNEXA, several options are available for Bambu, including the ability to adjust the Novel Discovery Rate (NDR) threshold, a key metric that balances sensitivity and precision (Chen et al., 2023). For instance, users can enhance sensitivity by increasing the default recommended NDR threshold with the `--bambu_threshold` option, and/or 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, while gene and transcript quantification (raw counts) are extracted using the `extractGeneExpression` function from the IsoformSwitchAnalyzeR program (Vitting-Seerup and Sandelin, 2019).
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 annotation, referred to as *Extended_annotations.full.gtf*.

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 (blue line, **Fig. 1.A**). To accelerate the FEELnc process, we previously demonstrated that training it with a subset of known mRNAs and lncRNAs from the reference annotation maintains high predictive performance (Wucher et al., 2017), therefore allowing ANNEXA to integrate FEELnc with 3,000 reference lncRNAs and mRNAs.
To assess the potential impact of genomic alterations (mutations) on newly identified protein-coding transcripts, ANNEXA employs TransDecoder (https://github.com/TransDecoder/TransDecoder) with the `--single_best_only` option, 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. 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/Transcription Start Sites (TSSs) (Tardaguila et al., 2018).
To filter the *Extended_annotations.full.gtf* and remove novel incomplete transcripts, ANNEXA implements two filtering strategies based on (i) the Novel Discovery Rate (NDR) cut-off from the Bambu tool and (ii) the transforKmer cut-off, applicable to both Bambu and StringTie-derived transcripts. The transforKmer cut-off in ANNEXA evaluates the likelihood of all novel transcript TSSs using a species specific deep learning model pre-trained with DNABERT (Ji et al., 2021) and fine-tuned on a classification task using labelled TSSs from the reference annotation (Karollus et al., 2024).
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 (See wiki ANNEXA).
At the end of this module, ANNEXA outputs an additional more stringent annotation, referred to as *Extended_annotations.filter.gtf*.

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) (See ANNEXA Wiki).

These features can be broadly categorized 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 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).
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 (for illustration 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. It is

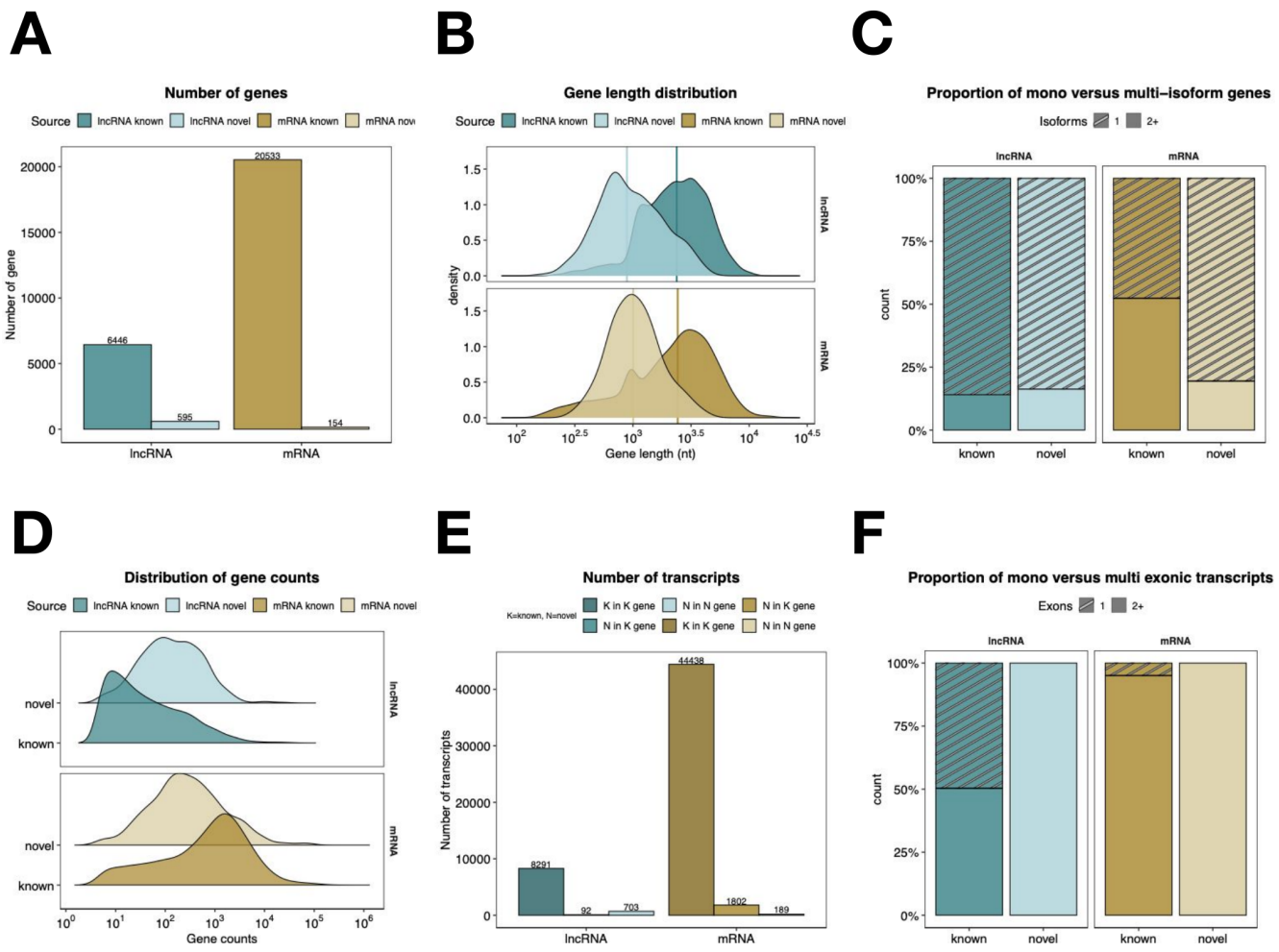


Figure 2 – Automatic QC report from ANNEXA. **A**- Number of known and novel genes. **B**- Gene length distribution. **C**- Proportion of gene mono- versus multi-isoform(s). **D**- Distribution of gene counts. **E**- Number of transcripts in known (K) and novel (N) genes. **F**- Proportion of transcript mono/single- versus multi-exonic.

available as 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).

Computational methods. Basecalling of `fast5` files was done with guppy (version 6.0.0) and the `nf-core/nanoseq` pipeline (version 3.1.0) 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 (**Table1**). Intersection between all genomics features (i.e. TSS and CAGE data) was done using bedtools (version 2.27.1).

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

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

Benchmark analyses

Reference annotations and genomes. Depending on the species, 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 (Wang et al., 2021) (Varabyou et al., 2023) (**Table 2**).

Species	Assembly	Reference Annotation	Genes	Transcripts
Dog	canFam3	Ensembl	30,951	60,994
Dog	canFam4	Ensembl	30,653	56,403
Dog	canFam4	UU (University Uppsala)	29,535	158,561
Dog	canFam4	Refseq	43,427	103,619
Dog	canFam5	Ensembl	26,037	48,737
Dog	canFam6	Ensembl	29,992	53,113
Human	GRCh38	Gencode	86,402	412,034
Human	GRCh38	CHESS	63,755	168,451

Table 2 – Species, Assembly, Reference Annotation with Gene and Transcript Information

Transcriptome reconstruction benchmarking. ANNEXA was run using Bambu and StringTie on canine samples with the canFam4 genome assembly and the Ensembl annotation, and on the human samples with the GRCh38 assembly and Chess or Gencode annotations. For the Bambu runs, we discarded single exon transcripts and used the recommended NDR of the tool. The full and filtered extended annotations produced after the runs were filtered to only keep genes and transcripts expressed in at least one of each sample (raw count >0) 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 level. At each level, reconstructed elements matching the reference annotation were considered true positives (TP), while those absent from the reference were classified as false positives (FP). Conversely,

features present in the reference but missing from the extended annotations were considered as false negatives (FN). We thus defined 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 to known genes from target annotation *e.g.* Gencode for query dog gene and Ensembl for human query gene).

Results

In a comparative oncology study, 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 test the robustness of the pipeline in its ability to extend and to quality control these annotations.

Effect of species-specific reference transcriptome 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, UU), with this difference being the most noticeable for the University of Uppsala (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 different being the strongest 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 complete than in dogs and also because our experimental design contained more LR-RNAseq samples in dogs (n=8) versus human (n=2).

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

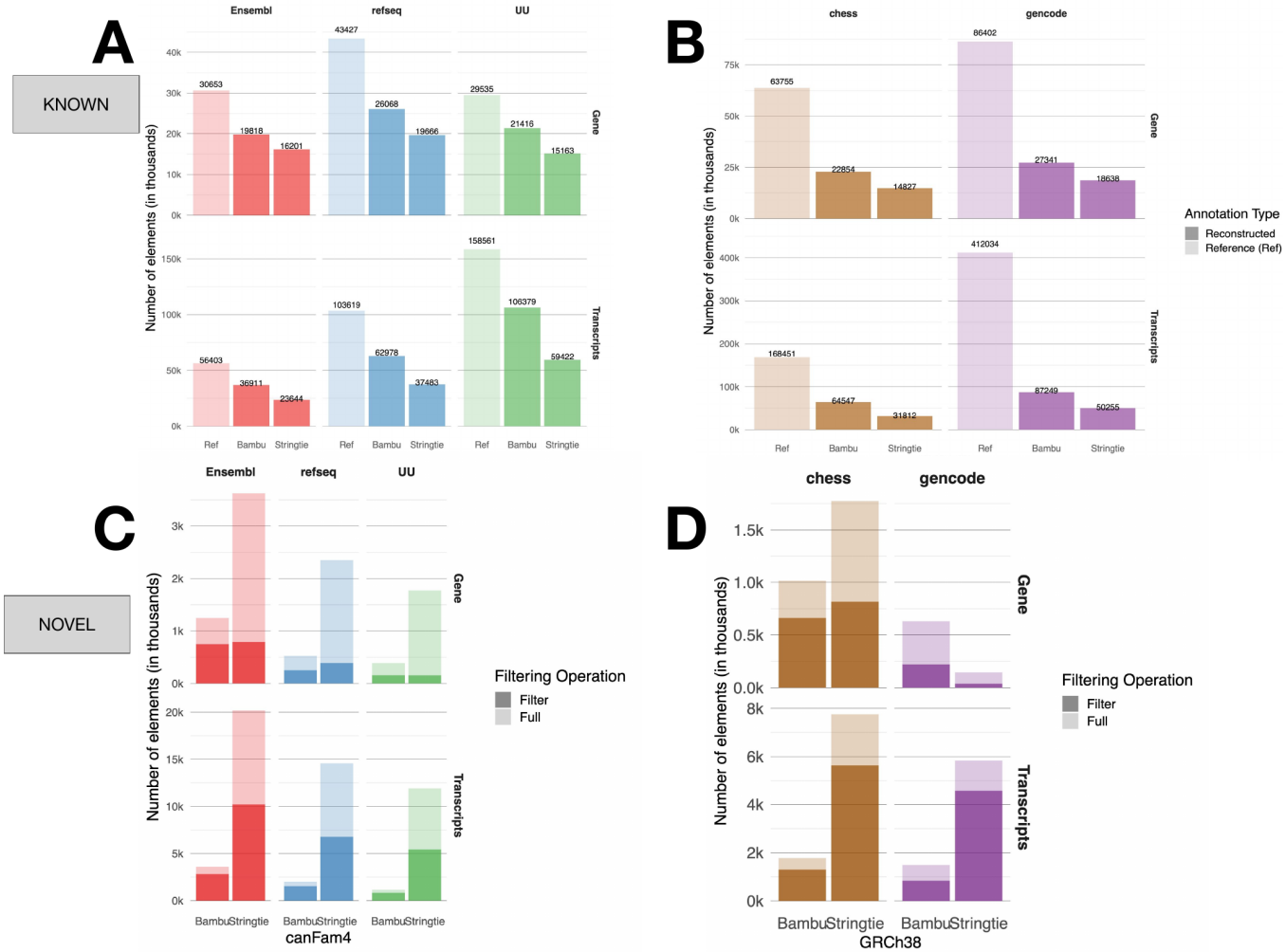


Figure 3 – Benchmark analysis between Bambu and StringTie for dog (left panels) and human (right panels) for known (top panel) and novel (bottom panel) genes/transcripts. For known elements (top), number of genes/transcripts from the reference annotation (light bar) are separated from reconstructed genes/transcripts (dark bar). **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. Full set of elements (light bar) are separated from ANNEXA's filtered (intersection) set (dark bar).

reducing the number of newly detected transcripts (**Fig.3.C**). In the human genome (GRCh38), using the Gencode annotation as reference produced substantially fewer novel elements than with CHESS, as expected given the higher number of genes and transcripts cataloged in the latest version of Gencode (**Fig.3.D**). Yet, at the gene level, we observed an intriguing pattern where StringTie annotates more novel genes with CHESS compared to Gencode, whereas the opposite trend occurs with Bambu (∼1000 novel genes for Bambu and 1,800 for StringTie with the CHESS annotation as compared to 631 and 145 with Gencode, respectively). Regarding ANNEXA filtering operations, which evaluated the full-lengthness of reconstructed genomic elements (see **Methods**), we observed that it dramatically reduced the number of novel elements identified by both tools. On average in dogs, 50% of Bambu and 16% of StringTie novel genes

are being conserved after ANNEXA's check for TSS validity. The proportion of StringTie's filtered transcripts is more pronounced with Ensembl (dog) and CHESS (human) annotations, suggesting that these novel transcripts may contain a higher proportion of incomplete transcripts. 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 completeness (**Supp. Fig. ??**)

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 (Kainth et al., 2023) (Derrien et al., 2012). 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 showed 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 categorized 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, 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**).

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.

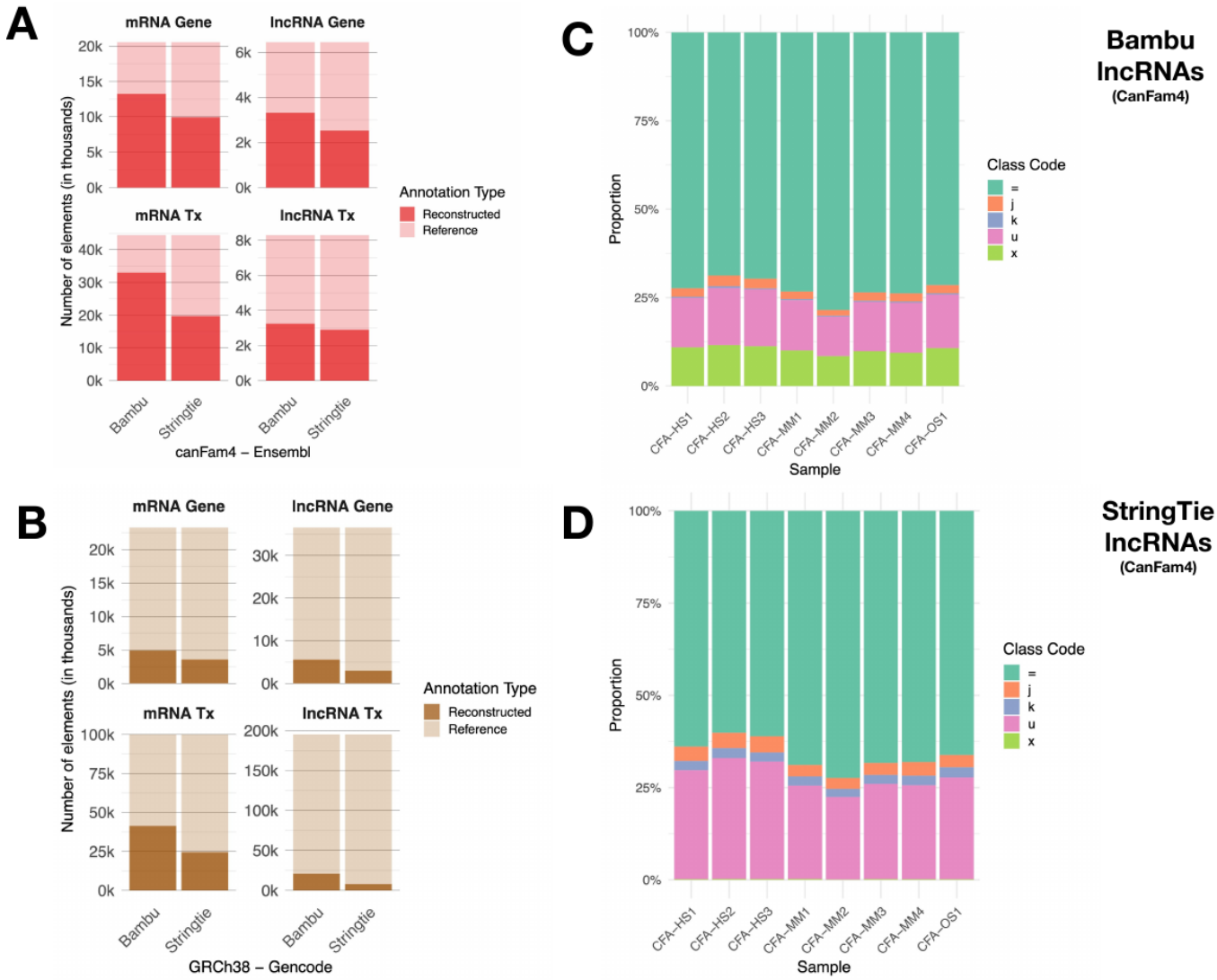


Figure 4 – Annotation of lncRNAs/mRNAs by Bambu and StringTie. **A**- Number of known canine genes (top) and transcripts (bottom) for mRNA (left) and lncRNAs (right) annotated by Ensembl. **B**- Number of known human genes (top) and transcripts (bottom) for mRNA (left) and lncRNAs (right) annotated by 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.

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, 3,268 (88%) correspond to human genes from the Gencode extended reference annotation and thus represent novel orthologous relationships between novel dog and known human genes. Notably,

we identified five novel canine lncRNA genes that also map to novel human genes (**Supp. Fig. ??**), with two of them being classified as protein coding in human. 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.

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. In addition, by uniquely integrating the comparative characterization of lncRNAs annotated from long-read RNA-seq data, ANNEXA enhances the profiling of these non-coding elements and refines their potential biological relevance. The comparison of various 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 coverage of the reference annotations. These findings align with the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium's observations, which demonstrated that tools based on reference genomes perform optimally in well-annotated genomes (Pardo-Palacios et al., 2024b). The impact of reference annotation choice is also evident in the analysis of our human 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.

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 patterns between the two tools. Bambu primarily categorised novel lncRNAs as intergenic (lincRNAs) or antisense, whereas StringTie assigned a higher proportion to lincRNAs and novel isoforms of known lncRNAs, with only a few lncRNAs classified as antisense. 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) may benefit more from Bambu's capabilities, whereas those aiming to extend known lncRNA annotations might prefer StringTie.

Our analyses further emphasizes 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

Acknowledgements

The authors thank the Genouest bioinformatic core facility for providing us 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), Cancéropôle Grand Ouest (CGO) 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., 2025))
Supplementary information is available online (https://doi.org/10.24072/fake3[Replace by the DOI of the webpage hosting the Supplementary information]; (**CharleMar_independant-trace-su2_2012**)[Re by the citation of the Supplementary information])

References

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**. Publisher: Nature Publishing Group, 1187–1195. https://doi.org/10.1038/s41592-023-01908-w. URL: https://www.nature.com/articles/s41592-023-01908-w (visited on 02/21/2025).

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/ (visited on 03/24/2025).

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 (visited on 03/14/2025).

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 Publisher: Cold Spring Harbor Lab, 1775–1789. https://doi.org/10.1101/gr.132159.111. URL: http://genome.cshlp.org/content/22/9/1775 (visited on 03/27/2025).

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**. Publisher: Nature Publishing Group, 316–319. https://doi.org/10.1038/nbt.3820. URL: https://www.nature.com/articles/nbt.3820 (visited on 02/20/2025).

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 (visited on 04/03/2025).

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 (visited on 03/14/2025).

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**. Publisher: Nature Publishing Group, 276–278. https://doi.org/10.1038/s41587-020-0439-x. URL: https://www.nature.com/articles/s41587-020-0439-x (visited on 04/02/2025).

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 (visited on 04/02/2025).

Hoffmann N, Derrien T, vlebars, Aurore-B, Lorthiois M (2025). IGDRion/ANNEXA: Annexa v.3.2.5. https://doi.org/10.5281/zenodo.15209931. URL: https://zenodo.org/records/15209931 (visited on 04/14/2025).

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**. Publisher: Nature Publishing Group, 9082. https://doi.org/10.1038/s41467-024-52798-1. URL: https://www.nature.com/articles/s41467-024-52798-1 (visited on 03/30/2025).

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/ (visited on 03/14/2025).

Kainth AS, Haddad GA, Hall JM, Ruthenburg AJ (2023). Merging short and stranded long reads improves transcript assembly. *PLOS Computational Biology* **19**. Publisher: Public Library of Science, e1011576. https://doi.org/10.1371/journal.pcbi.1011576. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011576 (visited on 03/27/2025).

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 Publisher: BioMed Central, 1–21. https://doi.org/10.1186/s13059-024-03221-x. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03221-x (visited on 03/14/2025).

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 (visited on 03/11/2025).

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 Publisher: BioMed Central, 1–13. https://doi.org/10.1186/s13059-019-1910-1. URL:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1 (visited on 02/21/2025).

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 (visited on 03/14/2025).

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/ (visited on 04/03/2025).

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**. Publisher: Nature Publishing Group, 793–797. https://doi.org/10.1038/s41592-024-02229-2. URL: https://www.nature.com/articles/s41592-024-02229-2 (visited on 03/14/2025).

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**. Publisher: Nature Publishing Group, 1349–1363. https://doi.org/10.1038/s41592-024-02298-3. URL: https://www.nature.com/articles/s41592-024-02298-3 (visited on 03/11/2025).

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 (visited on 03/14/2025).

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 Publisher: Multidisciplinary Digital Publishing Institute, 276. https://doi.org/10.3390/cancers14020276. URL: https://www.mdpi.com/2072-6694/14/2/276 (visited on 03/30/2025).

Prouteau A, André C (2019). Canine Melanomas as Models for Human Melanomas: Clinical, Histological, and Genetic Comparison. *Genes* **10**. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, 501. https://doi.org/10.3390/genes10070501. URL: https://www.mdpi.com/2073-4425/10/7/501 (visited on 03/30/2025).

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**. Publisher: Nature Publishing Group, 14908. https://doi.org/10.1038/s41598-019-51470-9. URL: https://www.nature.com/articles/s41598-019-51470-9 (visited on 04/01/2025).

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 (visited on 03/29/2025).

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/ (visited on 03/24/2025).

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**. Publisher: Nature Publishing Group, 601–614. https://doi.org/10.1038/nrg.2016.85. URL: https://www.nature.com/articles/nrg.2016.85 (visited on 03/31/2025).

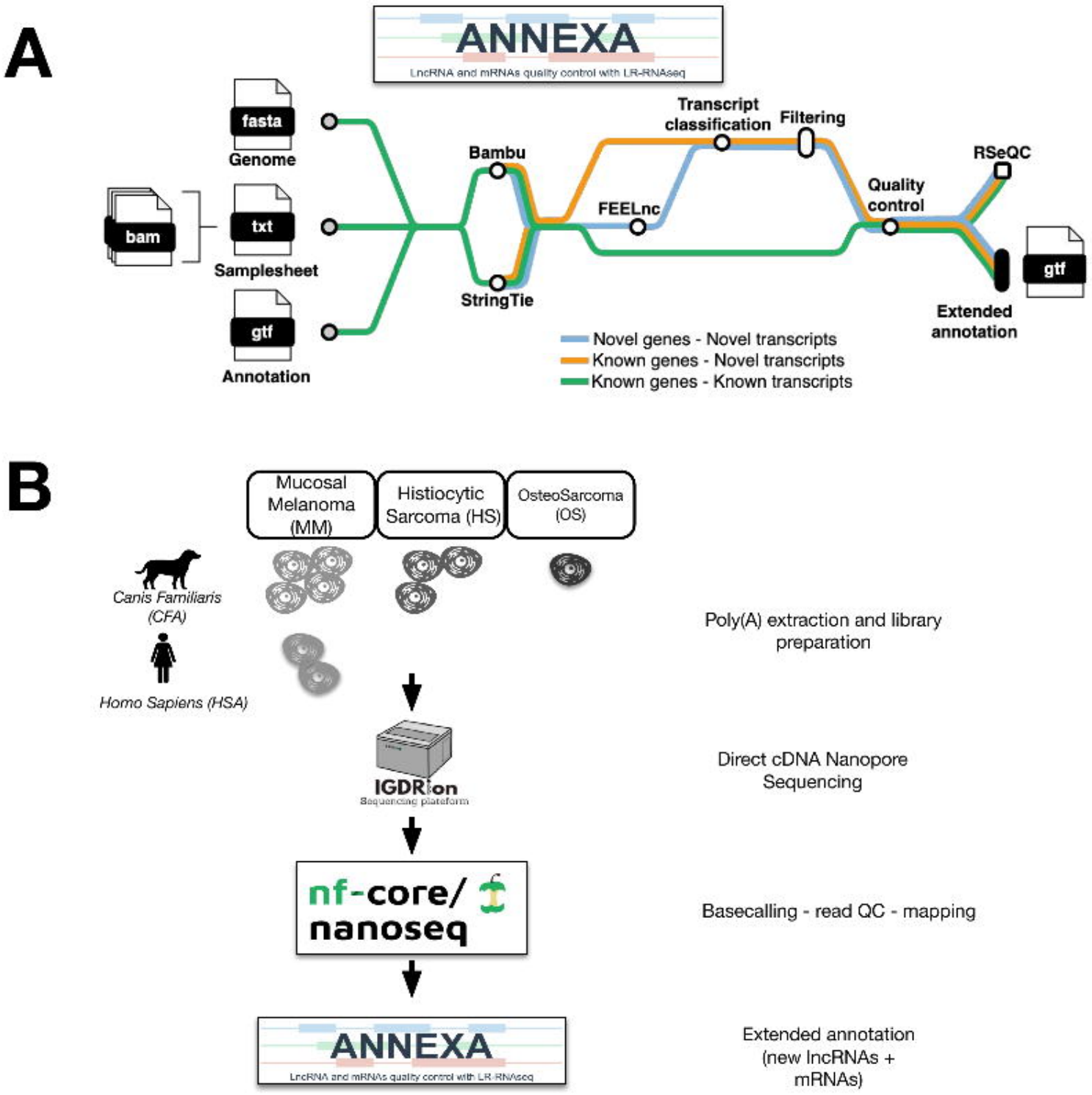
Varabyou A, Sommer MJ, Erdogdu B, Shinder I, Minkin I, Chao KH, Park S, Heinz J, Pockrandt C, Shumate A, Rincon N, Puiu D, Steinegger M, Salzberg SL, Pertea M (2023). CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, 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 (visited on 04/02/2025).

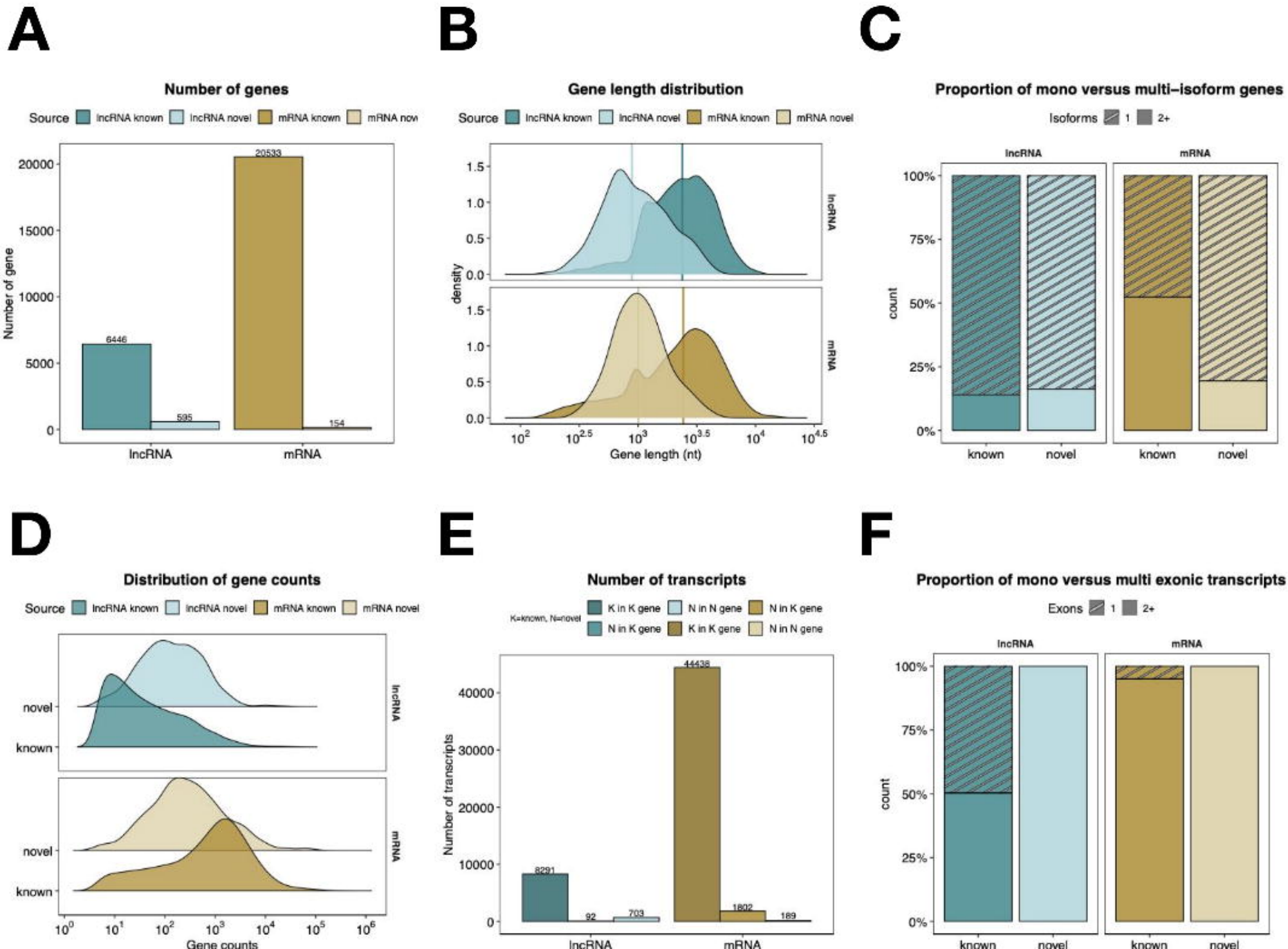
Vitting-Seerup K, Sandelin A (2019). IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics* **35**, 4469–4471. https://doi.org/10.1093/bioinformatics/btz247. URL: https://doi.org/10.1093/bioinformatics/btz247 (visited on 03/11/2025).

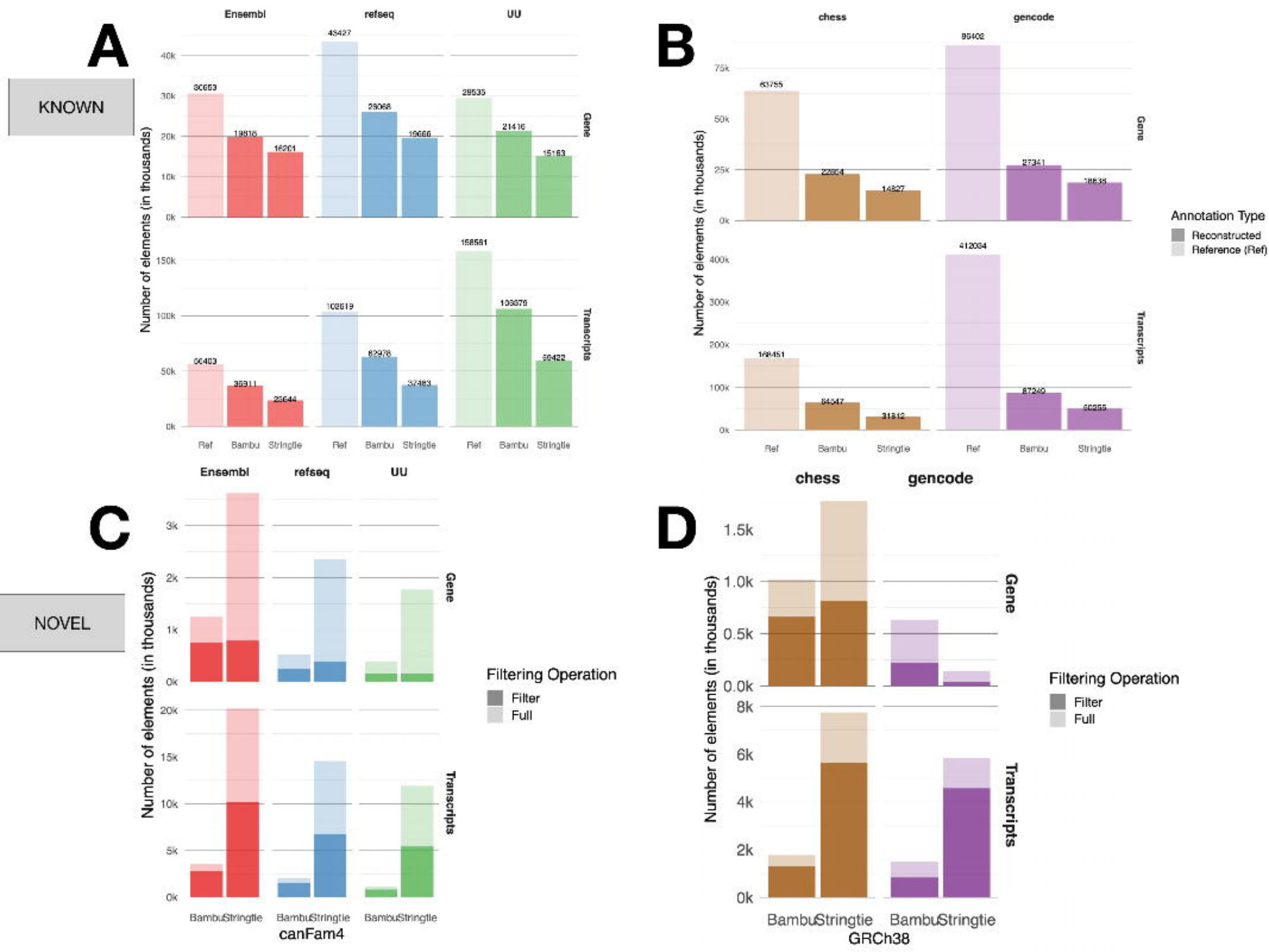
Wang C, Wallerman O, Arendt ML, Sundström E, Karlsson Å, Nordin J, Mäkeläinen S, Pielberg GR, Hanson J, Ohlsson Å, Saellström S, Rönnberg H, Ljungvall I, Häggström J, Bergström TF, Hedhammar Å, Meadows JRS, Lindblad-Toh K (2021). A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Communications Biology* **4**. Publisher: Nature Publishing Group, 1–11. https://doi.org/10.1038/s42003-021-01698-x. URL: https://www.nature.com/articles/s42003-021-01698-x (visited on 04/02/2025).

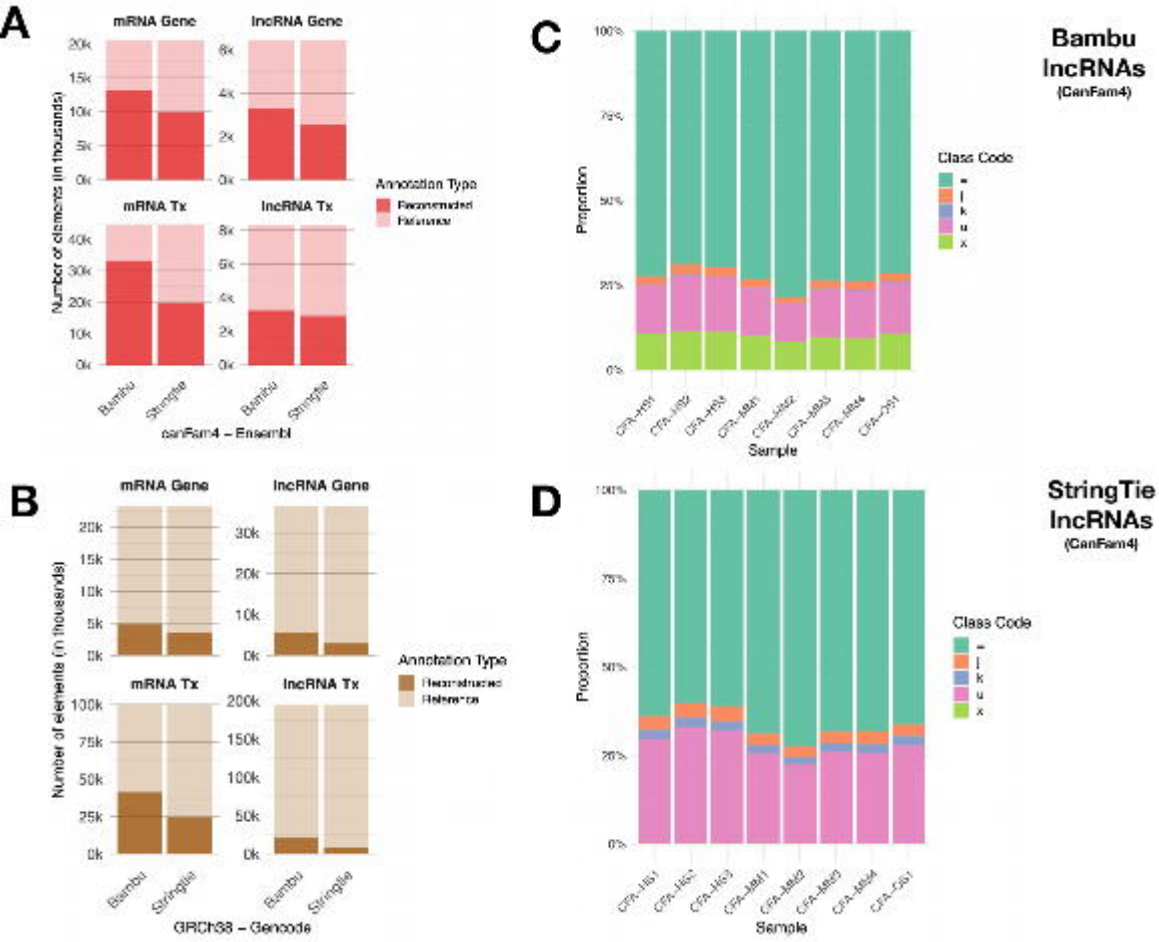
Wang L, Wang S, Li W (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185. https://doi.org/10.1093/bioinformatics/bts356. URL: https://doi.org/10.1093/bioinformatics/bts356 (visited on 03/14/2025).

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 (visited on 02/21/2025).


A
ANNEXA
LncRNA and mRNAs quality control with LR-RNAseq
fasta
Genome
bam
txt
Samplesheet
gtf
Annotation
Bambu
StringTie
FEELnc
Transcript classification
Filtering
Quality control
RSeQC
gtf
Extended annotation
Novel genes - Novel transcripts
Known genes - Novel transcripts
Known genes - Known transcripts
B
Mucosal Melanoma (MM)
Histiocytic Sarcoma (HS)
OsteoSarcoma (OS)
Canis Familiaris (CFA)
Homo Sapiens (HSA)
Poly(A) extraction and library preparation
IGDRion
Sequencing platform
Direct cDNA Nanopore Sequencing
nf-core/ nanoseq
Basecalling - read QC - mapping
ANNEXA
LncRNA and mRNAs quality control with LR-RNAseq
Extended annotation (new lncRNAs + mRNAs)


A
Number of genes
Source lncRNA known lncRNA novel mRNA known mRNA nov
Number of gene
20000
15000
10000
5000
0
6446
595
20533
154
lncRNA
mRNA
B
Gene length distribution
Source lncRNA known lncRNA novel mRNA known mRNA novel
density
1.5
1.0
0.5
0.0
lncRNA
mRNA
10^2
$10^{2.5}$
10^3
$10^{3.5}$
10^4
$10^{4.5}$
Gene length (nt)
C
Proportion of mono versus multi−isoform genes
Isoforms 1 2+
lncRNA
mRNA
count
100%
75%
50%
25%
0%
known
novel
known
novel
D
Distribution of gene counts
Source lncRNA known lncRNA novel mRNA known mRNA novel
novel
known
lncRNA
novel
known
mRNA
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Gene counts
E
Number of transcripts
K=known, N=novel
K in K gene
N in N gene
N in K gene
N in K gene
K in K gene
N in N gene
Number of transcripts
40000
30000
20000
10000
0
8291
92
703
44438
1802
189
lncRNA
mRNA
F
Proportion of mono versus multi exonic transcripts
Exons 1 2+
lncRNA
mRNA
count
100%
75%
50%
25%
0%
known
novel
known
novel


A
KNOWN
Ensembl
refseq
UU
Number of elements (in thousands)
Gene
Transcripts
Ref
Bambu
Stringtie
B
chess
gencode
Annotation Type
Reconstructed
Reference (Ref)
C
NOVEL
Filtering Operation
Filter
Full
canFam4
D
GRCh38


A
mRNA Gene
lncRNA Gene
mRNA Tx
lncRNA Tx
Number of elements (in thousands)
Annotation Type
Recommended
Reference
Bambu
Stringtie
canFam4 – Ensembl
B
mRNA Gene
lncRNA Gene
mRNA Tx
lncRNA Tx
Number of elements (in thousands)
Annotation Type
Recommended
Reference
Bambu
Stringtie
GRCh38 – Gencode
C
Bambu
lncRNAs
(CanFam4)
Proportion
Class Code
=
j
k
u
x
Sample
D
StringTie
lncRNAs
(CanFam4)
Proportion
Class Code
=
j
k
u
x
Sample