Exploring the (non-)coding genome through long-read sequencing

Thomas DERRIEN
(thomas.derrien@univ-rennes.fr)

CGT : Canine Genetics Team

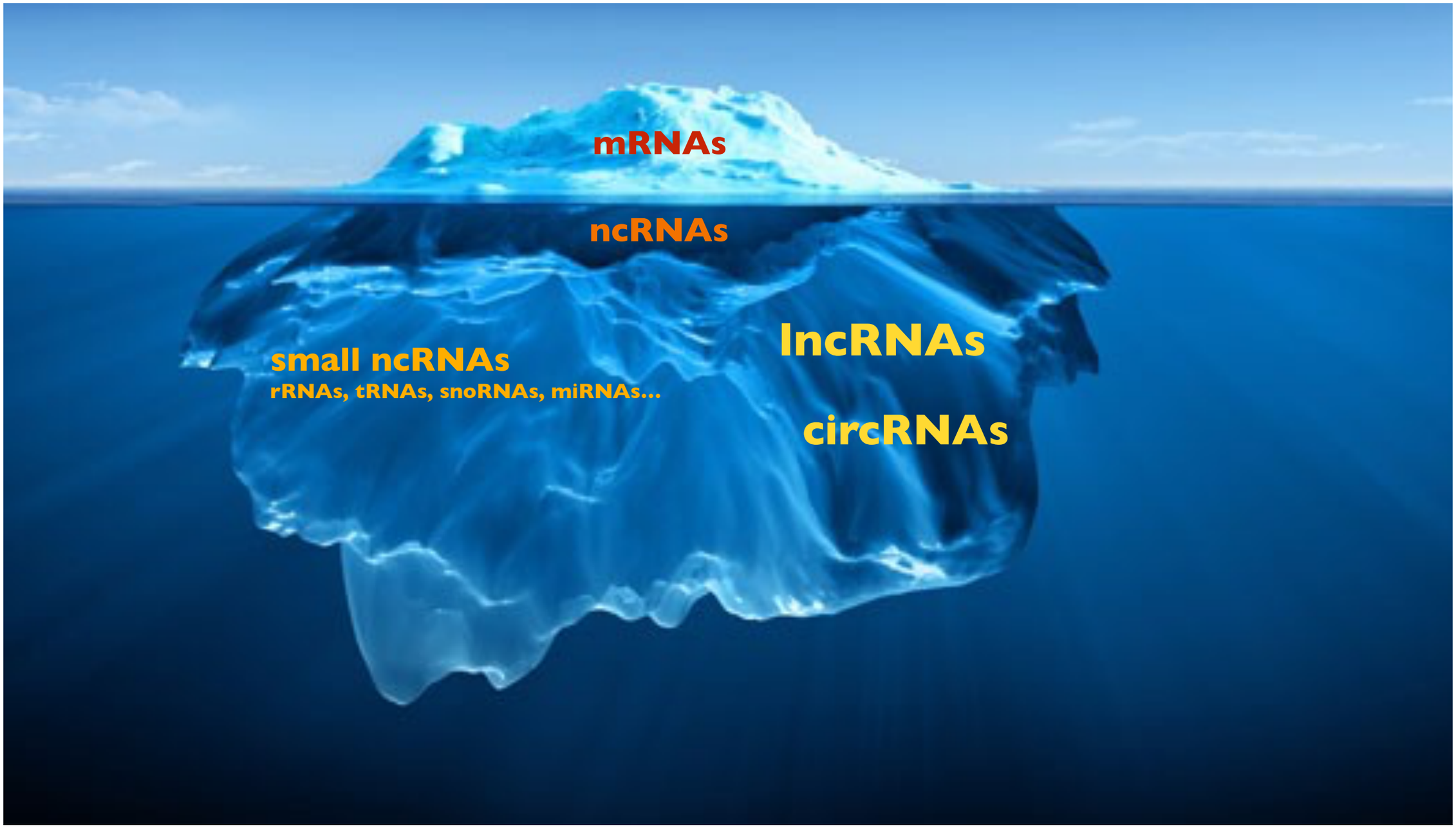
IGDRion sequencing PF

PEPI IBIS - Oct 2025

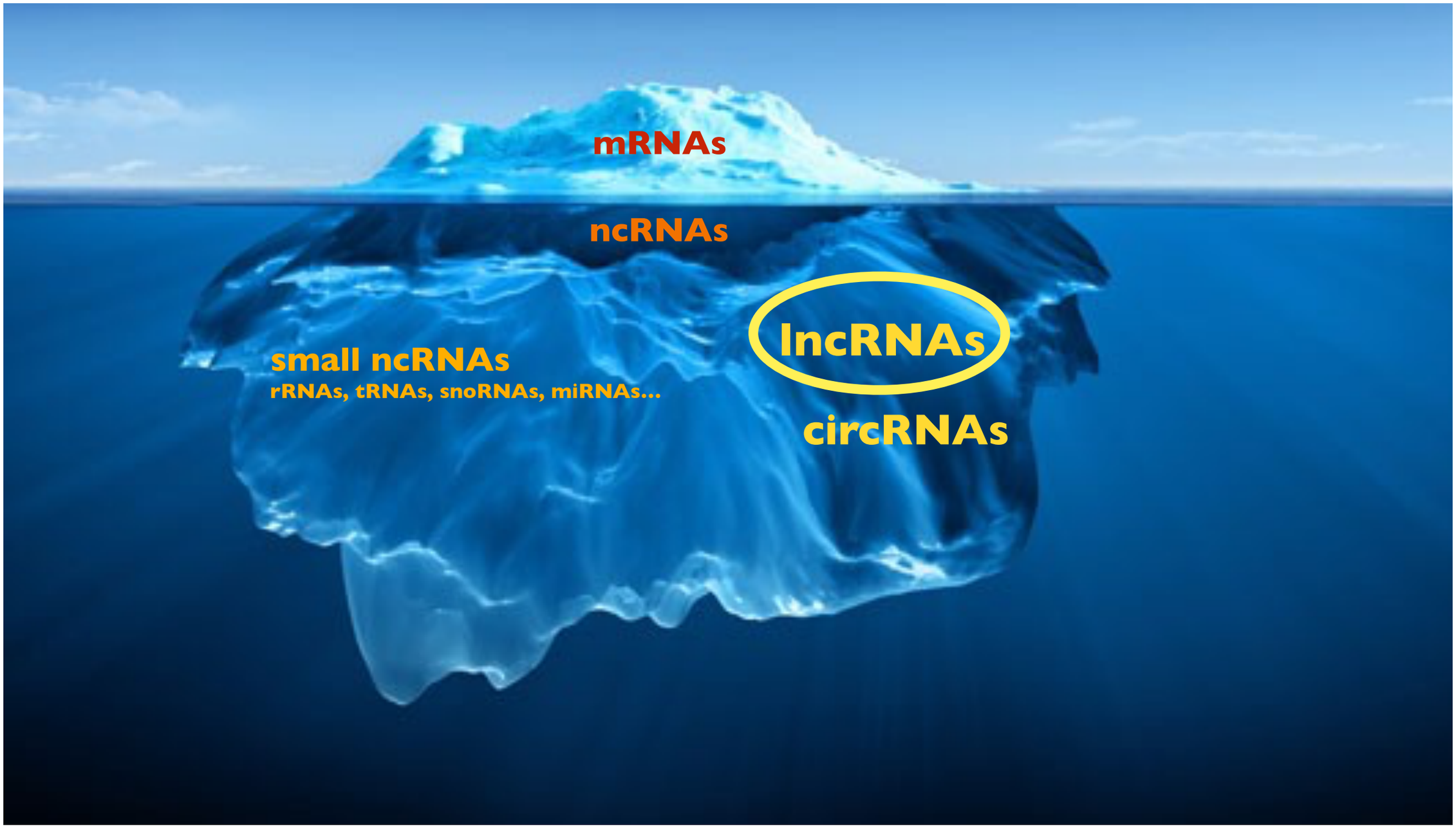
The cell as an RNA machine

The cell as an RNA machine

The cell as an RNA machine



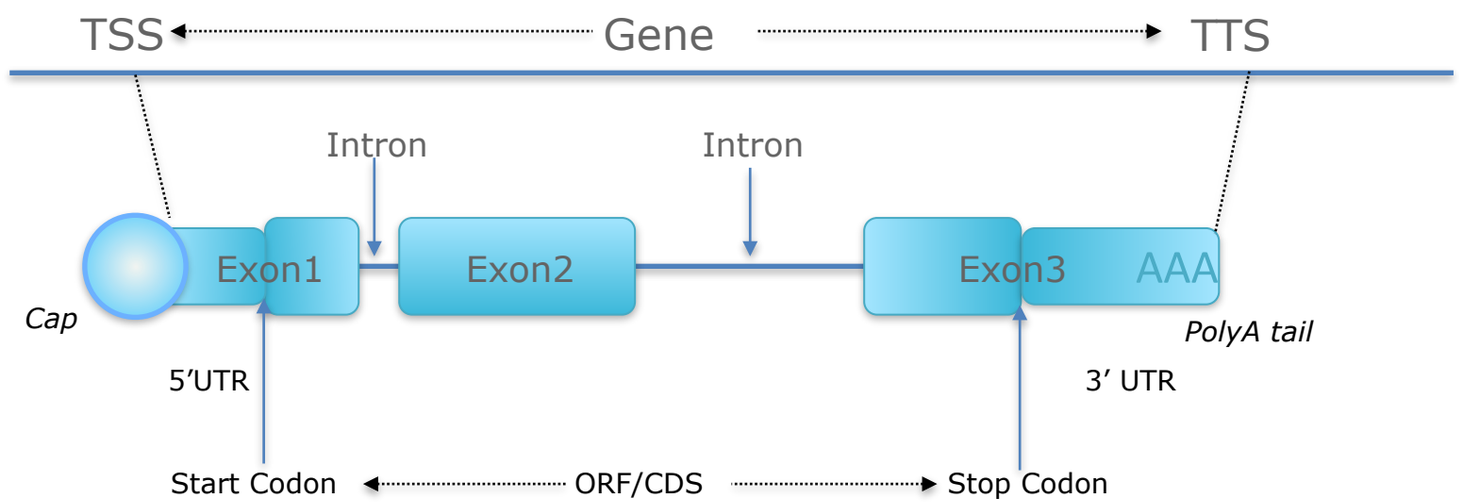
The cell as an RNA machine



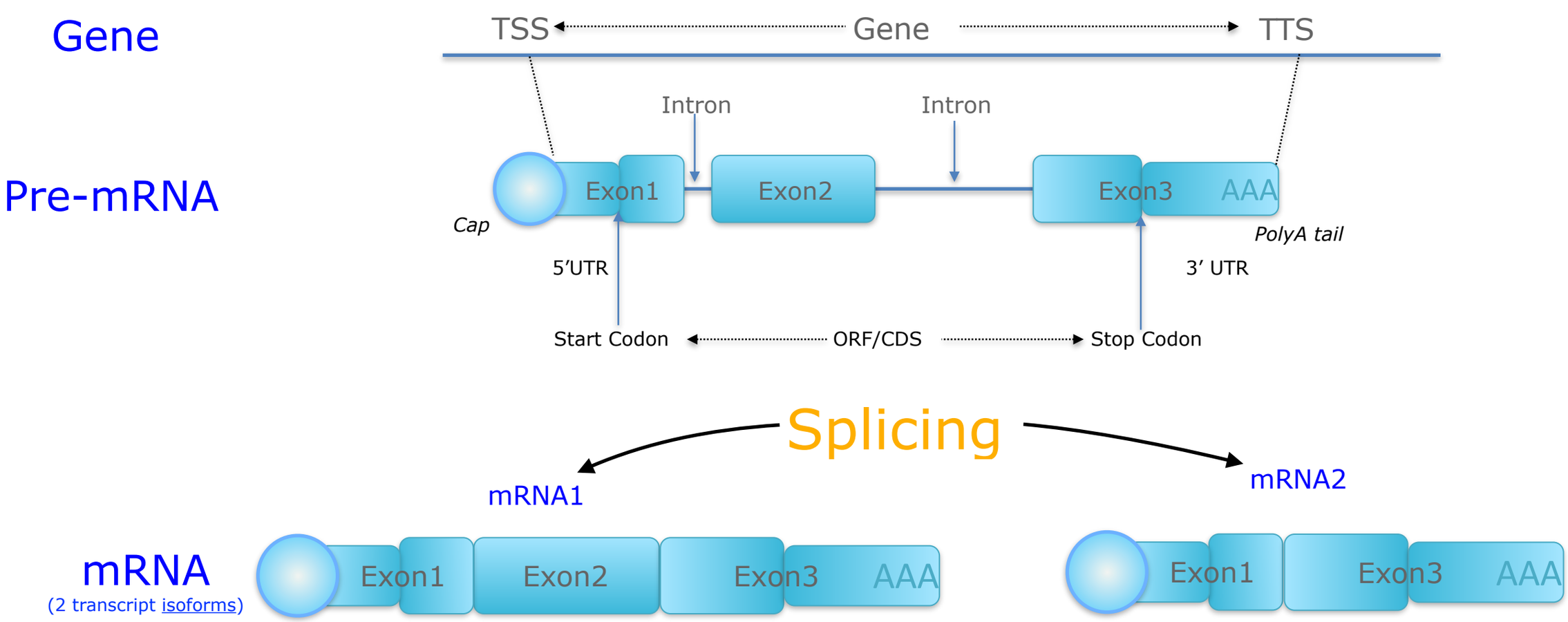
Back2Basics

Gene

Pre-mRNA

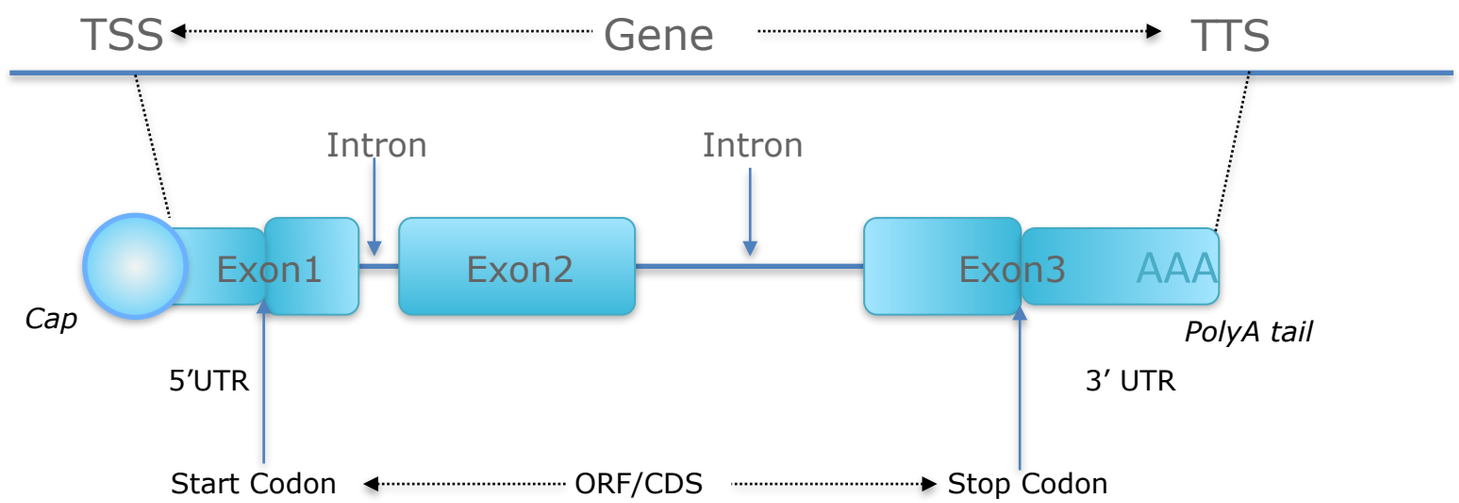


Back2Basics

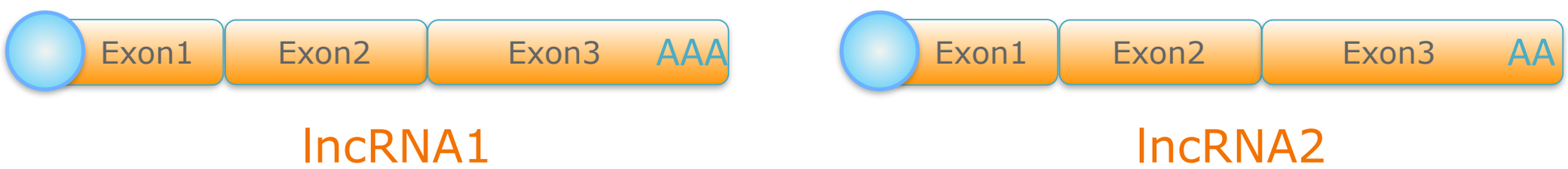


mRNAs versus lncRNAs

Gene



lncRNA
(Transcript/Isoform)

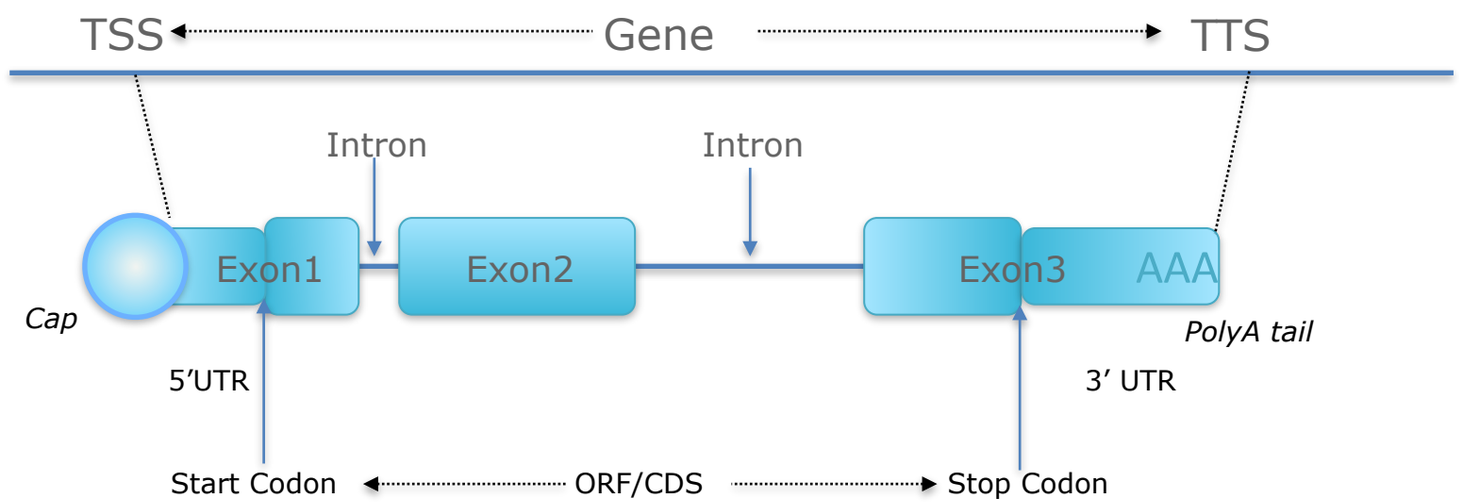


lncRNAs are "mRNA-like" ...

- ✅ Transcribed by RNA Polymerase II
- ✅ Undergo splicing
- ✅ 5' capped
- ✅ Most have a polyA tail

mRNAs versus lncRNAs

Gene



lncRNA
(Transcript/Isoform)

lncRNA1

lncRNA2

lncRNAs are "mRNA-like" ...

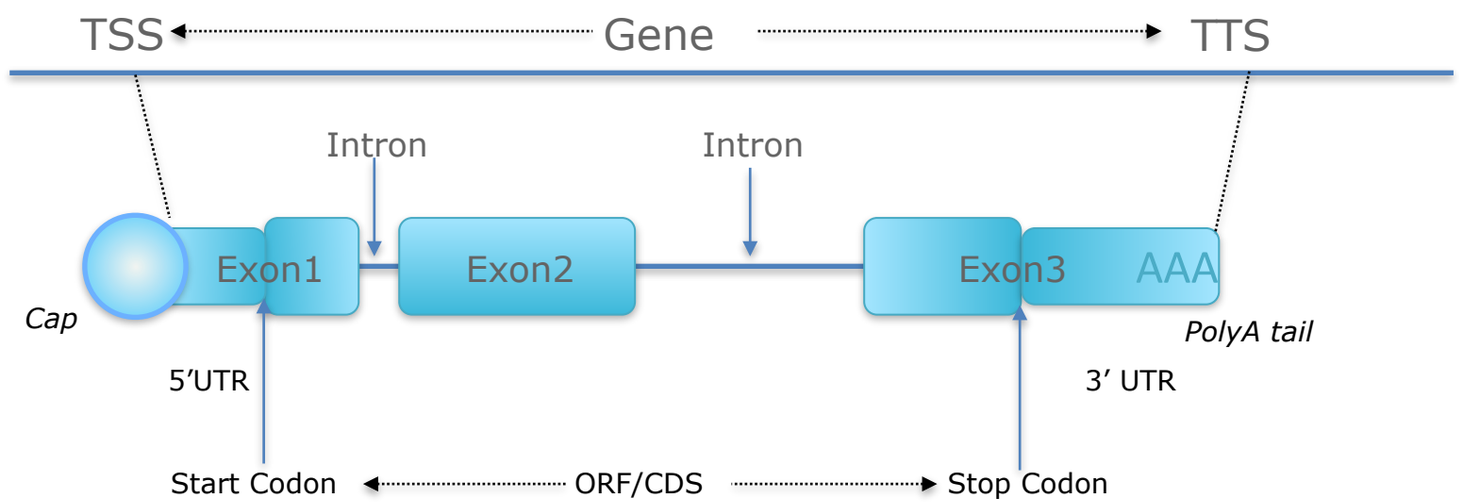
- ✅ Transcribed by RNA Polymerase II
- ✅ Undergo splicing
- ✅ 5' capped
- ✅ Most have a polyA tail

... but are not mRNAs

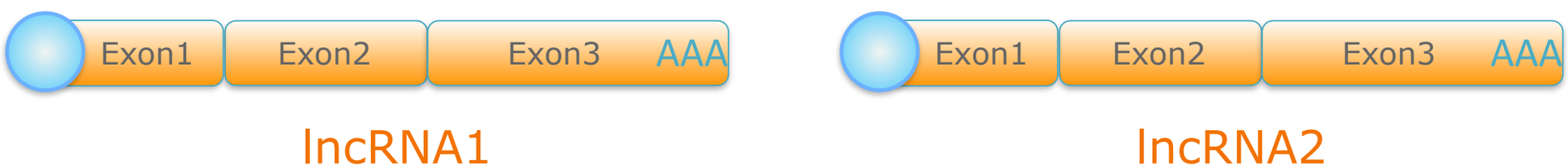
- 🚫 No coding potential

mRNAs versus lncRNAs

Gene



lncRNA
(Transcript/Isoform)



lncRNAs are "mRNA-like" ...

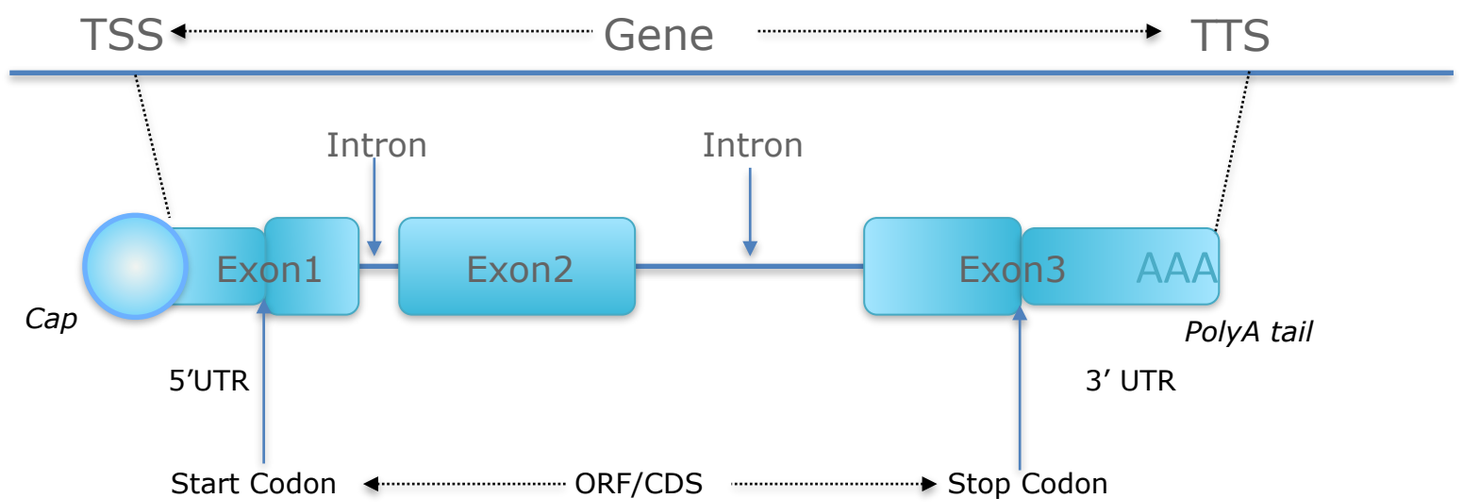
- ✅ Transcribed by RNA Polymerase II
- ✅ Undergo splicing
- ✅ 5' capped
- ✅ Most have a polyA tail

... but are not mRNAs

- 🚫 No coding potential
- 🚫 Lowly expressed, highly tissue-specific
- 🚫 Less conserved across species
- 🚫 Often overlap Transposable Elements (TEs)

mRNAs versus lncRNAs

Gene



lncRNA
(Transcript/Isoform)

lncRNA1 lncRNA2

lncRNAs are "mRNA-like" ...

- ✅ ◉ Transcript >200nt with low coding potential
- ✅ ◉ Classified vs mRNA locations : lincRNA vs antisense
- ✅ 5' capped
- ✅ Most have a polyA tail

... but are not mRNAs

- 🚫 Less conserved across species
- 🚫 Often overlap Transposable Elements (TEs)

lncRNA functions

doi:1

lincRNAs act in the circuitry contro pluripotency and differentiation

Cell

Transcription of Two Long Noncoding RNAs Mediates Mating-Type Control of Gametogenesis in Budding Yeast

Nature. 2012 Nov 15;491(7424):454-7. doi: 10.1038/nature11508. Epub 2012 Oct 14.

Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.

Epigenetic Regulation by Long Noncoding RNAs

Jeannie T. Lee

Science **338**, 1435 (2012);
DOI: 10.1126/science.1231776

Cell

A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA

Marcella Cesana,[1,6] Davide Cacchiarelli,[1,6] Ivano Legnini,[1] Tiziana Santini,[1] Olga Sthandier,[1] Mauro Chinappi,[2] Anna Tramontano,[2,3,4] and Irene Bozzoni[1,3,4,5,*]

The Emergence of lncRNAs in Cancer Biology

John R. Prensner and Arul M. Chinnaiyan

LETTER

doi:10.1038/nature17161

Melanoma addiction to the long non-coding RNA SAMMSON

erto Vendramin[1,2], Marco Spinazzi[2], Patrick Laurette[3], Mark Fiers[2], Jasper Wouters[4], Enrico Radaelli[5], rina Leonelli[8,9], Katrien Vanderheyden[8,9], Aljosja Rogiers[1,2], Els Hermans[10], Pieter Baatsen[2], Stein Aerts[11], n Van Aelst[12,13], Joost van den Oord[4], Bart de Strooper[2], Irwin Davidson[3], Denis L. J. Lafontaine[14], lesompele[8,9], Pieter Mestdagh[8,9]* & Jean-Christophe Marine[1,2]*

Biochim Biophys Acta. 2016 Jul;1863(7 Pt B):1953-60. doi: 10.1016/j.bbamcr.2015.11.026. Epub 2015 Nov 24.

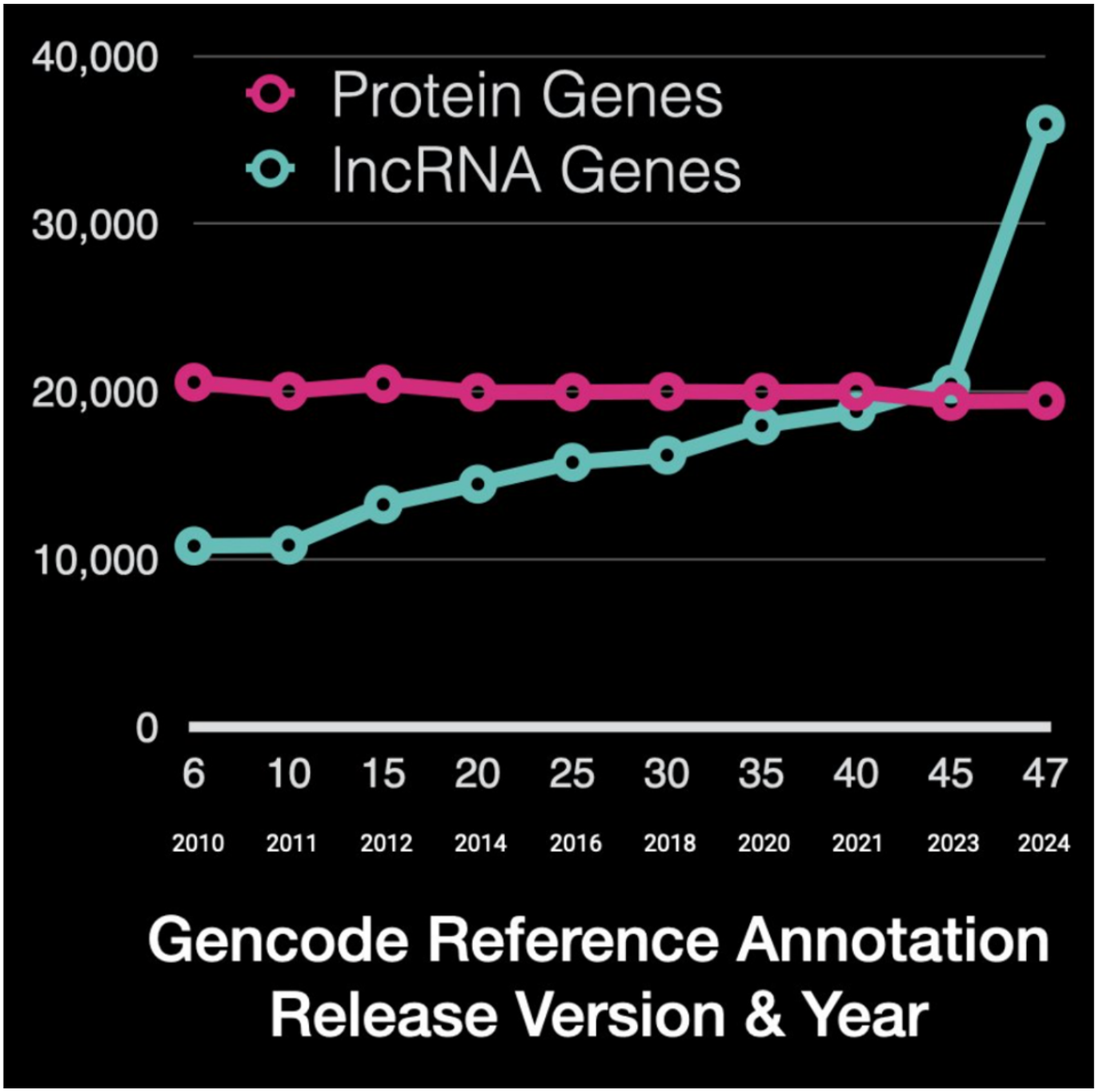
Super-enhancer lncs to cardiovascular development and disease.

Oun

Gene regulation by long non-coding RNAs and its biological functions

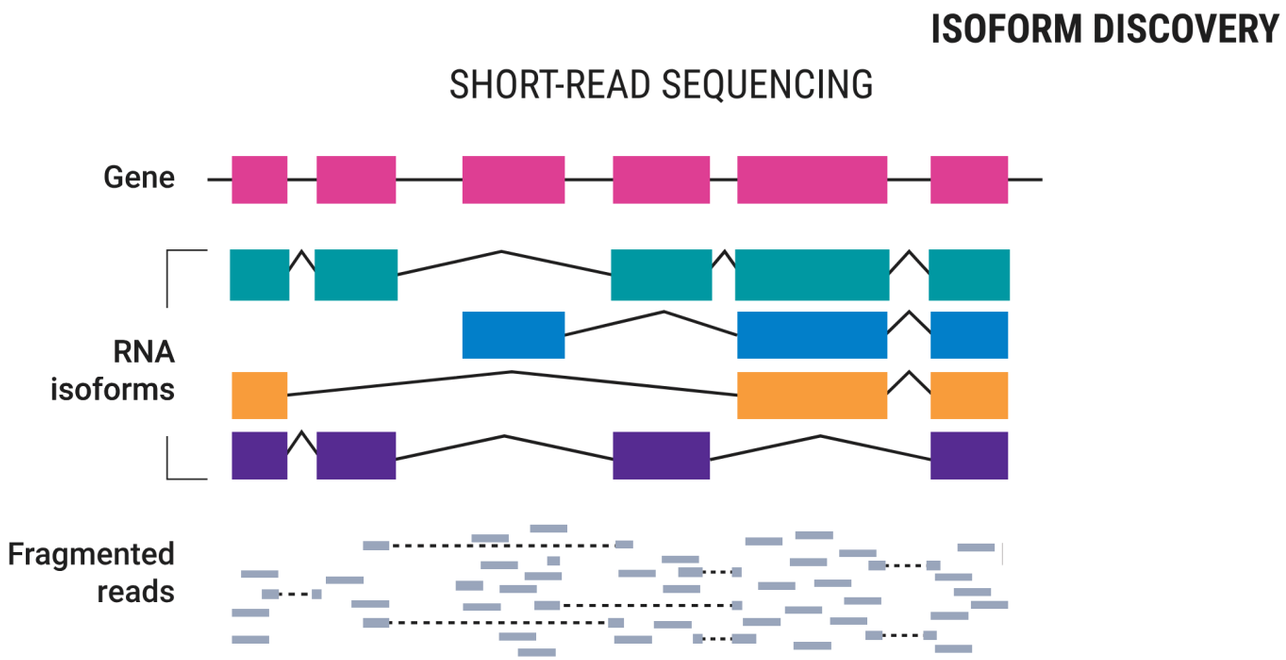
Luisa Statello[1,2,6], Chun-Jie Guo[3,6], Ling-Ling Chen[3,4,5] and Maite Huarte[1,2]

Reference annotation



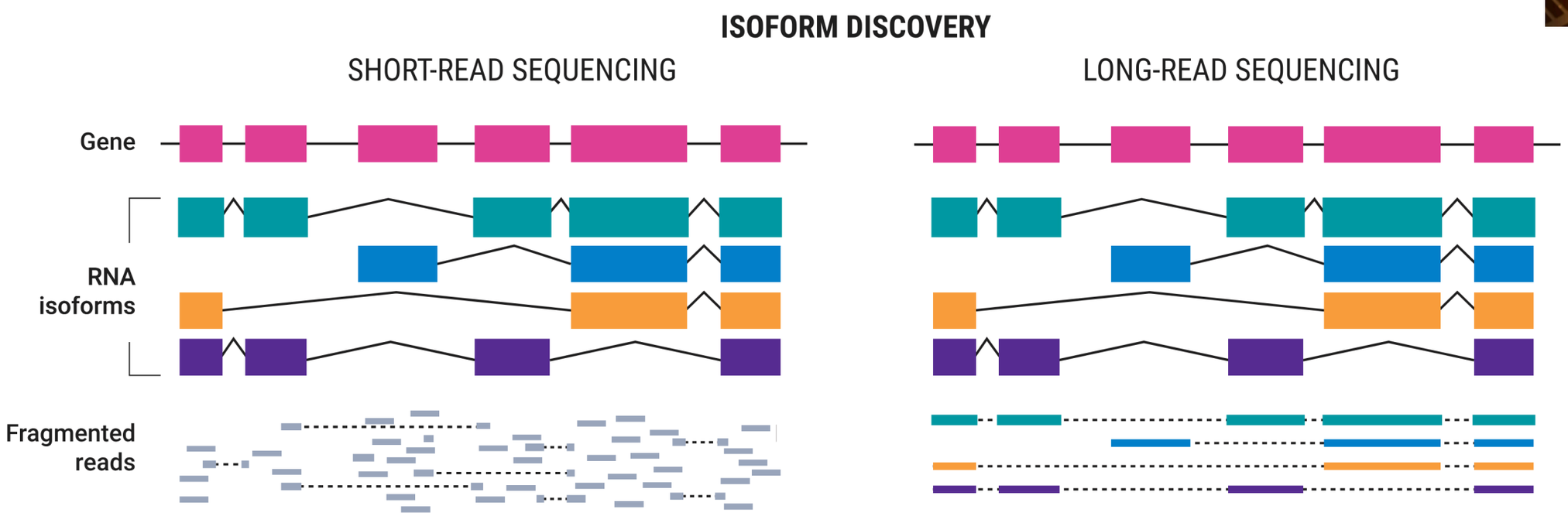
- Gold standard for reference annotation in human and mouse
- **Manual curation +** evidence-based transcript models (RNASeq)

Genome annotation with short-read RNAseq



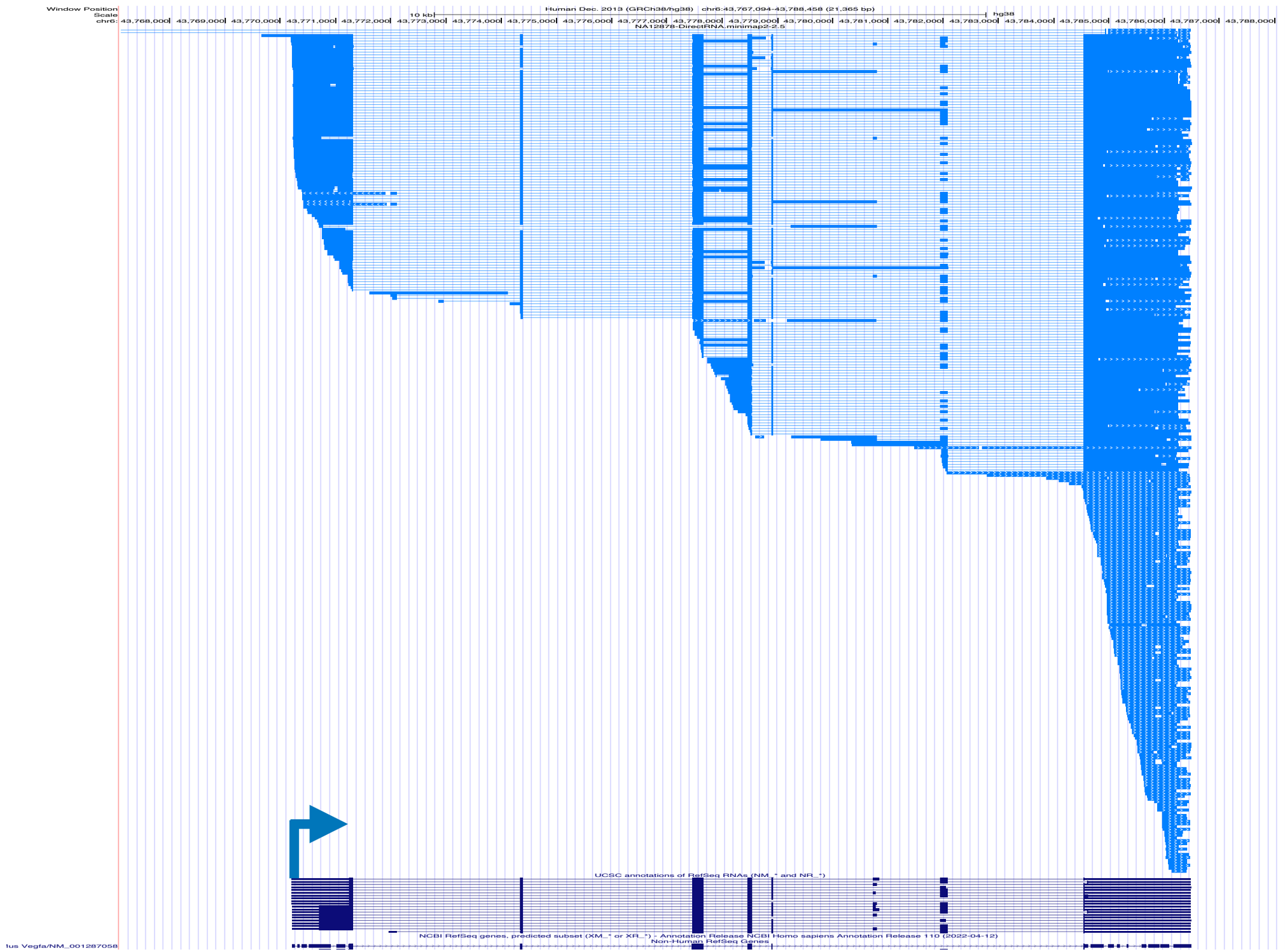
- 🚫 RNAs are fragmented
- 🚫 Assembly problem for both :
  - mRNAs and lncRNAs
  - novel isoforms of known genes and novel genes

Genome annotation with long-read RNAseq



- ✅ "No need to reassemble reads"
- ✅ Characterize and quantify "full length" transcripts (repeat-associated transcripts
- ✅ Direct exon connectivity
- ✅ Epitranscriptomics: identification of RNA modifications (m6A)

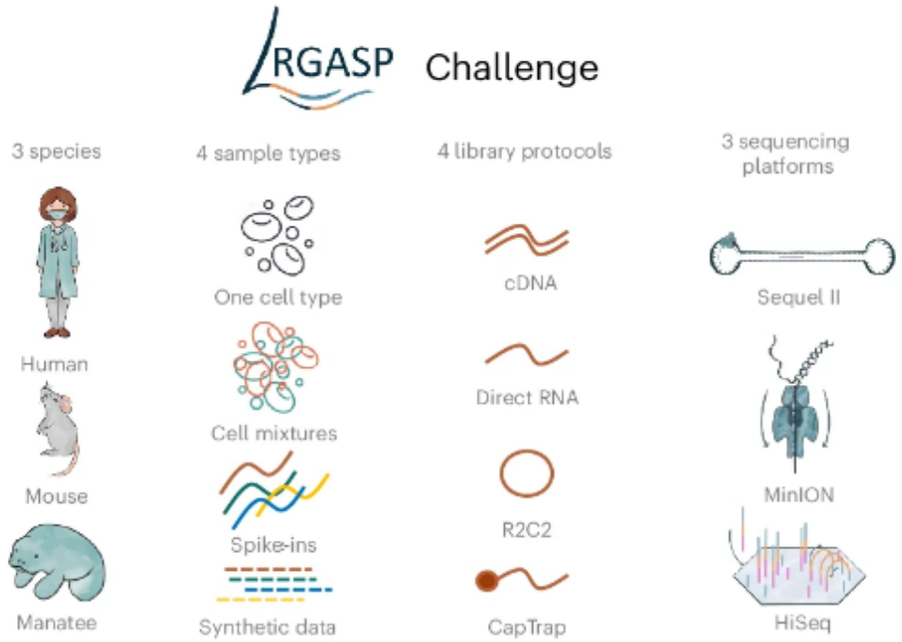
➡ Unfragmented view of the transcriptome



- LR-RNAseq is not totally long/full-length…

Sessegolo et al., 2019

Long-read RNA-seq Genome Annotation Assessment Project



Challenge 1: transcript isoform detection with a high-quality genome

Challenge 2: transcript isoform quantifcation

Challenge 3: *de novo* transcript isoform identifcation

nature methods

Explore content · About the journal · Publish with us

nature > nature methods > registered report > article

Registered Report | Open access | Published: 07 June 2024

Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

Francisco J. Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, Maite De María, Matthew S. Adams, Gabriela Balderrama-Gutierrez, Amit K. Behera, Jose M. Gonzalez Martinez, Toby Hunt, Julien Lagarde, Cindy E. Liang, Haoran Li, Marcus Jerryd Meade, David A. Moraga Amador, Andrey D. Prjibelski, Inanc Birol, Hamed Bostan, Ashley M. Brooks, Muhammed Hasan Çelik, Ying Chen, ... Angela N. Brooks + Show authors

Nature Methods **21**, 1349–1363 (2024) | Cite this article

◉Long-read methods

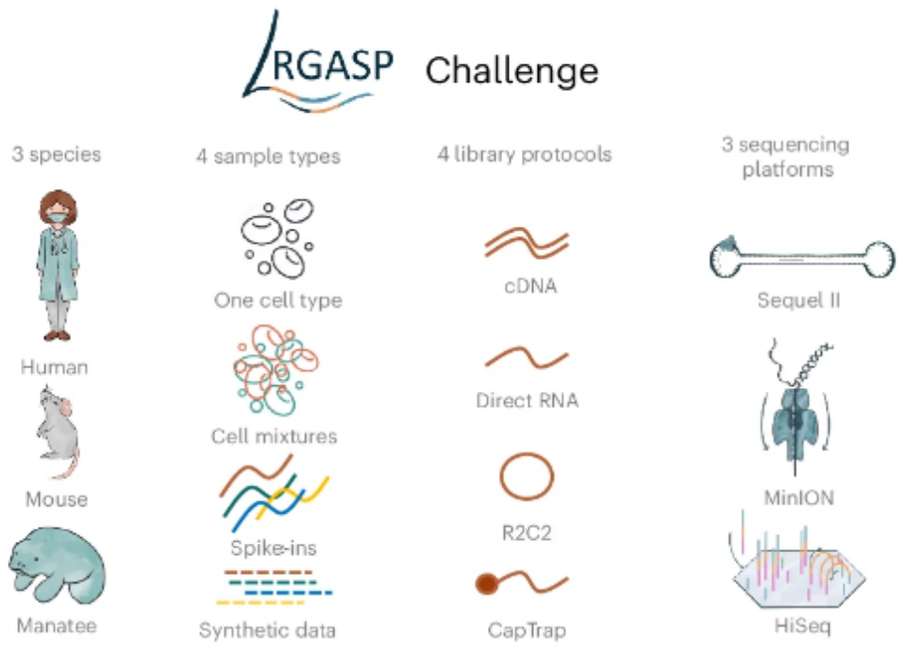
StringTie

Bambu

IsoQuant

...

Long-read RNA-seq Genome Annotation Assessment Project



Challenge 1: transcript isoform detection with a high-quality genome

Challenge 2: transcript isoform quantifcation

Challenge 3: *de novo* transcript isoform identifcation

nature methods

Explore content · About the journal · Publish with us

nature > nature methods > registered report > article

Registered Report | Open access | Published: 07 June 2024

Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

Francisco J. Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, Maite De María, Matthew S. Adams, Gabriela Balderrama-Gutierrez, Amit K. Behera, Jose M. Gonzalez Martinez, Toby Hunt, Julien Lagarde, Cindy E. Liang, Haoran Li, Marcus Jerryd Meade, David A. Moraga Amador, Andrey D. Prjibelski, Inanc Birol, Hamed Bostan, Ashley M. Brooks, Muhammed Hasan Çelik, Ying Chen, ... Angela N. Brooks + Show authors

Nature Methods **21**, 1349–1363 (2024) | Cite this article

🚫 *"Significant discrepancies among LR-RNAseq methods for transcript discovery."*

➡ Better multiple tools

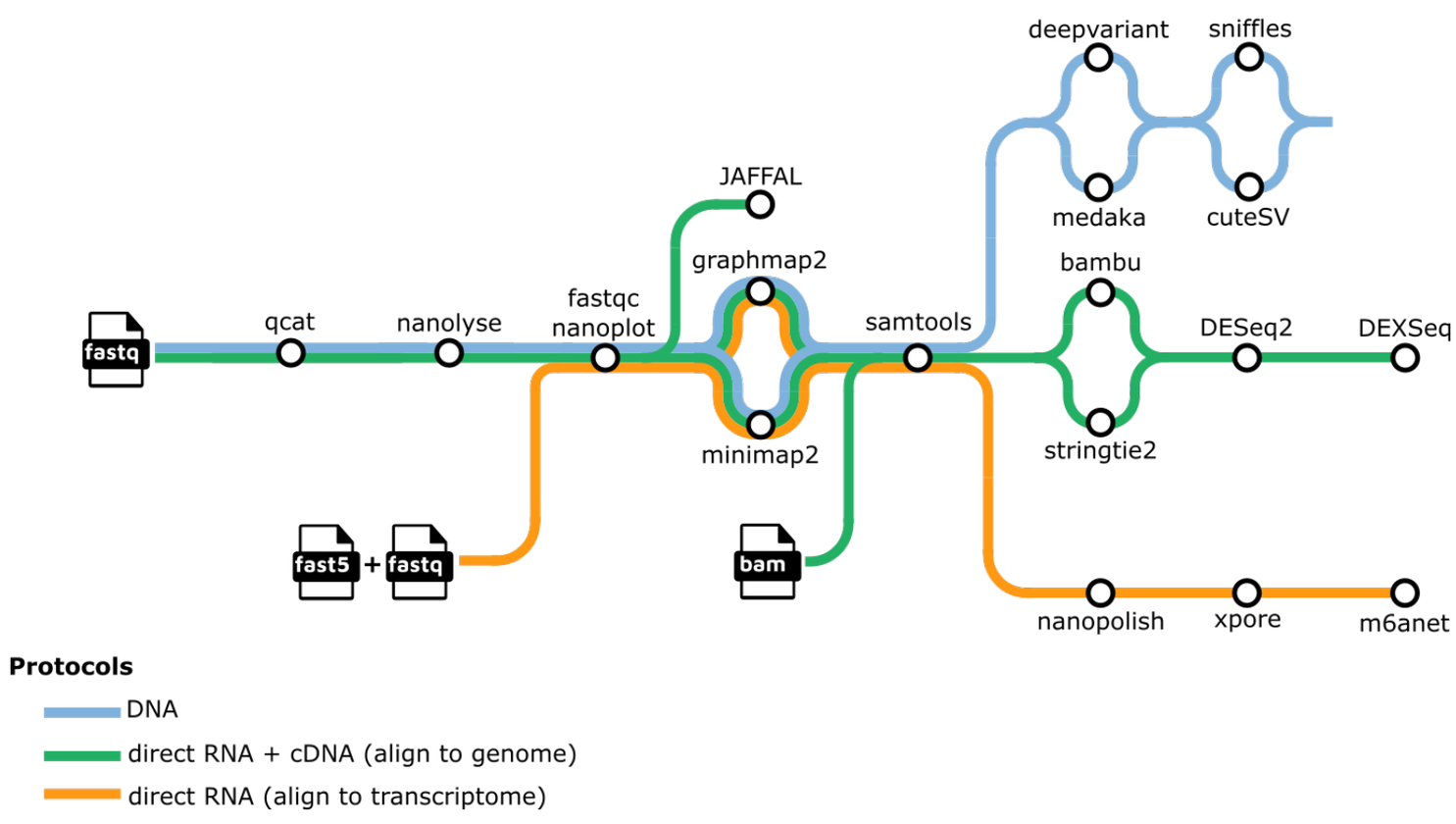
🚫 *No tool include lncRNA annotation (coding potential)*

Long-read RNA sequencing : state of the art

	Multiple tools for discovery	Annotate lncRNAs	Transcript classification	Full-length assessment	

Long-read RNA sequencing : state of the art

	Multiple tools for discovery	Annotate lncRNAs	Transcript classification	Full-length assessment	Links
nf-core/ nanoseq	✓ Bambu or Stringtie	✗	✗	✗	https://nf-co.re/nanoseq/



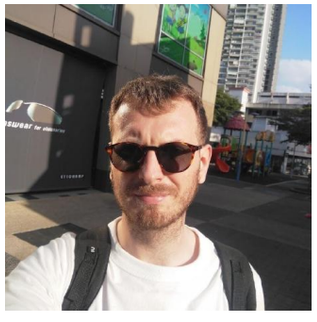
Long-read RNA sequencing : state of the art

	Multiple tools for discovery	Annotate lncRNAs	Transcript classification	Full-length assessment	Links
nf-core/ nanoseq	✓ Bambu or Stringtie	✗	✗	✗	https://nf-co.re/nanoseq/
wf-human-transcriptome	✗ Stringtie only	✗	✓ gffcompare	✗	https://github.com/epi2me-labs/wf-transcriptomes
TAGADA	✗ Stringtie only	✓ FEELnc	✓ FEELnc_classifier	✗	https://github.com/FAANG/analysis-TAGADA

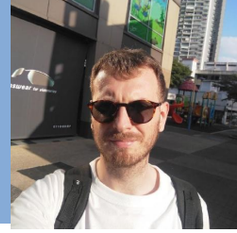
Long-read RNA sequencing : state of the art

	Multiple tools for discovery	Annotate lncRNAs	Transcript classification	Full-length assessment	Links
nf-core/ nanoseq	✓ Bambu or Stringtie	✗	✗	✗	https://nf-co.re/nanoseq/
wf-human-transcriptome	✗ Stringtie only	✗	✓ gffcompare	✗	https://github.com/epi2me-labs/wf-transcriptomes
TAGADA	✗ Stringtie only	✓ FEELnc	✓ FEELnc_classifier	✗	https://github.com/FAANG/analysis-TAGADA
ANNEXA	✓	✓	✓	✓	https://github.com/IGDRion/ANNEXA/

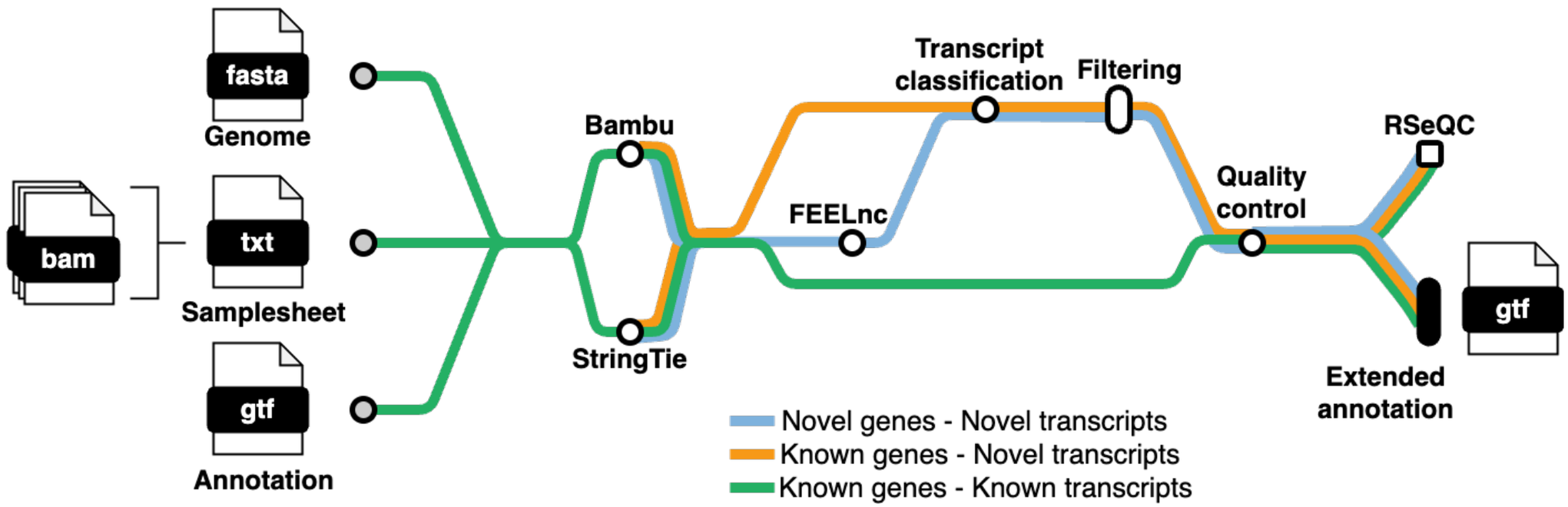
Nicolaï HOFFMANN



ANNEXA

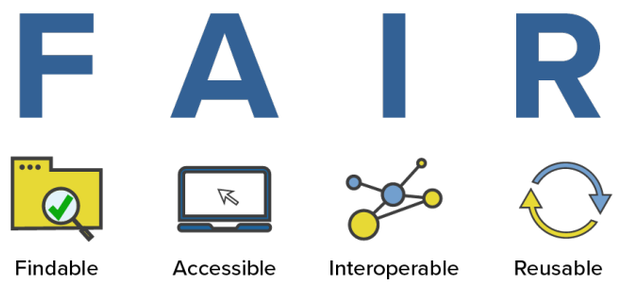


ANalysis with Nextflow for EXtended Annotation

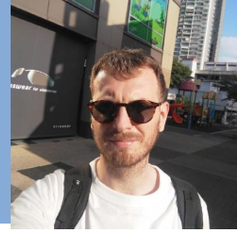


IGDRion/ANNEXA

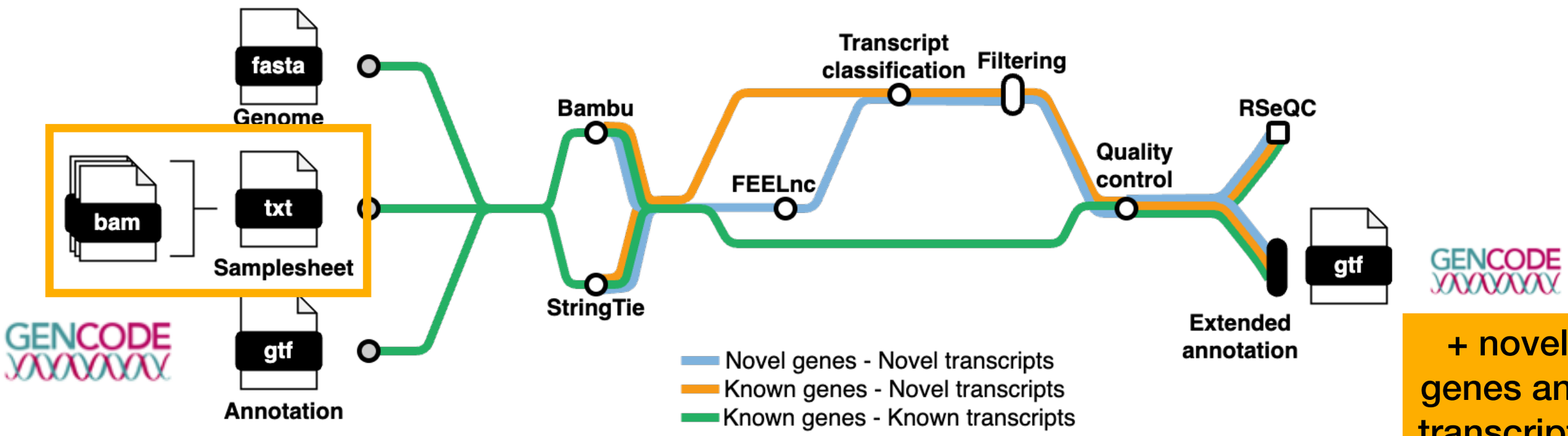
nextflow



ANNEXA



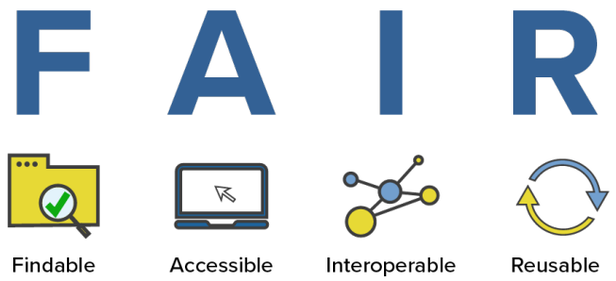
ANalysis with Nextflow for EXtended Annotation

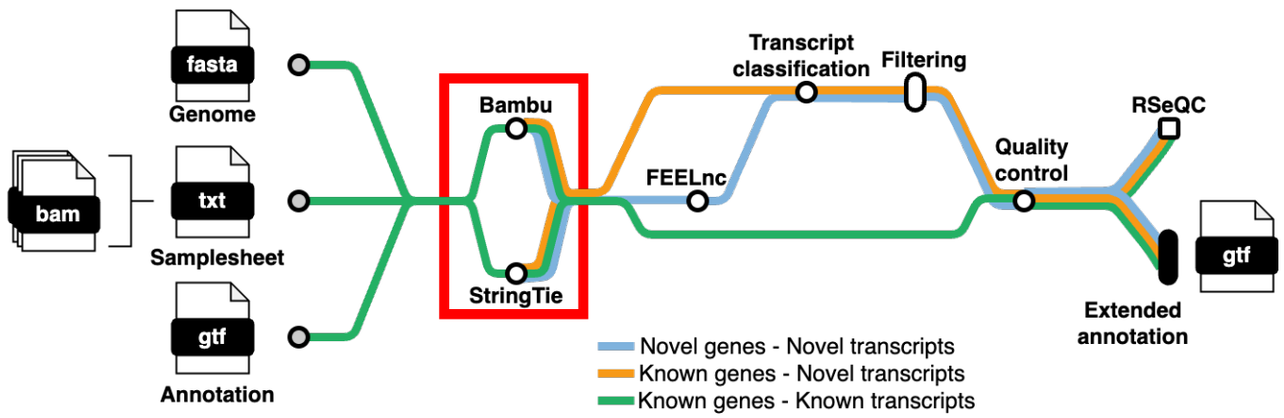


+ novel genes and transcripts

IGDRion/ANNEXA

nextflow

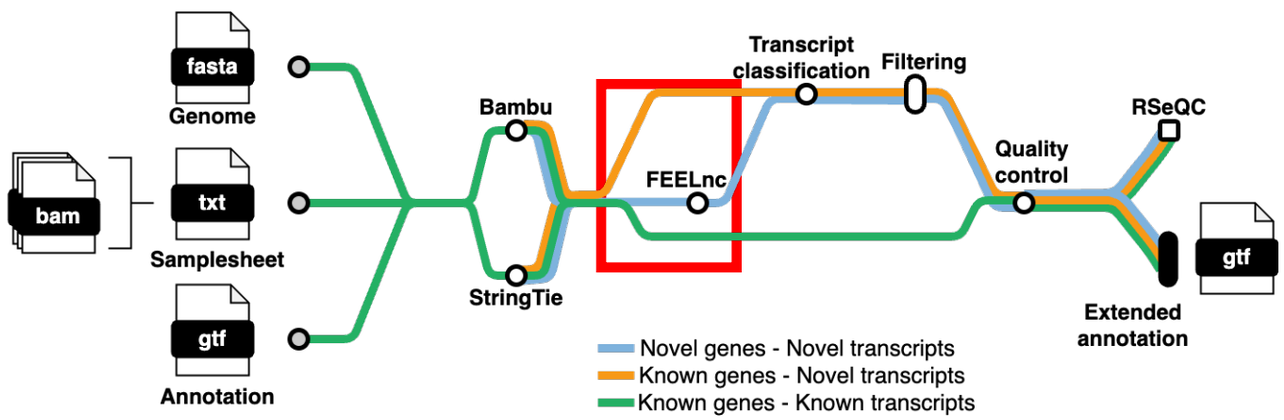




IGDRion/ANNEXA

- Bambu : better when reference annotation is more complete
- StringTie : better to identify novel isoforms

FEELnc: Wucher et al., 2017



IGDRion/ANNEXA

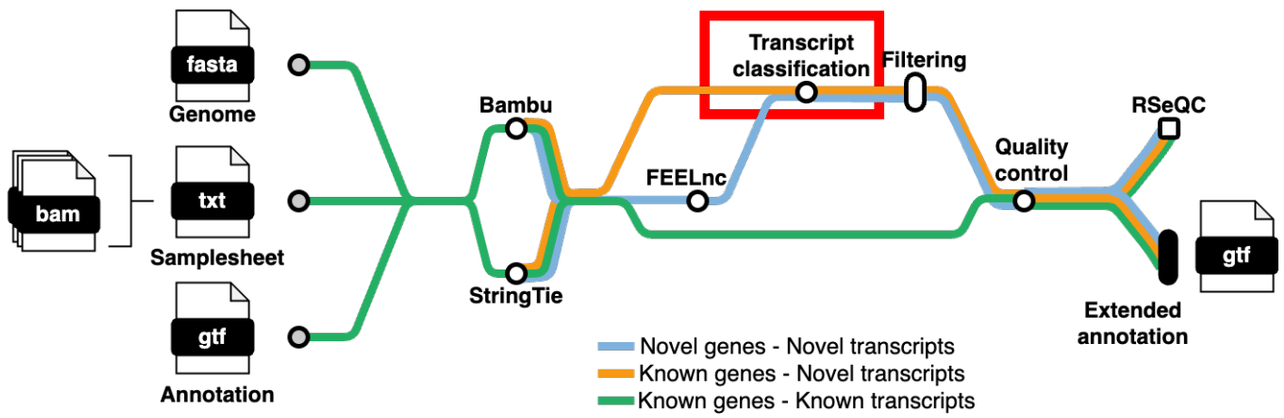
- **FEELnc** to compute the coding potential of all novel transcripts from novel genes

- **FEELnc** uses a RandomForest model trained with sequence features
  - ORF sizes
  - k-mer profiles
  - Transcript length

- **TransDecoder** allows CDS, 5'UTR and 3'UTR in novel mRNAs

➡ Allows to discriminate between novel mRNAs versus lncRNAs

gffCompare: Pertea G and Pertea M, 2020



IGDRion/ANNEXA

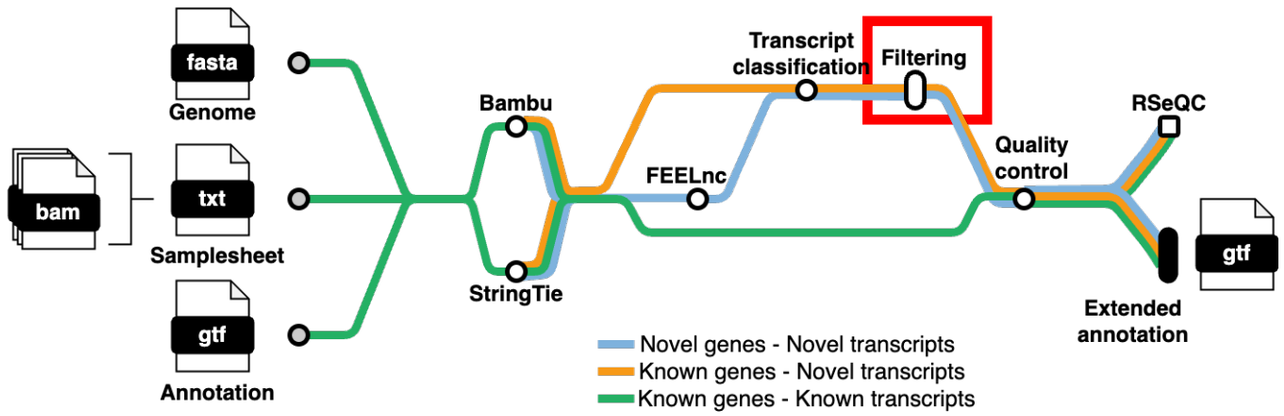
j	multi-exon with at least one junction match 
k	containment of reference (reverse containment) 
x	exonic overlap on the opposite strand (like ***o*** or ***e*** but on the opposite strand) 
u	none of the above (unknown, intergenic)

➡ Novel isoform of known gene

➡ Extension of known gene

➡ antisense lncRNA

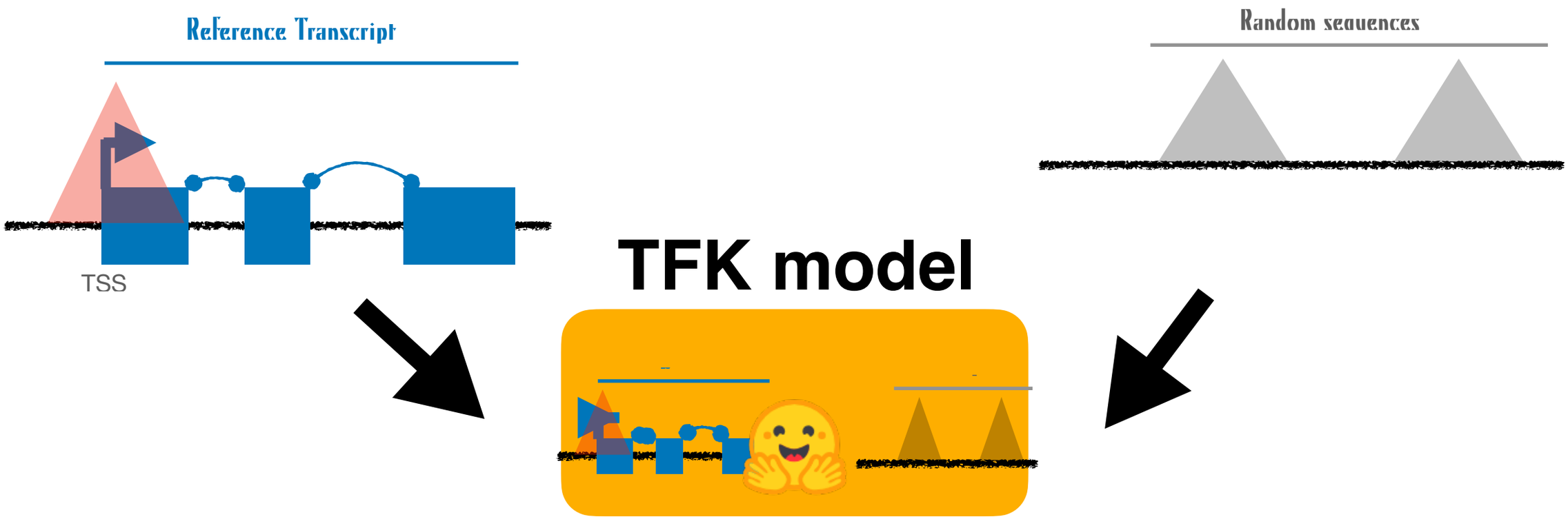
➡ Long intergenic ncRNA

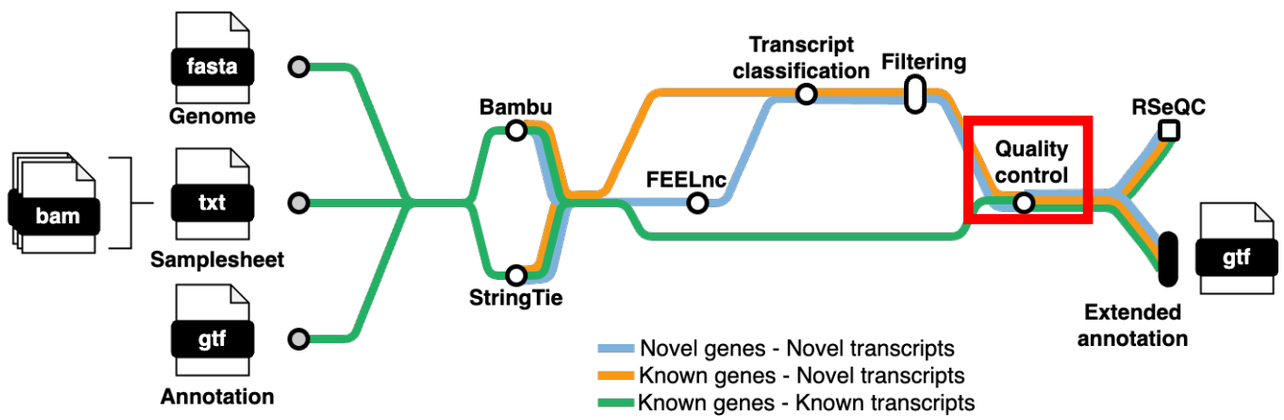


https://github.com/IGDRion/TransforKmers

IGDRion/ANNEXA

- Transformer-based model to classify DNA/RNA sequences.
  - Pretrain with species-specific genomic sequences (DNABERT)
  - Finetuning with CAGE-validated TSSs vs Random sequences

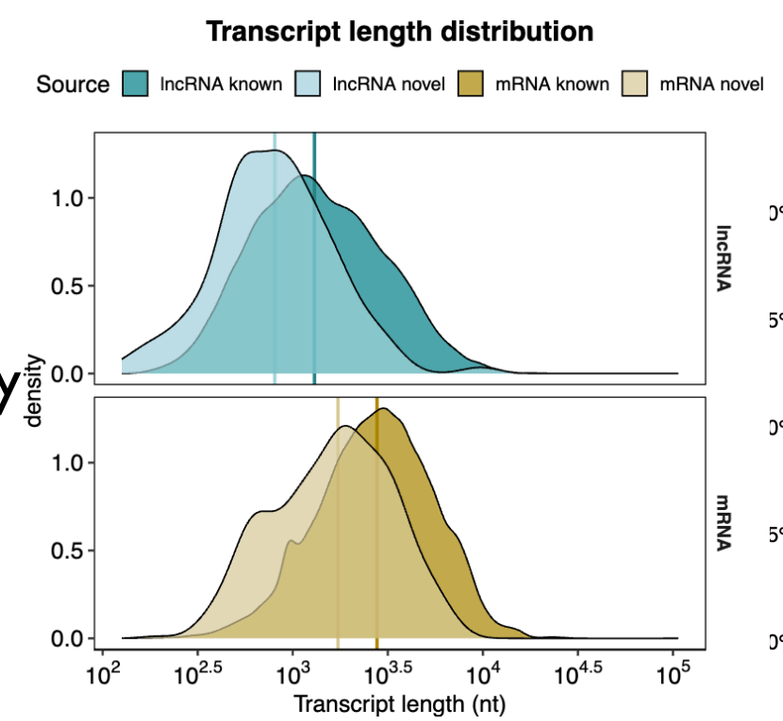


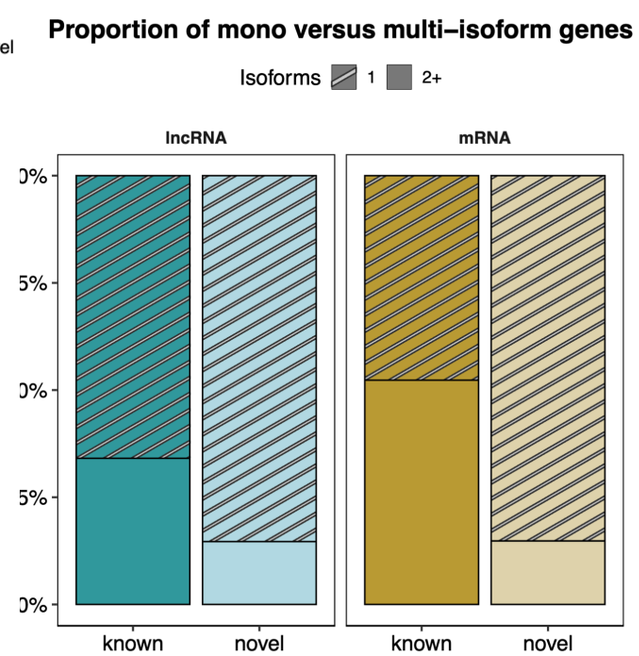


IGDRion/ANNEXA

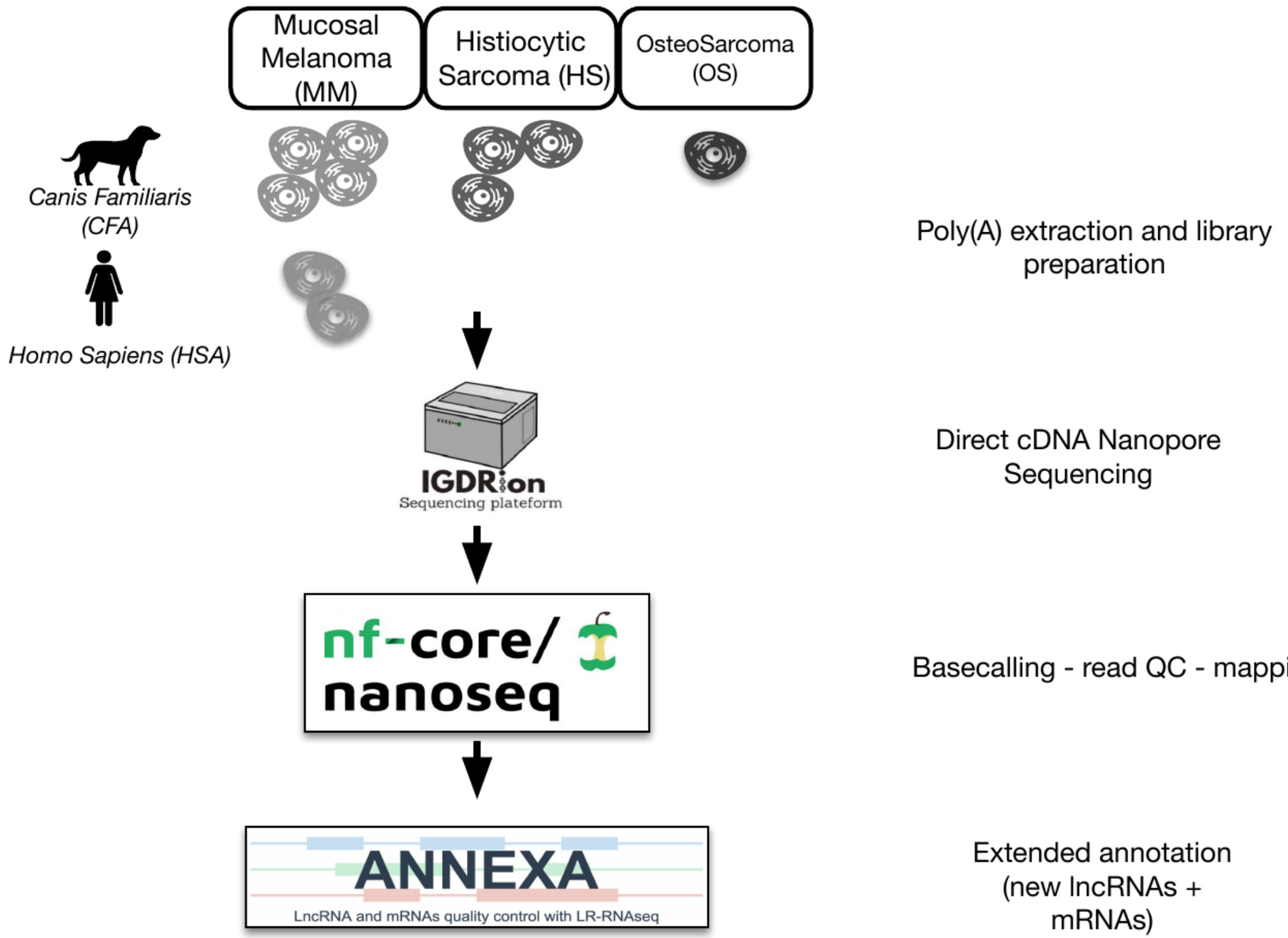
- Automatically generated .pdf report at 3 levels genes/transcripts/exons

- Multiple features :
  - Number of mRNAs versus lncRNAs
  - Number of single-exon genes
  - Breadth of expression /Tissue specificity
  - Gene extensions





➡ Allows to QC novel annotated genes/transcripts/exons


Mucosal Melanoma (MM)
Histiocytic Sarcoma (HS)
OsteoSarcoma (OS)
Canis Familiaris (CFA)
Homo Sapiens (HSA)
Poly(A) extraction and library preparation
IGDRion
Sequencing plateform
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)

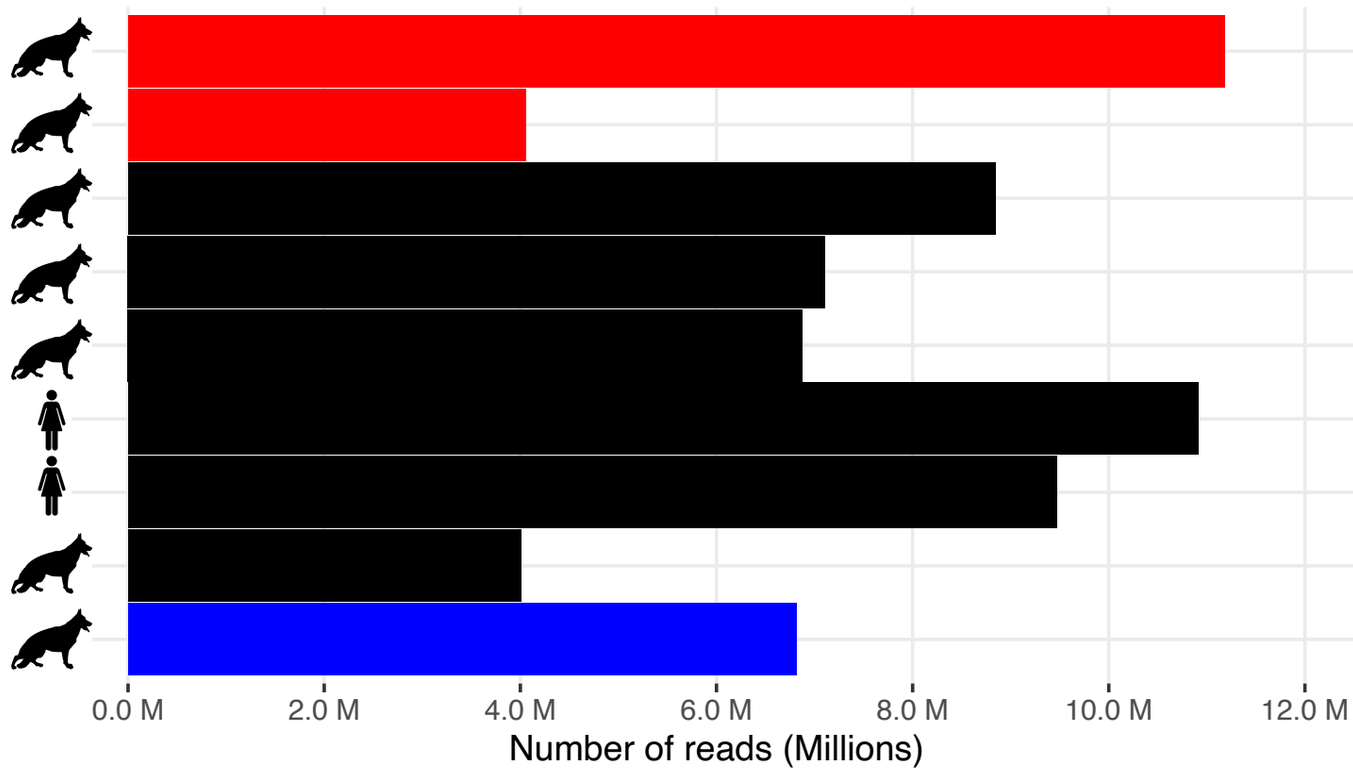
IGDRion : https://igdr.univ-rennes.fr/igdrion

INPUT SAMPLES

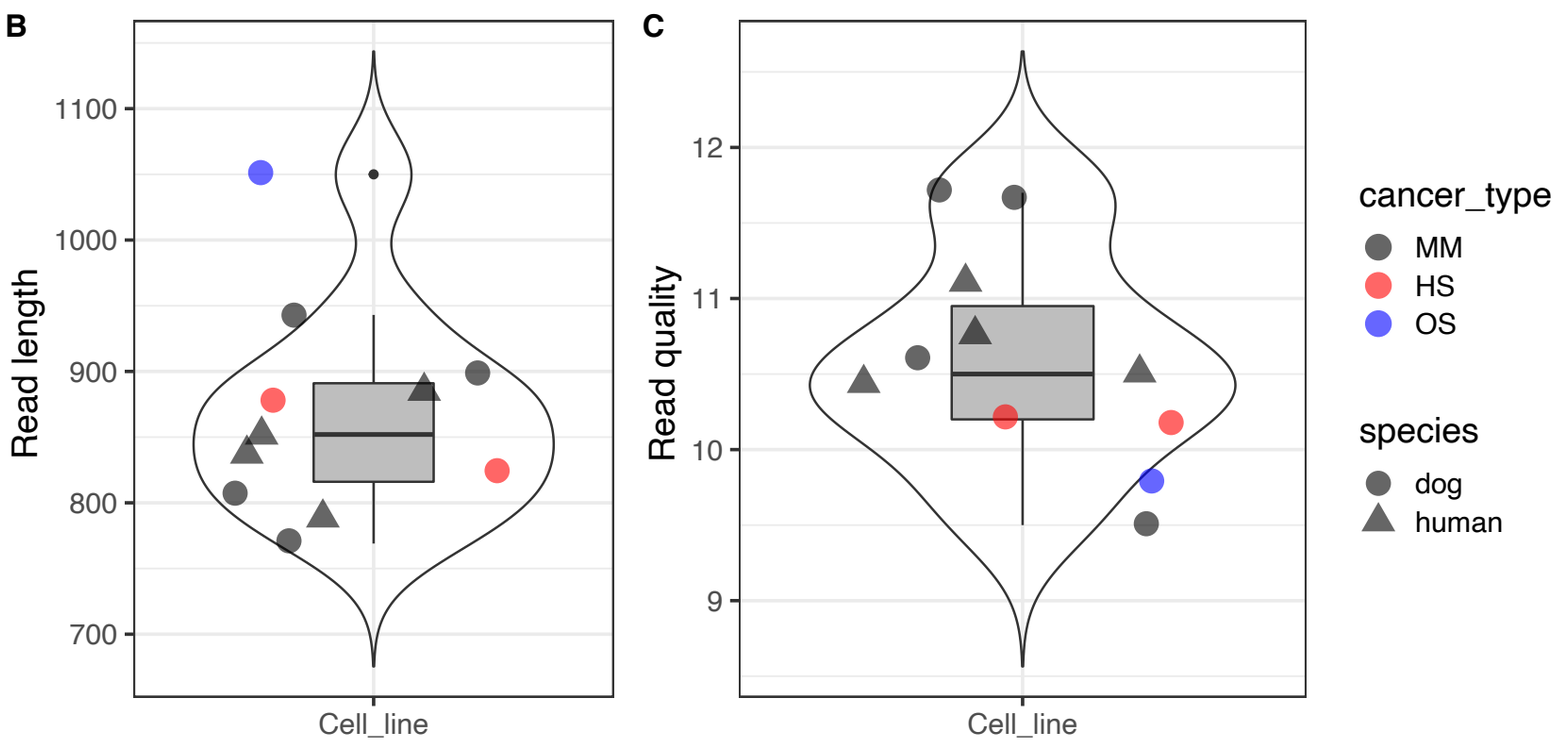
LIBRARY PREP.

SEQUENCING

BIOINFORMATIC ANALYSES



➡ ~6M reads
➡ Variable number of reads per sample



➡ length~900nt
➡ Qscore ~10.5 (old chemistry)

Reference genome :

canFam4

Reference genome :

GRCh38

Reference annotations (x3) :

Ensembl (Dyer et al., 2025)

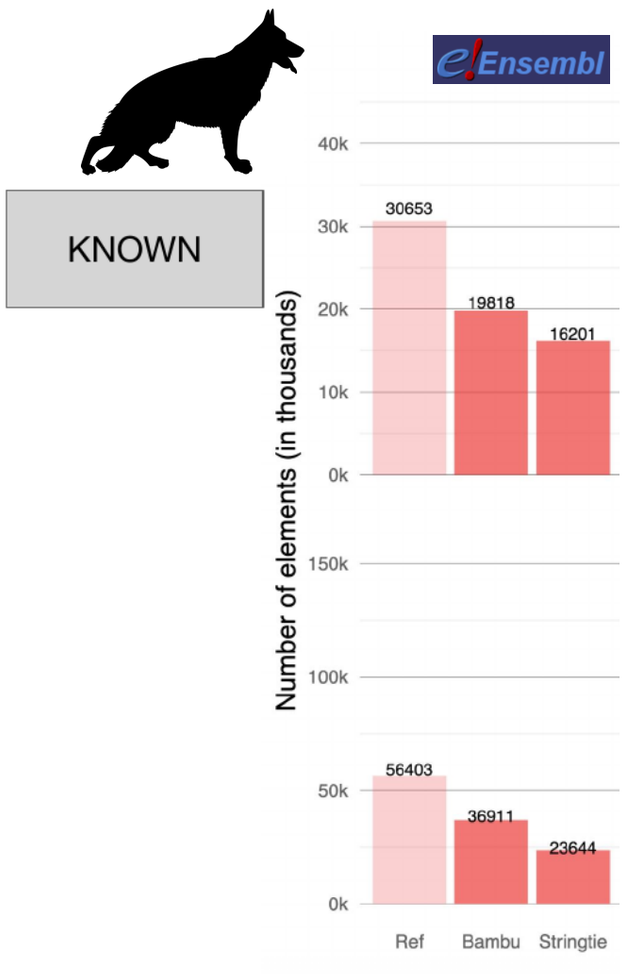
Refseq (O'Leary et al., 2016)

Uppsala Univ (Wang et al., 2021)

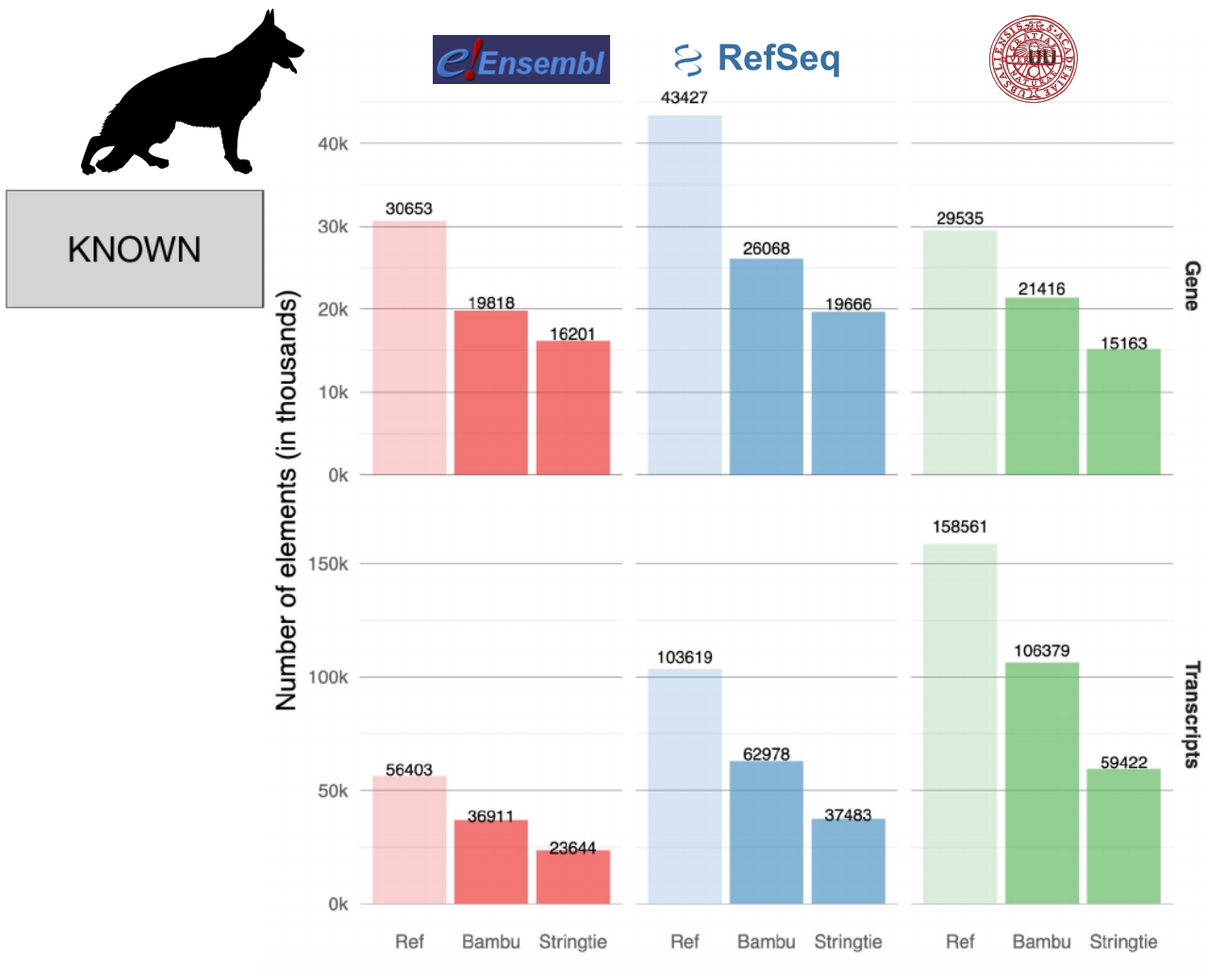
Reference annotations (x2):

Gencode (Mudge et al., 2025)

CHESS (Varabyou et al., 2023)

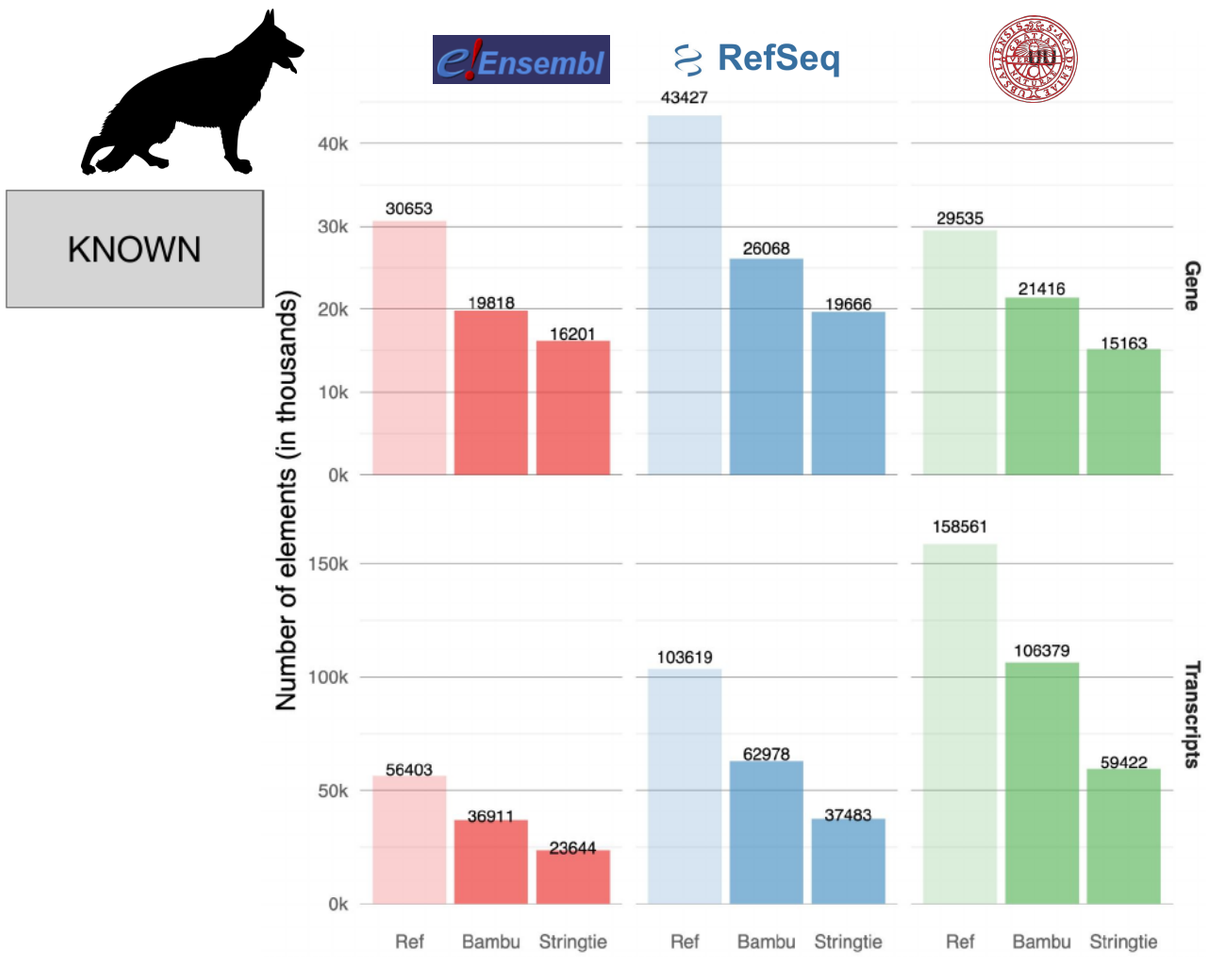


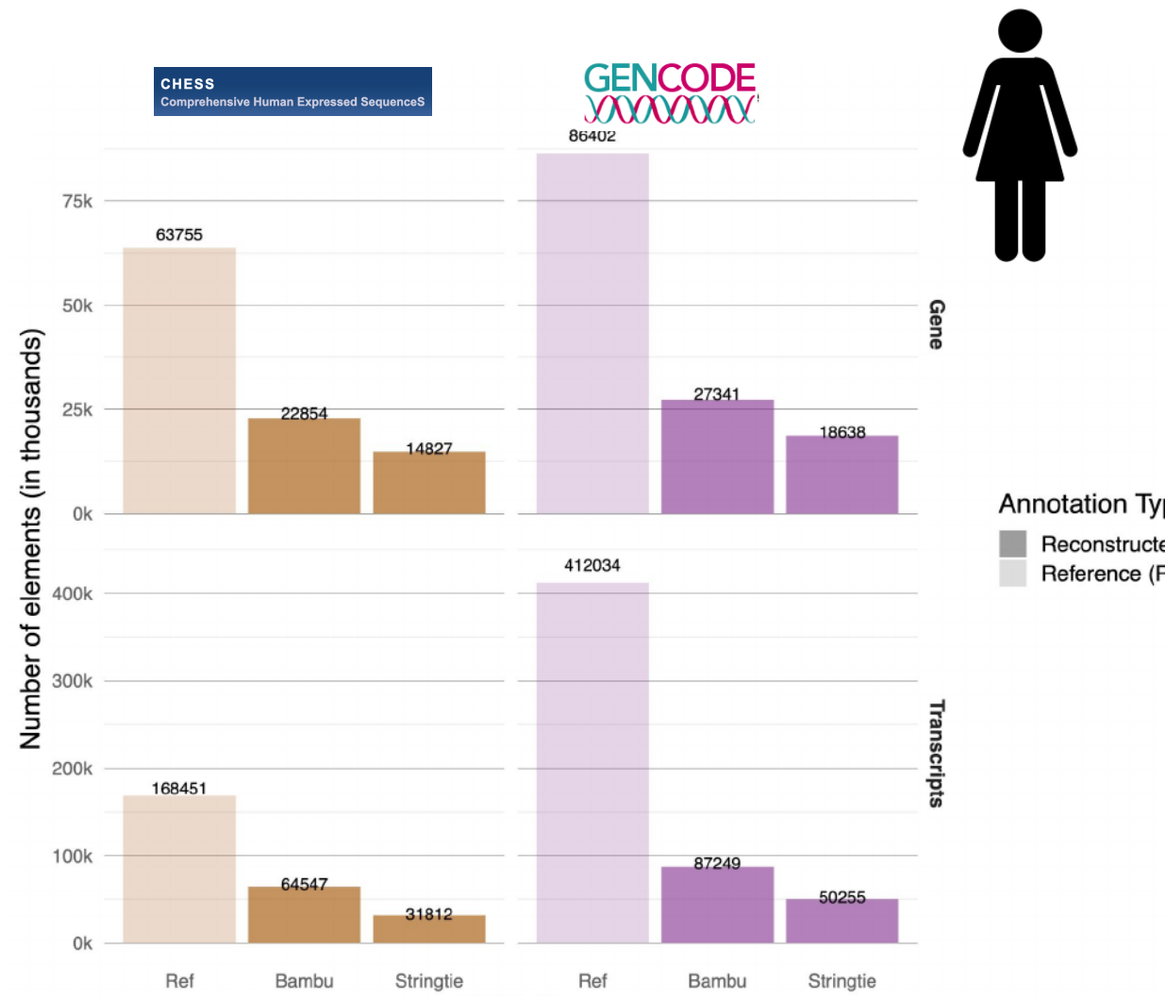
➡ Bambu reconstructs more **known** genes/transcripts that Stringtie consistently



➡ Bambu reconstructs more **known** genes/transcripts that Stringtie consistently

➡ Consistent pattern across the canine annotation



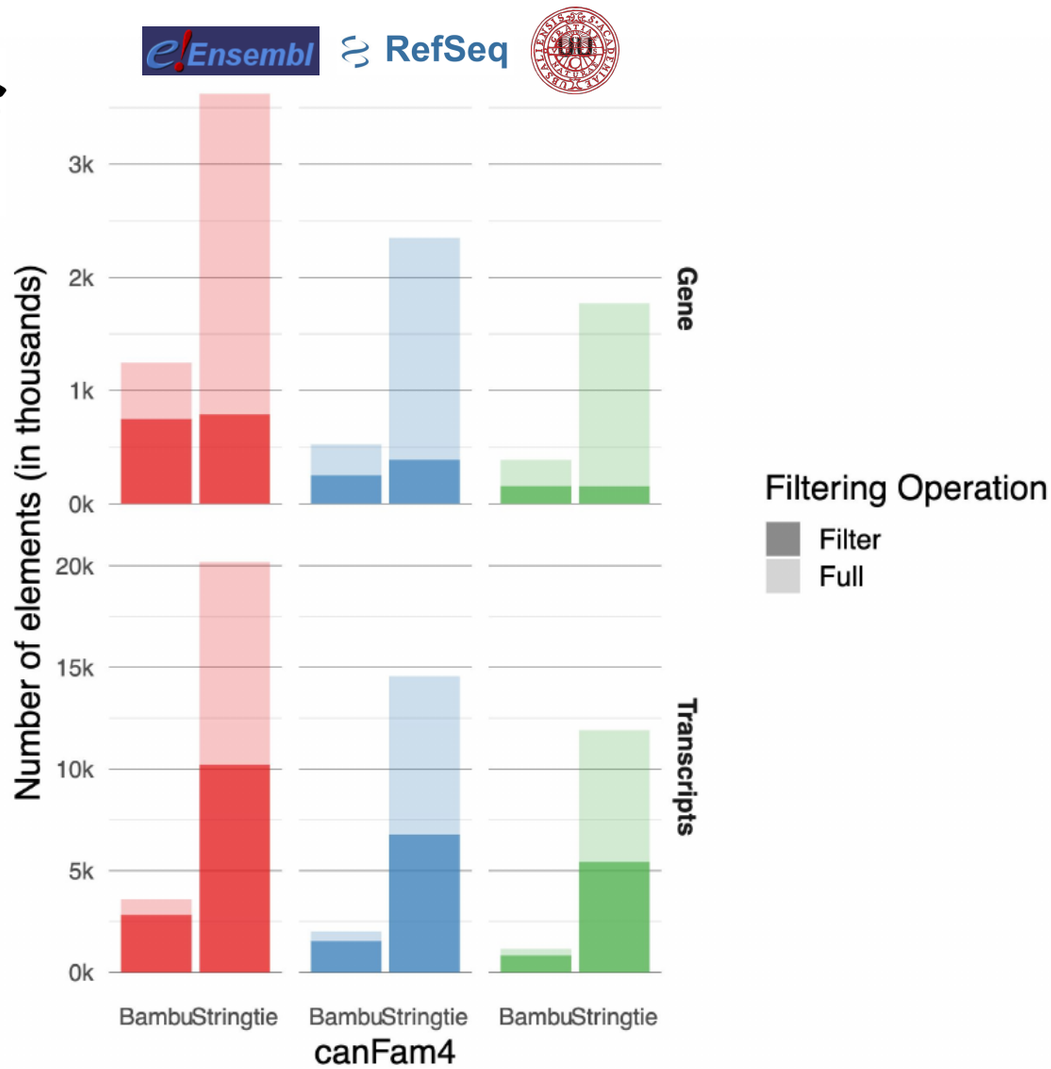


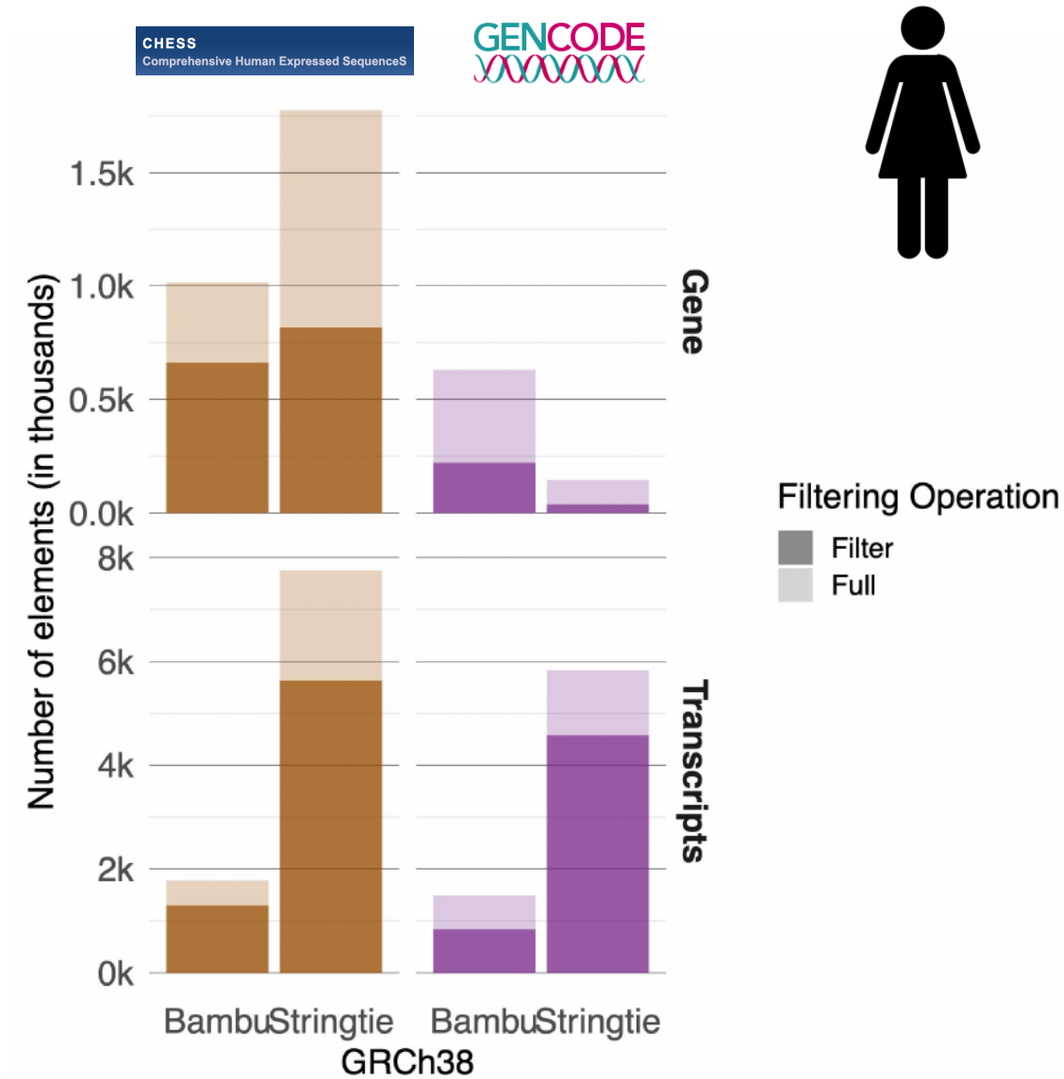
➡ Bambu reconstructs more **known** genes/transcripts that Stringtie consistently

➡ Consistent pattern across the canine annotations

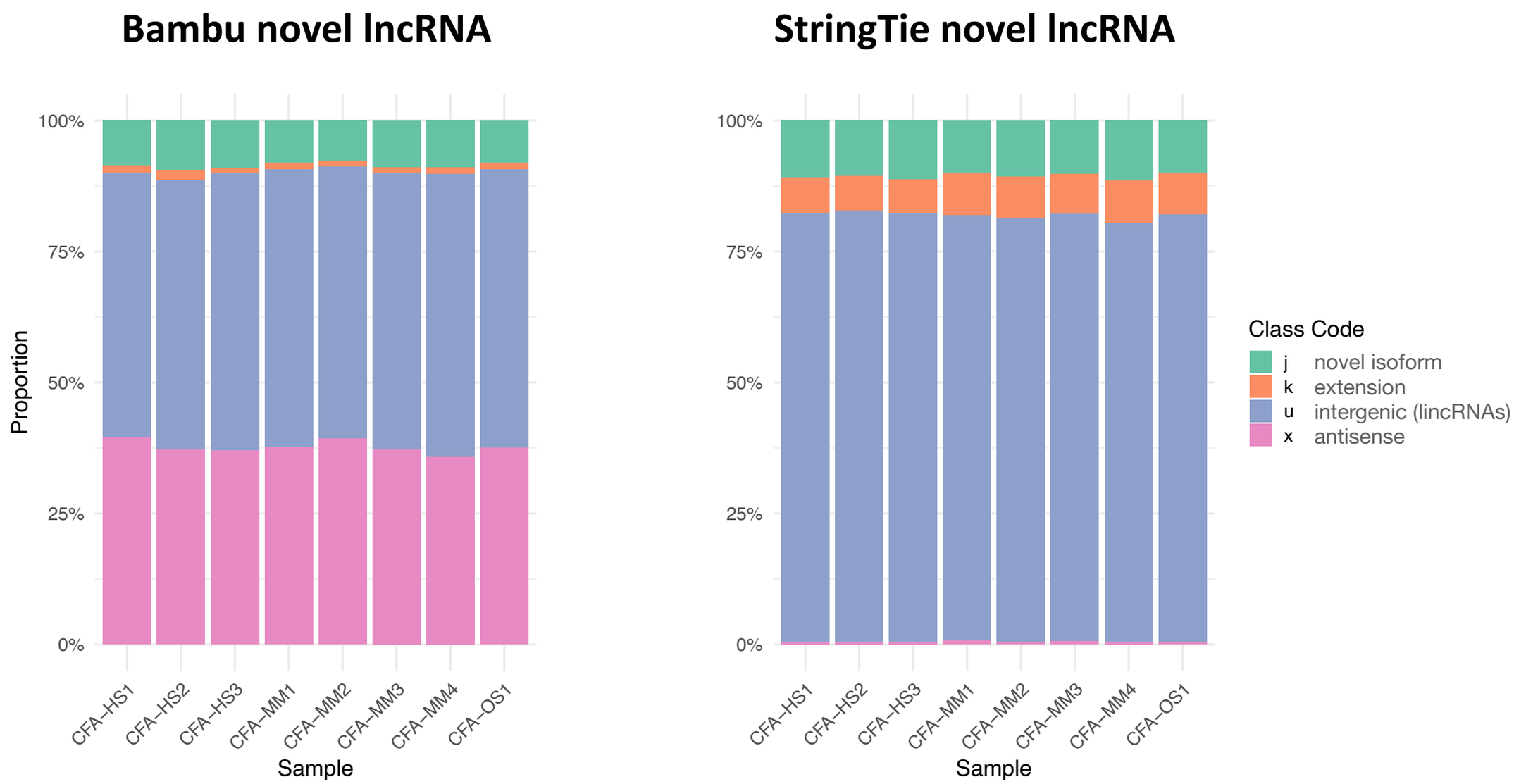
➡ As well as human annotations (although with lower rates)

NOVEL



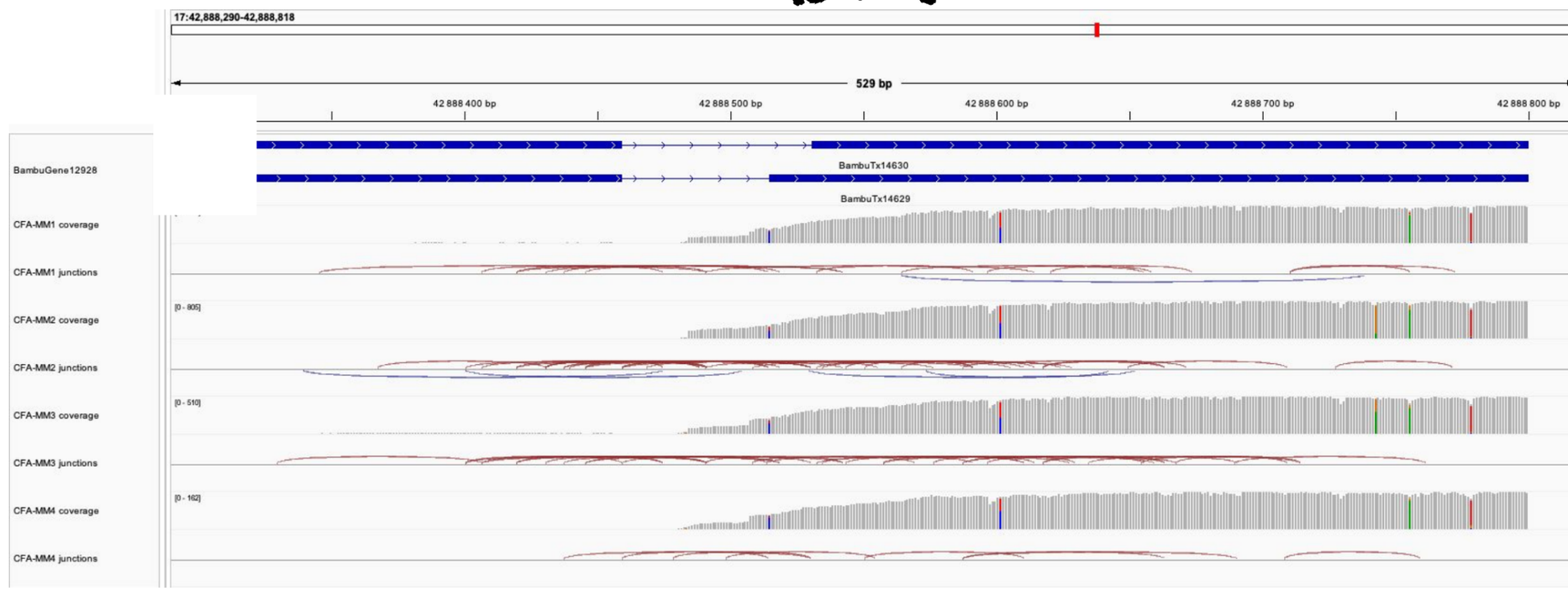


➡ StringTie reconstructs more **novel** elements than Bambu
➡ Filtering process dramatically reduces number of novel elements



Different patterns in classification:

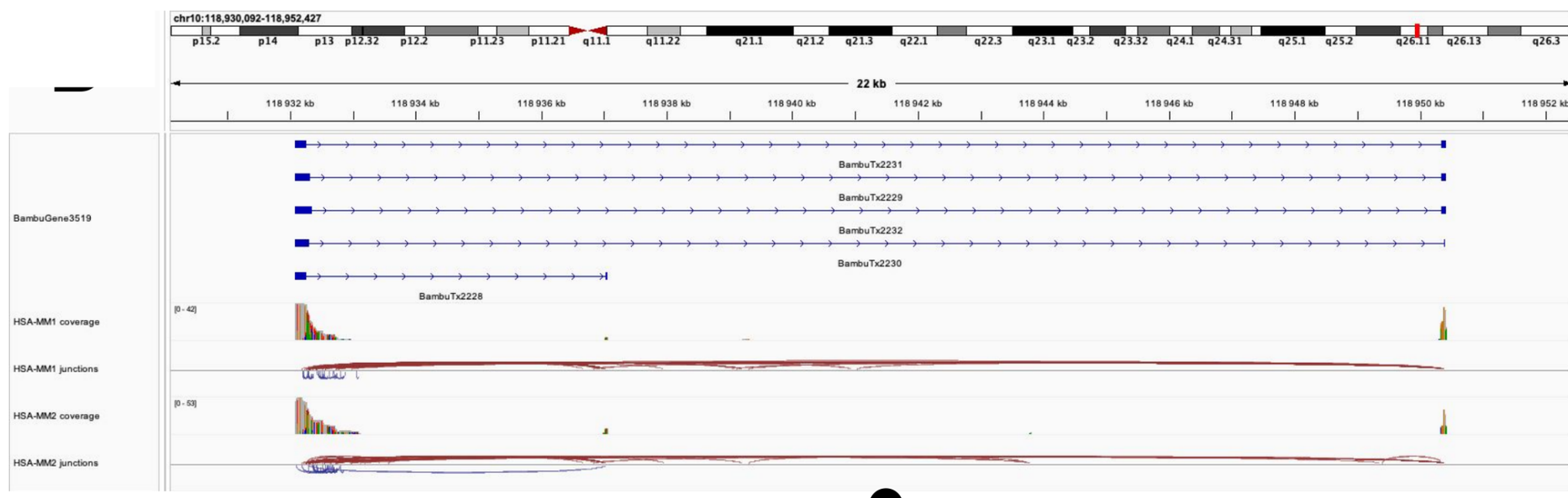
- **Bambu**: primarily intergenic (u) and antisense (x)
- **Stringtie**: intergenic (u) and extension of known lncRNAs (k), very few antisense



=> 5 novel genes ortholog bw human and dog (3 lncRNAs)

=> MM-specific

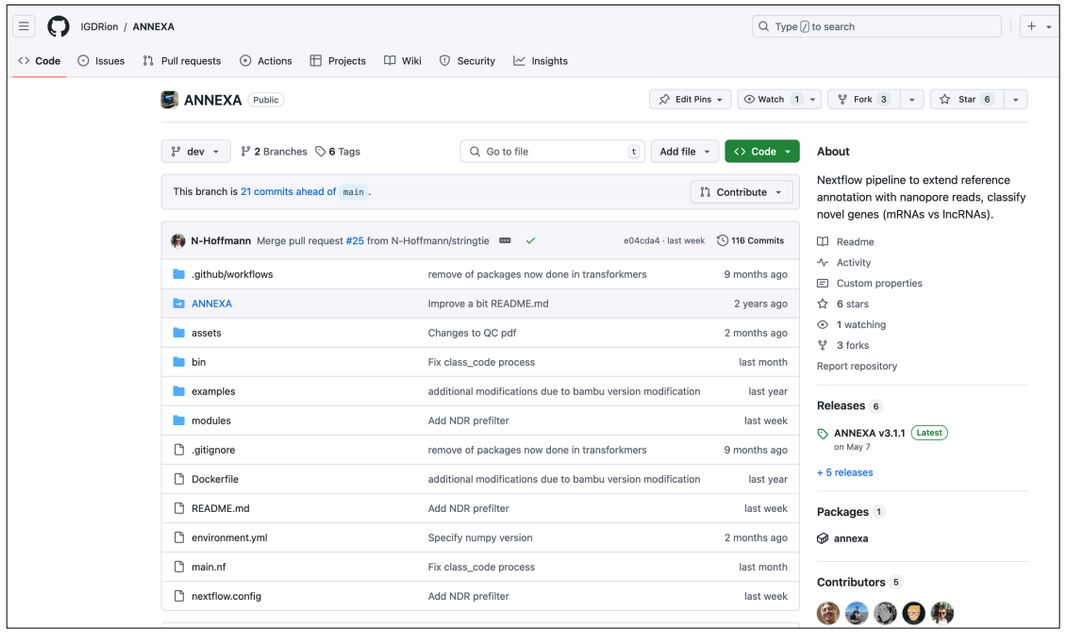
=> Potential candidates for future functional validation/biomarker



Conclusions

- Development of a new tool to process long-read RNA seq data and extend existing annotations in a **context-aware manner**.

- ANNEXA is **FAIR and freely available** (https://github.com/IGDRion/ANNEXA/)



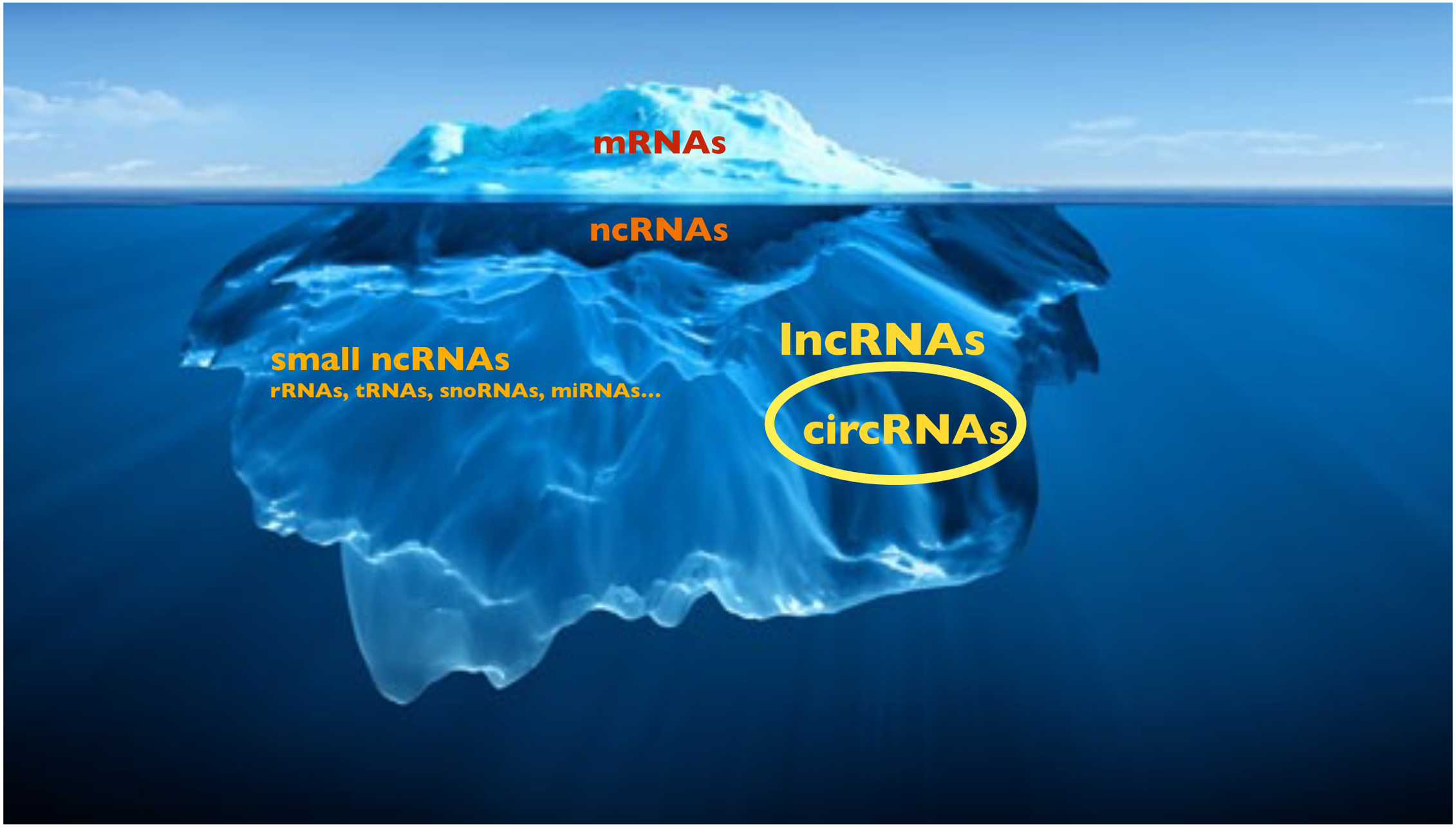
- Applied ANNEXA on pilot comparative oncology long-read RNA seq, highlighting tool specific annotation of lncRNA biotypes

MSD AVENIR
améliorer la Vie ENsemble
par l'Innovation et la Recherche

➡ 100 dog and 30 human MM samples

The cell as an RNA machine

Anastasia
RUSAKOVICH

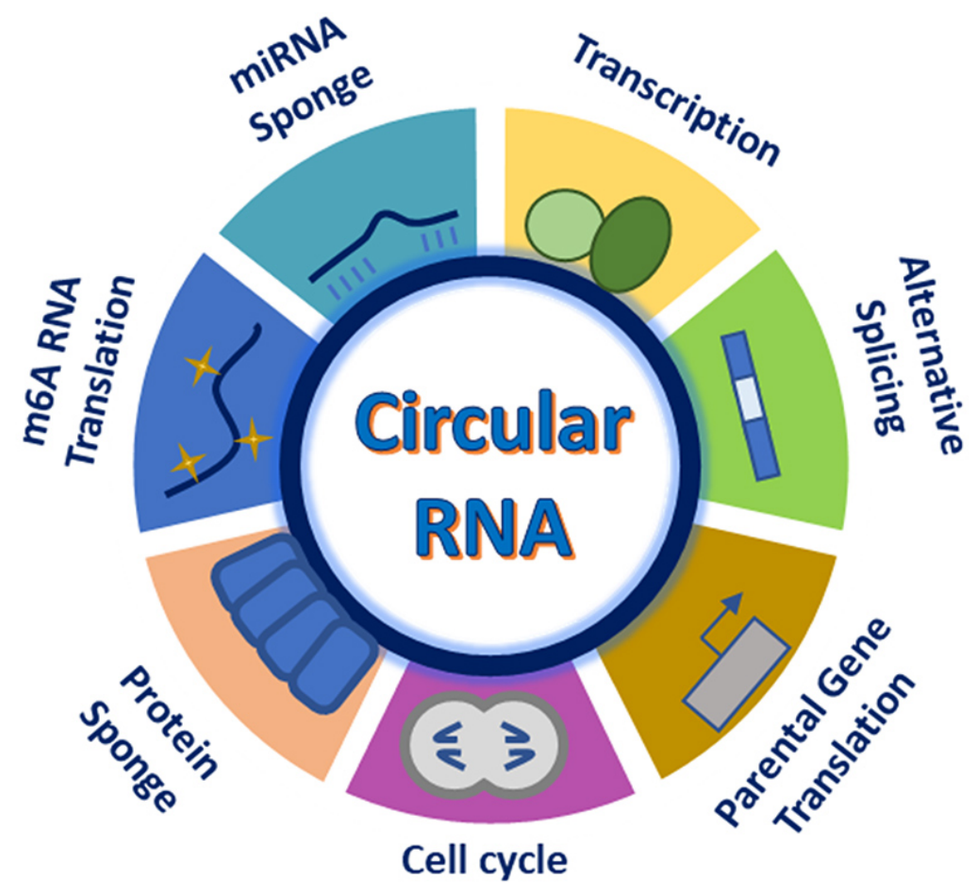


What are circRNA?

- **circRNAs** are a type of non-coding RNA forming closed-loop structures

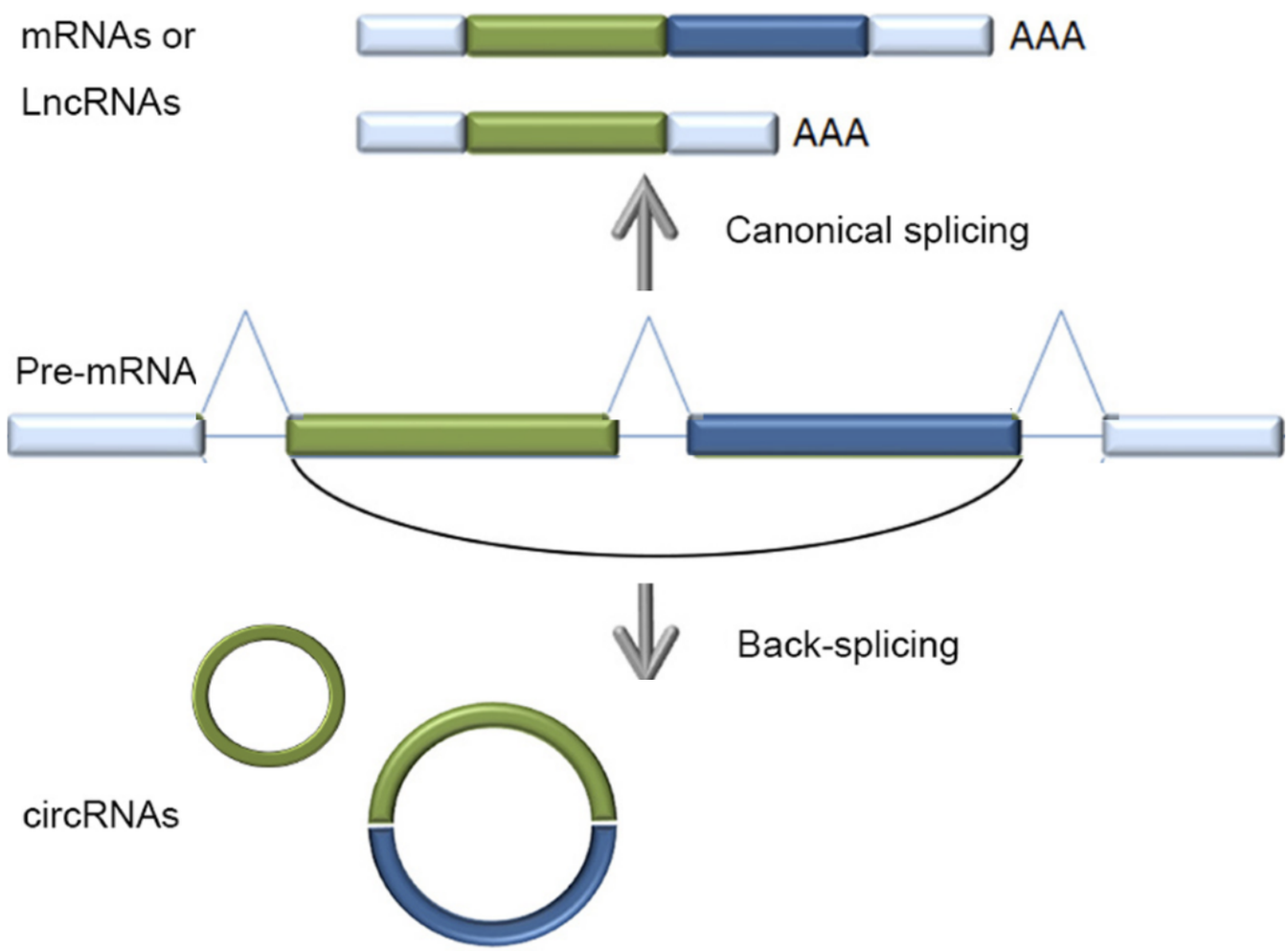
What are circRNA?

- **circRNAs** are a type of non-coding RNA forming closed-loop structures
- Highly **stable**, often **tissue-specific**, **potential disease biomarkers**
- Interest is growing due to their **roles in gene regulation and cancer**



How are they generated?

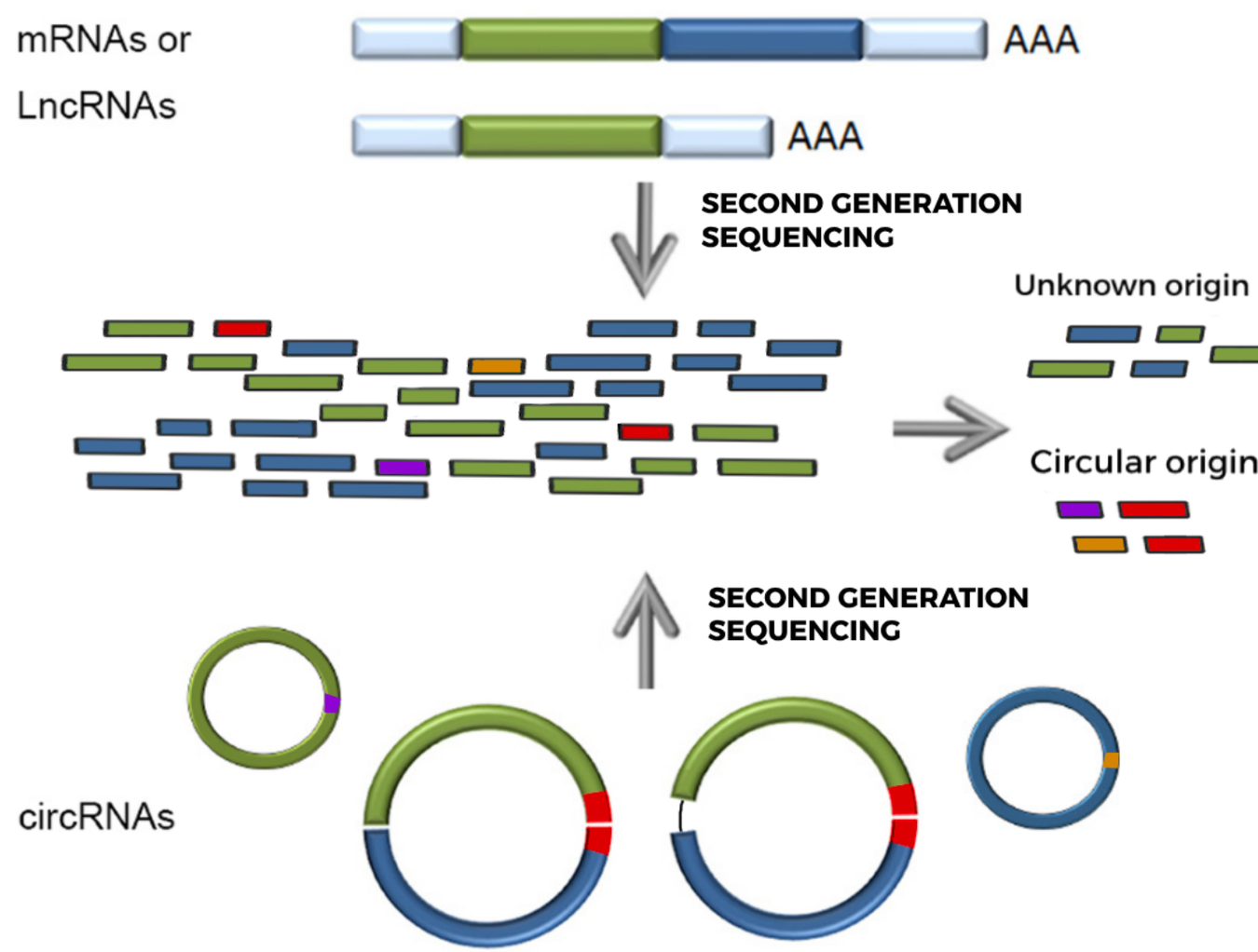
- Produced by **back-splicing**, a **non-canonical splicing event**
- **Multiple isoforms** can arise from the same linear RNA



Yu, CY., Kuo, HC. The emerging roles and functions of circular RNAs and their generation. *J Biomed Sci* 26, 29 (2019)

How are they generated?

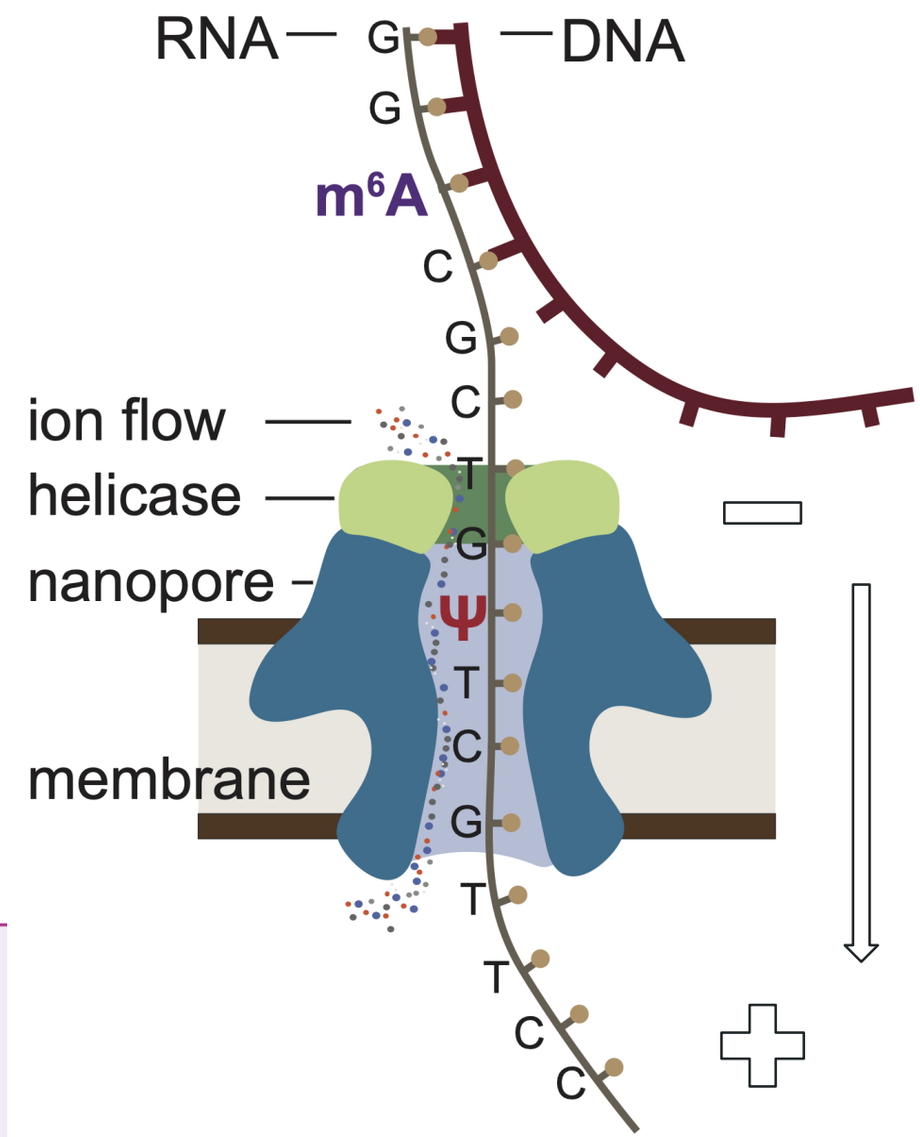
- Produced by **back-splicing**, a **non-canonical splicing event**
- **Multiple isoforms** can arise from the same linear RNA
- **Short-read sequencing struggles** to reconstruct full isoforms

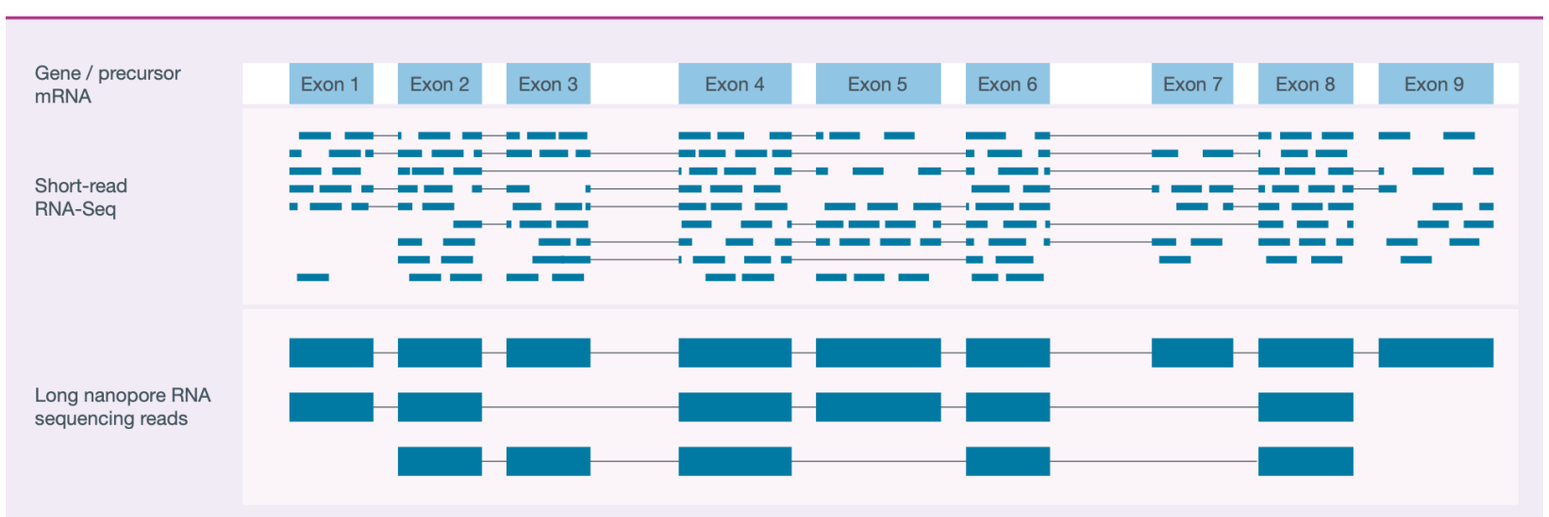


Yu, CY., Kuo, HC. The emerging roles and functions of circular RNAs and their generation. *J Biomed Sci* 26, 29 (2019)

Why long-read sequencing?

Oxford Nanopore allows sequencing of **full-length** molecules, capturing **internal structure** and **alternative splicing events**

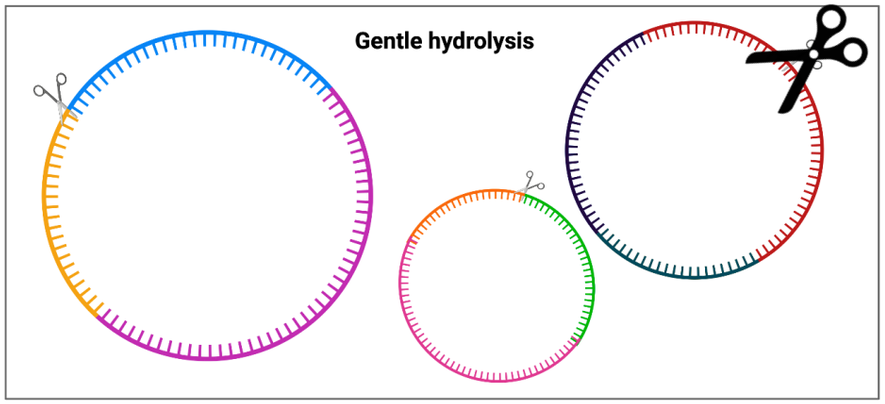


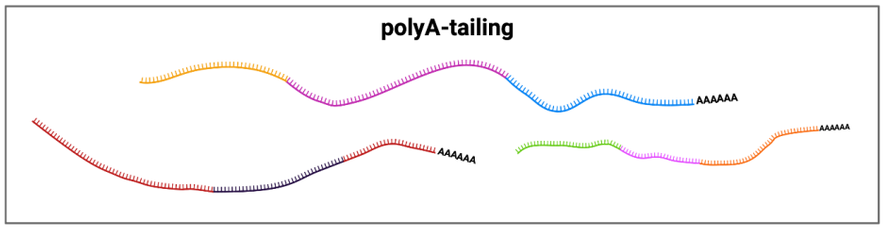


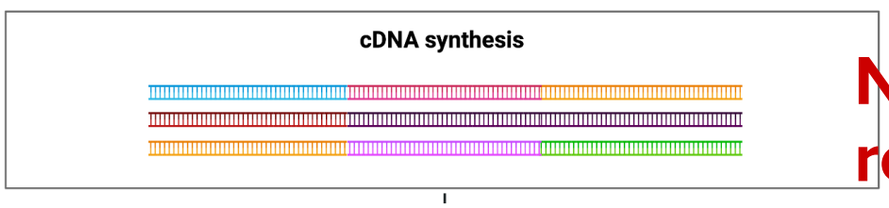
IGDRion
Sequencing plateform

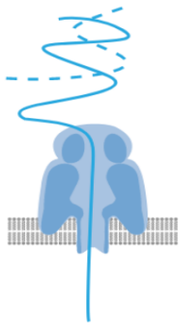
Protocols for linearization of circRNAs

Fragmentation based approach



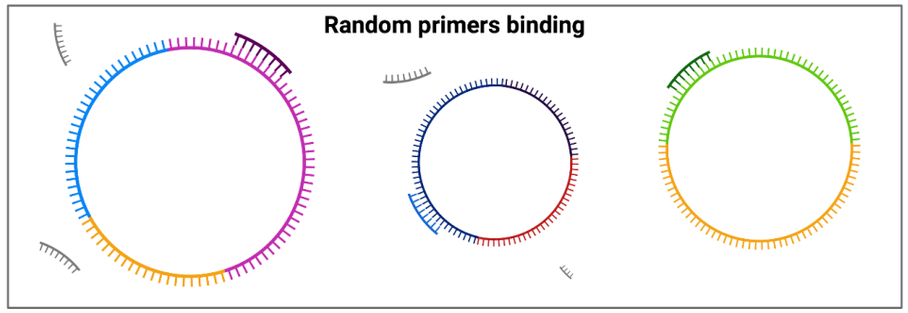


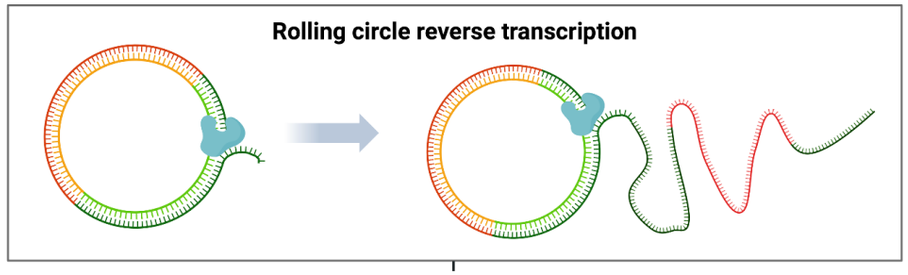


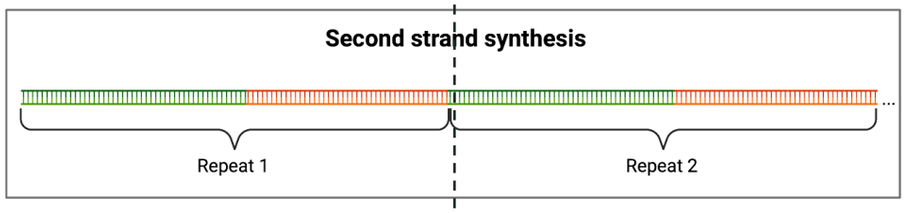


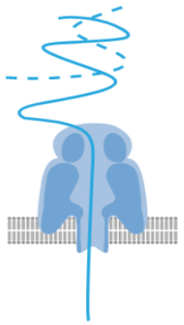
No repeats!

Rolling circle based approach









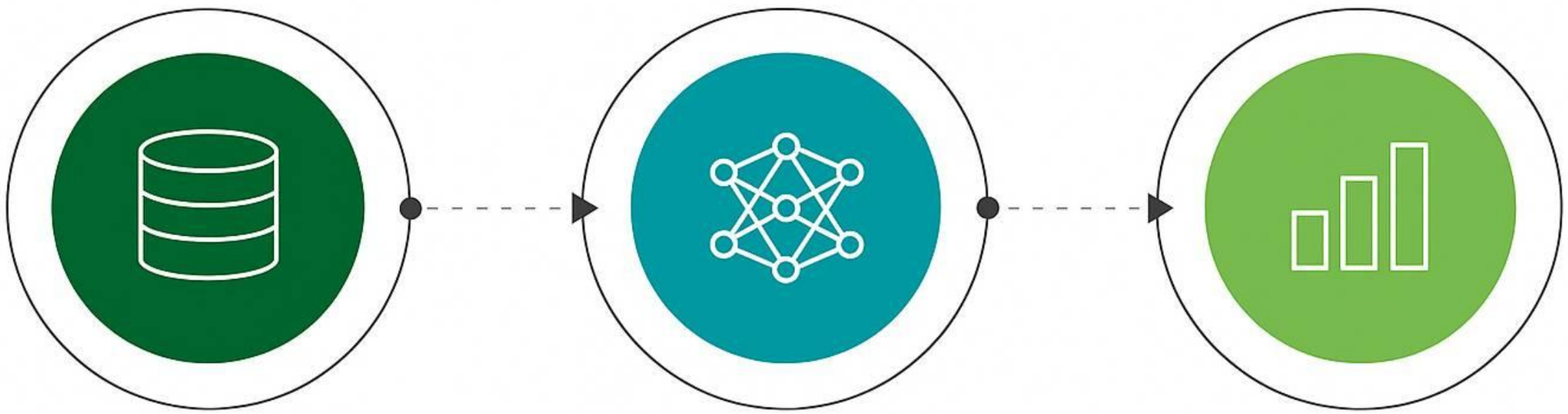
Have repeats!

Existing bioinformatic tools for circRNA long-read data

Step \ Tool	circNICK-lrs	isoCirc	CIRI-long
Expected Input	Fragmentation	Rolling circle amplification	Rolling circle reverse transcription
1 - Repeat Detection	✂	Tandem Repeat Finder	k-mer approach (k=8,11)
2 - Mapping	PBLAT	minimap2	minimap2, SPOA, mappy, bwapy
3 - BSJ detection -Iterative **alignment** -Multi-tiered scoring -Spatial constraints -Splice site usage -Database **reliance**	✘ ✘ ✓ ✘ ✓ (3 databases)	✘ ✓ ✓ ✓ ✘ (only classif 2 db)	✓ ✘ ✘ ✓ ✓ (for BSJ correction,1 db)

→ No comprehensive benchmark study comparing the bioinformatic tools developed for ONT sequencing data analysis

Our benchmark study



circRNA sequences generation

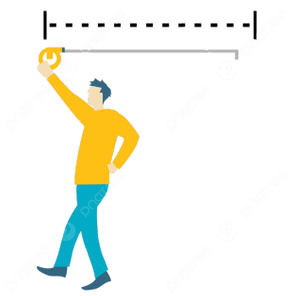
Oxford Nanopore reads generation

Tool application and evaluation

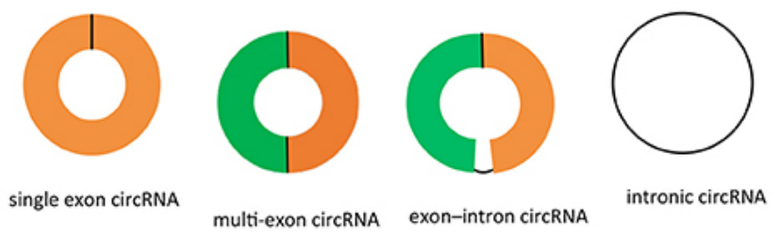
Requirements for simulation

Database features

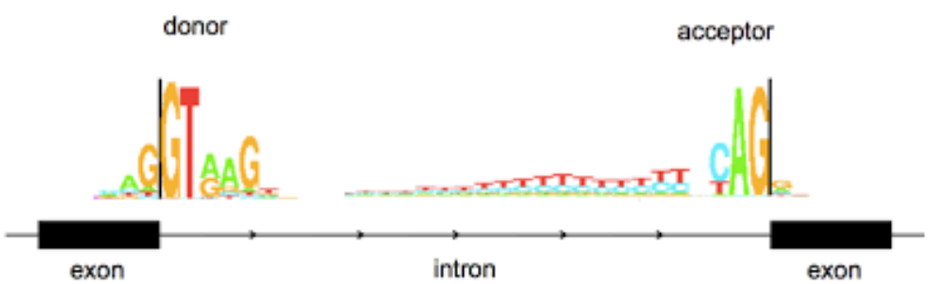
circAtlas [Wu et al.], **circBase** [Glažar et al.]



circRNA **lengths**



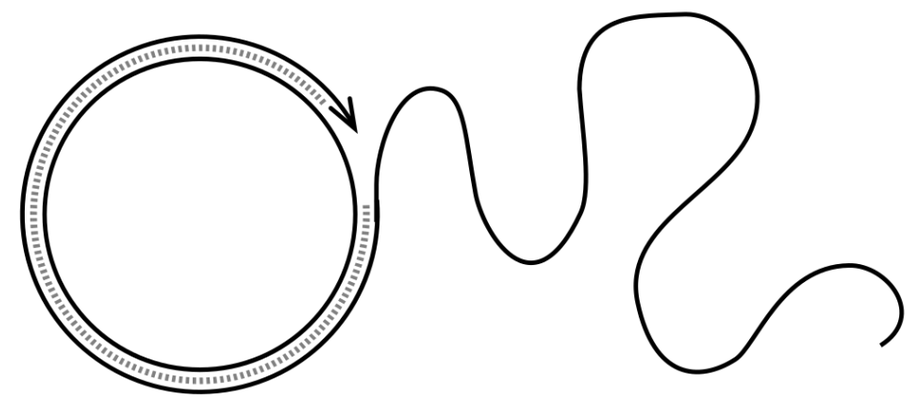
circRNA **type** and **inner structure**



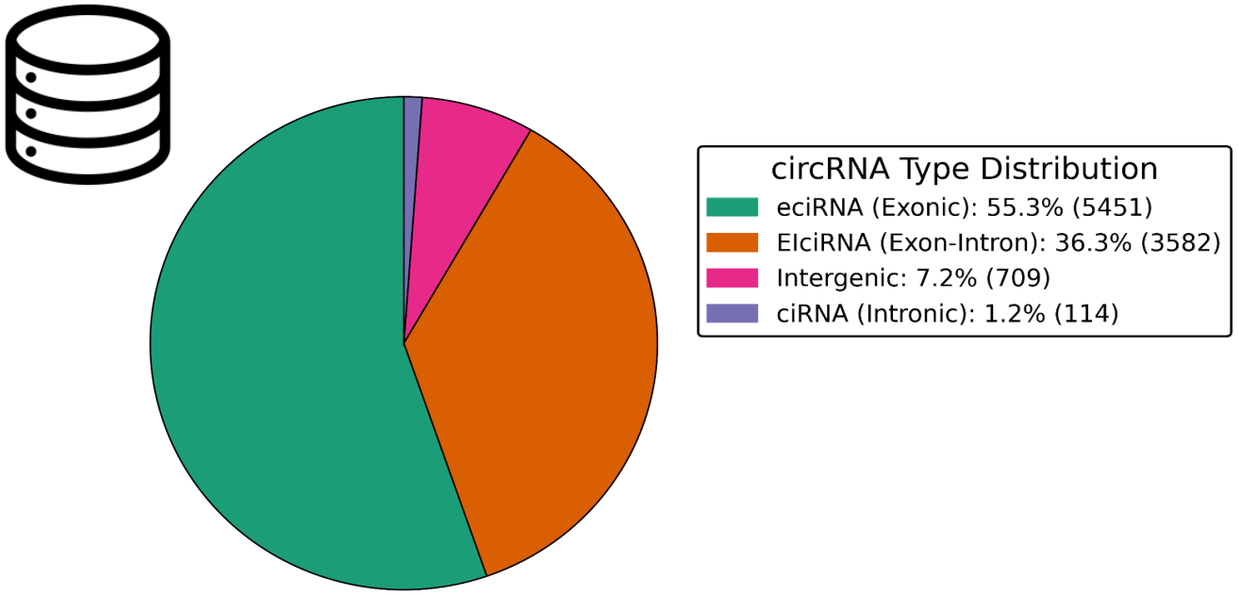
circRNA **splice sites**

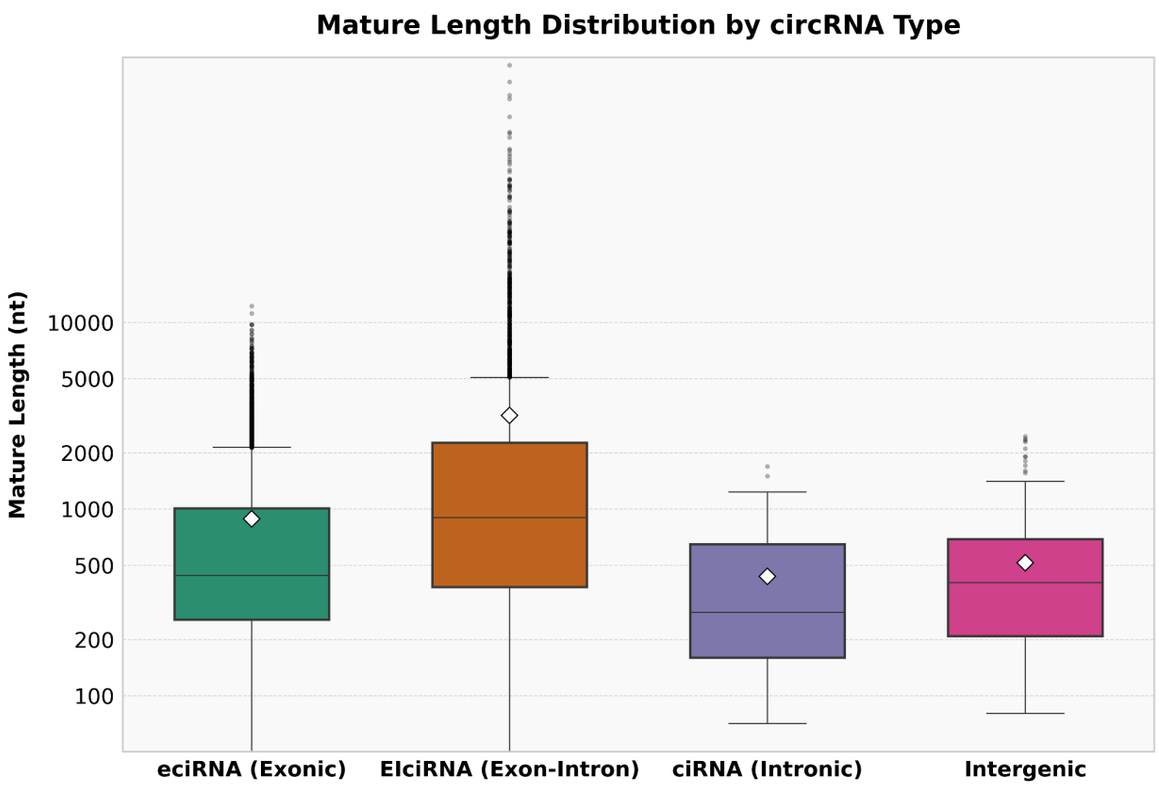
Protocol features

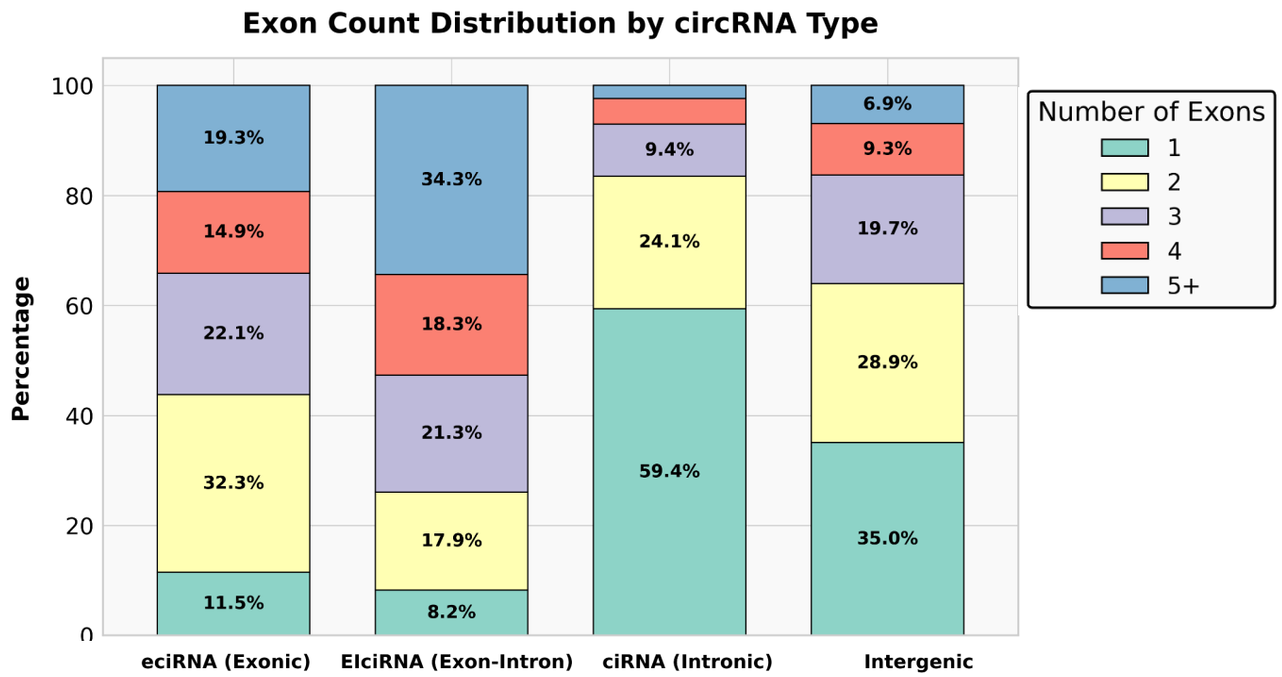
Mouse brain, ONT (CIRI-long protocol, SRA: CRR194180)

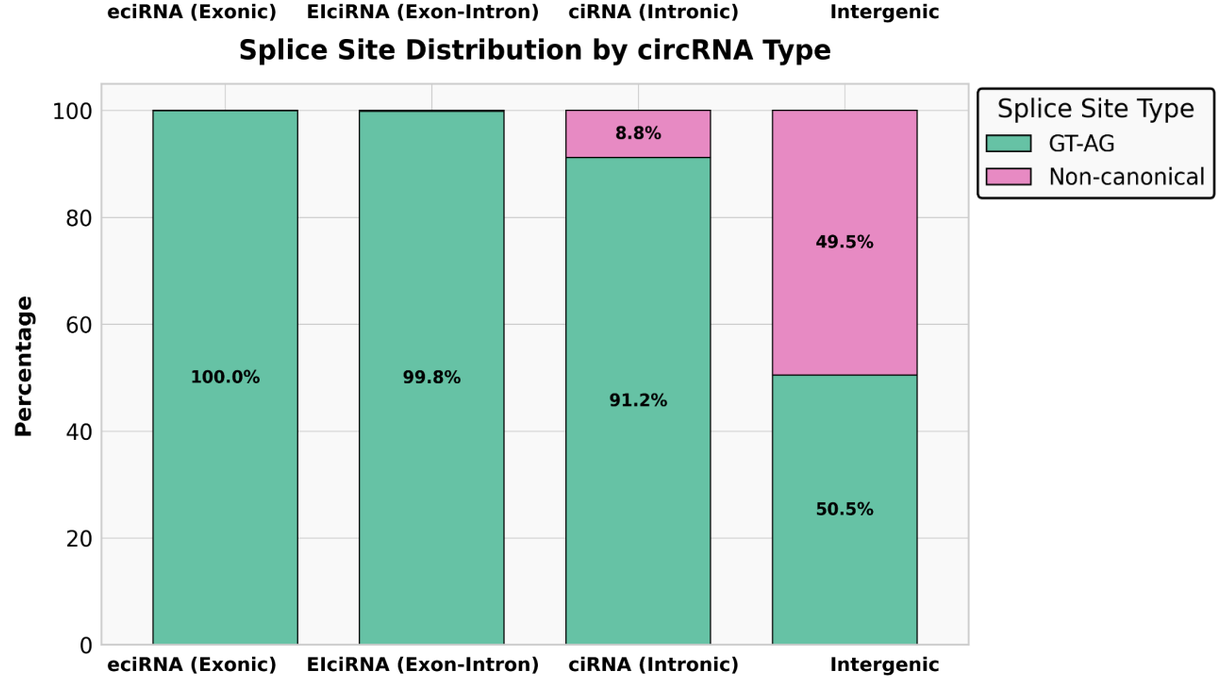


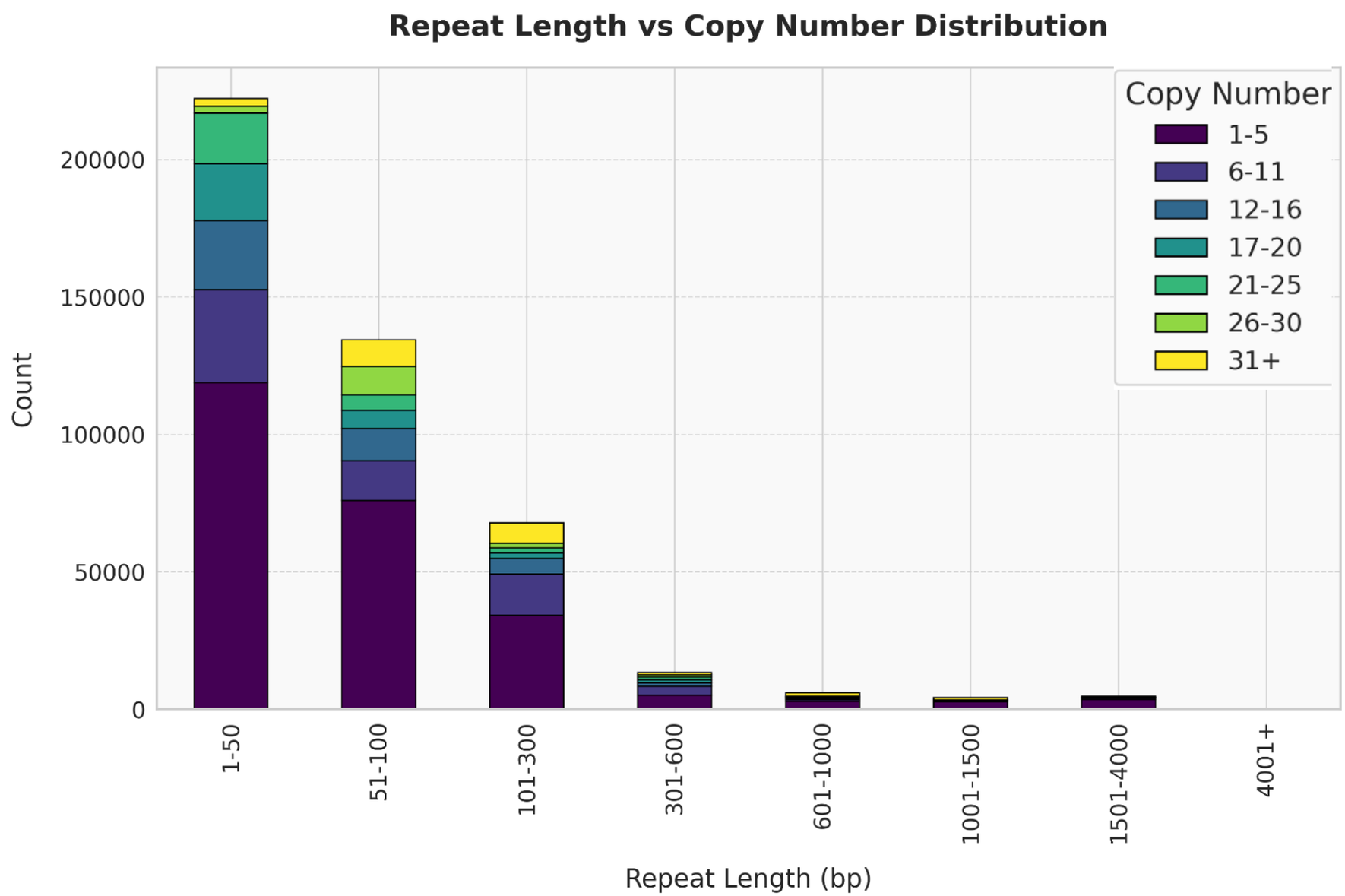
Rolling circles count and **repeat pattern**


circRNA Type Distribution
eciRNA (Exonic): 55.3% (5451)
EIciRNA (Exon-Intron): 36.3% (3582)
Intergenic: 7.2% (709)
ciRNA (Intronic): 1.2% (114)

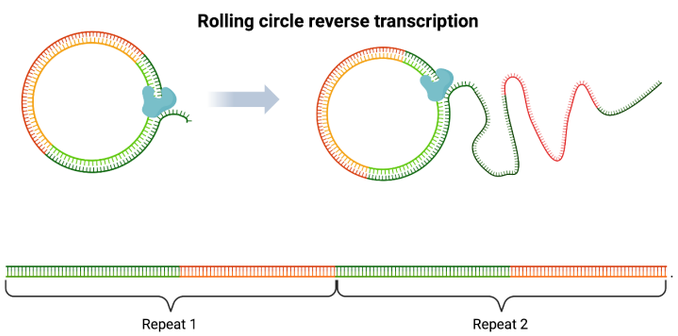

Mature Length Distribution by circRNA Type
Mature Length (nt)
10000
5000
2000
1000
500
200
100
eciRNA (Exonic)
EIciRNA (Exon-Intron)
ciRNA (Intronic)
Intergenic


Exon Count Distribution by circRNA Type
Percentage
100
80
60
40
20
0
19.3%
14.9%
22.1%
32.3%
11.5%
34.3%
18.3%
21.3%
17.9%
8.2%
9.4%
24.1%
59.4%
6.9%
9.3%
19.7%
28.9%
35.0%
eciRNA (Exonic)
EIciRNA (Exon-Intron)
ciRNA (Intronic)
Intergenic
Number of Exons
1
2
3
4
5+


Splice Site Distribution by circRNA Type
Percentage
100
80
60
40
20
0
100.0%
99.8%
8.8%
91.2%
49.5%
50.5%
eciRNA (Exonic)
EIciRNA (Exon-Intron)
ciRNA (Intronic)
Intergenic
Splice Site Type
GT-AG
Non-canonical

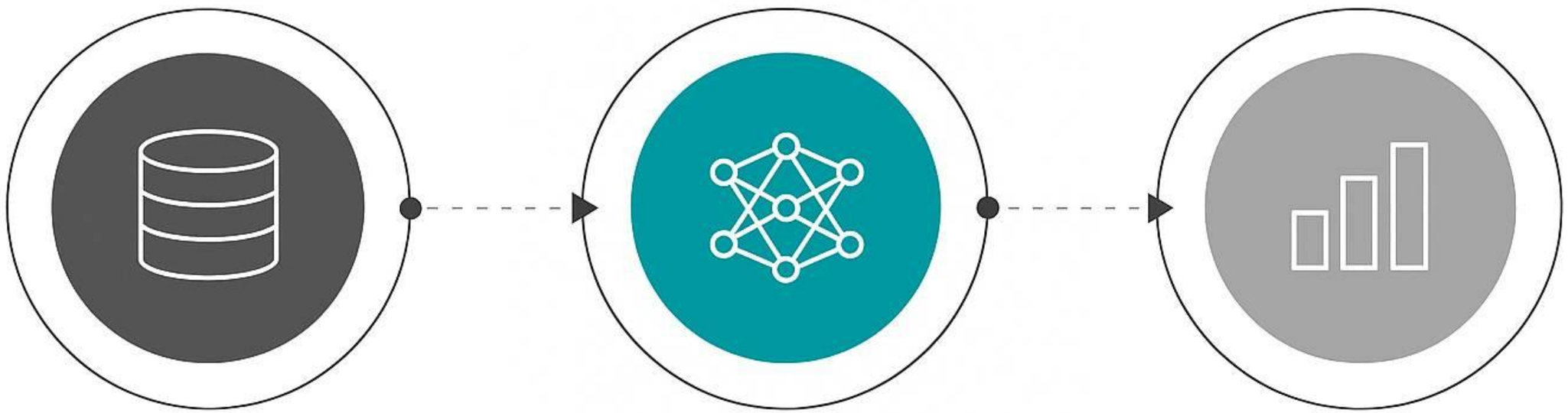


How many rolling circles protocol introduces?



1 repeat = 1 copy of circRNA

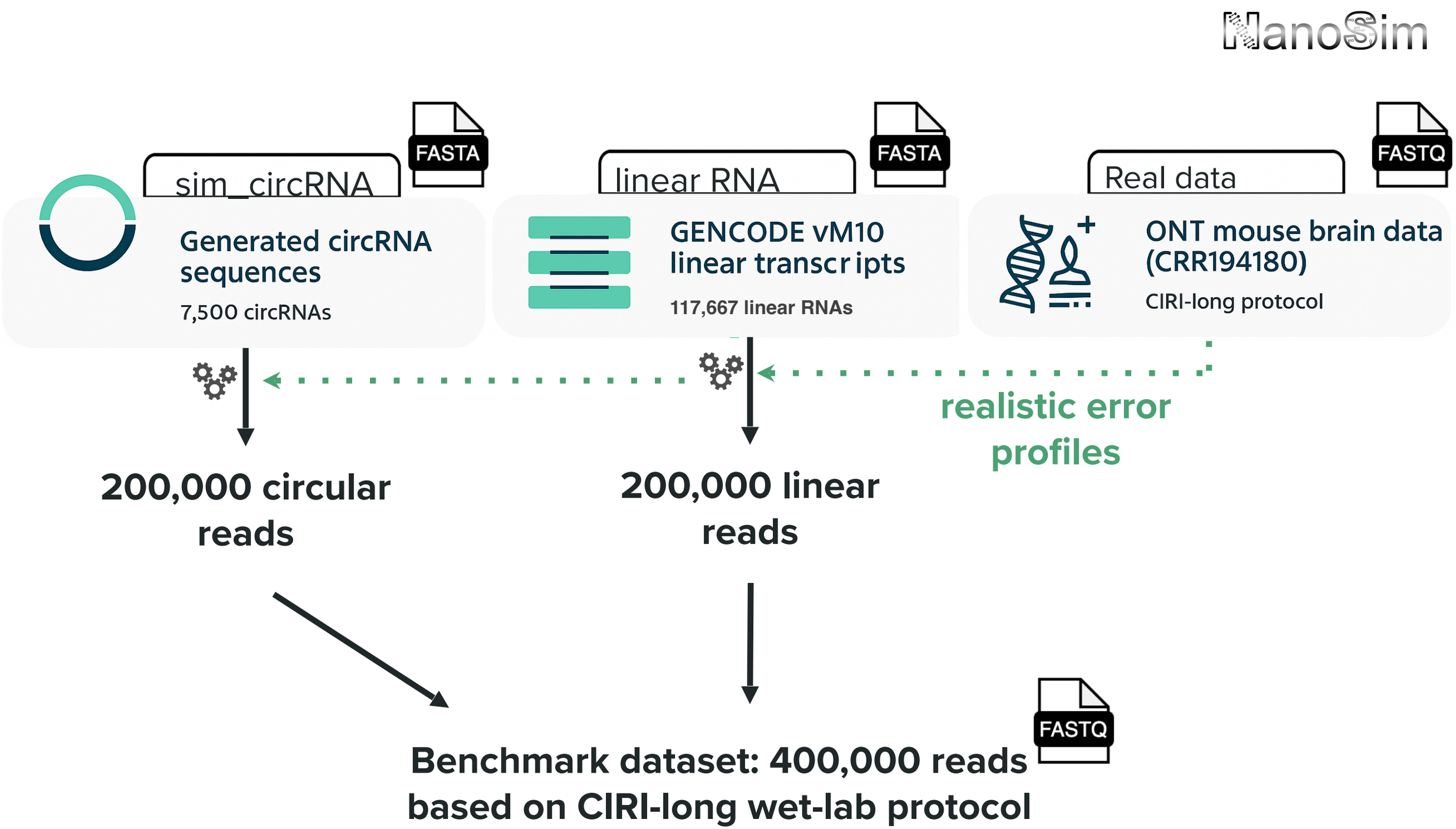
Our benchmark study



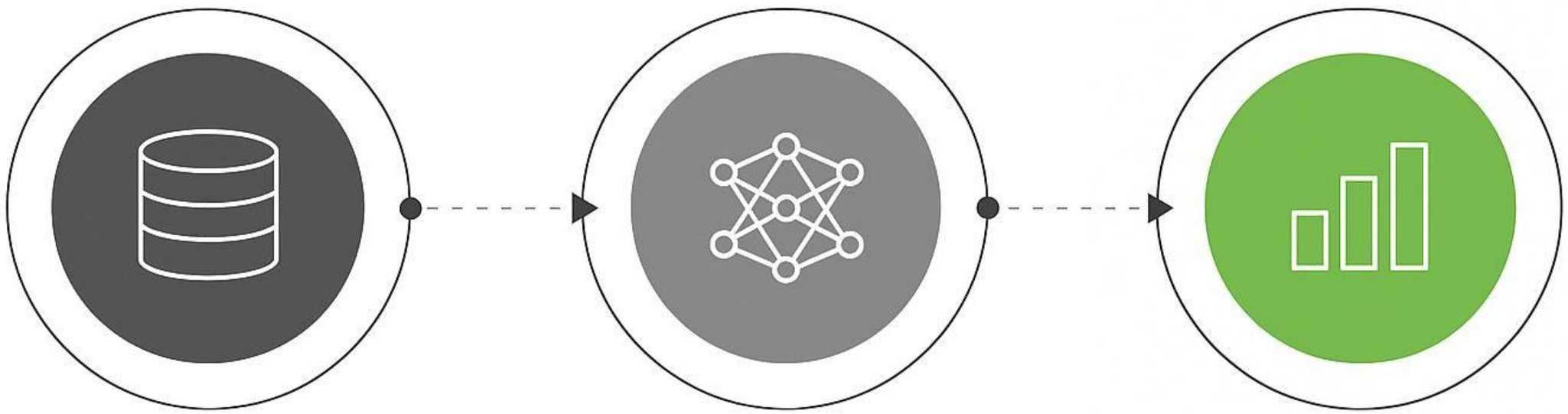
Oxford Nanopore reads generation

Using NanoSim tool
(Hafezqorani et al. Gigascience, 2020)

Benchmark dataset



Our benchmark study

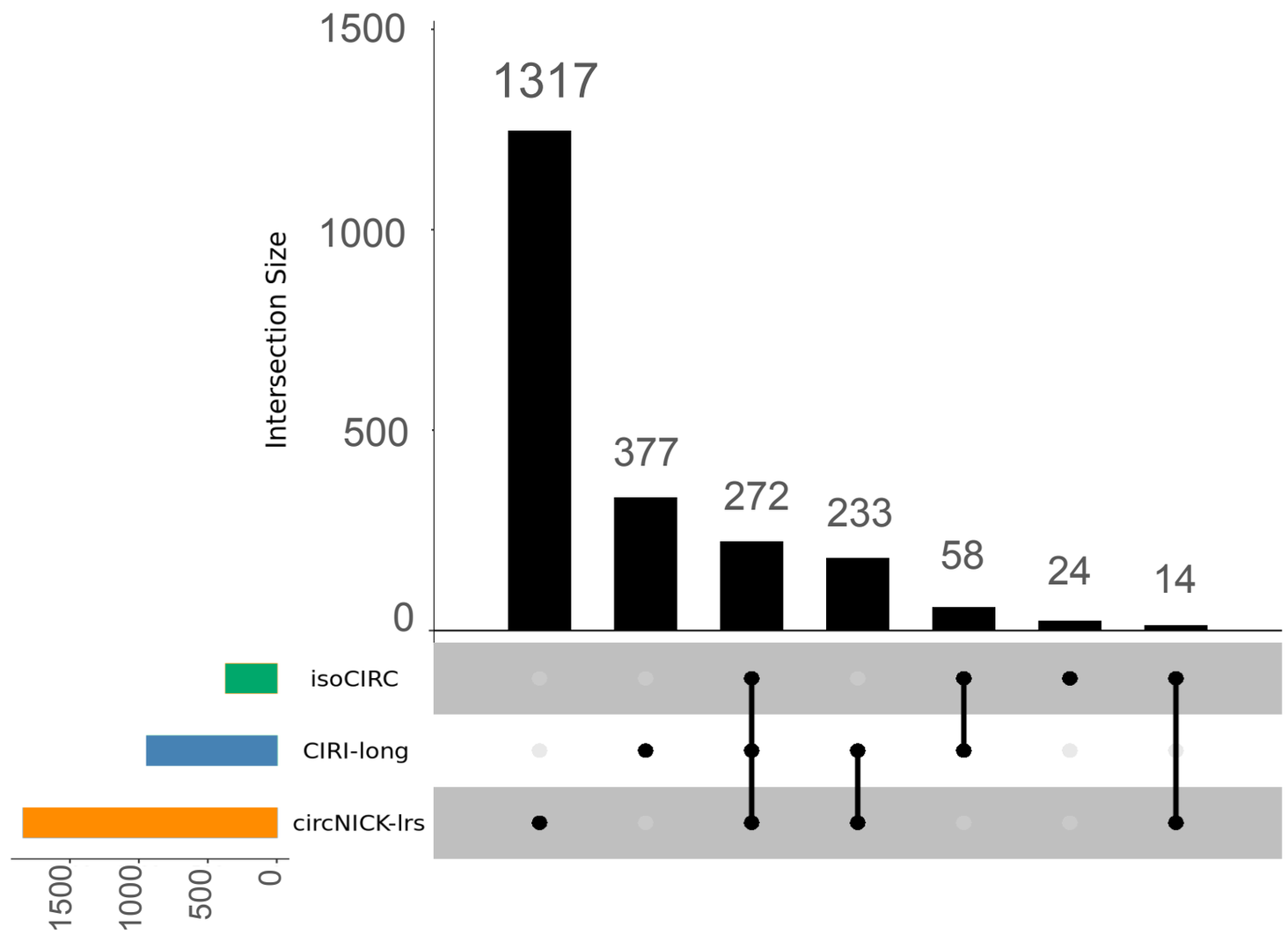


Tool application and evaluation

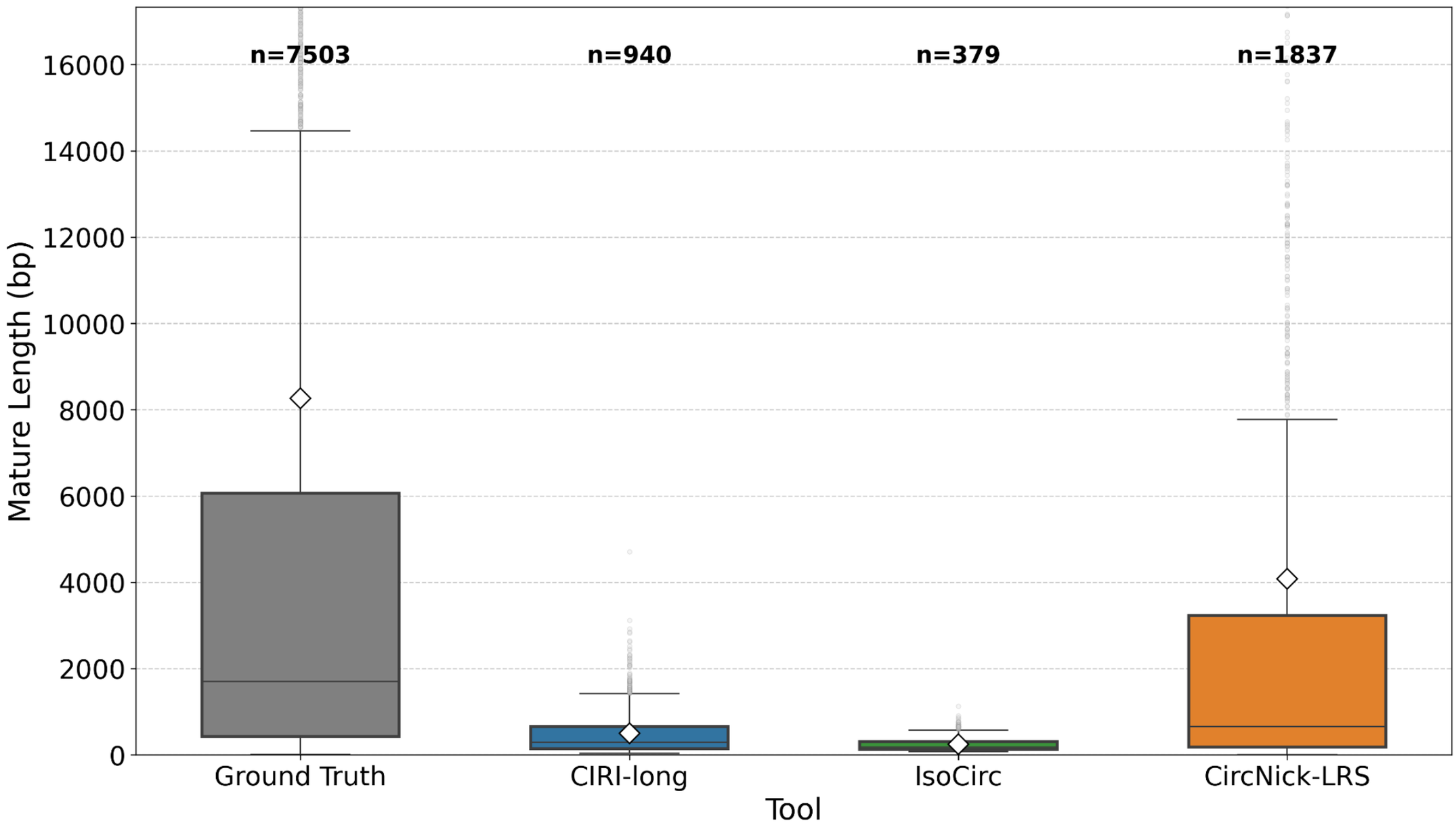
Descriptive and **performance** metrics

Descriptive metrics: Number of circRNAs detected

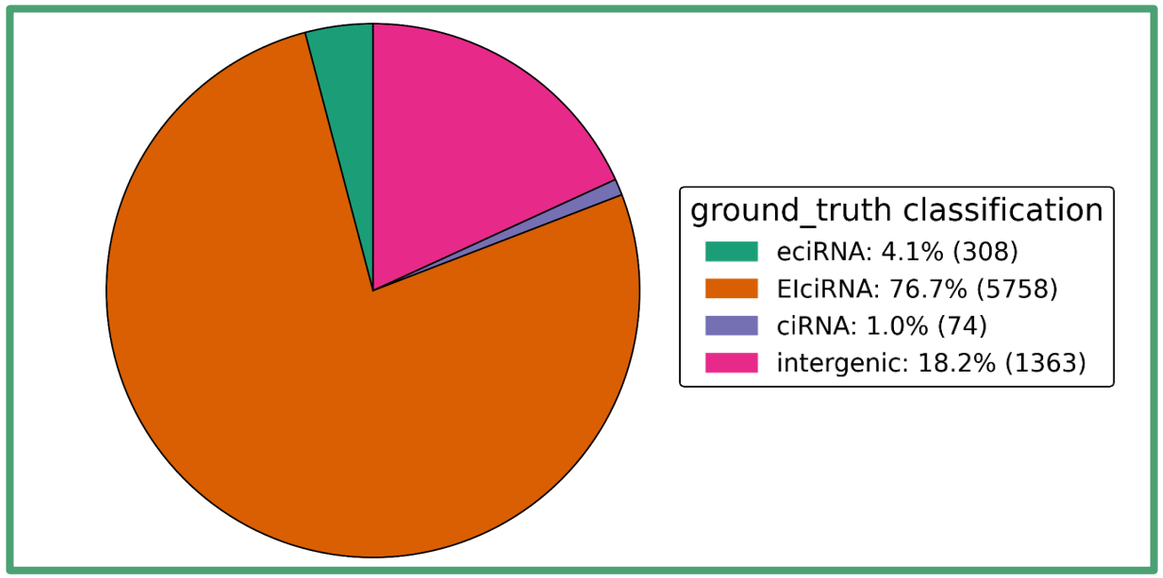
Total isoforms detected: 2295/7503 (30.58%)

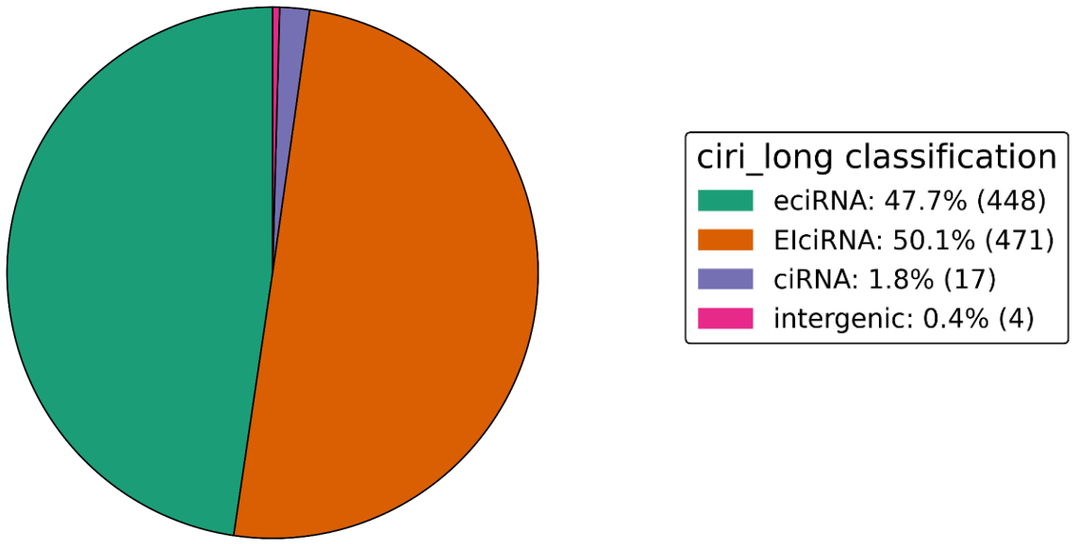


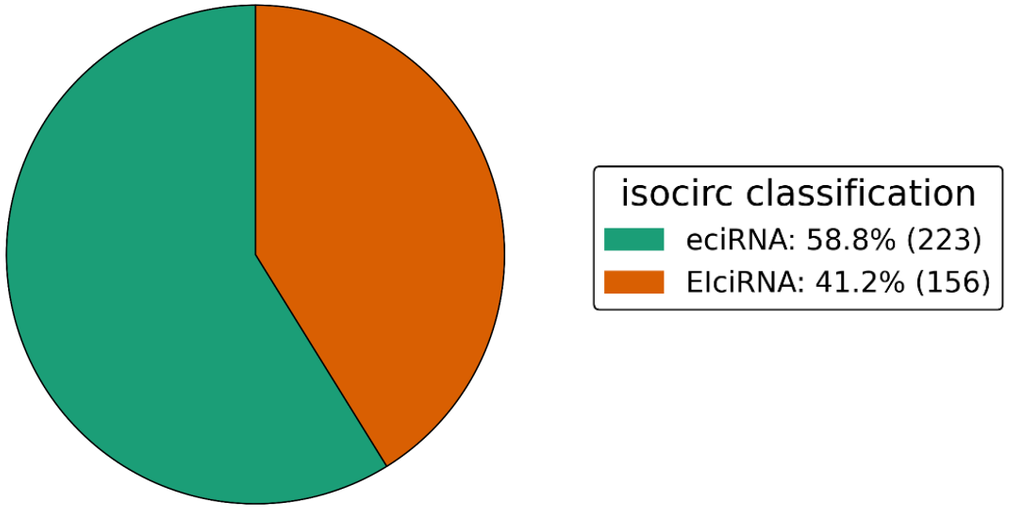
Descriptive metrics: Length distribution

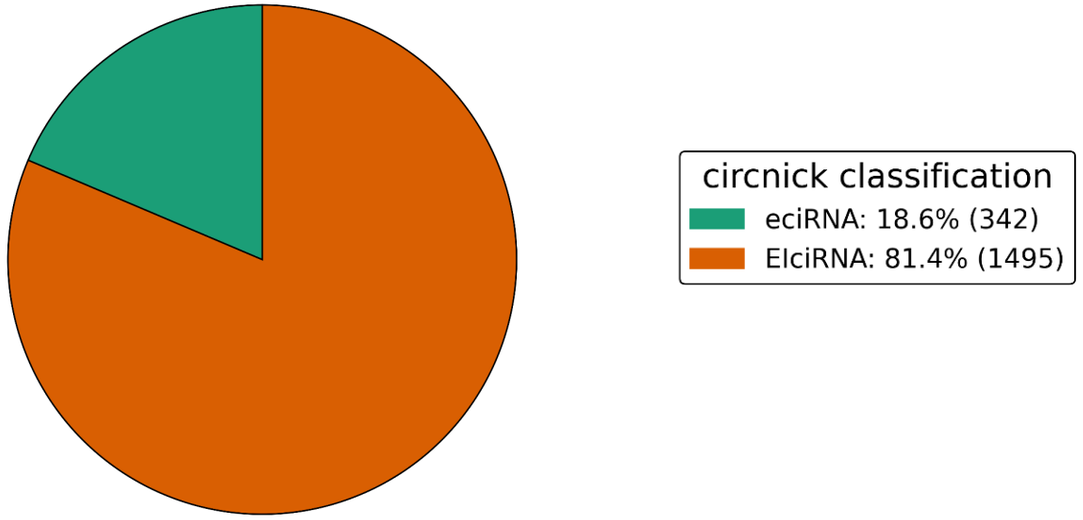


Descriptive metrics: circRNA type distribution

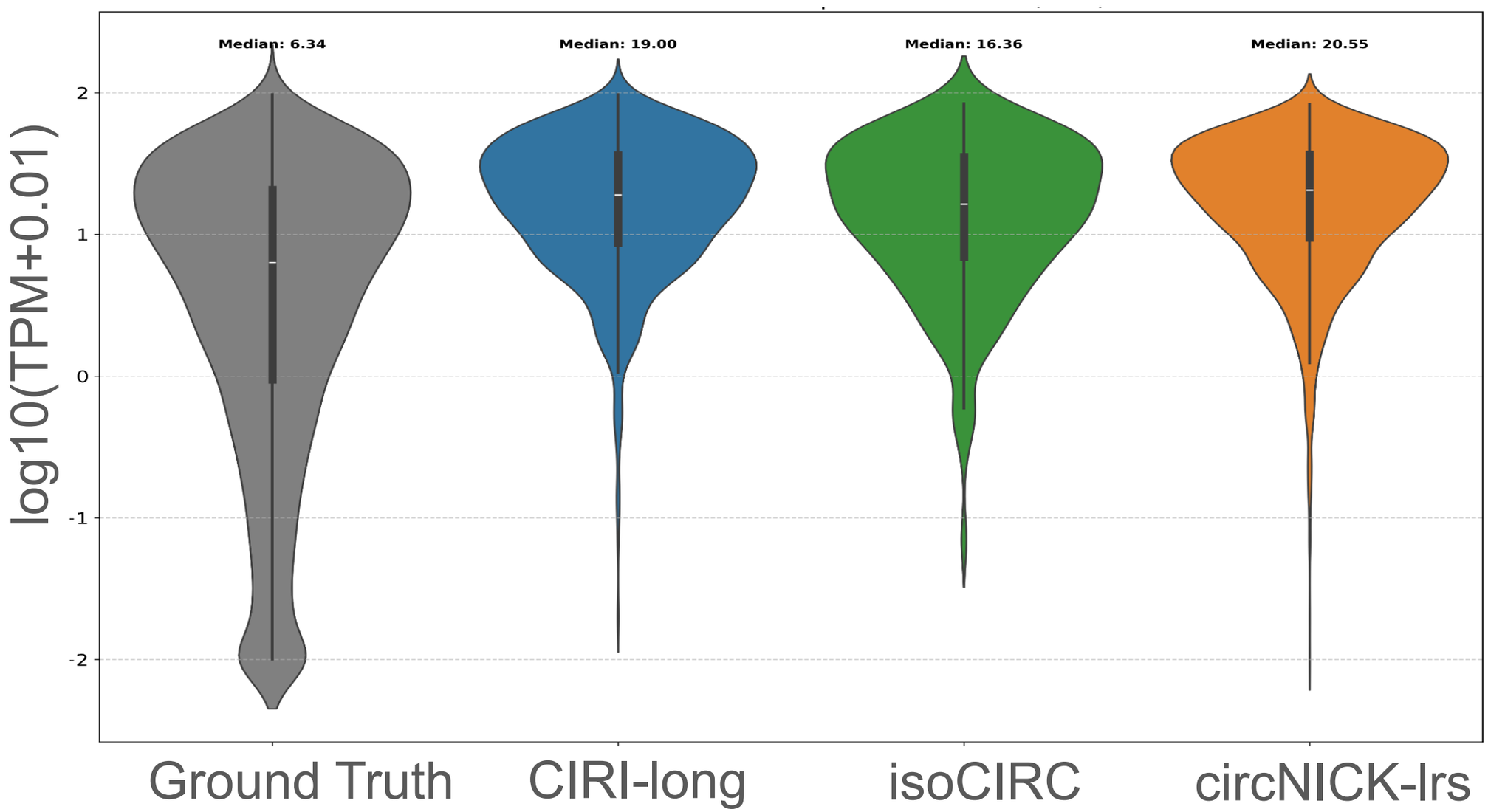






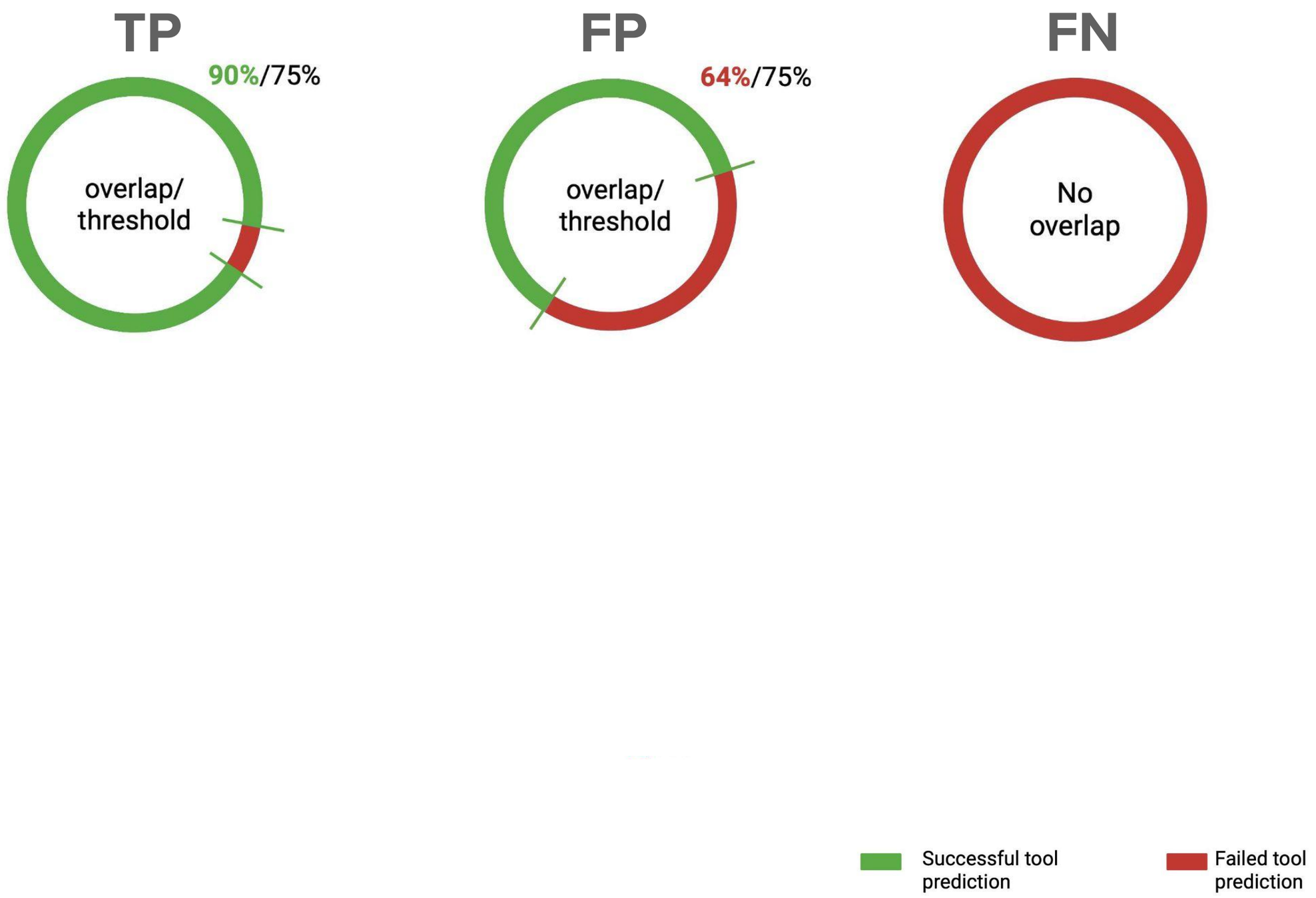


Descriptive metrics: Expression level (TPM)



Performance evaluation

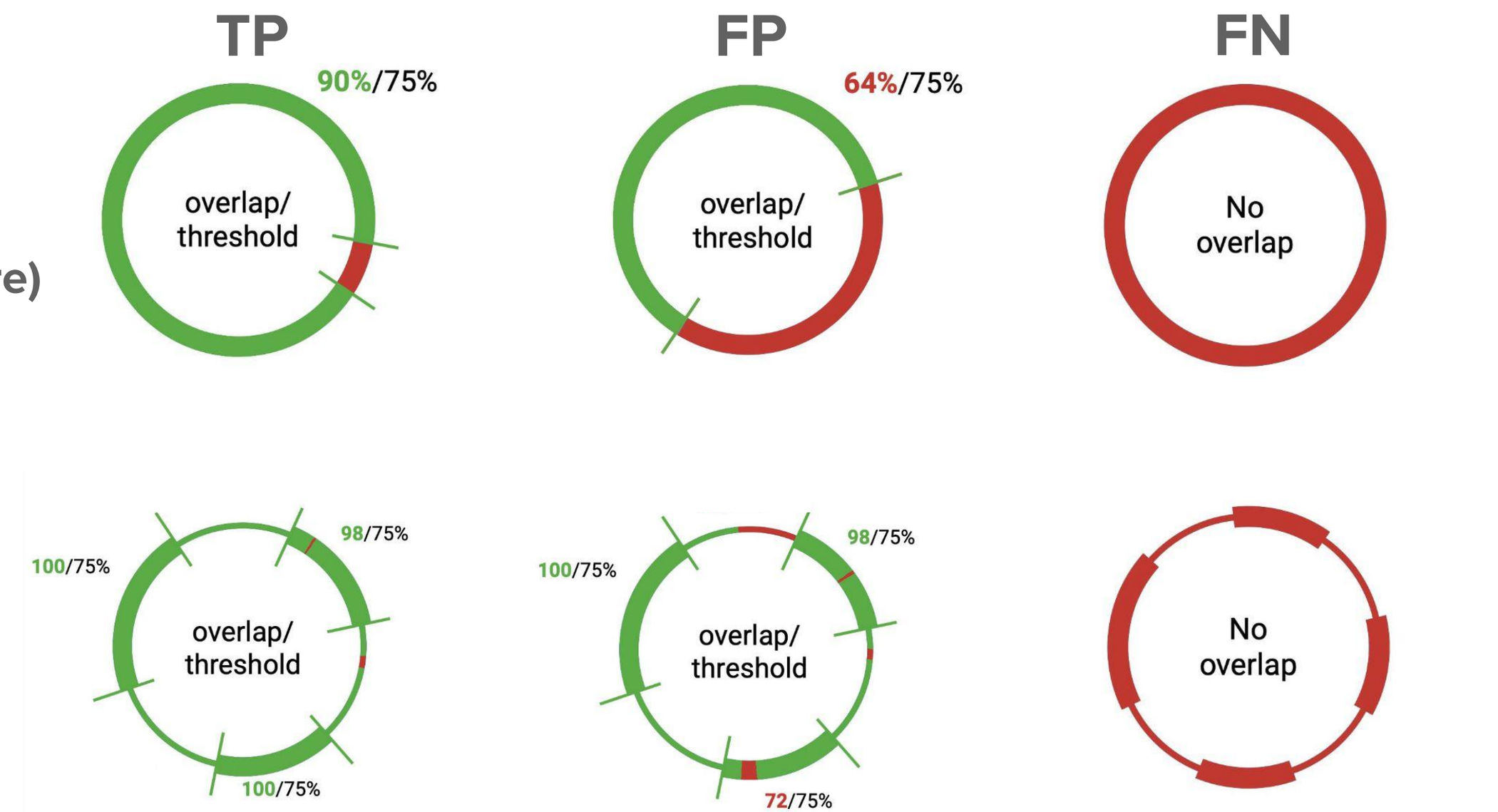
Transcript level
(not exon-aware)



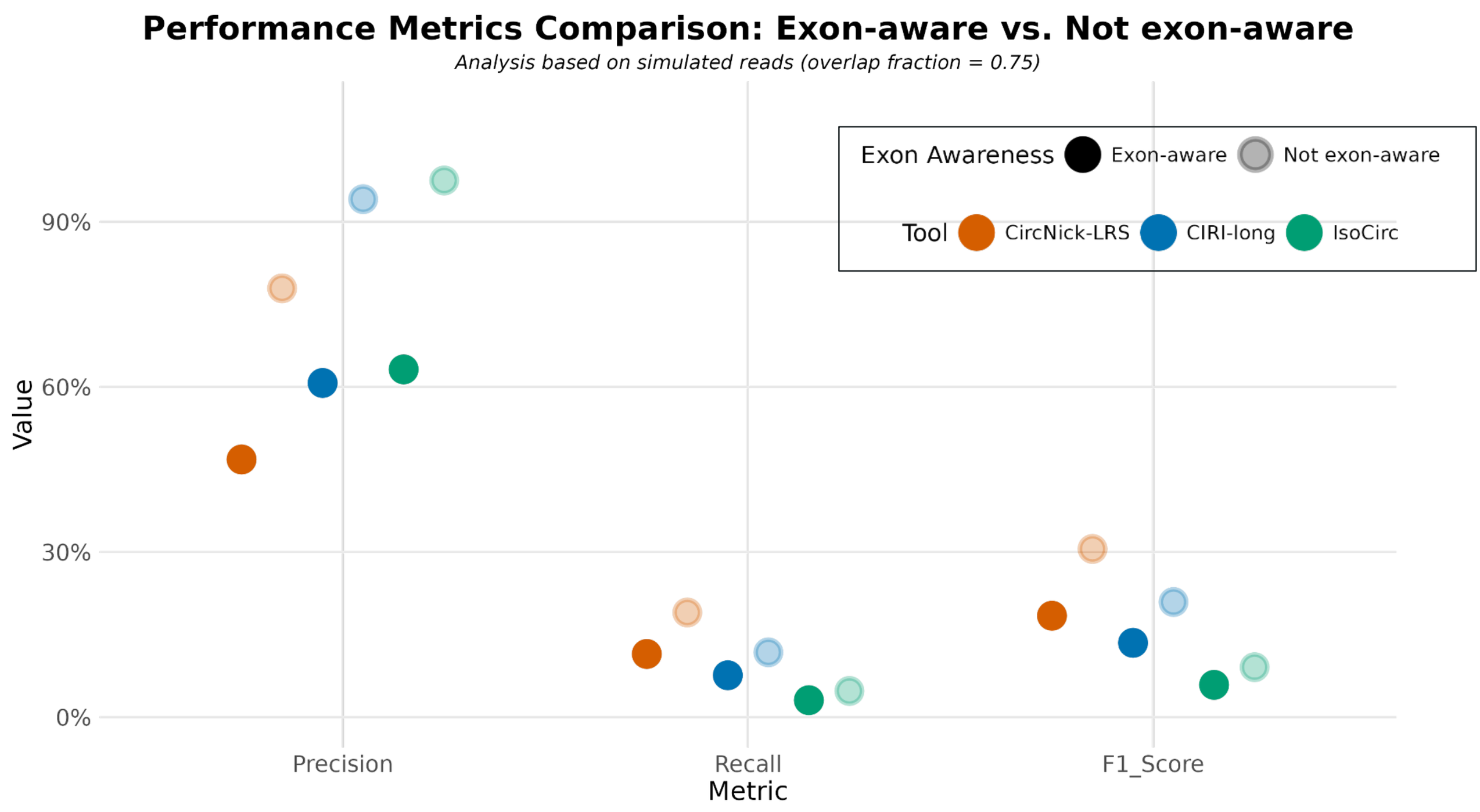
Performance evaluation

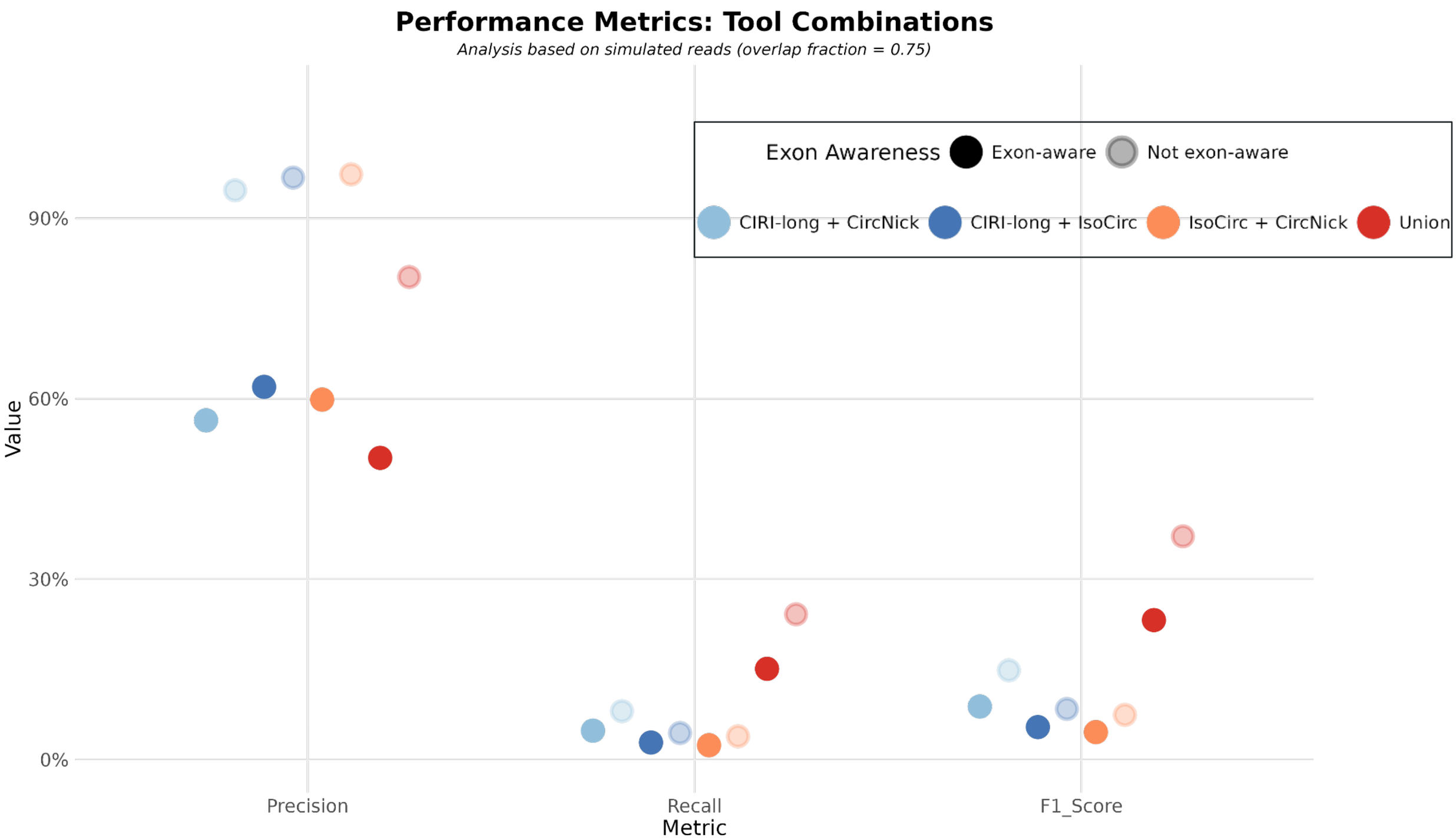
Transcript level
(not exon-aware)

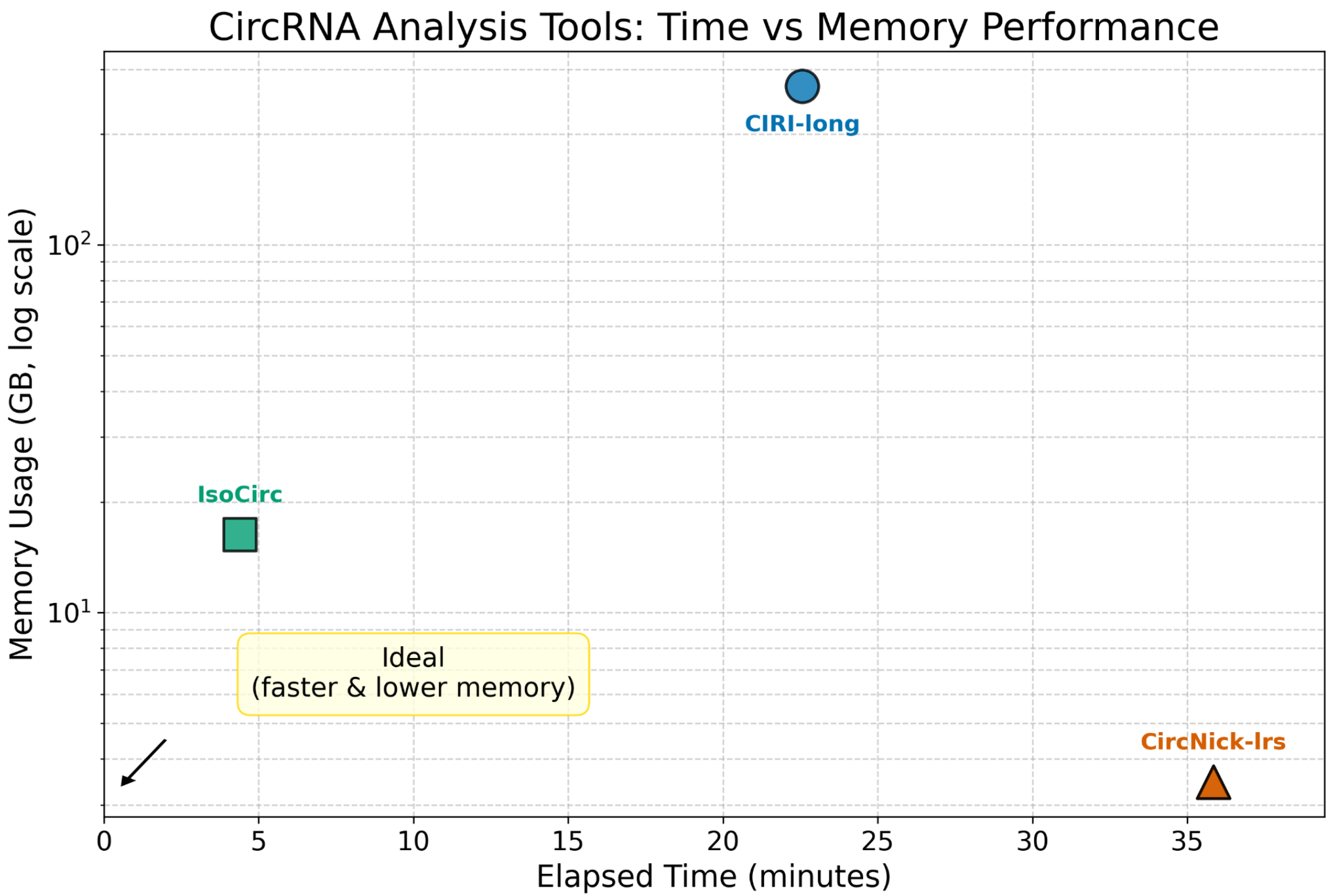
Block level
(exon-aware)



Recall/Sensitivity = TP / (TP+FN)
Precision s= TP / (TP+FP)


Performance Metrics Comparison: Exon-aware vs. Not exon-aware
Analysis based on simulated reads (overlap fraction = 0.75)
Exon Awareness
Exon-aware
Not exon-aware
Tool
CircNick-LRS
CIRI-long
IsoCirc
90%
60%
30%
0%
Value
Precision
Recall
F1_Score
Metric

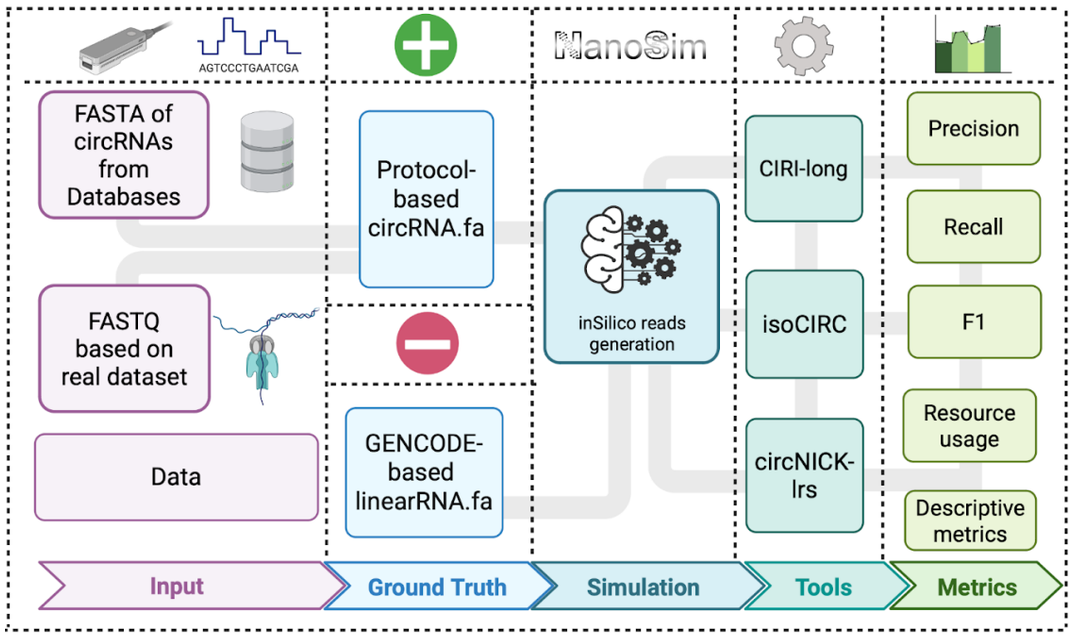

Performance Metrics: Tool Combinations
Analysis based on simulated reads (overlap fraction = 0.75)
Exon Awareness
Exon-aware
Not exon-aware
CIRI-long + CircNick
CIRI-long + IsoCirc
IsoCirc + CircNick
Union
90%
60%
30%
0%
Value
Precision
Recall
F1_Score
Metric


CircRNA Analysis Tools: Time vs Memory Performance
CIRI-long
IsoCirc
CircNick-lrs
Ideal
(faster & lower memory)
Memory Usage (GB, log scale)
10^2
10^1
0
5
10
15
20
25
30
35
Elapsed Time (minutes)

Main Findings

- Each tool has distinct strengths: **isoCirc** (precision and resources), **circNick-lrs** (recall and length), **CIRI-long** (balance and type assignment)
- Only **11.85% overlap** between all three tools - highlighting complementary detection
- All tools show **bias toward highly expressed circRNAs** and **lower sensitivity** for **exact inner structure**

Availability



Simulation framework is available on git

Perspectives

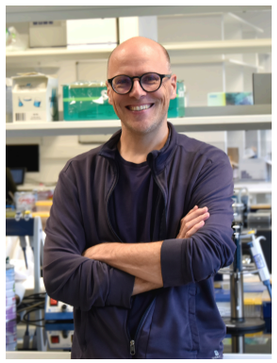
🔬 **Expand protocols** → tool assessment on different protocols and species

🐳 **Containerized pipeline** → tools application and result combination

📊 **Real data application** → Melanoma & hepatocellular carcinoma data generated in IGDR and OSS

Acknowledgements

IGDRion



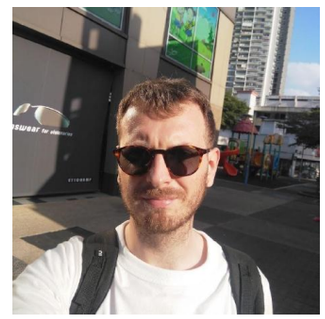
Edouard CADIEU

Aurore BESSON

Victor LE BARS

Enora CORLER

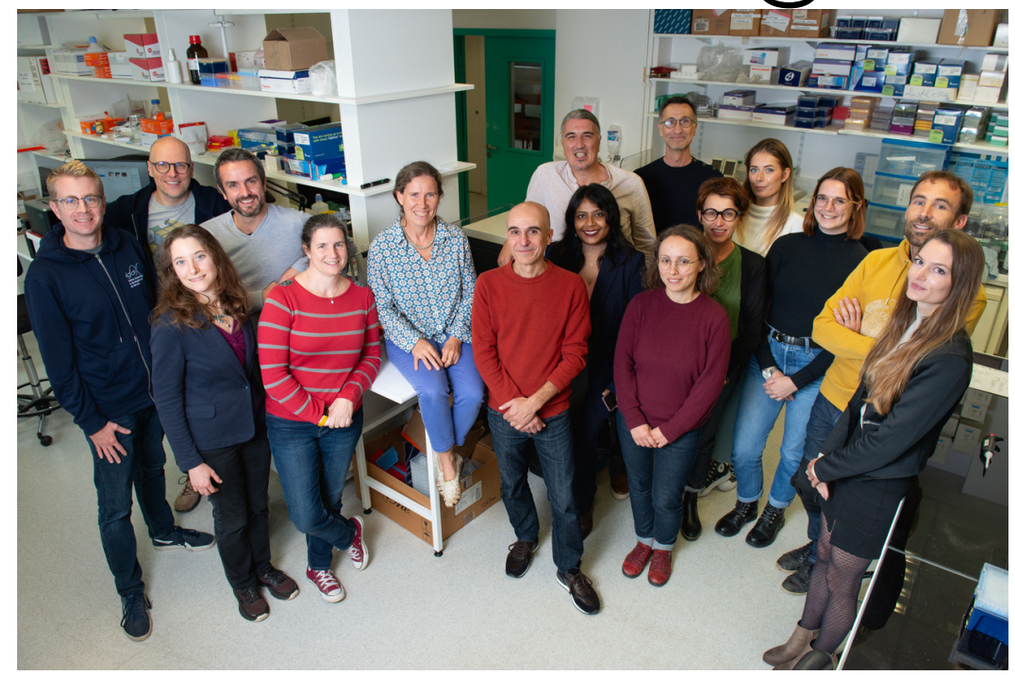
Matthias LORTHIOIS



Nicolaï HOFFMANN

Anastasia RUSAKOVICH

Canine Genetics Team @ IGDR



Catherine ANDRE

BINGO

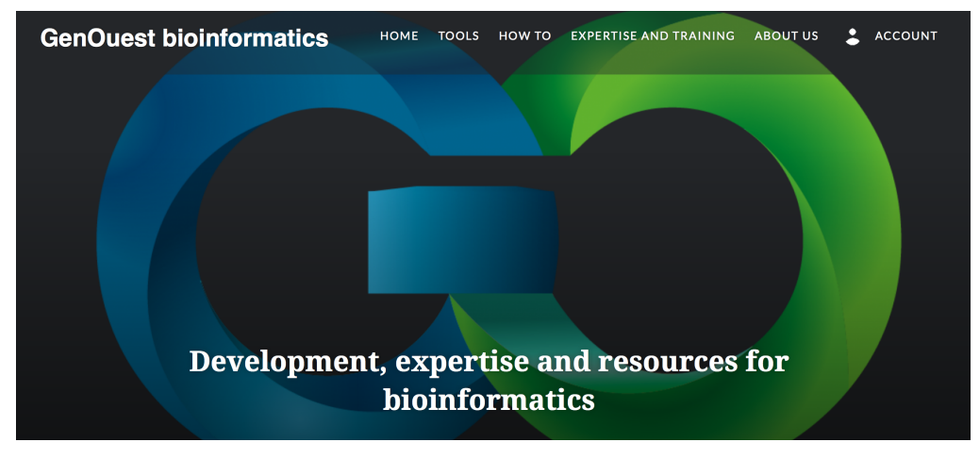
A C G T

>BIOLOGY IN SILICO for NON CODING GENOME and ONCOLOGY

Yuna BLUM

Christophe HELIGON

Genouest bioinformatic



JBISR

Journée de Biologie In silico de Rennes

25 November 2025 - Health Campus

PROGRAM

JBISR is a one-day symposium dedicated to **bioinformatics**, organized as part of the BIS2 (Biology In Silico) transversal axis at IGDR. The program highlights **cutting-edge research** in **AI-driven analysis** and **multi-modal data integration** in **health** and **cancer**.

09:30 - 10:00 Welcoming participants

10:00 - 10:10 **Introductory remarks**

10:10 - 11:45 **Session 1 – Integration of Multi-Modal Data in Biology and Cancer**

10:10 - 11:00 Keynote Speaker 1

DR. ANDREA RAU
Université Paris-Saclay, INRAE, AgroParisTech

Title: **Tackling the statistical challenges of multi-omic integration**

Dr. Andrea Rau, a leading expert in biostatistics and computational biology, will discuss cutting-edge methodologies for multi-omics data integration and highlight applications in oncology.

11:00 - 11:45 **3 selected presentations** *(15 minutes for each presentation)*

11:45 - 12:45 *Lunch Break*

12:45 – 13:45 **Poster all Sessions** with Coffee

13:45 – 15:15 **Session 2 – AI for Sequence and Image Analyses**

13:45 - 14:25 Keynote Speaker 2

DR. JULIETTE GRIFFIÉ
Science for Life Laboratory, Stockholm University, Sweden

Title: **Interpretable ML for data driven therapeutic design and diagnosis**

Dr. Juliette Griffié specializes in computational imaging and AI-driven biomedical modeling. Her talk will cover recent advances in microscopy image analysis, generative AI, and explainable AI in immunologyapplications in oncology.

14:25 - 15:10 **3 selected presentations** (15 minutes for each presentation)

15:15 – 15:45 *Coffee Break*

15:45 – 17:15 **Session 3 – Applications of AI and integrative approaches in Personalized Medicine and Health**

15:45 - 16:25 Keynote Speaker 3

DR. RÉMY NICOLLE
Centre de Recherche sur l'Inflammation (INSERM, CNRS, Université Paris Cité) Paris

Title: **Multimodal Tumor phenotyping for treatment recommendation in pancreatic cancer**

Dr. Rémy Nicolle, a renowned expert in onco-genomics and pancreatic cancer research, will discuss emerging multi-modal integrative approaches and AI methodologies for personalized medicine, focusing on patient stratification, biomarker discovery, and personalized therapeutic strategies in pancreatic cancer.

16:25 – 17:15 **3 selected presentations** (15 minutes for each presentation)

17:15 – 17:30 **Closing Session - Future Directions in Bioinformatics**

Closing Remarks and Farewell