Genome annotation of lncRNAs using long-read transcriptomics:

lessons from large-scale annotation projects

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

Fabien DEGALEZ
(fabien.degalez@institut-agro.fr)

BINGO team

IGDRion sequencing PF

UMR PEGASE, équipe GGA

JOBIM 2026

Why should we still care about genome annotation?

<u>3 layers:</u>

Human genome sequence

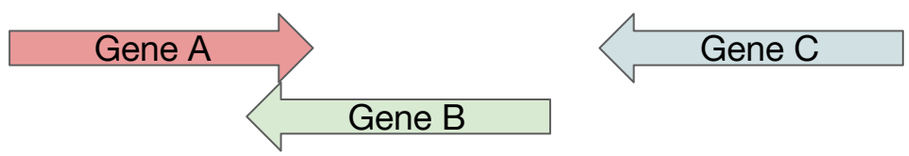
↓

Reference annotation

↓

Biological interpretation

…ATCGGTCGATCGATGTTAGCAGCTAACATCGATTTACGATGCAT…



RNA-seq quantification

Variant interpretation
GWAS
Single-cell RNA-seq
Comparative genomics

Every genomic analysis ultimately depends on the quality of the reference annotation.

Genome sequencing is no longer the bottleneck.
Today, the challenge is understanding what is actually transcribed

Why should we still care about genome annotation?

Genome annotation is far from complete

Human Genome Project

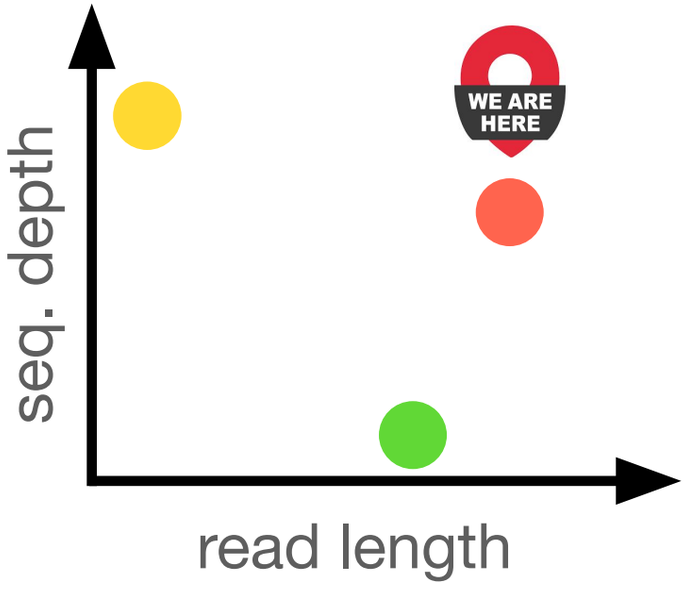
1977 : Sanger sequencing ***(First generation)***

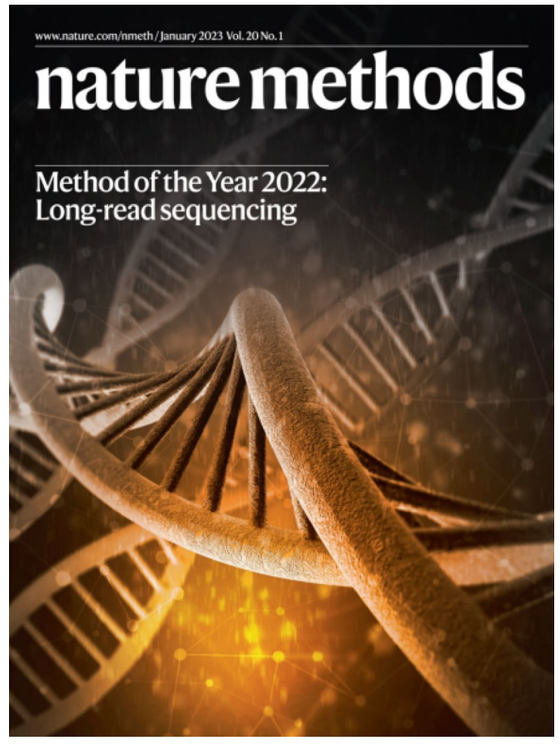
1990 — 1st “complete” genome — **2001**

- 3 billions nucleotides
- ~3 billions $

2007 : NGS - Short read sequencing ***(Seconde generation)***

2015 : Long read sequencing (1 molecule ?) ***(Third generation)***





The impact of read “length”

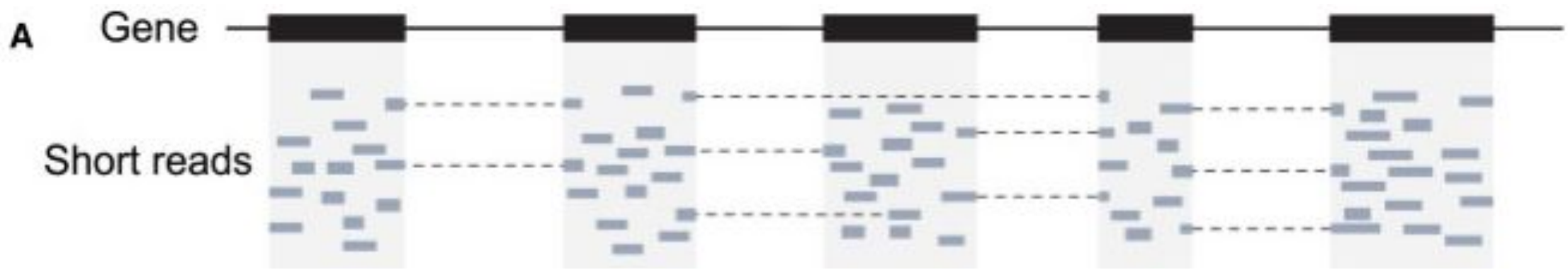
Genome annotation is far from complete

Ament, I.H., DeBruyne, N., Wang, F., Lin, L., 2025. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. Molecular Therapy 33, 883–894. https://doi.org/10.1016/j.ymthe.2024.11.025

The impact of read “length”

Genome annotation is far from complete

Ament, I.H., DeBruyne, N., Wang, F., Lin, L., 2025. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. Molecular Therapy 33, 883–894. https://doi.org/10.1016/j.ymthe.2024.11.025

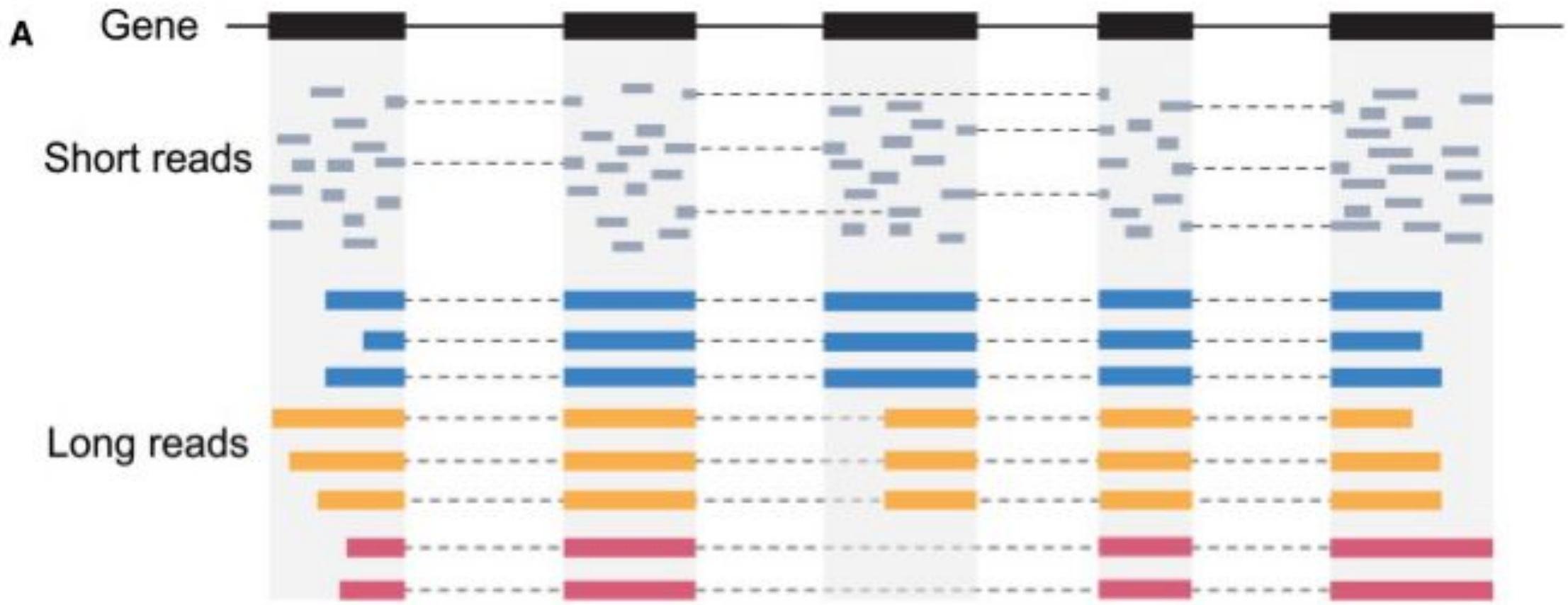


- 🚫 RNAs are fragmented
- 🚫 Assembly problem for both :
  - mRNAs and lncRNAs
  - novel isoforms of known genes and novel genes

The impact of read “length”

Genome annotation is far from complete

Ament, I.H., DeBruyne, N., Wang, F., Lin, L., 2025. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. Molecular Therapy 33, 883–894. https://doi.org/10.1016/j.ymthe.2024.11.025



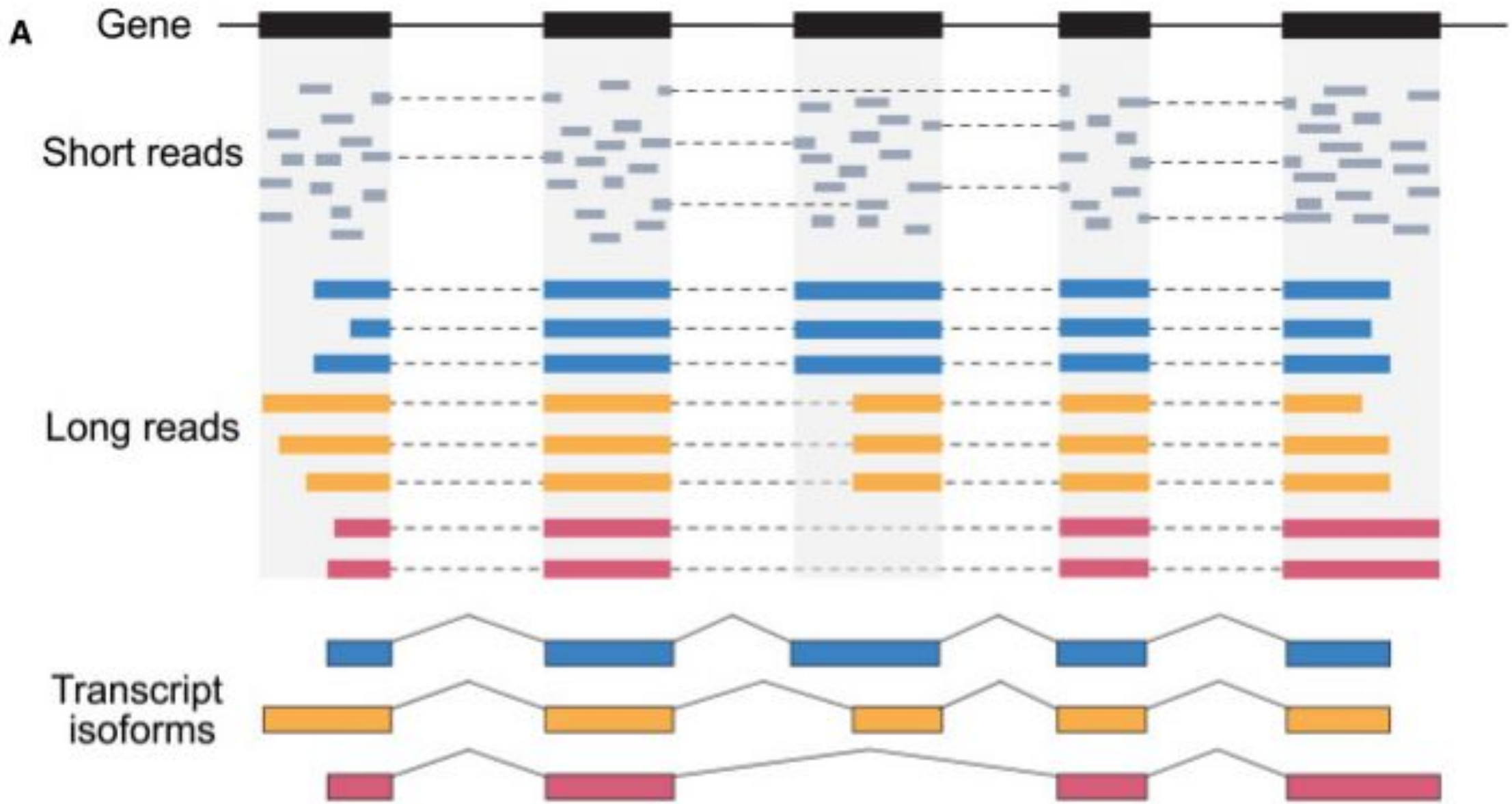
- ✅ "No need to reassemble reads"
- ✅ Characterize and quantify "full length" transcripts (repeat-associated transcripts
- ✅ Direct exon connectivity
- ✅ Epitranscriptomics: identification of RNA modifications (m6A)

➡ “Unfragmented” view of the transcriptome

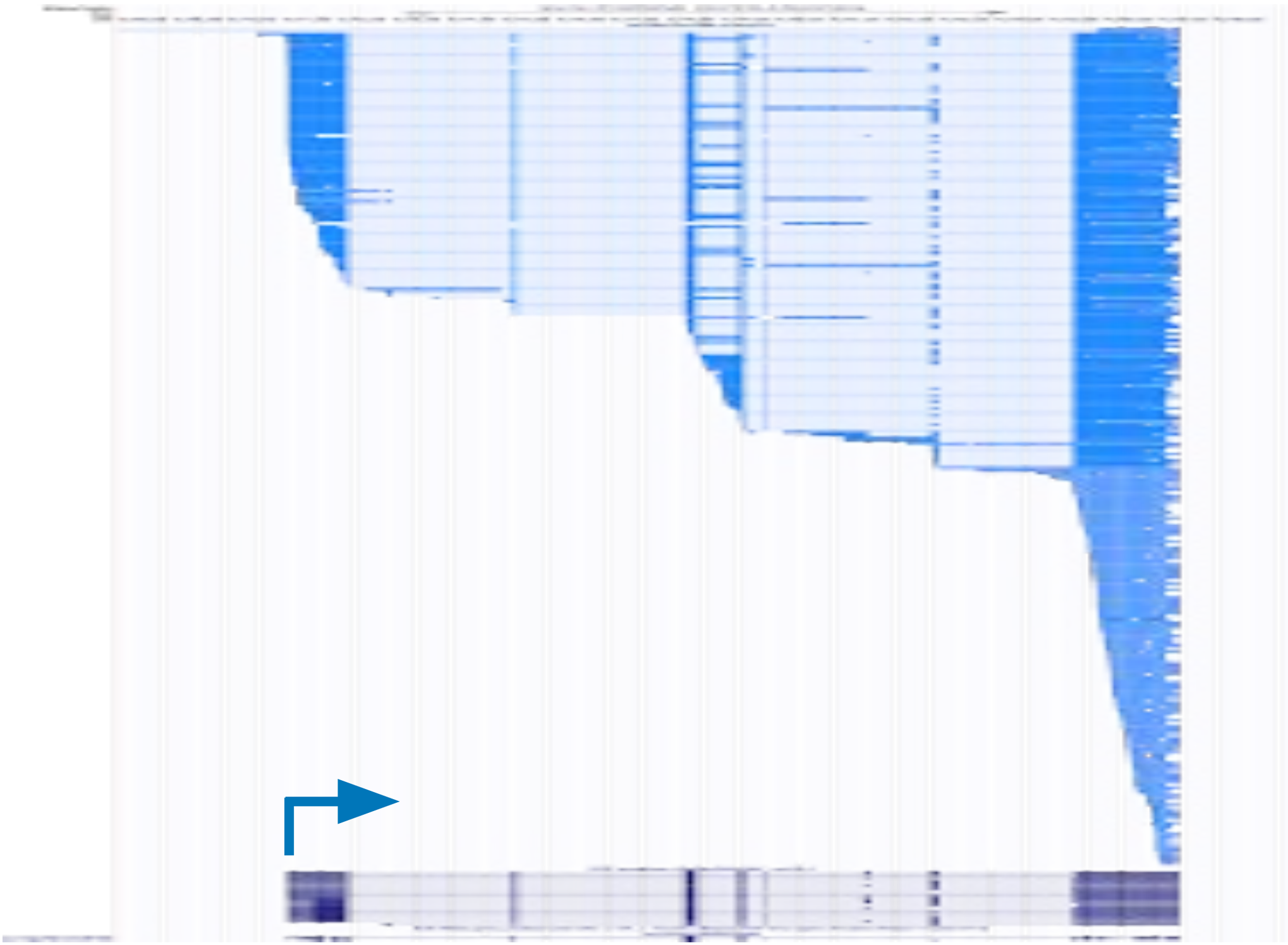
The impact of read “length”

Genome annotation is far from complete

Ament, I.H., DeBruyne, N., Wang, F., Lin, L., 2025. Long-read RNA sequencing: A transformative technology for exploring transcriptome complexity in human diseases. Molecular Therapy 33, 883–894. https://doi.org/10.1016/j.ymthe.2024.11.025

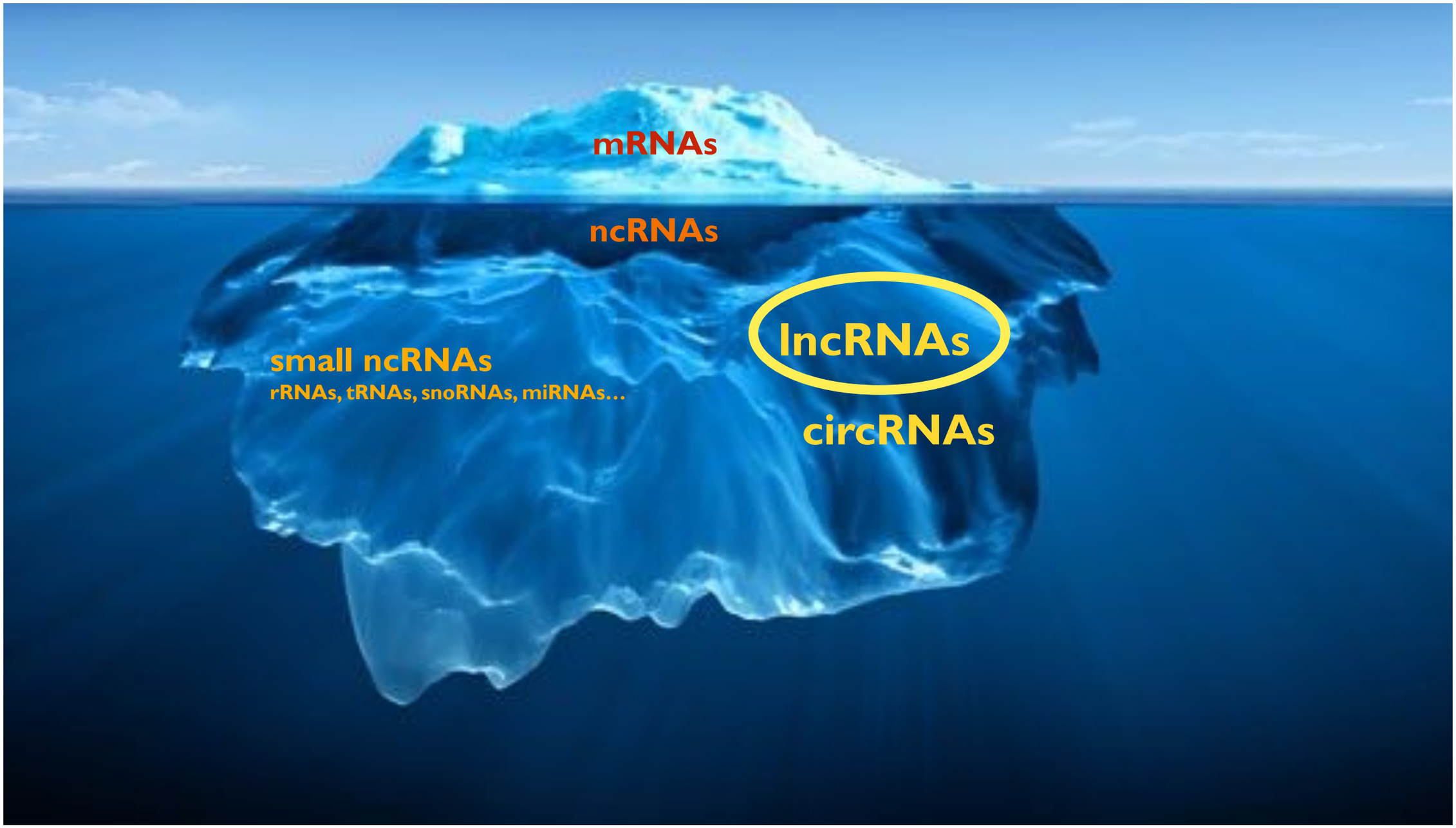


LR-RNASeq is not always long…



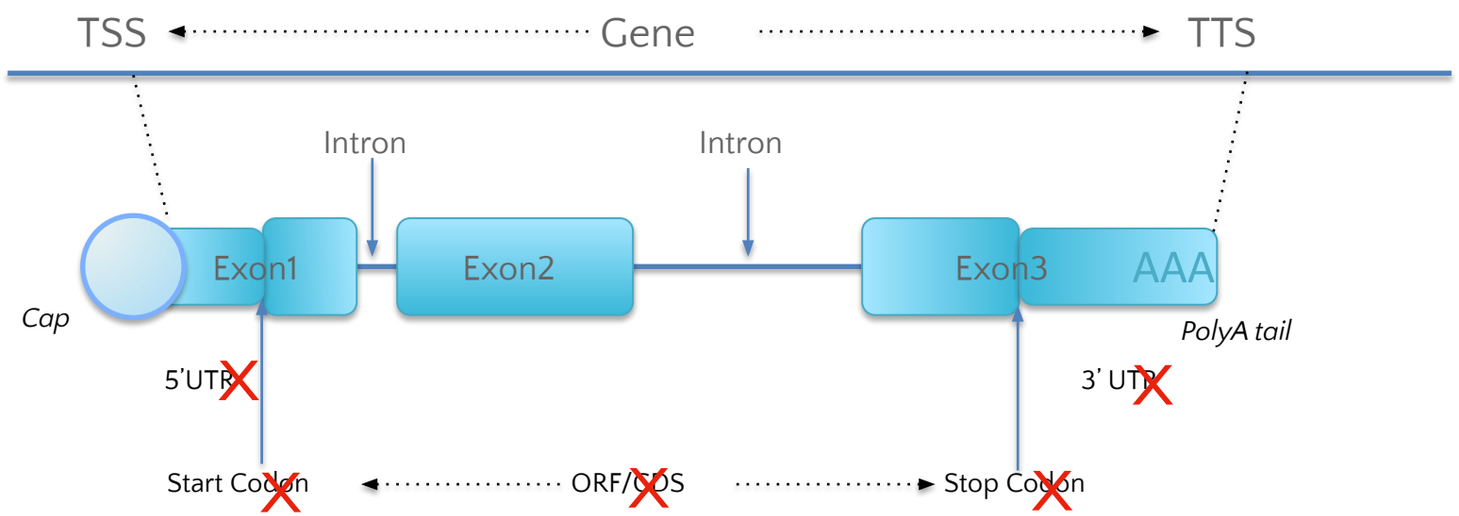
Sessegolo et al., 2019

Why are lncRNAs still missing?



Why are lncRNAs still missing?

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
- 🚫 Lowly expressed, highly tissue-specific
- 🚫 Less conserved across species
- 🚫 Often overlap Transposable Elements (TEs)

What do long reads actually change for lncRNAs?

1 - Can we reconstruct lncRNAs transcripts correctly?

(GENCODE-CLS / Annexa)

↓

2 - Can one annotation represent the entire diversity?

(PanTranscriptome)

↓

3 - How do we prioritize which lncRNAs matter?

(lncRNA_PutOrtho)

What do long reads actually change for lncRNAs?

1 - Can we reconstruct lncRNAs transcripts correctly?

(GENCODE-CLS / Annexa)

1 - Can we reconstruct lncRNAs transcripts correctly?

The GENCODE CLS project: massively expanding the lncRNA catalog through capture long-read RNA sequencing

Tamara Perteghella, Gazaldeep Kaur, Sílvia Carbonell-Sala, Jose Gonzalez-Martinez, Toby Hunt, Tomasz Mądry, Irwin Jungreis, Fabien Degalez, Carme Arnan, Ramil Nurtdinov, Julien Lagarde, Beatrice Borsari, Cristina Sisu, Yunzhe Jiang, Ruth Bennett, Andrew Berry, Marta Blangiewicz, Daniel Cerdán-Vélez, Kelly Cochran, Covadonga Vara, Claire Davidson, Sarah Donaldson, Cagatay Dursun, Silvia González-López, Sasti Gopal Das, Kathryn Lawrence, Daniel Nachun, Matthew Hardy, Zoe Hollis, Mike Kay, José Carlos Montañés, Pengyu Ni, Emilio Palumbo, Carlos Pulido-Quetglas, Marie-Marthe Suner, Xuezhu Yu, Dingyao Zhang, François Aguet, Kristin Ardlie, Stephen B. Montgomery, Jane E. Loveland, M. Mar Albà, Mark Diekhans, Andrea Tanzer, Jonathan M. Mudge, Paul Flicek, Fergal J Martin, Mark Gerstein, Manolis Kellis, Anshul Kundaje, Benedict Paten, Michael L. Tress, Rory Johnson, Barbara Uszczynska-Ratajczak, Adam Frankish, Roderic Guigó

doi: https://doi.org/10.1101/2024.10.29.620654

https://www.biorxiv.org/content/10.1101/2024.10.29.620654v2

GENCODE

= Annotation V47

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

CRG Centre for Genomic Regulation

Submitted to Nature in October 2025
Second round of review, resubmission soon

1 - Can we reconstruct lncRNAs transcripts correctly?

The experimental design

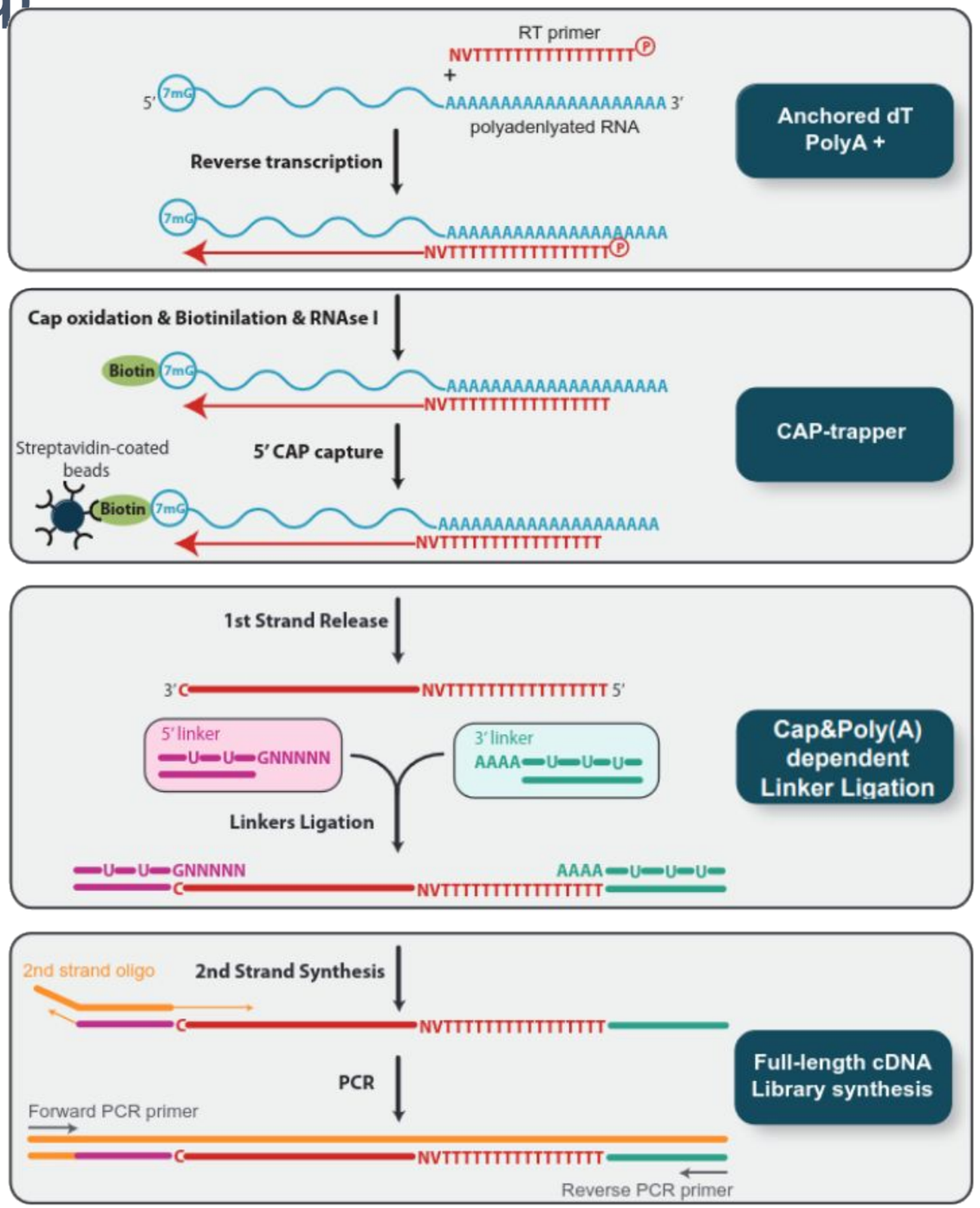
Step 1 - Improve cDNA Library (CapTrap-Seq)

nature communications

CapTrap-seq: a platform-agnostic and quantitative approach for high-fidelity full-length RNA sequencing

Sílvia Carbonell-Sala [1], Tamara Perteghella [1 2], Julien Lagarde [1 3], Hiromi Nishiyori [4], Emilio Palumbo [1], Carme Arnan [1], Hazuki Takahashi [4], Piero Carninci [4 5], Barbara Uszczynska-Ratajczak [6 7], Roderic Guigó [8 9]

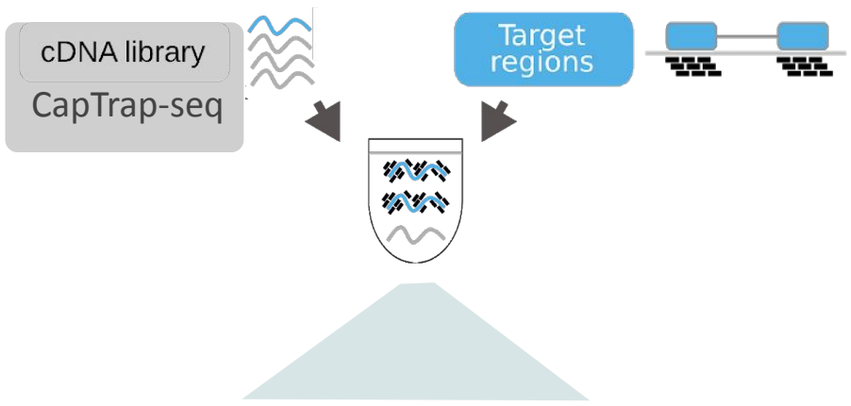
CapTrap-seq enhances full length long-read RNA sequencing for accurate genome annotation. By combining Cap-trapping and oligo(dT) priming to detect full-length 5' capped transcripts. Tested with ONT and PacBio show it as a competitive, platform-agnostic method.



1 - Can we reconstruct lncRNAs transcripts correctly?

The experimental design

Step 1 - Improve cDNA Library (CapTrap-Seq)
Step 2 - Enrich Target regions by CLS (Capture long-read Sequencing)



Capture Long-Read (CLS)

Target transcripts from genomic regions not explored by standard experimental approaches.

> Nat Genet. 2017 Dec;49(12):1731-1740. doi: 10.1038/ng.3988. Epub 2017 Nov 6.

High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing

Julien Lagarde [1 2], Barbara Uszczynska-Ratajczak [1 2], Silvia Carbonell [3], Sílvia Pérez-Lluch [1 2], Amaya Abad [1 2], Carrie Davis [4], Thomas R Gingeras [4], Adam Frankish [5], Jennifer Harrow [5], Roderic Guigo [1 2], Rory Johnson [1 2]

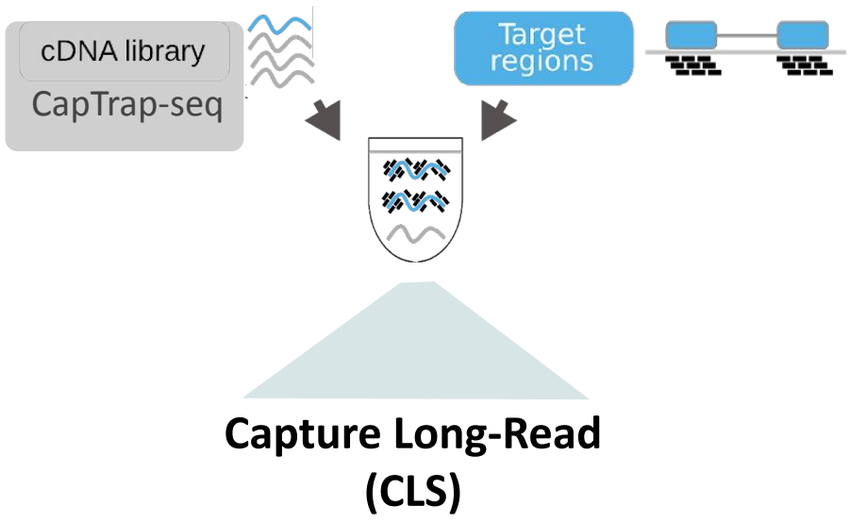
Affiliations + expand

PMID: 29106417 PMCID: PMC5709232 DOI: 10.1038/ng.3988

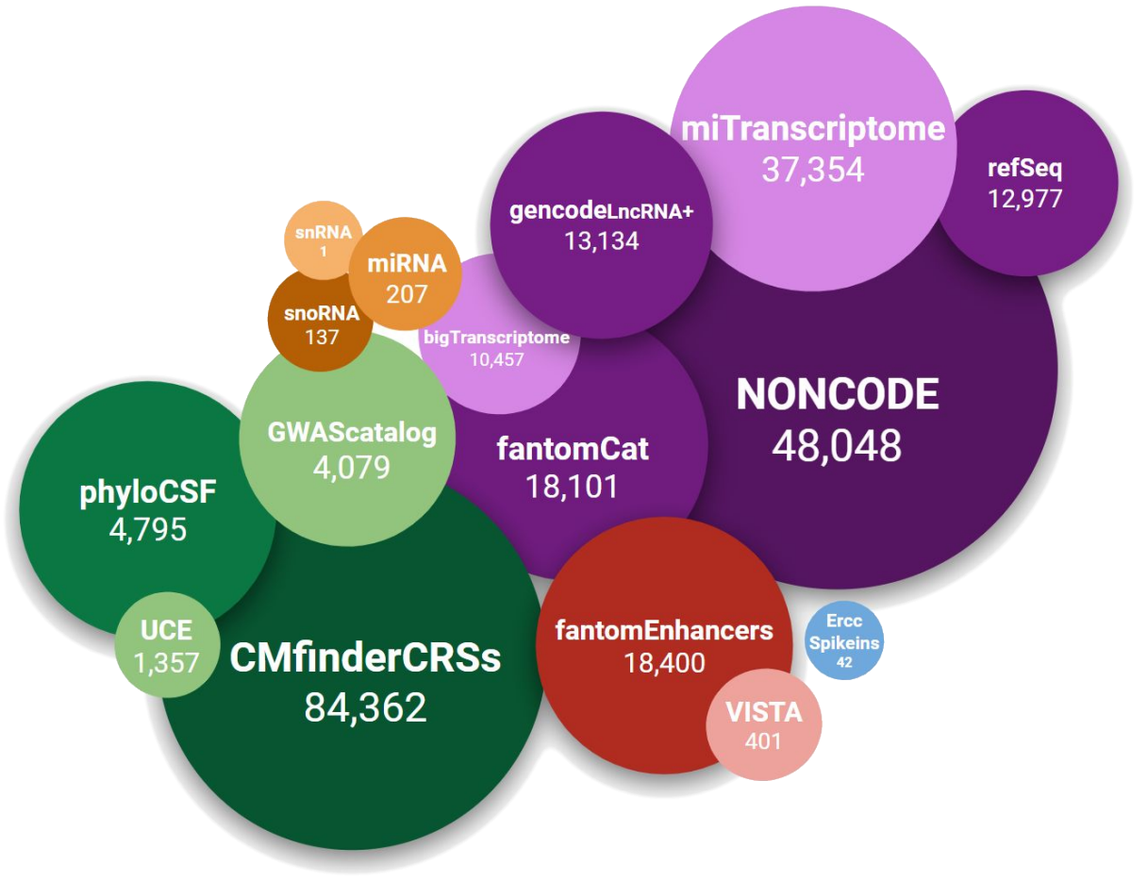
1 - Can we reconstruct lncRNAs transcripts correctly?

The experimental design

Step 1 - Improve cDNA Library (CapTrap-Seq)
Step 2 - Enrich Target regions by CLS (Capture long-read Sequencing)



Target transcripts from genomic regions not explored by standard experimental approaches.

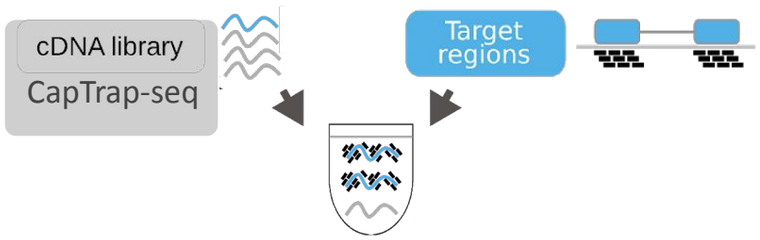


A total of ***176,435 features covering 84Mb*** *(2.8% of the genome)*.

1 - Can we reconstruct lncRNAs transcripts correctly?

The experimental design

Step 1 - Improve cDNA Library (CapTrap-Seq)
Step 2 - Enrich Target regions by CLS (Capture long-read Sequencing)
Step 3 - Multi species and tissues



88 long read samples for a total of 736M reads.

Capture Long-Read (CLS)

Capture		Short reads	Long reads	CapTrap-Seq cDNA				CLS-Enriched cDNA			
				pre-capture				post-capture			
Dev. Stage				Adult		Embryo		Adult		Embryo	
Species				Human	Mouse	Human	Mouse	Human	Mouse	Human	Mouse
Brain		4	16	*	*	*	*				
Heart		4	16	*	*	*	*				
Liver		4	16	*	*	*	*				
White Blood		4	8	*	*						
iPSC/ESC		4	8			*	*				
Testis		0	8								
Placenta		0	4								
Cell lines Pool		0	4								
Tissues Pool		0	8								
TOTALS		20	88	7	6	5	4	7	6	5	4

*Illumina short-reads sequencing provided

1 - Can we reconstruct lncRNAs transcripts correctly?

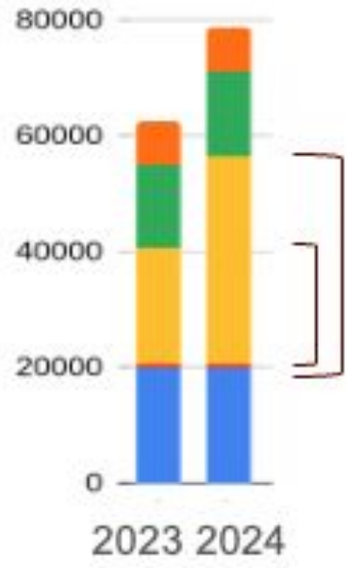
The experimental design

Step 1 - Improve cDNA Library (CapTrap-Seq)
Step 2 - Enrich Target regions by CLS (Capture long-read Sequencing)
Step 3 - Multi species and tissues
Step 4 - Stringent filters (using several methods for transcript identification), sometimes combined with manual selection, to identify lncRNAs

1 - Can we reconstruct lncRNAs transcripts correctly?

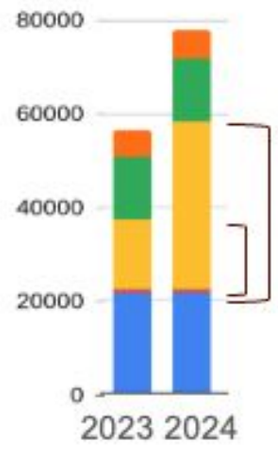
Results

lncRNA	v46	v47
Genes	20,424	**37,241**



lncRNA	Human v46	Mouse vM35
Novel genes	**16,817**	**22,210**
• spliced genes	15,283	20,537
• single-exon genes	1,534	1,673
Modified genes	9,919	7,655
Novel transcripts	**132,049**	**131,546**
• in novel genes	44,996	60,670
• in modified genes	87,053	70,876

	vM35	vM36
Genes	15,131	**37,341**

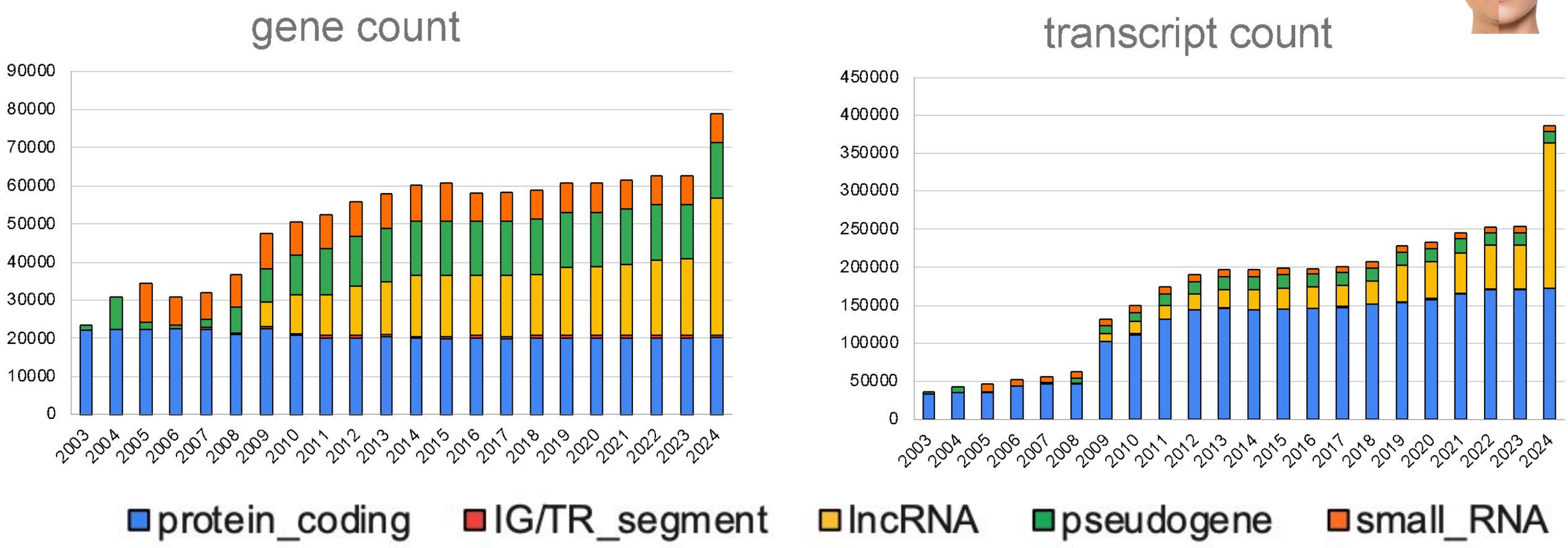


2× human transcript and gene for lncRNAs

6× mouse transcript and 2× gene for lncRNAs

1 - Can we reconstruct lncRNAs transcripts correctly?

Results



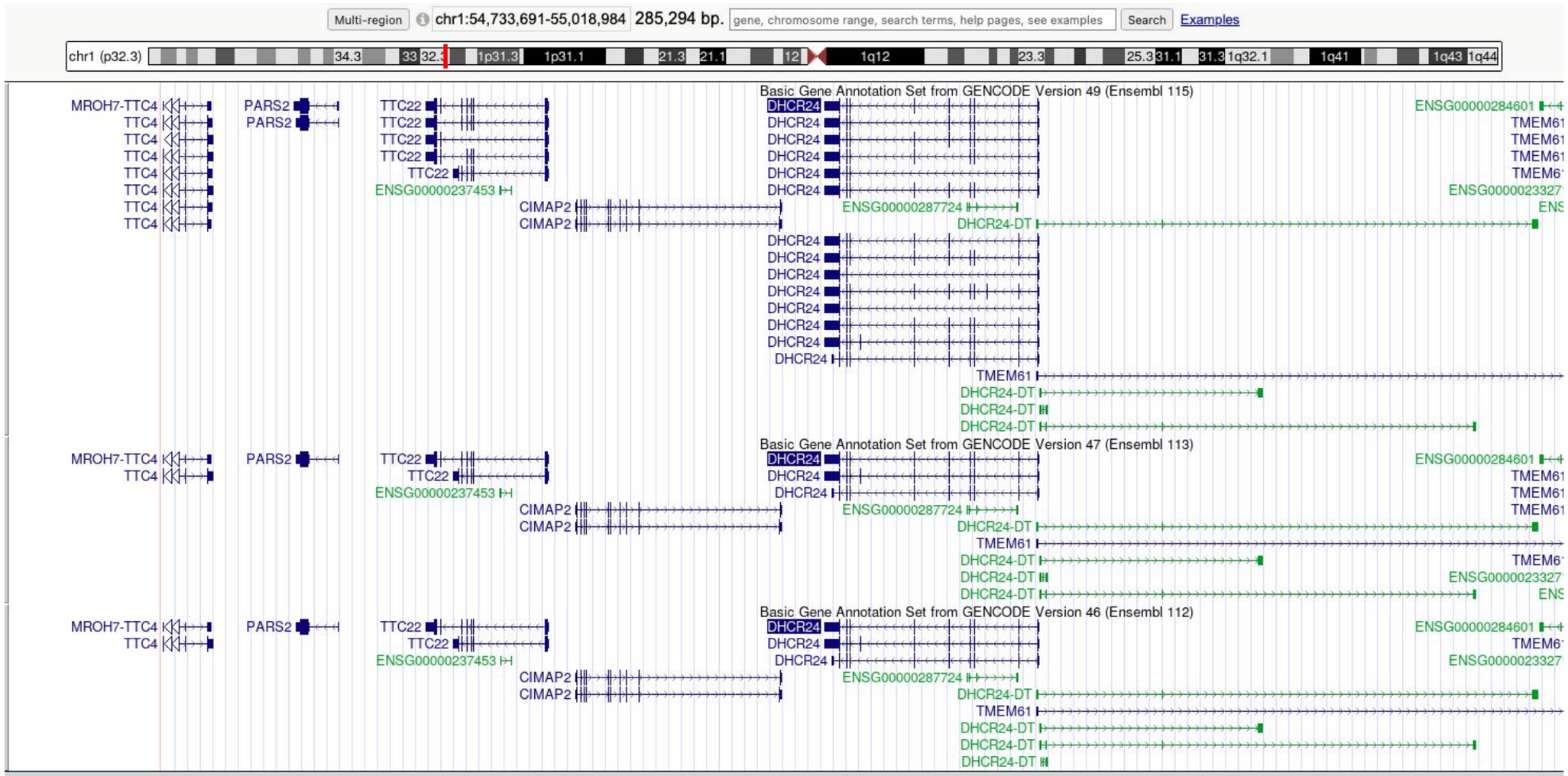
In the last decade, the number of **annotated human genes has remained stable**, while **annotated transcripts increased by 25%**, **“**mainly due to the addition of alternative lncRNA isoforms.”

Not so true anymore … with Gencode V49
for PCG 278,455 transcript (89,832 in V47)

1 - Can we reconstruct lncRNAs transcripts correctly?

Results

https://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr1%3A54733691%2D55018984&hgsid=440286989_PIMbeEfTLcxailT7e7a9konaKABg



Long reads can provide:

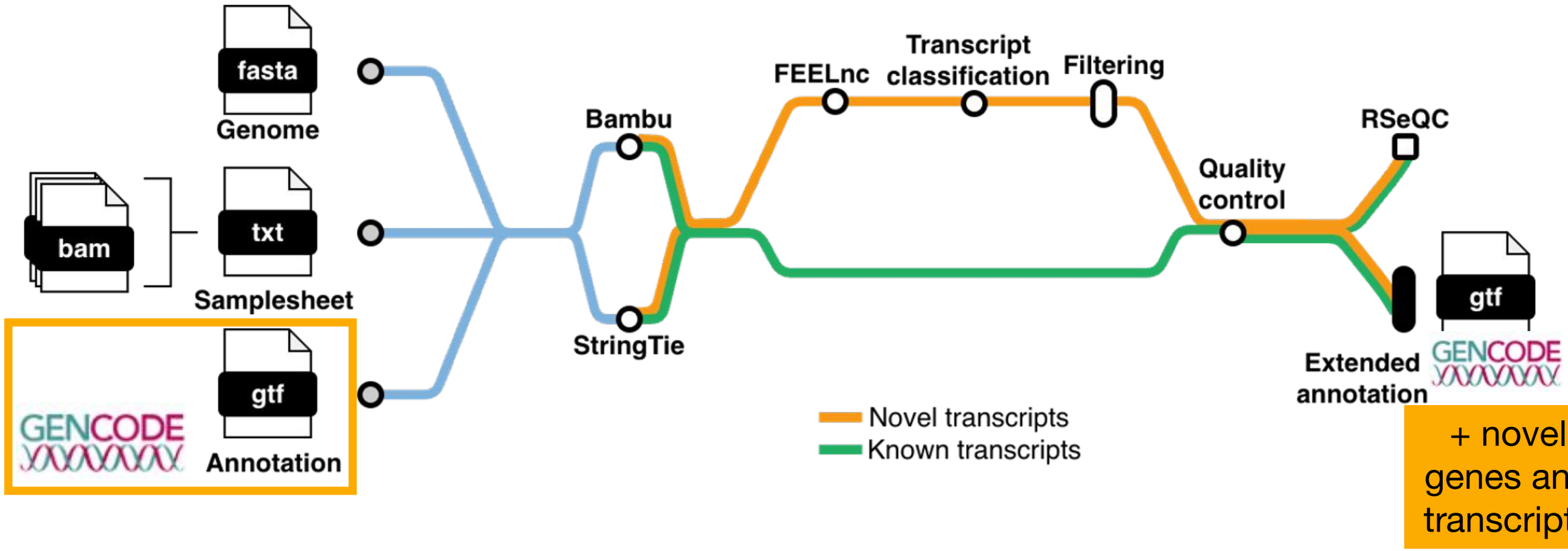
✓ complete exon chains

✓ transcript boundaries

✓ splice junction validation

1 - Can we reconstruct lncRNAs transcripts correctly?

ANEXA : ANalysis with Nextflow for EXtended Annotation



https://github.com/IGDRion/ANNEXA

nextflow

FAIR

Findable Accessible Interoperable Reusable

What do long reads actually change for lncRNAs?

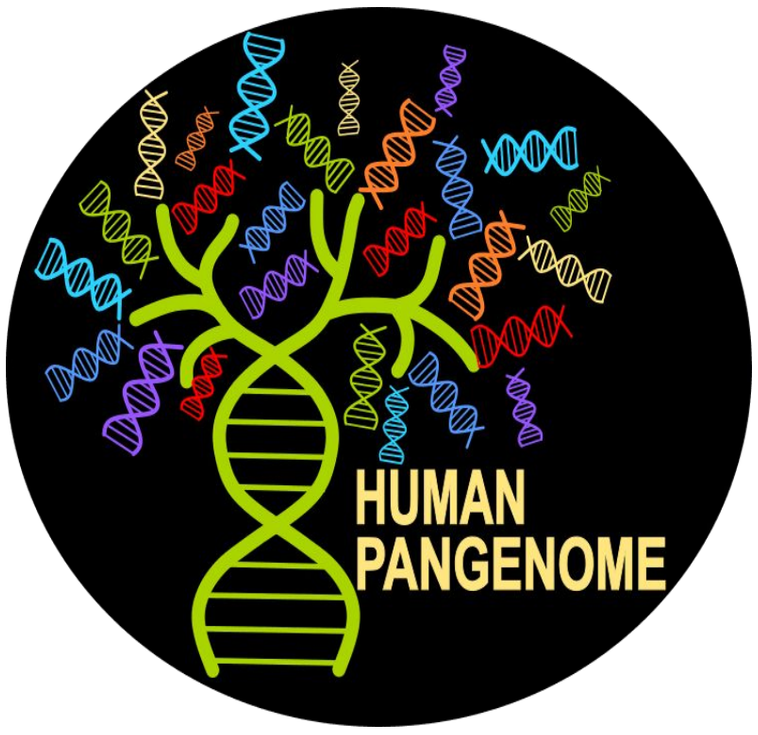
2 - Can one annotation represent the entire diversity?

(PanTranscriptome)

2 - Can one annotation represent the entire diversity?

Thousands of gene loci and transcript isoforms are still unknown.
Accurate and complete annotation is essential to decode the biological information stored in the genome.
Many transcripts might only be expressed in a **few cells** within tissues and they **can be population specific**…

What's Next?



Towards the definition of a human Pantranscriptome.

2 - Can one annotation represent the entire diversity?

The “Human Pantranscriptome”

Article | Open access | Published: 03 December 2025

Long-read transcriptomics of a diverse human cohort reveals ancestry bias in gene annotation

Pau Clavell-Revelles, Fairlie Reese, Sílvia Carbonell-Sala, Fabien Degalez, Carme Arnan, Winona Oliveros, Emilio Palumbo, Tamara Perteghella, Roderic Guigó ✉ & Marta Melé ✉

Nature Communications **16**, Article number: 10194 (2025) | Cite this article

4859 Accesses | **130** Altmetric | Metrics

https://www.nature.com/articles/s41467-025-66096-x

NIH NHGRI

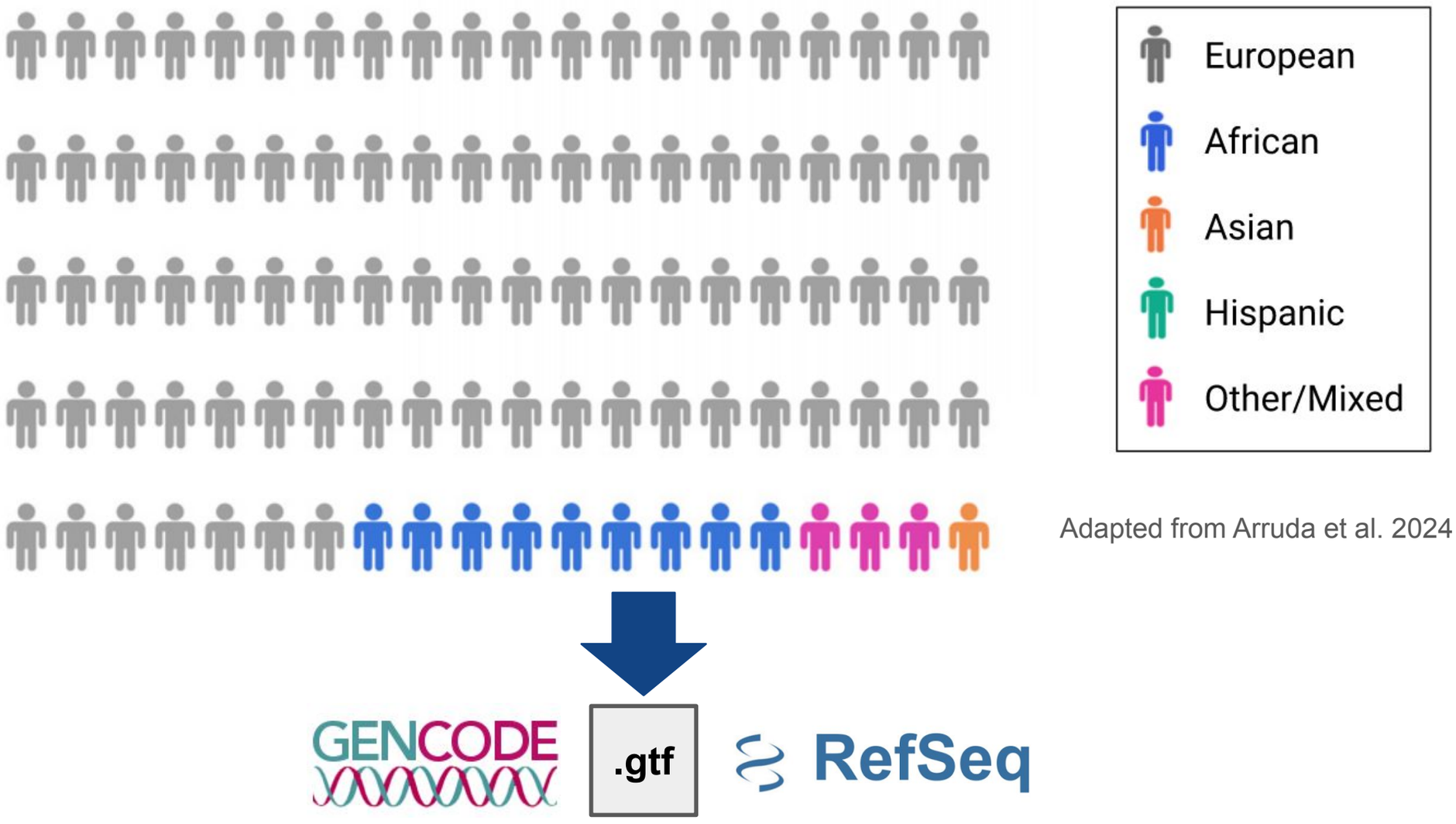
GENCODE

CRG Centre for Genomic Regulation

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

2 - Can one annotation represent the entire diversity?

RNA-seq studies overrepresent Europeans



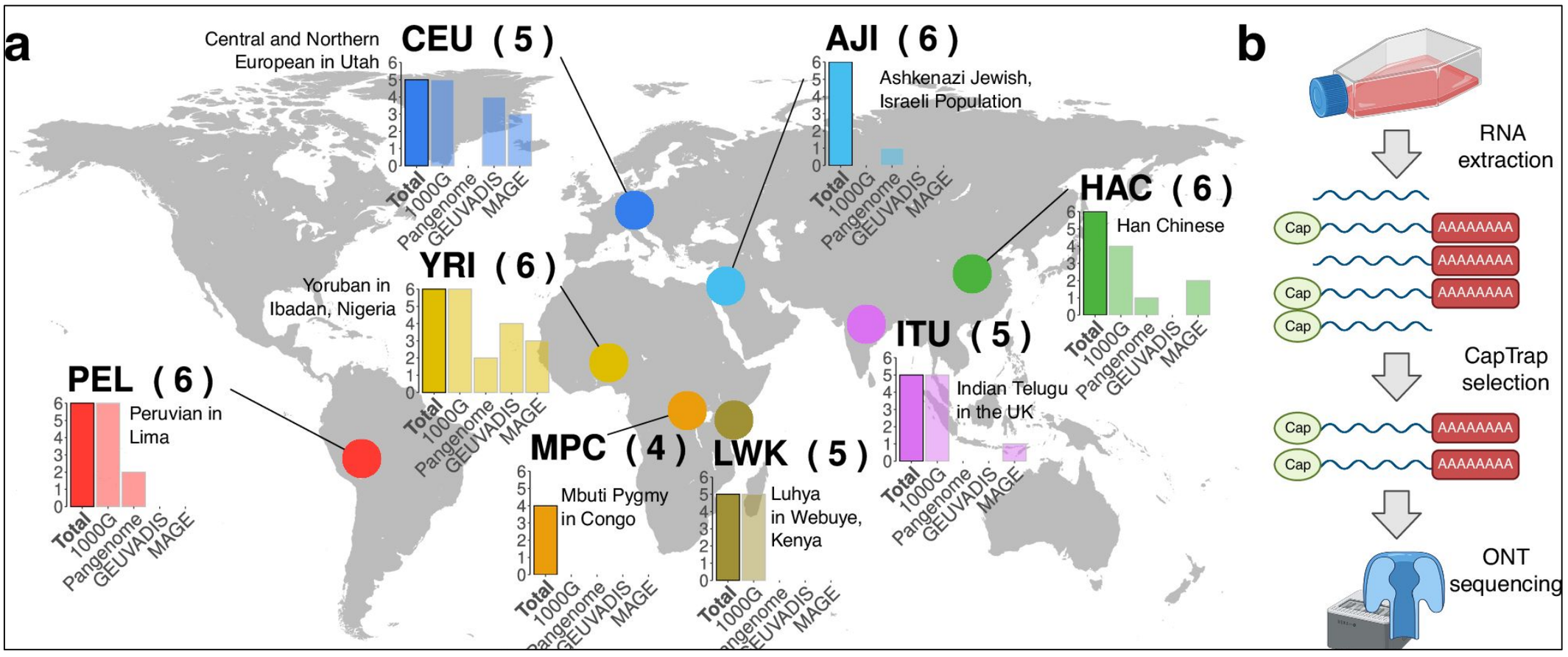
Adapted from Arruda et al. 2024

Human gene annotations are mainly built with European-descent individuals

2 - Can one annotation represent the entire diversity?

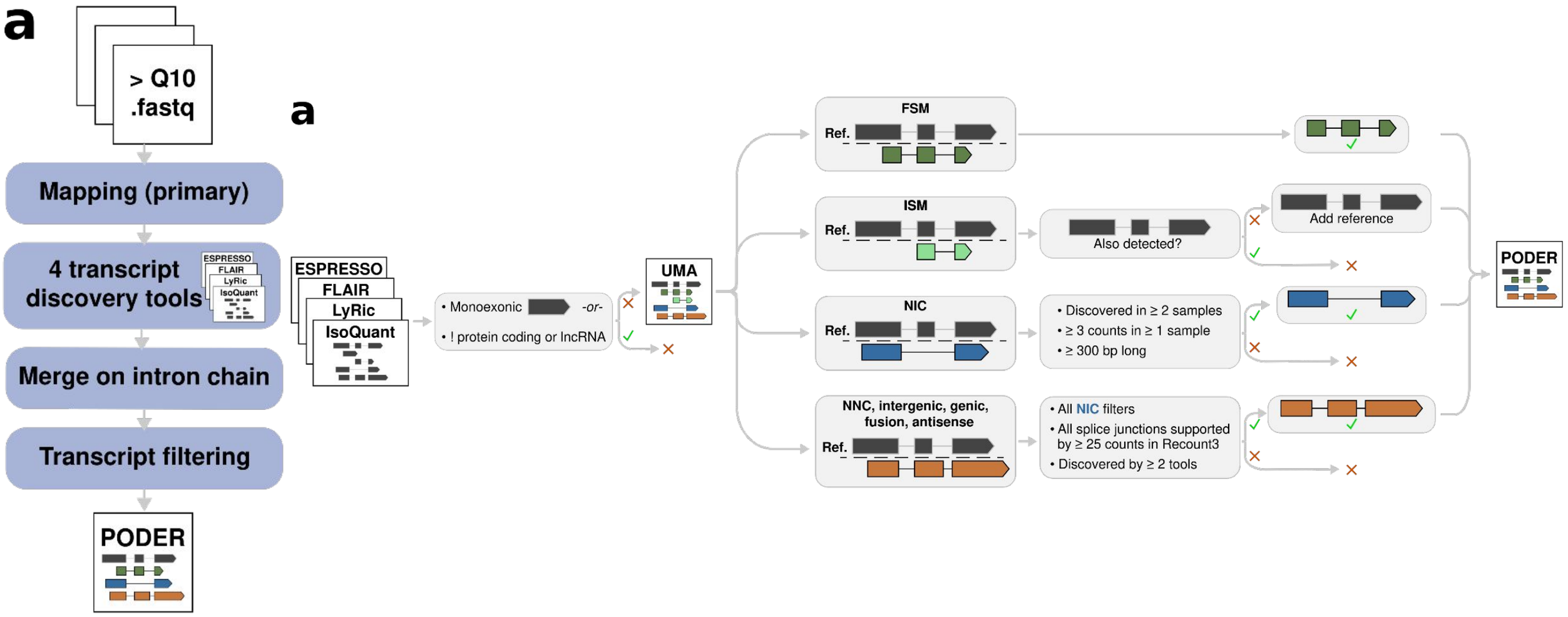
The dataset

Lymphoblastoid cell lines



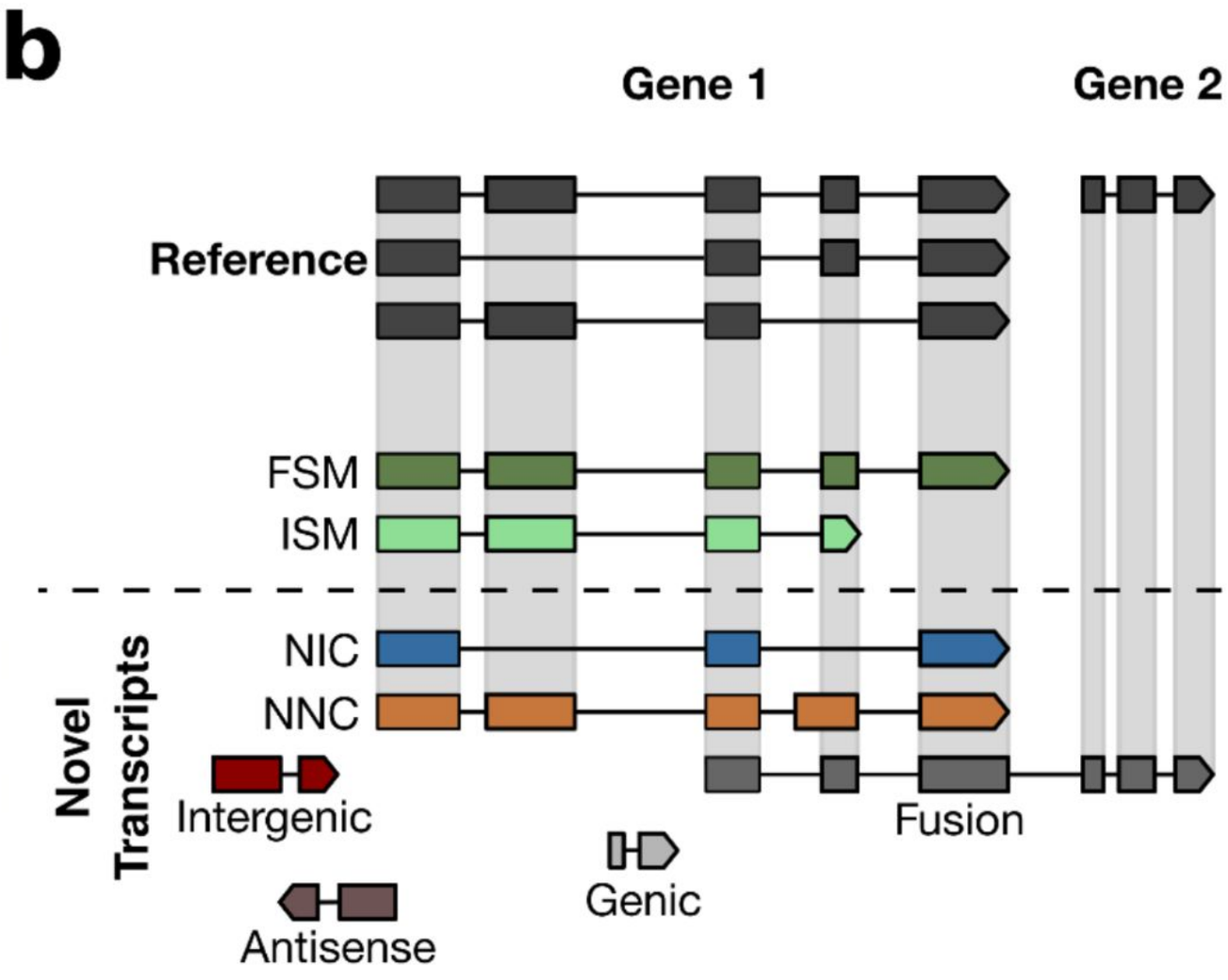
2 - Can one annotation represent the entire diversity?

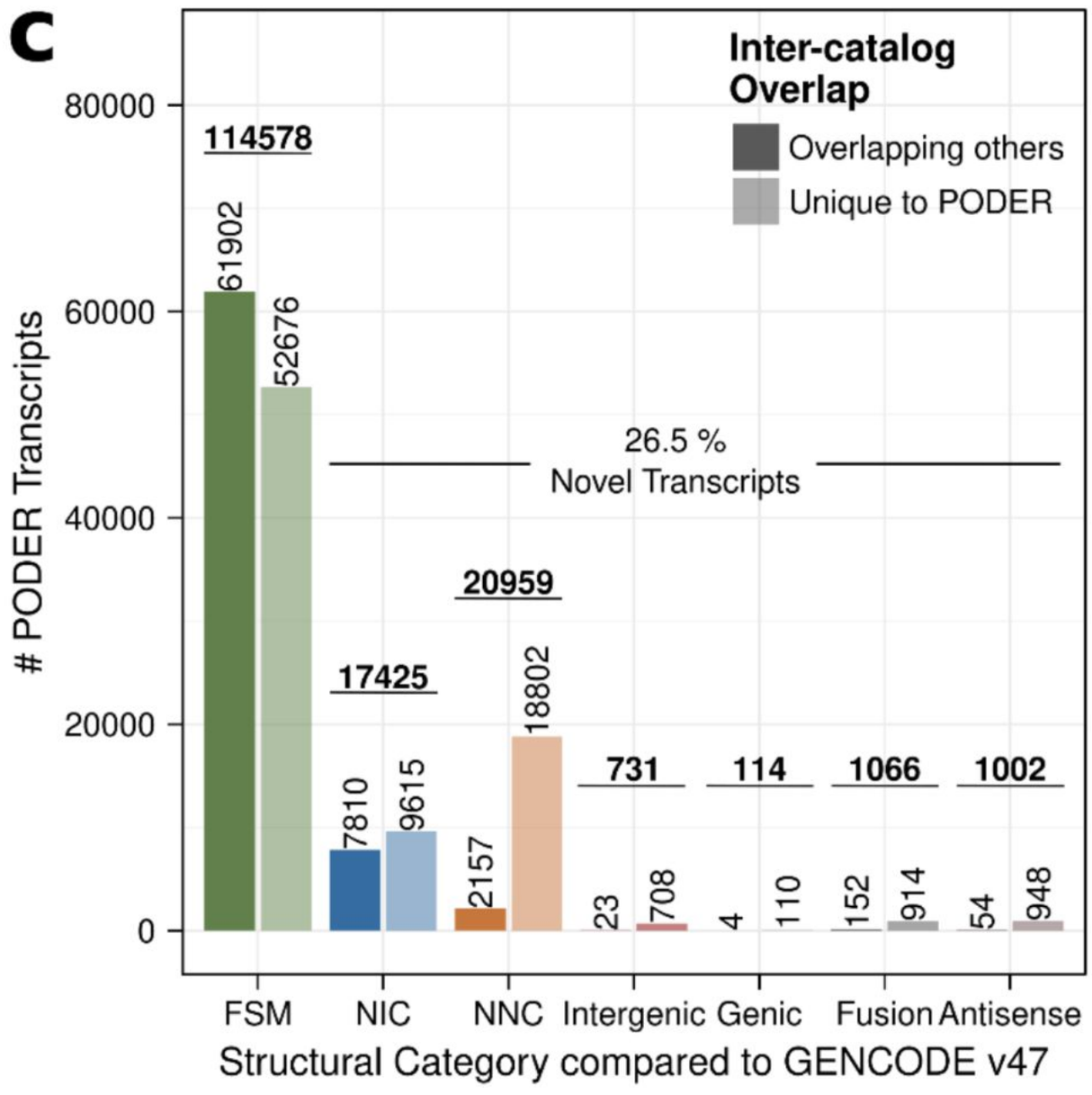
The method



2 - Can one annotation represent the entire diversity?

The result(s)





41,293 novel, high-confidence transcripts (both PCG and lncRNAs)

What do long reads actually change for lncRNAs?

3 - How do we prioritize which lncRNAs matter?

(lncRNA_PutOrtho)

3 - How do we prioritize which lncRNAs matter?

Long-read annotation

↓

Orthology inference

↓

Evolutionary conservation

↓

Candidate prioritization

Not every novel transcript is functional equally.

Comparative genomics helps prioritize candidates

3 - How do we prioritize which lncRNAs matter?

MANUSCRIPT TITLE

An integrative workflow for lncRNA orthology detection and its application to 13 evolutionarily diverse species

AUTHORS AND AFFILIATIONS

Fabien DEGALEZ[1,*], Coralie ALLAIN[1], Laetitia LAGOUTTE[1], Frédéric LECERF[1], Sandrine LAGARRIGUE[1,*]

https://www.biorxiv.org/content/10.1101/2024.10.03.616473v1

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

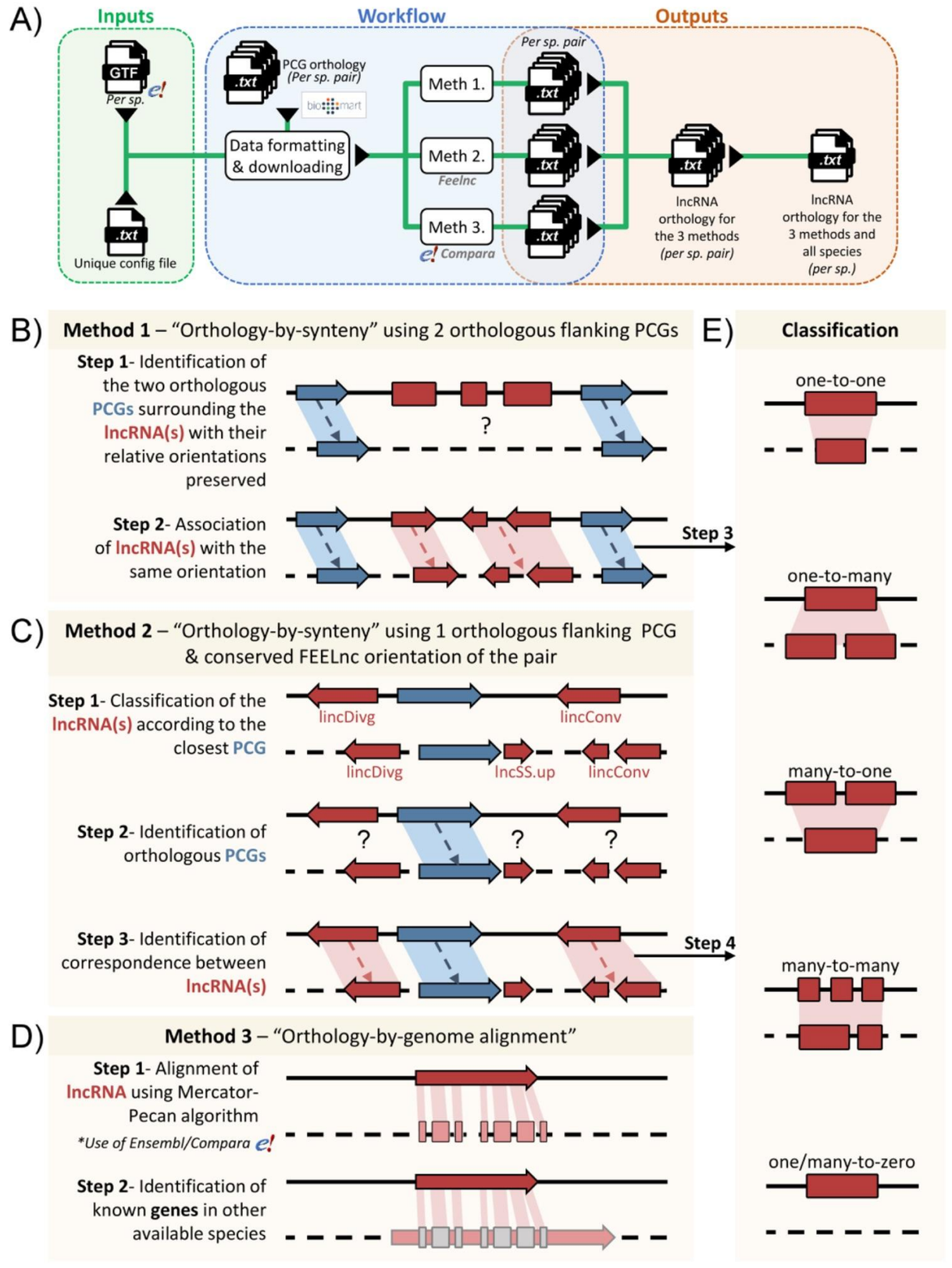
L'INSTITUT agro Rennes Angers
INRAE

Submitted to Nucleic Acid Research in October 2024

Accepted last week ! 🎉

3 - How do we prioritize which lncRNAs matter?

https://github.com/fdegalez/lncrna_PutOrtho



3 - How do we prioritize which lncRNAs matter?

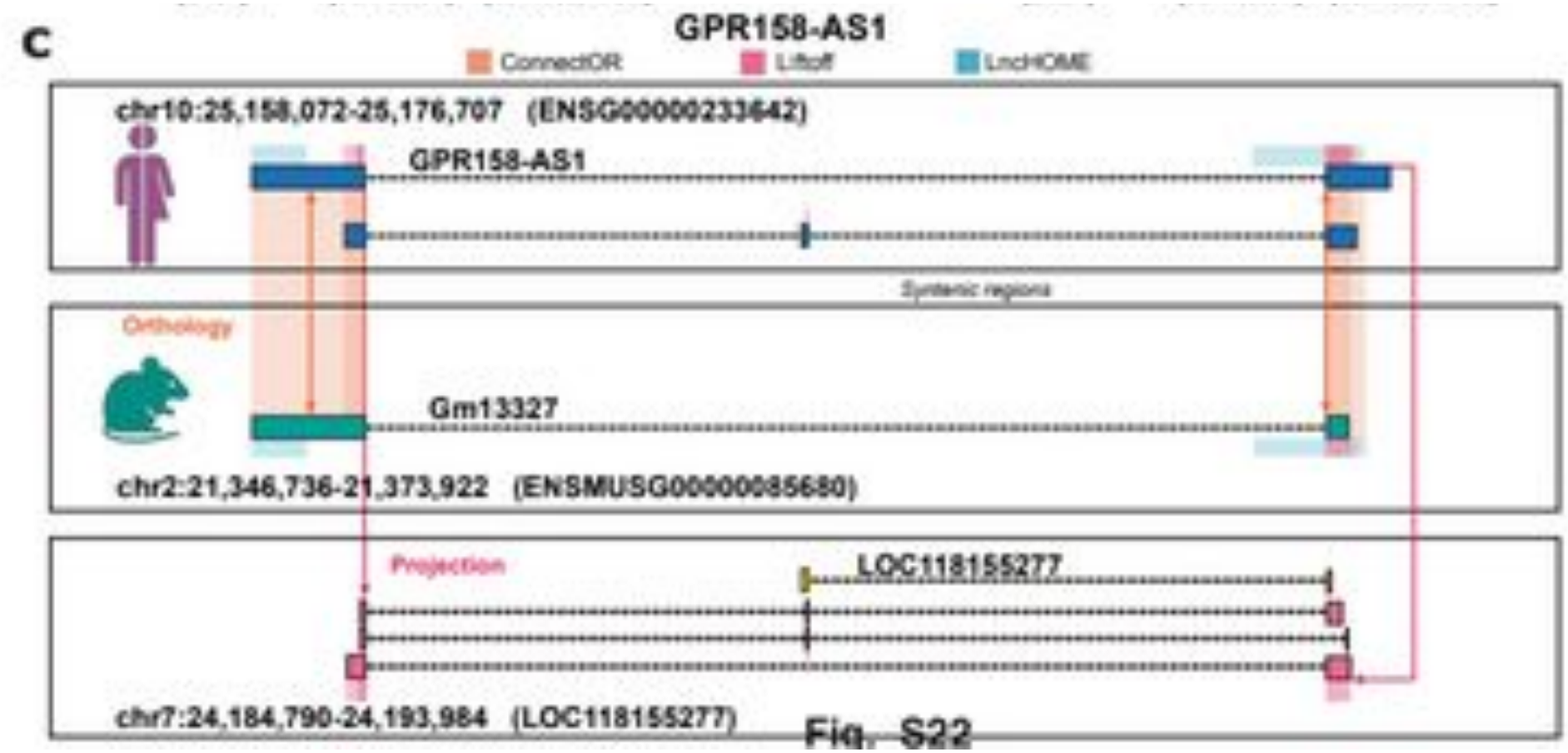


Fig. S22

GENCODE

Take-home messages

Take-home messages

1 - Can we reconstruct lncRNAs transcripts correctly?

(GENCODE-CLS / Annexa)

→ Long-read sequencing has transformed lncRNAs transcript reconstruction but different technical biais still exist

↓

2 - Can one annotation represent the entire diversity?

(PanTranscriptome)

→ Reference annotations remain incomplete because transcript diversity is biological as much as technical (especially for lncRNAs)

↓

3 - How do we prioritize which lncRNAs matter?

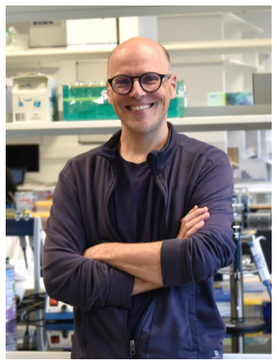
(lncRNA_PutOrtho)

Take-home messages

Better annotations are the foundation for functional genomics!

Acknowledgment

IGDRion



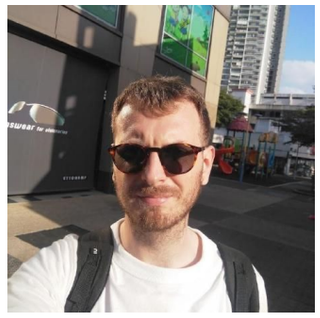
Edouard CADIEU

Aurore BESSON

Victor LE BARS

Enora CORLER

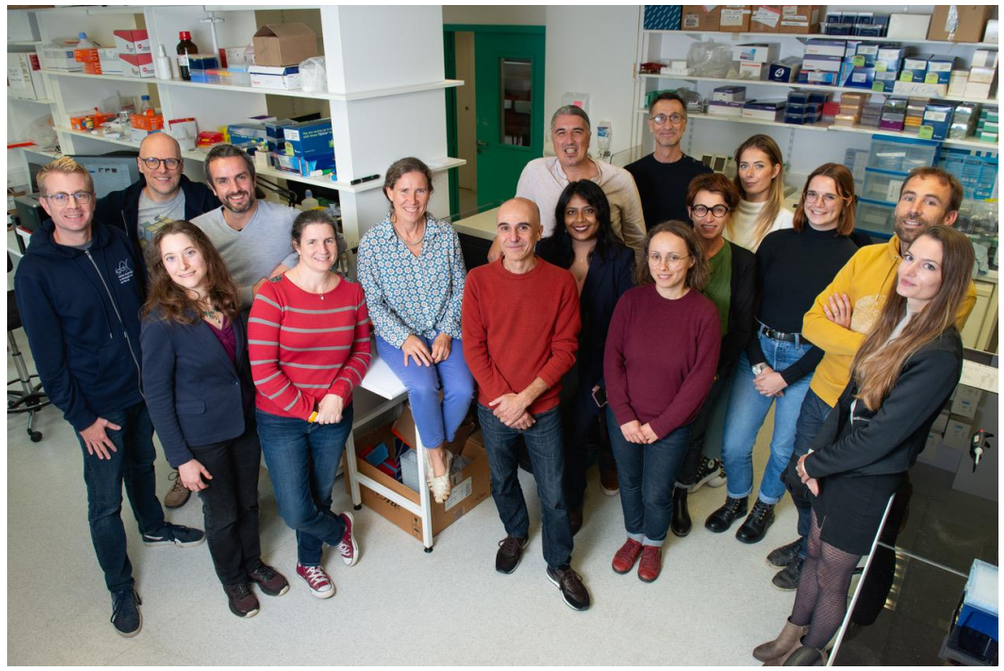
Matthias LORTHIOIS



Nicolaï HOFFMANN

Anastasia RUSAKOVICH

Canine Genetics Team @ IGDR



Catherine ANDRE

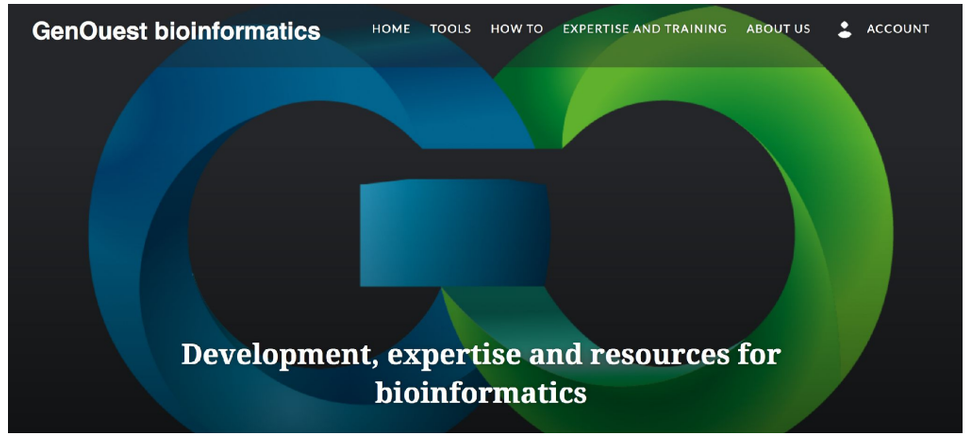
BINGO

>BIOLOGY IN SILICO for NON CODING GENOME and ONCOLOGY

Yuna BLUM

Christophe HELIGON

Genouest bioinformatic



Acknowledgment

Teams and Labs

L'INSTITUT agro Rennes Angers

INRAE

UMR PEGASE, équipe GGA

CRG Centre for Genomic Regulation

Guigo's lab

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Marta Melé's lab

Stanford University

Montgomery's lab

Institute of Bioorganic Chemistry PAS
IBCH PAS
Poznań, Poland

CobRNA

Barbara Uszczynska-Ratajczak's lab (Cobrna lab)

We found many of you.
Now we just need to
understand what you do!

Thank you!
Questions?

I'm not just background noise!

IncRNA

LONG. NON-CODING. IMPORTANT.

NEW ISOFORM?

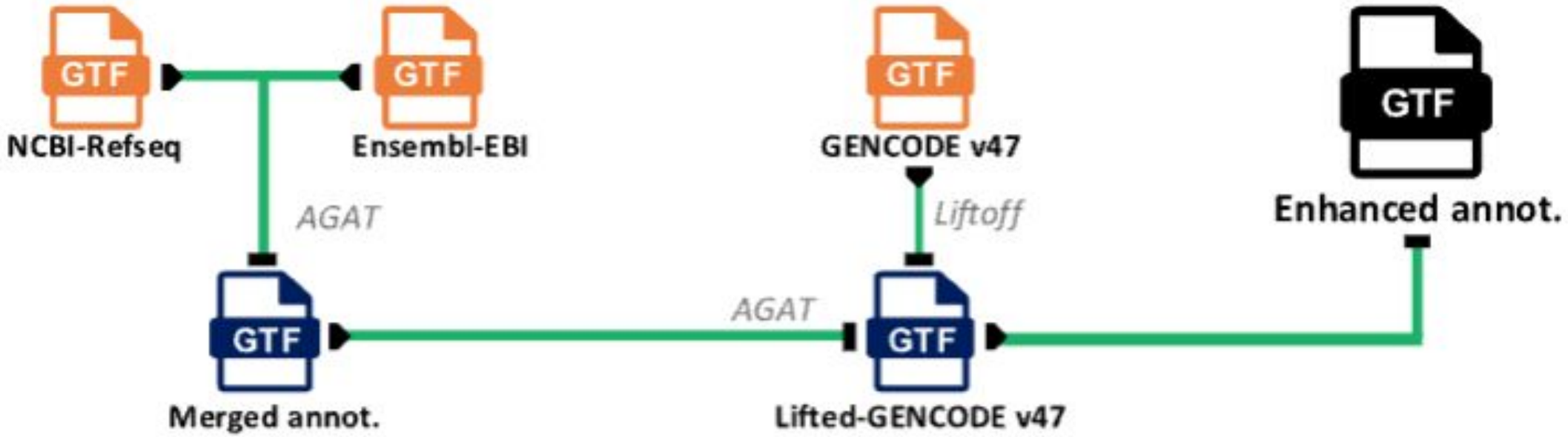
EVOLUTIONARY CONSERVATION?

FUNCTION?

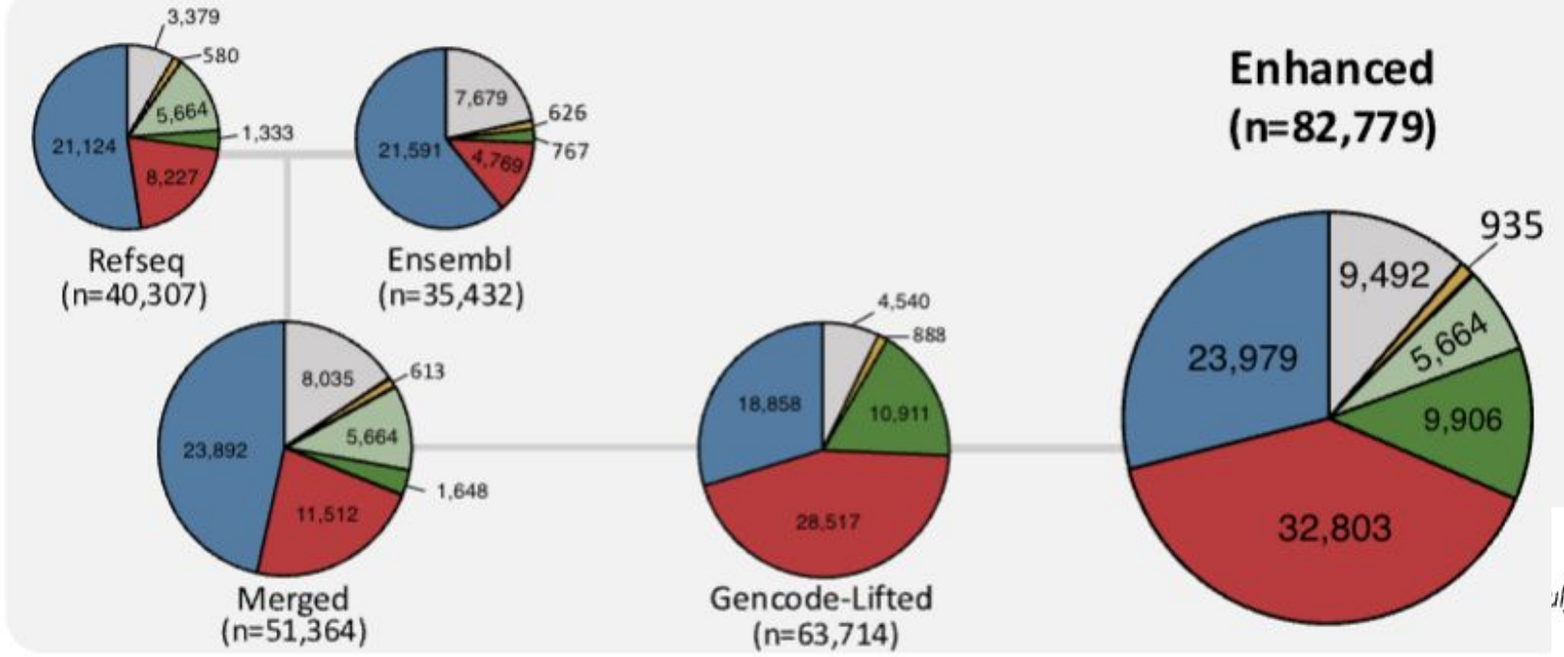
STAY TUNED!

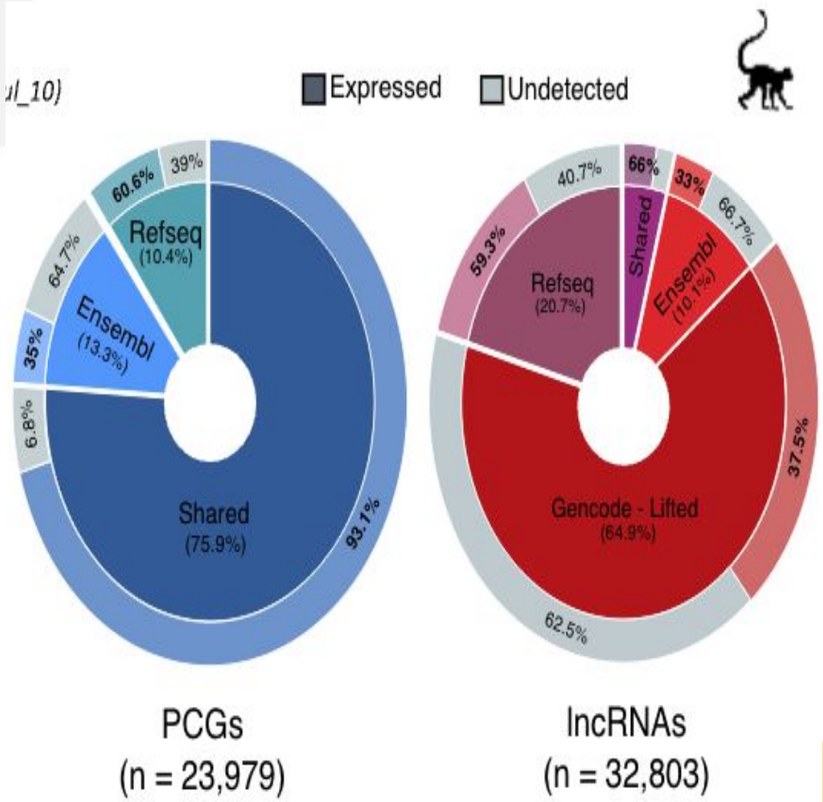
Supplementary elements

A)

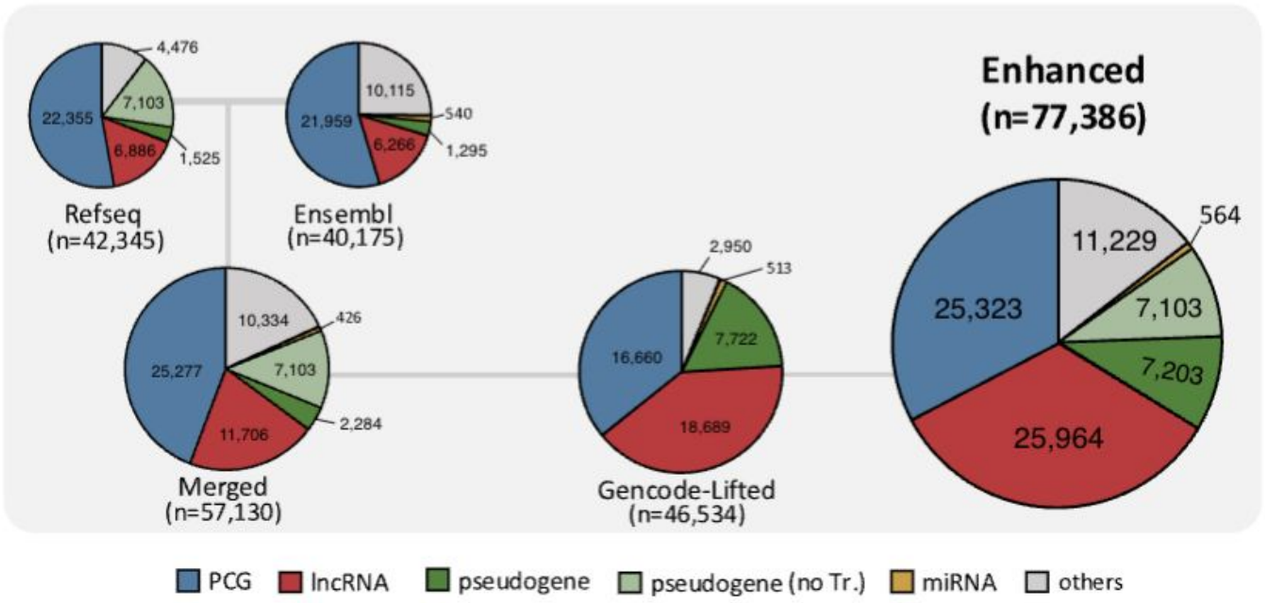

GTF
NCBI-Refseq
GTF
Ensembl-EBI
AGAT
GTF
Merged annot.
AGAT
GTF
GENCODE v47
Liftoff
GTF
Lifted-GENCODE v47
GTF
Enhanced annot.

B) **Macaque** – *Macaca mulatta (Mmul_10)*

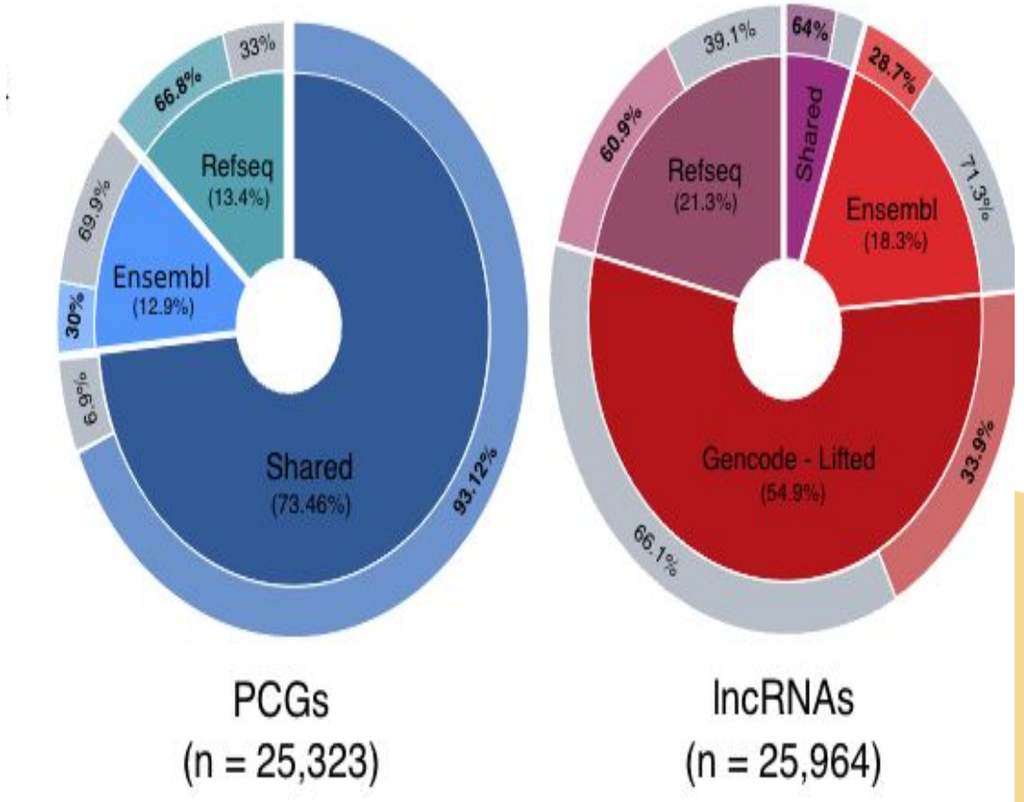




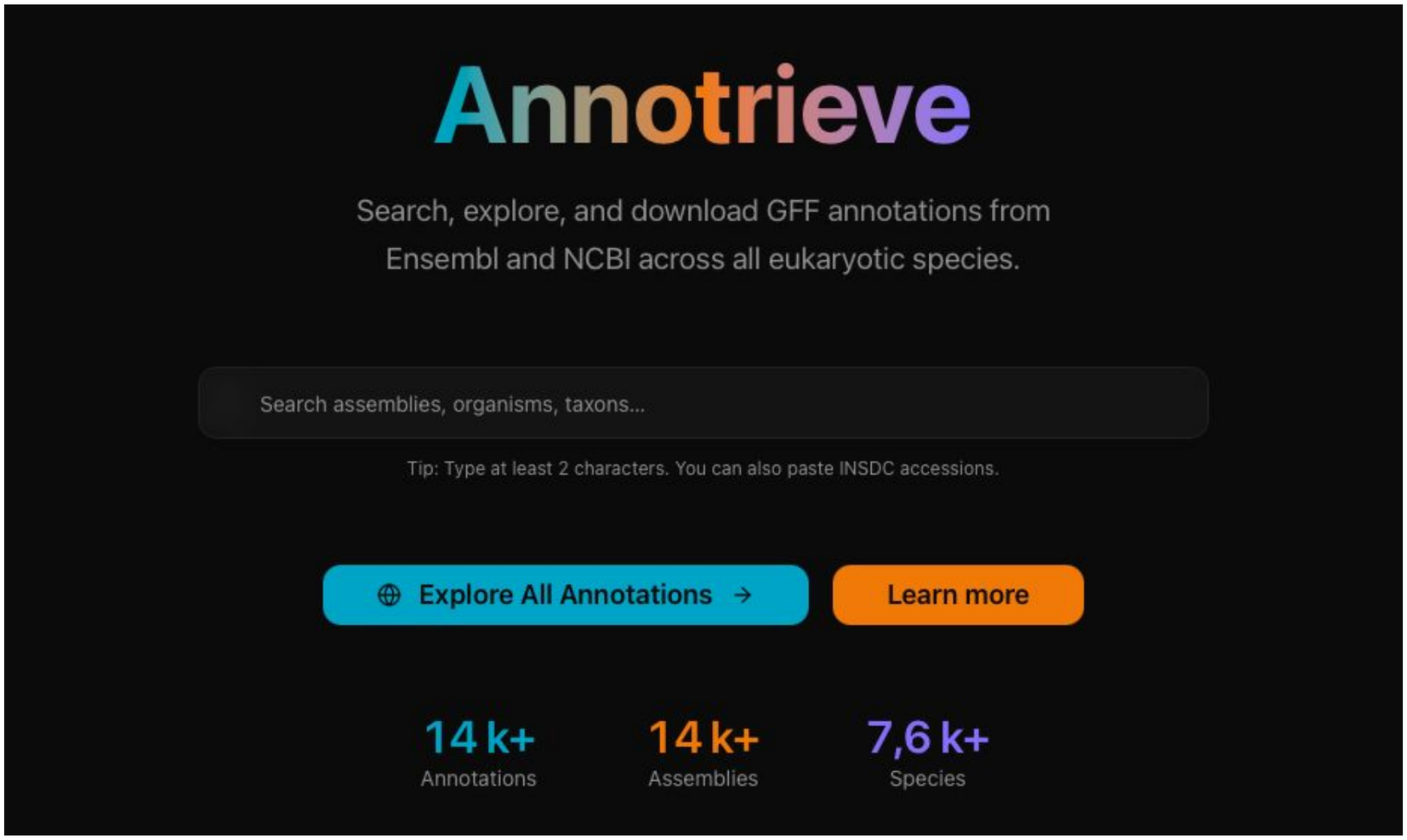
C) **Marmoset** – *Callithrix jacchus (mCalJa1.2.pat.X)*



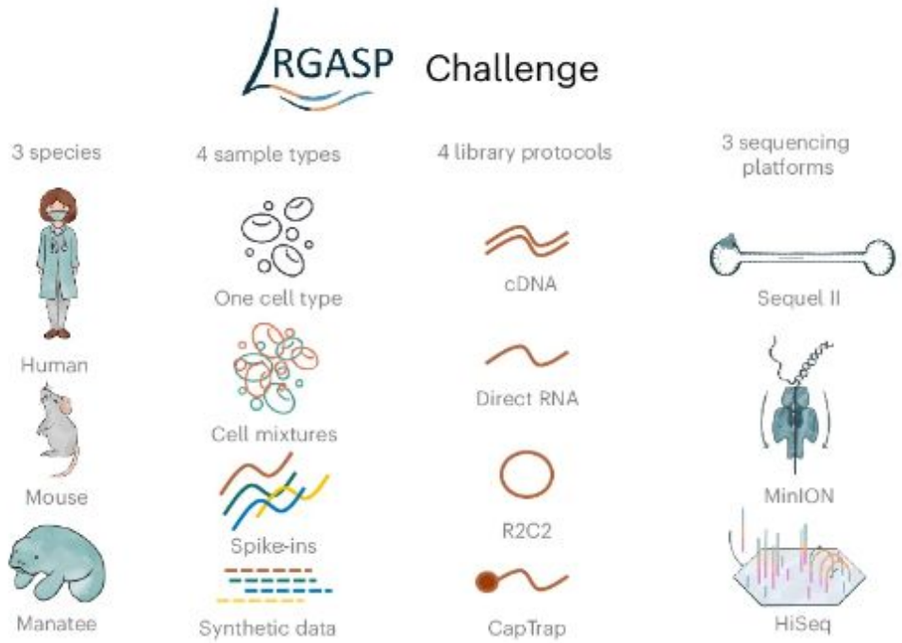
(mCalJa1.2.pat.X)



https://genome.crg.es/annotrieve/



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

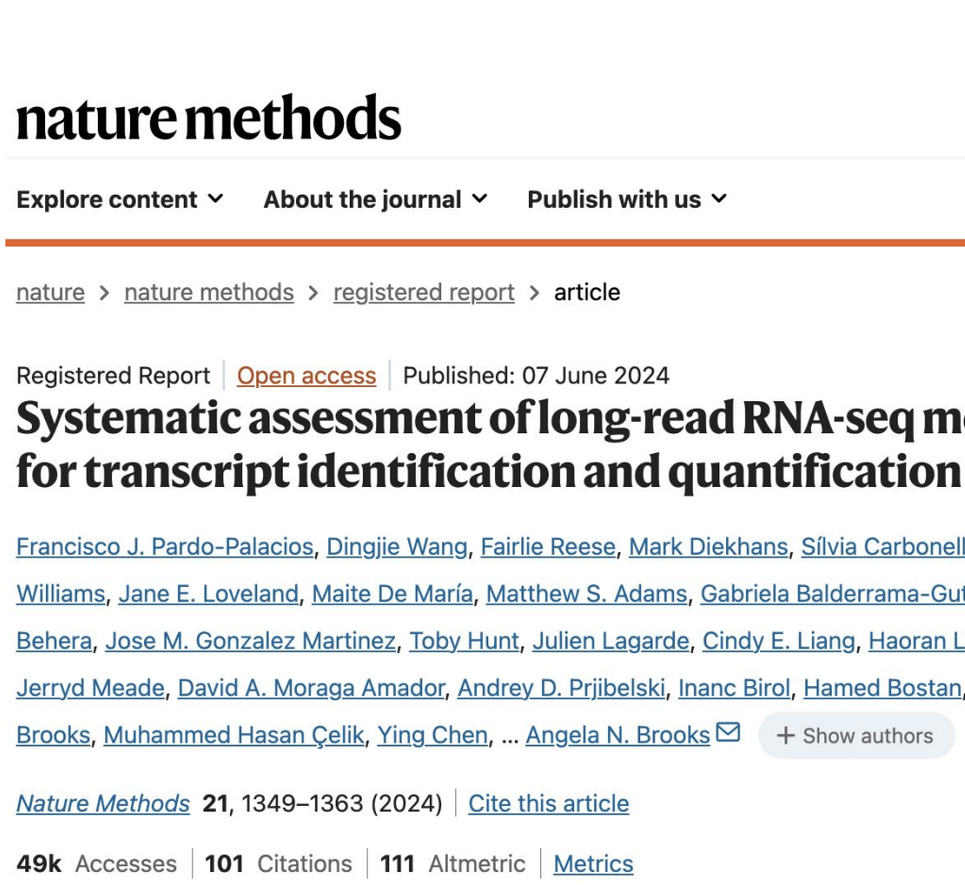
Francisco J. Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, Maite De María, Matthew S. Adams, Gabriela Balderrama-Gutierrez, Amit K. Behera, Jose M. Gonzalez Martinez, Toby Hunt, Julien Lagarde, Cindy E. Liang, Haoran Li, Marcus Jerryd 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


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

49k Accesses | **101** Citations | **111** Altmetric | Metrics

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

➡ Better multiple tools

Long-read RNA sequencing : state of the art

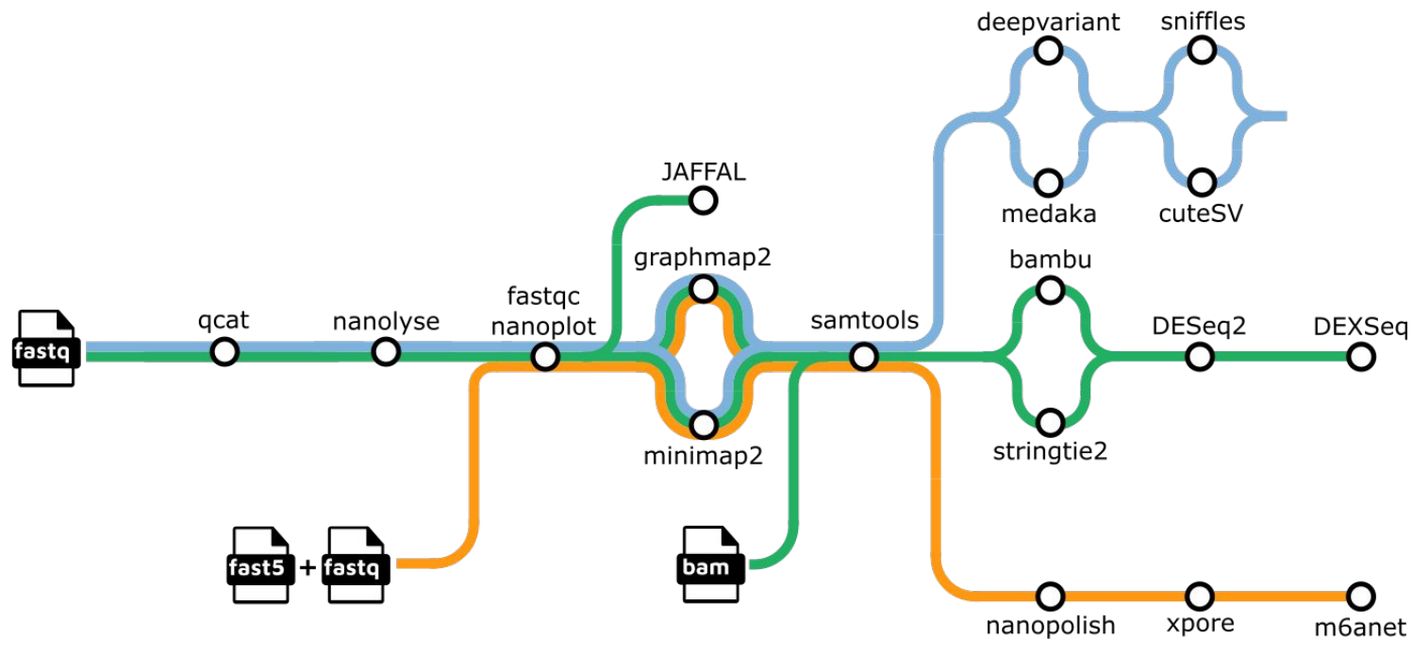
	Multiple tools for discovery	Annotate lncRNAs	Transcript classification	Full-length assessment	

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/

nf-core/nanoseq

Nanopore demultiplexing, QC, alignment, and downstream analysis



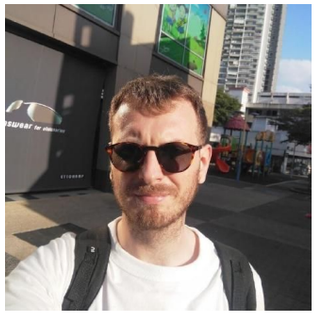
Long-read RNA sequencing : state of the art

	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

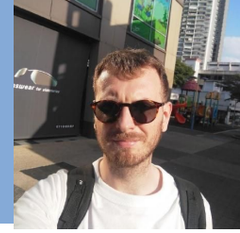
Long-read RNA sequencing : state of the art

	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/

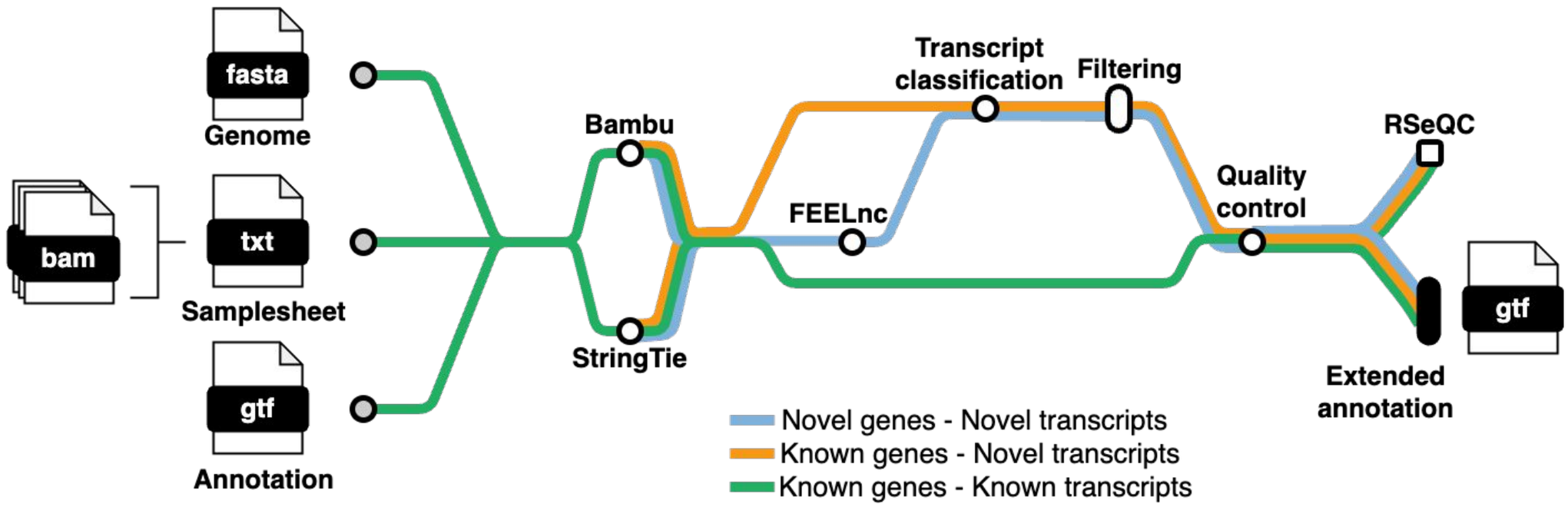
Nicolaï HOFFMANN



ANNEXA



ANalysis with Nextflow for EXtended Annotation



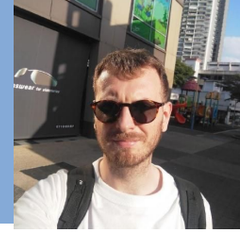
IGDRion/ANNEXA

nextflow

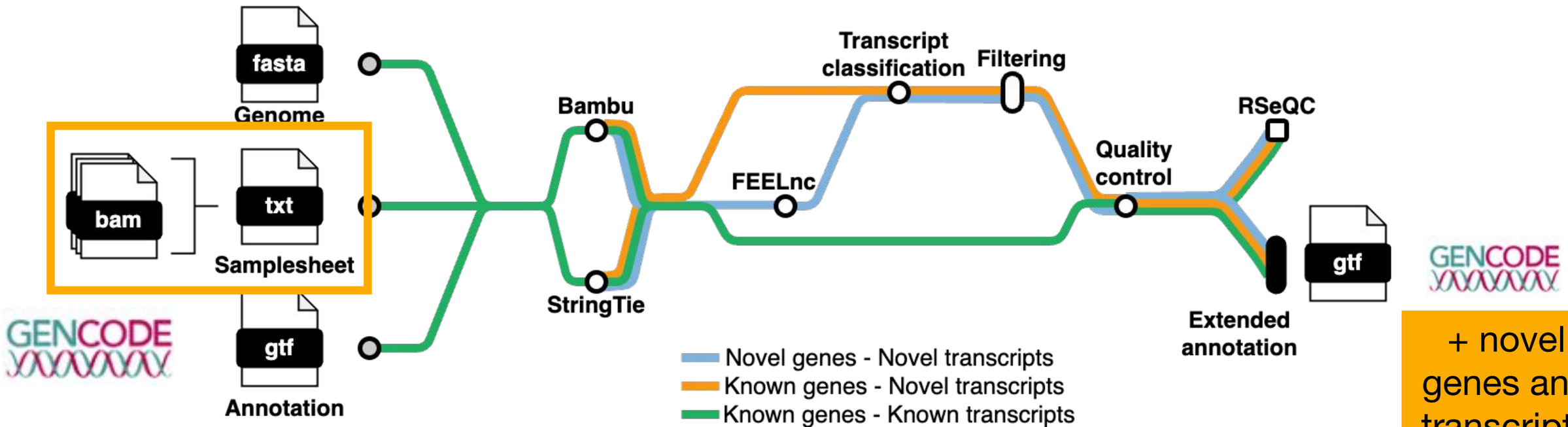
F A I R

Findable Accessible Interoperable Reusable

ANNEXA



ANalysis with Nextflow for EXtended Annotation



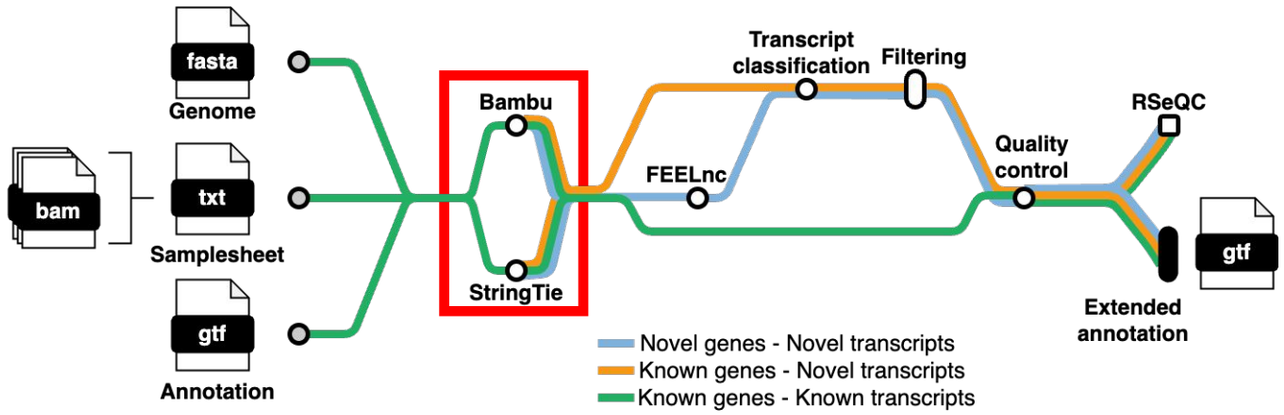
IGDRion/ANNEXA

nextflow

F A I R

Findable Accessible Interoperable Reusable

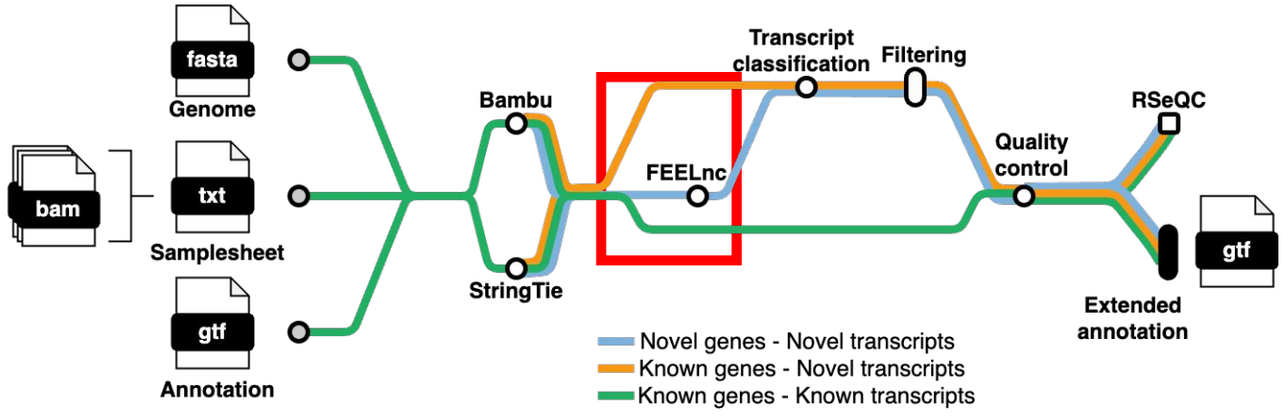
ANNEXA : Transcriptome discovery



IGDRion/ANNEXA

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

ANNEXA : Coding potential



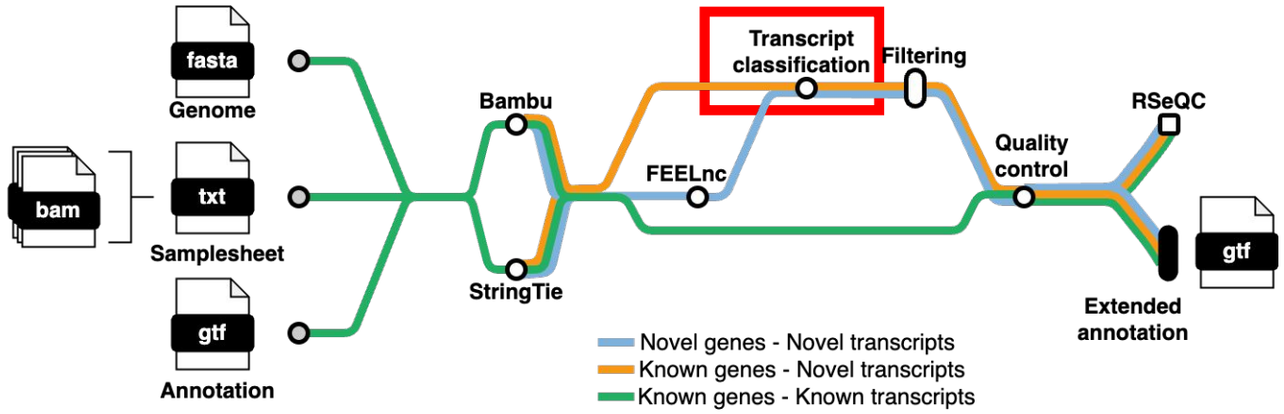
FEELnc: Wucher et al., 2017

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

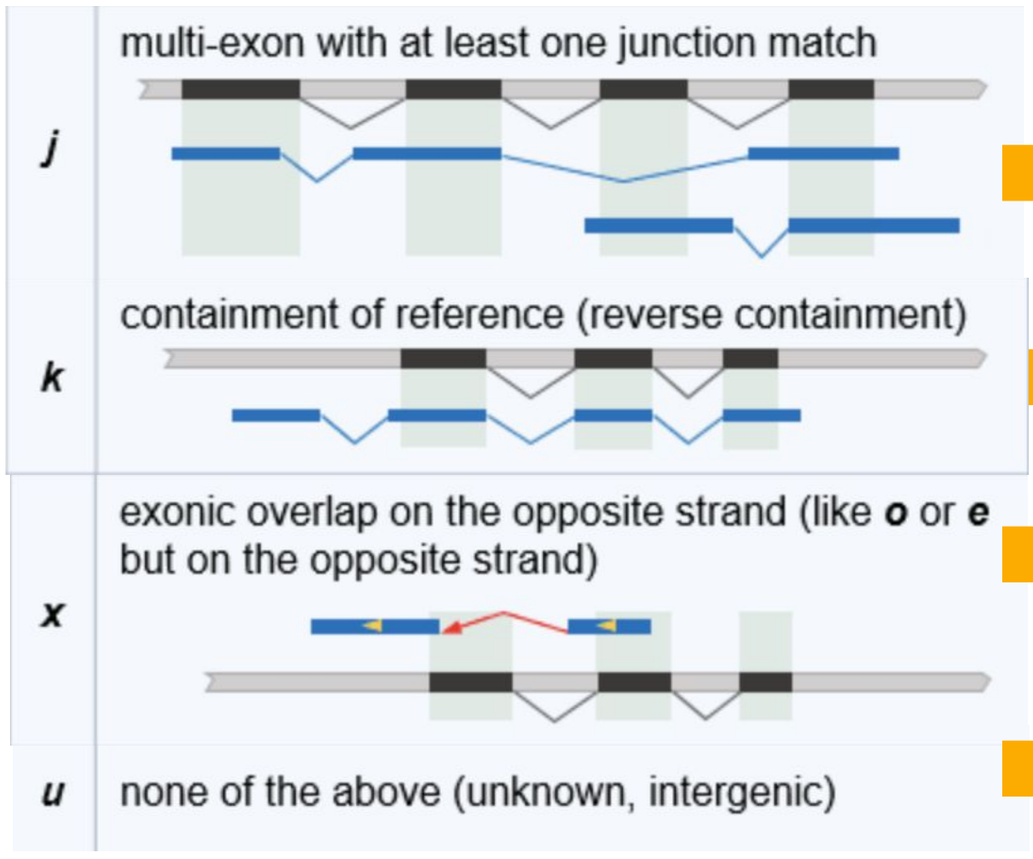
→ Allows to discriminate between novel mRNAs versus lncRNAs

ANNEXA : Coding potential



IGDRion/ANNEXA

gffCompare: Pertea G and Pertea M, 2020



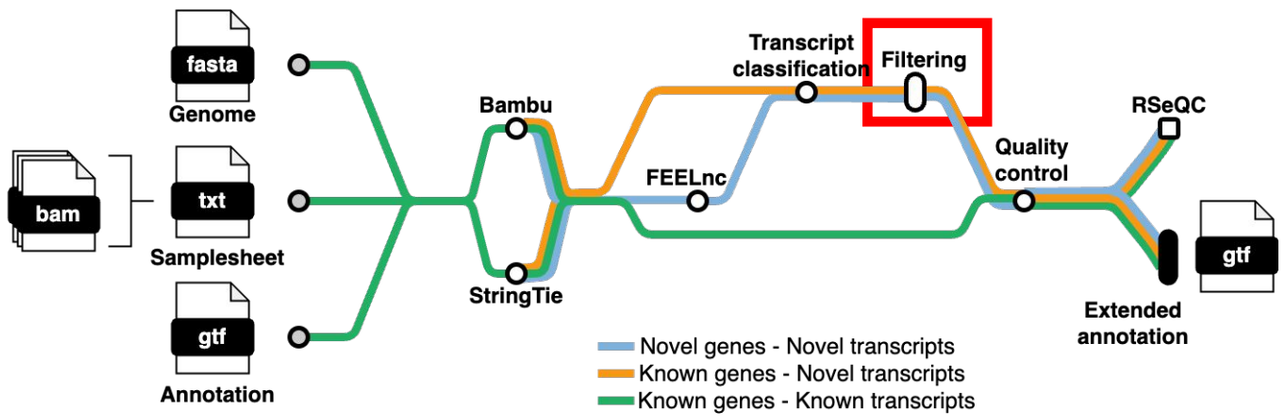
Novel isoform of known gene

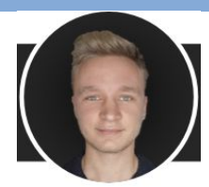
Extension of known gene

antisense lncRNA

Long intergenic ncRNA

ANNEXA : Full-lengthness of novel genes (TFK)

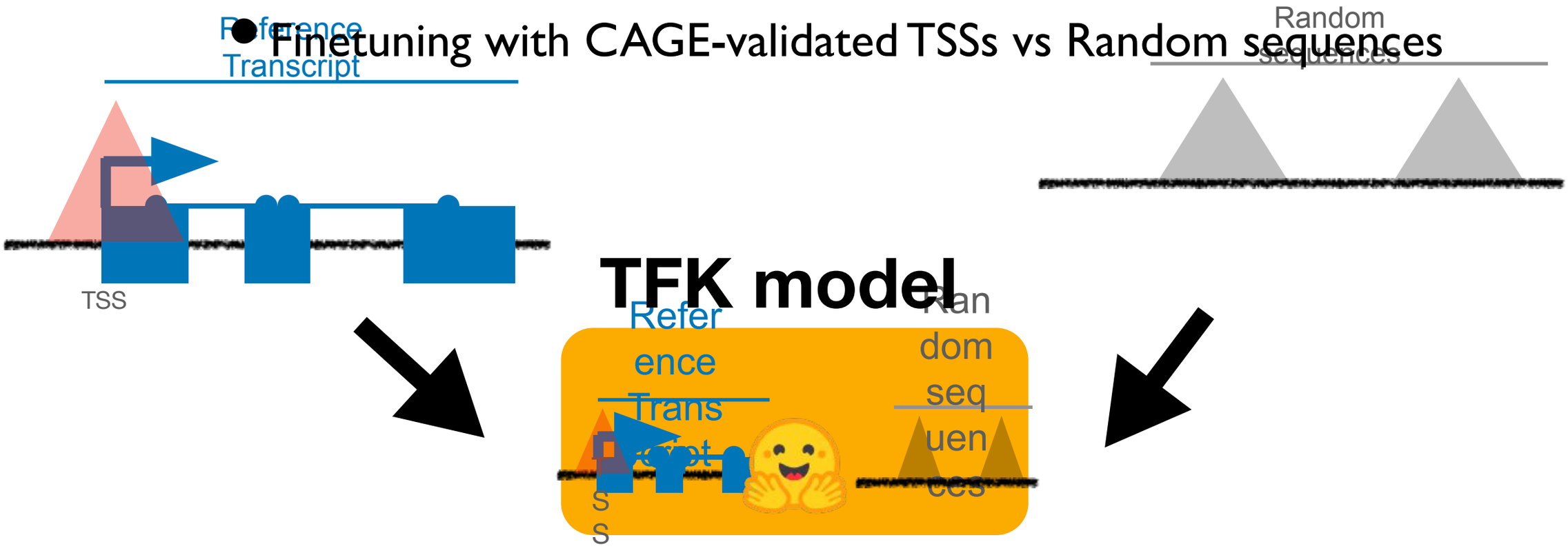




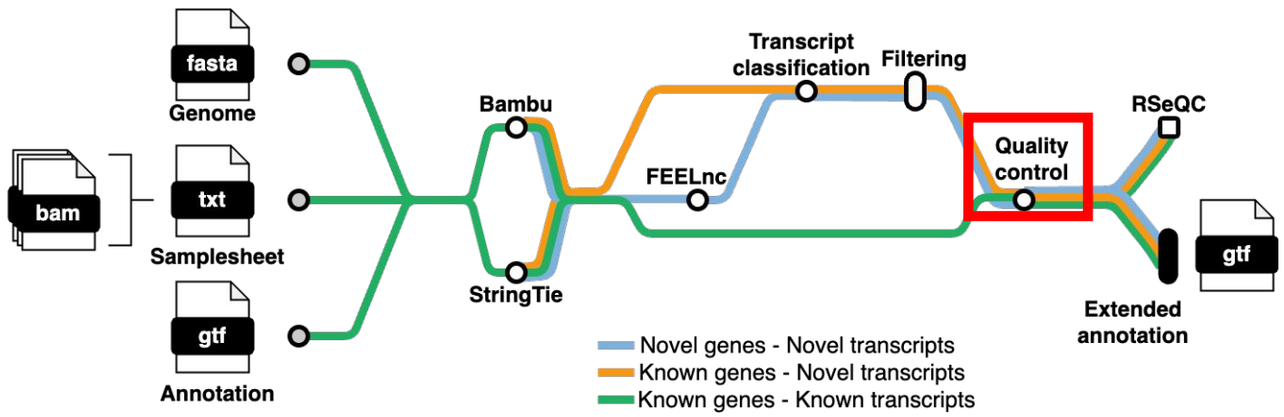
https://github.com/IGDRion/TransforKmers

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



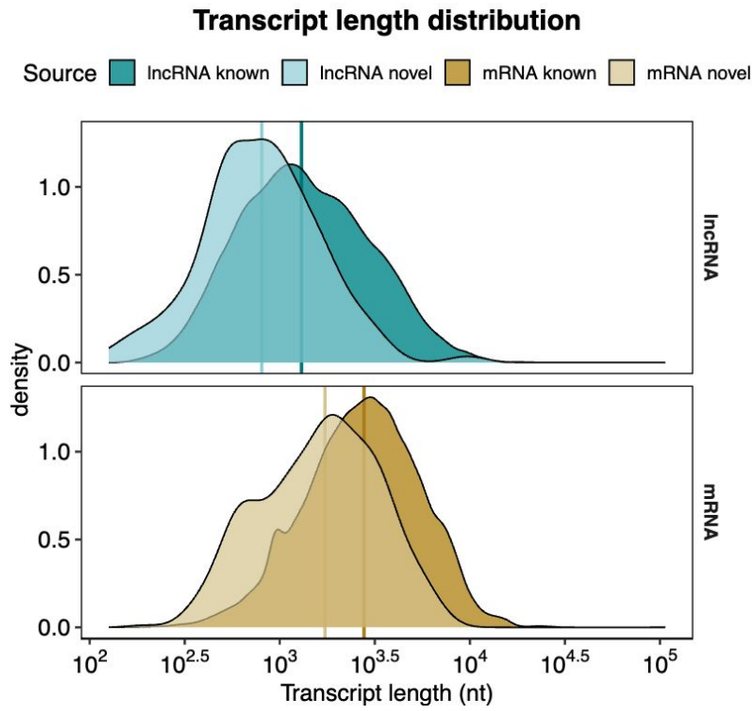
ANNEXA : Quality control

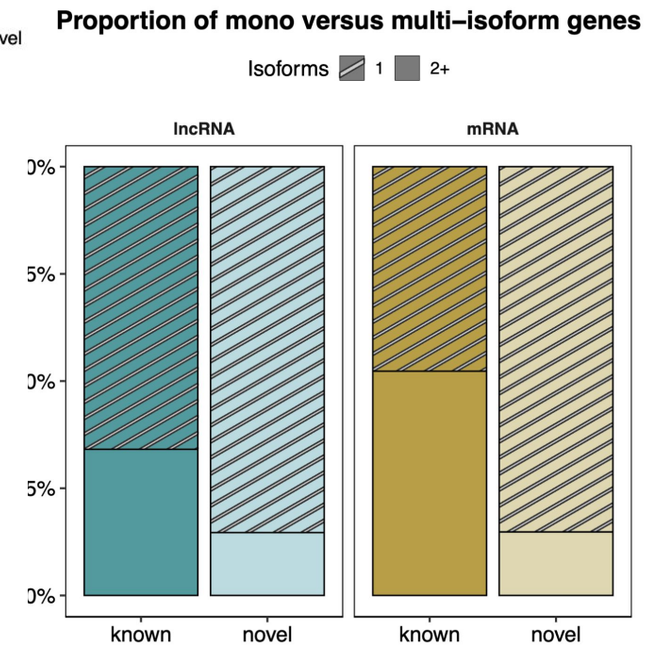


IGDRion/ANNEXA

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

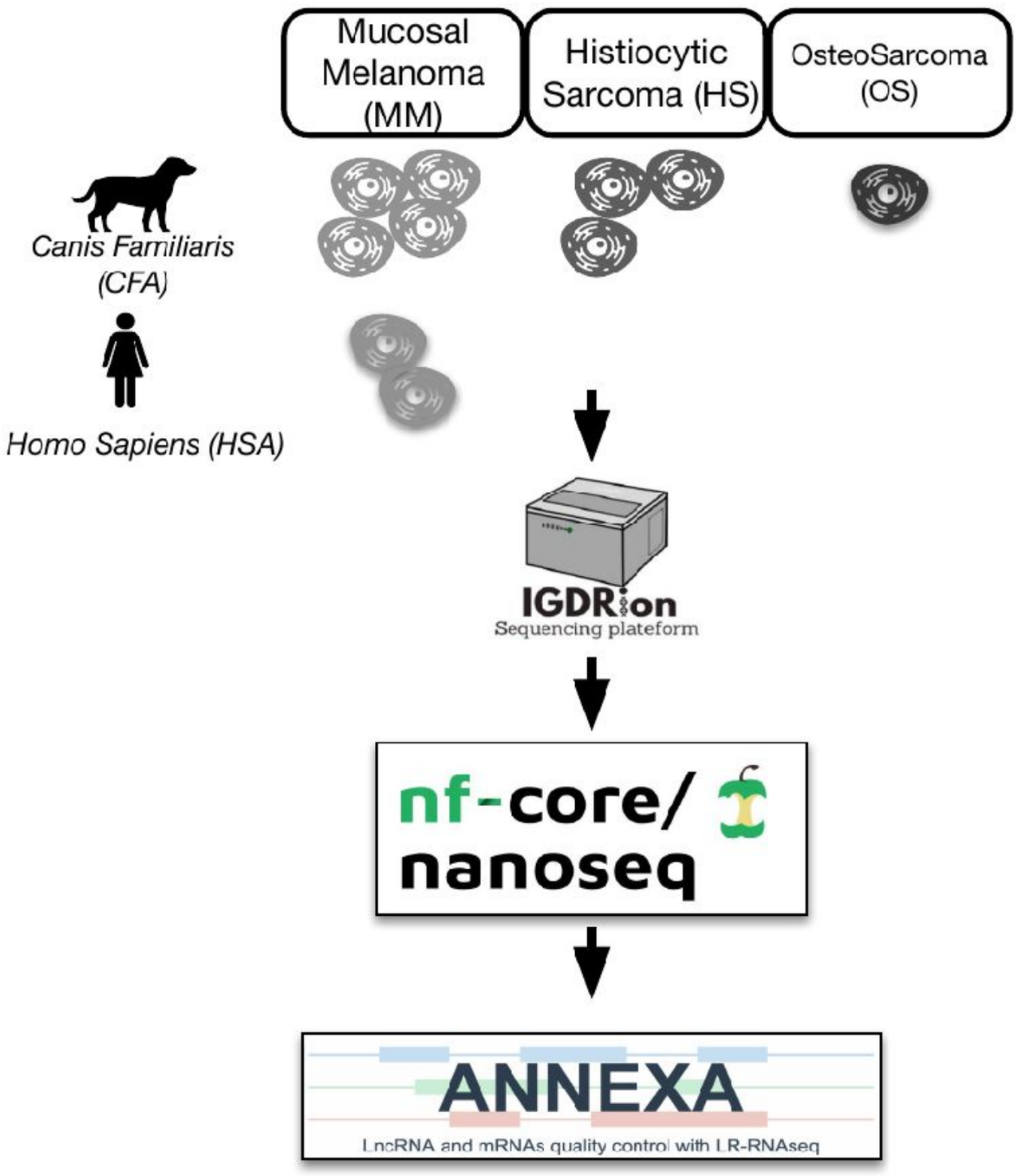
- Multiple features :
  - Number of mRNAs versus lncRNAs
  - Number of single-exon genes
  - Breadth of expression /Tissue specificity
  - Gene extensions





Allows to QC novel annotated genes/transcripts/exons

Applying ANNEXA to comparative oncology



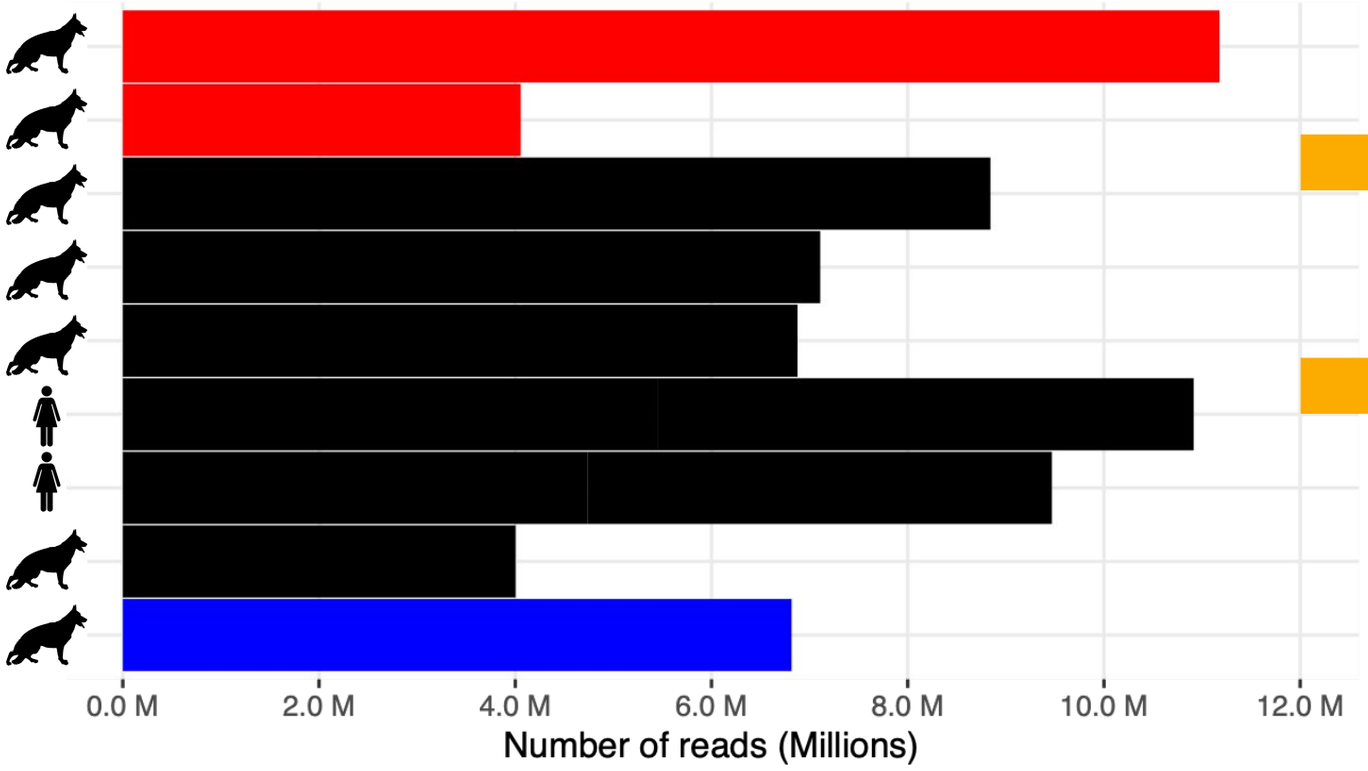
Poly(A) extraction and library preparation

Direct cDNA Nanopore Sequencing

Basecalling - read QC - mapping

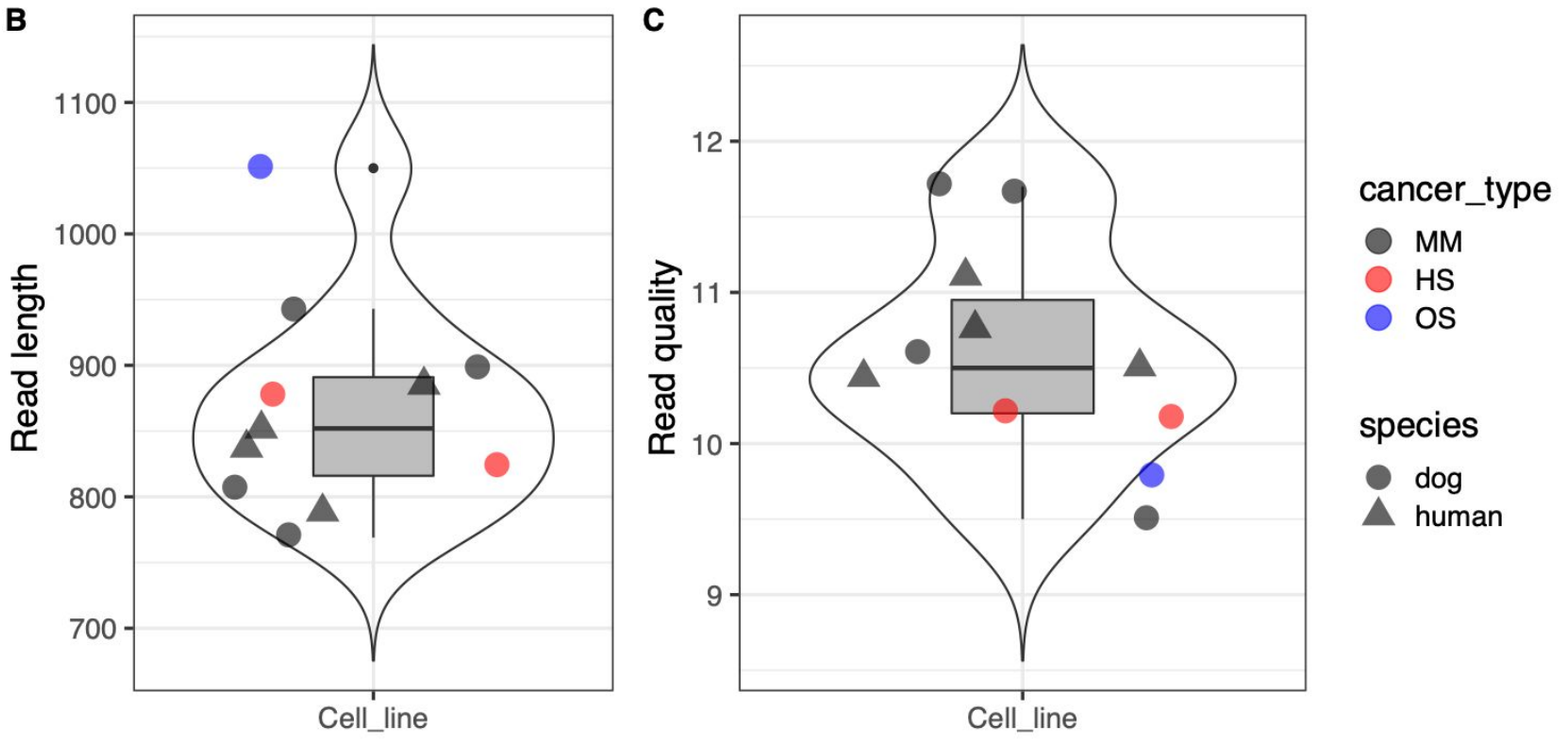
Extended annotation (new lncRNAs + mRNAs)

Applying ANNEXA : M&M



~6M reads

Variable number of reads per sample



length~900nt

Qscore ~10.5 (old chemistry)

Applying ANNEXA : M&M

Reference genome :

canFam4

Reference annotations (x3) :

- Ensembl (Dyer et al., 2025)

- RefSeq (O'Leary et al., 2016)

- UCSC (Wang et al., 2021)

Reference genome :

GRCh38

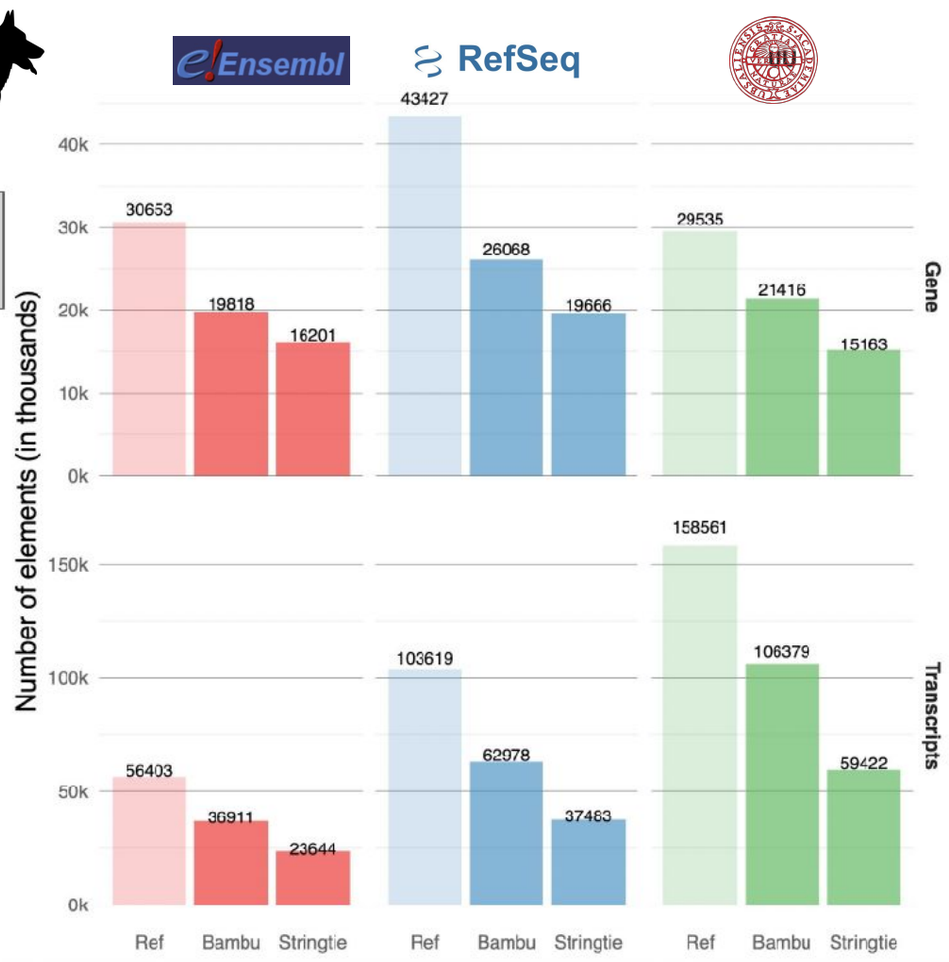
Reference annotations (x2):

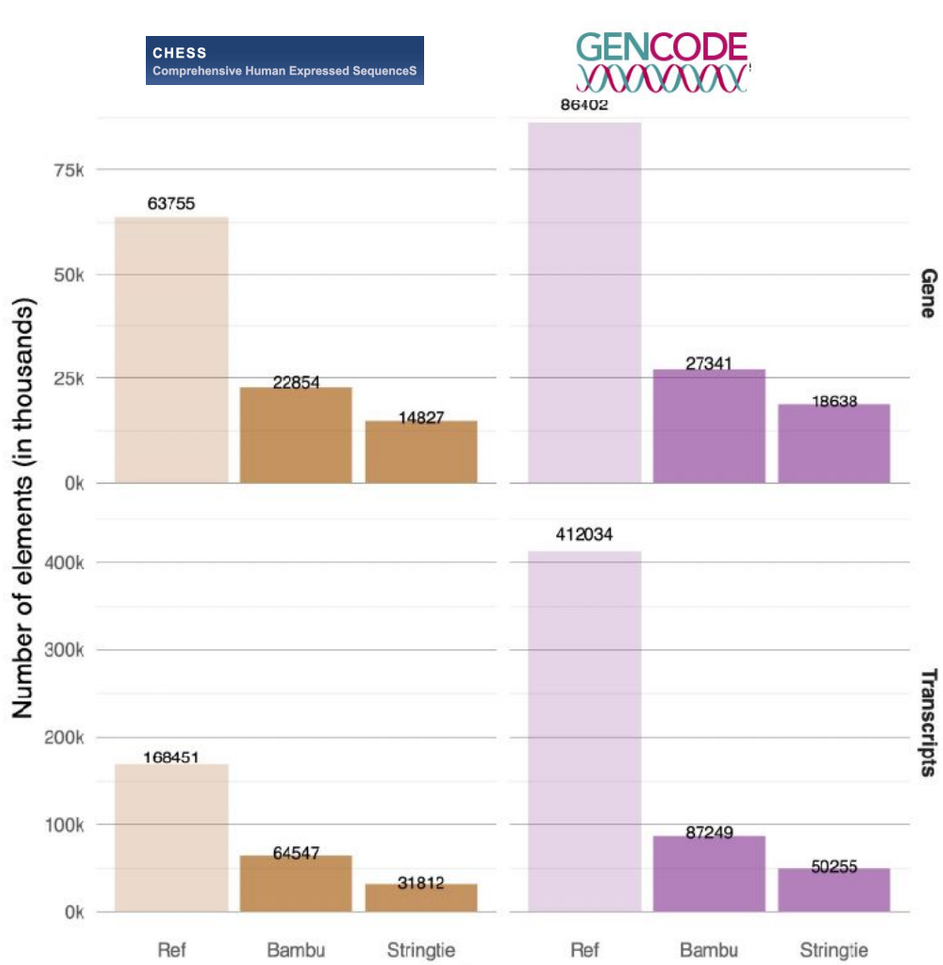
- Gencode (Mudge et al., 2025)

- CHESS (Varabyou et al., 2023)

Benchmarking Stringtie and Bambu for reconstruction – Known elements

KNOWN

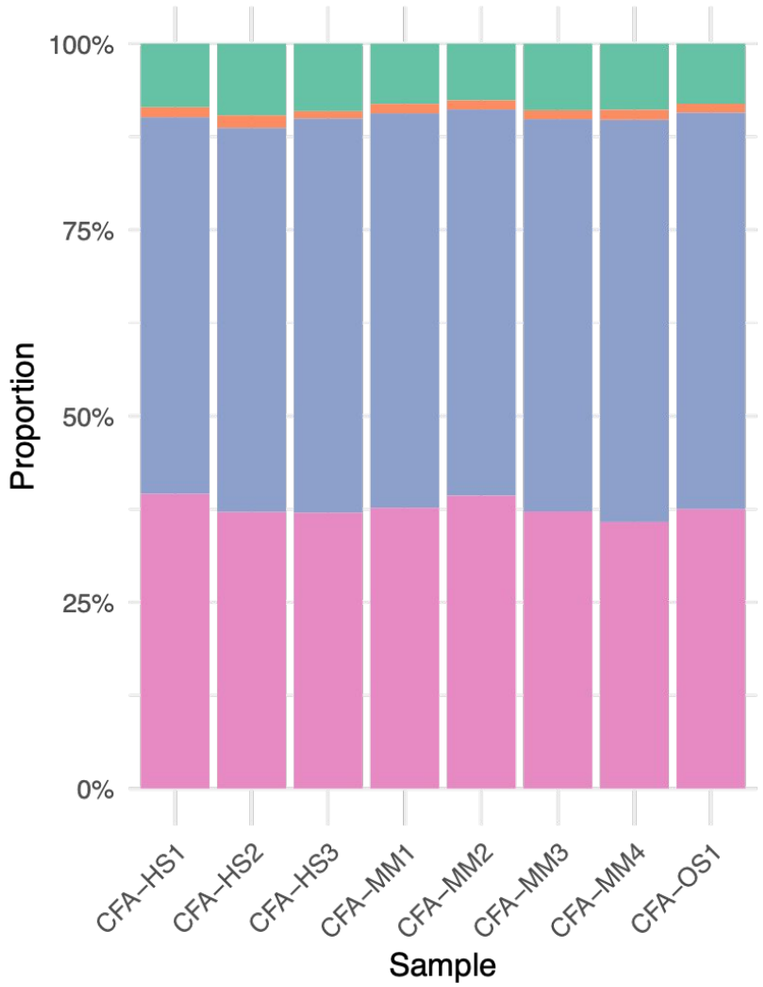




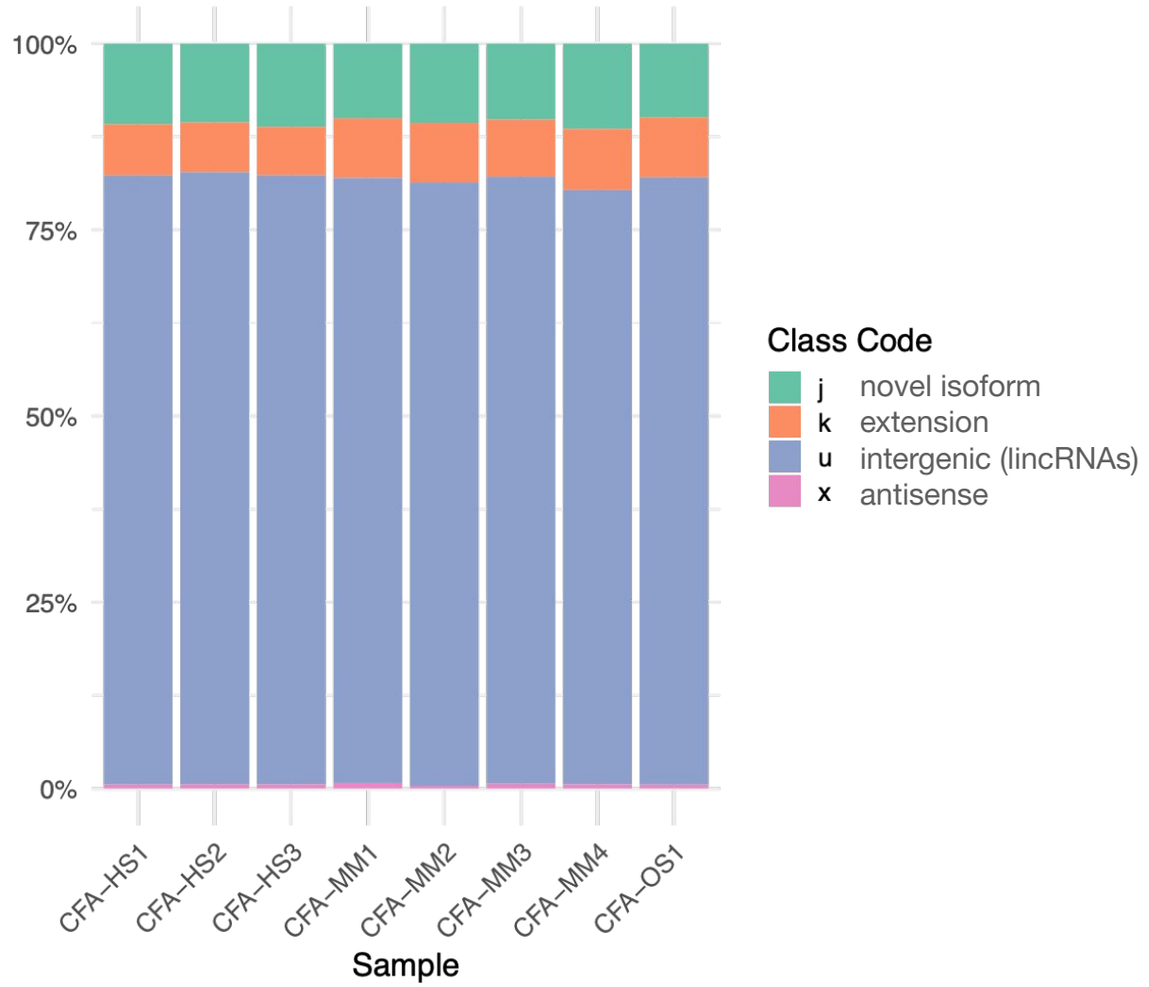
- Bambu reconstructs more **known** genes/transcripts that Stringtie consistently
- Consistent pattern across the canine annotations
- As well as human annotations (although with lower rates)

Comparing classification of novel lncRNAs

Bambu novel lncRNA



StringTie novel lncRNA



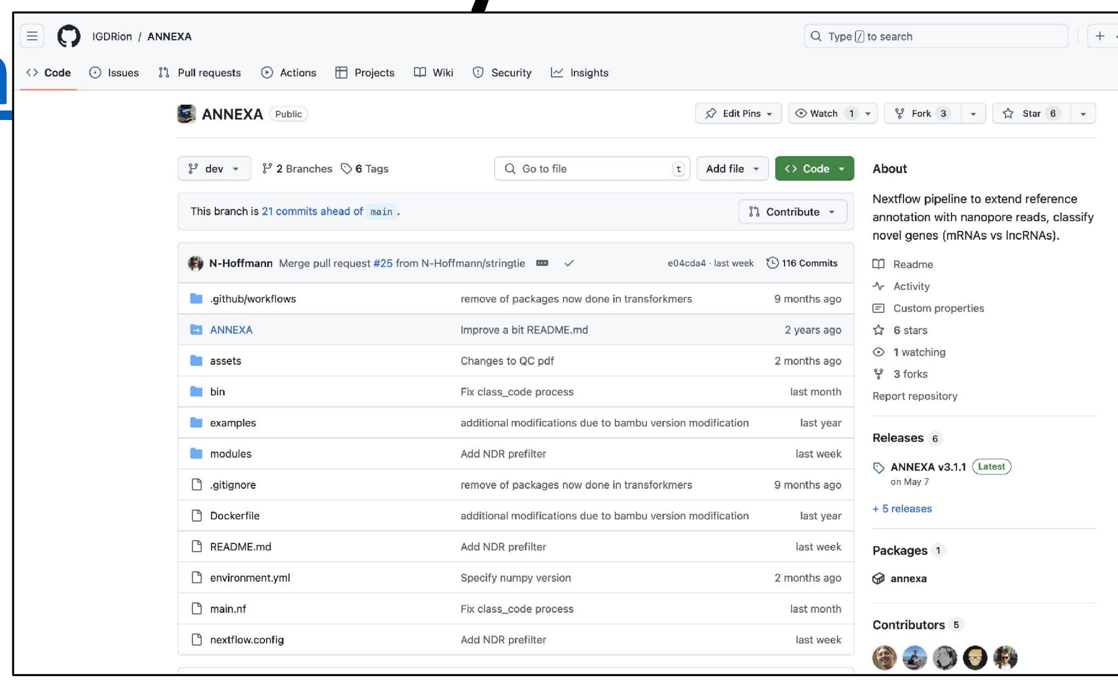
Different patterns in classification:

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

Conclusions

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

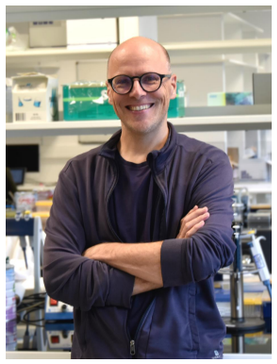
- ANNEXA is **FAIR and freely available** (https://gith)



- Applied ANNEXA on pilot comparative oncology long-read RNA seq, ➡ 100 dog and 30 human MM samples

Acknowledgements

IGDRion



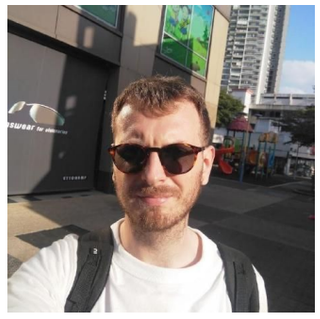
Edouard CADIEU

Aurore BESSON

Victor LE BARS

Enora CORLER

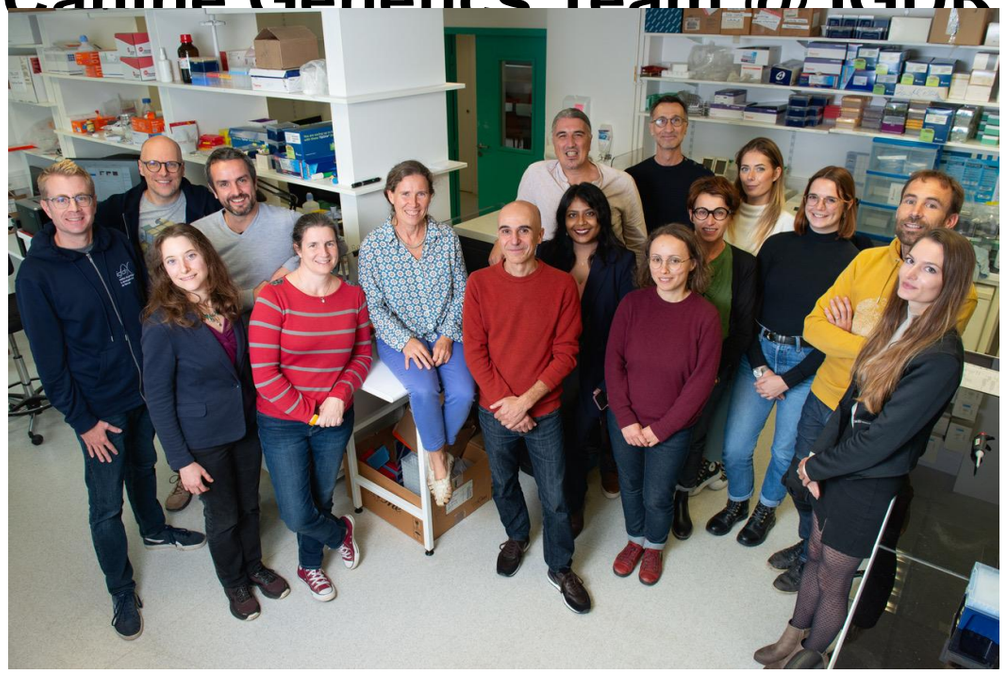
Matthias LORTHIOIS



Nicolaï HOFFMANN

Anastasia RUSAKOVICH

Canine Genetics Team @ IGDR



Catherine ANDRE

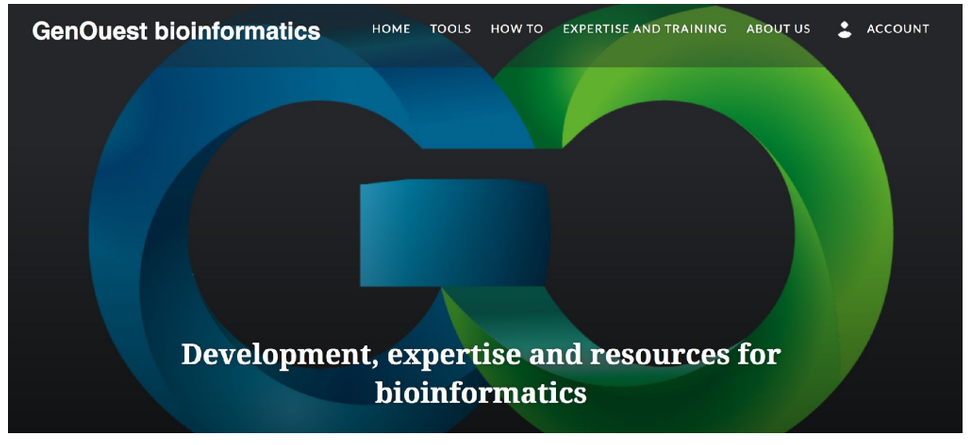
BINGO

>BIOLOGY IN SILICO for NON CODING GENOME and ONCOLOGY

Yuna BLUM

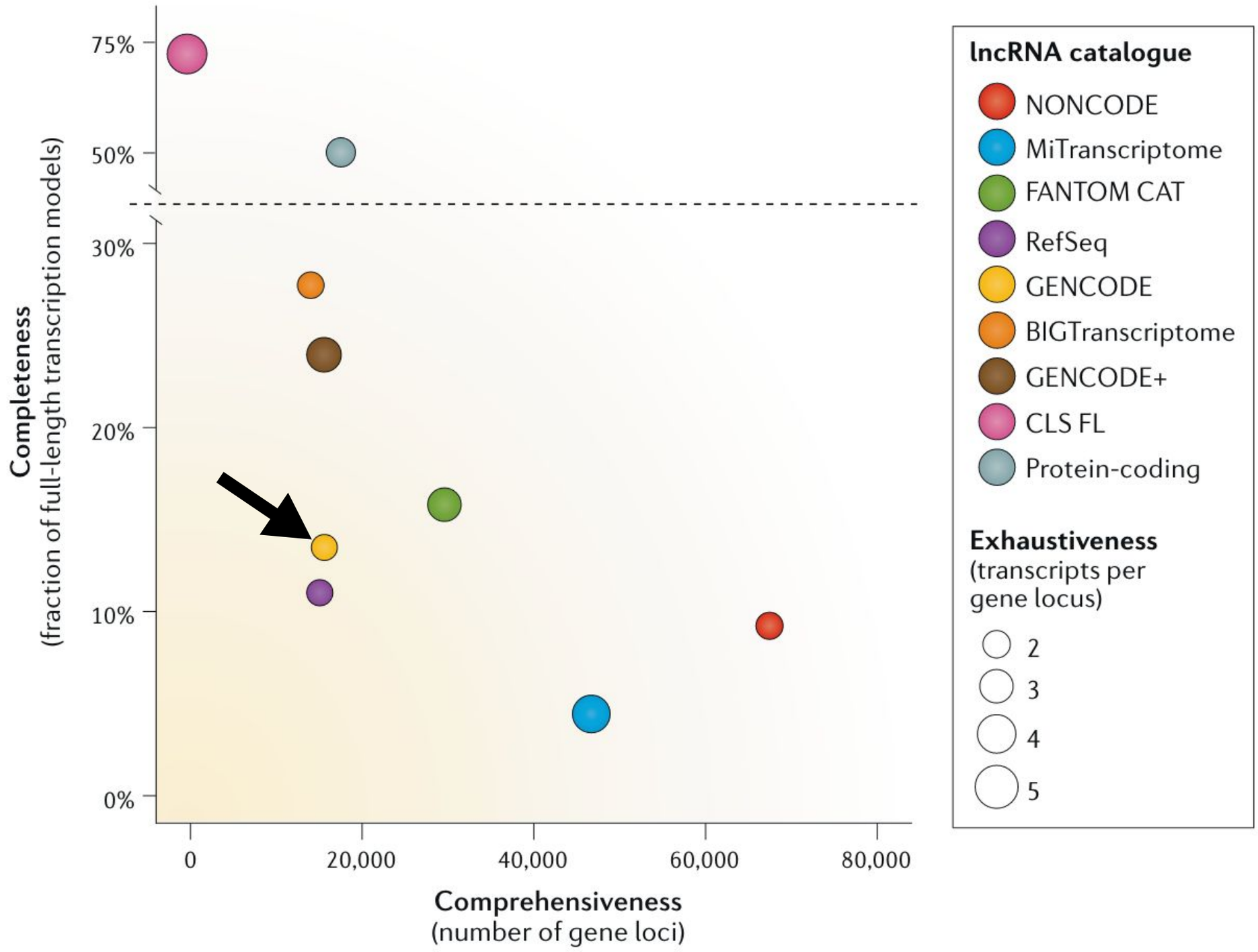
Christophe HELIGON

Genouest bioinformatic



lncRNA annotation resources

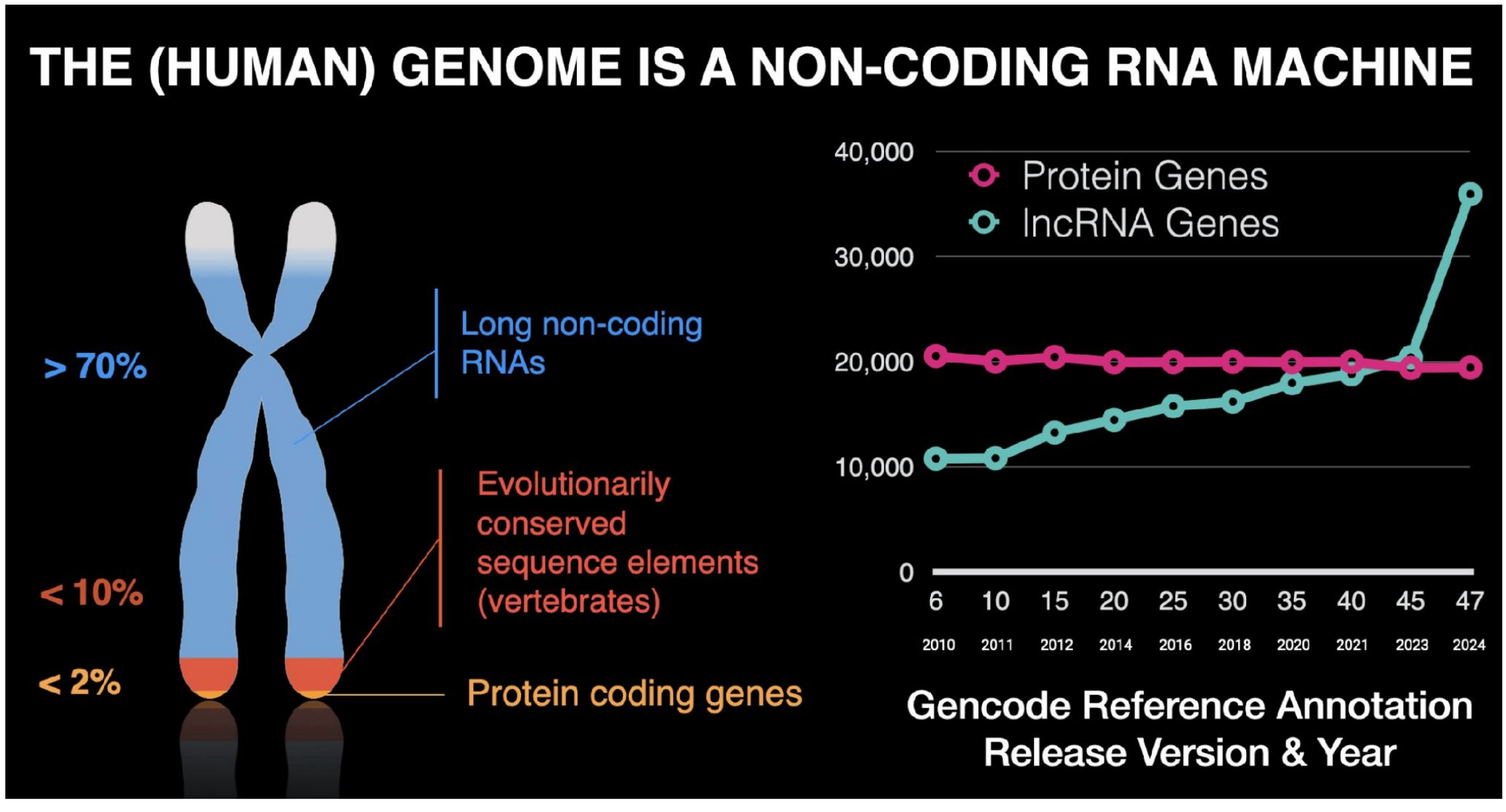
(Uszczynska-Ratajczak et al., 2018)



lncRNA resources vary in quality/quantity of the annotation

The non-coding genome

- **~70%** of the human genome is **transcribed into RNAs** with only **2% corresponding to proteins (mRNAs)**



➡ GOAL : Exploiting the strength of LR-RNAseq to correctly reconstruct novel genes/isoforms involved in therapy resistance

Repartition of reads in glioblastoma resistant samples