Structural annotations of eukaryotic genomes: Can do better !

Sophie Lemoine

MERIT GT-Annotation 25/03/17

Our facility background

- **Sequencing projects** that aim at **quantifying gene and transcript expression**
- **Interest in nanopore** since its availability (2017)

  ➡**Quantification of transcripts without the need of a model** to discriminate exon belonging

- Bulk RNASeq
  - Illumina and Nanopore data
- SingleCell 10X RNASeq
  - Nanopore and Illumina data

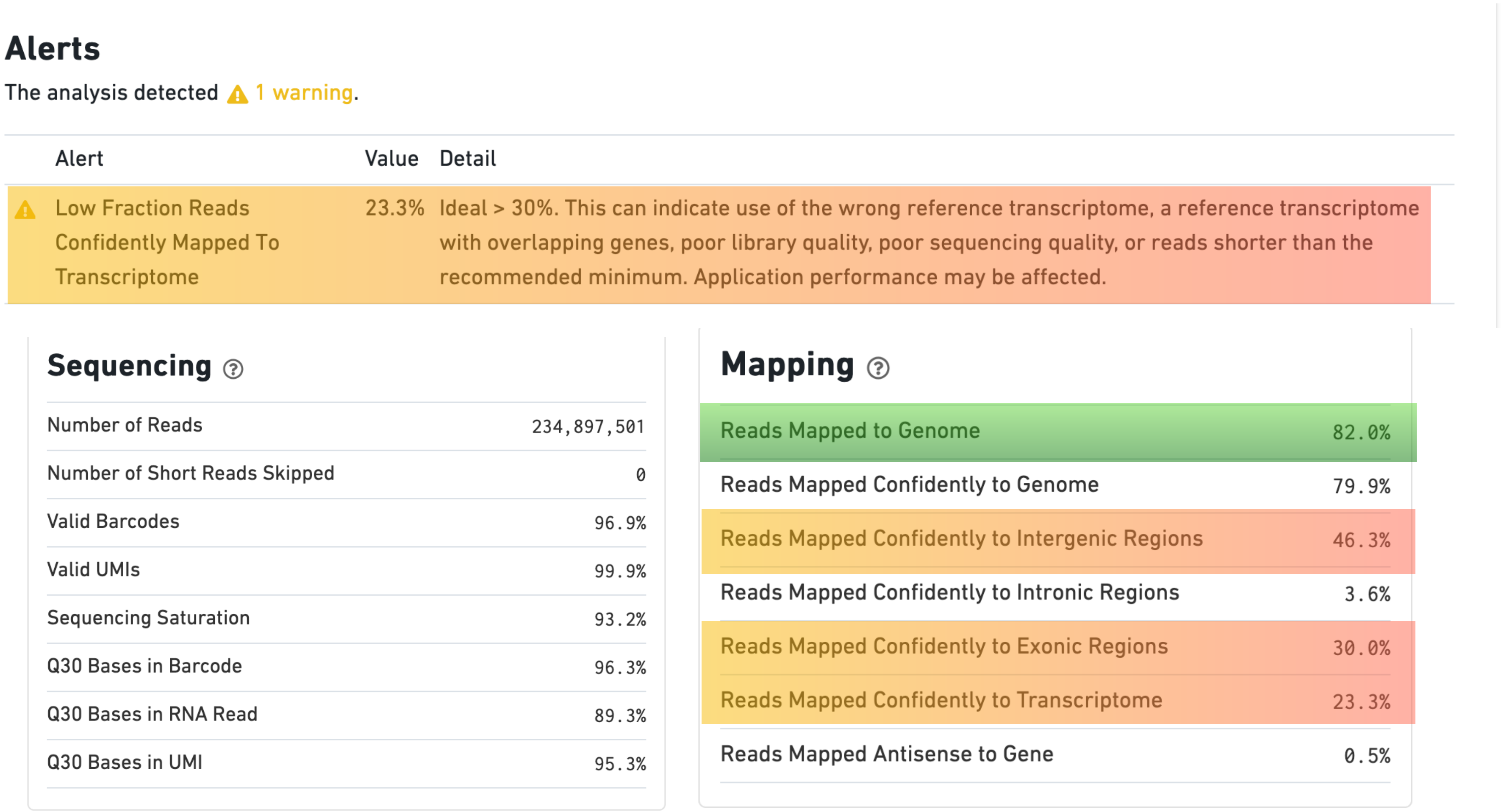
➡ **No particular need to perform structural annotation except since 2021…**

The turning point of 2021

SingleCell 10X Illumina project on an invertebrate

➡10X data : reads expected to map in the UTRs (here in 3')

Genome and structural annotation available and already used in a bulk RNASeq project performed in 2019



Alerts

The analysis detected ⚠ 1 warning.

	Alert	Value	Detail
⚠	Low Fraction Reads Confidently Mapped To Transcriptome	23.3%	Ideal > 30%. This can indicate use of the wrong reference transcriptome, a reference transcriptome with overlapping genes, poor library quality, poor sequencing quality, or reads shorter than the recommended minimum. Application performance may be affected.

Sequencing ⓘ

Number of Reads	234,897,501
Number of Short Reads Skipped	0
Valid Barcodes	96.9%
Valid UMIs	99.9%
Sequencing Saturation	93.2%
Q30 Bases in Barcode	96.3%
Q30 Bases in RNA Read	89.3%
Q30 Bases in UMI	95.3%

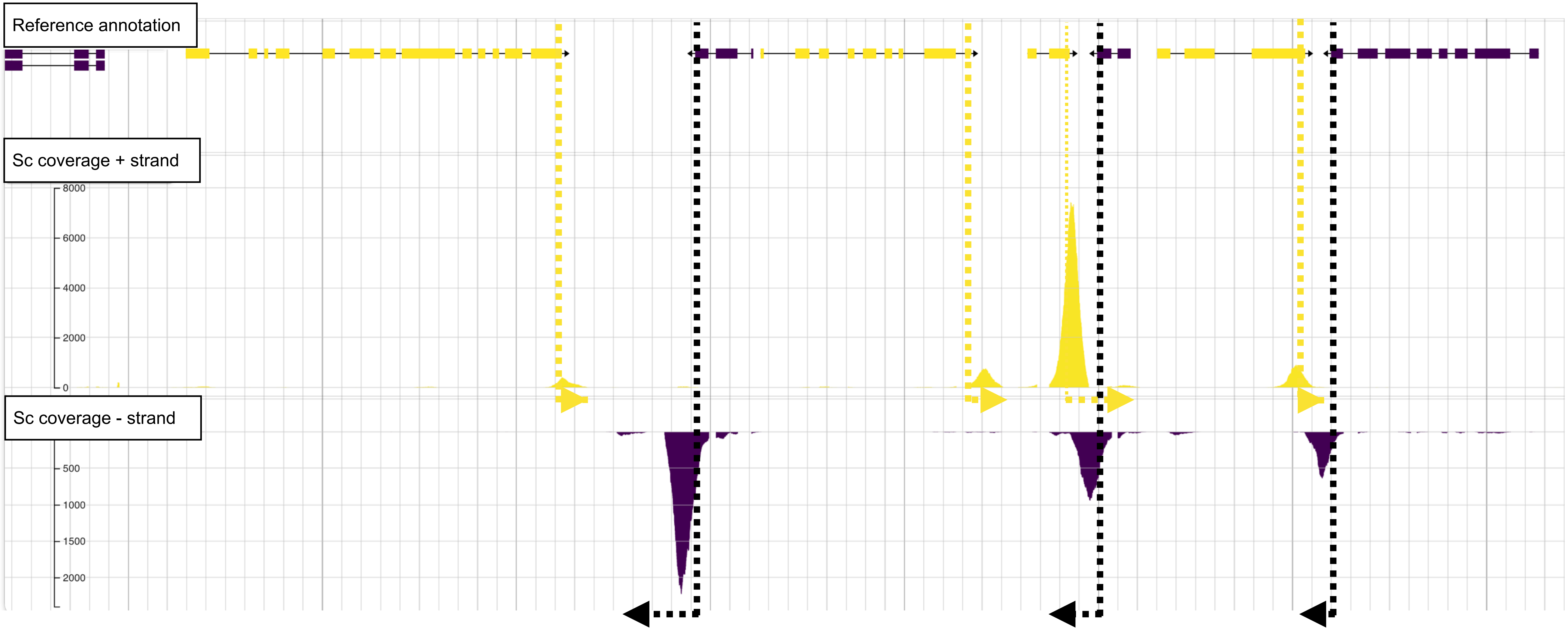
Mapping ⓘ

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	79.9%
Reads Mapped Confidently to Intergenic Regions	46.3%
Reads Mapped Confidently to Intronic Regions	3.6%
Reads Mapped Confidently to Exonic Regions	30.0%
Reads Mapped Confidently to Transcriptome	23.3%
Reads Mapped Antisense to Gene	0.5%

The CellRanger QC report indicates that **only 30% of the reads are mapped on exons and 46% are mapped between genes**

➡ **Why ?**

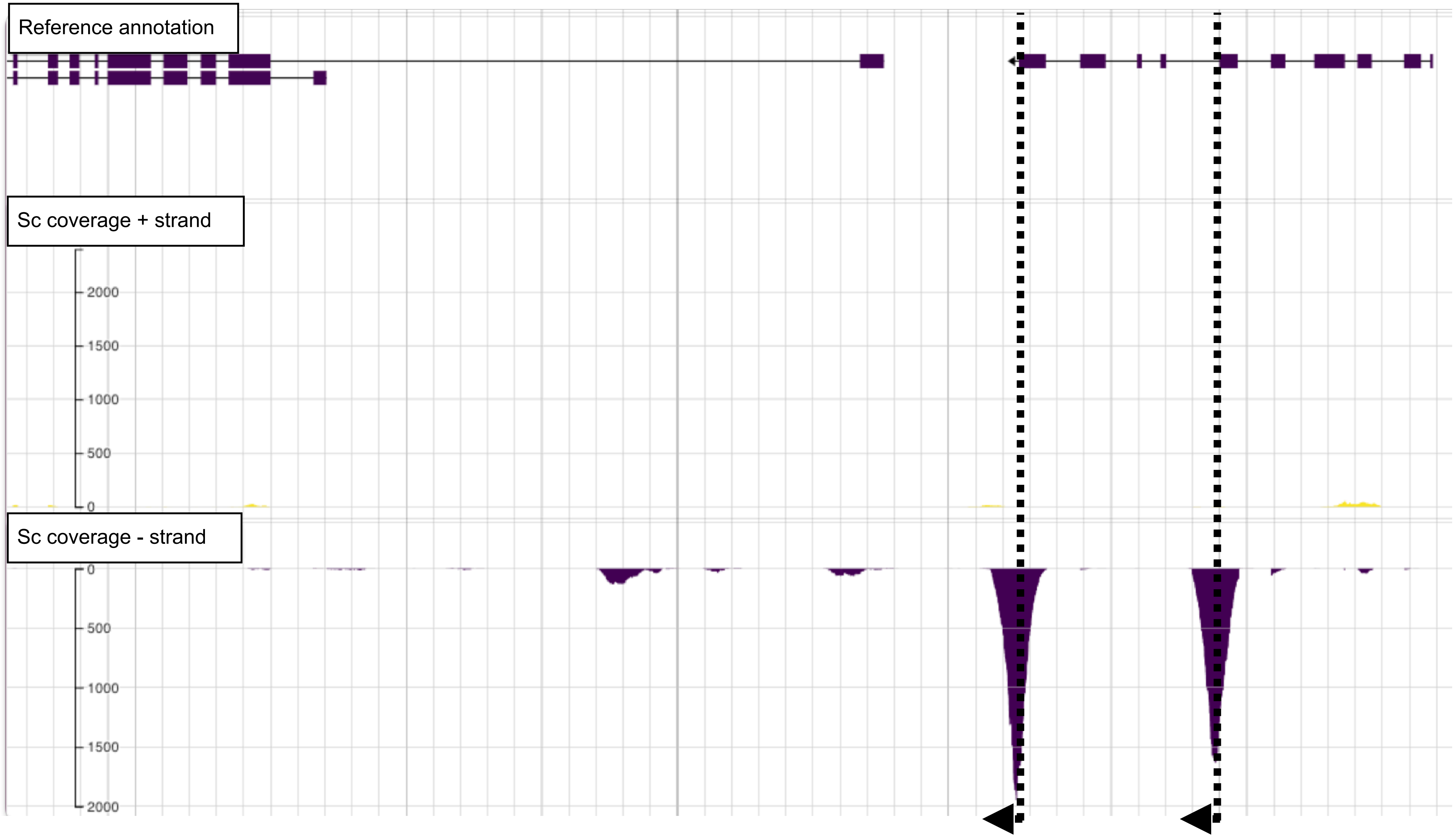
When in trouble, check your data in a genome viewer



➡ **Most of the 10X SC data has no intercept with the reference annotation**

➡ **SC data cannot be counted**

And again….



➡ **Probably 2 genes or 2 isoforms expressed here and one is not annotated**

Structural annotation has to be improved

- **Annotations rely a lot on proteins**

  GALBA

  GeneMark-ETP

  Augustus

  BRAKER

  ➡ **UTRs are often badly annotated**

- **It’s not a surprised to have missing genes on non model organisms**

StringTie2

A lots of pipelines use **StringTie2**

It is quite flexible and adapts to all types of data

➡ Uses short reads, long reads with annotation or without

It's easy and convenient

➡ Let's test StringTie2 to improve the structural annotation and provide accurate counts on the SC project

StringTie2 inputs available for this project:

➡ Illumina bulk RNASeq project sequenced in 2019 (**SR**)

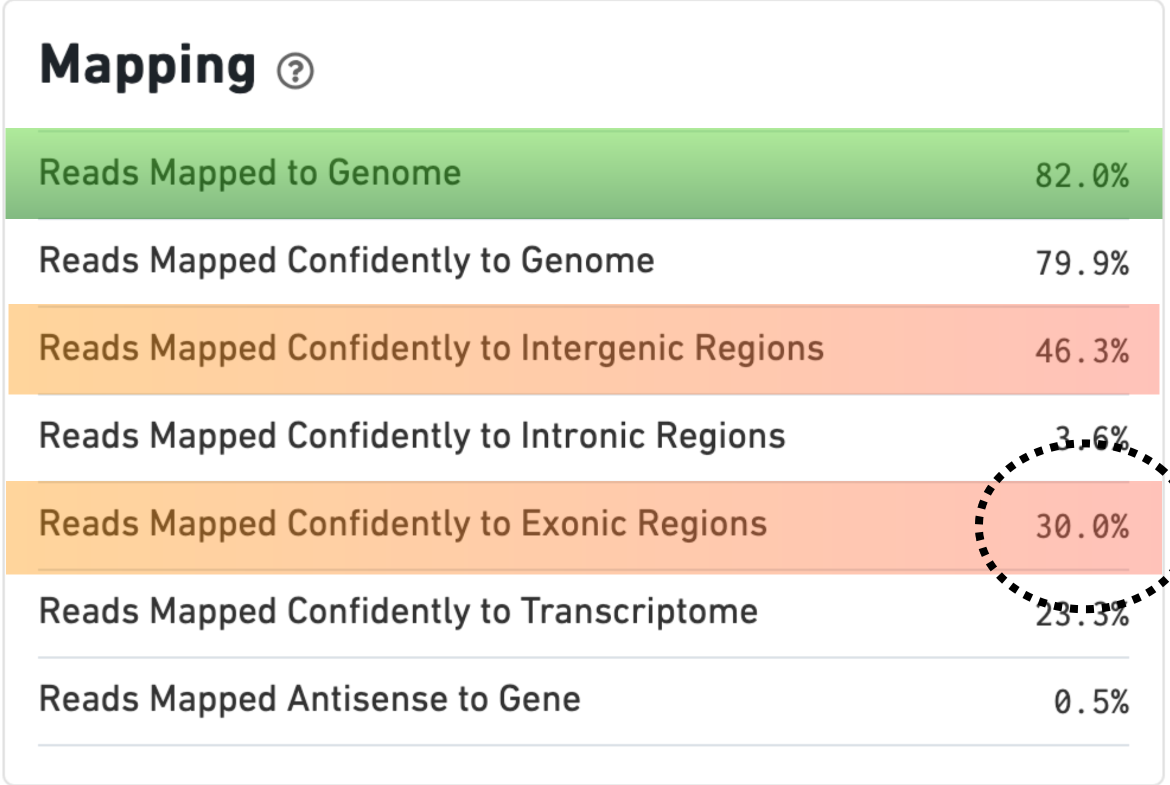
➡ Reference annotation to be improved (**Annot**)

➡ New Nanopore bulk RNASeq data being performed (**LR**)

PLOS Computational Biology 18, 6 (2022)

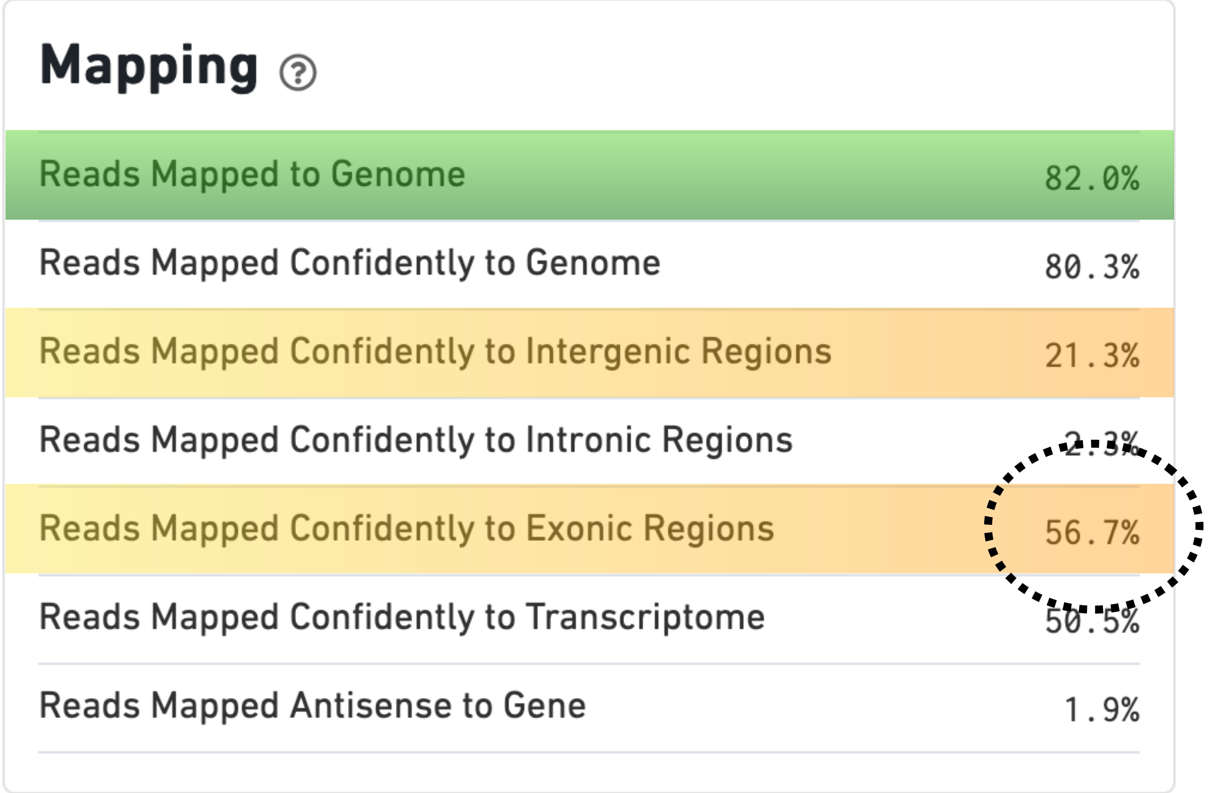
SC counts with StringTie2 - SR + Annot

SC on Reference annotation → SC on StringTie2 annotation (SR+annot)



Mapping ?

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	79.9%
Reads Mapped Confidently to Intergenic Regions	46.3%
Reads Mapped Confidently to Intronic Regions	3.6%
Reads Mapped Confidently to Exonic Regions	30.0%
Reads Mapped Confidently to Transcriptome	23.3%
Reads Mapped Antisense to Gene	0.5%



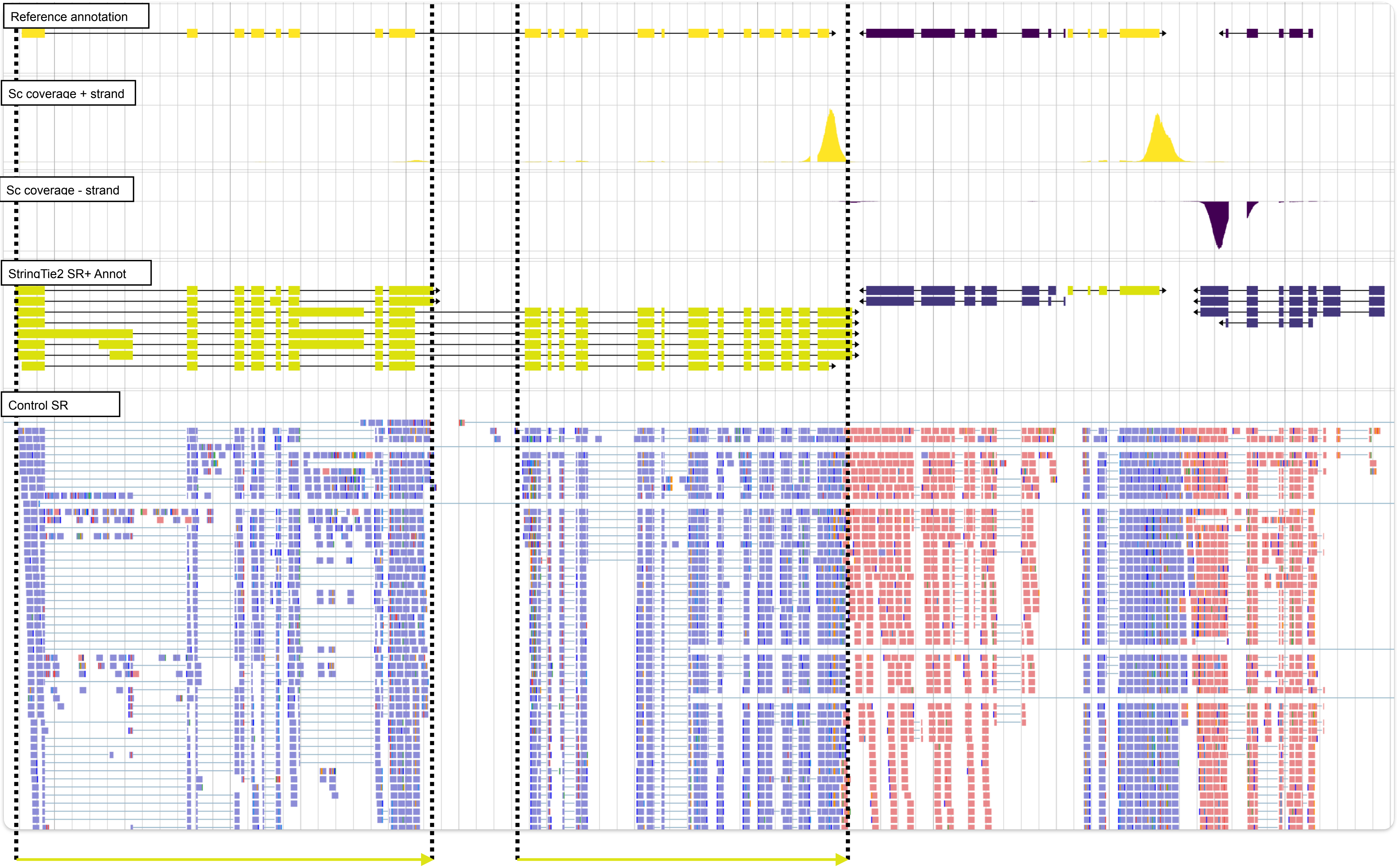
Mapping ?

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	80.3%
Reads Mapped Confidently to Intergenic Regions	21.3%
Reads Mapped Confidently to Intronic Regions	2.3%
Reads Mapped Confidently to Exonic Regions	56.7%
Reads Mapped Confidently to Transcriptome	50.5%
Reads Mapped Antisense to Gene	1.9%

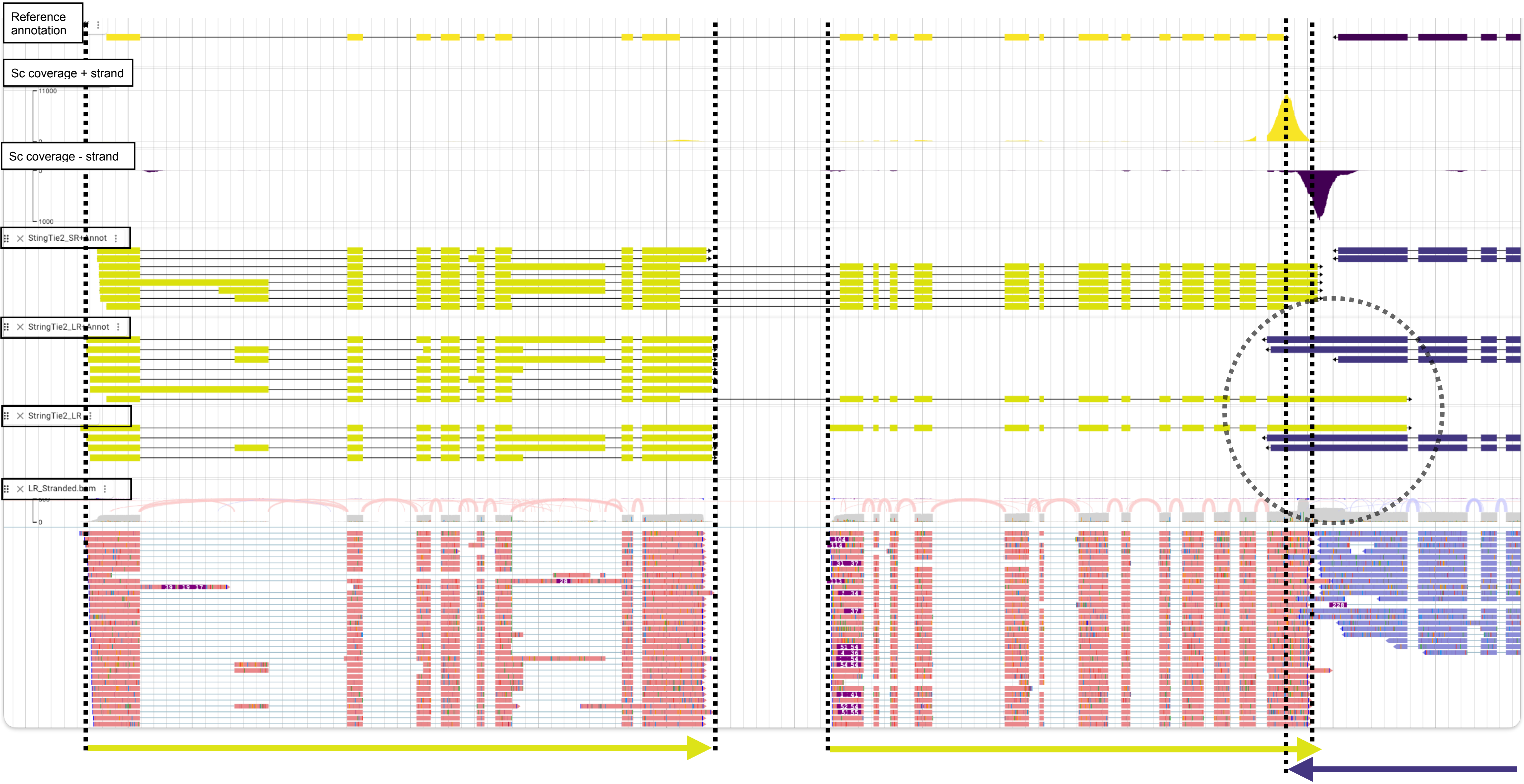
➡ **The number of reads considered in the analysis is increasing**

But how do they look like mapped on the genome ?

StringTie2 - SR + Annot



StringTie2 - SR, LR + Annot (1)

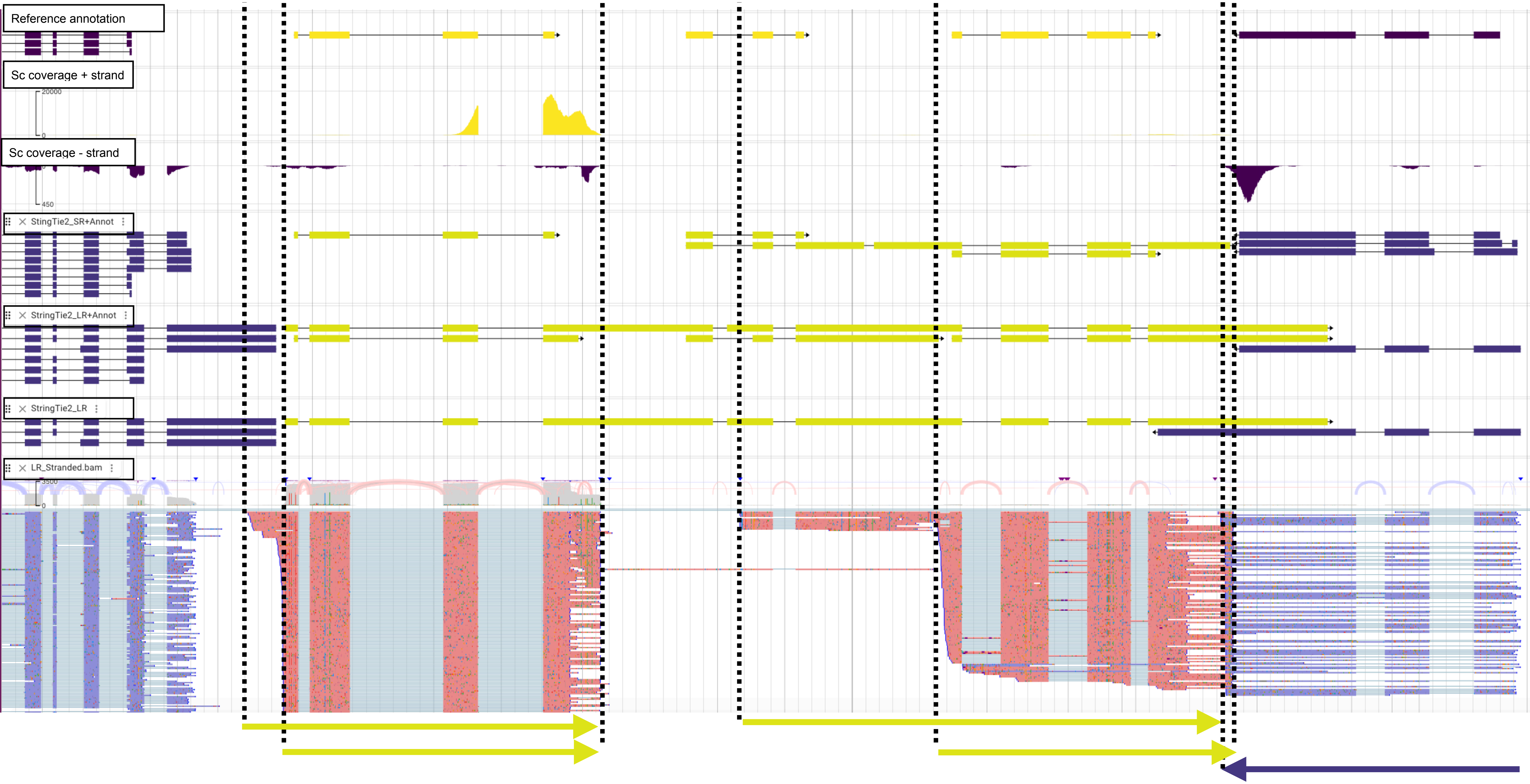


1- If the annotation is provided, it weighs on the model to the detriment of the data

➡ Fusion between 2 genes in the annotation remains despite evidence that they are 2 separate genes

2- Extension of UTRs without an evident link to the input data

StringTie2 - SR, LR + Annot (2)



1- If StringTie2 is provided with reference annotations, it systematically includes them and may also propose an alternative model.

2- Otherwise, it does anything without considering the input data

➡ **StringTie2 -whatever the inputs are- cannot be a good structural annotation tool**

StringTie2 is bundled in many annotation pipelines

- **Funnanotate**
- **PASA**
- **BRAKER3**
- **nf-core/nanoseq**

RNA-Seq data are then supplemented with protein data from closely related species and other evidences

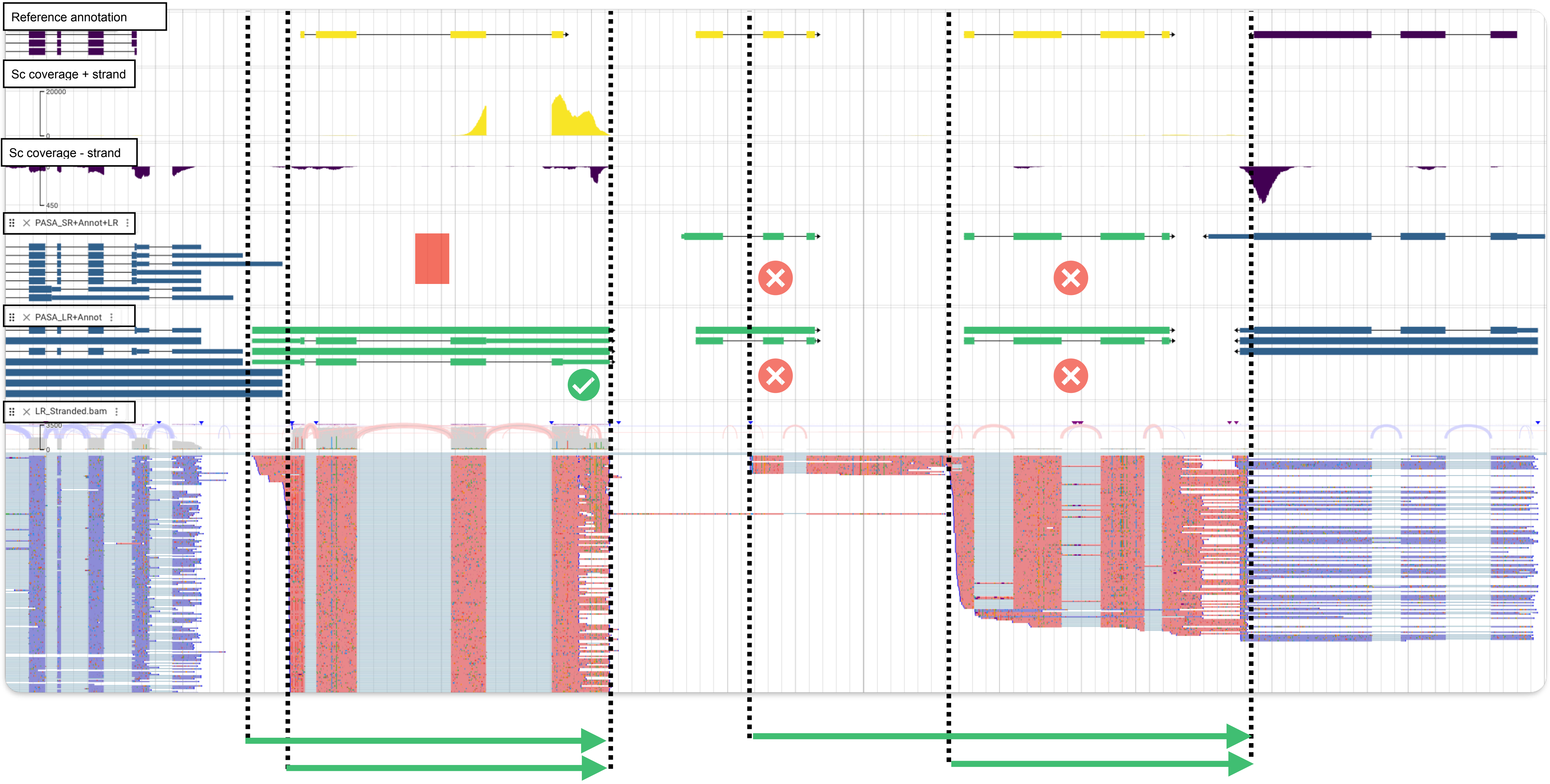
➡ **StringTie2 is then a step among others**

Example of PASA

- Ab initio gene finding: **GeneMarkHMM**, **FGENESH**, **Augustus**, and **SNAP**, **GlimmerHMM**
- Protein homology detection: **GeneWise**
- Alignment of known ESTs, full-length cDNAs, RNA-Seq assemblies **GSNAP**, **StringTie** and **Trinity**
- **EVidenceModeler (EVM)** to compute weighted consensus gene structure annotations

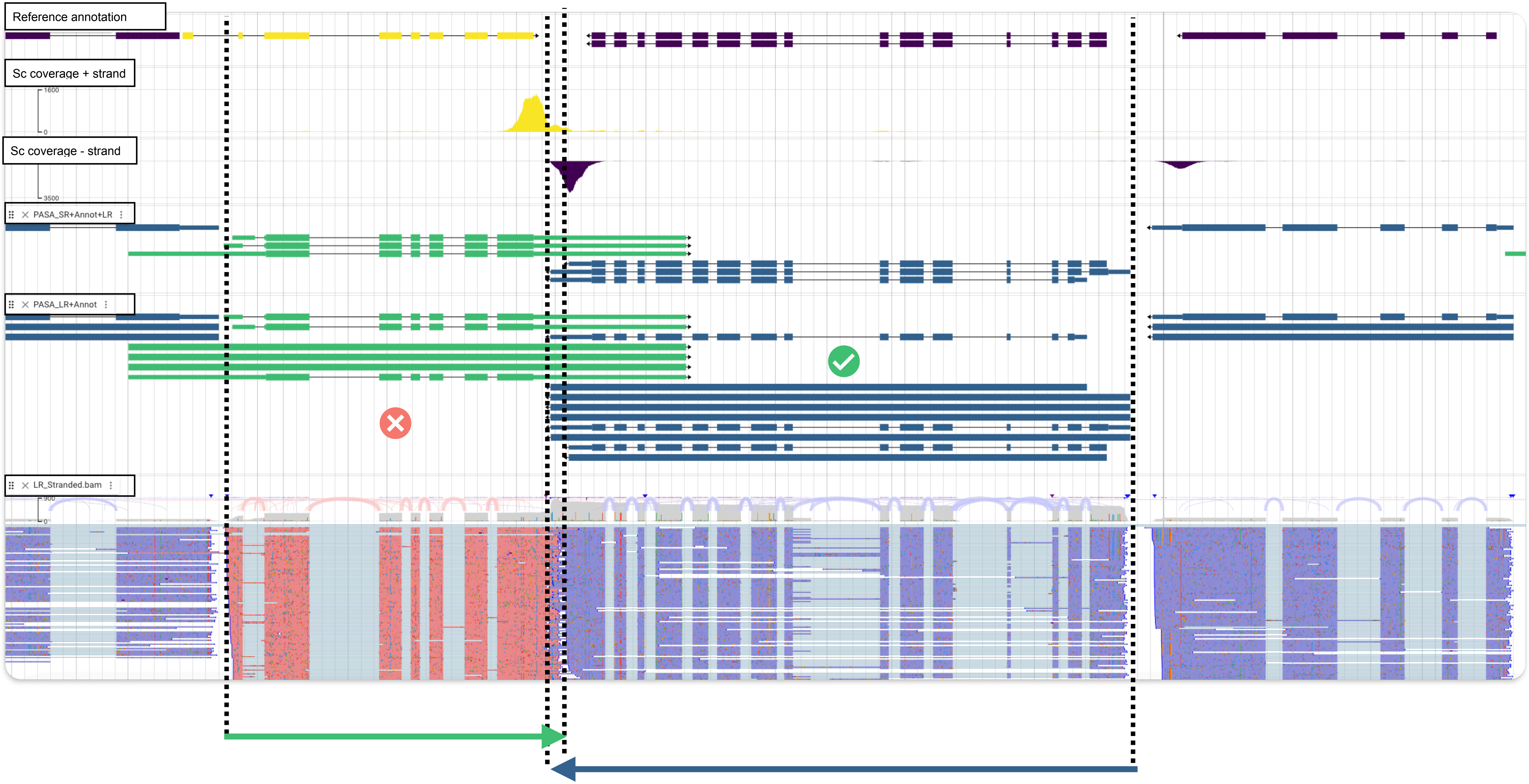
Nucleic Acids Res 2003 Oct 1;31(19):5654-66

PASA - SR+LR+Annot and LR+Annot (1)



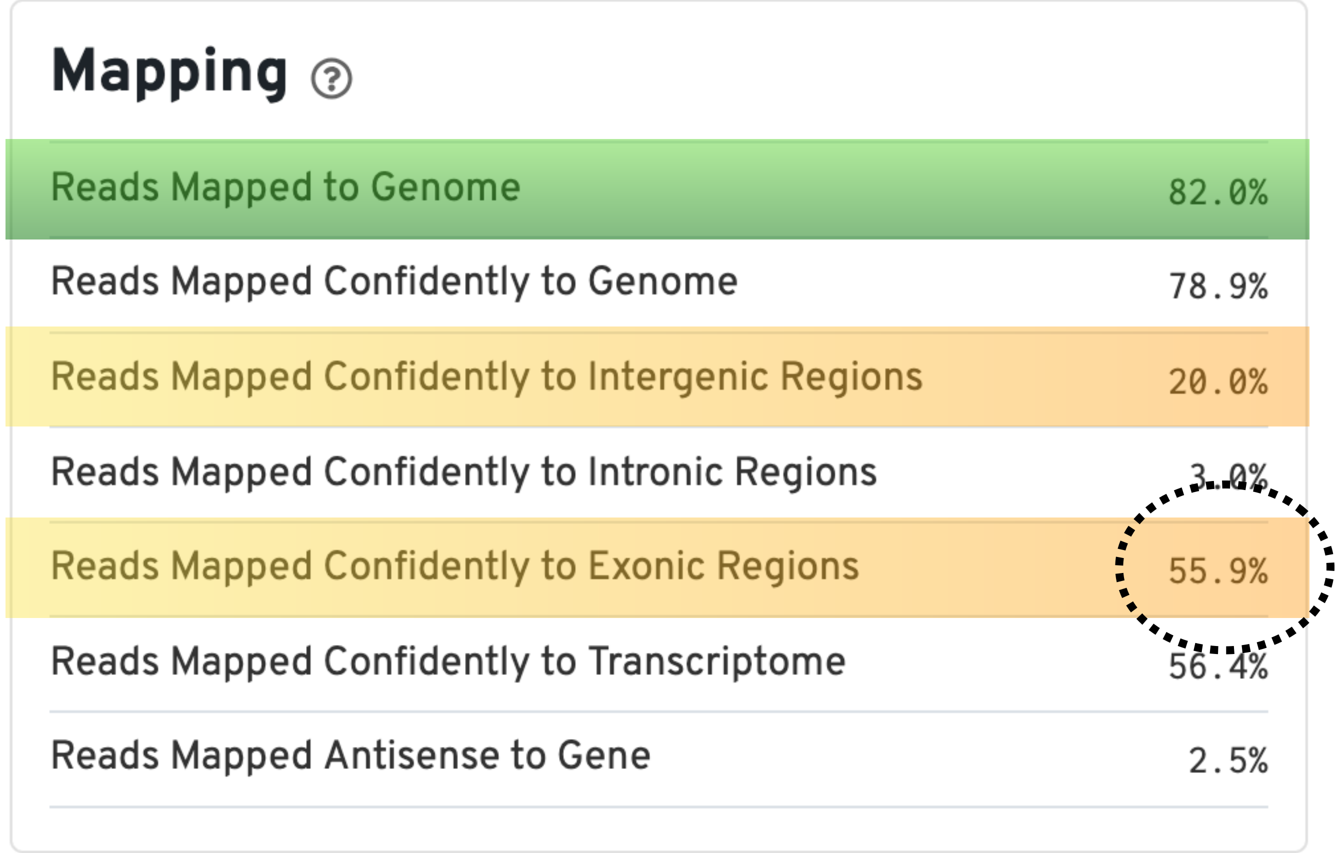
- Genes can missed, despite the large amount of data
- The reference annotation weighs more than the RNA-Seq data in the final model

PASA - SR+LR+Annot and LR+Annot (2)



PASA is very computationally intensive, complex to install and run, and yields only mediocre results despite the variety of input data

SC counts with PASA LR+Annot



Mapping

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	78.9%
Reads Mapped Confidently to Intergenic Regions	20.0%
Reads Mapped Confidently to Intronic Regions	3.0%
Reads Mapped Confidently to Exonic Regions	55.9%
Reads Mapped Confidently to Transcriptome	56.4%
Reads Mapped Antisense to Gene	2.5%

➡ **Like for StringTie2 (native), counts improve but they are not reliable**

If all sorts of StringTie2 are banned, what can we use ?

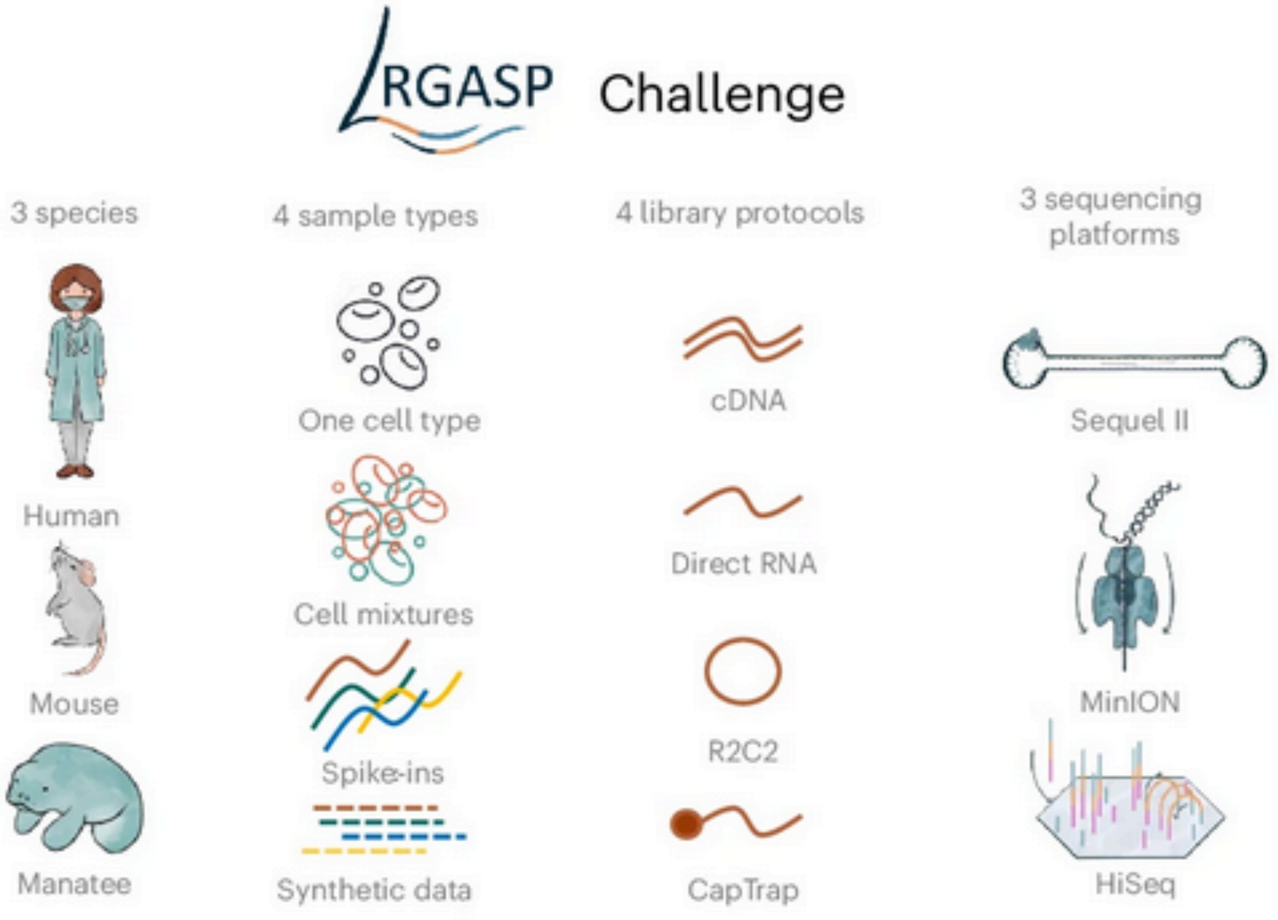
The LRGASP Encode Challenge (began in September 2020)

Evaluation of tools combining a large diversity of species, protocols and sequencing methods

« Characterizing long-read approaches to identify and quantify the transcriptomes of both model and non-model organisms »

3 topics:

1. **Transcript isoform detection with a high-quality genome**
2. **Transcript isoform quantification**
3. **Novo transcript isoform identification**

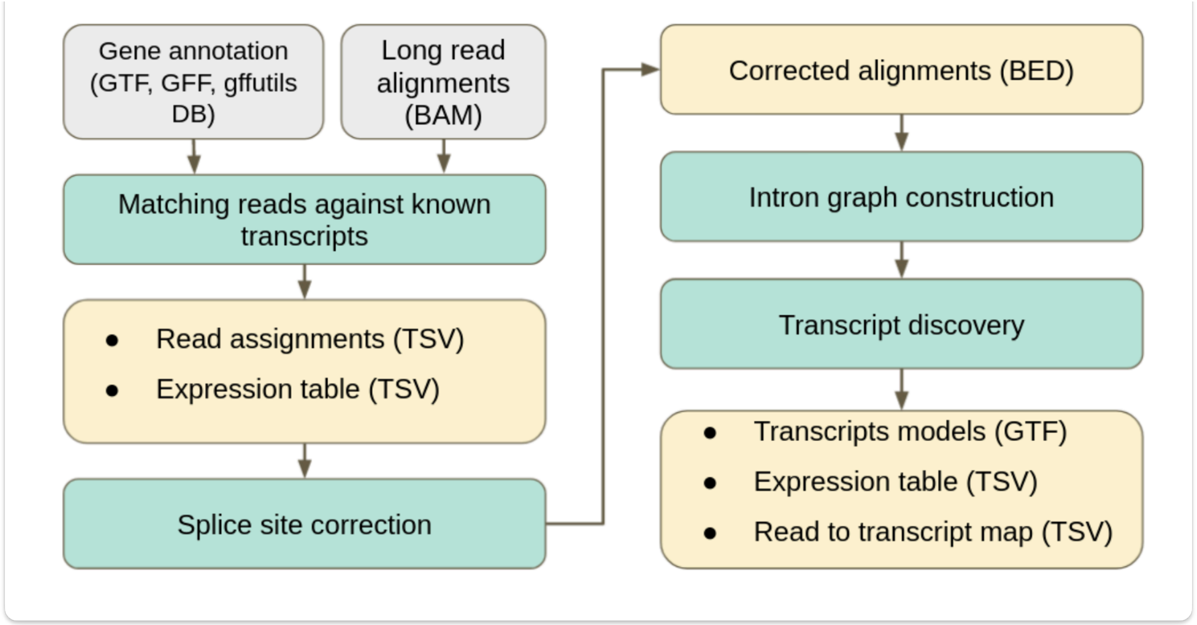


IsoQuant and RNABloom

IsoQuant

IsoQuant is a tool for the **genome-based analysis of long RNA reads**, such as PacBio or Oxford Nanopores.

Nat Biotechnol **41**, 915–918 (2023)


Gene annotation (GTF, GFF, gffutils DB)
Long read alignments (BAM)
Matching reads against known transcripts
• Read assignments (TSV)
• Expression table (TSV)
Splice site correction
Corrected alignments (BED)
Intron graph construction
Transcript discovery
• Transcripts models (GTF)
• Expression table (TSV)
• Read to transcript map (TSV)

RNABloom

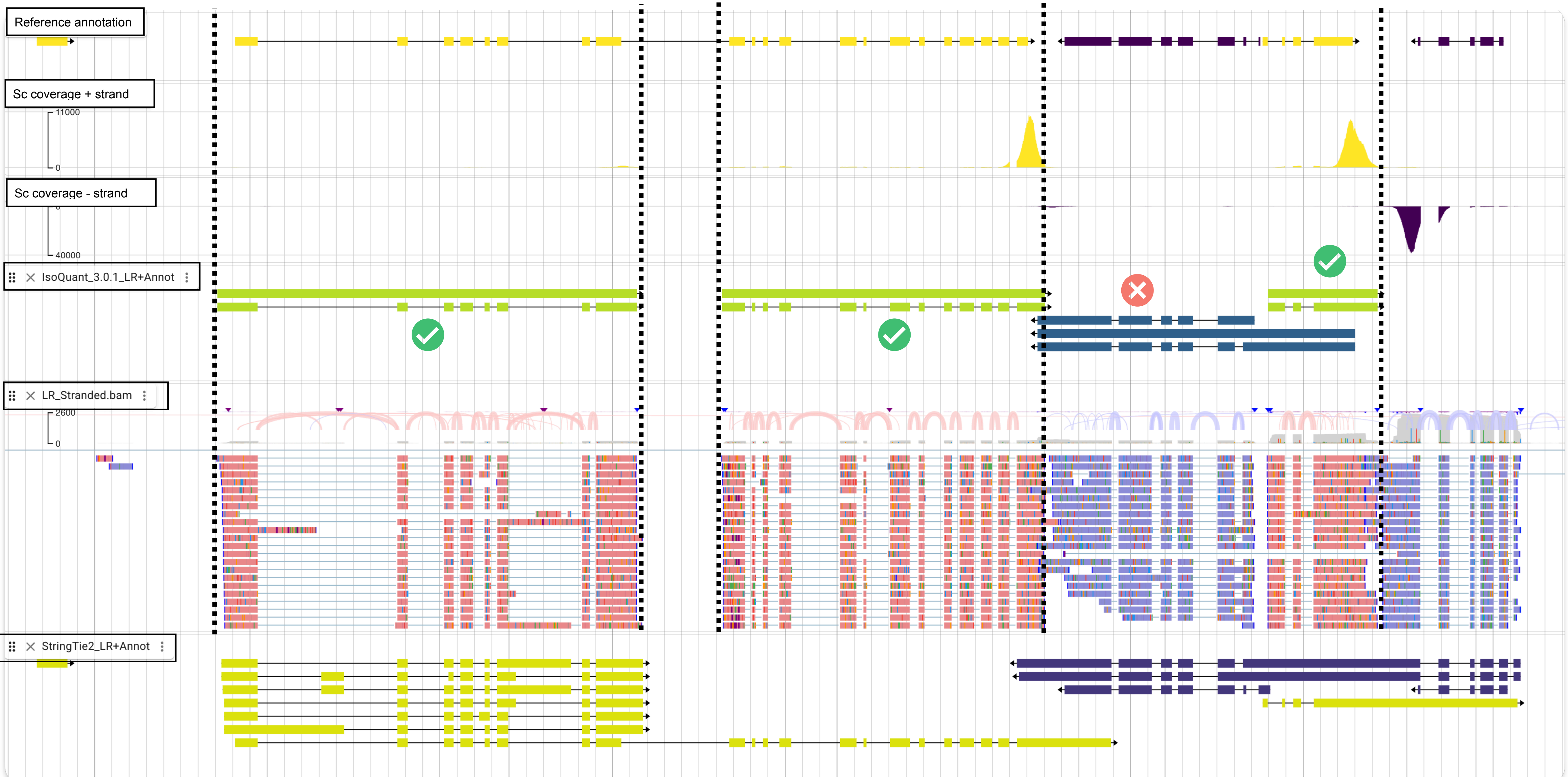
RNA-Bloom is a fast and memory-efficient *de* ***novo*** **transcript sequence assembler**

Nat Commun **14**, 2940 (2023)

➡Testing these two tools based on different and possibly complementary strategies.

IsoQuant - LR+Annot

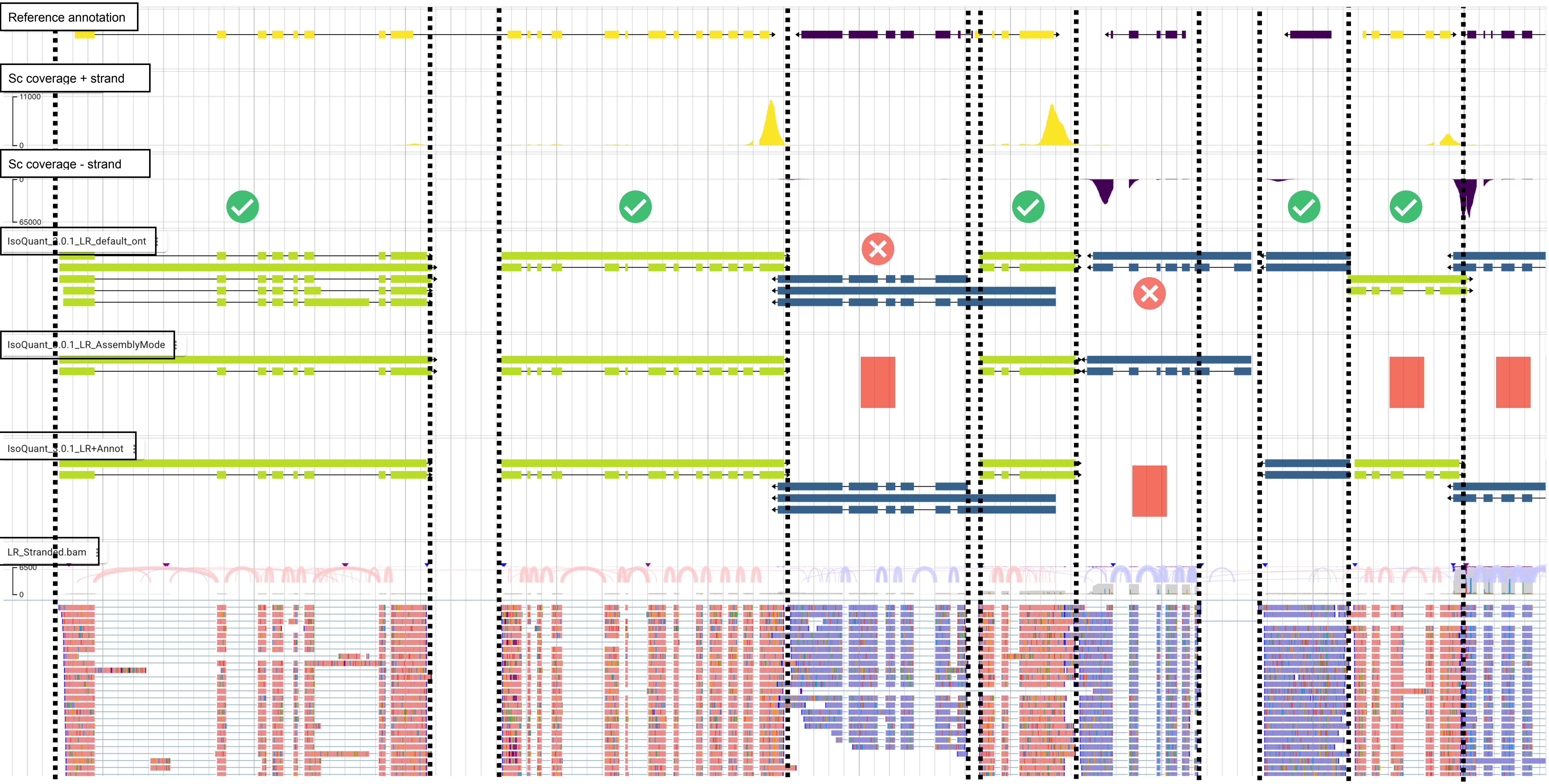
Transcript model construction=default ONT



The annotation does not outweigh the sequencing data.

IsoQuant still makes errors and misses genes in an incomprehensible manner.

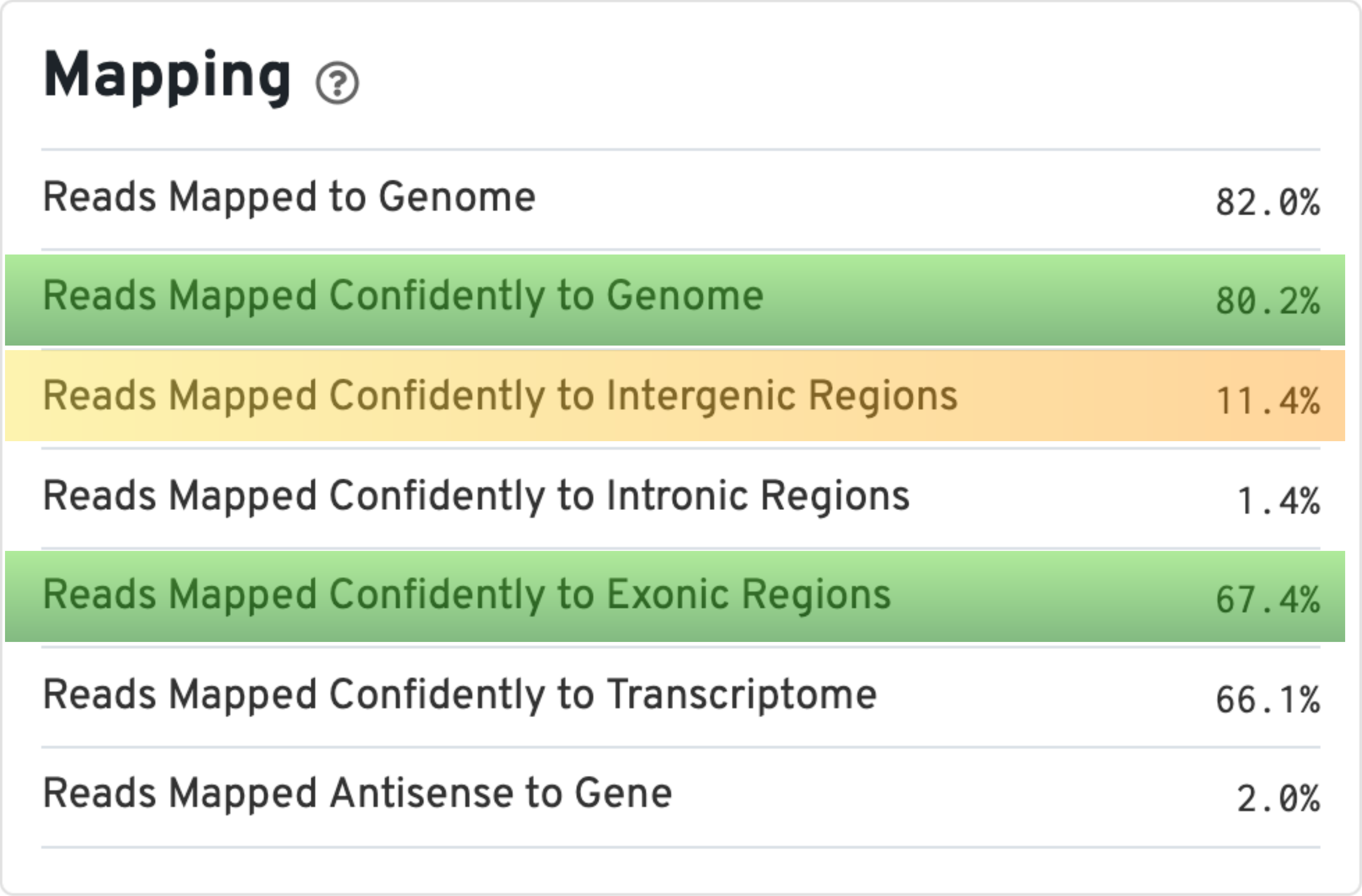
IsoQuant - LR+Annot (default ONT) / LR (default ONT / Assembly)



IsoQuant can be run with many different parameters

➡ the default mode for ONT data seems the most suitable

SC counts with IsoQuant LR defaultONT



Mapping ?

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	80.2%
Reads Mapped Confidently to Intergenic Regions	11.4%
Reads Mapped Confidently to Intronic Regions	1.4%
Reads Mapped Confidently to Exonic Regions	67.4%
Reads Mapped Confidently to Transcriptome	66.1%
Reads Mapped Antisense to Gene	2.0%

➡ **Counts improve a lot even if we have just seen the annotation is not perfect**

In my last exemple, annotations are missed but they are not misidentified

What am I looking for? A beautiful annotation or an annotation that allows me to count accurately ?

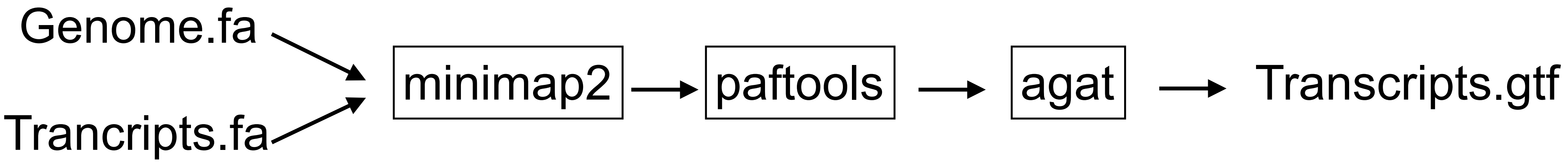
RNABloom

RNABloom is a de novo transcript assembly tool:

➡ It takes FASTQ files (LR and SR to polish junctions) as input and outputs FASTA files containing the transcripts

It is therefore necessary to align the generated transcripts to the genome and produce GFF/GTF files

To convert FASTA to GTF



Genome.fa

Trancripts.fa

minimap2

paftools

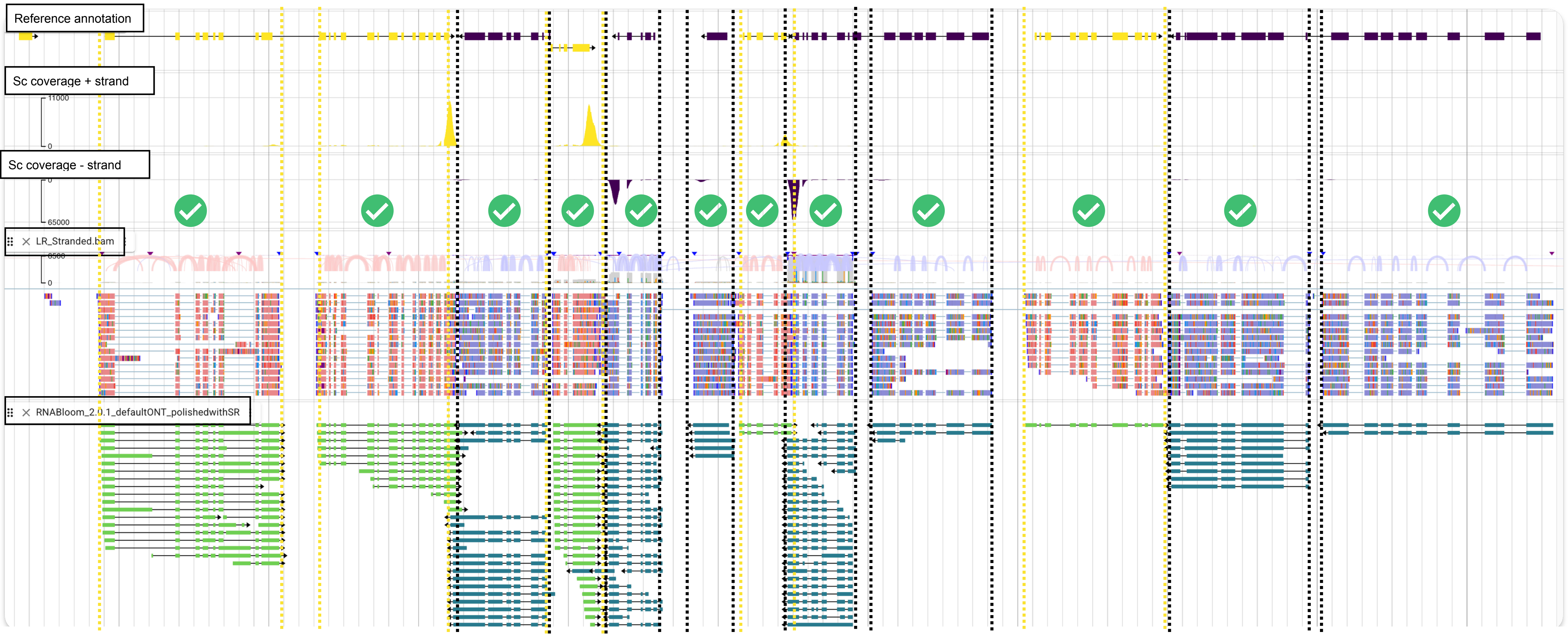
agat

Transcripts.gtf

New strategies to improve minimap2 alignment accuracy, *Bioinformatics*, Volume 37, Issue 23, December 2021, Pages 4572–4574

Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (Version v0. 7.0). Zendo. doi. 2023;10

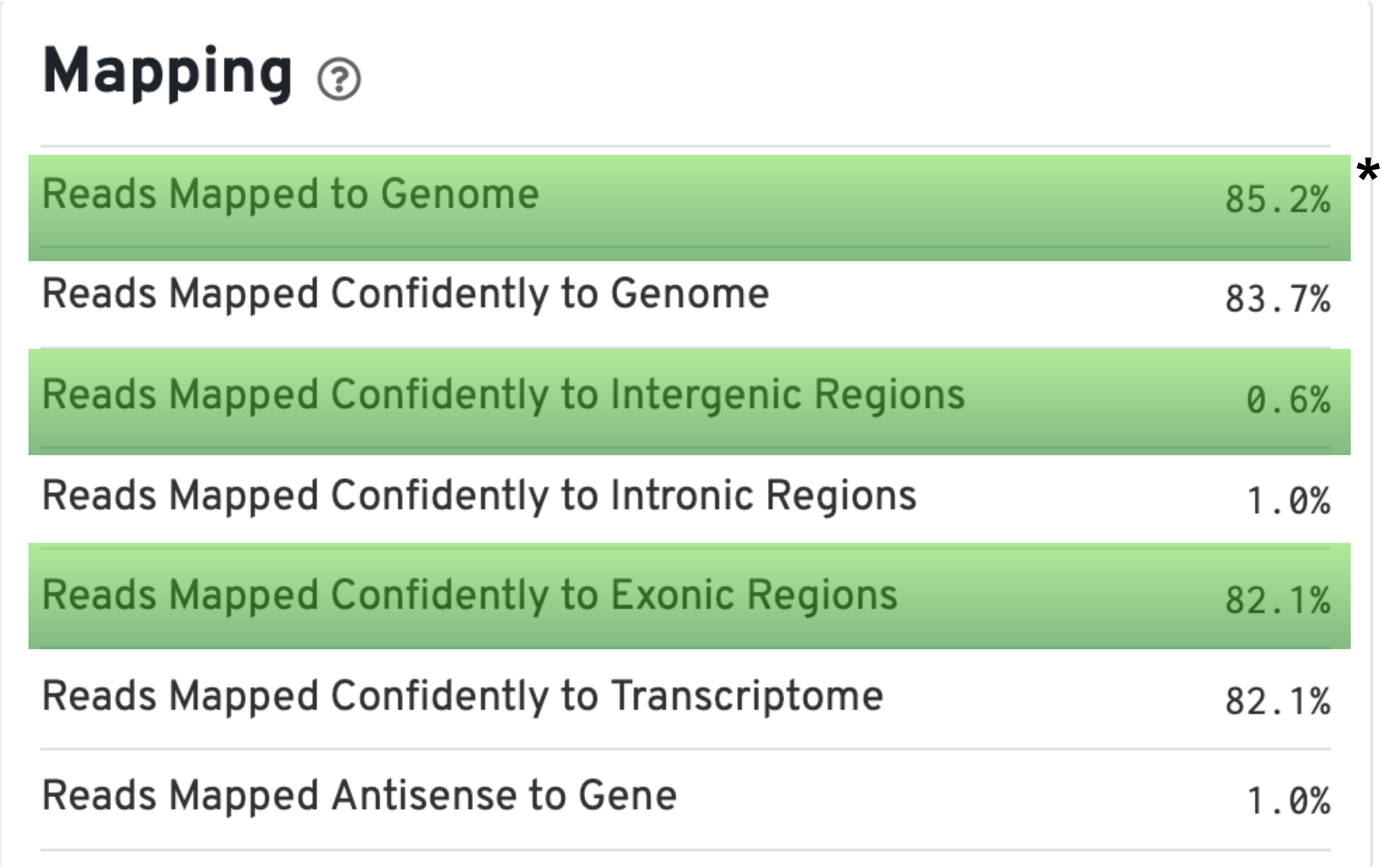
RNABloom (LR polished with SR)



All genes are annotated according to the input data but no filter is applied to the exons.

➡ One can perceive that it will be complicated to work at the isoform level

SC counts with RNABloom (LR polished with SR)



Mapping ⓘ

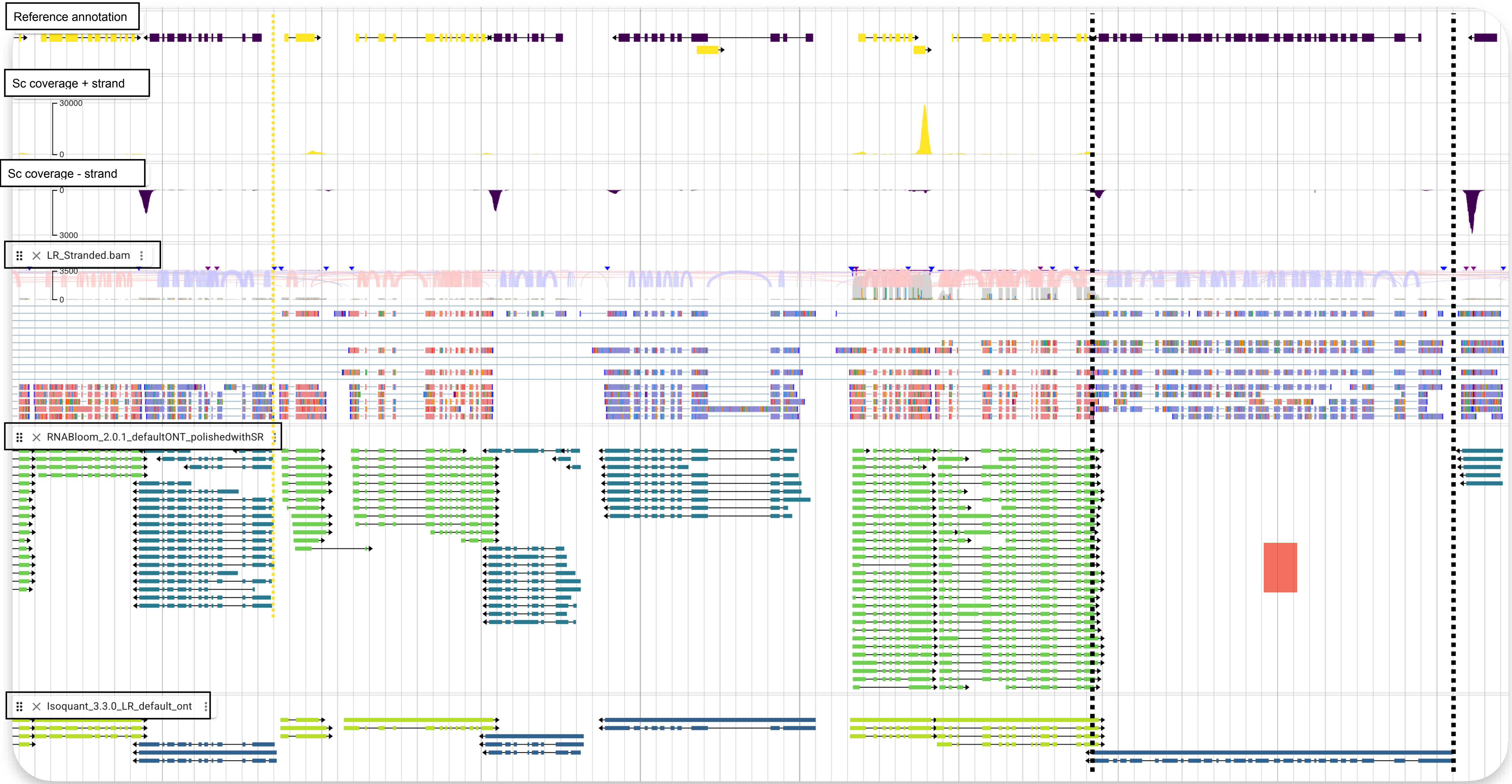
Reads Mapped to Genome	85.2%
Reads Mapped Confidently to Genome	83.7%
Reads Mapped Confidently to Intergenic Regions	0.6%
Reads Mapped Confidently to Intronic Regions	1.0%
Reads Mapped Confidently to Exonic Regions	82.1%
Reads Mapped Confidently to Transcriptome	82.1%
Reads Mapped Antisense to Gene	1.0%

*

➡ **RNABloom allows counting the entire SingleCell signal**

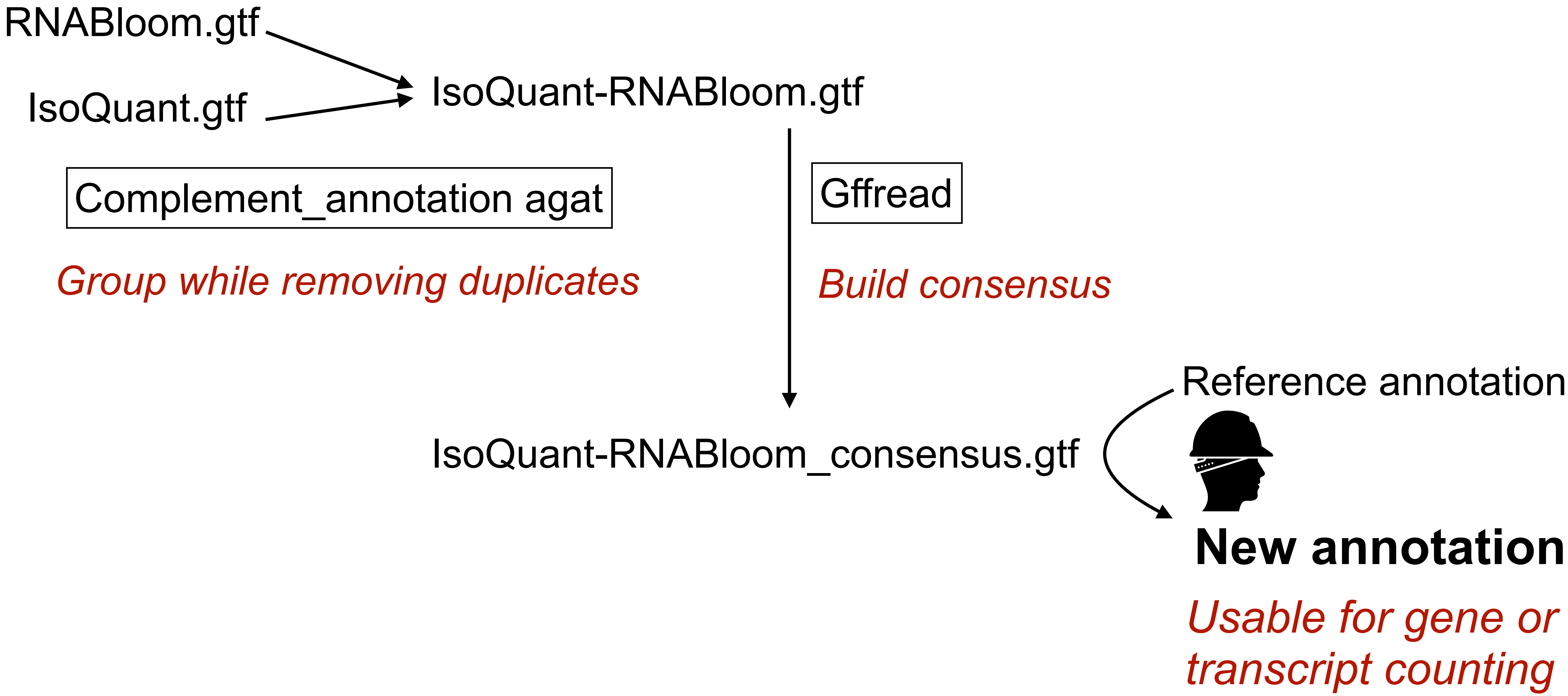
(Biological replicate of the 1st SC)

A mix of RNABloom and IsoQuant ?



- RNABloom returns almost all the transcripts revealed by the sequencing data (sometimes, IsoQuant can be a good complement)
- IsoQuant gives consensus to simplify RNABloom results at the isoform level

How to create a consensus between IsoQuant and RNABloom while incorporating gene names from the official annotation?



RNABloom.gtf

IsoQuant.gtf

IsoQuant-RNABloom.gtf

Complement_annotation agat

Group while removing duplicates

Gffread

Build consensus

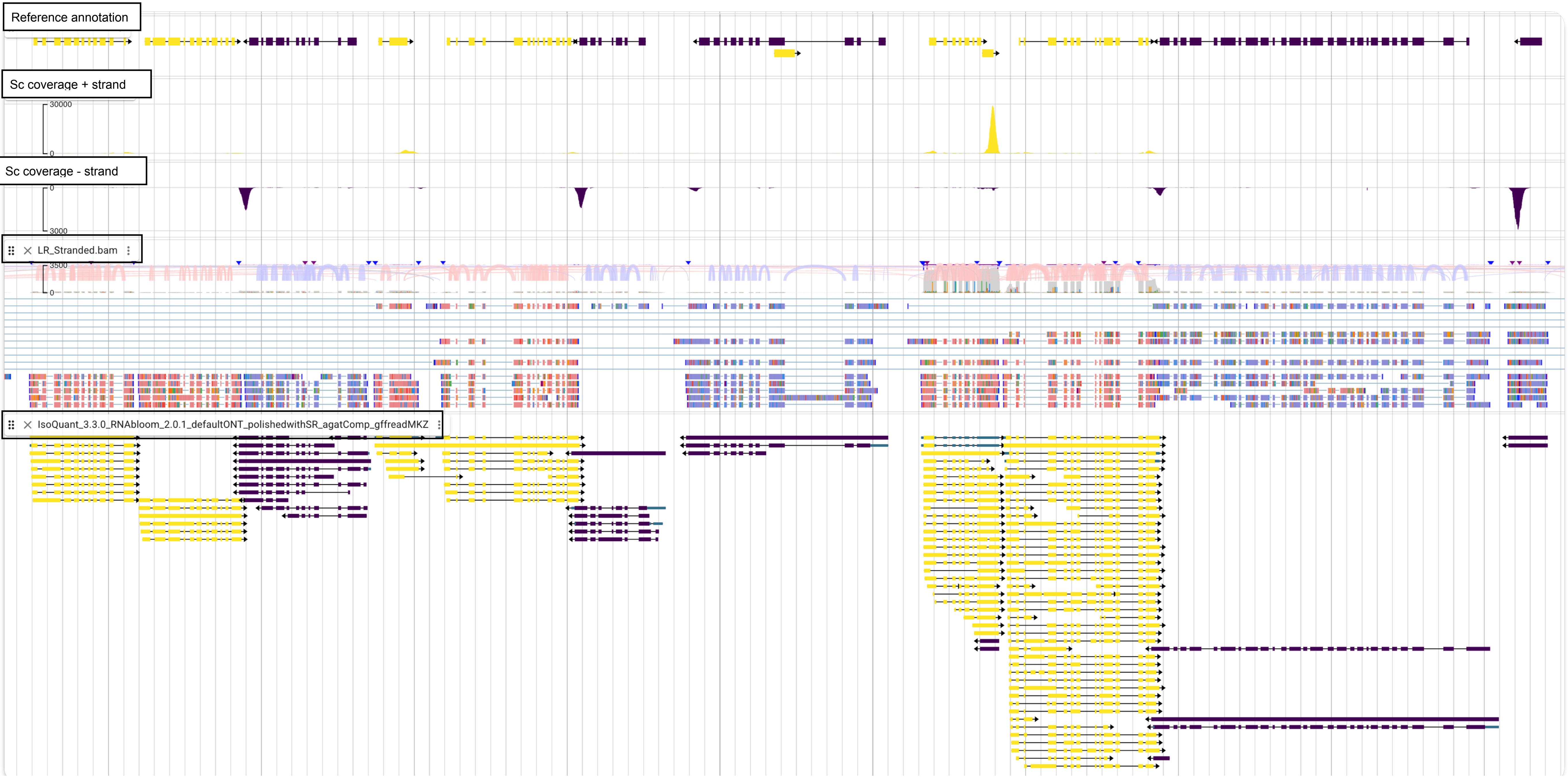
IsoQuant-RNABloom_consensus.gtf

Reference annotation

New annotation

Usable for gene or transcript counting

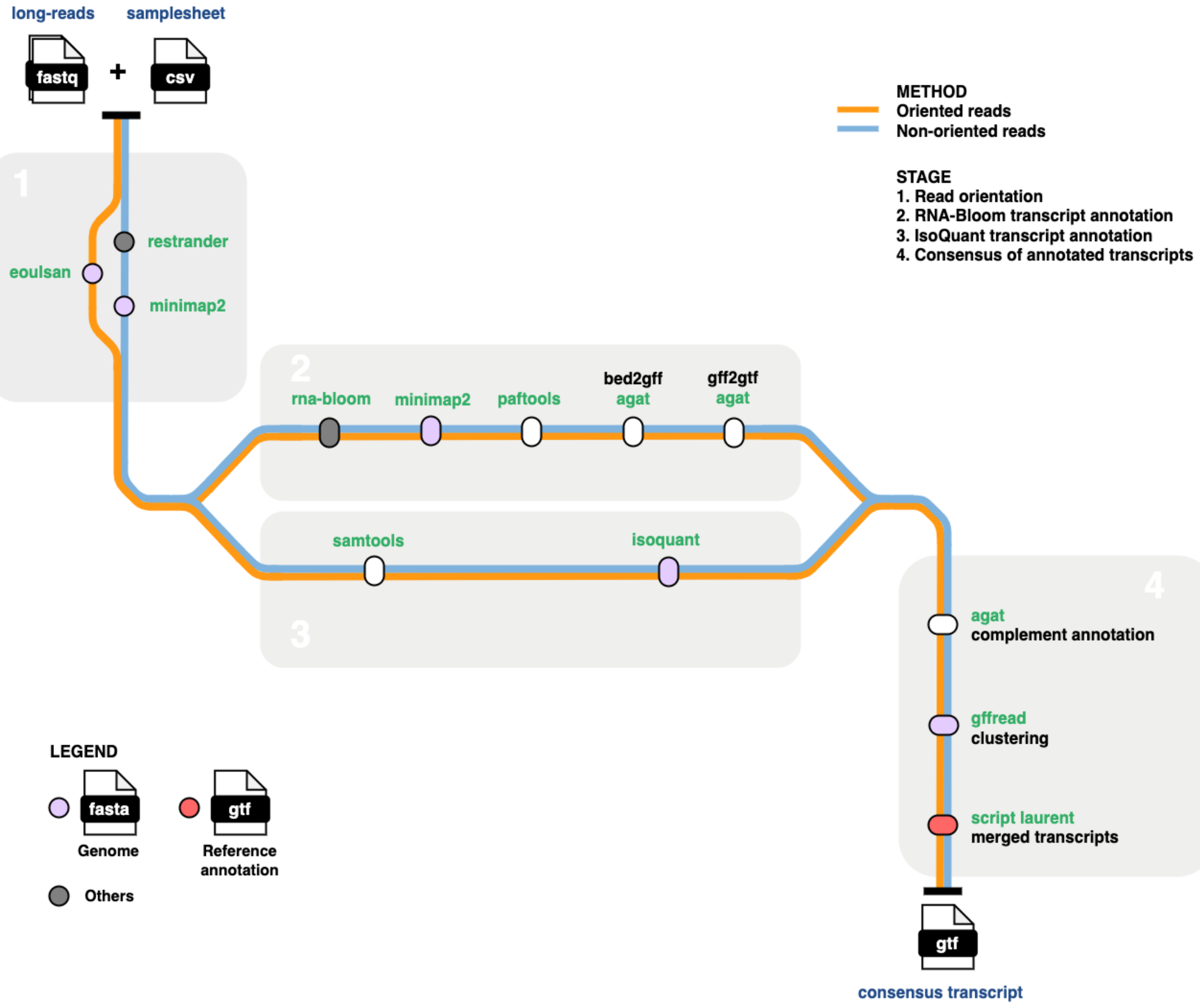
IsoQuant+RNABloom+Agat+Gffread



➡**The situation is not perfect, but we can work at the <u>gene level</u>**

Egzotec : a nextflow pipeline to automate annotation from fastq to gtf

➡ **We are working on a pipeline to automate these steps because annotation projects are piling up dangerously…**



1. Read orientation
   i. Oriented protocol (eoulsan)
   ii. Non-oriented protocol (restrander)
2. Transcript annotation with RNA-Bloom
   i. Transcript annotation (rna-bloom) (with optional short-read polishing)
   ii. Genome mapping (minimap2)
   iii. Bam to bed file conversion (minimap2-paftools)
   iv. Bed to gff file conversion (agat)
   v. Gff to gtf file conversion (agat)
3. Transcript annotation with Isoquant
   i. Genome mapping (minimap2)
   ii. Sam to bam file conversion (samtools)
   iii. Transcript annotation (isoquant)
4. Complement annotation (agat)
5. Clusterisation (gffread)
6. Merge annotation

GenomiqueENS/egzotek

Salomé Brunon

Take home messages

- Long reads restranded RNASeq data are good data sets
- Add the reference annotation at the very last moment to prevent it from taking over
- Tools designed for long reads do better than tools that are adapted to long reads
- Non model organisms can behave very differently than well annotated model organisms
- Even model organisms can need a reannotation
- Check your data and annotations in a genome browser: JBrowse2 or IGV

Perspectives

- Retrain Helixer model on insects and test it

- Add QC tools to Egzotek
  - ➡ Validation of the reference annotation gff file
  - ➡ Parts of SQANTI3
  - ➡ Functional annotation
  - ➡ BUSCO, Compleasm

Tested on other species (fungi or insects) but not retained tool

Annotation dedicated to long reads

- Flair
- Bambu

Junction validation

- Portcullis

Annotation pipeline initially based on short reads

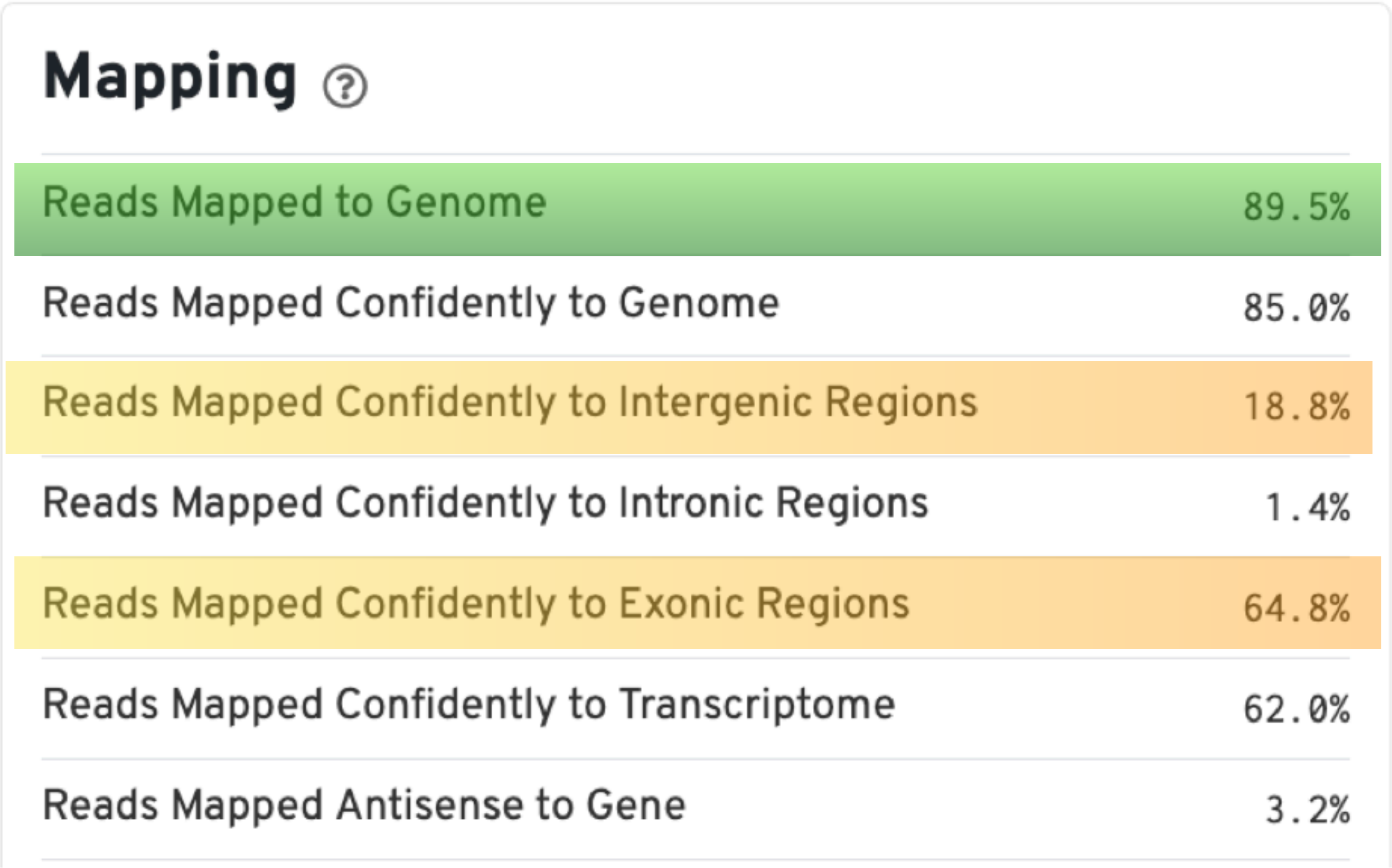
- Mikado
- Funannotate

Consensus building

- Tama-collapse
- Tmerge

What about BRAKER3 ?

Untested myself….but CellRanger results are not that impressive…



Mapping

Reads Mapped to Genome	89.5%
Reads Mapped Confidently to Genome	85.0%
Reads Mapped Confidently to Intergenic Regions	18.8%
Reads Mapped Confidently to Intronic Regions	1.4%
Reads Mapped Confidently to Exonic Regions	64.8%
Reads Mapped Confidently to Transcriptome	62.0%
Reads Mapped Antisense to Gene	3.2%

*

< than IsoQuant results

* Genome improved version

The GenomiqueENS team

https://genomique.biologie.ens.fr

Wet lab

Catherine Senamaud-Beaufort

Corinne Blugeon

Tiphaine Marvillet

Oumy Seydi

Morgane Thomas-Chollier

Bioinformatics

Ali Hamraoui

Laurent Jourdren

Sophie Lemoine

Salomé Brunon

Stéphane Le Crom