Atelier scRNA-seq

Checking signal with a genome browser (IGV)

Margot Tragin - Institut Imagine, Paris

D'après Sophie Lemoine,
IBENS - GenomiqueENS, Paris

École de bioinformatique EB3I 2025

Single cell analysis is about counts, so why visualizing the reads in a genome browser ?

- When to check?
- How to check?
- Solution(s) to improve results?
- Integrative Genomics Viewer

When to check ?

- You do not understand the counts on your favourite gene?
- The global results look weird?

→ **Your reads may not overlap annotations…**

When to check ?

- The mapping statistics are not good

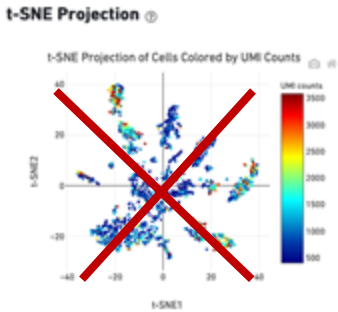
Alerts

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.

→ The clusters are then based on a very limited amount of reads and will not be reliable

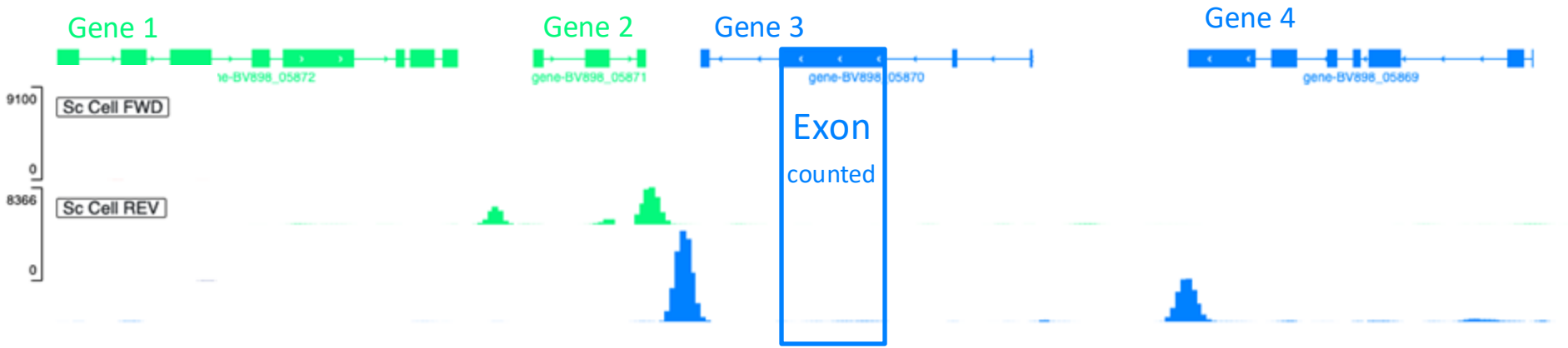
→ If the counted reads are low, the estimated number of cells will be smaller (relies on a smaller amount of BC+UMIs)



Bad counts result in poor and even fake results!!

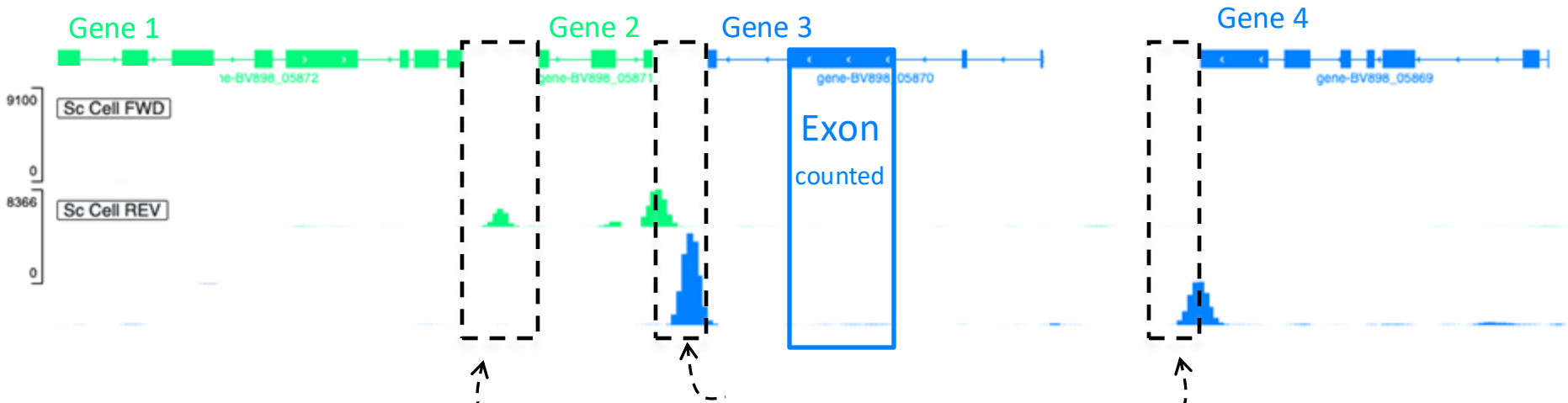
How to check ?

- Take a look at the signal



How to check ?

- Take a look at the signal



What is not annotated cannot be counted…

How to check ?

- Are the **annotations** adapted to the **biological question**?
- If the **gene annotations** are **deduced** from **the protein annotations**, the **5' and 3' UTRs** (10x data) are **not included**
- If you work on a **not so popular model organism** or **cancer data**, **annotations may not fit** your data

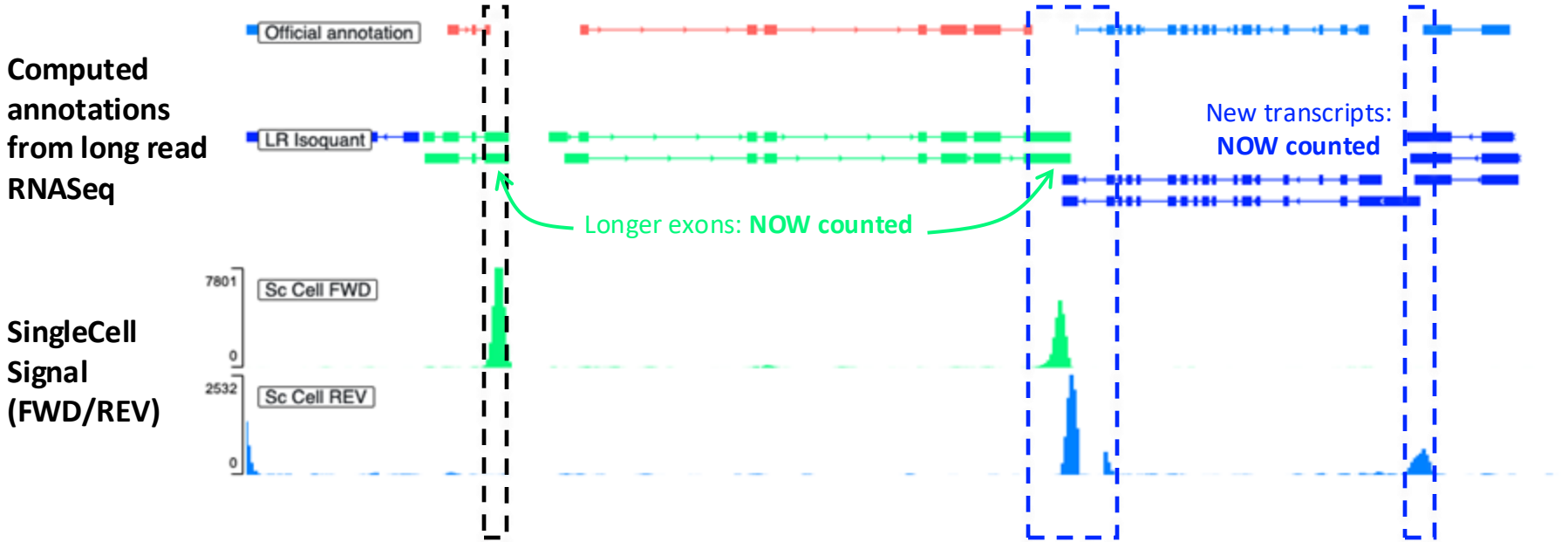
→ **It's an issue for single cell analyses!**

Annotation: position on the genome + informations (genes, exons, introns…)

- **CDS**: coding sequences
- **ORFs**: open reading frame
- **…**

Solution(s) to improve results

- Build **new annotation** using either **long or/and short read** bulk RNAseq (stranded if possible) and tools such as **IsoQuant**, **Stringtie2,** etc…



Solution(s) to improve results

- Before and after re-annotation analyses

	Official annotation	Isoquant annotation
Estimated Number of Cells	**2114**	**2624**
Reads Mapped to Genome	82	82
Reads Mapped Confidently to Genome	79,9	80,2
Reads Mapped Confidently to Intergenic Regions	**46,3**	**11,4**
Reads Mapped Confidently to Intronic Regions	3,6	1,4
Reads Mapped Confidently to Exonic Regions	**30**	**67,4**
Reads Mapped Confidently to Transcriptome	**23,3**	**66,1**
Reads Mapped Antisense to Gene	0,5	2

Take home messages

- If in **doubt** about single cell results → **have a look** at the signal!
- **Don’t interpret your data**, clusters and number of cells will not be right.
- Check if reads and/or coverage **match expected annotations**
- If not, **try new annotations**…

Integrative Genomics Viewer

- IGV is a java **multiplatform tool** : It will work under **Linux, macOSX and Windows**
- IGV is **open, free, lively** and maintained at the Broad Institute

IGV is available in multiple forms, including:

- the original **IGV** - a Java desktop application
- **IGV-Web** - a web application

https://igv.org

IGV - DATA list

- A **reference** genome (fasta file)
- An **annotation** file (gtf or gff file)

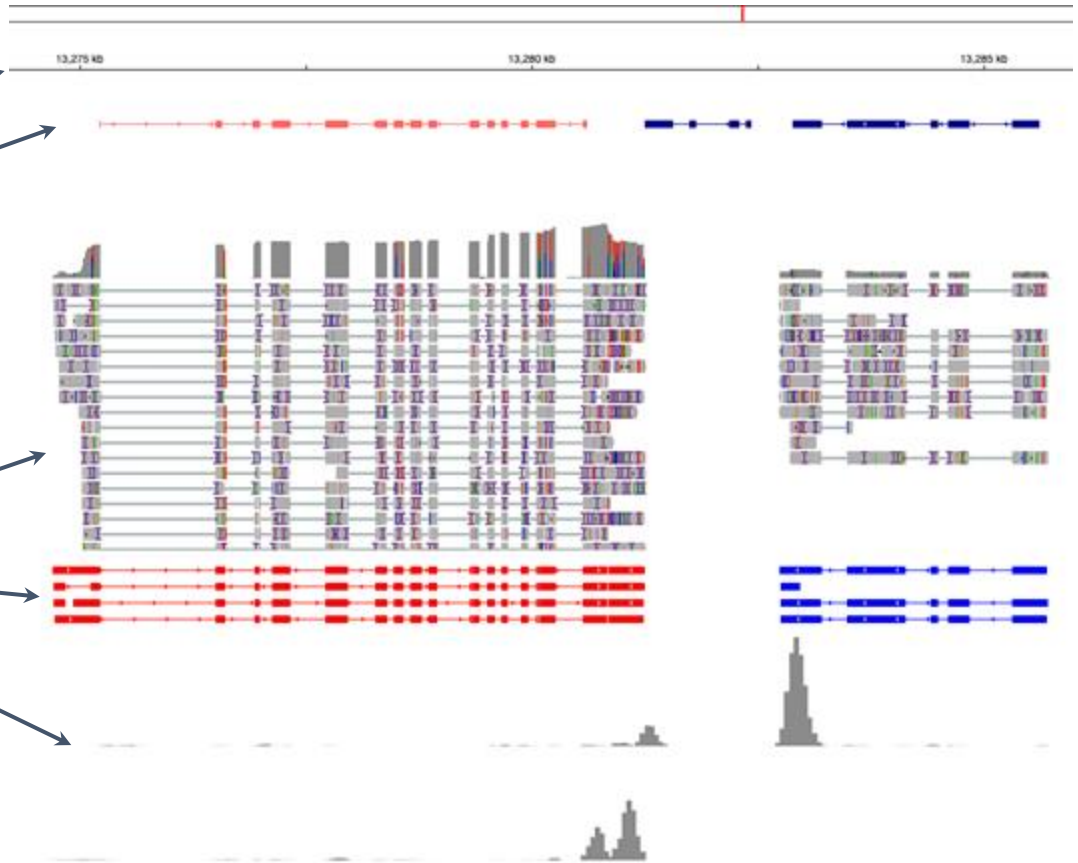
Already used to perform your SC analysis

The files resulting from **alignements**

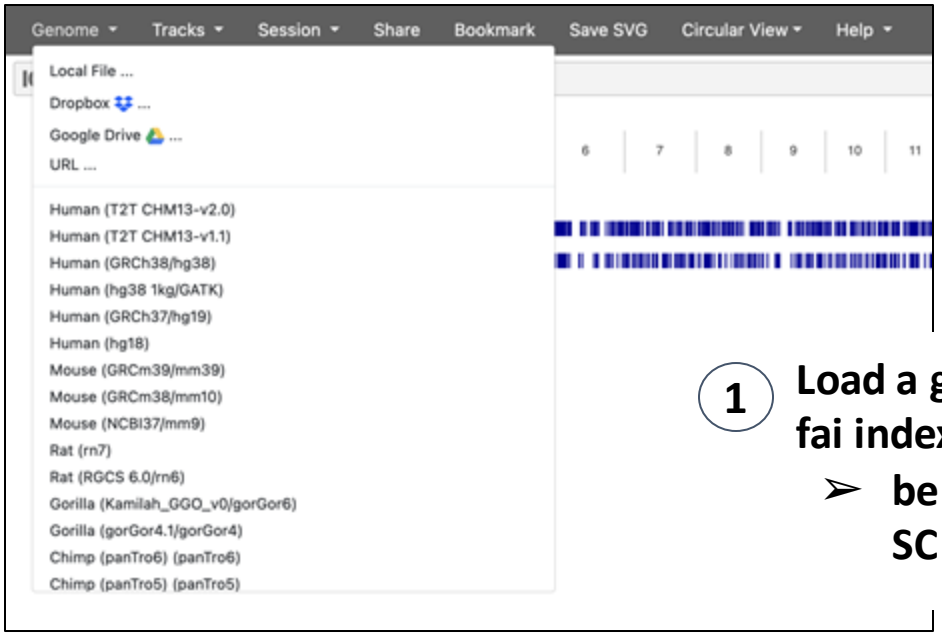
*.**bam** files (and bai index files)

*.**bed** files (read position files)

*.**bedgraph** files (coverage files)

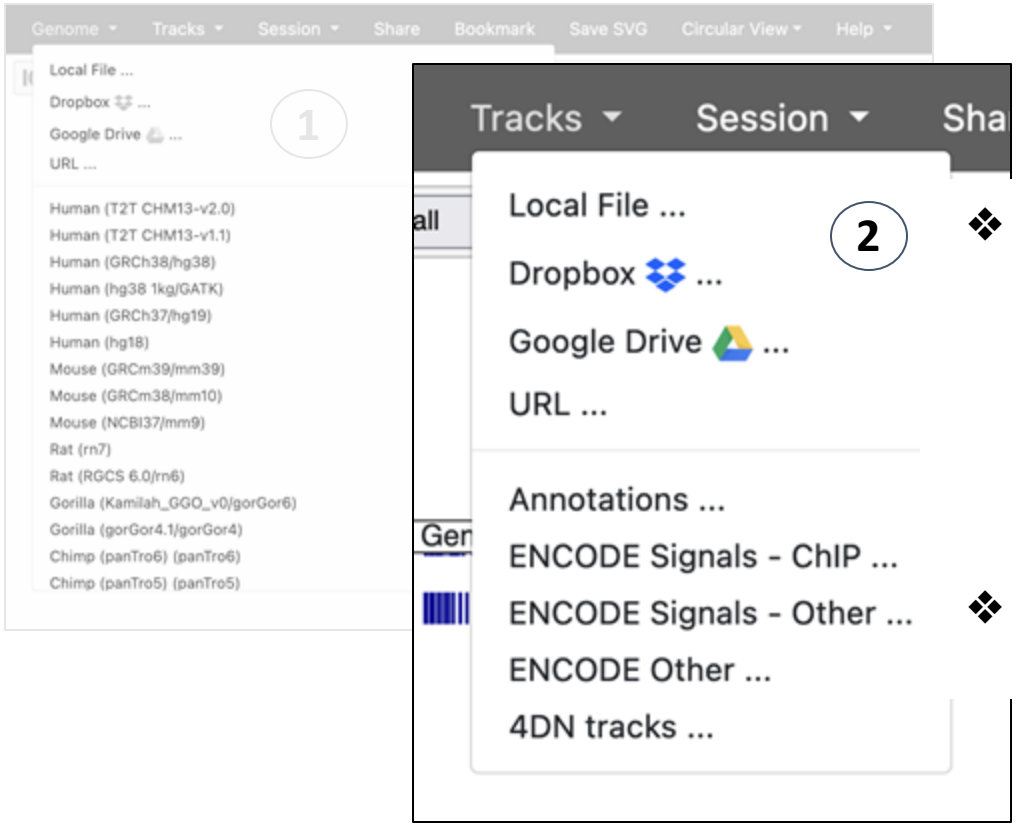


IGV - tutorial



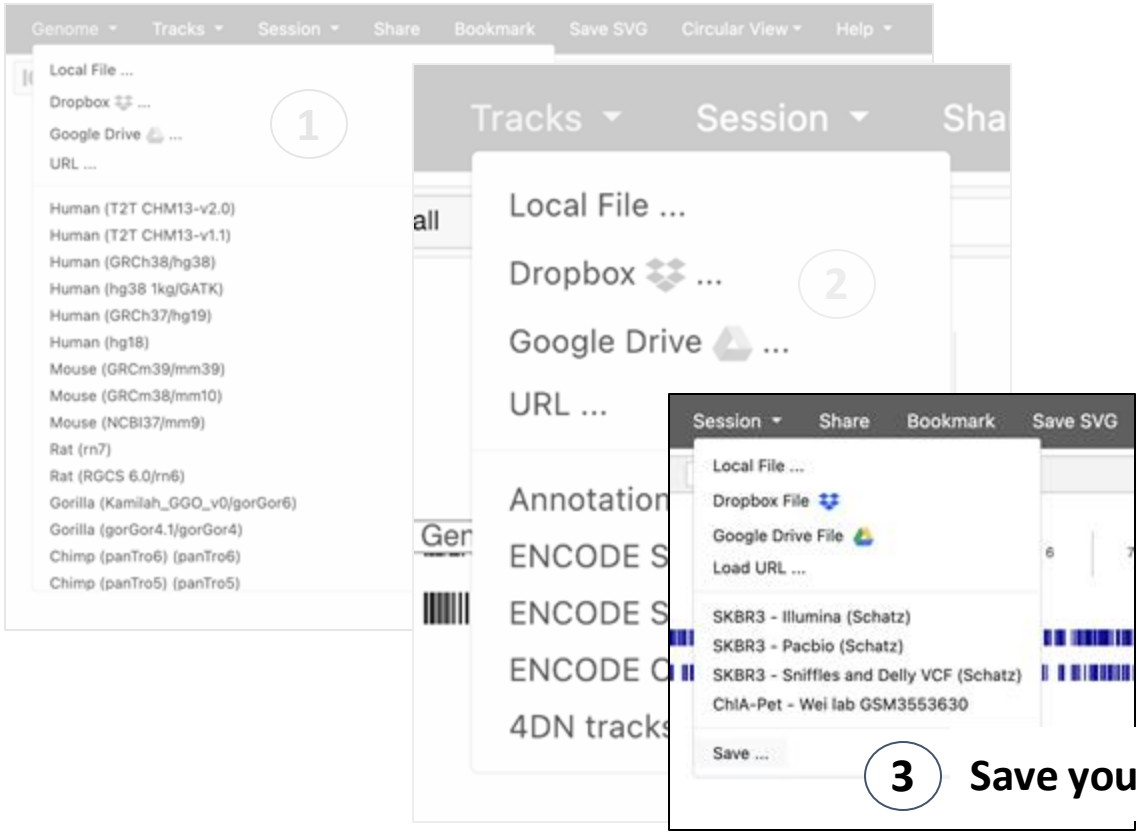
1. **Load a genome from the list or upload a fasta file (with fai index file)**
   - ➢ **be sure it's the same as the genome used for your SC analysis**

IGV - tutorial



- **Load an annotation from the list or upload a gtf file**
  - **Again be sure it's the same as the annotation used for your SC analysis**
  - **Be sure it goes with your genome file (Chromosome name…)**
- **Load your bam, bed, bedgraph files**

IGV - tutorial



3 **Save you session to avoid reloading your files**

References

Single cell analysis failure and gene annotation

Pool, AH., Poldsam, H., Chen, S. *et al.* Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references. *Nat Methods* **20**, 1506–1515 (2023).

IsoQuant

Prjibelski, A.D., Mikheenko, A., Joglekar, A. *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* **41**, 915–918 (2023)
https://github.com/ablab/IsoQuant

Stringtie2

Shumate A, Wong B, Pertea G, Pertea M Improved transcriptome assembly using a hybrid of long and short reads with StringTie, *PLOS Computational Biology* 18, 6 (2022)
https://github.com/gpertea/stringtie

IGV

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P.Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

https://igv.org/doc/desktop/