



Atelier scRNA-seq

Check your signal with a genome browser
IGV

Sophie Lemoine, IBENS - GenomiqueENS, Paris



Organisation of the scRNA-seq course

- From cells to nucleotide sequences (reads)
 - focus on the 10X genomics technology
 - how are the reads organised
- Preprocessing : from reads to raw count matrix
 - quality check (FASTQC)
 - mapping (STAR)
 - how is annotation used
 - barcode and UMI treatment
 - visualizing the reads
 - constructing the count matrix
 - call cells / empty droplets filtering

Single cell analysis is about counts

so why visualizing the reads in a genome browser ?


- You do not understand the counts on your favourite gene ?
 - The global results look weird ?
- Your reads may not overlap the gene positions...



Bad counts result in poor and even fake results

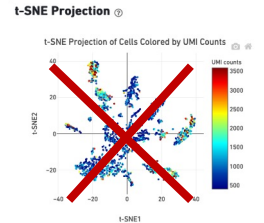
Alerts

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.

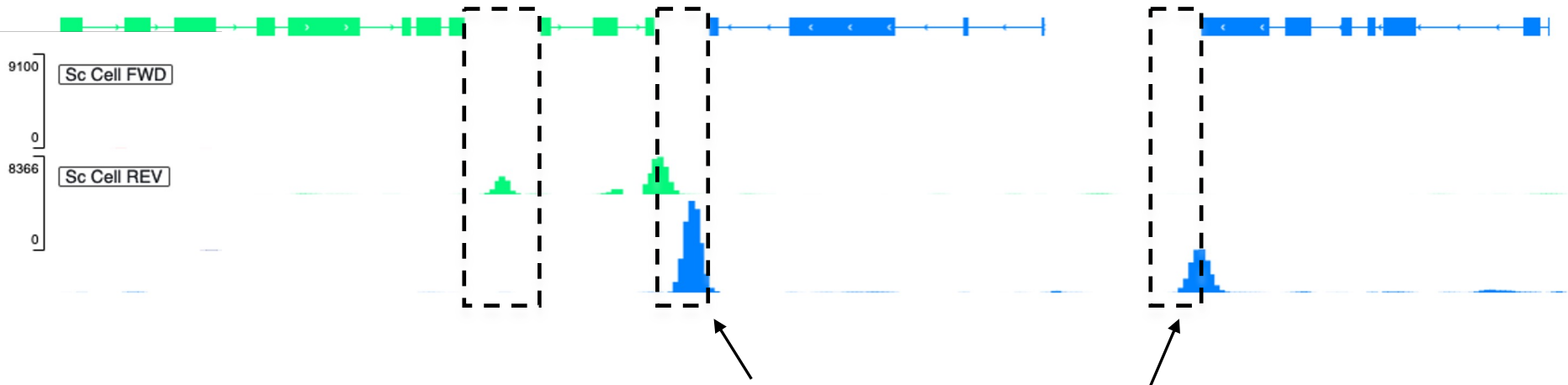
→ The clusters are then based on a very limited amount of reads and will not be reliable

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



What is not annotated cannot be counted

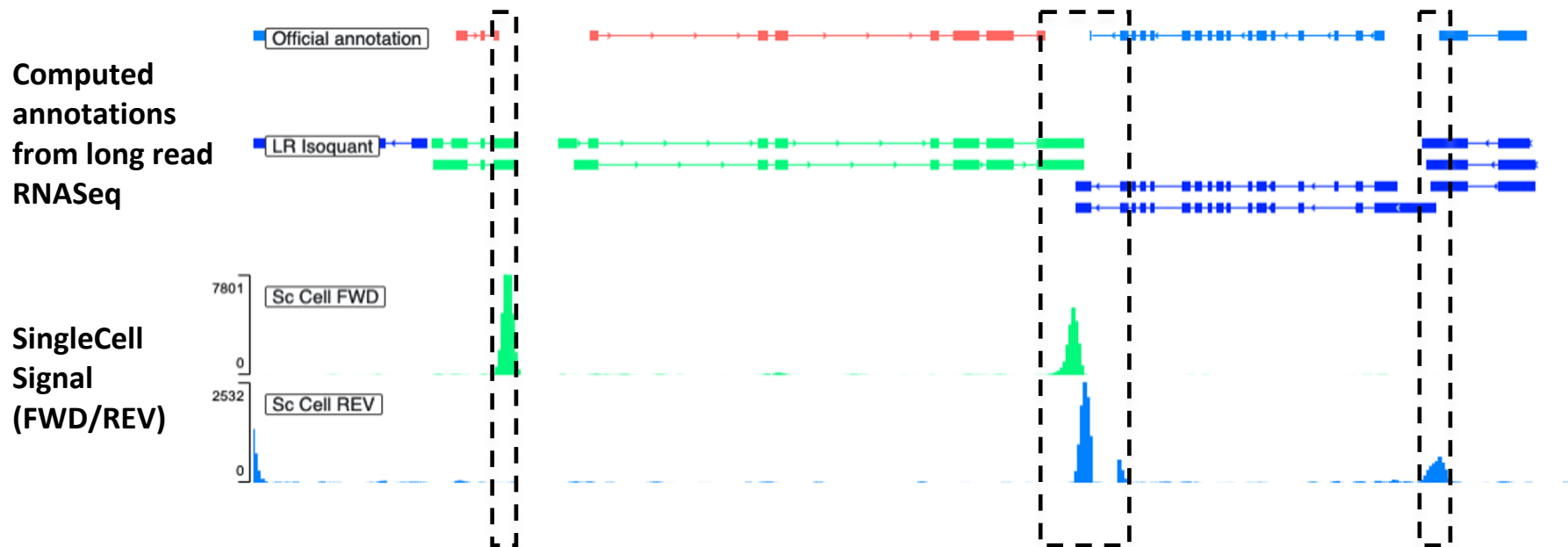
- ❖ If the **gene annotations** are deduced from the protein annotations, the 5' and 3' UTRs (10x data) are **not included**
- ❖ If you work on a **not so popular model organism** or **cancer data**, annotations may not **fit** your data
 - It's clearly **an issue for single cell analyses**



Only the reads out of the rectangles are counted
→ Most of the signal is excluded

Compute a new annotation using bulk RNASeq data (short and long reads)

You can build a **new annotation** using either **long or/and short read protocols** (but stranded if possible) and tools such as **Isoquant**, **Stringtie2**, etc...

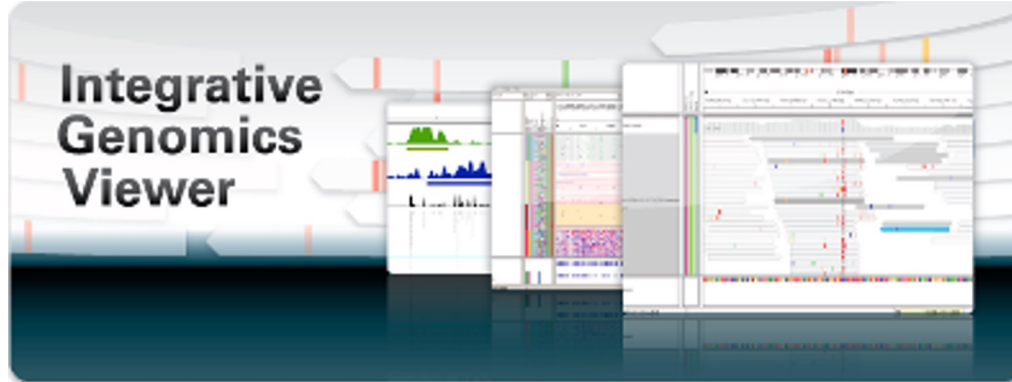


Before and after re-annotation analyses

	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

Integrative Genome viewer (IGV) is the most popular Genome Browser

- 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

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

<https://igv.org>

What do you need to use IGV ?

- ❖ A **reference genome** (fasta file)

- ❖ An **annotation file** (gtf or gff file)

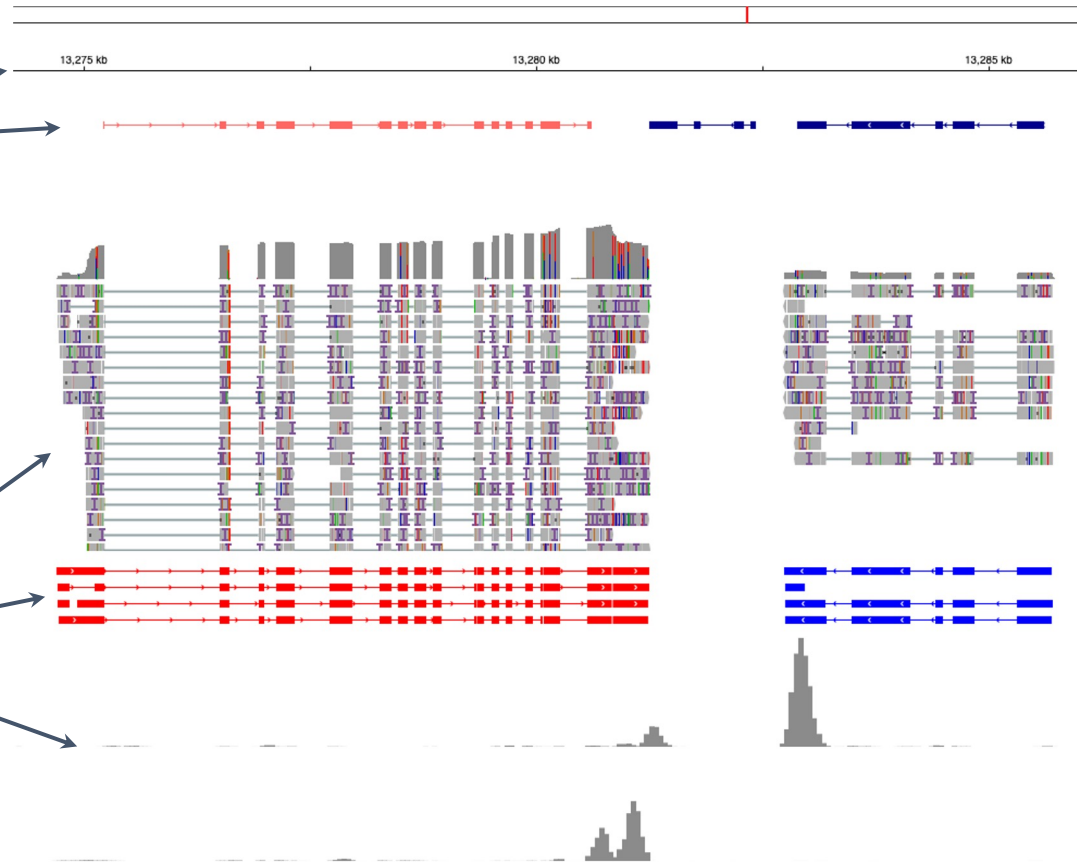
- Already used to perform your SC analysis

- ❖ The files resulting from **your alignments**

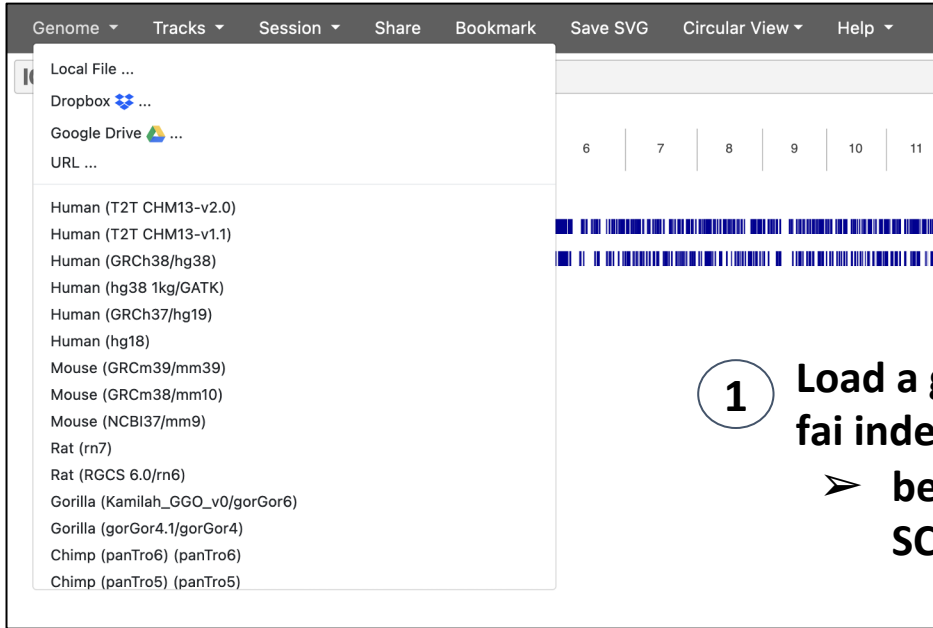
- **bam** files (and bai index files)

- **bed** files (read position files)

- **bedgraph** files (coverage files)



How to begin with IGV ?



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**

How to begin with IGV ?

The screenshot shows the IGV (Integrative Genomics Viewer) interface. At the top, there is a menu bar with options: Genome, Tracks, Session, Share, Bookmark, Save SVG, Circular View, and Help. Below the menu bar, there is a list of genome annotations. A red circle with the number '1' highlights the 'Local File ...' option in the list. To the right, a 'Tracks' dropdown menu is open, showing options: Local File ..., Dropbox ..., Google Drive ..., URL ..., Annotations ..., ENCODE Signals - CHIP ..., ENCODE Signals - Other ..., ENCODE Other ..., and 4DN tracks ... A red circle with the number '2' highlights the 'Local File ...' option in the dropdown menu. A red diamond symbol is placed to the right of the 'Local File ...' option in the dropdown menu.

Genome ▾ Tracks ▾ Session ▾ Share Bookmark Save SVG Circular View ▾ Help ▾

Local File ...
Dropbox ...
Google Drive ...
URL ...

Human (T2T CHM13-v2.0)
Human (T2T CHM13-v1.1)
Human (GRCh38/hg38)
Human (hg38 1kg/GATK)
Human (GRCh37/hg19)
Human (hg18)
Mouse (GRCm39/mm39)
Mouse (GRCm38/mm10)
Mouse (NCBI37/mm9)
Rat (rn7)
Rat (RGCS 6.0/rn6)
Gorilla (Kamilah_GGO_v0/gorGor6)
Gorilla (gorGor4.1/gorGor4)
Chimp (panTro6) (panTro6)
Chimp (panTro5) (panTro5)

Tracks ▾ Session ▾ Sha

Local File ...
Dropbox ...
Google Drive ...
URL ...

Annotations ...
ENCODE Signals - CHIP ...
ENCODE Signals - Other ...
ENCODE Other ...
4DN tracks ...

❖ 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

How to begin with IGV ?

The image shows a screenshot of the IGV (Integrative Genomics Viewer) interface. The top menu bar includes 'Genome', 'Tracks', 'Session', 'Share', 'Bookmark', 'Save SVG', 'Circular View', and 'Help'. A dropdown menu is open under 'Session', showing options: 'Local File ...', 'Dropbox File', 'Google Drive File', 'Load URL ...', and 'Save ...'. A second dropdown menu is open under 'Local File ...', showing options: 'Local File ...', 'Dropbox ...', 'Google Drive ...', and 'URL ...'. A third dropdown menu is open under 'Google Drive File', showing options: 'SKBR3 - Illumina (Schatz)', 'SKBR3 - Pacbio (Schatz)', 'SKBR3 - Sniffles and Delly VCF (Schatz)', and 'ChIA-Pet - Wei lab GSM3553630'. A fourth dropdown menu is open under 'Annotations ...', showing options: 'ENCODE Signals - C...', 'ENCODE Signals - C...', 'ENCODE Other ...', and '4DN tracks ...'. A list of genomes is visible on the left side of the interface, including Human (T2T CHM13-v2.0), Human (T2T CHM13-v1.1), Human (GRCh38/hg38), Human (hg38 1kg/GATK), Human (GRCh37/hg19), Human (hg18), Mouse (GRCm39/mm39), Mouse (GRCm38/mm10), Mouse (NCBI37/mm9), Rat (rn7), Rat (RGCS 6.0/rn6), Gorilla (Kamilah_GGO_v0/gorGor6), Gorilla (gorGor4.1/gorGor4), Chimp (panTro6) (panTro6), and Chimp (panTro5) (panTro5). The main view area shows a genomic track with blue bars representing data points. The number '6' is visible in the track area. A large text overlay on the right side of the image reads: 'Save you session to avoid reloading your files'. Three numbered circles (1, 2, 3) are overlaid on the interface to highlight specific steps: 1 is on the 'Local File ...' option in the 'Session' menu, 2 is on the 'Local File ...' option in the 'Local File ...' dropdown, and 3 is on the 'Save ...' option in the 'Session' menu.

1

2

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/>