Atlas single cell PDAC

Journée des utilisateurs Gricad

Lucie Lamothe - Équipe Aptikal - LIG

06 Novembre 2025

Summary

1 - Introduction

2 - Création d’un atlas single cell

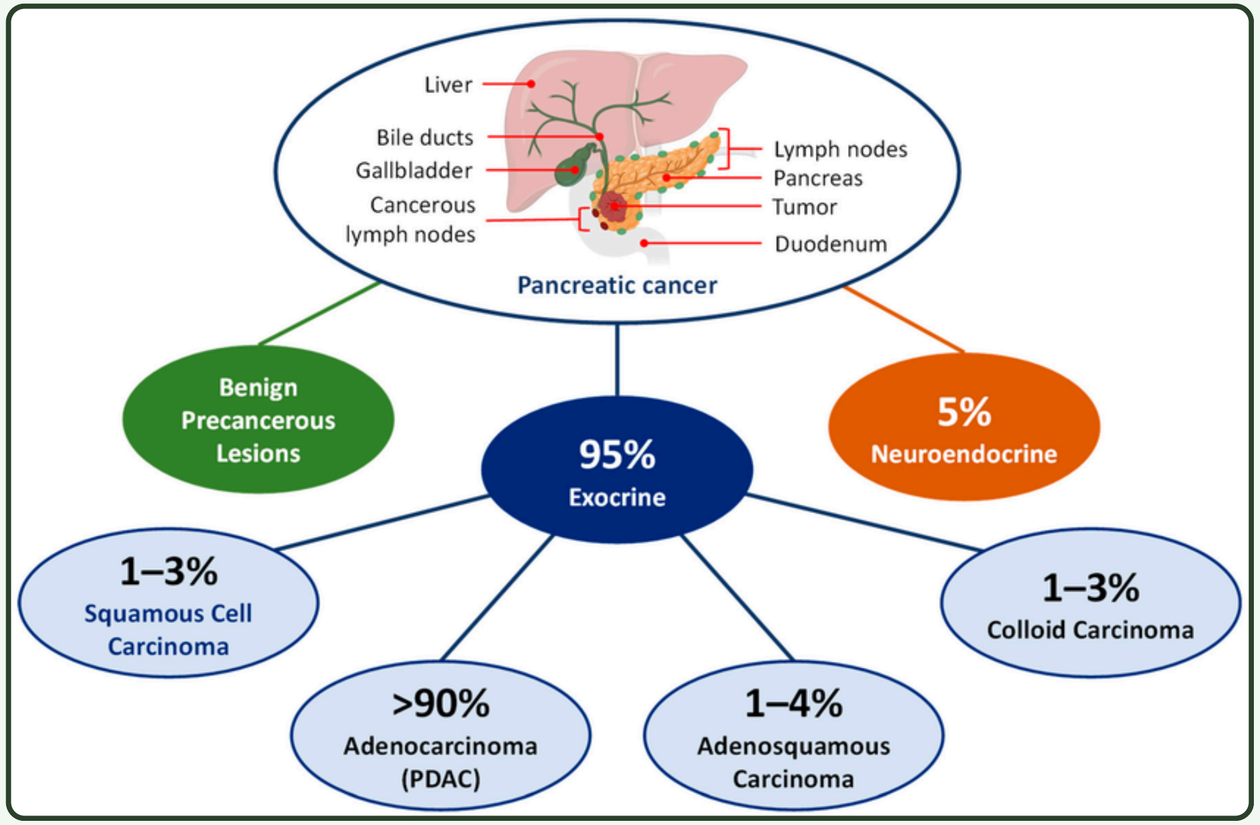
3 - Utilisation des ressources Gricad

Introduction

Sujet biologique & technologie

Cancer du pancréas

PDAC = pancreatic ductal adenocarcinoma



- Une des principales causes de mortalité liée au cancer
- Forte résistance au traitement
- Survie à 5 ans < 10 % (diagnostic tardif)

Classification of pancreatic cancer. PDAC - Pancreatic ductal adenocarcinom

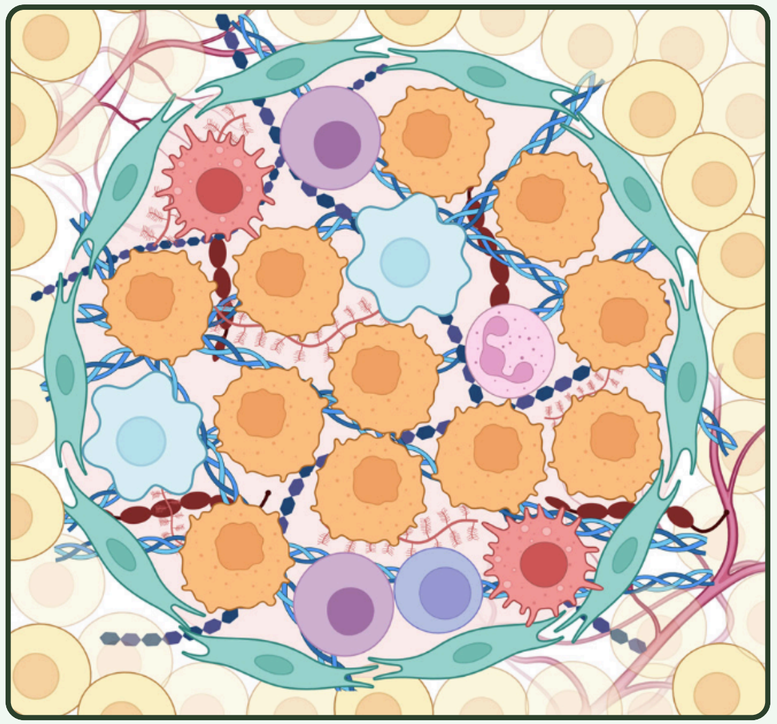
Xin Gu et al., Cancers 2024

Environment tumoral

Types cellulaires:

Cellules cancéreuses: avec différents sous-types (ex: classique, basal) et différent prognostiques.

Celluels saines: cellules pancréatiques.



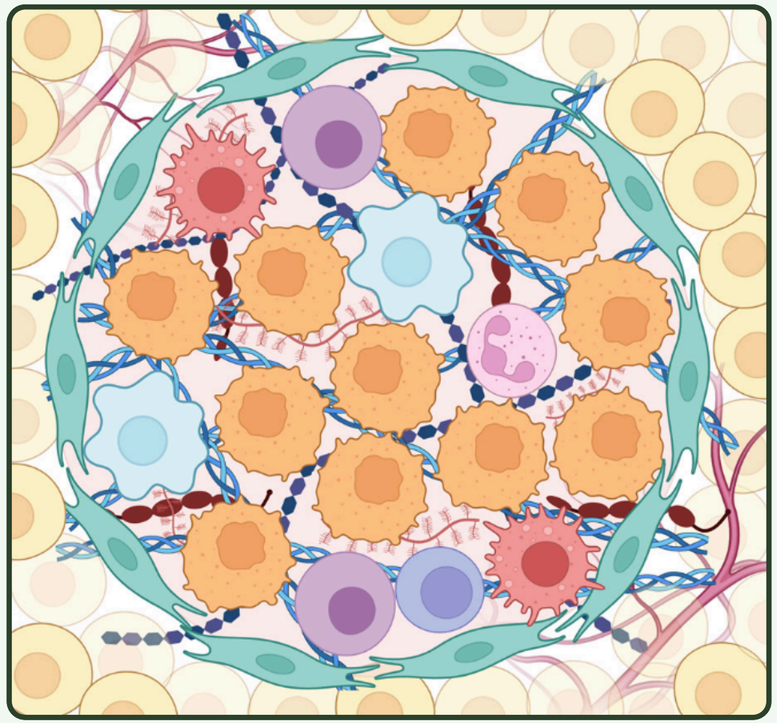
Impacts sur l'envrionnement tumoral :

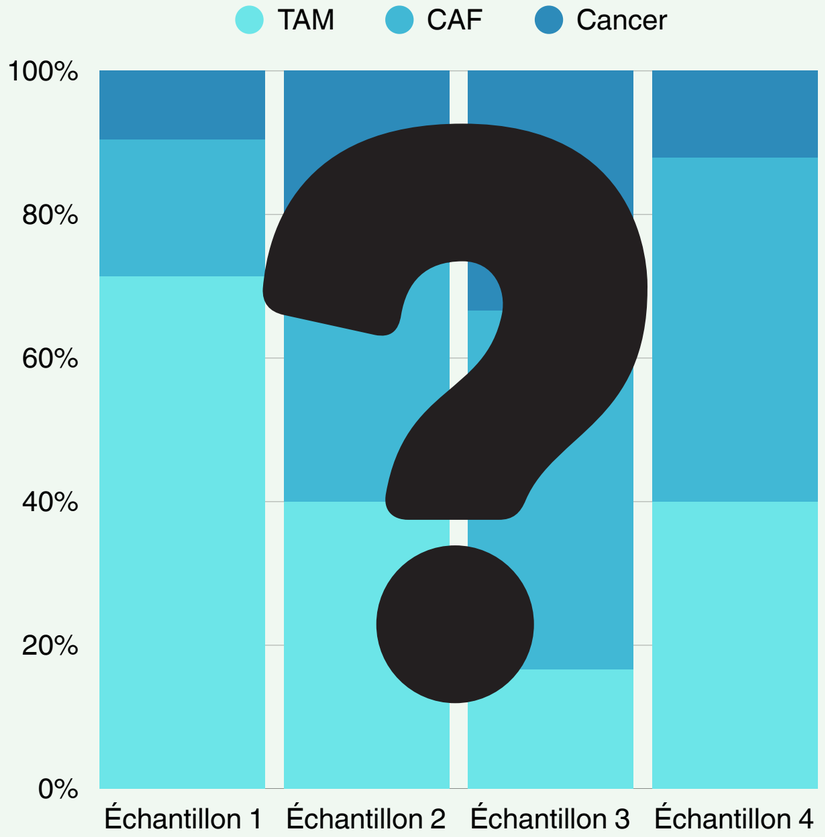
TAMs: type 1 (anti-tumoral) ou type 2 (effet immunosuppresseur et favorisant la prolifération des cellules cancéreuses).

CAFs: impliqués dans la résistance au traitements, densification de la tumeur et l'apparition de métastases.

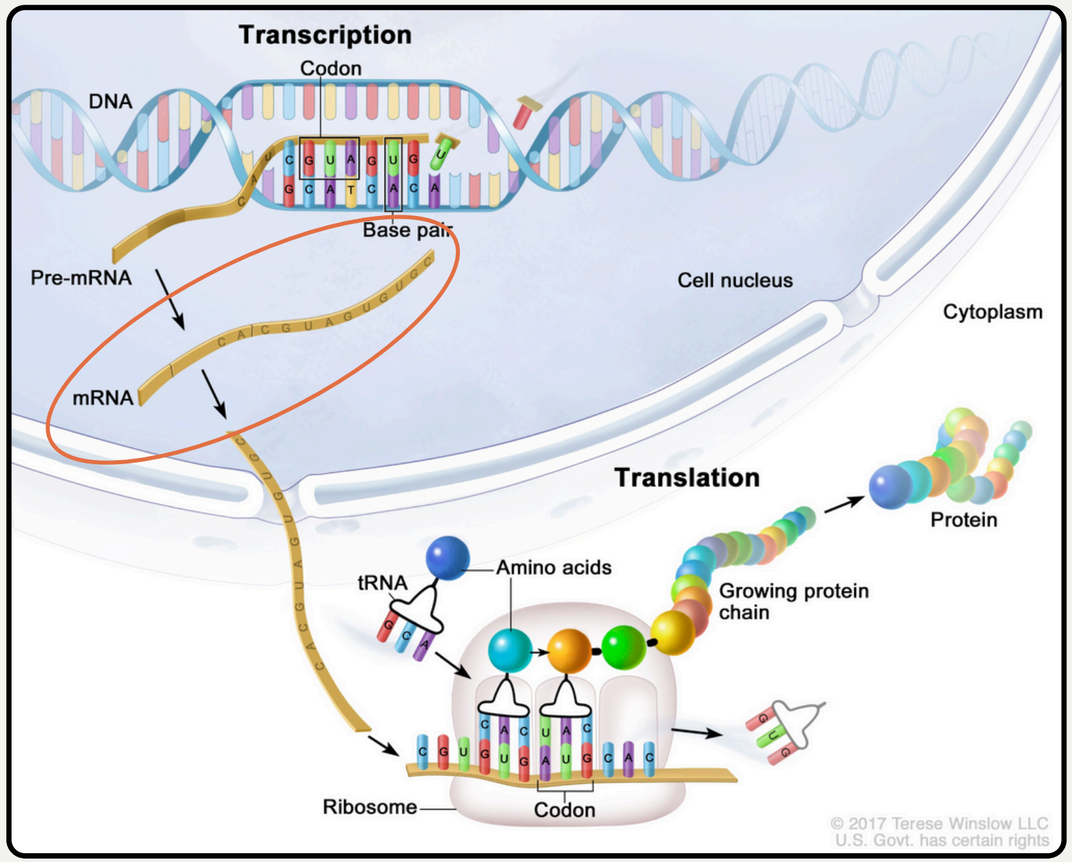
Schéma du micro-environnement du PDAC- Vitorakis, 2024

Environment tumoral





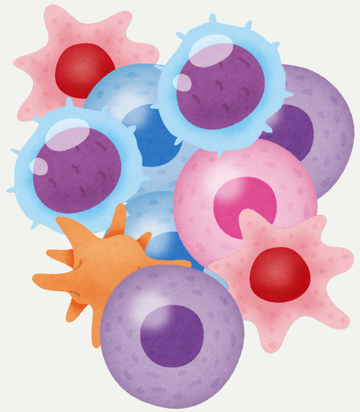
Sequençage ARN



- **ADN =** matériel génétique (identique pour toutes les cellules)
- **Protéines =** réalisant la plupart des fonctions biologiques
- **ARNm =** expression des gènes ; peut être considéré comme un proxy de l'identité cellulaire

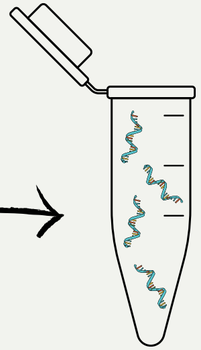
Séquençage bulk

Tissue samples

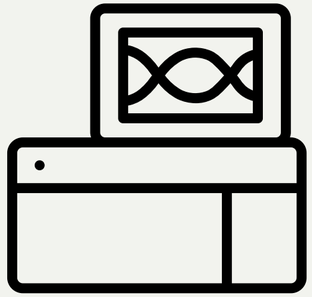


Mix

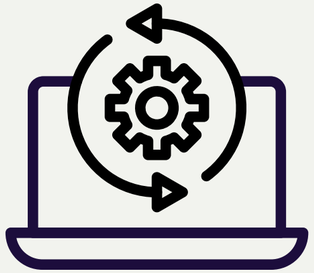
Extract mRNA

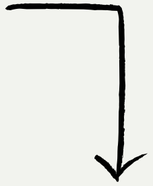


Sequence



Process raw files





```
(qiime2-2018.2) [yuuuhee@node01 0410_qiime2]$ head QY1-48.new.fastq
@QY38_1 M01056:42:000000000-A4R2M:1:1101:16586:1574
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCATGCAGGTGGTTTGTTAAGTCAGATGTGAAAGCCCTGGGCTCAACCTAGGAATCGCATTTGAAACTGACAAGCTAGAGTACTGTAGAGG
GTAGAATTTCAGGTGTAGCGGTGAAATGCGTAGAGATCTGAAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAGATACTGACACTCAGATGCGAAAGCGTGGGGAGCAAACAGG
+
HHHHGGGGGGGGHHHHGGGGGHHGGGGGHHHHHHHHGGGGHHHGGGGGHHHHHHGGGHGHHHHHHHHHHHHHHHHGHHHHHGF:B?GHHHHHHHHHHHHGGGGGHHHHHHGHHHHHHHHHHHHHHHHHHHHH
HHGHFHHHFGHHHHGGGGGHGHHHGGGGGFFFGFHHHFGHHGHGHGFFGG#GGGGGHGFFFFFGFGFFGFFFGFFHGHHHFHFFFBBBHEE/ACHGGEEACFEAHGHHFGHHFHFG
@QY12_1 M01056:42:000000000-A4R2M:1:1101:17149:1607
TACGTATGTCCCGAGCGTTATCCGGATTTATTGGGCGTAAAGCGAGCGCAGACGGTTGATTAAGTCTGATGTGAAAGCCCGGAGTTCAACTCCGGAATGGCATTGGAAACTGGTTAACTTGAGTGTTGTAGAGG
GTGGAACTCCATGTGTAGCGGTGGAATGCGTAGATATATGGAAGAACACCAGTGGCGAAGGCGGCTTACTGGACAACAACTGACGTTGAGGCTCGAAAGTGTGGGTAGCAAACAGG
+
GGGHHHHHHHHHHGGGGGGGEHHGGG?EEHHHHHHHFFGGGHHHGGGGGGGGFHCEFFFFGHDFHGHHHGGFHHGHEFFGHGGCGFHHHACFGGHGGGGHGHHHHHHHHFGHEBFGHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHGFGDDDHFEHGGGGGHHHHFGHGHGHHHGHGGGFFFHGGFE1HGGECGGF5BFG35HFHHHGGB5?GHHGHFGGEEEAAA5HHHGGGFFFFHGD3HHHHHFG
@QY38_2 M01056:42:000000000-A4R2M:1:1101:16704:1609
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCATGCAGGTGGTTTGTTAAGTCAGATGTGAAAGCCCTGGGCTCAACCTAGGAATCGCATTTGAAACTGACAAGCTAGAGTACTGTAGAGG
GTAGAATTTCAGGTGTAGCGGTGAAATGCGTAGAGATCTGAAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAGATACTGACACTCAGATGCGAAAGCGTGGGGAGCAAACAGG
```

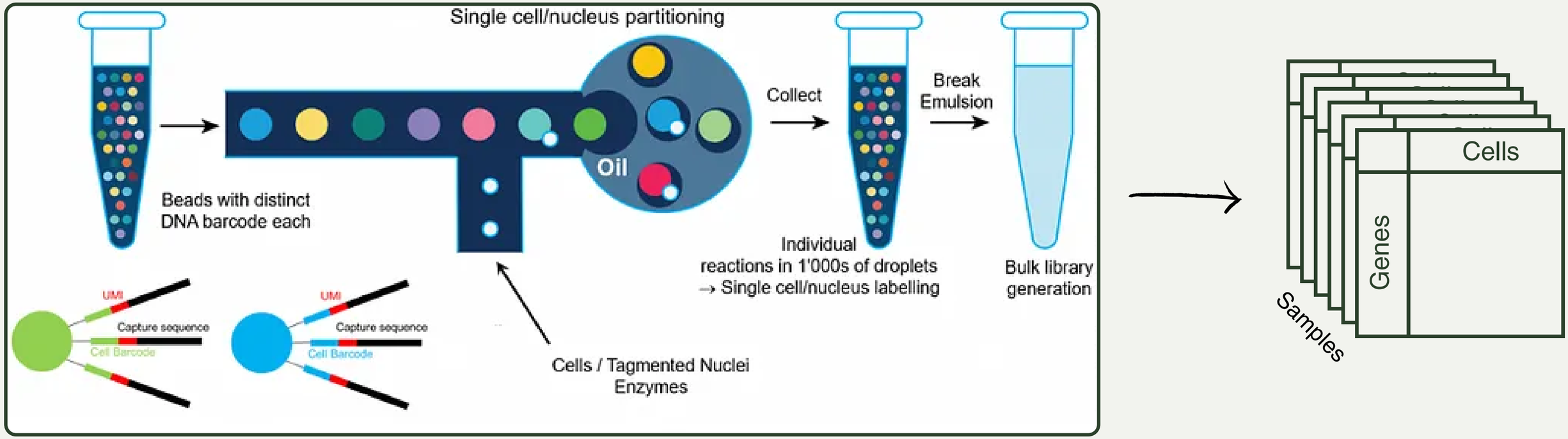
fastq sequence file

Align to reference genome

	Samples
Genes	

Expression matrix of mixed cells

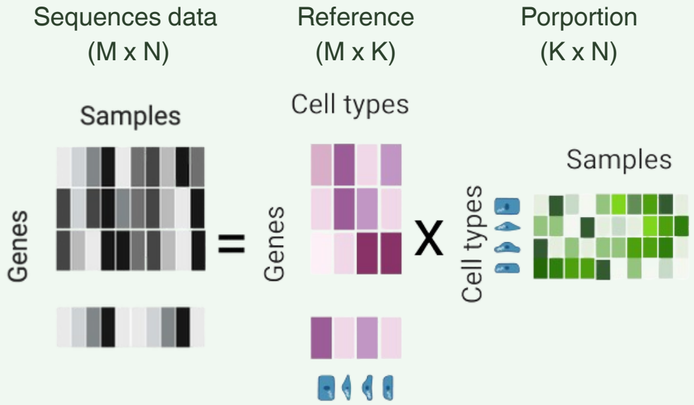
Séquençage single cell



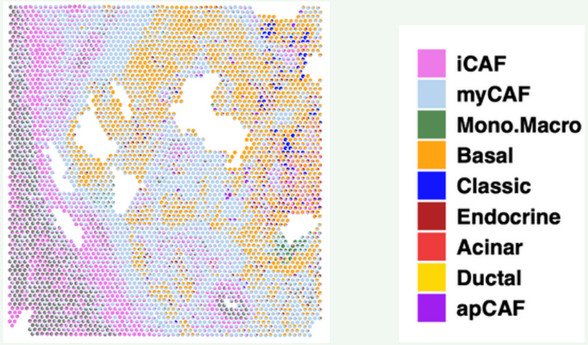
⇒ Nécessite de grosses ressources de puissance de calcul et de stockage

Quel besoin pour un Atlas?

- **Déconvolution données bulk :**
  Ce sont les données utilisées pour le diagnostique (prix, rapidité ...). Résultats plus performants avec une référence single-cell.[1]



- **Étude de données de spacial transcriptomique**



- **Découverte de nouveau markers**
- **Annotation de nouvelles données single cell**
- **Étude de nouvelles sous-population cellulaires.**

[1] Cobos, F.A. et al. Genome Biol, 2023

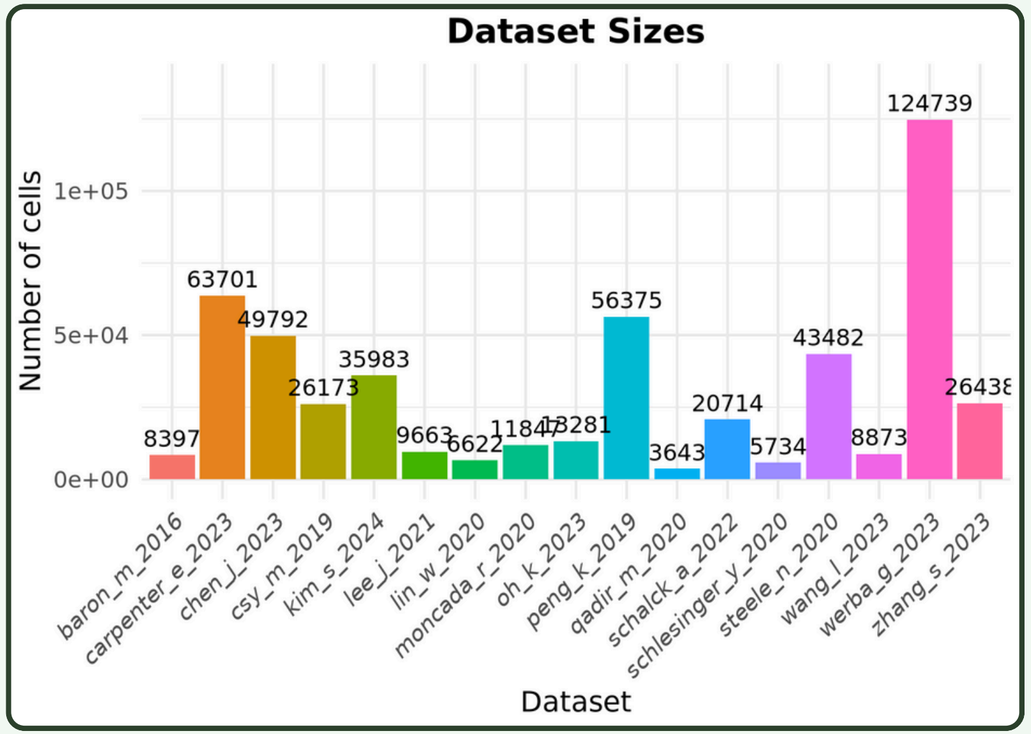
Création d’un atlas single cell

Intégration et annotation de datasets publics

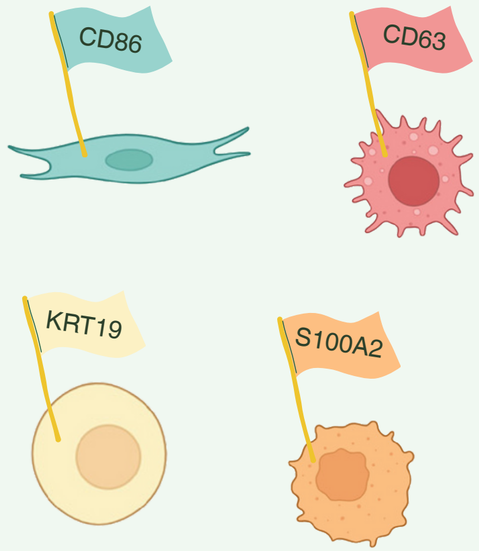
Données

18 datasets publiques

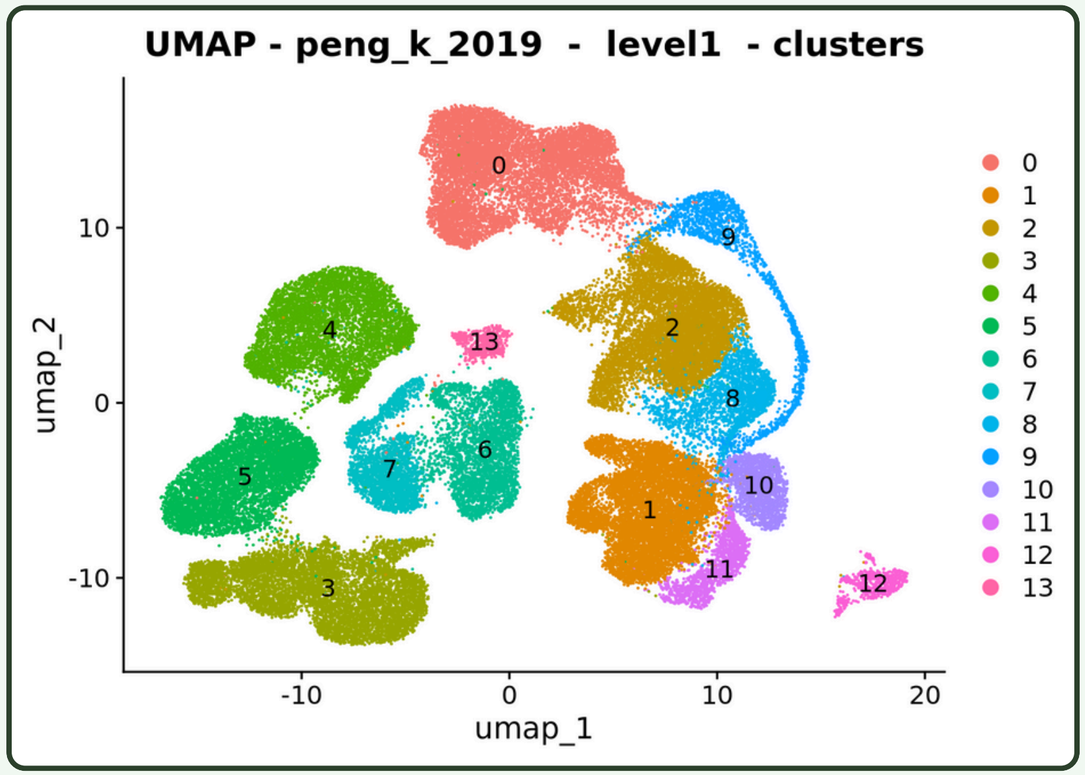
NCBI GEO Gene Expression Omnibus

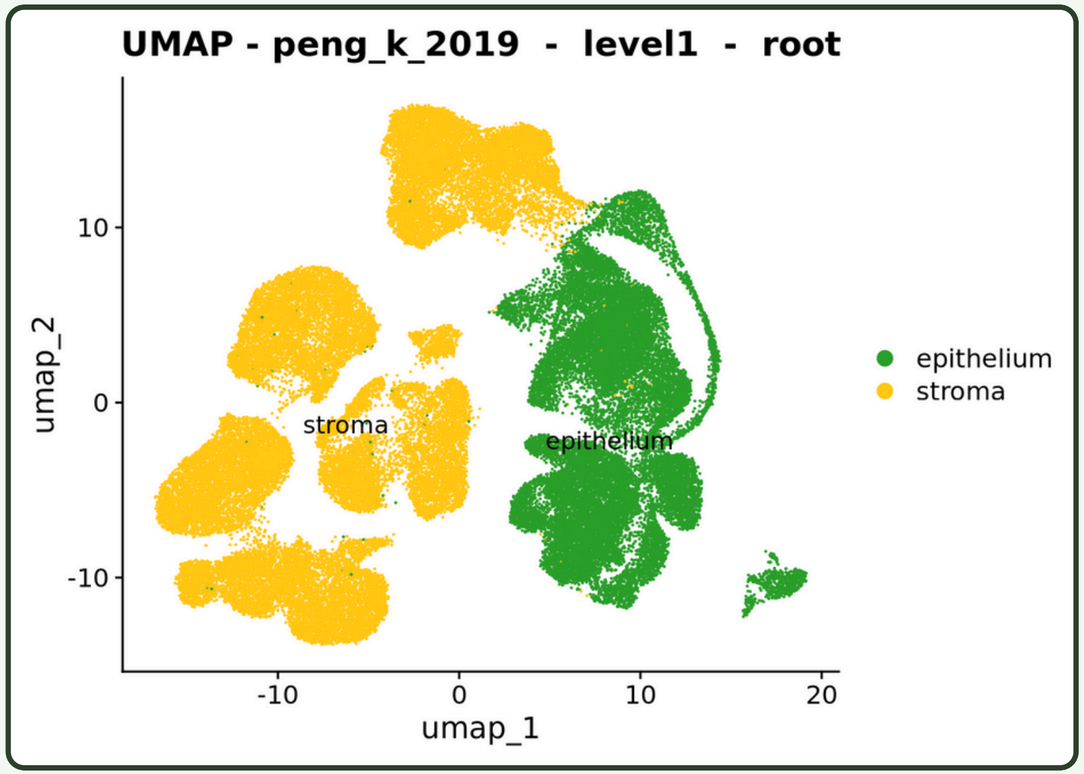


Liste de marqueurs de la littérature

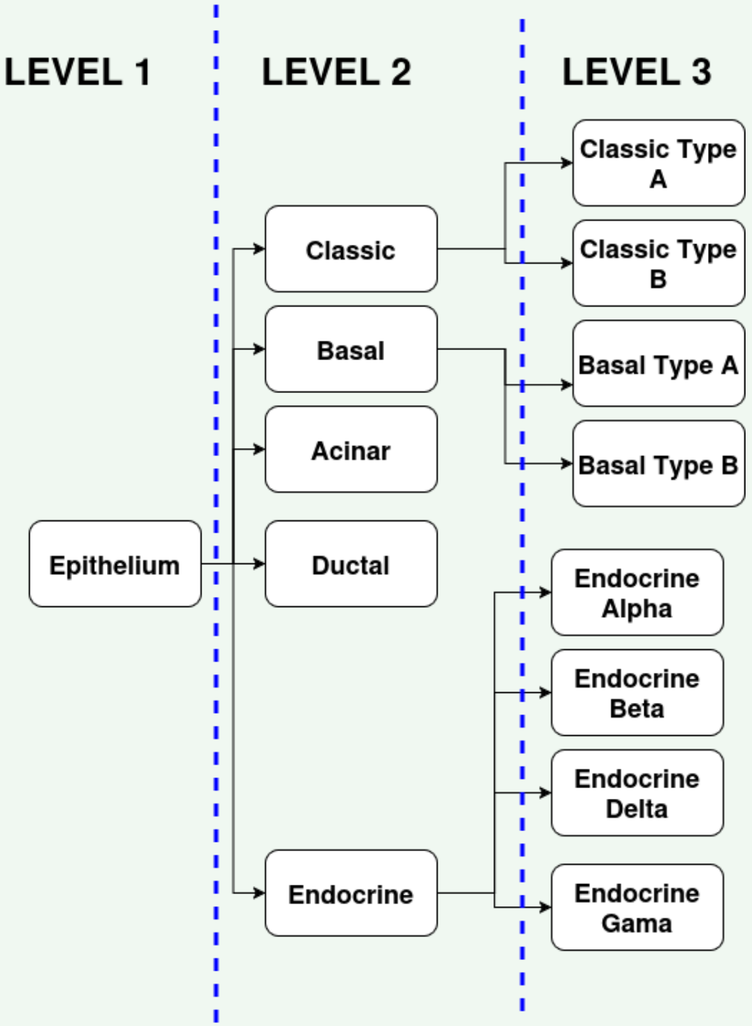


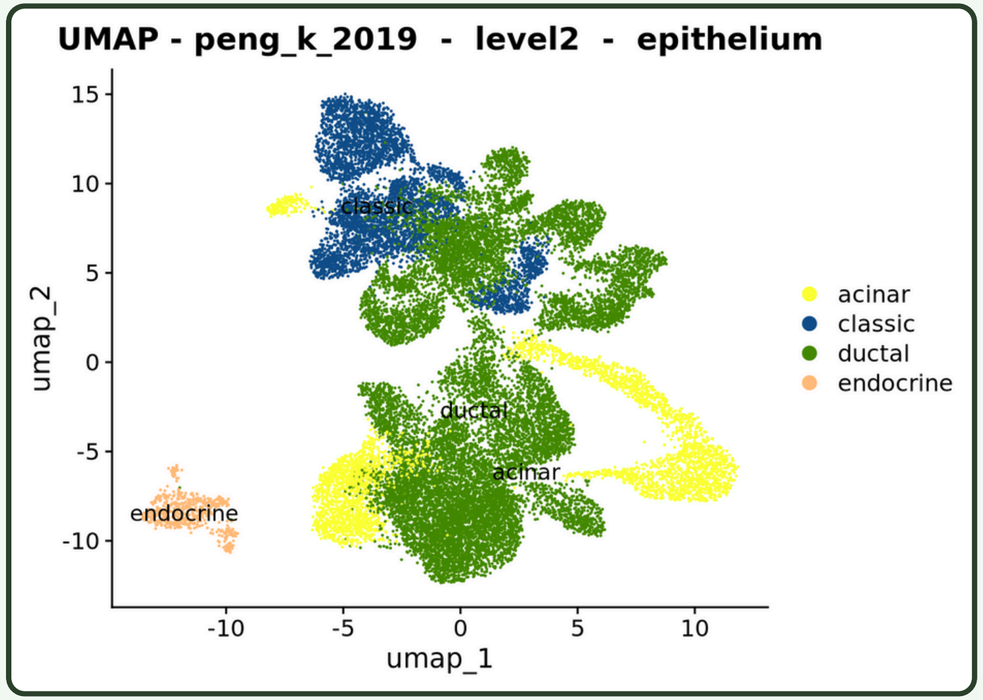
Identification cellulaire





Identification cellulaire

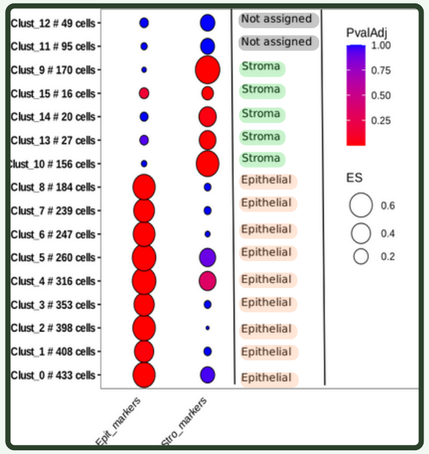




Identification cellulaire

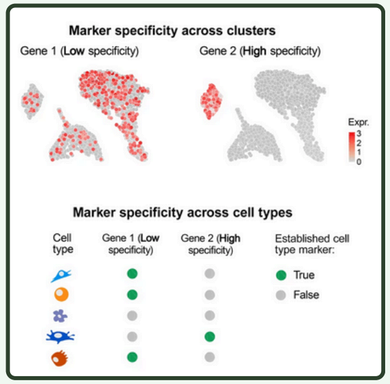
Home-made fgsea[1] based approach

statisticaly significant cell type enrichment score (ES) for each cluster



sctype[2]

Non statistical and marker based cell type score for each cluster

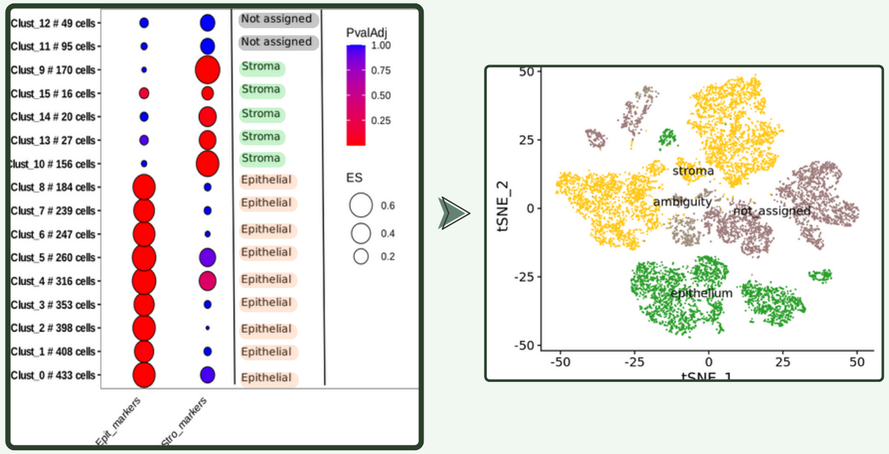


1. Gennady, K. (2021). bioRxiv, 060012.
2. Ianevski, A. (2022). Nature Communication, 14.

Identification cellulaire

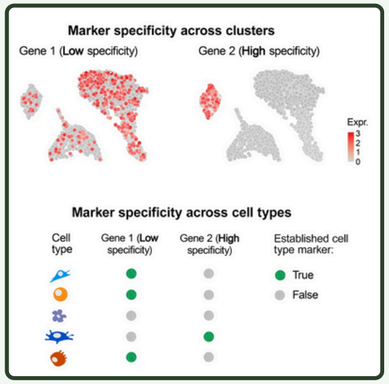
Home-made fgsea[1] based approach

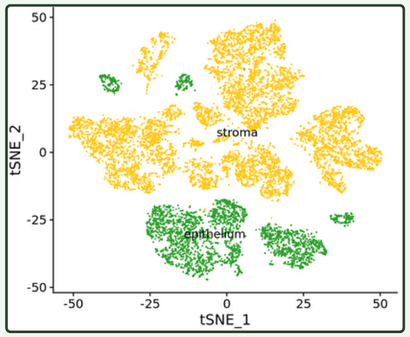
statisticaly significant cell type enrichment score (ES) for each cluster



sctype[2]

Non statistical and marker based cell type score for each cluster



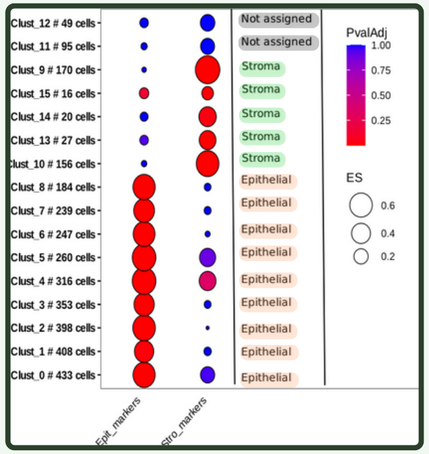


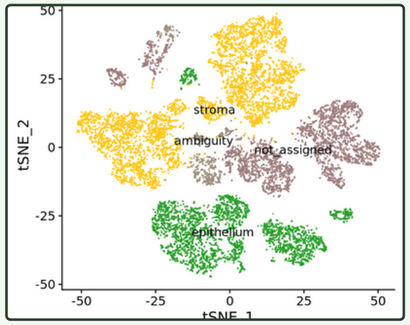
1. Gennady, K. (2021). bioRxiv, 060012.
2. Ianevski, A. (2022). Nature Communication, 14.

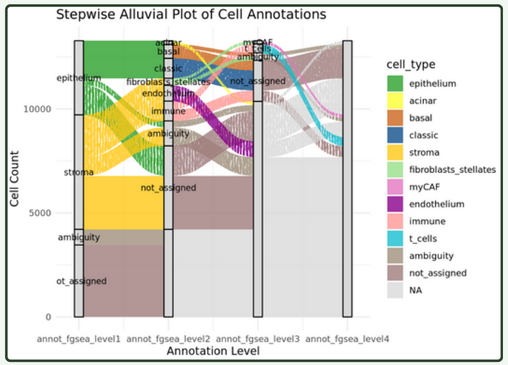
Identification cellulaire

Home-made fgsea[1] based approach

statisticaly significant cell type enrichment score (ES) for each cluster

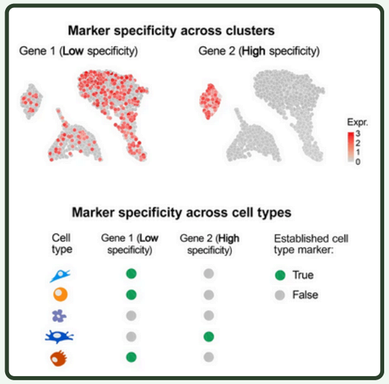


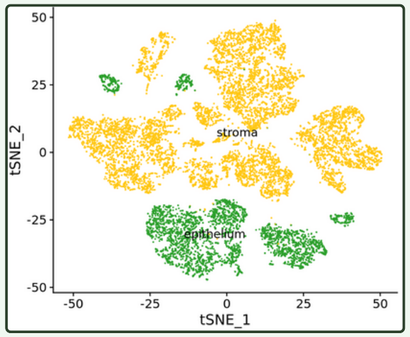


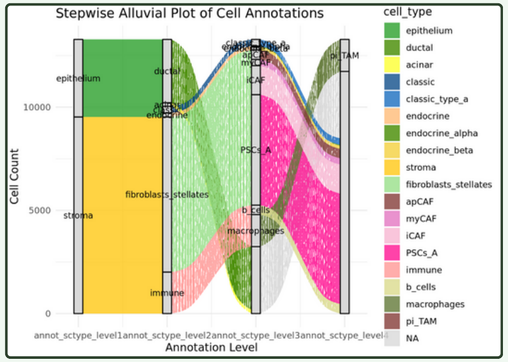


sctype[2]

Non statistical and marker based cell type score for each cluster





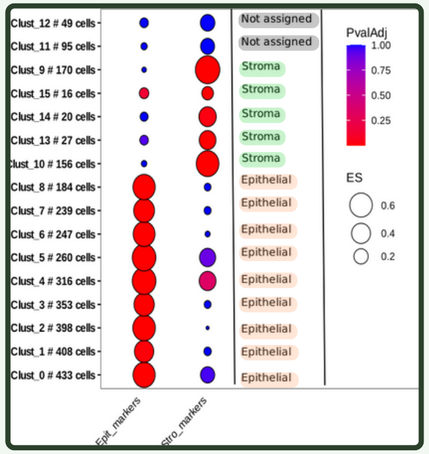


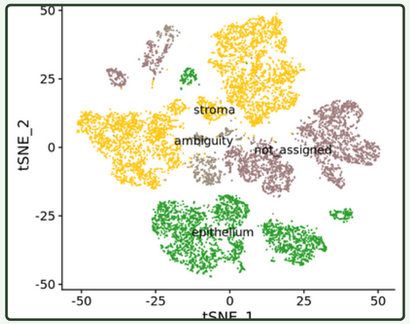
1. Gennady, K. (2021). bioRxiv, 060012.
2. Ianevski, A. (2022). Nature Communication, 14.

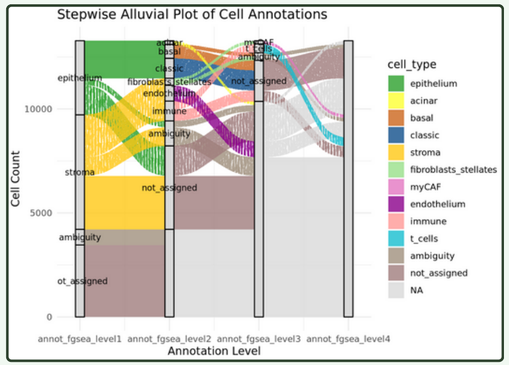
Identification cellulaire

Home-made fgsea[1] based approach

statisticaly significant cell type enrichment score (ES) for each cluster

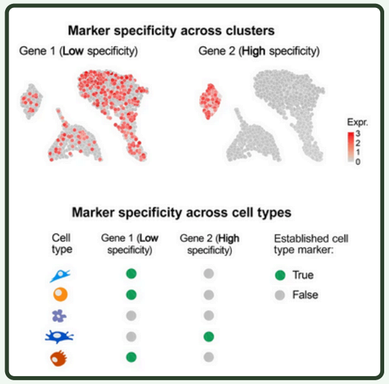


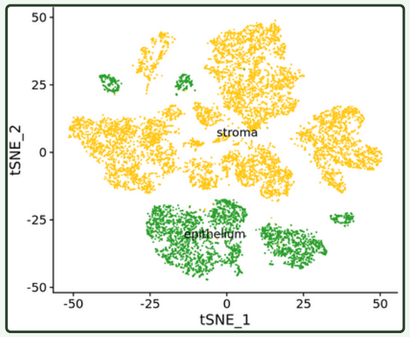


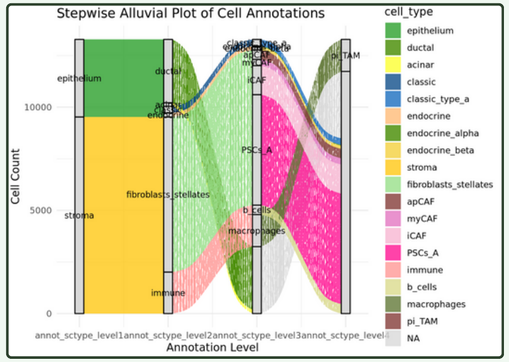


sctype[2]

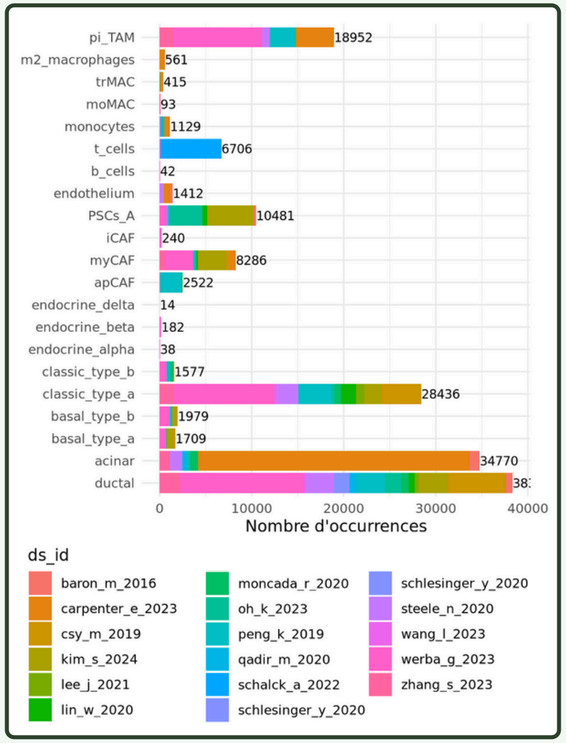
Non statistical and marker based cell type score for each cluster







Consensus annotation



⇨ Pour chaque dataset, une annotation consensus est obtenue en combinant les annotations de fgsea et sctype.

1. Gennady, K. (2021). bioRxiv, 060012.
2. Ianevski, A. (2022). Nature Communication, 14.

Cell type annotation

Une fois chaque dataset annotés,

⇒ Intégration des **18 datasets** dans un seul **atlas single cell robuste** pour tissue pdac.

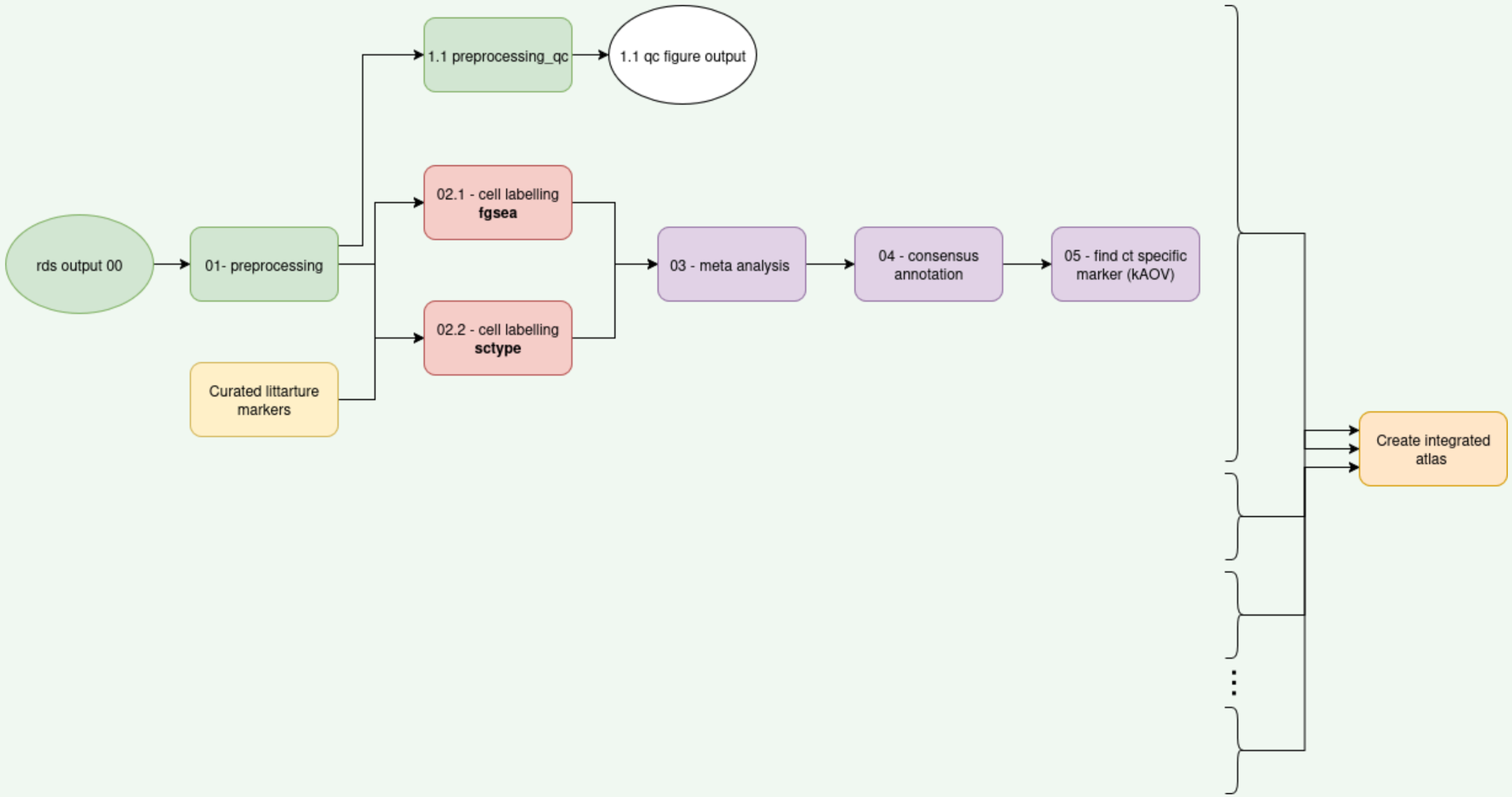
Il rassemble des cellules provenant principalement de tumeur primaire mais aussi de tissue metastatiques, adjacent et sain.

Notre Atlas contient **~ 200.000 cellules annotées.**

Utilisation des ressources Gricad

Implémentation d’un workflow snakemake

Diagramme workflow



Snakemake sur Dahu (bientôt kraken🐙)

Utilisation de snakemake pour lancer des jobs

```
snakemake -k --cluster "oarsub --project epimed -t fat -l /nodes=1/
core={resources.threads_fat},walltime={resources.walltime} -n {rule}_{jobid}" --latency-wait 60 --cores 30 --jobs 50 -s
Snakefile_Rmd.smk --jobname "{rule}_{jobid}"
```

Lancement dans un screen sur la frontale (oups🙈 j'ai appris récemment qu'il fallait pas)

Exemple de rule:

```
rule plot_per_ds:
    input:
        script = "07.2_plots_per_ds.Rmd",
        ds = BASE_OUTPUT + "/05_pool_output_" + TISSUE_TAG + "/05_" + ANNOT_TO_POOL + "_{ds}_" + TISSUE_TAG + ".rds"
    output:
        output_html = BASE_OUTPUT + "/07_plot_output_" + TISSUE_TAG + "/07.2_plot_" + ANNOT_TO_POOL + "_{ds}_" +
TISSUE_TAG + ".html"
    params:
        ds = "{ds}",
        output_dir = BASE_OUTPUT + "/07_plot_output_" + TISSUE_TAG + "/"
    resources:
        threads = 10,
        threads_fat = 1,
        walltime = "48:00:00"
    shell:
        """
        conda activate deconvpdac &&
        Rscript -e "rmarkdown::render('{input.script}', params = list(
            ds = '{input.ds}',
            ds_id = 'params.ds',
            output_html = '{output.output_html}'
        ), output_file = '{output.output_html}')"
        """
```

Snakemake sur Dahu (bientôt kraken🐙)

Utilisation de snakemake pour lancer des jobs

```
snakemake -k --cluster "oarsub --project epimed -t fat -l /nodes=1/
core={resources.threads_fat},walltime={resources.walltime} -n {rule}_{jobid}" --latency-wait 60 --cores 30 --jobs 50 -s
Snakefile_Rmd.smk --jobname "{rule}_{jobid}"
```

Lancement dans un screen sur la frontale (oups🙈 j'ai appris récemment qu'il fallait pas)

Exemple de rule:

```
rule plot_per_ds:
    input:
        script = "07.2_plots_per_ds.Rmd",
        ds = BASE_OUTPUT + "/05_pool_output_" + TISSUE_TAG + "/05_" + ANNOT_TO_POOL + "_{ds}_" + TISSUE_TAG + ".rds"
    output:
        output_html = BASE_OUTPUT + "/07_plot_output_" + TISSUE_TAG + "/07.2_plot_" + ANNOT_TO_POOL + "_{ds}_" +
TISSUE_TAG + ".html"
    params:
        ds = "{ds}",
        output_dir = BASE_OUTPUT + "/07_plot_output_" + TISSUE_TAG + "/"
    resources:
        threads = 10,
        threads_fat = 1,
        walltime = "48:00:00"
    shell:
        """
        conda activate deconvpdac &&
        Rscript -e "rmarkdown::render('{input.script}', params = list(
            ds = '{input.ds}',
            ds_id = 'params.ds',
            output_html = '{output.output_html}'
        ), output_file = '{output.output_html}')"
        """
```

Snakemake sur Dahu (bientôt kraken🐙)

Remarques:

- Sur les 11 rules besoin en RAM très hétérogènes ⇨ problèmes de dimensionnement récurrents.
- Utilisation de snakemake 7.22.0 - vieille version parce que recommandé par collègue.

L’équipe bio-informatique

Toute l’équipe se sert de ressources gricad:

- Utilisation uniquement en job interactif
- oarsub depuis bash
- Snakemake
- Kraken
- dahu workflow (snakemake)
- Bigfoot gpu

Merci pour votre attention!

Questions?