Paleofish Mitochondries

Objectif : Lancement de Eager sur référence majoritaire

28/08/2025

Point sur le travail réalisé avant mi juillet 2025 et liste des tâches à réaliser

Travail réalisé entre mai et mi-juillet 2025 :
* Lancement de FastQ Screen
* Comparaison des résultats avec BWA
* Ré-organisation des répertoires de travail Paleofish ShortGun afin de correspondre à l'organisation de Paleotrutta
* Nettoyage des fichiers afin de gagner de la place dans le work du projet CRUCIAL.
* Vérification/correction des barcodes. Ajout d'une étape win2dos
* Répartition des jeux de données par référence et préparation des scripts de lancement du pipeline pour chaque référence.
* Récupération des paramétrages spécifiques aux données ShortGun.
* Préparation des input.csv pour l'ensemble des jeux de données par référence.
* Pour tester, application du pipeline Eager modifié par Odile sur les 64 jeux de données classés dans NC_001960.1 , avec prise en compte des spécificités ShortGun.

A venir :
* Vérifier la liste des références, des jeux de données, et de leur répartition par référence.
* Lancer les scripts sur les jeux de données manquants c'est-à-dire hors les 64 jeux de données classés dans NC_001960.1

FastQ Screen – Hits on one genome

Lancement de FastQ Screen

```
/work/project/crucial/PALEOFISH/01_FastQScreen$ head out_L3_9_MB17_S4_R2_001_screen.txt
#Fastq_screen version: 0.15.3   #Aligner: bwa   #Reads in subset: 500000
Genome  #Reads_processed        #Unmapped       %Unmapped       #One_hit_one_genome     %One_hit_one_genome
_multiple_genomes       %One_hit_multiple_genomes       Multiple_hits_multiple_genomes  %Multiple_hits_multip
H.sapiens       531042  523809  98.64   9       0.00    12      0.00    2852    0.54    4360    0.82
C.lupus.familiaris      531042  523880  98.65   0       0.00    3       0.00    456     0.09    6703    1.26
O.mykiss        531042  514216  96.83   24      0.00    13      0.00    8100    1.53    8689    1.64
H.hucho 531042  510185  96.08   43      0.01    22      0.00    12714   2.39    8078    1.52
B.bubo  531042  524971  98.86   1       0.00    0       0.00    4359    0.82    1711    0.32
S.glanis        531042  506792  95.44   24      0.00    0       0.00    23741   4.47    485     0.09
B.scandiacus    531042  527529  99.34   1       0.00    0       0.00    2140    0.40    1372    0.26
L.idus  531042  510524  96.14   0       0.00    0       0.00    114     0.02    20404   3.84
```

Datamash sur les fichiers txt : sum #One_hit_one_genome + #Multiple_hits_one_genome

/work/project/crucial/PALEOFISH/01_FastQScreen$ tail -n +3 out_L3_9_MB17_S4_R2_001_screen.txt | awk '{print $1, $5 + $7}' | sort -k2,2nr | head -n 1
S.trutta 48065

Synthèse des résultats :

FastQ Screen – Hits on one genome : Comparaison FastQ Screen et BWA

Alignement sur référence majoritaire :
https://web-genobioinfo.toulouse.inrae.fr/~smaman/BOOTSTRAP/projet_crucial/align_stats.html

Mêmes résultats avec FastQ Screen et BWA sur multifasta

FastQ Screen – % Unmapped

/work/project/crucial/PALEOFISH/01_FastQScreen/analyze_fastqscreen.sh
→ mêmes résultats sauf pour les demultiplexed.


```
/work/project/crucial/PALEOFISH/01_FastQScreen$ more analyze_fastqscreen.sh
#!/bin/bash

# Fichier de sortie CSV avec en-tête
echo "Fichier,Min_%Unmapped,Espèce_Min_Unmapped,Max_Alignments,Espèce_Max_Alignments" > mapping_summary.csv

# Boucle sur les fichiers *_screen.txt
for f in *R1*_screen.txt; do
  # Nom sans extension
  nom=$(basename "$f" _screen.txt)

  # Nettoyer les espaces -> tabulations
  content=$(tail -n +3 "$f" | tr -s ' ' | tr ' ' '\t')

  # Filtrer les lignes avec %Unmapped > 0
  min_line=$(echo "$content" | awk -F'\t' '$4 > 0' | sort -t $'\t' -k4,4n | head -n1)
  espece_min=$(echo "$min_line" | cut -f1)
  min_unmapped=$(echo "$min_line" | cut -f4)

  # Identifier la ligne avec max($5 + $7)
  max_line=$(echo "$content" | awk -F'\t' '{sum=$5+$7; print $1 "\t" sum}' | sort -k2,2nr | head -n1)
  espece_max=$(echo "$max_line" | cut -f1)
  max_align=$(echo "$max_line" | cut -f2)

  # Ajouter au CSV
  echo "${nom},${min_unmapped},${espece_min},${max_align},${espece_max}" >> mapping_summary.csv
done

echo "✅ Résultat enregistré dans mapping_summary.csv"
```

FastQ Screen – % Unmapped vs %Mapped

Excellente question, S ! Les deux indicateurs renseignent des choses différentes, donc **le meilleur dépend de ce que tu cherches à évaluer**. Voici un petit comparatif clair :

🧪 **%Unmapped le plus bas**

- **Ce que ça mesure** : la proportion de séquences qui **n'ont pas pu être alignées** à un génome de référence.
- **Interprétation** : plus ce pourcentage est faible, plus il y a eu d'alignements — ce qui peut signaler **une proximité génétique** ou **une contamination biologique**.
- **Limite** : ce % ne distingue pas les bons alignements (exacts) des alignements partiels ou multiples.

🧬 **Somme $5 + $7**

- **Ce que ça mesure** : le **nombre total de reads mappés de façon unique** à une seule espèce (soit en one-hit, soit avec ambiguïté partielle sur un seul génome).
- **Interprétation** : c'est une mesure plus robuste du **signal spécifique pour une espèce donnée**, et donc plus pertinente pour identifier une **présence réelle dans l'échantillon**.
- **Limite** : peut sous-estimer l'abondance si les reads se distribuent entre plusieurs espèces proches.

🎯 **Conclusion pratique**

- Pour détecter **une espèce présente spécifiquement et majoritairement** dans un échantillon

👉 **préférer `$5 + $7` (max)**. - Pour évaluer si **l'échantillon contient beaucoup de reads non assignés** (bruit, contamination, mauvaise qualité)

👉 **regarder `%Unmapped`**.

🔍 **Croiser les deux** est idéal. Par exemple, une espèce avec : - un **%Unmapped bas** mais **peu de one-hit** peut refléter un alignement peu spécifique (ou une contamination générale), - alors qu'un **$5+$7 très élevé avec un Unmapped moyen** pointe vers un signal clair (présence vraie).

Barcodes et noms des fichiers

1/ Méthode d'extraction des barcodes : https://forge-dga.jouy.inra.fr/projects/paleofish_mito/wiki/Extraction_des_barcodes

Les 4 couples de tags les plus trouvés correspondent bien à ceux figurant dans le tableau excel pour les non-démultiplexés : GACGATT TCGCAGG, AACCTGC, CTCTGCA, GCCTACG GGATCAA, GTCCGGC CTCGATG

Et https://forge-dga.jouy.inra.fr/projects/paleofish_mito/wiki/Fichiers_barcodes_fw_et_rv

```
>barcode1
GACGATT
>barcode2
GTCCGGC
>barcode3
GCCTACG
>barcode4
AACCTGC
```

2/ Source d'informations sur les barcodes : https://web-genobioinfo.toulouse.inrae.fr/~smaman/BOOTSTRAP/projet_crucial/mapDammage.html Login : CRUCIAL - Pass : sigenae

```
zgrep -A 3 '1:N:0:GACGATT+TCGCAGG' Undetermined_S0_R1_001.fastq.gz > L1_1_AUD20212-2_Sxx_R1_001.fastq
zgrep -A 3 '2:N:0:GACGATT+TCGCAGG' Undetermined_S0_R2_001.fastq.gz > L1_1_AUD20212-2_Sxx_R2_001.fastq

zgrep -A 3 '1:N:0:AACCTGC+CTCTGCA' Undetermined_S0_R1_001.fastq.gz > L1_2_BC4_Sxx_R1_001.fastq
zgrep -A 3 '2:N:0:AACCTGC+CTCTGCA' Undetermined_S0_R2_001.fastq.gz > L1_2_BC4_Sxx_R2_001.fastq

zgrep -A 3 '1:N:0:GCCTACG+GGATCAA' Undetermined_S0_R1_001.fastq.gz > L1_4_AUD11764-47_Sxx_R1_001.fastq
zgrep -A 3 '2:N:0:GCCTACG+GGATCAA' Undetermined_S0_R2_001.fastq.gz > L1_4_AUD11764-47_Sxx_R2_001.fastq

zgrep -A 3 '1:N:0:GTCCGGC+CTCGATG' Undetermined_S0_R1_001.fastq.gz > L1_7_BC2_Sxx_R1_001.fastq
zgrep -A 3 '2:N:0:GTCCGGC+CTCGATG' Undetermined_S0_R2_001.fastq.gz > L1_7_BC2_Sxx_R2_001.fastq
```

3/ Si OK, renommage des fichiers.

Barcodes et noms des fichiers

```
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ ls
demultiplexed-barcode1_R1.fastq.gz   L10_4_BC8_S61_R1_
demultiplexed-barcode1_R2.fastq.gz   L10_4_BC8_S61_R2_
demultiplexed-barcode2_R1.fastq.gz   L10_5_BC9_S62_R1_
demultiplexed-barcode2_R2.fastq.gz   L10_5_BC9_S62_R2_
demultiplexed-barcode3_R1.fastq.gz   L11_1_SCO_98_S66_
demultiplexed-barcode3_R2.fastq.gz   L11_1_SCO_98_S66_
demultiplexed-barcode4_R1.fastq.gz   L11_2_SCO_81_S67_
demultiplexed-barcode4_R2.fastq.gz   L11_2_SCO_81_S67_
demultiplexed-unknown_R1.fastq.gz    L11_3_SCO_114_S68_
demultiplexed-unknown_R2.fastq.gz    L11_3_SCO_114_S68_
```

Renommage des fichiers :

```
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode1_R1.fastq.gz L1_1_AUD20212-2_Sxx_R1.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode1_R2.fastq.gz L1_1_AUD20212-2_Sxx_R2.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode2_R1.fastq.gz L1_7_BC2_Sxx_R1.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode2_R2.fastq.gz L1_7_BC2_Sxx_R2.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode3_R1.fastq.gz L1_4_AUD11764-47_Sxx_R1.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode3_R2.fastq.gz L1_4_AUD11764-47_Sxx_R2.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode4_R1.fastq.gz L1_2_BC4_Sxx_R1_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode4_R2.fastq.gz L1_2_BC4_Sxx_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode4_R1.fastq.gz L1_2_BC4_Sxx_R1_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv demultiplexed-barcode4_R2.fastq.gz L1_2_BC4_Sxx_R2_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv L1_1_AUD20212-2_Sxx_R1.fastq.gz L1_1_AUD20212-2_Sxx_R1_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv L1_1_AUD20212-2_Sxx_R2.fastq.gz L1_1_AUD20212-2_Sxx_R2_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv L1_7_BC2_Sxx_R1.fastq.gz L1_7_BC2_Sxx_R1_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv L1_7_BC2_Sxx_R2.fastq.gz L1_7_BC2_Sxx_R2_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv L1_4_AUD11764-47_Sxx_R1.fastq.gz L1_4_AUD11764-47_Sxx_R1_001.fastq.gz
/work/project/crucial/PALEOFISH/DATA/capture_jan22$ mv L1_4_AUD11764-47_Sxx_R2.fastq.gz L1_4_AUD11764-47_Sxx_R2_001.fastq.gz
```

Ré-Organisation de l'espace de travail

(mai 2005)

Ré-organisation du dossier PALEOFISH :

```
/work/project/crucial$ ls
MODERN   PALEOFISH   PALEOTROTA   programs   Ref_genomes
```

```
/work/project/crucial/PALEOFISH$ ls -ltrah
total 100K
-rw-r--r--   1 smaman CRUCIAL  111 Feb 13  2024 README
drwxr-sr-x  15 smaman CRUCIAL 4.0K Jan 16  2025 Sarah
drwxr-sr-x   5 jchat  CRUCIAL 4.0K Apr 11 17:33 Joelle
drwxrwsr-x   3 jchat  CRUCIAL  16K Apr 29 11:55 FASTQ_MITO
drwxrwsrwx   9 smaman CRUCIAL 4.0K May  6 11:02 .
drwxr-sr-x   4 jchat  CRUCIAL 4.0K May  6 14:01 DATA
drwxr-sr-x   4 jchat  CRUCIAL  16K May 12 10:17 FASTQ_shotgun
drwxr-sr-x  14 smaman CRUCIAL 4.0K May 19 10:45 02_eager
drwxrws---+  7 jchat  CRUCIAL 4.0K May 27 13:51 ..
drwxr-sr-x   2 smaman CRUCIAL  64K Jun 27 09:38 01_FastQScreen
```

Traitements FastQScreen :

```
/work/project/crucial/PALEOFISH$ ls 01_FastQScreen/
analyze_fastqscreen.sh                  out_L11_3_SCO_114_S68_R1_001_screen.png
fastq_screen.conf                       out_L11_3_SCO_114_S68_R1_001_screen.txt
fastq_screen.sh                         out_L11_3_SCO_114_S68_R2_001_screen.html
mapping_summary.csv                     out_L11_3_SCO_114_S68_R2_001_screen.png
```

Traitements Eager :

```
/work/project/crucial/PALEOFISH/02_eager$
```

```
GCF_018398675  GCF_901001165  NC_001960.1  NC_006531.1  NC_009263.1  NC_012928.1  NC_018341.1  NC_020762.1  NC_024032.1  scripts-archives
```

```
/work/project/crucial/PALEOFISH/02_eager$ ls scripts-archives/
01_script_eager_part1_REF.sh  02_rehead_bam_files_v2.sh  03_script_eager_only_genotyping_v2.sh  bam_tsv.sh  tsv_modif.sh
```

Pipeline et paramétrages ShortGun

Phase 1 : Premier lancement d'Eager
tsv_modif.sh : Génération du fichier d'entrée input.csv listant les FastQ en entrée d'Eager.
01_script_eager_part1_NC_001960.1.sh

Phase 2 : Filtre des BAM pour ne conserver que les séquences mito
02_rehead_bam_files_v2_NC_001960.1.sh : Enlever les séquences génomiques du BAM pour ne conserver que les séquences mito. Sinon fichier FASTQ trop gros qui fait planter le script 3.
02bis-bam_tsv.sh : Créer le fichier tabulé listant les BAM, en entrée de la phase 2 d'Eager.

Phase 3 : Second lancement d'Eager uniquement sur les mitochondries
03_script_eager_only_genotyping_v2.sh : Skip des étapes précédentes avec les étapes GATK et VCF.

Validation du pipeline et des paramétrages
Spécificités données shortgun

Ces informations ont été transmises en réunion avant mai 2025 :

* Pas de traitement UDG (udg_type none) car mapdammage a mis en évidence une erreur de désincorporation des bases plus grande, en début et en fin de séquence.

* Phase 2 d'Eager sur les séquences mito uniquement, ajouter l'option rescale (--run_mapdamage_rescaling true) permet de modifier la qualité de la base avec mapdammage selon la probabilité qu'elle soit endommagée.

* Données MySeq => --colour_chemistry 2

Validation de la liste des jeux de données

46 données Shortgun ? A valider avec Joelle.

Fichiers tsv descriptif des jeux de données

(mai 2025)

* Un fichier tsv par référence
* Les fichiers tsv ont été préparés en mai 2025.

```
/work/project/crucial/PALEOFISH/02_eager/NC_001960.1$ ls -ltrah input_*
-rw-r--r-- 1 smaman CRUCIAL 6.7K May 19 11:51 input_bam.tsv
-rw-r--r-- 1 smaman CRUCIAL 6.6K May 22 12:26 input_NC_001960.1.tsv
```

Reste à vérifier la répartition des jeux de données par références avant d'utiliser les fichiers inputs pour le pipeline Eager.

```
/work/project/crucial/PALEOFISH/02_eager$ more NC_001960.1/input_NC_001960.1.tsv
Sample_Name     Library_ID      Lane    Colour_Chemistry        SeqType Organism        Strandedness    UDG_Treatment   R1      R2      BAM
BC11    S38     5       2       PE      salmo_salar     double  none    /work/project/crucial/PALEOFISH/DATA/capture_jan22/L5_13_BC11_S38_R1_001.fastq.gz       /work/project/crucial/PA
LEOFISH/DATA/capture_jan22/L5_13_BC11_S38_R2_001.fastq.gz       NA
EG3     S73     2       2       PE      salmo_salar     double  none    /work/project/crucial/PALEOFISH/DATA/capture_jan22/L2_3_EG3_S73_R1_001.fastq.gz /work/project/crucial/PALEOFISH/
DATA/capture_jan22/L2_3_EG3_S73_R2_001.fastq.gz NA
```

Regroupement des jeux de données par référence

GCF_018398675 et GCF_901001165 vides : TODO ?

```
/work/project/crucial/PALEOFISH/02_eager/NC_001960.1$ ls *gz
L0_2_EG5_S82_R1_001.fastq.gz          L2_12_MPHB4_S80_R2_001.fastq.gz  L4_5_MD5_S15_R1_001.fastq.gz    L5_15_BC13_S40_R2_001.fastq.gz  L5_7_MD3_S32_R1_001.fastq.gz
L0_2_EG5_S82_R2_001.fastq.gz          L2_1_EG1_S71_R1_001.fastq.gz     L4_5_MD5_S15_R2_001.fastq.gz    L5_1_MD15_S26_R1_001.fastq.gz   L5_7_MD3_S32_R2_001.fastq.gz
L0_8_SH2_S88_R1_001.fastq.gz          L2_1_EG1_S71_R2_001.fastq.gz     L4_8_MB4_S18_R1_001.fastq.gz    L5_1_MD15_S26_R2_001.fastq.gz   L5_8_MD16_S33_R1_001.fastq.gz
L0_8_SH2_S88_R2_001.fastq.gz          L2_2_EG2_S72_R1_001.fastq.gz     L4_8_MB4_S18_R2_001.fastq.gz    L5_2_MD7_S27_R1_001.fastq.gz    L5_8_MD16_S33_R2_001.fastq.gz
L10_11_SCO_1338_S63_R1_001.fastq.gz   L2_2_EG2_S72_R2_001.fastq.gz     L4_9_MD12_S19_R1_001.fastq.gz   L5_2_MD7_S27_R2_001.fastq.gz    L8_15_HTMK99-XXII-0-4287_S49_R1_001.fastq.gz
L10_11_SCO_1338_S63_R2_001.fastq.gz   L2_3_EG3_S73_R1_001.fastq.gz     L4_9_MD12_S19_R2_001.fastq.gz   L5_3_MD8_S28_R1_001.fastq.gz    L8_15_HTMK99-XXII-0-4287_S49_R2_001.fastq.gz
L10_13_SCO_2764_S65_R1_001.fastq.gz   L2_3_EG3_S73_R2_001.fastq.gz     L5_12_MD9_S37_R1_001.fastq.gz   L5_3_MD8_S28_R2_001.fastq.gz    L9_1_26-52_S50_R1_001.fastq.gz
L10_13_SCO_2764_S65_R2_001.fastq.gz   L4_11_MD2_S21_R1_001.fastq.gz    L5_12_MD9_S37_R2_001.fastq.gz   L5_4_MD4_S29_R1_001.fastq.gz    L9_1_26-52_S50_R2_001.fastq.gz
L2_10_MPHB2_S78_R1_001.fastq.gz       L4_11_MD2_S21_R2_001.fastq.gz    L5_13_BC11_S38_R1_001.fastq.gz  L5_4_MD4_S29_R2_001.fastq.gz    L9_2_2777-17_S51_R1_001.fastq.gz
L2_10_MPHB2_S78_R2_001.fastq.gz       L4_13_MB5_S23_R1_001.fastq.gz    L5_13_BC11_S38_R2_001.fastq.gz  L5_5_MD10_S30_R1_001.fastq.gz   L9_2_2777-17_S51_R2_001.fastq.gz
L2_11_MPHB3_S79_R1_001.fastq.gz       L4_13_MB5_S23_R2_001.fastq.gz    L5_14_BC12_S39_R1_001.fastq.gz  L5_5_MD10_S30_R2_001.fastq.gz   L9_4_SM1_S52_R1_001.fastq.gz
L2_11_MPHB3_S79_R2_001.fastq.gz       L4_4_MB9_S14_R1_001.fastq.gz     L5_14_BC12_S39_R2_001.fastq.gz  L5_6_MB8_S31_R1_001.fastq.gz    L9_4_SM1_S52_R2_001.fastq.gz
L2_12_MPHB4_S80_R1_001.fastq.gz       L4_4_MB9_S14_R2_001.fastq.gz     L5_15_BC13_S40_R1_001.fastq.gz  L5_6_MB8_S31_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/02_eager/NC_006531.1$ ls *gz
L11_7_SCO_1354_S70_R1_001.fastq.gz  L11_7_SCO_1354_S70_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/02_eager/NC_009263.1$ ls *gz
L2_4_CD2_S74_R1_001.fastq.gz  L2_4_CD2_S74_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/02_eager/NC_009263.1$ ls *gz
L2_4_CD2_S74_R1_001.fastq.gz  L2_4_CD2_S74_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/02_eager/NC_012928.1$ ls *gz
L0_4_EG9_S84_R1_001.fastq.gz  L0_4_EG9_S84_R2_001.fastq.gz
```

Regroupement des jeux de données par référence

```
/work/project/crucial/PALEOFISH/02_eager/NC_018341.1$ ls *gz
L2_4_CD2_S74_R1_001.fastq.gz  L2_4_CD2_S74_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/02_eager/NC_020762.1$ ls *gz
L0_5_EG10_S85_R1_001.fastq.gz  L0_5_EG10_S85_R2_001.fastq.gz  L7_14_22d_S48_R1_001.fastq.gz  L7_14_22d_S48_R2_001.fastq.gz
```

```
/work/project/crucial/PALEOFISH/02_eager/NC_024032.1$ ls *gz
L0_1_EG4_S81_R1_001.fastq.gz          L11_6_SCO_116_S69_R1_001.fastq.gz  L3_9_MB17_S4_R1_001.fastq.gz    L6_11_SH2000-99-394-g_S46_R1_001.fastq.gz
L0_1_EG4_S81_R2_001.fastq.gz          L11_6_SCO_116_S69_R2_001.fastq.gz  L3_9_MB17_S4_R2_001.fastq.gz    L6_11_SH2000-99-394-g_S46_R2_001.fastq.gz
L0_3_EG6_S83_R1_001.fastq.gz          L2_6_TP1_S75_R1_001.fastq.gz       L4_10_BC14_S20_R1_001.fastq.gz  L6_13_SH2000-99-394-f_S47_R1_001.fastq.gz
L0_3_EG6_S83_R2_001.fastq.gz          L2_6_TP1_S75_R2_001.fastq.gz       L4_10_BC14_S20_R2_001.fastq.gz  L6_13_SH2000-99-394-f_S47_R2_001.fastq.gz
L0_6_EG13_S86_R1_001.fastq.gz         L2_7_TP2_S76_R1_001.fastq.gz       L4_12_MD1_S22_R1_001.fastq.gz   L6_3_SH1979-4-6843-e_S41_R1_001.fastq.gz
L0_6_EG13_S86_R2_001.fastq.gz         L2_7_TP2_S76_R2_001.fastq.gz       L4_12_MD1_S22_R2_001.fastq.gz   L6_3_SH1979-4-6843-e_S41_R2_001.fastq.gz
L0_7_OLG1_S87_R1_001.fastq.gz         L2_8_BC1_S77_R1_001.fastq.gz       L4_14_SM2_S24_R1_001.fastq.gz   L6_4_SH1979-4-6843-c_S42_R1_001.fastq.gz
L0_7_OLG1_S87_R2_001.fastq.gz         L2_8_BC1_S77_R2_001.fastq.gz       L4_14_SM2_S24_R2_001.fastq.gz   L6_4_SH1979-4-6843-c_S42_R2_001.fastq.gz
L10_12_SCO_2213_S64_R1_001.fastq.gz   L3_10_MB13_S5_R1_001.fastq.gz      L4_15_BC10_S25_R1_001.fastq.gz  L6_5_SH1979-4-6843-d_S43_R1_001.fastq.gz
L10_12_SCO_2213_S64_R2_001.fastq.gz   L3_10_MB13_S5_R2_001.fastq.gz      L4_15_BC10_S25_R2_001.fastq.gz  L6_5_SH1979-4-6843-d_S43_R2_001.fastq.gz
L10_1_BC5_S58_R1_001.fastq.gz         L3_11_MB18_S6_R1_001.fastq.gz      L4_1_MB14_S11_R1_001.fastq.gz   L6_6_SH2001-106_11-b_S44_R1_001.fastq.gz
L10_1_BC5_S58_R2_001.fastq.gz         L3_11_MB18_S6_R2_001.fastq.gz      L4_1_MB14_S11_R2_001.fastq.gz   L6_6_SH2001-106_11-b_S44_R2_001.fastq.gz
L10_2_BC6_S59_R1_001.fastq.gz         L3_12_MB6_S7_R1_001.fastq.gz       L4_2_MD13_S12_R1_001.fastq.gz   L6_8_SH2000-99-394-e_S45_R1_001.fastq.gz
L10_2_BC6_S59_R2_001.fastq.gz         L3_12_MB6_S7_R2_001.fastq.gz       L4_2_MD13_S12_R2_001.fastq.gz   L6_8_SH2000-99-394-e_S45_R2_001.fastq.gz
L10_3_BC7_S60_R1_001.fastq.gz         L3_13_MB11_S8_R1_001.fastq.gz      L4_3_MB16_S13_R1_001.fastq.gz   L9_13_SH2000-99-390-b_S55_R1_001.fastq.gz
L10_3_BC7_S60_R2_001.fastq.gz         L3_13_MB11_S8_R2_001.fastq.gz      L4_3_MB16_S13_R2_001.fastq.gz   L9_13_SH2000-99-390-b_S55_R2_001.fastq.gz
L10_4_BC8_S61_R1_001.fastq.gz         L3_14_MB10_S9_R1_001.fastq.gz      L4_6_MB2_S16_R1_001.fastq.gz    L9_14_SH1979-4-6843-a_S56_R1_001.fastq.gz
L10_4_BC8_S61_R2_001.fastq.gz         L3_14_MB10_S9_R2_001.fastq.gz      L4_6_MB2_S16_R2_001.fastq.gz    L9_14_SH1979-4-6843-a_S56_R2_001.fastq.gz
L10_5_BC9_S62_R1_001.fastq.gz         L3_15_MB12_S10_R1_001.fastq.gz     L4_7_MB1_S17_R1_001.fastq.gz    L9_15_SH2000-99-390-c_S57_R1_001.fastq.gz
L10_5_BC9_S62_R2_001.fastq.gz         L3_15_MB12_S10_R2_001.fastq.gz     L4_7_MB1_S17_R2_001.fastq.gz    L9_15_SH2000-99-390-c_S57_R2_001.fastq.gz
L11_1_SCO_98_S66_R1_001.fastq.gz      L3_1_MB15_S1_R1_001.fastq.gz       L5_10_MD11_S35_R1_001.fastq.gz  L9_6_SH1979-4-6843-b_S53_R1_001.fastq.gz
L11_1_SCO_98_S66_R2_001.fastq.gz      L3_1_MB15_S1_R2_001.fastq.gz       L5_10_MD11_S35_R2_001.fastq.gz  L9_6_SH1979-4-6843-b_S53_R2_001.fastq.gz
L11_2_SCO_81_S67_R1_001.fastq.gz      L3_5_SM3_S2_R1_001.fastq.gz        L5_11_MD14_S36_R1_001.fastq.gz  L9_7_SH2000-99-394-c_S54_R1_001.fastq.gz
L11_2_SCO_81_S67_R2_001.fastq.gz      L3_5_SM3_S2_R2_001.fastq.gz        L5_11_MD14_S36_R2_001.fastq.gz  L9_7_SH2000-99-394-c_S54_R2_001.fastq.gz
```

Validation de la liste des références

* Localisation : /work/project/crucial/Ref_genomes/mito_genome

* Nombre de références ?

/work/project/crucial/Ref_genomes/mito_genome$ ls *fasta | wc -l

45 - 1 (allRef_43_mito_genomes.fasta) = 44

=> 39 références ou 44 ? => Générer allRef_43_mito_genomes.fasta

* Non indexé ?

/work/project/crucial/Ref_genomes/mito_genome$ ls NC_037939*

NC_037939.1_Salmo_obtusirostris.fasta

Résultats du pipeline (en cours)

Comme les paramétrages spécifiques aux données shrotGun avaient été listés avant mai 2025, une partie des traitements ont déjà été réalisé en mai 2025 pour les 64 jeux de données référencés NC_001960.1 :

```
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 13 12:09 BAM_files
-rw-r--r--  1 smaman CRUCIAL  941 May 19 10:18 02bis-bam_tsv.sh
-rw-r--r--  1 smaman CRUCIAL 2.5K May 19 10:31 01_script_eager_part1_NC_
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 19 10:32 FastP
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 19 10:32 documentation
drwxr-sr-x  3 smaman CRUCIAL 4.0K May 19 10:32 adapterremoval
drwxr-sr-x  4 smaman CRUCIAL 4.0K May 19 10:32 fastqc
drwxr-sr-x 14 smaman CRUCIAL 4.0K May 19 10:45 ..
drwxr-sr-x  6 smaman CRUCIAL 4.0K May 19 11:04 reference_genome
drwxr-sr-x  3 smaman CRUCIAL 4.0K May 19 11:07 mapping
drwxr-sr-x  5 smaman CRUCIAL 4.0K May 19 11:07 samtools
drwxr-sr-x  3 smaman CRUCIAL 4.0K May 19 11:07 deduplication
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 19 11:08 endorspy
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 19 11:08 preseq
drwxr-sr-x  3 smaman CRUCIAL 4.0K May 19 11:08 mapdamage
drwxr-sr-x  3 smaman CRUCIAL 4.0K May 19 11:08 multiqc
-rw-r--r--  1 smaman CRUCIAL 7.1K May 19 11:08 pipeline_trace.txt
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 19 11:08 pipeline_info
-rw-r--r--  1 smaman CRUCIAL 125K May 19 11:08 slurm-20976860.out
-rw-r--r--  1 smaman CRUCIAL 1.1K May 19 11:18 02_rehead_bam_files_v2_NC
-rw-r--r--  1 smaman CRUCIAL  177 May 19 11:18 slurm-20978361.out
drwxr-sr-x  3 smaman CRUCIAL 4.0K May 19 11:18 reheader_bam
-rw-r--r--  1 smaman CRUCIAL    0 May 19 11:21 slurm-20979167.out
drwxr-sr-x  2 smaman CRUCIAL 4.0K May 19 11:21 03_eager_only_genotyping
-rw-r--r--  1 smaman CRUCIAL  740 May 19 11:48 02ter_tsv_modif.sh
-rw-r--r--  1 smaman CRUCIAL    0 May 19 11:51 slurm-20981660.out
-rw-r--r--  1 smaman CRUCIAL 6.7K May 19 11:51 input_bam.tsv
drwxr-sr-x 12 smaman CRUCIAL 8.0K May 20 09:47 03_script_eager_only_gen
-rw-r--r--  1 smaman CRUCIAL 6.6K May 22 12:26 input_NC_001960.1.tsv
```

Je poursuis pour les autres jeux, sous réserve de confirmation de la répartition des jeux de données par référence et des références, par Joelle.