Paleofish Mitochondries

Objectif : Lancement de Eager sur référence majoritaire

16/06/2025

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

Analyses FastQ Screen :

L11_7_SCO_1354_S70_R2_001	A.ruthenus	9163
L11_7_SCO_1354_S70_R1_001	A.ruthenus	9497
L0_5_EG10_S85_R2_001	C.clupeaformis	8299
L0_5_EG10_S85_R1_001	C.clupeaformis	8377
L7_14_22d_S48_R1_001	C.clupeaformis	104410
L7_14_22d_S48_R2_001	C.clupeaformis	104461
L2_4_CD2_S74_R1_001	H.sapiens	3826
L2_4_CD2_S74_R2_001	H.sapiens	4427
L0_4_EG9_S84_R2_001	T.thymallus	1262
L0_4_EG9_S84_R1_001	T.thymallus	1275

Les autres échantillons : Salmo salar ou Salmo trutta

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
```