Paleofish Mitochondries

Objectif : Lancement de Eager sur référence majoritaire

16/10/2025

CR réunion du 16 oct 2025

- Mise à disposition des scripts dans une foge INRAE : Forge : https://forge.inrae.fr/bios4biol/mt_paleofish
- Présentation des résultats du second multimapping Brachymystax
- Présentation des résultats de recherche des mitogénomes dans les whole génomes.
- Lancer des pipelines Eager avec les mitogénomes précédemment définis.

Rappel résultats des multimapping

4 Undeterminated
5 Autres que S. salar et S.trutta
32 S. salar
47 S. trutta

_____
88 échantillons

L11_7_SCO_1354_S70:NC_006531.1:182488 → Anguille
NC_006535.3 164611 Anguilla bicolor pacifica

L0_4_EG9_S84:NC_012928.1:19566 → Thymallus thymallus
NC_056303.1 12714 Thymallus ligericus

L2_4_CD2_S74:NC_018341.1:5799 → Brachymystax
NC_018341.1 727 Brachymystax lenok

L0_5_EG10_S85:NC_020762.1:38793 → Coregonus
L7_14_22d_S48:NC_020762.1:1255682
L0_5_EG10_S85 : NC_002646.1 6117 Coregonus lavaretus
L7_14_22d_S48 : NC_020765.1 186123 Coregonus oxyrinchus

Méthodologie pour la préparation des whole génomes

Deux cas de figure :

1/ Retrait des scaffolds avec seqtk puis ajout du mitogénome avec cat

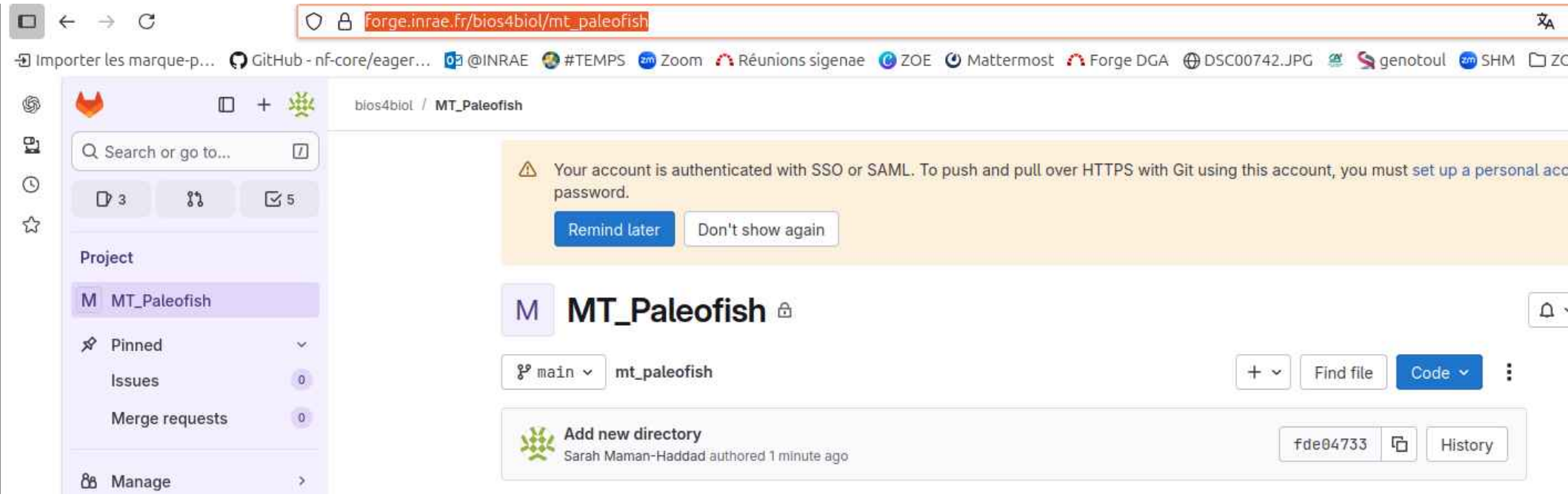
2/ Modification du nom de la circular target dans le pipeline de lancement Eager

Résultats du double pass Eager

Espèces	Liens vers les séquences consensus	Commentaires
Anguilla bicolor pacifica	smaman@genobioinfo1 /work/project/crucial/PALEOFISH/02_eager_notSalmo/Anguilla_bicolor_pacifica/03_eager_only_genotyping $ ls consensus_sequence/ SCO_1354.fasta.gz SCO_1354.fasta_refmod.fasta.gz SCO_1354.fasta_uncertainty.fasta.gz	sort S70.rmdup.reheader.bam (1) dos2unix et retrait ^$MS dans le mitogenome
Brachymystax lenox tsinlingensis	smaman@genobioinfo2 /work/project/crucial/PALEOFISH/02_eager_notSalmo/Brachymystax_lenox_tsinlingensis/03_eager_only_genotyping $ ls consensus_sequence/ CD2.fasta.gz CD2.fasta_refmod.fasta.gz CD2.fasta_uncertainty.fasta.gz	-
Thymallus_ligericus	smaman@genobioinfo1 /work/project/crucial/PALEOFISH/02_eager_notSalmo/Thymallus_ligericus/03_eager_only_genotyping $ ls consensus_sequence/ EG9_S84.fasta.gz EG9_S84.fasta_refmod.fasta.gz EG9_S84.fasta_uncertainty.fasta.gz	sort S84.rmdup.reheader.sam (1)
Coregonus lavaretus	smaman@genobioinfo1 /work/project/crucial/PALEOFISH/02_eager_notSalmo/Coregonus/Coregonus_lavaretus/03_eager_only_genotyping $ ls consensus_sequence/ EG10.fasta.gz EG10.fasta_refmod.fasta.gz EG10.fasta_uncertainty.fasta.gz	dos2unix et retrait ^$MS dans le mitogenome
Coregonus oxyrinchus	smaman@genobioinfo1 /work/project/crucial/PALEOFISH/02_eager_notSalmo/Coregonus/Coregonus_oxyrinchus/03_eager_only_genotyping $ ls consensus_sequence/ 22d.fasta.gz 22d.fasta_refmod.fasta.gz 22d.fasta_uncertainty.fasta.gz	dos2unix et retrait ^$MS dans le mitogenome

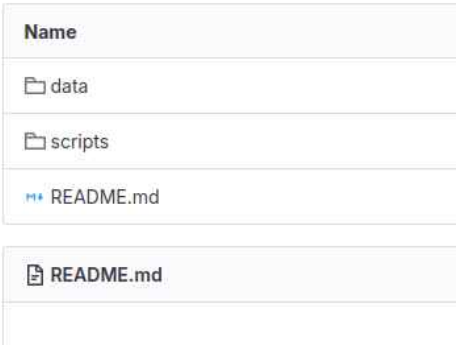
(1) BAM contient des reads non alignés (NO_COOR) qui ne sont pas regroupés à la fin du fichier, ce qui empêche la création de l'index

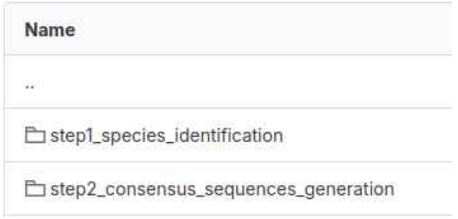
Partage des scripts dans la forge INRAE



https://forge.inrae.fr/bios4biol/mt_paleofish

Partage des scripts dans la forge INRAE





Arborescence inspirée du work du projet PALEOFISH.
Ajout de fichiers README et des scripts

Scripts de mise en forme de fichiers

La compilation des fichiers idxstats est disponible dans le fichier: /work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/STATS/compilation.txt généré par le script build_matrix.py

```
smaman@genobioinfo1 ~ $ more /work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/STATS/compilation.txt
sample  CM038646.1      DQ536427.1      KF913024.1      LC381866.1      NC_000860.1     NC_000861.1     NC_001606.1     NC_001717.1     NC_001960.1     NC_002980.1     NC_004379.1
NC_004593.1     NC_006291.1     NC_006531.1     NC_008615.1     NC_008648.1     NC_008654.1     NC_009263.1     NC_009575.1     NC_010959.1     NC_012576.1     NC_012928.1     NC_01426
1.1     NC_018341.1     NC_020355.1     NC_020356.1     NC_020762.1     NC_022844.1     NC_024032.1     NC_025589.1     NC_026313.1     NC_026533.1     NC_031540.1     NC_031561.1
NC_036392.1     NC_037502.1     NC_081187.1     OM736820.1      OX637958.1
L0_1_EG4_S81_R1_001     18      18      22      687     631     633     16      741     3895    759     68      74      14      72      805     18      16      829     44      714
48      428     64      983     14      18      343     717     42604   1136    49      46      14      18      729     624     34      458     34
L0_2_EG5_S82_R1_001     98      74      126     4085    3826    3161    85      3248    347642  3570    194     443     120     440     3601    126     113     3511    284     3556
172     2446    340     5375    140     116     3732    3794    18843   6140    23
L0_3_EG6_S83_R1_001     12      5       10      646     562     465     6       80
30      326     88      709     18      10      315     583     39843   724     36
L0_4_EG9_S84_R1_001     22      20      32      850     822     780     72      13
40      19566   42      978     30      22      874     834     1943    937     13
```

```
smaman@genobioinfo1 ~ $ more /work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/STATS/build_matrix.py
#!/usr/bin/env python3
import os
import glob

# Récupère tous les fichiers sans extension dans le dossier courant
files = [f for f in glob.glob("*") if os.path.isfile(f) and not '.' in os.path.splitext(f)[1]]

data = {}
genomes = set()

for filename in files:
    sample = filename.replace("idxstats_cleaned_", "")
    with open(filename) as f:
        for line in f:
            parts = line.strip().split()
            # Ignore la ligne si elle commence par un astérisque
            if len(parts) >= 3 and not parts[0].startswith("*"):
                genome = parts[0]
                value = parts[2]
                data.setdefault(sample, {})[genome] = value
                genomes.add(genome)

# Trie les génomes pour l'ordre des colonnes
genomes = sorted(genomes)

# Écriture du tableau dans un fichier .txt
with open("compilation.txt", "w") as out:
    out.write("sample\t" + "\t".join(genomes) + "\n")
    for sample in sorted(data.keys()):
        values = [data[sample].get(g, "0") for g in genomes]
        out.write(sample + "\t" + "\t".join(values) + "\n")

print("Fichier compilation.txt généré avec succès.")
```

Les nombre de reads par échantillon post fastp est disponible ici (généré pa
/work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/RES/count_reads_cleaned_sample.txt

Scripts de mise en forme de fichiers

Les nombre de reads par échantillon post fastp est disponible ici (généré par ./count_reads_cleaned_sample.sh):
/work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/RES/count_reads_cleaned_sample.txt

```
smaman@genobioinfo1 ~ $ more /work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/RES/count_reads_cleaned_sample.txt
Fichier Nombre_de_reads
cleaned_L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq        3072280
cleaned_L1_1_AUD20212-2_Sxx_R2_001_GACGATT+TCGCAGG.fastq        3072280
cleaned_L1_2_BC4_Sxx_R1_001_AACCTGC+CTCTGCA.fastq       797127
cleaned_L1_2_BC4_Sxx_R2_001_AACCTGC+CTCTGCA.fastq       797127
cleaned_L1_4_AUD11764-47_Sxx_R1_001_GCCTACG+GGATCAA.fastq       2466072
cleaned_L1_4_AUD11764-47_Sxx_R2_001_GCCTACG+GGATCAA.fastq       2466072
cleaned_L1_7_BC2_Sxx_R1_001_GTCCGGC+CTCGATG.fastq       2718209
cleaned_L1_7_BC2_Sxx_R2_001_GTCCGGC+CTCGATG.fastq       2718209
cleaned_others_underterminated_R1.fastq 16776989
cleaned_others_underterminated_R2.fastq 16776989
cleaned_L0_1_EG4_S81_R1_001.fastq.gz    54448
cleaned_L0_1_EG4_S81_R2_001.fastq.gz    54448
cleaned_L0_2_EG5_S82_R1_001.fastq.gz    313235
cleaned_L0_2_EG5_S82_R2_001.fastq.gz    313235
cleaned_L0_3_EG6_S83_R1_001.fastq.gz    91032
cleaned_L0_3_EG6_S83_R2_001.fastq.gz    91032
cleaned_L0_4_EG9_S84_R1_001.fastq.gz    62687
cleaned_L0_4_EG9_S84_R2_001.fastq.gz    62687
cleaned_L0_5_EG10_S85_R1_001.fastq.gz   129818
cleaned_L0_5_EG10_S85_R2_001.fastq.gz   129818
cleaned_L0_6_EG13_S86_R1_001.fastq.gz   86993
cleaned_L0_6_EG13_S86_R2_001.fastq.gz   86993
cleaned_L0_7_OLG1_S87_R1_001.fastq.gz   42669
cleaned_L0_7_OLG1_S87_R2_001.fastq.gz   42669
cleaned_L0_8_SH2_S88_R1_001.fastq.gz    63354
cleaned_L0_8_SH2_S88_R2_001.fastq.gz    63354
cleaned_L10_11_SCO_1338_S63_R1_001.fastq.gz     3730986
cleaned_L10_11_SCO_1338_S63_R2_001.fastq.gz     3730986
cleaned_L10_12_SCO_2213_S64_R1_001.fastq.gz     1001235
```

```
smaman@genobioinfo1 ~ $ more /work/project/crucial/PALEOFISH/01_detect_species_1st_mapping_against_39_MT_Salmo/RES/*sh
#!/bin/bash

echo -e "Fichier\tNombre_de_reads"

for file in *.fastq *.fastq.gz; do
  # Ignore si le fichier n'existe pas (évite les erreurs si aucun .fastq ou .fastq.gz)
  [ -e "$file" ] || continue

  if [[ "$file" == *.gz ]]; then
    # Fichier compressé : utiliser zcat
    count=$(zcat "$file" | awk 'END {print NR/4}')
  else
    # Fichier non compressé
    count=$(awk 'END {print NR/4}' "$file")
  fi

  echo -e "$file\t$count"
done
```