Paleofish Mitochondries

Objectif : Lancement de Eager sur référence majoritaire

15/01/2026

CR réunion du 27 novembre 2025

- Pour les 5 échantillons not salmo, travailler sur les espèces en génome de référence et non les sous espèces (même pour Thymallus): Anguilla anguilla, Thymallus thymallus, Coregonus clupeaformis, Brachymystax lenok --> Eager:

smaman@genobioinfo1 /work/project/crucial/PALEOFISH/ARCHIVES/archive_02_eager_44_mito_genomes $ ls

1-bwa_sur_ref_majoritaire.sh NC_001960.1 NC_006531.1 NC_010007.1 NC_012928.1 NC_018341.1 NC_024032.1 NC_025648.1 OLD REF_MAJORITAIRE scripts-archives

--> Joelle: voir multiQC, comparaison consensus double pass Eager sur espèces et sous-espèces.

- Pas besoin d'IGV.js pour la visualisation des consensus.

- Sarah : comparer consensus mpileup avec consensus double pass sur espèces.

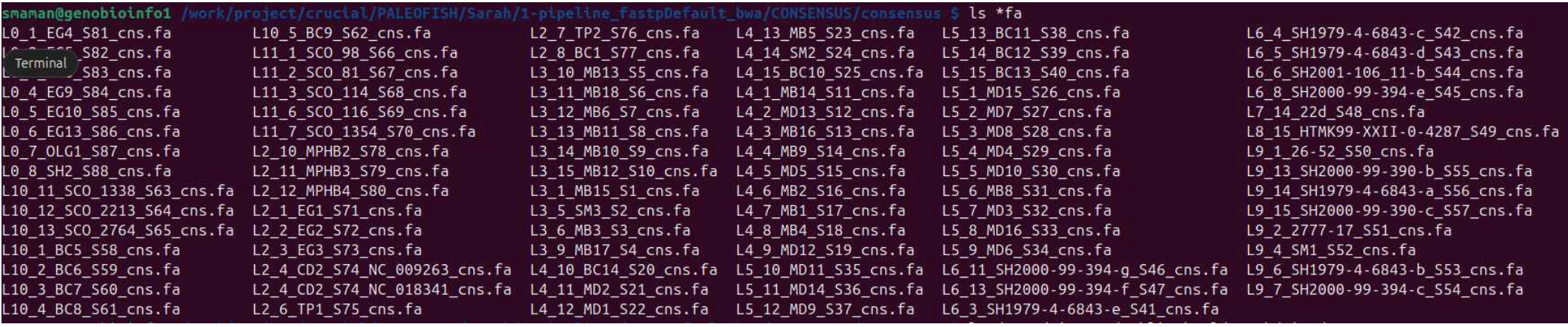
Autres points abordés lors d'une précédente réunion :

* Nettoyer les fichiers undeterminated pour les traiter avec Eager (pour fev/mars 26).

* Point projet : https://annuel.framapad.org/p/pnqb6r9hsu-ai09?lang=fr

Comparer consensus mpileup avec consensus double pass sur espèces.

Visualisation consensus mpileup



Réflexions autour du L11_7

Visualisation consensus mpileup L11 7 Anguilla

/save/sigenae/public_html/sarah/mito/consensus_mpileup/.

```
smaman@genobioinfo1 /save/sigenae/public_html/sarah/mito/consensus_mpileup $ ls ../consensus_not_salmo
Anguilla  Brachymystax_lenox_tsinlingensis  Coregonus  Thymallus_ligericus
smaman@genobioinfo1 /save/sigenae/public_html/sarah/mito/consensus_mpileup $ ls ../consensus_not_salmo/Anguilla
CM077320.2.fna      GCA_039654925.2_AngPac_1_genomic.fna          SCO_1354.realign.bai  SCO_1354.unifiedgenotyper.vcf.gz
CM077320.2.fna.fai  igvjs_SCO_1354_Anguilla_bicolor_pacifica.html  SCO_1354.realign.bam
```

```
smaman@genobioinfo2 /work/project/crucial/PALEOFISH/comparaison_mpileup_eager $ ls /work/project/crucial/PALEOFISH/ARCHIVES/archive_02_eager_44_mito_genomes/NC_006531.1/03_eager_only_g
enotyping/results/consensus_sequence/
SCO_1354.fasta.gz  SCO_1354.fasta_refmod.fasta.gz  SCO_1354.fasta_uncertainty.fasta.gz
```

Consensus Eager
L11 7 Anguilla

```
smaman@genobioinfo2 /work/project/crucial/PALEOFISH/comparaison_mpileup_eager $ ls /work/project/crucial/PALEOFISH/ARCHIVES/archive_02_eager_44_mito_genomes/NC_006531.1/03_eager_only_g
enotyping/results/consensus_sequence/
SCO_1354.fasta.gz  SCO_1354.fasta_refmod.fasta.gz  SCO_1354.fasta_uncertainty.fasta.gz
```

Comparaison des consensus Eager/mpileup L11 7 Anguilla

```
smaman@genobioinfo2 /work/project/crucial/PALEOFISH/comparaison_mpileup_eager $ zgrep 'N' /work/project/crucial/PALEOFISH/ARCHIVES/archive_02_eager_44_mito_genomes/NC_006531.1/03_eager_only_genotyping/results/consensus_sequence/SCO_1354.fasta.gz | wc -l
276
smaman@genobioinfo2 /work/project/crucial/PALEOFISH/comparaison_mpileup_eager $ grep 'N' /work/project/crucial/PALEOFISH/Sarah/1-pipeline_fastpDefault_bwa/CONSENSUS/consensus/L11_7_SCO_1354_S70_cns.fa | wc -l
1
```

mpileup / Eager => Comparaison des IGV ?

```
smaman@genobioinfo2 /work/project/crucial/PALEOFISH/comparaison
4_S70_cns.fa
>NC_006531.1 Anguilla anguilla mitochondrion, complete genome
GTTAACGTAGCTTAAACAAAAAGCATGGCACTGAAGATGCCAAGATGAGCCATAAAAAGC
TCCGATGACACAAAAGCCTGGTCCTGACTTTAACATCAGTTCTGGCCTGACTTACACATG
CAAGTACCCGCGCACCCGTGAGAATGCCCTATATCCCCTCCCGGGGAAAAGGAGCCGGCA
TCAGGCACACCAACGTAGCCCAAAACACCTTGCTTAACCACACCCCCAAGGGAATTCAGC
AGTGATAGACATTGAGCAATAAGCGCAAGCTTGACTTAGTCAAGGCCAAAAGAGTTGGTT
AATCTCGTGCCAGCCACCGCGGTTATACGAGTAACTCACATTAATACTTCACGGCGTAAA
```

```
>SCO_1354_draftN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTT
AATCTCGTGCCAGCCACCGCGGTTATACGAGTNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNACTACNAGCGCTAGCTTAAAACCCAAAGGACTTGGCGGTG
CCCCAGACCCCCCTAGAGGAGCCTGTTCTATAACCGATAATCCACGTTAAACCTNACCAC
```

Nettoyer les fichiers undeterminated pour les traiter avec Eager (pour fev/mars 26).

CR 11 sept 2025

Voici les points abordés:

Présentation des résultats et des slides d'avancement envoyés par mail ce jour au groupe de travail: multimapping sur 44 références mito, répartition des jeux de données capture_janv22 par lot, pipelines Eager sur chaque lot, chemin d'accès aux résultats, exclusion des fichiers Sxx issus des fastq undeterminated.

Lancement FastQScreen pour "coller" au pipeline Paleotrutta et pour se conforter sur les résultats du multimapping, sans a priori.

Il a été décidé:

Ne pas traiter les données shotgun car elles ne peuvent pas être exploitées par Joelle. Traiter uniquement les données capture. @Joelle : Confirmer le chemin d'accès et la liste des données à traiter : /work/project/crucial/PALEOFISH/DATA/capture_jan22

Ne pas traiter les jeux de données récupérés avec les barcodes depuis les fichiers Undeterminated car les fichiers contiennent des reads non pairés. Ces fichiers *Sxx* sont donc archivés dans un nouveau répertoire Undeterminated/ : smaman@genobioinfo2 /work/project/crucial/PALEOFISH/DATA/capture_jan22 $ ls Undeterminated/

L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq L1_2_BC4_Sxx_R2_001_AACCTGC+CTCTGCA.fastq
L1_7_BC2_Sxx_R1_001_GTCCGGC+CTCGATG.fastq Undetermined_S0_R2_001.fastq.gz

L1_1_AUD20212-2_Sxx_R2_001_GACGATT+TCGCAGG.fastq L1_4_AUD11764-47_Sxx_R1_001_GCCTACG+GGATCAA.fastq
L1_7_BC2_Sxx_R2_001_GTCCGGC+CTCGATG.fastq

L1_2_BC4_Sxx_R1_001_AACCTGC+CTCTGCA.fastq L1_4_AUD11764-47_Sxx_R2_001_GCCTACG+GGATCAA.fastq Undetermined_S0_R1_001.fastq.gz

o

Noms de paires de reads qui ne matchent pas => Retrait des *Sxx* du lot traité.

```
Command error:
  Trimming paired end reads ...
  Opening FASTQ file 'repaired_L1_2_BC4_Sxx_R1_001_AACCTGC+CTCTGCA.fastq.pG.fq.gz', line numbers start at 1
  Opening FASTQ file 'repaired_L1_2_BC4_Sxx_R2_001_AACCTGC+CTCTGCA.fastq.pG.fq.gz', line numbers start at 1
  ERROR: Unhandled exception in thread:
      Pair contains reads with mismatching names:
      - 'NB501044:604:HHNTFAFX3:1:11101:6394:7897'
      - 'NB501044:604:HHNTFAFX3:1:11101:2238:8023'

      Note that AdapterRemoval by determines the mate numbers as the digit found at
      the end of the read name, if this is preceded by the character '/'; if these
      data makes use of a different character to separate the mate number from the
      read name, then you will need to set the --mate-separator command-line option
      to the appropriate character.
  ERROR: AdapterRemoval did not run to completion;
         do NOT make use of resulting trimmed reads!
```

Fastq Sxx reconstruits à partir des fichiers Undeterminated

Séquences non reconnues par Eager :

```
/work/project/crucial/PALEOFISH/02_eager/NC_001960.1$ seqtk seq -A L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq | head
>NB501044:604:HHNTFAFX3:1:11101:1980:1042 1:N:0:GACGATT+TCGCAGG
NATTTTCTGTAGACAACGACACCCTAACACGATTTTTCGCCTTCCACTTCCTATTCCCATTCGTTATTGCAGCTGC
>NB501044:604:HHNTFAFX3:1:11101:20400:1043 1:N:0:GACGATT+TCGCAGG
NAGGCTCTTGGTCTTAGTGCTATCCTAAGCCCTTGTTATAAGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
>NB501044:604:HHNTFAFX3:1:11101:5386:1044 1:N:0:GACGATT+TCGCAGG
NCTTATAATTCAAGTAGCCCCAACTATCAACTCTTCTCTACTCATTGCGATAGGCCTTTTATCAATACTTGTAGGA
>NB501044:604:HHNTFAFX3:1:11101:19272:1045 1:N:0:GACGATT+TCGCAGG
AGATCCCCCCGGCTTCCGCGCGGCAAACCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGACGATTATCGCG
>NB501044:604:HHNTFAFX3:1:11101:6772:1045 1:N:0:GACGATT+TCGCAGG
TATAATTAAAGCTGCTTTCGTAGTAGATCTTCGTGCCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGACGA
/work/project/crucial/PALEOFISH/02_eager/NC_001960.1$ file L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq
L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq: ASCII text
/work/project/crucial/PALEOFISH/02_eager/NC_001960.1$ wc -l L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq
12785788 L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq
/work/project/crucial/PALEOFISH/02_eager/NC_001960.1$ cat -A L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq | head
@NB501044:604:HHNTFAFX3:1:11101:1980:1042 1:N:0:GACGATT+TCGCAGG$
NATTTTCTGTAGACAACGACACCCTAACACGATTTTTCGCCTTCCACTTCCTATTCCCATTCGTTATTGCAGCTGC$
+$
#AAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEAEEEEEE$
@NB501044:604:HHNTFAFX3:1:11101:20400:1043 1:N:0:GACGATT+TCGCAGG$
NAGGCTCTTGGTCTTAGTGCTATCCTAAGCCCTTGTTATAAGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC$
+$
#AAAAEEEEEEEEEEEEEEEEEEEEEE/EEEEEEEEEEAAEAEAEEEEEEEEEEE/EEEEEE/EEEE/EEEEA$
@NB501044:604:HHNTFAFX3:1:11101:5386:1044 1:N:0:GACGATT+TCGCAGG$
NCTTATAATTCAAGTAGCCCCAACTATCAACTCTTCTCTACTCATTGCGATAGGCCTTTTATCAATACTTGTAGGA$
```

Réparation des fichiers undeterminated pour les traiter avec Eager

```
smaman@genobioinfo2 /work/project/crucial/PALEOFISH/DATA/capture_jan22/Undeterminated $ ls
L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq  L1_2_BC4_Sxx_R2_001_AACCTGC+CTCTGCA.fastq        L1_7_BC2_Sxx_R1_001_GTCCGGC+CTCGATG.fastq  Undetermined_S0_R2_001.fastq.gz
L1_1_AUD20212-2_Sxx_R2_001_GACGATT+TCGCAGG.fastq  L1_4_AUD11764-47_Sxx_R1_001_GCCTACG+GGATCAA.fastq  L1_7_BC2_Sxx_R2_001_GTCCGGC+CTCGATG.fastq
L1_2_BC4_Sxx_R1_001_AACCTGC+CTCTGCA.fastq         L1_4_AUD11764-47_Sxx_R2_001_GCCTACG+GGATCAA.fastq  Undetermined_S0_R1_001.fastq.gz
```

```
smaman@n001 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED $ awk 'NR>=4 && NR<=8{print NR ":" $0}' L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq
4:#AAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEAEEEEEE
5:--
6:@NB501044:604:HHNTFAFX3:1:11101:20400:1043 1:N:0:GACGATT+TCGCAGG
7:NAGGCTCTTGGTCTTAGTGCTATCCTAAGCCCTTGTTATAAGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
8:+

SOLUTION : SUPPRESSION DES '--'

smaman@n032 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED $ sed '/^[[:space:]]*--[[:space:]]*$/d'  L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq > L1_1_AUD20212-2_Sxx_R1_cleaned.fastq
```

```
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED $ ls
L1_1_AUD20212-2_Sxx_R1_001_GACGATT+TCGCAGG.fastq  L1_2_BC4_Sxx_R1_cleaned.fastq                     L1_4_AUD11764-47_Sxx_R2_001_GCCTACG+GGATCAA.fastq  L1_7_BC2_Sxx_R2_cleaned.fastq
L1_1_AUD20212-2_Sxx_R1_cleaned.fastq              L1_2_BC4_Sxx_R2_001_AACCTGC+CTCTGCA.fastq         L1_4_AUD11764-47_Sxx_R2_cleaned.fastq              long_reads.sh
L1_1_AUD20212-2_Sxx_R2_001_GACGATT+TCGCAGG.fastq  L1_2_BC4_Sxx_R2_cleaned.fastq                     L1_7_BC2_Sxx_R1_001_GTCCGGC+CTCGATG.fastq          README
L1_1_AUD20212-2_Sxx_R2_cleaned.fastq              L1_4_AUD11764-47_Sxx_R1_001_GCCTACG+GGATCAA.fastq L1_7_BC2_Sxx_R1_cleaned.fastq
L1_2_BC4_Sxx_R1_001_AACCTGC+CTCTGCA.fastq         L1_4_AUD11764-47_Sxx_R1_cleaned.fastq             L1_7_BC2_Sxx_R2_001_GTCCGGC+CTCGATG.fastq
```

Résultats de l'alignement et de la couverture :

```
/work/project/crucial/PALEOFISH/02_eager/UNDETERMINATED_ref_majoritaire$ more STATS/*align
::::::::::::::
STATS/L1_1_AUD20212-2_Sxx_GACGATT+TCGCAGG.align
::::::::::::::
NC_024032.1     154514
NC_010007.1     161407
NC_001960.1     4690381
::::::::::::::
STATS/L1_2_BC4_Sxx_AACCTGC+CTCTGCA.align
::::::::::::::
NC_024032.1     25409
NC_010007.1     25667
NC_001960.1     1012069
::::::::::::::
STATS/L1_4_AUD11764-47_Sxx_GCCTACG+GGATCAA.align
::::::::::::::
NC_024032.1     40467
NC_010007.1     42482
NC_001960.1     1200198
::::::::::::::
STATS/L1_7_BC2_Sxx_GTCCGGC+CTCGATG.align
::::::::::::::
NC_001960.1     100882
NC_010007.1     1727220
NC_024032.1     2029130
```

Lancement des étapes Eager sur les fichiers undeterminated nettoyés

```
smaman@genobioinfo1 ~ $ cd /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED/EAGER
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED/EAGER $ ls *
NC_001960.1:
01_script_eager_part1_NC_001960.1.sh   02ter_tsv_modif.sh        crucial_nextflow_VersionPaleofish.config  fastqc                pipeline_trace.txt  work
02bis-bam_tsv.sh                       03_eager_only_genotyping  documentation                             input_NC_001960.1.tsv  reference_genome
02_rehead_bam_files_v2_NC_001960.1.sh  adapterremoval            FastP                                     pipeline_info          slurm-29462790.out

NC_024032.1:
01_script_eager_part1_NC_024032.1.sh   03_eager_only_genotyping              documentation  input_NC_024032.1.tsv  reference_genome
02bis-bam_tsv.sh                       adapterremoval                        FastP          pipeline_info          slurm-29462832.out
02_rehead_bam_files_v2_NC_024032.1.sh  crucial_nextflow_VersionPaleofish.config  fastqc         pipeline_trace.txt     work
```

Pipeline et paramétrages mito

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.

Lancement des étapes Eager sur les fichiers undeterminated nettoyés

Nom de la librairie (Sxx) insuffisant pour identifier les échantillon lors de l'étape rehead BAM.
=> Ajout Sxx_NomSample pour rendre l'information unique.

```
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED/EAGER/NC_001960.1/03_eager_only_genotyping $ ls consensus_sequence/
AUD11764-47.fasta.gz              AUD11764-47.fasta_uncertainty.fasta.gz  AUD20212-2.fasta_refmod.fasta.gz      BC4.fasta.gz              BC4.fasta_uncertainty.fasta.gz
AUD11764-47.fasta_refmod.fasta.gz AUD20212-2.fasta.gz                     AUD20212-2.fasta_uncertainty.fasta.gz BC4.fasta_refmod.fasta.gz
```

```
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED/EAGER/NC_024032.1/03_eager_only_genotyping $ ls
03_script_eager_only_genotyping_v2_NC_024032.1.sh  damage_rescaling  genotyping     multiqc                                                  pipeline_trace.txt  slurm-29752109.out
bcftools                                           documentation     input_bam.tsv  OLD_03_script_eager_only_genotyping_v2_NC_024032.1.sh  reference_genome    temp_input_bam.tsv
consensus_sequence                                 endorspy          mapdamage      pipeline_info                                            samtools            work
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED/EAGER/NC_024032.1/03_eager_only_genotyping $ ls consensus_sequence/
BC2.fasta.gz  BC2.fasta_refmod.fasta.gz  BC2.fasta_uncertainty.fasta.gz
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/DATA_UNDETERMINATED/EAGER/NC_024032.1/03_eager_only_genotyping $
```

Point projet

https://annuel.framapad.org/p/pnqb6r9hsu-ai09?lang=fr