Paleofish Mitochondries

Objectif : Lancement de Eager sur référence majoritaire

17/09/2025

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

@Odile : Transmettre à Sarah la version BWA utilisée pour l'indexation de allRef.fasta

Reprendre l'ensemble des traitements multimapping avec les 39 ref mito du fichier /work/project/crucial/Ref_genomes/mito_genome/previous/allRef.fasta comme indiqué dans le mail de Joelle et en réunion ce jour.

Présentation des résultats sur multimapping sur 44 références mito. Joelle se demande donc s'il est intéressant d'affiner sur les références publiées intra-genre pour choisir la plus proche.

Joelle propose de lancer les étapes Eager uniquement sur le saumon et la truite. Pour les autres genres, nous nous laissons une semaine de réflexion.

Pipeline multimapping pour rechercher de les références majoritaires

Dans le répertoire : /work/project/crucial/PALEOFISH/02_eager/REF_MAJORITAIRE

0/ Indexation du multifasta de référence (00-BWA-index-fasta.sh) :/work/project/crucial/Ref_genomes/mito_genome/allRef_43_mito_genomes.fasta

1/ Nettoyage fastp des fichiers undeterminated dos2unix (01-FASTP_fastp-0.23.2.sh)

2/ Alignement sur le multifasta pour trouver les références majoritaires (02-multiAlignPostFastp.sh)

3/ Convert SAM en BAM (03-convertSamtoBam.sh)

4/ Sort BAM (04-sortBam.sh)

5/ Index BAM trié (05-indexSortedbam.sh)

6/ Statistiques avec Samtools idxstats (06-samtools-idxstats.sh)

7/ Tri des statistiques (07-sort_stats.sh)

8/ Couverture des fichiers BED (08-coverage_bedtools.sh)

Removed -- 9/ Couverture médiane (09-median-coverage.sh)

10/ Liste des références majoritaires (10-sort_stats_MAJORITAIRE.sh)

Regroupement des échantillons par référence majoritaire avec 44 mito génomes

L0_2_EG5_S82	NC_001960.1
L0_8_SH2_S88	NC_001960.1
L10_11_SCO_1338_S63	NC_001960.1
L10_13_SCO_2764_S65	NC_001960.1
L1_1_AUD20212-2_Sxx_GACGATT+	NC_001960.1
L1_2_BC4_Sxx_AACCTGC+CTCTGC	NC_001960.1
L1_4_AUD11764-47_Sxx_GCCTACG+	NC_001960.1
L2_10_MPHB2_S78	NC_001960.1
L2_11_MPHB3_S79	NC_001960.1
L2_12_MPHB4_S80	NC_001960.1
L2_1_EG1_S71	NC_001960.1
L2_2_EG2_S72	NC_001960.1
L2_3_EG3_S73	NC_001960.1
L4_11_MD2_S21	NC_001960.1
L4_13_MB5_S23	NC_001960.1
L4_4_MB9_S14	NC_001960.1
L4_5_MD5_S15	NC_001960.1
L4_8_MB4_S18	NC_001960.1
L4_9_MD12_S19	NC_001960.1
L5_12_MD9_S37	NC_001960.1
L5_13_BC11_S38	NC_001960.1
L5_14_BC12_S39	NC_001960.1
L5_15_BC13_S40	NC_001960.1
L5_1_MD15_S26	NC_001960.1
L5_2_MD7_S27	NC_001960.1
L5_3_MD8_S28	NC_001960.1
L5_4_MD4_S29	NC_001960.1
L5_5_MD10_S30	NC_001960.1
L5_6_MB8_S31	NC_001960.1
L5_7_MD3_S32	NC_001960.1
L5_8_MD16_S33	NC_001960.1
L8_15_HTMK99-XXII-0-4287_S49	NC_001960.1
L9_1_26-52_S50	NC_001960.1
L9_2_2777-17_S51	NC_001960.1
L9_4_SM1_S52	NC_001960.1

Sample	Ref
L11_7_SCO_1354_S70	NC_006531.1

Sample	Ref
L2_4_CD2_S74	NC_018341.1

Sample	Ref
L0_1_EG4_S81	NC_010007.1
L0_3_EG6_S83	NC_010007.1
L0_6_EG13_S86	NC_010007.1
L0_7_OLG1_S87	NC_010007.1
L10_12_SCO_2213_S64	NC_010007.1
L2_6_TP1_S75	NC_010007.1
L2_7_TP2_S76	NC_010007.1
L3_10_MB13_S5	NC_010007.1
L3_11_MB18_S6	NC_010007.1
L3_13_MB11_S8	NC_010007.1
L3_14_MB10_S9	NC_010007.1
L3_6_MB3_S3	NC_010007.1
L3_9_MB17_S4	NC_010007.1
L4_2_MD13_S12	NC_010007.1
L4_7_MB1_S17	NC_010007.1
L5_10_MD11_S35	NC_010007.1
L5_11_MD14_S36	NC_010007.1
L5_9_MD6_S34	NC_010007.1
L6_11_SH2000-99-394-g_S46	NC_010007.1
L6_13_SH2000-99-394-f_S47	NC_010007.1
L6_3_SH1979-4-6843-e_S41	NC_010007.1
L6_4_SH1979-4-6843-c_S42	NC_010007.1
L6_6_SH2001-106_11-b_S44	NC_010007.1
L9_15_SH2000-99-390-c_S57	NC_010007.1
L9_6_SH1979-4-6843-b_S53	NC_010007.1
L9_7_SH2000-99-394-c_S54	NC_010007.1

Sample	Ref
L10_1_BC5_S58	NC_024032.1
L10_2_BC6_S59	NC_024032.1
L10_3_BC7_S60	NC_024032.1
L10_4_BC8_S61	NC_024032.1
L10_5_BC9_S62	NC_024032.1
L11_1_SCO_98_S66	NC_024032.1
L11_2_SCO_81_S67	NC_024032.1
L11_3_SCO_114_S68	NC_024032.1
L11_6_SCO_116_S69	NC_024032.1
L1_7_BC2_Sxx_GTCCGGC+CTCGAT	NC_024032.1
L2_8_BC1_S77	NC_024032.1
L3_12_MB6_S7	NC_024032.1
L3_15_MB12_S10	NC_024032.1
L3_1_MB15_S1	NC_024032.1
L3_5_SM3_S2	NC_024032.1
L4_10_BC14_S20	NC_024032.1
L4_12_MD1_S22	NC_024032.1
L4_14_SM2_S24	NC_024032.1
L4_15_BC10_S25	NC_024032.1
L4_1_MB14_S11	NC_024032.1
L4_3_MB16_S13	NC_024032.1
L4_6_MB2_S16	NC_024032.1
L6_5_SH1979-4-6843-d_S43	NC_024032.1
L6_8_SH2000-99-394-e_S45	NC_024032.1
L9_13_SH2000-99-390-b_S55	NC_024032.1
L9_14_SH1979-4-6843-a_S56	NC_024032.1

Sample	Ref
L0_4_EG9_S84	NC_012928.1

Sample	Ref
L0_5_EG10_S85	NC_025648.1
L7_14_22d_S48	NC_025648.1

repartition_sample_by_ref_majoritaire.ods

Regroupement des échantillons par référence majoritaire avec 39 mito génomes

```
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/01_Multimapping $ cut -f 2 -d ':' RESULTATS_MAJORITAIRES | sort -u
NC_001960.1
NC_006531.1
NC_012928.1
NC_018341.1
NC_020762.1
NC_024032.1
```

Comme convenu en réunion le jeudi 11 septembre, le pipeline Eager est à lancer uniquement pour les échantillons classés pour ces 2 références :

NC_024032.1 : Salmo trutta mitochondrion, complete genome

NC_001960.1 : Salmo salar mitochondrion, complete genome

Regroupement des échantillons par référence majoritaire avec 39 mito génomes

```
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/01_Multimapping $ cut -f 1,2 -d ':' RESULTATS_MAJORITAIRES | grep NC_001960.1
L0_2_EG5_S82:NC_001960.1
L0_8_SH2_S88:NC_001960.1
L10_11_SCO_1338_S63:NC_001960.1
L10_13_SCO_2764_S65:NC_001960.1
L1_1_AUD20212-2_Sxx_GACGATT+TCGCAGG:NC_001960.1
L1_2_BC4_Sxx_AACCTGC+CTCTGCA:NC_001960.1
L1_4_AUD11764-47_Sxx_GCCTACG+GGATCAA:NC_001960.1
L2_10_MPHB2_S78:NC_001960.1
L2_11_MPHB3_S79:NC_001960.1
L2_12_MPHB4_S80:NC_001960.1
L2_1_EG1_S71:NC_001960.1
L2_2_EG2_S72:NC_001960.1
L2_3_EG3_S73:NC_001960.1
L4_11_MD2_S21:NC_001960.1
L4_13_MB5_S23:NC_001960.1
L4_4_MB9_S14:NC_001960.1
L4_5_MD5_S15:NC_001960.1
L4_8_MB4_S18:NC_001960.1
L4_9_MD12_S19:NC_001960.1
L5_12_MD9_S37:NC_001960.1
L5_13_BC11_S38:NC_001960.1
L5_14_BC12_S39:NC_001960.1
L5_15_BC13_S40:NC_001960.1
L5_1_MD15_S26:NC_001960.1
L5_2_MD7_S27:NC_001960.1
L5_3_MD8_S28:NC_001960.1
L5_4_MD4_S29:NC_001960.1
L5_5_MD10_S30:NC_001960.1
L5_6_MB8_S31:NC_001960.1
L5_7_MD3_S32:NC_001960.1
L5_8_MD16_S33:NC_001960.1
L8_15_HTMK99-XXII-0-4287_S49:NC_001960.1
L9_1_26-52_S50:NC_001960.1
L9_2_2777-17_S51:NC_001960.1
L9_4_SM1_S52:NC_001960.1
others_underterminated_R1:NC_001960.1
Undetermined_S0:NC_001960.1
```

Regroupement des échantillons par référence majoritaire avec 39 mito génomes

```
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/01_Multimapping $ cut -f 1,2 -d ':' RESULTATS_MAJORITAIRES | grep NC_024032.1
L0_1_EG4_S81:NC_024032.1
L0_3_EG6_S83:NC_024032.1
L0_6_EG13_S86:NC_024032.1
L0_7_OLG1_S87:NC_024032.1
L10_12_SCO_2213_S64:NC_024032.1
L10_1_BC5_S58:NC_024032.1
L10_2_BC6_S59:NC_024032.1
L10_3_BC7_S60:NC_024032.1
L10_4_BC8_S61:NC_024032.1
L10_5_BC9_S62:NC_024032.1
L11_1_SCO_98_S66:NC_024032.1
L11_2_SCO_81_S67:NC_024032.1
L11_3_SCO_114_S68:NC_024032.1
L11_6_SCO_116_S69:NC_024032.1
L1_7_BC2_Sxx_GTCCGGC+CTCGATG:NC_024032.1
L2_6_TP1_S75:NC_024032.1
L2_7_TP2_S76:NC_024032.1
L2_8_BC1_S77:NC_024032.1
L3_10_MB13_S5:NC_024032.1
L3_11_MB18_S6:NC_024032.1
L3_12_MB6_S7:NC_024032.1
L3_13_MB11_S8:NC_024032.1
L3_14_MB10_S9:NC_024032.1
L3_15_MB12_S10:NC_024032.1
L3_1_MB15_S1:NC_024032.1
L3_5_SM3_S2:NC_024032.1
L3_6_MB3_S3:NC_024032.1
L3_9_MB17_S4:NC_024032.1
L4_10_BC14_S20:NC_024032.1
L4_12_MD1_S22:NC_024032.1
L4_14_SM2_S24:NC_024032.1
L4_15_BC10_S25:NC_024032.1
L4_1_MB14_S11:NC_024032.1
L4_2_MD13_S12:NC_024032.1
L4_3_MB16_S13:NC_024032.1
L4_6_MB2_S16:NC_024032.1
L4_7_MB1_S17:NC_024032.1
L5_10_MD11_S35:NC_024032.1
L5_11_MD14_S36:NC_024032.1
L5_9_MD6_S34:NC_024032.1
L6_11_SH2000-99-394-g_S46:NC_024032.1
L6_13_SH2000-99-394-f_S47:NC_024032.1
L6_3_SH1979-4-6843-e_S41:NC_024032.1
L6_4_SH1979-4-6843-c_S42:NC_024032.1
L6_5_SH1979-4-6843-d_S43:NC_024032.1
L6_6_SH2001-106_11-b_S44:NC_024032.1
L6_8_SH2000-99-394-e_S45:NC_024032.1
L9_13_SH2000-99-390-b_S55:NC_024032.1
L9_14_SH1979-4-6843-a_S56:NC_024032.1
L9_15_SH2000-99-390-c_S57:NC_024032.1
L9_6_SH1979-4-6843-b_S53:NC_024032.1
L9_7_SH2000-99-394-c_S54:NC_024032.1
```

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.

Résultats du pipeline pour S. salar et S. trutta

```
/work/project/crucial/PALEOFISH/02_eager_39_mito_genomes/crucial_nextflow_VersionPaleofish.config

/work/project/crucial/PALEOFISH/02_eager_39_mito_genomes/NC_001960.1:
01_script_eager_part1_NC_001960.1.sh   03_eager_only_genotyping                documentation  input_NC_001960.1.tsv  pipeline_info       reference_genome    slurm-24635376.out
02bis-bam_tsv.sh                       adapterremoval                          endorspy       mapdamage              pipeline_trace.txt  reheader_bam        slurm-24635526.out
02_rehead_bam_files_v2_NC_001960.1.sh  crucial_nextflow_VersionPaleofish.config FastP          mapping                preseq              samtools            work
02ter_tsv_modif.sh                     deduplication                           fastqc         multiqc                qualimap            slurm-24562555.out

/work/project/crucial/PALEOFISH/02_eager_39_mito_genomes/NC_024032.1:
01_script_eager_part1_NC_024032.1.sh   03_eager_only_genotyping  FastP                L4_6_MB2_slurm-24562820.out        pipeline_trace.txt  work
02bis-bam_tsv.sh                       adapterremoval            fastqc               OLD_with_Sxx_input_NC_024032.1.tsv  reference_genome
02_rehead_bam_files_v2_NC_024032.1.sh  documentation             input_NC_024032.1.tsv  pipeline_info                    slurm-24635219.out
smaman@genobioinfo1 /work/project/crucial/PALEOFISH/02_eager_39_mito_genomes/NC_024032.1 $
```