

une image = un projet

AG CATI BIOS4BioI 2022



Rappel : qu'est-ce qu'une production CATI ?

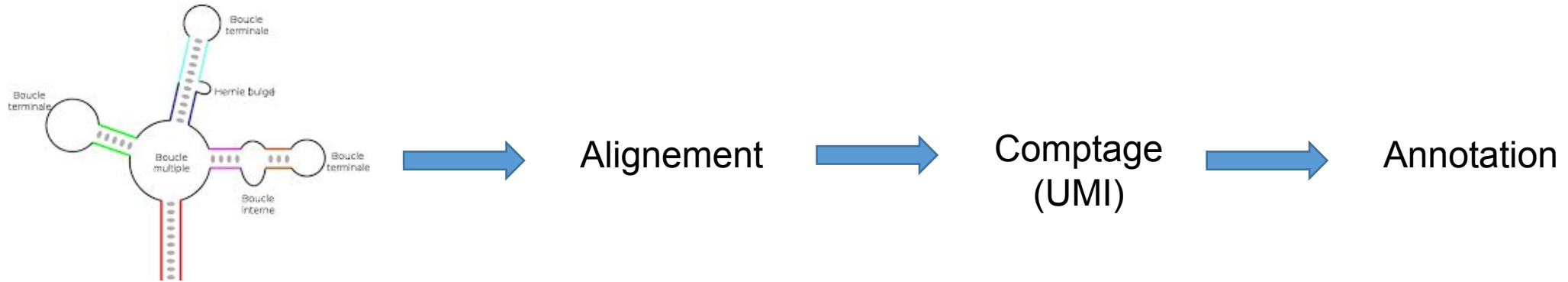
Les productions du CATI se définissent par des productions qui

- émanent de discussions/idées/rerelations d'agents au sein du CATI
 - sont des projets menés par plusieurs membres du CATI
-
- Une production CATI peut être une production issue d'une plateforme
 - Une personne/un groupe peut indiquer dans ce document plusieurs projets mais une diapo = une production
 - Modèle demandé = diapo 3
 - Exemple = diapo 4

pipelines, développements,
benchmarking

Pipeline PAQMir

Chaîne de traitement dédiée à l'analyse et l'annotation des petits ARNs



> [Epigenetics Chromatin](#). 2021 May 24;14(1):24. doi: 10.1186/s13072-021-00397-5.

Dynamics of cattle sperm sncRNAs during maturation, from testis to ejaculated sperm

Eli Sellem¹, Sylvain Marthey^{2,3}, Andrea Rau^{2,4}, Luc Jouneau^{5,6}, Aurelie Bonnet⁷,
Chrystelle Le Danvic⁷, Benoît Guyonnet⁸, Hélène Kiefer^{5,6}, Hélène Jammes^{5,6}, Laurent Schibler⁷

RESEARCH

Open Access

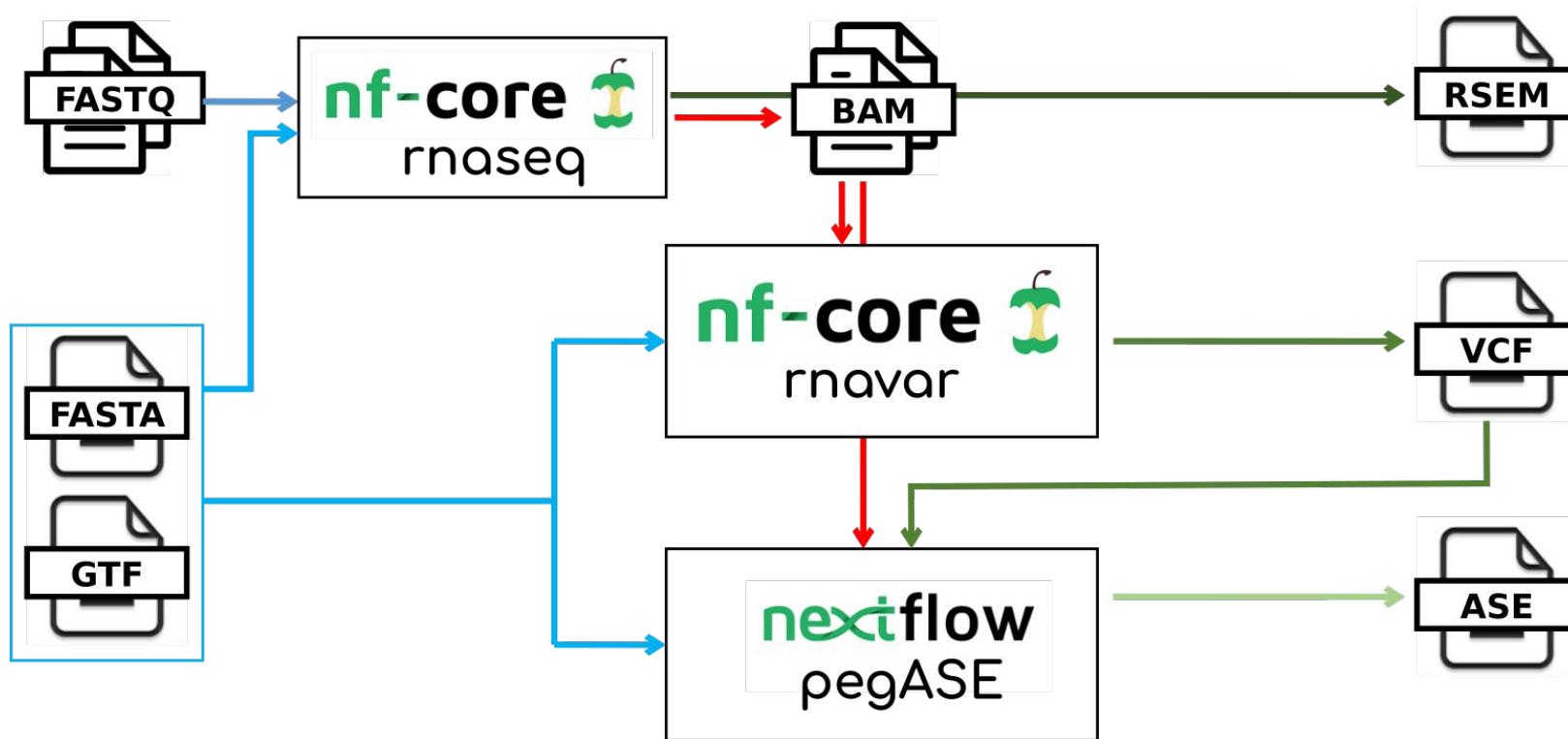
A comprehensive overview of bull sperm-borne small non-coding RNAs and their diversity across breeds

Eli Sellem^{1*}, Sylvain Marthey², Andrea Rau², Luc Jouneau^{3,4}, Aurelie Bonnet¹, Jean-Philippe Perrier^{3,4},
Sébastien Fritz^{1,2}, Chrystelle Le Danvic¹, Mekki Boussaha², Hélène Kiefer^{3,4}, Hélène Jammes^{3,4}
and Laurent Schibler¹

Auteurs : Sylvain Marthey, **Anne Frambourg**, Luc Jouneau, Eli Sellem, Valentin Costes

Analyse de variants RNA-seq 100% Nextflow

Utilisation / développement / amélioration de 3 pipelines pour l'analyse de variants à partir de données RNA-seq



Utilisation de nf-core rnaseq

Amélioration de nf-core rnavar
- Mapping en 2 passes amélioré
- Sortie gvcf

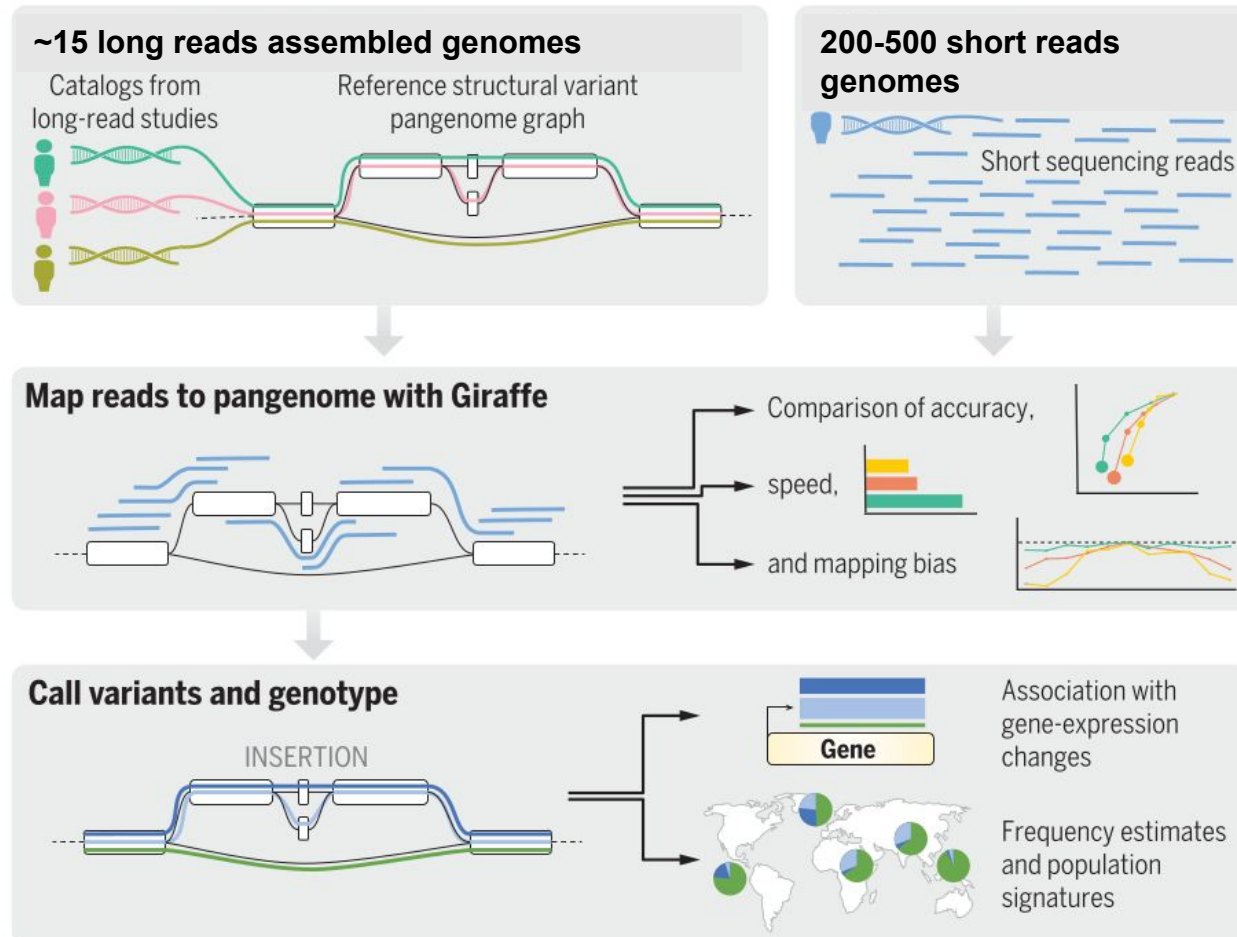
Création de pegASE
- Sur la base du travail (SnakeMake) de Maria Bernard
- Basé sur phASER
- Au standard nf-core

Validation à grande échelle sur les données du projet EFFICACE et CO-LOcATION

Auteurs : Maria Bernard, Mathieu Charles, Cervin Guyomar, Sarah Maman

WoodySV - Pangoak

Chaîne de traitement pour la construction de graphes de pangénomes afin de déterminer le rôle des variants structuraux dans l'adaptation locale chez le chêne et la domestication chez l'abricotier.



Modifié d'après Siren et al. 2021
10.1126/science.abg8871

Auteurs : Sukanya Denni, Quynh Trang-Bui, Véronique Decroocq, **Ludovic Duvaux**

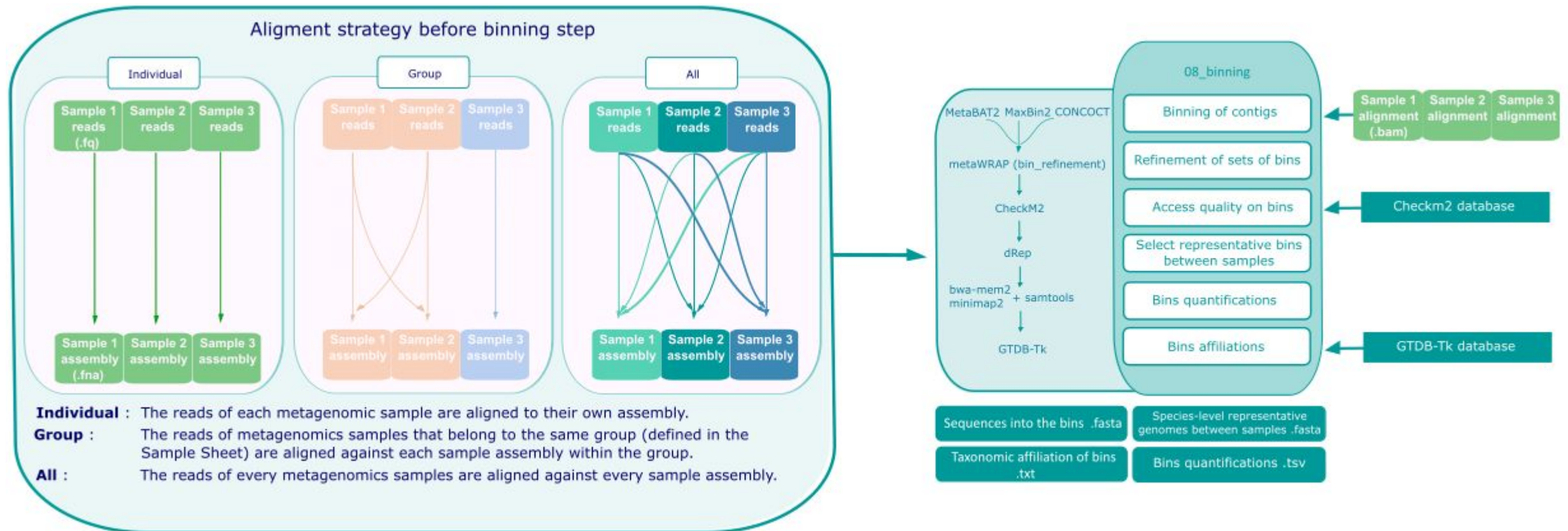
Ajout du binning dans le workflow metagWGS



metagWGS 2.3 permet l'analyse des short-reads ou des long-reads HiFi PacBio.

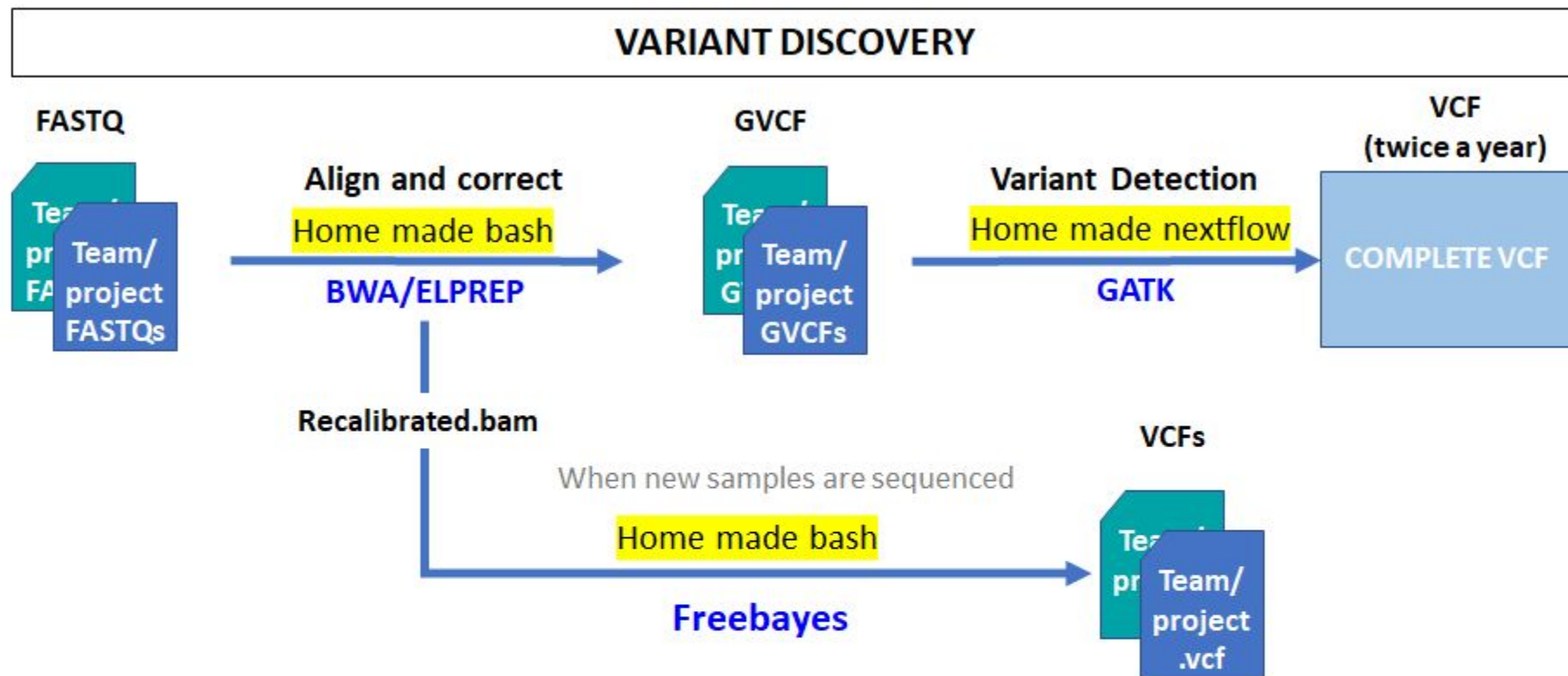
Nouvelles fonctionnalités majeures : le binning des contigs couplé à la possibilité d'aligner plusieurs échantillons sur plusieurs assemblages.

Amélioration des performances de l'étape de raffinement des bins : temps d'exécution divisé par 7.



Auteurs : Maina Vienne, Vincent Darbot, Jean Mainguy, Céline Noirot, Geraldine Pascal and Claire Hoede

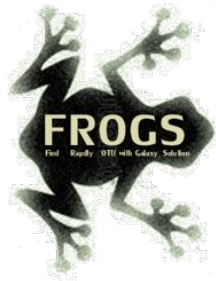
VarPipeline workflow description



Recherche de variants et génotypage sur de larges et nombreux jeux de données (nb animaux incrémentés en fonction de l'arrivée de nouveaux lots, ex: porc >100 animaux actuellement) de façon récurrente et complète à chaque nouvelle version d'assemblage.

Du fait de la taille des lots à traiter, il a été nécessaire de générer de nouvelles chaînes de traitement en s'inspirant de celles existantes, en mettant l'accent sur l'économie de l'espace de stockage phagocyté par les fichiers temporaires de ces dernières, et en optimisant le parallélisme.

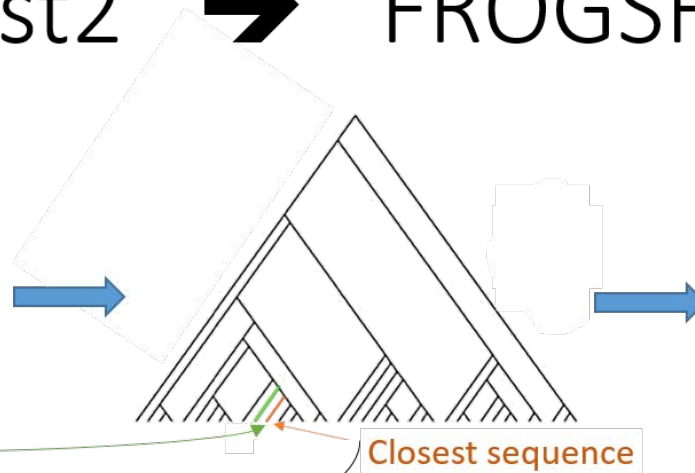
Scripts bash / Nextflow / Slurm / Genotoul / Singularity (portabilité)



+ Picrust2 → FROGSFUNC

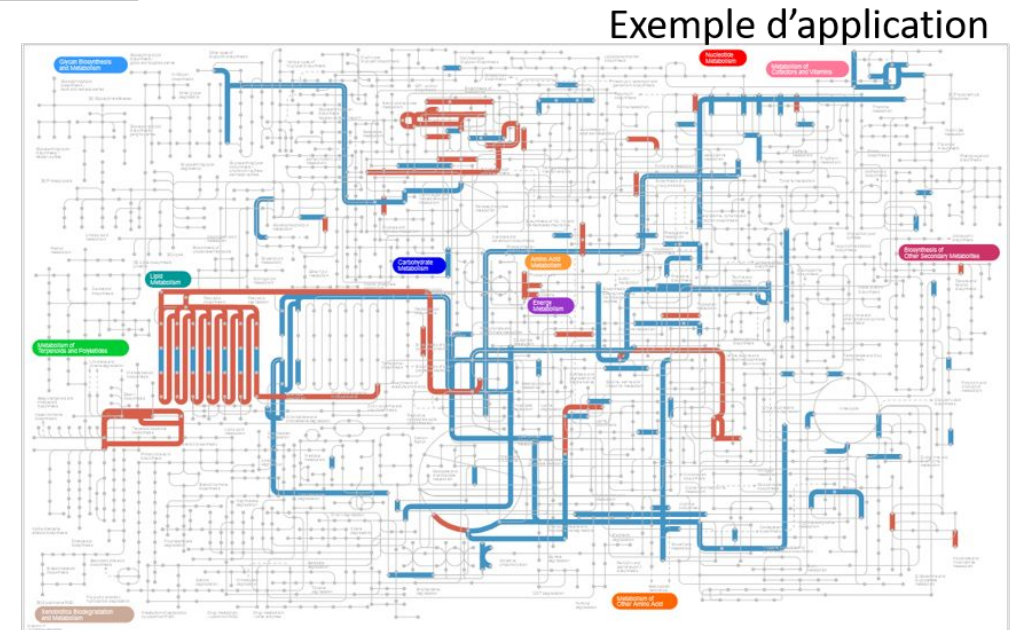
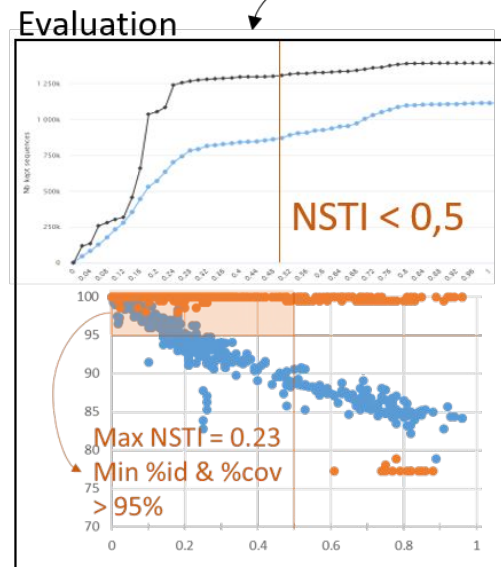


	Taxonomic affiliation	sequence	Sample1	Sample2	Sample3
OTU1	Species A	>ACCGC...	3500	6300	210
OTU2	Species B	>TGCGG...	0	460	36
OTU3	Species C	>ATTGT...	400	700	500



	Name	Sample1	Sample2	Sample3
Func1	Func A	1301	1695	2206
Func2	Func B	31	19	2
Func3	Func C	10221	7684	10419

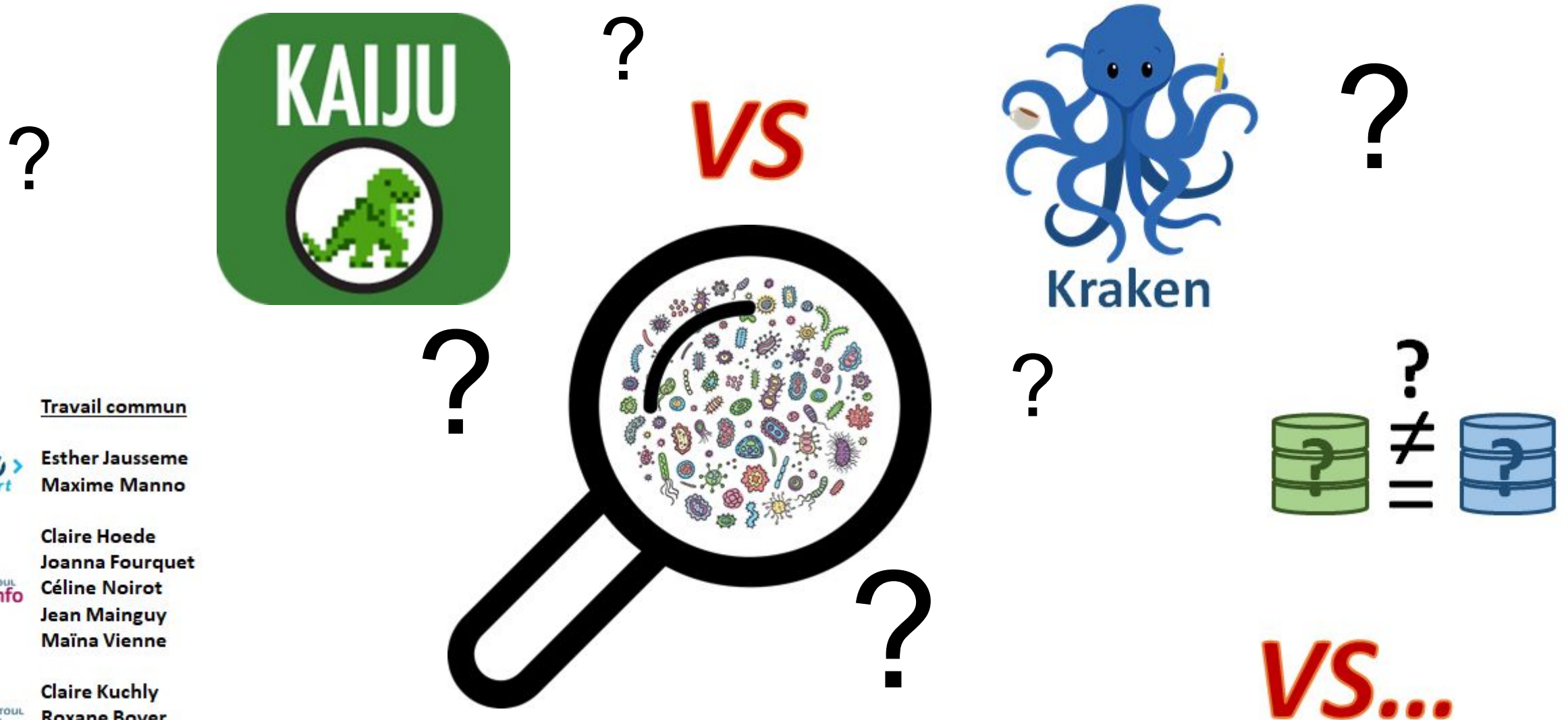
- Intégration des outils PICRUST2 (4 outils) pour la prédiction et quantification fonctionnelle des écosystèmes microbiens.
- Application au microbiote ceacal de poules divergentes sur l'efficacité alimentaire.



Auteurs : Vincent Darbot, Moussa Samb, Maria Bernard, Olivier Rué, Géraldine Pascal

Tests d'outils pour l'affiliation taxonomique de données WMS ou WGS

-> Échanges autour de l'affiliation taxonomique de données de **Whole Metagenome Sequencing** ou **Whole Genome Sequencing** en vue de faire de la recherche de contamination ou étudier la composition d'une communauté complexe.



Travail commun



Esther Jausseme
Maxime Manno

Claire Hoede
Joanna Fourquet
Céline Noirot
Jean Mainguy
Maïna Vienne



Claire Kuchly
Roxane Boyer
Eden Darnige

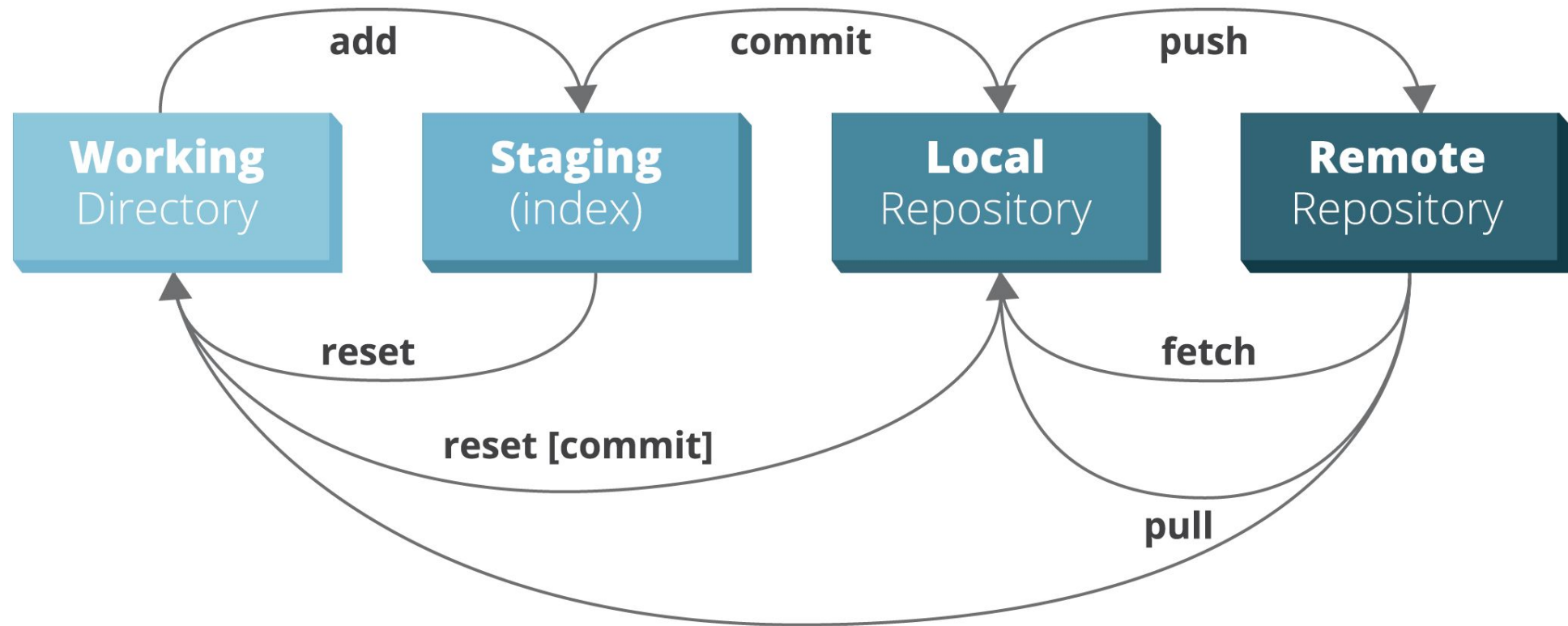


Moyens : Plusieurs réunions effectuées, discussions informelles et échanges de scripts

autour de la formation

Formation GIT

Organisation de 9 sessions de formation à GIT, à Toulouse et Jouy en Josas. 3 sessions débutant et 6 session avancée. 54 personnes formées dont 30 du CATI BIOS4Biol.

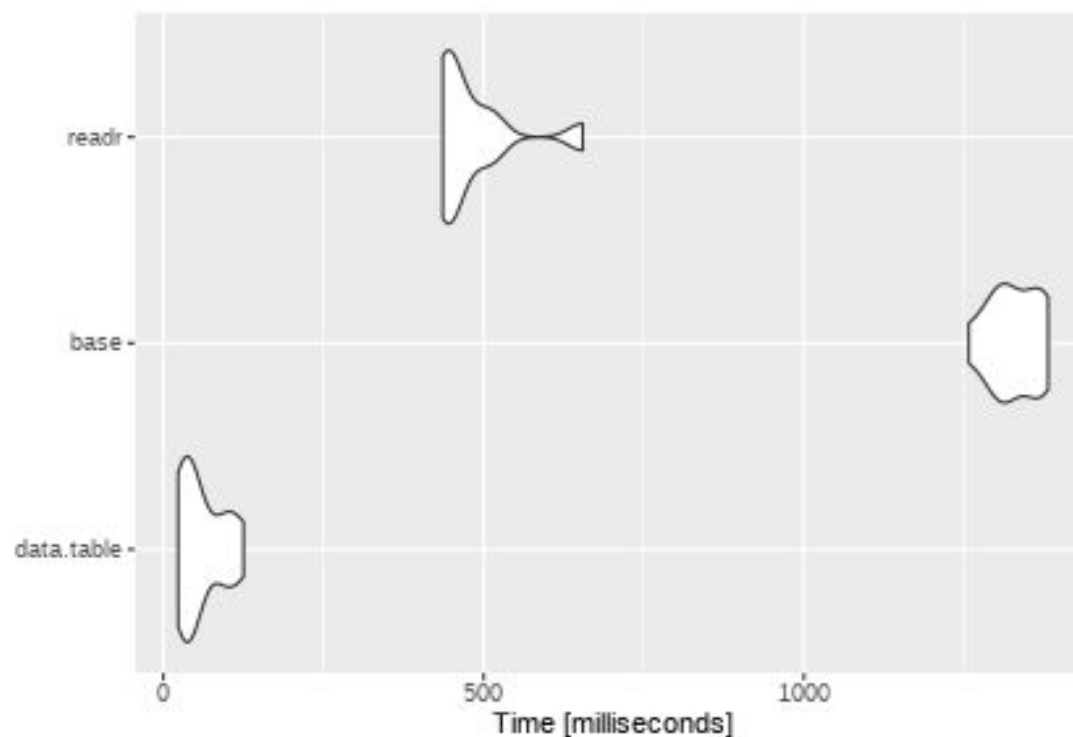
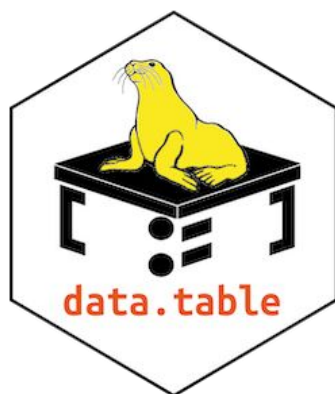


Auteurs : **Céline Noirot**, Estelle Ancelet, **Maria Bernard**, **Luc Jouneau**, formation donnée par Makina Corpus.

Session de formation R avancé : le package data.table

Organisation d'une formation interne (MIAT) sur 1 jour sur le package R data.table.

data.table = « high performance data.frame ». 7 personnes formées dont 3 du CATI bios4biol



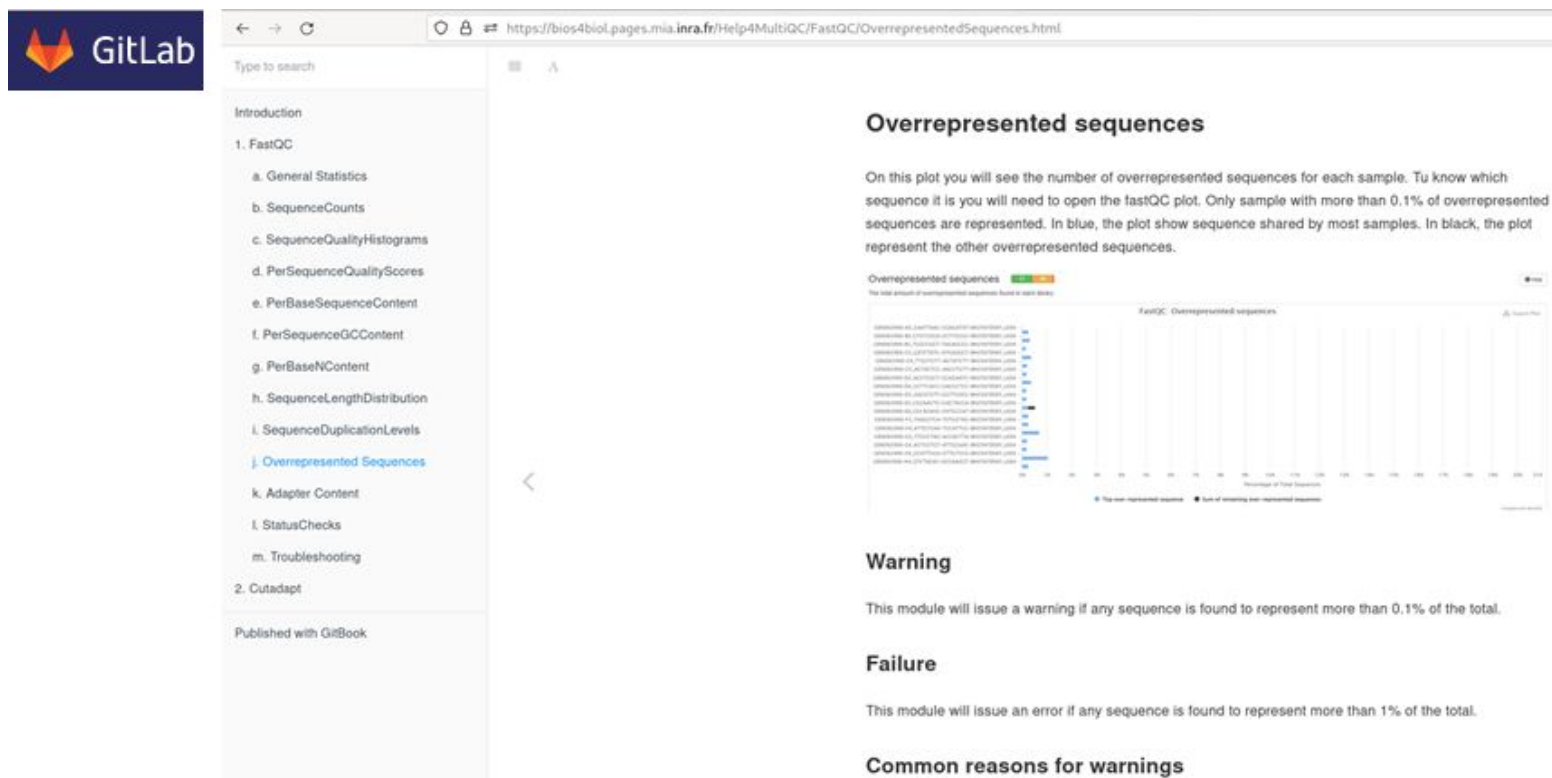
Auteurs : **Paul Tierzan**, Julien Henry et **Elise Maigné**

Help4MultiQC

→ Rédaction collaborative d'un GitBook d'aide à l'interprétation des résultats compilés par multiQCreport à partir de pipelines NGS standards tels que RNAseq, 16S metaGenomics, ...

→ Ajout du lien vers l'aide dans multiQCreport.

<https://bios4biol.pages.mia.inra.fr/Help4MultiQC>




The screenshot shows a web browser displaying a GitBook page. On the left is a navigation sidebar with a table of contents. The main content area on the right is titled 'Overrepresented sequences' and contains a paragraph of text, a legend, a bar chart, and three sections: 'Warning', 'Failure', and 'Common reasons for warnings'.

Table of Contents (Left Sidebar):

- Introduction
- 1. FastQC
 - a. General Statistics
 - b. SequenceCounts
 - c. SequenceQualityHistograms
 - d. PerSequenceQualityScores
 - e. PerBaseSequenceContent
 - f. PerSequenceGCContent
 - g. PerBaseContent
 - h. SequenceLengthDistribution
 - i. SequenceDuplicationLevels
 - j. Overrepresented Sequences**
 - k. Adapter Content
 - l. StatusChecks
 - m. Troubleshooting
- 2. Cutadapt
- Published with GitBook

Overrepresented sequences (Main Content):

On this plot you will see the number of overrepresented sequences for each sample. To know which sequence it is you will need to open the fastQC plot. Only sample with more than 0.1% of overrepresented sequences are represented. In blue, the plot show sequence shared by most samples. In black, the plot represent the other overrepresented sequences.

Overrepresented sequences 

The total amount of overrepresented sequences found in each block.

FastQC: Overrepresented sequences

Warning

This module will issue a warning if any sequence is found to represent more than 0.1% of the total.

Failure

This module will issue an error if any sequence is found to represent more than 1% of the total.

Common reasons for warnings

Auteurs : **Cervin Guyomar, Claire Hoede, Yannick Lippi, Sarah Maman.**

Le groupe FROGS s'agrandit

- Lucas Auer a rejoint « officiellement le groupe de travail FROGS ». support et support-utilisateurs en statistiques, formation en binôme (3 webinaires en 2022), développement du site web et production de pages de tutoriels des outils.
- Perspectives de développements de nouvelles méthodes à intégrer dans FROGS



Auteurs : **Lucas Auer, Vincent Darbot, Olivier Rué, Maria Bernard, Géraldine Pascal**

Formation alignement de lectures courtes et recherche de variants de petite taille



Exercice 5 : preprocessing et calling - GATK

1. Charger le module GATK.

Solution

```
search_module gatk
module load bioinfo/gatk-4.1.7.0
```

1. Marquer les duplicats (GATK MarkDuplicates).

Solution

```
\ls *.bam | perl -ln '$out=$_; $out=~s/\.bam//
s/\.bam$/.bam.dups/; echo $out' > LOGS/%x_%j.out
```

Solution compliquée

```
\ls -l *_R1.fastq.gz | perl -ln '$id=$_; $id=~s/\.*/; $out=$_; $out=~s/\.*/; $r2=$_; $r2=~s/\.*/;
print "bwa mem -R \"\@RG\t\tID:1\t\tSM:$id\t\tPL:illumina\t\tLB:$id\t\tPU:1\" -t4 Gallus_gallus-5.0.dna.toplevel.fa $_ $r2 |\
samtools sort - > $out.bam" > 1_bwamem.jobs
sarray -J 1_bwamem -e LOGS/%x_%j.err -o LOGS/%x_%j.out -c 5 1_bwamem.jobs
```



Programme sur 2 jours :

- qualité fastQC
- aln BWA
- manipulation SAM/BAM
- visualisation IGV
- calling GATK
- formats VCF/gVCF
- annotation SNPsift/SNPeff
- pipeline nf-core/sarek

Auteurs : Céline Noirot, Philippe Bardou, Cédric Cabau

Participation à l'hackathon bioinformatique inter-CATI omiques

- du 5 au 7 octobre 2021 à Sète
- 5 CATIs : BARIC, BIOS4Biol, BOOM, GREP, PlantBreed
- 4 ateliers :

Reproductibilité des workflows



DataScience



Text-mining



Développement d'API



Auteurs : 69 participants dont **21** du CATI Bios4Biol

https://forgemia.inra.fr/inter_cati_omics/hackathon_octobre2021

infra/systeme/BD

Acquisition cluster de calcul 2022

- Consultation des fournisseurs info.
- Rédaction du cahier des charges
- Analyse et comparaison des offres
- Signature du marché mi-juin
- Livraison / installation octobre 2022

- Matériel LENOVO
- Intégration AXIANS
- Hébergement DROCC

- 5000 cœurs, 16G RAM / cœur
- Stockage Spectrum (/work) 2.5 Po
- Nouveaux services : GPU, OOD



Auteurs : **Marie-Stéphane Trotard, Patrice Dehais, Didier Laborie**

Remplacement OTRS (gestionnaire de tickets)

- ❑ Outil libre community edition non maintenu depuis plusieurs mois, soumis à licence
- ❑ Utilisé par PF Bioinfo/SIGENAE/PlaGe (2 instances)
 - ❑ Evaluation / mise en place forks OTRS



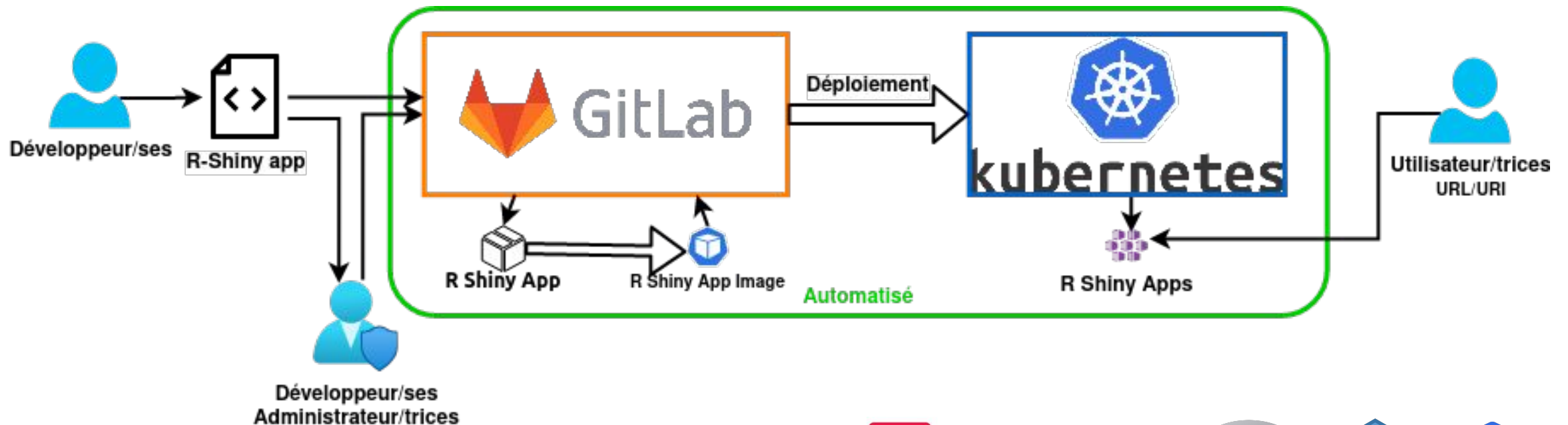
Auteurs : **Gérald Salin, Didier Laborie, Marie-Stéphane Trotard**

Lancement du service SK8



Infrastructure pour déployer des application R shiny.

16 demandes d'hébergement depuis ouverture à INRAE en avril 2022 :



Auteurs : ~30 personnes, 9 CATI représentés.

Porté par Jean-François Rey (CATI IMOTEP), pour le CATI Bios4Biol : **Elise Maigné**

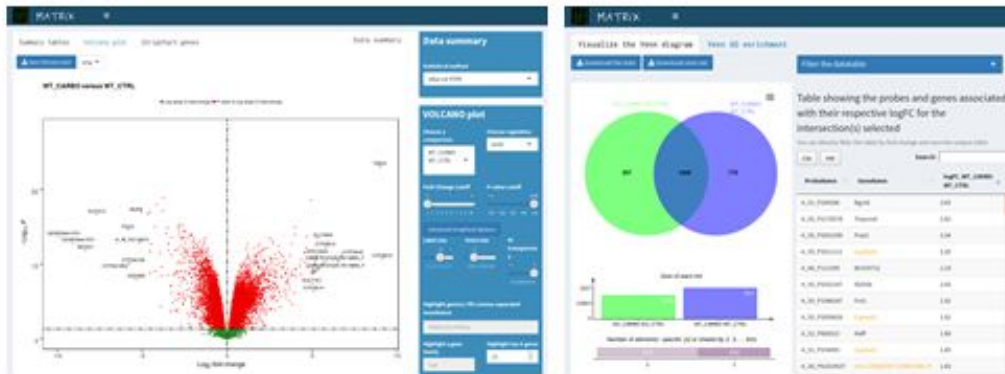
Déploiement de l'app matrix sur SK8



Pipeline automatisé dépôt gitlab SK8

<https://forgemia.inra.fr/sk8/sk8-apps/get-trix/matrixapp>

✓ Explore data/results



✓ Functional analysis





Actions :

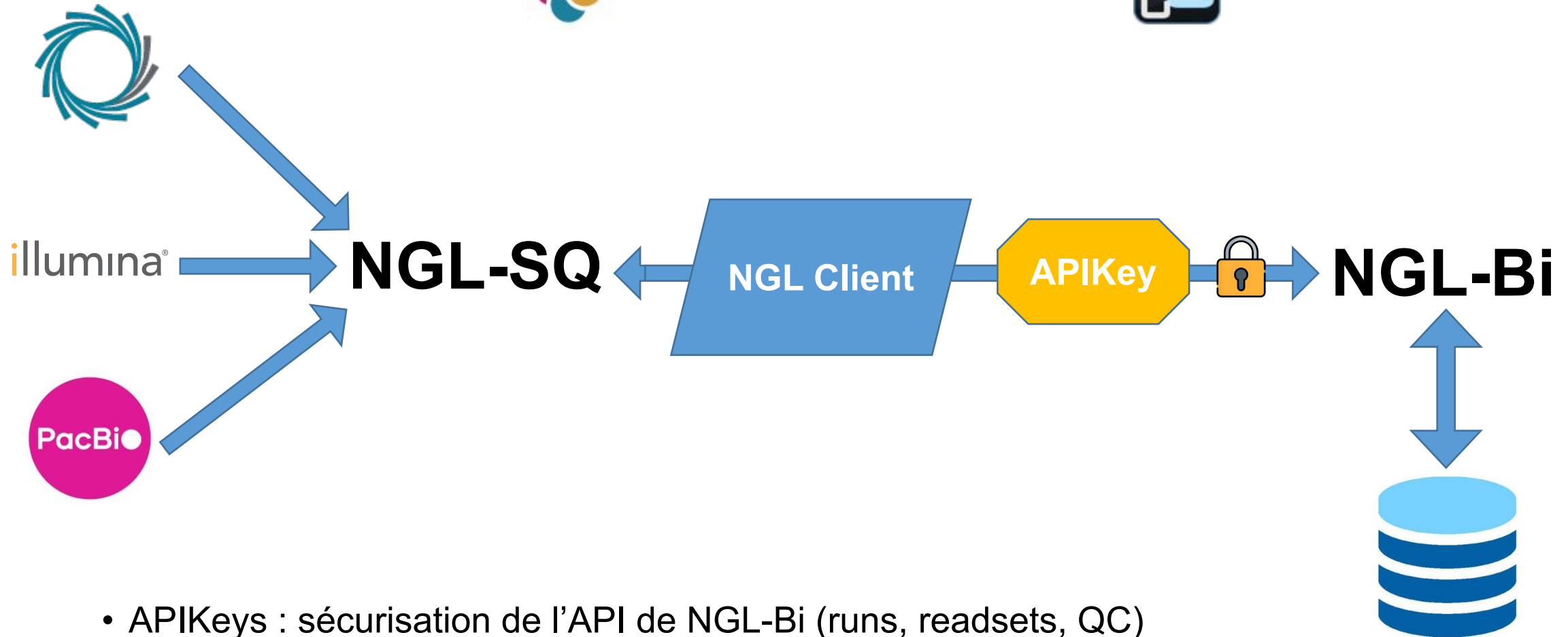
- Adaptation du Dockerfile
- Aide debug


Tests à venir :

- Authentification utilisateurs
- Montage volume de données

Auteurs : Yannick Lippi , Jean-François Rey (CATI IMOTEP) 

Évolution de  **NGL** | succession à 



- APIKeys : sécurisation de l'API de NGL-Bi (runs, readsets, QC)
- Intégration de QC : MultiQC, Krona, MinionQC, SMRTLink 

Auteurs : Eden Darnige, Claire Kuchly, Gerald Salin, Archimede Abdias Towe-Patipe, Jules Sabban, Céline Noirot, Romain Therville

Une base de données de fichiers

Gérer des fichiers de données hétérogènes (phénotypes, séquences, ...)

DATA

EXPLORE EFFICACE DATABASE



Show/Hide columns

Configure the table below by selecting column name(s)

Nothing selected

Reset

Add optional column(s) to the table.

Reset visible columns.



Search

Filter on specific field(s)

Nothing selected

Nothing selected

Enter a value.

Select a field.

Choose an operator.

Enter a value.



Facet

Explore data by facet

Datatype 5

- Individual 1
- Pedigree 1
- PhenoBW 1
- PhenoEggProductionByDay 1
- PhenoEggQuality 1

Filter

Project 1

Novo 5

Pop 1

A3A3 5

Batch 1

2016.1 5

AgeFactor 2

- NA 3
- 70w 2

Filter



Export

For the selected columns and the filtered rows export data...

Copy

Excel

CSV

The data.

Show 10 entries

Search:

Datatype	Project	Pop	Batch	AgeFactor	FileUrl	Download
Individual	Novo	A3A3	2016.1	NA	./data/upload/Novo_A3A3_2016.1_Individual_NA.csv	↓
Pedigree	Novo	A3A3	2016.1	NA	./data/upload/Novo_A3A3_2016.1_Pedigree_NA.csv	↓
PhenoBW	Novo	A3A3	2016.1	70w	./data/upload/Novo_A3A3_2016.1_PhenoBW_70w.csv	↓
PhenoEggProductionByDay	Novo	A3A3	2016.1	NA	./data/upload/Novo_A3A3_2016.1_PhenoEggProductionByDay_NA.csv	↓

Auteurs : P. Bardou, C. Cabau, M. Gachet, C. Klopp, Sandrine Lagarrigue, F. Lecerf