# DeFIS - Detection pipeline with k-mers analysis to identify species

Sarah Maman, Gaston Rognon, Chloé Bellanger, Léonard Ransan, Patrick Jacques, Christophe Klopp, Régis Debruyne, Brice Ephrem, Myriam Sternberg, Aurélie Manicki, et al.

# DEFIS - DETECTION PIPELINE WITH K-MERS ANALYSIS TO IDENTIFY SPECIES.

DeFIS « Detection of Fauna and flora: Identification of Species ».

## CONTEXT

- DNA highly degraded.
- Quick identification focuses on repeated zones.
- Conventional mapping tools inadequate.

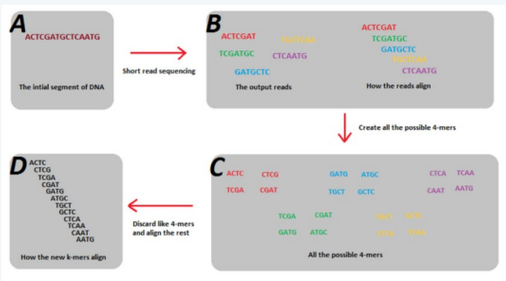
For:

- Modern and ancient DNA.
- Tested on several species and on genomes of different sizes.

## K-MERS DEFINITION

A k-mer is a subsequence of length k nucleotides extracted from a longer sequence.

Source: https://en.wikipedia.org/wiki/K-mer




## DETECTION PIPELINE PRINCIPE

Identifies species by comparing k-mers dictionaries between:

- a chosen reference genome,
- and the studied samples.

## PIPELINE STEPS & RESULTS

This pipeline includes :

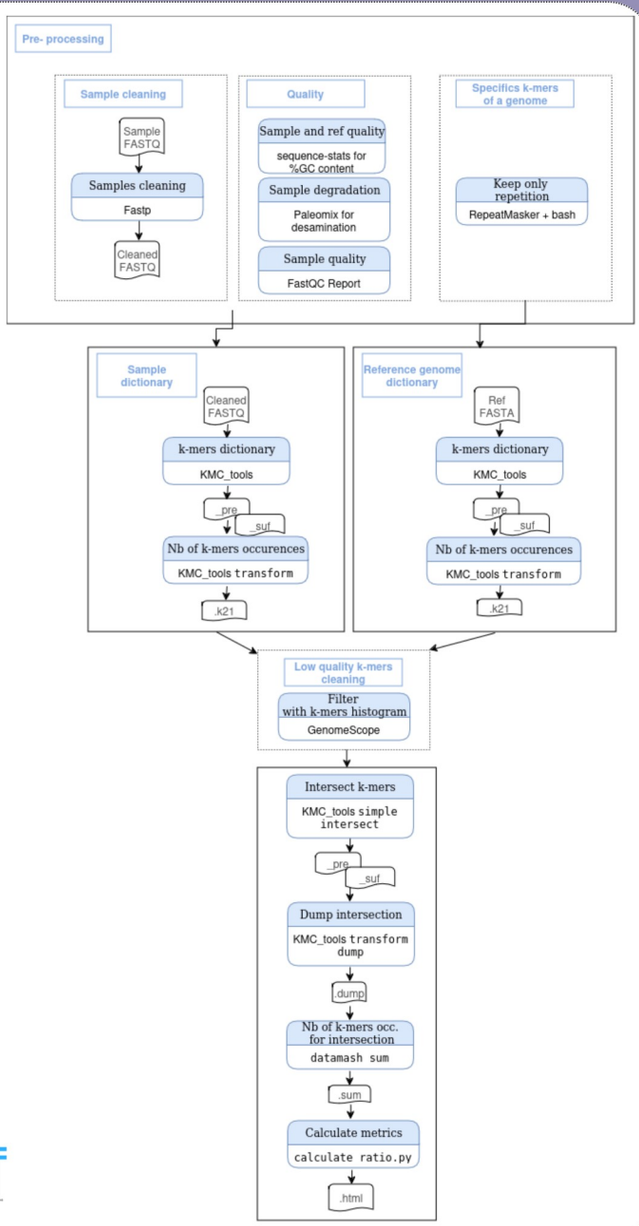
- Pre-process to clean and analyse quality as %GC content or degradation.
- KMC_tools [1] to generate k-mers dictionaries for reference genome(s) ans sample(s).
- Intersect these dictionaries.
- Calculate metrics as Jaccard similarity/dissimilarity indices [2], ratio.
- Determine the closest species.
- HTML report to display results.

This pipeline is being improved to save processing time and computing resources:

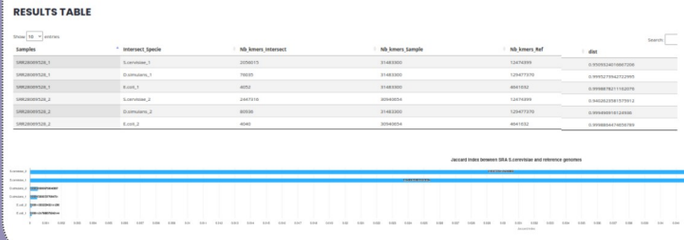
- Study the works of Camila Duitama (akmerbroom & decom) [3] [4]
- To minimize dictionaries sizes: filter out low-quality k-mers, erroneous k-mers.
- To focus on repeated zone:
  - Specific of a genome.
  - Provide an indication of resistance to DNA degradation.
  - Provide some precomputed sets of k-mers.

This potential impact of repeated zones is being studied by comparing small dictionaries, representing those most commonly found within a species.

This pipeline has already been employed with success to taxonomically identify fish archaeological remains, the oldest ones dating from the paleolithic period [5].




**RESULTS TABLE**






Sarah MAMAN1, Gaston ROGNON2, Chloé BELLANGER2, Léonard RANSAN2, Patrick JACQUES3, Christophe KLOPP4, Régis DEBRUYNE5, Brice EPHREM6, Myriam STERNBERG7, Aurélie MANICKI3, Joëlle CHAT3, Philippe BEAREZ5, Natacha NIKOLIC3,8

1 Sigenae, GenPhySE, Université de Toulouse, INRAE, ENVT, 24 chemin de Borde Rouge, F-31326, Castanet Tolosan, France. - 2 Université Paul Sabatier,118 route de Narbonne, 31062 Toulouse, France - 3 UMR INRAE-UPPA ECOBIOP, 64310 Saint Pée sur Nivelle, FRANCE - 4 Sigenae, MIAT, INRAE, 24 chemin de Borde Rouge, F-31326, Castanet Tolosan, France. - 5 CNRS/MNHN, UMR 7209, Paris, France - 6 CNRS, UMR 6566, CReAAH, Rennes, France - 7 CNRS, UMR 7299, Aix-en-Provence, France - 8 CRBE, Aquaeco, Université de Toulouse, France

Corresponding Author: sarah.maman@inrae.fr https://forgemia.inra.fr/sarah.maman-haddad/DeFiS/



1. Marek Kokot, Maciej Długosz, Sebastian Deorowicz, KMC 3: counting and manipulating k-mer statistics, Bioinformatics, Volume 33, Issue 17, September 2017, Pages 2759–2761. (10.1093/bioinformatics/btx304).

2.Shaopeng Liu, David Koslicki, CMash: fast, multi-resolution estimation of k-mer-based Jaccard and containment indices, Bioinformatics, Volume 38, Issue Supplement_1, July 2022, Pages i28–i35,. (10.1093/bioinformatics/btac237).

3.Camila Duitama González, Riccardo Vicedomini, Téo Lemane, Nicolas Rascovan, Hugues Richard, Rayan Chikhi : decOM: similarity-based microbial source tracking of ancient oral samples using k-mer-based methods. Microbiome 2023 Nov; 11(1): 243. (10.1186/s40168-023-01670-3).

4. Duitama González C, Rangavittal S, Vicedomini R, Chikhi R, Richard H : aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets.iScience 2023 Nov; 26(11): 108057. (10.1016/j.isci.2023.108057).

5. Patrick Jacques*, Sarah Maman*, Régis Debruyne, Brice Ephrem, Aurélie Manicki, Myriam Sternberg, Gaston Rognon, Léonard Ransan, Chloé Bellanger, Joelle Chat, Philippe Béarez, Natacha Nikolic. Decoding ancient genomes: Genomics approaches and innovative species recognition pipeline for diadromous fish. 20. Portugaliæ Genetica: DNA - Ancient and New, Mar 2024, Porto, Portugal. 2024. ⟨10.13140/RG.2.2.12906.53442⟩. ⟨hal-04568047⟩