



HAL
open science

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.

► To cite this version:

Sarah Maman, Gaston Rognon, Chloé Bellanger, Léonard Ransan, Patrick Jacques, et al.. DeFIS - Detection pipeline with k-mers analysis to identify species. JOBIM, Jun 2024, Toulouse, France. hal-04728545

HAL Id: hal-04728545

<https://hal.science/hal-04728545v1>

Submitted on 9 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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.

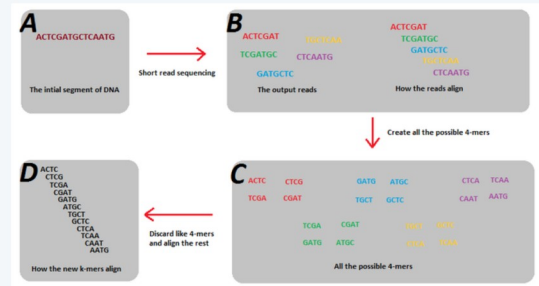
For:

- 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 PRINCIPLE

Identifies species by comparing k-mers dictionaries between:

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

PIPELINE STEPS & RESULTS

This pipeline includes :

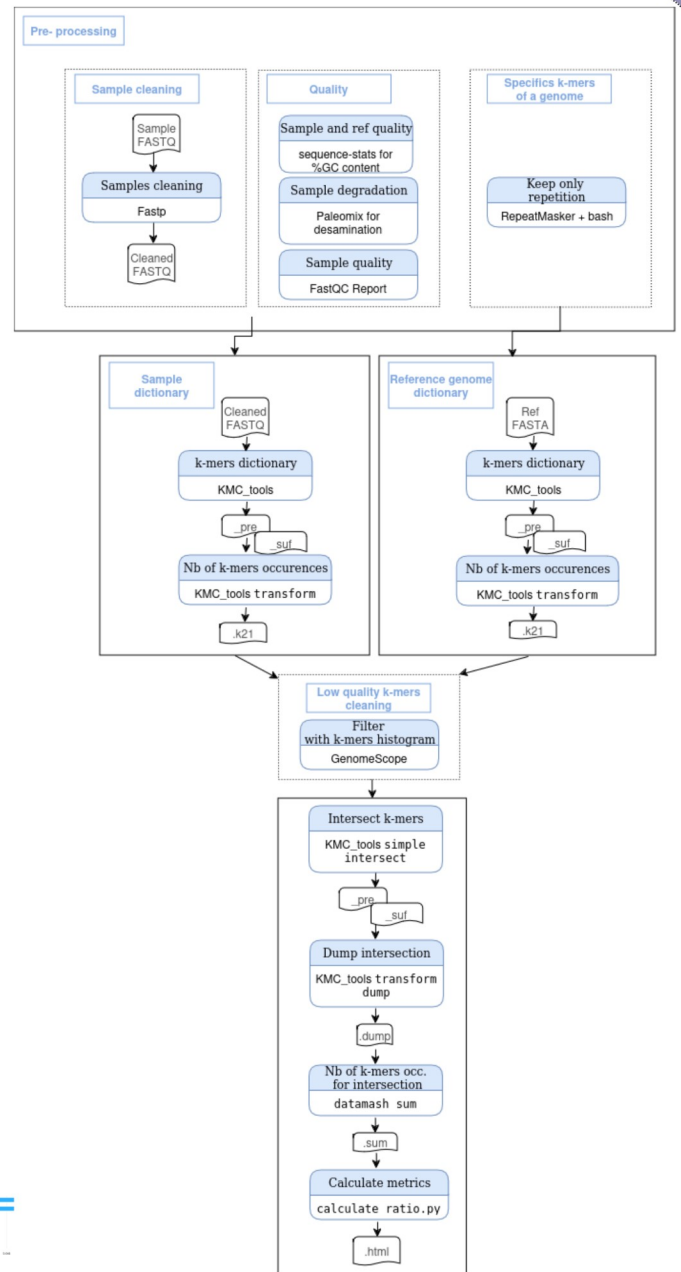
- Pre-process to clean and analyse quality as %GC content or degradation.
- KMC_tools [1] to generate k-mers dictionaries for reference genome(s) and sample(s).
- Intersect these dictionaries.
- Calculate metrics as Jaccard similarity/dissimilarity indices [2], ratio.
- Determine the closest species.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

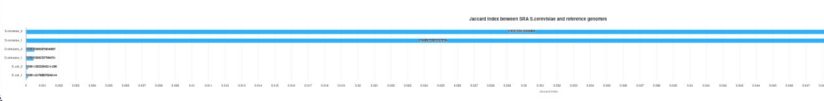
This potential impact of repeated zones is being studied by comparing small dictionaries, representing those most commonly found within a species.

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

Sample	Intersect_Species	Nb_Kmers_Intersection	Nb_Kmers_Sample	Nb_Kmers_Ref	Ratio
Sample000001_1	Sample000001_1	2000000	2000000	2000000	1.000000
Sample000001_2	Sample000001_1	1000000	2000000	2000000	0.500000
Sample000001_3	Sample000001_1	500000	2000000	2000000	0.250000
Sample000001_4	Sample000001_1	250000	2000000	2000000	0.125000
Sample000001_5	Sample000001_1	125000	2000000	2000000	0.062500
Sample000001_6	Sample000001_1	62500	2000000	2000000	0.031250
Sample000001_7	Sample000001_1	31250	2000000	2000000	0.015625
Sample000001_8	Sample000001_1	15625	2000000	2000000	0.007812



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 NIKOLIC^{3,8}
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 ECOBIO, 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, CRéAAH, Rennes, France - ⁷ CNRS, UMR 7299, Aix-en-Provence, France - ⁸ CRBE, Aqualco, 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, Téó 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, Joëlle Chat, Philippe Béarez, Natacha Nikolich. Decoding ancient genomes: Genomics approaches and innovative species recognition pipeline for diadromous fish. *20. Portugaliae Genetica: DNA - Ancient and New*, Mar 2024, Porto, Portugal. 2024. (10.13140/RG.2.2.12906.53442). (hal-04568047)