Detection/characterisation of undescribed genetically modified bacteria by statistical analysis of high throughput sequencing data

Julie HUREL

Supervision :

Fabrice TOUZAIN (bioinformatics)
Stéphanie BOUGEARD (statistics)
Mathieu Rolland (biology)

Viral Genetics and Biosecurity Unit (UGVB)
ANSES Ploufragan

15/03/2023

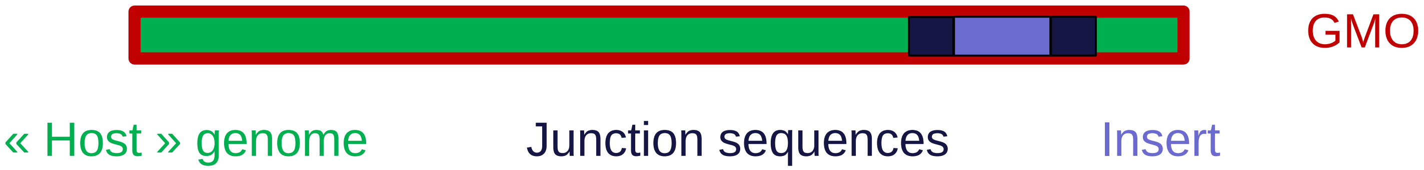
1. Introduction
2. Objectives
3. Data preparation
4. Calculation of the distances
5. Design of the prediction model
6. Results
7. Conclusion
8. Perspectives

1. **Introduction**
2. Objectives
3. Data préparation
4. Calculation of the distances
5. Design of the prediction model
6. Results
7. Conclusion
8. Perspectives

A – Genetically Modified Organism : GMO

- Living being whose genetic material has been modified in a non-natural way

- Simplified description of the structure of a GMO



GMO

« Host » genome

Junction sequences

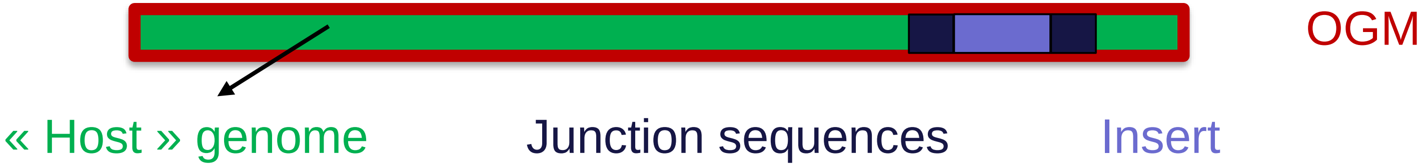
Insert

- Insert : often CDS(s) (coding sequence)

A – Genetically Modified Organism : GMO

- Living being whose genetic material has been modified in a non-natural way

- Simplified description of the structure of a GMO

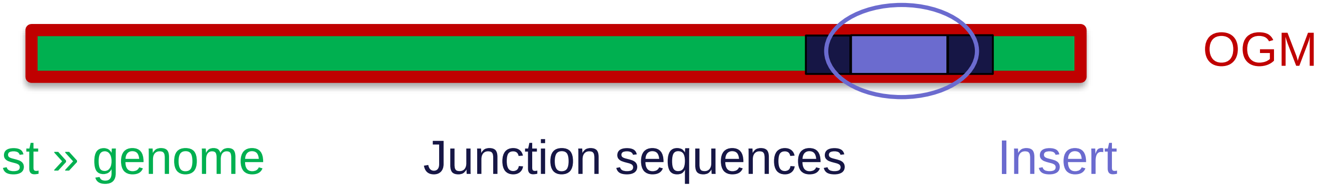


- Insert : often CDS(s) (coding sequence)

A – Genetically Modified Organism : GMO

- Living being whose genetic material has been modified in a non-natural way

- Simplified description of the structure of a GMO



- Insert : often CDS(s) (coding sequence)

B – Existing detection methods

- For known GMO
  - Methods based on protein detection or DNA detection
  - Use of qPCR : greater sensitivity and acurate quantification
  - GMOseek [Morisset et al. 2014]

Credit: Vit Kovalcik/Shutterstock.com

B – Existing detection methods

- For partially known GMO
  - New sequencing techniques coupled with
    - Molecular methods [Fraiture *et al.*, 2017, 2018]
    - Bioinformatics/Statistics [Willems *et al.*, 2016]
  - DNA walking [Fraiture *et al.*, 2015, 2018]
- For unknown GMO
  - No method available so far

C – State of the art

- Current detection limits

	Detection of known and partially known GMO		**Detection of unknown GMO**	
	Prokaryotes	Eukaryotes	Prokaryotes	Eukaryotes
Intergenic sequences	✔	✔	✘	✘
Truncated gene	✔	✔	~	~
Fused gene	✔	✔	~	~
Insertion/deletion in a gene	✔	✔	✘	✘
% of point mutations ≥ 9%	✔	✔	✘	✘
% of point mutations < 9%	✔	✔	✘	✘

1. Introduction
2. **Objectives**
3. Data preparation
4. Calculation of the distances
5. Design of the prediction model
6. Results
7. Conclusion
8. Perspectives

A – General information

- No method available for the detection of unknown GMO

- Creation of a method to address this issue
  - **DUGMO** : Detection of Unknown Genetically Modified Organism [Hurel *et al.*, BMC Bioinformatics 2020] https://github.com/ANSES-Ploufragan/DUGMO

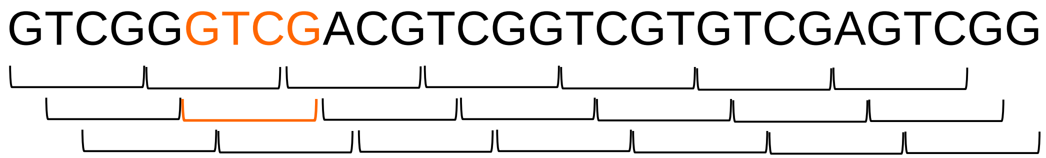
- Basic idea of the method
  - Identify the vocabulary differences between the host genome and the insert

B – General principle of DUGMO

- Particularity : use the CDS of the host genome
- Specific genomic vocabulary
  - Species-specific
  - Composed of words

Nucleotidic sequence

Species 1



GTCGGGTCGACGTCGGTCGTGTCGAGTCGG

4-letter words

B – General principle of DUGMO

- Particularity : use the CDS of the host genome
- Specific genomic vocabulary
  - Species-specific
  - Composed of words

Nucleotidic sequence

Species 1	GTCGGGTCGACGTCGGTCGTGTCGAGTCGG
	Over-represented 4 letter-words
Species 2	AGGCTCAGCTGAGCTTCAGTCGTGTACTAGC

C – Steps in DUGMO

1. Data preparation

    - Cleaning, assembly, annotation pipeline
    - Sorting of the CDSs in the sample
    - Filtering of the databank of known GMOs

2. Characterisation of the « host » genome CDSs based on their over-represented words

3. Calculation of the distances to support the comparaison to the host genome CDSs

4. Design of a prediction model

C – Steps in DUGMO

1. Data preparation

   - Cleaning, assembly, annotation pipeline
   - Sorting of the CDSs in the sample
   - Filtering of the databank of known GMOs

2. Characterisation of the host genome CDSs based on their over-represented words

3. Calculation of the distances to support the comparaison to the host genome CDSs

4. Design of a prediction model

C – Steps in DUGMO

1. Data preparation

    - Cleaning, assembly, annotation pipeline
    - Sorting of the CDSs in the sample
    - Filtering of the databank of known GMOs

2. Characterisation of the host genome CDSs based on their over-represented words

3. Calculation of the distances to support the comparaison to the host genome CDSs

4. Design of a prediction model

C – Steps in DUGMO

1. Data preparation

    - Cleaning, assembly, annotation pipeline
    - Sorting of the CDSs in the sample
    - Filtering of the databank of known GMOs

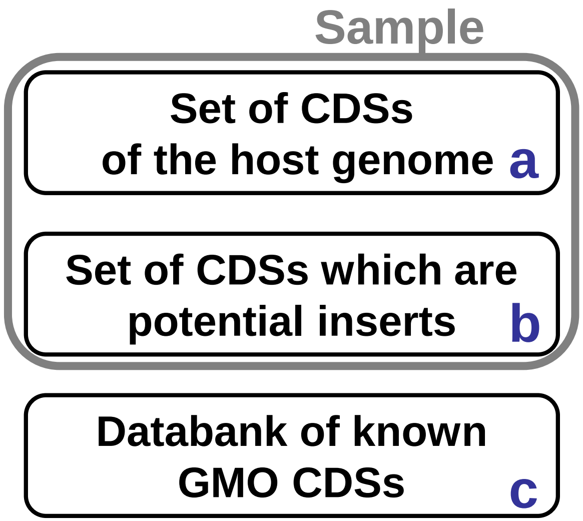
2. Characterisation of the host genome CDSs based on their over-represented words

3. Calculation of the distances to support the comparaison to the host genome CDSs

4. Design of a prediction model

C – Steps in DUGMO

- Data preparation : cleaning/assembly/annotation, sorting and filtering



Sample

Set of CDSs of the host genome a

Set of CDSs which are potential inserts b

Databank of known GMO CDSs c

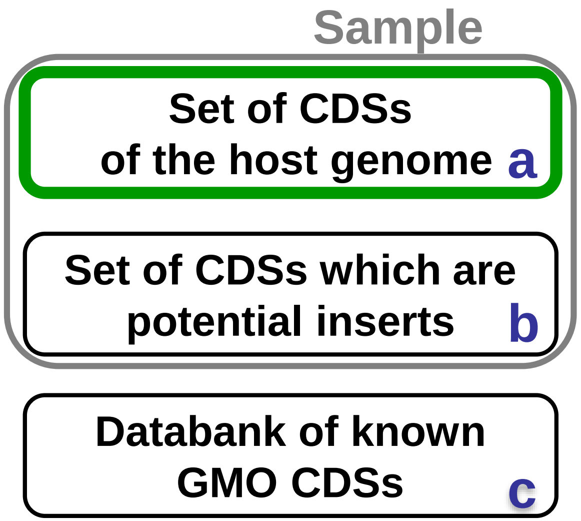
- Characterizes what we are not looking for
- Other CDSs, absent from the pangenome
- Non-species sequence variability

- Computation of distances to characterize :
  - the vocabulary of **a**
  - the vocabulary of **c**
  - the vocabulary of each sequence in the set **b**

- Design of a prediction model

C – Steps in DUGMO

- Data preparation : cleaning/assembly/annotation, sorting and filtering



Sample

Set of CDSs of the host genome **a**

Set of CDSs which are potential inserts **b**

Databank of known GMO CDSs **c**

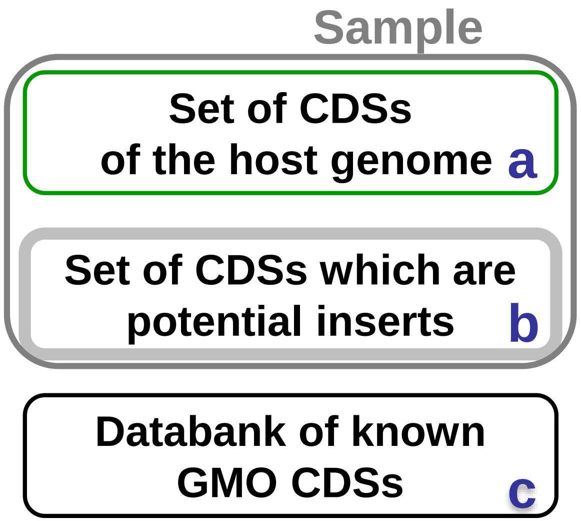
- Characterizes what we are not looking for
- Other CDSs, absent from the pangenome
- Non-species sequence variability

- Computation of distances to characterize :
  - the vocabulary of **a**
  - the vocabulary of **c**
  - the vocabulary of each sequence in the set **b**

- Design of a prediction model

C – Steps in DUGMO

- Data preparation : cleaning/assembly/annotation, sorting and filtering



Sample

Set of CDSs of the host genome a

Set of CDSs which are potential inserts b

Databank of known GMO CDSs c

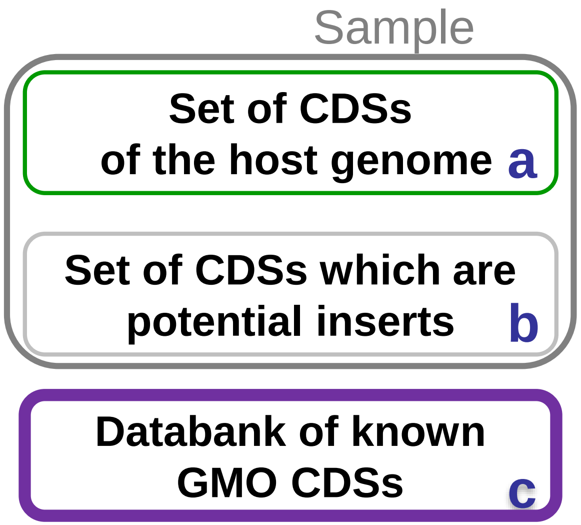
- Characterizes what we are not looking for
- Other CDSs, absent from the pangenome
- Non-species sequence variability

- Computation of distances to characterize :
  - the vocabulary of **a**
  - the vocabulary of **c**
  - the vocabulary of each sequence in the set **b**

- Design of a prediction model

C – Steps in DUGMO

- Data preparation : cleaning/assembly/annotation, sorting and filtering



Sample

Set of CDSs of the host genome a

Set of CDSs which are potential inserts b

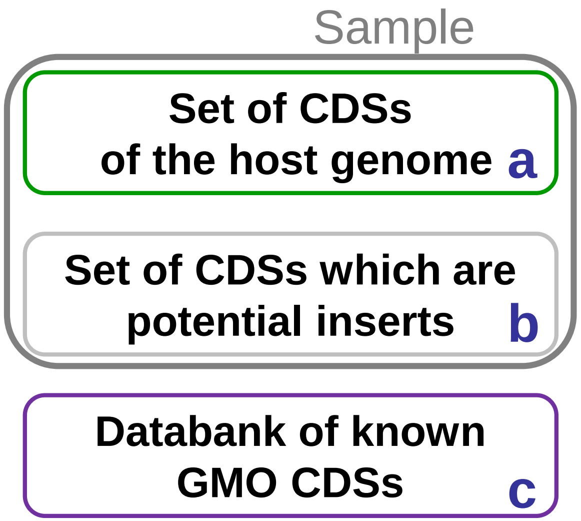
Databank of known GMO CDSs c

- Characterizes what we are not looking for
- Other CDSs, absent from the pangenome
- Non-species sequence variability

- Computation of distances to characterize :
  - the vocabulary of **a**
  - the vocabulary of **c**
  - the vocabulary of each sequence in the set **b**
- Design of a prediction model

C – Steps in DUGMO

- Data preparation : cleaning/assembly/annotation, sorting and filtering



Sample

Set of CDSs of the host genome a

Set of CDSs which are potential inserts b

Databank of known GMO CDSs c

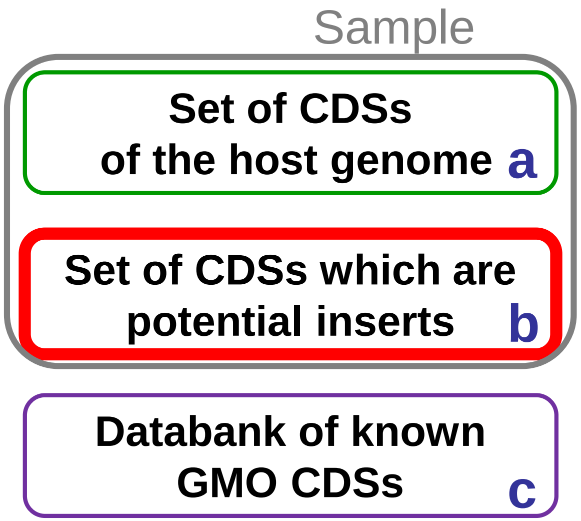
- Characterizes what we are not looking for
- Other CDSs, absent from the pangenome
- Non-species sequence variability

- Computation of distances to characterize :
  - the vocabulary of **a**
  - the vocabulary of **c**
  - the vocabulary of each sequence in the set **b**

- Design of a prediction model

C – Steps in DUGMO

- Data preparation : cleaning/assembly/annotation, sorting and filtering



Sample

Set of CDSs of the host genome a

Set of CDSs which are potential inserts b

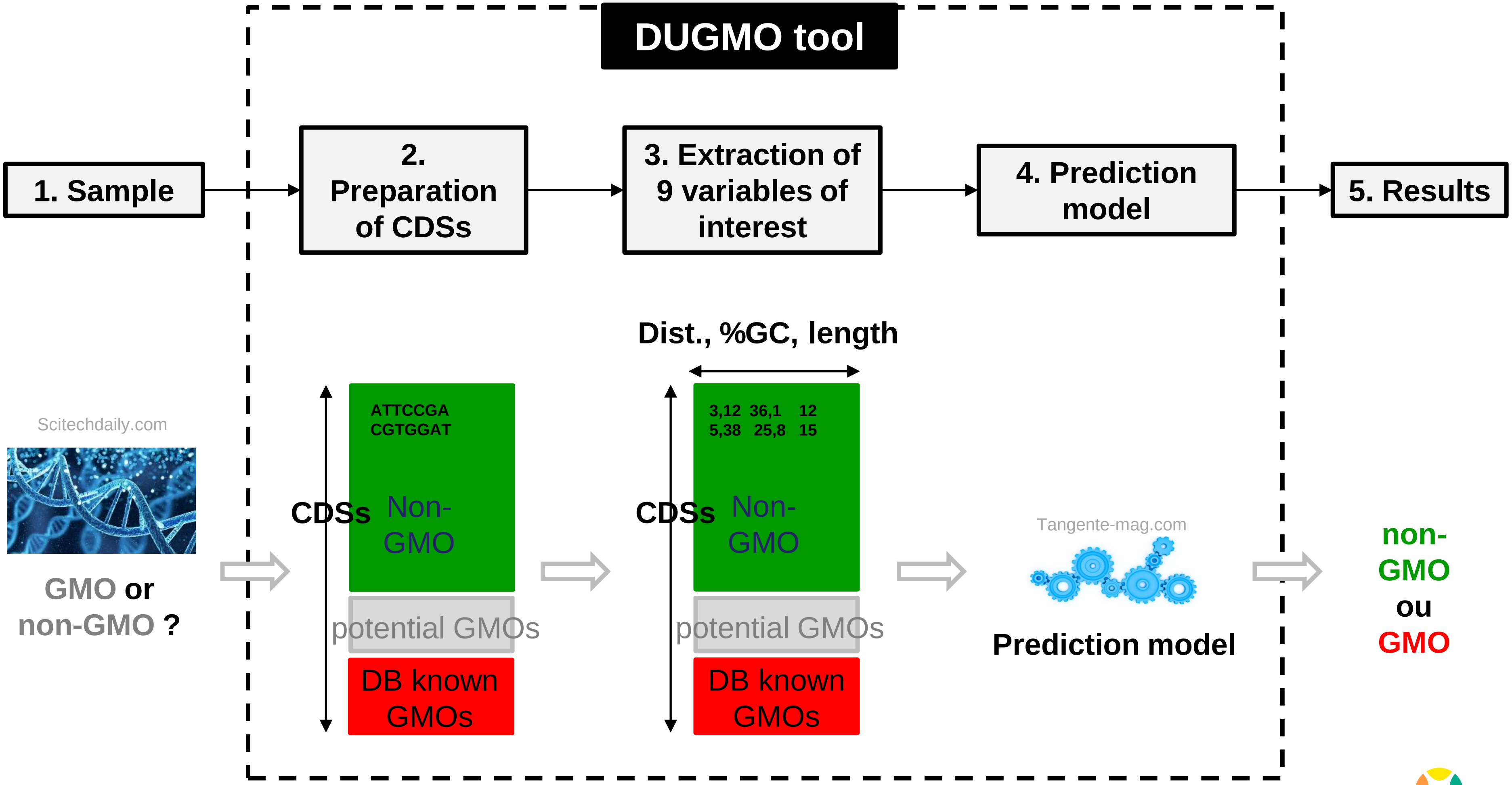
Databank of known GMO CDSs c

- Characterizes what we are not looking for
- Other CDSs, absent from the pangenome
- Non-species sequence variability

- Computation of distances to characterize :
  - the vocavulary of **a**
  - the vocavulary of **c**
  - the vocavulary of each sequence in the set **b**

- Design of a prediction model **=> GMO ?**

C – Steps in DUGMO



DUGMO tool

1. Sample

2. Preparation of CDSs

3. Extraction of 9 variables of interest

4. Prediction model

5. Results

Scitechdaily.com

GMO or non-GMO ?

CDSs

ATTCCGA
CGTGGAT

Non-GMO

potential GMOs

DB known GMOs

Dist., %GC, length

CDSs

3,12 36,1 12
5,38 25,8 15

Non-GMO

potential GMOs

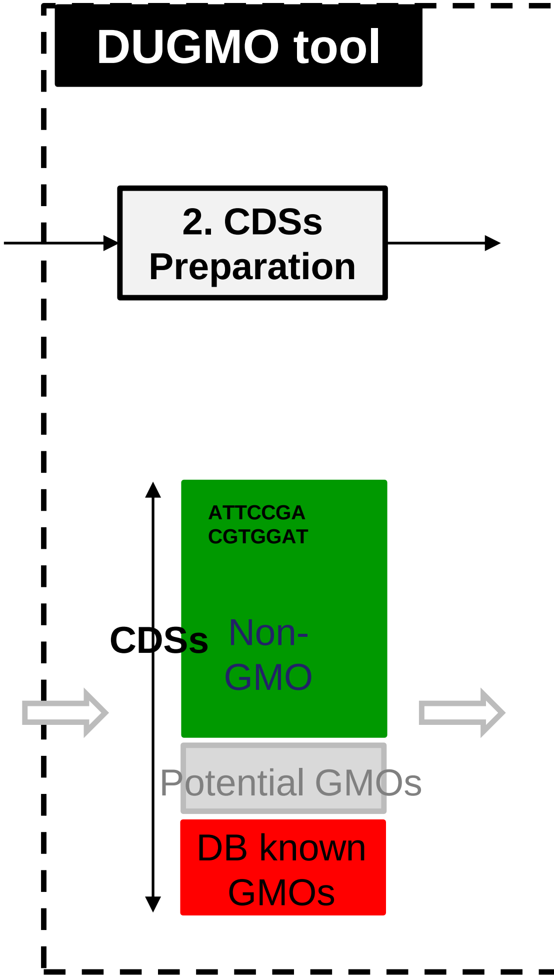
DB known GMOs

Tangente-mag.com

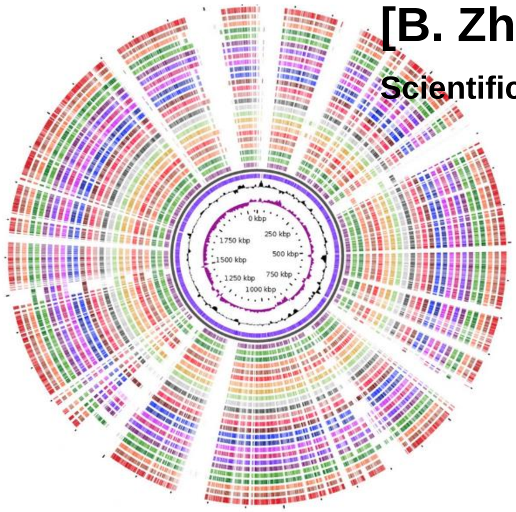
Prediction model

non-GMO
ou
GMO

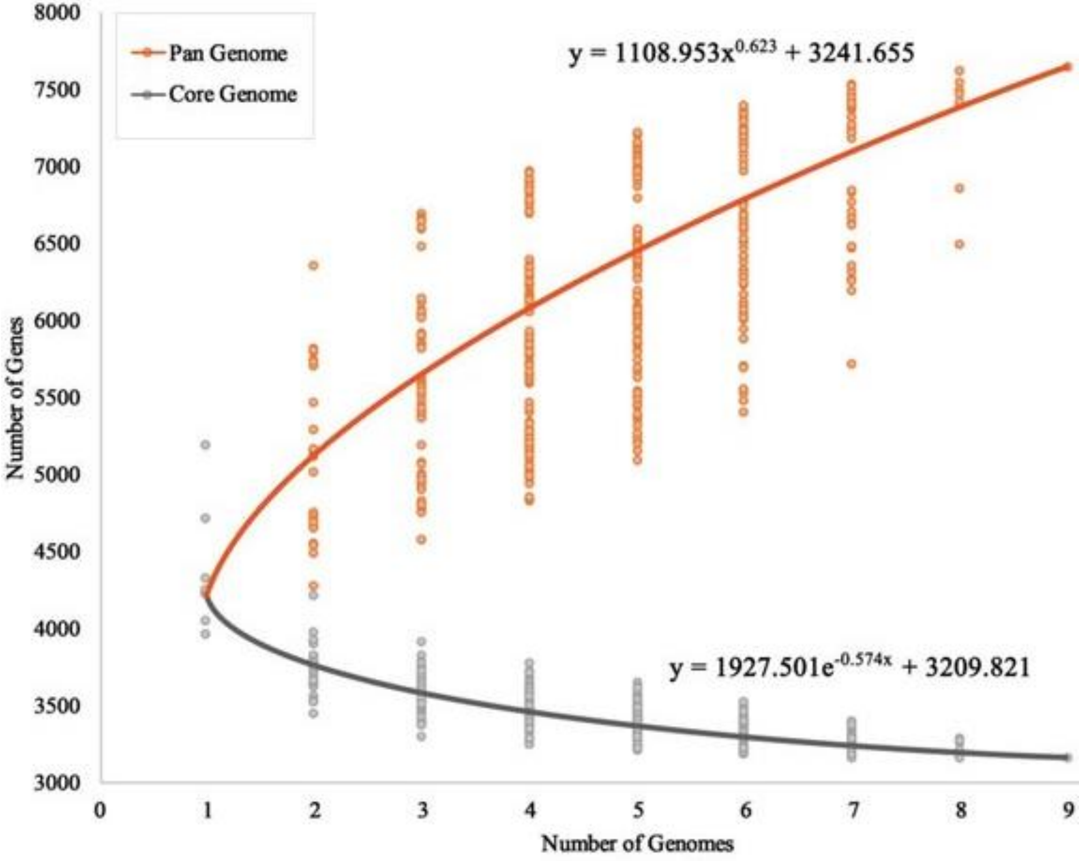
1. Introduction
2. Objectives of the PhD
3. **Data preparation**
4. Computation of distances
5. Design of a prediction model
6. Results
7. Conclusion
8. Perspectives

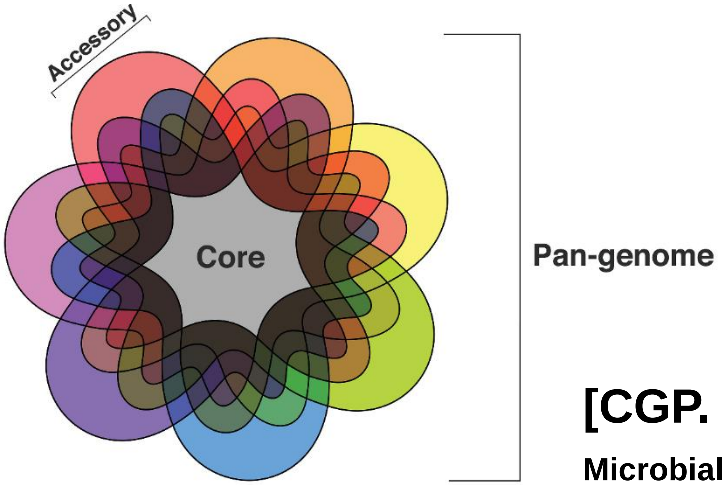


A – Pangenome: wild-type and representative of species



[B. Zheng et al.
ScientificReports 2017]





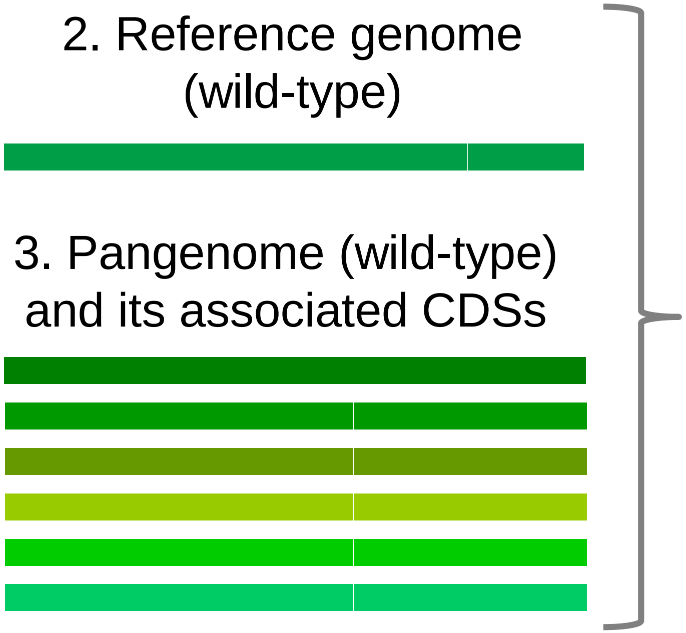
[CGP. McCarthy et al.
Microbial genomics 2019]

[D. Sinha et al.
BMC genomics 2021]

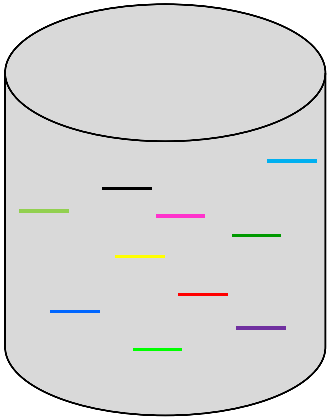
A – Data required as input to the software

illumina.com

1. High throughput sequencing data


2. Reference genome (wild-type)
3. Pangenome (wild-type) and its associated CDSs

Species of the potentially GM bacterium



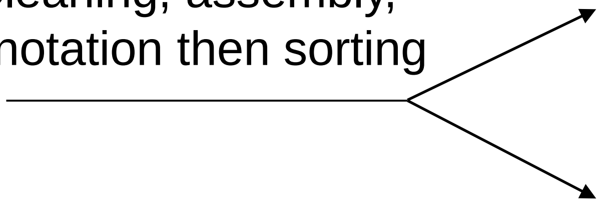
4. Databank of known GMO inserts

B – Objectives

- Data cleaning, assembly, annotation of sequencing data
- Sorting of the sample CDSs
- Filtering of the database of known GMO inserts

illumina.com

High throughput sequencing data



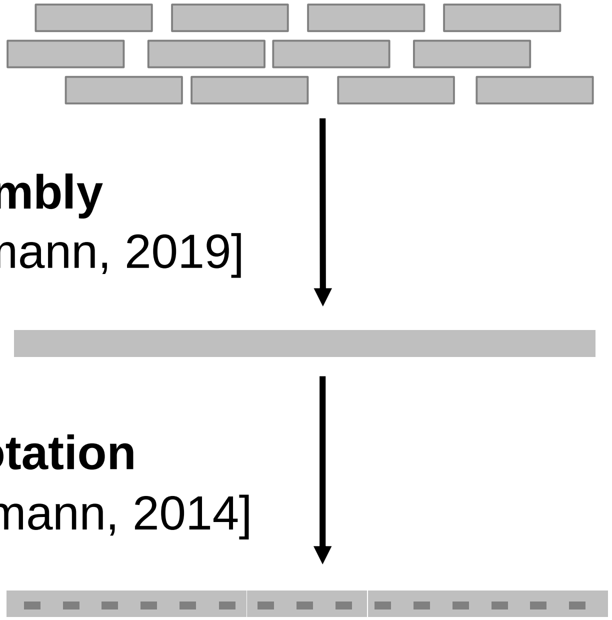
Cleaning, assembly, annotation then sorting

CDSs related to the host genome

CDSs which are potential GMO inserts

B – Cleaning pipeline

- Assembly and annotation of high throughput sequencing data



potentially GM sample

Reads

Assembly
Shovill [Seemann, 2019]

Whole genome
contigs/scaffolds

Annotation
Prokka [Seemann, 2014]

CDSs

C – Sorting of the sample CDSs

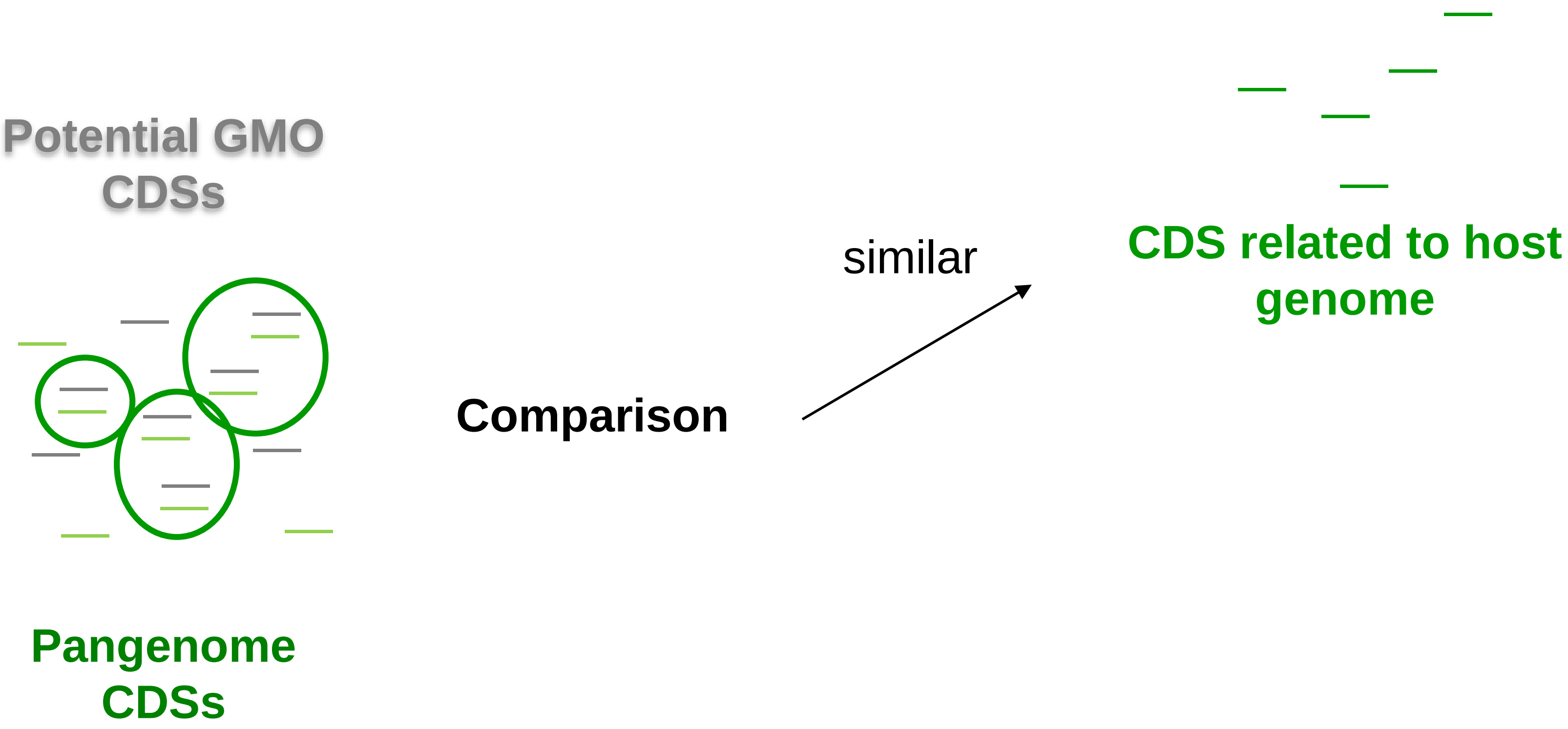
- Step 1 : Comparison of the potental GMO CDSs with pangenome's CDSs

Potential GMO CDSs

Pangenome CDSs

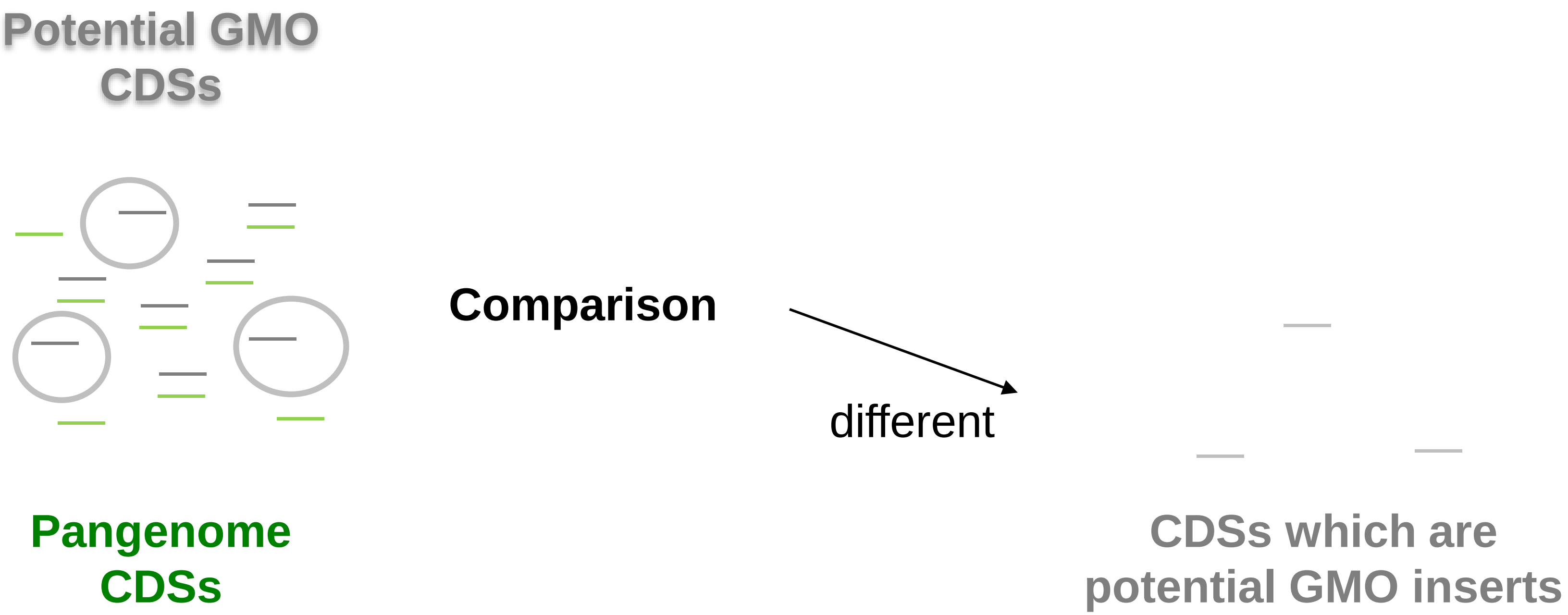
C – Sorting of the sample CDSs

- Step 1 : Comparison of the potental GMO CDSs with pangenome’s CDSs



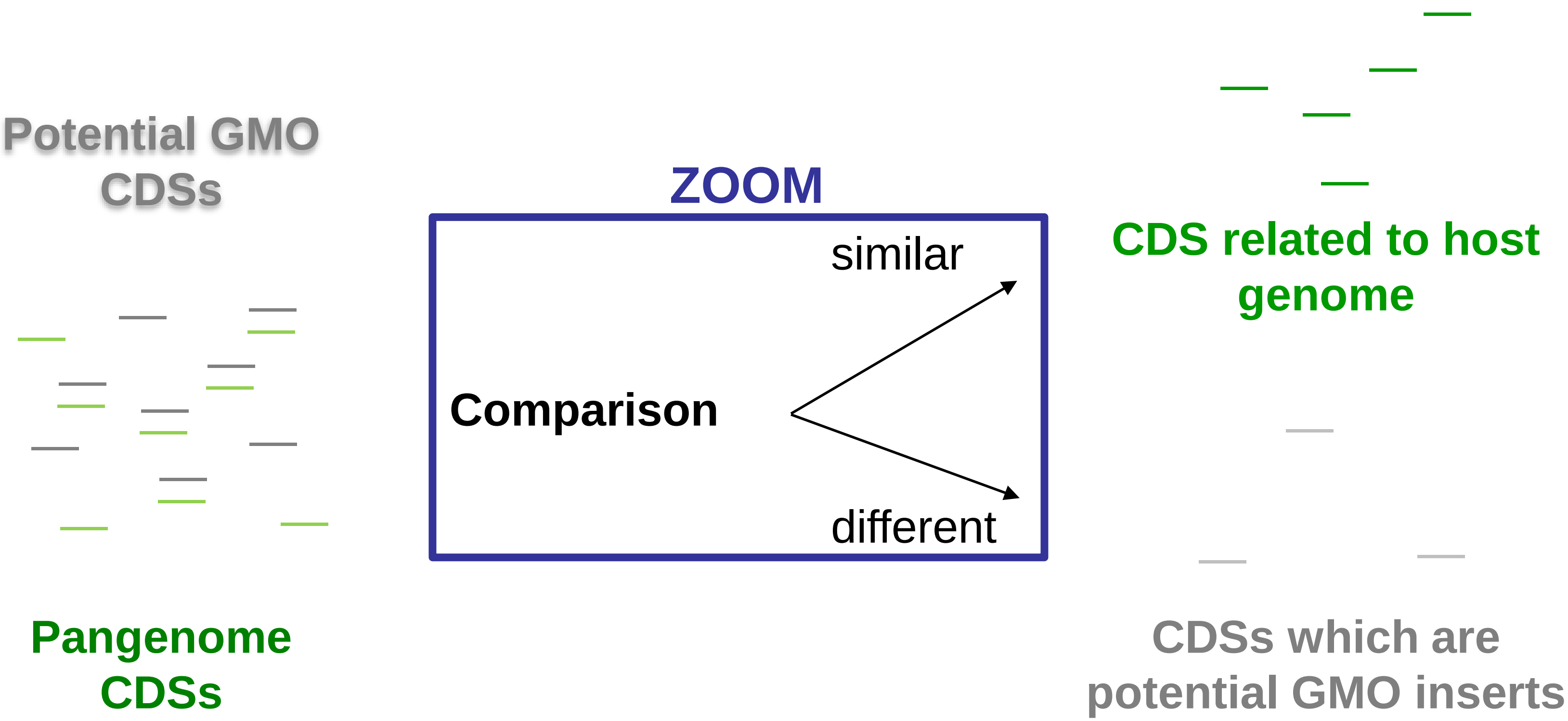
C – Sorting of the sample CDSs

- Step 1 : Comparison of the potental GMO CDSs with pangenome's CDSs



C – Sorting of the sample CDSs

- Step 1 : Comparison of the potental GMO CDSs with pangenome’s CDSs



Potential GMO CDSs

Pangenome CDSs

ZOOM

Comparison

similar

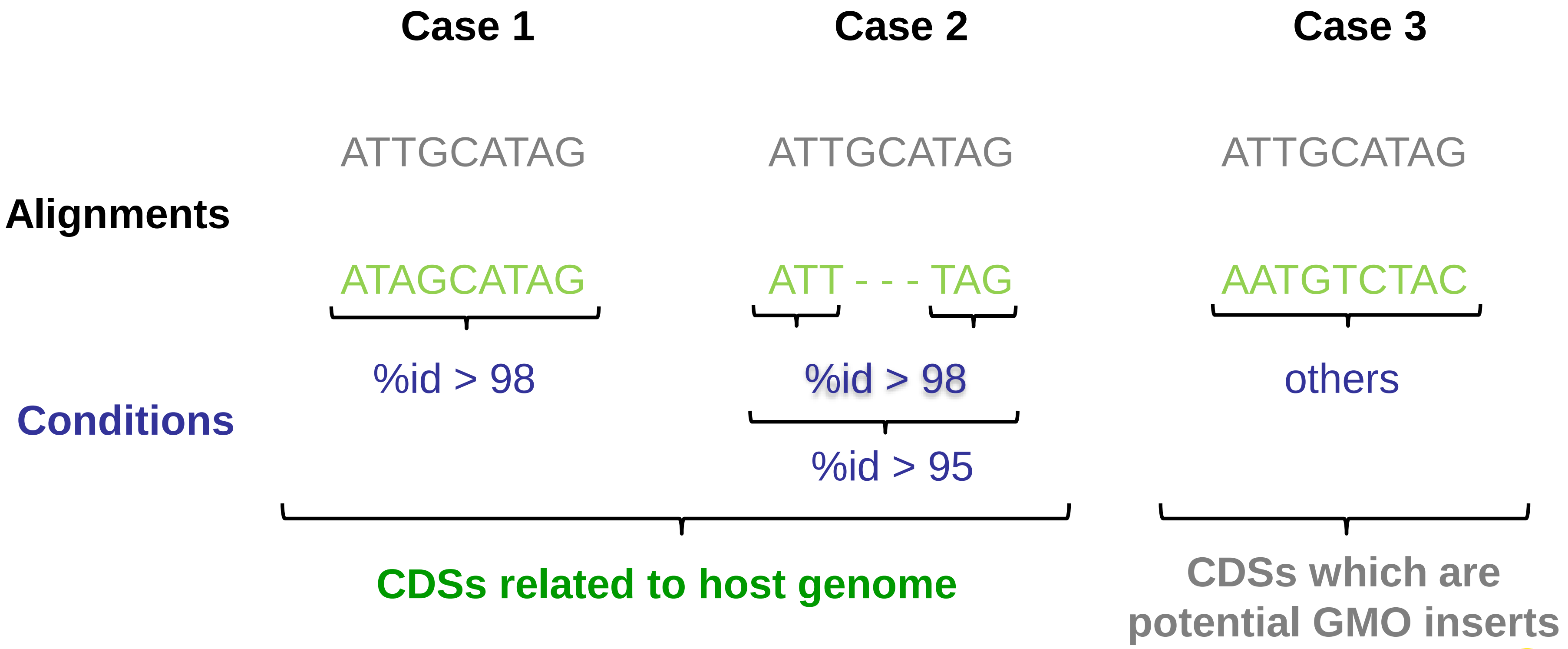
different

CDS related to host genome

CDSs which are potential GMO inserts

C – Sorting of the sample CDSs

- Step 1 : Comparison of the potental GMO CDSs with pangenome’s CDSs



Case 1 | **Case 2** | **Case 3**

Alignments

Case 1: ATTGCATAG / ATAGCATAG

Case 2: ATTGCATAG / ATT - - - TAG

Case 3: ATTGCATAG / AATGTCTAC

Conditions

Case 1: %id > 98

Case 2: %id > 98, %id > 95

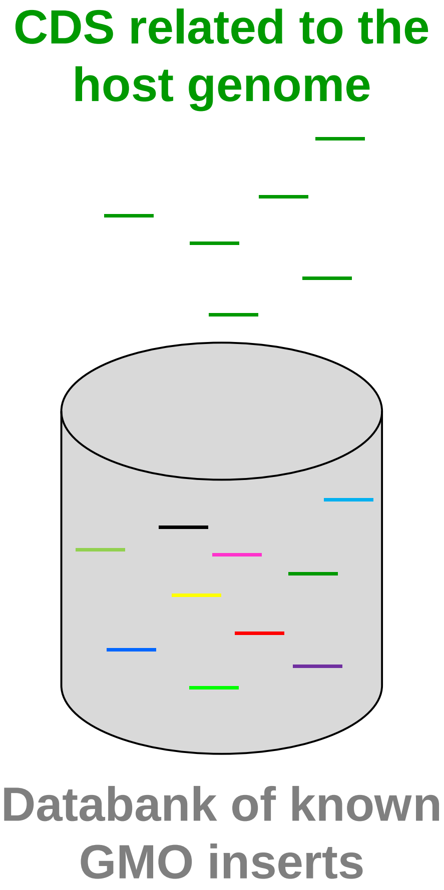
Case 3: others

CDSs related to host genome

CDSs which are potential GMO inserts

D – Filtering the GMO insert databank

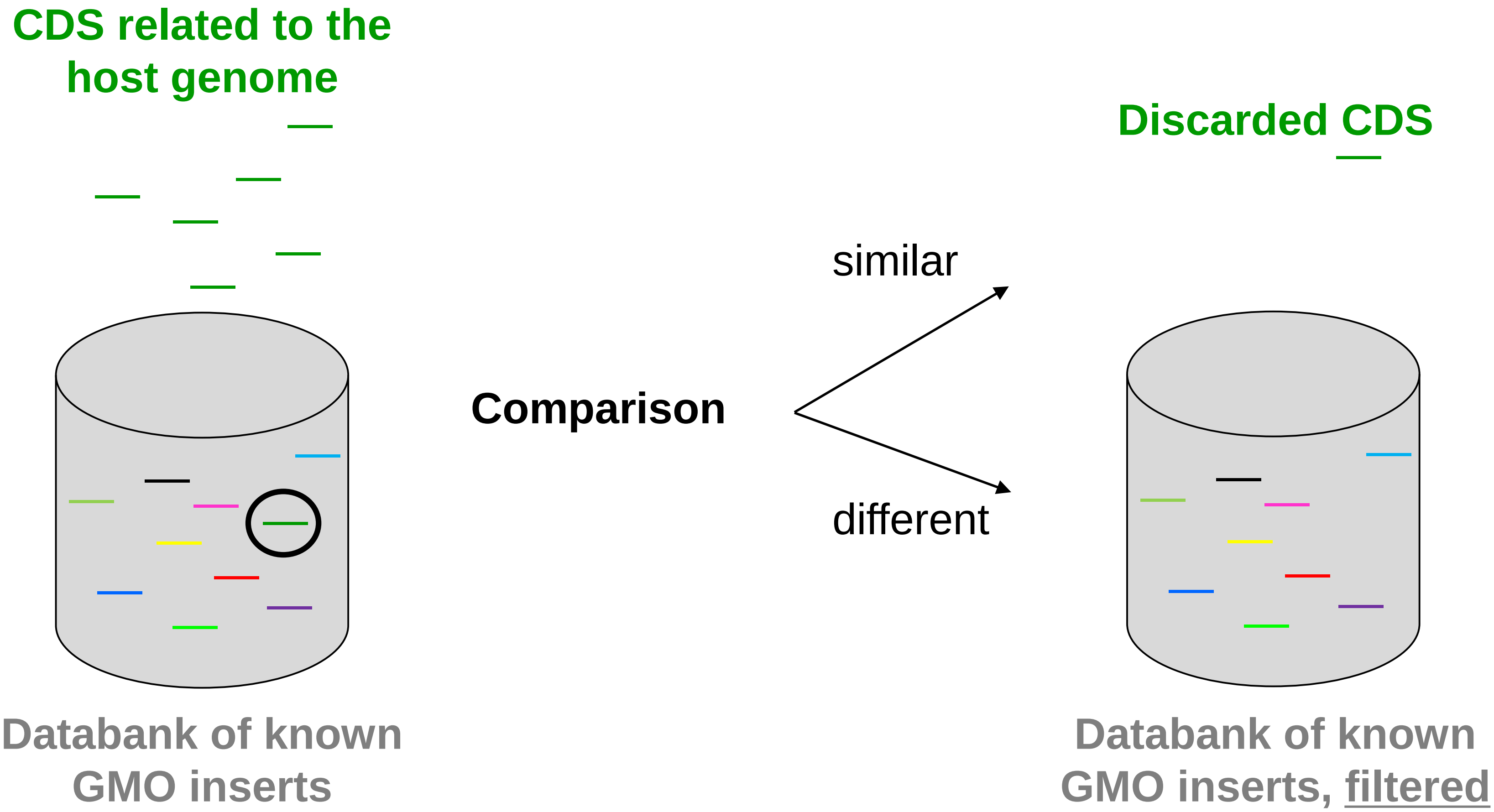
- Using the host genome CDSs



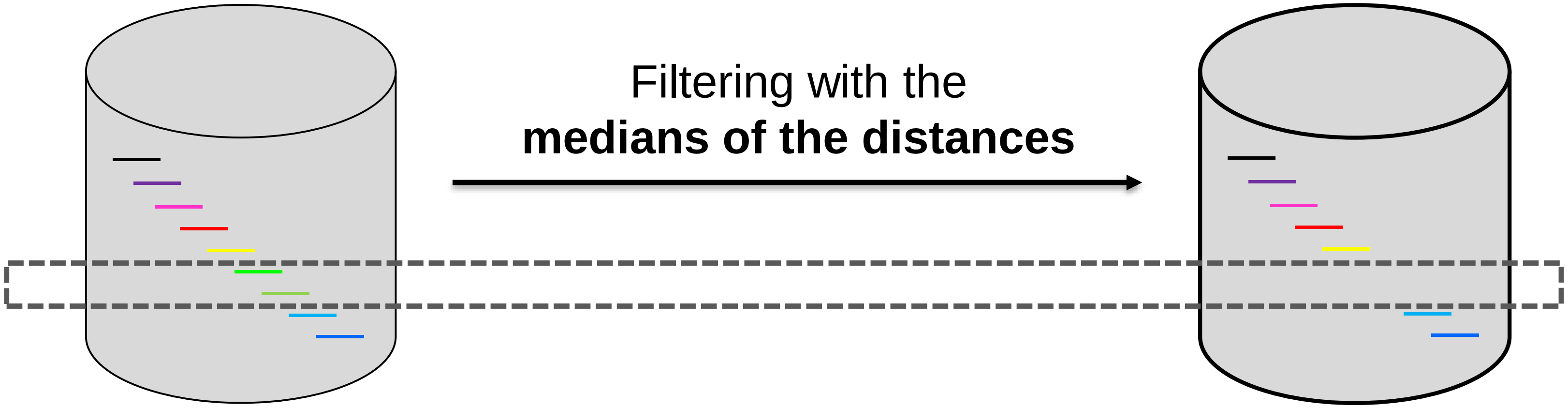
Comparison criteria identical to step 1

D – Filtering the GMO insert databank

- Using the host genome CDSs


CDS related to the host genome
Discarded CDS
similar
Comparison
different
Databank of known GMO inserts
Databank of known GMO inserts, filtered

D – Filtering the GMO insert databank

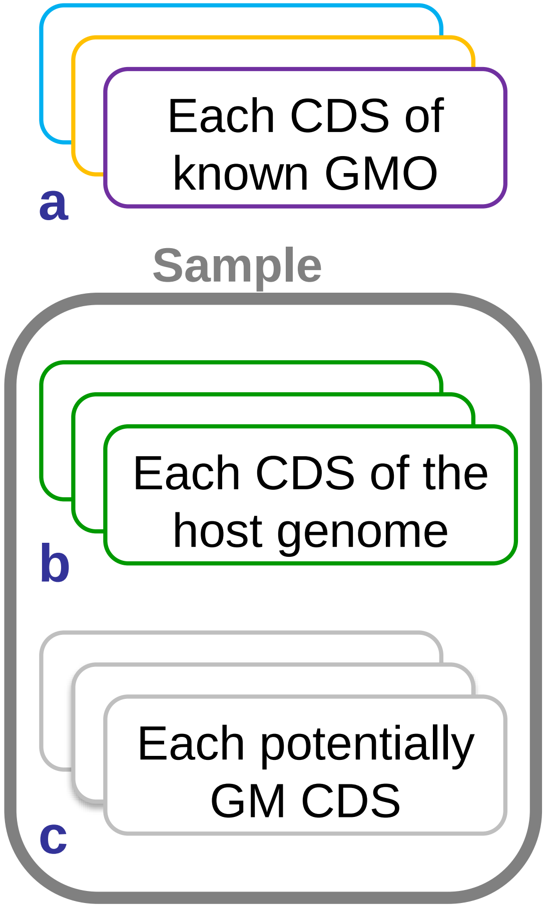


Filtering with the **medians of the distances**

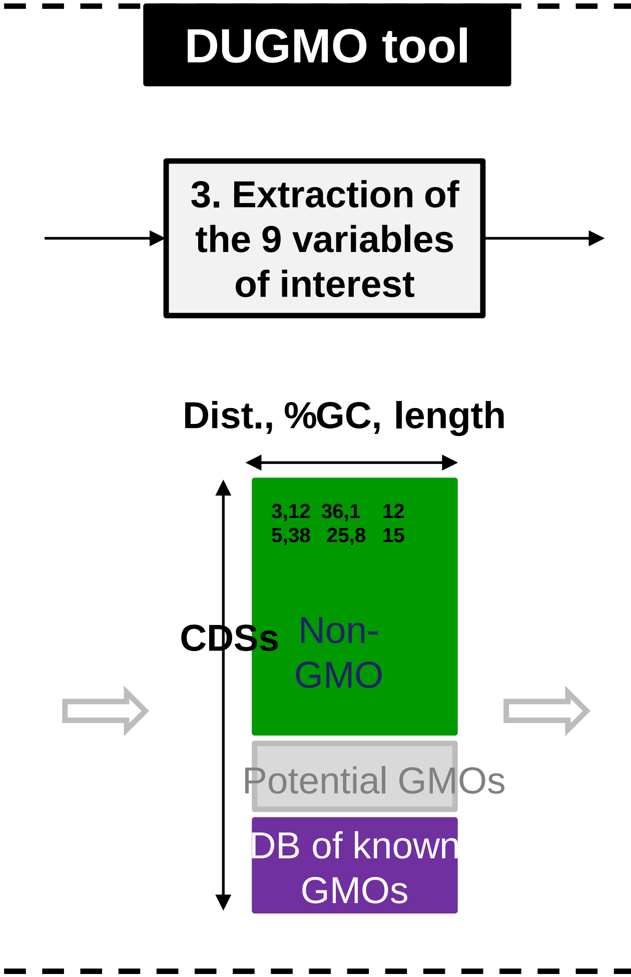
Databank of known GMO inserts

Databank of known GMO inserts, filtered

C – Get the three needed distinct datasets



1. Introduction
2. Objectives of the PhD
3. Data preparation
4. **Computation of the distances**
5. Design of a prediction model
6. Results
7. Conclusion
8. Perspectives



A – Objectives

- Objective : characterization of the host genome CDSs and of the two other CDS sets

- Approach : determine the best combinations of parameters

- Purpose : establish explanatory variables for the development of the prediction model

B – Characterization of the genomic vocabulary

- R’MES [Schbath, S. and Hoebeke, M. 2011]: Searches for the exceptional words in a sequence
  - Determine the number of occurrences of each word and its exceptional character
  - It defines a set of words that are over-represented in the host genome

- Three distance formula and two types of calculations
  - Euclidean distance
  - Kullback-Leibler distance [Trifonov et Rabadan, 2010]
  - Bray-Curtis distance

B – Characterization of the genomic vocabulary

- R’MES : Searches for the exceptional words in a sequence
    - Determine the number of occurrences of each word and its exceptional character
    - It defines a set of words that are over-represented in the host genome
- Three distance formula and two types of calculations
    - Euclidean distance
    - Kullback-Leibler distance [Trifonov et Rabadan, 2010]
    - Bray-Curtis ~distance (dissimilarity measurement)

B – Characterization of the genomic vocabulary

- Type of distance calculation named « in Frequencies »
  - Concatenation of the third codon positions into a new sequence

	Word size	Running R'mes
AGTACGTCAGGTAGTATCCAGCTAATG	27	impossible
TGATTCGAG	9	fast

  - 10% of over-represented words

- Type of distance calculation named « in Proportions »
  - Whole CDS
  - All words

B – Characterization of the genomic vocabulary

- Type of distance calculation named « in Frequencies »
  - Concatenation of the third codon positions into a new sequence

AGTACGTCAGGTAGTATCCAGCTAATG

TGATTCGAG

Arginine

CGU
CGC
CGA
CGG

  - 10% of over-represented words

- Type of distance calculation named « in Proportions »
  - Whole CDS
  - All words

B – Characterization of the genomic vocabulary

- Type of distance calculation named « in Frequencies »
  - Concatenation of the third codon positions into a new sequence

Arginine

CGU
CGC
CGA
CGG

AGTACGTCAGGTAGTATCCAGCTAATG

TGATTCGAG

  - 10% of over-represented words

- Type of distance calculation named « in Proportions »
  - Whole CDS
  - All words

C – Combinations of parameters

- List of tested parameters :
  - Word sizes
  - Order of the Markov model
  - Percentage of over-/under-represented words
  - Concatenation of the third codon positions

D – Explanatory variables retained in DUGMO for one CDS

- Bray-Curtis distances
  - « In frequencies »
    - F L9M7 : Word size 9 and Markov model order 7
  - « In proportions »
    - P L3M1 : Word size 3 and Markov model order 1
    - P L4M2 : Word size 4 and Markov model order 2
- Average exceptionality scores in the host genome for L4M2 et L9M7
- Count density per nucleotide for 4-letter and 9-letter words

$$\frac{\text{Sum of all word counts}}{\text{CDS length}}$$

- Percentage of GC
- Length

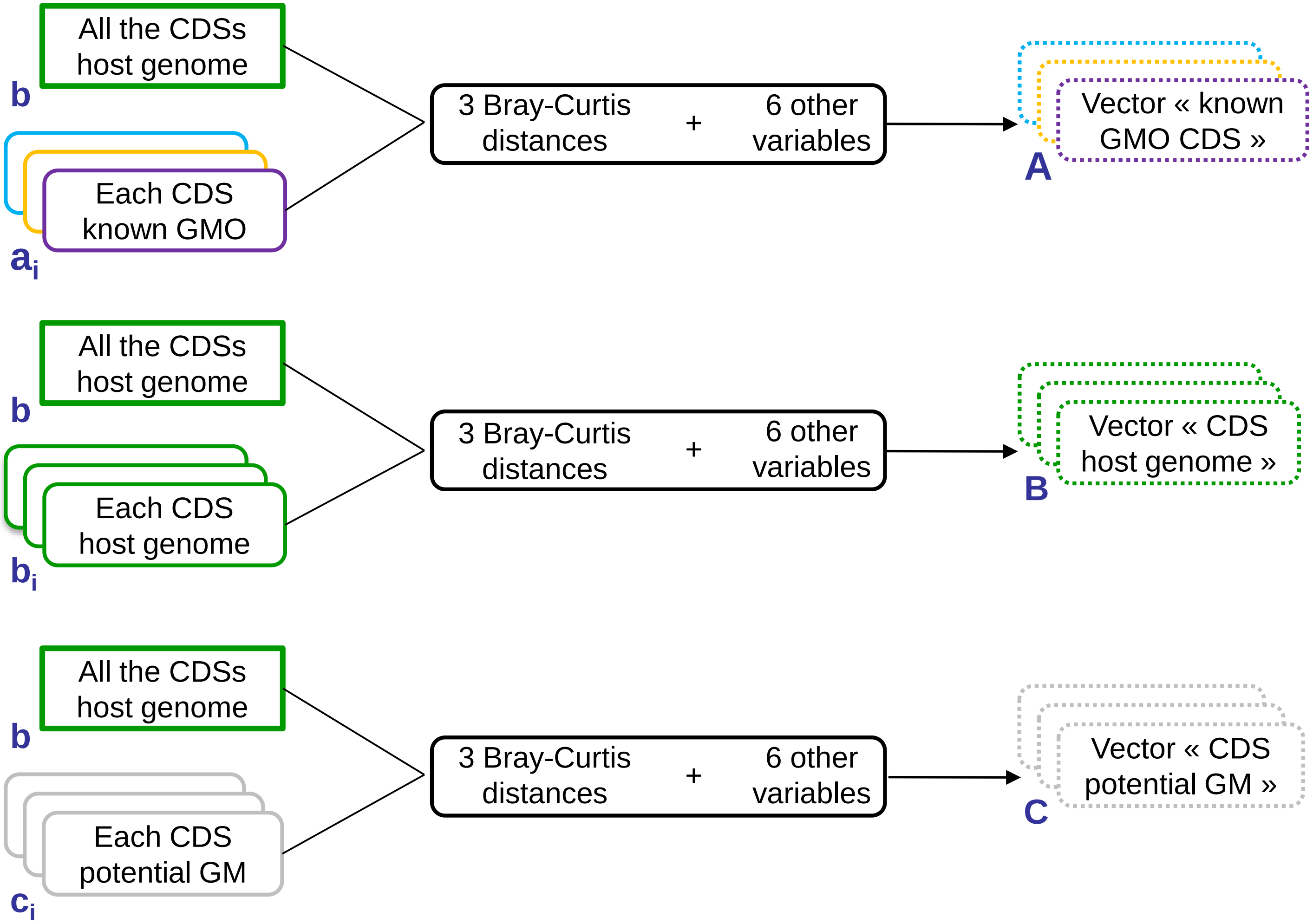
D – Explanatory variables retained in DUGMO for one CDS

- Bray-Curtis distances
  - « In frequencies »
    - F L9M7 : Word size 9 and Markov model order 7
  - « In proportions »
    - P L3M1 : Word size 3 and Markov model order 1
    - P L4M2 : Word size 4 and Markov model order 2
- Average exceptionality scores provided by R’MES in the host genome for L4M2 and L9M7
- Count density per nucleotide for 4-letter and 9-letter words

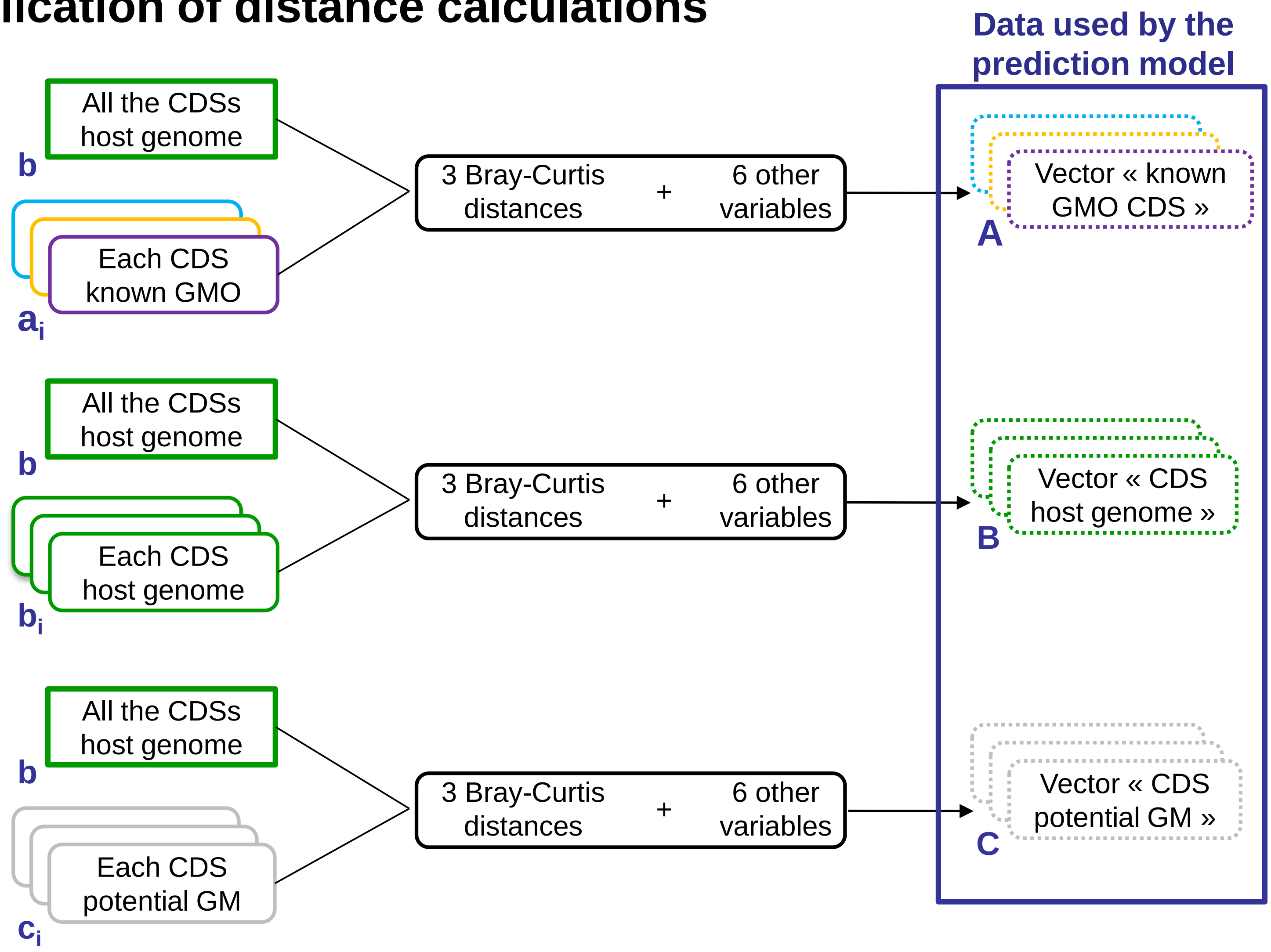
$$\frac{\text{Sum of all word counts}}{\text{CDS length}}$$

- Percentage of GC
- Length

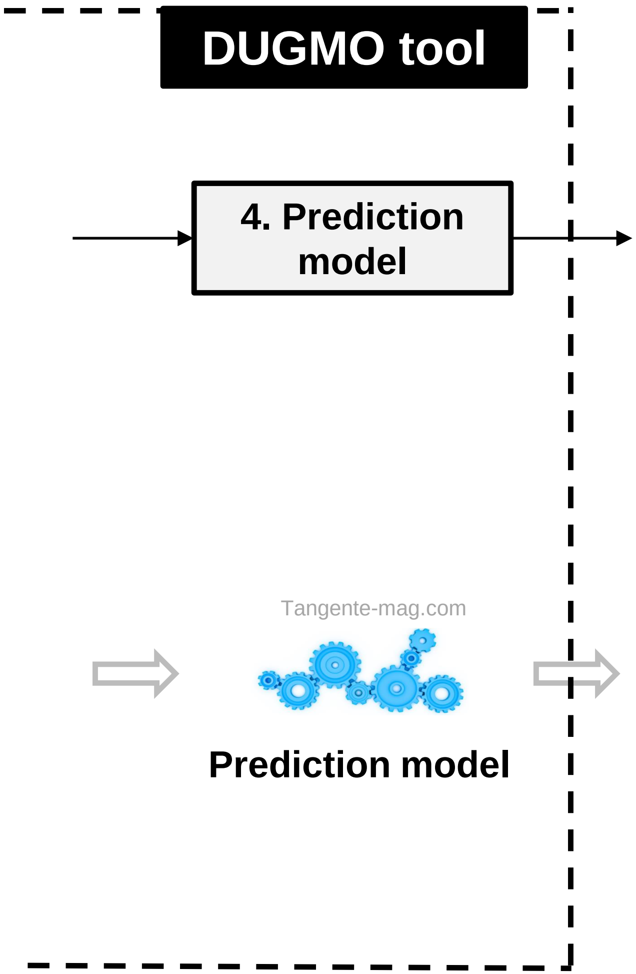
E – Application of distance calculations


All the CDSs host genome
b
Each CDS known GMO
a$_i$
3 Bray-Curtis distances + 6 other variables
Vector « known GMO CDS »
A
All the CDSs host genome
b
Each CDS host genome
b$_i$
3 Bray-Curtis distances + 6 other variables
Vector « CDS host genome »
B
All the CDSs host genome
b
Each CDS potential GM
c$_i$
3 Bray-Curtis distances + 6 other variables
Vector « CDS potential GM »
C

E – Application of distance calculations



1. Introduction
2. Objectives of the PhD
3. Préparation des données
4. Calcul de distances
5. **Design of the prediction model**
6. Results
7. Conclusion
8. Perspectives



DUGMO tool

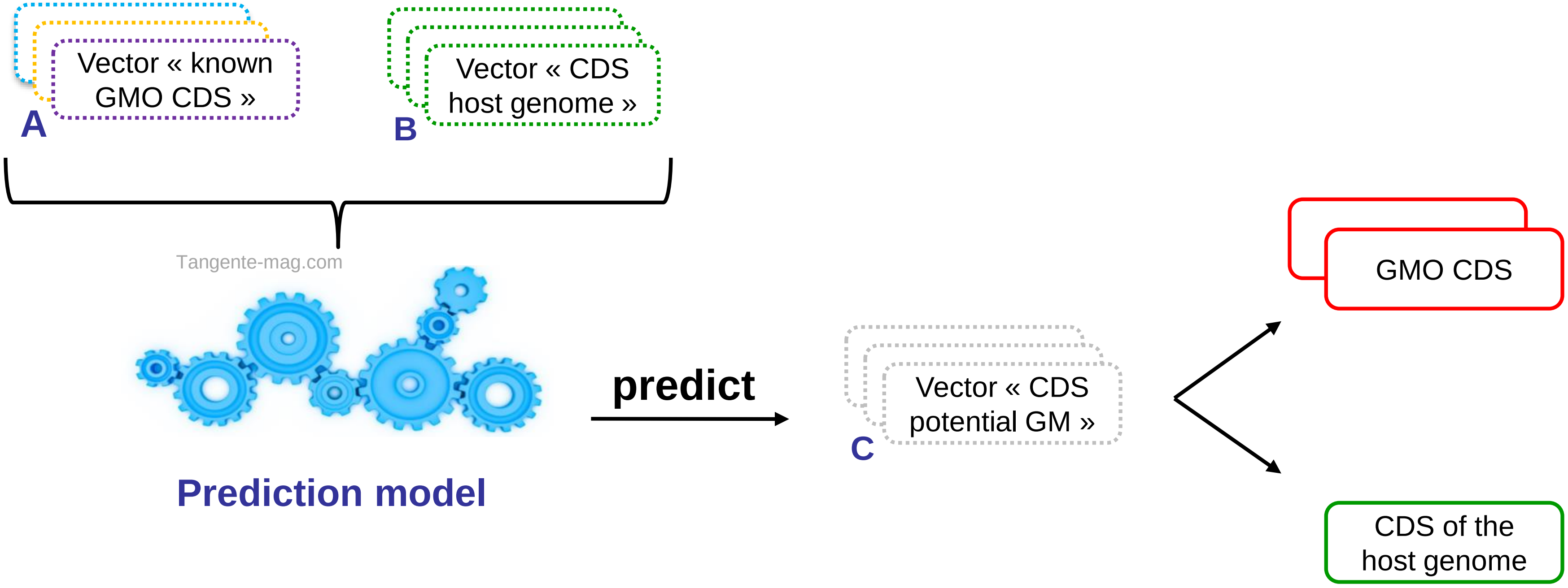
4. Prediction model

Tangente-mag.com

Prediction model

A – Objectives

- Approach : use of Machine Learning methods
- Purpose : predict proven CDSs of GMO inserts



A
Vector « known GMO CDS »

B
Vector « CDS host genome »

Tangente-mag.com

Prediction model

predict

C
Vector « CDS potential GM »

GMO CDS

CDS of the host genome

B – Machine Learning

- 12 tested methods

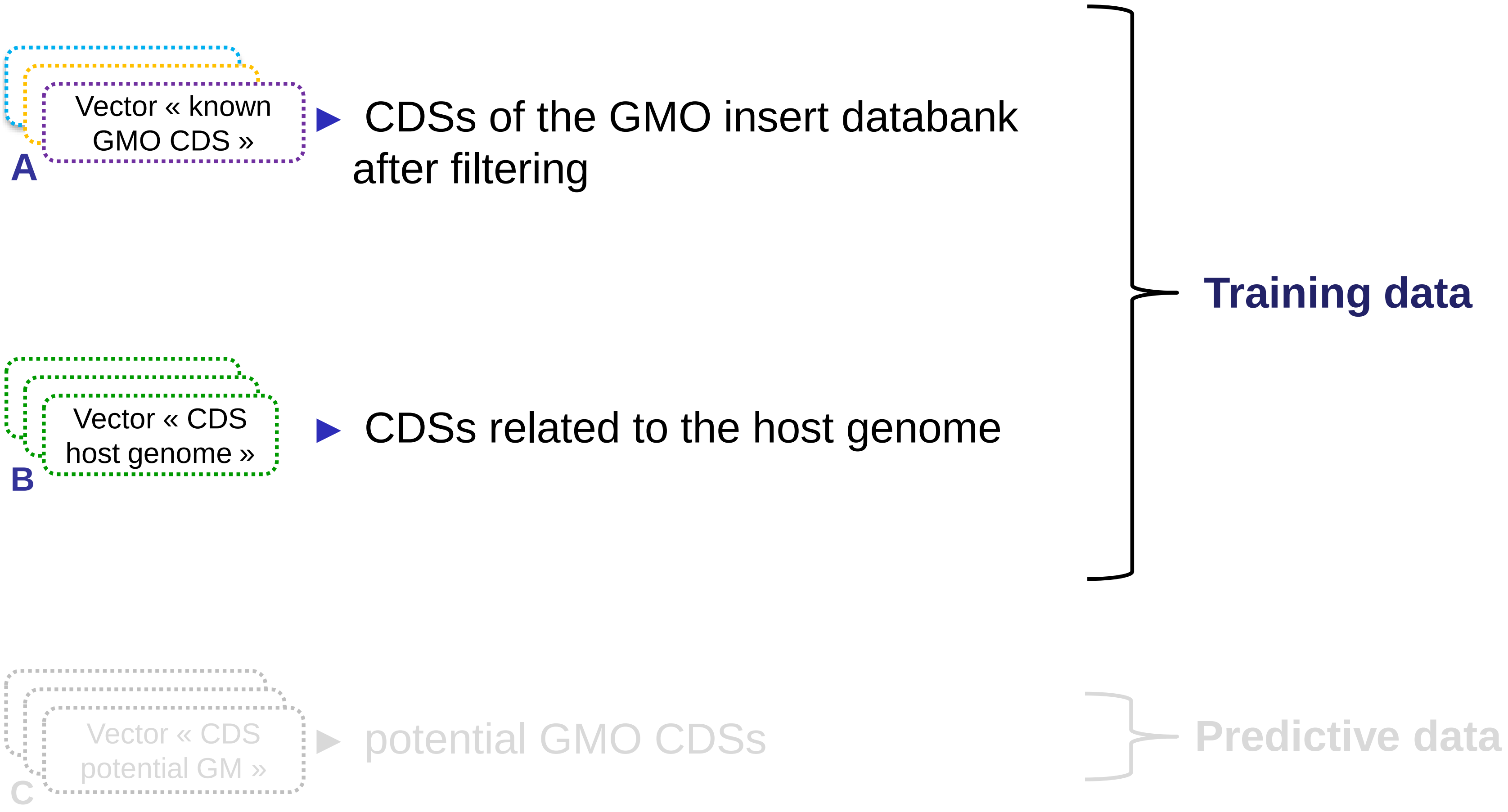
parametric

- Logit : generalized linear model
- StepLDA : linear discriminant analysis
- StepQDA : quadratic discriminante analysis
- Plsda : partial least squares regression

non parametric including decision trees

- NN : neural networks
- SvmRadial : support vector machines
- KNN : K nearest neighbors
- C5.0 : classification algorithm
- Rpart : recursive partitionig trees
- RF : random forests
- Treebag : classification trees with bagging
- Xgboost : extreme gradient boosting

C – Used data



A

Vector « known GMO CDS »

▶ CDSs of the GMO insert databank after filtering

B

Vector « CDS host genome »

▶ CDSs related to the host genome

Training data

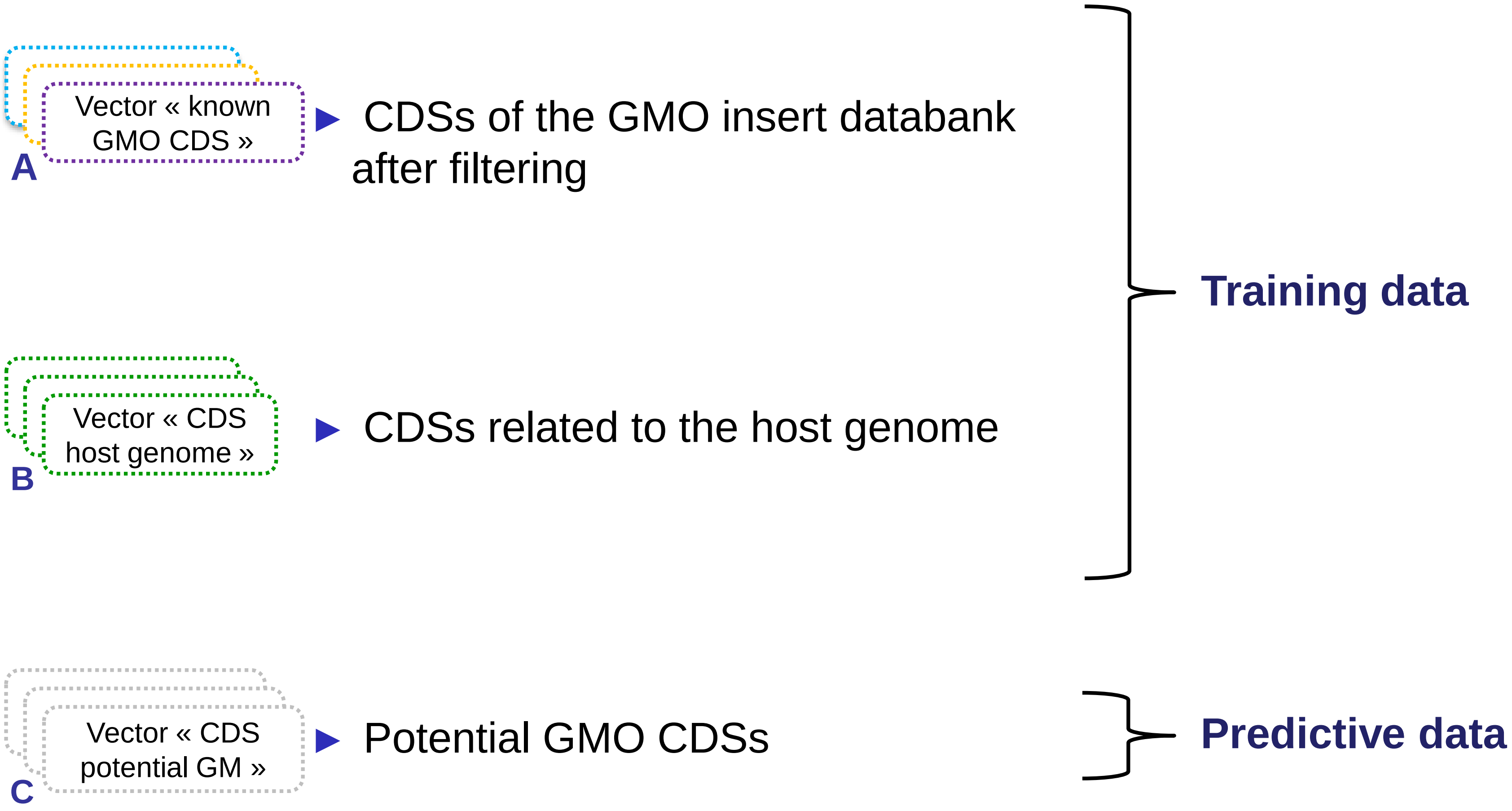
C

Vector « CDS potential GM »

▶ potential GMO CDSs

Predictive data

C – Used data



A – Vector « known GMO CDS » ▶ CDSs of the GMO insert databank after filtering

B – Vector « CDS host genome » ▶ CDSs related to the host genome

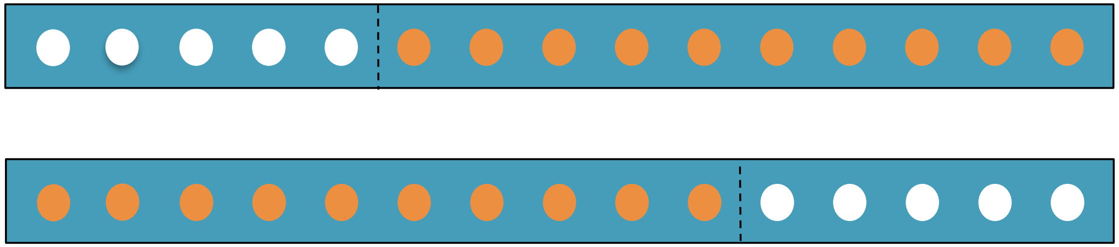
Training data

C – Vector « CDS potential GM » ▶ Potential GMO CDSs

Predictive data

D – Comparison of the tested methods

- Centered, reduced, and stratified data
- Optimisation of the parameters of each method in 10-fold cross-validation
- 2-fold cross-validation comparison of methods



2-fold cross-validation

E – Selection criteria

- Confusion matrix

		Predictions of a model	
		GMO	**Non-GMO**
Real data	**GMO**	True positives	False negatives
	Non-GMO	False positives	True negatives

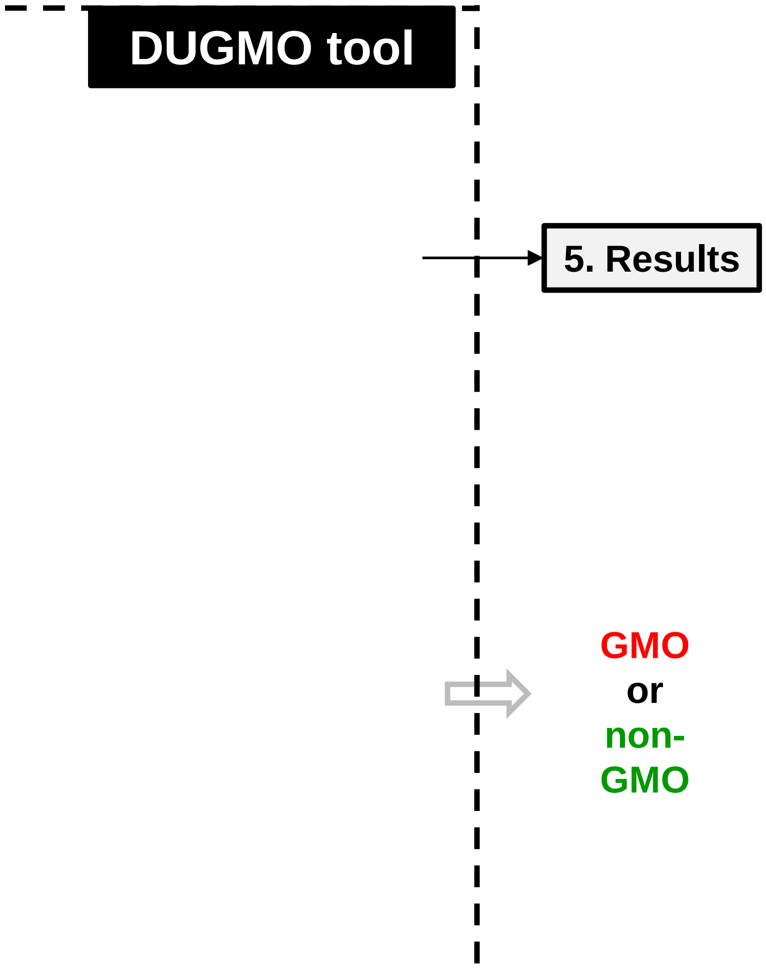
- **Specificity : ++**
- **False positive rate: --**
- **Sensitivity : ++**
- **False negative rate : +++**

F – Final choice

- Union of the results of two methods
  - **RF** : Random Forests
  - **Logit** : Generalized linear model

	Results for prediction data		
	Logit	**RF**	**Union of RF and Logit**
False negative rate	0.04	0.01	0.01
Specificity	0.94	0.98	0.99
Sensitivity	0.95	0.98	0.99
False positive rate	0.05	0.01	0.09

1. Introduction
2. Objectives of the thesis
3. Data preparation
4. Computation of the distances
5. Design of a prediction model
6. **Results**
7. Conclusion
8. Perspectives



DUGMO tool

5. Results

GMO
or
non-GMO

A – Tested data

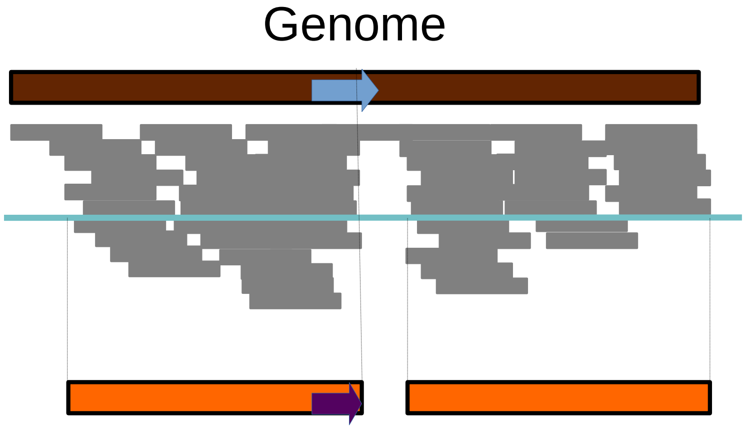
- **1 wild-type** genome and **1 GM** genome of *B. subtilis* bacterium
- **3 GM** genomes of *E. coli* bacterium
- **6 wild-type** genomes of bacteria
  - *Campylobacter jejuni*
  - *Lactococcus lactis*
  - *Listeria monocytogenes*
  - *Mycobacterium tuberculosis*
  - *Staphylococcus aureus*
  - *Salmonella* Typhimurium
- **42 synthetic GM** genomes
  - Combinations : 6 wild-type bacteria + 7 exogenous genes

B – Global results

- With the 2 *B. subtilis* genomes
  - WT: No insert ✓
  - GM: 25 detected inserts + ✓
    12 false negatives (maximum)
- With the 3 GM genomes of *E. coli*
  - **3** known inserts are detected ✓
  - **1 false positive** ✗ → ✓
- With the 48 synthetic WT genomes
  - **47**: No insert ✓
  - **1 false positive** (*M. tuberculosis)* ✗ → ✓
- With the 48 synthetic GM genomes
  - **47**: insert found ✓
  - **1 false negative** (*S. aureus* including a *C. jejuni* gene) ✗

C – GM *Escherichia coli*

- In the genome modified with a gene from *S. pyogenes* : 1 false positive
  - *arlS* gene
  - **Truncated gene because of the assembly**



Legend

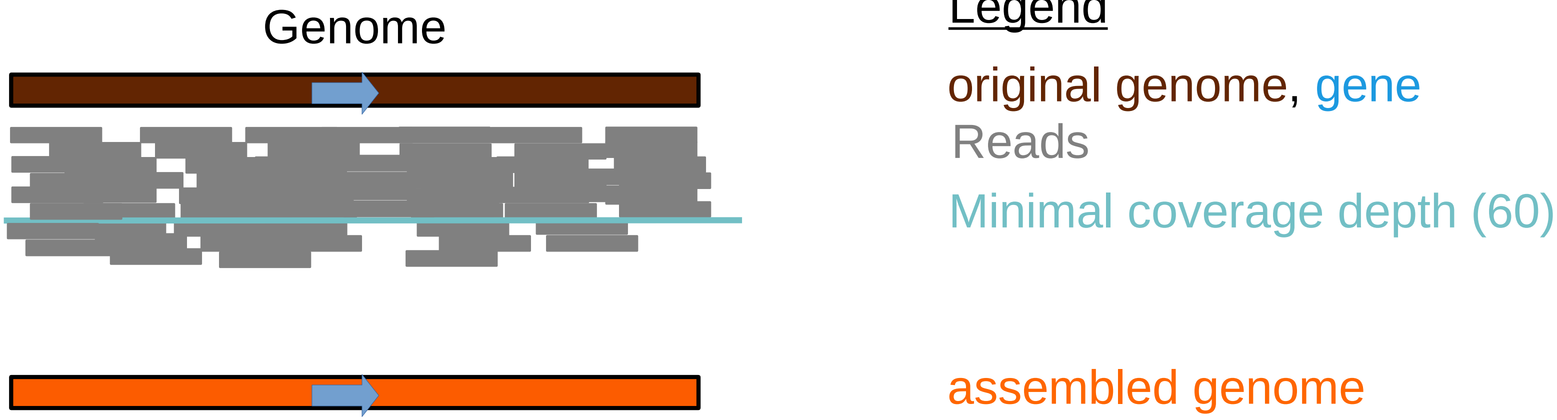
Original genome, gene
Reads
Minimal coverage depth (60)

assembled genome, truncated gene

C – GM *Escherichia coli*

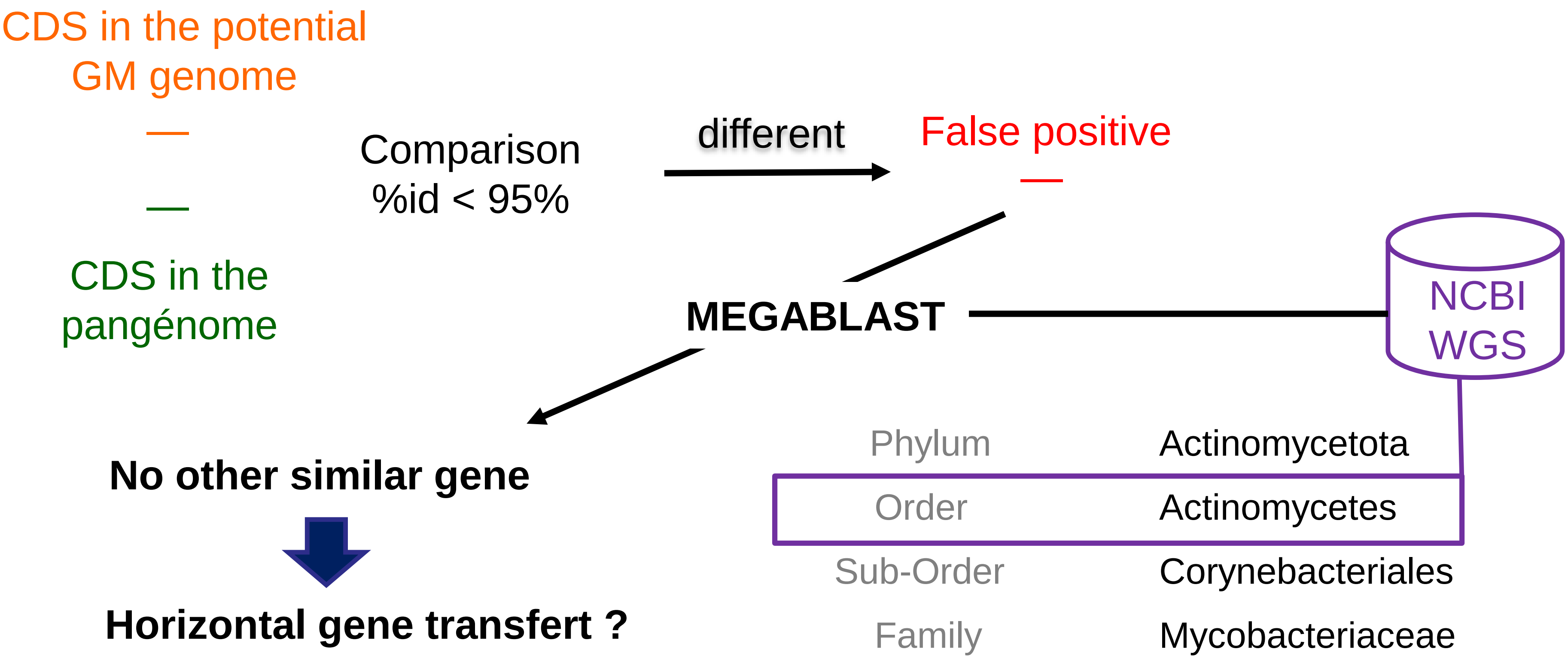
- In the genome modified with a gene from *S. pyogenes* : ~~1 false positive~~

- Solution : **minimal coverage depth of 60** (bacteria)



D – Wild-type *Mycobacterium tuberculosis*

- One false positive detected by DUGMO: **labelled GM** while it **is not**



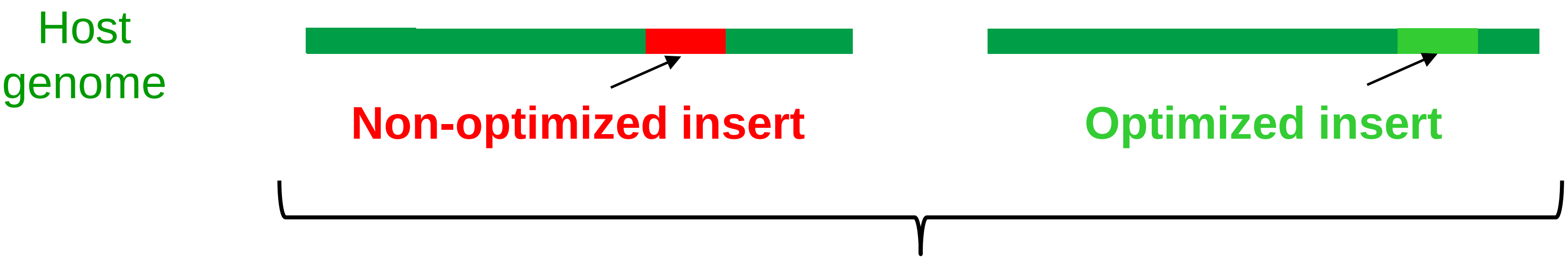
No other similar gene

Horizontal gene transfert ?

Phylum	Actinomycetota
Order	Actinomycetes
Sub-Order	Corynebacteriales
Family	Mycobacteriaceae
Genus	Mycobacterium

E – Synthetic data

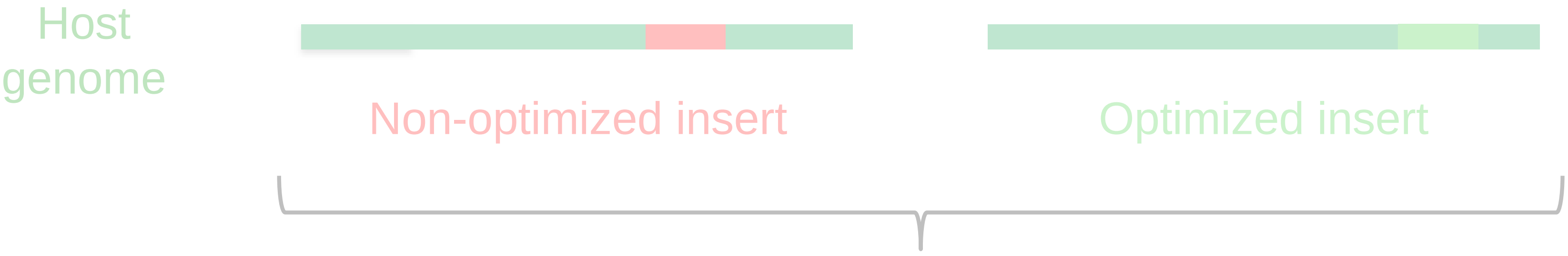
- Objectives
  - Test the sensitivity of the method to **dicodon optimization**



Detection by DUGMO ? **YES**

E – Synthetic data

- Objectives
  - Test the sensitivity of the method to dicodon optimization



Host genome

Non-optimized insert

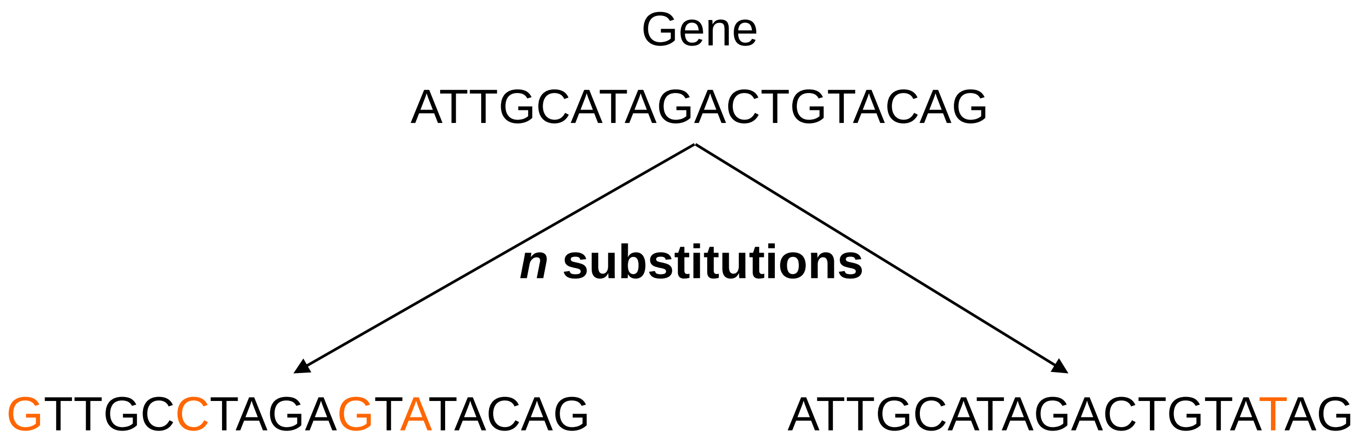
Optimized insert

Detection by DUGMO ? YES

  - Test the **detection threshold** of the method

E – Synthetic data

- Generation of **mutations in a wild-type gene** of *B. subtilis*



Gene

ATTGCATAGACTGTACAG

n **substitutions**

GTTGCCTAGAGTATACAG

ATTGCATAGACTGTATAG

DUGMO results	Detection	No detection

- Results : **substitution rate ≥ 9%** ➡ **detection** of the GM gene

1. Introduction
2. Objectives of the thesis
3. Data preparation
4. Computation of the distances
5. Design of a prediction model
6. Results
7. **Conclusion**
8. Perspectives

C – Summary: before

	Detection known or partially known GMO		Detection unkown GMO	
	Prokaryotes	Eukaryotes	Prokaryotes	Eukaryotes
Intergenic sequence	✔	✔	✘	✘
Truncated gene	✔	✔	~	~
Fused gene	✔	✔	~	~
Insertion/deletion in a gene	✔	✔	✘	✘
% of point mutations ≥ 9%	✔	✔	✘	✘
% of point mutations < 9%	✔	✔	✘	✘

C – Summary: after

	Detection known or partially known GMO		Detection unknown GMO	
	Prokaryotes	Eukaryotes	Prokaryotes	Eukaryotes
Intergenic sequences	✔	✔	✘	✘
Truncated gene	✔	✔	✔	⌛
Fused gene	✔	✔	✔	⌛
Insertion/deletion in a gene	✔	✔	✔	⌛
% of point mutations ≥ 9%	✔	✔	✔	⌛
% of point mutations < 9%	✔	✔	✘	✘

1. Introduction
2. Objectives of the thesis
3. Data preparation
4. Computation of the distances
5. Design of a prediction model
6. Results
7. Conclusion
8. **Perspectives**

A – General

- ▶ Adapt the method for application to other organisms

tokopedia.com Tetra Import Glowfish

- ▶ Possibility to provide assembly data as input (Illumina, Pacbio)

GloFish®

Ackowledgements

- **Machine Learning (Anses)**
  Stéphanie Bougeard

- **Biology : LSV Angers (Anses)**
  Vincent Hérau,
  Mathieu Rolland
  Anne-Laure Boutigny

- **Word statistics: INRAE of Jouy en Josas**
  Sophie Schbath

- **JRC at Ispra**
  Mauro Petrillo

European Commission
Joint Research Centre

- **CRA-W (Belgique)**
  Fredérique Debode

Wallonie recherche CRA-W

- **BVL (Allemagne)**
  Lutz Grohmann

Federal Office of Consumer Protection and Food Safety

Région BRETAGNE

- **Funders**
  Anses

Région Bretagne

Thank you for your attention

https://github.com/ANSES-Ploufragan/DUGMO

Answers reflect my personal opinion, they are not the expression of the ANSES opinion

18 months post-doc position in bioinformatics

DUGMO for Eukaryotic genomes

Conditions:
- out of France for 18 months from May 2020
- work in a P2+ lab

Contact: fabrice.touzain@anses.fr

https://github.com/ANSES-Ploufragan/DUGMO