Vers un portrait moléculaire complet d'une tumeur unique reposant uniquement sur l'ARN

Nikita LAGRANGE

Stage sous la direction de Daniel GAUTHERET, Professeur

M2 Bioinformatique & Modélisation (BIM-BMC)

Préparé au sein de l'Institut de Biologie Intégrative de la Cellule (I2BC)
Équipe Séquence, Structure et Fonction des ARN (SSFA)

30 juin 2022

Plan

1 - Introduction

2 – Matériel et méthodes

3 - Résultats et discussion

4 – Conclusion et perspectives

Plan

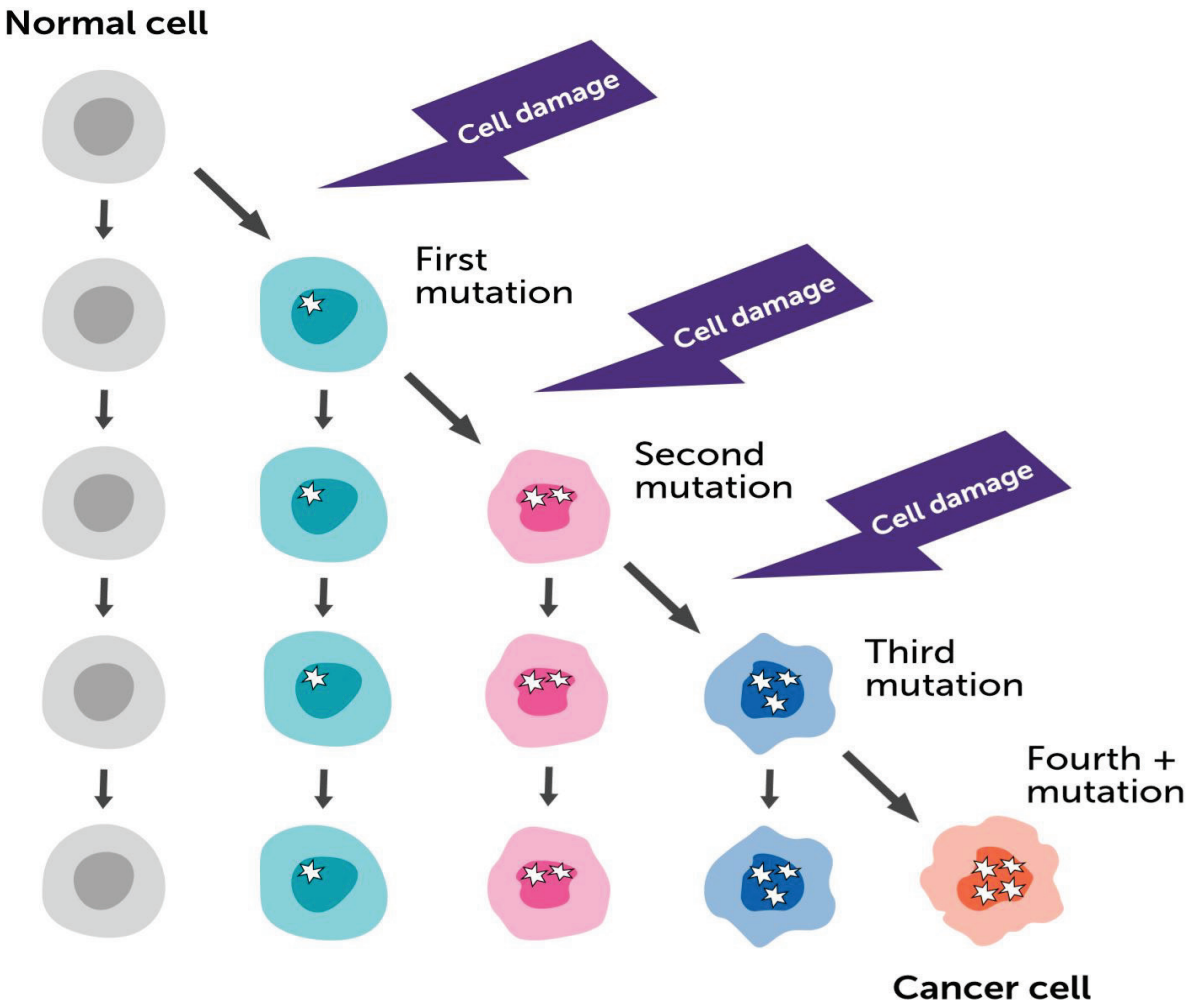
1 - Introduction

2 – Matériel et méthodes

3 - Résultats et discussion

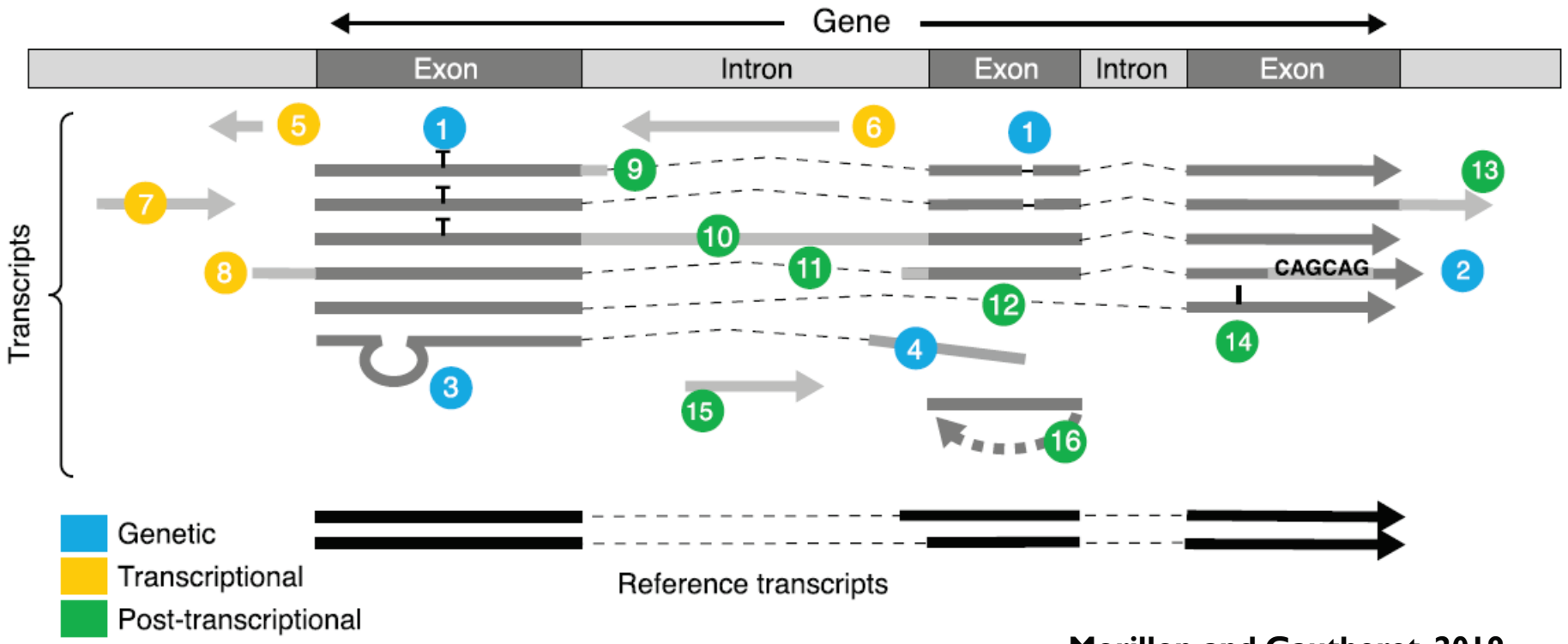
4 – Conclusion et perspectives

Le cancer : une maladie du génome



Adapted from "Understanding Gene Testing" - NIH 1995

RNA-Seq : moyen d'étude complet de la diversité transcriptionnelle



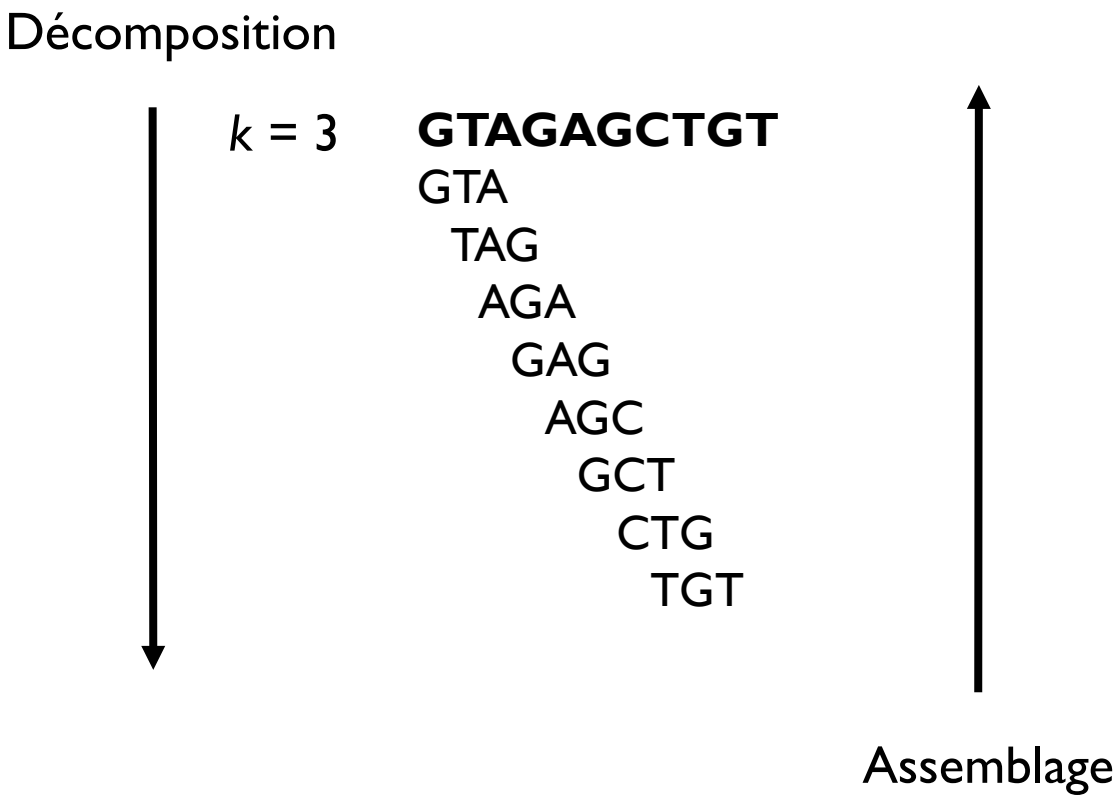
Morillon and Gautheret, 2019

Homo sapiens : ~ 60 *k* gènes et ~ 237 *k* transcrits annotés

$$\frac{Transcrits}{Gènes} \sim 4$$

Analyse du transcriptome : l'approche par k-mers

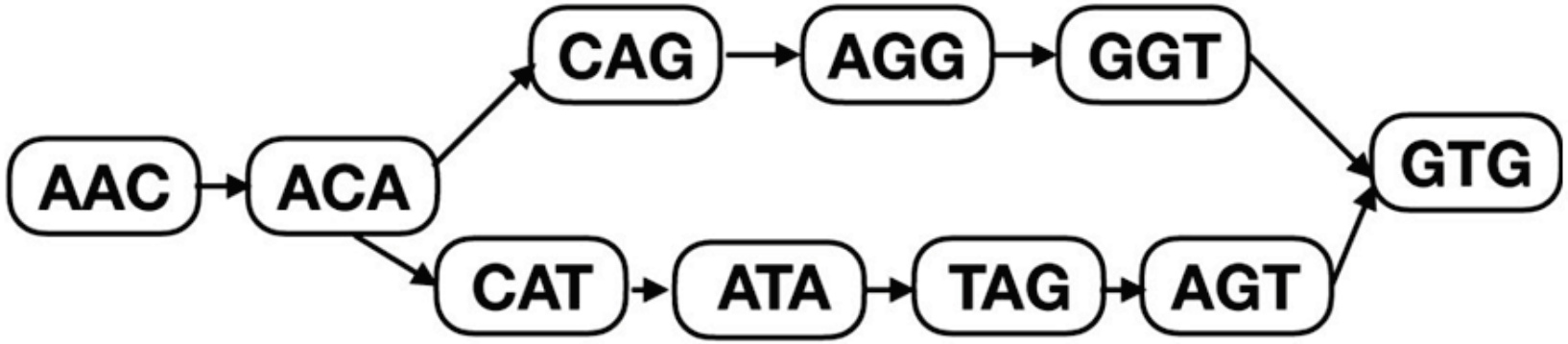
Toutes sous-chaînes de longueur *k* contenues dans une séquence (eg. ADN, read …)



Dans une séquence de taille L
L – k + 1 k-mers

Graphe de De Bruijn

Graphe orienté représentant les chevauchements de longueurs k-1 entre tous les k-mers d'une séquence
Nœuds : k-mer
Arrêtes : chevauchement exact suffixe-préfixe



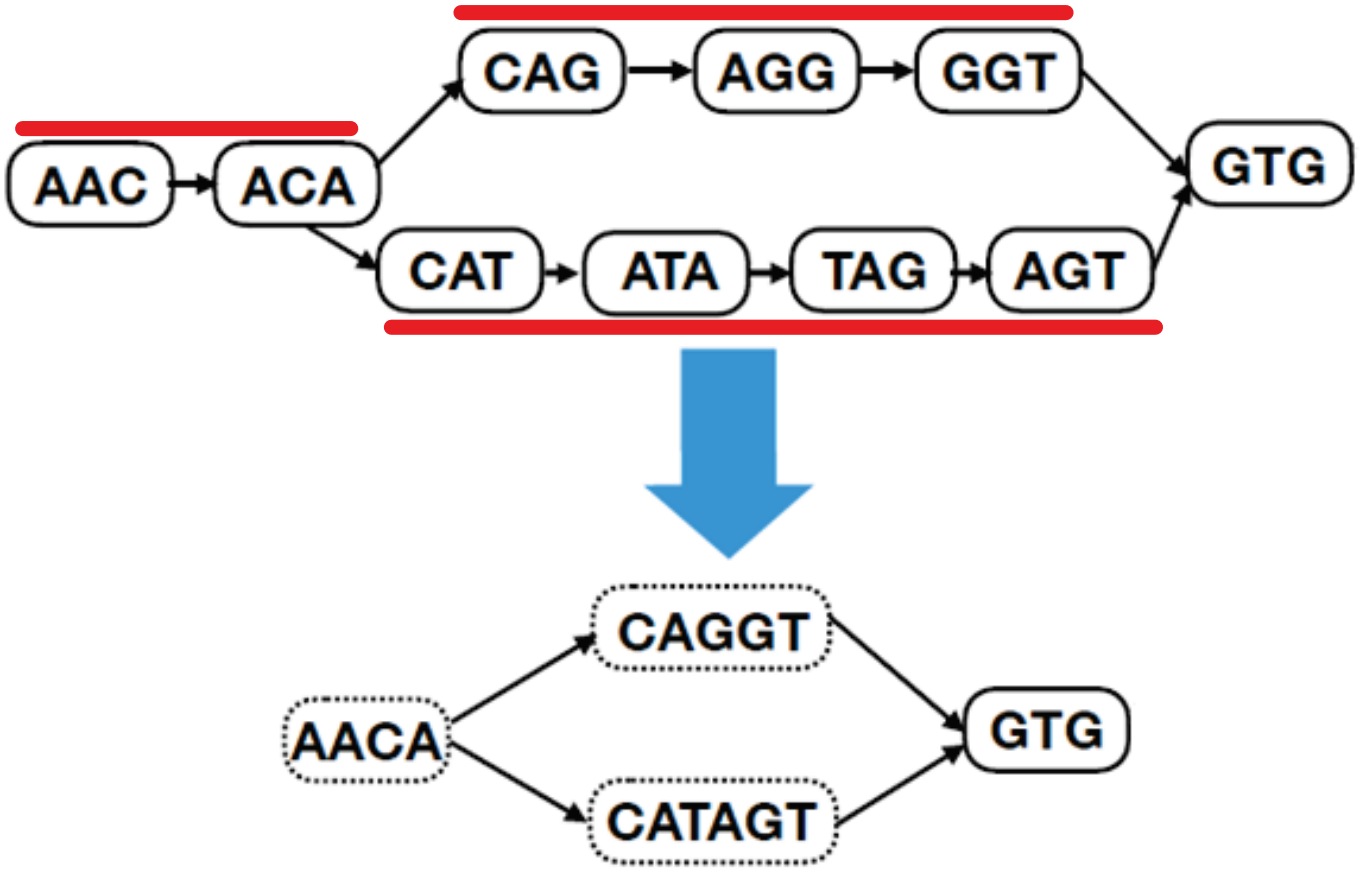
Deux séquences alternatives :

AACAGGTG

AACATAGTG

Graphe de De Bruijn compacté

Réduction des données → gain en temps et en mémoire (compression)
Nœuds : *unitig* (chemin maximal sans embranchement)
Arrêtes : chevauchement exact suffixe-préfixe



Problématique du stage

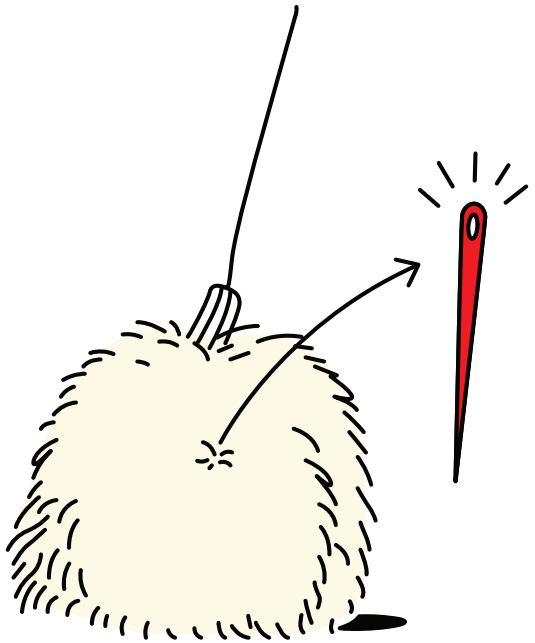
Distinguer ce qui est « normal » de ce qui est pathologique dans un transcriptome ?

Problématique du stage

Distinguer ce qui est « normal » de ce qui est pathologique dans un transcriptome ?

Problématique du stage

Distinguer ce qui est « normal » de ce qui est pathologique dans un transcriptome ?

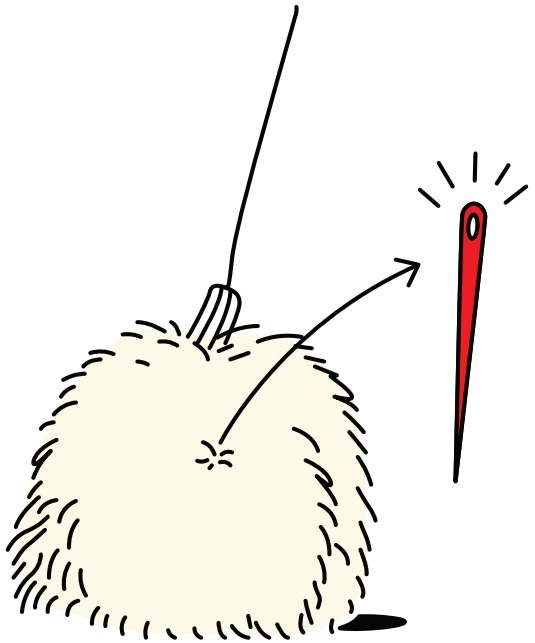


Ordre du million de séquences à analyser

Les transcrits « normaux » prédominent dans une tumeur

Problématique du stage

Distinguer ce qui est « normal » de ce qui est pathologique dans un transcriptome ?



Ordre du million de séquences à analyser

Les transcrits « normaux » prédominent dans une tumeur

Objectif : capturer l'univers des transcrits pathologiques en éliminant les transcrits « normaux »

Plan

1 - Introduction

2 – Matériel et méthodes

3 - Résultats et discussion

4 – Conclusion et perspectives

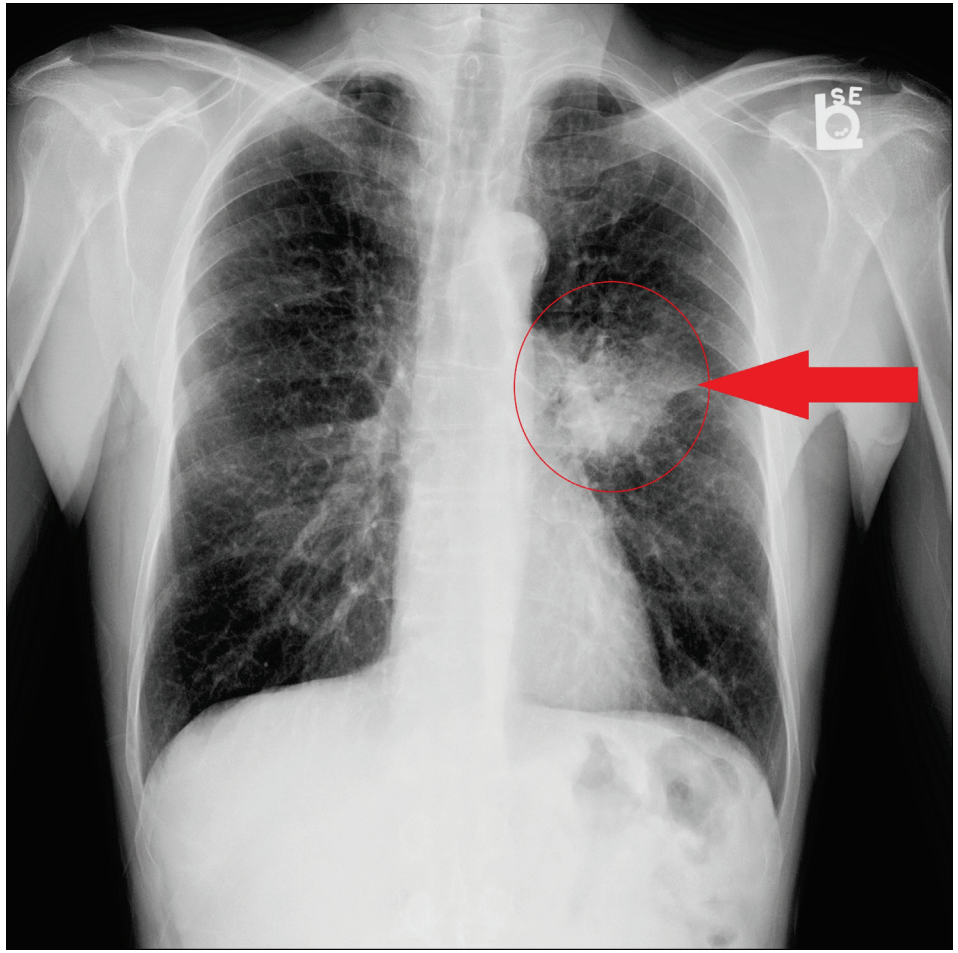
Données

Seo et al. (2012)

Cancer du poumon : adénocarcinome

RNA-seq paired-end et unstranded; 100 nt reads

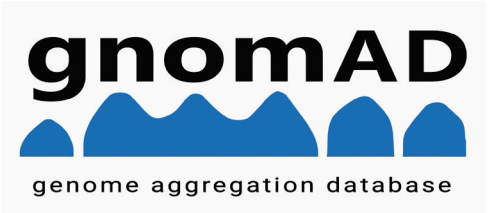
77 patients



Panel de données « normales »

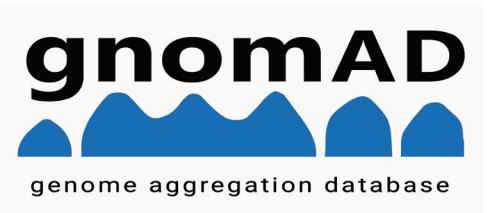
Transcriptome référence *Homo sapiens*

Panel de données « normales »



Recueil de variations polymorphiques exomes + mitochondriaux

Panel de données « normales »



Séquences brutes (FASTA)
Masquage discret/booléen
Tout (1) ou rien (0)

Panel de données « normales »

RNA-Seq « normaux » (136)

Séquences brutes (FASTA)
Masquage discret/booléen
Tout (1) ou rien (0)

Séquences brutes + abondances (FASTQ)
Masquage quantitatif/probabiliste
Masquage fin avec seuil

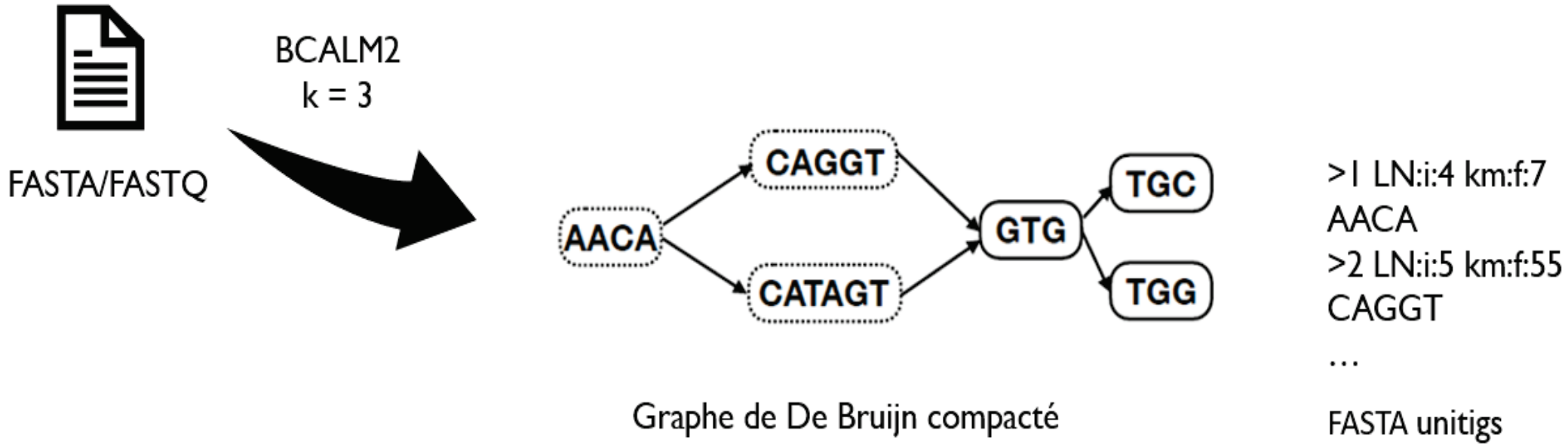

FASTQ

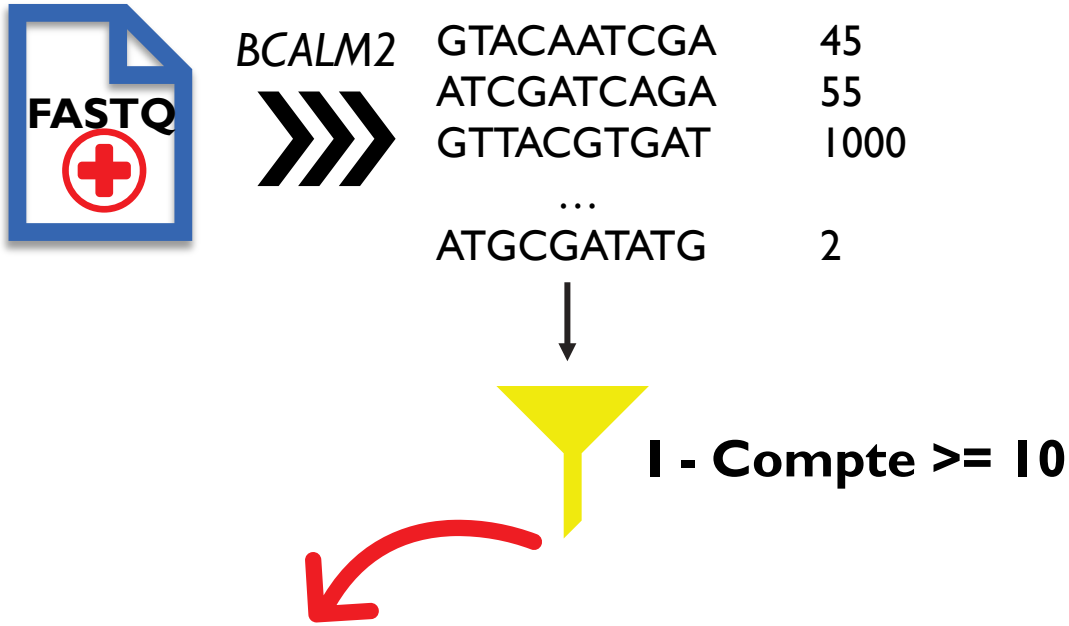
BCALM2

GTACAATCGA	45
ATCGATCAGA	55
GTTACGTGAT	1000
…	
ATGCGATATG	2

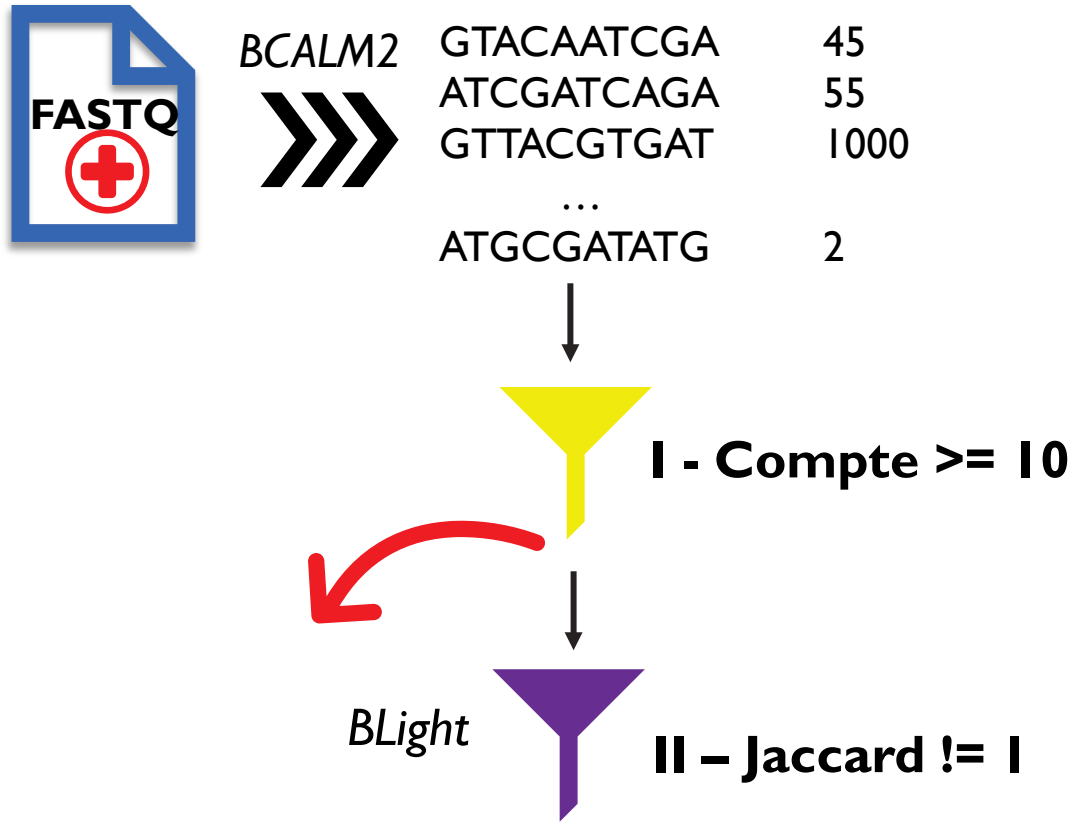
BCALM2 : assemblage local de novo + quantification

Bruijn CompAction in Low Memory 2 Chikhi et al., 2016

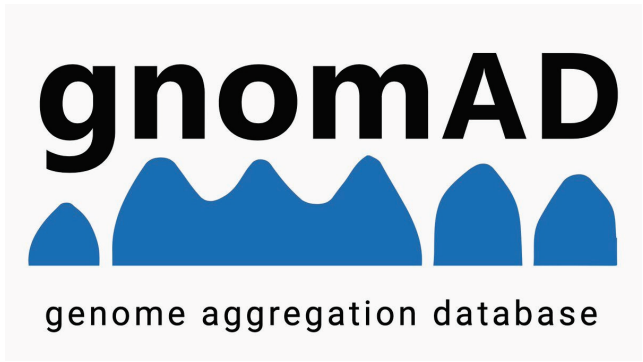




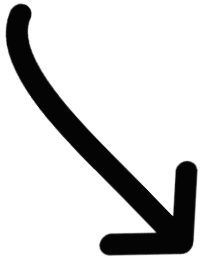
Erreur de séquençage/assemblage
Séquence faiblement exprimée

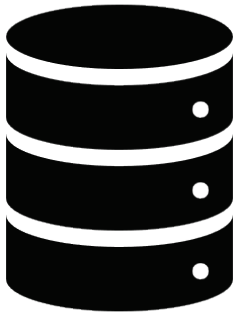

FASTQ
BCALM2
GTACAATCGA 45
ATCGATCAGA 55
GTTACGTGAT 1000
…
ATGCGATATG 2
I - Compte >= 10
BLight
II – Jaccard != 1

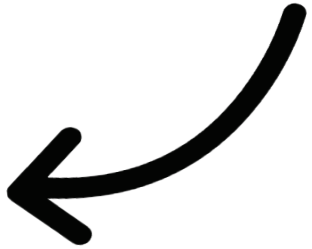
Index de transcrits référencés et polymorphiques



Fréquence allélique > 1 %







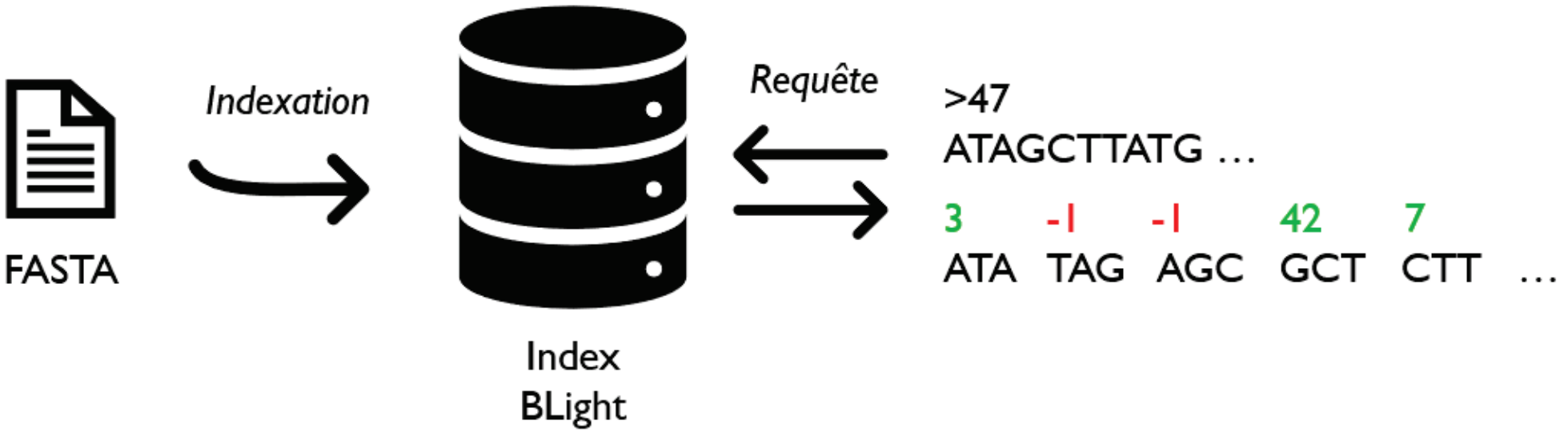
Index GENCODE/gnomAD

BLight : indexage et requête rapide de k-mer

Construction de l'index en quelques secondes/minutes

Marchet et al., 2021

1 millions de requête par seconde avec 1 CPU



Analyse des requêtes : index de Jaccard

A : ensemble des k-mers de la requête (unitig)

B : ensemble des k-mers de l'index GENCODE/gnomAD

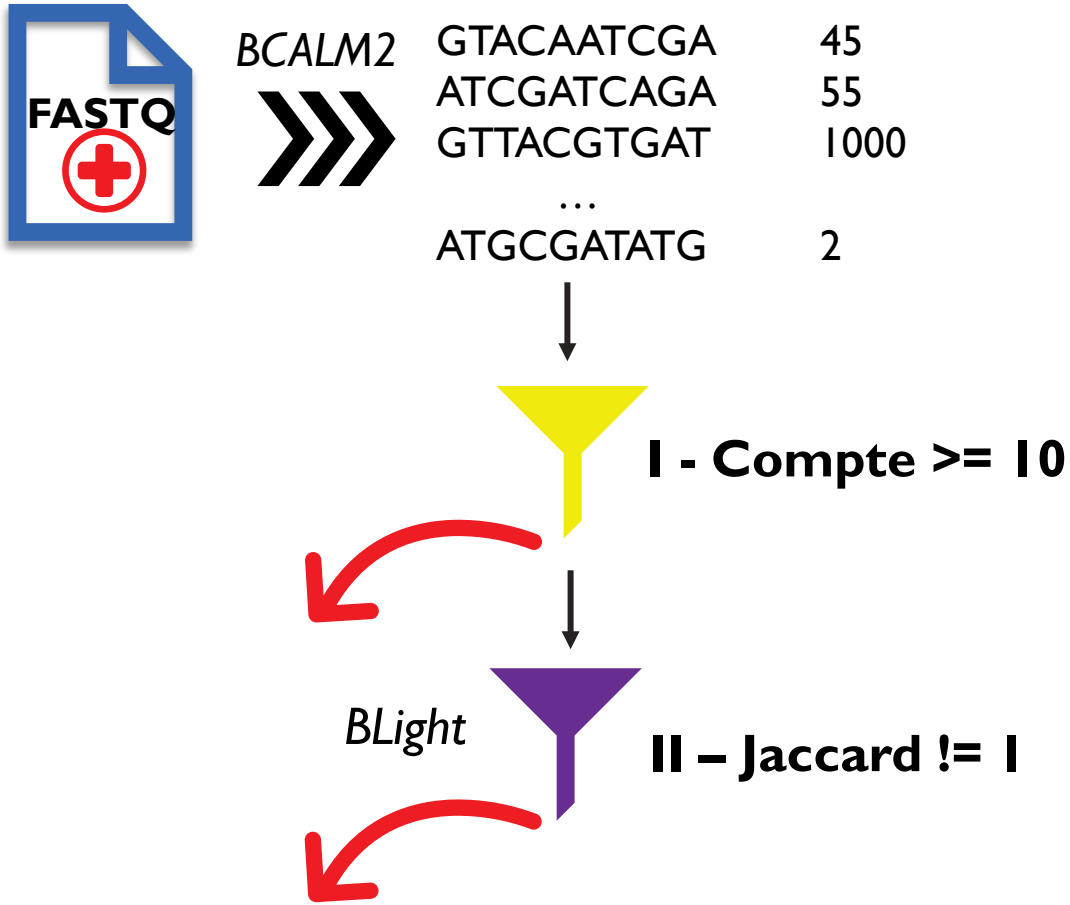
$$J(A,B) = \frac{|A \cap B|}{|A|}$$

$$J(A,B) \in [0,1]$$

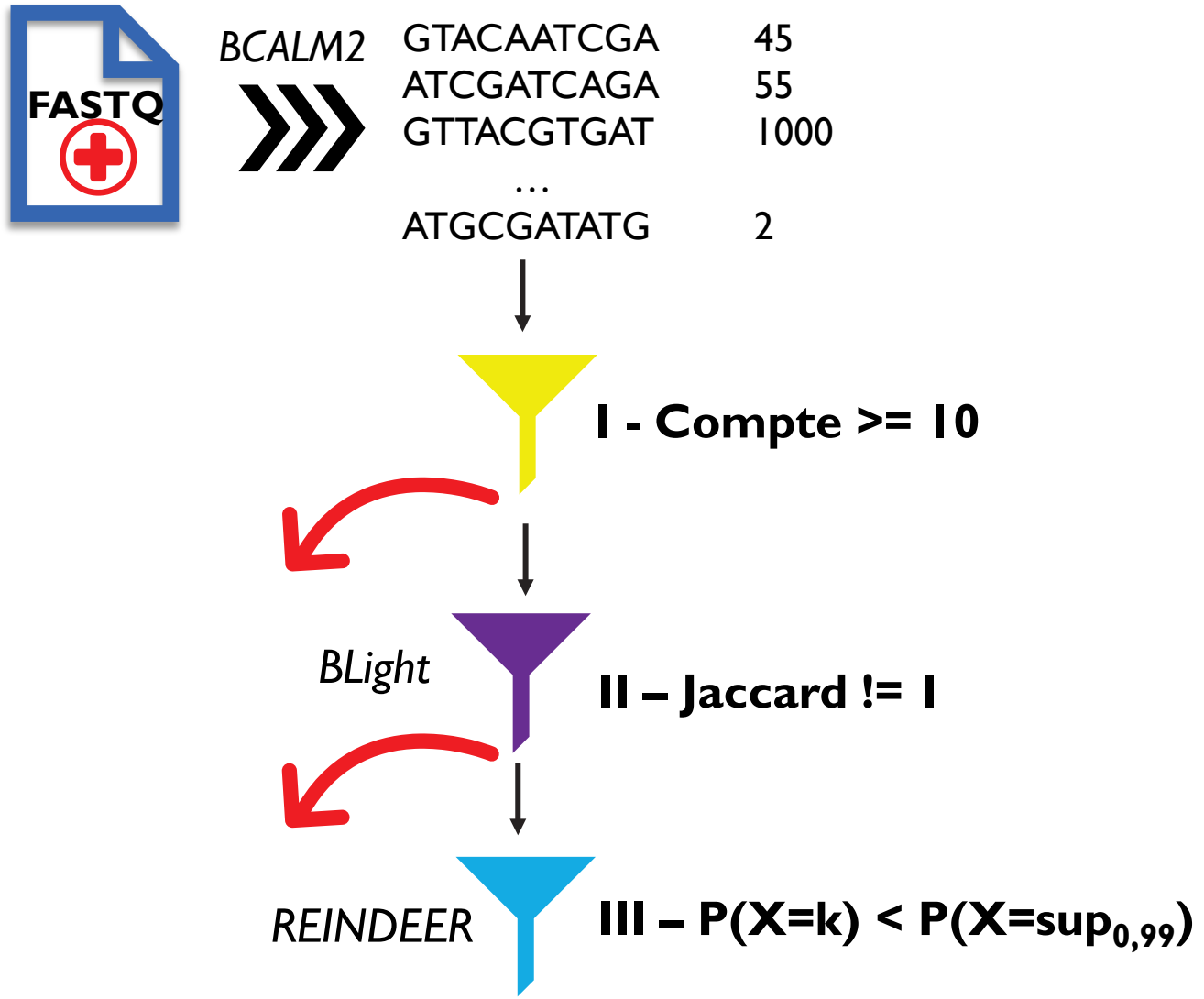
J(A, B) = 0 : requête totalement absente de l'index

J(A, B) = 1 : requête totalement présente dans l'index

0 < J(A, B) < 1 : requête partiellement présente dans l'index

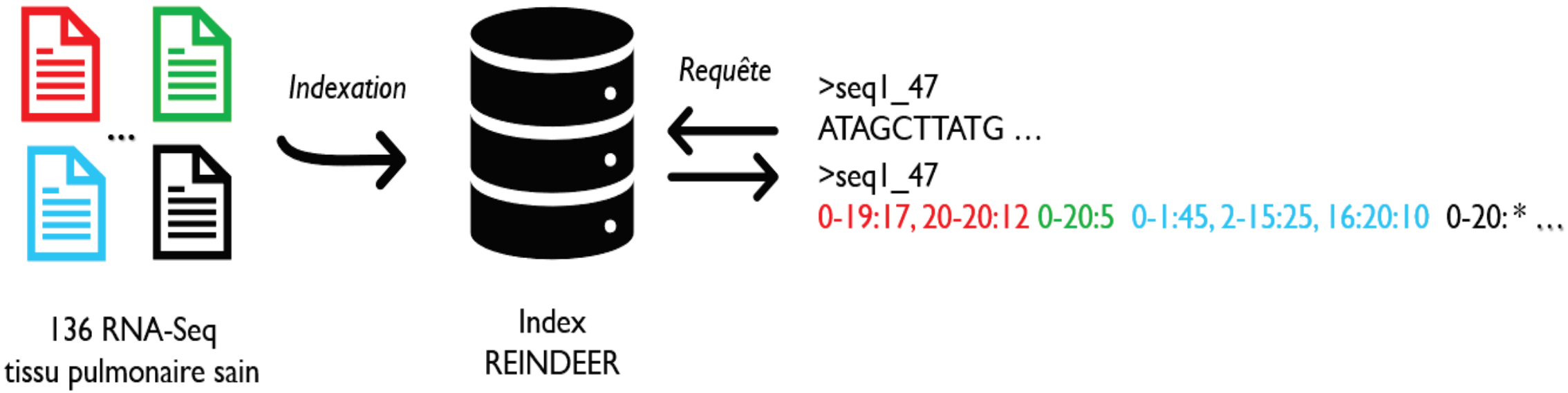


Séquence présente totalement dans l'index GENCODE/gnomAD

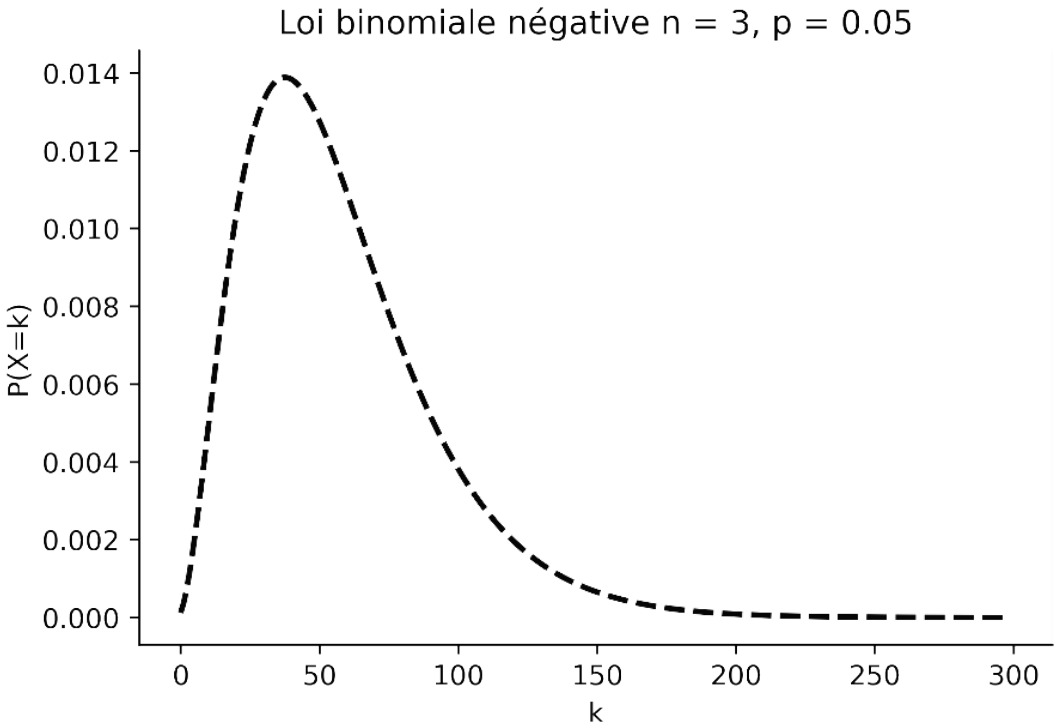

FASTQ
BCALM2
GTACAATCGA 45
ATCGATCAGA 55
GTTACGTGAT 1000
…
ATGCGATATG 2
I - Compte >= 10
BLight
II – Jaccard != 1
REINDEER
III – P(X=k) < P(X=sup0,99)

Index de librairies RNA-Seq « normales »

REad Index for abuNDancE quERy (REINDEER) Marchet et al., 2020



Modélisation de l'abondance par une loi binomiale négative

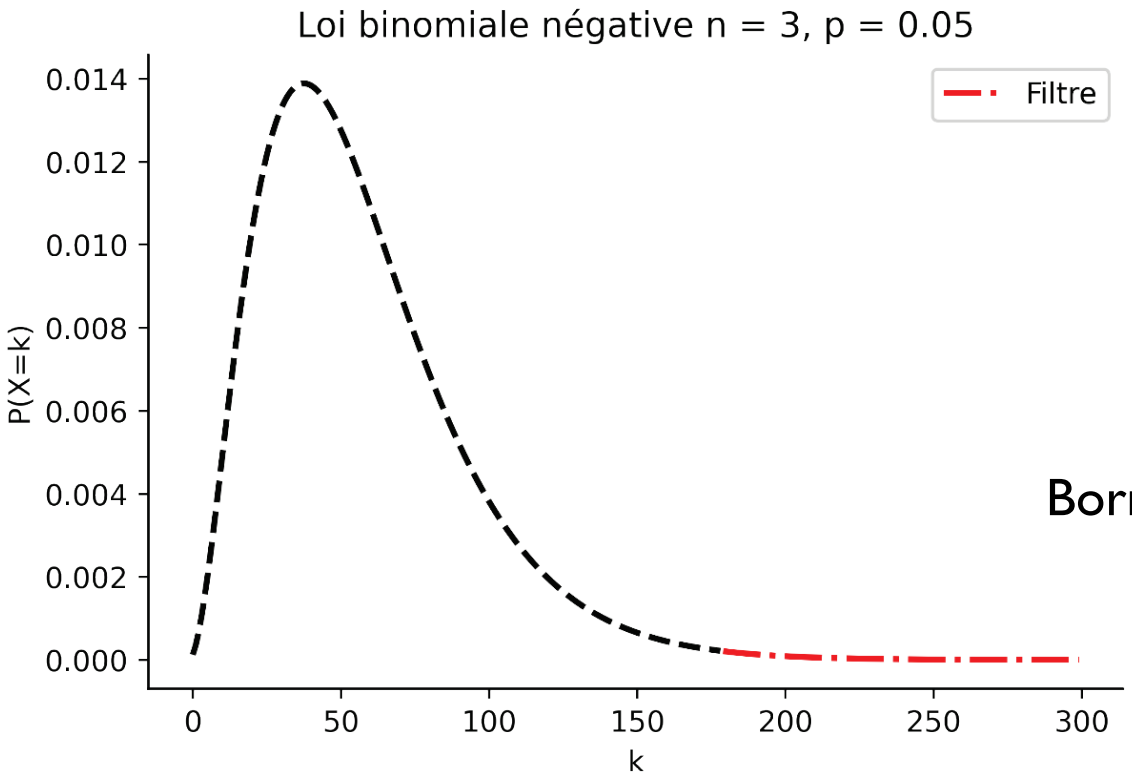


$$X \sim NB(n, p),$$

$$P(X = k) = \binom{k + n - 1}{k} p^n q^k \; \forall k = 0,1, \ldots$$

- n est le nombre d'échecs à observer
- p est la probabilité de succès de chaque expérience de Bernoulli

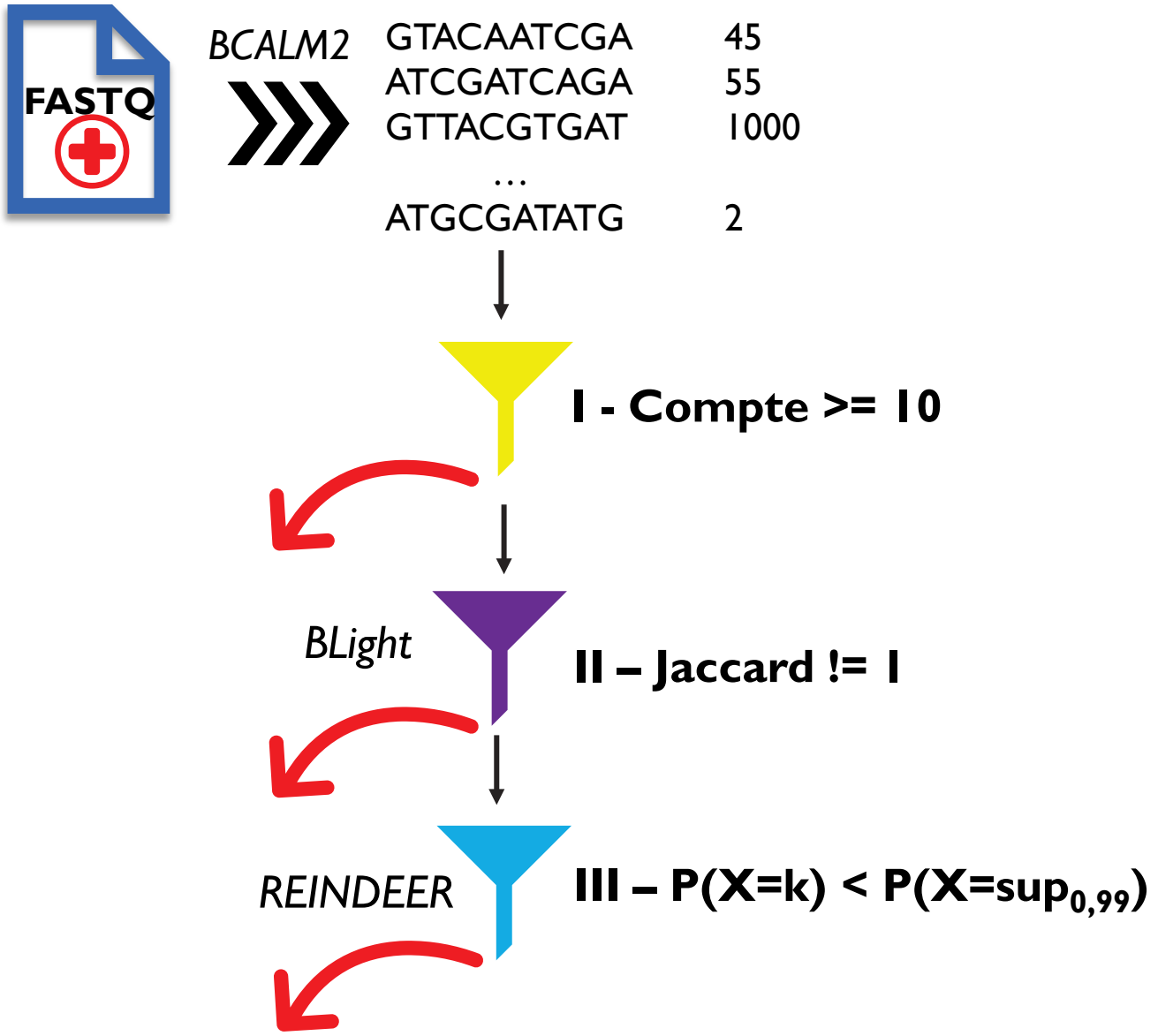
Filtre de la binomiale négative / REINDEER



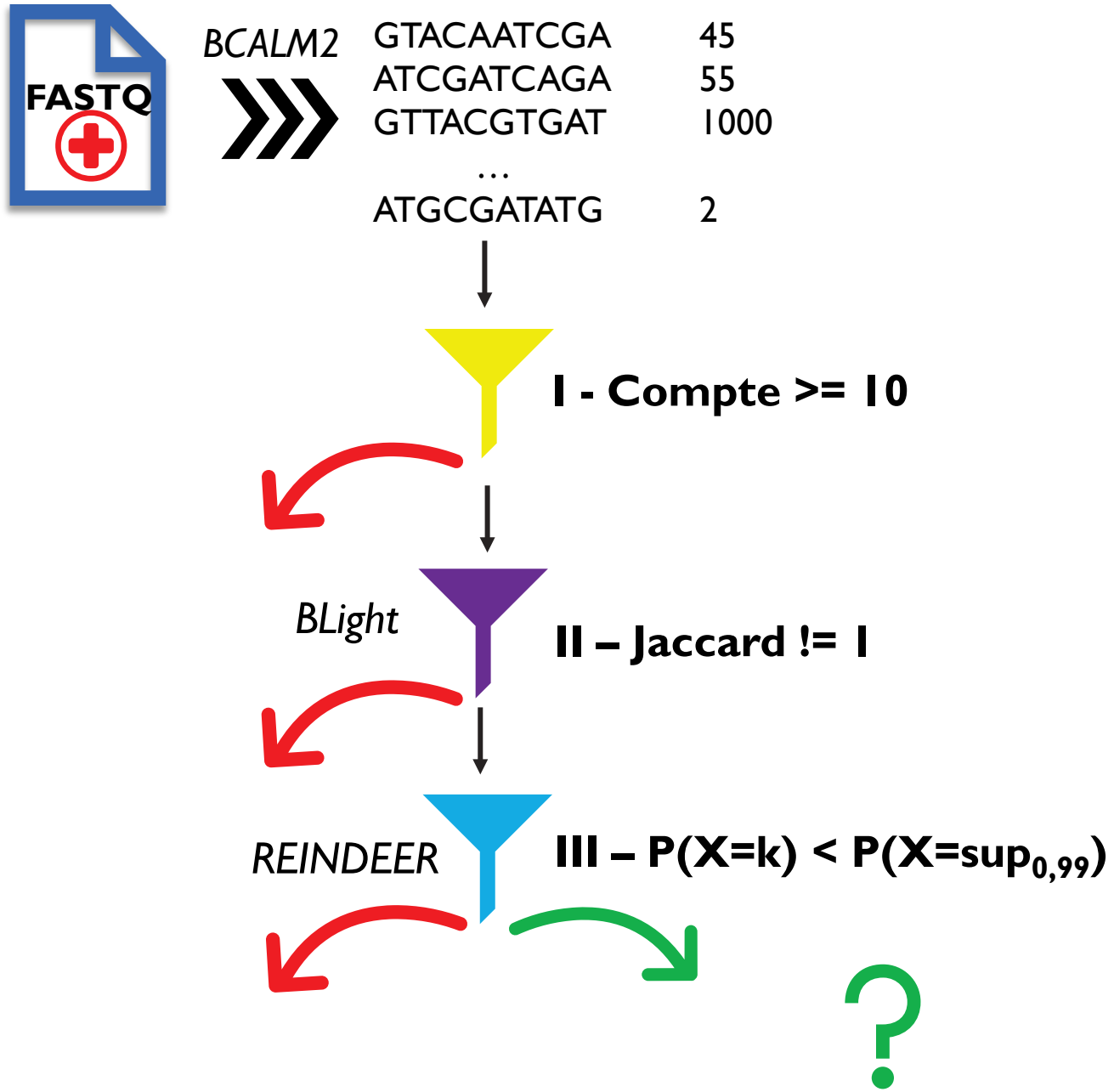
Ici borne supérieure k = 179

P(X=k) < P(X=sup$_{0,99}$)

Borne supérieure de l'intervalle qui contient 99 % de la distribution



Séquence présente dans l'index REINDEER et dont l'abondance est proche de la « normale »


FASTQ
BCALM2
GTACAATCGA 45
ATCGATCAGA 55
GTTACGTGAT 1000
…
ATGCGATATG 2
I - Compte >= 10
BLight
II – Jaccard != 1
REINDEER
III – P(X=k) < P(X=sup0,99)
?

Que reste-t-il … ?

- Séquence suffisamment exprimée (filtre 1)

Que reste-t-il … ?

- Séquence suffisamment exprimée (filtre I)

- Séquence non référencée et non polymorphique dans la population (filtre II)

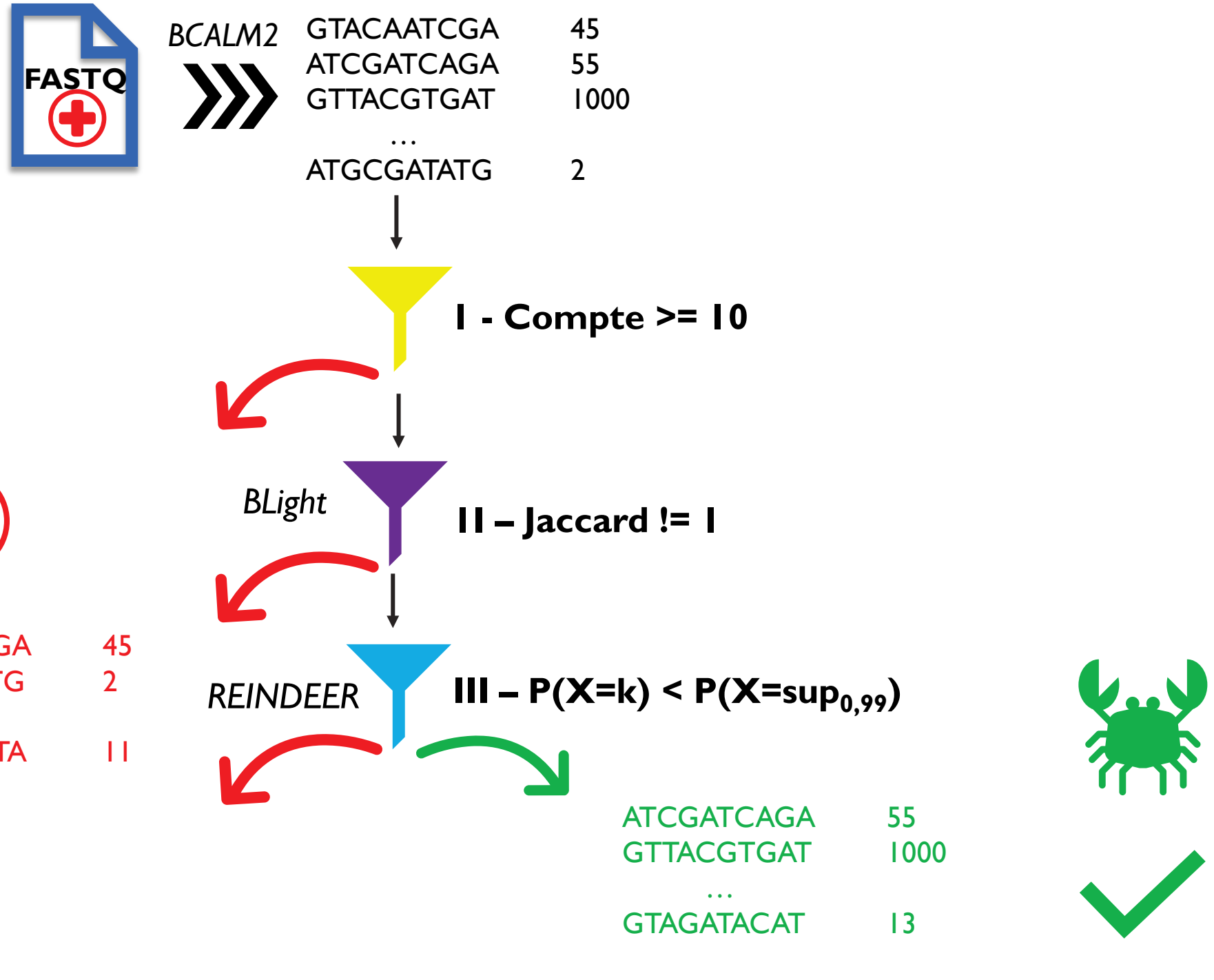
Que reste-t-il … ?

- Séquence suffisamment exprimée (filtre I)
- Séquence non référencée et non polymorphique dans la population (filtre II)
- Séquence absente dans les libraires RNA-Seq « normale » ou surexprimée par rapport à la normale (filtre III)

Que reste-t-il … ?

- Séquence suffisamment exprimée (filtre I)
- Séquence non référencée et non polymorphique dans la population (filtre II)
- Séquence absente dans les libraires RNA-Seq « normale » ou surexprimée par rapport à la normale (filtre III)

✓ Séquence qui n'est pas « normale »


FASTQ
BCALM2
GTACAATCGA 45
ATCGATCAGA 55
GTTACGTGAT 1000
…
ATGCGATATG 2
I - Compte >= 10
BLight
II – Jaccard != 1
REINDEER
III – P(X=k) < P(X=sup0,99)
GTACAATCGA 45
ATGCGATATG 2
…
TTGAATGCTA 11
ATCGATCAGA 55
GTTACGTGAT 1000
…
GTAGATACAT 13

Plan

1 - Introduction

2 – Matériel et méthodes

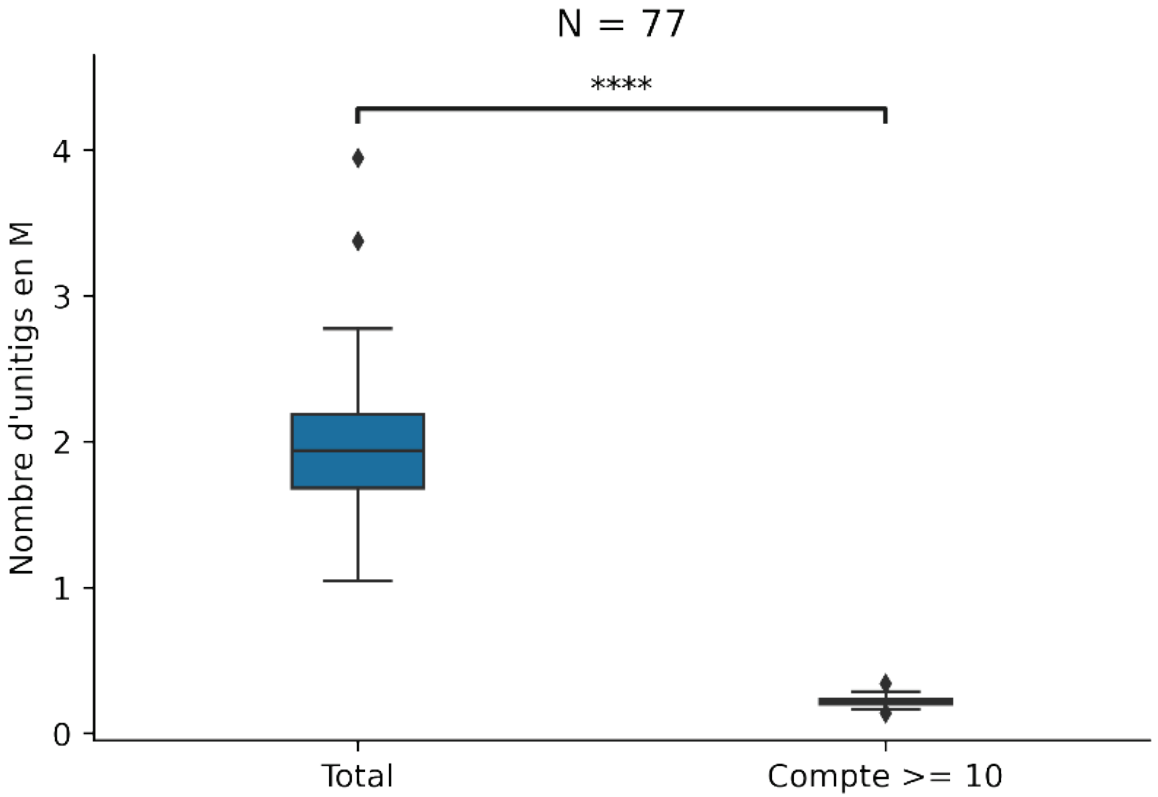
3 - Résultats et discussion

4 – Conclusion et perspectives

Données

	Nombre	Taille FASTQ	Total reads	Taille moyenne d'un read
Cancer	77	661 GB	6.4 G	93 nt
Normal	136	710 GB	11 G	74 nt

1er filtre : filtre de compte



Médiane Total : **2 M**

Médiane Compte >= 10 : **217 k**

**** : p<= 1e-04

Réduction de 89 %

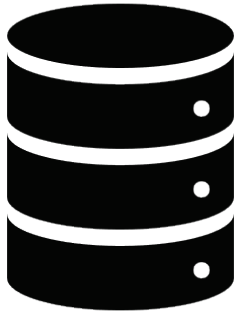
Index BLight GENCODE/gnomAD

162 701 116 k-mers de taille 31

382 MB

Index REINDEER RNA-Seq « normaux »

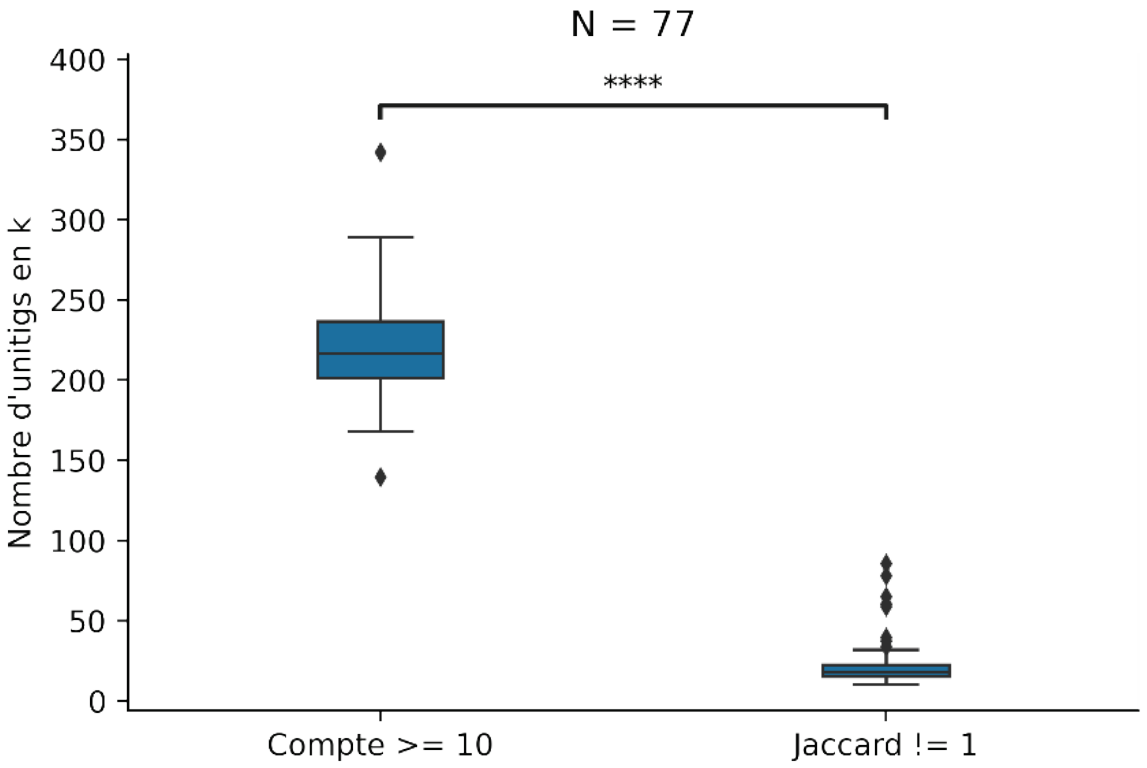
136 RNA-Seq « normaux » poumon



494 323 372 k-mers de taille 31

12 GB

2ème filtre : Index de Jaccard / BLight

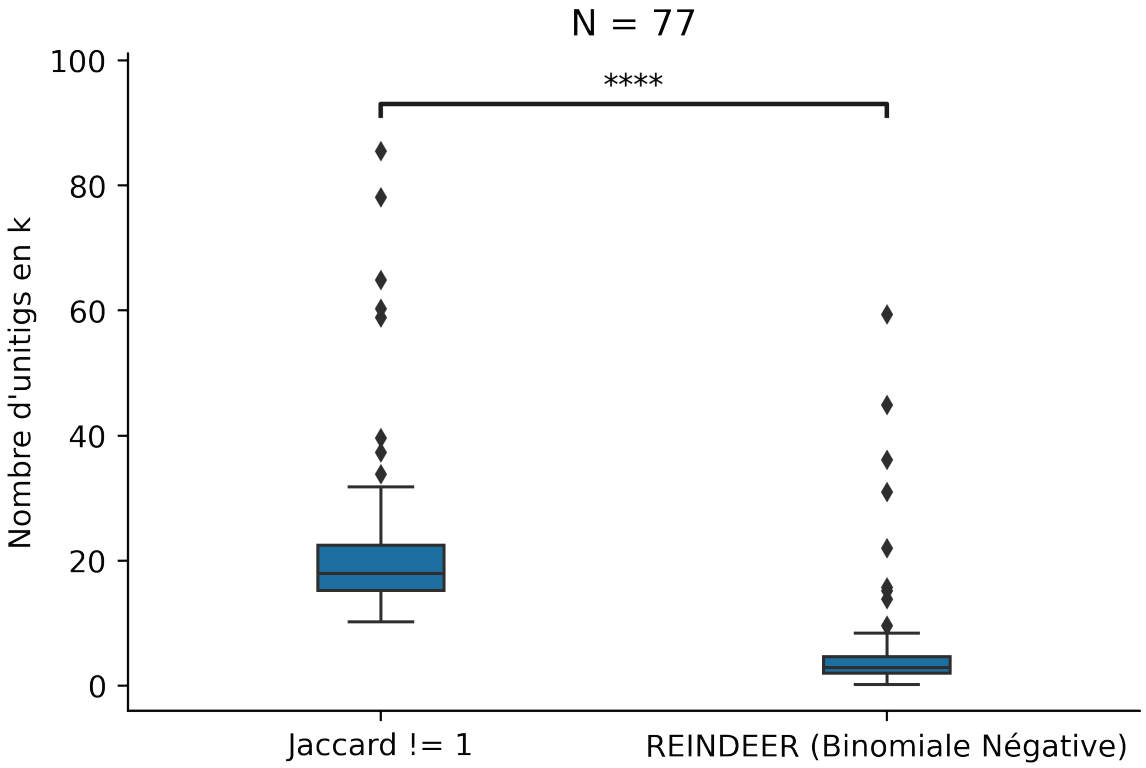


Médiane Compte >= 10 : **217 000**

Médiane Jaccard !=1 : **18 010**

Réduction de 92 %

3ème filtre : Binomiale négative / REINDEER



Médiane Jaccard !=1 : **18 010**

Médiane Binomiale négative : **2 955**

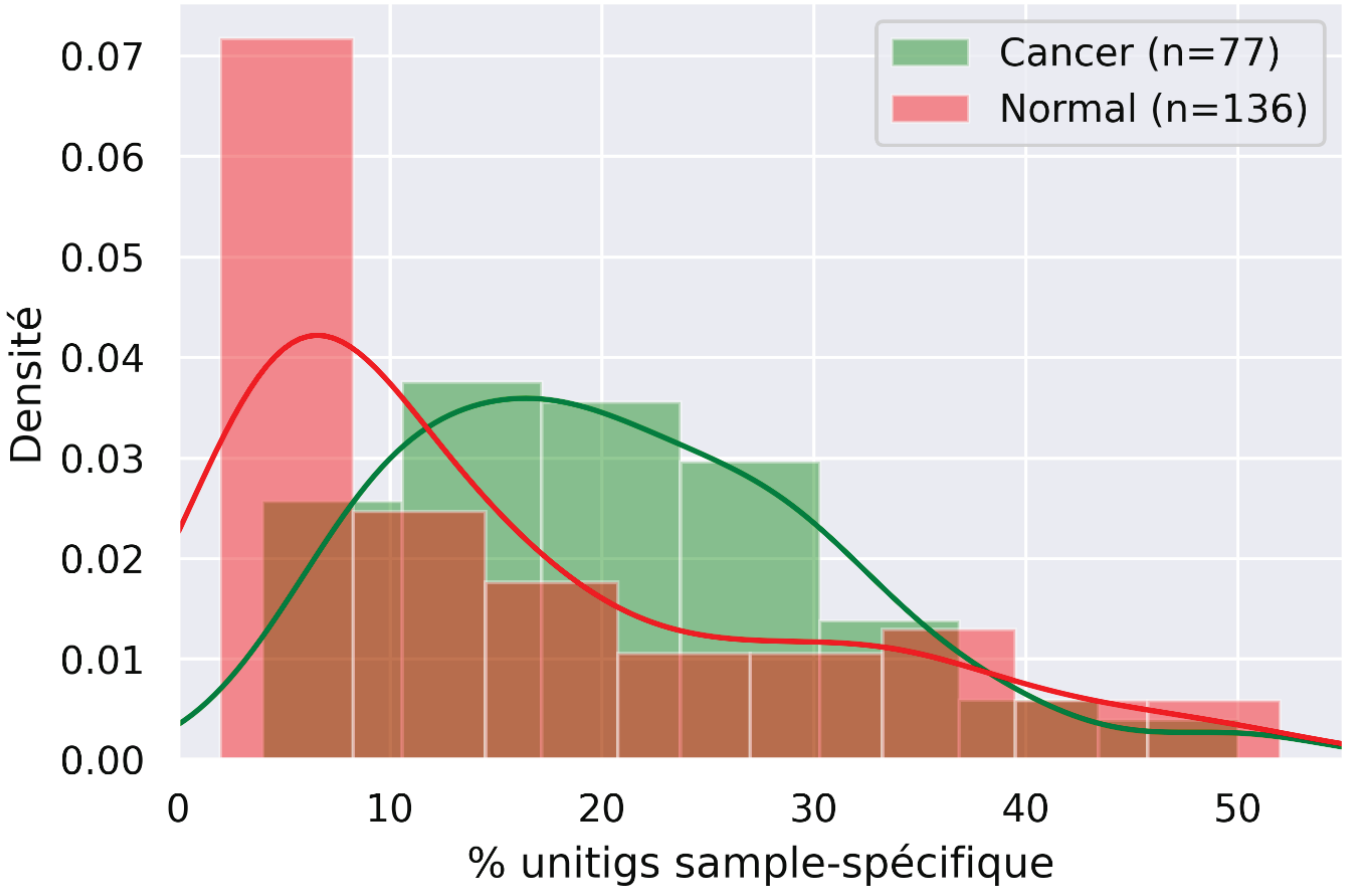
Réduction de 84 %

Comparaison avec les échantillons « normaux »

	Moyenne unitigs BCALM2	Moyenne unitigs filtre 1	Moyenne unitigs filtre 2	Moyenne unitigs filtre 3	Médiane séquences restantes	Médiane % séquences éliminées
Cancer	**1.97 M**	221 k	26 k	5 896	**2 955**	99.85 %
Normal	**1.98 M**	279 k	29 k	3 198	**1 372**	99.92 %

****

Comparaison avec les échantillons « normaux »



Médiane normal : 10.5 %

Médiane cancer : 19 %

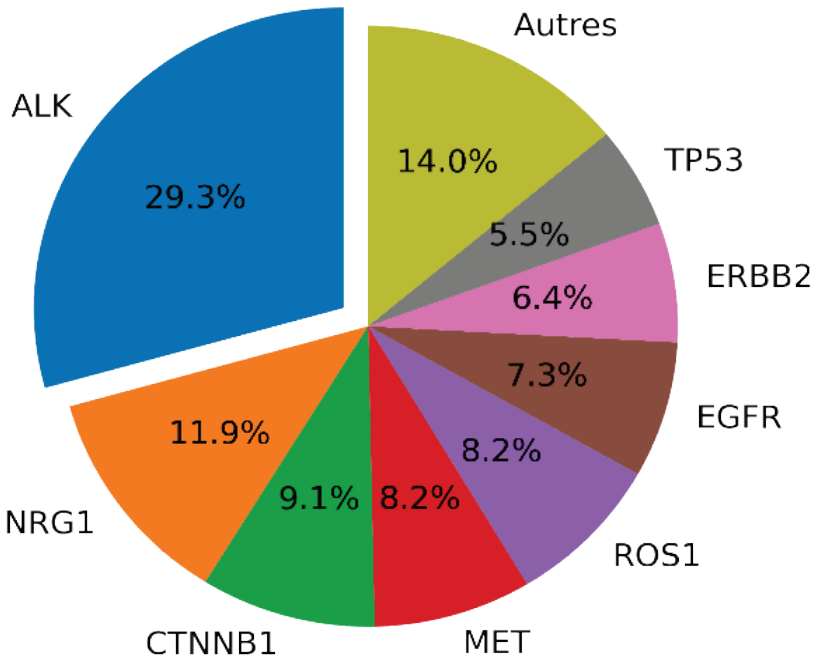
** : p = 2.1^{e}-03

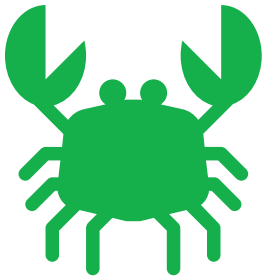
Sample-spécifique : unitig retrouvé uniquement dans une librairie RNA-Seq

Annotation des fragments d'ARN restant

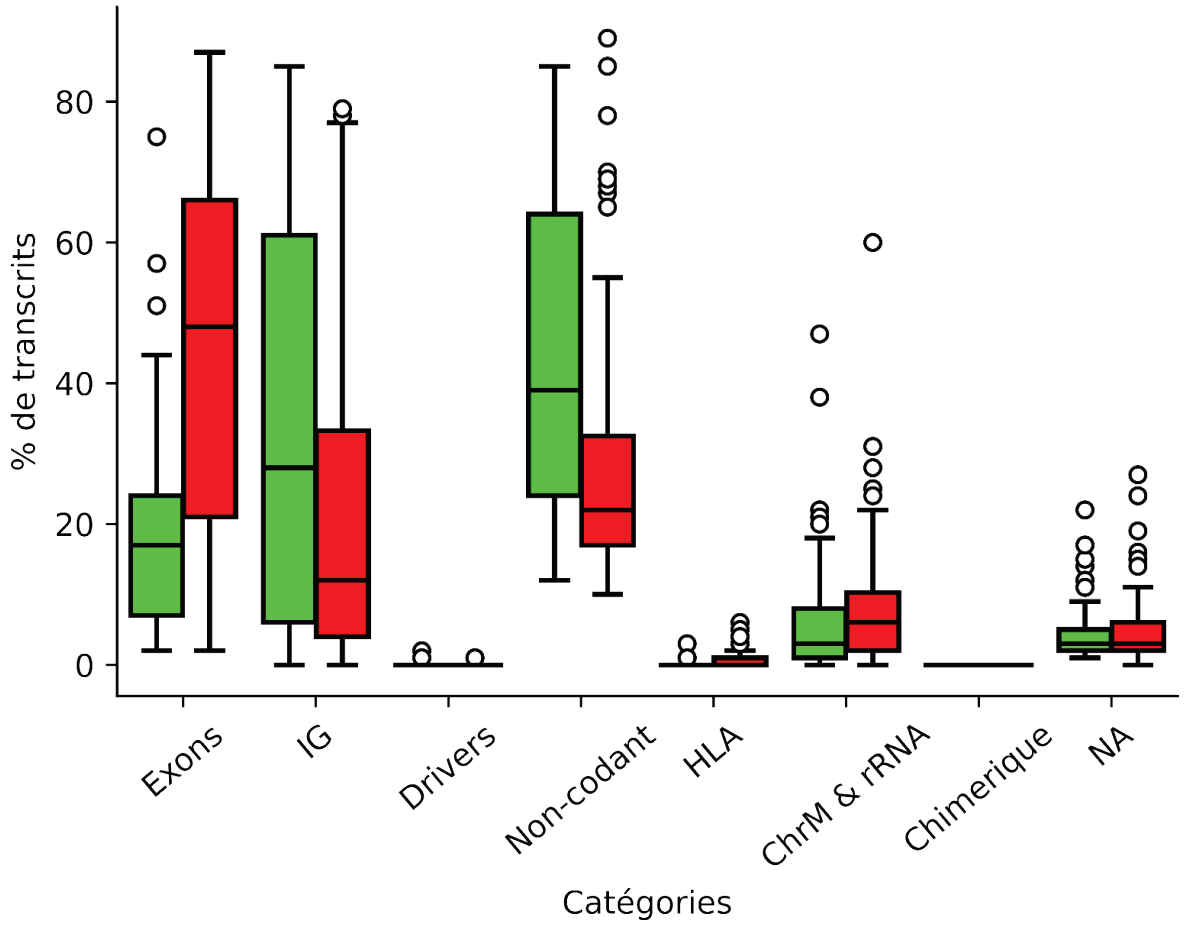
Absence de gènes drivers dans 17 échantillons cancers (22 %)

Absence de gènes drivers dans 51 échantillons normaux (37.5 %)



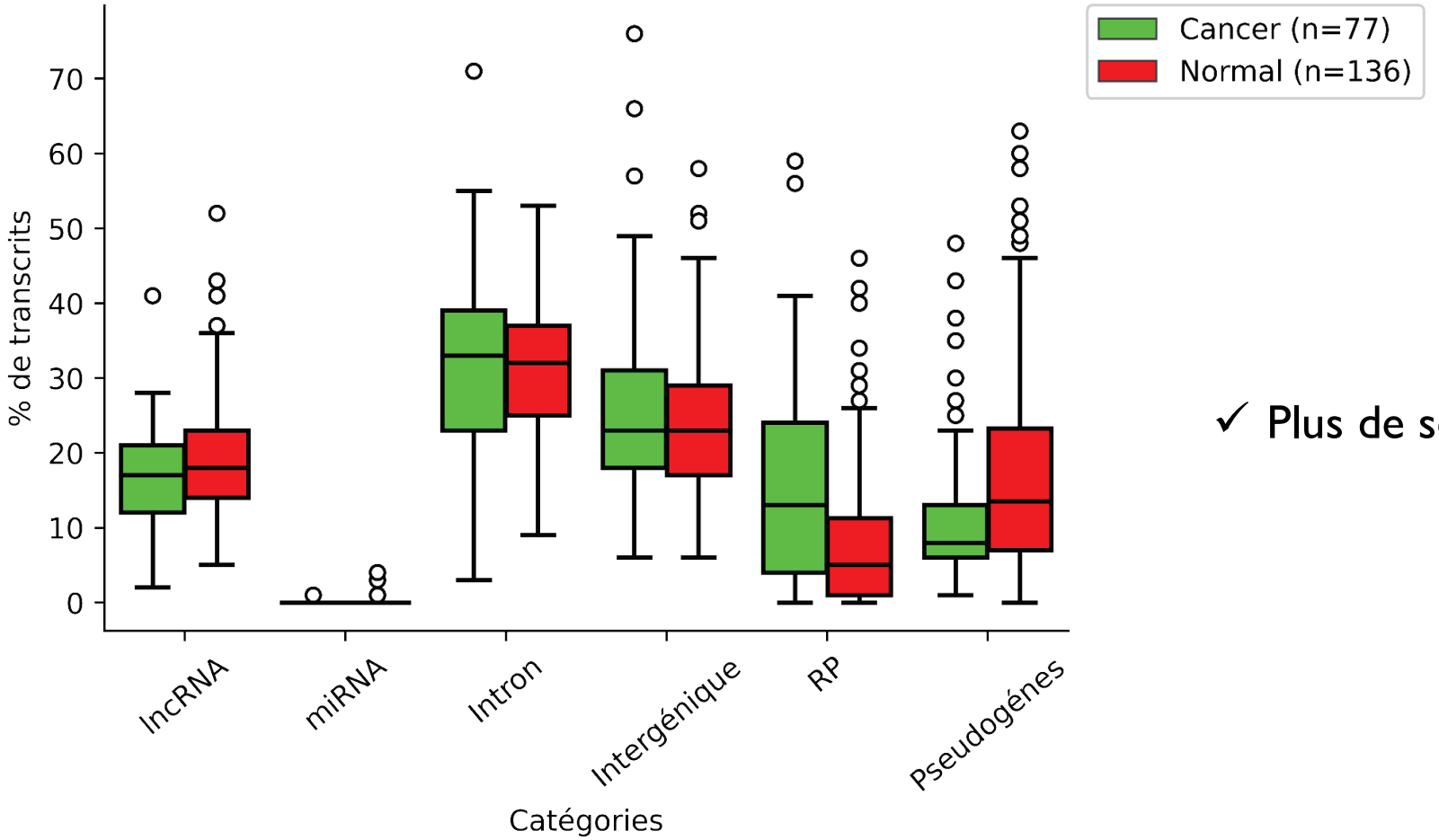


Annotation des fragments d'ARN restant



- ✓ Moins d'exons (17 % contre 48 %)
- ✓ Plus d'immunoglobulines (IG, 28 % contre 12 %)
- ✓ Plus de non-codant (39 % contre 22 %)

Annotation des fragments d'ARN restant



✓ Plus de séquences répétées (13 % contre 5 %)

Plan

1 - Introduction

2 – Matériel et méthodes

3 - Résultats et discussion

4 – Conclusion et perspectives

Conclusion

- Développement d'un nouvel outil d'analyse de données RNA-Seq
- Outil qui a pour objectif de conserver des transcrits spécifiques d'une tumeur à l'aide d'une approche couplée avec référence et sans référence
- Outil de recherche : découvrir de nouvelles altérations du cancer
- Elimination de 99.85 % des séquences de départ (2 millions à 3 000 unitigs)
- Séquences restantes constituent base d'un portrait moléculaire d'une tumeur d'un patient (mutations drivers, ARN non codant, séquences répétées ...)

Perspectives

✓ Étendre l'index REINDEER en intégrant de nouvelles données « normales » e.g. projet GTEx

✓ Index de Jaccard comme option au protocole

✓ Protocole généralisable à d'autres questions biologiques :

- Comparaison de transcriptome entre différents individus
- Comparaison de transcriptome entre différents organes
- Comparaison de transcriptome entre différents tissus (e.g. sain/pathologique (maladie héréditaire))