

# Session de Formation

## « analyse de données : génomes & transcriptomes »

### Mardi 28 Octobre 2014

Salle Informatique - Formation Permanente  
INRA Auzeville

Mardi 28

Module	Horaire	Durée
Introduction et présentation du portail BBRIC	9H00-10H15	1 h 15
Pause	10H15-10H30	15 min
Assemblage de novo de transcriptomes avec mesure de l'expression par banque	10H30-12H30	2 h
Repas – déjeuner	12H30-13H30	1 h
Mesure de l'expression a partir de données RNAseq	13H30-15H00	1h30
Pause	15H00-15H15	15 min
Détection de transferts horizontaux	15H15-16H15	1h

Modules	Responsable et intervenant principal	Expert	Relecteur
Introduction et présentation du portail BBRIC	Ludovic Legrand , Sébastien Carrere		
<b>Assemblage de novo de transcriptomes avec mesure de l'expression par banque</b>	Sébastien Carrere	Anthony Bretaudeau IGEPP/Rennes	Erika Sallet
Mesure de l'expression a partir de données RNAseq	Erika Sallet	Ludovic Legrand	Sébastien Carrere
Détection de transferts horizontaux	Ludovic Legrand	Corinne Rancurel ISA/Sophia	Erika Sallet



---

# Introduction

## Sébastien Carrere





- BBRIC est l'un des 7 CATI « orientés bioinformatique »

## Définition de la communauté servie BBRIC

Chercheurs/ingénieurs biologistes issus des laboratoires sous tutelle principale SPE et chercheurs/ingénieurs d'autres unités avec un agent affilié au CATI.

## Principes

- On ne peut pas tout faire
- On n'a pas vocation à faire ce que les autres doivent faire

# Domaines d'applications

- ▶ Assemblage (genomes/transcriptomes)
- ▶ Annotation structurale des génomes (gènes codants pour des protéines / ncRNA)
- ▶ Annotation fonctionnelle (genomes/transcriptomes)
- ▶ Analyse de l'expression (RNAseq, puces)
- ▶ Détection et analyse du polymorphisme
- ▶ Métagénomique
- ▶ Epigénomique
- ▶ Modélisation des réseaux métaboliques et de régulation.
- ▶ Gestion de collections (bactéries, insectes, etc.)
- ▶ Systématique: identification des espèces de groupes d'espèces d'intérêt
- ▶ Phylogénie, évolution
- ▶ Génomique des populations

- *Modèles biologiques: des dizaines (Plantes, insectes, bactéries, champignons, oomycètes, nématodes, virus, etc.)*

- Toujours en amont du dépôt des projets
  - avis à donner sur les données produites par rapport à un savoir faire ou des expériences.
  - avis à donner sur le calendrier et les moyens nécessaires
- Si possible avec les membres locaux du CATI, sinon avec le responsable du CATI.
- Arbitrage au fil de l'eau
- **Formation: rendre autonome les biologistes sur les activités d'analyse « maitrisées » et automatisées.**
  - 1<sup>ere</sup> Session de formation « génome & transcriptome » : 24 & 25 avril 2014 à Paris

# Objectifs des formations BBRIC

1. Illustrer à travers des exemples variés comment nous allons interagir avec les biologistes
  - à distance et sur un temps long
  - à travers quelques principes de fonctionnement
  - grâce aux outils que nous mettons à disposition.
2. Rendre autonomes les utilisateurs sur les tâches d'analyse de données récurrentes et automatisées.
3. Illustrer un savoir faire bioinformatique pour la conception de pipeline d'analyses bioinformatiques « ad hoc ».
  - Une session de formation par an à Paris avec la création d'un support pédagogique qui permettra, si nécessaire, de rejouer la formation dans les laboratoires.
  - Essayer d'être complémentaire des nombreuses autres formations autour de la bioinformatique (galaxy, analyse RNAseq, etc.)



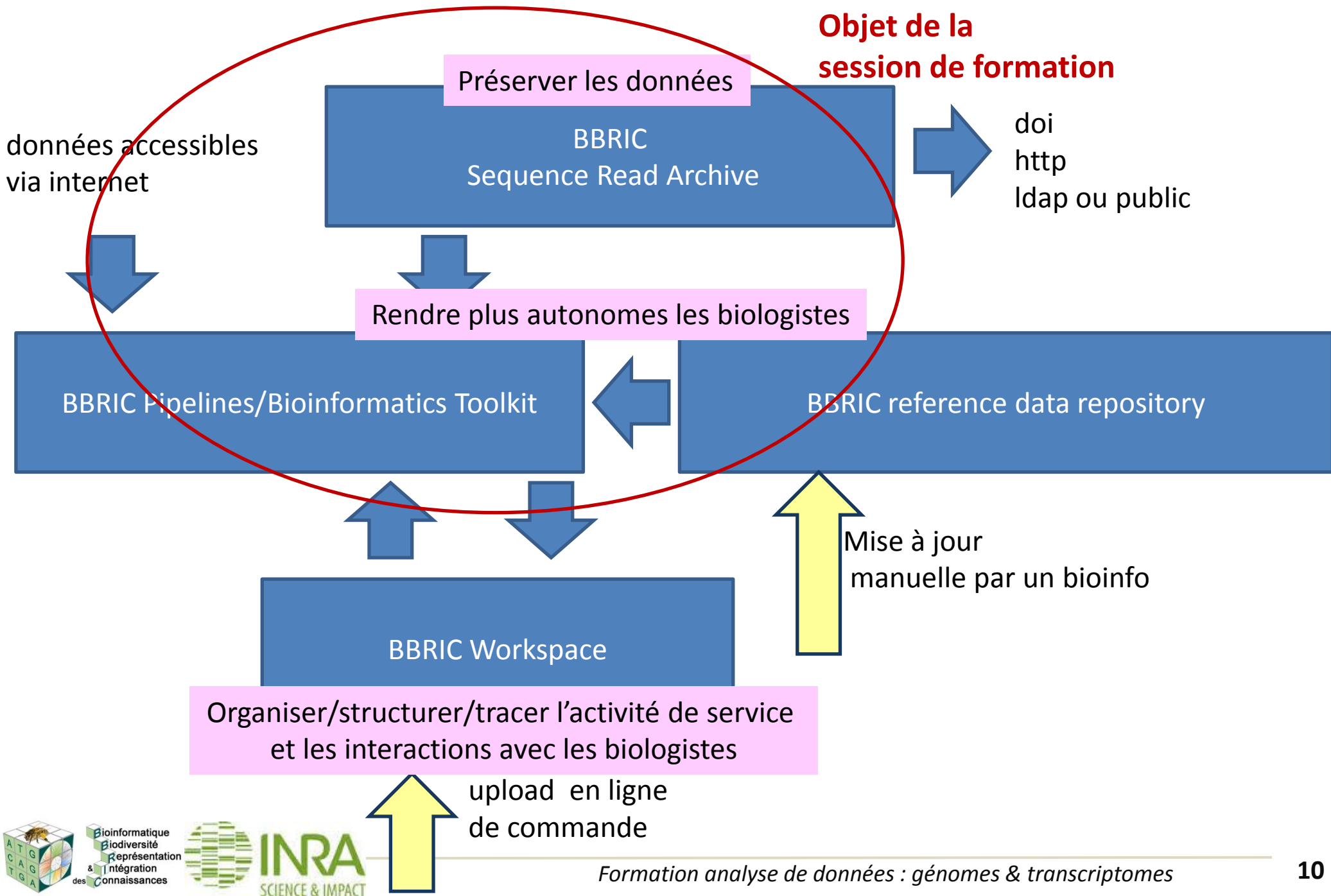
---

# Présentation de l'architecture bioinformatique et du portail BBRIC

Sébastien Carrere



# Architecture bioinformatique cible



- <https://bbric.toulouse.inra.fr>



*Portail Bioinformatique de la communauté BBRIC*



Welcome

Dejavu/Bioinfo	Sequence Archive	Protocols

## En termes d'analyse de données

- Il n'y a pas UNE façon de faire les choses
  - Pour répondre à une question, les données et les programmes changent au fil du temps
  - On n'a pas forcément *a priori* connaissance de là où se trouve l'outil qui va permettre de répondre à une question
- ➔ On interroge le système BBRIC a partir d'un problème (ex: Bactérie/Assemblage)



[Home](#)[Analysis Protocols](#)[Archive](#)[Login](#)

Quick search:

 Include BBRIC Archive[Query BBRIC protocols](#)

Species:

Insects (9)  
Plants (8)  
Metazoa (6)  
**Bacteria (6)**  
Fungi (6)

Applications:

Gene expression data processing (4)  
Transcriptome assembly (4)  
Gene and gene component prediction (3)  
File reformatting (2)  
**Sequence assembly (1)**  
Phylogenomics (1)  
Protein site detection (1)



Bioinformatique  
Biodiversité  
Représentation  
& Intégration  
des Connaissances





Home Analysis Protocols Archive Login

login

Quick search:  go

Include BBRIC Archive

Query BBRIC protocols Search results

title	description	url	keywords	categories	domains	authors	submitter
Small genome assembly	Small genome assembly from illumina paired-end reads	<a href="#">link</a>	assembly illumina genome	Sequence assembly	Bacteria	lipm.bioinfo@toulou	Ludovic,Legrand...

Le système propose la liste des programmes potentiellement pertinents avec un lien hypertexte qui renvoie vers le formulaire permettant de faire l'analyse

Galaxy / BBRIC

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

Get Data  
BBRIC protocols  
Blast

Workflows  
All workflows

Welcome to Galaxy

It appears that you found this tool from a link outside of Galaxy. If you're not familiar with Galaxy, please consider visiting the [welcome page](#). To learn more about what Galaxy is and what it can do for you, please visit the [Galaxy wiki](#).

Small genome assembly (version 1)

Paired-end libraries

Paired-end library 1

Read file 1:  
fastq,fastq.gz

Read file 2:  
fastq,fastq.gz

Add new Paired-end library

Scaffold prefix:  
Organism code without space like: ECOLI, ARATH...

Contig minimum length (nt):  
500



---

# Présentation de l'Archive Séquences BBRIC

Ludovic Legrand



## Objectifs

- Conserver sur le long terme les données « brutes » de séquence et les informations associées à la génération des données (métadonnées)
- Faciliter la soumission des séquences aux banques publiques lors des publications
- Fournir un minimum d'informations permettant l'analyse (automatique) des données

## Liste de vos données

ludovic.jegrand@toulouse.inra.fr  
llegrand@inra.fr

**BBRIC Archive Portal - Sequence Collection**

Recherche rapide

Sequences Account Admin Logout

Submit Search Browse Manage

Quick search  Envoyer (use \* for partial search, ex: 'phospho\*' will match with 'phosphorylase', 'phosphokinase')

Filter(s) Project Molecule

4 results

Contributor	Date	Title	Species	Molecule	Project	More Info
CATI BBRIC	20140314	Sb-2404-G1	Sinorhizobium bbric	genomic_DNA	BBRIC	+
CATI BBRIC	20140314	Sb-2404-33-R1	Sinorhizobium bbric	total_RNA	BBRIC	+
CATI BBRIC	20140314	Sb-2404-21-R1	Sinorhizobium bbric	total_RNA	BBRIC	+
CATI BBRIC	20140314	Sb-2404-36-R1	Sinorhizobium bbric	total_RNA	BBRIC	+

All files are public All files are shared or belong to you Some files are shared All files are private

Liste des données accessibles

Détails sur les données

## Détails

Visualisation des métadonnées

Téléchargement des métadonnées

Contributor	Date	Title	Species	Molecule	Project	Metadata
CATI BBRIC	20140314	Sb-2404-G1	Sinorhizobium bbric	genomic_DNA	BBRIC	
		S.bbric-250K-100x.2.fastq.gz	4.25 Mo	jerome.gouzy@toulouse.inra.fr	public	
		S.bbric-250K-100x.1.fastq.gz	4.24 Mo	jerome.gouzy@toulouse.inra.fr	public	

■ All files are public 
 ■ All files are shared or belong to you 
 ■ Some files are shared 
 ■ All files are private

Contributor	CATI BBRIC
Date	20140314
Date update	20140314
Institution	INRA
Project	BBRIC
Extract protocol	unknown
Molecule	genomic_DNA
Species	Sinorhizobium bbric
Strain	2404
Title	Sb-2404-G1
Read length	100 (x2)
Name	Sb-2404-G1
Format	fastq
Repeat	none
Instrument model	Illumina HiSeq 2000
Sequencing center	Genotoul PLAGE
Average insert	300
Standard deviation	10
Type	pe
Library construction protocol	art
Library strategy	OTHER
Submitter	jerome.gouzy@toulouse.inra.fr

Téléchargement des données

Résumé des métadonnées

Choix du type de librairie

Sequences

Submit

Select data type

Data Type: (unoriented) single end  
(unoriented) single end  
**(unoriented) paired end**  
oriented single end  
oriented paired end  
mate pair

continue

Personnes impliquées dans la production de séquence

Sequences Account Admin Logout

Submit Search Browse Manage

Specify Meta Data

File name	File size	Percent uploaded	Server Data
No files have been selected.			

Welcome ludovic.legrand@toulouse.inra.fr

Please select a group to share the data: ludovic.legrand@toulouse.inra.fr

Contributors

\*Contributor: CATI BBRIC  
Add other Contributor

Sample

Sequencing

Funding

Upload files and metadata

# Soumission (2/4): Description de l'échantillon biologique

Contributors

Sample

\*Title Sb-2404-G1

Summary

Source name

**Informations générales**

Organisms

Add other Organism

Organism

\*Species Sin

Strain Sinorhizobium meliloti

Genotype Sinorhizobium bbric

TaxID

Genus species format.

**Métadonnées concernant l'organisme**

Characteristics

Add other Characteristic

Molecule genomic DNA

Growth protocol

Treatment protocol

\*Extract protocol

Description

Type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.

**Protocoles appliqués à l'échantillon**

Sequencing





Informations  
provenant du  
centre de  
séquençage

Sample

Sequencing

Protocol

Library strategy: RNA-Seq

\*Library construction protocol: RNA-Seq, RNA-Seq (size fractionation), RNA-Seq (CAGE), RNA-Seq (RACE), CTS, ChIP-Seq, MNase-Seq, MBD-Seq, MRE-Seq, Bisulfite-Seq, Bisulfite-Seq (reduced representation), MeDIP-Seq, DNase-Hypersensitivity, **OTHER**

Library construction kit

Library sequencing kit

Data processing

Basecalling protocol: OTHER

Data processing steps

Genome build

Processed data files format and content

Library

Average insert

Standard deviation

A field that describes the sequencing technique for this library.

**Protocoles et informations concernant le séquençage**

Platform

\*Sequencing center: Genotoul PLAGE

\*Instrument model: Illumina HiSeq 2000

Collection

Repeat

Format

Add other Files

Files

\*Name

Orientation

File

\*File name: Select a file from your local directories

**Plateforme de séquençage**

Include one of the following models: Illumina Genome Analyzer, Illumina Genome Analyzer II, Illumina Genome Analyzer Ix, Illumina HiSeq 2000, Illumina HiSeq 1000, Illumina MiSeq, AB SOLiD System, AB SOLiD System 2.0, AB SOLiD System 3.0, AB SOLiD 4 Systemn AB SOLiD 4hq System, AB SOLiD PI System, AB SOLiD 5500xl SOLiD System, AB SOLiD 5500 SOLiD System, 454 GS, 454 GS 20, 454 GS FLX, 454 GS Junior, 454 GS FLX Titanium, Helicos HeliScope, PacBio RS, Complete Genomics, Ion Torrent PGM.



# Soumission (3b/4): Upload des données

Collection

Repeat

Format

Files

\*Name

Orientation

File

\*File name

S.bbric-250K-100x.1.fastq.gz : 4.45 Mo

\*Read length

File

\*File name

S.bbric-250K-100x.2.fastq.gz : 4.45 Mo

\*Read length

Nom du jeu de données

Taille des lectures

Sélection des fichiers

- directement à partir de son disque dur
- liste de fichiers après un transfert par ftp

	Sequences	Account	Admin	Logout
	Submit	Search	Browse	Manage

## Specify Meta Data

**File name File size Percent uploaded Server Data**

No files have been selected.

Welcome ludovic.legrand@toulouse.inra.fr

Please select a group to share the data

Contributors	*
Sample	*
Sequencing	*
Funding	*

Add other Source

Source

\*Project

Institution

Upload files and metadata

Informations sur le financement



- Validation des métadonnées et des données
  - contrôle de la taille des lectures
  - contrôle d'intégrité sur les paires pour les données paired-end et mate-pairs
    - ➔ Envoi d'un email (échec ou succès)
- Possibilité d'associer les données à des groupes d'utilisateurs




---

# Présentation de l'environnement web « Galaxy »

Sébastien Carrere



- 
- ❖ Portail web
  - ❖ Accès simplifié à de nombreux outils bioinformatiques
  - ❖ Moins puissant que la ligne de commande, mais suffisant pour beaucoup d'analyses
  - ❖ Projet international très actif, grande communauté
  - ❖ Aide : tutoriaux, vidéos d'exemple
    - <http://wiki.galaxyproject.org/Learn>



❖ Beaucoup de fonctionnalités

- Nombreuses formations dédiées à Galaxy

❖ Aujourd'hui

- Utilisation de workflows conçus par BBRIC
- Lien avec l'architecture BBRIC
  - Traitement de données issues de l'Archive


Liste d'outils

The screenshot shows the Galaxy web interface in Google Chrome. The browser address bar shows `bipaa-galaxy.genouest.org/root`. The interface includes a top navigation bar with 'Galaxy' and various menu items like 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. On the left, there is a 'Tools' sidebar with a search bar and a list of tool categories such as 'Get Data', 'Send Data', 'LEPIDODB', 'LepidoDB', 'APHIDBASE', 'Sequences', 'SYMBIOSE', 'Metagenomics benchmark', 'Text Manipulation', 'Fastq to OTUs', '454 data Manipulation', 'Motif', 'Primer design and test', 'Alignment', 'NGS: Assembly', 'MetaGenomics', 'Genome manipulation', 'Data management', 'NGS: SNPs analysis', 'BIOGENOUEST COMMUNITY', 'QTL Detection', 'Genetics Graphics', 'GENOMICS', 'Alignment', 'GENETICS', 'Primer design', 'contigs assembly', 'population genetics', 'PHYLOGENY', 'Text Manipulation', 'Next Generation Quality', 'Phylogenetic Tree', 'Alignment', 'NCBI Blast', and 'Blat'. The main workspace contains a green notification box that says 'Hello world! It's running...' and a workflow diagram titled 'WWFSMD?' with the subtitle 'grow noodly appendages...'. The workflow diagram shows several tool steps connected by arrows, including 'Filter', 'Join', 'Group', 'Sort', 'Join two Queries', and 'Select first'. Below the workflow is the 'usegalaxy.org' logo and a paragraph of text describing Galaxy as an open, web-based platform for data intensive biomedical research. On the right, there is a 'History' sidebar showing a list of recent jobs, including '92: megablast on data 91 and data 89', '91: Pasted Entry', '90: megablast on data 32 and data 89', '89: nucleotide BLAST database from data 82', '88: megablast on db', '87: Blast2GFF on data 86', '86: megablast on db', '84: megablast on file', '83: megablast on file', '82: Spodoptera frugiperda v3.0.fa', '80: Blast2GFF on data 79', '79: megablast on file', '78: megablast on db', '77: discoSnp on data 74: extra-files.html', '76: discoSnp on data 74: coherence.fa', '75: discoSnp on data 74: out.txt', '74: sample.fastq', '73: Scipio on data 31 and data 71', and '71: potato\_virus\_genome.fasta'. Each job entry includes an eye icon, a refresh icon, and a delete icon.

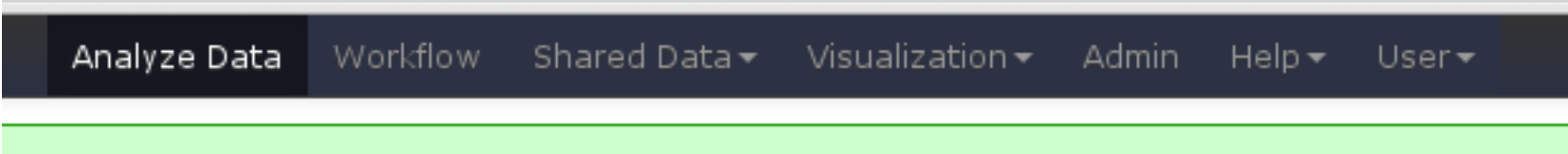
Historique qui contient des Datasets

Lancement des outils  
Visualisation des résultats  
...





---



Analyze Data

Workflow

Shared Data ▾

Visualization ▾

Admin

Help ▾

User ▾

## ❖ Analyze Data

- Lancement d'outil, visualisation de résultats

## ❖ Workflow

- Conception et **utilisation** de workflows

## ❖ Shared Data

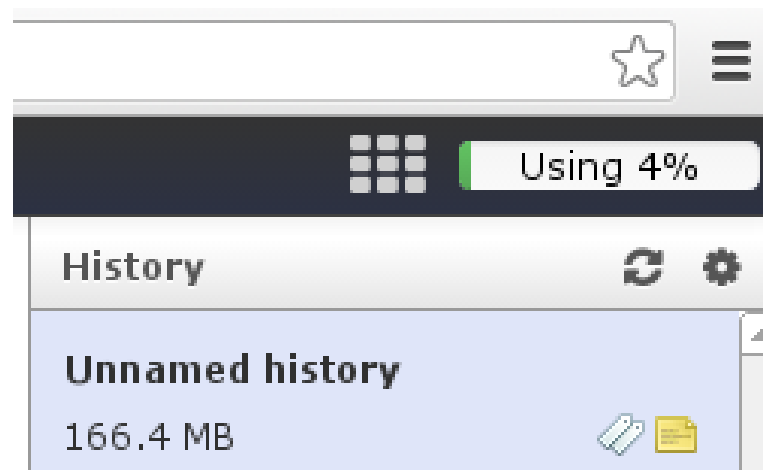
- Accès à des données mises à disposition par les administrateurs

## ❖ Sources de données multiples

- Upload depuis le portail web
- **Shared Data**
- **Archive BBRIC**
- FTP

## ❖ Volume total limité par utilisateur

- Quota défini par l'administrateur



Chaque dataset :

- Résultat d'un outil
- Donnée récupérée (upload, archive)

Les datasets s'empilent  
au fur et à mesure

Gris = outil pas encore lancé

Jaune = l'outil s'exécute

Vert = terminé avec succès

Rouge = terminé avec une erreur

Dataset Name	Status	View	Edit	Delete
Unnamed history	Grey			
92: megablast on data 91 and data 89	Green	Eye	Pencil	X
91: Pasted Entry	Green	Eye	Pencil	X
90: megablast on data 32 and data 89	Green	Eye	Pencil	X
89: nucleotide BLAST database from data 82	Green	Eye	Pencil	X
88: megablast on db	Green	Eye	Pencil	X
87: Blast2GFF on data 86	Green	Eye	Pencil	X
86: megablast on db	Green	Eye	Pencil	X

Aperçu du contenu

Modification

-nom

-type

Suppression

Détails d'un dataset : clic sur le titre

Télécharger le fichier

Infos sur l'origine

Relancer l'outil  
avec les mêmes paramètres

History

Unnamed history  
0 bytes

**4: sample.read2.fq** 20.8 MB  
format: fastq, database: ?  
uploaded fastq file

```
@FCB067AABXX:8:1101:1676:2039/2  
CGGGGGAAGGAACCTATCCCCAGAACCCACAGACCCTGT  
+  
HHHHHHHHHHH;HGHHHFHHHGHGHHGHGHHHHHHHHH  
@FCB067AABXX:8:1101:2106:2035/2  
GCTTGTATGGACTGTTAAAAACACAAGGGCCAGCTAT
```

**3: sample.read1.fq**

**2: sample.read2.fq**

**1: sample2.read1.fq**

Statistiques/aperçu

Nom de l'historique

History

Using 4%

Unnamed history  
166.4 MB

92: megablast on data 91 and data 89

91: Pasted Entry

90: megablast on data 32 and data 89

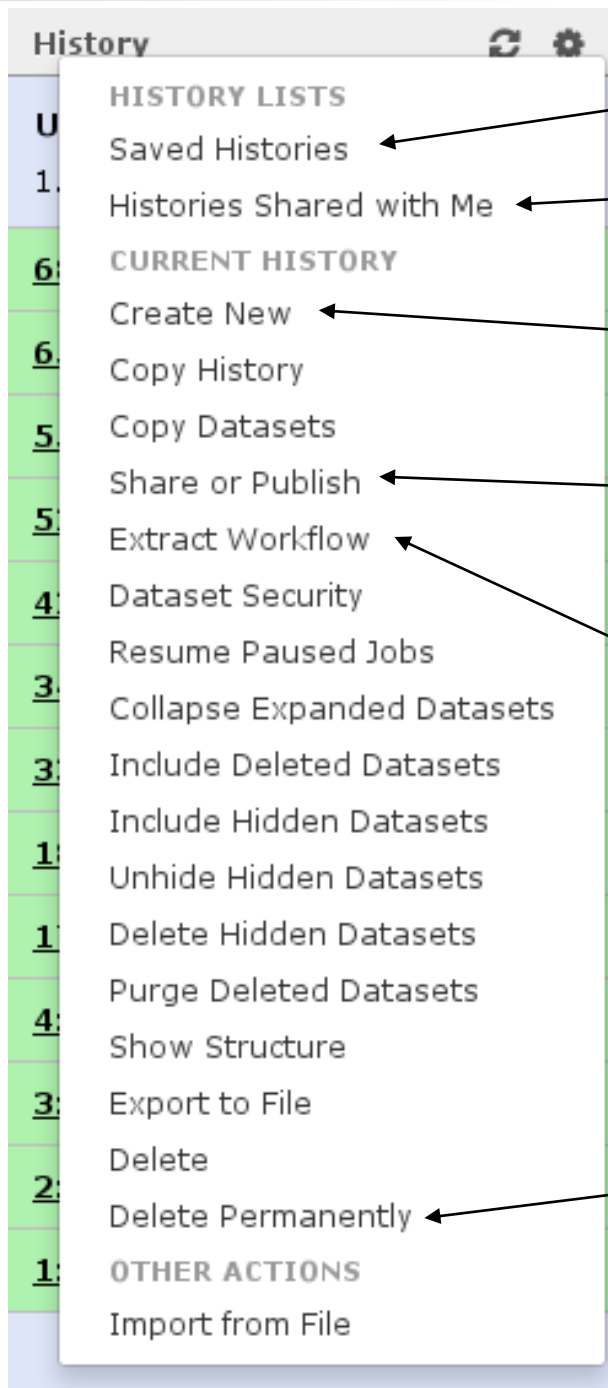
89: nucleotide BLAST database from data 82

88: megablast on db

87: Blast2GFF on data 86

86: megablast on db

Menu de l'historique



Liste de nos historiques

Historiques partagés par d'autres

Nouvel historique

Partage de l'historique actuel

Création de workflow

Suppression de l'historique actuel

# Saved Histories

*search history names and tags*

[Advanced Search](#)

<input type="checkbox"/>	<u>Name</u>	<u>Datasets</u>	<u>Tags</u>	<u>Sharing</u>	<u>Size on Disk</u>	<u>Created</u>	<u>Last Updated</u> ↑	<u>Status</u>
<input type="checkbox"/>	Unnamed history	13	<a href="#">0 Tags</a>		1.0 GB	Aug 08, 2013	6 days ago	<b>current history</b>
<input type="checkbox"/>	Unnamed history	5	<a href="#">0 Tags</a>		35.1 MB	Apr 10, 2014	Apr 11, 2014	
<input type="checkbox"/>	Unn histo		<a href="#">0 Tags</a>		0 bytes	Feb 14, 2014	Feb 14, 2014	

- Switch
- View
- Share or Publish
- Rename
- Delete
- Delete Permanently

For 0
me
Delete
Delete Permanently
Undelete

Histories that have not been updated for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.



# Lancement d'un outil

## 1. Choix d'un outil

**Tools**

blast

**Metagenomics benchmark**

[NCBI BLAST+ blastn](#) Search nucleotide database with nucleotide query sequence(s)

[blastFilter](#) Filter by E-value

**Genome manipulation**

[Filter fasta sequences](#) from a genome using tab file like a BLAST result list

**NCBIBlast**

[NCBI BLAST+ blastn](#) Search nucleotide database with nucleotide query sequence(s)

[NCBI BLAST+ tblastn](#) Search translated nucleotide database with protein query sequence(s)

[NCBI BLAST+ makeblastdb](#) Make BLAST database

[BLAST XML to tabular](#) Convert BLAST XML output to tabular

[NCBI BLAST+ blastp](#) Search protein database with protein query sequence(s)

NCBI BLAST+ blastn version 0.0.20

**Nucleotide query sequence(s):**  
42: Assembled transcripts

**Subject database/sequences:**  
Locally installed BLAST database

**Nucleotide BLAST database:**  
Apisum genome v2

**Type of BLAST:**

- megablast
- blastn
- blastn-short
- dc-megablast

**Set expectation value cutoff:**  
0.001

**Output format:**  
Tabular (extended 24 columns)

**Advanced Options:**  
Hide Advanced Options

Execute

## 2. Réglages des options

## 3. Lancement du calcul (crée un nouveau dataset)





---

❖ Comment créer les premiers datasets ?

- Depuis l'archive BBRIC
- Depuis les Shared data
- Upload d'un fichier
- ...

# Depuis l'archive : outil « BBRIC Archive »

The screenshot shows the Galaxy web interface by GenOuest. On the left, a 'Tools' panel is open, displaying a search bar and a list of tools. The tool 'BBRIC Archive server' is highlighted with a red rectangular box. Other visible tools include 'Fetch NCBI Entrez database with an Entrez query', 'Search SfruDB sequences', 'RIDA BBRIC Archive server', 'Decompress an archive in zip, gz, tar.gz, fastq.gz, fastq.bz2 or tar.bz2 format', 'LOCAL UPLOAD', and 'Upload File from your computer'. The main workspace on the right contains several icons: a green folder, a DNA double helix, a power button, and a logo for 'Bioinformatique Biodiversité Représentation & Intégration des Connaissances'.

# Depuis l'archive : choix d'un jeu de données

Galaxy by GenOuest

Analyze Data Workflow Shared Data Visualization Admin Help User Using 24%

Tools

search tools

Get Data

Fetch NCBI Entrez Fetch NCBI database with an Entrez query

Search SfruDB sequences

BIPAA BBRIC Archive server

BBRIC Archive server

Decompress an archive in zip, gz, tar.gz, fastq.gz, fastq.bz2 or tar.bz2 format

LOCAL UPLOAD

Upload File from your computer

GENOCLUSTER UPLOAD

GenoLink imports file from your Genoduster HOME to the current history WITHOUT copying

anthony.breiladeau@rennes.inra.fr  
abreiladeau@inra.fr

Bioinformatique Biodiversité Représentation & Intégration des Connaissances

BBRIC Archive Portal - Sequence Collection

Sequences Account Logout

Submit Search Browse Manage

Quick search  Submit Query (use \* for partial search, ex: 'phospho\*' will match with for 'phosphorylase', 'phosphokinase')

Filter(s) Project Molecule

4 results

Contributor	Date	Title	Species	Molecule	Project	More Info
CATI BBRIC	20140314	Sb-2404-G1	Sinorhizobium bbric	genomic_DNA	BBRIC	+
CATI BBRIC	20140314	Sb-2404-33-R1	Sinorhizobium bbric	total_RNA	BBRIC	+
CATI BBRIC	20140314	Sb-2404-21-R1	Sinorhizobium bbric	total_RNA	BBRIC	+
CATI BBRIC	20140314	Sb-2404-36-R1	Sinorhizobium bbric	total_RNA	BBRIC	+

All files are public All files are shared or belong to you Some files are shared All files are private

History

Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

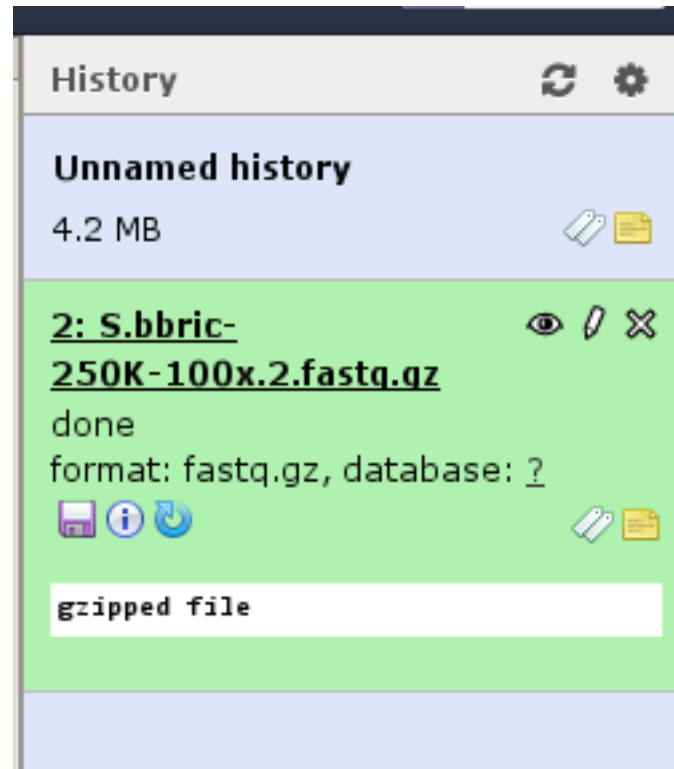
# Depuis l'archive : import dans galaxy

The screenshot shows the BBRIC Archive Portal interface. At the top, there is a navigation bar with options: Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User. The page title is "BBRIC Archive Portal - Sequence Collection". On the left, there is a sidebar with a vertical list of email addresses: anthony.bretaudeau@rennes.inra.fr and abretaudeau@inra.fr. The main content area features a table with columns: Contributor, Date, Title, Species, Molecule, Project, and Metadata. The table contains three rows of data. The first row is highlighted in blue. Below the table, there are four colored boxes representing file visibility: All files are public (blue), All files are shared or belong to you (green), Some files are shared (yellow), and All files are private (red). On the right side, there is a "History" panel showing "Unnamed history" with 0 bytes and a message: "Your history is empty. Click 'Get Data' on the left pane to start".

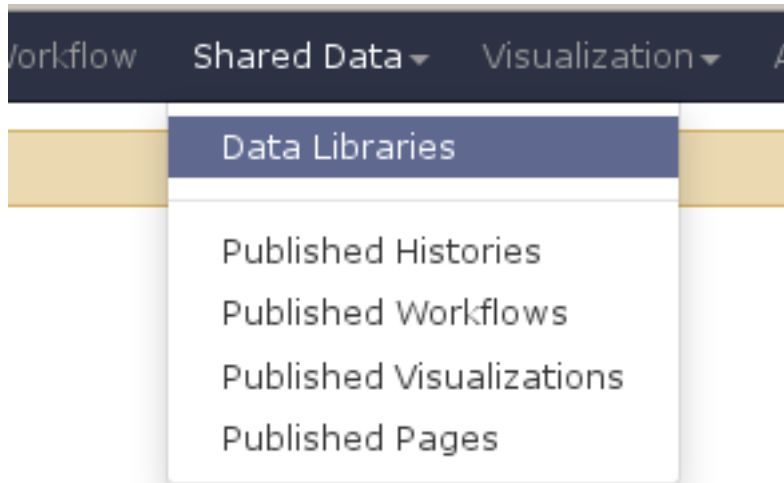
Contributor	Date	Title	Species	Molecule	Project	Metadata
CATI BBRIC	20140314	Sb-2404-G1	Sinorhizobium bbric	genomic_DNA	BBRIC	
S.bbric-250K-100x.2.fastq.gz		4.25 Mo	public, jerome.gouzy@toulouse.inra.fr			<a href="#">Send to galaxy</a>
S.bbric-250K-100x.1.fastq.gz		4.24 Mo	jerome.gouzy@toulouse.inra.fr, public			<a href="#">Send to galaxy</a>

Legend: ■ All files are public ■ All files are shared or belong to you ■ Some files are shared ■ All files are private

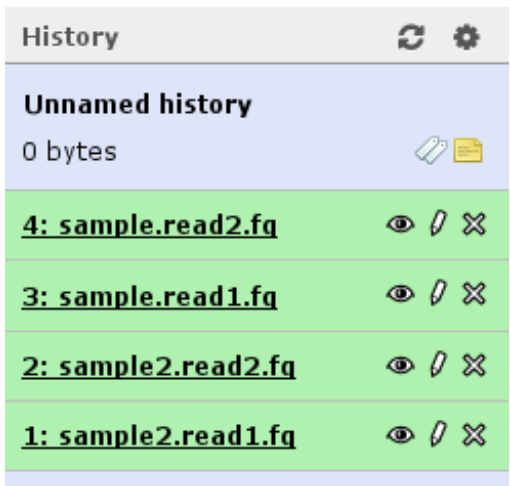
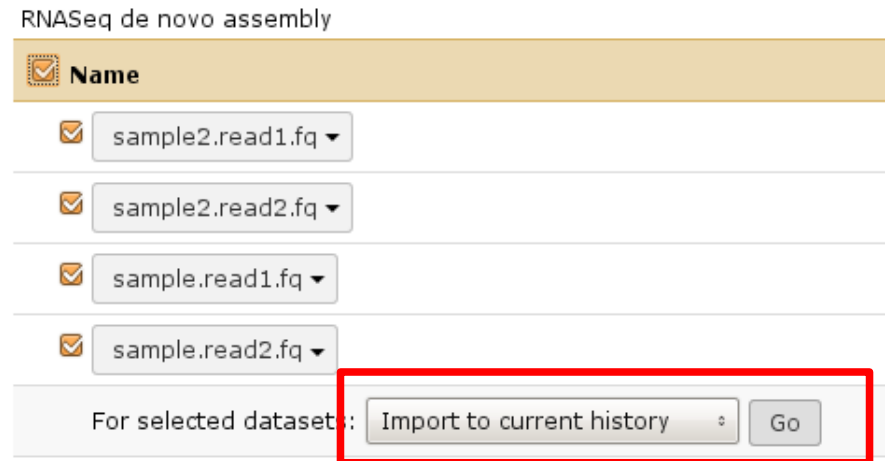
# Depuis l'archive BBRIC : nouveau dataset



# Depuis les « data libraries »



## Data Library "BBRIC"



# Upload de fichier

**Tools**

search tools

**Get Data**

- [Fetch NCBI Entrez](#) Fetch NCBI database with an Entrez query
- [Search SfrudB](#) sequences
- [BIPAA BBRIC Archive](#) server
- [BBRIC Archive](#) server
- [Decompress an archive](#) in zip, gz, tar.gz, fastq.gz, fastq.bz2 or tar.bz2 format

**LOCAL UPLOAD**

- Upload File from your computer**

**GENOCLUSTER UPLOAD**

- [GenoLink](#) imports file from your Genocluster HOME to the current history WITHOUT copying
- [GenoCopy](#) import a file or directory from your Genocluster HOME to the current history

**EXTERNAL UPLOAD**

- [UCSC Main](#) table browser
- [EBI SRA](#) ENA SRA
- [BioMart](#) Central server

**Upload File (version 1.1.3)**

**File Format:**

Auto-detect

Which format? See help below

**File:**

Browse... No file selected.

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail (ask your system administrator).

**URL/Text:**

Here you may specify a list of URLs (one per line) or paste the contents of a file

**Files uploaded via FTP:**

File	Size
Your FTP upload directory contains no files.	

This Galaxy server allows you to upload files via FTP. To upload some files, log in with your credentials (email address and password).

**Convert spaces to tabs:**

Yes

Use this option if you are entering intervals by hand.

**Genome:**

unspecified (?)

Execute



---

# ASSEMBLAGE *DE NOVO* DE TRANSCRIPTOMES AVEC MESURE DE L'EXPRESSION PAR BANQUE

(DONNÉES: LECTURES ILLUMINA RNASEQ)

Responsable et intervenant principal: Sébastien Carrere

Expert: Anthony Bretaudeau

Relecteur: Erika Sallet





- Introduction générale sur le RNASeq
  - Vue d'ensemble des grandes étapes dont l'analyse par assemblage de novo de transcriptome et la mesure de l'expression
- L'assemblage *de novo* de transcriptome
  - Traitement initial des reads
    - Élimination des artefacts
  - L'assemblage *de novo*
    - Filtrage/Normalisation des reads par couverture des k-mers
    - Cas de Trinity
    - Qualité de l'assemblage
  - Mesure de l'expression par banque
    - Mapping
    - Filtrage et comptage

## Définition générale

La technique d'analyse RNASeq permet de faire une étude qualitative et quantitative des différents transcrits d'un ou plusieurs échantillons. Elle comprend un séquençage sur toute ou une grande partie de la longueur de chaque transcrit.

## Applications

### ① Annotation

Identifier des transcrits/gènes, des exons, des jonctions intron/exon, des TSS, des sites polyA, ncRNAs, trans-splicing, etc.

### ② Quantification

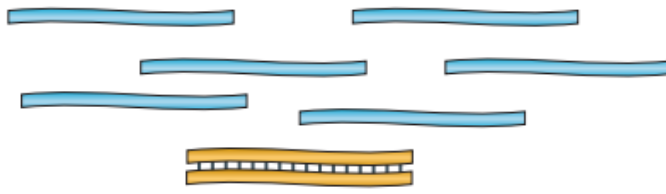
Mesurer des différences d'expression, de splicing alternatif, des TSS alternatifs, des sites polyA alternatifs entre un ou plusieurs groupes/traitements.

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
<i>Technology specifications</i>			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
<i>Application</i>			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
<i>Practical issues</i>			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

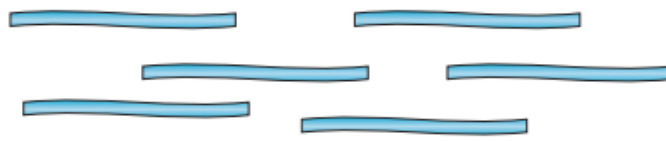
Wang Z et al. RNA-Seq: a revolutionary tool for transcriptomics, Nature Reviews Genetics 10, 57-63 (January 2009)

Le RNASeq permet d'identifier des nouveaux transcrits, en étudiant l'expression des gènes.

① mRNA or total RNA

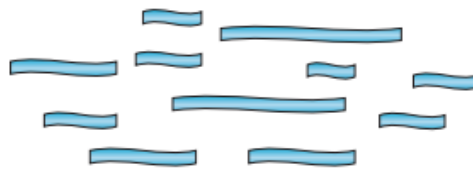


② Remove contaminant DNA

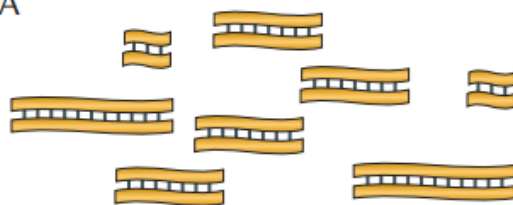


Remove rRNA?  
Select mRNA?

③ Fragment RNA



④ Reverse transcribe into cDNA



## Types de librairies:

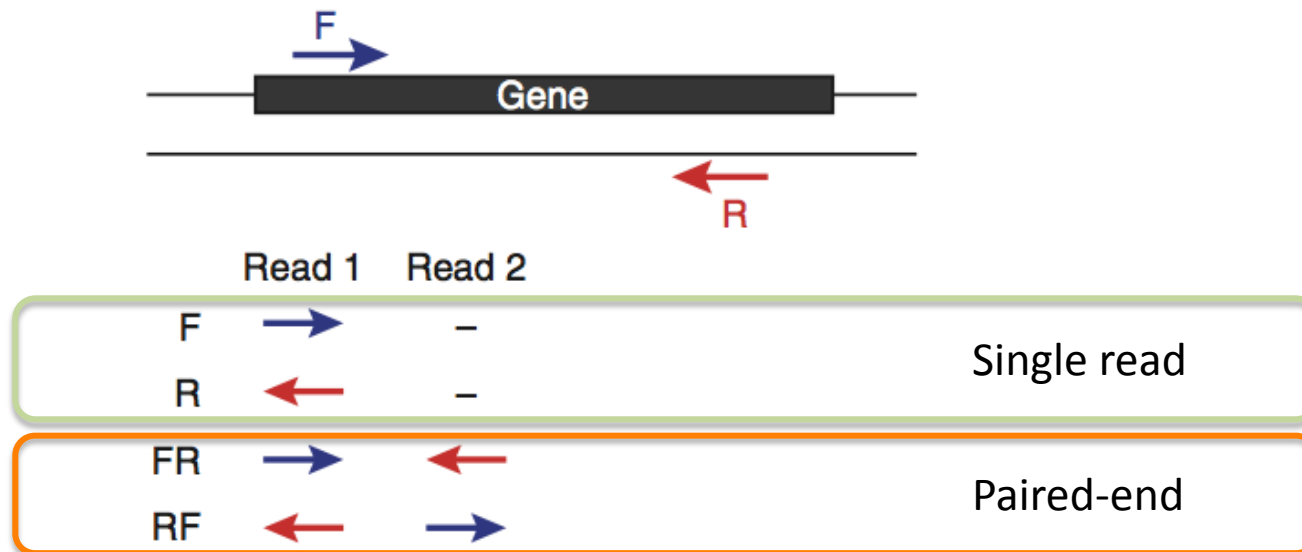
- Single read
- Paired-end

## Options du Paired-end

- ① Double brin
- ② Brin spécifique

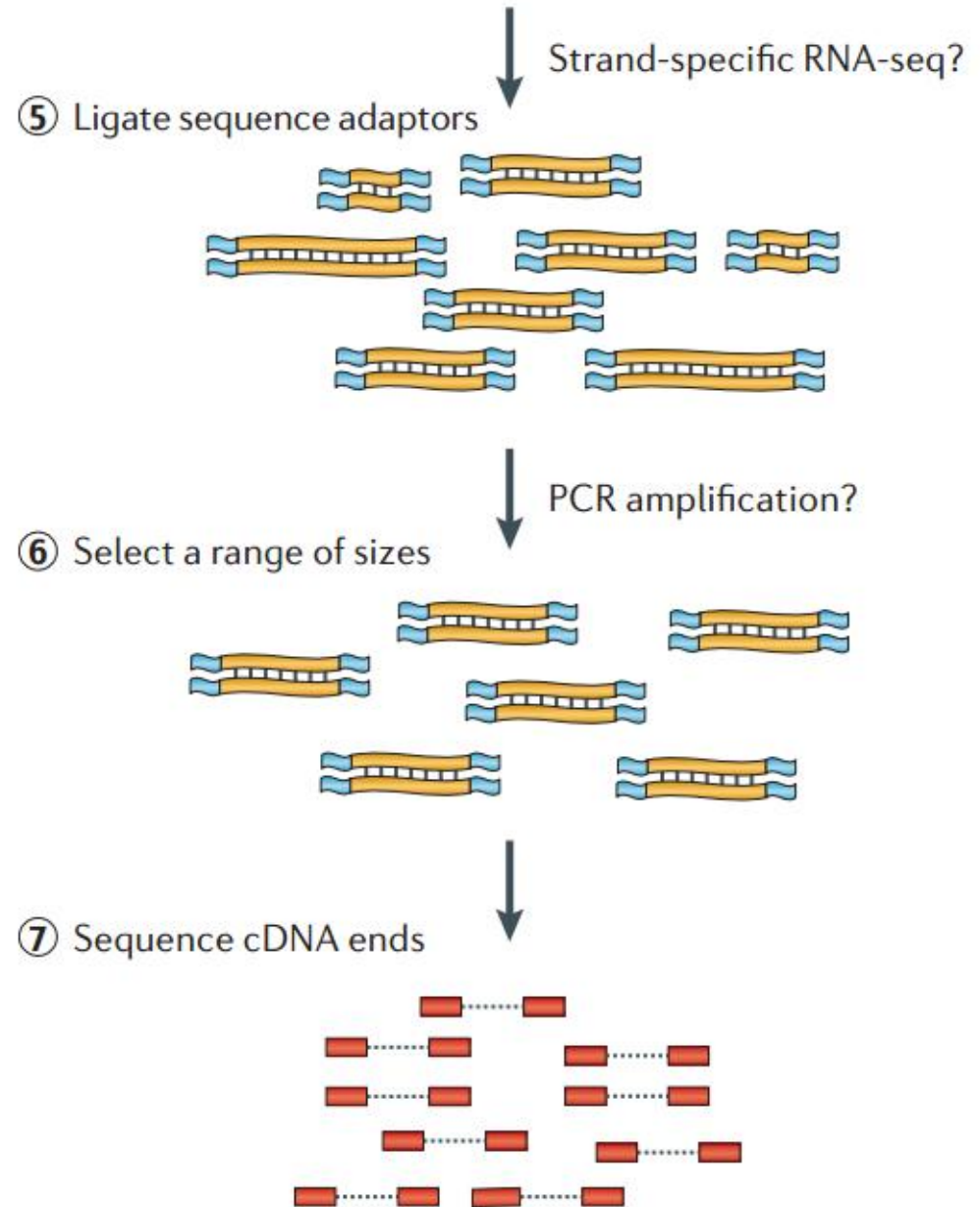
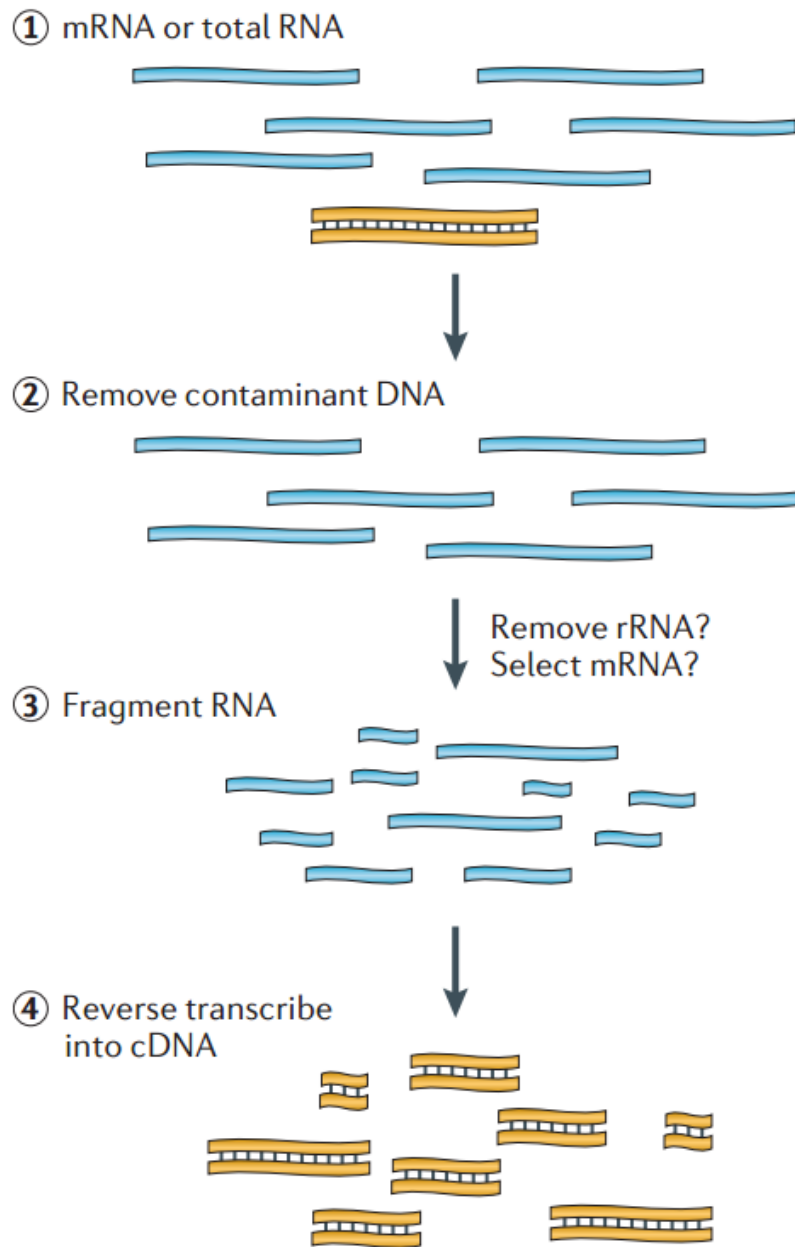
Martin & Wang (2011) Nat. Rev. Gen. 12,671

**Librairie brin spécifique:** les reads obtenues par séquençage conservent l'information du brin transcrit de l'ARNm dont elles sont issues.  
=> facilite la découverte de transcrits (sens et anti-sens), l'annotation des génomes, la quantification de l'expression.



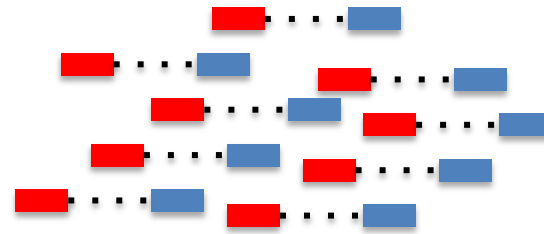
**Figure 4 |** Strand-specific library types. The left (/1) and right (/2) sequencing reads are depicted according to their orientations relative to the sense strand of a transcript sequence. The strand-specific library type (F, R, FR or RF) depends on the library construction protocol and is user-specified to Trinity via the '*--SS\_lib\_type*' parameter.

# Le RNASeq: protocole expérimental



Martin & Wang (2011) Nat. Rev. Gen. 12,671

Paired-end Raw reads



read.left.fq (/1)

read.right.fq (/2)

Format FASTQ

Nom

@61G9EAAXX100520:5:100:10000:5699/1

Séquence

ATTGAGGAATAGTAATAAACGGAGGACTATTTAACCTGTTTCCTTTCTTTACGTTTTT  
AAATCCTTT

+

Score

DDDDABBD?DDDCDDDDDD?DD5DDD:CB=DACBCCDDBB:BCCCBBA=?AABBABBB

Qualités

BBBB@B=BAAA

@61G9EAAXX100520:5:100:10000:5699/2

AGTCGCTGTGCCTTACATACAGCTGCTAAGGATCCTTTTCGATCTAAAATTGCTCCG  
GTGTAACAG

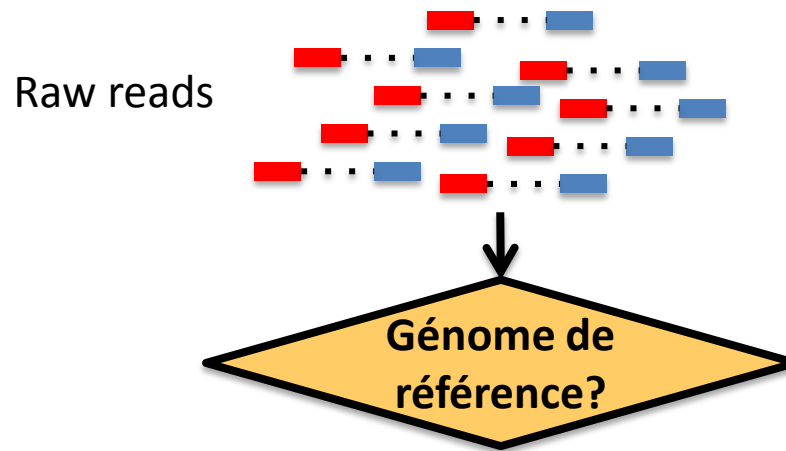
+

EDGFGGFGGGGBGGGGDFEGEFGGGGGGDGGDDFBDG?DFEFGEEFEFD?EFBDDGD  
GGD=AEEBEEDBD

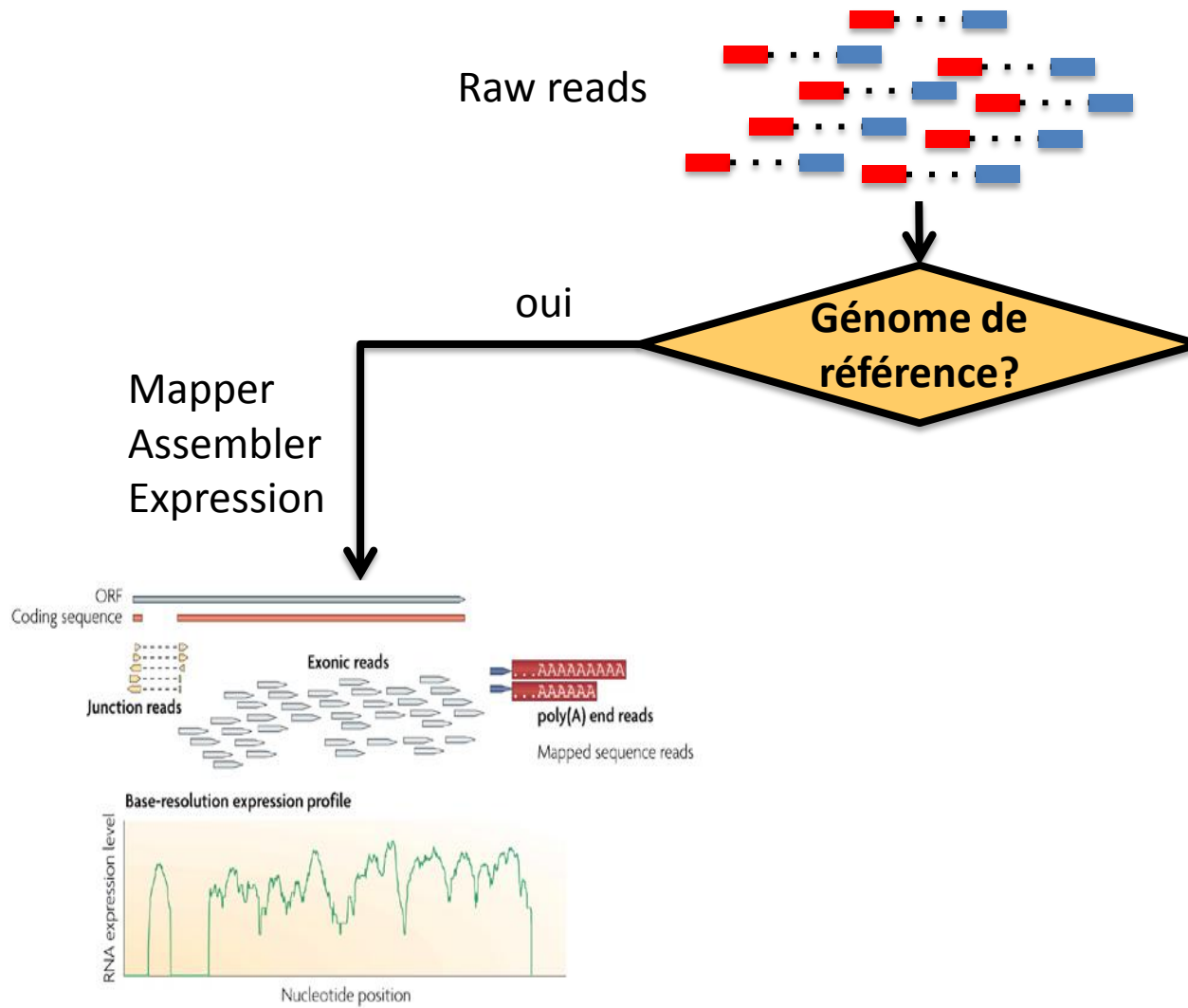






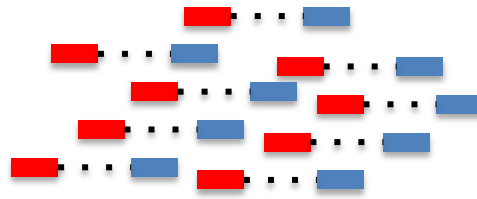


Wang Z et al. (January 2009) Nature Reviews Genetics 10, 57-63  
Martin & Wang (2011) Nat. Rev. Gen. 12,671



Wang Z et al. (January 2009) Nature Reviews Genetics 10, 57-63  
Martin & Wang (2011) Nat. Rev. Gen. 12,671

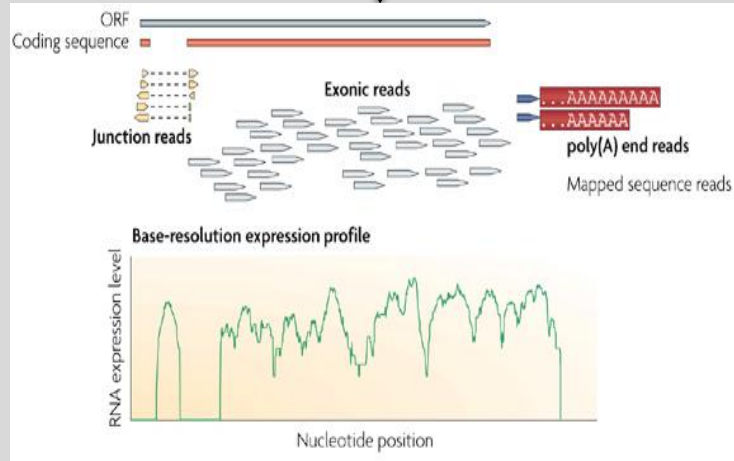
Raw reads



Génom de référence?

oui

Mapper  
Assembler  
Expression



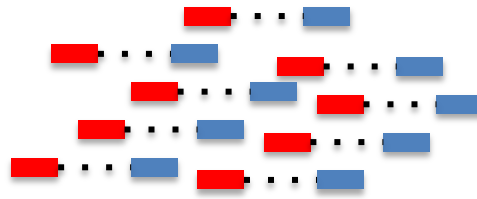
⇒ Annotation de génomes bactériens

⇒ Mesure de l'expression

Wang Z et al. (January 2009) Nature Reviews Genetics 10, 57-63

Martin & Wang (2011) Nat. Rev. Gen. 12,671

Raw reads



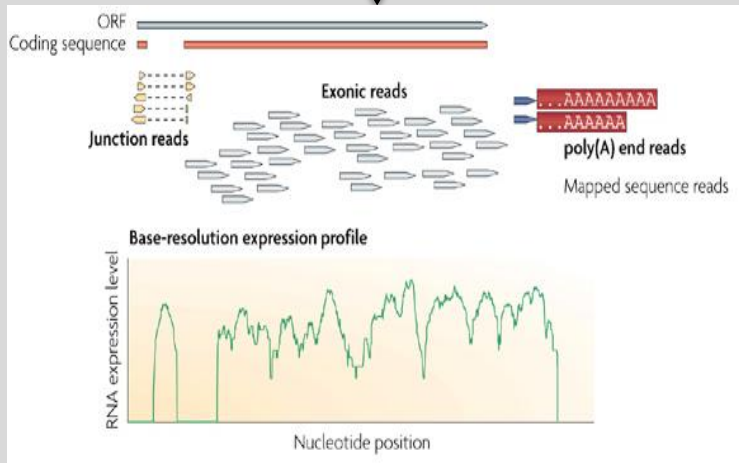
oui

Génome de référence?

non

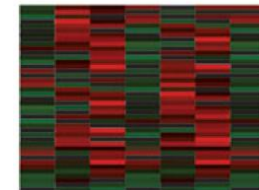
Mapper  
Assembler  
Expression

Assembler  
Mapper  
Expression



De novo assembly

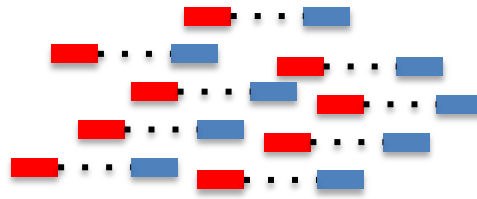
Quantification de l'expression



- ⇒ Annotation de génomes bactériens
- ⇒ Mesure de l'expression

Wang Z et al. (January 2009) Nature Reviews Genetics 10, 57-63  
Martin & Wang (2011) Nat. Rev. Gen. 12,671

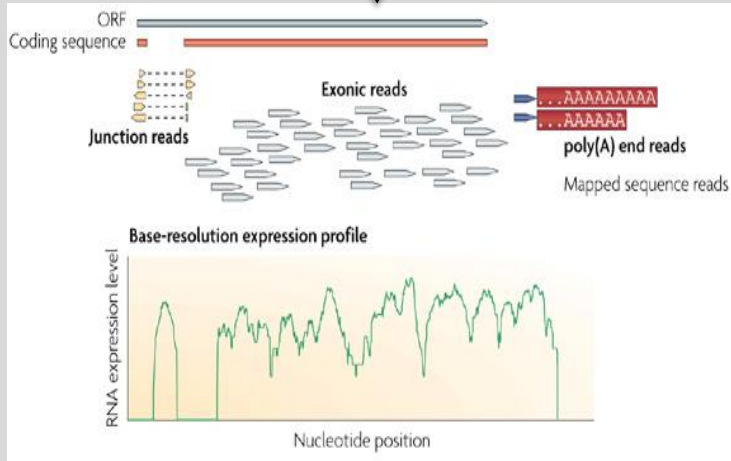
Raw reads



Génome de référence?

oui

Mapper  
Assembler  
Expression



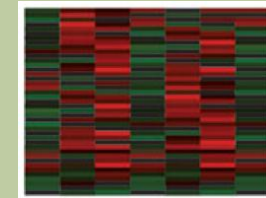
⇒ Module Annotation de génomes bactériens  
⇒ Mesure de l'expression

non

Assembler  
Mapper  
Expression

De novo assembly

Quantification de l'expression

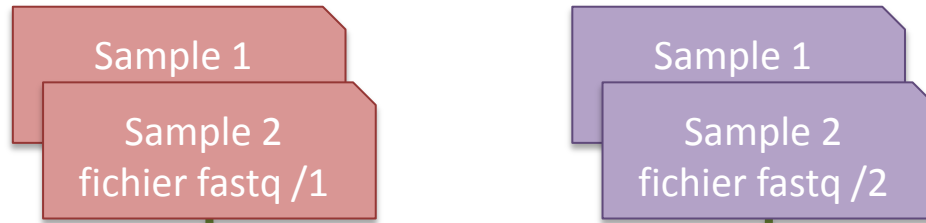


⇒ Ce module

Wang Z et al. (January 2009) Nature Reviews Genetics 10, 57-63  
Martin & Wang (2011) Nat. Rev. Gen. 12,671

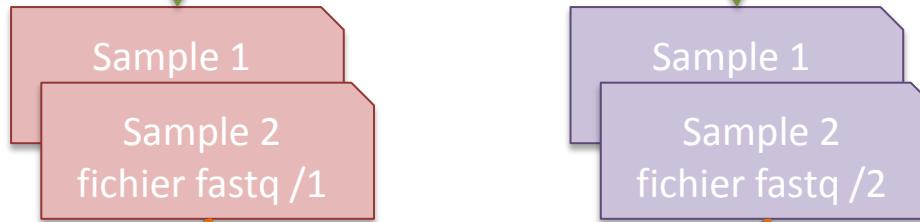
# Le pipeline: fonctionnement général

Librairies  
Paired-end  
brin spécifique



Traitement initial

Fichiers  
nettoyés



Assemblage  
de novo

Mesure de  
l'expression par  
banque

Transcriptome de  
référence

Comptage par  
transcrit

Comptage par  
gène

**Périmètre du pipeline  
d'analyse RNASeq**

- ① Pas de génome de référence
- ② Organisme eucaryote
- ③ Données Illumina

# Le RNASeq: étapes et choix pour le *de novo*

① Assembler de novo un transcriptome de référence

Choix technique

② Mesurer l'abondance des transcrits par banque

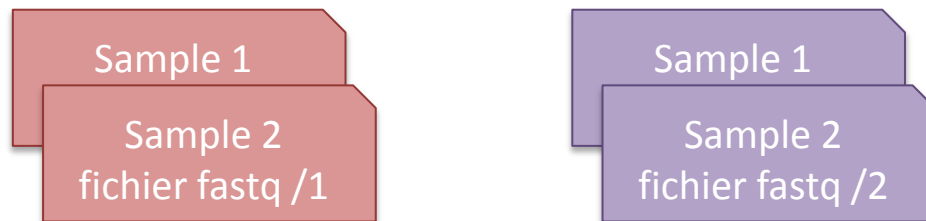
	Design expérimental	Réplicats	Design expérimental	
++				+++
Fonction du type d'ARN	RNA Prep	Nb méthodes d'enrichissements	RNA Prep	Eviter les normalisations DSN
+++	Librairies Prep	Librairie brin spécifique	Librairies Prep	++
+++		Reads Paired-end		++
++	Séquençage	Profondeur du séquençage	Séquençage	+++
+++		Multiplexage		+++
+++		Longueur des reads		++
+++	Analyse	Normalisation kmer	Analyse	---
+++				+++



- Introduction générale sur le RNASeq
  - Vue d'ensemble des grandes étapes dont l'analyse par assemblage de novo de transcriptome et la mesure de l'expression
- L'assemblage *de novo* de transcriptome
  - Traitement initial des reads
    - Élimination des artefacts
  - L'assemblage *de novo*
    - Filtrage/Normalisation des reads par couverture des k-mers
    - Cas de Trinity
    - Qualité de l'assemblage
  - Mesure de l'expression par banque
    - Mapping
    - Filtrage et comptage

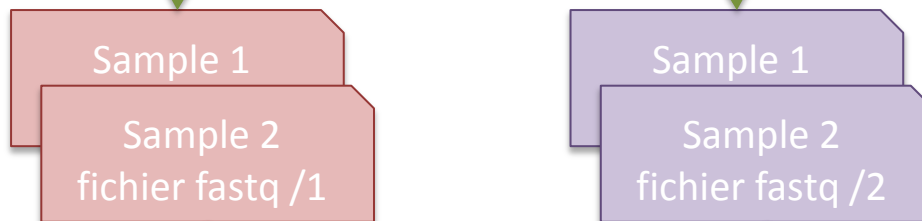


Librairies  
Paired-end  
brin spécifique



Traitement initial

Fichiers  
nettoyés



## Périmètre du pipeline d'analyse RNASeq

- ① Pas de génome de référence
- ② Organisme eucaryote
- ③ Données Illumina

Assemblage  
de novo

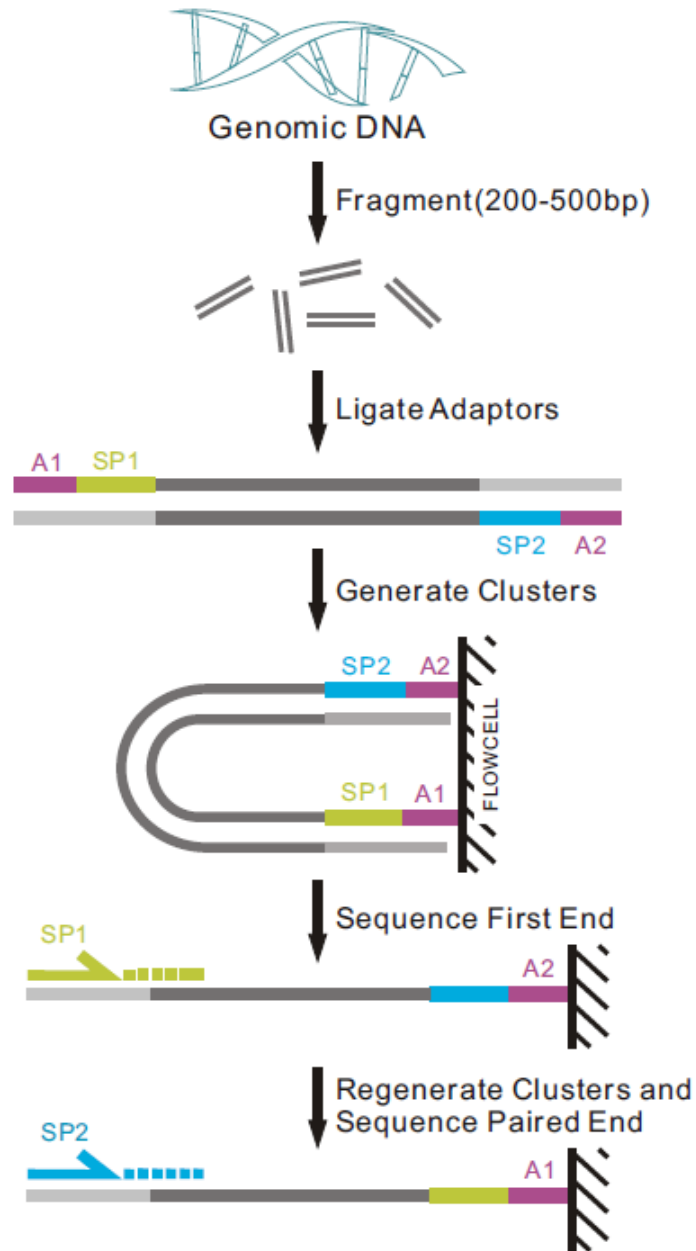
Mesure de  
l'expression par  
banque

Transcriptome de  
référence

Comptage par  
transcrit

Comptage par  
gène

# Élimination des artefacts: retrait des adaptateurs



**Exemple: adaptateurs/primers (brin spécifique)**

**P5: A1 / SP1**

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA  
CACGACGCTCTTCCGATCT

**Index barcode primer:**

5' CAAGCAGAAGACGGCATAACGAGAT

**P7: A2 / barcode (8nt) / SP2**

5' GATCGGAAGAGCACACGTCT [barcode]  
ATCTCGTATGCCGTCTTCTGCTTG

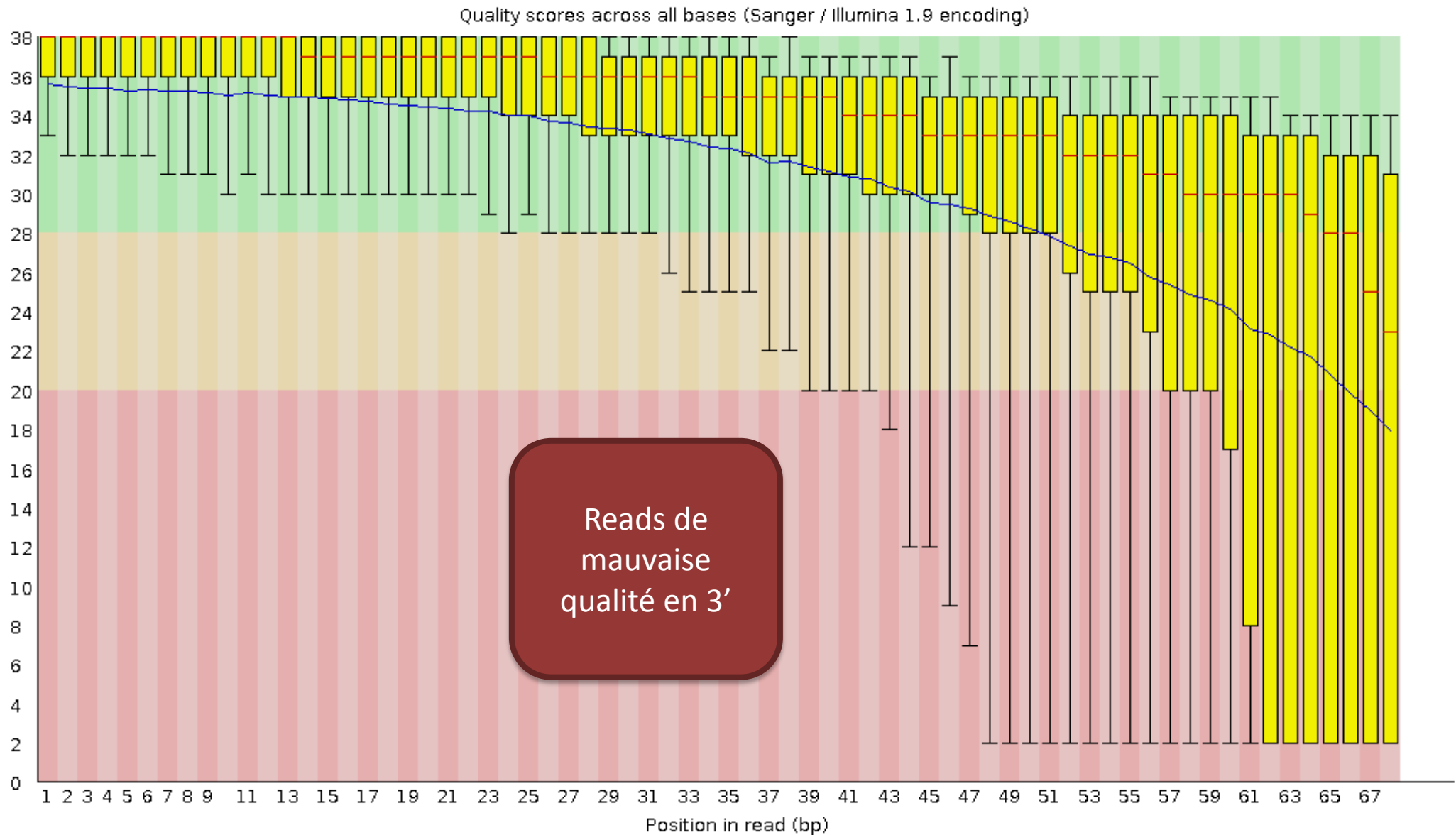
**Si adaptateurs non éliminés  
des reads**

**=> risque d'assembler des  
chimères car les reads  
provenant de différents  
transcrits peuvent se  
chevaucher sur ces  
séquences.**

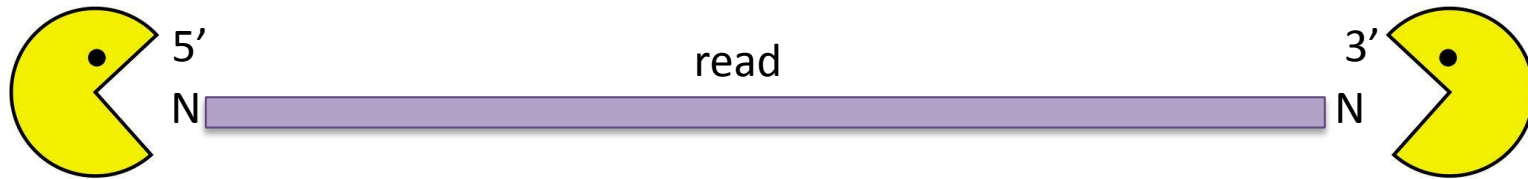
Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)



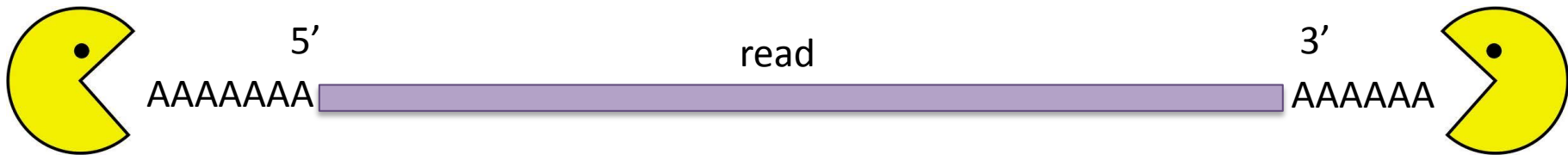
## Raw reads



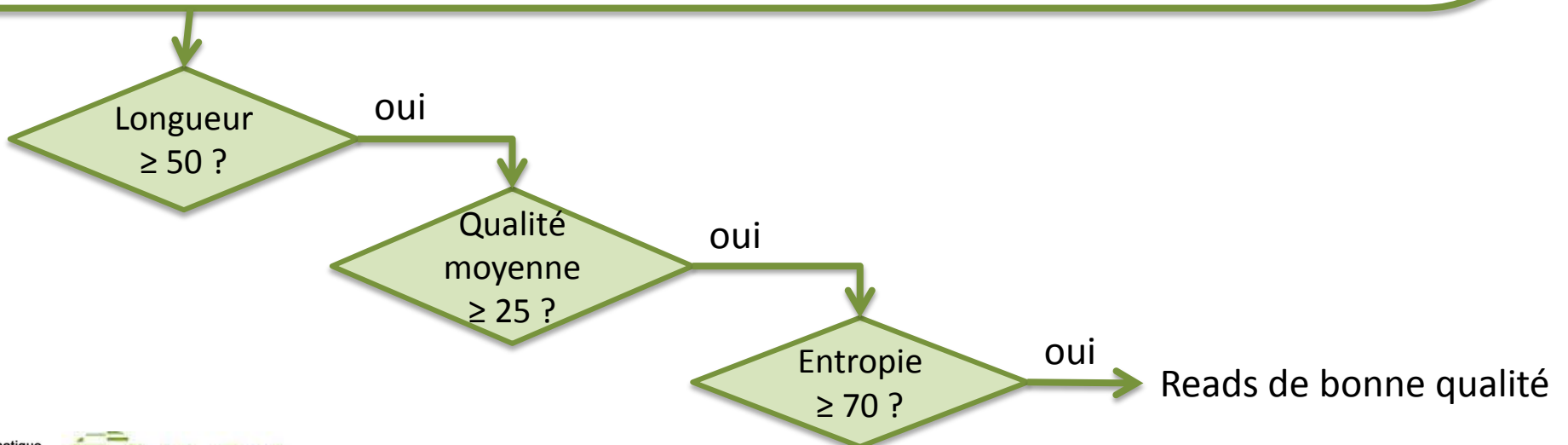
## Élimination des Ns en 5' et en 3' (si présents)



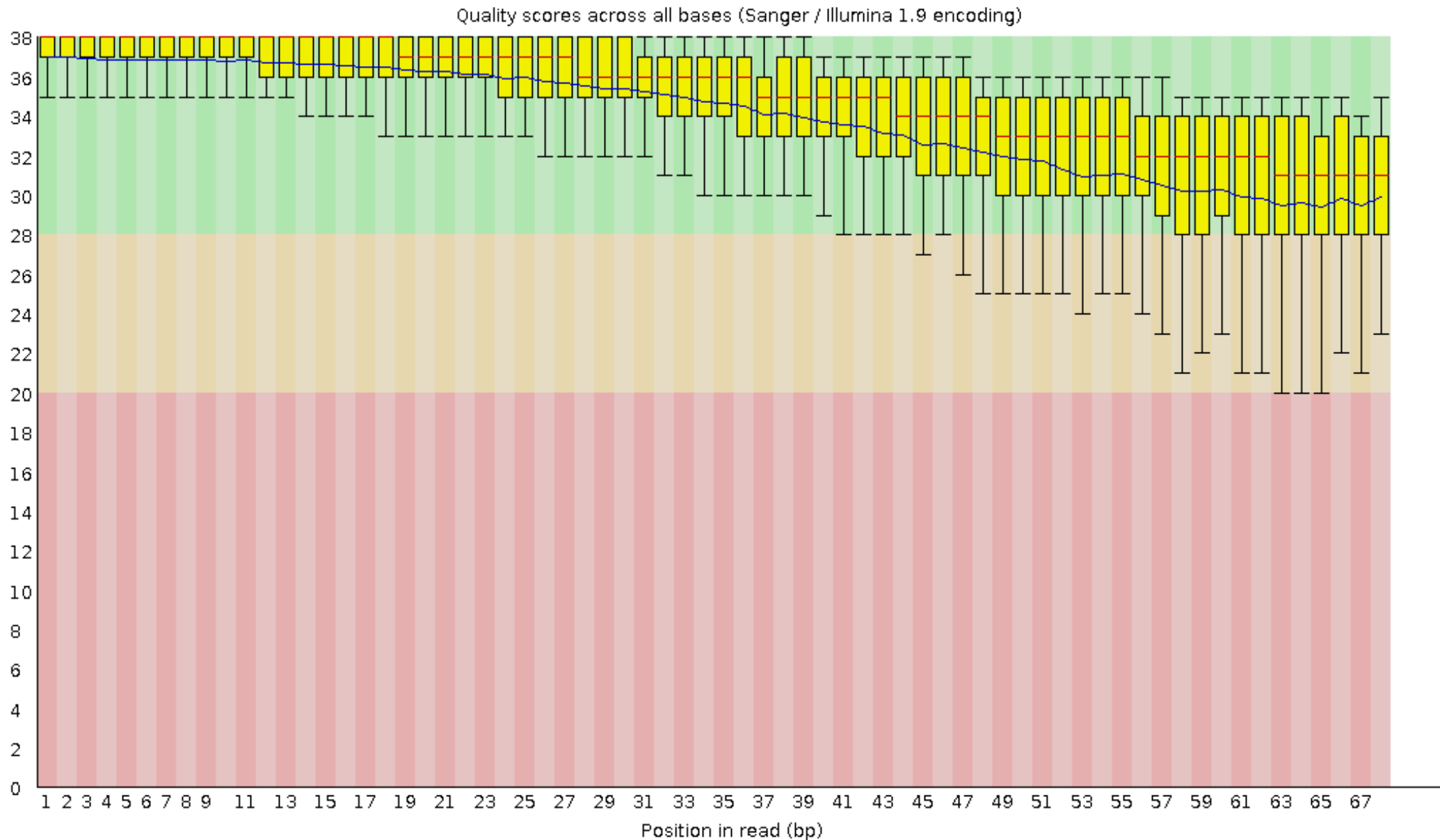
## Élimination des queues polyA/T en 5' et en 3' (si taille $\geq 5$ )



## Élimination des bases en 3' (si **qualité** $< 20$ )



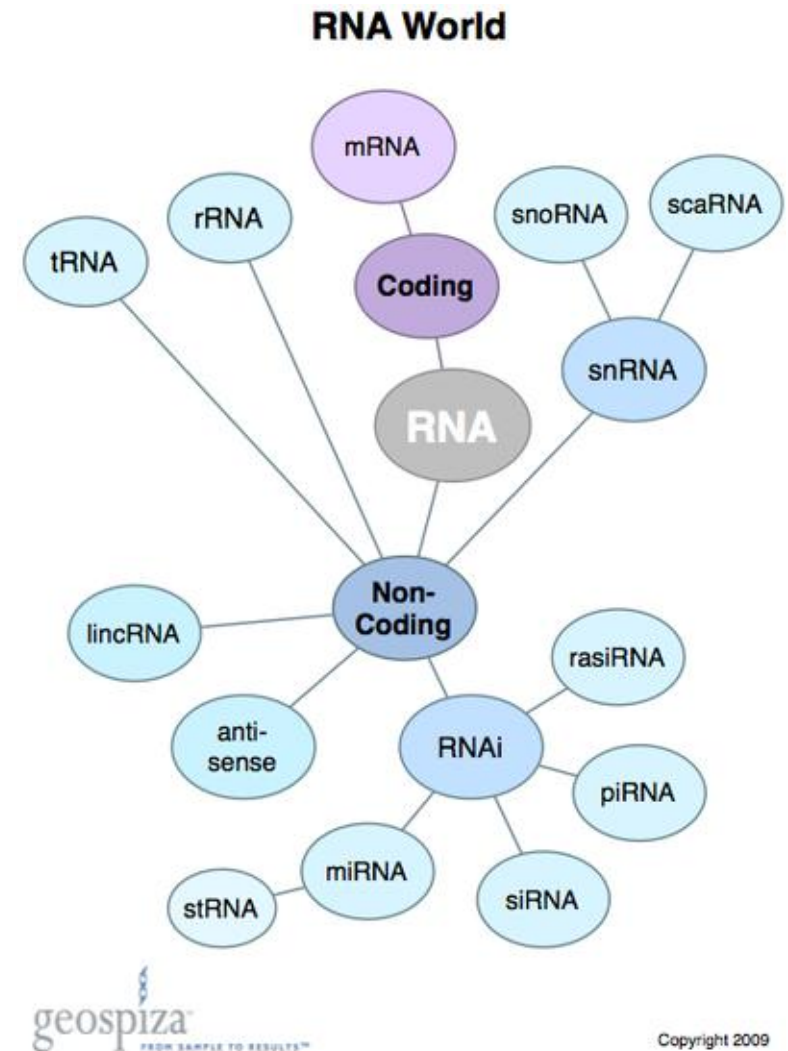
## Base-quality filtered reads



« Ribosomal RNA (rRNA) is the most highly abundant component of RNA, comprising the majority (>80% to 90%) of the molecules present in a total RNA sample. Depletion of this rRNA fraction is desirable prior to performing an RNA-seq reaction, so that sequencing capacity can be focused on more informative parts of the transcriptome.»

O'Neil D et al. (2013) Curr Protoc Mol Biol.

Élimination des reads qui matchent dans des bases de données de rRNA

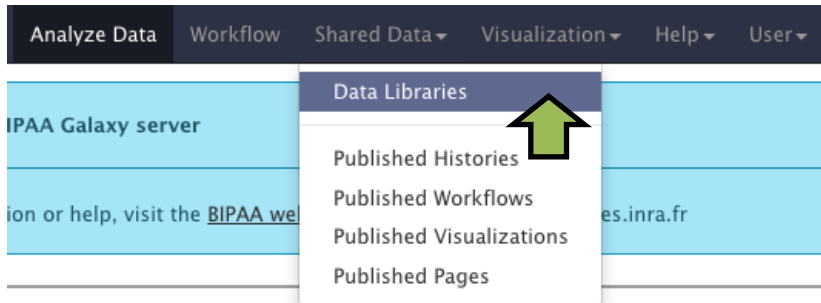


Todd Smith, "Small RNAs Get Smaller," Geospiza FinchTalk (2009)

# I. Traitement initial des reads

## I.1. Chargement des données dans galaxy

Aller dans les Data Libraries



Analyze Data Workflow Shared Data Visualization Help User

IPAA Galaxy server

ion or help, visit the [BIPAA website](#)

**Data Libraries**

- Published Histories
- Published Workflows
- Published Visualizations
- Published Pages

es.inra.fr

Cliquer sur la Data Library BBRIC

### Data Libraries

search dataset name, info, message, dbkey

Advanced Search

Data library name ↓

2013 BP2

**BBRIC**

IncrRNA

[Drosophila melanogaster](#)

Sélectionner tous les fichiers (cliquer sur la case Name) et les charger dans l'historique

### Data Library "BBRIC"

RNASeq de novo assembly

<input checked="" type="checkbox"/> Name	Message
<input checked="" type="checkbox"/> SRS167021_Spombe_gd_SRR097901_75k.left.fastq	None
<input checked="" type="checkbox"/> SRS167021_Spombe_gd_SRR097901_75k.right.fastq	None
<input checked="" type="checkbox"/> SRS167022_Spombe_plat_SRR097918_75K.left.fastq	None
<input checked="" type="checkbox"/> SRS167022_Spombe_plat_SRR097918_75K.right.fastq	None

For selected datasets: Import to current history

Go

### Data Library "BBRIC"

✓ 4 datasets imported into 1 history: Unnamed history

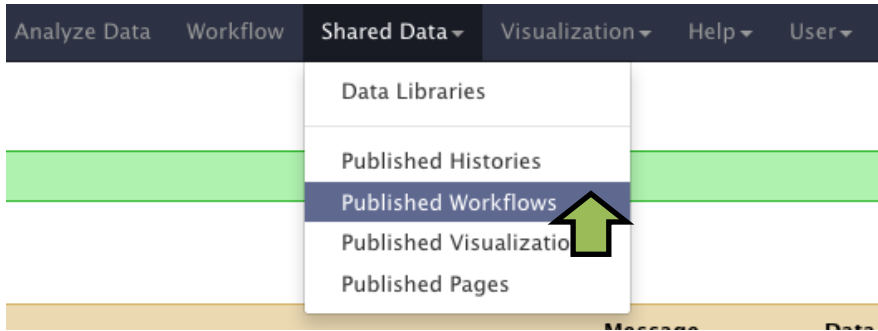




# I. Traitement initial des reads

## I.2. Chargement du pipeline cleaning

Aller dans les Published Workflows



Charger le workflow cleaning

### Published Workflows

[Advanced Search](#)

Name	Annotation
<input type="text" value="RNaseq de novo assembly: 3 expression"/>	RNaseq assembly using trinity. Third step: expression analysis
<input type="text" value="lncRNAs identification workflow"/>	
<input type="text" value="RNaseq de novo assembly: 2 assembly"/>	RNaseq assembly using trinity. Second step: trinity
<input type="text" value="RNaseq de novo assembly: 1 cleaning"/>	RNaseq assembly using trinity. Second step: trinity
<input type="text" value="RNA-Seq alignment on genome"/>	

Import  
Save as File

# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Cliquer sur le lien pour visualiser vos pipelines importés

Workflow "RNASeq de novo assembly: 1 cleaning" has been imported.  
You can [start using this workflow](#) or [return to the previous page](#).

Sélectionner Run pour utiliser le pipeline cleaning

### Your workflows

Name

imported: RNASeq de novo assembly: 1 cleaning

### Workflows shared with you

No workflows have been shared with you

### Other options

Configure your workflow menu

Edit  
Run  
Share or Publish  
Download or Export  
Copy  
Rename  
View  
Delete

Dans ce pipeline, les étapes qui requièrent un paramétrage sont les étapes 1 et 2 pour la sélection des données d'entrée, et enfin les étapes 5 et 6 pour le retrait des adaptateurs

# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Activer la sélection multiple de fichiers pour les étapes 1 et 2

### Running workflow "imported: RNASeq de novo assembly: 1 cleaning"

RNASeq assembly using trinity. Second step: trinity

Step 1: Input datas

Input Dataset

4: Enable/disable selection of multiple input files. Each selected file will have an instance of the workflow.  
SRR097918\_75K.right.fastq

Step 2: Input datas

Input Dataset

4: SRS167022\_Spombe\_plat\_SRR097918\_75K.right.fastq  
type to filter

Step 3: FASTQ Groomer (version 1.0.4)

Step 4: FASTQ Groomer (version 1.0.4)

Cliquer pour activer la sélection multiple de fichiers

History

Unnamed history

57.8 MB

4: SRS167022\_Spombe\_plat\_SRR097918\_75K.right.fastq

3: SRS167022\_Spombe\_plat\_SRR097918\_75K.left.fastq

2: SRS167021\_Spombe\_gd\_SRR097901\_75k.right.fastq

1: SRS167021\_Spombe\_gd\_SRR097901\_75k.left.fastq

Votre historique actuel doit contenir les jeux de données chargés précédemment

# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Sélectionner les jeux de données « left » pour l'étape 1 et « right » pour l'étape 2

### Running workflow "imported: RNASeq de novo assembly: 1 cleaning"

Expand All

Collapse

RNASeq assembly using trinity. Second step: trinity

#### Step 1: Input dataset

##### Input Dataset

- 1: SRS167021\_Spombe\_gd\_SRR097901\_75k.left.fastq
- 2: SRS167021\_Spombe\_gd\_SRR097901\_75k.right.fastq
- 3: SRS167022\_Spombe\_plat\_SRR097918\_75K.left.fastq
- 4: SRS167022\_Spombe\_plat\_SRR097918\_75K.right.fastq

type to filter, [enter] to select all

#### Step 2: Input dataset

##### Input Dataset

- 1: SRS167021\_Spombe\_gd\_SRR097901\_75k.left.fastq
- 2: SRS167021\_Spombe\_gd\_SRR097901\_75k.right.fastq
- 3: SRS167022\_Spombe\_plat\_SRR097918\_75K.left.fastq
- 4: SRS167022\_Spombe\_plat\_SRR097918\_75K.right.fastq

type to filter, [enter] to select all

Dans cette étape, vous sélectionnez 2 banques Spombe\_gd et Spombe\_plat et indiquez quels sont les lots de fichiers (left et right) pour chaque banque. La sélection des fichiers left et right est distincte car chacun est traité indépendamment.

# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Paramétrer l'étape 5 pour les fichiers « left » des banques sélectionnés à l'étape 1

Step 5: Cutadapt (version 1.1.a)

Fastq file to trim

Output dataset 'output\_file' from step 3

3' Adapters

5' or 3' (Anywhere) Adapters

5' or 3' (Anywhere) Adapters 1

Source

Enter custom sequence

Enter custom 5' or 3' adapter sequence

AATTGGCC

Remplacer la séquence de l'adaptateur par:  
AATGATACGGCGACCACCGAGATCTACACNNNNNNN  
NACACTCTTTCCCTACACGACGCTCTTCCGATCT

5' (Front) Adapters

Maximum error rate

0.1

Match times

1

Minimum overlap length

3

Remplacer la valeur du  
chevauchement minimum par:  
49

Match Read Wildcards

False

Do Not Match Adapter Wildcards

False

**Alternative:** les séquences d'adaptateurs peuvent aussi être retirées en utilisant d'autres outils comme FastQC ou prinseq.

# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Paramétrer l'étape 6 pour les fichiers « right » des banques sélectionnés à l'étape 2

### Step 6: Cutadapt (version 1.1.a)

#### Fastq file to trim

Output dataset 'output\_file' from step 4

#### 3' Adapters

##### 5' or 3' (Anywhere) Adapters

##### 5' or 3' (Anywhere) Adapters 1

#### Source

Enter custom sequence

Enter custom 5' or 3' adapter sequence

AATTGCC

Remplacer la séquence de l'adaptateur par:  
GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNN  
NNNNATCTCGTATGCCGTCTTCTGCTTG

#### 5' (Front) Adapters

#### Maximum error rate

0.1

#### Match times

1

#### Minimum overlap length

3

Remplacer la valeur du  
chevauchement minimum par:  
46

#### Match Read Wildcards

False

#### Do Not Match Adapter Wildcards

False

Cliquer sur « run workflow »  
pour exécuter le pipeline

Step 10: riboPicker (version 1.0.0)

Step 11: Get pairs (version 2012-11-

Send results to a new history

Run workflow

# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Le pipeline en cours d'exécution...

✓ Successfully ran workflow "imported: RNASeq de novo assembly: 1 cleaning". The following datasets have been added to the queue:

- 1: SRS167021\_Spombe\_gd\_SRR097901\_75k.left.fastq
- 2: SRS167021\_Spombe\_gd\_SRR097901\_75k.right.fastq
- 5: FASTQ Groomer on data 1
- 6: FASTQ Groomer on data 2
- 7: Cutadapt on data 5 (Report)
- 8: Cutadapt on data 5
- 9: Cutadapt on data 6 (Report)
- 10: Cutadapt on data 6
- 11: Cutadapt on data 5\_good.fastqsanger
- 12: Cutadapt on data 6\_good.fastqsanger
- 13: rRNA
- 14: Non-rRNA
- 15: rRNA
- 16: Non-rRNA
- 17: Clean left
- 18: Clean right
- 19: Non-rRNA.unpaired.fastqsanger
- 20: Non-rRNA.unpaired.fastqsanger

Instance du pipeline pour la banque Spombe\_gd

- 3: SRS167022\_Spombe\_plat\_SRR097918\_75K.left.fastq
- 4: SRS167022\_Spombe\_plat\_SRR097918\_75K.right.fastq
- 21: FASTQ Groomer on data 3
- 22: FASTQ Groomer on data 4
- 23: Cutadapt on data 21 (Report)
- 24: Cutadapt on data 21
- 25: Cutadapt on data 22 (Report)
- 26: Cutadapt on data 22
- 27: Cutadapt on data 21\_good.fastqsanger
- 28: Cutadapt on data 22\_good.fastqsanger
- 29: rRNA
- 30: Non-rRNA
- 31: rRNA
- 32: Non-rRNA
- 33: Clean left
- 34: Clean right
- 35: Non-rRNA.unpaired.fastqsanger
- 36: Non-rRNA.unpaired.fastqsanger

Instance du pipeline pour la banque Spombe\_plat

History  

Unnamed history  
57.8 MB  

36: Non-rRNA.unpaired.fastqsanger			
35: Non-rRNA.unpaired.fastqsanger			
34: Clean right			
33: Clean left			
32: Non-rRNA			
31: rRNA			
30: Non-rRNA			
29: rRNA			
28: Cutadapt on data 22 good.fastqsanger			
27: Cutadapt on data 21 good.fastqsanger			
26: Cutadapt on data 22			
25: Cutadapt on data 22 (Report)			
24: Cutadapt on data 21			
23: Cutadapt on data 21 (Report)			
22: FASTQ Groomer on data 4			
21: FASTQ Groomer on data 3			
20: Non-rRNA.unpaired.fastqsanger			
19: Non-rRNA.unpaired.fastqsanger			
18: Clean right			
17: Clean left			
16: Non-rRNA			
15: rRNA			





# I. Traitement initial des reads

## I.3. Paramétrage et exécution du pipeline cleaning

Fin d'exécution du pipeline cleaning: fichiers « left » et « right » nettoyés / banque

History	
Unnamed history 325.5 MB	
<u>34: Clean right</u>	👁️ ✎️ ✕
<u>33: Clean left</u>	👁️ ✎️ ✕
<u>18: Clean right</u>	👁️ ✎️ ✕
<u>17: Clean left</u>	👁️ ✎️ ✕
<u>4:</u> <u>SRS167022 Spombe plat SRR09791</u> <u>8 75K.right.fastq</u>	👁️ ✎️ ✕
<u>3:</u> <u>SRS167022 Spombe plat SRR09791</u> <u>8 75K.left.fastq</u>	👁️ ✎️ ✕
<u>2:</u> <u>SRS167021 Spombe gd SRR097901</u> <u>_75k.right.fastq</u>	👁️ ✎️ ✕
<u>1:</u> <u>SRS167021 Spombe gd SRR097901</u> <u>_75k.left.fastq</u>	👁️ ✎️ ✕

Fichiers de sortie pour la banque Spombe\_plat

Fichiers de sortie pour la banque Spombe\_gd



Cliquer pour éditer les attributs (dont le nom du fichier)

Editer le nom du fichier en rajoutant le nom de la banque par exemple

Edit Attributes

Name:

Spombe\_plat Clean right

fichiers « left » et « right » nettoyés et renommés / banque

History	
Unnamed history 325.5 MB	
<u>34: Spombe plat Clean right</u>	👁️ ✎️ ✕
<u>33: Spombe plat Clean left</u>	👁️ ✎️ ✕
<u>18: Spombe gd Clean right</u>	👁️ ✎️ ✕
<u>17: Spombe gd Clean left</u>	👁️ ✎️ ✕
<u>4:</u> <u>SRS167022 Spombe plat SRR09791</u> <u>8 75K.right.fastq</u>	👁️ ✎️ ✕
<u>3:</u> <u>SRS167022 Spombe plat SRR09791</u> <u>8 75K.left.fastq</u>	👁️ ✎️ ✕
<u>2:</u> <u>SRS167021 Spombe gd SRR097901</u> <u>_75k.right.fastq</u>	👁️ ✎️ ✕
<u>1:</u> <u>SRS167021 Spombe gd SRR097901</u> <u>_75k.left.fastq</u>	👁️ ✎️ ✕



# I. Traitement initial des reads

## Hors TP: contrôle qualité avec FastQC

**FastQC se trouve dans GALAXY TOOLS**

- NGS: QC and manipulation
- FASTQC: FASTQ/SAM/BA

**1** Dans le menu déroulant, sélectionner le fichier fastq,

**2** puis exécuter FastQC

**Exercice subsidiaire: exécuter FastQC sur les fichiers avant et après nettoyage et comparer les résultats**

**FastQC:Read QC (version 0.52)**

Short read data from your current history:  
34: Spombe\_plat Clean right

Title for the output file – to remind you what the job was for:  
FastQC

Contaminant list:  
Selection is Optional

Execute

**Purpose**  
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:  
Import of data from BAM, SAM or FastQ files (any variant)  
Providing a quick overview to tell you in which areas there may be problems  
Summary graphs and tables to quickly assess your data  
Export of results to an HTML based permanent report  
Offline operation to allow automated generation of reports without running the interactive application

**FastQC**  
This is a Galaxy wrapper. It merely exposes the external package `FastQC` which is documented at `FastQC`. Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the `Picard-tools` libraries for sam/bam processing.

The contaminants file parameter was borrowed from the independently developed `fastqcwrapper` contributed to the Galaxy Community Tool Shed by J. Johnson.

**Inputs and outputs**  
`FastQC` is the best place to look for documentation – it's very good. A summary follows below for those in a tearing hurry.

This wrapper will accept a Galaxy `fastq`, `sam` or `bam` as the input read file to check. It will also take an optional file containing a list of contaminants information, in the form of a tab-delimited file with 2 columns, name and sequence.

The tool produces a single HTML output file that contains all of the results, including the following:

Basic Statistics  
Per base sequence quality  
Per sequence quality scores  
Per base sequence content  
Per base GC content  
Per sequence GC content  
Per base N content  
Sequence Length Distribution  
Sequence Duplication Levels  
Overrepresented sequences  
Kmer Content

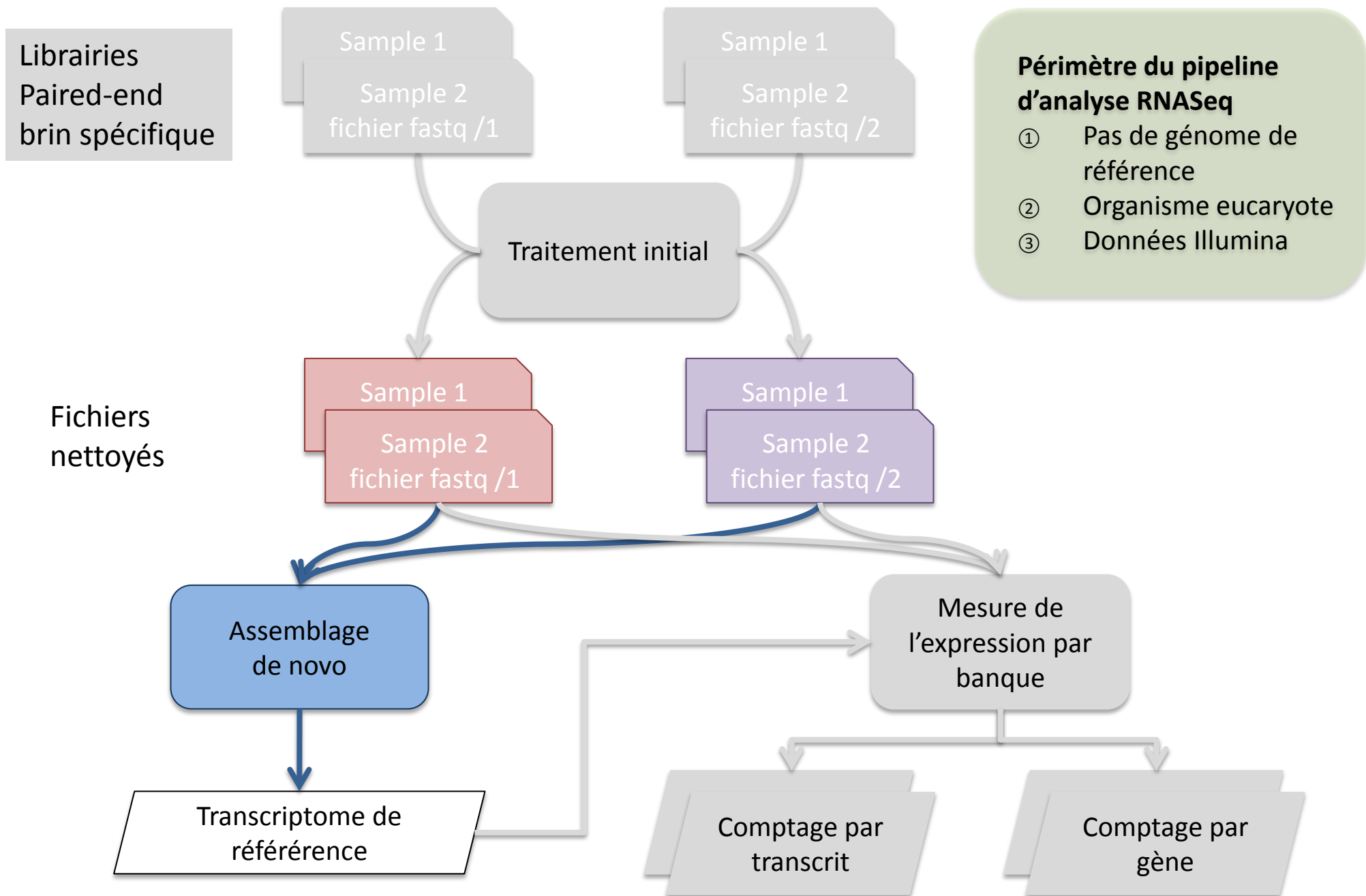
All except Basic Statistics and Overrepresented sequences are plots.

History  
Unnamed history  
325.5 MB

- 34: Spombe\_plat Clean right
- 33: Spombe\_plat Clean left
- 18: Spombe\_qd Clean right
- 17: Spombe\_qd Clean left
- 4: SRS167022\_Spombe\_plat\_SRR09791\_8\_75K.right.fastq
- 3: SRS167022\_Spombe\_plat\_SRR09791\_8\_75K.left.fastq
- 2: SRS167021\_Spombe\_qd\_SRR097901\_75k.right.fastq
- 1: SRS167021\_Spombe\_qd\_SRR097901\_75k.left.fastq



- Introduction générale sur le RNASeq
  - Vue d'ensemble des grandes étapes dont l'analyse par assemblage de novo de transcriptome et la mesure de l'expression
- L'assemblage *de novo* de transcriptome
  - Traitement initial des reads
    - Élimination des artefacts
  - L'assemblage *de novo*
    - Filtrage/Normalisation des reads par couverture des k-mers
    - Cas de Trinity
    - Qualité de l'assemblage
  - Mesure de l'expression par banque
    - Mapping
    - Filtrage et comptage



L'assemblage de novo est très sensible

- ① à la quantité de données, provenant notamment des gènes fortement exprimés,
- ② au nombre d'erreurs (indiscernables de vrais variations pour les transcrits peu abondants).

=> consommation mémoire et temps de calcul très importants

**Solution:** réduire le jeu de données à un volume utile sans perte importante d'information afin d'accélérer et d'améliorer l'étape d'assemblage.

=> Normalisation par couverture de kmer

- Découpage de toutes les lectures en sous séquences de même taille (**k**), appelées **K-mer**
- Deux K-mer vont être liés si leurs séquences ne divergent que par le premier nucléotide pour l'un et par le dernier pour l'autre

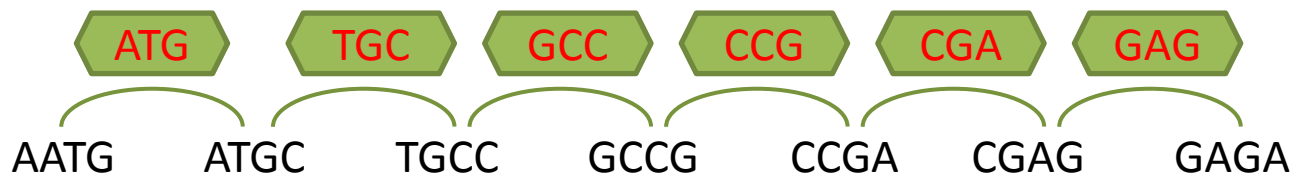
Seq1: AATGCCGA                      Seq2: TGCCGAGA

- Découpage des séquences en K-mer avec k=4

Seq1: AATG    ATGC    TGCC    GCCG    CCGA  
Seq2: TGCC    GCCG    CCGA    CGAG    GAGA

Les séquences ont des K-mer commun

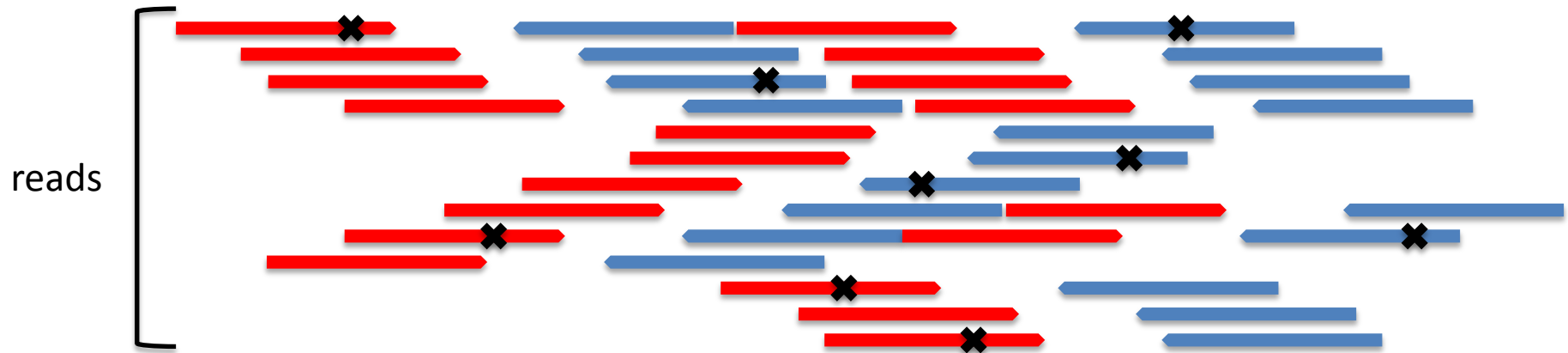
- Création du graphe de de Bruijn



Un K-mer est présent qu'une fois dans le graphe

- Lecture du graphe de de Bruijn

AATGCCGAGA

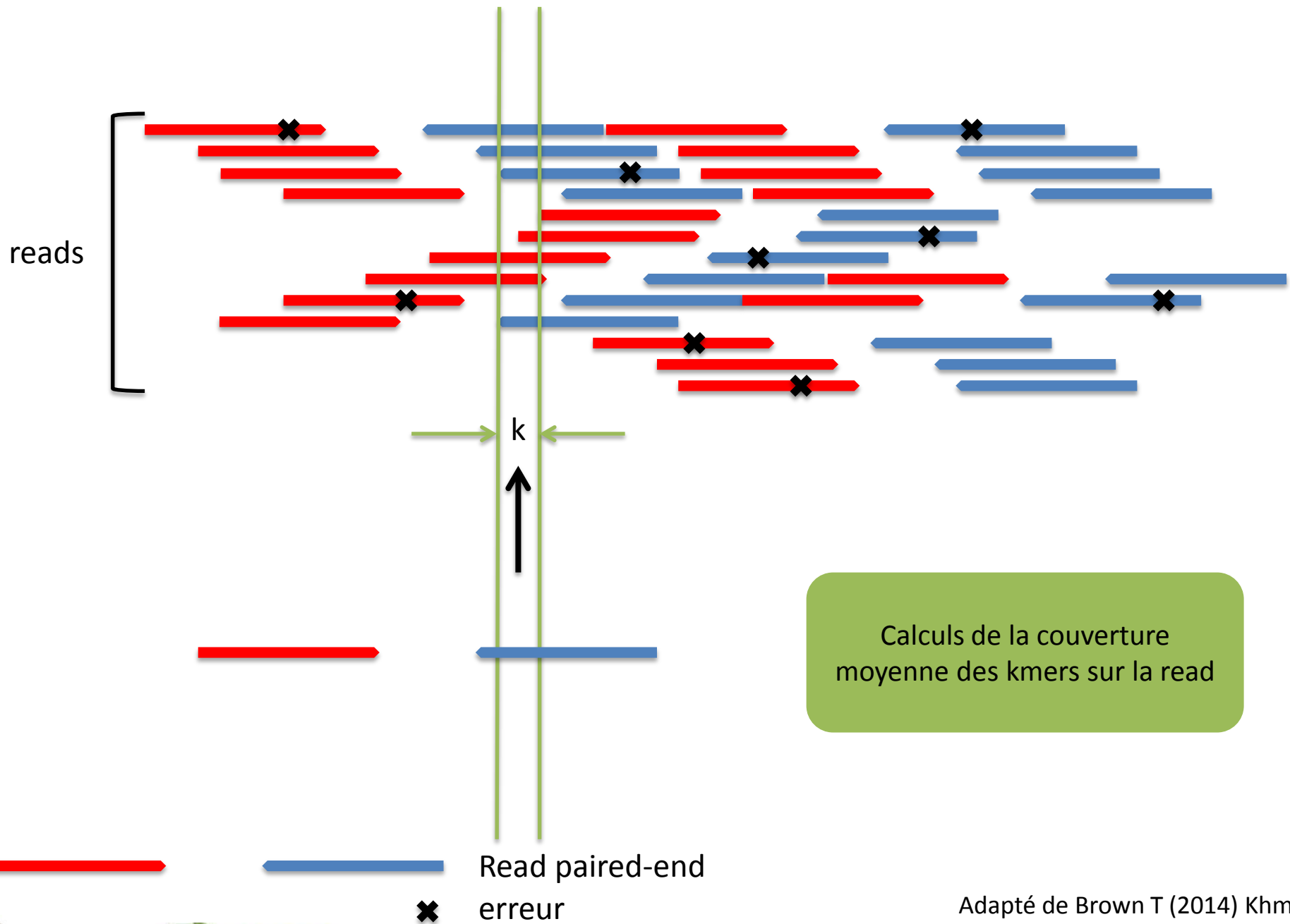


Read paired-end

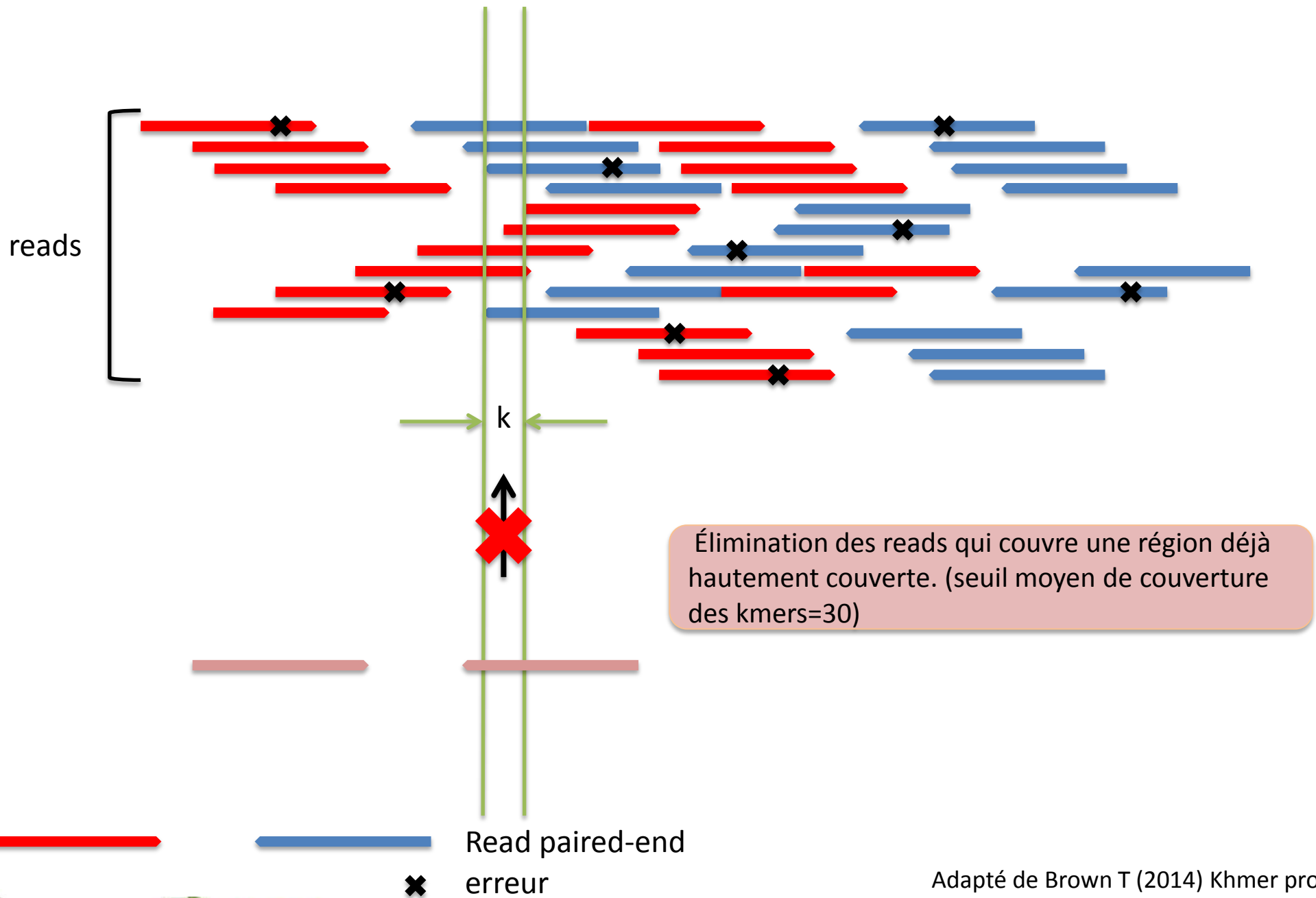


erreur

Adapté de Brown T (2014) Khmer protocols

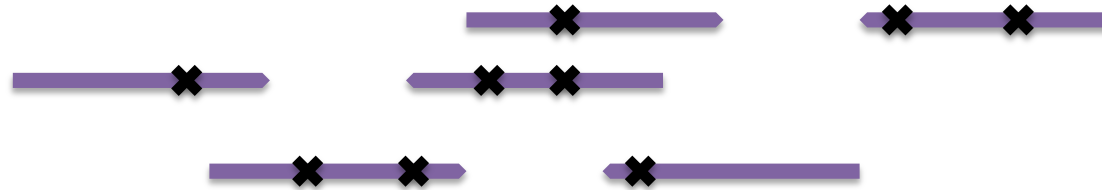
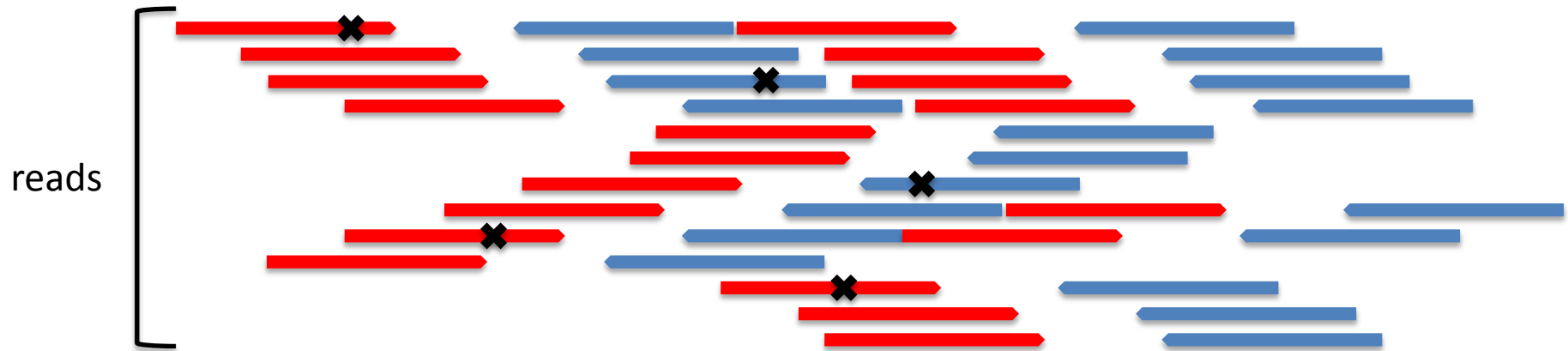


Adapté de Brown T (2014) Khmer protocols



Adapté de Brown T (2014) Khmer protocols





Élimination des reads qui contiennent des erreurs (région hautement couverte)



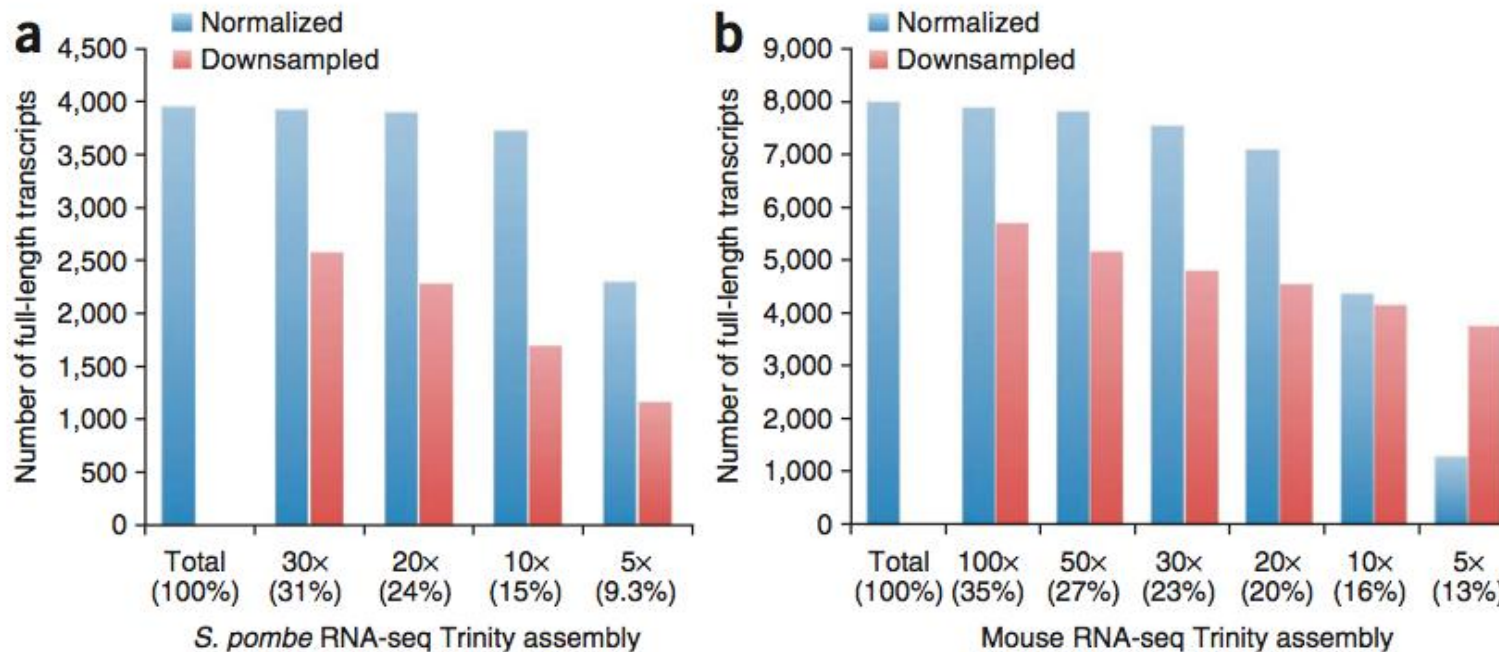
Read paired-end



erreur

Adapté de Brown T (2014) Khmer protocols

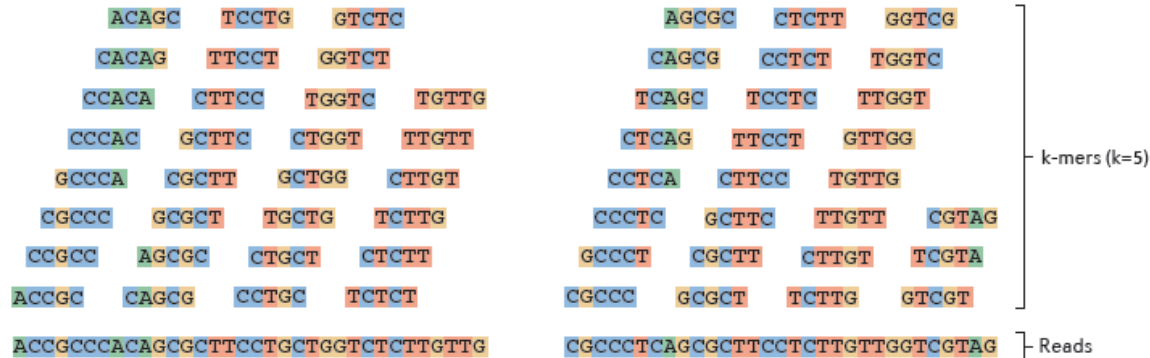




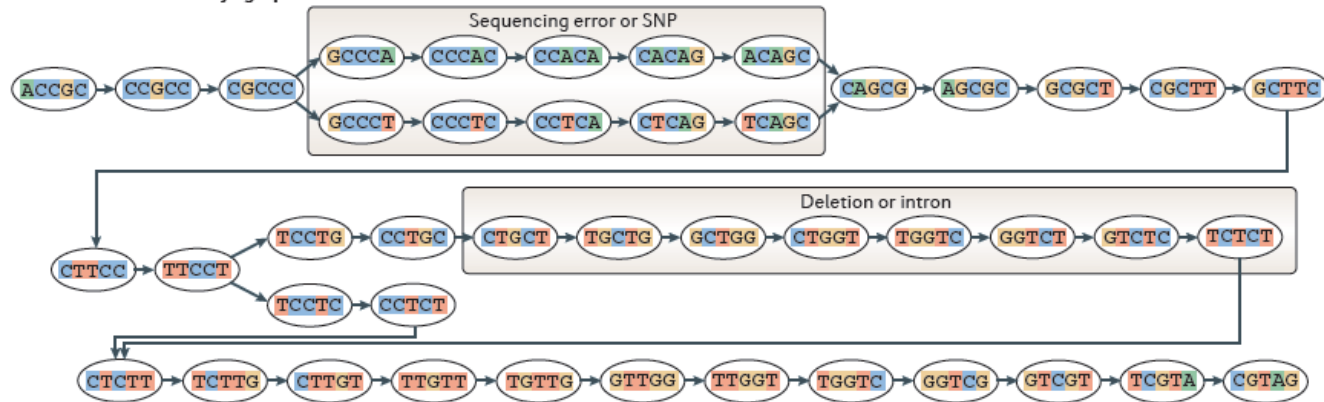
Haas B et al. (2013) Nature Protocols

La normalisation (par couverture des kmers) permet de réduire le jeu de données à une taille utile, sans perte importante d'information, afin d'accélérer et d'améliorer l'étape d'assemblage.

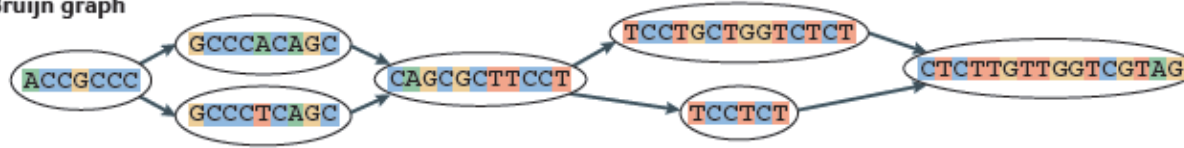
**a** Generate all substrings of length k from the reads



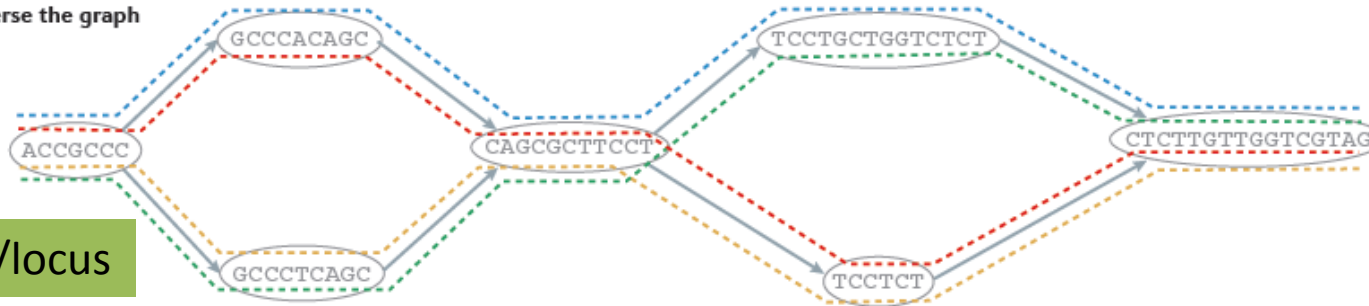
**b** Generate the De Bruijn graph



c Collapse the De Bruijn graph



d Traverse the graph



Composante/locus

e Assembled isoforms

- ① Isoformes
- ② Paralogues
- ③ variants alléliques

- - - - - ACCGCCACAGCGCTTCCTGCTGGTCTCTTGGTGGTCGTAG  
 - - - - - ACCGCCACAGCGCTTCCT - - - - - CTTGTTGGTCGTAG  
 - - - - - ACCGCCCTCAGCGCTTCCT - - - - - CTTGTTGGTCGTAG  
 - - - - - ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGGTGGTCGTAG

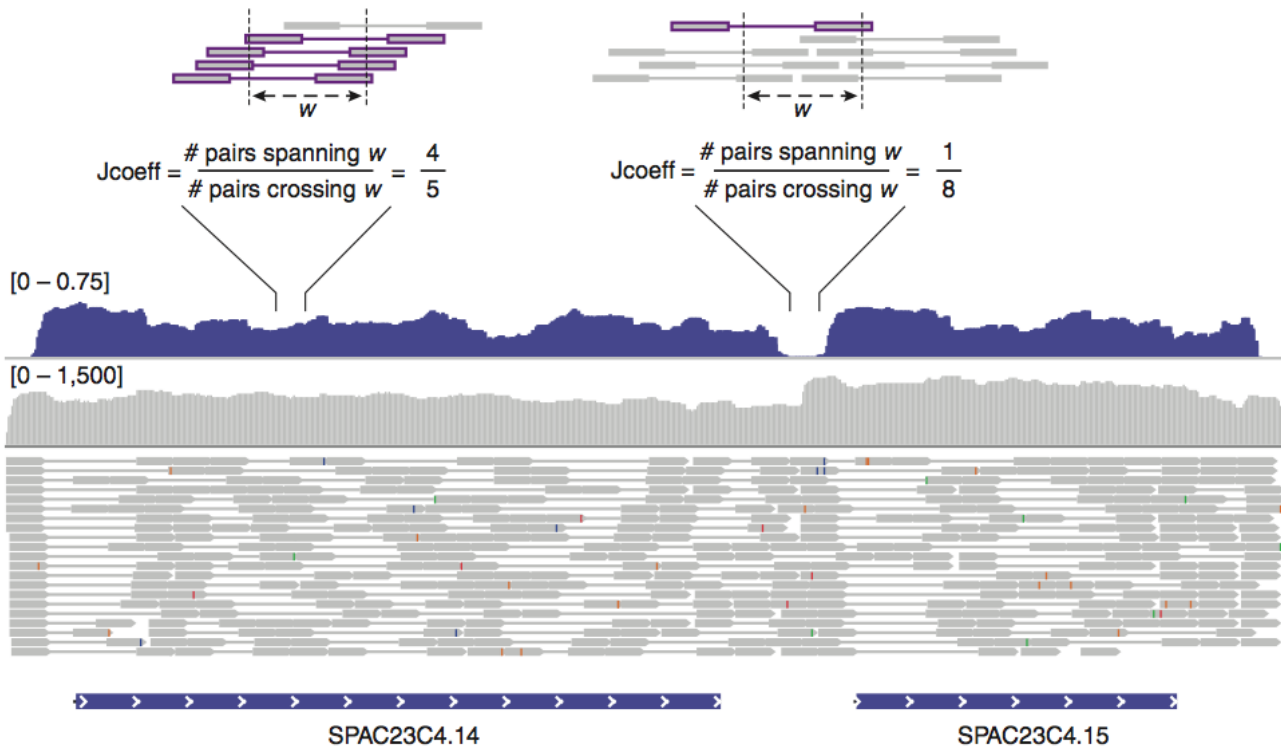
## Sortie de Trinity: Fichier fasta des transcrits assemblés

```
>comp18_c0_seq1 len=608 path=[7155:0-276 12558:277-339 11658:340-607]
CTCTCACTCAAAGTTGCACCTTAACCAAACAAATTAATCAAATGGGTCGTATGCATAG
CAAAGGAAAGGGTATCGCTTCTTCCGCTTTACCCTACGTTTCGCTCTCCCCCTGCTTGGTG
CAAGGCTGATGCCGACTCCGTTGTGCGAGCAAATTCTTAAGTTCTCCAAGAAGGGTATGTC
CCCTTCTCAAATTGGTGTGACTCTCCGTGACTCTCATGGAATTCCTCAAGTTCGTTTCAT
CACTGGTCAAAGATCATGCGTATCTTGAAGGCTAATGGTATGCCACTTTTACTTAGTTA
TTAAAGCCTCTGATTGGTTTTGTGTGCTGACAGAGTGATAGGTCTTGCTCCCGAGCTCCC
CGAGGATCTTACAATCTTATTAAGAAGGCTGTTTCCGTCCGCAAGCATTGGAACGTAA
CCGTAAGGATAAGGACTCCAAGTTCGTTTGATTCTTATTGAGTCTCGTATCCACCGTCT
TGCTCGTTACTACAGAAAGGTCGGTGCTCTTCCCCCTACCTGGAAGTACGAATCTGCTAC
TGCTTCTGCTTTGGTTGCTTAAGTTAGTAGAAAGTGAGCCCTTAATTGAAGCTTGCTTAG
GTCTTCTT
```

```
>comp18_c0_seq2 len=545 path=[7155:0-276 11658:277-544]
CTCTCACTCAAAGTTGCACCTTAACCAAACAAATTAATCAAATGGGTCGTATGCATAG
CAAAGGAAAGGGTATCGCTTCTTCCGCTTTACCCTACGTTTCGCTCTCCCCCTGCTTGGTG
CAAGGCTGATGCCGACTCCGTTGTGCGAGCAAATTCTTAAGTTCTCCAAGAAGGGTATGTC
CCCTTCTCAAATTGGTGTGACTCTCCGTGACTCTCATGGAATTCCTCAAGTTCGTTTCAT
CACTGGTCAAAGATCATGCGTATCTTGAAGGCTAATGGTCTTGCTCCCGAGCTCCCCGA
GGATCTCTACAATCTTATTAAGAAGGCTGTTTCCGTCCGCAAGCATTGGAACGTAACCG
TAAGGATAAGGACTCCAAGTTCGTTTGATTCTTATTGAGTCTCGTATCCACCGTCTTGC
TCGTTACTACAGAAAGGTCGGTGCTCTTCCCCCTACCTGGAAGTACGAATCTGCTACTGC
TTCTGCTTTGGTTGCTTAAGTTAGTAGAAAGTGAGCCCTTAATTGAAGCTTGCTTAGGTC
TTCTT
```



# Détection des fusions de transcrits



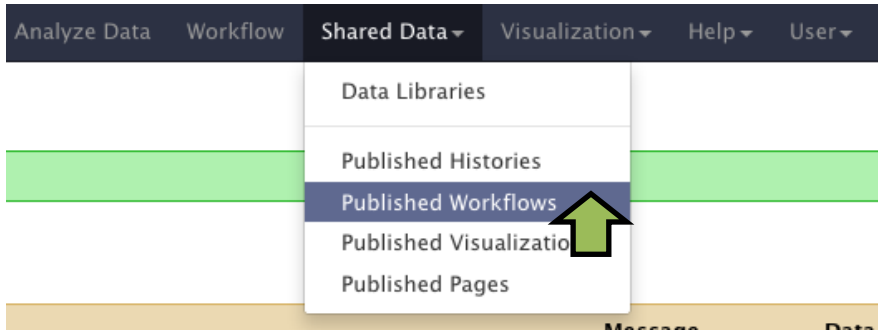
Haas B et al. (2013) Nature Protocols

Le coefficient de Jaccard permet de détecter les fusions de transcrits dans les génomes à forte densité de gènes qui peuvent se chevaucher sur leurs UTRs.

## II. Assemblage de novo

### II.1. Chargement du pipeline assembly

Aller dans les Published Workflows



The screenshot shows a navigation bar with the following items: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The 'Shared Data' dropdown menu is open, showing options: Data Libraries, Published Histories, Published Workflows (highlighted with a green arrow), Published Visualizations, and Published Pages.

Charger le workflow assembly

#### Published Workflows

search name, annotation, owner, and tag

[Advanced Search](#)

Name	Annotation
RNASeq de novo assembly: 3 expression	RNASeq assembly using trinity. Third step: expression analysis
lncRNAs identification workflow	
RNASeq de novo assembly: 2 assembly	RNASeq assembly using trinity. Second step: trinity
RNASeq de novo assembly:	RNASeq assembly using trinity. Second step: trinity
RNA-Seq alignment on genome with Tophat and Cufflinks	

The table shows a list of workflows. The workflow 'RNASeq de novo assembly: 2 assembly' is highlighted with a red box. A context menu is open over it, showing 'Import' and 'Save as File' options, with a green arrow pointing to 'Import'.

## II. Assemblage de novo

### II.1. Chargement du pipeline assembly

Cliquer sur le lien pour visualiser vos pipelines importés



Workflow "RNASeq de novo assembly: 2 assembly" has been imported.  
You can [start using this workflow](#) or [return to the previous page](#).

Sélectionner Run pour utiliser le pipeline assembly

#### Your workflows

Name

imported: RNASeq de novo assembly: 2 assembly

imported: RNASeq de novo assembly:

Edit

Run

Share or Publish

Download or Export

Copy

Rename

View

Delete

#### Workflows shared with

No workflows have been shared with you

#### Other options

Configure your workflow menu



## II. Assemblage de novo

### II.2. Paramétrage et exécution du pipeline assembly

Sélectionner les fichiers d'entrée pour les étapes 1 (Spombe\_gd/plat clean left) et 2 (Spombe\_gd/plat clean right)

#### Running workflow "imported: RNASeq de novo assembly: 2 assembly"

Expand All

Collapse

RNASeq assembly using trinity. Second step: trinity

##### Step 1: Concatenate multiple datasets (version 1.0.0)

###### Concatenate Datasets

17: Spombe\_gd Clean left  
18: Spombe\_gd Clean right  
33: Spombe\_plat Clean left  
34: Spombe\_plat Clean right

Don't unzip gzip or bzip2 files if possible

False

Action:

Hide output 'out\_file1'.

##### Step 2: Concatenate multiple datasets (version 1.0.0)

###### Concatenate Datasets

17: Spombe\_gd Clean left  
18: Spombe\_gd Clean right  
33: Spombe\_plat Clean left  
34: Spombe\_plat Clean right

Don't unzip gzip or bzip2 files if possible

False

Action:

Hide output 'out\_file1'.

Dans les étapes 1 et 2, vous sélectionnez les lots de fichiers clean left (étape 1) et clean right (étape 2) obtenus précédemment et indépendamment des banques.

Le pipeline compile tous les fichiers left puis right pour ne travailler qu'avec 1 seul fichier left et un seul fichier right et assembler le transcriptome

## II. Assemblage de novo

### II.2. Paramétrage et exécution du pipeline assembly

Paramétrage de l'étape 3 de normalisation par la couverture des kmers

Step 3: normalize by kmer coverage (version r2012-10-05)

**single or paired reads**

paired

**left reads fastq file**

Output dataset 'out\_file1' from step 1

**right reads fastq file**

Output dataset 'out\_file1' from step 2

**pairs\_together**

True

**SS\_lib\_type**

✓ None

RF

FR

30

**KMER\_SIZE**

25

**max\_pct\_stdev**

100

Les banques sont du type  
brin spécifique en RF  
(sélectionner cette option)

## II. Assemblage de novo

### II.2. Paramétrage et exécution du pipeline assembly

#### Paramétrage de l'étape 4 d'assemblage de novo avec Trinity

##### Step 4: Trinity (version 0.0.2)

###### Paired or Single-end data?

Paired

###### Left/Forward strand reads

Output dataset 'output\_left' from step 3

###### Right/Reverse strand reads

Output dataset 'output\_right' from step 3

###### Is it strand specific data?

Yes

Forward-Reverse  
✓ Reverse-Forward

###### Paired Fragment Length

300

###### Jaccard Clip options

###### Use additional Params?

No

- ① Les banques sont du type brin spécifique en Reverse-Forward (sélectionner cette option)
- ② Cocher l'option Jaccard Clip Options (Spombe est connu pour être un genome a forte densité de genes)

Cliquer sur « run workflow » pour exécuter le pipeline

Run workflow

##### Actions:

Rename output 'assembled\_transcripts' to 'Assembled transcripts'.

Hide output 'trinity\_log'.

# II. Assemblage de novo

## II.2. Paramétrage et exécution du pipeline assembly

Le pipeline en cours d'exécution...

Galaxy by GenOuest

Successfully ran workflow "imported: RNASeq de novo assembly: 2 assembly". The following datasets have been added to the queue:

- 37: Concatenate multiple datasets on data 33, data 17, and data 17
- 38: Concatenate multiple datasets on data 34, data 18, and data 18
- 39: normalized\_K25\_C30\_pctSD100 on Concatenate multiple datasets on data 33, data 17, and data 17
- 40: normalized\_K25\_C30\_pctSD100 on Concatenate multiple datasets on data 34, data 18, and data 18
- 41: Trinity on data 39 and data 40: log
- 42: Assembled transcripts

History

Unnamed history  
325.5 MB

- 42: Assembled transcripts
- 41: Trinity on data 39 and data 40: log
- 40: normalized\_K25\_C30\_pctSD100 on Concatenate multiple datasets on data 34, data 18, and data 18
- 39: normalized\_K25\_C30\_pctSD100 on Concatenate multiple datasets on data 33, data 17, and data 17
- 38: Concatenate multiple datasets on data 34, data 18, and data 18
- 37: Concatenate multiple datasets on data 33, data 17, and data 17
- 34: Spombe plat Clean right
- 33: Spombe plat Clean left
- 18: Spombe qd Clean right
- 17: Spombe qd Clean left
- 4: SRS167022 Spombe plat SRR097918\_75K.right.fastq
- 3: SRS167022 Spombe plat SRR097918\_75K.left.fastq
- 2: SRS167021 Spombe qd SRR097901\_75k.right.fastq
- 1: SRS167021 Spombe qd SRR097901\_75k.left.fastq



# II. Assemblage de novo

## II.2. Paramétrage et exécution du pipeline assembly

Visualisation du fichier fasta contenant les transcrits assemblés

The screenshot shows the Galaxy by GenOuest web interface. The main panel displays the output of an assembly pipeline, showing FASTA sequences for three samples:

```
>comp0_c0_seq1 len=304 path=[90:0-303]
GACGAAAAAGGTGAGAACGTTGGTCTGAACTTCGTAAGCGAGTCATCGAACACAATATT
CGTGTAGTCGCAAACTATTATCCCGCATTTCATTCAGTCTAGGAGTGTGTGGAT
ATGAGTCTCCGAAACAGAACAAATTTCTTGGCATCTTATTGCAAGCATCATTTCTAT
GCCAAGATTGATCGCCAGCTCAAGTTCATTCATTCAGAAATTCAAAATGTGCAAGAG
CAATTGAATGAGTGGGGTCCAATATTACTGAACTTCTTGGAAAATGGAAAAGGTAAGA
CAGG
>comp1_c0_seq1 len=821 path=[390:0-77 6095:78-78 952:79-820]
GTCAAAATTTTTTGGACTTAAACAATACGTTCTCTAAGATTTTGTTCCTAGTAGAGTA
GGAAGTTTAGTATATTCATTTTTTTTTAAAGTGCATAAGATAAAGAAACGCGATGCTC
TTTCATATTACGCCCATATGGATGTGGAAGGAACGGGCAACAACATATGCCTATTGCTG
ACTTGTGATAAAACAAAATCACTGCCATCGTCGACCCCTGCTGAGCCTGAGAGTGTATT
CCGTGCATAAAAAGAAAAACGGCGAAGAAAGAAATGATTTGCAGTACATTCGACCACT
CATCATCACTATGATCATGACGGCGCAACGAAGACATTTTGGACTACTTCCCTCTTTA
AAGGCTATGGCGGAAAAAATGCATCTGGTGTACATACACTCCCAAGATAAGGAGATT
TTCGAAGTTGGAGAAGTTCAAGTTGAGGCTCTTCATACGCCATGCCATACCCAAGACA
ATTTGTTATTACGTTTCCTCTCCTCAAAAAGAGCCGCTGTTACTGGAGATACTCTTT
ACATCAGGCTGTGGTCTGCTTTTTTGAAGGAGATGCCAAACAATGGACTACGCTTTGA
CAGTCCCTGTGCTTTACCCGATGATACCGTAACCTACCTGGCATGAATACACCA
TCAAATGCCAAGTTTAGTCCACGATCTTCTACTCTGAGTTCACAAAATTTGGTGG...
TTTTGCAAGAATCATGAATCTACTACTGGTCACTTACCATCGGTGACGAAAAAGATTG
AATCCATTCATGTGCTTGGGTAGAATCCGGTACAAAAGC
>comp2_c0_seq1 len=236 path=[1:0-81 85:82-235]
CAACGCTGCTCCGCACTGATGACTGTTCTGGCACTGTTGAGGAGTGTGAGCCTGAAG
CTGGAACGTGCACTACTACTGTATATTCGGCACTCAAGAGTACACTACCACACTTGCTA
CAGCCAGTGTGACTGTTTCAGGCACTGTTGAAGTAGTGGATACAGCAGCTGGAACCTGA
CTACTACAATTTACTCTGGAACCTACACCATTAAACACAACGCTTGTCTCCGCAACT
>comp3_c0_seq1 len=308 path=[168:0-273 2351:274-307]
CATCATCTGAAGATGAACCTGCTGCTTTCGATCCGTTTTCCCGACGGATCTCGGGCTG
TTCGAAGTTTAAAGAGGATGATACTGTTGAATCCGTTTTACAACATGTTGATTAATGTC
TTTTTGGAAAAGAGGAGCCTGAAGAATTCGGTCCGCTACTTCACTCTTAATCCGGTCA
CTCCACCAAGTACTATAAGCATGATTTTCAATTTTCAACTATATCTTCTTTGCCCCAGAG
CTTTGCTGAAACCATCAGTGGCTATTTCTACTAACAAGCCATCTTCCAAATGGCACTG
TGGTGGTG
>comp6_c0_seq1 len=757 path=[1032:0-719 5360:720-720 5366:721-756]
TTAGCTCTCTATATGGGTTAATTTGTTTCTTCGCTTTTGGTTTTTAGAGGATTCATA
CAAAAGCTACTGAATTCCTTTTCAACGGCTTATAGTTAAAAGAAATGCCAACGTTTTTA
ATGGATTTCTGGATTTGAAATCTTGATTTTTATTGACTAGATTTCCCTAACCTCATCTT
ATTCCTTGTAGTTAATGAGCGAGCAATTCATTATAAATGGAATGTTCTGCAATGA
CTATTTATCGAAAATGTGCTCGTTTACTTTTCGATTTGATGCTTAGAATGGGCAATTTG
TTACCTTTTTTATGAGCGTTGGGTGTATGATCTGCATATTGCATTAATTTTATCTGGC
TAGTTTTTATGCTATTAATAATCTGTTGCTTTTATGAGATTTGTTGAAACTT
GTTTCCCGGTCCTGACATTTTTTCTCTGTGTTTTTCTTTAATTTCCTAATTA
TTACTGACGTGTGCTATTTGCTTTGGGATGACAGTACATCTTATTTCTGTGCGGAG
ATTATAGTTATAATCTTCTTCTCTCATGAACCTTGGCTTTGGTTGCCAACCTTTAAAC
CTTCCGTCATGATATACATGATGTTTGTCAATATAAGTGATATTTATGTTGTCACCTTT
TTTGGCGAATGTTGAACCTATCTCTACGCTCAAGGAGTACATGCAATGTGCTTTATGCT
TTTTTTTTTTGAAATGTTTGCATTTTCAAGCCCTC
>comp7_c0_seq1 len=282 path=[1170:0-190 1994:191-281]
CAGCGATTCATTTGCAGTCTGTTTGGCATGCGCCATGATACACGTTCTTGGAGCGCAA
AATATGCTACTCCACTCAATGCTCTTCTCTTTTGAACAATCTGCTGGAAATGGCAAC
```

A central text box indicates: **1784 transcrits assemblés**

The right sidebar shows a history list with 42 items, including 'Assembled transcripts' (1,784 sequences) and 'Spombe plat Clean right' (1,784 sequences).



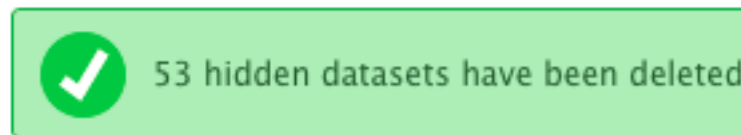
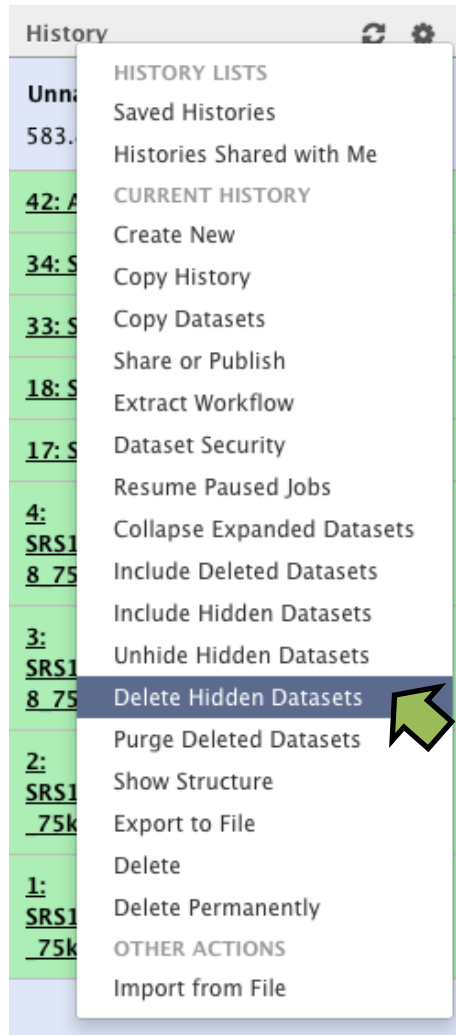
Bioinformatique  
Biodiversité  
Représentation  
& Intégration  
des Connaissances



## II. Assemblage de novo

### II.2. Paramétrage et exécution du pipeline assembly

Suppression des fichiers cachés



Sélectionner et confirmer la suppression

## II. Assemblage de novo

### II.2. Paramétrage et exécution du pipeline assembly

Purge des fichiers supprimés

History

- HISTORY LISTS
- Saved Histories
- 583. Histories Shared with Me
- CURRENT HISTORY
- 42: A Create New
- 34: S Copy History
- 33: S Copy Datasets
- Share or Publish
- 18: S Extract Workflow
- 17: S Dataset Security
- Resume Paused Jobs
- 4: Collapse Expanded Datasets
- 8 75 SRS1**
- 8 75 Include Deleted Datasets
- Include Hidden Datasets
- 3: Unhide Hidden Datasets
- 8 75 SRS1**
- 8 75 Delete Hidden Datasets
- 2: Purge Deleted Datasets**
- 2: Show Structure
- 75k Export to File
- Delete
- 1: Delete Permanently
- 75k SRS1**
- OTHER ACTIONS
- Import from File

✓ 57 datasets have been deleted permanently

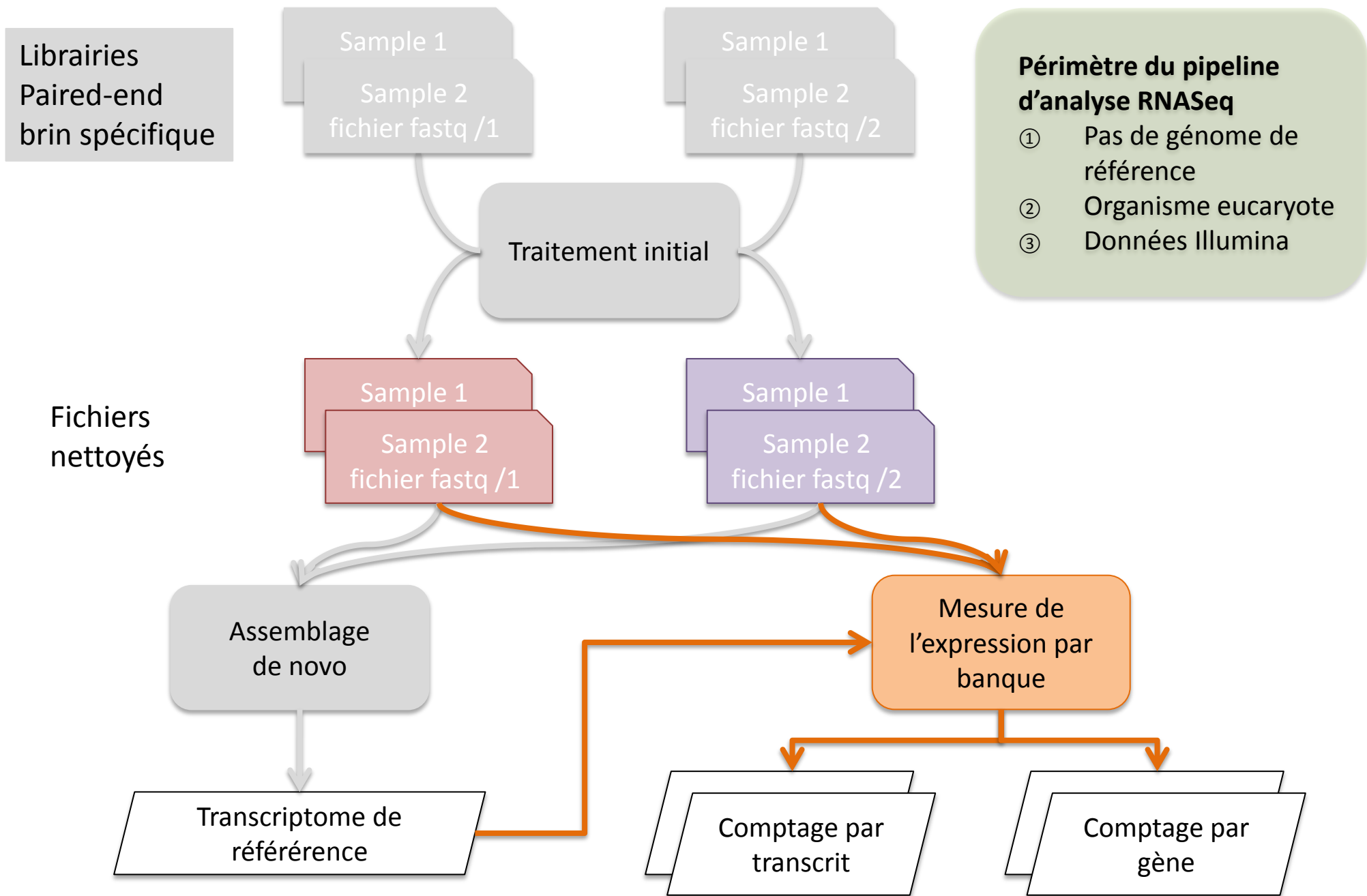
Sélectionner et confirmer la purge



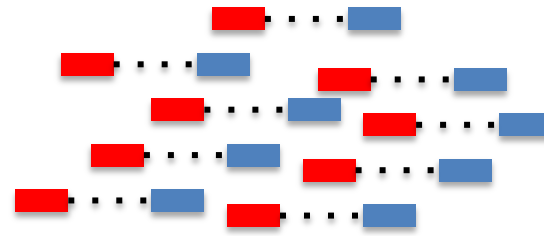
- Présentation et organisation
- Introduction générale sur le RNASeq
  - Vue d'ensemble des grandes étapes dont l'analyse par assemblage de novo de transcriptome et la mesure de l'expression
- L'assemblage *de novo* de transcriptome
  - Traitement initial des reads
    - Élimination des artefacts
  - L'assemblage *de novo*
    - Filtrage/Normalisation des reads par couverture des k-mers
    - Cas de Trinity
    - Qualité de l'assemblage
  - Mesure de l'expression par banque
    - Mapping
    - Filtrage et comptage



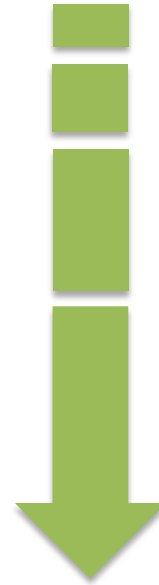
# Le pipeline: mesure de l'expression par banque



Paired-end reads



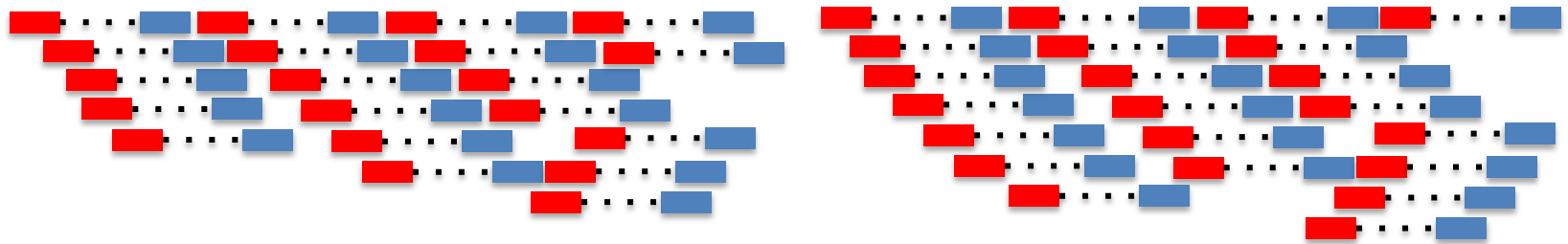
Alignement



Paramètres:

- ① Intervalle de la taille d'insert
- ② Nombre de hits

reads



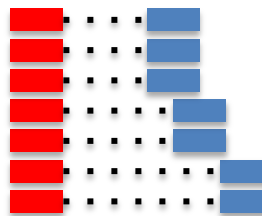
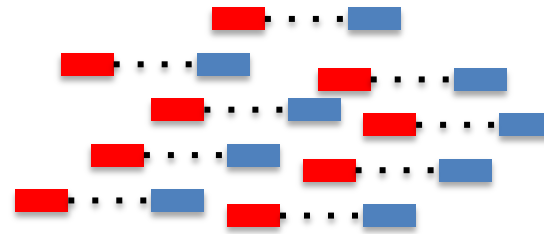
assemblage



Contig 1

Contig 2

Paired-end reads



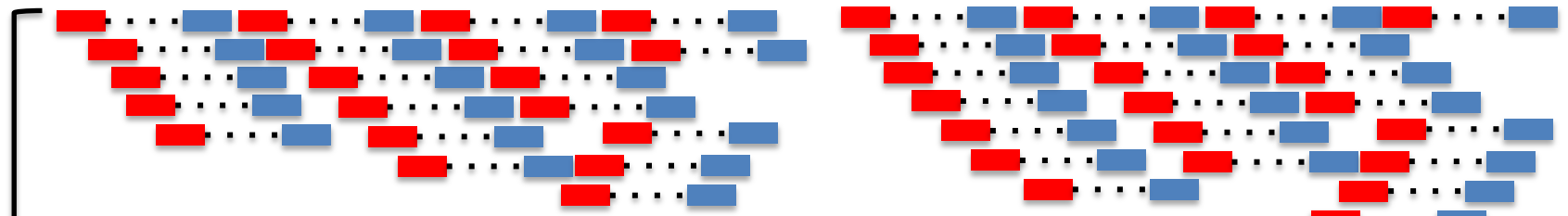
**Taille d'insert autorisé:**  
entre 100 et 500\*

\* Par défaut dans le pipeline

Alignement



reads



assemblage



Contig 1

Contig 2

Paired-end reads

Best hit unique

Best hits ex-aequo

**Nombre de hits autorisés:**  
un seul hit autorisé, best hit, pour augmenter la spécificité du comptage. Si N best hits, alignement aléatoire sur les un des N best hits.

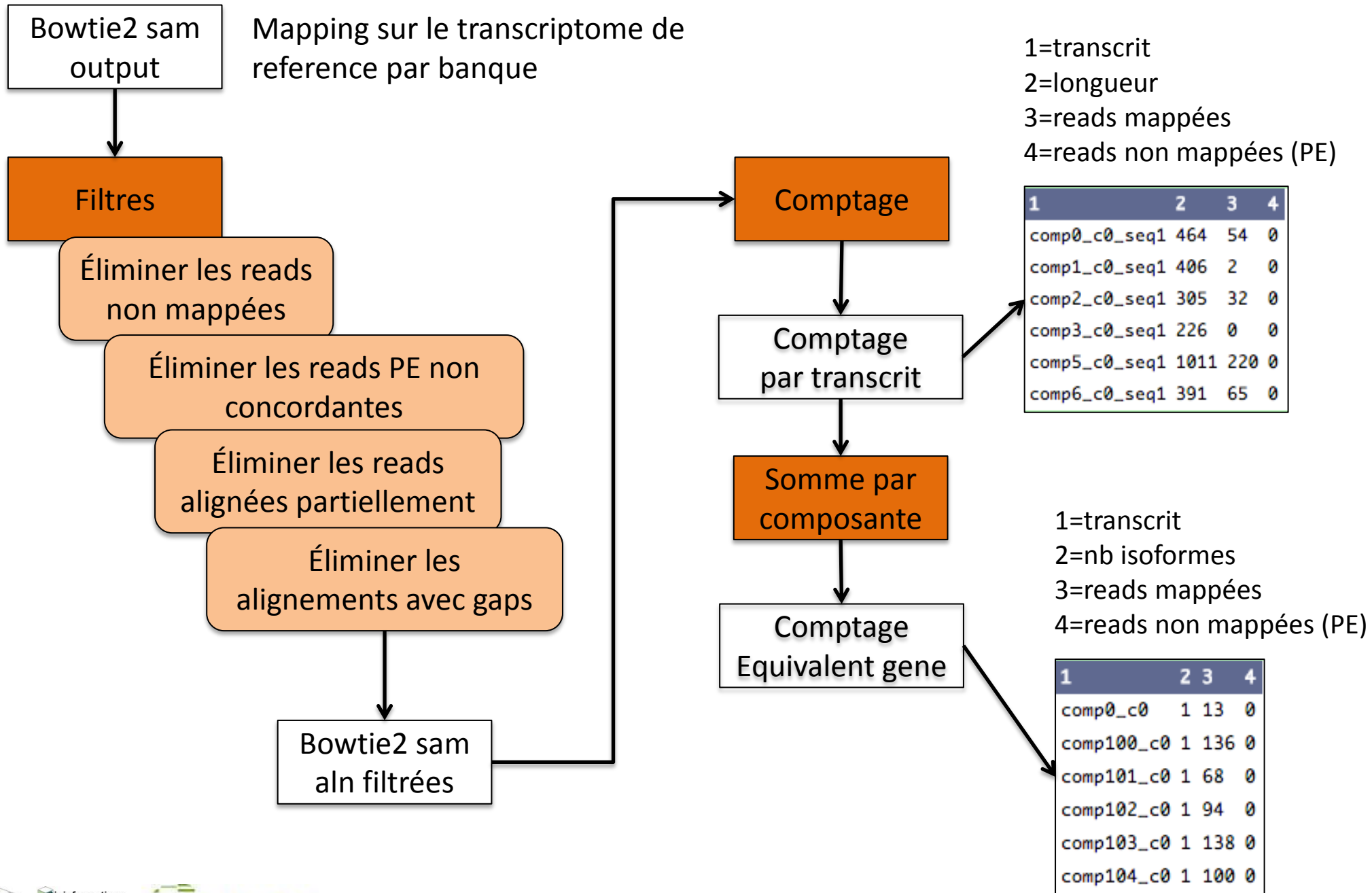
reads

assemblage

Isoforme 1

Isoforme 2

**Attention:** le comptage dans le cas des isoformes d'un même locus (composante) n'est pas représentatif du niveau d'expression de ces isoformes. Il vaut mieux se placer au niveau du locus (composante) en cumulant les comptages, uniques, des isoformes.



1=transcrit

2=nb isoformes

3=reads mappées/banque 1

4=reads non mappées/banque 1(PE)

5=reads mappées/banque 2

6=reads non mappées/banque 2(PE)

1	2	3	4	5	6
comp18_c0	2	172	0	174	0
comp19_c0	2	96	0	78	0
comp28_c0	2	11	0	40	0
comp32_c0	2	52	0	122	0
comp37_c0	3	134	0	368	0
comp44_c0	2	371	0	175	0



Tableau compilant les comptages pour chacune des banques

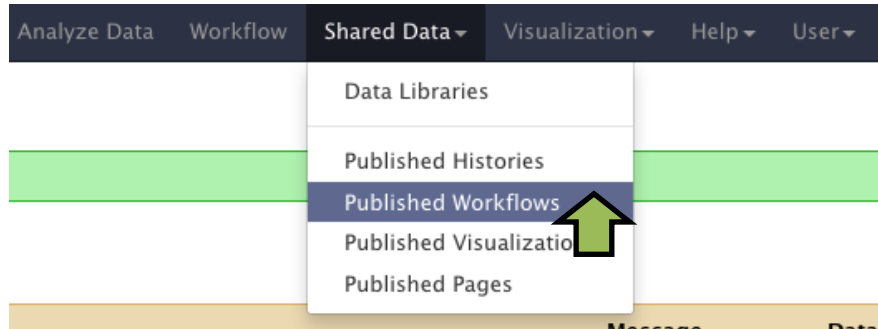
## Suite de l'analyse: Analyse différentielle

- ① DESeq/DESeq2
- ② EdgeR

# III. Mesure de l'expression

## III.1. Chargement du pipeline expression

Aller dans les Published Workflows



Charger le workflow expression

### Published Workflows

search name, annotation, owner, and tag

[Advanced Search](#)

Name	Annotation
RNASeq de novo assembly: 3 expression	RNASeq assembly using trinity. Third step: expression analysis
lncRNAs identification workflow	
RNASeq de novo assembly: 2 assembly	RNASeq assembly using trinity. Second step: trinity
RNASeq de novo assembly: 1 cleaning	RNASeq assembly using trinity. Second step: trinity
RNA-Seq alignment on genome with Tophat and Cufflinks	

# III. Mesure de l'expression

## III.1. Chargement du pipeline expression

Cliquer sur le lien pour visualiser vos pipelines importés



Workflow "RNASeq de novo assembly: 3 expression" has been imported. You can [start using this workflow](#) or [return to the previous page](#).

Sélectionner Run pour utiliser le pipeline assembly

### Your workflows

Name

imported: RNASeq de novo assembly: 3 expression ▾

imported: RNASeq de novo assembly: 2

imported: RNASeq de novo assembly: 1

- Edit
- Run
- Share or Publish
- Download or Export
- Copy
- Rename
- View

### Workflows shared with

No workflows have been shared with you.





# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

Activer la sélection multiple de fichiers pour les étapes 1 et 2

### Running workflow "imported: RNASeq de novo assembly: 3 expression"

Expand All

Collapse

RNASeq assembly using trinity. Third step: expression analysis

#### Step 1: Input dataset

Input Dataset

34: Spombe\_plat Clean right

type to filter

Cliquer pour activer la sélection multiple de fichiers

#### Step 2: Input dataset

Input Dataset

34: Spombe\_plat Clean right

type to filter

#### Step 3: Input dataset

Input Dataset

42: Assembled transcripts

type to filter

# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression



Sélectionner les jeux de données « Clean left » pour l'étape 1 et « Clean right » pour l'étape 2, et le transcriptome assemblé à l'étape 3

### Running workflow "imported: RNASeq de novo assembly: 3 expression"

Expand All Collapse


RNASeq assembly using trinity. Third step: expression analysis

#### Step 1: Input dataset



Input Dataset  

- 17: Spombe\_gd Clean left
- 18: Spombe\_gd Clean right
- 33: Spombe\_plat Clean left
- 34: Spombe\_plat Clean right

type to filter, [enter] to select all




#### Step 2: Input dataset


Input Dataset  

- 17: Spombe\_gd Clean left
- 18: Spombe\_gd Clean right
- 33: Spombe\_plat Clean left
- 34: Spombe\_plat Clean right

type to filter, [enter] to select all




#### Step 3: Input dataset

Input Dataset 

- 42: Assembled transcripts

type to filter



- Vous sélectionnez les 2 banques Spombe\_gd et Spombe\_plat et indiquez quels sont les lots de fichiers (clean left et clean right) pour chaque banque. La sélection des fichiers left et right est distincte car chacun est traité indépendamment.
- Le mapping se fait sur l'assemblage réalisé précédemment: assembled transcripts à sélectionner à l'étape 3.

# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

Paramétrage de l'étape 4 mapping des reads sur le transcriptome assemblé/banque

### Step 4: Bowtie2 (version 0.2)

**Is this library mate-paired?**

Paired-end

**FASTQ file**

Output dataset 'output' from step 1

**FASTQ file**

Output dataset 'output' from step 2

**Minimum insert size for valid paired-end alignments**

0

**Maximum insert size for valid paired-end alignments**

500

**Write unaligned reads to separate file(s)**

False

**Will you select a reference genome from your history or use a built-in index?**

Use one from the history

**Select the reference genome**

Output dataset 'output' from step 3

**Specify the read group for this file?**

No

**Parameter Settings**

Full parameter list

**Type of alignment**

End to end

**Preset option**

Sensitive

La taille maximale de l'insert est fixé à 500 (elle peut être adaptée pour tenir compte de la variabilité des tailles de fragment des banques, pour aller jusque 600-700)

Cliquer sur « run workflow » pour exécuter le pipeline

Run workflow



# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

Le pipeline en cours d'exécution...

Galaxy by GenOuest

Successfully ran workflow "imported: RNASeq de novo assembly: 3 expression". The following datasets have been added to the queue:

**Instance du pipeline pour la banque Spombe\_gd**

- 17: Spombe\_gd Clean left
- 18: Spombe\_gd Clean right
- 42: Assembled transcripts
- 43: Bowtie2 on data 42, data 18, and data 17: aligned reads
- 44: BAM-to-SAM on data 43: converted SAM
- 45: Filter SAM on data 44
- 46: Select on data 45
- 47: Filter on data 46
- 48: SAM-to-BAM on data 42 and data 47: converted BAM
- 49: Count all transcripts
- 50: Select on data 49
- 51: Convert on data 50
- 52: Add column on data 51
- 53: Merge Columns on data 52
- 54: Gene count

**Instance du pipeline pour la banque Spombe\_plat**

- 33: Spombe\_plat Clean left
- 34: Spombe\_plat Clean right
- 42: Assembled transcripts
- 55: Bowtie2 on data 42, data 34, and data 33: aligned reads
- 56: BAM-to-SAM on data 55: converted SAM
- 57: Filter SAM on data 56
- 58: Select on data 57
- 59: Filter on data 58
- 60: SAM-to-BAM on data 42 and data 59: converted BAM
- 61: Count all transcripts
- 62: Select on data 61
- 63: Convert on data 62
- 64: Add column on data 63
- 65: Merge Columns on data 64
- 66: Gene count

History

- 66: Gene count
- 65: Merge Columns on data 64
- 64: Add column on data 63
- 63: Convert on data 62
- 62: Select on data 61
- 61: Count all transcripts
- 60: SAM-to-BAM on data 42 and data 59: converted BAM
- 59: Filter on data 58
- 58: Select on data 57
- 57: Filter SAM on data 56
- 56: BAM-to-SAM on data 55: converted SAM
- 54: Gene count
- 53: Merge Columns on data 52
- 52: Add column on data 51
- 51: Convert on data 50
- 50: Select on data 49
- 49: Count all transcripts
- 48: SAM-to-BAM on data 42 and data 47: converted BAM



# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

Visualisation du fichier de comptage par transcrit pour la banque Spombe\_gd

### Comptage par transcrit

# Legende:

#1=transcrit 2=longueur 3=reads mappées 4=reads non mappées (PE)

comp0_c0_seq1	304	8	0
comp1_c0_seq1	821	26	0
comp2_c0_seq1	236	6	0
comp3_c0_seq1	308	2	0
comp6_c0_seq1	757	9	0
comp7_c0_seq1	282	0	0
comp8_c0_seq1	292	4	0
comp9_c0_seq1	354	6	0
comp10_c0_seq1	242	6	0
comp11_c0_seq1	215	2	0
comp12_c0_seq1	248	0	0
comp13_c0_seq1	267	2	0
comp14_c0_seq1	746	124	0
comp15_c0_seq1	730	88	0
comp16_c0_seq1	990	187	0
comp17_c0_seq1	1693	367	0

History

Unnamed history  
259.4 MB

- 90: Gene count
- 85: Count all transcripts
- 78: Gene count
- 73: Count all transcripts

empty  
format: tabular, database: ?  
no peek

# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

Visualisation du fichier de comptage par gene pour la banque Spombe\_gd

Comptage par  
equivalent gène

# Legende:

1=transcrit 2=nb\_isoformes 3=reads mappées 4=reads non mappées (PE)

comp0_c0	1	8	0
comp1000_c0	1	10	0
comp1001_c0	1	0	0
comp1002_c0	1	0	0
comp1003_c0	1	6	0
comp1004_c0	1	2	0
comp1005_c0	1	2	0
comp1006_c0	1	6	0
comp1007_c0	1	0	0
comp1008_c0	1	6	0
comp1009_c0	1	2	0
comp100_c0	1	1299	0
comp1010_c0	1	14	0
comp1011_c0	1	6	0
comp1012_c0	1	0	0
comp1013_c0	1	0	0
comp1014_c0	1	2	0
comp1015_c0	1	0	0
comp1016_c0	1	8	0
comp1017_c0	1	2	0
comp1018_c0	1	4	0

History

Unnamed history  
259.4 MB

90: Gene count

85: Count all transcripts

78: Gene count  
1,744 lines  
format: tabular, database: ?  
--Group by c8: count[c8] sum[c5]  
sum[c6]

1	2	3	4
comp0_c0	1	8	0
comp1000_c0	1	10	0
comp1001_c0	1	0	0
comp1002_c0	1	0	0
comp1003_c0	1	6	0
comp1004_c0	1	2	0

# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

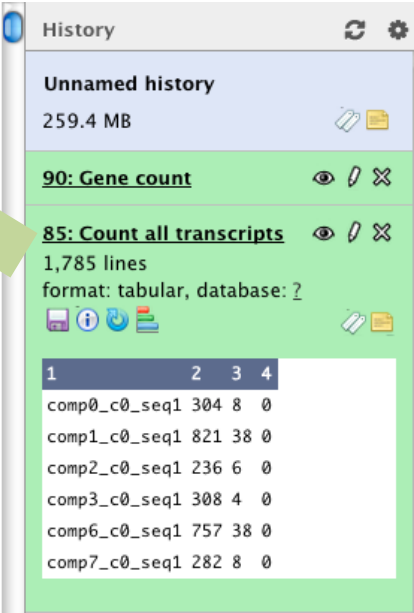
Visualisation du fichier de comptage par transcrit pour la banque Spombe\_plat

### Comptage par transcrit

# Legende:

#1=transcrit 2=longueur 3=reads mappées 4=reads non mappées (PE)

comp0_c0_seq1	304	8	0
comp1_c0_seq1	821	38	0
comp2_c0_seq1	236	6	0
comp3_c0_seq1	308	4	0
comp6_c0_seq1	757	38	0
comp7_c0_seq1	282	8	0
comp8_c0_seq1	292	4	0
comp9_c0_seq1	354	8	0
comp10_c0_seq1	242	2	0
comp11_c0_seq1	215	0	0
comp12_c0_seq1	248	0	0
comp13_c0_seq1	267	0	0
comp14_c0_seq1	746	68	0
comp15_c0_seq1	730	80	0
comp16_c0_seq1	990	279	0
comp17_c0_seq1	1693	14	0
comp18_c0_seq1	550	28	0
comp19_c0_seq1	2043	188	0



History

Unnamed history  
259.4 MB

90: Gene count

85: Count all transcripts  
1,785 lines  
format: tabular, database: ?

1	2	3	4
comp0_c0_seq1	304	8	0
comp1_c0_seq1	821	38	0
comp2_c0_seq1	236	6	0
comp3_c0_seq1	308	4	0
comp6_c0_seq1	757	38	0
comp7_c0_seq1	282	8	0

# III. Mesure de l'expression

## III.2. Paramétrage et exécution du pipeline expression

Visualisation du fichier de comptage par gène pour la banque Spombe\_plat

### Comptage par équivalent gène

# Legende:

#1=composante\_gene 2=nb\_isoformes 3=reads mappées 4=reads non mappées (PE)

comp0_c0	1	8	0
comp1000_c0	1	4	0
comp1001_c0	1	0	0
comp1002_c0	1	2	0
comp1003_c0	1	16	0
comp1004_c0	1	2	0
comp1005_c0	1	4	0
comp1006_c0	1	0	0
comp1007_c0	1	2	0
comp1008_c0	1	16	0
comp1009_c0	1	2	0
comp100_c0	1	295	0
comp1010_c0	1	8	0
comp1011_c0	1	10	0
comp1012_c0	1	0	0
comp1013_c0	1	0	0
comp1014_c0	1	0	0
comp1015_c0	1	2	0

History

Unnamed history  
259.4 MB

90: Gene count  
1,744 lines  
format: tabular, database: ?  
--Group by c8: count[c8] sum[c5]  
sum[c6]

1	2	3	4
comp0_c0	1	8	0
comp1000_c0	1	4	0
comp1001_c0	1	0	0
comp1002_c0	1	2	0
comp1003_c0	1	16	0
comp1004_c0	1	2	0
comp1004_c0	1	2	0



# III. Mesure de l'expression

## III.3. Combiner les fichiers de comptages par gene

Faire une jointure sur les fichiers de comptages par gene

### GALAXY TOOLS > Join, Subtract and Group > Column join

The screenshot displays the Galaxy web interface for the 'Column Join' tool. On the left is a navigation menu with categories like 'Tools', 'PHYLOGENY', 'METADATA MANAGEMENT', and 'GALAXY TOOLS'. Under 'GALAXY TOOLS', the 'Join, Subtract and Group' sub-category is selected, and a green arrow points to the 'Column Join' tool. The main area shows the tool's configuration page for version 1.1.0. It includes fields for 'Choose the first file for the join' (set to '90: Gene count'), 'Use this column and columns to left the 'hinge' (set to 'c1'), 'Include these column:' (a list containing 'c1', 'c2', 'c3', 'c4'), 'Fill empty columns:' (set to 'No'), and 'Choose the second file for the join:' (set to '90: Gene count'). There is an 'Execute' button at the bottom. Below the configuration is a 'What it does' section explaining the tool's function. On the right, a 'History' panel shows a list of previous tool runs, with the current run highlighted in green.

**Column Join (version 1.1.0)**

Choose the first file for the join:  
90: Gene count

Use this column and columns to left the 'hinge' (matching data for each join):  
c1

All columns to left of selected column (plus selected column) will be used. Select 2 for pileup

Include these column:  
c1  
c2  
c3  
c4

Multi-select list – hold the appropriate key while clicking to select multiple columns

Fill empty columns:  
No

Choose the second file for the join:  
90: Gene count

Additional Inputs  
Add new Additional Input

Execute

**What it does**

This tool allows you to join several files with the same column structure into one file, removing certain columns if necessary. The user needs to select a 'hinge', which is the number of left-most columns to match on. They also need to select the columns to include in the join, which should include the hinge columns, too.

Note that the files are expected to have the same number of columns. If for some reason the join column is missing (this only applies to the last column(s)), the tool attempts to handle this situation by inserting an empty item (or the appropriate filler) for that column on that row. This could lead to the situation where a row has a hinge but entirely empty or filled columns, if the hinge exists in at least one file but every file that has it is missing the join column. Also, note that the tool does not distinguish between a file missing the hinge altogether and a file having the hinge but missing the column (in both cases the column would be empty or filled). There is an example of this below.

**History**

Unnamed history  
259.4 MB

- 90: Gene count
- 85: Count all transcripts
- 78: Gene count
- 73: Count all transcripts
- 42: Assembled transcripts
- 34: Spombe\_plat Clean right
- 33: Spombe\_plat Clean left
- 18: Spombe\_gd Clean right
- 17: Spombe\_gd Clean left
- 4: SRS167022\_Spombe\_plat\_SRR097918\_75K.right.fastq
- 3: SRS167022\_Spombe\_plat\_SRR097918\_75K.left.fastq
- 2: SRS167021\_Spombe\_gd\_SRR097901\_75k.right.fastq
- 1: SRS167021\_Spombe\_gd\_SRR097901\_75k.left.fastq



# III. Mesure de l'expression

## III.3. Combiner les fichiers de comptages par gene

Sélectionner les paramètres de la jointure entre les 2 fichiers de comptages

Column Join (version 1.1.0)

Choose the first file for the join:  
78: Gene count

Use this column and columns to left the 'hinge' (matching data for each join):  
c2

All columns to left of selected column (plus selected column) will be used. Select 2 for pileup

Include these column:  
c1  
c2  
c3  
c4

Multi-select list – hold the appropriate key while clicking to select multiple columns

Fill empty columns:  
Yes

Fill Columns by:  
Single fill value

Fill value:  
.

Choose the second file for the join:  
90: Gene count

Additional Inputs  
Add new Additional Input

Execute

- ① Sélectionner le fichier de comptage par gene pour la banque Spombe\_gd (dataset 78)
- ② Sélectionner les colonnes pour faire la jointure (jusque c2)
- ③ Sélectionner les colonnes à inclure dans le fichier final (c1 à c3)
- ④ Remplir les colonnes vides avec une seule valeur (caractère .)
- ⑤ Sélectionner le fichier de comptage par gene pour la banque Spombe\_plat (dataset 90)

Cliquer sur Execute

# III. Mesure de l'expression

## III.3. Combiner les fichiers de comptages par gene

Le pipeline en cours d'exécution...

The screenshot shows the Galaxy by GenOuest web interface. The browser address bar displays the URL: `bipaa-galaxy.genouest.org/root?workflow_id=1364d3b9ef471763`. The interface includes a navigation menu with options like 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A central notification box with a green checkmark states: 'The following job has been successfully added to the queue: 91: Column Join on data 90 and data 78'. Below this, it explains that users can check the status of queued jobs in the History pane. On the right, the History pane shows a list of jobs, including '91: Column Join on data 90 and data 78' and several 'Spombe plat Clean' jobs. The left sidebar contains a 'Tools' menu with categories like 'PHYLOGENY', 'Text Manipulation', 'Next Generation Quality', 'Phylogenetic Tree', 'Alignment', 'NCBI Blast', 'Blat', 'Phylostatistics', 'METADATA MANAGEMENT', 'GALAXY TOOLS', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Wavelet Analysis', 'Graph/Display Data', 'Regional Variation', and 'Multiple regression'.



# III. Mesure de l'expression

## III.3. Combiner les fichiers de comptages par gène

Visualisation du fichier de comptage par gène combiné pour les banques Spombe\_gd/plat

### Comptage par équivalent gène

comp0_c0	1	8	8
comp1_c0	1	26	38
comp2_c0	1	6	6
comp3_c0	1	2	4
comp6_c0	1	9	38
comp7_c0	1	0	8
comp8_c0	1	4	4
comp9_c0	1	6	8
comp10_c0	1	6	2
comp11_c0	1	2	0
comp12_c0	1	0	0
comp13_c0	1	2	0
comp14_c0	1	124	68
comp15_c0	1	88	80
comp16_c0	1	187	279
comp17_c0	1	367	14
comp18_c0	1	91	28

# Legende:

# banque 1=Spmobe\_gd; banque 2=Spombe\_plat

#1=composante\_gene 2=nb\_isoformes 3=comptage banque 1 4=comptage banque 2



Suite de l'analyse: Analyse différentielle

- ① DESeq/DESeq2
- ② EdgeR

History

Unnamed history  
259.4 MB

91: Column Join on data 90 and data 78  
1,744 lines  
format: tabular, database: ?

1	2	3	4
comp0_c0	1	8	8
comp1_c0	1	26	38
comp2_c0	1	6	6
comp3_c0	1	2	4
comp6_c0	1	9	38
comp7_c0	1	0	8



# MESURE DE L'EXPRESSION A PARTIR DE DONNÉES RNASEQ

**Responsable et intervenant principal: Erika Sallet**

**Expert: Ludovic Legrand**

**Relecteur: Sébastien Carrere**



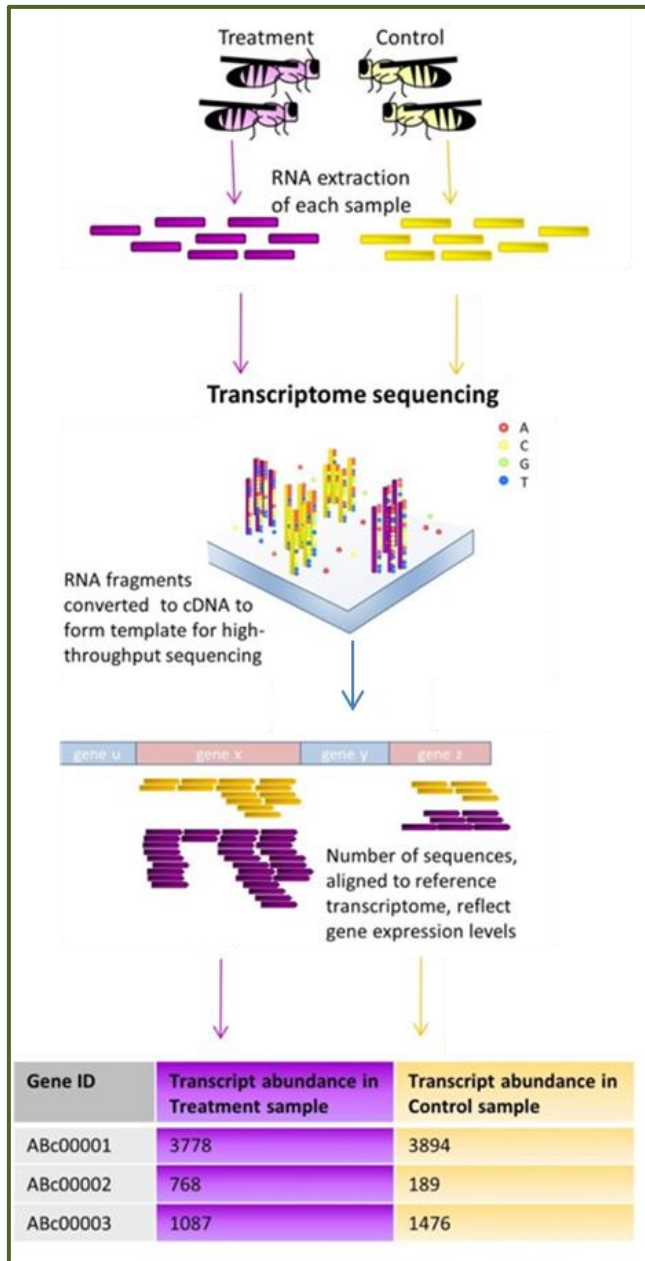
- Objectif : Mesure de l'expression

- Périmètre du pipeline :

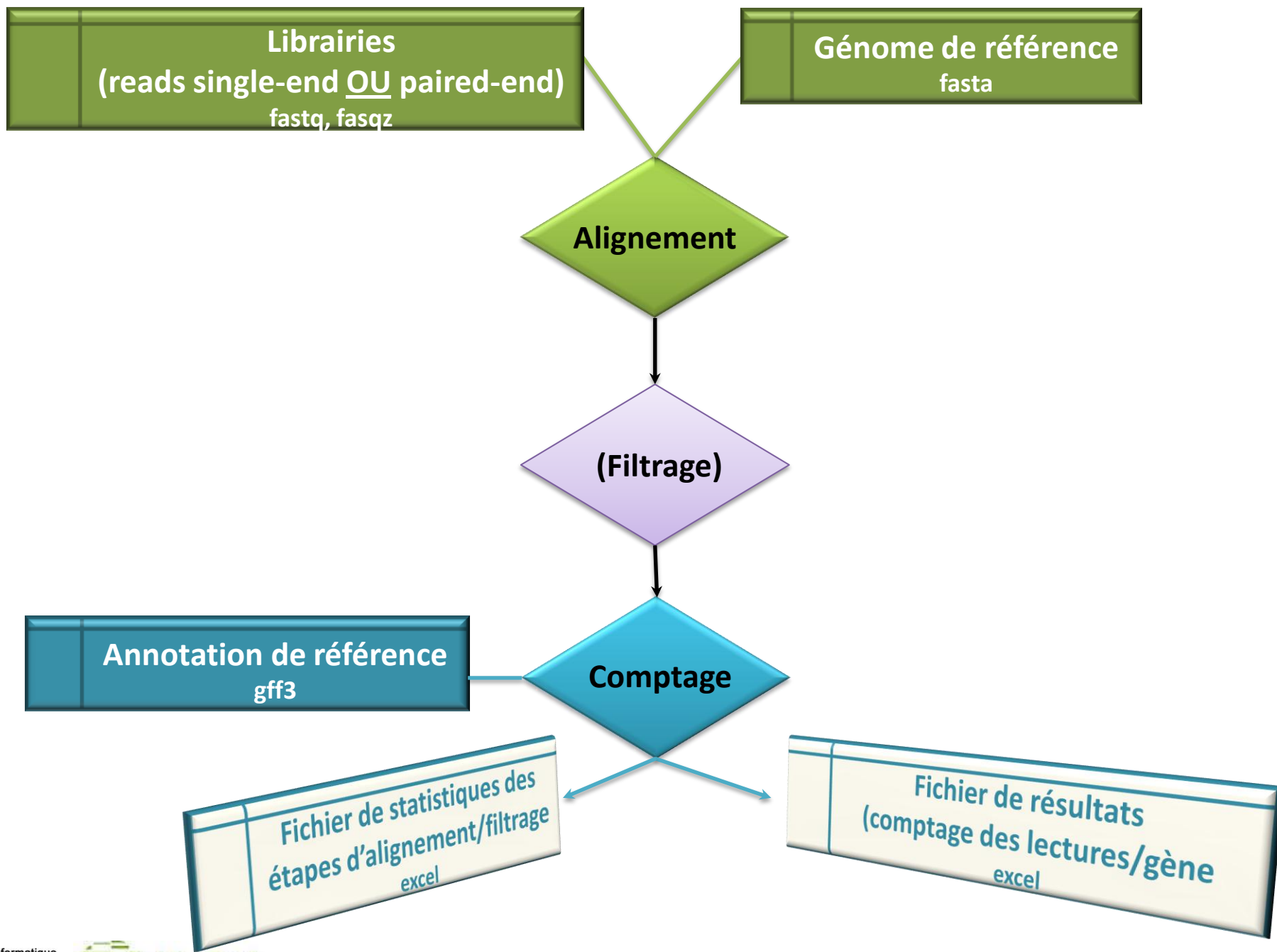
- Comptage des lectures sur des gènes définis dans le fichier d'annotation (pas de découverte de nouveaux transcrits)
- Pro /eucaryotes

- Fichiers requis :

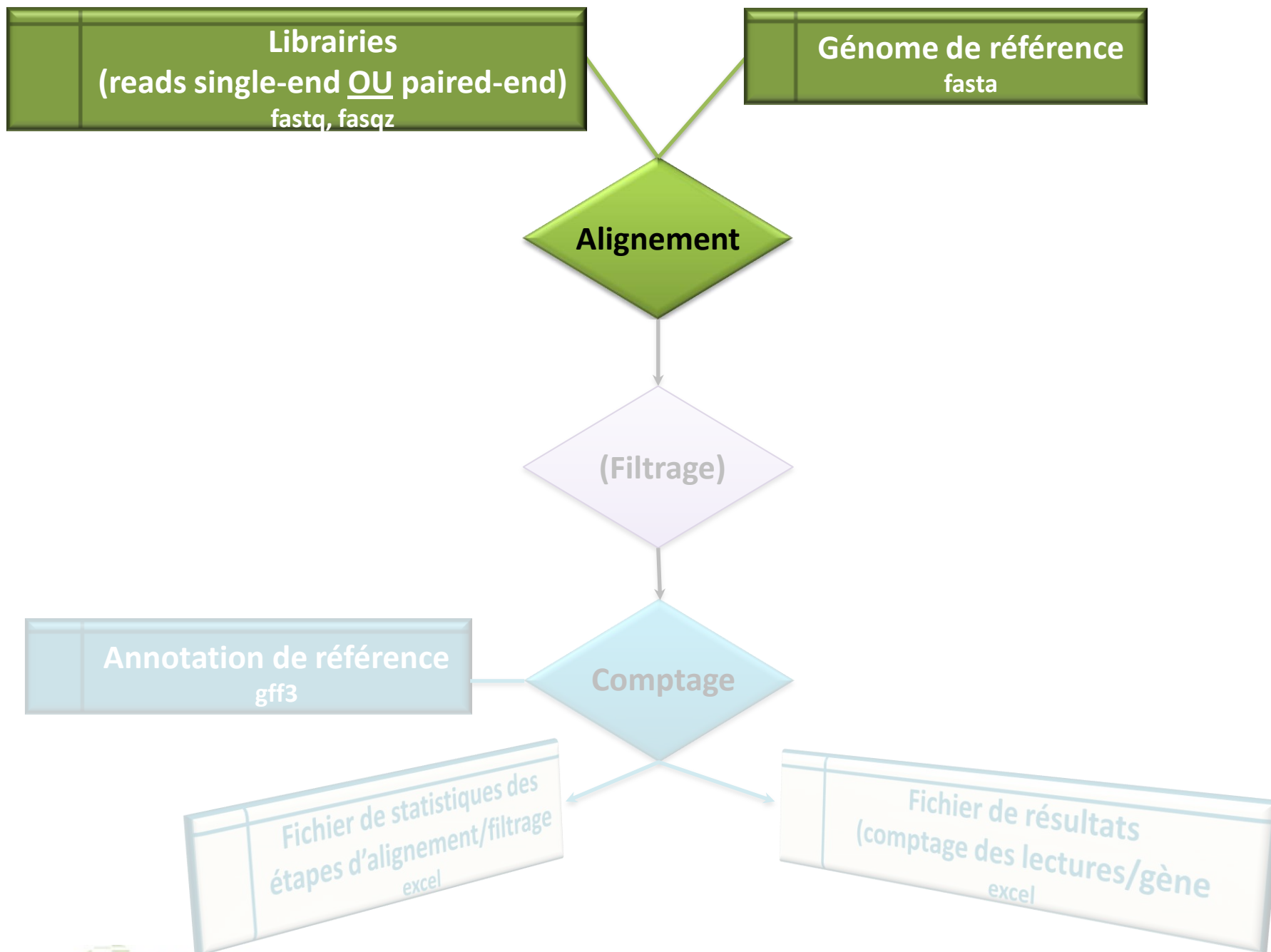
- Bibliothèques RNASeq paired-end ou single-end (fastq ou fastqz)
- Génome ou transcriptome (fasta)
- Annotation (GFF3)



Adapté de : Beyond the Gene List: Exploring Transcriptomics Data in Search for Gene Function, Trait Mechanisms and Genetic Architecture - By Bregje Wertheim

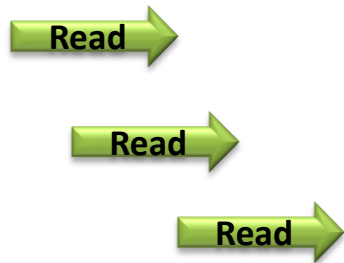




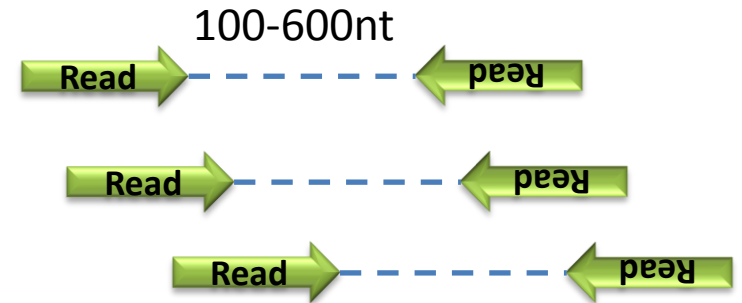




## Contig/Chromosome



Librairie Single-end



Librairie Paired-end

## Paramètres influant sur l'alignement :

- longueur minimale du hit
- nombre maximal de mismatches
- distance maximale entre les paires
- petit ARN non codant (small RNA)

## Longueur minimale du hit

- valeur par défaut : 18 nt
- compromis entre bruit et information
- perte des exons plus petits que ce paramètre

Exemple      read: 25nt      lmin: 18nt

Exon  
...TCAGAAGCAGCGGTGAGATCCTGGCTGTTCTGAAAGTGAGACGAGCGGATTTCTGCTG...  
Exon  
AGAAGCAGCGGTGAGATC~~GATTTCC~~  
  
CAGCGGTGAGATC~~GATTTCTGCTG~~ ← Alignement < 18nt

### Nombre maximal de mismatches

- valeur par défaut : 0
- erreurs de séquençage (~1%)
- variabilité avec la référence
- hétérozygotie

### Exemple: lecture de 100nt paired-end

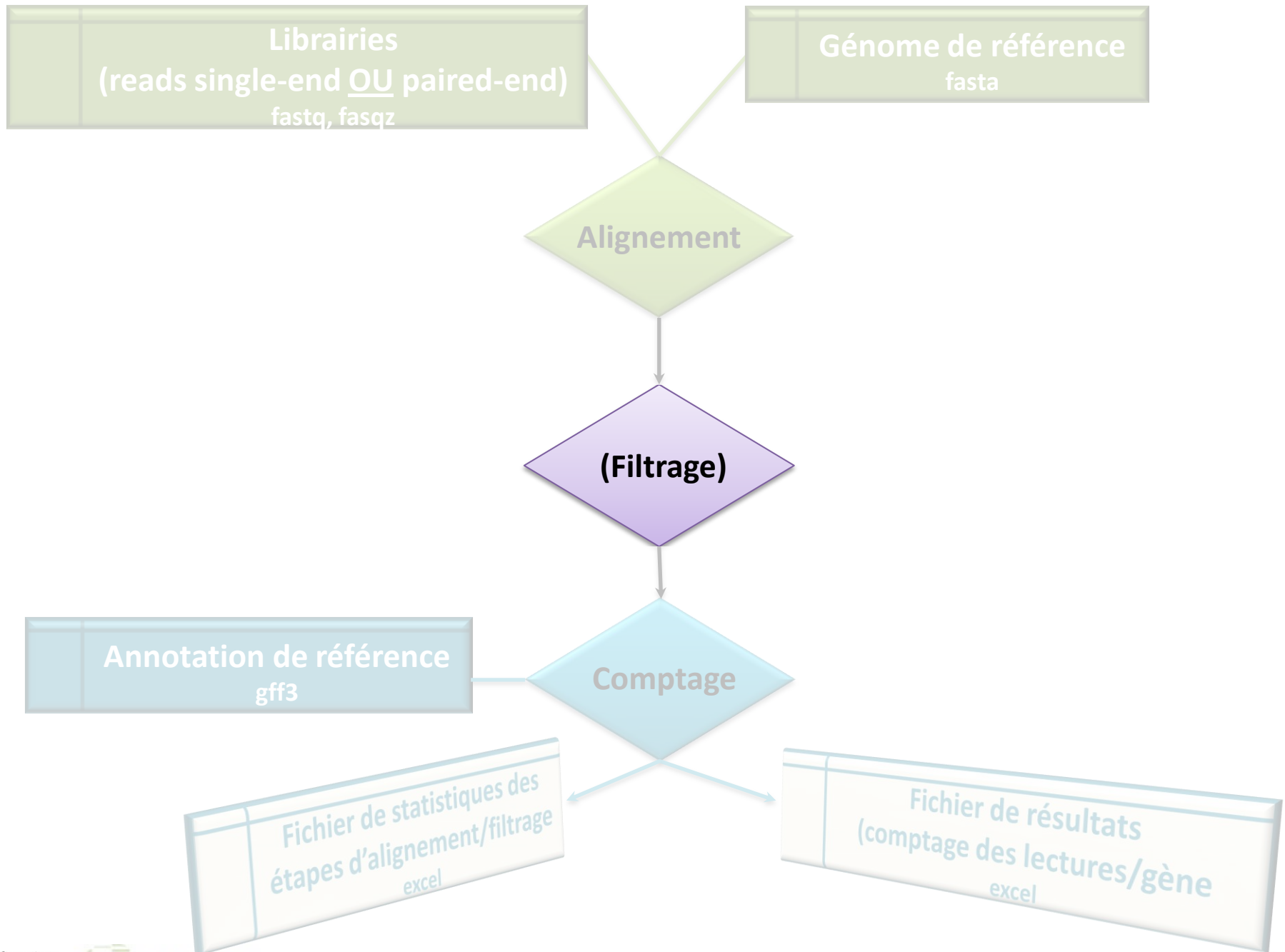
- 5 mismatches autorisés
  - prise en compte des variations alléliques et des erreurs
  - perte de spécificité d'alignement sur les lectures compensée par l'alignement de la paire

### Distance maximale entre les paires

- cohérence de l'alignement de 2 reads d'une paire
- généralement entre 100 et 600 nt

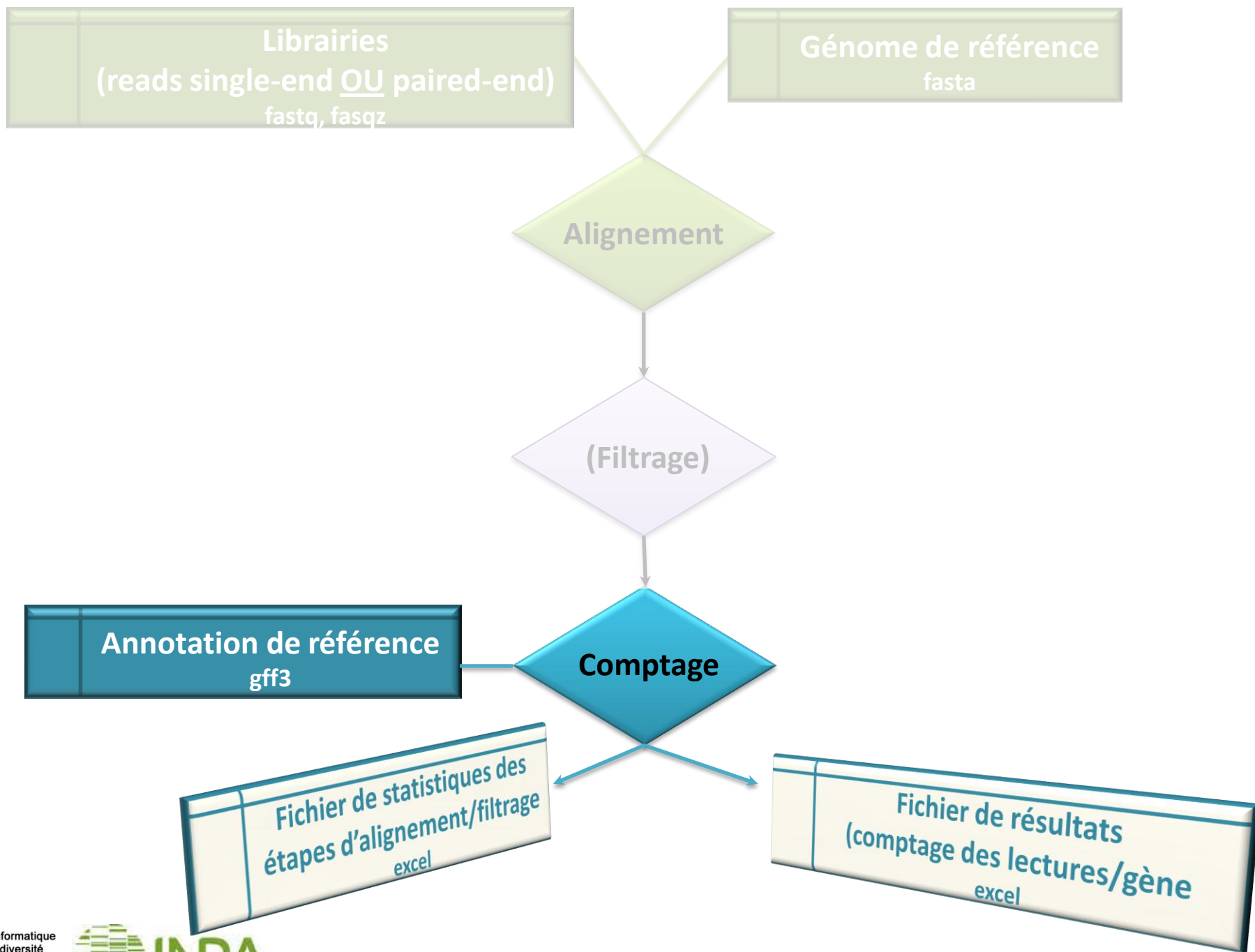
### Petit ARN non codant (small RNA)

- lecture entre 20 et 50 nt mais hits entre 18 et 30 nt
- 0 mismatch autorisé
  - pas d'informations complémentaires avec la paire
  - limite les alignements dus au hasard
- adaptation des seuils pour les petits alignements



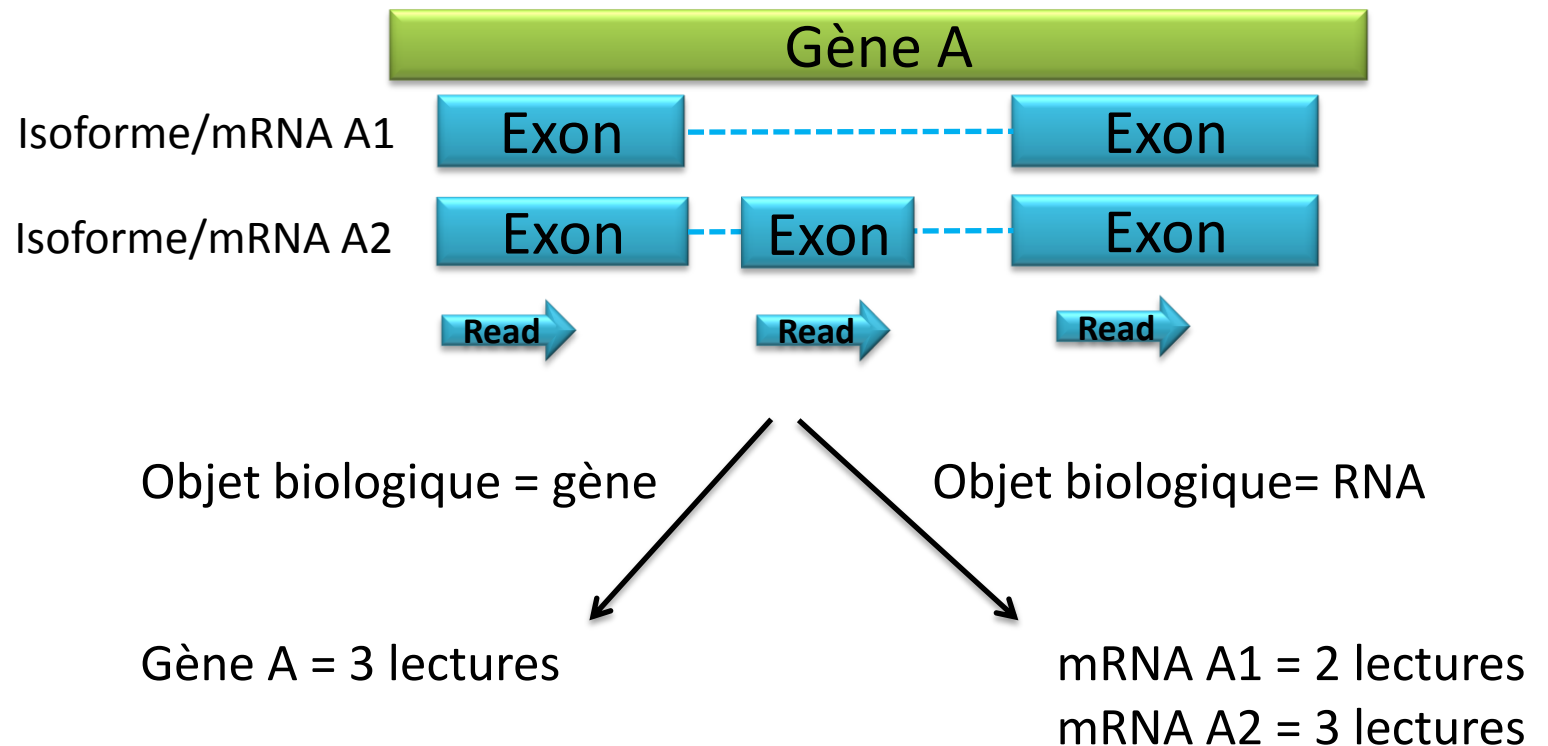
## Filtrage des lectures non spécifiques (ambiguës)

- alignement identique de score maximum sur plusieurs positions
  - duplication de gènes, transposons, gènes conservés
- ⇒ on préfère perdre de l'information en écartant les hits ambigus
- ⇒ garder les hits ambigus c'est additionner le niveau d'expression de N objets biologiques très probablement régulés différemment



# Comptage du nombre de lectures par objet biologique

*Exemple d'un gène A  
avec épissage alternatif*





### Comptage du nombre de lectures par gène

- moins sensible à la qualité de l'annotation
- ne tient pas compte de l'épissage alternatif

### Comptage du nombre de lectures par RNA

- sensible à l'annotation des exons
- tient compte de l'épissage alternatif

## Couverture min de la lecture sur l'objet biologique

- Valeur par défaut = 1 (100%)
  - suppose une annotation de bonne qualité

Gène/exon

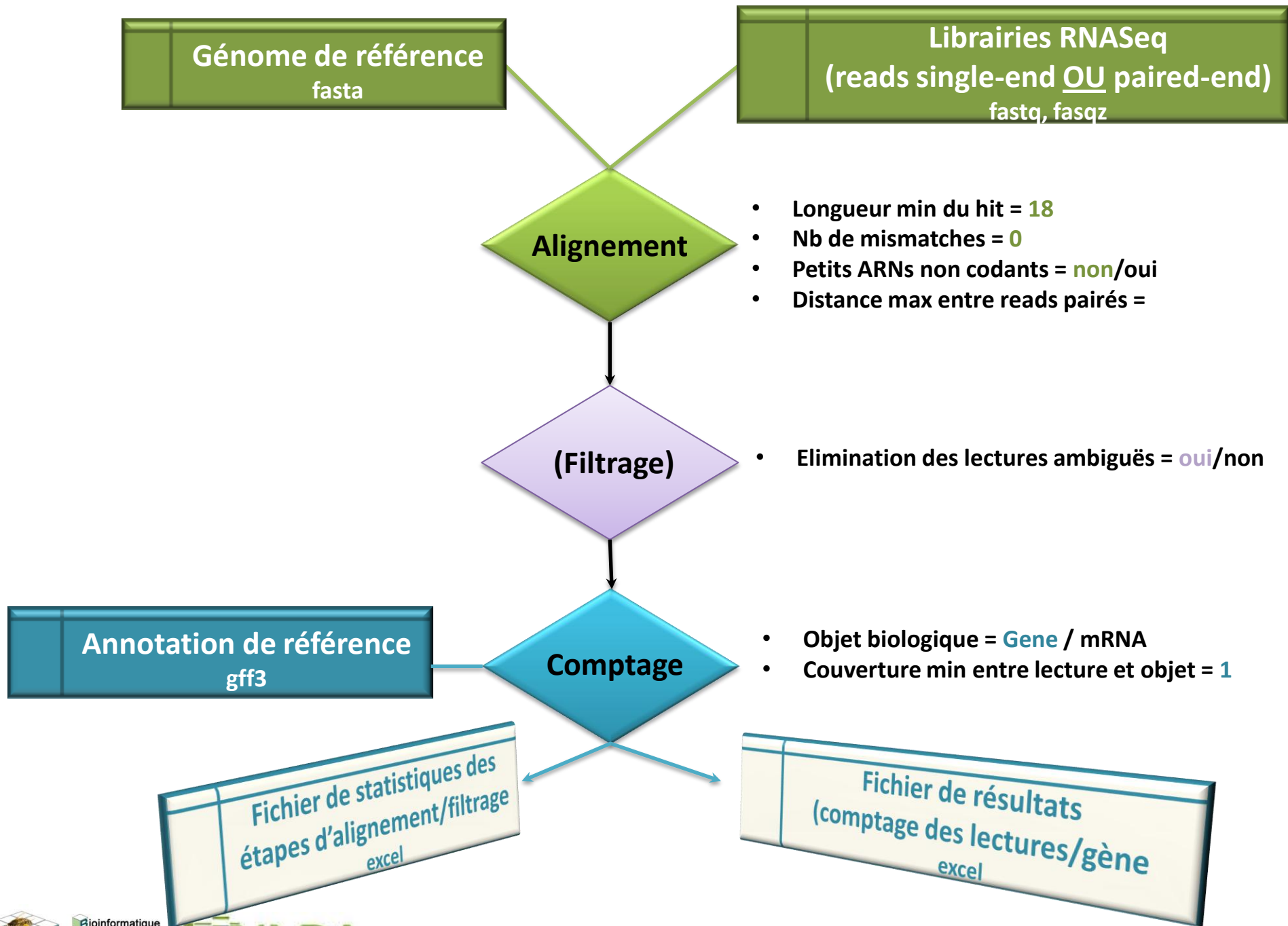


- diminuer la couverture
  - autoriser les hits partiels sur les objets
  - annotation de faible ou moyenne qualité

Gène/exon



# Fonctionnement : résumé étapes + paramètres modifiables



## Expression measure (version 1)

### Library type:

Paired-end ▾

### Paired-end libraries

#### Paired-end library 1

##### Read file 1:

11: S.bbric-RbmSmall-GGK36.ope.2.fastq.gz ▾  
fastq,fastq.gz

##### Read file 2:

11: S.bbric-RbmSmall-GGK36.ope.2.fastq.gz ▾  
fastq,fastq.gz

Add new Paired-end library

### Maximal distance between paired reads (nt):

### Reference genome file (fasta):

19: Genomic sequence from Annotation on bacterial genome on data 8, data 9, and others ▾

### Genome annotation file (GFF3):

12: Annotation on bacterial genome on data 8, data 9, and others ▾

### Expression reported for:

Gene ▾

Select biological objects (gff3 type) for which the expression is reported

### Small inserts analysis:

Change mapping parameters

### Minimal hit length:

### Maximum number of mismatches:

### Advanced parameters:

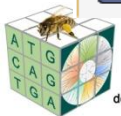
### Use no specific mapped reads:

Considering all mapped read/pairs (default: considering unambiguously mapped reads/pairs only)

### Minimum overlap:

Minimum overlap required as a fraction of the mapped read (intersecBed parameter -f)

Execute



# Interface Galaxy

Expression measure (version 1)

Library type:

Paired-end

Paired-end libraries

Paired-end library 1

Read file 1:

11: S.bbric-RbmSmall-GGK36.ope.2.fastq.gz  
fastq,fastq.gz

Read file 2:

11: S.bbric-RbmSmall-GGK36.ope.2.fastq.gz  
fastq,fastq.gz

Add new Paired-end library

Maximal distance between paired reads (nt):

- Distance max entre reads pairés

Reference genome file (fasta):

19: Genomic sequence from Annotation on bacterial genome on data 8, data 9, and others

Genome annotation file (GFF3):

12: Annotation on bacterial genome on data 8, data 9, and others

Expression reported for:

Gene

- Objet = Gene / RNA

Select biological objects (gff3 type) for which the expression is reported

Small inserts analysis:

- Petits ARNs non codants = non/oui

Change mapping parameters

Minimal hit length:

18

- Longueur min du hit = 18

Maximum number of mismatches:

0

- Nb de mismatches = 0

Advanced parameters:

Use no specific mapped reads:

- Elimination des lectures ambiguës = oui/non

Considering all mapped read/pairs (default: considering unambiguously mapped reads/pairs only)

Minimum overlap:

1.0

- Couverture min entre lecture et objet = 1 (100%)

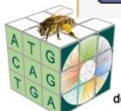
Minimum overlap required as a fraction of the mapped read (intersecBed parameter -f)

Execute

Librairies  
(reads single-end OU paired-end)  
fastq, fastqz

Génome de référence  
fasta

Annotation de référence  
gff3



Nombre d'alignements spécifiques sur génome

Nombre d'alignements sur génome

Nombre de lectures brutes

Nombre de lectures alignées sur le génome

Nombre d'alignements sur objet biologique

alignements sur objet biologique / alignements spécifiques sur génome

lib	specific_hits	mapping_hits	raw_reads/pairs_count	mapped_reads/pairs_count	feature_overlapping_hits	feature_overlapping_hits/specific_hits_percent
S.bbric_Rb mLong_GGK 21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_Rb mSmall_GG K36.ope	24405	24411	64272	24407	20674	84.71

## Contig/Chromosome



1181 hits totaux  
1179 lectures qui s'alignent

Nombre d'alignements spécifiques sur génome

Nombre d'alignements sur génome

Nombre de lectures brutes

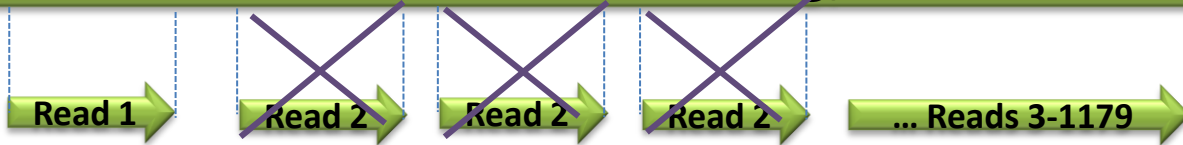
Nombre de lectures alignées sur le génome

Nombre d'alignements sur objet biologique

alignements sur objet biologique / alignements spécifiques sur génome

lib	specific_hits	mapping_hits	raw_reads/pairs_count	mapped_reads/pairs_count	feature_overlapping_hits	feature_overlapping_hits/specific_hits_percent
S.bbric_Rb mLong_GGK 21.ope	1178	1181	9697	1179	1153	97.88
S.bbric_Rb mSmall_GG K36.ope	24405	24411	64272	24407	20674	84.71

## Contig/Chromosome



1178 hits spécifiques =  
1178 lectures avec alignement spécifique

Gene/RNA ID

Contig Id

Positionnement du gène/RNA sur le contig

Début

Fin

Brin

Taille du Gène/RNA

Type (Gène/RNA)

Comptage brut (Lectures/Objet)

Comptage normalisé (RPKM)

id	0_seqid	1_start	2_end	3_strand	4_length	5_type	6_Note	S.bbric_RbmLong_GGK21.op e-count	S.bbric_RbmLong_GGK21.op e-rpkm	S.bbric_RbmSmall_GGK36.op e-count	S.bbric_RbmSmall_GGK36.op e-rpkm
SBBRIC1.1	SBBRIC1	384	685	-	302	gene		5	14054.58	2	260.41
SBBRIC1.10	SBBRIC1	9162	11163	+	2002	gene		14	5936.34	107	2101.63
SBBRIC1.100	SBBRIC1	90298	90594	+	297	gene		0	0.00	9	1191.58



# Comptage

- nombre de lectures alignées par objet
- utilisé pour les analyses d'expression différentielle après normalisation

# Normalisation RPKM

Nat Methods. 2008 Jul;5(7):621-8. doi: 10.1038/nmeth.1226. Epub 2008 May 30.  
**Mapping and quantifying mammalian transcriptomes by RNA-Seq.**  
Mortazavi A<sup>1</sup>, Williams BA, McCue K, Schaeffer L, Wold B.

- **Read Per Kilobase per Million** mapped reads

$$\text{RPKM (X)} = \frac{\text{Nb de lectures/gène (comptage brut)}}{\text{million lectures alignées} \times \text{taille objet biologique(kb)}}$$

Normalisation entre librairies

Normalisation entre objets

## A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies\*, Andrea Rau\*, Julie Aubert\*, Christelle Hennequet-Antier\*, Marine Jeanmougin\*, Nicolas Servant\*, Céline Keime\*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaeffer, Stéphane Le Crom\*, Mickaël Guedj\*, Florence Jaffrézic\* and on behalf of The French StatOmique Consortium

### Key points

- Normalization of RNA-seq data in the context of differential analysis is essential in order to account for the presence of systematic variation between samples as well as differences in library composition.
- The Total Count and RPKM normalization methods, both of which are still widely in use, are ineffective and should be definitively abandoned in the context of differential analysis.
- Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions, both of which are typical of real RNA-seq data.

Table 3: Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	–	+	++	–
RPKM	–	+	+	–	–

A '–' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

- Normalisation et analyses statistiques pour identifier les gènes différentiellement exprimés avec R
  - DESeq ou DESeq2 (bioconductor)
  - TMM (edgeR)

- Exemple d'utilisation basique de DESeq (bioconductor)
  - Pour normaliser les comptages et identifier des gènes différentiellement exprimés

### Dans une console R :

```
# la première fois, installer la librairie DESeq
source("http://www.bioconductor.org/biocLite.R")
biocLite("DESeq")

# appel de la librairie DESeq
library("DESeq")

# lecture du fichier de comptages bruts (c'est un fichier avec en-têtes ; colonnes=noms des librairies ; lignes=noms des gènes ; valeurs = comptages bruts)
alldata <- read.table("AllCounts.txt", header = TRUE, sep="\t", row.names=1)

# définir le design (ici, 7 librairies : 3 "WT" et 4 "Mut ")
conditions <- factor(c("WT", "WT", "WT", "Mut", "Mut", "Mut", "Mut"))

# création d'un objet « DESeq »
cds <- newCountDataSet(alldata, conditions)

# suppression des gènes pour lesquels la moyenne des comptages est inférieure ou égale à 5
cds <- cds[rowMeans(counts(cds))>5,]

# normalisation par DESeq
cds <- estimateSizeFactors(cds)
cds <- estimateDispersions(cds)

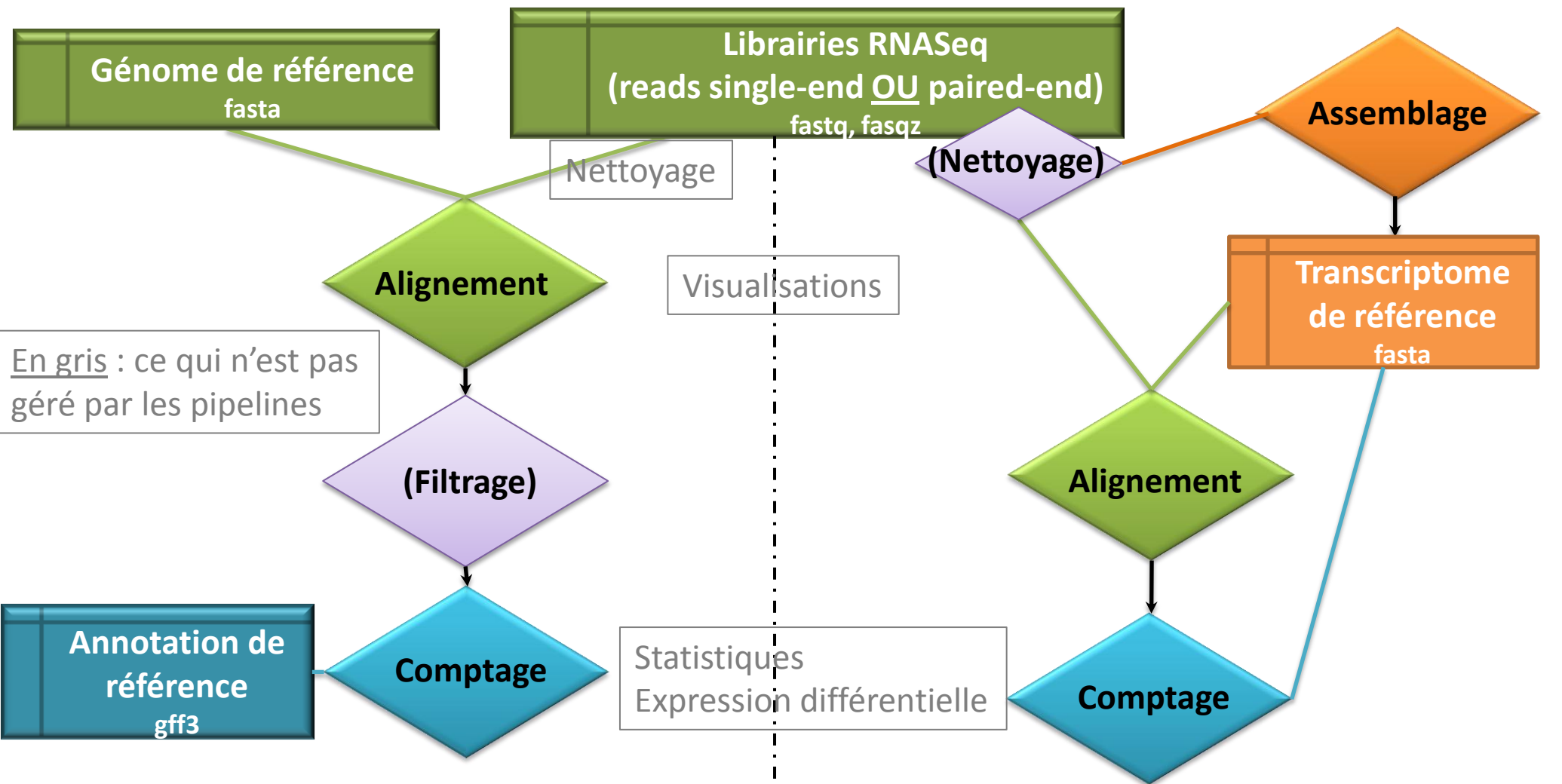
# test différentiel entre WT et Mut
res <- nbinomTest(cds, "WT", "Mut")

# écriture des résultats complets dans une table lisible par excel (pour trier sous excel les gènes avec adjpvalue<0,05 et log2(fold-change>1))
write.csv(res, file="WT_Mut.csv")
```

### Attention :

ceci est un exemple basique, et ne vous dispense pas :  
de suivre une formation adéquate  
de demander conseils aux statisticiens  
de chercher des tutos sur le web  
de regarder du côté de RStudio  
...

# Résumé du périmètre du pipeline



**Mesure de l'expression avec génome annoté**  
Application principale : Expression différentielle

**Mesure de l'expression avec assemblage *de novo***  
Applications : Expression différentielle, découverte de gènes, d'isoformes, annotation structurale.



# DÉTECTION DE TRANSFERTS HORIZONTAUX (DONNÉES: PROTEOME)

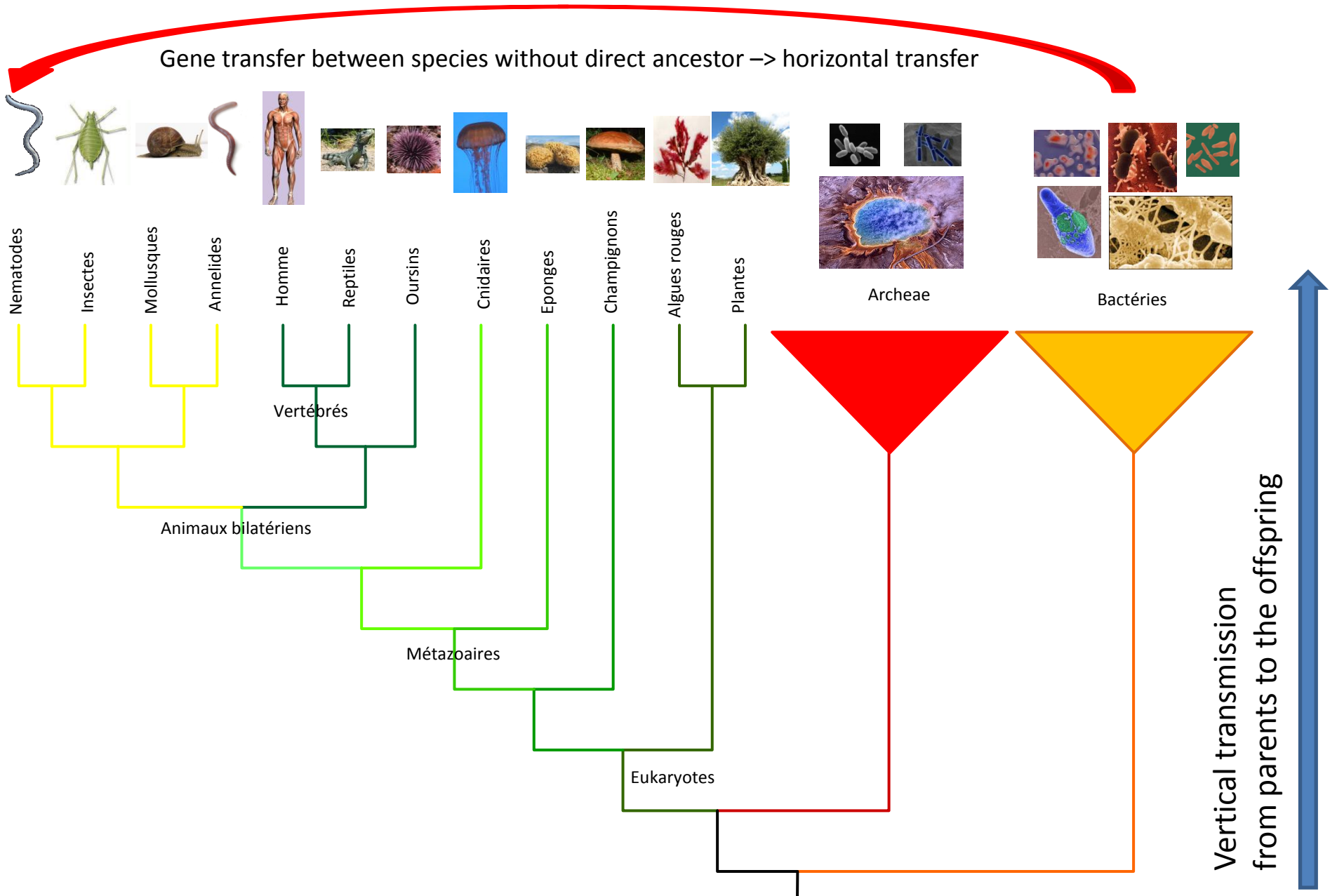
**Responsable et intervenant principal: Ludovic Legrand**

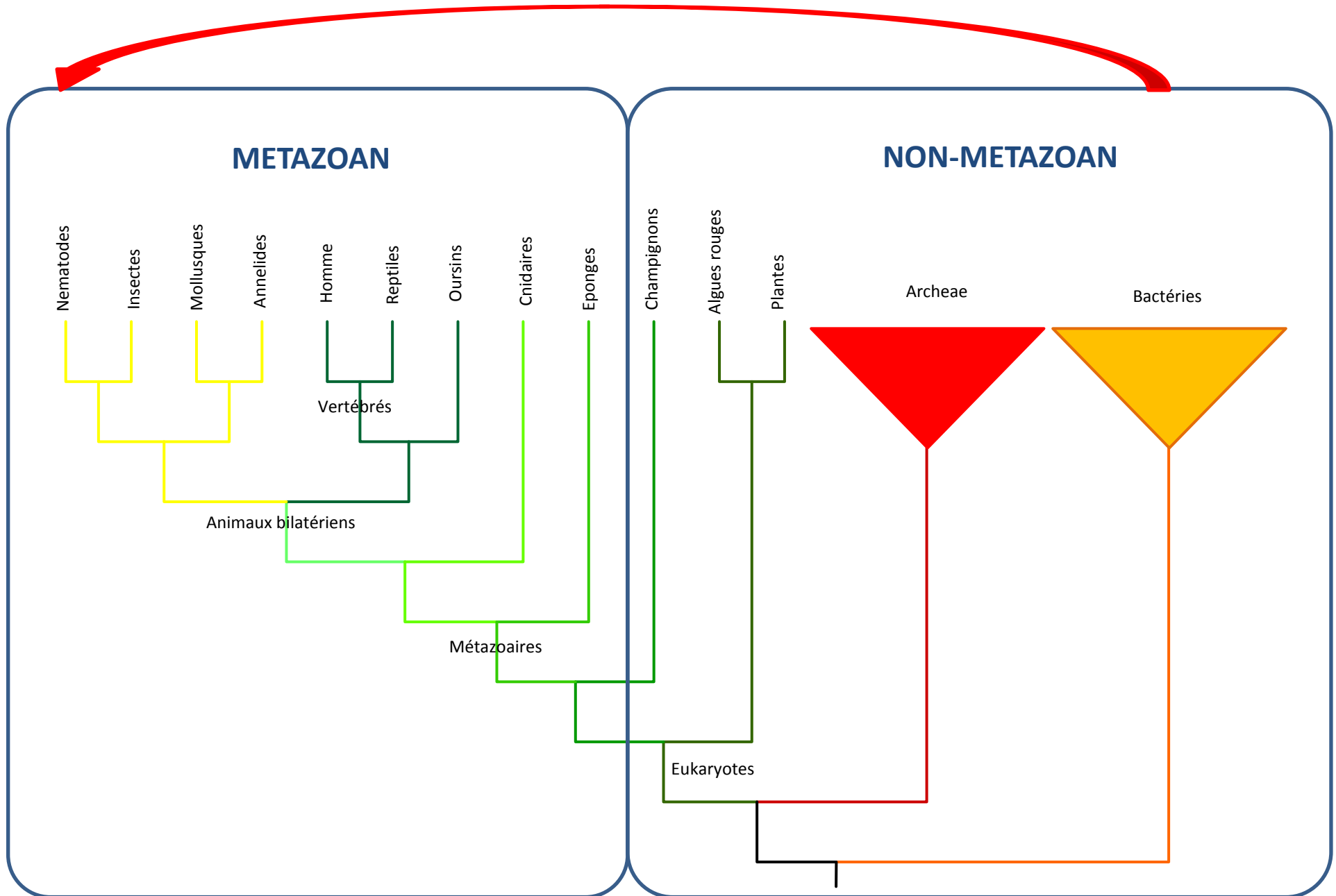
**Expert: Corinne Rancurel**

**Relecteurs : Sébastien Carrere**



# ACQUISITION DE GENES VIA DES TRANSFERTS HORIZONTALAUX







**Identifier à partir d'un protéome de métazoaire, les gènes  
issus d'un transfert horizontal provenant d'un organisme  
non métazoaire**

Pour détecter un transfert horizontal  
d'un **non-métazoaire** vers un **métazoaire**,  
Gladyshev et al, ont mis au point un modèle mathématique basé sur un  
différentiel d'e-value de blast.

Un calcul d'**Alien Index (AI)** résulte de cette étude

**Alien index (AI)**

=

$$\log( (\text{Best E-value for } \mathbf{Metazoa}) + e^{-200} ) - \log( (\text{Best E-value for } \mathbf{Non-Metazoa}) + e^{-200} )$$

**Massive Horizontal gene Transfer in Bdelloid Rotifers , Gladyshev et al. Science 2008**



Alien index (AI)

=

$$\log( (\text{Best E-value for Metazoa}) + e^{-200} ) - \log( (\text{Best E-value for Non-Metazoa}) + e^{-200} )$$

$AI > 0 \Leftrightarrow$  e-value non-métazoaire < e-value métazoaire

➡ anormal pour un gène classique de métazoaire

En utilisant **Blast**, il nous est possible de récupérer

l'e-value du meilleur hit Métazoaire

l'e-value du meilleur hit Non Métazoaire

Massive Horizontal gene Transfer in Bdelloid Rotifers , Gladyshev et al. Science 2008

## Alien index (AI)

=

$$\log( (\text{Best E-value for Metazoa}) + e^{-200} ) - \log( (\text{Best E-value for Non-Metazoa}) + e^{-200} )$$

# BLASTP 2.2.26+

# Query: ANUcomp28394\_c0\_seq1

# Database: nr

# Fields: query id, subject id, % identity, .... evalue, bit score, subject ids

# 250 hits found

ANUcomp28394_c0_seq1	gi 522006232	40.98	....	1e-37	144
.....					
ANUcomp28394_c0_seq1	gi 110671508	33.52	....	2e-24	105

**Massive Horizontal gene Transfer in Bdelloid Rotifers , Gladyshev et al. Science 2008**

## Alien index (AI)

=

$$\log(\text{Best E-value for Metazoa}) + e^{-200} - \log(\text{Best E-value for Non-Metazoa}) + e^{-200}$$

LOCUS WP\_020517503 377 aa linear BCT 09-JUL-2013  
 DEFINITION hypothetical protein [Actinoplanes globisporus].  
 ACCESSION WP\_020517503  
 VERSION WP\_020517503.1 **GI:522006232**  
 KEYWORDS RefSeq.  
 SOURCE Actinoplanes globisporus  
 ORGANISM Actinoplanes globisporus  
 Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;  
 Micromonosporineae; Micromonosporaceae; Actinoplanes.

ANUcomp28394_c0_seq1	gi 522006232	40.98	....	1e-37	144
.....					
ANUcomp28394_c0_seq1	gi 110671508	33.52	....	2e-24	105

LOCUS ABG82005 230 aa linear INV 30-JUL-2006  
 DEFINITION putative endoglucanase [Diaphorina citri].  
 ACCESSION ABG82005  
 VERSION ABG82005.1 **GI:110671508**  
 DBSOURCE accession DQ673432.1  
 KEYWORDS .  
 SOURCE Diaphorina citri (Asian citrus psyllid)  
 ORGANISM Diaphorina citri  
 Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;  
 Pterygota; Neoptera; Paraneoptera; Hemiptera; Sternorrhyncha;  
 Psylliformes; Psylloidea; Psyllidae; Diaphorina.

### Massive Horizontal gene



Bioinformatique  
 Biodiversité  
 Représentation  
 & Intégration  
 des Connaissances



## Alien index (AI)

=

$$\log(\text{Best E-value for Metazoa} + e^{-200}) - \log(\text{Best E-value for Non-Metazoa} + e^{-200})$$

S'il n'y a aucun hit trouvé pour l'un ou l'autre des groupes, la e-value est fixée à 1.

Le AI varie entre -460 et 460.

$$\underbrace{\log(0 + e^{-200})}_{-460,517} - \underbrace{\log(1 + e^{-200})}_0 = -460$$

$$\underbrace{\log(1 + e^{-200})}_0 - \underbrace{\log(0 + e^{-200})}_{-460,517} = +460$$

Massive Horizontal gene Transfer in Bdelloid Rotifers , Gladyshev et al. Science 2008

## Alien index (AI)

=

$$\log( (\text{Best E-value for Metazoa}) + e^{-200} ) - \log( (\text{Best E-value for Non-Metazoa}) + e^{-200} )$$

Il a été estimé qu'un **AI ≥45**, qui représente une différence de **magnitude ≥20** entre le meilleur hit **non-métazoaire** et le meilleur hit **métazoaire**, est un bon indicateur de transfert potentiel de gènes

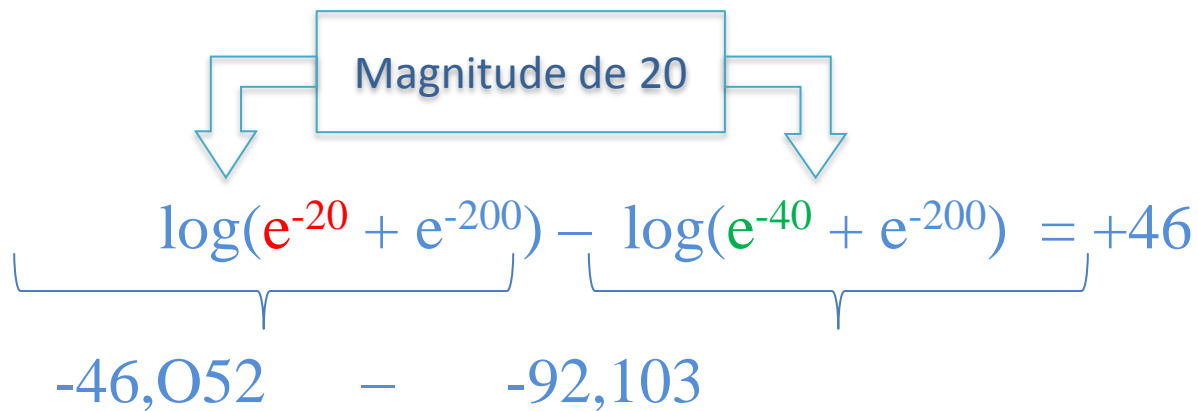
Massive Horizontal gene Transfer in Bdelloid Rotifers , Gladyshev et al. Science 2008



## Alien index (AI)

=

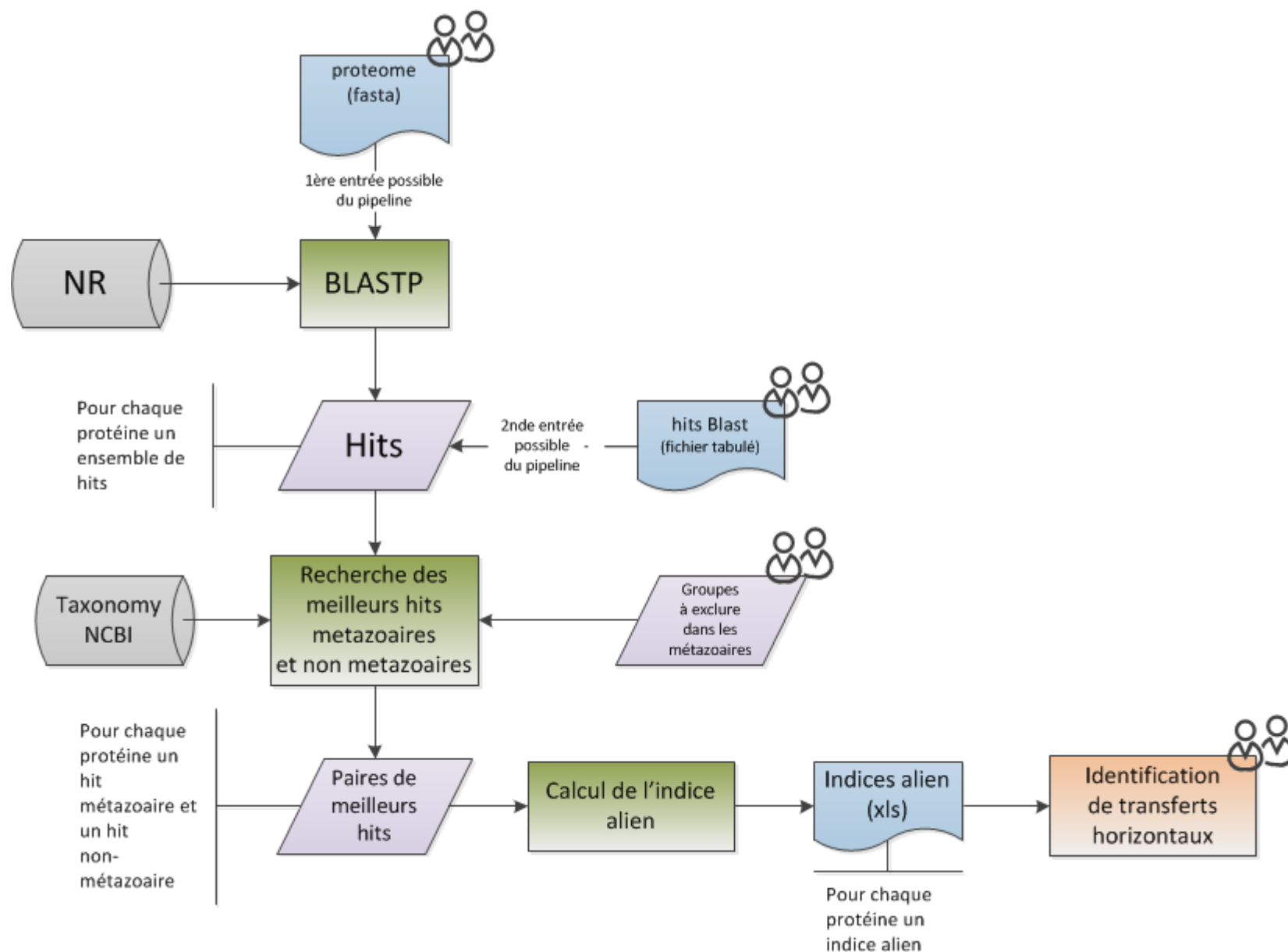
$$\log( (\text{Best E-value for Metazoa}) + e^{-200} ) - \log( (\text{Best E-value for Non-Metazoa}) + e^{-200} )$$

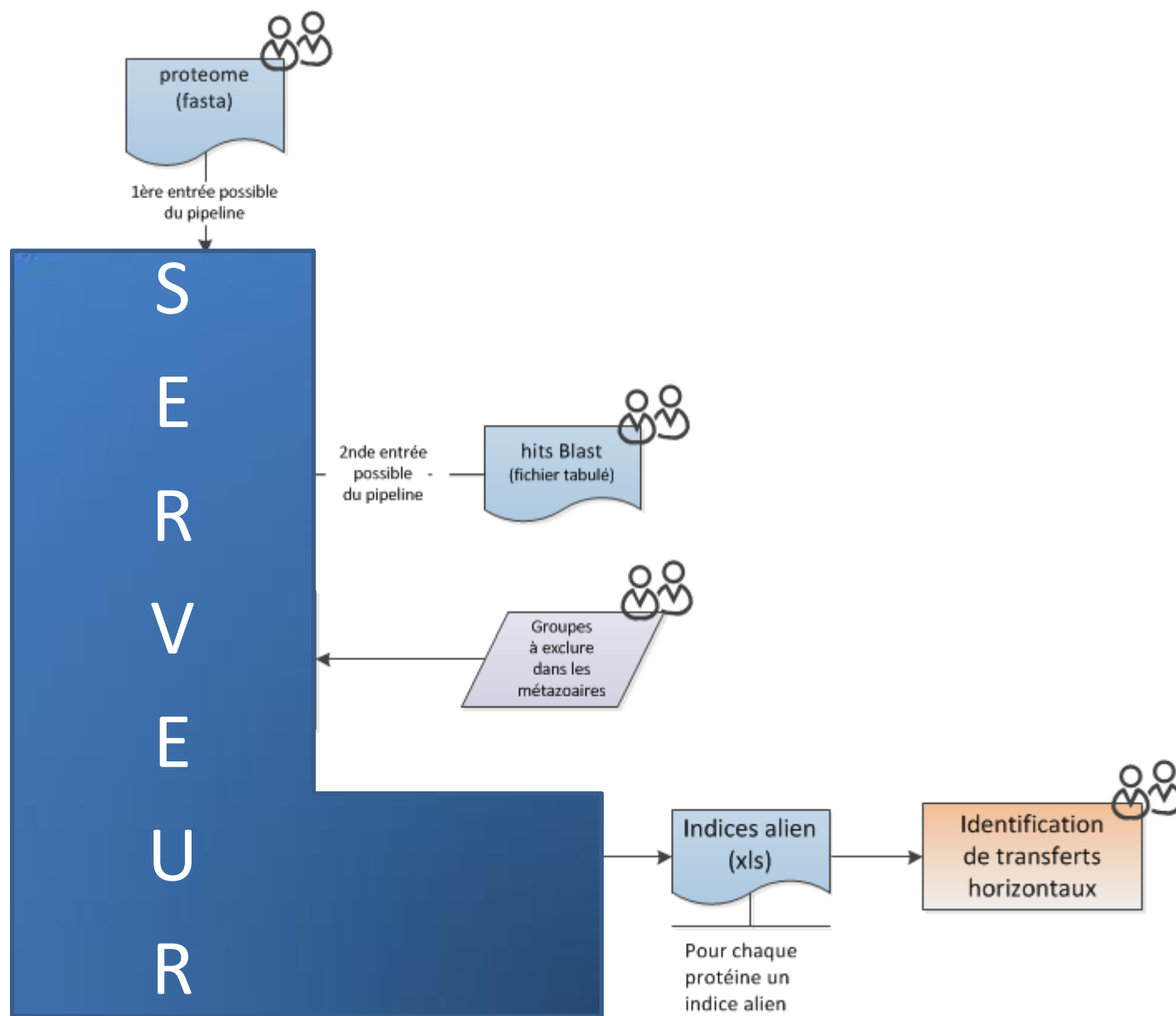


Massive Horizontal gene Transfer in Bdelloid Rotifers , Gladyshev et al. Science 2008









- Pour identifier les transferts horizontaux, il est important de pouvoir exclure l'organisme lui-même mais aussi les organismes trop proches

Horizontal Gene Transfers (version unknown)

**Input type:**  
Proteic fasta ▾  
-outfmt 7 = tabular with comment lines

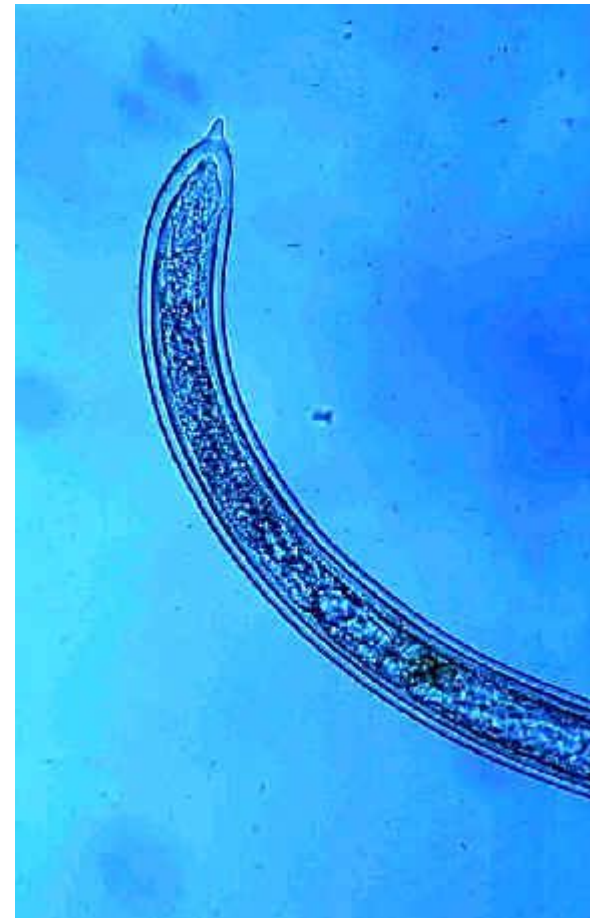
**Fasta file:**  
1: queries\_15\_xiph.fa ▾  
fasta

**Project name:**  
  
one word

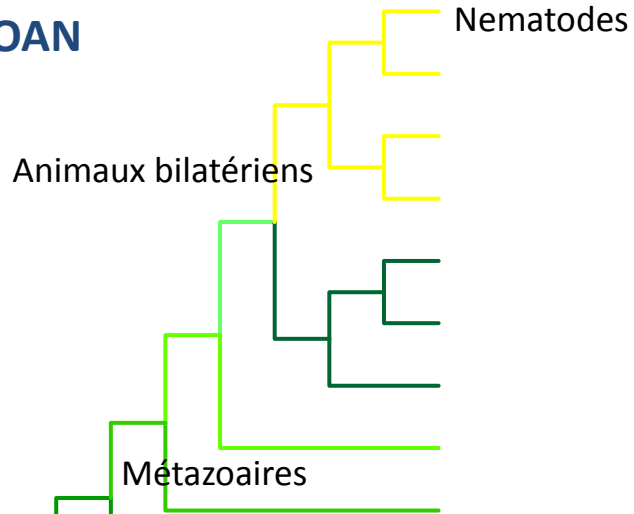
**metazoan group(s) to exclude:**  
  
Metazoan\_group-taxid ex:Longidorus-70230,Xiphidorus-243731,Xiphinema-46002

Execute

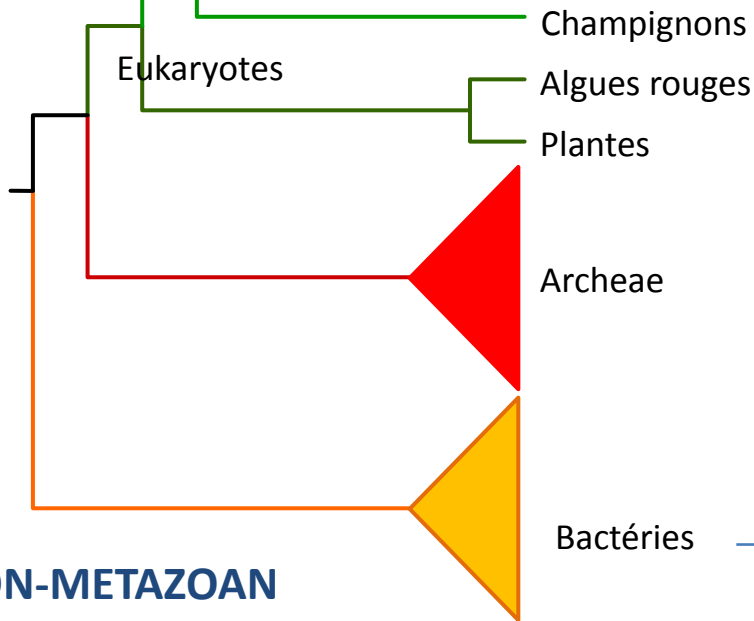
Exemple : on veut détecter les transferts horizontaux chez *Xiphinema index*



## METAZOAN

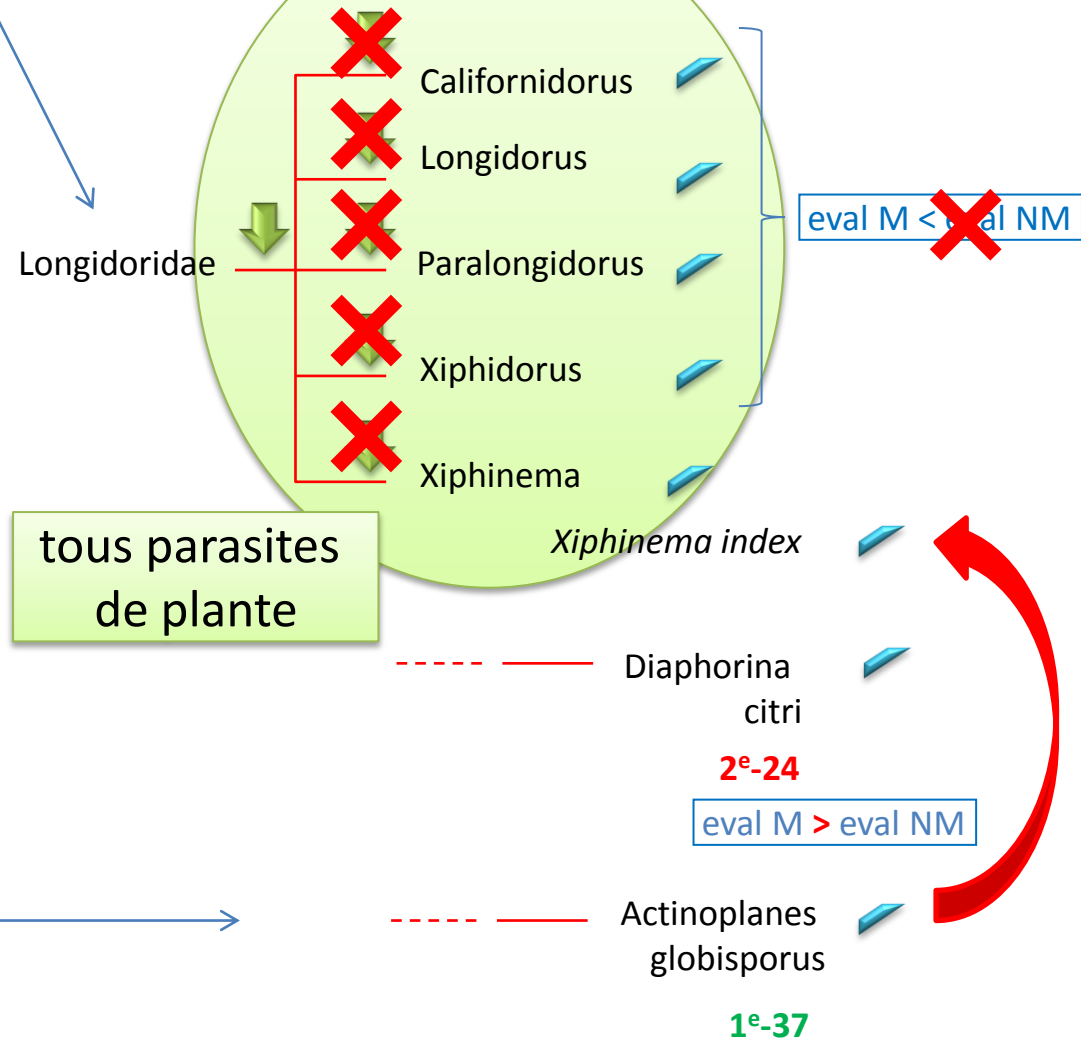


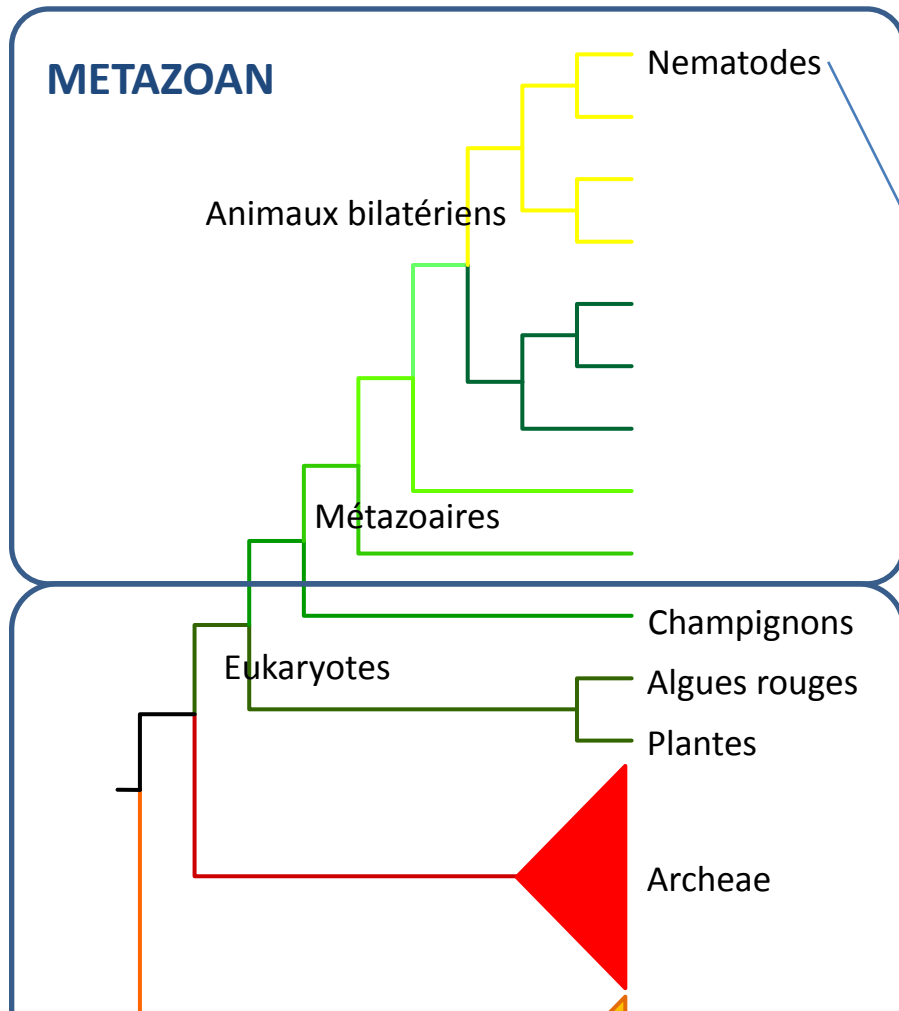
## NON-METAZOAN



A quel niveau veut-on détecter les transferts ?

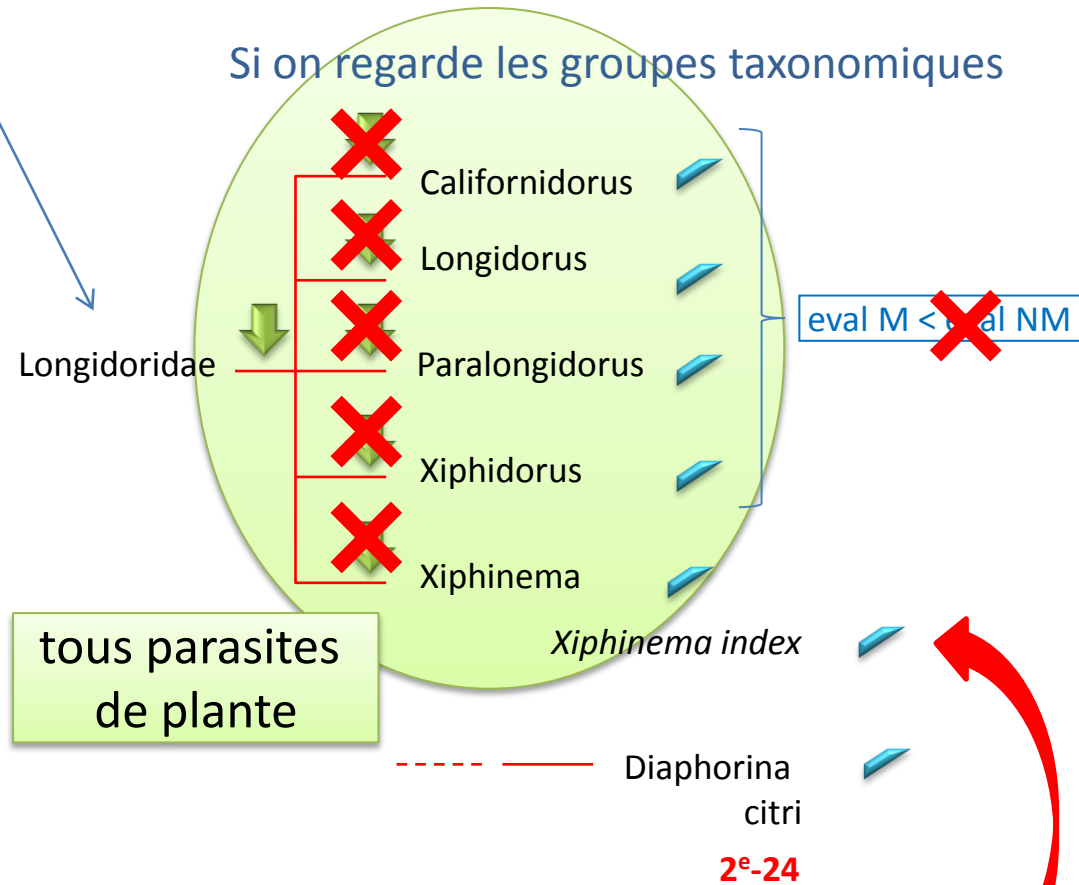
Si on regarde les groupes taxonomiques





A quel niveau veut-on détecter les transferts ?

Si on regarde les groupes taxonomiques

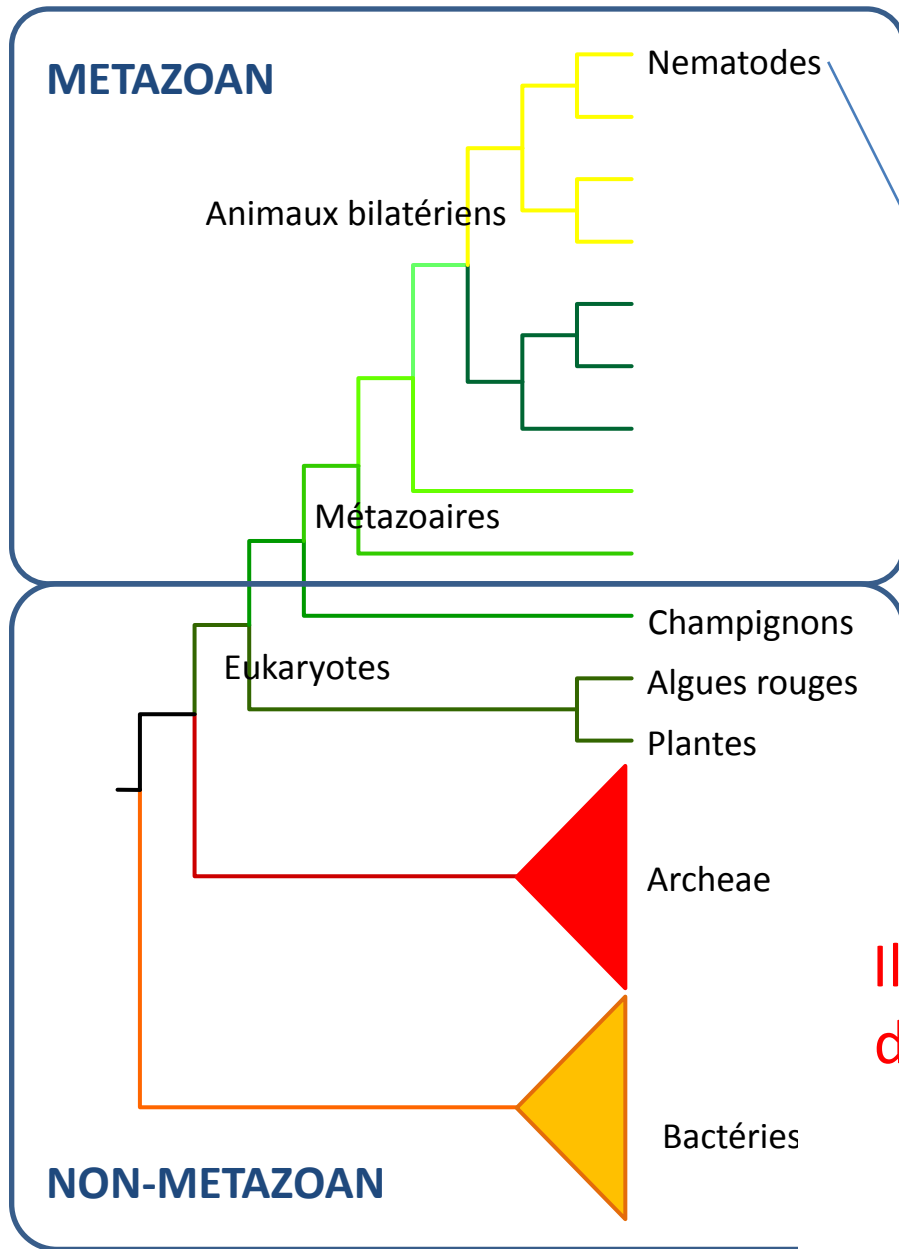


Groupes à exclure : à remplir pour l'option **exclude\_gp**

Longidoridae-46001

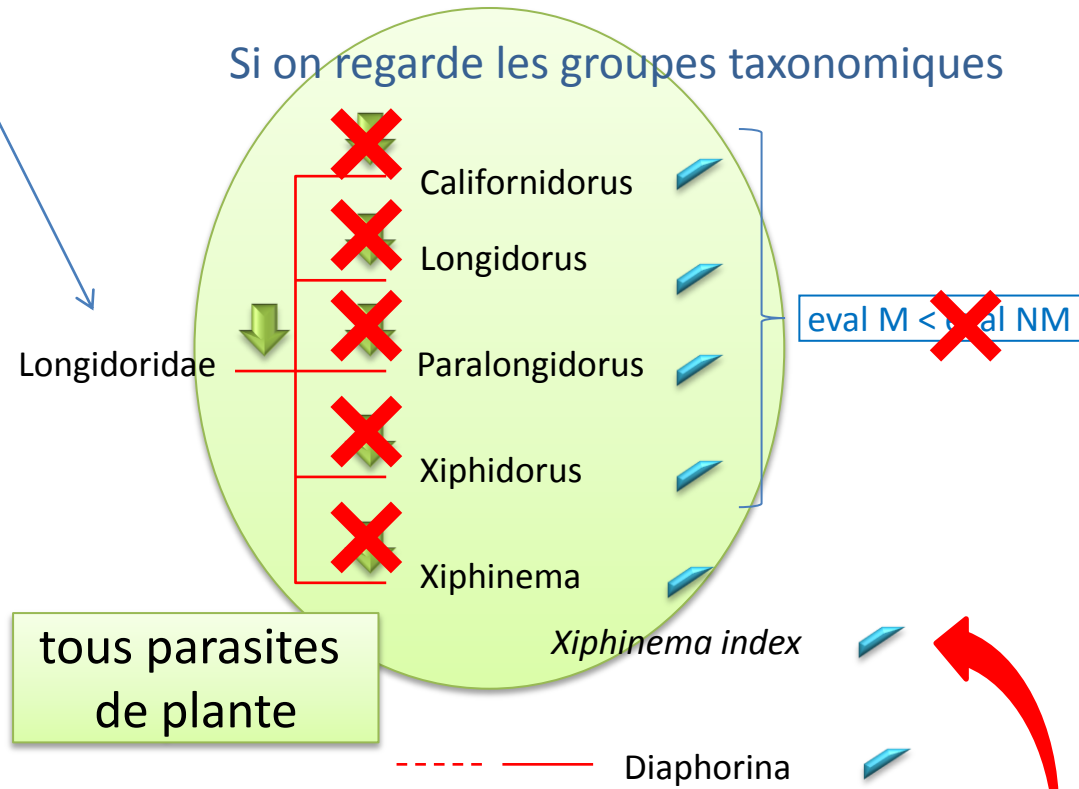
Ou

California-241686, Longidorus-70230, Xiphinema-46002, Paralongidorus-188096, Xiphidorus-243741



A quel niveau veut-on détecter les transferts ?

Si on regarde les groupes taxonomiques

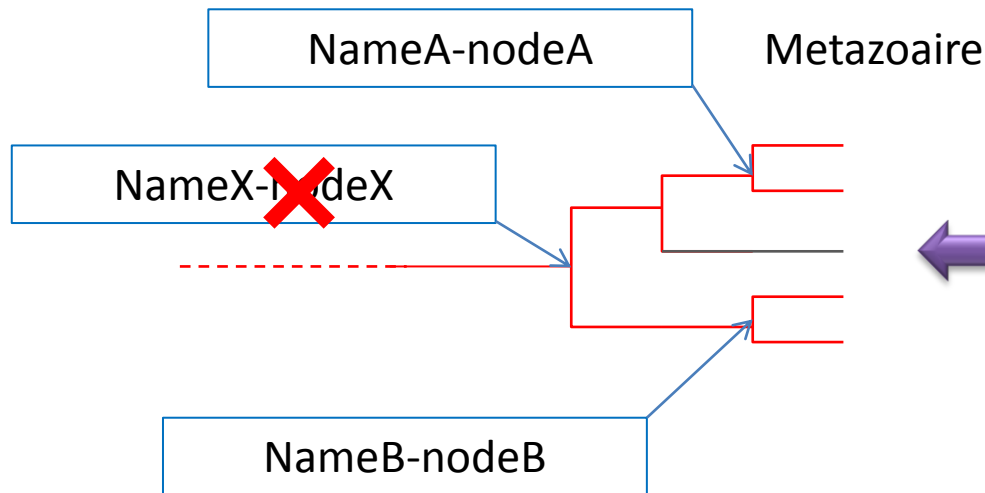


Il faut bien connaître son système, et déterminer la finesse de la détection

espèce spécifique

ou groupe spécifique

A quel niveau veut-on détecter les transferts ?



Un organisme peut être mal classifié !!!  
dans la classification sur laquelle on se base (ncbi ici)

**Groupes à exclure : à remplir pour l'option exclude\_gp**

NameX-nodeX

Ou

NameA-nodeA, NameB-nodeB

Il faut bien connaître son système, et déterminer la finesse de la détection  
espèce spécifique (ex: homme)  
ou groupe spécifique (ex: primates)



Deux formats de fichier au choix en entrée du pipeline :

- Le fichier fasta de protéines
- Le résultat blast au format tableur option -outfmt 7
  - Edit attributes -> Onglet DataType -> Choisir Blast7

Attributes Convert Format **Datatype** Permissions


Change data type

**New Type:**

blast7

This will change the datatype of the existing dataset but *not* modify its contents.

Save

 Galaxy / BBRIC
Analyze Data

**Tools**

**Get Data**  
[Upload File](#) from your computer  
[BBRIC Archive](#) Retrieve data from

**BBRIC protocols**  
[Small genome assembly](#)  
[Annotation on bacterial genome](#)  
 (genome and Illumina oriented RNA-Seq data)  
[Expression measure](#)  
[Biological File Converter](#)  
 Converting, filtering, checking your sequence and annotation files  
[Submission of annotated genomes](#) Genome file format conversion and validation for public database submission.  
[Horizontal Gene Transfers](#) Detection of Horizontal Gene transfers on non-Metazoan origin in Metazoan species  
[Signal Peptide Predictor](#) for plant and eukaryote organisms

**Blast**

**Workflows**  

- [All workflows](#)

**Horizontal Gene Transfers (version unknown)**

**Input type:**  

 -outfmt 7 = tabular with comment lines

**Blastp result:**  

 blast m7 (tabular with comment lines)

**Project name:**  
  
 one word

**metazoan group(s) to exclude:**  
  
 Metazoan\_group-taxid ex:Longidorus-70230,Xiphidorus-243731,Xiphinema-46002

**Program:** wrapper\_cht.pl

**Name:** Horizontal Gene Transfer

**Description:** Horizontal Gene Transfer

**Authors:** [corinne.rancurel@sophia.inra.fr](mailto:corinne.rancurel@sophia.inra.fr)

**Version:** unknown

---

**Inputs**

**infile\_aa** blast m7 (text format)

---

**Outputs**



**10: test2 - Statistics  
file**



**9: test2 - Queries  
without blast hits**



**8: test2 - Horizontal  
Gene Transfer results**



**7: test2 - Warnings  
file about lineage(not  
found, Other or Unclassified  
groups)**



```
20140411_091938 START
15      nb_queries
4619   nb_hits (all gi)
2559   nb_single_gi
679    nb_single_taxid
20140411_091938 END PART 1
20140411_092204 END PART 2
20140411_092204 END PART 3
20140411_092204 END
```

**10: test2 - Statistics  
file**



**9: test2 - Queries  
without blast hits**



**8: test2 - Horizontal  
Gene Transfer results**



**7: test2 - Warnings  
file about lineage(not  
found, Other or Unclassified  
groups)**



Loc\_1779\_Trans\_25-35\_Conf\_0.475\_Length\_1665  
ANUcomp19394\_c0\_seq1

Intéressant aussi pour savoir ce qu'il y a  
de spécifique à un protéome.

**10: test2 - Statistics  
file**



**9: test2 - Queries  
without blast hits**



**8: test2 - Horizontal  
Gene Transfer results**



**7: test2 - Warnings  
file about lineage(not  
found, Other or Unclassified  
groups)**



Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	312377228	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	312377228	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	312377228	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608		NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	328782104	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	328782104	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	328782104	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	328782104	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	328782104	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	328782104	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound
Loc_1_Trans_27351-88886_Conf_0.000_Length_3608	350584925	NO LINEAGE NotFound

Il n'existe pas de synchronisation parfaite dans la mise à jour des fichiers de taxonomie et la version de nr au ncbi.

A partir de certains numéros gi, il est impossible de retrouver le lineage.

**10: test2 - Statistics file**



**9: test2 - Queries without blast hits**



**8: test2 - Horizontal Gene Transfer results**



**7: test2 - Warnings file about lineage(not found, Other or Unclassified groups)**



AI	query hits number	query name
30.6267533894825	250	ANUcomp28394_c0_seq1
-460.517018598809	1553	Loc_1_Trans_27351-88886_Conf_0.000_Length_
-460.517018598809	250	ANUcomp29880_c0_seq1
-46.2748454111951	255	ANUcomp70659_c0_seq2
-172.693881974553	267	ANUcomp30234_c0_seq1
-460.517018598809	251	ANUcomp27134_c0_seq1
-26.0215832034945	250	ANUcomp9206_c0_seq1
15.7614207070196	253	ANUcomp30674_c0_seq1
146.266833662951	250	ANUcomp573545_c0_seq1
0	1	Loc_28783_Trans_1-1_Conf_0.667_Length_159
8.51719319141624	28	ANUcomp24489_c0_seq1
9.0925573363198	45	Loc_34766_Trans_1-4_Conf_0.600_Length_187
-367.315002590379	966	ANUcomp26606_c0_seq1

# FICHIERS DE SORTIE

query name	best evalue Archaea	best evalue Bacteria	best evalue Viroids	best evalue Viruses
ANUcomp28394_c0_seq1	-	1e-37	-	-
Loc_1_Trans_27351-88886_Conf_0.000_Length_	-	-	-	-
ANUcomp29880_c0_seq1	-	-	-	-
ANUcomp70659_c0_seq2	-	-	-	-
ANUcomp30234_c0_seq1	-	-	-	-
ANUcomp27134_c0_seq1	-	-	-	-
ANUcomp9206_c0_seq1	-	-	-	-
ANUcomp30674_c0_seq1	-	-	-	-
ANUcomp573545_c0_seq1	-	-	-	-
Loc_28783_Trans_1-1_Conf_0.667_Length_159	-	-	-	-
ANUcomp24489_c0_seq1	-	-	-	-
Loc_34766_Trans_1-4_Conf_0.600_Length_187	-	9e-08	-	-
ANUcomp26606_c0_seq1	-	-	-	-

query name	best evalue Viridiplantae	best evalue Fungi	best evalue Other	best evalue Metazoa
ANUcomp28394_c0_seq1	-	-	-	2e-24
Loc_1_Trans_27351-88886_Conf_0.000_Length_	-	-	-	0.0
ANUcomp29880_c0_seq1	-	-	-	0.0
ANUcomp70659_c0_seq2	-	-	5e-90	4e-110
ANUcomp30234_c0_seq1	-	-	-	1e-75
ANUcomp27134_c0_seq1	-	-	-	0.0
ANUcomp9206_c0_seq1	-	-	2e-108	1e-119
ANUcomp30674_c0_seq1	-	1e-71	-	7e-65
ANUcomp573545_c0_seq1	-	-	3e-64	-
Loc_28783_Trans_1-1_Conf_0.667_Length_159	-	-	-	-
ANUcomp24489_c0_seq1	-	-	1e-13	5e-10
Loc_34766_Trans_1-4_Conf_0.600_Length_187	-	-	-	8e-04
ANUcomp26606_c0_seq1	-	-	-	3e-160





# FICHIERS DE SORTIE

query name	best hit gi	best hit prct ident	best hit org nickname	best hit org full name	best hit taxo group
ANUcomp28394_c0_	110671508	33.52	A_glo	Actinoplanes globisporus	Bacteria
Loc_1_Trans_27351- ANUcomp29880_c0_	10444510 339244725	41.16 70.11	T_spi T_spi	Trichinella spiralis Trichinella spiralis	Eukaryota_Metazoa Eukaryota_Metazoa
ANUcomp70659_c0_	330800338	38.46	S_scr	Sus scrofa	Eukaryota_Metazoa
ANUcomp30234_c0_	321454577	43.56	D_pul	Daphnia pulex	Eukaryota_Metazoa
ANUcomp27134_c0_	541043055	60.00	A_suu	Ascaris suum	Eukaryota_Metazoa
ANUcomp9206_c0_s	470289190	62.66	B_ter	Bombus terrestris	Eukaryota_Metazoa
ANUcomp30674_c0_	380307969	55.08	S_rac	Syncephalastrum racemosum	Eukaryota_NO_Metazoa
ANUcomp573545_c0_	354549241	92.31	Phyto	Phytophthora sp. SH-2011	Eukaryota_NO_Metazoa
Loc_28783_Trans_1- ANUcomp24489_c0_	- 443725244	- 36.56	- Capsa	- Capsaspora owczarzaki ATCC 30864	- Eukaryota_NO_Metazoa
Loc_34766_Trans_1- ANUcomp26606_c0_	444727908 158296362	45.61 54.49	S_sub Anoph	Spirulina subsalsa Anopheles gambiae str. PEST	Bacteria Eukaryota_Metazoa

query name	best hit taxid	best hit lineage
ANUcomp28394_c0_	113565	;cellular organisms;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Micromonospor
Loc_1_Trans_27351- ANUcomp29880_c0_	6334 6334	;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Protostomia;Ecdysozoa;Nematoc ;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Protostomia;Ecdysozoa;Nematoc
ANUcomp70659_c0_	9823	;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Deuterostomia;Chordata;Craniat
ANUcomp30234_c0_	6669	;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Protostomia;Ecdysozoa;Panarthr
ANUcomp27134_c0_	6253	;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Protostomia;Ecdysozoa;Nematoc
ANUcomp9206_c0_s	30195	;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Protostomia;Ecdysozoa;Panarthr
ANUcomp30674_c0_	13706	;cellular organisms;Eukaryota;Opisthokonta;Fungi;Fungi incertae sedis;Early diverging fungal lineages;Mucc
ANUcomp573545_c0_	1100812	;cellular organisms;Eukaryota;Stramenopiles;Oomycetes;Peronosporales;Phytophthora;unclassified Phytop
Loc_28783_Trans_1- ANUcomp24489_c0_	- 595528	- ;cellular organisms;Eukaryota;Opisthokonta;Opisthokonta incertae sedis;Ichthyosporea;Capsaspora;Capsa
Loc_34766_Trans_1- ANUcomp26606_c0_	54311 180454	;cellular organisms;Bacteria;Cyanobacteria;Oscillatoriothycideae;Oscillatoriales;Spirulina;Spirulina subsalsa ;cellular organisms;Eukaryota;Opisthokonta;Metazoa;Eumetazoa;Bilateria;Protostomia;Ecdysozoa;Panarthr





- Il n'est pas possible de différencier « contamination » et « transfert » juste sur la base de l'alien index.

En se basant sur la colonne qui donne le pourcentage d'identité, et en regardant la fréquence à laquelle réapparaît cet organisme, il est possible de déceler une probable contamination.

- On est restreint aux protéines de la banque publique NCBI et de la taxonomie qu'il propose.
- L'outil est limité aux transferts de gènes d'organismes non métazoaires vers des organismes métazoaires.

- Généraliser le pipeline à d'autres groupes taxonomiques, non métazoaires, comme chez les plantes ou les champignons.
- Proposer un rendu html pour faciliter l'analyse des résultats
- Gérer la détection de la contamination
- Rapport sur l'origine des donneurs (quels groupes, quelles espèces, permettant de détecter des endosymbiontes par exemple)
- Le challenge : Confirmation par des phylogénies automatiques