

Formation sRNAseq Analyse des miRNAs sous Galaxy

- EXERCICES -



cutadapt

BWA

SAMtools

“FastQC is a quality control tool for high throughput sequence data.”

<http://www.bioinformatics.bbsrc.ac.uk/>

A tool that removes adapter sequences from DNA sequencing reads.

cutadapt removes adapter sequences from high-throughput sequencing data. <https://code.google.com/p/cutadapt/>

“Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome.” <http://bio-bwa.sourceforge.net>

“SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments.” <http://samtools.sourceforge.net>



Objectifs :

Cette formation a pour objectif de vous aider à traiter les séquences issues de projet de sRNAseq (miRNA). Vous y découvrirez les problématiques spécifiques de l'analyse des petits ARNs non codant, les outils liés et les mettrez en œuvre afin de détecter, annoter, prédire, quantifier, ... les miRNA.

Pour vous connecter à Galaxy sur l'instance de Roscoff:

- Instance de Roscoff : <http://galaxy.sb-roscoff.fr/>
- Si besoin, instance de Toulouse : <http://sigenae-workbench.toulouse.inra.fr>
(demande de compte sur <http://bioinfo.genotoul.fr/index.php?id=74>)
Comptes de formation suivants :
 - Logins : anemone , aster, bleuet, iris, muguet, narcisse, pensee, rose, tulipe, violette, lilas, pervenche, laurier, lavande, lis, capucine, coquelicot, geranium, liseron, arome, chardon.
 - Password : f1o2r3!

L'ensemble des outils utilisés dans ce TP sont regroupés dans la section « miRNA et sRNAseq » du menu « Tools ».

Exercice n°1 : Prise en main de votre instance Galaxy

1 – Notion de « dataset » et de « tool »

Un fichier téléchargé dans Galaxy se nomme « dataset » et est visible à droite de votre interface Galaxy, dans la fenêtre « history ».

Nous allons travailler à partir de données Illumina représentant 2 tissus. Ces jeux de données « [s2.fastq](#) » et « [s1.fastq](#) » peuvent être téléchargés dans Galaxy à partir de ce lien : http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/miRNA/Roscoff/ ou directement en récupérant les fichiers dans l'instance Galaxy de Roscoff : shared data -> data libraries -> tp-jeudi-mirna-sarah .

Pour que les fichiers « [s2.fastq](#) » et « [s1.fastq](#) » soient disponibles dans Galaxy pour vos traitements, veuillez utiliser l'outil « [Upload File from your computer](#) » en précisant les URL des fichiers à télécharger. Cet outil télécharge en copiant votre fichier sur le serveur Galaxy. Cette copie diminue donc votre quota sur Galaxy.



Si l'extension de votre fichier est « .fastq », alors ce fichier ne sera pas visible par les outils Galaxy. Pour ce faire, il est nécessaire de modifier l'extension de « fastq » en « fastqsanger » via l'icône « pencil » de votre dataset, menu « change datatype ».

Veuillez renommer les fichiers importés :

Renommer « http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/miRNA/Roscoff/s1.fastq » en « s1 ».

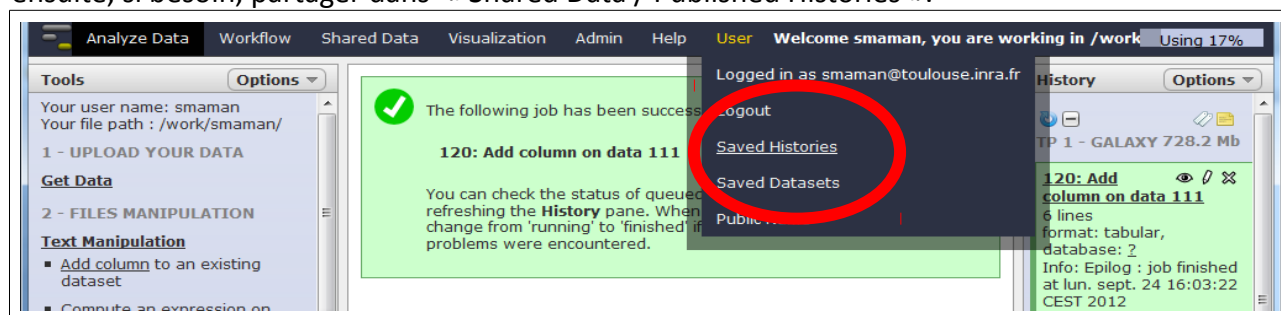
Renommer « http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/miRNA/Roscoff/s2.fastq » en « s2 ».

Puis changer le datatype de ces deux fichiers, de « fastq » à « fastqsanger ».

2 – Création et partage d'historique et de workflow

2.1 – Archivage automatique de vos traitements dans un historique

Au fur et à mesure que vous faites appel aux différents outils utiles au sein de votre interface depuis le menu « Analyse Data », l'ensemble des étapes sont enregistrées dans un historique qui est automatiquement archivé dans « User / Saved Histories » et que vous pouvez ensuite, si besoin, partager dans « Shared Data / Published Histories ».



Depuis le menu « User » / « Saved Histories », vous avez la possibilité de gérer vos historiques (delete, delete permanently, rename, undelete) en cliquant sur l'intitulé de l'historique.



A chaque connexion à votre instance Galaxy, un nouveau « current history » est automatiquement créé.

Tools Options ▾
Linux Username: sigenae
File Path : /work/sigenae/galaxy/
Serveur : galaxy
Get Data
Send Data
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
Multiple regression

Saved Histories

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑
☐ Unnamed history ▾		0 Tags		0 bytes	less than a minute ago	less than a minute ago
☐ RNA seq statistics ▾	2	4	0 Tags	34.9 Mb	~ 6 hours ago	6 minutes ago
☒ Test BWA fichiers Gnome ▾	4	0 Tags		9.0 Mb	Apr 06, 2012	1 day ago
☒ Test region promoters ▾	5	0 Tags		23.9 Mb	Mar 08, 2012	Apr 06, 2012
☐ Unnamed history ▾	3	0 Tags		60 bytes	Feb 23, 2012	Mar 09, 2012
☐ Unnamed history ▾	1	1	0 Tags	0 bytes	Mar 07, 2012	Mar 07, 2012
☐ Unnamed history ▾	1	1	0 Tags	16.0 Kb	Feb 22, 2012	Mar 05, 2012

For 2 selected histories: **Rename** **Delete** **Delete Permanently** **Undelete**

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.

Veillez nommer votre historique « miRNAs »,

2.2 – Partage de vos analyses grâce aux workflows

Using 24%

History

History Lists
Saved Histories
Histories Shared with Me
Current History
Create New
Clone
Copy Datasets
Share or Publish
Extract Workflow
Dataset Security
Show Deleted Datasets
Show Hidden Datasets

TP 1 - GAL
128: FAST
data NC 0
1 line
format: tab
Info: Epilog
24 16:47:2
1
gi | 22458183
ATACTTTAAAC
TTTTCTGCGCG

Pour lancer plusieurs jobs en une fois, vous pouvez générer un workflow à partir de votre historique (**History** panel / click **Options** → **Extract Workflow**) ou bien créer directement un workflow depuis une page blanche.

Veillez générer un workflow depuis votre historique « miRNA » et le nommer « miRNA-Votre/user/login »,

Puis partager le workflow que vous venez de générer avec un autre utilisateur Galaxy présent dans la salle ("Share with a user"). Pour cela, vous avez besoin de connaître son nom d'utilisateur sous Galaxy.

Faire un clone du workflow partagé.

Exercice n°2 : Analyse de la qualité et nettoyage (suppression des adaptateurs – suppression de la redondance intra « fastq »)

1 – Analyse de la qualité des données : « FastQC »

A partir du jeu de données s2, veuillez utiliser l'outil Galaxy « [Fastqc: Fastqc QC using FastQC from Babraham](#) » pour générer les graphiques d'analyse de la qualité.

Fastqc: Fastqc QC (version 0.4)

Short read data from your current history:

38: Suppression des a.. on data 25

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

FastQC

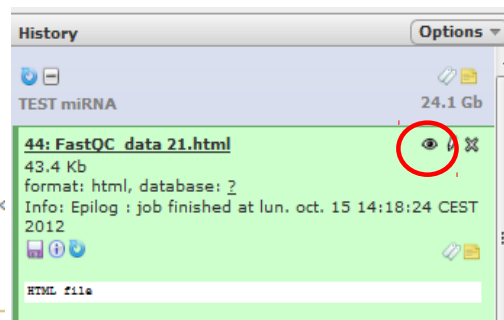
Contaminant list:

Selection is Optional

tab delimited file with 2 columns: name and sequence. For ex CAAGCAGAAGACGGCATACGA

Execute

En cliquant sur l'icône « œil », les statistiques s'affichent dans la fenêtre centrale :



Il vous est possible de récupérer :

- L'ensemble des résultats de FastQC en cliquant sur l'icône « disquette » du dataset.



Remarque : Pour les personnes connectées à l'instance Galaxy de Roscoff (uniquement), l'ouverture des rapports html générés dans le cadre de ce TP s'effectue via un click droit sur l'icône "save" de votre dataset, "enregistrer sous", puis ajouter ".html" au nom du fichier.

- Un graphique par un clic droit sur son intitulé, en fin de page :

Files created by FastQC

[duplication_levels.png \(19.9 KB\)](#)
[fastqc_data.txt \(24.2 KB\)](#)
[fastqc_report.html \(42.1 KB\)](#)
[kmer_profiles.png \(105.8 KB\)](#)
[per_base_gc_content.png \(31.4 KB\)](#)
[per_base_n_content.png \(7.9 KB\)](#)
[per_base_quality.png \(9.2 KB\)](#)

2 – Nettoyage des données :

2.1 – Suppression des adaptateurs

La suppression des adaptateurs s'effectue grâce au script « cutadapt » disponible dans Galaxy « [* Suppression des adaptateurs](#) ». Veuillez lancer ce traitement sur « s2 ».

Paramètres :

- adaptateur : ATCTCGTATGCCGTCTTCTGCTTG
- taille minimum : 18pb
- taille maximum : 25pb

Remarques :



- Attention, la taille minimum des séquences requise pour mirdeep2 est de 18 pb.
- Un astérisque * devant le nom de l'outil signifie qu'il s'agit d'un module Galaxy rajouté par Sigenae.

Sorties de « cutadapt » :

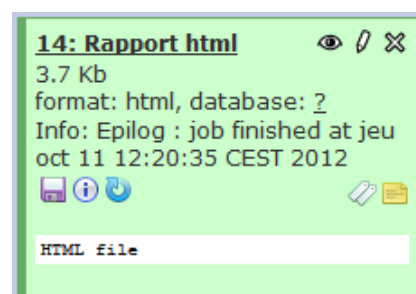
- fichier « fastq » contenant les séquences nettoyées de leurs adaptateurs
- fichier « log » contenant les informations de traitement

Veuillez renommer le fichier sortant « Suppression des adaptateurs on data 5 » en « cut_s2_log » pour le fichier de log et en « cut_s2_sequences » pour le fichier fastq.

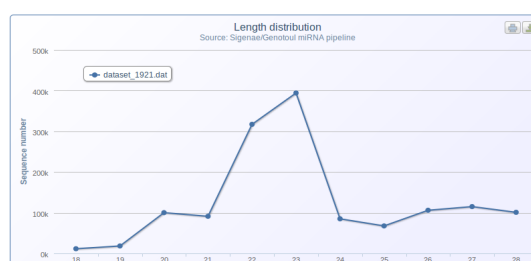
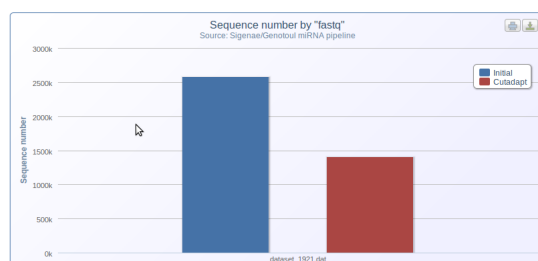
2.2 – Statistiques après suppression des adaptateurs

Lancer l'outil « [* Production du rapport](#) » après élimination des adaptateurs », sur « cut_s2_log » et « cut_s2_sequences », pour produire d'un rapport de traitement (au format html).

Veuillez renommer le fichier sortant « Votre rapport apres suppression des adaptateurs » en « Votre rapport apres suppression des adaptateurs sur s2 ».



Pour visualiser ces statistiques, veuillez enregistrer le fichier sur votre PC et ouvrir la page « html » dans votre navigateur.



Remarque : Pour les personnes connectées à l'instance Galaxy de Roscoff (uniquement), l'ouverture des rapports html générés dans le cadre de ce TP s'effectue via un click droit sur l'icône "save" de votre dataset, "enregistrer sous", puis ajouter ".html" au nom du fichier.

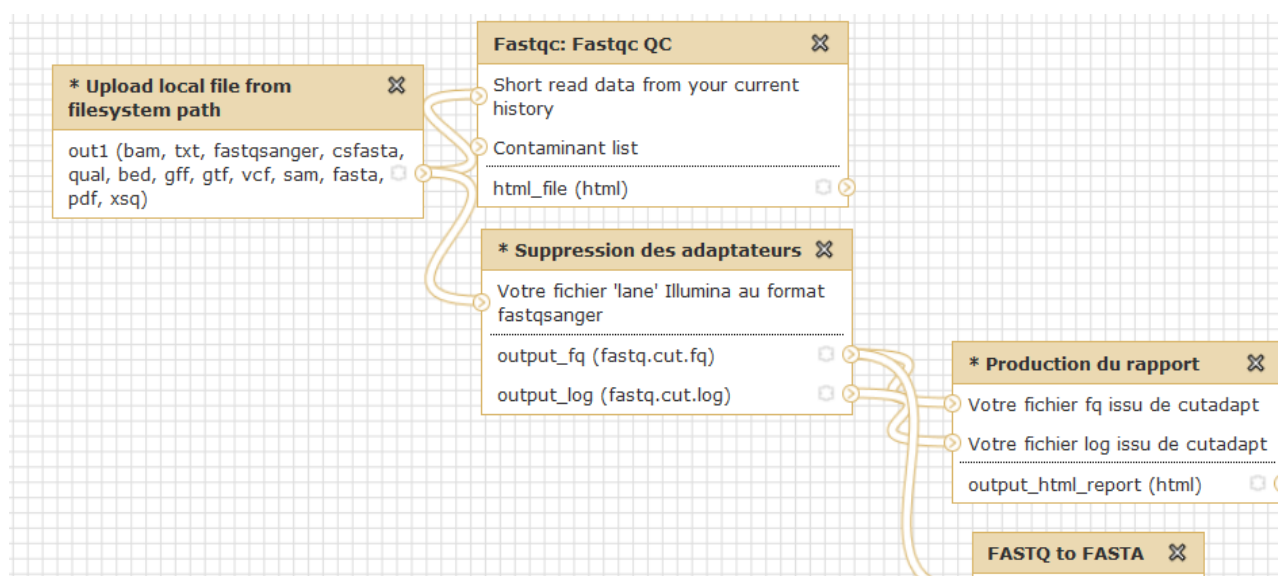


3- Exporter votre historique en workflow

Afin de pouvoir relancer ce workflow sur l'ensemble de vos « fastq », veuillez convertir l'historique que vous venez de générer en workflow « miRNA-WF1-qualite_et_nettoyage_des_donnees ».

Pour chaque jeu « fastq », il vous suffira alors de relancer ce WF1.

Relancer votre workflow sur le « fastq » s1 après avoir supprimé les deux boîtes d'upload des vos jeux de données (car ils sont déjà dans votre historique) et avoir vérifier que les outils sont bien paramétrés. Il est préférable de ne pas envoyer les fichiers générés par ce workflow vers un nouvel historique.

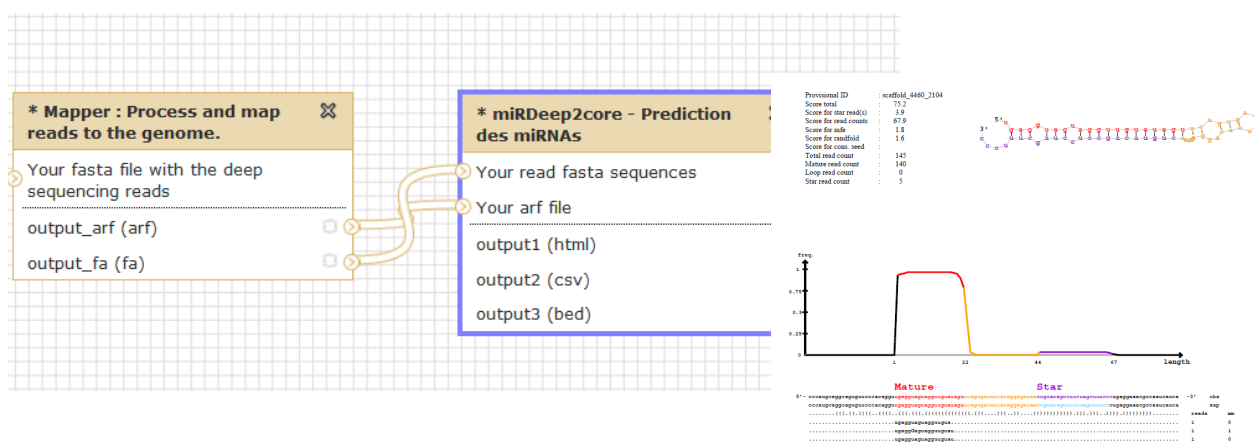


Pour plus de lisibilité, veuillez renommer les datasets suivantes :

Nom proposé par Galaxy	Renommage
* Suppression des adaptateurs on data 4 (format log)	cut_s1_log
* Suppression des adaptateurs on data 4 (format fastq)	cut_s1_sequences
* Votre rapport apres suppression des adaptateurs	Votre rapport apres suppression des adaptateurs sur s1

Exercice n°3 : Recherche de miRNAs avec mirdeep2

Le WF2 « miRNA-WF2-Mirdeep2 » exécute mirdeep2 à partir des fichiers générés par le l'outil « Suppression des adaptateurs » du WF1.



La première étape de mirdeep2 (« mapper ») permet d'aligner les lectures (fichier « fasta » précédent) sur le génome de référence passé en paramètre (V4_454Scaffolds_filter) : outil « [* Process and map reads to the genome.](#) »

La seconde étape de mirdeep2 (« core ») permet d'annoter les miRNAs appartenant à des familles d'ARN connus : outil « [* Annotation des miRNAs](#) ».

Voici vos étapes de traitement :

1. Au préalable, il est nécessaire de convertir vos fichiers « fastq » (« cut_s1_sequences » et « cut_s2_sequences ») en fichiers « fasta » à l'aide de l'outil « FASTQ to FASTA ». Veuillez renommer les fichiers sortants « [fastq -> fasta] Output File » en « s1_fasta » et « s2_fasta ».
2. Importation du WF2 intitulé « Galaxy-Workflow-miRNA-WF2-Mirdeep2.ga » sur http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/miRNA/Roscoff/workflows/Galaxy-Workflow-miRNA-WF2-Mirdeep2.ga
3. Visualiser (clic sur l'intitulé du workflow → Edit) le workflow importé pour sélectionner la banque V4_454Scaffolds_filter pour les 3 outils puis enregistrer ces modifications dans WF2.
4. Lancement du workflow WF2 (clic sur l'intitulé du workflow → Run) sur les fichiers entrants « s1_fasta » et « s2_fasta », dans l'historique en cours.
5. Visualisation des résultats : Veuillez observer les premières lignes du fichier « fasta » généré par l'outil « mapper » et les « pdfs » générés par l'outil d'annotation de mirdeep2.
6. Pour plus de lisibilité, veuillez renommer les datasets suivantes :

Nom proposé par Galaxy	Renommage
Fichier fasta généré par l'outil mapper de « Mapper : Process and map reads to the genome. on data 16 and data 18 »	« Mapper on s1 and s2 fasta »
Fichier arf généré par l'outil mapper de « Mapper : Process and map reads to the genome. on data 16 and data 18 »	« Mapper on s1 and s2 arf »
Fichier fasta généré par le convertisseur « * miRDeep2core - bed to fasta file on data 24 ».	« miRDeep2core fasta »





Exercice n°4 : Recherche des annotations fonctionnelles

Quelques liens (outils / base de données) :

- BWA : <http://bio-bwa.sourceforge.net>
- BWA man : <http://bio-bwa.sourceforge.net/bwa.shtml>
- SAMtools : <http://samtools.sourceforge.net>
- mirBase : <ftp://mirbase.org/pub/mirbase/CURRENT/> (hairpin.fa.gz)
- Rfam : <ftp://ftp.sanger.ac.uk/pub/databases/Rfam/CURRENT/> (Rfam.fasta.gz)
- tRNA : <http://gtrnadb.ucsc.edu/download.html> (eukaryotic-tRNAs.fa.gz)
- rRNA : <ftp://ftp.arb-silva.de/current/Exports/> ([LS]SUF108_tax_silva_trunc.fasta.tgz)

Dans le cadre de ce TP, les banques de référence sont déjà disponibles dans le menu déroulant de l'outil BWA.

L'objectif est de comparer les diagrammes de Venn obtenus après le mapper avec ceux obtenus après le mirdeep2core.

Voici vos étapes de traitement :

1. Annoter le « fasta » issu de Mapper (« Mapper on s1 and s2 fasta ») puis le « fasta » issu de miRDeep2core (« miRDeep2core fasta ») avec l'outil « *** Alignement bam et tri sur un fasta** ».

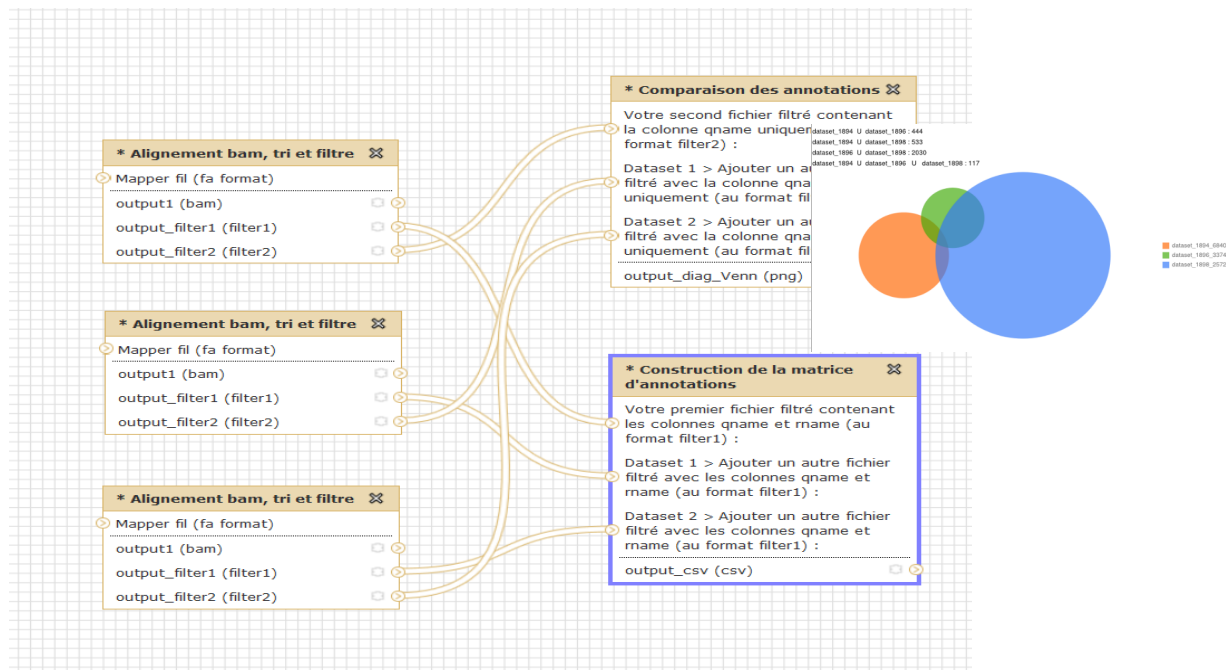
Chaque « fasta » doit être annoter 3 fois, une fois contre la banque mirBase (hairpin_T), puis contre la banque Rfam puis contre la banque tRNA. Chaque alignement doit être filtré selon ces paramètres : Flag 0 ou 16 – 1 mismatch autorisé. Le filtre est nécessaire car il n'est pas possible avec BWA de spécifier qu'on ne veut que les alignements avec un seul mismatch. Les options 0 et 16 permettent de garder les alignements obtenus en forward et en reverse.

2. Pour plus de lisibilité, veuillez renommer les datasets suivantes :

Nom proposé par Galaxy	Renommage
« * Alignement baw, tri et filtre on data 21 » - Fichier au format filter1, sur hairpin_T, issu du « Mapper on s1 and s2 fasta »	«mapper hairpin_T filter1 » A décliner pour Rfam et tRNA.
« * Alignement baw, tri et filtre on data 21 » - Fichier au format filter2 sur hairpin_T, issu du « Mapper on s1 and s2 fasta »	«mapper hairpin_T filter2 » A décliner pour Rfam et tRNA.
« * Alignement baw, tri et filtre on data 21 » - Fichier au format filter1, sur hairpin_T, issu du « miRDeep2core fasta »	«miRDeep2core hairpin_T filter1 » A décliner pour Rfam et tRNA.
« * Alignement baw, tri et filtre on data 21 » - Fichier au format filter2 sur hairpin_T, issu du « miRDeep2core fasta »	« miRDeep2core hairpin_T filter2 » A décliner pour Rfam et tRNA.

3. Puis lancer l'outil « *** Comparaison des annotations** », à partir des 3 fichiers au format « filter2 » générés par l'alignement. L'objectif est de construire un diagramme de Venn.

4. Puis lancer l'outil « [* Construction de la matrice d'annotations](#) », à partir des 3 fichiers au format « filter1 » générés par l'alignement.
5. Comparer les diagrammes de Venn selon que les traitements aient été lancés à partir du « fasta » issu du Mapper ou du « fasta » issu de miRDeep2core.



Cette suite d'outils « miRNA et sRNAseq » est disponible dans le Galaxy Tool Shed (<http://toolshed.g2.bx.psu.edu/>) avec le mot clés « miRNA » :

Galaxy Tool Shed Repositories Help User

2430 valid tools on Jan 11, 2013

Search

- Search for valid tools
- Search for workflows

All Repositories

- Browse by category

My Repositories and Tools

- Repositories I own
- My writable repositories
- My invalid tools

Available Actions

- Create new repository

Repositories

miRNA x search repository name, description

Name ↓	Synopsis	Metadata	Revisions	Tip	Revis
mirdeep2_and_targetspy	Finding miRNA in NGS data and finding the targets for those miRNA	1:798fe7ba8b5e	1:798fe7b		
mirna_mirdeep2	miRNA	3:12ab33cb3511	3:12ab33c		