# Mini-projets

## Les rapports sont à rendre avant le 28 février 2015

***Annotation et analyse des ARNnc:***

**1. Utiliser RNASpace pour l'annotation de génomes**

**2. Utiliser miRdeep2**

**3. Analyse comparative des ARNnc chez les Listeria**

**4. Développer un outil d'annotation de séquences sRNAseq**

**9. Créer un logiciel de nettoyage des données sRNAseq**

***GALAXY :***

**5. Recherche de variants génomiques dans GALAXY**

**6. Créer un wrapper GALAXY dédié à supprimer des séquences peu complexes**

**8. Créer un wrapper GALAXY modèle**

**10. Conception et création de videos tutoriels sur GALAXY**

# 1. Utiliser RNASpace pour l'annotation de génomes de mycoplasmes

**Référent :**

Christine Gaspin – Christine.Gaspin@toulouse.inra.fr

**Objectif :**

Utiliser l'environnement RNAspace pour annoter les ARNnc d'un groupe de mycoplasmes.

**Données :**

Récupérer au NCBI les séquences au format Fasta des génomes suivants

- Mycoplasma capricolum subsp. capricolum, ATCC 27343 (public)
- Mycoplasma mycoides subsp. mycoides SC, Gladysdale (public)

**Rendu :**

La démarche et l'ensemble des résultats feront l'objet d'un rapport court qui sera remis au référént. La page RNAspace permettant de visualiser les résultats des outils utilisés sera rendue disponible.

**Mise en oeuvre :**

La plateforme logicielle RNAspace est accessible à l'URL rnaspace.org. Pour chacun des génomes, il est demandé d'identifier les ARNnc appartenant à des familles connues à l'aide des logiciels disponibles dans RNAspace. Vous vous appuirez sur ces résultats pour identifier, à l'aide de l'approche comparative disponible dans RNAspace, des candidats appartenant à de nouvelles familles. A partir de l'ensemble de ces résultats vous construirez les profils phylogénomiques pour l'ensemble des génomes comparés, à l'aide d'un tableau en codant 1 (N>1) la présence de(s) l'ARNnc orthologue(s) et 0 son absence.

# 2. mirdeep2 - Compréhension d'un outil bioinformatique

**Référent :** Olivier Rué - olivier.rue@toulouse.inra.fr

**Objectif :** Faire tourner le pipeline mirdeep2 sur 2 jeux de données sRNAseq et obtenir une liste de prédictions de miRNA

**Données :** http://genoweb.toulouse.inra.fr/~formation/UPS/2/

- Oncorhynchus_mykiss_V1.0.fa : génome de référence de la truite
- s_2_uT12.fastq et s_5_uT14.fastq
- miRNA matures de miRbase : mature.fa
- précurseurs de miRNA de miRbase : hairpin.fa
- la publication de miRDeep2

Faites "copy link location" sur le lien et utilisez la commande wget pour récupérer les données dans votre espace de travail.

**Rendu :** Mini-rapport expliquant la démarche et contenant les réponses aux questions ci-dessous. Emplacement de votre espace de travail sur genotoul.

**Mise en oeuvre :**

Le logiciel mirdeep2 a été installé sur genotoul dans le répertoire **/usr/local/bioinfo/src/mirdeep2/current/**. Tous les scripts se trouvent dans ce répertoire. Un tutoriel est proposé, je vous conseille de le lire (ou de le suivre) avant de vous lancer !

Mapper les reads grâce au script mapper.pl

1) Comment le mapping est-il effectué d'après la documentation (outil, paramètres...) ? Détaillez les paramètres --best, --strata, -n, -e et -l du logiciel de mapping utilisé par mirdeep2.
2) Ouvrez le script mapper.pl et étudiez-le. Quel outil est utilisé pour convertir les fichiers FASTQ en fichiers FASTA. Reprenez le code correspondant et détaillez-le.
3) Expliquez les étapes successives du mapper. Quel est le gros défaut de ce code en terme d'optimisation ? Proposez une façon de faire plus adaptée aux gros jeux de données (pseudo-code)
4) Expliquez l'algorithme utilisé pour supprimer les adaptateurs. Vous parait-il correct ?
5) Détaillez le ficher obtenu à l'issue du mapper (.arf). Quels champs vous semblent inutiles, ou comment feriez-vous pour stocker l'information de manière plus intelligente ? Quel format standard d'alignement de séquences est communément utilisé pour stocker ces informations ?
6) Compte tenu des options passées au script parse_mappings.pl, proposez une expression régulière ou une commande unix qui fasse la même chose.

Annotation et quantification de miRNA connus

7) Lancez le module de quantification de mirdeep2 en utilisant uniquement les miRNA connus du Zebrafish.

Prédiction de nouveaux miRNA

8) À partir des fichiers hairpin.fa et mature.fa, créez deux fichiers ne contenant que les séquences Zebrafish (tni).
9) Lancez le module de prédiction de mirdeep2 en utilisant ces fichiers et le fichier .mrd issu du module de quantification.
10) La documentation autorise de passer en paramètre le fichier .mrd issu de l'annotation. Pourtant, il semble qu'une quantification soit lancée au début de la prédiction. Isolez le code correspondant à cette action et proposez un code (tout simple) qui puisse éviter de relancer la quantification.
11) Détaillez chaque étape effectuée lors de la prédiction. Aidez-vous de la publication.
12) En lisant le fichier report.log, vous constaterez que le logiciel n'a pas pu faire de liens BLAT vers l'USCS et le NCBI car le nom de l'espèce n'a pas été indiqué. Relancez le module de prédiction, mais pas sur tout le génome (temps de traitement long…) mais uniquement sur le scaffold_1. Détaillez votre méthodologie pour y parvenir.
13) Combien de nouveaux miR sont prédits sur le scaffold 1 ? Combien de miR connus sont retrouvés sur le scaffold 1 ?

**Conseils :**

- Avant de vous lancer dans le traitement avec les jeux de données complets, il est **TRÈS FORTEMENT** recommandé de vous créer des petits jeux de données sur lesquels tester vos commandes (un scaffold, quelques millers de reads par jeu).
- Les commandes avec les jeux de données réelles doivent être lancées par un qsub ou après un qlogin.

# 3. Analyse comparative des ARNnc chez les Listeria avec l'outil Blastn

**Référent :** Christine Gaspin – Christine.Gaspin@toulouse.inra.fr

**Objectif :** Utiliser l'outil d'alignement Blastn pour réaliser une analyse comparative des ARNnc chez les Listeria à partir d'une liste donnée d'ARNnc connus chez Listeria monocytogenes..

**Données :** Récupérer au NCBI les séquences des génomes complets de Listeria. La liste des ARNnc (~300) sera fournie par le référent.

**Rendu :** La démarche et l'ensemble des résultats feront l'objet d'un rapport qui sera remis au référént.

**Mise en oeuvre :**

L'outil Blastn est disponible en ligne de commande sur la plateforme bioinformatique GenoToul. Les ARNnc seront fournis sur la base de 3 classes: les ARNnc en trans, les riboswitch/thermosenseurs et les ARNnc antisens. Pour chaque ARNnc, il est demandé d'identifier ses ARNnc homologues dans les autres génomes de Listeria. A partir de l'ensemble de ces résultats vous construirez les profils phylogénomiques pour l'ensemble des génomes comparés, à l'aide d'un tableau en codant N (N>0) la présence de(s) l'ARNnc(s) homologue(s) et 0 son absence.

# 4. Développer un outil d'annotation de séquences sRNAseq

**Référents :** Olivier Rué - olivier.rue@toulouse.inra.fr

**Objectif :** Développer un "outil" permettant l'annotation de locus de sRNAseq

**Données :** http://genoweb.toulouse.inra.fr/~formation/UPS/4/
- les reads alignés : file.bam
- banque miRbase : mirbase.fa
- banque Rfam : rfam.fa
- banque tRNA : trna.fa
- banque Silva : silvaLSU.fa

Faites "copy link location" sur le lien et utilisez la commande wget pour récupérer les données dans votre espace de travail.

**Rendu :** rapport expliquant la démarche et outil.

**Mise en oeuvre :**

Spécifications :
Input :
- file.bam
- banques
- choix du format d'output (BED ou GFF3)

Logiciels à utiliser :
- /usr/local/bioinfo/bin/bwa
- /usr/local/bioinfo/bin/samtools

Output :
- fichier BED ou GFF3

**Conseils :**
- Lisez bien les spécifications du format SAM avant de vous lancer.
- Un locus est une suite de lectures alignées le long du génome qui se chevauchent d'au moins une base.
- Il doit être possible de spécifier en paramètre le nombre de nucléotides minimum pour distinguer deux locus (défaut : 1).
- L'annotation du locus est définie par l'annotation des séquences qui le composent.
- Une séquence annotée est une séquence mappée parfaitement sur une banque avec bwa (0 mismatch).
- Un locus peut avoir plusieurs annotations.
- Vous avez le droit d'écrire des fichiers temporaires (organisez-les intelligemment).
- Pensez à l'optimisation de votre code.

**Documentation :**

- format GFF3 : http://gmod.org/wiki/GFF3
- format BED : http://www.ensembl.org/info/website/upload/bed.html
- bwa : http://bio-bwa.sourceforge.net/
- samtools : http://samtools.sourceforge.net/

# 5. Recherche de variants génomiques dans GALAXY

**Référent :** Olivier Rué - olivier.rue@toulouse.inra.fr

**Objectif :** Effectuer une recherche variants génomiques dans l'environnement GALAXY

**Données :** historique publié GALAXY (TP SNP NOV-2014)

- NC_012125.1.fasta : génome de référence (Salmonella enterica)
- ERR003037.fastq (run Illumina Salmonella enterica)
- ERR000017.fastq (run Illumina Salmonella enterica)
- SRR007327.fastq (run 454 Salmonella enterica)

**Rendu :** historique partagé dans GALAXY, les réponses au questionnaire sous forme d'un rapport par mail.

**Documentation :**

- http://www.broadinstitute.org/gatk/
- http://bio-bwa.sourceforge.net/
- http://samtools.sourceforge.net/
- http://picard.sourceforge.net/
- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

**Mise en oeuvre :**

*Effectuer un contrôle qualité des fastq en utilisant FastQC.*

1) Comment interprétez-vous les graphiques produits par FastQC ? Discutez-les brièvement.
2) Combien de séquences contiennent au moins un N pour le run ER000017 ?
3) Discutez l'intérêt (ou non) de supprimer ces séquences ?

*Mapper les reads sur notre référence avec bwa.*

4) Faites attention à l'algorithme utilisé pour le mapping en fonction de la longueur des reads.
5) Présentez brièvement ces algorithmes
6) Décrivez brièvement les avantages de l'algorithme bwa mem.
7) Mapper les trois jeux de données sur la référence.

*Analyser le format SAM.*

7) Décrivez les différentes sections et champs présents dans le format SAM.
8) Combien de reads parfaitement alignés y a-t-il dans le fichier SAM généré à partir du run ERR000017 ?
9) Pour chaque valeur de qualité de mapping, associez le nombre de reads correspondant (ex : 30 -> 1500 reads)
10) Selon vous, faut-il garder les reads avec une qualité de mapping de 0 ?

*Manipuler les fichiers BAM.*

10) Convertissez les fichiers SAM en BAM

11) Les fichiers ERR000017 et ER003037 sont issus d'un même échantillon. Mergez les en un seul fichier.

12) Combien de reads sont mappés dans ce fichier mergé ?

13) Quels outils sont à votre disposition pour obtenir des statistiques sur l'alignement ?

14) Comment jugez-vous le taux de mapping pour ces jeux de données ?

*Nettoyer les fichiers BAM (SRR et merge).*

Nous allons suivre le pipeline GATK : http://www.broadinstitute.org/gatk/guide/best-practices
Certaines fonctionnalités ne sont pas présentes dans GALAXY.

15) Retirez les duplicats PCR des fichiers BAM. Quel est l'intérêt d'une telle procédure ?

16) Ajoutez les read groups à vos deux fichiers BAM. Quel est l'intérêt d'une telle procédure ?

17) Réalignez vos fichiers BAM. Expliquez en quoi consiste les deux étapes du réalignement GATK.

18) Recalibrez vos fichiers BAM. Expliquez en quoi consiste les deux étapes de la recalibration GATK. Attention : vous devez pour le CountCovariates passer un fichier VCF en paramètre. Quelle est l'utilité de ce fichier de SNP, que vous aurez généré grâce à l'outil proposé par GATK (vous ne garderez que les 10 000 meilleurs SNP).

19) Lancez une recherche de variants (SNP + INDEL).

*Analyser le format VCF.*

20) Expliquez les différents champs du format VCF.

21) Le biologiste qui vous a demandé de traiter ses données s'intéresse à une zone particulière : les 10000 premières positions du génome. Il ne veut pas du reste. Générez les fichiers BAM, FASTA et VCF réduits.

22) Combien de SNP hétérozygotes pour les 2 échantillons se retrouvent dans sa zone d'intérêt ?

*Partagez votre historique*

23) Renommez votre historique : UPS-P5-workflow. Puis partagez-le avec le user : orue@toulouse.inra.fr

**Conseils :**

- Pensez à renommer vos datasets de façon intelligible.

# 6. Créer un wrapper GALAXY dédié à supprimer des séquences peu complexes

**Référents :**
Sarah Maman (smaman@toulouse.inra.fr)
Olivier Rué (olivier.rue@toulouse.inra.fr)

**Objectif :** Créer un wrapper galaxy pour filtrer les séquences peu complexes d'un fichier fasta ou fastq. Le script de filtre devra générer en sortie un fichier du même type que le fichier d'entrée (fasta ou fastq) avec les séquences peu complexes en moins.

**Données :** http://genoweb.toulouse.inra.fr/~formation/UPS/6/

- file.fastq
- file.fasta

Faites "copy link location" sur le lien et utilisez la commande wget pour récupérer les données dans votre espace de travail.
Veuillez travailler sur votre instance locale de Galaxy.

**Rendu :** Un rapport expliquant la démarche et une archive contenant :

- Le script de filtre des séquences (fasta ou fastq).
- Le fichier xml de configuration du wrapper.

**Documentation :**

- Ressources documentaires du Galaxy Project : http://wiki.galaxyproject.org
- Description du format fasta : http://fr.wikipedia.org/wiki/FASTA_%28format_de_fichier%29
- Description du format fastq : http://genotoul.toulouse.inra.fr/~formation/UPS/9/Nucl.%20Acids%20Res.-2010-Cock-1767-71.pdf

**Mise en oeuvre :**
Une séquence peu complexe est une séquence qui possède au moins un de ces deux critères :

- Une séquence contenant moins de 3 nucléotides différents
- Une séquence contenant plus de 4 N (base non connues)

# 7. Créer un wrapper GALAXY modèle

**Référents :**
Sarah Maman (smaman@toulouse.inra.fr)

**Objectif :** Prise en main d'une instance locale de Galaxy et de la procédure de développement de wrappers spécifiques à cet environnement de travail.

**Données :**
* Guide de bonnes pratiques pour la conception de wrappers Galaxy : http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/wrappers_bonnes_pratiques.txt
* Code source de Galaxy : http://wiki.galaxyproject.org/Admin/Get%20Galaxy

**Rendu :** un rapport expliquant la démarche, un wrapper xml et un script en perl.

**Documentation :** Ressources documentaires du Galaxy Project : http://wiki.galaxyproject.org (voir particulièrement la page "Tools Synthax").

**Mise en oeuvre :**
* Installer une instance locale de Galaxy sur votre PC. Veuiller utiliser, de préférence, un environnement Linux et l'explorateur Firefox. Les souces de Galaxy sont disponibles depuis : http://galaxyproject.org/
* Prendre connaissance du code des principaux fichiers utiles au développement de wrappers : tool_conf.xml, les fichiers xml et pl du répertoire tools/.
* Faire un état de l'art, le plus exhaustif possible, des tags xml disponibles depuis la page "Tool Synthax" du Galaxy project (http://wiki.galaxyproject.org/Admin/Tools/ToolConfigSyntax?action=show&redirect=Admin%2FTools%2FTool+Config+Syntax) mais aussi (et surtout) à l'aide du wiki de Galaxy et des archives mails.
* Concevoir et développer deux fichiers (xml et perl) pour construire un wrapper modèle qui soit le plus exhaustif possible.
* Avant de transmettre votre wrapper, veuillez le tester en local sur votre instance de Galaxy.
* Vos contraintes :
- la ligne de commande lancée par ce wrapper prend un nombre indéfini d'inputs en entrée et, si possible, génère un nombre indéfini d'outputs en sortie (http://gmod.827538.n3.nabble.com/conditional-number-of-output-files-td1687778.html).
- Suivre le guide des bonnes pratiques (http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/wrappers_bonnes_pratiques.txt) pour une meilleure intégration de votre wrapper dans une instance Galaxy.
- Le tag <help> devra lister les points suivants : "Please cite", "description", "For further informations, please visite the balbla website." (avec un lien vers le site en question), "Example of command line" et autres informations qui vous semblent utiles et pertinentes.

# 8. Développer un outil de nettoyage de données sRNAseq

**Référents :** Olivier Rué - olivier.rue@toulouse.inra.fr

**Objectif :** Développer un “outil” de nettoyage de données sRNAseq sur votre instance locale de Galaxy.

**Données :** http://genoweb.toulouse.inra.fr/~formation/UPS/9/

- file.fastq : les reads sans adaptateur

Faites “copy link location” sur le lien et utilisez la commande wget pour récupérer les données dans votre espace de travail.

**Rendu :** un rapport expliquant la démarche et un outil à envoyer par mail

**Mise en oeuvre :**

Spécifications :

Input :

- file.fastq

Paramètres :

- Fichier d’entrée (--fastq)
- Taille de la graine (--seed-size)
- Valeur moyenne de qualité Phred au sein de la graine (--mean-seed-qual)
- Nombre maximal de N dans la séquence (--N)
- Nombre de nucléotides différents (--diff-nuc-count)

Output :

- fichier nettoyé au format fastq

Exemple :

mon_script --fastq file.fastq --seed-size 20 --mean-sead-qual 25 --N 3 --diff-nuc-count 3 --repeat-count 5

Le script mon_script travaille sur le fichier file.fastq. Il supprime les séquences dont la qualité de base moyenne de la graine de 20 est inférieure à 30. Il supprime également les séquences qui contiennent plus de 3 N et les séquences contenant des répétitions d’au moins 5 nucléotides (AAAAA par exemple).

**Conseils :**

- Lisez bien les spécifications du format FASTQ avant de vous lancer.
- Le fastq est encodé au format Illumina 1.5.
- La graine est la séquence des n premières bases d’une séquence.

**Documentation :**

- format FASTQ : http://genotoul.toulouse.inra.fr/~formation/UPS/9/Nucl.%20Acids%20Res.-2010-Cock-1767-71.pdf
- encodage qualité : http://en.wikipedia.org/wiki/FASTQ_format

# 9. Conception et création de videos tutoriels sur GALAXY

**Référents:**
Sarah Maman (smaman@toulouse.inra.fr)

**Objectif :** Prise en main de l'interface Galaxy et création de tutoriels sous Wink..

**Données :** Avec vos logins/mot de passe donnés en formation (penser à demander leur ré-activation en amont du mini-projet) :
* Accès à l'instance Sigenae de Galaxy : http://sigenae-workbench.toulouse.inra.fr
* Accès aux pages d'e-learning de Sigenae: http://sig-learning.toulouse.inra.fr/

**Rendu :** un rapport expliquant la démarche, les vidéos et les commentaires associés à chaque tutoriels.

**Documentation :** Ressources documentaires du Galaxy Project : http://wiki.galaxyproject.org. Voir particulièrement les pages "Support" (http://wiki.galaxyproject.org/Support) et "Tutorials" / "Videos".

**Mise en oeuvre :**
* Installer Wink (gratuiciel)
* Il n'est pas du tout nécessaire d'installer une instance locale de Galaxy sur votre PC.
* Veiller à demander un compte de formation ou un compte sur Genotoul (http://bioinfo.genotoul.fr/index.php?id=81) afin d'accéder à l'instance Galaxy de l'équipe Sigenae Toulouse.
* Prendre en main quelques outils Galaxy identifiés par une astérisque (à vous de voir combien de tutoriels vous pouvez créer). Attention, certains outils ont déjà une vidéo associée, ils sont repérés par l'information "(e-learning available)".
* Concevoir chaque video sous Wink avec des bulles explicatives.
* Veiller à prendre comme exemple les videos disponibles depuis les pages d'e-learning de Sigenae et à ne pas refaire un tutoriel déjà existant dans l'e-learning de Sigenae.

# 10. Poursuivre son wrapper du TP

**Référents:**
Sarah Maman (smaman@toulouse.inra.fr)

**Objectif :** Publier votre wrapper sur le Galaxy ToolShed international.

**Données :** Ce travail doit être réalisé sur l'instance Galaxy de test (/usr/local/bioinfo/src/galaxy/galaxy-dist/) avec un compte de formation. Seuls le répertoire tools/initial_VotreTool/ vous sera accessible.

**Rendu :** Votre code commenté, versionné, et un README.

**Documentation :** Ressources documentaires du Galaxy Project : http://wiki.galaxyproject.org.

**Mise en oeuvre :**
Des indications et code type vous ont été donné en cours et en TP. Vous pouvez choisir votre langage de programmation. N'oubliez pas de tester vos lignes de commande et wrapper en local avant de demander une mise en ligne.