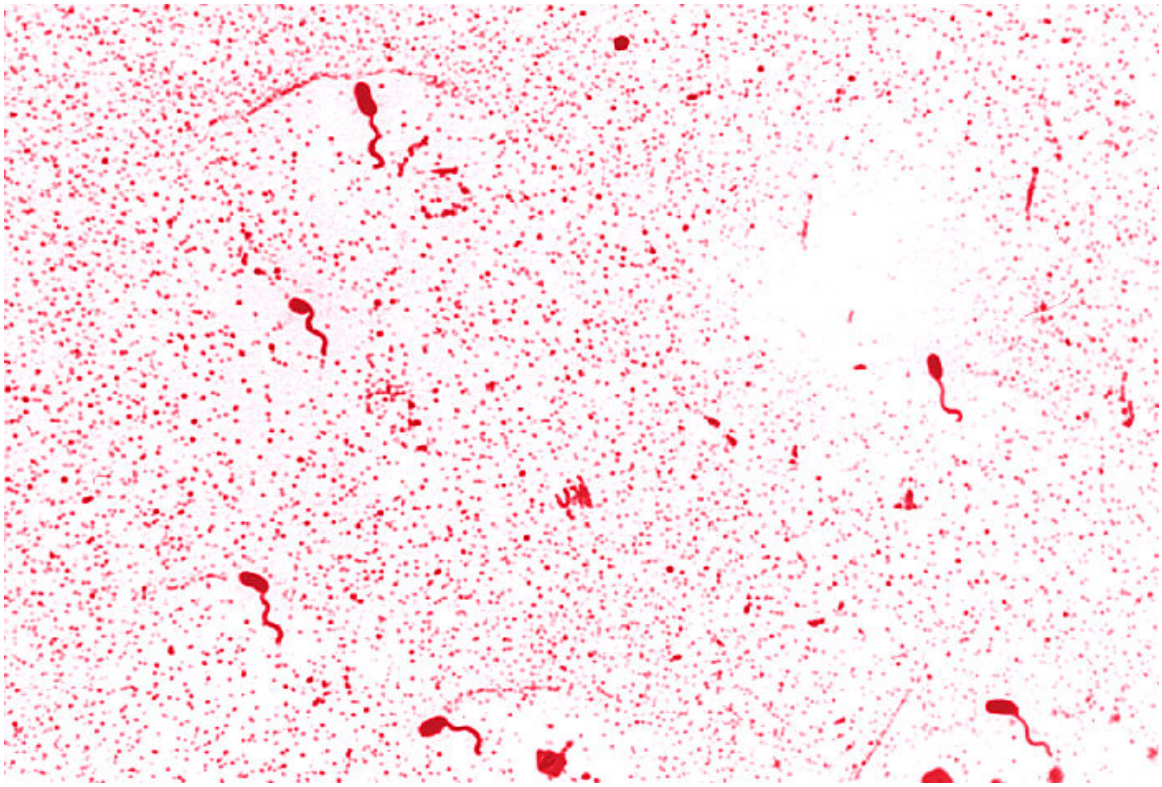
3I019 : Travaux Pratiques

Projet Analyse de génomes



Responsables. Hugues Richard (hugues.richard@upmc.fr) et Juliana Silva-Bernardes (juliana.silva_bernardes@upmc.fr).

Enseignants TP. Yasaman Karami (yasamankarami@gmail.com) et Ari Ugarte (ari.ugarte@gmail.com)

Le but de ce mini-projet est d'analyser de manière comparative le génome de deux bactéries : un organisme déjà bien connu *E. coli* (souche K12, qui n'est pas pathogène) et une bactérie pathogène des Entérococques (deux bactéries sont données au choix).

Le but du projet est d'analyser les propriétés générales du génome, de détecter les différentes régions de composition homogène, de pouvoir trouver les éléments génétiques qui lui confèrent sa toxicité, et de comprendre comment cette toxicité a été acquise.

Ce mini projet est organisé en trois parties successives pour pouvoir analyser les propriétés des génomes :

Partie A : propriété globales des génomes et détection des hétérogénéités dans la séquence.

Partie B : Analyse comparative, détection des îlots de pathogénicité, analyse des propriétés, annotation des gènes correspondants.

Partie C : Analyse de l'usage des codons et lien avec les propriétés des gènes et les îlots de pathogénicité.

Les deux génomes d'Entérobactéries sont les suivants :

- Groupe 1 : *Shigella dysenteriae*, une entérobactérie pathogène rencontrée chez l'homme (responsable de la dysenterie bacillaire, ou de la shigellose). Cette bactérie provoque des infections intestinales localisées essentiellement au gros intestin où les germes se multiplient en provoquant une inflammation de la muqueuse se traduisant par une diarrhée glaireuse et sanguinolente.
- Groupe 2 : Vibrio cholerae, est une bactérie gram négatif responsable du choléra chez l'Homme.

Partie A : Annotation des génomes

Dans un premier temps, nous allons utiliser les fonctions et les méthodes développées durant les TME précédents pour faire une première annotation des génomes.

Chaque groupe travaille avec le génome de *E. coli* et au choix le génome de *Shigella dysenteriae* ou *Vibrio cholerae* (juste un des deux, pas les deux !).
Vous pouvez télécharger les séquences sur le site du module.

Préliminaires, propriétés de base

1. Après avoir téléchargé les deux génomes que vous analyserez par la suite, produire une liste des premières propriétés en rapportant dans une table pour l'organisme choisi (cf TME3)
    a. Le nombre de chromosomes et de plasmides, leurs longueurs, la longueur totale du génome, le pourcentage en GC global et la composition en nucléotides.
    b. Découpez le génome en blocs 1kbp non chevauchants et calculer de % en GC dans chaque bloc. Faites un histogramme de la distribution du GC, que remarquez vous ? Est ce que les blocs avec un GC atypique ont tendance à être colocalisés ?

2. Annotons l'ensemble des ORFs du génome choisi avec la méthode vue au TME4 et comparons le résultat avec l'annotation disponible (fichier tab). Nous utiliserons un programme d'annotation automatique des gènes comme Glimmer.
    a. Produire un fichier d'annotation des gènes en utilisant Glimmer.
    b. Faites un histogramme des longueurs des gènes prédits avec Glimmer et un histogramme des longueurs des gènes du fichier tab. Y a-t-il des différences ?
    c. Adapter la fonction compare_intervalle pour comparer les annotations obtenues avec Glimmer avec le fichier tab d'annotation fourni sur les deux brins et renvoyer une matrice de confusion. Donnez les sensibilités et spécificités de la prédiction.

On travaille à partir de maintenant et pour toute la suite avec les annotations du fichier tab.

    d. Extraire les séquences des gènes et les écrire dans un fichier fasta `genes.fasta`. Pour cela vous utiliserez les fichiers avec les extensions .tab et .genome (cf TME4). Dans les en-tête du fichier fasta, rapportez comme identifiant du gène la colonne "GeneID" du fichier .tab.
    e. En analysant le fichier `genes.fasta`, on remarque qu'il y a des séquences qui possèdent un ou plusieurs stop codon en phase ou qui ne commencent pas par un start codon. Nous les analyserons plus tard. Pour l'instant séparez ces gènes dans un autre fichier que nous appellerons `genes_non_codants.fasta` et mettez dans le fichier `genes_codants.fasta` les gènes codants.
    f. A partir du fichier (`genes.fasta`) calculer le pourcentage en GC de chaque gène. Mettez à jour le fichier .tab en rajoutant une colonne à droite nommée "percGC".
    g. Faites un histogramme des pourcentages en GC des gènes. Que remarquez vous ? Comparez avec l'histogramme obtenu sur *coli*, commentez. Pouvez

vous faire le lien avec l'histogramme obtenu à la question 1.b ?

Annotation par homologie avec BLAST

Nous allons maintenant tenter d'annoter la fonction des gènes du génome choisi en utilisant d'abord l'annotation du génome de *E. coli*.

3. Traduisez les séquences des gènes codants (`genes_codants.fasta`) en séquences protéiques et faites un blast de toutes les protéines du génome choisi contre les protéines de *E. coli.* Quel type de BLAST doit-on utiliser ?
    a. A partir des résultats, fixer un seuil de E-value (10^{-3} par exemple) et ne garder que les alignements au dessous du seuil.
    b. Pour chaque alignement, vérifier la couverture de la séquence query et éliminer les alignements qui couvrent moins 80% du gène.
    c. Sauvegardez dans le fichier `correspondance.blast`, les alignements qui sont gardés après les filtrage effectués aux questions a et b.
    d. Rajoutez dans le fichier tab le nombre d'alignements significatifs (0,1,2..) obtenus pour chacun des gènes. A quoi correspondent les cas où un gène a plus d'un alignement ?
    e. Combien de gènes n'ont aucun alignement ? Faire un histogramme de leur pourcentage en GC, que remarquez vous ?
4. Alignons maintenant les séquences des gènes non codants (fichier `genes_non_codants.fasta`)
    a. Faire un alignement avec blastn sur la base de données nr.
    b. Pour quelles familles de gènes obtenons nous des alignements ?

Assignation des catégories fonctionnelles pour les gènes codants avec COG.

COG est une base de donnée de groupes de protéines orthologues. Ces groupes ont été générés en comparant les séquences protéiques de génomes complets, les groupes similaires ont été ensuite regroupés dans les classes fonctionnelles (voir la table ci-dessous).
Nous avons déjà détecté les COGs pour *E. coli* et les avons rangés dans le fichier `Escherichia_coli_COG_annotation.tsv`. Vous allez maintenant détecter les COGs pour le génome choisi en le comparant avec les annotations COG de *E. coli*.

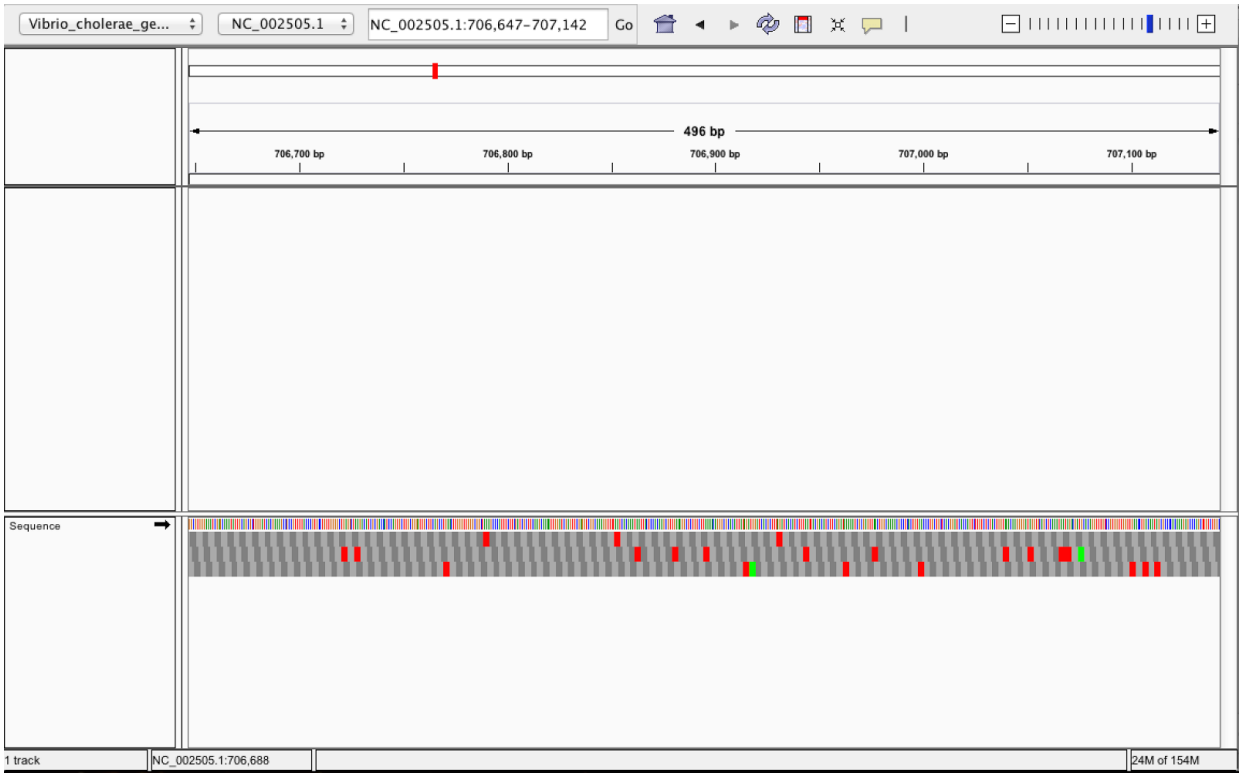
Code COG	Classe fonctionnelle
A	RNA processing and modification
B	Chromatin Structure and dynamics
C	Energy production and conversion
D	Cell cycle control and mitosis
E	Amino Acid metabolism and transport
F	Nucleotide metabolism and transport
G	Carbohydrate metabolism and transport
H	Coenzyme metabolis
I	Lipid metabolism
J	Tranlsation
K	Transcription

L	Replication and repair
M	Cell wall/membrane/envelop biogenesis
N	Cell motility
O	Post-translational modification, protein turnover, chaperone functions
P	Inorganic ion transport and metabolism
Q	Secondary Structure
T	Signal Transduction
U	Intracellular trafficing and secretion
Y	Nuclear structure
Z	Cytoskeleton
R	General Functional Prediction only
S	Function Unknown

5. A partir du fichier `correspondance.blast` récupérez pour chaque alignement l'identifiant du gène de *E. coli* et cherchez dans le fichier `Escherichia_coli_COG_annotation.tsv` le COG ID et la catégorie correspondent (colunnes COG.id et COG.function).
6. Ajoutez une colonne supplémentaire à `correspondance.blast` avec les deux informations : COG.id et COG.function. Combien de gènes ne possèdent pas de classe fonctionnelle, ou sont classés dans la classe S.
7. Faites un histogramme de classes fonctionnelles de COG pour le génome choisit. Quelles sont les classes les plus abondantes?

Visualisation et navigation dans un génome, analyse de la composition

8. Il sera intéressant de pouvoir visualiser les prédictions que nous ferons en face de l'information génomique et des différents types d'annotation. Pour cela nous allons utiliser un navigateur de génome (*genome browser* en anglais). Nous allons utiliser l'outil IGV développé à l'institut BROAD, qui a une prise en main relativement intuitive. Tous les détails sur le logiciel sont disponibles ici : http://software.broadinstitute.org/software/igv/userguide
   Pour lancer le programme, taper igv dans une fenêtre de terminal. Le programme s'installe facilement sur tous les systèmes d'exploitation (http://software.broadinstitute.org/software/igv/download)
   a. Charger le génome de l'organisme que vous étudiez (Vibrio_cholerae_genome.fasta ou Shigella_dysenteriae_genome.fasta) en allant dans
      Genomes → Load genome from file…
      Vous obtenez une vue du type montré en dessous, avec de haut en bas:
      - les informations de navigation: génome et chromosome utilisé, coordonnées de la région, boutons de navigation.
      - une graduation de la région, celle ci est navigable avec la souris
      - une case blanche où différents niveaux d'information peuvent être ajouté
      - l'information de la séquence.



Vous pouvez changer les options d'affichage pour chacun des élément en faisant un clic droit sur la région et en regardant les options.

b. Zoomer sur une zone d'un millier de bp, essayez de lire la séquence génomique à l'écran, puis cliquez sur Regions-> Region Navigator -> add, ça vous permet de sauvegarder une région d'intérêt pour pouvoir l'analyser plus tard.
c. Chargeons maintenant le fichier d'annotation pour le génome, l'annotation est disponible dans les fichiers liens ci dessous :
   - Vibrio cholerae
   ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/745/GCF_000006745.1_ASM674v1/GCF_000006745.1_ASM674v1_genomic.gff.gz
   - Shigella dysenteriae
   ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/012/005/GCF_000012005.1_ASM1200v1/GCF_000012005.1_ASM1200v1_genomic.gff.gz
   Ces fichiers contiennent la même information que celle du fichier tab dans un format lisible par IGV, et peuvent être chargés avec la commande File → Load from file… ou → Load from URL.
b. Vous pouvez sauvegarder votre configuration en cours en faisant File → Save session

9. Nous allons maintenant étudier la composition en nucléotides le long du génome en générant des fichiers qui seront chargés dans le navigateur.
   Nous allons coder l'information dans un fichier au format igv (à sauvegarder avec l'extension .igv) qui est simplement un tableau avec les colonnes séparées par des tabulations et formaté de la manière suivante (la colonne Feature n'est pas importante)

Chromosome	start	end	Feature	percGC	percA	percT
NC_007606.1	0	100	F1	68	15	17
NC_009344.1	210	300	F2	45	35	20

Ceci est un exemple pour deux annotations sur le chromosome de Shigella (NC_007606.1: de la première à la 100è base) et un de ses deux plasmides (NC_009344.1: bases 211 à 300) avec trois informations données en même temps.

Quelques détails auxquels il faut faire attention :

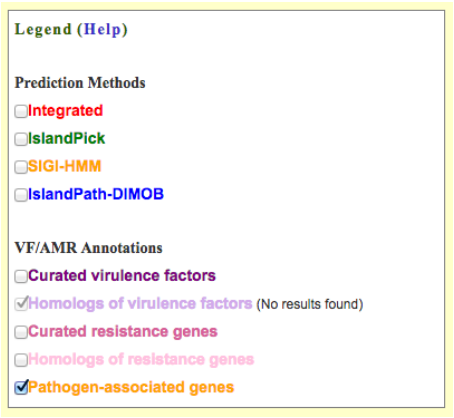
- le nom utilisé pour les chromosomes doit être exactement le même que celui donné dans les en-têtes du fichier fasta pour que l'affichage ait lieu.
- la première coordonnée sur un chromosome est la coordonnée 0. La dernière coordonnée d'une région est toujours **exclue.** Cette seconde convention simplifie le calcul de la longueur d'une région (et est similaire à python).
- Le fichier doit être trié par chromosome puis suivant les positions de début des régions.

a. Adaptez la fonction codée pour la question 1.b pour qu'elle renvoie un fichier au format IGV qui contient le pourcentage en GC donné par blocs de 50bp dans le génome.
b. Générez un second fichier qui contient les % en GC pour les gènes annotés et rajoutez le à votre visualisation. Vous pouvez à ce niveau jouer un peu avec les options (track height, color, scale...) pour rendre l'affichage plus agréable et pertinent.
c. Chargez le fichier dans IGV et visualisez quelques régions. Partez de l'analyse faite à la question 1.b et sélectionnez une dizaine de régions qui ont des pourcentages en GC atypique (plus bas ou plus haut que la moyenne).
d. Analysez les annotations des gènes et les catégories COGs pour les éléments qui sont dans ces régions. Que remarquez vous pour les annotations des gènes ?
e. Regardez les annotations aussi autour des régions.

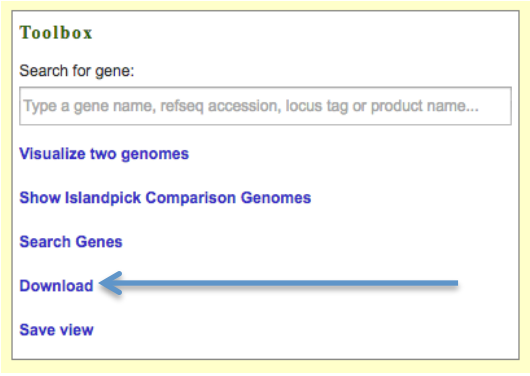
B – Annotation plus automatique des îlots de pathogénicité

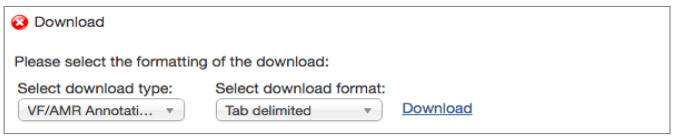
Nous allons utiliser le logiciel IslandViewer 3 pour détecter les ilots de pathogénicité dans le génome choisi.

a. Aller sur le site http://www.pathogenomics.sfu.ca/islandviewer/browse/ et chercher votre génome, Cliquez sur *"go to genome"*.
b. choisir uniquement Pathogen-associated genes



c. téléchargez les région (download of predicted regions)





d. Pour chaque gene dans le fichier _annotations.cvs du type PAG faite:
   1. Un BLAST contre E. coli
   2. Un BLAST contre tous les Bactéries
   3. Faire un tableau comparatif, quel est votre conclusion ?

Pour trouver les séquences de gènes taper leurs identifiant dans https://www.ncbi.nlm.nih.gov

Attention : Comme Vibrio cholerae O1 a deux chromosomes, vous devez faire les étapes a-d deux fois.