Tri de données – TP de Bioinformatique
Partie 2 – A l'échelle du génome

Licence professionnelle Génomique – automne 2015

Résumé

Ce deuxième TP se veut une continuation logique du premier TP que vous avez réalisé, en passant de l'échelle génétique à l'échelle génomique.
Les consignes concernant le rapport et la manière d'appréhender le TP restent les mêmes que précedement.
Le but scientifique de ce TP est de vous familiariser avec l'utilisation de logiciels de visualisation de données génomiques liée à différentes questions biologiques.

Documents nécessaires :
Tous les documents dont vous aurez besoin sont disponibles sur mon GitHub, un service web d'hébergement et de gestion de développement de logiciels, qui se veut aussi utile pour le partage de fichiers. Pour y accéder, taper l'adresse dans votre navigateur web : `https://github.com/tdelhomme`. Dans l'onglet "Popular Repository", cliquer sur "TP_bioinfo_L3_genomique". Les fichiers disponibles sont listés au centre de la page, vous pouvez les télécharger un par un en cliquant droit sur le lien du fichier, puis en cliquant sur "Save link as...", mais aussi les télecharger sous forme de zip (dossier compressé contenant tous les fichier), en cliquant sur "Download ZIP sur la droite de votre écran.

Exercice 1 : UCSC Genome Browser

Rendez vous sur la page d'accueil de l'UCSC Genome Browser, ici : `http://genome.ucsc.edu/`, en cliquant sur l'onglet *Genomes*.

Qu'est-ce-que l'UCSC ?
Qu'est ce qu'un "genome browser" ? En existe-il d'autre que celui maintenu par l'UCSC ?

Les SNP du gène Brca1 chez la souris

Qu'est-ce-qu'un SNP ?
Par définition, les SNPs peuvent tomber soit dans les gènes soit dans les régions intergéniques, et si dans les gènes, soit dans les régions codantes soit dans les régions non codantes.
Décrivez l'effet des SNP suivant leur localisation dans le génome.

Dans cette partie du TP, nous allons étudier un gène particulier chez la souris, Brca1.
Décrivez brievement la fonction de ce gène, ainsi que son lien avec la prédisposition au cancer.

A l'aide du "genome browser" de l'UCSC, lancez une requète pour visualiser la séquence du gène Brca1 chez la souris.
Dans la base de données de l'UCSC, on peut noter qu'il existe plusieurs annotation du même gène Brca1, combien au total ? qu'est-ce-qui différencie les deux premières annotations ? Pour répondre, cliquer sur le nom de chacune des annotations pour être rediriger vers une fiche détaillée.

A présent, nous allons discriminer les différents SNPs répértoriés chez Brca1.
Chaque SNP est représenté par son identifiant (rs...) et par un trait vertical le situant sur le génome.

En bas de la page, dans l'onglet *Variations and Repeats*, cliquez sur *Common SNPs(138)*.
Maintenant, changer les options des couleurs afin de faire apparaître les SNP synonymes en bleu et les non-synonymes en vert (veillez à ce que le champ "SNP feature for color specification" soit mis à "function"). Cliquer ensuite sur *submit* pour charger les modifications.

Que pouvez-vous remarquer quant à la disposition de ces deux types de SNPs ?
A quels SNPs correspondent les identifiants restés en couleur noire ?
Les SNP bleux et verts sont-ils plus ou moins fréquents que les noirs ? Pourquoi ?

Exercice 2 : CGView - Visualisation de génomes

Les Genome Browsers sont très utiles pour visualiser des génomes, mais pour une échelle réduite, de quelques milliers de paires de bases tout au plus. Au dela de ça, ce genre d'utilitaire n'est pas pratique, comme par exemple dans le cas de petits génomes qu'il peut être utile de visualiser dans leurs ensembles. Pour cela nous allons maintenant utiliser le logiciel CGView.

Rendez vous sur le serveur du software CGView, qui va nous permettre d'utiliser le logiciel directement en ligne : `http://stothard.afns.ualberta.ca/cgview_server/`.

Décrivez brièvement la fonction de CGView, et donner un exemple d'utilisation pour une étude biologique.
Récupérer sur mon github les 4 fichiers qui vont être nécessaires à l'exercice :

- NC_001823.gbk
- H_sapiens.fasta
- E_coli.fasta
- A_thaliana.fasta

NC_001823.gbk est l'annotation GenBank du génome mitochondrial de *Reclinomonas americana*, un petit protozoaire, bien connu (tout est relatif) car propriétaire du premier génome mitochondrial séquencé, début 90's.
A_thaliana.fasta contient les séquences des protéines mitochondriales d'*Arabidopsis thaliana*, une espèce de plantes.
E_coli.fasta contient les séquences de la bactérie *Escherichia coli* souche K12.
Pour finir, *H_sapiens* contient les séquences des protéines mitochondriales chez l'Homme.

1) Nous allons à présent représenter circulairement le génome mitochondrial du protozoaire, ainsi que les régions que nous possédont dans nos fichiers fasta.
Sur la page du serveur de GCSkew, rentrer une adresse mail valide (pas d'inquiétude c'est juste pour l'envoi des résultats), et uploader la séquence de *Reclinomonas americana* au format GenBank.
En vous référant à la table 1 à la fin de l'énoncé du TP, modifier les paramètres correspondants (les paramètres en question appartiennet aux onglets "BLAST analysis" et "Output settings").

Cliquez ensuite tout en bas de la page à droite sur "submit".
Dans quelques minutes vous allez recevoir votre image par mail. Je vous conseille donc de bien vérifier vos paramètres avant l'envoi de la requête pour ne pas recommencer cette étape plusieurs fois.

En attendant vos résultats vous pouvez lire le papier du séquençage du génome mitochondrial de *Reclinomonas americana* ici : `http://www.ncbi.nlm.nih.gov/pubmed/9168110`

Intégrer dans votre rapport l'image que vous avez créer. Et pour finir, discuter des résultats (5-10 lignes). Pour cela vous pouvez vous renseigner sur le web ou tout autre source.

TABLE 1 – Paramètres de GC skew (1)

Sequence file : NC_001823.gbk (format : genbank, length : 69034)
Global Blast Settings : query_split_size=50000, overlap_split_size=0
Blast 1 : A_thaliana.fasta, blastx, expect=0.00001, Standard, filter=Yes
Blast 2 : E_coli.fasta, blastx, expect=0.00001, Bacterial and Plant Plastid, filter=Yes
Blast 3 : H_sapiens.fasta, blastx, expect=0.00001, Standard, filter=Yes
Map title : Untitled
Show GenBank/EMBL features : Yes
Show GC content : No
Show GC skew : No
Draw divider rings : Yes
Use opacity for BLAST hits : Yes
Show labels : Yes
Show ORFs : No
Show start and stop codons : No
Show blast hits by reading frame : No
Show legend : Yes
Font size : small
Tick density : 0.6
Draw features as : Arrows
View : Full size