**TD - Annotation**

Jinxin Wang, Olga Permiakova

**A : Assemblage des différents contigs**

1. Rechercher les régions chevauchantes entre les différents contigs et assemblez la séquence du chromosome dans un fichier '*chromosome.fasta*’. Vous ferez attention au sens de chaque contig et aux zones de chevauchement.
2. Indiquer l’ordre des contigs dans le chromosome assemblé, leurs zones de chevauchements, et leur orientation
3. Préciser les lignes de commande BLAST utilisées.

**B : Recherche des ORF**

1. Quelle est la différence entre une ORF et une CDS ? Une CDS correspond-elle à un gène ?

ORF est une séquence des nucléotides d'ADN qui commence par un start-codon (AUG) et elle contient des codons codant potentiellement une proteine. Une ORF est un assez long fragmen codant qui n’a pas de stop-codon. Les CDs se trouvent dans ORF. A la difference d'ORF, les CDs se termine par un codon-stop, dans plupart des cas. Une CDs correspond à un gène codant une proteine.

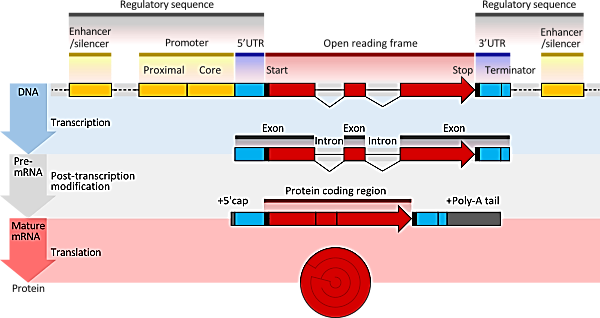
1. Open Reading Frame

Avec Artemis on a obtenu 153 ORF. La densité d’ORF est 2 ORF/kb.

|  |  |  |
| --- | --- | --- |
|  | La longueur moyenne | Le % GC moyen |
| ORF | 643 | 11,62% |
| Régions intergéniques | 338 | 8,21% |
| Chromosome | 77334 | 9,85% |

1. Quels gènes « échappent » à cette annotation ?

Utilisant cette annotation on considère seulement des fragments de la sequence qui codent des proteins. Ces sequences n'ont pas d'introns. Il y a aussi des gènes qui codent des ARN (ARNr, ARNt).



**D : Recherche des paralogues**

1. Qu’est ce que des gènes paralogues ?

Gènes paralogues sont des gènes homologues qui sont issus d’un évènement de duplication dans une même espèce. Ils peuvent effectuer des tâches différentes dans la cellule ou devenir des pseudogènes s’ils ont perdue la fonctionnalité, mais ils ont conservé une similarité de séquence.

1. Quels critères peut-on utiliser pour définir des paralogues ?

Pour définir des paralogues on utilise la E-value. Elle diminue lors que la similarité des gènes augmente. Pourtant des séquences courtes ont des E-values élevées parce qu'elles se produisent plus probablement dans base des données.

1. Définir des valeurs seuils pour chacun des critères utilisés

On conclure que deux genes sont paralogues si E-value est inférieur à .