**TD GENETIQUE : Recherche d’information dans les génomes**

**Sujet : Recherche de séquences codantes dans le génome de la levure *Candida glabrata***

*Candida glabrata* est un champignon unicellulaire pathogène opportuniste qui provoque, au niveau du tractus urogénital, des infections, chez les individus immunodéprimés (HIV positifs, transplantés, patients soumis à une chimiothérapie…).

Le génome de *Candida Glabrata* a été séquencé pour la première fois en 2004. Ce génome reste encore mal annoté puisque, pour plus de 90% des séquences codantes identifiées par des méthodes informatiques, la fonction de la protéine correspondante n’est pas connue. En comparaison, 80% des séquences codantes de la levure de boulangerie *Saccharomyces cerevisiae*, qui est un modèle très étudié, ont une fonction connue.

Il reste donc encore beaucoup de travail aux biologistes et aux bioinformaticiens pour pleinement comprendre la biologie du génome de *Candida glabrata*.

Dans ce TD, nous vous proposons d’explorer ce génome avec des méthodes classiques de recherche de gènes codants.

1. **Base de données NCBI sur le génome de *Candida glabrata***

**Base de données génomiques NCBI :** [**https://www.ncbi.nlm.nih.gov/genome**](https://www.ncbi.nlm.nih.gov/genome)

1. **Découverte des informations sur le génome de *Candida glabrata***

Utiliser la barre de recherche pour accéder aux données du génome de *Candida glabrata*

**Questions/Discussions :**

1- Combien de molécules d’ADN constituent le génome de *Candida glabrata*

2- Analyse des données pour le chromosome A : Combien de gènes sont présents sur le chromosome A ? Donner le nombre de gènes codants et de gènes non codants présents sur ce chromosome. Calculer le nombre de gènes codants présents par kb (kiloBase) sur le chromosome A.

1. **Format de la séquence chromosome A de Candida glabrata**

Cliquer sur le lien RefSeq du chromosome A puis afficher la séquence de ce chromosome au format FASTA.

**Questions/Discussions :**

1- Quelles sont les caractéristiques du format FASTA ?

2 - Expliquez pourquoi une unique chaine de caractères est suffisante pour décrire la séquence d’ADN du chromosome A

3 - Copier les 10 premiers nucléotides de la séquence du chromosome A en indiquant l’orientation de cette séquence.

4 - Ecrire la séquence des 10 premiers nucléotides du chromosome sous forme double brin en conservant l’orientation du brin 1 donnée par la base de données. La séquence obtenue pour le brin 2 est dite **complémentaire** par rapport au brin 1.

5 - Ecrire au format Fasta la séquence du brin 2. Cette séquence est dite **reverse-complémentaire**.

6 - Utiliser l’outil en ligne de [convertion de séquence](http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html#iupacdegeneracies) pour vérifier votre travail sur la séquence des 10 premières bases du chromosome A. A quoi correspondent les opérations « reverse », « complément » et « reverse complément » proposées par cet outil.

1. **Analyse du chromosome A de *Candida glabrata***
   1. **Recherche des séquences codantes sur le chromosome A de Candida glabrata.**

Outil de recherche de séquence : <https://www.ncbi.nlm.nih.gov/orffinder/>

Lien pour les informations sur le [code génétique](https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG1) utilisé par l’outil ORF finder.

Utiliser l’outil de ORF FINDER de NCBI pour rechercher des séquences codantes sur le chromosome A de *Candida glabrata*. Vous pouvez soit donner en entrée le numéro accession de ce chromosome (NC\_005967.2), soit copier la séquence Fasta associée.

**Pour cette recherche vous limiterez les coordonnées à analyser à la région située entre les coordonnées 50 000 et 75 000 du chromosome A.**

**Dans un premier temps lancer cette analyse avec les paramètres par défaut.**

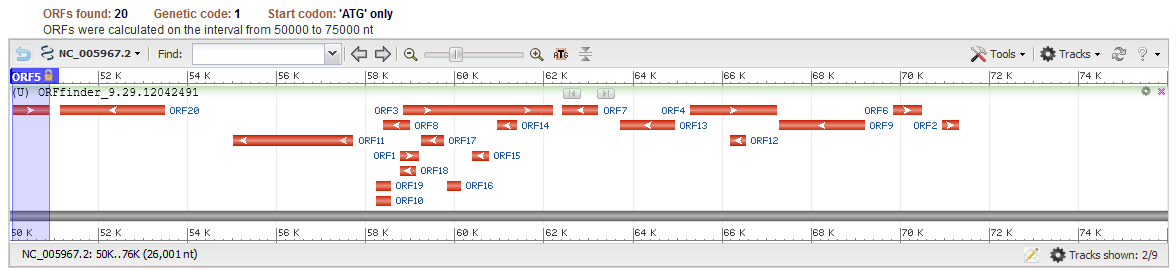
Cliquer sur le lien **Six-Frame Translations>** **Add six-frame translation track** situé en bas de la figure montrant la détection des ORF.

**Questions/Discussions :**

1- Observez la représentation Six-Frame Translations dans laquelle l’analyse de la distribution des codons initiateurs (barres vertes) et Stop a été effectuée sur la séquence. Pourquoi cette analyse est-elle faite sur 6 cadres de lecture ? Comment détecter des séquences codantes (CDS) à partir de cette représentation ?

2- Observer le résultat de l’analyse ORF FINDER. Combien de CDS ont été détectées avec les paramètres par défaut ? Que pensez-vous de cette analyse compte tenu du nombre moyen d’ORF par kb calculé précédemment ?

3- Quels paramètres pourraient être modifiés pour augmenter l’efficacité de détection des CDS ? Refaire l’analyse en modifiant les paramètres de la recherche qui vous semblent importants. Donnez les paramètres choisis pour obtenir la table de détection ci-dessous.

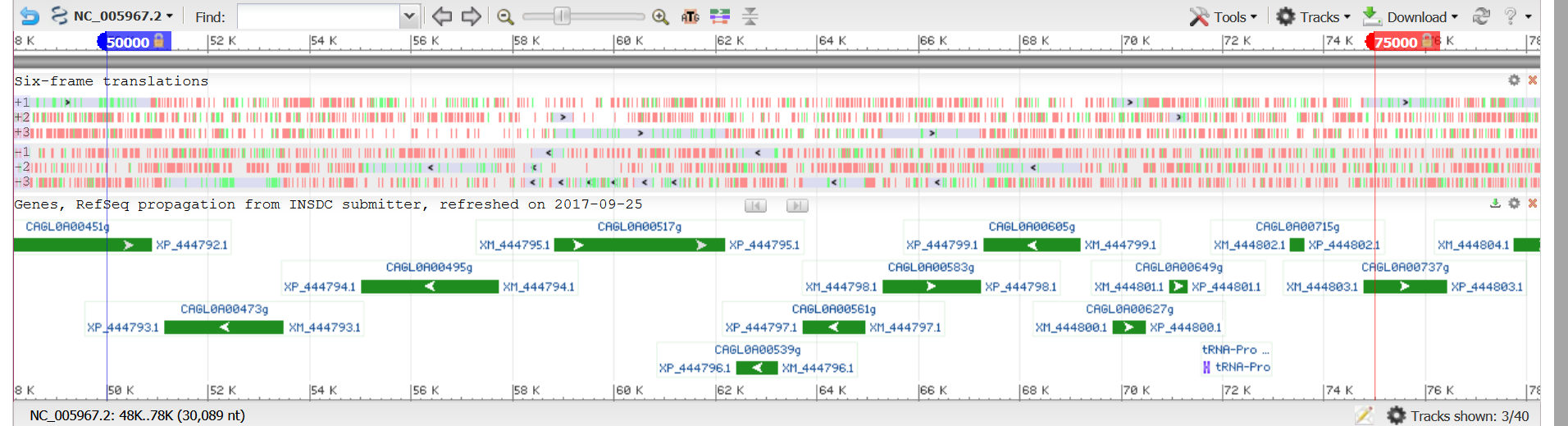


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Numéro** | **Brin** | **Cadre de lecture** | **Début** | **Fin** | **Taille (nt |aa)** |
| ORF5 | + | 1 | 50068 | 50880 | 813 | 270 |
| ORF20 | - | 3 | 53497 | 51137 | 2361 | 786 |
| ORF11 | - | 2 | 57722 | 55014 | 2709 | 902 |
| ORF19 | - | 3 | 58555 | 58220 | 336 | 111 |
| ORF10 | - | 2 | 58562 | 58230 | 333 | 110 |
| ORF1 | + | 2 | 58775 | 59191 | 417 | 138 |
| ORF3 | + | 3 | 58830 | 62198 | 3369 | 1122 |
| ORF8 | - | 1 | 58995 | 58378 | 618 | 205 |
| ORF18 | - | 3 | 59119 | 58760 | 360 | 119 |
| ORF17 | - | 3 | 59749 | 59249 | 501 | 166 |
| ORF16 | - | 3 | 60127 | 59813 | 315 | 104 |
| ORF15 | - | 3 | 60766 | 60386 | 381 | 126 |
| ORF14 | - | 3 | 61393 | 60947 | 447 | 148 |
| ORF7 | - | 1 | 63225 | 62416 | 810 | 269 |
| ORF13 | - | 3 | 64948 | 63710 | 1239 | 412 |
| ORF4 | + | 3 | 65286 | 67226 | 1941 | 646 |
| ORF12 | - | 3 | 66538 | 66188 | 351 | 116 |
| ORF9 | - | 2 | 69203 | 67281 | 1923 | 640 |
| ORF6 | + | 1 | 69832 | 70485 | 654 | 217 |
| ORF2 | + | 2 | 70937 | 71320 | 384 | 127 |

4- A partir de cette détection, tester l’option « Ignored Nested ORFs ». Sur la figure et la table précédente, indiquez les ORF impactées par cette option.

* 1. **Comparaison de la prédiction obtenue avec ORF FINDER avec l’annotation du génome.**

Site de visualisation d’annotation de génome : <https://www.ncbi.nlm.nih.gov/projects/sviewer/>

Visualisation centrée sur la région [[50000-75000] du chromosome A de Candida Glabrata](https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=NC_005967.2&tracks=%5bkey:sequence_track,name:T727902,display_name:Sequence,id:T727902,dbname:GenBank,category:Sequence,subcategory:Assembly,annots:NA,ShowLabel:false,ColorGaps:false,shown:true,order:1%5d%5bkey:six_frames_translation,name:T727920,display_name:Six-frame%20translations,id:T727920,dbname:GenBank,annots:Six-frame%20translation,ShowOption:All,OrfThreshold:100,HighlightCodons:true,AltStart:false,shown:true,order:38%5d%5bkey:gene_model_track,name:T765810,display_name:Genes\,%20RefSeq%20propagation%20from%20INSDC%20submitter\,%20refreshed%20on%202017-09-25,id:T765810,dbname:SADB,category:Genes,subcategory:NCBI%20Genes,annots:NA000139222.1,Options:MergeAll,CDSProductFeats:false,NtRuler:true,AaRuler:true,HighlightMode:2,ShowLabel:true,shown:true,order:41%5d&assm_context=GCF_000002545.3&mk=50000|50000|blue|9,75000|75000|red|9&v=46197:77574&c=993366&select=null&slim=0) 

**Questions/Discussions :**



Le site de visualisation des données d’annotation montre les CDS pour lesquelles un ARN a pu être détecté expérimentalement. Le cadre six frames translation montre en grisée les CDS de plus de 300 nucléotides détectés de manière informatique.

1- Comparez l’annotation présentée à la détection effectuée précédemment. Annoter la figure ci-dessus pour mettre en valeur : (1) les CDS correctement détectées, (2) les CDS détectées mais qui n’ont pas été confirmées par l’annotation, (3) Les CDS non détectées, (4) les gènes non codants annotés.

2- Faire un bilan des différences observées entre la prédiction obtenue avec ORF FINDER et l’annotation du génome à partir de données expérimentales ? Comment expliquer ces différences ?

1. **Analyse du chromosome mitochondrial de *Candida glabrata***
   1. **Observations de gènes annotés sur le chromosome mitochondrial**

Annotation du chromosome mitochondrial de *Candida glabrata* sur NCBI viewer : [Accession NC\_004691.1](https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=NC_004691.1&tracks=%5bkey:sequence_track,name:T727902,display_name:Sequence,id:T727902,dbname:GenBank,category:Sequence,subcategory:Assembly,annots:NA,ShowLabel:false,ColorGaps:false,shown:true,order:1%5d%5bkey:six_frames_translation,name:T727920,display_name:Six-frame%20translations,id:T727920,dbname:GenBank,annots:Six-frame%20translation,ShowOption:All,OrfThreshold:100,HighlightCodons:true,AltStart:false,shown:true,order:37%5d%5bkey:gene_model_track,name:T765810,display_name:Genes\\,%20RefSeq%20propagation%20from%20INSDC%20submitter\\,%20refreshed%20on%202017-09-25,id:T765810,dbname:SADB,category:Genes,subcategory:NCBI%20Genes,annots:NA000139222.1,Options:MergeAll,CDSProductFeats:false,NtRuler:true,AaRuler:true,HighlightMode:2,ShowLabel:true,shown:true,order:40%5d&assm_context=GCF_000002545.3&v=1:20063&c=CCFFCC&select=null&slim=0)

**Questions/Discussions :**



1 -Combien de gènes codants sont présents sur le génome mitochondrial de *C. glabrata* ?

2 -Le gène COX1 est formé d’une succession d’exons et d’introns. Cliquer sur le gène COX1 pour faire apparaitre la limite des exons et des introns dans le cadre « Six-Frame Translation ». Les exons de COX1 sont-ils tous codés dans le même cadre de lecture ? Comment l’élimination d’un intron peut-il permettre de changer de cadre de lecture ?

3- Le génome mitochondrial de C. glabrata contient de très nombreux gènes non codants. A quoi correspondent ces gènes non codants ?

* 1. **Utilisation du logiciel ORF FINDER pour détecter les CDS du chromosome mitochondrial de *Candida Glabrata*.**

Utiliser l’outil ORF FINDER en utilisant l’identifiant du chromosome mitochondrial de *Candida Glabrata* en entrée (NC\_004691.1).

**Questions/Discussions :**



1-Combien de séquences codantes sont détectées avec les paramètres par défaut ? Cette détection est-elle en cohérence avec l’annotation du génome mitochondrial ? Donner une explication au résultat obtenu ?

2-Quels paramètres modifier pour obtenir la détection la plus efficace des CDS mitochondriales de *Candida Glabrata* ?

3-Retourner sur la page d’annotation du génome mitochondrial de *C. Glabrata* ([Accession NC\_004691.1](https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=NC_004691.1&tracks=%5bkey:sequence_track,name:T727902,display_name:Sequence,id:T727902,dbname:GenBank,category:Sequence,subcategory:Assembly,annots:NA,ShowLabel:false,ColorGaps:false,shown:true,order:1%5d%5bkey:six_frames_translation,name:T727920,display_name:Six-frame%20translations,id:T727920,dbname:GenBank,annots:Six-frame%20translation,ShowOption:All,OrfThreshold:100,HighlightCodons:true,AltStart:false,shown:true,order:37%5d%5bkey:gene_model_track,name:T765810,display_name:Genes\,%20RefSeq%20propagation%20from%20INSDC%20submitter\,%20refreshed%20on%202017-09-25,id:T765810,dbname:SADB,category:Genes,subcategory:NCBI%20Genes,annots:NA000139222.1,Options:MergeAll,CDSProductFeats:false,NtRuler:true,AaRuler:true,HighlightMode:2,ShowLabel:true,shown:true,order:40%5d&assm_context=GCF_000002545.3&v=1:20063&c=CCFFCC&select=null&slim=0)). Comparer la détection informatique qui apparait dans le cadre six-frame translations (CDS de plus de 300 nucléotides grisées) à l’annotation basée sur des données expérimentales.

Quelles sont les limites de la méthode ORF FINDER mises en évidence par cette comparaison ? Donner des exemples de séquences codantes non détectées et expliquer pourquoi elles n’ont pas été correctement détectées.

4-Le génome mitochondrial de *C. glabrata* contient de nombreux gènes d’ARNt. Pourquoi ces gènes ne sont-ils pas détectés par ORF Finder ?

5- Le génome mitochondrial est dérivé du génome d’une bactérie ancestrale capable d’effectuer la respiration et qui est entrée en symbiose avec l’ancêtre de la cellule eucaryote. Au cours de l’évolution, les gènes mitochondriaux redondants avec les gènes de la cellule hôte ont été perdus, ce qui fait que la mitochondrie a perdu son autonomie. Votre analyse de ce génome vous permet-elle de comprendre pourquoi le génome mitochondrial a néanmoins conservé ses gènes d’ARNt ?