**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 TP, nous vous proposons d’explorer ce génome avec des méthodes classiques de recherche de gènes codants. Lors de votre projet, vous serez amenés à développer des outils pour découvrir des motifs permettant le contrôle de l’expression des séquences codantes de *Candida glabrata*.

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

**Base de données génomiques NCBI : <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 :*

*13 correspondent aux chromosomes et 1 correspond à la mitochondrie*

*Nombre total de molécule d’ADN : 13 + 1 = 14*

1. Analyse des données pour le chromosome A : Combien de gènes sont présents sur le chromosome A ?

Nombre de gène : 209

Donner le nombre de gènes codants et de gènes non codants présents sur ce chromosome.

Nombre de gène codant : 200

Nombre de gène non codant : 9

Calculer le nombre de gènes codants présents par kb (kiloBase) sur le chromosome A.

Chromosome A : 0.49 mb = 490 kb

le nombre de gènes codants présents par kb (kiloBase) sur le chromosome A:

Nombre de gène codant/taille de chromosome A = 200/490 = 0.4082

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 ?

Un fichier au format FASTA peut contenir plusieurs séquences. Chaque séquence est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">", et chaque ligne a un nombre de caractères limité.

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

Parce qu’un chromosome correspond à un molécule d’ADN. Même s’il a deux brins, on peut déduire les nucléotides d’un brin à partir de l’autre. Donc il suffit d’enregistrer une chaîne pour représenter un des deux brins.

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

TCAAAGGTAT

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’ TCAAAGGTAT 3’ brin 1

3’ AGTTTCCATA 5’ brin 2

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

5’ ATACCTTTGA 3’ brin2

6 - Utiliser l’outil en ligne de [conversion de séquence](http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html" \l "iupacdegeneracies) pour vérifier votre travail sur la séquence des 10 premières bases du chromosome A. Remplir le tableau ci-dessous indiquant à quoi correspondent les résultats (brin et orientation) des opérations « reverse », « complément » et « reverse complément ».

|  |  |  |
| --- | --- | --- |
| **Opération** | **Brin** | **Orientation** |
| Aucune : séquence initiale | Brin 1  TCAAAGGTAT | 5’>3’ |
| Reverse | Brin1  TATGGAAACT | 3’>5’ |
| Complement | Brin2  AGTTTCCATA | 3’>5’ |
| Reverse complement | Brin2  ATACCTTTGA | 5’>3’ |

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" \l "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. Une fois l’affichage obtenu vous pourrez configurer la longueur des ORFs à surligner en cliquant sur « six-Frame translation » au-dessus du diagramme affiché.

**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 ?

Parce qu’il y a 3 cadres de lecture pour chaque brin (car un codon est un triplet de nucléotides, on peut commencer par la première, deuxième ou troisième nucléotide), et on a 2 brins pour chaque chromosome, donc il y a 6 cadres.

Comment détecter des séquences codantes (CDS) à partir de cette représentation ?

Une séquence codante doit commencer par un codon Start (ATG) et se termine avec un codon Stop (TAA, TAG, UGA), représentés respectivement par des traits vert et rouge. Une fois la distance de deux traits, vert et rouge, est assez grande, il est possible qu’il y ait une CDS.

1. Observer le résultat de l’analyse ORF FINDER. Combien de CDS ont été détectées avec les paramètres par défaut ?

139 CDS

Que pensez-vous de cette analyse compte tenu du nombre moyen d’ORF par kb calculé précédemment ?

Le nombre de gènes codants présents par kb (kiloBase) sur le chromosome A : 0.4082, d’après la question 1)a)-2.

La région sélectionnée : 75 000 - 50 000 = 25 000 = 25 kb

Le nombre d’ORF présents par kb (kiloBase) sur la région sélectionnée :

Nombre d’ORF dans la région/ taille de la région = 139/25 = 5.56

Ce nombre est relativement élevé par rapport au nombre moyen d’apparition de CDS dans le chromosome A, cela veut dire que cette région a plus de CDS que les autres.

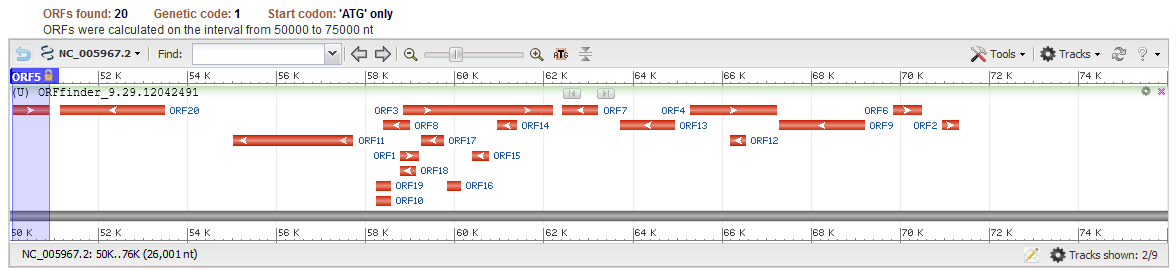
Quel est le nombre d’ORF qui devrait être attendu ?

(CDS sur le chromosome A) \* (taille de la région) = 0.4082\*25 = 10.25 ORF

1. Quels paramètres pourraient être modifiés pour augmenter l’efficacité de détection des CDS ?

Il faut augmenter la taille minimale de l’ORF, car plus grand elle est, plus possible qu’elle représente un CDS. Ici on choisit 300 comme taille minimale de l’ORF.

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.



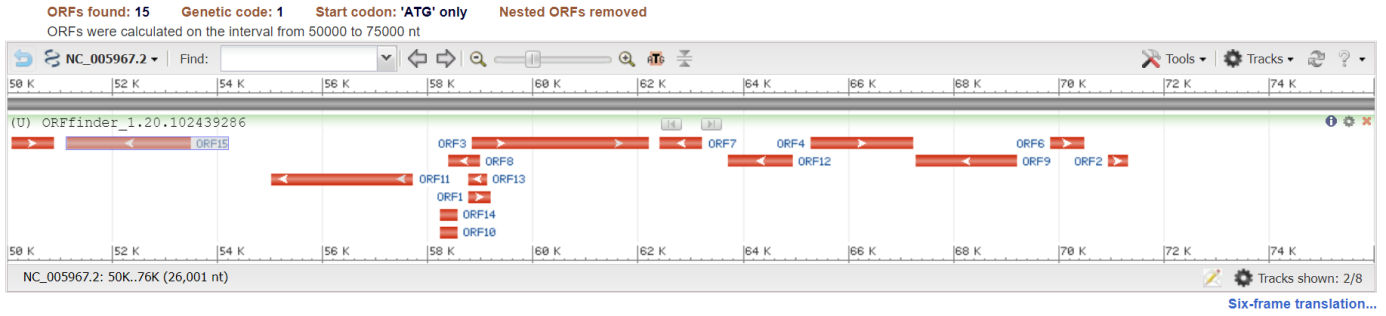
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Numéro** | **Brin** | **Cadre de lecture** | **Début** | **Fin** | **Taille (nt |aa)** |
| ORF5 | + | 1 | 50068 | 50880 | 813 | 270 |
| ORF20 | - | 3 | 53497 | 51137 | 2361 | 786 |
| ORF11 | - | 2 | 57722 | 55014 | 2709 | 902 |
| ORF19 | - | 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 |

1. 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 et expliquez sur quelle base sont éliminé certains ORFs  (Donner un exemple précis en utilisant les coordonnées de la table ci- dessus)?

Les ORF marquées en rouge sont impactées : ORF 12, 14 - 17, soit 5 ORFs.

Une ORF sont ignorée si elle :

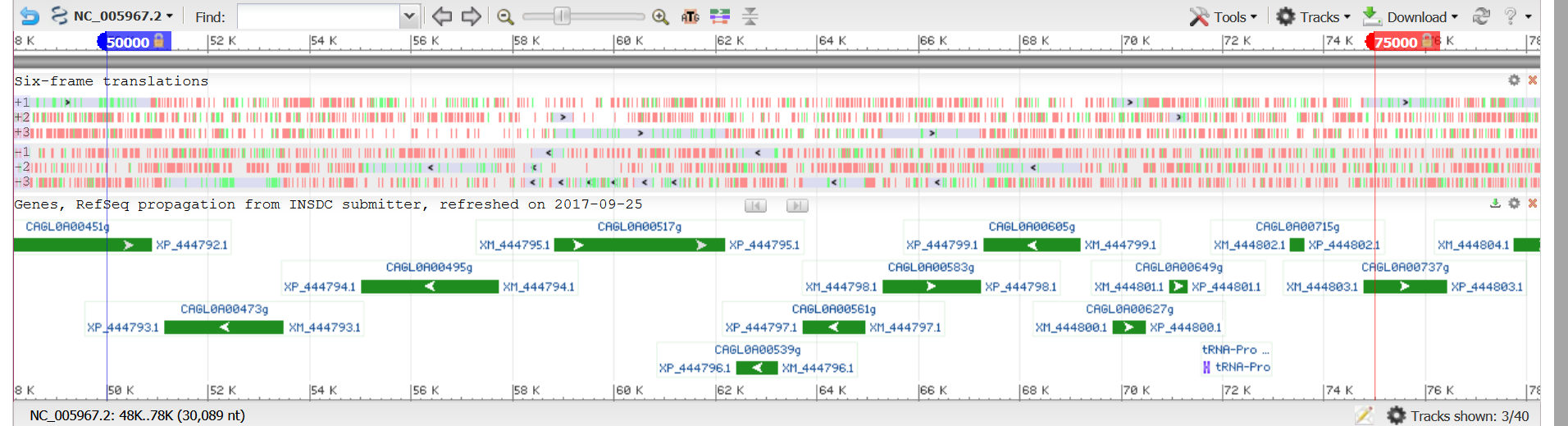
1. ont une ORF dans l’autre brin
2. Et cette ORF englobe une région plus grand (début plus petit qu’elle et fin plus grand qu’elle)
3. Cette ORF est dans le même lecture qu’elle



* 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](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) *[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)*



4

3

3

1

1

1

1

1

1

1

1

1

1

**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 (2a, question 3). Annoter la figure ci-dessus pour mettre en valeur :
2. les CDS correctement détectées : 10
3. les CDS détectées mais qui n’ont pas été confirmées par l’annotation : 10
4. Les CDS non détectées : 2
5. les gènes non codants annotés : 1

Avec 1, 3, 4 marqué dans la capture d’écran

1. Faire un bilan des différences observées entre la prédiction obtenue avec ORF FINDER et l’annotation du génome à partir de données expérimentales ?
2. Les CDS de taille assez grande sont probablement correctement détectées, alors que ceux de taille petit (prenons 0.3kb par exemple) détectés ont une possibilité considérable de ne pas être confirmés.
3. Il y a, même si le nombre n’est pas grand, des CDS non détectées et des gènes non détectés mais annotés.

Comment expliquer ces différences ?

1. Plus la taille d’une CDS détectée est grande, plus possible qu’elle représente un gène codant, et vice versa. Donc Même si des CDS de petite taille sont détectées, elles sont moins possible de représenter un gène codant, ce qui correspond au résultat qu’on a obtenu.
2. Les CDS non détectées ne commence pas par un codon Start, alors que les gènes non codants annotés correspondent aux autres choses que protéine(ici ARNt).
   1. **Analyse détaillée de CDS : de la séquence d’ADN à la protéine**

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

Nous vous proposons d’étudier deux des CDS détectées par l’outil ORF FINDER dont les coordonnées sont [55,014..57,722] et [58,830..62,198]. A modifier

Pour chacun de ces deux CDS, cliquer sur le bouton ATG pour zoomer au niveau de la séquence puis puis utiliser le cadre FIND pour vous placer soit au début soit à la fin de la séquence. Observer la séquence nucléotidique et celle de la protéine produite.



**Questions/Discussions :**



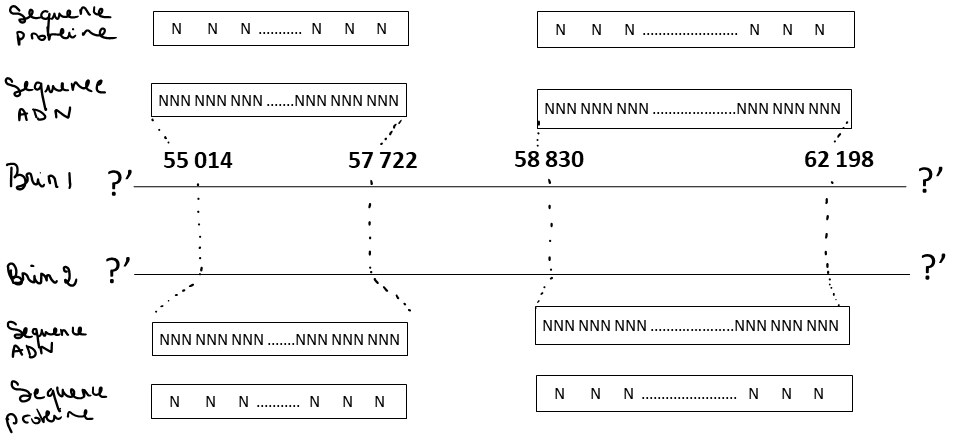
Compléter le schéma suivant à partir de votre analyse :



1- Indiquer l’orientation des deux brins (remplacer les ? par 5 ou 3)

2- Donner la séquence correspondant aux 3 premiers et derniers codons de chaque CDS pour les 2 brins : Remplacer les N par les lettres appropriées, cette séquence sera surlignée pour le brin codant

3- Donner les séquences en acides aminés correspondantes : Remplacer les N par les lettres appropriées uniquement dans le cadre correspondant au brin traduit en protéine (effacer l’autre cadre)



CCG GCT TGA

TAC GGT GTG

ATG CCA CAC

GTC GGA CAT

TTA GTT TTC

CAG CCT GTA

AAT CAA AAG

P A \*

CCG GCT TGA

M P H

\* N E

D S M

3

5

5

3



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* ?

Gènes codants : 11

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 ?

Non. 2 gènes sont dans le cadre de lecture 1, et 2 autres sont dans le cadre de lecture 2.

Comment l’élimination d’un intron peut-il permettre de changer de cadre de lecture ?

La longueur des introns peuvent NE PAS être un multiple de 3. Dans ce cas, les exons avant et après l’intron sont dans différents cadres de lecture. L’élimination de cet intron peuvent donc de changer le cadre de lecture.

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

Ils correspondent aux ARNt et ARNm.

4- 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.

A partir de cette analyse faire un bilan des limites de la méthode ORF FINDER mises en évidence par cette comparaison ?

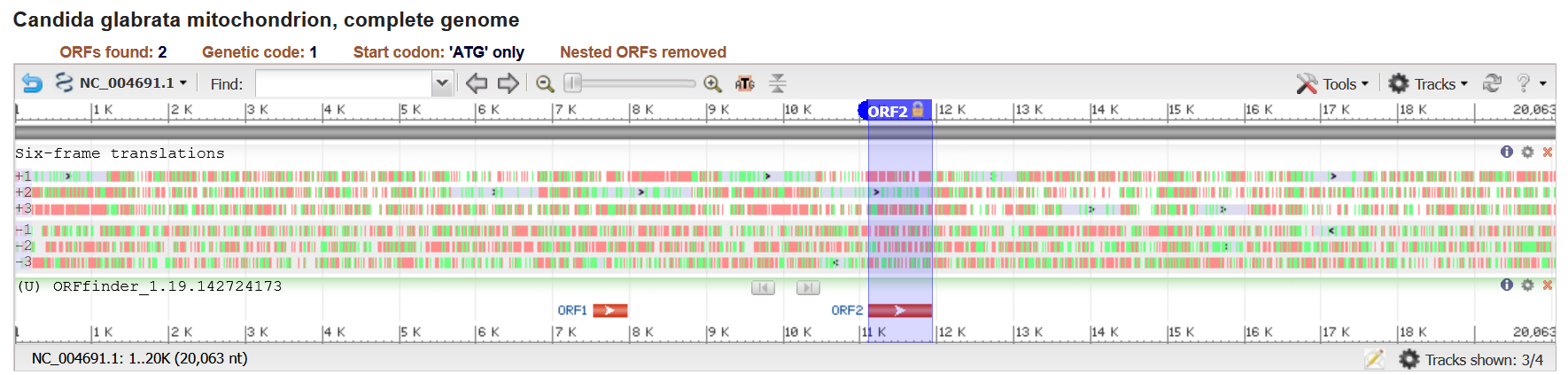
Les CDS détectées sont, dans la plupart, confirmées par l’annotation pour les gènes codants. En revanche, les gènes non codants ne sont pas détectés par ORF FINDER, car ils correspondent aux ARNt et ARNm, qui ne commence par un codon start.

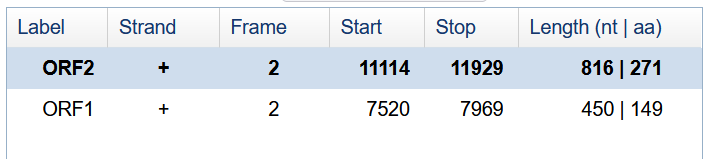
Donner des exemples de séquences codantes non détectées et expliquer pourquoi elles n’ont pas été correctement détectées.

Exemples : NP\_818782.1 de taille 147 nt et NP\_818784.1 de taille 231 nt. Elles ne sont pas détectées car leur taille sont trop petite (inférieur à la taille de filtrage).

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

Utilisez l’outil ORF FINDER avec l’identifiant du chromosome mitochondrial de *Candida glabrata* en entrée (NC\_004691.1) et une limite de taille d’ORF fixée à 300 nucléotides (les autres paramètres étant laissés aux valeurs par défaut). On obtient alors le résultat suivant :





Dans cette représentation les recherches d’ORF sont effectuée en utilisant le code génétique standard ce qui conduit à la détection de seulement 2 ORF (ORF1 et ORF2). La représentation 6-frame translations utilise quant à elle le code génétique associé à l’identifiant de la séquence (ici le code Yeast Mitochondrial Code : [code génétique](https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi" \l "SG1)

**Questions/Discussions :**



1-Commencer par obtenir avec l’outil ORF Finder la représentation présentée ci-dessus.

1. Pour comprendre les divergences observées entre la détection avec le code génétique standard et le code génétique utilisé au niveau des mitochondrie des levures, réalisez une étude approfondie des ORF1 et 2 détectées. Dans la boite Find, indiquez successivement la position des codons start et stop de chacune des ORF afin d’avoir un zoom au niveau ces sites.

ORFn° start stop

ORF1 7520 7969

ORF2 11114 11929

Vous pouvez aussi cliquer sur l’ORF d’intérêt et, dans le cadre en bas à gauche de la représentation ORF map, cliquer sur display as ORF.

Compléter le tableau ci- dessous avec vos observations :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numéro de l’ORF | Start | | | Stop | | |
| Codon | Traduction :  code standard | Traduction :  code mitochondrial | Codon | Traduction :  code standard | Traduction :  code mitochondrial |
| ORF1 | ATG | M | M | TGA | \* | W |
| ORF2 | ATG | M | M | TAA | \* | \* |

En vous aidant de cette analyse, indiquez quelles sont les différences entre les 2 codes génétiques qui perturbent la détection des CDS lorsque l’on utilise le code standard.

Pour codon stop, traduction code standard et celui de mitochondrial sont différentes, d’où la perturbation.

Pour chacun des deux ORF (ORF 1 et ORF2) proposez des explication expliquant pourquoi la detection avec le standard n’a pas permis d’identifier correctement l’ORF :

Pour ORF1, elle doit se terminer avec un codon stop dans le code mitochondrial, en occurrence TGA, mais la traduction standard prend TGA comme une protéine W. Donc la taille du gène prédit est plus grande que celle du gène annoté.

Pour ORF2, sur la séquence prédit, ORF FINDER ne trouve pas de codon stop dans la traduction du code standard, alors que le codon stop dans la traduction du code mitochondrial est très présent dans la séquence détectée mais non annotée.

Généralisez en expliquant pourquoi seulement 2 ORF ont été détectés avec le code standard.

Parce que la traduction du code standard et du code mitochondrial sont différentes, surtout les codons start et les codon stop. Cela fait que beaucoup d’ORF trouvées n’existent pas dans la réalité.

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

Parce que ces gènes ne commencent pas par un codon start pour la traduction du code standard.

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.

1. En théorie, compte tenu des divergences entre code standard et code mitochondrial, combien et quels tRNA devraient au minimum être codés par le génome mitochondrial ?

6 tRNA au minimum.

Est-ce le cas ?

Non

Si la cellule était incapable d’importer les tRNA dans la mitochondrie combien de tRNA devraient être codé dans le génome mitochondrial ?

62 tRNA.

Combien de tRNA sont codés dans le génome mitochondrial de *C. glabrata* ?

23 tRNA.

Conclure sur l’origine des tRNA permettant la production des protéine mitochondriale dans cette espèce.

Certaines (39) tRNA sont importées, d’autres (23) sont codées dans le génome mitochondrial.