Travail personnel associé au cours de Génétique

de VANHELLEPUTTE, Alyssa

Ce travail personnel a pour but de vous familiariser aux bases de données bioinformatiques que les chercheurs consultent lorsqu'ils ont besoin d'accéder aux informations sur l'organisation et la séquence des gènes et des génomes.

Connectez-vous à l'adresse http://www.ncbi.nlm.nih.gov/ du NCBI (National Center for Biotechnology Information). Cette page donne accès à un ensemble de bases de données très utiles pour la communauté scientifique, les enseignants et les étudiants. Par exemple, en cliquant dans le menu à droite sur le lien « PubMed », vous pouvez réaliser une recherche bibliographique sur un sujet particulier dans l'ensemble des revues scientifiques spécialisées dans le domaine des sciences du vivant. Si vous cliquez plutôt sur le lien « Bookshelf », vous pouvez réaliser cette recherche dans des livres disponibles gratuitement.

Cliquez dans le menu à droite sur le lien « Genome », vous aboutissez à un page fournissant une information régulièrement mise à jour sur le **séquençage des génomes**. La fenêtre d'interrogation vous permet d'accéder aux données génomiques d'un groupe biologique particulier ou d'une espèce. En dessous, le tableau grisé donne un aperçu des données (data) de génomique qui sont actuellement disponibles. On voit que +/- 2,5 millions de génomes ont été séquencés et que parmi ceux-ci, +/- 2 millions sont annotés (annotated). Cela signifie qu'à partir de la séquence brute d'ADN, des analyses bioinformatiques ont été réalisées pour identifier et recenser les gènes de chacun de ces +/- 2 millions de génomes. Ceci ne signifie pas que le génome de +/- 2 millions d'espèces différentes a été séquencé. En effet, pour une espèce donnée, c'est le génome de différentes variétés de cette espèce qui a le plus souvent été séquencé. D'où l'utilité, pour chaque espèce, de disposer d'une séquence de référence, unique. On voit ici qu'un peu moins de 40.000 séquences de référence sont disponibles, ce qui donc correspond au nombre approximatif d'espèces dont le génome a été séquencé. Si vous cliquez maintenant sur le groupe « Eukaryota », vous obtenez les chiffres correspondants mais pour les espèces eucaryotes seulement.

Insérez dans la fenêtre d'interrogation l'espèce ***Serratia rubidaea*** et cliquez sur « Search » à droite. Un tableau apparaît, il comporte une 30aine de lignes. Dans la colonne « Assembly », vous apercevez des codes, le premier est ASM1602673v1. Chaque code correspond à une séquence génomique assemblée de *S. rubidaea*. Autrement dit, le génome de cette bactérie a été séquencé une 30aine de fois. Plus précisément, l'analyse a été menée sur une 30aine de souches ou variétés de *S. rubidaea*. Le code marqué d'un « v » en vert indique que cette séquence génomique particulière est la **séquence de référence** pour cette espèce. Dans la colonne « Modifier », on voit que la souche ou variété (strain) correspondante est « FDAARGOS_926 ». Cliquez maintenant sur le lien ASM1602673v1, vous accédez à une page complète d'information sur cette séquence génomique de l'espèce. Sous le titre « Assembly statistics », sont présentées les informations fournies par deux plateformes de bioinformatique, RefSeq et GenBank. Pour votre travail personnel, on se concentrera sur les données fournies par **RefSeq**. Vous pouvez voir que le génome de *S. rubidaea* est formé d'un seul chromosome, d'une taille estimée à +/- 5 millions de bases (Mb), dont 59,5% de G :C. Un peu

plus bas, sous « Annotation details », on constate que RefSeq a identifé dans ce génome 4708 gènes dont 4558 codent pour une protéine. D'autres informations sont fournies sur cette page, dont des références d'article (dans le menu grisé à droite), que vous pouvez consulter pour en savoir plus sur cette espèce.

Cliquez maintenant sur le lien « view RefSeq annotation » en bas de la colonne « RefSeq » dans la section « Annotation details». La page qui apparait présente un tableau de l'ensemble des 4708 gènes que les bioinformaticiens ont recensés à partir de la séquence génomique de cette espèce. Le n° d'accession unique de cette séquence, visible dans la 1ère colonne, est NZ_CP065640.1. Les chiffres qui suivent directement ce n° indiquent la position des nucléotides délimitant la région codante de chaque gène (**ORF**, **open reading frame**), depuis l'ATG jusqu'au codon stop inclus. Dans la colonne « Orientation », est indiqué « minus » ou « plus », c'est-à-dire si l'ORF se trouve sur un brin ou l'autre de l'ADN. A noter que si deux ORFs consécutives sont sur des brins différents, leur sens de transcription sera opposé. En effet, la transcription s'effectue dans le sens 5' vers 3' et les deux brins d'ADN sont antiparallèles (5'-3' et 3-'5'). Par exemple, vous pouvez constater qu'entre les nucléotides 12310 et 13497, se trouve une ORF codant une protéine de 395 acides aminés (voir colonne « Length (aa)) ». Vous pouvez accéder à la séquence d'acides aminés de cette protéine, par exemple en cliquant sur le lien actif qui débute par « WP » dans la colonne « Proteins ». Dans le titre du haut de la page, on lit « NupC/NupG family nucleoside CNT transporter ». C'est la fonction qui est proposée pour cette protéine en se basant sur une ressemblance avec des protéines de la famille NupC/NpuG bien caractérisées chez d'autres espèces, mais il ne s'agit là que d'une prédiction et des expériences seraient nécessaires pour le confirmer. La séquence est présentée tout en bas de la page en utilisant le code à une lettre pour chaque acide aminé. Vous pouvez aussi l'obtenir en cliquant sur le lien « FASTA » qui apparait en dessous du titre de la page. Dans les fichiers de format « FASTA », bien connu des bioinformaticiens, la première ligne contient toujours le code de la protéine (ou de la séquence d'ADN) précédé du symbole « > ». La séquence débute à la 2de ligne. Revenez maintenant à la page qui listait tous les gènes de la bactérie. A noter que les deux gènes qui suivent celui qu'on vient d'analyser code pour des ARNs de transfert. Revenons à notre gène « NZ_CP065640.1:12310-13497».Dans la colonne « GeneID », se trouve le numéro de référence de ce gène au sein de la base de données « GeneID », 61762119. Cliquez sur ce lien. La page qui s'ouvre montre que le code de ce gène est « 6G83_RS00065». Tous ces codes sont assez complexes, mais on devine qu'ils sont indispensables pour organiser les informations sur les millions de séquences disponibles dans les bases de données bioinformatiques. Sous la barre grisée « Genomic context », vous pouvez observer une portion de chromosome (entre les nucléotides n° 10164 et 13834) comprenant l'ORF de 6G83_RS00065 (en couleur foncée) et les ORFs avoisinantes. Les flèches indiquent le sens de transcription des gènes correspondants, et vous pouvez ainsi déduire que le brin codant (ou brin sens) de l'ORF se terminant par 055 est situé sur le même brin d'ADN que notre gène 065, tandis que l'ORF se terminant par 060 se trouve sur le brin complémentaire, comme les deux petits gènes d'ARNt à côté de 065. Sous la barre grisée « Genomic regions, transcripts, and products », la barre rouge horizontale représente la région codante du gène 6G83_RS00065, les petites flèches blanches dans la barre indiquent le sens de transcription du gène (la graduation montre bien que le gène se situe environ entre les nucléotides 12300 et 13500 du chromosome bactérien). Les autres éléments qui en dehors de l'ORF constituent le gène (comme le **promoteur**, les **séquences régulatrices**, voir le podcast) ne sont pas représentés ici. Si vous positionnez le curseur de votre souris sur cette barre

rouge, une petite fenêtre grise contenant plusieurs informations s'ouvre. Elle vous indique par exemple que l'ORF (et le codon stop qui suit) s'étend (« Length ») sur 1188 nucléotides et que la protéine codée par ce gène (« Protein length ») a une longueur de 395 acides aminés. Fermez cette petite fenêtre grise pour revenir à la page du gène. Au-dessus de la barre rouge, vous voyez un menu avec des outils. En cliquant sur les flèches grises foncées (gauche et droite), vous pouvez « vous promener » le long du chromosome et analyser les gènes à proximité. En cliquant sur le + ou le – bordant la barre d'ajustement de taille (juste à droite des flèches grises), vous pouvez agrandir ou réduire la zone du chromosome. En réduisant la zone, vous constaterez que la **densité génique** de cette espèce bactérienne est très élevée (cf. podcast pour cette notion de « densité génique »). En agrandissant plusieurs fois la zone choisie, vous finissez par afficher la séquence nucléotidique du chromosome. Vous pouvez aussi accéder directement à cette séquence en cliquant sur l'outil « ATG ». Notez bien les polarités 5'-3' de chaque brin d'ADN indiquées aux extrémités de la séquence visible. Accéder à la séquence nucléotidique de ce gène serait par exemple utile pour concevoir les deux oligonucléotides amorces nécessaires pour amplifier ce gène par une réaction de PCR. Sous la barre rouge, est indiquée la séquence du brin codant (ou brin sens) organisée en codons successifs, ainsi que les acides aminés correspondants, indiqués à l'aide du code à une lettre.

Plus bas sur la même page, vous pouvez accéder à d'autres informations sur ce gène ainsi qu'à des liens vers d'autres bases de données. L'un d'eux est particulièrement intéressant et indiqué juste en face de «UniProtKB/TrEMBL», cliquez dessus. Vous avez ainsi basculé dans la base de données «**UniProt**», une des plus utilisées par les biologistes. On voit ici que les bioinformaticiens de UniProt considèrent également que cette protéine est du type « Nucleoside-transport system protein nupC ». La page renvoie à de multiples informations, par exemple dans la section « Structure » tout en bas on vous présente la structure tertiaire de la protéine prédite par **Alphafold**, un logiciel d'intelligence artificielle développé par Google/DeepMind. En plaçant et bougeant votre curseur (bouton de souris enfoncé) sur la protéine, vous pouvez la faire pivoter dans toutes les directions ☺. Dans la section «Family & Domains» juste en-dessous, se trouvent des liens multiples vers d'autres bases de données. Un est particulièrement utile, c'est « **Pfam** », à droite, qui est une base de données de tous les **domaines protéiques**. Cliquez sur « View protein in Pfam ». On peut voir ici que notre protéine contient trois domaines (en couleur), et si mettez votre curseur dessus, vous voyez qu'ils se situent entre les acides 10-82, 91-191 et 194-391.

Q1. Maintenant que vous savez comment le site du NCBI a organisé les informations sur les génomes des bactéries, connectez-vous à la page d'information sur le génome de l'espèce *Burkholderia pyrrocinia*. Combien de séquences génomiques sont disponibles pour cette bactérie?

Q2. Quel est le n° d'accession de la séquence génomique de référence de cette espèce ?

Q3. Quel est le nombre de gènes codant pour des protéines qui a été recensé par RefSeq dans la séquence génomique de référence de l'espèce ?

Q4. Consultez la page d'information de l'ORF qui débute au nucléotide n°40087. Pour quel type de protéine cette ORF est-elle prédite coder ?

Q5. Quelle est la longueur en pb de l'ORF de ce gène et quelle est la longueur en acides aminés de la protéine correspondante ?

Q6. Donnez la séquence des 3 codons qui suivent le codon « start » de l'ORF.

Q7. Regardez le gène (ORF) qui se trouve juste en aval de celui que vous venez d'examiner (donc du côté 3', au-delà du codon stop). Est-il transcrit dans le même sens? Justifiez votre réponse.

Q8. Donnez la séquence des 3 codons qui suivent le codon « start » de l'ORF analysée à la Q7.

Q9. Quel est, sur base des données fournies par Pfam, le nom complet du principal domaine ou, s'il y en a plusieurs, du plus long domaine de la protéine examinée aux questions 4 à 6, et entre quels acides aminés se situe-t-il?

Revenez maintenant à la page de départ qui fournit la fenêtre d'interrogation pour accéder aux données génomiques d'une espèce donnée. Insérez *Arabidopsis thaliana* et cliquez à droite sur Search. Vous constaterez que la séquence génomique qui sert de référence aux chercheurs a pour code « TAIR10.1 ». Mais vous pouvez aussi voir un peu plus bas que le génome d'*A. thaliana* a été séquencé (au moins partiellement) par plus de 250 projets additionnels. Vous pourriez vous demander pourquoi ce génome, comme celui des bactéries analysées avant, a été séquencé autant de fois ? En fait, grâce aux **méthodes Next Generation Sequencing** ou **NGS** (cf. cours), les scientifiques peuvent séquencer assez facilement le génome de variétés issues de zones géographiques différentes ou de différents mutants d'*A. thaliana*, et c'est ce type de données sur le **polymorphisme génétique** de l'espèce qui est accessible dans cette liste de projets. En cliquant sur le lien « TAIR10.1 », vous verrez que le génome d'*A. thaliana* a une longueur de 119,1 millions de paires de base (Mb) répartis sur 5 chromosomes. Plus bas dans la page est fournie une représentation graphique des différents chromosomes nucléaires, mais aussi du génome mitochondrial (MT) et du génome chloroplastique (Pltd). Pour chaque entité génétique, dans le tableau juste en dessous, la longueur en millions de paires de bases est fournie (Size, bp) est fournie, ainsi que le pourcentage en G:C. L'annotation bioinformatique de ce génome est une des plus abouties. Les différentes publications sur le séquençage du génome de *A. thaliana* sont également fournies.

Cliquez sur le chromosome 3, vous accédez à la représentation du chromosome fournie par l'outil « Genome Data Viewer » qui a été développé pour les espèces eucaryotes. Vous constaterez grâce à la barre graduée que le chromosome a une longueur d'un peu plus de 23 millions de pb. Placez votre curseur sur le chromosome (barre graduée), à hauteur de la position 17M (M = millions), et cliquez sur le bouton droit de la souris pour actionner plusieurs fois l'outil « Zoom in » de manière à identifier le gène qui se situe au niveau de la position nucléotidique 17.390.000 de ce chromosome. Une autre manière d'accéder à cette zone est de sélectionner, en enfonçant le bouton gauche de la souris, au niveau de la barre graduée, une fenêtre et de cliquer sur « Zoom on Range ». Vous verrez que le nom de ce gène est **PLC9**. Les gènes qui le précèdent et le suivent sur le chromosome, marqués en vert, sont aussi visibles. Si vous pointez votre curseur sur le gène, une fenêtre grisée apparait. Dans celle-ci, cliquez sur le lien en face de « Gene ID ». Vous accédez alors à une fiche d'information complète sur le gène PLC9, comme il en existe pour chacun des 28.500 gènes de *A. thaliana*. Vous retrouverez en particulier la représentation graphique du gène équivalente à celle montrée à la Figure 1 (à la fin de ce document). La zone incluant les différentes boites vertes successives correspond à l'**unité de transcription** (UT) du gène. Les boites vertes correspondent aux **exons** et les barres fléchées entre ces boites sont donc les **introns**. La zone en vert plus foncé correspond à la **région codante** du gène (répartie sur plusieurs exons), les zones en vert pâle aux

deux extrémités de l'UT correspondent aux **régions 5'UTR et 3'UTR** (untranslated region). La barre graduée au-dessus vous montre la numérotation des nucléotides le long du chromosome.

Positionnez le curseur sur la zone verte et cliquez encore. Vous observez maintenant une représentation en trois parties, de couleurs verte, mauve et rouge.

La partie en mauve correspond à l'**ARNm** du gène et la partie en rouge à la **région codante** du gène. Positionnez votre curseur au niveau de la zone verte, sans cliquer. Une fenêtre grise apparait. Vous y trouvez les **coordonnées chromosomiques** de l'unité de transcription, des nucléotides 17.387.883 à 17.390.960 du chromosome 3, l'unité s'étend ainsi (Length) sur 3078 nucléotides, qui est aussi la longueur de l'unité de transcription et donc du **transcrit primaire** (ou ARN prémessager). Il faut bien réaliser que les extrémités 5' et 3' de l'unité de transcription ne peuvent pas être prédites facilement par l'analyse bioinformatique de la séquence brute d'un gène. C'est en séquençant les ARNm du gène (ceux-ci sont en fait d'abord copiés sous forme d'ADN, appelé ADN complémentaire ou cDNA, puis clonés avant d'être séquencés) qu'on peut définir ces extrémités 5' et 3', ainsi que la position des introns et des exons (les introns étant manquants dans l'ARNm). Si vous placez le curseur sur la partie en mauve (ARNm), on voit dans la fenêtre grise qui s'ouvre que la longueur de l'ARNm (Sequence length) est de 2349 nucléotides. Enfin, si le curseur est placé sur la partie en rouge correspondant à la région codante, on peut lire au niveau de « CDS length » (CDS veut dire coding sequence) qu'elle a une longueur de 1596 nucléotides (codon stop compris) et qu'elle code une protéine de 531 acides aminés (Protein length).

Plus bas sur la même page, vous avez accès à de nombreuses autres informations sur ce gène, qui code pour une phospholipase agissant spécifiquement sur le phosphatidylinositol, une classe particulière de phospholipides. En particulier, une liste bibliographique est fournie, des liens vers « PubChem » qui est une base de données de voies métaboliques, et vous trouverez aussi le lien vers la base de données « UniProt ». Comme on l'a vu, celle-ci fournit des informations sur la protéine codée par ce gène, dont un modèle de sa structure tertiaire, sa localisation au sein de la cellule et un lien vers la base de données « Pfam » (vous pourrez y voir que la protéine dispose de 3 domaines particuliers). Dans les menus à droite, celui intitulé « Links to other resources » renvoie vers la base de données « TAIR » (The « Arabidopsis Information Resource »). Cette base de données est très utilisée par les chercheurs qui étudient *A. thaliana*. Parmi les informations utiles fournies par « TAIR » il y a les tissus de la plante où le gène est exprimé (« Atlas eFP browser »).

Q10. Vous êtes maintenant familiarisé à cet outil d'analyse de la structure des gènes eucaryotes. Retournez aux pages précédentes pour rechercher sur le chromosome n°3 de cette même plante, *Arabidopsis thaliana*, le gène qui se situe précisément au niveau de la position nucléotidique 18.996.000 de ce chromosome. Quel est le code de ce gène ?

Q11. Combien d'exons ce gène contient-il ?

Q12. Quelle est la longueur estimée de l'unité de transcription du gène ? (précisez les unités employées)

Q13. Quelle est la longueur estimée de l'ARNm de ce gène ? (précisez les unités employées)

Q14. Quelle est, au niveau de l'ARNm, la longueur de la région codante ? (précisez les unités employées)

Q15. Utilisez les différents outils qui sont à votre disposition dans la barre grise au-dessus, celle pour agrandir ou réduire la zone, les flèches pour se déplacer à gauche et à droite, et le bouton « ATG » ou « Zoom To Sequence » pour faire apparaitre la séquence des gènes. Positionnez-vous maintenant au niveau de l'exon comprenant le début de la région codante du gène que vous avez analysé lors des questions précédentes et zoomez plusieurs fois jusqu'à ce que la séquence de l'ADN apparaisse. Quelle est la séquence des 5 codons suivant le codon d'initiation de la traduction ?

Q16. Quel est la séquence des cinq premiers acides aminés de la protéine ? (utilisez le code à 3 lettres)

Q17. En amont du gène que vous avez examiné (c'est-à-dire dans la zone qui précède le promoteur), se trouve un autre gène de code AT3G51120 qui code pour une protéine. Examinez ce gène. A votre avis, le **brin codant** de ce gène se situe-t-il sur le même brin d'ADN du chromosome 3 que le brin codant du gène analysé plus haut ? Commentez votre réponse.

Q18. Examinez le début de la région codante du gène AT3G51120 et indiquez la séquence des 5 codons qui suivent le codon d'initiation de la traduction.

Q19. Comparer la longueur des **régions 5' et 3' non-traduites** (5' UTR et 3' UTR) de ce gène, laquelle est la plus longue ?

A ce stade de votre travail personnel, il est utile de préciser que les informations que vous venez d'extraire pour ces différents gènes ne sont pas nécessairement fiables à 100%. Par exemple, comme indiqué plus haut, les limites 5' et 3' d'une unité de transcription (ou de l'ARNm) ne peuvent être définies que par des moyens expérimentaux, et le degré de précision du résultat obtenu est limité. De même, les analyses bioinformatiques sont capables de prédire la position des introns et des exons d'un gène, mais la fiabilité de ces prédictions varie d'une espèce à l'autre et selon l'algorithme employé. Heureusement, ces prédictions sur la position des introns et des exons sont comparées aux résultats d'expériences (comme le séquençage des ARNm copiés en cDNA), mais ces derniers ne sont pas toujours disponibles ou complets. Il n'est donc pas toujours aisé pour les bioinformaticiens qui maintiennent à jour les bases de données de fournir à l'utilisateur une information unique et parfaitement précise pour chaque gène. D'ailleurs, pour de nombreux gènes, les limites 5' - 3' de l'U.T. ne sont tout simplement pas (encore) connues. Dans ce cas, on choisit provisoirement de faire coïncider ces limites 5' - 3' avec celles de la région codante du gène (bien qu'en réalité, on sache qu'elles se situent en amont et en aval de celle-ci), qui elles sont plus facilement prédictibles. On obtient alors une représentation graphique du gène du type de celle montrée à la Figure 2. C'est aussi ce principe qui a été adopté pour annoter les génomes bactériens.

En conclusion, le chercheur ou l'étudiant qui s'intéresse à un gène en particulier utilisera comme point de départ les bases de données bioinformatiques telles que celles que vous avez employées, mais il se donnera la peine de remontrer aux données sources afin de vérifier chaque information sur le gène qui l'intéresse. C'est pour cette raison que chaque page renvoie à tant d'autres pages et serveurs d'information.

Pour la dernière partie de ce travail personnel, vous allez examiner des gènes d'une espèce assez … singulière, *Homo sapiens* ☺ . Retournez d'abord sur la page de départ et introduisez ce nom d'espèce. La séquence génomique humaine de référence a pour code « GRCh38.14 », cliquez sur ce lien.

Q20. En bas de la page, vous voyez une représentation graphique des 24 chromosomes humains et du génome mitochondrial. Cliquez sur le chromosome n°4 (si cela ne fonctionne pas, allez plutôt dans le tableau juste en dessous et dans la colonne « Action », cliquez sur « Graphical view »). Dans la page qui s'ouvre, vous voyez en haut une image du chromosome et une graduation en unités M correspondant à des millions de pb. Comme expliqué plus haut et dans le podcast, vous pouvez grâce à la barre d'outils zoomer sur une région particulière du chromosome et faire apparaitre les gènes qui s'y trouvent. Avec les flèches gauche et droite de la barre d'outils, vous pouvez vous promener le long de vos chromosomes (n'est-ce pas fabuleux ? ☺). Au niveau de chaque gène, si vous avez suffisamment zoomé, vous verrez un code, propre à chaque gène. En positionnant le curseur de votre souris sur le gène, une fenêtre grise s'ouvre, et dans celle-ci, si vous cliquez sur le code en face de « GeneID », vous accédez à la fiche d'information complète de ce gène. Prenez le temps de consulter cette fiche, dont le format devrait maintenant vous être assez familier. Une chose intéressante qu'on y trouve, sous la rubrique « Expression », c'est un histogramme montrant le niveau relatif d'expression (quantité d'ARNm) du gène mesuré dans différents tissus.

Il faut savoir que la définition des gènes humains à partir des données brutes du séquençage a été établie indépendamment par plusieurs projets d'**annotation** du génome. C'est pourquoi plusieurs représentations de la succession des gènes le long des chromosomes sont proposées. La plus simple, celle que vous apercevez en haut, est celle fournie par le « MANE Project (release v1.2) » (Matched Annotation from the NCBI and EMBL-EBI), résultat d'une collaboration entre les bioinformaticiens américains du NCBI et européens de l'EMBL-EBI, les deux instituts de bioinformatique les plus connus. Juste en-dessous, apparait la position des gènes proposée par le NCBI (« NCBI RefSeq Annotation .. »). Elle est parfois plus complexe car pour la plupart des gènes, on distingue différentes unités de transcription, résultat de la présence probable de **promoteurs alternatifs** et du processus de **polyadénylation alternative**. De plus, certains exons ne sont parfois pas repris, ce qui suggère l'intervention d'un **épissage alternatif**. Plus bas encore, vous trouverez l'annotation fournie par la base de données « Ensembl » de l'EMBL-EBI. En parcourant le chromosome, vous pourrez constater que les annotations proposées par les différents instituts de bioinformatique divergent parfois, ce qui illustre bien que ces analyses bioinformatiques ne sont pas toujours aisées.

Vous allez maintenant examiner un gène en particulier de ce chromosome, le gène VEGFC. Revenez à la page correspondant à la séquence génomique humaine de référence « GRCh38.14 » et cliquez sur l'icône ronde « View annotated genes ». Dans la fenêtre « Filters », insérez VEGFC. Dans le tableau en bas, cliquez sur le code « Gene ID » correspondant qui apparait. La fenêtre qui s'ouvre ensuite vous présente une page d'information complète sur ce gène VEGFC. Une page de ce type est donc disponible pour chacun des ~20.000 gènes humains codant pour une protéine. Sur cette page, vous retrouvez en particulier la représentation graphique du gène (barre verte horizontale). Une différence saute aux yeux par rapport aux gènes que vous avez examinés chez *Arabidopsis thaliana* : les **introns** sont nettement plus grands que les **exons** et contribuent à la majeure partie de la très

grande longueur de **l'unité de transcription**. Quelle est cette longueur ? Précisez les unités employées.

Q21. Quelle est la taille de l'ARNm ?

Q22. Quelle est la longueur (en acides aminés) de la protéine ?

Q23. L'espèce humaine est celle dont le polymorphisme génétique a été le plus étudié à ce jour. Ainsi, pour chaque gène, on peut par exemple examiner la liste des « Single Nucleotide Polymorphisms » qui ont été identifiés. Ces **SNPs** (prononcez « snip » ☺) correspondent à des substitutions d'un seul nucléotide. Les SNPs les plus intéressants sont ceux qui provoquent un changement dans la séquence d'acides aminés de la protéine. Ces changements peuvent éventuellement altérer les propriétés fonctionnelles de la protéine et donc engendrer un phénotype. Dans certains cas, ce phénotype est une anomalie associée à une pathologie. L'étude des SNPs représente donc un énorme intérêt.

Voyons ce qu'il en est pour le gène VEGFC, qui code pour le facteur de croissance vasculaire C endothélial. Dans la page d'information de ce gène, vous trouverez dans la section « Variation » (ombrée en gris) un lien actif « See Variation Viewer (GRCh38) ». Cliquez sur ce lien, vous aboutissez à une page qui fournit toutes les variations génétiques recensées au sein du gène VEGFC par rapport à la séquence génomique GRCh38 de référence. Par exemple, pour accéder uniquement à la liste des SNPs, dans le menu à gauche « Filter by » cochez « dbSNP » et la liste des SNPs apparait. Plus bas dans le même menu, dans la section « Molecular consequences », vous pouvez sélectionner les variations génétiques qui provoquent une **mutation faux-sens** (missense variant) ou **non-sens**. Combien de mutations non-sens ont été identifiées dans ce gène VEGFC ?

Q24. Examinez maintenant l'organisation du gène **ST6GAL1** dans la version proposée par le « NCBI RefSeq Annotation ». On constate que cette organisation est plus complexe que celle du gène que vous avez examiné plus haut. En effet, quatre unités de transcription (UT) distinctes existent pour ce gène. La première UT code pour une protéine de 406 acides aminés et le codon start se trouve dans l'exon n°4. Dans la seconde, l'exon 4 est manquant, en raison d'un **épissage alternatif**, et la traduction débute alors dans l'exon suivant. La protéine (écourtée) a une longueur de 175 acides aminés. Les troisième et quatrième UT débutent en aval des deux autres, ce gène présente donc trois **promoteurs alternatifs**. Ces UT produisent des ARNm qui codent pour la même protéine de 406 acides aminés que l'UT n°1, mais la **région 5' UTR** de ces ARNm est différente. Examinez maintenant le gène humain **IMPA1** et son organisation proposée par le « NCBI RefSeq Annotation » : que pouvez-vous dire à propos de l'organisation de ce gène ?

Q25. Finalement, examinez le gène humain **OSGIN2** et son organisation proposée par le « NCBI Homo sapiens Annotation Release » : que pouvez-vous dire à propos de la terminaison de la transcription de ce gène ?

Portez une attention toute particulière aux dernières questions de ce formulaire, qui comptent le plus dans la pondération.

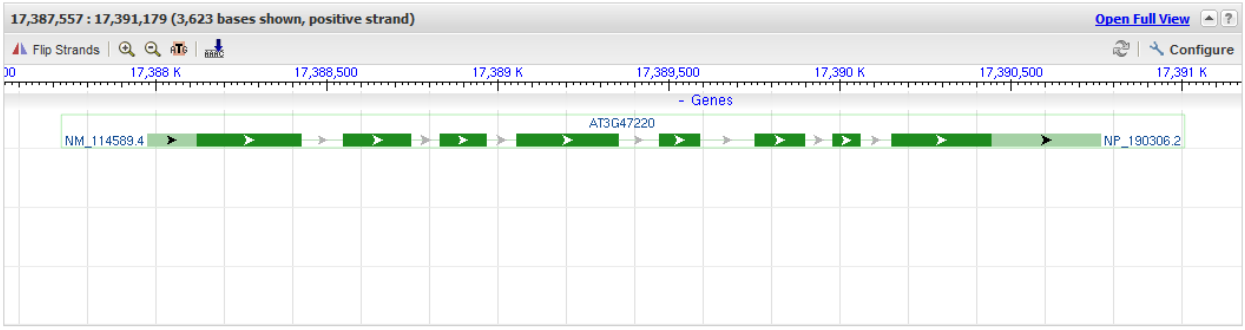

Figure 1. Représentation graphique d'un gène. En positionnant le curseur sur la souris sur la zone verte, on peut accéder à différentes informations concernant la structure de ce gène.

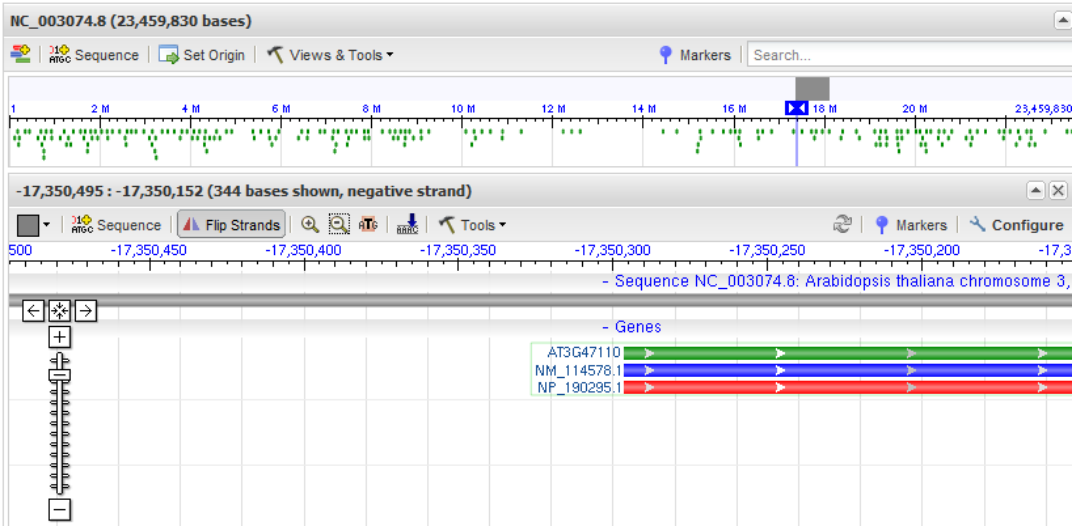

Figure 2. Dans la représentation graphique des gènes proposée par le site Entrez, quand les limites 5' et 3' de l'unité de transcription et de l'ARNm ne sont pas connues, elles coïncident avec celles de la région codante du gène.