

Module Analyse de Génomes 2010-2011

Master 2 Compétence Complémentaire en Bioinformatique.

Planning du Module :

Date	Heure	Salle
13/12	9h-13h	TD info 5.07
	13h-17h	TD info TA1Z bat 25
14/12	11h30-17h	TD info 5.04
15/12	9h00-11h15	TD info TA1Z bat 25
	11h30-15h00	TD info 5.10
	15h00-17h00	TD info 5.07
16/12	9h00-17h00	TD info TA1Z bat 25
17/12	9h00-17h00	TD info TA1Z bat 25

Prévoir:

Une présentation orale (présentation de votre travail de programmation 15-20 minutes) ainsi que des documents à rendre (rapport + scripts sur CD).

Important : Pour chaque site Web que vous visiterez, donnez :

- l'origine des données (source),
- l'organisme (privé ou public) distribuant les données,
- le domaine de spécificité de ce site,
- les liens vers l'extérieur qu'ils fournissent,
- toute autre information intéressante...

PARTIE 2 - Questions

a/ Quels sont les derniers génomes d'eucaryotes (vertébrés) séquencés (sources, liste, dates, taille du génome, nombre de chromosomes) ?

b/ La suite du module portera sur l'étude du génome du Pig (*Sus scrofa*) :

Espèce	Pig (<i>Sus scrofa</i>)
Chaîne	T cell Receptor Beta

Pour la suite vous complèterez le tableau en Annexe 2 en utilisant certaines instructions de l'Annexe 1

c/ Rechercher dans la base de données de références bibliographiques PubMed le ou les articles relatifs aux gènes TRBV, TRBD, TRBJ et TRBC de l'espèce.

d/ A partir de l'article de référence, récupérer les séquences nucléotiques (accessibles sur le site NCBI) et déterminer l'espèce la plus proche phylogénétiquement pour les loci TRB. Trouver les documents et informations allant dans ce sens.

e/ Dans Ensembl donner la localisation chromosomique approximative de la partie du locus TRB trouvé sur le chromosome, son orientation ainsi que les numéros des contigs correspondants.

f/ En vous appuyant sur les exemples de l'Annexe 3 et le protocole en Annexe 5 choisissez une séquence par une séquence par binôme pour répondre aux questions suivantes :

- Identifier le(s) V-GENE(s), D-GENE(s) et J(GENE)s présent(s). Délimiter (position de début et position de fin) chaque label constitutif (L-PART1, L-PART2, V-REGION, V-RS, J-REGION, D-REGION.....), les nommer et déterminer la fonctionnalité du (ou des) gène(s)(cf. IMGT® Scientific Chart – [Functionality](#)).
- Dessiner une carte de la portion de locus TRB trouvée en vous appuyant sur les représentations des locus présents dans IMGT.

Proposer la démarche et/ou les outils bioinformatiques nécessaires pouvant être réutilisés pour l'étude d'autres loci ?

g/ Pour chaque TRBV trouvé, donner le pourcentage d'identité avec les gènes de références de la souris et de l'homme à l'aide d'IMGT/V-QUEST (http://www.imgt.org/IMGT_vquest/share/textes/)

PARTIE 3 - Localisation des différents loci IG et TR de quelques espèces (Annexe 4) :

Indiquer, lorsque l'information est connue, pour chaque locus de chaque espèce, la localisation chromosomique et les positions correspondantes (avec le lien et/ou la référence qui vous ont permis de trouver l'information).

Remarque : Les informations pour l'homme sont vérifiées (source : IMGT Repertoire <http://www.imgt.org>). Il s'agit, donc pour les autres espèces, de vérifier, de préciser et de compléter le tableau ci-joint.

PARTIE 3 - Programmation.

But : Obtenir une carte complète du locus IGHV du rat (sur le chromosome 6)

Représentation graphique des gènes V à partir des données fournies (fichier sera distribué).
Automatisation de cette tâche (langage JAVA)

The column description is as follows:

Accession number
Gene name
Gene orientation/direction (1=+ve, -1=-ve)
Gene position (start) in bp
Gene position (stop) in bp
Functionnality

D'après la Charte IMGT®, il existe un code pour la fonctionnalité des gènes (rouge pour un pseudogène, jaune pour un ORF et vert pour un fonctionnel, cf. IMGT® Scientific Chart - [V-GENE prototypes with IMGT/LIGM-DB labels](#)).

Remarques :

- La possibilité de réutilisation de l'outil pour d'autres espèces devra être prise en compte.
- La création d'un diagramme de classe sera un plus.
- Chaque allèle ayant sa fonctionnalité, ce serait un plus de permettre au logiciel de gérer plusieurs fonctionnalités pour chaque gène.

Pour vous guider

- **IMGT:**

<http://www.imgt.org>

<http://www.imgt.org/textes/IMGTScientificChart/RepresentationRules/colorchart.html#LOCUS>

<http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/variable.html>

<http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/IMGTfunctionality.html>

<http://www.imgt.org/textes/IMGTScientificChart/Nomenclature/IMGTnomenclature.html>

IMGT/V-QUEST :

http://www.imgt.org/IMGT_vquest/share/textes/

- **Ensembl :**

<http://www.ensembl.org>

http://www.ensembl.org/Sus_scrofa/index.html

- **NCBI:**

<http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/>

<http://www.ncbi.nlm.nih.gov/Entrez/>

<http://www.ncbi.nlm.nih.gov/mapview/>

ANNEXE 1 :

Canis familiaris

Protocole pour la localisation des gènes IG Heavy, IG Light et TR

Date: 4 déc. 07

Auteurs: LUCAS Odile, BELLAHCENE Fatena

Localisation des loci sur le génome

- **Recherche du chromosome de chaque locus, via NCBI, soit par :**
 - Recherche bibliographique dans PubMed (<http://www.ncbi.nlm.nih.gov/sites/entrez/>)
 - Recherche de gènes dans **Entrez Gene** (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=>).

Saisie de mots clefs précis, exemple: *t cell receptor beta canis lupus*.

- Obtention d'une liste de numéro d'accès (LOC...) correspondant aux gènes recherchés. Dans le résultat de la requête (cf. exemple), nous obtenons la définition du gène, sa **localisation chromosomique**, sa position dans la séquence nucléotidique de référence, et son numéro d'identité (GeneID).

Exemple de résultats:

[LOC609058](#) similar to T-cell receptor beta chain V region 86T1 precursor [*Canis lupus familiaris*]

Chromosome: 16

Annotation: Chromosome 16, NC_006598.2 (9934782..9935332, complement)

GeneID: 609058

- Collecte de tous les LOC (séquences de gènes) pour les blaster, éventuellement, contre le génome du chien d'Ensembl, dans le cas où n'avons pas ou peu de séquences dans LIGM-DB.

- Recherche sur Ensembl, des Contigs contenant les gènes recherchés :

A partir des séquences de l'espèce *Canis familiaris* présentes dans LIGM-DB, réaliser un BLAST sur Ensembl (<http://www.ensembl.org/Multi/blastview>).

Collez la séquence en format FASTA.

Sélectionnez l'espèce : [Canis_familiaris](#).

The screenshot shows the Ensembl BLAST interface. At the top, there are tabs for 'new', 'SETUP', 'CONFIG', 'RESULTS', and 'DISPLAY'. On the right, there's a 'refresh' button and an 'Online Help' link. The main content area is titled 'Enter the Query Sequence' and has three options: 'Either Paste sequences (max 30 sequences) in FASTA or plain text', 'Or Upload a file containing one or more FASTA sequences', and 'Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)'. The first option is selected, and a FASTA sequence is pasted. Below this, there's a 'Parcourir...' button for uploading files. The 'Select the databases to search against' section has a dropdown menu for 'Select species:' with 'Canis_familiaris' selected. Below it, there are radio buttons for 'dna database' (selected) and 'peptide database'. The 'Select the Search Tool' section has a dropdown menu for 'Select the Search Tool:' with 'BLASTN' selected. At the bottom, there's a 'Search sensitivity:' dropdown menu with 'Near-exact matches' selected. On the right sidebar, there's a 'Summary' section with links for 'setup', 'configure', 'results', and 'display'. The 'setup' link is highlighted.

2) Visualiser le meilleur alignement, accessible *via* le lien : [\[A\]](#) (alignement)

Links	Query				Chromosome				Stats			
	Start	End	Ori		Name	Start	End	Ori	Score	E-val	%ID	Length
[A] [S] [G] [C]	246	497	+		Chr:8	5951940	5952191	+	252	6.5e-171	100.00	252
[A] [S] [G] [C]	1	174	+		Chr:8	5309561	5309734	+	170	6.3e-90	99.43	174
[A] [S] [G] [C]	1	171	+		Chr:8	5357376	5357546	+	159	2.3e-83	98.25	171
[A] [S] [G] [C]	1	171	+		Chr:8	5403505	5403675	+	155	5.8e-81	97.66	171
[A] [S] [G] [C]	1	171	+		Chr:8	5400390	5400560	+	139	2.1e-71	95.32	171

3) Détermination du contig grâce à *ContigView* (obtenu par le lien [C]).

Your Ensembl

- Login or Register
- About User Accounts

Chromosome 8

5,949,940 - 5,954,191

- View of Chromosome 8
- Graphical view
- Graphical overview
- View alignment with ...
- View alongside ...
- View Syntenic regions ...
- View region at UCSC
- View region at NCBI

Export data

- Export from region...
- Export Gene info in region
- Export SNP info in region

Chromosome 8

Chr. 8

5.90 Mb 5.92 Mb 5.94 Mb 5.96 Mb 5.98 Mb 6.00 Mb 6.02 Mb 6.04 Mb

contig_19031 > contig_19032 > contig_19034 > > > >

Overview

DNA(contigs)

Markers

Ensembl Genes

ncRNA Genes

Gene legend

- Ensembl Novel Pseudogene
- Ensembl Known Protein Coding
- RNAgene (Novel)
- Ensembl Known Proj Protein Coding
- Ensembl Novel Protein Coding

Obtenir le numéro d'accès (NCBI) du contig :

Remarque: Au préalable, il est nécessaire de centrer sur la région contenant le contig.

Cliquer sur le contig et choisir dans le menu s'affiche "[Centre on the contig](#)". Ainsi, nous obtiendrons **uniquement le numéro** du contig cible Map Viewer.

qui

d'accès
sur

La page de Map Viewer s'ouvre. A droite, dans le menu [Maps & Options](#), choisir d'afficher [Component](#).

S'affichent, alors, sur la carte les numéros d'accès :

- du contig du NCBI (colonne [Contig](#): séquence regroupant les components)
- du ou des gènes contenus dans la séquence (colonne [Genes_seq](#))
- de la séquence nucléotidique correspondant au contig d'Ensembl (colonne [Comp](#)).

Les étapes suivantes sont :

Récupération des séquences d'Ensembl: On récupère le format fasta des séquences d'ensembl pour les coller dans le champ texte d'un blast par exemple.

BLAST sur les bases d'IMGT :

Sur le site d'IMGT au Cines, nous pouvons effectuer un BLAST contre la base de données IMGT/LIGM-DB : <http://www.imgt.org/blast/blast.html>.

Ainsi il est possible de comparer nos séquences trouvées avec ce qui existe dans la base, mais des bases peuvent être faites (comme une base de données ne regroupant que les séquences de chien par exemple).

Localiser la région chromosomique du locus cible:

Récupérer les séquences nucléotidiques des contigs adjacents (en amont et en aval) du premier contig identifié comme contenant des gènes d'intérêts.

Pour chaque contig, réaliser le protocole précédemment décrit:

- Réaliser un BLAST sur les bases IMGT,
- Obtenir le numéro d'accès NCBI (si des gènes d'intérêt y sont localisés).

Synthèse des résultats :

Dans un tableau, noter :

- l'espèce, la race
- le chromosome,
- la position du locus sur le chromosome
- la taille du locus en Kb
- le nombre de paires de bases analysées (pour estimer si la recherche couvre une assez grande longueur chromosomique, en comparant avec les locus respectifs de l'homme)
- le nom des contig d'Ensembl, et le numéro d'accès du NCBI
- la position de chaque séquence sur le chromosome de 5' en 3'
- la taille des gaps entre chaque séquence
- les résultats du Blast pour déterminer la présence ou non de gènes d'intérêt.

ANNEXE 2 :
A remplir

Nom du gène	Fonctionalit e	EMBL Accession number	Ensembl Contig	Localisation chromosome	Position dans la séquence	% homologie avec l'espèce la plus proche	Sens

ANNEXE 3 - Exemples d'annotation

V-GENE

>X62106.0|HIVI2|Homo sapiens VI-2 gene for immunoglobulin heavy chain

tgagagctcc gttcctcacc atggactgga cctggaggat cctcttcttg gtggcagcag 60

ccacaggaa gaggtccct agtcccagtg atgagaaaga gattgagtcc agtccaggga 120

gatctcatcc acttctgtgt tctctccaca gaggccact ccaggtgca gctggtgcag 180

tctggggctg aggtgaagaa gcctggggcc tcagtgaagg tctctgcaa ggcttctgga 240

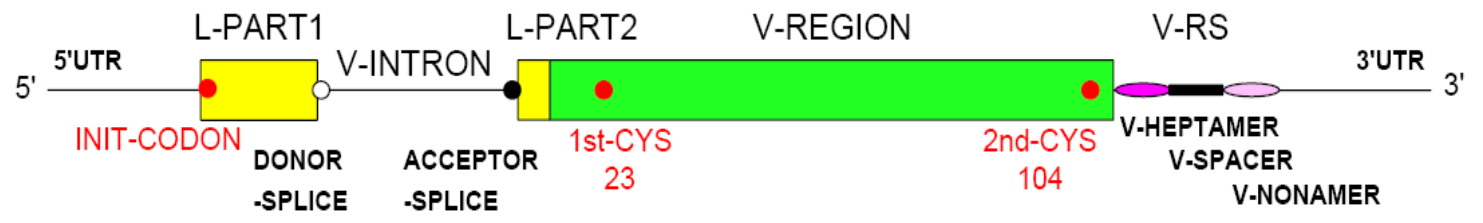
tacaccttca cgggtacta tatgactgg gtgcgacagg ccctggaca agggcttgag 300

tggatgggat ggatcaaccc taacagtggg ggcacaaact atgcacagaa gtttcagggc 360

agggtcacca tgaccaggga cagtcctatc agcacagcct acatggagct gagcaggctg 420

agatctgacg acacggccgt gtattactgt gcgagagaca cagtgtgaaa acccacatcc 480

tgaggggtg agaaacccaa gggaggaggc ag



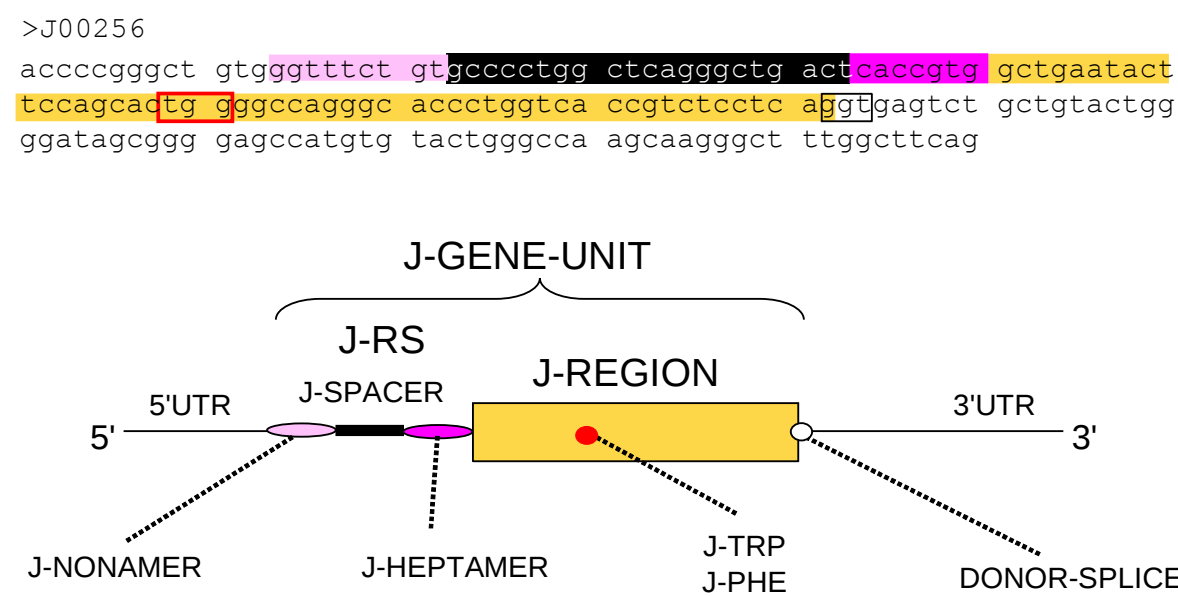
V-GENE (ProteinDisplay Mus musculus)

IGHV gene	FR1-IMGT (1-26)	CDR1-IMGT (27-38)	FR2-IMGT (39-55)	CDR2-IMGT (56-65)	FR3-IMGT (66-104)
AC073561,IGHV1-4	QVQLQQSGA.ELARPGASVKMSCKAS	GYTFTSYT....	MHWVKQRPGQGLEWIGY	INPSSGYT..	KYNQKFK.DKATLTADKSSSTAYMQLSSLTSEDS
AC090843,IGHV1-5	EVQLQSGT.VLARPGASVKMSCKTS	GYTFTSYW....	MHWVKQRPGQGLEWIGA	IYPGNSDT..	SYNQKFK.GKAKLTAVTSASTAYMELSSLTNEDS

D-GENE



J-GENE



ANNEXE 4 :

A remplir (quand ceci est possible)

	Human <i>Homo sapiens</i>	Mouse <i>Mus musculus</i>	Rat <i>Rattus</i> <i>norvegicus</i>	Cat <i>Felix catus</i>	Dog <i>Canis familiaris</i>	Zebrafish <i>Danio rerio</i>
IGH	14q32.33	12F1-F2				
IGK	2p11.2					
IGL	22q11.2					
TRA	14q11.2					
TRB	7q34					
TRG	7p14					
TRD	14q11.2		15p13			

[Le fichier texte](#)

ANNEXE 5 :

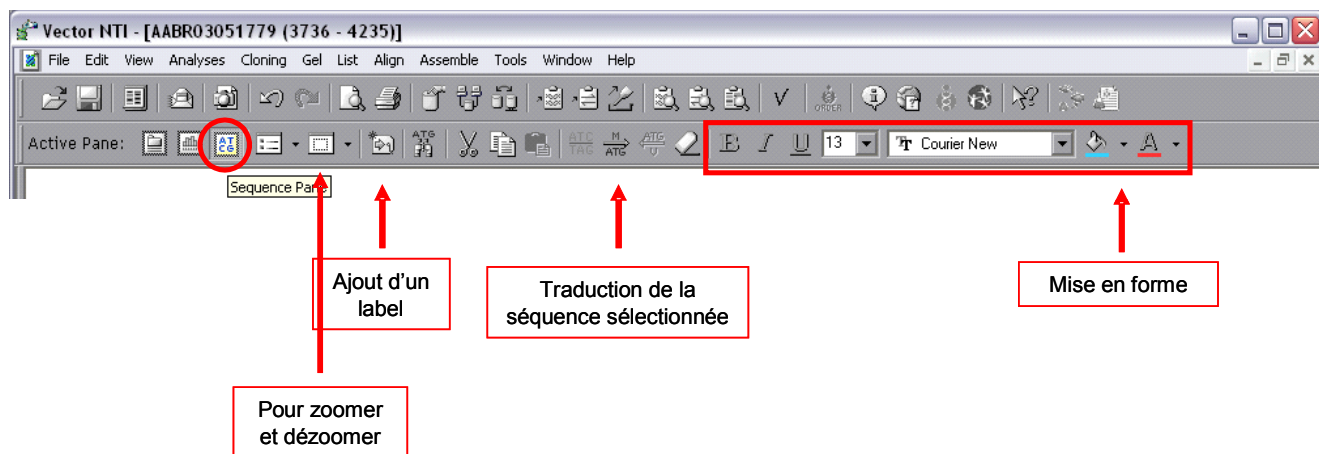
Utilisation du logiciel NTI Vector

Vous travaillez sur la **version DEMO** du logiciel, certaines fonctions sont limitées.

- Lancez le logiciel : Menu démarrer / Invitrogen / Vector NTI Advance 10 / **Vector NTI**
- Allez chercher la séquence à étudier dans GenBank : Tools / Open / Retrieve DNA-RNA by GenBank NID ...
- Entrez le numéro d'accèsion de votre séquence (AAEX...)
- Après chargement de la séquence, décochez la visualisation des sites de restriction (View / Display Setup / décochez "Restriction Map"), changez le code de traduction (View / Display Setup / Sequence Setup / choisissez "1-Letter AA Code"), et n'affichez que le brin sens (View / Display Setup / Sequence Setup / cochez "Prefer Single-Stranded Display").
- Afin de pouvoir enregistrer votre travail, vous devez, après importation de votre séquence, **faire une modification** (par exemple ajouter un label), puis **fermer le programme** NTI Vector. Il vous demandera s'il doit enregistrer les modifications. Acceptez l'enregistrement par défaut, la séquence est enregistrée à son nom (AAEX...) dans la base Database DNA/RNAs.
- Vous pouvez maintenant travailler sur votre séquence : ouvrez le fichier enregistré dans la base locale.

La mise en forme (couleurs, gras, italique, ...) et les traductions ne peuvent pas être enregistrées. Pensez à ajouter des "features" pour ne pas perdre votre travail lors de la fermeture du fichier.

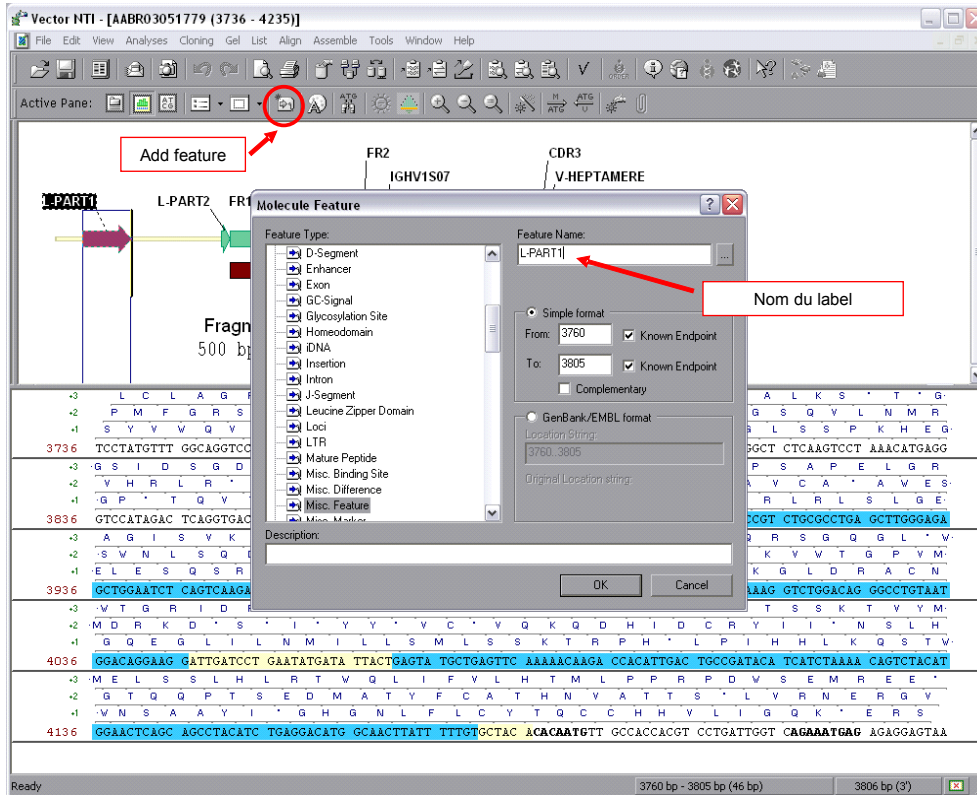
Principales fonctionnalités de la barre d'outil «Sequence pane» :



texte

Le fichier

Ajout d'un label :



Détermination des positions et de la taille des labels :

Vector NTI - [AABR03051779 (3736 - 4235)]

File Edit View Analyses Cloning Gel List Align Assemble Tools Window Help

Active Pane:

Fragment of AABR03051779
500 bp (molecule 11265 bp)

Diagram showing the structure of the fragment with regions labeled: L-PART1, L-PART2, FR1, CDR1, FR2, IGHV1S07, CDR2, FR3, CDR3, V-HEPTAMERE, and V-NONAMERE.

Sequence view showing the DNA sequence with a selection box highlighting a region from 3736 to 4235 bp. The sequence is displayed in a multi-line format with line numbers on the left and right.

Set Selection dialog box showing the selection range: From 3736 bp to To 4235 bp. The dialog box has OK and Cancel buttons.

Ready

3736 bp - 4235 bp (500 bp)

3736 bp (5')

Position et taille
de la sélection

Accession number	Gene Name	Orientation	Start position	End Position	Fonctionnalité
AABR03051417	IGHV5S6*01	1	138779786	138780078	P
AABR03050380	IGHV2S9*01	-1	138675881	138676176	P
AABR03054118	IGHV2S5*01	-1	138626807	138627097	F
AABR03050380	IGHV5S3*01	-1	138670380	138670670	P
AABR03051417	IGHV5S5*01	1	138776863	138777151	P
AABR03052910	IGHV2S12*01	-1	138677702	138677992	F
AABR03050380	IGHV2S11*01	-1	138661063	138661361	P
AABR03051316	IGHV5S2*01	-1	138620425	138620723	ORF
AABR03052450	IGHV2S14*01	1	138784488	138784780	F
AABR03053476	IGHV2S1*01	-1	138565901	138566193	F
AABR03050380	IGHV5S4*01	-1	138655570	138655861	P
AABR03051417	IGHV2S13*01	1	138771682	138771974	F
AABR03053476	IGHV2S2*01	-1	138564098	138564396	P
AABR03053042	IGHV2S7*01	-1	138642244	138642548	P
AABR03050380	IGHV2S10*01	-1	138662792	138663084	F
AABR03051316	IGHV5S1*01	-1	138621231	138621527	P
AABR03052372	IGHV2S8*01	-1	138645520	138645821	F
AABR03050045	IGHV5S7*01	1	138797897	138798192	ORF
AABR03052450	IGHV2S15*01	1	138786869	138787165	P
AABR03053042	IGHV2S6*01	-1	138644050	138644337	P