Annotation de séquences génomiques
Exemple d'une région du chromosome 1 de riz
autour du gène qSH1 (Os_1:36429001..36558000)

II) Annotation de gènes codant des protéines

1) Objectif du TD

L'objectif du TD est **d'identifier, sur une grande région génomique, l'ensemble des structures codant potentiellement pour des protéines**, au travers d'un ensemble de **méthodes d'annotation intrinsèques** (prédiction *ab initio* de structures codantes) et **extrinsèques** (faisant appel aux bases de données existantes).
La **comparaison des résultats** obtenus avec différentes méthodes bioinformatiques laisse apparaitre parfois des **divergences** sur le nombre de séquences codantes potentielles et/ou sur leurs bornes. L'utilisation de l'**éditeur Artémis** permet de mettre en évidence ces différences et de réaliser soi-même un travail de **correction manuelle de l'annotation**.
Au-delà d'**informations structurales** sur la région génomique considérée, il est possible d'acquérir des **informations fonctionnelles** au travers de méthodes extrinsèques par similarité des séquences et recherche de domaines protéiques conservés (signatures).
En fonction de la **significativité des résultats**, le résultat du produit des polypeptides va être attribué avec plus ou moins de confiance. L'éditeur Artémis permettra de **valider et d'enrichir** cette annotation fonctionnelle en fonction de l'**expertise** du bio-analyste.

Les modules bioinformatiques que nous allons utiliser pour l'annotation sont les suivants:

Méthodes intrinsèques

a. **Splicemachine** http://bioinformatics.psb.ugent.be/webtools/splicemachine/ prédit les sites d'épissage des introns par l'utilisation de la méthode dite « linear support vector machines » (LSVM) pour classifier les sites d'épissage actuels et pseudo-sites, à partir de données issues du génome d'*Arabidopsis thaliana* et du génome humain.
b. **EugeneIMM** utilise la méthode IMM **(**Interpolated Markov Modeler) pour interpréter les régions codantes et non codantes.
c. **FGenesh** http://www.softberry.com/berry.phtml est une méthode de prédiction de gènes *ab initio* basée sur des méthodes statistiques HMM (chaines de Markov cachées) avec une phase d'apprentissage supervisée.

Méthodes extrinsèques

a. **BLAST** (Basic Local Alignment Search Tool) http://www.ncbi.nlm.nih.gov/BLAST/ identifie des régions de similarité locale entre séquences. Le programme compare des séquences nucléotidiques ou protéiques et calcule la significativité des résultats.
   - **BLASTX** adresse une requête de type « nucléotide transcrit » sur des bases de données « protéines » type Swissprot ou Trembl.
   - **BLASTP** adresse une requête de type « protéine » sur des bases de données « protéines » type Swissprot ou Trembl.
   - **TBLASTN** adresse une requête de type « nucléotide transcrit » sur des bases de données « nucléotide transcrit », type NR (séquences non redondantes), EST (Expressed sequence Tag) ou des génomes complets.

b. **Genome Threader** http://www.genomethreader.org/ prédit des structures de gènes au travers de similarités avec des ADNc ou EST et/ou des séquences protéiques alignées (alignements consensus,

tenant compte des épissages). Il utilise un exciseur d'introns et un modèle « Baysian Splice Site Models » (BSSMs) pour identifier les limites exons-introns.

c. **Exonerate** http://www.ebi.ac.uk/~guy/exonerate/ est un outil d'alignement de séquences deux à deux. Il est capable de prendre en compte différents modèles d'alignements avec notamment la possibilité d'aligner un EST contre une séquence génomique ou bien une séquence protéique contre un génome.

EuGène (http://eugene.toulouse.inra.fr/) est un outil d'intégration des modules précédents dans le processus d'annotation. Il produit en sortie une prédiction de score maximal, c'est-à-dire la plus consistante possible avec les informations fournies par chacun des modules.

2) Executions de workflows sous Galaxy pour la prédiction automatique de gènes codant pour des protéines

*Récupération des données de séquence génomique

Sous Galaxy, dans le menu « Shared Data / Data Librairies », récupérer les fichiers du répertoire Formation / TD Annotation 2013 / Input :

- Os01_36429_36558.fna : Fichier fasta qui correspond à une séquence extraite du génome du riz que l'on va annoter.
- Os01_36429_36558.fna.raw.fg correspondant à la sortie du programme FGenesh.
- Os01_36429_36558.fna.repeat qui correspond à la sortie du programme RepeatMasker.

*Exécution de Workflows pour l'annotation sous Galaxy :

Importation du Workflow

- Dans le menu « Shared Data », cliquer sur le lien « Published Workflows »
- Cliquer sur le lien « EuGeneIMM3.2 Training 2013 »
- Importer le workflow dans son environnement
- Exécuter le workflow, puis l'éditer pour comprendre sa structure.

Ce workflow permet de prédire la structure et la fonction des séquences codant pour des protéines en se basant sur les modules précédemment cités.

Lancer le workflow à partir du fichier Os01_36429_36558.fna et du fichier Os01_36429_36558.fna.raw.fg (**ne cliquez qu'une fois**).

*Description du workflow :

Pour l'annotation structurale (Figure 1), 2 briques sont utilisées : « SpliceMachine » et « EuGene » (incluant EuGeneIMM). Le résultat d'une analyse réalisée sous FGenesh est également inclus dans Eugene, après conversion de format (« GNPAnnot Converters : FGenesH »).

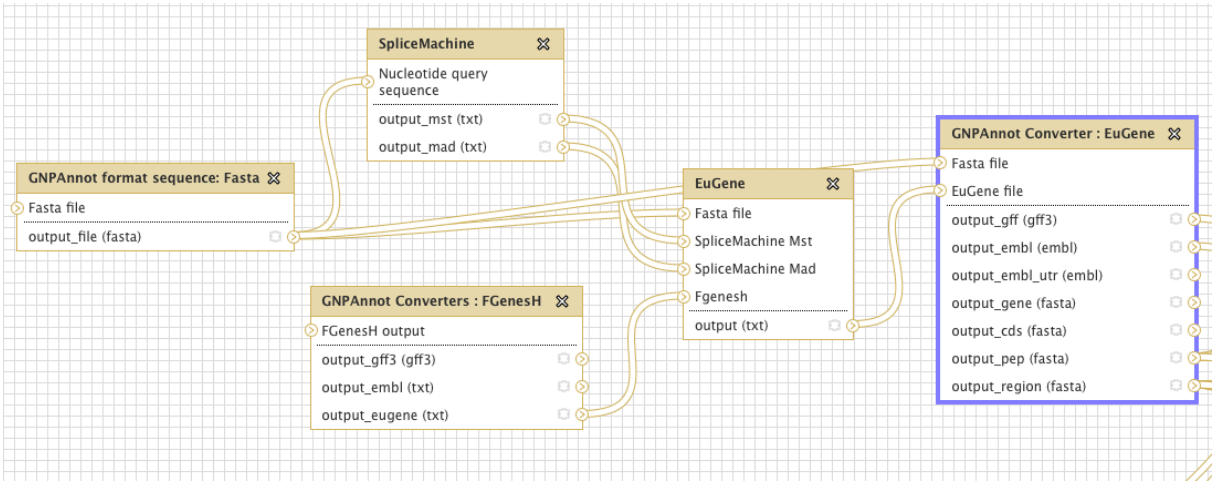

Figure 1 : Workflow Galaxy pour l'annotation structurale de séquence génomique

Le fichier résultant, EuGene result, correspond à la sortie brute de EuGene. Il sert de point de départ à l'annotation fonctionnelle. La brique « GNPAnnot Converter : Eugene » permet en effet d'extraire un fichier GFF3 contenant la structure des gènes prédits et les fichiers multi-fasta nécessaire à l'annotation fonctionnelle.
Cette brique produit en sortie les fichiers suivants :

- EuGene without functional annotation (gff3)
- EuGene without functional annotation (embl)
- Gene sequence with intron (fasta)
- Gene Coding Sequence intron less (fasta)
- Region around Gene (fasta)
- Translated Gene Coding sequence (fasta)

Annotation Fonctionnelle

Pour attribuer une fonction à un gène prédit par EuGene (Figure 2), la brique « GNPAnnot Converter : Blastp » combine les résultats de plusieurs sources de BLAST (SwissProt, MSU Rice genome annotation project =Rice MSUv6.1, Protéome Sorgho extrait de la base de donnée Phytozome) et transfère la fonction de la protéine la plus similaire ainsi identifiée.

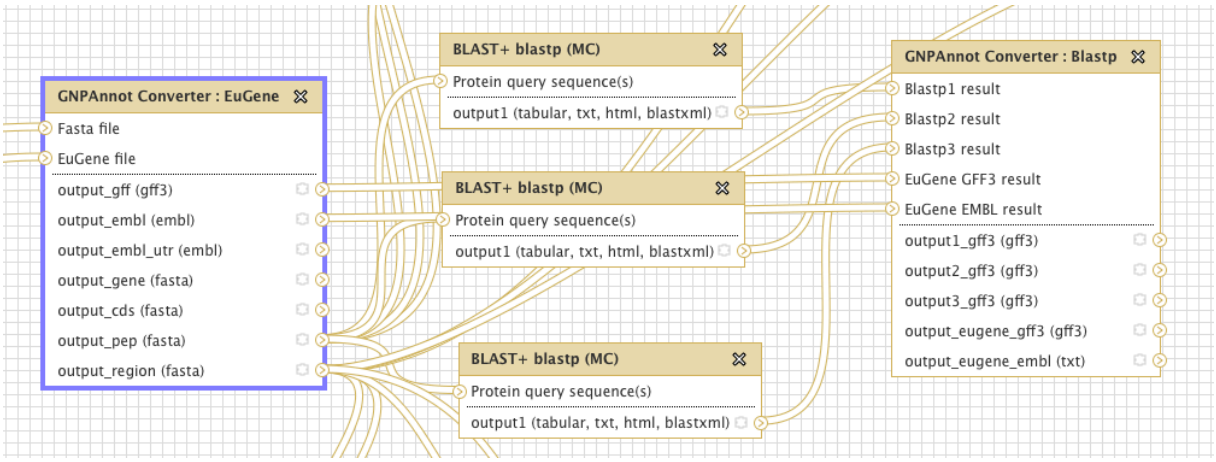

Figure 2 : Workflow Galaxy pour l'annotation fonctionnelle

Perfectionnement de l'annotation structurale :

Pour préciser la structure des gènes prédits (Figure 3), on utilise dans un premier temps une combinaison de TBLASTN et Exonerate sur les bases de données EST de riz (*Oryza sativa* et *Oryza glaberrima*) et de sorgho.
On utilise également en parallèle une combinaison de BLASTX/Exonerate et le programme Genome Threader, sur la séquence nucléique élargie entre gènes (Figure 4).

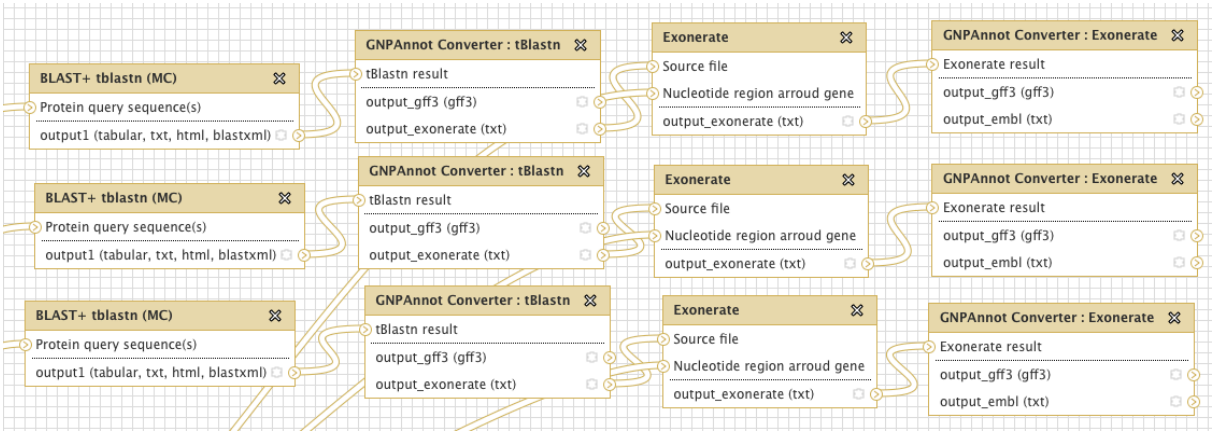

Figure 3 : Workflow Galaxy pour améliorer l'annotation structurale à partir des séquences protéiques des gènes prédits

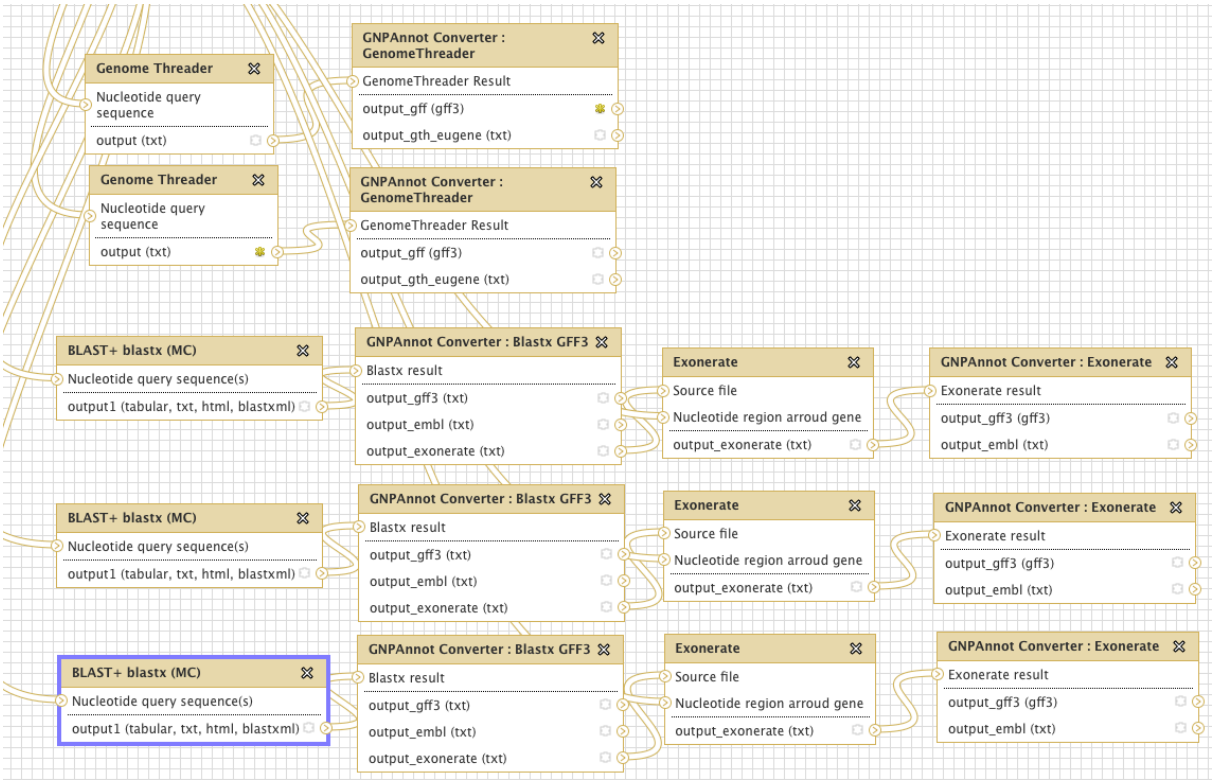

Figure 4 : Workflow Galaxy pour améliorer l'annotation structurale à partir des séquences nucléiques élargies des gènes.

*Récupération des fichiers de sortie du workflow:

Récupérer les fichiers de sortie suivants :

- FGenesH (embl) : Fichier au format EMBL du logiciel FGenesH
- EuGene (EMBL) : Fichier au format EMBL du programme EuGene
- Exonerate OG_ngs (EMBL) : Fichier EMBL correspondant à la combinaison des programmes tBlastn/Exonerate sur les contigs de Riz (ssp. glaberrima)
- Exonerate OS_mrnas (EMBL) : Fichier EMBL correspondant à la combinaison des programmes tBlastn/Exonerate sur la banque d'EST Riz (ssp japonica)
- Exonerate SB_mrnas (EMBL) : Fichier EMBL correspondant à la combinaison des programmes tBlastn/Exonerate sur le banque d'EST sorgho.
- Exonerate Rice (EMBL) : Fichier EMBL correspondant à la combinaison des programmes Blastx/Exonerate sur le protéome du Riz (MSU version 6.1)

- Exonerate SwissProt (EMBL): Fichier EMBL correspondant à la combinaison des programmes Blastx/ Exonerate sur la banque UniProtKB/SwissProt
- Exonerate Sorghum (EMBL): Fichier EMBL correspondant à la combinaison des programmes Blastx/ Exonerate sur le protéome du Sorgho

3) Visualisation des résultats sur navigateur de génome GBrowse et éditeur d'annotations Artemis

* Aller sur le site web de GNPAnnot.
http://www.gnpannot.org/
Puis dans ressources cliquer sur Sandbox 2.
http://www.gnpannot.org/content/gnpannot-sandbox-form-2-without-access-restriction
Générer la sandbox (**ne cliquez qu'une fois**).

* Launch your GBrowse!
Rechercher la région Os_1-36429001-36558000
Aller dans l'onglet « Select Tracks » pour choisir les pistes d'analyse à afficher
Revenez sur l'onglet « Browser » pour visualiser les prédictions affichées.
Cliquer sur un gène de la piste « Protein Coding Gene Model » et sur le lien « Edit with Artemis » du menu de la fenêtre Popup ou pour avoir toute la région la sélectionne au préalable et cliquer sur « Launch Artemis »
Ouvrir avec Java Web Start (défaut) -> OK
« Os_1-36429001-36558000_1..129000.jnlp est une application provenant d'un téléchargement depuis Internet. Voulez vous vraiment l'ouvrir » -> Ouvrir
Voulez vous exécuter l'application -> cocher « J'accepte le risque et je souhaite exécuter l'application » puis Exécuter
« Set Working Directory » -> OK
« Enter Database Address » -> remplissez l'identifiant « Password » puis cliquer sur OK
Artemis connecting -> Sequence loaded

* Le guide pratique d'Artemis connecté à Chado se trouve à l'adresse :
https://www.sanger.ac.uk/resources/software/artemis/#chado
ftp://ftp.sanger.ac.uk/pub/resources/software/artemis/database/chado.practical.guide.pdf

* A partir de la fenêtre d'édition de l'entrée Os_1-36429001-36558000 cliquez sur le menu File/Read An Entry
Fichiers du type : Tous les fichiers
Nom de fichier: Nom de fichier: Galaxy___-[FGenesH_(embl)].txt
A la question « there were warnings while reading - view now ? » répondez Non (ou oui si vous voulez voir les avertissements sur le format des annotations)
Ouvrir les fichiers
Nom de fichier: Galaxy___-[Exonerate_OG_ngs_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_OS_mrnas_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_SB_mrnas_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_Rice_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_Sorgho_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_SwissProt_(EMBL)].txt
Nom de fichier: Os01_36429_36558.fna.repeat

NB : Si vous avez besoin de retirer une entrée
Menu Entry/Remove An Entry/choisissez le fichier à retirer

* Pour faciliter la visualisation des résultats :
Clic droit sur la carte de la séquence
Cocher One Line Per Entry
Décocher Feature Labels

Q1 : Combien de structures codantes sont-elles prédites par Eugène ?

Cliquez sur l'objet CDS (exons en orange) du premier gène prédit par EuGène pour le sélectionnez
Menu Edit/Selected Features In Editor (Ctrl E)

Q2: Quel est le numéro du gène (identifiant ou locus_tag) ? Sur quel chromosome du Riz se trouve la région étudiée ?

4) Fgenesh

Nom de fichier: Galaxy___-[FGenesH_(embl)].txt

Q3: Quelles sont les différences de structure entre la prédiction EuGène et celle de Fgenesh ? A quoi cela peut-il être dû ?

5) TBLASTN / Exonerate contre les transcriptomes

Nom de fichier: Galaxy___-[Exonerate_OS_mrnas_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_OG_ngs_(EMBL)].txt
Nom de fichier: Galaxy___-[Exonerate_SB_mrnas_(EMBL)].txt

Q4: Peut-on émettre l'hypothèse que ce premier gène est exprimé Chez Glaberrima ? chez le sorgho ?
Q5: Quelles sont les différences de structure entre la prédiction EuGène et celles d'Exonerate ?

6) BLASTx / Exonerate contre protéome du sorgho

Nom de fichier: Galaxy___-[Exonerate_Sorghum_(EMBL)].txt

Q6: Comment exploiter ce résultat pour rechercher de la microsynténie entre cette région du riz et les chromosomes du Sorgho ?
Q7: Sur quel(s) chromosome(s) du sorgho se trouvent des régions synténiques potentielles ?
Q8: Quelles sont les différences de structure entre le premier gène prédit par EuGène et celle d'Exonerate ?

7) BLASTx / Exonerate contre UniprotKB/Swissprot

Nom de fichier: Galaxy___ Galaxy___-[Exonerate_SwissProt_(EMBL)].txt

Q9: Est-ce que les résultats attendus correspondent aux résultats observés ?
Q10: Quelles sont les différences de structure entre la prédiction EuGène et celle d'Exonerate ?

8) Annotation structurale dans Artemis

* Commencez par mettre de côté la séquence protéique du premier gène
Clic droit sur l'objet CDS (exons en orange)
Write/Amino acids of selected features
Select an output file name locus_tag_ori.faa (mettez le numéro du gène trouvé à la question 2).

*Editez le gène
Cliquer sur l'objet CDS (exons en orange)
Menu Edit/Selected Features In Editor (Ctrl E) pour ouvrir le Gene Builder d'Artemis.

* Corrigez la structure
Pour ajouter ou supprimer des exons, faites le absolument à partir de la fenêtre graphique après avoir cliqué à gauche sur les exons (auto, …)
Ajouter : sélectionner une région puis clic droit « Add to transcript in selected range » -> « exon » puis copier/coller les positions correctes dans location en respectant le format join(b1..e1,b2..e2,b3..e3,b4..e4,b5..e5)
Supprimer : sélectionner l'exon à supprimer puis clic droit « Delete Selected Exon ».
Cliquez sur OK

* Vérifiez la jonction GT / AG des exons créés
Double cliquez dans l'exon que vous venez de créer sur la carte de la séquence
Cela va positionner correctement la vue de l'ADN
Corrigez les bornes si nécessaire pour respecter la jonction GT / AG tout en respectant le cadre de lecture des exons (+1) : on ne doit pas voir de stop dans les exons (barre noire).
Pour cela, en positionnant le curseur sur l'extrémité d'un exon et en maintenant le bouton gauche appuyé vous pouvez étirer ou raccourcir l'exon.

* Cliquer sur Commit pour enregistrer les modification et attendez le message de l'inspecteur des annotations qui vérifie la cohérence de vos modifications.

Q11: Selon vous quelles sont les coordonnées correctes des exons du premier gène ?

9) BLASTp contre Uniprot / InterproScan

* Récupérez la séquence protéique du premier gène annoté manuellement
Clic droit sur l'objet CDS (exons en jaune)
View/Amino acids of selection as fasta
Copier la sequence sous le nom locus_tag_cor.faa.

* Lancez un navigateur, ouvrez deux onglets et aller à l'adresse suivante
http://www.expasy.ch/tools/blast/
ou http://www.uniprot.org/ onglet Blast
Copier-coller la séquence du fichier locus_tag_ori.faa et de locus_tag_cor.faa dans deux onglets séparés (à priori les multifasta ne sont pas acceptés)
Lancer le BLASTp en cliquant sur le bouton Run BLAST

De la même manière vous pouvez lancer un InterproScan pour la recherche de domaines protéiques
http://www.ebi.ac.uk/Tools/pfa/iprscan/

Q12 : Observez les alignements, votre annotation permet-elle d'améliorer l'alignement ? Quels indices vous permettent de conclure ?

10) Annotation fonctionnelle de LOC_Os01g62920 dans Artemis

* Editez et annotez ce gène (Deuxième sur le brin antisens, noté Os01b36429e36558_g0040 par Eugene)
Cliquez sur la CDS dont la structure a été annotée manuellement pour la sélectionner
Menu Edit/Selected Features In Editor (Ctrl E)

Analysez vos alignements blastp contre Uniprot

Q13: Quelle est l'accession Uniprot correspondant à votre gène ?
Q14: Quelle est l'accession Uniprot correspondant à une annotation de référence chez le riz ?
Q15: Grâce à cette annotation retrouvez la référence bibliographique permettant de valider la fonction expérimentale du polypeptide ?
Q16: Au vu de l'ensemble des ressources à votre disposition corrigez, complétez et finalisez l'annotation fonctionnelle du polypeptide en utilisant au maximum les *controlled vocabularies* termes de l'onglet CV du polypeptide dans le Gene Builder, le texte libre de l'onglet Core, vous pouvez remplir les champs correspondants.

Sauvez vos données.