TP :
Pipeline d’annotation des variants génétiques sous Galaxy

Ecole de Bioinformatique Roscoff 2013 – Atelier Annotation des SNPs
Sabrina Rodriguez et Maria Bernard

Objectif :

Cette formation a pour but de vous apprendre à annoter fonctionnellement des SNPs, annotations qui sont des critères de choix pour ensuite approfondir vos analyses. Vous avez vu dans les ateliers précédents « Détections des SNPs » comment identifier des SNPs. Nous verrons ici comment sélectionner des SNPs sur différents critères, comment prédire leurs conséquences, et enfin comment les annoter fonctionnellement.

Indications Pratiques :

Pour vous connecter à Galaxy :

- instance de Roscoff : http://galaxy.sb-roscoff.fr/
- instance de Toulouse : http://sigenae-workbench.toulouse.inra.fr/

Identifiant de formation pour l'instance de Toulouse :

- **login :**
  - anemone
  - aster
  - bleuet
  - iris
  - muguet
  - narcisse
  - pensee
  - rose
  - tulipe
  - violette
  - lilas
  - pervenche
  - laurier
  - lavande
  - lis
  - capucine
  - coquelicot
  - geranium
  - liseron
  - arome
  - chardon

- **Password**
  - f1o2r3!

Tous les outils concernant l'annotation sont classés dans la section « SNP annotation», dans le menu de gauche.
D'autres outils généraux vous seront très souvent utiles pour manipuler vos fichiers, n'hésitez pas à naviguer dans ces catégories.

ETAPE 1 : Initialisation de votre projet d'analyse

1) Les données

Pour la formation nous mettons à votre disposition un jeu de données test « Fichier_test_annot.vcf ». Le fichier est téléchargeable à cette adresse : http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/Annotation_SNP/Fichier_test_annot.vcf

Ces SNPs ont été détectés chez le bovin, c'est donc sur cet organisme que nous allons travailler.
Version du génome : Cow December 2009 (UMD 3.1 assembly)

2) Charger les données dans Galaxy : outil « Get Data »

Dans Galaxy pour accéder à vos fichiers vous avez 2 modes :

- *Upload from your computer :*

Cet outil vous permet de télécharger un fichier de votre ordinateur ou bien disponible via une URL sur internet (comme c'est le cas ici).
Une fois téléchargé vous pouvez renommer le nom du fichier grâce à l'outil « pencil » à côté de votre fichier. (exemple : « Fichier_test_annot.vcf »).

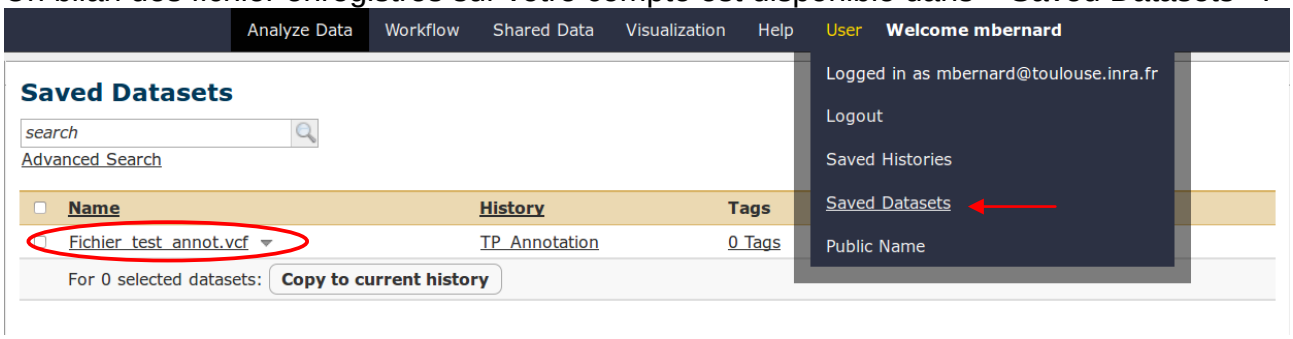
Attention : ce mode copie votre fichier sur le serveur Galaxy, il consomme donc une partie de votre quotat.

- ***Upload local file from filesystem path***

Cet outil permet de communiquer le chemin de votre fichier sur Genotoul/Roscoff au serveur Galaxy sans pour autant le copier. Ce mode vous permet d'économiser votre quotat sur le serveur Galaxy.
Pour Genotoul (comme c'est le cas ici), le fichier est déjà déposé sur vos comptes de formation il vous suffit d'indiquer les informations suivantes :

File Name : **Fichier_test_annot**
Filte Type: **VCF**
Path to file: **/work/USERNAME/galaxy/Fichier_test_annot.vcf**

Maintenant que notre fichier est chargé, notre projet d'analyse peut commencer, renommons donc notre historique dans le menu de droite,exemple « **TP_Annotation** ».
Un bilan des fichier enregistrés sur votre compte est disponible dans « Saved Datasets ».



3) Explorer et sélectionner des SNPs

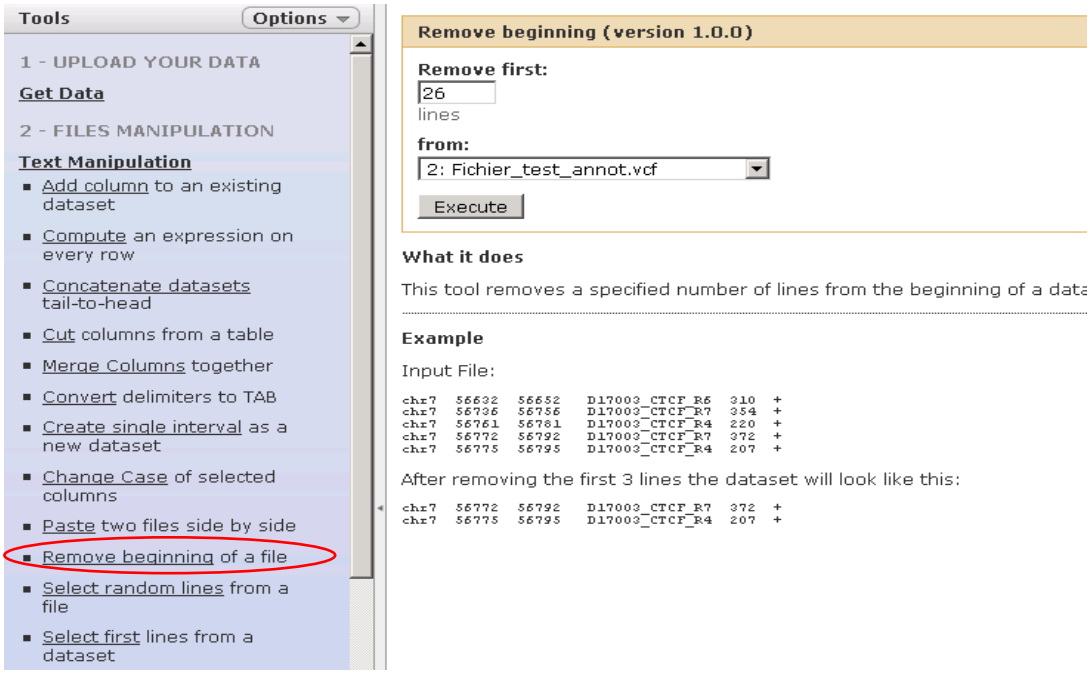
Vous pouvez visualiser un fichier en utilisant l'outil « eye » de ce fichier.

Exercice :

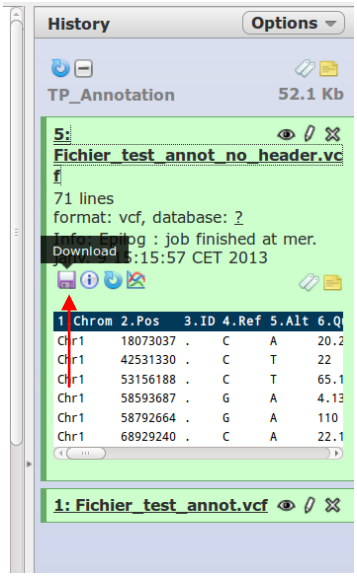
Supprimez les lignes d’en-tête qui vous gêneront lors de l'import du fichier dans Excel et téléchargez le fichier VCF sur votre ordinateur.

Créez un fichier qui ne contient pas le chromosome 8.

- Pour supprimer l'entête il suffit de supprimer les 26 premières lignes de notre fichier VCF grâce à l'outil « *Remove beginning* » de la catégorie « *Text Manipulation* »



Renommez le nouveau fichier créé puis utilisez l'outil « *download* » sur le nouveau fichier généré.



- Pour sélectionner uniquement certaines lignes, utilisez l'outil « *Filter* » de la catégorie « *Filter and Sort* ».

Cet outil permet de sélectionner des lignes en fonction des valeurs qu'elles contiennent dans 1 ou plusieurs colonnes. Votre fichier doit nécessairement être un fichier de type tabulé.

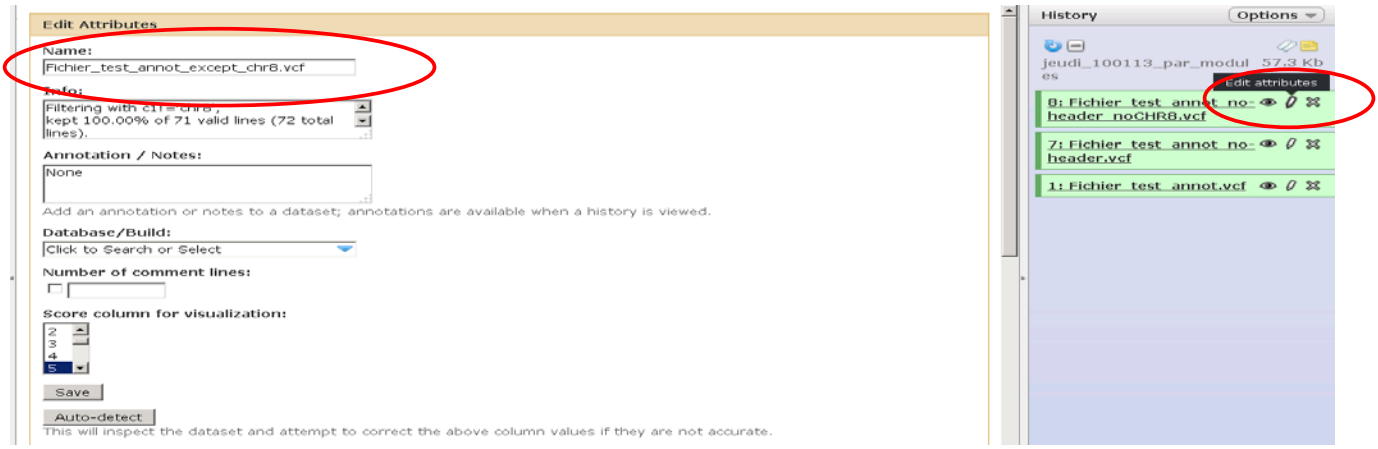
Dans le champ « With following condition: » écrire c1!= « Chr8».

Note :

- Si la valeur recherchée est une chaine de caractères, on la met entre simple côte
- S'il s'agit d'un nombre, il ne faut pas de côtes

N'oubliez pas de renommer votre nouveau fichier, **Fichier_test_annot_except_chr8.vcf**

Pour se faire, vous pouvez cliquez sur le « crayon » de la boite verte en haut à droite contenant le résultat de votre sélection de chromosomes, vous permettant d'éditer les attributs :



ETAPE 2 : Prédiction des conséquences

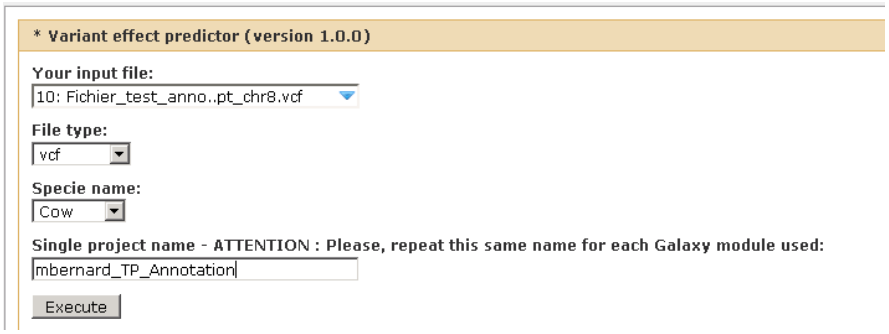
1) Génération des conséquences

Rappel

La première annotation que l'on peut générer est ce que l'on appelle la conséquence de chaque SNP. Cette annotation se fait grâce à l'outil : « *Variant_effect_Predictor* ».

Ce programme va parcourir un fichier VCF, et se connecter à Ensembl selon l'espèce du génome qu'on lui précisera. Il va ensuite produire des annotations de bases ou conséquences de chaque SNP en fonction de la localisation du SNP sur le génome dans Ensembl.

Lancement

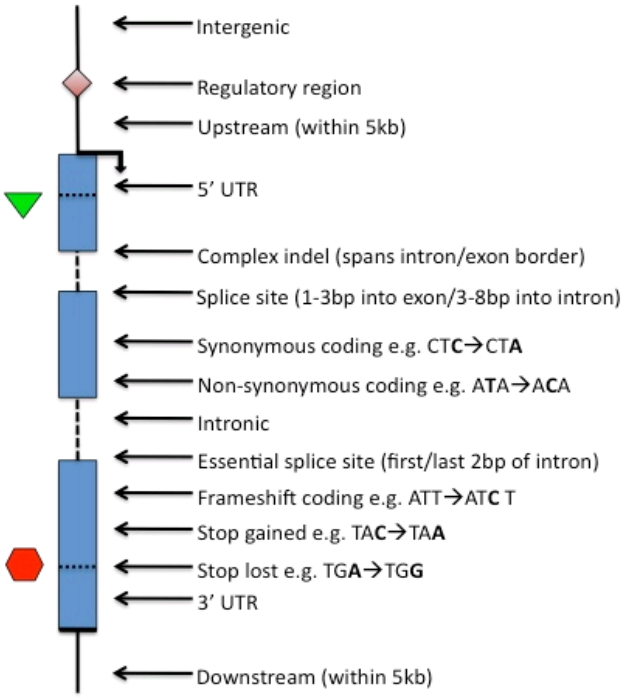


Attention de bien préciser un nom de projet. Celui-ci permettra à Galaxy de retrouver tous les fichiers de résultats lorsque vous voudrez les fusionner. Exemple :
« USERNAME_TP_Annotation ».
Vous devez garder ce nom de projet tout au long de l'analyse de votre échantillon (ou fichier .vcf), mais, ce nom ne pourra être utilisé que pour 1 analyse (pour une 2eme analyse, il faudra changer de nom de projet).

Ne pas oublier de préciser l'espèce : « cow ».

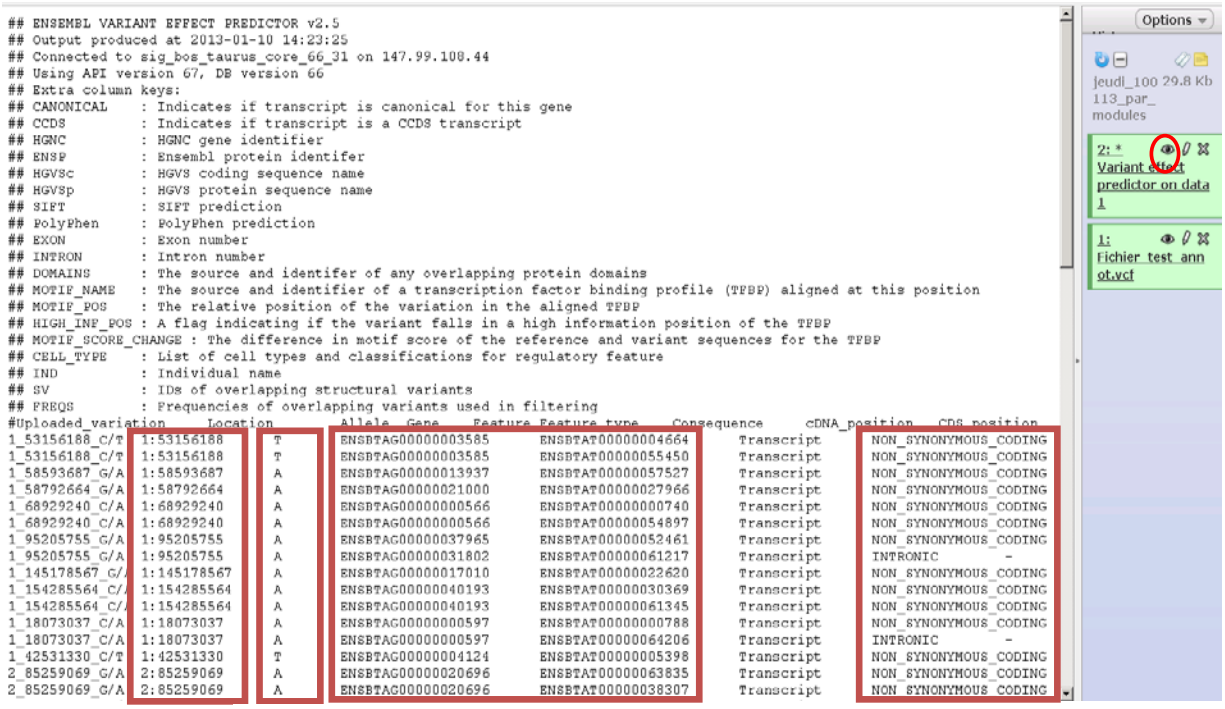
Résultat

Variant_effect_predictor retourne différentes annotations :



Others: Within non-coding gene, Within mature miRNA, NMD transcript

Il retourne également des informations concernant le gène impacté, et l'élément du gène impacté ainsi que des informations sur la position du SNP dans les différents éléments du gène et la conséquence au niveau de la séquence protéique.



```
## ENSEMBL VARIANT EFFECT PREDICTOR v2.5
## Output produced at 2013-01-10 14:23:25
## Connected to sig_bos_taurus_core_66_31 on 147.99.108.44
## Using API version 67, DB version 66
## Extra column keys:
## CANONICAL    : Indicates if transcript is canonical for this gene
## CCDS         : Indicates if transcript is a CCDS transcript
## HGNC         : HGNC gene identifier
## ENSP         : Ensembl protein identifer
## HGVSc        : HGVS coding sequence name
## HGVSp        : HGVS protein sequence name
## SIFT         : SIFT prediction
## PolyPhen     : PolyPhen prediction
## EXON         : Exon number
## INTRON       : Intron number
## DOMAINS      : The source and identifer of any overlapping protein domains
## MOTIF_NAME   : The source and identifier of a transcription factor binding profile (TFBP) aligned at this position
## MOTIF_POS    : The relative position of the variation in the aligned TFBP
## HIGH_INF_POS : A flag indicating if the variant falls in a high information position of the TFBP
## MOTIF_SCORE_CHANGE : The difference in motif score of the reference and variant sequences for the TFBP
## CELL_TYPE    : List of cell types and classifications for regulatory feature
## IND          : Individual name
## SV           : IDs of overlapping structural variants
## FREQS        : Frequencies of overlapping variants used in filtering
#Uploaded_variation     Location        Allele  Gene            Feature         Feature_type    Consequence     cDNA_position   CDS_position
1_53156188_C/T  1:53156188      T       ENSBTAG00000003585      ENSBTAT00000004664      Transcript      NON_SYNONYMOUS_CODING
1_53156188_C/T  1:53156188      T       ENSBTAG00000003585      ENSBTAT00000055450      Transcript      NON_SYNONYMOUS_CODING
1_58593687_G/A  1:58593687      A       ENSBTAG00000013937      ENSBTAT00000057527      Transcript      NON_SYNONYMOUS_CODING
1_58792664_G/A  1:58792664      A       ENSBTAG00000021000      ENSBTAT00000027966      Transcript      NON_SYNONYMOUS_CODING
1_68929240_C/A  1:68929240      A       ENSBTAG00000000566      ENSBTAT00000000740      Transcript      NON_SYNONYMOUS_CODING
1_68929240_C/A  1:68929240      A       ENSBTAG00000000566      ENSBTAT00000054897      Transcript      NON_SYNONYMOUS_CODING
1_95205755_G/A  1:95205755      A       ENSBTAG00000037965      ENSBTAT00000052461      Transcript      NON_SYNONYMOUS_CODING
1_95205755_G/A  1:95205755      A       ENSBTAG00000031802      ENSBTAT00000061217      Transcript      INTRONIC        -
1_145178567_G/  1:145178567     A       ENSBTAG00000017010      ENSBTAT00000022620      Transcript      NON_SYNONYMOUS_CODING
1_154285564_C/  1:154285564     A       ENSBTAG00000040193      ENSBTAT00000030369      Transcript      NON_SYNONYMOUS_CODING
1_154285564_C/  1:154285564     A       ENSBTAG00000040193      ENSBTAT00000061345      Transcript      NON_SYNONYMOUS_CODING
1_18073037_C/A  1:18073037      A       ENSBTAG00000000597      ENSBTAT00000000788      Transcript      NON_SYNONYMOUS_CODING
1_18073037_C/A  1:18073037      A       ENSBTAG00000000597      ENSBTAT00000064206      Transcript      INTRONIC        -
1_42531330_C/T  1:42531330      T       ENSBTAG00000004124      ENSBTAT00000005398      Transcript      NON_SYNONYMOUS_CODING
2_85259069_G/A  2:85259069      A       ENSBTAG00000020696      ENSBTAT00000063835      Transcript      NON_SYNONYMOUS_CODING
2_85259069_G/A  2:85259069      A       ENSBTAG00000020696      ENSBTAT00000038307      Transcript      NON_SYNONYMOUS_CODING
```


Localisation **Alternatif** **Géne et transcrit** **Conséquences des SNPs**

Renommez le fichier, par exemple : **Fichier_test_annot_except_chr8.vep**

Exercice :

Sélectionnez les SNPs qui sont catégorisés comme « NON_SYNONYMOUS_CODING » et « INTRONIC ».

Sachant que nous avons 1 SNP par ligne, comment savoir le nombre de SNPs sélectionnés ?

- Pour sélectionner les SNPs, vous pouvez utiliser « Select ».

L'outil « Select » est plus généraliste que l'outil « Filter » qui ne fonctionne que sur des fichiers tabulés, mais il ne vous permet pas de sélectionner la colonne sur laquelle vous voulez faire votre recherche.

Dans le champ « the pattern: » écrire : NON_SYNONYMOUS_CODING|INTRONIC

Attention ne pas mettre d'espace, si tel était le cas l'outil chercherait également ce caractère, or nous avons une tabulation entre chaque colonne !!!

Renommez votre fichier de sortie : exemple :
« **Fichier_test_annot_except_chr8_NCS_INTRONIC.vep** »

- Pour connaître le nombre de lignes d'un fichier, donc le nombre de SNPs,
    - Vous pouvez lancer l'outil « *Line/Word/Character count* » en cochant « line »,
    - Ou plus simplement cliquer sur le fichier d'intérêt et vous avez en dessous une brève description du fichier qui contient parfois le nombre de lignes.

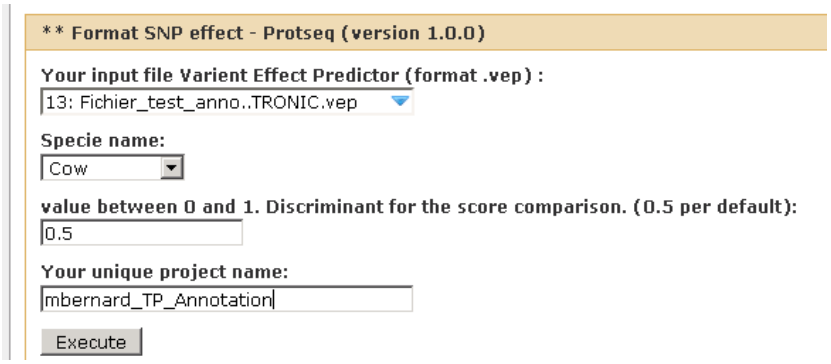
2) Formatage et génération des séquences protéiques pour la recherche des conséquences des SNPs

Rappel

Le formatage des données consistes à récupérer les séquences protéiques de chaque couple [allèle référence / allèle alternatif] de chaque SNP.
Dans le cas ou le SNP est multiallèlique (> 2 allèles), il peut y avoir plus de 2 couples d'allèles par SNP.
Si un SNP affecte plus de 2 transcrits, il y aura aussi plus de 2 couples, c'est à dire au moins 4 protéines générées (référence / alternatif 1 et référence / alternatif 2).
Les noms initiaux des séquences sont formés à partir de leur localisation sur le génome, du nom du transcrit auquel il appartient et des allèles référence et alternatif qu'il contient suivi de « alléle1 » pour la référence et de « allèle 2 » pour l'alternatif. Les séquences sont ensuite renommées avec un encodage de noms pour faciliter l'utilisation de certains outils d'annotation en utilisant la syntaxe « _seq1, seq_2... ».

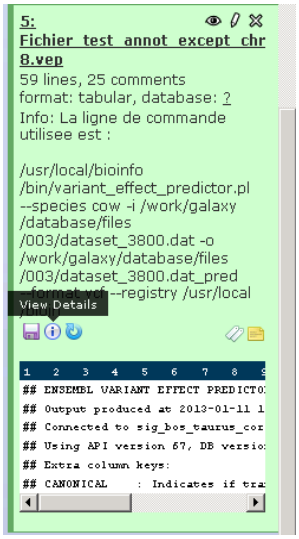
L'outil qui permet de faire ce formatage est l'outil « *Format SNP Effect - Protseq* ».

Lancement

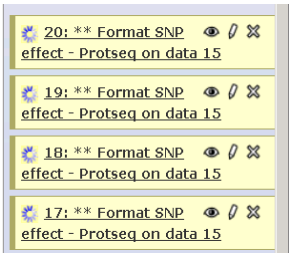


Attention utilisez le même nom de projet que pour « *Variant effect Predictor* » (cf : « USERNAME_TP_Annotation »).

Astuce double cliquez dans la zone de texte pour faire apparaître les noms de projets tapés précédemment ou cliquer sur votre fichier VEP initial, cliquer sur l'outil informations, vous trouverez les paramètres que vous avez utilisés pour générer ce fichier.



Résultats

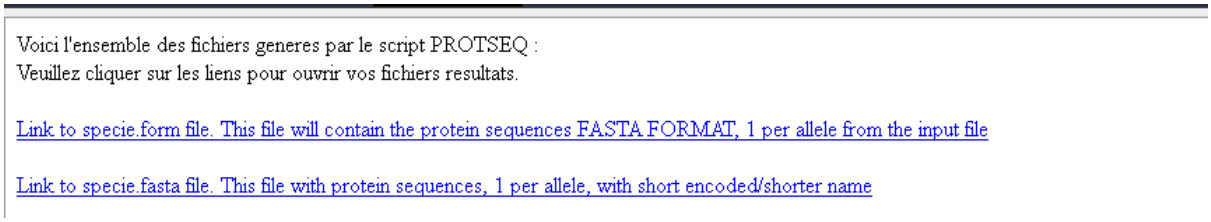


4 boites de dialogues sont lancées en même temps pour ce module.

En effet, Protseq va générer différents fichiers que vous allez devoir ensuite utiliser avec les modules de prédiction des conséquences.

Renommez vos fichiers en fonction de ce qu'ils contiennent :

– Le premier fichier généré est un fichier html bilan des différentes sorties du programme. Renommez le **« Protseq_bilan.html »**


Voici l'ensemble des fichiers generes par le script PROTSEQ :
Veuillez cliquer sur les liens pour ouvrir vos fichiers resultats.

Link to specie.form file. This file will contain the protein sequences FASTA FORMAT, 1 per allele from the input file

Link to specie.fasta file. This file with protein sequences, 1 per allele, with short encoded/shorter name

– Le second fichier est un fichier de séquences protéiques au format fasta. Renommez le **« sequences_prot.fasta »**

```
>_seq1
MWPLVVVLLLGSVRCGSAQLIFNAIKSVEYTLCNQTVVIPCFVNNVETKNITELYVRWKFKGENIFIFDGSQRMSKPSSNFSSAEIAPSELLRGIASLKMAKSDAVLGNYTCEVTELSREGETIIELKYRVVSWFSPNENIL
>_seq2
MWPLVVVLLLGSVRCGSAQLIFNAIKSVEYTLCNQTVVIPCFVNNVETKNITELYVRWKFKGENIFIFDGNQRMSKPSSNFSSAEIAPSELLRGIASLKMAKSDAVLGNYTCEVTELSREGETIIELKYRVVSWFSPNENIL
>_seq3
MWPLVVVLLLGSVRCGSAQLIFNAIKSVEYTLCNQTVVIPCFVNNVETKNITELYVRWKFKGENIFIFDGSQRMSKPSSNFSSAEIAPSELLRGIASLKMAKSDAVLGNYTCEVTELSREGETIIELKYRVVSWFSPNENIL
>_seq4
MWPLVVVLLLGSVRCGSAQLIFNAIKSVEYTLCNQTVVIPCFVNNVETKNITELYVRWKFKGENIFIFDGNQRMSKPSSNFSSAEIAPSELLRGIASLKMAKSDAVLGNYTCEVTELSREGETIIELKYRVVSWFSPNENIL
>_seq5
MSFVRVNRYGPRGGGRKTLKVKKKTSVKQEWDNTVTDLTVHRATPEDLIRRHEIHKSKNRALVHWELQEKALKRRWKKQKPEISNLEKRRLSIMKEILSDQYQLQDVLEKSDHLMATAKGLFVDFPRRRTGFPNVTMAPESS
```

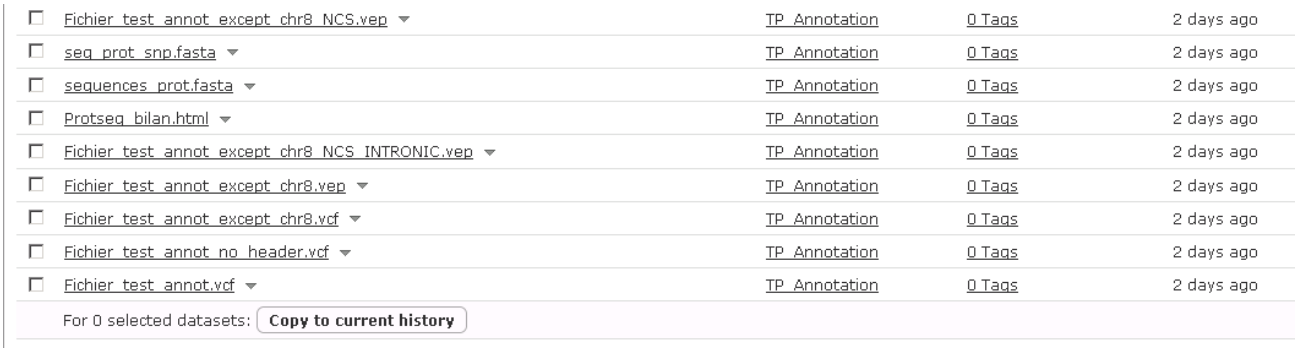
– Le troisième fichier permet la correspondance entre le nom des séquences protéiques et les différents allèles des SNPs « NON_SYNONYMOUS_CODING ». Renommez le **« seq_prot_snp.name. »**

```
>1_53156188_C/T_-_ENSBTAT00000004664[S/N-71]_allele1    _seq1   snp
>1_53156188_C/T_-_ENSBTAT00000004664[S/N-71]_allele2    _seq2   snp
>1_53156188_C/T_-_ENSBTAT00000055450[S/N-71]_allele1    _seq3   snp
>1_53156188_C/T_-_ENSBTAT00000055450[S/N-71]_allele2    _seq4   snp
>1_58593687_G/A_-_ENSBTAT00000057527[S/F-633]_allele1   _seq5   snp
```

– Le quatrième fichier est un fichier VEP (fichier de sortie de Variant Effect Predictor) qui ne contient **que** les SNPs « NON_SYNONYMOUS_CODING ». Renommons le **« Fichier_test_annot_except_chr8_NCS.vep »**

```
1_53156188_C/T  1:53156188      T       ENSBTAG00000003585      ENSBTAT00000004664      Transcript      NON_SYNONYMOUS_CODING   330     212
1_53156188_C/T  1:53156188      T       ENSBTAG00000003585      ENSBTAT00000055450      Transcript      NON_SYNONYMOUS_CODING   330     212
1_58593687_G/A  1:58593687      A       ENSBTAG00000013937      ENSBTAT00000057527      Transcript      NON_SYNONYMOUS_CODING   1977    1898
1_58792664_G/A  1:58792664      A       ENSBTAG00000021000      ENSBTAT00000027966      Transcript      NON_SYNONYMOUS_CODING   4856    4856
1_68929240_C/A  1:68929240      A       ENSBTAG00000000566      ENSBTAT00000000740      Transcript      NON_SYNONYMOUS_CODING   239     239
1_68929240_C/A  1:68929240      A       ENSBTAG00000000566      ENSBTAT00000054897      Transcript      NON_SYNONYMOUS_CODING   198     95
1_95205755_G/A  1:95205755      A       ENSBTAG00000037965      ENSBTAT00000052461      Transcript      NON_SYNONYMOUS_CODING   512     512
1_145178567_G/A 1:145178567     A       ENSBTAG00000017010      ENSBTAT00000022620      Transcript      NON_SYNONYMOUS_CODING   710     643
1_154285564_C/A 1:154285564     A       ENSBTAG00000040193      ENSBTAT00000030369      Transcript      NON_SYNONYMOUS_CODING   158     86
```

Bilan des dataSets enregistrés :

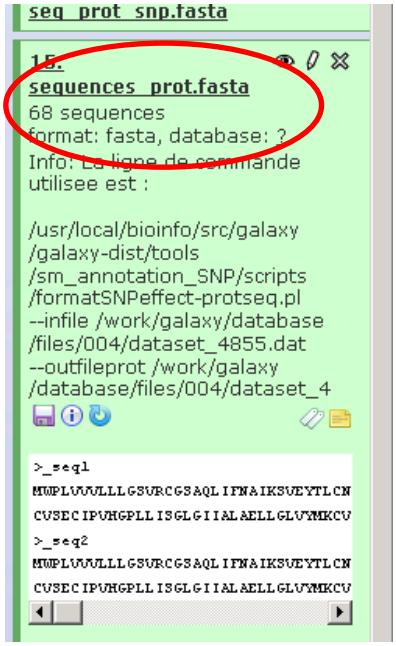


☐	Fichier_test_annot_except_chr8_NCS.vep	TP_Annotation	0 Tags	2 days ago
☐	seq_prot_snp.fasta	TP_Annotation	0 Tags	2 days ago
☐	sequences_prot.fasta	TP_Annotation	0 Tags	2 days ago
☐	Protseq_bilan.html	TP_Annotation	0 Tags	2 days ago
☐	Fichier_test_annot_except_chr8_NCS_INTRONIC.vep	TP_Annotation	0 Tags	2 days ago
☐	Fichier_test_annot_except_chr8.vep	TP_Annotation	0 Tags	2 days ago
☐	Fichier_test_annot_except_chr8.vcf	TP_Annotation	0 Tags	2 days ago
☐	Fichier_test_annot_no_header.vcf	TP_Annotation	0 Tags	2 days ago
☐	Fichier_test_annot.vcf	TP_Annotation	0 Tags	2 days ago

For 0 selected datasets: **Copy to current history**

Exercice

Combien y a-t-il de séquences dans le fichier fasta ?



- Pour connaître le nombre de séquence d'un fichier fasta

Il faut simplement cliquer sur le nom de ce fichier, et dans la brève description nous n'avons pas le nombre de lignes mais le nombre de séquences du fichier.

ETAPE 3 : Annotation fonctionnelle des SNPs « non-synonymous coding ».

Nous allons maintenant lancer les différents modules de prédictions des conséquences des SNPs « NON_SYNONYMOUS_CODING ».

Les fichiers de sortie indiqueront au niveau de la colonne « case » : « loss », « gain » ou vide. Les colonnes suivantes indiqueront les scores pour l'allèle de référence et l'allèle alternatif (sauf pour le module conservation). La perte « loss » d'une modification post-traductionnelle correspond à l'allèle 1 avec signal prédit et allèle 2 sans signal prédit pour le module d'annotation fonctionnelle concerné.

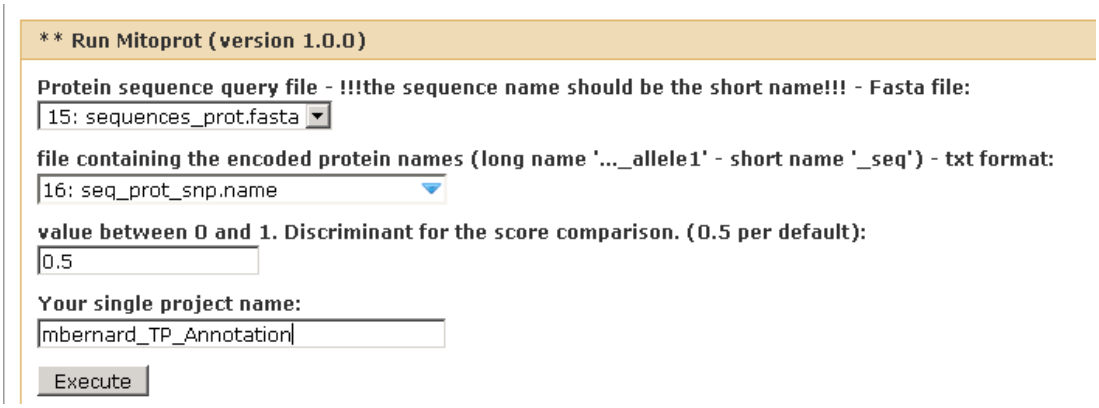
Le gain « gain » d'une modification post-traductionnelle correspond à l'allèle 2 avec signal et allèle 1 sans signal. A cette comparaison s'ajoute une comparaison des scores. Dans le cas ou le programme génère un score de prédiction mais, sans fournir de signal fort; si les 2 allèles ont fourni un tel score et que la différence de score est supérieure ou égale au seuil minimum fixée en paramètre (0.5), alors, une sortie avec « perte? » ou « gain? » est indiquée ainsi que les scores des allèles 1 et 2 ayant permit de générer l'hypothèse de prédiction.

Pour chaque module, sont générés :
- un fichier contenant les résultats « bruts » du programme
- un fichier tabulaire contenant les résultats interprétés que l'on nommera « [nom_de_module]_results.txt »
- un fichier contenant les titres des colonnes des résultats que l'on nommera « [nom_de_module]__title.txt ».

Les fichiers « _results.txt » contiennent la liste des SNPs avec les gains ou pertes de signaux. Le fichier « title » contient les noms des colonnes des informations trouvées dans le fichier « results».

1) Mitoprot

Pour étudier les signaux d'adressage à la mitochondrie, sélectionnez l'outil « *run_Mitoprot ».*



Les fichiers de sorties mitoprot

Renommez les fichiers de sorties en fonction des exemples ci dessous

- Extrait du fichier contenant les titres des colonnes de résultats « mitoprot_title.txt »

```
Sequence Name    Case    score allele1    score allele2
```

- Extrait du fichier contenant les résultats formatés « mitoprot_results.txt » :

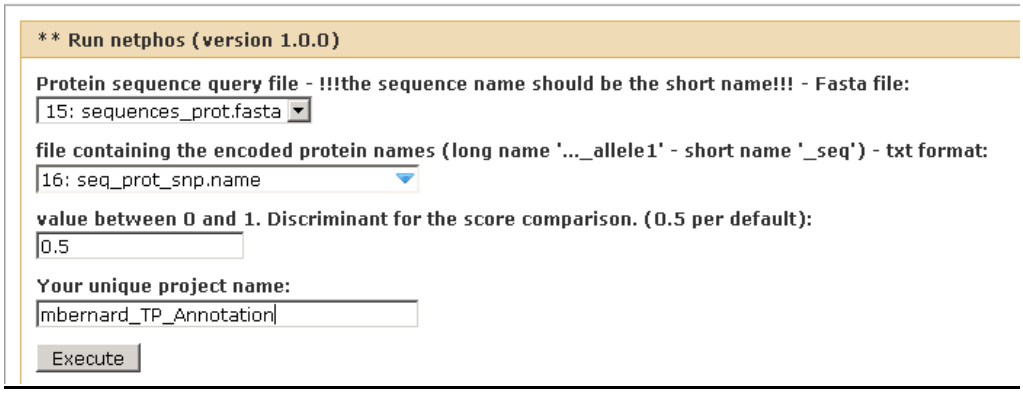
```
1_68929240_C/A_-_ENSBTAT00000054897[W/L-32]     loss    0.3961  0.4230
1_154285564_C/A_-_ENSBTAT00000030369[S/I-29]    loss    0.3758  0.4093
1_42531330_C/T_-_ENSBTAT00000005398[E/K-29]     loss    0.0900  0.6499
2_126676995_C/T_-_ENSBTAT00000055788[P/S-17]    gain    0.3499  0.3855
3_25969298_G/T_-_ENSBTAT00000039902[P/T-21]     loss    0.2367  0.3668
4_94912500_C/A_-_ENSBTAT00000017924[R/S-11]     loss    0.7161  0.3288
4_106869716_G/A_-_ENSBTAT00000064277[E/K-4]     gain    0.1166  0.4817
4_116237864_G/C_-_ENSBTAT00000011851[W/S-9]     loss    0.6446  0.4435
4_12724969_A/G_-_ENSBTAT00000065396[S/G-27]     loss    0.6108  0.5940
13_61584514_T/A_-_ENSBTAT00000027390[L/Q-11]    gain    0.1882  0.2982
```

- Extrait du fichier contenant les résultats « mitoprot_bruts.txt » :

```
>1_53156188_C/T_-_ENSBTAT00000004664[S/N-71]_allele1    not imported
0.2119
>1_53156188_C/T_-_ENSBTAT00000004664[S/N-71]_allele2    not imported
0.2119
>1_53156188_C/T_-_ENSBTAT00000055450[S/N-71]_allele1    not imported
0.2119
>1_53156188_C/T_-_ENSBTAT00000055450[S/N-71]_allele2    not imported
0.2119
>1_58593687_G/A_-_ENSBTAT00000057527[S/F-633]_allele1   imported
0.9538
>1_58593687_G/A_-_ENSBTAT00000057527[S/F-633]_allele2   imported
0.9538
```

2) Netphos

Pour prédire s'il y a acquisition ou perte de sites de phosphorylation, sélectionnez l'outil « *run_Netphos* ».



Les fichiers de sorties netphos

Renommez les fichiers de sorties en fonction des exemples ci-dessous.

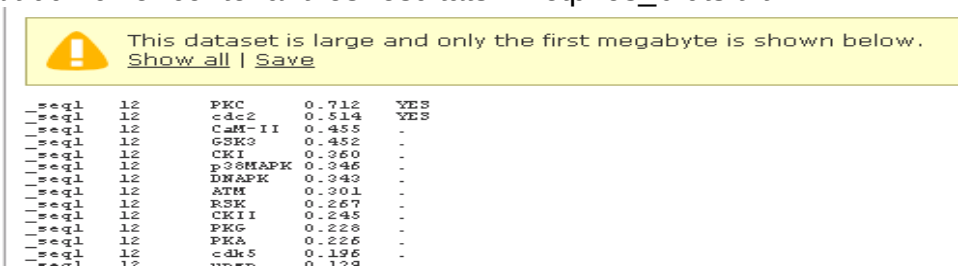
- Extrait du fichier contenant les titres des colonnes de résultats « netphos_title.txt »

```
Name     ATM result       ATM allele1 score        ATM allele2 score        CKI result
```

- Extrait du fichier contenant les résultats formatés « netphos_results.txt » :

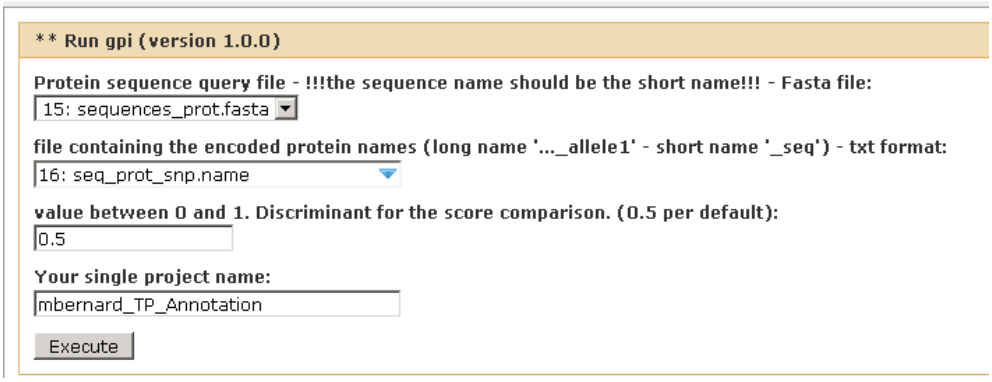
```
1_53156188_C/T_-_ENSBTAT00000055450[S/N-71]
1_58593687_G/A_-_ENSBTAT00000057527[S/F-633]     loss     0.553
1_58792664_G/A_-_ENSBTAT00000027966[S/F-1619]    loss     0.529
1_68929240_C/A_-_ENSBTAT00000000740[W/L-80]
1_68929240_C/A_-_ENSBTAT00000054897[W/L-32]
1_95205755_G/A_-_ENSBTAT00000052461[S/N-171]     loss     0.546
2_126676995_C/T_-_ENSBTAT00000055788[P/S-17]
2_133321047_C/T_-_ENSBTAT00000017333[S/F-83]     loss     0.548
2_17916660_G/A_-_ENSBTAT00000040194[G/S-342]     gain             0.546                          gain             0.536
2_85259069_G/A_-_ENSBTAT00000038307[P/S-441]
3_100498539_A/C_-_ENSBTAT00000017715[L/R-240]
```

- Extrait du fichier contenant les résultats « netphos_bruts.txt » :



3) Gpi

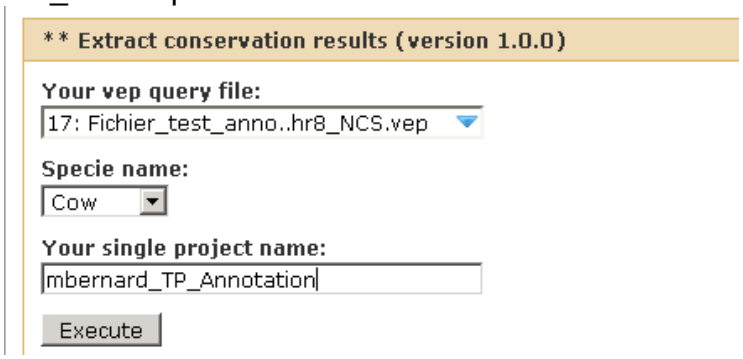
Pour analyser les changements d'ancrage GPI, selectionnez l'outil « Run_gpi ».



GPI produit 2 fichiers de sorties, un fichier title et un fichier « résults » comme précédemment.

4) Conservation

Pour analyser la conservation protéique entre référence et alternatif, utilisez l'outil « extract_conservation_results.pl »



Attention cette fois ci nous utilisons le fichier VEP produit par l'outil de formatage Protseq. Ce VEP ne contient que les NON_SYNONYMOUS_CODING SNPs :
« Fichier_test_annot_except_chr8_NCS.vep ».

Conservation produit 2 fichiers de sorties, un fichier « title » et un fichier « résults » comme précédemment, renommez les.

Pour cet outil, les résultats sont constitués d'un tableau à 4 colonnes:

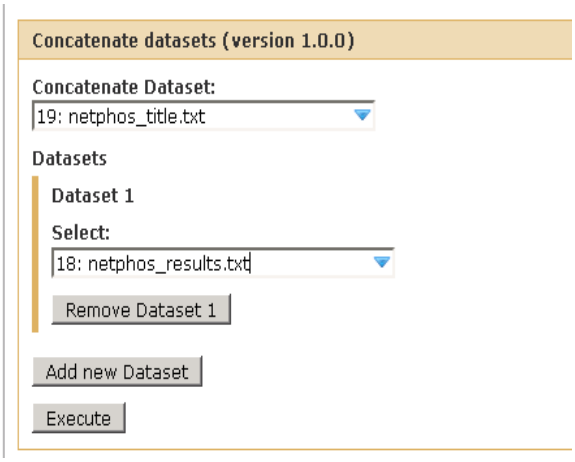
- 1 colonne pour le résultat généré à partir de la matrice Grantham de conservation des propriétés physico chimique au sein de la protéine entre référence et alternatif :
  - conservées (0-50)
  - modérément conservées (51-100),
  - modérément radicales (101-150)
  - radicales (≥ 151)
- et 3 colonnes concernant les résultats générés à partir de la base Ensembl (score attendu, score observé et différence de score) qui indique un taux de conservation de la mutation par rapport aux protéines orthologues.
  - Plus le score de différence est élevé, plus la position est conservée et donc plus une mutation à cette position a de fortes chances d'être délétère.

Exercice

Fusionnez tous les fichiers « title » avec leur correspondant « results ».
Renommez les par exemple : **netphos_final, mitoprot_final, conservation_final,** et **gpi_final.**

- La concaténation de plusieurs fichier se fait grâce à l'outil « Concatenate datasets »

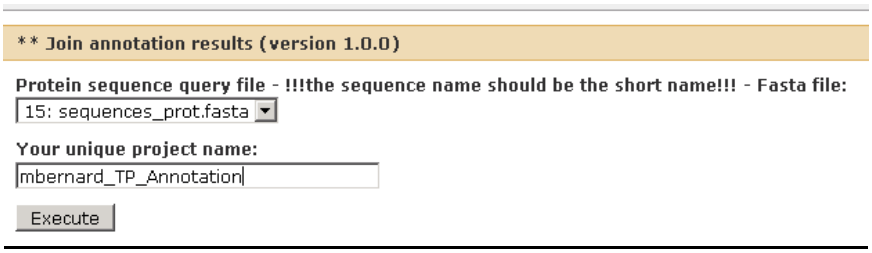
Selectionnez un fichier « title » suivi du fichier « result » correspondant, exécutez l'outil et renommez le résultat.



Etape 4 : fusion et export des résultats.

Join

Le choix du fichier d'entrée sert uniquement à retrouver le chemin du répertoire de projet « USERNAME_TP_Annotation », vous pouvez par exemple utiliser le fichier fasta des séquences protéiques « sequences_prot.fasta ».



Remarque

Cette fois nous avons 48 lignes dans notre fichier de sortie. Tous les SNPs contenus dans notre fichier VEP sont retournés, mais seulement les SNPs « NON_SYNONYMOUS_CODING » ont des résultats de score.

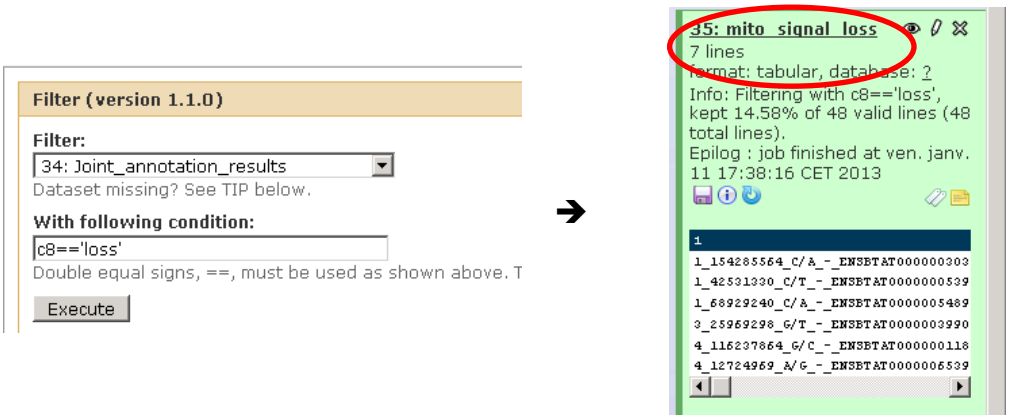
Le fichier de résultats est un fichier de type texte. Pour pouvoir faire des sélections sur les colonnes il est nécessaire de changer le type du fichier en « tabular ».
Utilisez l'outil « pencil » et sélectionnez « tabular » dans « New Type ».

Exercice:

Combien de SNPs provoquent une perte de signalisation à la mitochondrie ?
Combien de SNPs ont un score de conservation de Grantham > 100 ?

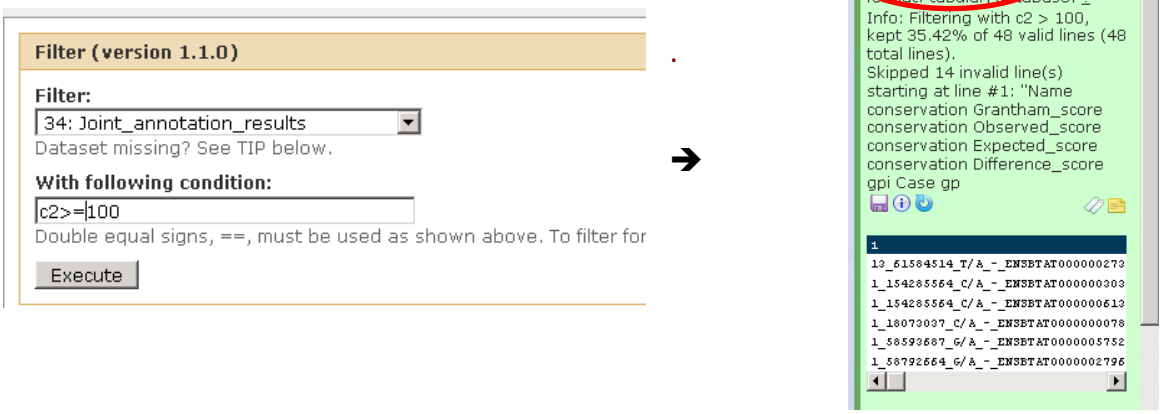
- Combien de SNPs provoquent une perte de signalisation à la mitochondrie ?

La sélection de la colonne correspondant aux résultats de perte de signalisation à la mitochondrie doit se faire grâce à une recherche du mot « loss » sur la 8e colonne, l'outil « Filter » est dédié à ce type de recherche.



- Combien de SNPs ont un score de conservation de Grantham > 100 ?

Cette fois ci il s'agit de faire une comparaison numérique sur la 2e colonne. L'outil « Filter » est toujours l'outil à utiliser.



Attention de ne pas mettre de côte lorsqu'il s'agit d'une comparaison numérique.

ETAPE 5 : Automatisation de l'analyse

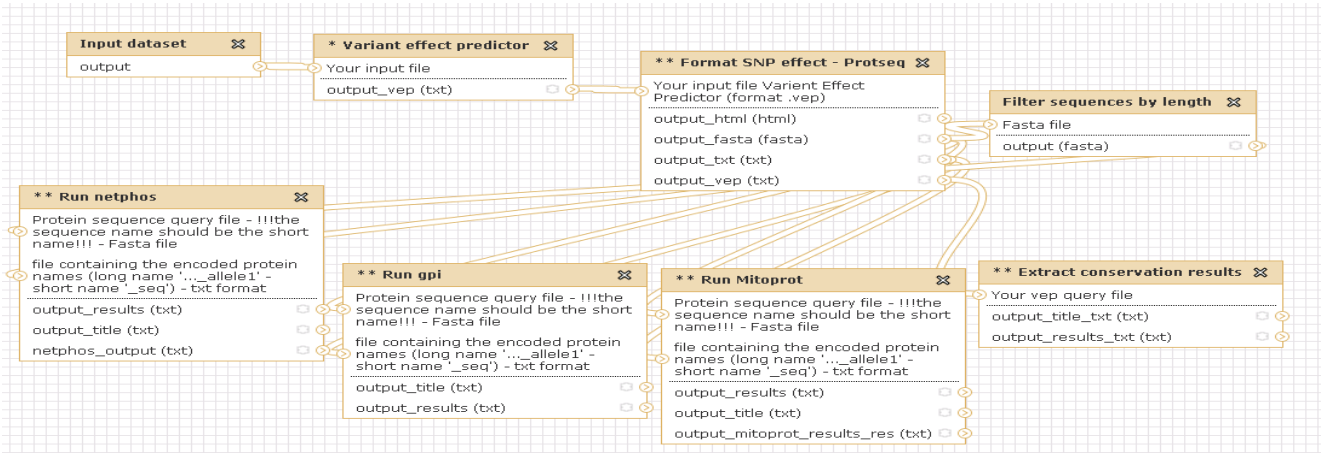
L'utilisation de workflow dans Galaxy (démo)

Un workflow est déjà disponible pour faire toutes ces analyses.

Dans l'onglet workflow, en haut à droit cliquez sur « upload or import workflow », et entrez l'URL suivante : http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/Annotation_SNP/Galaxy-Workflow-Pipeline_annotation.ga

Le pipeline automatisé apparaît dans votre onglet workflow.

Cliquez sur Edit (flèche à côté du nom du workflow), pour voir les différents éléments qui le composent et l'enchainement des traitements.



Tous les programmes d'analyses que nous avons lancé précédemment se retrouvent dans le workflow, sauf le programme Join qu'il faut ensuite lancer manuellement à la fin..

Pour lancer le workflow, créer un nouvel historique et réimporter notre fichier d'entrée VCF. Sur l'onglet workflow, sélectionner « run » du pipeline d'annotation. Sélectionné votre VCF ainsi qu'un nouveau nom de projet, enfin cliquez sur « Run workflow ».

ETAPE 6 : Conclusion

- Tous les résultats intermédiaires sont sur le site :

http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/Annotation_SNP/

- Source vers les sites des modules
  - Mitoprot
    http://ihg2.helmholtz-muenchen.de/ihg/mitoprot.html
    MITOPROT: Prediction of mitochondrial targeting sequence

  - Netphos
    http://www.cbs.dtu.dk/services/NetPhos/
    Sequence and structure based prediction of eukaryotic protein phosphorylation sites.

  - GPI
    http://gpi.unibe.ch/
    GPI-SOM: Identification of GPI-anchor signals by a Kohonen Self Organizing Map

  - Conservation
    - Matrice Grantham

    Amino acid difference formula to help explain protein evolution.
    - API Ensembl Compara

    http://www.ensembl.org/info/docs/api/compara/index.html

- Installation galaxy et workflow SNP Annotation

Si votre laboratoire installe une instance de Galaxy et que vous désirez installer ces outils et ce workflow vous trouverez toutes les instructions en tapant cet commande :
hg clone http://inra-sigenae-sarah-maman@toolshed.g2.bx.psu.edu/repos/inra-sigenae-sarah-maman/snp_annotation