EuGene

Pipeline automatique pour annoter des génomes eucaryotes

eugene.toulouse.inra.fr

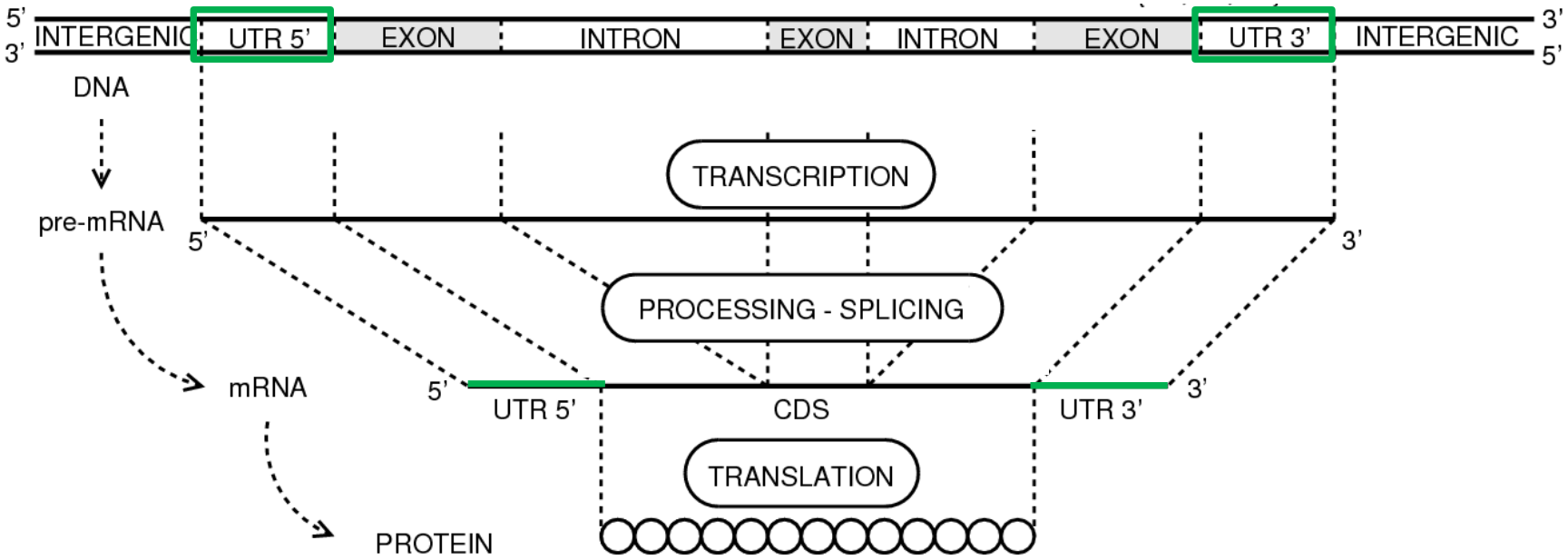
Erika Sallet [*], Jérôme Gouzy [*], Thomas Schiex[+]

(*) Laboratoire Interaction Plantes-Microorganismes (LIPM) Toulouse

(+) Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT)

Annotation structurale des gènes

- Codant pour des protéines
  - Intron/exon, CDS, UTR 5’ et 3’



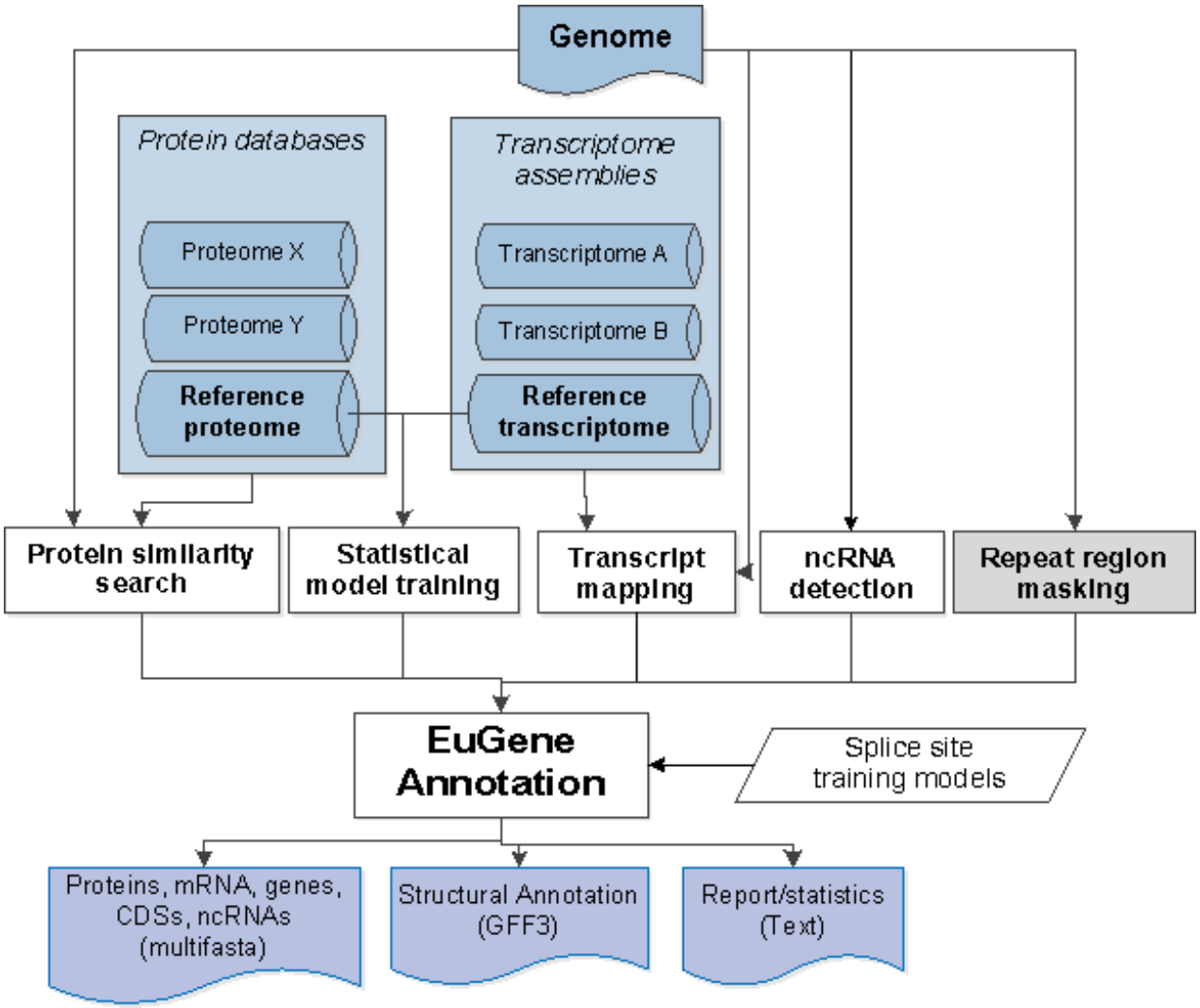
- Non codant
  - tRNA, rRNA, ncRNA

Objectifs

- Annoter **automatiquement, simplement** et **rapidement** un **génome** eucaryote

- Automatiser entièrement l'annotation
  - Réduire autant que possible le paramétrage manuel
- Optimiser les temps d'exécution
  - Optimiser les protocoles de certaines étapes
  - Paralléliser certaines tâches
  - Faciliter la 'reprise'
- Supprimer la dépendance à des logiciels sous licence

Etapes principales de EuGene



Stratégie d'intégration des données RNAseq

- Nécessite un pré-assemblage pour intégrer de la même façon les EST Sanger, le RNAseq illumina et les IsoSeq PacBio
  - Attention a bien nettoyer les assemblages trinity par exemple

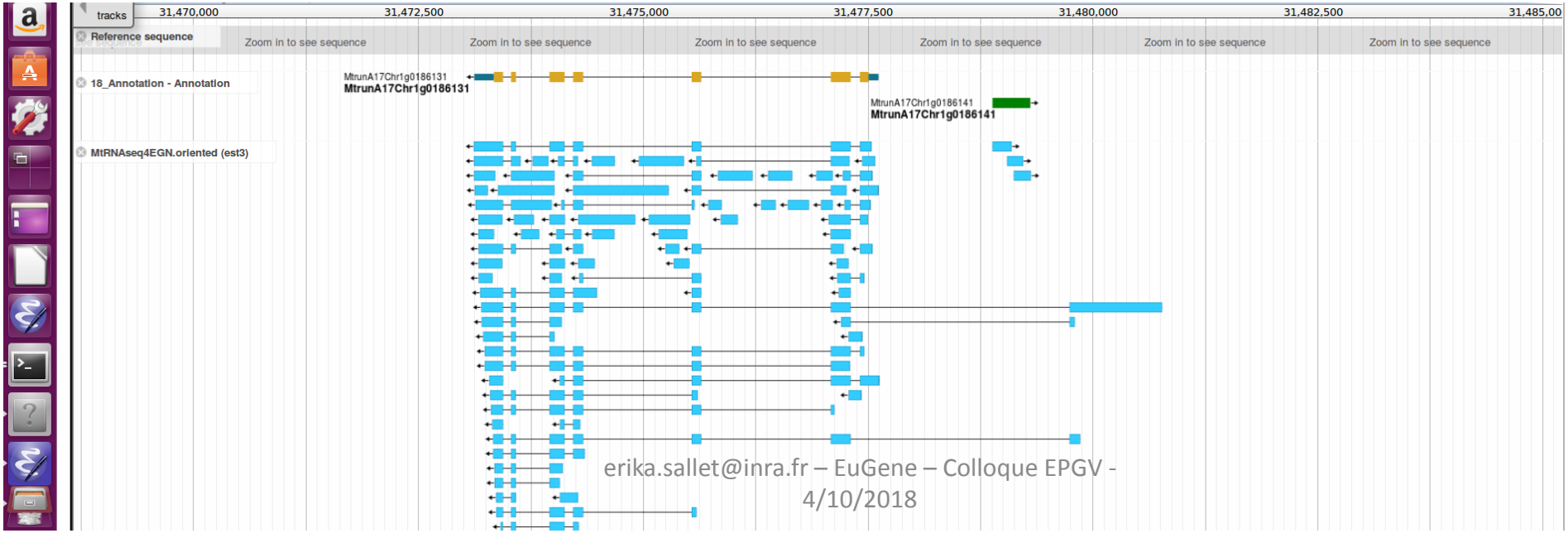
- Alignement avec *gmap* des transcrits assemblés http://research-pub.gene.com/gmap

- On donne beaucoup plus de poids à la prédiction **supportée par les données RNAseq** qu'à la prédiction *ab initio* (c'est-à-dire uniquement basée sur les modèles statistiques)

- On accorde plus de poids si les transcrits alignés sont **épissés**

Stratégie d'intégration des données RNAseq

- Gestion des incohérences locales
  - Filtrage des alignements des transcrits lorsqu'il y a des incohérences, dues par exemple à la présence de variants d'épissage ou à des ARNm partiellement maturés.
  - On privilégie **les introns les plus représentés**



Masquage des éléments répétés

- EuGene n'a pas pour objectif d'annoter les éléments répétés
  - D'autres outils sont dédiés à cela
  - On veut éviter d'annoter les gènes associés aux éléments transposables

  = > On veut travailler sur un **génome masqué**

- Principe :
  - **Masquage** sur le génome des éléments répétés détectés par:
    - Red
    - LTRharvest
    - BlastX contre une banque d'éléments repétés (= RepBase + protéines TE spécifiques de l'espèce)
  - **Démasquage** des régions transcrites « protégées » (hits avec une banque de transcrits ou avec une banque protéique, ou ncRNA)

- Annotation du génome masqué

Etapes complètement automatisées

- L'entrainement des modèles statistiques (IMM) pour différencier le codant du non codant

- L'entrainement des modèles statistiques (WAM) pour la détection des sites d'épissage
  - Construction de modèles « plantes dicot » à partir des sites d'épissage de plusieurs plantes (Ha, Rosa, At, Mt)
  - Modèles entrainés actuellement disponibles :
    - plantes (dicot), champignons, nématodes, oomycetes

- La détection de sites d'épissage 'non canoniques'
  - Recherche dans les résultats de mapping, au frontière exon/intron, des sites autres que GT/AG
  - Si un site est trouvé dans plus de 1% des cas (par défaut), on l'autorise

Optimisation des temps d'exécution

- Gmap est multithreadé et très rapide

- Optimisations de BlastX
  - Sur pseudomolécule : parallélisation via fenêtres glissantes
  - Sur petit scaffold : parallélisation en lançant plusieurs blastx en parallèle

    Remarque : les deux types de sequences peuvent être mixés
  - Sur chaque fenêtre on lance **ublast** pour sélectionner les protéines pouvant matcher et on lance **blastx** sur la sous banque sélectionnée
    - Ublast seul ne donne pas des frontières d'HSPs suffisamment fiables

– Lorsqu'un calcul est terminé, on crée un fichier .success. Si on relance le pipeline, le calcul n'est pas relancé si le fichier .success existe.
  - Utile par exemple si on ajoute une banque de transcrits ou protéique

Pour aller plus loin

- Intégration de tout type d'information supplémentaire
  - Exemples : mapping de données issues d'analyse protéomique, résultat de recherche de TSS (Transcription Start Site)
  - Un fichier GFF3 et un fichier de configuration additionnel

- Deux annotations « brin indépendant »
  - Permet la détection de gènes chevauchants sur des brins opposés, de ncRNA antisense.

- Configuration avancée :
  - Taille minimum des introns
  - Table d'usage des codons, ..

Utilisation simple

- Un unique fichier de configuration pour renseigner :

  1) les bases de données protéiques et les assemblages RNA-seq

```
# List the protein database numbers for blastX
blastx_db_list= 1 2 4

blastx_db_1_file=/path/of/myproteome1.fasta
blastx_db_1_weight=0.3
blastx_db_1_pcs=50
blastx_db_1_remove_repet=0
blastx_db_1_preserve=1
blastx_db_1_training=1

# List the trancriptome numbers
est_list=2 1

est_1_file=/path/of/mytranscriptome1.fasta
est_1_pcs=30
est_1_pci=97
est_1_remove_unspliced=0/1/2
est_1_preserve=1
est_1_training=1
```

Utilisation simple

2) d'éventuels informations complémentaires

```
additional_list=1
additional_1_file=%i/ADDITIONAL/mygenome.MQ.peptides.gff3
additional_1_cfg_template=%i/ADDITIONAL/plugin_AnnotaStruct_Proteomics.cfg
```

3) les chemins et paramètres des programmes

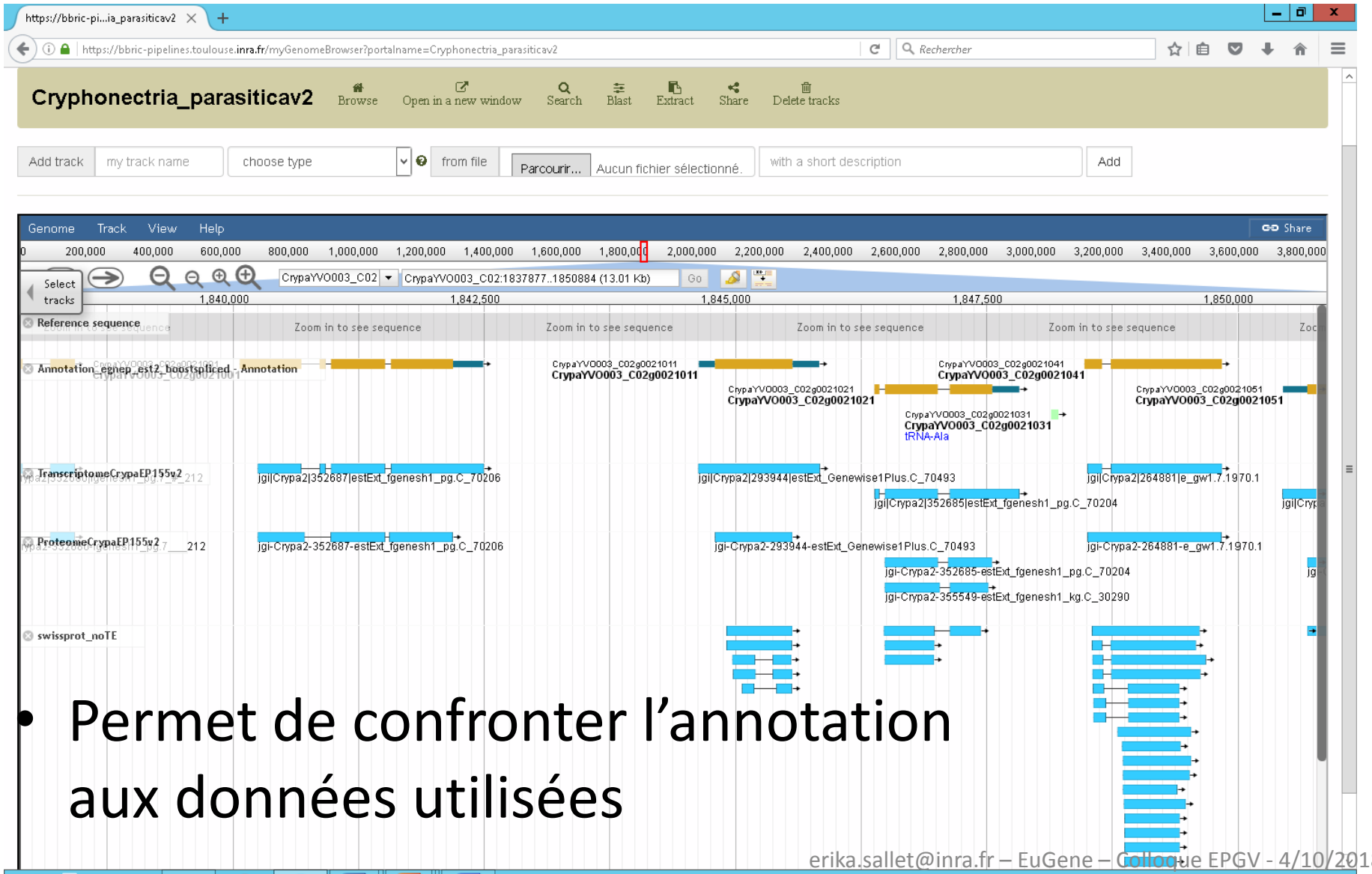
Une ligne de commande :

```
$EGNEP/bin/int/egn-euk.pl \
    --indir        /path/of/myindir \
    --outdir       /path/of/myoutdir \
    --workingdir   /path/of/myworkdir \
    --cfg          /path/of/myeugeneconffile.cfg
```

Résultats

- Annotation au format GFF3
- Fichiers fasta des gènes, mRNA, CDS, protéines, ncRNA
- Fichier de comptage
  - nombre de gènes, taille moyenne des gènes, GC% des régions, etc
- Information sur les étapes d'annotation
  - %age de transcrits alignés
  - Sites non canoniques détectés
  - %age de la séquence masqué par des régions répétées

Utilitaire pour transformer les fichiers créés par le pipeline en Genome Browser



- Permet de confronter l'annotation aux données utilisées

« Companion tools »

- **egn_build_wam.pl** : programme pour construire de nouveaux modèles statistiques pour la détection des sites d'épissage
  - Cas d'utilisation : aucun des modèles statistiques entrainés disponibles ne correspond à l'espèce que je veux annoter
  - Données nécessaires : plusieurs génomes complets + pour chacun d'eux un transcriptome de bonne qualité et suffisamment complet

- **egn_annotation_transfer.pl** : programme pour transférer une annotation sur une autre séquence
  - Objectif : transférer une annotation entre génotypes/souches d'une même espèce en préservant la correspondance entre nom de gène
  - Pas de découverte de gène
  - Les gènes annotés sur la nouvelle séquence ont une structure valide

Limitations

- Le modèle de gène actuel ne permet pas d'annoter un ncRNA ou un gène à l'intérieur d'un intron sur le même brin.

- Même si EuGene peut le faire, la configuration par défaut du pipeline n'autorise pas la prédiction de variants d'épissage et ne gère pas la présence de "frameshift" dans les CDS.

- EuGene n'est pas adapté pour annoter des séquences de chloroplastes ou de mitochondries.

Plus d'information

- Le protocole détaillé d'installation et d'utilisation sera bientôt disponible dans un volume de Methods in Molecular Biology

- Jusqu'à présent, nous avons formé une dizaine de personnes à l'utilisation du pipeline EuGene
  - Formation individuelle ou en petit groupe à Toulouse de 1 jour
  - Chacun configure le pipeline et l'exécute sur ses données (génome, transcriptomes, etc)

- Pour nous contacter, écrivez à eugene-help@groupes.renater.fr

- Pour recevoir des news, abonnez vous à : https://groupes.renater.fr/sympa/info/eugene-info